WWS 508c, Econometrics and Public Policy **Prof. Tom Vogl**

Spring 2015 **Princeton University**

LECTURE NOTE 1: REVIEW OF PROBABILITY AND STATISTICS

1 Introduction

The goal of this lecture is to give you a refresher on key concepts in probability and statistics. We will review material you covered in 507c, with some extensions and new approaches. To start, we will examine properties of random variables, and we will then move on to properties of samples. We will conclude with an overview of hypothesis testing.

2 Random Variables

Before we begin, let's set some notation. Let X and Y be random variables, and let a and b be constants. We will call x and y *realizations* of X and Y.

2.1 Expectations

The *expectation*, or *expected value*, or *mean*, of X is a measure of its central tendency. If X is a continuous random variable with probability density function $f(\cdot)$, then the expectation of X is:

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

If, on the other hand, X is a discrete random variable that takes on values x_k with probability p_k for $k = 1, 2, \cdots, K$, then the expectation of X is:

$$E[X] = \sum_{k=1}^{K} p_k x_k$$

where K can equal infinity. In reality, these two definitions are the same, since summation is a form of integration. Alternative notation for the mean is μ.

The *conditional expectation* of Y given X, written $E[Y|X]$, is the expectation of Y as a function of X. If $X = x$, then the conditional expectation of Y is $E[Y|X = x]$.

Expectations have several convenient properties:

1. Expectation of a scalar: $E[a] = a$.

2. Scalar addition and multiplication: $E[a + bX] = a + bE[X]$.
3. Addition of random variables: $E[X + Y] = E[X] + E[Y]$.
4. Multiplication of random variables: $E[XY] = E[X]E[Y]$ if X and Y are independent.
5. Law of iterated expectations (also called the law of total expectation): $E[Y] = E[E[Y|X]]$.

2.2 Variances and Covariances

The *variance* of X is a measure of its dispersion, defined as the expected square deviation of X from its mean:

$$V[X] = E\left[(X - E[X])^2\right]$$

We can apply the expectation properties above to obtain an alternative expression for the variance of X:

$$\begin{aligned} V[X] &= E\left[(X - E[X])^2\right] \\ &= E\left[X^2 - 2XE[X] + (E[X])^2\right] \\ &= E\left[X^2\right] - 2E\left[XE[X]\right] + (E[X])^2 \\ &= E\left[X^2\right] - 2E[X]E[X] + (E[X])^2 \\ &= E\left[X^2\right] - (E[X])^2 \end{aligned}$$

The *standard deviation* of X is the square root of its variance. Often, we write σ_X for the standard deviation of X and σ_X^2 for the variance of X.

As with expectations, we can also define the *conditional variance* of Y given X, which gives the variance of Y as a function of X. It is written $V[Y|X]$. The definition is analogous to the variance:

$$V[Y|X] = E\left[(Y - E[Y|X])^2 |X\right] = E[Y^2|X] - E[Y|X]^2$$

The *covariance* of X and Y measures the degree of comovement between the two variables:

$$\sigma_{XY} = cov(X, Y) = E\left[(X - E[X])(Y - E[Y])\right]$$

Similar to the above result for the variance, we can derive an alternative expression for the covariance:

$$\begin{aligned} cov(X, Y) &= E\left[(X - E[X])(Y - E[Y])\right] \\ &= E[XY] - E\left[XE[Y]\right] - E\left[YE[X]\right] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

Note that $V[X] = cov(X, X)$, so the variance is a special case of the covariance.

The concept of *correlation* is closely related to the covariance. The correlation between X and Y is:

$$\rho_{XY} = corr(X, Y) = \frac{cov(X, Y)}{\sqrt{V(X)V(Y)}}$$

It can be shown that $-1 \leq \rho_{XY} \leq 1$

We can derive other properties of variances and covariances:

1. Variance of a scalar: $V[a] = 0$.
2. Scalar addition and multiplication:

$$\begin{aligned} V[a + bX] &= E\left[(a + bX)^2\right] - E\left[a + bX\right]^2 \\ &= E\left[a^2 + 2abX + b^2X^2\right] - (a + bE[X])^2 \\ &= a^2 + 2abE[X] + b^2E[X^2] - a^2 - 2abE[X] - b^2E[X]^2 \\ &= b^2\left(E[X^2] - E[X]^2\right) \\ &= b^2V[X] \end{aligned}$$

3. Addition of random variables:

$$\begin{aligned} V[X + Y] &= E\left[(X + Y)^2\right] - E\left[X + Y\right]^2 \\ &= E[X^2 + 2XY + Y^2] - (E[X] + E[Y])^2 \\ &= E[X^2] + 2E[XY] + E[Y^2] - E[X]^2 - 2E[X]E[Y] - E[Y]^2 \\ &= \left(E\left[X^2\right] - (E[X])^2\right) + \left(E\left[Y^2\right] - (E[Y])^2\right) + 2\left(E[XY] - E[X]E[Y]\right) \\ &= V[X] + V[Y] + 2cov(X, Y) \end{aligned}$$

4. Covariance of X and Y when the conditional expectation of Y given X is linear: Let $E[Y|X] = a + bX$. Then we have:

$$\begin{aligned} cov(x, y) &= E[XY] - E[X]E[Y] \\ &= aE[X] + bE[X^2] - E[X]\left(a + bE[X]\right) \\ &= b\left(E[X^2] - (E[X])^2\right) = bV[X] \end{aligned}$$

We can rearrange to obtain $b = \frac{cov(X,Y)}{V[X]}$. Note also that by iterated expectations, we have that $E[Y] = E\left[E[Y|X]\right] = a + bE[X]$, so that $a = E[Y] - \frac{cov(X,Y)}{V[X]}E[X]$. These expressions for a and b will

be a cornerstone for the course.

5. Law of total variance:

$$\begin{aligned} V[Y] &= E[Y^2] - E[Y]^2 \\ &= E\left[E[Y^2|X]\right] - E\left[E[Y|X]\right]^2 \\ &= E\left[V[Y|X] + E[Y|X]^2\right] - E\left[E[Y|X]\right]^2 \\ &= E\left[V[Y|X]\right] + \left[E\left[E[Y|X]^2\right] - E\left[E[Y|X]\right]^2\right] \\ &= E\left[V[Y|X]\right] + V\left[E[Y|X]\right] \end{aligned}$$

This is also called the variance decomposition formula. It separates the variance of Y into an *unexplained component,* $E\left[V[Y|X]\right]$, and an *explained component*, $V\left[E[Y|X]\right]$.

3 Sampling and Estimation

The properties of random variables we reviewed in the previous section (expectation, conditional expectation, variance, covariance) are called *population parameters*. Throughout the course, we will be interested in the properties of *samples*, which are called *statistics*. Here, we consider the properties of a sample of N observations of the random variable X: $X_1, X_2, \cdots, X_N$. The observations are mutually independent draws from the distribution of X, which has mean μ_X and variance σ_X^2. We will say they are are "independently and identically distributed," or "i.i.d." Note the implication: *each observation has its own probability distribution*, and we assume that all of these individual distributions are identical. We will derive statistics of the sample to estimate population parameters for the distribution of X. We will call such statistics *estimators*.

At this point, a quick digression on notation will be useful. Usually, we use Greek letters to refer to parameters and either Greek letters with hats or Roman letters to refer to their corresponding estimators. For example, suppose we are interested in a regression parameter β. We will typically use $\hat{\beta}$ or b to refer to an estimator for β. For general discussions about parameters and estimators, we will use the parameter θ and the estimator $\hat{\theta}$.

We often want $\hat{\theta}$ to satisfy some of the following properties:

1. Unbiasedness: $E[\hat{\theta}] = \theta$.
2. Consistency: As the sample size grows, $\hat{\theta}$ gets closer and closer to θ. Mathematically, we express this property in the following way: as $N \to \infty$, $Pr\left[|\hat{\theta} - \theta| < \varepsilon\right] \to 1$ for any constant ε. We say that $\hat{\theta}$ "converges in probability" to θ, and we write $\hat{\theta} \xrightarrow{p} \theta$.
3. Efficiency (or Precision): $\hat{\theta}$ has the smallest possible variance $V[\hat{\theta}] = E\left[\left(\hat{\theta} - E[\hat{\theta}]\right)^2\right]$.

Properties (1) and (2) both represent the accuracy of $\hat{\theta}$, so we will typically seek estimators that satisfy either (1) and (3) or (2) and (3). Note that all three properties, especially (1) and (3), are features of the distribution of $\hat{\theta}$. This fact bears repeating: $\hat{\theta}$ has its own distribution. Make sure you understand this abstract concept, for it underlies much of the rest of the course.

3.1 Estimating the Mean

We begin by deriving an estimator for the mean of X that satisfies unbiasedness and precision. Let us consider the class of estimators that are linear in the data, i.e., $\hat{\mu} = a_1X_1 + a_2X_2 + \cdots + a_NX_N$ for some constants $a_1, a_2, \cdots, a_N$. We will try to find the appropriate a_k's to satisfy unbiasedness and precision.

Let us first examine the expectation of $\hat{\mu}_X$:

$$\begin{aligned} E[\hat{\mu}_X] &= E[a_1X_1 + a_2X_2 + \cdots + a_NX_N] \\ &= a_1E[X_1] + a_2E[X_2] + \cdots + a_NE[X_N] \\ &= a_1\mu_X + a_2\mu_X + \cdots + a_N\mu_X \\ &= \mu_X(a_1 + a_2 + \cdots + a_N) \end{aligned}$$

Thus if $\mu_X \neq 0$, $\hat{\mu}_X$ is unbiased if and only if $\sum_{i=1}^{N} a_i = 1$.

Now let us examine the variance of $\hat{\mu}_X$. Because the X_i's are independent, we can leave out all covariance terms:

$$\begin{aligned} V[\hat{\mu}_X] &= a_1^2V[X_1] + a_2^2V[X_2] + \cdots + a_N^2V[X_N] \\ &= \sigma_X^2 \sum_{i=1}^{N} a_i^2 \end{aligned}$$

To maximize precision, we want to minimize the variance of the estimator subject to the unbiasedness constraint that the a_i's add up to one:

$$\underset{a_1,a_2,\cdots,a_N}{min} \sigma_X^2 \sum_{i=1}^{N} a_i^2 \text{ subject to } \sum_{i=1}^{N} a_i = 1$$

If we were to write down the Lagrangian for this problem (you can do it at home), the first order conditions would imply that $a_1 = a_2 = \cdots = a_N$. Combined with the unbiasedness constraint, this implies that the minimum variance estimator has $a_i = \frac{1}{N}$ for all i.

Thus, the minimum variance unbiased linear estimator of the mean is the well-known sample average:

$$\hat{\mu}_X = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

The variance of this estimator is $V[\hat{\mu}_X] = V[\bar{X}] = \frac{1}{N}\sigma_X^2$.

Two other results regarding the sample average are worthy of note. We will state them here without proof:

1. Law of large numbers: $\bar{X} \xrightarrow{p} \mu_X$. In words: the sample average converges in probability to the population mean. Thus, in addition to being unbiased and precise, the sample average is also consistent.
2. Central limit theorem: As N approaches infinity, the distribution of $\bar{X}$ approaches a normal distribution with mean μ_X and variance $\frac{1}{N}\sigma_X^2$. As a result, in large samples, $\frac{\sqrt{N}}{\sigma_X}\left(\bar{X} - \mu_X\right)$ is approximately distributed $\mathcal{N}(0,1)$. The definition of "large sample" varies, but a common rule of thumb cutoff is a sample size of 30. We can use the central limit theorem to compute the probability that $\bar{X}$ will fall within $\delta > 0$ of the population mean μ_X:

$$\Pr\left[\bar{X} \in [\mu_X - \delta, \mu_X + \delta]\right] \approx 1 - 2\Phi\left[-\frac{\sqrt{N}}{\sigma_X}\delta\right] = 2\Phi\left[\frac{\sqrt{N}}{\sigma_X}\delta\right] - 1$$

   where Φ is the standard normal cumulative distribution function.

3.2 Estimating the Variance

Now that we have an estimator for the mean, we can derive an estimator for the variance quite easily. In particular, the fact that $V[X] = E\left[(X - E[X])^2\right] = E\left[X^2\right] - (E[X])^2$ suggests $\hat{\sigma}_X^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2 = \frac{1}{N}\sum_{i=1}^{N} X_i^2 - \bar{X}^2$ as an estimator ot the variance.[1] Let's check whether it is unbiased:

$$\begin{aligned}
E\left[\hat{\sigma}_X^2\right] &= \frac{1}{N}\sum_{i=1}^{N} E\left[X_i^2\right] - E\left[\bar{X}\right]^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(V\left[X_i\right] + E\left[X_i\right]^2\right) - \left(V\left[\bar{X}\right] + E\left[\bar{X}\right]^2\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\sigma_X^2 - \mu_X^2\right) - \left(\frac{\sigma_X^2}{N} - \mu_X^2\right) \\
&= \sigma_X^2\left(1 - \frac{1}{N}\right)
\end{aligned}$$

Thus, this estimator of the variance is biased. It is fairly straightforward to see, however, that $\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$ is an unbiased estimator of σ_X^2; you can try this at home. Nonetheless, the unbiased estimator is not obviously superior to $\hat{\sigma}_X^2$ as defined above. It can be shown that $\hat{\sigma}_X^2$ is a consistent estimator, and it has a

[1] In the textbook, Stock and Watson refer to $\hat{\sigma}_X^2$ as s_X^2. We will use both types of notation in this course.

lower variance than the unbiased estimator. Consequently, we have a *tradeoff between bias and variance*. In any case, we can see above that the bias term shrinks to zero as the sample becomes large; this is known as *asymptotic unbiasedness*.

4 Hypothesis Testing

Suppose we compute the sample average $\bar{X}$ in a sample of 100 observations and want to test the hypothesis that the mean of X, μ_X, equals 0. In this case, $\theta = \mu_X$ and $\hat{\theta} = \bar{X}$. For now, assume that we already know that the variance of X is 225. Because we have a large sample, the Central Limit Theorem tells us that $\frac{\sqrt{N}}{\sigma_X}\left(\bar{X} - \mu_X\right)$ is approximately distributed $\mathcal{N}(0,1)$. We can compute the p-value, or the probability of observing a sample average at least as extreme as ours, given that the null hypothesis ($\mu_X = 0$) is true. Let $\mu_{X,0}$ be the hypothesized value of μ_X, and let $\bar{X}^{actual}$ be the value of $\bar{X}$ we observe in our sample. We have $N = 100$ and $\sigma_X^2 = 225$, so $\sigma_X/\sqrt{N} = 15/10 = 1.5$. Then:

$$\begin{aligned} p &= \Pr\left[\left|\frac{\bar{X} - \mu_{X,0}}{\sigma_X/\sqrt{N}}\right| \geq \left|\frac{\bar{X}^{actual} - \mu_{X,0}}{\sigma_X/\sqrt{N}}\right|\right] = 2\Phi\left[-\left|\frac{\bar{X}^{actual} - \mu_{X,0}}{\sigma_X/\sqrt{N}}\right|\right] \\ &= \Pr\left[\left|\frac{\bar{X} - 0}{1.5}\right| \geq \left|\frac{\bar{X}^{actual} - 0}{1.5}\right|\right] = 2\Phi\left[-\left|\frac{\bar{X}^{actual}}{1.5}\right|\right] \end{aligned}$$

We take absolute values and multiply by two because we are performing a two-sided test. Suppose we observe a sample mean of 3 in our sample. Then the p-value for the null hypothesis $\mu_{X,0} = 0$ is 0.046. Since this p-value is less than 0.05, we often say that our sample mean is statistically different from 0 at the 5% level. Matters are a bit more complicated when σ_X^2 is unknown, but fortunately, in large samples, we can replace σ_X with $\hat{\sigma}_X^2$ in the formalas above.

In the above expressions, the denominators represent the estimated standard deviation of the distribution of the estimator $\bar{X}$. We refer to this value as the *standard error* of the estimator, and we write $SE[\bar{X}] = \sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{N}}$. For an arbitrary estimator $\hat{\theta}$, we write either $SE[\hat{\theta}]$ or $\sigma_{\hat{\theta}}$. In large samples, the statistic:

$$t = \frac{\hat{\theta} - \theta_0}{SE[\hat{\theta}]}$$

has a standard normal distribution, so that $p = 2\Phi\left[-\left|\frac{\hat{\theta}-\theta_0}{SE[\hat{\theta}]}\right|\right]$ for the test of the null hypothesis that $\theta = \theta_0$ against the alternative that $\theta \neq \theta_0$. We call t a t-statistic.

Above, we tested whether the mean μ_X equals zero. But often, we will be most interested in comparing means across populations. For instance, suppose X represents hourly earnings, and the two populations are white and African-American men in the United States. We can write μ_W and μ_B for their respective means

and σ_W^2 and σ_B^2 for their respective variances. The central limit theorem tells us that in large samples, $\bar{Y}_W \sim \mathcal{N}[\mu_W, \sigma_W^2/N_W]$ and $\bar{Y}_B \sim \mathcal{N}[\mu_B, \sigma_B^2/N_B]$. You'll recall from 507c that the sum (or difference) of two normally distributed random variables is itself normally distributed. (Technically, this statement is true if the variables are *jointly* normally distributed or independent, as is the case here.) Furthermore, from Section 2.2 above, we have the result that $V\left[\bar{Y}_W - \bar{Y}_B\right] = V\left[\bar{Y}_W\right] + V\left[\bar{Y}_B\right] - 2cov(\bar{Y}_W, \bar{Y}_B)$, but since W and B are independent samples, $cov(\bar{Y}_W, \bar{Y}_B) = 0$. Thus, in large samples:

$$\bar{Y}_W - \bar{Y}_B \sim \mathcal{N}\left[\mu_W - \mu_B, \frac{\sigma_W^2}{N_W} + \frac{\sigma_B^2}{N_B}\right]$$

If the variances σ_W^2 and σ_B^2 are unknown (as will almost always be the case), we can estimate them using the sample variances of the two samples, $\hat{\sigma}_W^2$ and $\hat{\sigma}_B^2$. In this case, the standard error of $\bar{Y}_W - \bar{Y}_B$ is:

$$SE\left[\bar{Y}_W - \bar{Y}_B\right] = \sqrt{\frac{\hat{\sigma}_W^2}{N_W} + \frac{\hat{\sigma}_B^2}{N_B}}$$

Now we can perform hypothesis tests as before. For instance, suppose we want to test whether black and white men have equal earnings. Our null hypothesis is $\mu_W - \mu_B = 0$, and our alternative is $\mu_W - \mu_B \neq 0$. We calculate our t-statistic as $t = \frac{\bar{Y}_W - \bar{Y}_B}{SE\left[\bar{Y}_W - \bar{Y}_B\right]}$ and compute p-values as before.

5 Confidence Intervals

Until now, we have been concerned with *point estimation*. A *point estimator* gives a unique value (the *point estimate*) that approximates the population parameter. An alternative to point estimation is *interval estimation*, which provides a range of values that contain the population parameter. One such interval estimator is the *confidence interval*. The confidence interval for a parameter θ at confidence level γ (say, 95%) will contain the θ in 100γ out of every 100 samples. Note that this concept bears a close relation to hypothesis testing. In particular, to determine whether we can reject the null hypothesis $\theta = \theta_0$ in favor of the alternative $\theta \neq \theta_0$ at significance level α, we can look to see whether the confidence interval with confidence level $1 - \alpha$ contains θ_0. Indeed, in large samples, the confidence interval for θ is $\hat{\theta} \pm \Phi^{-1}[1 - \frac{\alpha}{2}]SE[\hat{\theta}]$, where $\Phi^{-1}[\cdot]$ is the inverse of the standard normal cumulative distribution function. When $\gamma = 0.95$ (i.e., $\alpha = 0.05$), $\Phi^{-1}[1 - \frac{\alpha}{2}] = 1.96$. For example, the confidence interval for μ_X is $\bar{X} \pm 1.96SE[\bar{X}] = \bar{X} \pm 1.96\hat{\sigma}_X/\sqrt{N}$.

LECTURE NOTE 2: TWO-WAY CONTINGENCY TABLES AND PEARSON'S CHI-SQUARED TEST

1 Introduction

Until now, we have primarily considered continuous variables. Many variables of interest are categorical, however. As you learned last semester, in large samples, we can deal with binary variables using normal approximations, but we cannot apply such methods to variables with more than two categories. This lecture note describes simple methods for analyzing relationships between categorical variables.

2 Contingency Tables

You must have seen contingency tables (a.k.a. cross-tabulations) at least a few times in your life. A (two-way) contingency table has rows representing one categorical variable and columns representing another. The table reports the number (or proportion) of observations in each cell.

Consider a contingency table with R rows (indexed $i = 1, \cdots, R$) and C columns (indexed $j = 1, \cdots, C$). Let n_{ij} equal the number of observations in cell i, j. It will also be convenient to define the following sums (the number of observations in each row, each column, and the whole sample, respectively): $N_{i\cdot} = \sum_{j=1}^{C} n_{ij}$, $N_{\cdot j} = \sum_{i=1}^{R} n_{ij}$, and $N = \sum_{i=1}^{R} \sum_{j=1}^{C} n_{ij}$. For example, a two-by-two table would appear as follows:

	Column 1	Column 2	Total
Row 1	n_{11}	n_{12}	$N_{1\cdot} = n_{11} + n_{12}$
Row 2	n_{21}	n_{22}	$N_{2\cdot} = n_{21} + n_{22}$
Total	$N_{\cdot 1} = n_{11} + n_{21}$	$N_{\cdot 2} = n_{12} + n_{22}$	$N = n_{11} + n_{12} + n_{21} + n_{22}$

3 Pearson's Chi-Squared Statistic

Over a century ago, the English mathematician Karl Pearson set out to test the hypothesis that the row variable is independent of the column variable. He started by noting that under this null hypothesis, the expected number of obserations in cell i, j is:

$$E_{ij} = \frac{N_{i\cdot} N_{\cdot j}}{N}$$

You might find this expression conceptually easier when it is written $E_{ij} = N_{i\cdot}\frac{N_{\cdot j}}{N}$. This (slight) modification makes clear that we are multiplying the total number of observations in row i by the fraction of the sample that is in column j.

Pearson went on to prove that the statistic:

$$X^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

$$X^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

has an asymptotic χ^2 distribution with $(R-1)(C-1)$ degrees of freedom. (X is the upper-case version of the greek letter χ, or *chi*.) X^2 measures the sum of the squared deviations from expected cell size, normalized by the expected cell size. We can also write the statistic as:

$$X^2 = N\sum_{i=1}^{R}\sum_{j=1}^{C}\left(\frac{\left\{p_{ij} - \left(\sum_{k=1}^{C} p_{ik}\right)\left(\sum_{k=1}^{R} p_{kj}\right)\right\}^2}{\left(\sum_{k=1}^{C} p_{ik}\right)\left(\sum_{k=1}^{R} p_{kj}\right)}\right)$$

where p_{ij} is the fraction of the sample in cell i, j. This second expression highlights the fact that the test statistic (and therefore the significance level) increases with the sample size (provided the null hypothesis is not exactly correct, which would imply $X^2 = 0$ for all N). Holding the relative frequencies constant, the test statistic is proportional to the sample size. The X^2 statistic is attractive because it is easy to calculate with a pen and paper (or Excel, if you prefer more cutting-edge technology). The next time you read a policy brief or survey report that contains a cross-tabulation and little else, you'll easily be able to assess statistical significance.

In a two-by-two table, a back-of-the-envelope calculation is especially doable. The above expressions reduce to:

$$\begin{aligned} X^2_{2\times 2} &= N\frac{(n_{11}n_{22} - n_{12}n_{21})^2}{(n_{11}+n_{12})(n_{12}+n_{22})(n_{11}+n_{21})(n_{21}+n_{22})} \\ &= N\frac{(p_{11}p_{22} - p_{12}p_{21})^2}{(p_{11}+p_{12})(p_{12}+p_{22})(p_{11}+p_{21})(p_{21}+p_{22})} \end{aligned}$$

In the two-by-two case, the chi-squared test is equivalent to the test of equality of proportions using a normal approximation (the `prtest` command in Stata).

Before wrapping up, we should take note of a caveat. The Pearson chi-squared statistic does not perform well in the presence of very small expected cell sizes. Small expected cell sizes can occur if (1) the sample size

is small or (2) the data are very unbalanced across cells. The cutoff for "small" is a matter of some controversy, and it varies slightly with the number of cells, but a good rule of thumb is that you should check your results with an alternative test if any of the E_{ij}'s are smaller than 10, and you should not try using the Pearson chi-squared test if any of the E_{ij}'s are smaller than 5. When these circumstances arise, Fischer's exact test provides a good alternative. Fischer's exact test calculates p-values using combinatorics; the p-values hold exactly for any sample size, large or small. For large samples that are balanced across cells, however, the exact test becomes compuationally burdensome.

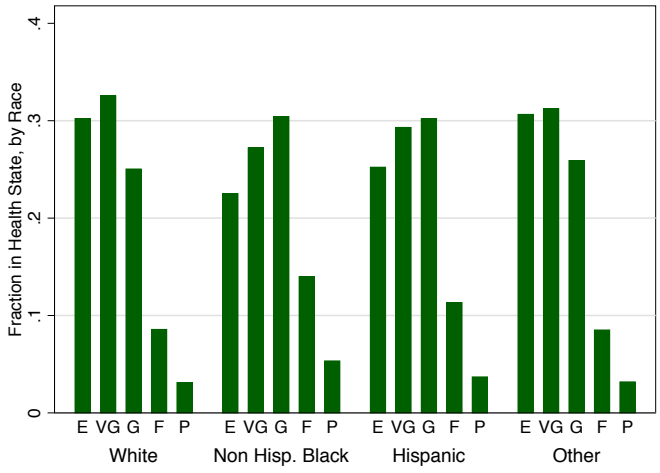
Both Pearson's and Fischer's tests are easy to implement in Stata. For Pearson's chi-squared test, type: `tab var1 var2, chi2`. For Fischer's exact test, type: `tab var1 var2, exact`.

