

POL 572
Multivariate Political Analysis

Gregory Wawro
Associate Professor
Department of Political Science
Columbia University
420 W. 118th St.
New York, NY 10027

phone: (212) 854-8540

fax: (212) 222-0598

email: gjw10@columbia.edu

ACKNOWLEDGEMENTS

These notes are adapted from lectures prepared by Professors Lucy Goodhart and Nolan McCarty. The course also draws from the following works.

- Baltagi, Badi H. 1995. *Econometric Analysis of Panel Data*. New York: John Wiley & Sons.
- Beck, Nathaniel, and Jonathan N. Katz. 1995. "What To Do (and Not To Do) with Time-Series Cross-Section Data in Comparative Politics." *American Political Science Review* 89:634-647.
- Davidson, Russell and James G. MacKinnon. 1993. *Estimation and Inference in Econometrics*. New York: Oxford University Press.
- Eliason, Scott R. 1993. *Maximum Likelihood Estimation: Logic and Practice*. Newbury Park, CA: Sage.
- Fox, John. 2002. *An R and S-Plus Companion to Applied Regression*. Thousand Oaks: Sage Publications.
- Greene, William H. 2003. *Econometric Analysis*, 5th ed. Upper Saddle River, N.J.: Prentice Hall
- Gujarati, Damodar N., *Basic Econometrics*, 2003, Fourth Edition, New York: McGraw Hill.
- Hsiao, Cheng. 1986. *Analysis of Panel Data*. New York: Cambridge University Press.
- Kennedy, Peter. 2003. *A Guide to Econometrics*, Fifth Edition. Cambridge, MA: MIT Press.
- King, Gary. 1989. *Unifying Political Methodology*. New York: Cambridge University Press
- Lancaster, Tony. 1990. *The Econometric Analysis of Transition Data*. Cambridge: Cambridge University Press.
- Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks: Sage Publications.
- Maddala, G. S. 2001. *Introduction to Econometrics*. Third Edition, New York: John Wiley and Sons.

Wooldridge, Jeffrey M. 2002. *Introductory Econometrics: A Modern Approach*. Cincinnati, OH: Southwestern College Publishing.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press.

TABLE OF CONTENTS

LIST OF FIGURES	ix
LIST OF TABLES	x
I Asymptotics and Violations of Gauss-Markov Assumptions in the Classical Linear Regression Model	1
1 Large Sample Results and Asymptotics	2
1.1 Why do we Care about Large Sample Results (and what does this mean?)	2
1.2 What are Desirable Large Sample Properties?	4
1.3 How Do We Figure Out the Large Sample Properties of an Estimator?	6
1.3.1 The Consistency of $\hat{\beta}_{OLS}$	6
1.3.2 Asymptotic Normality of OLS	9
1.4 Large Sample Properties of Test Statistics	11
1.5 Desirable Large Sample Properties of ML Estimators	12
1.6 How Large Does n Have To Be?	13
2 Heteroskedasticity	14
2.1 Heteroskedasticity as a Violation of Gauss-Markov	14
2.1.1 Consequences of Non-Spherical Errors	15
2.2 Consequences for Efficiency and Standard Errors	16
2.3 Generalized Least Squares	16
2.3.1 Some intuition	17
2.4 Feasible Generalized Least Squares	18
2.5 White-Consistent Standard Errors	19
2.6 Tests for Heteroskedasticity	20
2.6.1 Visual Inspection of the Residuals	21
2.6.2 The Goldfeld-Quandt Test	21
2.6.3 The Breusch-Pagan Test	22
3 Autocorrelation	23
3.1 The Meaning of Autocorrelation	23
3.2 Causes of Autocorrelation	24
3.3 Consequences of Autocorrelation for Regression Coefficients and Standard Errors . .	24
3.4 Tests for Autocorrelation	25
3.4.1 The Durbin-Watson Test	26
3.4.2 The Breusch-Godfrey Test	27
3.5 The consequences of autocorrelation for the variance-covariance matrix	27
3.6 GLS and FGLS under Autocorrelation	30
3.7 Non-AR(1) processes	32

3.8	OLS Estimation with Lagged Dependent Variables and Autocorrelation	33
3.9	Bias and “Contemporaneous Correlation”	35
3.10	Measurement Error	35
3.11	Instrumental Variable Estimation	37
3.12	In the general case, why is IV estimation unbiased and consistent?	38
4	Simultaneous Equations Models and 2SLS	40
4.1	Simultaneous Equations Models and Bias	40
4.1.1	Motivating Example: Political Violence and Economic Growth	40
4.1.2	Simultaneity Bias	41
4.2	Identifying the Endogenous Variables	42
4.3	Identification	44
4.3.1	The Order Condition	44
4.4	IV Estimation and Two-Stage Least Squares	45
4.4.1	Some Important Observations	46
4.5	Recapitulation of 2SLS and Computation of Goodness-of-Fit	47
4.6	Computation of Standard Errors in 2SLS	48
4.7	Three-Stage Least Squares	50
4.8	Different Methods to Detect and Test for Endogeneity	51
4.8.1	Granger Causality	51
4.8.2	The Hausman Specification Test	52
4.8.3	Regression Version	53
4.8.4	How to do this in Stata	54
4.9	Testing for the Validity of Instruments	54
5	Time Series Modeling	56
5.1	Historical Background	56
5.2	The Auto-Regressive and Moving Average Specifications	56
5.2.1	An Autoregressive Process	57
5.3	Stationarity	58
5.4	A Moving Average Process	59
5.5	ARMA Processes	60
5.6	More on Stationarity	61
5.7	Integrated Processes, Spurious Correlations, and Testing for Unit Roots	61
5.7.1	Determining the Specification	65
5.8	The Autocorrelation Function for AR(1) and MA(1) processes	65
5.9	The Partial Autocorrelation Function for AR(1) and MA(1) processes	66
5.10	Different Specifications for Time Series Analysis	67
5.11	Determining the Number of Lags	69
5.12	Determining the Correct Specification for your Errors	70
5.13	Stata Commands	71

II	Maximum Likelihood Estimation	72
6	Intro to Maximum Likelihood	73
7	Maximum Likelihood In Depth	78
7.1	Asymptotic Properties of MLEs	81
7.2	Iterative Process of Finding the MLE	84
III	Models for Repeated Observations Data—Continuous Dependent Variables	86
8	Fixed Effects Estimators	87
8.1	LSDV as Fixed Effects	87
8.2	Application: Economic growth in 14 OECD countries	93
9	Random Effects Estimators	97
9.1	Intro	97
9.2	Deriving the random effects estimator	98
9.3	GLS Estimation	100
9.4	Maximum Likelihood Estimation	104
9.5	Fixed v. Random Effects	105
9.6	Testing between Fixed and Random Effects	105
9.7	Application	107
10	Non-Spherical Errors	109
10.1	Introduction	109
10.2	The Method of PCSEs	109
10.3	Robust Estimation of Asymptotic Covariance Matrices	111
10.4	Costs of ignoring unit effects revisited	115
10.5	Heteroskedasticity in FE and RE models	118
10.6	Serial Correlation in RE and FE models	120
10.7	Robust standard error estimation with unit effects	122
10.7.1	Arellano robust standard errors	122
10.7.2	Kiefer robust standard errors	123
10.8	Application: Garrett data	124
IV	Qualitative and Limited Dependent Variable Models Based on the Normal Regression Model	128
11	Introduction	129
11.1	Linear Regression Model	129

12 Probit	130
12.1 Interpretation of Coefficients	132
12.2 Goodness of fit measures	134
12.3 Voting Behavior Example	136
12.4 Obstruction and Passage of Legislation Example	140
12.5 Heteroskedasticity and the Probit Model	143
13 Ordered Probit	146
14 Censored Regression	151
14.1 Reparameterization for tobit model	156
14.2 Marginal Effects	159
14.3 Heteroskedasticity and the tobit model	161
15 Truncated Regression	162
15.1 Marginal Effects	164
16 Sample and Self-Selection Models	165
V Probabilistic Choice Models	177
17 Introduction	178
18 The Multinomial Logit Model	179
18.1 Identification Issue	180
18.2 Normalization	181
19 The Conditional Logit Model	184
19.1 Equivalence of the MNL model and conditional logit model	187
19.2 Independence of Irrelevant Alternatives	188
19.3 IIA test	191
20 The Nested Logit Model	192
21 The Multinomial Probit Model	195
21.1 Identification of variance-covariance matrix	197
21.2 Application: Alvarez and Nagler “Economics, Issues, and the Perot Candidacy: Voter Choice in the 1992 Presidential Election” ’95 <i>AJPS</i>	198
VI Duration Models	202
22 Introduction	203

23 Functions for Analyzing Duration Data	204
24 Proportional Hazard Model	212
25 Partial Likelihood Estimator	214
26 Nonparametric Approaches	217
VII Event Count Models	219
27 Introduction	220
28 Poisson Regression Model	221
28.1 Dispersion	223
28.2 Tests for Overdispersion	225
29 Gamma Model for Overdispersion	227
30 Binomial Model for Underdispersion	229
31 Generalized Event Count Model	231
32 Hurdle Poisson Models	234
VIII Models for Repeated Observations—Dichotomous Dependent Variables	236
33 Introduction	237
33.1 Fixed Effect Logit	239
33.2 Random Effects Probit	242
33.3 Correlated Random Effects Probit	245
33.4 CRE Probit Application: PAC contributions and roll call votes	248
34 Binary Time-Series Cross-Section (BTSCS) Data	252
35 Generalized Estimating Equations (GEEs)	256
35.1 GLMs for Correlated Data	257
35.2 Options for specifying within-cluster correlation	258
35.3 “Robust” Standard Errors	259
35.4 GEE2	260
35.5 Application: Panel Decision-making on the Court of Appeals	261

LIST OF FIGURES

7.1	The Natural Log of the Normal Distribution	85
12.1	Simulated Probabilities of Passage and Coalition Size	142
13.1	Probabilities for categories	148
14.1	Censoring	152
14.2	Censoring and OLS	153
14.3	Marginal Effects and Censoring	160
19.1	PDF and CDF for Extreme Value Distribution	186
23.1	Plots of hazard functions	206
23.2	Lognormal Hazard Function ($\lambda = .5$)	207

LIST OF TABLES

10.1	Results from Monte Carlo experiments involving time invariant variables	117
12.1	Probability of Voting for the Incumbent Member of Congress	138
12.2	Marginal Effects on Probability of Voting for the Incumbent Member of Congress .	139
12.3	Probit analysis of passage of obstructed measures, 1st–64th Congresses	141
12.4	Simulated Probabilities of Passage and Coalition Sizes, 1st–64th Congresses	141
23.1	Distributions, Hazard Functions, and Survival Functions	205
35.1	GEE analysis of judges' votes in Appeals Court decisions	262

Part I

Asymptotics and Violations of Gauss-Markov Assumptions in the Classical Linear Regression Model

Section 1

Large Sample Results and Asymptotics

1.1 Why do we Care about Large Sample Results (and what does this mean?)

- Large sample results for any estimator, $\hat{\theta}$, are the properties that we can say hold true as the number of data points, n , used to estimate $\hat{\theta}$ becomes “large.”
- Why do we care about these large sample results? We have an OLS model that, when the Gauss-Markov assumptions hold, has desirable properties. Why would we ever want to rely on the more difficult mathematical proofs that involve the limits of estimators as n becomes large?
- Recall to mind the Gauss-Markov assumptions:
 1. The true model is a linear functional form of the data: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.
 2. $E[\boldsymbol{\varepsilon}|\mathbf{X}] = \mathbf{0}$
 3. $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] = \sigma^2\mathbf{I}$
 4. \mathbf{X} is $n \times k$ with rank k (i.e., full column rank)
 5. $\boldsymbol{\varepsilon}|\mathbf{X} \sim N[\mathbf{0}, \sigma^2\mathbf{I}]$
- Recall that if we are prepared to make a further, simplifying assumption, that \mathbf{X} is fixed in repeated samples, then the expectations conditional on \mathbf{X} can be written in unconditional form.
- There are two main reasons for the use of large sample results, both of which have to do with violations of these of these assumptions.
 1. Errors are not distributed normally
 - If assumption 5, above, does not hold, then we cannot use the small sample results.
 - We can establish that $\hat{\boldsymbol{\beta}}_{OLS}$ is unbiased and “best linear unbiased estimator” without any recourse to the normality assumption and we also established that the variance of $\hat{\boldsymbol{\beta}}_{OLS} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

- Used normality to show that $\hat{\boldsymbol{\beta}}_{OLS} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$ and that $(n - k)s^2/\sigma^2 \sim \chi_{n-k}^2$ where the latter involved showing that $(n - k)s^2/\sigma^2$ can also be expressed as a quadratic form of $\boldsymbol{\varepsilon}/\sigma$ which is distributed standard normal.
- Both of these results were used to show that we could calculate test statistics that were distributed as t and F . It is these results on test statistics, and our ability to perform hypothesis tests, that are invalidated if we cannot assume that the true errors are normally distributed.

2. Non-linear functional forms

- We may be interested in estimating non-linear functions of the original model.
- E.g., suppose that you have an unbiased estimator, π^* of

$$\pi = 1/(1 - \beta)$$

but you want to estimate β . You cannot simply use $(1 - 1/\pi^*)$ as an unbiased estimate β^* of β .

- To do so, you would have to be able to prove that $E(1 - 1/\pi^*) = \beta$. It is not true, however, that the expected value of a non-linear function of π is equal to the non-linear function of the expected value of π (this fact is also known as “Jensen’s Inequality”).
- Thus, we can’t make the last step that we would require (in small samples) to show that our estimator of β is unbiased. As Kennedy says, “the algebra associated with finding (small sample) expected values can become formidable whenever non-linearities are involved.” This problem disappears in large samples.
- The models for discrete and limited dependent variables that we will discuss later in the course involve non-linear functions of the parameters, $\boldsymbol{\beta}$. Thus, we cannot use the G-M assumptions to prove that these estimates are unbiased. However, ML estimates have attractive *large sample* properties. Thus, our discussion of the properties of those models will always be expressed in terms of large sample results.

1.2 What are Desirable Large Sample Properties?

- Under finite sample conditions, we look for estimators that are unbiased and efficient (i.e., have minimum variance for any unbiased estimator). We have also found it useful, when calculating test statistics, to have estimators that are normally distributed. In the large sample setting, we look for analogous properties.
- The large sample analog to unbiasedness is *consistency*. An estimator $\hat{\beta}$ is *consistent* if it converges in probability to β , that is if its probability limit as n becomes large is β ($\text{plim } \hat{\beta} = \beta$).
- The best way to think about this is that the sampling distribution of $\hat{\beta}$ collapses to a spike around the true β .
- Two Notions of Convergence

1. “Convergence in probability”:

$$X_n \xrightarrow{P} X \text{ iff } \lim_{n \rightarrow \infty} \Pr(|X(\omega) - X_n(\omega)| \geq \varepsilon) = 0$$

where ε is some small positive number. Or

$$\text{plim } X_n(\omega) = X(\omega)$$

2. “Convergence in quadratic mean or mean square”: If X_n has mean μ_n and variance σ_n^2 such that $\lim_{n \rightarrow \infty} \mu_n = c$ and $\lim_{n \rightarrow \infty} \sigma_n^2 = 0$, then X_n converges in mean square to c .

Note: convergence in quadratic mean implies convergence in probability ($X_n \xrightarrow{\text{qm}} c \Rightarrow X_n \xrightarrow{P} c$).

- An estimator may be biased in small samples but consistent. For example, take an estimate of $\beta = \beta + 1/n$. In small samples this is biased, but as the sample size becomes infinite, $1/n$ goes to zero ($\lim_{n \rightarrow \infty} 1/n = 0$).

- Although an estimator may be biased yet consistent, it is very hard (read impossible) for it to be unbiased but inconsistent. If $\hat{\beta}$ is unbiased, there is nowhere for it to collapse to but β . Thus, the only way an estimator could be unbiased but inconsistent is if its sampling distribution never collapses to a spike around the true β .
- We are also interested in *asymptotic normality*. Even though an estimator may not be distributed normally in small samples, we can usually appeal to some version of the central limit theorem (see Greene, Fifth Edition, Appendix D.2.6 or Kennedy Appendix C) to show that it will be distributed normally in large samples.
- More precisely, the different versions of the central limit theorem state that the mean of any random variable, whatever the distribution of the underlying variable, will in the limit be distributed such that:

$$\sqrt{n}(\bar{x}_n - \mu) \xrightarrow{d} N[0, \sigma^2]$$

Thus, even if x_i is not distributed normal, the sampling distribution of the average of an independent sample of the x_i 's will be distributed normal.

- To see how we arrived at this result, begin with the variance of the mean: $\text{var}(\bar{x}_n) = \frac{\sigma^2}{n}$.
- Next, by the CLT, this will be distributed normal: $\bar{x}_n \sim N(\mu, \frac{\sigma^2}{n})$
- As a result,

$$\frac{\bar{x}_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \sqrt{n} \frac{(\bar{x}_n - \mu)}{\sigma} \sim N(0, 1)$$

Multiplying the expression above by σ will get you the result on the asymptotic distribution of the sample average.

- When it comes to establishing the asymptotic normality of estimators, we can usually express that estimator in the form of an average, as a sum of values divided by the number of observations n , so that we can then apply the central limit theorem.

- We are also interested in *asymptotic efficiency*. An estimator is asymptotically efficient if it is consistent, asymptotically normally distributed, and has an asymptotic covariance matrix that is not larger than the asymptotic covariance matrix of any other consistent, asymptotically normally distributed estimator.
- We will rely on the result that (under most conditions) the maximum likelihood estimator is asymptotically efficient. In fact, it attains the smallest possible variance, the Cramer-Rao lower bound, if that bound exists. Thus, to show that any estimator is asymptotically efficient, it is sufficient to show that the estimator in question either *is* the maximum likelihood estimate or has identical asymptotic properties.

1.3 How Do We Figure Out the Large Sample Properties of an Estimator?

- To show that any estimator of any quantity, $\hat{\theta}$, is consistent, we have to show that $\text{plim } \hat{\theta} = \theta$. The means of doing so is to show that any bias approaches zero as n becomes large and the variance in the sampling distribution also collapses to zero.
- To show that $\hat{\theta}$ is asymptotically normal, we have to show that its sampling distribution can be expressed as the sampling distribution of a sample average pre-multiplied by \sqrt{n} .
- Let's explore the large sample properties of $\hat{\beta}_{OLS}$ w/o assuming that $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$.

1.3.1 The Consistency of $\hat{\beta}_{OLS}$

- Begin with the expression for $\hat{\beta}_{OLS}$:

$$\hat{\beta}_{OLS} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$$

- Instead of taking the expectation, we now take the probability limit:

$$\text{plim } \hat{\beta}_{OLS} = \boldsymbol{\beta} + \text{plim } (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$$

We can multiply both sides of the equation by $n/n = 1$ to produce:

$$\text{plim } \hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta} + \text{plim } (\mathbf{X}'\mathbf{X}/n)^{-1}(\mathbf{X}'\boldsymbol{\varepsilon}/n)$$

- For the next step we need the Slutsky Theorem: For a continuous function $g(x_n)$ that is not a function of n , $\text{plim } g(x_n) = g(\text{plim } x_n)$.
- An implication of this thm is that if x_n and y_n are random variables with $\text{plim } x_n = c$ and $\text{plim } y_n = d$, then $\text{plim } (x_n \cdot y_n) = c \cdot d$.
- If \mathbf{X}_n and \mathbf{Y}_n are random matrices with $\text{plim } \mathbf{X}_n = \mathbf{A}$ and $\text{plim } \mathbf{Y}_n = \mathbf{B}$ then $\text{plim } \mathbf{X}_n \mathbf{Y}_n = \mathbf{AB}$.
- Thus, we can say that:

$$\text{plim } \hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta} + \text{plim } (\mathbf{X}'\mathbf{X}/n)^{-1} \text{plim } (\mathbf{X}'\boldsymbol{\varepsilon}/n)$$

- Since the inverse is a continuous function, the Slutsky thm enables us to bring the first plim into the parenthesis:

$$\text{plim } \hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta} + (\text{plim } \mathbf{X}'\mathbf{X}/n)^{-1} \text{plim } (\mathbf{X}'\boldsymbol{\varepsilon}/n)$$

- Let's assume that

$$\lim_{n \rightarrow \infty} (\mathbf{X}'\mathbf{X}/n) = \mathbf{Q}$$

where \mathbf{Q} is a finite, positive definite matrix. In words, as n increases, the elements of $\mathbf{X}'\mathbf{X}$ do not increase at a rate greater than n and the explanatory variables are not linearly dependent.

- To fix ideas, let's consider a case where this assumption would not be valid:

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t$$

which would give

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} T & \sum_{t=1}^T t \\ \sum_{t=1}^T t & \sum_{t=1}^T t^2 \end{bmatrix} = \begin{bmatrix} T & T(T+1)/2 \\ T(T+1)/2 & T(T+1)(2T+1)/6 \end{bmatrix}$$

Taking limits gives

$$\lim_{t \rightarrow \infty} (\mathbf{X}'\mathbf{X}/t) = \begin{bmatrix} 1 & \infty \\ \infty & \infty \end{bmatrix}.$$

- More generally, each element of $(\mathbf{X}'\mathbf{X})$ is composed of the sum of squares and the sum of cross-products of the explanatory variables. As such, the elements of $(\mathbf{X}'\mathbf{X})$ grow larger with each additional data point, n .
- But if we assume the elements of this matrix do not grow at a rate faster than n and the columns of \mathbf{X} are not linear dependent, then dividing by n , gives convergence to a finite number.
- We can now say that $\text{plim } \hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta} + \mathbf{Q}^{-1}\text{plim } (\mathbf{X}'\boldsymbol{\varepsilon}/n)$ and the next step in the proof is to show that $\text{plim } (\mathbf{X}'\boldsymbol{\varepsilon}/n)$ is equal to zero. To demonstrate this, we will prove that its expectation is equal to zero and that its variance converges to zero.
- Think about the individual elements in $(\mathbf{X}'\boldsymbol{\varepsilon}/n)$. This is a $k \times 1$ matrix in which each element is the sum of all n observations of a given explanatory variable multiplied by each realization of the error term. In other words:

$$\frac{1}{n}\mathbf{X}'\boldsymbol{\varepsilon} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \varepsilon_i = \frac{1}{n} \sum_{i=1}^n \mathbf{w}_i = \bar{\mathbf{w}} \quad (1.1)$$

where $\bar{\mathbf{w}}$ is a $k \times 1$ vector of the sample averages of $x_{ki}\varepsilon_i$. (Verify this if it is not clear to you.)

- Since we are still assuming that \mathbf{X} is non-stochastic, we can work through the expectations operator to say that:

$$E[\bar{\mathbf{w}}] = \frac{1}{n} \sum_{i=1}^n E[\mathbf{w}_i] = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i E[\varepsilon_i] = \frac{1}{n} \mathbf{X}' E[\boldsymbol{\varepsilon}] = \mathbf{0} \quad (1.2)$$

In addition, using the fact that $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \sigma^2\mathbf{I}$ we can say that:

$$\text{var}[\bar{\mathbf{w}}] = E[\bar{\mathbf{w}}\bar{\mathbf{w}}'] = \frac{1}{n} \mathbf{X}' E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] \mathbf{X} \frac{1}{n} = \frac{\sigma^2}{n} \frac{\mathbf{X}'\mathbf{X}}{n}$$

- In the limit, as $n \rightarrow \infty$, $\frac{\sigma^2}{n} \rightarrow 0$, $\frac{\mathbf{X}'\mathbf{X}}{n} \rightarrow \mathbf{Q}$, and thus

$$\lim_{n \rightarrow \infty} \text{var}[\bar{\mathbf{w}}] = 0 \cdot \mathbf{Q}$$

- Therefore, we can say that $\bar{\mathbf{w}}$ converges in mean square to 0 \Rightarrow $\text{plim}(\mathbf{X}'\boldsymbol{\varepsilon}/n)$ is equal to zero, so that:

$$\text{plim} \hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta} + \mathbf{Q}^{-1} \cdot \mathbf{0} = \boldsymbol{\beta}$$

Thus, the OLS estimator is consistent as well as unbiased.

1.3.2 Asymptotic Normality of OLS

- Let's show that the OLS estimator, $\hat{\boldsymbol{\beta}}_{OLS}$, is also asymptotically normal. Start with

$$\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$$

and then subtract $\boldsymbol{\beta}$ from each side and multiply through by \sqrt{n} to yield (use $\sqrt{n} = n/\sqrt{n}$):

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) = \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} \left(\frac{1}{\sqrt{n}}\right) \mathbf{X}'\boldsymbol{\varepsilon}$$

We've already established that the first term on the right-hand side converges to \mathbf{Q}^{-1} . We need to derive the limiting distribution of the term $\left(\frac{1}{\sqrt{n}}\right) \mathbf{X}'\boldsymbol{\varepsilon}$.

- From equations 1.1 and 1.2, we can write

$$\left(\frac{1}{\sqrt{n}}\right) \mathbf{X}'\boldsymbol{\varepsilon} = \sqrt{n}(\bar{\mathbf{w}} - E[\bar{\mathbf{w}}])$$

and then find the limiting distribution of $\sqrt{n}\bar{\mathbf{w}}$.

- To do this, we will use a variant of the CLT (called Lindberg-Feller), which allows for variables to come from different distributions.

From above, we know that

$$\bar{\mathbf{w}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \varepsilon_i$$

which means $\bar{\mathbf{w}}$ is the average of n independent vectors $\mathbf{x}_i \varepsilon_i$ with means $\mathbf{0}$ and variances

$$\text{var}[\mathbf{x}_i \varepsilon_i] = \sigma^2 \mathbf{x}_i \mathbf{x}_i' = \sigma^2 \mathbf{Q}_i$$

- Thus,

$$\begin{aligned}\text{var}[\sqrt{n}\bar{\mathbf{w}}] &= \sigma^2 \bar{\mathbf{Q}}_n = \sigma^2 \left(\frac{1}{n} \right) [\mathbf{Q}_1 + \mathbf{Q}_2 + \cdots + \mathbf{Q}_n] \\ &= \sigma^2 \left(\frac{1}{n} \right) \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \\ &= \sigma^2 \left(\frac{\mathbf{X}'\mathbf{X}}{n} \right)\end{aligned}$$

- Assuming the sum is not dominated by any particular term and that $\lim_{n \rightarrow \infty} (\mathbf{X}'\mathbf{X}/n) = \mathbf{Q}$, then

$$\lim_{n \rightarrow \infty} \sigma^2 \bar{\mathbf{Q}}_n = \sigma^2 \mathbf{Q}$$

- We can now invoke the Lindberg-Feller CLT to formally state that if the ε are iid w/ mean 0 and finite variance, and if each element of \mathbf{X} is finite and $\lim_{n \rightarrow \infty} (\mathbf{X}'\mathbf{X}/n) = \mathbf{Q}$, then

$$\left(\frac{1}{\sqrt{n}} \right) \mathbf{X}'\boldsymbol{\varepsilon} \xrightarrow{d} N[0, \sigma^2 \mathbf{Q}]$$

It follows that:

$$\mathbf{Q}^{-1} \left(\frac{1}{\sqrt{n}} \right) \mathbf{X}'\boldsymbol{\varepsilon} \xrightarrow{d} N[\mathbf{Q}^{-1} \cdot 0, \mathbf{Q}^{-1}(\sigma^2 \mathbf{Q})\mathbf{Q}^{-1}]$$

[Recalling that if a random variable \mathbf{X} has a variance equal to σ^2 , then $k\mathbf{X}$, where k is a constant, has a variance equal to $k^2\sigma^2$].

- Combining terms, and recalling what it was we were originally interested in:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) \xrightarrow{d} N[0, \sigma^2 \mathbf{Q}^{-1}]$$

- How do we get from here to a statement about the normal distribution of $\hat{\boldsymbol{\beta}}_{OLS}$? Divide through by \sqrt{n} on both sides and add $\boldsymbol{\beta}$ to show that the OLS estimator is asymptotically distributed normal:

$$\hat{\boldsymbol{\beta}}_{OLS} \overset{a}{\sim} N \left[\boldsymbol{\beta}, \frac{\sigma^2}{n} \mathbf{Q}^{-1} \right]$$

- To complete the steps, we can also show that $s^2 = \frac{e'e}{n-k}$ is consistent for σ^2 . Thus, $s^2(\mathbf{X}'\mathbf{X})^{-1}$ is consistent for $\sigma^2\mathbf{Q}^{-1}$. As a result, a consistent estimate for the asymptotic variance of $\hat{\beta}_{OLS}(= \frac{\sigma^2}{n}\mathbf{Q}^{-1} = \frac{\sigma^2}{n}(\frac{\mathbf{X}'\mathbf{X}}{n})^{-1})$ is $s^2(\mathbf{X}'\mathbf{X})^{-1}$.
- Thus, we can say that $\hat{\beta}_{OLS}$ is normally distributed and a consistent estimate of its asymptotic variance is given by $s^2(\mathbf{X}'\mathbf{X})^{-1}$, *even when the error terms are not distributed normally*. We have gone through the rigors of large sample proofs in order to show that in large samples OLS retains desirable properties that are similar to what it has in small sample when all of the G–M conditions hold.
- To conclude, the desirable properties of OLS do not rely on the assumption that the true error term is normally distributed. We can appeal to large sample results to show that the sampling distribution will still be normal as the sample size becomes large and that it will have variance that can be consistently estimated by $s^2(\mathbf{X}'\mathbf{X})^{-1}$.

1.4 Large Sample Properties of Test Statistics

- Given the normality results and the consistent estimate of the asymptotic variance given by $s^2(\mathbf{X}'\mathbf{X})^{-1}$, hypothesis testing proceeds almost as normal. Hypotheses on individual coefficients can still be estimated by constructing a t -statistic:

$$\frac{\hat{\beta}_k - \beta_k}{SE(\hat{\beta}_k)}$$

- When we did this in small samples, we made clear that we had to use an estimate, $\sqrt{s^2(\mathbf{X}'\mathbf{X})_{kk}^{-1}}$ for the true standard error in the denominator, equal to $\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})_{kk}^{-1}}$.

- As the sample size becomes large, we can replace this estimate by its probability limit, which is the true standard error, so that the denominator just becomes a constant. When we do so, we have a normally distributed random variable in the numerator divided by a constant, so the test-statistic is distributed as z , or standard normal. Another way to think of this intuitively is that the t -distribution converges to the z distribution as n becomes large.
- Testing joint hypothesis also proceeds via constructing an F -test. In this case, we recall the formula for an F -test, put in terms of the restriction matrix:

$$F[J, n - K] = \frac{(\mathbf{Rb} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{q})/J}{s^2}$$

- In small samples, we had to take account of the distribution of s^2 itself. This gave us the ratio of two chi-squared random variables, which is distributed F .
- In large samples, we can replace s^2 by its probability limit, σ^2 which is just a constant value. Multiplying both sides by J , we now have that the test statistic JF is composed of a chi-squared random variable in the numerator over a constant, so the JF statistic is the large sample analog of the F -test. If the JF statistic is larger than the critical value, then we can say that the restrictions are unlikely to be true.

1.5 Desirable Large Sample Properties of ML Estimators

- Recall that MLEs are *consistent*, *asymptotically normally distributed*, and *asymptotically efficient*, in that they always achieve the Cramer-Rao lower bound, when this bound exists.
- Thus, MLEs always have desirable large sample properties although their small sample estimates (as of σ^2) may be biased.
- We could also have shown that OLS was going to be consistent, asymptotically normally distributed, and asymptotically efficient by indicating that OLS *is* the MLE for the classical linear regression model.

1.6 How Large Does n Have To Be?

- Does any of this help us if we have to have an infinity of data points before we can attain consistency and asymptotic normality? How do we know how many data points are required before the sampling distribution of $\hat{\beta}_{OLS}$ becomes approximately normal?
- One way to check this is via Monte Carlo studies. Using a non-normal distribution for the error terms, we simulate draws from a distribution of the model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, using a different set of errors each time.
- We then calculate the $\hat{\beta}_{OLS}$ from repeated samples of size n and plot these different sample estimates $\hat{\beta}_{OLS}$ on a histogram.
- We gradually enlarge n , checking how large n has to be in order to give us a sampling distribution that is approximately normal.

Section 2

Heteroskedasticity

2.1 Heteroskedasticity as a Violation of Gauss-Markov

- The third of the Gauss-Markov assumptions is that $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \sigma^2\mathbf{I}_n$. The variance-covariance matrix of the true error terms is structured as:

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \begin{bmatrix} E(\varepsilon_1^2) & E(\varepsilon_1\varepsilon_2) & \dots & E(\varepsilon_1\varepsilon_N) \\ \vdots & \ddots & & \vdots \\ E(\varepsilon_N\varepsilon_1) & E(\varepsilon_N\varepsilon_2) & \dots & E(\varepsilon_N^2) \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

- If all of the diagonal terms are equal to one another, then each realization of the error term has the same variance, and the errors are said to be *homoskedastic*. If the diagonal terms are not the same, then the true error term is *heteroskedastic*.
- Also, if the off-diagonal elements are zero, then the covariance between different error terms is zero and the errors are *uncorrelated*. If the off-diagonal terms are non-zero, then the error terms are said to be *auto-correlated* and the error term for one observation is correlated with the error term for another observation.
- When the third assumption is violated, the variance-covariance matrix of the error term does not take the special form, $\sigma^2\mathbf{I}_n$, and is generally written, instead as $\sigma^2\boldsymbol{\Omega}$. The disturbances in this case are said to be *non-spherical* and the model should then be estimated by *Generalized Least Squares*, which employs $\sigma^2\boldsymbol{\Omega}$ rather than $\sigma^2\mathbf{I}_n$.

2.1.1 Consequences of Non-Spherical Errors

- The OLS estimator, $\hat{\beta}_{OLS}$, is still unbiased and (under most conditions) consistent.

Proof for Unbiasedness:

As before:

$$\hat{\beta}_{OLS} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$$

and we still have that $E[\boldsymbol{\varepsilon}|\mathbf{X}] = 0$, thus we can say that $E(\hat{\beta}_{OLS}) = \beta$.

- The estimated variance of $\hat{\beta}_{OLS}$ is $E[(\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))']$, which we can estimate as:

$$\begin{aligned} \text{var}(\hat{\beta}_{OLS}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\boldsymbol{\Omega})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

- Therefore, if the errors are normally distributed,

$$\hat{\beta}_{OLS} \sim N[\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}]$$

- We can also use the formula we derived in the lecture on asymptotics to show that $\hat{\beta}_{OLS}$ is consistent.

$$\text{plim } \hat{\beta}_{OLS} = \beta + \mathbf{Q}^{-1}\text{plim } (\mathbf{X}'\boldsymbol{\varepsilon}/n)$$

To show that $\text{plim } (\mathbf{X}'\boldsymbol{\varepsilon}/n) = 0$, we demonstrated that the expectation of $(\mathbf{X}'\boldsymbol{\varepsilon}/n)$ was equal to zero and that its variance was equal to $\frac{\sigma^2}{n} \frac{\mathbf{X}'\mathbf{X}}{n}$, which becomes zero as n grows to infinity.

- In the case where $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \sigma^2\boldsymbol{\Omega}$, the variance of $(\mathbf{X}'\boldsymbol{\varepsilon}/n)$ is equal to $\frac{\sigma^2}{n} \frac{(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})}{n}$. So long as this matrix converges to a finite matrix, then $\hat{\beta}_{OLS}$ is also consistent for β .
- Finally, in most cases, $\hat{\beta}_{OLS}$ is asymptotically normally distributed with mean β and its variance-covariance matrix is given by $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$.

2.2 Consequences for Efficiency and Standard Errors

- So what's the problem with using OLS when the true error term may be heteroskedastic? It seems like it still delivers us estimates of the coefficients that have highly desirable properties.
- The true variance of $\hat{\beta}_{OLS}$ is no longer $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, so that any inference based on $s^2(\mathbf{X}'\mathbf{X})^{-1}$ is likely to be “misleading”.
- Not only is the wrong matrix used, but s^2 may be a biased estimator of σ^2 .
- In general, there is no way to tell the direction of bias, although Goldberger (1964) shows that in the special case of only one explanatory variable (in addition to the constant term), s^2 is biased downward if high error variances correspond with high values of the independent variable.
- Whatever the direction of the bias, we should not use the standard equation for the standard errors in hypothesis tests on $\hat{\beta}_{OLS}$.
- More importantly, OLS is no longer efficient, since another method called **Generalized Least Squares** will give estimates of the regression coefficients, $\hat{\beta}_{GLS}$, that are unbiased and have a smaller variance.

2.3 Generalized Least Squares

- We assume, as we always do for a variance matrix, that $\mathbf{\Omega}$ is positive definite and symmetric. Thus, it can be factored into a set of matrices containing its characteristic roots and vectors.

$$\mathbf{\Omega} = \mathbf{C}\mathbf{\Lambda}\mathbf{C}'$$

- Here the columns of \mathbf{C} contain the characteristic vectors of $\mathbf{\Omega}$ and $\mathbf{\Lambda}$ is a diagonal matrix containing its characteristic roots. We can “factor” $\mathbf{\Omega}$ using the square root of $\mathbf{\Lambda}$, or $\mathbf{\Lambda}^{1/2}$, which is a matrix containing the square roots of the characteristic roots on the diagonal.
- Then, if $\mathbf{T} = \mathbf{C}\mathbf{\Lambda}^{1/2}$, $\mathbf{T}\mathbf{T}' = \mathbf{\Omega}$. The last result holds because $\mathbf{T}\mathbf{T}' = \mathbf{C}\mathbf{\Lambda}^{1/2}\mathbf{\Lambda}^{1/2}\mathbf{C}' = \mathbf{C}\mathbf{\Lambda}\mathbf{C}' = \mathbf{\Omega}$. Also, if we let $\mathbf{P}' = \mathbf{C}\mathbf{\Lambda}^{-1/2}$, then $\mathbf{P}'\mathbf{P} = \mathbf{\Omega}^{-1}$.

- It can be shown that the characteristic vectors are all orthogonal and for each characteristic vector, $\mathbf{c}'_i \mathbf{c}_i = 1$ (Greene, p. 825). It follows that $\mathbf{C}\mathbf{C}' = \mathbf{I}$, and $\mathbf{C}'\mathbf{C} = \mathbf{I}$, a fact that we will use below.
- GLS consists of estimating the following equation, using the standard OLS solutions for the regression coefficients:

$$\mathbf{P}\mathbf{y} = \mathbf{P}\mathbf{X}\boldsymbol{\beta} + \mathbf{P}\boldsymbol{\varepsilon}$$

or

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*.$$

This $\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}^{*\prime} \mathbf{X}^*)^{-1} (\mathbf{X}^{*\prime} \mathbf{y}^*) = (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{y})$

- It follows that the variance of $\boldsymbol{\varepsilon}^*$, is equal to $E[\boldsymbol{\varepsilon}^* \boldsymbol{\varepsilon}^{*\prime}] = \mathbf{P}\sigma^2 \boldsymbol{\Omega} \mathbf{P}'$, and:

$$\begin{aligned} \mathbf{P}\sigma^2 \boldsymbol{\Omega} \mathbf{P}' &= \sigma^2 \mathbf{P} \boldsymbol{\Omega} \mathbf{P}' = \sigma^2 \boldsymbol{\Lambda}^{-1/2} \mathbf{C}' \boldsymbol{\Omega} \mathbf{C} \boldsymbol{\Lambda}^{-1/2} \\ &= \sigma^2 \boldsymbol{\Lambda}^{-1/2} \mathbf{C}' \mathbf{C} \boldsymbol{\Lambda}^{1/2} \boldsymbol{\Lambda}^{1/2} \mathbf{C}' \mathbf{C} \boldsymbol{\Lambda}^{-1/2} \\ &= \sigma^2 \boldsymbol{\Lambda}^{-1/2} \boldsymbol{\Lambda}^{1/2} \boldsymbol{\Lambda}^{1/2} \boldsymbol{\Lambda}^{-1/2} \\ &= \sigma^2 \mathbf{I}_n \end{aligned}$$

2.3.1 Some intuition

- In the standard case of heteroskedasticity, GLS consists of dividing each observation by the square root of its own element in the $\boldsymbol{\Omega}$ matrix, $\sqrt{\omega_i}$. The nice thing about this is that the variance of $\boldsymbol{\varepsilon}^*$, equal to $E[\boldsymbol{\varepsilon}^* \boldsymbol{\varepsilon}^{*\prime}] = \mathbf{P}\sigma^2 \boldsymbol{\Omega} \mathbf{P}' = \sigma^2 \mathbf{I}$.
- In effect, we've removed the heteroskedasticity from the residuals, and we can go ahead and estimate the variance of $\hat{\boldsymbol{\beta}}_{GLS}$ using the formula $\sigma^2 (\mathbf{X}^{*\prime} \mathbf{X}^*)^{-1} = \sigma^2 (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1}$. This is also known as the Aitken estimator after the statistician who originally proposed the method in 1935.
- We can also show, using the standard methods, that $\hat{\boldsymbol{\beta}}_{GLS}$ is unbiased, consistent, and asymptotically normally distributed.
- We can conclude that $\hat{\boldsymbol{\beta}}_{GLS}$ is BLUE for the generalized model in which the variance of the errors is given by $\sigma^2 \boldsymbol{\Omega}$. The result follows by applying the Gauss-Markov theorem to model $\mathbf{y}^* = \mathbf{X}^* \boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$.

- In this general case, the maximum likelihood estimator will be the GLS estimator, $\hat{\beta}_{GLS}$, and that the Cramer-Rao lower bound for the variance of $\hat{\beta}_{GLS}$ is given by $\sigma^2(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}$.

2.4 Feasible Generalized Least Squares

- All of the above assumes that $\mathbf{\Omega}$ is a known matrix, which is usually not the case.
- One option is to estimate $\mathbf{\Omega}$ in some way and to use $\hat{\mathbf{\Omega}}$ in place of $\mathbf{\Omega}$ in the GLS model above. For instance, one might believe that the true error term was a function of one (or more) of the independent variables. Thus,

$$\varepsilon_i = \gamma_0 + \gamma_1 x_i + u_i$$

- Since $\hat{\beta}_{OLS}$ is consistent and unbiased for β , we can use the OLS residuals to estimate the model above. The procedure is that we first estimate the original model using OLS, and then use the residuals from this regression to estimate $\hat{\mathbf{\Omega}}$ and $\sqrt{\hat{\omega}_i}$.
- We then transform the data using these estimates, and use OLS again on the transformed data to estimate $\hat{\beta}_{GLS}$ and the correct standard errors for hypothesis testing.
- One can also use the estimated errors terms from the last stage to conduct FGLS again and keep on doing this until the error model begins to converge, so that the estimated residuals barely change as one moves through the iterations.
- Some examples:
 - If $\sigma_i^2 = \sigma^2 x_i^2$, we would divide all observations through by x_i .
 - If $\sigma_i^2 = \sigma^2 x_i$, we would divide all observations through by $\sqrt{x_i}$.
 - If $\sigma_i^2 = \sigma^2(\gamma_0 + \gamma_1 X_i + u_i)$, we would divide through all observations by $\sqrt{\gamma_0 + \gamma_1 x_i}$.

- **Essential Result:** So long as our estimate of $\mathbf{\Omega}$ is consistent, the FGLS estimator will be consistent and will be asymptotically efficient.
- **Problem:** Except for the simplest cases, the finite-sample properties and exact distributions of FGLS estimators are unknown. Intuitively, $\text{var}(\hat{\boldsymbol{\beta}}_{FGLS}) \neq \text{var}(\hat{\boldsymbol{\beta}}_{GLS})$ because we also have to take into account the uncertainty in estimating $\mathbf{\Omega}$. Since we cannot work out the small sample distribution, we cannot even say that $\hat{\boldsymbol{\beta}}_{FGLS}$ is unbiased in small samples.

2.5 White-Consistent Standard Errors

- If you can find a consistent estimator for $\mathbf{\Omega}$, then go ahead and perform FGLS if you have a large number of data points available.
- Otherwise, and in most cases, analysts estimate the regular OLS model and use “White-consistent standard errors”, which are described by Leamer as “White-washing heteroskedasticity”.
- White’s heteroskedasticity consistent estimator of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{OLS}$ is recommended whenever OLS estimates are being used for inference in a situation in which heteroskedasticity is suspected, but the researcher is unable to consistently estimate $\mathbf{\Omega}$ and use FGLS.
- White’s method consists of finding a consistent estimator for the true OLS variance below:

$$\text{var}[\hat{\boldsymbol{\beta}}_{OLS}] = \frac{1}{n} \left(\frac{1}{n} \mathbf{X}'\mathbf{X} \right)^{-1} \left(\frac{1}{n} \mathbf{X}'[\sigma^2 \mathbf{\Omega}]\mathbf{X} \right) \left(\frac{1}{n} \mathbf{X}'\mathbf{X} \right)^{-1}$$

- The trick to White’s estimation of the asymptotic variance-covariance matrix is to recognize that what we need is a consistent estimator of:

$$\mathbf{Q}_* = \frac{\sigma^2 \mathbf{X}'\mathbf{\Omega}\mathbf{X}}{n} = \frac{1}{n} \sum_{i=1}^n \sigma_i^2 \mathbf{x}_i \mathbf{x}_i'$$

- Under very general conditions

$$\mathbf{S}_0 = \frac{1}{n} \sum_{i=1}^n e_i^2 \mathbf{x}_i \mathbf{x}_i'$$

(which is a $k \times k$ matrix with $k(k+1)/2$ original terms) is consistent for \mathbf{Q}_* .
(See Greene p. 198)

- Because $\hat{\boldsymbol{\beta}}_{OLS}$ is consistent for $\boldsymbol{\beta}$, we can show that White's estimate of $\text{var}[\hat{\boldsymbol{\beta}}_{OLS}]$ is consistent for the true asymptotic variance.
- Thus, without specifying the exact nature of the heteroskedasticity, we can still calculate a consistent estimate of $\text{var}[\hat{\boldsymbol{\beta}}_{OLS}]$ and use this in the normal way to derive standard errors and conduct hypothesis tests. This makes the White-consistent estimator extremely attractive in a wide variety of situations.
- The White heteroskedasticity consistent estimator is

$$\begin{aligned} \text{Est. asy. var}[\boldsymbol{\beta}_{OLS}] &= \frac{1}{n} \left(\frac{1}{n} \mathbf{X}'\mathbf{X} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n e_i^2 \mathbf{x}_i \mathbf{x}_i' \right) \left(\frac{1}{n} \mathbf{X}'\mathbf{X} \right)^{-1} \\ &= n (\mathbf{X}'\mathbf{X})^{-1} \mathbf{S}_0 (\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

- Adding the **robust** qualifier to the regression command in Stata produces standard errors computed from this estimated asymptotic variance matrix.

