Welcome to Economics 20

What is Econometrics?

Why study Econometrics?

- Rare in economics (and many other areas without labs!) to have experimental data
- Need to use nonexperimental, or observational, data to make inferences
- Important to be able to apply economic theory to real world data

Why study Econometrics?

- An empirical analysis uses data to test a theory or to estimate a relationship
- A formal economic model can be tested
- Theory may be ambiguous as to the effect of some policy change – can use econometrics to evaluate the program

Types of Data – Cross Sectional

- Cross-sectional data is a random sample
- Each observation is a new individual, firm, etc. with information at a point in time
- If the data is not a random sample, we have a sample-selection problem

Types of Data – Panel

- Can pool random cross sections and treat similar to a normal cross section. Will just need to account for time differences.
- Can follow the same random individual observations over time – known as panel data or longitudinal data

Types of Data – Time Series

- Time series data has a separate observation for each time period – e.g. stock prices
- Since not a random sample, different problems to consider
- Trends and seasonality will be important

The Question of Causality

- Simply establishing a relationship between variables is rarely sufficient
- Want to the effect to be considered causal
- If we've truly controlled for enough other variables, then the estimated ceteris paribus effect can often be considered to be causal
- Can be difficult to establish causality

Example: Returns to Education

- A model of human capital investment implies getting more education should lead to higher earnings
- In the simplest case, this implies an equation like

$$Earnings = \beta_0 + \beta_1 education + u$$

Example: (continued)

- The estimate of β_1, is the return to education, but can it be considered causal?
- While the error term, u, includes other factors affecting earnings, want to control for as much as possible
- Some things are still unobserved, which can be problematic

The Simple Regression Model

$y = \beta_0 + \beta_1 x + u$

Some Terminology

- In the simple linear regression model, where $y = \beta_0 + \beta_1 x + u$, we typically refer to y as the
  - Dependent Variable, or
  - Left-Hand Side Variable, or
  - Explained Variable, or
  - Regressand

Some Terminology, cont.

- In the simple linear regression of y on x, we typically refer to x as the
  - Independent Variable, or
  - Right-Hand Side Variable, or
  - Explanatory Variable, or
  - Regressor, or
  - Covariate, or
  - Control Variables

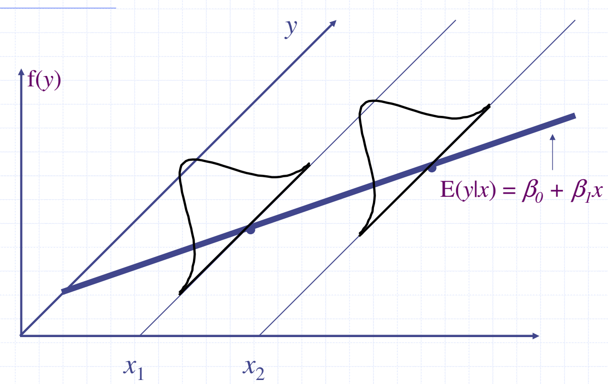
A Simple Assumption

- The average value of u, the error term, in the population is 0. That is,
- $E(u) = 0$
- This is not a restrictive assumption, since we can always use β_0 to normalize $E(u)$ to 0

Zero Conditional Mean

- We need to make a crucial assumption about how u and x are related
- We want it to be the case that knowing something about x does not give us any information about u, so that they are completely unrelated. That is, that
- $E(u|x) = E(u) = 0$, which implies
- $E(y|x) = \beta_0 + \beta_1 x$

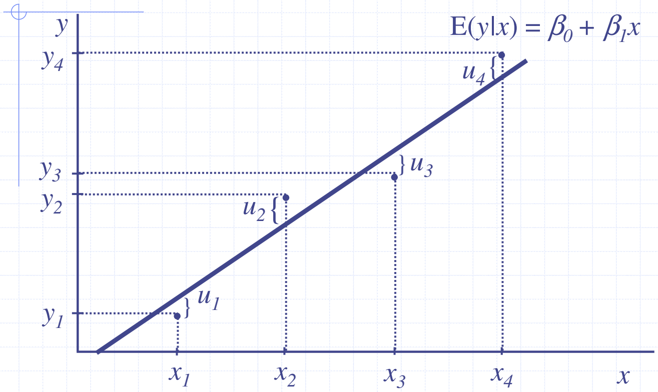
E(y|x) as a linear function of x, where for any x the distribution of y is centered about E(y|x)



Ordinary Least Squares

- Basic idea of regression is to estimate the population parameters from a sample
- Let $\{(x_i,y_i): i=1, \ldots,n\}$ denote a random sample of size n from the population
- For each observation in this sample, it will be the case that
- $y_i = \beta_0 + \beta_1 x_i + u_i$

Population regression line, sample data points and the associated error terms



Deriving OLS Estimates

- To derive the OLS estimates we need to realize that our main assumption of $E(u|x) = E(u) = 0$ also implies that
- $\mathrm{Cov}(x,u) = E(xu) = 0$
- Why? Remember from basic probability that $\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$

Deriving OLS continued

- We can write our 2 restrictions just in terms of x, y, β_0 and β_1 , since $u = y - \beta_0 - \beta_1 x$
- $E(y - \beta_0 - \beta_1 x) = 0$
- $E[x(y - \beta_0 - \beta_1 x)] = 0$
- These are called moment restrictions

Deriving OLS using M.O.M.

- The method of moments approach to estimation implies imposing the population moment restrictions on the sample moments
- What does this mean? Recall that for E(X), the mean of a population distribution, a sample estimator of E(X) is simply the arithmetic mean of the sample

More Derivation of OLS

- We want to choose values of the parameters that will ensure that the sample versions of our moment restrictions are true
- The sample versions are as follows:

$$n^{-1}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

$$n^{-1}\sum_{i=1}^{n} x_i\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

More Derivation of OLS

- Given the definition of a sample mean, and properties of summation, we can rewrite the first condition as follows

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

or

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

More Derivation of OLS

$$\sum_{i=1}^{n} x_i \left(y_i - \left(\bar{y} - \hat{\beta}_1 \bar{x} \right) - \hat{\beta}_1 x_i \right) = 0$$

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

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

So the OLS estimated slope is

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

provided that $\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 > 0$

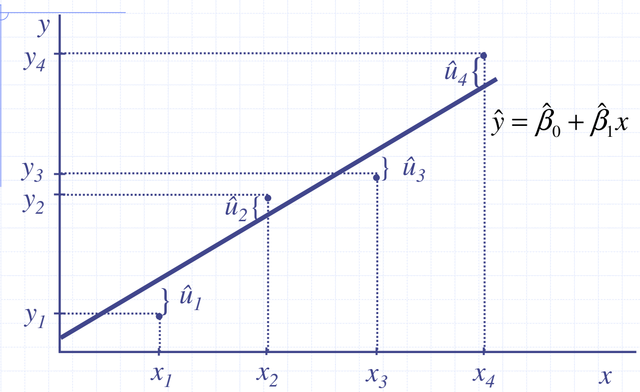
Summary of OLS slope estimate

- The slope estimate is the sample covariance between x and y divided by the sample variance of x
- If x and y are positively correlated, the slope will be positive
- If x and y are negatively correlated, the slope will be negative
- Only need x to vary in our sample

More OLS

- Intuitively, OLS is fitting a line through the sample points such that the sum of squared residuals is as small as possible, hence the term least squares
- The residual, $\hat{u}$, is an estimate of the error term, u, and is the difference between the fitted line (sample regression function) and the sample point

Sample regression line, sample data points and the associated estimated error terms



Alternate approach to derivation

- Given the intuitive idea of fitting a line, we can set up a formal minimization problem
- That is, we want to choose our parameters such that we minimize the following:

$$\sum_{i=1}^{n}(\hat{u}_i)^2 = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

Alternate approach, continued

- If one uses calculus to solve the minimization problem for the two parameters you obtain the following first order conditions, which are the same as we obtained before, multiplied by n

$$\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

$$\sum_{i=1}^{n} x_i\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

Algebraic Properties of OLS

- The sum of the OLS residuals is zero
- Thus, the sample average of the OLS residuals is zero as well
- The sample covariance between the regressors and the OLS residuals is zero
- The OLS regression line always goes through the mean of the sample

Algebraic Properties (precise)

$$\sum_{i=1}^{n}\hat{u}_i = 0 \text{ and thus, } \frac{\sum_{i=1}^{n}\hat{u}_i}{n} = 0$$

$$\sum_{i=1}^{n} x_i\hat{u}_i = 0$$

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

More terminology

We can think of each observation as being made up of an explained part, and an unexplained part,

$y_i = \hat{y}_i + \hat{u}_i$ We then define the following:

$\sum(y_i - \bar{y})^2$ is the total sum of squares (SST)

$\sum(\hat{y}_i - \bar{y})^2$ is the explained sum of squares (SSE)

$\sum\hat{u}_i^2$ is the residual sum of squares (SSR)

Then SST = SSE + SSR

Proof that SST = SSE + SSR

$$\sum(y_i - \bar{y})^2 = \sum\left[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})\right]^2$$

$$= \sum\left[\hat{u}_i + (\hat{y}_i - \bar{y})\right]^2$$

$$= \sum\hat{u}_i^2 + 2\sum\hat{u}_i(\hat{y}_i - \bar{y}) + \sum(\hat{y}_i - \bar{y})^2$$

$$= \text{SSR} + 2\sum\hat{u}_i(\hat{y}_i - \bar{y}) + \text{SSE}$$

and we know that $\sum\hat{u}_i(\hat{y}_i - \bar{y}) = 0$

Goodness-of-Fit

- How do we think about how well our sample regression line fits our sample data?
- Can compute the fraction of the total sum of squares (SST) that is explained by the model, call this the R-squared of regression
- $R^2 = SSE/SST = 1 – SSR/SST$

Using Stata for OLS regressions

- Now that we've derived the formula for calculating the OLS estimates of our parameters, you'll be happy to know you don't have to compute them by hand
- Regressions in Stata are very simple, to run the regression of y on x, just type
- reg y x

Unbiasedness of OLS

- Assume the population model is linear in parameters as $y = \beta_0 + \beta_1 x + u$
- Assume we can use a random sample of size n, $\{(x_i, y_i): i=1, 2, \ldots, n\}$, from the population model. Thus we can write the sample model $y_i = \beta_0 + \beta_1 x_i + u_i$
- Assume $E(u|x) = 0$ and thus $E(u_i|x_i) = 0$
- Assume there is variation in the x_i

Unbiasedness of OLS (cont)

- In order to think about unbiasedness, we need to rewrite our estimator in terms of the population parameter
- Start with a simple rewrite of the formula as

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) y_i}{s_x^2}, \text{ where}$$

$$s_x^2 \equiv \sum (x_i - \bar{x})^2$$

Unbiasedness of OLS (cont)

$$\sum (x_i - \bar{x}) y_i = \sum (x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i) =$$

$$\sum (x_i - \bar{x})\beta_0 + \sum (x_i - \bar{x})\beta_1 x_i$$

$$+ \sum (x_i - \bar{x}) u_i =$$

$$\beta_0 \sum (x_i - \bar{x}) + \beta_1 \sum (x_i - \bar{x}) x_i$$

$$+ \sum (x_i - \bar{x}) u_i$$

Unbiasedness of OLS (cont)

$$\sum (x_i - \bar{x}) = 0,$$

$$\sum (x_i - \bar{x}) x_i = \sum (x_i - \bar{x})^2$$

so, the numerator can be rewritten as

$$\beta_1 s_x^2 + \sum (x_i - \bar{x}) u_i, \text{ and thus}$$

$$\hat{\beta}_1 = \beta_1 + \frac{\sum (x_i - \bar{x}) u_i}{s_x^2}$$

Unbiasedness of OLS (cont)

let $d_i = (x_i - \bar{x})$, so that

$$\hat{\beta}_i = \beta_1 + \left(1 \big/ s_x^2 \right) \sum d_i u_i \text{, then}$$

$$E(\hat{\beta}_1) = \beta_1 + \left(1 \big/ s_x^2 \right) \sum d_i E(u_i) = \beta_1$$

Unbiasedness Summary

- The OLS estimates of β_1 and β_0 are unbiased
- Proof of unbiasedness depends on our 4 assumptions – if any assumption fails, then OLS is not necessarily unbiased
- Remember unbiasedness is a description of the estimator – in a given sample we may be "near" or "far" from the true parameter

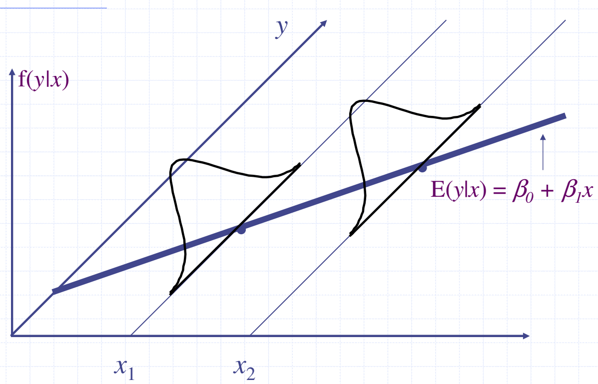
Variance of the OLS Estimators

- Now we know that the sampling distribution of our estimate is centered around the true parameter
- Want to think about how spread out this distribution is
- Much easier to think about this variance under an additional assumption, so
- Assume Var(u|x) = σ^2 (Homoskedasticity)

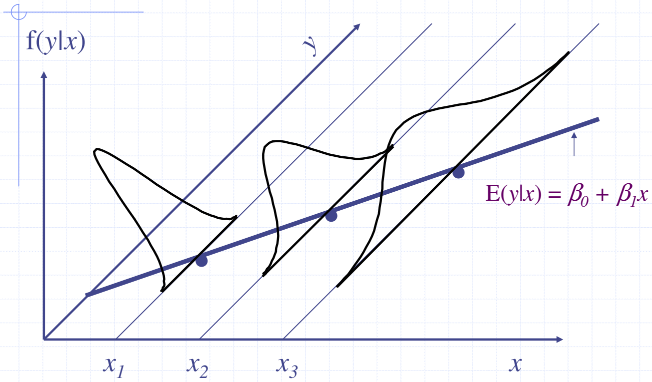
Variance of OLS (cont)

- Var(u|x) = E(u^2|x)-[E(u|x)]^2
- E(u|x) = 0, so σ^2 = E(u^2|x) = E(u^2) = Var(u)
- Thus σ^2 is also the unconditional variance, called the error variance
- σ, the square root of the error variance is called the standard deviation of the error
- Can say: E(y|x)=β_0 + $\beta_1 x$ and Var(y|x) = σ^2

Homoskedastic Case



Heteroskedastic Case



Variance of OLS (cont)

$$Var\left(\hat{\beta}_1\right) = Var\left(\beta_1 + \left(1/s_x^2\right)\sum d_i u_i\right) =$$

$$\left(1/s_x^2\right)^2 Var\left(\sum d_i u_i\right) = \left(1/s_x^2\right)^2 \sum d_i^2 Var(u_i)$$

$$= \left(1/s_x^2\right)^2 \sum d_i^2 \sigma^2 = \sigma^2 \left(1/s_x^2\right)^2 \sum d_i^2 =$$

$$\sigma^2 \left(1/s_x^2\right)^2 s_x^2 = \sigma^2/s_x^2 = Var\left(\hat{\beta}_1\right)$$

