Financial Econometrics

Assumptions and biases

Maximilian Rohrer

Overview

- Assumptions of single OLS regression
- Properties of OLS estimator
- Additional assumptions of multiple regression
- Standard error correction
- Biases

Disclaimer

- A lot of math in this slide set
- Intuition most important

Lest squares regression

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

- Linear model
- Estimation by minimizing squared error
- β_1 as marginal effect
- Estimation on sample of population

Least squares assumptions

Least squares assumptions

- LS.1 The conditional distribution of u given x has mean zero, that is $E(u_i|x_i) = 0$
- LS.2 (x_i, y_i) for $i = 1, \ldots, N$ are identically and independently distributed (i.i.d)
- LS.3 Large outliers in x and/or y are rare

LS.1 Zero conditional mean - derivation

- Equation 4.25

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

- Note,

$$y_i = \beta_0 + \beta_1 x_i + u_i \;\Rightarrow\; y_i - \bar{y} = \beta_1(x_i - \bar{x}) + (u_i - \bar{u})$$

- Insert back

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\left[\beta_1(x_i - \bar{x}) + (u_i - \bar{u})\right]}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

LS.1 Zero conditional mean - derivation

- Regroup

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\beta_1(x_i - \bar{x}) + \sum_{i=1}^{n}(x_i - \bar{x})(u_i - \bar{u})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})(u_i - \bar{u})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})(u_i - \bar{u})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

LS.1 Zero conditional mean - derivation

- Regroup the numerator of last part

$$\sum_{i=1}^{n}(x_i - \bar{x})(u_i - \bar{u}) = \sum_{i=1}^{n}(x_i - \bar{x})u_i - \sum_{i=1}^{n}(x_i - \bar{x})\bar{u}$$

- Note,

$$\sum_{i=1}^{n}(x_i - \bar{x})\bar{u} = (\sum_{i=1}^{n} x_i - n\bar{x})\bar{u} = \left(\sum_{i=1}^{n} x_i - n\frac{1}{n}\sum_{i=1}^{n} x_i\right)\bar{u} = 0 * \bar{u}$$

- Put back

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

LS.1 Zero conditional mean - derivation

- Equation 4.29: Under what condition is the expected value of the estimate given the data equal to the population parameter?

$$E(\hat{\beta}_1|X_1,\ldots,X_n) = E\left[\beta_1 + \frac{\sum_{i=1}^{n}(x_i-\bar{x})u_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}|X_1,\ldots,X_n\right]$$

$$E(\hat{\beta}_1|X_1,\ldots,X_n) = \beta_1 + \frac{\sum_{i=1}^{n}(x_i-\bar{x})E(u_i|X_i,\ldots,X_n)}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

If var(x) $\neq 0$,

$$E(\hat{\beta}_1|X_1,\ldots,X_n) = \beta_1 \quad \text{only if} \quad E(u_i|x_i) = 0$$

LS.1 Zero conditional mean - Illustrative example

- Think about the effect of class-rooms on educational outcomes
- What are other factors that might correlated with educational outcomes? Educational outcome parents, wealth, number of imigrants
- All these reasons, are captured by the error term u_i
- If there is some correlation between these factors and the main dependent variable, our estimate $\hat{\beta}_1$ will be biased
- Example: wealthy parents put their kids in schools with small class room size. Hence the effect attributed to class size might actually be parental wealth.

