Multiple Regression Analysis

1. Motivation
2. Estimation
3. Expected Value
4. Variance

Motivation

- Examples
- Ceteris Paribus: Public vs. Private University
  - `https://www.youtube.com/watch?v=iPBV3BlV7jk&list=PL-uRhZ_p-BM5ovNRg-G6hDib27OCvcyW8&index=2`
- Interpretation

Estimation

Ass?s: $\boxed{E(u) = 0}$

$$E(u|x_1, x_2, \ldots, x_k) = E(u) = 0$$

$\boxed{E(x_1 u) = 0}$, $\boxed{E(x_2 u) = 0}$ ... $\boxed{E(x_k u) = 0}$

- Model with k independent variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

- Objective: estimate $\beta_0, \beta_1, \ldots \beta_k$

$$u = y - \beta_0 - \beta_1 x_1 - \ldots - \beta_k x_k$$

$$E(y - \beta_0 - \beta_1 x_1 - \ldots - \beta_k x_k) = 0$$

$$E[x_1(y - \beta_0 - \beta_1 x_1 - \ldots - \beta_k x_k)] = 0$$

$$\vdots$$

$$E[x_k(y - \beta_0 - \beta_1 x_1 - \ldots - \beta_k x_k)] = 0$$

Estimation (cont.)

e.g. $x_1 \rightarrow$ educ

$exper_i \rightarrow x_{i2}$ $\qquad x_2 \rightarrow$ exper

$educ_i \rightarrow x_{i1}$

- Sample analogs

$$n^{-1}\sum_{i=1}^{n}\left(y_i-\widehat{\beta}_0-\widehat{\beta}_1 x_{i1}-\ldots-\widehat{\beta}_k x_{ik}\right) = 0$$

$$n^{-1}\sum_{i=1}^{n}x_{i1}\left(y_i-\widehat{\beta}_0-\widehat{\beta}_1 x_{i1}-\ldots-\widehat{\beta}_k x_{ik}\right) = 0$$

$$\vdots$$

$$n^{-1}\sum_{i=1}^{n}x_{ik}\left(y_i-\widehat{\beta}_0-\widehat{\beta}_1 x_{i1}-\ldots-\widehat{\beta}_k x_{ik}\right) = 0$$

x_{ij} : observation i for variable x_j

Estimation (cont.)

$$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$$

- OLS estimates:
- Fitted value:
- Residual:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik}$$

$$\hat{u}_i = y_i - \hat{y}_i$$

Estimation (cont.)

Example

y (wage)	x_1 (educ)	x_2 (exper)
3.1	11	2
3.2	12	22
3	11	2
6	8	44
5.3	12	7
8.8	16	9
11	18	15
5	12	5
3.6	12	26
18	17	22

Estimation (cont.)

$$\overline{y} - \widehat{\beta}_0 - \widehat{\beta}_1\overline{x_1} - \widehat{\beta}_2\overline{x_2} = 0$$
$$\overline{x_1 y} - \widehat{\beta}_0\overline{x_1} - \widehat{\beta}_1\overline{(x_1)^2} - \widehat{\beta}_2\overline{x_1 x_2} = 0$$
$$\overline{x_2 y} - \widehat{\beta}_0\overline{x_2} - \widehat{\beta}_1\overline{x_1 x_2} - \widehat{\beta}_2\overline{(x_2)^2} = 0$$

$$6.742 - \widehat{\beta}_0 - 12.9\widehat{\beta}_1 - 15.4\widehat{\beta}_2 = 0$$
$$97.234 - 12.9\widehat{\beta}_0 - 175.1\widehat{\beta}_1 - 190.4\widehat{\beta}_2 = 0$$
$$115.064 - 15.4\widehat{\beta}_0 - 190.4\widehat{\beta}_1 - 396.8\widehat{\beta}_2 = 0$$

$\widehat{\beta}_0 = -12.317 \qquad \widehat{\beta}_1 = 1.312 \qquad \widehat{\beta}_2 = 0.138$

Estimation (cont.)

Sum / avg. value of residuals $= 0$

Correl. b/w residuals & each explanatory variable $= 0$

Properties of OLS

correl. b/w residuals & fitted value $= 0$

If each $x_j = \bar{x}_j$, $\hat{y} = \bar{y}$.