4 Example: Race and Self-Reported Health

An example might help at this point. In the 2000 National Health Interview Survey (which you will analyze later in the semester), respondents were asked to rate their health on a 5-point scale, from "excellent" to "poor." The following contingency table shows how respondents of different racial and ethnic backgrounds responded to the question:

	Excellent	Very Good	Good	Fair	Poor	Total
White	12,254	13,205	10,158	3,476	1,261	40,354
Non-Hispanic Black	1,814	2,194	2,452	1,128	430	8,018
Hispanic	2,883	3,350	3,453	1,294	422	11,402
Other	886	903	749	246	92	2,876
Total	17,837	19,652	16,812	6,144	2,205	62,650

As a first step to interpreting a table of frequencies like this, I like to divide the cell entries by the total number of observations per race. Here is that calculation, in bar graph form:



This graph makes clear that whites and "others" (mostly Asians) report being in much better health than blacks and Hispanics. But are these differences statistically significant? By plugging the numbers from the table into the formula for Pearson's chi-squared statistic, we obtain $X^2 = 743.5$. We have 4 rows and 5 columns, which implies that X^2 has an approximate χ^2 distribution with $4 \times 3 = 12$ degrees of freedom. For a significance level of 0.001, the $\chi^2(12)$ distibution has a critical value of 33. Since our statistic (overwhelmingly!) exceeds 33, we can reject the hypothesis that the distribution of self-reported health is independent of race, with $p < 0.001$.

LECTURE NOTE 3: BIVARIATE REGRESSION

1 Introduction

Upon seeing a scatterplot of Y on X, most people wonder what line would best describe that scatterplot. This lecture is about how to estimate the line that best fits the data we observe on Y and X. We will seek to estimate a linear function, $\hat{Y} = b_0 + b_1 X$, that provides this best fit for the data. Several possible definitions of "best fit" exist, but we will focus on minimizing the sum of squared residuals. The "residual" (or "error term") measures how far Y deviates from its predicted value: $U = Y - \hat{Y} = Y - b_0 - b_1 X$. We will find the b_0 and b_1 that solve:

$$\min_{b_0, b_1} E\left[(Y - b_0 - b_1 X)^2\right] \tag{1}$$

which has the empirical analogue:

$$\min_{\hat{b}_0, \hat{b}_1} \sum_{i=1}^{N} \left(Y_i - \hat{b}_0 - \hat{b}_1 X_i\right)^2 \tag{2}$$

These minimization problems underlie ordinary least squares (OLS) regression. We will denote the solutions to minimization problem (1) β_0 and β_1, and we will denote the solutions to minimization problem (2) $\hat{\beta}_0$ and $\hat{\beta}_1$.[1] Note that $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimators for the parameters of the statistical model $Y_i = \beta_0 + \beta_1 X_i + U_i$. Also note that the estimates we obtain through OLS do not necessarily have a causal interpretation. They merely express Y as a linear function of X, and only with additional assumptions can we say anything about the causal effect of X on Y.

2 Deriving the OLS Estimator

In minimization problem (1), the first order condition for b_1 is:

$$\begin{aligned} E\left[-2X\left(Y - \beta_0 - \beta_1 X\right)\right] &= 0 \\ -E[XY] + \beta_0 E[X] + \beta_1 E[X^2] &= 0 \\ \beta_0 &= \frac{E[XY] - \beta_1 E[X^2]}{E[X]} \end{aligned}$$

[1] So $b_0^* = \beta_0$, $b_1^* = \beta_1$, $\hat{b}_0^* = \hat{\beta}_0$, and $\hat{b}_1^* = \hat{\beta}_1$

The first order condition for b_0 is:

$$\begin{aligned} E\left[-2\left(Y-\beta_0-\beta_1 X\right)\right] &= 0 \\ \beta_0 - E[Y-\beta_1 X] &= 0 \\ \beta_0 &= E[Y]-\beta_1 E[X] \end{aligned}$$

Combine these to obtain:

$$\begin{aligned} E[XY]-\beta_1 E[X^2] &= -\beta_1 E[X]^2 + E[Y]E[X] \\ \beta_1 &= \frac{E[XY]-E[Y]E[X]}{E[X^2]-E[X]^2} \\ &= \frac{cov(Y,X)}{V[X]} \end{aligned}$$

Thus we have:

$$\begin{aligned} \beta_0 &= E[Y]-\beta_1 E[X] \\ \beta_1 &= \frac{cov(Y,X)}{V[X]} \end{aligned}$$

We will estimate β_0 and β_1 using their empirical analogues:

$$\begin{aligned} \hat{\beta}_0 &= \bar{Y}-\hat{\beta}_1\bar{X} \\ \hat{\beta}_1 &= \frac{\frac{1}{N}\sum_i (Y_i-\bar{Y})(X_i-\bar{X})}{\frac{1}{N}\sum_i (X_i-\bar{X})^2} \end{aligned}$$

In your spare time, you can check that these expressions satisfy the first order conditions of minimization problem (2): $\min_{\hat{b}_0,\hat{b}_1}\sum_i\left(Y_i-\hat{b}_0-\hat{b}_1 X_i\right)^2$.

Let's check a few properties of the parameters β_0 and β_1. What is $E[U]$?

$$\begin{aligned} E[U] &= E\left[Y-\left(E[Y]-\frac{cov(Y,X)}{V[X]}E[X]\right)-\frac{cov(Y,X)}{V[X]}X\right] \\ &= E[Y]-E[Y]+\frac{cov(Y,X)}{V[X]}E[X]-\frac{cov(Y,X)}{V[X]}E[X] \\ &= 0 \end{aligned}$$

And what is $cov(U, X)$?

$$\begin{aligned}
cov(U,X) &= E[UX] - E[U]E[X] \\
&= E[YX] - \beta_0 E[X] - \beta_1 E[X^2] - E[Y]E[X] + \beta_0 E[X] + \beta_1 E[X]^2 \\
&= E[YX] - E[Y]E[X] + \frac{cov(Y,X)}{V[X]} E[X]^2 - \frac{cov(Y,X)}{V[X]} E[X^2] \\
&= cov(Y,X) + \frac{cov(Y,X)}{V[X]} \left(E[X]^2 - E[X^2] \right) \\
&= cov(Y,X) + \frac{cov(Y,X)}{V[X]} \left(-V[X] \right) \\
&= cov(Y,X) - cov(Y,X) = 0
\end{aligned}$$

Thus, linear regression coefficients satisfy the following two conditions: $E[U] = 0$ and $cov(U, X) = 0$. The empirical analogues $\frac{1}{N}\sum_i \hat{U}_i = 0$ and $\frac{1}{N}\sum_i (X_i - \bar{X})\hat{U}_i = 0$ also hold (where $\hat{U}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$).

3 The Gauss-Markov Assumptions

In the classical setup for ordinary least squares, X_i is considered non-random. This assumption is easily relaxed, but it simplifies the math slightly, so we will maintain it here. The result of this assumption is U_i is the only random variable that concerns us. We make the following assumptions about U_i:

1. $E[U_i] = 0$ for all i
2. $V[U_i] = \sigma^2$ for all i
3. $cov(U_i, U_j) = 0$ for all $i \neq j$

These assumptions are also known as the Gauss-Markov assumptions, after the mathematicians who proved the following seminal result:

Gauss Markov Theorem: *Under assumptions (1)-(3), the ordinary least squares estimator is the best linear unbiased estimator (BLUE).*

Each of the components of the BLUE acronym deserves further explanation. The word "best" means that the estimator is "efficient" or "minimum variance." The word "linear" means that the estimator is a linear function of $Y_1, \cdots, Y_N$. The word "unbiased" means that $E\left[\left(\hat{\beta}_0, \hat{\beta}_1\right)\right] = (\beta_0, \beta_1)$. Thus, a BLUE gives us the most precise answer that can be attained with a linear estimator. In Lecture Note 1, we proved that $\bar{X}$ is the BLUE of μ_X.

4 Inference for the OLS Estimator

We will now derive the variance of the OLS slope estimator under classical assumptions (1)-(3), and under the assumption that X_i is non-random. We will focus on the slope estimator $\hat{\beta}_1$ rather than the intercept $\hat{\beta}_0$ because our interest almost always lies in the former. To start, let's come up with a slightly different expression for $\hat{\beta}_1$. You'll see in a bit why the new expression will be useful.

$$\begin{aligned}
\hat{\beta}_1 &= \frac{1}{\sum_i (X_i - \bar{X})^2} \sum_i \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) \\
&= \frac{1}{\sum_i (X_i - \bar{X})^2} \sum_i \left(X_i - \bar{X}\right)\left(\beta_0 + \beta_1 X_i + U_i - \beta_0 - \beta_1 \bar{X} - \bar{U}\right) \\
&= \frac{1}{\sum_i (X_i - \bar{X})^2} \sum_i \left(X_i - \bar{X}\right)\left(\beta_1 X_i + U_i - \beta_1 \bar{X} - \bar{U}\right) \\
&= \frac{1}{\sum_i (X_i - \bar{X})^2} \sum_i \left(X_i - \bar{X}\right)\left(\beta_1 \left(X_i - \bar{X}\right) + \left(U_i - \bar{U}\right)\right) \\
&= \beta_1 + \frac{1}{\sum_i (X_i - \bar{X})^2} \sum_i \left(X_i - \bar{X}\right)\left(U_i - \bar{U}\right) \\
&= \beta_1 + \frac{1}{\sum_i (X_i - \bar{X})^2} \sum_i \left(X_i - \bar{X}\right) U_i
\end{aligned}$$

As a side note, this expression very directly implies that $E[\hat{\beta}_1] = \beta_1$, thus confirming the unbiasedness of the OLS estimator. It can also be shown that the expression $\frac{\sum_i (X_i - \bar{X}) U_i}{\sum_i (X_i - \bar{X})^2}$ converges in probability to zero, so that $\hat{\beta}_1$ is a consistent estimator of β_1.

Because X_i is non-random, we can treat all of the terms involving X_i as constant terms. Recall that if c_k is a series of constant terms and Z_k is a series of random variables, then $V\left[\sum_k c_k Z_k\right] = \sum_k c_k^2 V\left[Z_k\right] + \sum_k \sum_{l \neq k} c_k c_l cov(Z_k, Z_l)$. Thus, if we take the variance of the estimator above, we obtain:

$$\begin{aligned}
V\left[\hat{\beta}_1\right] &= V\left[\hat{\beta}_1 - \beta_1\right] \\
&= V\left[\frac{1}{\sum_i (X_i - \bar{X})^2} \sum_i \left(X_i - \bar{X}\right) U_i\right] \\
&= \left(\frac{1}{\sum_i (X_i - \bar{X})^2}\right)^2 \left[\sum_i \left(X_i - \bar{X}\right)^2 V\left[U_i\right] + \sum_i \sum_{j \neq i} \left(X_i - \bar{X}\right)\left(X_j - \bar{X}\right) cov\left(U_i, U_j\right)\right] \\
&= \left(\frac{1}{\sum_i (X_i - \bar{X})^2}\right)^2 \sum_i \left(X_i - \bar{X}\right)^2 \sigma^2 \\
&= \frac{\sigma^2}{\sum_i \left(X_i - \bar{X}\right)^2} \\
&= \frac{\sigma^2}{NV[X]}
\end{aligned}$$

The fourth line follows from assumptions (2) and (3).

We have an estimator for the denominator of this expression, but we still need to estimate σ^2. An unbiased (and consistent) estimator for σ^2 is:

$$s^2 = \hat{\sigma}^2 = \frac{1}{N-2}\sum_i U_i^2$$

This expression is very similar to the variance estimator we developed in Lecture Note 1. First, because the mean of U_i is zero, $\left(U_i - \bar{U}\right)^2 = U_i^2$. Second, in both cases, we divide the sum of squared deviations by the sample size minus the number of parameters we needed to estimate along the way. This denominator is known as the *degrees of freedom* adjustment. When we estimate the variance of X_i, we only need to estimate $\bar{X}$ before we perform our final computation, leaving us with $N-1$ degrees of freedom. When we estimate the variance of U_i, we need to estimate both β_0 and β_1, leaving us with $N-2$ degrees of freedom. In large samples, these adjustments will not matter much; an estimator that used N as the denominator would still be consistent. We call s^2 the *standard error of the regression.*

We are now equipped to derive an estimator for $V[\hat{\beta}_1]$:

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{s^2}{N\hat{\sigma}^2_X}$$

As in Lecture Note 1, the central limit theorem guarantees that as the sample size grows, $\hat{\beta}_1$ approaches a normal distribution with mean β_1 and variance $V\left[\hat{\beta}_1\right]$. The (asymptotic) standard error of $\hat{\beta}_1$ is $SE\left(\hat{\beta}_1\right) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}$, and the t-statistic $t = \frac{\hat{\beta}_1 - \beta_1^0}{SE\left(\hat{\beta}_1\right)}$ has a standard normal distribution for the null hypothesis that $\beta_1 = \beta_1^0$.

As a result, we can use this t-statistic to test hypotheses and construct confidence intervals as in Lecture Note 1. Let's calculate the large-sample p-value for the hypothesis that $\beta_1 = \beta_1^0$ against the alternative $\beta_1 \neq \beta_1^0$. We'll have to change notation slightly and call $\hat{\beta}_1^{act}$ the actual estimate we obtained in our data. Then the p-value will be the probability of observing a value of the estimator $\hat{\beta}_1$ that is at least as extreme as $\hat{\beta}_1^{act}$:

$$\begin{aligned}
p &= Pr\left[\left|\hat{\beta}_1 - \beta_1^0\right| \geq \left|\hat{\beta}_1^{act} - \beta_1^0\right|\right] \\
&= Pr\left[\frac{\left|\hat{\beta}_1 - \beta_1^0\right|}{SE\left(\hat{\beta}_1\right)} \geq \frac{\left|\hat{\beta}_1^{act} - \beta_1^0\right|}{SE\left(\hat{\beta}_1\right)}\right] \\
&= Pr\left[\frac{\left|\hat{\beta}_1 - \beta_1^0\right|}{SE\left(\hat{\beta}_1\right)} \geq t\right] \\
&= 2\Phi[-t] = 2(1-\Phi[t])
\end{aligned}$$

where the last line follows because t is a draw from a standard normal distribution in large samples. As you learned last semester, $\hat{\beta}_1$ is statistically different from β_1^0 at the 5% level if $|t| > 1.96$.

The confidence interval for β_1 follows similar logic. Let's go back to referring to our estimate of β_1 as $\hat{\beta}_1$. Then the confidence interval for confidence level $1 - 2\alpha$ is:

$$CI = \left[\hat{\beta}_1 - z_{\alpha/2} SE\left(\hat{\beta}_1\right), \hat{\beta}_1 + z_{\alpha/2} SE\left(\hat{\beta}_1\right)\right]$$

where $z_{\alpha/2}$ is the standard normal critical value for $1 - \frac{\alpha}{2}$. For instance, the 95% confidence interval uses $z_{\alpha/2} = \Phi^{-1}\left[1 - \frac{0.05}{2}\right] \approx 1.96$.

5 Common Alternative Assumptions for OLS

5.1 Normal Linear Model

All of the preceding results are asymptotic; currently, we know little about the distribution of $\hat{\beta}_1$ in small samples. We can obtain the *exact* sampling distribution of $\hat{\beta}_1$, however, if we make an additional assumption about the distribution of the residual U_i. If, in addition to assumptions (1)-(3) above, $U_i \sim \mathcal{N}(0, \sigma^2)$, then the t-statistic $t = \frac{\hat{\beta}_1 - \beta_1^0}{SE\left(\hat{\beta}_1\right)}$ is distributed according to a Student's t-distribution with $n - 2$ degrees of freedom. We will not derive this result in class, but it can come in handy when the sample size is too small to justify the premise that $N \to \infty$. In fact, Stata computes all of its p-values and confidence intervals using the $t(N - 2)$ distribution. In large samples, the adjustment does not matter, but we should note that there is no theoretical justification for it in the absence of the assumption that U_i has a normal distribution.

5.2 Random X's

The preceding results also assumed that X_i was non-random, which may strike you as implausible for many situations. Fortunately, if we condition on the X_i's, the preceding results still hold when X_i is random. Thus, the least squares assumptions become:

1. $E[U_i | X_1, X_2, \cdots, X_N] = 0$ for all i
2. $V[U_i | X_1, X_2, \cdots, X_N] = \sigma^2$ for all i
3. $cov(U_i, U_j | X_1, X_2, \cdots, X_N) = 0$ for all $i \neq j$

and the normality assumption in the normal linear model becomes $U_i | X_1, X_2, \cdots, X_N \sim \mathcal{N}(0, \sigma^2)$. The results also become conditional on the X_i's: for instance, $E[\hat{\beta}_1 | X_1, \cdots, X_N]$ and $V[\hat{\beta}_1 | X_1, \cdots, X_N]$. But aside from this more cumbersome notation, we introduce no new complications by allowing X_i to be random.

5.3 Heteroskedasticity-Robust Standard Errors

In practice, researchers rarely have good justification for assuming that the error terms have constant variance, i.e. that the errors are *homoskedastic*. Fortunately, an alternative estimator of the standard error of $\hat{\beta}_1$ allows for heteroskedasticity: $V[U_i|X_i]$ may vary with i. Chapter 4 of the textbook treats this as the main case of OLS, and indeed it is the workhorse regression method of empirical work in economics and other quantitative social sciences. For this method, we will make three assumptions:

1. $E[U_i|X_i] = 0$
2. (X_i, Y_i) are independently and identically distributed (i.i.d.) draws from their joint distribution.
3. X_i and Y_i have non-zero finite fourth moments, meaning that large outliers are unlikely.

Assumption (1) is identical to before. Assumption (2) is a slightly stricter and more intuitive version of our previous assumption that $cov(U_i, U_j|X_i) = 0$ for all $i \neq j$. The covariance assumption is a statistical condition that bears no obvious relation to the real world. In contrast, the i.i.d. assumption basically implies that the observations in our sample are unconnected but comparable. Assumption (3) is a technicality that is necessary for proofs we won't do in this class. The basic idea is that if extreme outliers are likely, $\hat{\beta}_1$ will not converge in distribution to a normal distribution.

Under these three alternative assumptions, the estimator for the variance of $\hat{\beta}_1$ is:

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{N} \frac{\frac{1}{N-2} \sum_i \left(X_i - \bar{X}\right)^2 U_i^2}{\left[\frac{1}{N} \sum_i (X_i - \bar{X})^2\right]^2}$$

and $SE\left(\hat{\beta}_1\right) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}$.[2] We can then proceed with testing hypotheses and constructing confidence intervals as before. In Stata, implementation is quite easy, with the `robust` option for the `regress` command.

6 Least Squares without I.I.D.

In all of the above approaches to OLS, we have assumed that the error terms were either mutually independent or mutually uncorrelated. Here, we will discuss inference under weaker independence assumptions. The mutual independence of the observations in our sample might break down if, for example, (1) the sample comes from a survey with a cluster design, (2) the sample is a group of students who are separated into classrooms, or (3) the sample includes multiple observations per unit (e.g., per person). None of these scenarios guarantees a violation of the i.i.d. assumption, but they make such a violation more likely.

[2] Technically, I should write the variance and standard error as conditional on $X_1, X_2, \cdots, X_N$ (e.g., $SE\left(\hat{\beta}_1|X_1, \cdots, X_N\right)$). To simplify the exposition, I omit this detail. Stock and Watson do the same.

To understand these issues better, it is useful to rewrite our regression specification in vector form:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix} = \beta_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix} + \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_N \end{pmatrix}$$

In the future, we will use matrix notation:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_N \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_N \end{pmatrix}$$

or:

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

As a preview of results to come, the OLS estimator for the vector of coefficients $\hat{\beta}$ is $[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Y}$. More important for the current discussion, however, is the fact that the variance of the OLS estimator is a function of the variance of the vector $\mathbf{U}$. Without any independence or covariance assumptions, the variance of $\mathbf{U}$ is a long string of variances and covariances. These variances and covariances can be represented by the $N \times N$ matrix:

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1N} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1N} & \sigma_{2N} & \cdots & \sigma_N^2 \end{pmatrix} \tag{3}$$

where $\sigma_i^2 = V[U_i]$ and $\sigma_{ij} = cov(U_i, U_j)$. We refer to this matrix as a *variance-covariance matrix*. Under the i.i.d. assumption (or our original assumption that $\sigma_{ij} = 0$ for $i \neq j$), this matrix simplifies to:

$$\begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_N^2 \end{pmatrix} \tag{4}$$

And with a homoskedasticity assumption ($\sigma_i^2 = \sigma^2$ for all i), it simplifies even further to:

$$\begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 \mathbf{I}_N \tag{5}$$

where $\mathbf{I}_N$ is the $N \times N$ identity matrix.

Matrices (3), (4), and (5) impose progressively more assumptions. Matrix (3) is completely assumption free, but it does not allow us to do any inference; with completely arbitrary covariance terms, our sample is no longer a sample but rather one giant observation. Matrix (4) is the heteroskedastic case, in which we allow σ_i^2 to vary with i but still assume i.i.d. Matrix (5) is the homoskedastic case, which also assumes i.i.d.

But there are also variance-covariance matrices that land somewhere between matrices (3), (4), and (5). These intermediate cases would arise in the above examples of possible independence violations. For instance, suppose Y were a worker's labor earnings and X were her years of education, and suppose we observed two years of earnings data for each worker. In this case, the i.i.d. assumption would be unreasonable. Instead, we can make the following assumption on the variance-covariance matrix for the residuals:

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & 0 & 0 & \cdots & 0 & 0 \\ \sigma_{12} & \sigma_2^2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \sigma_3^2 & \sigma_{34} & \cdots & 0 & 0 \\ 0 & 0 & \sigma_{34} & \sigma_4^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \sigma_{N-1}^2 & \sigma_{N-1,N} \\ 0 & 0 & 0 & 0 & \cdots & \sigma_{N-1,N} & \sigma_N^2 \end{pmatrix} \tag{6}$$

This matrix is a *block matrix*, and we will typically refer to each of its blocks as a *cluster* or *group*. Instead of using only an i subscript, we will now use the subscripts it, where i indexes the individual and t indexes the year of observation: $Y_{it} = \beta_0 + \beta_1 X_{it} + U_{it}$, with $i = 1, \cdots, N$ and $t = 1, 2$. Matrix (6) results from an i.i.d. assumption across the individuals (clusters) i but not across the years t within i. Within each individual, the covariances of the error terms are completely unrestricted. For example, if earnings never changed, then we would have $\sigma_{12} = \sigma_{34} = \cdots = \sigma_{N-1,N} = 1$ because each worker would have the same error term in both years.

In this case, the heteroskedasticity- and cluster-robust estimator of the variance of $\hat{\beta}_1$ becomes:

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{NT} \frac{\frac{1}{NT-2}\sum_i^N \left\{ \sum_t^T \sum_{s=1}^T \left(X_{it} - \bar{X}\right)\left(X_{is} - \bar{X}\right) U_{it} U_{is} \right\}}{\left[\frac{1}{NT}\sum_i^N \sum_t^T (X_{it} - \bar{X})^2\right]^2}$$

where N is the number of individuals, and T is the number or years (in this case, 2). Compare this expression with the expression under heteroskedasticity and i.i.d., and see if you can spot the similarities. The cluster-robust standard error can be larger or smaller than the standard robust standard error. If the error terms from observations within each cluster are positively correlated, the cluster-robust standard error will be larger, and if they are negatively correlated, it will be smaller.

If the X_{it}'s are constant within each i, as they are in this case (with $X_{it} = X_i =$ years of education), then an alternative to the cluster-robust standard error is to take averages within each i. Then we run: $\bar{Y}_i = \beta_0 + \beta_1 X_i + \bar{U}_i$. In this regression, as in the cluster-robust case, we are only imposing independence across i.

7 Weighted Least Squares

Sometimes we wish to weight observations differently in our estimation procedures. Suppose we have weight w_i for each observation i. Then we can solve the *weighted least squares* minimization problem:

$$\min_{\hat{b}_0^w, \hat{b}_1^w} \sum_{i=1}^N w_i \left(Y_i - \hat{b}_0^w - \hat{b}_1^w X_i\right)^2$$

The solution for the slope coefficient is $\hat{\beta}_1^w = \frac{\sum_i w_i (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_i w_i (X_i - \bar{X})^2}$.

Weighting comes up in a couple of situations you are likely to encounter. One is when you have a known form of heteroskedasticity. An example of known heteroskedasticity occurs when individuals i are in groups g, each with group size N_g. (The groups might be cities, states, countries, classrooms, etc.) Suppose that the individual-level model is $Y_{ig} = \beta_0 + \beta_1 X_{ig} + U_{ig}$, with $V[U_{ig}|X_{ig}] = \sigma^2$. But suppose also that we only observe group level averages, $\bar{Y}_g = \frac{1}{N_g}\sum_{i=1}^{N_g} Y_{ig}$ and $\bar{X}_g = \frac{1}{N_g}\sum_{i=1}^{N_g} X_{ig}$, so that we can only run the group-level regression $\bar{Y}_g = \beta_0 + \beta_1 \bar{X}_g + \bar{U}_g$. This group-level regression is similar to the regression at the end of Section 6 above. Note, however, that $V[\bar{U}_g|\bar{X}_g] = \frac{\sigma^2}{N_g}$. If group sizes vary across groups, then by using group averages, we have introduced heteroskedasticity. If we compute standard errors assuming homoskedasticity, our standard errors will be incorrect. We can fix this problem by computing heteroskedasticity-robust standard errors, but these come with an efficiency loss; since we no longer meet the Gauss-Markov assumptions, our regression estimator is no longer BLUE. But if we weight the regression using $w_g = \frac{1}{V[\bar{U}_g|\bar{X}_g]} = \frac{N_g}{\sigma^2}$, we restore

homoskedasticity and return to the efficient world of Gauss-Markov. In fact, since the denominator of w_g is the same for all g, we can just weight groups by N_g. This weighting scheme is intuitive; we give more weight to groups with more observations and therefore more precisely-estimated means.

