2

Statistik dan Ekonometrika Terapan

Aplikasi dengan STATA

Moch. Doddy Ariefianto
Irwan Trinugroho

UNDANG-UNDANG REPUBLIK INDONESIA
NOMOR 19 TAHUN 2002
TENTANG HAK CIPTA

PASAL 72
KETENTUAN PIDANA
SANKSI PELANGGARAN

1. Barangsiapa dengan sengaja dan tanpa hak mengumumkan atau memperbanyak suatu Ciptaan atau memberikan izin untuk itu, dipidana dengan pidana penjara paling singkat 1 (satu) bulan dan/atau denda paling sedikit Rp1.000.000,00 (satu juta rupiah), atau pidana penjara paling lama 7 (tujuh) tahun dan/atau denda paling banyak Rp5.000.000.000,00 (lima miliar rupiah).

2. Barangsiapa dengan sengaja menyerahkan, menyiarkan, memamerkan, mengedarkan, atau menjual kepada umum suatu Ciptaan atau barang hasil pelanggaran Hak Cipta atau Hak Terkait sebagaimana dimaksud pada ayat (1), dipidana dengan pidana penjara paling lama 5 (lima) tahun dan/atau denda paling banyak Rp500.000.000,00 (lima ratus juta rupiah).

2

Statistik dan Ekonometrika Terapan

Aplikasi dengan STATA

Moch. Doddy Ariefianto
Irwan Trinugroho

PENERBIT
PT. Penerbit Erlangga
Jl. H. Baping Raya no. 100
Ciracas, Jakarta 13740
Website: www.erlangga.co.id
(Anggota IKAPI)

007-310-005-0
978-623-266-550-7

STATISTIK DAN EKONOMETRIKA TERAPAN
Aplikasi dengan STATA
Jilid 2

Disusun oleh:
Dr. Moch. Doddy Ariefianto
Irwan Trinugroho, PhD

Editor:
Suryadi Saat

Desain Cover:
Maya Kumala

Buku ini diset dan dilayout oleh Bagian Produksi ***Penerbit Erlangga***
dengan Power MacPro

24 23 22 21 4 3 2 1

ENDORSEMENT

Ilmu ekonomi dan sosial berkembang makin pesat. Fenomena sosial juga berubah dengan cepat. Perubahan itu tidak lagi bisa hanya mengandalkan kepada penilaian proses dan pola yang berjalan. Dibutuhkan pula pengetahuan yang mendalam soal dampak dan proyeksi atas suatu perubahan. Pada titik inilah ilmu sosial (khususnya ekonomi) membutuhkan perangkat kuantitatif untuk membantu merumuskan fenomena yang terjadi. Statistik dan ekonometrik merupakan perkakas yang tidak lagi bisa dihindari pemanfaatannya demi memperoleh analisis yang presisi. Buku bagus ini merupakan salah satu instrumen yang hadir untuk memperbaiki kualitas analisis maupun kebijakan. Poficiat!

Ahmad Erani Yustika
Guru Besar Fakultas Ekonomi dan Bisnis - Universitas Brawijaya

Buku ini disusun bagi akademisi maupun praktisi yang ingin langsung mempraktikkan permodelan ekonometrika dengan STATA yang sudah menjadi perangkat standar untuk berbagai studi empiris di bidang ekonomi, keuangan, dan ilmu sains lainnya. Penulis sudah berhasil menyajikan konsep statistik dan metode ekonometrika yang kompleks dan rumit menjadi lebih sederhana dan mudah dicerna bagi pemula atau yang berpengalaman dan butuh referensi cepat. Semoga dengan kehadiran buku ini bisa mendorong riset-riset yang lebih berkualitas ke depan.

Inka B. Yusgiantoro, B.S.E., M.S., M.Sc., M.A., Ph.D.
Direktur Eksekutif, Kepala Departemen Riset Jasa Keuangan, Otoritas Jasa Keuangan

Buku ini merupakan bacaan yang lengkap dan ditulis dengan lugas untuk memberikan tuntunan penerapan berbagai metodologi ekonometrika kontemporer dalam analisis dan riset di bidang ekonomi dan keuangan. Tidak hanya sekedar berbagai contoh kasus penggunaan metodologi ekonometrika, buku ini juga memberikan landasan teori-teori yang relevan sebelum mengulas setiap contoh tersebut, yang memudahkan bagi pembaca baik yang masih relatif baru belajar ekonometrika maupun yang sudah berpengalaman. Software STATA, di antara berbagai pilihan software ekonometrika lainnya, juga merupakan pilihan yang bijak mengingat betapa banyak modul-modul ekonometrika dalam software ini, yang juga ditunjang dengan komunitas para peneliti pengguna STATA yang aktif dari berbagai belahan dunia. *Great works Dr. Doddy dan Dr. Irwan!*

Herman Saheruddin, Ph.D.
Direktur, Group Riset , Lembaga Penjamin Simpanan

TENTANG PENGARANG

Moch. Doddy Ariefianto (Doddy) adalah Dosen tetap Prodi Magister Akuntansi, Binus University sejak Agustus 2019. Doddy memperoleh gelar sarjana bidang ilmu Ekonomi Pembangunan dari Universitas Brawijaya pada tahun 1999. Gelar Magister dan Doktor Ilmu Ekonomi kemudian diraihnya pada tahun 2005 dan 2010 dari Universitas Indonesia. Di samping menjadi akademisi; Doddy juga telah berkarier sebagai praktisi selama lebih dari 18 tahun di Bank Mandiri (posisi terakhir sebagai Senior Economist) dan Lembaga Penjamin Simpanan (posisi terakhir sebagai Direktur Group Surveilans dan Stabilitas Sistem Keuangan merangkap Pgs Direktur Group Riset). Doddy telah menulis lebih dari 60 artikel di media populer dan ilmiah. Minat riset Doddy adalah Perbankan, Pasar Keuangan, Keuangan Korporasi, Keuangan Internasional, dan Regulasi Keuangan. Artikel ilmiahnya telah dimuat pada publikasi internasional dan nasional bereputasi seperti Economic Systems, Research in Business and Finance, Borsa Istanbul Review, Economic Bulletin, International Journal of Economic and Management, Bulletin of Monetary Economics and Banking, Indonesian Capital Market Review, Ventura dan Jurnal Keuangan Perbankan. Buku yang telah diterbitkan adalah Ekonometrika: Esensi dan Aplikasi dengan Menggunakan Eviews (Erlangga, 2012).

Irwan Trinugroho (Irwan) adalah dosen tetap di Fakultas Ekonomi dan Bisnis, Universitas Sebelas Maret (UNS), Surakarta. Irwan saat ini juga merupakan kepala Pusat Unggulan Iptek-Perguruan Tinggi (PUI-PT) Center for Fintech and Banking UNS serta Direktur Kerjasama Pengembangan, dan International UNS. Irwan memperoleh gelar sarjana bidang Manajemen, dengan konsentrasi Manajemen Keuangan, dari Universitas Sebelas Maret tahun 2006. Gelar Master of Science di bidang Manajemen, konsentrasi Manajemen Keuangan diperoleh dari Universitas Gadjah Mada pada tahun 2009, sedangkan gelar doktor bidang Perbankan dan Keuangan diperoleh dari University of Limoges, Perancis pada tahun 2014. Irwan produktif dalam menulis artikel ilmiah di jurnal internasional bereputasi di antaranya terpublikasi di Journal of Financial Stability, British Accounting Review, Global Finance Journal, Economic Systems, Research in International Business and Finance, Economics Bulletin, Emerging Markets Finance and Trade, Borsa Istanbul Review, Journal of Asia Business Studies, dan Singapore Economic Review. Minat riset Irwan adalah perbankan, finance and development, fintech dan digital finance, penjaminan simpanan serta corporate governance, dan institutional development. Irwan terpilih menjadi dosen berprestasi UNS pada tahun 2015 dan 2018 serta terpilih sebagai finalis dosen berprestasi nasional pada tahun 2018.

KATA PENGANTAR

Alhamdulillah dengan rahmat Allah SWT, kami dapat menyelesaikan buku berjudul "Statistik dan Ekonometrika Terapan: Aplikasi dengan STATA". Penulis memiliki passion yang tinggi terhadap statistika dan ekonometrika karena keduanya adalah "ilmu mengenai ilmu". Statistik dan ekonometrika memberikan panduan bagi kita untuk melakukan prosedur analisis secara sistematis dan ilmiah dalam mendukung atau memfalsifikasi suatu prinsip-teori di bidang ilmu tertentu (misalnya, ekonomi) dengan membandingkannya dengan fakta-data yang ada di lapangan (empirisme).

Terdapat banyak buku mengenai statistik dan ekonometrika, namun buku yang memiliki karakter "menyederhanakan" serta aplikatif dapat dikatakan cukup jarang. Istilah "menyederhanakan" ini merupakan suatu tantangan tersendiri. Harus diakui mempelajari statistik dan ekonometrika bukan pekerjaan mudah; selain tebal buku-bukunya juga penuh dengan simbol matematik, yang sering diistilahkan sebagai "cacing". Banyak sekali konsep yang abstrak dan membutuhkan perenungan berhari-hari untuk memahami apa maksudnya. Penulis berharap buku ini dapat menyampaikan "penyederhanaan" tersebut tanpa menimbulkan "misleading", sehingga para pembaca dapat lebih fokus pada substansi teori-prinsip keilmuan yang hendak diteliti.

Buku ini merupakan hasil sintesis pengalaman praktisi penelitian dan mengajar para penulis selama lebih dari 10 tahun. Dengan pengalaman tersebut, penulis berharap dapat memenuhi harapan tugas "menyederhanakan" ini. Pembahasan secara matematik-statistik tidak terelakkan; namun penulis berharap narasi yang dibuat cukup dapat menjelaskan apa yang dimaksud. Penulis juga selalu mencantumkan rujukan literatur bagi pembaca yang mungkin masih belum cukup jelas, atau ingin menelusuri konsepnya lebih lanjut untuk pendalaman.

Perkembangan teknologi komputasi yang sangat pesat telah melahirkan banyak sekali varian teknik analisis (disebut sebagai estimator) yang beradaptasi dengan asumsi dan kondisi dari desain empiris (pertanyaan riset) yang dimiliki peneliti. Dalam hal ini, STATA memiliki keunggulan kompetitif di mana terdapat cukup banyak komunitas atau individu yang menggunakan software tersebut, dan sangat aktif menghasilkan modul-modul baru (serta membaginya dengan gratis). Tentu saja, hal ini sangat menguntungkan bagi para peneliti. Membangun suatu modul statistik-ekonometrika membutuhkan tidak hanya pengetahuan tentang coding, tetapi juga pemahaman teori matematika statistika yang mendalam. Proses membangun modul juga memakan sumber daya finansial, energi, dan waktu yang tidak sedikit. Tersedianya komunitas yang aktif dan "sharing" seperti yang dimiliki STATA tentu merupakan nilai tambah yang sangat berharga.

Tentu saja, sikap "hanya pengguna" seperti ini tidak ideal; karena dapat menyebabkan modul-modul menjadi seperti "black box". Namun menurut penulis sepanjang modul-modul tersebut diperoleh dari sumber yang bonafide; risiko bias pada penggunaan dapat dimitigasi. STATA memahami hal ini dan bahkan membuat suatu sistem peer review yang robust, dan sekarang telah memiliki reputasi yang sangat baik dikalangan akademisi dunia (lihat STATA journal; https://www.stata-journal.com/). Kita sebagai pengguna tetap berkewajiban memahami modul-modul yang digunakan untuk studi, walaupun mungkin pemahaman tidak secara detail teknis; paling tidak pengertian secara filosofis-prinsip harus dapat diperoleh. Untuk maksud inilah buku ini dibuat. Setiap modul statistika akan diuraikan secara prinsip, maksud, dan tujuan formulasinya. Aplikasi lebih lanjut pada data diharapkan dapat memberikan pemahaman lebih mendalam.

Statistik dan ekonometrika adalah disiplin ilmu yang sudah cukup matang. Sangat banyak teknik analisis-estimator yang dihasilkan. Buku-buku teks statistik dan ekonometrika baik di tingkat sarjana maupun pasca sarjana dapat memiliki ketebalan lebih dari 1.000 halaman; belum

jurnal-jurnalnya. Dengan demikian, penulis melihat nilai tambah signifikan lain yang dapat diberikan; yakni buku berkarakter selektif. Buku ini ditulis untuk para akademisi dan praktisi yang memiliki minat penelitian pada rumpun ilmu ekonomi-bisnis dengan pendekatan statistik-ekonometrika. Penulis telah menelusuri literatur yang ada dan secara selektif memilih modul-modul yang dipandang relevan serta cukup sering digunakan. Modul-modul tersebut meliputi topik klasik (seperti Ordinary Least Squares = OLS) hingga yang kontemporer (seperti Machine Learning).

Singkatnya, penulis berharap buku ini dapat menjadi semacam "buku saku"; di mana para peneliti dapat mencari tahu metodologi apa yang paling cocok untuk menjawab pertanyaan riset tertentu. Buku ini diharapkan dapat memberikan gambaran umum mengenai aspek teknis (langkah-langkah dan teori) dari metodologi tersebut. Dari penelusuran penulis di pasar; buku semacam ini belum banyak tersedia. Untuk memperoleh manfaat yang optimal, pembaca diharapkan sudah memiliki pemahaman yang baik terhadap statistik dan matematika dasar pada tingkat perguruan tinggi.

Dalam penulisan buku ini, penulis memperoleh banyak dukungan dari berbagai pihak yang tidak dapat disebutkan satu per satu. Secara khusus, penulis ingin menyampaikan ucapan terima kasih kepada

1. Kedua orang tua penulis, yang telah membimbing dan membesarkan para penulis.
2. Keluarga penulis, istri dan anak tercinta yang telah mendampingi penulis dalam meniti karier.
3. Pimpinan Binus University dan Universitas Sebelas Maret
4. Mahasiswa-mahasiwa penulis yang telah memberikan inspirasi dan masukan
5. Rekan kerja dan mitra penelitian yang tidak dapat penulis sebutkan satu per satu.

Semoga buku ini dapat menjadi suatu karya yang membantu para civitas akademika dan praktisi untuk menggunakan metode statistik-

ekonometrika baik untuk meningkatkan kualitas aktivitas ilmiah maupun bisnis.

Akhir kata penulis menyadari bahwa tidak ada karya manusia yang sempurna. Buku ini memiliki kekurangan dan kesalahan yang seluruhnya diakui menjadi tanggung jawab penulis. Penulis terbuka dan sangat mengharapkan berbagai masukan dari para pembaca untuk kemungkinan pengembangan di masa depan. Segala kritik, saran, dan masukan dapat disampaikan dengan menghubungi penulis via email doddy_ariefianto@yahoo.com dan irwan_t@staff.uns.ac.id

Jakarta dan Surakarta,
September 2020

Doddy dan Irwan

DAFTAR ISI

Tentang Pengarang vi
Kata Pengantar viii
Daftar Isi xii
Daftar Gambar xiv
Daftar Tabel xv

BAB 12 MODEL REGRESI DERET WAKTU MULTIVARIAT 1
Representasi dan Estimasi ADL 2
Vector Auto Regression 9
Granger Causality Test 13
Impulse Response Function dan Variance Decomposition 15
Regime Switching Regression 24

BAB 13 KARAKTER NON-STATIONARY DAN KOINTEGRASI 32
Proses Stationary 34
Konsep Eksogenitas 38
Proses NonStationer 41
Pengujian Unit Root 45
Uji Dickey-Fuller 51
Perkembangan Uji Unit Root 53
Kointegrasi dan Error Correction Model 59
Kointegrasi: Konsep dan Pengujian 59
Error Correction Model 63
Vector Error Correction Model (VECM) 68

BAB 14 REGRESI DATA PANEL **75**
Representasi dan Estimasi OLS 77
Model Efek Tetap (PEM) 78
Model Efek Random 81
Uji Hausman 83
Metode Instrumental Variabel 89
Model Data Panel Dinamis 93
Model Long Data Panel 101

BAB 15 MODEL VARIABEL DEPENDEN YANG TERBATAS 117
Model Regresi Binary Response 119
Variabel Dependen Multikategori 130
Variabel Dependen Multinominal 130
Ordered Response 136
Regresi Poisson 140
Censored Regression 145
Prosedur Heckman 152

BAB 16 STRUCTURAL EQUATION MODELLING (SEM) 159
Representasi dan Path Model 160
Estimasi dan Evaluasi 163
Penggunaan SEM Builder 166
Pengembangan Model SEM 172
Cross Lagged Panel Design 174

BAB 17 PEMROGRAMAN DAN SIMULASI 177
Beberapa Unsur Penting 179
Ilustrasi: Pembuatan Program Sederhana: ARDL 181
Eksperimen Monte Carlo 182
Ilustrasi 1: Spurious Regression 186
Illustrasi 2: Dickey Fuller 189

BAB 18 MACHINE LEARNING DAN EKONOMETRIKA BAYESIAN 193
Machine Learning: Suatu Pengantar 195
LASSO Estimator 200
Penggunaan LASSO untuk Pemodelan Regresi 213
Ekonometrika Bayesian: Suatu Pengantar 217
Bayesian Linear Regression 222

BAB 19 PELAPORAN HASIL ANALISIS 229
Analisis Pendahuluan 231
Robustness Check 233
Laporan Regresi: Baseline dan Elaborasi 236
Tabel Rangkuman Regresi 237
Intrepretasi dan Analisis 242
Daftar Pustaka 246
Indeks 261

DAFTAR GAMBAR

Gambar 12.1. Pengujian Stabilitas VAR 20
Gambar 13.1. Series Stasioner $y_t = 0{,}9\ y_t - 1 + ut$; $ut \sim NIID(0,1)$ 38
Gambar 13.2. Regresi Palsu (*Spurious Regression*) 44
Gambar 13.3. Prosedur Pengujian Unit Root Doldado et al (1990) 49
Gambar 13.4 Grafik Garis USDIDR_L 54
Gambar. 13.5. Grafik Garis D_USDIDR_L 56
Gambar. 13.6. Grafik Garis dari Residual EG Tahap 1. 65
Gambar. 13.7. Pengujian Stabilitas VECM 73
Gambar 15.1. Fungsi Logistik 121
Gambar 15.2. Ordered Response 138
Gambar 15.3 Histogram Sebaran Medali 145
Gambar 15.4. Histogram Sebaran Jam Kerja (Hours), Wanita yang Telah Menikah 148
Gambar 16.1. Suatu Ilustrasi SEM yang Lengkap 161
Gambar 16.2. Non-Recursive; Left Panel dan (b) Recursive Model; Right Panel 163
Gambar 16.3. Akses Menu SEM Builder 166
Gambar 16.4. Menu SEM Builder dan Beberapa Icon Terpilih 167
Gambar 16.5 SEM Builder Model I. 169
Gambar 16.6. Hasil Estimasi Model 1 170
Gambar 16.7. SEM Builder Model II: Cross Lagged Panel. 174
Gambar 17.1. Syntax Command pada File Dickey_Fuller_Brooks.do. 190
Gambar 18.1 Bias dan Varians Trade Off (a) dan Optimal Training (b) .. 201
Gambar 18.2. 5-Fold Cross Validation 206
Gambar 18.3. Jalur Pergerakan Koefisien (Trayektori) Variabel Terpilih .. 210
Gambar 18.4. MSPE sebagai Fungsi dari Log Natural λ 212
Gambar 18.5. Beberapa Indikator Diagnosis Estimasi Bayesian 228

DAFTAR TABEL

Tabel 12.1. Regresi OLS Dep_rate Inflasi 6
Tabel 12.2. Regresi ADL(1,1) dan ADL(3,3) 8
Tabel 12.3. VAR dari Stock Watson (2003) 18
Tabel 12.4. Granger Causality Test 19
Tabel.12.5. Forecast Error Variance Decomposition 22
Tabel 12.6. Pengujian Struktural Break; di mana Break diketahui: Maret 2008 27
Tabel 12.7. Pengujian Structural Break; Asumsi Terjadi Satu Break. 28
Tabel 12.8. Hasil Estimasi Model SETAR untuk ER_ASII. 29
Tabel 12.9. Hasil Markov Switching Regression untuk ER_ASII. 30
Tabel 12.10. Transition Probabilities, Markov Switching Regression untuk ER_ASII. 31
Tabel 12.11. Expected Duration, Markov Switching Regression untuk ER_ASII. 31
Tabel 13.1. Statistik Dickey Fuller (DF) 52
Tabel 13.2. Pengujian ADF pada Series USDIDR_L 55
Tabel. 13.3 Pengujian ADF pada Series D_USDIDR_L. 57
Tabel. 13.4. Pengujian Phillip-Perron (a) dan KPSS (b) untuk USDIDR_L. 58
Tabel. 13.5. Regresi Tahap Pertama Prosedur EG, Variabel Dependen: btc 65
Tabel 13.6. Pengujian ADF untuk Series btc_r. 66
Tabel 13.7. Regresi Tahap Kedua Prosedur EG, Variabel Dependen: btc 67
Tabel 13.8. Penentuan Lag yang Optimal dari VECM 70
Tabel 13.9. Pengujian Kointegrasi dari VECM 71
Tabel 13.10. Pengujian Kointegrasi dari VECM 72
Tabel 14.1. Estimasi Pooled OLS 85
Tabel 14.2. Estimasi Fixed Effect Model. 86
Tabel 14.3. Estimasi Random Effect Model. 87
Tabel 14.4. Breusch Pagan Test REM 88

Tabel 14.5. Hausman Test FEM vs REM 89
Tabel 14.6. Hasil Estimasi Instrumental Variabel Data Panel: G2SLS Estimator 92
Tabel 14.7. Hasil Estimasi Instrumental Variabel Data Panel: EC2SLS Estimator 92
Tabel 14.8. Hasil Estimasi Xtivreg Dibandingkan Xtivreg2 94
Tabel 14.9. Hasil Estimasi dengan Menggunakan Teknik D-GMM; Xtabond2 98
Tabel 14.10. Hasil Estimasi dengan Menggunakan Teknik S-GMM; xtabond2 100
Tabel 14.11 Pilihan teknik estimasi Model Long Panel (diadaptasi dari Eberhardt dan Teal (2011)) 104
Tabel 14.12. Pengujian Panel Unit Root dengan Routine Multipurt, Panel Atas (Maddala dan Wu), Panel Bawah (CIPS). 110
Tabel 14.13. Pengujian Uji Kointegrasi Panel dengan Metode Westerlund (2008). 111
Tabel 14.14. Pengujian Cross Section Dependence dengan Metode Pesaran (2004). 112
Tabel 14.15. Hasil Estimasi Mean Group. 113
Tabel 14.16. Hasil Estimasi Mean Group - Common Correlated Effect. ... 114
Tabel 14.17. Hasil Estimasi Pooled Mean Group. 115
Tabel 15.1. Estimasi OLS (LPM) Pilihan Moda Transportasi. 125
Tabel 15.2. Estimasi Logit (Panel Atas) dan Probit (Panel Bawah) Pilihan Moda Transportasi 126
Tabel 15.3. Pengujian Hosmer and Lemeshow untuk Model Logit 127
Tabel 15.4. Tabel Klasifikasi Estimasi Logistik Moda Transportasi 128
Tabel 15.5. Prediksi Marginal Probabilitas Auto = 1 pad dtime = 2 (Panel Atas) dan dtime = 3 (Panel Bawah). 129
Tabel 15.6. Regresi Multinomial Logit Pilihan Pendidikan. 133
Tabel 15.7. Perhitungan Probabilitas Marjinal pada pilihan Pendidikan: 1 (Atas), 2 (Tengah) dan 3 (Bawah) untuk Nirat = 8 134
Tabel 15.8. Perhitungan Rata-rata Probabilitas Marjinal (AME) untuk Variabel Independen Nirat. 136
Tabel 15.9. Estimasi dengan Ordered Logit; Variabel Tergantung Pendidikan. 139
Tabel 15.10. Perhitungan Probabilitas Marjinal pada pilihan Pendidikan: 1 (Atas), 2 (Tengah) dan 3 (Bawah) untuk Nirat = 8 dengan menggunakan Ordered Logit. 141

Tabel 15.11. Perhitungan Rata-rata Probabilitas Marjinal (AME) untuk Variabel Independen Nirat Model Regresi Ordered Logit. 142
Tabel 15.12 Regresi Possion dengan Variabel Tergantung Medaltot. 143
Tabel 15.13. Regresi Possion dengan Variabel Tergantung Medaltot. 144
Tabel 15.14. Statistik Deskriptif untuk Wanita Tidak Bekerja (Atas) Versus Bekerja (Bawah) 148
Tabel 15.15. Regresi Tobit dengan Variabel Depnden Hours, Left Limit. 149
Tabel 15.16. Regresi OLS dengan Variabel Depnden Hours, Seluruh Sampel (Atas) dan Hanya Jika Hours > 0 (Bawah). 150
Tabel. 15.17. Dampak Marginal Total Jam Kerja 151
Tabel. 15.18. Dampak Marginal Total Jam Kerja untuk Data yang Terobservasi (Hours 0)..... 152
Tabel. 15.19. Regresi Tahap Pertama dari Prosedur Heckman 156
Tabel. 15.20. Regresi Tahap Kedua dari Prosedur Heckman 156
Tabel. 15.21. Regresi OLS Hanya pada Sampel Wanita Bekerja (Hours > 0)..... 157
Tabel. 15.22. Regresi Heckman Two Step Estimator 158
Tabel 16.1 Hasil Estimasi SEM Model 1..... 170
Tabel 16.2. Koefisien Determinasi Model 1..... 171
Tabel 16.3 Beberapa Statistik Goodness of Fit Model 1 172
Tabel 16.4 Modified Indices; Alternatif Pengembangan Model I 173
Tabel 16.5 Hasil Estimasi Model II..... 175
Tabel 16.6 Koefisien Determinasi Model II 175
Tabel 16.7. Beberapa Statistik Goodness of Fit Model II 176
Tabel. 17.1. Output Program ardl_reg 183
Tabel 17.2. Hasil Simulasi Monte Carlo Spurious Regression..... 189
Tabel 17.3. Hasil Running Program Dickey_Fuller_Brooks.ado..... 191
Tabel 18.1. Taksonomi Machine Learning..... 198
Tabel 18.2. Variabel yang Digunakan; Harrison dan Rubinfeld (1978) .. 208
Tabel 18.3 Estimasi LASSO 209
Tabel 18.4. Hasil Estimasi Cross Validation LASSO..... 211
Tabel 18.5. Hasil Estimasi Rigorous LASSO 213
Tabel 18.6. Hasil Estimasi PDS Lasso untuk Seleksi Variabel Kontrol dan Variabel Instrument pada Studi AJR (2001) 218
Tabel 18.7. Perbandingan Ekonometrika Frequentist Versus Ekonometrika Bayesian..... 219

Tabel 18.8. Keunggulan dan Kelemahan Ekonometrika Bayesian..... 220
Tabel 18.9. Hasil Estimasi Regresi OLS Model harga Mobil 224
Tabel 18.10. Hasil Estimasi Bayesian Default Prior, Regresi Linear Hodel Harga Mobil 225
Tabel 18.11. Hasil Estimasi Bayesian Zellner's g Prior, Regresi Linear Model Harga Mobil 227
Tabel 19.1. Statistik Deskriptif Univariat pada Ariefianto, Widuri, Abdurachman, dan Trinugroho (2020)..... 232
Tabel 19.2. Tabel Korelasi pada Ariefianto, Widuri, Abdurachman, dan Trinugroho (2020)..... 233
Tabel 19.3. Robustness Check; AlternatiF Estimator 234
Tabel 19.4. Robustness Check; Sequential Inclusion 236
Tabel 19.5. Elaborasi Regresi; Klasifikasi Negara..... 240
Tabel 19.6. Tabel Rangkuman Output..... 241
Tabel 19.7. Output Window Eksekusi Program File real_rate_outreg2.do
Tabel 19.8. Output File (real_rate.xls) eksekusi Program File real_rate_outreg2.do..... 242

Bab 12

Model Regresi Deret Waktu Multivariat

Dalam pemodelan ekonometrika dengan menggunakan berbagai variabel deret waktu, salah satu aspek yang harus diperhatikan adalah keberadaan dinamika. Di antara berbagai variabel tersebut sangat mungkin terdapat hubungan yang bersifat dinamis. Nilai suatu variabel dependen tidak hanya dipengaruhi oleh nilai variabel independen lainnya pada periode yang sama (disebut hubungan *contemporaneous*), tetapi juga oleh nilai variabel lainnya (baik variabel dependen maupun variabel independen) pada titik waktu yang berbeda.

Apabila model yang diestimasi tidak mempertimbangkan kemungkinan fenomena ini, maka sangat mungkin model akan mengalami masalah misspesifikasi (yang berimplikasi biasnya parameter) atau autokorelasi dan heterokedastisitas (yang berimplikasi pada kesulitan inferensial). Ada dua teknik yang dapat digunakan untuk mengakomodasi keberadaan dinamika, yakni model *Autoregressive Distributed Lag* (ADL) dan *Vector Auto Regression* (VAR).

REPRESENTASI DAN ESTIMASI ADL

Suatu model ADL dengan 1 variabel independen serta orde p, r, dan q dapat diberikan sebagai formulasi sebagai berikut

$$y_t = \alpha + \sum_{i=1}^{p} \beta_i y_{t-i} + \sum_{i=0}^{r} \gamma_i x_{t-i} + \sum_{i=1}^{q} \delta_i \varepsilon_{t-i} + \varepsilon_t \qquad (12.1)$$

di mana p adalah orde lag dari variabel dependen (komponen *autoregressive*), r adalah orde lag dari variabel independen, dan q adalah orde lag dari residual (komponen *moving average*).

Model yang diberikan oleh Persamaan 12.1 itu bersifat dinamis. Di sini dapat dilihat bahwa nilai variabel dependen saat ini: y_t tidak hanya dipengaruhi oleh nilai variabel independen saat ini x_t, tetapi juga oleh nilai variabel dependen dan variabel independen di masa lalu serta diskrepansi yang ada di antara hubungan keduanya (residual).

Terdapat beberapa penyebab mengapa terjadi pola yang bersifat dinamis, di antaranya (Heij et al, 2004, hal 638-639):

a. Mekanisme penyesuaian parsial (*partial adjustment*). Dari teori ekonomi diketahui bahwa konsumsi dipengaruhi oleh pendapatan. Lebih lanjut, kenaikan pendapatan tidak akan disertai dengan peningkatan konsumsi secara menyeluruh, melainkan bertahap. Hal ini dapat terjadi karena karakter kehati-hatian konsumen serta ketidakpastian mengenai sustainabilitas kenaikan pendapatan itu.

b. Ekspektasi adaptif (*adaptive expectation*). Dalam mengambil keputusan, seorang pelaku bisnis sering bergantung pada informasi yang diperoleh pada waktu yang lalu. Sebagai contoh, seorang petani yang mendapati harga komoditas yang ditanamnya mengalami penurunan tajam akan sangat mungkin mengurangi alokasi penanaman saat ini. Dengan demikian, tentu saja pada saat panen produksi tanaman dimaksud akan mengalami penurunan, dan jika keputusan ini dilakukan juga oleh petani lain, maka akan terdapat dampak agregat berupa penurunan pasokan atau *supply* (sehingga harga pasar berpotensi naik lagi).

c. Proses Koreksi Kesalahan (*Error Correction Model*). Beberapa keputusan bisnis/ekonomi tertentu sangat mungkin berpedoman

pada suatu pencapaian angka yang ideal. Apabila realisasi suatu variabel tidak sama dengan target ideal tersebut (surplus atau defisit), maka akan dilakukan upaya pemulihan. Keputusan-keputusan ini dapat ditunjukkan pada perilaku pengendalian persediaan, jumlah saldo kas, dan variabel fiskal.

Estimasi dan evaluasi model ADL sebenarnya bersifat pengembangan langsung (*straight forward*) dari teknik OLS. Namun demikian, beberapa asumsi tambahan digunakan untuk menjamin bahwa parameter yang ditemukan adalah tidak bias dan efisien.

Beberapa asumsi tambahan yang diperlukan itu adalah

a. Variabel penjelas (sisi sebelah kanan regresi) adalah bersifat eksogen. Hal ini berlaku baik bagi variabel x_t (dan lag-nya) serta lag dari variabel dependen (variabel *predetermined*). Apabila variabel-variabel ini tidak bersifat eksogen (yang berarti bersifat endogen), maka akan lebih baik jika estimasi dilakukan dengan menggunakan teknik estimasi sistem (multiequation seperti vector auto regression atau persamaan simultan)[1].
b. Seluruh parameter autoregressive β_i, $i = 1,..., p$ memiliki nilai absolut di bawah 1. Secara teknis, persyaratan ini disebut stasioneritas. Apabila variabel-variabel yang digunakan bersifat tidak stasioner, maka pemodelan yang lebih tepat digunakan adalah estimasi model koreksi kesalahan (*error correction model*). Pembahasan mengenai hal ini akan diuraikan lebih lanjut di Bab 13.

[1] Lihat Bab 10 untuk pemahaman mengenai dampak dari variabel yang tidak bersifat eksogen terhadap estimasi dengan menggunakan OLS.

Pada model regresi ini dapat dilakukan prosedur evaluasi standar seperti pada model OLS. Kita dapat menguji apakah parameter dinamis adalah signifikan, dengan menggunakan Wald Test. Koefisien determinasi: R^2, memiliki interpretasi sebagai varians variabel dependen yang dapat dijelaskan oleh variabel independen. Demikian juga, teknik deteksi autokorelasi dan heterokedastisitas sebagaimana dijelaskan pada Bab 9 di Jilid 1, juga dapat diterapkan pada residual model.

Dengan menggunakan model ADL kita dapat membedakan koefisien/parameter respons yang bersifat jangka pendek dan jangka panjang. Misalnya, kita menggunakan versi yang lebih sederhana dari ADL Persamaan 12.1 dengan menghilangkan parameter koreksi kesalahan atau

$$y_t = \alpha + \sum_{i=1}^{p} \beta_i y_{t-i} + \sum_{i=0}^{r} \gamma_i x_{t-i} + \varepsilon_t \qquad (12.2)$$

Koefisien γ_i, $i = 1,...,r$ disebut sebagai parameter respons jangka pendek. Koefisien respons jangka panjang (sebut saja λ) diperoleh ketika variabel dependen dan variabel independen sudah tidak lagi mengalami pergerakan: sistem berada dalam kondisi seimbang (ekuilibrium). Dengan kata lain, $y_t = y_{t-1} = .. .= y_{t-p}$ dan $x_t = x_{t-1} = x_{t-r}$. Secara matematis

$$\lambda = \frac{\sum_{i=0}^{r} \gamma_i}{1 - \sum_{i=1}^{p} \beta_i} \qquad (12.3)$$

Contoh 12.1

Kita akan menggunakan data pada file Inflasi_Bunga.dta. File ini berisi data frekuensi bulanan tingkat inflasi umum (inf) dan suku bunga rata-rata deposito Rupiah tenor 1 bulan dari Bank Umum (dep_rate) selama periode Januari 2010 hingga Desember 2019. Dari teori ekonomi uang dan bank, kita mengetahui bahwa sebagai suatu bentuk investasi imbal hasil Deposito minimal sama dengan inflasi. Namun demikian, dari teori bank kita juga mengetahui bahwa karena proses bisnis internal seperti strategi bisnis dan manajemen risiko, maka penyesuaian suku bunga tersebut tidak dapat dilakukan secara seketika.

Source	SS	df	MS	Number of obs	=	120
				F(1, 118)	=	29.61
Model	17.7824389	1	17.7824389	Prob > F	=	0.0000
Residual	70.8707512	118	.600599586	R-squared	=	0.2006
				Adj R-squared	=	0.1938
Total	88.65319	119	.74498479	Root MSE	=	.77498

dep_rate	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
inf	.2303207	.0423282	5.44	0.000	.1464993	.314142
_cons	5.596504	.2156098	25.96	0.000	5.169538	6.02347

TABEL 12.1. Regresi OLS Dep_rate Inflasi

Salah satu cara untuk memulai pemodelan ADL adalah dengan melakukan terlebih dahulu regresi contemporaneous dan melihat apakah regresi tersebut mengalami masalah serial korelasi. Hasil regresi OLS dep_rate dengan inf ditunjukkan pada Tabel 12.1. Pengujian serial korelasi (estat Dwatson) akan memberikan nilai statistik DW sebesar 0,063, yang berarti regresi mengalami

serial korelasi. Dengan demikian, pemodelan ADL mungkin dapat digunakan sebagai perbaikan spesifikasi.

Isu berikutnya adalah penentuan parameter untuk spesifikasi model ADL(*p*, *r*, dan *q*). Kita tidak akan menggunakan dahulu *q*, karena topik ini adalah tentang koreksi kesalahan (*error correction*) yang akan dibahas pada bab tersendiri (*error correction model*). Dengan demikian, apa yang dapat digunakan sebagai alternatif adalah lag variabel dependen (*autoregressive*, *p*) dan lag variabel independen (*distributed lag, r*). Kita dapat menggunakan *correlogram* yang telah dipelajari sebelumnya atau pertimbangan praktis (berdasarkan pengalaman; *judgment*). Di sini penulis memilih menggunakan judgment. Sebagai seorang mantan praktisi perbankan; penulis meyakini bahwa passthrough inflasi ke suku bunga simpanan akan berlangsung cukup cepat, yaitu antara 1-3 bulan. Dengan demikian, kita memiliki sembilan pilihan yakni ADL(1,1), ADL(1,2), ADL(1,3), ADL(2,1), ADL(2,2), ADL(2,3), ADL(3,1), ADL(3,2), dan ADL(3,3).

Ilustrasi hasil estimasi dua model ADL: ADL(1,1) dan ADL(3,3) diberikan pada Tabel 12.2. Tabel ini diperoleh dari perintah STATA berikut **reg dep_rate L.dep_rate L.inf inf** untuk ADL(1,1) dan **reg dep_rate L(1/3).dep_rate L(1/3).inf inf** untuk ADL(3,3). Ada pun nilai statistik Durbin Watson untuk kedua model tersebut adalah (masing-masing): 1,167 (ADL(1,1)) dan 2,065 (ADL(3,3)). Sedangkan nilai AIC untuk kedua model tersebut adalah –126.544 (ADL(1,1)) dan -152.896 (ADL(3,3)). Dengan demikian, berdasarkan evaluasi serial korelasi dan kriteria informasi, terlihat model ADL(3,3) lebih superior. Tentu saja, kesimpulan yang lebih tepat adalah melakukan estimasi dan evaluasi terhadap seluruh 9 model, dan kita akan kembali lagi membahas hal ini pada bab penyajian laporan.

Source	SS	df	MS	Number of obs	=	119
				F(3, 115)	=	1469.96
Model	86.254208	3	28.7514027	Prob > F	=	0.0000
Residual	2.24931713	115	.019559279	R-squared	=	0.9746
				Adj R-squared	=	0.9739
Total	88.5035251	118	.750029874	Root MSE	=	.13985

dep_rate	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
dep_rate						
L1.	.9518301	.016696	57.01	0.000	.9187585	.9849017
inf						
L1.	.0106927	.0233031	0.46	0.647	-.0354663	.0568517
--.	.0328993	.0225417	1.46	0.147	-.0117515	.0775501
_cons	.1042491	.1021579	1.02	0.310	-.0981061	.3066044

Source	SS	df	MS	Number of obs	=	117
				F(7, 109)	=	836.00
Model	86.8286016	7	12.4040859	Prob > F	=	0.0000
Residual	1.6172728	109	.014837365	R-squared	=	0.9817
				Adj R-squared	=	0.9805
Total	88.4458744	116	.762464435	Root MSE	=	.12181

dep_rate	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
dep_rate						
L1.	1.276607	.0924507	13.81	0.000	1.093373	1.459842
L2.	-.0732725	.1512981	-0.48	0.629	-.3731405	.2265955
L3.	-.2464132	.088379	-2.79	0.006	-.4215774	-.071249
inf						
L1.	-.0645268	.0349851	-1.84	0.068	-.1338661	.0048126
L2.	.067564	.0350501	1.93	0.057	-.0019041	.1370321
L3.	-.0305963	.0221351	-1.38	0.170	-.0744673	.0132746
--.	.0515923	.0211523	2.44	0.016	.0096691	.0935155
_cons	.1708525	.0923023	1.85	0.067	-.0120877	.3537926

TABEL 12.2. Regresi ADL(1,1) dan ADL(3,3)

VECTOR AUTO REGRESSION

Vector Auto Regression (VAR) adalah pengembangan dari model ADL. VAR melonggarkan asumsi variabel yang bersifat eksogen pada ADL. Dalam kerangka VAR, dimungkinkan untuk melakukan estimasi terhadap serangkaian variabel yang diduga mengalami endogenitas.

Metodologi VAR dikemukakan pertama kali oleh Sims (1980). Ia menganggap pendekatan persamaan struktural ekonometrika (yang sangat dominan waktu itu) rentan terhadap kritik Lucas. Hubungan yang dibangun atas dasar asumsi adanya variabel endogen dan eksogen bersifat temporer. Respons agen ekonomi setelah mengobservasi suatu kebijakan yang diambil berdasarkan hubungan tersebut dapat saja berubah, sehingga variabel yang semula eksogen akan menjadi endogen. Agar suatu *reduced form* dapat diestimasi secara tidak bias dan konsisten serta dapat digunakan sebagai alat perumusan kebijakan, maka variabel eksogen tidak cukup bersifat *strongly exogenous* tetapi harus *super exogenous*. Asumsi ini terlalu ketat dan sulit dipenuhi.

Hubungan antara variabel ekonomi itu bersifat kompleks dan teori ekonomi baru dapat mengungkapkan sebagian dari pola hubungan tersebut. Dengan demikian, derajat tertentu dari endogenitas akan terjadi sehingga asumsi *super exogeneity* tidak akan dapat dipenuhi. Model VAR dibangun untuk mengatasi hal ini di mana hubungan antarvariabel ekonomi dapat tetap diestimasi tanpa perlu menitikberatkan masalah eksogenitas. Dalam pendekatan ini semua variabel dianggap sebagai endogen dan estimasi dapat dilakukan secara serentak atau sekuensial.

Suatu VAR sederhana yang terdiri dari 2 variabel dan 1 lag dapat diformulasikan sebagai berikut:

$$y_{1t} = \beta_{10} + \beta_{11} y_{1t-1} + \alpha_{11} y_{2t-1} + u_{1t}$$
$$y_{2t} = \beta_{20} + \beta_{21} y_{2t-1} + \alpha_{21} y_{1t-1} + u_{2t} \quad (12.6)$$

atau dalam bentuk matriks

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} \beta_{10} \\ \beta_{20} \end{pmatrix} + \begin{pmatrix} \beta_{11} & \alpha_{11} \\ \alpha_{21} & \beta_{21} \end{pmatrix} \begin{pmatrix} y_{1t-1} \\ y_{2t-1} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} \quad (12.7)$$

$$\underset{2\times 1}{\boldsymbol{y_t}} = \underset{2\times 1}{\boldsymbol{\beta_0}} + \underset{(2\times 2)}{\beta_1} \underset{(2\times 1)}{y_{t-1}} + \underset{2\times 1}{u_t}$$

Dapat dilihat di sini bahwa VAR terdiri dari variabel endogen dengan indeks saat ini di sisi sebelah kiri serta suatu komponen konstanta dan komponen *lagged term* di sisi sebelah kanan. Dengan asumsi bahwa tidak ada korelasi silang di antara *error term* ($E(u_{1t}u_{2t}) = 0$), VAR dapat diestimasi dengan menggunakan OLS. Estimasi dilakukan secara sekuensial, seperti dengan mengestimasi terlebih dahulu persamaan untuk y_{1t} baru y_{2t}.

Model ini dapat digeneralisir untuk mencakup k variabel dan dengan lag term sebanyak p. Dalam bentuk matriks kita akan memiliki formulasi berikut

$$\begin{pmatrix} y_{1t} \\ y_{2t} \\ \cdots \\ y_{kt} \end{pmatrix} = \begin{pmatrix} \alpha_{10} \\ \alpha_{20} \\ \cdots \\ \alpha_{k0} \end{pmatrix} + \begin{pmatrix} \alpha_{11}^1 & \alpha_{11}^2 & \cdots & \alpha_{11}^k \\ \alpha_{21}^1 & \alpha_{21}^2 & \cdots & \alpha_{21}^k \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_{k1}^k & \alpha_{k1}^k & \cdots & \alpha_{k1}^k \end{pmatrix} \begin{pmatrix} y_{1t-1} \\ y_{2t-1} \\ \cdots \\ y_{kt-1} \end{pmatrix} + \ldots \quad (12.8)$$

$$\underset{kx1}{\boldsymbol{y_t}} = \underset{kx1}{\boldsymbol{\alpha_0}} + \underset{(kxk)(kx1)}{\boldsymbol{\alpha_1 y_{t-1}}} +$$

$$+\begin{pmatrix} \alpha_{1p}^{1} & \alpha_{1p}^{2} & \dots & \alpha_{1p}^{k} \\ \alpha_{2p}^{1} & \alpha_{2p}^{2} & \dots & \alpha_{2p}^{k} \\ \dots & \dots & \dots & \dots \\ \alpha_{kp}^{k} & \alpha_{kp}^{k} & \dots & \alpha_{kp}^{k} \end{pmatrix} \begin{pmatrix} y_{1t-p} \\ y_{2t-p} \\ \dots \\ y_{kt-p} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \\ \dots \\ u_{kt} \end{pmatrix};$$

$$+ \quad \boldsymbol{\alpha_p y_{t-p}} \quad + \quad \mathbf{u_t}$$

$$(kxk)(kx1) \qquad kx1$$

Indeks koefisien variabel di sisi sebelah kanan, α^{i}_{uv} terdiri atas indeks persamaan variabel dependen (dependent) $i = 1,\dots,k$, indeks variabel penjelas $u = 1,\dots, k$ dan indeks lag $v = 1,\dots,p$.

Perhatikan kita di sini memformulasikan VAR sebagai suatu bentuk sistem *reduced form* terhadap variabel lain dan variabel sendiri dengan lag (*lagged own variable*). Tidak ada hubungan kontemporer dalam artian variabel di sisi sebelah kiri hanya dipengaruhi oleh seluruh variabel dalam sistem menurut kondisi *lagged* (jadi merupakan *predetermined variables*). Kita selalu dapat mengkonversi bentuk hubungan kontemporer ke dalam bentuk *predetermined* ini (lihat Brooks, 2014, hal 336-337). Dengan bentuk seperti ini, teknik OLS selalu dapat digunakan.

Teori ekonomi biasanya tidak akan memberikan informasi yang sangat mendetail dalam operasionalisasi empiris. Pemilihan lag VAR adalah salah satunya. Di sisi lain, penggunaan lag yang tepat adalah sangat penting. Lag yang terlalu sedikit akan berpotensi menimbulkan masalah bias spesifikasi, sedangkan jika terlalu banyak akan menghabiskan derajat kebebasan atau *degree of freedom*, sehingga estimasi menjadi tidak efisien.

Terdapat dua cara untuk menentukan orde lag, di mana yang pertama adalah menggunakan uji restriksi koefisien yang merupakan

generalisasi dari uji restriksi pada persamaan regresi tunggal (Wald test). Sebagai ilustrasi, misalkan kita melakukan estimasi VAR bivariat dengan 8 lag. Kita menduga bahwa lag 5 s/d 8 adalah tidak signifikan. Jadi, dengan menotasikan VAR lag 1/sd 4 sebagai *restricted* VAR dan VAR dengan lag 5 s/d 8 sebagai *unrestricted* VAR, statistik uji berikut dapat dihitung

$$LR = T[\log|\hat{\Sigma}_r| - \log|\hat{\Sigma}_u|] \quad (12.9)$$

di mana $|\hat{\Sigma}_r|$ adalah determinan matriks varians kovarians dari *restricted* VAR dan $|\hat{\Sigma}_u|$ adalah untuk *unrestricted* VAR. Statistik uji ini akan memiliki distribusi χ^2 dengan derajat kebebasan jumlah restriksi (pada kasus ini adalah 16 = 2 variabel atau persamaan regresi x 2 variabel lag x 4 restriksi orde lag (= 5 – 8)).

Cara pemilihan lag lain yang juga sering digunakan adalah kriteria informasi. Prosedur pemilihan lag dengan kriteria informasi dapat dilakukan sebagai berikut:

1. Estimasi VAR dengan lag maksimum. Lag maksimum tergantung pada jumlah observasi (T) dan dapat dihitung dengan rumus yang diberikan oleh Said dan Dickey (1984), yakni $T^{1/3}$ (lag maksimum adalah akar tiga dari T).
2. Selanjutnya lag yang optimal dapat dilihat dari nilai statistik kriteria informasi yang dihitung bagi setiap lag. Lag yang optimal adalah lag dengan nilai statistik kriteria informasi yang terkecil.
3. Terdapat beberapa statistik kriteria informasi *multivariat* yang di antaranya *Akaike Information Criterion* (AIC), *Schwartz Information Criterion*, dan *Hannan Quin*. Penggunaan kriteria berganda dapat dilakukan untuk pencarian lag yang lebih optimal.

GRANGER CAUSALITY TEST

Terdapat suatu aplikasi terkait dengan VAR, yakni *Granger Causality Test*. Sebagai suatu uji sebab akibat, kita perlu membedakannya dengan arti sebab akibat secara harfiah. Sebab akibat menurut Granger tidak memiliki arti fundamental, yaitu dalam artian kita dapat menelusuri alur logika mengapa suatu kejadian (X) akan menyebabkan kejadian lain (Y).

Granger Causality adalah murni suatu konsep statistik. Dalam konsep ini X dikatakan menyebabkan Y jika realisasi X terjadi lebih dahulu daripada Y dan realisasi Y tidak terjadi mendahului realisasi X. Dengan demikian, secara empiris uji Granger Causality dapat dilakukan dengan menggunakan model VAR sebagaimana diberikan oleh Persamaan 12.8.

Sebagai ilustrasi dapat diperhatikan model VAR bivariat (dengan lag p) sebagai berikut:

$$\begin{aligned} y_{1t} &= \beta_{10} + \beta_{11} y_{1t-1} + \ldots + \beta_{1p} y_{1t-p} + \alpha_{11} y_{2t-1} + \ldots + \alpha_{1p} y_{2t-p} + u_{1t} \\ y_{2t} &= \beta_{20} + \beta_{21} y_{2t-1} + \ldots + \beta_{2p} y_{2t-p} + \alpha_{21} y_{1t-1} + \ldots + \alpha_{2p} y_{2t-p} + u_{2t} \end{aligned} \quad (12.10)$$

Suatu struktur hipotesis Granger Causality dapat disusun sebagai berikut

1. ***y_1 granger cause y_2***

   Jika struktur hipotesis berikut mengalami penolakan hipotesis null, yaitu dalam artian koefisien α_{2i}; i =1,..,p adalah signifikan.

$$\begin{aligned} H_0 &: \alpha_{21} = \ldots = \alpha_{2p} = 0 \\ H_1 &: \alpha_{21} \neq \ldots = \alpha_{2p} \neq 0 \end{aligned} \quad (12.11)$$

dan struktur hipotesis berikut mengalami hipotesis null tidak dapat ditolak dalam artian koefisien α_{1i}; i =1,..,p adalah tidak signifikan.

$$H_0 : \alpha_{11} = ... = \alpha_{1p} = 0$$
$$H_1 : \alpha_{11} \neq 0; ... = \alpha_{1p} \neq 0 \quad (12.12)$$

2. *y_2 granger cause y_1*

Jika struktur hipotesis berikut mengalami penolakan hipotesis null, yaitu dalam artian koefisien α_{1i}; i = 1, .., p adalah signifikan.

$$H_0 : \alpha_{11} = ... = \alpha_{1p} = 0$$
$$H_1 : \alpha_{11} \neq 0; ... = \alpha_{1p} \neq 0 \quad (12.13)$$

dan struktur hipotesis berikut mengalami hipotesis null tidak dapat ditolak dalam artian koefisien α_{2i}; = 1,..,p adalah tidak signifikan.

$$H_0 : \alpha_{21} = ... = \alpha_{2p} = 0$$
$$H_1 : \alpha_{21} \neq 0; ... = \alpha_{2p} \neq 0 \quad (12.14)$$

Pengujian dilakukan dengan menggunakan uji restriksi koefisien (Wald test). Karena pada prinsipnya *VAR* adalah sekelompok persamaan regresi yang saling independen, maka Wald Test dapat diterapkan pada masing-masing persamaan regresi. Dalam praktek, dapat saja terjadi deviasi dari dua pilihan tersebut. Di sini bisa saja y_1 *granger cause* y_2 dan y_2 juga *granger cause* y_1. Dengan kata lain, kedua variabel tersebut bersifat endogen satu dengan yang lain (saling simultan) atau y_1 dan y_2 tidak saling *granger cause* yang di sini berarti kedua variabel tersebut tidak memiliki hubungan.

IMPULSE RESPONSE FUNCTION DAN VARIANCE DECOMPOSITION

Terdapat dua aplikasi popular lain dari model VAR, yakni *Impulse-Response Function* (IRF) dan *Variance Decomposition* (VD). IRF melakukan penelusuran atas dampak suatu goncangan (*shock*) terhadap suatu variabel pada sistem (seluruh variabel) sepanjang durasi tertentu. Sebagai ilustrasi, perhatikan model VAR bivariat dengan lag 1 sebagai berikut:

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} 0{,}5 & 0{,}3 \\ 0{,}0 & 0{,}2 \end{bmatrix} \begin{bmatrix} y_{1t-1} \\ y_{2t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} \tag{12.15}$$

Matriks koefisien VAR dapat diperoleh melalui estimasi. Sekarang seandainya pada saat $t = 0$ terjadi goncangan pada y_1 sebesar 1, maka kita dapat menelusuri dampak goncangan dimaksud terhadap y_1 dan y_2 pada $t = 0,1,\ldots$ dst sebagai

$$\begin{aligned} y_0 &= \begin{bmatrix} u_{10} \\ u_{20} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ y_1 &= \begin{bmatrix} y_{11} \\ y_{21} \end{bmatrix} = \begin{bmatrix} 0{,}5 & 0{,}3 \\ 0{,}0 & 0{,}2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0{,}5 \\ 0{,}0 \end{bmatrix} \\ y_2 &= \begin{bmatrix} y_{12} \\ y_{22} \end{bmatrix} = \begin{bmatrix} 0{,}5 & 0{,}3 \\ 0{,}0 & 0{,}2 \end{bmatrix} \begin{bmatrix} 0{,}5 \\ 0{,}0 \end{bmatrix} = \begin{bmatrix} 0{,}25 \\ 0{,}0 \end{bmatrix} \\ &dst \end{aligned} \tag{12.16}$$

Dapat kita lihat di sini bahwa dampak goncangan adalah semakin kecil dari waktu ke waktu (sistem adalah stabil). Tidak ada pengaruh

dari goncangan y_1 terhadap y_2, karena memang y_2 tidak dipengaruhi oleh y_1 (koefisien y_1 pada regresi y_2 adalah nol, unsur matriks 21). Sedangkan dampak goncangan terhadap variabel y_2 adalah terjadi pada kedua variabel. Hal ini dapat dilihat melalui ilustrasi berikut:

$$y_0 = \begin{bmatrix} u_{10} \\ u_{20} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$y_1 = \begin{bmatrix} y_{11} \\ y_{21} \end{bmatrix} = \begin{bmatrix} 0{,}5 & 0{,}3 \\ 0{,}0 & 0{,}2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0{,}3 \\ 0{,}2 \end{bmatrix} \quad (12.17)$$

$$y_2 = \begin{bmatrix} y_{12} \\ y_{22} \end{bmatrix} = \begin{bmatrix} 0{,}5 & 0{,}3 \\ 0{,}0 & 0{,}2 \end{bmatrix} \begin{bmatrix} 0{,}3 \\ 0{,}2 \end{bmatrix} = \begin{bmatrix} 0{,}21 \\ 0{,}04 \end{bmatrix}$$

dst

Seperti halnya goncangan pada variabel y_1, dampak goncangan terhadap variabel y_2 adalah semakin kecil (*dampen out*).

VD melakukan dekomposisi atas perubahan nilai suatu variabel yang disebabkan oleh (a) goncangan variabel itu sendiri dan (b) goncangan dari variabel lain. Varians residual prediksi s (s = 1,2,..) langkah ke depan dipecah berdasarkan bagian yang bersumber dari variabel itu sendiri dengan yang bersumber dari variabel-variabel lain. Secara umum, kita tentunya mengharapkan proporsi varians yang terbesar adalah yang bersumber dari variabel itu sendiri.

Urutan variabel sangatlah penting dalam perhitungan IRF dan VD (lihat Brooks, 2014, hal. 343 untuk penjelasan). Teori ekonomi mungkin dapat memberikan masukan, terutama mana yang

menjadi penyebab dan mana yang menjadi akibat. Jika teori tidak menyediakan informasi, maka suatu teknik *trial and error* dapat dilakukan dengan mengubah-ubah urutan dan memilih di antaranya yang dianggap paling baik/masuk akal atau stabil[2].

Contoh 12.2

Kita akan menggunakan contoh dari Stock dan Watson (2003) untuk menunjukkan teknik estimasi VAR serta penggunaannya. Data yang digunakan adalah file SW_VAR.dta. File ini berisi data tingkat Inflasi (*inflation*), Tingkat Pengangguran (*unrate*) dan suku bunga kebijakan bank sentral (Fed_Funds) dari Amerika Serikat. Data memiliki frekuensi kuartalan pada periode 1960Q1 sampai dengan 2000Q4.

Estimasi VAR dapat dilakukan melalui ribbon menu statistiks/multivariate timeseries/Vector Autoregression. Setelah sampai ke dalam menu VAR kita dapat mengisi parameter-parameter yang diperlukan. Dalam contoh ini, kita akan mengikuti Stock and Watson (2003) yang menggunakan lag 4 dan urutan estimasi inflation un_rate fed_funds[3]. Dalam bentuk syntax, perintah VAR diberikan sebagai berikut **var inflation un_rate fed_funds, lags(1/4)**. Hasil dari estimasi VAR disajikan pada Tabel 12.3.

[2] Hal ini tidak terlalu menjadi perhatian jika kita dapat menjamin asumsi bahwa korelasi residual antar persamaan regresi adalah nol (lihat Lutkepohl, 1991, Bab 2).

[3] Alternatifnya pembaca dapat menggunakan rumus Said Dickey (1984) untuk memperoleh maksimum lag sebesar 5 (=floor(164)^1/3)); kemudian baik dengan menggunakan Wald Test maupun Kriteria Informasi maka diperoleh lag optimum = 3.

```
Vector autoregression

Sample:  1961q1 - 2000q4                        Number of obs     =        160
Log likelihood =  -381.3692                     AIC               =   5.254615
FPE            =   .0384676                     HQIC              =   5.558991
Det(Sigma_ml)  =   .0235998                     SBIC              =   6.004189

Equation           Parms      RMSE     R-sq      chi2     P>chi2
------------------------------------------------------------------
inflation             13    .985028   0.8547   940.9632   0.0000
un_rate               13    .226332   0.9796   7701.066   0.0000
fed_funds             13    .884461   0.9283   2071.513   0.0000
------------------------------------------------------------------

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
inflation    |
   inflation |
         L1. |   .6330409   .0781085     8.10   0.000      .479951    .7861307
         L2. |   .0534795   .0909963     0.59   0.557    -.1248699    .2318289
         L3. |   .1162038   .0923887     1.26   0.208    -.0648747    .2972823
         L4. |   .1925958   .0827329     2.33   0.020     .0304423    .3547493
             |
     un_rate |
         L1. |  -1.145122    .371053    -3.09   0.002    -1.872372   -.4178713
         L2. |   1.970371   .6375363     3.09   0.002     .7208225    3.219919
         L3. |  -1.573514   .6432297    -2.45   0.014    -2.834221   -.3128071
         L4. |   .5879851    .354803     1.66   0.097    -.1074116    1.283386
             |
   fed_funds |
         L1. |   .1577543   .0972981     1.62   0.105    -.0329465     .348455
         L2. |  -.1339647   .1297199    -1.03   0.302    -.3882111    .1202818
         L3. |   .0204681   .1275838     0.16   0.873    -.2295916    .2705278
         L4. |  -.0404852   .0977421    -0.41   0.679    -.2320562    .1510858
             |
       _cons |   .9470082   .3564853     2.66   0.008     .2483098    1.645707
-------------+----------------------------------------------------------------
un_rate      |
   inflation |
         L1. |   .0383619   .0179472     2.14   0.033     .0031861    .0735377
         L2. |  -.0269326   .0209084    -1.29   0.198    -.0679123    .0140472
         L3. |   .0161574   .0212284     0.76   0.447    -.0254494    .0577643
         L4. |  -.0216453   .0190097    -1.14   0.255    -.0589036    .0156131
             |
     un_rate |
         L1. |   1.493191   .0852577    17.51   0.000     1.326089    1.660294
         L2. |  -.5357773   .1464882    -3.66   0.000    -.8228889   -.2486656
         L3. |   .0041948   .1477964     0.03   0.977    -.2854809    .2938705
         L4. |  -.0132062   .0815233    -0.16   0.871    -.1729902    .1465778
             |
   fed_funds |
         L1. |   .0050053   .0223564     0.22   0.823    -.0388125    .0488231
         L2. |   .0439149   .0298061     1.47   0.141    -.0145039    .1023337
         L3. |  -.0250026   .0293152    -0.85   0.393    -.0824828    .0324308
         L4. |   .0088039   .0224584     0.39   0.695    -.0352138    .0528216
             |
       _cons |   .0588181   .0819105     0.72   0.473    -.1017235    .2193597
-------------+----------------------------------------------------------------
fed_funds    |
   inflation |
         L1. |   .0681366   .0701339     0.97   0.331    -.0693234    .2055966
         L2. |   .2264686   .0817059     2.77   0.006     .066328    .3866093
         L3. |  -.1010505   .0829562    -1.22   0.223    -.2636416    .0615406
         L4. |  -.0252986   .0742862    -0.34   0.733    -.1708969    .1202996
             |
     un_rate |
         L1. |  -1.642548   .3331699    -4.93   0.000    -2.295549   -.9895465
         L2. |   1.824396   .5724463     3.19   0.001     .7024221     2.94637
         L3. |  -.8537473   .5775584    -1.48   0.139    -1.985741    .2782465
         L4. |   .5511849    .318579     1.73   0.084    -.0732185    1.175588
             |
   fed_funds |
         L1. |    .946091   .0873643    10.83   0.000      .77486    1.117322
         L2. |  -.3669472    .116476    -3.15   0.002    -.595236   -.1386583
         L3. |    .339235    .114558     2.96   0.003     .1147054    .5637646
         L4. |   .0116979    .087763     0.13   0.894    -.1603144    .1837102
             |
       _cons |    .540692   .3200896     1.69   0.091    -.086672    1.168056
------------------------------------------------------------------------------
```


TABEL 12.3. VAR dari Stock Watson (2003)

Setelah melakukan estimasi VAR, kita dapat langsung melanjutkan ke pengujian Granger Causality. Hal ini dilakukan dengan perintah **vargranger**, yang dilakukan langsung setelah estimasi VAR. Hasil pengujian Granger Causality disajikan pada Tabel 12.4. Dapat dilihat di sini arah kausalitas satu arah (berdasarkan metodologi statistik) terjadi pada pasangan: (a) un_rate ? inflation dan (b) inflation ? fed_funds. Sedangkan kausalitas dua arah (simultanitas) terjadi di antara pasangan: un_rate ? fed_funds.