2.6 Tests for Heteroskedasticity

- Homoskedasticity implies that the error variance will be the same across different observations and should not vary with \mathbf{X} , the independent variables.
- Unsurprisingly, all tests for heteroskedasticity rely on checking how far the error variances for different groups of observations depart from one another, or how precisely they are explained by the explanatory variables.

2.6.1 Visual Inspection of the Residuals

- Plot them and look for patterns (residuals against fitted values, residuals against explanatory vars).

2.6.2 The Goldfeld-Quandt Test

- The Goldfeld-Quandt tests consists of comparing the estimated error variances for two groups of observations.
- First, sort the data points by one of the explanatory variables (e.g., country size). Then run the model separately for the two groups of countries.
- If the true errors are homoskedastic, then the estimated error variances for the two groups should be approximately the same.
- We estimate the true error variance by $s_g^2 = (\mathbf{e}'_g \mathbf{e}_g)/(n_g - k)$, where the subscript g indicates that this is the value for each group. Since each estimated error variance is distributed chi-squared, the test statistic below is distributed F .
- If the true error is homoskedastic, the statistic will be approximately equal to one. The larger the F -statistic, the less likely it is that the errors are homoskedastic. It should be noted that the formulation below is computed assuming that the errors are higher for the first group.

$$\frac{\mathbf{e}'_1 \mathbf{e}_1 / (n_1 - k)}{\mathbf{e}'_2 \mathbf{e}_2 / (n_2 - k)} = F[n_1 - k, n_2 - k]$$

- It has also been suggested that the test-statistic should be calculated using only the first and last thirds of the data points, excluding the middle section of the data, to sharpen the test results.

2.6.3 The Breusch-Pagan Test

- The main problem w/ Goldfeld-Quandt: requires knowledge of how to sort the data points.
- The Breusch-Pagan test uses a different approach to see if the error variances are systematically related to the independent variables, or to any subset or transformation of those variables.
- The idea behind the test is that for the regression

$$\sigma_i^2 = \sigma^2 f(\alpha_0 + \boldsymbol{\alpha}'\mathbf{z}_i)$$

where \mathbf{z}_i is a vector of independent variables, if $\boldsymbol{\alpha} = \mathbf{0}$ then the errors are homoskedastic.

- To perform the test, regress $e_i^2/(\mathbf{e}'\mathbf{e}/n - 1)$ on some combination of the independent variables (\mathbf{z}_i).
- We then compute a Lagrange Multiplier statistic as the regression sum of squares from that regression divided by 2. That is,

$$\frac{SSR}{2} \stackrel{a}{\sim} \chi_{m-1}^2,$$

where m is the number of explanatory variables in the auxiliary regression.

- Intuition: if the null hypothesis of no heteroskedasticity is true, then the SSR should equal zero. A non-zero SSR is telling us that $e_i^2/(\mathbf{e}'\mathbf{e}/n - 1)$ varies with the explanatory variables.
- Note that this test depends on the fact that $\boldsymbol{\beta}_{OLS}$ is consistent for $\boldsymbol{\beta}$, and that the errors from an OLS regression of y_i on \mathbf{x}_i , e_i will be consistent for ε_i .

Section 3

Autocorrelation

3.1 The Meaning of Autocorrelation

- Auto-correlation is often paired with heteroskedasticity because it is another way in which the variance-covariance matrix of the true error terms (if we could observe it) is different from the Gauss-Markov assumption, $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \sigma^2\mathbf{I}_n$.
- In our earlier discussion of heteroskedasticity, we saw what happened when we relax the assumption that the variance of the error term is constant. An equivalent way of saying this is that we relax the assumption that the errors are “identically distributed.” In this section, we see what happens when we relax the assumption that the error terms are *independent*.
- In this case, we can have errors that covary (e.g., if one error is positive and large the next error is likely to be positive and large) and are correlated. In either case, one error can give us information about another.
- Two types of error correlation:
 1. *Spatial correlation*: e.g., of contiguous households, states, or counties.
 2. *Temporal or Autocorrelation*: errors from adjacent time periods are correlated with one another. Thus, ε_t is correlated with $\varepsilon_{t+1}, \varepsilon_{t+2}, \dots$ and $\varepsilon_{t-1}, \varepsilon_{t-2}, \dots, \varepsilon_1$.
- The correlation between ε_t and ε_{t-k} is called autocorrelation of order k .
 - The correlation between ε_t and ε_{t-1} is the first-order autocorrelation and is usually denoted by ρ_1 .
 - The correlation between ε_t and ε_{t-2} is the second-order autocorrelation and is usually denoted by ρ_2 .

- What does this imply for $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']$?

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \begin{bmatrix} E(\varepsilon_1^2) & E(\varepsilon_1\varepsilon_2) & \dots & E(\varepsilon_1\varepsilon_T) \\ \vdots & \ddots & & \vdots \\ E(\varepsilon_T\varepsilon_1) & E(\varepsilon_T\varepsilon_2) & \dots & E(\varepsilon_T^2) \end{bmatrix}$$

Here, the off-diagonal elements are the covariances between the different error terms. Why is this?

- Autocorrelation implies that the off-diagonal elements are *not* equal to zero.

3.2 Causes of Autocorrelation

- Misspecification of the model
- Data manipulation, smoothing, seasonal adjustment
- Prolonged influence of shocks
- Inertia

3.3 Consequences of Autocorrelation for Regression Coefficients and Standard Errors

- As with heteroskedasticity, in the case of autocorrelation, the β_{OLS} regression coefficients are still unbiased and consistent. Note, this is true only in the case where the model does *not* contain a lagged dependent variable. We will come to this latter case in a bit.
- As before, OLS is inefficient. Further, inferences based on the standard OLS estimate of $\text{var}(\beta_{OLS}) = s^2(\mathbf{X}'\mathbf{X})^{-1}$ are wrong.
- OLS estimation in the presence of autocorrelation is likely to lead to an underestimate of σ^2 , meaning that our t -statistics will be inflated, and we are more likely to reject the null when we should not do so.

3.4 Tests for Autocorrelation

- Visual inspection, natch.
- Test statistics
 - There are two well-known test statistics to compute to detect autocorrelation.
 1. Durbin-Watson test: most well-known but does not always give unambiguous answers and is not appropriate when there is a lagged dependent variable (LDV) in the original model.
 2. Breusch-Godfrey (or LM) test: does not have a similar indeterminate range (more later) and can be used with a LDV.
 - Both tests use the fact, exploited in tests for heteroskedasticity, that since β_{OLS} is consistent for β , the residuals, e_i , will be consistent for ε_i . Both tests, therefore, use the residuals from a preceding OLS regression to estimate the actual correlations between error terms.
 - At this point, it is useful to remember the expression for the correlation of two variables:

$$\text{corr}(x, y) = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)}\sqrt{\text{var}(y)}}$$

- For the error terms, and if we assume that $\text{var}(\varepsilon_t) = \text{var}(\varepsilon_{t-1}) = \varepsilon_t^2$, then we have

$$\text{corr}(\varepsilon_t, \varepsilon_{t-1}) = \frac{E[\varepsilon_t \varepsilon_{t-1}]}{E[\varepsilon_t^2]} = \rho_1$$

- To get the sample average of this autocorrelation, we would compute this ratio for each pair of succeeding error terms, and divide by n .

3.4.1 The Durbin-Watson Test

- The Durbin-Watson test does something similar to estimating the sample autocorrelation between error terms. The DW statistic is calculated as:

$$d = \frac{\sum_{t=2}^T (e_t - e_{t-1})^2}{\sum_{t=1}^T e_t^2}$$

- We can write d as $d = \frac{\sum e_t^2 + \sum e_{t-1}^2 - 2 \sum e_t e_{t-1}}{\sum e_t^2}$
- Since $\sum e_t^2$ is approximately equal to $\sum e_{t-1}^2$ as the sample size becomes large, and $\frac{\sum e_t e_{t-1}}{\sum e_t^2}$ is the sample average of the autocorrelation coefficient, ρ , we have that $d \cong 2(1 - \hat{\rho})$.
- If $\rho = +1$, then $d = 0$, if $\rho = -1$, then $d = 4$. If $\rho = 0$, then $d = 2$. Thus, if d is close to zero or four, the residuals can be said to be highly correlated.
- The problem is that the exact sampling distribution of d depends on the values of the explanatory variables. To get around this problem, Durbin and Watson have derived upper (d_u) and lower limits (d_l) for the significance levels of d .
- If we were testing for positive autocorrelation, versus the null hypothesis of no autocorrelation, we would use the following procedure:
 - If $d < d_l$, reject the null of no autocorrelation.
 - If $d > d_u$, we do not reject the null hypothesis.
 - If $d_l < d < d_u$, the test is inconclusive.
- Can be calculated within Stata using **dwstat**.

3.4.2 The Breusch-Godfrey Test

- To perform this test, simply regress the OLS residuals, e_t , on all the explanatory variables, \mathbf{x}_t , and on their own lags, $e_{t-1}, e_{t-2}, e_{t-3}, \dots, e_{t-p}$.
- This test works because the coefficient on the lagged error terms will only be significant if the *partial* autocorrelation between e_t and e_{t-p} is significant. The partial autocorrelation is the autocorrelation between the error terms accounting for the effect of the \mathbf{X} explanatory variables.
- Thus, the null hypothesis of no autocorrelation can be tested using an F -test to see whether the coefficients on the lagged error terms, $e_{t-1}, e_{t-2}, e_{t-3}, \dots, e_{t-p}$ are jointly equal to zero.
- If you would like to be precise, and consider that since we are using e_t as an estimate of ε_t , you can treat this as a large sample test, and say that $p \cdot F$, where p is the number of lagged error terms restricted to be zero, is asymptotically distributed chi-squared with degrees of freedom equal to p . Then use the chi-squared distribution to give critical values for the test.
- Using what is called the Lagrange Multiplier (LM) approach to testing, we can also show that $n \cdot R^2$ is asymptotically distributed as χ_p and use this statistic for testing.

3.5 The consequences of autocorrelation for the variance-covariance matrix

- In the case of heteroskedasticity, we saw that we needed to make some assumptions about the form of the $E[\varepsilon\varepsilon']$ matrix in order to be able to transform the data and use FGLS to calculate correct and efficient standard errors.
- Under auto-correlation, the off-diagonal elements will not be equal to zero, but we don't know what those off-diagonal elements are. In order to conduct FGLS, we usually make some assumption about the form of autocorrelation and the process by which the disturbances are generated.

- First, we must write the $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']$ matrix in a way that can represent autocorrelation:

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \sigma^2\boldsymbol{\Omega}$$

- Recall that

$$\text{corr}(\varepsilon_t, \varepsilon_{t-s}) = \frac{E[\varepsilon_t\varepsilon_{t-s}]}{E[\varepsilon_t^2]} = \rho_s = \frac{\gamma_s}{\gamma_0}$$

where $\gamma_s = \text{cov}(\varepsilon_t, \varepsilon_{t-s})$ and $\gamma_0 = \text{var}(\varepsilon_t)$

- Let \mathbf{R} be the “autocorrelation matrix” showing the correlation between all the disturbance terms. Then $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \gamma_0\mathbf{R}$
- Second, we need to calculate the autocorrelation matrix. It helps to make an assumption about the process generating the disturbances or true errors. The most frequent assumption is that the errors follow an “autoregressive process” of order one, written as AR(1).

An AR(1) process is represented as:

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

where $E[u_t] = 0$, $E[u_t^2] = \sigma_u^2$, and $\text{cov}[u_t, u_s] = 0$ if $t \neq s$.

- By repeated substitution, we have:

$$\varepsilon_t = u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \dots$$

- Each disturbance, ε_t , embodies the entire past history of the u s, with the most recent shocks receiving greater weight than those in the more distant past.
- The successive values of u_t are uncorrelated, so we can estimate the variance of ε_t , which is equal to $E[\varepsilon_t^2]$, as:

$$\text{var}[\varepsilon_t] = \sigma_u^2 + \rho^2\sigma_u^2 + \rho^4\sigma_u^2 + \dots$$

- This is a series of positive numbers (σ_u^2) multiplied by increasing powers of ρ . If ρ is greater than one, the series will be infinite and we won't be able to get an expression for $\text{var}[\varepsilon_t]$. To proceed, we assume that $|\rho| < 1$.

- Here is a useful trick for a series. If we have an infinite series of numbers:

$$y = a^0 \cdot x + a^1 \cdot x + a^2 \cdot x + a^3 \cdot x + \dots$$

then

$$y = \frac{x}{1 - a}.$$

- Using this, we see that

$$\text{var}[\varepsilon_t] = \frac{\sigma_u^2}{1 - \rho^2} = \sigma_\varepsilon^2 = \gamma_0 \quad (3.1)$$

- We can also estimate the covariances between the errors:

$$\text{cov}[\varepsilon_t, \varepsilon_{t-1}] = E[\varepsilon_t \varepsilon_{t-1}] = E[\varepsilon_{t-1}(\rho \varepsilon_{t-1} + u_t)] = \rho \text{var}[\varepsilon_{t-1}] = \frac{\rho \sigma_u^2}{1 - \rho^2}$$

- And, since $\varepsilon_t = \rho \varepsilon_{t-1} + u_t = \rho(\rho \varepsilon_{t-2} + u_{t-1}) + u_t = \rho^2 \varepsilon_{t-2} + \rho u_{t-1} + u_t$ we have

$$\text{cov}[\varepsilon_t, \varepsilon_{t-2}] = E[\varepsilon_t \varepsilon_{t-2}] = E[\varepsilon_{t-2}(\rho^2 \varepsilon_{t-2} + \rho u_{t-1} + u_t)] = \rho^2 \text{var}[\varepsilon_{t-2}] = \frac{\rho^2 \sigma_u^2}{1 - \rho^2}$$

- By repeated substitution, we can see that:

$$\text{cov}[\varepsilon_t, \varepsilon_{t-s}] = E[\varepsilon_t \varepsilon_{t-s}] = \frac{\rho^s \sigma_u^2}{1 - \rho^2} = \gamma_s$$

- Now we can get back to the autocorrelations. Since $\text{corr}[\varepsilon_t, \varepsilon_{t-s}] = \frac{\gamma_s}{\gamma_0}$, we have:

$$\text{corr}[\varepsilon_t, \varepsilon_{t-s}] = \frac{\frac{\rho^s \sigma_u^2}{1 - \rho^2}}{\frac{\sigma_u^2}{1 - \rho^2}} = \rho^s$$

- In other words, the auto-correlations fade over time. They are always less than one and become less and less the farther two disturbances are apart in time.

- The auto-correlation matrix, \mathbf{R} , shows all the auto-correlations between the disturbances. Given the link between the auto-correlation matrix and $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \sigma^2\boldsymbol{\Omega}$, we can now say that:

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \sigma^2\boldsymbol{\Omega} = \gamma_0\mathbf{R}$$

or

$$\sigma^2\boldsymbol{\Omega} = \frac{\sigma_u^2}{1 - \rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \rho^2 & \dots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \rho & \dots & \rho^{T-3} \\ \vdots & & & & \dots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & & \dots & 1 \end{bmatrix}$$

3.6 GLS and FGLS under Autocorrelation

- Given that we now have the expression for $\sigma^2\boldsymbol{\Omega}$, we can in theory transform the data and estimate via GLS (or use an estimate, $\hat{\sigma}^2\hat{\boldsymbol{\Omega}}$, for FGLS).
- We are transforming the data so that we get an error term that conforms to the Gauss-Markov assumptions. In the heteroskedasticity case, we showed how we transformed the matrix using the factor for $\sigma^2\boldsymbol{\Omega}$.
- In the case of autocorrelation of the AR(1) type, what's nice is that the transformation is fairly simple, even though the matrix expression for $\sigma^2\boldsymbol{\Omega}$ may not look that simple.
- Let's start with a simple model with an AR(1) error term:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \quad t = 1, 2, \dots, T \quad (3.2)$$

where $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$

- Now lag the data by one period and multiply it by ρ :

$$\rho y_{t-1} = \beta_0 \rho + \beta_1 \rho x_{t-1} + \rho \varepsilon_{t-1}$$

- Subtracting this from Equation 3.2 above, we get:

$$y_t - \rho y_{t-1} = \beta_0(1 - \rho) + \beta_1(x_t - \rho x_{t-1}) + u_t$$

- But (given our assumptions) the u_t s are serially independent, have constant variance (σ_u^2), and the covariance between different u s is zero. Thus, the new error term, u_t , conforms to all the desirable features of Gauss-Markov errors.
- If we conduct OLS using this transformed data, and use our normal estimate for the standard errors, or $s^2(\mathbf{X}'\mathbf{X})^{-1}$, we will get correct and efficient standard errors.
- The transformation is:

$$y_t^* = y_t - \rho y_{t-1}$$

$$x_t^* = x_t - \rho x_{t-1}$$

- If we drop the first observation (because we don't have a lag for it), then we are following the *Cochrane-Orcutt* procedure. If we keep the first observation and use it with the following transformation:

$$y_1^* = \sqrt{1 - \rho^2} y_1$$

$$x_1^* = \sqrt{1 - \rho^2} x_1$$

then we are following something called the *Prais-Winsten* procedure. Both of these are examples of FGLS.

- But how do we do this if we don't "know" ρ ? Since $\hat{\beta}_{OLS}$ is unbiased, our estimated residuals are unbiased for the true disturbances, ε_t .
- In this case, we can estimate ρ using our residuals from an initial OLS regression, estimate ρ , and then perform FGLS using this estimate. The standard errors that we calculate using this procedure will be asymptotically efficient.
- To estimate ρ from the residuals, e_t compute:

$$\hat{\rho} = \frac{\sum e_t e_{t-1}}{\sum e_t^2}$$

We could also estimate ρ from a regression of e_t on e_{t-1} .

- To sum up this procedure:
 1. Estimate the model using OLS. Test for autocorrelation. If tests reveal this to be present, estimate the autocorrelation coefficient, ρ , using the residuals from the OLS estimation.
 2. Transform the data.
 3. Estimate OLS using the transformed data.
- For Prais-Winsten in Stata do:


```
prais depvar expvars
```

 or


```
prais depvar expvars, corc
```

 for Cochrane-Orcutt.

3.7 Non-AR(1) processes

- A problem with the Prais-Winsten and Cochrane-Orcutt versions of FGLS is that the disturbances may not be distributed AR(1). In this situation, we will have used the wrong assumptions to estimate the auto-correlations.
- There is an approach, analogous to the White-consistent standard errors, that directly estimates the correct OLS standard errors (from $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{\Omega})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$), using an estimate of $\mathbf{X}'(\sigma^2\mathbf{\Omega})\mathbf{X}$ based on very general assumptions about the autocorrelation of the error terms. These standard errors are called *Newey-West* standard errors.
- Stata users can do the following command to compute these:


```
newey depvar expvars, lag(#) t(varname)
```

 In this case, `lag(#)` tells Stata how many lags there are between any two disturbances before the autocorrelations die out to zero. `t(varname)` tells Stata the variable that indexes time.

3.8 OLS Estimation with Lagged Dependent Variables and Autocorrelation

- We said previously that autocorrelation, like heteroskedasticity, does not affect the unbiasedness of the OLS estimates, just the means by which we calculate efficient and correct standard errors. There is one important exception to this.
- Suppose that we use lagged values of the dependent variable, y_t as a regressor, so that:

$$y_t = \beta y_{t-1} + \varepsilon_t \quad \text{where } \beta < 1$$

and

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t$$

where $E[u_t] = 0$, $E[u_t^2] = \sigma_u^2$, and $\text{cov}[u_t, u_s] = 0$ if $t \neq s$.

- The OLS estimate of $\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ where $\mathbf{X} = y_{t-1}$. Let us also assume that y_t and y_{t-1} have been transformed so that they are measured as deviations from their means.

$$\begin{aligned} \hat{\beta}_{OLS} &= \frac{\sum_{t=2}^T y_{t-1}y_t}{\sum_{t=2}^T y_{t-1}^2} = \frac{\sum_{t=2}^T y_{t-1}(\beta y_{t-1} + \varepsilon_t)}{\sum_{t=2}^T y_{t-1}^2} \\ &= \beta + \frac{\sum_{t=2}^T y_{t-1}\varepsilon_t}{\sum_{t=2}^T y_{t-1}^2} \end{aligned}$$

- Thus,

$$E[\hat{\beta}_{OLS}] = \beta + \frac{\text{cov}(y_{t-1}, \varepsilon_t)}{\text{var}(y_t)}$$

- To show that $\hat{\beta}_{OLS}$ is biased, we need to show only that $\text{cov}(y_{t-1}, \varepsilon_t) \neq 0$, and to show that $\hat{\beta}_{OLS}$ is inconsistent, we need to show only that the limit of this covariance as $T \rightarrow \infty$ is not equal to zero.

$$\text{cov}[y_{t-1}, \varepsilon_t] = \text{cov}[y_{t-1}, \rho\varepsilon_{t-1} + u_t] = \rho\text{cov}[y_{t-1}, \varepsilon_{t-1}] = \rho\text{cov}[y_t, \varepsilon_t]$$

- The last step is true assuming the DGP is “stationary” and the u_t s are uncorrelated.
- Continuing:

$$\rho\text{cov}[y_t, \varepsilon_t] = \rho\text{cov}[\beta y_{t-1} + \varepsilon_t, \varepsilon_t] = \rho\{\beta\text{cov}[y_{t-1}, \varepsilon_t] + \text{cov}[\varepsilon_t, \varepsilon_t]\} \quad (3.3)$$

$$= \rho\{\beta\text{cov}[y_{t-1}, \varepsilon_t] + \text{var}[\varepsilon_t]\} \quad (3.4)$$

- Since $\text{cov}[y_{t-1}, \varepsilon_t] = \rho\text{cov}[y_t, \varepsilon_t]$ (from above) we have:

$$\text{cov}[y_{t-1}, \varepsilon_t] = \rho\beta\text{cov}[y_{t-1}, \varepsilon_t] + \rho\text{var}[\varepsilon_t]$$

so that

$$\text{cov}[y_{t-1}, \varepsilon_t] = \frac{\rho\text{var}[\varepsilon_t]}{(1 - \beta\rho)}.$$

- Given our calculation of $\text{var}[\varepsilon_t]$ when the error is AR(1) (see Eq. 3.1):

$$\text{cov}[y_{t-1}, \varepsilon_t] = \frac{\rho\sigma_u^2}{(1 - \beta\rho)(1 - \rho^2)}$$

- Thus, if ρ is positive, the estimate of β is biased upward (more of the fit is imputed to the lagged dependent variable than to the systematic relation between the error terms). Moreover, since the covariance above will not diminish to zero in the limit as $T \rightarrow \infty$, the estimated regression coefficients will also be inconsistent.
- It can also be shown that the estimate of ρ in the Durbin-Watson test will be biased downward, leading us to accept the null hypothesis of no autocorrelation too often. For that reason, when we include a lagged dependent variable in the model, we should be careful to use the Breusch-Godfrey test to determine whether autocorrelation is present.

3.9 Bias and “Contemporaneous Correlation”

- A more immediate question is what general phenomenon produces this bias and how to obtain unbiased estimates when this problem exists. The answer turns out to be a general one, so it is worth exploring.
- The bias (and inconsistency) in this case arose because of a violation of the G-M assumptions $E[\boldsymbol{\varepsilon}] = 0$ and \mathbf{X} is a known matrix of constants (“fixed in repeated samples”). We needed this so that we could say that $E[\mathbf{X}'\boldsymbol{\varepsilon}] = 0$. This yielded our proof of unbiasedness.

$$E[\hat{\boldsymbol{\beta}}_{OLS}] = \boldsymbol{\beta} + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}] = \boldsymbol{\beta}$$

- Recall we showed that we could relax the assumption that \mathbf{X} is fixed in repeated samples if we substituted the assumption that \mathbf{X} is “strictly exogenous”, so that $E[\boldsymbol{\varepsilon}|\mathbf{X}] = 0$. This will also yield the result that $E[\mathbf{X}'\boldsymbol{\varepsilon}] = 0$
- The problem in the case of the lagged dependent variable is that $E[\mathbf{X}'\boldsymbol{\varepsilon}] \neq 0$. When there is a large and positive error in the preceding period, we would be likely to get a large and positive ε_t and a large and positive y_{t-1} .
- Thus, we see a positive relationship between the current error and the regressor. This is all that we need to get bias, a “contemporaneous” correlation between a regressor and the error term, such that $\text{cov}[x_t, \varepsilon_t] \neq 0$.

3.10 Measurement Error

- Another case in which we get biased and inconsistent estimates because of contemporaneous correlation between a regressor and the error is measurement error.
- Suppose that the true relationship does not contain an intercept and is:

$$y = \beta x + \varepsilon$$

but x is measured with error as z , where $z = x + u$ is what we observe and $E[u_t] = 0$, $E[u_t^2] = \sigma_u^2$.

- This implies that x can be written as $z - u$ and the true model can then be written as:

$$y = \beta(z - u) + \varepsilon = \beta z + (\varepsilon - \beta u) = \beta z + \eta$$

The new disturbance, η , is a function of u (the error in measuring x). z is also a function of u . This sets up a non-zero covariance (and correlation) between z , our imperfect measure of the true regressor x , and the new disturbance term, $(\varepsilon - \beta u)$.

- The covariance between them, as we showed then, is equal to:

$$\text{cov}[z, (\varepsilon - \beta u)] = \text{cov}[x + u, \varepsilon - \beta u] = -\beta\sigma_u^2$$

- How does this lead to bias? We go back to the original equation for the expectation of the estimated $\hat{\beta}$ coefficient, $E[\hat{\beta}] = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$. It is easy to show:

$$E[\hat{\beta}] = E[(z'z)^{-1}(z'y)] = E[(z'z)^{-1}(z'(\beta z + \eta))] \quad (3.5)$$

$$= E[[\sum_i (x + u)'(x + u)]^{-1} \sum_i (x + u)'(\beta x + \varepsilon)] \quad (3.6)$$

$$= E\left(\frac{\sum_i (x + u)(\beta x + \varepsilon)}{\sum_i (x + u)^2}\right) = \frac{\beta x^2}{x^2 + \sigma_u^2} = \beta \frac{x^2}{x^2 + \sigma_u^2} \quad (3.7)$$

- Note: establishing a non-zero covariance between the regressor and the error term is sufficient to prove bias but is not the same as indicating the direction of the bias. In this special case, however, $\hat{\beta}$ is biased downward.
- To show consistency we must show that:

$$\text{plim } \hat{\beta} = \beta + \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} \left(\frac{\mathbf{X}'\varepsilon}{n}\right) = \beta \quad (\text{recall } \mathbf{Q}^* = \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1})$$

- This involves having to show that $\text{plim } \left(\frac{\mathbf{X}'\varepsilon}{n}\right) = 0$.
- In the case above,

$$\text{plim } \hat{\beta} = \left(\frac{\text{plim } (1/n) \sum_i (x + u)(\beta x + \varepsilon)}{\text{plim } (1/n) \sum_i (x + u)^2}\right) = \frac{\beta \mathbf{Q}^*}{\mathbf{Q}^* + \sigma_u^2}$$

- Describing the bias and inconsistency in the case of a lagged dependent variable with autocorrelation would follow the same procedure. We would look at the expectation term to show the bias and the probability limit to show inconsistency. See particularly Greene, p. 85.

3.11 Instrumental Variable Estimation

- In any case of contemporaneous correlation between a regressor and the error term, we can use an approach known as instrumental variable (or IV) estimation. The intuition to this approach is that we will find an instrument for X , where X is the variable correlated with the error term. This instrument should ideally be highly correlated with X , but uncorrelated with the error term.
- We can use more than one instrument to estimate β_{IV} . Say that we have one variable, x_1 , measured with error in the model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- We believe that there exists a set of variables, \mathbf{Z} , that is correlated with x_1 but not with ε . We estimate the following model:

$$x_1 = \mathbf{Z}\boldsymbol{\alpha} + u$$

where the $\boldsymbol{\alpha}$ are the regression coefficients on \mathbf{Z} where $\boldsymbol{\alpha} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{x}_1$.

- We then calculate the fitted or predicted values of x_1 , or \tilde{x}_1 , equal to $\mathbf{Z}\boldsymbol{\alpha}$, which is $\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{x}_1$. These fitted values should not be correlated with the error term because they are derived from instrumental variables that are uncorrelated with the error term.
- We then use the fitted values \tilde{x}_1 in the original model for x_1 .

$$y = \beta_0 + \beta_1 \tilde{x}_1 + \beta_2 x_2 + \varepsilon$$

This gives an unbiased estimate β_{IV} of β_1 .

In the case of a model with autocorrelation and a lagged dependent variable, Hatanaka (1974), suggests the following IV estimation for the model:

$$y_t = \mathbf{X}\boldsymbol{\beta} + \gamma y_{t-1} + \varepsilon_t$$

where $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$

- Use the predicted values from a regression of y_t on \mathbf{X}_t and \mathbf{X}_{t-1} as an estimate of y_{t-1} . The coefficient on \tilde{y}_{t-1} is a consistent estimate of γ , so it can be used to estimate ρ and perform FGLS.

3.12 In the general case, why is IV estimation unbiased and consistent?

- Suppose that

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where \mathbf{X} contains one regressor that is contemporaneously correlated with the error term and the other variables are uncorrelated. The intercept and the variables uncorrelated with the error can serve as their own (perfect) instruments.

- Each instrument is correlated with the variable of interest and uncorrelated with the error term. We have at least one instrument for the explanatory variable correlated with the error term. By regressing \mathbf{X} on \mathbf{Z} we get $\tilde{\mathbf{X}}$, the predicted values of \mathbf{X} . For each uncorrelated variable, the predicted value is just itself since it perfectly predicts itself. For the correlated variables, the predicted value is the value given by the first-stage model.

$$\tilde{\mathbf{X}} = \mathbf{Z}\boldsymbol{\alpha} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$$

- Then

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{IV} &= (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{y} \\ &= [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}) \end{aligned}$$

- Simplifying and substituting for \mathbf{y} we get:

$$\begin{aligned}
E[\hat{\boldsymbol{\beta}}_{IV}] &= E[(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{y}] \\
&= E[(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}))] \\
&= E\{[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})\boldsymbol{\beta} \\
&\quad + [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon})\} \\
&= \boldsymbol{\beta} + 0
\end{aligned}$$

since $E[\mathbf{Z}'\boldsymbol{\varepsilon}]$ is zero by assumption.

Section 4

Simultaneous Equations Models and 2SLS

4.1 Simultaneous Equations Models and Bias

- Where we have a system of *simultaneous equations* we can get biased and inconsistent regression coefficients for the same reason as in the case of measurement error (i.e., we introduce a “contemporaneous correlation” between at least one of the regressors and the error term).

4.1.1 Motivating Example: Political Violence and Economic Growth

- We are interested in the links between economic growth and political violence. Assume that we have good measures of both.
- We could estimate the following model

$$Growth = f(Violence, OtherFactors)$$

(This is the bread riots are bad for business model).

- We could also estimate a model

$$Violence = f(Growth, OtherFactors)$$

(This is the hunger and privation causes bread riots model).

- We could estimate either by OLS. But what if both are true? Then violence helps to explain growth and growth helps to explain violence.
- In previous estimations we have treated the variables on the right-hand side as *exogenous*. In this case, however, some of them are *endogenous* because they are themselves explained by another causal model. The model now has two equations:

$$G_i = \beta_0 + \beta_1 V_i + \varepsilon_i$$

$$V_i = \alpha_0 + \alpha_1 G_i + \eta_i$$

4.1.2 Simultaneity Bias

- What happens if we just run OLS on the equation we care about? In general, this is a bad idea (although we will get into the case in which we can). The reason is simultaneity bias.
- To see the problem, we will go through the following simulation.
 1. Suppose some random event sparks political violence (η_i is big). This could be because you did not control for the actions of demagogues in stirring up crowds.
 2. This causes growth to fall through the effect captured in β_1 .
 3. Thus, G_i and η_i are negatively correlated.
- What happens in this case if we estimate $V_i = \alpha_0 + \alpha_1 G_i + \eta_i$ by OLS without taking the simultaneity into account? We will mis-estimate α_1 because growth tends to be low when η_i is high.
- In fact, we are likely to estimate α_1 with a negative bias, because $E(\mathbf{G}'\boldsymbol{\eta})$ is negative. We may even produce the mistaken result that low growth produces high violence solely through the simultaneity.
- To investigate the likely direction and existence of bias, let's look at $E[\mathbf{X}'\boldsymbol{\varepsilon}]$ or the covariance between the explanatory variables and the error term.

$$E[\mathbf{G}'\boldsymbol{\eta}] = E[(\beta_0 + \beta_1 V + \boldsymbol{\varepsilon})\boldsymbol{\eta}] = E[(\beta_0 + \beta_1(\alpha_0 + \alpha_1 G + \boldsymbol{\eta}) + \boldsymbol{\varepsilon})\boldsymbol{\eta}]$$

- Multiplying through and taking expectations we get:

$$E[\mathbf{G}'\boldsymbol{\eta}] = E[(\beta_0 + \beta_1\alpha_1)'\boldsymbol{\eta}] + E[\beta_1\alpha_1 G'\boldsymbol{\eta}] + E[\beta_1\boldsymbol{\eta}'\boldsymbol{\eta}] + E[\boldsymbol{\varepsilon}'\boldsymbol{\eta}]$$

- Passing through the expectations operator we get:

$$E[\mathbf{G}'\boldsymbol{\eta}] = (\beta_0 + \beta_1\alpha_1)E[\boldsymbol{\eta}] + \beta_1\alpha_1 E[G'\boldsymbol{\eta}] + \beta_1 E[\boldsymbol{\eta}'\boldsymbol{\eta}] + E[\boldsymbol{\varepsilon}]E[\boldsymbol{\eta}]$$

($E[\boldsymbol{\varepsilon}'\boldsymbol{\eta}] = E[\boldsymbol{\varepsilon}]E[\boldsymbol{\eta}]$ because the two error terms are assumed independent).

- Since $E[\boldsymbol{\varepsilon}] = 0$ and $E[\boldsymbol{\eta}] = 0$ and $E[\boldsymbol{\eta}^2] = \sigma_{\boldsymbol{\eta}}^2$, we get:

$$E[G'\boldsymbol{\eta}] = \beta_1\alpha_1E[G'\boldsymbol{\eta}] + \beta_1\sigma_{\boldsymbol{\eta}}^2$$

So, $(1 - \beta_1\alpha_1)E[G'\boldsymbol{\eta}] = \beta_1\sigma_{\boldsymbol{\eta}}^2$ and $E[G'\boldsymbol{\eta}] = \frac{\beta_1\sigma_{\boldsymbol{\eta}}^2}{(1-\beta_1\alpha_1)}$

- Since this is non-zero, we have a bias in the estimate of α_1 , the effect of growth, G , on violence, V .
- The equation above indicates that the size of bias will depend on the variance of the disturbance term, η , and the size of the coefficient on V , β_1 . These parameters jointly determine the magnitude of the feedback effect from errors in the equation for violence, through the impact of violence, back onto growth.
- The actual computation of the bias term is not quite this simple in the case of the models above because we will have other variables in the matrix of right-hand side variables, \mathbf{X} .
- Nevertheless, the expression above can help us to deduce the likely direction of bias on the endogenous variable. When we have other variables in the model, the coefficients on these variables will also be affected by bias, but we cannot tell the direction of this bias a priori.
- How can we estimate the causal effects without bias in the presence of simultaneity? Doing so will involve re-expressing all the endogenous variables as a function of the exogenous variables and the error terms. This leads to the question of the “identification” of the system.

4.2 Identifying the Endogenous Variables

- In order to estimate a system of simultaneous equations (or any equation in that system) the model must be either “just identified” or “over-identified”. These conditions depend on the exogenous variables in the system of equations.

- Suppose we have the following system of simultaneous equations:

$$G_i = \beta_0 + \beta_1 V_i + \beta_2 X_i + \varepsilon_i$$

$$V_i = \alpha_0 + \alpha_1 G_i + \alpha_2 Z_i + \eta_i$$

- We can insert the second equation into the first in order to derive an expression for growth that does not involve the endogenous variable, violence:

$$G_i = \beta_0 + \beta_1(\alpha_0 + \alpha_1 G_i + \alpha_2 Z_i + \eta_i) + \beta_2 X_i + \varepsilon_i$$

Then you can pull all the expressions involving G_i onto the left-hand side and divide through to give:

$$G_i = \frac{\beta_0 + \beta_1 \alpha_0}{(1 - \beta_1 \alpha_1)} + \frac{\beta_1 \alpha_2}{(1 - \beta_1 \alpha_1)} Z_i + \frac{\beta_2}{(1 - \beta_1 \alpha_1)} X_i + \left(\frac{\varepsilon_i + \beta_1 \eta_i}{1 - \beta_1 \alpha_1} \right)$$

- We can perform a similar exercise to write the violence equation as:

$$V_i = \frac{\alpha_0 + \alpha_1 \beta_0}{(1 - \alpha_1 \beta_1)} + \frac{\alpha_1 \beta_2}{(1 - \alpha_1 \beta_1)} X_i + \frac{\alpha_2}{(1 - \alpha_1 \beta_1)} Z_i + \left(\frac{\eta_i + \alpha_1 \varepsilon_i}{1 - \alpha_1 \beta_1} \right)$$

- We can then estimate these *reduced form equations* directly.

$$G_i = \gamma_0 + \gamma_1 Z_i + \gamma_2 X_i + \varepsilon_i^*$$

$$V_i = \phi_0 + \phi_1 Z_i + \phi_2 X_i + \eta_i^*$$

Where,

$$\gamma_0 = \frac{\beta_0 + \beta_1 \alpha_0}{(1 - \beta_1 \alpha_1)} \quad \gamma_1 = \frac{\beta_1 \alpha_2}{(1 - \beta_1 \alpha_1)} \quad \gamma_2 = \frac{\beta_2}{(1 - \beta_1 \alpha_1)}$$

Also,

$$\phi_0 = \frac{\alpha_0 + \alpha_1 \beta_0}{(1 - \alpha_1 \beta_1)} \quad \phi_1 = \frac{\alpha_2}{(1 - \alpha_1 \beta_1)} \quad \phi_2 = \frac{\alpha_1 \beta_2}{(1 - \alpha_1 \beta_1)}$$

- In this case, we can solve backward from this for the original regression coefficients of interest, α_1 and β_1 . We can re-express the coefficients from the reduced form equations above to yield:

$$\alpha_1 = \frac{\phi_2}{\gamma_2} \quad \text{and} \quad \beta_1 = \frac{\gamma_1}{\phi_1}$$

- This is often labeled “Indirect Least Squares,” a method that is mostly of pedagogic interest.
- Thus, we uncover and estimate the true relationship between growth and violence by looking at the relationship between those two variables and the exogenous variables. We say that \mathbf{X} and \mathbf{Z} identify the model.
- Without the additional variables, there would be no way to estimate α_1 and β_1 from the reduced form. The model would be *under-identified*.
- It is also possible to have situations where there is more than one solution for α_1 and β_1 from the reduced form. This can occur if either equation has more than one additional variable.
- Since one way to estimate the coefficients of interest without bias is to estimate the reduced form and compute the original coefficients from this directly, it is important to determine whether your model is identified. We can estimate just-identified and over-identified models, but not under-identified models.

4.3 Identification

- There are two conditions to check for identification: the order condition and the rank condition. In theory, since the rank condition is more binding, one checks first the order condition and then the rank condition. In practice, very few people bother with the rank condition.

4.3.1 The Order Condition

- Let g be the number of endogenous variables in the system (here 2) and let k be the total number of variables (endogenous and exogenous) missing from the equation under consideration. Then:
 1. If $k = g - 1$, the equation is exactly identified.
 2. If $k > g - 1$, the equation is over-identified.
 3. If $k < g - 1$, the equation is under-identified.

- In general, this means that there must be at least one exogenous variable in the system, excluded from that equation, in order to estimate the coefficient on the endogenous variable that is included as an explanatory variable in that equation.
- These conditions are necessary for a given degree of identification. The Rank Condition is sufficient for each type of identification. The Rank Condition assumes the order condition and adds that the reduced form equations must each have full rank.

4.4 IV Estimation and Two-Stage Least Squares

1. If the model is just-identified, we could regress the endogenous variables on the exogenous variables and work back from the reduced form coefficients to estimate the structural parameters of interest.
 2. If the model is just or over-identified, we can use the exogenous variables as instruments for G_i and V_i . In this case, we use the instruments to form estimates of the two endogenous variables, \tilde{G}_i and \tilde{V}_i that are now uncorrelated with the error term and we use these in the original structural equations.
- The second method is generally easier. It is also known as Two-Stage Least Squares, or 2SLS because it involves the following stages:
 1. Estimate the reduced form equations using OLS. To do this, regress each endogenous variable on *all* the exogenous variables in the system. In our running example we would have:

$$V_i = \phi_0 + \phi_1 Z_i + \phi_2 X_i + \eta_i^*$$

and

$$G_i = \gamma_0 + \gamma_1 Z_i + \gamma_2 X_i + \varepsilon_i^*$$

2. From these first-stage regressions, estimate \tilde{G}_i and \tilde{V}_i for each observation. The predicted values of the endogenous variables can then be used to estimate the structural models:

$$\begin{aligned} V_i &= \alpha_0 + \alpha_1 \tilde{G}_i + \alpha_2 Z_i + \eta_i \\ G_i &= \beta_0 + \beta_1 \tilde{V}_i + \beta_2 X_i + \varepsilon_i \end{aligned}$$

- This is just what we would do in standard IV estimation, which is to regress the problem variable on its instruments and then use the predicted value in the main regression.

4.4.1 Some Important Observations

- This method gives you unbiased and consistent estimates of α_1 and β_1 . Another way of saying this is that \tilde{G}_i and \tilde{V}_i are good instruments for G_i and V_i .
 - A good instrument is highly correlated with the variable that we are instrumenting for and uncorrelated with the error term. In this case, \tilde{G}_i and \tilde{V}_i are highly correlated with G_i and V_i because we use all the information available to us in the exogenous variables to come up with an estimate.
 - Second, \tilde{G}_i and \tilde{V}_i are uncorrelated with the error terms η_i and ε_i because of the properties of OLS and the way we estimate the first stage. A property of OLS is that the estimated residuals are uncorrelated with the regressors and thus uncorrelated with the predicted values (\tilde{G}_i and \tilde{V}_i).
 - Thus, since the direct effect of growth on V_i , for example, in the first reduced form equation is part of the estimated residual, the predicted value \tilde{V}_i can be treated as exogenous to growth. We have broken the chain of connection that runs from one model to another. We can say that $E[\tilde{\mathbf{G}}'\boldsymbol{\eta}] = \mathbf{0}$ and $E[\tilde{\mathbf{V}}'\boldsymbol{\varepsilon}] = \mathbf{0}$.
- When the system is exactly identified, 2SLS will give you results that are identical to those you would obtain from estimated the reduced form equations and using those coefficients directly to estimate α_1 and β_1 .

- There is one case in which estimation of a system of simultaneous equations by OLS will not give you biased and inconsistent estimates. This is the case of recursive systems. The following is an example of a recursive system:

$$\begin{aligned}y_i &= \alpha_0 + \alpha_1 X_i + \varepsilon_i \\p_i &= \beta_0 + \beta_1 y_i + \beta_2 R_i + \eta_i \\q_i &= \delta_0 + \delta_1 y_i + \delta_2 p_i + \delta_3 S_i + \nu_i\end{aligned}$$

Here, all the errors are independent. In this system of recursive equations, substituting for the endogenous variables y_i and p_i will ultimately get you to the exogenous variable X_i , so we don't get the feedback loops and correlations between regressor and the error that we did in the earlier case.

4.5 Recapitulation of 2SLS and Computation of Goodness-of-Fit

Let us review the two-stage least square procedure:

1. We first estimated the reduced form equations for Violence and Growth using OLS.

$$V_i = \phi_0 + \phi_1 Z_i + \phi_2 X_i + \eta_i^*$$

and

$$G_i = \gamma_0 + \gamma_1 Z_i + \gamma_2 X_i + \varepsilon_i^*$$

We used the predicted value of G_i and V_i , or \tilde{G}_i and \tilde{V}_i , as instruments in the second stage.

2. We estimated the structural equations using the instruments:

$$V_i = \alpha_0 + \alpha_1 \tilde{G}_i + \alpha_2 Z_i + \eta_i$$

$$G_i = \beta_0 + \beta_1 \tilde{V}_i + \beta_2 X_i + \varepsilon_i$$

And get unbiased coefficients.

- We would normally compute R^2 as:

$$\frac{SSR}{SST} = \frac{\sum (x_i - \bar{x})^2 \hat{\beta}^2}{\sum (y_i - \bar{y})^2} \quad \text{or} \quad 1 - \frac{SSE}{SST} = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}}$$

- In the second stage, however, the estimated residuals are:

$$\begin{aligned}\hat{\eta}_i &= V_i - (\hat{\alpha}_0 + \hat{\alpha}_1 \tilde{G}_i + \hat{\alpha}_2 Z_i) \\ \hat{\varepsilon}_i &= G_i - (\hat{\beta}_0 + \hat{\beta}_1 \tilde{V}_i + \hat{\beta}_2 X_i)\end{aligned}$$

- If we use these residuals in our computation of R^2 we will get a statistic that tells us how well the model with the *instruments* fits the data. If we want an estimate of how well the original structural model fits the data, we should estimate the residuals using the true endogenous variables G_i and V_i . Thus, we use:

$$\begin{aligned}\hat{\eta}_i &= V_i - (\hat{\alpha}_0 + \hat{\alpha}_1 G_i + \hat{\alpha}_2 Z_i) \\ \hat{\varepsilon}_i &= G_i - (\hat{\beta}_0 + \hat{\beta}_1 V_i + \hat{\beta}_2 X_i)\end{aligned}$$

- This gives us an estimate of the fit of the structural model. There is one oddity in the calculated R^2 that may result. When we derived R^2 , we used the fact that the OLS normal equations estimated residuals such that $E[\mathbf{X}'\mathbf{e}] = 0$. This gave us the result that:

$$SST = SSR + SSE \text{ or } \sum (y_i - \bar{y})^2 = \sum (x_i - \bar{x})^2 \hat{\beta}^2 + \sum e_i^2$$

The variances in y can be completely partitioned between the variance from the model and the variance from the residuals. This is no longer the case when you re-estimate the errors using the real values of G_i and V_i rather than their instruments and you can get cases where $SSE > SST$. In this case, you can get negative values of R^2 in the second stage. This may be perfectly okay if the coefficients are of the right sign and the standard errors are small.

- It also matters in this case whether you estimate R^2 as $\frac{SSR}{SST}$ or $1 - \frac{SSE}{SST}$.

4.6 Computation of Standard Errors in 2SLS

- Let us denote all the original right-hand side variables in the two structural models as \mathbf{X} , the instrumented variables as $\tilde{\mathbf{X}}$, and the exogenous variables that we used to estimate $\tilde{\mathbf{X}}$ as \mathbf{Z} .