Variance of OLS Summary

- The larger the error variance, σ^2, the larger the variance of the slope estimate
- The larger the variability in the x_i, the smaller the variance of the slope estimate
- As a result, a larger sample size should decrease the variance of the slope estimate
- Problem that the error variance is unknown

Estimating the Error Variance

- We don't know what the error variance, σ^2, is, because we don't observe the errors, u_i
- What we observe are the residuals, $\hat{u}_i$
- We can use the residuals to form an estimate of the error variance

Error Variance Estimate (cont)

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

$$= (\beta_0 + \beta_1 x_i + u_i) - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

$$= u_i - \left(\hat{\beta}_0 - \beta_0\right) - \left(\hat{\beta}_1 - \beta_1\right)$$

Then, an unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{(n-2)} \sum \hat{u}_i^2 = SSR/(n-2)$$

Error Variance Estimate (cont)

$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ = Standard error of the regression

recall that $\mathrm{sd}\left(\hat{\beta}\right) = \sigma/s_x$

if we substitute $\hat{\sigma}$ for σ then we have

the standard error of $\hat{\beta}_1$,

$$\mathrm{se}\left(\hat{\beta}_1\right) = \hat{\sigma} / \left(\sum (x_i - \bar{x})^2\right)^{1/2}$$

Multiple Regression Analysis

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + u$

1. Estimation

Parallels with Simple Regression

- β_0 is still the intercept
- β_1 to β_k all called slope parameters
- u is still the error term (or disturbance)
- Still need to make a zero conditional mean assumption, so now assume that
- $E(u|x_1,x_2, \ldots,x_k) = 0$
- Still minimizing the sum of squared residuals, so have k+1 first order conditions

Interpreting Multiple Regression

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$, so

$\Delta\hat{y} = \Delta\hat{\beta}_1 x_1 + \Delta\hat{\beta}_2 x_2 + \ldots + \Delta\hat{\beta}_k x_k$,

so holding $x_2,\ldots,x_k$ fixed implies that

$\Delta\hat{y} = \Delta\hat{\beta}_1 x_1$, that is each β has

a *ceteris paribus* interpretation

A "Partialling Out" Interpretation

Consider the case where $k = 2$, i.e.

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$, then

$\hat{\beta}_1 = \left(\sum \hat{r}_{i1} y_i\right) / \sum \hat{r}_{i1}^2$, where $\hat{r}_{i1}$ are

the residuals from the estimated

regression $\hat{x}_1 = \hat{\gamma}_0 + \hat{\gamma}_2 \hat{x}_2$

"Partialling Out" continued

- Previous equation implies that regressing y on x_1 and x_2 gives same effect of x_1 as regressing y on residuals from a regression of x_1 on x_2
- This means only the part of x_{i1} that is uncorrelated with x_{i2} are being related to y_i so we're estimating the effect of x_1 on y after x_2 has been "partialled out"

Simple vs Multiple Reg Estimate

Compare the simple regression $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$

with the multiple regression $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

Generally, $\tilde{\beta}_1 \neq \hat{\beta}_1$ unless:

$\hat{\beta}_2 = 0$ (i.e. no partial effect of x_2) OR

x_1 and x_2 are uncorrelated in the sample

Goodness-of-Fit

We can think of each observation as being made up of an explained part, and an unexplained part, $y_i = \hat{y}_i + \hat{u}_i$ We then define the following:

$\sum (y_i - \bar{y})^2$ is the total sum of squares (SST)

$\sum (\hat{y}_i - \bar{y})^2$ is the explained sum of squares (SSE)

$\sum \hat{u}_i^2$ is the residual sum of squares (SSR)

Then SST = SSE + SSR

Goodness-of-Fit (continued)

- How do we think about how well our sample regression line fits our sample data?
- Can compute the fraction of the total sum of squares (SST) that is explained by the model, call this the R-squared of regression
- R^2 = SSE/SST = 1 – SSR/SST

Goodness-of-Fit (continued)

We can also think of R^2 as being equal to the squared correlation coefficient between the actual y_i and the values $\hat{y}_i$

$$R^2 = \frac{\left(\sum (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right)^2}{\left(\sum (y_i - \bar{y})^2\right)\left(\sum (\hat{y}_i - \bar{\hat{y}})^2\right)}$$

More about *R*-squared

- R^2 can never decrease when another independent variable is added to a regression, and usually will increase
- Because R^2 will usually increase with the number of independent variables, it is not a good way to compare models

Assumptions for Unbiasedness

- Population model is linear in parameters: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$
- We can use a random sample of size n, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i): i=1, 2, \ldots, n\}$, from the population model, so that the sample model is $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i$
- $E(u|x_1, x_2, \ldots x_k) = 0$, implying that all of the explanatory variables are exogenous
- None of the x's is constant, and there are no exact linear relationships among them

Too Many or Too Few Variables

- What happens if we include variables in our specification that don't belong?
- There is no effect on our parameter estimate, and OLS remains unbiased
- What if we exclude a variable from our specification that does belong?
- OLS will usually be biased

Omitted Variable Bias

Suppose the true model is given as $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$, but we estimate $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + u$, then

$$\tilde{\beta}_1 = \frac{\sum (x_{i1} - \bar{x}_1) y_i}{\sum (x_{i1} - \bar{x}_1)^2}$$

Omitted Variable Bias (cont)

Recall the true model, so that

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$, so the numerator becomes

$$\sum (x_{i1} - \bar{x}_1)(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i) =$$

$$\beta_1 \sum (x_{i1} - \bar{x}_1)^2 + \beta_2 \sum (x_{i1} - \bar{x}_1) x_{i2} + \sum (x_{i1} - \bar{x}_1) u_i$$

Omitted Variable Bias (cont)

$$\tilde{\beta} = \beta_1 + \beta_2 \frac{\sum (x_{i1} - \bar{x}_1) x_{i2}}{\sum \left((x_{i1} - \bar{x}_1)^2 \right)} + \frac{\sum (x_{i1} - \bar{x}_1) u_i}{\sum \left((x_{i1} - \bar{x}_1)^2 \right)}$$

since $\mathrm{E}(u_i) = 0$, taking expectations we have

$$E\left(\tilde{\beta}_1\right) = \beta_1 + \beta_2 \frac{\sum (x_{i1} - \bar{x}_1) x_{i2}}{\sum \left((x_{i1} - \bar{x}_1)^2 \right)}$$

Omitted Variable Bias (cont)

Consider the regression of x_2 on x_1

$\tilde{x}_2 = \tilde{\delta}_0 + \tilde{\delta}_1 x_1$ then $\tilde{\delta}_1 = \dfrac{\sum (x_{i1} - \bar{x}_1) x_{i2}}{\sum \left((x_{i1} - \bar{x}_1)^2 \right)}$

so $E\left(\tilde{\beta}_1\right) = \beta_1 + \beta_2 \tilde{\delta}_1$

Summary of Direction of Bias

	$\mathrm{Corr}(x_1, x_2) > 0$	$\mathrm{Corr}(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

Omitted Variable Bias Summary

- Two cases where bias is equal to zero
  - $\beta_2 = 0$, that is x_2 doesn't really belong in model
  - x_1 and x_2 are uncorrelated in the sample
- If correlation between x_2 , x_1 and x_2 , y is the same direction, bias will be positive
- If correlation between x_2 , x_1 and x_2 , y is the opposite direction, bias will be negative

The More General Case

- Technically, can only sign the bias for the more general case if all of the included *x*'s are uncorrelated
- Typically, then, we work through the bias assuming the *x*'s are uncorrelated, as a useful guide even if this assumption is not strictly true

Variance of the OLS Estimators

- Now we know that the sampling distribution of our estimate is centered around the true parameter
- Want to think about how spread out this distribution is
- Much easier to think about this variance under an additional assumption, so
- Assume $\mathrm{Var}(u|x_1, x_2,\ldots, x_k) = \sigma^2$ (Homoskedasticity)

Variance of OLS (cont)

- Let $\boldsymbol{x}$ stand for $(x_1, x_2,\ldots x_k)$
- Assuming that $\mathrm{Var}(u|\boldsymbol{x}) = \sigma^2$ also implies that $\mathrm{Var}(y|\boldsymbol{x}) = \sigma^2$
- The 4 assumptions for unbiasedness, plus this homoskedasticity assumption are known as the Gauss-Markov assumptions

Variance of OLS (cont)

Given the Gauss - Markov Assumptions

$$Var\left(\hat{\beta}_j\right) = \frac{\sigma^2}{SST_j\left(1 - R_j^2\right)}, \text{ where}$$

$SST_j = \sum\left(x_{ij} - \bar{x}_j\right)^2$ and R_j^2 is the R^2 from regressing x_j on all other x's

Components of OLS Variances

- The error variance: a larger σ^2 implies a larger variance for the OLS estimators
- The total sample variation: a larger SST_j implies a smaller variance for the estimators
- Linear relationships among the independent variables: a larger R_j^2 implies a larger variance for the estimators

Misspecified Models

Consider again the misspecified model

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1, \text{ so that } Var\left(\tilde{\beta}_1\right) = \frac{\sigma^2}{SST_1}$$

Thus, $Var\left(\tilde{\beta}_1\right) < Var\left(\hat{\beta}_1\right)$ unless x_1 and x_2 are uncorrelated, then they're the same

Misspecified Models (cont)

- While the variance of the estimator is smaller for the misspecified model, unless $\beta_2 = 0$ the misspecified model is biased
- As the sample size grows, the variance of each estimator shrinks to zero, making the variance difference less important

Estimating the Error Variance

- We don't know what the error variance, σ^2, is, because we don't observe the errors, u_i
- What we observe are the residuals, $\hat{u}_i$
- We can use the residuals to form an estimate of the error variance

Error Variance Estimate (cont)

$$\hat{\sigma}^2 = \left(\sum \hat{u}_i^2\right) / (n - k - 1) \equiv SSR/df$$

$$\text{thus, } se\left(\hat{\beta}_j\right) = \hat{\sigma} / \left[SST_j\left(1 - R_j^2\right)\right]^{1/2}$$

- $df = n - (k + 1)$, or $df = n - k - 1$
- df (i.e. degrees of freedom) is the (number of observations) – (number of estimated parameters)

The Gauss-Markov Theorem

- Given our 5 Gauss-Markov Assumptions it can be shown that OLS is "BLUE"
- Best
- Linear
- Unbiased
- Estimator
- Thus, if the assumptions hold, use OLS

Multiple Regression Analysis

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + u$
- 2. Inference

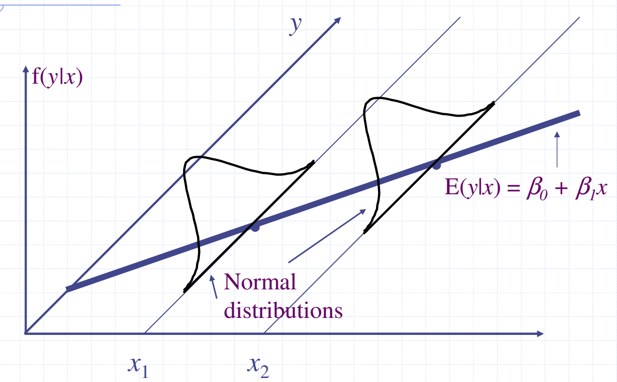
Assumptions of the Classical Linear Model (CLM)

- So far, we know that given the Gauss-Markov assumptions, OLS is BLUE,
- In order to do classical hypothesis testing, we need to add another assumption (beyond the Gauss-Markov assumptions)
- Assume that u is independent of $x_1, x_2, \ldots, x_k$ and u is normally distributed with zero mean and variance σ^2: $u \sim \text{Normal}(0,\sigma^2)$

CLM Assumptions (cont)

- Under CLM, OLS is not only BLUE, but is the minimum variance unbiased estimator
- We can summarize the population assumptions of CLM as follows
- $y|\mathbf{x} \sim \text{Normal}(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \sigma^2)$
- While for now we just assume normality, clear that sometimes not the case
- Large samples will let us drop normality

The homoskedastic normal distribution with a single explanatory variable



Normal Sampling Distributions

Under the CLM assumptions, conditional on the sample values of the independent variables

$$\hat{\beta}_j \sim \text{Normal}\left[\beta_j, Var\left(\hat{\beta}_j\right)\right], \text{ so that}$$

$$\left(\hat{\beta}_j - \beta_j\right) \Big/ sd\left(\hat{\beta}_j\right) \sim \text{Normal}(0,1)$$

$\hat{\beta}_j$ is distributed normally because it is a linear combination of the errors

The *t* Test

Under the CLM assumptions

$$\left(\hat{\beta}_j - \beta_j\right) \Big/ se\left(\hat{\beta}_j\right) \sim t_{n-k-1}$$

Note this is a t distribution (vs normal) because we have to estimate σ^2 by $\hat{\sigma}^2$

Note the degrees of freedom : $n-k-1$

The t Test (cont)

- Knowing the sampling distribution for the standardized estimator allows us to carry out hypothesis tests
- Start with a null hypothesis
- For example, H_0: $\beta_j = 0$
- If accept null, then accept that x_j has no effect on y, controlling for other x's

The t Test (cont)

To perform our test we first need to form

"the" t statistic for $\hat{\beta}_j : t_{\hat{\beta}_j} \equiv \hat{\beta}_j / se(\hat{\beta}_j)$

We will then use our t statistic along with a rejection rule to determine whether to accept the null hypothesis, H_0

t Test: One-Sided Alternatives

- Besides our null, H_0, we need an alternative hypothesis, H_1, and a significance level
- H_1 may be one-sided, or two-sided
- H_1: $\beta_j > 0$ and H_1: $\beta_j < 0$ are one-sided
- H_1: $\beta_j \neq 0$ is a two-sided alternative
- If we want to have only a 5% probability of rejecting H_0 if it is really true, then we say our significance level is 5%

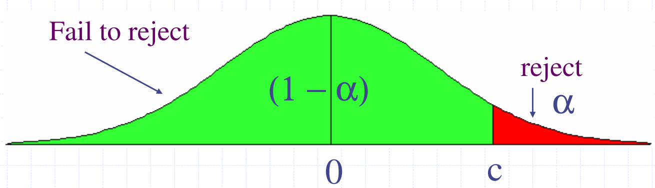
One-Sided Alternatives (cont)

- Having picked a significance level, α, we look up the $(1 - \alpha)^{th}$ percentile in a t distribution with $n - k - 1$ df and call this c, the critical value
- We can reject the null hypothesis if the t statistic is greater than the critical value
- If the t statistic is less than the critical value then we fail to reject the null

One-Sided Alternatives (cont)

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i$$

H_0: $\beta_j = 0$ H_1: $\beta_j > 0$



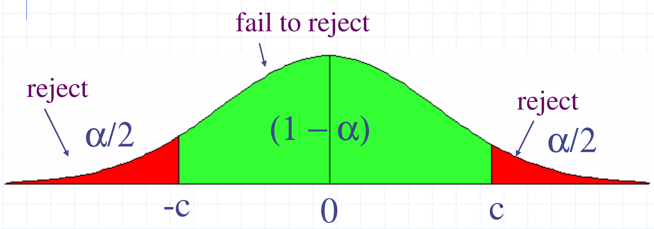
One-sided vs Two-sided

- Because the t distribution is symmetric, testing H_1: $\beta_j < 0$ is straightforward. The critical value is just the negative of before
- We can reject the null if the t statistic $< -c$, and if the t statistic $>$ than $-c$ then we fail to reject the null
- For a two-sided test, we set the critical value based on $\alpha/2$ and reject H_1: $\beta_j \neq 0$ if the absolute value of the t statistic $> c$

Two-Sided Alternatives

$y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik} + u_i$

$H_0: \beta_j = 0$ $H_1: \beta_j > 0$



Summary for H_0: $\beta_j = 0$

- Unless otherwise stated, the alternative is assumed to be two-sided
- If we reject the null, we typically say "x_j is statistically significant at the α % level"
- If we fail to reject the null, we typically say "x_j is statistically insignificant at the α % level"

Testing other hypotheses

- A more general form of the *t* statistic recognizes that we may want to test something like H_0: $\beta_j = a_j$
- In this case, the appropriate *t* statistic is

$$t = \left(\hat{\beta}_j - a_j\right) \Big/ se\left(\hat{\beta}_j\right), \text{ where}$$

$$a_j = 0 \text{ for the standard test}$$

Confidence Intervals

- Another way to use classical statistical testing is to construct a confidence interval using the same critical value as was used for a two-sided test
- A (1 - α) % confidence interval is defined as

$$\hat{\beta}_j \pm c \bullet se\left(\hat{\beta}_j\right), \text{ where c is the } \left(1 - \frac{\alpha}{2}\right) \text{ percentile}$$

in a t_{n-k-1} distribution

Computing *p*-values for *t* tests

- An alternative to the classical approach is to ask, "what is the smallest significance level at which the null would be rejected?"
- So, compute the *t* statistic, and then look up what percentile it is in the appropriate *t* distribution – this is the *p*-value
- *p*-value is the probability we would observe the *t* statistic we did, if the null were true

Stata and *p*-values, *t* tests, etc.

