ECON 241B

(Hopefully) Helpful Notes

Note: This material is purely supplemental, and is not a substitute for the actual course material. This does not include all course material, and contains additional material not necessary for the purposes of this course. If you find any typos please let me know at sherrard@ucsb.edu.

Section 1:

Tools:

Law(s) of Iterated Expectations:
If $\mathbb{E}[y] < \infty$:

$$\mathbb{E}[\mathbb{E}(y|\mathbf{x})] = \mathbb{E}[y]$$

$$\mathbb{E}[\mathbb{E}(y|\mathbf{x}_1, \mathbf{x}_2)|\mathbf{x}_1] = \mathbb{E}[y|\mathbf{x}_1]$$

Note: Remember that if there are multiple sets of conditioning variable the **smallest** information set wins.

Conditioning Theorem:
If $\mathbb{E}[y] < \infty$:

$$\mathbb{E}[g(\mathbf{x})y|\mathbf{x}] = g(\mathbf{x})\,\mathbb{E}[y|\mathbf{x}]$$

If, in addition, $\mathbb{E}[g(\mathbf{x})y] < \infty$ then

$$\mathbb{E}[g(\mathbf{x})y] = \mathbb{E}[g(\mathbf{x})\,\mathbb{E}(y|\mathbf{x})]$$

Material:

Conditional Expectation Function (CEF):

$$\mathbb{E}[y|x_1, x_2, ..., x_k] = m(x_1, x_2, ..., x_k) = \int_{-\infty}^{\infty} y f_{y|\mathbf{x}}(y|\mathbf{x})$$

where

$$f_{y|\mathbf{x}}(y|\mathbf{x}) = \frac{f(y, \mathbf{x})}{f_{\mathbf{x}}(\mathbf{x})}$$

Intuitively you can think of the conditional expectation as the mean of y at a fixed point of x. You are taking a (normalized) slice out of the contours of the joint density function at a fixed value of x and taking the mean of that.

Questions:

i Under what conditions is the CEF a random variable?

ii When can we be sure that m($\mathbf{x}$) exists?

Answers:

i It depends on x. If x is a random variable then the CEF is. Once x is fixed the CEF becomes fixed as well. e.g: $\mathbb{E}[wage|GradStudent] = 0$.

ii When $\mathbb{E}[y] < \infty$

CEF Error:
The CEF error is defined as the difference between y and the CEF evaluated at $\mathbf{x}$

$$e = y - m(\mathbf{x})$$

Note: Remember that the error is derived from the joint distribution of (y,$\mathbf{x}$)

Exercise: Show that the CEF error has a conditional mean of zero.

Answer:

$$\begin{aligned}
\mathbb{E}[e|\mathbf{x}] &= \mathbb{E}[(y - m(\mathbf{x})|\mathbf{x}] && \text{Def. of e} \\
&= \mathbb{E}[y|\mathbf{x}] - \mathbb{E}[m(\mathbf{x})|\mathbf{x}] && \text{Linearity of Expectation Operator} \\
&= m(\mathbf{x}) - m(\mathbf{x}) && \text{Conditioning Theorem} \\
&= 0
\end{aligned}$$

Properties of the CEF Error:
If $\mathbb{E}[y] < \infty$:

1. $\mathbb{E}[e|\mathbf{x}] = 0$
2. $\mathbb{E}[e] = 0$
3. If $\mathbb{E}[y]^r < \infty$ for $r \geq 1$ then $\mathbb{E}[e]^r < \infty$
4. For any function $h(\mathbf{x})$ such that $\mathbb{E}[h(\mathbf{x})e] < \infty$ then $\mathbb{E}[h(\mathbf{x})e) = 0$

Most of the properties of the CEF error are relatively self-explanatory. Note, however, that the implication of 3. is that if a moment exists for y then it must also exist for e.

Conditional Variance:

$$\sigma^2(\mathbf{x}) = Var(y|\mathbf{x}) = \mathbb{E}[(y - \mathbb{E}[y|\mathbf{x}])^2|\mathbf{x}]$$

Homoskedasticiy and Heteroskedasticity:

i. Homoskedasticity: $\mathbb{E}[e^2|\mathbf{x}] = \sigma^2$

ii. Heteroskedasticity: $\mathbb{E}[e^2|\mathbf{x}] = \sigma^2(\mathbf{x})$

Exercise:

i. Write down the mean-squared error of a predictor $h(\mathbf{x})$ for e^2.
ii. Show that $\sigma^2(\mathbf{x})$ minimizes the mean-squared error and is thus the best predictor.