- In section 3.12, we showed that:

$$\begin{aligned}\hat{\beta}_{IV} &= (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{y} = (\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta} + (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}(\tilde{\mathbf{X}}'\boldsymbol{\varepsilon})\end{aligned}$$

$$\text{var}[\hat{\beta}_{IV}] = E[(\hat{\beta}_{IV} - \boldsymbol{\beta})(\hat{\beta}_{IV} - \boldsymbol{\beta})'] = E[(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}]$$

- If the errors conform to the Gauss-Markov assumptions then, $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \sigma^2\mathbf{I}_N$ and

$$\text{var}[\hat{\beta}_{IV}] = \sigma^2(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}$$

- To estimate σ^2 we would normally use $\frac{\mathbf{e}'\mathbf{e}}{(n-K)}$.
- As with R^2 , we should use the estimated residuals from the structural model with the true variables in it rather than the predicted values. These are consistent estimates of the true disturbances.

$$\hat{\eta}_i = V_i - (\hat{\alpha}_0 + \hat{\alpha}_1 G_i + \hat{\alpha}_2 Z_i)$$

$$\hat{\varepsilon}_i = G_i - (\hat{\beta}_0 + \hat{\beta}_1 V_i + \hat{\beta}_2 X_i)$$

- The IV estimator can have very large standard errors, because the instruments by which \mathbf{X} is proxied are not perfectly correlated with it and your residuals will be larger.

4.7 Three-Stage Least Squares

- What if, in the process above, we became concerned that $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] \neq \sigma^2\mathbf{I}_N$?
- We would perform three-stage least squares.
 1. Estimate the reduced form equations in OLS and calculate predicted values of the endogenous variables \tilde{G}_i and \tilde{V}_i .
 2. Estimate the structural equations with the fitted values.
 3. Use the residuals calculated in the manner above (using the actual values of G_i and V_i) to test for heteroskedasticity and/or autocorrelation and compute the appropriate standard errors if either are present. In the case of heteroskedasticity, this would mean either FGLS and a data transformation or White-Corrected Standard Errors. In the case of autocorrelation, this would mean either FGLS and a data transformation or Newey-West standard errors.

4.8 Different Methods to Detect and Test for Endogeneity

1. A priori tests for endogeneity – **Granger Causality**.
2. A test to look at whether the coefficients under OLS are markedly different from the coefficients under 2SLS – **Hausman Specification Test**.

4.8.1 Granger Causality

- A concept that is often used in time series work to define endogeneity is “Granger Causality.” What it defines is really “pre-determinedness.” Granger causality is absent when we can say that:

$$f(x_t|x_{t-1}, y_{t-1}) = f(x_t|x_{t-1})$$

The definition states that in the conditional distribution, lagged values of y_t add no information to our prediction of x_t beyond that provided by lagged values of x_t itself.

- This is tested by estimating:

$$x_t = \beta_0 + \beta_1 x_{t-1} + \beta_2 y_{t-1} + \varepsilon_t$$

- If a t-test indicates that $\beta_2 = 0$, then we say that y does not Granger cause x . If y does not Granger cause x , then x is often said to be exogenous in a system of equations with y . Here, exogeneity implies only that prior movements in y do not lead to later movements in x .
- Kennedy has a critique of Granger Causality in which he points out that under this definition weather reports “cause” the weather and that an increase in Christmas card sales “cause” Christmas. The problem here is that variables that are based on expectations (that the weather will be rainy, that Christmas will arrive) cause earlier changes in behavior (warnings to carry an umbrella and a desire to buy cards).

4.8.2 The Hausman Specification Test

- If $\hat{\beta}_{OLS}$ and $\hat{\beta}_{IV}$ are “close” in magnitude, then it would appear that endogeneity is not producing bias.
- This intuition has been formalized into a test by the econometrician Jerry Hausman. The test is called a specification test because it tells you whether you were right to use 2SLS. In this case, however, it can also be used to test the original assumption of endogeneity. The logic is as follows:
- H_0 : There is no endogeneity
In this case both $\hat{\beta}_{OLS}$ and $\hat{\beta}_{IV}$ are consistent and $\hat{\beta}_{OLS}$ is efficient relative to $\hat{\beta}_{IV}$ (recall that OLS is BLUE).
- H_1 : There is endogeneity
In this case $\hat{\beta}_{IV}$ remains consistent while $\hat{\beta}_{OLS}$ is inconsistent. Thus, their values will diverge.
- The suggestion, then, is to examine $\mathbf{d} = (\hat{\beta}_{IV} - \hat{\beta}_{OLS})$. The question is how large this difference should be before we assume that something is up. This will depend on the variance of \mathbf{d} .
- Thus, we can form a Wald statistic to test the hypothesis above:

$$W = \mathbf{d}'[\text{Estimated Asymptotic Variance}(\mathbf{d})]^{-1}\mathbf{d}$$

- The trouble, for a long time, was that no-one knew how to estimate this variance since it should involve the covariances between $\hat{\beta}_{IV}$ and $\hat{\beta}_{OLS}$. Hausman solved this by proving that the covariance between an efficient estimator $\hat{\beta}_{OLS}$ of a parameter vector β and its difference from an inefficient estimator $\hat{\beta}_{IV}$ of the same parameter vector is zero (under the null). (For more details see Greene Section 5.5.)
- Based on this proof, we can say that:

$$\text{Asy. var}[\hat{\beta}_{IV} - \hat{\beta}_{OLS}] = \text{Asy. var}[\hat{\beta}_{IV}] - \text{Asy. var}[\hat{\beta}_{OLS}]$$

- Under the null hypothesis, we are using two different but consistent estimators of σ^2 . If we use s^2 as a common estimator of this, the Hausman statistic will be:

$$H = \frac{(\hat{\beta}_{IV} - \hat{\beta}_{OLS})'[(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}]^{-1}(\hat{\beta}_{IV} - \hat{\beta}_{OLS})}{s^2}$$

- This test statistic is distributed χ^2 but the appropriate degrees of freedom for the test statistic will depend on the context (i.e., how many of the variables in the regression are thought to be endogenous).

4.8.3 Regression Version

- The test statistic above can be automatically computed in most standard software packages. In the case of IV estimation (of which 2SLS is an example) there is a completely equivalent way of running the Hausman test using an “auxiliary regression”.
- Assume that the model has K_1 potentially endogenous variables, \mathbf{X} , and K_2 remaining variables, \mathbf{W} . We have predicted values of \mathbf{X} , $\tilde{\mathbf{X}}$, based on the reduced form equations. We estimate the model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \tilde{\mathbf{X}}\boldsymbol{\alpha} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

- The test for endogeneity is performed as an F -test on the K_1 regression coefficients $\boldsymbol{\alpha}$ being different from zero, where the degrees of freedom are K_1 and $(n - (K_1 + K_2) - K_1)$. If $\boldsymbol{\alpha} = \mathbf{0}$, then \mathbf{X} is said to be exogenous.
- The intuition is as follows. If the \mathbf{X} variables are truly exogenous, then the $\boldsymbol{\beta}$ should be unbiased and there will be no extra information added by the fitted values. If the \mathbf{X} variables are endogenous, then the fitted values will add extra information and account for some of the variation in \mathbf{y} . Thus, they will have a coefficient on them significantly different from zero.
- Greene does not go through the algebra to prove that the augmented regression is equivalent to running the Hausman test but refers readers to Davidson and MacKinnon (1993). Kennedy has a nice exposition of the logic above on p. 197–198.

4.8.4 How to do this in Stata

1. 2SLS or To perform any type of IV estimation, see **ivreg**.

The command is:

```
ivreg (depvar) [varlist1] [varlist2=varlist IV]
```

2. To perform 3SLS, see **reg3**

3. To perform a Hausman test, see **hausman**

The hausman test is used in conjunction with other regression commands.

To use it, you would:

Run the less efficient model (here IV or reg3)

```
hausman, save
```

Run the fully efficient model (here OLS)

```
hausman
```

4.9 Testing for the Validity of Instruments

- We stated previously that the important features of “good” instruments, \mathbf{Z} , are that they be highly correlated with the endogenous variables, \mathbf{X} , and uncorrelated with the true errors, $\boldsymbol{\varepsilon}$.
- The first requirement can be tested fairly simply by inspecting the reduced form model in which each endogenous variables is regressed on its instruments to yield predicted values, $\tilde{\mathbf{X}}$. In a model with one instrument, look at the t -statistic. In an IV model with multiple instruments, look at the F -statistic.
- For the second requirement, the conclusion that \mathbf{Z} and $\boldsymbol{\varepsilon}$ are uncorrelated in the case of one instrument must be a leap of faith, since we cannot observe $\boldsymbol{\varepsilon}$ and must appeal to theory or introspection.

- For multiple instruments, however, so long as we are prepared to believe that one of the instruments is uncorrelated with the error, we can test the assumption that remaining instruments are uncorrelated with the error term. This is done via an auxiliary regression and is known as “testing over-identification restrictions.”
- It is called this because, if there is more than one instrument, then the endogenous regressor is over-identified. The logic of the test is simple.
 - If at least one of the instruments is uncorrelated with the true error term, then 2SLS gives consistent estimates of the true errors.
 - The residuals from the 2SLS estimation can then be used as the dependent variable in a regression on all the instruments.
 - If the instruments are, in fact, correlated with the true errors, then this will be apparent in a significant F -statistic on the instruments being jointly significant for the residuals.
- Thus, the steps in the test for over-identifying restrictions are as follows:
 1. Estimate the 2SLS residuals using $\hat{\eta}_i = V_i - \hat{\alpha}_0 + \hat{\alpha}_1 G_i + \hat{\alpha}_2 Z_i$. Use $\hat{\eta}_i$ as your consistent estimate of the true errors.
 2. Regress $\hat{\eta}_i$ on all the instruments used in estimating the 2SLS coefficients.
 3. You can approximately test the over-identifying restrictions via inspection of the F -statistic for the null hypothesis that all instruments are jointly insignificant. However, since the residuals are only consistent for the true errors, this test is only valid asymptotically and you should technically use a large-sample test.
 - An alternative is to use $n \cdot R^2 \stackrel{a}{\sim} \chi^2$ w/ df = # of instruments – # of endogenous regressors. If reject the null \Rightarrow at least some of the IVs are not exogenous.

Section 5

Time Series Modeling

5.1 Historical Background

- Genesis of modern time series models: large structural models of the macro-economy, which involved numerous different variables, were poor predictors of actual economic outcomes.
- Box and Jenkins showed that the present and future values of an economic variable could often be better predicted by its own past values than by other variables—*dynamic models*.
- In political science: used in models of presidential approval and partisan identity.

5.2 The Auto-Regressive and Moving Average Specifications

- A time series is a sequence of numerical data in which each observation is associated with a particular instant in time.
- Univariate time series analysis: analysis of a single sequence of data
- Multivariate time series analysis: several sets of data for the same sequence of time periods
- The purpose of time series analysis is to study the dynamics, or temporal structure of the data.
- Two representations of the temporal structure that will allow us to describe almost all dynamics (for “stationary” sequences) are the auto-regressive and moving-average representations.

5.2.1 An Autoregressive Process

- The AR(1) process:

$$y_t = \mu + \gamma y_{t-1} + \varepsilon_t$$

is said to be auto-regressive (or self-regressive) because the current value is explained by past values, so that:

$$E[y_t] = \mu + \gamma y_{t-1}$$

- This AR(1) process also contains a per-period innovation of μ , although this is often set to zero. This is sometimes referred to as a “drift” term. A more general, p th-order autoregression or AR(p) process would be written:

$$y_t = \mu + \gamma y_{t-1} + \gamma_2 y_{t-2} + \dots + \gamma_p y_{t-p} + \varepsilon_t$$

- In the case of an AR(1) process we can substitute infinitely for the y terms on the right-hand side (as we did previously) to show that:

$$y_t = \mu + \gamma\mu + \gamma^2\mu + \dots + \varepsilon_t + \gamma\varepsilon_{t-1} + \gamma^2\varepsilon_{t-2} + \dots + \gamma^\infty\varepsilon_{t-\infty} = \sum_{i=0}^{\infty} \gamma^i \mu + \sum_{i=0}^{\infty} \gamma^i \varepsilon_{t-i}$$

- So, one way to remember an auto-regressive process is that your current state is a function of all your previous errors.
- We can present this information far more simply using the **lag operator** or L :

$$Lx_t = x_{t-1} \quad \text{and} \quad L^2x_t = x_{t-2} \quad \text{and} \quad (1 - L)x_t = x_t - x_{t-1}$$

- Using the lag operator, we can write the original AR(1) series as:

$$y_t = \mu + \gamma Ly_t + \varepsilon_t$$

so that:

$$(1 - \gamma L)y_t = \mu + \varepsilon_t$$

and

$$y_t = \frac{\mu}{(1 - \gamma L)} + \frac{\varepsilon_t}{(1 - \gamma L)} = \frac{\mu}{(1 - \gamma)} + \frac{\varepsilon_t}{(1 - \gamma L)} = \sum_{i=0}^{\infty} \gamma^i \mu + \sum_{i=0}^{\infty} \gamma^i \varepsilon_{t-i}$$

- The last step comes from something we encountered before: how to represent an infinite series:

$$A = x(1 + a + a^2 + \dots + a^n)$$

If $|a| < 1$, then the solution to this series is approximately $A = \frac{x}{(1-a)}$. In other words, the sequence is convergent (has a finite solution).

- Thus, for the series $\sum_{i=0}^{\infty} \gamma^i \varepsilon_{t-i} = \varepsilon_t + \gamma \varepsilon_{t-1} + \gamma^2 \varepsilon_{t-2} + \dots = \varepsilon_t + \gamma L \varepsilon_t + \gamma^2 L^2 \varepsilon_t + \dots$ we have that, $a = \gamma L$ and $\sum_{i=0}^{\infty} \gamma^i \varepsilon_{t-i} = \frac{\varepsilon_t}{(1-\gamma L)}$.
- For similar reasons, $\sum_{i=0}^{\infty} \gamma^i \mu = \frac{\mu}{(1-\gamma L)} = \frac{\mu}{(1-\gamma)}$ because $L\mu = \mu$. In other words, μ , the per-period innovation is not subscripted by time and is assumed to be the same in each period.

5.3 Stationarity

- So, an AR(1) process can be written quite simply as:

$$y_t = \frac{\mu}{(1-\gamma)} + \frac{\varepsilon_t}{(1-\gamma L)}$$

- Recall, though, that this requires that $|\gamma| < 1$. If $|\gamma| \geq 1$ then we cannot even define y_t . y_t keeps on growing as the error terms collect. Its expectation will be undefined and its variance will be infinite.
- That means that we cannot use standard statistical procedures if the autoregressive process is characterized by $|\gamma| > 1$. This is known as the **stationarity condition** and the data series is said to be stationary if $|\gamma| < 1$.
- The problem is that our standard results for consistency and for hypothesis testing requires that $(\mathbf{X}'\mathbf{X})^{-1}$ is a finite, positive matrix. This is no longer true. The matrix $(\mathbf{X}'\mathbf{X})$ will be infinite when the data series \mathbf{X} is non-stationary.

- In the more general case of an $AR(p)$ model, the only difference is that the lag function by which we divide the right-hand side, $(1 - \gamma L)$, is more complex and is often written as $C(L)$. In this case, the stationarity condition relies on the roots to this more complex expression are all greater than one, or “lie outside the unit circle.”

5.4 A Moving Average Process

- A first-order, moving average process $MA(1)$ is written as:

$$y_t = \mu + \varepsilon_t - \theta\varepsilon_{t-1}$$

In this case, your current state depends on only the current and previous errors.

Using the lag operator:

$$y_t = \mu + (1 - \theta L)\varepsilon_t$$

Thus,

$$\frac{y_t}{(1 - \theta L)} = \frac{\mu}{(1 - \theta L)} + \varepsilon_t$$

Once again, if $|\theta| < 1$, then we can invert the series and express y_t as an infinite series of its own lagged values:

$$y_t = \frac{\mu}{(1 - \theta)} - \theta y_{t-1} - \theta^2 y_{t-2} - \dots + \varepsilon_t$$

Now we have written our MA process as an AR process of infinite lag length p , describing y_t in terms of all its own past values and the contemporaneous error term. Thus, an $MA(1)$ process can be written as an infinite $AR(p)$ process.

- Similarly, when we expressed the $AR(1)$ function in terms of all past errors terms, we were writing it as an infinite $MA(p)$ process.
- Notice, that this last step once again relies on the condition that $|\theta| < 1$. This is referred to in this case as the **invertibility** condition, implying that we can divide through by $(1 - \theta L)$.

- If we had a more general, $MA(q)$ process, with more lags, we could go through the same steps, but we would have a more complex function of the lags than $(1 - \theta L)$. Greene's textbook refers to this function as $D(L)$. In this case, the invertibility condition is satisfied when the roots of $D(L)$ lie outside the unit circle (see Greene, 5th Ed., pp. 611–615).

5.5 ARMA Processes

- Time series can also be posited to contain both AR and MA terms. However, if we go through the inversion above, getting:

$$y_t = \frac{\mu}{(1 - \theta)} - \theta y_{t-1} - \theta^2 y_{t-2} - \dots + \varepsilon_t$$

and then substitute for the lagged y_t s in the AR process, we will arrive at an expression for y_t that is based only on a constant and a complex function of past errors. See Greene p. 611 (Fifth Edition) for an example.

- We could also write an ARMA(1,1) process as:

$$y_t = \frac{\mu}{(1 - \gamma)} + \frac{(1 - \theta L)}{(1 - \gamma L)} \varepsilon_t$$

- An ARMA process with p autoregressive components and q moving average components is called an ARMA(p, q) process.
- Where does this get us? We can estimate y_t and apply the standard proofs of consistency if the time series is stationary, so it makes sense to discuss what stationarity means in the context of time series data. If the time series is stationary and can be characterized by an AR process, then the model can be estimated using OLS. If it is stationary and characterized by an MA process, you will need to use a more complicated estimation procedure (non-linear least squares).

5.6 More on Stationarity

- There are two main concepts of stationarity applied in the literature.
 1. **Strict Stationarity:** For the process $y_t = \rho y_{t-1} + \varepsilon_t$, strict stationarity implies that:
 - $E[y_t] = \mu$ exists and is independent of t .
 - $\text{var}[y_t] = \gamma_0$ is a finite, positive constant, independent of t .
 - $\text{cov}[y_t, y_s] = \gamma(|t - s|)$ is a finite function of $|t - s|$, but not of t or s .
 - AND all other “higher order moments” (such as skewness or kurtosis) are also independent of t .
 2. **Weak Stationarity** (or covariance stationarity): removes the condition on the “higher order moments” of y_t .
- A stationary time series will tend to revert to its mean (*mean reversion*) and fluctuations around this mean will have a broadly consistent amplitude.
- Intuition: if we take two slices from the data series, they should have approximately the same mean and the covariance between points should depend only on the number of time periods that divide them. This will not be true of “integrated” series, so we will have to transform the data to make it stationary before estimating the model.

5.7 Integrated Processes, Spurious Correlations, and Testing for Unit Roots

- One of the main concerns w/ non-stationary series is *spurious correlation*.
- Suppose you have a non-stationary, highly trending series, y_t , and you regress it on another highly trending series, x_t . You are likely to find a significant relationship between y_t and x_t even when there was none, because we see upward movement in both produced in their own dynamics.
- Thus, when the two time series are non-stationary, standard critical values of the t and F statistics are likely to be highly misleading about true causal relationships (see Greene p. 778).

- The question is, what kind of non-stationary sequence do we have and how can we tell it's non-stationary. Consider the following types of non-stationary series:

1. The Pure Random Walk

$$y_t = y_{t-1} + \varepsilon_t$$

This DGP can also be written as:

$$y_t = y_0 + \sum \varepsilon_t$$

$$E[y_t] = E[y_0 + \sum \varepsilon_t] = y_0$$

- In similar fashion it can be shown that the variance of $y_t = t\sigma^2$. Thus, the mean is constant but the variance increases indefinitely as the number of time points grows.
- If you take the first difference of the data process, however, $\Delta y_t = y_t - y_{t-1}$, we get:

$$y_t - y_{t-1} = \varepsilon_t$$

The mean of this process is constant (and equal to zero) and its variance is also a finite constant. Thus, the first difference of a random walk process is a *difference stationary process*.

2. The Random Walk with Drift

$$y_t = \mu + y_{t-1} + \varepsilon_t$$

For the random walk with drift process, we can show that $E[y_t] = y_0 + t\mu$ and $\text{var}[y_t] = t\sigma^2$. Both the mean and the variance are non-constant.

- In this case, first differencing of the series also will give you a variable that has a constant mean and variance.

3. The Trend Stationary Process

$$y_t = \mu + \beta t + \varepsilon_t$$

- y_t is non-stationary because the mean of y_t is equal to $\mu + \beta t$, which is non-constant, although its variance is constant and equal to σ^2 .
 - Once the values of μ and β are known, however, the mean can be perfectly predicted. Therefore, if we subtract the mean of y_t from y_t , the resulting series will be stationary, and is thus called a *trend stationary process* in comparison to the *difference stationary* processes described above.
- Each of these series is characterized by a *unit root*, meaning that the coefficient on the lagged value of $y_t = 1$ in each process. For a trend stationary process, this follows because you can re-write the time series as $y_t = \mu + \beta t + \varepsilon_t = y_{t-1} + \beta + \varepsilon_t - \varepsilon_{t-1}$.
 - In each case, the DGP can be written as:

$$(1 - L)y_t = \alpha + \nu$$

where $\alpha = 0, \mu,$ and β respectively in each process and ν is a stationary process.

- In all cases, the data should be detrended or differenced to produce a stationary series. But which? The matter is not of merely academic interest, since detrending a random walk will induce autocorrelation in the error terms of an MA(1) type.

- A unit-root test (described in Greene, p. 637) is based on a model that nests the different processes above into one regression that you run to test the properties of the underlying data series:

$$y_t = \mu + \beta t + \gamma y_{t-1} + \varepsilon_t$$

- Next subtract y_{t-1} from both sides of the equation to produce the equation below. This produces a regression with a (difference) stationary dependent variable (even under the null of non-stationarity) and this regression forms the basis for Dickey-Fuller tests of a unit root:

$$y_t - y_{t-1} = \mu + \beta t + (\gamma - 1)y_{t-1} + \varepsilon_t$$

- A test of the hypothesis that $(\gamma - 1)$ is zero gives evidence for a random walk, because this $\Rightarrow \gamma = 1$.
 - If $(\gamma - 1) = 0$ and μ is significantly different from zero we have evidence for a random walk with drift.
 - If $(\gamma - 1)$ is significantly different from zero (and < 0) we have evidence of a stationary process.
 - If $(\gamma - 1)$ is significantly different from and less than zero, and β (the coefficient on the trend variable) is significant, we have evidence for a trend stationary process.
- There is one complication. Two statisticians, Dickey and Fuller (1979, 1981) showed that if the unit root is exactly equal to one, the standard errors will be under-estimated, so that revised critical values are required for the test statistic above.
 - For this reason, the test for stationarity is referred to as the Dickey-Fuller test. The augmented Dickey-Fuller test applies to the same equation above but adds lags of the first difference in y , $(y_t - y_{t-1})$.
 - One problem with the Dickey-Fuller unit-root test is that it has low power and seems to privilege the null hypothesis of a random walk process over the alternatives.

- To sum up: If your data looks like a random walk, you will have to difference it until you get something that looks stationary. If your data looks like it's trend stationary, you will have to de-trend it until you get something stationary. An ARMA model carried out on differenced data is called an ARIMA model, standing for "Auto-Regressive, Integrated, Moving-Average."

5.7.1 Determining the Specification

- You have data that is now stationary. How do you figure out which ARMA specification to use? How many, if any, AR terms should there be? How many, if any, MA terms?
- As a means of deciding the specification, analysts in the past have looked at the autocorrelations and partial autocorrelations between y_t and y_{t-s} . This is also known as looking at the autocorrelation function (ACF) and the partial autocorrelation function (or PACF).
- Recall that

$$\text{corr}(\varepsilon_t, \varepsilon_{t-s}) = \frac{\text{cov}(\varepsilon_t, \varepsilon_{t-s})}{\sqrt{\text{var}(\varepsilon_t)}\sqrt{\text{var}(\varepsilon_{t-s})}} = \frac{E[\varepsilon_t \varepsilon_{t-s}]}{E[\varepsilon_t^2]} = \frac{\gamma_s}{\gamma_0}$$

- If y_t and y_{t-s} are both expressed in terms of deviations from their means, and if $\text{var}(y_t) = \text{var}(y_{t-s})$ then:

$$\text{corr}(y_t, y_{t-s}) = \frac{\text{cov}(y_t, y_{t-s})}{\sqrt{\text{var}(y_t)}\sqrt{\text{var}(y_{t-s})}} = \frac{E[y_t y_{t-s}]}{E[y_t^2]} = \frac{\gamma_s}{\gamma_0}$$

5.8 The Autocorrelation Function for AR(1) and MA(1) processes

- We showed in the section on autocorrelation in the error terms that if $y_t = \rho y_{t-1} + \varepsilon_t$ then $\text{corr}[y_t, y_{t-s}] = \rho^s$.
- In this context, ε_t is *white noise*. $E[\varepsilon_t] = 0$, $E[\varepsilon_t^2] = \sigma_\varepsilon^2$, $E[\varepsilon_t \varepsilon_{t-1}] = 0$
- Thus, the autocorrelations for an AR(1) process tend to die away gradually.

- By contrast, the autocorrelations for an MA(1) process die away abruptly.
- Let

$$y_t = \varepsilon_t - \theta\varepsilon_{t-1}$$

Then

$$\gamma_0 = \text{var}[y_t] = E\{y_t - E[y_t]\}^2 = E[(\varepsilon_t - \theta\varepsilon_{t-1})^2] = E(\varepsilon_t^2) + \theta^2 E(\varepsilon_{t-1}^2) = (1 + \theta^2)\sigma_\varepsilon^2$$

and

$$\gamma_1 = \text{cov}[y_t, y_{t-1}] = E[(\varepsilon_t - \theta\varepsilon_{t-1})(\varepsilon_{t-1} - \theta\varepsilon_{t-2})] = -\theta E[\varepsilon_{t-1}^2] = -\theta\sigma_\varepsilon^2$$

- The covariances between y_t and y_{t-s} when $s > 1$ are zero, because the expression for y_t only involves two error terms. Thus, the ACF for an MA(1) process has one or two spikes and then shows no autocorrelation.

5.9 The Partial Autocorrelation Function for AR(1) and MA(1) processes

- The partial autocorrelation is the simple correlation between y_t and y_{t-s} minus that part explained by the intervening lags.
- Thus, the partial autocorrelation between y_t and y_{t-s} is estimated by the last coefficient in the regression of y_t on $[y_{t-1}, y_{t-2}, \dots, y_{t-s}]$. The appearance of the partial autocorrelation function is the reverse of that for the autocorrelation function. For a true AR(1) process,

$$y_t = \rho y_{t-1} + \varepsilon_t$$

- There will be an initial spike at the first lag (where the autocorrelation equals ρ) and then nothing, because no other lagged value of y is significant.
- For the MA(1) process, the partial autocorrelation function will look like a gradually declining wave, because any MA(1) process can be written as an infinite AR process with declining weights on the lagged values of y .

5.10 Different Specifications for Time Series Analysis

- We now turn to models in which y_t is related to its own past values and a set of exogenous, explanatory variables, \mathbf{x}_t .
- The pathbreaking time series work in pol. sci. concerned how presidential approval levels respond to past levels of approval and measures of presidential performance. Let's use A_t , standing for the current level of approval, as the dependent variable for our examples here.
 - More on this topic can be found in an article by Neal Beck, 1991, “Comparing Dynamic Specifications,” *Political Analysis*.

1. The Static Model

$$A_t = \beta X_t + \varepsilon_t$$

Comment from Beck, “This assumes that approval adjusts instantaneously to new information (“no stickiness”) and that prior information is of no consequence (“no memory”).

2. The Finite Distributed Lag Model

$$A_t = \sum_{i=0}^M \beta_i X_{t-i} + \varepsilon_t$$

This allows for memory but may have a large number of coefficients, reducing your degrees of freedom.

3. The Exponential Distributed Lag (EDL) Model

$$A_t = \sum_{i=0}^M (X_{t-i} \lambda^i) \beta + \varepsilon_t$$

Reduces the number of parameters that you now have to estimate. In the case in which T can be taken as infinite, this can also be written as:

$$A_t = \sum_{i=0}^{\infty} X_t \beta (\lambda L)^i + \varepsilon_t = \frac{X_t \beta}{(1 - \lambda L)} + \varepsilon_t$$

4. The Partial Adjustment Model

Multiply both sides of the EDL model above by the “Koyck Transformation” or $(1 - \lambda L)$. After simplification, this will yield:

$$A_t = X_t\beta + \lambda A_{t-1} + \varepsilon_t - \lambda\varepsilon_{t-1}$$

- This says that current approval is a function of current exogenous variables and the past value of approval, allowing for memory and stickiness. By including a lagged dependent variable, you are actually allowing for all past values of X to affect your current level of approval, with more recent values of X weighted more heavily.
- If the errors in the original EDL model were also AR(1), that is to say if errors in preceding periods also had an effect on current approval (e.g., $u_t = \lambda u_{t-1} + \varepsilon_t$), with the size of that effect falling in an exponential way, then the Koyck transformation will actually give you a specification in which the errors are iid. In other words:

$$A_t = X_t\beta + \lambda A_{t-1} + \varepsilon_t$$

This specification is very often used in applied work with a stationary variable whose current level is affected by memory and stickiness. If the error is indeed iid, then the model can be estimated using OLS.

5. Models for “Difference Stationary” Data: The Error Correction Model

Often, if you run an augmented Dickey-Fuller test and are unable to reject the hypothesis that your data has a unit root, you will wind up running the following type of regression:

$$\Delta A_t = \Delta X_t\beta + \varepsilon_t$$

As Beck notes (p. 67) this is equivalent to saying that only the information in the current period counts and that it creates an instantaneous change in approval. This can be restrictive.

- Moreover, it will frequently result in finding no significant results in first differences, although you have strong theoretical priors that the dependent variable is related to the independent variables.

- What you can do in this instance, with some justification, is to run an *Error Correction Model* that allows you to estimate both long-term and short-term dynamics.
- Let us assume, for the moment, that both A and X are integrated (of order one). Then we would not be advised to include them in a model in terms of levels. However, if they are in a long-term, equilibrium relationship with one another, then the errors

$$e_t = A_t - X_t\alpha$$

should be stationary. Moreover, if we can posit that people respond to their “errors” and that gradually the relationship comes back into equilibrium, then we can introduce the error term into a model of the change in approval. This is the Error Correction Model.

$$\Delta A_t = \Delta X_t\beta + \gamma(A_{t-1} - X_{t-1}\alpha) + \varepsilon_t$$

The γ coefficient in this model is telling you how fast errors are adjusted. An attraction of this model is that it allows you to estimate long-term dynamics (in levels) and short-term dynamics (in changes) simultaneously.

5.11 Determining the Number of Lags

- In order to select the appropriate number of lags of the dependent variable (in an AR(p)) model, you could use the “general to specific” methodology. Include a number of lags that you think is more than sufficient and take out of the model any lags that are not significant.
- Second, analysts are increasingly using adjusted measures of fit (analogous to R^2) that compensate for the fact that as you include more lags (to fit the data better) you are reducing the degrees of freedom.
- The two best known are the Akaike Information Criterion and the Schwartz Criterion. Both are based on the standard error of the estimate (actually on s^2). Thus, you want to *minimize* both, but you are penalized when you include additional lags to reduce the standard error of the estimate.

- The equations for the two criteria are:

$$AIC(K) = \ln \left(\frac{\mathbf{e}'\mathbf{e}}{n} \right) + \frac{2K}{n}$$

$$SC(K) = \ln \left(\frac{\mathbf{e}'\mathbf{e}}{n} \right) + \frac{K \ln n}{n}$$

The Schwartz Criterion penalizes degrees of freedom lost more heavily.

- The model with the smallest AIC or SC indicates the appropriate number of lags.

5.12 Determining the Correct Specification for your Errors

- Beck gives examples (p. 60 and p. 64) where the error term is autocorrelated even when you have built in lags of the dependent variable.
- This could occur if the effect of past errors dies out at a different rate than does the effect of past levels in the explanatory variables. If this is the case, you could still get an error term that is AR or MA, even when you have introduced a lagged dependent variable, although doing this will generally *reduce* the autocorrelation.
- The autocorrelation could be produced by an AR or an MA process in the error term. To determine which, you might again look at the autocorrelations and partial autocorrelations.
- We know how to transform the data in the case where the errors are AR(1) using the Cochrane-Orcutt transformation or using Newey-West standard errors. If the model was otherwise static, I would also suggest these approaches.

5.13 Stata Commands

- Dickey-Fuller Tests:
dfuller varname, **noconstant** **lags**(#) **trend** **regress**
- The Autocorrelation and Partial Autocorrelation Function:
corrgram varname, **lags**(#)
ac varname, **lags**(#)
pac varname, **lags**(#)
- ARIMA Estimation:
arima depvar [varlist], **ar**(numlist) **ma**(numlist)

Part II

Maximum Likelihood Estimation

Section 6

Intro to Maximum Likelihood

Our basic task: make inferences about the relationship between variables in a population based on what's going on in a sample

Fundamentals:

- Assume distribution for data—i.e., all observations are drawn from the same distribution.
- Given distribution we can write down a function that describes how the data for an individual were generated.
- Key assumption: observations in sample are independent and identically distributed (iid)—we get independence by assuming we have a random sample.

– If observations are independent, then the joint density of the observations is just the product of their marginal densities:

$$f(x_1, x_2, \dots, x_{n-1}, x_n) = f(x_1) \cdot f(x_2) \cdots f(x_{n-1}) \cdot f(x_n)$$

- This product gives the likelihood of the sample.
- ML means we choose the parameters that maximize the likelihood of observing the sample that we actually observe.
- This requires that we find the maximum of the likelihood function w.r.t. to the parameters of interest.

To fix ideas consider the following example:

- Flipping coins (Bernoulli trials):

$$Y = \begin{cases} 1 & \text{for a head} \\ 0 & \text{for a tail} \end{cases}$$

- Let $p = \Pr(Y = 1)$
- Flip coin n times to obtain the sample: Y_1, \dots, Y_n .
- Key assumption: coin flips are independent and prob of observing a head or a tail is the same at each trial.
- From the given sample, we want to estimate p .
- Note that p is population mean:

$$\begin{aligned} E(Y_i) &= \Pr(Y_i = 1) \cdot 1 + \Pr(Y_i = 0) \cdot 0 \\ &= \Pr(Y_i = 1) \\ &= p \end{aligned}$$

- To estimate the population mean we typically use the sample mean:

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}.$$

- Note that the proportion of heads is given by $\hat{p} = \frac{\sum_{i=1}^n Y_i}{n}$.

- Let $L_i =$ likelihood for observation i .
- What is likelihood for i th toss?
- For each observation could get a head which happens with prob p or could get a tail which happens with prob $1 - p$:

$$L_i = p^{Y_i} \cdot (1 - p)^{(1 - Y_i)} \quad (6.1)$$

- If we have iid sampling then

$$\begin{aligned} L &= \left[p^{Y_1} \cdot (1 - p)^{(1 - Y_1)} \right] \left[p^{Y_2} \cdot (1 - p)^{(1 - Y_2)} \right] \cdots \left[p^{Y_n} \cdot (1 - p)^{(1 - Y_n)} \right] \\ &= \prod_{i=1}^n p^{Y_i} \cdot (1 - p)^{(1 - Y_i)}. \end{aligned}$$

- Usually take logs because they are easier to deal with (Remember: $\ln a^b = b \ln a$ and $\ln(a \cdot b) = \ln a + \ln b$):

$$\begin{aligned} \ln L &= \sum_{i=1}^n \ln L_i \\ &= \sum_{i=1}^n \ln \left[p^{Y_i} (1 - p)^{(1 - Y_i)} \right] \\ &= \sum_{i=1}^n [Y_i \ln p + (1 - Y_i) \ln(1 - p)] \end{aligned}$$

- Want to choose value for p than maximizes $\ln L$. How do we do this?
- Some calculus:
 - To find the maximum of a function with respect to some parameter, we take the derivative of the function with respect to that parameter and then set result equal to zero and solve for the parameter.
 - To make sure it's a maximum we then check to make sure the second derivative is negative.

- For example, consider the function $f(x) = -x^2 + 2x$.

$$\frac{df(x)}{dx} = -2x + 2$$

- Setting this equal to zero and solving for x gives $x = 1$.
- Checking the second derivative:

$$\begin{aligned}\frac{d^2 f(x)}{d^2 x} &= \frac{d(-2x + 2)}{dx} \\ &= -2\end{aligned}$$

which implies we have a maximum rather than a minimum.

- Returning to our likelihood function:

$$\frac{d \ln L}{dp} = \sum_{i=1}^n \left[Y_i \frac{1}{p} - (1 - Y_i) \frac{1}{1-p} \right]$$

since

$$\frac{d \ln a}{da} = \frac{1}{a}$$

(Note: must also use the chain rule).

- Setting the derivative equal to zero and solving gives:

$$\sum_{i=1}^n \left[Y_i \frac{1}{p} - (1 - Y_i) \frac{1}{1-p} \right] = 0$$

$$\frac{\sum_{i=1}^n [Y_i(1-p) - (1-Y_i)p]}{p(1-p)} = 0$$

$$\sum_{i=1}^n [Y_i - pY_i - p + pY_i] = 0$$

$$np = \sum_{i=1}^n Y_i$$

$$p = \frac{\sum_{i=1}^n Y_i}{n}$$

- Checking the second derivative of $\ln L$ w.r.t. p we find that it is always negative:

$$\frac{d^2 \ln L}{d^2 p} = - \sum_{i=1}^n Y_i \frac{1}{p^2} - \sum_{i=1}^n (1 - Y_i) \frac{1}{(1-p)^2}$$

so $\hat{p} = \frac{\sum_{i=1}^n Y_i}{n}$ yields a global max. of the log likelihood function.

Section 7

Maximum Likelihood In Depth

- Suppose we have random sample of n observations where each observations has the density $f(x_i, \theta)$. Then the joint pdf is given by

$$\begin{aligned} f(x_1, \dots, x_n, \theta) &= f(x_1, \theta) \cdots f(x_n, \theta) \\ &= \prod_{i=1}^n f(x_i, \theta) \\ &= L(\theta) \end{aligned} \tag{7.1}$$

- We usually deal with the log of the likelihood function:

$$\ln L(\theta) = \sum_{i=1}^n \ln f(x_i, \theta)$$

- The MLEs of θ , denoted $\hat{\theta}$, are the values of the parameters that maximize $L(\theta)$.
- Note that $\hat{\theta}$ also maximizes $\ln L(\theta)$ because \ln is a monotonic function.
- The necessary condition for maximizing $\ln L(\theta)$ is

$$\frac{\partial \ln L(\theta)}{\partial \theta} = 0$$

- Suppose we have a sample from a normal distribution with mean μ and variance σ^2 . Then the likelihood function is

$$\begin{aligned}
 L(\mu, \sigma^2) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2}\right) \\
 &= \prod_{i=1}^n (2\pi\sigma^2)^{-1/2} e^{-(1/(2\sigma^2))(x_i - \mu)^2} \\
 &= (2\pi\sigma^2)^{-n/2} e^{-(1/(2\sigma^2)) \sum_{i=1}^n (x_i - \mu)^2}
 \end{aligned} \tag{7.2}$$

- Taking the log gives

$$\ln L(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

- Taking first derivatives gives

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2} \sum_i (x_i - \mu) \tag{7.3}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_i (x_i - \mu)^2 \tag{7.4}$$

- Setting (7.3) equal to zero and solving for μ , gives the MLE $\hat{\mu}$:

$$\hat{\mu} = \frac{1}{n} \sum_i x_i = \bar{x}$$

- To find the variance of the MLE, we need to take the second derivative w.r.t. the parameter of interest.
 - The variance-covariance matrix is a function of the **Hessian**—the matrix of second derivatives.
 - Specifically it is the inverse of the negative of the expectation of the second derivatives.
- From our example,

$$\begin{aligned}\text{var}(\hat{\mu}) &= \left[-E \left(\frac{\partial^2 \ln L}{\partial \mu^2} \right) \right]^{-1} \\ &= \frac{\sigma^2}{n}\end{aligned}$$

- Setting (7.4) equal to zero, inserting $\hat{\mu}$ for μ and solving for σ^2 gives the MLE:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2$$

- Note that the unbiased estimator for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$$

- The attractiveness of MLEs stems from their large sample or asymptotic properties.

7.1 Asymptotic Properties of MLEs

1. The MLE is consistent.

That is,

$$\lim_{n \rightarrow \infty} \Pr(|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}| > \varepsilon) = \mathbf{0}$$

where ε is some small positive number. Or simply

$$\text{plim } \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$$

2. The MLE is asymptotically normally distributed:

$$\hat{\boldsymbol{\theta}} \stackrel{a}{\sim} N[\boldsymbol{\theta}, \{\mathbf{I}(\boldsymbol{\theta})\}^{-1}]$$

where $\mathbf{I}(\boldsymbol{\theta})$ is the **information matrix**. The information matrix contains the negatives of the expected values of the second derivatives.

$$\begin{aligned} [\mathbf{I}(\boldsymbol{\theta})]^{-1} &= \left\{ -E \left[\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right] \right\}^{-1} \\ &= \left\{ E \left[\left(\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right) \right] \right\}^{-1} \end{aligned}$$

3. The MLE is asymptotically efficient and achieves the **Cramér-Rao lower bound** for consistent estimators:

$$\text{Asy. var}[\hat{\boldsymbol{\theta}}] = [\mathbf{I}(\boldsymbol{\theta})]^{-1}$$

- The **Cramér-Rao inequality** states that, given certain regularity conditions, if $\hat{\theta}$ is an unbiased estimator of θ then

$$\text{var}(\hat{\theta}) \geq \frac{1}{-E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]}$$

- The Cramér-Rao lower bound gives the lower bound for the variance of unbiased estimators.
- Since the MLE achieves this lower bound, we know that the MLE has the minimum variance achievable by a consistent estimator.

- Example: Consider $L(\mu, \sigma^2)$, the likelihood function in equation 7.2. Taking the second derivatives gives

$$\begin{aligned}\frac{\partial^2 \ln L}{\partial \mu^2} &= -\frac{n}{\sigma^2} \\ \frac{\partial^2 \ln L}{\partial (\sigma^2)^2} &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_i (x_i - \mu)^2 \\ \frac{\partial^2 \ln L}{\partial \mu \partial \sigma^2} &= -\frac{1}{\sigma^4} \sum_i (x_i - \mu)\end{aligned}$$

so the information matrix is

$$\mathbf{I}(\mu, \sigma^2) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

- Another important property of the MLE is **invariance**: if $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$ and if $g(\boldsymbol{\theta})$ is a continuous function, then the MLE of $g(\boldsymbol{\theta})$ is $g(\hat{\boldsymbol{\theta}})$.
- Note that the $\text{Asy. var}[\hat{\boldsymbol{\theta}}]$ is a function of the unknown parameters $\boldsymbol{\theta}$, so we have to come up with an estimator for it. Two consistent estimators are

$$\widehat{\text{Asy. var}}[\hat{\boldsymbol{\theta}}] = \left\{ - \left[\frac{\partial^2 \ln L(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}} \partial \hat{\boldsymbol{\theta}}'} \right] \right\}^{-1}$$

or

$$\widehat{\text{Asy. var}}[\hat{\boldsymbol{\theta}}] = \left\{ \left[\left(\frac{\partial \ln L(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}} \right) \left(\frac{\partial \ln L(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}'} \right) \right] \right\}^{-1}$$

7.2 Iterative Process of Finding the MLE

The models we will be dealing with here typically do not have closed-form solutions for the MLE. It becomes necessary to use some type of iterative “hill-climbing” algorithm, such as the following:

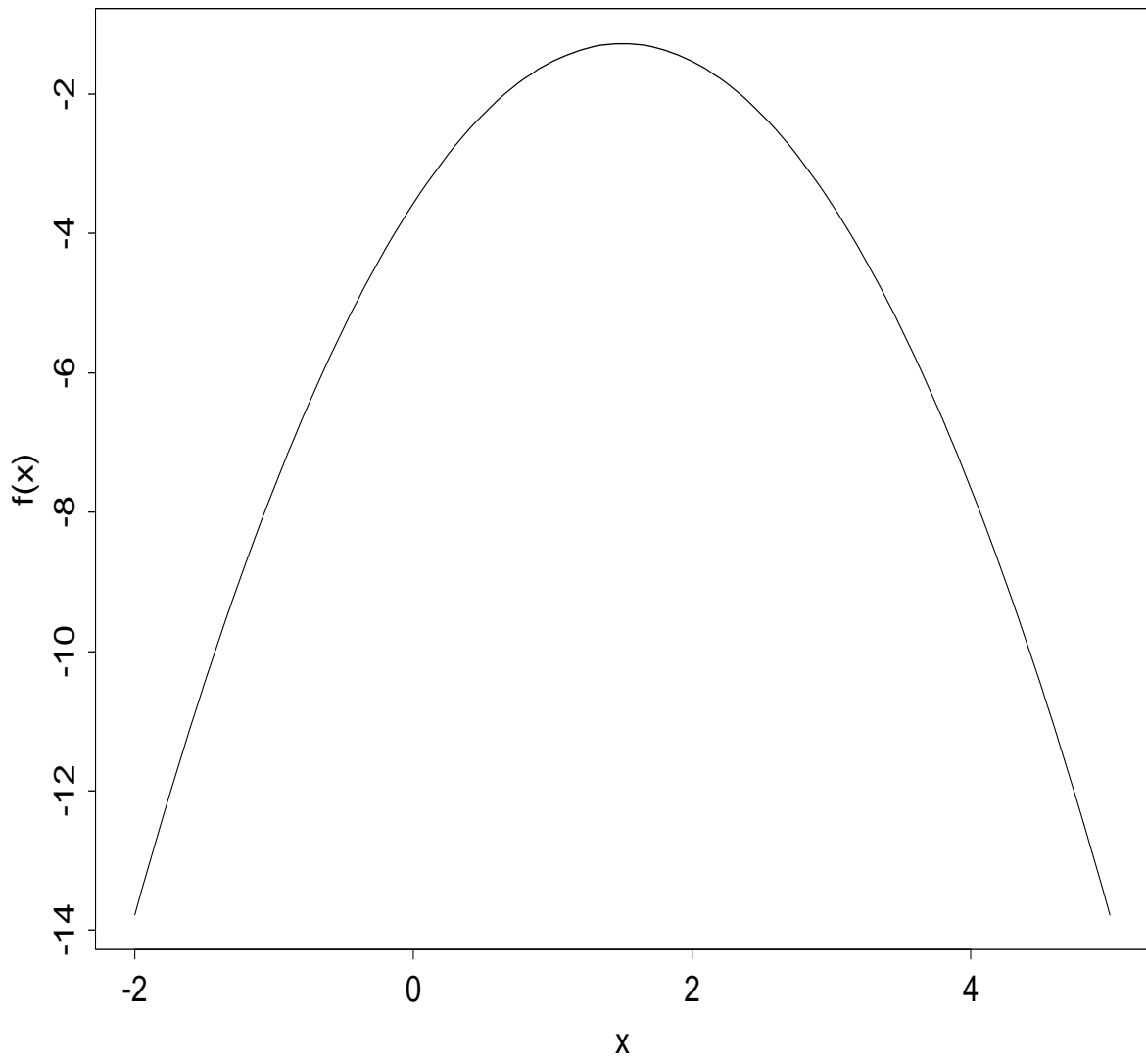
1. Compute starting values $\boldsymbol{\theta}_0$.
2. Compute

$$\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 + s_0[\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}\mathbf{S}(\boldsymbol{\theta}_0) \quad (7.5)$$

where s_0 is the **step size**, $\mathbf{S}(\boldsymbol{\theta}_0)$ is the **score vector** (i.e., vector of first derivatives of $\ln L$), $\mathbf{I}(\boldsymbol{\theta}_0)$ is the information matrix, and $\boldsymbol{\theta}_1$ is the updated vector of parameter values.

