ECON7772: Econometric Methods: Econometric Methods LECTURE NOTES PART 1

Arthur Lewbel, Department of Economics, Boston College

February 28, 2023

Contents

1 Lecture 01. Regression vs Correlation and Causation

1.1 Introduction

Suppose we have a sample of observations of variables Y and X. We can draw the sample as a "scatter plot." This plots each data point, where Y is on the vertical axis and X on the horizontal axis.

Suppose we believe Y and X are linearly related.

First objective: Draw the straight line that fits the data the best. Depends on what we mean by "best."

We can propose different rules for drawing a fitted line. For example:

1. connect the two lowest and highest points

2. minimize sum of total absolute deviations (distances of each point to the line)

3. minimize sum of squared differences of each point to the line point to the line. This is called a Deming regression.

4. minimize sum of total vertical absolute deviations (vertical distances of each point to the line). This is least absolute deviations (LAD) regression.

5. minimize sum of squared vertical distances of each point to the line. This is ordinary least squares (OLS) regression.

Choice of rule for drawing a fitted line corresponds choice of a "loss function," which describes how you penalize errors (that is, differences between the data points and the line).

Example: Suppose Y is sales (in millions of dollars) and X is advertising expenditures (in thousands of dollars). I want you to predict sales based on advertising. I'm going to deduct one dollar from your salary for every thousand dollars that your line's prediction of Y differs from the actual Y.

How should you fit the line?

In this case I've told you your loss function, i.e., how you're being penalized for errors.

First note I'm charging you for the error in Y, which means the vertical distance between the line and the points. So not choices 1, 2, or 3 above.

To minimize your losses (dollars deducted from salary), choose rule 4 above. Not rule 5, because that minimizes the squared difference in Y between fit and actual, while I'm charging you the absolute difference, not the square.

OLS penalizes large errors more, and small errors less, than LAD.

Using X to predict Y implicitly implies X in some way 'causes' Y. Standard way to write the model is $Y = a + bX + e$, where line is $a + bX$ and e is the (vertical) error.

A good fit means small e's, so the data points do lie close to a line. A good fit does NOT mean the model is right. In particular, we might get a good fit even if X does not cause Y.

Potential problems in fitting linear regression model:

1. Spurious correlation. When two phenomena seem to behave in a connected way but there is actually no real causality relationship between the two. Example: the width of the Atlantic ocean and the population of the world are positively correlated. Both are increasing over time, but for completely different reasons.

2. Direction of causality. You may think X causes Y, but maybe Y actually causes X, or both simultaneously help determine the other. Example: Story of a tribe in South America that began chasing away missionary doctors, because they noticed that doctors correlated with illness, and assumed the doctors caused the illnesses. Did they have the direction of causality wrong? Maybe not; Europeans did bring many diseases like small pox to the Americas. Moral: You have biases and beliefs about how the world works (direction of causality), but your model of the world might be wrong.

3. Third cause. It may be that neither X causes Y or Y causes X, but instead both are (partly) caused by some other variable Z. Example: Over long time periods, the total number of people NOT working in the US is positively correlated with US GDP. Why? Third cause: both increase due to population increasing.

A good or bad fit does NOT mean the model is right or wrong. (George Box: All models are wrong, but some are useful). All models are wrong because the real world is always more complicated than our models. But a model is useful if it correctly informs us about some aspects of the world.

A good or bad fit does not tell us if a model is useful or not. You ALWAYS need some theory, that is, some idea about how the data were generated.

Good model building requires:
1. Some reliable theory (economics and/or behavior and/or or randomization) of how the data are generated.
2. Statistics to help identify potential problems and errors with our theory.

Note: Machine Learning finds correlations, it finds patterns in data, but it doesn't tell us about causality or which models are correct. Randomized experiments can help determine causality, but they too don't tell us what models are useful.

1.2 Ordinary Least Squares

Consider the model:

$$Y_i = a + bX_i + e_i$$

Here i indexes observations. Assume $i = 1, ..., n$, so we have n observations. Writing the line this way, we are assuming Y depends on X.

Y is the dependent variable, X is the independent variable.

e_i is the model error. It is the difference between Y_i and the line $a + bX_i$.

The estimated values of parameters a and b are denoted $\widehat{a}$ and $\widehat{b}$.

$\widehat{Y}_i$ is the fitted value of Y_i. It is the predicted or estimated value of Y_i given X_i.

$\widehat{Y}_i = \widehat{a} + \widehat{b}X_i$ is the fitted regression line.

$\widehat{e}_i = Y_i - \widehat{Y}_i$, so $Y_i = \widehat{a} + \widehat{b}X_i + \widehat{e}_i$. $\widehat{e}_i$ is the estimated error. $\widehat{e}_i$ is called the residual.

Ordinary Least Squares (OLS) $\widehat{a}$ and $\widehat{b}$ are estimates based on the loss function of minimizing the sum of squared errors.

OLS: $\min \sum_{i=1}^{n} e_i^2$ or $\min \sum_{i=1}^{n} (y_i - a - bx_i)^2$
so that

$$(\hat{a}, \hat{b}) = \arg\min \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

To minimize, take the derivative with respect to a and b, set those derivatives equal to zero, and solve for a and b. Those solutions are $\widehat{a}$ and $\widehat{b}$. Those solutions take the form:

$$\begin{aligned} \hat{b} &= \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\widehat{Cov(X,Y)}}{\widehat{Var(X)}} \\ \hat{a} &= \overline{Y} - \hat{b}\overline{X} \end{aligned}$$

where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample average of X (and similar for $\overline{Y}$), and $\widehat{Cov(X,Y)}$ and $\widehat{Var(X)}$ are the estimated covariance of X and Y and estimated variance of X.

Notes: OLS regression chooses the constant $\widehat{a}$ to make the regression line go through the center of the data, i.e., to make $\overline{Y} = \hat{a} + \hat{b}\overline{X}$.

By the definition of a slope, $b = dY/dX$.

Example: let Y_i be GDP in billions of dollars of country i, and X_i be population in millions of country i. Suppose we have $n = 3$ observations of (Y_i, X_i), which are $(1, 3)$, $(6, 7)$ and $(2, 2)$.

If apply the above OLS formula, get $\widehat{b} = 13/14 = 0.929$ and $\widehat{a} = 3 - 4(13/14) = -0.714$ (try plugging in the data to see if you can get these values).

What do these estimates mean? $\widehat{b}$ is the slope. to interpret correctly, pay attention to the units Y and X are measured in. $\widehat{b} = 0.929$ says if population goes up by one million people, GDP increases by 0.929 billion dollars, so one more person increases GDP by \$929. Is this reasonable? Depends on how rich the countries are.

Intercept $\widehat{a} = -0.714$. Says if country had zero people, GDP would be -714 million dollars. Is this sensible? Remember that a regression is only fitting the range of the data. In our data, the country populations range from one million to six million. This is very far from zero people, so we don't expect the regression to fit zero people well. The intercept may be reasonable for the actual range of the data.

Other issues:

True relationship may not be linear.

Not sure about direction of causality (High GDP could encourage migration leading to a higher population). If we regressed X on Y we would get a different line.

Results depend on quantity of the data. With only 3 points, we could be way off. Just one more point could radically change the estimates.

Results depend on quality of the data. We don't know the quality of the surveys data are based on. The data are clearly rounded to the nearest million people and billion dollars, that alone could greatly affect the results.

1.3 Statistics Review

- **Notation**

  1. r.v. means random variable or random vector. A random vector or a random matrix is a vector of matrix, each element of which is a random variable.
  2. Capital letter X: random variable (or random vector or random matrix)
  3. Lower-case letter x: realization of X
  4. Example X denotes the role of a die (die is the singular of dice). So X is an r.v. that could equal any number from 1 to 6, each with probability 1/6. Suppose we roll the die and a three comes up. Then x, the realization, is three. Realizations are constants, not random variables.

- **Properties of Expectations**

  If X is a random $n \times m$ matrix (meaning each element X_{ij} is a random variable) then:

  1. $E(X) =$ the matrix of elements $E(X_{ij})$.
  2. Let $'$ denote matrix transpose. $E(X') = E(X)'$
  3. For any constant matrices A and B: $E(AXB') = AE(X)B'$
  4. For any random $n \times m$ matrices X and Z, if $E(X + Z) = E(X) + E(Z)$
  5. For any random $n \times m$ matrix X and any random $m \times k$ matrix Z, if every element of X has zero covariance with every element of Z, then $E(XZ) = E(X)E(Z)$
  6. For any random $n \times m$ matrix X and any random $m \times k$ matrix Z, $E(XZ \mid Z) = E(X \mid Z) Z$

- **Variance**

  1. Let $X = (X_1, X_2, ...X_J)'$ be a random $J \times 1$ vector. Then $Var(X)$ is a $J \times J$ matrix given by

  $$\begin{aligned} Var(X) &= E[(X - E(X))(X - E(X))'] \\ &= E(XX') - E(XE(X)') - E(E(X)X') + E(X)E(X)' \\ &= E(XX') - E(X)E(X)' \end{aligned}$$

  The ij element of $Var(X)$ equals $cov(X_i, X_j)$. The i'th element on the diagonal of $Var(X)$ equals $var(X_i)$

2. If A is a constant $K \times J$ matrix, X is a random $J \times 1$ matrix, b is a constant $K \times 1$ vector then
$$Var(AX + b) = AVar(X)A'$$

- **Covariance**

1. Suppose X is a random $J \times 1$ vector and Z is a random $K \times 1$ vector. Then $Cov(X, Z)$ is the $J \times K$ matrix $Cov(X, Z) = E[(X - E(X))(Z - E(Z))']$. The element of row j and column k of $Cov(X, Z)$ is $Cov(X_j, Z_k)$.
2. $E(XZ') = Cov(X, Z) + E(X)E(Z')$.
3. For constant matrics A and B and constant vectors c and d, we have $Cov(AX + c, BZ + d) = Acov(X, Z)B'$.

- **The Multivariate Normal Distribution**

1. Suppose X is a random $J \times 1$ vector. The notation $X \,\tilde{}\, N(\mu, \Omega)$ means that X has a multivariate normal distribution with $E(X) = \mu$ and $Var(X) = \Omega$.
2. If $X \,\tilde{}\, N(\mu, \Omega)$, then for constant matrix A and constant vector b, $AX + b \,\tilde{}\, N(A\mu + b, A\Omega A')$
3. If X and Z are jointly multivariate normal vectors (that is, if $(X', Z')'$ is a multivariate normal vector) and $Cov(X, Z) = 0$, then X and Z are independent of each other. This is a special property of normals. Given any two random vectors U and V having any distributions, if U and V are independent of each other, then $Cov(U, V) = 0$. But the converse of that statement is usually not true. Normals are an exception.
4. $X \,\tilde{}\, N(\mu, \Omega)$, and Ω is nonsingular, then $(X - \mu)' \Omega^{-1} (X - \mu)$ has a Chi-squared distribution with J degrees of freedom, denoted as $(X - \mu)' \Omega^{-1} (X - \mu) \,\tilde{}\, \chi^2_J$.
5. The multivariate standard normal distribution is $N(0, I_J)$, where I_J is the $J \times J$ identity matrix. It follows from above that if the random $J \times 1$ vector S has $S \,\tilde{}\, N(0, I_J)$, then the elements of S are each standard normal random variables, each independent of the others, and $S'S = \sum_{j=1}^{J} S_j^2 \,\tilde{}\, \chi^2_J$. So a Chi-squared statistic with J degrees of freedom equals the sum of J independent squared standard normals.

- **Derivatives**

1. The derivative of a scalar with respect to a J vector is called a gradient. The gradient equals the J vector of derivatives of the scalar with respect to each element of the J vector.
2. If a is a constant $J \times 1$ vector and X is a random $J \times 1$ vector, then $a'X = a_1X_1 + a_2X_2 + ... + a_JX_J$ and
$$\frac{\partial(a'X)}{\partial a} = \begin{bmatrix} \frac{\partial(a'X)}{\partial a_1} \\ \vdots \\ \frac{\partial(a'X)}{\partial a_J} \end{bmatrix} = \begin{bmatrix} X_1 \\ \vdots \\ X_J \end{bmatrix} = X$$

3. If X is a random $J \times J$ matrix and a is a constant $J \times 1$ vector then

$$\frac{\partial a'Xa}{\partial a} = (X + X')a$$

   If X is symmetric, so $X = X'$, then this gradient equals $2Xa$.

4. Let X be a random matrix and a be a constant vector. Assume function $g(X, a)$ is continuously differentiable in a, and that there exists a random vector Z with finite mean such that $|\partial g(X, a) / \partial a| \leq Z$. Then

$$E\left(\frac{\partial g(X, a)}{\partial a}\right) = \frac{\partial E(g(X, a))}{\partial a}$$

   so the derivative of an expectation equals the expectation of the derivative. This is closely related to Fubini's theorem. The above conditions are actually a little stronger than necessary.

- **Law of iterated expectation**

$$E(E(X \mid Z)) = E(X)$$

where X and Z are random vectors (aside: this is holding the information set for expectations fixed).

- **Moments**

If $E(|X|^r)$ is bounded for some $r > 0$, then $E(|X|^s)$ is bounded for any positive s smaller than r. This can be proved by observing that $|x|^s \leq 1 + |x|^r$ for any x.

- **Independence and Uncorrelatedness**

Let X and Z be random vectors with $E(X) = \mu_X$ and $E(Z) = \mu_Z$.

1. Suppose X and Z are independent. This is denoted. $X \perp Z$. Independence implies that knowing the realization of Z will tell you nothing about the realization of X. If X and Z are continuously distributed, then independence means for their probability density functions (pdf) $f_{X,Z}(x, z) = f_X(x)f_Z(z)$.
2. X is said to be "mean independent" of Z if $E(X \mid Z) = E(X) = \mu_X$. Mean independence says that knowing the realization of Z does not tell you anything about the mean of X, though it might tell you something about other features of the distribution of X
3. X is said to be uncorrelated with Z if $E(XZ') = E(X)E(Z') = \mu_X\mu_Z'$, or equivalently if $Cov(X, Z) = 0$. Uncorrelatedness means that X and Z are not linearly related. Roughly, knowing the realization of Z does not linearly provide information about the mean X, though the mean of X might be nonlinearly related to Z.

"Independent" is the strongest while "uncorrelated" is the weakest.

1. Independent$\Longrightarrow$Mean independent:

   If X and Z are independent, then X is mean independent of Z (the converse does not hold):

$$E(X|Z) = \int x f_{X|Z}(x|z)dx = \int x \frac{f_{X,Z}(x,z)}{f_Z(z)}dx = \int x \frac{f_X(x)f_Z(z)}{f_Z(z)}dx = \int x f_X(x)dx = E(X)$$

2. Mean independent$\Longrightarrow$Uncorrelated:

   If X is mean independent of Z, then X and Z are uncorrelated (the converse does not hold):

$$\begin{aligned} E(XZ') &= E(E(XZ'|Z)) = E(E(X|Z)Z') \\ &= E(E(X)Z') \text{ because mean independence says E(X|Z)=E(X)} \\ &= E(X)E(Z') \text{ because E(X) is constant, so it comes out of the expectation} \end{aligned}$$

- **Existence of Expectation**

  What does it mean to say that an expectation exists?

  Let $X^+ = \max(0, X)$ and $X^- = \min(0, X)$, then $X = X^+ + X^-$ by construction, and so $E(X) = E(X^+) + E(X^-)$

  1. If both $E(X^+)$ and $E(X^-)$ are finite, then $E(X)$ is finite
  2. If $E(X^+) = +\infty$ and $E(X^-)$ is finite, then $E(X) = +\infty$
  3. If $E(X^-) = -\infty$ and $E(X^+)$ is finite, then $E(X) = -\infty$
  4. If both $E(X^+)$ and $E(X^-)$ are infinite, then $E(X)$ does not exist.
  5. So $E(X)$ exists if either $E(X^+)$ is finite or $E(X^-)$ is finite.

2 Lecture 02. Finite Sample Properties of Estimators

Readings for this lecture is: Greene Appendix C.

2.1 Definitions and Concepts of an Estimator

- A random variable (rv) is denoted by capital letters, e.g., Z. Realizations are denoted by small letters, e.g., z.

  Random variables have probability distribution functions (PDF). Features of a pdf, like moments (mean, variance, etc) or quantiles (median, deciles), are constants.

  A sequence of random variables: $Z_1, Z_2, Z_3, \ldots$

  Corresponding sequence of realizations $z_1, z_2, z_3, \ldots$

  Each element of the sequence Z_i for $i = 1, 2, ...$ has a distribution. If each Z_i has the same distribution, we say the sequence is *identically* distributed. If each Z_i is independently distributed from the other elements of the sequence, then we say the sequence is *independently* distributed. A sequence that is both is called *iid (independently, identically distributed).*

- What is a statistic?

  A statistic is a function of random variables. A statistic therefore itself also a random variable. A statistic therefore itself has a distribution.

  A sequence of statistics is another sequence of random variables:

  $X_1 = h_1(Z_1)$, $X_2 = h_2(Z_1, Z_2)$, $X_3 = h_3(Z_1, Z_2, Z_3), \ldots$

  The distribution of each element of the sequence of statistics $X_1, X_2, X_3, \ldots$ depends on the distribution of the elements of the underlying sequence $Z_1, Z_2, Z_3, \ldots$

- What is an estimator? An estimator is a statistic that tells us something about a probability distribution.

  Example: suppose $E(Z_i) = \mu$. Note the mean μ is a constant, a feature of the pdf of each Z_i.

  The sample mean (average) is an estimator of μ: $X_n = \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Z_i$

  An estimate is the realization of an estimator: An estimator is: $X_n = h_n(Z_1, ..., Z_n)$, the corresponding estimate is: $x_n = h_n(z_1, ..., z_n)$

  An estimand is the object (constant) that an estimator is trying to estimate. Example: sample mean of iid $Z_1, Z_2, Z_3, \ldots Z_n$ is $\hat{\mu}$. The estimand is μ.

  Note, it is common to put a hat on an estimand to denote an estimator or an estimate. Above example estimand μ, and $\hat{\mu}$ can denote either the estimator or the estimate (so the notation is not clear about that).

  It is also common to use a subscript zero to denote the true value of an estimand. So we could say we want to estimate a parameter θ, we have an estimator $\widehat{\theta}$, and the unknown true value of the estimand is θ_0.

- Properties of estimators:

An estimator $X_n = h_n(Z_1, ..., Z_n)$ is a function of random variables $Z_1, ..., Z_n$, and so is itself a random variable, and so has a distribution. The estimand is a feature of the distribution of $Z_1, ..., Z_n$. Properties of the estimator X_n are features of the distribution of X_n, and how those features relate to the estimand.

Example: suppose we have data $z_1, z_2, z_3, \ldots z_{10}$. Suppose $Z_1, Z_2, Z_3, \ldots$ are iid with $E\left(Z_i\right) = \mu$.

Consider the estimator $X_{10} = \frac{1}{10}\sum_{i=1}^{10} Z_i$. The estimate is $x_{10} = \frac{1}{10}\sum_{i=1}^{10} z_i$. This estimate is a single number we calculate.

Hypothetical thought experiment: Suppose we could have millions of samples, each consisting of 10 realizations. Each sample would give us another estimate x_{10}. We could make a histogram of those millions of estimates of x_{10}. That histogram would be a very good approximation of the probability density function (pdf) or probability mass function (pmf) of the distribution of X_{10}. Properties of the estimator are features of that distribution (e.g., its mean, its variance, median, etc), and how those features relate to the estimand μ.

Note that if we had different sample size, say 9 or 11 instead of 10, then the estimator would be X_9 or X_{11}, and these would have different distributions from X_{10}. The properties of an estimator (features of its distribution) generally depend on the sample size.

- Two types of properties of estimators

  Finite sample properties: Features of the distribution of the estimator for a given sample size.

  Asymptotic properties: How the features of distribution of the estimator change, and what they become, as the sample size goes to infinity.

2.2 Unbiasedness

- Definition

  Unbiased: for any estimator $\hat{\theta}$ of θ, $\hat{\theta}$ is unbiased if $E(\hat{\theta}) = \theta$

  Bias: for any estimator $\hat{\theta}$ of θ, $bias(\hat{\theta}) = E(\hat{\theta}) - \theta$. Obviously, if $\hat{\theta}$ is unbiased, then $bias(\hat{\theta}) = 0$.

- Example:

  Theorem: Assume a sequence of r.v.'s $Z_1, Z_2, \ldots, Z_n$, where $E(Z_i) = \mu$ is finite (note we are not requiring independent or identical distributions, just the same mean for each). Then the sample mean $X_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$ is unbiased.

  Proof:

$$E(X_n) = E(\frac{1}{n}\sum_{i=1}^{n} Z_i) = \frac{1}{n}E\left(\sum_{i=1}^{n} Z_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Z_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}n\mu = \mu$$

  Don't be confused: $E(X_n)$ is the true mean of the estimator X_n. The unbiasedness property relates a feature of the distribution of the estimator (it's mean) to the estimand. In this case, the estimand also happens to be a mean, namely, μ, the mean of each Z_i.

  What does the unbiasedness property here, that $E(X_n) = \mu$, actually mean? Return to the thought experiment. Suppose we had an infinite number of samples, each of size n, and we calculate the realized sample mean x_n for each sample. For some samples we'd get $x_n > \mu$, other samples would give $x_n < \mu$, but the average value of x_n across the infinite number of samples would equal μ.

- Another example of an unbiased estimator is the sample variance. Suppose we have a sequence $Z_1, Z_2, Z_3, \ldots$ is iid with finite $E\left(Z_i\right) = \mu$ and finite $var\left(Z_i\right) = \sigma^2$. Let $\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$. Then it can be shown that

$$E\left(\frac{1}{n-1}\sum\nolimits_{i=1}^{n}\left(Z_i - \overline{Z}\right)^2\right) = \sigma^2$$

- $E(\hat{\theta}) - \theta$ is sometimes called mean bias. We can also define median bias equals $med(\hat{\theta}) - \theta$, the difference between the median of $\hat{\theta}$ and θ. People also sometimes, informally, use bias to mean the difference between the probability limit of $\hat{\theta}$ and θ, which for inconsistent estimators is nonzero (we will define probability limits and consistency later).

2.3 Variance, Efficiency, Mean Squared Error

- Bias is about the mean of the distribution of an estimator X_n. Another property of an estimator is its variance.

  Theorem: Assume $Z_1, Z_2, \ldots, Z_n$ are independent identically distributed (iid), with $E(Z_i) = \mu$ and $var(Z_i) = \sigma^2$, finite. Let X_n be the sample mean: $X_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$. Then $Var(X_n) = \sigma^2/n$.

  Proof:

$$\begin{aligned}
Var(X_n) &= E[(X_n - E(X_n))^2] \\
&= E\left[\left(\left(\frac{1}{n}\sum\nolimits_{i=1}^{n} Z_i\right) - \mu\right)^2\right] \\
&= E\left[\left(\left(\frac{1}{n}\sum\nolimits_{i=1}^{n} Z_i\right) - \left(\frac{1}{n}\sum\nolimits_{i=1}^{n} \mu\right)\right)^2\right] \\
&= E\left[\left(\frac{1}{n}\sum\nolimits_{i=1}^{n} \left(Z_i - \mu\right)\right)^2\right] \\
&= \frac{1}{n^2}E\left[\left(\sum\nolimits_{i=1}^{n} \left(Z_i - \mu\right)\right)^2\right] \\
&= \frac{1}{n^2}E\left[\left(\sum\nolimits_{i=1}^{n} \left(Z_i - \mu\right)\right)\left(\sum\nolimits_{j=1}^{n} \left(Z_j - \mu\right)\right)\right] \\
&= \frac{1}{n^2}E\left[\sum\nolimits_{i=1}^{n}\sum\nolimits_{j=1}^{n} \left(Z_i - \mu\right)\left(Z_j - \mu\right)\right] \\
&= \frac{1}{n^2}E\left[\left(\sum\nolimits_{i=1}^{n} \left(Z_i - \mu\right)^2\right) + \left(\sum\nolimits_{i=1}^{n}\sum\nolimits_{j\neq} \left(Z_i - \mu\right)\left(Z_j - \mu\right)\right)\right] \\
&= \frac{1}{n^2}\left[\left(\sum\nolimits_{i=1}^{n} E\left[\left(Z_i - \mu\right)^2\right]\right) + \left(\sum\nolimits_{i=1}^{n}\sum\nolimits_{j\neq} E\left[\left(Z_i - \mu\right)\left(Z_j - \mu\right)\right]\right)\right] \\
&= \frac{1}{n^2}\left[\left(\sum\nolimits_{i=1}^{n} var\left(Z_i\right)\right) + \left(\sum\nolimits_{i=1}^{n}\sum\nolimits_{j\neq} cov\left(Z_i, Z_j\right)\right)\right] \\
&= \frac{1}{n^2}\sum\nolimits_{i=1}^{n} \sigma^2 = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}
\end{aligned}$$

This shows the bigger the sample size, the smaller is the variance of X_n. Recall that X_n is an unbiased estimator of μ. So the bigger the sample size n, the smaller is the variance of X_n, and so the less likely it is that X_n is far from μ.

- Definition:

  Given two unbiased estimators $\hat{\theta}$ and $\tilde{\theta}$, we say $\hat{\theta}$ is "more efficient" than $\tilde{\theta}$ if $Var(\hat{\theta}) \leq Var(\tilde{\theta})$.

  Also, an unbiased estimator $\hat{\theta}$ is said to be "efficient" if $Var(\hat{\theta}) \leq Var(\tilde{\theta})$ for any other unbiased estimator $\tilde{\theta}$. In general, more efficient estimators are preferred over less efficient.

- Suppose we have two estimators. One is unbiased, but with a big variance. The other is biased, but with a small variance. Which one is better? One possible criterion for choosing is called "Mean Squared Error," or MSE.

  The MSE of an estimator $\hat{\theta}$ is defined by $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$. Another formula for MSE is:

$$\begin{aligned}
MSE(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] \\
&= E[(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2] \\
&= E[(\hat{\theta} - E(\hat{\theta}))^2 + 2(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta)) + (E(\hat{\theta}) - \theta)^2] \\
&= Var(\hat{\theta}) + 2E[\hat{\theta} - E(\hat{\theta})]bias(\hat{\theta}) + \left[bias(\hat{\theta})\right]^2 \\
&= Var(\hat{\theta}) + 0 + \left[bias(\hat{\theta})\right]^2 \\
&= Var(\hat{\theta}) + \left[bias(\hat{\theta})\right]^2
\end{aligned}$$

  So MSE is a measure of the trade-off between variance and bias of an estimator. In the earlier example, we might prefer the biased estimator over the unbiased estimator if the biased one has a smaller MSE.

