# Review of Estimation Principles for the Regression Model



## The Multiple Regression Model

- The multiple regression model has the following assumptions
  - The dependent variable is a linear function of the explanatory variables (A1)
  - The errors have a mean of zero (A2)
  - The errors have a constant variance (A3)
  - The errors are uncorrelated across observations (A3)
  - The error term is not correlated with any of the explanatory variables (A4)
  - No explanatory variable is an exact linear function of other explanatory variables (A5)
  - The errors are drawn from a normal distribution (A6)

# Estimation (Three Ways)

- Parameter estimates for $\beta_0$, $\beta_1$, $\beta_2$, ..., $\beta_k$ can be derived using
  - Ordinary least squares
  - Method of moments
  - Method of maximum likelihood
- All lead to the same estimators



# Estimation Principles

- Method of moments
  - equate the moments (mean, variance, *etc.)* implied by a statistical model of the population distribution with the actual moments observed in the sample and solve for the unknown parameter(s) $\theta$
- Method of maximum likelihood
  - data is drawn from some population distribution (*e.g.*, normal) with unknown parameter(s) $\theta$. Select estimate of $\theta$ so as to maximize the likelihood of seeing the sample actually observed
- Method of least squares
  - select the best estimator by minimizing the sum of the squared errors for the whole sample

# Ordinary Least Squares

- Select $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_1$, …, $\hat{\beta}_k$ to minimize

$$
\begin{aligned}
S &= \sum_{i=1}^{n} e_i^2 \\
&= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \\
&= \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} \ldots - \hat{\beta}_k X_{ki})^2
\end{aligned}
$$

  - Setting the partial derivatives supplies us with k + 1 normal equations



# Method of Moments: 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 $b_0$ to normalize E(*u*) to 0

## Method of Moments: 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
- $E(y|x) = b_0 + b_1x$



## Ordinary Least Squares

- Basic idea of regression is to estimate the population parameters from a sample
- Let $\{(x_i,y_i): i=1, ...,n\}$ denote a random sample of size $n$ from the population
- For each observation in this sample, it will be the case that
- $y_i = b_0 + b_1x_i + u_i$

# 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
- $Cov(x,u) = E(xu) = 0$
- Why? Remember from basic probability that $Cov(X,Y) = E(XY) – E(X)E(Y)$
  - If independent, $E(XY)=E(X)E(Y)$ and
  - In the case of $E(xu)$ ➔ $E(u)=0$ so $E(x)\,E(u)=0$



# Deriving OLS continued

- We can write our 2 restrictions just in terms of $x$, $y$, $b_0$ and $b_1$ , since $u = y – b_0 – b_1x$
- $E(y – b_0 – b_1x) = 0$
- $E[x(y – b_0 – b_1x)] = 0$
- These are called moment restrictions

## Deriving OLS using Method of Moments Principle

- 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



## Method of Moments 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:

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

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

## MoM Derivation of OLS

- Given the definition of a sample mean, and properties of summation (e.g., S*k*=*nk*), we can rewrite the first condition as follows

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

or

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



## MoM Derivation of OLS

(insert previous result in place of $\hat{b}_0$ in second moment equation)

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

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

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

## So the OLS estimated slope is

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

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



## Alternate approach to derivation: least squares

- 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:

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

## 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*

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

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



## Method of Moments Derivation of OLS – Matrix Version

- Shortest way to derive, and focuses on the important assumption

$$y = Xb + u \quad and \quad E(Xu) = 0$$

$$\text{replace with sample values } E[X(y - X\hat{b})] = 0$$

$$E[Xy - X'X\hat{b}] = 0 \quad \Rightarrow \quad Xy = X'X\hat{b}$$

$$\hat{b} = (X'X)^{-1}X'y$$

## Proof of unbiasedness of the OLS estimator – matrix version

Use the assumption that:

E($u$|$x$) = E($xu$)= E($u$) = 0

And note that

(X'X)$^{-1}$X'X=Identity matrix

$$\hat{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$$
$$= (\mathbf{X'X})^{-1}\mathbf{X'}(\mathbf{X}ß + u)$$
$$= b + (\mathbf{X'X})^{-1}\mathbf{X'}u$$
$$E(\hat{b}) = E\left[b + (\mathbf{X'X})^{-1}\mathbf{X'}u\right]$$
$$= b + (\mathbf{X'X})^{-1}\mathbf{X'}E(u)$$
$$= b + 0$$
$$= b$$



