Lecture 4: Heteroskedasticity and Dependence

OLS estimator:

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

$$\hat{\beta}_1 = \frac{1}{\sum_{i=1}^{N}(X_i-\bar{X})^2}\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})$$

$$\hat{\beta}_1 = \beta_1 + \frac{1}{\sum_i (X_i-\bar{X})^2}\sum_i (X_i-\bar{X})U_i$$

$E[U_i]=0 \rightarrow \hat{\beta}_1$ unbiased

$\rightarrow 0 \rightarrow \hat{\beta}_1$ consistent

$$V[\hat{\beta}_1] = \cancel{V[\beta_1]} + V\left[\frac{1}{\sum_i (X_i-\bar{X})^2}\sum_i (X_i-\bar{X})U_i\right]$$

$$V[\hat{\beta}_1] = \left(\frac{1}{\sum_i (X_i-\bar{X})^2}\right)^2\left[\sum_i (X_i-\bar{X})^2 V[U_i] + \sum_i\sum_{j\neq i}(X_i-\bar{X})(X_j-\bar{X})\,\mathrm{cov}(U_i,U_j)\right]$$

Classical model

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

- GM assumptions
  1. $E[U_i] = 0$
  2. $V[U_i] = \sigma^2$
  3. $\mathrm{cov}(U_i, U_j) = 0 \quad i \neq j$

$$V[\hat{\beta}_1] = \left(\frac{1}{\sum_i (X_i - \bar{X})^2}\right)^2 \left[\sum_i (X_i - \bar{X})^2 \overset{\sigma^2}{\cancel{V[U_i]}} + \cancel{\sum_i \sum_{j \neq i} (X_i - \bar{X})(X_j - \bar{X}) \,\mathrm{cov}(U_i, U_j)} \right]$$

homoskedasticity $\rightarrow$ simple math!

no assumptions about shape of U_i distribution

$\rightarrow$ $SE[\hat{\beta}_1]$ and then apply CLT

$\rightarrow$ large sample inference!

Normal linear model

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

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

$\rightarrow$ then $t = \dfrac{\hat{\beta}_1 - \beta_1^0}{SE[\hat{\beta}_1]}$ has $t(N-2)$ dist.

Random X's

$E[U_i] = 0 \Rightarrow E[U_i \mid X_1, X_2 \cdots X_N] = 0$

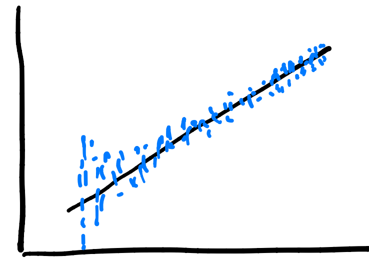
$E[\hat{\beta}_1] = \beta_1 \Rightarrow E[\hat{\beta}_1 \mid X_1, \cdots, X_N] = \beta_1$

Heteroskedasticity

① $E[U_i \mid X_i] = 0$

② (X_i, Y_i) are iid

③ outliers are unlikely



$$V[\hat{\beta}_1] = \left(\frac{1}{\sum_i (X_i - \bar{X})^2}\right)^2 \left[\sum_i (X_i - \bar{X})^2 V[U_i] + \cancel{\sum_i \sum_{j \neq i} (X_i - \bar{X})(X_j - \bar{X}) \operatorname{cov}(U_i, U_j)}\right]$$

→ still set cov( ) = 0, but more complicated formula

Dependence

→ observations are dependent within <u>clusters</u>, but ind. <u>across</u> clusters.

① clustered sample design

② group-level treatment

$$V[\hat{\beta}_1] = \left(\frac{1}{\sum_i (X_i-\bar{X})^2}\right)^2 \left[\sum_i (X_i-\bar{X})^2 \underbrace{V[U_i]}_{\text{still allow hetero}}^{U_i^2} + \sum_i \sum_{\cancel{j \neq i}} (X_i-\bar{X})(X_j-\bar{X}) \overbrace{\mathrm{cov}(U_i, U_j)}^{U_iU_j}\right]$$

$j \in i$'s cluster

vcov = 'hetero'

vcov = ~ clustervar

→ Another option: group data $\bar{Y}_k = \beta_0 + \beta_1 X_k + \bar{U}_k$

Back WLS

- Recall: $\hat{\beta}_1^{WLS} = \dfrac{\sum_i w_i (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_i w_i (X_i - \bar{X})^2}$

- If U_i is heteroskedastic, $w_i = \dfrac{1}{V[U_i]}$

- Grouped data example:

individual: $Y_{ig} = \beta_0 + \beta_1 X_{ig} + U_{ig}$ $\quad V[U_{ig} | X_{ig}] = \sigma^2$

group: $\bar{Y}_g = \beta_0 + \beta_1 \bar{X}_g + \bar{U}_g$ $\quad V[\bar{U}_g | \bar{X}_g] = \dfrac{\sigma^2}{N_g}$

- $w_g = \dfrac{1}{V[\bar{U}_g]} = \dfrac{N_g}{\sigma^2}$ ← constant → can set $w_g = N_g$