The other situation in which you may consider running weighted least squares is when you analyze complex survey data. Often, surveys have complicated stratified designs, and not all individuals in the population are sampled with equal probability. In other cases, non-response rates are different for different sorts of people, and survey statisticians provide ex-post estimates of each individual's probability of selection. The individuals selected with different probabilities may not have identical distributions, so for these situations, it is natural to think of weighting by the inverse of the probability of selection into the sample. That is to say, if individuals i are selected for the sample with probability p_i, we may want to use weights $w_i = \frac{1}{p_i}$. These weights work very well for estimating means. Indeed, in a weighted sample, the weighted average $\bar{X}^w = \frac{1}{N}\sum_i w_i X_i$ is an unbiased and consistent estimator of the overall population mean of X_i. Matters are more complicated for regression, however. If the regression model is correct, in the sense that all individuals have the same β_1, then there is no need to weight the regression; due to the Gauss-Markov Theorem, the unweighted regression is the most efficient approach. However, you may encounter cases of heterogeneity in the regression parameters (i.e., β_{1i}). In this case, weighted regression will lead to the parameter estimates that you would have estimated using a census of the population. But this estimand is not the same as $E[\beta_i]$, so the use of survey weights in regression does not always answer a question we wish to answer. The use of survey weights in regression is quite complicated; if you want a more in-depth treatment, consult the second chapter of Angus Deaton's 1997 book *The Analysis of Household Surveys*. You will spend more time discussing survey weights in precept.

These two reasons for weighted regression have an important distinction. In the case of known heteroskedasticity, weighted least squares alters the variance of our estimator but does not change its expectation or probability limit. As discussed above, this change in the variance is usually desirable. In the case of complex survey designs, however, weighted least squares may change the expectation or probability limit of our estimator. This change in the expectation or probability limit is not always desirable.

8 Bivariate Regression Miscellania

8.1 Interpreting Regression Coefficients

When both X and Y are continuous, β_1 is the slope of the regression line: $\partial\hat{Y}/\partial X$. When X is binary, β_1 equals the difference in means between the sub-populations with $X = 1$ and $X = 0$, or $E[Y|X = 1] - E[Y|X = 0]$. In this case, regression is identical to the two-sample t-test we covered in Lecture Note 1. When Y is binary, matters are a bit different; we will discuss this case later in the semester.

Researchers commonly transform their dependent or independent variables using the logarithms. Recall that $\frac{d \ln z}{dz} = \frac{1}{z}$, or $d \ln z = \frac{1}{z} dz$. The second equality shows that an infinitesimal change in the natural logarithm of a variable z is equal to a proportional change in z of the same amount. Consequently, β_1 has the following interpretations when logarithmic transformations are used:

1. $\ln Y = \beta_0 + \beta_1 X + U$: a one-unit change in X is associated with a $100 \times \beta_1$ percent change in Y. In this case, β_1 measures the *semi-elasticity* of Y with respect to X.
2. $Y = \beta_0 + \beta_1 \ln X + U$: a one percent change in X is associated with a change in Y of $0.01 \times \beta_1$.
3. $\ln Y = \beta_0 + \beta_1 \ln X + U$: a one percent change in X is associated with a β_1 percent change in Y. In this case, β_1 measures the *elasticity* of Y with respect to X.

8.2 Goodness of Fit

Researchers often use the R^2 statistic of a regression to assess its goodness of fit. The R^2 measures the fraction of the variance of Y that is explained by X. For the population, we write:

$$R^2_{pop} = \frac{\sigma^2_{\hat{Y}}}{\sigma^2_Y} = \frac{\beta^2 \sigma^2_X}{\sigma^2_Y} = 1 - \frac{\sigma^2_U}{\sigma^2_Y}$$

We estimate R^2_{pop} using the sample analogues of these population moments: $R^2 = \frac{\sum_i \left(\hat{\beta}_1 X_i\right)^2}{\sum_i \left(Y_i - \bar{Y}\right)^2} = 1 - \frac{\sum_i U_i^2}{\sum_i \left(Y_i - \bar{Y}\right)^2}$. Note that the R^2 is NOT a measure of statistical significance. In fact, in many cases, variables with both statistically and economically important effects have low R-squareds.

WWS 508c, Econometrics and Public Policy **Prof. Tom Vogl**

Spring 2015 **Princeton University**

LECTURE NOTE 4: MULTIPLE REGRESSION

1 Introduction

Lecture Note 3 discussed the regression of a dependent variable Y on a single predictor variable X. But often we will be interested in the regression of Y on many X's:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + U + \cdots + \beta_K X_K + U \\ &= \beta_0 + \textstyle\sum_{k=1}^{K} \beta_k X_k + U \end{aligned} \tag{1}$$

In this formulation, we say that β_k represents the relationship between X_k and Y, "controlling for" (or "conditional on," or "holding constant") all other $X_{\tilde{k}}$ with $\tilde{k} \neq k$. This lecture note reviews a few of the basics of multiple regression but focuses largely on areas that 507c either did not cover or did not cover in detail.

In the sample, we observe $(Y_i, X_{1i}, \cdots, X_{Ki})_{i=1}^{N}$. The least squares minimization problem is a simple extension of the bivariate case:

$$\min_{\hat{b}_0, \cdots, \hat{b}_K} \sum_{i=1}^{N} \left(Y_i - \hat{b}_0 - \hat{b}_1 X_{1i} - \cdots - \hat{b}_K X_{Ki} \right)^2$$

For this lecture note, we will use the least squares assumptions presented in the textbook as the standard case: i.i.d. observations with heteroskedastic errors. We take the three assumptions from Lecture Note 3 and add one more to accommodate multiple regressors.

1. $E[U_i | X_{1i}, \cdots, X_{Ki}] = 0$
2. $(Y_i, X_{1i}, \cdots, X_{Ki})$ are independently and identically distributed (i.i.d.) draws from their joint distribution.
3. $X_{1i}, \cdots, X_{Ki}$ and Y_i have non-zero finite fourth moments.
4. No perfect multicollinearity. This assumption is equivalent to saying that for any $k \leq K$, the regression of any X_k on all the other X variables has an R^2 of less than 1.

Much of Lecture Note 3 dealt with violations of assumption (2). This lecture note will have much more to say about violations of assumption (1). Assumption (1) implies that U_i has mean zero and is uncorrelated with the X_{ki}'s. Since assumption (2) implies that two separate observations have uncorrelated X_{ki}'s, we have that all U_i's are uncorrelated with all X_{ki}'s. The theme of a correlation between the error term and the predictor

variables underlies much of this lecture note. Assumption (3) is the same arcane statistical condition we used before, and it requires no new attention. Assumption (4) is new, and it implies that when the X_k's are "too" correlated with each other, we cannot compute the regression.

2 Residual Regression

The concept of "controlling for" (or "conditional on," or "holding constant") is somewhat abstract. The Frisch-Waugh Theorem, which you covered briefly in 507c, provides a more intuitive way to look at it. The essence of the Frisch-Waugh Theorem is that you can estimate β_1 in equation (1) above by first regressing Y on $X_2, \cdots, X_K$, then regressing X_1 on $X_2, \cdots, X_K$, and finally regressing the residuals from the first regression on the residuals from the second.

A simple example clarifies the reasoning. Suppose the conditional expectation of Y given X_1 and X_2 is linear: $E[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. Let $U = Y - E[Y|X_1, X_2]$. Then we can write:

$$E[Y|X_2] = \beta_0 + \beta_1 E[X_1|X_2] + \beta_2 X_2 + E[U|X_2]$$

And so:

$$Y - E[Y|X_2] = \beta_1 \left(X_1 - E[X_1|X_2]\right) + \underbrace{\left(U - E[U|X_2]\right)}_{\text{new residual}}$$

The new residual satisfies the least squares assumptions. As a result, we can estimate β_1 by regressing the residual from a regression of Y on X_2 on the residual from a regression of X_1 on X_2. For simplicity, we have assumed that the conditional expectation of Y is linear, but the same result holds for the linear projection, even if the conditional expectation is non-linear. Furthermore, one can easily extend the derivation above to account for more than two X_k's.

3 True Models, Estimated Models, and Bias

The most common problems that plague OLS stem from bias or inconsistency. Many of the methods we study in the second half of the course will attempt to deal with these sources of bias. We define the bias or inconsistency by specifying a "true" model for the dependent variable and then considering how the model we estimate with the data differs from the "true" model. In many cases - in fact, in both cases below - we can think of the biases as violations of assumption (1) above: $E[U_i|X_i] = 0$.

3.1 Omitted Variables Bias

Suppose the "true" model of Y is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U \qquad \text{(long regression)}$$

But we only observe (Y, X_1), and as a result, we can only estimate the following regression:

$$Y = \alpha_0 + \alpha_1 X_1 + V \qquad \text{(short regression)}$$

We would like to estimate the "true" coefficient β_1, but we can only estimate α_1. Under what conditions are they the same? To answer this question, we define a third regression:

$$X_2 = \gamma_0 + \gamma_1 X_1 + \varepsilon \qquad \text{(auxiliary regression)}$$

In all three cases, we construct the error terms such that they have mean zero and are uncorrelated with the X_k's in the same regression. Now we derive how α_1 in the short regression relates to the parameters in the long and auxiliary regressions:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U \\ &= \beta_0 + \beta_1 X_1 + \beta_2(\gamma_0 + \gamma_1 X_1 + \varepsilon) + U \\ &= \underbrace{(\beta_0 + \beta_2\gamma_0)}_{\alpha_0} + \underbrace{(\beta_1 + \beta_2\gamma_1)}_{\alpha_1} X_1 + (U + \beta_2\varepsilon) \end{aligned}$$

Thus, $\alpha_1 = \beta_1 + \beta_2\gamma_1$. For α_1 to equal β_1, either β_2 or γ_1 must equal zero. If $\beta_2 = 0$, then X_2 has no predictive power (conditional on X_1), so the short regression is equivalent to the long regression. Meanwhile, $\gamma_1 = 0$ is equivalent to saying that X_1 and X_2 are uncorrelated. Thus, even if $\beta_2 \neq 0$, the OLS estimator $\hat{\alpha}_1$ is an unbiased and consistent estimator of β_1 if X_1 and X_2 are uncorrelated. If X_1 and X_2 are correlated, however, then $\hat{\alpha}_1$ is biased and inconsistent. This result is known as omitted variables bias.

Many economists view this problem through the lens of the error term in the short regression. We can write $V = \beta_2 X_2 + U$. If $\beta_2 \neq 0$ and $\gamma_1 \neq 0$, then V is correlated with X_1, which implies a violation of least squares assumption (1).

3.2 Measurement Error

A related bias stems from measurement error in one of the predictor variables. The case we will study here is actually an application of bivariate regression rather than multivariate regression, but I have placed it in this lecture note to accompany other sources of bias.

Suppose we are interested in the relationship between two variables, Y^* and X^*:

$$Y^* = \beta_0 + \beta_1 X^* + U \tag{2}$$

where $E[U|X^*] = 0$. However, we do not observe Y^* and X^* directly. Instead, we measure these variables with error:

$$Y = Y^* + \eta \quad \text{and} \quad X = X^* + \nu \tag{3}$$

where we assume that the measurement errors η and ν are uncorrelated with each other and with Y^* and X^*. Measurement errors with these properties are known as *classical measurement errors*. If we try to estimate β_1 using an OLS regression of Y on X, we obtain a slope coefficient with the following probability limit:

$$\begin{aligned} \hat{\beta}_1 \xrightarrow{p} \frac{cov(X,Y)}{V[X]} &= \frac{cov(Y^* + \eta, X^* + \nu)}{V[X^* + \nu]} = \frac{cov(\beta_0 + \beta_1 X^* + U + \eta, X^* + \nu)}{V[X^* + \nu]} \\ &= \beta_1 \frac{V[X^*]}{V[X^*] + V[\nu]} = \beta_1 \frac{1}{1 + V[\nu]/V[X^*]} \end{aligned}$$

Since $V[\nu] > 0$, $\hat{\beta}_1$ will underestimate β_1. We say that measurement error has *attenuated* $\hat{\beta}_1$, and we call $\text{plim } \hat{\beta}_1 - \beta_1 = -\beta_1 \frac{V[\nu]}{V[X^*]+V[\nu]}$ the *attenuation bias*.

A few comments are in order. First, note that while measurement error in X leads to attenuation bias, measurement error in Y poses no problems. You can see this in the expression above because only $V[\nu]$, not $V[\eta]$, appears in the final expression. Measurement error in Y adds to the residual variance and thus increases our standard errors, but it does not bias our point estimator. Second, we can again express the problem as a correlation between the error term and X. Substitute equations (3) into equation (2) to get:

$$\begin{aligned} Y^* &= \beta_0 + \beta_1 X^* + U \\ Y - \eta &= \beta_0 + \beta_1 (X - \nu) + U \\ Y &= \beta_0 + \beta_1 X + \underbrace{(U - \beta_1 \nu + \eta)}_{\text{error term}} \end{aligned}$$

The presence of $\beta_1 \nu$ in the error term introduces a correlation between X and the error term, violating least squares assumption (1). Finally, the discussion in this sub-section deals exclusively with classical measurement

error. If the measurement errors are correlated with each other or with Y^* or X^*, the same results will not apply.

The above results apply to bivariate regression. In the multivariate case, matters get considerably more complicated. Measurement error in a variable X_k attenuates the coefficient on X_k but has quite complicated effects on the coefficients for other variables.

4 A Matrix Version

In Lecture Note 3, we explored a matrix representation of OLS. Here, we complete the picture for the multivariate case. Equation (1), which represents the population regression function, can be re-expressed as:

$$\begin{aligned} Y &= \begin{pmatrix} 1 & X_1 & X_2 & \cdots & X_K \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix} + U \\ &= X'\beta + U \end{aligned} \tag{4}$$

The solution to the population least squares problem is $\beta = E[XX']^{-1}E[XY]$. In the bivariate case, when $X' = (1\ X_1)$ and $\beta' = (\beta_0\ \beta_1)$, it can be shown that the matrix version is the same as the estimators derived in Lecture Note 3.

In the sample, we have many observations $(Y_i, X_{1i}, \cdots, X_{Ki}, U_i)_{i=1}^N$. We write the regression equation as:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{21} & \cdots & X_{K1} \\ 1 & X_{12} & X_{22} & \cdots & X_{K2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1N} & X_{2N} & \cdots & X_{KN} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{pmatrix} + \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_N \end{pmatrix} \tag{5}$$

or $\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}$. The least squares estimator for the vector of β_k's is $\hat{\beta} = [\mathbf{X'X}]^{-1}\mathbf{X'Y}$. Its variance is the matrix $V[\hat{\beta}|\mathbf{X}] = (\mathbf{X'X})^{-1}\mathbf{X}'E[\mathbf{UU}'|\mathbf{X}]\mathbf{X}(\mathbf{X'X})^{-1}$. The expression $E[\mathbf{UU}'|\mathbf{X}]$ is the variance-covariance matrix for $\mathbf{U}$ we discussed in Lecture Note 3. *After we condition on* $\mathbf{X}$, *the only quantity that affects our standard errors is the variance of the error terms.* In the case of a simple random sample (i.i.d.) with homoskedasticity, we have $E[\mathbf{UU}'|\mathbf{X}] = \sigma^2\mathbf{I}_N$, where $\mathbf{I}_N$ is the $N \times N$ identity matrix. Then $V[\hat{\beta}|\mathbf{X}] = (\mathbf{X'X})^{-1}\left(\mathbf{X}'\sigma^2\mathbf{I}_N\mathbf{X}\right)(\mathbf{X'X})^{-1} = \sigma^2(\mathbf{X'X})^{-1}(\mathbf{X'X})(\mathbf{X'X})^{-1} = \sigma^2(\mathbf{X'X})^{-1}$. The variance-covariance matrix

for $\hat{\beta}$ is a $(K+1) \times (K+1)$ matrix:

$$V[\hat{\beta}|\mathbf{X}] = \begin{pmatrix} \sigma^2_{\hat{\beta}_0} & \sigma_{\hat{\beta}_0\hat{\beta}_1} & \cdots & \sigma_{\hat{\beta}_0\hat{\beta}_K} \\ \sigma_{\hat{\beta}_0\hat{\beta}_1} & \sigma^2_{\hat{\beta}_1} & \cdots & \sigma_{\hat{\beta}_1\hat{\beta}_K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\hat{\beta}_0\hat{\beta}_K} & \sigma_{\hat{\beta}_1\hat{\beta}_K} & \cdots & \sigma^2_{\hat{\beta}_K} \end{pmatrix}$$

Note that in practice, we have to estimate this matrix using the residuals we observe in our data. The estimated variance-covariance matrix has hats over all the V's and σ's.

The distribution of $\hat{\beta}$ has the same properties as in the bivariate case. First, $\hat{\beta}$ is unbiased: $E[\hat{\beta}|\mathbf{X}] = \beta$. This result implies that each of the components $\hat{\beta}_k$ is also unbiased: $E[\hat{\beta}_k|\mathbf{X}] = \beta_k$. Second, $\hat{\beta}$ is consistent: $\hat{\beta} \xrightarrow{p} \beta$. This result implies that each of the components $\hat{\beta}_k$ is also consistent: $\hat{\beta}_k \xrightarrow{p} \beta_k$. Third, a multivariate version of the Central Limit Theorem holds. In particular, as the sample size grows to infinity, $\hat{\beta}$ approaches a multivariate normal distribution with variance-covariance matrix $V[\hat{\beta}|\mathbf{X}]$. We write $\hat{\beta} \xrightarrow{d} \mathcal{N}\left(\beta, V[\hat{\beta}|\mathbf{X}]\right)$, where $\xrightarrow{d}$ means "converges in distribution." The fact that all the $\hat{\beta}_k$'s have a joint normal distribution means that the marginal distribution of each $\hat{\beta}_k$ is also normal: $\hat{\beta}_k \xrightarrow{d} \mathcal{N}\left(\beta_k, \sigma^2_{\hat{\beta}_k}\right)$. Thus, for hypothesis tests on single coefficients, we can proceed as we did in Lecture Note 3.

5 Inference on Linear Combinations of Coefficients

Often, however, we will want to test hypotheses on linear combinations of coefficients. We have a $(1+K)\times 1$ coefficient vector $\hat{\beta}$ and its estimated $(K+1) \times (K+1)$ variance-covariance matrix $\hat{V}[\hat{\beta}|\mathbf{X}]$:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_K \end{pmatrix} \qquad \hat{V}[\hat{\beta}|\mathbf{X}] = \begin{pmatrix} \hat{\sigma}^2_{\hat{\beta}_0} & \hat{\sigma}_{\hat{\beta}_0\hat{\beta}_1} & \cdots & \hat{\sigma}_{\hat{\beta}_0\hat{\beta}_K} \\ \hat{\sigma}_{\hat{\beta}_0\hat{\beta}_1} & \hat{\sigma}^2_{\hat{\beta}_1} & \cdots & \hat{\sigma}_{\hat{\beta}_1\hat{\beta}_K} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_{\hat{\beta}_0\hat{\beta}_K} & \hat{\sigma}_{\hat{\beta}_1\hat{\beta}_K} & \cdots & \hat{\sigma}^2_{\hat{\beta}_K} \end{pmatrix}$$

When we wish to take a linear combination of our coefficients using the $(1+K)\times 1$ vector of constants l, then $l'\hat{\beta} \xrightarrow{p} l'\beta$, and $\hat{V}[l'\hat{\beta}|\mathbf{X}] = l'\hat{V}[\hat{\beta}|\mathbf{X}]l$. For instance, suppose we wanted to estimate $\beta_0 + 3\beta_1$. Then we have:

$$l'\hat{\beta} = \begin{pmatrix} 1 & 3 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_K \end{pmatrix} = \hat{\beta}_0 + 3\hat{\beta}_1$$

And:

$$\hat{V}[l'\hat{\beta}|\mathbf{X}] = \begin{pmatrix} 1 & 3 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \hat{\sigma}^2_{\hat{\beta}_0} & \hat{\sigma}_{\hat{\beta}_0\hat{\beta}_1} & \cdots & \hat{\sigma}_{\hat{\beta}_0\hat{\beta}_K} \\ \hat{\sigma}_{\hat{\beta}_0\hat{\beta}_1} & \hat{\sigma}^2_{\hat{\beta}_1} & \cdots & \hat{\sigma}_{\hat{\beta}_1\hat{\beta}_K} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_{\hat{\beta}_0\hat{\beta}_K} & \hat{\sigma}_{\hat{\beta}_1\hat{\beta}_K} & \cdots & \hat{\sigma}^2_{\hat{\beta}_K} \end{pmatrix} \begin{pmatrix} 1 \\ 3 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \hat{\sigma}^2_{\hat{\beta}_0} + 9\hat{\sigma}^2_{\hat{\beta}_1} + 6\hat{\sigma}_{\hat{\beta}_0\hat{\beta}_1}$$

Because the $\hat{\beta}_k$'s are jointly normally distributed in large samples, their linear combination is also normally distributed. This result is exactly analogous to the case when we add two random variables X and Y: $V[X+Y] = V[X] + V[Y] + 2cov(X,Y)$. The linear combination of coefficients has many applications. To generate predicted values for each individual, set the first entry in l to be 1, and assign each individual's true X_k to the $k+1^{th}$ entry in l. (This gives us $l'\hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_K X_{Ki}$.) To test whether $\beta_1 = \beta_2$, define $l' = \begin{pmatrix} 0 & 1 & -1 & 0 & \cdots & 0 \end{pmatrix}$, so that we are doing inference on $\hat{\beta}_1 - \hat{\beta}_2$. A test of whether $\beta_1 - \beta_2 = 0$ is equivalent to a test of whether $\beta_1 = \beta_2$. Finally, note that a hypothesis test for a single coefficient is a specific case in which the $k+1^{th}$ entry of l is 1 and all other entries are zero. To implement these tests (or to compute confidence intervals for linear combinations of coefficients), you can use Stata's `lincom` command after running a regression.

6 Other Topics in Multiple Regression

To finish the lecture note, we briefly touch on some concepts you learned in 507c.

6.1 Testing Multiple Hypotheses

We have discussed how to test single hypotheses about linear combinations of coefficients. For a test of multiple hypotheses (e.g., $\beta_1 = \beta_2 = \beta_3 = 0$), we use an F-test. The F-test for q restrictions on β is based on a

statistic (called the F-statistic) that, in large samples, is drawn from an $F_{q,\infty}$ distribution. The `reg` command in Stata automatically computes an overall F-statistic for the regression, which tests the null hypothesis that all of the β_k's equal zero. The overall F-statistic is a test of the regression's goodness-of-fit.

6.2 R-squared and Adjusted R-squared

The R^2 provides another measure of goodness-of-fit, although note that it is not a *test* of goodness-of-fit. Recall from Lecture Note 3 that the R-squared in the population is defined as $R^2_{pop} = \frac{\sigma^2_{\hat{Y}}}{\sigma^2_Y} = 1 - \frac{\sigma^2_U}{\sigma^2_Y}$. This definition still holds with multiple regressors, only now $\hat{Y}$ is a function of many X_k's, rather than only one. We can therefore calculate the same sample R^2 as before:

$$R^2 = 1 - \frac{\sum_i \hat{U}_i^2}{\sum_i (Y_i - \bar{Y})^2} = 1 - \frac{SSR}{TSS}$$

where SSR is the sum of squared residuals and TSS is the total sum of squares. R^2 is a consistent estimator for R^2_{pop}. In practice, however, a problem arises in finite samples because the addition of a covariate will *always* increase the R^2 unless the coefficient on the new covariate is *exactly* zero. An alternative is to calculate the "adjusted R^2," which we define as follows:

$$\bar{R}^2 = 1 - \frac{n-1}{n-K-1}\frac{SSR}{TSS} = 1 - \frac{\sum_i \hat{U}_i^2/(n-K-1)}{\sum_i (Y_i - \bar{Y})^2/(n-1)} = 1 - \frac{s^2_{\hat{U}}}{s^2_Y}$$

where K is the number of regressors. $\bar{R}^2$ incorporates the degrees of freedom adjustments we make to obtain unbiased estimators of the variances of $\hat{U}_i$ and Y_i. Like R^2, $\bar{R}^2$ is a consistent estimator for R^2_{pop}. But unlike R^2, $\bar{R}^2$ does not mechanically grow with the number of covariates. The addition of a covariate reduces $\frac{SSR}{TSS}$ but increases $\frac{n-1}{n-K-1}$, and the net effect is ambiguous. In this way, $\bar{R}^2$ adjusts for the mechanical increase in the R^2 from adding covariates. It provides a better measure of whether additional covariates have real explanatory power.

The R^2 measures the explanatory power of the independent variables, and the F-statistic tests the hypothesis that the independent variables have no explanatory power. In this sense, both are measures of goodness-of-fit; indeed, the two statistics are related. In the homoskedastic case, the overall F-statistic is a simple increasing function of the (unadjusted) R^2 and the sample size. That is to say, for a given R^2_{pop}, a larger sample increases the likelihood of rejecting the null hypothesis that all β_k's equal zero. Similarly, for a given sample size, a larger R^2_{pop} does the same.

6.3 Interactions

Interactions between two or more independent variables will be crucial to some topics later in the course. Because you already discussed interactions in 507c, here we'll just examine a single example. Suppose we want to know how the returns to education vary by race and sex. Let $earnings_i$ be a measure of worker i's earnings; let $male_i$ be a dummy variable for i being male; and let $black_i$ and $other_i$ be dummies for black and "other" race, respectively (the omitted category is white). The following "fully interacted" regression specification is a natural starting point:

$$\begin{aligned} ln(earnings_i) &= \beta_0 + \beta_1 edyrs_i + \beta_2 male_i + \beta_3 black_i + \beta_4 other_i + \\ &\quad \beta_5(edyrs_i \cdot male_i) + \beta_6(edyrs_i \cdot black_i) + \beta_7(edyrs_i \cdot other_i) + \\ &\quad \beta_8(male_i \cdot black_i) + \beta_9(male_i \cdot other_i) + \\ &\quad \beta_{10}(edyrs_i \cdot male_i \cdot black_i) + \beta_{11}(edyrs_i \cdot male_i \cdot other_i) + \varepsilon_i \end{aligned}$$

This regression allows the returns to education to vary across all race-gender groups. The return to education among white women is β_1, while the return among black women is $\beta_1 + \beta_6$. The return among white men is $\beta_1 + \beta_5$, while the return among black men is $\beta_1 + \beta_5 + \beta_6 + \beta_{10}$. Meanwhile, after adjusting for educational attainment (with race/sex-specific returns), the log earnings gap between black women and men of "other" racial background is $\beta_3 - \beta_2 - \beta_4 - \beta_9$. Make sure you understand these results; if you don't, spend some time with Chapter 8 of Stock and Watson.

