POLS W4912
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

This course draws liberally on lecture notes prepared by Professors Neal Beck, Lucy Goodhart, George Jakubson, Nolan McCarty, and Chris Zorn. The course also draws from the following works:

Aldrich John H. and Forrest D. Nelson. 1984. *Linear Probability, Logit and Probit Models*. Beverly Hills, CA: Sage.

Alvarez, R. Michael and Jonathan Nagler. 1995. "Economics, Issues and the Perot Candidacy: Voter Choice in the 1992 Presidential Election." *American Journal of Political Science* 39:714-744

Amemiya, Takeshi. 1985 *Advanced econometrics*. Cambridge: Harvard University Press.

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-SeriesCross-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.

Herron, Michael. 2000. "Post-Estimation Uncertainty in Limited Dependent Variable Models" *Political Analysis* 8: 8398.

Hsiao, Cheng. 2003. *Analysis of Panel Data.* 2nd ed. Cambridge: Cambridge University Press.

Keifer, Nicholas M. 1988. Economic duration data and hazard functions. *Journal of Economic Literature* 24: 646–679.

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.

Yamaguchi, Kazuo. 1991. *Event History Analysis.* Newbury Park, CA: Sage.

Zorn, Christopher J. W. 2001. "Generalized Estimating Equations Models for Correlated Data: A Review With Applications." *American Journal of Political Science* 45: 470–90.

TABLE OF CONTENTS

LIST OF FIGURES

Part I

The Classical Linear Regression Model

Section 1

Probability Theory, Estimation, and Statistical Inference as a Prelude to the Classical Linear Regression Model

1.1 Why is this review pertinent?

1. Most of the phenomena that we care about in the world can be characterized as random variables, meaning that we can conceive of their values as being determined by the outcome of a chance experiment. To understand the behavior of a random variable, we need to understand probability, which assigns a likelihood to each realization of that random variable.

2. In univariate statistics, much of the focus is on estimating the expectation (or mean) of a random variable. In multivariate analysis, we will focus on relating movements in something we care about to variables that can explain it—but we will still care about expectations.

3. We will focus on establishing the **conditional expectation** of a dependent variable Y. In other words, given the value of X (where X may be a single variable or a set of variables) what value do we expect Y to take on? That conditional expectation is often written as $\boldsymbol{\beta}X$.

4. Since we do not directly observe all the data that make up the world (and/or cannot run experiments), we must estimate $\boldsymbol{\beta}X$ using a sample. To understand what we can say about that estimate, we will have to employ statistical inference.

5. Finally, many of the statistics we estimate (where a statistic is simply any quantity calculated from sample data) will conform to particular known probability distributions. This will simplify the business of conducting hypothesis tests. To assist in understanding the rationale behind those hypothesis tests, it helps to review the distributions in question.

1.2 Probability theory

- The set of all possible outcomes of a random experiment is called the population or sample space and an event is one cell (or subset) of the sample space.
- If events cannot occur jointly, they are *mutually exclusive*. If they exhaust all of the things that could happen they are *collectively exhaustive*.

1.2.1 Properties of probability

1. $0 \leq P(A) \leq 1$ for every A.
2. If A, B, C constitute an exhaustive set of events, then $P(A + B + C) = 1$.
3. If A, B, C are mutually exclusive events, then $P(A + B + C) = P(A) + P(B) + P(C)$.

1.2.2 Random variables

- Definition: a variable whose value is determined by a chance experiment.
- Convention: a rv is denoted by an upper case letter (e.g., X) and realizations of a rv are denoted by lower case letters (e.g., x_1, x_2, ..., x_n).
- A random variable is **discrete** if it takes on a finite set (or countably infinite set) of values. Otherwise the rv is **continuous**.
- A **probability density function (PDF)** assigns a probability to each value (or event or occurrence) of a random variable X.
- Thus, a discrete PDF for a variable X taking on the values $x_1, x_2, x_3, \ldots, x_n$ is a function such that:

$$f_X(x) = P[X = x_i] \qquad \text{for } i = 1, 2, 3, \ldots, n.$$

and is zero otherwise (there is no probability of X taking on any other value).

- A continuous probability density function for a variable X that covers a continuous range is a function such that:

$$f(x) \geq 0 \text{ for all } x$$

and

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

$$P[a < X < b] = \int_a^b f(x)dx$$

- The **cumulative probability density function (CDF)** gives the probability of a random variable being less than or equal to some value: $F(x) = P[X \leq x]$.

- For a discrete rv

$$F(x) = \sum_{x_j \leq x} f(x_j).$$

- For a continuous rv

$$\begin{aligned} F(x) &= P[X \leq x] \\ &= P[-\infty \leq X \leq x] \\ &= \int_{-\infty}^{x} f(u)du. \end{aligned}$$

Figure 1.1: An Example of PDF

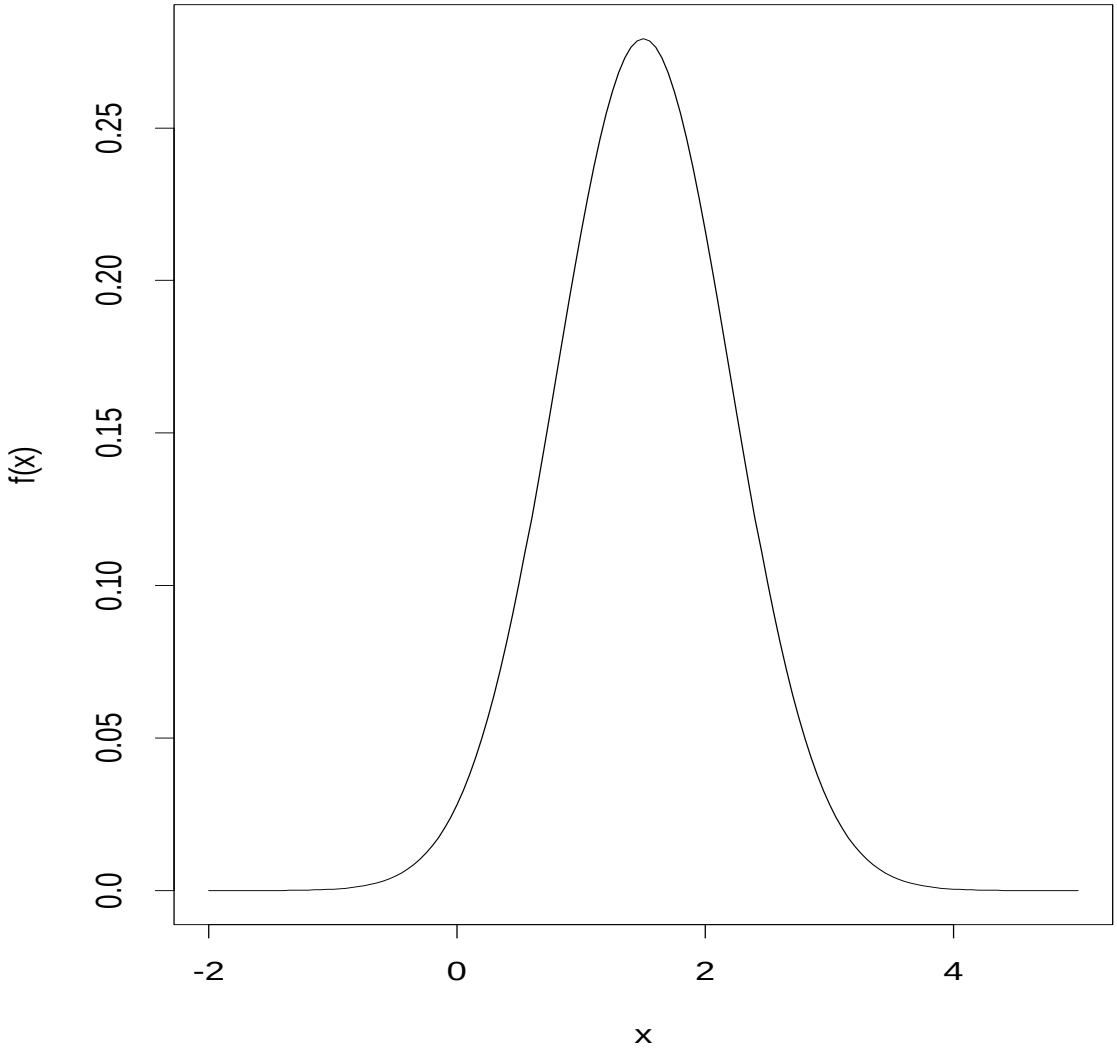
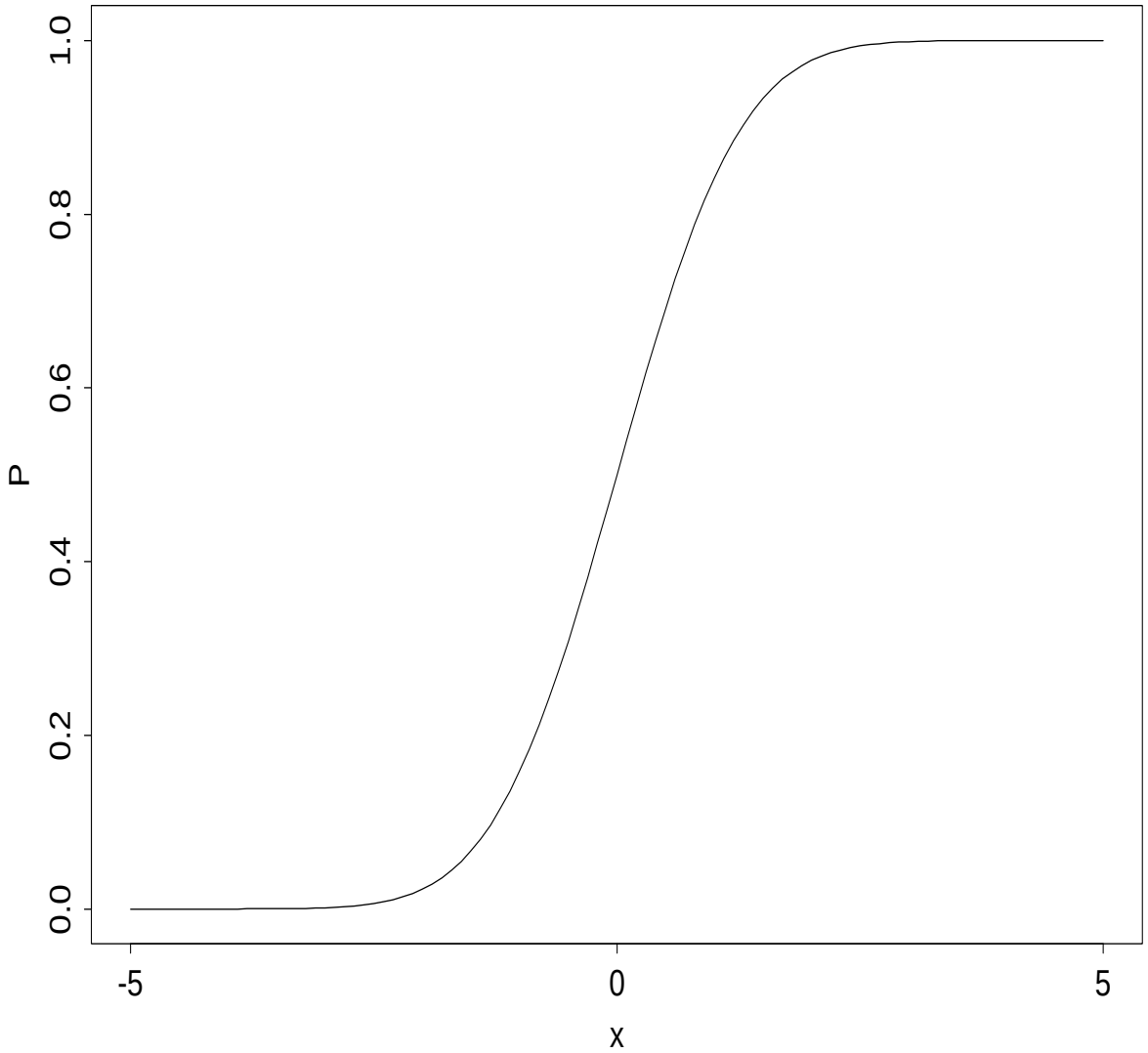


Figure 1.2: An example of a CDF



1.2.3 Joint probability density functions

- For all of the political science examples we will talk about, we will be interested in how one variable is related to another, so we will be interested in **joint probability density functions**:

$$f(x, y) = P[X = x \text{ and } Y = y]$$

- See Fig. 1.3 for an example of a joint PDF.
- The marginal probability density function, $f_X(x)$, can be derived from the joint probability density function, $f_{X,Y}(x, y)$, by summing/integrating over all the different values that Y could take on.

$$f_X(x) = \sum_y f_{X,Y}(x, y) \quad \text{if } f_{X,Y} \text{ is discrete} \tag{1.1}$$
$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy \quad \text{if } f_{X,Y} \text{ is continuous} \tag{1.2}$$

1.2.4 Conditional probability density functions

- Since we are interested in understanding how one variable is related to another, we will also consider the conditional PDF, which gives, for example, the probability that X has the realization x given that Y has the realization y. The **conditional PDF** is written as:

$$f_{X,Y}(x|y) = P(X = x|Y = y) \tag{1.3}$$
$$= \frac{f_{X,Y}(x, y)}{f_Y(y)} \tag{1.4}$$

- **Statistical Independence**: We say that X and Y are statistically independent if and only if

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$$

- See Fig. 1.4 for an example of a condit'l PDF and Fig. 1.5 for an example of a condit'l PDF for two independent variables.

Figure 1.3: Joint PDF

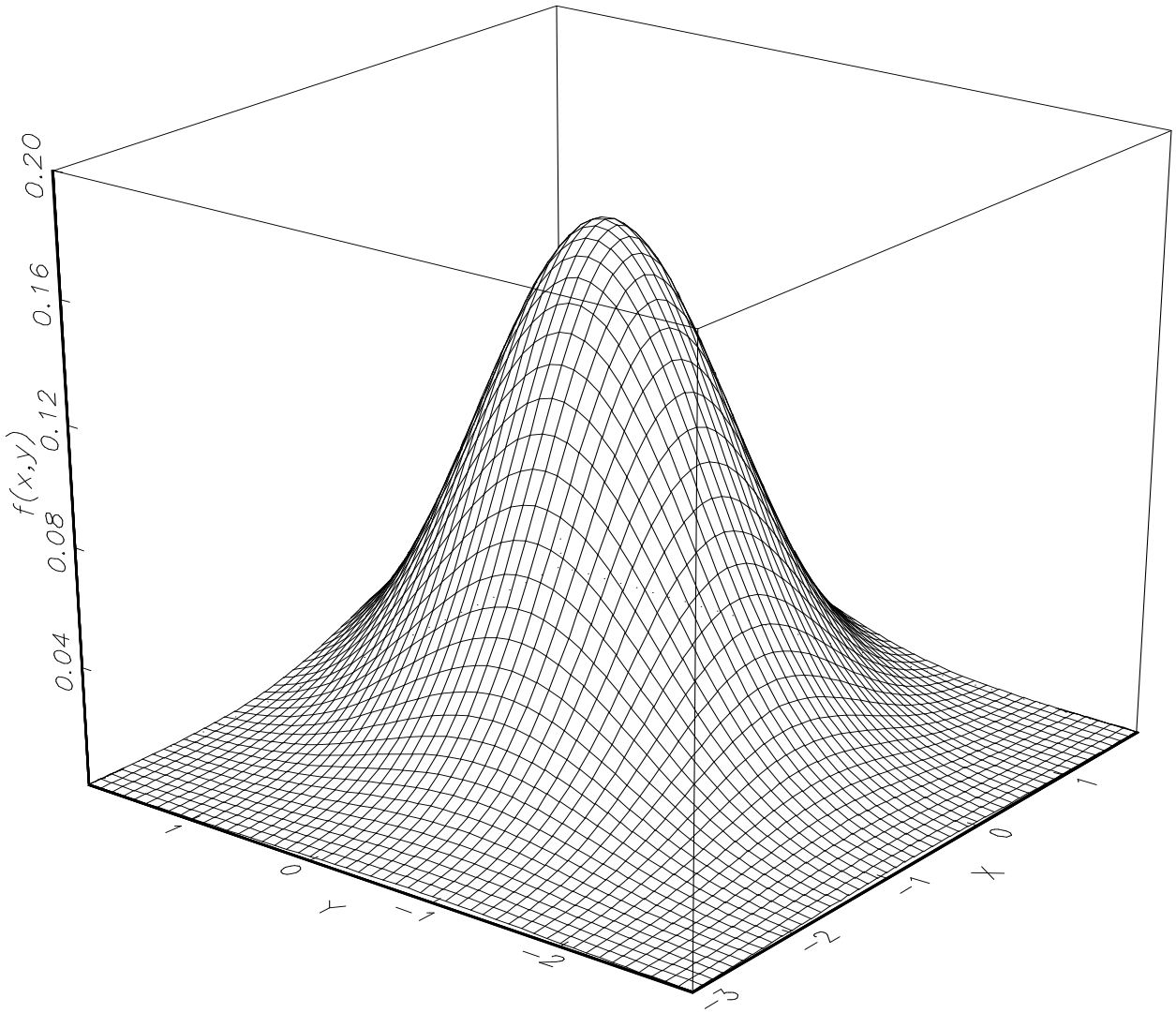


Figure 1.4: Conditional PDF

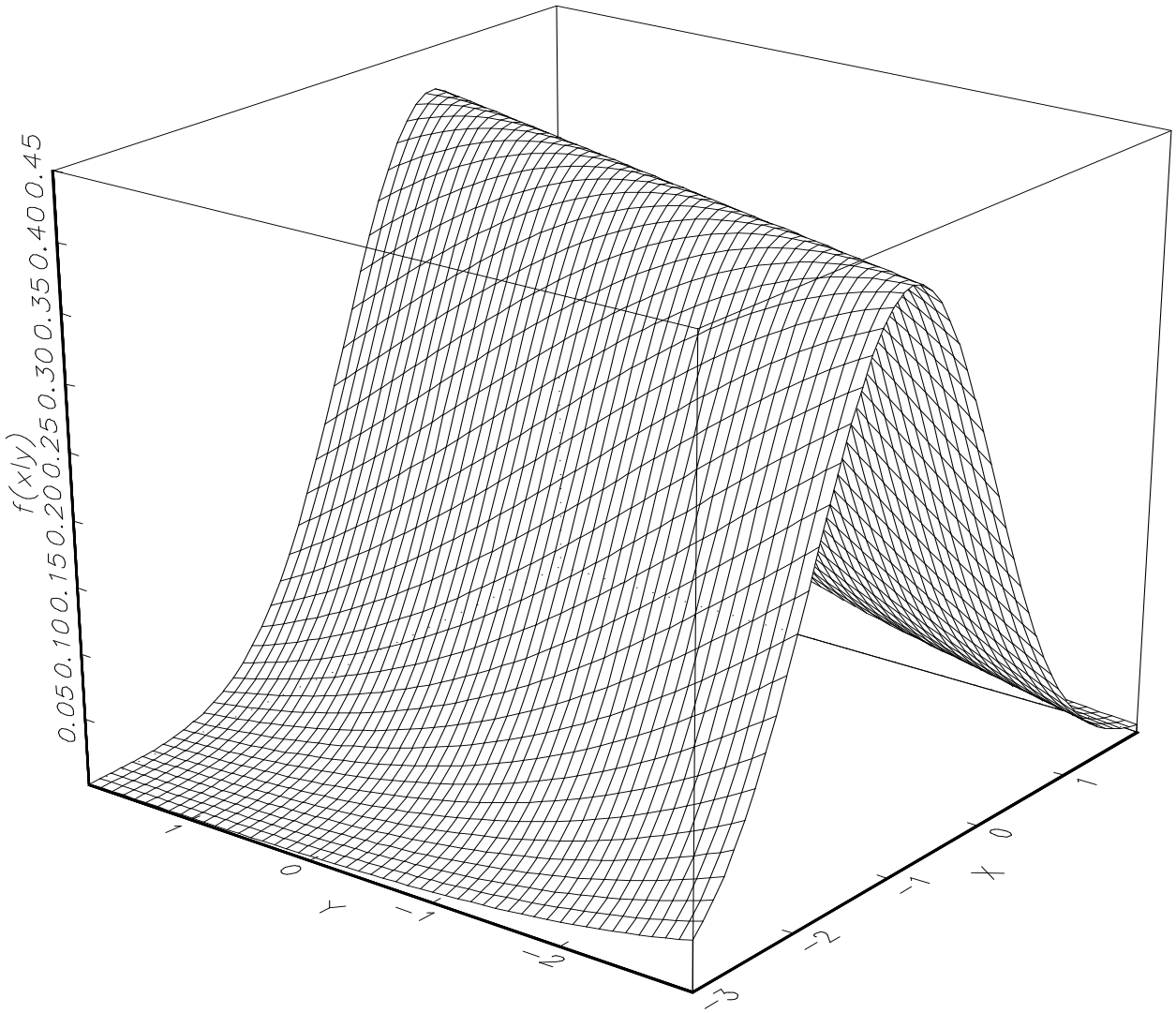
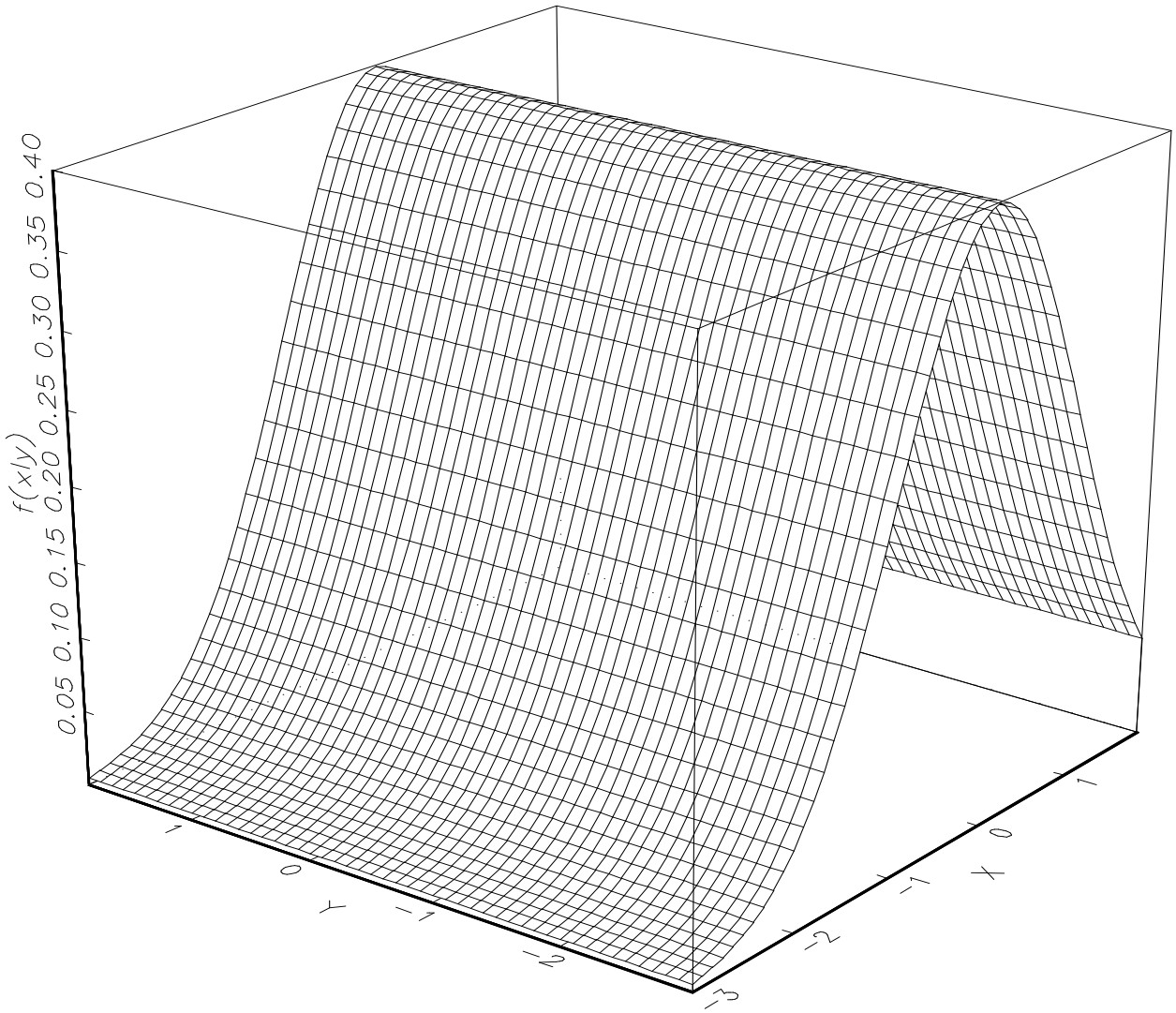


Figure 1.5: Conditional PDF, X and Y independent



1.3 Expectations, variance, and covariance

- We will often, want to know about the central tendency of a random variable. We are particularly interested in the expected value of the random variable, which is the average value that it takes on over many repeated trials.
- The expected value (μ), or population mean, is the first moment of X.

$$E[X] = \sum_{j=1}^{n} x_j f(x_j) \tag{1.5}$$

if X is discrete and has the possible values $x_1, x_2, \ldots, x_n$, and

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx \tag{1.6}$$

if X is continuous.

1.3.1 Properties of expected values

- If b is a constant, then $E(b) = b$.
- If a and b are constants, then $E(aX + b) = aE(X) + b$.
- If X and Y are independent, then $E(XY) = E(X)E(Y)$
- If X is a random variable with a PDF $f(x)$ and if $g(x)$ is any function of X, then

$$E[g(x)] = \begin{cases} \sum_x g(x)f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x)f(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

1.3.2 Variance

- The distribution of values of X around its expected value can be measured by the variance, which is defined as:

$$\text{var}(X) = \sigma_x^2 = E[(X - \mu)^2]$$

- The variance is the second moment of X. The **standard deviation** of X, denoted σ_X, is $+\sqrt{\text{var}[X]}$.

- The variance is computed as follows:

$$E[g(x)] = \begin{cases} \sum_x \left[(X-\mu)^2\right] f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} \left[(X-\mu)^2\right] f(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

- We can also express the variance as

$$\begin{aligned} \text{var}[X] &= E[(X - \mu_X)^2] \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

1.3.3 Properties of variance:

- If b is a constant, then $\text{var}[b] = 0$.

- If a and b are constants, and $Y = a + bX$ then

$$\text{var}[Y] = b^2\text{var}[X].$$

- If X and Y are independent random variables, then
$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$ and $\text{var}(X - Y) = \text{var}(X) + \text{var}(Y)$.

1.3.4 Covariance

- The **covariance** of two rvs, X and Y, with means μ_X and μ_Y is defined as

$$\begin{aligned} \text{cov}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] && (1.7) \\ &= E[XY] - \mu_X \mu_Y. && (1.8) \end{aligned}$$

1.3.5 Properties of covariance

- If X and Y are independent, then $E[XY] = E[X]E[Y]$ implying $\text{cov}(X, Y) = 0$.
- $\text{cov}(a + bX, c + dY) = bd \cdot \text{cov}(X, Y)$, where a, b, c, and d are constants.

1.3.6 Correlation

- The size of the covariance will depend on the units in which X and Y are measured. This has led to the development of the correlation coefficient, which gives a measure of statistical association that ranges between -1 and $+1$.
- The population correlation coefficient, ρ, is defined as:

$$\rho = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y}$$

1.3.7 Variance of correlated variables

$$\begin{aligned} \text{var}(X + Y) &= \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y) \\ \text{var}(X - Y) &= \text{var}(X) + \text{var}(Y) - 2\text{cov}(X, Y) \end{aligned}$$

1.3.8 Conditional expectation

- The conditional expectation of X, given $Y = y$, is defined as

$$E(X|Y=y) = \begin{cases} \sum_x x f(x|Y=y) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f(x|Y=y)dx & \text{if } X \text{ is continuous} \end{cases}$$

- The **Law of Iterated Expectations** indicates how the expected value of X relates to the conditional expectation of X: $E(X) = E_y[E(X|Y)]$.

1.4 Important distributions

1.4.1 The normal distribution

- A random variable X is normally distributed (denoted $X \sim N(\mu, \sigma^2)$) if it has the PDF

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) \tag{1.9}$$

  where $-\infty < x < \infty$, μ is the mean, and σ^2 is the variance of the distribution.

- Properties of the Normal Distribution:
  - ➢ It is symmetric around its mean value (so mean=median=mode).
  - ➢ Approximately 68% of the area under a normal curve lies between the values of $\mu \pm \sigma$, about 95% of the area lies between $\mu \pm 2\sigma$, and about 99.7% in the range $\mu \pm 3\sigma$.
  - ➢ A normal distribution is completely characterized by its two parameters.

- Any normally distributed variable can be transformed into a standard normal variable by subtracting its mean and dividing by its standard deviation.

  ➢ E.g., if $Y \sim N(\mu_Y, \sigma_Y^2)$, and $Z = (Y - \mu_Y)/\sigma_Y$, then $Z \sim N(0, 1)$. That is,

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \tag{1.10}$$

1.4.2 χ^2 (Chi-squared) distribution

- Let $Z_1, Z_2, \ldots, Z_k$ be independent standard normal variables. Then

$$U = \sum_{i=1}^{k} Z_i^2$$

  has a chi-squared distribution with k degrees of freedom (denoted $U \sim \chi_k^2$).

1.4.3 Student's t distribution

- Suppose $Z \sim N(0, 1)$, $U \sim \chi_k^2$, and U and Z are distributed independently of each other. Then the variable

$$t = \frac{Z}{\sqrt{U/k}}$$

  has a t distribution with k degrees of freedom.

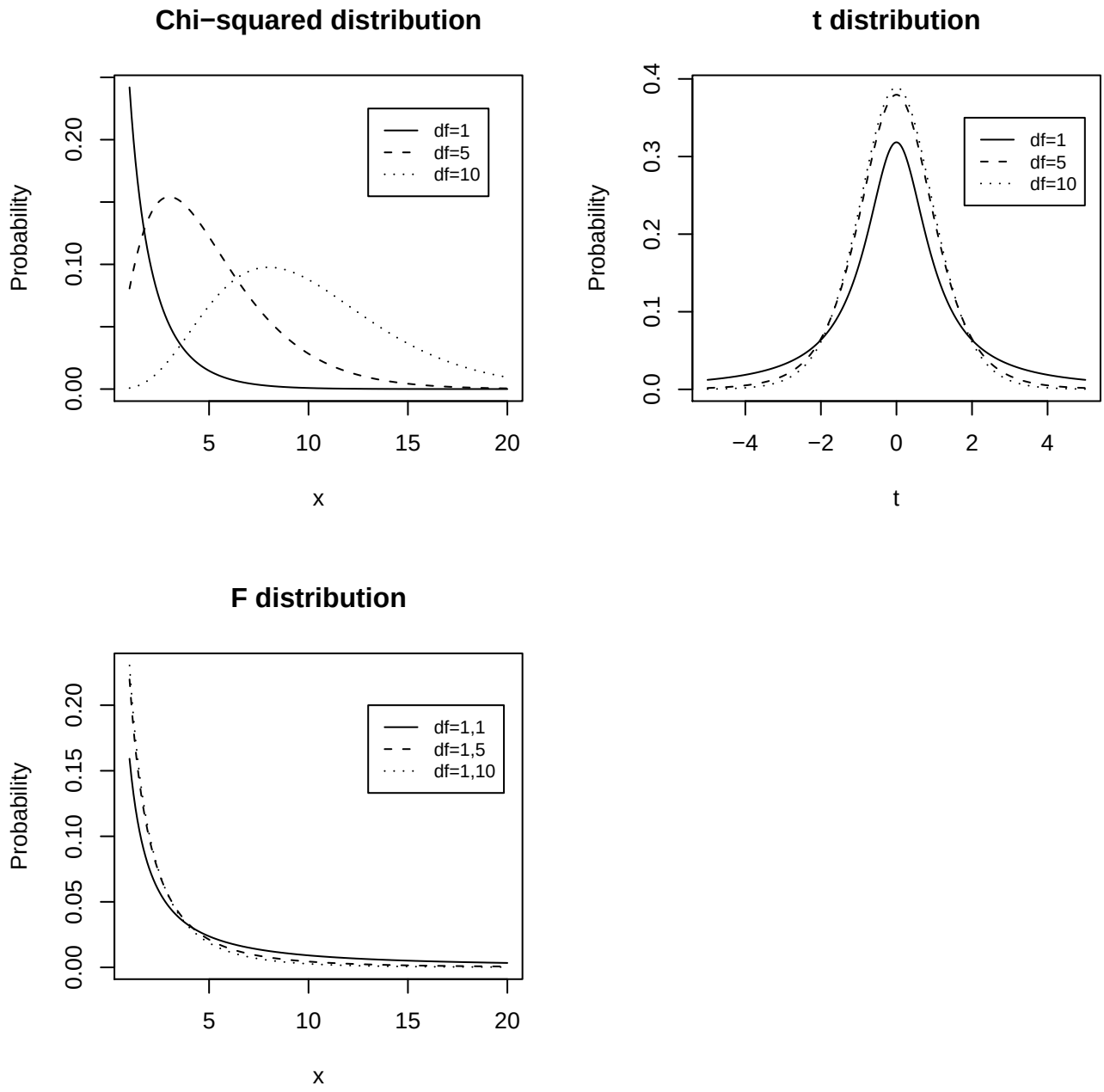
1.4.4 The F distribution

- The F distribution is the distribution of the ratio of two independent chi-squared random variables divided by their respective degrees of freedom.
- That is, if $U \sim \chi_m^2$ and $V \sim \chi_n^2$ then the variable

$$F = \frac{U/m}{V/n}$$

  has an F distribution with m and n degrees of freedom (denoted $F_{m,n}$).

