Econometrics V lecture 3

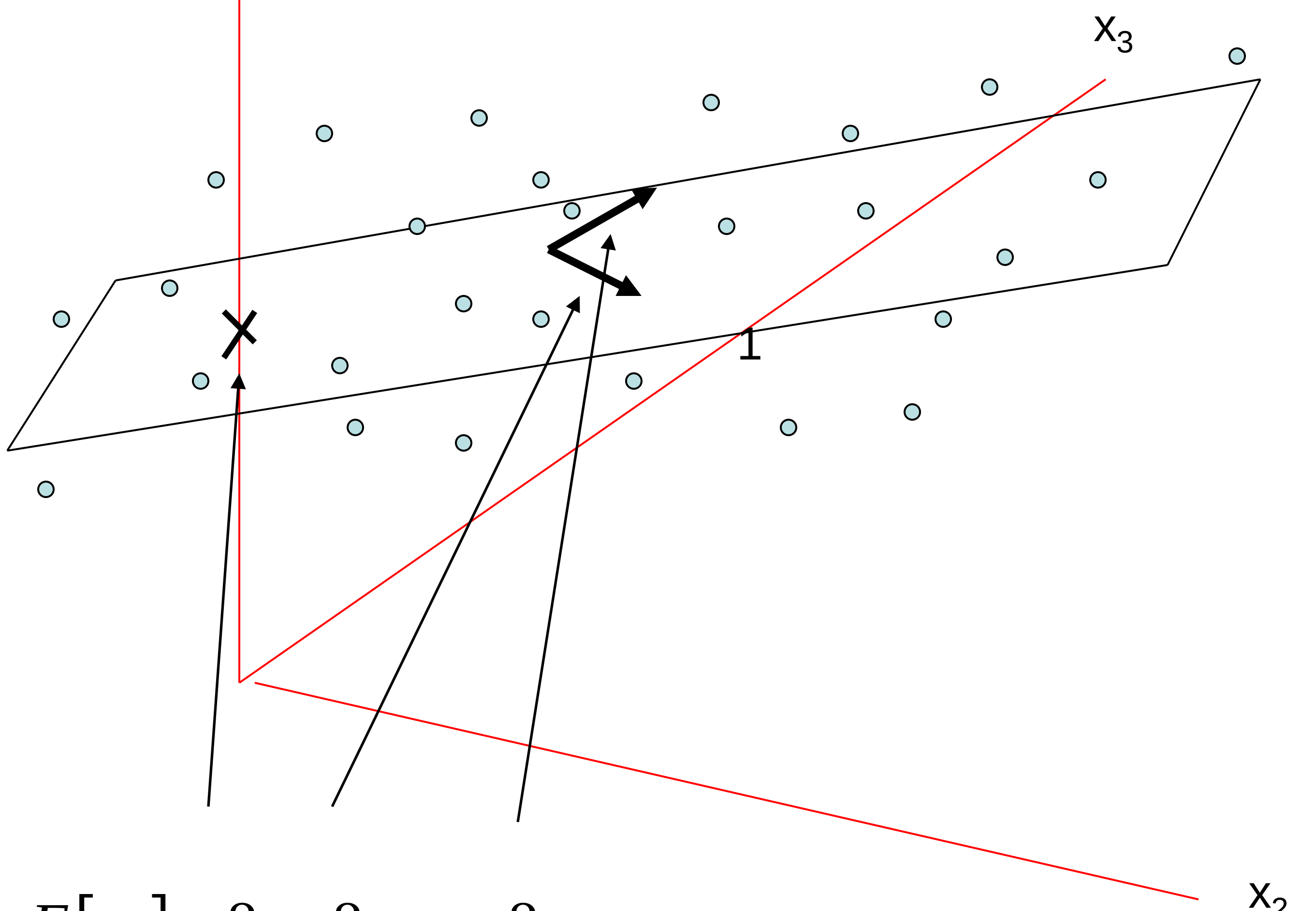
Multiple Regression Analysis

Econometrics V lecture 3

- Basic principle the same.
- OLS still minimises the sum of the squared residuals.
- Now there is more than 1 explanatory (right hand side) variable.
- Slide on next page illustrates idea when we have 2 explanatory variables – its 3D space and we fit a plane

Econometrics V lecture 3

- Econometrics V lecture 3



$$E[y_t] = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3}$$

Econometrics V lecture 3

- Start with a simple example:
-
- What might y and x1 and x2 be?
- y could be consumption
- x_2 could be income
- x_3 could be the interest rate
- Interpretation of parameters?
- Very similar to simple regression

$$\beta_2 = \frac{\partial y}{\partial x_2} \qquad \beta_3 = \frac{\partial y}{\partial x_3}$$

- What if we were using logarithms???