Granger causality Wald tests

Equation	Excluded	chi2	df	Prob > chi2
inflation	un_rate	15.476	4	0.004
inflation	fed_funds	2.6764	4	0.613
inflation	ALL	31.963	8	0.000
un_rate	inflation	5.4205	4	0.247
un_rate	fed_funds	17.56	4	0.002
un_rate	ALL	43.918	8	0.000
fed_funds	inflation	25.18	4	0.000
fed_funds	un_rate	32.44	4	0.000
fed_funds	ALL	45.405	8	0.000

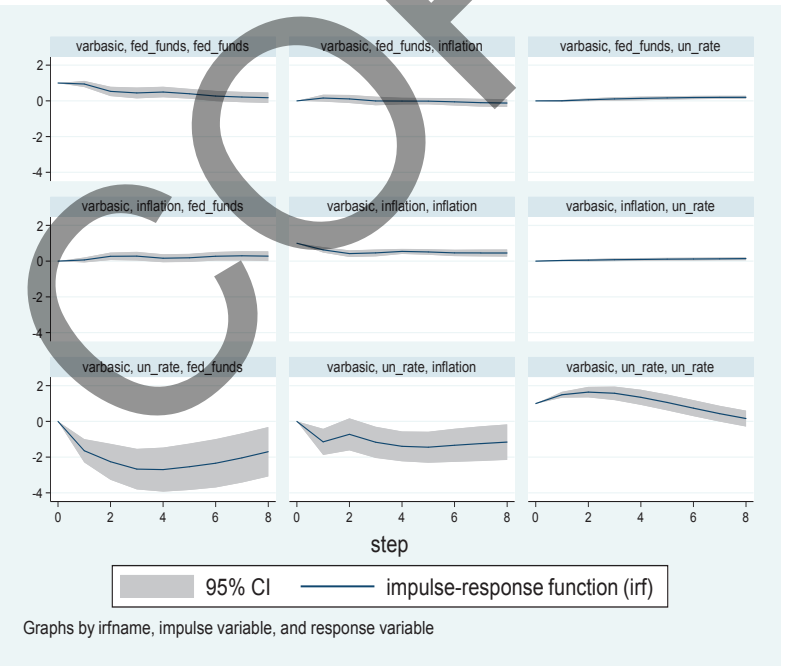
TABEL 12.4. Granger Causality Test

Kita dapat menguji stabilitas VAR dengan melihat apakah terdapat nilai eigen value dari sistem persamaan difference VAR yang lebih dari satu (*unit circle*). Hal ini dilakukan dengan perintah **varstable**. Impulse Response Function juga dapat digunakan untuk melihat stabilitas VAR; setelah terjadi shock atau goncangan, maka nilai variabel endogen VAR akan bergeser dari ekuilibriumnya. Pergeseran dari titik ekuilibrium ini bisa negatif atau positif, tetapi pada akhirnya akan die out (kembali ke titik deviasi = 0). Perintah Impulse Response Function adalah **irf graph irf**.

Eigenvalue stability condition

Eigenvalue	Modulus
.967199 + .05437487*i*	.968726
.967199 - .05437487*i*	.968726
.775581 + .1982162*i*	.80051
.775581 - .1982162*i*	.80051
-.0238767 + .6467808*i*	.647221
-.0238767 - .6467808*i*	.647221
.08549084 + .6012895*i*	.607337
.08549084 - .6012895*i*	.607337
-.4071293 + .1741432*i*	.442809
-.4071293 - .1741432*i*	.442809
.1388969 + .07348654*i*	.157139
.1388969 - .07348654*i*	.157139

All the eigenvalues lie inside the unit circle.
VAR satisfies stability condition.



GAMBAR 12.1. Pengujian Stabilitas VAR

Dari Gambar 12.1, pengujian stabilitas dengan menggunakan akar karakteristik (*eigen value*) menunjukkan seluruh nilainya berada dalam unit circle. Dengan kata lain, VAR yang diestimasi memenuhi kondisi stabilitas. Namun demikian, melalui Impulse Response Function kita dapat melihat bahwa reponse fed_funds (terhadap un_rate) dan un_rate (terhadap inflation) adalah tidak stabil. Nilai deviasi tidak tampak kembali ketitik nol bahkan setelah 8 kuartal.

Estimasi Forecast Error Variance Decomposition (FEVD) dapat dilakukan (dalam bentuk tabel) dengan perintah **irf table fevd**. Output dari perintah ini sangat panjang. Kita hanya mengambil suatu porsi untuk ilustrasi, yakni respons fed_funds terhadap impulse dari inflation dan un_rate (lihat Tabel 12.5). Setelah 12 kuartal, varians pada inflation dan un_rate dapat menjelaskan 64,46% (= 15,46% + 59,00%) varians dari fed_funds (lihat panel 3 dan 6).

Ilustrasi yang diberikan sebelumnya menunjukkan kemampuan VAR untuk digunakan dalam analisis kebijakan. Namun demikian, sebagai penutup bab ini penulis perlu mengingatkan kembali untuk jangan berlebihan (*overused*). VAR memiliki kelebihan dan kekurangan yang perlu diingat ketika seseorang menggunakan instrumen ini (Brooks, 2014, hal. 332-333).

Kelebihan VAR

1. VAR tidak memerlukan spesifikasi model, yaitu dalam artian mengidentifikasi variabel endogen-eksogen dan membuat persamaan-persamaan yang menghubungkannya. Semua variabel di dalam VAR bersifat endogen.
2. VAR bersifat sangat fleksibel, di mana pembahasan yang dilakukan hanya meliputi struktur autoregressive. Di sini pengembangan dapat dilakukan dengan memasukkan variabel yang dianggap murni eksogen (SVAR) dan/atau komponen

step	(3) fevd	(3) Lower	(3) Upper	(4) fevd	(4) Lower	(4) Upper
0	0	0	0	0	0	0
1	.015184	-.022423	.05279	0	0	0
2	.025399	-.028399	.079198	.067837	.005066	.130608
3	.063847	-.027652	.155346	.082841	-.000541	.166223
4	.085057	-.028938	.199052	.104413	.003	.205827
5	.082661	-.033737	.199058	.126324	.00723	.245418
6	.084705	-.037382	.206791	.145778	.006836	.284721
7	.095597	-.039731	.230924	.156886	.001134	.312638
8	.107838	-.041859	.257535	.161346	-.006922	.329614
9	.11807	-.044762	.280902	.162344	-.015729	.340418
10	.128988	-.047425	.305401	.161083	-.024561	.346727
11	.141647	-.049182	.332476	.158167	-.032614	.348948
12	.154642	-.050516	.3598	.154302	-.039393	.347998

step	(5) fevd	(5) Lower	(5) Upper	(6) fevd	(6) Lower	(6) Upper
0	0	0	0	0	0	0
1	.996467	.978114	1.01482	.19939	.089394	.309386
2	.996774	.987009	1.00654	.369377	.232897	.505856
3	.984671	.950073	1.01927	.448381	.293531	.60323
4	.957935	.887566	1.02831	.509855	.342537	.677172
5	.921831	.816311	1.02735	.556655	.380114	.733196
6	.875408	.739147	1.01167	.584595	.399099	.770092
7	.820531	.659521	.981541	.598145	.403767	.792524
8	.762607	.585869	.939346	.603414	.400686	.806142
9	.708109	.525663	.890555	.604706	.394137	.815275
10	.661758	.480708	.842808	.602357	.384253	.820461
11	.625213	.447782	.802644	.596933	.37162	.822247
12	.597425	.421989	.772861	.590051	.357943	.82216

95% lower and upper bounds reported
(1) irfname = varbasic, impulse = inflation, and response = inflation
(2) irfname = varbasic, impulse = inflation, and response = un_rate
(3) irfname = varbasic, impulse = inflation, and response = fed_funds
(4) irfname = varbasic, impulse = un_rate, and response = inflation
(5) irfname = varbasic, impulse = un_rate, and response = un_rate
(6) irfname = varbasic, impulse = un_rate, and response = fed_funds
(7) irfname = varbasic, impulse = fed_funds, and response = inflation
(8) irfname = varbasic, impulse = fed_funds, and response = un_rate
(9) irfname = varbasic, impulse = fed_funds, and response = fed_funds

TABEL 12.5. Forecast Error Variance Decomposition

moving average (VARMA). Dengan kata lain, VAR adalah suatu teknik ekonometrika struktural yang sangat kaya.

3. Kemampuan prediksi dari VAR adalah cukup baik. Beberapa kajian empiris (misalnya Sim, 1980 dan McNees, 1986) menunjukkan bahwa VAR memiliki kemampuan prediksi *out of sample* yang lebih tinggi daripada model makro struktural simultan.

Kelemahan VAR:

1. VAR bersifat ateoretis (tidak memiliki landasan teori). Hal ini disebabkan karena semua variabel dalam VAR bersifat endogen dan aspek struktur sebab akibat diabaikan.
2. Koefisien dalam VAR sulit untuk diinterpretasikan. Seperti yang dijelaskan sebelumnya, kegunaan VAR adalah untuk memprediksi dan menguji stabilitas hubungan sebab akibat (*impulse-response*). Jarang sekali perhatian diberikan pada masing-masing koefisien dalam VAR.
3. Estimasi dapat menjadi tidak efisien terutama jika jumlah sampel yang digunakan sedikit sedangkan variabel dan orde lag yang digunakan jumlahnya banyak (masalah *degree of freedom*). Jika terdapat g variabel endogen (berarti g persamaan regresi) serta orde lag sebanyak k, maka akan terdapat $g + kg^2$ parameter yang harus diestimasi. Sebagai ilustrasi, untuk VAR 3 variabel dengan orde lag 3, maka akan ada 30 parameter yang harus diestimasi.

REGIME SWITCHING REGRESSION

Pada Bab 8 di Jilid 1 telah diperkenalkan konsep stabilitas parameter. Secara khusus, pada regresi yang menggunakan data deret waktu, parameter hubungan yang diperoleh dapat berubah dari suatu subsampel (*window*) ke subsampel lainnya. Hal ini terjadi karena *structural break*: yaitu adanya perubahan yang substantial dari lingkungan asal data tersebut (*data generating process*–populasi). Sebagai ilustrasi, periode setelah krisis global tahun 2008 ditandai dengan kebijakan moneter, baik global maupun domestik, yang sangat longgar yang ditujukan untuk menstimulasi kinerja perekonomian. Hal ini sangat mungkin berpengaruh tidak hanya terhadap hubungan-hubungan variabel ekonomi moneter tetapi juga variabel makroekonomi lainnya (seperti penyerapan tenaga kerja dan ketimpangan).

Suatu cara memodelkan adanya structural break yang paling sederhana adalah dengan menggunakan variabel *dummy* (indikator). Misalnya, kita memiliki model regresi bivariat deret waktu sebagai berikut

$$y_t = \alpha_0 + \alpha_1 D_\tau + \beta_0 x_t + \beta_1 D_\tau * x_t + \varepsilon_t$$

$$D_\tau \begin{cases} 1 \text{ jika } t \geq T \\ 0 \text{ jika } t < T \end{cases} \qquad (12.18)$$

di mana T adalah periode awal terjadinya break. Persamaan 12.18 menyatakan adanya 2 persamaan regresi yang berbeda (disebut juga sebagai *regime*) yakni

$$y_t = \alpha_0 + \beta_0 x_t + \varepsilon_t \tag{12.19}$$

untuk periode sebelum break ($t < T$), dan

$$y_t = (\alpha_0 + \alpha_1) + (\beta_0 + \beta_1) x_t + \varepsilon_t \tag{12.20}$$

untuk periode setelah break ($t \geq T$). Kita dapat menguji apakah koefisien break α_1 dan β_1 adalah signifikan dengan menggunakan Wald test.

Dalam prakteknya, sangat mungkin kita tidak mengetahui di muka (*ex-ante*) *dating break* yang terjadi. Jika hal ini terjadi, kita dapat menggunakan *rolling regression* untuk mengambil sejumlah sampel t_1 di sekitar *dating T* yang diduga dari total sampel. Regresi dilakukan secara sekuensial dengan mendrop t_d awal sampel dan menambahkannya ke akhir sampel. Koefisien-koefisien dari rolling regression ini kemudian dibandingkan; dan ketika terjadi perubahan yang signifikan maka posisi tersebut dikatakan sebagai adanya *structural break*.

Pemodelan *break* lain yang mungkin lebih realistis adalah mengasumsikan proses yang berasal dari dalam sistem (endogen). Kebanyakan perubahan fundamental rezim terjadi karena perubahan dari satu atau lebih variabel-variabel yang ada dalam sistem. Tong (1978) mengusulkan suatu pemodelan endogen yang dapat diilustrasikan dengan model *threshold regression* berikut ini:

$$\begin{aligned} y_t &= \alpha_0 + x_t \beta + z_t \delta_1 + \varepsilon_t \text{ jika } -\infty < w_t < y \\ y_t &= \alpha_0 + x_t \beta + z_t \delta_2 + \varepsilon_t \text{ jika } \gamma < w_t < +\infty \end{aligned} \tag{12.21}$$

di mana vektor x_t adalah vektor variabel yang tidak terpengaruh oleh break (*regime invariant*); sedangkan z_t adalah vektor variabel yang

dipengaruhi oleh break (*regime variant*). Vektor w_t adalah vektor variabel penyebab perubahan *break*; yang dapat merupakan bagian dari x_t maupun z_t. Jika *lagged* variabel dependen juga merupakan bagian dari vektor x_t, maka Persamaan 12.21 disebut *threshold autoregression* (TAR); dan jika lagged variabel dependen juga merupakan bagian dari vektor z_t, maka Persamaan 12.21 disebut sebagai *self-exciting threshold autoregression* (SETAR). Estimasi Persamaan 12.21 dilakukan dengan menggunakan *conditional least squares*. Lihat Tong (2015) untuk refleksi dan rangkuman terkini dari keluarga estimator TAR.

Hamilton (1989) mengusulkan pemodelan perubahan regime (di sini disebut sebagai state) terjadi akibat berubahnya suatu variabel laten. Pemodelan yang dikenal sebagai Markov Switching Regression ini terdiri atas dua persamaan yakni persamaan regresi

$$y_t = \mu_S + x_t \beta + z_t \delta_s + \varepsilon_t \qquad (12.22)$$

di mana μ_S dan z_t adalah konstanta dan vektor variabel adalah yang dipengaruhi oleh regime. Sedangkan x_t adalah vektor variabel yang tidak terpengaruh oleh regime. Persamaan state yang cukup sederhana (dengan state = 2; dan state pada t hanya dipengaruhi oleh t-1) diberikan sebagai berikut

$$P(S_t = j|S_{t-1} = i) = p_{ij} = f(w_t) \qquad (12.23)$$

Estimasi Persamaan 12.22 dan 12.23 dilakukan secara serentak dengan menggunakan teknik Maximum Likelihood.

Prosedur regresi Markov Switching memiliki beberapa output lain yang menarik yakni (a) Matriks Transition Probabilities dan (b) Expected Duration. Transition probabilities memberikan

informasi kepada kita setiap p_{ij} yang berguna untuk memperkirakan probabilitas setiap state yang akan terjadi pada periode berikutnya jika kita tahu state saat ini. Sedangkan *expected duration* memberikan perhitungan rata-rata durasi (*in sample*) variabel dependen berada di setiap state.

Contoh 12.3

Kita akan menggunakan file CAPM_IHSG.dta. Sementara itu, model regresi yang diestimasi akan menggunakan ER_ASII sebagai variabel dependen dan IHSG_RP sebagai variabel penjelas (regresi CAPM). Datanya bersifat deret waktu dengan frekuensi bulanan yang mencakup periode Nov 2000 sampai dengan Oktober 2001. Periode data mencakup krisis global yang terjadi pada tahun 2008; misalnya, kita gunakan Maret 2008 sebagai penanda perubahan regime (terjadinya break). Kita ingin mengetahui apakah hasil regresi akan berbeda sebelum versus sesudah break (Maret 2008). Kita dapat melakukan pengujian dengan perintah berikut **estat sbknown, break(tm(2008m3)).**

```
Wald test for a structural break: Known break date

                                  Number of obs   =        228
Sample:      2000m11 - 2019m10
Break date:   2008m3
Ho: No structural break

             chi2(2)       =    3.1410
             Prob > chi2   =    0.2079

Exogenous variables:           IHSG_RP
Coefficients included in test: IHSG_RP _cons
```

TABEL 12.6. Pengujian Struktural Break, di mana Break Diketahui: Maret 2008

Hasil pengujiannya diberikan pada Tabel 12.7, yang menunjukkan bahwa hipotesis null tidak adanya structural break pada Maret 2008 tidak dapat ditolak. Dengan kata lain, regresi sebelum dan sesudah Maret 2008 kurang lebih sama. Selanjutnya kita mengasumsikan bahwa dalam periode itu data yang dimiliki paling banyak terdapat 1 break yang tidak diketahui datingnya. Pengujian terhadap hal ini dapat dilakukan dengan perintah **estat sbsingle**.

```
----+--- 1 ----+--- 2 ----+--- 3 ----+--- 4 ----+--- 5
..................................................    50
..................................................   100
..................................................   150
........

Test for a structural break: Unknown break date

                            Number of obs =       228

Full sample:                2000m11 - 2019m10
Trimmed sample:             2003m10 - 2016m12
Estimated break date:       2003m10
Ho: No structural break

     Test            Statistic            p-value
------------------------------------------------------------
     swald             17.6585             0.0034
------------------------------------------------------------
Exogenous variables:          IHSG_RP
Coefficients included in test: IHSG_RP _cons
```

TABEL 12.7. Pengujian *Structural Break*; Asumsi Terjadi Satu Break

Kita akan melakukan estimasi terhadap pemodelan SETAR di mana variabel lag dari ER_ASII digunakan sebagai threshold; dan variabel IHSG_RP digunakan sebagai variabel yang berubah karena regime. Perintah estimasi diberikan **threshold ER_ASII, threshvar(l. ER_ASII) regionvars(l.ER_ASII IHSG_RP)**.

Dapat dilihat pada Tabel 12.8 bahwa persamaan regresi CAPM ER_ASSI terhadap IHSG_RP terdiri dari 2 regime dengan threshold

```
Searching for threshold: 1
(Running 182 regressions)
..................................................    50
..................................................   100
..................................................   150
.............................

Threshold regression

                                                Number of obs     =        227
Full sample:    2000m12 - 2019m10               AIC               =  1989.1859
Number of thresholds =  1                       BIC               =  2009.7356
Threshold variable: L.ER_ASII                   HQIC              =  1997.4780

------------------------------------
Order      Threshold         SSR
------------------------------------
1         -90.5900        1.376e+06
------------------------------------
```

ER_ASII	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Region1						
ER_ASII						
L1.	-.0104045	.146571	-0.07	0.943	-.2976784	.2768694
IHSG_RP	1.787375	.14507	12.32	0.000	1.503043	2.071708
_cons	49.0068	29.32899	1.67	0.095	-8.47697	106.4906
Region2						
ER_ASII						
L1.	.0467603	.0554822	0.84	0.399	-.0619829	.1555035
IHSG_RP	1.395246	.0893106	15.62	0.000	1.220201	1.570292
_cons	-1.736219	6.444488	-0.27	0.788	-14.36618	10.89575

TABEL 12.8. Hasil Estimasi Model SETAR untuk ER_ASII

nilai ER_ASSI berada pada –90,59. Region 1 adalah untuk nilai ER_ASII di atas threshold; sedangkan region 2 adalah untuk yang di bawah threshold. Terlihat di sini bahwa area di atas threshold adalah area di mana excess return saham ASII lebih sensitif terhadap pergerakan pasar.

```
Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood = -1322.8274  (not concave)
Iteration 1:   log likelihood = -1320.8082  (not concave)
Iteration 2:   log likelihood = -1315.2613  (not concave)
Iteration 3:   log likelihood = -1309.3969  (not concave)
Iteration 4:   log likelihood = -1306.7679
Iteration 5:   log likelihood = -1305.2492
Iteration 6:   log likelihood = -1305.2159
Iteration 7:   log likelihood = -1305.2158

Markov-switching dynamic regression

Sample: 2000m11 - 2019m10                       No. of obs      =        228
Number of states =     2                        AIC             =    11.5107
Unconditional probabilities: transition         HQIC            =    11.5531
                                                SBIC            =    11.6160

Log likelihood = -1305.2158
```

ER_ASII	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
State1						
IHSG_RP	1.457638	.0650183	22.42	0.000	1.330204	1.585071
_cons	-.7600523	4.738364	-0.16	0.873	-10.04708	8.526971
State2						
IHSG_RP	3.252573	.6143803	5.29	0.000	2.04841	4.456736
_cons	254.8934	31.19156	8.17	0.000	193.759	316.0277
sigma	67.16169	3.498875			60.64251	74.38169
p11	.9734002	.014139			.9261919	.9907163
p21	.8491367	.1717274			.289142	.9873235

TABEL 12.9. Hasil Markov Switching Regression untuk ER_ASII

Kita memodelkan regresi Markov Switching dengan menggunakan spesifikasi ER_ASII sebagai variabel dependen; 2 state, tidak ada variabel yang state invarian, dan hanya ada satu variabel yang mempengaruhi state yakni IHSG-RP. Perintah estimasi diberikan sebagai berikut **mswitch dr ER_ASII, switch(IHSG_RP)**. Hasilnya

disajikan pada Tabel 12.9, di mana terdapat dua state dilihat dari konstanta yakni *standard return* (state 1) dan *high return* (state 2). State high return juga ditandai dengan sensitivitas yang tinggi terhadap pasar.

Matriks transition probabilities dan expected duration disajikan pada Tabel 12.10 dan 12.11. Dapat dilihat di sini bahwa terdapat kelembaman state p_{ij} di mana $i=j$ adalah tinggi; yaitu lebih dari 0,80. Dengan demikian, pola regresi CAPM untuk ER_ASII dapat dikatakan stabil. Saham ASII jauh lebih lama berada dalam *state standard return* (state 1 = 37,6 bulan) dibandingkan *high return* (state 2 = 1,2 bulan).

Number of obs = 228

Transition Probabilities	Estimate	Std. Err.	[95% Conf.	Interval]
p11	.9734002	.014139	.9261919	.9907163
p12	.0265998	.014139	.0092837	.0738081
p21	.8491367	.1717274	.289142	.9873235
p22	.1508633	.1717274	.0126765	.710858

TABEL 12.10. Transition Probabilities, Markov Switching Regression untuk ER_ASII

Number of obs = 228

Expected Duration	Estimate	Std. Err.	[95% Conf.	Interval]
State1	37.59421	19.983	13.54865	107.7156
State2	1.177667	.2381685	1.012839	3.458509

TABEL 12.11. Expected Duration, Markov Switching Regression untuk ER_ASII

Bab 13

Karakter Non-Stationary dan Kointegrasi

Pada bab ini kita akan menguraikan karakter terpenting dari data deret waktu yakni *nonstationarity* (suatu karakter yang telah diperkenalkan pada Bab 11 di Jilid 1). Terdapat berbagai bentuk ketidakstasioneran (*nonstationarity*), di antaranya yang terpenting dalam ekonometrika adalah *unit root*.

Realisasi data deret waktu dapat digambarkan sebagai suatu proses statistik: nilai (realisasi) suatu data ditentukan oleh pola (model) statistik tertentu. Terdapat banyak pola teoretis statistik yang mengkarakteristikkan proses data (disebut sebagai *data generating proses*/DGP)[1]. Di sini kita akan mengkarakteristikkannya sebagai suatu pola sederhana yang memiliki komponen deterministik yang linear serta komponen *stochastic* yang terdistribusi secara independen dan identik dengan rata-rata serta varians yang konstan. Meskipun sederhana, proses data semacam ini dianggap cukup untuk merepresentasikan berbagai variabel ekonomi yang ada.

Ekonometrika deret waktu memfokuskan pada karakterisasi proses data secara statistik. Sebagian besar teori ekonomi saat ini baru menunjukkan hubungan antara variabel ekonomi baik dalam kerangka sistem, optimisasi, maupun identitas. Hanya segelintir teori ekonomi yang memberikan penjelasan mengenai DGP suatu variabel, sehingga ekonometrika memilih untuk bertolak dari teori statistik dan mengasumsikan bahwa variabel tersebut mengikuti pola dimaksud (Hendry, Pagan, dan Sargan, 1984).

Di sini kita akan membahas dua tipe DGP, yakni stasioner dan tidak stasioner. Masing-masing sifat DGP itu memiliki implikasi teoretis dan praktis tersendiri bagi pemodelan ekonometris. Sebagai contoh, jika data yang dimiliki bersifat stasioner, maka pemodelan dengan menggunakan prosedur OLS standar yang telah dipelajari

[1] Lihat Harvey (1990) dan Hamilton (1994) untuk suatu survei.

selama ini sudah cukup memadai. Sebaliknya, jika data bersifat tidak stasioner, implementasi OLS berpotensi menghasilkan regresi palsu (*spurious regression*, Granger dan Newbold, 1974). Regresi palsu adalah fenomena di mana suatu persamaan regresi yang diestimasi memiliki signifikansi yang cukup baik, namun secara esensi tidak memiliki arti. Jadi, pemahaman atas karakteristik masing-masing DGP sangat diperlukan agar kita dapat melakukan pemodelan ekonometris yang tepat.

PROSES STATIONARY

Sebagai ilustrasi, kita akan memulai dari proses stokastik yang paling sederhana, yakni AR(1) sebagai berikut

$$y_t = \rho y_{t-1} + u_t$$
$$|\rho| < 1;\ u_t \sim IID(0, \sigma^2) \qquad (13.1)$$

Untuk kemudahan representasi persamaan matematis, kita akan menggunakan operator lag. Dengan operator lag maka y_{t-1} akan dituliskan sebagai Ly, dan y_{t-2} akan dituliskan sebagai L^2y, dan seterusnya (sedangkan $y_t = L^0y = y$). Dengan demikian, Persamaan 13.1 dapat dituliskan sebagai

$$y_t = \frac{1}{(1 - \rho L)} u_t \qquad (13.2)$$

Lebih lanjut, karena nilai ρ kita asumsikan memiliki nilai absolut kurang dari 1, maka proses AR(1) itu dapat direpresentasikan sebagai proses MA yang berorde tidak hingga

$$y_t = \frac{1}{(1-\rho L)} u_t = (1 + \rho L + \rho^2 L^2 + \ldots) u_t$$
$$= u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \ldots \quad (13.3)$$

Sifat ini dikenal juga sebagai *invertibility*. Dapat ditunjukkan bahwa nilai ekspektasi dan varians dari proses semacam ini bersifat konstan, yakni

$$E(y_t) = E(u_t) + \rho E(u_{t-1}) + \rho^2 E(u_{t-2}) + \ldots = 0$$
$$E[y_t - E(y_t)]^2 = E(y_t)^2 = \frac{\sigma^2}{(1-\rho^2)} \quad (13.4)$$

Sedangkan autocovariance antara observasi ke-t dan t-k dapat dituliskan sebagai[2]

$$\gamma_k = E[(y_t - \mu)(y_{t-k} - \mu)] = \rho^k \gamma_0 \quad (13.5)$$

dengan koefisien autokorelasi

$$r_{t,t-k} = \frac{E[(y_t - \mu)(y_{t-k} - \mu)]}{E[(y_t - \mu)]^2} = \frac{\gamma_k}{\gamma_0} = \rho^k \quad (13.6)$$

Dengan kata lain, koefisien autokorelasi juga bersifat konstan. Sifat DGP yang memenuhi Persamaan 13.4 dan 13.6 dikatakan sebagai data (*weakly*) stasioner. Selanjutnya Persamaan 13.1 dapat digeneralisir untuk melibatkan proses AR dengan lag p atau AR(p),

[2] Di sini γ_0 adalah varians $E(y_t)^2$.

atau

$$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + \ldots + \rho_p y_{t-p} + u_t$$
$$|\rho| < 1; u_t \sim IID(0, \sigma^2) \quad (13.7)$$

Proses ini dapat dinotasikan dengan menggunakan lag operator sebagai

$$A(L) y_t = u_t;$$
$$A(L) = (1 - \rho_1 L - \rho_2 L^2 - \ldots - \rho_p L^p) \quad (13.8)$$

Dengan penurunan aljabar yang rumit kita dapat menunjukkan bahwa untuk sembarang nilai p, proses ini bersifat stasioner. Proses ini memiliki nilai ekspektasi, varians, dan autokorelasi yang konstan[3]. DGP lainnya yang sering digunakan adalah proses Moving Average (MA) dengan orde q, yang dapat direpresentasikan sebagai

$$y_t = u_t + \theta_1 u_{t-1} + \ldots + \theta_q u_{t-q}$$
$$\theta_1; \ldots; \theta_q \in (0,1) \quad (13.9)$$

Dengan menggunakan lag operator,

$$y_t = B(L) u_t;$$
$$B(L) = (1 + \theta_1 L + \theta_2 L^2 - \ldots - \theta_q L^q) \quad (13.10)$$

[3] Lihat Harris dan Solis (2003) hal 5. untuk ilustrasi AR(2).

Proses ini juga bersifat stasioner. Sebagai ilustrasi, untuk MA(1) nilai ekspektasi, varians, dan autokorelasi dapat diberikan

$$E(y_t) = 0$$
$$E[y_t - E(y_t)]^2 = (1 + \theta^2)\sigma^2 \quad (13.11)$$
$$r_{t,t-k} = \frac{\theta}{(1 + \theta^2)}$$

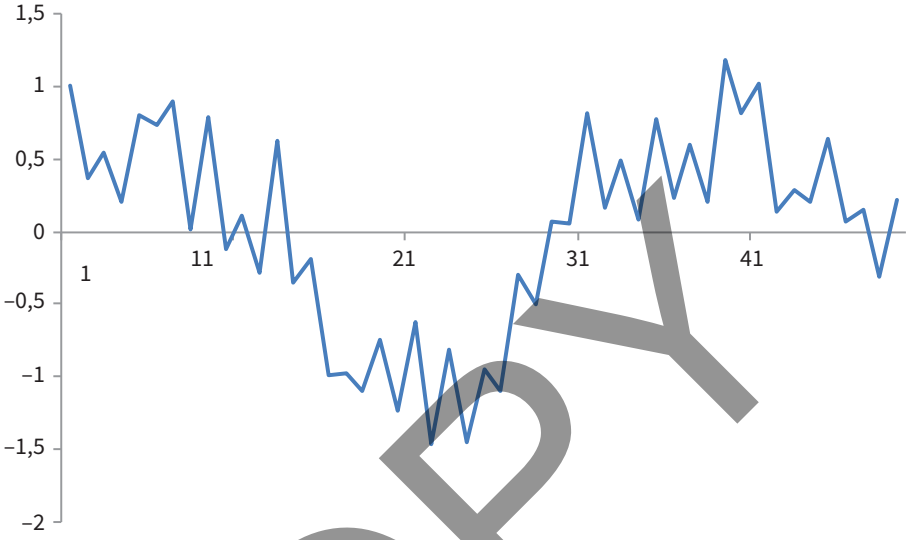
Akhirnya, suatu DGP juga dapat direpresentasikan sebagai Autoregressive Moving Average (ARMA) dengan orde p dan q (ARMA(p,q)) sebagai berikut:

$$A(L)y_t = B(L)u_t \quad (13.12)$$

Selama nilai ρ dan θ secara absolut kurang dari 1, dapat ditunjukkan bahwa DGP ini juga bersifat stasioner. Sebagai ilustrasi, untuk ARMA(1,1), nilai ekspektasi, varians, dan autokorelasi dapat diberikan sebagai

$$E(y_t) = (1 + \theta L)(1 - \rho L)^{-1}E(u_t) = 0$$
$$E[y_t - E(y_t)]^2 = \left(\frac{1 + \theta^2 + 2\rho\theta}{1 - \rho^2}\right)\sigma^2 \quad (13.13)$$
$$r_{t,t-k} = \frac{(1 + \rho\theta)(\rho + \theta)}{(1 + \theta^2 + 2\rho\theta)}$$

Secara grafis, suatu DGP yang bersifat stasioner dapat diberikan sebagai berikut



GAMBAR 13.1. Series Stasioner $y_t = 0{,}9y_{t-1} + u_t$; u_t ~ NIID(0,1)

Dapat dilihat di sini bahwa suatu DGP yang bersifat stasioner akan cenderung melakukan pembalikan arah yang cepat pada rata-rata yang konstan (*mean reversion*) serta memiliki pergerakan pada rentang yang juga kurang lebih konstan.

KONSEP EKSOGENITAS

Kita dapat memasukkan peran variabel lain dalam DGP suatu data (variabel). Satu ilustrasi yang sederhana dapat diberikan sebagai berikut:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 x_t + u_t \quad (13.14)$$

Asumsikan x_t memiliki sifat stokastik dengan DGP diberikan sebagai berikut:

$$x_t = \delta_1 x_{t-1} + v_t$$
$$|\delta| < 1; v_t \sim IID(0, \sigma_v{}^2) \quad (13.15)$$

Jika u_t dan v_t tidak berkorelasi, maka dimungkinkan untuk memperlakukan x_t sebagai variabel independen seperti pada prosedur OLS standar untuk mengestimasi Persamaan 13.14 (Harris dan Solis, 2003 hal. 7).

Formulasi seperti Persamaan 13.14 menyatakan bahwa x_t Granger Cause y_t atau, dengan kata lain, x_t bersifat eksogen terhadap y_t. *Strong exogeneity* adalah konsep yang lebih strict ketimbang *just exogeneity* yang baru saja diuraikan. Agar *strong exogeneity* berlaku maka formulasi berikut ini

$$x_t = \beta_0 + \beta_1 x_{t-1} + \beta_2 y_t + \varepsilon_t \quad (13.16)$$

tidak berlaku (dalam artian koefisien β_2 tidak signifikan). Perhatikan bahwa konsep *strong exogeneity* melarang keberadaan hubungan secara bersamaan (*contemporaneous*) antara y_t dan x_t.

Weak exogeneity adalah konsep yang lebih lemah yang memungkinkan pengaruh balik dari y_t ke x_t tetapi tidak secara bersamaan (*contemporaneous*). Di sini x_t adalah *weakly exogenous* jika formulasi berikut

$$x_t = \phi_0 + \phi_1 x_{t-1} + \phi_2 y_{t-1} + \eta_t \quad (13.17)$$

berlaku (dalam artian φ_2 adalah signifikan).

Akhirnya, dengan menggunakan konsep *weak exogeneity* dapat diuraikan konsep *super exogeneity*. *Super exogeneity* mensyaratkan tidak hanya *weak exogeneity* tercapai tetapi juga harus bersifat *structurally invariant*. Dengan kata lain, perubahan pada pola Persamaan 13.15 dan 13.17 tidak akan mempengaruhi Persamaan 13.14. Perubahan pada δ, φ_0, φ_1, dan φ_2 tidak akan mempengaruhi α_1 dan α_2.

Apa implikasi dari konsep eksogeneitas ini? Favero (2001, hal. 146) menunjukkan konsekuensi praktik dari konsep eksogeneitas:

1. Jika kita hanya tertarik pada inferensial parameter Persamaan 13.14, dan kita juga yakin x_t adalah *weakly exogenous*, maka estimasi Persamaan 13.14 adalah sudah memadai (tidak perlu mengestimasi Persamaan 13.17).
2. Jika kita memerlukan simulasi atas y_t, dan kita dapat meyakini bahwa x_t adalah *strongly exogenous*, maka estimasi Persamaan 13.14 sudah memadai.
3. Jika kita memerlukan pemodelan y_t untuk pengambilan kebijakan, maka penggunaan Persamaan 13.14 hanya dapat dilakukan jika x_t adalah *super exogenous*. Syarat ini diperlukan untuk menghindari kritik Lucas.

Seperti juga DGP univariat, Persamaan 13.14 dapat digeneralisasi menjadi model yang disebut *Autoregressive Distributed Lag* (ADL) berikut ini:

$$A(L)y_t = B(L)x_t + u_t \tag{13.18}$$

Jika kita ingin menggunakan model multivariabel yang lebih kompleks (y dan $x > 1$), maka skalar y dan x dapat diganti dengan notasi vektor.

Jika error term u_t dapat diasumsikan memiliki sifat klasik, yakni

1. Nilai ekspektasi/rata-rata sama dengan nol
2. Varians yang konstan
3. Tidak berkorelasi dengan term yang lampau (koefisien autokorelasi = 0),
4. Tidak berkorelasi dengan x_t.

Jadi, estimasi Persamaan 13.14 dapat dilakukan dengan OLS (lihat Johnston dan Dinardo, 1997 untuk bukti). Estimasi yang dihasilkan akan bersifat BLUE. Dengan demikian, pengujian atas sifat klasik dari u_t merupakan bagian yang esensial dalam pemodelan deret waktu. Di samping itu, secara implisit diskusi yang telah dilakukan menunjukkan bahwa estimasi model pada level seperti Persamaan 13.14 hanya dapat dilakukan jika variabel-variabel yang digunakan diasumsikan memiliki DGP yang bersifat stasioner.

PROSES NONSTASIONER

Seperti telah diuraikan sebelumnya, suatu series atau deret yang bersifat stasioner akan memiliki sifat nilai rata-rata serta varians yang konstan. Sebaliknya, suatu DGP yang bersifat nonstasioner memiliki rata-rata serta varians yang berubah (baik ditentukan secara deterministik/fungsional tertentu maupun random).

Suatu DGP yang tidak bersifat stasioner sederhana dapat diberikan dengan menggunakan Persamaan 13.1 tetapi dengan menetapkan $\rho = 1$, sehingga

$$y_t = y_{t-1} + u_t$$
$$u_t \sim IID(0, \sigma^2) \qquad (13.19)$$

Perhatikan bahwa DGP ini memiliki sifat stasioner pada diferens pertama, atau

$$\Delta y_t = y_t - y_{t-1} = u_t \tag{13.20}$$

Dengan mengasumsikan bahwa series atau deret bermula dari suatu nilai tertentu (y_0), substitusi berulang akan menghasilkan

$$\begin{aligned} y_T &= y_0 + u_1 + \ldots + u_T \\ &= y_0 + \sum_1^T u_t \end{aligned} \tag{13.21}$$

Rata-rata dan variansnya dapat dihitung sebagai

$$\begin{aligned} &E(y_T) = E(y_0) + E(u_1) + \ldots + E(u_T) = y_0 \\ &Var(y_T) = var(y_0 + \sum_{t=1}^T u_t) = t\sigma^2 \end{aligned} \tag{13.22}$$

Dapat dilihat di sini bahwa rata-rata masih konstan tetapi varians akan menuju tak hingga dengan semakin besarnya data.

Persamaan 13.19 bukanlah satu-satunya DGP yang bersifat tidak stasioner dan beberapa pola lainnya yang umum ditemukan adalah

1. *Random Walk with Drift*

$$y_t = \delta + y_{t-1} + u_t; u_t \sim IID(0, \sigma^2) \tag{13.23}$$

2. *Random Walk with Drift and Deterministic Trend*

$$y_t = \delta_0 + \delta_1 t + y_{t-1} + u_t; u_t \sim IID(0, \sigma^2) \quad (13.24)$$

3. *Stationary Around Deterministic Trend*

$$y_t = \delta_0 + \delta_1 t + \rho y_{t-1} + u_t; \quad \rho < 1; u_t \sim IID(0, \sigma^2) \quad (13.25)$$

Persamaan 13.19 dan 13.23 s/d 13.25 masih dapat digeneralisasi untuk mencakup AR(p). DGP seperti yang diberikan oleh model nonstasionary disebut juga sebagai proses *unit root* atau terintegrasi pada derajat pertama (*integrated degree* 1: I(1)).

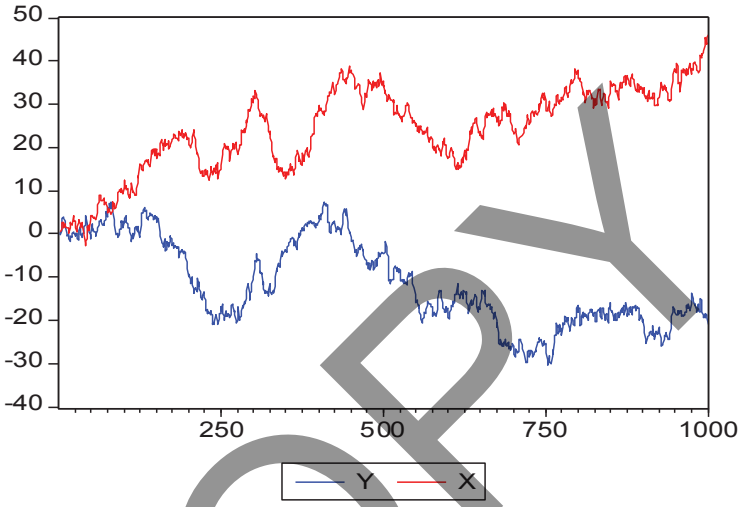
Penggunaan data deret waktu yang tidak bersifat stasioner memerlukan perlakuan khusus. Hal ini disebabkan potensi permasalahan *spurious regression* atau regresi palsu (Granger dan Newbold, 1974), yang timbul dari inferensial yang salah terhadap estimasi hubungan statistik di antara berbagai variabel. Secara sederhana, kita mengatakan bahwa terdapat hubungan yang bermakna di antara variabel *x* dan *y* yang sebenarnya tidak ada.

Sebagai ilustrasi, misalkan kita membuat suatu series hipotetis *y* dan *x* yang disusun sebagai

$$y_t = \alpha + y_{t-1} + e_t; e_t \sim NIID(0,1)$$
$$x_t = \beta + x_{t-1} + u_t; u_t \sim NIID(0,1) \quad (13.26)$$

Terlihat jelas bahwa kedua variabel ini tidak ada hubungan. Keduanya adalah series sintetis sebagai model *Random Walk* dengan *drift*. Jika kita membuat data set/sampel berukuran 1.000 observasi,

maka secara grafis satu realisasi dari kedua series tersebut dapat digambarkan berikut ini:



GAMBAR 13.2. Regresi Palsu (*Spurious Regression*)

Banarjee, Dolado, Galbraith, dan Hendry (1993, hal. 73-75) melakukan simulasi Monte Carlo dengan regresi berulang terhadap model seperti yang telah diberikan sebelumnya. Replikasi dilakukan sebanyak 10.000 kali dengan sampel berukuran 100. Dari eksperimen ini sebanyak 75,3% nilai statistik t koefisien X adalah lebih besar dari 1,96 (nilai kritis uji dua arah dengan α = 5%)[4]. Dengan demikian, sangat mungkin sekali bagi kita untuk menarik kesimpulan tentang adanya hubungan yang ”substantif” dari suatu estimasi yang sebenarnya adalah *spurious* (palsu).

Dengan demikian, dapat diambil kesimpulan bahwa ketika kita memiliki data yang bersifat deret waktu, maka suatu perlakuan

[4] Kita akan mereplikasi simulasi semacam ini pada Bab 17: Simulasi Monte Carlo.

khusus perlu dilakukan untuk memastikan bahwa data tidak bersifat tidak atau nonstasioner. Hal ini dilakukan dengan pengujian *unit root*.

Pada prinsipnya, pengujian unit root (atau disebut juga pengujian orde integrasi) adalah prosedur untuk memverifikasi bahwa koefisien ρ pada Persamaan 13.1 secara absolut lebih kecil dari 1. Nilai yang lebih kecil dari 1 ini diperlukan agar DGP yang dimiliki akan bersifat stabil: terdapat suatu tendensi bagi data untuk menjadi konvergen pada nilai tertentu. Pengujian unit root adalah suatu prosedur yang kompleks karena sangat tergantung pada bentuk DGP (komponen deterministik dan stokastik). Kita akan membahasnya secara khusus pada bagian berikutnya.

Perlakuan khusus lain bagi variabel yang memiliki karakteristik *nonstasionary* adalah yang menyangkut pemodelan ekonometris. Pemodelan ekonometris harus memperhatikan apakah di antara variabel-variabel yang diamati terdapat kointegrasi. Konsep kointegrasi serta pemodelan koreksi kesalahan (*error correction model*) yang menjadi implikasinya juga akan dibahas pada bagian tersendiri.

PENGUJIAN UNIT ROOT

Seperti telah diuraikan sebelumnya, pengujian stasioneritas data adalah hal yang penting dalam analisis data deret waktu. Pengujian yang tidak memadai dapat menyebabkan pemodelan menjadi tidak tepat sehingga hasil/kesimpulan yang diberikan dapat bersifat *spurious* (palsu).

Sejak tahun 1979, berawal dari karya Dickey dan Fuller, suatu prosedur formal untuk pengujian stasioneritas data telah disusun

(sering disebut juga sebagai uji *unit root* atau uji derajat integrasi: $I(d)$). Pada intinya, prosedur ini bertujuan untuk memverifikasi bahwa DGP adalah bersifat stasioner. DGP suatu data dapat mengambil berbagai bentuk yang dari uraian sebelumnya dapat dimodelkan sebagai suatu proses AR, MA, atau kombinasinya dengan orde p dan q tertentu.

Jika data bersifat stasioner, maka DGP dimaksud akan menunjukkan karakteristik rata-rata dan varians yang konstan serta nilai autokorelasi yang tidak tergantung pada titik waktu (*time invariant*). Hal yang sebaliknya akan terjadi jika data bersifat tidak stasioner. Pengujian ketidakstasioneran bukanlah suatu prosedur yang sederhana. Banyak hal yang perlu diperhatikan agar pengujian nonstasioneritas dapat bersifat valid. Beberapa aspek yang perlu diperhatikan di antaranya orde DGP (AR dan MA), keberadaan komponen deterministik (*drift* dan *trend*), *structural break*, dimensi data (tunggal atau panel) hingga *power* (probabilitas menolak hipotesis null yang salah), dan *size* (probabilitas menolak hipotesis null yang benar) alat uji itu sendiri. Pengembangan alat uji *unit root* adalah suatu area penelitian yang sangat aktif pada disiplin ilmu ekonometrika[5].

Berbagai alat pengujian derajat integrasi yang telah dikembangkan pada intinya bertanya apakah proses berikut:

$$\Delta y_t = \rho y_{t-1} + e_t; e_t \sim NIID(0, \sigma^2) \qquad (13.27)$$

bersifat stasioner. Stasioneritas mensyaratkan koefisien *autoregressive* memiliki nilai kurang dari 1 secara absolut. Kondisi ini dapat

[5] Fenomena *nonstationarity* sendiri masih menjadi perdebatan, apakah sifat pada populasi ataukah sampel? Lihat Nelson dan Plosser (1982) untuk pembahasan dan diskusi.

diperoleh dari solusi atas persamaan diferens berorde satu. Dapat diketahui dari teori Matematika, bahwa agar kondisi stabilitas tercapai (konvergen), maka syarat $|\rho| < 1$, harus terpenuhi[6].

Secara statistik, kita dapat *fitting* data terhadap model 1 secara langsung atau menempuh jalan yang dilakukan oleh Dickey dan Fuller (1976), yaitu melakukan modifikasi sebagai berikut:

$$\Delta y_t = (\rho - 1) y_{t-1} + e_t = \delta y_{t-1} + e_t \qquad (13.28)$$

dan menguji apakah δ adalah sama dengan nol. Jika hipotesis null tidak dapat ditolak, maka data yang diamati sangat kuat diduga bersifat tidak stasioner (terdapat unit root). Sebaliknya, jika hipotesis null dapat ditolak, kita akan lebih baik memodelkannya sebagai variabel stasioner. Kita akan menggunakan pendekatan yang terakhir di sini[7].

Dengan memasukkan komponen deterministik, yakni konstanta (*drift*) dan *time trend* serta komponen *stochastic* (AR dan MA), Persamaan 13.28 dapat digeneralisasi menjadi

$$\Delta yt = a0 + \gamma yt - 1 + a_2 t + \sum_{i=2}^{p} \beta i \Delta yt - i + 1 + \varepsilon_t \qquad (13.29)$$

di mana

[6] Untuk eksposisi matematis mengenai hal ini dapat dilihat pada Chiang and Wright (2005) dan Enders (2004).

[7] Masing-masing pendekatan memiliki kelebihan dan kelemahan masing-masing dan kita tidak akan membahasnya lebih lanjut di sini. Lihat Harris dan Solis (2003), Bab 3, untuk uraian lebih lanjut.

$$\gamma = -\left(1 - \sum_{i=1}^{p} ai\right)$$

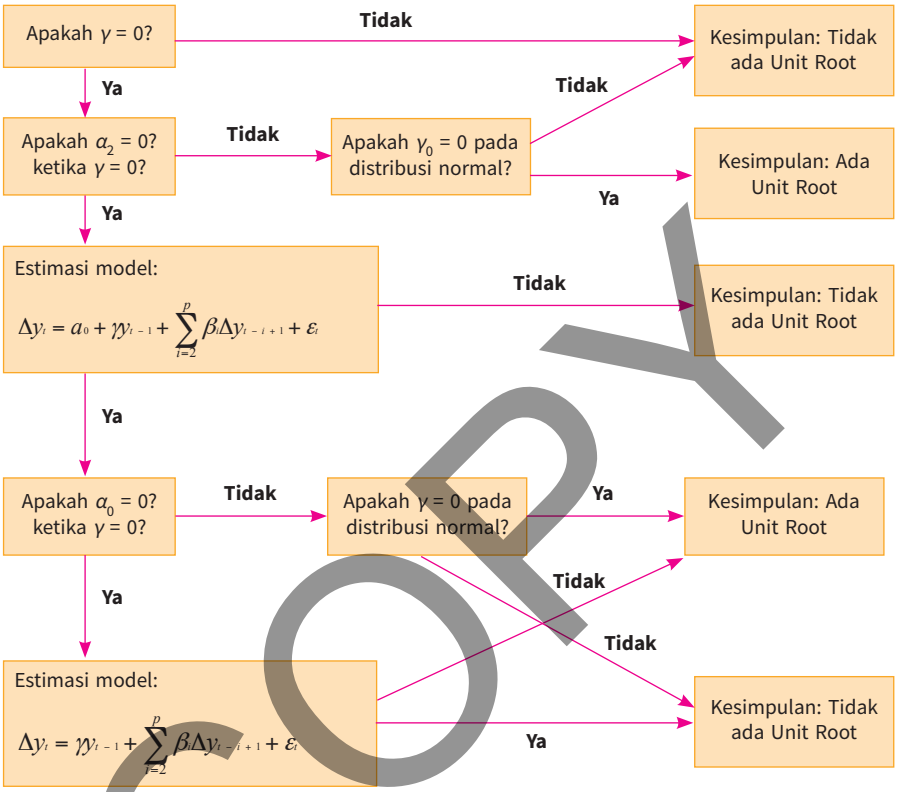
$$\beta i = \sum_{j=1}^{p} aj$$

Sekali lagi di sini titik perhatian adalah pada pengujian apakah koefisien γ adalah berbeda/tidak secara statistik dari nol (hipotesis null: ketidakstasioneran data).

Dickey dan Fuller (1979) telah menunjukkan bahwa koefisien γ tidak memiliki distribusi standar (distribusi t). Jadi, kita tidak dapat melakukan pengujian hipotesis dengan cara yang biasa (membandingkan nilai t dari model yang diestimasi dengan tabel statistik kritis: t). Lebih lanjut, nilai statistik kritis yang digunakan sangat tergantung pada bentuk model, yaitu komponen deterministik yang dimasukkan pada model.

Enders (2004) serta Harris dan Solis (2003) menguraikan beberapa masalah yang perlu diperhatikan terkait dengan penggunaan Persamaan 13.29, di antaranya:

1. *Trade off* komponen deterministik. Masuknya komponen deterministik akan menambah regresor sehingga mengurangi derajat kebebasan. Jika jumlah sampelnya kecil, hal ini akan menyebabkan inefisiensi (varians residual model yang semakin besar). Sedangkan mengeluarkannya akan berpotensi menimbulkan masalah misspesifikasi. Untuk mengatasi masalah ini suatu prosedur sekuensial telah disarankan oleh Doldado et al (1990); lihat Gambar 13.3.



GAMBAR 13.3. Prosedur Pengujian Unit Root Doldado et al (1990)

2. Penentuan lag yang optimum. DGP seperti yang diberikan Persamaan 13.29 dapat diperluas dengan memasukkan unsur MA. Selanjutnya, suatu proses MA dapat dimodelkan sebagai proses AR(∞) dengan syarat kondisi invertability terpenuhi. Namun demikian, tentunya kita tidak dapat mengestimasi suatu model dengan orde lag tak hingga seperti ini. Said dan Dickey (1984) menunjukkan proses dengan orde yang tak hingga ini dapat diaproksimasi dengan lag maksimal (*truncation lag*) =$T^{1/3}$,

di mana T adalah jumlah observasi[8]. Lag optimum yang dipakai sebagai model unit root selanjutnya dapat ditentukan melalui kriteria informasi, seperti AIC, SIC, dan HQIC.

3. Kemungkinan *multiple unit root*. Enders (2004) menunjukkan bahwa kebanyakan data ekonomi memiliki orde integrasi tidak lebih dari 2. Dickey dan Pantula (1987) memodifikasi Persamaan 13.29 untuk menyesuaikannya terhadap keberadaan derajat integrasi = 2. Ekstensi yang dilakukan cukup sederhana di mana kita terlebih dahulu membuat variabel diferens pertama, kemudian melakukan pengujian unit root (seperti telah diuraikan terdahulu) terhadap variabel dimaksud. Sebagai ilustrasi, modifikasi terhadap DGP AR(1) dapat dilakukan sebagai berikut:

$$\lambda \Delta^2 y_t = \lambda \Delta y_{t-1} + v_t; v_t \sim NIID(0,1) \qquad (13.30)$$

dan kita menguji apakah λ = 0.

4. Kemungkinan terjadinya *structural break*. Karena satu hal, seperti pergantian rezim kebijakan (dari ekspansioner ke kontraksi), perilaku suatu variabel mungkin mengalami perubahan. Perubahan perilaku ini mungkin dapat bersifat drastis sehingga nilai suatu variabel pada titik waktu (t) akan loncat dari periode sebelumnya (t – 1). Hal ini dapat dilihat seperti pada saat Devaluasi. Jika data yang diamati mencakup periode ini, maka pengujian *unit root* standar dapat memberikan hasil yang bias.

[8] Schwert (1989) memberikan alternatif rumus perhitungan lag yang optimum yakni l = int(12(T/100)$^{1/4}$), di mana int adalah operator untuk mencari integer terdekat (seperti int(1.4)=1).

Uji Dickey-Fuller

Pengujian *unit root* Dickey-Fuller (1979) dilakukan dengan menghitung nilai statistik hitung (statistik t) dari koefisien γ dan membandingkannya dengan nilai kritis. Nilai kritis di sini diperoleh bukan dari tabel distribusi t yang biasa digunakan dengan derajat kebebasan: jumlah observasi (T) dan level of significance (α) tertentu, melainkan tabel dari Dickey Fuller (1979) yang relevan.

Mengapa hal ini dilakukan? Dickey dan Fuller (1979) menunjukkan meskipun nilai kritis yang digunakan adalah dari tabel distribusi t, akan terjadi *over-rejection of null hypotheses*. Dengan kata lain, kita akan cenderung mengambil kesimpulan bahwa data yang diamati adalah bersifat stasioner padahal sebenarnya tidak.

Kesimpulan ini diperoleh dari hasil studi Monte Carlo sebagai berikut. Asumsikan DGP data adalah bentuk sederhana dari Persamaan 13.28 dengan koefisien $\rho = 1$ (yang berarti data adalah tidak stasioner). Selanjutnya, kita dapat membuat suatu series sintetis, y_t di mana y_t mengikuti formulasi berikut dan e_t diperoleh dari suatu proses random normal standar.

$$y_t = y_{t-1} + e_t; e_t \sim NIID(0,1) \quad (13.31)$$

Kita selanjutnya akan membuat suatu sampel yang terdiri dari T observasi y_t (misalnya, $T = 25$). Lalu, lakukan regresi

$$y_t = \rho y_{t-1} + e_t; e_t \sim NIID(0,1) \quad (13.32)$$

di mana sekarang nilai ρ adalah tidak lagi direstriksi = 1. Kita replikasi regresi semacam ini sebanyak katakan 10.000 kali (seperti yang dilakukan oleh Fuller, 1976). Dengan demikian, kita akan

memiliki 10.000 nilai ρ yang berbeda beserta nilai statistik hitungnya (sebut saja dengan τ). Nilai τ ini akan kita tabulasi secara kumulatif dan diperoleh persentile terbawah 1%, 5%, dan 10%. Di sini kita telah memperoleh nilai kritis dari proses unit root dengan model yang diberikan oleh Persamaan 13.32 untuk T = 25.

Proses ini dapat diulang untuk DGP yang berbeda-beda, yang merupakan bentuk yang lebih restriktif dari Persamaan 13.32 serta pada jumlah observasi yang berbeda-beda (misalnya, T = 50, 100, dan sebagainya). Fuller (1976) telah melakukan eksperimen ini untuk T = 25, 50, 100, 250, 500, dan ∞ untuk berbagai tingkat signifikansi (α). Terdapat tiga nilai statistik kritis (analog dengan t) yang dibedakan berdasarkan model uji unit root, yakni model dasar (tanpa konstanta dan tren waktu): τ, model dengan konstanta: τ_μ serta model dengan konstanta dan tren: τ_τ. Tabel 13.2 menunjukkan sebagian dari nilai statistik tersebut beserta nilai t statistik asimtotis (T = ∞).

Ukuran Sampel/ Tingkat Signifikansi	Nilai Kritis u/τ			Nilai Kritis u/τ_μ			Nilai Kritis u/τ_τ		
	0,01	0,05	0,10	0,01	0,05	0,10	0,01	0,05	0,10
25	-2,66	-1,95	-1,60	-3,75	-3,00	-2,63	-4,38	-3,60	-3,24
50	-2,62	-1,95	-1,61	-3,58	-2,93	-2,60	-4,15	-3,50	-3,18
100	-2,60	-1,95	-1,61	-3,51	-2,89	-2,58	-4,04	-3,45	-3,18
250	-2,58	-1,95	-1,62	-3,46	-2,88	-2,57	-3,99	-3,43	-3,13
500	-2,58	-1,95	-1,62	-3,44	-2,87	-2,57	-3,98	-3,42	-3,13
∞	-2,58	-1,95	-1,62	-3,43	-2,86	-2,57	-3,96	-3,41	-3,12
Distribusi *t*									
∞	-2,33	-1,65	-1,28	-2,33	-1,65	-1,28	-2,33	-1,65	-1,28

TABEL 13.1. Statistik Dickey Fuller (DF)

Perhatikan bahwa nilai kritis akan semakin kecil dengan semakin besarnya data dan α. Di samping itu, nilai statistik *t* juga selalu lebih kecil (secara absolut) dari statistik DF. Dengan demikian, seperti telah diuraikan sebelumnya penggunaan statistik *t* sebagai nilai kritis akan menghasilkan *over rejection null hypothesis nonstationarity*. Statistik DF tidak pernah konvergen ke arah distribusi *t* berapa pun besarnya sampel, sehingga penggunaan distribusi *t* untuk pengujian derajat integrasi adalah bias.

Perkembangan Uji Unit Root

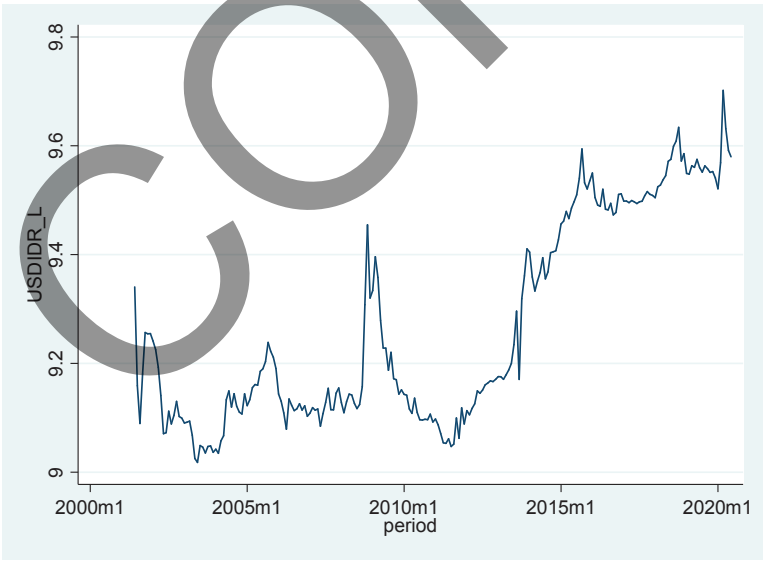
Sejak studi seminal Dickey dan Fuller (1976), pengujian ketidakstasioneran data telah menjadi area kajian yang sangat aktif. Berbagai metode pengujian baru telah ditawarkan dengan berbagai klaim kelebihannya. Di sini kita akan membahas beberapa metode dimaksud, tetapi hanya sebatas review. Kita akan membahas esensi dari metode itu dan tidak menyentuh detail teknisnya.

Salah satu uji unit root lain yang banyak digunakan adalah Uji Phillips dan Perron (1988). Uji Phillip dan Perron merupakan alternatif dari ADF, di mana ketimbang menggunakan *lag term* untuk menghilangkan autokorelasi yang ada pada DGP, lebih baik menggunakan suatu teknik nonparametris untuk mengkoreksi nilai *t* hitung. Penggunaan lag term dinilai akan mengurangi derajat kebebasan dan juga *power of test*. Dengan kata lain, uji Phillips-Perron mengklaim *power* yang lebih baik daripada uji ADF. Kwiatkowski, Phillips, Schmidt, dan Shin (KPSS, 1992) melakukan pengujian unit root bertolak dari hipotesis null *stationarity* (bukan *nonstationarity* seperti ADF dan Phillips-Perron). Pengujian ini dinilai memiliki kinerja yang biasa saja sehingga tidak banyak digunakan dalam penelitian empiris. Namun demikian, Harris dan Solis (2003)

menyarankan penggunaan ADF dan KPSS secara bersama pada pengujian unit root dalam kerangka uji konfirmasi.

