Econometrics V lecture 2

- WOOLDRIDGE CHAPTER TWO
- The simple regression model
  - Assumption 1 linearity in parameters

$$y = \beta_0 + \beta_1 x + u \qquad (1)$$

- Assumption 2 Random Sampling
  - We can use a random sample of size n (x,y);I = 1,2,3,..n) from the population model.
  - We can rewrite the previous equation

$$y = \beta_0 + \beta_1 x + u, \;\; i = 1, 2, \ldots n \qquad (2)$$

  - Where u_i contains the unobservables for observation i that affect y_i. Not to be confused with the residuals

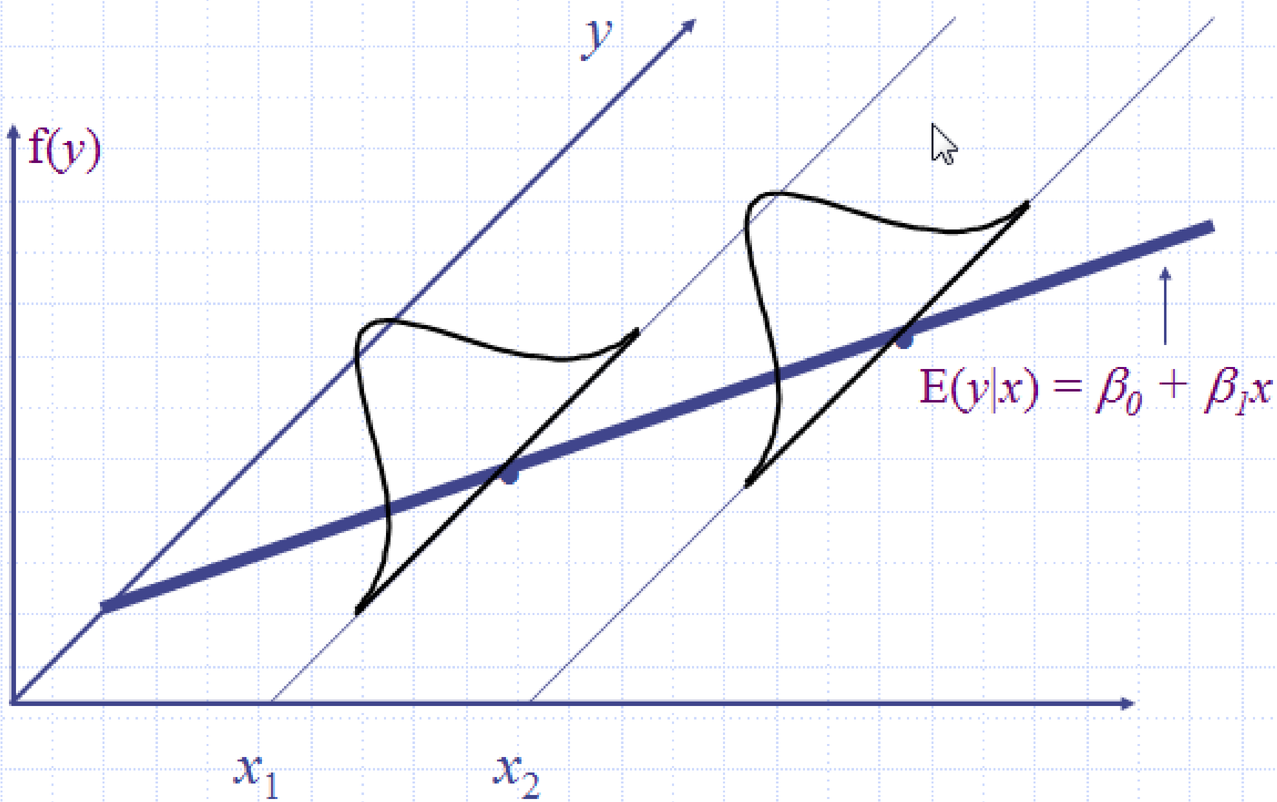
Econometrics V lecture 2

- Assumption 3 (Zero conditional mean)
- $E(u|x) = 0$
- For a random sample, this assumption implies that $E(u|x) = 0$, for all I = 1,2, …, n
- Technically, conditioning on the sample values of x, the independent variable is the same as treating the x_i as fixed in repeated sampling.