Figure 1.6: Plots of special distributions



1.5 Statistical inference

- In the real world, we typically do not know the true probability distribution of our population of interest; estimate it using a sample. Use laws of statistics.
- If we don't know, for instance, the true distribution of income in the population, we could take a random sample of individuals.
- Each individual that we poll, or observation that we record, can be viewed as a random variable, $X_1, X_2, \ldots, X_n$, because they have an income whose level is in part the result of an experiment. Each of those random draws (or variables) will, however, have a known value, $x_1, x_2, \ldots, x_n$.
- If each comes from the same population (i.e., w/ same likelihood of being rich or poor) then the observations are said to be **identically distributed**. If the selection of one person does not affect the chances of selection of another person, they are said to be **independent**. If the observations are both, they are described as **iid**—standard assumption.
- Once we have our sample, we can estimate a sample mean,

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

  and a sample variance,

$$s_y^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2.$$

- What we want to know is how good these estimates are as descriptions of the real world $\rightarrow$ use statistical inference.
- We can show that the sample avg. is an unbiased estimator of the true population mean, μ_Y:

$$E(\bar{Y}) = \frac{1}{N} \sum_{i=1}^{N} E(Y_i) = \mu_y$$

- The sample avg. is a rv: take a different sample, likely to get a different average. The hypothetical distribution of the sample avg. is the **sampling distribution**.
- We can deduce the variance of this sample avg.:

$$\text{var}(\bar{Y}) = \frac{\sigma_y^2}{N}.$$

- If the underlying population is distributed normal, we can also say that the avg. is distributed normal (why?). If the underlying population does not have a normal distribution, however, & income distribution is not normal, we cannot infer that the sampling distribution is normal.
- Can use two laws of statistics that can tell us about the precision and distribution of our estimator (in this case the avg.):
  - ➢ The **Law of Large Numbers** states that, under general conditions, $\bar{Y}$ will be near μ_y with very high probability when N is large. The property that $\bar{Y}$ is near μ_y with increasing prob. as $N \uparrow$ is called **convergence in probability** or, more concisely, **consistency**. We will return to consistency in the section on asymptotics.
  - ➢ The **Central Limit Theorem** states that, under general conditions, the distribution of $\bar{Y}$ is well approximated by a normal distribution when N is large.
  - ➢ $\therefore$ (under general conditions) can treat $\bar{Y}$ as normally distributed and use the variance of the average to construct hypothesis tests (e.g., "how likely it is that avg. income $<$ \$20,000 given our estimate?").
  - ➢ Can also construct confidence intervals, outlining the range of income in which we expect the true mean income to fall with a given probability.
- We will estimate statistics that are analogous to a sample mean and are also distributed normal. We can again use the normal as the basis of hypothesis tests. We will also estimate parameters that are distributed as χ^2, as t, and as F.

Section 2

Matrix Algebra Review

2.1 Introduction

- A matrix is simply a rectangular array of numbers, like a table. Often, when we collect data, it takes the form of a table (or excel spreadsheet) with variable names in the columns across the top and observations in the rows down the side. That's a typical matrix.
- A vector (a single-columned matrix) can also be thought of as a coordinate in space.
- Matrix algebra is the topic that covers the mathematical operations we can do with numbers. These include addition, subtraction, multiplication, division (via the inverse), differentiation and factoring. In doing these matrix operations, we just apply the particular operation to the matrix as a whole rather than to a single number.
- Matrix algebra is sometimes also called linear algebra because one of the major uses of matrix algebra is to solve systems of linear equations such as the following:

$$\begin{aligned} 2x + 4y &= 8 \\ x + 5y &= 7 \end{aligned}$$

- Easy to substitute to get expressions for x (or y) and solve for y (or x). Harder to do for more equations—motivation for using matrix algebra.
- In regression analysis, estimation of coefficients requires a solution to a set of N equations with k unknowns.

2.2 Terminology and notation

- Matrices are represented w/ bold capital letters (e.g., $\mathbf{A}$)
- Vectors are represented w/ bold lower case letters (e.g., $\mathbf{a}$).
- Each element in a matrix, $\mathbf{A}$, is represented as a_{ij}, where i gives the row and j gives the column.
- **Transpose**: switch the row and column positions of a matrix (so that the transposed matrix has a_{ji} for its elements). Example:

$$\begin{bmatrix} 4 & 2 \\ 1 & 5 \end{bmatrix}' = \begin{bmatrix} 4 & 1 \\ 2 & 5 \end{bmatrix}$$

- **Scalar**: a 1×1 matrix.
- A matrix is equal to another matrix if every element in them is identical. This implies that the matrices have the same dimensions.

2.3 Types of matrices

- **Square**: the number of rows equals the numbers of columns.
- **Symmetric**: A matrix whose elements off the main left-right diagonal are a mirror-image of one another ($\Rightarrow \mathbf{A}' = \mathbf{A}$). Symmetric matrices must be square.
- **Diagonal**: All off-diagonal elements are zero.
- **Identity**: A square matrix with all elements on the main diagonal $= 1$, $= 0$ everywhere else. This functions as the matrix equivalent to the number one.
- **Idempotent**: a matrix $\mathbf{A}$ is idempotent if $\mathbf{A}^2 = \mathbf{A}$ ($\Rightarrow \mathbf{A}^n = \mathbf{A}$). For a matrix to be idempotent it must be square.

2.4 Addition and subtraction

- As though we were dropping one matrix over another. We are adding the two exactly together. For that reason, we can add matrices only of exactly the same dimensions and we find the addition by adding together each equivalently located element. Example:

$$\begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 5 \\ 2 & 4 \end{bmatrix}$$

  In subtraction, it is simply as though we are removing one matrix from the other, so we subtract each element from each equivalent element. Example:

$$\begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 0 & 2 \end{bmatrix}$$

- Properties of Addition:

$$\begin{gathered} \mathbf{A} + \mathbf{0} = \mathbf{A} \\ (\mathbf{A} + \mathbf{B}) = (\mathbf{B} + \mathbf{A}) \\ (\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}' \\ (\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C}) \end{gathered}$$

2.5 Multiplying a vector or matrix by a constant

- For any constant, k, $k\mathbf{A} = ka_{ij} \quad \forall \quad i, j$.

2.6 Multiplying two vectors/matrices

- To multiply two vectors or two matrices together, they must be conformable—i.e., row dimension of one matrix must match the column dimension of the other.
- **Inner-product multiplication**. For two vectors, $\mathbf{a}$ and $\mathbf{b}$, the inner-product is written as $\mathbf{a}'\mathbf{b}$.

  Let $\mathbf{a} = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 2 \\ 4 \\ 3 \end{bmatrix}$, then

  $$\mathbf{a}'\mathbf{b} = \begin{bmatrix} 1 & 2 & 4 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 4 \\ 3 \end{bmatrix} = (1)(2) + (2)(4) + (4)(3) = 22$$

- Note: $\mathbf{a}'$ is 1×3, $\mathbf{b}$ is 3×1, and the result is 1×1.
- Note: if $\mathbf{a}'\mathbf{b} = 0$, they are said to be **orthogonal**. ($\Rightarrow$ the two vectors in space are at right angles to each other).

2.7 Matrix multiplication

- Same as vector multiplication except that we treat the first row of the matrix as the transpose vector and continue with the other rows and columns in the same way. An example helps:

  $$\begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} (2)(1) + (4)(1) & (2)(2) + (4)(1) \\ (1)(1) + (3)(1) & (1)(2) + (3)(1) \end{bmatrix} = \begin{bmatrix} 6 & 8 \\ 4 & 5 \end{bmatrix}$$

- Again, number of columns in the first matrix must be equal to the number of the rows in the second matrix, otherwise there will be an element that has nothing to be multiplied with.
- Thus, for $\mathbf{A}_{nm} \cdot \mathbf{B}_{jk}$, $m = j$. Note also that

  $$\underset{n \times m}{\mathbf{A}} \cdot \underset{m \times k}{\mathbf{B}} = \underset{n \times k}{\mathbf{C}}$$

- Makes a difference which matrix is pre-multiplying and which is post-multiplying (e.g., $\mathbf{A} \cdot \mathbf{B} = \mathbf{B} \cdot \mathbf{A}$ only under special conditions).
- Properties of Multiplication:

$$\begin{aligned} \mathbf{AI} &= \mathbf{A} \\ (\mathbf{AB})' &= \mathbf{B}'\mathbf{A}' \\ (\mathbf{AB})\mathbf{C} &= \mathbf{A}(\mathbf{BC}) \\ \mathbf{A}(\mathbf{B}+\mathbf{C}) &= \mathbf{AB}+\mathbf{AC} \quad \text{or} \quad (\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC}+\mathbf{BC} \end{aligned}$$

2.8 Representing a regression model via matrix multiplication

- We have a model: $y = \beta_0 + \beta_1 \mathbf{X} + \boldsymbol{\varepsilon}$, implying:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 X_1 + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 X_2 + \varepsilon_2 \end{aligned}$$

- This can be represented in matrix form by:

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

where $\mathbf{y}$ is $n \times 1$, $\mathbf{X}$ is $n \times 2$, $\boldsymbol{\beta}$ is 2×1, and $\boldsymbol{\varepsilon}$ is $n \times 1$.

2.9 Using matrix multiplication to compute useful quantities

- To get the average of four observations of a random variable: 30, 32, 31, 33.

  $\begin{bmatrix}1/4 & 1/4 & 1/4 & 1/4\end{bmatrix} \cdot \begin{bmatrix}30\\32\\31\\33\end{bmatrix} = \bar{x}$ and this can also be written as $1/n\mathbf{i}'\mathbf{x} = \bar{x}$, where $\mathbf{i}$ is a column vector of ones of dimensions $n \times 1$ and $\mathbf{x}$ is the data vector.

- Suppose we want to get a vector containing the deviation of each observation of $\mathbf{x}$ from its mean (why would that be useful?):

$$\begin{bmatrix}x_1 - \bar{x}\\x_2 - \bar{x}\\\vdots\\x_n - \bar{x}\end{bmatrix} = [\mathbf{x} - \mathbf{i}\bar{x}] = \left[\mathbf{x} - \frac{1}{n}\mathbf{i}\mathbf{i}'\mathbf{x}\right]$$

- Since $\mathbf{x} = \mathbf{I}\mathbf{x}$,

$$\left[\mathbf{x} - \frac{1}{n}\mathbf{i}\mathbf{i}'\mathbf{x}\right] = \left[\mathbf{I}\mathbf{x} - \frac{1}{n}\mathbf{i}\mathbf{i}'\mathbf{x}\right] = \left[\mathbf{I} - \frac{1}{n}\mathbf{i}\mathbf{i}'\right]\mathbf{x} = \mathbf{M}^0\mathbf{x}$$

- $\mathbf{M}^0$ has $(1 - 1/n)$ for its diagonal elements and $-1/n$ for all its off-diagonal elements ($\therefore$ symmetric). Also, $\mathbf{M}^0$ is equal to its square, $\mathbf{M}^0\mathbf{M}^0 = \mathbf{M}^0$ so it is idempotent.

2.10 Representing a system of linear equations via matrix multiplication

- Say that we have a fairly elementary set of linear equations:

$$\begin{aligned}2x + 3y &= 5\\3x - 6y &= -3\end{aligned}$$

- Matrix representation:
$$\begin{bmatrix} 2 & 3 \\ 3 & -6 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 5 \\ -3 \end{bmatrix}$$
- In general form, this would be expressed as: $\mathbf{Ax} = \mathbf{b}$
- Solve by using the **inverse**: $\mathbf{A}^{-1}$ where $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$.
- Pre-multiply both sides of the equation by $\mathbf{A}^{-1}$:
$$\begin{aligned} \mathbf{Ix} &= \mathbf{A}^{-1}\mathbf{b} \\ \mathbf{x} &= \mathbf{A}^{-1}\mathbf{b} \end{aligned}$$
- Only square matrices have inverses, but not all square matrices possess an inverse.
- To have an inverse (i.e., be non-singular), matrix must be "full rank" $\Rightarrow$ its determinant (e.g., $|\mathbf{A}|$) is not zero.
- For a 2×2 matrix, $\mathbf{A} = \begin{bmatrix} a & c \\ b & d \end{bmatrix}$, $|\mathbf{A}| = (ad - bc)$ and $\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

2.11 What is rank and why would a matrix not be of full rank?

- Let us consider a set of two linear equations with two unknowns:

$$3x + y = 7$$
$$6x + 2y = 14$$

$$\begin{bmatrix} 3 & 1 \\ 6 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 7 \\ 14 \end{bmatrix}$$

- Can't solve this: the second equation is not truly a different equation; it is a "linear combination" of the first $\Rightarrow$ an infinity of solutions.
- The matrix $\begin{bmatrix} 3 & 1 \\ 6 & 2 \end{bmatrix}$ does not have full "row rank."
- One of its rows can be expressed as a linear combination of the others ($\Rightarrow$ its determinant = 0).
- Does not have full column rank either; if a square matrix does not have full row rank it will not have full column rank.
- Rows and columns must be "linearly independent" for matrix to have full rank.

2.12 The Rank of a non-square matrix

- For non-square matrices, rank($\mathbf{A}$) = rank ($\mathbf{A}'$) $\leq$ min(number of rows, number of columns).
- For example, for the matrix $\begin{bmatrix} 5 & 1 \\ 2 & -3 \\ 7 & 4 \end{bmatrix}$, the rank of the matrix is at most two (the number of columns).
- Why is this? Let us look at the three rows separately and think of each of them as a vector in space. If we have two rows with two elements, we can produce the third row via a linear combination of the first two and we can do this for any third row that we could imagine. In technical language, the matrix only spans a vector space of two.

2.13 Application of the inverse to regression analysis

- Can't use this to directly solve

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

- We have the error vector and $\mathbf{X}$ is not square.
- But we will use this to get $\boldsymbol{\beta}$ by seeking a solution that minimized the sum of squared residuals; will get us a system of K equations from which we may derive K unknown $\boldsymbol{\beta}$ coefficients.

2.14 Partial differentiation

- To find the partial derivative of any vector, $\mathbf{x}$, or any function of that vector, f($\mathbf{x}$) $=\frac{\partial f(x)}{\partial(x_i)}$:

  ➢ take the derivative of each element of the matrix with respect to x_n.

- $\frac{\partial(\mathbf{a}'\mathbf{x})}{\partial\mathbf{x}} = \mathbf{a}$
- $\frac{\partial(\mathbf{x}'\mathbf{A}\mathbf{x})}{\partial\mathbf{x}} = 2\mathbf{A}\mathbf{x}$.

2.15 Multivariate distributions

- **Multivariate distribution** describes the joint distribution of any group of random variables X_1 to X_n.
- This set can be collected as a vector or matrix and we can write their moments (i.e., mean and variance) in matrix notation.
- E.g., if X_1 to X_n are written as $\mathbf{x}$, then $E(\mathbf{x}) = [E(X_1), E(X_2), \ldots, E(X_n)]$.
- For any $n \times 1$ random vector, $\mathbf{x}$, its variance-covariance matrix, denoted var($\mathbf{x}$), is defined as:

$$\text{var}(\mathbf{x}) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \ldots & \sigma_{2n} \\ \vdots & & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_n^2 \end{bmatrix} = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'] = \boldsymbol{\Sigma} \qquad (2.1)$$

- This matrix is symmetric. Why? In addition, if the individual elements of $\mathbf{x}$ are independent, then the off-diagonal elements are equal to zero. Why?
- $\mathbf{x} \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- The properties of the multivariate normal $\Rightarrow$ each element of this vector is normally distributed and that any two elements of $\mathbf{x}$ are independent iff they are uncorrelated.

Section 3

The Classical Regression Model

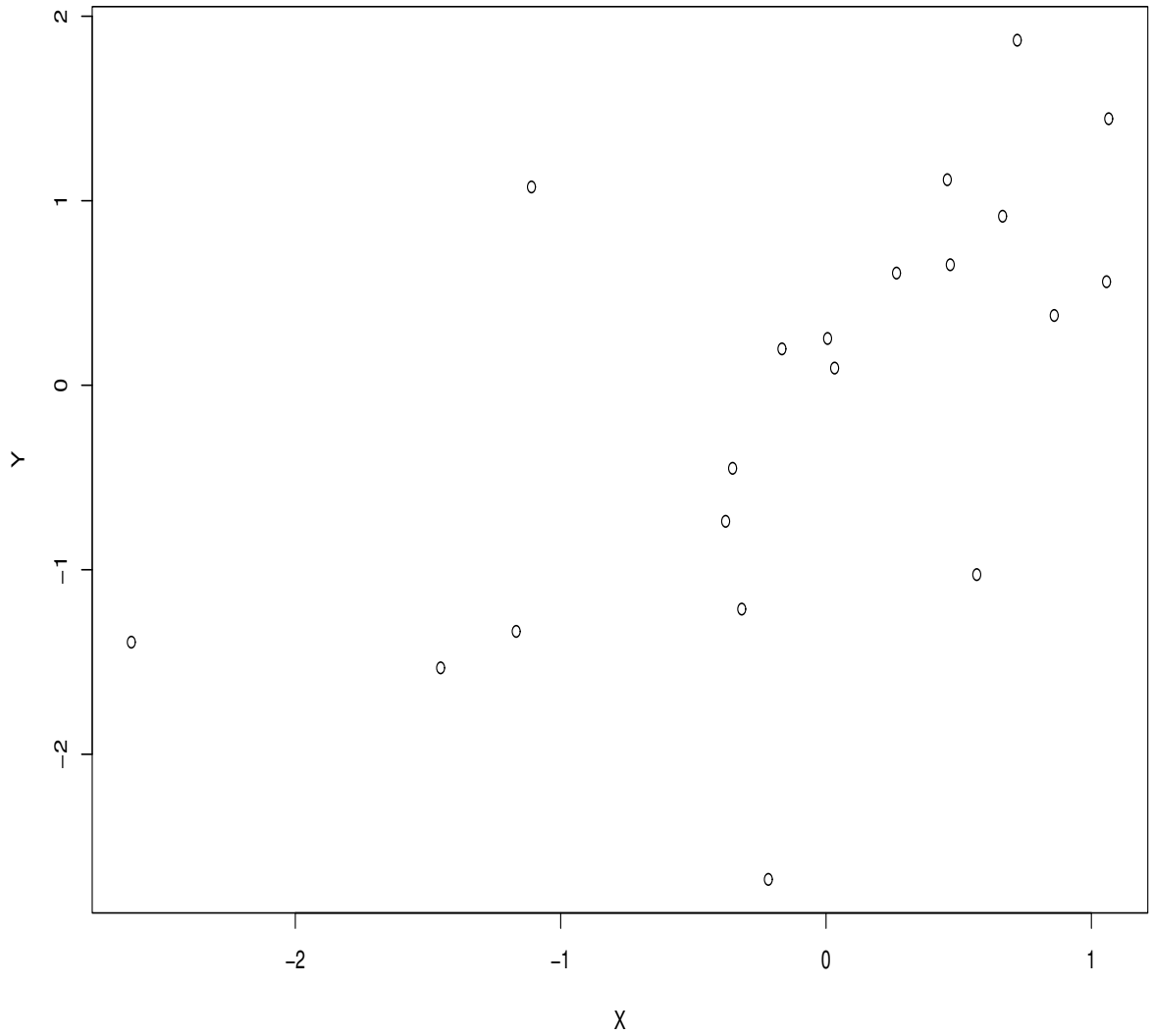
3.1 Overview of ordinary least squares

- OLS is appropriate whenever the Gauss-Markov assumptions are satisfied.
- Technically simple and tractable.
- Allows us to develop the theory of hypothesis testing.
- Regression analysis tells us what is the relationship between a dependent variable, y, an independent variable, x.

3.2 Optimal prediction and estimation

- Suppose that we observe some data (x_i, y_i) for $i = 1, \ldots, N$ observational units of individuals.
- Our theory tells us that x_i and y_i are related in some systematic fashion for all individuals.
- Use the tools of statistics to describe this relationship more precisely and to test its empirical validity.
- For now, no assumptions concerning the functional form linking the variables in the data set.
- First problem: how to predict y_i optimally given the values of x_i?
- x_i is known as an independent variable and y_i is the dependent or endogenous variable.
- The function that delivers the best prediction of y_i for given values of x_i is known as an optimal predictor.
  - ➢ The "best predictor" is a real-valued function $g(\cdot)$, not necessarily linear.

Figure 3.1: Data Plot



- How to define "best predictor"? One criterion we could apply is to minimize the mean-square error, i.e. to solve:

$$\min_{g(.)} E\left[(y_i - g(x_i, \boldsymbol{\beta}))^2\right]$$

- It turns out that the optimal predictor in this sense is the conditional expectation of y given x, also known as the regression function, $E(y_i|x_i) = g(x_i, \boldsymbol{\beta})$.
- $g(.)$ presumably involves a number of fixed parameters ($\boldsymbol{\beta}$) needed to correctly specify the theoretical relationship between y_i and x_i.
- Whatever part of the actually observed value of y_i is not captured in $g(x_i, \boldsymbol{\beta})$ in the notation of this equation must be a random component which has a mean of zero conditional on x_i.
- $\therefore$ another way to write the equation above in a way that explicitly captures the randomness inherent in any political process is as follows:

$$y_i = g(x_i,\beta) + \varepsilon_i$$

where ε_i is an error term or disturbance, with a mean of zero conditional on x_i. For the moment, we shall not say anything else about the distribution of ε_i.

3.3 Criteria for optimality for estimators

- Need to define what we mean by "good" or optimal indicators.
- Normally, there are two dimensions along which we compare estimators.

1. Unbiasedness, i.e.: $E(\hat{\beta}) = \beta$
   - Large sample analog of unbiasedness is "consistency".
2. Efficiency: $\text{var}(\hat{\boldsymbol{\beta}}) \leq \text{var}(\hat{\boldsymbol{\beta}}^*)$, where $\hat{\boldsymbol{\beta}}^*$ is any unbiased estimator for $\boldsymbol{\beta}$.
   - Large sample analog is asymptotic efficiency.

3.4 Linear predictors

- We often restrict $g(.)$ to being a linear function of x_i, or a linear function of non-linear transformations of x_i.
- More specifically, the best linear predictor (BLP) of y_i given values of x_i is denoted by:
$$E^*(y_i|x_i) = \beta_0 + \beta_1 x_i.$$
- Note that the BLP is not the same as the conditional expectation, unless the conditional expectation happens to be linear. As before, we can write the equation in an alternative but equivalent fashion:
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
- The assumption of linearity corresponds to a restriction that the true relationship between y_i and x_i be a line in (y_i, x_i) space. Such a line is known as a regression line.
- Moreover, the fact that we are trying to predict y_i given x_i requires that the error term be equal to the vertical distance between the true regression line and the data points.
- Thus, in order to estimate the true regression line—that is, to approximate it using the information contained in our finite sample of observations—one possible criterion will consist of minimizing some function of the vertical distance between the estimated line and the scattered points.
  - ➢ In doing so, we hope to exhaust all the information in x_i that is useful in order to predict y_i linearly.
- The problem that we are now confronted with consists of estimating the true parameters β_0 and β_1 and drawing statistical inferences from these estimates.
- One way: ordinary least squares estimation (OLS).
- Under special assumptions (the **Gauss-Markov assumptions** or assumptions of the classic linear regression model), OLS estimates are "optimal" in terms of unbiasedness and efficiency.

3.5 Ordinary least squares

- The estimates of β_0 and β_1 resulting from the minimization of the sum of squared vertical deviations are known as the OLS estimates.

3.5.1 Bivariate example

- The vertical deviations from the estimated line, or residuals e_i are as follows:

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i,$$

  where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates of the parameters that fully describe the regression line.

- Thus, the objective is to solve the following problem:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- The first order conditions for a minimum (applying the chain rule) are:

$$\frac{\partial \left(\sum_{i=1}^{N} e_i^2 \right)}{\partial \hat{\beta}_0} = \sum_{i=1}^{N} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1) = 0 \Rightarrow \sum_{i=1}^{N} e_i = 0$$

$$\frac{\partial \left(\sum_{i=1}^{N} e_i^2 \right)}{\partial \hat{\beta}_1} = \sum_{i=1}^{N} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i) = 0 \Rightarrow \sum_{i=1}^{N} x_i e_i = 0$$

- These are the "normal equations". They can be re-written by substituting for e_i and collecting terms. This yields the following two expressions.

$$\sum_{i=1}^{N} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{N} x_i$$

$$\sum_{i=1}^{N} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{N} x_i + \hat{\beta}_1 \sum_{i=1}^{N} x_i^2$$

- To obtain a solution, divide the first equation by n:

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

$\Rightarrow$ OLS regression line passes through the mean of the data (but may not be true if no constant in the model) and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Use the last expression in place of $\hat{\beta}_0$ in the solution for $\hat{\beta}_1$ and use $\sum_{i=1}^{N} x_i = N\bar{x}$:

$$\sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y} = \hat{\beta}_1 \left(\sum_{i=1}^{N} x_i^2 - N\bar{x}^2 \right)$$

or

$$\hat{\beta}_1 = \frac{\left(\sum_{i=1}^{N} x_1 y_1 \right) - N\bar{x}\bar{y}}{\left(\sum_{i=1}^{N} x_i^2 \right) - N\bar{x}^2} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

- Check to see if this is a minimum:

$$\begin{bmatrix} \frac{\partial^2 \left(\sum_{i=1}^{N} e_i^2 \right)}{\partial \hat{\beta}_0^2} & \frac{\partial^2 \left(\sum_{i=1}^{N} e_i^2 \right)}{\partial \hat{\beta}_0 \partial \hat{\beta}_1} \\ \frac{\partial^2 \left(\sum_{i=1}^{N} e_i^2 \right)}{\partial \hat{\beta}_0 \partial \hat{\beta}_1} & \frac{\partial^2 \left(\sum_{i=1}^{N} e_i^2 \right)}{\partial \hat{\beta}_1^2} \end{bmatrix} = \begin{bmatrix} 2N & 2N\bar{x} \\ 2N\bar{x} & 2\sum_{i=1}^{N} x_i^2 \end{bmatrix}$$

- Sufficient condition for a minimum: matrix must be positive definite.

- The two diagonal elements are positive, so we only need to verify that the determinant is positive:

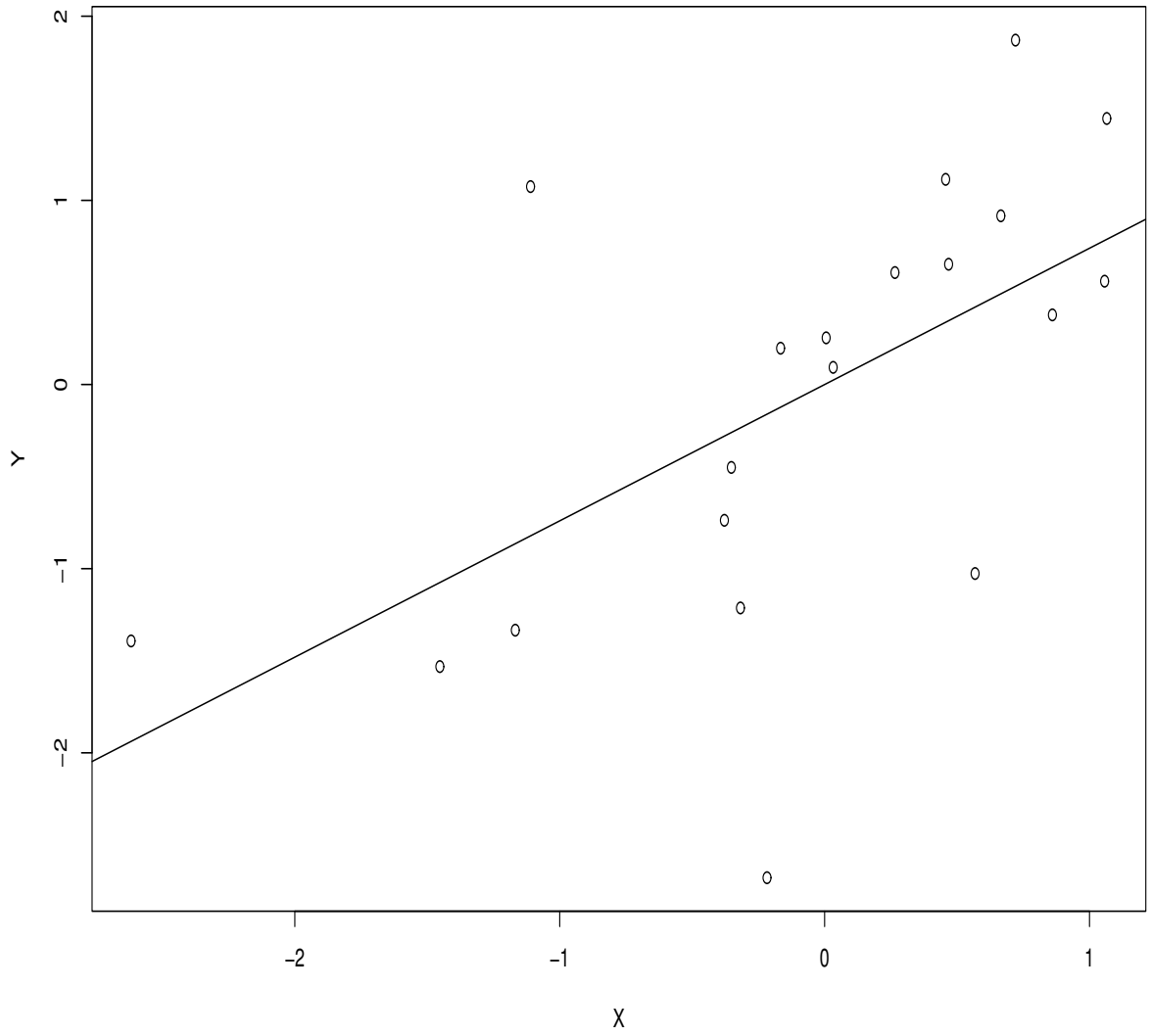
$$
\begin{aligned}
|D| &= 4N \sum_i x_i^2 - 4 \left(\sum_i x_i \right)^2 \\
&= 4N \sum_i x_i^2 - 4N^2 \bar{x}^2 \\
&= 4N \left(\sum_i x_i^2 - N \bar{x}^2 \right) \\
&= 4N \sum_i (x_i - \bar{x})^2 > 0 \quad \text{if variation in } x
\end{aligned}
$$

- If there are more regressors, one can add normal equations of the form

$$\frac{\partial \sum e_i^2}{\partial \hat{\beta}_k} = 0.$$

- Math is tedious; more convenient to use matrix notation.

Figure 3.2: Data Plot w/ OLS Regression Line



3.5.2 Multiple regression

- When we want to know the direct effects of a number of independent variables on a dependent variable, we use multiple regressions by assuming that

$$E(y_i|x_{1i},\ldots,x_{ki}) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}$$

(Notice that there is no x_{0i}. Actually, we assume that $x_{0i} = 1$ for all i.)

- May also be written as

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

- The sample regression function can be estimated by Least Squares as well and can be written as:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_k x_{ki}$$

or

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_k x_{ki} + e_i$$

- **Interpretation of Coefficients** β_k is the effect on y of a one unit increase of x_k holding all of the others xs constant.

  ➢ Example: Infant deaths per 1000 live births

  Infant mortality = 125−.972×(% Females who Read)−.002×(Per capita GNP)

  ➢ Holding per capita GDP constant, a one percentage point increase in female literacy reduces the expected infant mortality rate by .972 deaths per 1000 live births.

  ➢ Holding female literacy rates constant, a one thousand dollar increase in the per capita gross domestic product reduces the expected infant mortality rate by 2.234 deaths per 1000 live births.

3.5.3 The multiple regression model in matrix form

- To expedite presentation and for computational reasons, it is important to be able to express the linear regression model is matrix form. To this end, note that the observations of a multiple regression can be written as:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \ldots + \beta_k x_{k1} + \ldots + \beta_K x_{K1} + \varepsilon_1 \\ &\vdots \\ y_i &= \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \ldots + \beta_K x_{Ki} + \varepsilon_i \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \ldots + \beta_k x_{nk} + \ldots + \beta_K x_{nK} + \varepsilon_n \end{aligned}$$

- Now let $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$, $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_K \end{bmatrix}$, $\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$, and $\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1K} \\ \vdots & \ldots & \ldots & \vdots \\ 1 & x_{n1} & \ldots & x_{nK} \end{bmatrix}$.

- Now the entire model can be written as

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

- We will often want to refer to specific vectors within $\mathbf{X}$. Let $\mathbf{x}_k$ refer to a column in $\mathbf{X}$ (the kth independent variable) and $\mathbf{x}_i$ be a row in $\mathbf{X}$ (the independent variables for the ith observation).

- Thus, we can write the model for a specific observation as

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

3.5.4 OLS using matrix notation

- The sum of squared residuals is

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})^2$$

where $\hat{\boldsymbol{\beta}}$ is now a vector of unknown coefficients.

- The minimization problem can now be expressed as

$$\min_{\hat{\boldsymbol{\beta}}} \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

- Expanding this gives:

$$\mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

  or

$$\mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

- The first order conditions now become:

$$\frac{\partial \mathbf{e}'\mathbf{e}}{\partial \hat{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}$$

- The solution then satisfies the least squares normal equations:

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

- If the inverse of $\mathbf{X}'\mathbf{X}$ exists (i.e., assuming full rank) then the solution is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Section 4

Properties of OLS in finite samples

4.1 Gauss-Markov assumptions

- OLS estimation relies on five basic assumptions about the way in which the data are generated.
- If these assumptions hold, then OLS is BLUE (Best Linear Unbiased Estimator).

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. $\boldsymbol{\varepsilon}|\mathbf{X} \sim N[\mathbf{0}, \sigma^2\mathbf{I}]$

In English:

- Assumptions 1 and 2 $\Rightarrow E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}$. (Why?)
- Assumption 2: "strict exogeneity" assumption. $\Rightarrow E[\varepsilon_i] = 0$, since $E[\varepsilon_i] = E_X[E[\varepsilon_i|\mathbf{X}_i]] = 0$ (via the Law of Iterated Expectations).
- Assumption 3: errors are spherical/iid.
- Assumption 4: data matrix has full rank; is invertible (or "non-singular"); *not* characterized by perfect multicollinearity.
- Sometimes add to Assumption 4 "$\mathbf{X}$ is a non-stochastic matrix" or "$\mathbf{X}$ is fixed in repeated samples." This amounts to saying that $\mathbf{X}$ is something that we fix in an experiment. Generally not true for political science.

➢ We will assume $\mathbf{X}$ can be fixed or random, but it is generated by a mechanism unrelated to $\boldsymbol{\varepsilon}$.

- Note that

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

$$\text{var}(\varepsilon_i) = E[(\varepsilon_i - E(\varepsilon_i))^2] = E[\varepsilon_i^2]$$

and

$$\text{cov}(\varepsilon_i, \varepsilon_j) = E[(\varepsilon_i - E(\varepsilon_i))(\varepsilon_j - E(\varepsilon_j))] = E(\varepsilon_i\varepsilon_j)$$

- Thus, we can understand the assumption that $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\mathbf{I}_n$, as a statement about the variance-covariance matrix of the error terms, which is equal to:

$$\begin{bmatrix} \sigma^2 & 0 & ... & 0 \\ 0 & \sigma^2 & ... & 0 \\ \vdots & & \ddots & \\ 0 & 0 & ... & \sigma^2 \end{bmatrix}$$

- Diagonal elements ⇒ homoskedasticity; Off-diagonal elements ⇒ no auto-correlation.

4.2 Using the assumptions to show that $\hat{\boldsymbol{\beta}}$ is unbiased

- $\widehat{\boldsymbol{\beta}}$ is unbiased if $E[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \end{aligned}$$

- Taking expectations:

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

- So the coefficients are unbiased "conditioned" on the data set at hand. Using the law of iterated expectations, however, we can also say something about the unconditional expectation of the coefficients.
- By the LIE, the unconditional expectation of $\widehat{\boldsymbol{\beta}}$ can be derived by "averaging" the conditional expectation over all the samples of $\mathbf{X}$ that we could observe.

$$\begin{aligned} E[\widehat{\boldsymbol{\beta}}] &= E_X[E[\widehat{\boldsymbol{\beta}}|X]] \\ &= \boldsymbol{\beta} + E_X[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\boldsymbol{\varepsilon}|\mathbf{X}]] = \boldsymbol{\beta} \end{aligned}$$

The last step holds because the expectation over $\mathbf{X}$ of something that is always equal to zero is still zero.