Contoh 13.1

Kita akan menggunakan data USDIDR.dta dan memeriksa apakah variabel USDIDR memiliki karakteristik unit root. Kita akan menggunakan uji Dickey Fuller terlebih dahulu. Sebelum pengujian, mengingat data ini memiliki karakteristik numeris yang cukup besar (5 digit angka), maka variabel dikonversi ke dalam bentuk log dengan nama USDIDR_L. Membuat grafik garis dari variabel dapat digunakan sebagai langkah awal untuk menduga pola uji unit root (lihat Gambar 13.4).



GAMBAR 13.4 Grafik Garis USDIDR_L

Dari Gambar 13.4, tampaknya kita dapat menggunakan model dengan konstanta (drift). Lag yang digunakan dapat diperoleh dari formulasi

truncation lag Said-Dickey yakni $T^{1/3}$ = 6 (≈(229)^1/3). Perintah STATA untuk pengujian ADF diberikan sebagai **dfuller USDIDR_L, drift regress lags(6)**. Hasil pengujiannya ditunjukkan pada Tabel 13.2. Terlihat di sini bahwa USDIDR_L adalah suatu series atau deret yang mengalami unit root, kemungkinan orde 1. Koefisien pada variabel L1.USDIDR_L memiliki angka uji statistik $Z(t)$ sebesar –0,515 dengan p value sebesar 0,3034. Dengan demikian, hipotesis null unit root tidak dapat ditolak.

Augmented Dickey-Fuller test for unit root — Number of obs = 222

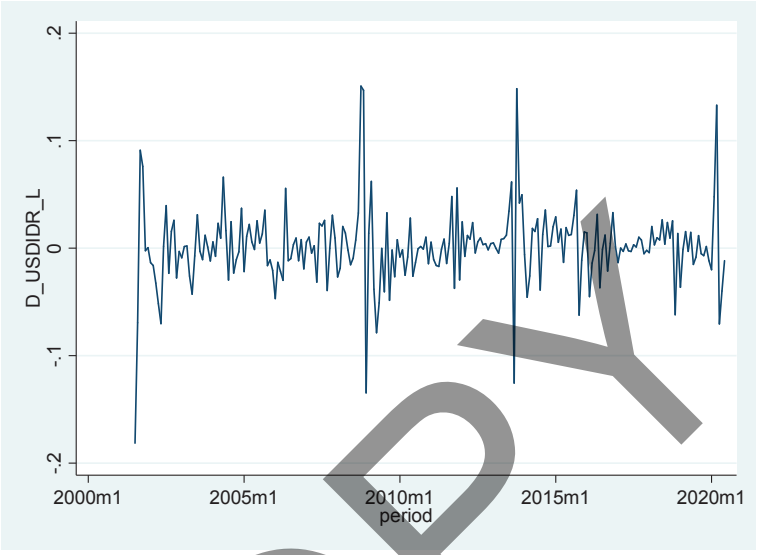
	Test Statistic	1% Critical Value	5% Critical Value	10% Critical Value
		Z(t) has t-distribution		
Z(t)	-0.515	-2.344	-1.652	-1.286

p-value for Z(t) = 0.3034

D.USDIDR_L	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
USDIDR_L						
L1.	-.0064565	.0125273	-0.52	0.607	-.0311492	.0182362
LD.	-.0513263	.0692069	-0.74	0.459	-.1877408	.0850883
L2D.	-.1008735	.0692988	-1.46	0.147	-.2374691	.0357221
L3D.	-.0074136	.0697335	-0.11	0.915	-.1448661	.1300388
L4D.	.0731186	.0708505	1.03	0.303	-.0665355	.2127728
L5D.	-.044085	.0697124	-0.63	0.528	-.1814959	.0933259
L6D.	-.0065062	.0658182	-0.10	0.921	-.1362412	.1232288
_cons	.0614997	.1160471	0.53	0.597	-.1672419	.2902414

TABEL 13.2. Pengujian ADF pada Series USDIDR_L

Selanjutnya untuk memastikan tingkat orde integrasi (yang diduga sama dengan satu), kita terlebih dahulu membuat variabel baru yang merupakan bentuk diferens dari USDIDR_L; sebut saja sebagai D_USDIDR_L. Series baru ini memiliki pola seperti yang ditunjukkan pada Gambar 13.5.



GAMBAR 13.5. Grafik Garis D_USDIDR_L

Pola yang terdapat pada Gambar 13.5 mengindikasikan kemungkinan USDIDR_L memang terintegrasi pada orde satu. Pengujian ADF pada series D_USDIDR_L dilakukan dengan spesifikasi tanpa konstanta; karena kita melihat variabel ini bergerak di sekitar angka nol. Perintah STATA diberikan sebagai **dfuller D_USDIDR_L, regress lags(6)**. Output dari pengujian ADF ini disajikan pada Tabel 13.3. Di sini nilai statistik uji $Z(t)$ sebesar –6,304 dengan p value sebesar 0,000. Dengan demikian, hipotesis null unit root ditolak.

Penerimaan hipotesis null unit root variabel di tingkat level (USDIDR_L) disertai dengan penolakan hipotesis null di tingkat differens (D_USDIDR_L), yang menghasilkan kesimpulan bahwa USDIDR_L memiliki orde integrasi sebesar 1 (I(1)). Kita dapat melakukan pengujian alternatif untuk mengkonfirmasi bahwa USDIDR_L mengalami masalah unit root seperti halnya dengan uji

Augmented Dickey-Fuller test for unit root Number of obs = 221

	Test Statistic	Interpolated Dickey-Fuller 1% Critical Value	5% Critical Value	10% Critical Value
Z(t)	-6.304	-3.470	-2.882	-2.572

MacKinnon approximate p-value for Z(t) = 0.0000

D.D_USDIDR_L	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
D_USDIDR_L						
L1.	-1.260271	.1999156	-6.30	0.000	-1.654337	-.8662047
LD.	.2031517	.1839062	1.10	0.271	-.1593577	.565661
L2D.	.0955824	.1667928	0.57	0.567	-.2331935	.4243584
L3D.	.0944309	.150328	0.63	0.531	-.2018901	.3907519
L4D.	.1679559	.1261554	1.33	0.184	-.0807171	.4166289
L5D.	.1009522	.0954729	1.06	0.292	-.0872407	.289145
L6D.	.0705595	.0652983	1.08	0.281	-.0581542	.1992732
_cons	.0019264	.0022892	0.84	0.401	-.0025861	.0064389

TABEL 13.3 Pengujian ADF pada Series D_USDIDR_L

Phillip Perron dan KPSS[9]. Hasil Pengujian dengan kedua prosedur ini untuk variabel USDIDR_L disajikan pada Tabel 13.4. Pada pengujian Phillip Perron (perintah: **pperron USDIDR_L**); statistik uji $Z(t)$ sebesar –0,853 dengan p value: 0,8030; sehingga menghasilkan kesimpulan bahwa series USDIDR_L mengalami unit root. Sedangkan pengujian KPSS (perintah: **kpss USDIDR_L**) memberikan hipotesis null tidak ada unit root; sementara series bersifat stasionari. Dapat dilihat bahwa statistik uji hingga lag yang optimal (ke-14) memiliki nilai yang lebih besar dari nilai kritis bahkan dengan α = 10%

[9] Pengujian KPSS bukan merupakan default routine STATA. Baum (2018) telah membuat routine untuk uji unit root tersebut dengan perintah kpss. Pembaca harus menginstalasi terlebih dahulu ssc install kpss.

(atau hipotesis null ditolak). Dengan demikian, pengujian KPSS juga memberikan kesimpulan yang sama: USDIDR_L adalah *unit root*.

Phillips-Perron test for unit root — Number of obs = 228; Newey-West lags = 4

	Test Statistic	Interpolated Dickey-Fuller 1% Critical Value	5% Critical Value	10% Critical Value
Z(rho)	-2.296	-20.227	-13.956	-11.171
Z(t)	-0.853	-3.467	-2.881	-2.571

MacKinnon approximate p-value for Z(t) = 0.8030

USDIDR_L	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
USDIDR_L L1.	.9869176	.0128704	76.68	0.000	.9615563	1.012279
_cons	.1223005	.1193104	1.03	0.306	-.1128027	.3574036

KPSS test for USDIDR_L

Maxlag = 14 chosen by Schwert criterion
Autocovariances weighted by Bartlett kernel

Critical values for H0: USDIDR_L is trend stationary

10%: 0.119 5% : 0.146 2.5%: 0.176 1% : 0.216

Lag order	Test statistic
0	3.06
1	1.59
2	1.09
3	.834
4	.679
5	.576
6	.503
7	.448
8	.405
9	.371
10	.343
11	.32
12	.3
13	.283
14	.269

TABEL 13.4. Pengujian Phillip-Perron (a) dan KPSS (b) untuk USDIDR_L

KOINTEGRASI DAN ERROR CORRECTION MODEL

Pada bagian sebelumnya telah diuraikan bahwa regresi OLS di antara variabel-variabel yang bersifat *nonstasionary* (terintegrasi orde d, $I(d)$) adalah palsu atau *spurious*. Tanpa perlakuan yang memadai kita tidak dapat membedakan apakah hubungan yang diperoleh adalah benar bemakna atau hanya sekedar disebabkan oleh interaksi DGP pada data.

Salah satu cara untuk mengidentifikasi hubungan di antara variabel yang bersifat *nonstasionary* adalah dengan melakukan pemodelan koreksi kesalahan. Dengan syarat bahwa pada sekelompok variabel *nonstationary* terdapat kointegrasi, pemodelan koreksi kesalahan adalah valid. Syarat ini dinyatakan dalam teorema representasi Engle-Granger (1987).

Dalam bagian ini kita akan membahas dua bentuk dari model koreksi kesalahan. Model yang pertama adalah model koreksi kesalahan persamaan tunggal (*single equation error correction model*). Model ini digunakan jika kita dapat mengidentifikasi dengan baik arah hubungan kointegrasi dan yang ada pada sekelompok variabel: mana yang variabel dependen dan mana yang variabel independen. Model kedua adalah *Vector Error Correction Model* (VECM); yang memiliki paradigma VAR di mana variabel-variabel dianggap endogen dan diestimasi secara serentak.

Kointegrasi: Konsep dan Pengujian

Sebelumnya telah dijelaskan bahwa langkah pertama dalam pemodelan koreksi kesalahan adalah memastikan bahwa variabel-variabel yang diamati telah terkointegrasi. Fenomena kointegrasi bukan merupakan kejadian yang umum. Suatu kombinasi linear

variabel yang bersifat *nonstationary* biasanya juga bersifat *nonstationary*, sedangkan kombinasi linear variabel yang *stationary* dan *nonstationary* juga akan bersifat *nonstationary* dengan derajat integrasi terbesar ada pada kelompok variabel tersebut (Brooks, 2014, hal. 387).

Suatu ilustrasi mungkin dapat membantu. Misalkan kita akan mengestimasi hubungan antara variabel y (yang diasumsikan sebagai variabel dependen) dan variabel independen x_1 dan x_2 sebagai berikut:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \quad (13.33)$$

u adalah suatu kombinasi linear dari y, x_1, dan x_2. Jika y, x_1, dan x_2 salah satunya/lebih bersifat *nonstationary* (terintegrasi dengan orde d; $I(d)$), maka u umumnya juga bersifat *nonstationary*, atau

$$u = y - \beta_0 - \beta_1 x_1 - \beta_2 x_2 \sim I(d) \quad (13.34)$$

Pengecualian akan terjadi jika variabel-variabel dimaksud adalah terkointegrasi.

Generalisasi terhadap kondisi ini dapat diuraikan sebagai berikut (Engle dan Granger, 1987): Notasikan w_t sebagai suatu vektor variabel yang berukuran $k \times 1$, maka komponen w_t disebut terkointegrasi pada orde (d,b) jika

a. Semua komponen w_t adalah orde $I(d)$
b. Paling tidak terdapat satu vektor koefisien α sedemikian rupa sehingga $\alpha' w_t$ memiliki derajat integrasi yang lebih rendah sebesar b: $I(d\text{-}b)$.

Pada contoh tersebut, jika y, x_1, dan x_2 adalah I(1), maka y, x_1, dan x_2 dikatakan terkointegrasi jika u adalah I(0).

Suatu hubungan kointegrasi dapat dipandang sebagai hubungan jangka panjang (ekuilibrium). Suatu set variabel dapat saja terdeviasi dari pola ekuilibrium, namun diharapkan terdapat mekanisme jangka panjang yang akan mengembalikan variabel-variabel dimaksud ke pola hubungan ekuilibrium.

Terdapat cukup banyak contoh hubungan-hubungan di antara variabel ekonomi yang memungkinkan terjadinya deviasi jangka pendek, namun dengan mekanisme penyeimbang kembali dalam jangka panjang. Doktrin PPP, misalnya, meskipun sangat jarang ditemui dalam jangka pendek (< 4 tahun), namun dari hasil konsensus tampaknya sepakat tentang kemungkinan tercapainya dalam jangka panjang (Rogoff, 1996). Hubungan kointegrasi lainnya dapat ditemui pada *spot* dan *forward rate*, nilai ekuitas dan dividen, real shock terhadap inflasi, dan sebagainya.

Sesuai dengan definisi kointegrasi yang telah disebutkan tampaknya pengujian kointegrasi adalah suatu teknik yang bersifat *straightforward*. Jika suatu kelompok variabel (yang seluruhnya adalah $I(d)$) diduga memiliki kointegrasi dengan bentuk linear tertentu, maka pengujian dilakukan dengan melihat apakah kombinasi linear tersebut adalah $I(d\text{-}b)$. Lebih lanjut, karena umumnya variabel makroekonomi adalah I(1), maka pengujian dilakukan dengan melihat apakah kombinasi linearnya adalah I(0).

Teknik pengujian ini disebut dengan uji kointegrasi pendekatan residual. Secara lebih konkret, misalkan kita hendak menguji apakah Persamaan 13.33 adalah suatu kointegrasi (di mana y, x_1, dan x_2 adalah I(1)), sehingga yang kita lakukan adalah menguji apakah u

(Persamaan 13.34) adalah I(0). Kita dapat menggunakan salah satu uji unit root yang telah dibahas sebelumnya yakni uji Dickey Fuller.

Namun demikian, harus diperhatikan bahwa kita tidak menerapkan ADF pada suatu series data melainkan pada residual model. Engle dan Granger (1987) menunjukkan nilai kritis ADF yang biasa digunakan adalah tidak valid sebagai alat uji hipotesis *nonstationarity*. Sebagai penggantinya, Mac Kinon (1991) telah melakukan simulasi Monte Carlo dan mentabulasikan nilai kritis yang lebih tepat dalam bentuk fungsi respons. Kerangka pengujian dengan menggunakan nilai kritis yang baru ini disebut uji Engle-Granger (EG).

Struktur hipotesis pengujian dilakukan pada model berikut ini:

$$\Delta\hat{u}_t = \gamma\hat{u}_{t-1} + \varepsilon_t \tag{13.35}$$

di mana struktur hipotesis adalah

$$\begin{aligned} H_0&: \gamma = 0; \quad \hat{u}_t \sim I(1) \\ H_1&: \gamma < 0; \quad \hat{u}_t \sim I(0) \end{aligned} \tag{13.36}$$

Statistik uji adalah nilai t statistik dari γ yang dibandingkan dengan tabel nilai kritis Mac Kinon (1991), bukan Dickey Fuller. Nilai statistik kritis dari Mac Kinon (1991) diperoleh melalui rumus sebagai berikut

$$C(p) = \varphi_\infty + \varphi_1 T^{-1} + \varphi_2 T^{-2} \tag{13.37}$$

di mana φ_∞, φ_1, dan φ_2 diperoleh melalui spesifikasi model yang relevan (berdasarkan asumsi adanya konstanta, tren waktu, derajat signifikansi, dan jumlah variabel independen), serta T adalah jumlah observasi.

Error Correction Model

Jika suatu hubungan kointegrasi terdeteksi pada kombinasi linear sekelompok variabel, maka kita dapat melakukan pemodelan koreksi kesalahan. Hal ini didasarkan pada teorema representasi Engle-Granger (1987).

Salah satu teknik yang sering digunakan untuk melakukan estimasi parameter model koreksi kesalahan adalah metode Engle-Granger 2 tahap. Tahap pertama dilakukan dengan memastikan bahwa semua variabel adalah I(1) dan melakukan regresi OLS sesuai dengan bentuk kointegrasi yang diasumsikan. Uji residual regresi akan dilakukan untuk melihat apakah terdapat suatu kointegrasi atau tidak. Jika hasilnya adalah positif, maka dapat dilanjutkan ke tahap 2.

Tahap kedua dilakukan dengan menggunakan residual yang diperoleh pada tahap pertama untuk mengestimasi model koreksi kesalahan berikut ini:

$$\Delta y_t = \beta_1 \Delta x_t + \beta_2 \hat{u}_{t-1} + v_t; \quad \hat{u}_{t-1} = y_{t-1} - \hat{\tau} x_{t-1} \tag{13.38}$$

Bagian kedua dari Persamaan 13.38 disebut dengan persamaan kointegrasi. Ini adalah estimasi residual persamaan yang diestimasi pada tahap 1. Model koreksi kesalahan akan dianggap valid dan stabil jika nilai β_2 adalah negatif dengan nilai absolut kurang dari satu (dan tentu saja signifikan). Dengan kata lain, terdapat mekanisme ekuilibrium yang menghilangkan suatu proporsi tertentu dari disekuilibrium antara y dan x yang terjadi dalam jangka pendek. Semakin besar nilai β_2 (lebih negatif), semakin cepat proses penyesuaian yang terjadi. Nilai β_2 yang tidak signifikan atau positif

signifikan dapat diinterpretasikan sebagai tidak adanya kointegrasi atau model tidak stabil (cenderung eksplosif). Sedangkan koefisien β_1 adalah respons jangka pendek dan temporer.

Jika kita gagal menemukan kointegrasi di antara sekelompok variabel, maka pemodelan yang tepat adalah bentuk *first difference*. Hal ini dilakukan dengan melakukan regresi OLS pada bentuk *first difference* variabel-variabel yang diamati. Pemodelan *first difference* tidak disarankan apabila suatu hubungan kointegrasi terdeteksi berpotensi menghilangkan kandungan informasi yang penting.

Contoh 13.2

Kita akan menggunakan data yang ada pada file Bitcoin.dta. Kita juga akan menerapkan metodologi Engle-Granger (EG, 1987) untuk menguji kointegrasi dan mengestimasi Error Correction Model (ECM). Demi efisiensi penyajian, kami akan menyerahkan kepada pembaca untuk melakukan pengujian unit root dan memverifikasi bahwa variabel-variabel yang digunakan: btc, vix, sp500, dan comm_idx adalah *unit root* (I(1)). Regresi OLS dengan variabel dependen btc dan regressor: vix, sp500, dan comm_idx disajikan kembali pada Tabel 13.5[10]. Regresi ini juga merupakan tahap pertama dari metodologi EG.

Pola residual dari regresi pada Tabel 3.5 (btc_r) ditunjukkan pada Gambar 13.6. Secara kualitatif kita dapat menduga bahwa series tersebut adalah "mungkin" stasioner. Pengujian formal dilakukan berdasarkan metode ADF dengan lag 4 (≈67^1/3) dan model

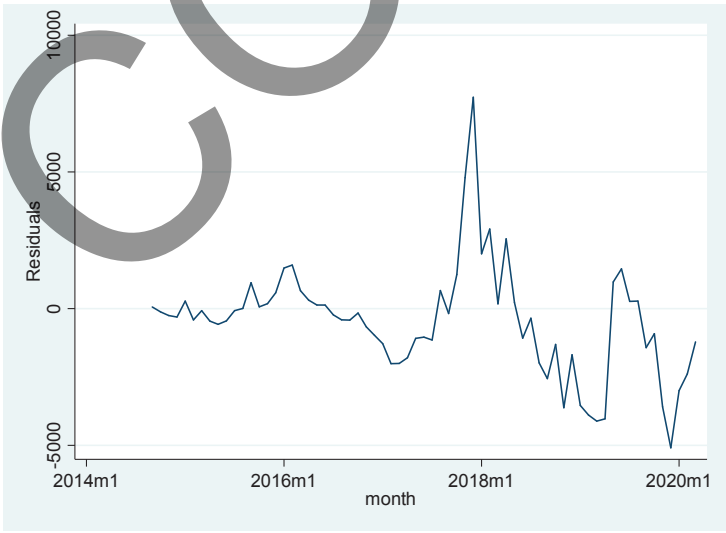
[10] Schaffer (2010) telah membuat suatu routine untuk melakukan estimasi ECM dengan metode EG; perintah **egranger**. Pembaca dapat melakukan instalasi terlebih dahulu sebelum menggunakan. Hasil yang diperoleh memiliki sedikit perbedaan dengan yang dilakukan secara manual pada contoh ini.

Source	SS	df	MS	Number of obs	=	67
				F(3, 63)	=	64.78
Model	739362467	3	246454156	Prob > F	=	0.0000
Residual	239673685	63	3804344.2	R-squared	=	0.7552
				Adj R-squared	=	0.7435
Total	979036151	66	14833881.1	Root MSE	=	1950.5

btc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
vix	61.96125	38.78667	1.60	0.115	-15.54772	139.4702
sp500	8.450175	.6695065	12.62	0.000	7.112273	9.788077
comm_idx	7.188383	4.96699	1.45	0.153	-2.737354	17.11412
_cons	-20733.15	2404.492	-8.62	0.000	-25538.14	-15928.15

TABEL 13.5. Regresi Tahap Pertama Prosedur EG, Variabel Dependen: btc

dengan konstanta diberikan melalui perintah: **dfuller btc_r, drift regress lags(4)**.



GAMBAR 13.6. Grafik Garis dari Residual EG Tahap 1

Hasil pengujian ADF diberikan pada Tabel 13.6. Di sini nilai statistik hitung $Z(t)$ sebesar –2,549 akan dibandingkan dengan nilai kritis Mac Kinon (1991) yang diberikan pada Persamaan 13.37. Nilai kritis ini dapat dihitung (dengan spesifikasi n = 3, T = 67, model dengan konstanta dan tren serta α = 5%) sebagai berikut $C(p) = -3{,}7429 - 8{,}352(67)^{-1} - 13{,}41(67)^{-2}$; yang sebesar –3,87. Jadi, nilai $Z(t)$ secara absolut lebih kecil dari nilai kritis, sehingga adanya kointegrasi di antara variabel-variabel yang dianalisis (btc, vix, sp500, dan comm_idx) tidak didukung oleh data.

Tahap kedua dari prosedur EG dilalui dengan melakukan regresi di mana variabel (dependen dan independen) dalam bentuk first difference memasukkan lag orde satu dari residual regresi tahap pertama (*Error Correction Term*; ECT). Perintah STATA diberikan sebagai berikut: **reg D.btc D.vix D.sp500 D.comm_idx L.btc_r**.

```
Augmented Dickey-Fuller test for unit root          Number of obs   =        62

                                 ------------ Z(t) has t-distribution ------------
                       Test          1% Critical       5% Critical      10% Critical
                     Statistic            Value             Value            Value
------------------------------------------------------------------------------------
 Z(t)                  -2.549              -2.395            -1.673           -1.297
------------------------------------------------------------------------------------
p-value for Z(t) = 0.0068
```

D.btc_r	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
btc_r						
L1.	-.3211184	.1259753	-2.55	0.014	-.5734773	-.0687595
LD.	-.0049821	.1506986	-0.03	0.974	-.3068676	.2969034
L2D.	.1609153	.1468073	1.10	0.278	-.1331749	.4550056
L3D.	.0452396	.1497144	0.30	0.764	-.2546744	.3451535
L4D.	.0182896	.1401338	0.13	0.897	-.2624321	.2990112
_cons	-159.4762	195.955	-0.81	0.419	-552.0212	233.0689

TABEL 13.6. Pengujian ADF untuk Series btc_r

Hasil dari prosedur EG tahap kedua disajikan pada Tabel 13.6. Koefisien ECT yang memiliki nilai sebesar –0,194 telah memenuhi

kondisi stabilitas. Statistik hitung (*t* statistics) dari koefisien ECT sebesar –2,29; kita melakukan pengujian satu arah dan memperoleh nilai kritis untuk α masing-masing sebesar 5% dan 1% (dengan perintah **display invttail(67, 0.05)** dan **display invttail(67, 0.01)**) adalah 1,67 dan 2,38. Nilai statistik hitung secara absolut lebih besar dari nilai kritis α = 5% tetapi lebih kecil dari α = 1%. Nilai *p* value hitung dapat diperoleh dengan perintah **display ttail(67,2.29)**, sebesar 0,0126.

Source	SS	df	MS	Number of obs	=	66
				F(4, 61)	=	1.56
Model	10803796.1	4	2700949.03	Prob > F	=	0.1962
Residual	105568478	61	1730630.79	R-squared	=	0.0928
				Adj R-squared	=	0.0334
Total	116372274	65	1790342.68	Root MSE	=	1315.5

D.btc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
vix						
D1.	-13.42172	54.30018	-0.25	0.806	-122.0016	95.1582
sp500						
D1.	1.005726	2.997782	0.34	0.738	-4.988709	7.00016
comm_idx						
D1.	5.759266	7.205269	0.80	0.427	-8.648559	20.16709
btc_r						
L1.	-.1944301	.0848366	-2.29	0.025	-.3640713	-.0247889
_cons	32.37024	181.3947	0.18	0.859	-330.3509	395.0914

TABEL 13.7. Regresi Tahap Kedua Prosedur EG, Variabel Dependen: btc

Dengan demikian, tahap kedua EG cenderung mengatakan adanya *Error Correction Mechanism* (ECM). Hasil ini bertolak belakang dengan pengujian kointegrasi yang dilakukan sebelumnya. Kita dapat mengambil sikap konservatif jika memperoleh hasil *mixed* seperti ini; yang tidak ada kointegrasi.

VECTOR ERROR CORRECTION MODEL (VECM)

Johansen (1991 dan 1995) mengembangkan kerangka kerja estimasi kointegrasi dan *error correction model* = ECM dari Engle dan Granger (1987) dengan menggunakan VAR yang diperkenalkan oleh Sims (1980). Model ECM mempostulasikan suatu hubungan (terkointegrasi) linear dari variabel (yang diasumsikan) dependen terhadap suatu set variabel (yang diasumsikan) eksogen (asumsi: arah kausalitas). Seperti dijelaskan oleh Sims (1980), arah kausalitas sulit diidentifikasi dan mungkin estimasi secara serentak harus dilakukan dengan mengasumsikan bahwa semua variabel yang bersifat endogen akan memiliki kinerja yang lebih baik. Di samping itu, penggunaan ECM juga mengharuskan pengujian unit root terlebih dahulu atas seluruh variabel yang digunakan; hal ini dilakukan untuk meyakini bahwa variabel-variabel tersebut adalah I(1) sehingga residual yang dihasilkan seharusnya I(0) untuk mendukung hipotesis kointegrasi (Enders. 2004).

Dengan *Vector Error Correction Model* (VECM), kedua isu tersebut ditangani secara simultan. Estimasi error correction model adalah berbentuk simultan; yaitu suatu VAR dengan variabel yang dapat memiliki derajat integrasi yang berbeda. Uji kointegrasi berdasarkan teknik *maximum likelihood* selanjutnya dilakukan dengan melihat peringkat matriks VAR untuk menguji hipotesis adanya kointegrasi (serta berapa jumlah kointegrasi yang ada). Suatu VECM sederhana yang terdiri dari 2 variabel endogen[11] dan lag yang sama sebesar *p* dapat direpresentasikan sebagai berikut:

[11] Routine VECM pada STATA cukup fleksibel. Pembaca dapat menentukan suatu set variabel sebagai *strongly exogenous* jika diinginkan. Estimasi VECM hanya dilakukan untuk set variabel endogen; sementara variabel eksogen akan diperlakukan seperti koefisien jangka pendek.

$$\Delta y_{1t} = \alpha_1(y_{2,t-1} - \beta y_{1,t-1}) + \sum_{p=1}^{P} \Delta y_{1,t-p} + \sum_{p=1}^{P} \Delta y_{2,t-p} + u_t$$
$$\Delta y_{2t} = \alpha_1(y_{2,t-1} - \beta y_{1,t-1}) + \sum_{p=1}^{P} \Delta y_{1,t-p} + \sum_{p=1}^{P} \Delta y_{2,t-p} + v_t \quad (13.39)$$

di mana kointegrasi diekspresikan sebagai

$$y_{2,t-1} = \beta y_{1,t-1} \quad (13.40)$$

Sistem Persamaan 13.39 diestimasi dengan menggunakan teknik *maximum likelihood*. Menurut Schopohl, Wichmann, dan Brooks (2019), hal pertama yang dilakukan adalah mencari jumlah lag yang optimal. Selanjutnya, dengan menggunakan lag yang optimal tersebut dilakukan pengujian kointegrasi; yaitu mengidentifikasi jumlah rank matriks VAR. Estimasi VECM dilakukan dengan menggunakan lag yang optimal dan jumlah kointegrasi yang telah diperoleh sebelumnya. Terakhir, terhadap VECM yang diperoleh dapat dilakukan pilihan *post estimation* seperti *stability*, *check*, *normality test*, dan *serial correlation*.

Contoh 13.3

Kita sekali lagi menggunakan data Bitcoin.dta. Di sini kita akan menspesifikasikan variabel btc, vix sp500, dan comm_idx sebagai variabel endogen. Lag maksimum yang digunakan adalah 12; karena datanya bersifat bulanan. Perintah untuk mencari lag yang optimal adalah **varsoc btc vix sp500 comm_idx, maxlag(12)**. Kita akan memperoleh hasil seperti yang disajikan pada Tabel 13.7. Dapat

dilihat di sini bahwa seluruh kriteria informasi AIC, HQIC, dan SBIC merekomendasikan penggunaan lag 12 (yang merupakan lag maksimum) sebagai lag yang optimal.

Selection-order criteria
Sample: 2015m9 - 2020m3 Number of obs = 55

lag	LL	LR	df	p	FPE	AIC	HQIC	SBIC
0	-1350.33				2.9e+16	49.2483	49.3048	49.3943
1	-1147.33	405.99	16	0.000	3.2e+13	42.4484	42.7307	43.1784
2	-1131.56	31.535	16	0.011	3.3e+13	42.4569	42.965	43.7708
3	-1121.93	19.272	16	0.255	4.2e+13	42.6883	43.4222	44.5861
4	-1113.4	17.053	16	0.382	5.8e+13	42.9601	43.9198	45.4419
5	-1091.88	43.044	16	0.000	5.1e+13	42.7593	43.9448	45.825
6	-1070.73	42.292	16	0.000	4.8e+13	42.5721	43.9835	46.2218
7	-1048.66	44.156	16	0.000	4.7e+13	42.3511	43.9883	46.5848
8	-1011.87	73.576	16	0.000	2.9e+13	41.5952	43.4582	46.4128
9	-995.849	32.037	16	0.010	4.3e+13	41.5945	43.6833	46.9961
10	-940.062	111.58	16	0.000	1.8e+13	40.1477	42.4623	46.1332
11	-869.767	140.59	16	0.000	6.4e+12	38.1733	40.7138	44.7428
12	-761.931	215.67*	16	0.000	1.1e+12*	34.8338*	37.6001*	41.9872*

Endogenous: btc vix sp500 comm_idx
Exogenous: _cons

TABEL 13.8. Penentuan Lag yang Optimal dari VECM

Selanjutnya, pengujian kointegrasi dapat dilakukan dengan menggunakan lag 12 dan untuk model kita hanya memasukkan konstanta. Perintah uji kointegrasi (rank VAR) adalah **vecrank btc vix sp500 comm_idx, trend(constant) lags(12) levela**. Hasil uji kointegrasi (Tabel 13.8) dibaca secara sekuensial dengan membandingkan nilai trace statistik dengan critical value (1% atau 5%). Maximum rank sebesar 0 adalah hipotesis null tidak ada kointegrasi (versus hipotesis alternatif terdapat persamaan

terkointegrasi minimal (1). Maximum rank sebesar 1 adalah hipotesis null paling banyak 1 persamaan terkointegrasi versus minimum 2, dan seterusnya. Di sini trace statistiks adalah 8,056 yang lebih besar dari nilai kritis rank 3 (baik menggunakan 1% maupun 5%); sehingga dapat disimpulkan bahwa rank matriks adalah 3.

Johansen tests for cointegration

Trend: constant — Number of obs = 55
Sample: 2015m9 - 2020m3 — Lags = 12

maximum rank	parms	LL	eigenvalue	trace statistic	5% critical value	1% critical value
0	180	-910.55812		297.2551	47.21	54.46
1	187	-825.72617	0.95426	127.5912	29.68	35.65
2	192	-785.95588	0.76453	48.0506	15.41	20.04
3	195	-765.9587	0.51673	8.0562	3.76	6.65
4	196	-761.93058	0.13625			

TABEL 13.9. Pengujian Kointegrasi dari VECM

Estimasi VECM dilakukan dengan menggunakan spesifikasi hanya konstanta; yaitu rank matriks VAR = 3 dan lag yang optimal = 12. Perintah untuk melakukan estimasi adalah **vec btc vix sp500 comm_idx, trend(constant) rank(3) lags(12)**. Hasil estimasi sangat panjang; yang pada Tabel 13.9 kita hanya menyampaikan beberapa yang dipandang cukup penting.

Bagian atas dari hasil estimasi menampilkan persamaan jangka pendek atas setiap variabel endogen. Dapat dilihat di sini nilai Chi Square untuk persamaan jangka pendek D.btc adalah sangat rendah; *p* value = 0,99. Hal ini menunjukkan bahwa ECM dengan variabel dependen btc tidak memiliki kemampuan berupa penjelasan statistik yang memadai. Kita juga melihat koefisien ECT (_ec1, _ec2, dan _ec3) dari persamaan jangka pendek btc di samping tidak

```
Sample:  2015m9 - 2020m3                        Number of obs      =        55
                                                AIC                =   34.94395
Log likelihood = -765.9587                      HQIC               =   37.69612
Det(Sigma_ml)  =  1.47e+07                      SBIC               =   42.06086

Equation           Parms      RMSE     R-sq      chi2     P>chi2
------------------------------------------------------------------
D_btc                 48   1970.27   0.7676   23.11431   0.9991
D_vix                 48   2.36826   0.9688   217.1399   0.0000
D_sp500               48   54.6193   0.9673   207.1547   0.0000
D_comm_idx            48   7.27655   0.9888   615.2491   0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
D_btc						
_ce1						
L1.	-1.109316	1.019115	-1.09	0.276	-3.106744	.8881125
_ce2						
L1.	-747.9608	760.8054	-0.98	0.326	-2239.112	743.1904
_ce3						
L1.	12.13283	11.23909	1.08	0.280	-9.895379	34.16104
btc						
LD.	.5010943	1.002425	0.50	0.617	-1.463623	2.465812
L2D.	.6695309	.9075542	0.74	0.461	-1.109243	2.448305
L3D.	.5402126	.8189923	0.66	0.510	-1.064983	2.145408
L4D.	.938982	.7138797	1.32	0.188	-.4601964	2.33816
L5D.	.9281619	1.191392	0.78	0.436	-1.406923	3.263247
L6D.	.4482359	2.154011	0.21	0.835	-3.773549	4.670021

... Output tidak ditampilkan

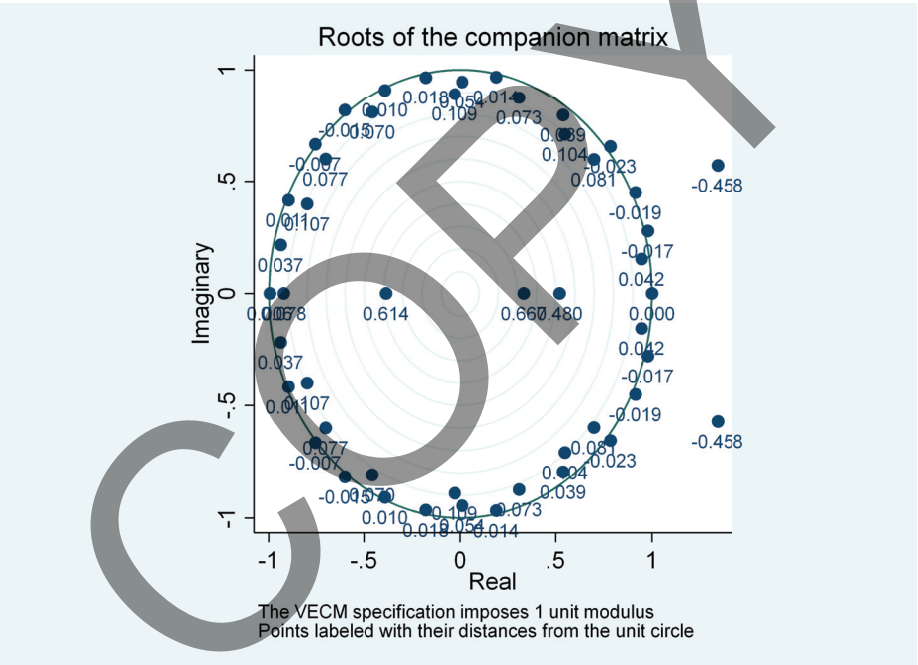
```
Cointegrating equations

Equation           Parms      chi2     P>chi2
------------------------------------------------
_ce1                   1   86.94489     0.0000
_ce2                   1   17.47184     0.0000
_ce3                   1   117.4797     0.0000
------------------------------------------------

Identification:  beta is exactly identified

                 Johansen normalization restrictions imposed
```

beta	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_ce1						
btc	1	.	.	.	.	.
vix	-1.14e-13	.	.	.	.	.
sp500	0	(omitted)				
comm_idx	-127.0598	13.62655	-9.32	0.000	-153.7673	-100.3522
_cons	41367.93	.	.	.	.	.
_ce2						
btc	8.81e-20	.	.	.	.	.
vix	1	.	.	.	.	.
sp500	0	(omitted)				
comm_idx	-.0366439	.0087666	-4.18	0.000	-.0538262	-.0194617
_cons	-8.812203	.	.	.	.	.
_ce3						
btc	-2.08e-17	.	.	.	.	.
vix	-7.11e-15	.	.	.	.	.
sp500	1	.	.	.	.	.
comm_idx	-13.78299	1.271634	-10.84	0.000	-16.27535	-11.29063
_cons	2251.552	.	.	.	.	.

TABEL 13.10. Pengujian Kointegrasi dari VECM

memenuhi persyaratan stabilitas (negatif dan memiliki nilai absolut di bawah satu) juga tidak signifikan. Segmen paling bawah dari output mengidentifikasi 3 persamaan jangka panjang (*cointegrating equation*) yakni (1) btc-vix-comm_idx, (2) vix-btc-comm_idx, dan (3) sp500-btc-vix-comm_idx.

Pengujian stabilitas dapat dilakukan dengan perintah **vecstable,**



GAMBAR 13.7. Pengujian Stabilitas VECM

graph dlabel segera setelah melakukan estimasi. Hasilnya diberikan dalam bentuk tabel dan grafik; di sini kita akan menggunakan grafik. VECM yang stabil akan memiliki maksimum *K-r* unit modulus; di mana *K* adalah jumlah variabel endogen dan *r* adalah jumlah rank matrik VAR (Enders, 2004). Dalam contoh ini karena kita memiliki

4 variabel endogen dan rank VAR = 3; maka jumlah modulus untuk persyaratan stabilitas VECM adalah maksimum 1 (= 4-3).

Bab 14

Regresi Data Panel

Salah satu bentuk struktur data yang sering digunakan dalam studi ekonometrika adalah data panel. Data dengan karakteristik panel adalah data yang berstruktur deret waktu sekaligus *cross section*. Data semacam ini dapat diperoleh seperti dengan mengamati serangkaian observasi *cross section* (antarindividu) selama suatu periode tertentu.

Data semacam ini memiliki keunggulan terutama karena bersifat robust terhadap beberapa jenis pelanggaran asumsi Gauss Markov, yakni heterokedastisitas dan normalitas (Wooldridge, 2019). Di samping itu, dengan perlakuan tertentu struktur data seperti ini juga dapat diharapkan memberikan informasi yang lebih banyak (*high informational content*) yaitu suatu aspek yang sangat diinginkan bagi penelitian empiris yang bernilai tinggi. Ada dua jenis struktur data panel, yakni (a) *short panel* di mana *cross section* unit (N) > time unit (T) dan (*b*) *long-panel* di mana *cross section* unit (N) < time unit (T).

Namun demikian, penggunaan data semacam ini bukannya tidak memberikan beban ekstra. Di samping biaya akuisisi yang cukup tinggi, beban ekstra juga timbul dari masalah kompleksitas analisis dan perlakuan data. Sebagai contoh, terdapat kompleksitas perlakuan estimasi yang berbeda di antara *short panel*, *long panel*, dan yang kurang lebih "*balanced*" (antara N dan T)[1]. Namun demikian, *trade off* yang terjadi akibat biaya yang lebih tinggi versus manfaat empiris dinilai masih cukup menguntungkan.

Pembahasan akan dilakukan terhadap teknik panel data konvensional; yang umumnya dilakukan terhadap $N \times T$ yang

[1] Penulis memberikan tanda petik "" pada kata *balanced* untuk membedakannya dengan istilah *balanced* yang dipahami dalam ekonometrika secara umum. Balanced adalah kondisi di mana setiap unit cross section *n* memiliki jumlah unit time series yang sama. Kondisi unbalanced akan terjadi jika hal ini tidak dipenuhi; ada cross section *n* yang memiliki jumlah unit time series yang tidak sama.

"*balanced*". Setelah itu, kita akan menjelaskan mengenai data panel yang bersifat spesifik: short panel dan long panel. Ekonometrika data panel termasuk topik yang mengalami perkembangan pesat akhir-akhir ini; uraian yang komprehensif mengenai hal ini dapat dirujuk ke Baltagi (2011), Wooldridge (2010), dan Pesaran (2015).

REPRESENTASI DAN ESTIMASI OLS

Suatu model data panel linear *k* variabel dapat direpresentasikan sebagai berikut:

$$y_{it} = \alpha_0 + \sum_{j=1}^{k} \alpha_j X_{j,it} + u_{it} \quad (14.1)$$

di mana index *i* dan *t* menunjukkan identifikasi observasi dari unsur cross section dan deret waktu. Karena data bersifat panel, maka residual dari regresi ini memiliki komponen yang umum dan spesifik: *cross section* dan deret waktu. Karakter ini secara matematis ditunjukkan oleh Persamaan 14.2.

$$u_{it} = e + v_i + w_t \quad (14.2)$$

di mana *e* adalah komponen residual yang bersifat umum (*common error component*), v_i adalah komponen yang spesifik dari *cross section* dan w_t adalah komponen yang spesifik dari deret waktu. Komponen v_i dan w_t disebut juga sebagai *unobserved heterogeneity* dan harus diestimasi melalui data.

Komponen yang spesifik dapat disebabkan oleh berbagai hal. Style atau gaya manajemen misalnya akan menyebabkan perilaku suatu variabel (seperti *Return on Equity*) akan berbeda-beda meskipun

perusahaan yang diamati berasal dari satu industri (heterogenitas *cross section*). Implementasi suatu regulasi atau adanya event atau peristiwa berpengaruh (krisis ekonomi) juga dapat menyebabkan suatu periode waktu memiliki dampak khusus.

Apabila dapat diasumsikan bahwa tidak ada komponen yang spesifik baik pada *cross section* maupun deret waktu, maka estimasi Persamaan 14.1 dapat dilakukan dengan metode OLS yang biasa digunakan (sering juga disebut sebagai *pooled OLS*). Interpretasi statistik juga dilakukan dengan cara standar seperti halnya penggunaan data yang berdimensi satu: *cross section* atau deret waktu saja.

Sebaliknya, apabila diyakini bahwa ada heterogenitas yang signifikan baik pada *cross section* dan/atau deret waktu, maka pemodelan residual harus dilakukan secara eksplisit. Komponen residual harus dimodelkan secara benar sesuai dengan spesifikasi empiris, karena model yang tidak tepat dapat menimbulkan bias. Hal ini terutama disebabkan karena akibat kurang tepatnya model, variabel independen akan memiliki korelasi dengan residual.

Apabila asumsi residual hanya terdiri dari *common component* tidak dapat diterima, maka ada dua alternatif model yang dapat digunakan yakni (a) Model efek tetap: *Fixed Effect Model* (FEM) dan (b) Model efek random: *Random Effect Model* (REM). Pemodelan ini didasarkan pada asumsi apakah karakter residual spesifik ini bersifat konstan atau random.

MODEL EFEK TETAP (FEM)

Suatu data panel dapat dipandang (diasumsikan) memiliki dua faktor tidak terobservasi yang akan mempengaruhi variabel dependen;

yakni yang bersifat (1) konstan antarobservasi *cross section* dan (2) konstan antarobservasi deret waktu. Dengan kata lain, dalam kasus di mana $t = T$ dan $i = N$, maka model data panel dengan k variabel independen dapat ditulis sebagai

$$y_{it} = \alpha_0 + \sum_{j=1}^{k} \alpha_j X_{j,it} + u_{it} \tag{14.3}$$

di mana

$$u_{it} = e + \sum_{i=1}^{N-1} D_i^c v_i + \sum_{t=1}^{T-1} D_t^\tau w_t \tag{14.4}$$

di mana D_i^c dan D_t^τ adalah dummy variabel sebanyak $N - 1$ dan $T - 1$ untuk mengidentifikasi komponen residual spesifik *cross section* dan deret waktu yang bersifat konstan. Dengan memasukkan Persamaan 14.4 ke 14.3, maka diperoleh

$$y_{it} = \alpha_0 + \sum_{j=1}^{k} \alpha_j X_{j,it} + \sum_{i=1}^{N-1} D_i^c v_i + \sum_{t=1}^{T-1} D_t^\tau w_t + e \tag{14.5}$$

Persamaan 14.5 adalah model variabel kategoris yang dapat diestimasi dengan OLS[2]. Jika kita dapat mengasumsikan bahwa v_i dan w_t tidak berkorelasi dengan variabel independen, maka estimator OLS akan bersifat tidak bias.

Pemodelan efek tetap (*fixed effect*) memiliki beberapa kelemahan (Gujarati, 2008 dan Heij et al, 2004) yakni:

a. Masalah berkurangnya derajat kebebasan (*degree of freedom*) akibat bertambahnya jumlah parameter yang harus diestimasi.

[2] Karena karakter ini, maka model efek tetap disebut juga sebagai *Least Square Dummy Variable* (LSDV).

Sebagai contoh, jika struktur data yang dimiliki terdiri atas 10 unit *cross section* dan 5 unit deret waktu, maka kita harus mengestimasi 13 variabel dummy tambahan. Rendahnya derajat kebebasan dapat menimbulkan inefisiensi pada parameter yang diestimasi.

b. Multikolinearitas yang diakibatkan oleh banyaknya variabel *dummy* yang diestimasi.

c. Keterbatasan kemampuan estimasi, terutama jika terdapat variabel yang bersifat tidak berubah berdasarkan waktu (*time invariant*). Parameter variabel tersebut tidak dapat diestimasi karena pengaruhnya tidak dapat dipisahkan dari dummy waktu.

d. Kemungkinan korelasi di antara komponen residual spesifik (*cross section* dan deret waktu).

Kita dapat menguji apakah pemodelan efek tetap jauh lebih baik dibandingkan dengan model residual gabungan (*pooled OLS*) melalui uji *F*. Apabila model dengan efek tetap ternyata lebih superior dari *pooled OLS*, maka nilai koefisien determinasi (R^2) model tersebut harus lebih tinggi secara signifikan. Secara eksplisit dapat digunakan rumus sebagai berikut:

$$F = \frac{(R^2_{UR} - R^2_R)/m}{(1 - R^2_{UR})/(n-k)} \qquad (14.6)$$

di mana R^2_{UR} adalah nilai koefisien determinasi model tanpa restriksi (FEM), sedangkan R^2_R adalah nilai koefisien determinasi model dengan restriksi (*pooled OLS*). Di sini m adalah jumlah variabel dummy, n adalah jumlah observasi, dan k adalah parameter pada FEM (konstanta, variabel penjelas, dan variabel dummy).

MODEL EFEK RANDOM

Salah satu kelemahan utama FEM adalah tidak dapat melakukan estimasi terhadap *time invariant* dan *cross section invariant variables*. Padahal mungkin dalam studi yang sedang dilakukan variabel-variabel tersebut harus diestimasi. Dengan demikian, diperlukan suatu alternatif teknik estimasi. Salah satu yang cukup popular adalah model efek random (REM). Misalkan kita melakukan estimasi terhadap suatu sistem data panel dengan k variabel independen sebagai berikut:

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it} \quad (14.7)$$

Model *efek random* digunakan ketika *unobserved effect* a_i, dapat diasumsikan tidak berkorelasi dengan satu/lebih variabel independen, atau

$$Cov(x_{itj}, a_i) = 0, t = 1,2,\ldots,T; j = 1,2,\ldots,k \quad (14.8)$$

Kita dapat memodelkan Persamaan 14.7 dengan menggunakan *composite error term* sebagai berikut:

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + v_{it} \quad (14.9)$$

Karena a_i selalu ada pada *composite error term* selama setiap periode waktu, maka v_{it} mengalami *serial correlation*. Jadi, dapat ditunjukkan bahwa

$$Corr(v_{it}, v_{is}) = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_u^2}; t \neq s \quad (14.10)$$

Kita dapat mengkoreksi keberadaan serial correlation dengan prosedur GLS. Namun demikian, agar prosedur ini efektif maka datanya harus memiliki N yang lebih besar ketimbang T (disebut *short panel*). GLS ditempuh dengan melakukan transformasi ke setiap regresor dan variabel dependen melalui koefisien λ, di mana

$$\lambda = 1 - \left(\frac{\sigma_u^2}{\sigma_u^2 + T\sigma_a^2} \right)^{1/2} \tag{14.11}$$

Dalam praktek, nilai λ tidak diketahui sehingga harus diestimasi melalui data. Terdapat beberapa usulan perhitungan λ namun di sini kita akan menggunakan yang disarankan oleh Wooldrige (2010), yakni

$$\hat{\lambda} = 1 - \left(\frac{1}{1 + T(\hat{\sigma}_a^2 / \hat{\sigma}_u^2)} \right)^{1/2};$$

$$\hat{\sigma}_a^2 = (NT(T-1)/2 - k)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \hat{v}_{it} \hat{v}_{is}; \tag{14.12}$$

$$\hat{\sigma}_u^2 = \hat{\sigma}_v^2 - \hat{\sigma}_a^2$$

di mana $\hat{\sigma}_v^2$ adalah kuadrat standar error dari pooled OLS.

Estimator λ ini selanjutnya digunakan untuk mentransformasikan Persamaan 14.7 menjadi

$$y_{it} - \hat{\lambda}\bar{y}_i = \beta_0(1 - \hat{\lambda}) + \beta_1(x_{it1} - \hat{\lambda}\bar{x}_{i1}) + \ldots + \beta_k(x_{itk} - \hat{\lambda}\bar{x}_{ik}) + (v_{it} - \hat{\lambda}\bar{v}_i) \tag{14.13}$$

Transformasi ini menghasilkan *quasi demeaned* data sementara estimator yang diperoleh dari regresi disebut *random effect estimator*[3].

Pengujian apakah terdapat Random Effect versus Common Error dilakukan dengan menggunakan Breusch-Pagan test (Hill, Griffiths, and Lim, 2017). Pada hipotesis null tidak adanya efek random, dihitung statistik uji (kuadrat Lagrange Multiplier) sebagai berikut

$$LM^2 = \frac{NT}{2(T-1)}\left\{\frac{\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\hat{e}_{it}\right)^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{e}_{it}^2} - 1\right\}^2 \qquad (14.14)$$

Statistik ini memiliki distribusi χ^2 dengan derajat kebebasan sama dengan satu.

UJI HAUSMAN

Pemilihan antara FEM atau REM didasarkan pada apakah heterogenitas bersifat konstan (dan berkorelasi dengan variabel independen) atau random. Namun demikian, dalam praktek hal ini sulit ditentukan secara apriori. Karena itu, diperlukan suatu test untuk menguji superioritas suatu model terhadap model lain.

Hausman (1978) mengajukan suatu test yang menggunakan REM sebagai acuan (*hipotesis null*). Dasar pemikiran yang digunakan adalah menguji adanya hubungan antara a_i dan x_{itj}. Jika terdapat korelasi di antara keduanya, maka penggunaan REM akan bersifat bias dan tidak konsisten. Dalam situasi ini, lebih baik menggunakan FEM. Sebaliknya, jika tidak ada korelasi, baik FEM maupun REM

[3] Formulasi ini merupakan suatu ilustrasi. Software analisis statistik (termasuk STATA) umumnya menggunakan variasi algoritma dari yang diberikan oleh Wooldridge atau buku teks ekonometrika lainnya.

bersifat tidak bias; tetapi FEM lebih tidak efisien (karena menyerap *degree of freedom* lebih banyak). Jika hal ini terjadi, maka penggunaan REM akan lebih baik.

Prosedur Hausman adalah suatu pengujian atas perbedaan sistematik dari dua estimator. Hipotesis null pada pengujian Hausman adalah tidak ada perbedaan yang sistematik. Dengan demikian, jika statistik uji menunjukkan penolakan hipotesis null, maka FEM adalah lebih tepat dan sebaliknya REM jika hipotesis null tidak dapat ditolak.

$$t = \frac{b_{FEM,k} - b_{REM,k}}{\left[se\left(b_{FEM,k}\right)^2 - se\left(b_{REM,k}\right)^2\right]^{1/2}} \qquad (14.15)$$

di mana $b_{FEM,k}$ dan $b_{REM,k}$ adalah vektor koefisien hasil estimasi FEM dan REM; serta $se(b_{FEM,k})^2$ dan $se(b_{REM,k})^2$ adalah vektor estimasi varians dari koefisien tersebut. Statistik uji diberikan pada Persamaan 14.15, yang sebenarnya merupakan varian dari uji restriksi koefisien (Wald Test). Namun demikian, berbagai hal teknis yang timbul akibat penggunaan selisih standar error dari parameter yang diestimasi memiliki implikasi distribusi dari statistik ini berupa $\chi 2(Ks)$; di mana *Ks* adalah koefisien dari variabel yang memiliki variasi *cross section* dan deret waktu (lihat Wooldridge 2010, halaman 328-334 untuk pembahasan lebih detail).

Contoh 14.1

Dalam contoh ini kita akan menggunakan sebagian data dari studi Ariefianto et al (2020); yang terdapat pada file cost intermediation. dta. File ini berisi dataset panel beberapa indikator sistem perbankan terpilih dari 212 yurisdiksi (ekuivalen) negara (*cross section unit*) di

dunia, yaitu frekuensi: tahunan dari periode 1996 sampai dengan 2016 (*time series unit*). Kita ingin melihat pola hubungan antara *spread* (selisih antara suku bunga pinjaman dan simpanan; variabel dependen) dan variabel penjelas: permodalan bank (CAR), likuiditas (LDR), profitabilitas (ROE), volatilitas sistem keuangan (stockvlty, dan PDB per kapital (gdppercap). Sebelum melakukan regresi kita harus terlebih dahulu mendeklarasikan bahwa data memiliki struktur panel. Hal ini dilakukan dengan perintah **xtset crossid t**.

Source	SS	df	MS	Number of obs	=	4,452
				F(5, 4446)	=	70.08
Model	16601.2649	5	3320.25298	Prob > F	=	0.0000
Residual	210654.19	4,446	47.3806096	R-squared	=	0.0731
				Adj R-squared	=	0.0720
Total	227255.455	4,451	51.0571681	Root MSE	=	6.8834

spread	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
car	.0263598	.0123035	2.14	0.032	.0022389	.0504808
ldr	.0091777	.0011914	7.70	0.000	.0068419	.0115135
roe	.0257314	.0061056	4.21	0.000	.0137614	.0377014
stockvlty	-.0144358	.0093325	-1.55	0.122	-.0327322	.0038606
gdppercap	-.0916033	.0058936	-15.54	0.000	-.1031576	-.080049
_cons	5.285882	.178952	29.54	0.000	4.935047	5.636717

TABEL 14.1. Estimasi Pooled OLS

Hasil regresi pooled OLS (yang dilakukan dengan perintah **reg spread car ldr roe stockvlty gdppercap**) disajikan pada Tabel 14.1. Dapat dilihat di sini bahwa kecuali variabel volatilitas sistem keuangan; variabel-variabel penjelas lainnya bersifat signifikan secara statistik. Sebagai pembanding, selanjutnya kita akan melakukan estimasi dengan menggunakan FEM, dengan perintah STATA: **xtreg spread car ldr roe stockvlty gdppercap, fe**.