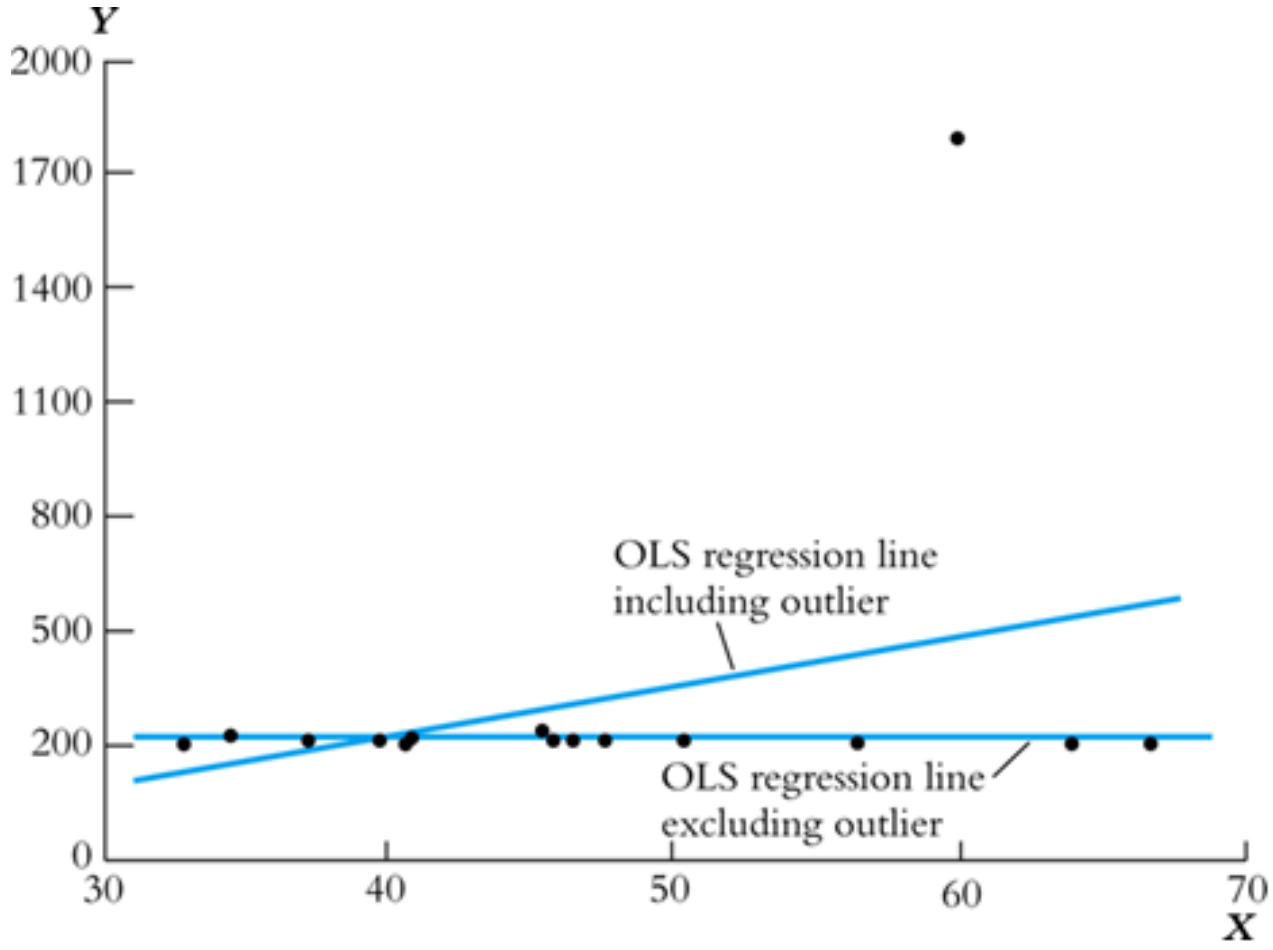
LS.2 (x_i, y_i) for $i = 1, \ldots, N$ are identically and independently distributed (i.i.d)

- Random sampling
  - Population: complete set of observations
  - Sample: subset of population
- If sample random, then representative for population

LS.2 Random Sampling - Illustrative example

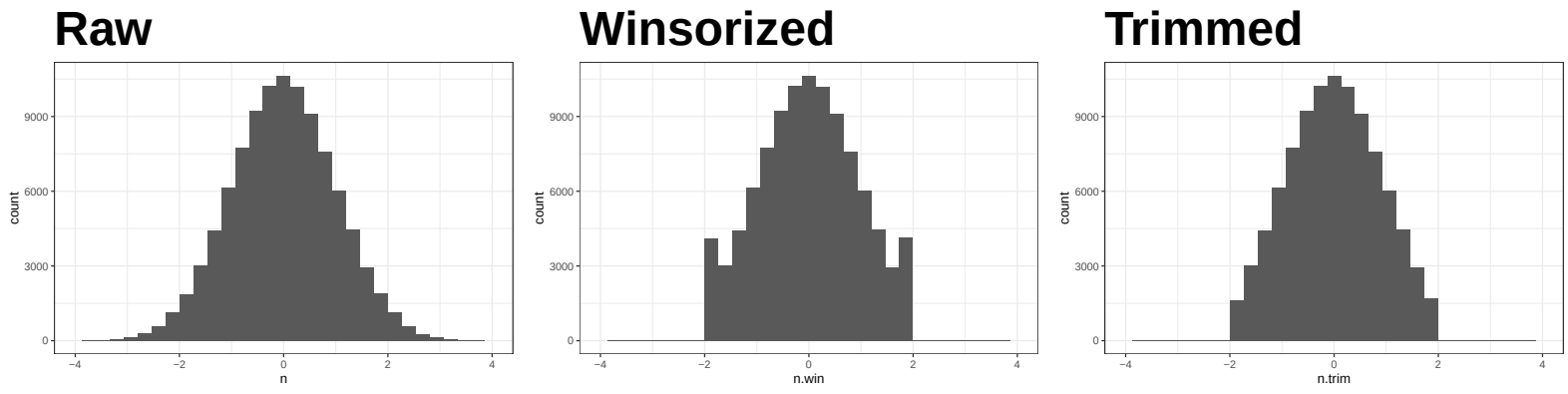
- You want to know the average income of Norwegians
- You travel to Kongsberg, go to the park, and ask all people sitting there at 11am in the morning what they get per month
- You take the average, which is the sample estimate of the average income of Norwegians
- Any problems?