4.3 Using the assumptions to find the variance of $\hat{\boldsymbol{\beta}}$

- Now that we have the expected value of $\widehat{\boldsymbol{\beta}}$, we should be able to find its variance using the equation for the variance-covariance matrix of a vector of random variables (see Eq. 2.1).

$$\text{var}(\widehat{\boldsymbol{\beta}}) = E[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})']$$

$$\begin{aligned} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta} \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} - \boldsymbol{\beta} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \end{aligned}$$

- Thus:

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

- Use the transpose rule $(\mathbf{AB})' = (\mathbf{B}'\mathbf{A}')$:

$$((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon})' = (\boldsymbol{\varepsilon}'\mathbf{X}[(\mathbf{X}'\mathbf{X})^{-1}]')$$

- The transpose of the inverse is equal to the inverse of the transpose:

$$[(\mathbf{X}'\mathbf{X})^{-1}]' = [(\mathbf{X}'\mathbf{X})']^{-1}$$

- Since $(\mathbf{X}'\mathbf{X})$ is a symmetric matrix:

$$[(\mathbf{X}'\mathbf{X})^{-1}]' = [(\mathbf{X}'\mathbf{X})']^{-1} = (\mathbf{X}'\mathbf{X})^{-1}$$

  which gives:

$$((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon})' = (\boldsymbol{\varepsilon}'\mathbf{X}[(\mathbf{X}'\mathbf{X})^{-1}]') = (\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1})$$

- Then,

$$\begin{aligned}\text{var}(\widehat{\boldsymbol{\beta}}) &= E[((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon})(\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1})] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]\end{aligned}$$

- Next, let's pass through the conditional expectations operator rather than the unconditional to obtain the conditional variance.

$$\begin{aligned}E[[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]|\mathbf{X}] &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\ &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\ &= \sigma^2[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

- But the variance of $\widehat{\boldsymbol{\beta}}$ depends on the unconditional expectation.
- Use a theorem on the decomposition of variance from Greene, 6th ed., p. 1008:

$$\text{var}(\widehat{\boldsymbol{\beta}}) = E_{\mathbf{X}}[\text{var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] + \text{var}_{\mathbf{X}}[E[\widehat{\boldsymbol{\beta}}|\mathbf{X}]]$$

- The second term in this decomposition is zero since $E[\widehat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta}$ for all $\mathbf{X}$, thus:

$$= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

4.4 Finding an estimate for σ^2

- Since $\sigma^2 = E[(\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon}))^2] = E[\boldsymbol{\varepsilon}^2]$

  and $\mathbf{e}$ is an estimate of $\boldsymbol{\varepsilon}$, $\frac{\sum e_i^2}{N}$ would seem a natural choice.

- But $e_i = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}} = \varepsilon_i - \mathbf{x}_i'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.

- Note that:

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}$$

- Let

$$\mathbf{M} = (\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$$

  $\mathbf{M}$ is a symmetric, idempotent matrix, like $\mathbf{M}^0$ ($\Rightarrow \mathbf{M} = \mathbf{M}'$ and $\mathbf{MM} = \mathbf{M}$).

- Note $\mathbf{MX} = \mathbf{0}$.

- We can also show (by substituting for $\mathbf{y}$) that:

$$\mathbf{e} = \mathbf{My} \Rightarrow \mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{M}'\mathbf{My} = \mathbf{y}'\mathbf{My} = \mathbf{y}'\mathbf{e} = \mathbf{e}'\mathbf{y}$$

  and (even more useful)

$$\mathbf{e} = \mathbf{My} = \mathbf{M}[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}] = \mathbf{MX}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\varepsilon} = \mathbf{M}\boldsymbol{\varepsilon}$$

- So $E(\mathbf{e}'\mathbf{e}) = E(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon})$.

- Now we need some matrix algebra: the trace operator gives the sum of the diagonal elements of a matrix. So, for example:

$$tr\begin{bmatrix}1 & 7\\ 2 & 3\end{bmatrix} = 4$$

- Three key results for trace operations:

  1. $tr(\mathbf{ABC}) = tr(\mathbf{BCA}) = tr(\mathbf{CBA})$
  2. $tr(c\mathbf{A}) = c \cdot tr(\mathbf{A})$
  3. $tr(\mathbf{A} - \mathbf{B}) = tr(\mathbf{A}) - tr(\mathbf{B})$

- For a 1×1 matrix, the trace is equal to the matrix (Why?). The matrix $\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}$ is a 1×1 matrix. As a result:

$$E(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}) = E[tr(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon})] = E[tr(\mathbf{M}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon})]$$

  using the trace operations above, and

$$E[tr(\mathbf{M}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon})] = tr(\mathbf{M}E[\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}]) = tr(\mathbf{M}\sigma^2\mathbf{I}) = \sigma^2 tr(\mathbf{M})$$

- Thus, $E(\mathbf{e}'\mathbf{e}) = \sigma^2 tr(\mathbf{M})$

$$\begin{aligned} tr(\mathbf{M}) &= tr(\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\ &= tr(\mathbf{I}_N) - tr(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\ &= N - tr(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\ &= N - tr(\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}) \\ &= N - tr(\mathbf{I}_K) = N - k \end{aligned}$$

- Putting this altogether: $E(\mathbf{e}'\mathbf{e}) = \sigma^2(N-k) \Rightarrow s^2 = \frac{\mathbf{e}'\mathbf{e}}{N-k}$ is an unbiased estimator for σ^2.
- Thus, $\widehat{\text{var}}[\widehat{\beta}] = s^2(\mathbf{X}'\mathbf{X})^{-1}$.

4.5 Distribution of the OLS coefficients

- Use Assumption Five to figure this out.
- In solving for $\widehat{\boldsymbol{\beta}}$, we arrived at an expression that was a linear function of the error terms. Earlier, we stated that:

  $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$, this is a linear function of $\boldsymbol{\varepsilon}$, analogous to $\mathbf{Az} + \mathbf{b}$, where $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{b} = \boldsymbol{\beta}$ and $\mathbf{z} = \boldsymbol{\varepsilon}$, a random vector.

- Next, we can use a property of the multi-variate normal distribution:

  > If $\mathbf{z}$ is a multivariate normal vector (i.e., $\mathbf{z} \sim N[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$), then
  > $\mathbf{Az} + \mathbf{b} \sim N[\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}']$

  This implies that $\widehat{\boldsymbol{\beta}}|\mathbf{X} \sim N[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$

- If we could realistically treat $\mathbf{X}$ as non-stochastic, then we could just say that $\widehat{\boldsymbol{\beta}}$ was distributed multivariate normal, without any conditioning on $\mathbf{X}$.
- Later, in the section on asymptotic results, we will be able to show that the distribution of $\widehat{\boldsymbol{\beta}}$ is approximately normal as the sample size gets large, without any assumptions on the distribution of the true error terms and without having to condition on $\mathbf{X}$.
- In the meantime, it is also useful to remember that, for a multivariate normal distribution, each element of the vector $\widehat{\boldsymbol{\beta}}$ is also distributed normal, so that we can say that:

$$\widehat{\beta}_k|\mathbf{X} \sim N[\beta_k, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}]$$

- Will use this result for hypothesis testing.

4.6 Efficiency of OLS regression coefficients

- To show this, we must prove that there is no other unbiased estimator of the true parameters, $\widetilde{\boldsymbol{\beta}}$, such that $\text{var}(\widetilde{\boldsymbol{\beta}}) < \text{var}(\widehat{\boldsymbol{\beta}})$.
- Let $\widetilde{\boldsymbol{\beta}}$ be a linear estimator of $\boldsymbol{\beta}$ calculated as $\mathbf{Cy}$, where $\mathbf{C}$ is a $K \times n$ matrix.
- If $\widetilde{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$, then:

$$E[\widetilde{\boldsymbol{\beta}}] = \boldsymbol{\beta} = E[\mathbf{Cy}] = \boldsymbol{\beta}$$

which implies that

$$E[\mathbf{C}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})] = \boldsymbol{\beta}$$

$$\begin{aligned}
&\Rightarrow \mathbf{CX}\boldsymbol{\beta} + \mathbf{C}E[\boldsymbol{\varepsilon}] = \boldsymbol{\beta} \\
&\Rightarrow \mathbf{CX}\boldsymbol{\beta} = \boldsymbol{\beta} \quad \text{and} \quad \mathbf{CX} = \mathbf{I}
\end{aligned}$$

We will use the last result when we calculate the variance of $\widetilde{\boldsymbol{\beta}}$.

$$\begin{aligned}
\text{var}[\widetilde{\boldsymbol{\beta}}] &= E[(\widetilde{\boldsymbol{\beta}} - E[\widetilde{\boldsymbol{\beta}}])(\widetilde{\boldsymbol{\beta}} - E[\widetilde{\boldsymbol{\beta}}])'] \\
&= E[(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})'] \\
&= E[(\mathbf{Cy} - \boldsymbol{\beta})(\mathbf{Cy} - \boldsymbol{\beta})'] \\
&= E[(\mathbf{CX}\boldsymbol{\beta} + \mathbf{C}\boldsymbol{\varepsilon}) - \boldsymbol{\beta})((\mathbf{CX}\boldsymbol{\beta} + \mathbf{C}\boldsymbol{\varepsilon}) - \boldsymbol{\beta})'] \\
&= E[(\boldsymbol{\beta} + \mathbf{C}\boldsymbol{\varepsilon} - \boldsymbol{\beta})(\boldsymbol{\beta} + \mathbf{C}\boldsymbol{\varepsilon} - \boldsymbol{\beta})'] \\
&= E[(\mathbf{C}\boldsymbol{\varepsilon})(\mathbf{C}\boldsymbol{\varepsilon})'] \\
&= \mathbf{C}E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']\mathbf{C}' \\
&= \sigma^2\mathbf{C}\mathbf{C}'
\end{aligned}$$

- Now we want to say something about this variance compared to the variance of $\widehat{\boldsymbol{\beta}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

- Let $\mathbf{C} = \mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$
- $\mathbf{D}$ could be positive or negative. We are not saying anything yet about the "size" of this matrix versus $(\mathbf{X}'\mathbf{X})^{-1}$.
- Now, $\text{var}[\widetilde{\boldsymbol{\beta}}] = \sigma^2\mathbf{C}\mathbf{C}' = \sigma^2(\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')'$
- Recalling that $\mathbf{C}\mathbf{X} = \mathbf{I}$, we can now say that:

$$(\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{X} = \mathbf{I}$$

so that,

$$\mathbf{D}\mathbf{X} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}$$

and

$$\mathbf{DX}+\mathbf{I}=\mathbf{I}$$

and

$$\mathbf{DX} = \mathbf{0}$$

- Given this last step, we can finally re-express the variance of $\widetilde{\boldsymbol{\beta}}$ in terms of $\mathbf{D}$ and $\mathbf{X}$.

$$\begin{aligned}
\text{var}[\widetilde{\boldsymbol{\beta}}] &= \sigma^2(\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \\
&= \sigma^2(\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{D}'+[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']') \qquad \text{(using } (\mathbf{A}+\mathbf{B})' = \mathbf{A}'+\mathbf{B}') \\
&= \sigma^2(\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{D}'+\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}) \\
&= \sigma^2(\mathbf{D}\mathbf{D}' + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}') \\
&= \sigma^2(\mathbf{D}\mathbf{D}' + (\mathbf{X}'\mathbf{X})^{-1}+(\mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1})')
\end{aligned}$$

Therefore:

$$\text{var}[\widetilde{\boldsymbol{\beta}}] = \sigma^2(\mathbf{D}\mathbf{D}' + (\mathbf{X}'\mathbf{X})^{-1}) = \text{var}(\widehat{\boldsymbol{\beta}}) + \sigma^2\mathbf{D}\mathbf{D}'$$

- The diagonal elements of $\text{var}[\widetilde{\boldsymbol{\beta}}]$ and $\text{var}(\widehat{\boldsymbol{\beta}})$ are the variance of the estimates. So it is sufficient to prove efficiency to show that the diagonal elements of $\sigma^2\mathbf{D}\mathbf{D}'$ are all positive.

- Since each element of $\mathbf{DD}'$ is made up of a sum of squares (i.e., $\sum_j d_{ij}^2$), no diagonal element could be negative.
- They could all be zero, but if so then $\mathbf{D} = \mathbf{0}$ and $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, meaning that $\widetilde{\boldsymbol{\beta}} = \mathbf{Cy} = \widehat{\boldsymbol{\beta}}$ and thus completing the proof.
- In conclusion, any other unbiased linear estimator of $\boldsymbol{\beta}$ will have a sampling variance that is larger than the sampling variance of the OLS regression coefficients.
- OLS is "good" in the sense that it produces regression coefficients that are "BLUE", the *best, linear unbiased estimator* of the true, underlying population parameters.

Section 5

Inference using OLS regression coefficients

- While we have explored the properties of the OLS regression coefficients, we have not explained what they tell us about the true parameters, $\boldsymbol{\beta}$, in the population. This is the task of statistical inference.
- Inference in multivariate regression analysis is conducted via hypothesis testing. Two approaches:
  1. **Test of significance** approach
  2. **Confidence interval** approach
- Some key concepts: the *size* and the *power* of a test, *Type I* and *Type II Errors*, and the *null* hypothesis and the *alternative* hypothesis.

5.1 A univariate example of a hypothesis test

- In the univariate case, we were often concerned with finding the average of a random distribution. Via the Central Limit theorem, we could say that $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$. Thus, it followed that:

$$Z = \frac{(\overline{X} - \mu)}{\sigma_{\overline{X}}} = \frac{(\overline{X} - \mu)}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N(0, 1)$$

- We could then use the Z distribution to perform hypothesis tests.
- E.g.: $H_0 : \mu = 4$ and $H_1 : \mu \neq 4$.
- Next, let us decide the size of the test. Suppose, we want to have 95% "confidence" with regard to our inferences, implying a confidence coefficient of (1-α)=0.95, so α, which is the size of the test, is equal to 0.05. The size of the test gives us a critical value of $Z = \pm Z_{\alpha/2}$
- Then the test of significance approach is to calculate $\frac{(\overline{X}-\mu)}{\sigma/\sqrt{n}}$directly.

- If $\frac{(\overline{X}-\mu)}{\sigma/\sqrt{n}} > +Z_{\alpha/2}$ or $\frac{(\overline{X}-\mu)}{\sigma/\sqrt{n}} < -Z_{\alpha/2}$, we reject the null hypothesis. Intuitively, if $\overline{X}$ is very different from what we said it is under the null, we reject the null.

- The confidence interval approach is to construct the 95% confidence interval for μ.

$$\Pr\left[\overline{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 0.95$$

- If the hypothesized value of $\mu = 4$ lies in the confidence interval, then we accept the null. If the hypothesized value does not lie in the confidence interval, then we say reject H_0 with 95% confidence. Intuitively, we reject the null hypothesis if the hypothesized value is far from the likely range of values of μ suggested by our estimate.

5.2 Hypothesis testing of multivariate regression coefficients

- As demonstrated earlier, the regression coefficients are distributed multivariate normal conditional on $\mathbf{X}$:

$$\widehat{\boldsymbol{\beta}}|\mathbf{X} \sim N\left[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right]$$

  and each individual regression coefficient is distributed normal conditional on $\mathbf{X}$:

$$\widehat{\boldsymbol{\beta}}_k|\mathbf{X} \sim N\left[\beta_k, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}\right]$$

- Let us assume, at this point, that we can use the relationship between the marginal and conditional multivariate normal distributions to say simply that the OLS regression coefficients are normally distributed.

$$\widehat{\beta}_k \sim N\left[\beta_k, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}\right]$$

- Thus, we should be able to use

$$Z = \frac{\widehat{\beta}_k - \beta_k}{\sigma_{\widehat{\beta}_k}} = \frac{\widehat{\beta}_k - \beta_k}{\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} = \frac{\widehat{\beta}_k - \beta_k}{\sigma\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} \sim N(0,1)$$

  to conduct hypothesis tests, right?

- So, why are regression results always reported with the t-statistic? The problem is that we don't know σ^2 and have to estimate it using $s^2 = \frac{\mathbf{e}'\mathbf{e}}{N-k}$.

- Aside: as the sample size becomes large, we can ignore the sampling distribution in s^2 and proceed as though $s^2 = \sigma^2$. Given our assumptions that the error is normally distributed, however, we can use the t-distribution.

5.2.1 Proof that $\frac{\widehat{\beta}_k - \beta_k}{se(\widehat{\beta}_k)} = \frac{\widehat{\beta}_k - \beta_k}{\sqrt{s^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} \sim t_k$

- Recall: if $Z_1 \sim N(0,1)$ and another variable $Z_2 \sim \chi_k$ and is independent of Z_1, then the variable defined as:

$$t = \frac{Z_1}{\sqrt{(Z_2/k)}} = \frac{Z_1\sqrt{k}}{\sqrt{Z_2}}$$

  is said to follow the student's t distribution with k degrees of freedom.

- We have something that is distributed like $Z_1 = \frac{\widehat{\beta}_k - \beta_k}{\sigma\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{kk}}}$.

- What we need is to divide it by something that is chi-squared. Let us assume for now that $\frac{(N-k)s^2}{\sigma^2} \sim \chi^2_{N-k}$.

- If that were true, we could divide one by another to get:

$$t_k = \frac{\frac{\widehat{\beta}_k - \beta_k}{\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}}}{\sqrt{[(N-k)s^2/\sigma^2]/(N-k)}} = \frac{\widehat{\beta}_k - \beta_k}{\sigma\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} \cdot \frac{\sigma}{s}$$
$$= \frac{\widehat{\beta}_k - \beta_k}{s\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} = \frac{\widehat{\beta}_k - \beta_k}{se(\widehat{\beta}_k)}$$

5.2.2 Proof that $\frac{(N-k)s^2}{\sigma^2} \sim \chi^2_{N-k}$

- This proof will be done by figuring out whether the component parts of this fraction are summed squares of variables that are distributed standard normal. The proof will proceed using something we used earlier ($\mathbf{e} = \mathbf{M}\boldsymbol{\varepsilon}$).

$$\frac{(N-k)s^2}{\sigma^2} = \frac{\mathbf{e}'\mathbf{e}}{\sigma^2} = \frac{\boldsymbol{\varepsilon}'\mathbf{M}\mathbf{M}\boldsymbol{\varepsilon}}{\sigma^2} = \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)' \mathbf{M} \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)$$

- We know that $\left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right) \sim N(0,1)$

- Thus, if $\mathbf{M} = \mathbf{I}$ we have $\left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)' \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right) \sim \chi^2_N$. This holds because the matrix multiplication would give us N sums of squared standard normal random variables. However, it can also be shown that if $\mathbf{M}$ is any idempotent matrix, then:

$$\left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)' \mathbf{M} \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right) \sim \chi^2[tr(\mathbf{M})]$$

- We showed before that $tr(\mathbf{M}) = N - k$

$$\left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)' \mathbf{M} \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right) \sim \chi^2_{N-k}$$

- This completes the proof. In conclusion, the test statistic, $\frac{\widehat{\beta}_k - \beta_k}{se(\widehat{\beta}_k)} \sim t_{N-k}$

5.3 Testing the equality of two regression coefficients

- This section will introduce us to the notion of testing a "restriction" on the vector of OLS coefficients.
- Suppose we assume that the true model is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

- Suppose also that we want to test the hypothesis $H_0 : \beta_2 = \beta_3$.
- We can also express this hypothesis as: $H_0 : \beta_2 - \beta_3 = 0$.
- We test this hypothesis using the estimator $(\widehat{\beta}_2 - \widehat{\beta}_3)$. If the estimated value is very different from the hypothesized value of zero, then we will be able to reject the null. We can construct the t-statistic for the test using:

$$t = \frac{(\widehat{\beta}_2 - \widehat{\beta}_3) - (\beta_2 - \beta_3)}{se(\widehat{\beta}_2 - \widehat{\beta}_3)}$$

- What is the standard error? $\widehat{\beta}_2$ and $\widehat{\beta}_3$ are both random variables. From section 1.3.7, $\text{var}(X - Y) = \text{var}(X) + \text{var}(Y) - 2\text{cov}(XY)$. Thus:

$$se(\widehat{\beta}_2 - \widehat{\beta}_3) = \sqrt{\text{var}(\widehat{\beta}_2) + \text{var}(\widehat{\beta}_3) - 2\text{cov}(\widehat{\beta}_2, \widehat{\beta}_3)}$$

- How do we get those variances and covariances? They are all elements of $\text{var}(\widehat{\boldsymbol{\beta}})$, the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ (Which elements?).

5.4 Expressing the above as a "restriction" on the matrix of coefficients

- We could re-express the hypothesis as a linear restriction on the $\boldsymbol{\beta}$ matrix:

$$\begin{bmatrix} 0 & 0 & 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = 0 \quad \text{or } \mathbf{R}'\boldsymbol{\beta} = \mathbf{q} \text{ where } \mathbf{R} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \end{bmatrix} \text{ and } \mathbf{q} = 0$$

- The sample estimate of $\mathbf{R}'\boldsymbol{\beta}$ is $\mathbf{R}'\widehat{\boldsymbol{\beta}}$ and the sample estimate of $\mathbf{q} = \widehat{\mathbf{q}}$ and is equal to $\mathbf{R}'\widehat{\boldsymbol{\beta}}$.
- Consistent with the procedure of hypothesis tests, we could calculate $t = \dfrac{\widehat{\mathbf{q}} - \mathbf{q}}{se(\widehat{\mathbf{q}})}$.
- To test this hypothesis, we need $se(\widehat{\mathbf{q}})$. Since $\widehat{\mathbf{q}}$ is a linear function of $\widehat{\boldsymbol{\beta}}$, and since we have estimated the variance of $\widehat{\boldsymbol{\beta}} = s^2(\mathbf{X}'\mathbf{X})^{-1}$, we can estimate $\widehat{\mathbf{q}}$'s variance as $\text{var}(\widehat{\mathbf{q}}) = \mathbf{R}'[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}$.
- This is just the matrix version of the rule that $\text{var}(a\mathbf{X}) = a^2\text{var}(\mathbf{X})$ (see section 1.3.3).
- Thus, $se(\widehat{\mathbf{q}}) = \sqrt{\mathbf{R}'[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}}$
- Given all of this, the t-statistic for the test of the equality of β_2 and β_3 can be expressed as:

$$t = \frac{\widehat{\mathbf{q}} - \mathbf{q}}{se(\widehat{\mathbf{q}})} = \frac{\mathbf{R}'\widehat{\boldsymbol{\beta}} - \mathbf{q}}{\sqrt{\mathbf{R}'[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}}}$$

- How does that get us exactly what we had above?

  $\mathbf{R}'\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_2 - \widehat{\beta}_3)$ and $\mathbf{q} = 0$.

- More complicatedly:

$$\mathbf{R}'[s^2(\mathbf{X}'\mathbf{X})^{-1}] = \begin{bmatrix} 0 & 0 & 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} \sigma^2_{\widehat{\beta}_0} & \sigma_{\widehat{\beta}_0,\widehat{\beta}_1} & \sigma_{\widehat{\beta}_0,\widehat{\beta}_2} & \sigma_{\widehat{\beta}_0,\widehat{\beta}_3} \\ \sigma_{\widehat{\beta}_1,\widehat{\beta}_0} & \sigma^2_{\widehat{\beta}_1} & \sigma_{\widehat{\beta}_1,\widehat{\beta}_2} & \sigma_{\widehat{\beta}_1,\widehat{\beta}_3} \\ \sigma_{\widehat{\beta}_2,\widehat{\beta}_0} & \sigma_{\widehat{\beta}_2,\widehat{\beta}_1} & \sigma^2_{\widehat{\beta}_2} & \sigma_{\widehat{\beta}_2,\widehat{\beta}_3} \\ \sigma_{\widehat{\beta}_3,\widehat{\beta}_0} & \sigma_{\widehat{\beta}_3,\widehat{\beta}_1} & \sigma_{\widehat{\beta}_3,\widehat{\beta}_2} & \sigma^2_{\widehat{\beta}_3} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{\widehat{\beta}_2,\widehat{\beta}_0} - \sigma_{\widehat{\beta}_3,\widehat{\beta}_0} & \sigma_{\widehat{\beta}_2,\widehat{\beta}_1} - \sigma_{\widehat{\beta}_3,\widehat{\beta}_1} & \sigma^2_{\widehat{\beta}_2} - \sigma_{\widehat{\beta}_3,\widehat{\beta}_2} & \sigma_{\widehat{\beta}_2,\widehat{\beta}_3} - \sigma^2_{\widehat{\beta}_3} \end{bmatrix}$$

And

$$\begin{aligned} &\mathbf{R}'[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R} \\ &= \begin{bmatrix} \sigma_{\widehat{\beta}_2,\widehat{\beta}_0} - \sigma_{\widehat{\beta}_3,\widehat{\beta}_0} & \sigma_{\widehat{\beta}_2,\widehat{\beta}_1} - \sigma_{\widehat{\beta}_3,\widehat{\beta}_1} & \sigma^2_{\widehat{\beta}_2} - \sigma_{\widehat{\beta}_3,\widehat{\beta}_2} & \sigma_{\widehat{\beta}_2,\widehat{\beta}_3} - \sigma^2_{\widehat{\beta}_3} \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \end{bmatrix} \\ &= \sigma^2_{\widehat{\beta}_2} - \sigma_{\widehat{\beta}_3,\widehat{\beta}_2} - \sigma_{\widehat{\beta}_2,\widehat{\beta}_3} + \sigma^2_{\widehat{\beta}_3} \\ &= \text{var}(\widehat{\beta}_2) + \text{var}(\widehat{\beta}_3) - 2\text{cov}(\widehat{\beta}_2, \widehat{\beta}_3) \end{aligned}$$

Thus, $se(\widehat{q}) = \sqrt{\mathbf{R}'[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}} = \sqrt{\text{var}(\widehat{\beta}_2) + \text{var}(\widehat{\beta}_3) - 2\text{cov}(\widehat{\beta}_2, \widehat{\beta}_3)}$

- We will come back to this when we do multiple hypotheses via F-tests.

Section 6

Goodness of Fit

6.1 The R-squared measure of goodness of fit

- The original fitting criteria used to produce the OLS regression coefficients was to minimize the sum of squared errors. Thus, the sum of squared errors itself could serve as a measure of the *fit* of the model. In other words, how well does the model fit the data?
- Unfortunately, the sum of squared errors will always rise if we add another observation or if we multiply the values of $\mathbf{y}$ by a constant. Thus, if we want a measure of how well the model fits the data we might ask instead whether variation in $\mathbf{X}$ is a good predictor of variation in $\mathbf{y}$.
- Recall the mean deviations matrix from Section 2.9, which is used to subtract the column means from every column of a matrix. E.g.,

$$\mathbf{M}^0\mathbf{X} = \begin{bmatrix} x_{11} - \bar{x}_1 & \dots & x_{1K} - \bar{x}_K \\ \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & \dots & x_{nK} - \bar{x}_K \end{bmatrix}.$$

- Using $\mathbf{M}^0$, we can take the mean deviations of both sides of our sample regression equation:

$$\mathbf{M}^0\mathbf{y} = \mathbf{M}^0\mathbf{X}\widehat{\beta} + \mathbf{M}^0\mathbf{e}$$

- To square the sum of deviations, we just use

$$(\mathbf{M}^0\mathbf{y})'(\mathbf{M}^0\mathbf{y}) = \mathbf{y}'\mathbf{M}^0\mathbf{M}^0\mathbf{y} = \mathbf{y}'\mathbf{M}^0\mathbf{y}$$

and

$$\begin{aligned} \mathbf{y}'\mathbf{M}^0\mathbf{y} &= (\mathbf{M}^0\mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{M}^0\mathbf{e})'(\mathbf{M}^0\mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{M}^0\mathbf{e}) \\ &= ((\mathbf{M}^0\mathbf{X}\widehat{\boldsymbol{\beta}})' + (\mathbf{M}^0\mathbf{e})')(\mathbf{M}^0\mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{M}^0\mathbf{e}) \\ &= (\widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{M}^0 + \mathbf{e}'\mathbf{M}^0)(\mathbf{M}^0\mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{M}^0\mathbf{e}) \\ &= \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{M}^0\mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{e}'\mathbf{M}^0\mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{M}^0\mathbf{e} + \mathbf{e}'\mathbf{M}^0\mathbf{e} \end{aligned}$$

- Now consider, what is the term $\mathbf{M}^0\mathbf{e}$? The $\mathbf{M}^0$ matrix takes the deviation of a variable from its mean, but the mean of $\mathbf{e}$ is equal to zero, so $\mathbf{M}^0\mathbf{e} = \mathbf{e}$ and $\mathbf{X}'\mathbf{M}^0\mathbf{e} = \mathbf{X}'\mathbf{e} = \mathbf{0}$.
- Why does the last hold true? Intuitively, the way that we have minimized the OLS residuals is to set the estimated residuals orthogonal to the data matrix – there is no information in the data matrix that helps to predict the residuals.
- You can also prove that $\mathbf{X}'\mathbf{e} = \mathbf{0}$ using the fact that $\mathbf{e} = \mathbf{M}\boldsymbol{\varepsilon}$ and that $\mathbf{M} = (\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$:

$$\mathbf{X}'\mathbf{e} = \mathbf{X}'(\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{\varepsilon} = (\mathbf{X}' - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{\varepsilon} = \mathbf{X}'\boldsymbol{\varepsilon} - \mathbf{X}'\boldsymbol{\varepsilon} = 0.$$

- Given this last result, we can say that:

$$\mathbf{y}'\mathbf{M}^0\mathbf{y} = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{M}^0\mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{e}'\mathbf{M}^0\mathbf{e}$$

- The first term in this decomposition is the *regression sum of squares* (or the variation in $\mathbf{y}$ that is explained) and the second term is the *error sum of squares*. What we have shown is that the total variation in $\mathbf{y}$ can be fully decomposed into the explained and unexplained portion. That is,

$$SST = SSR + SSE$$

- Thus, a good measure of fit is the coefficient of determination or R^2:

$$R^2 = \frac{SSR}{SST}$$

- It follows that:

$$1 = \frac{\widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{M}^0\mathbf{X}\widehat{\boldsymbol{\beta}}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}} + \frac{\mathbf{e}'\mathbf{M}^0\mathbf{e}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}} = R^2 + \frac{\mathbf{e}'\mathbf{M}^0\mathbf{e}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}}$$

  and

$$R^2 = 1 - \frac{\mathbf{e}'\mathbf{M}^0\mathbf{e}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}} = 1 - \frac{SSE}{SST}$$

- *By construction*, R^2 can vary between 1 (if the model fully explains $\mathbf{y}$ and all points lie along the regression plane) and zero (if all the $\widehat{\boldsymbol{\beta}}$ are equal to zero).

6.1.1 The uses of R^2 and three cautions

- R^2 is frequently used to compare different models and decide between them. Other things equal, we would prefer a model that explains more to a model that explains less. There are three qualifications that need to be borne in mind, however, when using R^2, one of which motivates the use of the adjusted R^2.

1. R^2 (almost) always rises when you add another regressor
   - Consider a model $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$.
   - Add to this model an explanatory variable, z, which I know a priori to be completely independent of $\mathbf{y}$. Unless the estimated regression coefficient on z is exactly equal to zero, I will be adding some information to the model. Thus, I will automatically lower the estimated residuals and increase R^2, leading to a temptation to throw ever more variables into the model.
   - Alternative: Adjusted R^2:

$$\overline{R}^2 = 1 - \frac{\mathbf{e}'\mathbf{M}^0\mathbf{e}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}} \cdot \frac{n-1}{n-K}$$

   - As the number of regressors increases, the error term decreases, but the adjustment term increases. Thus, the adjusted R^2 can go up or down when you add a new regressor. The adjusted R^2 can also be written as:

$$\overline{R}^2 = 1 - \frac{n-1}{n-K}(1 - R^2).$$

2. R^2 is not comparable across different dependent variables

   - If you are comparing two models on the basis of R^2 or the adjusted R^2, the sample size, n, and the dependent variable must be the same.
   - Say you used $\ln \mathbf{y}$ as the dependent variable in your next model. R^2 measures the proportion of the variation in the dependent variable that is explained by the model, but in the first case it will be measuring the variation in $\mathbf{y}$ and in the second case it will be measuring the variation in $\ln \mathbf{y}$. The two are not expected to be the same and cannot be taken as the basis for model comparison. Also, clearly, as the number of observations changes, adjusted R^2 will also change.
   - The standard error of the regression is recommended instead of R^2 in these cases.

3. R^2 does not necessarily lie between zero and one if we do not include a constant in the model.

   - Inclusion of a constant gives the result in the normal equation that $\overline{y} = \overline{\mathbf{x}}'\widehat{\boldsymbol{\beta}}$ and this also implied that $\overline{\mathbf{e}} = 0$. If you do not include a constant term in the regression model, this does not necessarily hold, which means that we cannot simplify the variation in $\mathbf{y}$ as we did above.
   - As a result, if you do not include a constant, it is possible to get an R^2 of less than zero. Generally, you should exclude the constant term only if you have strong theoretical reasons to believe the dependent variable is zero when all of the explanatory variables are zero.

- R^2 is often abused. We will cover alternative measures that can be used to judge the adequacy of a regression model. These include Akaike's Information Criterion, the Bayes Information Criterion and Amemiya's Prediction Criterion.

6.2 Testing Multiple Hypotheses: the F-statistic and R^2

- Can use the restriction approach to test multiple hypotheses. E.g., $H_0 : \beta_2 = 0$ and $\beta_3 = 0$.
- The method we will use can accommodate any case in which we have $J < K$ hypotheses to test jointly, where K is the number of regressors including the constant.
- J restrictions $\Rightarrow$ J rows in the restriction matrix, $\mathbf{R}$. We can express the null hypothesis associated w/ the example above as a set of linear restrictions on the $\boldsymbol{\beta}$ vector, $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$.

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{q} \Rightarrow \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- Given our estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, our interest centers on how much $\mathbf{R}\hat{\boldsymbol{\beta}}$ deviates from $\mathbf{q}$ and we base our hypothesis test on this deviation.
- Let $\mathbf{m} = \mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}$. Intuition: will reject H_0 if $\mathbf{m}$ is too large. Since $\mathbf{m}$ is a vector, we must have some way to measure distance in multi-dimensional space. One possibility—suggested by Wald—is to use the normalized sum of squares, or $\mathbf{m}'\{\text{var}(\mathbf{m})\}^{-1}\mathbf{m}$.
- This is equivalent to estimating $\left(\frac{\mathbf{m}}{\sigma_{\mathbf{m}}}\right)' \left(\frac{\mathbf{m}}{\sigma_{\mathbf{m}}}\right)$.
- If $\mathbf{m}$ can be shown to be normal w/ mean zero, then the statistic above would be made up of a sum of squared standard normal variables (with the number of squared standard normal variables given by the number of restrictions) $\Rightarrow \chi^2$ (Wald statistic).
- We know that $\hat{\boldsymbol{\beta}}$ is normally distributed $\Rightarrow$ $\mathbf{m}$ is normally distributed (Why?).