3. Repeat Step 2 until convergence.

Figure 7.1: The Natural Log of the Normal Distribution



Part III

Models for Repeated Observations Data—Continuous Dependent Variables

Section 8

Fixed Effects Estimators

8.1 LSDV as Fixed Effects

- Least squares dummy variable estimation is also known as **fixed effects**, because it assumes that the unobserved effect for a given cross-sectional unit or time period can be estimated as a given, *fixed* effect.
- Can think of this as fixed in repeated samples (e.g., France is France) as opposed to randomly drawn.
- Let the original model be

$$y_{it} = \alpha_i^* + \beta' \mathbf{x}_{it} + u_{it} \quad (8.1)$$

- We can rewrite this in vector form as

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_N \end{bmatrix} = \begin{bmatrix} \mathbf{e} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} \alpha_1^* + \begin{bmatrix} \mathbf{0} \\ \mathbf{e} \\ \vdots \\ \mathbf{0} \end{bmatrix} \alpha_2^* + \dots + \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{e} \end{bmatrix} \alpha_N^* + \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_N \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_N \end{bmatrix} \quad (8.2)$$

where

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{bmatrix}, \quad \mathbf{X}_i = \begin{bmatrix} x_{1i1} & x_{2i1} & \dots & x_{Ki1} \\ x_{1i2} & x_{2i2} & \dots & x_{Ki2} \\ \vdots & \vdots & & \vdots \\ x_{1iT} & x_{2iT} & \dots & x_{KiT} \end{bmatrix}, \quad \mathbf{e} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix},$$

$$\mathbf{u}_i = \begin{bmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{iT} \end{bmatrix},$$

$$E[\mathbf{u}_i] = \mathbf{0}, \quad E[\mathbf{u}_i \mathbf{u}_i'] = \sigma_u^2 \mathbf{I}_T, \quad E[\mathbf{u}_i \mathbf{u}_j'] = \mathbf{0} \text{ if } i \neq j.$$

(\mathbf{I}_T denotes the $T \times T$ identity matrix).

- These assumptions regarding u_{it} mean that the OLS estimator for eq. 8.2 is BLUE.

- To obtain the OLS estimators of α_i^* and $\boldsymbol{\beta}$, we minimize:

$$S = \sum_{i=1}^N \mathbf{u}_i' \mathbf{u}_i = \sum_{i=1}^N (\mathbf{y}_i - \mathbf{e} \alpha_i^* - \mathbf{X}_i \boldsymbol{\beta})' (\mathbf{y}_i - \mathbf{e} \alpha_i^* - \mathbf{X}_i \boldsymbol{\beta}).$$

- Take partial derivatives wrt to α_i^* , set equal to zero and solve to get:

$$\hat{\alpha}_i^* = \bar{y}_i - \boldsymbol{\beta}' \bar{\mathbf{x}}_i \quad (8.3)$$

where

$$\bar{y}_i = \sum_{t=1}^T y_{it}/T, \quad \bar{\mathbf{x}}_i = \sum_{t=1}^T \mathbf{x}_{it}/T, \quad .$$

- Substitute our estimate for $\hat{\alpha}_i^*$ in S , take partial derivatives wrt $\boldsymbol{\beta}$, set equal to zero and solve:

$$\hat{\boldsymbol{\beta}}_{\text{CV}} = \left[\sum_{i=1}^N \sum_{t=1}^T (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' \right]^{-1} \left[\sum_{i=1}^N \sum_{t=1}^T (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(y_{it} - \bar{y}_i) \right]$$

- Including separate dummies for each cross-sectional unit will produce estimates of the unit-specific effects.
- While this may be desirable, it does come at some cost—possibly inverting a large matrix of 0s and 1s.

- Another way to compute this estimator w/o including dummies is to subtract off the time means:

$$\bar{y}_i = \alpha_i^* + \boldsymbol{\beta}'\bar{\mathbf{x}}_i + \bar{u}_i \quad (8.4)$$

- If we run OLS on this regression, it will produce what is known as the “Between Effects” estimator, or $\boldsymbol{\beta}_{BE}$, which shows how the mean level of the dependent variable for each cross-sectional unit varies with the mean level of the independent variables.
- Subtracting eq. 8.4 from eq. 8.1 gives

$$(y_{it} - \bar{y}_i) = (\alpha_i^* - \alpha_i^*) + \boldsymbol{\beta}'(\mathbf{x}_{it} - \bar{\mathbf{x}}_i) + (u_{it} - \bar{u}_i)$$

or

$$(y_{it} - \bar{y}_i) = \boldsymbol{\beta}'(\mathbf{x}_{it} - \bar{\mathbf{x}}_i) + (u_{it} - \bar{u}_i)$$

- Running OLS on this equation gives results identical to LSDV.
- Sometimes called the *within-group estimator*, because it uses only the variation in y_{it} and \mathbf{x}_{it} within each cross-sectional unit to estimate the $\boldsymbol{\beta}$ coefficients.
- Any variation between cross-sectional units is assumed to spring from the unobserved fixed effects.

- Another way to approach this is to pre-multiply each cross-sectional unit equation ($\mathbf{y}_i = \mathbf{e}\alpha_i^* + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i$) by a $T \times T$ idempotent “sweep” matrix:

$$\mathbf{Q} = \mathbf{I}_t - \frac{1}{T}\mathbf{e}\mathbf{e}'$$

- This has the effect of sweeping out the α_i^* s and transforming the variables so that the values for each individual are measured in terms of deviations from their means over time:

$$\mathbf{Q}\mathbf{y}_i = \mathbf{Q}\mathbf{e}\alpha_i^* + \mathbf{Q}\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Q}\mathbf{u}_i \quad (8.5)$$

$$= \mathbf{Q}\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Q}\mathbf{u}_i \quad (8.6)$$

- Running OLS on this regression gives

$$\hat{\boldsymbol{\beta}}_{\text{CV}} = \left[\sum_{i=1}^N \mathbf{X}_i' \mathbf{Q} \mathbf{X}_i \right]^{-1} \left[\sum_{i=1}^N \mathbf{X}_i' \mathbf{Q} \mathbf{y}_i \right]$$

- The variance-covariance matrix is

$$\text{var}[\boldsymbol{\beta}_{\text{CV}}] = \sigma_u^2 \left[\sum_{i=1}^N \mathbf{X}_i' \mathbf{Q} \mathbf{X}_i \right]^{-1}$$

- Properties of β_{CV} : unbiased and consistent whether N or T or both tend to infinity.
- Note that the OLS estimate of α_i^* is unbiased, but is consistent only as $T \rightarrow \infty$.
 - With LSDV consistency is an issue: **incidental parameters problem**.
- A key advantages of FE estimators: can have correlation between \mathbf{x}_{it} and α_i^* .
- A key drawback: if time-invariant regressors are included in the model, the standard FE estimator will not produce estimates for the effects of these variables (perfect collinearity in LSDV).
 - There is an IV approach to produce estimates, but requires some exogeneity assumptions that may not be met in practice.
- The effects of slow-moving variables can also be estimated very imprecisely due to collinearity.

8.2 Application: Economic growth in 14 OECD countries

- Garrett examines the political component of economic performance in his '98 book *Partisan Politics in the Global Economy*.
- Question: how does labor centralization and left control of the government affect economic growth (esp. in terms of an interaction effect)?
 - Tests the social democratic corporatist model of economic performance: congruence b/t political and labor bargaining enhances performance.
 - I.e., centralized labor bargaining goes well w/ politically powerful left parties; decentralized bargaining is more congruent when right parties are in power.
- Data: 14 OECD countries observed annually from 1966–1990 (i.e., $N = 14$; $T = 25$).
- Dependent variable: GDP.
- Explanatory vars:
 - Oil dependence (OIL),
 - Overall OECD GDP growth, weighted for each country by its trade with the other OECD nations (DEMAND),
 - proportion of cabinet posts occupied by left parties (LEFTLAB),
 - degree of centralized labor bargaining as a measure of corporatism (CORP),
 - interaction b/t CORP and LEFTLAB (CLINT).

- OLS on this equation gives:

```
. regress gdp oild demand corp leftlab clint
```

Source	SS	df	MS			
Model	291.283034	5	58.2566069	Number of obs =	350	
Residual	1773.90603	344	5.15670357	F(5, 344) =	11.30	
				Prob > F =	0.0000	
				R-squared =	0.1410	
				Adj R-squared =	0.1286	
				Root MSE =	2.2708	
Total	2065.18906	349	5.91744717			

gdp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
oild	-15.2321	4.572497	-3.33	0.001	-24.22567	-6.238529
demand	.0049977	.000999	5.00	0.000	.0030328	.0069625
corp	-1.139716	.3043989	-3.74	0.000	-1.738433	-.5409982
leftlab	-1.483549	.3844653	-3.86	0.000	-2.239747	-.7273499
clint	.4547183	.1233779	3.69	0.000	.2120482	.6973883
_cons	5.919865	.7356383	8.05	0.000	4.47295	7.36678

- Including dummies for each country (except one) gives:

```
. regress gdp oild demand corp leftlab clint Icc_2 Icc_3 Icc_4 Icc_5 Icc_6 Icc_
> 7 Icc_8 Icc_9 Icc_10 Icc_11 Icc_12 Icc_13 Icc_14;
```

Source	SS	df	MS	Number of obs =	350
Model	686.921905	18	38.1623281	F(18, 331) =	9.16
Residual	1378.26716	331	4.16394912	Prob > F =	0.0000
-----				R-squared =	0.3326
-----				Adj R-squared =	0.2963
Total	2065.18906	349	5.91744717	Root MSE =	2.0406

gdp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
oild	-25.59808	5.946569	-4.30	0.000	-37.29592	-13.90025
demand	.0084949	.001129	7.52	0.000	.006274	.0107158
corp	-.2500641	.6654194	-0.38	0.707	-1.559048	1.05892
leftlab	-1.172257	.4468775	-2.62	0.009	-2.051335	-.2931785
clint	.5030912	.1596682	3.15	0.002	.1889988	.8171836
Icc_2	-.4136903	.5988751	-0.69	0.490	-1.591772	.7643909
Icc_3	-2.090873	.7898403	-2.65	0.009	-3.644613	-.5371336
Icc_4	-2.159732	.7633224	-2.83	0.005	-3.661307	-.658157
Icc_5	-2.587796	1.091167	-2.37	0.018	-4.734293	-.4412985
Icc_6	.6289216	.8440104	0.75	0.457	-1.031379	2.289222
Icc_7	-1.796217	1.255454	-1.43	0.153	-4.265892	.6734584
Icc_8	-3.993015	1.891938	-2.11	0.036	-7.714754	-.2712759
Icc_9	-.8709414	1.02877	-0.85	0.398	-2.894693	1.15281
Icc_10	-1.449112	1.301281	-1.11	0.266	-4.008935	1.110711
Icc_11	-3.893792	1.607724	-2.42	0.016	-7.056438	-.7311463
Icc_12	-3.489515	1.29017	-2.70	0.007	-6.027481	-.9515491
Icc_13	-3.10808	1.477907	-2.10	0.036	-6.015355	-.200806
Icc_14	2.929627	.6076861	4.82	0.000	1.734213	4.125041
_cons	3.374094	1.365808	2.47	0.014	.6873365	6.060852

- The F -test to determine if the dummies should be included gives:

```
. test Icc_2=Icc_3=Icc_4=Icc_5=Icc_6=Icc_7=Icc_8=Icc_9=Icc_10=Icc_11=Icc_12=Icc
> _13=Icc_14=0;
```

```
      F( 13,   331) =    7.31
      Prob > F =    0.0000
```

- Finally, estimating the fixed effects model:

```
. xtreg gdp oild demand corp leftlab clint, fe ;
```

```
Fixed-effects (within) regression           Number of obs   =       350
Group variable (i): country                 Number of groups =        14

R-sq:  within = 0.2315                      Obs per group:  min =       25
        between = 0.0461                      avg =       25.0
        overall = 0.0424                      max =       25

corr(u_i, Xb) = -0.7104                      F(5,331)        =       19.94
                                                Prob > F         =       0.0000
```

```
-----+-----
```

	gdp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
	oild	-25.59808	5.946569	-4.30	0.000	-37.29592	-13.90025
	demand	.0084949	.001129	7.52	0.000	.006274	.0107158
	corp	-.2500641	.6654194	-0.38	0.707	-1.559048	1.05892
	leftlab	-1.172257	.4468775	-2.62	0.009	-2.051335	-.2931785
	clint	.5030912	.1596682	3.15	0.002	.1889988	.8171836
	_cons	1.78165	1.961666	0.91	0.364	-2.077255	5.640556

```
-----+-----
```

	sigma_u	1.9296773					
	sigma_e	2.0405757					
	rho	.47208946	(fraction of variance due to u_i)				

```
-----+-----
```

```
F test that all u_i=0:      F(13, 331) =    7.31          Prob > F = 0.0000
```

Section 9

Random Effects Estimators

9.1 Intro

- Fixed effects is completely appropriate if we believe that the unit-specific effects are indeed fixed, estimable amounts that we can calculate for each cross-sectional observation.
- Thus, we believe that Sweden will always have an intercept of 1.2 units (for instance). If we were able to take another sample, we would once again estimate the same intercept for Sweden. There are cases, however, where we may not believe that we can estimate some fixed amount for each country.
- In particular, assume that we have a panel data model run on 20 countries, but which should be generalizable to 100 different countries. We cannot estimate the given intercept for each country or each type of country because we don't have all of them in the sample for which we estimate the model.
- In this case, we might want to estimate the β s on the explanatory variables taking into account that there could be country-specific effects that would enter as a random shock from a known distribution.
- If we go this route, we will be estimating a random effects model.

9.2 Deriving the random effects estimator

- Also called variance components b/c we can set up the disturbance in this way:

$$v_{it} = \alpha_i + \lambda_t + u_{it}.$$

- We make the following crucial assumptions about the variance and covariances of these components:

$$E[\alpha_i] = E[\lambda_t] = E[u_{it}] = 0$$

$$E[\alpha_i \lambda_t] = E[\alpha_i u_{it}] = E[\lambda_t u_{it}] = 0$$

$$E[\alpha_i \alpha_j] = \begin{cases} \sigma_\alpha^2 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

$$E[\lambda_t \lambda_s] = \begin{cases} \sigma_\lambda^2 & \text{if } t = s \\ 0 & \text{if } t \neq s \end{cases}$$

$$E[u_{it} u_{js}] = \begin{cases} \sigma_u^2 & \text{if } i = j, t = s \\ 0 & \text{otherwise} \end{cases}$$

$$E[\alpha_i \mathbf{x}'_{it}] = E[\lambda_t \mathbf{x}'_{it}] = E[u_{it} \mathbf{x}'_{it}] = \mathbf{0}$$

- Note that the $\text{var}[y_{it} | \mathbf{x}_{it}] = \sigma_y^2 = \sigma_\alpha^2 + \sigma_\lambda^2 + \sigma_u^2$.

- Let's add a general intercept to our model and set $\lambda_t = 0 \forall t$:

$$y_{it} = \mu + \boldsymbol{\beta}' \mathbf{x}_{it} + \alpha_i + u_{it} \quad (9.1)$$

- We can rewrite this in vector form:

$$\mathbf{y}_i = \tilde{\mathbf{X}}_i \boldsymbol{\delta} + \mathbf{v}_i \quad (9.2)$$

where

$$\tilde{\mathbf{X}}_i = \begin{matrix} T \times (K+1) \\ \left[\mathbf{e} \quad \mathbf{X}_i \right] \end{matrix}, \quad \boldsymbol{\delta} = \begin{matrix} (K+1) \times 1 \\ \left[\begin{matrix} \mu \\ \boldsymbol{\beta} \end{matrix} \right] \end{matrix}, \quad \mathbf{v}_i = \begin{matrix} T \times 1 \\ \left[\begin{matrix} v_{i1} \\ v_{i2} \\ \vdots \\ v_{iT} \end{matrix} \right] \end{matrix}, \quad v_{it} = \alpha_i + u_{it}$$

- The variance-covariance matrix of the T disturbance terms \mathbf{v}_i is:

$$\begin{aligned} \mathbf{V} = E[\mathbf{v}_i \mathbf{v}_i'] &= \begin{bmatrix} (\sigma_u^2 + \sigma_\alpha^2) & \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & (\sigma_u^2 + \sigma_\alpha^2) & \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ \vdots & & & \ddots & \vdots \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & (\sigma_u^2 + \sigma_\alpha^2) \end{bmatrix} \\ &= \sigma_u^2 \mathbf{I}_T + \sigma_\alpha^2 \mathbf{e} \mathbf{e}' \end{aligned}$$

- Note that

$$\mathbf{V}^{-1} = \frac{1}{\sigma_u^2} \left[\mathbf{I}_T - \frac{\sigma_\alpha^2}{\sigma_u^2 + T\sigma_\alpha^2} \right] \mathbf{e} \mathbf{e}'.$$

- The full variance-covariance matrix for all the NT observations is:

$$\boldsymbol{\Omega} = \begin{bmatrix} \mathbf{V} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{V} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{V} \end{bmatrix} = \mathbf{I}_N \otimes \mathbf{V}$$

- To produce parameter estimates, we could proceed in the manner that we did with the FE estimator—i.e., pre-multiply by \mathbf{Q} and run OLS.
 - This will give unbiased and consistent estimates.
 - However, if the α_i are assumed to be random rather than fixed, the CV estimator is not BLUE; the GLS estimator is.
 - Intuition: v_{it} and v_{is} both contain α_i , inducing correlation in the disturbances. Need to use GLS to get efficient estimates.

9.3 GLS Estimation

- The normal equations for the GLS estimator are

$$\left[\sum_{i=1}^N \tilde{\mathbf{X}}_i' \mathbf{V}^{-1} \tilde{\mathbf{X}}_i \right] \hat{\boldsymbol{\delta}}_{\text{GLS}} = \sum_{i=1}^N \tilde{\mathbf{X}}_i \mathbf{V}^{-1} \mathbf{y}_i$$

- We could write the GLS estimator simply as

$$\hat{\boldsymbol{\delta}}_{\text{GLS}} = \left[\sum_{i=1}^N \tilde{\mathbf{X}}_i' \mathbf{V}^{-1} \tilde{\mathbf{X}}_i \right]^{-1} \sum_{i=1}^N \tilde{\mathbf{X}}_i \mathbf{V}^{-1} \mathbf{y}_i \quad (9.3)$$

- But let's unpack this to get a better understanding of what is going on w/ this estimator. We can rewrite the inverse of the variance-covariance matrix as

$$\mathbf{V}^{-1} = \frac{1}{\sigma_u^2} \left[\mathbf{I}_T - \frac{1}{T} \mathbf{e} \mathbf{e}' + \psi \cdot \frac{1}{T} \mathbf{e} \mathbf{e}' \right] = \frac{1}{\sigma_u^2} \left[\mathbf{Q} + \psi \cdot \frac{1}{T} \mathbf{e} \mathbf{e}' \right]$$

where

$$\psi = \frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}$$

- Economizing on notation:

$$[W_{\tilde{x}\tilde{x}} + \psi B_{\tilde{x}\tilde{x}}] \begin{bmatrix} \hat{\mu} \\ \hat{\beta} \end{bmatrix}_{\text{GLS}} = W_{\tilde{x}y} + \psi B_{\tilde{x}y} \quad (9.4)$$

where

$$\begin{aligned} W_{\tilde{x}\tilde{x}} &= T_{\tilde{x}\tilde{x}} - B_{\tilde{x}\tilde{x}} & W_{\tilde{x}y} &= T_{\tilde{x}y} - B_{\tilde{x}y} \\ T_{\tilde{x}\tilde{x}} &= \sum_{i=1}^N \tilde{\mathbf{X}}_i' \tilde{\mathbf{X}}_i & T_{\tilde{x}y} &= \sum_{i=1}^N \tilde{\mathbf{X}}_i' \mathbf{y}_i \\ B_{\tilde{x}\tilde{x}} &= \frac{1}{T} \sum_{i=1}^N (\tilde{\mathbf{X}}_i' \mathbf{e} \mathbf{e}' \tilde{\mathbf{X}}_i) & B_{\tilde{x}y} &= \frac{1}{T} \sum_{i=1}^N (\tilde{\mathbf{X}}_i' \mathbf{e} \mathbf{e}' \mathbf{y}_i) \end{aligned}$$

- Some intuition: These matrices contain the sum of squares and the sums of cross products between groups ($B_{\tilde{x}\tilde{x}}$ & $B_{\tilde{x}y}$), within groups ($W_{\tilde{x}\tilde{x}}$ & $W_{\tilde{x}y}$), and for total variation ($T_{\tilde{x}\tilde{x}}$ & $T_{\tilde{x}y}$).
- Solving 9.4 gives

$$\begin{aligned} & \begin{bmatrix} \psi NT & \psi T \sum_{i=1}^N \bar{\mathbf{x}}_i' \\ \psi T \sum_{i=1}^N \bar{\mathbf{x}}_i & \sum_{i=1}^N \mathbf{X}_i' \mathbf{Q} \mathbf{X}_i + \psi T \sum_{i=1}^N \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i' \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\beta} \end{bmatrix}_{\text{GLS}} \\ &= \begin{bmatrix} \psi NT \bar{y} \\ \sum_{i=1}^N \mathbf{X}_i' \mathbf{Q} \mathbf{y}_i + \psi T \sum_{i=1}^N \bar{\mathbf{x}}_i \bar{y}_i \end{bmatrix} \end{aligned}$$

- Taking the partitioned inverse and solving for the parameters gives

$$\begin{aligned} \hat{\beta}_{\text{GLS}} &= \left[\frac{1}{T} \sum_{i=1}^N \mathbf{X}_i' \mathbf{Q} \mathbf{X}_i + \psi \sum_{i=1}^N (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})' \right]^{-1} \\ & \quad \times \left[\frac{1}{T} \sum_{i=1}^N \mathbf{X}_i' \mathbf{Q} \mathbf{y}_i + \psi \sum_{i=1}^N (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{y}_i - \bar{y}) \right] \\ \hat{\mu}_{\text{GLS}} &= \bar{y} - \hat{\beta}'_{\text{GLS}} \bar{\mathbf{x}} \end{aligned}$$

- We can rewrite $\hat{\beta}_{\text{GLS}}$ as

$$\Delta \hat{\beta}_b + (\mathbf{I}_k - \Delta) \hat{\beta}_{\text{CV}}$$

where

$$\begin{aligned} \Delta &= \psi T \left[\sum_{i=1}^N \mathbf{X}_i' \mathbf{Q} \mathbf{X}_i + \psi T \sum_{i=1}^N (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})' \right]^{-1} \\ &\quad \times \left[\sum_{i=1}^N (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})' \right], \\ \hat{\beta}_b &= \left[\sum_{i=1}^N (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})' \right]^{-1} \left[\sum_{i=1}^N (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{y}_i - \bar{y}) \right]. \end{aligned}$$

- In words: the GLS estimator is a weighted average of the b/t group estimator and the w/in group estimator, w/ ψ indicating the weight given to b/t group variation. Recall

$$\psi = \frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}$$

- As $\psi \rightarrow 1$, $\hat{\beta}_{\text{GLS}} \rightarrow T_{\tilde{x}\tilde{x}}^{-1} T_{\tilde{x}y}$ (i.e., the OLS estimator). This means that little variance is explained by the unit effects.
- As $\psi \rightarrow 0$, $\hat{\beta}_{\text{GLS}} \rightarrow$ the w/in estimator. This happens as either
 1. the unit-specific effects dominate the disturbance u_{it} .
 2. $T \rightarrow \infty$ (intuition: the α_i are like fixed parameters since we have so much data on the T dimension).
- GLS then is an intermediate approach b/t OLS and FE (which uses no b/t group variation).

- The variance of the GLS estimator is

$$\text{var} \left[\hat{\boldsymbol{\beta}}_{\text{GLS}} \right] = \sigma_u^2 \left[\sum_{i=1}^N \mathbf{X}_i' \mathbf{Q} \mathbf{X}_i + \psi T \sum_{i=1}^N (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})' \right]^{-1}$$

- Recall the variance for the w/in group estimator:

$$\text{var}[\boldsymbol{\beta}_{\text{CV}}] = \sigma_u^2 \left[\sum_{i=1}^N \mathbf{X}_i' \mathbf{Q} \mathbf{X}_i \right]^{-1}.$$

- The difference b/t these var-cov matrices is a p.d. matrix (assuming $\psi > 0$).
- Thus, as $T \rightarrow \infty$, $\psi \rightarrow 0$, and $\text{var} \left[\sqrt{T} \hat{\boldsymbol{\beta}}_{\text{GLS}} \right] \rightarrow \text{var} \left[\sqrt{T} \hat{\boldsymbol{\beta}}_{\text{CV}} \right]$ (assuming our cross-product matrices converge to finite p.d. matrices).
- Since we typically do not know σ_u^2 and σ_α^2 , they must be estimated. We can do two-step GLS (i.e., obtain consistent estimates of the variance components and then plug these in to compute 2nd stage parameter estimates).
- If either $N \rightarrow \infty$ or $T \rightarrow \infty$, 2-step GLS has the same asymptotic properties as GLS w/ known variance components.
- Can use w/in and b/t group residuals to compute estimates:

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^N \sum_{t=1}^T \left[(y_{it} - \bar{y}_i) - \hat{\boldsymbol{\beta}}'_{\text{CV}} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i) \right]^2}{N(T-1) - K} \quad (9.5)$$

$$\hat{\sigma}_\alpha^2 = \frac{\sum_{i=1}^N \left[\bar{y}_i - \tilde{\mu} - \tilde{\boldsymbol{\beta}}' \bar{\mathbf{x}}_i \right]^2}{N - (K+1)} - \frac{1}{T} \sigma_u^2 \quad (9.6)$$

where $\tilde{\mu}$ and $\tilde{\boldsymbol{\beta}}$ are obtained from $B_{\tilde{x}\tilde{x}}^{-1} B_{\tilde{x}y}$.

9.4 Maximum Likelihood Estimation

- RE estimates can also be computed by ML.
- To obtain the MLE, assume u_{it} and α_i are normally dist'd and start w/ the log of the likelihood function:

$$\begin{aligned} \ln L = & -\frac{NT}{2} \ln 2\pi - \frac{N}{2} \ln |\mathbf{V}| \\ & - \frac{1}{2} \sum_{i=1}^N (\mathbf{y}_i - \mathbf{e}\mu - \mathbf{X}_i\boldsymbol{\beta})' \mathbf{V}^{-1} (\mathbf{y}_i - \mathbf{e}\mu - \mathbf{X}_i\boldsymbol{\beta}) \end{aligned}$$

- To obtain the MLE $\hat{\boldsymbol{\delta}}' = (\mu, \boldsymbol{\beta}', \sigma_u^2, \sigma_\alpha^2)$, we take partial derivatives wrt each of these parameters, set to zero and solve.
- This gives four equations that we must solve simultaneously, which can be difficult.
- Instead we can use a sequential iterative procedure, alternating back and forth b/t μ and $\boldsymbol{\beta}$ and the variance components σ_u^2 and σ_α^2 .
- For N fixed and $T \rightarrow \infty$, the MLEs of μ , $\boldsymbol{\beta}'$, and σ_u^2 are consistent and \rightarrow CV estimator. The MLE of σ_α^2 is inconsistent (insufficient variation b/c of fixed N).
- With simultaneous solution of σ_α^2 , it's possible to obtain a negative value. It's also possible to obtain a boundary solution, although the prob. of this $\rightarrow 0$ as either T or $N \rightarrow \infty$.

9.5 Fixed v. Random Effects

- Makes no difference for large T .
- Can make a big difference when T is fixed and N is large.
- Does it make sense to treat one source of unobserved variation as random (u_{it}) and another as fixed (α_i)?
- Conditional v. marginal inference:
 - FE is often thought of as an approach where inferences are made conditional on effects that are in the sample.
 - RE can be thought of as making inferences that are unconditional or marginal wrt the pop. of all effects.
- Perhaps the most important consideration is whether we think the unit effects are correlated w/ explanatory variables—if so, RE is not appropriate (although may not make much difference in certain situations).

9.6 Testing between Fixed and Random Effects

- If α_i is uncorrelated with the explanatory variables \mathbf{x}_{it} :
 - GLS is unbiased/consistent and will achieve the Cramer-Rao lower bound (i.e., is efficient).
 - CV is unbiased/consistent but is inefficient.
- If α_i is correlated with any of the explanatory variables:
 - GLS is biased/inconsistent.
 - CV is unbiased/consistent.

- This sets us up for a Hausman test:

$H_0: E[\alpha_i \mathbf{x}_{it}] = \mathbf{0}$; Random effects appropriate $\Rightarrow \hat{\boldsymbol{\beta}}_{\text{GLS}}$ is approximately equal to $\hat{\boldsymbol{\beta}}_{\text{CV}}$ but is more efficient (has smaller standard errors).

$H_1: E[\alpha_i \mathbf{x}_{it}] \neq \mathbf{0}$; Random effects is not appropriate $\Rightarrow \hat{\boldsymbol{\beta}}_{\text{GLS}}$ will be different from $\hat{\boldsymbol{\beta}}_{\text{CV}}$ (and inconsistent).

- In this setting, the Hausman test statistic is calculated as:

$$m = (\hat{\boldsymbol{\beta}}_{\text{CV}} - \hat{\boldsymbol{\beta}}_{\text{GLS}})' \left(\text{var}[\hat{\boldsymbol{\beta}}_{\text{CV}}] - \text{var}[\hat{\boldsymbol{\beta}}_{\text{GLS}}] \right)^{-1} (\hat{\boldsymbol{\beta}}_{\text{CV}} - \hat{\boldsymbol{\beta}}_{\text{GLS}})$$

- $m \sim \chi_K^2$.
- If m is larger than its appropriate critical value, then we reject random effects as the appropriate specification.

9.7 Application

- Let's try random effects on the Garrett data:

```
xtreg gdp oild demand corp leftlab clint, re ;
```

```
Random-effects GLS regression                Number of obs    =    350
Group variable (i): country                 Number of groups  =    14

R-sq:  within = 0.2225                      Obs per group:  min =    25
        between = 0.0007                      avg =    25.0
        overall = 0.1255                      max =    25

Random effects u_i ~ Gaussian               Wald chi2(5)     =    87.18
corr(u_i, X) = 0 (assumed)                 Prob > chi2      =    0.0000
```

gdp	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
oild	-20.44602	5.394257	-3.79	0.000	-31.01857	-9.873467
demand	.0075601	.0010875	6.95	0.000	.0054286	.0096915
corp	-1.210037	.420998	-2.87	0.004	-2.035178	-.3848961
leftlab	-1.256097	.4275844	-2.94	0.003	-2.094147	-.4180465
clint	.4653267	.1481581	3.14	0.002	.1749422	.7557112
_cons	5.19839	1.111886	4.68	0.000	3.019134	7.377646
sigma_u	.98722663					
sigma_e	2.0405757					
rho	.18966702	(fraction of variance due to u_i)				

- Note the big differences in the coefficient values compared w/ FE (for FGLS MLE, replace `re` w/ `mle`).

- Let's run a Hausman test. The commands for this model would be

```
xtreg gdp oild demand corp leftlab clint, fe ;
```

```
est store garrettfe ;
```

```
xtreg gdp oild demand corp leftlab clint, re ;
```

```
est store garrettre ;
```

```
hausman garrettfe garrettre;
```

- The results:

	---- Coefficients ----			
	(b)	(B)	(b-B)	sqrt(diag(V_b-V_B))
	garrettfe	garrettre	Difference	S.E.
oild	-25.59808	-20.44602	-5.152068	2.502733
demand	.0084949	.0075601	.0009349	.0003033
corp	-.2500641	-1.210037	.9599731	.5153092
leftlab	-1.172257	-1.256097	.0838399	.1298888
clint	.5030912	.4653267	.0377645	.0595242

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)
 = 15.39
 Prob>chi2 = 0.0088

Section 10

Non-Spherical Errors

10.1 Introduction

- Up to this point, we have assumed that our errors were spherical.
- A good deal of attention, however, has been paid to issues of non-spherical errors in panel and TSCS data (esp. the latter).

10.2 The Method of PCSEs

- Key motivation for using panel corrected standard errors (PCSEs): improve inferences made from TSCS data by taking into account the complexity of the error process, but not ask too much of data.
- Non-standard error structures (TSCS):
 1. Contemporaneous correlation: errors across cross-sect'l units are correlated due to common shocks in a given time period.

$$E(u_{it}, u_{js}) = \begin{cases} \sigma_i^2 & \text{if } i = j \text{ and } s = t \\ \sigma_{ij} & \text{if } i \neq j \text{ and } s = t \\ 0 & \text{otherwise} \end{cases}$$

2. Panel heteroskedasticity: error var. differs across cross-sect'l units due to characteristics unique to the units.

$$E(u_{it}, u_{js}) = \begin{cases} \sigma_i^2 & \text{if } i = j \text{ and } s = t \\ 0 & \text{otherwise} \end{cases}$$

3. Serial correlation: errors w/ units are temporally correlated.

$$u_{it} = \rho u_{i,t-1} + \varepsilon_{it}.$$

- OLS not BLUE & can produce incorrect SEs when the errors are nonspherical.
- GLS is BLUE & gives correct SEs.
- But assumes that the var-cov matrix ($\mathbf{\Omega}$), used to weight the data is known.
- Can do FGLS using $\hat{\mathbf{\Omega}}$.
- Beck & Katz '95 *APSR* show, however, that the FGLS method advocated by Parks and Kmenta produces incorrect SEs when applied to TSCS data.
- FGLS gives overconfident SEs—does not fully take into account the variability in the estimates of the error parameters (rely too heavily on asymptotic properties).
- Beck & Katz '95 *APSR*: superior way to handle complex error structures w/ TSCS analysis is to estimate coefficients by OLS & compute PCSEs.
- Intuition: OLS with TSCS data will be unbiased but will produce incorrect standard errors.

10.3 Robust Estimation of Asymptotic Covariance Matrices

- In order to understand how PCSEs work, let's review robust estimators for cross-sectional data.
- The asymptotic covariance matrix of the OLS estimator of β is given by

$$\text{var}[\hat{\beta}] = \frac{1}{N} \left(\frac{1}{N} \mathbf{X}'\mathbf{X} \right)^{-1} \left(\frac{1}{N} \mathbf{X}'\Sigma\mathbf{X} \right) \left(\frac{1}{N} \mathbf{X}'\mathbf{X} \right)^{-1} \quad (10.1)$$

where $\Sigma = E[\mathbf{u}\mathbf{u}'|\mathbf{X}]$.

- The problem here is how to estimate $N^{-1}\mathbf{X}'\Sigma\mathbf{X}$, since Σ contains $N(N+1)/2$ unknown parameters and we have only N data points.
- Fortunately, all we need is an estimator of the $K(K+1)/2$ unknown elements of

$$\text{plim } \mathbf{Q}_* = \text{plim } \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \sigma_{ij} \mathbf{x}_i \mathbf{x}_j' \quad (10.2)$$

- \mathbf{Q}_* is a matrix of the sums of squares and cross-products that involves σ_{ij} and the rows of \mathbf{X} .
- Since OLS $\hat{\beta}$ is consistent for β , then the OLS residuals \hat{u}_i will be consistent for u_i , and can be used to construct estimates of σ_{ij} .
- For the case of heteroskedasticity, we want to estimate

$$\mathbf{Q}_* = \frac{1}{N} \sum_{i=1}^N \sigma_i^2 \mathbf{x}_i \mathbf{x}_i' \quad (10.3)$$

- White (1980) has shown that for

$$\mathbf{S}_0 = \frac{1}{N} \sum_{i=1}^N \hat{u}_i^2 \mathbf{x}_i \mathbf{x}_i', \quad (10.4)$$

$$\text{plim } \mathbf{S}_0 = \text{plim } \mathbf{Q}_*$$

under very general conditions.

- This gives the White heteroskedasticity consistent estimator:

$$\begin{aligned}\widehat{\text{var}}[\hat{\boldsymbol{\beta}}] &= \frac{1}{N} \left(\frac{1}{N} \mathbf{X}'\mathbf{X} \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^N \hat{u}_i^2 \mathbf{x}_i \mathbf{x}_i' \right) \left(\frac{1}{N} \mathbf{X}'\mathbf{X} \right)^{-1} \\ &= N (\mathbf{X}'\mathbf{X})^{-1} \mathbf{S}_0 (\mathbf{X}'\mathbf{X})^{-1}\end{aligned}\quad (10.5)$$

- Another way to get to this result is to consider the asymptotic distribution of the OLS estimator:

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \stackrel{a}{\sim} N(\mathbf{0}, E[\mathbf{x}'\mathbf{x}]^{-1} \cdot E[u^2 \mathbf{x}'\mathbf{x}] \cdot E[\mathbf{x}'\mathbf{x}]^{-1}) \quad (10.6)$$

which implies

$$\text{asy. var}[\hat{\boldsymbol{\beta}}] = N^{-1} E[\mathbf{x}'\mathbf{x}]^{-1} \cdot E[u^2 \mathbf{x}'\mathbf{x}] \cdot E[\mathbf{x}'\mathbf{x}]^{-1} \quad (10.7)$$

- To consistently estimate $E[\mathbf{x}'\mathbf{x}]$, we use the sample averages:

$$N^{-1} \sum_{i=1}^N \mathbf{x}_i' \mathbf{x}_i = (\mathbf{X}'\mathbf{X}/N).$$

- By the law of large numbers, $\text{plim } N^{-1} \sum_{i=1}^N u_i^2 \mathbf{x}_i' \mathbf{x}_i = E[u^2 \mathbf{x}'\mathbf{x}]$. Replacing u_i with OLS residuals gives a consistent estimator of this expectation.
- Putting this altogether gives the estimator of the asymptotic variance:

$$\widehat{\text{asy. var}}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{i=1}^N \hat{u}_i^2 \mathbf{x}_i' \mathbf{x}_i \right) (\mathbf{X}'\mathbf{X})^{-1} \quad (10.8)$$

- For PCSEs, $\boldsymbol{\Omega}$ is an $NT \times NT$ band diagonal matrix with cross-sectional variances along the diagonal and contemporaneous correlations in the bands.
- For example, if $N = 2$ and $T = 3$, then

$$\boldsymbol{\Omega} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \sigma_{12} & 0 & 0 \\ 0 & \sigma_1^2 & 0 & 0 & \sigma_{12} & 0 \\ 0 & 0 & \sigma_1^2 & 0 & 0 & \sigma_{12} \\ \sigma_{12} & 0 & 0 & \sigma_2^2 & 0 & 0 \\ 0 & \sigma_{12} & 0 & 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_{12} & 0 & 0 & \sigma_2^2 \end{bmatrix}$$

- More generally,

$$\mathbf{\Omega} = \begin{bmatrix} \sigma_1^2 \mathbf{I}_T & \sigma_{12} \mathbf{I}_T & \cdots & \sigma_{1N} \mathbf{I}_T \\ \sigma_{21} \mathbf{I}_T & \sigma_2^2 \mathbf{I}_T & \cdots & \sigma_{2N} \mathbf{I}_T \\ \vdots & & \ddots & \vdots \\ \sigma_{N1} \mathbf{I}_T & \sigma_{N2} \mathbf{I}_T & \cdots & \sigma_N^2 \mathbf{I}_T \end{bmatrix}$$

- Let

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1N} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \cdots & \sigma_{2N} \\ \vdots & & & \ddots & \vdots \\ \sigma_{1N} & \sigma_{2N} & \sigma_{3N} & \cdots & \sigma_N^2 \end{bmatrix}$$

- Use OLS residuals, denoted e_{it} for unit i at time t (in Beck and Katz's notation), to estimate the elements of $\mathbf{\Sigma}$:

$$\hat{\Sigma}_{ij} = \frac{\sum_{t=1}^T e_{it} e_{jt}}{T}, \quad (10.9)$$

which means the estimate of the full matrix $\hat{\Sigma}$ is

$$\hat{\Sigma} = \frac{\mathbf{E}'\mathbf{E}}{T}$$

where \mathbf{E} is a $T \times N$ matrix of the re-shaped $NT \times 1$ vector of OLS residuals, such that the columns contains the $T \times 1$ vectors of residuals for each cross-sectional unit (or conversely, each row contains the $N \times 1$ vector of residuals for each cross-sectional in a given time period) :

$$\mathbf{E} = \begin{bmatrix} e_{11} & e_{21} & \cdots & e_{N1} \\ e_{12} & e_{22} & \cdots & e_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ e_{1T} & e_{2T} & \cdots & e_{NT} \end{bmatrix}$$

Then

$$\hat{\mathbf{\Omega}} = \frac{\mathbf{E}'\mathbf{E}}{T} \otimes \mathbf{I}_T,$$

- Compute SEs using the square roots of the diagonal elements of

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, \quad (10.10)$$

where \mathbf{X} denotes the $NT \times k$ matrix of stacked vectors of explanatory variables, \mathbf{x}_{it} .

- Intuition behind why PCSEs do well: similar to White's heteroskedasticity-consistent standard errors for cross-sect'l estimators, but better b/c take advantage of info provided by the panel structure.
- Good small sample properties confirmed by Monte Carlo studies.
- Potential problems:
 - Note that this methods ignores unit effects—doesn't have to, though.
 - PCSEs solves the problems of panel heteroskedasticity and contemporaneous correlation—not serial correlation. Serial correlation must be removed before applying this fix.
 - How to correct for serial correlation? Lags are recommended, but that introduces a potential problem that we'll discuss when we turn to dynamic specifications.
 - Can also do Prais-Winsten.

10.4 Costs of ignoring unit effects revisited

- A dispute has arisen about the value of FE estimators when theory tells you that time-invariant random variables should be included in your specification—better to just do OLS w/ PCSEs?
- We can run into trouble if unit effects are present in the data, but we ignore them for the sake of including time-invariant variables; the problem is that we don't know for sure what the true specification is.
- Methods for robust standard error estimation can help shed light on this question.
- Suppose the DGP is described by the following equation:

$$y_{it} = \boldsymbol{\beta}' \mathbf{x}_{it} + \alpha_i + u_{it}, \quad (10.11)$$

where α_i indicate the unit effects and we assume u_{it} is spherical.

- If we ignore the unit effects then we are estimating the above model with the disturbance $v_{it} = \alpha_i + u_{it}$.
- If α_i is correlated with \mathbf{x}_{it} , then this leads to bias and inconsistency in OLS estimates of $\boldsymbol{\beta}$.
- If it is not correlated there can still be problems with inferences: relegating α_i to the disturbance term in essence induces serial correlation in the errors.
- The variance-covariance matrix for $\hat{\boldsymbol{\beta}}$ is given by

$$\text{var}[\hat{\boldsymbol{\beta}}] = E[(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{v} \mathbf{v}' \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1}], \quad (10.12)$$

where \mathbf{X} and \mathbf{v} are the \mathbf{x}_{it} and \mathbf{v}_{it} stacked over all i and t .

- Let $\boldsymbol{\Sigma} = E[\mathbf{X}' \mathbf{v} \mathbf{v}' \mathbf{X}]$. For the case of repeated cross-sections, this can be rewritten as

$$\boldsymbol{\Sigma} = E \left[\sum_i \sum_j \sum_t \sum_s \mathbf{x}_{it} v_{it} v_{js} \mathbf{x}'_{js} \right] \quad (10.13)$$

- If $v_{it} = \alpha_i + u_{it}$, then

$$\Sigma = E \left[\sum_i \sum_t \sum_s \mathbf{x}_{it} \mathbf{x}'_{it} v_{it}^2 \right] + 2E \left[\sum_i \sum_{t>s} \mathbf{x}_{it} \mathbf{x}'_{is} \alpha_i^2 \right] \quad (10.14)$$

- Even if the u_{it} are spherical, the standard OLS estimator for the variance-covariance matrix will be wrong, since the second term in eq. 10.14 will be ignored if the α_i are not accounted for in the model specification.
- PCSEs will also ignore this term, leading to wrong standard errors.
- Interestingly, this presents a problem particularly for time invariant variables b/c the std. errs. for the coefficients on such variables will generally be too small, possibly leading to type I errors.
- Consider the second term of eq. 10.14 when $\mathbf{x}_{it} = (x_{it} \ z_i)$:

$$2E \left[\sum_i \sum_{t \neq s} \begin{bmatrix} x_{it} x_{is} & x_{it} z_i \\ x_{is} z_i & z_i z_i \end{bmatrix} \alpha_i^2 \right] \quad (10.15)$$

- It is possible that the x s are uncorrelated across cross-sectional units and time periods and are uncorrelated w/ the z s \Rightarrow 0s in the first diagonal element and the off-diagonals—do not contribute anything to the standard errors in expectation.
- However, the z s are perfectly correlated within $i \Rightarrow$ positive number when multiplied by $\alpha^2 \Rightarrow$ larger standard errors than what we would get from OLS.
- To confirm this analytical result, conduct Monte Carlo analysis: generate the data with an explanatory variable, x_{it} , and a unit effect, but then estimate a model that replaces the unit effect with a randomly generated z_i , which is independent of x_{it} and the unit effect.
 - This set-up mimicks a scenario where a researcher forsakes the FE approach to include a time invariant variable, even though—unbeknown to the researcher—the time invariant variable actually has no effect.