3 Lecture 03. Asymptotic Properties

Readings for this lecture are: Eric Zivot's Primer, Greene Appendix D, and The Newey and McFadden handbook chapter.

3.1 Convergence in Mean Square and in Probability

- Consider a sequence of rv's $X_1, X_2, ..., X_n$. Assume the mean and variance of each X_i is finite. Let $\mu_i = E(X_i)$ and $\sigma_i^2 = Var(X_i)$. Note we did NOT assume iid, so each X_i can have its own mean and variance.

- Convergence in mean square

  Let $\mu_n = E(X_n)$ and $\sigma_n^2 = Var(X_n)$. Then X_n converges in mean square to a constant c if

  $$\lim_{n\to\infty} \mu_n = c \text{ and } \lim_{n\to\infty} \sigma_n^2 = 0$$

  Usually, written as $X_n \overset{ms}{\to} c$.

  So X_n converges in mean square to c if, as n goes to infinity, the mean of X_n approaches c, and the variance of X_n goes to zero.

  Equivalently, X_n converges in mean square to c if

  $$\lim_{n\to\infty} E(X_n - c)^2 = 0$$

  Example: Suppose $Z_1, Z_2, ..., Z_n$ is an iid sequence of random variables, with finite mean and variance $E(Z_i) = \mu$ and $Var(Z_i) = \sigma^2$. Consider the statistic $X_n = \frac{1}{n}\sum_{i=1}^n Z_i$

  $$\lim_{n\to\infty} E(X_n) = \lim_{n\to\infty} \mu = \mu \text{ and } \lim_{n\to\infty} Var(X_n) = \lim_{n\to\infty} \frac{1}{n}\sigma^2 = 0$$

  so $X_n \overset{ms}{\to} \mu$

- Convergence in probability

  X_n converges in probability to a constant c if

  $$\lim_{n\to\infty} \Pr(|X_n - c| > \varepsilon) = 0 \text{ for any } \varepsilon > 0$$

  Usually, written as $plim\,(X_n) = c$ or $X_n \overset{p}{\to} c$.

  This definition says, pick any small number ε. The probability that X_n is more than ε away from c goes to zero as the sample size goes to infinity.

- Theorem: if $X_n \overset{ms}{\to} c$ then $X_n \overset{p}{\to} c$.

  Convergence in mean square implies convergence in probability. The converse is not true.

Suppose $X_n \overset{ms}{\to} c$. By the Markov (Chebyshev) inequality, we have

$$\begin{aligned} Pr\left(|X_n - c| > \varepsilon\right) &= Pr((X_n - c)^2 > \varepsilon^2) \\ &\leq \frac{E[(X_n - c)^2]}{\varepsilon^2}. \end{aligned}$$

By the definition of convergence in mean square, we have $\lim_{n\to\infty} E(X_n - c)^2 = 0$.

$$\begin{aligned} Pr\left(|X_n - c| > \varepsilon\right) &\leq \frac{E[(X_n - c)^2]}{\varepsilon^2} \\ \lim_{n\to\infty} Pr\left(|X_n - c| > \varepsilon\right) &\leq \lim_{n\to\infty} \frac{E[(X_n - c)^2]}{\varepsilon^2} \\ &= \frac{0}{\varepsilon^2} \\ &= 0 \end{aligned}$$

Therefore $X_n \xrightarrow{p} c$
This result is extremely useful, because showing convergence in mean squared error is often easier than showing convergence in probability.

- Example: Suppose we have sequence of independent random variables $X_1, X_2, ...$ Assume $X_n = 0$ with probability $1 - 1/n$ and $X_n = n^2$ with probability $1/n$.

  Then $\lim_{n\to\infty} \Pr(|X_n - 0| > \varepsilon) = \lim_{n\to\infty} \Pr(X_n = n^2) = \lim_{n\to\infty} 1/n = 0$ so then $X_n \xrightarrow{p} 0$.

  But $E(X_n) = 0(1 - 1/n) + n^2\,(1/n) = n \to \infty$.

  So $X_n \overset{ms}{\nrightarrow} 0$. This is an example of sequence that converges in probability (to zero) but does not converge in mean square. Note also the contrast that $X_n \xrightarrow{p} 0$ even though $E(X_n) = n$.

- Convergence to a random variable. Consider a sequence of rv's $X_1, X_2, ..., X_n$, and a random variable Y. We say $X_n \overset{ms}{\to} Y$ if $(X_n - Y) \overset{ms}{\to} 0$. Similarly, we say then $X_n \xrightarrow{p} Y$ if $(X_n - Y) \xrightarrow{p} 0$.

3.2 Consistency

Definition: an estimator $\hat{\theta}$ is a consistent estimator for θ if $\hat{\theta} \xrightarrow{p} \theta$.

Example: Assume Z_i are a sequence of i.i.d. random variables with $E(Z_i) = \mu$ and $Var(Z_i) = \sigma^2 > 0$ finite.
Consider the sample average $X_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$. Is X_n a consistent estimator of μ?
We showed $E(X_n) = \mu$ and $Var(X_n) = \sigma^2/n$.
It follows that $\lim_{n\to\infty} E\left(X_n\right) = \lim_{n\to\infty} \mu = \mu$ and $\lim_{n\to\infty} Var(X_n) = \lim_{n\to\infty} \sigma^2/n = 0$.
Therefore $X_n \overset{m.s}{\to} \mu$, which means that $X_n \xrightarrow{p} \mu$. So we have proven the average X_n is a consistent estimator of the true mean μ.

Note: while convergence in mean square is a stronger condition than convergence in probability, it is often much easier to prove convergence in mean square (when it is true). So often the way we prove estimators are consistent is to prove they converge in mean square to the estimand.

Note: Bias, efficiency, and MSE are finite sample properties of estimators. Consistency is an asymptotic property.

We showed above that the sample average estimator is both unbiased and consistent. Other combinations are possible.

Example 1: Continue to assume that Assume Z_i are a sequence of i.i.d. random variables with $E(Z_i) = \mu$ and $Var(Z_i) = \sigma^2 > 0$ finite, and $X_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$. Consider the estimator for μ given by $Y_n = \frac{1}{n+3}\sum_{i=1}^{n} Z_i$. Is it biased? Is it consistent?
$Y_n = \frac{n}{n+3}\frac{1}{n}\sum_{i=1}^{n} Z_i = \frac{n}{n+3}X_n$

Biasedness: $E(Y_n) = E(\frac{n}{n+3}X_n) = \frac{n}{n+3}E(X_n) = \frac{n}{n+3}\mu \neq \mu$. This estimator is biased.

Consistency: $\lim_{n\to\infty} E\left(Y_n\right) = \lim_{n\to\infty}\frac{n}{n+3}\mu = \mu$,
$Var(Y_n) = Var(\frac{n}{n+3}X_n) = \frac{n^2}{(n+3)^2}Var(X_n) = \frac{n^2}{(n+3)^2}\frac{\sigma^2}{n} = \frac{n}{(n+3)^2}\sigma^2$
So $\lim_{n\to\infty} Var(Y_n) = \lim_{n\to\infty}\frac{n}{(n+3)^2}\sigma^2 = 0$. Therefore $Y_n \overset{m.s}{\to} \mu \Rightarrow Y_n \overset{p}{\to} \mu$. This estimator is consistent.

You might think that Y_n is an inferior estimator, because it's biased, However, Y_n also has smaller variance than the sample average X_n. So depending on n, it is possible that Y_n has a smaller MSE than X_n, and so it might be preferred to X_n (depending on what one's loss function is). This Y_n is an example of what's called a shrinkage estimator, i.e., an estimator that one biases towards zero to reduce variance and so hopefully improve MSE.

Example 2: Continue to assume that Assume Z_i are a sequence of i.i.d. random variables with $E(Z_i) = \mu$ and $Var(Z_i) = \sigma^2 > 0$ finite. Consider the estimator for μ given by $W_n = Z_1$ (just the first observation). This estimator is unbiased but inconsistent.
Unbiasedness: $E(W_n) = E(Z_1) = \mu$.
Inconsistent: Z_1 is a random variable with a strictly positive variance, so there must exist an ε such that $\Pr(|Z_1 - \mu| > \varepsilon) > 0$, and this expression does not depend on n so it holds in the limit as $n \to \infty$
Note that this W_n also does not converge in mean square. But while convergence in mean square implies convergence in probability, it is still possible to be consistent but not converge in mean square, so we had to directly check the convergence in probability condition.

3.3 Law of Large Numbers (LLN)

- A law of large numbers (LLN) is a condition under which a sample average $\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$ converges to the true mean.
- Example: we showed if the sequence Z_i is i.i.d. with $E(Z_i) = \mu$, $Var(Z_i) = \sigma^2$ finite then $\overline{Z} \overset{ms}{\to} \mu$, which implies consistency: $\overline{Z} \overset{p}{\to} \mu$
- The condition that variance be finite is stronger than necessary for consistency of $\overline{Z}$ with iid data. The weak law of large numbers (WLLN) says that if the sequence Z_i is i.i.d. with $E(Z_i) = \mu$, and $E(|Z_i|)$ finite, then $\overline{Z} \overset{p}{\to} \mu$. This allows Z_i to have slightly thicker tails than is the case with finite variance. (note: Kolmogorov showed this condition suffices for a stronger form of convergence called 'almost sure' convergence, and Khinchin showed a slightly weaker necessary condition for convergence in probability)

- The Chebyshev LLN: Suppose we have a sequence of Z_i that are NOT i.i.d. Let $E(Z_i) = \mu_i$, $Var(Z_i) = \sigma_i^2$, and $Cov(Z_i, Z_j) = \sigma_{ij}$, which we assume are finite. If $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mu_i = \mu_0$ is finite and and $\lim_{n\to\infty} Var(\overline{Z}) = 0$, then $\overline{Z} \xrightarrow{p} \mu_0$.

  Proof: $\lim_{n\to\infty} E\left(\overline{Z}\right) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mu_i = \mu_0$. And $\lim_{n\to\infty} Var(\overline{Z}) = 0$. So $\overline{Z} \xrightarrow{m.s} \mu_0$, and therefore $\overline{Z} \xrightarrow{p} \mu_0$

  Note that $Var(\overline{Z}) = \frac{1}{n^2}\sum_{i=1}^n \left(\sigma_i^2 + \sum_{j\neq i}\sigma_{ij}\right)$. The assumption $Var(\overline{Z}) \to 0$ means that $Var(Z_n)$ cannot grow to quickly with n, and that most covariances σ_{ij} to be zero (or go to zero as i and j increase).

3.4 The Continuous Mapping Theorem

- The Continuous Mapping Theorem: Given a sequence of random vectors X_n, For any continuous function g that does not depend on n we have

$$plim\,[g(X_n)] = g(plim\,(X_n))$$

  The probability limit of a function equals that function of the probability limit.

  This implies if $\hat{\theta} \xrightarrow{p} \theta$, then $g(\hat{\theta}) \xrightarrow{p} g(\theta)$. If $\hat{\theta}$ is a consistent estimator of θ, then $g(\hat{\theta})$ is a consistent estimator of $g(\theta)$.

- Example: consider the regression model $Y_i = bW_i + e_i$

  When is OLS consistent? The OLS estimator of b is

$$\hat{b}_{OLS} = \frac{\sum_i W_iY_i}{\sum_i W_i^2} = \frac{\sum_i W_i(bW_i + e_i)}{\sum_i W_i^2} = b + \frac{\sum_i W_ie_i}{\sum_i W_i^2} = b + \frac{\frac{1}{n}\sum_i W_ie_i}{\frac{1}{n}\sum_i W_i^2}$$

  Here X_n is the 2 element vector $X_n = (X_{1n}, X_{2n}) = \left(\frac{1}{n}\sum_i W_ie_i, \frac{1}{n}\sum_i W_i^2\right)$ and g is the function $g\left(X_n\right) = b + \frac{X_{1n}}{X_{2n}}$ so, by the theorem

$$plim\left(b + \frac{X_{1n}}{X_{2n}}\right) = b + \frac{plim\,(X_{1n})}{plim\,(X_{2n})}$$

$$plim(\hat{b}_{OLS}) = plim\left(b + \frac{\frac{1}{n}\sum_i W_ie_i}{\frac{1}{n}\sum_i W_i^2}\right) = b + \frac{plim\left(\frac{1}{n}\sum_i W_ie_i\right)}{plim\left(\frac{1}{n}\sum_i W_i^2\right)}$$

  So OLS is consistent if $plim\left(\frac{1}{n}\sum_i W_ie_i\right) = 0$ and $plim\left(\frac{1}{n}\sum_i W_i^2\right) \neq 0$. Sufficient is if $E\left(W_ie_i\right) = 0$, $E\left(W_i^2\right) \neq 0$, and a law of large numbers holds for these averages.

3.5 Convergence in Distribution

- For a sequence $X_1, X_2, \ldots$ let $F_n(x)$ be the distribution function of X_n. So by definition, $F_n(x) = \Pr\left(X_n \leq x\right)$.

  Let Y be a random variable, with distribution function $F\left(y\right)$.

  Definition: $X_n \xrightarrow{d} Y$, or X_n converges in distribution to Y, if

$$\lim_{n\to\infty} F_n(y) = F(y)$$

  at every value of y where the function F is continuous.

- Note that the actual random variable Y is irrelevant to this definition. All the matters is F, the distribution function of Y. So, e.g., if $Y \sim N(0,1)$ (a standard normal), instead of saying $X_n \xrightarrow{d} Y$ we could equivalently say $X_n \xrightarrow{d} N(0,1)$

- Remark: if $X_n \xrightarrow{p} X$, then $X_n \xrightarrow{d} X$. The proof is an application of the continuous mapping theorem. Intuitively, $X_n \xrightarrow{p} Y$ or equivalently $X_n - Y \xrightarrow{p} 0$, means X_n gets closer and closer to being the same random variable as Y, while $X_n \xrightarrow{d} Y$ only says that the distribution function of X_n gets closer and closer to being the same distribution function that Y has.

- The Slutsky Theorem: For a constant c, a random variable X, and random sequences $X_1, X_2, ..$ and $Y_1, Y_2, ..$

  1. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then $X_n Y_n \xrightarrow{d} Xc$ and $X_n + Y_n \xrightarrow{d} X + c$
  2. If $X_n \xrightarrow{d} X$ and $X_n - Y_n \xrightarrow{p} 0$, then $Y_n \xrightarrow{d} X$.

  Note: There is also a completely different Slutsky theorem in microeconomics. Eugen Slutsky did both theorems - he was first an economist, and later worked in probability theory.

- The continuous mapping theorem also applies to convergence in distribution: If $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$.

3.6 Central Limit Theorem (Lindeberg-Levy CLT)

Assume $Z_1, Z_2, \ldots$ is an i.i.d. sequence of random variables. Assume $E(Z_i) = \mu$ and $Var(Z_i) = \sigma^2$ are finite. Let $X_n = \frac{1}{n} \sum_{i=1}^{n} Z_i$.

We showed that $E(X_n) = \mu$ and $Var(X_n) = \sigma^2/n$.
We we take an average, as n grows, the distribution of the average narrows (its variance shrinks).

Consider a new sequence of random variables $Y_1, Y_2, ...$ defined by $Y_n = \sqrt{n}(X_n - \mu)$.

Can check that Y_n are generally not independent or identically distributed, but

$$E(Y_n) = E\left(\sqrt{n}(X_n - \mu)\right) = \sqrt{n}(E(X_n) - \mu) = \sqrt{n}(\mu - \mu) = 0 \quad \text{and}$$

$$Var(Y_n) = Var\left(\sqrt{n}(X_n - \mu)\right) = \sqrt{n}^2 Var(X_n - \mu)) = nVar(X_n)) = n\sigma^2/n = \sigma^2$$

$Y_n = \sqrt{n}(X_n - \mu)$ is an example of a "stabilizing transformation." It converts the sequence of statistics $X_1, X_2 ...$ where each element has a different variance, into a new sequence $Y_1, Y_2, ..$ where each element has the same variance (and in this case, also the same mean zero).

What Y_n does is take X_n, recenter around zero by subtracting off μ, and scales it up by multiplying by $\sqrt{n}$. This scaling by $\sqrt{n}$ is exactly the amount needed to offset the shrinking of the variance of X_n as n grows. The distribution of X_n narrows as n grows - the stabilizing transformation stretches the distribution back out again,

We also showed $X_n \xrightarrow{ms} \mu$, because X_n is unbiased, and has a variance that shrinks to zero.

Unlike $X_1, X_2...$, the sequence $Y_1, Y_2, ..$ can't converge in mean square to a constant, because it's variance doesn't shrink as $n \to \infty$. However, $Y_1, Y_2, ..$ does converge in distribution, by the CLT:

Lindeberg-Levy Central Limit Theorem (CLT): Assume $Z_1, Z_2, \ldots$ is an i.i.d. sequence of random variables. Assume $E(Z_i) = \mu$ and $Var(Z_i) = \sigma^2$ are finite. Let $X_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$. Then

$$\sqrt{n}(X_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

We already knew that $\sqrt{n}(X_n - \mu)$ had mean zero and variance σ^2. What the CLT does is say that no matter what the shape of the distribution of each Z_i is, the shape of the distribution of Y_n gets closer and closer to normal as n grows to infinity.

The process of averaging (X_n) and then stretching the distibution back out again (Y_n) also smooths the resulting distribution, making it close to normal.

X_n is the estimator that we're interested in. But the limiting distribution of Y_n is what we know. We can use the limiting distribution of Y_n to approximate the distribution of X_n, as follows:

We have $X_n = Y_n/\sqrt{n} + \mu$, so X_n is a linear function of Y_n.
Linear functions of normals are normal.
If Y_n was actually normal, then X_n would be normal.
Since X_n has mean μ and variance σ^2/n, if X_n were normal, it would be $N\left(\mu, \frac{\sigma^2}{n}\right)$.
Since Y_n is approximately normal when n is large, X_n is approximately $N\left(\mu, \frac{\sigma^2}{n}\right)$.

The way we denote this is to say that, since $\sqrt{n}(X_n - \mu) \xrightarrow{d} N(0, \sigma^2)$, X_n is "asymptotically normal," denoted

$$X_n \overset{a}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

You can think of the a in $\overset{a}{\sim}$ as either meaning "asymptotic" or "approximate." The formal convergence result is $Y_n \xrightarrow{d} N(0, \sigma^2)$, but since we care about X_n and not Y_n, we use the corresponding asymptotic approximation: $X_n \overset{a}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$.

The Lindeberg-Levy CLT extends to vectors. Assume $Z_1, Z_2, \ldots$ is a sequence of i.i.d. random $J-$vectors and that $E(Z_i Z_i^{'})$ is finite. Then, letting $E(Z_i) = \mu$ (the J vector of means) and $Var(Z_i) = \Omega$ (the $J \times J$ variance matrix) then $\sqrt{n}(X_n - \mu) \xrightarrow{d} N(0, \Omega)$.

Remark on terminology: Suppose for some estimator $\hat{\theta}$ and some constant $b > 0$ we have $n^b(\hat{\theta} - \theta) \xrightarrow{d} N(0, V)$. This implies that $\hat{\theta}$ is consistent, and is asymptotically normal, with a "rate of convergence" n^b. The CLT theorem gives us a rate $n^{1/2}$, called root-n. In this case we say the sample average is "root-n CAN," which stands for "root-n consistent and asymptotically normal.

3.7 Delta Method

- The Delta method provides an alternative to the Slutsky Theorem for considering smooth functions of variables that converge in distribution. The Delta method says: Consider a sequence of random vectors X_1,X_2,... Assume for a constant $b > 0$, that $n^b(X_n - a) \xrightarrow{d} Y$, and we have a vector valued function g such that (i) $g(a)$ is continuously differentiable in a neighborhood of a, (ii) $dg(a)/da'$ is finite and non-zero and (iii) g is not a function of n, then

  $$n^b[g(X_n) - g(a)] \xrightarrow{d} \frac{dg(a)}{da'}Y$$

  Note if g is dimension p, and a is dimension q, then $\frac{dg(a)}{da'}$ is a $p \times q$ Jacobian matrix, and Y is dimension q.

- Example 1:

  Suppose $\sqrt{n}(X_n - a) \xrightarrow{d} N(0, \Omega)$. This might hold because of a CLT, for example. Then by the delta method, we have

  $$\sqrt{n}(g(X_n) - g(a)) \xrightarrow{d} N\left(0, \frac{dg(a)}{da'}\Omega\frac{dg(a)'}{da}\right)$$

  This says that if a stabilizing transformation of X_n is asymptotically normal, then smooth functions of X_n are also asymptotically normal. Combined with the CLT, we get that smooth functions of sample averages are asymptotically normal. This is convenient for econometrics, because many estimators, like OLS, are themselves smooth functions of averages.

- Example 2: Comparing the Delta Method to the Slutsky theorem:

  Suppose $\sqrt{n}(Z_n - \alpha) \xrightarrow{d} N(0, 1)$, and we consider the function $g(x) = x^2$. Then, since the square of a standard normal has a chi-squared distribution with one degree of freedom, we have by Slusky therorem

  $$(\sqrt{n}(Z_n - \alpha))^2 \xrightarrow{d} \chi^2_1$$

  In contrast, by the Delta method,

  $$\sqrt{n}(Z_n^2 - \alpha^2) \xrightarrow{d} N(0, 2\alpha \cdot 1 \cdot 2\alpha) = N(0, 4\alpha^2)$$

  How can the square of a normal be asymptotically normal? because of the stabilizing transformation. Z_n itself is converging to a constant. In the Slutsky case here, we are first applying the stabilizing transformation to Z_n, then squaring it. In the Delta method application here, we are squaring first, then applying the stabilizing transformation.

3.8 Central Limit Theorem (Lindeberg-Feller CLT)

- The Lindeberg-Feller CLT is for non-identically distributed data.

  Let $Z_1, Z_2, \ldots$ be a sequence of random variables that are independent but not necessarily identically distributed (this is denoted i.n.i.d). Let $E(Z_i) = \mu_i$, $Var(Z_i) = \sigma_i^2$, and define the following values:

$\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mu_i \quad \bar{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2$

Suppose that, for some $\eta > 2$, $E(|Z_i|^\eta)$ is bounded. That is there exists a finite constant c such that for some $\eta > 2$ we have $E(|Z_i|^\eta) < c$ for all i, then

$$\sqrt{n}\left(\frac{\bar{Z} - \bar{\mu}}{\bar{\sigma}}\right) \xrightarrow{d} N(0,1).$$

- The tail condition $E(|Z_i|^\eta) < c$ above is stronger than the finite variance condition needed for the Lindeberg-Levy CLT. In particular, it holds if third moments are bounded, $E\left(|Z_i|^3\right) < c$ for all i. Alternatively, it holds if the random variable Z_i is itself is bounded.

- Lindeberg Feller can also be extended to allow for a sequence of random vectors of $Z_1, Z_2, \ldots$ A sufficient condition in that case is that, for any three elements Z_{ij}, Z_{ik}, and Z_{il} of Z_i (including cases where two or all three elements are the same element), we have $E[|Z_{ij}||Z_{ik}||Z_{il}|] < c$ for all i.

- There exist many other CLT's that we will not cover here, including for observations that are not independent, for observations that change as n grows (called "double array CLT's), and for functions rather than vectors.

3.9 Little o and Little o_p notation

- Little o is used to provide information on the rates at which sequences are growing or shrinking.

  Let $c_1, c_2, \ldots$ be a sequence of constants. If $\lim\limits_{n\to\infty} c_n = 0$, then we write $c_n = o(1)$, "c_n is little o of one."

  By the definition of a limit, this says that for any $\varepsilon > 0$, there exists $N > 0$ such that $|c_n|$ $< \varepsilon$ for all $n > N$.

  For a given constant k, if $\frac{c_n}{n^k} = o(1)$, then we write $c_n = o(n^k)$. This is read as, "c_n is little o of n^k."

  If $k < 0$ then $c_n = o(n^k)$ means the sequence c_n is shrinking towards zero faster than n^k.

  If $k > 0$ then $c_n = o(n^k)$ means that the sequence c_n either is shrinking towards zero, or is converging to a constant, or is growing to infinity, but is doing so more slowly than n^k grows to infinity.

- Little o_p :

  Let $X_1, X_2, \ldots$ be a sequence of random variables. If $plim_{n\to\infty}X_n = 0$, then we write $X_n = o_p(1)$, "X_n is little o_p of one."

  By the definition of a probability limit, this says that for any $\delta > 0$ and any $\varepsilon > 0$, there exists an N such that

  $$\Pr(|X_n| > \varepsilon) < \delta$$

  holds for all $n > N$.

  $X_n = o_p(1)$ means that the probability that X_n lies outside a small neighborhood of zero goes to zero as $n \to \infty$.

For a given constant k, if $\frac{X_n}{n^k} = o_p(1)$, then we write $X_n = o_p(n^k)$, which means $plim_{n\to\infty}\frac{X_n}{n^k} = 0$.

So o_p is like, o but for sequences of random variables.

3.10 Bounded in Probability

Let $X_1, X_2, \ldots$ be a sequence of random variables. Suppose for any given small constant $\delta > 0$, there exists some (possibly very large) constant k_δ and N such that, for all $n > N$ we have

$$P(|X_n| > k_\delta) < \delta$$

Then the sequence $X_1, X_2, \ldots$ is said to be "bounded in probability."

Being bounded in probability basically means that the tails of the distribution of X_n are not getting arbitrarily fat as n goes to infinity.

Roughly ,if a sequence $X_1, X_2, \ldots$ is NOT bounded in probability, then the probability that $|X_n|$ is bigger than any large constant k grows as $n \to \infty$.

Examples:

1. If $X_n \xrightarrow{d} Y$ for some ordinary random variable Y, then X_n is bounded in probability.

2. Suppose X_n ~ $N(0,1)$ if n is even and X_n ~ $N(1,1)$ if n is odd. Then X_n does not converge in distribution to anything, but still X_n is bounded in probability.

3. Suppose X_n ~ $N(0,n)$. Then X_n is not bounded in probability.

3.11 Big O and big O_p

- Big O: Let $c_1, c_2, \ldots$ be a sequence of constants.