- Further, $\text{var}(\mathbf{m}) = \text{var}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}) = \mathbf{R}\{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\}\mathbf{R}'$ and so the statistic

$$W = \mathbf{m}'\{\text{var}(\mathbf{m})\}^{-1}\mathbf{m} = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'[\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}) \sim \chi^2$$

- Since, we do not know σ^2 and must estimate it by s^2, we cannot use the Wald statistic. Instead, we derive the sample statistic:

$$F = \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})' \left[\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})/J}{\left[(n-k)s^2/\sigma^2\right]/(n-k)}$$

  [Notice that we are dividing by the same correction factor that we used to prove the appropriateness of the t-test in the single hypothesis case]

- Recall that

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-k} = \frac{\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}}{n-k}.$$

- Since under the null hypothesis $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, then $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q} = \mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{R}\boldsymbol{\beta} = \mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$.

- Using these expressions, which are true under the null, we can transform the sample statistic above so that it is a function of $\boldsymbol{\varepsilon}/\sigma$.

$$F = \frac{\{\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}/\sigma\}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\{\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}/\sigma\}/J}{[\mathbf{M}(\boldsymbol{\varepsilon}/\sigma)]'[\mathbf{M}(\boldsymbol{\varepsilon}/\sigma)]/(n-k)}$$

- The F-statistic is the ratio of two quadratic forms in $\boldsymbol{\varepsilon}/\sigma$.

- Thus, both the numerator and the denominator are, in effect, the sum of "squared" standard normal random variables (where the $\boldsymbol{\varepsilon}/\sigma$ are the standard normal variables) and are therefore distributed χ^2.

- Since it can be shown that the two quadratic forms are independent (see Greene, 6th Ed., p. 85) $\Rightarrow$ the sample statistic above, is distributed as F with degrees of freedom $[J, n-k]$, where J is the number of restrictions and $n-k$ is the degrees of freedom in the estimation of s^2.

- Cancel the σs to get

$$F = \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})/J}{\mathbf{e}'\mathbf{e}/(n-k)}$$

or

$$F = \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'[s^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - q)}{J}$$

- Question posed by this test: how far is the estimated value, $\mathbf{R}\hat{\boldsymbol{\beta}}$, from its value under the null, where this distance is "scaled" in the F test by its estimated variance, equal to $s^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$.

- The F statistic from our running example is:

$$F_{[2,n-k]} = \frac{\begin{bmatrix}\hat{\beta}_2 & \hat{\beta}_3\end{bmatrix} \cdot \begin{bmatrix}\text{var}(\hat{\beta}_2) & \text{cov}(\hat{\beta}_2\hat{\beta}_3) \\ \text{cov}(\hat{\beta}_3\hat{\beta}_2) & \text{var}(\hat{\beta}_3)\end{bmatrix}^{-1} \cdot \begin{bmatrix}\hat{\beta}_2 \\ \hat{\beta}_3\end{bmatrix}}{2}$$

  The sub-scripts indicate that its degrees of freedom are two (for the number of restrictions) and $(n-k)$ for the number of free data points used to estimate σ^2.

- F-statistic can be handy when confronted w/ high collinearity.

6.3 The relationship between F and t for a single restriction

- Assume that in the example above we now have only one restriction reflected in the $\mathbf{R}$ matrix: $\beta_2 = 0$. The relevant F-statistic is given by:

$$F = \frac{(\hat{\beta}_2)'[\text{var}(\hat{\beta}_2)]^{-1}(\hat{\beta}_2)}{1}.$$

  Since $[\text{var}(\hat{\beta}_2)]$ is a matrix containing one element, its inverse is equal to $1/\sigma^2_{\hat{\beta}_2}$.

- Thus, $F = \frac{\hat{\beta}_2^2}{[\text{var}(\hat{\beta}_2)]}$, which is just equal to the square of the estimated t-statistic,

6.4 Calculation of the F-statistic using the estimated residuals

- Another approach to using the F-test is to regard it as a measure of "loss of fit." We'll show that this is equivalent although it may be easier to implement computationally.
- Imposing any restrictions on the model is likely to increase the sum of the squared errors, simply because the minimization of the error terms is being done subject to a constraint. On the other hand, if the null hypothesis is true, then the restrictions should not substantially increase the error term. Therefore, we can build a test of the null hypothesis from the estimated residuals, which will result in an F-statistic equivalent to the one above.
- Let $\mathbf{e}^*$ equal the estimated residuals from the restricted model $\Rightarrow \mathbf{e}^* = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^*$.
- Adding and subtracting $X\hat{\boldsymbol{\beta}}$ from the restricted model we get:

$$\mathbf{e}^* = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) = \mathbf{e} - \mathbf{X}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}})$$

- The sum of squared residuals in the restricted model is:

$$\mathbf{e}^{*\prime}\mathbf{e}^* = (\mathbf{e}' - (\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}})'\mathbf{X}')(\mathbf{e} - \mathbf{X}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}})) = \mathbf{e}'\mathbf{e} + (\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \geq \mathbf{e}'\mathbf{e}$$

- The loss of fit is equal to:

$$\mathbf{e}^{*\prime}\mathbf{e}^* - \mathbf{e}'\mathbf{e} = (\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})$$

- In other words, the loss of fit can also be seen as (most of) the usual numerator from the F-statistic. Under the null, $\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}}$, thus $(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})$ is equal to $(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}})$ and $\mathbf{R}$ is equal to the identity matrix. Thus, we simply need to introduce our s^2 estimate and divide by J, the number of restrictions, to obtain an F-statistic.

$$F_{J,n-k} = \frac{(\mathbf{e}^{*\prime}\mathbf{e}^* - \mathbf{e}'\mathbf{e})/J}{\mathbf{e}'\mathbf{e}/(n-k)}$$

- Thus, an equivalent procedure for calculating the F-statistic is to run the restricted model, calculate the residuals, run the unrestricted model, calculate this set of residuals, and construct the F-statistic based on the expression above.
- Since $\frac{\mathbf{e}'\mathbf{e}}{SST} = 1 - R^2$, we can also divide through the above expression by the SST, to yield:

$$F_{J,n-k} = \frac{(R^2 - R^{*2})/J}{(1 - R^2)/(n - k)}$$

  where R^{*2} is the coefficient of determination in the restricted model.
- If we want to test the significance of the model as a whole, we use:

$$F_{J,n-k} = \frac{(R^2)/(k - 1)}{(1 - R^2)/(n - k)}$$

6.5 Tests of structural change

- We often pool different sorts of data in order to test hypotheses of interest. For instance, we are liable to cumulate different U.S. Presidents in a time-series used to test the relationship between the economy and approval and to pool different countries when testing for the relationship between democracy and income levels.
- In some cases, this pooling is inappropriate because we would not expect the causal mechanisms that underlie our model to function identically across different countries or time-periods.
- Tests for structural change of model stability (generally known as Chow Tests) allow us to test whether the model is significantly different for two different time periods or different groups of observations (e.g., countries).

- Suppose that we are interested in how a set of explanatory variables affects a dependent variable before and after some critical juncture. Denote the data before the juncture as $\mathbf{y}_1$ and $\mathbf{X}_1$ and after as $\mathbf{y}_2$ and $\mathbf{X}_2$. We also allow the coefficients across the two periods to vary.

- We wish to test an unrestricted model

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix}$$

  against the restricted model

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \boldsymbol{\beta}^* + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix}$$

  This is just a special case of linear restrictions $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ where $\mathbf{R} = \begin{bmatrix} \mathbf{I} & -\mathbf{I} \end{bmatrix}$ and $\mathbf{q} = 0$.

- The restricted sum of squares is $\mathbf{e}^{*\prime}\mathbf{e}^*$, which we estimate using all $n_1 + n_2$ observations (so that this is just the normal sum of squared errors). The unrestricted sum of squares is $\mathbf{e}_1'\mathbf{e}_1 + \mathbf{e}_2'\mathbf{e}_2$ which is estimated from each sub-sample separately. Thus, we can construct an F-test from:

$$F_{[J,n-k]} = \frac{(\mathbf{e}^{*\prime}\mathbf{e}^* - \mathbf{e}_1'\mathbf{e}_1 - \mathbf{e}_2'\mathbf{e}_2)/K}{(\mathbf{e}_1'\mathbf{e}_1 + \mathbf{e}_2'\mathbf{e}_2)/(n_1 + n_2 - 2K)}$$

- In this case, the number of restrictions, J, is equal to the number of coefficients in the restricted model, K, because we are assuming that each one of these coefficients is the same across sub-samples. If the F-statistic is sufficiently large, we will reject the hypothesis that the coefficients are equal.

- This test can also be performed using dummy variables.

6.5.1 The creation and interpretation of dummy variables

- So far, we have mainly discussed continuous explanatory variables. Now we turn to the creation and interpretation of dummy variables that take on the value zero or one to indicate membership of a binary category and discuss their uses as explanatory variables.
- Suppose that we are trying to determine whether money contributed to political campaigns varies across regions of the country (e.g., northeast, south, midwest, and west).
- We can estimate: $Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + \varepsilon_i$

  Where

  - ➢ $D_1 = 1$ if the contributor is from the south.
  - ➢ $D_2 = 1$ if the contributor is from the midwest.
  - ➢ $D_3 = 1$ if the contributor is from the west.
- This type of statistical analysis is called ANOVA (or analysis of variance) and informs you about how contribution levels vary by region.
- What is the interpretation of the intercept in this regression and what is the interpretation of the dummy variables and their t-statistics?

6.5.2 The "dummy variable trap"

- We do not include all the dummy variables in the regression above. Why?
- There is a constant in the regression above. The constant appears in the $\mathbf{X}$ data matrix as a column of ones. Each of the dummy variables appears in the data matrix as a column of zeroes and ones. Since they are mutually exclusive and collectively exhaustive, if we were to introduce them all into the data matrix, the sum:

$$D_{1i} + D_{2i} + D_{3i} + D_{4i} = 1$$

  for all observations.

- If we include all of them in the regression *with* a constant, then the data matrix $\mathbf{X}$ does not have full rank, it has rank at most of $K-1$ (perfect collinearity). In this case, the matrix $(\mathbf{X}'\mathbf{X})$ is also not of full rank and has rank of at most $K-1$.
- Another way of putting this, is that we do not have enough information to estimate the constant separately if we include all the regional dummies because, once we have them in, the overall model constant is completely determined by the individual regional intercepts and the weight of each region. It cannot be separately estimated, because it is already given by what is in the model.
- If you include a constant in the model, you can only include $M-1$ of M mutually exclusive and collectively exhaustive dummy variables. Or could drop the constant.

6.5.3 Using dummy variables to estimate separate intercepts and slopes for each group

- The model above allowed us to estimate different intercepts for each region in a model of political contributions. Suppose I am now running a model that looks at the effect of gender on campaign contributions. We might also expect personal income to affect the level of donations and include it in our model as well.
- Once we include both continuous and categorical variables into the same model, we are no longer performing ANOVA but ANCOVA (analysis of covariance). If we were interested only in estimating a model looking at the effects of gender and income on political contributions, our model would take the following form:

$$y_i = \beta_0 + \beta_1 d_{m,i} + \beta_2 X_i + \varepsilon_i$$

  This model simply estimates a different intercept for men.

- Suppose, however, we think that men will contribute more to political campaigns as their income rises than will women. To test this hypothesis, we would estimate the model:

$$Y_i = \beta_0 + \beta_1 d_{m,i} + \beta_2 X_i + \beta_3 d_{m,i} \cdot X_i + \varepsilon_i$$

  To include the last term in the model, simply multiply the income variable by the gender dummy. If income is measured in thousands of dollars, what is the marginal effect of an increase of $1,000 dollars for men on expected contributions to political campaigns?

- How do we test whether gender has any impact on the marginal effect of income on political contributions? How do we test the hypothesis that gender has no impact on the behavior underlying political contributions?

- It is generally a bad idea to estimate interaction terms without including the main effects.

6.6 The use of dummy variables to perform Chow tests of structural change

- One of the more common applications of the F-test is in tests of structural change (or parameter stability) due to Chow. The test basically asks where the behavior that one is modeling varies across different sub-sets of observations.

- We often estimate a causal model using samples that combine men and women, rich and poor countries, and different time periods. Yet we can hypothesize that the behavior that we are modeling varies by gender, income level and time period. If that is correct, then restricting the different sub-sets of observations is inappropriate.

- For example, in the model of political contributions, if I assume that giving to political campaigns does depend on gender, I would estimate the "unrestricted" model:

$$Y_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 d_{m,i} + \boldsymbol{\beta}_2 X_i + \boldsymbol{\beta}_3 d_{m,i} \cdot X_i + \varepsilon_i$$

- If, however, I impose the restriction that gender does not matter, then I estimate the restricted model:

$$Y_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_2 X_i + \varepsilon_i$$

  In this case, I am restricting $\boldsymbol{\beta}_1$ to be equal to zero and $\boldsymbol{\beta}_3$ to be equal to zero.

- In general terms, we wish to test the unrestricted model where the OLS coefficients differ by gender.

- The Chow test to determine whether the restriction is valid is performed using an F-test and can be accomplished using the approach outlined above.

- This version of the Chow Test is exactly equivalent to running the unrestricted model and then performing an F-test on the joint hypothesis that $\boldsymbol{\beta}_1 = 0$ and $\boldsymbol{\beta}_3 = 0$. Why is this?

- We saw previously that an F-test could also be computed using:

$$F_{J,n-k} = \frac{(\mathbf{e}^{*\prime}\mathbf{e}^* - \mathbf{e}'\mathbf{e})/J}{\mathbf{e}'\mathbf{e}/(n-k)}$$

  This is just the same as we did above. We are simply getting the unrestricted error sum of squares from the regressions run on men and women separately.

- This form of the F-test could be derived as an extension of testing a restriction on the OLS coefficients.

- Under the restriction, the $\boldsymbol{\beta}$ vector of regression coefficients was equal to $\boldsymbol{\beta}^*$. In the case outlined above:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \boldsymbol{\beta}_3 \end{bmatrix} \text{ and } \boldsymbol{\beta}^* = \begin{bmatrix} \boldsymbol{\beta}_0 \\ 0 \\ \boldsymbol{\beta}_2 \\ 0 \end{bmatrix}$$

- But the null hypothesis is given by $\boldsymbol{\beta} = \boldsymbol{\beta}^*$. In terms of our notation for expressing a hypothesis as a restriction on the matrix of coefficients, $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, we have $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, so $\mathbf{R}$ is equal to the identity matrix. Moreover, the distance $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q}$ is now given by $\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}$.

- As a result, the quantity $(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) = \mathbf{e}^{*\prime}\mathbf{e}^* - \mathbf{e}'\mathbf{e}$ can be written as:
$$(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{q})$$
- This is most of the F-statistic, except that we now have to divide by s^2, equal to $\mathbf{e}'\mathbf{e}/(n-k)$, our estimate of the variance of the true errors, and divide by J to get the same F-test that we would get above.
- Testing this restriction on the vector of OLS coefficients, however, is exactly the same thing as testing that $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_3$ are jointly equal to zero, and the latter computing procedure is far easier in statistical software.

6.6.1 A note on the estimate of s^2 used in these tests

- In using these Chow tests, we have assumed that σ^2 is the same across sub-samples, so that it is okay if we use the s^2 estimate from the restricted or unrestricted model (although in practice we use s^2 from the unrestricted model).
- If this is not persuasive, there is an alternative. Suppose that $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ are two normally distributed estimators of a parameter based on independent sub-samples with variance-covariance matrices $\mathbf{V}_1$ and $\mathbf{V}_2$. Then, under the null hypothesis that the two estimators have the same expected value, $\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2$ has mean $\mathbf{0}$ and variance $\mathbf{V}_1 + \mathbf{V}_2$. Thus, the Wald statistic:
$$W = (\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2)'[\mathbf{V}_1 + \mathbf{V}_2]^{-1}(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2)$$
has a chi-square distribution with K degrees of freedom and a test that the difference between the two parameters is equal to zero can be based on this statistic.
- What remains for us to resolve is whether it is valid to base such a test on estimates of $\mathbf{V}_1$ and $\mathbf{V}_2$. We shall see, when we examine asymptotic results from infinite samples that we can do so as our sample becomes sufficiently large.

- Finally, there are some more sophisticated techniques which allow you to be agnostic about the locations of the breaks (i.e., let the data determine where they are).

Section 7

Partitioned Regression and Bias

7.1 Partitioned regression, partialling-out and applications

- We will first examine the OLS formula for the regression coefficients using "partitioned regression" as a means of better understanding the effects of excluding a variable that should have been included in the model.
- Suppose our regression model contains two sets of variables, $\mathbf{X}_1$ and $\mathbf{X}_2$. Thus:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{bmatrix} \cdot \begin{bmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{bmatrix} + \boldsymbol{\varepsilon}$$

- What is the algebraic solution for $\hat{\boldsymbol{\beta}}$, the estimate of $\boldsymbol{\beta}$ in this partitioned version of the regression model?
- To find the solution, we use the "normal equations," $(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{y})$ employing the partitioned matrix, $\begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{bmatrix}$ for $\mathbf{X}$.
- In this case, $(\mathbf{X}'\mathbf{X}) = \begin{bmatrix}\mathbf{X}_1' \\ \mathbf{X}_2'\end{bmatrix} \cdot \begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2\end{bmatrix}$
- The derivation of $(\mathbf{X}'\mathbf{y})$ proceeds analogously.
- The normal equations for this partitioned matrix give us two expressions with two unknowns:

$$\begin{bmatrix}\mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2\end{bmatrix} \cdot \begin{bmatrix}\hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y}\end{bmatrix}$$

- We can solve directly for $\hat{\boldsymbol{\beta}}_1$, the first set of coefficients, by multiplying through the first row of the partitioned matrix and then using the trick of pre-multiplying by $(\mathbf{X}_1'\mathbf{X}_1)^{-1}$:

$$(\mathbf{X}_1'\mathbf{X}_1)\hat{\boldsymbol{\beta}}_1 + (\mathbf{X}_1'\mathbf{X}_2)\hat{\boldsymbol{\beta}}_2 = \mathbf{X}_1'\mathbf{y}$$

  and

$$\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} - (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{y} - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2)$$

- We can conclude from the above that if $\mathbf{X}_1'\mathbf{X}_2 = 0$—i.e., if $\mathbf{X}_1$ and $\mathbf{X}_2$ are orthogonal—then the estimated coefficient $\hat{\boldsymbol{\beta}}_1$ from a regression of $\mathbf{y}$ on $\mathbf{X}_1$ will be the same as the estimated coefficients from a regression of $\mathbf{y}$ on $\mathbf{X}_1$ and $\mathbf{X}_2$.

- If not, however, and if we mistakenly omit $\mathbf{X}_2$ from the regression, then reporting the coefficient of $\hat{\boldsymbol{\beta}}_1$ as $(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$ will over-state the true coefficient by $(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\hat{\boldsymbol{\beta}}_2$.

- The solution for $\hat{\boldsymbol{\beta}}_2$ is found by multiplying through the second row of the matrix and substituting the above expression for $\hat{\boldsymbol{\beta}}_1$. This method for estimating the OLS coefficients is called "partialling out."

$$(\mathbf{X}_2'\mathbf{X}_1)\hat{\boldsymbol{\beta}}_1 + (\mathbf{X}_2'\mathbf{X}_2)\hat{\boldsymbol{\beta}}_2 = \mathbf{X}_2'\mathbf{y}$$

- Now substitute for $\hat{\boldsymbol{\beta}}_1$.

$$(\mathbf{X}_2'\mathbf{X}_1)[(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} - (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\hat{\boldsymbol{\beta}}_2] + (\mathbf{X}_2'\mathbf{X}_2)\hat{\boldsymbol{\beta}}_2 = \mathbf{X}_2'\mathbf{y}$$

  Multiplying through and rearranging gives:

$$[\mathbf{X}_2'\mathbf{X}_2 - \mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2]\hat{\boldsymbol{\beta}}_2 = \mathbf{X}_2'\mathbf{y} - \mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$$
$$\mathbf{X}_2'[\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1']\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{X}_2'[\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1']\mathbf{y}$$

  Solving for $\hat{\boldsymbol{\beta}}_2$ gives

$$\begin{aligned}\hat{\boldsymbol{\beta}}_2 &= [\mathbf{X}_2'(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{X}_2]^{-1}[\mathbf{X}_2'(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{y}] \\ &= (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}(\mathbf{X}_2'\mathbf{M}_1\mathbf{y})\end{aligned}$$

- Here, the matrix $\mathbf{M}$ is once again the "residual-maker," with the sub-script indicating the relevant set of explanatory variables. Thus:

  $\mathbf{M}_1\mathbf{X}_2$ = a vector of residuals from a regression of $\mathbf{X}_2$ on $\mathbf{X}_1$.

  $\mathbf{M}_1\mathbf{y}$ = a vector of residuals from a regression of $\mathbf{y}$ on $\mathbf{X}_1$.

- Using the fact that $\mathbf{M}_1$ is idempotent, so that $\mathbf{M}_1'\mathbf{M}_1 = \mathbf{M}_1$, and setting $\mathbf{X}_2^* = \mathbf{M}_1\mathbf{X}_2$ and $\mathbf{y}^* = \mathbf{M}_1\mathbf{y}$, we can write the solution for $\hat{\boldsymbol{\beta}}_2$ as:

$$\hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_2^{*'}\mathbf{X}_2^*)^{-1}\mathbf{X}_2^{*'}\mathbf{y}^*$$

- Thus, we have an entirely separate approach by which we can estimate the OLS coefficient $\hat{\boldsymbol{\beta}}_2$ when $\mathbf{X}_2$ is a single variable:

  1. Regress $\mathbf{X}_2$ on $\mathbf{X}_1$ and calculate the residuals.
  2. Regress $\mathbf{y}$ on $\mathbf{X}_1$ and calculate the residuals.
  3. Regress the second set of residuals on the first. The first round of regressions "partial out" the effect of $\mathbf{X}_1$ on $\mathbf{X}_2$ and $\mathbf{y}$, allowing us to calculate the marginal effect of $\mathbf{X}_2$ on $\mathbf{y}$, accounting for $\mathbf{X}_1$.

- The process can also be carried out in reverse, yielding:

$$\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1^{*'}\mathbf{X}_1^*)^{-1}\mathbf{X}_1^{*'}\mathbf{y}^*$$

7.2 R^2 and the addition of new variables

- Using the partitioning approach, we can prove that R^2 will always rise when we add a variable to the regression.
- Start with the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$ where $\mathbf{b} = \hat{\boldsymbol{\beta}}$
- Add a single variable $\mathbf{z}$ with the coefficient c so that $\mathbf{y} = \mathbf{X}\mathbf{d} + \mathbf{z}c + \mathbf{u}$ ($\mathbf{d} \neq \mathbf{b}$ if $\mathbf{X}_1'\mathbf{X}_2 \neq 0$).
- R^2 will increase if $\mathbf{u}'\mathbf{u} < \mathbf{e}'\mathbf{e}$. To find out if this inequality holds, we have to find an expression relating $\mathbf{u}$ and $\mathbf{e}$.

- From the formula for partitioned regression above, we can say that:

$$\mathbf{d} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \mathbf{z}c) = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}c$$

  and we know that $\mathbf{u} = \mathbf{y} - \mathbf{X}\mathbf{d} - \mathbf{z}c$.

- Inserting the expression for $\mathbf{d}$ into this equation for the residuals, we get:

$$\mathbf{u} = \mathbf{y} - \mathbf{X}\mathbf{b} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}c - \mathbf{z}c \tag{7.1}$$

- Since $\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{M}$ and $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{I} - \mathbf{M}$, we can rewrite Eq. 7.1 as:

$$\mathbf{u} = \mathbf{e} - \mathbf{M}\mathbf{z}c = \mathbf{e} - \mathbf{z}_*c$$

- Thus,

$$\mathbf{u}'\mathbf{u} = \mathbf{e}'\mathbf{e} + c^2(\mathbf{z}_*'\mathbf{z}_*) - 2c\mathbf{z}_*'\mathbf{e}$$

  $\mathbf{e} = \mathbf{M}\mathbf{y} = \mathbf{y}_*$ and $\mathbf{z}_*'\mathbf{e} = \mathbf{z}_*'\mathbf{y}_*$. Since $c = (\mathbf{z}_*'\mathbf{z}_*)^{-1}(\mathbf{z}_*'\mathbf{y}_*)$, we can say that $\mathbf{z}_*'\mathbf{e} = \mathbf{z}_*'\mathbf{y}_* = c(\mathbf{z}_*'\mathbf{z}_*)$.

- Therefore:

$$\mathbf{u}'\mathbf{u} = \mathbf{e}'\mathbf{e} + c^2(\mathbf{z}_*'\mathbf{z}_*) - 2c^2(\mathbf{z}_*'\mathbf{z}_*) = \mathbf{e}'\mathbf{e} - c^2(\mathbf{z}_*'\mathbf{z}_*)$$

  Since the last matrix is composed of a sum of squares, it is positive definite.

- Thus, $\mathbf{u}'\mathbf{u} < \mathbf{e}'\mathbf{e}$ unless $c = 0$. In other words, adding another variable to the regression will always result in a higher R^2 unless the OLS coefficient on the additional variable is exactly equal to zero.

7.3 Omitted variable bias

- The question of which variables should be included in a regression model is part of the larger question of selecting the right specification. Other issues in specification testing include whether you have the right functional form and whether the parameters are the same across sub-samples.
- We will now use a different approach to prove that the OLS coefficients will normally be biased when you exclude a variable from the regression that is part of the true, population model.
- Suppose that the correctly specified regression model would be:

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

- If we don't know this and assume that the correct model is $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$, then we will regress $\mathbf{y}$ on $\mathbf{X}_1$ only. Then our estimate of $\boldsymbol{\beta}_1$will be:

$$\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$$

- Now substitute $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$ from the true equation for $\mathbf{y}$ in the expression for $\hat{\boldsymbol{\beta}}_1$:

$$\hat{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\boldsymbol{\varepsilon}$$

- To check for bias (i.e. whether $E(\hat{\boldsymbol{\beta}}_1) \neq \boldsymbol{\beta}_1$), take the expectation:

$$E(\hat{\boldsymbol{\beta}}_1) = \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2$$

- Thus, unless $(\mathbf{X}_1'\mathbf{X}_2) = \mathbf{0}$ or $\boldsymbol{\beta}_2 = \mathbf{0}$, $\hat{\boldsymbol{\beta}}_1$ is biased. Also, note that $(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2$ gives the coefficient(s) in a regression of $\mathbf{X}_2$ on $\mathbf{X}_1$.

7.3.1 Direction of the Bias

- If we have just two explanatory variables in our data set, then the coefficient from a regression of X_2 on X_1 will be equal to $\frac{\text{cov}(X_1 X_2)}{\text{var}(X_1)}$.
- Thus, if we had a good idea of the sign of the covariance and the sign of $\boldsymbol{\beta}_2$, we would be able to predict the direction of the bias.
- If there are more than two explanatory variables in the regression, then the direction of the bias on $\hat{\boldsymbol{\beta}}_1$ when we exclude X_2 will depend on the partial correlation between X_1 and X_2 (since this always has the same sign as the regression coefficient in the OLS regression model) and the sign of the true $\boldsymbol{\beta}_2$.

7.3.2 Bias in our estimate of σ^2

- In order to calculate the standard error of the coefficients, we need to estimate σ^2. If we did this just from the regression including X_1 we would get:

$$s^2 = \frac{\mathbf{e}_1'\mathbf{e}_1}{n - K_1}$$

but $\mathbf{e}_1 = \mathbf{M}_1\mathbf{y} = \mathbf{M}_1(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}) = \mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{M}_1\boldsymbol{\varepsilon}$.

(Note that $\mathbf{M}_1\mathbf{X}_1 = \mathbf{0}$.)

- Does $E(\mathbf{e}_1'\mathbf{e}_1) = \sigma^2(n - K_1)$?

$$\begin{aligned} E(\mathbf{e}_1'\mathbf{e}_1) &= E[(\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{M}_1\boldsymbol{\varepsilon})'(\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{M}_1\boldsymbol{\varepsilon})] \\ &= E[(\boldsymbol{\beta}_2'\mathbf{X}_2'\mathbf{M}_1' + \boldsymbol{\varepsilon}'\mathbf{M}_1')(\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{M}_1\boldsymbol{\varepsilon})] \\ &= \boldsymbol{\beta}_2'\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 + E(\boldsymbol{\beta}_2'\mathbf{X}_2'\mathbf{M}_1\boldsymbol{\varepsilon}) + E(\boldsymbol{\varepsilon}'\mathbf{M}_1'\mathbf{X}_2\boldsymbol{\beta}_2) + E(\boldsymbol{\varepsilon}'\mathbf{M}_1\boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta}_2'\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 + E(\boldsymbol{\varepsilon}'\mathbf{M}_1\boldsymbol{\varepsilon}) \end{aligned}$$

This last step involves multiplying out $\mathbf{M}_1\mathbf{X}_2$ and using $E(\boldsymbol{\varepsilon}'\mathbf{X}_2) = 0$ and $E(\mathbf{X}_1'\boldsymbol{\varepsilon}) = 0$, both of which we know from the Gauss-Markov assumptions.

- From here we proceed as we did when working out the expectation of the sum of squared errors originally:

$$E(\mathbf{e}_1'\mathbf{e}_1) = \boldsymbol{\beta}_2'\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 + \sigma^2 tr(\mathbf{M}_1) = \boldsymbol{\beta}_2'\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 + (n - K_1)\sigma^2$$

The first term in the expression is the increase in the error sum of squares (SSE) that results when $\mathbf{X}_2$ is dropped from the regression.

- We can show that $\boldsymbol{\beta}_2^{'}\mathbf{X}_2^{'}\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2$ is positive, so the s^2 derived from the regression of $\mathbf{y}$ on X_1 will be biased upward (Intuition: the shorter version of the model leaves more variance unaccounted for).
- Omitted variable bias is a result of a violation of the Gauss-Markov assumptions: we assumed we had the correct model.

7.3.3 Testing for omitted variables: The RESET test

- The RESET test (due to Ramsey) checks for patterns in the estimated residuals against the predicted values of $\mathbf{y}$, $\hat{\mathbf{y}}$.
- Specifically, the test is conducted by estimating the original regression, computing the predicted values, and then using powers of these predicted values, $\hat{y}_i^2$, $\hat{y}_i^3$ and of the original regressors, $\mathbf{X}$, as explanatory variables in an expanded regression. An F-test on these additional regressors is then employed to detect whether or not they are jointly significant.

7.4 Including an irrelevant variable

- What if the true model is $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1+\boldsymbol{\varepsilon}$ and we estimate $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1+\mathbf{X}_2\boldsymbol{\beta}_2+\boldsymbol{\varepsilon}$? We are likely to find that the estimated coefficient $\hat{\boldsymbol{\beta}}_2$ cannot be distinguished from zero.
- $\hat{\boldsymbol{\beta}}_1$ will not be biased, nor will s^2. [To verify, use $\boldsymbol{\beta}_2 = 0$ in the above equations].
- This does not mean we should estimate "kitchen sink" regressions.
- The cost of including an irrelevant variable is a loss in efficiency. It can be shown that the variance-covariance matrix of the coefficients always rises when we add a variable, which means we will estimate the sampling distributions of the coefficients less precisely. Our standard errors will rise and our t-statistics will fall.

7.5 Model specification guidelines

- Think about theoretically likely confounding variables; alternative hypotheses.
- Avoid "stepwise regression".
  - ➢ A common technique of model building is to build it iteratively by adding regressors sequentially and keeping those that are statistically significant. This is a very problematic technique.
  - ➢ Consider a simple case of adding columns from $\mathbf{X}$ sequentially and discarding those whose estimated coefficient is not statistically different from zero. First, we estimate $\mathbf{y} = \beta_0 + \mathbf{x}_1\beta_1 + \boldsymbol{\varepsilon}$, a simple model with an intercept and one explanatory variable.
  - ➢ We will keep $\mathbf{x}_1$ in if $\frac{|\hat{\beta}_1|}{se(\hat{\beta}_1)} > t_{\alpha/2}$ (the t-statistic is bigger than the critical value).
  - ➢ The problem is that unless $\mathbf{y} = \beta_0 + \mathbf{x}_1\beta_1 + \boldsymbol{\varepsilon}$ is the true model, $\hat{\beta}_1$ and $se(\hat{\beta}_1)$ will be biased. So our decision of what to include will be based on biased estimates.
- Avoid "Data Mining"
  - ➢ Data mining refers to the practice of estimating lots of different models to see which has the best fit and gives statistically significant coefficients.
  - ➢ The problem is that if you are using a 95% confidence interval (size of the test, α, is .05) then the probability that you will reject the null hypothesis that $\boldsymbol{\beta} = 0$, *when it is true*, is 5%.
  - ➢ If you try out twenty different variables, all of which actually have true $\boldsymbol{\beta} = 0$, your probability of finding one that looks significant in this sample is approximately one. Thus, when you are doing repeated tests, your probability of incorrectly rejecting the null is actually much higher than the given size of the test.

- Try "Cross Validation"
  - ➢ Randomly divide your data set in half. Select your model using the first data set and guided by theory.
  - ➢ Next, after deciding on the right model using the first data set, apply it to the second data set. If any of the variables are not significant in your second data set, there's a good chance that they were incorrectly included to begin with.
- Let theory be your guide
  - ➢ Include variables that are theoretically plausible or necessary for accounting for relevant (but theoretically uninteresting) phenomenon.
  - ➢ If theory is not a good guide, for instance when it comes to determining the number of lags in a time series model, you could try using one of the criterion used in the profession for model selection, including the Akaike Information Criterion.

7.6 Multicollinearity

- Partitioned regression can be used to show that if you add another variable to a regression, the standard errors of the previously included coefficients will rise and their t-statistics will fall.
- Magnitude of change depends on covariance b/t variables.
- The estimated variance of $\hat{\beta}_k$ is given by

$$\widehat{\text{var}}(\hat{\beta}_k) = s^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}$$

  $\Rightarrow$ esimated variance depends on the presence of other variables in the regression via their influence on the kth diagonal element in the inverse matrix $(\mathbf{X}'\mathbf{X})^{-1}_{kk}$.
- We can use the formula for the inverse of a partitioned matrix to say something about what this element will be.