LS.3 Large outliers are rare



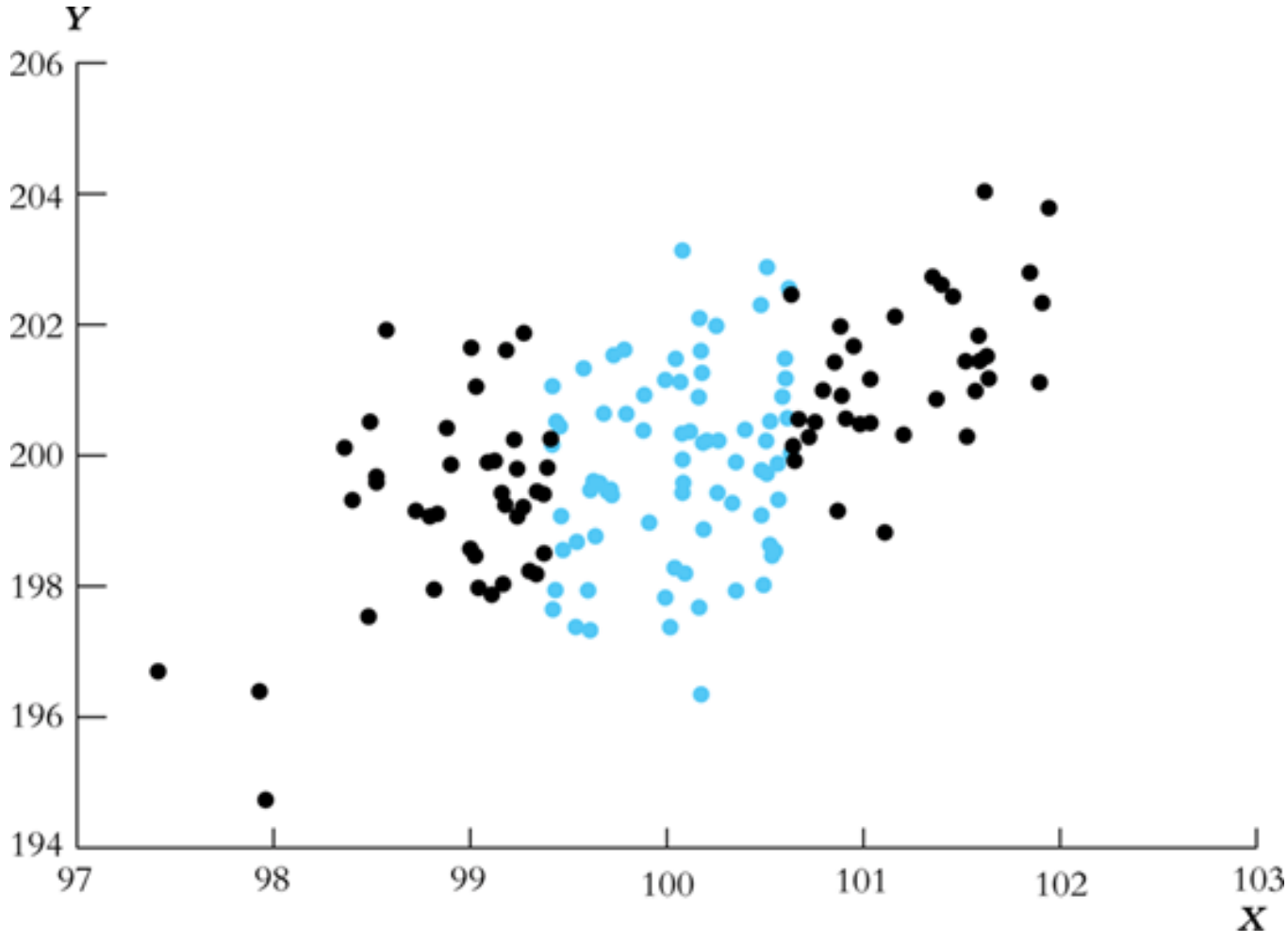
LS.3 What to do if there are outliers?