```
Fixed-effects (within) regression               Number of obs      =       4,452
Group variable: crossid                         Number of groups   =         212

R-sq:                                           Obs per group:
     within  = 0.0332                                         min =          21
     between = 0.0036                                         avg =        21.0
     overall = 0.0115                                         max =          21

                                                F(5,4235)          =       29.11
corr(u_i, Xb)  = -0.1013                        Prob > F           =      0.0000
```

spread	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
car	-.1106892	.0106567	-10.39	0.000	-.1315819	-.0897965
ldr	.0026999	.0010794	2.50	0.012	.0005837	.0048161
roe	-.0071287	.0047333	-1.51	0.132	-.0164084	.002151
stockvlty	-.0279157	.0098185	-2.84	0.004	-.0471651	-.0086662
gdppercap	-.0262821	.0126697	-2.07	0.038	-.0511213	-.0014429
_cons	6.521109	.2155131	30.26	0.000	6.09859	6.943628
sigma_u	5.6151526					
sigma_e	4.5159979					
rho	.60723041	(fraction of variance due to u_i)				

```
F test that all u_i=0: F(211, 4235) = 28.88                     Prob > F = 0.0000
```

TABEL 14.2. Estimasi Fixed Effect Model

Hasil regresi FEM disajikan pada Tabel 14.2. Secara default perintah estimasi FEM akan melampirkan hasil pengujian Wald terhadap Fixed Effect dari cross section unit (u_i) yang dapat dilihat pada bagian paling bawah. Di sini akan diperoleh nilai statistik hitung sebesar 28,88 dengan *p* value sebesar 0,000. Dengan demikian, terdapat *fixed effect* dari *composite residual*; di mana model FEM lebih baik daripada model pooled OLS. Estimasi yang terakhir ini bersifat bias.

Selanjutnya, kita melakukan estimasi dengan menggunakan model REM. Perintah STATA yang diperlukan adalah **xtreg spread car ldr roe stockvlty gdppercap, re**. Hasil dari regresi disajikan

```
Random-effects GLS regression                   Number of obs      =       4,452
Group variable: crossid                         Number of groups   =         212

R-sq:                                           Obs per group:
     within  = 0.0325                                         min =          21
     between = 0.0216                                         avg =        21.0
     overall = 0.0248                                         max =          21

                                                Wald chi2(5)       =      145.18
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =      0.0000
```

spread	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
car	-.0992274	.0105258	-9.43	0.000	-.1198575	-.0785973
ldr	.0033638	.001064	3.16	0.002	.0012784	.0054491
roe	-.0047531	.0047159	-1.01	0.313	-.013996	.0044898
stockvlty	-.0254368	.0095294	-2.67	0.008	-.0441141	-.0067595
gdppercap	-.043856	.0107817	-4.07	0.000	-.0649877	-.0227242
_cons	6.548554	.3994748	16.39	0.000	5.765598	7.33151
sigma_u	5.0437803					
sigma_e	4.5159979					
rho	.55504088	(fraction of variance due to u_i)				

TABEL 14.3. Estimasi Random Effect Model

pada Tabel 14.3. Pengujian apakah ada efek random dalam composite error (Breusch-Pagan test) langsung dilakukan dengan perintah **test xttest0**, setelah estimasi REM. Hasil pengujian atas efek random disajikan pada Tabel 14.4. Statistik hitung kuadratik LM sebesar 13170,21 dengan *p* value sebesar 0,000. Hasil pengujian adalah menolak hipotesis null tidak ada efek random. Dengan demikian, pemodelan dengan spesifikasi REM lebih baik dibandingkan pooled OLS.

Hasil estimasi dan pengujian yang telah dilakukan menunjukkan bahwa baik FEM maupun REM lebih superior dibandingkan pooled OLS. Jadi, isu terakhir yang harus diklarifikasi adalah apakah parameter yang diperoleh melalui REM tidak berkorelasi dengan residual. Apabila situasi ini terpenuhi, maka baik FEM maupun REM

```
Breusch and Pagan Lagrangian multiplier test for random effects

        spread[crossid,t] = Xb + u[crossid] + e[crossid,t]

        Estimated results:
                         |       Var      sd = sqrt(Var)
                ---------+------------------------------
                  spread |    51.05717         7.14543
                       e |    20.39424        4.515998
                       u |    25.43972         5.04378

        Test:   Var(u) = 0
                                 chibar2(01) = 13170.21
                              Prob > chibar2 =   0.0000
```

TABEL 14.4. Breusch Pagan Test REM

adalah estimator yang sama-sama tidak bias, sehingga tentu saja memiliki nilai yang hampir serupa. Perbedaan output estimasi di antara kedua estimator tersebut lebih disebabkan karena masalah variasi sampling.

Pengujian Hausman dilakukan dengan melihat apakah koefisien yang diestimasi melalui FEM dan REM memiliki perbedaan yang sistematis. Karena itu, kita harus menyimpan setiap parameter dari FEM dan REM untuk diperbandingkan. Hal ini dilakukan dengan perintah **estimates store (nama)**. Dalam contoh ini, kita akan menggunakan ***fe*** untuk menyimpan hasil estimasi FEM dan ***re*** untuk hasil estimasi REM. Selanjutnya, perintah Hausman dilakukan dengan membandingkan estimator yang dianggap "tidak bias dan selalu konsisten" dengan "yang lebih efisien, jika sama-sama tidak bias dan konsisten" yakni **hausman fe re**.

Hasil pengujian disajikan pada Tabel 14.5. Dapat dilihat di sini bahwa uji Hausman bekerja dengan membandingkan setiap parameter dan variansnya. Di sini hipotesis null: tidak adanya perbedaan yang sistematis di antara kedua estimator dapat ditolak

	Coefficients (b) fe	(B) re	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
car	-.1106892	-.0992274	-.0114618	.0016654
ldr	.0026999	.0033638	-.0006638	.0001819
roe	-.0071287	-.0047531	-.0023756	.0004058
stockvlty	-.0279157	-.0254368	-.0024789	.0023651
gdppercap	-.0262821	-.043856	.0175738	.006654

b = consistent under Ho and Ha; obtained from xtreg
B = inconsistent under Ha, efficient under Ho; obtained from xtreg

Test: Ho: difference in coefficients not systematic

chi2(5) = (b-B)'[(V_b-V_B)^(-1)](b-B)
= 54.75
Prob>chi2 = 0.0000

TABEL 14.5. Hausman Test FEM vs REM

karena statistik hitung yang sebesar 54,75 memiliki *p* value sebesar 0,000. Dengan kata lain, terdapat perbedaan yang sistematis di antara hasil estimasi FEM dan REM. Mengingat FEM adalah estimator yang tidak bias dan konsisten; maka hasil Hausman menunjukkan bahwa spesifikasi FEM lebih baik digunakan dibandingkan REM.

METODE INSTRUMENTAL VARIABEL

Seperti halnya struktur tunggal (*cross section* atau *time series*); data panel juga sangat mungkin mengalami endogenitas. Endogenitas adalah fitur model (diturunkan-implikasi dari teori) sekaligus data (akibat *omitted variables*, kesalahan spesifikasi, serta pengukuran). Dengan demikian, jika terdeteksi ada endogenitas, maka harus ditangani dengan metode khusus yang salah satunya adalah

instrumental variabel. Misalkan kita memiliki persamaan linear sebagai berikut

$$y_{it} = \alpha + Z\delta + u_i \tag{14.16}$$

$$u_i = Z_\mu \mu + v_{it} \tag{14.17}$$

di mana Z = [Y, X] adalah vektor gabungan dari vektor variabel endogen (Y) dan vektor variabel eksogen (X). Selanjutnya asumsikan bahwa vektor variabel eksogen (X) terdiri dari variabel eksogen yang berada dalam model (berada dalam Z) yang disebut sebagai X_1 dan variabel eksogen yang berada di luar model (disebut X_2). Variabel-variabel pada X_2 adalah kandidat instrumen bagi Y. Persamaan 14.17 menunjukkan bahwa berbeda dengan model panel sebelumnya, di sini residual adalah fungsi linear dari Z. Tanpa penanganan khusus, estimasi OLS (atau FE dan RE standar) pada Persamaan 14.16 akan menghasilkan estimator yang bias dan tidak konsisten.

Derivasi dari koreksi atas adanya endogenitas pada data panel cukup kompleks dan berada di luar cakupan buku ini. Terdapat dua pilihan yang umum digunakan yakni *Error Component Two Stage Least Squares* (EC2SLS, Baltagi 1981).

$$\delta_{EC2SLS} = \left[\frac{\tilde{Z}' P_{\bar{X}} \tilde{Z}}{\sigma^2_{v_{11}}} + \frac{\tilde{Z}' P_{\bar{X}} \tilde{Z}}{\sigma^2_{1_{11}}}\right]^{-1} \left[\frac{\tilde{Z}' P_{\bar{X}} \tilde{y}}{\sigma^2_{v_{11}}} + \frac{\tilde{Z}' P_{\bar{X}} \tilde{y}}{\sigma^2_{1_{11}}}\right] \tag{14.18}$$

dan *Generalized Two Stage Least Squares* (Balestra dan Varadharajan-Khrisnakumar, 1987)

$$\delta_{G2SLS} = (Z^{*\prime} P_{x*} Z^{*})^{-1} (Z^{*\prime} P_{x*} y^{*}) \qquad (14.19)$$

Kedua estimator tersebut dapat diterapkan pada model asumsi komponen residual yang bersifat fixed effect maupun random effect. Selanjutnya, Arrelano (1987) dan Wooldrige (2010) mengusulkan modifikasi pada kedua estimator tersebut dengan menggunakan teknik estimasi serentak, yakni Maximum Likelihood dan *General Method of Moment* = GMM (Hansen 1982).

Contoh 14.2

Di sini akan digunakan data dari studi Layard and Nickell (1986); yang juga digunakan dalam studi Arrelano dan Bond (1991). Kita simpan data ini sebagai LN_1986.dta. Dataset terdiri dari variabel gaji (*w*), modal saham (*k*), output perusahaan (*Ys*), dan penyerapan tenaga kerja (*n*). Struktur dataset adalah *unbalanced* yang terdiri dari cross section unit: 140 perusahaan di Inggris dan time unit: tahunan, 1976-1984. Data ini dapat didownload dari Internet dengan menggunakan perintah **webuse abdata**.

Kita akan mengestimasi regresi *Ys* terhadap *k* dan *n* (fungsi produksi linear); di mana *n* adalah endogen serta lag 2 dan 3 dari *n* digunakan sebagai instrumen dengan model FE. Dapat dilihat dari help file bahwa estimasi instrumental variabel pada STATA menggunakan Balestra dan Varadharajan-Khrisnakumar (1987) sebagai default. Perintah untuk estimator default (G2SLS) diberikan dengan **xtivreg ys k (n=l2.n l3.n),** sedangkan untuk EC2SLS adalah **xtivreg ys k (n=l2.n l3.n), ec2sls**.

Hasil estimasi disajikan pada Tabel 14.6 (untuk G2SLS) dan 14.7 (untuk EC2SLS). Dalam kondisi default, kedua estimator tersebut

```
G2SLS random-effects IV regression              Number of obs      =       611
Group variable: id                              Number of groups   =       140

R-sq:                                           Obs per group:
     within  = 0.0050                                         min =         4
     between = 0.0385                                         avg =       4.4
     overall = 0.0097                                         max =         6

                                                Wald chi2(2)       =     19.42
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0001
```

ys	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
n	-.0211639	.0058228	-3.63	0.000	-.0325764	-.0097515
k	.0216127	.0050101	4.31	0.000	.011793	.0314324
_cons	4.625355	.0085937	538.23	0.000	4.608512	4.642198
sigma_u	0					
sigma_e	.06215015					
rho	0	(fraction of variance due to u_i)				

```
Instrumented:   n
Instruments:    k L2.n L3.n
```

TABEL 14.6. Hasil Estimasi Instrumental Variabel Data Panel: G2SLS Estimator

```
EC2SLS random-effects IV regression             Number of obs      =       611
Group variable: id                              Number of groups   =       140

R-sq:                                           Obs per group:
     within  = 0.0393                                         min =         4
     between = 0.0310                                         avg =       4.4
     overall = 0.0123                                         max =         6

                                                Wald chi2(2)       =     10.40
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0055
```

ys	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
n	-.0114523	.0056334	-2.03	0.042	-.0224936	-.000411
k	.0139012	.0048679	2.86	0.004	.0043603	.0234422
_cons	4.612007	.008346	552.60	0.000	4.595649	4.628365
sigma_u	0					
sigma_e	.06215015					
rho	0	(fraction of variance due to u_i)				

```
Instrumented:   n
Instruments:    k L2.n L3.n
```

TABEL 14.7. Hasil Estimasi Instrumental Variabel Data Panel: EC2SLS Estimator

adalah RE estimator. Terdapat perbedaan yang cukup signifikan pada koefisien variabel eksogen (*k*) dan endogen (*n*); di mana hasil dari EC2SLS secara absolut adalah 2 kali lipat dari G2SLS.

Evaluasi terhadap variabel endogen dan instrumen menurut perintah xtivreg masih harus dilakukan secara adhoc. Schaffer (2010) telah membuat routine pelengkap (*wrapper*) terhadap perintah xtivreg dengan nama **xtivreg2**. Ada pun kelengkapan tambahan yang dimasukkan adalah (a) estimasi dengan menggunakan maximum likelihood (LIML dan *k* class), general methods of moments, dan pengujian (*underidentification*, *weak instruments*, dan *overidentification*). Ringkasnya, program ini adalah gabungan (*wrapper*) dari **xtivreg** dan **ivreg2** (lihat Bab 10 di Jilid 1).

Kita akan mengulang estimasi spesifikasi sebelumnya (Tabel 14.6 dan 14.7); tetapi sekarang dengan menggunakan model FE. Hasil estimasinya disajikan pada Tabel 14.8; di sini terlihat bahwa kedua perintah tersebut memberikan hasil yang sama. Namun demikian, xtivreg2 memberikan tambahan output berupa underidentification test, weak instrument test, dan overidentification test. Underidentification test menghasilkan angka LM statistik sebesar 30,649; dengan *p* value: 0,000 sehingga hipotesis null model tidak teridentifikasi dapat ditolak. Statistik Cragg-Donald Wald F berada di bawah nilai kritis sebesar 10%, sehingga instrumen yang digunakan dapat dikatakan "masih lemah". Pengujian *overidentification* juga menolak hipotesis null validitas instrumen; yaitu Sargan statistik sebesar 6,089 memiliki *p* value sebesar 0,0136.

MODEL DATA PANEL DINAMIS

Salah satu model yang popular dalam menangani struktur short-panel ($N < T$) adalah data panel dinamis: *Dynamic Panel Data* =

```
Fixed-effects (within) IV regression            Number of obs      =       611
Group variable: id                              Number of groups   =       140

R-sq:                                           Obs per group:
     within  = 0.3569                                         min =         4
     between = 0.0020                                         avg =       4.4
     overall = 0.0082                                         max =         6

                                                F(142,469)         =    117.48
corr(u_i, Xb)  = -0.9897                        Prob > F           =    0.0000

------------------------------------------------------------------------------
          ys |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           n |    .105914   .0865132     1.22   0.221    -.0640873    .2759154
           k |   .1352526   .0638731     2.12   0.035     .0097397    .2607655
       _cons |   4.558421   .1159595    39.31   0.000     4.330556    4.786285
-------------+----------------------------------------------------------------
     sigma_u |  .34516869
     sigma_e |  .06215015
         rho |  .96859745   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F  test that all u_i=0:     F(139,469) =     2.48                 Prob > F    = 0.0000
------------------------------------------------------------------------------
Instrumented:   n
Instruments:    k L2.n L3.n
------------------------------------------------------------------------------

FIXED EFFECTS ESTIMATION
------------------------

Number of groups =       140                    Obs per group: min =         4
                                                               avg =       4.4
                                                               max =         6

IV (2SLS) estimation
--------------------

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

                                                Number of obs =       611
                                                F(  2,   469) =    117.48
                                                Prob > F      =    0.0000
Total (centered) SS     =  2.817151731          Centered R2   =    0.3569
Total (uncentered) SS   =  2.817151731          Uncentered R2 =    0.3569
Residual SS             =  1.811578915          Root MSE      =    .06215

------------------------------------------------------------------------------
          ys |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           n |    .105914   .0865132     1.22   0.221    -.0640873    .2759154
           k |   .1352526   .0638731     2.12   0.035     .0097397    .2607655
------------------------------------------------------------------------------
Underidentification test (Anderson canon. corr. LM statistic):          30.649
                                                Chi-sq(2) P-val =       0.0000
------------------------------------------------------------------------------
Weak identification test (Cragg-Donald Wald F statistic):               16.287
Stock-Yogo weak ID test critical values: 10% maximal IV size            19.93
                                         15% maximal IV size            11.59
                                         20% maximal IV size             8.75
                                         25% maximal IV size             7.25
Source: Stock-Yogo (2005).  Reproduced by permission.
------------------------------------------------------------------------------
Sargan statistic (overidentification test of all instruments):           6.089
                                                Chi-sq(1) P-val =       0.0136
------------------------------------------------------------------------------
Instrumented:         n
Included instruments: k
Excluded instruments: L2.n L3.n
------------------------------------------------------------------------------
```

TABEL 14.8. Hasil Estimasi xtivreg Dibandingkan xtivreg2

DPD (Roodman, 2009). Di samping *short panel*; model DPD juga digunakan untuk spesifikasi regresi linear di mana (a) terdapat karakter persistensi yang diakibatkan oleh variabel dependen, (b) kemungkinan adanya endogenitas pada satu set variabel penjelas, (c) fixed effect pada cross section, dan (d) adanya autokorelasi serta heterokedastisitas pada *intra cross section unit*. Dengan kata lain, model DPD adalah suatu model ekonometrika yang cukup *generalized*. Mengingat banyak sekali variabel ekonomi yang bersifat *autoregressive* (Nelson dan Plosser, 1982), maka model DPD sering digunakan sebagai alternatif pengembangan desain estimasi bagi metode instrumental variabel standar yang telah dibahas sebelumnya.

Misalkan kita memiliki regresi linear dengan vektor regressor X serta persistensi pada variabel dependen (y) di lag 1, sebagai berikut

$$y_{it} = \alpha + \delta y_{i,t-1} + X_{it}\beta + u_i + v_{it} \tag{14.20}$$

Dengan mengambil first difference terhadap Persamaan 14.20 kita memperoleh

$$\Delta y_{it} = \rho \Delta y_{i,t-1} + X_{it}\gamma + \Delta e_{it} \tag{14.21}$$

Perhatikan bahwa akibat dari *first differencing*, koefisien menjadi konstan dan individual effect menjadi hilang. Namun demikian, dampak sampingan dari transformasi ini adalah adanya korelasi antara *regressor* (lag pertama *y* bentuk *first difference*) dan residual (Nickell, 1981). Permasalahan ini dapat ditangani dengan mengambil instrumen (lag ke-2, 3, atau lebih) dari difference variabel dependen.

Cara estimasi seperti ini diusulkan oleh Anderson dan Hsiao (1982). Arellano dan Bond (1991) mengembangkan estimator ini lebih lanjut dengan memasukkan persyaratan kondisi orthogonality antara regressor dan residual (*General Method of Moment* = GMM; Hansen 1982). Di sini Arrelano dan Bond (1991) melakukan generalisasi terhadap Persamaan 14.20 dan 14.21 menjadi

$$y_{it} = X_{it}\beta + Z_{it}\gamma + u_i + e_{it} \quad (14.22)$$

di mana *X* sekarang adalah vektor *strictly exogenous regressor* dan vektor *Z* disebut *predetermined regressor* yang tidak hanya berisi (lag 2, 3, dan seterusnya) *first difference* dari variabel dependen, tetapi juga bentuk lag difference dari variabel endogen. Vektor *Z* adalah vektor instrumental variabel dari persistensi variabel dependen dan variabel endogen lainnya. Baum (2013) menyebutnya sebagai vektor internal instrumen karena instrumen ini diperoleh dari variabel-variabel yang sudah ada dalam model. Untuk mengurangi dampak hilangnya *degree of freedom* karena instrumentasi, digunakan konstruksi matriks yang disarankan oleh Holtz-Eakin, Newey, dan Rosen (1988). Estimator yang telah diuraikan sebelumnya dikenal juga sebagai *Difference* GMM; D-GMM.

Estimator Arellano and Bond (1991) memiliki kelemahan yakni kinerja instrumentasi yang rendah terutama jika variabel-variabel dalam vektor *Z* berada pada level (Arrelano dan Bover, 1995 dan Blundell dan Bond, 1998). Karena itu, mereka mengusulkan modifikasi dengan menggunakan variabel instrumen *lagged level* dan *lagged differences*. Estimator Arrelano-Bond yang memiliki matriks variabel instrumen dalam bentuk seperti ini dikenal dengan nama *System GMM*; S-GMM.

Estimator-estimator yang telah diuraikan sebelumnya merupakan metode instrumental. Jadi, setelah melakukan estimasi harus dievaluasi (a) reliabilitas instrumen dalam arti memiliki korelasi yang erat dengan variabel endogen dan (b) validitas instrumen dalam arti tidak memiliki korelasi dengan residual. Terutama pada model DPD, juga harus dilakukan evaluasi tambahan berupa persyaratan stabilitas dan uji autokorelasi. Persyaratan stabilitas akan dipenuhi jika koefisien lag dari variabel dependen memiliki nilai absolut di bawah satu. Berdasarkan konstruksi Persamaan 14.17, residual dari persamaan *first difference* akan memiliki autokorelasi derajat pertama (AR1). Namun demikian, jika asumsi residual persamaan level: yaitu Persamaan 14.16 tidak memiliki korelasi serial, maka persamaan *first difference* tidak boleh menunjukkan karakter serial korelasi derajat dua (AR2). Jika terdeteksi ada AR(2), maka lag ke-2 dari variabel endogen tidak dapat digunakan sebagai variabel instrumen atas nilai saat ini (Baum, 2013 hal. 20). Roodman(2009) menambahkan untuk juga memperhatikan jumlah instrumen yang digunakan[4].

Contoh 14.3

Kita masih menggunakan LN_1986 dataset. Dengan mereplikasi Baum (2013), hal. 29, kita akan melakukan estimasi terhadap model dengan variabel dependen *n* yang diregresi atas pilihan regressor sebagai berikut: *n* (lag satu atau dua), *w* (lag nol atau satu), *k* (lag satu sampai dengan dua), *ys* (lag satu sampai dengan dua), dan dummy tahun-year effect. Kita akan menggunakan routine yang dikembangkan oleh Roodman (2009) yang bernama

[4] Routine xtabond2 akan melaporkan *warning message* apabila jumlah instrumen tidak proporsional dengan jumlah observasi; terdapat *overfitting model*.

```
Favoring speed over space. To switch, type or click on mata: mata set matafavor space, perm.
Warning: Two-step estimated covariance matrix of moments is singular.
  Using a generalized inverse to calculate robust weighting matrix for Hansen test.
  Difference-in-Sargan/Hansen statistics may be negative.

Dynamic panel-data estimation, one-step difference GMM
------------------------------------------------------------------------------
Group variable: id                              Number of obs      =       611
Time variable : year                            Number of groups   =       140
Number of instruments = 41                      Obs per group: min =         4
F(19, 140)     =     87.61                                     avg =      4.36
Prob > F       =     0.000                                     max =         6
------------------------------------------------------------------------------
                             Robust
           n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           n |
         L1. |   .6862261    .147303     4.66   0.000     .3950001    .9774522
         L2. |  -.0853582   .0570649    -1.50   0.137    -.1981786    .0274621
             |
           w |
         --. |  -.6078208   .1815439    -3.35   0.001    -.9667429   -.2488987
         L1. |   .3926237   .1711403     2.29   0.023     .0542702    .7309772
             |
           k |
         --. |   .3568456    .060126     5.93   0.000     .2379734    .4757179
         L1. |  -.0580012   .0745506    -0.78   0.438    -.2053918    .0893895
         L2. |  -.0199475   .0333255    -0.60   0.550    -.0858337    .0459388
             |
          ys |
         --. |   .6085073   .1757634     3.46   0.001     .2610136     .956001
         L1. |  -.7111651   .2360572    -3.01   0.003    -1.177863   -.2444673
         L2. |   .1057969   .1438474     0.74   0.463    -.1785971    .3901909
             |
      yr1976 |           0  (omitted)
      yr1977 |           0  (omitted)
      yr1978 |           0  (omitted)
      yr1979 |   .0095545   .0104824     0.91   0.364    -.0111697    .0302787
      yr1980 |   .0220152   .0180422     1.22   0.224    -.0136552    .0576856
      yr1981 |  -.0117743   .0300607    -0.39   0.696    -.0712058    .0476572
      yr1982 |  -.0270588   .0298235    -0.91   0.366    -.0860215    .0319039
      yr1983 |  -.0213204   .0310305    -0.69   0.493    -.0826694    .0400285
      yr1984 |  -.0077033   .0319991    -0.24   0.810    -.0709672    .0555606
------------------------------------------------------------------------------
Instruments for first differences equation
  Standard
    D.(w L.w k L.k L2.k ys L.ys L2.ys yr1976 yr1977 yr1978 yr1979 yr1980
    yr1981 yr1982 yr1983 yr1984)
  GMM-type (missing=0, separate instruments for each period unless collapsed)
    L(1/8).L.n
------------------------------------------------------------------------------
Arellano-Bond test for AR(1) in first differences: z =  -3.60  Pr > z =  0.000
Arellano-Bond test for AR(2) in first differences: z =  -0.52  Pr > z =  0.606
------------------------------------------------------------------------------
Sargan test of overid. restrictions: chi2(22)   =  67.59  Prob > chi2 =  0.000
  (Not robust, but not weakened by many instruments.)
Hansen test of overid. restrictions: chi2(22)   =  31.38  Prob > chi2 =  0.089
  (Robust, but weakened by many instruments.)

Difference-in-Hansen tests of exogeneity of instrument subsets:
  iv(w L.w k L.k L2.k ys L.ys L2.ys yr1976 yr1977 yr1978 yr1979 yr1980 yr1981 yr1982 yr1983 yr198
> 4)
    Hansen test excluding group:     chi2(8)    =  12.01  Prob > chi2 =  0.151
    Difference (null H = exogenous): chi2(14)   =  19.37  Prob > chi2 =  0.151
```


TABEL 14.9. Hasil Estimasi dengan Menggunakan Teknik D-GMM; xtabond2

xtabond2. Karena ini bukan merupakan routine default STATA, maka pembaca dapat menginstallkannya terlebih dahulu dengan **ssc install xtabond2**.

Pertama kita akan mengestimasi dengan menggunakan teknik Arellano-Bond (1991); D-GMM. Perintah yang dijalankan adalah **xtabond2 n L(1/2).n L(0/1).w L(0/2).(k ys) yr*, gmm(L.n) iv(L(0/1).w L(0/2).(k ys) yr*) nolevel robust small**. Terdapat dua jenis instrumen yakni GMM dan IV; yang diberikan setelah tanda koma. Instrumen GMM digunakan untuk variabel yang dianggap endogen: lag variabel dependen dan variabel lain. Instrumen IV digunakan untuk variabel selain instrumen GMM. Dapat diperhatikan di sini bahwa teknik D-GMM menggunakan semua variabel independen di luar lag variabel dependen sebagai instrumen. Dengan xtabond2; D-GMM dikenali dari keberadaan term **nolevel**. Term robust dan small; bersifat opsional untuk melakukan koreksi terhadap keberadaan heterokedastisitas dan *small sample*.

Dari hasil estimasi (Tabel 14.9) terlihat bahwa model menggunakan 41 instrumen. Koefisien lag (satu dan dua) variabel dependen memiliki nilai absolut di bawah satu (0,686 dan –0,085); sehingga model itu memenuhi persyaratan stabilitas. Nilai statistik Hansen (overidentification test) yang sebesar 31,38; dengan *p* value sebesar 0,089 menunjukkan masih adanya isu endogenitas pada spesifikasi ini. Pengujian autokorelasi menunjukkan bahwa spesifikasi tidak memiliki order AR(2) yang signifikan sesuai dengan yang dipersyaratkan. Pembaca dapat melakukan eksperimen dengan mengubah parameter lag untuk mendapatkan spesifikasi terbaik.

Kita akan memberikan ilustrasi tentang estimasi S-GMM dengan melakukan modifikasi pada contoh sebelumnya. Di sini *w* dan *k* dianggap sebagai endogen, dan hanya menggunakan lag satu

```
Favoring speed over space. To switch, type or click on mata: mata set matafavor space, perm.
Warning: Two-step estimated covariance matrix of moments is singular.
  Using a generalized inverse to calculate robust weighting matrix for Hansen test.
  Difference-in-Sargan/Hansen statistics may be negative.

Dynamic panel-data estimation, one-step system GMM
------------------------------------------------------------------------------
Group variable: id                              Number of obs      =       891
Time variable : year                            Number of groups   =       140
Number of instruments = 113                     Obs per group: min =         6
F(14, 139)    =    987.19                                      avg =      6.36
Prob > F      =     0.000                                      max =         8
------------------------------------------------------------------------------
                             Robust
           n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           n |
         L1. |   .9356053   .0265995    35.17   0.000     .8830134    .9881973
             |
           w |
         --. |  -.6309761   .1194204    -5.28   0.000    -.8670915   -.3948607
         L1. |   .4826203   .1384721     3.49   0.001     .2088363    .7564042
             |
           k |
         --. |   .4839299   .0544906     8.88   0.000     .3761923    .5916676
         L1. |  -.4243928   .0591559    -7.17   0.000    -.5413546   -.3074311
             |
      yr1976 |          0  (omitted)
      yr1977 |          0  (omitted)
      yr1978 |    .006405    .01942     0.33   0.742    -.0319918    .0448018
      yr1979 |   .0214058   .0223268     0.96   0.339    -.0227383    .0655499
      yr1980 |   .0066578   .0221916     0.30   0.765    -.037219     .0505346
      yr1981 |  -.019471    .0275339    -0.71   0.481    -.0739105    .0349685
      yr1982 |   .0144379   .0277471     0.52   0.604    -.040423     .0692989
      yr1983 |   .0278705   .0263394     1.06   0.292    -.0242071    .0799481
      yr1984 |   .0240573   .0297311     0.81   0.420    -.0347263    .0828409
       _cons |   .5281438   .2042459     2.59   0.011     .1243133    .9319742
------------------------------------------------------------------------------
Instruments for first differences equation
  GMM-type (missing=0, separate instruments for each period unless collapsed)
    L(1/8).(L.n L.w L.k)
Instruments for levels equation
  Standard
    yr1976 yr1977 yr1978 yr1979 yr1980 yr1981 yr1982 yr1983 yr1984
    _cons
  GMM-type (missing=0, separate instruments for each period unless collapsed)
    D.(L.n L.w L.k)
------------------------------------------------------------------------------
Arellano-Bond test for AR(1) in first differences: z =  -5.46  Pr > z =  0.000
Arellano-Bond test for AR(2) in first differences: z =  -0.25  Pr > z =  0.804
------------------------------------------------------------------------------
Sargan test of overid. restrictions: chi2(98)   = 157.54  Prob > chi2 =  0.000
  (Not robust, but not weakened by many instruments.)
Hansen test of overid. restrictions: chi2(98)   = 110.70  Prob > chi2 =  0.179
  (Robust, but weakened by many instruments.)

Difference-in-Hansen tests of exogeneity of instrument subsets:
  GMM instruments for levels
    Hansen test excluding group:     chi2(77)   =  84.33  Prob > chi2 =  0.266
    Difference (null H = exogenous): chi2(21)   =  26.37  Prob > chi2 =  0.193
  iv(yr1976 yr1977 yr1978 yr1979 yr1980 yr1981 yr1982 yr1983 yr1984, eq(level))
    Hansen test excluding group:     chi2(91)   = 107.79  Prob > chi2 =  0.110
    Difference (null H = exogenous): chi2(7)    =   2.91  Prob > chi2 =  0.893
```

TABEL 14.10. Hasil Estimasi dengan Menggunakan Teknik S-GMM; xtabond2

pada variabel dependen dan lag nol sampai dengan satu untuk variabel *w* dan *k*. Perintah estimasi diberikan sebagai **xtabond2 n L.n L(0/1).(w k) yr*, gmm(L.(n w k)) iv(yr*, equation(level)) robust small**. Estimasi S-GMM ditunjukkan dengan keberadaan term **equation(level)**.

Seperti dapat dilihat pada Tabel 14.10, S-GMM menggunakan lebih banyak instrumen dibandingkan D-GMM: 113 instrumen. Koefisien lag (satu dan dua) variabel dependen memiliki nilai absolut di bawah satu (0,936); sehingga model memenuhi persyaratan stabilitas. Nilai statistik Hansen (*overidentification test*) sebesar 110,70 dengan *p* value sebesar 0,179 menunjukkan bahwa model telah dispesifikasikan dengan baik (dari aspek endogenitas). Pengujian autokorelasi menunjukkan bahwa spesifikasi tidak memiliki order AR(2) yang signifikan sesuai dengan yang dipersyaratkan.

MODEL LONG DATA PANEL

Apabila jumlah unsur deret waktu (*T*) cukup banyak, jauh lebih banyak dibandingkan unsur cross section (*N*), maka kita memiliki struktur data *long panel*. Struktur data *long panel* sering juga disebut sebagai *nonstationary panel, heterogenous panel*, atau *macro-panel*. Untuk struktur data semacam ini terdapat beberapa karakter yang perlu diperhatikan dan memerlukan perlakuan khusus (Eberhardt, 2011). Karakter tersebut adalah (a) heterogenitas parameter, (b) *nonstationary,* dan (c) *cross-section correlation*.

Apabila kita memiliki dataset dengan T yang banyak, Eberhardt (2011a) memberikan *rule of thumb* T > 20 dan N < 100, sehingga setiap cross section unit (*i*) dapat diperlakukan seolah-olah

memiliki regresinya sendiri (*heterogenous constant* dan *slope*). Dengan kata lain, estimasi koefisien (konstanta dan *slope regressor*) serta matriks varians-kovarians dari indeks *i* juga menjadi fokus studi. Seperti ditunjukkan oleh Pesaran dan Smith (1995) serta Im, Pesaran dan Shin (2003), estimator DPD (2SLS, D-GMM, dan S-GMM) mengasumsikan bahwa homogenitas slope/kemiringan parameter mungkin tidak terpenuhi. Penggunaan estimator DPD akan menghasilkan koefisien yang bias dan tidak konsisten.

Selanjutnya *T* yang cukup besar juga berpotensi menimbulkan isu *nonstationary* pada masing-masing variabel: dependen dan independen. Dengan demikian, tanpa perlakuan khusus estimasi terhadap hubungan di antara variabel-variabel tersebut berpotensi menghasilkan *spurious regression*. Karena itu, pengujian *unit root* dan kointegrasi pada variabel-variabel yang digunakan perlu dilakukan terlebih dahulu. Jika variabel-variabel tersebut bersifat nonstationary, maka spesifikasi yang tepat harus mengakomodasi *error correction model* untuk meyakini pemenuhan persyaratan stabilitas regresi.

Struktur panel heterogenitas juga memiliki isu kemungkinan adanya korelasi di antara unsur-unsur *cross section*. Hal ini terjadi akibat adanya variabel yang mempengaruhi pergerakan semua unsur cross section: *common factor* atau *spill-over effect*. *Common factor* tersebut mungkin tidak dapat dimasukkan secara eksplisit ke dalam regresi karena merupakan variabel laten (tidak terobservasi; *unobserved*). Di sisi lain, model-model regresi panel yang telah dibahas pada bagian sebelumnya mengasumsikan independensi dari unit *cross section*. Jika isu ini tidak ditangani dengan baik, maka estimator yang diperoleh menjadi kurang akurat (*imprecise*) dan bukan tidak mungkin regresi akan mengalami masalah identifikasi: yang sebenarnya tidak dapat diestimasi.

Eberhardt dan Teal (2011) menyatakan bahwa matriks pilihan teknik estimasi macro panel akan dipengaruhi oleh karakter (a) heterogenitas *cross section unit* dan (b) heterogenitas *common factor*. Suatu model regresi nonstationary panel dapat direpresentasikan sebagai berikut:

$$y_{it} = x_{it}\beta + u_{it} \; ; \tag{14.23}$$

$$u_{it} = \alpha_i + F_t\gamma_i + \varepsilon_{it} \tag{14.24}$$

$$x_{mit} = \pi_{mi} + G_{mt}\delta_{mi} + f_{nmt}\rho_{nmi} + v_{mit} \tag{14.25}$$

Persamaan 14.23 adalah regresi linear yang menjadi fokus studi. Persamaan 14.24 adalah dekomposisi residual yang sekarang tidak hanya dipengaruhi oleh *cross section fixed effect* (α_i), tetapi juga (kemungkinan) oleh keberadaan suatu vektor *common factor* yang akan mempengaruhi setiap cross section unit (dengan derajat yang berbeda-beda pada suatu titik waktu); term F_t. Persamaan 14.25 mengakomodasi keberadaan heterogenitas slope regressor (yang diberi indeks m = 1,.., k*).* Di sini terdapat komponen umum (π_{mi}), komponen spesifik regressor (δ_{mi}), dan komponen spesifik cross section unit (ρ_{nmi}). Dengan memperhatikan struktur Persamaan 14.23–14.25, pilihan estimasi disajikan pada Tabel 14.11. Parameter γ_i, δ_{mi}, dan ρ_{nmi} disebut *cross section specific factor loading*; sedangkan vektor F_t dan G_{mt} menunjukkan karakter *common factor*.

Cross Section Factor Loading		
Common Factor Character	**Homogenous**	**Heterogenous**
Homogenous	Pooled OLS, FE, RE	Common Correlated Effect Mean Pooled (CCEP)
Heterogenous	Mean Group (MG); Pooled Mean Group (PMG)	Augmented Mean Group (AMG), Common Correlated Effect Mean Group (CCEMG), Dynamic Common Correlated Effect (DCCE)

TABEL 14.11 Pilihan Teknik Estimasi Model Long Panel (diadaptasi dari Eberhardt dan Teal (2011))

Estimator-estimator yang berada di kuadran 2, 3, dan 4 itu bersifat nonlinear. Semuanya diestimasi oleh teknik Maximum Likelihood (Pesaran, Shin, dan Smith, 1999). Lebih lanjut, estimator-estimator ini juga dapat direpresentasikan untuk persamaan jangka panjang (*long term*) dan jangka pendek (*short term*), *error correction model* (ECM). Misalkan kita merepresentasikan suatu model long term panel sebagai bentuk panel ARDL(P,Q) berikut ini:

$$y_{it} = \sum_{p=1}^{P} \lambda_{ip} y_{i,t-p} + \sum_{q=1}^{Q} \eta_{iq} X_{i,t-q} + \mu_i + \epsilon_{it} \qquad (14.26)$$

Jika variabel-variabel dalam Persamaan 14.26 bersifat *nonstationary* (I(1)) dan terkointegrasi, maka komponen residual komposit bersifat

stasioner (I(0)) untuk semua *i*. Dengan demikian, Persamaan 14.26 dapat direparameterisasi menjadi

$$\Delta y_{it} = \phi\left(y_{i,t-1} - X_{i,t-1}\theta\right) + \sum_{p=1}^{P-1} \lambda_{ip}^{*} \Delta y_{i,t-p} + \sum_{q=1}^{Q-1} \eta_{iq}^{*} \Delta X_{i,t-q} + \mu_i + \epsilon_{it} \quad (14.27)$$

di mana ϕ adalah koefisien error correction term (ECT); θ adalah koefisien regresi jangka panjang, serta λ_{ip}^{*} dan η_{iq}^{*} adalah koefisien regresi jangka pendek. Seperti juga model time series, kita mengharapkan koefisien ϕ memiliki nilai negatif di bawah satu dan signifikan untuk menandai adanya stabilitas sekaligus proses penyesuaian deviasi terhadap ekuilibrium. Jika konstanta, slope/kemiringan, dan matriks varians-kovarians residual dapat diasumsikan berbeda-beda di antara cross section (*i*), maka estimator disebut sebagai *Mean Group* = MG (Pesaran dan Smith, 1995). Pada estimator MG, koefisien slope diperoleh sebagai rata-rata dari slope/kemiringan individual cross section. Jika koefisien-koefisien estimator tersebut diasumsikan (atau direstriksi) sebagai kombinasi (pooling atau averaging), maka estimator itu disebut *Pooled Mean Group* = PMG (Pesaran, Shin, dan Smith, 1999).

Tentu saja, sebelum melakukan estimasi kita harus melakukan analisis prasyarat berikut (a) uji unit root, (b) kointegrasi, dan (c) uji independensi cross section unit. Pengujian unit root pada panel data sebenarnya merupakan pengembangan langsung dari model time series. Levin, Lin, dan Chu (1992), Im, Pesaran, dan Shin-IPS (2003), serta Cross Sectionally Augmented IPS-CIPS: Pesaran (2007) melakukan uji unit root dengan menggunakan atas teknik *Augmented Dickey Fuller* yang dimodifikasi. Maddala dan Wu (1999),

Hadri (2000), dan Breitung and Das (2005) mengusulkan uji unit root dengan menggunakan beberapa alternatif.

Berbeda dengan uji unit root pada time series, kesimpulan jauh lebih sulit digeneralisir pada panel unit root. Kita hanya akan memperoleh inferensial yang berseberangan yakni (a) seluruh cross section unit dalam suatu variabel bersifat stasioner versus (b) paling tidak ada satu cross section unit yang tidak bersifat stasioner. Jika kita masuk ke dalam kategori (b), maka dalam praktek kita akan sering mengalami kesulitan untuk mengambil kesimpulan mengenai karakter nonstasioner dari variabel tersebut. Namun variabel tersebut masih dapat disebut stasioner meskipun ada beberapa cross section unit yang bersifat nonstasioner (sehingga inferensial menjadi ambigu). Pesaran (2012) merekomendasikan pengujian proporsi cross section unit yang bersifat *nonstationary* untuk membantu inferensial.

Di samping ambiguitas inferensial; Eberhardt (2011) juga merangkum beberapa kelemahan uji panel unit root yang ada saat ini. Kelemahan tersebut adalah

a. Low power pada kasus near unit root
b. Inferensial yang sensitif terhadap spesifikasi: yaitu autokorelasi dan penggunaan konstanta serta time trend.
c. Sensitivitas terhadap structural break.
d. Power test tergantung pada jumlah observasi deret waktu (*time span*).
e. Spesifikasi model adalah berkarakter nonlinear.

Seperti uji unit root, uji kointegrasi data panel juga merupakan pengembangan langsung dari model time series. Beberapa teknik pengujian yang cukup popular adalah Kao (1999) dan Pedroni (2004)

yang menggunakan kerangka kerja Engle Granger. Pengembangan lebih lanjut yang mengakomodasi adanya *cross section dependence* adalah Westerlund (2008) dan Gengenbach, serta Urbain dan Westerlund (2009). Heterogenitas pada cross section juga menimbulkan isu ambiguitas pada uji kointegrasi. Eberhardt (2011) mengusulkan kerangka kerja inferensial sekuensial sebagai berikut

$$\begin{aligned} y_{it} &= \alpha_i + \beta_i x_{it} + u_{it} \\ x_{it} &= \mu + x_{it-1} + v_{it} \\ u_{it} &= \rho_i u_{it-1} + \varepsilon_{it} \end{aligned} \tag{14.28}$$

Jika $\rho_i = 1\Lambda_i$, maka tidak ada kointegrasi di antara *x* dan *y*. Sebaliknya, jika $\rho_i < 1\Lambda_i$, maka ada kointegrasi di antara *x* dan *y*. Jika $\rho_i < 1\Lambda_i$ dan $\beta_i = \beta$, maka terdapat homogenous cointegration, dan sebaliknya bersifat heterogenous cointegration. Jika sebenarnya berlaku heterogenous cointegration tetapi spesifikasinya "dipaksakan" homogenous ($\beta_i = \beta$), maka residual akan bersifat nonstationary.

Pesaran (2004) mengusulkan pengujian cross section dependensi (CD test) dengan formulasi berikut

$$CD = \sqrt{\left(\frac{2}{N(N-1)}\right)}\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\sqrt{T_{ij}}\hat{\rho}_{ij}\right) \tag{14.29}$$

Statistik CD memiliki distribusi standar normal dengan hipotesis null tidak ada *cross section dependence*. Pengujian cross section dependence dapat dilakukan sebelum regresi (ex-ante) atau sesudah regresi (ex-post).

Contoh 14.4

Kita akan mereplikasi studi Eberhardt, Helmers, dan Strauss (2013). Data STATA dapat didownload dari situs: https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/21735. Di sini file telah disimpan dengan nama EHS_2013.dta. Variabel utama dalam data ini (semua dalam bentuk log natural) adalah real value added dari output nasional (lny), jumlah jam kerja pegawai di bidang riset (lnl), jumlah capital stock (lnk), dan jumlah stock modal R&D (lnrd). Datanya bersifat *unbalanced panel* yang meliputi 10 negara pada periode 1980-2005: 2637 observasi.

Pertama, kita akan melakukan pengujian unit root terhadap variabel lny, lnl, lnk, dan lnrd. Pengujian unit root dilakukan dengan metode Maddala dan Wu-MW (1999) dan Pesaran-CIPS (2007) yang dilakukan dengan perintah **multipurt.** Ini bukan merupakan routine default STATA; di mana program (ado.file) multipurt dibuat oleh Eberhardt(b) (2011). Pembaca harus melakukan instalasi terlebih dahulu dengan perintah **ssc install multipurt**. Dengan menggunakan lag = 1; program ini dijalankan dengan syntax **multipurt lny lnl lnk lnrd, lags(1)**.

Dapat dilihat dari Tabel 14.12, dan dengan menggunakan spesifikasi tanpa trend (hanya konstanta) maka pengujian unit root yang menggunakan CIPS akan menunjukkan hipotesis null: non-stationary pada seluruh variabel tidak dapat ditolak. Sedangkan jika menggunakan MW, kesimpulan sebaliknya adalah hipotesis null ditolak (variabel adalah stasioner). Namun demikian, karena MW mengasumsikan cross section dependence seperti akan kita lihat nanti dilanggar, maka kesimpulan yang digunakan adalah hasil dari CIPS.

First and Second Generation Panel Unit Root Tests

Variables tested: lny lnl lnk lnrd
Group variable: id
Number of groups: 119
Total # of observations: 2637+
Average # of observations: 23.76+
Panel is unbalanced and has gaps
+ Full sample statistics prior to testing.

(A) Maddala and Wu (1999) Panel Unit Root test (MW)

	Specification without trend		
Variable	lags	chi_sq	p-value
lny	0	434.941	0.000
lny	1	387.979	0.000
lnl	0	275.122	0.049
lnl	1	356.499	0.000
lnk	0	475.552	0.000
lnk	1	353.653	0.000
lnrd	0	837.782	0.000
lnrd	1	360.628	0.000

	Specification with trend		
Variable	lags	chi_sq	p-value
lny	0	272.891	0.060
lny	1	471.180	0.000
lnl	0	182.008	0.997
lnl	1	458.859	0.000
lnk	0	218.261	0.816
lnk	1	381.984	0.000
lnrd	0	115.956	1.000
lnrd	1	548.092	0.000

(B) Pesaran (2007) Panel Unit Root test (CIPS)

	Specification without trend			
Variable	lags	Zt-bar	p-value	t-bar
lny	0	2.334	0.990	.
lny	1	2.501	0.994	.
lnl	0	3.459	1.000	.
lnl	1	-0.238	0.406	.
lnk	0	8.010	1.000	.
lnk	1	8.435	1.000	.
lnrd	0	9.452	1.000	.
lnrd	1	7.129	1.000	.

	Specification with trend			
Variable	lags	Zt-bar	p-value	t-bar
lny	0	1.106	0.866	.
lny	1	-3.296	0.000	.
lnl	0	3.451	1.000	.
lnl	1	-1.595	0.055	.
lnk	0	8.012	1.000	.
lnk	1	-2.619	0.004	.
lnrd	0	10.264	1.000	.
lnrd	1	0.573	0.717	.

Null for MW and CIPS tests: series is I(1).
MW test assumes cross-section independence.
CIPS test assumes cross-section dependence is in
form of a single unobserved common factor.

-multipurt- uses Scott Merryman's -xtfisher- and
Piotr Lewandowski's -pescadf-.

TABEL 14.12. Pengujian Panel Unit Root dengan Routine Multipurt, Panel Atas (Maddala dan Wu), Panel Bawah (CIPS).

Uji kointegrasi dilakukan dengan menggunakan teknik Westerlund (2008). Di sini variabel lny digunakan sebagai variabel dependen; sedangkan lnl, lnk, dan lnrd digunakan sebagai variabel independen. Perintah untuk menjalankan uji kointegrasi diberikan sebagai **xtcointtest westerlund lny lnl lnk lnrd**. Dapat dilihat dari Tabel 14.13, bahwa hasil uji kointegrasi menunjukan hipotesis null tidak ada kointegrasi pada seluruh cross section dapat ditolak. Dengan demikian, paling tidak ada satu cross section yang memiliki kointegrasi.


```
Westerlund test for cointegration
------------------------------------------------------------------------------
Ho: No cointegration                   Number of panels       =    119
Ha: Some panels are cointegrated       Avg. number of periods =  22.16

Cointegrating vector: Panel specific
Panel means:          Included
Time trend:           Not included
AR parameter:         Panel specific
------------------------------------------------------------------------------
                                        Statistic         p-value
------------------------------------------------------------------------------
 Variance ratio                          -6.3012          0.0000
------------------------------------------------------------------------------
```

TABEL 14.13. Pengujian Uji Kointegrasi Panel dengan Metode Westerlund (2008)

Pengujian cross section dependence (CD test) dilakukan dengan program (ado.file) xtcd. Program ini bukan merupakan default STATA, sehingga pembaca perlu melakukan instalasi terlebih dahulu dengan perintah **ssc install xtcd**. Pengujian CD Test dilakukan dengan perintah berikut **xtcd lny lnl lnk lnrd**. Hasilnya disajikan pada Tabel 14.14, di mana hipotesis null *no cross-section dependence* pada variabel lny, lnl, lnk, dan lnrd adalah ditolak (dengan *p* value: 0,000).

```
Average correlation coefficients & Pesaran (2004) CD test

Variables series tested: lny lnl lnk lnrd
                                 Group variable: id
                               Number of groups: 119
                     Average # of observations: 19.53
                                     Panel is: unbalanced
```

Variable	CD-test	p-value	corr	abs(corr)
lny	110.44	0.000	0.290	0.589
lnl	105.45	0.000	0.295	0.567
lnk	199.00	0.000	0.550	0.770
lnrd	149.64	0.000	0.398	0.781

```
Notes: Under the null hypothesis of cross-section
       independence CD ~ N(0,1)
```

TABEL 14.14. Pengujian Cross Section Dependence dengan Metode Pesaran (2004)

Selanjutnya, kita akan melakukan estimasi dengan menggunakan Mean Group. Blackburne III dan Frank (2007) telah membuat program **xtmg** dan **xtpmg** yang memberikan output tidak hanya dalam bentuk long run (jangka panjang), tetapi juga *short run* (jangka pendek = ECM)[5]. Estimasi Mean Group (dengan tren) dilakukan dengan perintah **xtmg lny lnl lnk lnrd, trend**. Estimasi ini tidak memiliki representasi ECM, karena koefisien slope atau kemiringan diperoleh secara adhoc, yaitu sebagai rata-rata slope individu cross section.

[5] Estimasi nonheterogenous panel dapat menjadi sangat kompleks dan membutuhkan banyak memori. Pastikan STATA yang dimiliki setidaknya versi MP, dan alokasikan ukuran memori dengan perintah **set matvar** setidaknya 10000 dan ukuran matriks dengan perintah **set matsize** setidaknya 5000 sebelum melakukan regresi. Penulis dengan menggunakan laptop spesifikasi icore3 dan RAM 4GB melakukan estimasi xtpmg pada contoh ini yang membutuhkan waktu sekitar 10 menit.

```
Pesaran & Smith (1995) Mean Group estimator

All coefficients represent averages across groups (group variable: id)
Coefficient averages computed as unweighted means

Mean Group type estimation                    Number of obs      =      2,637
Group variable: id                            Number of groups   =        119

                                              Obs per group:
                                                           min =         11
                                                           avg =       22.2
                                                           max =         26

                                              Wald chi2(3)       =      87.33
                                              Prob > chi2        =     0.0000

------------------------------------------------------------------------------
         lny |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnl |     .56774   .0864325     6.57   0.000     .3983355    .7371445
         lnk |    .116649   .1221103     0.96   0.339    -.1226829    .3559808
       lnrd |   -.0577263   .0792549    -0.73   0.466     -.213063    .0976104
       trend |    .0224639   .0075591     2.97   0.003     .0076484    .0372794
       _cons |    4.468307   .7936524     5.63   0.000     2.912777    6.023837
------------------------------------------------------------------------------
Root Mean Squared Error (sigma): 0.0507
Variable trend refers to the group-specific linear trend terms.
Share of group-specific trends significant at 5% level: 0.504 (= 60 trends)
```

TABEL 14.15. Hasil Estimasi Mean Group

Hasil estimasi Mean Group yang disajikan pada Tabel 14.15 adalah koefisien jangka panjang. Dapat dilihat di sini bahwa hanya variabel ln yang signifikan. Sementara itu, estimasi yang mengasumsikan adanya *common correlated effect* dapat dilakukan dengan perintah **xtmg lny lnl lnk lnrd, cce.** Hasilnya disajikan pada Tabel 14.16. Dengan mengasumsikan adanya *common factor* mulai terlihat perbaikan pada efisiensi di mana standar error dari parameter variabel independen mengalami penurunan.

Terakhir, kita akan melakukan estimasi dengan menggunakan teknik *Pooled Mean Group*. Dengan teknik ini, hubungan di antara

```
Pesaran (2006) Common Correlated Effects Mean Group estimator

All coefficients represent averages across groups (group variable: id)
Coefficient averages computed as unweighted means

Mean Group type estimation                    Number of obs     =       2,637
Group variable: id                            Number of groups  =         119

                                              Obs per group:
                                                           min =          11
                                                           avg =        22.2
                                                           max =          26

                                              Wald chi2(3)      =       81.74
                                              Prob > chi2       =      0.0000
```

lny	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lnl	.5989646	.0665471	9.00	0.000	.4685346	.7293945
lnk	.2438184	.1432445	1.70	0.089	-.0369356	.5245724
lnrd	.0353425	.079482	0.44	0.657	-.1204393	.1911243
lny_avg	1.053031	.1245706	8.45	0.000	.8088772	1.297185
lnl_avg	-1.071046	.1882613	-5.69	0.000	-1.440031	-.7020605
lnk_avg	.1262476	.4042894	0.31	0.755	-.6661451	.9186404
lnrd_avg	-.4133752	.1054467	-3.92	0.000	-.6200469	-.2067034
_cons	1.240475	2.306341	0.54	0.591	-3.279869	5.760819

```
Root Mean Squared Error (sigma): 0.0374
Cross-section averaged regressors are marked by the suffix avg.
```

TABEL 14.16. Hasil Estimasi Mean Group - Common Correlated Effect

variabel dependen akan memiliki dua karakter: jangka panjang dan jangka pendek. Hubungan jangka panjang direstriksi bagi semua cross section; sedangkan slope/kemiringan jangka pendek dapat bervariasi. Estimasi dilakukan dengan perintah berikut **xtpmg d.lny d.lnl d.lny d.lnrd, lr(l.lny lnl lnk lnrd) ec(ec)**. Dapat dilihat di sini bahwa seperti halnya model time series, variabel jangka pendek ditransformasikan sebagai bentuk difference; sedangkan level digunakan untuk persamaan jangka panjang (term **lr()**). Koefisien

ECT diberi nama ec; dan pembaca dapat memberikan nama lain jika diinginkan.

Hasil estimasi disajikan pada Tabel 14.17. Output terdiri dari dua bagian di mana bagian atas adalah persamaan jangka panjang,

```
Iteration 0:   log likelihood =  4639.0952  (not concave)
Iteration 1:   log likelihood =  4683.8561
Iteration 2:   log likelihood =  4688.2616
Iteration 3:   log likelihood =  4693.5722
Iteration 4:   log likelihood =  4693.7581
Iteration 5:   log likelihood =  4693.7853
Iteration 6:   log likelihood =  4693.7854

Pooled Mean Group Regression
(Estimate results saved as pmg)

Panel Variable (i): id                    Number of obs      =      2518
Time Variable (t): year                   Number of groups   =       119
                                          Obs per group: min =        10
                                                        avg =      21.2
                                                        max =        25

                                          Log Likelihood     =  4693.785
```

D.lny	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ec						
lnl	-.2094011	.0298868	-7.01	0.000	-.2679782	-.150824
lnk	.4819969	.0277768	17.35	0.000	.4275555	.5364384
lnrd	.0655186	.0095554	6.86	0.000	.0467903	.0842469
SR						
ec	-.227076	.0238444	-9.52	0.000	-.2738103	-.1803418
lnl						
D1.	.6154335	.0514169	11.97	0.000	.5146583	.7162087
lnk						
D1.	.1456646	.102456	1.42	0.155	-.0551454	.3464746
lnrd						
D1.	.0249017	.0760565	0.33	0.743	-.1241663	.1739698
_cons	1.20337	.1256734	9.58	0.000	.9570543	1.449685

TABEL 14.17. Hasil Estimasi Pooled Mean Group

sedangkan bagian bawah adalah persamaan jangka pendek. Kita dapat menulis output dalam format sebagai berikut (dengan pembulatan 3 digit):

$$\Delta lny_{it} = 0{,}615^{***}\Delta lnl_{it} + 0{,}146\Delta lnk_{it} + 0{,}024\Delta lnrd_{it}$$
$$-0{,}227^{***}\, y_{i,t-1} + 0{,}209^{***}\, lnk_{i,t-1} - 0{,}482^{***}\, lnl_{i,t-1} - 0{,}065^{***}\, lnrd_{i,t-1} + u_{it}$$

(14.30)

di mana tanda *** menunjukkan bahwa koefisien terkait adalah signifikan pada α = 1%. Koefisien ECT adalah negatif dengan nilai absolut di bawah satu (dan signifikan secara statistik).

Bab 15

Model Variabel Dependen yang Terbatas

Model regresi linear yang telah dibahas sebelumnya adalah menggunakan variabel dependen atau tergantung yang bersifat numeris dan diasumsikan dapat mengambil nilai apa saja (*unbounded*). Asumsi yang terakhir ini pada beberapa penelitian dapat bersifat kurang realistis dan tidak dapat diterapkan. Penelitian dengan variabel dependen yang bersifat kualitatif (kategoris) ini misalnya adalah keputusan membeli atau tidak suatu produk yang dikaitkan dengan serangkaian variabel independen (demografis, daya beli, dan psikologis). Dalam hal ini, nilai variabel dependen hanyalah 1 (jika beli) dan 0 (jika tidak). Model regresi yang digunakan untuk data semacam ini disebut model *binary response*, yang di antaranya adalah model probabilitas linear, logit, dan probit.

Selanjutnya juga akan dibahas mengenai pengembangan variabel dependen menjadi lebih dari satu pilihan (multikategori). Apabila variabel dependen memiliki lebih dari satu kategori, maka sifat pilihan dapat berupa nominal (klasifikasi tidak memiliki arti urutan) dan ordinal (klasifikasi memiliki arti urutan). Untuk variabel yang bersifat nominal, dapat digunakan model multinominal logit sedangkan jika bersifat ordinal dapat digunakan model *ordered response*.

Sifat variabel dependen lainnya yang menghambat penerapan OLS adalah *count data*. Di sini nilai variabel dependen atau tergantung harus bersifat integer dan nonnegatif. Variabel semacam ini misalnya adalah frekuensi kunjungan, jumlah anak, dan pembelian kendaraan bermotor (oleh seorang individu). Regresi Poisson dapat mengakomodasikan variabel semacam ini.

Jika nilai variabel dependen bersifat kontinu tetapi hanya terbatas pada range atau rentang tertentu, maka ini juga merupakan hambatan bagi penerapan OLS secara langsung. Variabel semacam

ini meliputi Indeks Prestasi, persentase kepesertaan pensiun, dan nilai TOEFL. Data yang dimiliki disebut *censored* jika nilai variabel dependen dibatasi. Model untuk mengatasi masalah ini disebut *censored regression*. Semua teknik yang digunakan untuk mengatasi permasalahan yang disebut tersebut termasuk kelas Model Regresi dengan Variabel Dependen yang Terbatas: *Limited Dependent Variable Model*.

MODEL REGRESI BINARY RESPONSE

Berbeda dengan regresi yang telah dipelajari sebelumnya, interpretasi hubungan antara variabel dependen dan independen pada model *binary response* bersifat probabilistik. Dengan kata lain, jika kita menotasikan $y = 1$ sebagai terjadinya suatu event (dan $y = 0$, bukan event tersebut), maka regresi OLS

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u \quad (15.1)$$

harus diinterpretasikan sebagai probabilitas terjadinya $y = 1$, dengan syarat x_j bernilai tertentu, atau

$$P(y = 1|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \quad (15.2)$$

Jika kita menggunakan *Linear Probability Model* (LPM), maka Persamaan 15.1 akan diestimasi dari data dengan menggunakan teknik OLS. Semua prosedur dan interpretasi yang dilakukan adalah sama dengan yang telah dipelajari sebelumnya.

Model ini memiliki 2 kelemahan. Pertama, ada pembatasan yang bersifat *adhoc*. Ini terjadi apabila *fitted value* variabel dependen lebih dari 1, sehingga ia dianggap 1 dan sebaliknya jika di bawah

0, maka akan dianggap 0 (1 dan 0 adalah batas atas dan batas bawah "yang dipaksakan" kepada nilai variabel dependen). Dengan demikian, *fitted value* = 1,50 dianggap sama dengan *fitted value* = 1,05, yaitu sama-sama memiliki probabilitas terjadinya $y = 1$. Kelemahan lain adalah model ini mengalami heterokedastisitas (melanggar asumsi Gauss-Markov). Meskipun demikian, model ini tetap banyak digunakan dan cukup valid terutama jika nilai dari variabel independen adalah terdistribusi di sekitar rata-rata (tidak terlalu menyebar).

Seperti telah diuraikan sebelumnya, kelemahan utama dari LPM adalah adanya batas atas dan bawah yang bersifat adhoc. Salah satu model alternatif yang dapat mengatasi masalah ini adalah menggunakan fungsi kumulatif densitas atau fungsi asimtotik (antara 0 dan 1) pada fungsi objektifnya. Salah satu model semacam ini adalah model logit dan probit. Dalam bentuk umum model dengan fungsi densitas kumulatif dapat ditulis sebagai

$$Pr(y = 1|x) = G(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) = G(\beta_0 + x\beta) \quad (15.3)$$

di mana $x\beta$ menunjukkan term perkalian vektor untuk meringkas $\sum \beta_j x_j$.

Pada model logit, G(.) adalah fungsi logistik yang memiliki bentuk

$$G(z) = \frac{e^z}{1 + e^z} \quad (15.4)$$

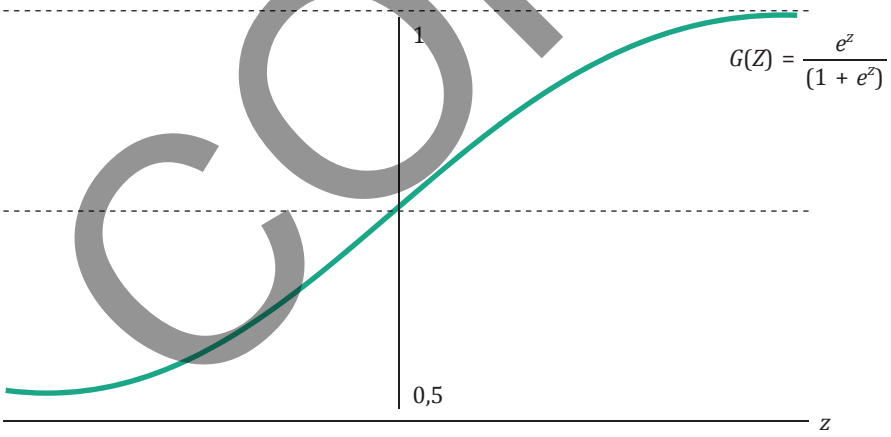
Sedangkan pada model probit G(.) adalah fungsi densitas kumulatif normal, yakni

$$G(z) = \Phi(z) = \int_{-\infty}^{z} \varphi(v)dv \qquad (15.5)$$

di mana

$$\varphi(z) = (2\pi)^{-1/2} e^{(-z^2/2)} \qquad (15.6)$$

Dapat ditunjukkan di sini bahwa baik fungsi Persamaan 15.4 maupun 15.5 adalah asimtotik ke arah 0 dan 1; yakni ($G(z) \rightarrow 0$ ketika z $-\infty$ dan $G(z) \rightarrow 1$ ketika $z \rightarrow \infty$). Sebagai ilustrasi, Gambar 15.1 menunjukkan suatu fungsi logistik.



GAMBAR 15.1. Fungsi Logistik

Modifikasi lebih lanjut terhadap Persamaan 15.4 dapat dilakukan sehingga kita memperoleh hasil sebagai berikut

$$P(y=1) = P = G(z) = G(\beta_0 + x\beta) = \frac{e^{(\beta_0 + x\beta)}}{1 + e^{\beta_0 + x\beta}} = \frac{1}{1 + e^{-(\beta_0 + x\beta)}} \quad (15.7)$$

serta nilai 1 – p dapat dihitung sebagai

$$1 - P = 1 - \frac{1}{1 + e^{-(\beta_0 + x\beta)}} = \frac{e^{-(\beta_0 + x\beta)}}{1 + e^{-(\beta_0 + x\beta)}} \quad (15.8)$$

sehingga dapat dihitung

$$ln\left(\frac{P}{1-P}\right) = ln\left(\frac{1}{1 + e^{-(\beta_0 + x\beta)}} \times \frac{1 + e^{-(\beta_0 + x\beta)}}{e^{-(\beta_0 + x\beta)}}\right) = ln\left(\frac{1}{e^{-(\beta_0 + x\beta)}}\right)$$
$$ln\left(\frac{P}{1-P}\right) = z = \beta_0 + x\beta \quad (15.9)$$

Rumus pada Persamaan 15.9 disebut dengan *odd ratio*, yakni rasio antara probabilitas $y = 1$ (terjadinya event) dan $y = 0$ (tidak terjadinya event).

Kita dapat menurunkan suatu model logit atau probit melalui variabel laten, yang ditentukan sebagai

$$y^* = \beta_0 + x\beta + e, y = 1[y^* > 0] \quad (15.10)$$

Persamaan 15.10 menunjukkan bahwa $y = 1$ jika $y^* > 0$.

Seperti biasa, perhatian kita terutama tertuju pada apa dampak dari perubahan satu/lebih variabel independen terhadap variabel dependen. Hal ini dapat dihitung sebagai berikut:

$$\frac{\partial p(x)}{\partial x_j} = g(\beta_0 + \boldsymbol{x\beta})\beta_j \tag{15.11}$$

di mana

$$g(z) = \frac{dG}{dz}(z) \tag{15.12}$$

Jika x_j adalah variabel dummy (misalnya, 0 dan 1), dampak parsial terjadinya perubahan variabel tersebut dari nol ke satu dapat dihitung sebagai

$$G(\beta_0 + \beta_1 x_1 + \ldots + \beta_j + \ldots + \beta_k x_k) - G(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) \tag{15.13}$$

Karena sifat $G(z)$ yang nonlinear, maka estimasi parameter model logit dan probit dilakukan melalui prosedur *Maximum Likelihood Estimation* (MLE).

Uji signifikansi pada parameter dilakukan dengan melihat nilai *p* value yang dibandingkan dengan α (*level of significance*) yang digunakan pada hipotesis null dua arah. Sedangkan untuk *overall significance*, kita menggunakan *likelihood ratio statistiks* (LR statistiks). Statistik LR dapat dihitung dengan rumus berikut ini:

$$LR = 2(\ell_{ur} - \ell_0) \tag{15.14}$$

di mana ℓ_{ur} dan ℓ_0 adalah nilai log likelihood masing-masing untuk fungsi unrestricted (model lengkap) dan restricted (hanya intersep).

Nilai log likelihood umumnya negatif di mana ℓ_{ur} lebih tidak negatif dari $\ell_0 (\ell_{ur} \geq \ell_0)$. Nilai LR mengikuti distribusi χ^2 dengan $df = k$.

Untuk menilai kelaikan suai (*goodness of fit*) model ini dapat digunakan dua kriteria, yakni

a. **Percent Correctly Predicted**, yang menunjukkan persentase prediksi yang benar dengan threshold/cut off tertentu (biasanya 0,5). Di sini semua nilai $P(x) > 0{,}5$ akan dikategorikan sebagai prediksi terjadinya event atau peristiwa ($Y = 1$); kemudian statistik Hosmer-Lemeshow dapat dihitung dengan formulasi berikut.

$$H = \sum_{g=1}^{G} \frac{(O_{1g} - E_{1g})^2}{N_g \pi_g (1 - \pi_g)} \qquad (15.15)$$

di mana O1g adalah observasi pada event $Y = 1$; dan E_{1g} adalah nilai ekspektasi pada event $Y = 1$, N_g adalah jumlah event $Y = 1$ pada kelompok g dan π_g adalah probabilitas terjadinya event $Y = 1$ pada kelompok g. G adalah jumlah kelompok yang dipandang memiliki karakteristik probabilitas $Y = 1$ yang serupa. Statistik H ini memiliki distribusi $\chi 2$ dengan derajat kebebasan G-2 dengan hipotesis null: tidak ada perbedaan yang sistematis di antara nilai prediksi dengan nilai yang terobservasi pada setiap event null untuk setiap kelompok g.

b. **Pseudo R-Squared** (Mc Faden, 1974). Ini adalah ukuran yang analog dengan R^2 pada estimasi OLS yang biasa[1]. Ada pun rumus yang digunakan adalah

[1] Meskipun analog, namun Pseudo R^2 tidak memiliki arti persentase varians variabel dependen yang dapat dijelaskan oleh variabel independen. Dengan demikian, manfaat dari statistik ini terbatas untuk mengevaluasi *goodness of fit* dari berbagai pilihan model.

$$Pseudo\ R^2 = 1 - \frac{\ell_{ur}}{\ell_0} \tag{15.16}$$

Contoh 15.1

Kita akan menggunakan file transport.dta dari Ben-Akiva dan Lerman (1985). File ini berisi respons dari 21 orang sampel mengenai pilihan moda transportasi (auto; 1 jika memilih naik mobil pribadi dan 0 jika memilih naik bis umum). Pilihan moda transportasi tersebut dipengaruhi oleh variabel selisih waktu (naik bis umum-mobil) perjalanan (dalam satuan 10 menit; dtime). Regresi dengan OLS (*Linear Probability Model*) yang dilakukan dengan perintah **reg auto dtime** dapat dilihat pada Tabel 15.1.

Source	SS	df	MS	Number of obs	=	21
				F(1, 19)	=	29.88
Model	3.20218144	1	3.20218144	Prob > F	=	0.0000
Residual	2.0359138	19	.107153358	R-squared	=	0.6113
				Adj R-squared	=	0.5909
Total	5.23809524	20	.261904762	Root MSE	=	.32734

auto	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
dtime	.0703099	.0128616	5.47	0.000	.0433902	.0972296
_cons	.4847951	.0714494	6.79	0.000	.3352497	.6343404

TABEL 15.1. Estimasi OLS (LPM) Pilihan Moda Transportasi

Dapat dilihat di sini bahwa sesuai hipotesis, kenaikan selisih waktu tempuh secara signifikan akan mempengaruhi probabilitas pilihan naik mobil pribadi (dengan koefisien sebesar 0,070 dan *p* value sebesar 0,000). Estimasi yang menggunakan logit dan probit

diberikan dengan perintah (masing-masing): **logit auto dtime** dan **probit auto dan dtime**. Hasil estimasi ini disajikan pada Tabel 15.2.