- 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
- $E(y|x) = \beta_0 + \beta_1 x$

Econometrics V lecture 2

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

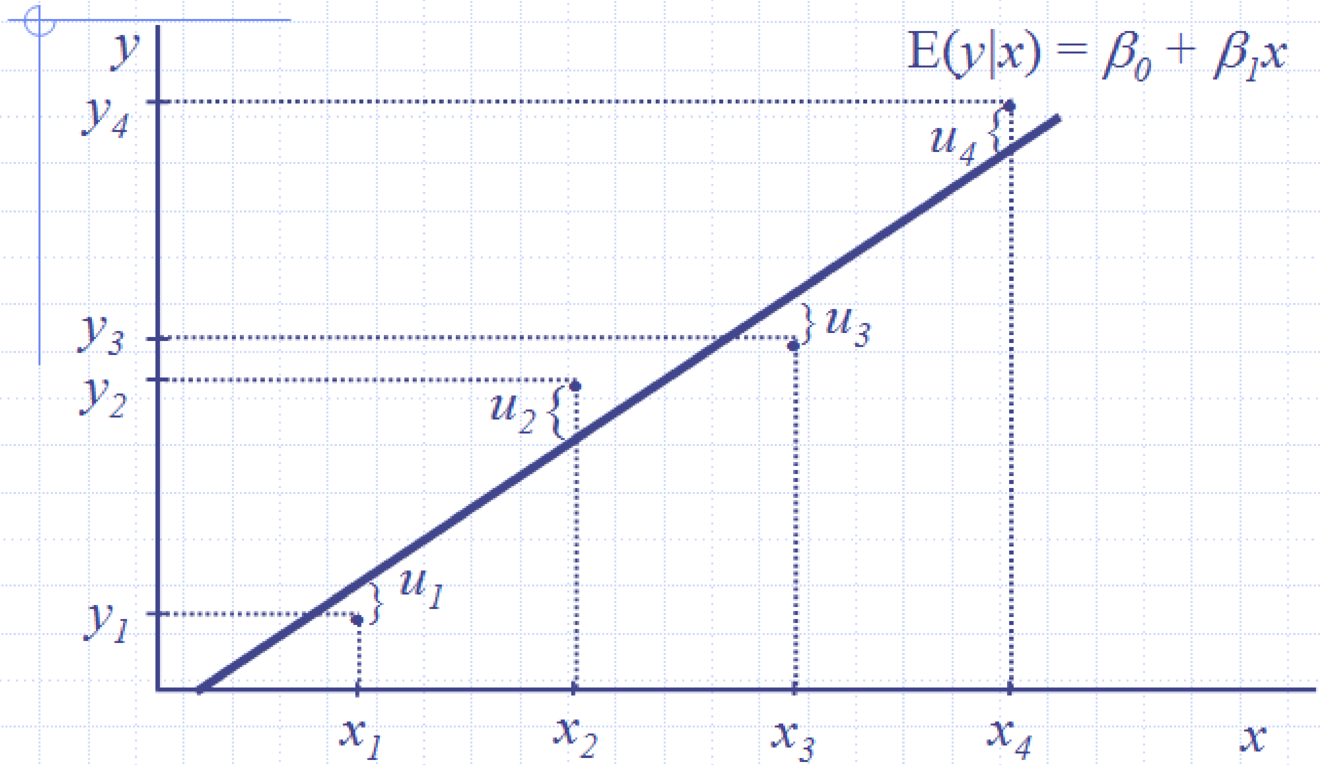


- Basic idea of regression is to estimate the population parameters from a sample
- Let $\{(x_i, y_i): i = 1, \ldots, n\}$ denote a random sample of size n from the population
- For each observation in this sample, it will be the case that
- $y_i = \beta_0 + \beta_1 x_i + u_i$