Econometrics V lecture 3

- Assumptions?
- Essentially those of simple regression – but now we have to add to the assumption that x is not random as there are more than 1 x variables (explanatory variables).
- We now have to assume that the x variables are not perfectly linearly related to each other ie $x_2 \neq 4x_3$ – this is the assumption of no perfect multicollinearity.

Econometrics V lecture 3

- SR1 $E\left[y_t\right]=\beta_1+\beta_2 x_{t2}+\beta_3 x_{t3}$
- SR2. $E(e)=0 \Leftrightarrow E(y)=\beta_1+\beta_2 x$
- SR3. $\operatorname{var}(e)=\sigma^2=\operatorname{var}(y)$
- SR4. $\operatorname{cov}(e_i,e_j)=\operatorname{cov}(y_i,y_j)=0$
- SR5. The variable *x* is not random and are not exact linear functions of each other.
- SR6. (*optional*) The values of *e* are *normally distributed* about their mean

$$e \sim N(0,\sigma^2)$$

Econometrics V lecture 3

- Multicollinearity
- M is important – think how it might arise.
- If we do have perfect M ols will fail – try it – make up some data and make one of the x's a linear function of the other – then try running it.
- The software will fail – - "singular matrix" or some such nonsense?!

Econometrics V lecture 3

- We will look at the effects of M later in the lecture – but the logic is simple – suppose you are trying to explain the weight of individuals – you use two explanatory variables – height and inside leg measurement.
- OLS will have problems sorting out the effects of the 2 exp. Variables because they are likely to be highly collinear.

Econometrics V lecture 3

- Estimation

- You will not have to do any calculations “by hand” for the multiple regression model – computers only and in the exam we will focus on interpretation and use of results – not parameter calculation.
- Method the same:

$$E\left[y_t\right] = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3}$$

$$Min^m Q = \sum (y - \hat{\beta}_1 - \hat{\beta}_2 - \hat{\beta}_3)^2$$

$$w.r.t. \hat{\beta}_{1,} \hat{\beta}_2 , \hat{\beta}_3$$

Econometrics V lecture 3

- So as before – differentiate w.r.t. to each parameter, set equal to zero – we will get 3 equations in 3 unknowns, the normal equations - which can then be solved.
- So given data on y, x_1 and x_2 we can use the method of ols to obtain estimates of the parameters β_1, β_2 and β_3 which we will call

$$\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$$

Econometrics V lecture 3

- OK just like simple regression
- We have estimators of the parameters
- Once again the estimators are random variables.
- Again to make it useful we need to know their means and variances
- Once again it turns out expected values equal population values – so we just need formulae for the variances.
- Once again the first step is the error variance

Econometrics V lecture 3

- And once again

$$\hat{\sigma}^2 = \frac{\sum \hat{e}^2}{T-k}$$

- So the estimator of the error variance uses the same formulae as for simple regression but now k is not always equal to 2
- k is the number of estimated parameters in the regression model so varies according to how many explanatory variables there are.

Econometrics V lecture 3

- Now, and again without proof, we can work out expressions for the variances and covariances of the estimated parameters.
- These formulae involve σ^2 and we can use our estimate of this.
- For our example model

$$E\left[y_t\right] = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3}$$

- We can write these down. For example:

$$\hat{\text{var}}(\hat{\beta}_2) = \frac{\sigma^2}{\sum_1^T (x_2 - \bar{x}_2)^2 (1 - r_{23}^2)}$$

$$r_{12} = \frac{\sum (x_2 - \bar{x}_2)(x_3 - \bar{x}_3)}{\sqrt{\sum (x_2 - \bar{x}_2)^2 \sum (x_3 - \bar{x}_3)^2}}$$

which is the correlation coefficient between x_2 and x_3

Econometrics V lecture 3

- I don’t expect you to remember that – but there are some things about it I do expect you to remember.
- Smaller variance is important. The smaller the parameter variance is the more likely the estimate is going to be close to its true value.
- If the variance is large (and since the test of significance of the coefficient is coefficient divided by se) then the estimated coefficient is not likely to be statistically significant.