```
Logistic regression                               Number of obs     =         21
                                                  LR chi2(1)        =      16.73
                                                  Prob > chi2       =     0.0000
Log likelihood = -6.1660422                       Pseudo R2         =     0.5757

------------------------------------------------------------------------------
        auto |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       dtime |   .5310983   .2064228     2.57   0.010     .126517    .9356795
       _cons |  -.2375754   .7504766    -0.32   0.752    -1.708483    1.233332
------------------------------------------------------------------------------

Probit regression                                 Number of obs     =         21
                                                  LR chi2(1)        =      16.73
                                                  Prob > chi2       =     0.0000
Log likelihood = -6.1651585                       Pseudo R2         =     0.5758

------------------------------------------------------------------------------
        auto |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       dtime |   .2999898   .1028673     2.92   0.004     .0983735    .5016061
       _cons |  -.0644338   .3992438    -0.16   0.872    -.8469372    .7180696
------------------------------------------------------------------------------
```

TABEL 15.2. Estimasi Logit (Panel Atas) dan Probit (Panel Bawah) Pilihan Moda Transportasi

Sejalan dengan hasil regresi LPS, estimasi yang menggunakan logit dan probit juga menghasilkan koefisien yang positif bagi variabel dtime. Semakin besar selisih waktu, semakin besar probabilitas seseorang akan memilih menggunakan mobil pribadi. Amemiya (1981) menyarankan prosedur *rule of thumb* untuk membuat koefisien logit dapat dibandingkan dengan LPM, yakni dengan mengalikan parameter variabel independen dengan 0,25 dan khususnya kepada konstanta ditambahkan kembali dengan 0,5[2].

[2] Wooldridge (2019) menyarankan perkalian dengan angka 0,4 untuk hasil estimasi probit.

Statistik LR Chi2 pada model logit dan probit sama-sama memberikan nilai 16,73 dengan *p* value hitung sebesar 0,000. Dengan kata lain, variabel independen memiliki pengaruh yang secara statistik signifikan dalam menjelaskan variasi pada variabel dependen. Untuk melakukan evaluasi atas *goodness of fit* lebih lanjut kita dapat menggunakan tabel klasifikasi; dan melakukan pengujian Hosmer and Lemeshow. Perintah STATA diberikan sebagai **estat gof, group(10)**[3] yang dilakukan setelah regresi logistik.

```
Logistic model for auto, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)

      number of observations =        21
            number of groups =        10
    Hosmer-Lemeshow chi2(8) =      9.18
                Prob > chi2 =      0.3270
```

TABEL 15.3. Pengujian Hosmer and Lemeshow untuk Model Logit

Seperti ditunjukkan oleh Tabel 15.4, statistik hitung Hosmer and Lemeshow untuk model logit (pembaca dipersilahkan mencoba sendiri model probit) tidak dapat menolak hipotesis null (*p* value hitung sebesar 0,3270) tidak adanya perbedaan yang sistematis antara nilai prediksi model (untuk auto = 1) dan nilai observasinya. Secara kualitatif kita dapat melihat dukungan terhadap hasil pengujian statistik ini melalui tabel klasifikasi (Tabel 15.4).

Seperti ditunjukkan oleh Tabel 15.4, model yang diestimasi telah dapat mengklasifikasi dengan benar event $Y = 1$. Prediksi akan diklasifikasikan sebagai benar jika (a) nilai prediksi (P(auto = 1) ≥

[3] Jumlah group = 10 adalah default dari STATA; kecuali ada pertimbangan tertentu angka default ini dapat digunakan.

```
Logistic model for auto

                    True
Classified |         D            ~D |        Total
-----------+------------------------+-----------------
     +     |         9             1 |           10
     -     |         1            10 |           11
-----------+------------------------+-----------------
   Total   |        10            11 |           21

Classified + if predicted Pr(D) >= .5
True D defined as auto != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   90.00%
Specificity                     Pr( -|~D)   90.91%
Positive predictive value       Pr( D| +)   90.00%
Negative predictive value       Pr(~D| -)   90.91%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    9.09%
False - rate for true D         Pr( -| D)   10.00%
False + rate for classified +   Pr(~D| +)   10.00%
False - rate for classified -   Pr( D| -)    9.09%
--------------------------------------------------
Correctly classified                        90.48%
--------------------------------------------------
```

TABEL. 15.4. Tabel Klasifikasi Estimasi Logistik Moda Transportasi

0,5) dan individu yang bersangkutan memang menggunakan mobil pribadi dan (b) nilai prediksi (P(auto = 1) < 0,5) serta individu yang bersangkutan menggunakan bis umum. Selain dua kategori tersebut, model juga telah melakukan kesalahan klasifikasi. Dalam kasus ini, telah terjadi kesalahan klasifikasi sebesar 9,52%.

Sebagai penutup, perlu diperhatikan bahwa koefisien-koefisien pada model regresi tidak dapat diinterpretasikan sebagai sensitivitas (dy/dy) layaknya regresi standar. Untuk mengetahui dampak suatu variabel independen terhadap variabel dependen akan sangat tergantung pada titik di mana ia dievaluasi. Sebagai contoh, pada Tabel 15.5 disajikan prediksi probabilitas penggunaan mobil pribadi pada 2 titik variabel penjelas (dtime); yakni 2 (20 menit) dan 3 (30

menit). Dapat dilihat di sini bahwa delta perubahan probabilitas hampir sebesar 0,1 (dari 0,69 ke 0,80)[4]; yang berbeda dari koefisien model logistik yang sebesar 0,531 (lihat Tabel 15.2 panel atas).

```
Adjusted predictions                              Number of obs     =         21
Model VCE    : OIM

Expression   : Pr(auto), predict()
at           : dtime           =           2
```

	Delta-method Margin	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	.6952164	.1684049	4.13	0.000	.3651489	1.025284

```
Adjusted predictions                              Number of obs     =         21
Model VCE    : OIM

Expression   : Pr(auto), predict()
at           : dtime           =           3
```

	Delta-method Margin	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	.7950631	.1451628	5.48	0.000	.5105493	1.079577

TABEL 15.5. Prediksi Marginal Probabilitas auto = 1 pad dtime = 2 (panel atas) dan dtime = 3 (panel bawah)

[4] Angka-angka ini diperoleh dari fitted value logistic function berikut: P(dtime=2;)

$$\frac{1}{1+e^{-(-0,24+0,53*2)}} = \frac{1}{1+e^{-0,82}} = 0,694 \text{ dan P(dtime=3; } \frac{1}{1+e^{-(-0,24+0,53*3)}} = \frac{1}{1+e^{-1,45}} = 0,80)$$

VARIABEL DEPENDEN MULTIKATEGORI

Dalam melakukan penelitian terkadang suatu variabel dependen yang bersifat multikategori akan digunakan. Variabel ini meliputi jenis pekerjaan, jenjang pendidikan, dan preferensi penggunaan moda transportasi. Kita ingin mengetahui bagaimana pilihan variabel dependen dipengaruhi oleh berbagai variabel independen lainnya. Untuk itu perlu pemodelan ekonometrika secara khusus.

Pilihan pemodelan dan teknik estimasi sangat tergantung pada karakter kategori variabel dependen. Jika karakter kategori bersifat nominal, yaitu klasifikasi tidak memiliki arti urutan superior-inferior, maka dapat digunakan model multinominal logit. Jika karakter kategori bersifat ordinal, yaitu terdapat arti urutan, maka digunakan model *ordered response*. Kita akan membahas terlebih dahulu variabel dependen multinominal baru kemudian *ordered response*.

Variabel Dependen Multinominal

Model variabel dependen multinominal pada dasarnya adalah pengembangan langsung dari model binary logit (atau probit). Analog dengan model variabel dependen kategoris yang pernah dibahas sebelumnya, di sini kita harus menentukan satu kategori sebagai acuan. Selanjutnya, estimasi dilakukan secara sekuensial sehingga untuk *j* kategori akan terdapat *j* – 1 model logit bivariat. Setiap *odd ratio* $(p_i/(1 - p_i)$ yang diperoleh sekarang harus diinterpretasikan sebagai probabilitas terjadinya event/kategori ke-*i* versus kategori acuan. Secara formal, hal ini dapat ditunjukkan sebagai berikut

$$p_{ij} = \frac{e^{x_i'\beta_j}}{1 + \sum_{h=2}^{m} e^{x_i'\beta_h}} \qquad (15.17)$$

Sebagai ilustrasi, misalkan kita akan mengestimasi hubungan antara suatu variabel kualitatif (Y: dengan 3 kategori) dan 2 variabel penjelas (X_1 dan X_2). Selanjutnya ditentukan bahwa kategori A adalah acuan sehingga terdapat 2 model logit biner yang akan diestimasi sebagai suatu sistem multinominal logit sebagai berikut:

$$pi_1 = \frac{1}{1 + e^{\beta_{20} + \beta_{21} x_1 + \beta_{22} x_2} + e^{\beta_{30} + \beta_{31} x_1 + \beta_{32} x_2}} \tag{15.18}$$

$$pi_2 = \frac{e^{\beta_{20} + \beta_{21} x_1 + \beta_{22} x_2}}{1 + e^{\beta_{20} + \beta_{21} x_1 + \beta_{22} x_2} + e^{\beta_{30} + \beta_{31} x_1 + \beta_{32} x_2}} \tag{15.19}$$

$$pi_3 = \frac{e^{\beta_{30} + \beta_{31} x_1 + \beta_{32} x_2}}{1 + e^{\beta_{20} + \beta_{21} x_1 + \beta_{22} x_2} + e^{\beta_{30} + \beta_{31} x_1 + \beta_{32} x_2}} \tag{15.20}$$

Dampak marjinal dari variabel penjelas harus dihitung secara tersendiri, mengingat sekarang parameter variabel pada suatu model logit bivariat tidak lagi dapat diinterpretasikan secara individual, tetapi harus mempertimbangkan parameter yang diperoleh pada model bivariat lainnya. Secara matematis hal ini dapat dirumuskan sebagai berikut

$$\frac{\partial P(y_i = j)}{\partial x_i} = p_{ij}(\beta_j - \sum_{h=2}^{m} p_{i_h} \beta_k) \tag{15.21}$$

Estimasi model multinominal logit dilakukan dengan menggunakan teknik maximum likelihood dan mengasumsikan bahwa residual di antara model logit bivariat adalah terdistribusi secara independen serta identik. Di samping itu, dalam menggunakan model logit

multinominal akan diterapkan asumsi "*independence of irrelevant alternatives*". Dengan kata lain, setiap kategori pada variabel dependen adalah unik dan tidak memiliki korelasi dengan kategori lainnya. Pelanggaran terhadap asumsi ini akan menimbulkan bias pada hasil estimasi.

Contoh 15.2

Kita akan menggunakan data pada file pendidikan.dta; yang dimodifikasi dari Adkins dan Hill (2011). File ini memuat pilihan pendidikan yang akan ditempuh dari 500 responden hipotetis pascalulus SMA. Pilihan yang diambil adalah 1 langsung kerja, 2 ambil pendidikan kejuruan Non Sarjana (D2 atau D3), dan 3 ambil S1 universitas. Pilihan yang diambil dimodelkan sebagai fungsi dari (a) nilai rata-rata mata ajaran utama (Nirat, skala 0-10), (b) penghasilan keluarga (pnhl_kel, dalam Rp Juta per tahun), (c) pendidikan orang tua (didik_ortu, 1 jika sarjana dan 0 jika tidak), dan (d) Jenis kelamin (Jen_Kel; 1 jika laki).

Kita akan melakukan estimasi model dengan menggunakan kategori langsung kerja (pendidikan = 1) sebagai kategori dasar (baseoutcome). Estimasi dengan menggunakan multinominal logit (dilakukan dengan perintah: **mlogit pendidikan nirat jen_kel pnhl_kel didik_ortu, baseoutcome(1)**) akan menghasilkan output regresi yang disajikan pada Tabel 15.6. Variabel didik ortu terlihat tidak memberikan pengaruh yang signifikan terhadap probabilitas relatif kategori 2 (versus bukan 2) dan kategori 3 (versus bukan 3). Hasil regresinya memberikan nilai LR2 sebesar 167,84 dengan *p* value hitung sebesar 0,000. Dengan kata lain, variabel-variabel independen secara bersama memiliki kemampuan menjelaskan

yang secara signifikan lebih baik dibandingkan model *unconditional average* (regresi hanya dengan konstanta)[5].

```
Multinomial logistic regression                 Number of obs     =        500
                                                LR chi2(8)        =     167.84
                                                Prob > chi2       =     0.0000
Log likelihood = -428.97341                     Pseudo R2         =     0.1636
```

pendidikan	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
1	(base outcome)					
2						
nirat	.3850121	.1010393	3.81	0.000	.1869786	.5830455
jen_kel	-.5574942	.286862	-1.94	0.052	-1.119733	.004745
pnhl_kel	.0134014	.0061254	2.19	0.029	.0013957	.025407
didik_ortu	.0446423	.4020069	0.11	0.912	-.7432768	.8325614
_cons	-1.771384	.596282	-2.97	0.003	-2.940075	-.6026925
3						
nirat	.8569792	.103001	8.32	0.000	.6551009	1.058857
jen_kel	-.1000473	.28654	-0.35	0.727	-.6616554	.4615608
pnhl_kel	.0186002	.00606	3.07	0.002	.0067228	.0304777
didik_ortu	.562078	.3817585	1.47	0.141	-.186155	1.310311
_cons	-4.549949	.6364944	-7.15	0.000	-5.797455	-3.302443

TABEL 15.6. Regresi Multinominal Logit Pilihan Pendidikan

Dapat dilihat di sini bahwa kategori hasil regresi baseoutcome akan kosong. Koefisien-koefisien regresi yang diperoleh bukanlah hal yang mudah diinterpretasikan. Hal itu menunjukkan dampak suatu variabel independen terhadap probabilitas relatif terjadinya kategori *j* terhadap kategori bukan *j*. Dengan demikian, suatu slope yang positif pada variabel nirat panel 2 (kategori: pilihan Pendidikan Kejuruan Non Sarjana) yang berarti kenaikan nilai rata-rata akan meningkatkan probabilitas untuk memilih kategori 2 versus 1 dan 3. Namun, perhatikan bahwa variabel nirat memiliki koefisien yang lebih besar pada regresi kategori 3; sehingga dapat dikatakan

[5] Kita tidak lagi dapat menghitung tabel klasifikasi serta statistik Hosmer dan Lemeshow pada model multikategoris (nominal maupun ordered); karena variabel dependen bukan lagi merupakan suatu klasifikasi melainkan suatu probabilitas relatif.

secara kualitatif dampaknya (dalam penentuan pilihan) lebih besar dibandingkan kategori 2.

```
Predictive margins                              Number of obs     =        500
Model VCE    : OIM

Expression   : Pr(pendidikan==1), predict(outcome(1))
at           : nirat           =           8
```

	Delta-method Margin	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	.0354347	.0113251	3.13	0.002	.013238	.0576314

```
Predictive margins                              Number of obs     =        500
Model VCE    : OIM

Expression   : Pr(pendidikan==2), predict(outcome(2))
at           : nirat           =           8
```

	Delta-method Margin	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	.1658586	.0280475	5.91	0.000	.1108864	.2208307

```
. margins, predict(outcome(3)) at(nirat=8)

Predictive margins                              Number of obs     =        500
Model VCE    : OIM

Expression   : Pr(pendidikan==3), predict(outcome(3))
at           : nirat           =           8
```

	Delta-method Margin	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	.7987067	.0307638	25.96	0.000	.7384107	.8590027

TABEL 15.7. Perhitungan Probabilitas Marjinal pada Pilihan Pendidikan: 1 (atas), 2 (tengah), dan 3 (bawah) untuk nirat = 8

Kita dapat menghitung probabilitas marjinal suatu nilai variabel independen tertentu dengan asumsi nilai variabel independen yang lain konstan (ceteris paribus). Misalkan kita ingin mengetahui probabilitas marjinal untuk nirat = 8; bagaimana dampaknya terhadap probabilitas pilihan 1, 2, dan 3. Hal ini dapat dihitung dengan perintah sebagai berikut:

```
margins, predict(outcome(1)) at(nirat=8)
margins, predict(outcome(2)) at(nirat=8)
margins, predict(outcome(1)) at(nirat=8)
```

Hasil dari perhitungan probabilitas marjinal disajikan pada Tabel 15.7. Di sini terlihat, sebagaimana telah ditunjukkan oleh Tabel 15.6, bahwa probabilitas memilih kategori 3 (melanjutkan hingga Sarjana) adalah tertinggi untuk responden yang memiliki nirat tinggi. Pembaca dapat membuat gambaran yang komprehensif dengan menghitung semua probabilitas marjinal pada variabel dan rentang nilai yang relevan, serta menyajikannya dalam bentuk tabel dan grafik untuk suatu laporan yang lebih komprehensif.

STATA juga dapat menghitung rata-rata probabilitas marjinal (*average marginal probability* = AME). Statistik AME adalah rata-rata dari probabilitas marjinal seluruh nilai sampel suatu variabel tertentu. Pada contoh sebelumnya, kita hanya menghitung probabilitas marjinal pada nirat = 8. Perintah untuk melakukan perhitungan AME bagi nirat diberikan sebagai **margins, dydx(nirat)**. Hasilnya disajikan pada Tabel 15.8

Seperti dapat dilihat pada Tabel 15.8, kenaikan nilai nirat berasosiasi dengan penurunan probabilitas seseorang berada dalam kategori 1 (versus 2 dan 3) dan 2 (versus 1 dan 3), serta meningkatkan probabilitas individu tersebut berada pada kategori 3

```
Average marginal effects                        Number of obs     =        500
Model VCE    : OIM

dy/dx w.r.t. : nirat
1._predict   : Pr(pendidikan==1), predict(pr outcome(1))
2._predict   : Pr(pendidikan==2), predict(pr outcome(2))
3._predict   : Pr(pendidikan==3), predict(pr outcome(3))
```

	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
nirat						
_predict						
1	-.0807636	.0100848	-8.01	0.000	-.1005295	-.0609976
2	-.0332752	.0111075	-3.00	0.003	-.0550455	-.0115049
3	.1140388	.0093603	12.18	0.000	.095693	.1323846

TABEL 15.8. Perhitungan Rata-rata Probabilitas Marjinal (AME) untuk Variabel Independen Nirat

(versus 1 dan 2). Keseluruhan rangkaian analisis ini dapat direplikasi dengan menggunakan model probit. Perintah STATA yang digunakan hampir serupa, di mana pembaca hanya perlu menggantikan **mlogit** dengan **mprobit**. Pembaca diharapkan dapat mengulangi sendiri untuk meningkatkan pemahaman.

Ordered Response

Pada model *ordered response*, hasil dari regresi (*fitted value* variabel dependen) merupakan suatu variabel laten (semacam indeks). Apabila *fitted value* variabel dependen ini melebihi batas tertentu, maka termasuk dalam kategori *j*. Secara formal

$$y_i^* = x_i'\beta + \varepsilon_i; E(\varepsilon_i) = 0 \qquad (15.22)$$

di mana

$$\begin{aligned} y_i &= 1; -\infty < y_i^* < \tau_1 \\ y_i &= j; \tau_{j-1} < y_i^* < \tau_j \;\; j = 2, \ldots, m-1 \\ y_i &= m; \tau_{m-1} < y_i^* < \infty \end{aligned} \tag{15.23}$$

Nilai indeks y_i^* tidak terobservasi (*latent variable*), yang kita amati hanyalah Y_i akan bernilai j jika indeks berada pada rentang τ_{j-1} dan τ_j.

Dampak marjinal dari perubahan variabel penjelas ke-i dapat diberikan sebagai berikut[6]

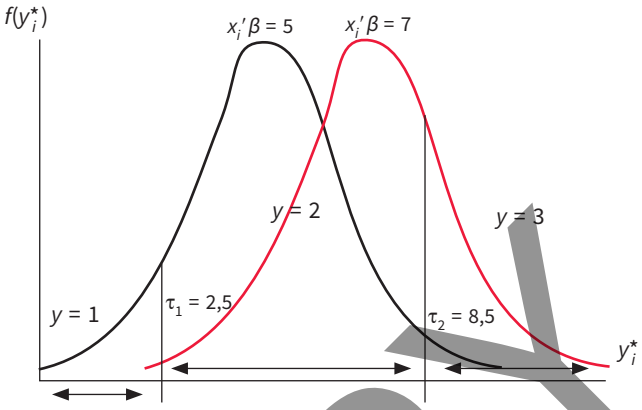
$$\frac{\partial P(y_i=j)}{\partial x_i} = \left(f(\tau_{j-1} - x_i'\beta) - f(\tau_j - x_i'\beta)\right)\beta \tag{15.24}$$

di mana f adalah fungsi densitas dari ε_i.

Ketika nilai $x_i'\beta$ meningkat, maka nilai indeks y_i^* juga akan meningkat. Setelah melewati batas tertentu, probabilitas $y_i = 1$ akan menurun sedangkan probabilitas $y_i = m$ akan meningkat. Probabilitas $y_i = 1$ dapat meningkat atau menurun tergantung pada posisi distribusi. Secara lebih spesifik hal ini dapat diilustrasikan secara grafis pada Gambar 15.2.

Gambar 15.2 menunjukkan skala indeks dan distribusi bagi suatu variabel laten dengan kategori (m) sebanyak 3. Batas pertama bernilai 2,5 dan batas kedua bernilai 8,5. Dengan demikian, sepanjang indeks variabel laten dimaksud bernilai sama atau kurang dari 2,5 maka variabel dependen akan cenderung memiliki nilai 1, cenderung bernilai 2 jika berada di antara 2,5 dan 8,5, serta cenderung bernilai 3 jika lebih dari 8,5.

[6] Untuk derivasi lihat Heij et al (2004) hal. 475.



GAMBAR 15.2. Ordered Response

Kurva berwarna hitam menunjukkan fungsi densitas untuk $x_i'\beta = 5$. Dapat dilihat di sini bahwa jika indeks variabel laten sama dengan lima, maka probabilitas memberikan respons 1 maupun 3 sangatlah kecil. Ketika $x_i'\beta = 7$ maka probabilitas respon $y = 1$ nyaris nol, sedangkan probabilitas memberikan respon $y = 2$ adalah tertinggi, dan memberikan respons $y = 3$ sudah meningkat.

Contoh 15.3

Kita kembali menggunakan file pendidikan.dta. Misalkan kita sekarang mengasumsikan bahwa mencapai gelar pendidikan yang lebih tinggi adalah sesuatu yang lebih diinginkan. Dengan kata lain, terdapat urutan preferensi Pendidikan = 3 > Pendidikan = 2 > Pendidikan = 1; simbol >menunjukkan hierarki preferensi. Estimasi dengan menggunakan ordered logit (dengan spesifikasi yang sama seperti pada Tabel 15.1) dilakukan atas perintah **ologit pendidikan nirat jen_kel pnhl_kel didik_ortu**.

```
Ordered logistic regression                     Number of obs      =        500
                                                LR chi2(4)         =     158.26
                                                Prob > chi2        =     0.0000
Log likelihood = -433.76318                     Pseudo R2          =     0.1543
```

pendidikan	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
nirat	.6107782	.0643628	9.49	0.000	.4846294	.7369269
jen_kel	.0870803	.1921365	0.45	0.650	-.2895003	.463661
pnhl_kel	.0106462	.0035055	3.04	0.002	.0037755	.0175168
didik_ortu	.4009576	.2486372	1.61	0.107	-.0863624	.8882777
/cut1	2.414616	.3997419			1.631136	3.198095
/cut2	4.067078	.428247			3.227729	4.906427

TABEL 15.9. Estimasi dengan Ordered Logit; Variabel Dependen Pendidikan

Tidak seperti regresi multinominal logit, di sini kita tidak perlu membuat suatu kategori acuan. Probabilitas kategori *j* akan dilihat melalui posisi nilai *fitted variable* terhadap *cut off*. Tabel 15.9 memperlihatkan 2 cut off (secara umum berlaku untuk kategori *j* akan ada *j* – 1 *cut off*). Cut off pertama yang sebesar 2,415 adalah transisi dari kategori Pendidikan = 1 ke Pendidikan = 2; sedangkan cut off kedua yang sebesar 4,067 adalah transisi dari kategori Pendidikan = 2 ke Pendidikan = 3. Di sini interpretasi dari koefisien regresi variabel independen adalah bersifat *straightforward*. Jika koefisien variabel tersebut positif, maka semakin tinggi nilainya semakin besar nilai variabel laten, dan akan semakin cenderung menjadi bagian dari kategori yang lebih tinggi.

Variabel nirat dan pnhl _kel adalah dua variabel yang memiliki koefisien positif dan secara statistik signifikan. Semakin tinggi nilai kedua variabel tersebut (ceteris paribus), semakin besar kemungkinan seorang responden berada dalam kategori lebih tinggi. Kita dapat mengkonversi kembali fitted value (yang berupa

variabel laten) menjadi probabilitas (yang lebih relevan untuk analisis kebijakan) dengan menggunakan perintah yang sama dengan multinominal logit. Sebagai contoh, kita ingin mengetahui probabilitas bahwa seseorang akan berada dalam kategori *j* ketika ia memiliki nirat sebesar 8 dan hal itu dapat diberikan pada perintah

```
margins, predict(outcome(1)) at(nirat=8)
margins, predict(outcome(2)) at(nirat=8)
margins, predict(outcome(1)) at(nirat=8)
```

Hasil estimasinya disajikan pada Tabel 15.10. Pembaca dapat membandingkan hasil dari kedua metode regresi (mlogit dan ologit) serta memperoleh secara kualitatif bahwa sebenarnya hal itu memberikan interpretasi yang sama. Nilai nirat yang tinggi berasosiasi dengan probabilitas seseorang akan berada pada kategori Pendidikan = 3. Kita juga dapat menghitung AME untuk model ologit dengan perintah **margins, dydx(nirat)** dengan hasil seperti yang ditunjukkan pada Tabel 15.11. Sekali lagi kita memperoleh interpretasi yang secara kualitatif serupa dengan multinominal logit.

REGRESI POISSON

Regresi Poisson digunakan ketika variabel dependen memiliki sifat *count data*, dan hanya dapat mengambil nilai *non-negative integer* (0, 1, 2, ...). Contoh variabel semacam ini adalah jumlah anak dari seorang wanita, berapa kali seorang ditahan dalam setahun, dan jumlah paten yang diajukan.

```
Predictive margins                                Number of obs     =        500
Model VCE    : OIM

Expression   : Pr(pendidikan==1), predict(outcome(1))
at           : nirat           =           8
```

	Delta-method					
	Margin	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	.045255	.0098375	4.60	0.000	.0259739	.064536

```
Predictive margins                                Number of obs     =        500
Model VCE    : OIM

Expression   : Pr(pendidikan==2), predict(outcome(2))
at           : nirat           =           8
```

	Delta-method					
	Margin	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	.1497014	.0216065	6.93	0.000	.1073534	.1920494

```
Predictive margins                                Number of obs     =        500
Model VCE    : OIM

Expression   : Pr(pendidikan==3), predict(outcome(3))
at           : nirat           =           8
```

	Delta-method					
	Margin	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	.8050436	.0291023	27.66	0.000	.7480042	.862083

```
. margins, dydx(nirat) at(nirat=8) predict(outcome(3))
```

TABEL 15.10. Perhitungan Probabilitas Marjinal pada Pilihan Pendidikan: 1 (atas), 2 (tengah), dan 3 (bawah) untuk Nirat = 8 dengan Menggunakan Ordered Logit

```
Average marginal effects                        Number of obs     =        500
Model VCE    : OIM

dy/dx w.r.t. : nirat
1._predict   : Pr(pendidikan==1), predict(pr outcome(1))
2._predict   : Pr(pendidikan==2), predict(pr outcome(2))
3._predict   : Pr(pendidikan==3), predict(pr outcome(3))
```

	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
nirat						
_predict						
1	-.0834518	.0081611	-10.23	0.000	-.0994472	-.0674563
2	-.0313506	.0044853	-6.99	0.000	-.0401416	-.0225597
3	.1148024	.0085142	13.48	0.000	.0981148	.13149

TABEL 15.11. Perhitungan Rata-rata Probabilitas Marjinal (AME) untuk Variabel Independen Nirat Model Regresi *Ordered Logit*

Model regresi Poisson adalah berbentuk eksponensial, yakni

$$E(y|x_1,\ldots,x_k) = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k} \quad (15.25)$$

Dengan mengambil nilai log terhadap sisi sebelah kanan dan kiri Persamaan 15.29, masing-masing parameter dapat diinterpretasikan sebagai

$$log(E(y|x_1,\ldots,x_k)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$
$$\%\Delta E(y|x) \approx (100\beta_j)\Delta x_j \quad (15.26)$$

Dengan kata lain, koefisien regresi akan diinterpretasikan sebagai persentase perubahan variabel dependen akibat perubahan 1 unit variabel independen.

Model Persamaan 15.26 adalah bersifat nonlinear sementara distribusi dari variabel dependen (y) adalah nonnormal (yakni Poisson Distribution). Dengan demikian, diperlukan suatu teknik khusus untuk melakukan estimasi terhadap parameter model itu. Di sini digunakan *Quasi Maximum Likelihood Estimation* (QMLE). Kita tidak akan menguraikan bagaimana estimasi dilakukan karena sangat kompleks dan berada di luar pembahasan, lihat Wooldridge (2005) Bab 19.

Contoh 15.4

Kita akan menggunakan data dari file Olympics.dta (Hill, Griffiiths, dan Lim, 2017). File ini berisi data jumlah medali (medaltot) yang diperoleh negara-negara peserta Olimpiade tahun 1996 beserta log dari GDP (lgdp) dan log Populasi (lpop). Jumlah medali adalah suatu bilangan integer, sehingga model regresi Poisson juga dapat digunakan. Spesifikasi regresi adalah medaltot fungsi dari nilai log PDB dan log populasi. Perintah estimasi diberikan sebagai **poisson medaltot lgdp lpop.**


```
Poisson regression                              Number of obs     =        192
                                                LR chi2(2)        =    1857.87
                                                Prob > chi2       =     0.0000
Log likelihood = -619.22586                     Pseudo R2         =     0.6000
```

medaltot	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lgdp	.5348365	.0217721	24.57	0.000	.4921639	.5775091
lpop	.1707622	.0290882	5.87	0.000	.1137503	.2277741
_cons	-1.201575	.1024046	-11.73	0.000	-1.402284	-1.000865

TABEL 15.12 Regresi Possion dengan Variabel Dependen Medaltot

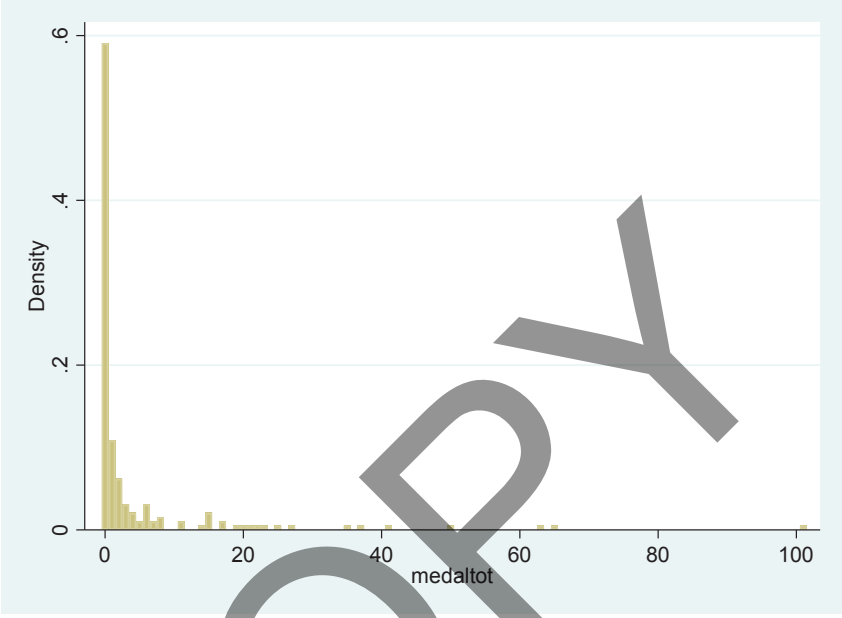
Sebagaimana pada Tabel 15.12, hasil regresi menunjukkan koefisien yang positif (dan signifikan) dari variabel lgdp dan lpop. Negara-negara besar dan berpenghasilan tinggi akan cenderung memperoleh banyak medali dalam Olimpiade. Mungkin setelah melihat data, pembaca menduga OLS dapat juga digunakan untuk melakukan estimasi medaltot. Hasil estimasinya sebagai suatu perbandingan disajikan pada Tabel 15.13.

Source	SS	df	MS		
				Number of obs	= 192
				F(2, 189)	= 41.52
Model	8457.74724	2	4228.87362	Prob > F	= 0.0000
Residual	19247.7319	189	101.839851	R-squared	= 0.3053
				Adj R-squared	= 0.2979
Total	27705.4792	191	145.054865	Root MSE	= 10.092

medaltot	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lgdp	2.945877	.4710654	6.25	0.000	2.016655	3.875098
lpop	-.1723301	.5139971	-0.34	0.738	-1.186238	.8415781
_cons	-2.076863	1.016495	-2.04	0.042	-4.081996	-.0717302

TABEL 15.13. Regresi Possion dengan Variabel Dependen Medaltot

Di sini hasil regresi dan interpretasi yang diperoleh memiliki perbedaan yang signifikan. Sebagai contoh, koefisien lpop menjadi negatif, tetapi tidak signifikan. Kita dapat melihat pola sebaran medali dengan perintah **histogram medaltot, discrete density.** Dapat dilihat dari Gambar 15.3 bahwa distribusi sangat jauh dari normal. Perolehan medali hanya terkonsentrasi pada segelintir negara (sekitar 20-30) dari 198 yang mengikuti Olimpiade. Tentu saja, estimasi OLS (yang mengasumsikan distribusi normal) akan memberikan hasil yang bias.



GAMBAR 15.3 Histogram Sebaran Medali

CENSORED REGRESSION

Model *censored regression* dilakukan ketika, karena satu hal, kita harus membatasi nilai yang dapat diambil oleh suatu variabel dependen. Sebagai contoh, dalam penelitian yang bersifat survei terhadap variabel pengeluaran per bulan di mana variabel ini akan bersifat kategoris. Tentu saja, kuesioner tidak akan mencantumkan setiap pilihan jumlah pengeluaran yang mungkin, dan praktek yang umum dilakukan adalah membuat batas atas dan batas bawah.

Batas atas terjadi apabila dalam kuesioner terdapat pilihan lebih dari 5 juta per bulan (*right censoring*) dan di bawah 1 juta

per bulan (*left censoring*). Ketika seorang responden memilih opsi ini kita tidak akan mengetahui dengan akurat pengeluaran yang sebenarnya, ia mungkin memiliki pengeluaran sebesar Rp5,5 juta tetapi bisa juga Rp100 juta. Tentu saja, implikasi analisis dari dua data semacam ini sangat berbeda tetapi kita telah mengabaikannya.

Suatu (right) *censored regression* dapat dimodelkan sebagai

$$Y_i = \beta_0 + \sum_{i=1}^{K} \beta_i x_i + u_i; u_i \sim N(0, \sigma^2)$$
$$w_i = min(y_i, c_i) \quad (15.27)$$

di mana c_i adalah batas atas. Dengan demikian, nilai variabel y adalah mana yang lebih kecil y_i atau c_i.

Parameter regresi dapat diestimasi dengan menggunakan teknik MLE, di mana observasi yang disensor diharapkan memiliki probabilitas sebagai berikut:

$$f(w|x_i, c_i) = 1 - \Phi[(c_i - \boldsymbol{x_i\beta})/\sigma]; w = c_i$$
$$= \frac{1}{\sigma}\varphi[(w - \boldsymbol{x_i\beta})/\sigma]; w < c_i \quad (15.28)$$

Teknik estimasi ini dikenal juga dengan nama Tobit.

Perlu diperhatikan juga bahwa estimator Tobit bukanlah estimator linear; kita tidak dapat menginterpretasikan koefisien-koefisien tersebut layaknya OLS. Besaran dampaknya akan sangat tergantung pada titik evaluasi dan akan terdiri dari dua komponen (yang dikenal dengan nama McDonald and Moffit decomposition) sebagai berikut

$$\frac{\partial E(y|x)}{\partial x} = Prob(y > 0)\frac{\partial E(y|x, y > 0)}{\partial x} + (y|x, y > 0)\frac{\partial E(y > 0)}{\partial x}$$

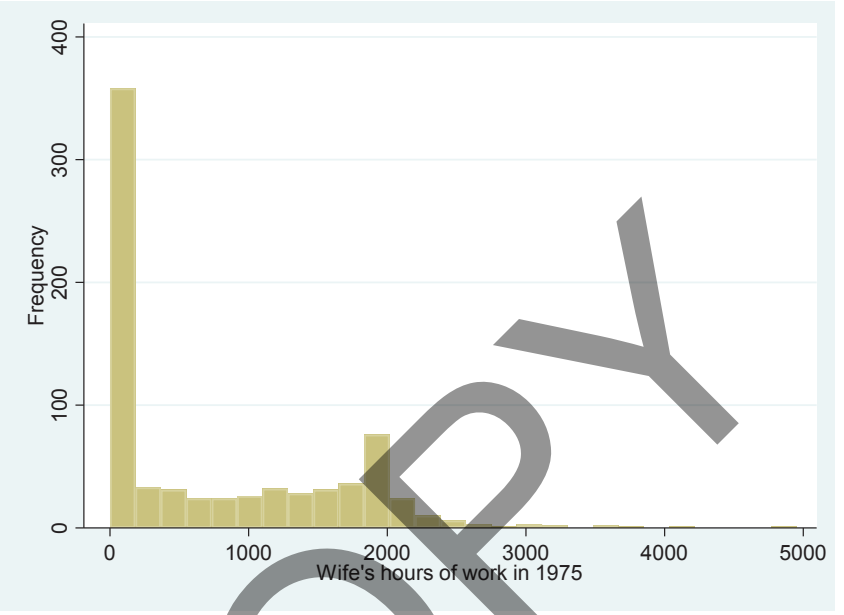
(15.29)

Term pertama di sebelah kanan Persamaan 15.29 adalah dampak marjinal dari perubahan variabel dependen (akibat perubahan variabel penjelas) terhadap variabel yang terobservasi ($y > 0$). Sedangkan term kedua adalah dampak dari perubahan variabel dependen akibat pergeseran perilaku (menjadi $y > 0$) variabel yang tidak terobservasi ($y = 0$).

Contoh 15.5

Dalam contoh ini kita akan menggunakan data dari studi Mroz (1987) mengenai partisipasi kerja para wanita yang telah menikah. File yang digunakan adalah mroz.dta, yang berisi respons dari 753 responden. Data yang tersedia adalah jumlah jam kerja (hours), pendidikan (dalam tahun, educ), pengalaman kerja (dalam tahun, exper), usia (dalam tahun, age), dan jumlah anak di bawah usia 6 tahun (kidsl6). Untuk mengetahui apakah model regresi yang akan diestimasi memerlukan perlakuan/treatment khusus; yaitu censored regression, sehingga kita perlu terlebih dahulu memahami pola data (sebaran jam/hours) yang dilakukan dengan perintah **histogram hours, frequency** (lihat Gambar 15.4).

Dari Gambar 15.4, terlihat bahwa 325 responden adalah tidak bekerja (hours = 0). Jumlah ini adalah 43,16% dari seluruh observasi, sehingga situasi ini akan mendistorsi estimasi; jika dilakukan tanpa treatment (misalnya, OLS). Selanjutnya kita dapat melihat karakter



GAMBAR 15.4. Histogram Sebaran Jam Kerja (Hours), Wanita yang Telah Menikah

Variable	Obs	Mean	Std. Dev.	Min	Max
educ	325	11.79692	2.181995	5	17
exper	325	7.461538	6.918567	0	45
age	325	43.28308	8.467796	30	60
kidsl6	325	.3661538	.6368995	0	3

Variable	Obs	Mean	Std. Dev.	Min	Max
hours	428	1302.93	776.2744	12	4950
educ	428	12.65888	2.285376	5	17
exper	428	13.03738	8.055923	0	38
age	428	41.97196	7.721084	30	60
kidsl6	428	.1401869	.3919231	0	2

TABEL 15.14. Statistik Deskriptif untuk Wanita Tidak Bekerja (atas) versus Bekerja (bawah)

```
Tobit regression                                  Number of obs     =        753
                                                  Uncensored        =        428
Limits: lower = 0                                 Left-censored     =        325
        upper = +inf                              Right-censored    =          0

                                                  LR chi2(4)        =     255.50
                                                  Prob > chi2       =     0.0000
Log likelihood = -3827.1433                       Pseudo R2         =     0.0323
```

hours	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	73.29082	20.47454	3.58	0.000	33.0965	113.4851
exper	80.53514	6.287792	12.81	0.000	68.19135	92.87893
age	-60.76768	6.888176	-8.82	0.000	-74.29011	-47.24526
kidsl6	-918.9165	111.6604	-8.23	0.000	-1138.121	-699.7119
_cons	1349.876	386.2982	3.49	0.001	591.5204	2108.233
var(e.hours)	1285263	95371.19			1111034	1486814

TABEL 15.15. Regresi Tobit dengan Variabel Dependen Hours, Left Limit

data dari mereka yang bekerja (hours > 0) dengan yang tidak bekerja sebagai berikut: **summarize educ exper age kidsl6 if (hours==0) dan summarize hours educ exper age kidsl6 if (hours>0)**.

Jika kita membandingkan Tabel 15.14, panel atas dan yang bawah terlihat bahwa wanita yang tidak bekerja memiliki rata-rata pengalaman kerja yang lebih rendah; pendidikan yang sedikit lebih rendah, usia yang sedikit lebih tua, dan memiliki lebih banyak anak di bawah usia 6 tahun. Perbedaan karakter ini akan mempengaruhi hasil regresi. Mengingat terdapat porsi yang substansial pada observasi hours = 0; yang berada di sebelah kiri rata-rata hours (atau mediannya: 288), kita akan menggunakan regresi Tobit dengan *left hand side censored (left limit)*. Hal ini dilakukan dengan perintah **tobit hours educ exper age kidsl6, ll**.

Regresi Tobit pada Tabel 15.15 memberitahukan kepada kita bahwa terdapat 325 observasi yang ter"sensor" (hours = 0).

Source	SS	df	MS	Number of obs	=	753
				F(4, 748)	=	64.71
Model	146771295	4	36692823.7	Prob > F	=	0.0000
Residual	424138429	748	567029.985	R-squared	=	0.2571
				Adj R-squared	=	0.2531
Total	570909724	752	759188.463	Root MSE	=	753.01

hours	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	27.08568	12.23989	2.21	0.027	3.057054	51.1143
exper	48.03981	3.641804	13.19	0.000	40.89044	55.18919
age	-31.30782	3.96099	-7.90	0.000	-39.0838	-23.53184
kidsl6	-447.8547	58.41252	-7.67	0.000	-562.5267	-333.1827
_cons	1335.306	235.6487	5.67	0.000	872.6945	1797.918

Source	SS	df	MS	Number of obs	=	428
				F(4, 423)	=	15.12
Model	32193987.4	4	8048496.86	Prob > F	=	0.0000
Residual	225117032	423	532191.566	R-squared	=	0.1251
				Adj R-squared	=	0.1168
Total	257311020	427	602601.92	Root MSE	=	729.51

hours	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	-16.46211	15.58083	-1.06	0.291	-47.0876	14.16339
exper	33.93637	5.009185	6.77	0.000	24.09038	43.78237
age	-17.10821	5.457674	-3.13	0.002	-27.83575	-6.380677
kidsl6	-305.309	96.44904	-3.17	0.002	-494.8881	-115.7299
_cons	1829.746	292.5356	6.25	0.000	1254.741	2404.75

TABEL 15.16. Regresi OLS dengan Variabel Dependen Hours, Seluruh Sampel (Atas) dan Hanya Jika Hours > 0 (Bawah)

Seluruh variabel penjelas memiliki tingkat signifikansi yang tinggi dan sesuai dengan hipotesis (*common logic*). Variabel pendidikan dan pengalaman akan meningkatkan jumlah jam kerja; sedangkan usia dan jumlah anak di bawah 6 tahun akan menekan jumlah jam kerja. Nilai LR $\chi 2$ adalah 255,50 yang signifikan secara statistik (*p*

value = 0,000). Kita dapat membandingkan hasil yang diperoleh ini dengan teknik OLS (diberikan pada Tabel 15.16).

Dapat dilihat dari Tabel 15.16 bahwa jika melakukan estimasi OLS atas seluruh data sampel (panel atas), kita akan memperoleh koefisien dengan tanda aljabar dan tingkat signifikansi statistik yang dapat dikatakan serupa. Namun, ada perbedaan yang sangat signifikan pada besaran koefisien. Sebaliknya, jika regresi dilakukan hanya untuk observasi di mana hours > 0, maka terdapat perubahan yang signifikan pada besaran koefisien; bahkan ada yang berubah tanda aljabarnya yakni educ, sehingga estimator OLS dapat dikatakan bias dan tidak konsisten.

Kita dapat menghitung dampak total (*marginal effect*) untuk nilai educ = 12,29, exper = 10,63, age = 42,25, dan kidsl6 = 1. Perintah STATA adalah **margins, dydx(educ) at(educ=12.29 exper=10.63 age=42.5 kidsl6=1) predict(ystar(0,.))**. Hasilnya disajikan pada Tabel 15.17.


```
Conditional marginal effects                    Number of obs     =        753
Model VCE    : OIM

Expression   : E(hours*|hours>0), predict(ystar(0,.))
dy/dx w.r.t. : educ
at           : educ            =       12.29
               exper           =       10.63
               age             =        42.5
               kidsl6          =           1
```

	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
educ	26.66452	7.563667	3.53	0.000	11.84	41.48903

TABEL 15.17. Dampak Marjinal Total Jam Kerja

Sedangkan dampaknya terhadap data yang terobservasi (hours > 0) dapat diberikan dengan perintah berikut **margins, dydx(educ) at(educ=12.29 exper=10.63 age=42.5 kidsl6=1) predict(e(0,.))**. Hasilnya disajikan pada Tabel 15.18. Dengan demikian, dampak perubahan jam kerja sebagai akibat dari pergeseran perilaku terobservasi adalah 26,66 – 21,57 = 5,09%.

```
Conditional marginal effects                    Number of obs     =        753
Model VCE    : OIM

Expression   : E(hours|hours>0), predict(e(0,.))
dy/dx w.r.t. : educ
at           : educ            =      12.29
               exper           =      10.63
               age             =       42.5
               kidsl6          =          1
```

	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
educ	21.57398	5.999968	3.60	0.000	9.814258	33.3337

TABEL 15.18. Dampak Marjinal Total Jam Kerja untuk Data yang Terobservasi (hours > 0)

PROSEDUR HECKMAN

Salah satu permasalahan ekonometrika yang sering ditemui di lapangan adalah bias seleksi (*selection bias*). Hal ini terutama disebabkan oleh desain sampling atau penarikan sampel yang tidak random. Apabila sampling dilakukan dengan teknik random (atau variannya), maka setiap unsur populasi memiliki probabilitas yang "relatif" sama untuk terpilih menjadi unsur sampel. Karena

satu hal dan lainnya (biasanya pertimbangan biaya dan feasibilitas teknis), *random sampling* tidak digunakan. Jika ini terjadi, maka terdapat suatu karakter yang sistematis dari keberadaan elemen/unsur pada suatu sampel. Karakter sistematis ini berbentuk kedekatan dengan periset, kemudahan akuisisi, dan karakter yang terobservasi. Terjadinya karakter ini akan menyebabkan bias yang disebut *selection bias*[7].

Suatu ilustrasi spesifik yang dapat diberikan adalah penelitian tentang ketenagakerjaan (Hill, Griffiths, and Guay, 2017 hal. 723). Peneliti ingin melakukan studi mengenai faktor-faktor yang mempengaruhi penghasilan yang diterima oleh pekerja wanita. Sampel yang representatif harus terdiri dari wanita yang bekerja dan wanita yang tidak bekerja (misalnya, memilih sebagai ibu rumah tangga). Tentu saja, wanita ibu rumah tangga akan mengisi respons nol pada pertanyaan berapa jumlah penghasilan. Strategi empiris yang sering ditempuh adalah mengeluarkan wanita yang tidak bekerja dari sampel, sehingga sampel menjadi tidak random. Hal ini tentu menimbulkan bias karena pasti ada karakteristik wanita bekerja (misalnya, berpendidikan tinggi) pada ibu rumah tangga yang seharusnya dapat memberikan informasi atas dampak pendidikan terhadap gaji pada wanita. Mengeluarkan ibu rumah tangga karena penghasilan tidak terobservasi adalah karakter sistematis yang menyebabkan sampel menjadi bias.

Salah satu solusi terhadap permasalahan ini diajukan oleh Heckman (1979); yang disebut juga regresi Heckitt. Prosedur ini terdiri dari dua tahap, yakni

---

[7] Bias yang mirip dan sering ditemui dalam literatur keuangan adalah *survivorship bias*. Misalnya, dalam penelitian mengenai kinerja saham, kita biasanya hanya akan menggunakan saham-saham yang listed sepanjang periode sampel. Prosedur Heckman juga sering digunakan untuk mengatasi masalah missing observations (Little dan Rubin, 2019).

a. Estimasi probit atas kategori (z_i) seorang wanita menjadi pekerja versus tidak (disebut persamaan seleksi; *selection equation*)

$$z_i^* = \delta + W\gamma + u_i \quad (15.30)$$

di mana z_i* adalah variabel laten; sebagai indikator z_i = 1 jika z_i* > 0 dan z_i = 0 jika tidak. *W* adalah vektor regressor yang dapat menjelaskan z_i. Persamaan ini diestimasi dengan seluruh sampel (*N*) yang terdiri dari wanita bekerja dan ibu rumah tangga.

b. Regresi Least Squares penghasilan terhadap satu set variabel yang relevan dan satu variabel yang bernama *Inverse Mills Ratio* = IMR. IMR diperoleh dari estimasi probit untuk mengkoreksi adanya bias pada sampel yang diakibatkan tidak dimasukkannya ibu rumah tangga.

$$E(y_i)(y_i|z_i^* > 0) = \alpha + X\beta + \theta\lambda_i + v_i \quad (15.31)$$

di mana *X* adalah vektor variabel penjelas dari regresi fokus studi, dan λi adalah IMR, yang dapat diestimasi dari tahap regresi yaitu tahap pertama melalui formulasi berikut

$$\lambda_i = \frac{\phi(\delta + W\hat{\gamma})}{\Phi(\delta + W\hat{\gamma})} \quad (15.32)$$

di mana ϕ adalah fungsi densitas distribusi normal standar dan ϕ adalah fungsi kumulatif distribusi normal standar. Estimasi pada model Persamaan 15.31 hanya dilakukan kepada *n* (yakni unsur sampel di mana z_i = 1).

Contoh 15.6

Kita kembali menggunakan data Mroz (1987). Dari data tersebut sebanyak 428 wanita adalah bekerja dan 325 tidak bekerja. Selanjutnya, kita ingin memodelkan gaji (*wage*) sebagai fungsi dari Pendidikan (educ) dan pengalaman (exper) dengan melakukan prosedur Heckman untuk menangani bias yang timbul akibat proporsi wanita yang tidak bekerja yang substansial. Regresi tahap pertama (probit) dilakukan dengan memodelkan pilihan wanita untuk bekerja (versus tidak; labor force participation-lfp) sebagai fungsi dari usia (age), pendidikan (educ), variabel dummy memiliki anak (kids), dan tarif pajak marjinal (mtr). Ingat variabel dummy memiliki anak diperoleh dengan perintah **generate kids = (kidsl6+kids618>0)**.

Regresi tahap pertama dilakukan dengan perintah berikut **probit lfp age educ kids mtr**, dan hasilnya dapat dilihat pada Tabel 15.19. Nilai variabel IMR diperoleh dengan perintah **predict w, xb** dan diikuti dengan **generate imr = normalden(w)/normal(w)**. Di sini normalden() adalah syntax STATA untuk menghitung nilai PDF variabel distribusi standar normal dan normal() adalah untuk menghitung nilai CDF variabel distribusi standar normal.

Setelah memperoleh variabel IMR, kita dapat melakukan regresi tahap kedua dengan perintah **regress lwage educ exper imr**. Hasilnya disajikan pada Tabel 15.20. Dapat dilihat di sini bahwa seluruh koefisien variabel penjelas (educ dan exper) memiliki tanda aljabar yang sesuai hipotesis dan signifikan secara statistik. Kita dapat mengetahui adanya bias akibat dari sampling (selection bias) dengan melihat signifikansi koefisien variabel IMR. Di sini kita memperoleh koefisien IMR sebesar –0,866 dengan *p* value sebesar 0,008.