Edith Cowan University
Faculty of Business and Law

Econometrics V lecture 2

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
- $Cov(x,u) = E(xu) = 0$
- Why? Remember from basic probability that $Cov(X,Y) = E(XY) – E(X)E(Y)$

- 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

- 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

Econometrics V lecture 2

- 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:

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

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

- 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

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

Econometrics V lecture 2

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

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

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

Econometrics V lecture 2

- So the OLS estimated slope is

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

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

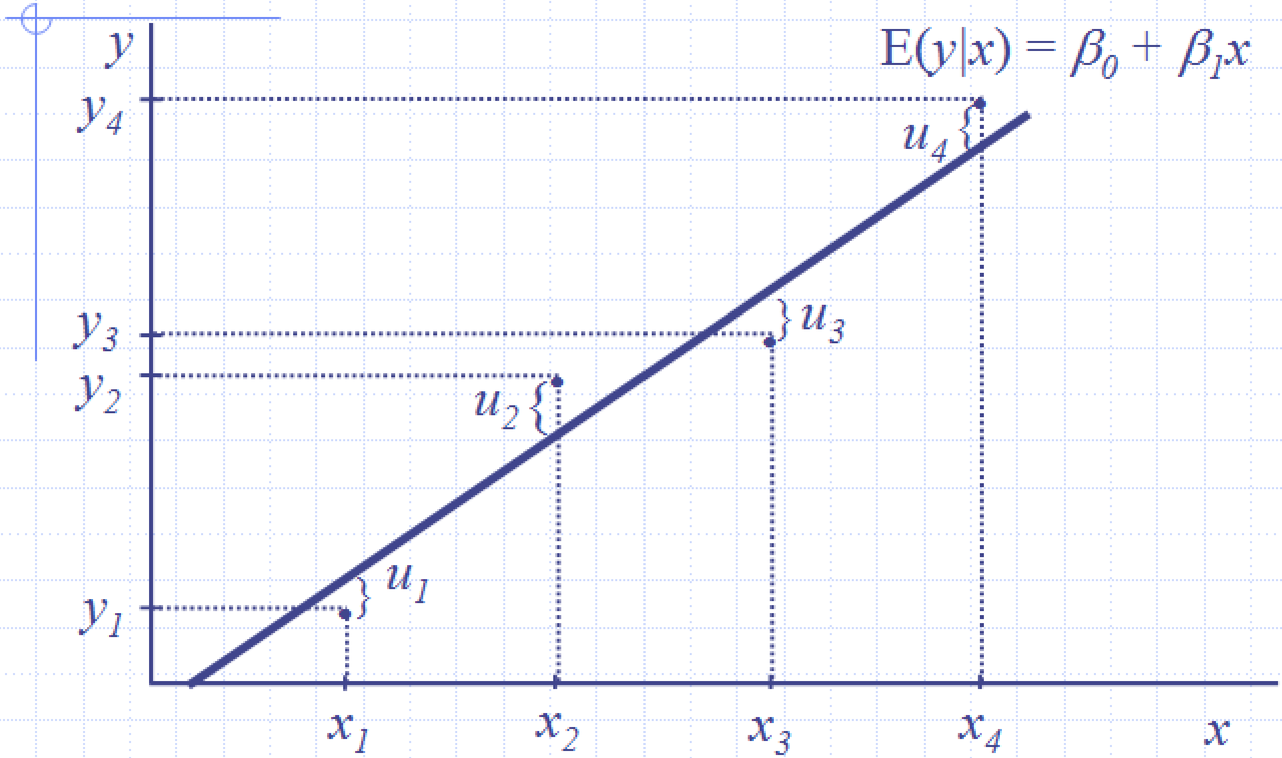
Econometrics V lecture 2

- 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

- 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

Econometrics V lecture 2

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



Econometrics V lecture 2

- 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:

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

Econometrics V lecture 2

- 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

Econometrics V lecture 2

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

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

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

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 $\text{SST} = \text{SSE} + \text{SSR}$