Table 10.1: Results from Monte Carlo experiments involving time invariant variables

	No lag	One lag	Two lags
Mean OLS γ	-0.005	-0.003	0.007
Mean Between γ	-0.002	0.000	0.007
Proportion of sig. OLS γ_s	0.573	0.991	0.976
Proportion of sig. OLS γ_s , PCSEs	0.693	0.972	0.949
Proportion of sig. OLS γ_s , Robust	0.134	0.958	0.948
Proportion of sig. Between γ_s	0.073	0.071	0.074
% reject H_0 of no autocorrelation	99	96	10

Notes: $N = 15; T = 20$. 1000 simulations each.

- We estimated the model via standard OLS as well as with PCSEs. To verify that the problem is a kind of serial correlation induced by omission of the unit effects, we computed standard errors that are robust to serial correlation using

$$\widehat{\text{var}}(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{i=1}^N \sum_{t=1}^T \sum_{s=1}^T \mathbf{x}_{it}\mathbf{x}'_{is} \hat{v}_{it}\hat{v}_{is} \right) (\mathbf{X}'\mathbf{X})^{-1}, \quad (10.16)$$

where the \hat{v}_{it} are the residuals from a pooled OLS regression.

10.5 Heteroskedasticity in FE and RE models

- If we treat the α_i as random, then heteroskedasticity can occur if we have $\sigma_{\alpha i}^2$ or σ_{ui}^2 or both:

$$\mathbf{V}_i = E[\mathbf{v}_i \mathbf{v}_i'] = \sigma_{ui}^2 \mathbf{I}_T + \sigma_{\alpha i}^2 \mathbf{e} \mathbf{e}'$$

- To do GLS, we use \mathbf{V}_i instead of \mathbf{V} .
- Since we typically do not know either $\sigma_{\alpha i}^2$ or σ_{ui}^2 , we can resort to a 2-step FGLS approach.
 - Problem: cannot get a consistent estimate for $\sigma_{\alpha i}^2$ even as $T \rightarrow \infty$; only one realization of α_i (incidental parameters).
 - Can get a consistent estimate of σ_{ui}^2 as $T \rightarrow \infty$.
 - W/ finite T , cannot get consistent estimates of either $\sigma_{\alpha i}^2$ or σ_{ui}^2 , even if $N \rightarrow \infty$.
 - If N and T are getting big, then we can get consistent estimates of σ_{ui}^2 .

- To do FGLS:

1. Run OLS or CV on $\mathbf{y}_i = \tilde{\mathbf{X}}_i \boldsymbol{\delta} + \mathbf{v}_i$ and compute the estimated residuals $\hat{\mathbf{V}}_i$.
2. Compute $\hat{\mathbf{V}}_i$ either by
 - (a) assuming $\sigma_{\alpha_i}^2 = \sigma_\alpha^2 \forall i$ and using

$$\hat{\sigma}_{ui}^2 = \frac{1}{T-1} \sum_{t=1}^T (\hat{v}_{it} - \hat{v}_i)^2, \quad (10.17)$$

or

- (b) assuming the variance of α_i conditional on \mathbf{x}_i has the same functional form across individuals $\text{var}[\alpha_i | \mathbf{x}_i] = \sigma^2 \mathbf{x}_i$; $\hat{\sigma}_{ui}^2$ is estimated as in eq. 10.17 (see Roy '02 *IER* for a method of this type for heteroskedasticity of unknown form).
3. Then do

$$\hat{\boldsymbol{\delta}}_{\text{FGLS}} = \left[\sum_{i=1}^N \tilde{\mathbf{X}}_i' \hat{\mathbf{V}}_i^{-1} \tilde{\mathbf{X}}_i \right]^{-1} \sum_{i=1}^N \tilde{\mathbf{X}}_i \hat{\mathbf{V}}_i^{-1} \mathbf{y}_i$$

and approximate the asymptotic var-cov matrix of $\hat{\boldsymbol{\delta}}_{\text{FGLS}}$ by

$$\left(\sum_{i=1}^N \tilde{\mathbf{X}}_i' \hat{\mathbf{V}}_i^{-1} \tilde{\mathbf{X}}_i \right)^{-1}$$

- Could also do a feasible weighted LS method, weighting each obs. by the reciprocal of $\hat{\sigma}_{ui}$ and then apply the CV estimator to the transformed data.

10.6 Serial Correlation in RE and FE models

- Including α_i is in essence a way to account for unobserved persistence in the data.
- Unobserved persistence can also show up in the form of serial correlation in the disturbances.
- Suppose we have the classic AR(1) disturbance:

$$u_{it} = \rho u_{i,t-1} + \varepsilon_{it} \quad (10.18)$$

where the ε_{it} are iid w/ mean zero and variance σ_ε^2 .

- If we knew ρ then we could write down a standard variance components model:

$$y_{it} - \rho y_{i,t-1} = \mu(1 - \rho) + \boldsymbol{\beta}'(\mathbf{x}_{it} - \rho \mathbf{x}_{i,t-1}) + (1 - \rho)\alpha_i + \varepsilon_{it} \quad (10.19)$$

- All we are doing here is substituting in $y_{i,t-1} - \mu - \boldsymbol{\beta}'\mathbf{x}_{i,t-1} - \alpha_i$ for $u_{i,t-1}$.

- Can get an asymptotically efficient estimator of β by doing the following:

1. De-mean the data to take out the α_i :

$$(y_{it} - \bar{y}_i) = \beta'(\mathbf{x}_{it} - \bar{\mathbf{x}}_i) + (u_{it} - \bar{u}_i)$$

2. Compute an estimate of ρ :

(a) Use the LS residual from the de-meaned regression, or

(b) regress $(y_{it} - \bar{y}_i)$ on $(y_{i,t-1} - \bar{y}_{i,-1})$ and $(\mathbf{x}_{i,t-1} - \bar{\mathbf{x}}_{i,-1})$ and use the coefficient on $(y_{i,t-1} - \bar{y}_{i,-1})$ as an estimate of ρ (note: $\bar{y}_{i,-1} = (1/T) \sum_{t=1}^T y_{i,t-1}$; we assume we have measures of y_{i0} and x_{i0}).

3. Compute estimates of σ_ε^2 and σ_α^2 :

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \left\{ (y_{it} - \bar{y}_i) - (1 - \hat{\rho})\hat{\mu} - \hat{\rho}(y_{i,t-1} - \bar{y}_{i,-1}) - \hat{\beta}' [(\mathbf{x}_{it} - \bar{\mathbf{x}}_i) - (\mathbf{x}_{i,t-1} - \bar{\mathbf{x}}_{i,-1})\hat{\rho}] \right\}^2$$

$$\hat{\sigma}_\alpha^2 = \frac{1}{(1 - \hat{\rho})^2} \cdot \frac{1}{N} \sum_{i=1}^N \left[\bar{y}_i - \hat{\mu}(1 - \hat{\rho}) - \hat{\rho}\bar{y}_{i,-1} - \hat{\beta}'(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_{i,-1}\hat{\rho}) \right]^2 - \frac{1}{T} \hat{\sigma}_\varepsilon^2$$

Intuition: see equations for computing GLS variance estimates under assumption of spherical disturbances (i.e., eq. 9.5 and 9.6).

4. Plug in our estimate for $\hat{\rho}$ and use $\hat{\sigma}_\varepsilon^2$ and $\hat{\sigma}_\alpha^2$ to compute the var-cov matrix of $(1 - \hat{\rho})\alpha_i + \varepsilon_{it}$.

10.7 Robust standard error estimation with unit effects

- The properties of the FGLS estimators just described depend crucially on asymptotics—may not do well in finite samples.
- Might be better off computing some kind of robust estimator of the var-cov matrix to get our standard errors (i.e., robust to both heteroskedasticity and serial correlation).
- Robust estimators for FE models exist; although they also rely on asymptotic properties, they may do better in finite samples since they demand less of the data.

10.7.1 Arellano robust standard errors

- Let $\mathbf{y}_i^+ = \mathbf{Q}\mathbf{y}_i$, $\mathbf{X}_i^+ = \mathbf{Q}\mathbf{X}_i$, $\mathbf{u}_i^+ = \mathbf{Q}\mathbf{u}_i$.
- Arellano ('87 *Oxford Bulletin of Economics and Statistics*) suggests the following robust estimator:

$$(\mathbf{X}^{+'}\mathbf{X}^+)^{-1} \left(\sum_{i=1}^N \mathbf{X}_i^{+'} \hat{\mathbf{u}}_i^+ \hat{\mathbf{u}}_i^{+'} \mathbf{X}_i^+ \right) (\mathbf{X}^{+'}\mathbf{X}^+)^{-1} \quad (10.20)$$

where $\hat{\mathbf{u}}_i^+$ are the estimated residuals obtained from running OLS on the transformed equation (i.e., $\hat{\mathbf{u}}_i^+ = \mathbf{y}_i^+ - \mathbf{X}_i^+ \hat{\boldsymbol{\beta}}_{\text{CV}}$).

10.7.2 Kiefer robust standard errors

- Arellano also suggests (following Kiefer) the robust estimator:

$$(\mathbf{X}^{+\prime}\mathbf{X}^+)^{-1} \left(\sum_{i=1}^N \mathbf{x}_i^{+\prime} \hat{\boldsymbol{\Omega}}_i^+ \mathbf{x}_i^+ \right) (\mathbf{X}^{+\prime}\mathbf{X}^+)^{-1} \quad (10.21)$$

where

$$\hat{\boldsymbol{\Omega}}_i^+ = \frac{1}{N} \sum_{i=1}^N \hat{\mathbf{u}}_i^+ \hat{\mathbf{u}}_i^{+\prime}$$

- Both of these estimators assume T fixed and $N \rightarrow \infty$.
- In some Monte Carlo work I've done, the Arellano estimator does better than Kiefer for TSCS data.
- FE robust std. errs. can be obtained in Stata by using the `robust` and `cluster` options in the `xtreg` routine.
- Note that the Arellano approach is essentially the same as doing LSDV and then computing PCSEs.

10.8 Application: Garrett data

- The Garrett model of GDP growth estimated w/ OECD data is exactly the kind of case where we might expect non-spherical errors.

```
. xtpcse gdp oild demand corp leftlab clint ;
```

Linear regression, correlated panels corrected standard errors (PCSEs)

```
Group variable:  country                Number of obs      =      350
Time variable:  year                    Number of groups   =       14
Panels:         correlated (balanced)    Obs per group: min =       25
Autocorrelation: no autocorrelation      avg                =       25
                                                max                =       25

Estimated covariances      =      105      R-squared           =      0.1410
Estimated autocorrelations =         0      Wald chi2(5)       =      42.01
Estimated coefficients     =         6      Prob > chi2        =      0.0000
```

	Panel-corrected						
gdp	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]		

oild	-15.2321	5.228694	-2.91	0.004	-25.48015	-4.984049	
demand	.0049977	.0015394	3.25	0.001	.0019804	.0080149	
corp	-1.139716	.2234088	-5.10	0.000	-1.577589	-.7018424	
leftlab	-1.483549	.2755847	-5.38	0.000	-2.023685	-.9434125	
clint	.4547183	.0839526	5.42	0.000	.2901741	.6192624	
_cons	5.919865	.5833949	10.15	0.000	4.776432	7.063298	

- PCSEs, correcting for serial correlation:

```
. xtpcse gdp oild demand corp leftlab clint, correlation(ar1) ;
```

Prais-Winsten regression, correlated panels corrected standard errors (PCSEs)

```
Group variable:   country                Number of obs   =   350
Time variable:   year                    Number of groups =   14
Panels:          correlated (balanced)    Obs per group: min =   25
Autocorrelation: common AR(1)           avg =   25
                                                    max =   25
Estimated covariances =   105            R-squared       =   0.1516
Estimated autocorrelations =   1         Wald chi2(5)    =   31.55
Estimated coefficients =   6             Prob > chi2     =   0.0000
```

	Panel-corrected					
gdp	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
oild	-13.77226	6.587739	-2.09	0.037	-26.684	-.8605317
demand	.0060806	.0016414	3.70	0.000	.0028635	.0092977
corp	-1.177445	.2934019	-4.01	0.000	-1.752502	-.6023874
leftlab	-1.46776	.3623476	-4.05	0.000	-2.177949	-.757572
clint	.4488461	.1112233	4.04	0.000	.2308525	.6668397
_cons	5.814019	.807692	7.20	0.000	4.230972	7.397066
rho	.2958842					

- Fixed effects w/ robust standard errors à la Arellano:

```
. xtreg gdp oild demand corp leftlab clint, fe robust cluster(country)
```

```
Fixed-effects (within) regression      Number of obs      =      350
Group variable (i): country            Number of groups   =       14

R-sq:  within = 0.2315                  Obs per group: min =       25
      between = 0.0461                    avg =              25.0
      overall = 0.0424                    max =              25

corr(u_i, Xb) = -0.7104                  F(5,345)           =      14.29
                                          Prob > F            =      0.0000
```

(Std. Err. adjusted for 14 clusters in country)

	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
gdp						
oild	-25.59808	14.10159	-1.82	0.093	-56.06271	4.866543
demand	.0084949	.0020412	4.16	0.001	.0040851	.0129047
corp	-.2500641	.9970869	-0.25	0.806	-2.404139	1.904011
leftlab	-1.172257	.3593795	-3.26	0.006	-1.948649	-.3958644
clint	.5030912	.1346974	3.73	0.002	.2120951	.7940873
_cons	1.78165	3.051424	0.58	0.569	-4.810551	8.373852
sigma_u	1.9296773					
sigma_e	2.0405757					
rho	.47208946	(fraction of variance due to u_i)				

- Note that

```
areg gdp oild demand corp leftlab clint, absorb(country) cluster(country)
```

gives the same results.

- For kicks, let's do FGLS:

```
. xtgls gdp oild demand corp leftlab clint, panels(correlated);
```

Cross-sectional time-series FGLS regression

Coefficients: generalized least squares

Panels: heteroskedastic with cross-sectional correlation

Correlation: no autocorrelation

Estimated covariances	=	105	Number of obs	=	350
Estimated autocorrelations	=	0	Number of groups	=	14
Estimated coefficients	=	6	Time periods	=	25
			Wald chi2(5)	=	108.70
Log likelihood	=	-627.671	Prob > chi2	=	0.0000

gdp	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
oild	-14.01453	2.347816	-5.97	0.000	-18.61616	-9.412892
demand	.0045992	.0008527	5.39	0.000	.0029279	.0062705
corp	-1.107322	.1370615	-8.08	0.000	-1.375958	-.8386865
leftlab	-1.247823	.1348838	-9.25	0.000	-1.512191	-.9834559
clint	.4114256	.0479446	8.58	0.000	.3174559	.5053952
_cons	5.41023	.276808	19.55	0.000	4.867696	5.952763

Part IV

Qualitative and Limited Dependent Variable Models Based on the Normal Regression Model

Section 11

Introduction

2 parts:

1. Latent model: what's really going on.
2. Observation scheme: determines what we get to see.

- Latent model for an individual:

$$y_i^* = \boldsymbol{\beta}' \mathbf{x}_i + u_i$$

where y_i^* is a scalar, $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$, $\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ik} \end{bmatrix}$, and $u_i \sim N[0, \sigma^2]$.

$$\boldsymbol{\beta}' \mathbf{x}_i = [\beta_1 \cdots \beta_k] \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ik} \end{bmatrix} = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

11.1 Linear Regression Model

- Observation Scheme #1: $y_i = y_i^*$
- That is, we get to see everything, so we have the classic normal linear regression model.

Section 12

Probit

- Observation Scheme #2:

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

$$y_i^* > 0 \Rightarrow \boldsymbol{\beta}'\mathbf{x}_i + u_i > 0 \Rightarrow u_i > -\boldsymbol{\beta}'\mathbf{x}_i$$

$$\begin{aligned} \Pr(y_i^* > 0 | \mathbf{x}_i) &= \Pr(y_i = 1 | \mathbf{x}_i) = \Pr(u_i > -\boldsymbol{\beta}'\mathbf{x}_i) \\ &= \Pr\left(\frac{u_i}{\sigma} > -\frac{\boldsymbol{\beta}'\mathbf{x}_i}{\sigma}\right) \end{aligned}$$

- Given the distribution we've assumed for u_i , we can rewrite this probability as

$$\begin{aligned} \Pr(y_i = 1 | \mathbf{x}_i) &= 1 - \Phi\left(-\frac{\boldsymbol{\beta}'\mathbf{x}_i}{\sigma}\right) = \Phi\left(\frac{\boldsymbol{\beta}'\mathbf{x}_i}{\sigma}\right) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\boldsymbol{\beta}'\mathbf{x}_i/\sigma} \exp\left(-\frac{u^2}{2}\right) du \end{aligned}$$

- Similarly

$$\Pr(y_i = 0|\mathbf{x}_i) = 1 - \Phi\left(\frac{\boldsymbol{\beta}'\mathbf{x}_i}{\sigma}\right)$$

- An identification issue: in the probit model we cannot estimate $\boldsymbol{\beta}$ and σ separately because they enter as a ratio (only the ratio is identifiable). So we just set $\sigma^2 = 1$.
- The likelihood function for the probit model then is

$$L = \prod_{i=1}^n [\Phi(\boldsymbol{\beta}'\mathbf{x}_i)]^{y_i} [1 - \Phi(\boldsymbol{\beta}'\mathbf{x}_i)]^{1-y_i}$$

while the log likelihood is

$$\ln L = \sum_{i=1}^n \{y_i \ln \Phi(\boldsymbol{\beta}'\mathbf{x}_i) + (1 - y_i) \ln [1 - \Phi(\boldsymbol{\beta}'\mathbf{x}_i)]\}$$

- Differentiating w.r.t. $\boldsymbol{\beta}$ gives

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n \frac{y_i - \Phi(\boldsymbol{\beta}'\mathbf{x}_i)}{\Phi(\boldsymbol{\beta}'\mathbf{x}_i) [1 - \Phi(\boldsymbol{\beta}'\mathbf{x}_i)]} \phi(\boldsymbol{\beta}'\mathbf{x}_i) \mathbf{x}_i$$

- Taking the second derivatives gives

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = - \sum_{i=1}^n \phi \left[y_i \frac{\phi + \Phi \boldsymbol{\beta}'\mathbf{x}_i}{\Phi^2} + (1 - y_i) \frac{\phi - (1 - \Phi) \boldsymbol{\beta}'\mathbf{x}_i}{[1 - \Phi]^2} \right] \mathbf{x}_i \mathbf{x}_i'$$

where $\phi = \phi(\boldsymbol{\beta}'\mathbf{x}_i)$ and $\Phi = \Phi(\boldsymbol{\beta}'\mathbf{x}_i)$.

12.1 Interpretation of Coefficients

- Estimated coefficients from the probit models do not have the same interpretation as they do in the linear regression model.

- In the linear model:

$$\tilde{y}_i = \boldsymbol{\beta}' \mathbf{x}_i + u_i \quad \text{so} \quad \frac{\partial \tilde{y}_i}{\partial \mathbf{x}_i} = \boldsymbol{\beta}$$

- In the probit model we have

$$\frac{\partial \Pr(y_i = 1 | \mathbf{x}_i)}{\partial \mathbf{x}_i} = \frac{\partial \Phi(\boldsymbol{\beta}' \mathbf{x}_i)}{\partial \mathbf{x}_i} = \phi(\boldsymbol{\beta}' \mathbf{x}_i) \boldsymbol{\beta}$$

- The effect of a unit change in x_{ik} on the probability that $y_i = 1$ depends on the value of x_{ik} as well as the values of other explanatory variables in the model.
- Note that to compute this probability you must choose some values for the x variables (e.g., medians or means for the sample).
- Another standard technique is to set the explanatory variables equal to certain values and then change the value of a particular variable of interest to see how it changes the probability of an event occurring.

- To be thorough, one should report measures of uncertainty for marginal effects (see Herron “Post-Estimation Uncertainty in Limited Dependent Variable Models” ’00 *Political Analysis*).
- For example, we can use simulation methods to compute confidence intervals for predicted probabilities. To do this, we’re going to get a bit Bayesian, by assuming $\boldsymbol{\beta} \stackrel{a}{\sim} N(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}})$, where $\hat{\boldsymbol{\Sigma}}$ denotes the estimated covariance matrix.
 1. Estimate $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}$.
 2. Repeat the following steps S times (e.g., $S = 1000$), indexing each iteration by l .
 - (a) Draw a vector $\tilde{\boldsymbol{\beta}}$ from a multivariate normal distribution with mean vector $\hat{\boldsymbol{\beta}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$.
 - (b) Calculate and save $\hat{p}_i^l = \Phi(\tilde{\boldsymbol{\beta}}' \mathbf{x}_i)$.
 3. Use the percentiles of the distribution of $\hat{p}_i^l, l = 1, \dots, S$, to form confidence intervals of the desired size (e.g., 0.025 and 0.975 percentiles for a 95% confidence interval).

12.2 Goodness of fit measures

- We can use a Likelihood Ratio (LR) test to compute a goodness of fit statistic similar to the F statistic in the linear regression context that tests the null hypothesis that all of the slope coefficients are simultaneously equal to zero.
- Let $\boldsymbol{\theta}$ be a vector of parameters to be estimated and H_0 specify some restrictions on this vector. The LR test is given by

$$-2 \left[\ln \hat{L}_R - \ln \hat{L}_U \right]$$

where \hat{L}_R and \hat{L}_U are the likelihood evaluated at $\hat{\boldsymbol{\theta}}_R$ (the restricted parameter vector) and $\hat{\boldsymbol{\theta}}_U$ (the unrestricted parameter vector).

$$-2 \left[\ln \hat{L}_R - \ln \hat{L}_U \right] \sim \chi_k^2$$

where k is the number of restrictions imposed.

- For the null that the slope coefficients are equal to zero

$$\ln \hat{L}_R = \ln \hat{L}_0 = n_0 \ln \left(\frac{n_0}{n} \right) + n_1 \ln \left(\frac{n_1}{n} \right)$$

where n is the sample size, n_0 is the number of observations where $y = 0$ and n_1 is the number of observations where $y = 1$

- Some packages will report a pseudo R^2 statistic, which has a similar “goodness of fit” interpretation as the R^2 statistic in the linear regression context.

For example, the statistic

$$1 - \frac{\ln \hat{L}_U}{\ln \hat{L}_0}$$

ranges between 0 and 1 as does R^2 in the linear regression model.

- Most stats packages will also report some measure of how well the model does at predicting the occurrence of an event, such as the percentage of observations which are “correctly classified”.
- Herron recommends using *expected percent correctly predicted* (ePCP), which surmounts some of the problems of using just percent correctly predicted.

$$\text{ePCP} = \frac{1}{N} \left(\sum_{y_i=1}^N \hat{p}_i + \sum_{y_i=0}^N (1 - \hat{p}_i) \right) \quad (12.1)$$

12.3 Voting Behavior Example

- Consider the following model of vote choice in congressional elections

$$y_i = \beta' \mathbf{x}_i + u_i$$

where

x_{2i} = party identification match incumbent's dummy

x_{3i} = national economic performance, retrospective evaluation

x_{4i} = personal financial situation, retrospective evaluation

x_{5i} = recall incumbent's name

x_{6i} = recall challenger's name

x_{7i} = quality of challenger

- Table 12.1 reports the results of probit estimation of this model using data from the American National Election Survey (1982–1988).
- Table 12.2 reports marginal effects using derivatives.
- We can also report marginal effects by simulating probabilities of observing an event. For example, we first compute the probability of voting for an incumbent given that she does not face a quality challenger. For the probit estimates, this is

$$\begin{aligned} \Pr(y_i = 1 | x_{7i} = 0) &= \Phi(\beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + \beta_7 x_{7i}) \\ &= \Phi(.184 + 1.355 \times 1 - .114 \times .5 + .095 \times .5 \\ &\quad + .324 \times 0 - .677 \times 0 - .339 \times 0) \\ &= .936 \end{aligned}$$

Then compute this probability given a quality challenger:

$$\Pr(y_i = 1 | x_{7i} = 1) = .881$$

Then the difference between these two gives the percentage increase in the probability of voting for the incumbent when she does not face a quality challenger (i.e., $.936 - .881 = .055$).

Table 12.1: Probability of Voting for the Incumbent Member of Congress

variable	Probit MLEs
Intercept	.184 (.058)
Party identification	1.35 (.056)
National economic performance (Retrospective Judgment)	-.114 (.069)
Personal financial situation (Retrospective Judgment)	.095 (.068)
Recall incumbent's name	.324 (.0808)
Recall challenger's name	-.677 (.109)
Quality of challenger	-.339 (.073)

Notes: Standard errors in parentheses.
 $N = 3341$. $-2 \ln L = 760.629$ Percent correctly predicted = 78.5%

Table 12.2: Marginal Effects on Probability of Voting for the Incumbent Member of Congress

variable	$\hat{\beta}_j \phi(\hat{\beta}' \mathbf{x}_i)$
Party identification	.251
National economic performance (Retrospective Judgment)	-.021
Personal financial situation (Retrospective Judgment)	.018
Recall incumbent's name	.060
Recall challenger's name	-.126
Quality of challenger	-.063

Notes: Explanatory variables are set equal to their medians in the sample.

12.4 Obstruction and Passage of Legislation Example

We model the passage of legislation ($y_i = 1$ if passed, $= 0$ if failed) as a function of coalition sizes opposing obstruction of legislation in the U.S. Senate.

- Table 12.3 reports probit estimation results.
- Table 12.4 reports simulated probabilities (with confidence intervals) for specific values of the coalition size variable.
- Figure 12.1 plots simulated probabilities (with confidence intervals) across a range of values for the coalition size variable.

Table 12.3: Probit analysis of passage of obstructed measures, 1st–64th Congresses

Variable	Coefficient	Std. Err.
Constant	−1.671	0.962
Coalition size	6.155	2.224
Coalition size × end of session	−1.944	0.690
Likelihood ratio test	12.84	($p = 0.002$)
% correctly predicted	72	

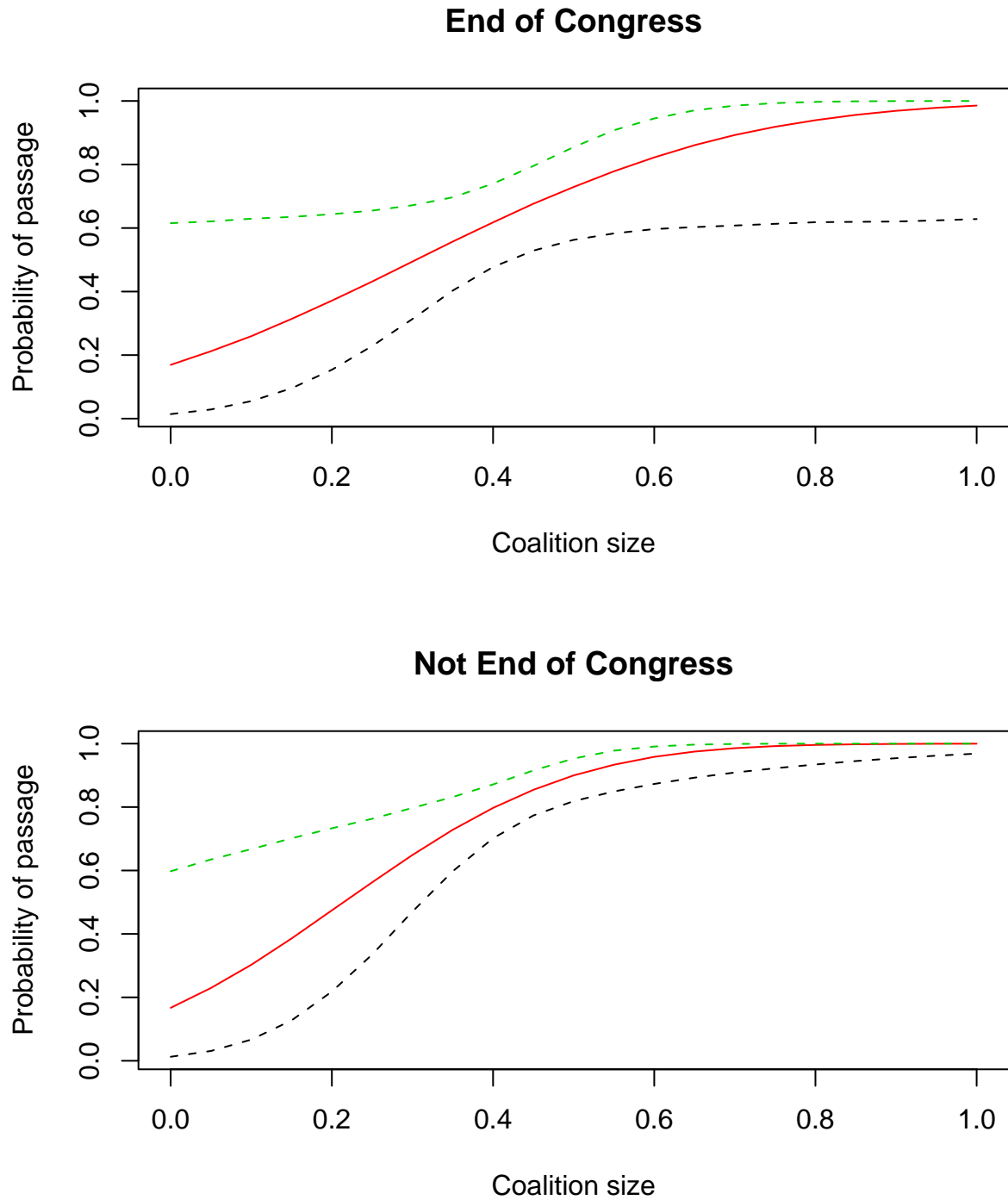
Note: $N = 114$.

Table 12.4: Simulated Probabilities of Passage and Coalition Sizes, 1st–64th Congresses

Coalition size (%)	Probability of Passage	
	Not end of session	End of session
51	.92 (.84, .97)	.67 (.47, .83)
60	.99 (.88, .99)	.80 (.54, .95)
67	.99 (.91, .99)	.87 (.54, .98)

Note: 95% confidence intervals in parentheses

Figure 12.1: Simulated Probabilities of Passage and Coalition Size



12.5 Heteroskedasticity and the Probit Model

- In the probit model heteroskedastic disturbances render the MLEs inconsistent and the variance-covariance matrix inappropriate.
- A general form of heteroskedasticity is

$$\sigma_i^2 = \exp(\boldsymbol{\gamma}'\mathbf{z}_i)$$

where $\mathbf{z}_i = [1, \mathbf{q}_i]$ and $\boldsymbol{\gamma} = [\ln \sigma^2, \boldsymbol{\alpha}]$.

- The log likelihood for this heteroskedastic probit model is

$$\ln L = \sum_{i=1}^n \left\{ y_i \ln \Phi \left(\frac{\boldsymbol{\beta}'\mathbf{x}_i}{\exp(\boldsymbol{\gamma}'\mathbf{z}_i)} \right) + (1 - y_i) \ln \left[1 - \Phi \left(\frac{\boldsymbol{\beta}'\mathbf{x}_i}{\exp(\boldsymbol{\gamma}'\mathbf{z}_i)} \right) \right] \right\} \quad (12.2)$$

- Note: \mathbf{q}_i should not contain a constant term.
- To test for heteroskedasticity we can use a LR test, where $\ln \hat{L}_u$ is given by (12.2) and for $\ln \hat{L}_R$ we restrict $\boldsymbol{\gamma} = \mathbf{0}$.
- A statistically significant χ^2 statistic \Rightarrow heteroskedasticity is present.

- Application: Alvarez and Brehm, “American Ambivalence Towards Abortion Policy: Development of a Heteroskedastic Probit Model of Competing Values.” ’95 *AJPS*.
- Individuals possess a variance of attitudes—variance stems from difficulty individuals have in resolving conflict between “core” beliefs.
- Some individuals will have resolved such conflict to a greater degree than others—i.e., they will be less ambivalent.
- ambivalence = f (motivation and ability to process information, the degree that the respondents evaluate the policy choice under both pro-life and pro-choice dimensions)
- Process that generates responses to questions about abortion policy is heterogeneous \Rightarrow iid assumption may be incorrect.
- Seek to model the process that produces ambivalence by using a heteroskedastic probit—use model given above with:
 - Dependent variable: response to question about whether it should be legal or not for a woman in a particular situation to obtain an abortion.
 - \mathbf{x}_i : race, gender, religious affiliation (Catholic or not), religious intensity, church attendance, knowledge and support of ERA.
 - \mathbf{z}_i : firmness of opinion, importance of issue, knowledge of issue, pro and con count interaction term

Table 1. Heteroskedastic Probit Estimates of Attitudes Toward Abortion Policy, 1982 General Social Survey

	Mothers' Health	Rape	Birth Defect	Too Poor	No More Children	Single	Any Reason
Percent Yes	90.4	83.9	82.1	49.0	45.8	45.5	38.5
Choice Model							
Constant	2.55 (.46)	1.92 (.40)	2.02 (.40)	.02 (.01)	.03 (.08)	.11 (.09)	-.07 (.13)
Black	-.51 (.14)	-.47 (.13)	-.54 (.15)	-.09 (.06)	-.11 (.06)	-.23 (.10)	-.15 (.09)
Male	-.08 (.11)	-.20 (.09)	-.21 (.11)	-.04 (.04)	-.02 (.03)	-.06 (.05)	-.13 (.07)
Catholic	-.52 (.13)	-.15 (.10)	-.33 (.12)	.01 (.04)	.02 (.04)	-.03 (.04)	.05 (.07)
Religious Intensity	-.39 (.20)	-.17 (.14)	-.51 (.19)	-.17 (.10)	-.13 (.69)	-.18 (.09)	-.22 (.12)
Attend Church	-1.04 (.25)	-.99 (.23)	-.91 (.24)	-.35 (.17)	-.43 (.17)	-.47 (.20)	-.79 (.26)
Know What ERA Means	-.18 (.17)	-.14 (.15)	.01 (.16)	.10 (.08)	.09 (.07)	.09 (.08)	.12 (.10)
Support ERA	.33 (.17)	.12 (.14)	.40 (.18)	.22 (.12)	.31 (.13)	.31 (.13)	.51 (.17)
Variance Model							
Pro Count	-.14 (.07)	-.19 (.09)	-.06 (.08)	-.25 (.22)	-.26 (.18)	-.34 (.17)	-.22 (.15)
Con Count	.17 (.09)	.20 (.12)	.37 (.12)	-.50 (.19)	-.58 (.17)	-.41 (.16)	-.48 (.14)
Pro Count × Con Count	-.44 (.04)	-.03 (.05)	-.09 (.05)	.19 (.11)	.25 (.09)	.21 (.08)	.22 (.08)
Importance	.51 (.15)	.17 (.15)	-.14 (.16)	-.16 (.31)	-.18 (.26)	-.24 (.25)	-.30 (.25)
Information	.37 (.13)	-.13 (.14)	.05 (.14)	-.32 (.29)	-.28 (.25)	-.28 (.24)	.68 (.23)
Firmness of Opinion	-.37 (.16)	-.58 (.17)	-.61 (.16)	.60 (.58)	.47 (.43)	1.81 (.67)	.63 (.38)
Heteroskedasticity Test							
Likelihood Ratio Test ($\chi^2_{df=1}$)	47.4†	46.7†	41.2†	12.5	19.9†	27.2†	25.9†
N	1312	1302	1294	1291	1289	1293	1295
Goodness of Fit ($\chi^2_{df=13}$)	126.12†	173.66†	181.29†	142.30†	182.86†	193.54†	180.86†

Note: Standard errors are in parentheses below coefficients. † indicates a χ^2 significant at the $p \leq .05$ level.

Section 13

Ordered Probit

- Observation Scheme #3: Suppose dependent variable can take on three ordered values.

$$y_i = \begin{cases} 1 & \text{if } y_i^* < \gamma_1 \\ 2 & \text{if } \gamma_1 \leq y_i^* < \gamma_2 \\ 3 & \text{if } \gamma_2 \leq y_i^* \end{cases}$$

- Remember our latent model is

$$y_i^* = \boldsymbol{\beta}'\mathbf{x}_i + u_i$$

where $u_i \sim N[0, \sigma^2]$.

- In addition to estimating $\boldsymbol{\beta}$, we also need to estimate the γ parameters.
- The γ s are cutpoints that indicate what discrete category the latent variable falls into.
- As with the dichotomous probit, we normalize $\sigma = 1$. Then the probabilities of observing the different values for y_i are as follows.

$$\begin{aligned} \Pr(y_i = 1) &= \Pr(y_i^* < \gamma_1) \\ &= \Pr(\boldsymbol{\beta}'\mathbf{x}_i + u_i < \gamma_1) \\ &= \Pr(u_i < \gamma_1 - \boldsymbol{\beta}'\mathbf{x}_i) \\ &= \Phi(\gamma_1 - \boldsymbol{\beta}'\mathbf{x}_i) \end{aligned}$$

$$\begin{aligned}
\Pr(y_i = 2) &= \Pr(\gamma_1 \leq y_i^* < \gamma_2) \\
&= \Pr(\gamma_1 \leq \boldsymbol{\beta}'\mathbf{x}_i + u_i < \gamma_2) \\
&= \Pr(u_i < \gamma_2 - \boldsymbol{\beta}'\mathbf{x}_i) - \Pr(u_i \leq \gamma_1 - \boldsymbol{\beta}'\mathbf{x}_i) \\
&= \Phi(\gamma_2 - \boldsymbol{\beta}'\mathbf{x}_i) - \Phi(\gamma_1 - \boldsymbol{\beta}'\mathbf{x}_i)
\end{aligned}$$

$$\begin{aligned}
\Pr(y_i = 3) &= \Pr(y_i^* \geq \gamma_2) \\
&= \Pr(\boldsymbol{\beta}'\mathbf{x}_i + u_i \geq \gamma_2) \\
&= \Pr(u_i \geq \gamma_2 - \boldsymbol{\beta}'\mathbf{x}_i) \\
&= 1 - \Pr(u_i \leq \gamma_2 - \boldsymbol{\beta}'\mathbf{x}_i) \\
&= 1 - \Phi(\gamma_2 - \boldsymbol{\beta}'\mathbf{x}_i)
\end{aligned}$$

- Note that $\gamma_1 < \gamma_2$ in order for the probabilities to be positive.
- The likelihood for an individual is

$$L_i = [\Phi(\gamma_1 - \boldsymbol{\beta}'\mathbf{x}_i)]^{z_{i1}} [\Phi(\gamma_2 - \boldsymbol{\beta}'\mathbf{x}_i) - \Phi(\gamma_1 - \boldsymbol{\beta}'\mathbf{x}_i)]^{z_{i2}} [\Phi(\boldsymbol{\beta}'\mathbf{x}_i - \gamma_2)]^{z_{i3}}$$

where

$$z_{ij} = \begin{cases} 1 & \text{if } y_i = j \\ 0 & \text{otherwise} \end{cases} \quad \text{for } j = 1, 2, 3.$$

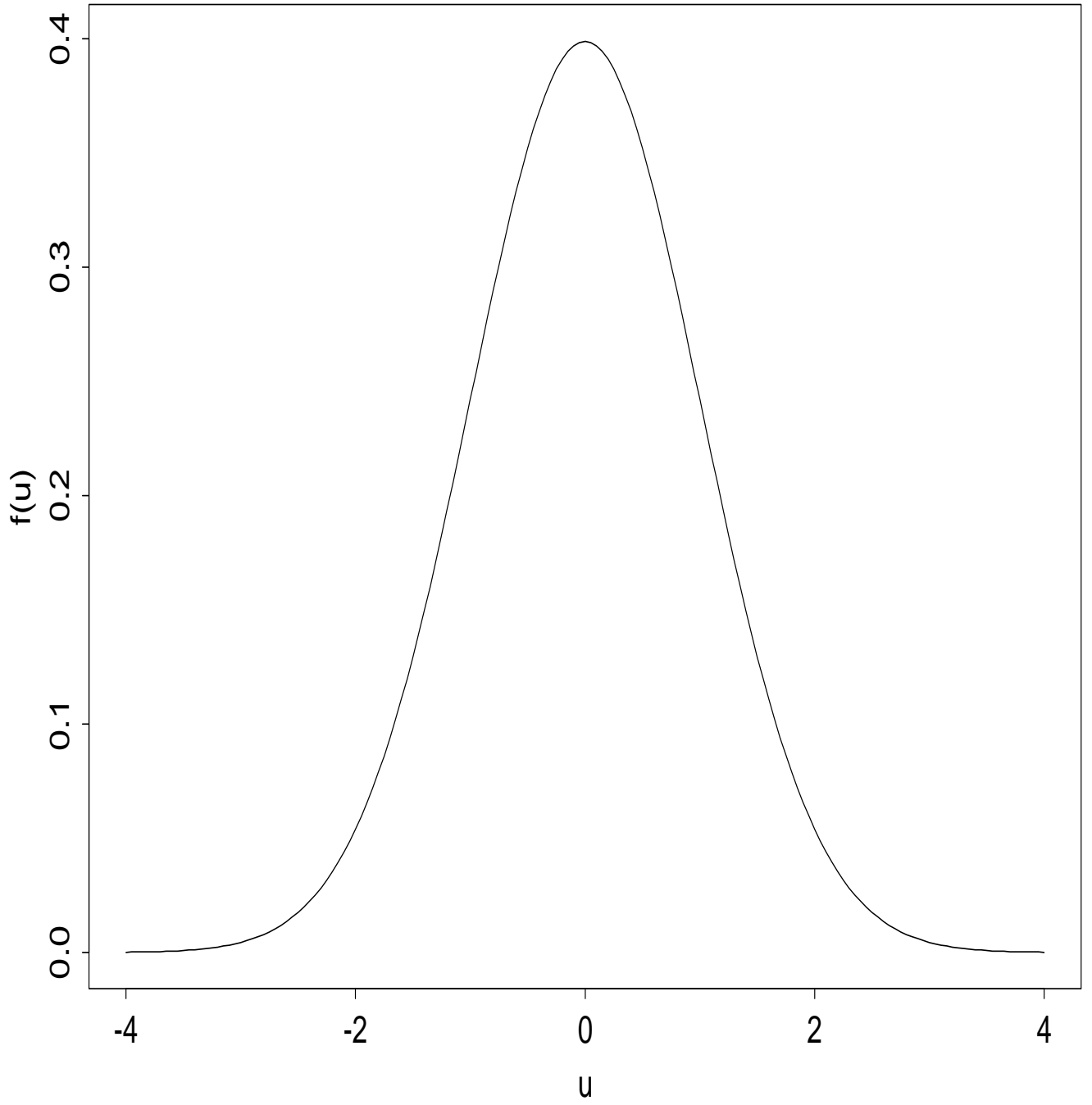


Figure 13.1: Probabilities for categories

- We can generalize this to m categories

$$y_i = \begin{cases} 1 & \text{if } y_i^* < \gamma_1 \\ 2 & \text{if } \gamma_1 < y_i^* \leq \gamma_2 \\ 3 & \text{if } \gamma_2 < y_i^* \leq \gamma_3 \\ \vdots & \\ m & \text{if } \gamma_{m-1} < y_i^* \end{cases}$$

- This gives the probabilities

$$\begin{aligned} \Pr(y_i = 1) &= \Phi(\gamma_1 - \boldsymbol{\beta}'\mathbf{x}_i) \\ \Pr(y_i = 2) &= \Phi(\gamma_2 - \boldsymbol{\beta}'\mathbf{x}_i) - \Phi(\gamma_1 - \boldsymbol{\beta}'\mathbf{x}_i) \\ \Pr(y_i = 3) &= \Phi(\gamma_3 - \boldsymbol{\beta}'\mathbf{x}_i) - \Phi(\gamma_2 - \boldsymbol{\beta}'\mathbf{x}_i) \\ &\vdots \\ \Pr(y_i = m) &= 1 - \Phi(\gamma_{m-1} - \boldsymbol{\beta}'\mathbf{x}_i) \\ &= \Phi(\boldsymbol{\beta}'\mathbf{x}_i - \gamma_{m-1}) \end{aligned}$$

- To write down the likelihood let

$$z_{ij} = \begin{cases} 1 & \text{if } y_i = j \\ 0 & \text{otherwise} \end{cases} \quad \text{for } j = 1, \dots, m.$$

- Then

$$\Pr(z_{ij} = 1) = \Phi(\gamma_j - \boldsymbol{\beta}'\mathbf{x}_i) - \Phi(\gamma_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i)$$

- The likelihood for an individual is

$$\begin{aligned} L_i &= [\Phi(\gamma_1 - \boldsymbol{\beta}'\mathbf{x}_i)]^{z_{i1}} [\Phi(\gamma_2 - \boldsymbol{\beta}'\mathbf{x}_i) - \Phi(\gamma_1 - \boldsymbol{\beta}'\mathbf{x}_i)]^{z_{i2}} \\ &\quad \cdots [1 - \Phi(\gamma_{m-1} - \boldsymbol{\beta}'\mathbf{x}_i)]^{z_{im}} \\ &= \prod_{j=1}^m [\Phi(\gamma_j - \boldsymbol{\beta}'\mathbf{x}_i) - \Phi(\gamma_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i)]^{z_{ij}} \end{aligned}$$

- The likelihood function for the sample then is

$$L = \prod_{i=1}^n \prod_{j=1}^m [\Phi(\gamma_j - \boldsymbol{\beta}'\mathbf{x}_i) - \Phi(\gamma_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i)]^{z_{ij}}$$

- We compute marginal effects just like for the dichotomous probit model:

$$\begin{aligned} \frac{\partial \Pr(y_i = 1)}{\partial \mathbf{x}_i} &= -\phi(\gamma_1 - \boldsymbol{\beta}'\mathbf{x}_i)\boldsymbol{\beta} \\ \frac{\partial \Pr(y_i = 2)}{\partial \mathbf{x}_i} &= (\phi(\gamma_1 - \boldsymbol{\beta}'\mathbf{x}_i) - \phi(\gamma_2 - \boldsymbol{\beta}'\mathbf{x}_i))\boldsymbol{\beta} \\ \frac{\partial \Pr(y_i = 3)}{\partial \mathbf{x}_i} &= \phi(\gamma_2 - \boldsymbol{\beta}'\mathbf{x}_i)\boldsymbol{\beta} \end{aligned}$$

- Applications:

- Fiorina, *Retrospective Voting in American National Elections*: What factors affect the strength of party identification?
- Krehbiel and Rivers '88 *AJPS*: What factors determine a legislator's votes on paired legislative proposals?

Section 14

Censored Regression

- Observation Scheme #4:

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

- Case 1: $y_i = 0 \Rightarrow u_i < -\boldsymbol{\beta}'\mathbf{x}_i$.

$$\Pr(u_i < -\boldsymbol{\beta}'\mathbf{x}_i) = \Pr\left(\frac{u_i}{\sigma} > \frac{\boldsymbol{\beta}'\mathbf{x}_i}{\sigma}\right) = 1 - \Pr\left(\frac{u_i}{\sigma} < \frac{\boldsymbol{\beta}'\mathbf{x}_i}{\sigma}\right)$$

– Information content: $1 - \Phi\left(\frac{\boldsymbol{\beta}'\mathbf{x}_i}{\sigma}\right)$.