Estimation (cont.)

Goodness-of-fit

- R-squared

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

  Non-decreasing in the number of independent variables, k

- Adjusted R-squared

$$\bar{R}^2 = 1 - \frac{\frac{SSR}{(n-k-1)}}{\frac{SST}{(n-1)}}$$

  As $k \uparrow$, $SSR \downarrow$ but $(n-k-1)$ also $\downarrow$. $\bar{R}^2$ may $\downarrow$ or $\uparrow$

Expected Value

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + u$

price sq.ft. age # BRs quality

- Under certain assumptions, the OLS estimators are unbiased so that

$wage = \beta_0 + \beta_1\, age + \beta_2\, educ + \beta_3\, exper + \dots + \beta_k x_k + u$

$$E(\hat{\beta}_j) = \beta_j \qquad j = 0, 1, \dots, k$$

- Assumptions
  - (MLR.1) Linear in Parameters
  - (MLR.2) Random Sampling
  - (MLR.3) No Perfect Collinearity
  - (MLR.4) Zero Conditional Mean

variation in each regressor

No linear relationship among explanatory vars.

$n \geq k+1$

$age = 6 + educ + exper$

$E(u \mid x_1, x_2, \dots, x_k) = 0$

Expected Value (cont.)

Omitted Variable Bias

- Will You Make More Going to a Private University?
  - https://www.youtube.com/watch?v=6YrIDhaUQOE
- Population model satisfying the assumptions (MLR.1) to (MLR.4)

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \\
\text{wage} &= \beta_0 + \beta_1 \text{educ} + \beta_2 \text{IQ} + u \\
\text{bweight} &= \beta_0 + \beta_1 \text{smoking} + \beta_2 \text{alcohol} + u
\end{aligned}
$$

- Omitting a relevant variable

OLS estimator of β_1 : likely biased

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + u \\
\text{wage} &= \beta_0 + \beta_1 \text{educ} + u \\
\text{bweight} &= \beta_0 + \beta_1 \text{smoking} + u
\end{aligned}
$$

Bias depends on β_2 and correl. b/w x_1 and x_2

Expected Value (cont.)

- Fitted values from the regression where x_2 is omitted $\longrightarrow \tilde{\beta}_0, \tilde{\beta}_1$.

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$
$$\widetilde{\text{wage}} = \tilde{\beta}_0 + \tilde{\beta}_1 \, \text{educ}$$
$$\widetilde{\text{bweight}} = \tilde{\beta}_0 + \tilde{\beta}_1 \, \text{smoking}$$

- Fitted values from the regression if x_2 is included $\longrightarrow \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$
$$\widehat{\text{wage}} = \hat{\beta}_0 + \hat{\beta}_1 \, \text{educ} + \hat{\beta}_2 \, \text{IQ}$$
$$\widehat{\text{bweight}} = \hat{\beta}_0 + \hat{\beta}_1 \, \text{smok} + \hat{\beta}_2 \, \text{alcoh.}$$

- Fitted values from the regression of x_2 on x_1

$$\widetilde{x_2} = \tilde{\delta}_0 + \tilde{\delta}_1 x_1$$
$$\widetilde{\text{IQ}} = \tilde{\delta}_0 + \tilde{\delta}_1 \, \text{educ}$$
$$\widetilde{\text{alcohol}} = \tilde{\delta}_0 + \tilde{\delta}_1 \, \text{smoking}$$

Expected Value (cont.)

- Relationship between $\tilde{\beta}_1$ and $\hat{\beta}_1$

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1$$

- If β_1 is estimated with x_2 omitted

$$E(\tilde{\beta}_1) = \beta_1 + \beta_2 \tilde{\delta}_1$$
$$\text{Bias} = \beta_2 \tilde{\delta}_1$$

- Bias depends on

β_2 & correl. b/w x_1 and x_2

Bias = 0 if $\beta_2 = 0$ or $\tilde{\delta}_1 = 0$

Expected Value (cont.)