Answers:

i. $\mathbb{E}[(e^2 - h(\mathbf{x}))^2]$
ii.

$$\begin{aligned}
\mathbb{E}[(e^2 - h(\mathbf{x}))^2] &= \mathbb{E}[(e^2 - \sigma^2(\mathbf{x}) + \sigma^2(\mathbf{x}) - h(\mathbf{x}))^2] \\
&= \mathbb{E}[(e^2 - \sigma^2(\mathbf{x}))^2 + (\sigma^2(\mathbf{x}) - h(\mathbf{x}))^2 + 2(e^2 - \sigma^2(\mathbf{x}))(\sigma^2(\mathbf{x}) - h(\mathbf{x}))] \\
&= \mathbb{E}[(e^2 - \sigma^2(\mathbf{x}))^2] + \mathbb{E}[(\sigma^2(\mathbf{x}) - h(\mathbf{x}))^2] + \mathbb{E}[2(e^2 - \sigma^2(\mathbf{x}))(\sigma^2(\mathbf{x}) - h(\mathbf{x}))] && \text{Linearity of Expectation} \\
&= \mathbb{E}[(e^2 - \sigma^2(\mathbf{x}))^2] + \mathbb{E}[(\sigma^2(\mathbf{x}) - h(\mathbf{x}))^2] + \mathbb{E}[\mathbb{E}(2(e^2 - \sigma^2(\mathbf{x}))(\sigma^2(\mathbf{x}) - h(\mathbf{x}))|\mathbf{x})] && \text{LIE} \\
&= \mathbb{E}[(e^2 - \sigma^2(\mathbf{x}))^2] + \mathbb{E}[(\sigma^2(\mathbf{x}) - h(\mathbf{x}))^2] + 2\,\mathbb{E}[(\sigma^2(\mathbf{x}) - h(\mathbf{x}))\,\mathbb{E}[(e^2 - \sigma^2(\mathbf{x}))|\mathbf{x}]] && \text{Conditioning Theorem} \\
&= \mathbb{E}[(e^2 - \sigma^2(\mathbf{x}))^2] + \mathbb{E}[(\sigma^2(\mathbf{x}) - h(\mathbf{x}))^2] + 2\,\mathbb{E}[(\sigma^2(\mathbf{x}) - h(\mathbf{x}))(\mathbb{E}[e^2|\mathbf{x}] - \sigma^2(\mathbf{x}))] && \text{Conditioning Theorem} \\
&= \mathbb{E}[(e^2 - \sigma^2(\mathbf{x}))^2] + \mathbb{E}[(\sigma^2(\mathbf{x}) - h(\mathbf{x}))^2] + 2\,\mathbb{E}[(\sigma^2(\mathbf{x}) - h(\mathbf{x}))(|\mathbf{x}] - \sigma^2(\mathbf{x}))] && (\mathbb{E}[e^2|\mathbf{x}] = \sigma^2(\mathbf{x})) \\
&= \mathbb{E}[(e^2 - \sigma^2(\mathbf{x}))^2] + \mathbb{E}[(\sigma^2(\mathbf{x}) - h(\mathbf{x}))^2]
\end{aligned}$$

which is clearly minimized at $h(x) = \sigma^2(\mathbf{x})$.

Section 2:

Best Linear Predictor Coefficient

The Best Linear Predictor Coefficient, β_{lpc} is defined as the β which minimizes the mean-squared prediction error:

$$S(\beta) = \mathbb{E}[(y - \mathbf{x}^T\beta)^2]$$

Various maths give us:

$$\mathbb{E}[\mathbf{x}\mathbf{x}^T]\beta = \mathbb{E}[\mathbf{x}y]$$

Which, given certain conditions, will give us:

$$\beta_{lpc} = \mathbb{E}[\mathbf{x}\mathbf{x}^T]^{-1}\,\mathbb{E}[\mathbf{x}y]$$

Question:

i. Under what condition is β_{lpc} uniquely identified?

Answer:

i. When $\mathbb{E}[\mathbf{x}\mathbf{x}^T]$ is invertible.

Notes on the Error

$$y = \mathbb{E}[y|\mathbf{x}] + e$$
$$y = \mathbf{x}^T\beta + u$$
$$0 = \mathbb{E}[y|\mathbf{x}] - \mathbf{x}^T\beta - u + e$$
$$u = \underbrace{\mathbb{E}[y|\mathbf{x}] - \mathbf{x}^T\beta}_{\text{Approximation Error}} + \underbrace{e}_{\text{CEF error}}$$

Exercise:

i. Show why $\mathbb{E}[u] = 0$ when we include an intercept.

Answer:

First note that we define the coefficients of the affine OLS regressor thusly:

$$(\beta_0, \beta) = \underset{(b_0,b)}{\arg\min}\,\mathbb{E}[\big(y - (b_0 + \mathbf{x}^T b)\big)^2]$$

For the purposes of this exercise we need only focus on β_0. Taking FOC:

$$\mathbb{E}[-2(y - \beta_0 - \mathbf{x}^T\beta)] = 0$$

Plugging in our definition of u:

$$\mathbb{E}[-2(u)] = 0$$

Which clearly simplifies to:

$$\mathbb{E}[u] = 0$$

Omitted Variable Bias Short and Long Regressions

Long Regression:

$$y = \mathbf{x}_1^T\beta_1 + \mathbf{x}_2^T\beta_2 + u$$

Which would have the corresponding estimand:

$$\beta = \mathbb{E}[\mathbf{x}\mathbf{x}^T]^{-1}\,\mathbb{E}[\mathbf{x}y]$$

Short Regression:

$$y = \mathbf{x}_1^T\gamma_1 + u$$

Which would have the corresponding estimand:

$$\gamma_1 = \mathbb{E}[\mathbf{x}_1\mathbf{x}_1^T]^{-1}\,\mathbb{E}[\mathbf{x}_1 y]$$

Exercise:

i. Show that $\gamma_1 \neq \beta_1$.