- Case 2: $y_i > 0$.

– Information content: $f(y_i|y_i^* > 0) \cdot \Pr(y_i^* > 0)$

– What is $f(y_i|y_i^* > 0)$?

– By the definition of conditional probability we have

$$f(y_i|y_i^* > 0) = \frac{f(y_i)}{\Pr(y_i^* > 0)}.$$

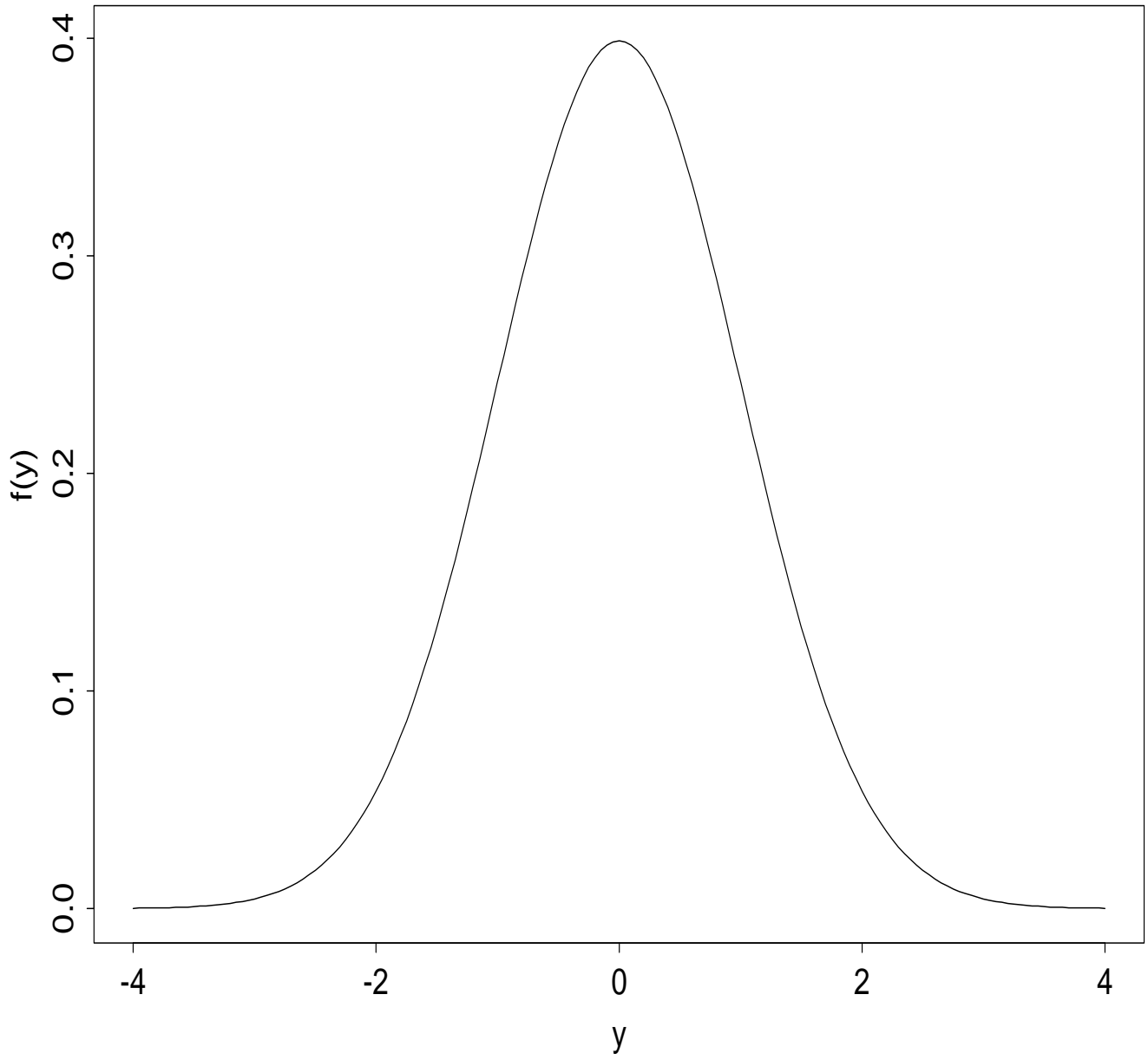


Figure 14.1: Censoring

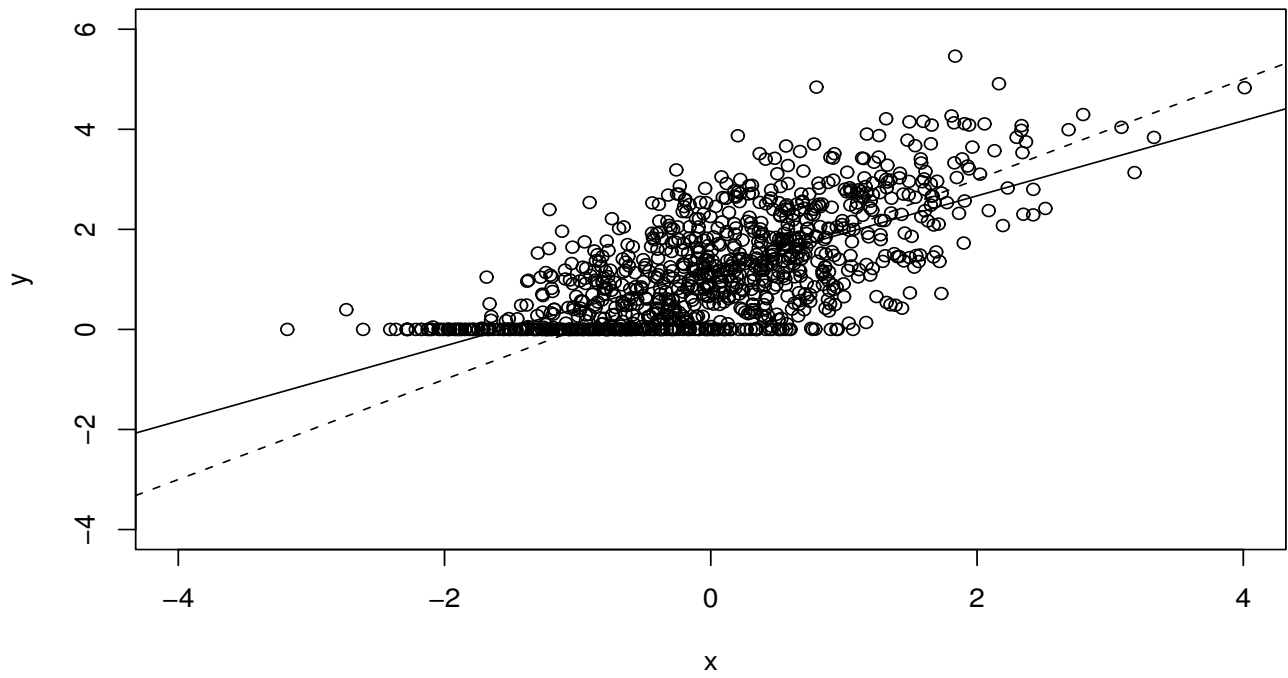
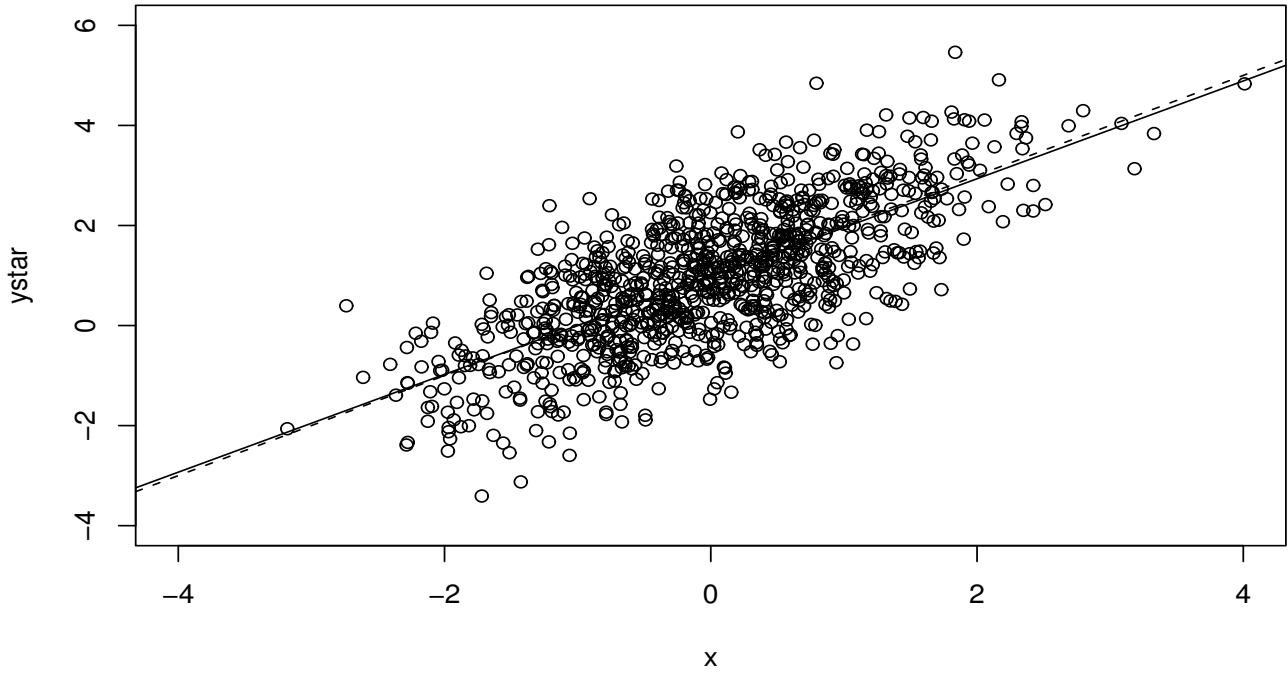


Figure 14.2: Censoring and OLS

– Given our assumptions about u_i we have

$$\begin{aligned} f(y_i) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(y_i - \boldsymbol{\beta}'\mathbf{x}_i)^2}{\sigma^2}\right) \\ &= \frac{1}{\sigma} \phi\left(\frac{y_i - \boldsymbol{\beta}'\mathbf{x}_i}{\sigma}\right) \end{aligned}$$

and

$$\begin{aligned} \Pr(y_i^* > 0) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\boldsymbol{\beta}'\mathbf{x}_i/\sigma} \exp\left(-\frac{u^2}{2}\right) du \\ &= \Phi\left(\frac{\boldsymbol{\beta}'\mathbf{x}_i}{\sigma}\right) \end{aligned}$$

– So for Case 2 the information content is

$$\begin{aligned} f(y_i|y_i^* > 0) \cdot \Pr(y_i^* > 0) &= \frac{f(y_i)}{\Pr(y_i^* > 0)} \cdot \Pr(y_i^* > 0) \\ &= f(y_i) \\ &= \frac{1}{\sigma} \phi\left(\frac{y_i - \boldsymbol{\beta}'\mathbf{x}_i}{\sigma}\right) \end{aligned}$$

• The likelihood function for an individual then is

$$L_i = \left[\frac{1}{\sigma} \phi\left(\frac{y_i - \boldsymbol{\beta}'\mathbf{x}_i}{\sigma}\right) \right]^{d_i} \left[1 - \Phi\left(\frac{\boldsymbol{\beta}'\mathbf{x}_i}{\sigma}\right) \right]^{1-d_i}$$

where

$$d_i = \begin{cases} 1 & \text{if } y > 0 \\ 0 & \text{otherwise} \end{cases}$$

- We can write the log likelihood for this sample as follows:

$$\ln L = \sum_{y_i > 0} -\frac{1}{2} \left[\ln(2\pi) + \ln \sigma^2 + \frac{(y_i - \boldsymbol{\beta}' \mathbf{x}_i)^2}{\sigma^2} \right] \\ + \sum_{y_i = 0} \ln \left[1 - \Phi \left(\frac{\boldsymbol{\beta}' \mathbf{x}_i}{\sigma} \right) \right]$$

14.1 Reparameterization for tobit model

- Let $\boldsymbol{\gamma} = \boldsymbol{\beta}/\sigma$ and $\theta = 1/\sigma$. The likelihood then becomes

$$\ln L = \sum_{y_i > 0} -\frac{1}{2} [\ln(2\pi) - \ln \theta^2 + (\theta y_i - \boldsymbol{\gamma}' \mathbf{x}_i)^2] \\ + \sum_{y_i = 0} \ln [1 - \Phi(\boldsymbol{\gamma}' \mathbf{x}_i)]$$

- After estimating this model we can get the parameter estimates of interest back from $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\gamma}}/\hat{\theta}$ and $\hat{\sigma} = 1/\hat{\theta}$.
- We also need the standard errors of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$, but we can't just get them in the standard way.
- The method that is used is called the **delta method**; use it to obtain the asymptotic variance-covariance matrix.
- The delta method is used to obtain the asymptotic variance-covariance matrix of a function of a random variable by employing the Taylor approximation.

- For example, suppose we have a sample y_1, \dots, y_n where the y s are iid with $E[y_i] = \mu$ and $\text{var}[y_i] = \sigma^2$.
- Let $\bar{y} = \frac{1}{n} \sum y_i$.
- Recall the CLT says: $\sqrt{n}(\bar{y} - \mu) \xrightarrow{D} N[0, \sigma^2]$.

- From the Taylor approximation:

$$g(\bar{y}) \approx g(\mu) + g'(\mu)(\bar{y} - \mu) + \text{higher order terms}$$

- Rearranging and setting to equal gives

$$g(\bar{y}) - g(\mu) = g'(\mu)(\bar{y} - \mu)$$

- Multiply through by \sqrt{n} to get

$$\sqrt{n}(g(\bar{y}) - g(\mu)) = g'(\mu)\sqrt{n}(\bar{y} - \mu)$$

- By the CLT we know

$$g'(\mu)\sqrt{n}(\bar{y} - \mu) \xrightarrow{D} g'(\mu) \cdot N(0, \sigma^2) \stackrel{D}{=} N(0, (g'(\mu))^2 \sigma^2)$$

- So

$$\sqrt{n}(g(\bar{y}) - g(\mu)) \xrightarrow{D} N(0, (g'(\mu))^2 \sigma^2)$$

- In higher dimensions, this will look like

$$\sqrt{n}(\bar{\mathbf{y}} - \boldsymbol{\mu}) \xrightarrow{D} N[\mathbf{0}, \boldsymbol{\Sigma}]$$

- So

$$\sqrt{n}(\mathbf{g}(\bar{\mathbf{y}}) - \mathbf{g}(\boldsymbol{\mu})) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Omega})$$

where

$$\boldsymbol{\Omega} = \frac{\partial \mathbf{g}(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} \boldsymbol{\Sigma} \frac{\partial \mathbf{g}(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}'}$$

where $\mathbf{g}(\boldsymbol{\mu})$ is a set of p continuous functions.

- Using this method for the reparameterized tobit model will give

$$\text{Asy. var} \left[\hat{\boldsymbol{\beta}}, \hat{\sigma}^2 \right] = \mathbf{J} \mathbf{V} \mathbf{J}'$$

where \mathbf{V} is the negative inverse of the estimated Hessian and

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \beta}{\partial \gamma'} & \frac{\partial \beta}{\partial \theta} \\ \frac{\partial \sigma}{\partial \gamma'} & \frac{\partial \sigma}{\partial \theta} \end{bmatrix} = \begin{bmatrix} \frac{1}{\theta} \mathbf{I} & \frac{-1}{\theta^2} \boldsymbol{\gamma} \\ \mathbf{0}' & \frac{-1}{\theta^2} \end{bmatrix}$$

14.2 Marginal Effects

We can compute marginal effects by taking derivatives—but what we take the derivative of depends on the question of interest. Some possibilities are:

- If interested in the marginal effect on the latent variable y^* do:

$$\frac{\partial E[y_i^*|\mathbf{x}_i]}{\partial \mathbf{x}_i} = \boldsymbol{\beta}$$

- If interested in the marginal effect on y , taking into account the censoring:

$$\begin{aligned} E[y_i|\mathbf{x}_i] &= 0 \cdot \Pr(y_i = 0|\mathbf{x}_i) + E[y_i|y_i > 0, \mathbf{x}_i] \cdot \Pr(y_i > 0|\mathbf{x}_i) \\ &= (1 - \Phi(-\boldsymbol{\beta}'\mathbf{x}_i/\sigma)) (\boldsymbol{\beta}'\mathbf{x}_i + E[u_i|u_i > -\boldsymbol{\beta}'\mathbf{x}_i]) \\ &= (1 - \Phi(-\boldsymbol{\beta}'\mathbf{x}_i/\sigma)) \left(\boldsymbol{\beta}'\mathbf{x}_i + \sigma \frac{\phi(\boldsymbol{\beta}'\mathbf{x}_i/\sigma)}{\Phi(\boldsymbol{\beta}'\mathbf{x}_i/\sigma)} \right) \end{aligned}$$

Taking the derivative gives

$$\frac{\partial E[y_i|\mathbf{x}_i]}{\partial \mathbf{x}_i} = \boldsymbol{\beta} \Phi \left(\frac{\boldsymbol{\beta}'\mathbf{x}_i}{\sigma} \right)$$

- If interested in the marginal effect on probability that y exceeds threshold:

$$\frac{\partial \Pr[y_i > 0|\mathbf{x}_i]}{\partial \mathbf{x}_i} = \boldsymbol{\beta} \phi \left(\frac{\boldsymbol{\beta}'\mathbf{x}_i}{\sigma} \right)$$

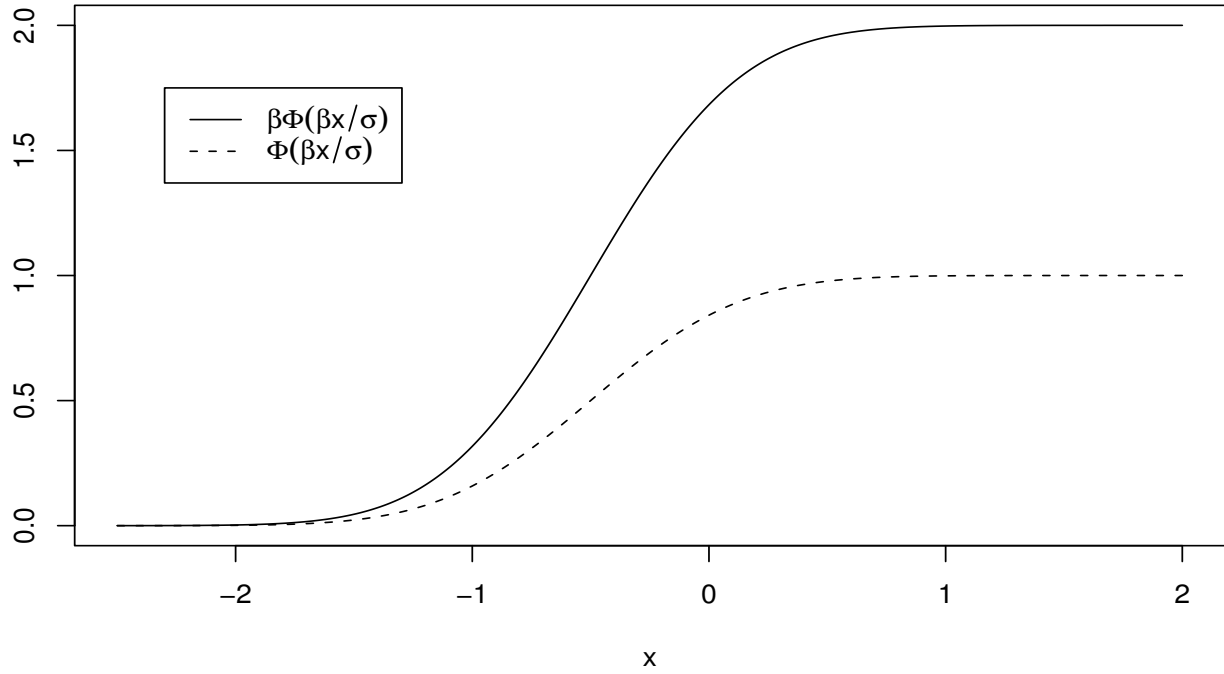
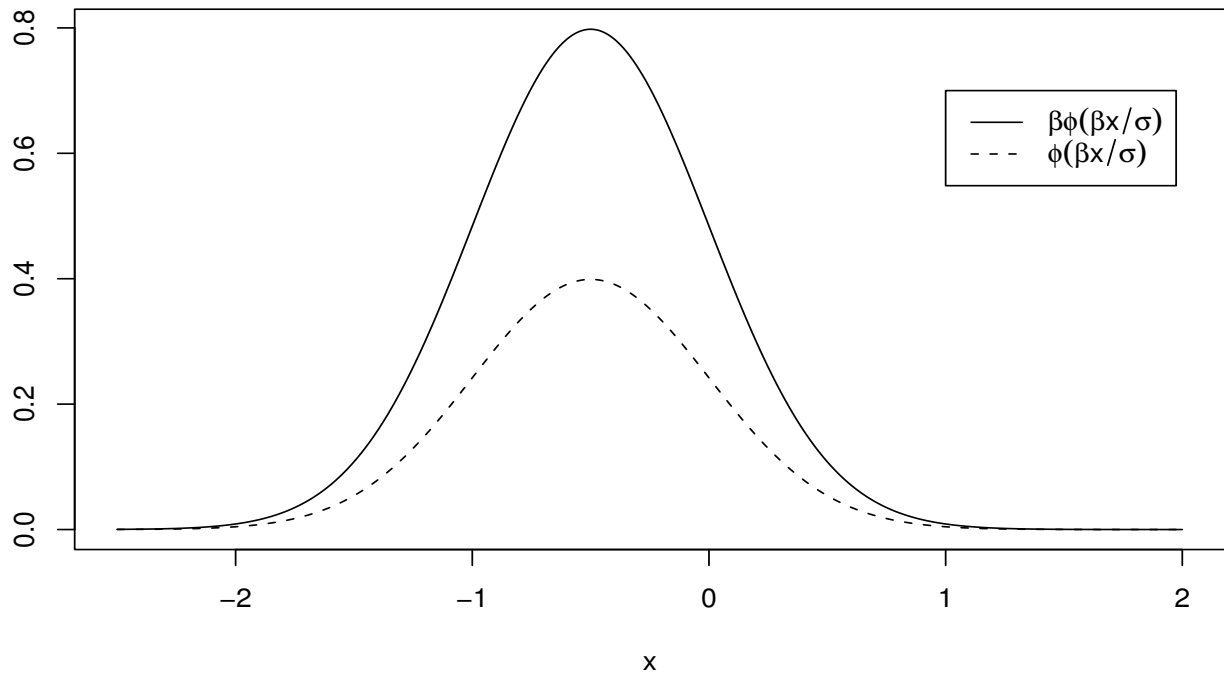
Marginal effect on y , taking into account the censoring**Marginal effect on probability that y exceeds threshold**

Figure 14.3: Marginal Effects and Censoring

14.3 Heteroskedasticity and the tobit model

- In the presence of heteroskedasticity, the tobit MLE is inconsistent, with the severity of the inconsistency dependent in part on the number of limit observations.
- Heteroskedasticity can be handled in the same way as with the probit model—i.e., specify the form of the heteroskedasticity such as

$$\sigma_i^2 = \sigma^2 \exp(\boldsymbol{\psi}' \mathbf{w}_i)$$

and then estimate the model. Use LR tests to determine whether heteroskedastic model is correct.

Section 15

Truncated Regression

- Observation Scheme #5:

$$y_i = \begin{cases} y^* & \text{if } y_i^* > 0 \\ ? & \text{if } y_i^* \leq 0 \end{cases}$$

- Case 1: $y_i > 0$.

– Information content:

$$f(y_i | y_i^* > 0) = \frac{\frac{1}{\sigma} \phi\left(\frac{y_i - \beta' \mathbf{x}_i}{\sigma}\right)}{\Pr(y_i^* > 0)}$$

- Case 2: $y_i = ?$

– Information content: Nada.

- The individual's contribution to the likelihood function then is

$$L_i = \frac{\frac{1}{\sigma} \phi\left(\frac{y_i - \beta' \mathbf{x}_i}{\sigma}\right)}{1 - \Phi(-\beta' \mathbf{x}_i / \sigma)}$$

- The log-likelihood then is

$$\begin{aligned} \ln L = & -\frac{n}{2} [\ln(2\pi) + \ln \sigma^2] - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta' \mathbf{x}_i)^2 \\ & - \sum_{i=1}^n \ln \left[1 - \Phi\left(\frac{-\beta' \mathbf{x}_i}{\sigma}\right) \right] \end{aligned}$$

- The same reparameterization that was recommended for the tobit model is also recommended here.

15.1 Marginal Effects

- From the moments of the truncated normal we know:

$$\begin{aligned} E[y_i | y_i > 0, \mathbf{x}_i] &= \left(\boldsymbol{\beta}' \mathbf{x}_i + \sigma \frac{\phi(-\boldsymbol{\beta}' \mathbf{x}_i / \sigma)}{1 - \Phi(-\boldsymbol{\beta}' \mathbf{x}_i / \sigma)} \right) \\ &= \boldsymbol{\beta}' \mathbf{x}_i + \sigma \lambda(\alpha_i) \end{aligned}$$

where $\alpha_i = \boldsymbol{\beta}' \mathbf{x}_i / \sigma$ and $\lambda(\cdot) = \phi(\cdot) / [1 - \Phi(\cdot)]$.

$$\begin{aligned} \frac{\partial E[y_i | y_i > 0, \mathbf{x}_i]}{\partial \mathbf{x}_i} &= \boldsymbol{\beta} + \left(\frac{d\lambda_i}{d\alpha_i} \right) \frac{\partial \alpha_i}{\partial \mathbf{x}_i} \\ &= \boldsymbol{\beta} + \sigma \left(\lambda_i^2 - \alpha_i \lambda_i \left(\frac{-\mathbf{x}_i}{\sigma} \right) \right) \\ &= \boldsymbol{\beta} (1 - \lambda_i^2 + \alpha_i \lambda_i) \\ &= \boldsymbol{\beta} (1 - \delta(\alpha_i)) \end{aligned}$$

- Note that $0 < \delta(\alpha_i) < 1$ which implies that truncation leads to an attenuation of the marginal effect that we would see in a linear regression.

Section 16

Sample and Self-Selection Models

- Consider the following research questions:
 1. How do certain firm characteristics affect the dollar amount of fines the firm pays for violating environmental regulations?
 2. What factors affect the vote share of incumbents, taking into account that they have to decide whether or not to run for reelection?
 3. What factors affect the number of troops that a country commits to a U.N. peacekeeping mission?
- While these questions look similar to the questions for which we would use truncated or censored regression models, they differ in a key way.
- While the dependent variable in each of these cases is continuous, we will only get to observe this dependent variable for certain outcomes of a stochastic process.
 - The kind of sample we get is determined by a stochastic **selection** process.
- Another way to think about this is that the truncation points here may be more appropriately thought of as random rather than constant.

- There is some prior decision that we need to account for:
 - In Case 1, our sample is determined by an agency's decisions to investigate firms.
 - In Case 2, our sample is determined by members' decisions to seek reelection or retire.
 - In Case 3, our sample is determined by countries' decisions to participate in the peacekeeping mission.

- These are examples where **sample selection models** would be appropriate.

- To answer these questions, we use an empirical model that consists of more than one equation.

- A Model:

Latent structure:

$$\begin{aligned}y_{1i}^* &= \boldsymbol{\beta}'_1 \mathbf{x}_{1i} + u_{1i} \\ y_{2i}^* &= \boldsymbol{\beta}'_2 \mathbf{x}_{2i} + u_{2i}\end{aligned}$$

Observed structure:

$$y_{1i} = \begin{cases} 1 & \text{if } y_{1i}^* > 0 \\ 0 & \text{if } y_{1i}^* \leq 0 \end{cases} \quad y_{2i} = \begin{cases} y_{2i}^* & \text{if } y_{1i} = 1 \\ ? & \text{otherwise} \end{cases}$$

- Assume distribution for u s:

$$\begin{bmatrix} u_{1i} \\ u_{2i} \end{bmatrix} \sim N \left(\mathbf{0}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right)$$

- That is, we assume that (u_{1i}, u_{2i}) have a bivariate normal distribution:

$$\begin{aligned}f(u_{1i}, u_{2i}) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\ &\times \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{u_{1i}}{\sigma_1} \right)^2 - 2\rho \frac{u_{1i}u_{2i}}{\sigma_1\sigma_2} + \left(\frac{u_{2i}}{\sigma_2} \right)^2 \right] \right\}\end{aligned}$$

where $\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}$.

- What is the likelihood function for this model?

Case 1: $y_{1i} = 0$, so $y_{2i} = ?$.

Individual's contribution to the likelihood is
 $\Pr(y_{1i} = 0 | \mathbf{x}_{1i}) = 1 - \Phi(\boldsymbol{\beta}'_1 \mathbf{x}_{1i} / \sigma_1)$.

Case 2: $y_{1i} = 1$ and we get to see y_{2i} .

To figure out what an individual's contribution to the likelihood is in this case will be we need to think about the distribution of u_{2i} given $y_{1i} = 1$:

$$\int_c^\infty f(u_{1i}, u_{2i}) du_{1i} \quad (16.1)$$

where c is the threshold u_{1i} must exceed to observe y_{2i} (i.e., $c = -\boldsymbol{\beta}'_1 \mathbf{x}_{1i}$).

Trick: a joint distribution can be written as a conditional distribution times a marginal distribution:

$$f(u_{1i}, u_{2i}) = f(u_{2i})f(u_{1i}|u_{2i}).$$

Then (16.1) can be written as

$$\begin{aligned} \int_c^\infty f(u_{1i}, u_{2i}) du_{1i} &= \int_c^\infty f(u_{2i})f(u_{1i}|u_{2i}) du_{1i} \\ &= f(u_{2i}) \int_c^\infty f(u_{1i}|u_{2i}) du_{1i} \\ &= f(u_{2i}) \Pr(u_{1i} > c|u_{2i}) \end{aligned}$$

Given our assumptions, $u_{1i}|u_{2i} \sim N\left(\frac{\sigma_{12}}{\sigma_2^2}u_{2i}, (1 - \rho^2)\sigma_1^2\right)$.

What is $\Pr(u_{1i} > c|u_{2i})$?

$$\frac{u_{1i} - \frac{\sigma_{12}}{\sigma_2^2}u_{2i}}{(1 - \rho^2)^{1/2}\sigma_1} > \frac{c - \frac{\sigma_{12}}{\sigma_2^2}u_{2i}}{(1 - \rho^2)^{1/2}\sigma_1}$$

Let $c = -\boldsymbol{\beta}'_1 \mathbf{x}_{1i}$ and $u_{2i} = y_{2i} - \boldsymbol{\beta}'_2 \mathbf{x}_{2i}$. Then

$$\frac{c - \frac{\sigma_{12}}{\sigma_2^2}u_{2i}}{(1 - \rho^2)^{1/2}\sigma_1} = \frac{-\boldsymbol{\beta}'_1 \mathbf{x}_{1i} - \frac{\sigma_{12}}{\sigma_2^2}(y_{2i} - \boldsymbol{\beta}'_2 \mathbf{x}_{2i})}{(1 - \rho^2)^{1/2}\sigma_1}$$

- Thus an individual's contribution to the likelihood if $y_{1i} = 1$ is

$$\frac{1}{\sigma_2} \phi \left(\frac{y_{2i} - \beta'_2 \mathbf{x}_{2i}}{\sigma_2} \right) \left[1 - \Phi \left(\frac{-\beta'_1 \mathbf{x}_{1i} - \frac{\sigma_{12}}{\sigma_2^2} (y_{2i} - \beta'_2 \mathbf{x}_{2i})}{(1 - \rho^2)^{1/2} \sigma_1} \right) \right]$$

- The log likelihood function for this model is

$$\begin{aligned} & \sum_{y_{1i}=0} \ln [1 - \Phi(\beta'_1 \mathbf{x}_{1i}/\sigma_1)] + \sum_{y_{1i}=1} \ln \left(\frac{1}{\sigma_2} \phi \left(\frac{y_{2i} - \beta'_2 \mathbf{x}_{2i}}{\sigma_2} \right) \right) \\ & + \sum_{y_{1i}=1} \ln \left[1 - \Phi \left(\frac{-\beta'_1 \mathbf{x}_{1i} - \frac{\sigma_{12}}{\sigma_2^2} (y_{2i} - \beta'_2 \mathbf{x}_{2i})}{(1 - \rho^2)^{1/2} \sigma_1} \right) \right] \end{aligned}$$

- As you can well imagine, this log likelihood is not the easiest thing to maximize.
 - Life would be easier if we had starting values that put us close to the maximum.
 - **Heckman's two-step method** gives us such starting values, but this method is often used instead of ML.
- Consider this model in a regression setting:

$$E[y_{1i} | \mathbf{x}_{1i}] = \Pr(y_{1i} = 1 | \mathbf{x}_{1i}) = \Phi(\beta'_1 \mathbf{x}_{1i} / \sigma_1)$$

$$\begin{aligned} E[y_{2i} | \mathbf{x}_{2i}, y_{1i} = 1] &= E[\beta'_2 \mathbf{x}_{2i} + u_{2i} | \mathbf{x}_{2i}, y_{1i} = 1] \\ &= \beta'_2 \mathbf{x}_{2i} + E[u_{2i} | \mathbf{x}_{2i}, y_{1i} = 1] \\ &= \beta'_2 \mathbf{x}_{2i} + E[u_{2i} | u_{1i} > -\beta'_1 \mathbf{x}_{1i}] \end{aligned}$$

- How might this be problematic for using OLS?
 - if $E[u_{2i}|u_{1i} > -\beta'_1 \mathbf{x}_{1i}] = 0$, no bias.
 - if $E[u_{2i}|u_{1i} > -\beta'_1 \mathbf{x}_{1i}] = a$ where a is any nonzero constant, then we get bias in the intercept but not the slope coefficients.
 - if $E[u_{2i}|u_{1i} > -\beta'_1 \mathbf{x}_{1i}] = f(\beta'_2 \mathbf{x}_{2i})$, then we get inconsistency in the intercept and the slope coefficients.

- We can use the **law of iterated expectations** to figure out what $E[u_{2i}|u_{1i} > -\beta'_1 \mathbf{x}_{1i}]$ is.
- Simple version of the law of iterated expectations (from Wooldridge's *Econometric Analysis of Cross Section and Panel Data*, p. 29):

$$E[y] = E_{\mathbf{x}}[E[y|\mathbf{x}]].$$

- Think of \mathbf{x} as a discrete vector taking on possible values $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_M$, with probabilities p_1, p_2, \dots, p_M . Then the LIE says:

$$E[y] = p_1 E[y|\mathbf{x} = \mathbf{c}_1] + p_2 E[y|\mathbf{x} = \mathbf{c}_2] + \dots + p_M E[y|\mathbf{x} = \mathbf{c}_M]$$

- That is, $E[y]$ is simply a weighted average of the $E[y|\mathbf{x} = \mathbf{c}_j]$, where the weight p_j is the probability that \mathbf{x} takes on the value of \mathbf{c}_j . In other words, a weighted average of averages.
- E.g., suppose we are interested in average IQ generally, but we have measures of average IQ by gender. We could figure out the quantity of interest by weighting average IQ by the relative proportions of men and women.

- For our problem:

$$E[u_{2i}|u_{1i} > -\boldsymbol{\beta}'_1 \mathbf{x}_{1i}] = E_{u_1} \left\{ E[u_{2i}|u_{1i} > -\boldsymbol{\beta}'_1 \mathbf{x}_{1i}, u_{1i}] | u_{1i} > -\boldsymbol{\beta}'_1 \mathbf{x}_{1i} \right\}$$

- Take the inner expectation:

$$\begin{aligned} E[u_{2i}|u_{1i} > -\boldsymbol{\beta}'_1 \mathbf{x}_{1i}, u_{1i}] &= E[u_{2i}|u_{1i}] \\ &= \tau u_{1i} \end{aligned}$$

where

$$\tau = \frac{\text{cov}(u_{1i}, u_{2i})}{\text{var}(u_{1i})} = \frac{\sigma_{12}}{\sigma_1^2}$$

- The outer expectation then becomes

$$\begin{aligned} E[\tau u_{1i}|u_{1i} > -\boldsymbol{\beta}'_1 \mathbf{x}_{1i}] &= \tau E[u_{1i}|u_{1i} > -\boldsymbol{\beta}'_1 \mathbf{x}_{1i}] \\ &= \tau \sigma_1 E \left[\frac{u_{1i}}{\sigma_1} \mid \frac{u_{1i}}{\sigma_1} > \frac{-\boldsymbol{\beta}'_1 \mathbf{x}_{1i}}{\sigma_1} \right] \\ &= \rho \sigma_2 \frac{\phi(\boldsymbol{\beta}'_1 \mathbf{x}_{1i}/\sigma_1)}{\Phi(\boldsymbol{\beta}'_1 \mathbf{x}_{1i}/\sigma_1)} \end{aligned}$$

- Putting this altogether gives

$$\begin{aligned} E[y_{2i} | \mathbf{x}_{2i}, y_{1i} = 1] &= \boldsymbol{\beta}'_2 \mathbf{x}_{2i} + \rho \sigma_2 \frac{\phi(\boldsymbol{\beta}'_1 \mathbf{x}_{1i} / \sigma_1)}{\Phi(\boldsymbol{\beta}'_1 \mathbf{x}_{1i} / \sigma_1)} \\ &= \boldsymbol{\beta}'_2 \mathbf{x}_{2i} + \beta_\lambda \lambda(\alpha_i) \end{aligned}$$

where $\alpha_i = \boldsymbol{\beta}'_1 \mathbf{x}_{1i} / \sigma_1$, $\lambda(\cdot) = \phi(\cdot) / \Phi(\cdot)$, and $\rho \sigma_2 = \beta_\lambda$.

- Why exactly is estimating outcome equation by OLS a bad idea?
 1. We have an omitted variables problem.
 2. $\lambda(\alpha_i)$ is a function of \mathbf{x}_{1i} and so will be correlated with \mathbf{x}_{1i} . If any elements of \mathbf{x}_{1i} are correlated with elements of \mathbf{x}_{2i} , the OLS estimator will be inconsistent.
- Identification issue: what if $\mathbf{x}_1 = \mathbf{x}_2$?

- This derivation provides the basis for Heckman's two-step method.

Step 1: Take full sample and estimate y_1 equation by probit.

Step 1.5: Using the (consistent) estimates from this equation, construct $\hat{\lambda} = \lambda(\hat{\beta}'_1 \mathbf{x}_{1i})$ for selected subsample where $y_{1i} = 1$.

Step 2: In selected sample, regress y_{2i} onto \mathbf{x}_{2i} and $\hat{\lambda}$ to obtain estimates of β_2 and β_λ .

- Unfortunately, we are not done after Step 2.
- It's smart to test whether or not sample selectivity exists. We can do this using the regression in Step 2.
 - Null hypotheses: $H_0 : \tau = 0$.
 - Under this null, $\frac{\tau}{\text{se}(\tau)} \stackrel{a}{\sim} N(0, 1)$.

If we fail to reject H_0 , then we can just use OLS.

- If we reject H_0 , then things are a lot more complicated. We have two problems:

1. Heteroskedasticity:

$$(y_{2i} | \mathbf{x}_{2i}, y_{1i} = 1) = \boldsymbol{\beta}'_2 \mathbf{x}_{2i} + \beta_\lambda \lambda_i + v_i$$

$$\text{var}[v_i] = \sigma_2^2(1 - \rho^2 \delta_i)$$

where $\delta_i = \lambda_i^2 - (\boldsymbol{\beta}'_1 \mathbf{x}_{1i} / \sigma_1) \lambda_i$.

- This follows from a theorem regarding the behavior of bivariate normal random variables.
- Assume for the moment that we have the true λ_i and δ_i (i.e., we have $\boldsymbol{\beta}_1$ rather than $\hat{\boldsymbol{\beta}}_1$).
- Let $\mathbf{x}_{2i}^* = [\mathbf{x}_{2i}, \lambda_i]$ and $\boldsymbol{\beta}_2^*$ be the least squares coefficient vector in the regression of y_{2i} on \mathbf{x}_{2i}^* for $y_{1i} = 1$. To correct for heteroskedasticity we compute:

$$\begin{aligned} \text{var}[\boldsymbol{\beta}_2^*] &= \sigma_2^2 [\mathbf{X}_2^{*'} \mathbf{X}_2^*]^{-1} \left[\sum_{i=1}^n (1 - \rho^2 \delta_i) \mathbf{x}_{2i}^* \mathbf{x}_{2i}^{*'} \right] [\mathbf{X}_2^{*'} \mathbf{X}_2^*]^{-1} \\ &= \sigma_2^2 [\mathbf{X}_2^{*'} \mathbf{X}_2^*]^{-1} [\mathbf{X}_2^{*'} (\mathbf{I} - \rho^2 \boldsymbol{\Delta}) \mathbf{X}_2^*] [\mathbf{X}_2^{*'} \mathbf{X}_2^*]^{-1} \end{aligned}$$

where

$$\mathbf{I} - \rho^2 \boldsymbol{\Delta} = \begin{bmatrix} 1 - \rho^2 \delta_1 & 0 & 0 & \cdots & 0 \\ 0 & 1 - \rho^2 \delta_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 - \rho^2 \delta_n \end{bmatrix}$$

2. Unfortunately, instead of β_1 we have $\hat{\beta}_1$ which implies that we are really estimating

$$(y_{2i} | \mathbf{x}_{2i}, y_{1i} = 1) = \beta_2' \mathbf{x}_{2i} + \beta_\lambda \hat{\lambda}_i + \varepsilon_i$$

where

$$\varepsilon_i = v_i + \beta_\lambda (\lambda_i - \hat{\lambda}_i)$$

- Since in finite samples $\beta_1 \neq \hat{\beta}_1$, then $\lambda_i \neq \hat{\lambda}_i$. This implies there is correlation across observations because the “measurement error” in estimating β_1 is common to all individuals in the subsample.
- To correct for this problem we compute

$$\text{var}[\hat{\beta}_2, \hat{\beta}_\lambda] = \hat{\sigma}_2^2 [\mathbf{X}_2^{*'} \mathbf{X}_2^*]^{-1} [\mathbf{X}_2^{*'} (\mathbf{I} - \rho^2 \Delta) \mathbf{X}_2^* + \mathbf{Q}] [\mathbf{X}_2^{*'} \mathbf{X}_2^*]^{-1}$$

where

$$\mathbf{Q} = \hat{\rho}^2 (\mathbf{X}_2^{*'} \hat{\Delta} \mathbf{X}_1) \text{Est. Var}[\hat{\beta}_1] (\mathbf{X}_1' \hat{\Delta} \mathbf{X}_2^*)$$

Part V

Probabilistic Choice Models

Section 17

Introduction

- The next set of models we'll focus on are based on the logistic distribution rather than the normal.
- To fix ideas, consider the binary logit model.

$$y_i^* = \boldsymbol{\beta}'\mathbf{x}_i + u_i$$

- Instead of assuming u_i follows a normal distribution, we assume it follows a logistic distribution w/ mean 0 and variance $\pi^2/3$.
- Going through derivations that are similar to what we did with the probit model we end up with

$$\Pr(y_i = 1) = \frac{\exp(\boldsymbol{\beta}'\mathbf{x})}{1 + \exp(\boldsymbol{\beta}'\mathbf{x})}$$

- The likelihood function for the logit model then is

$$L = \prod_{i=1}^n \left[\frac{\exp(\boldsymbol{\beta}'\mathbf{x})}{1 + \exp(\boldsymbol{\beta}'\mathbf{x})} \right]^{y_i} \left[\frac{1}{1 + \exp(\boldsymbol{\beta}'\mathbf{x})} \right]^{1-y_i}$$

- We also use models based on the logistic distribution when we have dependent variables that are categorical and the categories are unordered. For example:
 - Transportation choices: bus, car, train.
 - Voting choices in elections with more than two candidates.
 - Occupational categories: craft, blue collar, white collar, professional.

Section 18

The Multinomial Logit Model

- The set-up:
 - n decision makers indexed by i
 - J alternatives: $1, 2, \dots, J$.
 - We observe $y_i = j$ if i chooses alternative j .
 - We observe a vector of characteristics \mathbf{z}_i for each i .
 - The object of interest is $P_{ij} = \Pr(y_i = j | \mathbf{z}_i)$.
- We know $P_{ij} \in [0, 1]$, so we want P_{ij} to be proportional to some function that is nonnegative:

$$P_{ij} \propto \exp(\tilde{\gamma}'_j \mathbf{z}_i)$$

- The probabilities across choices must sum to one so

$$\sum_{j=1}^J P_{ij} = 1 \Rightarrow P_{ij} = \frac{\exp(\tilde{\gamma}'_j \mathbf{z}_i)}{\sum_{k=1}^J \exp(\tilde{\gamma}'_k \mathbf{z}_i)}$$

18.1 Identification Issue

- Unknown parameters in model are $\tilde{\gamma}_1, \tilde{\gamma}_2, \dots, \tilde{\gamma}_J$, but can't estimate all of these parameters.
- With J alternatives, if know the choice probabilities for $J - 1$ of these alternatives, then know the probability of choosing the remaining alternative.
- Don't have J independent pieces of information regarding the choice probabilities.
- \Rightarrow there can't be J separate parameter vectors to estimate—only have $J - 1$ identifiable parameter vectors.

18.2 Normalization

- We need some parameter restrictions to knock out one of the probabilities as an unknown piece of information:

$$\begin{aligned}
 P_{ij} &= \frac{\exp(\tilde{\gamma}'_j \mathbf{z}_i) / \exp(\tilde{\gamma}'_1 \mathbf{z}_i)}{\sum_{k=1}^J \exp(\tilde{\gamma}'_k \mathbf{z}_i) / \exp(\tilde{\gamma}'_1 \mathbf{z}_i)} \\
 &= \frac{\exp([\tilde{\gamma}'_j - \tilde{\gamma}'_1] \mathbf{z}_i)}{[\exp(\tilde{\gamma}'_1 \mathbf{z}_i) + \exp(\tilde{\gamma}'_2 \mathbf{z}_i) + \dots + \exp(\tilde{\gamma}'_J \mathbf{z}_i)] / \exp(\tilde{\gamma}'_1 \mathbf{z}_i)} \\
 &= \frac{\exp([\tilde{\gamma}'_j - \tilde{\gamma}'_1] \mathbf{z}_i)}{1 + \sum_{k=2}^J \exp([\tilde{\gamma}'_k - \tilde{\gamma}'_1] \mathbf{z}_i)} \\
 &= \frac{\exp(\gamma'_j \mathbf{z}_i)}{1 + \sum_{k=2}^J \exp(\gamma'_k \mathbf{z}_i)}
 \end{aligned}$$

where $\gamma'_j = \tilde{\gamma}'_j - \tilde{\gamma}'_1$ for $j = 2, 3, \dots, J$.