- Most computer packages will compute the *p*-value for you, assuming a two-sided test
- If you really want a one-sided alternative, just divide the two-sided *p*-value by 2
- Stata provides the *t* statistic, *p*-value, and 95% confidence interval for H_0: $\beta_j = 0$ for you, in columns labeled "t", "P > |t|" and "[95% Conf. Interval]", respectively

Testing a Linear Combination

- Suppose instead of testing whether β_1 is equal to a constant, you want to test if it is equal to another parameter, that is $H_0 : \beta_1 = \beta_2$
- Use same basic procedure for forming a t statistic

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{se\left(\hat{\beta}_1 - \hat{\beta}_2\right)}$$

Testing Linear Combo (cont)

Since

$$se\left(\hat{\beta}_1 - \hat{\beta}_2\right) = \sqrt{Var\left(\hat{\beta}_1 - \hat{\beta}_2\right)}, \text{ then}$$

$$Var\left(\hat{\beta}_1 - \hat{\beta}_2\right) = Var\left(\hat{\beta}_1\right) + Var\left(\hat{\beta}_2\right) - 2Cov\left(\hat{\beta}_1, \hat{\beta}_2\right)$$

$$se\left(\hat{\beta}_1 - \hat{\beta}_2\right) = \left\{\left[se\left(\hat{\beta}_1\right)\right]^2 + \left[se\left(\hat{\beta}_2\right)\right]^2 - 2s_{12}\right\}^{1/2}$$

where s_{12} is an estimate of $Cov\left(\hat{\beta}_1, \hat{\beta}_2\right)$

Testing a Linear Combo (cont)

- So, to use formula, need s_{12}, which standard output does not have
- Many packages will have an option to get it, or will just perform the test for you
- In Stata, after reg y x1 x2 … xk you would type test x1 = x2 to get a *p*-value for the test
- More generally, you can always restate the problem to get the test you want

Example:

- Suppose you are interested in the effect of campaign expenditures on outcomes
- Model is $voteA = \beta_0 + \beta_1 \log(expendA) + \beta_2 \log(expendB) + \beta_3 prtystrA + u$
- H_0: $\beta_1 = - \beta_2$, or H_0: $\theta_1 = \beta_1 + \beta_2 = 0$
- $\beta_1 = \theta_1 - \beta_2$, so substitute in and rearrange $\Rightarrow$ $voteA = \beta_0 + \theta_1 \log(expendA) + \beta_2 \log(expendB - expendA) + \beta_3 prtystrA + u$

Example (cont):

- This is the same model as originally, but now you get a standard error for $\beta_1 - \beta_2 = \theta_1$ directly from the basic regression
- Any linear combination of parameters could be tested in a similar manner
- Other examples of hypotheses about a single linear combination of parameters:
  - $\beta_1 = 1 + \beta_2$; $\beta_1 = 5\beta_2$; $\beta_1 = -1/2\beta_2$; etc

Multiple Linear Restrictions

- Everything we've done so far has involved testing a single linear restriction, (e.g. $\beta_1 = 0$ or $\beta_1 = \beta_2$)
- However, we may want to jointly test multiple hypotheses about our parameters
- A typical example is testing "exclusion restrictions" – we want to know if a group of parameters are all equal to zero

Testing Exclusion Restrictions

- Now the null hypothesis might be something like H_0: $\beta_{k-q+1} = 0, \ldots, \beta_k = 0$
- The alternative is just H_1: H_0 is not true
- Can't just check each *t* statistic separately, because we want to know if the *q* parameters are jointly significant at a given level – it is possible for none to be individually significant at that level

Exclusion Restrictions (cont)

- To do the test we need to estimate the "restricted model" without $x_{k-q+1}, \ldots, x_k$ included, as well as the "unrestricted model" with all *x*'s included
- Intuitively, we want to know if the change in SSR is big enough to warrant inclusion of $x_{k-q+1}, \ldots, x_k$

$$F \equiv \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}, \text{ where}$$

r is restricted and ur is unrestricted

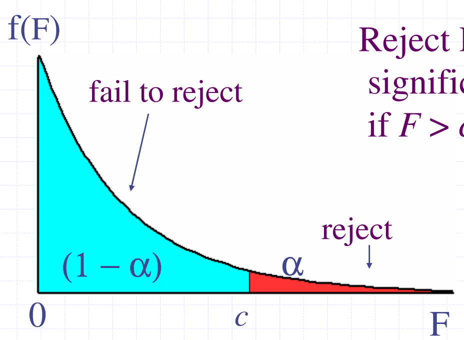
The *F* statistic

- The *F* statistic is always positive, since the SSR from the restricted model can't be less than the SSR from the unrestricted
- Essentially the *F* statistic is measuring the relative increase in SSR when moving from the unrestricted to restricted model
- q = number of restrictions, or $df_r - df_{ur}$
- $n - k - 1 = df_{ur}$

The *F* statistic (cont)

- To decide if the increase in SSR when we move to a restricted model is "big enough" to reject the exclusions, we need to know about the sampling distribution of our *F* stat
- Not surprisingly, $F \sim F_{q,n-k-1}$, where q is referred to as the numerator degrees of freedom and $n - k - 1$ as the denominator degrees of freedom

The *F* statistic (cont)



Reject H_0 at α significance level if $F > c$

The R^2 form of the *F* statistic

- Because the SSR's may be large and unwieldy, an alternative form of the formula is useful
- We use the fact that $SSR = SST(1 - R^2)$ for any regression, so can substitute in for SSR_u and SSR_{ur}

$$F \equiv \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/(n-k-1)}, \text{ where again}$$

r is restricted and ur is unrestricted

Overall Significance

- A special case of exclusion restrictions is to test H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$
- Since the R^2 from a model with only an intercept will be zero, the *F* statistic is simply

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

General Linear Restrictions

- The basic form of the *F* statistic will work for any set of linear restrictions
- First estimate the unrestricted model and then estimate the restricted model
- In each case, make note of the SSR
- Imposing the restrictions can be tricky – will likely have to redefine variables again

Example:

- Use same voting model as before
- Model is $voteA = \beta_0 + \beta_1\log(expendA) + \beta_2\log(expendB) + \beta_3 prtystrA + u$
- now null is H_0: $\beta_1 = 1$, $\beta_3 = 0$
- Substituting in the restrictions: $voteA = \beta_0 + \log(expendA) + \beta_2\log(expendB) + u$, so
- Use $voteA - \log(expendA) = \beta_0 + \beta_2\log(expendB) + u$ as restricted model

F Statistic Summary

- Just as with *t* statistics, p-values can be calculated by looking up the percentile in the appropriate *F* distribution
- Stata will do this by entering: display fprob(q, $n - k - 1$, F), where the appropriate values of F, q, and $n - k - 1$ are used
- If only one exclusion is being tested, then $F = t^2$, and the *p*-values will be the same

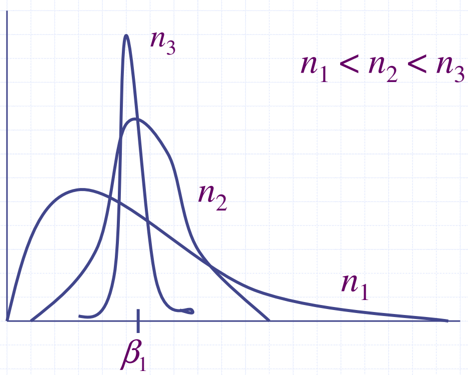
Multiple Regression Analysis

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + u$
- 3. Asymptotic Properties

Consistency

- Under the Gauss-Markov assumptions OLS is BLUE, but in other cases it won't always be possible to find unbiased estimators
- In those cases, we may settle for estimators that are consistent, meaning as $n \to \infty$, the distribution of the estimator collapses to the parameter value

Sampling Distributions as $n \uparrow$



Consistency of OLS

- Under the Gauss-Markov assumptions, the OLS estimator is consistent (and unbiased)
- Consistency can be proved for the simple regression case in a manner similar to the proof of unbiasedness
- Will need to take probability limit (plim) to establish consistency

Proving Consistency

$$\hat{\beta}_1 = \left(\sum (x_{i1} - \bar{x}_1) y_i\right) \Big/ \left(\sum (x_{i1} - \bar{x}_1)^2\right)$$
$$= \beta_1 + \left(n^{-1} \sum (x_{i1} - \bar{x}_1) u_i\right) \Big/ \left(n^{-1} \sum (x_{i1} - \bar{x}_1)^2\right)$$
$$\text{plim}\, \hat{\beta}_1 = \beta_1 + Cov(x_1, u) / Var(x_1) = \beta_1$$
$$\text{because } Cov(x_1, u) = 0$$

A Weaker Assumption

- For unbiasedness, we assumed a zero conditional mean – $E(u|x_1, x_2, \ldots, x_k) = 0$
- For consistency, we can have the weaker assumption of zero mean and zero correlation – $E(u) = 0$ and $Cov(x_j, u) = 0$, for $j = 1, 2, \ldots, k$
- Without this assumption, OLS will be biased and inconsistent!

Deriving the Inconsistency

- Just as we could derive the omitted variable bias earlier, now we want to think about the inconsistency, or asymptotic bias, in this case

$$\text{True model}: y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + v$$

$$\text{You think}: y = \beta_0 + \beta_1 x_1 + u, \text{ so that}$$

$$u = \beta_2 x_2 + v \text{ and, } \operatorname{plim}\tilde{\beta}_1 = \beta_1 + \beta_2 \delta$$

$$\text{where } \delta = Cov(x_1, x_2)/Var(x_1)$$

Asymptotic Bias (cont)

- So, thinking about the direction of the asymptotic bias is just like thinking about the direction of bias for an omitted variable
- Main difference is that asymptotic bias uses the population variance and covariance, while bias uses the sample counterparts
- Remember, inconsistency is a large sample problem – it doesn't go away as add data

Large Sample Inference

- Recall that under the CLM assumptions, the sampling distributions are normal, so we could derive *t* and *F* distributions for testing
- This exact normality was due to assuming the population error distribution was normal
- This assumption of normal errors implied that the distribution of *y*, given the *x*'s, was normal as well

Large Sample Inference (cont)

- Easy to come up with examples for which this exact normality assumption will fail
- Any clearly skewed variable, like wages, arrests, savings, etc. can't be normal, since a normal distribution is symmetric
- Normality assumption not needed to conclude OLS is BLUE, only for inference

Central Limit Theorem

- Based on the central limit theorem, we can show that OLS estimators are asymptotically normal
- Asymptotic Normality implies that $P(Z<z) \rightarrow \Phi(z)$ as $n \rightarrow \infty$, or $P(Z<z) \approx \Phi(z)$
- The central limit theorem states that the standardized average of any population with mean μ and variance σ^2 is asymptotically ~N(0,1), or

$$Z = \frac{\bar{Y} - \mu_Y}{\sigma / \sqrt{n}} \overset{a}{\sim} N(0,1)$$

Asymptotic Normality

$$\text{Under the Gauss-Markov assumptions,}$$

$$\text{(i) } \sqrt{n}\left(\hat{\beta}_j - \beta_j\right) \overset{a}{\sim} \text{Normal}\left(0, \sigma^2 / a_j^2\right),$$

$$\text{where } a_j^2 = \operatorname{plim}\left(n^{-1} \sum \hat{r}_{ij}^2\right)$$

$$\text{(ii) } \hat{\sigma}^2 \text{ is a consistent estimator of } \sigma^2$$

$$\text{(iii) } \left(\hat{\beta}_j - \beta_j\right) / se\left(\hat{\beta}_j\right) \overset{a}{\sim} \text{Normal}(0,1)$$

Asymptotic Normality (cont)

- Because the t distribution approaches the normal distribution for large df, we can also say that

$$\left(\hat{\beta}_j - \beta_j\right) / se\left(\hat{\beta}_j\right) \overset{a}{\sim} t_{n-k-1}$$

- Note that while we no longer need to assume normality with a large sample, we do still need homoskedasticity

Asymptotic Standard Errors

- If u is not normally distributed, we sometimes will refer to the standard error as an asymptotic standard error, since

$$se\left(\hat{\beta}_j\right) = \sqrt{\frac{\hat{\sigma}^2}{SST_j\left(1 - R_j^2\right)}},$$

$$se\left(\hat{\beta}_j\right) \approx c_j / \sqrt{n}$$

- So, we can expect standard errors to shrink at a rate proportional to the inverse of $\sqrt{n}$

Lagrange Multiplier statistic

- Once we are using large samples and relying on asymptotic normality for inference, we can use more that t and F stats
- The Lagrange multiplier or LM statistic is an alternative for testing multiple exclusion restrictions
- Because the LM statistic uses an auxiliary regression it's sometimes called an nR^2 stat

LM Statistic (cont)

- Suppose we have a standard model, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + u$ and our null hypothesis is
- H_0: $\beta_{k-q+1} = 0, \ldots, \beta_k = 0$
- First, we just run the restricted model

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \ldots + \tilde{\beta}_{k-q} x_{k-q} + \tilde{u}$$

Now take the residuals, $\tilde{u}$, and regress

$\tilde{u}$ on $x_1, x_2, \ldots, x_k$ (i.e. *all* the variables)

$LM = nR_u^2$, where R_u^2 is from this reg

LM Statistic (cont)

$LM \overset{a}{\sim} \chi_q^2$, so can choose a critical value, c, from a χ_q^2 distribution, or just calculate a p-value for χ_q^2

- With a large sample, the result from an F test and from an LM test should be similar
- Unlike the F test and t test for one exclusion, the LM test and F test will not be identical

Asymptotic Efficiency

- Estimators besides OLS will be consistent
- However, under the Gauss-Markov assumptions, the OLS estimators will have the smallest asymptotic variances
- We say that OLS is asymptotically efficient
- Important to remember our assumptions though, if not homoskedastic, not true

Multiple Regression Analysis

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + u$
- 4. Further Issues

Redefining Variables

- Changing the scale of the y variable will lead to a corresponding change in the scale of the coefficients and standard errors, so no change in the significance or interpretation
- Changing the scale of one x variable will lead to a change in the scale of that coefficient and standard error, so no change in the significance or interpretation

Beta Coefficients

- Occasional you'll see reference to a "standardized coefficient" or "beta coefficient" which has a specific meaning
- Idea is to replace y and each x variable with a standardized version – i.e. subtract mean and divide by standard deviation
- Coefficient reflects standard deviation of y for a one standard deviation change in x

Functional Form

- OLS can be used for relationships that are not strictly linear in x and y by using nonlinear functions of x and y – will still be linear in the parameters
- Can take the natural log of x, y or both
- Can use quadratic forms of x
- Can use interactions of x variables

Interpretation of Log Models

- If the model is $\ln(y) = \beta_0 + \beta_1 \ln(x) + u$
- β_1 is the elasticity of y with respect to x
- If the model is $\ln(y) = \beta_0 + \beta_1 x + u$
- β_1 is approximately the percentage change in y given a 1 unit change in x
- If the model is $y = \beta_0 + \beta_1 \ln(x) + u$
- β_1 is approximately the change in y for a 100 percent change in x

Why use log models?