  If there exists a constant $c > 0$ such that $\lim_{n\to\infty} \sup |c_n| \leq c$, meaning that $|c_n|$ is bounded by c, then we write $c_n = O(1)$, which you read as: "c_n is big O of one."

  We say $c_n = O(n^k)$ if $\frac{c_n}{n^k} = O(1)$.

- Big O_p: Let $X_1, X_2, \ldots$ be a sequence of random variables.

- We write $X_n = O_p(1)$, "c_n is big O_p of one," if the sequence is bounded in probability.

- $X_n = O_p(n^k)$ if $\frac{X_n}{n^k} = O_p(1)$.

  USE OF BIG AND LITTLE O:

  Adding a constant: For any constant λ, if $c_n = o(1)$ then $c_n + \lambda = O(1)$. If $X_n = o_p(1)$ then $X_n + \lambda = O_p(1)$.

  Notation: We often will just use O_p or o_p in place of a sequence that has that property. So, e.g., instead of writing something like $Y_n = X_n + S_n$ where $S_n = o_p\left(n^{1/2}\right)$, we might instead just write $Y_n = X_n + o_p\left(n^{1/2}\right)$

3.12 Asymptotic Linearity

- Let $Z_1, Z_2, ...$ be a sequence of random vectors. Consider a sequence of estimators $\widehat{\theta} = f_n(Z_1, ..., Z_n)$, for some functions f_n. Let θ_0 denote the unknown true value of θ.

- The estimator $\widehat{\theta}$ is defined to be "asymptotically linear" if there exists sequences of random vectors $R_1, R_2...$ and $S_1, S_2...$ such that

$$\sqrt{n}(\widehat{\theta} - \theta_0) = \sqrt{n}\bar{R} + S_n$$

where $\bar{R} = \frac{1}{n}\sum_i R_i$, $S_n \xrightarrow{p} 0$, and $\sqrt{n}\bar{R} \xrightarrow{d} N(0, \Omega)$ for some variance matrix Ω. Note this last condition says that the sample average $\bar{R}$ satisfies a Central Limit Theorem.

Theorem: if $\widehat{\theta}$ is asymptotically linear, then $\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0, \Omega)$, so $\widehat{\theta}$ is root-n CAN.

Proof: we had a theorem that said if $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then $X_n + Y_n \xrightarrow{d} X + c$. Apply this result with $X_n = \sqrt{n}\bar{R}$, $Y_n = S_n$, and $c = 0$.

Notes:

- Asymptotic linearity says that $\widehat{\theta}$ is almost the same as an average, and so $\widehat{\theta} - \theta_0$ has the same limiting distribution as the average $\bar{R}$.

- The sequence $S_1, S_2...$ is said to be "asymptotically neglible" here, since it has no effect on the limiting distribution of $\widehat{\theta}$. Note since $S_n \xrightarrow{p} 0$, two other equivalent ways to write $\sqrt{n}(\widehat{\theta} - \theta_0) = \sqrt{n}\bar{R} + S_n$ are $\sqrt{n}(\widehat{\theta} - \theta_0) = \sqrt{n}\bar{R} + o_p(1)$, and $\widehat{\theta} - \theta_0 = \bar{R} + o_p\left(n^{1/2}\right)$

- Each R_i must be a function of the data and of θ_0, that is, $R_i = g_i(Z_1, ..., Z_i, \theta_0)$, for some set of function g_i. R_i is call the "influence function."

- Many estimators in econometrics are asymptotically linear. A common method of deriving the asymptotic distribution of estimators is to show they are asymptotically linear, calculate their influence function, and then apply a central limit theorem to the influence function to obtain the estimator's root-n CAN limiting distribution.

- Once we have shown that $\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0, \Omega)$, to apply this result we also need to estimate Ω. Suppose we can find an estimator $\widehat{\Omega} \xrightarrow{p} \Omega$. Then we can say $\widehat{\theta} \overset{a}{\sim} N\left(\theta_0, \frac{\widehat{\Omega}}{n}\right)$ and use this to construct standard errors, confidence intervals, and hypothesis tests. In particular, standard errors are the square root of the diagonal of $\widehat{\Omega}/n$. If we can calculate the influence function, then the typical way to construct the estimate $\widehat{\Omega}$ would be to let $\widehat{R}_i = g_i(Z_1, ..., Z_i, \widehat{\theta})$ and let $\widehat{\Omega}$ equal the sample variance of $\widehat{R}_1$,...,$\widehat{R}_n$. We will see many examples of this procedure later.

3.13 Standard consistency conditions

- Let θ_0 denote the unknown true value of a parameter vector θ. We already gave one way to prove consistency of an estimator $\hat{\theta}$, which is to show convergence in mean square. But that required being able to directly calculate the mean and variance of $\hat{\theta}$.

- Many estimators in econometrics can be written as the maximum value of some function. For example, maximum likelihood estimation. Also, ordinary least squares maximizes the negative of the sum of squared errors. Here we we give a general theorem for proving consistency of such estimators, based on properties of the function that is being maximized.

- $\hat{\theta}$ is defined to be an "extremum estimator" for θ if $\hat{\theta}$ maximizes some function $Q_n(\theta)$, that is,
$$\hat{\theta} = \arg\max_{\theta} Q_n(\theta)$$
Note that Q_n in general depends on data, and so is a random function, meaning that for any value that you plug in for θ, $Q_n(\theta)$ is a random variable. Denote $Q_0(\theta) = plim Q_n(\theta)$, assume:

  1. Identification: $Q_0(\theta)$ exists and is maximized only at $\theta = \theta_0$.
  2. Smoothness: $Q_n(\theta)$ is differentiable in θ.
  3. Compactness: Θ, the set of possible values of θ, is compact.
  4. Stochastic-equicontinuity. The formal definition is complicated (see, e.g., Newey 1991), but a sufficient condition for stochastic-equicontinuity is that
  $$\sup_{\theta\in\Theta}\left|\frac{\partial Q_n(\theta)}{\partial\theta}\right|$$
  is bounded in probability, where the constants k_δ and N in the definition of boundedness in probability don't depend on θ.

     Intuition: To find $\hat{\theta}$, we maximize $Q_n(\theta)$ by setting $\frac{\partial Q_n(\theta)}{\partial\theta} = 0$. Stochastic equicontinuity ensures that this random function is well behaved, e.g., the tails of the distribution of this derivative don't get arbitrarily fat as n grows.

     Formally, stochastic-equicontinuity along with other assumptions, ensures "uniform convergence", meaning
  $$\sup_{\theta\in\Theta}|Q_n(\theta) - Q_0(\theta)| \xrightarrow{p} 0$$
  Assumption 1 includes "pointwise convergence" which is that $|Q_n(\theta) - Q_0(\theta)| \xrightarrow{p} 0$ for each $\theta \in \Theta$.

     What's the difference between pointwise and uniform convergence? Define $g(\theta) = |Q_n(\theta) - Q_0(\theta)|$. Consider a sequence of values $\theta_1, \theta_2, ...$. Pointwise convergence says that each element of the corresponding sequence, $g(\theta_1), g(\theta_2), ...$ goes to zero. Uniform convergence would require that the limit of the sequence $g(\theta_1), g(\theta_2), ...$ also goes to zero.

Theorem: If the conditions 1 to 4 hold, then $\hat{\theta} \xrightarrow{p} \theta_0$.

- Example: Consistency of OLS.

  Suppose $Y_i = b_0 W_i + e_i$ where $|b_0| \leq 1,000,000$. Assume i.i.d. observations of the vector $(Y_i, W_i)'$, $E(e_i W_i) = 0$, Y_i and W_i are bounded, and $E\left(W_i^2\right)$ is nonzero.

Define

$$Q_n(b) = -\frac{1}{n}\sum\nolimits_{i=1}^{n}(Y_i - bW_i)^2$$
$$\hat{b} = \arg\max Q_n(b)$$

This $\hat{b}$ is the OLS estimator. If the conditions of the consistency theorem are met, then $\hat{b} \xrightarrow{p} b_0$. We will now check each of the conditions.

1. Identification: Boundedness of Y_i, W_i and b means that $E\left[(Y_i - bW_i)^4\right]$ exists for any possible value of b (recall b is also bounded). So for any possible value of b, $Q_n(b) \xrightarrow{ms} Q_0(b)$ where $Q_0(b) = -E[(Y - bW)^2]$ which exists and is finite. This also means $Q_0(b) = plimQ_n(b)$. We have

   $$\begin{aligned} Q_0(b) &= -E[(Y-bW)^2] \\ &= -E(Y^2 - 2bWY + b^2W^2) \\ &= -b^2E(W^2) + 2bE(WY) - E(Y^2) \end{aligned}$$

   We now need $Q_0(b)$ to be maximized only at the value $b = b_0$. The maximizing value is given by the first order conditions

   $$\begin{aligned} \frac{\partial Q_0(b)}{\partial b} = 0 &\Rightarrow -2bE(W^2) + 2E(WY) = 0 \\ &\Rightarrow bE(W^2) = E(WY) = E(W\,(b_0W + e)) \\ &\Rightarrow bE(W^2) = b_0E(W^2) + E(We) \\ &\Rightarrow b = b_0 \end{aligned}$$

   and the second order condition is

   $$\frac{\partial^2 Q_0(b)}{\partial b^2} = -2E(W^2) < 0$$

   So the only value of b that maximizes $Q_0(b)$ is b_0.
2. Smoothness: $Q_n(b)$ is differentiable in b because it's quadratic.

   $$\frac{\partial Q_n(b)}{\partial b} = \frac{1}{n}\sum\nolimits_{i=1}^{n}(-2W_iY_i + 2bW_i^2)$$
3. Compactness: follows by assumption that $|b_0| \leq 1,000,000$ (any set of values that is a closed interval is compact).
4. Stochastic-equicontinuity: Now

   $$\begin{aligned} \sup_{|b|<1,000,000}\left|\frac{dQ_n(b)}{db}\right| &= \sup_{|b|\leq 1,000,000}\left|\frac{1}{n}\sum\nolimits_{i=1}^{n}(-2W_iY_i + 2bW_i^2)\right| \\ &\leq \frac{1}{n}\left[\sum\nolimits_{i=1}^{n}|(-2W_iY_i)| + \left|2,000,000W_i^2)\right|\right] \\ &\leq \sup|-2WY| + 2,000,000\sup W_i^2 \end{aligned}$$

   Since this is bounded, it must also be bounded in probability. So Stochastic Equicontinuity holds.

We have shown that all four condition needed for consistency hold, and therefore $\hat{b} \xrightarrow{p} b_0$.

- In this example, we didn't actually need the theorem. We could have instead shown consistency $\hat{b} \xrightarrow{p} b_0$ directly by solving for $\hat{b}$, then applying the rule that the plim of a function is the function of the plims, and then applying the LLN, as follows:

$$\begin{aligned} plim\left(\hat{b}\right) &= plim\left(\frac{\frac{1}{n}\sum_{i=1}^{n} W_i Y_i}{\frac{1}{n}\sum_{i=1}^{n} W_i^2}\right) = plim\left(b_0 + \frac{\frac{1}{n}\sum_{i=1}^{n} W_i e_i}{\frac{1}{n}\sum_{i=1}^{n} W_i^2}\right) = b_0 + \frac{plim\left(\frac{1}{n}\sum_{i=1}^{n} W_i e_i\right)}{plim\left(\frac{1}{n}\sum_{i=1}^{n} W_i^2\right)} \\ &= b_0 + \frac{E\left(We\right)}{E\left(W^2\right)} = b_0 + \frac{0}{E\left(W^2\right)} = b_0 \end{aligned}$$

  But this simpler way of proving consistency was only possible because we could explicitly, analytically solve for $\hat{b} = \arg\max Q_n(b)$. The theorem is mainly useful for applications where the function $Q_n(b)$ is too complicated for us to solve for $\hat{b}$ analytically. These are problems where we would instead have a computer search numerically for $\hat{b}$. Examples are many maximum likelihood estimators and nonlinear least squares.

3.14 Glivenko-Cantelli theorem

Let $Z_1, Z_2, ...$ be iid random variable with cumulative distribution function (cdf) $F(z)$. Recall that, by definition, $F(z)$ is just the probability that any Z_i is less than or equal to the value z. Suppose $F(z)$ is unknown and we want to estimate it. The empirical distribution function $\hat{F}_n(z)$ is defined by

$$\hat{F}_n(z) = \frac{1}{n}\sum\nolimits_{i=1}^{n} I(Z_i \leq z)$$

where $I(Z_i \leq z)$ is defined to be the indicator function that equals 1 if its argument, $Z_i \leq z$, is true and zero if its false. The function $\sum_{i=1}^{n} I(Z_i \leq z)$ therefore just equals the number of observations i that have Z_i less than or equal to z. So the estimator $\hat{F}_n(z)$ just equals the fraction of our sample that is less than or equal to z, which is the obvious way to estimate $F\left(z\right)$.

Since $\hat{F}_n(z)$ is a sample average, we can apply the law of large numbers to it. You can use results we already have to prove it converges in mean square, and therefore that $\hat{F}_n(z)$ is consistent, so $\hat{F}_n(z) \xrightarrow{p} F(z)$.

The Glivenko-Cantelli theorem says that $\hat{F}_n(z)$ is "uniformly consistent", that is, we have the uniform convergence

$$\sup_z \; |\hat{F}_n(z) - F(z)| \; \xrightarrow{p} 0.$$

4 Lecture 04. Classical Regression Model

Readings for this lecture is: Greene Chapter 4.

4.1 OLS Estimation

We can write the least squares model three ways:

$$Y_i = X_{i1}\beta_1 + X_{i2}\beta_2 + ...X_{iK}\beta_K + e_i$$

or vector form:

$$Y_i = X_i'\beta + e_i$$

or matrix form:

$$Y = \mathbf{X}\beta + e$$

To write it in matrix form, we have

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{11} & ... & X_{K1} \\ \vdots & & \vdots \\ X_{1n} & ... & X_{Kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$$

The rows of Y, $\mathbf{X}$ and e are observations, columns of X are different variables. Row i of the matrix $\mathbf{X}$ is the transpose of the vector X_i. Usually, the first column of $\mathbf{X}$, (the elements X_{1i} for $i = 1, ..., n$) are onese, making β_1 be the constant term in the regression.

- The OLS estimator $\hat{\beta}$ is defined as the value of β that minimizes the sum of squared errors:

$$\begin{aligned} \hat{\beta} &= \arg\min_{\beta} \sum\nolimits_{i=1}^{n} e_i^2 = \arg\min_{\beta} e'e \\ &= \arg\min_{\beta} (Y - \mathbf{X}\beta)' (Y - \mathbf{X}\beta) \\ &= \arg\min_{\beta} \left(Y' - (\mathbf{X}\beta)'\right) (Y - \mathbf{X}\beta) \\ &= \arg\min_{\beta} (Y' - \beta'\mathbf{X}') (Y - \mathbf{X}\beta) \\ &= \arg\min_{\beta} (Y'Y - \beta'\mathbf{X}'Y - Y'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta) \\ &= \arg\min_{\beta} (Y'Y - 2\beta'\mathbf{X}'Y + \beta'\mathbf{X}'\mathbf{X}\beta) \end{aligned}$$

  To minimize this expression, we take its derivative with respect to β, and $\hat{\beta}$ is the value of β that makes this derivative equal to zero. Note that expression (the sum of squared errors) is a scalar, and we are taking its derivative with respect to a $K-$ vector β, so this is really a gradient. We then set the resulting $K-$ vector of derivatives equal to a $K-$ vector of zeros. Solving for $\hat{\beta}$ is really solving K equations for K unknowns:

$$\begin{aligned} 0 &= \frac{\partial \left(Y'Y - 2\hat{\beta}'\mathbf{X}'Y + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta}\right)}{\partial \beta} \\ &= 0 - 2\mathbf{X}'Y + 2\mathbf{X}'\mathbf{X}\hat{\beta} \\ &\Rightarrow \mathbf{X}'Y = \mathbf{X}'\mathbf{X}\hat{\beta} \\ &\Rightarrow \hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y \end{aligned}$$

This solution assumes $\mathbf{X}'\mathbf{X}$ is nonsingular, so it can be inverted. Note that to confirm this is really the minimum (as opposed to a maximum, point of inflection, or saddle point), we can check the second order conditions.

Note by plugging in for Y we get another expression for $\hat{\beta}$:

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{X}\beta + e\right) = \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'e$$

In the special case where $Y_i = a + bZ_i + e_i$ (so $K = 2$, $X_{1i} = 1$, $X_{2i} = Z_i$, $\beta_1 = a$, and $\beta_2 = b$) the above simplifies to:

$$\begin{aligned} \hat{b} &= \frac{\sum_{i=1}^{n}(Z_i - \overline{Z})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(Z_i - \overline{Z})^2} = \frac{\widehat{Cov(Z,Y)}}{\widehat{Var(X)}} \\ \hat{a} &= \overline{Y} - \hat{b}\overline{Z} \end{aligned}$$

- The Classical assumptions, called Gauss-Markov (GM) assumptions:

  1. True model is linear: $Y = \mathbf{X}\beta_0 + e$
  2. $\mathbf{X}$ is constant
  3. $\mathbf{X}'\mathbf{X}$ is non-singular (equivalently, $n \geq K$ and $rank(\mathbf{X}) = K$)
  4. e is random, $E(e) = 0$ and $Var(e) = \sigma^2 I_n$ where I_n is the $n \times n$ identity matrix, which has ones on the diagonal and zeros everywhere else.

     This is equivalent to: $E(e_i) = 0$, $E(e_i^2) = var\,(e_i) = \sigma^2$, and $E(e_ie_j) = cov(e_ie_j) = 0$ for all $i \neq j$.

- Interpretation of and notes on the GM assumptions:

  1. The model being linear, with true coefficient vector β_0, by itself is not a restriction, since for any given Y, $\mathbf{X}$, β_0 could define e by $e = Y - \mathbf{X}\beta_0$. So this 'model is linear' assumption only is a restriction when combined with the other assumptions.
  2. When we do statistics, we think of our sample as being one possible draw from the underlying population. The distribution of a statistic or estimator is based on imagining that we could draw many samples, and calculate our statistic or estimator separately in equal sample. So for example $E\left(\hat{\beta}\right)$ is defined to be the average value of $\hat{\beta}$ when calculated separately for each of those infinitely many hypothetical samples (each of size n), and averaged across all those samples.

     The assumption that $\mathbf{X}$ is constant means that, for each one of those hypothetical samples, the matrix $\mathbf{X}$ stays the same as the one we actually observed, so only e and Y would change.

     Saying $\mathbf{X}$ is constant is equivalent to saying that we are conditioning on that value of $\mathbf{X}$. So for example it means that $E\left(W\right)$ where W is any function of our data is the same as the conditional expectation $E\left(W \mid X\right)$.
  3. Having $rank(\mathbf{X}) = K$ means that none of our regressor variables X_{ki} is a perfect linear function of the other regressors.

4. Condition 4, when combined with condition 2, implies that $E(e_i \mid X) = 0$, $E(e_i^2 \mid X) = \sigma^2$ the same for all i (homoscedasticity) and $E(e_i e_j \mid X) = 0$ for all $i \neq j$ (no autocorrelation).

- Other implications of the Classical GM assumptions:

$$\begin{aligned}
E\left(\hat{\beta}\right) &= E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{X}\beta + e\right)\right] = E\left[\beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'e\right] \\
&= \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'E\left(e\right) = \beta
\end{aligned}$$

$$\begin{aligned}
var\left(\hat{\beta}\right) &= E\left[\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)'\right] = E\left[\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'e\right)\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'e\right)'\right] \\
&= E\left[\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'e\right)\left(e\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1\prime}\right)\right] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'E\left(ee'\right)\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1\prime} \\
&= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\sigma^2 I\right)\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1\prime} \\
&= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}
\end{aligned}$$

4.2 Gauss Markov Theorem

- Theorem

  Under the Gauss-Markov assumption, the OLS estimator $\hat{\beta}$ is BLUE (Best Linear Unbiased Estimator).

- The word linear here is NOT referring to the model being linear, it's saying that the estimator is linear. An estimator is a function of data, so an estimator must be a function of $\mathbf{X}$ and Y. A "linear" estimator is defined to be an estimator that is linear in random variables. Under the GM assumptions $\mathbf{X}$ is constant, so the random variables are Y. So a linear estimator is any estimator $\tilde{\beta}$ that is linear in Y, i.e., $\tilde{\beta} = \mathbf{C}Y$ for some matrix of constants $\mathbf{C}$. Since $\mathbf{X}$ is constant, the matrix $\mathbf{C}$ can be any function of $\mathbf{X}$. Since β is a K vector and Y is an n vector, the matrix $\mathbf{C}$ must by $K \times n$.

- The "best" here means most efficient, i.e., lowest variance. So what the theorem, "$\hat{\beta}$ is BLUE" means is that the OLS estimator $\hat{\beta}$ is an unbiased and linear estimator, and that among all possible linear, unbiased estimators, $\hat{\beta}$ is most efficient.

- It is possible that more efficient estimators could exist if the GM assumptions don't hold, and it is possible that even if the GM assumptions do hold, more efficient estimators could exist that are NOT linear estimators. Examples are GLS (generalized least squares) estimators when GM assumptions are violated, and ARCH (auto-regressive conditionally heteroscedastic) estimators that are not linear estimators.

- Proof of the Gauss Markov Theorem

  1. Let $\tilde{\beta} = \mathbf{C}Y$ where $\mathbf{C}$ is some constant matrix. So $\tilde{\beta}$ is any linear estimator.

2. First note that

$$\begin{aligned}E\left(\tilde{\beta}\right) &= E\left(\mathbf{C}Y\right)\\ &= \mathbf{C}E\left(\mathbf{X}\beta_0 + e\right)\\ &= \mathbf{C}E\left(\mathbf{X}\beta_0\right) + \mathbf{C}E\left(e\right)\\ &= \mathbf{C}\mathbf{X}\beta_0\end{aligned}$$

but unbiasedness means $E\left(\tilde{\beta}\right) = \beta_0$. So (since each element of β_0 might be nonzero) a linear estimator $\tilde{\beta}$ is unbiased if and only if

$$\mathbf{CX} = I_K$$

where I_K is the $K \times K$ identity matrix.

3. The best would mean variance is lowest among other linear unbiased estimators. What is the variance of $\tilde{\beta}$?

$$\begin{aligned}Var\left(\tilde{\beta}\right) &= Var\left(\mathbf{C}Y\right) = Var\left(\mathbf{CX}\beta_0 + \mathbf{C}e\right)\\ &= Var\left(\mathbf{C}e\right) = E\left[\mathbf{C}e\left(\mathbf{C}e\right)'\right]\\ &= \mathbf{C}E\left(ee'\right)\mathbf{C}' = \mathbf{C}\sigma^2 I_n \mathbf{C}'\\ &= \sigma^2\mathbf{CC}'\end{aligned}$$

4. Now consider OLS: $\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'Y$. This is a linear estimator since it equals $\mathbf{CY}$ with $\mathbf{C}$ equal to $\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$. OLS is unbiased, because, for this choice of $\mathbf{C}$, we have $\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X} = I_K$. And for this choice of $\mathbf{C}$ we get variance

$$\begin{aligned}Var\left(\hat{\beta}\right) &= \sigma^2\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]'\\ &= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\\ &= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\end{aligned}$$

5. So far, we have shown that OLS is LUE (a linear, unbiased estimator), and we have calculated the variance of any linear estimator, and of OLS in particular. To finish proving the theorem, we need to show that, among all linear unbiased estimators $\tilde{\beta} = \mathbf{C}Y$, the one with the lowest variance is $\hat{\beta}$.

6. To do this last step, define the matrix $\mathbf{D}$ by $\mathbf{D} = \mathbf{C} - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$. Observe that, since both $\tilde{\beta}$ and $\hat{\beta}$ are unbiased, $\mathbf{DX} = 0$. Now by construction $\mathbf{C} = \mathbf{D} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ so

$$\begin{aligned}Var\left(\tilde{\beta}\right) &= \sigma^2\left[\mathbf{D} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\left[\mathbf{D} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]'\\ &= \sigma^2\left[\mathbf{D} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\left[\mathbf{D}' + \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right]\\ &= \sigma^2\left[\mathbf{DD}' + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}' + \mathbf{DX}\left(\mathbf{X}'\mathbf{X}\right)^{-1} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right]\\ &= \sigma^2\left[\mathbf{DD}' + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\right] \geq \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} = Var\left(\hat{\beta}\right)\end{aligned}$$

The terms $\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}' + \mathbf{DX}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ are zero because $\mathbf{DX} = 0$, and the inequality holds because for any matrix $\mathbf{D}$, the matrix $\mathbf{DD}'$ is positive semidefinite.

