ECON 113 DI A/B

Dizhi Wang

OH : Nov 20th 10:30am – 12:30 pm E2 403B

Multi-variable Regression

- $y = \beta_0 + \beta_1 x_1 + \beta_2 \log(x_2) + u$

- How would you interpret β_1?
  - A one unit increase in x_1 is associated with a β_1 units increase of y, **keeping all else fixed**.

- How would you interpret β_2?
  - A 1% increase in x_2 is associated with an increase of $\frac{\beta_2}{100}$ units in y, **keeping all else equal**.

Multi-variable Regression

- Why do we want to do a multi-variable regression?
  - **Adding controls to avoid OVB:**
    - $E[u|x_1] \neq 0$ but $E[v|x_1, x_2] = 0$

Multi-variable Regression

- Why do we want to do a multi-variable regression?

  - **Adding controls to avoid OVB:**
    - $E[u|x_1] \neq 0$ but $E[v|x_1, x_2] = 0$

  - OVB **only** caused by confounding factor:
    - Confounding factor: some factors x_i that is
      - affecting y
      - Correlated with the original regressor x_1

  - If there are some factors x_i that affect y but $cov(x_1, x_i) = 0$
    - it won't cause OVB ⇒ The estimation of the coefficient in front of x_1 ($\widehat{\beta_1}$) will be unbiased
    - However, the standard error of the estimation $se(\widehat{\beta_1})$ will be influenced

  - If there are some factors x_i that doesn't affect y
    - it won't cause OVB
    - If $cov(x_1, x_i) = 0$
      - standard error of the estimation $se(\widehat{\beta_1})$ will not be influenced
    - If $cov(x_1, x_i) \neq 0$
      - standard error of the estimation $se(\widehat{\beta_1})$ will be influence