- Log models are invariant to the scale of the variables since measuring percent changes
- They give a direct estimate of elasticity
- For models with $y > 0$, the conditional distribution is often heteroskedastic or skewed, while $\ln(y)$ is much less so
- The distribution of $\ln(y)$ is more narrow, limiting the effect of outliers

Some Rules of Thumb

- What types of variables are often used in log form?
- Dollar amounts that must be positive
- Very large variables, such as population
- What types of variables are often used in level form?
- Variables measured in years
- Variables that are a proportion or percent

Quadratic Models

- For a model of the form $y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$ we can't interpret β_1 alone as measuring the change in y with respect to x, we need to take into account β_2 as well, since

$$\Delta\hat{y} \approx \left(\hat{\beta}_1 + 2\hat{\beta}_2 x\right)\Delta x, \text{ so}$$

$$\frac{\Delta\hat{y}}{\Delta x} \approx \hat{\beta}_1 + 2\hat{\beta}_2 x$$

More on Quadratic Models

- Suppose that the coefficient on x is positive and the coefficient on x^2 is negative
- Then y is increasing in x at first, but will eventually turn around and be decreasing in x

For $\hat{\beta}_1 > 0$ and $\hat{\beta}_2 < 0$ the turning point will be at $x^* = \left|\hat{\beta}_1 / \left(2\hat{\beta}_2\right)\right|$

More on Quadratic Models

- Suppose that the coefficient on x is negative and the coefficient on x^2 is positive
- Then y is decreasing in x at first, but will eventually turn around and be increasing in x

For $\hat{\beta}_1 < 0$ and $\hat{\beta}_2 > 0$ the turning point will be at $x^* = \left|\hat{\beta}_1 / \left(2\hat{\beta}_2\right)\right|$, which is the same as when $\hat{\beta}_1 > 0$ and $\hat{\beta}_2 < 0$

Interaction Terms

- For a model of the form $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u$ we can't interpret β_1 alone as measuring the change in y with respect to x_1, we need to take into account β_3 as well, since

$$\frac{\Delta y}{\Delta x_1} = \beta_1 + \beta_3 x_2, \text{ so to summarize}$$

the effect of x_1 on y we typically evaluate the above at $\bar{x}_2$

Adjusted R-Squared

- Recall that the R^2 will always increase as more variables are added to the model
- The adjusted R^2 takes into account the number of variables in a model, and may decrease

$$\bar{R}^2 \equiv 1 - \frac{[SSR/(n-k-1)]}{[SST/(n-1)]}$$

$$= 1 - \frac{\hat{\sigma}^2}{[SST/(n-1)]}$$

Adjusted R-Squared (cont)

- It's easy to see that the adjusted R^2 is just $(1 - R^2)(n - 1) / (n - k - 1)$, but most packages will give you both R^2 and adj-R^2
- You can compare the fit of 2 models (with the same y) by comparing the adj-R^2
- You cannot use the adj-R^2 to compare models with different y's (e.g. y vs. $\ln(y)$)

Goodness of Fit

- Important not to fixate too much on adj-R^2 and lose sight of theory and common sense
- If economic theory clearly predicts a variable belongs, generally leave it in
- Don't want to include a variable that prohibits a sensible interpretation of the variable of interest – remember ceteris paribus interpretation of multiple regression

Standard Errors for Predictions

- Suppose we want to use our estimates to obtain a specific prediction?
- First, suppose that we want an estimate of $E(y|x_1=c_1,\ldots x_k=c_k) = \theta_0 = \beta_0+\beta_1 c_1+ \ldots + \beta_k c_k$
- This is easy to obtain by substituting the x's in our estimated model with c's , but what about a standard error?
- Really just a test of a linear combination

Predictions (cont)

- Can rewrite as $\beta_0 = \theta_0 - \beta_1 c_1 - \ldots - \beta_k c_k$
- Substitute in to obtain $y = \theta_0 + \beta_1 (x_1 - c_1) + \ldots + \beta_k (x_k - c_k) + u$
- So, if you regress y_i on $(x_{ij} - c_{ij})$ the intercept will give the predicted value and its standard error
- Note that the standard error will be smallest when the c's equal the means of the x's

Predictions (cont)

- This standard error for the expected value is not the same as a standard error for an outcome on y
- We need to also take into account the variance in the unobserved error. Let the prediction error be

$$\hat{e}^0 = y^0 - \hat{y}^0 = \left(\beta_0 + \beta_1 x_1^0 + \ldots + \beta_k x_k^0\right) + u^0 - \hat{y}_0$$

$$E\left(\hat{e}^0\right) = 0 \text{ and } Var\left(\hat{e}^0\right) = Var\left(\hat{y}^0\right) + Var\left(u^0\right) =$$

$$Var\left(\hat{y}^0\right) + \sigma^2 \text{, so } se\left(\hat{e}^0\right) = \left\{\left[se\left(\hat{y}^0\right)\right]^2 + \hat{\sigma}^2\right\}^{\frac{1}{2}}$$

Prediction interval

$$\hat{e}^0 / se\left(\hat{e}^0\right) \sim t_{n-k-1} \text{, so given that } \hat{e}^0 = y^0 - \hat{y}^0$$

we have a 95% prediction interval for y^0

$$\hat{y}^0 \pm t_{.025} \bullet se\left(\hat{e}^0\right)$$

- Usually the estimate of s^2 is much larger than the variance of the prediction, thus
- This prediction interval will be a lot wider than the simple confidence interval for the prediction

Residual Analysis

- Information can be obtained from looking at the residuals (i.e. predicted vs. observed)
- Example: Regress price of cars on characteristics – big negative residuals indicate a good deal
- Example: Regress average earnings for students from a school on student characteristics – big positive residuals indicate greatest value-added

Predicting *y* in a log model

- Simple exponentiation of the predicted ln(*y*) will underestimate the expected value of *y*
- Instead need to scale this up by an estimate of the expected value of exp(*u*)

$$E(\exp(u)) = \exp(\sigma/2) \text{ if } u \sim N(0,\sigma^2)$$

In this case can predict y as follows

$$\hat{y} = \exp(\hat{\sigma}^2/2)\exp(\hat{\ln y})$$

Predicting *y* in a log model

- If *u* is not normal, E(exp(*u*)) must be estimated using an auxiliary regression
- Create the exponentiation of the predicted ln(*y*), and regress *y* on it with no intercept
- The coefficient on this variable is the estimate of E(exp(*u*)) that can be used to scale up the exponentiation of the predicted ln(*y*) to obtain the predicted *y*

Comparing log and level models

- A by-product of the previous procedure is a method to compare a model in logs with one in levels.
- Take the fitted values from the auxiliary regression, and find the sample correlation between this and *y*
- Compare the R^2 from the levels regression with this correlation squared

Multiple Regression Analysis

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + u$
- 5. Dummy Variables

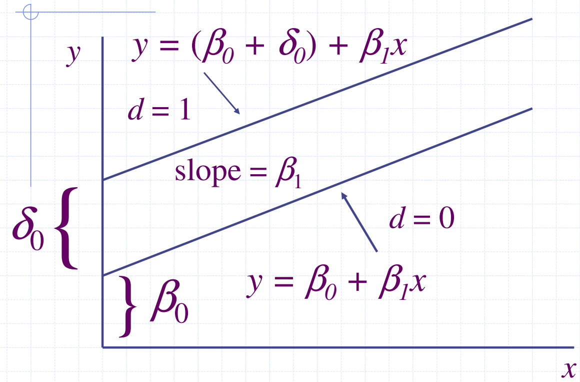
Dummy Variables

- A dummy variable is a variable that takes on the value 1 or 0
- Examples: male (= 1 if are male, 0 otherwise), south (= 1 if in the south, 0 otherwise), etc.
- Dummy variables are also called binary variables, for obvious reasons

A Dummy Independent Variable

- Consider a simple model with one continuous variable (x) and one dummy (d)
- $y = \beta_0 + \delta_0 d + \beta_1 x + u$
- This can be interpreted as an intercept shift
- If d = 0, then $y = \beta_0 + \beta_1 x + u$
- If d = 1, then $y = (\beta_0 + \delta_0) + \beta_1 x + u$
- The case of d = 0 is the base group

Example of $\delta_0 > 0$



Dummies for Multiple Categories

- We can use dummy variables to control for something with multiple categories
- Suppose everyone in your data is either a HS dropout, HS grad only, or college grad
- To compare HS and college grads to HS dropouts, include 2 dummy variables
- hsgrad = 1 if HS grad only, 0 otherwise; and colgrad = 1 if college grad, 0 otherwise

Multiple Categories (cont)

- Any categorical variable can be turned into a set of dummy variables
- Because the base group is represented by the intercept, if there are n categories there should be n – 1 dummy variables
- If there are a lot of categories, it may make sense to group some together
- Example: top 10 ranking, 11 – 25, etc.

Interactions Among Dummies

- Interacting dummy variables is like subdividing the group
- Example: have dummies for male, as well as hsgrad and colgrad
- Add male*hsgrad and male*colgrad, for a total of 5 dummy variables –> 6 categories
- Base group is female HS dropouts
- hsgrad is for female HS grads, colgrad is for female college grads
- The interactions reflect male HS grads and male college grads

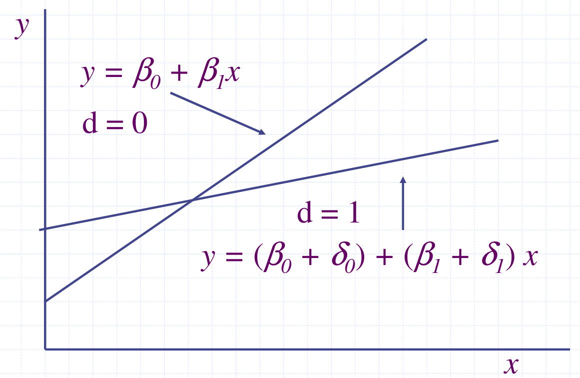
More on Dummy Interactions

- Formally, the model is $y = \beta_0 + \delta_1 male + \delta_2 hsgrad + \delta_3 colgrad + \delta_4 male*hsgrad + \delta_5 male*colgrad + \beta_1 x + u$, then, for example:
- If male = 0 and hsgrad = 0 and colgrad = 0
- $y = \beta_0 + \beta_1 x + u$
- If male = 0 and hsgrad = 1 and colgrad = 0
- $y = \beta_0 + \delta_2 hsgrad + \beta_1 x + u$
- If male = 1 and hsgrad = 0 and colgrad = 1
- $y = \beta_0 + \delta_1 male + \delta_3 colgrad + \delta_5 male*colgrad + \beta_1 x + u$

Other Interactions with Dummies

- Can also consider interacting a dummy variable, d, with a continuous variable, x
- $y = \beta_0 + \delta_1 d + \beta_1 x + \delta_2 d*x + u$
- If $d = 0$, then $y = \beta_0 + \beta_1 x + u$
- If $d = 1$, then $y = (\beta_0 + \delta_1) + (\beta_1 + \delta_2)\, x + u$
- This is interpreted as a change in the slope

Example of $\delta_0 > 0$ and $\delta_1 < 0$



Testing for Differences Across Groups

- Testing whether a regression function is different for one group versus another can be thought of as simply testing for the joint significance of the dummy and its interactions with all other x variables
- So, you can estimate the model with all the interactions and without and form an F statistic, but this could be unwieldy

The Chow Test

- Turns out you can compute the proper F statistic without running the unrestricted model with interactions with all k continuous variables
- If run the restricted model for group one and get SSR_1, then for group two and get SSR_2
- Run the restricted model for all to get SSR, then

$$F = \frac{\left[SSR - \left(SSR_1 + SSR_2\right)\right]}{SSR_1 + SSR_2} \bullet \frac{\left[n - 2(k+1)\right]}{k+1}$$

The Chow Test (continued)

- The Chow test is really just a simple *F* test for exclusion restrictions, but we've realized that $SSR_{ur} = SSR_1 + SSR_2$
- Note, we have $k + 1$ restrictions (each of the slope coefficients and the intercept)
- Note the unrestricted model would estimate 2 different intercepts and 2 different slope coefficients, so the df is $n - 2k - 2$

Linear Probability Model

- $P(y = 1|x) = E(y|x)$, when y is a binary variable, so we can write our model as
- $P(y = 1|x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$
- So, the interpretation of β_j is the change in the probability of success when x_j changes
- The predicted *y* is the predicted probability of success
- Potential problem that can be outside [0,1]

Linear Probability Model (cont)

- Even without predictions outside of [0,1], we may estimate effects that imply a change in *x* changes the probability by more than +1 or –1, so best to use changes near mean
- This model will violate assumption of homoskedasticity, so will affect inference
- Despite drawbacks, it's usually a good place to start when *y* is binary

Caveats on Program Evaluation

- A typical use of a dummy variable is when we are looking for a program effect
- For example, we may have individuals that received job training, or welfare, etc
- We need to remember that usually individuals choose whether to participate in a program, which may lead to a self-selection problem

Self-selection Problems

- If we can control for everything that is correlated with both participation and the outcome of interest then it's not a problem
- Often, though, there are unobservables that are correlated with participation
- In this case, the estimate of the program effect is biased, and we don't want to set policy based on it!