7. Suppose $K = 1$. Then another way to prove the GM theorem is as follows:

$$\begin{aligned}
Y_i &= X_i b_0 + e_i \\
\tilde{b} &= \sum\nolimits_{i=1}^{n} C_i Y_i \\
E(\tilde{b}) &= \sum\nolimits_{i=1}^{n} C_i X_i b_0 \\
Var(\tilde{b}) &= \sum\nolimits_{i=1}^{n} C_i^2 \sigma^2
\end{aligned}$$

Unbiasedness requires: $\sum_i C_i X_i = 1$. To find the BLUE estimator, we can search for the constants C_1,...,C_n that minimize $Var(\tilde{b})$, subject to the unbiasedness constraint. To minimize the variance given the constriant, set up the lagrangian:

$$\begin{aligned}
\mathcal{L} &= \sum\nolimits_{i=1}^{n} C_i^2 \sigma^2 - \lambda(\sum\nolimits_{i=1}^{n} C_i X_i - 1) \\
\frac{\partial \mathcal{L}}{\partial C_i} &= 0 = 2C_i \sigma^2 - \lambda X_i \\
C_i &= \frac{\lambda X_i}{2\sigma^2}
\end{aligned}$$

Plugging in $\sum_{i=1}^{n} C_i X_i - 1 = 0$

$$\begin{aligned}
\sum\nolimits_{i=1}^{n} \frac{\lambda X_i}{2\sigma^2} X_i - 1 = 0 \\
C_i = \frac{X_i}{\sum_{i=1}^{n} X_i^2}
\end{aligned}$$

and this variance minimizing choice of C_i gives the OLS estimator when $K = 1$:

$$\hat{\beta} = \sum\nolimits_{i=1}^{n} \frac{X_i}{(\sum_{i=1}^{n} X_i^2)} Y_i = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}.$$

4.3 Variance

Suppose the model is $Y_i = a + bX_i + e_i$. Then $Y = \mathbf{X}\beta + e$ with

$$\mathbf{X} = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}, \quad \beta = \begin{pmatrix} a \\ b \end{pmatrix}$$

$$\begin{aligned}
\mathbf{X'X} &= \begin{pmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{pmatrix} \\
(\mathbf{X'X})^{-1} &= \frac{1}{n\sum_{i=1}^{n}(X_i - \overline{X})^2} \begin{pmatrix} \sum_{i=1}^{n} X_i^2 & -\sum_{i=1}^{n} X_i \\ -\sum_{i=1}^{n} X_i & n \end{pmatrix}
\end{aligned}$$

$$\text{We showed } Var\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \sigma^2 (\mathbf{X'X})^{-1}, \text{ so:}$$

$$\begin{aligned}
Var\left(\hat{b}\right) &= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
Var\left(\hat{a}\right) &= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \left(\frac{\sum_{i=1}^n X_i^2}{n}\right) \\
Cov\left(\hat{a}, \hat{b}\right) &= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \left(-\overline{X}\right)
\end{aligned}$$

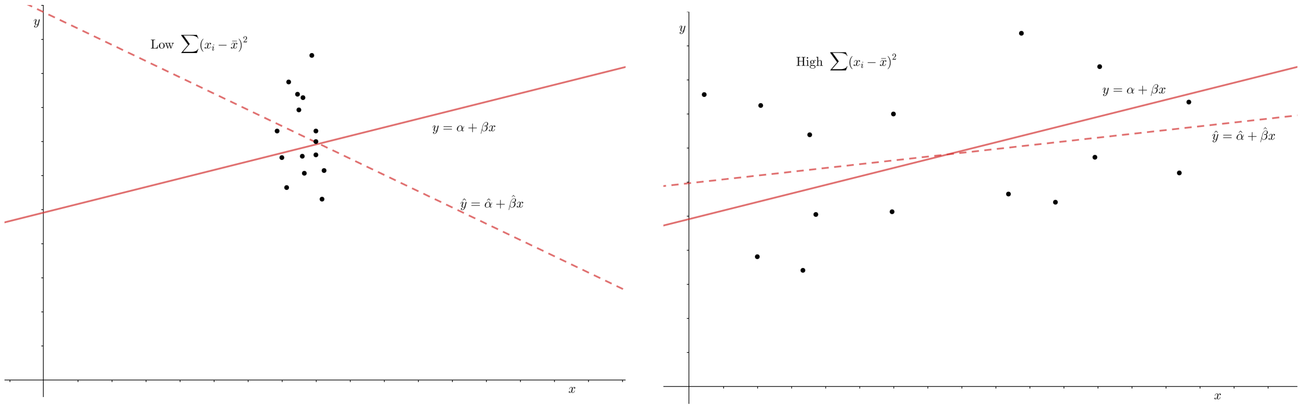
The lower are these variances, the more accurate the estimates are likely to be. We can examine each component of these variances.

1. The smaller the numerator σ^2 is (which means the smaller the magnitudes of e tend to be, and hence the closer the data points are to the true line $a + bX_i$), then the smaller $var(\hat{a})$, $var(\hat{b})$, and $Cov\left(\hat{a}, \hat{b}\right)$ will be.

2. The greater is the number of observations n, the larger is the denominator $\sum_{i=1}^n (X_i - \overline{X})^2$, and hence the smaller $var(\hat{a})$, $var(\hat{b})$, and $Cov\left(\hat{a}, \hat{b}\right)$ will be.

3. The further each X_i is from $\overline{X}$ on average (meaning the larger is the sample variance of X, i.e., the more spread out the X data are), the larger is each term in the denominator $\sum_{i=1}^n (X_i - \overline{X})^2$, and hence the smaller $var(\hat{a})$, $var(\hat{b})$, and $Cov\left(\hat{a}, \hat{b}\right)$ will be.

   Summarizing: what makes estimates more accurate is smaller errors e, large sample size n, and spread out values of X_i. It's obvious what small errors and large sample size help. The figure below shows why spread of X values helps: you can draw a line more accurately if you see points near the line that that are far apart from each other.



1. Now consider the additional components of the above variance formulas that just affect $var(\hat{a})$ and $Cov\left(\hat{a}, \hat{b}\right)$:

2. The smaller the magnitudes of each X_i (meaning the closer they are to zero), the smaller is $var(\hat{a})$ and $Cov\left(\hat{a}, \hat{b}\right)$. This is because we can more accurately estimate the intercept $\hat{a}$ on the vertical axis if our data are closer to that axis.

3. If $\overline{X} > 0$ then $cov(\hat{a}, \hat{b}) < 0$. This negative covariance means that if b is overestimated, then a is likely to be underestimated. This is because, if $\overline{X} > 0$, then the data is mostly to the right of the vertical axis, and in that region, if you draw a line through the middle of the data that is too steep (meaning $\hat{b}$ is too big), it will intersect the vertical axis at a point that is too low (meaning $\hat{a}$ is too small), and vice versa. Remember that the estimated regression line always goes through the middle of the data, that is, the point $\left(\overline{X}, \overline{Y}\right)$.

4.4 Residuals and Standard Errors

We would like to estimate the $K \times K$ matrix $Var\left(\hat{\beta}\right)$, to have a sense of how precise our estimates are. Under the GM assumptions,

$$Var\left(\hat{\beta}\right) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

and we observe $\mathbf{X}$ and so can construct $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$. So to estimate $Var\left(\hat{\beta}\right)$, the only other thing we need is an estimate of σ^2. Now $\sigma^2 = Var\left(e_i\right)$ for all observations i. So we need an estimate for the variance of the errors.

The residual vector $\hat{e}$ is defined as the estimated error vector, that is, $e = Y - \mathbf{X}\boldsymbol{\beta}$ and $\hat{e}$ is given by

$$\hat{e} = Y - \mathbf{X}\hat{\beta}$$

The way we usually estimate σ^2 is by s^2 defined by

$$s^2 = \frac{1}{n-K}\sum\nolimits_{i=1}^{n} \hat{e}_i^2 = \frac{\hat{e}'\hat{e}}{n-K} = \frac{\left(Y - \mathbf{X}\hat{\beta}\right)'\left(Y - \mathbf{X}\hat{\beta}\right)}{n-K}$$

The way we usually estimate a variance is to divide by $n-1$, but in this case we divide by the "degrees of freedom" $n-K$. The usual argument for why we divide by $n-K$ is because we used the n data points to estimate K parameters (the elements of $\hat{\beta}$). But that's not a real explanation. The real reason we divide by $n-K$ is that makes s^2 and unbiased estimator of σ^2, that is, $E\left(s^2\right) = \sigma^2$. If we had divided by n or $n-1$, it would be biased. At the end of this subsection, we prove that $E\left(s^2\right) = \sigma^2$.

The "Standard Error of the Regression," denoted by s, and sometimes abbreviated either SER or just SE, is defined by

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-K}\sum\nolimits_{i=1}^{n} \hat{e}_i^2}$$

Recall our goal was to estimate $Var\left(\hat{\beta}\right)$. Since X is constant, we now have an unbiased estimator:

$$\widehat{Var\left(\hat{\beta}\right)} = s^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} \text{ and } E\left[\widehat{Var\left(\hat{\beta}\right)}\right] = Var\left(\hat{\beta}\right)$$

Don't get confused: $Var\left(\hat{\beta}\right)$ is the true variance of the estimator $\hat{\beta}$, while $\widehat{Var\left(\hat{\beta}\right)}$ is the estimator of the true variance of the estimator $\hat{\beta}$.

The k'th element on the diagonal of $Var\left(\hat{\beta}\right)$ is $Var\left(\hat{\beta}_k\right)$, the variance of the k'th element of the vector $\hat{\beta}$. So the square root of $Var\left(\hat{\beta}_k\right)$ is the standard deviation of the estimated coefficient

$\hat{\beta}_k$. The standard error of $\hat{\beta}_k$ is defined to be the estimate of this standard deviation, that is,

$$\text{The standard error of } \hat{\beta}_k \text{ is } \widehat{stddev}\left(\hat{\beta}_k\right) = \sqrt{\widehat{Var}\left(\hat{\beta}_k\right)}.$$

This is as far as we can get with just the Gauss-Markov assumptions. To summarize, we've shown that the OLS estimator $\hat{\beta}$ is BLUE, we calculated its variance, and we can construct an unbiased estimator of the variance of $\hat{\beta}$. To make any additional statistical statements (and in particular, to construct confidence intervals or do hypothesis tests, we'll need to make more assumptions.

To end this section, here the proof that $E\left(s^2\right) = \sigma^2$, so s^2 is an unbiased estimate of σ^2. First, some preliminaries

Let $\mathbf{M} = I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Observe that

$$\mathbf{MX} = (I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{X} = \mathbf{X} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{X} - \mathbf{X}I = 0$$

and

$$\begin{aligned}\hat{e} &= Y - \mathbf{X}\hat{\beta} = Y - \mathbf{X}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y) = (I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')Y = \mathbf{M}Y \\ &= \mathbf{M}\left(\mathbf{X}\beta + e\right) = \mathbf{MX}\beta + \mathbf{M}e = 0\beta + \mathbf{M}e = \mathbf{M}e\end{aligned}$$

A couple more things:

1. The matrix $\mathbf{M}$ is *idempotent*, which means $\mathbf{M}^2 = \mathbf{M}$ and $\mathbf{M}' = \mathbf{M}$ (you can check these).

2. Under the GM assumptions, X is a matrix of constants, and therefore $\mathbf{M}$ is also a matrix of constants.

3. The *trace* of a square matrix A, denoted $tr\left(A\right)$, is the sum of the elements on the diagonal of A. Some relevant properties of the trace are:

For random square A, $E\left[tr\left(A\right)\right] = tr\left[E\left(A\right)\right]$.

For $j \times j$ square matrices A and B, $tr\left(A + B\right) = tr\left(A\right) + tr\left(B\right)$.

If matrix A is $j \times k$ and B is $k \times j$, then $tr(AB) = tr(BA)$.

The trace of a scalar (i.e., a 1×1 matrix) just equals that scalar.

4. In the derivation below, $e'\mathbf{M}e$ is a scalar, A is e', and B is $\mathbf{M}e$. We then use this trace of a product rule again with A being X and B being $\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$.

$$\begin{aligned}E(\hat{e}'\hat{e}) &= E(e^{'}\mathbf{M}'\mathbf{M}e) = E(e'\mathbf{M}e) \\ &= E\left[tr(e'\mathbf{M}e)\right] = E\left[tr(\mathbf{M}ee')\right] \\ &= tr\left[\mathbf{M}\left(E(ee')\right)\right] = tr(\mathbf{M}\sigma^2 I_n) = \sigma^2 tr\left(\mathbf{M}\right) \\ &= \sigma^2[tr(I_n) - tr(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')] \\ &= \sigma^2[tr(I_n) - tr(I_K)] = \sigma^2(n - K)\end{aligned}$$

So we have shown that $E\left(s^2\right) = \sigma^2$.

4.5 The Distribution of $\hat{\beta}$ with GM and Normal errors

Given the Gauss-Markov (GM) theorem, what we know about the distribution of $\hat{\beta}$ is properties of its mean and variance, and we can estimate its variance. But we don't know the distribution of $\hat{\beta}$, and so we can't construct confidence intervals or hypothesis tests. To do more, we need additional assumptions. These assumptions can either be about finite sample properties, or asymptotic properties. One possible finite sample property is normality.

Let the GM assumptions hold. Suppose in addition we assume that the vector e is multivariate normal.

Since GM assumes $E(e) = 0$ and $E(ee') = \sigma^2 I_n$, this means that $e \sim N(0, \sigma^2 I_n)$.

Given GM, $\mathbf{X}\beta_0$ is a constant vector, and $Y = \mathbf{X}\beta_0 + e$. Since Y equals a constant plus a normal, we have that Y is normal: $Y \sim N(\mathbf{X}\beta_0, \sigma^2 I)$.

Also, under GM, $\hat{\beta}$ equals a constant matrix $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ times a normal Y, so $\hat{\beta}$ is also normal. And under GM the mean of $\hat{\beta}$ equals the true β_0 and the variance of $\hat{\beta}$ is $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, so with GM plus normal errors we get that

$$\hat{\beta} \sim N\left(\beta_0, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right)$$

This is a nice feature of linear estimators like OLS; we get to exploit the fact that linear functions of normals are normal.

Let β_k be the k-th element of the K-vector β. Let V_k be the (k,k) element of $(\mathbf{X}'\mathbf{X})^{-1}$. We therefore have that

$$\hat{\beta}_k \sim N\left(\beta_{0k}, \sigma^2 V_k\right) \qquad \text{and so we get the statistic:} \qquad \frac{\hat{\beta}_k - \beta_{0k}}{\sqrt{\sigma^2 V_k}} \sim N(0,1)$$

where β_{0k} is the unknown true value of β_k. Now $\sqrt{\sigma^2 V_k}$ is the standard deviation of $\hat{\beta}_k$, but we don't know σ^2. However, we do know s^2, which is an unbiased estimate of σ^2. However, if replace σ^2 with s^2 in the above formula, we are adding an additional source of randomness to the statistic, which will change it's distribution. What we get instead of a normal is:

$$\frac{\hat{\beta}_k - \beta_{0k}}{\sqrt{s^2 V_k}} \sim t_{n-K}$$

where t_{n-K} is the students t distribution with $n - K$ degrees of freedom. Here $\sqrt{s^2 V_k}$ is the standard error of $\hat{\beta}_k$.

Suppose we want to test whether the true β_{0k} equals some particular value β_k that we choose. Here I just want to give the intuition for the t-test, which you should already know how to formally do. First, we can calculate the value of the t-statistic:

$$\widehat{t} = \frac{\hat{\beta}_k - \beta_k}{\sqrt{s^2 V_k}}$$

If the null hypothesis $H_0 : \beta_{0k} = \beta_k$ is true, then $\widehat{t}$ will be a draw from the t distribution t_{n-K}. If not, then $\widehat{t}$ will have some other distribution. So we can look at $\widehat{t}$, and ask how likely are we to see a value like $\widehat{t}$ if it's distribution really is t_{n-K}. The t-distribution is pretty similar to a standard normal, and in particular, there's a roughly 95% chance of a t distributed random variable taking a value between -2 and 2 (the exact chance depends on the value of $n-K$). So if $\widehat{t}$ is much smaller than 2 in absolute value, we'd think H_0 might be true. On the other hand, it is very unlikely that a draw from a t-distribution would take a much larger number like 4 or -6, so if $\widehat{t}$ equalled a large value like that, we would reject the null hypothesis H_0.

The intuition for confidence intervals is similar. We have from the t distribution that

$$\Pr\left(-2 \le \frac{\hat{\beta}_k - \beta_{0k}}{\sqrt{s^2 V_k}} \le 2\right) \approx 0.95$$

which we can rewrite as

$$\Pr\left(\hat{\beta}_k - 2\sqrt{s^2 V_k} \le \beta_{0k} \le \hat{\beta}_k + 2\sqrt{s^2 V_k}\right) \approx 0.95$$

This says that there's a roughly 95% chance that the true value β_{0k} lies between the estimated values $\hat{\beta}_k - 2\sqrt{s^2 V_k}$ and $\hat{\beta}_k + 2\sqrt{s^2 V_k}$. Note what's random here: β_{0k} is a constant. It's the estimated end points of this interval that is random. So, really the correct statement is that there's a roughly 95% chance that the random interval $\left(\hat{\beta}_k - 2\sqrt{s^2 V_k}, \hat{\beta}_k + 2\sqrt{s^2 V_k}\right)$ contains the true constant β_{0k}.

And of course, these statements about t-tests and confidence intervals are only valid if the GM assumption hold and the errors e are normal. So e.g. we might reject a null hypothesis not because it's false, but because the errors weren't really normal, or some GM assumption was violated.

Note: Normality is not the only possible assumption we could have made about the distribution of e or Y. But:

1. The OLS estimator $\hat{\beta}$ has some particularly nice properties under normality, including having above distribution. We will see later that normality will also make $\hat{\beta}$ asymptotically efficient.

2. There are sometimes good reasons to think errors might be normal. For example, if errors are themselves each an average of a large number of individual small unobserved effects, then the errors might be approximately normal by the central limit theorem.

3. Sometimes normality is not a good assumption. For example, if each Y_i is an integer, then e can't be normal (an alternative estimator called Poisson regression is often used in this case). Another example is that, in many financial data sets, the tails of the distribution of Y are often observed to be too thick to be normal.

4.6 The Asymptotic Distribution of $\hat{\beta}$ with GM

Suppose we keep the GM assumptions but do NOT assume normality. If we add a few asymptotic assumptions, then we can derive an asymptotic distribution for $\hat{\beta}$.

Assume that the GM assumptions hold for each sample size n (above some minimum sample size). Define the $K \times K$ matrix $\mathbf{Q}_n$ by

$$\mathbf{Q}_n = \frac{\mathbf{X}'\mathbf{X}}{n} = \frac{\sum_{i=1}^{n} X_i X_i'}{n}$$

Assume the limiting matrix $\mathbf{Q}$ exists, defined by

$$\mathbf{Q} = \lim_{n\to\infty} \mathbf{Q}_n$$

This is an ordinary limit, not a probability limit, because by the GM assumptions each matrix $\mathbf{Q}_n$ is a matrix of constants.

Assume all the elements of $\mathbf{Q}$ are finite, and that $\mathbf{Q}$ is a non-singular matrix.

By the GM assumptions, each $\mathbf{Q}_n$ is nonsingular, but that is not sufficient to ensure that the limit $\mathbf{Q}$ is nonsingular, so we need that as a separate, additional assumption. We have by GM that $E(\hat{\beta}) = \beta_0$ and $Var(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \frac{\sigma^2}{n}\mathbf{Q}_n^{-1}$.

We can now consider some asymptotic properties $\hat{\beta}$. First:

$$\begin{aligned}\lim_{n\to\infty} E\left(\hat{\beta}\right) &= \lim_{n\to\infty} \beta_0 = \beta_0 \\ \lim_{n\to\infty} Var\left(\hat{\beta}\right) &= \lim_{n\to\infty} \frac{\sigma^2}{n}\mathbf{Q}_n^{-1} = \lim_{n\to\infty} \frac{\sigma^2}{n} \lim_{n\to\infty} \mathbf{Q}_n^{-1} \\ &= 0 \cdot \mathbf{Q}^{-1} = 0\end{aligned}$$

This show that $\hat{\beta} \overset{ms}{\to} \beta_0$ and therefore $\hat{\beta}$ is a consistent estimator of β.

Let $Z_i = X_i e_i$. Then $\mathbf{X}'e = \sum_{i=1}^n Z_i$. Assume $\sum_{i=1}^n Z_i/n$ satisfies a Central Limit Theorem.

Example: suppose the e_i are i.n.i.d. (independent, not necessarily identically distributed), and that the sequences of X_i and e_i are bounded. Then since the X_i are constants, Z_i is also i.n.i.d, and bounded, and soand so $\sum_{i=1}^n Z_i/n$ satisfies the Lindeberg-Feller CLT.

Given these assumptions

$$\sqrt{n}\frac{1}{n}\sum\nolimits_{i=1}^{n} Z_i \overset{d}{\longrightarrow} N\left(0, \mathbf{R}\right)$$

for some variance matrix $\mathbf{R}$. Note the mean is zero because, under GM, $E\left(Z_i\right) = E\left(X_i e_i\right) = X_i E\left(e_i\right) = 0$. We then have

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'Y = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{X}\beta_0 + e\right) = \beta_0 + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'e$$

$$\hat{\beta} - \beta_0 = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'e = \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\frac{\mathbf{X}'e}{n}$$

$$\begin{aligned}\sqrt{n}\left(\hat{\beta} - \beta_0\right) &= \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\sqrt{n}\frac{\mathbf{X}'e}{n}\right) \\ &= \mathbf{Q}_n^{-1}\left(\sqrt{n}\frac{1}{n}\sum\nolimits_{i=1}^{n} Z_i\right) \overset{d}{\longrightarrow} \mathbf{Q}^{-1}N\left(0, \mathbf{R}\right) = N(0, \mathbf{Q}^{-1}\mathbf{R}\mathbf{Q}^{-1})\end{aligned}$$

This says that $\hat{\beta}$ converges in distribution to a normal, which means that $\hat{\beta}$ is asymptotically normal, so

$$\hat{\beta} \overset{a}{\sim} N\left(E\left(\hat{\beta}\right), Var\left(\hat{\beta}\right)\right)$$

and, since we know by GM that $E\left(\hat{\beta}\right) = \beta_0$ and $Var\left(\hat{\beta}\right) = \frac{\sigma^2}{n}\mathbf{Q}_n^{-1}$, we can conclude that

$$\hat{\beta} \overset{a}{\sim} N\left(\beta_0, \frac{\sigma^2}{n}\mathbf{Q}_n^{-1}\right)$$

This approximate, or asymptotic distribution for $\hat{\beta}$ is the same as the exact distribution we found for $\hat{\beta}$ when the errors were normal. So now we can say that under Gauss-Markov, $\hat{\beta}$ *either* has exactly this distribution if e is normal, *or* it has approximately this same distribution if n is large and we have the additional asymptotic assumptions we made above.

Also, in either case we can estimate the variance $\frac{\sigma^2}{n}\mathbf{Q}_n^{-1}$ using $\frac{s^2}{n}\mathbf{Q}_n^{-1}$ (the square root of the diagonal elements of this matrix are the standard errors).

5 Lecture 05. Specification Issues

Readings for this lecture are: Greene Chapters 5, and 6.

5.1 Randomness of X and Heteroskedasticity

The assumption that $\mathbf{X}$ is constant and not random is often not reasonable for economic data. We now consider what happens to our estimator $\hat{\boldsymbol{\beta}}$ when we drop this assumption. Instead of the Gauss-Markov assumptions, consider this alternative list of assumptions, where $\mathbf{X}$ is now random instead of constant (or equivalently, where we do not condition on the value of $\mathbf{X}$).

We will also generalize the previous results by allowing for heteroskedasticity of the errors.

Assume the following:

1. $Y_i = X_i'\beta + e_i$ or equivalently in matrix form, $Y = \mathbf{X}\beta + e$. The true value of β is β_0.

2. The $K+1$ vectors $\binom{X_i}{e_i}$ for $i = 1, ..., n$ are i.n.i.d (independent, not identically distributed). This implies that $\binom{X_i}{e_i}$ is independent of $\binom{X_j}{e_j}$ for all $i \neq j$.

3. Let $\mathbf{Q}_n = \mathbf{X}'\mathbf{X}/n = \sum_{i=1}^{n} X_i X_i'/n$. Assume $\mathbf{Q}_n$ is nonsingular.

4. $E(e_i) = 0$ and $E(e_i X_i) = 0$. Unlike GM, it is no longer the case that $E(e_i X_i) = 0$ follows automatically from $E(e_i) = 0$, because now X_i is random. So now we need $E(e_i X_i) = 0$ as a separate assumption.

These assumptions allow e_i to be heteroskedastic. In particular, we can define $\sigma_i^2 = E(e_i^2 \mid X_i)$.

What properties does OLS have under these assumptions? Consider finite sample properties first:

We can write the OLS estimator as

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y = \mathbf{Q}_n^{-1}\frac{\sum_{i=1}^{n} X_i Y_i}{n} = \beta_0 + \mathbf{Q}_n^{-1}\frac{\sum_{i=1}^{n} X_i e_i}{n}$$

So

$$\begin{aligned} E\left(\hat{\beta}\right) &= \beta_0 + E\left(\mathbf{Q}_n^{-1}\frac{\sum_{i=1}^{n} X_i e_i}{n}\right) \\ &= \beta_0 + \frac{1}{n}\sum_{i=1}^{n} E\left(\mathbf{Q}_n^{-1} X_i e_i\right) \end{aligned}$$

Given our assumptions, we know that $E(X_i e_i) = 0$, but we do NOT know if $E(\mathbf{Q}_n^{-1} X_i e_i) = 0$, because $\mathbf{Q}_n^{-1}$ also contains X_i in it. ($\mathbf{Q}_n^{-1}$ also has every X_j for $j \neq i$, but we know those X_j's are independent of X_i and e_i, by the inid assumption).

The assumptions we wrote above are NOT sufficient to be able to tell if OLS is unbiased! And since it might not be unbiased, it also doesn't have the other finite sample properties, e.g., not BLUE.

To prove unbiasedness, we would need a stronger assumption like $E\left(e_i \mid X_i\right)=0$. If this was true, then we would also have $E\left(e_i \mid X_1,...,X_n\right)=0$ by the iid assumption. And we would then be able to apply the law of iterated expectations to prove unbiasedness, using $E\left(\mathbf{Q}_n^{-1}X_ie_i\right)=E\left[E\left(\mathbf{Q}_n^{-1}X_ie_i \mid X_1,...,X_n\right)\right]=E\left[\mathbf{Q}_n^{-1}X_iE\left(e_i \mid X_1,...,X_n\right)\right]=E\left(\mathbf{Q}_n^{-1}X_i0\right)=0$.

Next consider asymptotic properties. Start with assumptions 1 to 4 above, and make the following additional asymptotic assumption:

5. Let $\mathbf{Q}=plim\ \mathbf{Q}_n$. Assume $\mathbf{Q}$ exists, is finite, and nonsingular. Assume $plim\ \frac{1}{n}\sum_{i=1}^n X_ie_i=0$.

Earlier, when $\mathbf{X}$ was constant, $\mathbf{Q}_n$ was constant and so we could take an ordinary limit to define $\mathbf{Q}$. Now $\mathbf{X}$ is random, so $\mathbf{Q}_n$ is random, so we have to take a probability limit. Note that by its definition, $\mathbf{Q}_n$ is an average, so existence of a probability limit $\mathbf{Q}$ holds if we can apply a law of large numbers. Similarly, if we can apply a law of large numbers to the average of $X_i'e_i$, then $plim\ \frac{1}{n}\sum_{i=1}^n X_ie_i=lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n E\left(X_ie_i\right)=lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n 0=0$.

