Econometrics Midterm Concepts

Ryan Safner

ECON 480

Ordinary Least Squares (OLS) Regression

- Bivariate data and associations between variables (e.g. X and Y)
  - Apparent relationships are best viewed by looking at a scatterplot
    - Check for associations to be positive/negative, weak/strong, linear/nonlinear, etc
    - Y: dependent variable
    - X: independent variable
  - Correlation coefficient (r) can quantify the strength of an association

$$r = \frac{1}{n-1} \sum^{n} \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right) = \frac{\sum^{n} Z_X Z_Y}{n-1}$$

    - $-1 \leq r \leq 1$ and r only measures *linear* associations
    - $|r|$ closer to 1 imply stronger correlation (near a perfect straight line)
    - Correlation does not imply causation! Might be confounding or lurking variables (e.g. Z) affecting X and/or Y

- Population regression model

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

  - β_1: $\frac{\Delta Y}{\Delta X}$: the slope between X and Y, number of units of Y from a 1 unit change in X
  - β_0 is the Y-intercept: literally, value of Y when $X = 0$
  - u_i is the error or residual, difference between actual value of $Y|X$ vs. predicted value of $\hat{Y}$

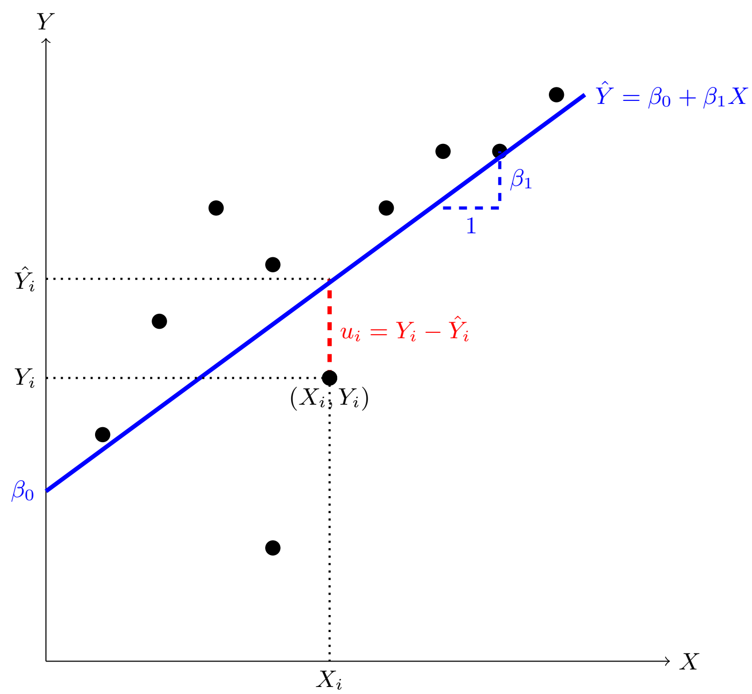
- Ordinary Least Squares (OLS) regression model

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

  - Least square estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ estimate population regression line from sample data
  - Minimize sum of squared errors (SSE) $min \sum^{n} u_i^2$ where $u_i = Y_i - \hat{Y}_i$
  - OLS regression line

$$\hat{\beta}_1 = \frac{cov(X,Y)}{var(X)} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = r_{X,Y} \frac{s_Y}{s_X}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$



Y
$\hat{Y} = \beta_0 + \beta_1 X$
β_1
1
$\hat{Y}_i$
$u_i = Y_i - \hat{Y}_i$
Y_i
(X_i, Y_i)
β_0
X
X_i

- Measures of Fit
  - R^2: fraction of total variation on Y explained by variation in X according to model

$$R^2 = \frac{ESS}{TSS}$$

$$R^2 = 1 - \frac{SSE}{TSS}$$

$$R^2 = r_{X,Y}^2$$

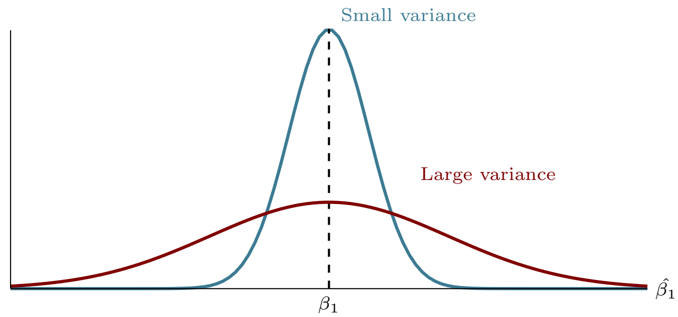
    - $ESS = \sum(\hat{Y}_i - \bar{Y})^2$
    - $TSS = \sum(Y_i - \bar{Y})^2$
    - $SSE = \sum \hat{u}_i{}^2$
  - Standard error of the regression (SER): average size of u_i, average distance from regression line to data points

$$SER = \frac{1}{n-2}\sum \hat{u}_i{}^2 = \frac{SSE}{n-2}$$

- Hypothesis testing of β_1
  - $H_0 : \beta_1 = \beta_{1,0}$, often $H_0 : \beta_1 = 0$
  - Two sided alternative $H_1 : \beta_1 \neq 0$
  - One sided alternatives $H_1 : \beta_1 > 0$, $H_2 : \beta_1 < 0$
  - t-statistic

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE[\hat{\beta}_1]}$$

  - Compare t against critical value $t*$, or compute p-value as usual
  - Confidence intervals (95%): $\hat{\beta}_1 \pm 1.96(SE[\hat{\beta}_1])$



$\hat{\beta}_1$ is a random variable, so it has its own sampling distribution with mean $E[\hat{\beta}_1]$ and standard error $se[\hat{\beta}_1]$

- Mean of OLS estimator $\hat{\beta}_1$ & Bias: Endogeneity & Exogeneity
  - X is **exogenous** if it is not correlated with the error term

$$corr(X, u) = 0$$

    - Equivalently, knowing X should not give you any information about u:

$$E[u|X] = 0$$

    - If X is exogenous, OLS estimate on X is unbiased:

$$E[\hat{\beta}_1] = \beta_1$$

- X is **endogenous** if it is correlated with the error term

$$corr(X, u) \neq 0$$

  * Equivalently, knowing X gives you information about u:

$$E[u|X] \neq 0$$

  * If X is endogenous, OLS estimate on X is biased:

$$E[\hat{\beta}_1] = \beta_1 + corr(X, u)\frac{\sigma_u}{\sigma_X}$$

    · Can measure strength and direction (+ or −) of bias
    · Note: if unbiased, $corr(X, u) = 0$, so $E[\hat{\beta}_1] = \beta_1$

- Variance of OLS estimator $\hat{\beta}_1$, measuring precision of estimate

$$var[\hat{\beta}_1] = \frac{\hat{\sigma}^2}{n \times var(X)}$$

and standard error

$$se[\hat{\beta}_1] = \sqrt{\frac{\hat{\sigma}^2}{n \times var(X)}}$$

  – Affected by 3 major factors:
    1. Model fit, where SER=$\hat{\sigma}$
    2. Sample size n
    3. Variation in X_j

- Heteroskedasticity and homoskedasticity
  – Homoskedastic errors (u) have the same variance over all values of X
  – Heteroskedastic errors (u) have different variance over values of X
    * Heteroskedasticity does *not* bias our estimates, but incorrectly lowers variance & standard errors (inflating t-statistics and significance!)
    * Can correct for heteroskedasticity by using robust standard errors