- Proof that SST = SSE + SSR

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

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

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

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

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

- 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

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

Econometrics V lecture 2

- 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

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

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

Econometrics V lecture 2

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

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

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

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

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

Econometrics V lecture 2

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

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

so, the numerator can be rewritten as

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

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

Econometrics V lecture 2

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

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

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

- 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

- 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 $\mathrm{Var}(u|x) = \sigma^2$ (Homoskedasticity)

- $\mathrm{Var}(u|x) = \mathrm{E}(u^2|x)-[\mathrm{E}(u|x)]^2$
- $\mathrm{E}(u|x) = 0$, so $\sigma^2 = \mathrm{E}(u^2|x) = \mathrm{E}(u^2) = \mathrm{Var}(u)$
- 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: $\mathrm{E}(y|x)=\beta_0 + \beta_1 x$ and $\mathrm{Var}(y|x) = \sigma^2$

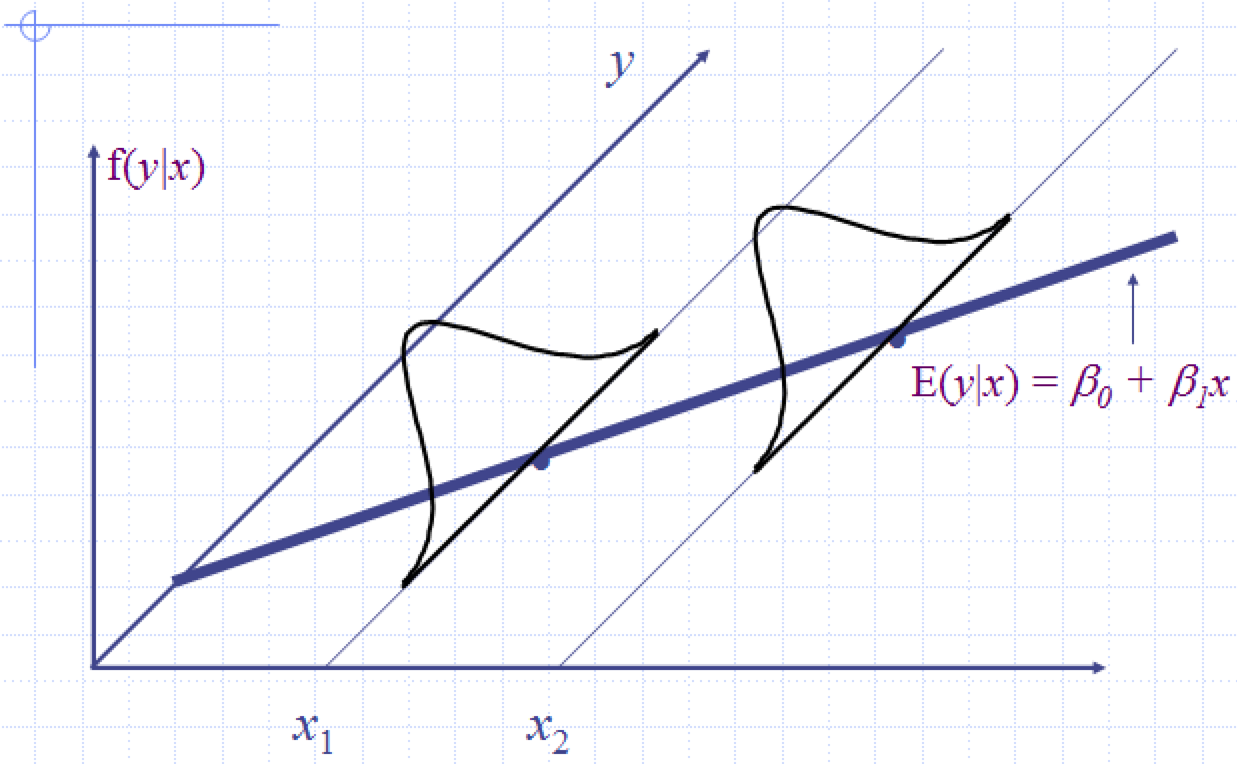
Econometrics V lecture 2

- Back to our Assumption 4
- Sample variation in the independent variable
  - The independent variables x_i, I = 1,2,…,n
  - Are not equal to the same constant. This requires some variation in x in the population to do our estimation.
  - This implies