Under assumptions 1 to 5 above, we have (using the rule that the plim of a function equals the function of the plims):

$$\begin{aligned} plim\ \hat{\beta} &= plim\left(\beta_0+\mathbf{Q}_n^{-1}\frac{\sum_{i=1}^n X_ie_i}{n}\right) \\ &= \beta_0+\left(plim\mathbf{Q}_n\right)^{-1}plim\left(\frac{\sum_{i=1}^n X_ie_i}{n}\right) \\ &= \beta_0+\mathbf{Q}^{-1}0=\beta_0 \end{aligned}$$

So $\hat{\beta}$ is consistent (even though $\hat{\beta}$ may be biased).

Can we derive the asymptotic distribution of $\hat{\beta}$? Yes, but we need more assumptions. Define $\widetilde{R}_n$ by

$$\widetilde{\mathbf{R}}_n=\frac{1}{n}\sum_{i=1}^n X_iX_i'e_i^2$$

6. Assume $\mathbf{R}=plim\ \widetilde{\mathbf{R}}_n$ exists and is finite, by satisfying a law of large numbers, so $\mathbf{R}=lim_{n\to\infty}E\left(\widetilde{\mathbf{R}}_n\right)$.

Notice that

$$\begin{aligned} \mathbf{R} &= lim_{n\to\infty}E\left(\widetilde{\mathbf{R}}_n\right)=lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n E\left(X_iX_i'e_i^2\right) \\ &= lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n E\left[X_iX_i'E\left(e_i^2 \mid X_i\right)\right] \\ &= lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n E\left[X_iX_i'\sigma_i^2\right] \end{aligned}$$

7. Assume, as we did at the end of the previous section, that the sample average of X_ie_i satisfies the Lindeberg-Feller CLT.

Before we can calculate the limiting distribution of $\hat{\beta}$, we will need the limiting distribution of $\frac{\sqrt{n}}{n}\sum_{i=1}^{n} X_i e_i$. We have

$$E\left(\frac{\sqrt{n}}{n}\sum\nolimits_{i=1}^{n} X_i e_i\right) = \frac{\sqrt{n}}{n}\sum\nolimits_{i=1}^{n} E\left(X_i e_i\right) = 0$$

$$\begin{aligned} Var\left(\frac{\sqrt{n}}{n}\sum\nolimits_{i=1}^{n} X_i e_i\right) &= \frac{1}{n}E\left(\left(\sum\nolimits_{i=1}^{n} X_i e_i\right)\left(\sum\nolimits_{j=1}^{n} X_j e_j\right)'\right) \\ &= \frac{1}{n}E\left(\sum\nolimits_{i=1}^{n}\sum\nolimits_{j=1}^{n} X_i X_j' e_i e_j\right) \\ &= \frac{1}{n}E\left(\sum\nolimits_{i=1}^{n} X_i X_i' e_i^2\right) = E\left(\widetilde{\mathbf{R}}_n\right) \end{aligned}$$

So $lim_{n\to\infty} Var\left(\frac{\sqrt{n}}{n}\sum_{i=1}^{n} X_i e_i\right) = lim_{n\to\infty} E\left(\widetilde{\mathbf{R}}_n\right) = \mathbf{R}$. We can now apply the Lindeberg-Feller CLT to say

$$\frac{\sqrt{n}}{n}\sum\nolimits_{i=1}^{n} X_i e_i \xrightarrow{d} N(0, \mathbf{R})$$

Now we're ready to give the limiting distribution of $\hat{\beta}$:

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) = \mathbf{Q}_n^{-1}\left[\frac{\sqrt{n}}{n}\sum\nolimits_{i=1}^{n} X_i e_i\right]$$

The term $\mathbf{Q}_n^{-1} \xrightarrow{p} \mathbf{Q}^{-1}$, and the term $\frac{\sqrt{n}}{n}\sum_{i=1}^{n} X_i e_i \xrightarrow{d} N(0, \mathbf{R})$, so applying the rule about the product of two terms, where one converges in probability to a constant and the other converges in distribution, we have that $\sqrt{n}\left(\hat{\beta} - \beta_0\right)$ converges in distribution to $\mathbf{Q}^{-1}$ times a normal $N(0, \mathbf{R})$, which is $N(0, \mathbf{Q}^{-1}\mathbf{R}\mathbf{Q}^{-1})$. So we get

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N(0, \mathbf{Q}^{-1}\mathbf{R}\mathbf{Q}^{-1}).$$

So, with random $\mathbf{X}$, and heteroskedasticity, OLS is still root-n CAN, even if it may not be unbiased.

Case 1: Homoskedastic errors.

IF the errors are homoskedastic, then $\mathbf{R}$ simplifies to $\mathbf{R} = lim_{n\to\infty}\sigma^2\frac{1}{n}\sum_{i=1}^{n} E\left[X_i X_i'\right] = \sigma^2\mathbf{Q}$. So in the special case of homoskedasticity, we get $\mathbf{Q}^{-1}\mathbf{R}\mathbf{Q}^{-1} = \sigma^2\mathbf{Q}^{-1}$, and so $\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N(0, \sigma^2\mathbf{Q}^{-1})$ which is the same limiting distribution we had with under homoskedasticity with constant X or with normal errors. Which also means that, in this case, we have the same asymptotic distribution as before, that is,

$$\hat{\beta} \overset{a}{\sim} N\left(\beta_0, \frac{\sigma^2}{n}\mathbf{Q}_n^{-1}\right)$$

And we can estimate the variance $\frac{\sigma^2}{n}\mathbf{Q}_n^{-1}$ using our usual formula $\frac{s^2}{n}\mathbf{Q}_n^{-1}$ (the square root of the diagonal elements of this matrix are the standard errors).

Case 2: Heteroskedastic errors. With errors that are, or may be, heteroskedastic, we get the asymptotic distribution

$$\hat{\beta} \overset{a}{\sim} N\left(\beta_0, \frac{1}{n}\mathbf{Q}^{-1}\mathbf{R}\mathbf{Q}^{-1}\right)$$

To estimate the variance $\frac{1}{n}\mathbf{Q}^{-1}\mathbf{R}\mathbf{Q}^{-1}$ we need estimates of $\mathbf{Q}$ and $\mathbf{R}$. We can use $\mathbf{Q}_n$ to consistently estimate $\mathbf{Q}$. Similarly, $\widetilde{\mathbf{R}}_n$ would be a consistent estimate of $\mathbf{R}$, but $\widetilde{\mathbf{R}}_n$ depends on e, and we don't know the value of the true errors e. However, we do observe residuals, so instead of $\widetilde{\mathbf{R}}_n$, we can use the estimator

$$\widehat{\mathbf{R}}_n = \frac{1}{n}\sum\nolimits_{i=1}^{n} X_i X_i' \widehat{e}_i^2$$

and then estimate the variance matrix $\frac{1}{n}\mathbf{Q}^{-1}\mathbf{R}\mathbf{Q}^{-1}$ using the formula $\frac{1}{n}\mathbf{Q}_n^{-1}\widehat{\mathbf{R}}_n\mathbf{Q}_n^{-1}$. The square root of the diagonal elements of this matrix are standard errors that allow for unknown heteroskedasticity. These are called heteroskedasticity consistent standard errors, or White standard errors (after Hal White, in 1980). They are also known as Eicker-White standard errors, or Huber-Eicker-White standard errors, since both Eicker and Huber found special cases of them in 1967, before White showed the general case.

Summary: With random X we can get all the same asymptotic results we had with constant X, but not the small sample results (e.g., OLS is still root-n-CAN, but maybe not BLUE). We can also generalize to heteroskedastic errors, with a more general standard error formula (this generalization also works with constant X). But warning: this generalization is only asymptotic. White standard errors are only correct asymptotically, and so may be very inaccurate in small samples.

We end this section with some Nonstandard examples that violate the assumptions:

- Suppose $Y_i = X_i\beta + e_i$ with a scalar X_i and

  1. $E(e_i) = 0$
  2. $E(e_i^2) = \sigma^2$ for all i
  3. $E(e_ie_j) = 0$ for all $i \neq j$
  4. The true model is linear
  5. $x_i = \sqrt{i}$

In this case all of the Gauss-Markov assumptions hold, including X constant. So OLS is BLUE, including $\hat{\beta}$ unbiased. We have $\mathbf{Q}_n = \frac{1}{n}\sum x_i^2 = \sum \frac{i}{n} = \frac{n+1}{2}$, and $\mathbf{R}_n = \sigma^2\mathbf{Q}_n$. Given this, what is the limiting variance?

$$\lim_{n\to\infty} Var(\hat{\beta}) = \lim_{n\to\infty} \frac{1}{n}\mathbf{Q}_n^{-1}\mathbf{R}_n\mathbf{Q}_n^{-1} = \lim_{n\to\infty} \frac{2\sigma^2}{n(n+1)} = 0$$

Therefore we can conclude that $\hat{\beta}$ converges in mean square and so is also consistent. What is more noteworthy is the rate at which it converges.

$$\lim_{n\to\infty} Var(\sqrt{n}(\hat{\beta}-\beta_0)) = \lim_{n\to\infty} nVar((\hat{\beta}-\beta_0)) = \lim_{n\to\infty} \frac{n2\sigma^2}{n(n+1)} = 0$$

In this case OLS is "too accurate"! The usual stabilizing transformation $\sqrt{n}(\hat{\beta}-\beta_0)$ does not yield a (limit normal) random variable; it converges in mean square to zero. If we instead look at $n(\hat{\beta}-\beta_0)$, we get an asymptotic variance of $2\sigma^2$, which is finite and nonzero.

In this example, OLS is converging at the rate n instead of $\sqrt{n}$, and we cannot apply our usual CLT to get the asymptotic distribution. The asymptotic assumption that $\mathbf{Q}$ (the probability limit of $\mathbf{Q}_n$) is finite is what's violated here.

This is because X_i is growing (it is nonstationary). Recall from Lecture 4 that the greater is the sample variance of the observed X's, so the more spread out the X observations are, the smaller is the variance of $\hat{\beta}$. In this example, the X's are quickly spreading out, so the sample variance of X is rapidly increasing with n.

Another nonstandard example:

Suppose we have the model $Y_i = X_i'\boldsymbol{\beta} + e_i$, which we will write in matrix form as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$. Suppose $\mathbf{X}$ is constant, $E(\mathbf{e}) = 0$, and assume the sample mean $\overline{X} = \sum_{i=1}^{n} X_i/n$ has a finite limit $\mu = \lim_{n\to\infty}\overline{X}$. Let Assumption 5 above hold.

Assume $E(e_ie_j) = \begin{cases} 1 \text{ if } i \neq j \\ \sigma^2 \text{ if } i = j \end{cases}$.

Write the OLS estimator as estimator as $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}$. OLS is unbiased:

$$E\left(\hat{\boldsymbol{\beta}}\right) = E\left(\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}\right) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{e}) = \boldsymbol{\beta}$$

But now look at the variance of $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{ee})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &\mathbf{Q}_n^{-1}\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} X_iX_j'E(e_ie_j)\right)\mathbf{Q}_n^{-1} \\ &= \mathbf{Q}_n^{-1}\overline{XX}'\mathbf{Q}_n^{-1} + \mathbf{Q}_n^{-1}\frac{1}{n^2}\sum X_iX_i'(\sigma^2-1)\mathbf{Q}_n^{-1} \end{aligned}$$

If we take the limit we get

$$\lim_{n\to\infty} Var(\hat{\boldsymbol{\beta}}) = \mathbf{Q}^{-1}\mu\mu'\mathbf{Q}^{-1} \neq 0$$

The OLS estimator is unbiased, but does not converge in mean square, and so it is not consistent. Why not?

The problem is too much autocorrelation; all of the errors are positively correlated with each other. Intuitively, if the first error is positive, then the majority of the other errors will also likely be positive, shifting the estimated line away from true line even asymptotically.

But then why are we still unbiased? Remember that bias looks at averaging across many hypothetical samples, each with fixed n, while consistency looks at one sample with n going to infinity. The above logic about consistency applies to one sample. But with many samples, some will have the first error positive, others negative (and similarly for the other errors), so while each sample is likely to have a line that is too high or too low, those biases average out across samples, making the overall bias be zero.

Why might errors be positively correlated in this way? Could be omitted variables, or social interactions.

5.2 Multicollinearity

Consider the model with two regressors and a constant: $Y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + e_i$. So the matrix $\mathbf{X}$ has three columns, the first is ones, the second contains each X_i and the third contains each Z_i. If you look at the middle element of $Var\left(\hat{\beta}\right) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$

$$Var\left(\hat{\beta}_2\right) = \frac{\sigma^2}{\sum_i \left(X_i - \overline{X}\right)^2 (1 - r^2)}$$

where r equals the sample correlation between X and Z. Remember that the correlation between two variables ranges between -1 and 1. So the higher in magnitude the correlation r is, the larger is the variance of $\hat{\beta}_2$ (and similarly for $\hat{\beta}_3$). If X and Z are perfectly correlated, meaning $r^2 = 1$, then the variance becomes infinite. In this case, $rank\,(X) < 3$, and $\mathbf{X}'\mathbf{X}$ is singular. This is an example of what is called perfect collinearity. If r^2 is close to one, so $\mathbf{X}'\mathbf{X}$ is close to singular, then we have an example of what is called multicollinearity.

Perfect Collinearity:

A regression is defined to have perfect collinearity if $\mathbf{X}'\mathbf{X}$ is singular, which happens when the rank of the matrix $\mathbf{X}$ is lower than the number of variables K. Equivalently, this is when one regressor (one column of $\mathbf{X}$) exactly equals a linear function of the other regressors.

Perfect collinearity violates the GM assumption that $\mathbf{X}'\mathbf{X}$ be non-singular. If you have perfect collinearity, you will know it immediately when you try to run the regression, because the computer will be unable to invert $\mathbf{X}'\mathbf{X}$. Some regression packages will automatically discard some regressors when this happens.

The only solution to perfect collinearity is to drop some regressors, or change the model.

Usually perfect collinearity is a mistake: you've accidently constructed/chosen data where one regressor is by construction or definition linear in others.

Multicollinearity:

Multicollinearity is when $\mathbf{X}'\mathbf{X}$ is close to singular. This makes elements of $Var\left(\hat{\beta}\right) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$ be very large. Multicollinearity does not violate any GM conditions, so OLS keeps all the properties we've discussed. However, multicollineary implies high variances and hence imprecise estimates of β with large standard errors.

Why does multicollinearity cause these problems? Each element of β is defined to be the effect of changing one regressor on Y, holding the other regressors fixed. When regressors are highly correlated with each other and so tend to move together, it is difficult to estimate the separate effects of each regressor.

- Examples.

  1. suppose we have variables x and z as regressors, where x is average wage per hour and z is average wage per day. If workday is 8 hours, then $z = 8x$. Perfect collinearity. But if length of work day varies, then this is just multicollinearity.
  2. suppose we have variables x, z and w as regressors, where x is average wage per hour; z is average hours per week, and w is money earned per week. Have $w = zx$. This is a perfectly deterministic relationship, but it is not a linear relationship. So multicollinearity is likely, but we don't have perfect collinearity.

3. If in previous example, the regressors were $logw$, $logz$, and $logx$ then we do have perfect collinearity, because $logw = logz + logx$ is a linear relationship.
4. suppose we use x and x^2 as regressors. Again one is a deterministic functionn of the other, but not a linear function of the other. Not perfect collinearity, but multicollinearity is likely.

How can we tell if we have multicollinearity?

Multicollinearity is NOT a zero-one, you either have it or not problem. The issue is one of severity. Is $\mathbf{X}'\mathbf{X}$ close enough to singular so that coefficient estimates are very imprecise? The closer to singular, the worse the potential problem

Possible evidence of a multicollinearity problem:
1. High correlation between regressors.
2. Determinant of $\mathbf{X}'\mathbf{X}$ is near zero.
3. If dropping one or more regressors changes coefficient estimates a lot, and makes standard errors much smaller, while not changing R^2 much.

What is R^2 of a regression?

$$R^2 = 1 - \frac{\hat{e}'\hat{e}}{(y-\bar{y})'(y-\bar{y})}$$

R^2 compares how much errors vary relative to how much y varies.
R^2 ranges from zero to one.
Essentially R^2 equals how much of the variation in y is explained by the regression model, vs how much by the error term.
Example, $R^2 = .9$ says that 90% of the variation in y is explained by variation in the regressors, and 10% is due to variation in the errors.
R^2 is only meaningful if the regression includes a constant term.
Adding regressor always makes R^2 increase (unless the coefficient of the regressor is zero)

There is also adjusted R^2, or R-bar-squared:

$$\bar{R}^2 = 1 - \frac{\hat{e}'\hat{e}/(n-k)}{(y-\bar{y})'(y-\bar{y})/(n-1)}$$

Instead of the relative variation in e vs y, $\bar{R}^2$ compares the estimated variance of e to the estimated variance of y.
Adding a regressor can make $\bar{R}^2$ decrease if the degrees of freedom correction in the numerator (increasing k) more than offsets the reduction in residual $\hat{e}$ from including the regressor.

If $\bar{R}^2$ and R^2 change very little when a given regressor is added or removed, then that regressor is explaining very little additional variation in y.

What to do about multicollinearity? Possible answers:
1. Do nothing. multicollinearity does not violate any assumptions, so just live with the large standard errors.
2. Drop one or more variables from the model

3. Change the specification of the model in some other way.

What happens if we drop a variable?

Suppose $Y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + e_i$ satisfies GM. Let $\widetilde{\beta}_2$ be the OLS estimate of β_2 if we drop Z from the regression, and so only regress Y_i on a constant and on X. Then can show

$$plim\left(\widetilde{\beta}_2\right) = \beta_2 + \frac{cov(X,Z)}{var(X)}\beta_3$$

So dropping Z_i causes $\widetilde{\beta}_2$ to be inconsistent, unless either X and Z are uncorrelated or if $\beta_3 = 0$.

Multicollinearity usually means $cov(X,Z)$ large, so get large bias from dropping Z, unless true β_3 is close to zero.

However, dropping Z can make the variance (standard errors) of estimate of β_2 shrink a great deal.

Have a mean squared error tradeoff: dropping a multicollinear variable increases bias and decreases variance.

We can see how much variance decreases (compare standard errors before vs after dropping Z). But we don't know bias.

Similar math applies to dropping a vector of regressors. Suppose true model is $Y = \mathbf{X}B_2 + \mathbf{Z}B_3 + e$ for matrices of variables $\mathbf{X}$ and $\mathbf{Z}$ and vectors of coefficients B_2 and B_3. If drop $\mathbf{Z}$ so just estimate B_2 by $\widetilde{B}_2 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$, then

$$plim\left(\widetilde{B}_2\right) = B_2 + plim\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\right)B_3$$

There exist statistical methods for helping to decide whether to drop variables or not.

1. Information criterion tests (e.g., using changes in $\bar{R}^2$ to decide)

2. Bayesian methods (ridge regression, Stein shrinkage estimators): reduces estimated coefficients, instead of setting one or more of them equal to zero.

2. LASSO (least absolute shrinkage and selection operator). Popular in machine learning, this tries to drop variables having coefficients that are estimated to be small.

Bottom line: including or excluding a regressor comes down to judgement.

Whether to include a variable in a multicollinear model is often the wrong question.

Better is to ask: can I redefine my variables or my model to reduce the problem?

Example: Suppose Z was population. Maybe instead of dropping it, I can incorporate it into the model in some other way, like making all my variables be per capita. Or maybe a different variable like number of households will be less collinear with other regressors.

Similar arguments could apply to collections of potentially multicollinear regressors like prices, exchange rates, income, etc.

5.3 Coefficient and Elasticity

Under the model $Y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + e_i$, coefficients are

$$\beta_2 = \frac{\partial Y_i}{\partial X_i} \text{ and } \beta_3 = \frac{\partial Y_i}{\partial Z_i}$$

Suppose we wanted to know the elasticity of Y with respect to X.

$$\varepsilon_{Y,X} \approx \frac{\%\Delta Y}{\%\Delta X} = \frac{\Delta Y/Y}{\Delta X/X}, \quad \varepsilon_{Y,X} = \frac{d\ln Y}{d\ln X} = \frac{dY/Y}{dX/X} = \frac{dY}{dX}\frac{X}{Y}$$

So the elasticity of Y with respect to X at the realization (x, y) is

$$\varepsilon_{y,x} = \beta_2 \frac{x}{y}$$

We could evaluate this at the current values of x, y to have the estimate the current value of the elasticity.
The elasticity at the mean of the data is

$$\varepsilon_{\overline{y},\overline{x}} = \beta_2 \frac{\overline{x}}{\overline{y}}$$

The average elasticity over our sample is

$$\varepsilon = \beta_2 \frac{1}{n} \sum_{i=1}^{n} \frac{x_i}{y_i}$$

If we believe the elasticity is constant, then we should instead estimate the log model: $logY_i = \alpha_0 + \alpha_1 logX_i + \alpha_2 logz_i + e_i$. This model says that the elasticity is α_1, a constant at all points of the data.

What you shouldn't do: say you want to know the elasticity, so estimate the model in logs.

What you should do: estimate the model in whatever form is most likely to satisfy the the GM assumptions (or other assumptions for consistency), and then calculate the elasticity from your estimates.

5.4 Units of Measurement

Consider two models, both estimated by OLS.

$$Y_i = \alpha + \beta X_i + e_i$$

$$Y_i = \alpha^* + \beta^* X_i^* + e_i^*$$

The only difference is that X is measured in thousands of dollars, while X^* is the same variable but measured in dollars. For example, if $X_i = 7$ then $X_i^* = 7000$. So $X_i^* = 1000 X_i$.

How do the OLS estimates compare? Answer:

$$\widehat{\beta} = 1000\widehat{\beta}^*, \quad \widehat{\alpha} = \widehat{\alpha}^*, \ \ \widehat{e}_i = \widehat{e}_i^*$$

Changing the units X is measured in changes the meaning of the coefficient, but doesn't change the constant in the regression.

Why? intuition is, this makes $\widehat{\beta}X_i = \widehat{\beta}^* X_i^*$. If they weren't equal, then one of the two regressions (whichever one had bigger errors) would not actually be minimizing the sum of squared errors.

Let $s_{\widehat{\beta}}$ be standard error of $\widehat{\beta}$ and $s_{\widehat{\beta}^*}$ be standard error of $\widehat{\beta}^*$.

How do the OLS standard errors compare? Answer: $s_{\widehat{\beta}} = 1000 s_{\widehat{\beta}^*}$, and $s_{\widehat{\alpha}^*} = s_{\widehat{\alpha}^*}$.

Why? because t-statistics testing if the coefficient is zero are: $\widehat{\beta}/s_{\widehat{\beta}}$ and $\widehat{\beta}^*/s_{\widehat{\beta}^*}$. Now β is zero if and only if β^* is zero, so intuition is testing if either is zero should have the same p-value, and hence the same t-statistic.

Now consider the same variables with X logged:

$$Y_i = \alpha + \beta \log X_i + e_i$$
$$Y_i = \alpha^* + \beta^* \log X_i^* + e_i^*$$

Then

$$\begin{aligned} Y_i &= \alpha + \beta \log (1000 X_i^*) + e_i \\ &= (\alpha + \beta \log (1000)) + \beta \log X_i^* + e_i \end{aligned}$$

So by the same logic as before that minimizing mean squared errors means $\widehat{e}_i = \widehat{e}_i^*$, we now get

$$\widehat{\beta} = \widehat{\beta}^*, \quad \widehat{\alpha} + \log (1000) \widehat{\beta} = \widehat{\alpha}^*,$$

So now the coefficient of X doesn't change, $\widehat{\beta} = \widehat{\beta}^*$. Will also get $s_{\widehat{\beta}} = s_{\widehat{\beta}^*}$. This is another example of elasticities not depending on the units of measurement. But the constant does change, and it might become more or less statistically different from zero.

General rule: don't drop the constant in a regression, even if it is statistically insignificant, unless theory strongly says it should be zero.

5.5 Modeling % changes

Suppose you want to include the growth rate of some variable Z_t in a model. Do NOT construct it as a percent change,

$$\frac{Z_t - Z_{t-1}}{Z_{t-1}}$$

Instead, define the growth rate as

$$X_t = \log \frac{Z_t}{Z_{t-1}} = \log Z_t - \log Z_{t-1}$$

This latter definition corresponds to a continuous rather than discrete rate of change, and so is not sensitive to the time units. For example, suppose you switched from quarterly data to annual. If you added together four quarters of percent changes measured by $\frac{Z_t - Z_{t-1}}{Z_{t-1}}$, that would not equal the annual percent change $\frac{Z_t - Z_{t-4}}{Z_{t-4}}$. But if you add together four quarters measured by $\log Z_t - \log Z_{t-1}$, you'll get the annual change $\log Z_t - \log Z_{t-4}$.

5.6 Non-linear terms

If we think the dependence of Y on X is quadratic instead of linear, we can estimate the model as

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + e_i$$

This is still a linear regression model, since it has Y linear in the regressors X_i and X_i^2. This does not violate GM assumptions, and in particular doesn't cause perfect collinearity, because X_i and X_i^2, though deterministically related, are not linearly related.

However, the model might have a multicollinearity problem.

Also, one must be careful about how to interpret the coefficients. We have:

$$\frac{dY}{dX} = \beta_1 + 2\beta_2 X$$

One can similarly have interaction terms, e.g.,

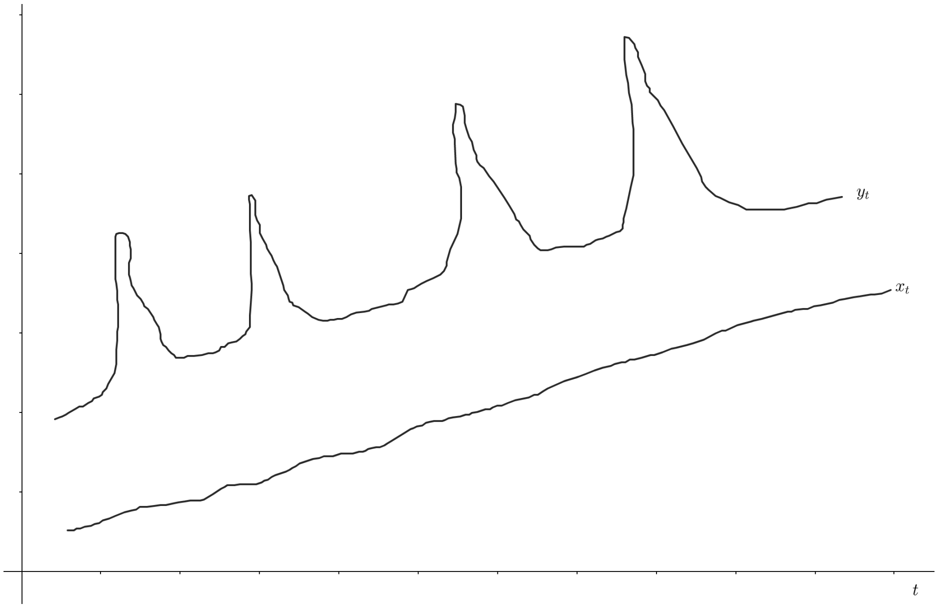
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i + e_i$$

again, multicollinearity is a potential concern, and one must interpret coefficients carefully.

$$\frac{dY}{dX} = \beta_1 + \beta_3 Z$$

5.7 Dummy Variables

Suppose we have y_t =sales in month t, and x_t=advertising in month t. Before estimating a model we observe the following graph of the data.