Answer:

i.

$$\begin{aligned}
\gamma_1 &= \mathbb{E}[\mathbf{x}_1\mathbf{x}_1^T]^{-1}\,\mathbb{E}[\mathbf{x}_1 y] \\
&= \mathbb{E}[\mathbf{x}_1\mathbf{x}_1^T]^{-1}\,\mathbb{E}[\mathbf{x}_1(\mathbf{x}_1^T\beta_1 + \mathbf{x}_2^T\beta_2 + u)] && \text{(Def. of y)} \\
&= \mathbb{E}[\mathbf{x}_1\mathbf{x}_1^T]^{-1}\left(\mathbb{E}[\mathbf{x}_1\mathbf{x}_1^T\beta_1] + \mathbb{E}[\mathbf{x}_1\mathbf{x}_2^T\beta_2] + \mathbb{E}[\mathbf{x}_1 u]\right) && \text{(Linearity of } \mathbb{E}\text{)} \\
&= \mathbb{E}[\mathbf{x}_1\mathbf{x}_1^T]^{-1}\left(\mathbb{E}[\mathbf{x}_1\mathbf{x}_1^T]\beta_1 + \mathbb{E}[\mathbf{x}_1\mathbf{x}_2^T]\beta_2\right) && (\mathbb{E}[\mathbf{x}_1 u] = 0) \\
&= \mathbb{E}[\mathbf{x}_1\mathbf{x}_1^T]^{-1}\,\mathbb{E}[\mathbf{x}_1\mathbf{x}_1^T]\beta_1 + \mathbb{E}[\mathbf{x}_1\mathbf{x}_1^T]^{-1}\,\mathbb{E}[\mathbf{x}_1\mathbf{x}_2^T]\beta_2 \\
&= \beta_1 + \mathbb{E}[\mathbf{x}_1\mathbf{x}_1^T]^{-1}\,\mathbb{E}[\mathbf{x}_1\mathbf{x}_2^T]\beta_2
\end{aligned}$$

Exercise:

i. (Hansen 2.15) Consider the intercept-only model $y = \alpha + e$ defined as the best linear predictor. Show that $\alpha = \mathbb{E}[y]$.

Answer:

i.

$$\begin{gathered}
y = \alpha + u \\
\min \mathbb{E}[(y-\alpha)^2] \\
-2\,\mathbb{E}[y - \alpha] = 0 \\
\alpha = \mathbb{E}[y]
\end{gathered}$$

Exercise:

i. Suppose that:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u$$

and

$$x_1 = \gamma_0 + \gamma_1 x_2 + u_1$$

Show that:

$$\beta_1 = \frac{Cov(u_1, y)}{Var(u_1)}$$

Answer:

i.

$$\begin{aligned}
Cov(u_1, y) &= Cov(u_1, \alpha + \beta_1 x_1 + \beta_2 x_2 + u) \\
&= \beta_1 Cov(u_1, x_1) + \beta_2 Cov(u_1, x_2) + Cov(u_1, u) \\
&= \beta_1 Cov(u_1, \gamma + \gamma_1 x_2 + u_1) + Cov(x_1 - \gamma - \gamma_1 x_2, u) \\
&= \beta_1 Cov(u_1, u_1) \\
&= \beta_1 Var(u_1)
\end{aligned}$$

Section 3:

Non-linear prediction example:

Suppose you have two groups, high-school educated people and college educated people, such that high-school educated people follow:

$$wage = \alpha + 0.5edu + u_1$$

while college educated people follow:

$$wage = \alpha = 0.8edu + u_2$$

Question:

i. What can we say about the conditional mean?

Answer:

i. The conditional mean is non-linear in education.

So, let's consider an alternate, better fitting specification:

$$wage = \alpha + \beta_1 edu + \beta_2 edu^2 + u$$

Question:

i. What happens, with this specification, as we increase education by 1 unit?

Answer:

i.

$$\frac{\partial P(y|x)}{\partial edu} = \beta_1 + 2\beta_2 edu$$

Application: Thinking about Causality

Consider the following linear approximation:

$$Health = \alpha + \beta Insured + u$$

where $Health$ is a self reported health metric and $Insured$ is an indicator (dummy) variable equal to 1 when a person is insured. Assume $\beta > 0$, i.e. insured people have better health outcomes.

Question:

i. Can we reasonably assume β captures a causal effect? What is the necessary condition?

Answer:

i. No, the CIA almost surely fails here. More specifically:

$$f(u|Insured) \neq f(u)$$

which implies:

$$\Delta_1 f(u|Insured) \neq 0$$

This is likely the case because insurance is correlated with other things which also affect health, such as income or wealth.

Now, suppose that we control for education, wealth, and some other easily observable characteristics such that we estimate:

$$Health = \alpha + \beta_1 Insured + \mathbf{x}^T \beta + u$$

Question:

i. Can we reasonably assume β_1 captures a causal effect? Or in other words, do you think

$$f(u|Insured, \mathbf{x}) = f(u|\mathbf{x})$$

Answer:

i. Probably not. Health in particular is likely subject to many unobserved things correlated with insurance such as risk preferences.