## Unbiasedness Summary

- The OLS estimates of $b_1$ and $b_0$ are unbiased
- Proof of unbiasedness depends on our assumptions about $u$ – 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, which is why the variance of an estimator also matters

## Variance of the OLS Estimators

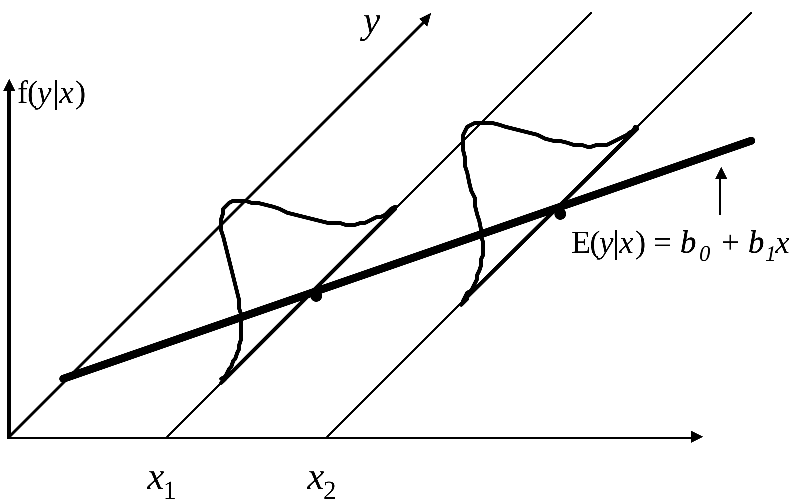
- Now we know that the sampling distribution of our estimate is centered around the true parameter
- Want to think about how spread out this distribution is
- Much easier to think about this variance under an additional assumption, so
- Assume Var($u|x$) = $s^2$



## Variance of OLS (cont)

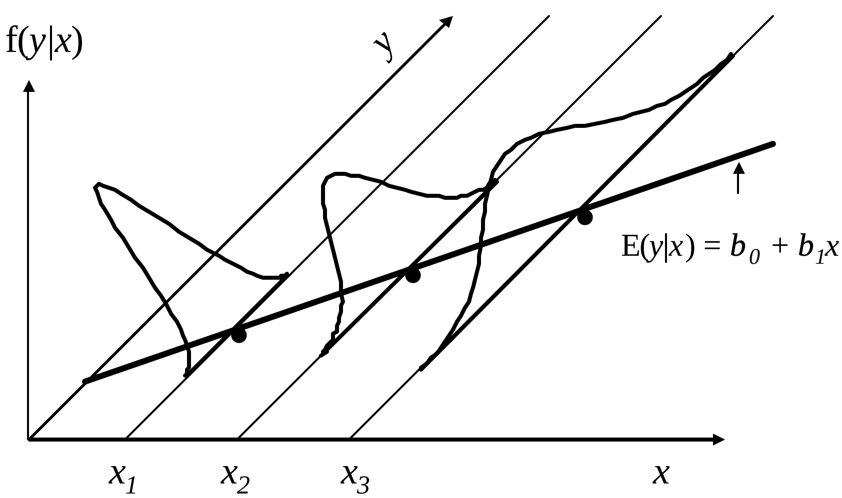
- Var($u|x$) = E($u^2|x$) - [E($u|x$)]$^2$
- E($u|x$) = 0, so $s^2$ = E($u^2|x$) = E($u^2$) = Var($u$)
- Thus $s^2$ is also the unconditional variance, called the error variance
- $s$, the square root of the error variance is called the standard deviation of the error
- Can say: E($y|x$)= $b_0$ + $b_1x$ and Var($y|x$)

## Homoskedastic Case






## Heteroskedastic Case



# Variance of OLS

$$Var(\hat{b}) \equiv E\left\{\left[\hat{b} - E(\hat{b})\right]\left[\hat{b} - E(\hat{b})\right]'\right\}$$

Using the fact that OLS is unbiased $E(\hat{b}) = b$

$$= E\left[(\hat{b} - b)(\hat{b} - b)'\right]$$

From the unbiasedness proof $\hat{b} - b = (\mathbf{X'X})^{-1}\mathbf{X'}u$

$$Var(\hat{b}) = \left[(\mathbf{X'X})^{-1}\mathbf{X'}uu'\mathbf{X}(\mathbf{X'X})^{-1}\right]$$