- To write down the likelihood function, first define a set of dummy variables, d_{ij} , where

$$d_{ij} = \begin{cases} 1 & \text{if } y_i = j \\ 0 & \text{otherwise} \end{cases}$$

- The likelihood function for an individual then is

$$\begin{aligned}
 L_i &= P_{i1}^{d_{i1}} P_{i2}^{d_{i2}} \dots P_{iJ}^{d_{iJ}} \\
 &= \prod_{j=1}^J P_{ij}^{d_{ij}}
 \end{aligned}$$

- The log likelihood for the sample is

$$\ln L = \sum_{i=1}^n \sum_{j=1}^J d_{ij} \ln P_{ij} \quad (18.1)$$

- Interpretation of the coefficients in this model is tricky:

$$\frac{\partial P_j}{\partial \mathbf{z}_i} = P_j \left[\gamma_j - \sum_{k=1}^J P_k \gamma_k \right]$$

- Another way to interpret effects is to compute log-odds ratios:

$$\begin{aligned} \frac{P_j}{P_k} &= \frac{\exp(\boldsymbol{\gamma}'_j \mathbf{z}_i) / \left(1 + \sum_{k=2}^J \exp(\boldsymbol{\gamma}'_k \mathbf{z}_i) \right)}{\exp(\boldsymbol{\gamma}'_k \mathbf{z}_i) / \left(1 + \sum_{k=2}^J \exp(\boldsymbol{\gamma}'_k \mathbf{z}_i) \right)} \\ &= \frac{\exp(\boldsymbol{\gamma}'_j \mathbf{z}_i)}{\exp(\boldsymbol{\gamma}'_k \mathbf{z}_i)} = \exp([\boldsymbol{\gamma}'_j - \boldsymbol{\gamma}'_k] \mathbf{z}_i) \end{aligned}$$

- Taking logs gives

$$\ln \left(\frac{P_j}{P_k} \right) = [\boldsymbol{\gamma}'_j - \boldsymbol{\gamma}'_k] \mathbf{z}_i$$

- Interpretation: we are taking a base case and comparing the probability of making other choices relative to the base case.
- Consider category 1, where $\boldsymbol{\gamma}_1 = \mathbf{0}$:

$$\ln \left(\frac{P_j}{P_1} \right) = \boldsymbol{\gamma}'_j \mathbf{z}_i$$

- So $\boldsymbol{\gamma}_j$ gives you the change in the log odds of being in category j instead of category 1.

Section 19

The Conditional Logit Model

- The conditional logit model is derived from random utility models.
- Key difference between the MNL model and the conditional logit model: the latter considers the effects of the choice characteristics on the determinants of choice probabilities.
- Suppose an individual i can choose from J alternatives, we can write down the utility of choice j as

$$U_{ij} = \boldsymbol{\beta}'\mathbf{x}_{ij} + \varepsilon_{ij}$$

- We assume that if i chooses j then the utility associated with alternative j is the largest of the utilities obtained from the various choices. So we are interested in

$$\Pr(U_{ij} > U_{ik}) \quad \text{for } k \neq j$$

- For example,

$$\begin{aligned}\Pr(y_i = 1) &= \Pr(U_{i1} > U_{i2}) \\ &= \Pr(\boldsymbol{\beta}'\mathbf{x}_{i1} + \varepsilon_{i1} > \boldsymbol{\beta}'\mathbf{x}_{i2} + \varepsilon_{i2}) \\ &= \Pr(\varepsilon_{i1} - \varepsilon_{i2} > \boldsymbol{\beta}'\mathbf{x}_{i2} - \boldsymbol{\beta}'\mathbf{x}_{i1})\end{aligned}$$

- To derive the statistical model, we make some distributional assumptions about ε_{ij} . Specifically, we assume it has an extreme value distribution:

$$F(\varepsilon_{ij}) = \exp(-\exp(-\varepsilon_{ij}))$$

$$f(\varepsilon_{ij}) = \exp(-\varepsilon_{ij}) \exp(-\exp(-\varepsilon_{ij}))$$

- Key things to note:
 1. We assume that the ε_{ij} are iid. This is a very strong assumption: not only are they iid across decision makers, but iid across alternatives for each decision-maker.
 2. We have common parameters across choices. Still have to normalize, though.
- Let Y_i indicate the choice of an individual. Then given our assumptions, we obtain the **conditional logit** model as

$$\Pr(Y_i = j) = \frac{\exp(\boldsymbol{\beta}' \mathbf{x}_{ij})}{\sum_{k=1}^J \exp(\boldsymbol{\beta}' \mathbf{x}_{ik})}$$

- The log likelihood for this model has the same form as the log likelihood for the MNL model (i.e., see eq. 18.1).

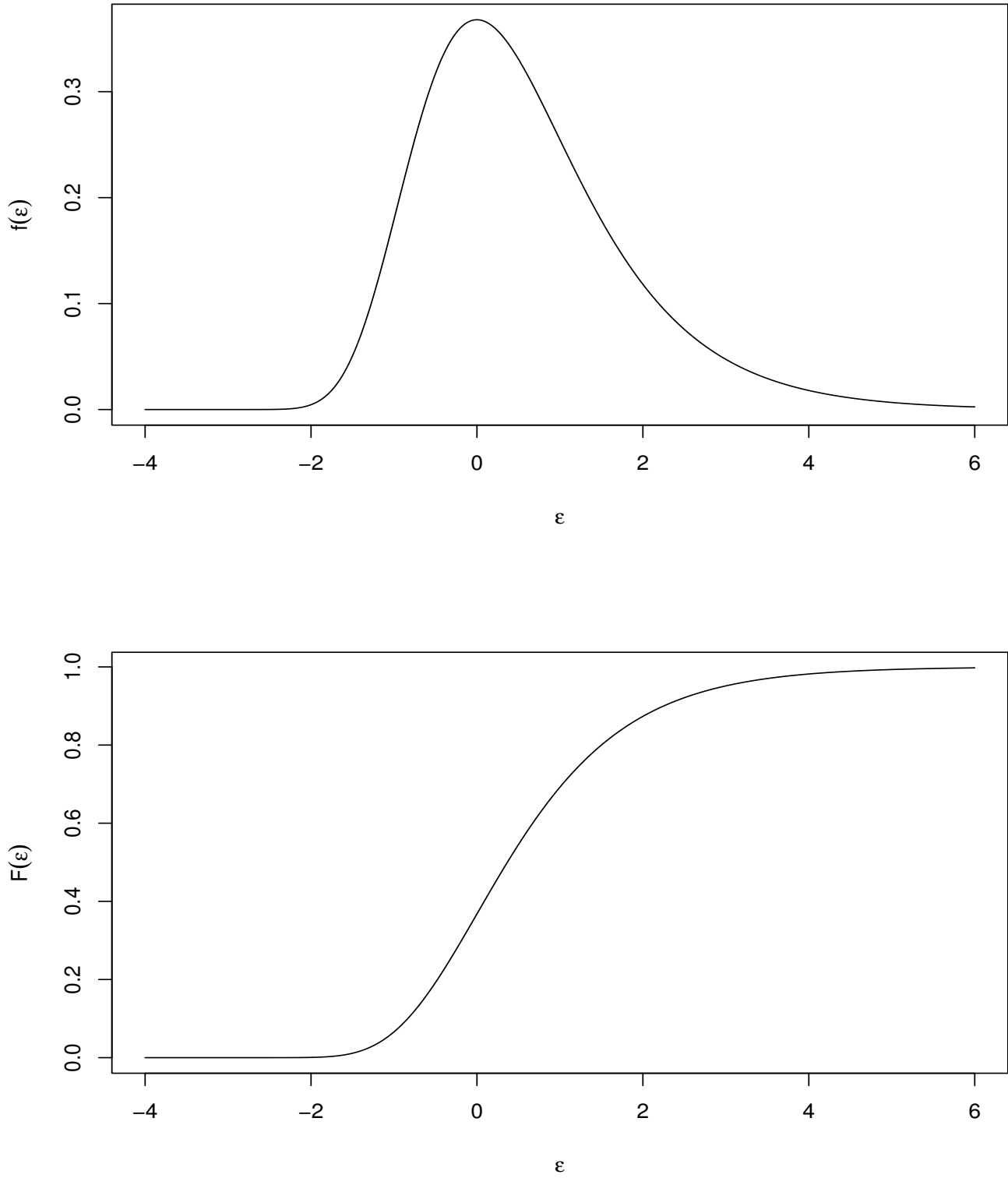


Figure 19.1: PDF and CDF for Extreme Value Distribution

19.1 Equivalence of the MNL model and conditional logit model

- It is important to note that the MNL model and conditional logit model are algebraically equivalent. To see this let

$$\gamma'_2 = (\mathbf{0} \ \boldsymbol{\beta} \ \mathbf{0}) \quad \gamma'_1 = (\boldsymbol{\beta} \ \mathbf{0} \ \mathbf{0}) \quad \mathbf{z}_i = \begin{pmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \\ \mathbf{x}_{i3} \end{pmatrix}$$

- Then

$$\begin{aligned} \frac{P_2}{P_1} &= \exp \left[((\mathbf{0} \ \boldsymbol{\beta} \ \mathbf{0}) - (\boldsymbol{\beta} \ \mathbf{0} \ \mathbf{0})) \begin{pmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \\ \mathbf{x}_{i3} \end{pmatrix} \right] \\ &= \exp \left[(-\boldsymbol{\beta} \ \boldsymbol{\beta} \ \mathbf{0}) \begin{pmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \\ \mathbf{x}_{i3} \end{pmatrix} \right] \\ &= \exp [\boldsymbol{\beta}(\mathbf{x}_{i2} - \mathbf{x}_{i1})] \end{aligned}$$

- We could combine the MNL and conditional logit models to get something like

$$P_{ij} = \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{ij} + \gamma'_j\mathbf{z}_i)}{\sum_{k=1}^J \exp(\boldsymbol{\beta}'\mathbf{x}_{ik} + \gamma'_k\mathbf{z}_i)}$$

19.2 Independence of Irrelevant Alternatives

- The multinomial logit model has the property known as **independence of irrelevant alternatives** which follows from the iid assumption for the disturbances.

- From above

$$\ln \left(\frac{P_j}{P_k} \right) = [\gamma'_j - \gamma'_k] \mathbf{z}_i$$

- This means that the odds ratio between any two choices does not depend on the other choices. This can be problematic. To see why consider the classic “red bus, blue bus” problem.
- Suppose individuals have 2 modes of transportation they can choose from:

1. Car.
2. Red bus.

- Assume individuals are indifferent between traveling by car and by bus. This implies

$$\Pr(\text{car}) = \Pr(\text{red bus}) = 1/2.$$

- The odds of taking the car v. the red bus:

$$\Pr(\text{car}) / \Pr(\text{red bus}) = 1.$$

- Suppose we add the choice of a blue bus. IIA requires

$$\Pr(\text{car}) = \Pr(\text{red bus}) = \Pr(\text{blue bus}) = 1/3$$

since $\Pr(\text{car})/\Pr(\text{red bus})$ must still equal 1.

- But if individuals simply split evenly in their choice of the red v. blue bus (which makes sense), then

$$\Pr(\text{red bus}) = \Pr(\text{blue bus}) = 1/4$$

- But this means

$$\frac{\Pr(\text{car})}{\Pr(\text{red bus})} = 2$$

- Intuition: we have to be careful what we define as an “irrelevant alternative.”
- The event that an individual chooses the car over the red bus should make the event that an individual chooses the car over the blue bus more likely.
 - * But we’ve essentially ruled this out by assuming iid errors.

- The IIA assumption renders the MNL model highly problematic for many settings in the social sciences.
- For example:
 - Suppose a country has a liberal and conservative party, and a new conservative party enters.
 - IIA \Rightarrow entrance of the 2nd conservative party should not affect the relative probability of an individual choosing b/t the liberal party and the first conservative party.
 - Not likely if voters view the two conservative parties as similar.

19.3 IIA test

- Intuition: if IIA holds, then you can estimate the model leaving out one of the alternatives and the coefficient estimates will not change significantly.
- If you estimate the model excluding an irrelevant alternative then the parameter estimates will be consistent, but not efficient.
- But if the relative odds ratios between the choices that remain are not truly independent, then we will get inconsistent estimates after we eliminate the choice.
- To perform this test we compute the statistic

$$\chi_k^2 = \left(\hat{\beta}_s - \hat{\beta}_f \right)' \left[\hat{\mathbf{V}}_s - \hat{\mathbf{V}}_f \right]^{-1} \left(\hat{\beta}_s - \hat{\beta}_f \right)$$

where s indicates the estimator using the subset of alternatives, f indicates the estimator using the full set of alternatives, and $\hat{\mathbf{V}}_s$ and $\hat{\mathbf{V}}_f$ are the respective asymptotic covariance matrices ($k =$ number of rows in $\hat{\beta}_s$).

- Note that if $\hat{\mathbf{V}}_s - \hat{\mathbf{V}}_f$ is not positive semidefinite, the statistic we compute can be negative. This is usually taken as evidence that IIA holds.
- Test is generally of low power.

Section 20

The Nested Logit Model

- IIA is a problem for the MNL model when some of the alternatives are closely related. The nested MNL (NMNL) alleviates this problem to a certain extent.
- One way to think about the NMNL model is that we are treating decisions as sequential.
- For example, suppose we want to study votes for candidates in multimember districts with multiple parties. We can think of voters first deciding which party they will vote for and then which of the party's candidates to vote for (\Rightarrow tree-like structure to the decision making process).
- Suppose individuals can choose from parties indexed by $i = 1, 2, \dots, C$ and candidates indexed by $j = 1, 2, \dots, N_i$ from party i .
- Let

$$U_{ij} = V_{ij} + \varepsilon_{ij}$$

where

- U_{ij} is the voter's utility for alternative (i, j) .
- V_{ij} is a function of the measured attributes of this alternative.
- ε_{ij} includes other factors not measured that affect utility.

- If we assume ε_{ij} has an extreme value distribution then the probability P_{ij} that the (i, j) th alternative will be chosen is given by

$$P_{ij} = \frac{\exp(V_{ij})}{\sum_{m=1}^C \sum_{n=1}^{N_m} \exp(V_{mn})}$$

- Let

$$V_{ij} = \boldsymbol{\beta}'\mathbf{x}_{ij} + \boldsymbol{\gamma}'\mathbf{z}_i$$

where \mathbf{x}_{ij} is a vector of variables that vary with **both** party and candidates and \mathbf{z}_i is a vector of variables that vary only with the party.

- We can write P_{ij} as $P_{j|i} \cdot P_i$ where

$$\begin{aligned} P_{j|i} &= \frac{\exp(V_{ij})}{\sum_{k=1}^{N_i} \exp(V_{ik})} \\ &= \frac{\exp(\boldsymbol{\gamma}'\mathbf{z}_i) \exp(\boldsymbol{\beta}'\mathbf{x}_{ij})}{\exp(\boldsymbol{\gamma}'\mathbf{z}_i) \sum_{k=1}^{N_i} \exp(\boldsymbol{\beta}'\mathbf{x}_{ik})} = \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{ij})}{\sum_{k=1}^{N_i} \exp(\boldsymbol{\beta}'\mathbf{x}_{ik})} \end{aligned}$$

and

$$\begin{aligned} P_i &= \frac{\sum_{j=1}^{N_i} \exp(V_{ij})}{\sum_{m=1}^C \sum_{n=1}^{N_m} \exp(V_{mn})} \\ &= \frac{\exp(\boldsymbol{\gamma}'\mathbf{z}_i) \sum_{j=1}^{N_i} \exp(\boldsymbol{\beta}'\mathbf{x}_{ij})}{\sum_{m=1}^C \exp(\boldsymbol{\gamma}'\mathbf{z}_m) \sum_{n=1}^{N_m} \exp(\boldsymbol{\beta}'\mathbf{x}_{mn})} \end{aligned}$$

- Next, define an inclusive value for the i th choice:

$$I_i = \ln \sum_{j=1}^{N_i} \exp(\boldsymbol{\beta}'\mathbf{x}_{ij})$$

- Then the above equations become

$$P_{j|i} = \frac{\exp(\boldsymbol{\beta}' \mathbf{x}_{i,j})}{\exp(I_i)}$$

$$P_i = \frac{\exp(\boldsymbol{\gamma}' \mathbf{z}_i + I_i)}{\sum_{m=1}^C \exp(\boldsymbol{\gamma}' \mathbf{z}_m + I_m)}$$

- We can rewrite P_i as

$$P_i = \frac{\exp(\boldsymbol{\gamma}' \mathbf{z}_i + \tau_i I_i)}{\sum_{m=1}^C \exp(\boldsymbol{\gamma}' \mathbf{z}_m + \tau_i I_m)}$$

where τ_i is a parameter that we will estimate. Note that above we restricted $\tau_i = 1$. If we let τ vary then we have the nested logit model.

- To estimate this model we can use a 2-stage procedure:
 1. Estimate $\boldsymbol{\beta}$ in $P_{j|i}$ by doing conditional logit.
 2. Compute the inclusive values using the estimates of $\boldsymbol{\beta}$ for all i categories. Estimate $\boldsymbol{\gamma}$ and τ by conditional logit.
- As with Heckman's two step method for selection models, we need to correct the standard errors from the second stage.
- Another alternative would be to just do full information maximum likelihood, which involves maximizing the log-likelihood

$$\ln L = \sum_n \ln (P_{j|i} \cdot P_i)$$

- Note that this model can be generalized to more choice-levels.

Section 21

The Multinomial Probit Model

- MNL/CL are unattractive because they assume $\varepsilon_{ij} \perp \varepsilon_{ik}$ for $j \neq k$.
- The **multinomial probit (MNP) model** relaxes this assumption, though as we will see, not without substantial cost.
- The derivation of the MNP model is similar to the derivation of the conditional logit model in that we start from a random utility model. To keep the derivation simple, we will look at the special case of three alternatives:

$$U_1 = \boldsymbol{\beta}'\mathbf{x}_1 + \boldsymbol{\gamma}'_1\mathbf{z} + \varepsilon_1$$

$$U_2 = \boldsymbol{\beta}'\mathbf{x}_2 + \boldsymbol{\gamma}'_2\mathbf{z} + \varepsilon_2$$

$$U_3 = \boldsymbol{\beta}'\mathbf{x}_3 + \boldsymbol{\gamma}'_3\mathbf{z} + \varepsilon_3$$

(Note we are suppressing the i subscript.)

- Let $\bar{U}_j = \boldsymbol{\beta}'\mathbf{x}_j + \boldsymbol{\gamma}'_j\mathbf{z}$.

- Assume $(\varepsilon_1, \varepsilon_2, \varepsilon_3) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}$

- The probability of choosing the first alternative is

$$\Pr(U_1 > U_2, U_1 > U_3) = \Pr(\varepsilon_2 - \varepsilon_1 < \bar{U}_1 - \bar{U}_2, \varepsilon_3 - \varepsilon_1 < \bar{U}_1 - \bar{U}_3)$$

- Let $\eta_{21} = \varepsilon_2 - \varepsilon_1$ and $\eta_{31} = \varepsilon_3 - \varepsilon_1$. Then η_{21} and η_{31} have a bivariate normal distribution with covariance matrix

$$\Omega_1 = \begin{bmatrix} \sigma_1^2 + \sigma_2^2 - 2\sigma_{12} & \sigma_1^2 - \sigma_{13} - \sigma_{12} + \sigma_{23} \\ \sigma_1^2 - \sigma_{13} - \sigma_{12} + \sigma_{23} & \sigma_1^2 + \sigma_3^2 - 2\sigma_{13} \end{bmatrix}$$

- Then the probability of choosing category 1 is

$$P_1 = \int_{-\infty}^{\bar{U}_1 - \bar{U}_2} \int_{-\infty}^{\bar{U}_1 - \bar{U}_3} f(\eta_{21}, \eta_{31}) d\eta_{31}, d\eta_{21}$$

- The probabilities for the other categories are defined similarly.
- The likelihood function for the MNP model looks the same as the likelihood function for the MNL and conditional logit models.
- Traditionally, the main barrier to using the MNP model has been its computational difficulty.
- Some advances have been made recently using simulation methods, but there is still not much progress beyond 3 categories (at least in political science).
- Fundamental identification issues remain.

21.1 Identification of variance-covariance matrix

- Not all of the parameters in Σ can be estimated.
- For the model written in utility differences, it was traditionally thought that $J(J - 1)/2$ free parameters could be estimated. More recent treatments claim that it is possible to estimate only $J(J - 1)/2 - 1$ elements. So need to make assumptions about restrictions.
- Easy way to test if you've got an identification problem: nonsensical parameters or parameters on the boundary.
- A standard assumption is $\sigma_j^2 = 1 \forall j$ (i.e., homoskedasticity across choices).

21.2 Application: Alvarez and Nagler “Economics, Issues, and the Perot Candidacy: Voter Choice in the 1992 Presidential Election” ’95 *AJPS*

- Key question of concern: how did the candidacy of Ross Perot—a rare serious 3rd party candidate—affect the outcome of the 1992 presidential election
- Did Perot take away votes from Clinton, Bush, or both?
- Obviously, a model that assumes IIA would have difficulty answering these questions.
- They estimate a 3-choice MNP using Nat’l Election Survey data. Their model includes both choice-specific and individual-specific characteristics.
- Normalize on the Perot choice, which makes sense given the stated goals of the study.
- See table for results.
- In a similar study, Lacy and Burden (“The Vote-Stealing and Turnout Effects of Ross Perot in the 1992 U.S. Presidential Election” ’99 *AJPS*) include the choice to vote in the election.
- However, some questions are raised about the model’s identification.

**Table 3. Multinomial Probit Estimates for a Three-Candidate Model
(Perot Coefficients Normalized to Zero)**

Independent Variables	Coefficients for	
	Bush	Clinton
Ideological Distance		-.09*
		.02
Constant	.50	-.44
	.44	.58
Felt personal finances were worse	-.04	.02
	.05	.05
Felt national economy was worse	-.14**	.21*
	.08	.10
Oppose government jobs	.07	-.01
	.05	.05
Oppose government health care	.10*	.06
	.05	.04
Oppose government minority assistance	.01	-.17*
	.05	.05
Abortion	-.35*	.01
	.14	.11
Region (East)	-.15	.32
	.17	.21
Region (South)	.25	.50*
	.18	.19
Region (West)	-.11	-.03
	.18	.21
New or returning voter	.28**	-.23*
	.15	.17
Term limits	.06	.08
	.13	.11
Felt deficit was a major problem	-.58*	-.003
	.22	.18
Democrat	-.19	1.34*
	.17	.28
Republican	1.00*	-.74
	.43	.46
Gender (Female)	.38*	.21
	.19	.14
Respondent's education	.14*	.004
	.07	.06
Age: 18-29	-.86*	-.57*
	.41	.26
Age: 30-44	-.64*	-.54*
	.30	.19
Age: 45-59	-.51*	-.10
	.24	.21
σ_{BC}		-.08
		.28
σ_{BP}		.27
		.54
σ_{CP}		-.07
		.26

Note: Maximum-likelihood estimates with their estimated standard errors below.

LL = -568.18; % correct = 70.6; number of observations = 909.

*indicates an estimate significant at the $p = .05$ level.

**indicates an estimate significant at the $p = .10$ level.

Table 3. Multinomial Probit Estimates for Four-Choice Model

	Clinton	Perot	Abstain
Constant	.313 (.519)	-1.510 (.981)	1.820** (.602)
Personal Finances (worse)	.102* (.046)	.141* (.061)	.007 (.051)
National Economy (worse)	.207** (.067)	.094 (.082)	.133* (.064)
National Health Care (oppose)	-.167** (.042)	-.193** (.054)	-.137** (.046)
Abortion (pro-choice)	.320** (.069)	.309** (.101)	.119 (.076)
Democrat	.917** (.179)	0.00 —	0.00 —
Independent	0.00 —	.399** (.202)	0.00 —
Liberal	.781** (.218)	.405 (.275)	.356 (.226)
Conservative	-.716** (.180)	-.702** (.234)	-.528** (.201)
No Ideology	-.078 (.203)	-.437 (.306)	.289 (.203)
Voted in 1988	0.00 —	.137 (.309)	-1.86** (.437)
Clinton Moral (disagree)	-.468** (.103)	0.00 —	0.00 —
South	-.122 (.154)	-.486* (.225)	.072 (.155)
Gender (male)	-.290* (.147)	-.561** (.203)	-.348* (.156)
Race (African-American)	1.040** (.286)	-.287 (.597)	.478* (.289)
Income	-.030* (.014)	-.007 (.019)	-.057* (.017)
Age 18–29	.134 (.234)	1.02** (.368)	-.022 (.246)
Age 30–44	.059 (.220)	.133 (.283)	.231 (.263)
Age 45–59	-.022 (.202)	.410 (.270)	.279 (.209)
College Educated	0.00 —	0.00 —	-.429* (.191)
External Efficacy	0.00 —	0.00 —	-.076* (.046)
$\hat{\sigma}_{Bush, Perot}$	-.99 (.91)		
$\hat{\sigma}_{Clinton, Perot}$.11 (.47)		
$\hat{\sigma}_{Bush, Abstain}$.23 (.30)		
$\hat{\sigma}_{Clinton, Abstain}$	-.09 (.42)		
$\hat{\sigma}_{Perot, Abstain}$	-.59 (.96)		

(continued)

**Table 3. Multinomial Probit Estimates
for Four-Choice Model (*continued*)**

Final Log Likelihood	-1314.12
$\chi^2(57)$	1450.24**
Number of Iterations	35
Percent Correctly Predicted	50.6
Number of Cases	1471

Note: Bush coefficients normalized to zero. Coefficients are maximum likelihood estimates with standard errors in parentheses. Parameters fixed at zero are indicated by 0.00 coefficients and (—) standard errors. * indicates $p < .05$, ** indicates $p < .01$, one-tailed.

Source: 1992 National Election Study

Part VI
Duration Models

Section 22

Introduction

- These models are used when we are interested in the time it takes for some event to occur.
- Duration data typically consists of a sample of cross-sections measured at a set of time intervals: t_1, t_2, \dots, t_K .
- Duration models essentially model the probability of an event occurring at time t given that the event has not occurred prior to t .
- Censoring: one problem that duration models have to account for is that the “spells” for some individuals may not end before the maximum number of time periods is reached.
- Right censoring v. left censoring
- Let T be a continuous random variable with pdf $f(t)$, where t denotes the realization of T . The probability distribution of the duration then is

$$\Pr(T < t) = F(t) = \int_0^t f(s)ds$$

Section 23

Functions for Analyzing Duration Data

- Given this specification, there are several functions that are useful for analyzing duration data:

1. The **survival function**

$$S(t) = \Pr(T \geq t) = 1 - F(t)$$

which gives the probability that the spell is of length at least t .

2. The **hazard rate** gives the rate at which spells are completed at duration t , given that they lasted until t .

$$\begin{aligned}\lambda(t) &= \lim_{\Delta \rightarrow 0} \frac{\Pr(t \leq T < t + \Delta | T \geq t)}{\Delta} \\ &= \lim_{\Delta \rightarrow 0} \frac{F(t + \Delta) - F(t)}{\Delta S(t)} \\ &= \frac{f(t)}{S(t)}\end{aligned}$$

3. The **integrated hazard function**

$$\Lambda(t) = \int_0^t \lambda(s) ds$$

is used in specification tests.

- The next step is to pick a distribution for $F(\cdot)$. The choice of distribution depends on what we are trying to model.
- The hazard functions of some popular distributions appear in Table 23. Plots of these hazards appear in Figures 23.1 and 23.2.

Table 23.1: Distributions, Hazard Functions, and Survival Functions

Distribution	Hazard Function	Survival Function
Exponential	λ	$\exp[-\lambda t]$
Weibull	$\lambda p(\lambda t)^{p-1}$	$\exp[-(\lambda t)^p]$
Lognormal	$(p/t)\phi[p \ln(\lambda t)]$	$\Phi[-p \ln(\lambda t)]$
Log-logistic	$\frac{\lambda p(\lambda t)^{p-1}}{1+(\lambda t)^p}$	$\frac{1}{1+(\lambda t)^p}$

Figure 23.1: Plots of hazard functions

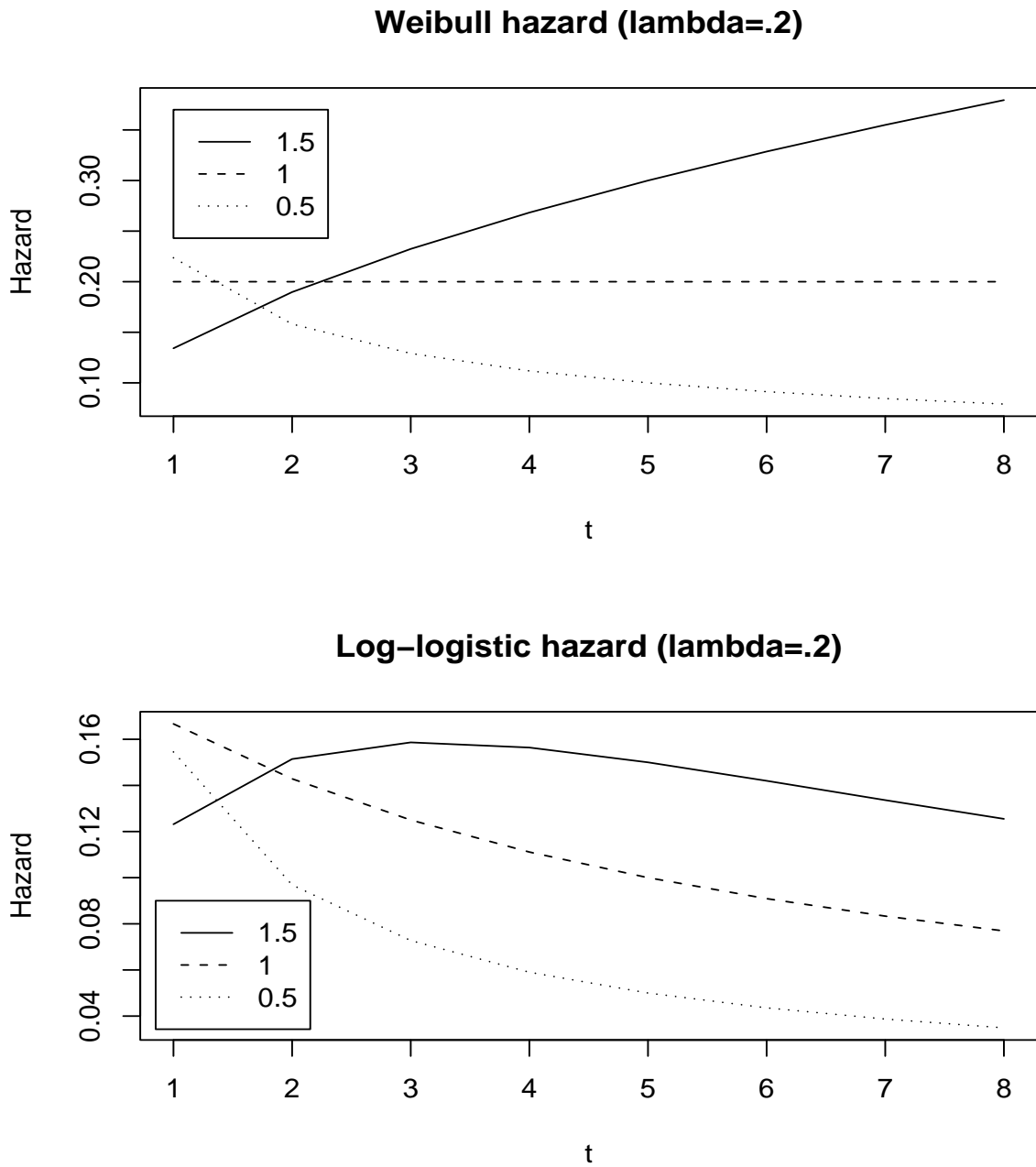
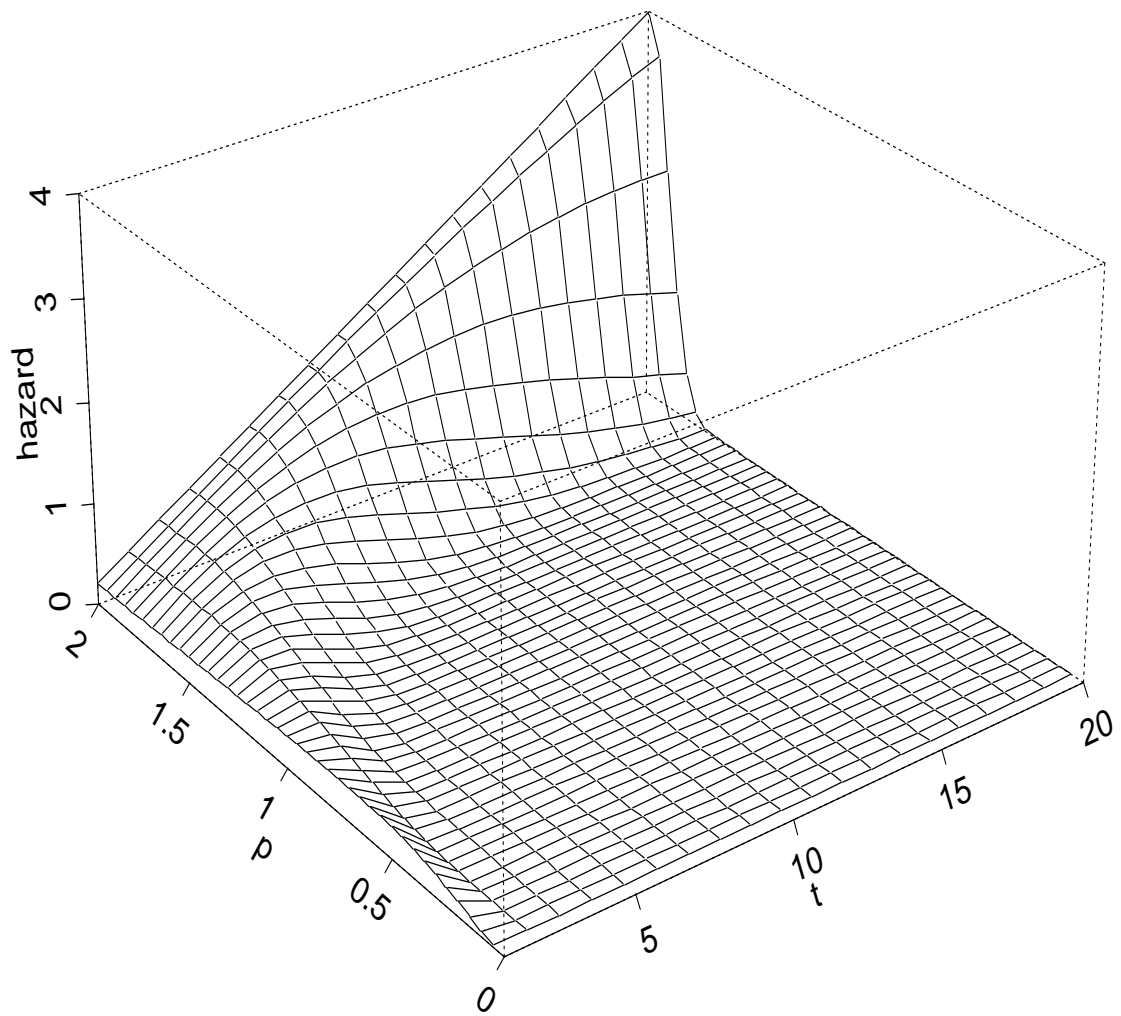


Figure 23.2: Lognormal Hazard Function ($\lambda = .5$)

- Let $d_i = 1$ if the i th spell is not censored, $d_i = 0$ if censored. Then the log-likelihood function is

$$\ln L(\theta) = \sum_{i=1}^n d_i \ln f(t_i, \theta) + \sum_{i=1}^n (1 - d_i) \ln S(t_i, \theta)$$

where $\theta = (p, \lambda)$.

- Sometimes it is more convenient to write the log-likelihood in terms of the hazard function:

$$\ln L(\theta) = \sum_{i=1}^n d_i \ln \lambda(t_i, \theta) - \sum_{i=1}^n \Lambda(t_i, \theta)$$

which follows from the fact that $f(t, \theta) = \lambda(t, \theta)S(t, \theta)$ and $\ln S(t, \theta) = -\Lambda(t, \theta)$.

- How do we go about picking the distribution?
 - One of the main things we want to consider is how do we think the probability of a failure changes over time.
- If we think the probability of a failure is time invariant, then we might want to use the exponential model. The hazard function incorporating explanatory variables is

$$\lambda(t) = \exp(\beta_0 + \boldsymbol{\beta}'_k \mathbf{x}_i)$$

- Typically, we hypothesize that the hazard is not time invariant, so we should use a distribution that depends on t .
- If the corresponding hazard function of a distribution slopes upward, it has **positive duration dependence**. If it slopes downward, it has **negative duration dependence**.

- The Weibull distribution is popular choice when we want to allow for time dependence in the model. The hazard function for the Weibull model with covariates can be specified as

$$\lambda(t) = \exp(\boldsymbol{\beta}' \mathbf{x}_i + \alpha \ln t)$$

This hazard is monotonically increasing or decreasing in t .

- Other distributions allow for more complex time dependencies.
- It is advisable to check whether it is valid to impose this amount of structure on the data. We can do residual analysis similar to what we do with OLS.
- The generalized residuals are given by the integrated hazard function:

$$\epsilon = \Lambda(t|\theta) = -\ln S(t|\theta)$$

- It can be shown that the generalized residuals will follow a unit exponential distribution. So we can compare the estimated integrated hazard function against the actual integrated hazard function for the unit exponential—which is just the 45 degree line.
- If the estimated integrated hazard deviates substantially from this line, this suggests there is unaccounted for heterogeneity in the data.

- Another easy specification check is to split the sample data into separate groups and estimate the model for each group.
- Greene gives some numerical tests based on conditional moments.

Section 24

Proportional Hazard Model

- The distribution we choose will generally have a substantial impact on the inferences we make. It would be nice if we had a method where we didn't have to specify or parameterize the nature of the dependence on time.
- One way to do this is the **proportional hazard model** developed by Cox. In this model the hazard function depends on a vector of explanatory variables \mathbf{x} with unknown coefficients $\boldsymbol{\beta}$ and λ_0 that we estimate.

- We factor the hazard function as follows

$$\lambda(t, \mathbf{x}, \boldsymbol{\beta}, \lambda_0) = g(\mathbf{x}, \boldsymbol{\beta})\lambda_0(t)$$

where $\lambda_0(t)$ is a *baseline* hazard, corresponding to $g(\cdot) = 1$.

- The baseline hazard can be chosen from a parametric family and written as

$$\lambda_0 = \lambda_0(t, \alpha)$$

which will give the log-likelihood function

$$\ln L(\alpha, \boldsymbol{\beta}) = \sum_{i=1}^n d_i \ln \lambda(t_i, \mathbf{x}_i, \alpha, \boldsymbol{\beta}) - \sum_{i=1}^n \Lambda(t_i, \mathbf{x}_i, \alpha, \boldsymbol{\beta})$$

- One way to go about estimating this model is to specify $\lambda_0(t, \alpha)$ and $g(\cdot)$.

- For the **exponential regression** specification we choose

$$g(\mathbf{x}, \boldsymbol{\beta}) = \exp(\boldsymbol{\beta}'\mathbf{x}).$$

and assume $\lambda_0(t, \alpha)$ is constant and normalize it to equal one.

- The log-likelihood function is

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^n d_i \boldsymbol{\beta}'\mathbf{x}_i - \sum_{i=1}^n t_i \exp(\boldsymbol{\beta}'\mathbf{x}_i)$$

Section 25

Partial Likelihood Estimator

- We can use the **partial likelihood estimator** to estimate $\boldsymbol{\beta}$ without having to estimate λ_0 . Suppose we can order the completed durations as $t_1 < t_2 < \dots < t_K$ and let's assume there is no censoring and only one individual exits at each period.

- The conditional probability that one and only one individual exits at t_1 given that any of the individuals could have exited at this time is

$$\frac{\lambda(t_1, \mathbf{x}_1, \boldsymbol{\beta})}{\sum_{i=1}^K \lambda(t_1, \mathbf{x}_i, \boldsymbol{\beta})} \quad (25.1)$$

- If we assume $\lambda(t, \mathbf{x}, \boldsymbol{\beta}) = \exp(\boldsymbol{\beta}'\mathbf{x})\lambda_0(t)$ then (25.1) becomes

$$\frac{\exp(\boldsymbol{\beta}'\mathbf{x}_1)}{\sum_{i=1}^K \exp(\boldsymbol{\beta}'\mathbf{x}_i)}$$

This is the contribution of the shortest duration to the partial likelihood function.

- The contribution of the j th shortest duration is

$$\frac{\exp(\boldsymbol{\beta}'\mathbf{x}_j)}{\sum_{i=j}^K \exp(\boldsymbol{\beta}'\mathbf{x}_i)}$$

- Things to note:
 - With this specification we have taken out the baseline hazard functions.
 - The contribution to the likelihood for each duration is the ratio of the hazard of the individual whose spell is completed at duration t_i divided by the sum of the hazards of those individuals who were “at risk” of having their spell completed at t_i .
- The partial log-likelihood then is

$$\ln L = \sum_{i=1}^K \left[\boldsymbol{\beta}' \mathbf{x}_i - \ln \sum_{j=i}^K \exp(\boldsymbol{\beta}' \mathbf{x}_j) \right]$$

- It is a substantial complication to incorporate an exit/failure by more than one individual at time t_i . The general idea is that we replace \mathbf{x}_i and \mathbf{x}_j with the sums of the explanatory variables of the individuals that fail at each time and account for the different permutations of failures in the denominator.
- This model can also be altered to account for censoring—see Lancaster 1990, Ch. 9.

- The partial likelihood approach has the following drawbacks:
 1. It can perform poorly in small samples relative to the straight ML approach in terms of the precision of the parameter estimates.
 2. If there are many ties in failure times ($> 5\%$ of the sample), can get bias due to computer programs using an approximation to the exact partial likelihood.
 3. We cannot directly analyze the form of time dependence.
 4. It has weaker theoretical foundations than ML.

- Given the drawbacks of the partial likelihood method, we may want to just choose a distribution for the hazard model and do ML.

Section 26

Nonparametric Approaches

- If we were very concerned about the distortions that parametric approaches might induce, we could use nonparametric methods. One such method is the **Kaplan-Meier product limit estimator**.
- As with the partial likelihood estimator, we order the completed durations in our sample from shortest to longest:
 $t_1 < t_2 < \dots < t_K$.
- Define the following terms:
 - h_j : the number of completed spells of durations t_j .
 - m_j : the number of observations censored between t_j and t_{j+1} .
 - n_j : the number of spells neither completed or censored before duration t_j

$$n_j = \sum_{i \geq j}^K (m_i + h_i)$$

- The estimator for the hazard function is

$$\hat{\lambda}(t_j) = \frac{h_j}{n_j}$$

- The estimator for the survival function is

$$\begin{aligned}\hat{S}(t_j) &= \prod_{i=1}^j \frac{(n_i - h_i)}{n_i} \\ &= \prod_{i=1}^j (1 - \hat{\lambda}_i)\end{aligned}$$

- Some other issues:
 - Time varying covariates.
 - Multiple failure models.
 - Competing risk models.
 - Split population models.
- Political science applications abound; two notable ones are Katz and Sala, “Careerism, Committee Assignments, and the Electoral Connection,” ’96 *APSR*; and Krehbiel and Kessler, “Dynamics of Cosponsorship,” ’96 *APSR*

Part VII
Event Count Models

Section 27

Introduction

- **Count data** is data where the dependent variable assumes nonnegative integer values $(0, 1, 2, \dots)$ for each of n observations. These values represent the number of times an event occurs within a fixed observation period.
- Examples of count data would include number of presidential vetoes per congressional session, annual number of presidential nominations for the Supreme Court, and the number of military conflicts between countries.
- Events occur at an unobserved expected rate of event occurrence during the observation period. We get to see the number of events that occurred during the period only at the end of the period.
- Least squares regression does not handle these kinds of data very well:
 - The linearity assumption is inappropriate for count data.
 - Count data typically come from distributions that are heteroskedastic. As a result least squares tends to be inefficient and it gives inconsistent standard errors.

Section 28

Poisson Regression Model

- Given these problems, a likelihood approach based on probability models is attractive. The **Poisson regression model** has been popular for analyzing count data.
- For the Poisson model, we assume that the i th event count—denoted y_i —is a draw from a Poisson distribution with parameter λ_i .
- To build the likelihood function we start with

$$\Pr(Y_i = y_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \dots$$

- We model the mean λ_i as a function of explanatory variables:

$$\ln(\lambda_i) = \boldsymbol{\beta}'\mathbf{x}_i$$

- Thus the expected number of events per period is given by

$$E[y_i|\mathbf{x}_i] = \lambda_i = \exp(\boldsymbol{\beta}'\mathbf{x}_i)$$

- A feature of the Poisson distribution is that its mean is equal to its variance. So $\text{var}[y_i|\mathbf{x}_i] = \exp(\boldsymbol{\beta}'\mathbf{x}_i)$ as well.

- Marginal effects are given by

$$\frac{\partial E[y_i | \mathbf{x}_i]}{\partial \mathbf{x}_i} = \lambda_i \boldsymbol{\beta}$$

- The log-likelihood function for the Poisson model is

$$\ln L = \sum_{i=1}^n [-\lambda_i + y_i \boldsymbol{\beta}' \mathbf{x}_i - \ln y_i!]$$

28.1 Dispersion

- The derivation of the Poisson model requires the assumption that the probability of an event occurring at any instant is constant within a given period and independent of all previous events during that observation period.
- These assumptions of homogeneity and independence are necessary if the data generating process is to have the feature of the Poisson distribution that the $E[y_i] = \text{var}[y_i]$.
- But we may have reason to believe that these assumptions do not hold.
- Suppose we were counting the number of presidential vetoes in a congressional session.
- If presidents were more likely to veto at certain times during the session (e.g., near the end of the fiscal year when it is necessary to pass appropriations bills) or if a successful veto might lead to President to veto more legislation then the assumptions of homogeneity and independence would be violated (i.e., there is heterogeneity and contagion in the data generating process).