Exercise (Hansen 2.22):

i. Take the homoskedastic model

$$y = \mathbf{x}_1'\beta_1 + \mathbf{x}_2'\beta_2 + e$$
$$\mathbb{E}[e|\mathbf{x}_1, \mathbf{x}_2] = 0$$
$$\mathbb{E}[e^2|\mathbf{x}_1, \mathbf{x}_2] = \sigma^2$$
$$\mathbb{E}[\mathbf{x}_2|\mathbf{x}_1] = \Gamma \mathbf{x}_1$$
$$\Gamma \neq 0$$

Suppose that the parameter β_1 is of interest. We know that hte exclusion of $\mathbf{x}_2$ creates omitted variable bias in the projection coefficient on $\mathbf{x}_2$. It also changes the equation error. What is the effect on the homoskedastic property of the induced equation error? Does the exclusion of $\mathbf{x}_2$ induce heteroskedasticity?

Answer:

i. First let's write down the short regression we are estimating

$$y = \gamma \mathbf{x}_1 + u$$

ii. Second, show that $Var(u|\mathbf{x}_1) = Var(y)$

$$\begin{aligned} Var(u|\mathbf{x}_1) &= Var(y - \gamma\mathbf{x}_1|\mathbf{x}_1) \\ &= Var(y|\mathbf{x}_1) \end{aligned}$$

iii. Finally, test for heteroskedasticity

$$\begin{aligned} Var(y|\mathbf{x}_1) &= Var(\mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2 + e|\mathbf{x}_1) \\ &= \beta_2^2 Var(\mathbf{x}_2|\mathbf{x}_1) + Var(e|\mathbf{x}_1) \end{aligned}$$

Now, note that $Var(\mathbf{x}_2|\mathbf{x}_1)$ is a function of $\mathbf{x}_1$ as $\mathbb{E}[\mathbf{x}_2|\mathbf{x}_1]$. Thus we have heteroskedasticity.

Section 4-5:

Projection and Annihilator Matrices:

To begin, we define the projection matrix:

$$P = X(X^TX)^{-1}X^T$$

Which has the following properties:

1. P is symmetric.
2. P is idempotent.
3. $tr(P) = rank(P) = k$
4. $PX = X$
5. $Py = \hat{y}$

Next, we define the annihilator matrix:

$$M = I_n - P$$

Which has the following properties:

1. M is symmetric.
2. M is idempotent.
3. $tr(M) = rank(M) = n - k$
4. $MX = 0$
5. $My = \hat{e} = Me$

These matrices are useful when thinking about estimating the error variance and when considering regression components/residual regressions.

Leverage Values

Moving forward we will continue to encounter leverage values, the diagonal elements of P.

$$h_{ii} = \mathbf{x}_i^T(X^TX)^{-1}\mathbf{x}_i$$

which have the following properties:

$0 \leq h_{ii} \leq 1$

$h_{ii} > \frac{1}{n}$ if X includes an intercept.

$\sum_{i=1}^{n} h_{ii} = k$

Intuitively we can think of the leverage value as giving us a measure for how similar a particular observation is to others in the sample. These will become important when we get to heteroscedastic-robust standard error estimators.

OLSE General Setting:

The observations $(y_i, \mathbf{x}_i)$ satisfy the linear regression equation:

$$y_i = \mathbf{x}_i^T \beta + u_i$$

$$\mathbb{E}[u_i | x_i] = 0$$

The variables have finite second moments:

$$\mathbb{E}[y_i^2] < \infty$$

$$\mathbb{E}\,||\mathbf{x}_i||^2 < \infty$$

and an invertible design matrix:

$$\mathbb{E}[\mathbf{x}_i \mathbf{x}_i^T] > 0$$

(Aka it is positive definite) Also, note that these assumptions imply:

$$\mathbb{E}[y_i | X] = \mathbb{E}[y_i | \mathbf{x}_i] = \mathbf{x}_i^T \beta$$

OLSE Examples: Questions: Consider the following model:

$$y_i = \mathbf{x}_i^T \beta + u_i$$

State the assumptions necessary for:

a) $\mathbb{E}[y_i | \mathbf{x}_i| = \mathbf{x}_i^T \beta$

b) $Median(y_i | \mathbf{x}_i) = \mathbf{x}^T \beta$

Answers:

a) $\mathbb{E}[u_i | x_i] = 0$

b)

$$\begin{aligned} P(y_i \leq Median(y_i | \mathbf{x}_i) | \mathbf{x}_i) &= 0.5 \\ P(\mathbf{x}_i^T \beta + u_i \leq \mathbf{x}_i^T \beta | \mathbf{x}_i) &= 0.5 \\ P(u_i \leq 0 | \mathbf{x}_i) &= 0.5 \end{aligned}$$

Thus:

$$Median(u_i | \mathbf{x}_i) = 0$$

is required.