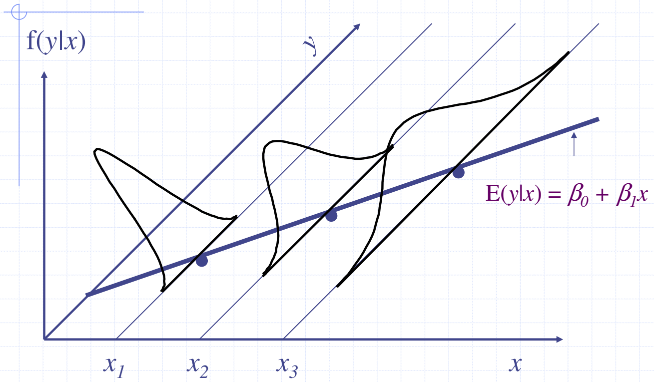
Multiple Regression Analysis

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + u$
- 6. Heteroskedasticity

What is Heteroskedasticity

- Recall the assumption of homoskedasticity implied that conditional on the explanatory variables, the variance of the unobserved error, *u*, was constant
- If this is not true, that is if the variance of *u* is different for different values of the *x*'s, then the errors are heteroskedastic
- Example: estimating returns to education and ability is unobservable, and think the variance in ability differs by educational attainment

Example of Heteroskedasticity



Why Worry About Heteroskedasticity?

- OLS is still unbiased and consistent, even if we do not assume homoskedasticity
- The standard errors of the estimates are biased if we have heteroskedasticity
- If the standard errors are biased, we can not use the usual *t* statistics or *F* statistics or *LM* statistics for drawing inferences

Variance with Heteroskedasticity

For the simple case, $\hat{\beta}_1 = \beta_1 + \frac{\sum (x_i - \bar{x}) u_i}{\sum (x_i - \bar{x})^2}$, so

$Var(\hat{\beta}_1) = \frac{\sum (x_i - \bar{x})^2 \sigma_i^2}{SST_x^2}$, where $SST_x = \sum (x_i - \bar{x})^2$

A valid estimator for this when $\sigma_i^2 \neq \sigma^2$ is

$\frac{\sum (x_i - \bar{x})^2 \hat{u}_i^2}{SST_x^2}$, where $\hat{u}_i$ are are the OLS residuals

Variance with Heteroskedasticity

For the general multiple regression model, a valid estimator of $Var(\hat{\beta}_j)$ with heteroskedasticity is

$\hat{Var}(\hat{\beta}_j) = \frac{\sum \hat{r}_{ij}^2 \hat{u}_i^2}{SST_j^2}$, where $\hat{r}_{ij}$ is the i^{th} residual from regressing x_j on all other independent variables, and SST_j is the sum of squared residuals from this regression

Robust Standard Errors

- Now that we have a consistent estimate of the variance, the square root can be used as a standard error for inference
- Typically call these robust standard errors
- Sometimes the estimated variance is corrected for degrees of freedom by multiplying by $n/(n - k - 1)$
- As $n \rightarrow \infty$ it's all the same, though

Robust Standard Errors (cont)

- Important to remember that these robust standard errors only have asymptotic justification – with small sample sizes t statistics formed with robust standard errors will not have a distribution close to the t, and inferences will not be correct
- In Stata, robust standard errors are easily obtained using the robust option of reg

A Robust *LM* Statistic

- Run OLS on the restricted model and save the residuals $\breve{u}$
- Regress each of the excluded variables on all of the included variables (q different regressions) and save each set of residuals $\check{r}_1, \check{r}_2, \ldots, \check{r}_q$
- Regress a variable defined to be = 1 on $\check{r}_1 \breve{u}, \check{r}_2 \breve{u}, \ldots, \check{r}_q \breve{u}$, with <u>no</u> intercept
- The LM statistic is $n - SSR_1$, where SSR_1 is the sum of squared residuals from this final regression

Testing for Heteroskedasticity

- Essentially want to test H_0: $Var(u|x_1, x_2, \ldots, x_k) = \sigma^2$, which is equivalent to H_0: $E(u^2|x_1, x_2, \ldots, x_k) = E(u^2) = \sigma^2$
- If assume the relationship between u^2 and x_j will be linear, can test as a linear restriction
- So, for $u^2 = \delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k + v$) this means testing H_0: $\delta_1 = \delta_2 = \ldots = \delta_k = 0$

The Breusch-Pagan Test

- Don't observe the error, but can estimate it with the residuals from the OLS regression
- After regressing the residuals squared on all of the x's, can use the R^2 to form an F or LM test
- The F statistic is just the reported F statistic for overall significance of the regression, $F = [R^2/k]/[(1 - R^2)/(n - k - 1)]$, which is distributed $F_{k,\, n - k - 1}$
- The LM statistic is $LM = nR^2$, which is distributed χ^2_k

The White Test

- The Breusch-Pagan test will detect any linear forms of heteroskedasticity
- The White test allows for nonlinearities by using squares and crossproducts of all the x's
- Still just using an F or LM to test whether all the x_j, x_j^2, and $x_j x_h$ are jointly significant
- This can get to be unwieldy pretty quickly

Alternate form of the White test

- Consider that the fitted values from OLS, $\hat{y}$, are a function of all the x's
- Thus, $\hat{y}^2$ will be a function of the squares and crossproducts and $\hat{y}$ and $\hat{y}^2$ can proxy for all of the x_j, x_j^2, and $x_j x_h$, so
- Regress the residuals squared on $\hat{y}$ and $\hat{y}^2$ and use the R^2 to form an F or LM statistic
- Note only testing for 2 restrictions now

Weighted Least Squares

- While it's always possible to estimate robust standard errors for OLS estimates, if we know something about the specific form of the heteroskedasticity, we can obtain more efficient estimates than OLS
- The basic idea is going to be to transform the model into one that has homoskedastic errors – called weighted least squares

Case of form being known up to a multiplicative constant

- Suppose the heteroskedasticity can be modeled as $\mathrm{Var}(u|\boldsymbol{x}) = \sigma^2 h(\boldsymbol{x})$, where the trick is to figure out what $h(\boldsymbol{x}) \equiv h_i$ looks like
- $E(u_i/\sqrt{h_i}|\boldsymbol{x}) = 0$, because h_i is only a function of $\boldsymbol{x}$, and $\mathrm{Var}(u_i/\sqrt{h_i}|\boldsymbol{x}) = \sigma^2$, because we know $\mathrm{Var}(u|\boldsymbol{x}) = \sigma^2 h_i$
- So, if we divided our whole equation by $\sqrt{h_i}$ we would have a model where the error is homoskedastic

Generalized Least Squares

- Estimating the transformed equation by OLS is an example of generalized least squares (GLS)
- GLS will be BLUE in this case
- GLS is a weighted least squares (WLS) procedure where each squared residual is weighted by the inverse of $\mathrm{Var}(u_i|\boldsymbol{x}_i)$

Weighted Least Squares

- While it is intuitive to see why performing OLS on a transformed equation is appropriate, it can be tedious to do the transformation
- Weighted least squares is a way of getting the same thing, without the transformation
- Idea is to minimize the weighted sum of squares (weighted by $1/h_i$)

More on WLS

- WLS is great if we know what $\mathrm{Var}(u_i|\boldsymbol{x}_i)$ looks like
- In most cases, won't know form of heteroskedasticity
- Example where do is if data is aggregated, but model is individual level
- Want to weight each aggregate observation by the inverse of the number of individuals

Feasible GLS

- More typical is the case where you don't know the form of the heteroskedasticity
- In this case, you need to estimate $h(\boldsymbol{x}_i)$
- Typically, we start with the assumption of a fairly flexible model, such as
- $\mathrm{Var}(u|\boldsymbol{x}) = \sigma^2\exp(\delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k)$
- Since we don't know the δ, must estimate

Feasible GLS (continued)

- Our assumption implies that $u^2 = \sigma^2\exp(\delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k)v$
- Where $E(v|\boldsymbol{x}) = 1$, then if $E(v) = 1$
- $\ln(u^2) = \alpha_0 + \delta_1 x_1 + \ldots + \delta_k x_k + e$
- Where $E(e) = 1$ and e is independent of $\boldsymbol{x}$
- Now, we know that $\hat{u}$ is an estimate of u, so we can estimate this by OLS

Feasible GLS (continued)

- Now, an estimate of h is obtained as $\hat{h} = \exp(\hat{g})$, and the inverse of this is our weight
- So, what did we do?
- Run the original OLS model, save the residuals, $\hat{u}$, square them and take the log
- Regress $\ln(\hat{u}^2)$ on all of the independent variables and get the fitted values, $\hat{g}$
- Do WLS using $1/\exp(\hat{g})$ as the weight

WLS Wrapup

- When doing F tests with WLS, form the weights from the unrestricted model and use those weights to do WLS on the restricted model as well as the unrestricted model
- Remember we are using WLS just for efficiency – OLS is still unbiased & consistent
- Estimates will still be different due to sampling error, but if they are very different then it's likely that some other Gauss-Markov assumption is false

Multiple Regression Analysis

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + u$
- 7. Specification and Data Problems

Functional Form

- We've seen that a linear regression can really fit nonlinear relationships
- Can use logs on RHS, LHS or both
- Can use quadratic forms of x's
- Can use interactions of x's
- How do we know if we've gotten the right functional form for our model?

Functional Form (continued)

- First, use economic theory to guide you
- Think about the interpretation
- Does it make more sense for x to affect y in percentage (use logs) or absolute terms?
- Does it make more sense for the derivative of x_1 to vary with x_1 (quadratic) or with x_2 (interactions) or to be fixed?

Functional Form (continued)

- We already know how to test joint exclusion restrictions to see if higher order terms or interactions belong in the model
- It can be tedious to add and test extra terms, plus may find a square term matters when really using logs would be even better
- A test of functional form is Ramsey's regression specification error test (RESET)

Ramsey's RESET

- RESET relies on a trick similar to the special form of the White test
- Instead of adding functions of the x's directly, we add and test functions of $\hat{y}$
- So, estimate $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_1 \hat{y}^3 + error$ and test
- H_0: $\delta_1 = 0$, $\delta_2 = 0$ using $F \sim F_{2,n-k-3}$ or $LM \sim \chi^2_2$

Nonnested Alternative Tests

- If the models have the same dependent variables, but nonnested x's could still just make a giant model with the x's from both and test joint exclusion restrictions that lead to one model or the other
- An alternative, the Davidson-MacKinnon test, uses $\hat{y}$ from one model as regressor in the second model and tests for significance

Nonnested Alternatives (cont)

- More difficult if one model uses y and the other uses $\ln(y)$
- Can follow same basic logic and transform predicted $\ln(y)$ to get $\hat{y}$ for the second step
- In any case, Davidson-MacKinnon test may reject neither or both models rather than clearly preferring one specification

Proxy Variables

- What if model is misspecified because no data is available on an important x variable?
- It may be possible to avoid omitted variable bias by using a proxy variable
- A proxy variable must be related to the unobservable variable – for example: $x_3^* = \delta_0 + \delta_3 x_3 + v_3$, where * implies unobserved
- Now suppose we just substitute x_3 for x_3^*

Proxy Variables (continued)

- What do we need for for this solution to give us consistent estimates of β_1 and β_2?
- $E(x_3^* \mid x_1, x_2, x_3) = E(x_3^* \mid x_3) = \delta_0 + \delta_3 x_3$
- That is, u is uncorrelated with x_1, x_2 and x_3^* and v_3 is uncorrelated with x_1, x_2 and x_3
- So really running $y = (\beta_0 + \beta_3\delta_0) + \beta_1 x_1 + \beta_2 x_2 + \beta_3\delta_3 x_3 + (u + \beta_3 v_3)$ and have just redefined intercept, error term x_3 coefficient

Proxy Variables (continued)

- Without out assumptions, can end up with biased estimates
- Say $x_3^* = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + v_3$
- Then really running $y = (\beta_0 + \beta_3\delta_0) + (\beta_1 + \beta_3\delta_1)\, x_1 + (\beta_2 + \beta_3\delta_2)\, x_2 + \beta_3\delta_3 x_3 + (u + \beta_3 v_3)$
- Bias will depend on signs of β_3 and δ_j
- This bias may still be smaller than omitted variable bias, though

Lagged Dependent Variables

- What if there are unobserved variables, and you can't find reasonable proxy variables?
- May be possible to include a lagged dependent variable to account for omitted variables that contribute to both past and current levels of y
- Obviously, you must think past and current y are related for this to make sense

Measurement Error

- Sometimes we have the variable we want, but we think it is measured with error
- Examples: A survey asks how many hours did you work over the last year, or how many weeks you used child care when your child was young
- Measurement error in y different from measurement error in x

Measurement Error in a Dependent Variable

- Define measurement error as $e_0 = y - y^*$
- Thus, really estimating $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u + e_0$
- When will OLS produce unbiased results?
- If e_0 and x_j, u are uncorrelated is unbiased
- If $E(e_0) \neq 0$ then β_0 will be biased, though
- While unbiased, larger variances than with no measurement error

Measurement Error in an Explanatory Variable

- Define measurement error as $e_1 = x_1 - x_1^*$
- Assume $E(e_1) = 0$, $E(y| x_1^*, x_1) = E(y| x_1^*)$
- Really estimating $y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1)$
- The effect of measurement error on OLS estimates depends on our assumption about the correlation between e_1 and x_1
- Suppose $Cov(x_1, e_1) = 0$
- OLS remains unbiased, variances larger

Measurement Error in an Explanatory Variable (cont)

- Suppose $Cov(x_1^*, e_1) = 0$, known as the classical errors-in-variables assumption, then
- $Cov(x_1, e_1) = E(x_1 e_1) = E(x_1^* e_1) + E(e_1^2) = 0 + \sigma_e^2$
- x_1 is correlated with the error so estimate is biased

$$\text{plim}\left(\hat{\beta}_1\right) = \beta_1 + \frac{Cov(x_1, u - \beta_1 e_1)}{Var(x_1)} = \beta_1 - \frac{\beta_1 \sigma_e^2}{\sigma_{x^*}^2 + \sigma_e^2}$$

$$= \beta_1 \left(1 - \frac{\sigma_e^2}{\sigma_{x^*}^2 + \sigma_e^2}\right) = \beta_1 \left(\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_e^2}\right)$$

Measurement Error in an Explanatory Variable (cont)

- Notice that the multiplicative error is just $Var(x_1^*)/Var(x_1)$
- Since $Var(x_1^*)/Var(x_1) < 1$, the estimate is biased toward zero – called attenuation bias
- It's more complicated with a multiple regression, but can still expect attenuation bias with classical errors in variables

Missing Data – Is it a Problem?