LECTURE NOTE 5: MAXIMUM LIKELIHOOD ESTIMATION

1 Introduction

In the next few lectures, we will be discussing methods for analyzing categorical outcome variables, such as logit and probit models. Although one can estimate these models using a non-linear variant of least squares, most statistical software uses another method called maximum likelihood. In this lecture note, we will first develop a general framework for maximum likelihood estimation. Then we will explore the use of maximum likelihood methods in two familiar applications: a Bernoulli random variable and the normal linear model. In Lecture Note 6, we will apply these methods to the binomial discrete choice problem.

2 General Case

Maximum likelihood methods represent a significant departure from the way we have been approaching estimation. The least squares approach asked: what parameter value is least wrong in describing the observed data points? In contrast, maximum likelihood asks: what parameter values make the observed data points most likely? In this section, we develop a general framework for maximum likelihood estimation (MLE).

Let X be a random variable with probability density function $f(x;\theta)$, where θ is a vector of parameters. Suppose we have an i.i.d. sample of observations $\{X_i\}_{i=1}^N$, and each observation i has realization x_i. The maximum likelihood estimator of θ chooses the value $\hat{\theta}$ that maximizes the joint probability of observing $X_i = x_i$ for all i. Because the X_i's are independent, we can write this joint probability as:

$$L = \prod_{i=1}^{N} f(x_i;\theta)$$

L is known as the *likelihood* or the *likelihood function*. Unfortunately, θ is unknown, so we cannot calculate L directly. But we can try to plug in candidate values C to see how likely the observed data are, given the value of $\tilde{\theta}$. The maximum likelihood estimator $\hat{\theta}$ is the value of $\tilde{\theta}$ that maximizes the likelihood. In other words, $\hat{\theta}$ is the solution to:

$$\max_{\tilde{\theta}} L = \max_{\tilde{\theta}} \prod_{i=1}^{N} f(x_i;\tilde{\theta})$$

In practice, it's quite difficult to maximize a very long product. We can simplify the problem by taking the logarithm of the likelihood. Because the logarithm is a strictly increasing transformation, the value of $\tilde{\theta}$ that maximizes L is the same as the value of $\tilde{\theta}$ that maximizes $\ln(L)$. Thus, we can write the MLE problem as:

$$\max_{\tilde{\theta}} \ln L = \max_{\tilde{\theta}} \sum_{i=1}^{N} \ln\left(f(x_i;\tilde{\theta})\right)$$

The solution $\hat{\theta}$ satisfies $\frac{\partial \ln L}{\partial \hat{\theta}} = 0$.

Under a fairly innocuous set of assumptions, the MLE satisfies three useful properties:

1. Consistency: $\hat{\theta} \xrightarrow{p} \theta$.
2. Asymptotic Normality: $\hat{\theta} \xrightarrow{d} \mathcal{N}(\theta, \Sigma)$. That is to say, in large samples, $\hat{\theta}$ is approximately normally distributed with a variance-covariance matrix that we here call Σ. There is an explicit solution for Σ, but you do not need to know it.[1]
3. Asymptotic Efficiency: No other asymptotically unbiased estimator has a smaller variance than$\hat{\theta}$.

Consistency is a basic property we want for nearly every estimator we use. Asymptotic normality implies that we can compute p-values and CIs using the same normal approximations we used with OLS. Asymptotic efficiency tells us that the MLE is in some sense a "best" estimator. It is the reason that most statistical software estimates logits and probits using maximum likelihood rather than non-linear least squares. In general, maximum likelihood is more efficient.

3 Examples

3.1 Maximum Likelihood Estimation of p from a Bernoulli Random Variable

Consider a Bernoulli random variable X_i that takes on value 1 with probability p and 0 with probability $1-p$. Suppose we observe a sample with three i.i.d. observations, $(1,1,0)$. The likelihood function is:

$$\begin{aligned} L &= Pr[X_1 = 1] \cdot Pr[X_2 = 1] \cdot Pr[X_3 = 0] \\ &= p \cdot p \cdot (1-p) \\ &= p^2(1-p) \end{aligned}$$

Thus, the MLE problem becomes:

$$\max_{\tilde{p}} L = \max_{\tilde{p}} \tilde{p}^2(1-\tilde{p})$$

[1] The formula for Σ is beyond the scope of this class, but it you are interested, it is equal to minus the second derivative matrix of the log likelihood function. This result makes sense. The second derivative measures the curvature of the likelihood function. When the likelihood function is very curved, we can be most confident in the optimum we chose as our MLE.

Or in terms of the log-likelihood:

$$\max_{\tilde{p}} \ln L = \max_{\tilde{p}} 2\ln(\tilde{p}) + \ln(1-\tilde{p})$$

Both maximization problems lead to the same first order condition, but just for example, consider the first order condition for the second maximization problem: $\frac{d\ln L}{d\hat{p}} = \frac{2}{\hat{p}} - \frac{1}{1-\hat{p}} = 0$. Rearranging terms, we obtain $\hat{p} = \frac{2}{3}$, which is exactly the same as our usual estimator for p, $\bar{X}$.

One can generalize this example to an i.i.d. sample of N Bernoulli random variables with probability of success p. Using similar logic to the $N = 3$ example above, if we observe S successes in the sample, the likelihood is:

$$L = p^S (1-p)^{N-S}$$

We then maximize the log-likelihood:

$$\max_{\tilde{p}} \ln L = \max_{\tilde{p}} S\ln(\tilde{p}) + (N-S)\ln(1-\tilde{p})$$

which gives us the first order condition $\hat{p} = \frac{S}{N}$: again exactly the same as the sample average.

3.2 Maximum Likelihood Estimation of the Normal Linear Model

Let $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and ε_i i.i.d. and independent of X_i. Equivalently, we also have that $Y_i - \beta_0 - \beta_1 X_i \sim \mathcal{N}(0, \sigma^2)$. Then the likelihood for the i^{th} observation is:

$$f\left(Y_i | X_i; \beta_0, \beta_1, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2\right)$$

which is the probability density function of a normal random variable. Then the likelihood for the entire sample is the product of the individual likelihoods:

$$L = \prod_{i=1}^{N} f\left(Y_i | X_i; \beta_0, \beta_1, \sigma^2\right) = \prod_{i=1}^{N} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2\right)\right\}$$

That is a beast of a product, so now you can see why we take logs. The log-likelihood is:

$$\ln L = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(Y_i - \beta_0 - \beta_1 X_i\right)^2$$

The MLE problem is to choose values $\hat{\beta}_0$ and $\hat{\beta}_1$ that maximize $\ln L$. The first order conditions are:

$$\begin{aligned}\frac{\partial \ln L}{\partial \hat{\beta}_1} = 0 \quad &= - \quad \frac{1}{2\sigma^2} \sum_{i=1}^{N} 2\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) X_i \\ 0 \quad &= \quad \sum_{i=1}^{N} \left(Y_i X_i - \hat{\beta}_0 X_i\right) - \hat{\beta}_1 \sum_{i=1}^{N} X_i^2\end{aligned}$$

And:

$$\begin{aligned}\frac{\partial \ln L}{\partial \hat{\beta}_0} = 0 \quad &= \quad \frac{1}{2\sigma^2} \sum_{i=1}^{N} 2\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) \\ 0 \quad &= \quad \sum_{i=1}^{N} Y_i - N\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{N} X_i \\ \hat{\beta}_0 \quad &= \quad \frac{1}{N} \sum_{i=1}^{N} Y_i - \hat{\beta}_1 \left(\frac{1}{N} \sum_{i=1}^{N} X_i\right) \\ &= \quad \bar{Y} - \hat{\beta}_1 \bar{X}\end{aligned}$$

Now plug back into the first FOC:

$$\begin{aligned}0 \quad &= \quad \sum_{i=1}^{N} Y_i X_i - \hat{\beta}_0 \sum_{i=1}^{N} X_i - \hat{\beta}_1 \sum_{i=1}^{N} X_i^2 \\ &= \quad \sum_{i=1}^{N} Y_i X_i - \left(\sum_{i=1}^{N} Y_i - \hat{\beta}_1 \sum_{i=1}^{N} X_i\right) \sum_{i=1}^{N} X_i - \hat{\beta}_1 \sum_{i=1}^{N} X_i^2 \\ &= \quad \sum_{i=1}^{N} Y_i X_i - \left(\sum_{i=1}^{N} Y_i\right)\left(\sum_{i=1}^{N} X_i\right) - \hat{\beta}_1 \left[\sum_{i=1}^{N} X_i^2 - \left(\sum_{i=1}^{N} X_i\right)^2\right] \\ \hat{\beta}_1 \quad &= \quad \frac{\sum_{i=1}^{N} \left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right)}{\sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2}\end{aligned}$$

The maximum likelihood estimates of the coefficients from the normal linear model are the same as those obtained through OLS!

4 Maximum Likelihood Estimation in Practice

In the examples above, we derived analytic solutions to the maximum likelihood problem. But analytic solutions are not always possible; indeed, most of the models we estimate using maximum likelihood lack a closed-form solution. We have three options for finding the optimal $\hat{\theta}^{MLE}$:

1. Analytic optimization: As already mentioned, we can differentiate the likelihood function, set the first derivative to zero, and solve. We should check the second derivative to make sure we have found a

maximum (rather than a minimum).

2. (Undirected) grid search: If we know that θ lies in some range $[\underline{\theta}, \bar{\theta}]$, then we can calculate L for all the values of a grid over $[\underline{\theta}, \bar{\theta}]$ and choose the value that produces the highest L.
3. (Directed) numerical optimization: A number of algorithms exist to use the properties of the likelihood function (e.g., the first and second derivatives) to direct numerical search. So we choose an initial value, $\hat{\theta}^0$, and we then allow the algorithm to find the optimum.

The most attractive method is (1), analytic optimization, but this option is often impossible. Option (2) can be extremely time-intensive, for two reasons. First, we may not have strong prior beliefs about a narrow range $[\underline{\theta}, \bar{\theta}]$, in which case we would have to compute L for many possible parameter values. Second, θ may be a vector of many parameters, in which case we would have to search over a complex, multi-dimensional grid. As a result of these limitations of options (1) and (2), we use option (3) for most maximum likelihood estimation. Numerical optimization works quite well if the likelihood function has only one optimum, as it does in most of the models we study in this course. Of course, if the likelihood function has many local maxima, then we run the risk of choosing the wrong maximum. In practice, this concern is not important for the applications we'll study.

LECTURE NOTE 6: BINARY DEPENDENT VARIABLES

1 Introduction

In this lecture note, we take a fresh look at the estimation of models with binary dependent variables. We will first discuss the linear probability model, exploring estimation details that you did not discuss in 507c. We will then move on to the estimation of probit and logit models by maximum likelihood.

2 Linear Probability Model

The simplest approach to estimate an equation with a binary dependent variable is the linear probability model. Suppose D_i takes on values 0 or 1. We write the model:

$$\begin{aligned} D_i &= \beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + \varepsilon_i \\ &= X_i'\beta + \varepsilon_i \end{aligned}$$

where the error term ε_i has expectation zero conditional on $X_{1i}, \cdots, X_{Ki}$. We interpret $\hat{D}_i = X_i'\beta = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki}$ as the expected probability that $D_i = 1$ given $X_{1i}, \cdots, X_{Ki}$. Thus, we interpret $\beta_k = \frac{\partial Pr[D_i=1]}{\partial X_{ki}}$ as the derivative of the expected probability with respect to X_{ki}. If X_{ki} is a binary variable, then $\beta_k = Pr[D_i = 1|X_{ki} = 1, X_{ji} = x_{ji} \forall j \neq k] - Pr[D_i = 1|X_{ki} = 0, X_{ji} = x_{ji} \forall j \neq k]$. (We hold all other X_{jt} constant at some value x_{jt}, although the value does not matter.) We can then use our coefficient estimates to generate a predicted probability: $\widehat{Pr}[D_i = 1|X_i] = X_i'\hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_K X_{Ki}$.

Two main problems plague the linear probability model. The first, probably better known, problem is that $\widehat{Pr}[D_i = 1]$ may lie outside the $[0, 1]$ interval. This issue results from the linearity of the functional form, and we cannot fix it while still estimating a linear probability model.

The second problem deals with heteroskedasticity, and we can fix it. To see the heteroskedasticity problem, note that given X_i, the error term has only two possible values:

$$\varepsilon_i = \begin{cases} 1 - X_i'\beta & \text{if } D_i = 1 \\ -X_i'\beta & \text{if } D_i = 0 \end{cases}$$

Now consider the conditional variance of the error term. Because the $E[\varepsilon_i|X_i] = 0$, we can write:

$$\begin{aligned} V[\varepsilon_i|X_i] &= E[\varepsilon_i^2|X_i] \\ &= X_i'\beta\left(1 - X_i'\beta\right)^2 + \left(1 - X_i'\beta\right)\left(-X_i'\beta\right)^2 \\ &= X_i'\beta\left(1 - X_i'\beta\right) \end{aligned}$$

(Try the math yourself.) This heteroskedasticity implies that the classical OLS estimator is not efficient and gives incorrect standard errors. As we discussed in Lecture Note 3, we have two options. First, we can estimate robust standard errors, which will not be efficient. Second, we can perform weighted least squares, using estimates of $\frac{1}{V[\varepsilon_i|X_i]}$ as our weights. To do so, we carry out a two-step procedure. In the first step, we estimate the model using OLS, and we use our vector of coefficients $\hat{\beta}^{OLS}$ to estimate $\hat{V}[\varepsilon_i|X_i] = X_i'\hat{\beta}^{OLS}\left(1 - X_i'\hat{\beta}^{OLS}\right)$. In the second step, we run weighted least squares, using $w_i = \frac{1}{\hat{V}[\varepsilon_i|X_i]}$ as our regression weights. In Stata, we would type: `reg D X1 X2 X3 [aw=w]`. Because we have adjusted for a known form of heteroskedasticity, we can assume homoskedastic errors in the weighted regression, and we will once again obtain a BLUE. A significant limitation of this approach is that we cannot assign observations negative weight, so any observations with predicted values outside the unit interval $[0, 1]$ will be dropped from the weighted regression. These dropped observations will be non-randomly selected, so their omission is likely to bias our estimates. As a result, the weighting procedure is useful only when few observations have predicted probabilities that lie outside $[0, 1]$.

3 Probit and Logit Models

Lecture Note 5 discussed maximum likelihood estimation of a Bernoulli random variable. If D_i is a binary variable that equals 1 with probability p and equals 0 with probability $1 - p$, then the likelihood is:

$$L = \prod_{i=1}^{N} p^{D_i}\left(1 - p\right)^{(1 - D_i)}$$

and the log-likelihood is:

$$\ln L = \sum_{i=1}^{N} \left\{D_i \ln(p) + \left(1 - D_i\right)\ln\left(1 - p\right)\right\}$$

In Lecture Note 5, we were okay with having a single p for the overall sample, Now, however, we are interested in allowing p_i to be individual-specific in a way that depends on X_i. To do so, let us define a function G and

parameters β such that $p_i = G(X_i^{'}\beta)$. Then the likelihood becomes:

$$L = \prod_{i=1}^{N} G(X_i'\beta)^{D_i} \left[1 - G(X_i'\beta)\right]^{(1-D_i)}$$

and the log-likelihood becomes:

$$\ln L = \sum_{i=1}^{N} \left\{D_i \ln \left[G(X_i'\beta)\right] + (1 - D_i) \ln \left[1 - G(X_i'\beta)\right]\right\}$$

Now we only need to choose an appropriate function G.

Because p_i (and therefore also $G(X_i'\beta)$) is a probability, we want the function G to satisfy a few properties. First, its range should be the interval $[0, 1]$. Second, as $X_i'\beta$ approaches infinity, $G(X_i'\beta)$ should approach 1. Third (and analogously), as $X_i'\beta$ approaches negative infinity, $G(X_i'\beta)$ should approach 0. Note that these properties are central features of cumulative distribution functions (CDFs). Indeed, the two most common choices for G are the CDFs for the normal distribution and the logistic distribution. We write $\Phi[X_i'\beta]$ for the standard normal distribution, and $\Lambda[X_i'\beta]$ for the logistic distribution. The method using the standard normal distribution is called *probit regression*, and the method using the logistic distribution is called *logit* or *logistic regression*. The next two subsections describe the likelihood functions for both methods. In neither case does the maximum likelihood problem have a closed-form solution, so statistical software uses numerical optimization methods to find the maximum of the likelihood function.

3.1 Probit Likelihood

The probit model uses the standard normal CDF:

$$Pr[D_i = 1|X_i] = \Phi[X_i'\beta] = \int_{-\infty}^{X_i'\beta} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(x'\beta\right)^2\right\} dx$$

Thus, the likelihood for the probit is:

$$L = \prod_{i=1}^{N} \Phi\left[X_i'\beta\right]^{D_i} \left[1 - \Phi[X_i'\beta]\right]^{(1-D_i)}$$

and the log-likelihood is:

$$\begin{aligned} \ln L &= \sum_{i=1}^{N} \left\{D_i \ln \left[\Phi[X_i'\beta]\right] + (1 - D_i) \ln \left[1 - \Phi[X_i'\beta]\right]\right\} \\ &= \sum_{i=1}^{N} \left\{D_i \ln \left[\Phi[X_i'\beta]\right] + (1 - D_i) \ln \left[\Phi[-X_i'\beta]\right]\right\} \end{aligned}$$

where the second line follows because the normal distribution is symmetric.

3.2 Logit Likelihood

The logit model uses the logistic CDF:

$$Pr[D_i = 1|X_i] = \Lambda[X_i^{'}\beta] = \frac{\exp\{X_i'\beta\}}{1+\exp\{X_i'\beta\}} = \frac{1}{1+\exp\{-X_i'\beta\}}$$

where $\exp\{X_i'\beta\}$ is alternative notation for $e^{X_i'\beta}$. The likelihood for the logit is:

$$L = \prod_{i=1}^{N} \Lambda\left[X_i'\beta\right]^{D_i} \left[1-\Lambda[X_i'\beta]\right]^{(1-D_i)}$$

and the log-likelihood is:

$$\begin{aligned} \ln L &= \sum_{i=1}^{N} \{D_i \ln\left[\Lambda[X_i'\beta]\right] + (1-D_i)\ln\left[1-\Lambda[X_i'\beta]\right]\} \\ &= \sum_{i=1}^{N} \{D_i \ln\left[\Lambda[X_i'\beta]\right] + (1-D_i)\ln\left[\Lambda[-X_i'\beta]\right]\} \end{aligned}$$

where the second line follows because the logistic distribution is symmetric.

4 Alternative Representations of the Binomial Discrete Choice Problem

Section 3 characterized logit and probit models using purely probabilistic representations: generalizations of the maximum likelihood problem for a Bernoulli random variable. But other approaches to the problem of modeling binary dependent variables exist. Two representations are especially popular in the social science literature. Because social scientists tend to be interested in human behavior, they describe the problem as the *binomial discrete choice problem*.

4.1 Latent Variables Representation

As one alternative way to conceptualize the problem, we can think of the binary dependent variable as a discretization of an underlying (unobserved) continuous random variable. We suppose that a continuous variable Y_i exists, but we cannot observe it. Y_i is determined as follows:

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

where ε_i has either a logistic or a normal distribution. However, we can only observe the binary variable D_i:

$$D_i = \begin{cases} 1 & \text{if } Y_i > 0 \\ 0 & \text{if } Y_i \leq 0 \end{cases}$$

In this setup, D_i is just a coarse version of the underlying latent variable Y_i.

We can derive the probability that $D_i = 1$:

$$\begin{aligned} Pr[D_i = 1|X_i] &= Pr[Y_i > 0|X_i] \\ &= Pr[X_i'\beta + \varepsilon_i > 0|X_i] \\ &= Pr[\varepsilon_i > -X_i'\beta|X_i] \\ &= 1 - F[-X_i'\beta] \\ &= F[X_i'\beta] \end{aligned}$$

where $F[\cdot]$ is either the standard normal CDF or the logistic CDF. The last step follows from the fact that both distributions have mean zero and are symmetric. This derivation shows that the latent variables representations are equivalent to the probit and logit models we originally specified.

4.2 Random Utility Representation

Yet another conceptualization of the problem involves a bit of economic theory. We consider an individual i choosing between two options $j = 1, 2$. The options could be occupations, cars, school-leaving ages, and so forth. (The next lecture note will discuss methods to analyze decisions with more than two options.) The individual's utility from option j is:

$$U_{ij} = X_{ij}'\beta_j + \varepsilon_{ij}$$

and the individual chooses option 2 over option 1 if $U_{i2} > U_{i1}$, or $X_{i2}'\beta_2 + \varepsilon_{i2} > X_{i1}'\beta_1 + \varepsilon_{i1}$, or $X_{i2}'\beta_2 - X_{i1}'\beta_1 > \varepsilon_{i1} - \varepsilon_{i2}$. We allow β_j to vary by option, but we do not require it to do so. For example, if we are modeling the decision to enroll in a state or private university, we might think of net tuition (tuition minus financial aid) as having the same (negative) slope coefficient for both options. But parental income may differentially affect the attractiveness of the two options, making the slope coefficient on parental income option-specific. In fact, since parental income is fixed for each individual, we *must* assume option-specific coefficients. To identify our model, we need either $X_{i2} - X_{i1}$ or $\beta_2 - \beta_1$ (or both) to be non-zero.

Given the decision rule above, the probability the the individual chooses option 2 is:

$$Pr[U_{i2} > U_{i1}|X_i] = Pr[\varepsilon_{i1} - \varepsilon_{i2} < X'_{i2}\beta_2 - X'_{i1}\beta_1]$$

We typically assume that the ε_{ij}'s are i.i.d. random variables with either a standard normal distribution or a type 1 extreme-value distribution (a distribution you don't need to know). As you already know, the difference of two normal random variables is itself normal, so the normality assumption leads us to the probit model. Just as conveniently, the difference of two random variables with type 1 extreme-value distributions has a logistic distribution, so the extreme-value assumption leads us to the logit model.

5 Interpreting the Results of Probit and Logit Models

Despite their limitations, linear probability have the attractive property that we can interpret their coefficients as changes in the absolute probability of the event that $D_i = 1$. In contrast, probit and logit results are more difficult to interpret. In those models, the coefficients depend on the variance we choose for the latent variable Y_i, which is arbitrary. As a result, the signs and relative sizes of the coefficients within a regression are meaningful, but one generally cannot compare the sizes of coefficients across regressions. Furthermore, even within a single regression, the changes in probability implied by the coefficients are not always obvious. Nonetheless, we can use the probit and logit coefficients to calculate more interpretable quantities. I outline several approaches below. In all cases, we will take interest in assessing the *magnitude* of the results. For significance tests, we can rely on the original probit or logit coefficients and standard errors.

An example will help illustrate the various approaches below. Suppose we are studying the choice of an MPA1 to take b-track or c-track econometrics, and imagine we have a Stata dataset consisting of three variables:

Name	**Definition**	**Mean**
`ctrack`	1 if c-track, 0 if b-track	0.50
`male`	1 if male, 0 if female	0.45
`foreign`	1 if foreign, 0 if from US	0.33
`gre`	math GRE score	700

5.1 Predicted Probabilities

One option is to predict $\widehat{Pr}[D_i = 1|X_{ki} = x_k, X_{ji} = \bar{X}_{ji} \forall j \neq k]$ for various values of x_k. This method holds all other X_{ji}'s constant at their sample means and then predicts the probability that $D_i = 1$ given that $X_{ki} = x_k$. The easiest way to take this approach in Stata involves the predict command. Suppose we wish

to generate predicted probabilities of taking the c-track for foreign and non-foreign students, after adjusting for gender and GRE scores. First, save your dataset: `save temp,replace`. Second, run `logit ctrack male foreign gre,robust` or `probit ctrack male foreign gre,robust`. Third, replace `male` and `gre` by their means: `replace male = 0.45 and replace gre = 700`. Fourth, predict the probabilities: `predict phat`. This procedure will generate a variable `phat` that takes on two values, one for `foreign` = 1 and one for `foreign` = 0. You can create a graph with these predicted values, or you can summarize them. After you finish this procedure, type `use temp,clear` to return to your original dataset with the correct values for `male` and `gre`.

5.2 Marginal Effects

A second option, by far the most common in the economics literature, is to compute what are called the "marginal effect" of each X_{ki}, holding the other X_{ji}'s constant at their means. The marginal effect corresponds to the derivative $\frac{\partial Pr[D_i=1]}{\partial X_{ki}}$, which is exactly the same estimand we obtain from the linear probability model. We write:

$$\frac{\partial Pr[D_i = 1]}{\partial X_{ki}} = \frac{\partial F[X_i'\beta]}{\partial X_{ki}} = F'[X_i'\beta]\beta_k = f(X_i'\beta)\beta_k$$

where $F[\cdot]$ and $f(\cdot)$ are the CDF and PDF for either the normal or the logistic distribution. You can see clearly here that the marginal effect of X_{ki} depends on the values of the other X_{ji}'s. This dependency is due to the non-linearity of the model; it does not arise in OLS. As already mentioned, we usually choose to set each X_{ji} equal to its mean. In Stata, one can compute marginal effects by typing `mfx compute` after running a probit or logit regression. Stata also has a command `dprobit` that automatically reports marginal effects rather than probit coefficients. Stata does not have a similar built-in command for logits, but you can download `dlogit2` from the internet if you wish. In practice, logits and probits almost always produce nearly identical marginal effects estimates.