We observe that every December, y_t spikes upward because of the holiday season. Therefore if we want to forecast the effect advertising has on sales, we may want to control for the presence of

December in our data. We can do this with a dummy variable, that is, a variable that only takes the values zero and one.

$$z_t = \begin{cases} 1 \text{ if month is december} \\ 0 \text{ otherwise} \end{cases}$$

The model is

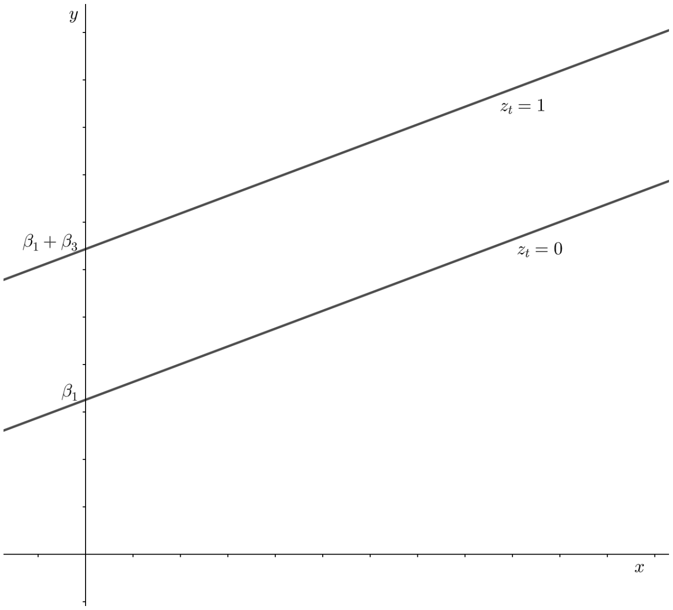
$$y_t = \beta_1 + \beta_2 x_t + \beta_3 z_t + e_t$$

How to interpret β_3? So how does this effect our model? It means that the intercept of our model changes to account for the jump in sales that come about in December. If it is not December, then $z_t = 0$ and

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 \cdot 0 + e_t = y_t = \beta_1 + \beta_2 x_t + e_t.$$

If it is December, then $z_t = 1$ and

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 \cdot 1 + e_t = y_t = (\beta_1 + \beta_3) + \beta_2 x_t + e_t.$$

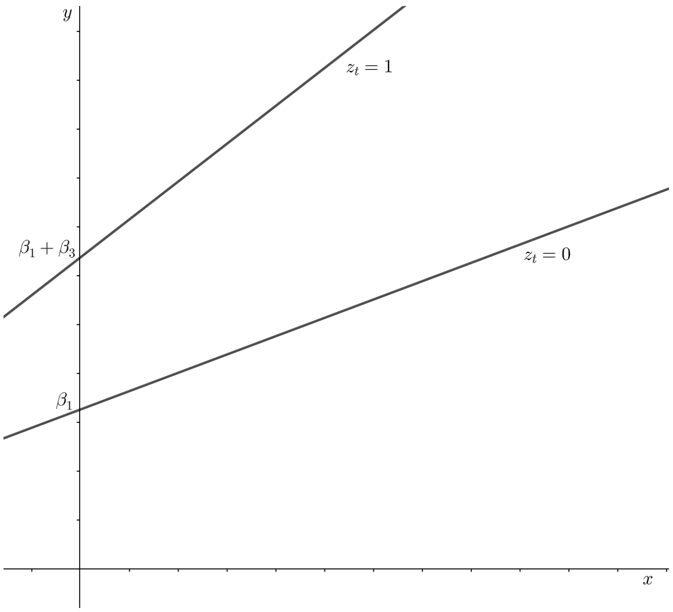


Suppose we think advertising has a different effect in December from other months. Then we could add an interaction term, to account for this hypothesized interaction between x_t and z_t. The model is then

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 z_t + \beta_4 x_t z_t + e_t.$$

Now the effect z_t has on our model is not only shifting the intercept, but it also changes the slope of the line to account for an increased effectiveness of advertising in the month of December.

$$\begin{array}{rr} z_t = 0 & y_t = \beta_1 + \beta_2 x_t + e_t \\ z_t = 1 & y_t = (\beta_1 + \beta_3) + (\beta_2 + \beta_4) x_t + e_t \end{array}$$



In terms of coefficient estimates, this is the same as running two regressions: one with non december data, having intercept β_1 and slope β_2, and one with just decembers, having intercept $\beta_1 + \beta_3$ and slope $\beta_2 + \beta_4$. However, when estimated as one regression with dummy variables and interactions, we get different standard errors, than if we estimated as two separate regressions.

Forget the interaction term now. Suppose we thought the intercept was different in every month. Could we add a dummy for every month? No, because the sum of those dummies across all months would equal the constant term, meaning we'd have perfect collinearity. We could instead have a dummy for every month but one (say, January). Then β_1 would be the intercept in January, and β_1 plus the coefficient of any other month's dummy would be that month's intercept.

5.8 Difference-in-Difference

Difference-in-difference (DiD) is a model where a 'treatment' (e.g., an event like a natural disaster) happens that affects one group of people and not another. Let group one be the group of people who are treated (e.g., live where the natural disaster occurred), and group zero be the control group. We have some outcome Y. We create two dummy variables, D and T:

$$D_i = \begin{cases} 1 & \text{for people } i \text{ in group 1} \\ 0 & \text{for people } i \text{ in group 0} \end{cases}$$

$$T_i = \begin{cases} 1 & \text{people } i \text{ who are observed after the event} \\ 0 & \text{people } i \text{ who are observed before the event} \end{cases}$$

The DiD model is then just a regression with these dummy variables and their interaction:

$$Y_i = \beta_1 + \beta_2 T_i + \beta_3 D_i + \beta_4 T_i D_i + e_i$$

β_2 is the average change in outcome (before vs after the event) in the control group.
$\beta_2 + \beta_4$ is the average change in outcome (before vs after the event) in the treatedl group.

DiD makes the "parallel trends" assumption: Suppose the average change in outcome Y over time would have been the same in both groups, if neither had been treated.

If parallel trends is true, then β_4 equals the difference between the average outcome in the treated group, and what the average outcome in the treated group would have been, had they had not been treated. This is called the "Average Treatment effect on the Treated," or ATT. This will be discussed further in lecture 13.

6 Lecture 06. Maximum Likelihood Estimation

Readings for this lecture are: Greene Chapter 14, and 17.1-17.3.

6.1 Maximum Likelihood Estimation

We first showed OLS is BLUE. We then added the assumption of normal errors, so we could do inference without asymptotics.

Question: Could we use the assumption of normality to get a better estimator than OLS?

More generally, how can we use knowledge of distributions to construct estimators?

This is what Maximum Likelihood Estimation (MLE) is about.

6.1.1 Motivating example

Roughly speaking the idea of MLE is to choose as your estimate the value of the parameters that maximizes the probability of seeing the data that you actually observed.

Example 1: Suppose you have a coin, and want to estimate the probability p that a coin flip comes up heads.

Suppose you flip the coin 100 times and get 100 heads.

It is possible the coin is fair, so $p = 1/2$.
It is more likely that the coin is unfair, and favors heads, maybe $p = 3/4$.
But the most likely thing, given the data, is that it's a two-headed coin, so $p = 1$.

In this case, the MLE estimate is $\widehat{p} = 1$, because it is more likely to get 100 heads with a 100 flips with $\widehat{p} = 1$ than with any other value of p.

Formally, assume the coin flips are IID. Then the probability of getting 100 heads is p^{100}, and for p in the range of zero to one (as required for a probability), p^{100} is maximized at $p = 1$.

Now suppose what we got was 99 heads, but then 1 tail. The probability of first getting 99 heads, then one tail, is $p^{99}(1-p)$. The MLE estimate $\widehat{p}$ is then

$$\widehat{p} = \arg\max_{p\in[0,1]} p^{99}(1-p)$$

The first order condition for maximizing the probability of 99 heads then one tail is:

$$\begin{aligned}\frac{d}{dp}\left(p^{99} - p^{100}\right) = 0 &\Rightarrow \left(99p^{98} - 100p^{99}\right) = 0\\ &\Rightarrow p^{98}(99 - 100p) = 0\\ &\Rightarrow \widehat{p} = 0.99 \text{ or } \widehat{p} = 0\end{aligned}$$

Checking the second order condition, you can verify that $\widehat{p} = 0$ is a minimum, and $\widehat{p} = 0.99$ is the maximum. This formally shows the obvious answer: if 99% of the flips come up heads, then the MLE estimate of p is 0.99.

More generally, suppose our sample is a realization of the random vectors $Z_1, \ldots, Z_n$. Suppose this data has some joint distribution that depends on a vector of unknown parameters θ. If the Z's are continuously distributed, then they have some joint pdf (probability density function) $f(Z_1, \ldots, Z_n \mid \theta)$. Alternatively, if the Z's are discretely distributed (like in the coin flip example), then they have some joint pmf (probability mass function) $f(Z_1, \ldots, Z_n \mid \theta)$.

The likelihood function L is defined to be

$$L(\theta \mid Z_1, \ldots, Z_n) = f(Z_1, \ldots, Z_n \mid \theta)$$

That is, the likelihood function is just the pdf or pmf, except that we treat the parameters θ as the unknown, conditional on the data, instead of the other way around.

The maximum likelihood estimate is then

$$\hat{\theta}_{ML} = \arg\max_{\theta} L(\theta \mid Z_1, \ldots, Z_n)$$

6.1.2 MLE Assumptions

Here we will make assumptions that allow us to derive attractive asymptotic properties of MLE. These assumptions are somewhat stronger than necessary, but make derivations easier.

ASSUMPTION 1: Z_i are i.i.d. continuously distributed random vector. Let $f(Z_i \mid \theta)$ be the pdf of Z_i, where θ is a $K \times 1$ vector. Let Ω_Z denote the support of Z_i.

Notes: The support of a random vector is just the set of all possible values that the vector can equal.

Assumption 1 implies that

$$L(\theta \mid Z_1, \ldots, Z_n) = f(Z_1 \mid \theta) f(Z_2 \mid \theta) ... f(Z_n \mid \theta)$$

which, implies

$$\frac{1}{n} \ln L(\theta \mid Z_1, \ldots, Z_n) = \frac{1}{n} \sum\nolimits_{i=1}^{n} \ln f(Z_i \mid \theta)$$

maximizing L is the same as maximizing $\frac{1}{n} \ln L$, and this is easier to work with because, with IID data, $\frac{1}{n} \ln L$ equals an average.

Almost all the same math goes through if Z_i are discrete instead of continuous by letting f be the probability mass function. One can also do MLE with some elements of Z being continuous and other discrete.

A key requirement for doing MLE is knowing the likelihood function, which means knowing the density function f. The only thing unknown about the distribution of the data is the parameter vector θ.

EXAMPLE: In the coin flipping example, let Z_i be one if coin flip number i was a head, otherwise let Z_i be zero. In this case $\Omega_Z = \{0, 1\}$, since those are the only values that each Z_i can equal. In our data, the realizations z_i of the random variables Z_i were $z_i = 1$ for $i = 1, 2, ..., 99$, and $z_{100} = 0$. Let θ be the unknown probability p of a head. Then we have the pmf

$$f(Z_i \mid \theta) = \theta^{Z_i} (1 - \theta)^{1 - Z_i}$$

since this expression equals θ when $Z_i = 1$ and it equals $1 - \theta$ when $Z_i = 0$. Therefore

$$\frac{1}{n} \ln L(\theta \mid Z_1, \ldots, Z_n) = \frac{1}{100} \sum\nolimits_{i=1}^{100} Z_i \ln \theta + (1 - Z_i) \ln(1 - \theta)$$

The first order condition for maximizing this expression is θ

$$\begin{aligned} 0 &= \frac{\partial \frac{1}{n} \ln L(\theta \mid Z_1, \ldots, Z_n)}{\partial \theta} \\ &= \frac{1}{100} \sum\nolimits_{i=1}^{100} \frac{Z_i}{\theta} - \frac{1 - Z_i}{1 - \theta} \\ &= \frac{1}{100} \left(\frac{99}{\theta} - \frac{1}{1 - \theta} \right) \end{aligned}$$

which, if you solve for θ, gives $\theta = 99/100$.

COVARIATES: Suppose, as in our GM linear regression model, we have $Z = (Y, X)$ where Y is random and X is either fixed, or is random but we condition on X. Assume the distribution of X does not depend on θ. let $f_{y,x}$, $f_{y|x}$, and f_x, denote, respectively, the joint distribution of Y and X, the conditional distribution of Y given X, and the marginal distribution of X. Then

$$f_{y,x}(Y_i, X_i \mid \theta) = f_{y|x}(Y_i \mid X_i, \theta) f_x(X_i, \mid \theta)$$
$$\ln f_{y,x}(Y_i, X_i \mid \theta) = \ln f_{y|x}(Y_i \mid X_i, \theta) + \ln f_x(X_i)$$

so

$$\frac{1}{n} \ln L(\theta \mid Z_1, \ldots, Z_n) = \left[\frac{1}{n} \sum\nolimits_{i=1}^{n} \ln f_{y|x}(Y_i \mid X_i, \theta) \right] + \left[\frac{1}{n} \sum\nolimits_{i=1}^{n} \ln f_x(X_i) \right]$$

The second sum here doesn't depend on θ, so maximizing this likelihood function is the same as just maximizing $\frac{1}{n} \sum_{i=1}^{n} \ln f_{y|x}(Y_i, \mid X_i, \theta)$. We can therefore ignore the distribution of X, and just define the likelihood function in terms of the conditional distribution of Y given X.

Now continue with our MLE assumptions:

ASSUMPTION 2: The true value of θ is θ_0. $\theta_0 \in \Theta$, where Θ is compact set.

Note: If each element of θ lies in some closed interval, then that is an example of a compact set. In the coin flipping example, θ is a probability and so must lie in the closed interval $0 \leq \theta \leq 1$, so in that example Assumption 2 is satisfied with $\Theta = [0, 1]$.

ASSUMPTION 3: $\ln f(Z \mid \theta)$ is three times differentiable in θ, with bounded first, second, and third own and cross derivatives, for all $Z \in \Omega_Z$ and all $\theta \in \Theta$.

Note: This assumption is stronger than necessary, but will make some proofs simpler later. One implication of this assumption is that the expected absolute values of these derivatives are also bounded.

ASSUMPTION 4:

$$E\left(\frac{\partial \ln f(Z \mid \theta)}{\partial \theta} \right) \neq E\left(\frac{\partial \ln f(Z \mid \theta_0)}{\partial \theta} \right) \qquad \text{for all } \theta \neq \theta_0 \text{ where } \theta \in \Theta$$

Note: This will later be used to ensure identification, based on the first order conditions.

Aside: Don't be confused by the fact that we are taking an expection, and that f is a density. The random variable here is Z. The derivative $\frac{\partial \ln f(Z|\theta)}{\partial \theta}$ is just a particular function of Z, and we are taking the expectation of that function. Note that by definition of an expectation, we have

$$E\left(\frac{\partial \ln f(Z \mid \theta)}{\partial \theta}\right)=\int_{\Omega_Z} \frac{\partial \ln f(Z \mid \theta)}{\partial \theta} f(Z \mid \theta_0)\, dZ$$

Pay attention to where we write θ and where we write θ_0 above! Expection of a function is defined as the integral of the function muliplied by the true density, which is $f(Z \mid \theta_0)$.

ASSUMPTION 5: Either Ω_Z does not depend on θ, or $f(Z_i \mid \theta)=0$ on the boundary of Ω_Z.

Note this assumption rules out a uniform distribution. For example, suppose Z_i is uniform, having the density function $f(Z \mid \theta)=\theta$ for all $Z \in \Omega_Z$. Then, because the area under a density must equal one, Ω_Z must be an interval of length θ (or multiple intervals with lengths that add to θ). So in this case Ω_Z does depend on θ, and the density does not equal zero at the end points of the interval Ω_Z, so Assumption 5 is violated.

ASSUMPTION 6: $Var\left(\frac{\partial \ln f(Z|\theta_0)}{\partial \theta}\right)$ is non-singular.

6.1.3 Score function

Before we derive the asymptotic theory of ML estimators, we need to look at some properties of likelihood function derivatives. Define the function $s(Z_i \mid \theta)$ by

$$s(Z_i \mid \theta)=\frac{\partial \ln f(Z_i \mid \theta)}{\partial \theta}$$

$s(Z_i \mid \theta)$ is a $K \times 1$ random vector. Note that since Z_i are i.i.d, it follows that $s(Z_i \mid \theta)$ is also i.i.d.

This function s is closely related to the score function, defined by

$$Score \text{ function}=\frac{\partial \ln L(\theta \mid Z_1,\ldots,Z_n)}{\partial \theta}$$

To see the connection, observe that with iid data

$$Score \text{ function}=\frac{\partial \sum_{i=1}^n \ln f(Z_i \mid \theta)}{\partial \theta}=\sum_{i=1}^n s(Z_i \mid \theta)$$

We now derive an interesting property of the function s, starting from the fact that densities integrate to one. For any $\theta \in \Theta$:

$$1=\int_{\Omega_Z} f(Z \mid \theta)\, dZ$$

Now take derivative of this expression with respect to θ:

$$0=\frac{d1}{d\theta}=\frac{d}{d\theta}\int_{\Omega_Z} f(Z \mid \theta)\, dZ$$

By Leibniz rule we can take the derivative inside the integral. However, by Leibniz rule we have to consider boundary terms. However, Assumption 5 ensures that these boundary terms are zero.

For example, suppose Z is a scalar and Ω_Z equals the interval from a to b. Then

$$\frac{d}{d\theta}\int_a^b f(Z \mid \theta)\, dZ = \int_a^b \frac{df(Z \mid \theta)}{d\theta} dZ + f(b \mid \theta)\frac{db}{d\theta} - f(a \mid \theta)\frac{da}{d\theta}$$

But by Assumption 5, either the pdf on the boundary, $f(b \mid \theta)$ and $f(a \mid \theta)$, equal zero, or Ω_Z (and therefore b and a) do not depend on θ, so $\frac{db}{d\theta}$ and $\frac{da}{d\theta}$ equal zero.

We therefore have

$$0 = \int_{\Omega_Z} \frac{df(Z \mid \theta)}{d\theta} dZ$$

Now, doing a little algebra, we get

$$\begin{aligned} 0 &= \int_{\Omega_Z} \frac{df(Z \mid \theta)}{d\theta} \frac{f(Z \mid \theta)}{f(Z \mid \theta)} dZ \\ &= \int_{\Omega_Z} \frac{d\ln f(Z \mid \theta)}{d\theta} f(Z \mid \theta)\, dZ \\ &= \int_{\Omega_Z} s(Z \mid \theta)\, f(Z \mid \theta)\, dZ \end{aligned}$$

This is true for any $\theta \in \Theta$. In particular, it holds for θ_0:

$$0 = \int_{\Omega_Z} s(Z \mid \theta_0)\, f(Z \mid \theta_0)\, dZ$$

But observe that by the definition of an expection:

$$E[s(Z \mid \theta)] = \int_{\Omega_Z} s(Z \mid \theta)\, f(Z \mid \theta_0)\, dZ$$

The expectation depends on the true distribution of Z, which depends on the true θ_0. Combining the above results, we get that

$$E[s(Z \mid \theta_0)] = 0$$

and, by Assumption 4, it must also be that

$$E[s(Z \mid \theta)] \neq 0 \text{ for any } \theta \neq \theta_0 \text{ where } \theta \in \Theta.$$

So the mean of the s function is zero at the true θ_0, and by assumption must not equal zero for any other value of θ. Notice that since the *Score* function $= \sum_{i=1}^{n} s(Z_i \mid \theta)$, the same is true of the score function.

6.1.4 Hessian Matrix

In addition to the score vector, we will also use a Hessian matrix, defined as

$$H(Z \mid \theta) = \frac{ds(Z \mid \theta)}{d\theta'} = \frac{d^2 \ln f(Z \mid \theta)}{d\theta d\theta'}$$

The score function is a vector. When we take the derivative of a column vector $s(Z \mid \theta)$ with a row vector θ', we get a matrix. The j, k element of the matrix $ds(Z \mid \theta)/d\theta'$ is the derivative of the j'th element of the score vector $s(Z \mid \theta)$ with respect to the k'th element of θ. Since $s(Z \mid \theta)$ itself is the gradient vector $df(Z \mid \theta)/d\theta$, we get that the Hessian $H(Z \mid \theta)$ is a matrix of second derivatives.

We will now derive another connection between $H(Z \mid \theta)$ and $s(Z \mid \theta)$. Start where we left off with the score function

$$0 = \int_{\Omega_Z} s(Z \mid \theta) f(Z \mid \theta) dZ$$

And now take the derivative of this with respect to θ'.

$$\frac{d0}{d\theta'} = \frac{d}{d\theta'} \int_{\Omega_Z} s(Z \mid \theta) f(Z \mid \theta) dZ = \int_{\Omega_Z} \frac{d}{d\theta'} s(Z \mid \theta) f(Z \mid \theta) dZ$$

Note again there are no boundary terms when we take the derivative inside the integral above, because of Assumption 5. Next apply the rule for derivative of a product:

$$\begin{aligned} 0 &= \int_{\Omega_Z} \frac{ds(Z \mid \theta)}{d\theta'} f(Z \mid \theta) + s(Z \mid \theta) \frac{df(Z \mid \theta)}{d\theta'} dZ \\ &= \int_{\Omega_Z} H(Z \mid \theta) f(Z \mid \theta) + s(Z \mid \theta) \frac{d \ln f(Z \mid \theta)}{d\theta'} f(Z \mid \theta) dZ \\ &= \int_{\Omega_Z} H(Z \mid \theta) f(Z \mid \theta) + s(Z \mid \theta) s(Z \mid \theta)' f(Z \mid \theta) dZ \\ &= \int_{\Omega_Z} \left[H(Z \mid \theta) + s(Z \mid \theta) s(Z \mid \theta)'\right] f(Z \mid \theta) dZ \end{aligned}$$

Note we have again used the fact that $\frac{df(Z|\theta)}{d\theta} = s(Z \mid \theta) f(Z \mid \theta)$.

If $\theta = \theta_0$, then

$$\begin{aligned} 0 &= \int_{\Omega_Z} \left[H(Z \mid \theta_0) + s(Z \mid \theta_0) s(Z \mid \theta_0)'\right] f(Z \mid \theta_0) dZ \\ &= E\left[H(Z \mid \theta_0) + s(Z \mid \theta_0) s(Z \mid \theta_0)'\right] \end{aligned}$$

so that

$$-E(H(Z \mid \theta_0)) = E\left[s(Z \mid \theta_0) s(Z \mid \theta_0)'\right] = Var\left[s(Z \mid \theta_0)\right] = J_0$$

At the true θ_0, the Hessian equals the negative of the variance of the score function. Call this matrix J_0.

Assumption 6 says that J_0 is nonsingular.

6.1.5 Information matrix

Given a likelihood function, the information matrix $I_n(\theta_0)$ is defined by

$$I_n(\theta_0) = -E\left[\frac{\partial^2 \ln L(\theta_0 \mid Z_1, \ldots, Z_n)}{\partial\theta\partial\theta'}\right]$$

In the special case where the data $Z_1, \ldots, Z_n$ are i.i.d., the information matrix simplifies to

$$I_n(\theta_0) = -nE(H(Z \mid \theta_0))$$

so that

$$\frac{I_n(\theta_0)}{n} = J_0$$

6.1.6 Consistency of MLE

Back in lecture 3, we gave a general theorem for showing consistency of an extremum estimator. Here we verify that the assumptions we made regarding the MLE imply that the MLE is consistent.

Define $Q_n(\theta)$ and $Q_0(\theta)$ by

$$\begin{aligned} Q_n(\theta) &= \frac{1}{n}\ln L = \frac{1}{n}\sum\nolimits_{i=1}^{n} \ln f(Z_i \mid \theta) \\ Q_0(\theta) &= E[\ln f(Z \mid \theta)] \end{aligned}$$

So the MLE is an extremum estimator, with the estimator $\hat{\theta}_{ML}$ given by

$$\hat{\theta}_{ML} = \arg\max_{\theta} Q_n(\theta)$$

and by the LLN

$$Q_n(\theta) \xrightarrow{p} Q_0(\theta)$$

1. the identification condition requires that $Q_0(\theta)$ be uniquely maximized at the true θ_0. The first order condition (FOC) for maximizing $Q_0(\theta)$ is

$$\begin{aligned} 0 &= \frac{dQ_0(\theta)}{d\theta} \\ &= \frac{d}{d\theta}E[\ln f(Z \mid \theta)] = E[s(Z \mid \theta)] \end{aligned}$$

which we showed is satisfied at $\theta = \theta_0$, that is, we showed $E[s(Z \mid \theta_0)] = 0$. The second order condition (SOC) is that $\frac{d^2}{d\theta d\theta'}E[\ln f(Z \mid \theta_0)] = E[H(Z \mid \theta_0)]$ be negative definite. This holds because $E[H(Z \mid \theta_0)] = -Var[s(Z \mid \theta_0)]$ and variance matrices are positive semidefinite, and we assumed $E[H(Z \mid \theta_0)]$ is nonsingular, which means it's negative definite.

This shows that $Q_0(\theta)$ is maximized at θ_0. But is it uniquely maximized at θ_0? We have by Assumption 4 that

$$E[s(Z \mid \theta)] \neq 0 \text{ for any } \theta \neq \theta_0 \text{ where } \theta \in \Theta.$$

which means that no value of θ other than θ_0 satisfies the FOC.