- Let the true model be equal to $Y_i = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$.
- Measure all the data as deviations from means, so that the transformed model is $Y_i^* = \alpha_1 X_{1i}^* + \alpha_2 X_{2i}^* + \varepsilon_i$, where $\alpha_1 = \beta_1$,$\alpha_2 = \beta_2$ and $X_{1i}^* = (x_{1i} - \bar{x}_1)$, etc.
- You can think of this as the case in which we had first "partialled out" the intercept, by regressing all the variables on an intercept, and using the residuals from this regression which amount to the deviations from the mean.
- The data matrix, $\mathbf{X}$, is now $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1^* & \mathbf{X}_2^* \end{bmatrix}$ and $\begin{bmatrix} \mathbf{X}_1^{*\prime}\mathbf{X}_1^* & \mathbf{X}_1^{*\prime}\mathbf{X}_2^* \\ \mathbf{X}_2^{*\prime}\mathbf{X}_1^* & \mathbf{X}_2^{*\prime}\mathbf{X}_2^* \end{bmatrix}$
- Multiplying out this last matrix,

$$(\mathbf{X}'\mathbf{X}) = \begin{bmatrix} \sum_i (x_{1i} - \bar{x}_1)^2 & \sum_i (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \\ \sum_i (x_{2i} - \bar{x}_2)(x_{1i} - \bar{x}_1) & \sum_i (x_{2i} - \bar{x}_2)^2 \end{bmatrix}$$

- To find the variance of $\hat{\alpha}_2 = \hat{\beta}_2$, we would need to know $(\mathbf{X}'\mathbf{X})_{22}^{-1}$, the bottom right element of the inverse of the matrix above.
- To find this, we will use a result on the inverse of a partitioned matrix (See Greene, 6th Ed., pp. 966). The formula for the bottom right element of the inverse of a partitioned matrix is given by: $\mathbf{F}_2 = (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}$, where the $\mathbf{A}$'s just refer to the elements of the original partitioned matrix.
- Using this for the above matrix we have:

$$\begin{aligned}
(\mathbf{X}'\mathbf{X})_{22}^{-1} &= \left[\sum (x_{2i} - \bar{x}_2)^2 - \sum (x_{2i} - \bar{x}_2)(x_{1i} - \bar{x}_1) \cdot \frac{1}{\sum (x_{1i} - \bar{x}_1)^2} \right. \\
&\quad \left. \cdot \sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \right]^{-1} \\
&= \left[\mathrm{var}(X_2) \cdot (n-1) - \frac{\mathrm{cov}(X_1, X_2) \cdot (n-1) \cdot \mathrm{cov}(X_1, X_2) \cdot (n-1)}{\mathrm{var}(X_1) \cdot (n-1)} \right]^{-1} \\
&= \left[\mathrm{var}(X_2) \cdot (n-1) - \frac{\mathrm{cov}(X_1, X_2)^2 \cdot (n-1)}{\mathrm{var}(X_1)} \right]^{-1}
\end{aligned}$$

- Multiply through by $1 = \frac{\text{var}(X_2)}{\text{var}(X_2)}$ on both sides of the equation to give:

$$\begin{aligned}(\mathbf{X}'\mathbf{X})^{-1}_{22} &= \left[\left(1 - \frac{(\text{cov}(X_1, X_2))^2}{\text{var}(X_1)\text{var}(X_2)}\right) \cdot \text{var}(X_2)(n-1)\right]^{-1} \\ &= \left[(1 - r^2_{12}) \cdot \text{var}(X_2)(n-1)\right]^{-1} \\ &= \frac{1}{(1 - r^2_{12})\text{var}(X_2)(n-1)}\end{aligned}$$

- In this expression, r^2_{12} is the squared partial correlation coefficient between X_1 and X_2.
- It follows then that: $\widehat{\text{var}}(\hat{\beta}_2) = s^2(\mathbf{X}'\mathbf{X})^{-1}_{kk} = \frac{s^2}{(1-r^2_{12})\text{var}(X_2)(n-1)}$ Thus, the variance of the OLS coefficient rises when the two included variables have a higher partial correlation.
- Intuition: the partialling out formula for the OLS regression coefficients demonstrated that each regression coefficient $\hat{\beta}_k$ is estimated from the covariance between $\mathbf{y}$ and $\mathbf{X}_k$ once you have netted out the effect of the other variables. If $\mathbf{X}_k$ is highly correlated with another variable, then there is less "information" remaining once you net out the effect of the remaining variables in your data set.

  This result can be generalized for a multivariate model with K regressors (including the constant) and in which data are entered in standard form, and not in deviations:

$$\text{var}(\hat{\beta}_k) = \frac{s^2}{(1 - R^2_{k.\mathbf{X}-k})\text{var}(\mathbf{X}_k)(n-1)}$$

  where $R^2_{k.X-k}$ is the R^2 from a regression of $\mathbf{X}_k$ on all the other regressors, $\mathbf{X}_{-k}$.

- It follows that the variance of $\hat{\beta}_k$:
  - ➢ rises with σ^2 the variance of ε and its estimate, s^2.
  - ➢ falls as the variance of $\mathbf{X}_k$ rises and falls with the number of data points.
  - ➢ rises when $\mathbf{X}_k$ and the other regressors are highly correlated.
- **Implications for adding another variable**: As you add another variable to the data set, $R^2_{k.X-k}$ will almost always rise $\Rightarrow$ estimating a kitchen sink regression raises the standard errors and reduces the efficiency of the estimates.
- **Implications for the bias in the standard errors when you have an omitted variable**: omitted relevant variable $\Rightarrow s^2$ is biased upwards.
  - ➢ Direction of bias for standard errors is unclear, however.
  - ➢ While s^2 in the numerator increases, the $R^2_{k.X-k}$ will be smaller than it would have been if we had estimated the full model (just know std. errs. are likely to be wrong).
- **Implications for the standard errors when the variables are collinear**: $R^2_{k.X-k}$ will be high, so that the variance and standard errors will be high and thus the t-statistics will be low (but check F-test). If perfectly collinear, $R^2_{k.X-k} = 1$.

7.6.1 How to tell if multicollinearity is likely to be a problem

1. Check the bivariate correlations between variables using the R command `cor`.
2. If overall R^2 is high, but few if any of the t-statistics surpass the critical value $\Rightarrow$ high collinearity.
3. Look at the Variance Inflation Factors (VIFs). The VIF for each variable is equal to:
$$\text{VIF}_k = \frac{1}{1 - R^2_{k.X-k}}$$
Rules of thumb: evidence of multicollinearity if the largest VIF is greater than 10 or the mean of all the VIFs is considerably larger than one. VIFs can be computed in R using Fox's `car` library.

7.6.2 What to do if multicollinearity is a problem

1. Nothing. Report the F-tests as an indication that your variables are jointly significant and the bivariate correlations or VIFs as evidence that multicollinearity is likely to explain the lack of individual significance.
2. Get more data.
3. Amalgamate variables that are very highly collinear into an index of the under-lying phenomenon. The only problem with this approach is that you lose the clarity of interpretation that comes with entering variables individually.

Section 8

Regression Diagnostics

- First commandment of multivariate analysis: "First, know your data."

8.1 Before you estimate the regression

- It is often useful to plot variables in model against each other to get a sense of bivariate correlations (other useful plots: histograms, nonparametric kernel density plots, lowess curves in plots).
- `R`'s graphical capabilities are tremendous, providing numerous canned routines for displaying data. Fox's `car` library augments what is in the standard `R` package. E.g., see Figure 8.1, which was produced by the `pairs` command.

8.2 Outliers, leverage points, and influence points

- An *outlier* is any point that is far from the fitted values. In other words, it has a large residual. Why might we care about checking particular outliers? Two main reasons, although analysts also don't enjoy large outliers because they reduce the fit of the model:
  - ➢ Outliers may signal an error in transcribing data.
  - ➢ A pattern in the outliers may imply an incorrect specification—a missing variable or a likely quadratic relationship in the true model.

8.2.1 How to look at outliers in a regression model

- Plot the residuals against the fitted values $\hat{y}$. If the model is properly specified, no patterns should be visible in the residuals. E.g., see Fig. 8.2.
- You might also want to graph the residuals against just one of the original explanatory variables.

Figure 8.1: Data on Presidential Approval, Unemployment, and Inflation

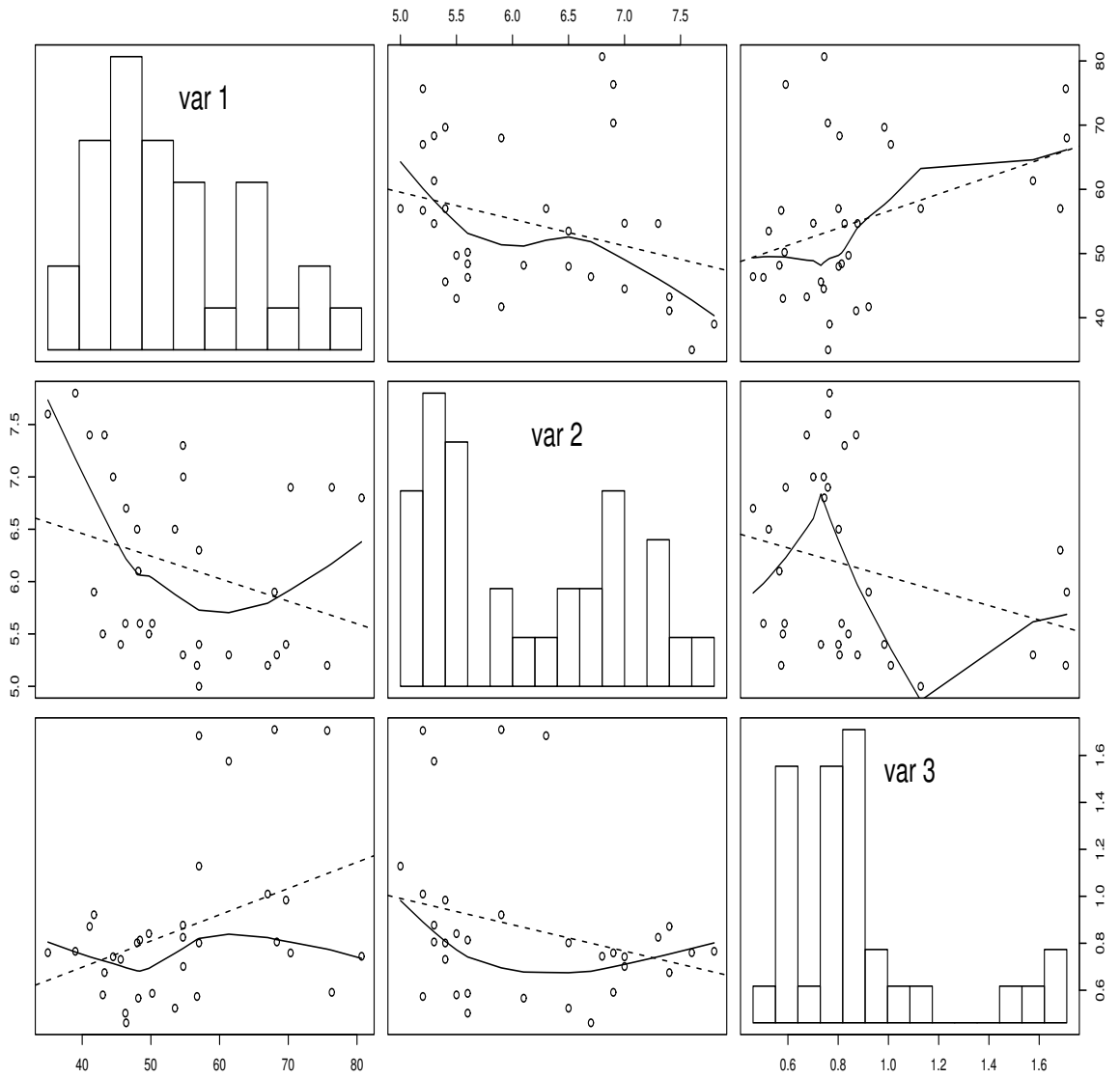
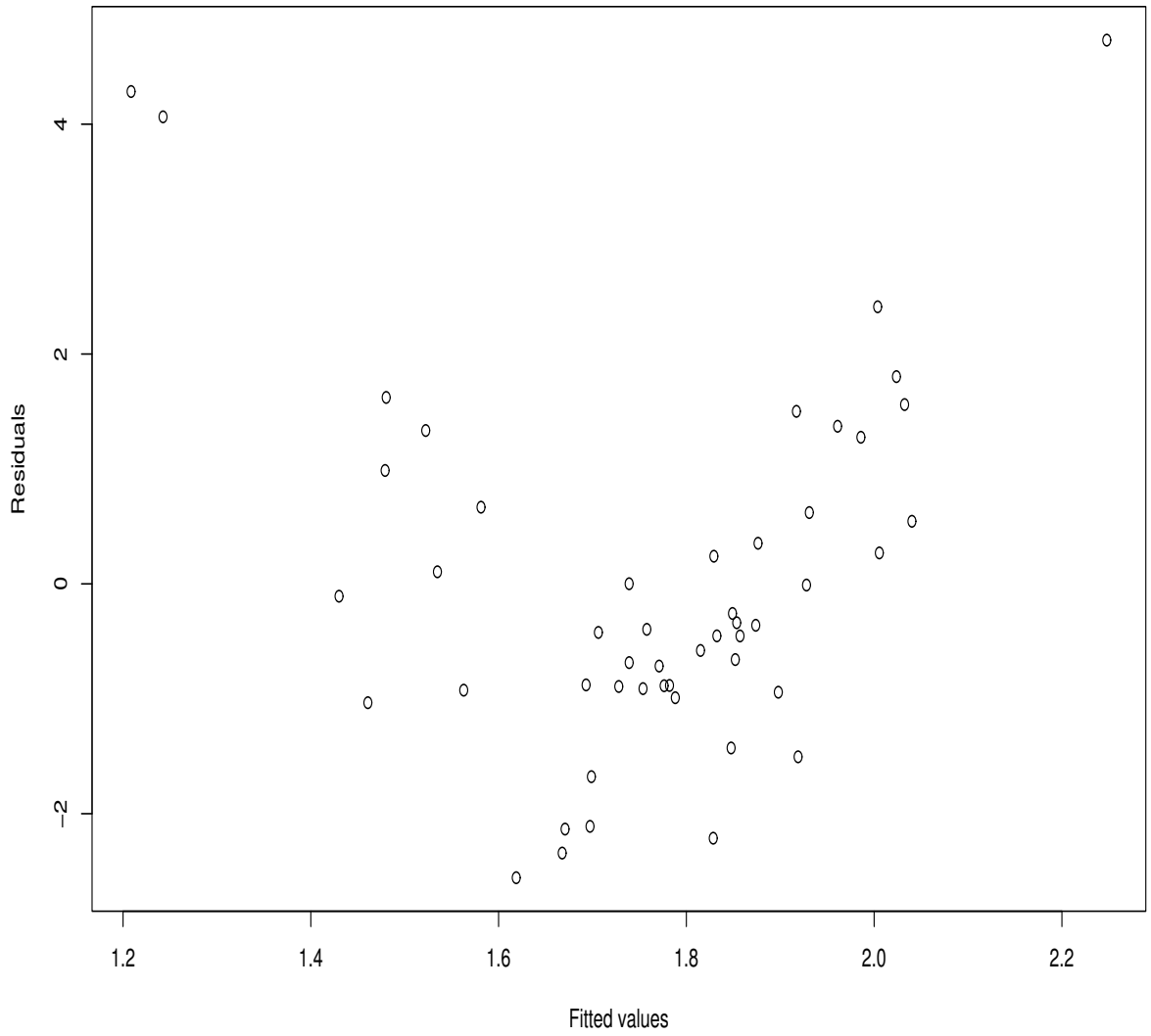


Figure 8.2: A plots of fitted values versus residuals



- It would be nice if we had a method that enabled us to identify influential pairs or subsets of observations. This can be done with an "added variable plot," which reduces a higher-dimension regression problem to a series of two-dimensional plots.
- An added variable plot is produced for a given regressor x_k by regressing both the dependent variable and x_k on all of the other $K-1$ regressors. The residuals from these regressions are then plotted against each other.
- Fig. 8.3 shows an example of some added variable plots from Fox, using the `car` library. This was produced by the commands:

```
library(car)

Duncan <- read.table('/usr/lib/R/library/car/data/Duncan.txt',header=T)

attach(Duncan)

mod.duncan <- lm(prestige~income+education)

av.plots(mod.duncan,labels=row.names(Duncan),ask=F)
```

8.3 How to look at leverage points and influence points

- A leverage point is any point that is far from the mean value of $\mathbf{X} = \bar{\mathbf{X}}$. A point with high leverage is capable of exerting influence over the estimated regression coefficients (the slope of the regression line).
- An influence point is a data point that actually does change the estimated regression coefficient, such that if that point were excluded, the estimated coefficient would change substantially.
- Why do points far from the mean of $\mathbf{X} = \bar{\mathbf{X}}$ have a bigger impact of the fit of the line? Recall that if a constant is included, $\bar{y} = \bar{\mathbf{X}}\beta$, so that $\bar{\mathbf{X}}$, becomes the fulcrum around which the line can tilt. See Fig. 8.4.
- We could also have worked out which were the large residuals in the estimated regression using the `lm` command:

Figure 8.3: Added variable plots

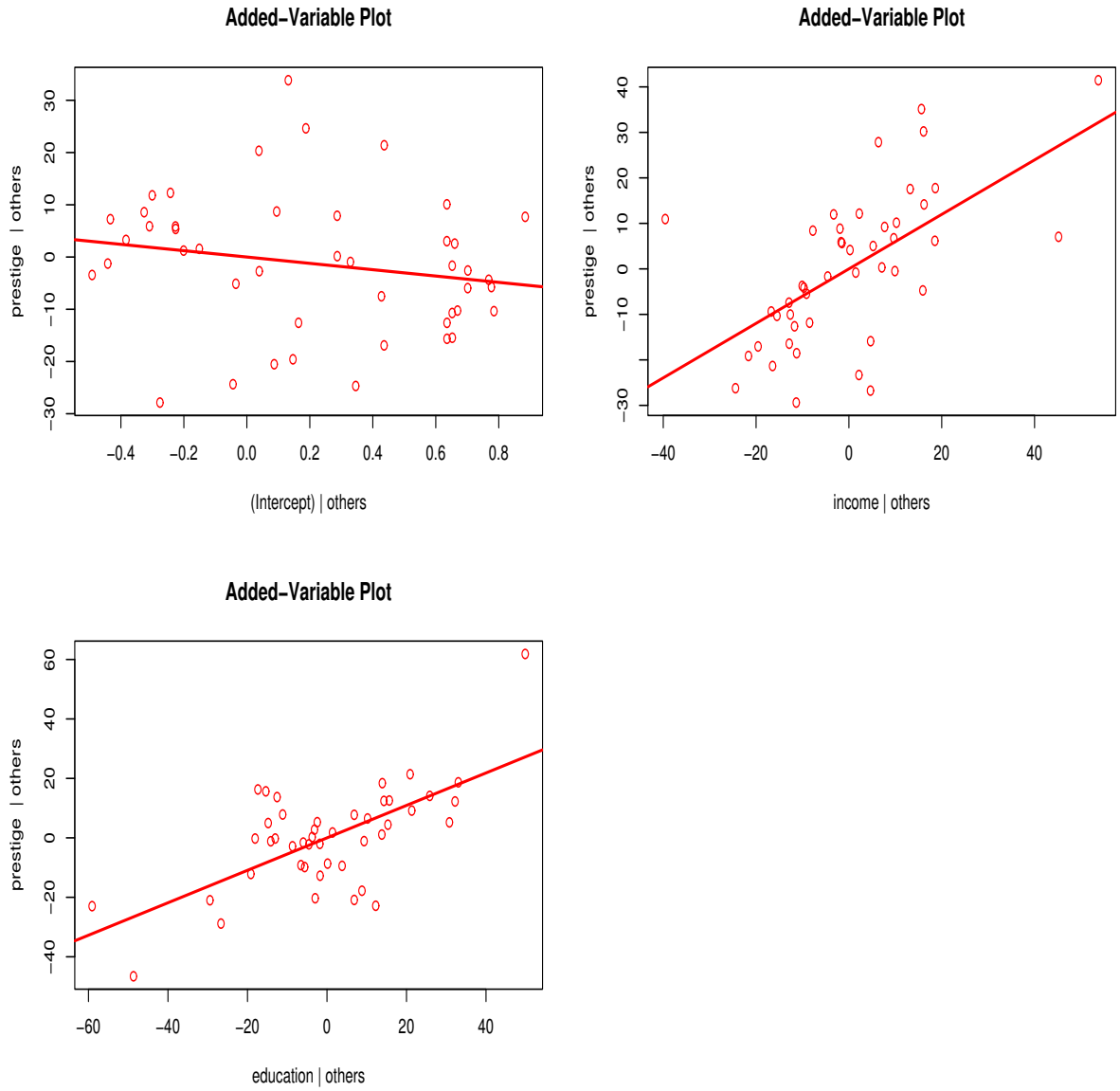
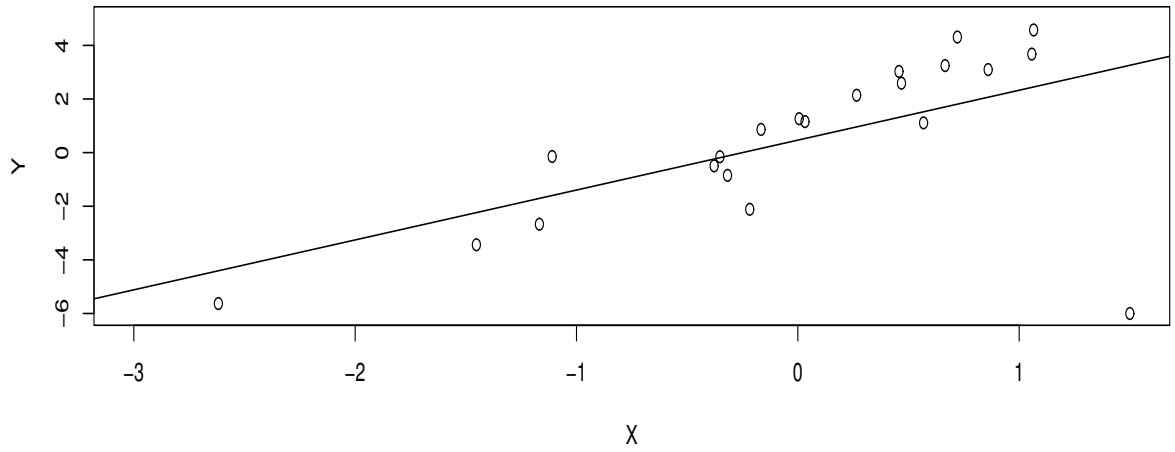
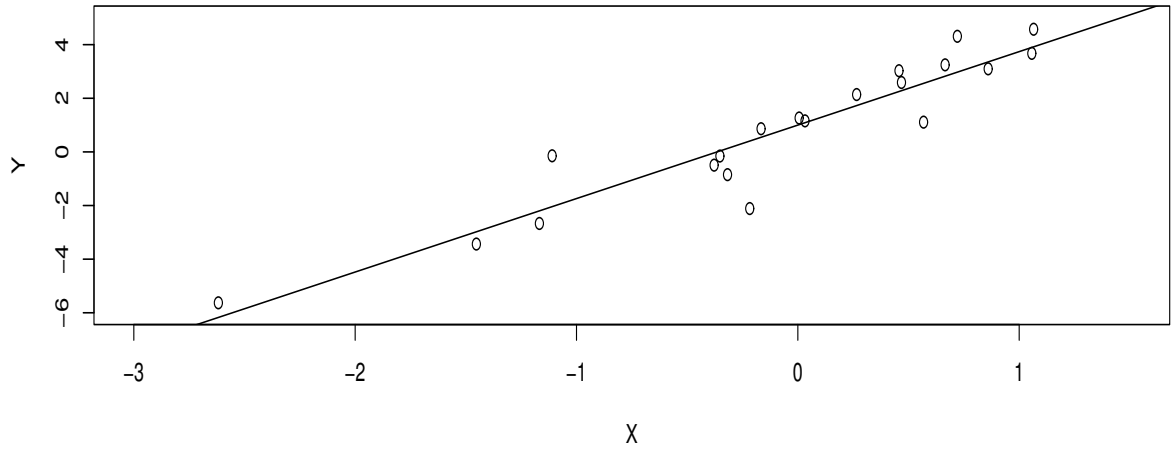


Figure 8.4: An influential outlier





```
xyreg <- lm(Y~X)

xyresid <- resid(xyreg)
```

and then sorting on errors to get the data points organized from highest to lowest residual. This would also allow us to tell the $\mathbf{X}$ values for which the residuals were largest.

8.4 Leverage

- We should draw a distinction between outliers and high leverage observations.
- The regressor values that are most informative (i.e., have the most influence on the coefficient estimates) and lead to relatively large reductions in the variance of coefficient estimates are those that are far removed from the majority of values of the explanatory variables.
- Intuition: regressor values that differ substantially from average values contribute a great deal to isolating the effects of changes in the regressors.
- Most common measure of influence comes from the "hat matrix":

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

- For any vector $\mathbf{y}$, $\mathbf{Hy}$ is the set of fitted values (or $\hat{\mathbf{y}}$ in the least squares regression of y on $\mathbf{X}$. In matrix language, it projects any $n \times 1$ vector into the column space of $\mathbf{X}$.
- For our purposes, $\mathbf{Hy} = \mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\mathbf{y}}$. Recall also that $\mathbf{M} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = (\mathbf{I} - \mathbf{H})$
- The least squares residuals are:

$$\mathbf{e} = \mathbf{My} = \mathbf{M}\boldsymbol{\varepsilon} = (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}$$

- Thus, the variance-covariance matrix for the least squares residual vector is:

$$E[\mathbf{ee}'] = \mathbf{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{M}' = \mathbf{M}\sigma^2\mathbf{M}' = \sigma^2\mathbf{MM}' = \sigma^2\mathbf{M} = \sigma^2(\mathbf{I} - \mathbf{H})$$

- The diagonal elements of this matrix are the variance of each individual estimated residual. Since these diagonal elements are not generally the same, the residuals are heteroskedastic. Since the off-diagonal elements are not generally equal to zero, we can also say that the individual residuals are correlated (they have non-zero covariance). This is the case even when the true errors are homoskedastic and independent, as we assume under the Gauss-Markov assumptions.

- For an individual residual, e_i, the variance is equal to $\text{var}(e_i) = \sigma^2(1 - h_i)$. Here, h_i is one of the diagonal elements in the hat matrix, $\mathbf{H}$. This is considered a measure of *leverage*, since h_i increases in the value of $\mathbf{x}$.

- Average hat values are given by $\bar{h} = (k + 1)/n$ where k is the number of coefficients in the model (including the constant). $2\bar{h}$ and $3\bar{h}$ are typically treated as thresholds that hat values must exceed to be noteworthy.

- Thus, observations with the greatest leverage have corresponding residuals with the smallest variance. Observations with high leverage tend to pull the regression line toward them, decreasing their size and variance.

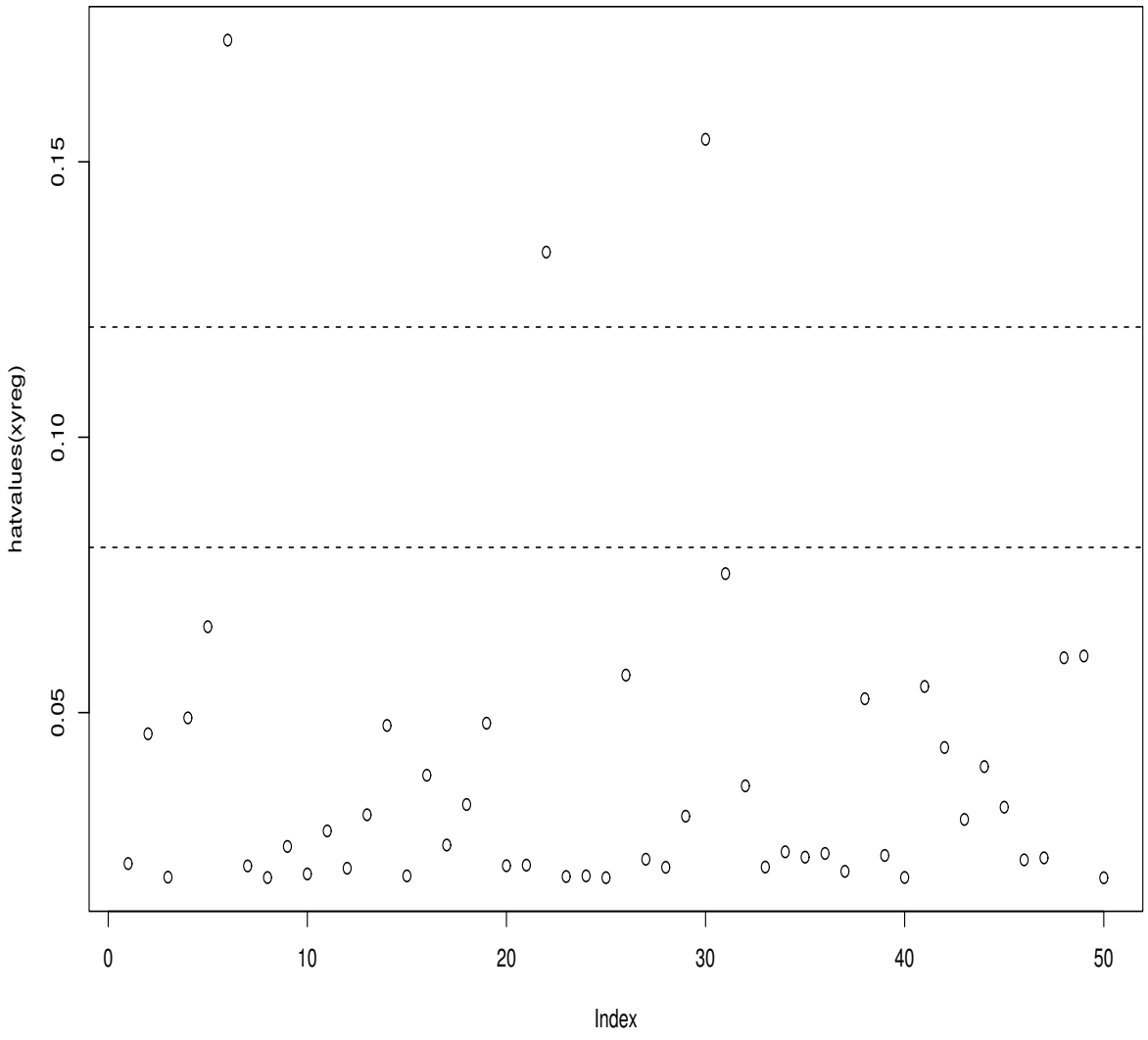
- Note that high leverage observations are not necessarily bad: they contribute a great deal of info about the coefficient estimates (assuming they are not coding mistakes). They might indicate misspecification (different models for different parts of the data).

- Fig. 8.5 contains a plot of hat values. This was produced using the R commands:

```
plot(hatvalues(xyreg))

abline(h=c(2,3)*2/N,lty=2)
```

Figure 8.5: Hat values



8.4.1 Standardized and studentized residuals

- To identify which residuals are significantly large, we first standardize them by dividing by the appropriate standard error (given by the square root of each diagonal element of the variance-covariance matrix):

$$\hat{e}_i = \frac{e_i}{[s^2(1-h_i)]^{1/2}}$$

  This is gives the **standardized residuals**.

- Since each residual has been standardized by its own standard error, we can now compare the standardized residuals against a critical value (e.g., ± 1.96) to see if it is truly "large".

- A related measure is the "studentized residuals." Standardized residuals use s^2 as the estimate for σ^2. Studentized residuals are calculated in exactly the same way, except that for the estimate of σ^2 we use s_i^2 calculated using all the data points except i.

- Let $\hat{\boldsymbol{\beta}}(i) = [\mathbf{X}(i)'\mathbf{X}(i)]^{-1}\mathbf{X}(i)'\mathbf{y}$ be the OLS estimator obtained by omitting the ith observation. The variance of the residual $y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}(i)$ is given by $\sigma(i)^2\{1 + \mathbf{x}_i'[\mathbf{X}(i)'\mathbf{X}(i)]^{-1}\mathbf{x}_i'\}$.

- Then

$$e_i^* = \frac{y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}(i)}{\hat{\sigma}(i)\{1 + \mathbf{x}_i'[\mathbf{X}(i)'\mathbf{X}(i)]^{-1}\mathbf{x}_i'\}^{1/2}}.$$

  follows a t-distribution with $N - K - 1$ df (assuming normal errors). Note that

$$\hat{\sigma}(i)^2 = \frac{[\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(i)]'[\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(i)]}{N - K - 1}$$

- Studentized residuals can be interpreted as the t-statistic on a dummy variable equal to one for the observation in question and zero everywhere else. Such a dummy variable will effectively absorb the observation and so remove its influence in determining the other coefficients in the model.

- Studentized residuals that are greater than 2 in absolute value are regarded as outliers.

- Studentized residuals are generally preferred to standardized residuals for the detection of outliers since they give us an idea of which residuals are large while purging the estimation of the residuals of the effect of that one data point.
- Related measures that combine the index of leverage taken from the diagonal elements of the hat matrix and the studentized residuals include DFITS, Cook's Distance, and Welsch Distance.
- This studentized residual still does not, however, tell us how much the inclusion of a single data point has affected the line. For that, we need a measure of influence.

8.4.2 DFBETAS

- A measure that is often used to tell how much each data point affects the coefficient is the DFBETAS, due to Belsley, Kuh, and Welch (1980).
- DFBETAS focuses on the difference between coefficient values when the ith observation is included and excluded, scaling the difference by the estimated standard error of the coefficient:

$$\text{DFBETAS}_{ki} = \frac{\hat{\beta}_k - \hat{\beta}_k(i)}{\hat{\sigma}(i)a_{kk}^{-1/2}}$$

  where $a_{kk}^{-1/2}$ is the kth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$.
- Belsley et al. suggest that observations with

$$|\text{DFBETA}_{ki}| > \frac{2}{\sqrt{N}}$$

  as deserving special attention (accounts for the fact that single observations have less of an effect as sample size grows). It is also common practice simply to use an absolute value of one, meaning that the observation shifted the estimated coefficient by at least one standard error.