Useful $\hat{\beta}$ Decomposition

In addition to:

$$\hat{\beta} = (X^T X)^{-1} (X^T y)$$

We can show:

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T y \\ &= (X^T X)^{-1} X^T (X\beta + u) \\ &= \beta + (X^T X)^{-1} X^T u \end{aligned}$$

Variance of the Least Squares Estimator

First, let us consider the conditional covariance matrix of the error u. We often write it thusly:

$$Var(u|X) = \mathbb{E}[uu^T|X] = D$$

However, it is worth recalling why this is the case:

$$Var(u|X) = \mathbb{E}[(u - \mathbb{E}[u|X])(u - \mathbb{E}[u|X])^T|X]$$

But we know that:

$$\mathbb{E}[u|X] = 0$$

Simplifying to the above expression. Now, note that the i^{th} diagonal element of D is:

$$\mathbb{E}[u_i^2|X] = \mathbb{E}[u_i^2|x_i] = \sigma_i^2$$

and, under the assumptions we've used thus far, we know that the off-diagonal elements are equal to 0. Thus, our matrix D is:
And, consequently, we can find the variance of our OLS estimator under the general condition of heteroskedasticity:

$$\begin{aligned} Var(\hat{\beta}|X) &= Var(\beta + (X^TX)^{-1}X^Tu|X) \\ &= Var((X^TX)^{-1}X^Tu|X) \\ &= (X^TX)^{-1}X^TVar(u|X)X(X^TX)^{-1} \\ &= (X^TX)^{-1}(X^TDX)(X^TX)^{-1} \end{aligned}$$

Now, from here we can derive the Eicker-White robust standard errors. First note that we can write:

$$D = \mathbb{E}[uu^T|X] = \mathbb{E}[\tilde{D}|X]$$

where $\tilde{D} = diag(u_1^2, u_2^2 ... u_n^2)$ Now, note that:

$$X^T\tilde{D}X = \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T u_i^2$$

Thus, if u_i was observable, we could calculate the estimator:

$$\begin{aligned} \hat{Var}(\hat{\beta}|X) &= (X^TX)^{-1}(X^T\tilde{D}X)(X^TX)^{-1} \\ &= (X^TX)^{-1}\left(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T u_i^2\right)(X^TX)^{-1} \end{aligned}$$

Which would be an unbiased estimator of $Var(\hat{\beta})$. However, given that we can't observe u_i, we instead can estimate:

$$(X^TX)^{-1}\left(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T \hat{u}_i^2\right)(X^TX)^{-1}$$

Which is our first feasible heteroskedastic-consistent standard error estimator. This, however, is not the estimator that various statistical programs will estimate. There are a few more corrections involved in those, as we will see later.

Variance of $\hat{\beta}$ under homoskedasticity:
Using our alternative decomposition of $\hat{\beta}$:

$$\begin{aligned} Var(\hat{\beta}|X) &= Var(\beta + (X^TX)^{-1}X^Tu|X) \\ &= Var((X^X)^{-1}X^Tu|X) \\ &= (X^TX)^{-1}X^TVar(u|X)X(X^TX)^{-1} \\ &= (X^TX)^{-1}X^T\sigma^2I_nX(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1} \end{aligned}$$

Note: we cannot calculate this variance. We can only estimate it.

Sections 6-7:

Convergence in Probability:
A sequence of Random Variables $\{\bar{Y}_1, \bar{Y}_2, ...\}$ converges in probability to a real number μ if, for any $\epsilon > 0$

$$\lim_{n\to\infty} P\left(|\bar{Y}_n - \mu| < \epsilon\right) = 1$$

To help provide some intuition, consider the following example I stole from Wikipedia: First, pick a random person in the street. Let μ be his/her height, which is ex ante a random variable. Then ask other people to estimate this height by eye. Let $\bar{Y}_n$ be the average of the first n responses. Then (provided there is no systematic error) by the law of large numbers, the sequence $\bar{Y}_n$ will converge in probability to the random variable μ.

Consistency:
An estimator $\hat{\theta}$ of a parameter θ is consistent if

$$\hat{\theta} \xrightarrow{p} \theta \text{ as } n \to \infty$$

(A) Weak Law of Large Numbers:
If Y_i are iid and $\mathbb{E}[Y] < \infty$ then as $n \to \infty$

$$\bar{Y} \xrightarrow{p} \mathbb{E}[Y]$$

Proof:
Chebyshev's inequality tells us

$$P\left(|\bar{Y} - \mathbb{E}[\bar{Y}]| > \delta\right) \leq \frac{\sigma^2}{\delta^2} \cdot \frac{1}{n}$$

Note that we can rewrite this as:

$$P\left(|\bar{Y} - \mu| \leq \delta\right) \geq 1 - \frac{1}{n}\frac{\sigma^2}{\delta^2}$$

Now, we can see that as $n \to \infty$ the final term will go to zero, giving us our desired result by the definition of convergence in probability.

Almost Sure Convergence:

A sequence of Random Variables $\{\bar{Y}_1, \bar{Y}_2, ...\}$ converges almost surely to a real number μ if, for any $\epsilon > 0$

$$\lim_{N\to\infty} P\left(\sup_{n>N} |\bar{Y}_n - \mu| < \epsilon\right) = 1$$

$$P\left(\lim_{n\to\infty} \bar{Y}_n = \mu\right) = 1$$

To help provide some intuition, consider another example I stole from Wikipedia: Consider an animal of some short-lived species. We record the amount of food that this animal consumes per day. This sequence of numbers will be unpredictable, but we may be quite certain that one day the number will become zero, and will stay zero forever after.

Note: Convergence almost surely implies convergence in probability.