With homoscedasticity $E[uu'] = s^2\mathbf{I}_n$

$$Var(\hat{b}) = s^2\left[(\mathbf{X'X})^{-1}\mathbf{X'I}_n\mathbf{X}(\mathbf{X'X})^{-1}\right]$$

$$Var(\hat{b}) = s^2(\mathbf{X'X})^{-1}$$



# Variance of OLS Summary

- The larger the error variance, $s^2$, 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

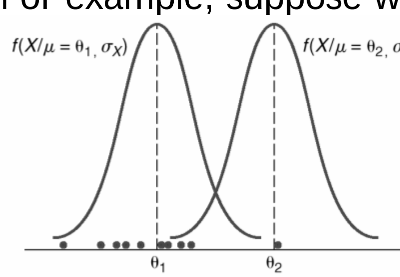
# Estimating the Error Variance

- We don't know what the error variance, $s^2$, 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
  - Assumption made about homoscedasticity can then matter a great deal (next week)



# Method of Maximum Likelihood

- Select estimate of unknown parameters so as to maximize the likelihood of seeing the sample actually observed
- For example, suppose we assume our data is




The horizontal axis shows the actual data drawn in the sample. These data are more likely to arise if the population distribution is as characterized by the distribution on the left.

# Why is Maximum Likelihood Estimation Useful?

- Many problems in microeconometrics have data that do not obey classical assumptions
  - Integer or dichotomous dependent variables
  - Censored or truncated variables
  - Stronger distributional assumptions are needed to estimate parameters of the underlying (latent) model from such data
    - Assumption of normality and use of MLE helps
- MLE are asymptotically efficient
  - Consistent, asymptotically normal and covariance matrix is “smaller” than cov matrix of any other consistent, asy-Normal estimator
  - So if can show that an estimator is an MLE or has the properties of an MLE, it is also showing that this estimator is the most efficient



# Maximum Likelihood

- If assumption A6 holds then
  - $Y_i \sim N(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki}, \sigma_\varepsilon^2)$
- The observations would have a likelihood function of

$$L(b_0, b_1, \ldots, b_k, s_e^2) = f(y_1) f(y_2) f(y_3) \ldots f(y_n)$$

- $Y_i \sim N(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki}, \sigma_\varepsilon^2)$

implies

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon}} e^{\left\{-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - \beta_3 X_{3i} - \ldots - \beta_k X_{ki}}{\sigma_\varepsilon}\right)^2\right\}}$$

# Maximum Likelihood

$$L(\beta_0, \beta_1, \sigma_\varepsilon^2) = f(y_1)f(y_2)f(y_3)\ldots f(y_n)$$

$$= \frac{1}{\sqrt{2\pi\sigma_\varepsilon}} e^{\left\{-\frac{1}{2}\left(\frac{Y_1-\beta_0-\beta_1X_{11}-\beta_2X_{21}-\beta_3X_{31}-\ldots-\beta_kX_{k1}}{\sigma_\varepsilon}\right)^2\right\}}$$

$$\cdot \frac{1}{\sqrt{2\pi\sigma_\varepsilon}} e^{\left\{-\frac{1}{2}\left(\frac{Y_2-\beta_0-\beta_1X_{12}-\beta_2X_{22}-\beta_3X_{32}-\ldots-\beta_kX_{k2}}{\sigma_\varepsilon}\right)^2\right\}} \cdot$$

$$\ldots \cdot \frac{1}{\sqrt{2\pi\sigma_\varepsilon}} e^{\left\{-\frac{1}{2}\left(\frac{Y_n-\beta_0-\beta_1X_{1n}-\beta_2X_{2n}-\beta_3X_{3n}-\ldots-\beta_kX_{kn}}{\sigma_\varepsilon}\right)^2\right\}}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma_\varepsilon}}\right)^n e^{\left\{\frac{-1}{2\sigma_\varepsilon}\left(\sum_{i=1}^{n} Y_i-\beta_0-\beta_1X_{1i}-\beta_2X_{2i}-\beta_3X_{3i}-\ldots-\beta_kX_{ki}\right)^2\right\}}$$



# Maximum Likelihood

- Taking logs yields

$$L(b_0, b_1, \ldots, b_k, s_e^2)$$

$$= n \ln\left(\frac{1}{\sqrt{2\pi\sigma_\varepsilon}}\right)\frac{-1}{2\sigma_\varepsilon}\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1X_{1i}-\beta_2X_{2i}-\beta_3X_{3i}-\ldots-\beta_kX_{ki})^2$$

  - Taking the derivative of the log likelihood function results in the same estimators as ordinary least squares