- Winsorizing: tranforming extreme values
- Triming: deleting extreme values
- Example Winsorize/Trim at 5% interval (affects 2.5% of the empirical distribution to the right and left)



- Need to understand the nature of the outlier
  - CEO salary on performance. What about the $\leq$ 1USD-salary CEOs (Steve Jobs, Elon Musk, etc.)

Sample variation in the explanatory variable



Properties of OLS

Properties of OLS

- Given the assumptions LS.1 to LS.3 and for large N, OLS estimator has . . . properties

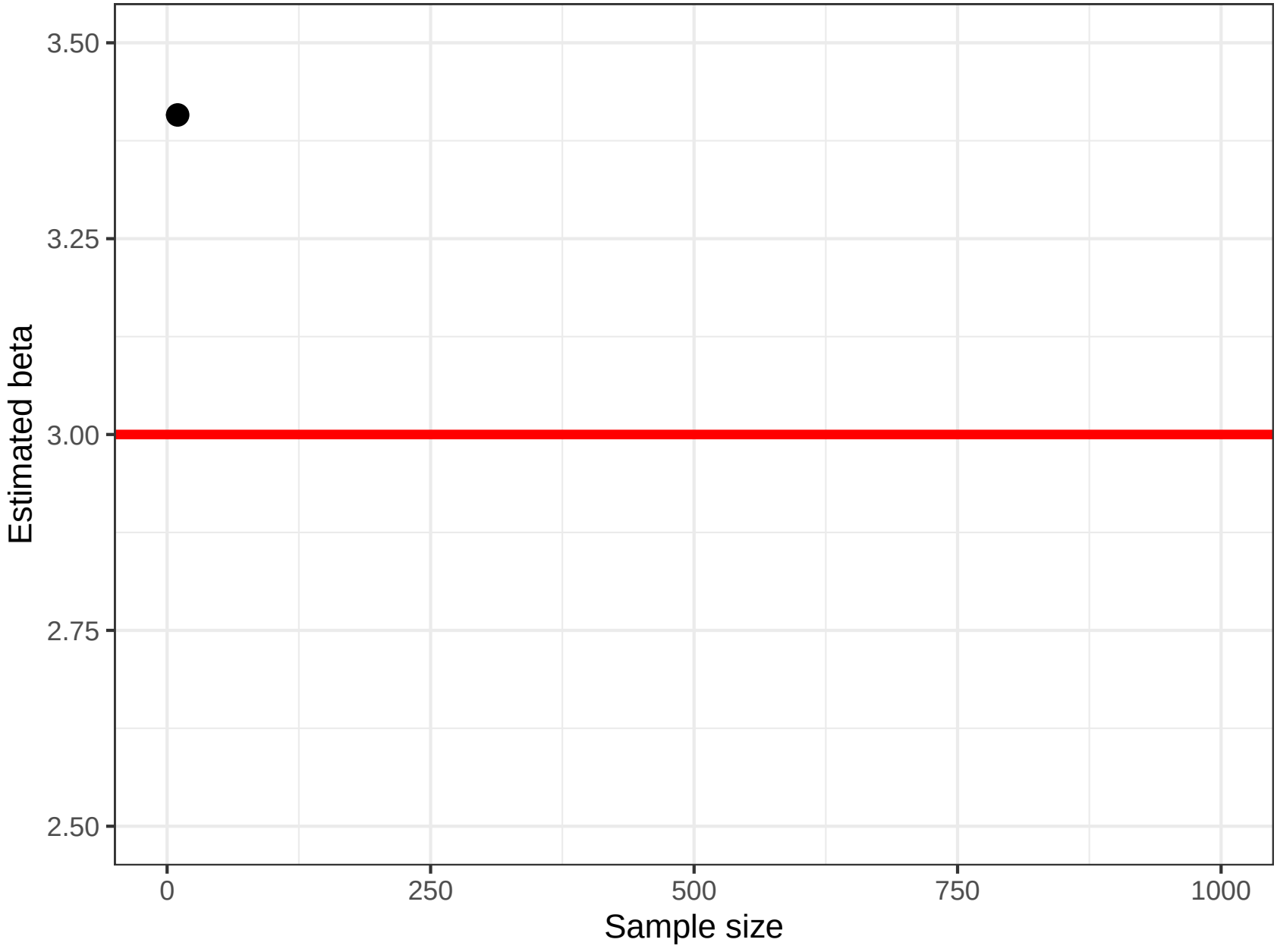
Thought experiment

- We know that in the population, the relationship between x and y can be described as:

$$y = 0 + 3 * x + u$$

- In the population $\beta = 3$
- The econometrician does not know, he collects 10 observations and estimates an OLS model with x and y