Multi-variable Regression

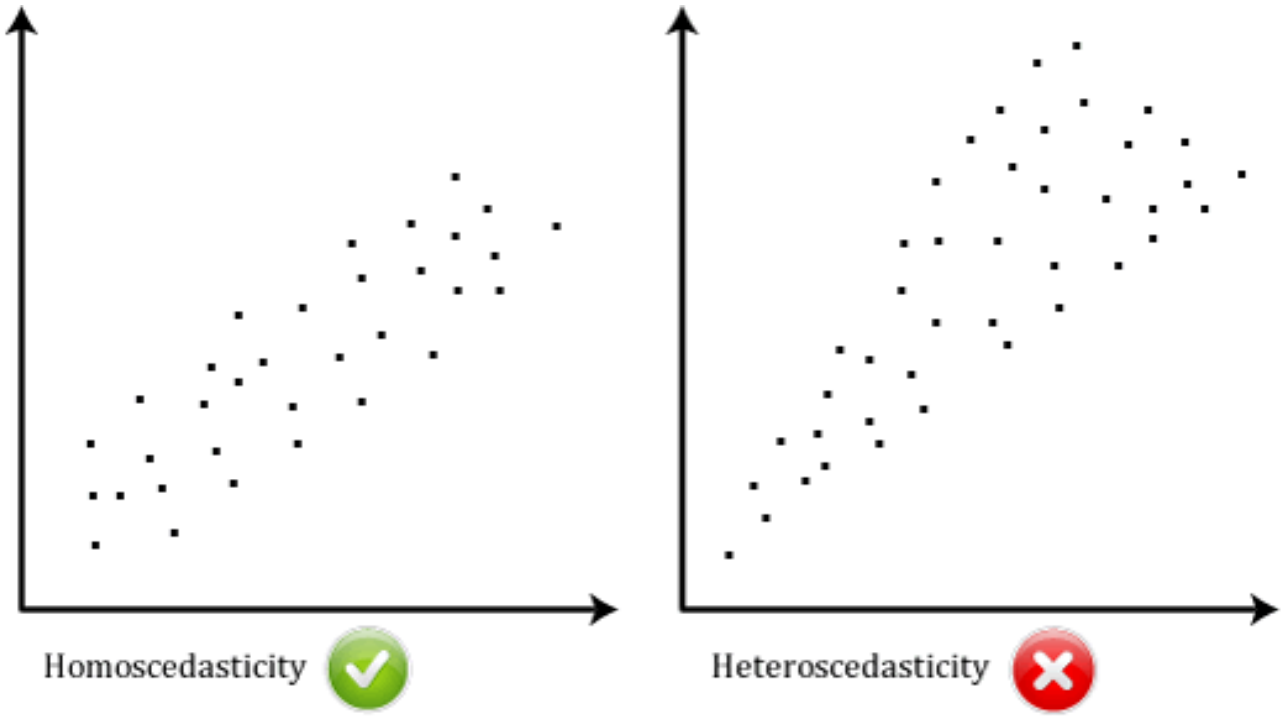
- How do confounding factors cause OVB?
  - Suppose the true model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$
  - Which means $E[u|x_1, x_2] = 0$
    - $E[u|x_1] = E[E[u|x_1, x_2]|x_1] = 0 = E[u|x_1] = E[u]$
  - $E[\widehat{\beta_1}] = E\left[\frac{\sum_i (x_{i,1} - \overline{x_1})(y_i - \bar{y})}{\sum_i (x_{i,1} - \bar{x})^2}\right] = E\left[\frac{\sum_i (x_{i,1} - \overline{x_1})y_i}{\sum_i (x_{i,1} - \bar{x})^2}\right] = E\left[\frac{\sum_i (x_{i,1} - \overline{x_1})(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + u_i)}{\sum_i (x_{i,1} - \bar{x})^2}\right]$
  - $E[\widehat{\beta_1}] = \beta_1 E\left[\frac{\sum_i (x_{i,1} - \overline{x_1})x_{i,1}}{\sum_i (x_{i,1} - \bar{x})^2}\right] + \beta_2 E\left[\frac{\sum_i (x_{i,1} - \overline{x_1})x_{i,2}}{\sum_i (x_{i,1} - \bar{x})^2}\right] + E[\frac{\sum_i (x_{i,1} - \overline{x_1})u}{\sum_i (x_{i,1} - \bar{x})^2}]$
  - $E[\widehat{\beta_1}] = \beta_1 E\left[\frac{\sum_i (x_{i,1} - \overline{x_1})(x_{i,1} - \overline{x_1})}{\sum_i (x_{i,1} - \bar{x})^2}\right] + \beta_2 E\left[\frac{\sum_i (x_{i,1} - \overline{x_1})(x_{i,2} - \overline{x_2})}{\sum_i (x_{i,1} - \bar{x})^2}\right] + E[\frac{\sum_i (x_{i,1} - \overline{x_1})u_i}{\sum_i (x_{i,1} - \bar{x})^2}]$
  - $E[\widehat{\beta_1}] = \beta_1 + \beta_2 \frac{cov(x_1, x_2)}{Var(x_1)}$ when the number of observations is large enough
  - When $cov(x_1, x_2) = 0 \Rightarrow Unbiased$
  - When $cov(x_1, x_2) \neq 0 \Rightarrow Biased$