Econometrics V lecture 3

- So its good to know what factors affect the size of the variance of the estimated coefficients.
- Lets list them – you keep looking back at the slide with the formulae

$$\hat{\text{var}}(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{1}^{T}(x_2 - \bar{x}_2)^2(1 - r_{23}^2)}$$

Econometrics V lecture 3

- The larger σ^2 the larger the variance of the least squares estimators. This is to be expected since σ^2 measures the overall uncertainty in the model specification. If σ^2 is large, then data values may be widely spread about the regression function

$$E[y_t] = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3}$$

- and there is less information in the data about the parameter values.

Econometrics V lecture 3

- The larger sample size *T* the smaller the variances. The sum in the denominator is

$$\sum_{t=1}^{T}(x_{t2}-\bar{x}_2)^2$$

- The larger is the sample size *T* the larger is this sum and thus the smaller is the variance.
- <u>More observations yield more precise parameter estimation.</u>

Econometrics V lecture 3

- In order to estimate β_2 precisely we would like there to be a large amount of variation in x_{t2}, $\sum_{t=1}^{T}(x_{t2} - \bar{x}_2)^2$. The intuition here is that it is easier to measure β_2, the change in *y* we expect given a change in x_2, the more sample variation (change) in the values of x_2 that we observe.

- In the denominator of var(b_2) is the term where r_{23} is the correlation between the sample values of x_{t2} and x_{t3}. Recall that the correlation coefficient measures the linear association between two variables. If the values of x_{t2} and x_{t3} are correlated then

$$1 - r_{23}^2$$

- is a fraction that is less than 1.

Econometrics V lecture 3

- The larger the correlation between x_{t2} and x_{t3} the larger is the variance of the least squares estimator b_2. The reason for this fact is that variation in x_{t2} adds most to the precision of estimation when it is not connected to variation in the other explanatory variables.
- If the two variables are linearly related the correlation coefficient = 1 and ols fails. This is perfect multicollinearity

Econometrics V lecture 3

- So just as with simple regression the variances and covariance's between the estd parameters will be of interest and we could write formulae down for them.
- We will not be writing them down but the computer will be calculating them.
- Once again we can get the computer to print the variance/covariance matrix for the estimated coefficients.

Econometrics V lecture 3

$$y=\beta 0+\beta_1+\beta_2+\beta_3+\beta_4+e$$

	b_o	b_1	b_2	b_3	b_4
b_o	var(bo)				
b_1					
b_2					
b_3					
b_4		$cov(b_4,b_1)$			

Econometrics V lecture 3

- Now the next bits are easy.
- If the error term is assumed normally distributed, then just as in the simple regression case:

$$t = \frac{\hat{\beta}_k - \beta_k}{se(\hat{\beta}_k)} \sim t_{T-k}$$

and

$$\hat{\beta}_k \pm t_c.se(\hat{\beta}_k)$$

gives us a confidence interval

Econometrics V lecture 3

- R^2 and R(bar)^2
- How is it calculated
- Why is that not much help in the multiple regression case.
- If we have T-1 variables $R^2 = 1$

$$\bar{R}^2 = 1 - \frac{SSE/(T-k)}{SST(T-1)}$$

- Now no longer % total variation explained
- Think of it as fit
- Uses – and abuses – don't use to pick variables for inclusion

Econometrics V lecture 3

- We have used our t test to test simple hypothesis about single coefficients.
- It can be used for slightly more complex hypothesis.
- Example
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$
- Ho : = 4
- Ha : $\neq$ 4
- Is straight forward