- To access the DFBETAS in R (e.g., for plotting) do something like

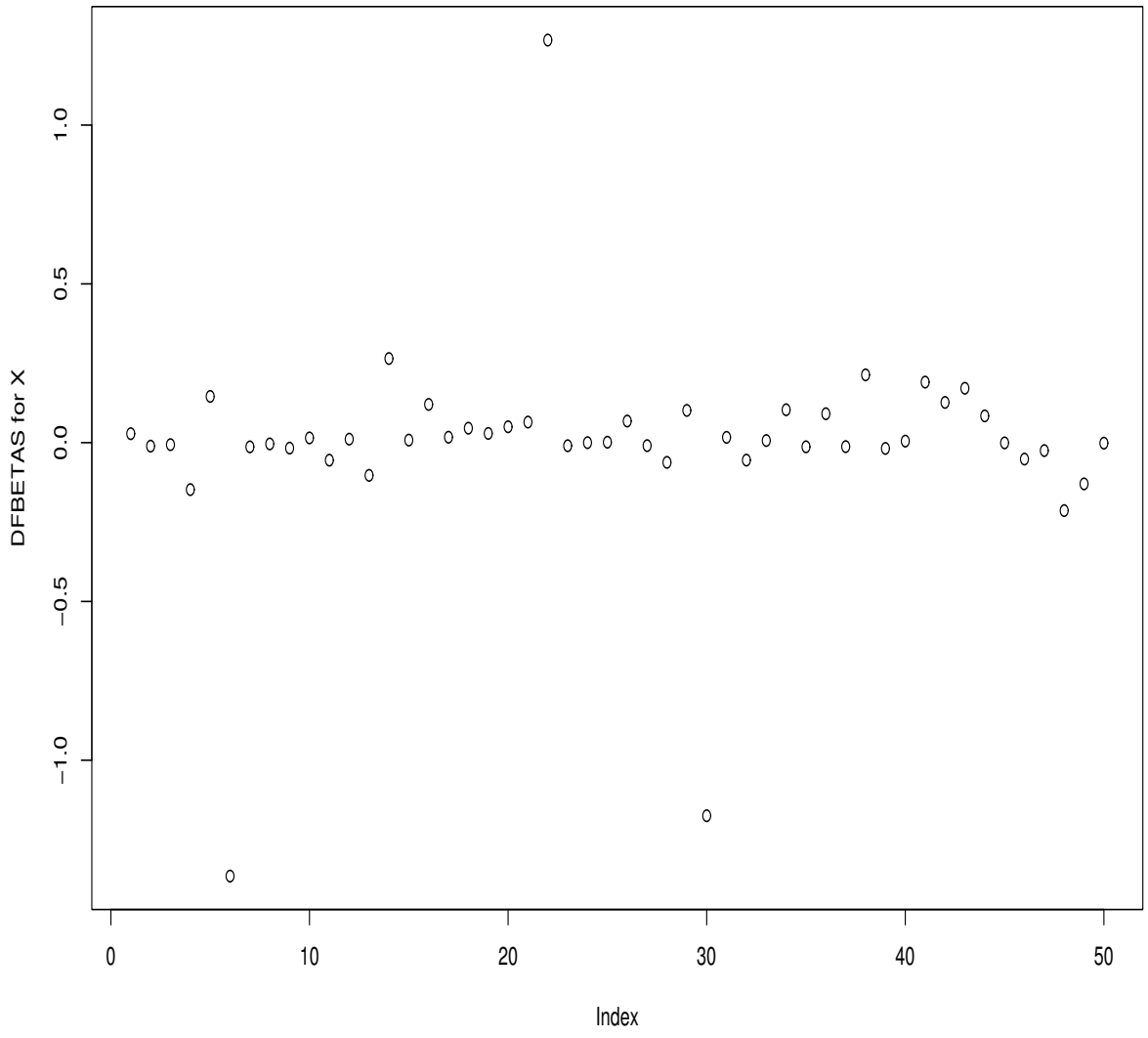
  ```
  dfbs.xyreg <- dfbetas(xyreg)
  ```

  where `xyreg` contains the output from the `lm` command. To plot them for the slope coefficient for our fake data, do

  ```
  plot(dfbs.xyreg[,2],ylab="DFBETAS for X")
  ```

  See Fig. 8.6 for the resulting plot.

Figure 8.6: Plot of DFBETAS



Section 9

Presentation of Results, Prediction, and Forecasting

9.1 Presentation and interpretation of regression coefficients

- Some rules of thumb for presenting results of regression analysis:
  - ➢ Provide a table of descriptive statistics (space permitting).
  - ➢ A table of OLS regression results should include coefficient estimates and standard errors (or t statistics), measures of model fit (adjusted R^2, F-tests, standard error of the regression), information about the sample (size), and estimation approach.
  - ➢ Make your tables readable. Do not simply cut and paste regression output. No one needs to see coefficients/std. errs reported to more than the second or third decimal place. Refrain from using scientific notation (take logs, divide by some factor if necessary). "Big" models make for unattractive tables.
  - ➢ Discuss substantive significance, providing ancillary tables or plots to help with inferences about the magnitude of effects.
- Substantive significance is easier to assess in linear models than in nonlinear models, although walking readers through some simulations is often helpful.
  - ➢ For example, what is the expected change in our dependent variable for a given change in the explanatory variable, where the given change is substantively interesting or intuitive. This could include, for instance, indicating how much the dependent variable changes for a one standard deviation movement up or down in the explanatory variable.
  - ➢ Relate this back to descriptive statistics.

9.1.1 Prediction

- We know what is the expected effect on y of a given change in $\mathbf{x}_k$ from $\hat{\boldsymbol{\beta}}$. But what is the predicted value of y, or $\hat{y}$, for a given vector of values of $\mathbf{x}$?
- **In-Sample Prediction** tells us how the level of y would change for a change in the $\mathbf{X}$s for a particular observation in our sample.
- For example, at a given value of each $\mathbf{X}$ variable, $\mathbf{x}_0$, $\hat{y}$, is equal to:

  $$\hat{y} = \mathbf{x}_0\hat{\beta}$$

  This is the conditional mean of y, conditional on the explanatory variables and the estimated parameters of the model.
- Let us say that we have a growth model estimated on 100 countries including Namibia and Singapore. An example of in-sample prediction would be to predict how Namibia's expected (and predicted) growth rate would change if Namibia had Singapore's level of education.
- When we do in-sample prediction we don't have to be concerned about "fundamental uncertainty" (arising from the error term which alters the value of y) because we are assuming that the true random error attached to Namibia would stay the same and we only change the average level of education.
- We also do not have to be concerned with what is called "estimation uncertainty" arising from the fact that we estimated $\hat{\boldsymbol{\beta}}$ because we are not generalizing out of the sample.
- **Out-of-sample prediction** or **forecasting** gives the prediction for the dependent variable for observations that are not included in the sample. When the prediction refers to the future, it is known as a forecast.
- For given values of the explanatory variables, $\mathbf{x}_0$, the predicted value is still, $\hat{y} = \mathbf{x}_0\hat{\beta}$. To get an idea of how precise this forecast is, we need a measure of the variance of y_0, the actual value of y at $\mathbf{x}_0$, around $\hat{y}$.

  $$y_0 = \mathbf{x}_0\boldsymbol{\beta} + \varepsilon_0$$

- Therefore the forecast error is equal to:

$$y_0 - \hat{y}_0 = e_0 = \mathbf{x}_0\beta + \varepsilon_0 - \mathbf{x}_0\hat{\beta} = \mathbf{x}_0(\beta - \hat{\beta}) + \varepsilon_0$$

- The expected value of this is equal to zero. Why?

- The variance of the forecast error is equal to:

$$\begin{aligned}
\text{var}[e_0] &= E[e_0 e_0'] = E[\mathbf{x}_0(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \varepsilon_0)(\mathbf{x}_0(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \varepsilon_0)'] \\
&= E[(\mathbf{x}_0(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \varepsilon_0)((\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{x}_0' + \varepsilon_0')] \\
&= E[\mathbf{x}_0(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{x}_0 + \varepsilon_0\varepsilon_0' + \varepsilon_0(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{x}_0' + \varepsilon_0'\mathbf{x}_0(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})] \\
&= \mathbf{x}_0 E[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})']\mathbf{x}_0' + E[\varepsilon_0\varepsilon_0'] \\
&= \mathbf{x}_0[\text{var}(\hat{\boldsymbol{\beta}})]\mathbf{x}_0' + \sigma^2 \\
&= \mathbf{x}_0[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{x}_0' + \sigma^2
\end{aligned}$$

  This is the variance of the forecast error, and thus the variance of the actual value of y around the predicted value.

- If the regression contains a constant term, then an equivalent expression (derived using the expression for partitioned matrices is:

$$\text{var}[e^0] = \sigma^2 \left[1 + \frac{1}{n} + \sum_{j=1}^{K-1}\sum_{k=1}^{K-1}(x_j^0 - \bar{x}_j)(x_k^0 - \bar{x}_k)(\mathbf{Z}'\mathbf{M}^0\mathbf{Z})^{jk}\right]$$

  where $\mathbf{Z}$ is the $K-1$ columns of $\mathbf{X}$ that do not include the constant and $(\mathbf{Z}'\mathbf{M}^0\mathbf{Z})^{jk}$ is the jkth element of the inverse of the $(\mathbf{Z}'\mathbf{M}^0\mathbf{Z})$ matrix.

- Implications for the variance of the forecast:

  ➢ will go up with the true variance of the errors, σ^2, and this effect is still present no matter how many data points we have.

  ➢ will go up as the distance of $\mathbf{x}_0$ from the mean of the data rises. Why is this? As $\mathbf{x}_0$ rises, any difference between β and $\hat{\beta}$ is multiplied by a larger number. Meanwhile, the true regression line and the predicted regression line both go through the mean of the data.

- ➢ will fall with the number of data points because the inverse will be smaller and smaller (intuition—as you get more data points, you will estimate $\hat{\beta}$ more precisely).

- The variance above can be estimated using:

$$\text{var}[e_0] = \mathbf{x}_0[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{x}_0' + s^2$$

  The square root of this is the standard error of the forecast.

- A confidence interval can then be formed using: $\hat{y} \pm t_{\alpha/2} se(e_0)$.

- **Clarify** software by King et. al (available for Stata) can be used to help with presenting predictions as well as other kinds of inferences (see "Making the Most of Statistical Analysis: Improving Interpretation and Presentation" by Gary King, Michael Tomz, and Jason Wittenberg, *American Journal of Political Science*, 44:4, April 2000, pp. 341–355; this article and the software described in the article can be found at http://gking.harvard.edu/clarify/docs/clarify.html.

9.2 Encompassing and non-encompassing tests

- All of the hypothesis tests we have examined so far have been what are known as "encompassing tests." In other words, we have always stated the test as a linear restriction on a given model and have tested $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{q}$ versus $H_1 : \mathbf{R}\boldsymbol{\beta} \neq \mathbf{q}$.

- For example, testing Model A below against Model B is a question of running an F-test on the hypothesis that $\beta_3 = \beta_4 = 0$:

  Model A: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i$

  Model B: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$

- This approach can be restrictive. In particular, it does not allow us to test which of two possible sets of regressors is more appropriate. In many cases, we are interested in testing a non-nested hypothesis of the form $H_0 : \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_0$ versus $H_1 : \mathbf{y} = \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\varepsilon}_1$.

- For example, how do we choose between Model C and Model D?

  Model C: $Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \varepsilon_{0i}$

  Model D: $Y_i = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \varepsilon_{1i}$

- The problem in conducting tests of this sort is to transform the two hypotheses above so that we get a "maintained" model (i.e., reflecting the null hypothesis, H_0) that *encompasses* the alternative. Once we have an encompassing model, we can base our test on restrictions to this model.

- Note that we could combine the two models above as Model F:

  Model F: $Y_i = \lambda_0 + \lambda_1 X_{1i} + \lambda_2 X_{2i} + \lambda_3 Z_{1i} + \lambda_4 Z_{2i} + u_i$

- We could then proceed to run an F-test on $\lambda_3 = \lambda_4 = 0$, as a test of C against D, or an F-test of $\lambda_1 = \lambda_2 = 0$, as a test of D against C.

- Problem: the results of the model comparisons may depend on which model we treat as the reference model and the standard errors of all the coefficients (and thus the precision of the F-tests) will be affected by the inclusion of these additional variables. This approach also assumes the same error term for both models.

- A way out of this problem, known as the J-test, is proposed by Davidson and MacKinnon (1981) and uses a non-nested F-test procedure. The test proceeds as follows:

  1. Estimate Model D and obtain the predicted y values, $\hat{Y}_i^D$.
  2. Add these predicted values as an additional regressor to Model C and run the following model. This model is an example of the encompassing principle.
  $$Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 \hat{Y}_i^D + u_i$$
  3. Use a t-test to test the hypothesis that $\alpha_3 = 0$. If this null is not rejected, it means that the predicted values from Model D add no significant information to the model that can be used to predict Y_i, so that Model C is the true model and *encompasses* Model D. Here, "encompasses" means that it encompasses the information that is in D. If the null hypothesis is rejected, then Model C cannot be the true model.

4. Reverse the order of estimations. Estimate Model C first, obtain the predicted values, $\hat{Y}_i^C$, and use these as an additional explanatory variable in Model D. Run a t-test on the null hypothesis that β_3, the coefficient on $\hat{Y}_i^C$, is equal to zero. If you do not reject the null, then Model D is the true model.

- One potential problem with this test is that you can find yourself rejecting both C and D or neither. This is a small sample problem, however.
- Another popular test is the Cox test (see Greene, 5th ed., pp. 155–159, for details).

Section 10

Maximum Likelihood Estimation

10.1 What is a likelihood (and why might I need it)?

- OLS is simply a method for fitting a regression line that, under the Gauss-Markov assumptions, has attractive statistical properties (unbiasedness, efficiency). However, it has no particular statistical justification.
- Maximum likelihood (ML) estimation has a theoretical justification and, if we are prepared to make assumptions about the population model that generates our data, it has "attractive statistical properties" (consistency, efficiency) whether or not the Gauss-Markov assumptions hold.
- In addition, ML estimation is one method for estimating regression models that are not linear.
- Also, if an unbiased minimum variance estimator (something BLUE) exists, then ML estimation chooses it and ML estimation will achieve the Cramer-Rao lower bound (more explanation on this jargon later).
- Where does the notion of a likelihood come from? It's a partial answer to the statistician's search for "inverse probability."
- If we know (or can assume) the under-lying model in the population, we can estimate the probability of seeing any particular sample of data. Thus, we could estimate $P(\text{data}|\text{model})$ where the model consists of parameters and a data generating function.
- This is the notion of conditional probability that underlies hypothesis testing, "How likely is it that we observe the data y, if the null is true?" In most cases, however, statisticians know the data, what they want to find out about the parameters that generate the data.
- Often, statisticians will feel comfortable assigning a particular type of distribution to the DGP (e.g., the errors are distributed normally). This is the same as saying that we are confident about the "functional form."

- What is unknown are the parameters of the DGP. For a single, normally distributed random variable, those unknown parameters would be the mean and the variance. For the dependent variable in a regression model, y, those parameters would include the regression coefficients, $\boldsymbol{\beta}$, and the variance of the errors, σ^2.
- What we want is a measure of "inverse probability," a means to estimate $P(\text{parameters}|\text{data})$ where the parameters are what we don't know about the model. In this inverse probability statement, the data are taken as given and the probability is a measure of absolute uncertainty over various sets of coefficients.
- We cannot derive this measure of absolute uncertainty, but we can get close to it, using the notion of likelihood, which employs Bayes Theorem.
- Bayes Theorem: $P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(A \cap B)}{P(B)}$
- Let's paraphrase this for our case of having data and wanting to find out about coefficients:

$$P(\text{parameters}|\text{data}) = \frac{P(\text{data}|\text{parameters})\ P(\text{parameters})}{P(\text{data})}$$

- For the purposes of estimation, we treat the marginal probability of seeing the data, $P(\text{data})$, as additional information that we use only to scale our beliefs about the parameters. Thus, we can say more simply:

$$P(\text{parameters}|\text{data}) \propto P(\text{data}|\text{parameters})\ P(\text{parameters}) \tag{10.1}$$

  where $\propto$ means "is proportional to".
- The conditional probability of the parameters given the data, or $P(\text{parameters} \mid \text{data})$, is also known as the likelihood function (e.g., $L(\mu, \sigma^2|x)$, where μ refers to the mean and σ^2 refers to the variance).
- The likelihood function may be read as the likelihood of any value of the parameters, μ and σ^2 in the univariate case, given the sample of data that we observe. As our estimates of the parameters of interest, we might want to use the values of the parameters that were most likely given the data at hand.

- We can maximize the right-hand side of Eq. 10.1, $P(\text{data}|\text{parameters})$, with respect to μ and σ^2 to find the values of the parameters that maximize the likelihood of getting the particular data sample.
- We can't calculate the absolute probability of any particular value of the parameters, but we can tell something about the values of the parameters that make the observed sample most likely.
- Given the proportional relationship, the values of the parameters that maximize the likelihood of getting the particular data sample are also the values of the parameters that are most likely to obtain given the sample.
- Thus, we can maximize $P(\text{data}|\text{parameters})$ with respect to the parameters to find the maximum likelihood estimates of the parameters.
- We treat $P(\text{parameters})$ as a prior assumption that we make independent of the sample and often adopt what is called a "flat prior" so that we place no particular weight on any value of the parameters.
- To estimate the parameters of the causal model, we start with the statistical function that is thought to have generated the data (the "data generating function"). We then use this model to derive the expression for the probability of observing the sample, and we maximize this expression with respect to the parameters to obtain the maximum likelihood estimates.

10.2 An example of estimating the mean and the variance

- The general form of a normal distribution with mean μ and variance σ^2 is:

$$f(x|\mu,\sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} e^{-1/2[(x_i-\mu)^2/\sigma^2]}$$

  This function yields a particular value of x from a normal distribution with mean μ and variance σ^2.

- If we had many, independent random variables, all generated by the same underlying probability density function, then the function generating the sample as a whole would be given by:

$$f(x_1,\ldots,x_i,\ldots x_n|\mu,\sigma^2) = \prod_{i=1}^{n}\left(\frac{1}{2\pi\sigma^2}\right)^{1/2} e^{-1/2[(x-\mu)^2/\sigma^2]}$$

  Note that there is an implicit iid assumption, which enables us to write the joint likelihood as the product of the marginals.

- The method of finding the most likely values of μ and σ^2, given the sample data, is to maximize the likelihood function. The expression above is the likelihood function since it is the area under the curve generated by the statistical function that gives you a probability and since this is proportional to $P(\text{parameters}|\text{data})$.

- An easier way to perform the maximization, computationally, is to maximize the log likelihood function, which is just the natural log of the likelihood function. Since the log is a monotonic function, the values that maximize L are the same as the values that maximize $\ln L$.

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

- To find the values of μ and σ^2, we take the derivatives of the log likelihood with respect to the parameters and set them equal to zero.

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0$$

- To solve the likelihood equations, multiply both sides by σ^2 in the first equation and solve for $\hat{\mu}$, the estimate of μ. Next insert this in the second equation and solve for σ^2. The solutions are:

$$\hat{\mu}_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}_n$$

$$\hat{\sigma}^2_{ML} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2$$

- You should recognize these as the sample mean and variance, so that we have some evidence that ML estimation is reliable. Note, however, that the denominator of the calculated variance is n rather than $(n-1) \Rightarrow$ ML estimate of the variance is biased (although it is consistent).

10.3 Are ML and OLS equivalent?

- We would be particularly perturbed if OLS and ML gave us different estimates of the regression coefficients, $\hat{\beta}$.
- Since the error terms under the Gauss-Markov assumptions are assumed to be distributed normal, we can set up the regression model as an ML problem, just as we did in the previous example, and check the estimated parameters, $\hat{\beta}$ and $\hat{\sigma}^2$.
- The regression model is:

  $y_i = \beta_0 + \beta_1 x_i + u_i$, where $u_i \sim$ iid $N(0, \sigma^2)$.

- This implies that the y_i are independently and normally distributed with respective means $\beta_0 + \beta_1 x_i$ and a common variance σ^2. The joint density of the observations is therefore:

$$f(y_1, \ldots, y_i, \ldots y_n | \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{2\pi\sigma^2} \right)^{1/2} e^{-1/2[(y_i - \beta_0 - \beta_1 x_i)^2/\sigma^2]}$$

  and the log likelihood is equal to:

$$\ln L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2} \sum_{i=1}^{n} \left[\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^2} \right]$$

- We will maximize the log likelihood function with respect to β_0 and β_1 and then with respect to σ^2.

- Note that only the last term in the log likelihood function involves β_0 and β_1 and that maximizing this is the same as minimizing the sum of squared errors, since there is a negative sign in front of the whole term.

- Thus, the ML estimators of β_0 and β_1, equal to $\hat{\boldsymbol{\beta}}_{ML}$ are the same as the least squares estimates that we have been dealing with throughout.

- Substituting $\hat{\boldsymbol{\beta}}_{ML}$ into the log likelihood and setting $\hat{Q} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$ which is the SSE, we get the log likelihood function in terms of σ^2 only:

$$\ln L(\sigma) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{\hat{Q}}{2\sigma^2} = \text{ a constant } - \frac{n}{2} \cdot \ln \sigma^2 - \frac{\hat{Q}}{2\sigma^2}$$

- Differentiating this with respect to σ and setting the derivative equal to zero we get:

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\hat{\sigma}} + \frac{\hat{Q}}{\hat{\sigma}^3} = 0$$

  This gives the ML estimator for $\sigma^2 = \hat{\sigma}^2 = \frac{\hat{Q}}{n} = \frac{SSE}{n}$

- This estimator is different from the unbiased estimator that we have been using, $= \hat{\sigma}^2_{OLS} = \frac{\hat{Q}}{n-k} = \frac{SSE}{n-k}$.

- For large N, however, the two estimates will be very close. Thus, we can say that the ML estimate is consistent for σ^2, and we will define the meaning of consistency explicitly in the future.

- We can use these results to show the value of the log likelihood function at its maximum, a quantity that can sometimes be useful for computing test statistics. We will continue to use the short-hand that $\hat{Q} = SSE$ and will also substitute in the equation for the log likelihood, $\ln L$, that $\hat{\sigma}^2_{ML} = \frac{\hat{Q}}{n}$. Substituting this into the equation for the log likelihood above, we get:

$$\begin{aligned}\ln L(\beta, \sigma^2) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\frac{\hat{Q}}{n}\right) - \frac{\hat{Q}}{2(\hat{Q}/n)} \\ &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\hat{Q} + \frac{n}{2}\ln(n) - \frac{n}{2}\end{aligned}$$

- Putting all the terms that do not rely on $\hat{Q}$ together, since they are constant and do not rely on the values of the estimated regression coefficients, $\hat{\boldsymbol{\beta}}_{ML}$, we can say that the maximum value of the log likelihood is:

$$\max \ln L = \text{ a constant } - \frac{n}{2}\ln\hat{Q} = \text{ a constant } - \frac{n}{2}\ln(SSE)$$

- Taking the anti-log, we can say that the maximum value of the likelihood function is:

$$\max L = \text{ a constant } \cdot (SSE)^{-n/2}$$

- This shows that, for a given sample size, n, the maximum value of the likelihood and log likelihood functions will rise as *SSE* falls. This can be handy to know, as it implies that measures of "goodness of fit" can be based on the value of the likelihood or log likelihood function at its maximum.

- **In conclusion**, we can say that ML is equivalent to OLS for the classical linear regression model. The real power of ML, however, is that we can use it in many cases where we do not assume that the errors are normally distributed or that the model is equal to $y_i = \beta_0 + \beta_1 x_i + u_i$, so that the model is far more general.

- For instance, models with binary dependent variables are estimated via ML, using the logistic distribution for the data if a logit model is chosen and a normal if the probit is used. Event count models are estimated via ML, assuming that the original data is distributed Poisson. Thus, ML is the operative method for estimating a number of models that are frequently employed in political science.

10.4 Inference and hypothesis testing with ML

- The tests that one can perform, and the statistics we can compute, on the ML regression coefficients are analogous to the tests and statistics we can form under OLS. The difference is that they are "large sample" tests.

10.4.1 The likelihood ratio test

- The likelihood ratio (LR) test is analogous to an F-test calculated using the residuals from the restricted and un-restricted models. Let $\boldsymbol{\theta}$ be the set of parameters in the model and let $L(\boldsymbol{\theta})$ be the likelihood function.
- Hypotheses such as $\theta_1 = 0$ impose restrictions on the set of parameters $\boldsymbol{\theta}$. What the LR test says is that we first obtain the maximum of $L(\boldsymbol{\theta})$ without any restrictions and we then calculate the likelihood with the restrictions imposed by the hypothesis to be tested.
- We then consider the ratio

$$\lambda = \frac{\max L(\theta) \text{ under the restrictions}}{\max L(\theta) \text{ without the restrictions}}$$

- λ will in general be less than one since the restricted maximum will be less than the unrestricted maximum.
- If the restrictions are not valid, then λ will be significantly less than one. If they are exactly correct, then λ will equal one. The LR test consists of using $-2 \ln \lambda$ as a test statistic $\sim \chi^2_k$, where k is the number of restrictions. If this amount is larger than the critical value of the chi-square distribution, then we reject the null.

- Another frequently used test statistic that is used with ML estimates is the Wald test, which is analogous to the t-statistic:

$$W = \frac{\hat{\boldsymbol{\beta}}_{ML} - \tilde{\boldsymbol{\beta}}}{se(\hat{\boldsymbol{\beta}}_{ML})}$$

  where $\tilde{\boldsymbol{\beta}}$ is the hypothesized value of $\hat{\boldsymbol{\beta}}_{ML}$ and $se(\hat{\boldsymbol{\beta}}_{ML})$ are the standard errors of the $\hat{\boldsymbol{\beta}}_{ML}$, calculated using $\hat{\sigma}^2_{ML}(\mathbf{X}'\mathbf{X})^{-1}$.

- The only difference from a regular t-test is that this is a large sample test. Since in large samples, the t-distribution becomes equivalent to the standard normal, the Wald statistic is distributed standard normal.

10.5 The precision of the ML estimates

- If the likelihood is very curved at the maximum, then the ML estimate of any parameter contains more information about what the true parameters are likely to be.

- If the likelihood is quite flat at the maximum, then the maximum is not telling you much about the likelihood as a whole and the maximum is not a great summary of the likelihood. *Hence, a measure of the likelihood function's curvature is also a measure of the precision of the ML estimate.*

- What is a reasonable measure of the likelihood function's curvature? The second derivative of any function tells us how much the slope (or gradient) of the function changes as we move along the x axis.

- At the maximum, the slope is changing from positive, through zero, to negative. The larger the second derivative, however, the more curved the likelihood function is at the maximum and the more information is given by the ML estimates.

- Information about curvature and precision can therefore be captured by the second derivatives of the likelihood function. Indeed, the **information matrix** of the likelihood function is used for this purpose and is given by:

$$I(\hat{\boldsymbol{\theta}}|y) = -E\left[\frac{\partial^2 \ln L(\tilde{\boldsymbol{\theta}}|y)}{\partial\tilde{\boldsymbol{\theta}}\partial\tilde{\boldsymbol{\theta}}'}\right]_{\tilde{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}}$$

  where $\hat{\boldsymbol{\theta}}$ are the ML estimates (e.g., this would include $\hat{\boldsymbol{\beta}}_{ML}$ and $\hat{\sigma}^2$ for the linear regression model). Thus, the information matrix is the negative of the expectation of the second derivatives of the likelihood function estimated at the point where $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}$.

- If $\boldsymbol{\theta}$ is a single parameter, then the larger is $I(\hat{\boldsymbol{\theta}}|y)$ then the more curved the likelihood (or log likelihood) and the more precise is $\hat{\boldsymbol{\theta}}$. If $\boldsymbol{\theta}$ is a vector of parameters (the normal case) then $I(\hat{\boldsymbol{\theta}}|y)$ is a $K \times K$ matrix with diagonal elements containing the information on each corresponding element of $\hat{\boldsymbol{\theta}}$.

- Because of the expectations operator, the information matrix must be estimated. The most intuitive is:

$$I(\hat{\boldsymbol{\theta}}|y) = -\left[\frac{\partial^2 \ln L(\tilde{\boldsymbol{\theta}}|y)}{\partial\tilde{\boldsymbol{\theta}}\partial\tilde{\boldsymbol{\theta}}'}\right]_{\tilde{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}}$$

  In other words, one just uses the data sample at hand to estimate the second derivatives.

- The information matrix is closely related to the asymptotic variance of the ML estimates, $\hat{\boldsymbol{\theta}}$. It can be shown that the asymptotic variance of the ML estimates, across an infinite number of hypothetically repeated samples is:

$$V(\hat{\boldsymbol{\theta}}) \approx \lim_{n\to\infty}[I(\hat{\boldsymbol{\theta}}|y)/n]^{-1} = \lim_{n\to\infty}\frac{1}{n}E\left[\frac{-\partial^2 \ln L(\tilde{\boldsymbol{\theta}}|y)}{\partial\tilde{\boldsymbol{\theta}}\partial\tilde{\boldsymbol{\theta}}'}\right]^{-1}_{\tilde{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}}$$

- Intuition: the variance is inversely related to the curvature. The greater the curvature, the more that the log likelihood resembles a spike around the ML estimates $\hat{\boldsymbol{\theta}}$ and the lower the variance in those estimates will be.

- In a given sample of data, one can estimate $\mathrm{V}(\hat{\boldsymbol{\theta}})$ on the basis of the estimated information:

$$V(\hat{\boldsymbol{\theta}}) \approx -\frac{1}{n}\left[\frac{-\partial^2 \ln L(\tilde{\boldsymbol{\theta}}|y)}{\partial\tilde{\boldsymbol{\theta}}\partial\tilde{\boldsymbol{\theta}}'}\right]^{-1}_{\tilde{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}}$$

- This expression is also known as the the Cramer-Rao lower bound on the variance of the estimates. The Cramer-Rao lower bound is the lowest value that the asymptotic variance on any estimator can take.

- Thus, the ML estimates have the very attractive property that they are asymptotically efficient. In other words, in large samples, they achieve the lowest possible variance (thus highest efficiency) of any potential estimate of the true parameters.

- For the linear regression case, the estimate of $V(\hat{\boldsymbol{\theta}})$ is $\hat{\sigma}^2_{ML}(\mathbf{X}'\mathbf{X})^{-1}$. This is the same as the variance-covariance matrix of the $\hat{\boldsymbol{\beta}}_{OLS}$ except that we use the ML estimate of σ^2 to estimate the true variance of the error term.

Part II

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

Section 11

Large Sample Results and Asymptotics

11.1 What are large sample results and why do we care about them?

- 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 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 established 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 $(n-k)s^2/\sigma^2 \sim \chi^2_{n-k}$ 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.

11.2 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 $\hat{\boldsymbol{\beta}}$ is *consistent* if it converges in probability to $\boldsymbol{\beta}$, that is if its probability limit as n becomes large is $\boldsymbol{\beta}$ (plim $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$).
- The best way to think about this is that the sampling distribution of $\hat{\boldsymbol{\beta}}$ collapses to a spike around the true $\boldsymbol{\beta}$.
- Two Notions of Convergence
  1. "Convergence in probability":

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

     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 $\lim_{n\to\infty}\mu_n = c$ and $\lim_{n\to\infty}\sigma_n^2 = 0$, then X_n converges in mean square to c.

     Note: convergence in quadratic mean implies convergence in probability ($X_n \xrightarrow{\text{qm}} c \Rightarrow X_n \xrightarrow{\text{P}} c$).
- 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 ($\lim_{n\to\infty} 1/n = 0$).

- Although an estimator may be biased yet consistent, it is very hard (read impossible) for it to be unbiased but inconsistent. If $\hat{\boldsymbol{\beta}}$ is unbiased, there is nowhere for it to collapse to but $\boldsymbol{\beta}$. 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, 6th Ed., 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.

11.3 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 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 $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

11.3.1 The consistency of $\hat{\boldsymbol{\beta}}_{OLS}$

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

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

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

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

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, 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 plim $x_n = c$ and plim $y_n = d$, then plim $(x_n \cdot y_n) = c \cdot d$.
- If $\mathbf{X}_n$ and $\mathbf{Y}_n$ are random matrices with plim $\mathbf{X}_n = \mathbf{A}$ and plim $\mathbf{Y}_n = \mathbf{B}$ then 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

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

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 depedent.

- 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

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

- 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 plim $\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta} + \mathbf{Q}^{-1}$plim $(\mathbf{X}'\boldsymbol{\varepsilon}/n)$ and the next step in the proof is to show that 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:

$$\frac{1}{n}\mathbf{X}'\boldsymbol{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i = \frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_i = \bar{\mathbf{w}} \tag{11.1}$$

  where $\bar{\mathbf{w}}$ is a $k \times 1$ vector of the sample averages of $x_{ki}\boldsymbol{\varepsilon}_i$. (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:

$$E[\bar{\mathbf{w}}] = \frac{1}{n}\sum_{i=1}^{n}E[\mathbf{w}_i] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i E[\varepsilon_i] = \frac{1}{n}\mathbf{X}'E[\boldsymbol{\varepsilon}] = \mathbf{0} \tag{11.2}$$

  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 $n \to \infty$, $\frac{\sigma^2}{n} \to 0$, $\frac{\mathbf{X}'\mathbf{X}}{n} \to \mathbf{Q}$, and thus

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

- Therefore, we can say that $\bar{\mathbf{w}}$ converges in mean square to $0 \Rightarrow$ 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.

11.3.2 The asymptotic normality of OLS

- Let's show that the OLS estimator, $\hat{\beta}_{OLS}$, 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 11.1 and 11.2, 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 indepedent vectors $\mathbf{x}_i\varepsilon_i$ with means $\mathbf{0}$ and variances

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

- 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 $\lim_{n\to\infty}(\mathbf{X}'\mathbf{X}/n) = \mathbf{Q}$, then

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

- 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 $\lim_{n\to\infty}(\mathbf{X}'\mathbf{X}/n) = \mathbf{Q}$, 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, $s^2(\mathbf{X}'\mathbf{X})/n)^{-1}$ is consistent for $\sigma^2\mathbf{Q}^{-1}$. As a result, a consistent estimate for the asymptotic variance of $\hat{\boldsymbol{\beta}}_{OLS}(= \frac{\sigma^2}{n}\mathbf{Q}^{-1} = \frac{\sigma^2}{n}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1})$ is $s^2(\mathbf{X}'\mathbf{X})^{-1}$.

- Thus, we can say that $\hat{\boldsymbol{\beta}}_{OLS}$ 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 $s^2(\mathbf{X}'\mathbf{X})^{-1}$.

11.4 The 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, $\sqrt{s^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}$ for the true standard error in the denominator, equal to $\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}$.

- 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.

11.5 The 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.

11.6 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 $\hat{\boldsymbol{\beta}}_{OLS}$ 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 $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, using a different set of errors each time.
- We then calculate the $\hat{\boldsymbol{\beta}}_{OLS}$ from repeated samples of size n and plot these different sample estimates $\hat{\boldsymbol{\beta}}_{OLS}$ 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 12

Heteroskedasticity

12.1 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:

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

- 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 $\sigma^2\mathbf{I}_n$.

12.1.1 Consequences of non-spherical errors

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

  Proof for Unbiasedness:

  As before:

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

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

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

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

- Therefore, if the errors are normally distributed,

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

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

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

  To show that 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 $\hat{\boldsymbol{\beta}}_{OLS}$ is also consistent for $\boldsymbol{\beta}$.

- Finally, in most cases, $\hat{\boldsymbol{\beta}}_{OLS}$ is asymptotically normally distributed with mean $\boldsymbol{\beta}$ and its variance-covariance matrix is given by $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$.

12.2 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 $\hat{\boldsymbol{\beta}}_{OLS}$ 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 $\hat{\boldsymbol{\beta}}_{OLS}$.
- More importantly, OLS is no longer efficient, since another method called **Generalized Least Squares** will give estimates of the regression coefficients, $\hat{\boldsymbol{\beta}}_{GLS}$, that are unbiased and have a smaller variance.

12.3 Generalized Least Squares

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

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

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

- It can be shown that the characteristic vectors are all orthogonal and for each characteristic vector, $\mathbf{c}_i'\mathbf{c}_i = 1$ (Greene, 6th ed. p. 968–969). 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

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

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

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

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

12.3.1 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 $E[\boldsymbol{\varepsilon}^*\boldsymbol{\varepsilon}^{*'}] = \mathbf{P}\sigma^2\boldsymbol{\Omega}\mathbf{P}' = \sigma^2\mathbf{I}$.
- 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 $\sigma^2(\mathbf{X}^{*'}\mathbf{X}^*)^{-1} = \sigma^2(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}$. 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, $\hat{\boldsymbol{\beta}}_{GLS}$, and that the Cramer-Rao lower bound for the variance of $\hat{\boldsymbol{\beta}}_{GLS}$ is given by $\sigma^2(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}$.