2. Compactness: By assumption 2, we have $\theta \in \Theta$ and Θ is compact.

3. Smoothness: $Q_n(\theta)$ is differentiable, by Assumption 3 that $\ln f(Z \mid \theta)$ is differentiable.

4. Stochastic Equicontinuity / Uniform convergence: requires that $\sup_\theta \left|\frac{\partial Q_n(\theta)}{\partial \theta}\right| = O_p(1)$, where the δ and ε used to define the boundedness in probability does not depend on θ. Now by LLN (and plim of a function is the function of plim),

$$\left|\frac{1}{n}\sum\nolimits_{i=1}^{n} s(Z_i \mid \theta)\right| \xrightarrow{P} |E[s(Z \mid \theta)]|$$

and, by Assumption 3, derivatives being bounded means that $s(Z \mid \theta)$ and hence $|E[s(Z \mid \theta)]|$ is bounded for all θ. Together this implies not only boundedness in probability, but

$$\sup_\theta \left|\frac{\partial Q_n(\theta)}{\partial \theta}\right| = \text{constant } + o_p(1)$$

which is stronger than $O_p(1)$.

So all the conditions for consistency are satisfied, and we have $\hat{\theta}_{ML} \xrightarrow{P} \theta_0$

6.1.7 Asymptotic Normality of MLE

Given consistency of the MLE $\hat{\boldsymbol{\theta}}_{ML}$, we now derive its limiting distribution. To do so, we need to find an expression for $\sqrt{n}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0)$. By the definition of $\hat{\boldsymbol{\theta}}_{ML}$ and the first order condition for maximizing the likelihood function, we have that

$$\frac{1}{n}\sum\nolimits_{i=1}^{n} s(Z_i \mid \hat{\boldsymbol{\theta}}_{ML}) = 0$$

We first multiply this expresion by $\sqrt{n}$, and then invoke the mean value theorem (Taylor expansion) of $\hat{\boldsymbol{\theta}}_{ML}$ around $\boldsymbol{\theta}_0$. Using the fact that H is the derivative of s with respect to θ, we have

$$\frac{\sqrt{n}}{n}\sum\nolimits_{i=1}^{n}\left[s(Z_i \mid \boldsymbol{\theta}_0) + H(Z_i \mid \tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0)\right] = 0$$

where $\tilde{\boldsymbol{\theta}}$ lies between $\hat{\boldsymbol{\theta}}_{ML}$ and $\boldsymbol{\theta}_0$. Solving the above expression for $\sqrt{n}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0)$ gives:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0) = \underbrace{\left[-\frac{1}{n}\sum\nolimits_{i=1}^{n} H(Z_i \mid \tilde{\boldsymbol{\theta}})\right]^{-1}}_{\text{term 1}} \underbrace{\left[\frac{\sqrt{n}}{n}\sum\nolimits_{i=1}^{n} s(Z_i \mid \boldsymbol{\theta}_0)\right]}_{\text{term 2}}$$

To determine what $\sqrt{n}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0)$ converges to, we need to evaluate what term 1 and term 2 converge to.

Term 1: We cannot apply a LLN to $\frac{1}{n}\sum_{i=1}^{n} H(Z_i \mid \tilde{\boldsymbol{\theta}})$, because $\tilde{\boldsymbol{\theta}}$ is changing with n. Instead we can again use another mean value theorem (Taylor expansion). For some $\breve{\boldsymbol{\theta}}$ between $\boldsymbol{\theta}_0$ and $\tilde{\boldsymbol{\theta}}$ we have

$$-\frac{1}{n}\sum\nolimits_{i=1}^{n} H(Z_i \mid \tilde{\boldsymbol{\theta}}) = \underbrace{\left[-\frac{1}{n}\sum\nolimits_{i=1}^{n} H(Z_i \mid \boldsymbol{\theta}_0)\right]}_{\text{term 1a}} + \underbrace{\left[\frac{-1}{n}\sum\nolimits_{i=1}^{n}\sum\nolimits_{k=1}^{K}\frac{\partial H(Z_i \mid \breve{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_k}(\tilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_{0_k})\right]}_{\text{term 1b}}$$

Note the sums over the K elements of θ is needed because we are taking the derivative of each element of the matrix H with respect to each element of θ.

We can apply the LLN to term 1a to get

$$-\frac{1}{n}\sum\nolimits_{i=1}^{n} H(Z_i \mid \boldsymbol{\theta}_0) \xrightarrow{p} -E\left[H\left(Z \mid \boldsymbol{\theta}_0\right)\right] = J_0$$

For term 1b, we've assumed that third derivatives of f are all bounded, so $\left|\frac{\partial H(Z_i|\breve{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_k}\right| \leq c$ for some constant c. Therefore we have

$$|\frac{1}{n}\sum_i\sum_j \frac{\partial H(Z_i \mid \breve{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_k}(\tilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_{0_k})| \;\leq \frac{1}{n}\sum_i\sum_j c|\tilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_{0_k}| \;\xrightarrow{p} 0$$

where we have used the fact that $\hat{\boldsymbol{\theta}}_{ML}$ is consistent, so $|\hat{\boldsymbol{\theta}}_{MLk} - \boldsymbol{\theta}_{0_k}| \xrightarrow{p} 0$, and $\tilde{\boldsymbol{\theta}}_k$ lies between $\hat{\boldsymbol{\theta}}_{MLk}$ and $\boldsymbol{\theta}_{0_k}$, so $|\tilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_{0_k}| \xrightarrow{p} 0$. We have therefore shown that term 1b converges in probability to zero, or equivalently, that term 1b $= o_p(1)$.

Putting together term 1a and term 1b, we get

$$-\frac{1}{n}\sum\nolimits_{i=1}^{n} H(Z_i \mid \tilde{\boldsymbol{\theta}}) \xrightarrow{p} J_0$$

and, since the plim of a function is the function of the plims, for term 1 we get

$$\left[-\frac{1}{n}\sum\nolimits_{i=1}^{n} H(Z_i \mid \tilde{\boldsymbol{\theta}})\right]^{-1} \xrightarrow{p} J_0^{-1}$$

Term 2: For term 2, we have that Z_i is iid and therefore $s(Z_i \mid \boldsymbol{\theta}_0)$ is iid. By our assumptions it also satisfies the conditions to apply the Lindeberg Levy CLT, and using the fact that $var\left(s(Z_i \mid \boldsymbol{\theta}_0)\right) = J_0$, we get

$$\frac{\sqrt{n}}{n}\sum\nolimits_{i=1}^{n} s(Z_i \mid \boldsymbol{\theta}_0) \xrightarrow{d} N(0, J_0)$$

Putting terms 1 and 2 together, we get that $\sqrt{n}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0)$ converges to J_0^{-1} times a normal $N(0, J_0)$, and therefore

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, J_0^{-1}J_0J_0^{-1}) = N(0, J_0^{-1}).$$

Showing that MLE is root N CAN, with a limiting variance given by J_0^{-1}.

6.1.8 Estimating the variance

We have the limiting distribution $\sqrt{n}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, J_0^{-1})$. To construct tests and confidence intervals for $\hat{\boldsymbol{\theta}}_{ML}$, we therefore need a consistent estimate of J_0. There are three different estimator we can consider.

1. We have that $J_0 = -E\left[H(\mathbf{Z} \mid \boldsymbol{\theta}_0)\right]$, so we can use the negative sample average of the Hessian evaluated at $\hat{\boldsymbol{\theta}}$, call it $\overline{H}$, as an estimator.

$$\overline{H} = -\frac{1}{n}\sum\nolimits_{i=1}^{n} H(Z_i \mid \hat{\boldsymbol{\theta}})$$

2. We have that $J_0 = Var[S(\mathbf{Z} \mid \boldsymbol{\theta}_0)]$, so we can take the sample variance of the score function evaluated at $\hat{\boldsymbol{\theta}}$ as an estimator.

$$\widetilde{J} = \widehat{Var(S(Z_i} \mid \hat{\boldsymbol{\theta}}))$$

   Which of the above to choose depends partly on whether one calculated derivatives numerically or analytically. Since H involves second derivatives of the likelihood while S only involves first derivatives, one might prefer to use $\widetilde{J}$, especially if the computer is doing numerical derivatives.

3. Lastly we could combine these two estimates, using

$$\widehat{J}^{-1} = \overline{H}^{-1} \widetilde{J} \overline{H}^{-1}$$

   The advantage of this estimator is that it mimics the derivation of the MLE variance. The true limiting variance of the MLE was $[E(H)]^{-1} Var(S) [E(H)]^{-1}$ (where S and H are evaluated at the true θ_0) and it's just an algebraic coincidence that $-E(H)$ and $Var(S)$ at the true θ_0 both equal J_0.

   There is a class of estimators called Quasi-MLE. These are estimators in which the likelihood function is misspecified, but are consistent anyway (we will see later that OLS with non-normal errors is an example). For these estimators, $\widehat{J}^{-1}$ remains a consistent estimator of the limiting variance, while $\overline{H}^{-1}$ or $\widetilde{J}^{-1}$ is not. For these reasons, this third estimator is popular.

6.1.9 Other properties of MLE

asymptotic distribution Based on the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0)$, we have that the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ is

$$\hat{\boldsymbol{\theta}}_{ML} \overset{a}{\sim} N\left(\boldsymbol{\theta}_0, \frac{J_0^{-1}}{n}\right), \text{ From which we obtain } \hat{\boldsymbol{\theta}}_{ML} \overset{a}{\sim} N\left(\boldsymbol{\theta}_0, I_n(\boldsymbol{\theta}_0)^{-1}\right)$$

The asymptotic variance of $\hat{\boldsymbol{\theta}}_{ML}$ is the inverse of the information matrix. (this is generally true for non - iid data also, as long as a CLT holds). We can also plug in to estimate the variance, calculating standard errors using $\hat{\boldsymbol{\theta}}_{ML} \overset{a}{\sim} N\left(\boldsymbol{\theta}_0, I_n\left(\hat{\boldsymbol{\theta}}_{ML}\right)^{-1}\right)$.

Cramer-Rao Lower Bound The "Cramer-Rao Lower Bound" is defined as the smallest possible variance among regular, consistent estimators. An estimator that achieves this lower bound is therefore efficient (among all regular estimators).

It can be shown that the inverse of the information matrix, $I_n(\theta_0)^{-1}$, equals the "Cramer-Rao Lower Bound." This means that MLE is "asymptotically efficient" among regular and consistent estimators.

However, the variance of MLE may not exactly equal $I_n(\theta_0)^{-1}$ for any given sample size n. MLE only has this variance asymptotically, meaning that the larger n is, the closer MLE is to having this variance. So MLE might not be efficient, since it's possible that, for any given sample size n, a more efficient estimator (one that achieves the lower bound) is possible. But it does mean that MLE is asymptotically efficient, and so gets closer to being efficient as the sample size grows.

Invariance properties For any bounded, continuous function $g(\theta)$ (that doesn't depend on the sample size n) if $\gamma = g(\theta)$, then $\hat{\gamma}_{ML} = g\left(\hat{\theta}_{ML}\right)$. So functions of MLE's are themselves MLE. E.g., if we estimated a coefficient β by MLE, and we instead wanted to know $\gamma = \ln(\beta)$, then the maximum likelihood estimate of $\ln(\beta)$ would be $\hat{\gamma}_{ML} = \ln\left(\widehat{\beta}_{ML}\right)$.

If g is sufficiently smooth, then the limiting distribution of of $\hat{\gamma}_{ML}$ can be obtained from the limiting distribution of $\hat{\theta}_{ML}$ using the delta method.

bias and finite sample properties Note MLE is consistent, but it might also be biased. For example, if Z_i are iid with Z_i ˜ $N(\mu, \sigma^2)$, the MLE estimator for σ^2 is

$$\hat{\sigma}^2_{ML} = \frac{1}{n}\sum\nolimits_{i=1}^{n}(Z_i - \overline{Z})^2$$

which is biased. As we showed earlier, the unbiased estimator is $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \overline{Z})^2$.

More generally, ALL of the nice properties of MLE we have derived are asymptotic properties.

Connection to Bayesian estimation Bayes estimators treat θ as a random variable, and tries to estimates the distribution, not of an estimator $\hat{\theta}$, but of θ itself. By Bayes rule,

$$\Pr(\theta|\text{data}) = \frac{\Pr(\theta \text{ and data})}{\Pr(\text{data})} = \frac{\Pr(\theta)\Pr(\text{data}|\theta)}{\Pr(\text{data})}$$

$\Pr(\theta)$ is the prior distribution, i.e., what you thought the distribution of θ was before you saw any data. The goal is then to estimate the so-called posterior distribution $\Pr(\theta|\text{data})$, that is, what you think the distribution of θ is given both your prior and the data. Your actual estimate of θ could be, e.g., the mean or some other measure of central tendency of the posterior distribution.

The connection to maximum likelihood estimation is that $\Pr(\text{data}|\theta)$ is the likelihood function. So the more data you have, the more the Bayes estimator will resemble a narrow distribution near maximum likelihood estimate. For example, if the prior distribution is uniform, the mode of the posterior distribution will equal the MLE.

6.2 OLS and MLE

Consider the linear regression model $Y_i = X_i'\beta + e_i$.

To obtain the finite sample distribution of the OLS estimator $\widehat{\beta}$, we assumed the X's were constant (or we conditioned on them) and assumed the errors e_i were normal.

Question: could we use the normality assumption to construct a better estimator than OLS?

Let's see what the MLE $\widehat{\beta}_{ML}$ is when e_i are iid and e_i ˜ $N(0, \sigma^2)$.

If we condition on the X's, or assume them constant, then Y_i ˜ $N(X_i'\beta, \sigma^2)$. The vector of parameters θ of this distribution consists of both the vector β and the scalar σ^2.

We have $Z = (Y, X)$ where Y is random and X is either fixed, or is random but we condition on X. Assume the distribution of X does not depend on θ. let $f_{y,x}$, $f_{y|x}$, and f_x, denote, respectively, the joint distribution of Y and X, the conditional distribution of Y given X, and the marginal distribution of X. Then

$$f_{y,x}\left(Y_i, X_i \mid \theta\right) = f_{y|x}\left(Y_i \mid X_i, \theta\right) f_x\left(X_i, \mid \theta\right)$$

$$\ln f_{y,x}\left(Y_i, X_i \mid \theta\right) = \ln f_{y|x}\left(Y_i \mid X_i, \theta\right) + \ln f_x\left(X_i\right)$$

so

$$\frac{1}{n}\ln L\left(\theta \mid Z_1, \dots, Z_n\right) = \left[\frac{1}{n}\sum\nolimits_{i=1}^{n} \ln f_{y|x}\left(Y_i \mid X_i, \theta\right)\right] + \left[\frac{1}{n}\sum\nolimits_{i=1}^{n} \ln f_x\left(X_i\right)\right]$$

The second sum here doesn't depend on θ, so maximizing this likelihood function is the same as just maximizing $\frac{1}{n}\sum_{i=1}^{n} \ln f_{y|x}\left(Y_i \mid X_i, \theta\right)$. We can therefore ignore the distribution of X, and just define the likelihood function in terms of the conditional distribution of Y given X.

Now $f_{y|x}\left(Y_i, \mid X_i, \theta\right)$ is the normal density function $N\left(X_i'\beta, \sigma^2\right)$. So

$$\begin{aligned}\frac{1}{n}\sum\nolimits_{i=1}^{n} \ln f_{y|x}\left(Y_i \mid X_i, \theta\right) &= \frac{1}{n}\sum\nolimits_{i=1}^{n} \ln \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-\left(Y_i - X_i'\beta\right)^2}{2\sigma^2}} \\ &= \frac{1}{n}\sum\nolimits_{i=1}^{n} -\frac{\ln\left(2\pi\right)}{2} - \frac{\ln\left(\sigma^2\right)}{2} - \frac{\left(Y_i - X_i'\beta\right)^2}{2\sigma^2}\end{aligned}$$

We can see that maximizing this log likelihood expression with respect to β is equivalent to minimizing $\sum_{i=1}^{n}\left(Y_i - X_i'\beta\right)^2$, which then gives the OLS estimator $\widehat{\beta}$. So, with normal errors, we get $\widehat{\beta} = \widehat{\beta}_{ML}$, so

$$\widehat{\beta}_{ML} = \left(\sum\nolimits_{i=1}^{n} X_i X_i'\right)^{-1}\left(\sum\nolimits_{i=1}^{n} X_i Y_i\right)$$

Taking the derivative with respect to σ^2 and setting the result equal to zero (FOC for maximizing the likelihood) gives

$$\widehat{\sigma}_{ML}^2 = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left(Y_i - X_i'\widehat{\beta}_{ML}\right)^2$$

Implications:

We cannot get a better asymptotic estimator for β than OLS by using the assumption that the errors are normal.

If errors are normal, then OLS is asymptotically efficient (attaining the Cramer-Rao lower bound).

In practice, we can test if errors are normal, by testing if the residuals $\widehat{e}_i$ are normal (e.g. apply a Jarque-Bera test). If not normal, then OLS is still root-N-CAN, but is no longer MLE, and so might not be asymptotically efficient.

The MLE for σ^2 is not the same as our usual unbiased estimator s^2. Our usual estimator is the same except it divides by $n - K$ instead of n.

Why do we use s^2 and not $\widehat{\sigma}_{ML}^2$? Recall all of the useful features of MLE are asymptotic. Now $s^2 = \frac{n}{n-K}\widehat{\sigma}_{ML}^2$ and $\frac{n}{n-K} \to 1$ as $n \to \infty$. So s^2 and $\widehat{\sigma}_{ML}^2$ have the same asymptotic properties, e.g., s^2 is asymptotically efficient. But in addition, as we showed earlier s^2 is unbiased, while $E\left(\widehat{\sigma}_{ML}^2\right) = \frac{n-K}{n}\sigma^2$ is biased.

6.3 Binary Choice (Binomial Response) Models

6.3.1 Binary Choice Maximum Likelihood

Suppose Y_i is a dummy variable. It can only take values 0 or 1. Let $Z_i = (Y_i, X_i)$, and we observe Z_1,...,Z_n. Assume Y_i conditional on X_1,...,X_n are iid. Assume the model that, for some known function P and unknown parameter vector β_0,

$$\Pr\left(Y_i = 1 \mid X_1, X_2, \ldots, X_n\right) = P\left(X_i, \beta_0\right)$$

Notes

In the special case where there are no X's, and if we let the resulting constant $P\left(\beta_0\right) = p$, then this model reduces to the coin flipping example at the start of the Maximum Likelihood section.

We will do our analysis assuming P is any nice smooth function that you know. Some examples: If $P\left(X_i, \beta\right) = X_i'\beta$, this is called the Linear Probability Model, or LPM. If $P\left(X_i, \beta\right) = \Phi\left(X_i'\beta\right)$ where Φ is the cumulative normal distribution function, then this is called the Binary Probit model.

Now, we have

$$\begin{aligned} E\left(Y_i \mid X_1, \ldots, X_n\right) &= \Pr\left(Y_i = 1 \mid X_1, \ldots, X_n\right) \cdot 1 + \Pr\left(Y_i = 0 \mid X_1, \ldots, X_n\right) \cdot 0 \\ &= P\left(X_i, \beta_0\right) \end{aligned}$$

This shows that the expectation of a dummy variable is the same as the probability that the dummy variable equals one.

Consider maximum likelihood estimation of β. The likelihood function is given by

$$\begin{aligned} L &= \Pr\left(Y_1, \ldots, Y_n \mid X_1, \ldots, X_n, \beta\right) \\ &= \Pr\left(Y_1 \mid X_1, \beta\right) \Pr\left(Y_2 \mid X_2, \beta\right) \cdots \Pr\left(Y_n \mid X_n, \beta\right) \end{aligned}$$

so

$$\ln L = \sum\nolimits_{i=1}^{n} \ln\left[\Pr\left(Y_i \mid X_i, \beta\right)\right]$$

Note that since Y_i is discrete, the likelihood function is defined in terms of the probability mass function $\Pr\left(Y_i \mid X_i, \beta\right)$ rather than a probability density function. Now what is $\Pr\left(Y_i \mid X_i, \beta\right)$? The probability that $Y_i = 1$ is $P\left(X_i, \beta\right)$. And the probability that $Y_i = 0$ is $1 - P\left(X_i, \beta\right)$. Putting these together, we can say that

$$\Pr\left(Y_i \mid X_i, \beta\right) = P\left(X_i, \beta\right)^{Y_i} \left[1 - P\left(X_i, \beta\right)\right]^{1-Y_i}.$$

You can verify this expression is correct by plugging in $Y_i = 1$ or $Y_i = 0$ and see what you get. Plugging this expression for $\Pr\left(Y_i \mid X_i, \beta\right)$ into the log likelihood function gives

$$\ln L = \sum\nolimits_{i=1}^{n} \left[Y_i \ln P\left(X_i, \beta\right) + \left(1 - Y_i\right) \ln\left(1 - P\left(X_i, \beta\right)\right)\right]$$

Our MLE $\widehat{\beta}_{ML}$ is obtained by finding the value of β that maximizes this log likelihood function.

Now that we have the log likelihood function, we can now calculate the score function s, and from that, obtain the matrix J_0 (and the corresponding information matrix) which we need to calculate the limiting distribution for the MLE of β.

$$\begin{aligned} s\left(Z_i \mid \beta\right) &= \frac{\partial \ln\left[\Pr\left(Y_i \mid X_i,\beta\right)\right]}{\partial\beta} = \frac{\partial\left[Y_i \ln P\left(X_i,\beta\right) + \left(1-Y_i\right)\ln\left(1-P\left(X_i,\beta\right)\right)\right]}{\partial\beta} \\ &= \left[\frac{Y_i}{P\left(X_i,\beta\right)} - \frac{1-Y_i}{1-P\left(X_i,\beta\right)}\right]\frac{\partial P\left(X_i,\beta\right)}{\partial\beta} \end{aligned}$$

Note that the term in the square brackets is a scalar, while $\partial P\left(X_i,\beta\right)/\partial\beta$ is a gradient vector.

To simplify notation, let $P_{i0} = P(X_i,\beta_0)$. Then

$$\begin{aligned} J_0 &= E\left[s\left(Z \mid \beta_0\right)s\left(Z \mid \beta_0\right)'\right] \\ &= E\left(\left[\frac{Y_i^2}{P_{i0}^2} - 2\frac{Y_i}{P_{i0}}\frac{1-Y_i}{1-P_{i0}} + \frac{\left(1-Y_i\right)^2}{\left(1-P_{i0}\right)^2}\right]\frac{\partial P_{i0}}{\partial\beta}\frac{\partial P_{i0}}{\partial\beta'}\right) \end{aligned}$$

Again the term in the square brackets is a scalar, the rest is a matrix (a column gradient vector times a row gradient vector. Now, using the fact that Y_i can only equal zero or one, we have that $Y_i^2 = Y_i$, $Y_i\left(1-Y_i\right) = 0$, and $\left(1-Y_i\right)^2 = \left(1-Y_i\right)$, so the above simplifies to

$$\begin{aligned} J_0 &= E\left(\left[\frac{Y_i}{P_{i0}^2} + \frac{\left(1-Y_i\right)}{\left(1-P_{i0}\right)^2}\right]\frac{\partial P_{i0}}{\partial\beta}\frac{\partial P_{i0}}{\partial\beta'}\right) \\ &= E\left(E\left(\left[\frac{Y_i}{P_{i0}^2} + \frac{\left(1-Y_i\right)}{\left(1-P_{i0}\right)^2}\right]\frac{\partial P_{i0}}{\partial\beta}\frac{\partial P_{i0}}{\partial\beta'} \mid X_1, X_2, \ldots, X_n\right)\right) \\ &= E\left(\left[\frac{E\left(Y_i \mid X_1, X_2, \ldots, X_n\right)}{P_{i0}^2} + \frac{1-E\left(Y_i \mid X_1, X_2, \ldots, X_n\right)}{\left(1-P_{i0}\right)^2}\right]\frac{\partial P_{i0}}{\partial\beta}\frac{\partial P_{i0}}{\partial\beta'}\right) \end{aligned}$$

where the second equality above comes from applying the law of iterated expectations. Next, plug in $E\left(Y_i \mid X_1, X_2, \ldots, X_n\right) = P\left(X_i,\beta_0\right) = P_{i0}$ to get

$$\begin{aligned} J_0 &= E\left(\left[\frac{P_{i0}}{P_{i0}^2} + \frac{1-P_{i0}}{\left(1-P_{i0}\right)^2}\right]\frac{\partial P_{i0}}{\partial\beta}\frac{\partial P_{i0}}{\partial\beta'}\right) \\ & E\left(\frac{1}{P_{i0}\left[1-P_{i0}\right]}\frac{\partial P_{i0}}{\partial\beta}\frac{\partial P_{i0}}{\partial\beta'}\right) \end{aligned}$$

So, given a function P, we can plug it into the above expression to calculate J_0, and thereby obtain the asymptotic distribution $\sqrt{n}(\widehat{\beta}_{ML} - \beta_0) \xrightarrow{d} N(0, J_0^{-1})$. To consistently estimate J_0, we have

$$\widehat{J}_0 = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left[\frac{1}{P\left(X_i,\widehat{\beta}_{ML}\right)\left[1-P\left(X_i,\widehat{\beta}_{ML}\right)\right]}\right]\frac{\partial P\left(X_i,\widehat{\beta}_{ML}\right)}{\partial\beta}\frac{\partial P\left(X_i,\widehat{\beta}_{ML}\right)}{\partial\beta'}$$

and so

$$\widehat{\beta}_{ML} \overset{a}{\sim} N\left(\beta_0, \frac{\widehat{J}_0^{-1}}{n}\right)$$

which we can use for asymptotic standard errors, inference, and confidence intervals.

6.3.2 Threshold Crossing Models

A popular class of binary choice models are called "Threshold Crossing" models. These models take the form

$$Y_i = I\left(X_i'\beta + e_i \geq 0\right)$$

For some unobserved error e_i. More precisely, this is the "Linear Index Threshold Crossing" model, where $X_i'\beta$ is the linear index. One could have more general threshold crossing models where this is replaced with some nonlinear function of covariates and parameters.

The idea of the threshold crossing model is that $X_i'\beta + e_i$ is an indicator of how much you like choice $Y_i = 1$. If you like this choice enough so that $X_i'\beta + e_i$ is positive, you choose $Y_i = 1$, other wise choose $Y_i = 0$. For example, if Y_i is the decision to buy a car, then $X_i'\beta + e_i$ could be the your willingness to pay for the car minus its price, or the benefits of owning the car minus the costs of owning it. $X_i'\beta + e_i$ could also equal the utility of choosing $Y_i = 1$ minus the utility of choosing $Y_i = 0$.