Econometrics V lecture 3

- But what about
- Ho : $\beta_1 = 2\beta_2$

- This can be done in the usual way but will need some manipulation

- How about
- Ho : $\beta_1 + \beta_2 = 4$

Econometrics V lecture 3

- What about
- Ho : $\beta_2 = \beta_3 = 0$
- Can we do this using a t test – how about 2 t tests?
- If we cannot reject Ho $\beta_2 = 0$ and we cannot reject Ho $\beta_3 = 0$ then we cannot reject the above?
- Not necessarily – the estimators of β_2 and β_3 are correlated – a test of joint hypothesis such as the above should take this into account
- Think of example of equation with multicollinearity??????
- Effects of mutlicollinearity – signs of multicollinearity….

Econometrics V lecture 3

- So rule of thumb
- If there is a single = sign it can be done as a t test (you need the VCV matrix – if this is not there take that as a hint that you have to do it a different way)
- If there is more than 1 equals sign (multiple restrictions) it cannot be done as a t test – we will look in a future slide at the F test which is how we will do it.

- Multiple Regression Analysis

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

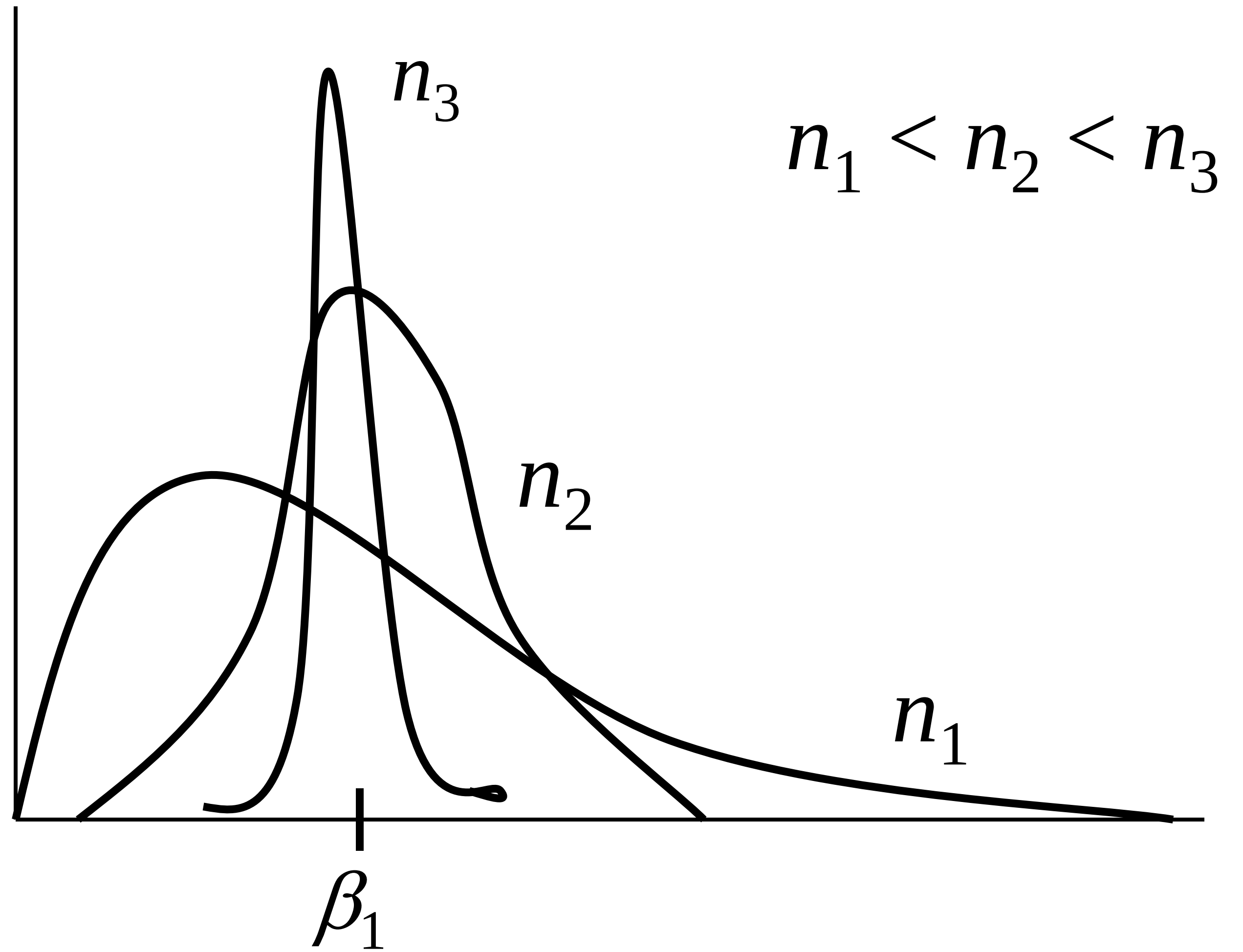
- 3. Asymptotic Properties

Econometrics V lecture 3

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

Econometrics V lecture 3
Sampling Distributions as *n* ↑



Econometrics V lecture 3

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

Econometrics V lecture 3
Proving Consistency

$$\hat{\beta}_1 = \left(\sum (x_{i1} - \bar{x}_1) y_i\right) / \left(\sum (x_{i1} - \bar{x}_1)^2\right)$$

$$= \beta_1 + \left(n^{-1} \sum (x_{i1} - \bar{x}_1) u_i\right) / \left(n^{-1} \sum (x_{i1} - \bar{x}_1)^2\right)$$

$$\text{plim} \hat{\beta}_1 = \beta_1 + Cov(x_1, u) / Var(x_1) = \beta_1$$

$$\text{because } Cov(x_1, u) = 0$$

Econometrics V lecture 3

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!

Econometrics V lecture 3

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

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

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

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

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

Econometrics V lecture 3

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

Econometrics V lecture 3

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

Econometrics V lecture 3

- If there are *n* data points to estimate parameters of both models from, then one can calculate the *F* statistic, given by

$$F = \frac{\left(\dfrac{RSS_1 - RSS_2}{p2 - p1}\right)}{\dfrac{RSS_2}{n - p2}}$$

- where RSS_i is the residual sum of squares of model *i*.

Econometrics V lecture 3

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 $P(Z<z) \to \Phi(z)$ as $n \to \infty$, 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 $\sim N(0,1)$, or

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

Econometrics V lecture 3
Asymptotic Normality

Under the Gauss-Markov assumptions,

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

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

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

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

Econometrics V lecture 3

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 $\sqrt{n}$