5.3 Odds Ratios

Finally, the logit model has a convenient property for measuring the proportional effects of a change in X_{ki}. Recall that is an event has probability p, then the odds of the even are $\frac{p}{1-p}$. In the logit model, we have $p = \frac{\exp\{X_i'\beta\}}{1+\exp\{X_i'\beta\}}$, so that we can write the logarithm of the odds (or the "log-odds") as:

$$\ln(\text{odds}) = \ln\left(\frac{\frac{\exp\{X_i'\beta\}}{1+\exp\{X_i'\beta\}}}{1-\frac{\exp\{X_i'\beta\}}{1+\exp\{X_i'\beta\}}}\right) = \ln\left(\exp\{X_i'\beta\}\right) = X_i'\beta$$

The log-odds are linear in X_i in the logit model. (For this reason, the log-odds of an event are often called the logit of an event.) As a result, β_k measures the derivative of the log-odds with respect to X_{ki}. Note that the equation above implies that we can write the odds in the logit model as:

$$\text{odds} = e^{X_i'\beta} = e^{\beta_0+\beta_1 X_{1i}+\cdots+\beta_K X_{Ki}}$$

Consider the effect of changing X_{1i} from 0 to 1. Above, we represented this effect in terms of the *absolute* change in probability. As an alternative, we can calculate the *proportional* change in the odds. We do so using an *odds ratio*:

$$\begin{aligned}
\text{odds ratio} &= \frac{\frac{Pr[D_i=1|X_{1i}=1,X_{2i},\cdots,X_{Ki}]}{1-Pr[D_i=1|X_{1i}=1,X_{2i},\cdots,X_{Ki}]}}{\frac{Pr[D_i=1|X_{1i}=0,X_{2i},\cdots,X_{Ki}]}{1-Pr[D_i=1|X_{1i}=0,X_{2i},\cdots,X_{Ki}]}} \\
&= \frac{\exp\{\beta_0+\beta_1\cdot 1+\beta_2 X_{2i}+\cdots+\beta_K X_{Ki}\}}{\exp\{\beta_0+\beta_1\cdot 0+\beta_2 X_{2i}+\cdots+\beta_K X_{Ki}\}} \\
&= \frac{\exp\{\beta_0\}}{\exp\{\beta_0\}}\cdot\frac{\exp\{\beta_1\cdot 1\}}{\exp\{\beta_1\cdot 0\}}\cdot\frac{\exp\{\beta_2 X_{2i}+\cdots+\beta_K X_{Ki}\}}{\exp\{\beta_2 X_{2i}+\cdots+\beta_K X_{Ki}\}} \\
&= e^{\beta_1}
\end{aligned}$$

So the exponentiated logit coefficient is the odds ratio! Consequently, many researchers report the exponentiated coefficient rather than the actual coefficient or the marginal effect. This representation is especially popular outside of economics, perhaps because economists like thinking about marginal changes, but researchers in many other fields like thinking about discrete changes. For instance, when epidemiologists want to know whether exposure to a toxin is associated with birth defects, they will often run logit regressions and report the exponentiated coefficient on the dummy for toxin exposure. That exponentiated coefficient represents the proportional change in the odds of a birth defect that is associated with exposure to a toxin. If the odds ratio is below 1, then toxins are negatively associated with the odds of a birth defect; if the odds ratio is above 1, the association is positive. Notably, the proportional change in the odds is *not* the same as the proportional change in the probability (i.e., the risk of a birth defect). The proportional change in the probability, $\frac{Pr[D_i=1|X_{1i}=1,X_{2i},\cdots,X_{Ki}]}{Pr[D_i=1|X_{1i}=0,X_{2i},\cdots,X_{Ki}]}$, is called the *relative risk*. The relative risk and the odds ratio always fall on the same side of 1, but the odds ratio is more extreme than the relative risk. That is to say, $|OR-1| \geq |RR-1|$.

In Stata, you can obtain the odds ratio in two ways. First, you can run the original logit command with the odds ratio option: `logit ctrack male foreign gre,robust or`. Second, you can use the `logistic` command, which automatically reports odds ratios: `logistic ctrack male foreign gre,robust`.

LECTURE NOTE 7: OTHER LIMITED DEPENDENT VARIABLES

1 Introduction

In Lecture Note 6, we considered estimation techniques to deal with dependent variables that take on only two values. Binary outcome variables are part of a class of dependent variables called *limited dependent variables*. This lecture note gives a tour of estimation techniques dealing with a range of other (non-binary) limited dependent variables. The lecture merely scratches the surface of a huge body of statistical knowledge. Students interested in learning more about these methods should take WWS 509 next fall.

We will begin with two generalizations of the logit and probit models we studied in Lecture Note 6. These generalizations involve *multinomial outcomes*: that is, categorical dependent variables that take on more than two values. In one case, the categories can be ordered (or ranked) in some meaningful way; in the other case, no clear ordering exists. After studying these models, we will briefly explore one type of dependent variable that involves a mixture of categorical and continuous random variables.

2 Ordered Multinomial Dependent Variables

Consider a random variable D_i that can take on values $k = 1, 2, \cdots, K$. Suppose that a natural ordering of the values of D_i exists, so we can write:

$$(D_i = 1) < (D_i = 2) < \cdots < (D_i = K)$$

For instance, D_i could be health status, where 1 represents poor health and 5 represents excellent health. This setup is called the *ordered multinomial discrete choice* problem.

2.1 Ordered Linear Probability Model

One estimation option, as always, is OLS: $D_i = X_i'\beta + \varepsilon_i$. Such a regression is called an *ordered linear probability model*. The predicted values, $\hat{D}_i = X_i'\hat{\beta}$, can be interpreted as the expected value of D_i given the independent variables. However, as with the binary linear probability model, several problems arise. Two problems are analogous: (1) the model may predict values $\hat{D}_i$ that lie outside the range of D_i, and (2) the

error term can only take on k values:

$$\varepsilon_i = k - X_i'\beta \qquad \text{if } D_i = k$$

(In the linear probability model, $k = 2$, so the error term can take on two values.) A third problem is new: the ordered linear probability model assumes that the units of D_i have a cardinal interpretation when we have no good reason to presume that they do.

2.2 Ordered Probit Model

An alternative to this approach is the *ordered probit model*. To motivate the ordered probit model, we return to the latent variable model we developed in Lecture Note 6. Let $Y_i = X_i'\beta + \varepsilon_i$ be a continuous, unobserved random variable, and let D_i be determined as follows:

$$\begin{aligned} D_i &= 1 \qquad \text{iff } Y_i < \lambda_1 \\ D_i &= k \qquad \text{iff } \lambda_{k-1} \leq Y_i < \lambda_k \text{ for } k \in [2, K-1] \\ D_i &= K \qquad \text{iff } Y_i \geq \lambda_{K-1} \end{aligned}$$

where $\lambda_1, \cdots, \lambda_{K-1}$ are a series of unobserved thresholds. As Y_i crosses each threshold, D_i moves to the next-highest category.

As in the binomial probit case, we assume that ε_i is drawn from a standard normal distribution. Then:

$$\begin{aligned} Pr[D_i &= 1|X_i] = Pr[Y_i < \lambda_1|X_i] = \Phi[\lambda_1 - X_i'\beta] \\ Pr[D_i &= k|X_i] = Pr[\lambda_{k-1} \leq Y_i < \lambda_k|X_i] = \Phi[\lambda_k - X_i'\beta] - \Phi[\lambda_{k-1} - X_i'\beta] \qquad \text{for } k \in [2, K-1] \\ Pr[D_i &= K|X_i] = Pr[Y_i \geq \lambda_K|X_i] = 1 - \Phi[\lambda_K - X_i'\beta] \end{aligned}$$

where $\Phi[\cdot]$ is the standard normal CDF. Given these probabilities, we can write down a log-likelihood function and optimize to find the values $\hat{\beta}, \hat{\lambda}_1, \cdots, \hat{\lambda}_K$ that maximize the likelihood. You can estimate an ordered probit model in Stata using the `oprobit` command. Note that the $K = 2$ case of the ordered probit is exactly the binomial probit. As in the case of the binomial probit, the signs and relative magnitudes of the coefficients are meaningful, but the implication for changes in probabilities is not obvious. As such, predicted probabilities will often aid interpretation.

2.3 Interval Regression

A special case of the ordered probit model arises when we have a continuous dependent variable that has been intervalled. For example, in surveys, respondents often report their incomes by choosing from several per-determined income ranges ($0-$19,999, $20,000-$30,000, etc.), rather than disclosing the actual number. If we suppose that the outcome Y_i is determined by the normal linear model ($Y_i = X_i'\beta + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$), then we have an ordered probit model with known cutoff points $\lambda_1, \cdots, \lambda_{K-1}$. In this case:

$$\begin{aligned}
Pr[Y_i \text{ in interval } 1|X_i] &= Pr[Y_i < \lambda_1|X_i] = \Phi\left[\frac{\lambda_1 - X_i'\beta}{\sigma}\right] \\
Pr[Y_i \text{ in interval } k|X_i] &= Pr[\lambda_{k-1} \leq Y_i < \lambda_k|X_i] = \Phi\left[\frac{\lambda_k - X_i'\beta}{\sigma}\right] - \Phi\left[\frac{\lambda_{k-1} - X_i'\beta}{\sigma}\right] \qquad \text{for } k \in [2, K-1] \\
Pr[Y_i \text{ in interval } K|X_i] &= Pr[Y_i \geq \lambda_K|X_i] = \Phi\left[\frac{\lambda_K - X_i'\beta}{\sigma}\right]
\end{aligned}$$

As in the case with unknown cutoff points, we can use these probabilities to write down a log-likelihood function and optimize to find the values $\hat{\beta}, \hat{\sigma}^2$ that maximize the likelihood. You can estimate an interval regression in Stata using the `intreg` command.

3 Unordered Multinomial Dependent Variables

We now move to the case of an *unordered* categorical dependent variable. For example, consider the choice of a Princeton affiliate to eat dinner at Hoagie Haven (option 1), Triumph (option 2), or Mediterra (option 3). Unlike the health status example in the previous section, there is no natural ordering of these options. As with the other limited dependent variables we have studied, we can conceptualize this estimation problem in several ways, but the random utility framework provides probably the clearest representation. We will call this representation the *unordered multinomial discrete choice problem.* We assume that each individual i obtains utility U_{ij} option from option $j = 1, \cdots, J$. U_{ij} is determined as follows:

$$\begin{aligned}
U_{ij} &= X_{ij}'\beta_j + \varepsilon_{ij} \\
&= W_i'\alpha_j + Z_{ij}'\gamma + \varepsilon_{ij}
\end{aligned}$$

We decompose X_{ij} into two subsets of variables, one of which is individual-specific (W_i) and one of which is option-specific (Z_{ij}). In the restaurant example, $J = 3$. Some examples of components of W_i are (1) whether the individual is an undergrad, grad student, staff member, or faculty; (2) whether the individual is on a diet; (3) the individual's income. Some examples of components of Z_{ij} are (1) the distance between the individual's home and the restaurant; (2) the price of the meal; (3) the cleanliness of the restaurant.

As in the binomial discrete choice case, the individual chooses the option that gives her the highest utility. In other words, for any two options j and k, the individual chooses option j if and only if $U_{ij} - U_{ik} \geq 0$. To find the maximum, we consider $J-1$ utility differences. In the restaurant example, we have:

$$\begin{aligned} S_{i2} &= U_{i2} - U_{i1} = W_i'(\alpha_2 - \alpha_1) + (Z_{i2} - Z_{i1})'\gamma + (\varepsilon_{i2} - \varepsilon_{i1}) \\ S_{i3} &= U_{i3} - U_{i1} = W_i'(\alpha_3 - \alpha_1) + (Z_{i3} - Z_{i1})'\gamma + (\varepsilon_{i3} - \varepsilon_{i1}) \end{aligned}$$

Both these utility differences take option 1 (Hoagie Haven) as the base category. The individual's decision rule is:

$$\begin{aligned} \text{Choose 1 if } \max(0, S_{i2}, S_{i3}) &= 0 \\ \text{Choose 2 if } \max(0, S_{i2}, S_{i3}) &= S_{i2} \\ \text{Choose 3 if } \max(0, S_{i2}, S_{i3}) &= S_{i3} \end{aligned}$$

3.1 Conditional and Multinomial Logit Models

In this random utility setup, we assume that the error terms (ε_{ij}) have independent extreme value distributions. As you'll recall from Lecture Note 6, this assumption guarantees that the difference of the error terms has a logistic distribution. With more than one option, however, the independence assumption has an additional implication: *independence of irrelevant alternatives* (IIA). The IIA property implies that the probability of choosing option 2 over option 1 depends only on the properties of options 1 and 2; the removal or addition of option 3 is irrelevant to the choice between 1 and 2. This property is reasonable in some situations, but not in others. For example, in the restaurant example, if Hoagie Haven closed down, most Hoagie Haven enthusiasts would probably go to Triumph, not Mediterra. So removing the Hoagie Haven option might alter their probability of choosing Mediterra over Triumph. The IIA property can be a bit difficult to understand, but a rough idea of the concept is useful.

Because $(\varepsilon_{ij} - \varepsilon_{i1})$ has a logistic distribution (for option $j \neq 1$), the probability of choosing option j is:

$$\begin{aligned} Pr[D_i = j | W_i, Z_{ij}] &= \frac{\exp\left(W_i'\alpha_j + Z_{ij}'\gamma\right)}{\sum_{k=1}^{K} \exp\left(W_i'\alpha_k + Z_{ik}'\gamma\right)} \\ &= \frac{\exp\left(W_i'(\alpha_j - \alpha_1) + (Z_{ij} - Z_{i1})'\gamma\right)}{1 + \sum_{k=2}^{K} \exp\left(W_i'(\alpha_k - \alpha_1) + (Z_{ik} - Z_{i1})'\gamma\right)} \\ &= \frac{\exp\left(W_i'\delta_j + (Z_{ij} - Z_{i1})'\gamma\right)}{1 + \sum_{k=2}^{K} \exp\left(W_i'\delta_k + (Z_{ik} - Z_{i1})'\gamma\right)} \end{aligned}$$

where the last line defines $\delta_j = \alpha_j - \alpha_1$ to make clear that we are only measuring differences relative to option

1. We can write down a likelihood function analogous to that of the binomial logit model, and we can then implement a similar estimation procedure. When the model only has individual-specific covariates (W_i), we call it the *multinomial logit model*; when the model only has option-specific covariates (Z_{ij}), we call it the *conditional logit model*; and when it has both, we call it the *mixed conditional logit model*. In Stata, you can estimate a conditional logit using the `clogit` command, and you can estimate a multinomial logit using the `mlogit` command. In the problem set, I only ask you to run a multinomial logit, as this method is probably relevant for more policy applications.

The coefficients from these models have the same problem as those from the binomial logit model; they are difficult to interpret. Fortunately, we can again obtain more interpretable quantities by exponentiating them. To see this result, consider the probabilities of choosing options j and m:

$$\begin{aligned} Pr[D_i = j|W_i, Z_{ij}] &= \frac{\exp\left(W_i'\delta_j + (Z_{ij} - Z_{i1})'\gamma\right)}{1 + \sum_{k=2}^{K} \exp\left(W_i'\delta_k + (Z_{ik} - Z_{i1})'\gamma\right)} \\ Pr[D_i = m|W_i, Z_{ij}] &= \frac{\exp\left(W_i'\delta_m + (Z_{im} - Z_{i1})'\gamma\right)}{1 + \sum_{k=2}^{K} \exp\left(W_i'\delta_k + (Z_{ik} - Z_{i1})'\gamma\right)} \end{aligned}$$

The ratio of these probabilities is:

$$\begin{aligned} \frac{Pr[D_i = j|W_i, Z_{ij}]}{Pr[D_i = m|W_i, Z_{ij}]} &= \frac{\exp\left(W_i'\delta_j + (Z_{ij} - Z_{i1})'\gamma\right)}{\exp\left(W_i'\delta_m + (Z_{im} - Z_{i1})'\gamma\right)} \\ &= \exp\left(W_i'(\delta_j - \delta_m) + (Z_{ij} - Z_{im})'\gamma\right) \end{aligned}$$

In the multinomial logit model, we consider only W_i, and by construction $\delta_1 = \alpha_1 - \alpha_1 = 0$, so that

$$\begin{aligned} \frac{Pr[D_i = j|W_i, Z_{ij}]}{Pr[D_i = 1|W_i, Z_{ij}]} &= \exp\left(W_i'\delta_j\right) \\ &= e^{W_{1i}\delta_{j1} + W_{2i}\delta_{j2} + \cdots} \end{aligned}$$

As a result, a one-unit change in W_{1i} affects the relative probability of choosing option j over option 1 by $e^{\delta_{j1}}$. When D_i is binary, the ratio of the probabilities is the odds, so $e^{\delta_{j1}}$ is the odds ratio. In the multinomial case, because $e^{\delta_{j1}}$ measures the change in the relative risk of choosing option j over option 1, $e^{\delta_{j1}}$ is often called the *relative risk ratio*. You can obtain this quantity in Stata by implementing the `rrr` option after the `mlogit` command. As in the ordered probit case, however, you will often find it illuminating to predict probabilities and graph them.

4 Censored Dependent Variables

Suppose we have a continuous dependent variable Y_i that is determined as follows:

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

where $\varepsilon_i \sim \mathcal{N}(o, \sigma^2)$. However, we only *observe* the true value of Y_i when it is less than some threshold $\overline{Y}$; when $Y_i \geq \overline{Y}$, we only observe $\bar{Y}$. This scenario is called *right censoring*. One common example of right censoring is the top-coding of income in surveys, where individuals report their actual incomes up to some threshold, after which they simply check off "over $100,000." Another common example arises in the modeling of durations, such as the length of an unemployment spell, the age at marriage, or the age at death. Take the example of age at death. If we observe a living person at age 35, we know only that the person's age at death is greater than 35.[1]

Left censoring, in which we we only observe the true value of Y_i when it is greater than some threshold $\underline{Y}$, also occurs. For example, suppose we are modeling an individual's labor supply. When the individual's optimal labor supply is negative, we only observe the corner solution where labor supply equals zero. In this case, Y_i would be optimal hours worked, and $\underline{Y} = 0$.

Finally, *double censoring*, in which we we only observe the true value of Y_i when it lands in the range $\left[\underline{Y}, \overline{Y}\right]$, can arise as well. One instance is a badly written exam that results in a large number of scores of 0 or 100.

Because we have assumed that ε_i is normally distributed, all three types of censoring present a mixture of the normal linear model and the probit model. The *censored normal regression model* (or *Tobit model*, after its discoverer James Tobin) solves the maximum likelihood problem for this mixture of continuous and discrete distributions. Under the assumptions that $\varepsilon_i \sim \mathcal{N}(o, \sigma^2)$, the model provides consistent estimates of β, which cannot be obtained through OLS. However, the assumptions of normality and heteroskedasticity turn out to be quite important; the results can be quite misleading when these assumptions do not hold. The Stata command `tobit` estimates the model when the censoring thresholds are fixed across observations. The Stata command `cnreg` carries out the same basic estimation procedure but allows the censoring thresholds to differ across observations.

[1] For the analysis of durations, researchers typically use a class of models called *survival analysis models*. We will not cover these models in WWS 508c, but WWS 509 covers them in depth.

LECTURE NOTE 8: TIME SERIES

1 Introduction

Up to this point in the course, we have focused on collections of independent and identically distributed units with subscript i. We refer to such data as *cross-sectional data*. In this lecture, we study an important counterpart to cross-sectional data, *time series*. Instead of studying a large number of units (or individuals), we study a single unit over time. Modern time series analysis involves quite complicated math and (in my opinion) has less capability than cross-sectional analysis to shed light on causal relationships. But time series analyses are useful in the study of some macroeconomic causal issues, and they are also quite useful for forecasting. (Forecasting does not require causality. For instance, we could look outside and count the number of umbrellas to forecast the probability of rain, even if umbrellas do not cause rain.)

Our starting point for time series will be similar to the cross-sectional model we studied:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_K X_{Kt} + U_t$$

with $t = 1, \cdots, T$. The only difference is that we have replaced i's with t's. But this difference has an important implication: observations now come in a sequence, so that t is closer to $t+1$ than t is to $t+100$. If we gather more data (i.e., expand T), we do not get more of the same. Furthermore, the sequencing of the observations makes the error terms unlikely to be independent. Unobservable variables are probably correlated over time, so $corr(U_t, U_{t-1}) \neq 0$. At the same time, we might be willing to assume that $corr(U_t, U_{t-100}) \approx 0$. Finally, we might now take interest in modeling today's outcome as a function of yesterday's outcome: $Y_t = \beta_0 + \beta_1 Y_{t-1} + U_t$. All of these features make time series analysis quite a different creature from cross-sectional analysis.

2 Serial Correlation

We first address *serial correlation* (or *autocorrelation*) in the error terms: the property that $corr(U_t, U_{t-1}) \neq 0$. To analyze this problem, we must specify the form of serial correlation in the error terms. A common

approach is to assume that U_t is *autoregressive of order 1*, or AR(1):

$$U_t = \rho U_{t-1} + \nu_t$$

where ν_t is an random variable that is i.i.d. across time periods (so $corr(\nu_s, \nu_t) = 0$ and $V[\nu_s] = V[\nu_t] = \sigma_\nu^2$ for all s, t). The parameter ρ measures the dependence of U_t on its value one period ago. Note that if U_t is AR(1), then:

$$U_t = \rho^2 U_{t-2} + \rho\nu_{t-1} + \nu_t = \rho^3 U_{t-3} + \rho^2\nu_{t-2} + \rho\nu_{t-1} + \nu_t = \cdots$$

Although we have specified U_t only as a function of U_{t-1}, U_t in fact depends on all past values of ν_t.

The dependence of U_t on all past values of ν_t complicates variance calculations:

$$\begin{aligned} V[U_t] &= \rho^2 V[U_{t-1}] + \sigma_\nu^2 \\ &= \rho^4 V[U_{t-2}] + (1+\rho)\sigma_\nu^2 \\ &\quad \rho^6 V[U_{t-3}] + (1+\rho+\rho^2)\sigma_\nu^2 \\ &\vdots \\ &= \rho^{2k} V[U_{t-k}] + (1+\rho+\rho^2+\cdots+\rho^{2k-2})\sigma_\nu^2 \end{aligned}$$

This variance formula leads to the concern that the variance of U_t changes over time, or that U_t is a *non-stationary process*, meaning that its distribution changes over time. For example, if $\rho = 1$, then:

$$V[U_t] = V[U_{t-k}] + k\sigma_\nu^2 > V[U_{t-k}]$$

for all $K > 0$. As a result, the variance of the error term is growing over time.

If U_t *is* stationary, meaning that its distribution does not change over time, then $V[U_t] = V[U_{t-1}]$. Therefore:

$$V[U_t] = \rho^2 V[U_{t-1}] + \sigma_\nu^2 = \frac{\sigma_\nu^2}{1-\rho^2}$$

Note that if $|\rho| \geq 1$, this expression is negative, which violates a basic property of the variance. So a process can be stationary only if $|\rho| < 1$. We can also derive the correlation between two error terms in a stationary error process. The covariance between U_t and U_{t-k} is:

$$cov(U_t, U_{t-k}) = cov\left(\rho^k U_{t-k} + \sum_{s=0}^{k-1} \rho^s \nu_{t-s}, U_{t-k}\right) = \rho^k V[U_{t-k}]$$

Because U_t is stationary, $V[U_t] = V[U_{t-k}]$. Thus:

$$corr(U_t, U_{t-k}) = \frac{cov(U_t, U_{t-k})}{\sqrt{V[U_t]V[U_{t-k}]}} = \frac{\rho^k V[U_{t-k}]}{V[U_{t-k}]} = \rho^k$$

We can see that no two errors are independent of one another. But because a stationary process has $|\rho| < 1$, $\lim_{k\to\infty} corr(U_t, U_{t-k}) = 0$.

3 Estimating Standard Errors with Serial Correlation

This section gives an overview of two strategies for estimating standard errors in the presence of serial correlation. Throughout this discussion, we assume that U_t is stationary.

3.1 HAC Standard Errors

For simplicity, suppose we are analyzing a univariate time series model:

$$Y_t = \beta_0 + \beta_1 X_t + U_t$$

The method we develop below also works for the multivariate case, but the math is more complicated. By focusing on the univariate case, we can refer to the following result from Lecture Note 3:

$$\hat{\beta}_1^{OLS} = \beta_1 + \frac{1}{\sum_t (X_t - \bar{X})^2} \sum_t \left(X_t - \bar{X}\right) U_t$$

The variance of this estimator is:

$$V\left[\hat{\beta}_1^{OLS}\right] = \frac{1}{\sum_t (X_t - \bar{X})^2} \sum_t \sum_s \left(X_t - \bar{X}\right)\left(X_s - \bar{X}\right) cov(U_t, U_s)$$

(If U_t follows a stationary AR(1) process, then $cov(U_t, U_s) = \rho^{t-s} V[U_t]$.) We can estimate each $cov(U_t, U_s)$ as we did in the cross-sectional setting for heteroskedasticity-robust and cluster-robust standard errors: $\hat{\sigma}_{ts} = \hat{U}_t \hat{U}_s$. In the cross-sectional setting, we had natural reasons to assume that some of these covariances were zero (think of the off-diagonal terms of the variance-covariance matrix). In the time series setting, no obvious analogue exists; indeed, we just saw that in the AR(1) case, $cov(U_t, U_s)$ is *never* zero. Nonetheless, a commonly-used method assumes that the covariances satisfy:

$$cov(U_t, U_s) = \begin{cases} \sigma_{ts} & \text{if } |t-s| < \Delta \\ 0 & \text{if } |t-s| \geq \Delta \end{cases}$$

where Δ is a number of time periods. In words, we are assuming that error terms more than Δ time periods apart are uncorrelated. This assumption is arbitrary and inconsistent with most error processes (including AR(1)). However, one could argue that because $\lim_{t-s\to\infty} corr(U_t, U_s) = 0$ when U_t is stationary, the model may provide a reasonable approximation of reality if we set Δ to be large. So we can estimate the variance of the OLS coefficient using $\hat{\sigma}_{ts} = \hat{U}_t\hat{U}_s$ as an estimator of σ_{ts} when $|t-s| < \Delta$, and assuming that the covariance is zero otherwise. The resulting standard errors are called *heteroskedasticity- and autocorrelation-robust standard errors,* or *HAC standard errors*, or *Newey-West standard errors.* You can implement them in Stata by typing `newey y x,lag(1)`, where the number inside `lag()` indicates the number of lagged error terms in the error process. In the AR(1) case, that number is one.