Suppose we have a threshold crossing model, and assume that e_i is independent of X_i. Then

$$\begin{aligned} P\left(X_i, \beta_0\right) &= \Pr\left(Y_i = 1 \mid X_i\right) = \Pr\left(X_i\beta_0 + e_i \geq 0 \mid X_i\right) \\ &= \Pr\left(-e_i \leq X_i'\beta_0 \mid X_i\right) = \Pr\left(-e_i \leq X_i'\beta_0\right) \\ &= F\left(x_i'\beta_0\right) \end{aligned}$$

Where F is the cumulative distribution function of $-e_i$. The middle line above holds because we assumed e_i is independent of X_i. This shows that the threshold crossing model is a special case of our general binary choice model where the function $P\left(X_i, \beta\right) = F\left(X_i'\beta\right)$.

If e_i follows normal distribution, this is called a probit model.
If e_i has logistic distribution, this called logit model.

Our general formula for the score function s and the matrix J_0 simplifies a little when we have a linear index threshold crossing model. Let f be the probability density function of $-e_i$. We had

$$\begin{aligned} J_0 &= E\left(\frac{1}{P\left(X_i, \beta_0\right)\left[1 - P\left(X_i, \beta_0\right)\right]}\frac{\partial P\left(X_i, \beta_0\right)}{\partial \beta}\frac{\partial P\left(X_i, \beta_0\right)}{\partial \beta'}\right) \\ &= E\left(\frac{1}{F\left(X_i'\beta_0\right)\left[1 - F\left(X_i'\beta_0\right)\right]}\frac{\partial F\left(X_i'\beta_0\right)}{\partial \beta}\frac{\partial F\left(X_i'\beta_0\right)}{\partial \beta'}\right) \\ &= E\left(\frac{1}{F\left(X_i'\beta_0\right)\left[1 - F\left(X_i'\beta_0\right)\right]}\left[f\left(X_i'\beta\right)X_i\right]\left[f\left(X_i'\beta\right)X_i\right]'\right) \\ & \quad E\left(\frac{\left[f\left(X_i'\beta_0\right)\right]^2}{F\left(X_i'\beta_0\right)\left[1 - F\left(X_i'\beta_0\right)\right]}X_iX_i'\right) \end{aligned}$$

$$\widehat{J}_0 = \frac{1}{n}\sum_{i=1}^{n} w_i^2 X_i X_i' \quad \text{where} \quad w_i^2 = \frac{\left[f\left(X_i'\widehat{\beta}_{ML}\right)\right]^2}{F\left(X_i'\widehat{\beta}_{ML}\right)\left[1 - F\left(X_i'\widehat{\beta}_{ML}\right)\right]}$$

Note the similarity of the limiting distribution to that of linear weighted least squares, with weights given by the scalar w_i.

7 Lecture 07. Inference and Hypothesis Tests

Readings for this lecture are: Greene Chapters 5, and 14.6.

7.1 Linear Restrictions

Suppose we have an estimator $\hat{\beta}$ of some k vector β. Suppose $\hat{\beta}$ is normally distributed, so that $\left(\hat{\beta}-\beta_0\right) \,\tilde{}\, N(0,V)$ for some variance matrix V. Note β doesn't need to be regression coefficients, it could be any estimated vector. The key is that we are assuming $\hat{\beta}$ is exactly normal (not just asymptotically normal, though we'll get to that later). So if $\hat{\beta}$ is a vector of regression coefficients, then we need the errors to be normal.

We wish to test a set of q linear restrictions regarding β. The number of restrictions q is the number of equal signs needed to write the null hypothesis. For example, suppose want to test $H_0 : \beta_2 = 7$ and $\beta_4 = 2\beta_3$. In this example, $q = 2$.

We can write any null hypothesis of linear restrictions like these as

$$H_0 : A\beta = C$$

where A is a $q \times k$ matrix and C is a $q \times 1$ vector.

Example: Suppose $k = 4$, and want to test $H_0 : \beta_2 = 7$ and $\beta_4 = 2\beta_3$. Then writing this as $A\beta = C$ gives

$$H_0 : \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & -2 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} 7 \\ 0 \end{bmatrix}.$$

A property of normals is that linear functions of normals are normal, so $\hat{\beta} \,\tilde{}\, N(\beta_0, V)$ means that

$$\begin{aligned} \left(A\hat{\beta}-C\right) &\,\tilde{}\, N\left(E\left(A\hat{\beta}-C\right), Var\left(\left(A\hat{\beta}-C\right)\right)\right) \\ \left(A\hat{\beta}-C\right) &\,\tilde{}\, N\left(A\beta_0 - C, AVA'\right) \end{aligned}$$

So in particular, if H_0 is true then $\left(A\hat{\beta}-C\right) \,\tilde{}\, N(0, AVA')$.

Another property of normals is: If θ is a q vector with $\theta \,\tilde{}\, N(0,\Omega)$, then $\theta'\Omega^{-1}\theta \sim \chi^2_q$, a chi-squared distribution with q degrees of freedom. This means that:

$$\begin{aligned} \text{If } H_0 \text{ is true then } \widehat{W} &\sim \chi^2_q \\ \text{where } \widehat{W} &= \left(A\hat{\beta}-C\right)'\left(AVA'\right)^{-1}\left(A\hat{\beta}-C\right) \end{aligned}$$

So $\widehat{W}$ is a test statistic. Given our estimate $\hat{\beta}$, we can calculate $\widehat{W}$. If the value of $\widehat{W}$ is large, and so very unlikely to be a draw from a χ^2_q distribution, then we reject H_0, and if $\widehat{W}$ is small enough, then we fail to reject. This type of test is an example of what is called a Wald test.

Problem: usually we don't know V, and we then need to estimate it. But if we replace V with an estimate $\widehat{V}$ in the definition of $\widehat{W}$, then that changes the distribution of $\widehat{W}$ (because now, instead of just $\hat{\beta}$ being random in the expression for $\widehat{W}$, we now have two random components, both $\hat{\beta}$ and $\widehat{V}$.

There are two solutions, finite sample and asymptotic. In finite samples, we need to use properties of the distribution of $\widehat{V}$ to calculate the distribution of $\widehat{W}$.

Alternatively, we can have conditions hold that make the randomness in $\widehat{V}$ be asymptotically irrelevant, so that $\widehat{W}$ is asymptotically chi-squared. Asymptotically, it's also not necessary for $\hat{\beta}$ to be normal, sufficient is that $\hat{\beta}$ converge in distribution to a normal. And we can handle tests of nonlinear restrictions asymptotically, using the delta method.

Below we'll see examples of each.

7.2 Finite Sample Linear Regression Tests - The F test

Let the model be $Y = X\beta + e$, let the classical Gauss-Markov Assumptions hold, and assume the errors e are normally distributed. We want to test $H_0 : A\beta = C$. Consider the usual OLS estimator $\hat{\beta} = (X'X)^{-1} X'Y$.

Then $\left(\hat{\beta} - \beta_0\right) \tilde{} \; N(0, V)$ where $V = \sigma^2 (X'X)^{-1}$. Plugging in this V we get

$$\begin{aligned} \text{If } H_0 \text{ is true then } \widehat{W} &\sim \chi^2_q \\ \text{where } \widehat{W} &= \frac{\left(A\hat{\beta} - C\right)' \left(A (X'X)^{-1} A'\right)^{-1} \left(A\hat{\beta} - C\right)}{\sigma^2} \end{aligned}$$

We don't know σ^2, but we have our usual estimator

$$s^2 = \frac{\left(Y - X\hat{\beta}\right)' \left(Y - X\hat{\beta}\right)}{n - k}$$

Since the errors in the regression are normal, the sum of squared errors in the numerator of s^2 is, after suitable scaling, itself chi-squared distributed.

So, if we replace σ^2 with s^2 in the test statistic, and suitably rescale, we get a new test statistic, called an F-statistic:

$$\begin{aligned} \text{If } H_0 \text{ is true then } \widehat{F} &\sim F_{q,n-k} \\ \text{where } \widehat{F} &= \frac{\left(A\hat{\beta} - C\right)' \left(A (X'X)^{-1} A'\right)^{-1} \left(A\hat{\beta} - C\right)/q}{s^2} \end{aligned}$$

An F statistic is the ratio of two independent chi-squared statistics, each scaled by their own degrees of freedom. So an $F_{q,n-k}$ statistic equals a χ^2_q/q divided by a $\chi^2_{n-k}/(n-k)$. Note the $n-k$ in the above definition of $\widehat{F}$ is inside s^2).

We can therefore construct the statistic $\widehat{F}$ and use it to test H_0 by seeing if the value of $\widehat{F}$ is one that is reasonable to be a draw from an $F_{q,n-k}$ distribution.

This F test is a special case of a general class of tests called Wald tests.

Note: Suppose we wanted to test if single coefficient takes some value, e.g., test H_0: $\beta_1 = 7$. Then we could construct our usual t-statistic: $\widehat{t} = \left(\widehat{\beta}_1 - 7\right)/s_1$ where s_1 is the standard error of $\widehat{\beta}_1$. Or alternatively, we could construct the above F test. How do these compare?

Answer: both will yield the same test result, with the same p-values. In fact, we will get $\widehat{F} = \widehat{t}^2$, that is, the F statistic will equal the square of the t statistic. And the F statistic $F_{1,n-k}$, has the same distribution as the square of t-statistic with $n - k$ degrees of freedom.

7.3 Another F-test

There is another way to construct an F test in linear regressions. Continue to make the same assumptions: $Y = X\beta + e$, the classical Gauss-Markov Assumptions hold, and the errors e are normally distributed. We want to test $H_0 : A\beta = C$.

Suppose we believed H_0 was true. Then we might want to estimate β imposing the restriction that$A\beta = C$. That is, we could estimate $\hat{\beta}_r$ (the subscript r stands for 'restricted'), defined by

$$\hat{\beta}_r = \arg\min \left(Y - X\beta\right)' \left(Y - X\beta\right), \text{ such that } A\beta = C$$

This minimization could be done using Lagrange multipliers. We could then define the n vector of restricted residuals $\hat{e}_r = Y - X\hat{\beta}_r$.

Let us also define the n vector of unrestricted residuals $\hat{e}_u = Y - X\hat{\beta}$, where $\hat{\beta} = \left(X'X\right)^{-1} X'Y$ is the usual unrestricted coefficients estimator.

By construction, the unrestricted sum of squared residuals $\hat{e}_u'\hat{e}_u = \sum_i \hat{e}_{ui}^2$ must be smaller than or equal to the restricted sum of squared residuals $\hat{e}_r'\hat{e}_r$.

Intuitively, if H_0 is true so the restriction is valid, then the difference between $\hat{e}_u'\hat{e}_u$ and $\hat{e}_r'\hat{e}_r$ should be random and will disappear asymptotically. Formally, we may construct an F test based on this intuition

$$\hat{F} = \frac{\sum_i \left[\hat{e}_{ri}^2 - \hat{e}_{ui}^2\right]/q}{\sum_i \hat{e}_{ui}^2/\left(n-k\right)}$$

It can be shown that if H_0 is true then this $\hat{F} \sim F_{q,n-k}$.

Observe that the denominator of this $\hat{F}$ is just s^2 from the unrestricted model. The numerator and denominator are essentially sums of squares normals, and so have chi-squared distributions. If H_0 is true then these are independent chi-squareds, giving us an F distribution. But if not, then these chi-squareds are correlated.

This test is a special case of what is called a Likelihood Ratio (LR) test.

7.4 Asymptotic Tests

Instead of focusing on linear restrictions, and requiring normality, we will now allow for testing general nonlinear restrictions, and only assume that our estimator is asymptotically normal.

Assume we are estimating a general k vector of parameters θ, and that we have an (unrestricted) estimator $\hat{\theta}_u$ that is $\sqrt{n} - CAN$, so $\sqrt{n}\left(\hat{\theta}_u - \theta_0\right) \xrightarrow{d} N(0, V)$.

The null hypothesis now is

$$H_0 : g(\theta_0) = 0$$

where g is a q-vector valued, continuously differentiable function in a neighborhood of θ_0 and has a derivatives that, at θ_0, are finite and nonzero.

We will construct three different kinds of tests of H_0, known as Wald tests, LR (Likelihood ratio) tests, and LM (Lagrange Multiplier) or efficient score tests.

7.4.1 Wald Tests

Given our assumptions, $\hat{\theta}_u$ is $\sqrt{n} - CAN$ with $\sqrt{n}\left(\hat{\theta}_u - \theta_0\right) \xrightarrow{d} N(0, V)$, and we can apply the Delta method to say that $g\left(\hat{\theta}_u\right)$ is also $\sqrt{n} - CAN$. So, if $H_0 : g(\theta_0) = 0$ is true then $\sqrt{n}g\left(\hat{\theta}_u\right)$ converges in distribution to a mean zero normal. Therefore $\widetilde{W} \overset{a}{\sim} \chi^2_q$ where, using the Delta method to calculate the variance,

$$\begin{aligned}\widetilde{W} &= g\left(\hat{\theta}_u\right) Var\left(g\left(\hat{\theta}_u\right)\right)^{-1} g\left(\hat{\theta}_u\right) \\ &= g\left(\hat{\theta}_u\right) \left[\frac{dg(\theta_0)}{d\theta} \frac{V}{n} \frac{dg(\theta_0)'}{d\theta}\right]^{-1} g\left(\hat{\theta}_u\right).\end{aligned}$$

We don't know V, but assume we can replace V with a $\sqrt{n}$ consistent estimate $\widehat{V}$. Similarly, we don't know θ_0, but we can estimate the gradient $dg(\theta_0)/d\theta$ using $dg\left(\hat{\theta}_u\right)/d\theta$. This gives us the test statistic

$$\widehat{W} = g\left(\hat{\theta}_u\right) \left[\frac{dg\left(\hat{\theta}_u\right)}{d\theta} \frac{\widehat{V}}{n} \frac{dg\left(\hat{\theta}_u\right)'}{d\theta}\right]^{-1} g\left(\hat{\theta}_u\right)$$

Suppose, as is usually the case, we can show that (under the null hypothesis) $\widehat{W} = \widetilde{W} + o_p(1)$. We'll have that, if H_0 is true, then $\widehat{W} \overset{a}{\sim} \chi^2_q$. This $\widehat{W}$ is called a Wald statistic, and this test is a Wald test.

Notes on Wald tests:

1. In the special case of linear restrictions in a linear regression, the first F statistic we gave earlier is essentially the same as this Wald statistic. Recall that an $F_{q,n-k}$ statistic equals a χ^2_q/q divided by a $\chi^2_{n-k}/(n-k)$. As $n \to \infty$ this denominator becomes a constant, and so, apart from scaling, the F statistic becomes a χ^2_q statistic asymptotically. So the finite sample F test based on $\widehat{F}$ earlier asymptotically becomes equivalent to the Wald chi-squared statistic. Our usual t-tests are also asymptotically equivalent to Wald tests.

2. The Wald test is based on the distribution under the null of the unrestricted estimate $\hat{\theta}_u$. It does not require obtaining a restricted estimate. For example, when we do a t-test checking if a coefficient is zero, we don't need to reestimate the model imposing the restriction that the coefficient is zero. This feature of Wald tests may be convenient when estimation of θ while imposing the restriction $g(\theta) = 0$ is difficult. For example, $\hat{\theta}_u$ could be linear OLS regression coefficients, and if $g(\theta)$ is nonlinear, then estimation of θ imposing the restriction would require doing nonlinear least squares.

3. The Wald test has a drawback of non-invariance to reparameterization. For example, let θ_{01} be the first element of θ_0, and suppose the null hypothesis is $\theta_{01} = 1$, so $g(\theta_0) = \theta_{01} - 1$. Another way to express the same restriction would be $\ln(\theta_{01}) = 0$, so $g(\theta_0) = \ln(\theta_{01})$. In general, we could get different values of the test statistic $\widehat{W}$ using these two different ways of expressing the null hypothesis H_0: $g(\theta_0) = 0$. It is possible we could reject writing g one way and fail to reject writing it the other way. Note this is just a finite sample problem - asymptotically these tests will agree.

4. There is often more than one way to estimate V. Different estimators $\widehat{V}$ will also give different values for the test statistic $\widehat{W}$. For example, if θ is a vector of coefficients in a linear regression, we might use our standard formula for V, or the White (heteroskedasticity rebust) formula for V.

7.4.2 Likelihood Ratio (LR) tests

Suppose we obtain an unrestricted estimate $\hat{\theta}_u$ by maximizing a log likelihood function $\ln L(\theta)$. Suppose we wanted to test the null hypothesis $H_0 : \theta_0 = \widetilde{\theta}$ for some given value $\widetilde{\theta}$. Consider a Taylor expansion of $\ln L\left(\widetilde{\theta}\right)$ around $\hat{\theta}_u$:

$$\ln L\left(\widetilde{\theta}\right) \approx \ln L\left(\hat{\theta}_u\right) + \frac{\partial \ln L\left(\hat{\theta}_u\right)}{\partial \theta}' \left(\widetilde{\theta} - \hat{\theta}_u\right) + \frac{1}{2}\left(\widetilde{\theta} - \hat{\theta}_u\right)' \frac{\partial^2 \ln L\left(\hat{\theta}_u\right)}{\partial \theta \partial \theta'} \left(\widetilde{\theta} - \hat{\theta}_u\right)$$

Now $\frac{\partial \ln L\left(\hat{\theta}_u\right)}{\partial \theta} = 0$ is the score function, and $\frac{-\partial^2 \ln L\left(\hat{\theta}_u\right)}{\partial \theta \partial \theta'}$ is the estimated information matrix, so $\frac{-\partial^2 \ln L\left(\hat{\theta}_u\right)}{\partial \theta \partial \theta'} = \left(\widehat{V}/n\right)^{-1}$. We therefore get

$$2\left[\ln L\left(\hat{\theta}_u\right) - \ln L\left(\widetilde{\theta}\right)\right] \approx \left(\hat{\theta}_u - \widetilde{\theta}\right)' \left(\widehat{V}/n\right)^{-1} \left(\hat{\theta}_u - \widetilde{\theta}\right)$$

Observe that the term on the right equals the Wald statistic for testing the null hypothesis $H_0 : \theta_0 = \widetilde{\theta}$. The term on the left is an alternative test statistic that, as this derivation suggests, may be asymptotically the same as the Wald statistic. To prove this, we'd need to show the remainder term in the above expansion is asymptotically negligible.

More generally, suppose we want to test the null hypothesis $H_0 : g(\theta) = 0$. We already have the unrestricted estimate

$$\hat{\theta}_u = \arg\max \ln L(\theta).$$

We can define a corresponding restricted estimate $\hat{\theta}_r$ defined by

$$\hat{\theta}_r = \arg\max \ln L(\theta) \text{ such that } g\left(\hat{\theta}_r\right) = 0.$$

So the restricted estimator $\hat{\theta}_r$ is the vector θ that maximizes the liklihood function under the constraint that $\hat{\theta}_r$ satisfies $g\left(\hat{\theta}_r\right) = 0$.

Using an extension of the above derivation called Wilk's Theorem, it can be shown more generally that, if the null hypothesis $H_0 : g(\theta) = 0$ is true, then

$$2\left[\ln L\left(\hat{\theta}_u\right) - \ln L\left(\hat{\theta}_r\right)\right] \overset{a}{\sim} \chi^2_q$$

This is statistic is called the Likelihood Ratio (or LR) statistic, noting that it can equivalently be written as

$$2 \ln \left(\frac{L\left(\hat{\theta}_u\right)}{L\left(\hat{\theta}_r\right)} \right)$$

Some intuition: If the null hypothesis $g(\theta) = 0$ is true, then imposing the estimate satisfy this constraint shouldn't hurt the maximum attainable value for the likelihood function, meaning that $\ln L\left(\hat{\theta}_u\right)$ shouldn't be much larger than $\ln L\left(\hat{\theta}_r\right)$. Asymptotically, the difference between the two will be zero if the null is true. So if the difference between $\ln L\left(\hat{\theta}_u\right)$ and $\ln L\left(\hat{\theta}_r\right)$ is large (improbably large for a χ^2_q statistic) then we reject the null, otherwise we fail to reject.

Notes on LR tests:

1. In the special case of linear restrictions in a linear regression with normal iid errors, the second F statistic we gave earlier (involving restricted and unrestricted sums of squared residuals) is asymptotically equivalent to an LR statistic. In particular, in linear regression with normal iid errors we have the log likelihood function

$$\ln L(\theta) = \sum\nolimits_{i=1}^{n} -\frac{\ln(2\pi)}{2} - \frac{\ln(\sigma^2)}{2} - \frac{(Y_i - X_i'\beta)^2}{2\sigma^2}$$

where $\theta = \{\beta, \sigma^2\}$, so

$$2\left[\ln L\left(\hat{\theta}_u\right) - \ln L\left(\hat{\theta}_r\right)\right] = \sum\nolimits_{i=1}^{n} \frac{e_{ir}^2}{\widehat{\sigma}_r^2} - \frac{e_{iu}^2}{\widehat{\sigma}_u^2} + \ln\left(\widehat{\sigma}_r^2\right) - \ln\left(\widehat{\sigma}_u^2\right)$$

Now if the null hypothesis is true then $\widehat{\sigma}_r^2$ and $\widehat{\sigma}_u^2$ are both consistent and therefore asymptotically equivalent, and so we can replace both with any consistent estimator like the unrestricted s^2. With that replacement, this LR statistic becomes equivalent, up scaling by q, to the F statistic based on restricted and unrestricted sums of squared errors we gave earlier.

2. The LR test requires estimating both the restricted and unrestricted model. However, since each model is estimated by maximizing the likelihood function, just by doing the estimation you will already have calculated the terms $L\left(\hat{\theta}_u\right)$ and $L\left(\hat{\theta}_r\right)$ needed to construct the LR statistic.

3. The LR test does not have the "non-invariance to reparameterization" problem of the Wald test.

4. Tests analogous to the LR test can be constructed for general extremum estimators. Suppose

$$\begin{aligned}\hat{\theta}_u &= \arg\max Q(\theta) \quad \text{and} \\ \hat{\theta}_r &= \arg\max Q(\theta) \text{ such that } g\left(\hat{\theta}_r\right) = 0.\end{aligned}$$

For some general (smooth, positive) objective function $Q(\theta)$. Then if $H_0 : g(\theta) = 0$ is true we get

$$\left[\ln Q\left(\hat{\theta}_u\right) - \ln Q\left(\hat{\theta}_r\right)\right] c \overset{a}{\sim} \chi^2_q$$

for some constant c. The value of c (which is 2 for likelihood functions) will depend on the particular extremum estimator.

7.4.3 Lagrange Multiplier (LM) and Efficient Score Tests

How does one actually calculate the restricted estimate $\hat{\theta}_R$? To maximize a function $Q(\theta)$ subject to equality constraints $g(\theta) = 0$, We use Lagrange multipliers. The Lagrangian is:

$$L = Q(\theta) - \lambda' g(\theta)$$

where λ is a $q-$ vector of lagrange multipliers. $\hat{\theta}_R$ then satisfies the first order conditions

$$\begin{aligned}\frac{\partial L}{\partial \theta} &= \frac{\partial Q\left(\hat{\theta}_R\right)}{\partial \theta} - \lambda' \frac{\partial g\left(\hat{\theta}_R\right)}{\partial \theta} = 0 \\ \frac{\partial L}{\partial \lambda} &= g\left(\hat{\theta}_R\right) = 0\end{aligned}$$

If the null hypothesis $H_0 : g(\theta) = 0$ is true, then asymptotically the constraint should not be binding, so asymptotically $\lambda \to 0$. So, assuming asymptotic normality, we should get

$$\frac{\partial Q\left(\hat{\theta}_R\right)}{\partial \theta} \xrightarrow{d} N\left(0, Var\left(\frac{\partial Q\left(\hat{\theta}_R\right)}{\partial \theta}\right)\right)$$

and therefore, if the null hypothesis is true,

$$\frac{\partial Q\left(\hat{\theta}_R\right)}{\partial \theta}' \left[\widehat{Var\left(\frac{\partial Q\left(\hat{\theta}_R\right)}{\partial \theta}\right)}\right]^{-1} \frac{\partial Q\left(\hat{\theta}_R\right)}{\partial \theta} \overset{a}{\sim} \chi^2_q$$

This is called the Lagrange Multiplier (or LM) test. It can be equivalently expressed in terms of testing if the estimate of λ is zero.

In the special case where Q is a log likelihood function, and we are testing $H_0 : \theta = \theta_0$, then $\hat{\theta}_R = \theta_0$ and we get

$$\frac{\partial \ln L\left(\hat{\theta}_R\right)}{\partial \theta}' \left[I\left(\hat{\theta}_R\right)\right]^{-1} \frac{\partial \ln L\left(\hat{\theta}_R\right)}{\partial \theta} \overset{a}{\sim} \chi^2_q$$

where $I\left(\hat{\theta}_R\right)$ is the restricted estimate of the information matrix. Recall that $\partial \ln L(\theta) / \partial \theta$ is the score function. So, if the null hypothesis is true, then $\hat{\theta}_R$ is an asymptotically efficient estimate of

θ, and therefore $\partial \ln L\left(\hat{\theta}_R\right)/\partial\theta$ is an asymptotically efficient estimate of the score function, which is why this test is also called the efficient score test.

Notes on LM tests:

1. The LM test only requires estimating the restricted model. This is sometimes convenient because it may require many fewer parameters than the unrestricted model, particularly when q is large.

2. Like the Wald test, there may be many different ways to estimate the variance, each of which would yield a valid LM test, even though the test statistics could be numerically different.

7.4.4 Which test?

The Wald, LR, and LM tests all have the same asymptotic distribution χ_q^2. But they can be numerically different. And even more numerically different Wald and LM tests with the same limit distribution are possible, by using different consistent estimators of the limiting variance. It is entirely possible that, with the same data, same model, and same null hypothesis, some of these tests could reject the null and others fail to reject.

Are some of these tests better than others? No. Sometimes Wald is the most accurate, sometimes LR, and sometimes LM, and in general there's no way to know.

Best practice is to use more than one test. If they disagree, then you just have to recognize greater uncertainty in whether to reject the null or not.