Econometrics V lecture 3

Lagrange Multiplier statistic

- With large samples, by relying on asymptotic normality for inference, we can use more than *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

Econometrics V lecture 3

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

Econometrics V lecture 3

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

Multiple Regression Analysis

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

Econometrics V lecture 4
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

Econometrics V lecture 4

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

Econometrics V lecture 3

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*

Econometrics V lecture 3

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

Econometrics V lecture 3

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

Econometrics V lecture 3

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

Econometrics V lecture 3

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

Econometrics V lecture 3

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

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

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

Econometrics V lecture 3
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 $x^* = \left|\hat{\beta}_1 / \left(2\hat{\beta}_2\right)\right|$

Econometrics V lecture 3

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*

$$\text{For } \hat{\beta}_1 < 0 \text{ and } \hat{\beta}_2 > 0 \text{ the turning point}$$

$$\text{will be at } x^* = \left| \hat{\beta}_1 / \left(2\hat{\beta}_2 \right) \right|, \text{ which is}$$

$$\text{the same as when } \hat{\beta}_1 > 0 \text{ and } \hat{\beta}_2 < 0$$

Econometrics V lecture 3

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

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

$$\text{the effect of } x_1 \text{ on } y \text{ we typically}$$

$$\text{evaluate the above at } \bar{x}_2$$

Econometrics V lecture 3

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

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

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

Econometrics V lecture 3

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. $\ln(y)$)

Econometrics V lecture 3

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

Econometrics V lecture 3

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

Econometrics V lecture 3

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

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

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

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

Econometrics V lecture 3
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

Econometrics V lecture 3
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

Econometrics V lecture 3

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*)

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

In this case can predict y as follows

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

Econometrics V lecture 3

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 <u>no intercept</u>
- 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*

Econometrics V lecture 3

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