3.2 Quasi-Differencing

A second method provides an improvement over the HAC standard errors if we can confidently assume a model with AR(1) errors:

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 X_t + U_t \\ U_t &= \rho U_{t-1} + \nu_t \end{aligned}$$

Again, the method we develop in this section can be used with more than one regressor, but we focus on one for simplicity. In this case, we take *quasi-differences* of the regression equation:

$$Y_t - \rho Y_{t-1} = (1-\rho)\beta_0 + \beta_1(X_t - \rho X_{t-1}) + (U_t - \rho U_{t-1})$$

It turns out that this quasi-differenced error term is homoskedastic. Thus, if we can estimate ρ, we can obtain efficient estimates of β_1 using the quasi-differenced model.

We follow three steps:

1. Estimate $Y_t = \beta_0 + \beta_1 X_t + U_t$ by OLS.
2. Using the estimated residuals, calculate $\hat{\rho} = \frac{1}{T-1}\sum_{t=2}^{T} \hat{U}_t\hat{U}_{t-1}$.
3. Estimate the quasi-differenced regression of $Y_t - \hat{\rho}Y_{t-1}$ on $X_t - \hat{\rho}X_{t-1}$.

This procedure is known as the *Cochrane–Orcutt* procedure, and you can implement it in Stata by typing `prais y x,corc twostep`. If you do not type `twostep`, then Stata will loop through steps (2) and (3) until it converges. If you do not type `corc`, then Stata will carry out the *Prais-Winsten* procedure, which is identical to the Cochrane–Orcutt procedure, except it uses the first observation. (Cochrane–Orcutt discards the first observation to take quasi-differences.) Either procedure works fine, with or without the `twostep` option.

4 Trends

The Stock and Watson textbook emphasizes two sources of non-stationarity: trends and breaks. A break is a one-time change in the population regression function. We will not explore breaks here, but Section 14.7 of Stock and Watson has a nice discussion. Various types of trends can be represented by the following specification:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 t + U_t$$

where U_t is i.i.d. over time. This setup embeds many different types of time series variation:

- *Deterministic trend*: β_0 =any value, $\beta_1 = 0$, $\beta_2 \neq 0$. The mean of Y_t is changing in a nonrandom way over time, such that $E[Y_t|Y_{t-1}, Y_{t-2}, \cdots] = \beta_0 + \beta_2 t$. No other aspects of the distribution of Y_t are changing.
- *Random walk*: $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = 0$. Conditional on all past values $Y_{t-1}, Y_{t-2}, \cdots$, the expectation of Y_t is its previous value, Y_{t-1}. That is to say, $E[Y_t|Y_{t-1}, Y_{t-2}, \cdots] = Y_{t-1}$. Furthermore, as discussed in Section 2, the variance of Y_t is also growing: $V[Y_t] = t\sigma_U^2$.
- *Random walk with a drift*: $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 \neq 0$. This process combines the previous two. We have $E[Y_t|Y_{t-1}, Y_{t-2}, \cdots] = Y_{t-1} + \beta_2$, and $V[Y_t] = t\sigma_U^2$.

The last two processes are also called *stochastic trends*, representing the fact that the expectation of Y_t is changing over time. If $\beta_2 = 0$ and $|\beta_1| < 1$, then Y_t is stationary and does not contain a trend.

The prescriptions for dealing with trends vary with the model of interest. A very common case occurs in the model $Y_t = \beta_0 + \beta_1 X_t + U_t$, if we are concerned that both Y_t and X_t have (either deterministic or random) trends. These trends can generate very misleading results if we estimate the model by OLS. For example, both GDP and government spending have been increasing over time for a host of reasons, so the positive relationship between them is not meaningful. We can deal with these trends in two ways. First, we can include time as an independent variable, $Y_t = \beta_0 + \beta_1 X_t + \gamma t + U_t$. This approach essentially treats t as an omitted variable that is correlated with both Y_t and X_t. Second, we can first-difference the equation: $Y_t - Y_{t-1} = \beta_1(X_t - X_{t-1}) + \gamma + (U_t - U_{t-1})$. As an example of this second approach, we might regress GDP growth on the growth in government spending, rather than regressing levels on levels.

Now suppose we want to estimate an *autoregression*: $Y_t = \beta_0 + \beta_1 Y_{t-1} + U_t$. If $|\beta_1| < 1$, then Y_t is stationary, and we have no problem. However, if $\beta_1 \geq 1$, we get non-stationarity, which gives rise to estimation problems. When $\beta_1 = 1$, we say that Y_t has a *unit root*. In this case, we can take first differences on the left hand side to restore stationarity: $\Delta Y_t = Y_t - Y_{t-1} = \beta_0 + (\beta_1 - 1)Y_{t-1} + U_t$. To look for a unit root in this regression, we test whether the coefficient on Y_{t-1} equals zero. This test is called the Dickey-Fuller test, and you can implement it in Stata using the `dfuller` command.

5 Forecasting

A natural use of time series analysis is forecasting: given what we know at present, what is our expectation of a future value of Y_t? This problem is distinct from the prediction problem we encountered in the cross-sectional setting. Our predictions were entirely within-sample, whereas our forecasts for the future are necessarily out-of-sample.

Consider the model $Y_t = \beta_0 + \beta_1 X_{t-1} + U_t$, where U_t is stationary, and X_{t-1} is either a separate lagged variable or a lagged value of Y_t. Our forecast for Y_{t+1} given Y_t is $\hat{Y}_{t+1|t} = \hat{\beta}_0 + \hat{\beta}_1 X_t$. The forecast error is:

$$\hat{Y}_{t+1|t} - Y_{t+1} = U_{t+1} + (\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1) X_t$$

which has the variance:

$$V[\hat{Y}_{t+1|t} - Y_{t+1}] = \sigma_U^2 + V[\hat{\beta}_0 + \hat{\beta}_1 X_t]$$

The square root of this quantity is called the *mean root squared forecast error*. It is equal to the variance of the error term plus the variance of the prediction.

6 Lags and Impulse Response Functions

As already suggested, we often want to control for lagged values of Y_t and X_t. We say that the n^{th} lag of Y_t is Y_{t-n}, and similarly for X_t. A regression that controls for lagged values of Y_t is called an *autoregression*, while a regression that controls for lagged values of X_t is called a *distributed lag model*. When the two are combined, we have an *autoregressive distributed lag model*:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \cdots + \beta_j X_{t-j} + \gamma_1 Y_{t-1} + \cdots + \gamma_k Y_{t-k} + U_t$$

If we are able to consistently estimate the parameters of this model, it is interesting to work through the effects of increasing X_t by one unit in period t. This "impulse" will set off a sequence of changes:

- Y_t increases by β_0.
- Y_{t+1} increases by $\beta_1 + \gamma_1 \beta_0$.
- Y_{t+2} increases by $\beta_2 + \gamma_1(\beta_1 + \gamma_1 \beta_0) + \gamma_2 \beta_0$.

And so on. The change in X_t will have direct effects on Y until period $t + j$ and will have indirect effects on Y forever.

7 Additional Time Series Topics

This lecture note represents a tiny slice of time series econometrics. Two additional topics that are worthy of note are *vector autoregression (VAR)* and the *autoregressive general heteroskedasticity model (ARCH)*. These methods are most frequently used for forecasting, for instance by central banks and other financial institutions. A vector autoregression of order p (written VAR(p)) is a set of k time series regressions for the variables $Y_{1t}, Y_{2t}, \cdots, Y_{kt}$. Each variable is regressed on a set of p of its own lags and p lags of all the other variables. An ARCH model of order p (written ARCH(p)) allows the variance of the error term U_t to depend on p past values of the error term $U_{t-1}, U_{t-2}, \cdots, U_{t-p}$. This type of model can be useful for modeling the clustering of stock market volatility, for example. Chapter 16 of Stock and Watson reviews these models in greater (but still not much) detail.

LECTURE NOTE 9: PANEL DATA

1 Introduction

To this point in the course, we have studied two dimensions of data: cross-sections of many units and time series of a single unit over time. In this lecture note, we combine the two sources of variation into *panel data*. The key ingredient to panel data is that observations are grouped in some important way. Often, we will follow a group of units (individuals, states, countries, etc.) over time, in which case we will write Y_{it} to denote the outcome of unit i in time period t. Such data are also called *longitudinal data*. In other instances, we will study groups of observations that have no ordering we consider relevant, for example siblings, classmates, and residents of the same neighborhood. One can think of possible orderings of these groups (birth order, class rank, etc.), but in many applications, it will be reasonable to ignore them. To distinguish the unordered groups from the ordered groups, we will write Y_{ij} to denote the outcome of (unordered) observation j from group i. The methods below apply to both types of panel data, but we will focus first on the unordered case because the algebra is somewhat simpler.

To clarify the discussion, we will frequently refer to two examples. For the unordered case, we will consider the determinants of earnings in a sample of sibling groups (which lay people also call "families"). For the ordered case, we will consider the relationship between traffic regulations (speed limits, seat belt laws, etc.) and motor vehicle fatalities in a sample of states over time.

2 Error Components Models

We begin by studying a large number of groups i, each consisting of a collection of observations j. In the siblings example introduced above, i indexes families, and j indexes siblings within each family. Our basic regression model is:

$$Y_{ij} = \alpha + X_{ij}'\beta + Z_i'\gamma + U_{ij} \quad i = 1, \cdots, N \quad j = 1, \cdots, J \tag{1}$$

Y_{ij} is the outcome for the j^{th} observation from group i (e.g., log earnings in adulthood), X_{ij} is a vector of observation-specific characteristics (e.g., educational attainment), and Z_i is a vector of group-wide characteristics (e.g., parental education and wealth). Note that we have assumed a constant number of observations per group (siblings per family), J. If a panel satisfies this assumption, we call it a *balanced panel*; if it does not

(so that instead of J we have J_i), we call it an *unbalanced panel*. One can apply the methods in this lecture to both balanced and unbalanced panels. However, in a balanced panel, the overall number of observations is NJ, which is much more convenient than the corresponding number for an unbalanced panel. To simplify the algebra, we will assume a balanced panel in most of our discussion.

We might be tempted to run regression (1) using OLS, but the grouping of observations makes standard OLS inappropriate. To see the problem, it is useful to rewrite the residual U_{ij} as the sum of two components:

$$U_{ij} = \delta_i + \varepsilon_{ij} \tag{2}$$

where δ_i is the part of the residual that is the same for all observations within a group, and ε_{ij} is the part of the residual that is observation-specific. We assume that δ_i and ε_{ij} have mean zero and are uncorrelated with each other.[1] In the siblings example, we would call δ_i a "family effect;" it would contain unobserved family-level variables such as shared ability, parental work ethic, parenting skills, and so on. Meanwhile, ε_{ij} would contain unobserved sibling-specific variables like a sibling's own IQ, own school experience, and so on.

We can rewrite equations (1) and (2) as follows:

$$Y_{ij} = \alpha + X_{ij}'\beta + Z_i'\gamma + \delta_i + \varepsilon_{ij} \tag{3}$$

We call equation (3) an *error components model*. The panel data methods we study below address two issues that arise in the estimation of the equation (3):

1. Error terms are positively correlated within each group: $cov(U_{ij}, U_{ik}) = V[\delta_i] > 0$ for $j \neq k$.
2. The δ_i component *may* be correlated with covariates X_{ij} and Z_j: $cov(\delta_i, X_{ij}) \neq 0$ or $cov(\delta_i, Z_i) \neq 0$.

Issue (1), which is *always* a concern in panel data, only affects the variance (standard errors) of our coefficient estimators. In the absence of a correlation between the error term and the covariates (i.e., when issue (2) is not a concern), OLS still delivers unbiased and consistent coefficient estimates, but they are inefficient, and the standard errors are incorrect. We have already learned one way to deal with within-group correlation of the error terms: clustering. The clustered standard errors we studied in Lecture Note (3) allow for any within-group serial correlation, including the form that arises in issue (1).[2] However, the OLS estimator with clustered standard errors is not efficient. Below, we will examine an alternative method, *random effects estimation*, which has better efficiency properties.

Issue (2) is a form of omitted variables bias: a correlation between the error term and the covariates. As we will see below, the panel structure of the data allows us to correct for this omitted variables bias in a very

[1] Note that $V[U_{ij}] = V[\delta_i] + V[\varepsilon_{ij}]$.

[2] In Stata, one can implement clustered standard errors with the following command: `reg y x z,cluster(groupvar)`.

sensible way. The method is called *fixed effects estimation.*

Before we continue, the distinction between random effects estimation and fixed effects estimation is worth repeating. We use random effects methods when our observations are grouped in some important way, but we are confident that there are no relevant group-level omitted variables. We use fixed effects methods when our observations are grouped in some important way, and we are concerned that there *are* relevant group-level omitted variables. Fixed effects estimation is easier to understand than random effects estimation, so we will start there.

3 Fixed Effects Estimation

The idea of fixed effects estimation is that we control for observation ij's membership in group i, so that our coefficient estimates are based only on *within-group* variation. We can do so in two ways. The first way, often called the "brute force approach," involves two steps. First, generate $N-1$ dummy variables:

$$D_i = \begin{cases} 1 & \text{if observation is in group } i \\ 0 & \text{otherwise} \end{cases}$$

for $i = 2, \cdots, N$. Then control for the dummies directly:

$$Y_{ij} = \alpha + X'_{ij}\beta + \sum_{i=2}^{N} D_i + \varepsilon_{ij} \tag{4}$$

Note that once we control for D_i, we cannot estimate coefficients for Z_i because it is collinear with the full vector of D_i dummies. Thus, fixed effects estimation can only handle covariates that vary within group. We interpret β as the association between Y_{ij} and X_{ij}, holding fixed all variables that are shared within each group.

The "brute force" approach works well when we have relatively few groups, but when N is large, regression (4) becomes computationally burdensome. In that case, we can use the "finesse" method, which involves first demeaning Y_{ij} and X_{ij} within each group and then running a regression using the demeaned data. To understand this method, first collapse equation (3) to group-level means:

$$\bar{Y}_i = \alpha + \bar{X}'_i\beta + \bar{Z}'_i\gamma + \delta_i \tag{5}$$

Here, we have averaged out all of the within-group variation in our data. Equation (5) is known as the *between*

regression because it uses only between-group variation. Now subtract equation (5) from equation (3):

$$Y_{ij} - \bar{Y}_i = \left(X_{ij} - \bar{X}_i\right)' \beta + \varepsilon_{ij} \tag{6}$$

This equation is known as the *within regression* because it uses only within-group variation. Both Z_i and δ_i have dropped out, so omitted variables bias from δ_i is no longer a concern. We can estimate the within regression by first estimating group-level means of Y_{ij} and X_{ij}, then calculating each observation's deviation from its group-level mean, $Y^*_{ij} = Y_{ij} - \bar{Y}_i$ and $X^*_{ij} = X_{ij} - \bar{X}_i$, and then running a regression of Y^*_{ij} on X^*_{ij}. This approach yields *identical* coefficient estimates to the "brute force" approach.

However, the second-stage regression in the "finesse" approach uses the incorrect number of degrees of freedom in calculating its standard errors. In particular, the second stage regression uses:

$$\hat{\sigma}^2_{OLS} = \frac{SSR}{NJ - K}$$

to estimate the variance of ε_{ij}. SSR is the sum of squared residuals, NJ is the number of individuals, and K is the number of elements in X_{ij}. But recall that we computed N group-level means in the first stage, thus using up N more degrees of freedom. So we should really use the estimator:

$$\hat{\sigma}^2_{FE} = \frac{SSR}{NJ - K - N} = \frac{SSR}{N(J-1) - K}$$

Because it uses the correct degrees-of-freedom adjustment, $\hat{\sigma}^2_{FE}$ is an unbiased estimator of the variance of ε_{ij}. We thus inflate all of the entries in $V[\hat{\beta}]$ by:

$$\omega = \frac{\hat{\sigma}^2_{FE}}{\hat{\sigma}^2_{OLS}} = \frac{NJ - K}{N(J-1) - K}$$

to obtain unbiased variance estimates. More directly, we multiply the standard errors from the second-stage regression by $\sqrt{\omega}$.

Stata offers several commands for fixed effects estimation:

- `reg y x i.groupvar` (Depending on the version of Stata, you may need to type `xi:` before this command.)
- `areg y x,a(groupvar)`
- `xtreg y x,i(groupvar) fe`

4 Random Effects Estimation

When δ_i is uncorrelated with X_{ij} and Z_i, we can use both between and within variation to identify $\hat{\beta}$. OLS estimation of equation (1) uses both sources of variation, but the standard errors are incorrect, and the coefficient estimators are not efficient. Random effects estimation weights the between and within components of the dependent and independent variables in a way that minimizes the variance of $\hat{\beta}$ and $\hat{\gamma}$. It places weight 1 on within variation and weight C on between variation:

$$\left((Y_{ij} - \bar{Y}_i) + C\bar{Y}_i\right) = \alpha + \left((X_{ij} - \bar{X}_i) + C\bar{X}_i\right)'\beta + CZ_i'\gamma + C\delta_i + \varepsilon_{ij} \tag{7}$$

where:

$$C = \sqrt{\frac{V[\varepsilon_{ij}]}{V[\varepsilon_{ij}] + JV[\delta_i]}}$$

This optimal weight has intuitive implications. As the variance of δ_i increases relative to the variance of ε_{ij}, the random effects model places less weight on between variation and more weight on within variation. In our siblings example, this result implies that when the residual variance is large across families but small within families, we place more weight on variation coming from within families.

Now note that:

$$(Y_{ij} - \bar{Y}_i) + C\bar{Y}_i = Y_{ij} - (1 - C)\bar{Y}_i = Y_{ij} - D\bar{Y}_i$$

where:

$$D = 1 - C = 1 - \sqrt{\frac{V[\varepsilon_{ij}]}{V[\varepsilon_{ij}] + JV[\delta_i]}}$$

So in addition to equation (7), we can estimate a random effects model by quasi-differencing:

$$Y_{ij} - D\bar{Y}_i = (1 - D)\alpha + (X_{ij} - D\bar{X}_i)'\beta + (1 - D)Z_i'\gamma + (1 - D)\delta_i + \varepsilon_{ij} \tag{8}$$

This quasi-differencing procedure is similar in spirit to the quasi-differenced models we discussed in the time series lecture. Both models are examples of *generalized least squares*, or GLS. We say "generalized" because we need to estimate D before estimating the main regression. To do so, we first estimate equation (1) by OLS (which will give us unbiased but inefficient estimates), then estimate $V[\delta_i]$ and $V[\varepsilon_{ij}]$ from the OLS residuals, and then calculate D using our estimates $\hat{V}[\delta_i]$ and $\hat{V}[\varepsilon_{ij}]$. Some complications arise in computing standard errors for the random effects model, but those complications are beyond the scope of this course.

In Stata, you can estimate a random effects model by typing: `xtreg y x z,i(groupvar) re`.

5 RE or FE?

When students first learn about random effects and fixed effects estimation, they are often confused about which method to use in any given situation. Sections 3 and 4 suggest a couple of general lessons.

- If one is interested in the relationship between a group-level variable (Z_i) and the outcome (Y_{ij}), a random effects model is preferable to a fixed effects model because the latter does not allow for the estimation of $\hat{\gamma}$. In our siblings example, one cannot use fixed effects (and should therefore use random effects) to estimate the relationship between parents' education and their children's earnings in adulthood.
- If one is interested in the relationship between a variable that is not constant within group (X_{ij}) and the outcome (Y_{ij}), then both random effects and fixed effects are feasible. The choice between the two depends on whether one thinks that the group-level error component is correlated with the regressors of interest (X_{ij}). In our siblings example, suppose that all siblings in our sample were entered into a lottery for college financial aid. The assignment of financial aid was random, so in measuring its effects on subsequent educational attainment or earnings, we need not worry about omitted variables. But siblings are still more similar than non-siblings, which random effects estimation very appealing. On the other hand, if financial aid is conditioned on family income and other family attributes, then family-level covariates are important, making fixed effects estimation more appropriate.

In the second case, when the choice between random effects and fixed effects is non-trivial, we can also draw on a test called the Hausman test. To determine whether fixed effects are necessary, the Hausman test adds the demeaned covariates to the random effects specification and asks whether their estimated coefficients are jointly significantly different from zero. In other words, it runs:

$$Y_{ij} - D\bar{Y}_i = (1-D)\alpha + (X_{ij} - D\bar{X}_i)'\beta + (X_{ij} - \bar{X}_i)'\tilde{\beta} + (1-D)Z_i'\gamma + (1-D)\delta_i + \varepsilon_{ij} \tag{9}$$

and then performs an F-test of the hypothesis that $\tilde{\beta} = 0$.

To implement the Hausman test in Stata, you need to save your random effects results and fixed effects results and then have Stata test between them:

```
xtreg y x, i(groupvar) re
estimates store re
xtreg y x, i(groupvar) fe
estimates store fe
hausman fe re
```

6 Time in Panel Data

Up to this point, we have primarily discussed observations that are grouped but unordered. But we will often encounter panel data that are both grouped and ordered, especially when we observe a single entity (person, state, country, etc.) over time. We typically use the subscripts i and t when describing this type of panel data, which is also called longitudinal data. All of the models described above also apply to longitudinal data. However, we may now wish to also include controls that vary over time but not across entities, as well as a time component in the error term:

$$Y_{it} = \alpha + X_{it}'\beta + Z_i'\gamma + W_t'\lambda + \delta_i + \tau_t + \varepsilon_{it} \tag{10}$$

In our state traffic laws example (where Y_{it} is motor vehicle fatalities), Z_i and δ_i vary across states but not over time, while W_t and τ_t vary over time but are shared across states. Z_i and δ_i might capture state differences in driving culture, while W_t and τ_t might capture changes in automotive technology or changes in national traffic laws.[3] In practice, we usually control for time variation by including time fixed effects: that is, including a separate dummy variable for each year (excluding the first as the base category). When we control for both state and time fixed effects, the estimating equation appears as:

$$Y_{it} = \alpha + X_{it}'\beta + \delta_i + \tau_t + \varepsilon_{it} \tag{11}$$

In the traffic laws example, the δ_i fixed effect controls for all time-invariant characteristics of state i, while the τ_t fixed effect controls for all national time trends that are shared by all states. Thus, the estimator for β in the above regression is identified from within-state, within-time variation. If X_{it} is a dummy for a seat-belt law, then β measures a state's change in fatalities that is associated with a change in seat-belt laws, net of national trends in fatalities over the same period. We are effectively using time trends in states that did not change their seat-belt laws as controls. Section 8 below will make this notion more explicit.

The Stata commands from Section 3 apply here, with the added wrinkle that we control for time fixed effects:

- `reg y x i.statevar i.timevar`
- `areg y x i.timevar,a(statevar)`
- `xtreg y x i.timevar,i(statevar) fe`

[3] Z_i and W_t are measured, while δ_i and τ_t are not.

7 First Differencing

For longitudinal data, first difference estimation offers an alternative to fixed effects estimation. Consider a first-differenced version of equation (11):

$$\begin{aligned}
\Delta Y_{it} &= Y_{it} - Y_{i,t-1} \\
&= (\alpha + X'_{it}\beta + \delta_i + \tau_t + \varepsilon_{it}) - \left(\alpha + X'_{i,t-1}\beta + \delta_i + \tau_{t-1} + \varepsilon_{i,t-1}\right) \\
&= (X_{it} - X_{i,t-1})'\beta + (\tau_t - \tau_{t-1}) + (\varepsilon_{it} - \varepsilon_{it-1}) \\
&= \Delta X'_{it}\beta + \Delta\tau_t + \Delta\varepsilon_{it}
\end{aligned}$$

The δ_i fixed effect drops out of the regression, so we can also use first differencing to estimate β from within variation. To account for the $\Delta\tau_t$ component of the error term, we include dummies for each time period t. In the two-period case, the first difference and fixed effects estimators are identical. With more than two periods, the estimators yield different results, but both are consistent. Fixed effects estimation is more efficient when the ε_{it} error terms are i.i.d., while first difference estimation is more efficient when the ε_{it} error terms follow a random walk. In most situations, the reality lies somewhere between those two extremes, so neither method is generally superior to the other. One practical consideration is that first differencing becomes difficult in unbalanced samples (i.e., when some time periods are missing for some states).

The easiest way to carry out first difference estimation in Stata is to first declare the structure of your panel data:

```
xtset statevar timvar
```

Then we can use time-series operators as follows:

```
reg D.y D.x i.timevar
```

The `D.` operator takes the first difference of the variable within each state.

8 Difference-in-Difference Estimation

Policy analysts very commonly use a related panel data technique called difference-in-difference estimation. A new example will be useful. Suppose we are studying the effect of the Massachusetts health reform on adult insurance coverage. We have a sample of many adults (i) living in all 50 states (s) over several years (t) before and after health reform implementation. The variable Y_{ist} equals 1 if individual i from state s in year t has insurance coverage, 0 if not.

Many policy analyses proceed with either a cross-sectional comparison or a time-series comparison. The cross-sectional approach would involve comparing insurance coverage among Massachusetts residents with

insurance coverage among residents of other states during the post-policy roll-out period:

$$\Delta_s = \bar{Y}_{MA,POST} - \bar{Y}_{OTHER,POST}$$

The time-series approach would involve looking at Massachusetts residents over time:

$$\Delta_t = \bar{Y}_{MA,POST} - \bar{Y}_{MA,PRE}$$

Both of these approaches have problems. The cross-sectional comparison is biased by the many other state-level differences in the determinants of insurance access, while the time-series comparison is biased by national economic trends, for example.