```
Iteration 0:   log likelihood =  -514.8732
Iteration 1:   log likelihood = -494.21411
Iteration 2:   log likelihood = -494.14614
Iteration 3:   log likelihood = -494.14614

Probit regression                                 Number of obs     =        753
                                                  LR chi2(4)        =      41.45
                                                  Prob > chi2       =     0.0000
Log likelihood = -494.14614                       Pseudo R2         =     0.0403
```

lfp	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	-.0206155	.0070447	-2.93	0.003	-.0344229	-.0068082
educ	.0837753	.023205	3.61	0.000	.0382943	.1292563
kids	-.3138848	.1237108	-2.54	0.011	-.5563535	-.0714162
mtr	-1.393853	.6165751	-2.26	0.024	-2.602318	-.1853878
_cons	1.192296	.7205439	1.65	0.098	-.2199443	2.604536

TABEL 15.19. Regresi Tahap Pertama dari Prosedur Heckman

Source	SS	df	MS	Number of obs	=	428
				F(3, 424)	=	27.37
Model	36.2307253	3	12.0769084	Prob > F	=	0.0000
Residual	187.096716	424	.441265841	R-squared	=	0.1622
				Adj R-squared	=	0.1563
Total	223.327442	427	.523015086	Root MSE	=	.66428

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0584579	.0238495	2.45	0.015	.01158	.1053358
exper	.0163202	.0039984	4.08	0.000	.0084612	.0241793
imr	-.8664386	.3269855	-2.65	0.008	-1.509153	-.2237242
_cons	.8105417	.4944723	1.64	0.102	-.1613804	1.782464

TABEL 15.20. Regresi Tahap Kedua dari Prosedur Heckman

Untuk memperoleh gambaran kualitatif yang lebih jauh, kita lakukan regresi dengan menggunakan sampel wanita bekerja saja. Hal ini dilakukan dengan perintah **reg lfp educ exper if (hours>0)**. Hasilnya disajikan pada Tabel 15.21. Dapat dilihat di sini bahwa koefisien educ pada regresi tahap kedua hanya setengah dari regresi OLS. Standard error pada koefisien tahap kedua tidak layak karena adanya variabel IMR yang merupakan hasil estimasi.

Source	SS	df	MS	Number of obs	=	428
				F(2, 425)	=	37.02
Model	33.132458	2	16.566229	Prob > F	=	0.0000
Residual	190.194984	425	.447517609	R-squared	=	0.1484
				Adj R-squared	=	0.1444
Total	223.327442	427	.523015086	Root MSE	=	.66897

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.1094888	.0141672	7.73	0.000	.0816423	.1373353
exper	.0156736	.0040191	3.90	0.000	.0077738	.0235733
_cons	-.4001744	.1903682	-2.10	0.036	-.7743548	-.0259939

TABEL 15.21. Regresi OLS Hanya pada Sampel Wanita Bekerja (hours > 0)

Kedua isu tersebut adalah serupa dengan permasalahan estimasi variabel endogen (lihat Bab 10 di Jilid 1). Untuk mengatasi kedua isu tersebut dilakukanlah estimasi secara serentak dengan perintah hackman two step estimator atau maximum likelihood. Untuk contoh ini kita dapat lakukan dengan perintah **heckman lwage educ exper, select(lfp=age educ kids mtr) twostep**[8].

[8] Alternatifnya pembaca dapat mencoba syntax **heckman lwage educ exper, select(lwage educ kids mtr)** untuk estimator heckman maximum likelihood.

Heckman selection model -- two-step estimates (regression model with sample selection)

Number of obs	= 753
Selected	= 428
Nonselected	= 325
Wald chi2(2)	= 19.53
Prob > chi2	= 0.0001

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lwage						
educ	.0584579	.0296354	1.97	0.049	.0003737	.1165422
exper	.0163202	.0042022	3.88	0.000	.0080842	.0245563
_cons	.8105418	.6107985	1.33	0.185	-.3866012	2.007685
lfp						
age	-.0206155	.0070447	-2.93	0.003	-.0344229	-.0068082
educ	.0837753	.023205	3.61	0.000	.0382943	.1292563
kids	-.3138848	.1237108	-2.54	0.011	-.5563535	-.0714162
mtr	-1.393853	.6165751	-2.26	0.024	-2.602318	-.1853878
_cons	1.192296	.7205439	1.65	0.098	-.2199443	2.604536
/mills						
lambda	-.8664387	.3992843	-2.17	0.030	-1.649022	-.0838559
rho	-0.92910					
sigma	.93255927					

TABEL 15.22. Regresi Heckman Two Step Estimator

Bab 16

Structural Equation Modelling (SEM)

Mayoritas pembahasan sebelumnya adalah tentang model linear yang terdiri dari satu variabel dependen dan beberapa variabel penjelas. Di bab ini akan dibahas suatu model statistik yang sangat fleksibel, yang dapat menganalisis hubungan antara beberapa variabel dependen dan beberapa variabel penjelas secara simultan. Teknik statistik ini dikenal dengan nama *structural equation modelling* (SEM). Sebenarnya, regresi linear itu sendiri adalah bentuk spesifik (varian) dari SEM, di mana variabel dependen sama dengan satu. Pembaca yang berminat mengeksplorasi teknik ini dapat merujuk pada Acock (2013) dan Schumacer dan Lomax (2016).

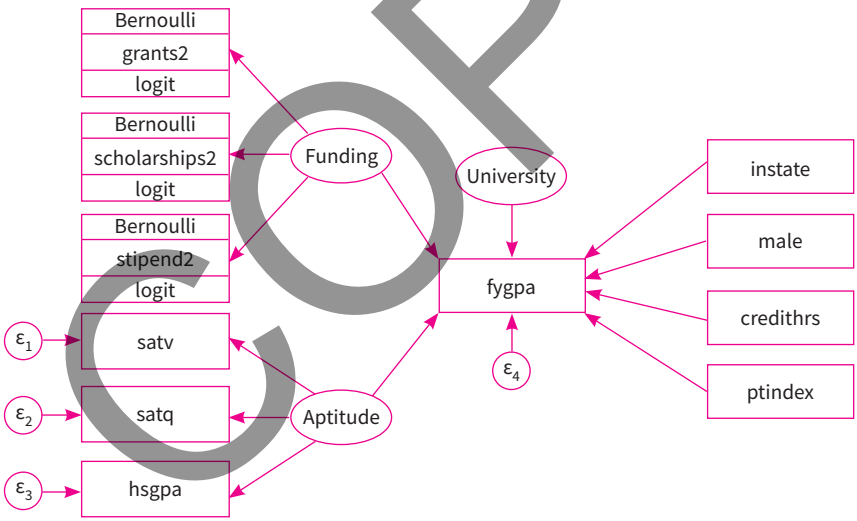
Menurut Huber (2014), SEM adalah desain empiris yang bersifat integratif yang mencakup cara berpikir, cara menulis model, dan cara mengestimasi. Desain empiris ini terdiri atas dua komponen yakni (a) model pengukuran (*confirmatory factor analysis*) dan model struktural (*path analysis*). SEM adalah analisis multivariate yang memungkinkan estimasi atas suatu desain empiris berupa sistem persamaan di mana di dalamnya terdapat variabel yang terobservasi (*observed*), tidak terobservasi (*latent* variabel), dan residual.

REPRESENTASI DAN PATH MODEL

Analisis yang dilakukan dengan menggunakan SEM bersifat struktural yang digambarkan dalam suatu diagram path yang menunjukkan arah pemikiran. Di sini kita akan mempelajari terlebih dahulu beberapa terminologi penting dalam analisis SEM. Gambar 16.1 memberikan ilustrasi tentang model SEM yang cukup lengkap, yang dapat digunakan untuk menjelaskan beberapa konsep penting.

Variabel yang terobservasi adalah variabel yang skalanya (nominal/ordinal atau numeris) dapat diperoleh dari pengamatan/

observasi. Dalam SEM variabel ini digambarkan dengan bentuk kotak. Contoh variabel ini adalah satq, instate, dan credithrs. Lawan dari variabel yang terobservasi adalah variabel laten, yaitu variabel yang tidak dapat diamati secara langsung. Dalam SEM, variabel ini digambarkan dengan bentuk oval, dan kita dapat menduga nilai variabel ini melalui besaran variabel-variabel terobservasi yang menyusunnya. Contoh variabel laten adalah Aptitude di mana besaran variabel ini diukur dengan menggunakan instrumen seperti faktor analisis dengan error tertentu (yang ditunjukkan oleh g1, g2, dan g3).



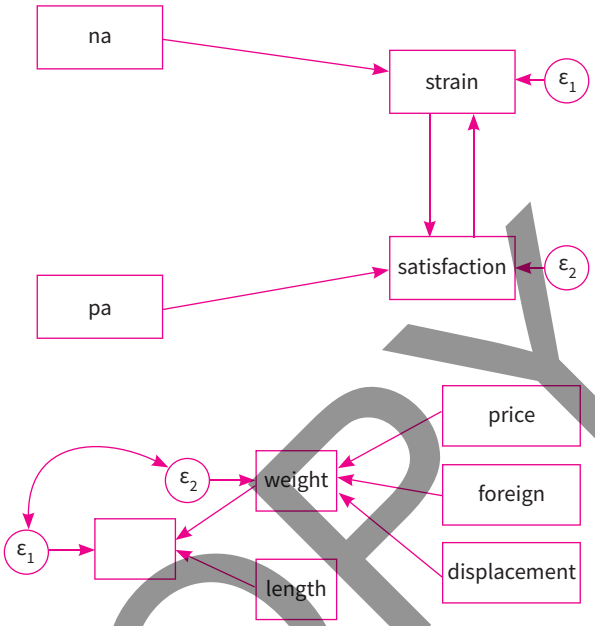
GAMBAR 16.1. Suatu Ilustrasi SEM yang Lengkap (sumber Huber, 2014)

Terkait dengan cara penentuan nilainya, variabel dapat dikategorikan menjadi dua jenis. Jenis yang pertama adalah variabel endogen: yakni variabel yang nilainya ditentukan di dalam sistem. Variabel endogen ditunjukkan dalam SEM oleh tanda panah yang menuju

variabel tersebut. Jenis yang kedua adalah variabel eksogen: yakni variabel yang nilainya ditentukan di luar sistem. Dalam ilustrasi sebelumnya, variabel fygpa adalah variabel yang bersifat endogen sedangkan variabel seperti instate dan credithrs adalah variabel yang bersifat eksogen.

Jalur (path) berguna untuk menunjukkan arah hubungan antar dua variabel, yang ditunjukkan oleh tanda panah. Jalur dapat bersifat satu arah (*one way*) dan bersifat dua arah (*two way*; *simultaneous*). Jika bersifat satu arah, maka pengaruhnya diasumsikan terlihat jelas di mana satu variabel dianggap sebagai penyebab (*cause*) sedangkan variabel lainnya (tujuan arah panah) dianggap sebagai akibat (*response*). Jika bersifat dua arah, maka kedua variabel (biasanya variabel laten) tersebut nilainya ditentukan secara simultan atau terdapat *feedback loop*; yaitu jika satu variabel berubah yang satunya lagi juga berubah. Sifat hubungan dua arah ini sering disebut sebagai covariance/kovarians.

Variabel ini kemudian dibentuk berdasarkan kombinasi dari karakter observasi dan penentuan nilainya. Dengan demikian, ada variabel (a) observed-exogen, (b) observed-endogen, (c) latent-exogen, dan (d) latent-endogen. Selanjutnya berdasarkan bentuk pemodelan SEM (yang ditunjukkan oleh arah tanda panah; lihat Gambar 13.2), kita akan memiliki (a) model nonrecursive, yaitu model yang tidak memiliki komponen *feedback loop*: correlated error, dan (b) recursive model; yaitu model yang memiliki komponen *feedback loop*. Akhirnya, apabila dalam struktur model tersebut ada suatu/sekelompok variabel endogen yang memodifikasi (memperbesar-amplify atau mereduksi-dampen) dampak hubungan antar dua variabel, maka dikatakan model itu memiliki variabel mediasi.



GAMBAR 16.2 Non-Recursive; Left Panel dan (b) Recursive Model; Right Panel (sumber Huber, 2014)

ESTIMASI DAN EVALUASI

Analisis SEM dilakukan pertama kali dengan mengidentifikasi seluruh komponen yang dibutuhkan (a) variabel-variabel (observed-exogen, observed-endogen, latent-exogen, dan latent-endogen), (b) arah hubungan (one way dan covariance), (c) keberadaan variabel mediasi, dan (d) error pengukuran. Dalam notasi matematis hal ini secara ringkas dapat diformulasikan sebagai berikut

$$Y = BY + \Gamma X + \alpha + \zeta \qquad (16.1)$$

di mana Y adalah matriks variabel endogen (terobservasi dan laten); X adalah matriks variabel eksogen (terobservasi dan laten), B adalah matriks koefisien variabel endogen, Γ adalah matriks koefisien variabel eksogen, α adalah vektor intercept, dan ζ adalah vektor residual. Dengan asumsi (a) model telah dispesifikasi dengan benar, (b) jumlah sampel cukup besar (rule of thumb dari Huber, 2014 adalah lebih dari 100), dan (c) residual dari model memiliki distribusi multivariat normal, sehingga dapat digunakan teknik Maximum Likelihood untuk mengestimasi koefisien-koefisien variabel endogen serta eksogen ketika mengidentifikasi SEM secara lengkap. STATA menyediakan SEM Builder sebagai instrumen visual dalam membangun model.

Setelah melakukan estimasi terhadap semua parameter yang diperlukan, langkah terakhir adalah melakukan evaluasi apakah model SEM itu. Evaluasi dilakukan dengan menghitung sekelompok statistik untuk membandingkan model SEM yang telah dibangun (disebut dengan *specified model*) dengan dua model acuan yakni (a) saturated model: yaitu model di mana seluruh variabel diasumsikan berkorelasi satu dengan yang lain, dan (b) baseline model: yaitu model di mana korelasi hanya terjadi antara variabel endogen dan eksogen. Beberapa statistik yang sering digunakan (Acock, 2013) adalah

a. Koefisien determinasi

$$R^2 = 1 - \frac{\det(\hat{\Psi})}{\det(\hat{\Sigma})} \qquad (16.2)$$

di mana det adalah operator determinan dari matriks $\hat{\Psi}$ (varians-kovarians residual) dan $\hat{\Sigma}$ (matriks varians variabel eksogen).

Nilai R^2 akan berada antara 0 dan 1; dengan angka semakin mendekati 1 menunjukkan goodness of fit atau kelaikan suai yang semakin baik.

b. Root Mean Square Error of Approximation (RMSEA)

$$RMSEA = \sqrt{\frac{(\chi^2_{ms} - df_{ms})}{(N-1)df_{ms}}} \quad (16.3)$$

di mana $\chi^2_{ms} = 2(\log L_s - \log L_m)$; L_s dan L_m masing-masing adalah nilai loglikelihood dari saturated model dan specified model serta $df_{ms} = df_s - df_m$ adalah selisih dari degree of freedom saturated model dan specified model. Nilai RMSEA dikonversi ke dalam *exact level of significance* (p value); sementara goodness of fit dikatakan memadai jika p value model lebih besar dari 0,05.

c. Comparative Fit Index

$$CFI = 1 - \frac{\chi^2_{ms} - df_{ms}}{\chi^2_{bs} - df_{bs}} \quad (16.4)$$

di mana $\chi^2_{ms} = 2(\log L_s - \log L_b)$; L_b dan df_{bs} adalah (masing-masing) nilai *loglikelihood* dan *degree of freedom* dari baseline model. *Goodness of fit* dianggap memadai jika CFI lebih besar dari 0,90.

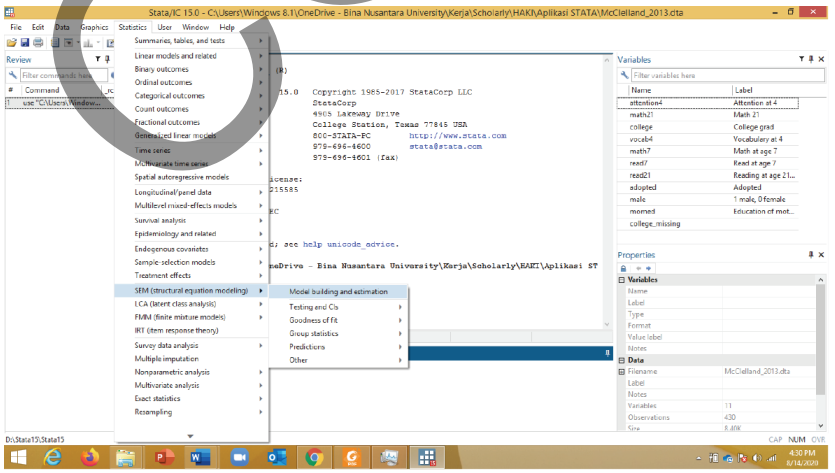
d. Tucker Lewis Index

$$TLI = \frac{\left(\chi^2_{bs}/df_{bs}\right) - \left(\chi^2_{ms}/df_{ms}\right)}{\left(\chi^2_{bs}/df_{bs}\right) - 1} \quad (16.5)$$

di mana index *s*, *b*, dan *m* adalah merujuk ke (masing-masing) saturated, baseline, dan specified model. Goodness of fit ditunjukkan oleh angka TLI yang lebih besar dari 0,95.

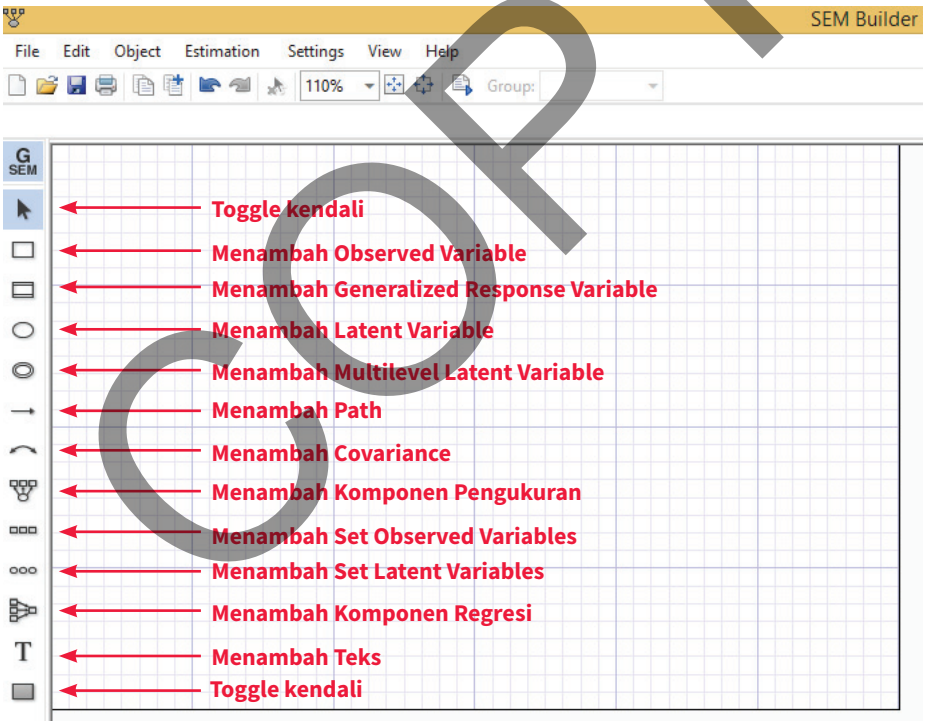
PENGGUNAAN SEM BUILDER

STATA memiliki interface yang sangat friendly untuk membuat suatu model SEM disebut dengan *SEM Builder*. Dengan menggunakan SEM Builder, identifikasi, hubungan, serta estimasi di antara variabel-variabel yang digunakan akan dilakukan secara virtual dengan cara *drawing* serta *drag and drop*. Setelah melakukan estimasi, pada output window dan log window akan terlihat bagaimana syntax yang diperlukan. SEM adalah prosedur yang rumit atau *complicated* sehingga tidak disarankan untuk melakukan konstruksi dan estimasi melalui syntax. Menu ini dapat diakses melalui ribbon menu **Statistiks/SEM(Structural Equation Modelling)/Model Building and estimates** (lihat Gambar 16.3).



GAMBAR 16.3. Akses Menu SEM Builder

Setelah mengklik Model Building and estimates kita akan masuk ke dalam suatu layar menu konstruksi dan estimasi SEM (**SEM Builder**) lihat Gambar 16.4. Panel icon sebelah kiri adalah pilihan yang dapat diambil dalam konstruksi SEM. Icon-icon yang akan sangat sering digunakan adalah penambahan variabel (terobservasi dan laten), pembuatan jalur (path), dan covariance/kovarians untuk SEM. Untuk model *Generalized SEM*, icon generalized response dan multilevel latent variable juga akan sering digunakan.



GAMBAR 16.4. Menu SEM Builder dan Beberapa Icon Terpilih

Icon-icon tersebut digunakan seperti kita membuat suatu diagram (sketsa) alur. Pertama kita tentukan berapa jumlah serta nama-nama variabel: terobservasi dan laten serta arah hubungan kausalitas serta kovarians. Setelah sketsa selesai, dapat dilakukan estimasi dengan mengklik menu **estimation** yang ada pada ribbon menu di SEM Builder. Kita akan tunjukkan bagaimana hal ini dilakukan dengan suatu ilustrasi.

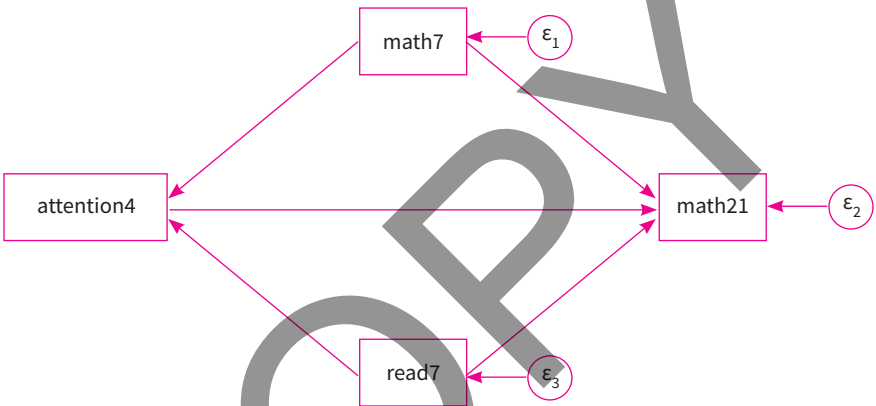
Contoh 16.1

Kita akan menggunakan data dari studi McClelland et al (2013), di mana nama filenya adalah McClelland_2013.dta. Studi ini memodelkan dan mengestimasi hubungan antara berbagai kemampuan kognitif masa kecil dengan kemampuan kognitif tersebut diusia dewasa. File ini mengandung berbagai variabel yang relevan dari 430 responden.

Pertama, kita akan membangun terlebih dahulu modelnya dengan menggunakan SEM builder (perintah STATA **sembuilder**). Misalnya, kita memodelkan bahwa kemampuan memperhatikan pada usia 4 tahun (attention4) akan berdampak terhadap kemampuan kognitif (seperti matematika) di usia 21 tahun (math21). Terdapat dua variabel mediasi (yang akan memperkuat pengaruh kemampuan memperhatikan) yakni kemampuan membaca di usia 7 tahun (read7) dan matematika di usia 7 tahun (math7). Variabel attention4 bersifat eksogen sedangkan read7, math7, dan math21 adalah variabel endogen terobservasi.

Pada submenu sembuilder kita dapat terlebih dahulu membuat empat kotak (sebagai icon variabel eksogen) yakni attention4, read7, math7, dan math21. Terdapat hubungan sebab akibat antara

attention4 dan read7, math7, dan langsung math21. Selanjutnya, read7 dan math7 akan mempengaruhi math21. Hubungan-hubungan ini dapat dibuat dengan icon path yang berbentuk tanda panah. Kita sebut saja pola hubungan ini (Gambar 16.5) sebagai model I untuk membedakan ilustrasi-ilustrasi yang akan kita buat belakangan.



GAMBAR 16.5 SEM Builder Model I

Kita dapat langsung melakukan estimasi dari SEM Builder dengan mengklik estimation yang ada pada ribbon menu di sisi atas. Kita pilih metode maximum likelihood with missing values (mlmv) dan pada tab reporting kita tandai display standardized coefficients and values. Setelah melakukan eksekusi perintah; kita akan memperoleh dua tipe output yakni (a) tabel (Tabel 16.1) dan (b) diagram (Gambar 16.6)[1].

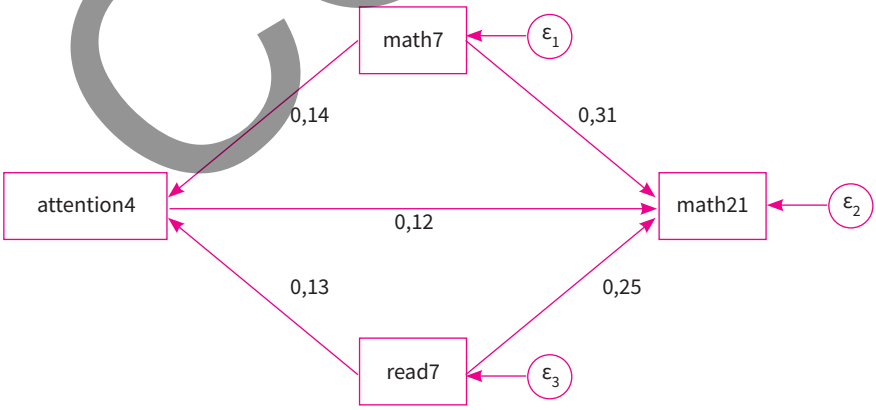
[1] Apa yang dilakukan di sini adalah prosedur basic saja. Pembaca dapat merujuk ke Acock (2013) untuk melakukan refinement terhadap diagram.

```
Structural equation model                       Number of obs     =        430
Estimation method  = mlmv
Log likelihood     =  -4246.557
```

Standardized	Coef.	OIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
Structural						
math7						
attention4	.141458	.0486307	2.91	0.004	.0461437	.2367723
_cons	3.04888	.3344304	9.12	0.000	2.393408	3.704351
math21						
math7	.3075685	.0481426	6.39	0.000	.2132108	.4019262
read7	.2520422	.0489132	5.15	0.000	.156174	.3479104
attention4	.1171187	.0467622	2.50	0.012	.0254664	.208771
_cons	1.380531	.361878	3.81	0.000	.6712636	2.089799
read7						
attention4	.1289838	.0491968	2.62	0.009	.0325598	.2254077
_cons	3.163475	.3383925	9.35	0.000	2.500238	3.826712
var(e.math7)	.9799896	.0137584			.9533913	1.00733
var(e.math21)	.8075246	.0341705			.7432537	.8773531
var(e.read7)	.9833632	.0126912			.9588009	1.008555

```
LR test of model vs. saturated: chi2(1)   =     27.56, Prob > chi2 = 0.0000
```

TABEL 16.1 Hasil Estimasi SEM Model 1



GAMBAR 16.6. Hasil Estimasi Model 1

Seberapa baik pemodelan ini? Kita dapat menggunakan beberapa statistik evaluasi: goodness of fit. Pertama adalah koefisien determinasi, di mana kita dapat memperoleh statistik ini dengan perintah **estat eqgof**. Tabel 16.2 menunjukkan model yang telah kita buat (attention4 hanya sebagai satu-satunya variabel eksogen) tidak terlalu bagus (overall R^2= 0,051). Kemampuan attention4 dalam menjelaskan varians variabel endogen paling tinggi adalah pada math21, yang mencapai 0,192).

Equation-level goodness of fit

depvars	Variance fitted	Variance predicted	Variance residual	R-squared	mc	mc2
observed						
math7	7.621122	.1525014	7.46862	.0200104	.141458	.0200104
math21	6.920939	1.33211	5.588828	.1924754	.4387202	.1924754
read7	64.70388	1.076467	63.62742	.0166368	.1289838	.0166368
overall				.0515245		

mc = correlation between depvar and its prediction
mc2 = mc^2 is the Bentler-Raykov squared multiple correlation coefficient

TABEL 16.2. Koefisien Determinasi Model 1

Kesimpulan kualitatif yang serupa akan kita temui jika kita menggunakan statistik goodness of fit yang lain, seperti *p* value dari RMSEA, CFI, dan TLI. Perintah STATA untuk menghitung statistik-statistik ini adalah **estat gof, stats(all)**. Seperti disajikan pada Tabel 16.3, nilai *p* value RMSEA adalah 0,000 lebih kecil dari acuan 0,05. Sedangkan nilai CFI dan TLI adalah (masing-masing) 0,787 dan –0,276; yang mengindikasikan kurang memadainya *goodness of fit*.

Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(1)	27.561	model vs. saturated
p > chi2	0.000	
chi2_bs(6)	130.877	baseline vs. saturated
p > chi2	0.000	
Population error		
RMSEA	0.249	Root mean squared error of approximation
90% CI, lower bound	0.174	
upper bound	0.332	
pclose	0.000	Probability RMSEA <= 0.05
Information criteria		
AIC	8515.114	Akaike's information criterion
BIC	8559.816	Bayesian information criterion
Baseline comparison		
CFI	0.787	Comparative fit index
TLI	-0.276	Tucker-Lewis index
Size of residuals		
CD	0.052	Coefficient of determination

Note: SRMR is not reported because of missing values.

TABEL 16.3 Beberapa Statistik Goodness of Fit Model 1

PENGEMBANGAN MODEL SEM

STATA menyediakan tools untuk memberikan saran demi meningkatkan kinerja model yang dimiliki saat ini. Dalam model ini kita memiliki degree of freedom sebesar 1, sehingga nilai Chi Square harus turun di bawah 3,84 agar kita dapat memperoleh *p* value RMSEA yang lebih besar dari 0,100. Dengan menggunakan perintah **estat mindices**, STATA akan memberikan berbagai alternatif path di antara variabel yang akan menurunkan nilai Chi Square (lihat Tabel 16.4).

Modification indices

	MI	df	P>MI	EPC	Standard EPC
Structural					
math7					
math21	26.885	1	0.00	1.091552	1.040202
read7	26.885	1	0.00	.0899778	.2621748
read7					
math7	26.885	1	0.00	.7665476	.2630773
math21	26.885	1	0.00	2.615316	.8553455
cov(e.math7,e.read7)	26.885	1	0.00	5.725053	.2626257

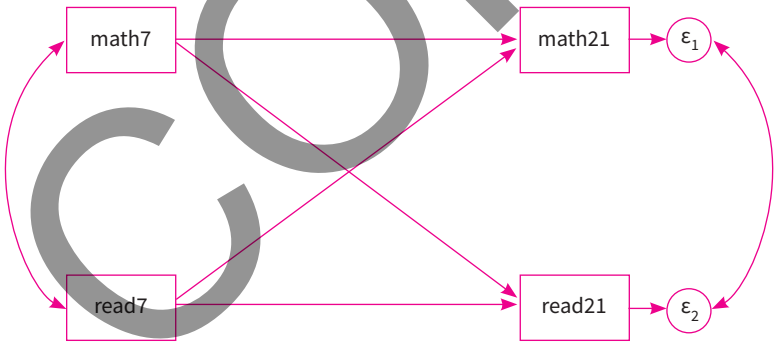
EPC = expected parameter change

TABEL 16.4 Modified Indices; Alternatif Pengembangan Model I

Terdapat 5 alternatif pengembangan yakni (a) path dari math7 ke read7, (b) path dari read7 ke math7, (c) path dari math21 ke read7, (d) path dari math21 ke math7, dan (e) korelasi antara residual math7 dan read7. Tentu saja, pilihan c dan d adalah tidak masuk akal: di mana nilai masa akan datang tidak dapat mempengaruhi nilai masa lalu. Dengan demikian, pilihan yang tersedia adalah (a), (b), dan (e). Misalnya, kita memilih poin e, sehingga kita akan memiliki model *correlated residual*. Suatu catatan perlu diberikan di sini mengingat kita hanya memiliki satu degree of freedom; dengan memasukkan kovarians atau covariance di antara residual read7 dan math7, sehingga model menjadi perfect fit. Akibatnya, evaluasi goodness of fit tidak akan memiliki arti.

CROSS LAGGED PANEL DESIGN

Mengingat data kita juga memiliki karakteristik longitudinal: yaitu variabel diukur dari responden yang sama pada waktu yang berbeda (usia 7 tahun dan usia 21 tahun), maka kita dapat membuat *cross lagged panel design*. Misalkan kita ingin mengetahui dampak dari read7 dan math7 terhadap (masing-masing) read21 dan math21. Diagram SEM bagi desain ini disajikan pada Gambar 16.7 (sebut saja sebagai model II). Hasil estimasi dari model II ditunjukkan pada Tabel 16.5. Dapat dilihat di sini bahwa pencapaian kemampuan kognitif pada usia muda (read7 dan math 7) berpengaruh positif dan signifikan terhadap kemampuan kognitif pada usia dewasa (read21 dan math 21).



GAMBAR 16.7. SEM Builder Model II: Cross Lagged Panel

Selanjutnya dari Tabel 16.6 kita dapat melihat model II memiliki koefisien determinasi (overall) sebesar 0,377. Terdapat perbedaan kemampuan antara read7 dan math7 dalam menjelaskan variasi read21 dan math 21. Tabel 16.7 menyajikan beberapa statistik *goodness of fit* yang terpilih untuk model II. Di sini terlihat bahwa

```
Structural equation model                       Number of obs     =        416
Estimation method  = mlmv
Log likelihood     = -4369.2178
```

Standardized	Coef.	OIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
Structural						
math21						
math7	.3103359	.0469761	6.61	0.000	.2182645	.4024074
read7	.2571059	.0473963	5.42	0.000	.1642109	.3500008
_cons	1.960534	.2806437	6.99	0.000	1.410482	2.510585
read21						
math7	.1102662	.0460747	2.39	0.017	.0199614	.200571
read7	.4931898	.0411844	11.98	0.000	.4124698	.5739098
_cons	6.343015	.4020612	15.78	0.000	5.55499	7.131041
mean(math7)	3.887371	.146786	26.48	0.000	3.599676	4.175067
mean(read7)	3.919858	.1489989	26.31	0.000	3.627825	4.21189
var(e.math21)	.7941389	.0379745			.7230916	.872167
var(e.read21)	.7149913	.0404764			.6399019	.7988922
var(math7)	1	.			.	.
var(read7)	1	.			.	.
cov(e.math21,e.read21)	.1735063	.0515344	3.37	0.001	.0725007	.2745119
cov(math7,read7)	.2722754	.0464696	5.86	0.000	.1811967	.3633541

```
LR test of model vs. saturated: chi2(0)    =      0.00, Prob > chi2 =      .
```

TABEL 16.5 Hasil Estimasi Model II

```
Equation-level goodness of fit
```

depvars	Variance fitted	Variance predicted	Variance residual	R-squared	mc	mc2
observed						
math21	7.173452	1.476734	5.696717	.2058611	.4537192	.2058611
read21	71.36437	20.33946	51.02491	.2850087	.533862	.2850087
overall				.3767532		

```
mc  = correlation between depvar and its prediction
mc2 = mc^2 is the Bentler-Raykov squared multiple correlation coefficient
```

TABEL 16.6 Koefisien Determinasi Model II

hasil yang "mungkin" terlalu bagus, yaitu *p* value RMSEA berada jauh di atas 0,05 sedangkan CFI dan TFI masing-masing berada di angka 1,00. Sama dengan model I; kita memiliki permasalahan terbatasnya *degree of freedom* yang di sini 2 variabel eksogen digunakan untuk mengestimasi 6 parameter: read7-read21, read7-math21, math7-read21, math7-math21, covarians residual math21 read21 dan covarians read7 math7.

Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(0)	0.000	model vs. saturated
p > chi2	.	
chi2_bs(5)	211.573	baseline vs. saturated
p > chi2	0.000	
Population error		
RMSEA	0.000	Root mean squared error of approximation
90% CI, lower bound	0.000	
upper bound	0.000	
pclose	1.000	Probability RMSEA <= 0.05
Information criteria		
AIC	8766.436	Akaike's information criterion
BIC	8822.865	Bayesian information criterion
Baseline comparison		
CFI	1.000	Comparative fit index
TLI	1.000	Tucker-Lewis index
Size of residuals		
CD	0.377	Coefficient of determination

Note: SRMR is not reported because of missing values.

TABEL 16.7. Beberapa Statistik Goodness of Fit Model II

Bab 17

Pemrograman dan Simulasi

Pada bagian awal telah dijelaskan bahwa salah satu keunggulan utama STATA adalah adanya komunitas yang sangat aktif dalam melakukan pengembangan serta sharing kepada para pengguna *routine-routine* baru yang sangat berguna untuk penelitian. Pengembangan suatu routine yang dilakukan melalui pembuatan program bisa sangat kompleks dan menghabiskan banyak sumber daya. Di sisi lain, hal ini diperlukan mengingat routine yang ada kurang memadai untuk menangani desain empiris yang kita miliki. Jika kita cukup rajin melakukan eksplorasi di Internet dan fleksibel dalam penyesuaian desain; peluang kita untuk memperoleh *community developed routine* cukup besar.

Namun demikian, tidaklah realistis untuk menganggap bahwa komunitas STATA yang ada di Internet akan dapat menjawab setiap tantangan dalam desain empiris yang kita miliki. Karena itu, kita harus mengembangkan kemampuan membuat suatu program yang dapat memenuhi kebutuhan kita. Secara umum, kita melakukan pemrograman untuk tiga keperluan berikut:

a. Substitusi, atas routine atau command yang secara default tidak disediakan oleh STATA atau belum dibuat oleh komunitas.
b. Efisiensi, terutama meringkas aktivitas yang bersifat berulang (*repetitive*) baik untuk suatu proyek maupun untuk penggunaan di masa yang akan datang.
c. Simulasi, untuk melakukan evaluasi atas karakteristik estimator terutama pada sampling terbatas

STATA menyediakan kemampuan yang sangat ekstensif bagi pemrograman di mana pengguna dapat memilih melakukannya dalam bahasa yang high level; yaitu melalui ado file atau bahasa yang lumayan dasar yakni matriks; dan melalui mata. Bahasan di bab

ini hanya bersifat memperkenalkan. Sementara untuk uraian yang ekstensif penulis menyarankan Baum (2016) untuk pemrograman dan Carsey and Harden (2013) serta Adkins dan Gade (2013) untuk simulasi.

BEBERAPA UNSUR PENTING

Suatu program adalah kumpulan instruksi-instruksi yang dirangkai secara terstruktur dan sistematis sehingga dapat dipahami serta dijalankan (*executable*) oleh komputer demi mencapai tujuan tertentu. Kata-kata kunci di sini adalah instruksi, terstruktur, sistematis, *executable*, dan tujuan tertentu.

Seperti pemrograman pada aplikasi lain, pemrograman pada STATA juga terdiri dari empat elemen/unsur penting, yakni

a. Deklarasi
b. Looping
c. Branching
d. Struktur

Deklarasi adalah tahap pertama dalam pemrograman. Di sini akan dibuat pernyataan yang menegaskan status objek-objek yang digunakan dalam suatu proyek program. Objek-objek yang umum adalah variabel, observasi, dan *macro*. Variabel dan observasi telah cukup jelas dipahami, tetapi dalam pemrograman STATA diperlukan deklarasi agar variabel-observasi tersebut dikenali dengan baik sehingga dapat digunakan dalam instruksi. Suatu teknik deklarasi yang sering digunakan dalam STATA adalah macro. **Macro** adalah teknik deklarasi yang fleksibel bagi suatu variabel, observasi, skalar, matriks, bahkan perintah sederhana yang dapat bersifat lokal (hanya

digunakan sebagai subkomponen untuk program tertentu), atau global (dapat digunakan sebagai komponen dari program lain secara berulang). Macro dapat digunakan untuk keperluan permanen atau temporer.

Looping adalah suatu set perintah untuk mengulangi aktivitas selama suatu kondisi berlaku. Terdapat tiga tipe looping di dalam STATA yakni **forvalue**, **foreach**, dan **while**. Perintah forvalue biasa digunakan untuk suatu pengulangan yang harmonis yang dilakukan selama titik waktu tertentu (misalnya, estimasi regresi dengan lag). Sedangkan *foreach* digunakan untuk suatu set tertentu (misalnya, elemen/unsur sampel cross section: negara), dan *while* untuk ekspresi (misalnya, selama kondisi tertentu terpenuhi). Juga mungkin menggunakannya sebagai substitusi pada kondisi tertentu di antara perintah-perintah looping.

Branching adalah suatu set perintah yang dilakukan atas dasar terpenuhinya kondisi tertentu (*conditional*). Hal ini dilakukan dengan syntax if ... else. Setelah mengetikkan if disusul dengan suatu ekspresi yang menunjukkan kondisi yang harus terpenuhi (seperti apakah variabel $X = K$). Setelah ekspresi kondisi kemudian dispesifikasikan perintah yang akan dijalankan jika kondisinya terpenuhi. Term else digunakan sebagai perintah jika kondisi tidak terpenuhi. Selain itu, juga dimungkinkan untuk melakukan branching secara berulang, karena adanya lebih dari dua kondisi; yang berarti dapat lebih banyak implikasi-alternatif perintah.

Program lalu disusun sebagai suatu rangkaian instruksi yang dimulai dari pembuatan prakondisi, biasanya berupa pembersihan lingkungan (dengan perintah clear dan/atau drop). Kemudian dilakukan deklarasi variabel, macro, dan spesifikasi lain yang diperlukan sesuai konteks program (seperti penentuan speed, akses

memori, *seed*, dan tracing atau penelusuran untuk debug). Instruksi-instruksi operasi terhadap variabel dan macro dilakukan sebagai kombinasi perintah plain vanilla; *looping*, dan *branching*. Sangat disarankan untuk memecah suatu program menjadi subroutine-subroutine demi memudahkan review dan penyelesaian jika terjadi masalah (*crash*).

Terdapat dua cara membuat program di dalam STATA, di mana yang pertama dengan perintah program dan yang kedua dengan ado file. Terdapat dua tipe program yakni R Class yang digunakan untuk aktivitas pemrograman yang mencakup penyimpanan dan penggunaan (storing dan retrieving) berbagai output estimasi-inferensial statistik, dan E Class yang digunakan untuk aktivitas selain itu. Sedangkan ado file digunakan untuk pemrograman yang kompleks (*elaborated*) yang terdiri dari beberapa subprogram.

ILUSTRASI: PEMBUATAN PROGRAM SEDERHANA: ARDL

Kita kembali ke contoh 12.1 di mana kita sedang mengestimasi berbagai model ADDL pilihan lag (1 sampai dengan 3) untuk variabel dependen dep_rate dan variabel penjelas berbagai lag dari variabel dep_rate dan inf. Seperti telah dijelaskan, secara keseluruhan akan terdapat 9 model regresi: ADL(1,1), ADL(1,2), ADL(1,3), ADL(2,1), ADL(2,2), ADL(2,3), ADL(3,1), ADL(3,2), dan ADL(3,3). Kita dapat melakukan regresi tersebut secara manual dan berulang. Tetapi di sini akan diilustrasikan suatu program sederhana untuk melakukan estimasi atas ke-9 regresi tersebut secara sekaligus. Programnya diberikan sebagai berikut

```
. capt prog drop ardl_reg
. prog define ardl_reg
  1. forvalues i=1(1)3{
  2. forvalues j=1(1)3{
  3. reg dep_rate L(`i'/3).dep_rate L(`j'/3).inf
  4. }
  5. }
  6. end
```

Dengan mengetik ardl_reg pada command window, akan diperoleh hasil seperti tersaji pada Tabel 17.1. Di sini penulis hanya menampilkan 1 regresi saja untuk efisiensi; dan pembaca dapat melihat sendiri keseluruhan regresi itu. STATA telah memiliki *built in routine* untuk melakukan estimasi dan evaluasi atas regresi ADL (perintah ardl); sementara itu, Kripfganz dan Scheneider (2016) telah membuat pengembangan atas routine tersebut[1]. Jadi, latihan dilakukan secara lebih mendalam untuk tujuan ilustrasi.

EKSPERIMEN MONTE CARLO

Monte Carlo adalah suatu eksperimen statistik yang dilakukan untuk mengevaluasi karakter sampel kecil dari berbagai *competing estimator* untuk suatu masalah estimasi (Kennedy, 2003 hal. 24). Kemampuan evaluasi dari eksperimen Monte Carlo akan diperoleh jika kita dapat mengendalikan lingkungan statistik di mana estimator-estimator tersebut didapat (Adkins dan Gade, 2013). Di buku ini, banyak estimator yang dibahas memiliki karakter: koefisien, matriks

[1] Kripfganz dan Scheneider (2016), membuat pengembangan dengan nama command default STATA: ardl. Penulis menyarankan untuk dilakukan secara hati-hati mengingat jika dilakukan replace, maka command default STATA akan digantikan dengan *command community user*.

```
. ardl_reg

      Source |       SS           df       MS      Number of obs   =       117
-------------+----------------------------------   F(6, 110)       =    932.40
       Model |  86.740332          6  14.456722    Prob > F        =    0.0000
    Residual |  1.70554246       110  .015504931   R-squared       =    0.9807
-------------+----------------------------------   Adj R-squared   =    0.9797
       Total |  88.4458744       116  .762464435   Root MSE        =    .12452

------------------------------------------------------------------------------
    dep_rate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    dep_rate |
         L1. |   1.253675   .0940176    13.33   0.000     1.067354    1.439996
         L2. |   -.038387   .1539717    -0.25   0.804    -.3435227    .2667488
         L3. |  -.2613838   .0901271    -2.90   0.005    -.4399946    -.082773
             |
         inf |
         L1. |   .0034821   .0216007     0.16   0.872    -.0393254    .0462896
         L2. |   .0375714   .0335525     1.12   0.265    -.0289217    .1040646
         L3. |  -.0191103   .0221095    -0.86   0.389    -.0629262    .0247056
             |
       _cons |   .2009357   .0935097     2.15   0.034     .0156213     .38625
------------------------------------------------------------------------------

      Source |       SS           df       MS      Number of obs   =       117
-------------+----------------------------------   F(5, 111)       =   1128.77
       Model |  86.739929          5  17.3479858   Prob > F        =    0.0000
    Residual |  1.70594537       111  .015368877   R-squared       =    0.9807
-------------+----------------------------------   Adj R-squared   =    0.9798
       Total |  88.4458744       116  .762464435   Root MSE        =    .12397

------------------------------------------------------------------------------
    dep_rate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    dep_rate |
         L1. |   1.256992   .0913351    13.76   0.000     1.076005    1.437978
         L2. |  -.0435567   .1499327    -0.29   0.772    -.3406583     .253545
         L3. |  -.2594362    .088921    -2.92   0.004     -.435639   -.0832334
             |
         inf |
         L2. |   .0418755   .0202303     2.07   0.041     .0017878    .0819633
         L3. |  -.0202645    .020826    -0.97   0.333    -.0615326    .0210036
             |
       _cons |   .2019051   .0929058     2.17   0.032      .017806    .3860042
------------------------------------------------------------------------------

      Source |       SS           df       MS      Number of obs   =       117
-------------+----------------------------------   F(4, 112)       =   1369.73
       Model |  86.6740788          4  21.6685197   Prob > F        =    0.0000
    Residual |  1.77179559       112  .015819603   R-squared       =    0.9800
-------------+----------------------------------   Adj R-squared   =    0.9793
       Total |  88.4458744       116  .762464435   Root MSE        =    .12578
```

(… tidak dilanjutkan)

TABEL 17.1. Output Program ardl_reg

varians-kovarians, statistik uji, dan statistik kritis yang sesuai teori hanya apabila serangkaian asumsi-asumsinya dipenuhi.

Suatu contoh estimator yang sederhana adalah estimator OLS. Seperti telah dijelaskan pada Bab 7 (di Jilid 1), OLS dapat dianggap sebagai *Best Linear Unbiased Estimator* (BLUE) jika asumsi-asumsi seperti (a) model telah dispesifikasi dengan benar (mana variabel yang endogen, mana variabel yang eksogen, dan tidak ada korelasi antara variabel eksogen dan residual), (b) tidak ada serial korelasi, (c) homokedastis, (d) korelasi yang rendah di antara variabel eksogen, dan (e) residual terdistribusi normal sudah terpenuhi. Ketika kita menghadapi realita bahwa data yang diperoleh sangat mungkin tidak memenuhi satu atau lebih asumsi-asumsi tersebut, yang menyebabkan penggunaan OLS menjadi bias, tidak konsisten, dan/atau tidak efisien.

Pelanggaran-pelanggaran terhadap asumsi ini mungkin dapat diatasi dengan desain estimasi alternatif yang diperoleh dari kajian analitis. Solusi Cohcrane-Orcutt terhadap isu serial korelasi adalah salah satu contohnya. Namun demikian, banyak sekali dalam pekerjaan di lapangan solusi analitis tidak dapat diperoleh. Jika hal ini ditemui, salah satu metode yang dilakukan adalah menggunakan Monte Carlo.

Monte Carlo akan memungkinkan kita untuk menspesifikasi kembali desain ekonometrika yang dimiliki dengan suatu dataset yang dapat "dikendalikan". Sebagai ilustrasi, Granger dan Newbold (1974) membuat teori mengenai regresi palsu (*spurious regression*). Jika kita melakukan regresi antara dua variabel yang memiliki karakteristik *nonstationary* (mengalami unit root), maka kita akan sangat mungkin memperoleh hasil regresi yang koefisien variabel independennya (x) adalah signifikan. Seperti dijelaskan pada Bab

13, Granger dan Newbold membuktikan hal ini dengan membuat 2 series variabel sintetis nonstationary dan melakukan regresi di antara keduanya. Regresi itu diulang (simulasi), katakan 10.000 kali, dengan sampel (misalnya, 100) tertentu dan mencatat statistik *t* yang diperoleh pada setiap simulasi. Di sini dapat ditunjukkan rejection/penolakan terhadap hipotesis null: $\beta_X = 0$ adalah sangat mungkin mencapai lebih dari 90%; tergantung pada jumlah simulasi dan sampel. Dengan demikian, regresi OLS akan memberikan hasil yang *misleading*; *false positive*, yaitu adanya dampak dari suatu variabel independen terhadap variabel dependen yang sebenarnya tidak ada.

Perhatikan di sini bahwa Monte Carlo memberikan kita keleluasaan untuk menspesifikasikan bentuk stasionaritas yang ada. Pola yang umum adalah *pure random walk*, *random walk with a drift,* dan *random walk with a trend*. Masing-masing pola tersebut tentu akan memiliki implikasi sendiri. Jika diinginkan, kita bahkan dapat memasukkan pola *nonstationary* dengan derajat integrasi lebih dari 1 atau berupa pecahan (*fraction integrated*). Tujuan yang ingin dicapai juga dapat dikembangkan, tidak hanya dengan melihat berapa proporsi *false positive; rejection of true null hypothesis* tetapi juga dampaknya terhadap R^2, uji *F*, atau bias-inkonsistensi pada koefisien itu sendiri.

Di samping menunjukkan implikasi atau identifikasi karakter suatu estimator akibat adanya fitur *data generating process* (DGP) tertentu (dalam kasus ini adalah *nonstationary*); Monte Carlo juga digunakan untuk membuat solusi alternatif (akibat adanya fitur tersebut). Sebagai contoh, dari Bab 13 kita mengetahui bahwa distribusi *t* standar tidak lagi dapat digunakan untuk memverifikasi hipotesis null adanya unit root. Dengan simulasi Monte Carlo, kita

dapat menunjukkan bahwa nilai distribusi *t* pada percentile 1%, 5% dan 10% terlalu rendah untuk digunakan sebagai statistik kritis dalam spesifikasi regresi unit root (situasi yang dikenal dengan nama *undersized*). Simulasi Monte Carlo juga dapat memberikan tabel statistik kritis yang harus digunakan agar kita dapat mengambil kesimpulan dengan *p* value yang lebih benar. Hal ini dilakukan dengan membuat distribusi *t* sintetis dari koefisien regresi unit root yang memiliki nilai percentile tertentu (yang biasanya akan mengikuti α, yakni 1%, 5%, dan 10%).

Kedua ilustrasi tersebut akan dibuat secara lebih konkret, yaitu dalam bentuk *coding*. Namun, diharapkan pembaca telah memperoleh gambaran atas kesimpulan uraian. Eksperimen Monte Carlo tidak hanya memungkinkan kita untuk mengidentifikasi implikasi-permasalahan dari adanya fitur DGP tertentu pada variabel-variabel yang digunakan terhadap estimator, tetapi juga menawarkan solusi.

ILUSTRASI 1: SPURIOUS REGRESSION

Ilustrasi Spurious Regression akan menggunakan fitur program yang dibuat dengan mengadopsi Baum (2013b). Coding lengkap dari program tersebut disajikan berikut ini:

```
.capt prog drop irwd
. prog irwd, rclass
1. version 12
2. drop _all
3. set obs $numobs
4. g double x=0 in 1
```

```
  5. g double y=0 in 1
  6. replace x=x[_n-1]+$trcoef*2+rnormal() in 2/1
  7. replace y=y[_n-1]+$trcoef*0.5+rnormal() in 2/1
  8. reg y x
  9. ret sca b=_b[x]
 10. ret sca se=_se[x]
 11. ret sca t=_b[x]/_se[x]
 12. ret sca r2=abs(return(t))>invttail($numobs-2,0.025)
 13. end
.glo numsim=10000
. glo numobs=1000
. glo trcoef=0
.simulate b=r(b) se=r(se) t=r(t) reject=r(r2), reps($numsim)
saving(spurious, replace) nolegend nodots:irwd
```


Programnya bernama irwd dan memiliki karakter sebagai rclass. Tipe program rclass sangat diperlukan karena program akan menggunakan output dari estimasi dan regresi. Baris ketiga mendeklarasikan jumlah observasi yang bersifat global; yang diberi tanda $ di depan nama (numobs). Baris ke-4 hingga 7 mendefinisikan pola random walk yang dispesifikasikan berikut ini:

$$y_t = y_{t-1} + 0.5 + u_t$$
$$x_t = x_{t-1} + 2 + v_t \qquad (17.1)$$

di mana baik u_t maupun v_t diasumsikan terdistribusi secara standar normal. Baris ke-8 merupakan perintah melakukan regresi OLS,

$$y_t = \beta_0 + \beta_1 x_t + e_t \qquad (17.2)$$

Baris ke-9 hingga 12 merupakan perintah untuk menyimpan hasil dari regresi yang menjadi fokus studi yakni berupa koefisien variabel *x* (vektor *b*), standar error koefisien variabel *x* (vektor *x*), statistik *t* (vektor *t*), dan perintah kategoris (memberikan angka 1 jika nilai absolut dari *t* lebih nilai *t* kritis (derajat kebebasan: jumlah observasi-2 dan α = 0,025%; vektor r2).

Setelah membuat program (berbentuk perintah bernama irwd), akan dilakukan simulasi. Akan tetapi, sebelumnya kita harus mendeskripsikan beberapa parameter simulasi yang diperlukan. Pertama kita masukkan *seed*[2] yang merupakan angka di mana simulasi akan mulai bekerja. Kita tentukan jumlah simulasi (numsim) sebanyak 1.000 dan jumlah observasi (yang digunakan untuk setiap regresi; numobs) sebesar 10.000. Kita akan melakukan eksperimen dengan model random walk murni, sehingga nilai trcoef ditentukan sebesar 0.

Perintah melakukan simulasi (*simulate*) diberikan dengan memasukkan variabel-variabel yang menjadi fokus studi: *b*, *se*, *t*, dan reject. Perhatikan bahwa dalam perintah simulasi kita mengubah return *r*2 menjadi nama reject. Parameter lain yang diperlukan adalah jumlah simulasi; yang sudah ditentukan sebelumnya melalui parameter numsim (= 1.000). Selanjutnya, hasil simulasi akan disimpan dalam file STATA (ekstensi .dta) yang bernama spurious, dengan perintah replace untuk menggantikan isi yang ada saat ini. Perintah terakhir (mean) adalah menghitung rata-rata dari sejumlah variabel yang akan dilaporkan dalam bentuk tabel: mean standard deviation dan 95% confidence interval.

[2] Dalam simulasi ini, seed tidak terlalu diperlukan sehingga kita bisa mengisi dengan sembarang angka. Untuk beberapa jenis simulasi, algorithma yang digunakan mungkin tidak *convex* (suatu fungsi yang patah-patah) sehingga angka di mana simulasi dimulai akan mempengaruhi hasilnya.

```
Mean estimation                   Number of obs    =      1000
```

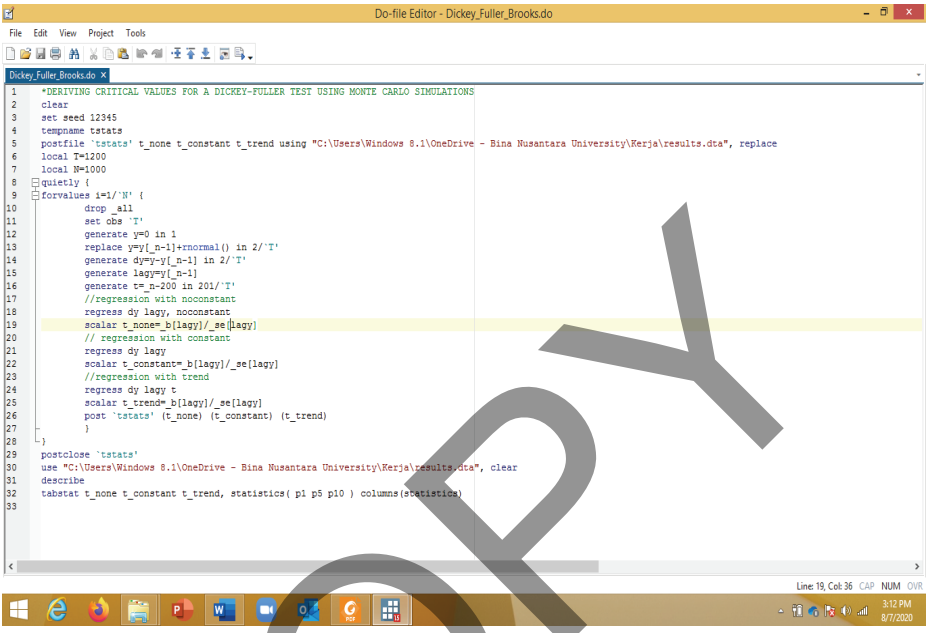
	Mean	Std. Err.	[95% Conf.	Interval]
b	-.0305688	.019545	-.0689226	.0077851
se	.0097193	.0001883	.0093496	.0100889
t	-1.210499	2.435943	-5.990652	3.569653
reject	.979	.0045365	.9700979	.9879021

TABEL 17.2. Hasil Simulasi Monte Carlo Spurious Regression

Hasil simulasi Monte Carlo spurious regression disajikan pada Tabel 17.1. Dapat dilihat di sini bahwa meskipun rata-rata *t* koefisien variabel *x* dari 1.000 regresi sebesar –1,210, tetapi memiliki deviasi standar yang sangat besar yaitu 2,435. Statistik reject yang sebesar 0,979 menunjukkan bahwa dari 1.000 regresi yang disimulasikan, sebanyak 979 di antaranya memiliki *t* yang secara absolut lebih besar dari *t* uji hipotesis 2 arah dengan α = 2,5%. Kita akan sangat mungkin (probabilitas 97,9%) mengatakan adanya hubungan antara variabel *x* dan *y* melalui regresi OLS yang sebenarnya tidak ada. Kedua variabel tersebut adalah hasil simulasi; tidak ada substansi sama sekali.

ILUSTRASI 2: DICKEY FULLER

Apabila program yang dibuat cukup kompleks yaitu lebih dari 20 baris, maka sangat disarankan melakukan *coding* melalui ado.file. Pembuatan do.file editor dapat dimulai dengan mengklik icon do-file editor yang berada di bawah ribbon menu, yaitu sebelah kiri data



GAMBAR 17.1. Syntax Command pada File Dickey_Fuller_Brooks.do

editor. Dalam ilustrasi sekarang kita akan menggunakan program untuk menghasilkan uji Dickey Fuller (Schopohl, Wichmann, dan Brooks, 2019). Program ini dibuat dengan do-file karena cukup kompleks; nama file Dickey_Fuller_Brooks.do.

Rangkaian instruksi pada file program Dickey_Fuller_Brooks.do dapat dilihat pada Gambar 17.1. Pada baris 4 dan 5 ditunjukkan bahwa program menggunakan temporary file bernama tstats yang hasilnya kemudian dimasukkan pada file **results.dta**. Perhatikan di sini bahwa kita harus memasukkan nama dan lokasi (directory) file yang akan digunakan sebagai host temporary file. Selanjutnya pada baris ke-6 dan 7 dideklarasikan jumlah observasi (T) dan jumlah simulasi (N); kedua parameter tersebut ditentukan bersifat lokal.

Baris ke-8 dan ke-27, adalah jantung dari program ini. Instruksi yang diberikan adalah melakukan *looping* sebanyak *N* simulasi. Aktivitas looping ini terdiri atas pembuatan series random dan

```
. do "C:\Users\Windows 8.1\OneDrive - Bina Nusantara University\Kerja\Scholarly\HAKI\Aplikasi STA
> TA\Dickey_Fuller_Brooks.do"

. *DERIVING CRITICAL VALUES FOR A DICKEY-FULLER TEST USING MONTE CARLO SIMULATIONS
. clear

. set seed 12345

. tempname tstats

. postfile `tstats' t_none t_constant t_trend using "C:\Users\Windows 8.1\OneDrive - Bina Nusanta
> ra University\Kerja\results.dta", replace

. local T=1200

. local N=1000

. quietly {

. postclose `tstats'

. use "C:\Users\Windows 8.1\OneDrive - Bina Nusantara University\Kerja\results.dta", clear

. describe

Contains data from C:\Users\Windows 8.1\OneDrive - Bina Nusantara University\Kerja\results.dta
  obs:             1,000
 vars:                 3                          8 Aug 2020 15:48
 size:            12,000
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
t_none          float   %9.0g
t_constant      float   %9.0g
t_trend         float   %9.0g
-------------------------------------------------------------------------------
Sorted by:

. tabstat t_none t_constant t_trend, statistics( p1 p5 p10 ) columns(statistics)

    variable |        p1        p5       p10
-------------+------------------------------
      t_none | -2.500893 -1.787141 -1.527512
  t_constant | -3.299719 -2.872216  -2.54882
     t_trend | -4.101599 -3.393682 -3.111774
----------------------------------------------