- If any observation is missing data on one of the variables in the model, it can't be used
- If data is missing at random, using a sample restricted to observations with no missing values will be fine
- A problem can arise if the data is missing systematically – say high income individuals refuse to provide income data

Nonrandom Samples

- If the sample is chosen on the basis of an x variable, then estimates are unbiased
- If the sample is chosen on the basis of the y variable, then we have sample selection bias
- Sample selection can be more subtle
- Say looking at wages for workers – since people choose to work this isn't the same as wage offers

Outliers

- Sometimes an individual observation can be very different from the others, and can have a large effect on the outcome
- Sometimes this outlier will simply be do to errors in data entry – one reason why looking at summary statistics is important
- Sometimes the observation will just truly be very different from the others

Outliers (continued)

- Not unreasonable to fix observations where it's clear there was just an extra zero entered or left off, etc.
- Not unreasonable to drop observations that appear to be extreme outliers, although readers may prefer to see estimates with and without the outliers
- Can use Stata to investigate outliers

Time Series Data

- $y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t$
- 1. Basic Analysis

Time Series vs. Cross Sectional

- Time series data has a temporal ordering, unlike cross-section data
- Will need to alter some of our assumptions to take into account that we no longer have a random sample of individuals
- Instead, we have one realization of a stochastic (i.e. random) process

Examples of Time Series Models

- A static model relates contemporaneous variables: $y_t = \beta_0 + \beta_1 z_t + u_t$
- A finite distributed lag (FDL) model allows one or more variables to affect *y* with a lag: $y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t$
- More generally, a finite distributed lag model of order *q* will include *q* lags of *z*

Finite Distributed Lag Models

- We can call δ_0 the impact propensity – it reflects the immediate change in *y*
- For a temporary, 1-period change, y returns to its original level in period *q*+1
- We can call $\delta_0 + \delta_1 + \ldots + \delta_q$ the long-run propensity (LRP) – it reflects the long-run change in *y* after a permanent change

Assumptions for Unbiasedness

- Still assume a model that is linear in parameters: $y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t$
- Still need to make a zero conditional mean assumption: $E(u_t|\mathbf{X}) = 0$, $t = 1, 2, \ldots$, n
- Note that this implies the error term in any given period is uncorrelated with the explanatory variables in all time periods

Assumptions (continued)

- This zero conditional mean assumption implies the x's are strictly exogenous
- An alternative assumption, more parallel to the cross-sectional case, is $E(u_t|\mathbf{x}_t) = 0$
- This assumption would imply the x's are contemporaneously exogenous
- Contemporaneous exogeneity will only be sufficient in large samples

Assumptions (continued)

- Still need to assume that no x is constant, and that there is no perfect collinearity
- Note we have skipped the assumption of a random sample
- The key impact of the random sample assumption is that each u_i is independent
- Our strict exogeneity assumption takes care of it in this case

Unbiasedness of OLS

- Based on these 3 assumptions, when using time-series data, the OLS estimators are unbiased
- Thus, just as was the case with cross-section data, under the appropriate conditions OLS is unbiased
- Omitted variable bias can be analyzed in the same manner as in the cross-section case

Variances of OLS Estimators

- Just as in the cross-section case, we need to add an assumption of homoskedasticity in order to be able to derive variances
- Now we assume $\mathrm{Var}(u_t|\mathbf{X}) = \mathrm{Var}(u_t) = \sigma^2$
- Thus, the error variance is independent of all the x's, and it is constant over time
- We also need the assumption of no serial correlation: $\mathrm{Corr}(u_t, u_s| \mathbf{X})=0$ for $t \neq s$

OLS Variances (continued)

- Under these 5 assumptions, the OLS variances in the time-series case are the same as in the cross-section case. Also,
- The estimator of σ^2 is the same
- OLS remains BLUE
- With the additional assumption of normal errors, inference is the same

Trending Time Series

- Economic time series often have a trend
- Just because 2 series are trending together, we can't assume that the relation is causal
- Often, both will be trending because of other unobserved factors
- Even if those factors are unobserved, we can control for them by directly controlling for the trend

Trends (continued)

- One possibility is a linear trend, which can be modeled as $y_t = \alpha_0 + \alpha_1 t + e_t$, $t = 1, 2, \ldots$
- Another possibility is an exponential trend, which can be modeled as $\log(y_t) = \alpha_0 + \alpha_1 t + e_t$, $t = 1, 2, \ldots$
- Another possibility is a quadratic trend, which can be modeled as $y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + e_t$, $t = 1, 2, \ldots$

Detrending

- Adding a linear trend term to a regression is the same thing as using "detrended" series in a regression
- Detrending a series involves regressing each variable in the model on t
- The residuals form the detrended series
- Basically, the trend has been partialled out

Detrending (continued)

- An advantage to actually detrending the data (vs. adding a trend) involves the calculation of goodness of fit
- Time-series regressions tend to have very high R^2, as the trend is well explained
- The R^2 from a regression on detrended data better reflects how well the x_t's explain y_t

Seasonality

- Often time-series data exhibits some periodicity, referred to seasonality
- Example: Quarterly data on retail sales will tend to jump up in the 4th quarter
- Seasonality can be dealt with by adding a set of seasonal dummies
- As with trends, the series can be seasonally adjusted before running the regression

Stationary Stochastic Process

- A stochastic process is stationary if for every collection of time indices $1 \le t_1 < \ldots < t_m$ the joint distribution of $(x_{t1}, \ldots, x_{tm})$ is the same as that of $(x_{t1+h}, \ldots x_{tm+h})$ for $h \ge 1$
- Thus, stationarity implies that the x_t's are identically distributed and that the nature of any correlation between adjacent terms is the same across all periods

Covariance Stationary Process

- A stochastic process is covariance stationary if $E(x_t)$ is constant, $Var(x_t)$ is constant and for any $t, h \ge 1$, $Cov(x_t, x_{t+h})$ depends only on h and not on t
- Thus, this weaker form of stationarity requires only that the mean and variance are constant across time, and the covariance just depends on the distance across time

Weakly Dependent Time Series

- A stationary time series is weakly dependent if x_t and x_{t+h} are "almost independent" as h increases
- If for a covariance stationary process $Corr(x_t, x_{t+h}) \rightarrow 0$ as $h \rightarrow \infty$, we'll say this covariance stationary process is weakly dependent
- Want to still use law of large numbers

An MA(1) Process

- A moving average process of order one [MA(1)] can be characterized as one where $x_t = e_t + \alpha_1 e_{t-1}$, $t = 1, 2, \ldots$ with e_t being an iid sequence with mean 0 and variance σ^2_e
- This is a stationary, weakly dependent sequence as variables 1 period apart are correlated, but 2 periods apart they are not

An AR(1) Process

- An autoregressive process of order one [AR(1)] can be characterized as one where $y_t = \rho y_{t-1} + e_t$, $t = 1, 2, \ldots$ with e_t being an iid sequence with mean 0 and variance σ_e^2
- For this process to be weakly dependent, it must be the case that $|\rho| < 1$
- $Corr(y_t, y_{t+h}) = Cov(y_t, y_{t+h})/(\sigma_y \sigma_y) = \rho_1^h$ which becomes small as h increases

Trends Revisited

- A trending series cannot be stationary, since the mean is changing over time
- A trending series can be weakly dependent
- If a series is weakly dependent and is stationary about its trend, we will call it a trend-stationary process
- As long as a trend is included, all is well

Assumptions for Consistency

- Linearity and Weak Dependence
- A weaker zero conditional mean assumption: $E(u_t|\mathbf{x}_t) = 0$, for each t
- No Perfect Collinearity
- Thus, for asymptotic unbiasedness (consistency), we can weaken the exogeneity assumptions somewhat relative to those for unbiasedness

Large-Sample Inference

- Weaker assumption of homoskedasticity: $\text{Var}\ (u_t|\mathbf{x}_t) = \sigma^2$, for each t
- Weaker assumption of no serial correlation: $E(u_t u_s|\ \mathbf{x}_t, \mathbf{x}_s) = 0$ for $t \neq s$
- With these assumptions, we have asymptotic normality and the usual standard errors, *t* statistics, *F* statistics and *LM* statistics are valid

Random Walks

- A random walk is an AR(1) model where $\rho_1 = 1$, meaning the series is not weakly dependent
- With a random walk, the expected value of y_t is always y_0 – it doesn't depend on t
- $\text{Var}(y_t) = \sigma_e^2 t$, so it increases with t
- We say a random walk is highly persistent since $E(y_{t+h}|y_t) = y_t$ for all $h \geq 1$

Random Walks (continued)

- A random walk is a special case of what's known as a unit root process
- Note that trending and persistence are different things – a series can be trending but weakly dependent, or a series can be highly persistent without any trend
- A random walk with drift is an example of a highly persistent series that is trending

Transforming Persistent Series

- In order to use a highly persistent series and get meaningful estimates and make correct inferences, we want to transform it into a weakly dependent process
- We refer to a weakly dependent process as being integrated of order zero, [I(0)]
- A random walk is integrated of order one, [I(1)], meaning a first difference will be I(0)

Time Series Data

- $y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t$
- 2. Further Issues

Testing for AR(1) Serial Correlation

- Want to be able to test for whether the errors are serially correlated or not
- Want to test the null that $\rho = 0$ in $u_t = \rho u_{t-1} + e_t$, $t = 2, \ldots, n$, where u_t is the model error term and e_t is iid
- With strictly exogenous regressors, the test is very straightforward – simply regress the residuals on lagged residuals and use a t-test

Testing for AR(1) Serial Correlation (continued)

- An alternative is the Durbin-Watson (DW) statistic, which is calculated by many packages
- If the DW statistic is around 2, then we can reject serial correlation, while if it is significantly < 2 we cannot reject
- Critical values are difficult to calculate, making the t test easier to work with

Testing for AR(1) Serial Correlation (continued)

- If the regressors are not strictly exogenous, then neither the t or DW test will work
- Regress the residual (or y) on the lagged residual and all of the x's
- The inclusion of the x's allows each x_{tj} to be correlated with u_{t-1}, so don't need assumption of strict exogeneity

Testing for Higher Order S.C.

- Can test for AR(q) serial correlation in the same basic manner as AR(1)
- Just include q lags of the residuals in the regression and test for joint significance
- Can use F test or LM test, where the LM version is called a Breusch-Godfrey test and is $(n-q)R^2$ using R^2 from residual regression
- Can also test for seasonal forms

Correcting for Serial Correlation

- Start with case of strictly exogenous regressors, and maintain all G-M assumptions except no serial correlation
- Assume errors follow AR(1) so $u_t = \rho u_{t-1} + e_t$, $t = 2, \ldots, n$
- $\mathrm{Var}(u_t) = \sigma^2_e/(1-\rho^2)$
- We need to try and transform the equation so we have no serial correlation in the errors

Correcting for S.C. (continued)

- Consider that since $y_t = \beta_0 + \beta_1 x_t + u_t$, then $y_{t-1} = \beta_0 + \beta_1 x_{t-1} + u_{t-1}$
- If you multiply the second equation by ρ, and subtract if from the first you get
- $y_t - \rho y_{t-1} = (1 - \rho)\beta_0 + \beta_1(x_t - \rho x_{t-1}) + e_t$, since $e_t = u_t - \rho u_{t-1}$
- This quasi-differencing results in a model without serial correlation

Feasible GLS Estimation

- Problem with this method is that we don't know ρ, so we need to get an estimate first
- Can just use the estimate obtained from regressing residuals on lagged residuals
- Depending on how we deal with the first observation, this is either called Cochrane-Orcutt or Prais-Winsten estimation

Feasible GLS (continued)

- Often both Cochrane-Orcutt and Prais-Winsten are implemented iteratively
- This basic method can be extended to allow for higher order serial correlation, AR(q)
- Most statistical packages will automatically allow for estimation of AR models without having to do the quasi-differencing by hand

Serial Correlation-Robust Standard Errors

- What happens if we don't think the regressors are all strictly exogenous?
- It's possible to calculate serial correlation-robust standard errors, along the same lines as heteroskedasticity robust standard errors
- Idea is that want to scale the OLS standard errors to take into account serial correlation

Serial Correlation-Robust Standard Errors (continued)

- Estimate normal OLS to get residuals, root MSE
- Run the auxiliary regression of x_{t1} on $x_{t2}, \ldots , x_{tk}$
- Form $\hat{a}_t$ by multiplying these residuals with $\hat{u}_t$
- Choose g – say 1 to 3 for annual data, then

$$\hat{v} = \sum_{t=1}^{n} \hat{a}_t^2 + 2\sum_{h=1}^{g} [1 - h/(g+1)] \left(\sum_{t=h+1}^{n} \hat{a}_t \hat{a}_{t-h} \right)$$

and $se(\hat{\beta}_1) = [SE/\hat{\sigma}]^2 \sqrt{\hat{v}}$, where SE is the usual OLS standard error of $\hat{\beta}_j$

Panel Data Methods

- $y_{it} = \beta_0 + \beta_1 x_{it1} + \ldots \beta_k x_{itk} + u_{it}$

A True Panel vs. A Pooled Cross Section

- Often loosely use the term panel data to refer to any data set that has both a cross-sectional dimension and a time-series dimension
- More precisely it's only data following the same cross-section units over time
- Otherwise it's a pooled cross-section

Pooled Cross Sections

- We may want to pool cross sections just to get bigger sample sizes
- We may want to pool cross sections to investigate the effect of time
- We may want to pool cross sections to investigate whether relationships have changed over time

Difference-in-Differences

- Say random assignment to treatment and control groups, like in a medical experiment
- One can then simply compare the change in outcomes across the treatment and control groups to estimate the treatment effect
- For time 1,2, groups A, B $(y_{2,B} - y_{2,A})$ - $(y_{1,B} - y_{1,A})$, or equivalently $(y_{2,B} - y_{1,B})$ - $(y_{2,A} - y_{1,A})$, is the difference-in-differences

Difference-in-Differences (cont)

- A regression framework using time and treatment dummy variables can calculate this difference-in-difference as well
- Consider the model: $y_{it} = \beta_0 + \beta_1 treatment_{it} + \beta_2 after_{it} + \beta_3 treatment_{it}*after_{it} + u_{it}$
- The estimated β_3 will be the difference-in-differences in the group means

Difference-in-Differences (cont)

- When don't truly have random assignment, the regression form becomes very useful
- Additional x's can be added to the regression to control for differences across the treatment and control groups
- Sometimes referred to as a "natural experiment" especially when a policy change is being analyzed

Two-Period Panel Data

- It's possible to use a panel just like pooled cross-sections, but can do more than that
- Panel data can be used to address some kinds of omitted variable bias
- If can think of the omitted variables as being fixed over time, then can model as having a composite error

Unobserved Fixed Effects

- Suppose the population model is $y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}$
- Here we have added a time-constant component to the error, $\upsilon_{it} = a_i + u_{it}$
- If a_i is correlated with the x's, OLS will be biased, since we a_i is part of the error term
- With panel data, we can difference-out the unobserved fixed effect

First-differences

- We can subtract one period from the other, to obtain $\Delta y_i = \delta_0 + \beta_1 \Delta x_{i1} + \ldots + \beta_k \Delta x_{ik} + \Delta u_i$
- This model has no correlation between the x's and the error term, so no bias
- Need to be careful about organization of the data to be sure compute correct change

Differencing w/ Multiple Periods

- Can extend this method to more periods
- Simply difference adjacent periods
- So if 3 periods, then subtract period 1 from period 2, period 2 from period 3 and have 2 observations per individual
- Simply estimate by OLS, assuming the Δu_{it} are uncorrelated over time

Fixed Effects Estimation

- When there is an observed fixed effect, an alternative to first differences is fixed effects estimation
- Consider the average over time of $y_{it} = \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}$
- The average of a_i will be a_i, so if you subtract the mean, a_i will be differenced out just as when doing first differences

Fixed Effects Estimation (cont)

- If we were to do this estimation by hand, we'd need to be careful because we'd think that $df = NT - k$, but really is $N(T - 1) - k$ because we used up *df*s calculating means
- Luckily, Stata (and most other packages) will do fixed effects estimation for you
- This method is also identical to including a separate intercept or every individual

First Differences vs Fixed Effects

- First Differences and Fixed Effects will be exactly the same when T = 2
- For T > 2, the two methods are different
- Probably see fixed effects estimation more often than differences – probably more because it's easier than that it's better
- Fixed effects easily implemented for unbalanced panels, not just balanced panels

Random Effects

- Start with the same basic model with a composite error, $y_{it} = \beta_0 + \beta_1 x_{it1} + \ldots \beta_k x_{itk} + a_i + u_{it}$
- Previously we've assumed that a_i was correlated with the x's, but what if it's not?
- OLS would be consistent in that case, but composite error will be serially correlated

Random Effects (continued)

- Need to transform the model and do GLS to solve the problem and make correct inferences
- Idea is to do quasi-differencing with the

Random Effects (continued)

- Need to transform the model and do GLS to solve the problem and make correct inferences
- End up with a sort of weighted average of OLS and Fixed Effects – use quasi-demeaned data

$$\lambda = 1 - \left[\sigma_u^2 / \left(\sigma_u^2 + T\sigma_a^2\right)\right]^{1/2}$$

$$y_{it} - \lambda \bar{y}_i = \beta_0(1-\lambda) + \beta_1\left(x_{it1} - \lambda \bar{x}_{i1}\right) + \ldots + \beta_k\left(x_{itk} - \bar{x}_{ik}\right) + \left(\nu_{it} - \bar{\nu}_i\right)$$

Random Effects (continued)

- If $\lambda = 1$, then this is just the fixed effects estimator
- If $\lambda = 0$, then this is just the OLS estimator
- So, the bigger the variance of the unobserved effect, the closer it is to FE
- The smaller the variance of the unobserved effect, the closer it is to OLS
- Stata will do Random Effects for us

Fixed Effects or Random?