- Now we have taken fully care of $E(\widehat{\beta_1})$
  - $E(\widehat{\beta_1})$ is related to biasedness of our estimator

- But there is some extra complication related to $Var(\widehat{\beta_1})$
  - We care about $\mathrm{Var}(\hat{\beta}_1)$ because it will determine the precision of our estimation
  - which affects the width of confidence intervals and the power of hypothesis tests

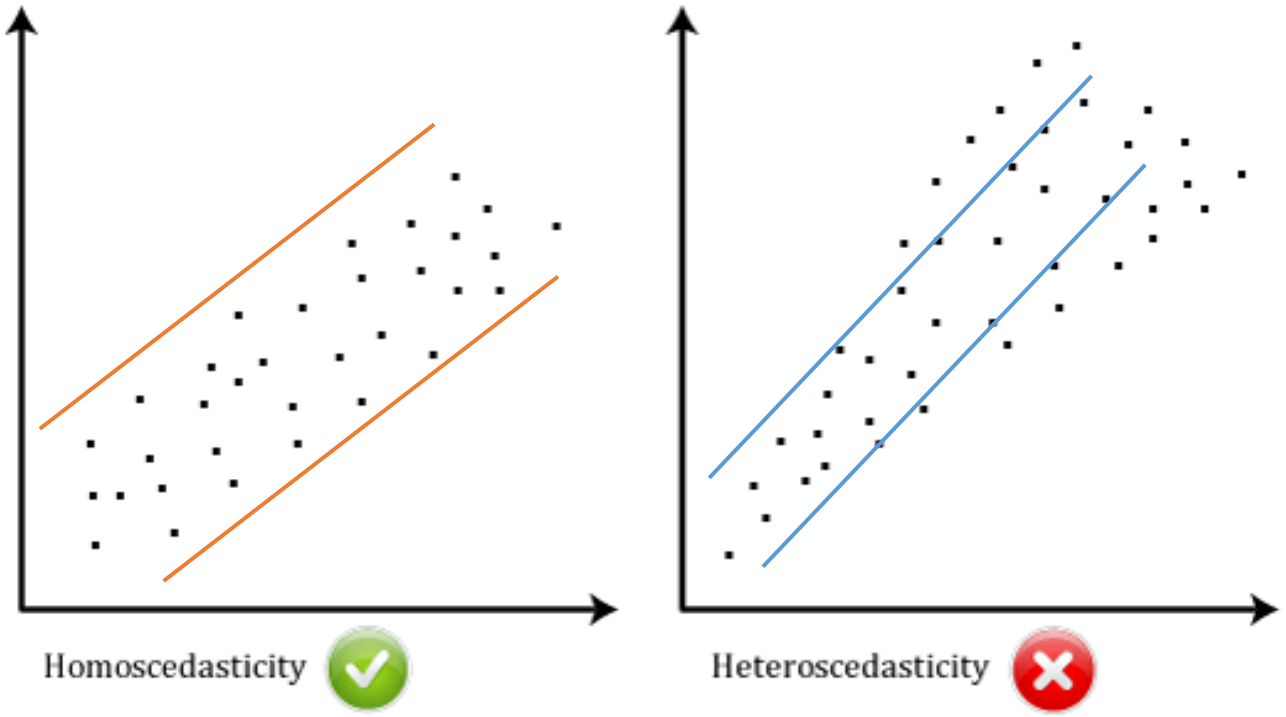
Why $Var(\widehat{\beta_1})$ matters

- Each observation is a point
  - Our estimator will give the line that fits those points as much as possible
  - But it will never fit all points perfectly



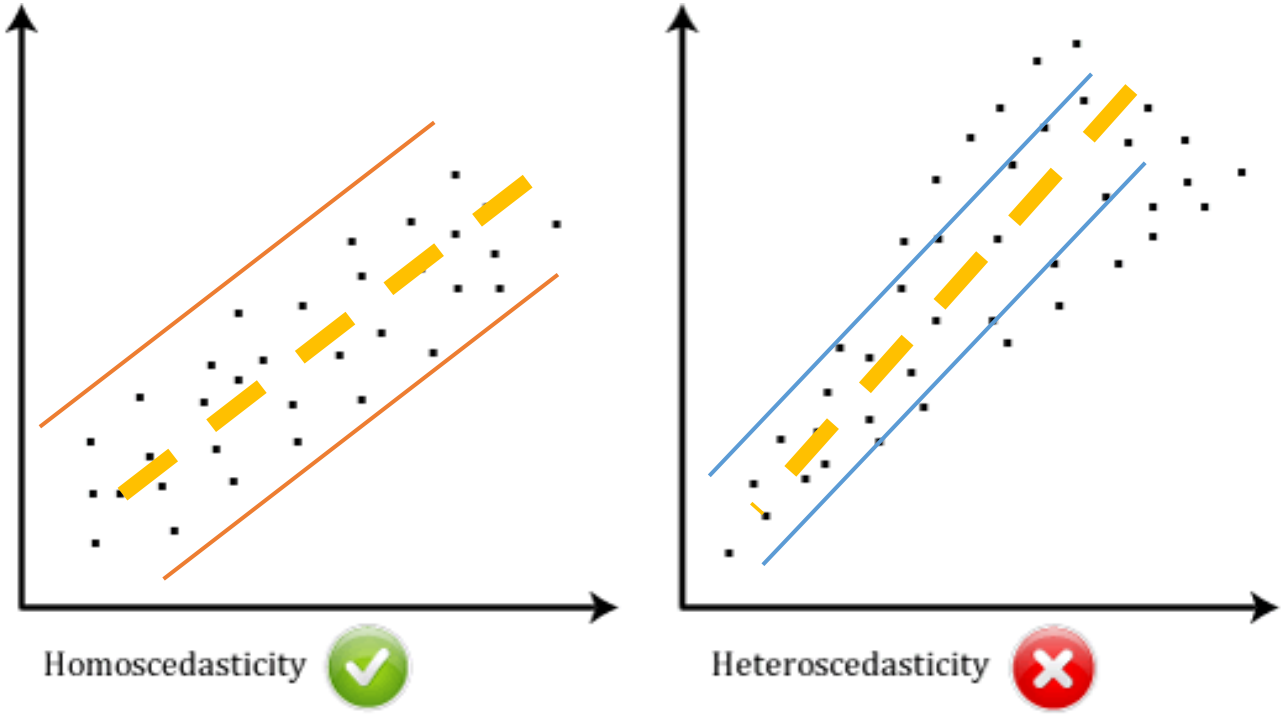
Why $Var(\widehat{\beta_1})$ matters

- Each observation is a point
  - Our estimator will give the line that fits those points as much as possible
  - But it will never fit all points perfectly
- We need a confidence interval