.
end of do-file
```

TABEL 17.3. Hasil Running Program Dickey_Fuller_Brooks.ado

deklarasi variabel (baris 11 sampai baris 16). Selanjutnya dilakukan estimasi atas tiga jenis regresi: yaitu pure random walk (baris 18), random walk with a drift (baris 21), dan random walk with a time trend (baris 24). Untuk setiap jenis regresi selanjutnya dihitung statistik *t* untuk koefisien unit root (y_{t-1}): t_none, t_constant, dan t_trend serta disimpan pada temporary file tstats.

Baris 30 hingga 32 memerintahkan untuk menghitung percentile 1, 5, dan 10 untuk statistik *t* koefisien unit root setiap spesifikasi regresi (t_none, t_constant, dan t_trend) dari seluruh simulasi (=1000). Hasil dari eksekusi program disajikan pada Tabel 17.3, yang menunjukkan *t* kritis uji koefisien unit root pada α = 1%, 5%, dan 10% untuk spesifikasi regresi *pure random walk*, *random walk with drift*, dan *random walk with trend*. Jumlah observasi dan simulasi yang dilakukan cukup besar dan dapat disandingkan dengan jumlah sampel asimtotik (T□∞).

Bab 18

Machine Learning dan Ekonometrika Bayesian

Gelombang industri 4.0 adalah suatu era yang ditandai dengan kemajuan pesat-terobosan pada teknologi informasi dan komputasi. Terobosan di bidang teknologi tersebut telah membawa dampak positif terhadap berbagai disiplin ilmu termasuk statistik-ekonometrika yang bermanifestasi dalam bentuk *machine learning* (*statistical learning*)[1]. *Machine learning* (ML) adalah bagian dari *Artificial Intelligence*, yang merupakan bidang ilmu yang mempelajari bagaimana membangun algoritma (suatu alur pemecahan masalah yang logis-matematis yang dilakukan oleh komputer) untuk belajar dari data yang ada dalam memecahkan masalah tertentu. Dalam perkembangannya; ML kemudian bergabung dengan berbagai metode pembelajaran data lainnya (mulai dari yang konvensional seperti matematika, statistik, dan sampling hingga yang kontemporer: *computer-based algorithm*) menjadi disiplin ilmu tersendiri yang dikenal dengan nama *data science* (Athey, 2017).

Adopsi ML oleh statistik dan ekonometrika adalah salah satu topik riset yang saat ini sedang *booming* (Athey (2017) dan Cameron (2019)). Seperti akan dijelaskan nanti, ML memberikan cara pandang baru dalam menyelesaikan berbagai masalah statistika dan ekonometrika. Cara pandang baru tersebut mencakup area-area penting seperti prediksi, inferensial, dan spesifikasi model. Dengan dukungan kemampuan komputasi (*hardware* dan *software*) yang meningkat pesat selama dua dekade belakangan ini serta ketersediaan data secara masif (*big data*), ML memberikan insight

[1] Apa perbedaan Machine Learning dan Statistikal Learning? Keduanya adalah algoritma yang dibangun untuk mengidentifikasi pola yang ada pada data. Khususnya statistikal learning, berangkat dari model-teori statistik. Perbedaan ini sebenarnya masih ambigu; seperti yang dijelaskan oleh Robert Tibshirani; yang secara bercanda menjelaskan perbedaan besaran grant riset di mana machine learning: maksimum USD 1 juta sedangkan statistikal learning: USD 50 ribu (dikutip dari Ahrens, 2019). Dalam bab ini, kita menggunakan machine learning sebagai sinomin statistical learning.

baru dalam implementasi metode-metode yang ada atau *existing*; bahkan dari yang paling standar seperti OLS.

Terobosan di bidang teknologi komputasi juga membuka jalan bagi cabang ekonometrika lain yakni *Bayesian Econometrics*. Berbeda dengan prinsip ekonometrika yang dibahas di buku ini (dikenal sebagai *frequentist econometrics*); ekonometrika Bayesian memulai analisis dengan suatu asumsi mengenai distribusi estimator (*prior distribution*). Datanya akan digunakan untuk meningkatkan kualitas presisi serta memberikan pernyataan probabilistik terhadap estimator. Ekonometrika Bayesian membutuhkan banyak *computing power*; yaitu sesuatu yang saat ini tersedia dengan harga yang relatif murah. Dengan demikian, para periset mulai banyak menggunakan pendekatan ini dalam melakukan penelitian.

Uraian di bab ini hanya bersifat perkenalan dan memberikan gambaran selintas mengenai peran ML serta Bayesian analysis pada statistik-ekonometrika. Untuk uraian lengkap setingkat buku teks, pembaca dapat merujuk pada Gareth, Witten, Hastie, dan Tibsharani (2013), serta Efron dan Hastie (2016). Varian (2014), Belloni, Chernozhukov dan Hansen (2014), Mullainathan dan Spies (2017), dan Athey (2017) memberikan gambaran mengenai adopsi *machine learning* terhadap disiplin ilmu ekonomi (melalui ekonometrika). Untuk ekonometrika Bayesian, Geweke, Koop dan Van Dijk (2011), serta Greenberg (2013), Sanchez (2017), dan Chan et al (2019) adalah beberapa rekomendasi literatur.

MACHINE LEARNING: SUATU PENGANTAR

Perkembangan ML tidak terlepas dari kamampuan kita untuk mengakuisisi data. Dengan kemampuan teknologi saat ini maka

data dalam struktur yang super kompleks yakni *Big Data* dapat disediakan. Ahrens, Aitken, dan Schaffer (2020) mendeskripsikan big data sebagai struktur data yang sangat elaborated (memiliki jumlah dan atribut-dimensi yang sangat banyak) yang dapat diklasifikasikan sebagai (a) *high dimensional*; jumlah observasi (N) lebih kecil dari jumlah atribut (p), dan (b) *long data*; jumlah observasi sangat banyak tetapi hanya sedikit atribut. *Machine learning* (ML) adalah suatu algorithma khusus yang didesain agar dapat bekerja secara efisien dan efektif pada struktur data seperti ini.

Suatu ilustrasi tentang *high dimensional data* dapat dilihat dari implementasi *Internet of things*. Sangat banyak aktivitas seorang individu yang dilakukan melalui aplikasi Internet mulai dari bersosialisasi, belanja, perbankan, investasi, belajar, hingga rekreasi. Setiap aktivitas tersebut akan menghasilkan informasi mengenai diri individu sendiri. Dapat dibayangkan jumlah atribut (karakter) yang saat ini dapat diperoleh dari seorang individu telah melonjak secara eksponensial dibandingkan, misalnya 10 tahun yang lalu. Dengan demikian, sangat mungkin untuk mempostulasikan suatu hipotesis mengenai perilaku membeli barang X dari seseorang; tidak lagi hanya dipengaruhi oleh variabel konvensional seperti penghasilan, profil demografi, dan lokasi. Faktor-faktor seperti seberapa aktif individu tersebut disosial media, aktivitas pembelian barang terkait lain (komplemen dan substitusi), intensitas penggunaan kartu kredit, dan banyak lainnya dapat menjadi kandidat variabel penjelas. Dalam kondisi ini, bukan tidak mungkin jumlah atribut dapat melebihi jumlah individu.

Kemampuan instrumen dan *storage* data sekarang telah memungkinkan perekaman data dengan frekuensi sangat tinggi. Dalam kepustakaan ilmu keuangan, riset terhadap pergerakan aset

keuangan pada frekuensi detik telah dilakukan. Hanya untuk periode 1 tahun saja, perekaman tersebut akan memberikan kita sebanyak 31.536.000 observasi. Ini adalah suatu ilustrasi *long data*, yaitu sesuatu yang mungkin belum terbayangkan keberadaannya (paling tidak secara massal) 20 tahun yang lalu.

Selanjutnya, berdasarkan bentuk prosesnya *machine learning* dibedakan menjadi *supervised* dan *unsupervised*. *Supervised learning* dilakukan dalam bentuk di mana kita telah memiliki suatu variabel acuan (variabel dependen) dan suatu set kandidat variabel penjelas (prediktor). Berbeda dengan ekonometrika standar, dalam *supervised learning* nilai suatu spesifikasi regresi dititikberatkan pada kemampuan memprediksi *out of sample*. Dalam versi yang kompleks, data yang digunakan untuk estimasi regresi (disebut juga *training data*), akan digunakan untuk melewati beberapa tahap *out of sample data* (*validating data*). Koefisien-koefisien regresi "dilatih" untuk dapat mencapai nilai kriteria yang optimal (biasanya formulasi *prediction error*). Dengan persyaratan fungsional dan numeris tertentu, biasanya solusi yang optimal dapat dicapai setelah ribuan (atau bahkan jutaan) iterasi.

Sedangkan dalam *unsupervised learning*; tidak ada tujuan tertentu yang ingin dicapai. Kita memiliki sejumlah *N* observasi dengan *P* atribut; selanjutnya ML dilakukan untuk melihat apakah di dalam dataset matriks *N* x *P* terdapat pola-pola seperti group, cluster, dan kelompok data lainnya. Unsupervised ML memang berasal dari Cluster dan Factor Analysis (lihat Bab 6 di Jilid 1) yang ditujukan untuk mengurangi kompleksitas data.

Cameron (2019) memberikan suatu tabel taksonomi metode ML yang cukup popular di penelitian ekonomi dan bisnis (lihat Tabel 18.1). Secara umum, ML bersifat lebih pragmatis dibandingkan

Kelas Fungsi (dan Parameterisasi); $\mathcal{F}$	Regularizer; $R(f)$
Prediktor Global Parametrik Linear $X\beta$ (dan generalisasinya)	Subset selection: $\|\beta\|_0 = \sum_{j=1}^{b} I_{\beta_j=0}$ LASSO: $\|\beta\|_1 = \sum_{j=1}^{k} \|\beta_j\|$ Ridge: $\|\beta\|_2^2 = \sum_{j=1}^{k} \beta_j^2$ Elastic net: $\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2$
Local/Prediktor Nonparametrik Decision/regression trees Random forest (kombinasi linear dari trees) Nearest neighbors Kernel regression	Kedalaman, jumlah nodes/leaves, minimum ukuran leaf, information gains at splits Jumlah trees, jumlah variabel yang digunakan pada setiap tree, besar sampel bootstrap, kompleksitas tree Jumlah neighbor Kernel Bandwith
Prediktor Campuran (*Mixed Predictor*) Deep learning, neural nets, convolutional neural networks. Splines	Jumlah level, jumlah neuron per level, konektivitas antarneuron. Jumlah knots dan order
Prediktor Kombinasi Bagging: unweighted average of predictors from bootstrap draws Boosting: kombinasi linear dari prediksi residual Ensemble: kombinasi tertimbag dari berbagai prediktor	Jumlah penarikan, ukuran sampel bootstrap dan parameter regularisasi individual Learning rate: jumlah iterasi dan parameter regularisasi individual

TABEL 18.1. Taksonomi Machine Learning (diadaptasi dari Cameron, (2019)

ekonometrika. Dalam ekonometrika kemampuan substantif suatu model: yaitu menegaskan adanya hubungan antara variabel *y* dan *x* sama pentingnya dengan kemampuan prediksi. Sedangkan ML lebih kepada kemampuan prediksi; di mana model terbaik adalah yang memiliki kriteria *prediction error* yang terkecil.

Dengan perkembangan ketersediaan data yang semakin elaborated seperti yang dijelaskan sebelumnya, kalangan peneliti melihat potensi sinergi dari kedua teknik analisis tersebut. Teori ekonomi biasanya tidak cukup detail dalam memberikan arah bagi spesifikasi model ekonometrika. Di sisi lain, kemampuan komputasi yang ada telah memungkinkan untuk melakukan estimasi terhadap model-model yang kompleks. Sebagai gambaran, tantangan bagi teori ekonomi tidak hanya menspesifikasikan hubungan antara variabel Y dan X_1, X_2, serta X_3 secara plain vanilla. Berbagai pengembangan model tersebut bisa dilakukan seperti mengambil bentuk log, lag-forward value dari variabel, interaksi, bentuk fungsional, dan frekuensi alternatif. Ringkasnya pilihan desain ekonometrika telah jauh melampaui kemampuan abstraksi dari teori, sehingga arah penemuan ilmu sekarang kembali dibalik: yaitu dari data ke prinsip ilmu.

Perkembangan ini juga sejalan dengan kritik dari Simon, Nelson, dan Simonsohn (2011):

> *"It is common (and accepted practice) for researchers to explore various analytic alternatives, to search for a combination that yields 'statistikal significance,' and to then report only what 'worked.'"*

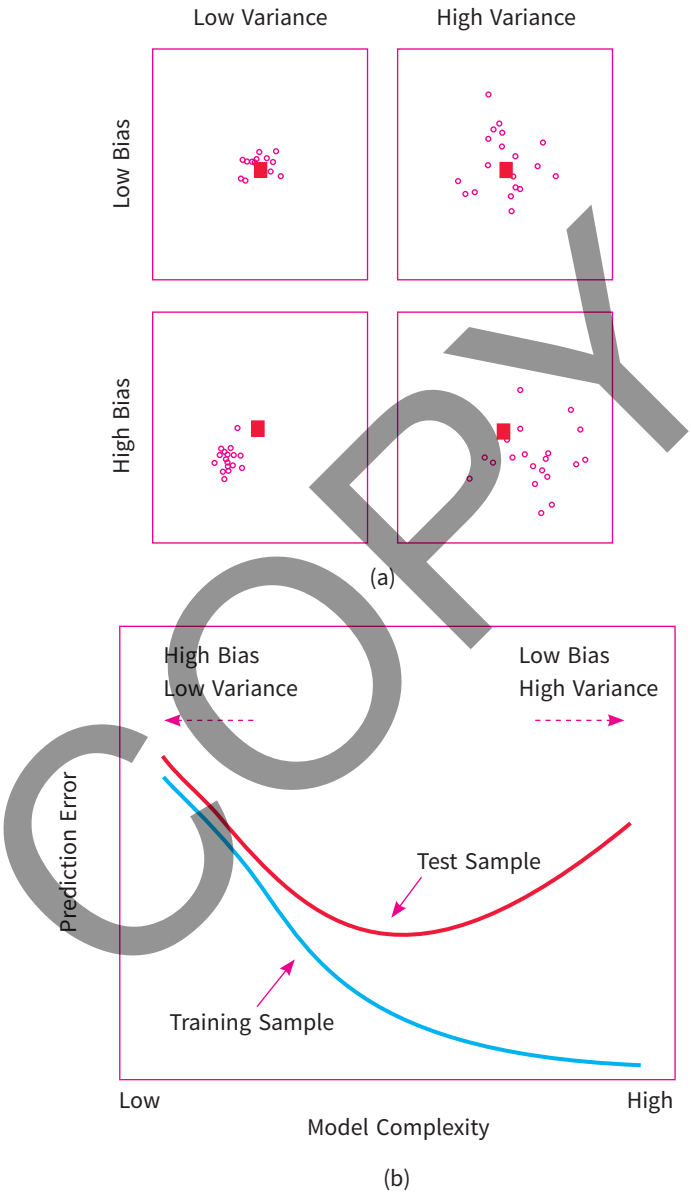
Dengan demikian, metode ML dipandang lebih baik (paling tidak lebih sistematis dan objektif) dibandingkan *self-selection*

dalam suatu penelitian yang diawalnya dianggap sebagai objektif. Dua bagian berikut ini akan mengilustrasikan bagaimana style ML dalam ekonometrika dilakukan.

LASSO ESTIMATOR

Least Absolute Shrinkage Selector Operator (LASSO) diperkenalkan secara terpisah oleh Frank dan Friedman (1993) serta Tibshirani (1996). LASSO adalah suatu estimator ML yang merupakan pengembangan langsung dari regesi linear yang disebut juga sebagai *regularized regression*. Seperti halnya OLS, parameter dari *regularized regression* juga diperoleh dari fungsi objektif meminimumkan residual kuadrat ditambah dengan regularization penalty yang ditujukan untuk membatasi kompleksitas model. Output dari regularized regression lebih ditujukan pada kemampuan prediksi; sementara parameter-parameter tidak dapat diinterpretasikan sebagai suatu hubungan sebab akibat serta memiliki inferensial yang lebih rumit.

Keunggulan regularized regression berupa kemampuan prediksi berasal dari optimalisasi bias-varians trade off. Gambar 18.1 mengilustrasikan bias-varians trade off. Icon kotak kecil berwarna merah menunjukkan nilai sesungguhnya (populasi) dari parameter. Suatu estimator yang memiliki bias dan varians yang kecil ditunjukkan oleh hasil estimasi yang menggerombol (secara sempit) di sekitar kotak merah. Tibshirani (1996) menunjukkan bahwa varians cenderung meningkat dengan semakin kompleknya model; namun di sisi lain, bias cenderung menurun dengan semakin tingginya kompleksitas model. Dengan demikian, memang terdapat *trade off* bias-varians terkait kompleksitas model.



GAMBAR 18.1 Bias dan Varians Trade Off (a) dan Optimal Training (b) (diadaptasi dari Gareth et al 2013)

Dengan mengurangi kompleksitas model, *regularized regression* memiliki *out of sample prediction error* yang lebih rendah dibandingkan OLS. Dengan demikian, pencarian parameter regresi dapat digambarkan sebagai "*optimal training*". Di sini set observasi akan dipecah menjadi 2 bagian di mana yang pertama adalah estimasi model regresi (disebut test sample; *training data*) dan yang kedua digunakan untuk "melatih" atau "*fine tuning*" regresi tersebut (*validating data*) demi mendapatkan titik yang *optimal* (ditandai dengan garis vertikal berwarna biru). Pada titik yang optimal, kompleksitas model akan memberikan kriteria terbaik (secara bias-varians trade off) dari segi *in sample* dan *out of sample*.

Bagaimana melakukan *finetuning* tersebut? Kita harus menspesifikasikan permasalahan yang dihadapi sebagai formulasi matematis (LASSO) berikut ini:

$$\hat{\beta}_{LASSO}(\lambda) = arg\ min\ \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \frac{\lambda}{n}\sum_{j=1}^{p}\psi_j|\beta_j| \qquad (18.1)$$

Persamaan 18.1 adalah perumusan estimasi LASSO. Term pertama pada persamaan tersebut adalah objective function dari OLS (minimumkan residual kuadrat); sedangkan term kedua adalah penalti yang dikenakan akibat bertambahnya kompleksitas model. Penalti ini adalah suatu fungsi dari λ; parameter *fine tuning*, dan *ψj* yang merupakan *predictor specific penalty loading*.

Terdapat beberapa alternatif bagi term penalti; di mana salah satu yang popular adalah bentuk kuadratik yakni

$$\hat{\beta}_{RIDGE}(\lambda) = arg\ min\ \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \frac{\lambda}{n}\sum_{j=1}^{p}\psi_j^2\beta_j^2 \qquad (18.2)$$

Estimator dengan *term penalty* berbentuk kuadratik dikenal dengan nama *ridge regression* (Tikhonov, 1963 serta Hoerl dan Kennard (1970)). Regresi LASSO memiliki keunggulan dalam hal seleksi variabel, sedangkan ridge regression memiliki kelebihan dalam menangani struktur data yang *high dimension* (Ahrens, Hansen, dan Schaffer, 2020). Zhou dan Hastie (2005) mengkombinasikan term penalti LASSO dan ridge yang menghasilkan *estimator elastic net* (EN) seperti yang diberikan pada Persamaan 18.3. Estimator ini sangat fleksibel; di mana α dapat disesuaikan terhadap intensitas permasalahan: yaitu kebutuhan akan seleksi variabel versus *high dimensionality*.

$$\hat{\beta}_{EN}(\lambda) = arg\ min \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \frac{\lambda}{n}\left[\alpha\sum_{j=1}^{p}\psi_j\left|\beta_j\right| + (1-\alpha)\sum_{j=1}^{p}\psi_j^2\beta_j^2\right] \tag{18.3}$$

Zhou (2006) mengusulkan LASSO adaptif yang merupakan pengembangan dari LASSO standar. Keunggulan LASSO adaptif adalah memiliki kinerja yang cukup baik meskipun terdapat derajat yang *high dimensionality*. LASSO adaptif adalah LASSO yang menggunakan term penalti $\psi_j = 1/|\hat{\beta}_{0,j}|^{\theta}$ di mana $\hat{\beta}_{0,j}$ adalah suatu estimator awal. Belloni, Chernozhukov, dan Wang (2011) mengusulkan *square root* LASSO yang memiliki keunggulan dalam hal adanya heterokedastisitas dan prosedur pemilihan λ yang lebih sederhana.

$$\hat{\beta}_{\sqrt{LASSO}}(\lambda) = arg\ min \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i'\beta\right)^2} + \frac{\lambda}{n}\sum_{j=1}^{p}\psi_j|\beta_j| \tag{18.4}$$

Pengenaan term penalti secara adhoc seperti yang diberikan pada rumus *regularized regression* akan menyebabkan potensi peningkatan bias (Belloni, Chernozhukov, dan Hansen, 2014). Penulis tersebut selanjutnya mengusulkan penggunaan LASSO hanya untuk seleksi variabel; di mana variabel yang terpilih akan digunakan dalam regresi standar (OLS). Estimator yang diperoleh dengan cara ini dikenal sebagai *post-estimation* OLS.

$$\hat{\beta}_{POST)} = \ arg\ min\ \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\beta)^2\ subject\ to\ \beta_j = 0\ if\ \tilde{\beta}_j = 0 \quad (18.5)$$

di mana $\tilde{\beta}_j$ adalah *sparse* estimator tahap pertama (contohnya adalah LASSO).

Setelah merumuskan permasalahan LASSO, pilihan parameter fine tuning λ dan α (untuk estimator *EN*) akan sangat tergantung pada (a) tujuan analisis (prediksi atau seleksi model), (b) konstrain komputasi, dan (c) apakah asumsi independent and identically distributed (IID) mengalami pelanggaran (Ahrens, Hansen, dan Schaffer, 2020). Terdapat tiga pilihan yang direkomendasikan, yakni

a. Kriteria informasi
b. *Cross validation*
c. *Theory driven*

Seperti halnya regresi standar, penggunaan kriteria informasi atas pilihan model adalah yang memiliki nilai terkecil. Ada dua kriteria informasi yang disarankan untuk memilih parameter *fine tuning* yakni Akaike Information Criterion = AIC dan Bayesian Information Criterion = BIC. Apabila jumlah data tidak banyak (*small sample*),

disarankan untuk menggunakan *corrected* AIC-AIC$_C$ (Hurvich and Tsai, 1989). Sedangkan jika p berjumlah banyak ($p > 1.000$), maka disarankan untuk menggunakan extended BIC (EBIC, Chen and Chen 2008).

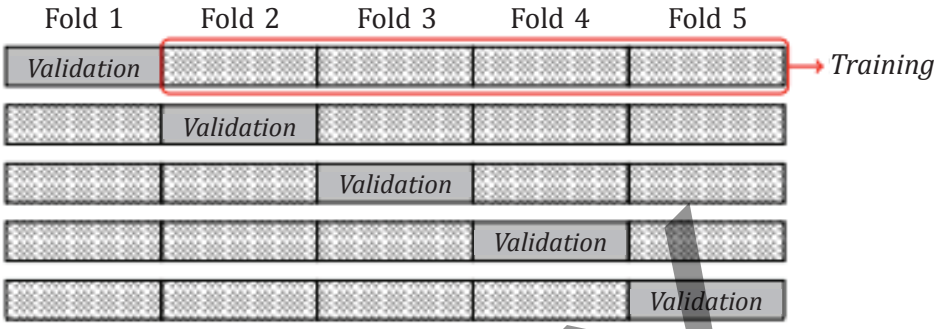
Tujuan dari cross validation adalah untuk melihat kinerja: kinerja prediktif (*predictive performance*) model atas data baru (*unseen data*). Cross validation dilakukan melalui aktivitas berulang (iteratif) yang membagi data ke dalam dua set: *training data* (data yang digunakan untuk estimasi) dan *validation data* (data yang digunakan untuk evaluasi kinerja prediktif). Metode cross validation yang paling popular adalah K-fold Cross Validation.

Bagaimana *cross validation* bekerja? Kita akan menggunakan ilustrasi sederhana seperti K = 5. Kita juga menggunakan *Mean Square Prediction Error* (MSPE) sebagai kriteria kinerja

$$MSPE_K(\lambda, \alpha) = \frac{1}{n_K} \sum_{i \epsilon K_k} \left(y_i - x_i' \beta_k(\lambda, \alpha)\right)^2 \quad (18.6)$$

Kita memiliki suatu dataset sejumlah (observasi) n. Kita dapat membuat partisi (fold) terhadap seluruh observasi (lihat Gambar 18.2). Selanjutnya, kita dapat melakukan estimasi regresi pada training data: partisi ke–2–5, yang kemudian hasilnya akan dievaluasi dengan menggunakan partisi 1. Kemudian kita menggunakan data partisi 1 dan 3–5 untuk melakukan estimasi; yaitu data pada partisi ke-2 digunakan untuk validasi. Demikian seterusnya di mana data setiap partisi akan digunakan satu kali untuk validasi.

Algoritma K-Fold Cross Validation adalah mencari angka K yang akan meminimumkan rata-rata MSPE, yakni



GAMBAR 18.2. 5-Fold Cross Validation (sumber: Ahrens, 2019)

$$L^{CV} = \frac{1}{K}\sum_{k-1}^{K} MSPE_k(\lambda, \alpha) \tag{18.7}$$

Chernozhukv et al (2016) memperkenalkan suatu kelas LASSO (disebut rigorous LASSO) di mana pencarian parameter penalti λ dan α dilakukan berdasarkan fungsi tujuan: yaitu konsistensi prediksi dan estimasi parameter. Estimator ini akan memungkinkan inferensial kausalitas dalam kondisi banyak variabel instrumen dan kontrol. Di samping itu, rigorous LASSO juga bersifat robust terhadap heterokedastisitas (Belloni et al, 2012). Rigorous LASSO didasarkan pada pemenuhan persyaratan regularisasi

$$\frac{\lambda}{n} \geq \max_{1 \leq j \leq p} \left|\psi_j^{-1} S_j\right| \; di\ mana\ S_j = \frac{2}{n}\sum_{i=1}^{n} x_{ij}\epsilon_i \tag{18.8}$$

Akan terpenuhi secara simtotik; atau

$$P(\max_{i \le j \le p} c|S_j| \le \frac{\lambda \psi_j}{n}) \to 1 \; ketika \; n \to \infty; \gamma \to \infty \quad (18.9)$$

Jika penalti level dan loadings ditentukan menurut rumus

$$Homokedastis: \lambda = 2c\sigma\sqrt{n}\Phi^{-1}\left(1 - \frac{\gamma}{(2p)}\right), \quad \psi_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2} \quad (18.10)$$

$$Heterokedastis: \lambda = 2c\sigma\sqrt{n}\Phi^{-1}\left(1 - \frac{\gamma}{(2p)}\right), \quad \psi_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 \epsilon_i^2} \quad (18.11)$$

Contoh 18.1

Kita akan menggunakan data dari studi Harrison dan Rubinfeld (1978) mengenai perilaku harga rumah di kota Boston US. File yang digunakan adalah housing.dta yang berisi data seperti yang disajikan pada Tabel 18.2.

Pertama kita akan menggunakan LASSO estimator dengan pemilihan model berdasarkan kriteria informasi: EBIC. Hal ini dilakukan dengan perintah **lasso2 medv crim-lstat, lic(ebic)**. Term – pada crim-lstat menunjukkan bahwa estimasi LASSO dilakukan pada seluruh variabel independen.

Seperti dapat dilihat pada Tabel 18.3, output dari LASSO dengan menggunakan kriteria informasi EBIC untuk memilih λ akan menghasilkan 2 tipe output. Pada panel atas akan ditampilkan

Name	*Description*
`crim`	per capita crime rate by town
`zn`	proportion of residential land zoned for lots over 25,000 sq.ft.
`indus`	proportion of non-retail business acres per town
`chas`	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
`nox`	nitrit oxides concentration (parts per 10 million)
`rm`	average number of rooms per dwelling
`age`	proportion of owner-occupied units builts prior to 1940
`dis`	weighted distances to five Boston employment centres
`rad`	indexs of accessibility to radial highways
`tax`	full-value property-tax rate per $10,000
`pratio`	pupil-teacher ratio by own
`b`	1000(Bk – 0,63)2 where Bk is the proportion of blacks by town
`lstat`	% lower status of the population
`medv`	Median value of owner-occupied homes in $1000's

TABEL 18.2. Variabel yang Digunakan; Harrison dan Rubinfeld (1978)

sejumlah tertentu (satuan: knot) progresi dari pencarian model; yaitu memasukkan-mengeluarkan variabel. Kriteria EBIC (minimum pada λ = 16,218) LASSO memilih 11 dari 13 variabel independen untuk dimasukkan ke dalam regresi OLS (*Post Estimation* OLS). Hasil estimasi Post Estimation OLS (serta perbandingannya dengan LASSO Standar) dengan menggunakan 11 variabel tersebut diberikan pada panel di bagian bawah. Kita dapat membuat grafik jalur pergerakan (trayektori) nilai koefisien variabel terpilih sebagai fungsi dari λ. Hal ini dilakukan dengan perintah **lasso2 medv crim-lstat, plotpath(lnlambda) plotopt(legend(off)) plotlabel plotvar(rm chas rad lstat ptratio dis)** dengan hasil yang dapat dilihat pada Gambar 18.3.

Selanjutnya, estimator *Cross Validation* (CV) LASSO dapat dilaksanakan dengan perintah **cvlasso medv crim-lstat, lopt plotcv**. Term lopt adalah opsi untuk mencari λ yang akan meminimumkan

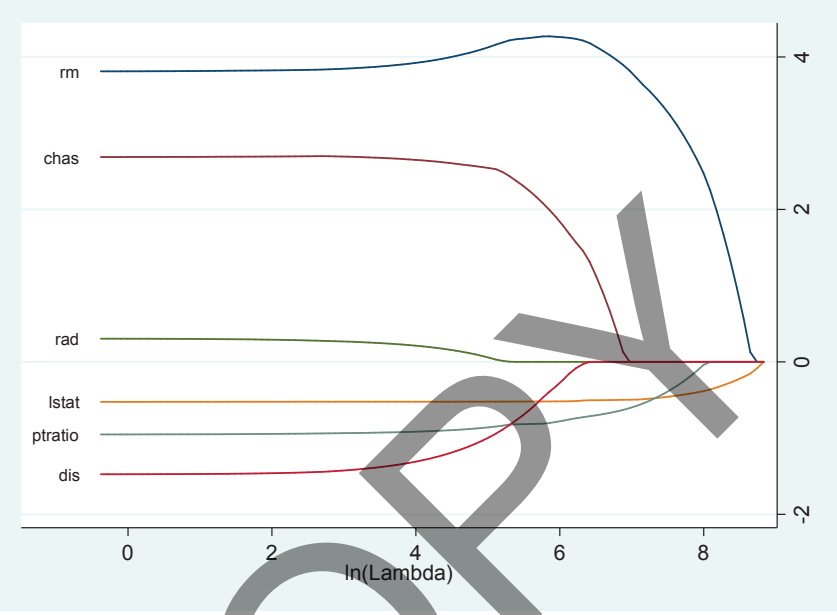
Knot	ID	Lambda	s	L1-Norm	EBIC	R-sq	Action
1	1	6858.98549	1	0.00000	2250.74087	0.0000	Added _cons.
2	2	6249.65212	2	0.08440	2207.91747	0.0924	Added lstat.
3	3	5694.45025	3	0.28099	2166.62026	0.1737	Added rm.
4	10	2969.09108	4	2.90443	1902.66627	0.5156	Added ptratio.
5	20	1171.07070	5	4.79923	1738.09475	0.6544	Added b.
6	22	972.24347	6	5.15524	1727.95402	0.6654	Added chas.
7	26	670.12971	7	6.46233	1709.14647	0.6815	Added crim.
8	28	556.35346	8	6.94988	1705.73465	0.6875	Added dis.
9	30	461.89442	9	8.10548	1698.65787	0.6956	Added nox.
10	34	318.36590	10	13.72934	1679.28783	0.7106	Added zn.
11	39	199.94307	12	18.33493	1671.61672	0.7219	Added indus rad.
12	41	165.99625	13	20.10742	1669.76857	0.7263	Added tax.
13	47	94.98916	12	23.30144	1645.44345	0.7359	Removed indus.
14	67	14.77724	13	26.71617	1642.91756	0.7405	Added indus.
15	82	3.66043	14	27.44510	1648.83626	0.7406	Added age.

Use **long** option for full output.

Use lambda=16.2179985686768 (selected by EBIC).

Selected	Lasso	Post-est OLS
crim	-0.1028391	-0.1084133
zn	0.0433716	0.0458449
chas	2.6983217	2.7187163
nox	-16.7712501	-17.3760234
rm	3.8375782	3.8015788
dis	-1.4380341	-1.4927115
rad	0.2736598	0.2996085
tax	-0.0106973	-0.0117780
ptratio	-0.9373015	-0.9465246
b	0.0091412	0.0092908
lstat	-0.5225124	-0.5225535
Partialled-out*		
_cons	35.2705784	36.3411450

TABEL 18.3 Estimasi LASSO



GAMBAR 18.3. Jalur Pergerakan Koefisien (Trayektori) Variabel Terpilih

MSPE, sedangkan term plotcv akan memberikan grafik pergerakan MSPE sebagai fungsi dari log natural λ. Dengan menjalankan perintah cvlasso secara default otomatis akan melakukan partisi (K) sebanyak 100. Untuk setiap partisi, akan dihitung λ, MSPE, dan deviasi standar dari MSPE. Perintah ini akan memberikan tanda * pada λ terpilih (terletak di sebelah kanan tabel pada baris λ optimal); yang dalam contoh ini adalah pada K = 64 dan λ = 19,535. Kemudian atas variabel terpilih dihitung nilai koefisiennya (terletak pada panel bawah).

Kita dapat melihat pergerakan MSPE sebagai fungsi dari log natural λ; yang ditunjukkan pada Gambar 18.4. Terdapat dua garis fungsi yakni MSPE sebagai fungsi log natural (garis solid warna

```
K-fold cross-validation with 10 folds. Elastic net with alpha=1.
Fold 1 2 3 4 5 6 7 8 9 10
```

	Lambda	MSPE	st. dev.	
1	6858.9855	84.302552	5.7124689	
2	6249.6521	77.022038	5.5626292	
3	5694.4503	70.352232	5.3037622	
4	5188.571	63.914422	4.9182026	
5	4727.6326	58.319707	4.4997251	
6	4307.6428	53.676524	4.1600126	
7	3924.9637	49.823174	3.886599	
8	3576.2807	46.625425	3.6691123	
9	3258.5738	43.964849	3.4996305	
10	2969.0911	41.613694	3.4044528	

…output tidak ditampilkan.

58	34.137332	23.447448	3.1439097	
59	31.104666	23.438247	3.1410603	
60	28.341413	23.430679	3.1385082	
61	25.82364	23.424869	3.1361665	
62	23.529539	23.421433	3.1339816	
63	21.43924	23.419627	3.1318222	
64	19.534637	23.418936	3.1298345	*
65	17.799233	23.419177	3.1280904	
66	16.217999	23.419668	3.1266575	
67	14.777236	23.42057	3.1253492	
68	13.464467	23.421575	3.124199	

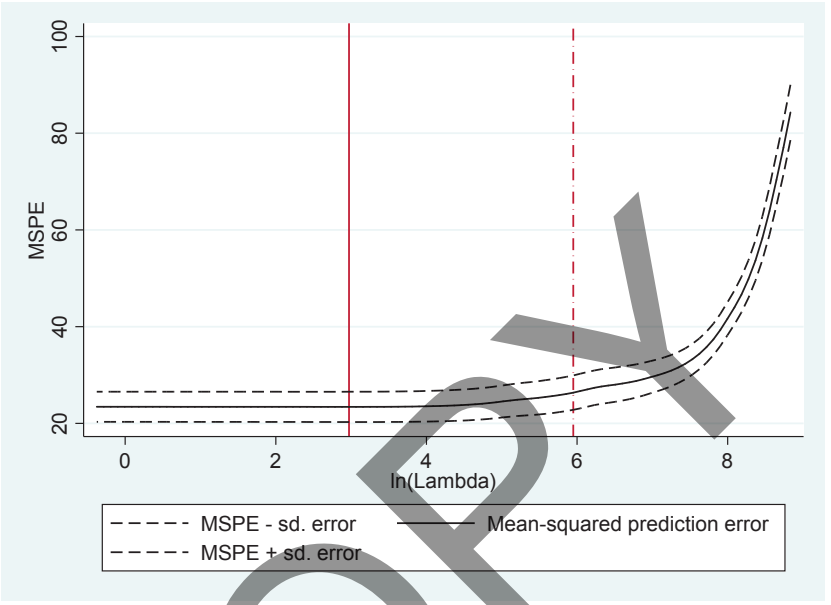
…output tidak ditampilkan.

97	.90671752	23.440956	3.1135392	
98	.82616723	23.441147	3.1134729	
99	.75277281	23.441321	3.1134126	
100	.68589855	23.441481	3.1133578	

```
* lopt = the lambda that minimizes MSPE.
  Run model: cvlasso, lopt
^ lse = largest lambda for which MSPE is within one standard error of the minimal MSPE.
  Run model: cvlasso, lse
Estimate lasso with lambda=19.535 (lopt).
```

Selected	Lasso	Post-est OLS
crim	-0.1016991	-0.1084133
zn	0.0428658	0.0458449
chas	2.6941509	2.7187163
nox	-16.6475718	-17.3760234
rm	3.8449402	3.8015788
dis	-1.4268524	-1.4927115
rad	0.2683532	0.2996085
tax	-0.0104763	-0.0117780
ptratio	-0.9354154	-0.9465246
b	0.0091106	0.0092908
lstat	-0.5225040	-0.5225535
Partialled-out*		
_cons	35.0516438	36.3411450

TABEL 18.4. Hasil Estimasi Cross Validation LASSO



GAMBAR 18.4. MSPE sebagai Fungsi dari log natural λ

hitam) serta fungsi +/1 1 deviasi standar MSPE (garis putus-putus warna hitam). Garis solid vertikal warna merah mengindikasikan λ optimal; sedangkan garis putus-putus vertikal menunjukkan yang berada pada +/- 1 deviasi standar MSPE.

Terakhir, kita dapat mengestimasi *rigorous lasso* (rlasso) dengan perintah **rlasso medv crim-lstat, supscore**. Term supscore adalah opsi untuk menjalankan uji hipotesis berganda (*sup-score test*); yakni hipotesis null seluruh β dari model terpilih adalah sama dengan nol.

Pelaksanaan perintah rlasso hanya akan menghasilkan output bagi model terpilih beserta uji hipotesis sup-score. Seperti dapat dilihat pada Tabel 18.5 bahwa dari 13 pilihan variabel independen, perintah rlasso "hanya" memilih 5 variabel. Uji hipotesis berganda

Selected		Lasso	Post-est OLS
chas		0.6614715	3.3200251
rm		4.0224500	4.6522737
ptratio		-0.6685444	-0.8582708
b		0.0036058	0.0101119
lstat		-0.5009804	-0.5180622
_cons	*	14.5986084	11.8535885

*Not penalized

Sup-score test H0: beta=0
CCK sup-score statistic 16.59 p-value= 0.000
CCK 5% critical value 3.18 (asympt bound)

TABEL 18.5. Hasil Estimasi Rigorous LASSO

(sup-score) memberikan statistik hitung jauh di atas nilai kritis, sehingga hipotesis null $\beta = 0$ dapat ditolak pada p value: 0,000.

PENGGUNAAN LASSO UNTUK PEMODELAN REGRESI

Uraian sebelumnya menunjukkan penggunaan LASSO untuk prediksi. Ahrens, Schaffer, dan Hansen (2018) juga merekomendasikan LASSO untuk melakukan spesifikasi model. Seperti yang dijelaskan sebelumnya, keluarga estimator LASSO digunakan terutama untuk struktur data yang *high dimensional* (p/n mendekati 1 atau lebih dari 1). Spesifikasi model ini khususnya dilakukan untuk memilih variabel kontrol dan variabel instrumen. Pemilihan variabel kontrol ditujukan untuk memitigasi masalah *omitted variable*, sedangkan pencarian instrumen ditujukan untuk menangani

masalah endogenitas. Permasalahan ini dapat diformulasikan dengan persamaan matematis berikut

$$y_i = \underbrace{\alpha d_i}_{\text{VIR; mungkin endogenous}} + \underbrace{\beta_1 x_{i1} + \cdots + \beta_p x_{ip}}_{\text{kontrol}} + \epsilon_i \qquad (18.12)$$

Variabel (atau dapat juga berupa vektor) d_i adalah *variables of interest* (VIR) yang mungkin mengalami endogenitas. Sedangkan x_1 sampai dengan x_p adalah suatu set kandidat variabel kontrol. Apabila d_i mengalami endogenitas, maka harus dilakukan pencarian instrumen yang memenuhi kondisi orthogonality dengan residual regresi. Selanjutnya, sangat mungkin jumlah kandidat instrumen dan kontrol melebihi jumlah observasi; dengan kata lain, struktur data adalah *high dimension*. Terdapat dua metode untuk mengelola permasalahan ini, yakni

a. Post Double Selection LASSO (Belloni, Chernozhukov, dan Hansen, 2014a, PDS LASSO).
b. Double Orthogonalization (Chernozhukov, Hansen, dan Spindler, 2015; CHS).

PDS LASSO dilakukan dengan tahapan sebagai berikut

i. Melakukan estimasi Persamaan 18.12 tanpa memasukkan VIR; hanya variabel kontrol dengan menggunakan LASSO. Sebut saja set variabel kontrol terpilih dengan A.

ii. Melakukan estimasi regresi VIR terhadap variabel kontrol dengan LASSO. Sebut saja set variabel kontrol terpilih sebagai B

$$d_i = \gamma_1 x_{i1} + \cdots + \gamma_p x_{ip} + \varepsilon_i \quad (18.13)$$

iii. Melakukan regresi berikut

$$y_i = \alpha d_i + w_i'\delta + \mu_i \quad (18.14)$$

di mana $w_i' = A \cup B$ adalah set gabungan antara A dan B.

CHS secara esensi adalah sama, namun di sini yang digunakan hanya set variabel yang diperoleh dari langkah (ii) dan langkah (iii); atau secara matematis

$$y_i = B_i'\pi + A_i'\omega + \nu_i \quad (18.15)$$

Contoh 18.2

Di sini kita akan menggunakan data dari studi Acemoglu, Johnson, dan Robinson (2001); file AJR.dta. Dataset ini memiliki 64 observasi (negara), dan 23 kandidat variabel kontrol serta 5 kandidat variabel instrumen sehingga dapat dikatakan bersifat "agak" *high dimension*. Selanjutnya, misalkan kita memodelkan regresi variabel log PDB negara pada tahun 1995 (dalam denominasi *Purchasing Power Parity* = PPP) terhadap variabel avexpr (indeks proteksi terhadap risiko ekspropriasi), di mana variabel avexpr bersifat endogenous.

Kita mempertimbangkan variabel lain yang ada pada data sebagai variabel kontrol dan variabel instrumen (bagi avexpr). Secara

spesifik, variabel-variabel lat_abst, edes1975, avelf, temp*, humid*, steplow-oilres adalah kandidat variabel kontrol. Sedangkan variabel logem4 euro1900-cons00a adalah kandidat variabel instrumen. Dapat dicatat di sini bahwa (nama variabel)* adalah suatu wildcard, di mana ada beberapa variabel yang memiliki nama sebelum * yang sama. Tanda "-" di antara dua nama variabel berarti menyatakan variabel-variabel yang urutannya dalam *variables window* di antara dua nama variabel tersebut juga digunakan (dalam hal steplow-oilres sebagai variabel kontrol).

Suatu perintah pdslasso[2] yang digunakan untuk memilih kandidat variabel instrumen dan kontrol diberikan sebagai **pdslasso logpgp95 (avexpr=logem4 euro1900-cons00a) (lat_abst edes1975 avelf temp* humid* steplow-oilres), partial(logem4 lat_abst)**. Term dalam tanda kurung pertama adalah pernyataan variabel endogen dan kandidat instrumennya, sedangkan tanda kurung kedua menyatakan kandidat variabel kontrol. Opsi partial() adalah melakukan partialling out terhadap variabel yang ada dalam tanda kurung.

Dari output perintah (Tabel 18.5) dapat dilihat bahwa PDS LASSO memilih 1 variabel instrumen dari 5 kandidat, yakni logem4. Sedangkan dari 24 kandidat variabel kontrol dipilih 2 yakni edes1975 dan humid3. Hasil post OLS LASSO (panel bawah) memperlihatkan sesuai hipotesis bahwa avexpr memiliki pengaruh positif terhadap kinerja ekonomi 1995; namun demikian, variabel kontrol bersifat tidak signifikan.

[2] Perintah pdslasso dapat diganti dengan ivlasso. Sebenarnya kedua perintah ini memiliki tujuan yang sama, tetapi ada sedikit perbedaan pada restriksi. Lihat Ahrens, Schaffer, dan Hansen (2018) halaman 51.

EKONOMETRIKA BAYESIAN: SUATU PENGANTAR

Metode ekonometrika yang telah dibahas di buku ini disebut dengan ekonometrika *frequentist*. Di sini kita akan memulai analisis dengan suatu dataset dan memperoleh estimator. Dengan asumsi eksperimen dapat diulangi menuju tak hingga, maka berdasarkan teori limit sentral secara asimtotik kita berharap estimator yang diperoleh dari dataset tersebut mencerminkan populasi. Sebaliknya, dalam *Bayesian econometrics*; jika kita memiliki suatu dataset (serta estimator yang diperoleh darinya) dan asumsi awal mengenai distribusi estimator (*prior belief*), maka pertanyaan yang hendak dijawab adalah berapa probabilitas bahwa data yang dimiliki berasal dari *prior distribution*.

Jika terdapat diskrepansi yang signifikan dalam arti data tidak mungkin berasal dari *prior distribution* (yang ditunjukkan dengan probabilitas akseptansi yang rendah), maka dari data dan *prior belief* tersebut dapat dikombinasi distribusi (dan estimator baru) yang lebih sesuai. Distribusi baru ini disebut *posterior distribution*; dan proses penggabungan informasi dari data dan *prior belief* disebut sebagai *updating*. Di samping perbedaan paradigma seperti yang diuraikan sebelumnya, ada perbedaan lain antara frequentist dan Bayesian yang selengkapnya dirangkum pada Tabel 18.6 (Sanchez, 2017).

Perlu ditegaskan di sini bahwa penggunaan ekonometrika Frequentist atau Bayesian adalah tergantung pada konteks: yaitu pertanyaan riset yang hendak dijawab (Geweke, Koop, dan Van Dijk, 2011 hal. 87). Apabila pertanyaan penelitian adalah ingin mengetahui apakah suatu parameter terletak pada selang tertentu, maka untuk menjawabnya akan digunakan ekonometrika Bayesian. Sedangkan jika kita ingin mengetahui berapa besar suatu parameter (dan inferensial statistiknya) dalam sampel berulang, maka hal ini dijawab

```
Estimation results:

Specification:
Regularization method:                  lasso
Penalty loadings:                       homoskedastic
Number of observations:                 59
Endogenous (1):                         avexpr
High-dim controls (23):                 edes1975 avelf temp1 temp2 temp3 temp4 temp5 humid1
                                        humid2 humid3 humid4 steplow deslow stepmid desmid drystep
                                        drywint landlock goldm iron silv zinc oilres
Selected controls, PDS (2):             edes1975 humid3
Selected controls, CHS-L (2):           edes1975 humid3
Selected controls, CHS-PL (2):          edes1975 humid3
Partialled-out controls (1):            lat_abst
High-dim instruments (5):               euro1900 democ1 cons1 democ00a cons00a
Selected instruments (1):               logem4
Partialled-out instruments (1):         logem4

Structural equation:

IV using CHS lasso-orthogonalized vars
------------------------------------------------------------------------------
    logpgp95 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avexpr |   .7335102   .1666875     4.40   0.000     .4068087    1.060212
------------------------------------------------------------------------------

IV using CHS post-lasso-orthogonalized vars
------------------------------------------------------------------------------
    logpgp95 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avexpr |   .5317677   .1522796     3.49   0.000     .2333051    .8302302
------------------------------------------------------------------------------

IV with PDS-selected variables and full regressor set
------------------------------------------------------------------------------
    logpgp95 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avexpr |   .8735301   .2929925     2.98   0.003     .2992753    1.447785
    edes1975 |   .0013051   .0083678     0.16   0.876    -.0150955    .0177057
      humid3 |   .0031166   .0087269     0.36   0.721    -.0139878     .020221
    lat_abst |  -.5020038   1.254352    -0.40   0.689    -2.960489    1.956482
       _cons |   2.253524   1.528477     1.47   0.140    -.7422353    5.249283
------------------------------------------------------------------------------
Standard errors and test statistics valid for the following variables only:
    avexpr
------------------------------------------------------------------------------
```


TABEL 18.6. Hasil Estimasi PDS Lasso untuk Seleksi Variabel Kontrol dan Variabel Instrumen pada Studi AJR(2001)

dengan ekonometrika frequentist. Meskipun demikian, terdapat area di mana kedua pendekatan ekonometrika ini *overlapping* (Chan et al, 2019). Keunggulan dan kelemahan ekonometrika Bayesian dapat dirangkum pada Tabel 18.7 (Sanchez, 2017).

Ekonometrika Frequentist	Ekonometrika Bayesian
a. Hasil estimasi berasal dari parameter populasi yang tidak diketahui; tetapi bersifat konstan. b. Data dianggap sebagai suatu skema sampling yang (secara hipotetis) dapat dilakukan berulang-ulang. c. Penggunaan data untuk memperoleh estimasi atas parameter populasi yang tidak diketahui d. Kualitas hasil estimasi sangat tergantung pada apakah data memenuhi asumsi yang digunakan pada model e. Kesimpulan didasarkan pada distribusi statistik yang diperoleh dari sampel random; dengan mengasumsikan parameter populasi tidak diketahui tetapi bersifat konstan.	a. Hasil estimasi diperoleh dari distribusi probabilitas suatu parameter populasi yang bersifat random. b. Data diasumsikan bersifat konstan c. Hasilnya diperoleh kombinasi data dengan prior belief atas parameter populasi. d. Distribusi posterior digunakan untuk membuat pernyataan probabilistik yang eksplisit mengenai hasil estimasi. e. Analisis Bayesian menjawab pertanyaan penelitian mengenai parameter populasi kondisional (berdasarkan) data yang dimiliki.

TABEL 18.7. Perbandingan Ekonometrika Frequentist versus Ekonometrika Bayesian (diadaptasi dari Sanchez, 2017)

Ekonometrika Bayesian dimulai dari suatu hukum (*Bayesian Law*) yang secara matematis diberikan sebagai berikut

$$p(B|A) = \frac{p(A \cap B)}{p(A)} = \frac{p(A|B)p(B)}{p(A)} \tag{18.16}$$

Keunggulan	Kelemahan
a. Dapat digunakan untuk semua model parametrik; karena prinsipnya berbasis Aturan Bayes. b. Inferensial bersifat eksak; di mana estimasi dan prediksi berasal dari distribusi posterior yang merupakan gabungan informasi dari data dan *prior belief*. c. Memberikan hasil yang intuitif khususnya untuk probabilitas hasil (*credible interval*). d. Tidak tergantung pada sampel.	a. Memiliki karakter subjektif yang cukup dominan; karena adanya *prior belief*. b. Sangat menantang dalam kebutuhan komputasi. c. Merumuskan permasalahan dan spesifikasi model untuk menjawab pertanyaan penelitian dapat menjadi suatu proses yang sangat kompleks.

TABEL 18.8. Keunggulan dan Kelemahan Ekonometrika Bayesian (diadaptasi dari Sanchez, 2017)

di mana $\boldsymbol{p(B|A)}$ adalah probabilitas terjadinya event B kondisional terjadinya A; di mana $\boldsymbol{p(A \cap B)}$ adalah probabilitas terjadinya joint event A dan B, $p(A)$ adalah probabilitas unconditional event A dan $p(B)$ adalah probabilitas unconditional event B. Secara lebih konkret dalam spesifikasi model ekonometrika kita notasikan y sebagai vektor data dan θ adalah parameter populasi yang akan diestimasi dari data, sehingga Persamaan 18.16 dapat diekspresikan kembali sebagai model Bayesian berikut

$$f(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{f(y)} \tag{18.17}$$

Selanjutnya $\boldsymbol{f(y|\theta)}$ dapat dikonstruksikan sebagai fungsi likelihood, yaitu observasi data y kondisional atas parameter tertentu $\boldsymbol{\theta}$.

$$L(y|\theta) = f(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta) \tag{18.18}$$

Kemudian $\boldsymbol{\theta}$ yang bersifat random diasumsikan dapat diekspresikan sebagai distribusi probabilitas ($\boldsymbol{\pi(\theta)}$). Fungsi distribusi marjinal y dapat diekspresikan sebagai fungsi dari y dan $\boldsymbol{\theta}$ sebagai berikut:

$$m(y) = \int f(y|\theta)\, \pi(\theta) d\theta \tag{18.19}$$

Perhatikan bahwa komponen $\boldsymbol{f(y)}$ pada Persamaan 18.17 tidak tergantung pada $\boldsymbol{\theta}$ sehingga berlaku

$$p(\theta|y) \propto L(y|\theta)\pi(\theta) \tag{18.20}$$

Persamaan 18.20 adalah persamaan fundamental dari analisis Bayesian. Persamaan ini mengatakan bahwa probabilitas suatu vektor parameter kondisional terhadap data bersifat proporsional terhadap perkalian fungsi likelihood y kondisional terhadap parameter dengan probabilitas *unconditional* parameter. Term pertama di sisi kanan Persamaan 18.20 diestimasi dari data; sedangkan yang term kedua berasal dari *prior belief*.

Aplikasi prinsip dasar yang telah dijelaskan di sini dapat diterapkan ke berbagai persoalan statistik. Namun, dalam buku ini kita hanya akan terapkan pada regresi linear, yang akan diuraikan pada bagian berikutnya.

BAYESIAN LINEAR REGRESSION

Prinsip dasar analisis Bayesian (Persamaan 18.20) dapat diterapkan secara langsung sebagai regresi. Maximum Likelihood Estimation (MLE) adalah suatu pilihan estimator standar yang ada diberbagai software statistik (termasuk STATA). Tantangannya adalah bagaimana melakukan integrasi dengan *prior belief*. Von Neuman (1951) mengusulkan teknik penolakan *sampling* yang digunakan sebagai dasar integrasi. Idenya adalah mengkonstruksi *prior distribution* sampel lain yang lebih mudah diperoleh (biasanya) dari varian-modifikasi suatu distribusi teoretis. Selanjutnya, apakah sampling ini telah memadai atau belum akan digunakan suatu kriteria akseptasi-penolakan tertentu.

Suatu cara yang efektif untuk mencapai tujuan yang disebutkan sebelumnya adalah menggunakan *Monte Carlo Markov Chain* = MCMC (Tanner dan Wong, 1987). MCMC adalah metode untuk menghasilkan nilai-nilai dari suatu kernel transisi sedemikian rupa sehingga nilai-nilai tersebut dapat konvergen dengan target distribusi tertentu. Metode ini akan mensimulasikan Markov Chain dengan target distribusi tertentu sebagai acuan ekuilibrium. Markov Chain adalah serangkaian nilai atau states di mana nilai suatu variabel observasi *t* hanya tergantung pada nilai variabel yang dihasilkan satu titik (*t*-1) sebelumnya. Versi awal dari algoritma MCMC dibangun oleh Metropolis dan Ulam (1949), yang kemudian dikembangkan oleh Hastings (1970) menjadi salah satu metode MCMC yang banyak digunakan: Metropolis-Hasting (MH). Pengembangan lebih lanjut atas MCMC yang cukup penting adalah Gibbs Sampling (Gelfand et al, 1990) dan Adaptive Random Walk MH (Giordani and Kohn, 2010); lihat Chan et al (2019) untuk survei terkini.

Gelman, Gilks, and Roberts (1997) serta Gelman et al (2014) mengusulkan 2 kriteria untuk menilai apakah MCMC telah berhasil melaksanakan tugasnya dengan baik. Kriteria pertama adalah konvergensi yang ditunjukkan dengan statistik akseptasi. Statistik akseptasi yang dapat diterima adalah yang berada di antara nol dan satu. Apabila statistik akseptasi bernilai mendekati nol, berarti MCMC gagal menemukan region-region sampling yang dapat mendekati target distribusi. Sebaliknya, jika mendekati satu, berarti MCMC tidak berhasil beranjak keluar dari titik awal, yang disebabkan rantai terdekat mungkin terlalu jauh. Untuk analisis univariat; statistik akseptasi diharapkan bernilai di kisaran 0,45 sedangkan untuk analisis multivariat diharapkan berada di kisaran 0,23. Kriteria kedua adalah autokorelasi yang minimal, yang ditunjukkan dengan koefisien autokorelasi yang tidak signifikan setelah lag 10.

STATA menyediakan banyak pilihan estimasi Bayes mulai dari regresi linear (*linear regression*) hingga *limited dependent variable* (dapat dilihat pada buku manual). Estimasi dengan menggunakan metode Bayesian akan dilakukan melalui prefiks **bayes:** sebelum perintah regresi.

Contoh 18.3

Kita akan menggunakan dataset yang ada pada situs STATA: auto.dta. Data tersebut dapat diakses dengan perintah **webuse auto.dta** (lalu kita simpan sebagai auto.dta). File ini berisi berbagai variabel (seperti harga, konsumsi BBM, produsen, dan sebagainya) dari 74 merek mobil yang ada di pasar US pada tahun 1978. Kita akan melakukan regresi OLS dengan menggunakan variabel dependen harga (price) dan variabel penjelas: panjang mobil (length) dan

Source	SS	df	MS	Number of obs	=	74
				F(2, 71)	=	16.35
Model	200288930	2	100144465	Prob > F	=	0.0000
Residual	434776467	71	6123612.21	R-squared	=	0.3154
				Adj R-squared	=	0.2961
Total	635065396	73	8699525.97	Root MSE	=	2474.6

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
length	90.21239	15.83368	5.70	0.000	58.64092	121.7839
foreign						
Foreign	2801.143	766.117	3.66	0.000	1273.549	4328.737
_cons	-11621.35	3124.436	-3.72	0.000	-17851.3	-5391.401

TABEL 18.9. Hasil Estimasi Regresi OLS Model Harga Mobil

produsen (dummy variabel: domestic atau luar negeri; foreign=1). Perintah untuk estimasi model harga mobil ini diberikan sebagai berikut **reg price length i.foreign**.

Selanjutnya kita akan melakukan regresi Bayesian dengan model yang *default prior*. Karakter *default prior* ini dapat ditunjukkan melalui penggunaan distribusi normal yang memiliki rata-rata nol dan varians sebanyak 10.000 (distribusi ini juga merupakan salah satu default STATA). Mengingat variabel dependen (price) memiliki skala yang jauh lebih besar dari length dan foreign, maka estimasi Bayesian dilakukan dengan menggunakan *burn-in period* sebesar 5.000 (naik 2 kali lipat dari default: yang 2.500). Perintah untuk melakukan estimasi noninformatif Bayesian ini adalah **bayes, burnin(5000): regress price length i.foreign**.

Dapat kita lihat bahwa meskipun default, penggunaan distribusi normal dengan rata-rata nol dan deviasi standar sebesar 10.000 bersifat *informative prior*. Akibat dari penggunaan prior distribution

```
Burn-in ...
Simulation ...

Model summary
------------------------------------------------------------------------------
Likelihood:
  price ~ regress(xb_price,{sigma2})

Priors:
  {price:length 1.foreign _cons} ~ normal(0,10000)                      (1)
                        {sigma2} ~ igamma(.01,.01)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_price.

Bayesian linear regression                      MCMC iterations  =     15,000
Random-walk Metropolis-Hastings sampling        Burn-in          =      5,000
                                                MCMC sample size =     10,000
                                                Number of obs    =         74
                                                Acceptance rate  =      .2981
                                                Efficiency:  min =      .04829
                                                             avg =      .09653
Log marginal-likelihood = -699.27101            max =      .1983

------------------------------------------------------------------------------
             |                                                 Equal-tailed
             |      Mean   Std. Dev.     MCSE     Median  [95% Cred. Interval]
-------------+----------------------------------------------------------------
price        |
      length |   33.15864   1.818462   .070729   33.15326    29.52992    36.688
             |
     foreign |
     Foreign |    25.3389   96.70311   3.56878     26.432  -162.3509    220.223
       _cons |  -7.111081   103.0203   4.68819  -12.97717  -206.4171   201.5714
-------------+----------------------------------------------------------------
      sigma2 |   7536289    1269867   28515.5    7393458    5479752   1.04e+07
------------------------------------------------------------------------------
Note: Default priors are used for model parameters.
```

TABEL 18.10. Hasil Estimasi Bayesian Default Prior, Regresi Linear Model Harga Mobil

ini besaran koefisien length menjadi hanya sekitar sepertiganya; sementara dampak terhadap slope foreign dan koefisien bahkan lebih besar lagi. Hal ini terjadi karena kita menambahkan informasi bahwa pada awalnya variabel-variabel itu memiliki rata-rata nol dan

deviasi standar sebesar 100. Dari data itu kita dapat mengetahui bahwa kondisi tersebut tidak benar: rata-rata price = 6165 (deviasi standar = 2949) sedangkan rata-rata length = 188 (deviasi standar = 22). Tentu saja, diskrepansi antara prior belief dan data yang dikombinasikan ke dalam *posterior distribution* akan menghasilkan reduksi yang signifikan pada hasil estimasi parameter.

Zellner (1986) mengusulkan suatu prior distribution yang bersifat objektif bagi estimasi regresi linear multivariabel. Prior distribution ini membutuhkan tiga input yakni jumlah koefisien pada regresi, derajat kebebasan atau degree of freedom tertentu, dan varians. Nilai rata-rata diasumsikan nol. Perintah estimasi Bayesian dengan menggunakan jumlah koefisien=3, degree of freedom sebesar 50, dan varians yang dihitung dari sampel adalah **bayes, prior({price:}, zellnersg0(3, 50, {sigma2})) burnin(5000): regress price length i.foreign**.

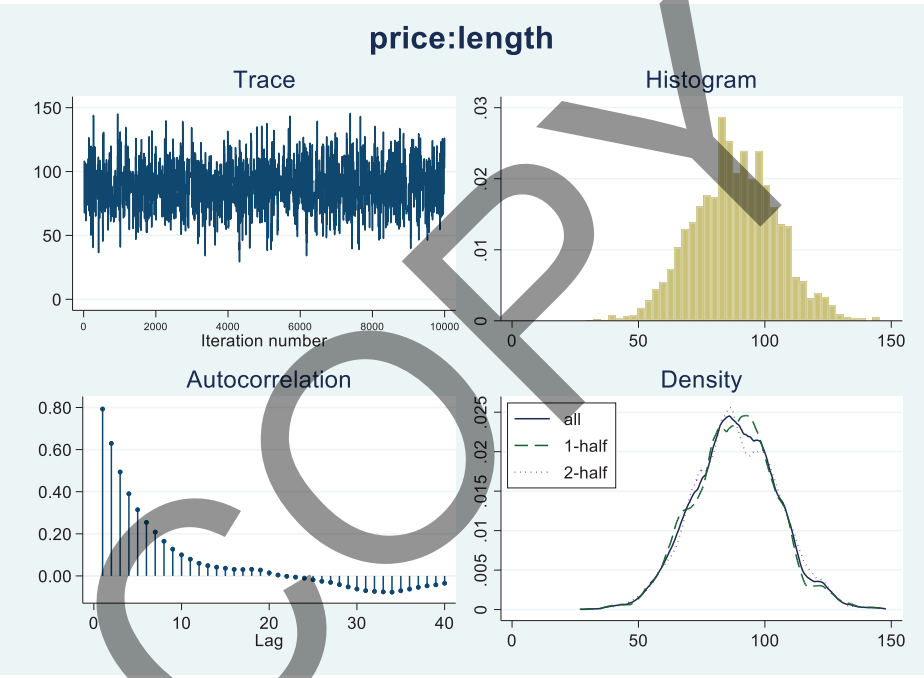
Dapat dilihat dari Tabel 18.10 bahwa dengan menggunakan probabilitas yang objektif, terlihat hasil estimasi menyerupai regresi OLS (Tabel 18.8). Kita akan melakukan evaluasi terhadap hasil yang diperoleh dari estimasi dengan menggunakan Zellner g prior ini (lihat Tabel 18.11). Dari kriteria akseptasi terlihat angka yang sedikit lebih tinggi dari acuan (= 0,234) tetapi masih dapat diterima. Diagnosis lebih lengkap (dalam bentuk visual) dapat dilakukan dengan perintah **bayesgraph diagnostics _all**. Perintah ini akan memberikan grafik trace, histogram, autocorrelation, dan density dari setiap koefisien regresi (dalam contoh ini konstanta, length, dan foreign) serta varians (sigma).

Kita akan menampilkan satu hasil diagnosis Bayesian untuk variabel length. Secara umum, terlihat bahwa estimasi Bayesian sudah cukup memadai. Grafik Trace terlihat cukup acak dengan pola

```
Burn-in ...
Simulation ...

Model summary
------------------------------------------------------------------------------
Likelihood:
  price ~ regress(xb_price,{sigma2})

Priors:
  {price:length 1.foreign _cons} ~ zellnersg(3,50,0,{sigma2})              (1)
                        {sigma2} ~ igamma(.01,.01)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_price.