12.4 Feasible Generalized Least Squares

- All of the above assumes that $\boldsymbol{\Omega}$ is a known matrix, which is usually not the case.
- One option is to estimate $\boldsymbol{\Omega}$ in some way and to use $\hat{\boldsymbol{\Omega}}$ in place of $\boldsymbol{\Omega}$ 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 $\hat{\boldsymbol{\beta}}_{OLS}$ is consistent and unbiased for $\boldsymbol{\beta}$, 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 $\hat{\boldsymbol{\Omega}}$ and $\sqrt{\omega_i}$.
- We then transform the data using these estimates, and use OLS again on the transformed data to estimate $\hat{\boldsymbol{\beta}}_{GLS}$ 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.

12.5 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, 6th ed., p. 162–3)

- 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

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

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

12.6 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.

12.6.1 Visual inspection of the residuals

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

12.6.2 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.

12.6.3 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,

$$\frac{SSR}{2} \overset{a}{\sim} \chi^2_{m-1},$$

  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 13

Autocorrelation

13.1 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 $\varepsilon_{t+1}, \varepsilon_{t+2}, \ldots$ and $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_1$.
- 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.

13.2 Causes of autocorrelation

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

13.3 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.

13.4 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.

13.4.1 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**.

13.4.2 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, $e_{t-1}, e_{t-2}, e_{t-3}, \ldots, e_{t-p}$.
- 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}, $e_{t-3}, \ldots, e_{t-p}$ 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.

13.5 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 $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']$ 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 common 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:

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

- Each disturbance, ε_t, embodies the entire past history of the us, 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:

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

- 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:

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

then

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

- Using this, we see that

$$\text{var}[\varepsilon_t] = \frac{\sigma_u^2}{1-\rho^2} = \sigma_\varepsilon^2 = \gamma_0 \tag{13.1}$$

- 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

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

13.6 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:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \quad t = 1, 2, \dots, T \tag{13.2}$$

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 13.2 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_ts are serially independent, have constant variance (σ_u^2), and the covariance between different us 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.

13.7 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 use 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.

13.8 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:

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

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.

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

$$= \beta + \frac{\sum_{t=2}^{T} y_{t-1} \varepsilon_t}{\sum_{t=2}^{T} y_{t-1}^2}$$

- 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 $T \to \infty$ 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_ts are uncorrelated.

- Continuing:

$$\rho\text{cov}[y_t, \varepsilon_t] = \rho\text{cov}[\beta y_{t-1} + \varepsilon_t, \varepsilon_t] = \rho\{\beta\text{cov}[y_{t-1}, \varepsilon_t] + \text{cov}[\varepsilon_t, \varepsilon_t]\} \quad (13.3)$$

$$= \rho\{\beta\text{cov}[y_{t-1}, \varepsilon_t] + \text{var}[\varepsilon_t]\} \quad (13.4)$$

- 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. 13.1):

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

- 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 $T \to \infty$, 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.

13.9 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.

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

- 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 $\text{cov}[x_t, \varepsilon_t] \neq 0$.

13.10 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 $\hat{\boldsymbol{\beta}}$ coefficient, $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$. It is easy to show:

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}] &= E[(z'z)^{-1}(z'y)] = E[(z'z)^{-1}(z'(\beta z + \eta))] && (13.5)\\ &= E[[(x+u)'(x+u)]^{-1}(x+u)'(\beta x + \varepsilon)] && (13.6)\\ &= E\left(\frac{\sum_i (x+u)(\beta x + \varepsilon)}{\sum_i (x+u)^2}\right) = \frac{\beta x^2}{x^2 + \sigma_u^2} = \beta\frac{x^2}{x^2 + \sigma_u^2} && (13.7) \end{aligned}$$

- 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:

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

- This involves having to show that plim $\left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}\right) = 0$.

- 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, 6th ed., pp. 325–327.

13.11 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.

13.12 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 14

Simultaneous Equations Models and 2SLS

14.1 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).

14.1.1 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:

$$G_i = \beta_0 + \beta_1 V_i + \varepsilon_i$$
$$V_i = \alpha_0 + \alpha_1 G_i + \eta_i$$

14.1.2 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:

$$E[G'\boldsymbol{\eta}] = (\beta_0 + \beta_1\alpha_1)E[\boldsymbol{\eta}] + \beta_1\alpha_1 E[G'\boldsymbol{\eta}] + \beta_1 E[\boldsymbol{\eta}^2] + E[\boldsymbol{\varepsilon}]E[\boldsymbol{\eta}]$$

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

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

$$E[G'\boldsymbol{\eta}] = \beta_1\alpha_1 E[G'\boldsymbol{\eta}] + \beta_1\sigma^2_{\boldsymbol{\eta}}$$

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

- 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.

14.2 Reduced form equations

- 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,

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

  Also,

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

- In this case, we can we 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.

14.3 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.

14.3.1 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.

14.4 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 inolves 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:

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

- 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.

14.4.1 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:

$$y_i = \alpha_0 + \alpha_1 X_i + \varepsilon_i$$
$$p_i = \beta_0 + \beta_1 y_i + \beta_2 R_i + \eta_i$$
$$q_i = \delta_0 + \delta_1 y_i + \delta_2 p_i + \delta_3 S_i + \nu_i$$

  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.

14.5 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:

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

- 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:

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

- 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 $1 - \frac{SSE}{SST}$.

14.6 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 13.12, we showed that:

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

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

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

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

- 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.

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

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

- 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.

14.7 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.

14.8 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**.

14.8.1 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).

14.8.2 The Hausman specification test

- If $\hat{\boldsymbol{\beta}}_{OLS}$ and $\hat{\boldsymbol{\beta}}_{IV}$ 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 $\hat{\boldsymbol{\beta}}_{OLS}$ and $\hat{\boldsymbol{\beta}}_{IV}$ are consistent and $\hat{\boldsymbol{\beta}}_{OLS}$ is efficient relative to $\hat{\boldsymbol{\beta}}_{IV}$ (recall that OLS is BLUE).
- H_1: There is endogeneity

  In this case $\hat{\boldsymbol{\beta}}_{IV}$ remains consistent while $\hat{\boldsymbol{\beta}}_{OLS}$ is inconsistent. Thus, their values will diverge.
- The suggestion, then, is to examine $\mathbf{d} = (\hat{\boldsymbol{\beta}}_{IV} - \hat{\boldsymbol{\beta}}_{OLS})$. 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 $\hat{\boldsymbol{\beta}}_{IV}$ and $\hat{\boldsymbol{\beta}}_{OLS}$. Hausman solved this by proving that the covariance between an efficient estimator $\hat{\boldsymbol{\beta}}_{OLS}$ of a parameter vector $\boldsymbol{\beta}$ and its difference from an inefficient estimator $\hat{\boldsymbol{\beta}}_{IV}$ of the same parameter vector is zero (under the null). (For more details see Greene, 6th ed., Section 12.4.)
- Based on this proof, we can say that:

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

- 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:

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

- 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).

14.8.3 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.

14.8.4 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

14.9 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 know 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 $n \cdot R^2 \overset{a}{\sim} \chi^2$ w/ df = # of instruments − # of endogenous regressors. If reject the null ⇒ at least some of the IVs are not exogenous.

Section 15

Time Series Modeling

15.1 Historical background

- Genesis of modern time series models: large structural models of the macroeconomy, 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.

15.2 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.

15.2.1 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, pth-order autoregression or AR(p) process would be written:

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

- 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:

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

- 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:

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

  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:

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

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 $\sum_{i=0}^{\infty} \gamma^i \varepsilon_{t-i} = \varepsilon_t + \gamma\varepsilon_{t-1} + \gamma^2\varepsilon_{t-2} + ... = \varepsilon_t + \gamma L\varepsilon_t + \gamma^2 L^2\varepsilon_t + ...$ 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.

15.3 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 auto-regressive 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 requires that the roots of this more complex expression "lie outside the unit circle."

15.4 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:

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

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, 6th ed., pp. 718–721).

15.5 ARMA processes

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

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

  and then substitute for the lagged y_ts 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, 6th ed., p. 717 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).

15.6 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.

15.7 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.

- 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 $\alpha = 0$, μ, 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 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."

15.7.1 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:

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

15.8 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.

15.9 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 $[y_{t-1}, y_{t-2}, ..., y_{t-s}]$. 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.

15.10 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.

15.11 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.

15.12 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.

15.13 `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 III

Special Topics

Section 16

Time-Series Cross-Section and Panel Data

- Consider the model

$$y_{it} = \beta_0 + \mathbf{x}_{it}\beta + \varepsilon_{it}$$

where

➢ $i = 1,\ldots, N$ cross-sectional units (e.g., countries, states, households)

➢ $t = 1, \ldots, T$ time units (e.g., years, quarters, months)

➢ NT = total number of data points.

- Oftentimes, this model will be estimated pooling the data from different units and time periods.

- Advantages: can give more leverage in estimating parameters; consistent w/ "general" theories

16.1 Unobserved country effects and LSDV

- In the model:

$$Govt\ Spending_{it} = \beta_0 + \beta_1 Openness_{it} + \beta_2 Z_{it} + \varepsilon_{it}$$

It might be argued that the level of government spending as a percentage of GDP differs for reasons that are specific to each country (e.g., solidaristic values in Sweden). This is also known as cross-sectional heterogeneity.

- If these unit-specific factors are correlated with other variables in the model, we will have an instance of omitted variable bias. Even if not, we will get larger standard errors because we are not incorporating sources of cross-country variation into the model.

- We could try to explicitly incorporate all the systematic factors that might lead to different levels of government spending across countries, but places high demands in terms of data gathering.

- Another way to do this, which may not be as demanding data-wise, is to introduce a set of country dummies into the model.

$$Govt\ Spending_{it} = \alpha_i + \beta_1 Openness_{it} + \beta_2 Z_{it} + \varepsilon_{it}$$

  This is equivalent to introducing a country-specific intercept into the model. Either include a dummy for all the countries but one, and keep the intercept term, or estimate the model with a full set of country dummies and no intercept.

16.1.1 Time effects

- There might also be time-specific effects (e.g., government spending went up everywhere in 1973–74 in OECD economies because the first oil shock led to unemployment and increased government unemployment payments). Once again, if the time-specific factors are not accounted for, we could face the problem of bias.
- To account for this, introduce a set of dummies for each time period.

$$Govt\ Spending_{it} = \alpha_i + \delta_t + \beta_1 Openness_{it} + \beta_2 Z_{it} + \varepsilon_{it}$$

- The degrees of freedom for the model are now $NT - k - N - T$. The significance, or not, of the country-specific and time-specific effects can be tested by using and F-test to see if the country (time) dummies are jointly significant.
- The general approach of including unit-specific dummies is known as *Least Squares Dummy Variables* model, or *LSDV*.
- Can also include $(T - 1)$ year dummies for time effects. These give the difference between the predicted causal effect from $\mathbf{x}_{it}\boldsymbol{\beta}$ and what you would expect for that year. There has to be one year that provides the baseline prediction.

16.2 Testing for unit or time effects

- For LSDV (including an intercept), we want to test the hypothesis that

$$\alpha_1 = \alpha_2 = ... = \alpha_{N-1} = 0$$

- Can use an F-test:

$$F(N-1, NT-N-K) = \frac{(R^2_{UR} - R^2_R)/(N-1)}{(1-R^2_{UR})/(NT-N-K)}$$

  In this case, the unrestricted model is the one with the country dummies (and hence different intercepts); the restricted model is the one with just a single intercept. A similar test could also be performed on the year dummies.

16.2.1 How to do this test in `Stata`

- After the `regress` command you type:

  1. If there are (N-1) country dummies and an intercept
     test `dummy1=dummy2=dummy3=dummy4=...=dummyN-1=0`
  2. If there are N country dummies and no intercept
     test `dummy1=dummy2=dummy3=dummy4=...=dummyN`

16.3 LSDV as fixed effects

- Least squares dummy variable estimation is also known as *Fixed Effects*, because it assumes that the variation in the dependent variable, y_{it}, for given countries or years can be estimated as a given, *fixed* effect.

- Before we go into the justification for this, let us examine which part of the variation in y_{it} is used to calculate the remaining $\boldsymbol{\beta}$ coefficients under fixed effects.

- A fixed effects model can be estimated by transforming the data. To do this, calculate the country mean of y_{it} for all the different countries. Let the group mean of a given country, i, be represented as $\bar{y}_{i.}$

- Let the original model be

$$y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta} + \varepsilon_{it} \tag{16.1}$$

Then:

$$\bar{y}_{i.} = \alpha_i + \bar{\mathbf{x}}_{i.}\boldsymbol{\beta} + \bar{\varepsilon}_{i.}$$

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 country varies with the mean level of the independent variables.

Substracting this from eq. 16.1 gives

$$(y_{it} - \bar{y}_{i.}) = (\alpha_i - \alpha_i) + (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})\boldsymbol{\beta} + (\varepsilon_{it} - \bar{\varepsilon}_{i.})$$

or

$$(y_{it} - \bar{y}_{i.}) = (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})\boldsymbol{\beta} + (\varepsilon_{it} - \bar{\varepsilon}_{i.})$$

- If we run OLS on this regression it will produce what is known as the "Fixed Effects" estimator, or $\boldsymbol{\beta}_{FE}$.
- It is indentical to LSDV and is sometimes called the *within-group estimator*, because it uses only the variation in y_{it} and $\mathbf{x}_{it}$ within each group (or country) to estimate the $\boldsymbol{\beta}$ coefficients. Any variation between countries is assumed to spring from the unobserved fixed effects.
- Note that if time-invariant regressors are included in the model, the standard FE estimator will not produce estimates for the effects of these variables. Similar issue w/ LSDV.

  ➢ IV approach to produce estimates, but requires some exgoneity assumptions that may not be met in practice.

- The effects of slow-moving variables can be estimated very imprecisely due to collinearity.

16.4 What types of variation do different estimators use?

- Let us now determine the sum of squares ($\mathbf{X}'\mathbf{X}$) and cross-products ($\mathbf{X}'\mathbf{y}$) for the OLS estimator and within-group estimator in order to clarify which estimator uses what variation to calculate the $\boldsymbol{\beta}$ coefficients.
- Let S_{xx} be the sum of squares and let S_{xy} be the cross-products. Let the overall means of the data be represented as $\bar{\mathbf{y}}$ and $\bar{\mathbf{x}}$.
- Then the *total* sum of squares and cross-products (which define the variation that we use to estimate $\hat{\boldsymbol{\beta}}_{OLS}$) is:

$$S_{xx}^T = \sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it} - \bar{\mathbf{x}})(\mathbf{x}_{it} - \bar{\mathbf{x}})'$$

$$S_{xy}^T = \sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it} - \bar{\mathbf{x}})(y_{it} - \bar{y})$$

- The within-group sum of squares and cross-products (used to estimate $\hat{\boldsymbol{\beta}}_{FE}$) is:

$$S_{xx}^W = \sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})(\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})'$$

$$S_{xy}^W = \sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})(y_{it} - \bar{y}_{i.})$$

- The between-group sum of squares and cross-products (used to estimate $\hat{\boldsymbol{\beta}}_{BE}$) is:

$$\mathbf{S}_{xx}^B = \sum_{i=1}^{N}\sum_{t=1}^{T}(\bar{\mathbf{x}}_{i.} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{i.} - \bar{\mathbf{x}})'$$

$$\mathbf{S}_{xy}^B = \sum_{i=1}^{N}\sum_{t=1}^{T}(\bar{\mathbf{x}}_{i.} - \bar{\mathbf{x}})(\bar{y}_{i.} - \bar{y})$$

- It is easy to verify that:

$$\mathbf{S}_{xx}^{T} = \mathbf{S}_{xx}^{W} + \mathbf{S}_{xx}^{B}$$

and:

$$\mathbf{S}_{xy}^{T} = \mathbf{S}_{xy}^{W} + \mathbf{S}_{xy}^{B}$$

- We also have that:

$$\hat{\boldsymbol{\beta}}_{OLS} = [\mathbf{S}_{xx}^{T}]^{-1}[\mathbf{S}_{xy}^{T}] = [\mathbf{S}_{xx}^{W} + \mathbf{S}_{xx}^{B}]^{-1}[\mathbf{S}_{xy}^{W} + \mathbf{S}_{xy}^{B}]$$

and

$$\hat{\boldsymbol{\beta}}_{FE} = [\mathbf{S}_{xx}^{W}]^{-1}[\mathbf{S}_{xy}^{W}]$$

while,

$$\hat{\boldsymbol{\beta}}_{BE} = [\mathbf{S}_{xx}^{B}]^{-1}[\mathbf{S}_{xy}^{B}]$$

- The standard $\hat{\boldsymbol{\beta}}_{OLS}$ uses all the variation in y_{it} and $\mathbf{x}_{it}$ to calculate the slope coefficients while $\hat{\boldsymbol{\beta}}_{FE}$ just uses the variation across time and $\hat{\boldsymbol{\beta}}_{BE}$ just uses the variation across countries.
- We can show that $\hat{\boldsymbol{\beta}}_{OLS}$ is a weighted average of $\hat{\boldsymbol{\beta}}_{FE}$ and $\hat{\boldsymbol{\beta}}_{BE}$. In fact:

$$\hat{\boldsymbol{\beta}}_{OLS} = \mathbf{F}^{W}\hat{\boldsymbol{\beta}}_{FE} + \mathbf{F}^{B}\hat{\boldsymbol{\beta}}_{BE}$$

where $\mathbf{F}^{W} = [\mathbf{S}_{xx}^{W} + \mathbf{S}_{xx}^{B}]^{-1}\mathbf{S}_{xx}^{W}$ and $\mathbf{F}^{B} = [\mathbf{I} - \mathbf{F}^{W}]$

16.5 Random effects estimation

- Fixed effects is completely appropriate if we believe that the country-specific effects are indeed fixed, estimable amounts that we can calculate for each country.
- 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 βs 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.
- The appropriate model that accounts for cross-national variation is *random effects*:

  $$y_{it} = \alpha + \mathbf{x}_{it}\boldsymbol{\beta} + u_i + \varepsilon_{it}$$

  In this model, α is a general intercept and u_i is a time-invariant, random disturbance characterizing the ith country. Thus, country-effects are treated as country-specific shocks. We also assume in this model that:

  $$E[\varepsilon_{it}] = E[u_i] = 0$$

  $$E[\varepsilon_{it}^2] = \sigma_\varepsilon^2, E[u_i^2] = \sigma_u^2$$

  $$E[\varepsilon_{it}u_j] = 0 \; \forall \; i, t, j; E[\varepsilon_{it}\varepsilon_{js}] = 0 \; \forall \; t \neq s, i \neq j; E[u_i u_j] = 0 \text{ for } i \neq j.$$

- For each country, we have a separate error term, equal to w_{it}, where:

  $$w_{it} = \varepsilon_{it} + u_i$$

  and

  $$E[w_{it}^2] = \sigma_\varepsilon^2 + \sigma_u^2, \text{ and } E[w_{it}w_{is}] = \sigma_u^2 \text{ for } t \neq s.$$

- It is because the RE model decomposes the disturbance term into different components that it is also known as an *error components model*.
- For each panel (or country), the variance-covariance matrix of the T disturbance terms will take the following form:

$$\boldsymbol{\Sigma} = \begin{bmatrix} (\sigma_\varepsilon^2 + \sigma_u^2) & \sigma_u^2 & \sigma_u^2 & \dots & \sigma_u^2 \\ \sigma_u^2 & (\sigma_\varepsilon^2 + \sigma_u^2) & \sigma_u^2 & \dots & \sigma_u^2 \\ \vdots & & & \ddots & \vdots \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \dots & (\sigma_\varepsilon^2 + \sigma_u^2) \end{bmatrix} = \sigma_\varepsilon^2 \mathbf{I}_T + \sigma_u^2 \mathbf{i}_T \mathbf{i}_T'$$

  where $\mathbf{i}_T$ is a $(T \times 1)$ vector of ones.
- The full variance-covariance matrix for all the NT observations is:

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

- The way in which the RE model differs from the original OLS estimation (with no fixed effects) is *only in the specification of the disturbance term.* When the model differs from the standard G-M assumptions only in the specification of the errors, the regression coefficients can be consistently and efficiently estimated by Generalized Least Squares (GLS) or (when we don't exactly know $\boldsymbol{\Omega}$) by FGLS).
- Thus, we can do a transformation of the original data that will create a new var-cov matrix for the disturbances that conforms to G-M.

16.6 FGLS estimation of random effects

- The FGLS estimator is

$$\hat{\boldsymbol{\beta}}_{FGLS} = (\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\hat{\boldsymbol{\Omega}}\mathbf{y})$$

  To estimate this we will need to know $\boldsymbol{\Omega}^{-1} = [\mathbf{I}_N \otimes \boldsymbol{\Sigma}]^{-1}$, which means that we need to estimate $\boldsymbol{\Sigma}^{-1/2}$:

$$\boldsymbol{\Sigma}^{-1/2} = \frac{1}{\sigma_\varepsilon}\left[\mathbf{I} - \frac{\theta}{T}\mathbf{i}_T\mathbf{i}_T^{'}\right]$$

  where

$$\theta = 1 - \frac{\sigma_\varepsilon}{\sqrt{T\sigma_u^2 + \sigma_\varepsilon^2}}$$

- Then the transformation of $\mathbf{y}_i$ and $\mathbf{X}_i$ for FGLS is

$$\boldsymbol{\Sigma}^{-1/2}y_{it} = \frac{1}{\sigma_\varepsilon}\begin{bmatrix} y_{i1} - \theta\bar{y}_{i.} \\ y_{i2} - \theta\bar{y}_{i.} \\ \vdots \\ y_{iT} - \theta\bar{y}_{i.} \end{bmatrix}$$

  with a similar looking expression for the rows of $\mathbf{X}_i$.

- It can be shown that the GLS estimator, $\hat{\boldsymbol{\beta}}_{RE}$, like the OLS estimator, is a weighted average of the within (FE) and between (BE) estimators:

$$\hat{\boldsymbol{\beta}}_{RE} = \hat{F}^W\hat{\boldsymbol{\beta}}_{FE} + (I - \hat{F}^W)\hat{\boldsymbol{\beta}}_{BE}$$

  where:

$$\hat{F}^W = [S_{xx}^W + \lambda S_{xx}^B]^{-1}S_{xx}^W$$

  and

$$\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_u^2} = (1-\theta)^2$$

- If $\lambda = 1$, then the RE model reduces to OLS. There is essentially no country-specific disturbance term, so the regression coefficients are most efficiently estimated using the OLS method.

- If $\lambda = 0$, the country-specific shocks dominate the other parts of the disturbance. Then the RE model reduces to the fixed effects model. We attribute all cross-country variation to the country-specific shock, with none attributed to the random disturbance ε_{it} and just use the cross-time variation to estimate the slope coefficients.
- To the extent that λ differs from one, we can see that the OLS estimation involves an inefficient weighting of the two least squares estimators (within and between) and GLS will produce more efficient results.
- To estimate the RE model when $\mathbf{\Omega}$ is unknown, we use the original OLS results, which are consistent, to get estimates of σ_ε^2 and σ_u^2.

16.7 Testing between fixed and random effects

- If the random, country-specific disturbance term, u_i, is correlated with any of the other explanatory variables, $\mathbf{x}_{it}$, then we will get biased estimates in the OLS stage because the regressors and the disturbance will be contemporaneously correlated.
- The coefficient on $\mathbf{x}_{it}$ will be biased and inconsistent, which means the OLS estimates of the residuals will be biased and the $\hat{\boldsymbol{\beta}}_{RE}$ will be biased. This sets us up for a Hausman test:

  H_0: $E[u_i x_{it}] = 0$; RE appropriate $\Rightarrow$ $\hat{\boldsymbol{\beta}}_{RE}$ is approximately equal to $\hat{\boldsymbol{\beta}}_{FE}$ but is more efficient (has smaller standard errors).

  H_1: $E[u_i x_{it}] \neq 0$; RE is not appropriate $\Rightarrow$ $\hat{\boldsymbol{\beta}}_{RE}$ will be different from $\hat{\boldsymbol{\beta}}_{FE}$ (and inconsistent).

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

$$W = \chi_K^2 = [\hat{\boldsymbol{\beta}}_{FE} - \hat{\boldsymbol{\beta}}_{RE}]' \hat{\mathbf{\Sigma}}^{-1} [\hat{\boldsymbol{\beta}}_{FE} - \hat{\boldsymbol{\beta}}_{RE}]$$

  where

$$\hat{\mathbf{\Sigma}} = \text{var}[\hat{\boldsymbol{\beta}}_{FE}] - \text{var}[\hat{\boldsymbol{\beta}}_{RE}]$$

  If the Hausman test statistic is larger than its appropriate critical value, then we reject RE as the appropriate specification.

- Greene, 6th ed., p. 205–206, also shows how to perform a Breusch-Pagan test for RE based on the residuals from the original OLS regression. This tests for the appropriateness of OLS versus the alternative of RE. It does not test RE against FE.

16.8 How to do this in `Stata`

- **xtreg** depvar [varlist], **re** for RE

  xtreg depvar [varlist], **fe** for FE

  xtreg depvar [varlist], **be** for between effects

  To perform the hausman test, type

  xthausman

  After **xtreg** depvar [varlist], **re**

  To run the Breusch-Pagan test for RE versus OLS, type

  xttest0

  After **xtreg** depvar [varlist], **re**

16.9 Panel regression and the Gauss-Markov assumptions

- OLS is BLUE if the errors are iid, implying that $E[\varepsilon_{it}] = 0, \quad E[\varepsilon_{it}^2] = \sigma_\varepsilon^2$, and $E[\varepsilon_{it}\varepsilon_{js}]$=0 for $t \neq s$ *or* $i \neq j$
- The errors in panel regressions, however, are particularly unlikely to be spherical:
  1. We might expect the errors in time-series, cross-sectional settings to be contemporaneously correlated (e.g., economies are linked).
  2. We might also expect the errors in panel models to show "panel heteroskedasticity" (e.g., the scale of the dependent variable may vary across countries).
  3. The errors may show first-order serial correlation (or autocorrelation) or some other type of temporal dependence.

- Consider the following example to see how this will show up in the variance-covariance matrix of the errors. Assume that $N = 2$ and $T = 2$ and that the data are stacked by country and then by time period:
- The covariance matrix for spherical errors is:

$$\mathbf{\Omega} = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix}$$

where

$$\mathbf{\Omega} = \begin{bmatrix} E(\varepsilon_{11}\varepsilon_{11}) & E(\varepsilon_{11}\varepsilon_{12}) & E(\varepsilon_{11}\varepsilon_{21}) & E(\varepsilon_{11}\varepsilon_{22}) \\ E(\varepsilon_{12}\varepsilon_{11}) & E(\varepsilon_{12}\varepsilon_{12}) & E(\varepsilon_{12}\varepsilon_{21}) & E(\varepsilon_{12}\varepsilon_{22}) \\ E(\varepsilon_{21}\varepsilon_{11}) & E(\varepsilon_{21}\varepsilon_{12}) & E(\varepsilon_{21}\varepsilon_{21}) & E(\varepsilon_{21}\varepsilon_{22}) \\ E(\varepsilon_{22}\varepsilon_{11}) & E(\varepsilon_{22}\varepsilon_{12}) & E(\varepsilon_{22}\varepsilon_{21}) & E(\varepsilon_{22}\varepsilon_{22}) \end{bmatrix}$$

- This is the covariance matrix with contemporaneous correlation:

$$\mathbf{\Omega} = \begin{bmatrix} \sigma^2 & 0 & \sigma_{12} & 0 \\ 0 & \sigma^2 & 0 & \sigma_{12} \\ \sigma_{12} & 0 & \sigma^2 & 0 \\ 0 & \sigma_{12} & 0 & \sigma^2 \end{bmatrix}$$

- This is the covariance matrix with contemporaneous correlation and panel-specific heteroskedasticity:

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

- This is the covariance matrix with all of the above and first-order serial correlation:

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

- The most popular method for addressing these issues is known as panel corrected standard errors (PCSEs) which is due to Beck and Katz (APSR '95).
- In other cases of autocorrelation and/or heteroskedasticity, we have suggested GLS or FGLS. We would transform the data so that the errors become spherical by multiplying $\mathbf{x}$ and $\mathbf{y}$ by $\hat{\mathbf{\Omega}}^{-1/2}$. The FGLS estimates of $\boldsymbol{\beta}$ are now equal to $(\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{y})$.
- FGLS is most often performed by first using the OLS residuals to estimate ρ (to implement the Prais-Winsten method) and then using the residuals from an OLS regression on this data to estimate the contemporaneous correlation. This is known as the Parks method and is done in `Stata` via **xtgls**.
- Beck and Katz argue, however, that unless $T >> N$, then very few periods of data are being used to compute the contemporaneous correlation between each pair of countries.
- In addition, estimates of the panel-specific autocorrelation coefficient, ρ_i, are likely to be biased downwards when they are based on few T observations.
- In other words, FGLS via the Parks method has undesirable *small sample properties* when T is not many magnitudes greater than N.
- To show that FGLS leads to biased estimates of the standard errors in TSCS, Beck and Katz use Monte Carlo Simulation

    1. Simulate 1000 different samples of data with known properties. Each of the 1000 samples will be "small" in terms of T and N.
    2. Compute the β coefficients and the standard errors of those coefficients using FGLS on the 1000 different runs of data.
    3. Compare the variance implied by the calculated standard errors with the actual variance found in the data to see if the calculated standard errors are correct.

- PCSEs are built on the same approach as White-Consistent Standard Errors in the case of heteroskedasticity and Newey-West Standard Errors in the case of autocorrelation. Instead of transforming the data, we use the following equation to compute standard errors that account for non-sphericity:

$$\text{var}[\boldsymbol{\beta}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

  where $\boldsymbol{\Omega}$ denote the variance-covariance matrix of the errors.

- Beck and Katz assert that the kind of non-sphericity in TSCS produces the following $\boldsymbol{\Omega}$:

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

- Let

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

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

$$\hat{\mathbf{\Sigma}}_{ij} = \frac{\sum_{t=1}^{T} e_{it} e_{jt}}{T}, \tag{16.2}$$

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

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

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) :

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

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

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{\Omega}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, \tag{16.3}$$

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.
- Note that this solves the problems of panel heteroskedasticity and contemporaneous correlation—not serial correlation. Serial correlation must be removed before applying this fix.
  - ➢ This can be done using the Prais-Winsten transformation or by including a lagged dependent variable (Beck and Katz are partial to the latter).

Section 17

Models with Discrete Dependent Variables

17.1 Discrete dependent variables

- There are many cases where the observed dependent variable is not continuous, but instead takes on discrete values. For example:
  - ➢ A binary or dichotomous dependent variable: $y_i = 0$ or 1 where these values can stand for any qualitative measure. We will estimate this type of data using logit or probit models.
  - ➢ Polychotomous but *ordered* data: $y_i = 0, 1, 2, 3, \ldots$ where each value is part or a ranking, for example [did not graduate high school, graduated HS, some college, BA, MA etc]. We will estimate this type of data using ordered logit or probit models.
  - ➢ Polychotomous or multinomial and *unordered* data, where there is no particular ranking to the data. We will estimate this type of data using multinomial or conditional logit models.
  - ➢ Count data, $y_i = 0, 1, 2, 3, \ldots$ but where the different values might represent the number of events per period rather than different qualitative categories. We will estimate this type of data using count models such as Poisson or negative-binomial models.
  - ➢ Duration data where we are interested in the time it takes until an event occurs. The dependent variable indicates a discrete time period.
- With continuous data, we are generally interested in estimating the expectation of y_i based on the data. With discrete dependent variables, we are generally interested in the probability of seeing a particular value and how that changes with the explanatory variables.

17.2 The latent choice model

- One of the most common ways of modeling and motivating discrete choices is the latent choice model:

$$y_i^* = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i,$$

where we observe:

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

- Then, the probability that we observe a value of one for the dependent variable is:

$$\Pr[y_i = 1|\mathbf{x}_i] = \Pr[y_i^* \geq 0|\mathbf{x}_i] = \Pr[\mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i \geq 0] = \Pr[\varepsilon_i \geq -\mathbf{x}_i\boldsymbol{\beta}].$$

- Let $F(\cdot)$ be the cumulative distribution function (cdf) of ε, so that $F(\varepsilon) = \Pr[\varepsilon \leq \varepsilon_i]$.

$$\Pr[y_i = 1|\mathbf{x}_i] = 1 - F(-\mathbf{x}_i\boldsymbol{\beta})$$
$$\Pr[y_i = 0|\mathbf{x}_i] = F(-\mathbf{x}_i\boldsymbol{\beta})$$

If $f(\cdot)$, the pdf, is symmetric, then:

$$\Pr[y_i = 1|\mathbf{x}_i] = F(\mathbf{x}_i\boldsymbol{\beta}) \text{ and } \Pr[y_i = 0|\mathbf{x}_i] = 1 - F(\mathbf{x}_i\boldsymbol{\beta})$$

- One way to represent the probability of $y_i = 1$ is as the *Linear Probability Model*:

$$\Pr[y_i = 1|\mathbf{x}_i] = F(\mathbf{x}_i\boldsymbol{\beta}) = \mathbf{x}_i\boldsymbol{\beta}$$

- This is the model that we implicitly estimate if we regress a column of ones and zeroes on a set of explanatory variables using ordinary least squares. The probability of $y_i = 1$ simply rises in proportion to the explanatory variables.

- There are two problems with this way of estimating the probability of $y_i = 1$.
  1. The model may predict values of y_i outside the zero-one range.
  2. The linear probability model is heteroskedastic. To see this note that, if $y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$, then $\varepsilon_i = 1 - \mathbf{x}_i\boldsymbol{\beta}$ when y_i=1 and $\varepsilon_i = -\mathbf{x}_i\boldsymbol{\beta}$ when y_i=0. Thus:

$$\begin{aligned}\text{var}[\varepsilon_i] &= E\left[\Pr[y_i = 1] \cdot (1 - \mathbf{x}_i\boldsymbol{\beta})^2 + \Pr[y_i = 0] \cdot (-\mathbf{x}_i\boldsymbol{\beta})^2 | \mathbf{x}_i\right] \\ &= \mathbf{x}_i\boldsymbol{\beta}(1 - \mathbf{x}_i\boldsymbol{\beta})^2 + (1 - \mathbf{x}_i\boldsymbol{\beta})(-\mathbf{x}_i\boldsymbol{\beta})^2 \\ &= \mathbf{x}_i\boldsymbol{\beta}(1 - \mathbf{x}_i\boldsymbol{\beta})\end{aligned}$$

- Given these two drawbacks, it seems wise to select a non-linear function for the predicted probability that will not exceed the values of one or zero and that will offer a better fit to the data, where the data is composed of zeroes and ones.
- For this reason, we normally assume a cumulative density function for the error term where:

$$\lim_{\mathbf{x}\boldsymbol{\beta} \to +\infty} \Pr(Y = 1) = 1$$

  and

$$\lim_{\mathbf{x}\boldsymbol{\beta} \to -\infty} \Pr(Y = 1) = 0$$

- The two most popular choices for a continuous probability distribution that satisfy these requirements are the normal and the logistic.
- The normal distribution function for the errors gives rise to the **probit** model: $F(\mathbf{x}_i\boldsymbol{\beta}) = \Phi(\mathbf{x}_i\boldsymbol{\beta})$ where Φ is the cumulative normal distribution.
- The logistic function gives rise to the **logit** model: $F(\mathbf{x}_i\boldsymbol{\beta}) = \frac{e^{\mathbf{x}\boldsymbol{\beta}}}{1+e^{\mathbf{x}\boldsymbol{\beta}}} = \Lambda(\mathbf{x}_i\boldsymbol{\beta})$.
- The vector of coefficients, $\boldsymbol{\beta}$, is estimated via maximum likelihood.