- If $E[y_i] \neq \text{var}[y_i]$ we will get consistent but inefficient estimates and inconsistent standard errors.
- Let $\text{var}[y_i] = \lambda_i \sigma^2$, where σ^2 is the **dispersion parameter**. Then we have the following possible situations:

$$\begin{array}{lll} \text{Poisson dispersion} & \text{if } \sigma^2 = 1 & \Rightarrow E[y_i] = \text{var}[y_i] \\ \text{Overdispersion} & \text{if } \sigma^2 > 1 & \Rightarrow E[y_i] < \text{var}[y_i] \\ \text{Underdispersion} & \text{if } 0 < \sigma^2 < 1 & \Rightarrow E[y_i] > \text{var}[y_i] \end{array}$$

- Overdispersion can occur if there is heterogeneity in the data. For example, certain classes of bills that pass Congress might have a higher probability of being vetoed than others.
- **Positive contagion** can also lead to overdispersion. This occurs when the occurrence of previous events increases the likelihood of future events.
- **Negative contagion**—when the number of previous events decreases the likelihood of future events—induces underdispersion. Underdispersion is much less common in social science settings.

28.2 Tests for Overdispersion

1. **Regression-based test:** The hypotheses we test are

$$H_0 : \text{var}[y_i] = E[y_i]$$

$$H_1 : \text{var}[y_i] = E[y_i] + \alpha g(E[y_i])$$

We estimate the regression

$$z_i = \delta_0 + \varepsilon_i$$

or

$$z_i = \delta_1 \lambda_i + \varepsilon_i$$

where

$$z_i = \frac{(y_i - \lambda_i)^2 - y_i}{\lambda_i \sqrt{2}}$$

and λ_i is the predicted value from the Poisson model. We then do a t test to see whether $\hat{\delta}_0$ or $\hat{\delta}_1$ is statistically different from zero. If they are, we reject H_0 .

2. **Lagrange Multiplier (LM) test:** This test is similar to a likelihood ratio test (in fact, the two tests are asymptotically equivalent), except that the LM test requires us to estimate only the restricted model.

- The intuition behind the LM test is that if the restrictions are valid then the restricted estimator should be close to the point that maximizes the likelihood function, which implies that the slope of the log-likelihood function should be near zero at the restricted estimator.
- The test is based on the slope of the log-likelihood at the point where the function is maximized subject to the restriction.
- The restricted model in this case is the Poisson model. The LM statistic is given by

$$\text{LM} = \left[\frac{\sum_{i=1}^n [(y_i - \hat{\lambda}_i)^2 - y_i]}{\sqrt{2 \sum_{i=1}^n \hat{\lambda}_i^2}} \right]^2 = \frac{(\mathbf{e}'\mathbf{e} - n\bar{y})^2}{2\hat{\boldsymbol{\lambda}}'\hat{\boldsymbol{\lambda}}}$$

This follows a chi-squared distribution with the degrees of freedom equal to the number of restrictions.

Section 29

Gamma Model for Overdispersion

- In the case of overdispersion, our likelihood function will be based on the negative binomial distribution.
- To account for heterogeneity, we generalize the Poisson model by assuming that λ_i is a random variable which follows a **gamma distribution**:

$$f_\gamma(\lambda|\phi_i, \sigma^2) = \frac{\lambda^{\phi_i/(\sigma^2-1)-1} \exp(-\lambda/(\sigma^2-1))}{\Gamma[\phi_i/(\sigma^2-1)] (\sigma^2-1)^{\phi_i/(\sigma^2-1)}}$$

where λ , ϕ , and σ^2 are all greater than zero and

$$\Gamma(x) = \int_0^\infty z^{x-1} \exp(-z) dz$$

- With this parameterization $E(\lambda_i) = \phi$ and $\text{var}(\lambda_i) = \phi(\sigma^2 - 1)$.
- We can write the joint distribution of Y_i and λ as

$$f(y_i, \lambda_i|\phi_i, \sigma^2) = f_p(y_i|\lambda_i) f_\gamma(\lambda_i|\phi_i, \sigma^2)$$

where f_p denotes the Poisson distribution.

- After a rather messy derivation we get the negative binomial distribution

$$f_{nb}(y_i|\lambda_i, \sigma^2) = \frac{\Gamma\left(\frac{\lambda_i}{\sigma^2-1} + y_i\right)}{\Gamma\left(\frac{\lambda_i}{\sigma^2-1}\right) y_i!} \left(\frac{\sigma^2-1}{\sigma^2}\right)^{y_i} (\sigma^2)^{\frac{-\lambda_i}{\sigma^2-1}}$$

where $\lambda_i > 0$ and $\sigma^2 > 1$.

- Now $E(Y_i) = \lambda_i$ and $\text{var}(Y_i) = \lambda_i\sigma^2$.
- We can include explanatory variables in the same manner as before: $\lambda_i = \exp(\boldsymbol{\beta}'\mathbf{x}_i)$.
- The log-likelihood function is

$$\ln L = \sum_{i=1}^n \left\{ \ln \Gamma\left(\frac{\lambda_i}{\sigma^2-1} + y_i\right) - \ln \Gamma\left(\frac{\lambda_i}{\sigma^2-1}\right) + y_i \ln(\sigma^2-1) - \ln \sigma^2 \left(y_i + \frac{\lambda_i}{\sigma^2-1}\right) \right\}$$

Section 30

Binomial Model for Underdispersion

- In the case of underdispersion, our likelihood function is based on a binomial distribution:

$$f_b(y_i|\lambda_i, \sigma^2) = \frac{\Gamma\left(\frac{-\lambda_i}{\sigma^2-1} + 1\right)}{y_i! \Gamma\left(\frac{-\lambda_i}{\sigma^2-1} - y_i + 1\right)} (1 - \sigma^2)^{y_i} (\sigma^2)^{\frac{-\lambda_i}{\sigma^2-1} - y_i}$$

where $E(Y_i) = \lambda_i > 0$, $\text{var}(Y_i) = \lambda_i \sigma^2$ and $0 < \sigma^2 < 1$.

- One problem with this distribution is that if $-\lambda_i/(\sigma^2 - 1)$ is not an integer then

$$\sum_{m=0}^{\infty} f_b(y_i|\lambda_i, \sigma^2) \neq 1$$

- This is easily fixed simply by dividing $f_b(\cdot)$ by a scaling factor that assures the probabilities sum to one:

$$f_{cpb}(y_i|\lambda_i, \sigma^2) = \frac{f_b(y_i|\lambda_i, \sigma^2)}{D_i}$$

where

$$D_i = \sum_{m=0}^{\left[\frac{-\lambda_i}{\sigma^2-1}\right]+1} f_b(m|\lambda_i, \sigma^2)$$

Note that this imposes a theoretical maximum on the count variable: $-\lambda_i/\sigma^2 - 1$.

- The log-likelihood function corresponding to the **continuous parameter binomial** distribution is

$$\begin{aligned} \ln L = \sum_{i=1}^n & \left\{ \ln \Gamma \left(\frac{-\lambda_i}{\sigma^2 - 1} + 1 \right) \right. \\ & - \ln \Gamma \left(\frac{-\lambda_i}{\sigma^2 - 1} - y_i + 1 \right) + y_i \ln (1 - \sigma^2) \\ & \left. + \left(\frac{-\lambda_i}{\sigma^2 - 1} - y_i \right) \ln(\sigma^2) - \ln(D_i) \right\} \end{aligned}$$

- One way to proceed would be to do some sort of pretest that tells us the type of dispersion and then estimate the appropriate likelihood function. This has the drawback that if the test is incorrect, then we can make incorrect inferences.

Section 31

Generalized Event Count Model

- The **Generalized Event Count (GEC) model** does not require us to make any assumptions about heterogeneity or contagion—i.e., we don't have to decide the nature of the dispersion a priori. Instead we estimate it.
- This model is derived from the GEC distribution, which is a general distribution which includes the Poisson, negative binomial, and continuous parameter binomial distributions as special cases.
- The count probabilities of the GEC satisfy the following difference equation:

$$f_{gec}(y_i|\lambda_i, \sigma^2) = \left[\frac{\lambda_i + (\sigma^2 - 1)(y_i - 1)}{\sigma^2 y_i} \right] f_{gec}(y_i - 1|\lambda_i, \sigma^2)$$

- We can recursively substitute values for y_i to obtain specific probabilities:

$$\begin{aligned} \Pr(Y_i = 2|\lambda_i, \sigma^2) &= \left[\frac{\lambda_i + (\sigma^2 - 1)1}{\sigma^2 2} \right] f_{gec}(1|\lambda_i, \sigma^2) \\ &= \left[\frac{\lambda_i + (\sigma^2 - 1)1}{\sigma^2 2} \right] \left[\frac{\lambda_i + (\sigma^2 - 1)0}{\sigma^2 1} \right] f_{gec}(0|\lambda_i, \sigma^2) \end{aligned}$$

- This generalizes to

$$f_{gec}(y_i|\lambda_i, \sigma^2) = f_{gec}(0|\lambda_i, \sigma^2) \prod_{j=1}^{y_i} \left[\frac{\lambda_i + (\sigma^2 - 1)(j - 1)}{\sigma^2 j} \right] \quad (31.1)$$

for $y_i \in \{1, 2, 3, \dots\}$.

- In order to get $\Pr(Y_i = 0|\lambda_i, \sigma^2) = f_{gec}(0|\lambda_i, \sigma^2)$ we can use the axiom of probability that

$$\sum_{m=0}^{\infty} f_{gec}(m|\lambda_i, \sigma^2) = 1$$

(i.e., y_i takes on some value in a given period with probability 1).

- Summing over both sides of Eq. 31.1 and solving for $f_{gec}(0|\lambda_i, \sigma^2)$ gives

$$f_{gec}(0|\lambda_i, \sigma^2) = \left[\sum_{m=1}^{\infty} \prod_{j=1}^m \left(\frac{\lambda_i + (\sigma^2 - 1)(j - 1)}{\sigma^2 j} \right) \right]^{-1}$$

- We can then use standard results on the convergence of infinite series to solve for the probability of zero events. These terms will vary with the value of σ^2 :

$$f_{gec}(0|\lambda_i, \sigma^2) = \begin{cases} \exp(-\lambda_i) & \text{for } y_i = 0 \text{ and } \sigma^2 = 1 \\ (\sigma^2)^{-\lambda_i/(\sigma^2-1)} & \text{for } y_i = 0 \text{ and } \sigma^2 > 1 \\ (\sigma^2)^{-\lambda_i/(\sigma^2-1)} D_i^{-1} & \text{for } y_i = 0 \text{ and } 0 < \sigma^2 < 1 \end{cases}$$

- Collecting all this together gives the following distribution:

$$f_{gec}(y_i|\lambda_i, \sigma^2) = \begin{cases} f_{gec}(0|\lambda_i, \sigma^2) \prod_{j=1}^{y_i} \left[\frac{\lambda_i + (\sigma^2 - 1)(j-1)}{\sigma^{2j}} \right] & \text{for } y_i = 1, 2, \dots \\ \exp(-\lambda_i) & \text{for } y_i = 0 \text{ and } \sigma^2 = 1 \\ (\sigma^2)^{-\lambda_i/(\sigma^2-1)} & \text{for } y_i = 0 \text{ and } \sigma^2 > 1 \\ (\sigma^2)^{-\lambda_i/(\sigma^2-1)} D_i^{-1} & \text{for } y_i = 0 \text{ and } 0 < \sigma^2 < 1 \\ & \text{and } y_i \leq \left[\frac{-\lambda_i}{\sigma^2-1} \right] + 1 \\ 0 & \text{otherwise} \end{cases}$$

- The log-likelihood of for this distribution (reduced to sufficient statistics) is

$$\ln L(\beta, \sigma^2) = \sum_{i=1}^n \left\{ C_i - y_i \ln \sigma^2 + \sum_{j=1}^{y_i} \ln[\exp(\beta' \mathbf{x}_i) + (\sigma^2 - 1)(j - 1)] \right\}$$

where

$$C_i = \begin{cases} -\exp(\beta' \mathbf{x}_i) & \text{for } \sigma^2 = 1 \\ -\exp(\beta' \mathbf{x}_i) \ln \sigma^2 (\sigma^2 - 1)^{-1} & \text{for } \sigma^2 > 1 \\ -\exp(\beta' \mathbf{x}_i) \ln \sigma^2 (\sigma^2 - 1)^{-1} - \ln(D_i) & \text{for } 0 < \sigma^2 < 1 \end{cases}$$

Section 32

Hurdle Poisson Models

- In some situations, it is unreasonable to assume that the process generating counts of zero is the same as the one generating positive outcomes.
- In such cases, we would want to alter standard count models to try to model these separate processes. Essentially, we want to model the contributions to the likelihood of the zero observations differently from the contributions of the positive ones.
- One way to do this is to have a binary probability model that determines whether a zero or positive outcome occurs and then use a truncated Poisson distribution for the positive outcomes. For example:

$$\begin{aligned}\Pr(Y_i = 0) &= \exp(-\theta) \\ \Pr(Y_i = y_i) &= \frac{(1 - \exp(-\theta)) \exp(-\lambda_i) \lambda_i^{y_i}}{y_i! (1 - \exp(-\lambda_i))}, \quad y_i = 1, 2, \dots\end{aligned}$$

- Note:
 - Probability of non-zero outcomes is rescaled, so that all sum to 1.
 - Probability of zero outcomes is not a function of covariates—can generalize this though.
 - Negative binomial version also exists.

- For a political science perspective on Poisson models see King, “Statistical Models for Political Science Event Counts” ’88 *AJPS*.
- For the development of the GEC model see King, “Variance Specification in Event Count Models: From Restrictive Assumptions to a Generalized Estimator.” ’89 *AJPS*.
- For political science applications of event count models see
 - Canon, “Sacrificial Lambs or Strategic Politicians: Political Amateurs in U.S. House Elections.” ’93 *AJPS*.
 - Krause, “Federal Reserve Policy Decision Making.” ’94 *AJPS*.
 - Wang et al. “Inequality and Political Violence Revisited.” ’93 *APSR*.

Part VIII

Models for Repeated Observations—Dichotomous Dependent Variables

Section 33

Introduction

- The basic set up of the repeated observations model for dichotomous dependent variables is similar to the standard models.

$$y_{it}^* = \beta \mathbf{x}_{it} + \alpha_i + u_{it}, \quad (33.1)$$

where

$$y_{it} = \begin{cases} 1 & \text{if } y_{it}^* > 0 \\ 0 & \text{if } y_{it}^* \leq 0 \end{cases}$$

and $i = 1, \dots, N$ and $t = 1, \dots, T$, and u_{it} is assumed to be iid with mean zero and variance σ_u^2 .

- The choice we make about the distribution for the disturbance term matters a lot with repeated observations data and is based on our beliefs about the correlation between the explanatory variables and the individual specific effect.
- If we think there is no correlation between α_i and \mathbf{x}_{it} then we can estimate a **random effects model**. With this model we assume that $\alpha \sim IID(0, \sigma_\alpha^2)$.
- If we think there is correlation between α_i and \mathbf{x}_{it} then we can estimate a **fixed effects model**. Here we assume that the α_i are fixed parameters to be estimated.

- If $T \rightarrow \infty$, then it is possible to get consistent estimates of β and α_i .
- However, if T is fixed and $N \rightarrow \infty$, then we have the incidental parameters problem—i.e., since the number of parameters increases with N , we cannot consistently estimate α_i for fixed T .
- Unfortunately, the inconsistency in α_i is transmitted to β .
- In the linear regression case we take out the individual effect by subtracting off the means of the variables. That is, if we take the average over time of eq. 33.1, we get

$$\bar{y}_i^* = \beta \bar{x}_i + \alpha_i + \bar{u}_i, \quad (33.2)$$

where $\bar{y}_i^* = \sum_{t=1}^T y_{it}^*/T$. Subtracting eq. 33.2 from eq. 33.1 gives

$$y_{it}^* - \bar{y}_i^* = \beta(x_{it} - \bar{x}_i) + (u_{it} - \bar{u}_i), \quad (33.3)$$

- But this transformation is not valid with a qualitative limited dependent variable model.
- Such models are nonlinear and so mean-differencing will not work to remove the individual specific effects.

33.1 Fixed Effect Logit

- Chamberlain ('80 *Rev. of Econ. Studies*) has derived a conditional maximum likelihood estimator (CMLE) that works in a similar way. The conditional likelihood function can be written as

$$L = \prod_{i=1}^N \Pr \left(y_{i1}, \dots, y_{iT} \mid \sum_{t=1}^T y_{it} \right)$$

- Consider the case where $T = 2$. The unconditional likelihood is

$$L = \prod_{i=1}^N \Pr(y_{i1}) \Pr(y_{i2})$$

- Note that:

$$\begin{aligned} \Pr[y_{i1} = 0, y_{i2} = 0 \mid y_{i1} + y_{i2} = 0] &= 1 \\ \Pr[y_{i1} = 1, y_{i2} = 1 \mid y_{i1} + y_{i2} = 2] &= 1 \end{aligned}$$

which means these probabilities add no information to the conditional log likelihood so we can ignore them.

- But

$$\begin{aligned} \Pr[y_{i1} = 0, y_{i2} = 1 | y_{i1} + y_{i2} = 1] &= \frac{\Pr[y_{i1} = 0, y_{i2} = 1 \text{ and } y_{i1} + y_{i2} = 1]}{\Pr[y_{i1} + y_{i2} = 1]} \\ &= \frac{\Pr[y_{i1} = 0, y_{i2} = 1 \text{ and } y_{i1} + y_{i2} = 1]}{\Pr[y_{i1} = 0, y_{i2} = 1] + \Pr[y_{i1} = 1, y_{i2} = 0]} \end{aligned}$$

- If we assume that the data follow a logistic distribution then we can rewrite this as

$$\frac{\frac{1}{1+\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i1})} \frac{\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i2})}{1+\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i2})}}{\frac{1}{1+\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i1})} \frac{\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i2})}{1+\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i2})} + \frac{\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i1})}{1+\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i1})} \frac{1}{1+\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i2})}}$$

which simplifies to

$$\frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{i2})}{\exp(\boldsymbol{\beta}'\mathbf{x}_{i1}) + \exp(\boldsymbol{\beta}'\mathbf{x}_{i2})}$$

- The expression for the remaining probability is similarly derived. These probabilities then constitute the conditional log-likelihood function, which can be maximized using standard techniques.
- This can be extended to T of arbitrary size but the computations are excessive for $T > 10$.

- Can't use standard specification test like LR for checking unit heterogeneity b/c likelihoods are not comparable (CML uses a restricted data set).
- But can use this estimator to do a Hausman test for the presence of individual effects.
- Intuition: in the absence of individual specific effects, both the Chamberlain estimator and the standard logit maximum likelihood estimator are consistent, but the former is inefficient. If individual specific effects exist, then the Chamberlain estimator is consistent while the standard logit MLE is inconsistent.
 - Inefficiency is due to loss of information/throwing away observations.
- We compute the following statistic for the test:

$$\chi_k^2 = \left(\hat{\beta}_{\text{CML}} - \hat{\beta}_{\text{ML}} \right)' \left[\mathbf{V}_{\text{CML}} - \mathbf{V}_{\text{ML}} \right]^{-1} \left(\hat{\beta}_{\text{CML}} - \hat{\beta}_{\text{ML}} \right)$$

If we get a significant χ^2 value we reject the null of no individual specific effects.

33.2 Random Effects Probit

- The probit model does not lend itself to the fixed effects treatment because there is no way to sweep out the individual specific effects. But we can estimate a probit model if we assume random effects.
- Let

$$\varepsilon_{it} = \alpha_i + u_{it}$$

and assume $\alpha_i \sim N(0, \sigma_\alpha^2)$; $u_{it} \sim N(0, \sigma_u^2)$, and α_i and u_{it} are independent of each other. Then

$$\text{var}[\varepsilon_{it}] = \sigma_u^2 + \sigma_\alpha^2 = 1 + \sigma_\alpha^2$$

and

$$\text{corr}[\varepsilon_{it}, \varepsilon_{is}] = \rho = \frac{\sigma_\alpha^2}{1 + \sigma_\alpha^2}$$

for $t \neq s$. This implies $\sigma_\alpha^2 = \rho/(1 - \rho)$.

- We can write the probability associated with an observation as

$$\Pr[y_{it}] = \int_{-\infty}^{q_{it}\boldsymbol{\beta}'\mathbf{x}_{it}} f(\varepsilon_{it})d\varepsilon_{it} = \Phi[q_{it}\boldsymbol{\beta}'\mathbf{x}_{it}]$$

where $q_{it} = 2y_{it} - 1$.

- Because of the α_i , the T observations for i are jointly normally distributed. The individual's contribution to the likelihood is

$$\begin{aligned} L_i &= \Pr[y_{i1}, y_{i2}, \dots, y_{iT}] \\ &= \int_{-\infty}^{q_{i1}\boldsymbol{\beta}'\mathbf{x}_{i1}} \int_{-\infty}^{q_{i2}\boldsymbol{\beta}'\mathbf{x}_{i2}} \cdots \int_{-\infty}^{q_{iT}\boldsymbol{\beta}'\mathbf{x}_{iT}} f(\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{iT})d\varepsilon_{iT} \cdots \varepsilon_{i2}\varepsilon_{i1} \end{aligned}$$

- Rather than evaluating multiple integrals, a simplification is possible. Consider the joint density:

$$f(\varepsilon_{i1}, \dots, \varepsilon_{iT}, \alpha_i) = f(\varepsilon_{i1}, \dots, \varepsilon_{iT} | \alpha_i) f(\alpha_i)$$

- We can then integrate over α_i :

$$f(\varepsilon_{i1}, \dots, \varepsilon_{iT}) = \int_{-\infty}^{\infty} f(\varepsilon_{i1}, \dots, \varepsilon_{iT} | \alpha_i) f(\alpha_i) d\alpha_i$$

- Conditioned on α_i , the ε_i s are independent:

$$f(\varepsilon_{i1}, \dots, \varepsilon_{iT}) = \int_{-\infty}^{\infty} \prod_{t=1}^T f(\varepsilon_{it} | \alpha_i) f(\alpha_i) d\alpha_i$$

- Letting $r_i = \alpha_i / (\sigma_\alpha \sqrt{2})$ and making the change of variable gives i 's contribution to the likelihood as

$$L_i = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-r_i^2} \left\{ \prod_{t=1}^T \Phi[q_{it}(\boldsymbol{\beta}' \mathbf{x}_{it} + \theta r_i)] \right\} dr_i \quad (33.4)$$

$$\theta = \sqrt{\frac{2\rho}{1-\rho}} \quad (33.5)$$

- Things to note:
 - The assumption that the α_i and \mathbf{x}_{it} are uncorrelated is very restrictive. We are also assuming that the within-cross section correlation is the same across all time periods.
 - ρ can be interpreted as the proportion of the variance contributed by the unit effects.
 - We can test for unit heterogeneity by checking the statistical significance of ρ . One way to do this is with a likelihood ratio test of the random effects probit and pooled probit models.
 - The standard way to evaluate the integral in the likelihood is by Gauss-Hermite quadrature. This raises some concerns about how the size of T and N affect the accuracy of the quadrature approximation, and some checks of the performance of the approximation are in order.
 - **Stata's** `xtprobit` command can be used to estimate this model.
 - We could derive this model for the logistic distribution rather than the normal distribution.

33.3 Correlated Random Effects Probit

- The big drawback of the random effects model is the unattractive assumption that the α_i and \mathbf{x}_i are uncorrelated.
- Chamberlain has proposed a correlated random effects (CRE) model that gets around this problem by assuming a specific functional relationship between α_i and \mathbf{x}_i . That is,

$$\alpha_i = \sum_{t=1}^T \mathbf{a}'_t \mathbf{x}_{it} + \eta_i = \mathbf{a}' \mathbf{x}_i + \eta_i$$

where $\mathbf{a}' = (\mathbf{a}'_1, \dots, \mathbf{a}'_T)$, $\mathbf{x}'_i = (\mathbf{x}'_{i1}, \dots, \mathbf{x}'_{iT})$, and η_i is normally distributed (with mean zero and variance σ_η^2) and is independent of the \mathbf{x}_{it} .

- The \mathbf{a}_t are vectors of parameters to be estimated and capture the nature of the relationship between α_i and \mathbf{x}_{it} .
- Equation (33.1) then becomes

$$y_{it}^* = \boldsymbol{\beta}' \mathbf{x}_{it} + \sum_{t=1}^T \mathbf{a}'_t \mathbf{x}_{it} + \varepsilon_{it}$$

where $\varepsilon_{it} = \eta_i + u_{it}$.

- Estimation of the CRE model proceeds in a sequential manner:
 1. Estimate separate probit equations by maximum likelihood for each time period, regressing the dependent variable in each period on all of the leads and lags of the explanatory variables.
 2. Stack the estimates from each of these probits into a vector $\hat{\pi}$ and construct the joint covariance matrix of all of the estimates. The vector $\hat{\pi}$ is a vector of reduced form estimates.
 3. Use a minimum distance estimator to impose restrictions on $\hat{\pi}$ to back out estimates of the structural parameters β and \mathbf{a}_t . Let $\theta = (\beta', \mathbf{a}')$ and choose θ to minimize

$$[\hat{\pi} - \mathbf{f}(\theta)]' \hat{\Omega}^{-1} [\hat{\pi} - \mathbf{f}(\theta)],$$

where $\hat{\Omega}$ is an estimate of the asymptotic variance-covariance matrix for the reduced-form estimates.

4. We then conduct our standard hypothesis tests to make inferences about the effects of the variables of interest.
5. Marginal effects can be determined by simulating probabilities:

$$\Pr(y_{it} = 1) = \Phi \left[(1 + \sigma_\eta^2)^{-1/2} (\beta' \mathbf{x}_{it} + \sum_{t=1}^T \mathbf{a}'_t \mathbf{x}_{it}) \right] \quad (33.6)$$

- One key advantage of the CRE estimator: allows for an explicit test for correlation between the individual specific effect and the explanatory variables.
- One drawback of the CRE estimator: requires us to impose a good deal of structure on the relationship between α_i and \mathbf{x}_{it} and restricts us to including only time-varying variables in \mathbf{x}_{it} .
- Including time invariant variables in \mathbf{x}_{it} in effect induces perfect collinearity since the values of the leads and the lags of these variables will be the same within each cross-sectional unit.
- Its drawbacks aside, this estimator is part of a class of generalized method of moments estimators that give substantial efficiency gains over pooled probit estimators, which ignore individual specific effects.
- For details of the estimation procedure for the CRE model, see Hsiao's treatment in *Analysis of Panel Data*, which is more accessible than Chamberlain's original derivation.
- Can be estimated w/ standard software with some more restrictive assumptions; **GAUSS** code is available for the general model.

33.4 CRE Probit Application: PAC contributions and roll call votes

- A major concern is the effects of campaign contributions on the behavior of members of the U.S. Congress.
- Assessing these effects is complicated because it is methodologically difficult to account for members' predispositions to vote in favor of PACs' interests.
- CRE probit model can help overcome this problem because it enables us to account for individual specific effects, such as the predisposition to vote for or against a particular piece of legislation, which are too costly or impossible to measure.
- I conducted a CRE probit analysis of roll-call votes and PAC contributions using sets of roll-call votes that are of particular interest to certain types of PACs (Wawro '01 *AJPS*).
 - Tables 3 and 4 report results of the CRE estimation.
 - Figures 2 & 4 plot simulated probabilities based on the CRE estimates.

TABLE 3 Panel Probit Model Results for Voting Behavior on AFL-CIO Roll-Call Votes, 102nd–104th Congresses

Variable	102nd Congress		103rd Congress		104th Congress	
	1st Session	2nd Session	1st Session	2nd Session	1st Session	2nd Session
	<i>Estimated coefficients*</i>					
In Labor PAC contributions	-0.006 (0.004)	0.002 (0.011)	-0.012 (0.009)	0.010 (0.006)	0.008 (0.001)	0.012 (0.009)
In Corporate PAC contributions	0.031 (0.003)	0.036 (0.011)	0.006 (0.007)	0.010 (0.006)	0.018 (0.002)	0.030 (0.011)
Unemployment rate	-0.003 (0.004)	-0.008 (0.011)	0.001 (0.008)	-0.006 (0.007)	-0.011 (0.002)	-0.025 (1.617)
	<i>Tests of no correlation between individual effect and regressors†</i>					
In Labor PAC contributions	923.348 (< 0.001)	371.868 (< 0.001)	342.397 (< 0.001)	481.322 (< 0.001)	4,384.273 (< 0.001)	159.388 (< 0.001)
In Corporate PAC contributions	170.824 (< 0.001)	33.342 (< 0.001)	106.950 (< 0.001)	32.075 (< 0.001)	3,986.840 (< 0.001)	102.342 (< 0.001)
Unemployment rate	112.456 (< 0.001)	3.222 (.666)	117.859 (< 0.001)	31.739 (< 0.001)	892.259 (< 0.001)	12.804 (.025)
N	315	319	356	332	324	343
T	8	5	7	7	9	5

Note: *Standard errors in parentheses.

† Entries are χ^2_T statistics. *p* values in parentheses.

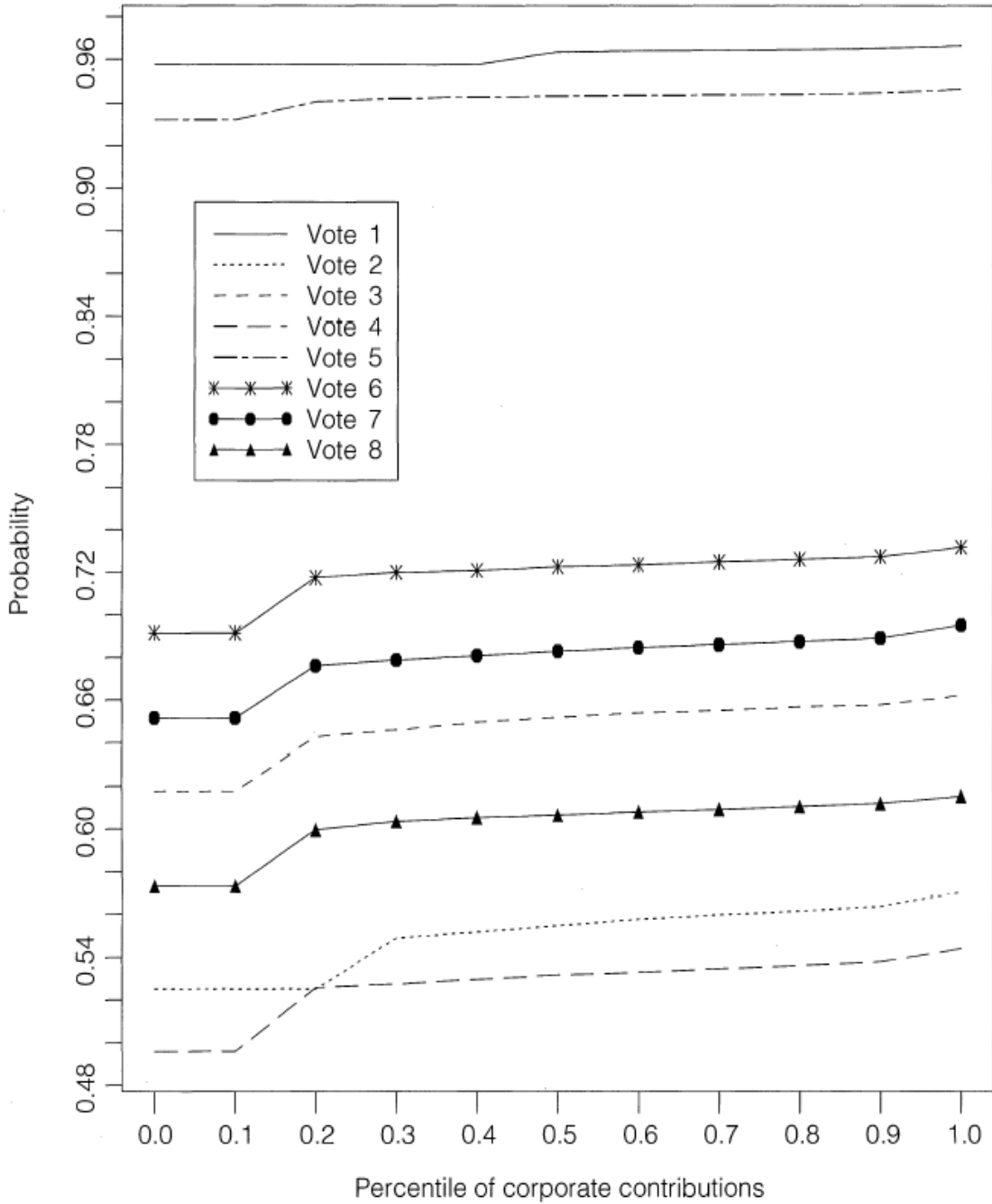
TABLE 4 Panel Probit Model Results for Voting Behavior on USCC Roll-Call Votes, 102nd–104th Congresses

Variable	102nd Congress		103rd Congress		104th Congress	
	1st Session	2nd Session	1st Session	2nd Session	1st Session	2nd Session
	<i>Estimated coefficients*</i>					
In Labor PAC contributions	0.025 (0.004)	-0.005 (0.007)	-0.012 (0.003)	0.043 (0.011)	0.008 (0.001)	0.022 (0.004)
In Corporate PAC contributions	-0.030 (0.005)	0.004 (0.006)	0.003 (0.004)	0.010 (0.009)	-0.011 (0.001)	0.015 (0.004)
Unemployment rate	0.002 (0.005)	-0.010 (0.007)	0.023 (0.004)	0.018 (0.011)	0.004 (0.001)	0.002 (0.005)
	<i>Tests of no correlation between individual effect and regressors†</i>					
In Labor PAC contributions	676.705 (< 0.001)	267.680 (< 0.001)	119.077 (< 0.001)	55.355 (< 0.001)	6,441.995 (< 0.001)	606.078 (< 0.001)
In Corporate PAC contributions	459.730 (< 0.001)	189.141 (< 0.001)	87.816 (< 0.001)	37.028 (< 0.001)	16,767.967 (< 0.001)	387.557 (< 0.001)
Unemployment rate	106.871 (< 0.001)	223.512 (< 0.001)	82.972 (< 0.001)	15.364 (0.031)	7,427.728 (< 0.001)	237.892 (< 0.001)
N	302	288	357	344	340	338
T	8	7	7	7	10	8

Note: *Standard errors in parentheses.

† Entries are χ^2_T statistics. *p* values in parentheses.

FIGURE 4 Effect of Corporate PAC Contributions on the Probability of Voting in Favor of the USCC Positions, 104th Congress, 2nd Session



Section 34

Binary Time-Series Cross-Section (BTSCS) Data

- The methods above are appropriate when N is large and T is small. Beck, Katz, and Tucker ('98 *AJPS*) derive a method for when T is large.
- The method is based on the observation that BTSCS data is identical to grouped duration data. That is, we get to observe whether an event occurred or not only after the end of some discrete period (e.g., a year).
- Thus, we can use duration methods to correct for the problem of temporal dependence.
- Start from the hazard rate for the continuous time Cox proportional hazard model:

$$\lambda(t) = \exp(\boldsymbol{\beta}'\mathbf{x}_{it})\lambda_0(t)$$

- The survival function is given by

$$S(t) = \exp\left(-\int_0^t \lambda(\tau)d\tau\right)$$

- Assuming we get to observe only whether or not an event occurred between time $t_k - 1$ and t_k , we can write

$$\begin{aligned}\Pr(y_{it_k} = 1) &= 1 - \exp\left(-\int_{t_k-1}^{t_k} \lambda_i(\tau) d\tau\right) \\ &= 1 - \exp\left(-\int_{t_k-1}^{t_k} \exp(\boldsymbol{\beta}' \mathbf{x}_{it}) \lambda_0(t) d\tau\right) \\ &= 1 - \exp\left(-\exp(\boldsymbol{\beta}' \mathbf{x}_{it}) \int_{t_k-1}^{t_k} \lambda_0(t) d\tau\right)\end{aligned}$$

- Let

$$\begin{aligned}\alpha_{t_k} &= \int_{t_k-1}^{t_k} \lambda_0(t) d\tau \\ \kappa_{t_k} &= \ln(\alpha_{t_k})\end{aligned}$$

- Then

$$\begin{aligned}\Pr(y_{it_k} = 1) &= 1 - \exp\left(-\exp(\boldsymbol{\beta}' \mathbf{x}_{it}) \alpha_{t_k}\right) \\ &= 1 - \exp\left(-\exp(\boldsymbol{\beta}' \mathbf{x}_{it} + \kappa_{t_k})\right)\end{aligned}$$

- This is a binary model with a complimentary log-log (cloglog) link. The cloglog link is identical to a logit link function when the probability of an event is small ($< 25\%$) and extremely similar when the probability of an event is moderate ($< 50\%$).

- For ease of application then, Beck, Katz, and Tucker recommend using the logistic analogue

$$\Pr(y_{it} = 1 | \mathbf{x}_{it}) = \frac{1}{1 + \exp(-(\boldsymbol{\beta}' \mathbf{x}_{it} + \kappa_{t-t_0}))}$$

where κ_{t-t_0} is a dummy variable marking the length of the sequence of zeros that precede the current observation. For example,

t	1	2	3	4	5	6	7	8	9
y	0	0	0	1	0	1	1	0	0
κ	κ_1	κ_2	κ_3	κ_4	κ_1	κ_2	κ_1	κ_1	κ_2

- The intuition behind why ordinary logit is inadequate for BTSCS data is that it doesn't allow for a nonconstant baseline hazard.
- Including the κ dummies allows duration dependence by allowing for a time-varying baseline hazard.
- To see how the κ dummies are interpretable as baseline probabilities or hazards, note

$$\Pr(y_{it} = 1 | \mathbf{x}_{it} = 0, t_0) = \frac{1}{1 + \exp(-\kappa_{t-t_0})}$$

- The κ dummies are essentially time fixed effects that account for duration dependence. Thus when we estimate the model we need to create a matrix of dummies and concatenate it with the matrix of explanatory variables. For the example given above, this matrix would look like

$$\mathbf{K}_i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} .$$

Note there are 4 columns because the longest spell is 4 periods long.

$$\Pr(y_{it} = 1 | \mathbf{x}_{it}) = \frac{1}{1 + \exp(-(\boldsymbol{\beta}' \mathbf{x}_{it} + \kappa_{t-t_0}))}$$

Section 35

Generalized Estimating Equations (GEEs)

- This class of models allows us to account for unobserved correlation among observations, without including unit-specific effects.
- GEEs relax assumptions about the independence of observations: observations are grouped into clusters and then parameters are estimated to model the correlations among observations in the cluster.
- This class of models is derived from the General Linear Model (GLM) approach.
 - Models in which one specifies the “link” function $E(Y_i) = \mu_i = h(\boldsymbol{\beta}'\mathbf{x}_i)$ and the relationship between the mean and the variance (e.g. $\mathbf{V}_i = g(\mu_i)$).
 - Standard GLMs obtain estimates by solving the “score equations”:

$$\mathbf{U}_k(\boldsymbol{\beta}) = \sum_{i=1}^N \mathbf{D}_i \mathbf{V}_i^{-1} (Y_i - \mu_i) = 0 \quad (35.1)$$

where $\mathbf{D}_i = \frac{\partial \mu_i}{\partial \boldsymbol{\beta}}$ and \mathbf{V}_i is the variance matrix.

35.1 GLMs for Correlated Data

- Consider $E(Y_{it}) = \mu_{it} = g(\boldsymbol{\beta}'\mathbf{x}_{it})$, $T > 1$.
- Must make some provision for dependence within i , across t .
- Specify the conditional within-unit or within-cluster correlation:
- Define the “working” $T \times T$ correlation matrix $\mathbf{R}_i(\alpha)$ as a function of α .
- Structure (but not the elements) of $\mathbf{R}_i(\alpha)$ determined by the investigator.
- Then redefine the variance matrix in Equation 35.1 as:

$$\mathbf{V}_i = \frac{(\mathbf{A}_i)^{\frac{1}{2}}\mathbf{R}_i(\alpha)(\mathbf{A}_i)^{\frac{1}{2}}}{\phi} \quad (35.2)$$

where the \mathbf{A}_i are $T \times T$ diagonal matrices w/ $g(\mu_{it})$ along the diagonal.

- Intuition:
 - Choose β so that μ_{it} is “close” to Y_{it} on average,
 - Optimally weight each residual $(Y_{it} - \mu_{it})$ by the inverse of $\text{cov}(Y_i)$.

35.2 Options for specifying within-cluster correlation

- “Working Independence”: $\mathbf{R}_i(\alpha) = \mathbf{I}$
 - No within-unit correlation (completely pooling observations).
 - Just like standard logit/probit.
- “Exchangeable”: $\mathbf{R}_i(\alpha) = \rho$
 - Within-unit correlation is the same across units.
 - Similar to “random effects,” in the sense that we’re assuming constant within-unit marginal covariance; but do not need orthogonality assumption of RE.
- Autoregressive : $\mathbf{R}_i(\alpha) = \rho^{|t-s|}$
 - Here, AR(1).
 - Correlation decays over “time.”
 - Can also do “banded” / “stationary” correlations.
- “Unstructured” : $\mathbf{R}_i(\alpha) = \alpha_{st}, t \neq s$
 - α is a $T \times T$ matrix.
 - Allows $\frac{T(T-1)}{2}$ unique pairwise correlations.
 - Very flexible, but can be hard to estimate, esp. w/ large T .

35.3 “Robust” Standard Errors

- $\hat{\beta}_{GEE}$ is robust to misspecification of $\mathbf{R}_i(\alpha)$.

- $\widehat{\text{var}}(\hat{\beta}_{GEE})$ is not.

- Can compute “sandwich” estimator: $\widehat{\text{var}}(\hat{\beta}_{GEE}) =$

$$N \left(\sum_{i=1}^N \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i \right)^{-1} \left(\sum_{i=1}^N \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{S}}_i \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i \right) \left(\sum_{i=1}^N \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i \right)^{-1} \quad (35.3)$$

where $\hat{\mathbf{S}}_i = (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)'$.

- Similar to the Huber/White estimator.
- One drawback of GEE is that it does not produce standard errors for elements of $\mathbf{R}_i(\alpha) \Rightarrow$ we should be cautious about drawing inferences about these parameters (treated as “nuisance”).
- A number of software packages can estimate GEEs: **xtgee** in **Stata**, **proc genmod w/ repeated option** in **SAS**, **gee**, **geepack** in **S-Plus/R**, **GEE** in Riemann Library for **GAUSS**.

35.4 GEE2

- Other options exist for estimating the $m = \frac{T(T-1)}{2}$ elements of $\mathbf{R}_i(\alpha)$.
- Can do this with a separate estimating equation:

$$\mathbf{U}_m(\alpha) = \sum_{i=1}^N \mathbf{E}'_i \mathbf{W}_i^{-1} (\mathbf{Z}_i - \boldsymbol{\eta}_i) \quad (35.4)$$

where

- $\mathbf{Z}'_i = (Z_{i12}, Z_{i13}, \dots, Z_{i1T}, Z_{i23}, \dots, Z_{iT,(T-1)})$; the $\frac{T(T-1)}{2}$ “observed” sample pairwise correlation.
 - $\boldsymbol{\eta}_i$ is the column vector of the expected values of the pairwise intracluster correlation for observation i .
 - $\mathbf{E}_i = \frac{\partial \boldsymbol{\eta}_i}{\partial \alpha_i}$
 - \mathbf{W}_i is a square diagonal matrix of rank $\frac{T(T-1)}{2}$ containing the variances and covariances of the \mathbf{Z}_i s.
- Can be estimated either separately from $\mathbf{U}_k(\boldsymbol{\beta})$, assuming $\text{cov}[U_k(\boldsymbol{\beta}), U_m(\alpha)] = 0$, or allowing the two to covary.
 - Drawbacks: Requires correct specification of $\mathbf{R}_i(\alpha)$ for consistent estimates of $\hat{\boldsymbol{\beta}}$.

35.5 Application: Panel Decision-making on the Court of Appeals

- Question of interest: do gender and race affect the decisions of judges on the U.S. Court of Appeals?
- Almost all cases are decided by (more or less) randomly selected panels of three judges.
- “Norm of unanimity” suggests there will be significant correlation among judges on the same panel.
- GEE approach is useful for accounting for correlation, treating each three judge panel as a cluster.
- Farhang and Wawro (*JLEO* '04) use this approach to examine decisions in employment discrimination cases.

Table 35.1: GEE analysis of judges' votes in Appeals Court decisions

Variable	Logit		GEE	
	Coefficient	Std. Err.	Coefficient	Std. Err.
Judge and Panel Level Variables				
Intercept	1.126	0.362	1.226	0.613
Gender	0.877	0.269	0.791	0.276
One female colleague (female judge)	-0.373	0.454	-0.398	0.594
One female colleague (male judge)	0.800	0.180	0.789	0.274
Two female colleagues (male judge)	-0.337	0.575	-0.362	0.595
Race	-0.329	0.363	-0.212	0.359
One nonwhite colleague (nonwhite judge)	-0.033	0.558	-0.458	0.764
One nonwhite colleague (white judge)	-0.136	0.236	-0.151	0.357
Two nonwhite colleagues (white judge)	-0.443	0.720	-0.859	1.081
NOMINATE score	-0.570	0.192	-0.643	0.176
Panel colleagues' NOMINATE scores	-1.098	0.243	-1.100	0.342
Author	0.122	0.188	-0.023	0.040
Gender \times author	-0.174	0.421	0.054	0.120
Race \times author	0.345	0.549	0.069	0.117
NOMINATE score \times author	-0.173	0.338	0.069	0.082
Case Specific Variables				
Race discrimination	0.285	0.197	0.294	0.310
Gender discrimination	-0.190	0.185	-0.171	0.308
Harassment	0.761	0.206	0.717	0.354
Age discrimination	-0.062	0.190	-0.047	0.330
Religious discrimination	-0.752	0.624	-0.715	1.138
Nationality discrimination	0.069	0.468	0.045	0.784
Reverse gender discrimination	0.576	0.391	0.487	0.716
Reverse race discrimination	-1.282	0.609	-1.388	1.016
Government defendant	0.139	0.159	0.091	0.268
EEOC plaintiff	0.150	0.620	-0.083	1.423
Plaintiff appeal	-1.597	0.235	-1.570	0.390
Both appeal	0.578	0.282	0.586	0.486
Posture	-0.079	0.207	-0.093	0.316
Circuit Level Variables				
1st Circuit dummy	-1.048	0.409	-1.102	0.641
2nd Circuit dummy	-0.624	0.373	-0.636	0.650
3rd Circuit dummy	-1.358	0.549	-1.386	0.773
4th Circuit dummy	-1.885	0.578	-1.973	0.839
5th Circuit dummy	-2.053	0.400	-2.038	0.688
6th Circuit dummy	0.093	0.397	-0.023	0.703
7th Circuit dummy	-1.112	0.320	-1.156	0.541
8th Circuit dummy	-0.870	0.326	-0.923	0.564
10th Circuit dummy	-1.315	0.377	-1.344	0.657
11th Circuit dummy	-0.207	0.334	-0.252	0.583
D.C. Circuit dummy	-1.410	0.489	-1.350	0.794
$\hat{\rho}$	—	—	0.893	—

Note: Table entries are GEE estimates. $N = 1200$.