(A) Strong Law of Large Numbers:

If Y_i are iid and $\mathbb{E}[Y] < \infty$ then as $n \to \infty$

$$\bar{Y} \xrightarrow{A.S} \mathbb{E}[Y]$$

Convergence in Distribution:

Let $\bar{Y}_n$ be a sequence (vector) of Random Variables with distribution $F_{\bar{Y}_n}(u)$. We say that $\bar{Y}_n$ converges in distribution to μ if as $n \to \infty$

$$\lim_{n\to\infty} F_{\bar{Y}_n}(u) = F_\mu(u)$$

Once again, to gain some intuition, consider the following example from Wikipedia: Let X_n be the fraction of heads after tossing up an unbiased coin n times. Then X_1 has the Bernoulli distribution with expected value $\mu = 0.5$ and variance $\sigma^2 = 0.25$. The subsequent random variables $X_2, X_3, ...$ will all be distributed binomially. As n grows larger, this distribution will gradually start to take shape more and more similar to the bell curve of the normal distribution.

Central Limit Theorem (Lindberg-Levy):

If y_i are iid and $\mathbb{E}[y_i^2] < \infty$ then as $n \to \infty$

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

Continuous Mapping Theorem:

If $Y_n \xrightarrow{p} a$ and $g(\cdot)$ is continuous at a then

$$g(Y_n) \xrightarrow{p} g(a)$$

Some useful properties. Let $Y_n \xrightarrow{p} \mu$ and $X_n \xrightarrow{p} \alpha$:

$$Y_n + X_n \xrightarrow{p} \mu + \alpha$$

$$Y_n X_n \xrightarrow{p} \mu\alpha$$

$$\frac{Y_n}{X_n} \xrightarrow{p} \frac{\mu}{\alpha} \text{ if } \alpha \neq 0$$

Slutsky's Theorem:

Let $Y_n \xrightarrow{p} \mu$ and $X_n \xrightarrow{d} X$

$$Y_n + X_n \xrightarrow{d} \mu + X$$

$$Y_n X_n \xrightarrow{d} \mu X$$

$$\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{\mu} \text{ if } \mu \neq 0$$

Delta Method:

Let $\hat{\mu}$ be an estimator for μ such that

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} N(0, V)$$

and let $g(\cdot)$ be continuously differentiable. Then

$$\sqrt{n}\left(g(\hat{\mu}) - g(\mu)\right) \xrightarrow{d} N(0, G^T V G) \tag{1}$$

where

$$G(\mu) = \frac{\partial}{\partial \mu} g(\mu)$$

Proof (Univariate Case):

First, note that the mean value theorem tells us:

$$g(\hat{\mu}) = g(\mu) + G(\tilde{\mu})(\hat{\mu} - \mu)$$

where $\hat{\mu} \xrightarrow{p} \mu$ and $\tilde{\mu}$ is some point such that $\hat{\mu} < \tilde{\mu} < \mu$. Because $g(\cdot)$ is assumed to be continuous we know that:

$$\hat{\mu} \xrightarrow{p} \mu \implies \tilde{\mu} \xrightarrow{p} \mu$$

Furthermore, the CMT tells us:

$$G(\tilde{\mu}) \xrightarrow{p} G(\mu)$$

Rearranging (1) and multiplying by $\sqrt{n}$ gives us:

$$\sqrt{n}[g(\hat{\mu}) - g(\mu)] = \sqrt{n} G(\tilde{\mu})[\hat{\mu} - \mu]$$

Now because we know:

$$\sqrt{n}[\hat{\mu} - \mu] \xrightarrow{d} N(0, V)$$

Slutsky's theorem tells us:

$$\sqrt{n}\left(g(\hat{\mu}) - g(\mu)\right) \xrightarrow{d} N(0, G^T V G)$$

Section 8:

Delta Method Example:

Consider the following affine transformation:

$$g(\mu) = a\mu + b$$

where a and b are some known constants. We know that:

$$G(\mu) = a$$

Thus if we know that

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} N(0, V)$$

then by the Delta Method

$$\sqrt{n}\left(g(\hat{\mu}) - g(\mu)\right) \xrightarrow{d} N(0, a^2 V)$$

Now, consider the example in our (upcoming) homework assignment where

$$g(\mu) = \mu^2$$

clearly we know that

$$G(\mu) = 2\mu$$

Thus (using the same assumptions as before) by the delta method:

$$\sqrt{n}\left(g(\hat{\mu}) - g(\mu)\right) \xrightarrow{d} N(0, 4\mu^2 V)$$

Bias vs. Consistency

In order to help delineate the difference between bias and consistency consider the following examples (also stolen from Wikipedia):

i. **Unbiased but not consistent:**

An estimator can be unbiased but not consistent. For example, for an iid sample $\{x_1, ..., x_n\}$ one can use $\hat{\theta} = x_n$ as the estimator of the mean $\mathbb{E}[x]$. Note that here the sampling distribution of is the same as the underlying distribution, so $\mathbb{E}[\hat{\theta}] = \mathbb{E}[x]$ and it is unbiased, but it does not converge to any value and clearly is not a consistent estimator of $\mathbb{E}[x]$.

ii. **Biased but consistent**: Alternatively, an estimator can be biased but consistent. For example, if the mean is estimated by $\frac{1}{n}\sum x_i + \frac{1}{n}$, we know that $\frac{1}{n}\sum x_i + \frac{1}{n}$ is biased, but as $n \to \infty$, it approaches the correct value, and so it is consistent. For an obvious example of such an estimator one need only look at the sample variance!