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u \rightarrow$ political activism

x_1: env. reg.; β_0: FDI; x_2: input prices; x_3: infrastructure

- Additional explanatory variables: more complicated derivation of bias
- Other sources of bias
- Inclusion of irrelevant regressors
  - Exercise caution
  - May affect the variance of OLS estimators

if u corr. with x_1 but not x_2 and x_3

OLS estimators of all β's biased if x_1 corr. w/ x_2 and x_3.

$wage = \beta_0 + \beta_1 educ + \beta_2 IQ + \beta_3 exper + \beta_4 discrim. + \beta_5 occupation + u$

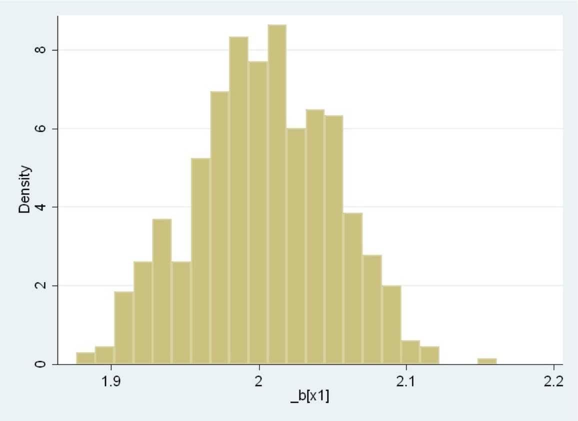
measurement error in x or y (e.g. crime)

simultaneity: x police PC $\leftrightarrow$ y crime PC

sample selection (e.g. trade): data observed if $y >$ threshold

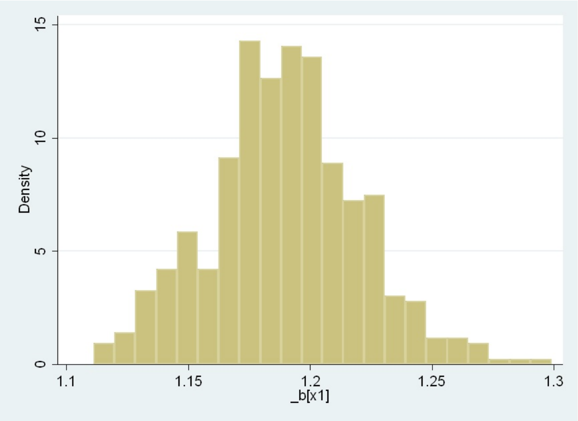
Expected Value (cont.)

- $n = 500$, reps $= 500$, corr$(x_1, x_2 = 0.4)$, corr$(x_1, u = 0)$, corr$(x_2, u = 0)$
- $y = 1 + 2x_1 + x_2 + u$



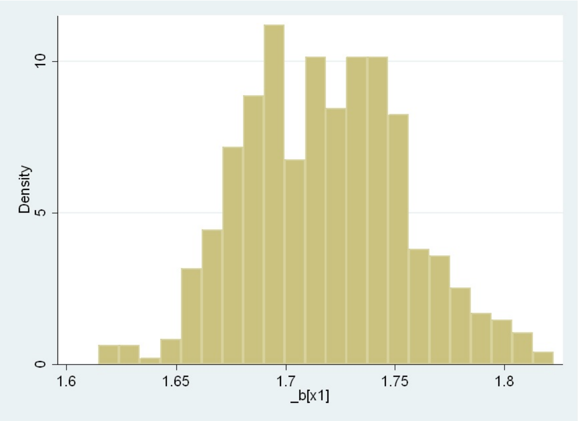
Expected Value (cont.)

- $n = 500$, reps $= 500$, corr$(x_1, x_2 = 0.4)$, corr$(x_1, u = -0.6)$, corr$(x_2, u = 0.2)$
- $y = 1 + 2x_1 + x_2 + u$



Expected Value (cont.)

- $n = 500$, reps $= 500$, $\text{corr}(x_1, x_2 = 0.4)$, $\text{corr}(x_1, u = 0)$, $\text{corr}(x_2, u = 0.6)$
- $y = 1 + 2x_1 + x_2 + u$



Variance

- Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

- (MLR.5) Homoskedasticity

$$Var(u|x_1, x_2, ..., x_k) = \sigma^2$$

- Alternatively

$$Var(y|x_1, x_2, ..., x_k) = \sigma^2$$

- If $Var(u|x_1, x_2, ..., x_k)$ depends on x_j $\longrightarrow$ heteroskedasticity

Variance (cont.)