- More usual to think need fixed effects, since think the problem is that something unobserved is correlated with the *x*'s
- If truly need random effects, the only problem is the standard errors
- Can just adjust the standard errors for correlation within group

Other Uses of Panel Methods

- It's possible to think of models where there is an unobserved fixed effect, even if we do not have true panel data
- A common example is where we think there is an unobserved family effect
- Can difference siblings
- Can estimate family fixed effect model

Additional Issues

- Many of the things we already know about both cross section and time series data can be applied with panel data
- Can test and correct for serial correlation in the errors
- Can test and correct for heteroskedasticity
- Can estimate standard errors robust to both

Instrumental Variables & 2SLS

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + u$
- $x_1 = \pi_0 + \pi_1 z + \pi_2 x_2 + \ldots \pi_k x_k + v$

Why Use Instrumental Variables?

- Instrumental Variables (IV) estimation is used when your model has endogenous x's
- That is, whenever Cov(x,u) $\neq 0$
- Thus, IV can be used to address the problem of omitted variable bias
- Additionally, IV can be used to solve the classic errors-in-variables problem

What Is an Instrumental Variable?

- In order for a variable, z, to serve as a valid instrument for x, the following must be true
- The instrument must be exogenous
- That is, Cov(z,u) = 0
- The instrument must be correlated with the endogenous variable x
- That is, Cov(z,x) $\neq 0$

More on Valid Instruments

- We have to use common sense and economic theory to decide if it makes sense to assume Cov(z,u) = 0
- We can test if Cov(z,x) $\neq 0$
- Just testing H_0: $\pi_1 = 0$ in $x = \pi_0 + \pi_1 z + v$
- Sometimes refer to this regression as the first-stage regression

IV Estimation in the Simple Regression Case

- For $y = \beta_0 + \beta_1 x + u$, and given our assumptions
- Cov(z,y) = β_1Cov(z,x) + Cov(z,u), so
- β_1 = Cov(z,y) / Cov(z,x)
- Then the IV estimator for β_1 is

$$\hat{\beta}_1 = \frac{\sum (z_i - \bar{z})(y_i - \bar{y})}{\sum (z_i - \bar{z})(x_i - \bar{x})}$$

Inference with IV Estimation

- The homoskedasticity assumption in this case is $E(u^2|z) = \sigma^2 = Var(u)$
- As in the OLS case, given the asymptotic variance, we can estimate the standard error

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{n\sigma_x^2 \rho_{x,z}^2}$$

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{SST_x R_{x,z}^2}$$

IV versus OLS estimation

- Standard error in IV case differs from OLS only in the R^2 from regressing x on z
- Since $R^2 < 1$, IV standard errors are larger
- However, IV is consistent, while OLS is inconsistent, when $\mathrm{Cov}(x,u) \neq 0$
- The stronger the correlation between z and x, the smaller the IV standard errors

The Effect of Poor Instruments

- What if our assumption that $\mathrm{Cov}(z,u) = 0$ is false?
- The IV estimator will be inconsistent, too
- Can compare asymptotic bias in OLS and IV
- Prefer IV if $\mathrm{Corr}(z,u)/\mathrm{Corr}(z,x) < \mathrm{Corr}(x,u)$

$$\text{IV}: \operatorname{plim}\hat{\beta}_1 = \beta_1 + \frac{Corr(z,u)}{Corr(z,x)} \bullet \frac{\sigma_u}{\sigma_x}$$

$$\text{OLS}: \operatorname{plim}\tilde{\beta}_1 = \beta_1 + Corr(x,u) \bullet \frac{\sigma_u}{\sigma_x}$$

IV Estimation in the Multiple Regression Case

- IV estimation can be extended to the multiple regression case
- Call the model we are interested in estimating the structural model
- Our problem is that one or more of the variables are endogenous
- We need an instrument for each endogenous variable

Multiple Regression IV (cont)

- Write the structural model as $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$, where y_2 is endogenous and z_1 is exogenous
- Let z_2 be the instrument, so $\mathrm{Cov}(z_2,u_1) = 0$ and
- $y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_2$, where $\pi_2 \neq 0$
- This reduced form equation regresses the endogenous variable on all exogenous ones

Two Stage Least Squares (2SLS)

- It's possible to have multiple instruments
- Consider our original structural model, and let $y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2$
- Here we're assuming that both z_2 and z_3 are valid instruments – they do not appear in the structural model and are uncorrelated with the structural error term, u_1

Best Instrument

- Could use either z_2 or z_3 as an instrument
- The best instrument is a linear combination of all of the exogenous variables, $y_2^* = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3$
- We can estimate y_2^* by regressing y_2 on z_1, z_2 and z_3 – can call this the first stage
- If then substitute $\hat{y}_2$ for y_2 in the structural model, get same coefficient as IV

More on 2SLS

- While the coefficients are the same, the standard errors from doing 2SLS by hand are incorrect, so let Stata do it for you
- Method extends to multiple endogenous variables – need to be sure that we have at least as many excluded exogenous variables (instruments) as there are endogenous variables in the structural equation

Addressing Errors-in-Variables with IV Estimation

- Remember the classical errors-in-variables problem where we observe x_1 instead of x_1*
- Where $x_1 = x_1^* + e_1$, and e_1 is uncorrelated with x_1* and x_2
- If there is a z, such that Corr(z,u) = 0 and Corr(z,x_1) $\neq$ 0, then
- IV will remove the attenuation bias

Testing for Endogeneity

- Since OLS is preferred to IV if we do not have an endogeneity problem, then we'd like to be able to test for endogeneity
- If we do not have endogeneity, both OLS and IV are consistent
- Idea of Hausman test is to see if the estimates from OLS and IV are different

Testing for Endogeneity (cont)

- While it's a good idea to see if IV and OLS have different implications, it's easier to use a regression test for endogeneity
- If y_2 is endogenous, then v_2 (from the reduced form equation) and u_1 from the structural model will be correlated
- The test is based on this observation

Testing for Endogeneity (cont)

- Save the residuals from the first stage
- Include the residual in the structural equation (which of course has y_2 in it)
- If the coefficient on the residual is statistically different from zero, reject the null of exogeneity
- If multiple endogenous variables, jointly test the residuals from each first stage

Testing Overidentifying Restrictions

- If there is just one instrument for our endogenous variable, we can't test whether the instrument is uncorrelated with the error
- We say the model is just identified
- If we have multiple instruments, it is possible to test the overidentifying restrictions – to see if some of the instruments are correlated with the error

The OverID Test

- Estimate the structural model using IV and obtain the residuals
- Regress the residuals on all the exogenous variables and obtain the R^2 to form nR^2
- Under the null that all instruments are uncorrelated with the error, LM ~ χ_q^2 where q is the number of extra instruments

Testing for Heteroskedasticity

- When using 2SLS, we need a slight adjustment to the Breusch-Pagan test
- Get the residuals from the IV estimation
- Regress these residuals squared on all of the exogenous variables in the model (including the instruments)
- Test for the joint significance

Testing for Serial Correlation

- When using 2SLS, we need a slight adjustment to the test for serial correlation
- Get the residuals from the IV estimation
- Re-estimate the structural model by 2SLS, including the lagged residuals, and using the same instruments as originally
- Can do 2SLS on a quasi-differenced model, using quasi-differenced instruments

Simultaneous Equations

- $y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1$
- $y_2 = \alpha_2 y_1 + \beta_2 z_2 + u_2$

Simultaneity

- Simultaneity is a specific type of endogeneity problem in which the explanatory variable is jointly determined with the dependent variable
- As with other types of endogeneity, IV estimation can solve the problem
- Some special issues to consider with simultaneous equations models (SEM)

Supply and Demand Example

- Start with an equation you'd like to estimate, say a labor supply function
- $h_s = \alpha_1 w + \beta_1 z + u_1$, where
- w is the wage and z is a supply shifter
- Call this a structural equation – it's derived from economic theory and has a causal interpretation where w directly affects h_s

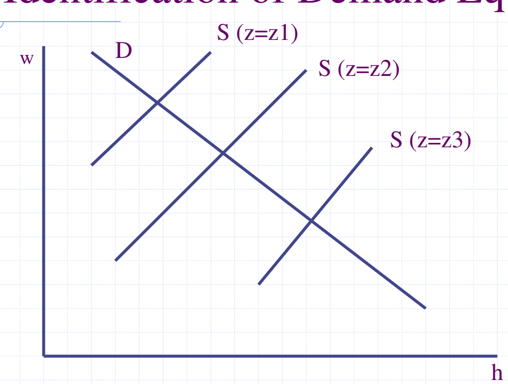
Example (cont)

- Problem that can't just regress observed hours on wage, since observed hours are determined by the equilibrium of supply and demand
- Consider a second structural equation, in this case the labor demand function
- $h_d = \alpha_2 w + u_2$
- So hours are determined by a SEM

Example (cont)

- Both h and w are endogenous because they are both determined by the equilibrium of supply and demand
- z is exogenous, and it's the availability of this exogenous supply shifter that allows us to identify the structural demand equation
- With no observed demand shifters, supply is not identified and cannot be estimated

Identification of Demand Equation



Using IV to Estimate Demand

- So, we can estimate the structural demand equation, using z as an instrument for w
- First stage equation is $w = \pi_0 + \pi_1 z + v_2$
- Second stage equation is $h = \alpha_2 \hat{w} + u_2$
- Thus, 2SLS provides a consistent estimator of α_2, the slope of the demand curve
- We cannot estimate α_1, the slope of the supply curve

The General SEM

- Suppose you want to estimate the structural equation: $y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1$
- where, $y_2 = \alpha_2 y_1 + \beta_2 z_2 + u_2$
- Thus, $y_2 = \alpha_2(\alpha_1 y_2 + \beta_1 z_1 + u_1) + \beta_2 z_2 + u_2$
- So, $(1 - \alpha_2\alpha_1)y_2 = \alpha_2\, \beta_1 z_1 + \beta_2 z_2 + \alpha_2\, u_1 + u_2$, which can be rewritten as
- $y_2 = \pi_1 z_1 + \pi_2 z_2 + v_2$

The General SEM (continued)

- By substituting this reduced form in for y_2, we can see that since v_2 is a linear function of u_1, y_2 is correlated with the error term and α_1 is biased – call it simultaneity bias
- The sign of the bias is complicated, but can use the simple regression as a rule of thumb
- In the simple regression case, the bias is the same sign as $\alpha_2/(1 - \alpha_2\alpha_1)$

Identification of General SEM

- Let z_1 be all the exogenous variables in the first equation, and z_2 be all the exogenous variables in the second equation
- It's okay for there to be overlap in z_1 and z_2
- To identify equation 1, there must be some variables in z_2 that are not in z_1
- To identify equation 2, there must be some variables in z_1 that are not in z_2

Rank and Order Conditions

- We refer to this as the rank condition
- Note that the exogenous variable excluded from the first equation must have a non-zero coefficient in the second equation for the rank condition to hold
- Note that the order condition clearly holds if the rank condition does – there will be an exogenous variable for the endogenous one

Estimation of the General SEM

- Estimation of SEM is straightforward
- The instruments for 2SLS are the exogenous variables from both equations
- Can extend the idea to systems with more than 2 equations
- For a given identified equation, the instruments are all of the exogenous variables in the whole system

Limited Dependent Variables

- $P(y = 1|\mathbf{x}) = G(\beta_0 + \mathbf{x}\beta)$
- $y^* = \beta_0 + \mathbf{x}\beta + u$, $y = max(0,y^*)$

Binary Dependent Variables

- Recall the linear probability model, which can be written as $P(y = 1|\mathbf{x}) = \beta_0 + \mathbf{x}\beta$
- A drawback to the linear probability model is that predicted values are not constrained to be between 0 and 1
- An alternative is to model the probability as a function, $G(\beta_0 + \mathbf{x}\beta)$, where $0<G(z)<1$

The Probit Model

- One choice for $G(z)$ is the standard normal cumulative distribution function (cdf)
- $G(z) = \Phi(z) \equiv \int\phi(v)dv$, where $\phi(z)$ is the standard normal, so $\phi(z) = (2\pi)^{-1/2}\exp(-z^2/2)$
- This case is referred to as a probit model
- Since it is a nonlinear model, it cannot be estimated by our usual methods
- Use maximum likelihood estimation

The Logit Model

- Another common choice for G(z) is the logistic function, which is the cdf for a standard logistic random variable
- $G(z) = \exp(z)/[1 + \exp(z)] = \Lambda(z)$
- This case is referred to as a logit model, or sometimes as a logistic regression
- Both functions have similar shapes – they are increasing in z, most quickly around 0

Probits and Logits

- Both the probit and logit are nonlinear and require maximum likelihood estimation
- No real reason to prefer one over the other
- Traditionally saw more of the logit, mainly because the logistic function leads to a more easily computed model
- Today, probit is easy to compute with standard packages, so more popular

Interpretation of Probits and Logits (in particular vs LPM)