Bayesian linear regression                      MCMC iterations  =     15,000
Random-walk Metropolis-Hastings sampling        Burn-in          =      5,000
                                                MCMC sample size =     10,000
                                                Number of obs    =         74
                                                Acceptance rate  =      .3339
                                                Efficiency:  min =      .06219
                                                             avg =      .08404
Log marginal-likelihood = -697.80425                         max =       .1383

------------------------------------------------------------------------------
             |                                                Equal-tailed
             |      Mean   Std. Dev.     MCSE     Median  [95% Cred. Interval]
-------------+----------------------------------------------------------------
price        |
      length |   88.31948   16.80984   .669267   87.97144    57.45512   123.0787
             |
     foreign |
     Foreign |   2743.706    817.073   30.3274   2722.814     1244.72   4397.302
       _cons |  -11355.47   3317.063   133.008  -11286.04   -18274.87  -5244.39
-------------+----------------------------------------------------------------
      sigma2 |   6899520    1197187   32192.6    6751245     4971981    9651593
------------------------------------------------------------------------------
Note: Default priors are used for some model parameters.
```

TABEL 18.11. Hasil Estimasi Bayesian Zellner's g Prior, Regresi Linear Model Harga Mobil

histogram dan densitas (*density*) yang kurang lebih normal (sesuai dengan yang diasumsikan baik pada *prior belief* maupun OLS). Grafik autokorelasi juga menunjukkan penurunan yang progresif di mana setelah lag 10, koefisien autokorelasi tidak lagi signifikan.



GAMBAR 18.5. Beberapa Indikator Diagnosis Estimasi Bayesian

Bab 19

Pelaporan Hasil Analisis

Buku ini ditutup dengan bab yang membahas mengenai sistematika serta cara penyampaian analisis statistik ekonometrika yang telah diuraikan di bab-bab sebelumnya. Seperti telah diuraikan di Bab 1 (di Jilid 1), statistik dan ekonometrika adalah suatu tools untuk memverifikasi hipotesis yang dibangun dari proses elaborasi literatur (teori dan empiris) yang dipandu oleh permasalahan penelitian dan logika berpikir. Perpaduan dari semua unsur ini adalah berbentuk desain empiris, yang akan diimplementasikan dengan metodologi yang salah satu alternatifnya adalah statistik-ekonometrika.

Dalam melakukan implementasi desain empiris dan pelaporan, kita harus mengikuti suatu alur yang sistematis. Alur yang sistematis ini sangat diperlukan agar para pembaca mendapat pesan dengan baik menyangkut aktivitas riset-studi yang telah dilakukan. Suatu laporan studi yang menggunakan statistik-ekonometrika biasanya terdiri atas tahapan-tahapan berikut:

a. Analisis Pendahuluan
b. Laporan Regresi: Baseline dan Elaborasi
c. *Robustness Check*
d. Interpretasi dan Analisis

Tahapan-tahapan b dan c sudah merupakan hal yang standar dilakukan, khususnya untuk paper yang akan dikirimkan kepada publikasi terindeks internasional bereputasi.

Pembahasan di bab ini bersifat ilustratif dengan mengambil contoh yang relevan dari satu paper yang telah penulis publikasikan: "The Effect of Competition and Capacity on Intermediation Cost: a Country Level Study"; Ariefianto, Widuri, Abdurachman, dan Trinugroho (2020). Paper ini berisi studi terhadap kinerja

intermediasi perbankan (yang diukur melalui selisih bunga pinjaman dan simpanan; Net Interest Margin). Sebagai variabel penjelas adalah *capacity to lend (2 proxy)*, *competition (3 proxy)*, *characteristics (2 proxy)*, *financial stability (2 proxy)*, dan *structural factors (2 proxy)*. Sementara itu, unit analisis adalah negara (atau konsep yang ekuivalen) yang diperoleh dari Global Financial Development Database Bank Dunia. Dataset ini bersifat panel yang tidak *balanced*; yang terdiri dari 212 negara dan 21 deret waktu (1.889 observasi).

ANALISIS PENDAHULUAN

Analisis pendahuluan (*preliminary analysis*) dilakukan untuk menyampaikan gambaran umum tentang dataset serta teknik persiapan yang digunakan untuk mempersiapkan dataset tersebut. Sebagai suatu gambaran umum, analisis pendahuluan bersifat deskriptif (profiling). Dengan memperhatikan profil (pola) deskriptif datanya, periset dan pembaca hasil riset diharapkan dapat melihat batasan-catatan yang perlu diperhatikan dalam memanfaatkan hasil studi. Tentu saja, batasan-catatan adalah hal terakhir di mana periset harus melakukan upaya semaksimal mungkin atas penanganan profil data yang dapat mendistorsi hasil estimasi (merupakan topik bahasan pada Bab 4 di Jilid 1).

Terdapat dua jenis analisis deskriptif yakni (a) univariat dan (b) asosiatif; sebagaimana telah dibahas pada Bab 4 (di Jilid 1). Di bab ini, kita akan singgung kembali topik tersebut dengan mengambil ilustrasi dari paper yang telah disebutkan diawal. Tabel 19.1 melaporkan statistik deskriptif dari variabel-variabel yang digunakan. Di sini beberapa variabel memiliki isu kemencengan (slope) yang dapat dilihat dari perbedaan yang cukup substansial

antara mean dan media. Variabel-variabel tersebut meliputi LDR, proxy kompetisi (Boone, HINDEX, dan Lerner), serta GDPPERCAP. Namun, karena dataset yang diperoleh telah melewati suatu *cleansing*, maka profil ini diperlakukan sebagai suatu catatan.

	Mean	Median	Maximum	Minimum	Std. Dev.	Skewness	Kurtosis	Jarque-Bera	Probability
NIM	4.514	3.809	18.634	0.125	2.887	1.075	4.335	504.247	0.000
CAR	16.422	15.567	48.600	1.755	5.162	1.603	7.382	2319.867	0.000
LDR	105.002	95.725	879.662	15.335	62.189	5.611	56.469	234932.000	0.000
BOONE	-5.922	0.000	160.662	-5981.630	138.432	-42.626	1840.109	266000000.000	0.000
HINDEX	4.159	0.000	92.500	-8.670	15.766	3.896	17.237	20732.820	0.000
LERNER	7.312	0.000	153.407	-160.869	15.697	1.845	21.608	28324.340	0.000
ROE	12.485	12.606	160.344	-117.673	14.050	-0.641	25.513	40019.980	0.000
OHCTOTA	3.661	2.807	81.900	0.041	3.437	8.395	159.484	1949524.000	0.000
ZSCORE	13.915	12.343	95.279	-0.241	9.345	1.723	8.584	3388.152	0.000
STOCKVLTY	14.228	14.222	99.030	0.000	13.444	1.022	5.233	721.430	0.000
DEPOTOGDP	56.540	44.973	472.049	0.000	52.698	3.214	17.301	19347.920	0.000
GDPPERCAP	17.258	8.313	111.968	0.218	20.610	1.712	5.879	1574.520	0.000

TABEL 19.1. Statistik Deskriptif Univariat pada Ariefianto, Widuri, Abdurachman, dan Trinugroho (2020)

Analisis pendahuluan asosiatif dalam paper ini dilakukan dengan menggunakan korelasi bivariat sederhana (Pearson Correlation). Sebagaimana ditunjukkan pada Tabel 19.2, terdapat korelasi positif yang cukup substansial antara NIM-CAR, NIM-ROE, dan NIM-OHCTOTA. Sedangkan korelasi negatif terlihat pada NIM-LDR, NIM-STOCKVLTY, NIM-DEPOTOGDP, dan NIM-GDPPERCAP. Koefisien-koefisien korelasi ini memberikan indikasi awal tentang hasil estimasi regresi (tanda aljabar dari variabel independen). Dengan menggunakan rule of thumb 0,60–0,70 tampaknya tidak ada potensi isu multikolinearitas di antara variabel independen itu.

Correlation	NIM	CAR	LDR	BOONE	HINDEX	LERNER	ROE	OHCTOTA	ZSCORE	STOCKVLTY	DEPOTOGDP	GDPPERCAP
NIM	1.000											
CAR	0.382	1.000										
LDR	-0.166	-0.141	1.000									
BOONE	-0.023	-0.045	0.009	1.000								
HINDEX	0.054	0.015	-0.034	0.002	1.000							
LERNER	0.051	0.063	-0.052	0.000	0.253	1.000						
ROE	0.359	0.092	-0.155	-0.007	0.058	0.036	1.000					
OHCTOTA	0.591	0.233	-0.087	-0.004	0.046	0.012	0.122	1.000				
ZSCORE	-0.117	-0.087	-0.079	-0.022	0.014	0.009	0.043	-0.166	1.000			
STOCKVLTY	-0374	-0.294	0.168	0.020	-0.082	-0.067	-0.244	-0.182	-0.076	1.000		
DEPOTOGDP	-0.507	-0.153	-0.160	-0.023	-0.050	0.001	-0.168	-0.363	0.275	0.226	1.000	
GDPPERCAP	-0.587	-0.191	0.129	-0.011	-0.031	-0.058	-0.149	-0.393	0.102	0.273	0.552	1.000

TABEL 19.2. Tabel Korelasi pada Ariefianto et al (2020)

ROBUSTNESS CHECK[1]

Setelah melakukan estimasi regresi, biasanya akan dilakukan *robustness check* untuk meyakini bahwa hasil estimasi regresi yang diperoleh tidak gampang berubah. Kesimpulan yang diperoleh adalah masih valid; yaitu dalam rentang perubahan spesifikasi yang wajar. Variasi dalam spesifikasi *robustness check* yang umum dilakukan adalah:

a. Penggunaan estimator yang berbeda, tetapi masih satu "keluarga" atau memiliki tujuan empiris yang sama.
b. *Sequential inclusion*; yaitu memasukkan secara bertahap variabel-variabel yang menjadi fokus studi (*variables of interest*).
c. Penggunaan variabel proxy yang sedikit berbeda; baik variabel independen maupun dependen.

[1] Pembahasan konseptual dalam bagian ini disarikan dari Lu dan White (2014).

Jika hasil studi ternyata bersifat *robust*, maka seharusnya tidak boleh terjadi perubahan yang terlalu substansial atas hasil estimasi. Kriteria substansial ini bersifat relatif; yaitu sebagai *rule of thumb* dapat digunakan (a) tidak adanya perubahan tanda aljabar dan (b) koefisien tidak berbeda lebih dari 20% dari baseline (jika estimator yang digunakan masih dalam "keluarga" yang sama).

Dalam paper penulis, *robustness check* dilakukan dengan menggunakan ketiga metode yang telah disebutkan sebelumnya. Tabel 19.3 menunjukkan hasil estimasi regresi baseline dengan

No	Variable/Proxies	FE			RE			Pool		
		1	2	3	4	5	6	7	8	9
1	CAR	0.053***	0.053***	0.053***	0.068***	0.067***	0.068***	0.096***	0.097***	0.097***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
2	LDR	-0.002	-0.002	-0.002	-0.002**	-0.002**	-0.002**	-0.003***	-0.003***	-0.003***
		(0.137)	(0.136)	(0.134)	(0.014)	(0.013)	(0.014)	(0.000)	(0.000)	(0.000)
3	BOONE	0.000			0.000			0.000		
		(0.000)			(0.119)			(0.000)		
	HINDEX		-0.003			-0.003			0.000	
			(0.213)			(0.174)			(0.894)	
	LERNER			-0.001			0.000			0.001
				(0.602)			(0.907)			(0.720)
4	OHCTOTA	0.119**	0.119**	0.119**	0.140***	0.141***	0.140***	0.285***	0.286***	0.286***
		(0.046)	(0.046)	(0.046)	(0.000)	(0.000)	(0.000)	(0.003)	(0.003)	(0.003)
5	ROE	0.022***	0.022***	0.022***	0.025***	0.025***	0.025***	0.042***	0.042***	0.042***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
6	STOCKVLTY	0.005	0.005	0.005	0.001	0.002	0.001	-0.020***	-0.020***	-0.020***
		(0.314)	(0.289)	(0.316)	(0.676)	0.646	0.679	(0.000)	(0.000)	(0.000)
7	ZSCORE	0.042**	0.042**	0.042	0.027***	0.027***	0.027***	-0.001	-0.001	-0.001
		(0.019)	(0.019)	(0.019)	(0.000)	(0.000)	(0.000)	(0.910)	(0.922)	(0.921)
8	DEPOTOGDP	-0.009**	-0.009**	-0.009**	-0.011***	-0.011***	-0.011***	-0.009***	-0.009***	-0.009***
		(0.026)	(0.026)	(0.027)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
9	GDPPERCAP	-0.051***	-0.051***	-0.051***	-0.051***	-0.051***	-0.051***	-0.037***	-0.037***	-0.037***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
R2		0.844	0.844	0.844	0.282	0.281	0.281	0.624	0.624	0.624
F stat		66.393***	66.415***	66.306	81.867***	81.648***	81.481***	46.969***	46.549***	46.569***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
FE Test (Chi Square stat)		1657.034***	58.981***	1656.305***						
		(0.000)	(0.000)	(0.000)						
LM Test (RE Test)*					54.839***	54.882***	54.948***			
*Cross Section Standardized Honda (Honda, 1991)					(0.000)	(0.000)	(0.000)			
Hausman Test					117.401**	118.18***	118.186***			
					(0.000)	(0.000)	(0.000)			

TABEL 19.3. Robustness Check; Alternatif Estimator

menggunakan tiga estimator: yaitu pooled OLS, Fixed Effect, dan Random Effect. Pengujian Wald, LM Test, dan Hausman mengindikasikan preferensi pada spesifikasi FE. Namun demikian, seperti dapat dilihat pada Tabel tersebut, paling tidak tanda aljabar dari variabel yang menjadi fokus studi: yaitu CAR, LDR, dan Competition Proxy tidak mengalami perubahan.

Pelaksanaan robustness check dengan menggunakan *alternative proxy* juga dapat dilihat pada Tabel 19.3. Penggunaan berbagai estimator yang masih dalam satu "keluarga" ini yang meliputi Pooled OLS, FE, dan RE, tampaknya tidak menyebabkan perubahan koefisien yang substantial; paling tidak jika dilihat dari tanda aljabar. Konsistensi tanda aljabar pada hasil estimasi juga terlihat dari variabel kontrol, sehingga dapat dikatakan bahwa hasil estimasi cukup robust.

Robustness check dengan sequential inclusion disajikan pada Tabel 19.4 untuk competition proxy: Boone. Kolom pertama pada setiap panel merupakan regresi baseline. Dapat dilihat di sini bahwa tersedia contoh untuk variabel LDR dan Boone tetapi untuk CAR tidak ada; tetapi kita memperoleh koefisien regresi yang secara kualitatif tidak berbeda dengan baseline (model 10 versus model 11). Kesimpulan yang serupa juga akan diperoleh jika kita membandingkan model di mana variabel LDR "dikeluarkan" versus baseline (model 12 versus 10).

Berdasarkan hasil pengujian robustness check yang telah diuraikan sebelumnya, dapat dikatakan bahwa hasil estimasi dengan menggunakan spesifikasi FE bisa dikembangkan lebih lanjut. Hal ini akan dibahas pada bagian berikutnya.

No	Variable/Proxies	FE				RE				Pool			
		10	11	12	13	14	15	16	17	18	19	20	21
1	CAR	0,053***		0,055***	0,053***	0,068***		0,070***	0,068***	0,096***		0,099***	0,097***
		(0.000)		(0.000)	(0.000)	(0.000)		(0.000)	(0.000)	(0.000)		(0.000)	(0.000)
2	LDR	-0.002	-0.003		-0.002	-0.002**	-0,003***		-0.002**	-0.003***	-0.003***		-0.003***
		(0.137)	(0.053)		(0.136)	(0.014)	(0.001)		(0.014)	(0.000)	(0.000)		(0.000)
3	BOONE	0.000	0.000	0.000		0.000	0.000	0.000		0.000	-0.001	0.000	
		(0.000)	(0.000)	(0.000)		(0.119)	(0.113)	(0.119)		(0.000)	(0.000)	(0.000)	
4	OHCTOTA	0.119**	0.126**	0.120**	0.120**	0.140***	0.150***	0.142***	0.140***	0.285***	0.307***	0.289***	0.286***
		(0.046)	(0.040)	(0.045)	(0.046)	(0.000)	(0.000)	(0.000)	(0.000)	(0.003)	(0.002)	(0.003)	(0.003)
5	RDE	0.022***	0.021***	0.023***	0.022***	0.025***	0.023***	0.025***	0.025***	0.042***	0.041***	0.043***	0.042***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
6	STOCKVLTY	0.005	0.003	0.004	0.005	0.001	-0.002	0.001	0.001	-0.020***	-0.029***	-0.021***	-0.020***
		(0.314)	(0.552)	(0.346)	(0.314)	(0.675)	(0.551)	(0.839)	(0.683)	(0.000)	(0.000)	(0.000)	(0.000)
7	ZSCORE	0.042**	0.049***	0.042**	0.042**	0.027***	0.034***	0.027***	0.027***	-0.001	-0.002	-0.001	-0.001
		(0.019)	(0.009)	(0.019)	(0.019)	(0.000)	(0.000)	(0.000)	(0.000)	(0.910)	(0.735)	(0.907)	(0.920)
8	DEPOTOGDP	-0.009**	-0.009**	-0.009**	-0.009**	-0.011***	-0.011***	-0.010***	-0.011***	-0.009***	-0.009***	-0.008***	-0.009***
		(0.026)	(0.028)	(0.031)	(0.026)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
9	GDPPERCAP	-0.051***	-0.043***	-0.053***	-0.051***	-0.051***	-0.051***	-0.051***	-0.051***	-0.037***	-0.038***	-0.039***	-0.037***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
R2		0.844	0.841	0.844	0.844	0.282	0.253	0.280	0.281	0.624	0.599	0.622	0.624
F stat		66.393***	65.312***	66.793***	66.804***	81.867***	79.571***	91.412***	91.953***	346.969***	350.321***	385.999***	390.071***
		0.844	0.841	0.844	0.844	0.282	0.253	0.280	0.281	0.624	0.599	0.622	0.624
FE Test (Chi Square LR Ratio)		1657.034***	1744.332***	1668.211***	1656.141***								
		(0.000)	(0.000)	(0.000)	(0.000)								
LM Test (RE Test)*						54.839***	54.929***	54.739***	54.815***				
*Standardized Honda (Honda, 1991)						(0.000)	(0.000)	(0.000)	(0.000)				
Hausman Test						117.401***	122.787***	121.709***	118.300***				
						(0.000)	(0.000)	(0.000)	(0.000)				

TABEL 19.4. Robustness Check; Sequential Inclusion

LAPORAN REGRESI: BASELINE DAN ELABORASI

Standar penelitian saat ini, terutama yang masuk dalam publikasi terindeks internasional bereputasi, umumnya bersifat elaborasi (*elaborated*). Di samping model baseline yang merupakan bentuk regresi pertama untuk menguji hipotesis-hipotesis, biasanya laporan riset juga akan memasukkan hasil pengembangan. Tujuan dari pengembangan adalah melihat tingkat relevansi hipotesis (model regresi baseline) pada berbagai konteks dan perspektif yang berbeda.

Pengembangan model (elaborasi model) biasanya dilakukan dengan

a. Memasukkan berbagai kategori sampel (negara, perusahaan, daerah, dan lainnya)
b. Memasukkan adanya event tertentu (adanya krisis, perubahan rezim regulasi, pergantian ownership unsur sampel, dan sebagainya).

Secara substansial sebenarnya elaborasi model juga merupakan *robustness check* yang lebih digeneralisasi. Di sini kita mencoba melihat apakah hipotesis masih relevan (dalam artian koefisien hasil regresi tidak berubah secara substantial) jika konteksnya "sedikit" diubah.

Dalam paper ini model elaborasi regresi dilakukan dengan memasukkan dampak krisis tahun 2008-2009 (event dummy 2008 = 1) dan klasifikasi negara (berpenghasilan rendah, menengah, dan tinggi). Tabel 19.5 menyajikan suatu contoh elaborasi regresi dengan memasukkan klasifikasi negara. Perhatikan bahwa dalam melakukan elaborasi regresi, variabel gdppercap kita keluarkan karena variabel ini akan mengalami masalah multikolinearitas dengan variabel kategori negara. Setelah dilakukan estimasi kembali, yaitu model 55 sampai dengan 63, ternyata koefisien ini signifikan dan sejalan dengan hasil regresi gdppercap. Terlihat bahwa rata-rata *unconditional* (intersep) NIM pada negara berpenghasilan tinggi adalah lebih rendah (sekitar –0,81 (= –1,67 – (–0,86); lihat model 55) dibandingkan negara berpenghasilan rendah.

TABEL RANGKUMAN REGRESI

Kita dapat menyimpan hasil estimasi yang dilakukan dengan perintah **estimates store** (nama output). Kemudian setelah melakukan

No	Variable/Proxies	FE			RE			Pool		
		55	56	57	58	59	60	61	62	63
1	CAR	0.066***	0.065***	0.066***	0.081***	0.080***	0.081***	0.095***	0.096***	0.095***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
2	LDR	-0.002	-0.002	-0.002	-0.002**	-0.002**	-0.002**	-0.002***	-0.002***	-0.002***
		(0.216)	(0.217)	(0.215)	(0.046)	(0.044)	(0.045)	(0.000)	(0.000)	(0.000)
3	BOONE	0.000			0.000			0.000		
		(0.000)			(0.000)			(0.000)		
	H-INDEX		-0.004			-0.003			0.000	
			(0.167)			(0.179)			(0.896)	
	LERNER			-0.001			0.001			0.001
				(0.792)			(0.733)			(0.527)
4	OHCTOTA	(0.095)	(0.095)*	(0.095)*	0.113***	0.113***	0.113***	0.208**	0.208**	0.208**
		(0.060)	(0.061)	(0.061)	(0.000)	(0.000)	(0.000)	(0.011)	(0.011)	(0.011)
5	ROE	0.027***	0.027***	0.027***	0.029***	0.029***	0.029***	0.039***	0.039***	0.039***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
6	STOCKVLTY	-0.001	-0.001	-0.001	-0.006	-0.006	-0.006	-0.017***	-0.017***	-0.017***
		(0.748)	(0.825)	(0.746)	(0.162)	(0.172)	(0.164)	(0.000)	(0.000)	(0.000)
7	ZSCORE	0.048**	0.048**	0.048**	0.035***	0.035***	0.035***	0.012*	0.012*	0.012*
		(0.043)	(0.042)	(0.042)	(0.000)	(0.000)	(0.000)	(0.054)	(0.054)	(0.052)
8	DEPOTOGDP	-0.018***	-0.018***	-0.018***	-0.019***	-0.019***	-0.019***	-0.014***	-0.014***	-0.014***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
9	Dummy Low Middle	-0.859***	-0.876***	-0.857***	-0.930***	-0.939***	-0.931***	-1.034***	-1.029***	-1.030***
		(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
10	Dummy Up Middle	-1.339***	-1.348***	-1.339***	-1.567***	-1.570***	-1.569***	-1.694***	-1.693***	-1.695***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
11	Dummy High	-1.671***	-1.663***	-1.670***	-2.378***	-2.327***	-2.382***	-2.860***	-2.858***	-2.856***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
R2		0.835	0.835	0.835	0.336	0.336	0.335	0.644	0.644	0.644
F stat		49.569***	49.623***	49.479***	61.649***	61.548***	61.443***	220.131***	219.961***	220.018***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
FE Test (Chi Square stat)		1039.125***	1041.012***	1037.505***						
		(0.000)	(0.000)	(0.000)						
LM Test (RE Test)*					43.608***	43.636***	43.761***			
*Cross Section Standardized Hinda (Honda, 1991)					(0.000)	(0.000)	(0.000)			
Hausman Test					80.491***	81.799***	81.539***			
					(0.000)	(0.000)	(0.000)			

TABEL 19.5. Elaborasi Regresi; Klasifikasi Negara

beberapa estimasi, mungkin kita ingin membandingkan output-output tersebut. Hal ini bisa dilakukan dengan perintah **estimates table (nama output 1) (nama output 2) … , stats(st1 st2 …)**. Term st1 dan seterusnya adalah nama statistik dalam kodifikasi STATA. Nama statistik tersebut dapat dilihat pada help file dari command terkait, baik estimasi maupun *post estimation*.

Kita akan mengilustrasikan penggunaannya dengan file Real_ Rate_IDN.dta.Misalkan kita akan melakukan tiga model regresi (OLS) yang berbeda hanya pada variabel penjelas sebagai berikut:

a. Model 1, REAL_RATE VOL_INF GINI RGDP_L FIN_DEPTH
b. Model 2, REAL_RATE GINI RGDP_L FIN_DEPTH
c. Model 3, REAL_RATE VOL_INF RGDP_L FIN_DEPTH

Di samping koefisien (*slope*) variabel, kita juga ingin mengetahui kriteria statistik komparatif: R2, AIC, dan BIC. Perintah regresi-regresi ini[2] dimasukkan secara sekuensial pada *command window* sebagai berikut

```
quietly reg REAL_RATE VOL_INF GINI RGDP_L FIN_DEPTH
estimates store Model1
quietly reg REAL_RATE GINI RGDP_L FIN_DEPTH
estimates store Model2
quietly reg REAL_RATE VOL_INF RGDP_L FIN_DEPTH
estimates store Model3
estimates table Model1 Model2 Model3, stats(2 aic bic)
```

Pelaksanaan perintah tersebut akan memberikan hasil sebagaimana terlihat pada Tabel 19.6.

Wada (2014) telah membuat suatu program (ado.file) yang bernama **outreg2** yang dapat sangat membantu dalam hal pelaporan hasil estimasi ekonometrika dengan menggunakan STATA. Seperti dijelaskan sebelumnya, dalam praktek sudah merupakan hal yang biasa untuk melakukan estimasi dan "membandingkan" banyak regresi. Outreg2 adalah suatu program yang menyimpan dan mengkompilasi hasil-hasil estimasi tersebut dalam bentuk output file tertentu (misalnya, MS Excel). Selanjutnya, hasil kompilasi tersebut

[2] Kita menggunakan quietly sebelum setiap regresi untuk mencegah hasil regresi terlihat di output window.

Variable	Model1	Model2	Model3
VOL_INF	.3170838		.34570158
GINI	-.26674957	-.70795817	
RGDP_L	19.626459	20.348091	12.864056
FIN_DEPTH	-.04802497	-.09549383	-.02778758
_cons	-299.33392	-292.14773	-202.46682
r2	.51274364	.17413377	.49426708
aic	85.769214	93.79442	84.476362
bic	90.491408	97.572176	88.254118

TABEL 19.6. Tabel Rangkuman Output

akan dibuat dalam bentuk tabel dan disertai dengan fitur seperti tanda */**/*** yang menunjukkan signifikansi. Ringkasnya, outreg2 membantu pelaporan hasil ekonometrika ke dalam bentuk yang *publishing quality ready*. Mengingat perintah ini bersifat community developed, maka pembaca harus terlebih dahulu melakukan instalasinya dengan perintah **ssc install outreg2**.

Kita kembali menggunakan contoh yang diberikan sebelumnya dan kita dapat menuliskan instruksi berikut pada new do.file (yang disimpan dengan nama real_rate_outreg2.do):

```
quietly reg REAL_RATE VOL_INF GINI RGDP_L FIN_DEPTH
outreg2 using real_rate.xls, replace ctitle(Model 1)
quietly reg REAL_RATE GINI RGDP_L FIN_DEPTH
outreg2 using real_rate.xls, append ctitle(Model 2)
quietly reg REAL_RATE VOL_INF RGDP_L FIN_DEPTH
outreg2 using real_rate.xls, append ctitle(Model 3)
```

Eksekusi do.file (dengan mengklik pada window do.file, **tools/ execute selection** atau **ctrl D)** akan menghasilkan output sebagai berikut:

```
. do "C:\Users\Windows 8.1\OneDrive - Bina Nusantara University\Kerja\Scholarly\HAKI\Aplikasi STA
> TA\real_rate_outreg2.do"

. quietly reg REAL_RATE VOL_INF GINI RGDP_L FIN_DEPTH

. outreg2 using real_rate.xls, replace ctitle(Model 1)
real_rate.xls
dir : seeout

. quietly reg REAL_RATE GINI RGDP_L FIN_DEPTH

. outreg2 using real_rate.xls, append ctitle(Model 2)
real_rate.xls
dir : seeout

. quietly reg REAL_RATE VOL_INF RGDP_L FIN_DEPTH

. outreg2 using real_rate.xls, append ctitle(Model 3)
real_rate.xls
dir : seeout

.
end of do-file
```

TABEL 19.7. Output Window Eksekusi Program File real_rate_outreg2.do

Dengan mengklik tulisan real_rate.xls pada output window tersebut, akan terbuka file real_rate.xls yang memuat kompilasi dari hasil regresi seperti yang terlihat pada Tabel 19.7.

Tentu saja, tampilan dalam bentuk Tabel 19.8 akan sangat membantu. Di sini hasil regresi telah disusun dalam bentuk tabel yang teratur disertai fitur evaluasi statistik yang dibutuhkan. Bentuk tabelnya mengikuti standar pelaporan publikasi ilmiah di mana setelah koefisien di bawahnya dilaporkan standar error dari koefisien tersebut. Jika koefisien dimaksud signifikan dengan menggunakan α tertentu, maka akan diberikan tanda *,** dan *** sesuai p value yang relevan: < 0,1, < 0,05, atau < 0,01. Tampilan seperti ini juga

	(1)	(2)	(3)
VARIABLES	Model 1	Model 2	Model 3
VOL_INF	0.317***		0.346***
	(0.102)		(0.0923)
GINI	-0.267	-0.708	
	(0.366)	(0.425)	
RGDP_L	19.63	20.35	12.86
	(17.89)	(22.50)	(15.05)
FIN_DEPTH	-0.0480	-0.0955	-0.0278
	(0.286)	(0.360)	(0.280)
Constant	-299.3	-292.1	-202.5
	(269.3)	(338.7)	(230.5)
Observations	19	19	19
R-squared	0.513	0.174	0.494
Standard errors in parentheses			
*** p<0.01, ** p<0.05, * p<0.1			

TABEL 19.8. Output File (real_rate.xls) Eksekusi Program File real_rate_outreg2.do

akan sangat membantu dalam analisis. Misalnya, kita dapat melihat begitu variabel VOL_INF dikeluarkan dari spesifikasi model yang lengkap (model 1), maka R2 langsung jatuh dari 0,513 menjadi 0,174 pada model 2.

INTERPRETASI DAN ANALISIS

Setelah melakukan estimasi regresi baseline, robustness check, dan elaborasi regresi kita dapat melakukan interpretasi serta analisis (banyak juga yang disebut sebagai segmen hasil dan diskusi). Pelaksanaan interpretasi dan analisis ini mencakup hal-hal sebagai berikut:

a. Membandingkan koefisien variables of interests dengan konstruksi hipotesis
b. Melakukan evaluasi terhadap variabel kontrol
c. Membahas kelaikan suai (*goodness of fit*) dan hasil pengujian spesifikasi lainnya yang relevan
d. Membahas model elaborasi

Koefisien variables of interests akan dibandingkan dengan konstruksi hipotesis untuk melihat apakah hasil estimasi telah sesuai; yaitu sejalan dengan teori-literatur empiris yang ada. Jika ada perbedaan dengan konstruksi hipotesis, maka peneliti harus melakukan review terhadap spesifikasi-estimasi model terlebih dahulu. Hal ini harus dilakukan untuk memastikan bahwa hasil tersebut bukan akibat dari sesuatu yang bersifat teknis ekonometris. Apabila spesifikasi-estimasi telah dijalankan sesuai standar, maka hasil yang berbeda tersebut dapat dikatakan reliable. Di sini yang kemudian dilakukan adalah menganalisis secara fundamental serta membangun argumen atas faktor yang menyebabkan diskrepansi tersebut.

Evaluasi terhadap variabel kontrol harus dilakukan untuk meyakini bahwa model regresi tidak mengalami masalah misspesifikasi. Meskipun bukan merupakan fokus utama studi, variabel-variabel kontrol hendaknya memiliki tanda aljabar yang sama dengan studi-studi serupa. Perbedaan yang substantial juga akan memerlukan review terhadap pelaksanaan spesifikasi dan estimasi regresi.

Kelaikan suai (*goodness of fit*) adalah suatu evaluasi syarat untuk meyakini bahwa regresi yang diestimasi telah memiliki mutu yang memadai. Mutu yang memadai dianggap telah tercapai

apabila variabel independen dan spesifikasi regresi telah mampu menjelaskan varians pada variabel dependen sesuai dengan standar teknik ekonometrika yang dipakai. Evaluasi kelaikan suai tergantung pada teknik ekonometrika yang digunakan; yaitu untuk keluarga least squares adalah R2 dan Uji F; sedangkan untuk estimator maximum likelihood akan menggunakan nilai Chi Square (Uji LM). Teknik yang lebih elaborate akan membutuhkan evaluasi yang lebih kompleks lagi; biasanya disebut uji spesifikasi. Misalnya, teknik data panel dinamis (DPD) akan membutuhkan uji *overidentification* dan uji relevansi instrumen sebagai persyaratan kelayakan spesifikasi.

Setelah membahas *baseline*, *robustness check*, kelaikan-suai, dan spesifikasi, maka selanjutnya dapat dilakukan pembahasan tentang model elaborasi. Seperti halnya pembahasan model baseline, analisis pada model elaborasi juga dilakukan dengan melihat alignment pada konstruksi hipotesis. Setiap diskrepansi harus ditindaklanjuti dengan mereview spesifikasi dan estimasi regresi terlebih dahulu sebelum mengemukakan argumen yang bersifat fundamental.

Langkah-langkah yang telah diuraikan sebelumnya dapat berlangsung sebagai suatu bentuk *looping* (*iterative*). Dalam praktek, riset ketidakselarasan dengan hipotesis yang telah dibangun sangat mungkin berasal dari kurang memadainya-kesalahan bentuk spesifikasi, seperti *omitted variable*, *redundant variable*, atau bentuk fungsional. Periset harus melakukan estimasi kembali (dan mungkin beberapa kali) dan setelah melakukan respesifikasi baru menemukan model yang *aligned*. Apabila evaluasi spesifikasi-estimasi telah dilakukan dan periset merasa yakin telah dilakukan sesuai pedoman, maka perlu dilakukan studi literatur lebih lanjut.

Dalam disiplin ilmu ekonomi-bisnis, banyak teori (dan implikasi hipotesisnya) belum mencapai konsensus bahkan banyak yang masih

bersifat ambigu. Dengan demikian, hasil yang diperoleh mungkin mendukung hipotesis alternatif. Ingat kembali bahwa analisis data akan selalu masuk ke dalam salah satu tipe kesalahan statistik: tipe satu atau tipe dua. Kita tidak akan pernah mengetahui dengan presisi 100%, sifat dari suatu fenomena data yang ada. Hal yang dapat dilakukan adalah mengoptimalisasi probabilitas kesalahan (melalui pemilihan teknik ekonometrika yang sesuai) dan membangun konstruksi logika yang bersifat *sound* dan *robust* untuk mendukung argumen.

DAFTAR PUSTAKA

1. Acock, A. C., (2018), Gentle Introduction to Stata (6th Edition), Stata Press.
2. Acock, A. C., (2013), Discovering Structural Equation Modelling Using STATA, Stata Press.
3. Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The Colonial Origins of Comparative Development: An Empirical Investigation. *American Economic Review*, *91*(5), 1369-1401.
4. Adkins, L. C., & Hill, R. C. (2011). Using Stata for Principles of Econometrics. John Wiley & Son.
5. Adkins, L. C., & Gade, M. N. (2012). Monte Carlo Experiments Using Stata: a Primer with Examples. In *30th Anniversary Edition*. Emerald Group Publishing Limited.
6. Ahrens, A, 2019, An Introduction to Machine Learning with STATA, Presentation in Italian User Group Meeting.
7. Ahrens, A., Hansen, C. B., & Schaffer, M. E. (2018). LASSOPACK and PDSLASSO: Prediction, Model Selection and Causal Inference with Regularized Regression. In *London Stata Conference 2018* (No. 12). Stata Users Group.
8. Ahrens, A., Aitken, C., & Schaffer, M. E. (2020). Using Machine Learning Methods to Support Causal Inference in Econometrics. In *Behavioral Predictive Modelling in Economics* (pp. 23-52). Springer, Cham.
9. Ahrens, A., Hansen, C. B., & Schaffer, M. E. (2020). lassopack: Model Selection and Prediction with Regularized Regression in Stata. *The Stata Journal*, *20*(1), 176-235.
10. Alexander, C., 2008, Practical Financial Econometrics, John Wiley & Sons, New York.

11. Amemiya, T. (1981). Qualitative Response Models: A Survey. *Journal of Economic Literature*, *19*(4), 1483-1536.
12. Anderson, T. W., & Hsiao, C. (1982). Formulation and Estimation of Dynamic Models Using Panel Data. *Journal of Econometrics*, *18*(1), 47-82.
13. Anderson D. R, D. J. Sweeney, Williams, T. A., Camm, J.D., and Cochran, J.J., 2017, Statistics For Business and Economics, Thomson South Western, Ohio, 13th Edition.
14. Ariefianto, M. D., 2012, Ekonometrika: Esensi dan Aplikasi dengan menggunakan Eviews, Penerbit Erlangga.
15. Ariefianto, M. D., & Trinugroho, I. (2020a). The Role of Structural Factors in Real Interest Rate Behaviour: A Cross-Country Study. *Jurnal Keuangan dan Perbankan*, *24*(3).
16. Ariefianto,M.D. dan Trinugroho, I, (2020b), Commercial Bank's Loan Loss Provisioning Behaviour in Reconciling Era: Evidence from Selected Emerging Economies, Working Paper.
17. Ariefianto, M. D., Widuri, R., Abdurachman, E., & Trinugroho, I. (2020). The Effect of Competition and Capacity on Intermediation Cost: A Country Level Study. *International Journal of Economics & Management*, *14*(1).
18. Arellano, M. (1987). Computing Robust Standard Errors for Within-Groups Estimators. *Oxford Bulletin of Economics and Statistics*, *49*(4), 431-434.
19. Arellano, M., & Bond, S. (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *The review of Economic Studies*, *58*(2), 277-297.
20. Arellano, M., & Bover, O. (1995). Another Look at the Instrumental Variable Estimation of Error-Components Models. *Journal of Econometrics*, *68*(1), 29-51.
21. Athey, S. (2017). Beyond Prediction: Using Big Data for Policy Problems. *Science*, *355*(6324), 483-485.
22. Athey, Susan (2018), "The Impact of Machine Learning on Economics" http://www.nber.org/chapters/c14009.pdf.

23. Athey. S. and G. Imbens (2019), "Machine Learning Methods Economists Should Know About."
24. Azevedo, J. P., 2003. "FACTORTEST: Stata Module to Perform Tests for Appropriateness of Factor Analysis," Statistical Software Components S436001, Boston College Department of Economics, Revised 27 Aug 2006.
25. Bailey, R. E. (2005). *The economics of Financial Markets*. Cambridge University Press.
26. Balestra, P., & Varadharajan-Krishnakumar, J. (1987). Full information Estimations of a System of Simultaneous Equations with Error Component Structure. *Econometric Theory*, 223-246.
27. Baltagi, B. H. (1981). Simultaneous Equations with Error Components. *Journal of Econometrics*, *17*(2), 189-200.
28. Baltagi, B., 2011, Econometric Analysis of Panel Data, John Wiley & Sons, New York.
29. Banarjee A., Dolado, J.J., Galbraith, J.W. dan D.F. Hendry, 1993, Co-integration, Error-Correction and The Econometric Analysis of Non Stationary Data, Advanced Texts in Econometrics, Oxford University Press, Oxford.
30. Baum, C., Schaffer, M., & Stillman, S. (2003). IVREG2: Stata Module for Extended Instrumental Variables/2SLS and GMM Estimation (Statistical Software Component No. S425401). *Boston College*.
31. Baum, Christopher F., 2006, *An Introduction to Modern Econometrics Using Stata*. Stata Press.
32. Baum, C. (2013a), Dynamic Panel Data Estimators, STATA Discussion Paper.
33. Baum, C. (2013b), Simulation for Estimation and Testing, STATA Discussion Paper.
34. Baum, C. (2014), ARCH and MGARCH Models, STATA Discussion Paper.
35. Baum, C. F. (2016). An Introduction to Stata Programming, Stata Press.
36. Baum, C. (2018). FCSTATS: Stata Module to Compute Time Series Forecast Accuracy Statistics.

37. Bauwens, L., Laurent, S., & Rombouts, J. V. (2006). Multivariate GARCH Models: a Survey. *Journal of Applied Econometrics*, *21*(1), 79-109.
38. Belloni, A., Chernozhukov, V., & Wang, L. (2011). Square-Root Lasso: Pivotal Recovery of Sparse Signals via Conic Programming. *Biometrika*, *98*(4), 791-806.
39. Belloni, A., V. Chernozhukov, and C. Hansen. (2014a). Inference on Treatment Effects After Selection Among High-Dimensional Controls. *Review of Economic Studies* 81: 608-650.
40. Belloni, A., Chernozhukov, V., & Hansen, C. (2014b). High-dimensional Methods and Inference on Treatment and Structural Effects in Economics. *Journal of Economic Perspectives*, *28*(2), 29-50.
41. Ben-Akiva, M., & Lerman, S. R. (1985). Discrete Choice Analysis: Theory and Application to Travel Demand, MIT Press.
42. Blackburne III, E. F., & Frank, M. W. (2007). Estimation of Nonstationary Heterogeneous Panels. *The Stata Journal*, *7*(2), 197-208.
43. Blundell, R., & Bond, S. (1998). Initial Conditions and Moment Restrictions in Dynamic Panel Data Models. *Journal of Econometrics*, *87*(1), 115-143.
44. Bollerslev, T., 1986," Generalized Autoregressive Conditional Heterocedasticity", *Journal of Econometrics*, Vol. 31, hal. 307-327.
45. Bollerslev, T., Engle, R. F., & Wooldridge, J. M. (1988). A capital asset Pricing Model with Time-Varying Covariances. *Journal of Political Economy*, *96*(1), 116-131.
46. Bollerslev, T. (1990). Modelling the Coherence in Short-Run Nominal Exchange Rates: a Multivariate Generalized ARCH Model. *The Review of Rconomics and Statistics*, 498-505.
47. Bound, J., Jaeger, D. A., & Baker, R. (1993). *The cure can be worse than the disease: A cautionary tale regarding instrumental variables* (No. t0137). National Bureau of Economic Research.
48. Breitung, J., & Das, S. (2005). Panel Unit Root Tests Under Cross-Sectional Dependence. *Statistica Neerlandica*, *59*(4), 414-433.
49. Breusch, T. S., & Pagan, A. R. (1980). The Lagrange multiplier test and its applications to model specification in econometrics. *The Review of Rconomic Studies*, *47*(1), 239-253.

50. Brooks, C., 2014, Introductory Econometrics for Finance, Cambrige University Press, 3rd Edition, Cambridge.
51. Cameroon, A. C.. & Triverdi, P. K, 2005, *Microeconometrics: Methods and Applications*, Cambridge, New York.
52. Cameron, A. C. (2019). Machine Learning Methods in Economics. *Machine Learning*, (1/67).
53. Campbell, J. Y., Lo, A. W., dan A. C. MacKinlay, 1997, The Econometrics of Financial Market, *Princeton University Press*, New Jersey.
54. Carnot, N., Koen, V dan B. Tissot, 2005, Economic Forecasting, Palgrave Mac Millan, London.
55. Carsey, T. M., & Harden, J. J. (2013). *Monte Carlo Simulation and Resampling Methods for Social Science*. Sage Publications.
56. Chan, J., Koop, G., Poirier, D. J., & Tobias, J. L. (2019). *Bayesian Econometric Methods* (Vol. 7). Cambridge University Press.
57. Charemza, W, W., dan D.F. Deadman, (1992), New Directions in Econometric Practices, Edward-Elgar, London.
58. Chen, J., & Chen, Z. (2008). Extended Bayesian Information Criteria for Model Selection with Large Model Spaces. *Biometrika*, *95*(3), 759-771.
59. Chernozhukov, V., Hansen, C., & Spindler, M. (2015). Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review*, *105*(5), 486-90.
60. Chiang, A. C dan Wainwright, K., 2005, Fundamental Methods of Mathematical Economics, Mc Graw-Hill, Singapore.
61. Clemens, M. P. dan D. F. Hendry, 1998, Forecasting Econometric Time Series, Cambridge University Press, Cambridge.
62. Cleff, T. (2019). *Applied Statistics and Multivariate Data Analysis for Business and Economics*. Springer International Publishing.
63. Danielson, J. (2011). Financial Risk Forecasting: The Theory and Practice of Forecasting Market Risk with Implementation in R and Matlab, Wiley.
64. Davidson, R., & MacKinnon, J. G. (1993). Estimation and Inference in Econometrics. Oxford University Press.

65. Diebold, F. X. (2019), *Econometric Data Science: A Predictive ModeLling Approach*, Department of Economics, University of Pennsylvania.
66. Dickey, D. A. dan Fuller, W. A., "Distribution Of The Estimators For Autoregressive Time Series With A Unit Root", *Journal of the American Statistical Association*, 1979, Vol. 74, hal. 427-431.
67. Dickey, D. A., & Pantula, S. G. (1987). Determining the Order of Differencing in Autoregressive Processes. *Journal of Business & Economic Statistics*, *5*(4), 455-461.
68. Doldado, Juan, Tim Jenkinson, dan Simon Sosvilla-Rivera, "Cointegration and Unit Roots", *Journal Of Economic Surveys*, April 1990, hal 249-73.
69. Eberhardt, M. (2011a). Panel Time-Series ModeLling: New Tools for Analyzing xt Data. In *2011 UK Stata Users Group meeting*.
70. Eberhardt, M.., (2011b). "*MULTIPURT: Stata module to run 1st and 2nd generation panel unit root tests for multiple variables and lags*," Statistical Software Components S457239, Boston College Department of Economics, revised 08 Feb 2011.
71. Eberhardt (c), M. 2011. "XTCD: Stata Module to Investigate Variable/Residual Cross-Section Dependence," Statistical Software Components S457237, Boston College Department of Economics.
72. Eberhardt, M., & Teal, F. (2011). Econometrics for grumblers: a new look at the literature on cross-country growth empirics. *Journal of Economic Surveys*, *25*(1), 109-155.
73. Eberhardt, M., Helmers, C., & Strauss, H. (2013). Do spillovers matter when estimating private returns to R&D?. *Review of Economics and Statistics*, *95*(2), 436-448.
74. Efron, B., & Hastie, T. (2016). *Computer Age Statistical Inference* (Vol. 5). Cambridge University Press.
75. Enders, W. E., 2004, Applied Econometrics Time Series, 2nd Ed, John Wiley & Sons, New York.
76. Engle, R. F., 1982, "Autoregressive Conditional Heterocedasticity with the Estimates of the Variance of United Kingdom Inflation", Econometrica, hal 987-1007.

77. Engle, R. (2001). GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics. *Journal of Economic Perspectives*, *15*(4), 157-168.
78. Engle, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, *20*(3), 339-350.
79. Engle, Robert F. dan Granger, C. W. J., 1987, "Cointegration and Error Correction: Representation, Estimation, and Testing", *Econometrica*, pp. 257-76.
80. Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in Regression Analysis: the Problem Revisited. *The Review of Economic and Statistics*, 92-107.
81. Favero, C. A. (2001). *Applied macroeconometrics*. Oxford University Press.
82. Frank, l. E., and J. H. Friedman. 1993. A Statistical View of Some Chemometrics Regression Tools. *Technometrics* 35(2): 109{135).
83. Gareth. J., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 112, p. 18). New York: Springer.
84. Gengenbach, C., Urbain, J. P., & Westerlund, J. (2009). Panel Error Correction Testing with Global Stochastic Trends (Updated Version of Research Memorandum 2008-051). *Maastricht: METEOR, Maastricht Research School of Economics of Technology and Organization*.
85. Gelfand, A. E., S. E. Hills, A. Racine-Poon, and A. F. M. Smith. 1990. Illustration of Bayesian Inference in Normal Data Models using Gibbs Sampling. *Journal of the American Statistical Association* 85: 972–985.
86. Gelman, A., W. R. Gilks, and G. O. Roberts. 1997. Weak convergence and optimal scaling of random walk Metropolis algorithms. Annals of Applied Probability 7: 110–120.
87. Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2014. Bayesian Data Analysis. 3rd ed. Boca Raton, FL: Chapman & Hall/CRC.
88. Geweke, J., Horowitz, J. L., & Pesaran, M. H. (2006). Econometrics: a Bird's Eye View, CESifo Working Paper No. 1870.

89. Geweke, J., Koop, G., and van Dijk, H. (2011). *The Oxford Handbook of Bayesian Econometrics*. Oxford University Press.
90. Giordani, P., and R. J. Kohn. 2010. Adaptive Independent Metropolis–Hastings by Fast Estimation of Mixtures of Normals. *Journal of Computational and Graphical Statistics* 19: 243–259.
91. Greenberg, E. (2012). *Introduction to Bayesian Econometrics*. Cambridge University Press.
92. Greene, W. H. (2018) *Econometric Analysis, Eighth Edition*, Boston: Prentice-Hall, 36–38.
93. Glosten, L. R., Jagannathan, R., dan D. E. Runkle, 1993, "On the relationship between the expected value and the volatility of the nominal excess return on stocks", *Journal of Finance*, Vol. 48, hal. 1779-1801.
94. Goldberger, A. S., (1991). *A Course in Econometrics*. Harvard University Press.
95. Granger, C. W. J, dan Newbold, P., 1974, "Spurious Eegressions in Econometrics", Journal of Econometrics, Vol. 2, hal. 111-120.
96. Gujarati, D., 2008, Basic Econometrics, McGraw-Hill, Singapore.
97. Hadri, K. (2000). Testing for Stationarity in Heterogeneous Panel Data. *The Econometrics Journal*, *3*(2), 148-161.
98. Hamilton, J. D., 1994, Time Series Analysis, Princeton University Press.
99. Hamilton, J. D. (1989), 'A new approach to the economic analysis of nonstationary time series and the business cycle', *Econometrica* 57(2), 357–384.
100. Harvey, A. C., 1990, The Econometric Analysis of Time Series, 2nd Ed, Phillip Allan, Oxford, UK.
101. Harris, R. dan R. Solis, 2003, Applied Time Series Modelling dan Forecasting, John Wiley & Sons, New York.
102. Hausman, J. A. (1978). Specification Tests in Econometrics. *Econometrica: Journal of the Econometric Society*, 1251-1271.
103. Hayashi, F., 2000, Econometrics, Princeton University Press, New Jersey.

104. Hansen, L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica: Journal of the Econometric Society*, 1029-1054.
105. Harkleroad, D. (1996). Actionable Competitive Intelligence, Society of Competitive Intelligence Professionals (Ed.), Annual International Conference & Exhibit Conference Proceedings. Alexandria/Va, 43–52.
106. Harrison, D. and Rubinfeld, D. L. 'Hedonic Prices and the Demand for Clean Air', J. Environ. Economics & Management, vol.5, 81-102, 1978
107. Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and their Applications. Oxford University Press.
108. Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica: Journal of the Econometric Eociety*, 153-161.
109. Heij, C., Heij, C., de Boer, P., Franses, P. H., Kloek, T., & van Dijk, H. K. (2004). *Econometric Methods with Applications in Business and Economics*. Oxford University Press.
110. Hendry, David F., Adrian R. Pagan, and J. Denis Sargan, 1984, "Dynamic Specification," Chapter 18 in Handbook of Econometrics, Vol. 1, ed. by Zvi Griliches and Michael D. Intriligator (New York: North-Holland, 2nd ed.), pp. 1023-1100.
111. Hendry, D. F. (1995). *Dynamic Econometrics*. Oxford University Press.
112. Hill, R. C, Griffiths W. E., and C. L. Guay, (2017), Principles of Econometrics (5th Edition), Wiley.
113. Hoerl, A. E., and R. W. Kennard. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 12(1): 55–67.
114. Holtz-Eakin, D., Newey, W., & Rosen, H. S. (1988). Estimating Vector Autoregressions with Panel Data. *Econometrica: Journal of the Econometric Society*, 1371–1395.
115. Hogg, R. V., McKean, J., & Craig, A. T. (2019). *Introduction to Mathematical statistics*. Pearson Education, 8th Edition.
116. Huber, C. (2014). Introduction to Structural Equation ModelLing using Stata. *California Association for Institutional Research*.
117. Hurvich, C. M., & Tsai, C. L. (1989). Regression and Time Series Model Selection in Small Samples. *Biometrika*, *76*(2), 297–307.

118. Im, K. S., Pesaran, M. H., & Shin, Y. (2003). Testing for Unit Roots in Heterogeneous Panels. *Journal of Econometrics*, *115*(1), 53-74.
119. Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica: Journal of the Econometric Society*, 1551-1580.
120. Johansen, S. (1995). *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press on Demand.
121. Johnson, R. A., dan Wichern, D. W. (2014). *Applied Multivariate Statistical Analysis* (Vol. 5, No. 8). Upper Saddle River, NJ: Prentice hall.
122. Johnston, J. dan J. Dinardo, 1997, Econometric Methods, 4th ed, Mc Graw-Hill International, New York.
123. Judd, C. M., McClelland, G. H., & Ryan, C. S. (2011). *Data Analysis: A Model Comparison Approach*. Routledge.
124. Kao, C. (1999). Spurious Regression and Residual-Based Tests for Cointegration in Panel Data. *Journal of Econometrics*, *90*(1), 1-44.
125. Kennedy, P. (2003). *A Guide to Econometrics*. MIT Press.
126. Koenker, R., & Bassett Jr, G. (1978). Regression Quantiles. *Econometrica: Journal of the Econometric Society*, 33-50.
127. Koenker, R. W., & d'Orey, V. (1987). Algorithm AS 229: Computing Regression Quantiles. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *36*(3), 383-393.
128. Koenker, R. (2005). Quantile Regression, Cambridge University Press.
129. Kripfganz, S., & Schneider, D. C. (2016, July). ardl: Stata Module to Estimate Autoregressive Distributed Lag Models. In *Stata Conference, Chicago*.
130. Kwiatkowski, D., Phillips, P. C. B, Schmidt, P. dan Y. Shin, 1992, Testing the Null Stationarity Againts the Alternative of a Unit Root, Journal of Econometrics, 54, hal 159-178.
131. Layard, R., & Nickell, S. (1986). Unemployment in Britain. *Economica*, *53*(210), S121–S169.
132. Levin, A., Lin, C. F., & Chu, C. (1992). Unit Root Tests in Panel Data: Asymptotic and Finite Sample Properties. USCD Discussion Paper

133. Lindsey, C., & Sheather, S. (2010). Variable Selection in Linear Regression. *The Stata Journal*, *10*(4), 650-669.
134. Little, R. J., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data* (Vol. 793). John Wiley & Sons.
135. Lovell, M.C. (2008) "A Simple Proof of the FWL Theorem," *Journal of Economic Education*, Winter 2008, 88–91.
136. Lu, X., & White, H. (2014). Robustness checks and robustness tests in applied economics. *Journal of Econometrics*, *178*, 194-206.
137. Lutkepohl, Helmut., 1991, *Introduction to Multiple Time Series Analysis*, Springer Verlag.
138. MacKinon, J., 1991, Critical Values for Cointegration Test, dalam R.F. Engle dan C.W.J. Granger (eds), Long Run Economic Relationship, hal 267-276, Oxford University Press, Oxford.
139. Maddala, G. S., & Wu, S. (1999). A comparative study of unit root tests with panel data and a new simple test. *Oxford Bulletin of Economics and Statistics*, *61*(S1), 631-652.
140. Malhotra, N. K. (2010). Marketing Research. An Applied Approach, 6th Global Edition. London: Pearson.
141. McClelland, M. M., Acock, A. C., Piccinin, A., Rhea, S. A., & Stallings, M. C. (2013). Relations between preschool attention span-persistence and age 25 educational outcomes. *Early Childhood Research Quarterly*, *28*(2), 314–324.
142. Medeiros, R. (2016). Handling missing data in Stata: Imputation and likelihood-based approaches. In *2016 Swiss Stata Users Group Meeting. StataCorp LP*.
143. Mendenhall, W dan T. Sincich, 1996, A Second Course in Statistics: Regression Analysis, 5th Ed, Upper Saddle River, NJ: Prentice-Hall.
144. Metropolis, N., and S. Ulam. 1949. The Monte Carlo method. Journal of the American Statistical Association 44: 335–341.
145. Mitchell, M. N. (2012). *A Visual Guide to Stata Graphics*. 3rd Edition, Stata Press.
146. Mitchell, M. N. (2020). *Data Management Using Stata: A Practical Handbook* (No. 005.369 M5.). 2nd Edition, College Station, TX: Stata Press.

147. Montgomery, D. C., and E. A Peck. 1982. Introduction to Linear Regression Analysis. Wiley and Son, *New York*.
148. Mroz, T. A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica: Journal of the Econometric Cociety*, 765-799.
149. Mullainathan, S., & Spies, J. (2017). Machine Learning: an Applied Econometric Approach. *Journal of Economic Perspectives*, *31*(2), 87-106.
150. Nelson, C.R. dan Plosser, Charles."Trends and Random Walks In Macroeconomic Time Series" *Journal Of Monetary Economics*, 1982, Vol 10. hal. 139-162.
151. Nelson, D., 1991, "Conditional Heterocedasticity in Asset Returns: A New Approach", *Econometrica*, Vol. 59, hal 347-370.
152. Newey, W. K., & West, K. D. (1987). A Simple, Positive-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* 55: 703 708.
153. Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica: Journal of the Econometric Society*, 1417-1426.
154. Osborn, D. R., "A Survey Of Seasonality In UK Macroeconomic Variables", *International Journal Of Forecasting*, 1990, 6, hal. 327-36.
155. Patterson, K., 2000, An Introduction to Applied Econometrics: A Time Series Approach, Palgrave, New York.
156. Pedroni, P. (2004). Panel cointegration: asymptotic and finite sample properties of pooled time series tests with an application to the PPP hypothesis. *Econometric Theory*, 597-625.
157. Pesaran, M. H., & Smith, R. (1995). Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics*, *68*(1), 79-113.
158. Pesaran, M. H., Shin, Y., & Smith, R. P. (1999). Pooled mean group estimation of dynamic heterogeneous panels. *Journal of the American Statistical Association*, *94*(446), 621–634.

159. Pesaran, H. M. (2004). General diagnostic tests for cross-sectional dependence in panels. *University of Cambridge, Cambridge Working Papers in Economics*, *435*.
160. Pesaran, M. H. (2007). A simple panel unit root test in the presence of cross-section dependence. *Journal of Applied Econometrics*, *22*(2), 265-312.
161. Pesaran, M. H. (2012). On the interpretation of panel unit root tests. *Economics Letters*, *116*(3), 545-546.
162. Pesaran, M. H. (2015). *Time series and Panel Data Econometrics*. Oxford University Press.
163. Phillips, P. C. B, and Perron P. "Testing for Unit Root in Time Series Regression", *Biometrika*, 1988, Vol 75. hal. 335-436.
164. Popper, K, (1934). *The Logic of Scientific Discovery*. Routledge.
165. Pyndick, R. dan D. L. Rubinfeld, 2000, Econometric Model & Economic Forecast, McGraw Hill, 4th Edition, Singapore.
166. Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, *31*(2), 350-371.
167. Rogoff, K. (1996). The Purchasing Power Parity Puzzle. *Journal of Economic literature*, *34*(2), 647-668.
168. Romer, D., "Openness and Inflation: Theory and Evidence," *Quarterly Journal of Economics, CVIII* (November 1993), 869-903.
169. Roodman, D. (2009). How to do xtabond2: An Introduction to Difference and System GMM in Stata. *The Stata Journal*, *9*(1), 86-136.
170. Said, S. and David Dickey, "Testing for Unit Roots in Autoregressive-Moving Average Models with Unknown Order", *Biometrica*, 1984, 71, hal 599-607.
171. Sánchez, G. (2017). Introduction to Bayesian Analysis in Stata, Stata Press.
172. Schaffer, M. E., 2010. xtivreg2: Stata Module to Perform Extended IV/2SLS, GMM and AC/HAC, LIML and k-class regression for panel data models. http://ideas.repec.org/c/boc/bocode/s456501.html
173. Schaffer, M. E. (2010). EGRANGER: Stata module to perform Engle-Granger cointegration tests and 2-step ECM estimation.

174. Schopohl, L., Wichmann, R., & Brooks, C. (2019). Stata Guide to Accompany Introductory Econometrics for Finance.
175. Schumacker, R. E., & Lomax, R. G. (2004). *A Beginner's Guide to Structural Equation Modeling*. Psychology Press.
176. Schwert, G. W. (1989) "Tests for unit roots: A Monte Carlo investigation" Journal of Business and Economics Statistics 7, 147–159.
177. Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359-1366.
178. Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica: Journal of the Econometric Society*, 1-48.
179. Smith, G. (2018). Step Away from Stepwise. *Journal of Big Data*, *5*(1), 32.
180. Stathis, P. and Martin, C., 2010, The Routledge Companion to Philosophy of Science, London, Routledge. hal. 129-38.
181. Stock, J. H., Wright, J. H., & Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, *20* (4), 518-529.
182. Stock, J. H. dan M. W. Watson, 2003, Introduction To Econometrics, Addison Wesley, Boston.
183. Taylor, M. A. (2018). Simulating the Central Limit Theorem. *The Stata Journal*, *18*(2), 345-356.
184. Tibshirani, R. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 58(1): 267-288.
185. Tikhonov, A. N. 1963. On the solution of ill-posed problems and the method of regularization. In *Doklady Akademii Nauk*, vol. 151, 501{504. Russian Academy of Sciences.
186. Tanner, M. A., and W. H. Wong. 1987. The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association 82*: 528–550.
187. Tong, H., 1978. On a Threshold Model. In: Chen, C. H. (Ed.), Pattern Recognition and Signal Processing. Sijthoff and Noordhoff, Amsterdam

188. Tong, H. (2015). Threshold Models in Time Series Analysis—Some Reflections. *Journal of Econometrics*, *189*(2), 485-491.
189. Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, *28*(2), 3-28.
190. Vogelvang, B. (2005). *Econometrics: Theory and Applications with Eviews*. Pearson Education.
191. Von Neuman, J. (1951). Various Techniques Used in Connection with Random Digits in Monte Carlo Method. *Applied Mathematics Series*, *12*.
192. Wada, R. (2014). OUTREG2: Stata Module to Arrange Regression Outputs into an Illustrative Table.
193. Wang, Q. (2015). Fixed-Effect Panel Threshold Model Using Stata. *The Stata Journal*, *15*(1), 121-134.
194. Westerlund, J. (2008). Panel Cointegration Tests of the Fisher Effect. *Journal of Applied Econometrics*, *23*(2), 193-233.
195. White, Halbert. "A Heteroskedasticiy-Consistent Covariance Matrix and a Direct Test for Heterokedasticiy", *Econometrica*, 1982, 48, hal 817-838.
196. Williams, R. 2007. Stata Tip 46: Step We Gaily, on We Go. *Stata Journal* 7: 272–274.
197. Wooldridge, J. M, 2019, Introductory Econometric: A Modern Approach, 7th Edition, Addison Wesley, Boston.
198. Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.
199. Young, G. A. (2013). Fundamental Theory of Statistical Inference. Imperial College.
200. Yu-Jun, L. (2014). WINSOR2: Stata Module to Winsorize Data.
201. Zellner, A., 1986, "On Assessing Prior Distributions and Bayesian Regression Analysis with g Prior Distribution", In Goel, P.; Zellner, A. (eds), Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti. Studies in Bayesian Econometrics and Statistics, New York: Elsevier.

INDEKS

A
Akaike Information Criterion (AIC) 12
Analisis pendahuluan (*preliminary analysis*) 231
Autoregressive Moving Average (ARMA) 37

B
Bayesian Econometrics 195
Bias seleksi (*selection bias*) 152
Branching 180

C
Censored regression 119, 145
Cross validation 205

D
Data generating proses 33
Data panel 76
Deklarasi 179
Dickey Fuller 189
Difference GMM; D-GMM 96

E
ECM 68
Ekonometrika Bayesian 195
Ekspektasi adaptif 3

F
Foreach 180
Forvalue 180
Frequentist econometrics 195

G
Granger Causality Test 13

H
Hannan Quin 12
High dimensional 196

I
Impulse-Response Function (IRF) 15
Inverse Mills Ratio = IMR 154

K
Kelaikan suai (*goodness of fit*) 243
K-fold Cross Validation 205
Kriteria informasi 204

L
Least Absolute Shrinkage Selector Operator (LASSO) 200
Long data 196
Looping 180

M
Machine learning (ML) 194
Macro 179
Mean Group = MG 105
Mekanisme penyesuaian parsial 3
Model binary response 118
Model Data Panel Dinamis 93
Model efek random (REM) 81
Model koreksi kesalahan 63
Model Long Data Panel 101
Monte Carlo 182

O
Odd ratio 122
Ordered Response 136

P
Percent Correctly Predicted 124
Persamaan seleksi; selection equation 154
Pooled Mean Group = PMG 105
Post-estimation OLS 204
Prosedur Heckman 152
Proses Koreksi Kesalahan 3
Pseudo R-Squared 124

R
Regresi palsu 34
Regresi Poisson 140
Regularized regression 200
Robustness Check 233

S
Schwartz Information Criterion 12
SEM Builder 166
Spurious Regression 186
Statistik AME 135
Strong exogeneity 39
Structural break 24
Structural equation modelling (SEM) 160
Super exogeneity 40
System GMM; S-GMM 96

T
Tobit 146

U
Uji Dickey-Fuller 51
Uji Hausman 83
Uji Phillip dan Perron 53
Unit root 43

V
Variabel Dependen Multinominal 130
Variabel yang terobservasi 160
Variance Decomposition (VD) 15
Vector Auto Regression (VAR) 9
Vector Error Correction Model (VECM) 59

W
Weak exogeneity 39
While 180