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

Econometrics V lecture 2

Homoskedastic Case



Econometrics V lecture 2

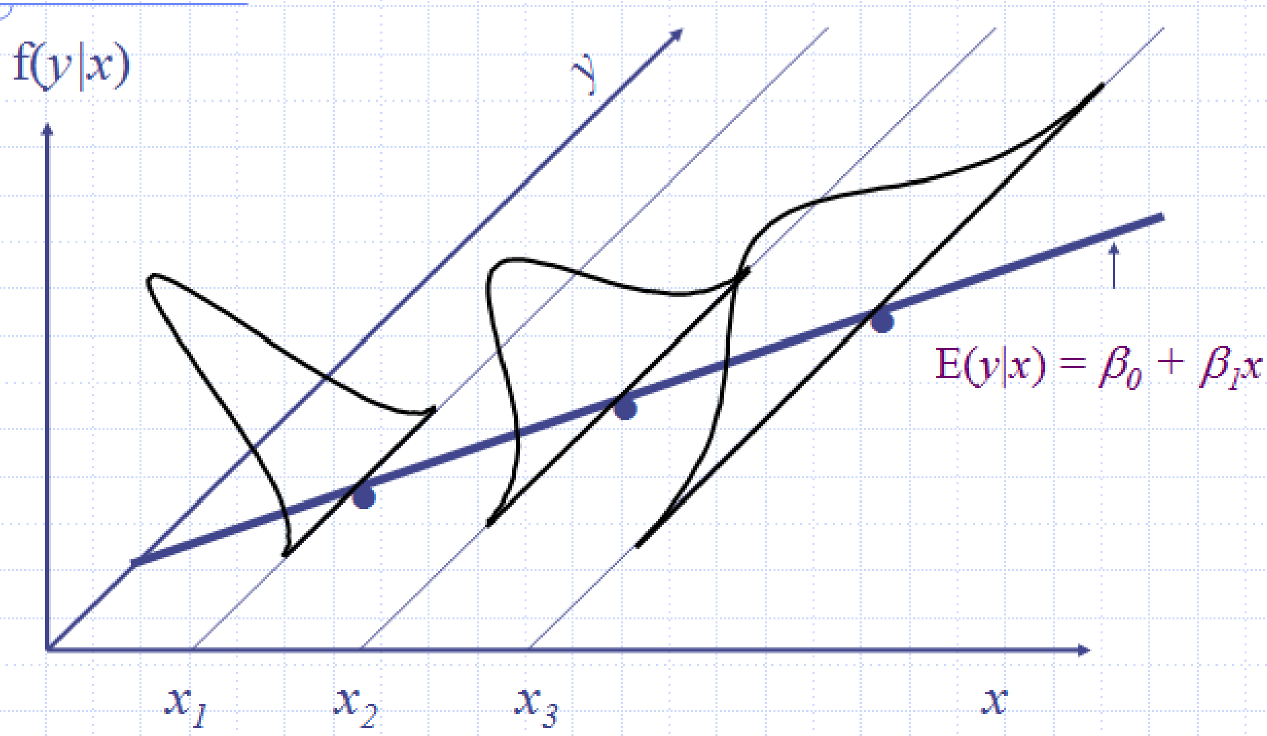
- Assumption 5 Homoscedasticity

u_i has the same variance at all values of *X*i.
this is referred to as “homoskedasticity”

ii. the predicted values of Y are equally good at all levels of X

Econometrics V lecture 2

Heteroskedastic Case



- Variance of OLS

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

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

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

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

- 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

- 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

Econometrics V lecture 2

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

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

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

Then, an unbiased estimator of σ^2 is

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

Econometrics V lecture 2

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

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

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

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

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