- Under Assumptions MLR.1 to MLR.5 (Gauss Markov assumptions)

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1-R_j^2)} \text{ except } j = 0$$

  - $SST_j = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$
  - R_j^2: → R^2 from regression of x_j on other x's

- Example

$$\text{wage} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{IQ} + \beta_3 \text{exper} + u$$

educ : avg. educ.

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_{educ}(1-R^2_{educ})}$$

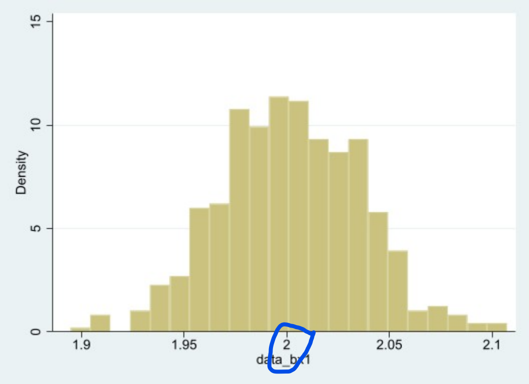
  - $SST_{educ} = \sum_{i=1}^{n} (educ_i - \overline{educ})^2$
  - R^2_{educ}: → R^2 from reg. of educ. on IQ and exper.

Variance (cont.)

- $Var(\hat{\beta}_j) \uparrow$ with σ^2 and thus *may* $\downarrow$ with additional regressors
- $Var(\hat{\beta}_j) \downarrow$ with SST_j and thus likely to $\downarrow$ with n
- $Var(\hat{\beta}_j) \uparrow$ with R_j^2
  - R_j^2 "close" to one - the "problem" of multicollinearity; does not violate ass?s reqd. for unbiasedness
  - $R_j^2 = 1$ ruled out by MLR.3

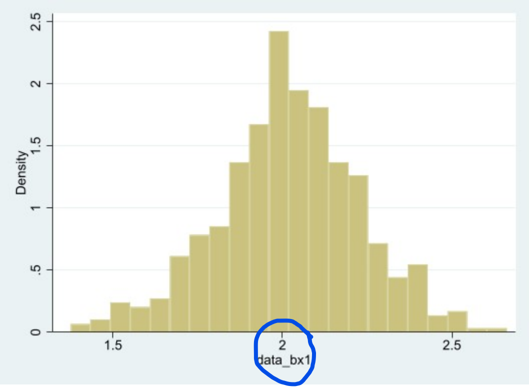
Variance (cont.)

- $n = 1000$, reps $= 500$, $\text{corr}(x_1, x_2 = 0.4)$, $\text{corr}(x_1, u = 0)$, $\text{corr}(x_2, u = 0)$
- $y = 1 + 2x_1 + x_2 + u$



Variance (cont.)

- $n = 1000$, reps $= 500$, $\text{corr}(x_1, x_2 = 0.99)$, $\text{corr}(x_1, u = 0)$, $\text{corr}(x_2, u = 0)$
- $y = 1 + 2x_1 + x_2 + u$



Variance (cont.)

- Additional thoughts on the inclusion of irrelevant regressors
  - May $\uparrow Var(\hat{\beta}_j)$ if R_j^2 high
  - Likely to $\downarrow Var(\hat{\beta}_j)$ " " low

Variance (cont.)

k : # of regressors

$n-(k+1)$: n − # of β's

e.g. SLR : $k=1$

- Under MLR.1 to MLR.5, $\hat{\sigma}^2$ is an unbiased estimator of σ^2

$$\hat{\sigma}^2 = (n-k-1)^{-1} \sum_{i=1}^{n} \hat{u}_i^2$$

- $\hat{\sigma}$: standard error of the regression
- Standard error of each $\hat{\beta}_j$

$$se(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{SST_j(1-R_j^2)}} \text{ except } j = 0$$