- In general we care about the effect of x on $P(y = 1|\mathbf{x})$, that is, we care about $\partial p/ \partial x$
- For the linear case, this is easily computed as the coefficient on x
- For the nonlinear probit and logit models, it's more complicated:
- $\partial p/ \partial x_j = g(\beta_0 + \mathbf{x}\beta)\beta_j$, where $g(z)$ is dG/dz

Interpretation (continued)

- Clear that it's incorrect to just compare the coefficients across the three models
- Can compare sign and significance (based on a standard *t* test) of coefficients, though
- To compare the magnitude of effects, need to calculate the derivatives, say at the means
- Stata will do this for you in the probit case

The Likelihood Ratio Test

- Unlike the LPM, where we can compute *F* statistics or *LM* statistics to test exclusion restrictions, we need a new type of test
- Maximum likelihood estimation (MLE), will always produce a log-likelihood, $\mathcal{L}$
- Just as in an F test, you estimate the restricted and unrestricted model, then form
- $LR = 2(\mathcal{L}_{ur} - \mathcal{L}_r) \sim \chi^2_q$

Goodness of Fit

- Unlike the LPM, where we can compute an R^2 to judge goodness of fit, we need new measures of goodness of fit
- One possibility is a pseudo R^2 based on the log likelihood and defined as $1 - \mathcal{L}_{ur}/\mathcal{L}_r$
- Can also look at the percent correctly predicted – if predict a probability >.5 then that matches y = 1 and vice versa

Latent Variables

- Sometimes binary dependent variable models are motivated through a latent variables model
- The idea is that there is an underlying variable y*, that can be modeled as
- $y^* = \beta_0 + \boldsymbol{x}\beta + e$, but we only observe
- $y = 1$, if $y^* > 0$, and $y = 0$ if $y^* \leq 0$,

The Tobit Model

- Can also have latent variable models that don't involve binary dependent variables
- Say $y^* = \boldsymbol{x}\beta + u$, $u|\boldsymbol{x} \sim \text{Normal}(0,\sigma^2)$
- But we only observe $y = \max(0, y^*)$
- The Tobit model uses MLE to estimate both β and σ for this model
- Important to realize that β estimates the effect of $\boldsymbol{x}$ on y*, the latent variable, not y

Interpretation of the Tobit Model

- Unless the latent variable y^* is what's of interest, can't just interpret the coefficient
- $E(y|\boldsymbol{x}) = \Phi(\boldsymbol{x}\beta/\sigma)\boldsymbol{x}\beta + \sigma\phi(\boldsymbol{x}\beta/\sigma)$, so
- $\partial E(y|\boldsymbol{x})/\partial x_j = \beta_j \Phi(\boldsymbol{x}\beta/\sigma)$
- If normality or homoskedasticity fail to hold, the Tobit model may be meaningless
- If the effect of x on $P(y>0)$ and $E(y)$ are of opposite signs, the Tobit is inappropriate

Censored Regression Models & Truncated Regression Models

- More general latent variable models can also be estimated, say
- $y = \boldsymbol{x}\beta + u$, $u|\boldsymbol{x},c \sim \text{Normal}(0,\sigma^2)$, but we only observe $w = \min(y,c)$ if right censored, or w = max(y,c) if left censored
- Truncated regression occurs when rather than being censored, the data is missing beyond a censoring point

Sample Selection Corrections

- If a sample is truncated in a nonrandom way, then OLS suffers from selection bias
- Can think of as being like omitted variable bias, where what's omitted is how were selected into the sample, so
- $\text{E}(y|z, s = 1) = \boldsymbol{x}\beta + \rho\lambda(z\gamma)$, where
- $\lambda(c)$ is the inverse Mills ratio: $\phi(c)/\Phi(c)$

Selection Correction (continued)

- We need an estimate of λ, so estimate a probit of s (whether y is observed) on z
- These estimates of γ can then be used along with z to form the inverse Mills ratio
- Then you can just regress y on $\boldsymbol{x}$ and the estimated λ to get consistent estimates of β
- Important that $\boldsymbol{x}$ be a subset of z, otherwise will only be identified by functional form

Limited Dependent Variables

- $P(y = 1|\mathbf{x}) = G(\beta_0 + \mathbf{x}\beta)$
- $y^* = \beta_0 + \mathbf{x}\beta + u,\ y = max(0,y^*)$

Binary Dependent Variables

- Recall the linear probability model, which can be written as $P(y = 1|\mathbf{x}) = \beta_0 + \mathbf{x}\beta$
- A drawback to the linear probability model is that predicted values are not constrained to be between 0 and 1
- An alternative is to model the probability as a function, $G(\beta_0 + \mathbf{x}\beta)$, where $0<G(z)<1$

The Probit Model

- One choice for $G(z)$ is the standard normal cumulative distribution function (cdf)
- $G(z) = \Phi(z) \equiv \int\phi(v)dv$, where $\phi(z)$ is the standard normal, so $\phi(z) = (2\pi)^{-1/2}\exp(-z^2/2)$
- This case is referred to as a probit model
- Since it is a nonlinear model, it cannot be estimated by our usual methods
- Use maximum likelihood estimation

The Logit Model

- Another common choice for G(z) is the logistic function, which is the cdf for a standard logistic random variable
- $G(z) = \exp(z)/[1 + \exp(z)] = \Lambda(z)$
- This case is referred to as a logit model, or sometimes as a logistic regression
- Both functions have similar shapes – they are increasing in z, most quickly around 0

Probits and Logits

- Both the probit and logit are nonlinear and require maximum likelihood estimation
- No real reason to prefer one over the other
- Traditionally saw more of the logit, mainly because the logistic function leads to a more easily computed model
- Today, probit is easy to compute with standard packages, so more popular

Interpretation of Probits and Logits (in particular vs LPM)

- In general we care about the effect of x on $P(y = 1|\mathbf{x})$, that is, we care about $\partial p/ \partial x$
- For the linear case, this is easily computed as the coefficient on x
- For the nonlinear probit and logit models, it's more complicated:
- $\partial p/ \partial x_j = g(\beta_0 + \mathbf{x}\beta)\beta_j$, where $g(z)$ is dG/dz

Interpretation (continued)

- Clear that it's incorrect to just compare the coefficients across the three models
- Can compare sign and significance (based on a standard *t* test) of coefficients, though
- To compare the magnitude of effects, need to calculate the derivatives, say at the means
- Stata will do this for you in the probit case

The Likelihood Ratio Test

- Unlike the LPM, where we can compute *F* statistics or *LM* statistics to test exclusion restrictions, we need a new type of test
- Maximum likelihood estimation (MLE), will always produce a log-likelihood, $\mathcal{L}$
- Just as in an F test, you estimate the restricted and unrestricted model, then form
- $LR = 2(\mathcal{L}_{ur} - \mathcal{L}_r) \sim \chi^2_q$

Goodness of Fit

- Unlike the LPM, where we can compute an R^2 to judge goodness of fit, we need new measures of goodness of fit
- One possibility is a pseudo R^2 based on the log likelihood and defined as $1 - \mathcal{L}_{ur}/\mathcal{L}_r$
- Can also look at the percent correctly predicted – if predict a probability >.5 then that matches y = 1 and vice versa

Latent Variables

- Sometimes binary dependent variable models are motivated through a latent variables model
- The idea is that there is an underlying variable y*, that can be modeled as
- $y^* = \beta_0 + \boldsymbol{x}\beta + e$, but we only observe
- $y = 1$, if $y^* > 0$, and $y = 0$ if $y^* \leq 0$,

The Tobit Model

- Can also have latent variable models that don't involve binary dependent variables
- Say $y^* = \boldsymbol{x}\beta + u$, $u|\boldsymbol{x} \sim \text{Normal}(0,\sigma^2)$
- But we only observe $y = \max(0, y^*)$
- The Tobit model uses MLE to estimate both β and σ for this model
- Important to realize that β estimates the effect of $\boldsymbol{x}$ on y*, the latent variable, not *y*

Interpretation of the Tobit Model

- Unless the latent variable *y** is what's of interest, can't just interpret the coefficient
- $E(y|\boldsymbol{x}) = \Phi(\boldsymbol{x}\beta/\sigma)\boldsymbol{x}\beta + \sigma\phi(\boldsymbol{x}\beta/\sigma)$, so
- $\partial E(y|\boldsymbol{x})/\partial x_j = \beta_j \Phi(\boldsymbol{x}\beta/\sigma)$
- If normality or homoskedasticity fail to hold, the Tobit model may be meaningless
- If the effect of *x* on $P(y>0)$ and $E(y)$ are of opposite signs, the Tobit is inappropriate

Censored Regression Models & Truncated Regression Models

- More general latent variable models can also be estimated, say
- $y = \boldsymbol{x}\beta + u$, $u|\boldsymbol{x},c \sim \text{Normal}(0,\sigma^2)$, but we only observe $w = \min(y,c)$ if right censored, or w = max(y,c) if left censored
- Truncated regression occurs when rather than being censored, the data is missing beyond a censoring point

Sample Selection Corrections

- If a sample is truncated in a nonrandom way, then OLS suffers from selection bias
- Can think of as being like omitted variable bias, where what's omitted is how were selected into the sample, so
- $E(y|z, s = 1) = \boldsymbol{x}\beta + \rho\lambda(z\gamma)$, where
- $\lambda(c)$ is the inverse Mills ratio: $\phi(c)/\Phi(c)$

Selection Correction (continued)

- We need an estimate of λ, so estimate a probit of s (whether y is observed) on z
- These estimates of γ can then be used along with z to form the inverse Mills ratio
- Then you can just regress y on $\boldsymbol{x}$ and the estimated λ to get consistent estimates of β
- Important that $\boldsymbol{x}$ be a subset of z, otherwise will only be identified by functional form

Testing for Unit Roots

- Consider an AR(1): $y_t = \alpha + \rho y_{t-1} + e_t$
- Let H_0: $\rho = 1$, (assume there is a unit root)
- Define $\theta = \rho - 1$ and subtract y_{t-1} from both sides to obtain $\Delta y_t = \alpha + \theta y_{t-1} + e_t$
- Unfortunately, a simple t-test is inappropriate, since this is an I(1) process
- A Dickey-Fuller Test uses the t-statistic, but different critical values

Testing for Unit Roots (cont)

- We can add p lags of Δy_t to allow for more dynamics in the process
- Still want to calculate the t-statistic for θ
- Now it's called an augmented Dickey-Fuller test, but still the same critical values
- The lags are intended to clear up any serial correlation, if too few, test won't be right

Testing for Unit Roots w/ Trends

- If a series is clearly trending, then we need to adjust for that or might mistake a trend stationary series for one with a unit root
- Can just add a trend to the model
- Still looking at the t-statistic for θ, but the critical values for the Dickey-Fuller test change

Spurious Regression

- Consider running a simple regression of y_t on x_t where y_t and x_t are independent I(1) series
- The usual OLS t-statistic will often be statistically significant, indicating a relationship where there is none
- Called the spurious regression problem

Cointegration

- Say for two I(1) processes, y_t and x_t, there is a β such that $y_t - \beta x_t$ is an I(0) process
- If so, we say that y and x are cointegrated, and call β the cointegration parameter
- If we know β, testing for cointegration is straightforward if we define $s_t = y_t - \beta x_t$
- Do Dickey-Fuller test and if we reject a unit root, then they are cointegrated

Cointegration (continued)

- If β is unknown, then we first have to estimate β, which adds a complication
- After estimating β we run a regression of $\Delta \hat{u}_t$ on $\hat{u}_{t-1}$ and compare t-statistic on $\hat{u}_{t-1}$ with the special critical values
- If there are trends, need to add it to the initial regression that estimates β and use different critical values for t-statistic on $\hat{u}_{t-1}$

Forecasting

- Once we've run a time-series regression we can use it for forecasting into the future
- Can calculate a point forecast and forecast interval in the same way we got a prediction and prediction interval with a cross-section
- Rather than use in-sample criteria like adjusted R^2, often want to use out-of-sample criteria to judge how good the forecast is

Out-of-Sample Criteria

- Idea is to note use all of the data in estimating the equation, but to save some for evaluating how well the model forecasts
- Let total number of observations be $n + m$ and use n of them for estimating the model
- Use the model to predict the next m observations, and calculate the difference between your prediction and the truth

Out-of-Sample Criteria (cont)

- Call this difference the forecast error, which is $\hat{e}_{n+h+1}$ for $h = 0, 1, \ldots, m$
- Calculate the root mean square error (RMSE)

Out-of-Sample Criteria (cont)

- Call this difference the forecast error, which is $\hat{e}_{n+h+1}$ for $h = 0, 1, \ldots, m$
- Calculate the root mean square error and see which model has the smallest, where

$$RMSE = \left(m^{-1} \sum_{h=0}^{m-1} \hat{e}_{n+h+1}^2 \right)^{1/2}$$

Summary and Conclusions

Carrying Out an Empirical Project

Choosing a Topic

- Start with a general area or set of questions
- Make sure you are interested in the topic
- Use on-line services such as EconLit to investigate past work on this topic
- Narrow down your topic to a specific question or issue to be investigated
- Work through the theoretical issue

Choosing Data

- Want data that includes measures of the things that your theoretical model imply are important
- Investigate what type of data sets have been used in the past literature
- Search for what other data sets are available (for example, ICPSR)
- Consider collecting your own data

Using the Data

- Create variables appropriate for analysis
- For example, create dummy variables from categorical variables, create hourly wages, etc.
- Check the data for missing values, errors, outliers, etc.
- Recode as necessary, be sure to report what you did

Estimating a Model

- Start with a model that is clearly based in theory
- Test for significance of other variables that are theoretically less clear
- Test for functional form misspecification
- Consider reasonable interactions, quadratics, logs, etc.

Estimating a Model (continued)

- Don't lose sight of theory and the *ceteris paribus* interpretation – you need to be careful about including variables that greatly alter the interpretation
- For example, effect of bedrooms conditional on square footage
- Be careful about putting functions of y on the right hand side – affects interpretation

Estimating a Model (continued)

- Once you have a well-specified model, need to worry about the standard errors
- Test for heteroskedasticity
- Correct if necessary
- Test for serial correlation if there is a time component
- Correct if necessary

Other Problems

- Often you have to worry about endogeneity of the key explanatory variable
- Endogeneity could arise from omitted variables that are not observed in the data
- Endogeneity could arise because the model is really part of a simultaneous equation
- Endogeneity could arise due to measurement error

Other Problems (continued)

- If you have panel data, can consider a fixed effects model (or first differences)
- Problem with FE is that need good variation over time
- Can instead try to find a perfect instrument and perform 2SLS
- Problem with IV is finding a good instrument

Interpreting Your Results

- Keep theory in mind when interpreting results
- Be careful to keep ceteris paribus in mind
- Keep in mind potential problems with your estimates – be cautious drawing conclusions
- Can get an idea of the direction of bias due to omitted variables, measurement error or simultaneity

Further Issues

- Some problems are just too hard to easily solve with available data
- May be able to approach the problem in several ways, but something wrong with each one
- Provide enough information for a reader to decide whether they find your results convincing or not

Further Issues (continued)

- Don’t worry if you don’t “prove” your theory
- With unexpected results, you want to be careful in thinking through potential biases
- But, ff you have carefully specified your model and feel confident you have unbiased estimates, then that’s just the way things are