To eliminate both types of bias, we can estimate a difference-in-difference model:

$$\Delta\Delta = \left(\bar{Y}_{MA,POST} - \bar{Y}_{MA,PRE}\right) - \left(\bar{Y}_{OTHER,POST} - \bar{Y}_{OTHER,PRE}\right)$$

Just as in the first-differenced models in Section 7, this estimator eliminates time-invariant differences across states, as well as time effects that are shared by all states. We can use the following regression to estimate $\Delta\Delta$:

$$Y_{ist} = \beta_0 + \beta_1 POST_t + \beta_2 MA_s + \beta_3 POST_t * MA_s + U_{ist} \tag{12}$$

where $POST_t$ is a dummy for the post-reform era (in all states), MA_s is a dummy for Massachusetts, and $POST_t * MA_s$ is their interaction. To see how this regression relates to difference-in-difference estimation, we consider each of the means in the expression for $\Delta\Delta$ separately:

- $\bar{Y}_{OTHER,PRE} = \beta_0$
- $\bar{Y}_{OTHER,POST} = \beta_0 + \beta_1$
- $\bar{Y}_{MA,PRES} = \beta_0 + \beta_2$
- $\bar{Y}_{MA,POST} = \beta_0 + \beta_1 + \beta_2 + \beta_3$

Now calculate:

$$\Delta\Delta = \left(\bar{Y}_{MA,POST} - \bar{Y}_{MA,PRE}\right) - \left(\bar{Y}_{OTHER,POST} - \bar{Y}_{OTHER,PRE}\right) = \beta_3$$

So the coefficient on the interaction term equals the difference-in-difference estimator.

A more general version of equation (12) is:

$$Y_{ist} = \tau_t + \delta_s + \beta POLICY_{st} + \varepsilon_{ist} \tag{13}$$

where $POLICY_{st}$ equals one in state s only after the health policy has been implemented in that state. τ_t and δ_s are year and state fixed effects. β in equation (13) has the same interpretation as β_3 in equation (12), but we can now study states that implemented health policy reform at different times.

In the 1980s and 1990s, researchers often estimated equation (13) under the assumption that ε_{ist} is i.i.d. across observations. But it may not be i.i.d. for several reasons. First, individuals from the same state and year may have correlated errors, so that $cov(\varepsilon_{ist}, \varepsilon_{jst}) \neq 0$ for $i \neq j$. In this case, we would want to cluster observations at the state-year level:

```
egen state_year_category = group(state year)
reg y policy i.state i.year,cluster(state_year_category)
```

Second, because policy changes are very persistent, the error terms may be serially correlated over time within each state. First differencing offered one way to deal with this problem, but another is to cluster by state across all time periods (and across all individuals within each state):

```
reg y policy i.state i.year,cluster(state)
```

This command represents the "state of the art" in difference-in-difference analysis. Note that we may have millions of individual observations in our dataset, but we are assuming only 50 independent entities. Clustering works well if the data have at least 40 clusters but can be unreliable with a smaller number of clusters. For more information on this matter, consult Angrist and Pischke's book, *Mostly Harmless Econometrics*.

LECTURE NOTE 10: CAUSALITY

1 Introduction

A great deal of social science and policy research takes interest in the "causal effect" of some policy, program, treatment, or experience. However, in this course, we have in large part maintained a conservative view of regression as linear projection (i.e., correlation or association). We have at times discussed whether our coefficients reflect "causal effects," but we have used this term rather loosely. This lecture note aims to clarify the meaning of causality.

2 Potential Outcomes

The most common approach to defining causality involves the concept of potential outcomes. This approach is called the *potential outcomes framework*, or the *Rubin Causal Model*, after its creator Don Rubin, currently a professor of statistics at Harvard.[1] The key idea is that each individual has her own set of potential outcomes, $Y_i(t)$, where t is a treatment level. In general, t can take on many values, but we will primarily focus on a binary treatment, so that each individual has two potential outcomes, $Y_i(1)$ and $Y_i(0)$. These potential outcomes reflect the outcome the individual would experience with and without the treatment, respectively. The *causal effect* of t for individual is $\alpha_i = Y_i(1) - Y_i(0)$. Note that causal effects may vary across individuals; i.e., they may be heterogeneous.

Unfortunately, we do not observe both $Y_i(1)$ and $Y_i(0)$ for each individual, so we cannot calculate the causal effect for each individual. To see this, let T_i denote the treatment level assigned to i. (In the binary case, $T_i = 0$ or 1.) For each individual, we observe only the realized outcome:

$$Y_i = Y_i(T_i) = T_i Y_i(1) + (1 - T_i) Y_i(0)$$

We wish to make inferences about $\alpha_i = Y_i(1) - Y_i(0)$, but we only have data on (Y_i, T_i).

[1] The attribution of the potential outcomes framework to Don Rubin is a little unfair to Jerzy Neyman, who conceived of a very similar model in the 1920s.

3 Common Estimands of Interest

In making inferences about the distribution of α_i, we will focus on two averages:

- The *Average Treatment Effect (ATE)* for the whole population:

$$ATE = E[Y_i(1) - Y_i(0)] = E[\alpha_i]$$

- The average treatment effect on the individuals who were treated. This is known as either the *mean effect of Treatment on the Treated (TOT)* or the *Selected Average Treatment Effect (SATE)*:

$$TOT = E[Y_i(1) - Y_i(0)|T_i = 1] = E[\alpha_i|T_i = 1]$$

When the treatment effect is homogeneous in the population (so that $\alpha_i = \alpha$ for all i), the ATE and the TOT are the same. But when treatment effects are heterogeneous, the two estimands are different. For example, suppose we are estimating the effect of college education on earnings. If high-ability children are more likely to attend college but also gain more from attending, then the TOT will be larger than the ATE. If poor children have high returns from attending college but are unable to attend because of credit constraints, then the TOT will be smaller than the ATE. Note that these differences are not due to bias; they simply reflect averages of different distributions of treatment effects.

4 Selection Bias

We are now equipped to characterize selection bias. Suppose we observe treatment status and an outcome for a sample of individuals. Intuition might lead us to try estimating the effect of the treatment on the outcome by taking the difference in mean outcomes between the treated and untreated groups:

$$E[Y_i|T_i = 1] - E[Y_i|T_i = 0]$$

If we rewrite this expression using the underlying potential outcomes, we will notice a potential problem:

$$\begin{aligned}
E[Y_i|T_i = 1] - E[Y_i|T_i = 0] &= E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 0] \\
&= (E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 0]) + (E[Y_i(0)|T_i = 1] - E[Y_i(0)|T_i = 1]) \\
&= \underbrace{(E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 1])}_{TOT} + \underbrace{(E[Y_i(0)|T_i = 1] - E[Y_i(0)|T_i = 0])}_{selection\ bias} \\
&= \underbrace{E[\alpha_i|T_i = 1]}_{TOT} + \underbrace{(E[Y_i(0)|T_i = 1] - E[Y_i(0)|T_i = 0])}_{selection\ bias}
\end{aligned}$$

So the difference in the group-level means is equal to the TOT plus a selection bias term. The selection bias term reflects differences in the distributions of baseline outcomes ($Y_i(0)$) between individuals with $T_i = 0$ and $T_i = 1$. For the difference in means to have a causal interpretation, we must assume:

$$E[Y_i(0)|T_i = 1] = E[Y_i(0)|T_i = 0]$$

which implies that baseline outcomes for the treatment and control groups have the same means. Under this condition, the difference in means is an unbiased estimator of the mean effect of treatment on the treated. Note that $E[Y_i(0)|T_i = 1]$ is unobservable, so the assumption above is fundamentally untestable.[2]

5 Unconfoundedness

The condition that $E[Y_i(0)|T_i = 1] = E[Y_i(0)|T_i = 0]$ is somewhat abstract, so researchers usually make a more general assumption:

$$(Y_i(0), Y_i(1)) \perp T_i$$

where $\perp$ denotes independence. This assumption is called the *unconfoundedness assumption*. It states that potential outcomes are independent of treatment, and it implies that $E[Y_i(0)|T_i = 1] = E[Y_i(0)|T_i = 0]$. A properly-implemented randomized trial (in which T_i is randomly assigned to members of the population) guarantees unconfoundedness.

Sometimes we wish to condition on covariates, as in a regression setting. The unconfoundedness assumption can easily accommodate covariates:

$$(Y_i(0), Y_i(1)) \perp T_i \big| X_i$$

This slightly-expanded condition states that, conditional on covariates, potential outcomes are independent of treatment. The expanded unconfoundedness assumption has several alternative names: *ignorable treatment assignment*, the *conditional independence assumption*, and *selection on observables.* As with the conditional mean assumption in Section 4, these assumptions are untestable.

6 Connection to Linear Regression

Previously, we would have tried to estimate the effect of T_i on Y_i by running the regression:

$$Y_i = \alpha T_i + X_i'\beta + U_i$$

[2] Although the assumption is untestable, we can still shed light on it by testing for mean differences between the treatment and control groups in variables that should not have been affected by the treatment.

Lecture Notes 3 and 4 suggest that α has a causal interpretation if $E[U_i|X_i, T_i] = 0$ (implying that the error term is uncorrelated with X_i and T_i). This assumption's connection with unconfoundedness is not obvious.

In fact, the OLS assumption about the conditional mean of the error term bears a very close relationship with the unconfoundedness assumption. To see this relationship, we rewrite the linear regression specification above in the potential outcomes framework. We begin by writing $Y_i(0)$ as a linear function of the covariates:

$$Y_i(0) = X_i'\beta + U_i$$

where $E[U_i|X_i] = 0$ by construction. Next, we note that in the linear regression specification above, the effect of T_i is the same for all i: $\alpha_i = \alpha$ for all i. As a result, we can write $Y_i(1)$ as the sum of $Y_i(0)$ and a constant treatment effect:

$$Y_i(1) = \alpha + Y_i(0) = \alpha + X_i'\beta + U_i$$

These expressions for the potential outcomes $Y_i(1)$ and $Y_i(0)$ imply:

$$\begin{aligned} Y_i &= Y_i(0) + \alpha T_i \\ &= \alpha T_i + X_i'\beta + U_i \end{aligned}$$

where Y_i is the observed outcome for individual i. Thus, there is a direct mapping from the potential outcomes framework to linear regression.

Now consider with the unconfoundedness assumption:

$$(Y_i(0), Y_i(1)) \perp T_i \big| X_i$$

This assumption implies:

$$E[Y_i(0)|X_i, T_i = 1] - E[Y_i(0)|X_i, T_i = 0] = 0$$

which guarantees the absence of selection bias. In the model above, we can re-write this expression as:

$$E[X_i'\beta + U_i|X_i, T_i = 1] - E[X_i'\beta + U_i|X_i, T_i = 0] = 0$$

This condition is implied by the OLS error term assumption:

$$E[U_i|X_i, T_i] = 0$$

showing that the unconfoundedness assumption is closely tied to the OLS assumption about the conditional mean of the error term.

7 Randomized Experiments

Many research designs, including cross-sectional and panel regression, can be interpreted through the unconfoundedness assumption. But the method most closely identified with unconfoundedness is the randomized experiment. This section describes two approaches to randomization in policy research and discusses the estimation issues that arise in each. Importantly, in both instances, we will assume that randomization does not change the pool of applicants or their behavior.

One common type of policy experiment is the *randomized social experiment*. In such an experiment, the experimenter controls T_i directly. For example, if a social program is over-subscribed (so the number of applicants exceeds the capacity of the program), the program administrator might randomly choose which applicants to accept. Applicants have already expressed their interest in program participation, so if they are accepted, they will participate.[3] As a result, the difference in means between the treatment (accepted) and control (unaccepted) groups provides an unbiased estimator of the TOT.

Another common type of policy experiment is eligibility randomization. Here, the experimenter randomly chooses individuals in the population to be given eligibility to participate in the program. Some treatment group individuals will decide to participate; some will not. In consequence, the difference in means between the treatment and control groups measures the effect of *eligibility*, not *treatment*. The average effect of eligibility is known as the *intent-to-treat effec*t:

$$ITT = E[Y_i|E_i = 1] - E[Y_i|E_i = 0]$$

where E_i denotes program eligibility. Under eligibility randomization, we can still retrieve the TOT. We need three assumptions:

1. *Eligibility is randomly assigned.* This assumption guarantees identification of the ITT.
2. *The effect of group assignment on outcomes only operates through treatment.* This assumption is important because it allows us to exclude E_i from the conditional expectation of Y_i:

$$E[Y_i|E_i, T_i] = E[Y_i|T_i]$$

Conditional on treatment status, eligibility does not affect the conditional expectation of Y_i. More

[3] If many accepted applicants end up not participating in the program, then the experiment will be more similar to eligibility randomization, described below.

broadly, we might define a potential outcome $Y_i(e,t)$ that depends on both the eligibility level e and the treatment level t. The assumption says that we can exclude e from the potential outcome function without losing any information:

$$Y_i(e,t) = Y_i(t)$$

The potential outcomes assumption guarantees the conditional expectation assumption, but not vice versa.

3. *Ineligibles cannot participate in the program.* In other words, $Pr[T_i = 1|E_i = 0] = 0$.

Under assumption (3), we can separate individuals into two groups: compliers and never-takers. The compliers have $T_i = 1$ if $E_i = 1$ and $T_i = 0$ if $E_i = 0$. The never-takers have $T_i = 0$ regardless of the value of E_i. In the eligible group, we can directly identify compliers and never-takers based on who opts into treatment. In the ineligible group, we cannot directly identify compliers and never-takers, but we know they exist. By asumption (1), the fraction of compliers (called the "compliance rate") is the same in the eligible and ineligible groups. Then we can write:

$$\begin{aligned} ITT \quad = \quad & (\text{mean effect of eligibility on outcomes of compliers})\,(\text{compliance rate}) \\ & + (\text{mean effect of eligibility on outcomes of never-takers})\,(1 - \text{compliance rate}) \end{aligned}$$

Note that the mean effect of eligibility on the on the outcomes of compliers is the TOT. Furthermore, by assumption (2), the mean effect of eligibility on the on the outcomes of never-takers is zero. Thus:

$$ITT = TOT \cdot Pr[T_i = 1|E_i = 1] + 0 \cdot Pr[T_i = 0|E_i = 1]$$

Rearrange to obtain:

$$TOT = \frac{ITT}{Pr[T_i = 1|E_i = 1]}$$

So the TOT equals the ITT divided by the compliance rate.

8 Internal and External Validity

In discussions of statistical studies, people often invoke the concepts of internal and external validity. A study is called *internally valid* if it leads to correct inferences about the population being studied. In the Rubin Causal Model above, internal validity basically requires the unconfoundedness assumption to be true. Threats to internal validity can arise from non-random assignment or non-random attrition from the sample, for example. Experimentalists also worry about two other threats to internal validity:

- The *Hawthorne effect*, which refers to situations in which subjects change their behavior just as a result of being studied (irrespective of treatment status).
- The *John Henry effect*, which refers to situations in which subjects in the control group change their behavior to offset the inequality in treatment status they observe.

Most broadly, we say a study is internally valid if the assumptions of its statistical methods are met.

A study is called *externally valid* if its conclusions can be generalized to other populations. In this lecture note, we have discussed the fact that the TOT does not always equal the ATE, implying that the TOT is not externally valid for the whole population. Furthermore, the ATE in one setting may be different from the ATE in another. Economists particularly like talking about one class of threats to external validity: *general equilibrium effects*. General equilibrium effects refer to the scalability of the results. For example, the Project STAR experiment gave internally valid estimates of the effects of class size reductions. But the class size reductions took place in partial equilibrium; the classes involved in the experiment were a small subset of all classes in the state of Tennessee. If Tennessee's state government decided to reduce class sizes in the whole state, the state-wide class size reduction would require a large expansion in the number of teachers. That expansion of the teacher workforce would probably change average teaching quality in some important way, so that the general equilibrium effects of small classes would be different from the effect estimates from Project STAR.

LECTURE NOTE 11: INSTRUMENTAL VARIABLES

1 Introduction

Suppose we are interested in estimating the causal effect of some variable X on another variable Y. We can compare the values of Y among individuals with different X's, but we are concerned about omitted variables and reverse causality. We theorize that the following system underlies the joint distribution of X and Y:

$$\begin{array}{ccc} X & \leftrightarrow & Y \\ \updownarrow & & \updownarrow \\ \multicolumn{3}{c}{\textit{other vars}} \end{array}$$

We wish to disentangle the arrow(s) pointing from X to Y from the other arrows in the causal system. In this lecture note, we study a method that takes advantage of an additional variable, Z, which causes X but bears no direct relation to any other variable in the system:

$$\begin{array}{ccccc} Z & \rightarrow & X & \leftrightarrow & Y \\ & & \updownarrow & & \updownarrow \\ & & \multicolumn{3}{c}{\textit{other vars}} \end{array}$$

Z is called an *instrument*. It can help us learn about the causal effect of X on Y because its association with Y is solely mediated by its effect on X. Specifically, any relationship between Z and Y *must* be mediated by the following causal path:

$$\begin{array}{ccccc} Z & \rightarrow & X & \rightarrow & Y \\ & & \downarrow & & \uparrow \\ & & \multicolumn{3}{c}{\textit{other vars}} \end{array}$$

All arrows pointing *toward* X drop out of the system, leaving a "direct" effect of X on Y and an "indirect" effect that is mediated by other variables. Thus, Z can help us learn about the overall causal effect of X on Y, but it cannot help us disentangle mechanisms. The *method of instrumental variables* (IV) uses variation in Z to identify the causal effect of X on Y.

What sorts of variables are instruments? One basic example is treatment assignment in a randomized experiment. In such a setting, we *know* that treatment assignment is correlated with the dependent variable

only because of the effect of the treatment on the dependent variable. Outside of randomized experiments, instruments are harder to find, and their validity is more open to debate. Here are a few examples:

1. Suppose we want to estimate the price elasticity of the demand for corn. We can look at the relationship between the equilibrium price and the equilibrium quantity sold, but that relationship is driven by both supply *and* demand. Ideally, we would like shift the supply curve while holding the demand curve fixed, which would allow us to follow movements *along* the demand curve. Weather in corn-growing regions provides a reasonable instrument for this purpose. In particular, bad rainfall in corn-growing regions affects the supply of corn but is unlikely to affect the demand for corn. Any relationship between weather in corn-growing regions and the price of corn is thus likely to be mediated by the movement of the supply curve against a fixed demand curve.
2. Suppose we want to estimate the effect of schooling on earnings. We can look at the relationship between schooling and earnings (either between or within families), but both schooling and earnings are correlated with ability, which we cannot perfectly observe. The literature has suggested several instruments for schooling, including the following two early examples:
   - *Quarter of birth*: Because teenagers in most U.S. states are required to stay in school until their sixteenth birthdays, the age at school entry has an effect on educational attainment. A child who entered school at age 4 must wait until the 11th grade before she is free to drop out, whereas a child who entered school at age 5 may leave in the 10th grade. Due to school entry regulations, a child born in December may start kindergarten at age 4, but a child born in January must typically wait until age 5. Thus, children born in the first quarter of the year obtain less schooling than children born in the fourth quarter. Angrist and Krueger (1991) argue that any differences in earnings by quarter of birth must be mediated by the effect of quarter of birth on education. By this reasoning, quarter of birth can serve as an instrument for educational attainment.
   - *Distance from college*: Many young adults enroll in college only if one is available close to home. As a result, the distance of the childhood home from the nearest college is a determinant of educational attainment. Card (1995) argues that any association between earnings and distance from college must be mediated by the effect of distance from college on college enrollment. By this reasoning, distance from college can serve as an instrument for educational attainment.
3. Suppose we want to estimate the effect of the number of children on a mother's labor supply. We can compare labor force participation rates of mothers with different numbers of children, but because labor supply and fertility both reflect maternal decisions, their relationship does not tell us about the causal effect of interest. Before the growth of fertility treatment technology, the birth of twins might be treated as close to random, so that twin births can serve as an instrument for family size. Based on

this argument, Angrist and Evans (1998) use twins to measure the effect of the number of children on maternal labor supply.

Although these instruments *plausibly* adhere to the causal system drawn above, all of them have potential problems. Can you think of some potential problems?

The remainder of this lecture note details how we can use instruments to uncover causal effects. We begin with the case in which the causal effect of X is the same for all individuals, and we then discuss the case in which the causal effect of X is heterogeneous in the population.

2 Instrumental Variables with Homogeneous Causal Effects

2.1 Instrumental Variables with a Single Regressor and a Single Instrument

Suppose we wish to estimate the effect of X_i on Y_i in the following equation:

$$Y_i = \beta_0 + \beta_1 X_i + U_i$$

However, X_i is endogenous, meaning that it is correlated with U_i. For instance, when X_i is years of schooling, and Y_i is earnings, researchers worry that both X_i and Y_i are correlated with "innate" ability, an unobserved variable that appears in the error term U_i. This identification problem fits naturally into the diagrams in the introduction, but we can draw a diagram that is a bit closer to our statistical model:

$$\begin{array}{ccc} X_i & \rightarrow & Y_i \\ \uparrow & \nearrow & \\ U_i & & \end{array}$$

This diagram describes the same causal system as the diagram in the introduction, but it captures the endogeneity of X_i entirely through the error term U_i. Because $cov(X_i, U_i) \neq 0$, we cannot estimate the model by OLS.

As described in the introduction, our solution is to find an instrument Z_i that causes X_i but is otherwise unrelated to the other variables in the system:

$$\begin{array}{ccccc} Z_i & \rightarrow & X_i & \rightarrow & Y_i \\ & & \uparrow & \nearrow & \\ & & U_i & & \end{array}$$

As the diagram suggests, two conditions are necessary for Z_i to be a valid instrument:

1. *Instrument relevance*: $cov(Z_i, X_i) \neq 0$. In words, Z_i must be correlated with X_i.

2. *Instrument exogeneity*: $cov(Z_i, U_i) = 0$. In words, Z_i must be uncorrelated with all other determinants of Y_i. This condition is also known as an *exclusion restriction* because, conditional on X_i, Z_i can be excluded from the causal system.

Most disagreements about instrument validity concern the second assumption, regarding the exogeneity of the instrument.

How exactly can we use Z_i to estimate β_1? The answer becomes apparent when we derive the covariance between Z_i and Y_i:

$$cov(Z_i, Y_i) = cov(Z_i, \beta_0 + \beta_1 X_i + U_i) = \beta_1 cov(Z_i, X_i) + cov(Z_i, U_i)$$

By assumption, $cov(Z_i, U_i) = 0$. So we can rearrange terms and obtain:

$$\beta_1 = \frac{cov(Z_i, Y_i)}{cov(Z_i, X_i)} = \frac{cov(Z_i, Y_i)/V[Z_i]}{cov(Z_i, X_i)/V[Z_i]}$$

The second equality tells us that β_1 is the ratio of two coefficients. The numerator is the coefficient on Z_i from a regression of Y_i on Z_i, while the denominator is the coefficient on Z_i from a regression of X_i on Z_i. This expression suggests the following procedure:

1. Run a regression of Y_i on Z_i. This regression is often called the *reduced form regression*.
2. Run a regression of X_i on Z_i. This regression is often called the *first stage regression*.
3. Take the ratio of the coefficient on Z_i in the reduced form regression to the coefficient on Z_i in the first stage regression.

The ratio of the reduced form coefficient to the first stage coefficient is called an instrumental variables estimator. It is a consistent estimator of β_1.

We can obtain the same estimator by another procedure, called *two-stage least squares* (*TSLS* or *2SLS*). Appropriately, two-stage least squares involves two steps:

1. *First stage*: Estimate the first stage regression, $X_i = \pi_0 + \pi_1 Z_i + V_i$, by OLS and generate a predicted value for X_i, $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$.
2. *Second stage*: Estimate the causal model of interest by OLS, using the predicted value of X_i instead of the actual value of X_i. In other words, run: $Y_i = \beta_0^{TSLS} + \beta_1^{TSLS} \hat{X}_i + \varepsilon_i$.

To see how the TSLS estimator relates to the IV estimator above, we write:

$$\begin{aligned}
\hat{\beta}_1^{TSLS} &= \frac{\widehat{cov}(Y_i, \hat{X}_i)}{\hat{V}[\hat{X}_i]} \\
&= \frac{\widehat{cov}(Y_i, \hat{\pi}_0 + \hat{\pi}_1 Z_i)}{\hat{V}[\hat{\pi}_0 + \hat{\pi}_1 Z_i]} \\
&= \frac{\hat{\pi}_1 \widehat{cov}(Y_i, Z_i)}{\hat{\pi}_1^2 \hat{V}[Z_i]} \\
&= \frac{\widehat{cov}(Y_i, Z_i)}{\hat{\pi}_1 \hat{V}[Z_i]} \\
&= \frac{\widehat{cov}(Y_i, Z_i)}{\left(\frac{\widehat{cov}(X_i, Z_i)}{\hat{V}[Z_i]}\right) \hat{V}[Z_i]} \\
&= \frac{\widehat{cov}(Y_i, Z_i)}{\widehat{cov}(X_i, Z_i)}
\end{aligned}$$

Thus, in this case of a single endogenous variable and a single instrument, the TSLS estimator is the same as the ratio of the reduced form coefficient to the first stage coefficient.

The variance of the TSLS estimator must take into account uncertainty from both the first and second stage regressions. As a result, if one estimates the second stage regression using the `reg` command in Stata, the standard errors will be too small. We will not dwell on the details of TSLS variance estimation in this course, but for your information, the TSLS estimator has the following asymptotic distribution:

$$\hat{\beta}_1^{TSLS} \sim \mathcal{N}\left(\beta_1, V\left[\hat{\beta}_1^{TSLS}\right]\right)$$

where:

$$V\left[\hat{\beta}_1^{TSLS}\right] = \frac{1}{N} \frac{V\left[(Z_i - E[Z_i])U_i\right]}{\left(cov(Z_i, X_i)\right)^2}$$

The standard error of $\hat{\beta}_1^{TSLS}$ is the square root of $V\left[\hat{\beta}_1^{TSLS}\right]$. In large samples, we can compute p-values and confidence intervals as usual.