Size and Power:

The size of a test is equal to the probability of committing a Type I error, mathematically we say the size α of a test is:

$$\alpha = P(\text{Reject } H_0 | H_0 \text{ is true})$$

In contrast, the power of a test is equal to the probability of **avoiding** a Type II error, or:

$$\text{Power} = P(\text{Reject } H_0 | H_0 \text{ is false})$$

Thus it should be clear why we are concerned if we think our test may be under-powered. Namely it becomes possible that the null hypothesis is actually false (usually we are testing a null of $H_0 = 0$ so this corresponds to there being an statistically significant effect) and yet we fail to reject.

Section 9:

Asymptotic Distribution of the OLS Estimator:

Under the assumptions of linearity, ergodic stationarity, predetermined regressors, rank condition, and that X_tU_t is a martingale difference sequence with finite second moment, we can show that:

(a) $\hat{\beta} \xrightarrow{p} \beta$ as $n \to \infty$

(b) $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V(\hat{\beta}))$ as $n \to \infty$
where $V(\hat{\beta}) = \sum_{XX}^{-1} S \sum_{XX}^{-1}$, $S = \mathbb{E}[(X_tU_t)(X_tU_t)^T]$

Now, let's consider how to prove this. First note that we have shown:

$$\hat{\beta} - \beta = \left(\frac{1}{n} \sum_{t=1}^{n} X_t X_t^T \right)^{-1} \frac{1}{n} \sum_{t=1}^{n} X_t U_t$$

Now, we know that, under our assumption that $\{X_t, X_t^T\}$ is ergodic stationary, we know:

$$\frac{1}{n} \sum_{t=1}^{n} X_t X_t^T \xrightarrow{p} \sum_{XX}$$

We also know that $\sum_{XX}^{-1}$ is invertible due to the rank condition, so this will also hold true for the inverse. Now, note that:

$$\frac{1}{n} \sum_{t=1}^{n} X_t U_t \xrightarrow{p} \mathbb{E}[X_t U_t]$$

which, under our assumption of predetermined regressors, is equal to 0. We thus know by the CMT that:

$$\hat{\beta} - \beta \xrightarrow{p} 0$$

or equivalently:

$$\hat{\beta} \xrightarrow{p} \beta$$

as $n \to \infty$. Now, note that the ergodic stationary Martingale Difference CLT tells us that:

$$\sqrt{n} \frac{1}{n} \sum_{t=1}^{n} X_t U_t \xrightarrow{d} N(0, S)$$

Thus, by the Slutsky Theorem, we know that:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V(\hat{\beta})) \text{ as } n \to \infty$$

We, however, do not observe S, so we must estimate it with $\hat{S}$ (which changes based on the situation), which we take to be a consistent estimator of S. Thus we estimate $V(\hat{\beta})$ with $\hat{V}(\hat{\beta})$ wherein $\hat{S}$ replaces S.

Robust t-ratio

Having established the asymptotic distribution of our OLS estimator, we can construct (various) test statistics. Taking our above results, applying Slutsky's Theorem to the t-ratio gives us:

$$\frac{\sqrt{n}(\hat{\beta} - \beta)}{\sqrt{\hat{V}(\hat{\beta})}} \xrightarrow{d} N(0, 1)$$

Returning to Endogeneity:

Having further established the basic framework within which we, as applied econometricians, usally work. Let's revisit the problem of endogeneity. Consider the following regression model:

$$y_i = \beta x_i + u_i$$

where:

$$Cov(x_i, u_i) \neq 0$$

and

$$\mathbb{E}[x_i u_i] \neq 0$$

We know that we can write our estimator in the following way:

$$\hat{\beta}_{OLS} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} = \beta + \frac{\sum_{i=1}^{n} u_i x_i}{\sum_{i=1}^{n} x_i^2}$$

Now, note that under our given assumptions:

$$\mathbb{E}[\hat{\beta}_{OLS}|x_i] = \beta + \frac{\sum_{i=1}^{n} x_i \, \mathbb{E}[u_i|x_i]}{\sum_{i=1}^{n} x_i^2}$$

but $\mathbb{E}[u_i|x_i] \neq 0$, so $\hat{\beta}_{OLS}$ is not consistent! This is, of course, problematic. We often refer to this inconsistency of the OLS estimator as endogeneity bias or estimation bias. One way we can fix this problem is by using Instrumental Variables. Intuitively an instrument is an exogenous variable which can be used to proxy for an endogenous regressor, but to understand this further consider the following model. Let

$$y_i = x_i \beta_1 + u_i$$

where

$$\mathbb{E}[x_i u_i] \neq 0$$

In words, x_i is endogenous. We will define an instrumental variable as a random variable z_i such that

$$\mathbb{E}[z_i u_i] = 0$$

$$\mathbb{E}[z_i x_i] \neq 0$$

Thus our instrument is a variable which is exogenous from our error, but related to the endogenous variable. So what do we do with our instrument? One common (and very general) method is two-stage least squares (2SLS). With 2SLS our estimation process takes a few steps. First, we regress our instrument on the endogenous variable:

$$x_i = \alpha z_i + \epsilon_i$$

in order to get predicted values of x_i:

$$\hat{x}_i = \hat{\alpha} z_i$$

where, as we know:

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} x_i z_i}{\sum_{i=1}^{n} z_i^2}$$

Next, we regress our outcome on $\hat{x}_i$:

$$y_i = \gamma \hat{x}_i + u_i$$

giving us:

$$\hat{\gamma} = \frac{\sum_{i=1}^{n} y_i \hat{x}_i}{\sum_{i=1}^{n} \hat{x}_i^2}$$

But how does this fix our endogeneity problem? Note:

$$\begin{aligned}
\hat{\gamma} &= \frac{\sum_{i=1}^{n} y_i \hat{x}_i}{\sum_{i=1}^{n} \hat{x}_i^2} \\
&= \frac{\sum_{i=1}^{n} y_i \hat{\alpha} z_i}{\sum_{i=1}^{n} \hat{\alpha}^2 z_i^2} \\
&= \frac{\sum_{i=1}^{n} y_i z_i}{\hat{\alpha} \sum_{i=1}^{n} z_i^2} \\
&= \frac{\sum_{i=1}^{n} z_i^2}{\sum_{i=1}^{n} x_i z_i} \cdot \frac{\sum_{i=1}^{n} y_i z_i}{\sum_{i=1}^{n} z_i^2} \\
&= \frac{\sum_{i=1}^{n} y_i z_i}{\sum_{i=1}^{n} z_i x_i}
\end{aligned}$$

Now, that we have simplified our estimator a bit, we must show it is consistent. Note that plugging in for y_i:

$$\hat{\gamma} = \frac{\sum_{i=1}^{n}(\beta x_i + u_i)z_i}{\sum_{i=1}^{n} x_i z_i} = \beta + \frac{\sum_{i=1}^{n} z_i u_i}{\sum_{i=1}^{n} x_i z_i}$$

Using the weak law of large numbers we can show that:

$$\frac{1}{n}\sum_{i=1}^{n} z_i u_i \xrightarrow{p} \mathbb{E}[z_i u_i] = 0$$

and

$$\frac{1}{n}\sum_{i=1}^{n} z_i x_i \xrightarrow{p} \mathbb{E}[z_i x_i] \neq 0$$

Thus the continuous mapping theorem tells us:

$$\hat{\gamma} \xrightarrow{p} \beta + 0 = \beta$$

Giving us that $\hat{\gamma}$ is a consistent estimator for β.

Section 10:

Let's think about measurement error. Consider the following model:

$$Y_i = \beta X_i^* + u_i$$

$$X_i = X_i^* + v_i$$

where v_i is an unspecified measurement error. We can rearrange these equations to get:

$$y_i = \beta x_i + u_i - \beta v_i$$

We can show that:

$$\hat{\beta}_{OLS} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} = \beta + \frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2} - \beta\frac{\sum_{i=1}^{n} x_i v_i}{\sum_{i=1}^{n} x_i^2}$$

Now, we can see if our estimator is unbiased:

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_{OLS}] &= \beta + \mathbb{E}\left[\frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2}\right] - \mathbb{E}\left[\beta\frac{\sum_{i=1}^{n} x_i v_i}{\sum_{i=1}^{n} x_i^2}\right] \\
&= \beta + \mathbb{E}\left[\mathbb{E}\left[\frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2}|x_i\right]\right] - \beta\,\mathbb{E}\left[\frac{\sum_{i=1}^{n} x_i v_i}{\sum_{i=1}^{n} x_i^2}\right] \\
&= \beta - \beta\,\mathbb{E}\left[\frac{\sum_{i=1}^{n} x_i v_i}{\sum_{i=1}^{n} x_i^2}\right]
\end{aligned}$$

Note: We cannot take a conditional expectation here because x_i is a function of both x_i^* and v_i and we do not observe x_i^*
Now, while we can't simplify further, to get some intuition note that we can approximate this with:

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_{OLS}] &\approx \beta\left[1 - \frac{Cov(x_i, v_i)}{Var(x_i)}\right] \\
&= \beta\left[1 - \frac{Cov(x_i^* + v_i, v_i)}{Var(x_i^* + v_i)}\right] \\
&= \beta\left[1 - \frac{Var(v_i)}{Var(x_i^*) + Var(v_i)}\right] \\
&= \beta\left[\frac{Var(x_i^*)}{Var(x_i^*) + Var(v_i)}\right] \\
&< \beta
\end{aligned}$$

Thus we can see that we have attenuation bias. Now, let's consider what happens if we have measurement error in our dependent variable. Consider the following model where we observe y_i with error:

$$y_i = y_i^* + v_i$$

$$y_i^* = \beta x_i + u_i$$

Where x_i is independent from both v_i and u_i. Thus:

$$y_i = \beta x_i + u_i - v_i$$

Giving us:

$$\hat{\beta}_{OLS} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} = \beta + \frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2} + \frac{\sum_{i=1}^{n} x_i v_i}{\sum_{i=1}^{n} x_i^2}$$

Taking an expectation to test for bias, we get:

$$\mathbb{E}[\hat{\beta}_{OLS}] = \beta + 0 + 0$$

Thus, can can see measurement error in the dependent variable does not result in bias. However, it is worth nothing that the variance will likely be greater:

$$Var(y_i) = Var(y_i^*) + Var(v_i) \geq Var(y_i^*)$$

.