Homoskedasticity assumption

- Given $y = \beta_0 + \beta_1 x_1 + u$
- Homoskedasticity assumption requires: $Var(u|X) = \sigma^2$
  - Conditional variance of u on the regressor X must be constant
  - We can have an unbiased estimator ($E[u|X] = 0$) but violate the homoskedasticity assumption (wrong std err for the estimation)



Conditional Expectation and Variance

- $E[XY|X] = ?$

Conditional Expectation and Variance

- $E[XY|X] = ?$
- Because it is conditional on X, we can treat X as known
  - $E[XY|X] = XE[X|Y]$

Conditional Expectation and Variance

- $E[XY|X] = ?$
- Because it is conditional on X, we can treat X as known
  - $E[XY|X] = XE[X|Y]$

- $Var[XY|X] = ?$

Conditional Expectation and Variance

- $E[XY|X] = ?$
- Because it is conditional on X, we can treat X as known
  - $E[XY|X] = XE[X|Y]$

- $Var[XY|X] = ?$
  - $Var[XY|X] = X^2 Var[Y|X]$

Conditional Expectation and Variance

- $E[XY|X] = ?$
- Because it is conditional on X, we can treat X as known
  - $E[XY|X] = XE[X|Y]$

- $Var[XY|X] = ?$
  - $Var[XY|X] = X^2 Var[Y|X]$

- $Var[Y|X] = E[Y^2|X] - E[Y|X]^2$
  - Recall $Var[Y] = E[Y^2] - E[Y]^2$

Violation of Homoskedasticity assumption

- $Price_i = \beta_0 + \beta_1 Size_i + u_i$
- $u_i = Size_i \times e_i$, where e_i is a r.v
  - $Var(e_i) = \sigma^2$
  - $E[e_i|Size_i] = 0$
- What is $E[u_i|Size_i]$?

Violation of Homoskedasticity assumption

- $Price_i = \beta_0 + \beta_1 Size_i + u_i$
- $u_i = Size_i \times e_i$, where e_i is a r.v
  - $Var(e_i) = \sigma^2$
  - $E[e_i|Size_i] = 0 \Rightarrow e_i \perp Size_i \Rightarrow \mathrm{E[e_i] = E[e_i|Size_i]}$
- What is $E[u_i|Size_i]$?

Violation of Homoskedasticity assumption

- $Price_i = \beta_0 + \beta_1 Size_i + u_i$
- $u_i = Size_i \times e_i$, where e_i is a r.v
  - $Var(e_i) = \sigma^2$
  - $E[e_i|Size_i] = 0$
- What is $E[u_i|Size_i]$?
  - $E[u_i|Size_i] = E[Size_i \times e_i|Size_i] = Size_i E[e_i|Size_i] = 0$
  - Independence assumption is good
- How about $Var[u_i|Size_i]$?

Violation of Homoskedasticity assumption

- $Price_i = \beta_0 + \beta_1 Size_i + u_i$
- $u_i = Size_i \times e_i$, where e_i is a r.v
  - $Var(e_i) = \sigma^2$
  - $E[e_i|Size_i] = 0 \Rightarrow \mathrm{e_i} \perp Size_i$ and $\mathrm{E[e_i] = E[E[e_i|Size_i]] = 0}$
- What is $E[u_i|Size_i]$?
  - $E[u_i|Size_i] = E[Size_i \times e_i|Size_i] = Size_i E[e_i|Size_i] = 0$
  - Independence assumption is good
- How about $Var[u_i|Size_i]$?
  - $Var[u_i|Size_i] = \mathrm{Var}[Size_i \times e_i|Size_i] = Size_i^2 Var[e_i|Size_i]$
  - $Var[e_i|Size_i] = E\left[e_i^2\middle|Size_i\right] - E[e_i|Size_i]^2 = E\left[e_i^2\middle|Size_i\right] = E[e_i^2]$
  - $Var[u_i|Size_i] = Size_i^2 \sigma^2$
  - Which is no a constant. So homoskedasticity assumption is violated