The next two sub-sections describe special cases of instrumental variables regression with a single endogenous regressor and a single instrument.

2.1.1 Wald Estimator

When Z_i is a binary variable, the IV estimator simplifies considerably. Define $\bar{Y}_0$ and $\bar{Y}_1$ as the means of Y_i in the sub-samples with $Z_i = 0$ and $Z_i = 1$, respectively. Define $\bar{X}_0$ and $\bar{X}_1$ similarly. Then:

$$\beta_1^{Wald} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{X}_1 - \bar{X}_0}$$

is a consistent estimator for β_1. In the quarter of birth example, the denominator is the difference in average educational attainment between individuals born in the fourth and first quarters, while the numerator is the difference in average earnings between the same groups. This ratio is known as the *Wald estimator*.

The Wald estimator highlights a link between IV estimation and the analysis of randomized controlled trials. To see this link, consider a program evaluation with eligibility randomization. Let Z_i be eligibility, and let X_i be actual participation in the program. Under the assumptions discussed in Lecture Note 10, ineligibles cannot participate in the program, so $\bar{X}_0 = 0$. Meanwhile, $\bar{X}_1$ is the participation rate among eligibles, and $\bar{Y}_1 - \bar{Y}_0$ is the ITT, so β_1^{Wald} is a consistent estimator of the TOT.

2.1.2 Instrumental Variables and Measurement Error

IV helps us deal with all sources of a correlation between X_i and U_i. Until now, we have mainly considered the roles of omitted variables and reverse causality, but recall from Lecture Note 4 that measurement error in X_i also induces a correlation between X_i and U_i. If we have two measurements of X_i, we can instrument one measurement for the other to eliminate attenuation bias from measurement error.

To see how IV helps us deal with measurement error, suppose we wish to estimate the model:

$$Y_i = \beta_0 + \beta_1 X_i^* + U_i$$

where $E[U_i|X_i^*] = 0$. Instead of observing X_i^* directly, we have two measurements of X_i^*:

$$X_i^1 = X_i^* + \nu_i \quad \text{and} \quad X_i^2 = X_i^* + \omega_i$$

where ν_i and ω_i are independent measurement errors with mean zero. For example, if X_i^* is schooling, we might have separate reports from the individual in our sample and her sibling. If X_i^* is blood pressure, we might have measurements from two separate doctor's visits.

In Lecture Note 4, we discovered that a regression of Y_i on X_i^1 leads to a coefficient with the following probability limit:

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 \left(\frac{1}{1 + \sigma_\nu^2 / \sigma_{X^*}^2} \right)$$

However, if we carry out TSLS using X_i^2 as the instrument, the TSLS coefficient for β_1 is consistent:

$$\hat{\beta}_1 \xrightarrow{p} \frac{cov(X_i^2, Y_i)}{cov(X_i^2, X_i^1)} = \frac{cov(X_i^* + \omega_i, Y_i)}{cov(X_i^* + \omega_i, X_i^* + \nu_i)} = \frac{cov(X_i^*, Y_i)}{V[X_i^*]} = \beta_1$$

The second equality comes from the fact that ν_i and ω_i are independent of each other and independent of X_i^*.

2.2 General Instrumental Variables Model

We now consider a model that allows for K endogenous variables, R exogenous control variables, and M instruments (where $K, R, M \geq 1$):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + \beta_{K+1} W_{1i} + \cdots + \beta_{K+R} W_{Ri} + U_i$$

$$cov(Z_{1i}, U_i) = 0 \quad cov(Z_{2i}, U_i) = 0 \quad \cdots \quad cov(Z_{Mi}, U_i) = 0$$

Because $W_{1i}, \cdots, W_{Ri}$ are exogenous (meaning they are uncorrelated with the error term), we do not need instruments to estimate $\beta_{K+1}, \cdots, \beta_{K+R}$. But $X_{1i}, \cdots, X_{Ki}$ are endogenous, and we thus must use $Z_{1i}, \cdots, Z_{Mi}$ as instruments to estimate $\beta_1, \cdots, \beta_K$. We assume that $Z_{1i}, \cdots, Z_{Mi}$ are all uncorrelated with U_i, so all satisfy instrument exogeneity. In order to estimate the model, the number of instruments must be at least as large as the number of endogenous variables: $M \geq K$.

We can use two-stage least squares to estimate $\beta_1, \cdots, \beta_K$ in this general setting:

1. Run first stage regressions of each X_{ki} on $Z_{1i}, \cdots, Z_{Mi}$ and $W_{1i}, \cdots, W_{Ri}$. Use the results to generate predicted values $\hat{X}_{ki}$.
2. Run a second stage regression using the predicted $\hat{X}_{ki}$'s:

$$Y_i = \beta_0^{TSLS} + \beta_1^{TSLS} \hat{X}_{1i} + \cdots + \beta_K^{TSLS} \hat{X}_{Ki} + \beta_{K+1}^{TSLS} W_{1i} + \cdots + \beta_{K+R}^{TSLS} W_{Ri} + \varepsilon_i$$

The resulting $\hat{\beta}_1^{TSLS}, \cdots, \hat{\beta}_K^{TSLS}$ are consistent estimators for $\beta_1, \cdots, \beta_K$.

As mentioned above, the assumption about the covariances of the Z_{mi}'s and U_i implies that $Z_{1i}, \cdots, Z_{Mi}$ satisfy *instrument exogeneity*. But recall from Section 2.1 that $Z_{1i}, \cdots, Z_{Mi}$ must also satisfy *instrument relevance*. In the general IV model, instrument relevance requires that the Z_{mi}'s be jointly predictive of all the X_{ki}'s, in the sense that the first stage predictions $\hat{X}_{1i}, \cdots, \hat{X}_{Ki}$ and the other covariates $1, W_{1i}, \cdots, W_{Ri}$ are not perfectly collinear. When the model include one instrument, one endogenous variable, and no control variables (i.e., the model in Section 2.1), the instrument relevance requirement is equivalent to a non-zero coefficient on the instrument in the first stage regression. One can assess instrument relevance by testing the hypothesis that all of the coefficients on $Z_{1i}, \cdots, Z_{Mi}$ in the first stage regression equal zero. In the case of a single endogenous regressor, if the F-statistic for this test is less than 10, the instruments are said to be *weak*, and the TSLS estimates are unreliable. Chapter 12 of the Stock and Watson book has further discussion on this point.

To implement TSLS in Stata 12, use the `ivregress` command. For three endogenous variables, four instruments, and one control variable, type: `ivregress 2sls y (x1 x2 x3 = z1 z2 z3 z4) w1`. You can

use the `robust` and `cluster` options as in all the other regression commands we have discussed. To see the first stage regression, add the option `first` to the command.

3 Instrumental Variables with Heterogeneous Causal Effects

We now consider the meaning of the IV (or TSLS) estimator when the causal effect of X_i on Y_i is heterogeneous. To do so, we return to the potential outcomes framework of Lecture Note 10. For simplicity, we start with the case of a single binary instrument, Z_i, and a single endogenous regressor, X_i. Because we are assuming that X_i is binary, we will refer to X_i as individual i's treatment status. Denote the *instrument level* as z and the *treatment level* as x. Let $Y_i(x, z)$ be the potential outcome for individual i at treatment level x and instrument level z, and let $X_i(z)$ be the potential treatment status for individual i at instrument level z. As in Lecture Note 10, every individual has her own set of potential outcomes and potential treatment statuses. In our data, we observe only $Y_i = Y_i(X_i(Z_i), Z_i)$, $X_i = X_i(Z_i)$, and Z_i.

Suppose we run two-stage least squares, using Z_i as an instrument for X_i. We know how to interpret $\hat{\beta}_1^{TSLS}$ when the effect of X_i is the same for all individuals, but the interpretation with heterogeneous effects is slightly more complex. To make such an interpretation possible, we make three assumptions:

1. Independence: $\{Y_i(x, z), X_i(z)\} \perp Z_i$. Potential outcomes and potential treatment levels are independent of the instrument.
2. Exclusion restriction: $Y_i(x, 0) = Y_i(x, 1)$. Conditional on x, the value of z does not affect the outcome. We can thus exclude z from the potential outcomes function: $Y_i(x, z) = Y_i(x)$.
3. Monotonicity: Either $X_i(0) \geq X_i(1)$ for all i or $X_i(0) \leq X_i(1)$ for all i. The instrument level affects the treatment level in weakly the same direction for all individuals. Individuals who are induced be be treated (i.e., those for whom $X_i(0) \neq X_i(1)$) are known as *compliers*.

Assumption (1) implies that the reduced form regression and the first stage regression are *identified*, meaning that we can interpret the coefficients on Z_i as causal. Assumption (2) implies that the effect of Z_i on Y_i is entirely mediated by X_i. Note that although assumption (2) is called an exclusion restriction, it is different from the exclusion restriction in Section 2 (which we also called the instrument exogeneity assumption). The exclusion restriction in Section 2 required that $cov(Z_i, U_i) = 0$. Here, assumption (2) is insufficient for that requirement. Assumptions (1) and (2) together guarantee that $cov(Z_i, U_i) = 0$.

Assumption (3), the monotonicity assumption, is key. It states that if Z_i increases X_i for any one individual, then for no individual does Z_i decrease X_i. In the quarter of birth example, monotonicity implies that being born in the fourth quarter rather than the first quarter either increases or has no effect on educational attainment. In a policy experiment with eligibility randomization, monotonicity implies that eligibility does not prevent any individual from participating in the program. Lecture Note 10 assumed that ineligibles could

not participate in the program, in which case monotonicity is guaranteed.

Under these assumptions, the TSLS estimator converges in probability to an estimand called the *local average treatment effect* ($LATE$). The $LATE$ is the average treatment effect among compliers:

$$\hat{\beta}_1^{TSLS} \xrightarrow{p} LATE = E[Y_i(1) - Y_i(0)|X_i(0) \neq X_i(1)]$$

The $LATE$ is closely related to the TOT in a randomized experiment. In particular, both the $LATE$ and the TOT measure the average treatment effect among individuals who were induced to be treated. When ineligibles cannot participate in the program, individuals who were induced to be treated *are* the treated population. For the $LATE$, we allow for the existence of treated individuals who would have been treated even in the absence of the experiment. (We call such individuals *always-takers*.) In this case, individuals who were induced to be treated are a subset of the treated population. But in the absence of always-takers, the $LATE$ and the TOT are the same.

We can generalize the local average treatment effect to non-binary instruments and non-binary endogenous variables, as well as multiple instruments and multiple endogenous variables. When Z_i is non-binary, we define a series of local average treatment effects, each for a different pair of instrument values:

$$LATE_{z_1,z_2} = E[Y_i(1) - Y_i(0)|X_i(z_1) = 1, X_i(z_2) = 0]$$

The TSLS estimator converges in probability to a weighted average of these local average treatment effects. When X_i is also non-binary, the TSLS estimator again converges to a particular weighted average treatment effect. The result is easiest to see through the lens of the first and second stage regressions:

$$\begin{aligned} X_i &= \pi_{0i} + \pi_{1i} Z_i + V_i \\ Y_i &= \beta_{0i} + \beta_{1i} X_i + U_i \end{aligned}$$

Two-stage least squares leads to:

$$\hat{\beta}_1^{TSLS} \xrightarrow{p} \frac{E\left[\beta_{1i}\pi_{1i}\right]}{E\left[\pi_{1i}\right]}$$

which is a weighted average of β_{1i}, giving more weight to individuals whose X_i was more affected by the instrument. We can derive similar results for multiple instruments and multiple endogenous variables, but those results require matrix algebra. The intuition is the same throughout. With heterogeneous causal effects, instrumental variables techniques uncover a weighted average of the causal effects that gives more weight to individuals who are more sensitive to the instrument.

The distinction between the $LATE$ and the ATE can be important. Consider the two instruments we have discussed for estimating the returns to schooling. The quarter of birth instrument affects schooling because it *forces* potential dropouts to stay in school. Proximity to college affects schooling because it *allows* eager students to go to college. The effects of additional schooling on these two groups of individuals are likely to be quite different. (Of course, a year of college may also have different effects from a year of high school.) In general, when you encounter an instrumental variables result, you should always think carefully about who the compliers are.

LECTURE NOTE 12: REGRESSION DISCONTINUITY DESIGNS

1 Introduction

Regression discontinuity (RD) designs take advantage of precise rules governing treatment assignment. The setup involves a *running variable* X_i and a *cutoff* or *threshold* c, such that the probability of receiving some treatment T_i changes discontinuously as X_i crosses c. *Sharp* RDs deal with the case in which an individual is treated if and only if $X_i \geq c$. *Fuzzy* RDs deal with the case in which individuals below the cutoff can access the treatment, but individuals above the cutoff have greater access. We will discover that fuzzy RDs can be seen as an application of instrumental variables methods.

2 Sharp Regression Discontinuity Designs

To develop our intuition for sharp RDs, we return to the potential outcomes setup of Lecture Note 10. Let $Y_i(t)$ be the potential outcome for treatment level t. Here, we consider a binary treatment $t \in \{0, 1\}$, such that the realized outcome Y_i can be expressed as follows:

$$Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0) = \begin{cases} Y_i(0) & \text{if } T_i = 0 \\ Y_i(1) & \text{if } T_i = 1 \end{cases}$$

where T_i is a dummy for treatment status. The following threshold rule determines T_i:

$$T_i = \begin{cases} 1 & \text{if } X_i \geq c \\ 0 & \text{if } X_i < c \end{cases}$$

This rule expresses the idea that an individual i is treated if and only if X_i exceeds the cutoff.

To use the threshold rule to identify the effect of T_i on Y_i we make the following continuity assumption:

$$E[Y_i(0)|X_i = x] \quad \text{and} \quad E[Y_i(1)|X_i = x] \quad \text{are continuous in } x.$$

The conditional expectations of the potential outcomes are continuous in X_i. As a result, were it not for the change in treatment status at c, average outcomes would change continuously at c. Then the discontinuity in

the conditional expectation of Y_i given X_i at c is can be interpreted as an average treatment effect:

$$\begin{aligned}
\alpha_{SRD} &= \lim_{x \downarrow c} E[Y_i | X_i = x] - \lim_{x \uparrow c} E[Y_i | X_i = x] \\
&= \lim_{x \downarrow c} E[Y_i(1) | X_i = x] - \lim_{x \uparrow c} E[Y_i(0) | X_i = x] \\
&= E[Y_i(1) - Y_i(0) | X_i = c]
\end{aligned}$$

That is to say, the discontinuity in the conditional expectation of Y_i given X_i is equal to the average treatment effect at the cutoff. This result is conceptually similar to the TOT and the $LATE$; we are identifying an average treatment effect among compliers.

For illustration, we list a few applications of this research design. First, many university admissions or academic honors criteria use a cutoff score or grade point average. Suppose we are studying such a case, and we want to know the effect of being admitted on later earnings. The discontinuity in the conditional expectation of earnings at the cutoff score is the average effect for the marginal admitted person. Second, elections use a cutoff rule. Suppose we want to know the effect of a Democratic mayor on a city's fiscal outcomes. Assuming that each election has a single Democratic candidate and a single Republican candidate (an assumption that is easily relaxed), we can look at the discontinuity in the conditional expectation of fiscal outcomes at the Democratic vote share of 50%. That discontinuity is the effect of electing a Democratic mayor in a close election. Third, many social programs use cutoff rules based on age. For instance, in the U.S., the legal drinking age is 21. Suppose we want to know the effect of being of legal drinking age on motor vehicle fatalities. If we assume that no other eligibility criteria change discontinuously at 21, then the discontinuity in motor vehicle fatality rates at age 21 has exactly that interpretation.

In all these cases, a key assumption is that outcomes would not change discontinuously in the absence of the threshold rule. As in past lecture notes, we may be a bit uncomfortable with this arcane statistical assumption that bears no obvious relation to reality. Lee (2008, also reviewed in Lee and Lemieux 2010) has suggested a more intuitive framework that motivates the continuity assumption. The details of his framework are beyond the scope of this course, but the basic idea is that if individuals have imperfect control over X_i, then we can think of RD as being based on local random assignment around $X_i = c$. Lee's framework applies well to the first two examples above. A student can influence her test score by studying, but some component of her final score will be random. Similarly, a politician can influence the distribution of votes by campaigning, but in a democratic system with a sufficiently large electorate, some component of the final vote distribution will be random. In both cases, assignment to either side of c can be treated as random, so that individuals with X_i just below c are similar to individuals with X_i just above c. However, the third example above does not fit as cleanly into Lee's framework because age is deterministic (i.e., it has no random component). For

the age discontinuity, the original continuity assumption is more natural than the local random assignment framework. In a few paragraphs, we will see that Lee's framework suggests ways to check the validity of an RD design.

To implement the sharp RD design, we will use the following regression specification:

$$\begin{aligned} Y_i &= \alpha T_i + f(X_i) + U_i \\ &= \alpha 1[X_i \geq c] + f(X_i) + U_i \end{aligned}$$

where $1[X_i \geq c]$ is a dummy variable that equals 1 if and only if $X_i \geq c$, and $f(X_i)$ is a continuous function of X_i. A major issue concerns how we approximate the function $f(X_i)$. Two approaches are common: (1) fit a global polynomial using all the data, and (2) fit a local linear regression only using data near the cutoff. The global polynomial approach involves running the regression:

$$Y_i = \alpha T_i + \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_K X_i^K + \beta_{K+1} T_i \cdot X_i + \beta_{K+2} T_i \cdot X_i^2 + \cdots + \beta_{K+K} T_i \cdot X_i^K + U_i$$

where K is the order of the polynomial. Usually, researchers use a cubic ($K = 3$) or a quartic ($K = 4$) polynomial. By interacting each of the polynomial terms with T_i, we allow the shape of the polynomial to differ above and below the cutoff. The local linear approach involves running the regression:

$$Y_i = \alpha T_i + \beta_0 + \beta_1 X_i + \beta_2 T_i \cdot X_i + U_i \quad \text{for} \quad X_i \in [c - h, c + h]$$

where h is called the *bandwidth*. Basically, we estimate the regression only using data within a window of the cutoff. By interacting X_i with T_i, we allow the slope of the regression line to be different to the right and left of the cutoff. Econometricians have developed algorithms to determine the optimal bandwidth h, but those algorithms are beyond the scope of this course, and in any case, one should always check the robustness of the results to a wide range of bandwidths. In the literature, you will occasionally see discussions about the choice of the *kernel* for the local linear regression. A *kernel* weights data points differently based on how close they are to the cutoff. When we run OLS on the sample close to the cutoff, we are using the *rectangular*kernel, which weights all data points equally. This approach is quickly becoming the norm in the RD literature.

Lee's framework suggests some useful ways to check the validity of an RD design:

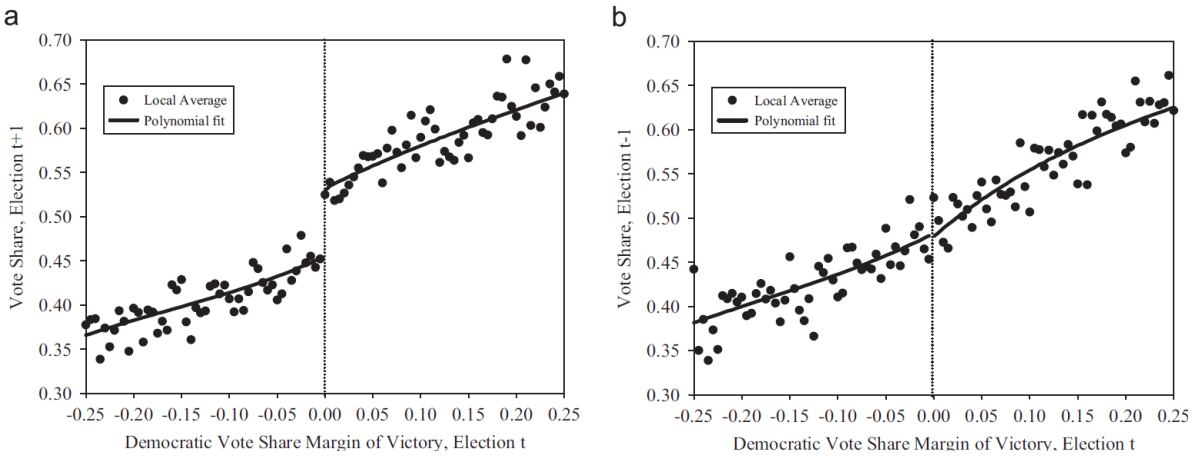
1. If local random assignment holds, then the density of X_i should in most cases be continuous at c. For example, if we observe a discontinuity in the test score distribution at an admissions cutoff, then we might conclude that at least some individuals have *perfect* control over their test scores, which would violate local random assignment. Similarly, if Democrats are disproportionately likely to win close

elections, we might be concerned about post-election manipulation of the vote returns, which would again violate local random assignment. McCrary (2008) provides a technical way to test whether the density of the running variable is discontinuous at c, but that method is beyond the scope of this course. Often, a simple histogram of the running variable is instructive enough. When plotting the histogram, one should always make sure to have separate bins above and below the cutoff. If the bars just above and below the cutoff look very different, the RD may not be valid.

2. Local random assignment also implied that the conditional expectations of predetermined variables will be continuous at c. We can test this implication by running an RD with a predetermined variable as the dependent variable. If predetermined variables change discontinuously at the cutoff, we might be concerned about a violation of local random assignment.
3. Similar to the logic in (2), our RD estimates should be robust to the inclusion of the predetermined variables as controls.

All three approaches are useful checks regardless of whether the local random assignment framework is natural for the RD design under consideration. (So they are also useful for deterministic running variables like age.)

RD designs are quite powerful because they can by illustrated graphically. Usually, researchers will plot estimates of the conditional expectation function, along with a series of local means for bins of X_i. For example, Lee (2008) examines the political party incumbency advantage in U.S. congressional elections by estimating the effect of a Democratic victory on the Democratic vote share in the next election. His running variable is the Democratic margin of victory, or the Democratic vote share minus the vote share of the best-performing non-Democrat. A Democrat wins the election if and only if the Democratic margin of victory is greater than zero. In graph (a) below, Lee plots the conditional expectation for the Democratic vote share in the next election. The observed discontinuity represents the effect of a Democratic victory in a close election. In graph (b), Lee plots the conditional expectation for the Democratic vote share in the *last* election. Since the last election outcome is predetermined, we expect to see no discontinuity, and indeed, we do not.



3 Fuzzy Regression Discontinuity Designs

Sometimes a threshold rule is not binding, so that some individuals with $X_i < c$ might have $T_i = 1$ and some individuals with $X_i \geq c$ might have $T_i = 0$. In this case, we can think of the threshold as affecting the *probability* of treatment:

$$\lim_{x \downarrow c} Pr[T_i = 1 | X_i = x] > \lim_{x \uparrow c} Pr[T_i = 1 | X_i = x]$$

In the university admissions example, the sharp RD identifies the effect of being *admitted*. But we might want to know the effect of *attending* the university, and some admitted students might not attend. This situation is similar to an eligibility experiment with non-compliance. Because the threshold rule does not change the probability of attending from 0 to 1, we call the estimation strategy a *fuzzy* RD.

To estimate a meaningful average causal effect using fuzzy RD, we make the following monotonicity assumption for potential treatment status, $T_i(x)$:

$$T_i(x) \text{ is non-decreasing in } x \text{ at } x = c.$$

This assumption is similar to the monotonicity condition in Lecture Note 11. It says that if crossing from below to above the cutoff increases T_i for any individual, then for no individual does it decrease T_i. We refer to individuals induced into treatment when they cross c as *compliers*. Compliers are characterized by:

$$\lim_{x \downarrow c} T_i(x) = 1 \quad \text{and} \quad \lim_{x \uparrow c} T_i(x) = 0$$

We can then obtain a local average treatment effect by dividing the discontinuity in Y_i by the discontinuity in X_i. This procedure is intuitively similar to dividing the ITT by the compliance rate in an eligibility experiment and to dividing the reduced form coefficient by the first stage coefficient in a more general IV setting. Specifically, we have:

$$\begin{aligned} \alpha_{FRD} &= \frac{\lim_{x \downarrow c} E[Y_i | X_i = x] - \lim_{x \uparrow c} E[Y_i | X_i = x]}{\lim_{x \downarrow c} E[T_i | X_i = x] - \lim_{x \uparrow c} E[T_i | X_i = x]} \\ &= E[Y_i(1) - Y_i(0) | X_i = c \text{ and } i \text{ is a complier}] \end{aligned}$$

As with IV, we can take two approaches: (1) estimate the numerator and the denominator separately and then divide, and (2) run two-stage least squares (TSLS). For the TSLS option, the first stage would be:

$$T_i = \pi 1[X_i \geq c] + g(X_i) + V_i$$

where $g(X_i)$ is a continuous function of X_i. We would then predict $\hat{T}_i$ and run the second stage:

$$Y_i = \alpha \hat{T}_i + f(X_i) + U_i$$

where $f(X_i)$ is a continuous function of X_i. We *always* use the same technique to estimate $g(X_i)$ and $f(X_i)$.

4 External Validity and the RD Design

Both sharp and fuzzy RD designs are only able to estimate average treatment effects among compliers who are exactly at the cutoff. This subpopulation is very specific. In fact, if X_i is continuous, then technically, no individual in our sample will have X_i exactly equal to c. So in a sense, RD is estimating an average treatment effect for a subpopulation that does not exist. In David Lee's local random assignment setup, it is possible to reframe the RD estimator. Rather than viewing it as the average effect *at* $X_i = c$, we can instead interpret it as a weighted average treatment effect for the population, where observations are weighted by their probabilities of being close to the cutoff. However, while this reframing does allow us to estimate an average treatment effect for a real population, the weighting remains very specific.

The specificity of the RD estimator is important for interpretation. In the test score example, RD measures the average treatment effect among individuals who marginally passed the test. In the election example, RD measure the average treatment effect in close elections. These subpopulations may have different treatment effects from the general population.