- Treating each observation, y_i, as an independent random draw from a given distribution, we can write out the likelihood of seeing the entire sample as:

$$\begin{aligned}\Pr(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n) &= \Pr(Y_1 = y_1)\Pr(Y_2 = y_2)\cdots\Pr(Y_n = y_n) \\ &= \prod_{y_i=0}[1 - F(\mathbf{x}_i\boldsymbol{\beta})]\prod_{y_i=1}F(\mathbf{x}_i\boldsymbol{\beta}).\end{aligned}$$

- This can be more conveniently written as:

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

- We take the natural log of this to get the log likelihood:

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

- Maximizing the log likelihood with respect to $\boldsymbol{\beta}$ yields the values of the parameters that maximize the likelihood of observing the sample. From what we learned of maximum likelihood earlier in the course, we can also say that these are the most probable values of the coefficients for the sample.

- The first order conditions for the maximization require that:

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}\left[\frac{y_i f_i}{F_i} + (1 - y_i)\frac{-f_i}{(1 - F_i)}\right]\mathbf{x}_i = 0$$

  where f_i is the probability distribution function, $dF_i/d(\mathbf{x}_i\boldsymbol{\beta})$

- The *standard errors* are estimated by directly computing the *information matrix* of the log likelihood using the data at hand. The information matrix is given by the inverse of the negative "Hessian". The Hessian is the matrix containing all the second derivatives of the function with respect to the parameters, or

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}.$$

- Maximum likelihood estimators have attractive large sample properties (consistency, asymptotic efficiency, asymptotic normality). We therefore treat our estimate of σ^2 as consistent for the real parameter and converging inexorably toward it. Thus, we use the z distribution (which assumes that we know σ^2) as the source of our critical values.

17.3 Interpreting coefficients and computing marginal effects

- A positive sign on a slope coefficient obtained from a logit or probit indicates that the probability of $y_i = 1$ increases as the explanatory variable increases.
- We would like to know how much the probability increases as x_i increases. In other words, we would like to know the *marginal effect*. In OLS, this was given by $\boldsymbol{\beta}$
- Recall that for OLS:
$$\frac{\partial E(y|\mathbf{x})}{\partial \mathbf{x}} = \boldsymbol{\beta}$$
This is not true for logit and probit.
- In binary models $\frac{\partial E(y|\mathbf{x})}{\partial \mathbf{x}} = f(\mathbf{x}_i\boldsymbol{\beta})\boldsymbol{\beta}$
- Thus, for probit models:
$$\frac{\partial E(y|\mathbf{x})}{\partial \mathbf{x}} = \phi(\mathbf{x}_i\boldsymbol{\beta})\boldsymbol{\beta}$$
for logit models:
$$\frac{\partial E(y|\mathbf{x})}{\partial \mathbf{x}} = \frac{e_i^{\mathbf{x}}\boldsymbol{\beta}}{(1+e^{\mathbf{x}_i\boldsymbol{\beta}})^2}\boldsymbol{\beta} = \Lambda(\mathbf{x}_i\boldsymbol{\beta})[1-\Lambda(\mathbf{x}_i\boldsymbol{\beta})]\boldsymbol{\beta}$$
- Two points to note about the marginal effects:
  1. They are not constant.
  2. They are maximized at $\mathbf{x}_i\boldsymbol{\beta} = 0$, which is where $E[y|\mathbf{x}] = 0.5$.

17.4 Measures of goodness of fit

- A frequently used measure of fit for discrete choice models is the Pseudo-R^2 which, as its name implies, is an attempt to tell you how well the model fits the data.

$$PseudoR^2 = 1 - \frac{\ln \hat{L}}{\ln \hat{L}_0}$$

  where $\ln \hat{L}$ is the log likelihood from the full model and $\ln \hat{L}_0$ is the log likelihood from the model estimated with a constant only. This will always be between 0 and 1.

- The log likelihood is always negative. The likelihood function that we start with is just the probability of seeing the entire sample:

$$\Pr[Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n] = \prod_{y_i=0} [1 - F(\mathbf{x}_i\boldsymbol{\beta})] \prod_{y_i=1} F(\mathbf{x}_i\boldsymbol{\beta}).$$

- As a probability, it must lie between zero and one. When we take the natural log of this, to get the log likelihood,

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

  we will get a number between negative infinity and zero. The natural log of one is the value to which e must be raised to get one, and $e^0 = 1$. The natural log of zero is negative infinity, because $e^{-\infty} = 0$. Thus, increases in the log likelihood toward zero imply that the probability of seeing the entire sample given the estimated coefficients is higher.

- The maximum likelihood method selects that value of $\boldsymbol{\beta}$ that maximizes the log likelihood and reports the value of the log likelihood at $\hat{\boldsymbol{\beta}}_{ML}$. A larger negative number implies that the model does not fit the data well.

- As you add variables to the model, you expect the log likelihood to get closer to zero (i.e., increase). $\ln \hat{L}_0$ is the log likelihood for the model with just a constant; $\ln \hat{L}$ is the log likelihood for the full model,so $\ln \hat{L}_0$ should always be larger and more negative than $\ln \hat{L}$.

- Where $\ln \hat{L}_0 = \ln \hat{L}$, the additional variables have added no predictive power, and the pseudo- R^2 is equal to zero. When $\ln \hat{L} = 0$, the model now perfectly predicts the data, and the pseudo- R^2 is equal to one.

- A similar measure is the likelihood ratio test estimating the probability that the coefficients on all the explanatory variables in the model (except the constant) are zero (similar to the F-test under OLS).

- Although these measures offer the benefit of comparison to OLS techniques, they do not easily get at what we want to do, which is to predict accurately when y_i is going to fall into a particular category. An alternative measure of the goodness of fit of the model might be the percentage of observations that were correctly predicted. This is calculated as:

  $$[(Observation = 1 \cap Prediction = 1) + (Observation = 0 \cap Prediction = 0)] / n$$

  The method of model prediction is to say that we predict a value of one for y_i if $F(\mathbf{x}_i\boldsymbol{\beta}) \geq 0.5$.

  Let $\hat{y}_i$ be the predicted value of y_i.

- The percent correctly predicted is:

$$\frac{1}{N}\sum_{i=1}^{N}(y_i\hat{y}_i + (1-y_i)(1-\hat{y}_i))$$

- This measure has some shortcomings itself, which are overcome by the *expected percent correctly predicted* (ePCP):

$$\text{ePCP} = \frac{1}{N}\left(\sum_{y_i=1}^{N}\hat{p}_i + \sum_{y_i=0}^{N}(1-\hat{p}_i)\right)$$

Section 18

Discrete Choice Models for Multiple Categories

18.1 Ordered probit and logit

- Suppose we are interested in a model where the dependent variable takes on more than two discrete values, but the values can be ordered. The latent model framework can be used for this type of model.

$$y_i^* = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i.$$

$$y_i = \begin{cases} 1 & \text{if } y_i^* < \gamma_1 \\ 2 & \text{if } \gamma_1 < y_i^* \leq \gamma_2 \\ 3 & \text{if } \gamma_2 < y_i^* \leq \gamma_3 \\ \vdots & \\ m & \text{if } \gamma_{m-1} < y_i^* \end{cases}$$

- This gives the probabilities

$$\begin{aligned} \Pr(y_i = 1) &= \Phi(\gamma_1 - \mathbf{x}_i\boldsymbol{\beta}) \\ \Pr(y_i = 2) &= \Phi(\gamma_2 - \mathbf{x}_i\boldsymbol{\beta}) - \Phi(\gamma_1 - \mathbf{x}_i\boldsymbol{\beta}) \\ \Pr(y_i = 3) &= \Phi(\gamma_3 - \mathbf{x}_i\boldsymbol{\beta}) - \Phi(\gamma_2 - \mathbf{x}_i\boldsymbol{\beta}) \\ &\vdots \\ \Pr(y_i = m) &= 1 - \Phi(\gamma_{m-1} - \mathbf{x}_i\boldsymbol{\beta}) \\ &= \Phi(\mathbf{x}_i\boldsymbol{\beta} - \gamma_{m-1}) \end{aligned}$$

- To write down the likelihood let

$$z_{ij} = \begin{cases} 1 & \text{if } y_i = j \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } j = 1, \ldots, m.$$

- Then
$$\Pr(z_{ij} = 1) = \Phi(\gamma_j - \mathbf{x}_i\boldsymbol{\beta}) - \Phi(\gamma_{j-1} - \mathbf{x}_i\boldsymbol{\beta})$$

- The likelihood for an individual is
$$\begin{aligned} L_i &= \left[\Phi(\gamma_1 - \mathbf{x}_i\boldsymbol{\beta})\right]^{z_{i1}} \left[\Phi(\gamma_2 - \mathbf{x}_i\boldsymbol{\beta}) - \Phi(\gamma_1 - \mathbf{x}_i\boldsymbol{\beta})\right]^{z_{i2}} \\ &\quad \cdots \left[1 - \Phi(\gamma_{m-1} - \mathbf{x}_i\boldsymbol{\beta})\right]^{z_{im}} \\ &= \Pi_{j=1}^{m} \left[\Phi(\gamma_j - \mathbf{x}_i\boldsymbol{\beta}) - \Phi(\gamma_{j-1} - \mathbf{x}_i\boldsymbol{\beta})\right]^{z_{ij}} \end{aligned}$$

- The likelihood function for the sample then is
$$L = \Pi_{i=1}^{n}\Pi_{j=1}^{m} \left[\Phi(\gamma_j - \mathbf{x}_i\boldsymbol{\beta}) - \Phi(\gamma_{j-1} - \mathbf{x}_i\boldsymbol{\beta})\right]^{z_{ij}}$$

- We compute marginal effects just like for the dichotomous probit model. For the three categoy case these would be:
$$\begin{aligned} \frac{\partial \Pr(y_i = 1)}{\partial \mathbf{x}_i} &= -\phi(\gamma_1 - \mathbf{x}_i\boldsymbol{\beta})\boldsymbol{\beta} \\ \frac{\partial \Pr(y_i = 2)}{\partial \mathbf{x}_i} &= (\phi(\gamma_1 - \mathbf{x}_i\boldsymbol{\beta}) - \phi(\gamma_2 - \mathbf{x}_i\boldsymbol{\beta}))\boldsymbol{\beta} \\ \frac{\partial \Pr(y_i = 3)}{\partial \mathbf{x}_i} &= \phi(\gamma_2 - \mathbf{x}_i\boldsymbol{\beta})\boldsymbol{\beta} \end{aligned}$$

18.2 Multinomial Logit

- Suppose we are interested in modeling vote choices in multi-party systems. In this setting, the dependent variable (party choice) might be $y \in [1, 2, 3]$ but there is no implicit ordering, they are just different choices. Hence, this model is different from *ordered* logit or probit, which we will discuss later.

- The multinomial logit or probit is motivated through a re-working of the latent variable model. The latent variable that we don't see now becomes the "utility" to the voter of choosing a particular party. Because the utility also depends on an individual shock, the model is called the random utility model.

- In the current example, we have three different choices for voter i, where those choices are sub-scripted by j.

$$U_{ij} = \mathbf{x}_i \boldsymbol{\beta}_j + \varepsilon_{ij}$$

  In this case $\mathbf{x}_i$ are the attributes of the voters selecting a party and $\boldsymbol{\beta}_j$ are choice-specific coefficients determining how utility from a party varies with voter attributes (e.g., union members are presumed to derive more utility from the Democratic party). U_{ij} is the utility to voter i of party j.

- The voter will chose party one over party two if:

$$U_{i1} > U_{i2} \Rightarrow \mathbf{x}_i \boldsymbol{\beta}_1 + \varepsilon_{i1} > \mathbf{x}_i \boldsymbol{\beta}_2 + \varepsilon_{i2} \Rightarrow (\varepsilon_{i1} - \varepsilon_{i2}) > \mathbf{x}_i (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)$$

- The likelihood of this being true can be modeled using the logistic function if we assume that the original errors are distributed log-Weibull. This produces a logistic distribution for $(\varepsilon_{i1} - \varepsilon_{i2})$ because if two random variables, ε_{i1} and ε_{i2}, are distributed log-Weibull, then the difference between the two random variables is distributed according to the logistic distribution.

- When there are only two choices (party one or party two) the probability of a voter choosing party 1 is given by:

$$\Pr(y_i = 1|\mathbf{x}_i) = \frac{e^{\mathbf{x}_i(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)}}{1 + e^{\mathbf{x}_i(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)}}$$

  Thus, the random utility model in the binary case gives you something that looks very like the regular logit.

- We notice from this formulation that the two coefficients, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ cannot be estimated separately. One of the coefficients, for instance $\boldsymbol{\beta}_1$, serves as a base and the estimated coefficients $(\boldsymbol{\beta}_j - \boldsymbol{\beta}_1)$ tell us the relative utility of choice j compared to choice 1. We normalize $\boldsymbol{\beta}_1$ to zero, so that $U_{i1} = 0$ and we measure the probability that parties 2 and 3 are selected over party 1 as voter characteristics vary.

- Thus, the probability of selecting party j compared to party 1 depends on the relative utility of that choice compared to one. Recalling that $e^{\mathbf{x}_i\boldsymbol{\beta}_1} = e^0 = 1$, and substituting into the equation above, we find that the probability of selecting party 2 over party 1 equals:

$$\frac{\Pr(y_i = 2)}{\Pr(y_i = 1)} = e^{\mathbf{x}_i\boldsymbol{\beta}_2}$$

- In just the same way, if there is a third party option, we can say that the relative probability of voting for that party over party 1 is equal to:

$$\frac{\Pr(y_i = 3)}{\Pr(y_i = 1)} = e^{\mathbf{x}_i\boldsymbol{\beta}_3}$$

  Finally, we could use a ratio of the two expressions above to give us the relative probability of picking Party 2 over Party 3.

- What we may want, however, is the probability that you pick Party 2 (or any other party). In other words, we want $\Pr(y_i = 2)$ and $\Pr(y_i = 3)$. To get this, we use a little simple algebra. First:

$$\Pr(y_i = 2) = e^{\mathbf{x}_i\boldsymbol{\beta}_2} \cdot \Pr(y_i = 1)$$

$$\Pr(y_i = 3) = e^{\mathbf{x}_i\boldsymbol{\beta}_3} \cdot \Pr(y_i = 1)$$

- Next, we use the fact that the three probabilities must sum to one (the three choices are mutually exclusive and collectively exhaustive):

$$\Pr(y_i = 1) + \Pr(y_i = 2) + \Pr(y_i = 3) = 1$$

So that:

$$\Pr(y_i = 1) + e^{\mathbf{x}_i\boldsymbol{\beta}_2} \Pr(y_i = 1) + e^{\mathbf{x}_i\boldsymbol{\beta}_3} \Pr(y_i = 1) = 1$$

And:

$$\Pr(y_i = 1) \left[1 + e^{\mathbf{x}_i\boldsymbol{\beta}_2} + e^{\mathbf{x}_i\boldsymbol{\beta}_3}\right] = 1$$

so that

$$\Pr(y_i = 1) = \frac{1}{(1 + e^{\mathbf{x}_i\boldsymbol{\beta}_2} + e^{\mathbf{x}_i\boldsymbol{\beta}_3})}$$

- Using this expression for the probability of the first choice, we can say that:

$$\Pr(y_i = 2) = \frac{e^{\mathbf{x}_i\boldsymbol{\beta}_2}}{1 + e^{\mathbf{x}_i\boldsymbol{\beta}_2} + e^{\mathbf{x}_i\boldsymbol{\beta}_3}}$$

$$\Pr(y_i = 3) = \frac{e^{\mathbf{x}_i\boldsymbol{\beta}_3}}{1 + e^{\mathbf{x}_i\boldsymbol{\beta}_2} + e^{\mathbf{x}_i\boldsymbol{\beta}_3}}$$

$$\Pr(y_i = 1) = \frac{1}{1 + e^{\mathbf{x}_i\boldsymbol{\beta}_2} + e^{\mathbf{x}_i\boldsymbol{\beta}_3}}$$

- The likelihood function is then estimated and computed using these probabilities for each of the three choices. The model is estimated in `Stata` using **mlogit**.

- When the choice depends on attributes of the alternatives instead of the attributes of the individuals, the model is estimated as a *Conditional Logit*. This model is estimated in `Stata` using **clogit**.

18.2.1 Interpreting coefficients, assessing goodness of fit

- We can enter the coefficients into the expressions above to get the probability of a vote for any particular party. Further, we can easily tell something about the relative probabilities of picking one party over another and how that changes with a one-unit change in an x variable.
- Recall: $\frac{\Pr(y_i=2)}{\Pr(y_i=1)} = e^{\mathbf{x}_i\boldsymbol{\beta}_2}$
- Thus, $e^{\boldsymbol{\beta}_2}$ will tell you the change in the relative probabilities of a one-unit change in x. This is also known as the relative risk or the relative risk ratios.

18.2.2 The IIA assumption

- The drawback to multinomial logit is the *Independence of Irrelevant Alternatives* (IIA) assumption. In this model, the odds ratios are independent of the other alternatives.
- That is to say, the ratio of the likelihood of voting Democrat to voting Republican is independent of whether option Nader is an option and does not change if we add other options to the model. This is due to the fact that we treat the errors in the multinomial logit model as independent.
- Thus, the error term, ε_{i1}, which makes someone more likely to vote for the Democrats than the Republicans, is treated as being wholly independent of the error term that might make them more likely to vote for the Greens.
- The assumption aids estimation, but is not always warranted in behavioral situations where the types of alternatives available might have an affect of the relative utility of other choices. In other words, if you could vote for the Greens, you might be less likely to vote for the Democrats compared to Republicans.
- One test of whether the assumption is inappropriate in your model setting is to run the model with and without a particular category and use a Hausman test to see whether the coefficients for the explanatory variables in the remaining categories change.

- The IIA problem afflicts both the multinomial logit and the conditional logit models. An option is to use the multinomial probit model, although it comes with its own set of problems.

Section 19

Count Data, Models for Limited Dependent Variables, and Duration Models

19.1 Event count models and poisson estimation

- Suppose that we have data on the number of coups, or the number of instances of state-to-state conflict per year over a 20 year stretch. This data is clearly not continuous, and therefore should not be modeled using OLS, because we never see 2.5 conflicts.
- In addition, OLS could sometimes predict that we should see negative numbers of state-to-state conflict, something we want to rule out given that our data are always zero or positive.
- It should not, however, be modeled as an ordered probit or logit because the numbers represent more than different categories. They are count data, and changes in the count have a cardinal as well as an ordinal interpretation. That is, we can interpret a change of one in the dependent variable as meaning something about the magnitude of conflict, rather than just identifying a change in category.
- Enter the Poisson distribution. This is an important discrete probability distribution for a count for a count starting at zero with no upper bound.
- For data that we believe could be modeled as a Poisson distribution, each observation is a unit of time (one year in our data example) and the Poisson gives the probability of observing a particular count for that period.

- For each observation, the probability of seeing a particular count is given by:

$$\Pr(Y_i = y_i) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, ...$$

- In this model, λ is just a parameter that can be estimated, as μ is for the normal distribution. It can also be shown that the expected event count in each period, or $E(y_i)$ is equal to λ and that the variance of the count is also equal to λ. Thus, the variance of the event counts increases as their number increases.

- The distribution above implies that the expected event count is just the same in each period. This is the univariate analysis of state-to-state conflict. We are saying what the expected number of events is for each year.

- In multivariate analysis, however, we can do better. We can relate the number of events in each period to explanatory variables. This requires relating λ somehow to explanatory variables that we think should be associated with instances of state-to-state conflict. Then we will get a different parameter λ_i for each observation, related to the level of the explanatory variables.

- The standard way to do this for the Poisson model is to say that:

$$\lambda_i = e^{x_i\boldsymbol{\beta}} \text{ or equivalently } \ln \lambda_i = \mathbf{x}_i\boldsymbol{\beta}$$

- We can now say that the probability of observing a given count for period i is:

$$\Pr(Y_i = y_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, ...$$

  and $E[y_i|\mathbf{x}_i] = \lambda_i = e^{x_i\boldsymbol{\beta}}$

- Given this probability for a particular value at each observation, the likelihood of observing the entire sample is the following:

$$\Pr[Y|\lambda_i] = \prod_{i=1}^{n} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

- The log likelihood of this is:

$$\ln L = \sum_{i=1}^{n} [-\lambda_i + y_i \mathbf{x}_i \boldsymbol{\beta} - \ln y_i!]$$

  Since the last term does not involve $\boldsymbol{\beta}$, it can be dropped and the log likelihood can be maximized with respect to $\boldsymbol{\beta}$ to give the maximum likelihood estimates of the coefficients.

- This model can be done in `Stata` using the **poisson** command.

- For inference about the effects of explanatory variables, we can

  ➢ examine the predicted number of events based on a given set of values for the variables, which is given by $\lambda_i = e^{x_i \boldsymbol{\beta}}$.

  ➢ examine the factor change: for a unit change in x_k, the expected count changes by a factor of e^{β_k}, holding other variables constant.

  ➢ examine the percentage change in the expected count for a δ unit change in x_k: $100 \times [e^{(\beta_k \cdot \delta)} - 1]$, holding other variables constant.

- Marginal effects can be computed in `Stata` after **poisson** using **mfx compute**.

- The Poisson model offers no natural counterpart to the R^2 in a linear regression because the conditional mean (the expected number of counts in each period) is non-linear. *Many* alternatives have been suggested. Greene p. 908–909 offers a variety and Cameron and Trivedi (1998) give more. A popular statistic is the standard $Pseudo - R^2$, which compares the log-likelihood for the full model to the log-likelihood for the model containing only a constant.

- One feature of the Poisson distribution is that $E[y_i] = \text{var}[y_i]$. If $E[y_i] < \text{var}[y_i]$, this is called overdispersion and the Poisson is inappropriate. A negative binomial or generalized event count model can be used in this case.

19.2 Limited dependent variables: the truncation example

- The technical definition of a limited dependent variable is that it is limited because some part of the data is simply not observed and missing (truncation) or not observed and set at some limit (censoring).
- Although it is not often encountered, we start by discussing the truncation case because it establishes the intuition.
- Imagine that we have a random variable, x, with a normal distribution. However, we do not observe the distribution of x below some value a, because the distribution is truncated.
- If we wanted to estimate the expected value of x using the values we do observe, we would over-estimate the expected value, because we are not taking into account the observations that are truncated.
- How do we tell what the mean of x would have been had we seen all the observations?
- To establish the expected value, we need the probability density function (pdf) of x given the truncation:

$$f(x|x > a) = \frac{f(x)}{\Pr[x > a]}$$

  The denominator simply re-scales the distribution by the probability that we observe x, so that we will get a *pdf* that would integrate to one.

- In addition:

$$\Pr[x > a] = 1 - \Phi\left(\frac{a - \mu}{\sigma}\right) = 1 - \Phi(\alpha)$$

  where Φ is the cumulative standard normal distribution and the term in brackets simply transforms the normal distribution of x into a standard normal.

- Thus: $f(x|x > a) = \frac{f(x)}{1-\Phi(\alpha)}$

- To get the expected value of x, we use:

$$E[x|x > a] = \int_{a}^{\infty} x.f(x|x > a)dx = \mu + \sigma\lambda(\alpha)$$

  where $\alpha = \left(\frac{a-\mu}{\sigma}\right)$, $\phi(\alpha)$ is the standard normal density and:

  - ➢ $\lambda(\alpha) = \phi(\alpha)/[1 - \Phi(\alpha)]$ when truncation is at $x > a$
  - ➢ $\lambda(\alpha) = -\phi(\alpha)/\Phi(\alpha)$ when truncation is at $x < a$

- The function $\lambda(\alpha)$ is also known as the "inverse Mills ratio" and is relevant for succeeding models.

- Now, imagine that we have a dependent variable, y_i, which is truncated. For example, we only see values of $y_i > a$, or where $\mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i > a$, implying that $\varepsilon_i > a - \mathbf{x}_i\boldsymbol{\beta}$. In this case, the probability that we observe y_i is equal to $\Pr[\varepsilon_i > a - \mathbf{x}_i\boldsymbol{\beta}] = 1 - \Phi\left(\frac{a-\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)$.

- Given our assumptions about the distribution of ε_i, we can calculate this probability, meaning that we can correct the conditional mean we estimate.

- In most cases, we assume that the conditional mean of y_i, or $E[y_i|\mathbf{x}_i] = \mathbf{x}_i\boldsymbol{\beta}$. This is not the case for the truncated dependent variable, for the reasons expressed above.

- Instead, and following a similar logic:

$$E[y_i|y_i > a] = \mathbf{x}_i\boldsymbol{\beta} + \sigma\frac{\phi[(a - \mathbf{x}_i\boldsymbol{\beta})/\sigma]}{1 - \Phi[(a - \mathbf{x}_i\boldsymbol{\beta})/\sigma} = \mathbf{x}_i\boldsymbol{\beta} + \sigma\lambda(\alpha_i)$$

- What this implies, very basically, is that we need to add an additional term to a regression model where we have truncated data in order to correctly estimate our $\hat{\boldsymbol{\beta}}$ coefficients. If we don't include the term that adjusts for the probability that we see y_i to begin with, we will get an inconsistent estimate, similar to what would happen if we omitted a relevant variable.

- You can do this in `Stata` using the command **truncreg**. The truncated regression is estimated using maximum likelihood.

19.3 Censored data and tobit regression

- Suppose that instead of the data being simply missing, we observe the value of $y_i = a$, whenever $y_i < a$. We can imagine a latent variable, y_i*, but we observe a whenever $y_i* \leq a$, and y_i* otherwise. This is known as a case of "left-censoring."
- In this case: $E[y_i] = \Phi a + (1 - \Phi)(\mu + \sigma\lambda)$
- With probability Φ, $y_i = a$ and with probability $(1 - \Phi)$, $y_i = \mu + \sigma\lambda$.
- In the "nice" case, where $a = 0$, the expression above simplifies to:

$$E[y_i] = \Phi(\mu/\sigma)(\mu + \sigma\lambda)$$

- In a multivariate regression model, where normally $E[y_i|\mathbf{x}_i] = \mathbf{x}_i\boldsymbol{\beta}$, the censoring with $a = 0$ implies that:

$$E[y_i|\mathbf{x}_i] = \Phi\left(\frac{\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)(\mathbf{x}_i\boldsymbol{\beta} + \sigma\lambda_i)$$

  where $\lambda_i = \frac{\phi[(0-\mathbf{x}_i\boldsymbol{\beta})/\sigma}{1-\Phi[(0-\mathbf{x}_i\boldsymbol{\beta})/\mu]} = \frac{\phi(\mathbf{x}_i\boldsymbol{\beta}/\sigma)}{\Phi(\mathbf{x}_i\boldsymbol{\beta}/\sigma)}$
- If the censoring cut point is not equal to zero, and if we also have an upper-censoring cut point, then the conditional expectation of y_i will be more complicated, but the point remains that in estimating the conditional mean we must take account of the fact that y_i could be equal to $\mathbf{x}_i\boldsymbol{\beta}$ or to the censored value(s). If we don't take this into account, we are likely to get incorrect estimates of the $\hat{\boldsymbol{\beta}}$ coefficients just as we did in the case of the truncated regression.
- The censored model (with an upper or lower censoring point or both) is estimated in `Stata` using the **tobit** command.
- For an example of tobit in practice, see a paper presented at the American Politics seminar last year, "Accountability and Coercion: Is Justice Blind when It Runs for Office" by Greg Huber and Sanford Gordon, which recently appeared in the *American Journal of Political Science*.

19.4 Sample selection: the Heckman model

- Imagine that you want to estimate the impact of participating in a community group on one's level of satisfaction with democratic institutions. This is a key concern for scholars of social capital. Perhaps taking part in some community organization is likely to increase one's trust in government and to increase satisfaction with government performance. Thus, taking part in a community organization is like a treatment effect. We want to see how this treatment could affect outcomes for the "average" person."
- The problem is that the people participating in community groups are unlikely to be the same as the "average person." There may well be something about them that leads them to take part in community initiatives and, if this unobserved characteristic also makes them likely to trust government, we will bias our estimates of the treatment effect.
- The difficulty is that we only observe those people who joined the community groups.
- The situation is analogous to one of truncation. Here, the truncation is "incidental." It does not occur at given, definite values of y_i, but instead the truncation cuts out those people for whom the underlying attractiveness of community service did not reach some critical level.
- Let us call the underlying attractiveness of community service a random variable, z_i, and trust in government, our dependent variable of interest, y_i.
- The expression governing whether someone joins a community group or not is given by:

  $z_i^* = \mathbf{w}_i\boldsymbol{\gamma} + u_i$ and when $z_i^* > 0$ they join.
- The equation of primary interest is:

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$$

- We assume that the two error terms, u_i and ε_i are distributed bivariate normal, with standard errors, σ_u and σ_ε, and that the correlation between them is equal to ρ.

- Then we can show that:

$$\begin{aligned} E[y_i|y_i\, observed] &= E[y_i|z_i^* > 0] \\ &= E[y_i|u_i > -\mathbf{w}_i\boldsymbol{\gamma}] \\ &= \mathbf{x}_i\boldsymbol{\beta} + E[\varepsilon_i|u_i > -\mathbf{w}_i\boldsymbol{\gamma}] \\ &= \mathbf{x}_i\boldsymbol{\beta} + \rho\sigma_\varepsilon\lambda_i(\alpha_u) \\ &= \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\beta}_\lambda\lambda_i(\alpha_u) \end{aligned}$$

  where $\alpha_u = -\mathbf{w}_i\boldsymbol{\gamma}/\sigma_u$ and $\lambda_i(\alpha_u) = \phi(\mathbf{w}_i\boldsymbol{\gamma}/\sigma_u)/\Phi(\mathbf{w}_i\boldsymbol{\gamma}/\sigma_u)]$

- This sounds like a complicated model, but in fact it is not. The sample selection model is typically computed via the Heckman two-stage procedure.

  1. Estimate the probability that $z_i^* > 0$ by the probit model to obtain estimates of $\boldsymbol{\gamma}$. For each observation in the sample, compute $\hat{\lambda}_i(\alpha_u) = \phi(\mathbf{w}_i\hat{\boldsymbol{\gamma}}/\sigma_u)/\Phi(\mathbf{w}_i\hat{\boldsymbol{\gamma}}/\sigma_u)$.
  2. Estimate $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_\lambda$ by least squares regression of y_i on $\mathbf{x}_i$ and $\hat{\lambda}_i$.

- The model is estimated in `Stata` using the **heckman** command. The output that `Stata` gives you will indicate the coefficients in the model for $y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$, the coefficients in the model for $z_i^* = \mathbf{w}_i\boldsymbol{\gamma} + u_i$, and will also compute for you ρ and σ_ε.

19.5 Duration models

- Suppose the data that you have is on the length of civil wars or on the number of days of a strike. You are interested in the factors that prolong civil conflict or that bring strikes to an end speedily.

- You should not estimate this model using the Poisson distribution, because you are not interested in the number of days of unrest per se. You are interested in the probability that a war will end and this probability can depend on the time for which combatants have already been fighting.

- Thus, duration models, and the kinds of questions that go with them, should be investigated using techniques that permit you to include the effects of time as a chronological sequence, and not just as a numerical marker. These models will also allow you to test for the effect of other "covariates" on the probability of war ending.
- Censoring is also a common (but easily dealt with) problem with duration analysis.
- Let us begin with a very simple example in which we are examining the probability of a "spell" of war or strike lasting t periods. Our dependent variable is the random variable, T, which has a continuous probability distribution $f(t)$, where t is a realization of T.
- The cumulative probability is:

$$F(t) = \int_0^t f(s)ds = \Pr(T \leq t)$$

- Note that the normal distribution might not be a good choice of functional form for this distribution because the normal usually takes on negative values and time does not.

  The probability that a spell is of length at least t is given by the survival function:

$$S(t) = 1 - F(t) = \Pr(T \leq t)$$

- We are sometimes interested in a related issue. Given that a spell has lasted until time t, what is the probability that it ends in the next short interval of time, say Δt?:

$$l(t, \Delta t) = \Pr[t \leq T \leq t + \Delta t | T \geq t)$$

- A useful function for characterizing this aspect of the question is the hazard rate:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr[t \leq T \leq t + \Delta t | T \geq t]}{\Delta t} = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t S(t)} = \frac{f(t)}{S(t)}$$

Roughly, the hazard rate is the rate at which spells are completed after duration t, given that they last at least until t.

- Now, we build in different functional forms for $F(t)$ to get different models of duration. In the exponential model, the hazard rate, λ, is just a parameter to be estimated and the survival function, $S(t) = e^{-\lambda t}$. As such, the exponential model has a hazard rate that does not vary over time.

- The Weibull model allows the hazard rate to vary over the duration of the spell. In the Weibull model, $\lambda(t) = \lambda p(\lambda t)^{p-1}$, where p is also a parameter to be estimated and the survival function is $S(t) = e^{-(\lambda t)p}$. This model can "accelerate" or "decelerate" the effect of time and is thus called an *accelerated failure time* model.

- How do we relate duration to explanatory factors like how well the combatants are funded? As in the Poisson model, we derive a λ_i for each observation (war or strike), where $\lambda_i = e^{-\mathbf{x}_i\boldsymbol{\beta}}$.

- Either of these models can be estimated in `Stata` via maximum likelihood using the **streg** command. This command permits you to adopt the Weibull or the exponential or one of several other distributions to characterize your survival function.

- If you do not specify otherwise, your output will present the effect of your covariates on your "relative hazard rate" (i.e., how a one unit increase in the covariate moves the hazard rate up or down). If you use the command **streg depvar expvars, nohr**, your output will be in terms of the actual coefficients. In all cases, you can use the **predict** command after streg to get the predicted time until end of spell for each observation. In cases where the hazard is constant over time, you can also predict the hazard rate using **predict, hazard**.

- One of the problems that is sometimes encountered with duration data is individual level heterogeneity. That is, even if two conflicts had the same level of the explanatory factors, $\mathbf{x}$, they might differ in the length of conflict because of some non-observed, conflict specific factor (note the similarity to unobserved, country-specific factors in panel data).
- To estimate a duration model in this context, we use a "proportional hazards" model, such as the Cox proportional hazard model. This model allows us to say how hazard rates change as we alter the value of $\mathbf{x}$, but does not permit us to compute the absolute level of hazard (or *baseline hazard*) for each observation. This model is estimated in `Stata` using **stcox**.