Homoskedasticity & Std Err

- No matter the homoskedasticity assumption is violated or not
- We can calculate the "confidence interval" of our estimated $\widehat{\beta_1}$
  - confidence interval = $\widehat{\beta_1} \pm \text{critical}value se(\widehat{\beta_1})$
- If homoskedasticity assumption is violated
  - We get the wrong confidence interval
    - $\widehat{\beta_1} \pm \text{critical}value \times se(\widehat{\beta_1})$ gives us the right confidence interval
- If homoskedasticity assumption is NOT violated
  - We get the right confidence interval
    - $\widehat{\beta_1} \pm \text{critical}value \times se\left(\widehat{\beta_1}\right)$ doesn't give us the right confidence interval

Standard Error of Estimation

- Before we have derived:
- $Var(\widehat{\beta_1}) = \frac{Var(u_i)}{\sum_i (x_i - \bar{x})^2}$
  - Note $Var(u_i)$ as σ^2
  - $Var(\widehat{\beta_1}) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} = \frac{\sigma^2}{S_{xx}}$
- The estimation of σ^2 is noted as $\hat{\sigma}^2 = \frac{\sum_i \widehat{u_i}^2}{n-2} = \frac{SSR}{n-2}$
  - Theoretically one can derive $\hat{\sigma}^2$ formular as we derive $\widehat{\beta_1}$ formular, but it is way more complicated. So just remember this formular
- The estimation of $Var(\widehat{\beta_1})$, i.e, $\widehat{Var}(\widehat{\beta_1}) = \frac{\hat{\sigma}^2}{S_{xx}}$
  - $Standard\ Error = \sqrt{Estimated\ variance}$
  - So $se(\widehat{\beta_1}) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$
  - Note $Standard\ Deviation = \sqrt{True\ Variance} = \sqrt{\frac{\sigma^2}{S_{xx}}}$

Standard Error of Estimation (Multi-Variable)

- $Var(\widehat{\beta_j}) = \frac{\sigma^2}{SST_j(1-R_j^2)}$
  - $\sigma^2 = Var(u) = Var(u|x_1, \dots, x_n)$
  - $SST_j = \sum_i (x_{ij} - \overline{x_j})^2$: Total sample variation in x_j
  - R_j^2: R-square from regressing x_j on all other independent variables (including intercept)

Standard Error of Estimation (Multi-Variable)

- $Var(\widehat{\beta_j}) = \frac{\sigma^2}{SST_j(1-R_j^2)}$
  - $\sigma^2 = Var(u) = Var(u|x_1, \dots, x_n)$
  - $SST_j = \sum_i (x_{ij} - \bar{x_j})^2$: Total sample variation in x_j
  - R_j^2: R-square from regressing x_j on all other independent variables (including intercept)

- $\widehat{Var(\widehat{\beta_j})} = \frac{\hat{\sigma}^2}{SST_j(1-R_j^2)} \Rightarrow \text{se}(\widehat{\beta_j}) = \sqrt{\frac{\hat{\sigma}^2}{SST_j(1-R_j^2)}}$

- We have learned that, if we are excluding regressors that is
  - Affecting y
  - and correlated to x_1
- $E[\widehat{\beta_1}]$ will be influenced
- How $E[\widehat{\beta_1}]$ will be influenced will be related to the $cov(x_1, x_2)$
- If we excludes regressors that is
  - Not correlated to x_2
- $E[\widehat{\beta_1}]$ will NOT be influenced

- How does adding different kind of regressors influence $se(\widehat{\beta_1}) = \sqrt{\frac{\widehat{\sigma}^2}{SST_j(1-R_j^2)}}$?

- How does adding different kind of regressors influence $se(\widehat{\beta_1}) = \sqrt{\frac{\widehat{\sigma}^2}{SST_j(1-R_j^2)}}$?
- If we are adding regressors that is
  - Affecting y
    - The residuals will be smaller (model is fitted in a better way) $\Rightarrow \hat{\sigma}^2 \downarrow$
      - If the added regressor is correlated to x_1
        - $R_j^2 > 0 \Rightarrow (1 - R_j^2) \downarrow$
        - $se(\widehat{\beta_1})$ change direction is ambiguous
      - If the added regressor is NOT correlated to x_1
        - $R_j^2 \approx 0$ but nominator is decreasing so
        - $se(\widehat{\beta_1}) \downarrow$
  - Not Affecting y
    - The residuals will not change much
      - If the added regressor is correlated to x_1
        - $R_j^2 > 0 \Rightarrow (1 - R_j^2) \downarrow$
        - $se(\widehat{\beta_1}) \uparrow$
      - If the added regressor is NOT correlated to x_1
        - $R_j^2 \approx 0$ and nominator is not changing much so
        - $se(\widehat{\beta_1})$ won't change much