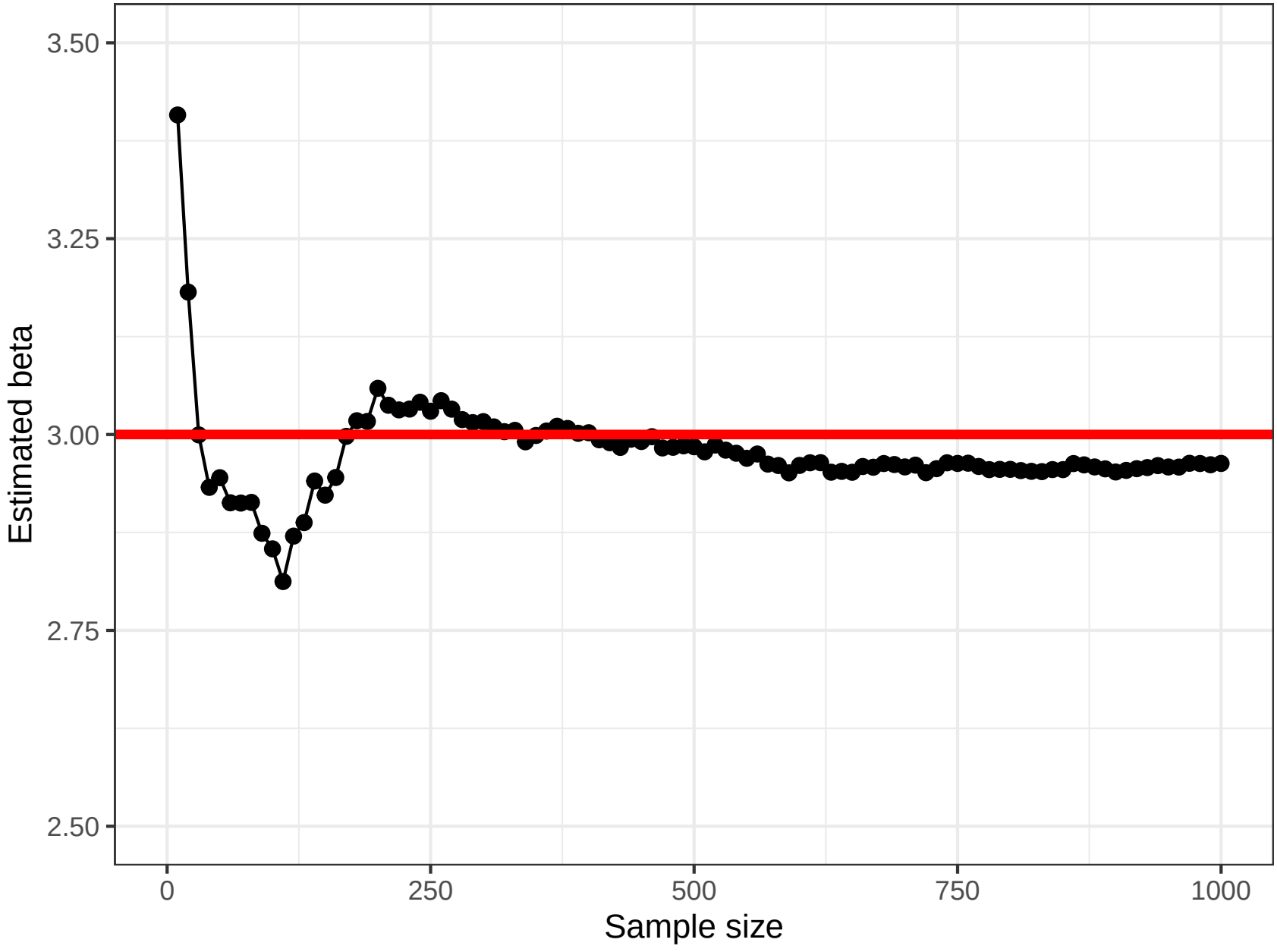
Thought experiment



Thought experiment

- The econometrician is not satisfied and gathers 10 additional observations and estimates the model again
- Thereafter, the econometrician is not satisfied and gathers 10 additional observations
- Thereafter, the econometrician is not satisfied and gathers 10 additional observations
- Thereafter, the econometrician is not satisfied and...
- Thereafter, the econometrician is...
- Thereafter,...

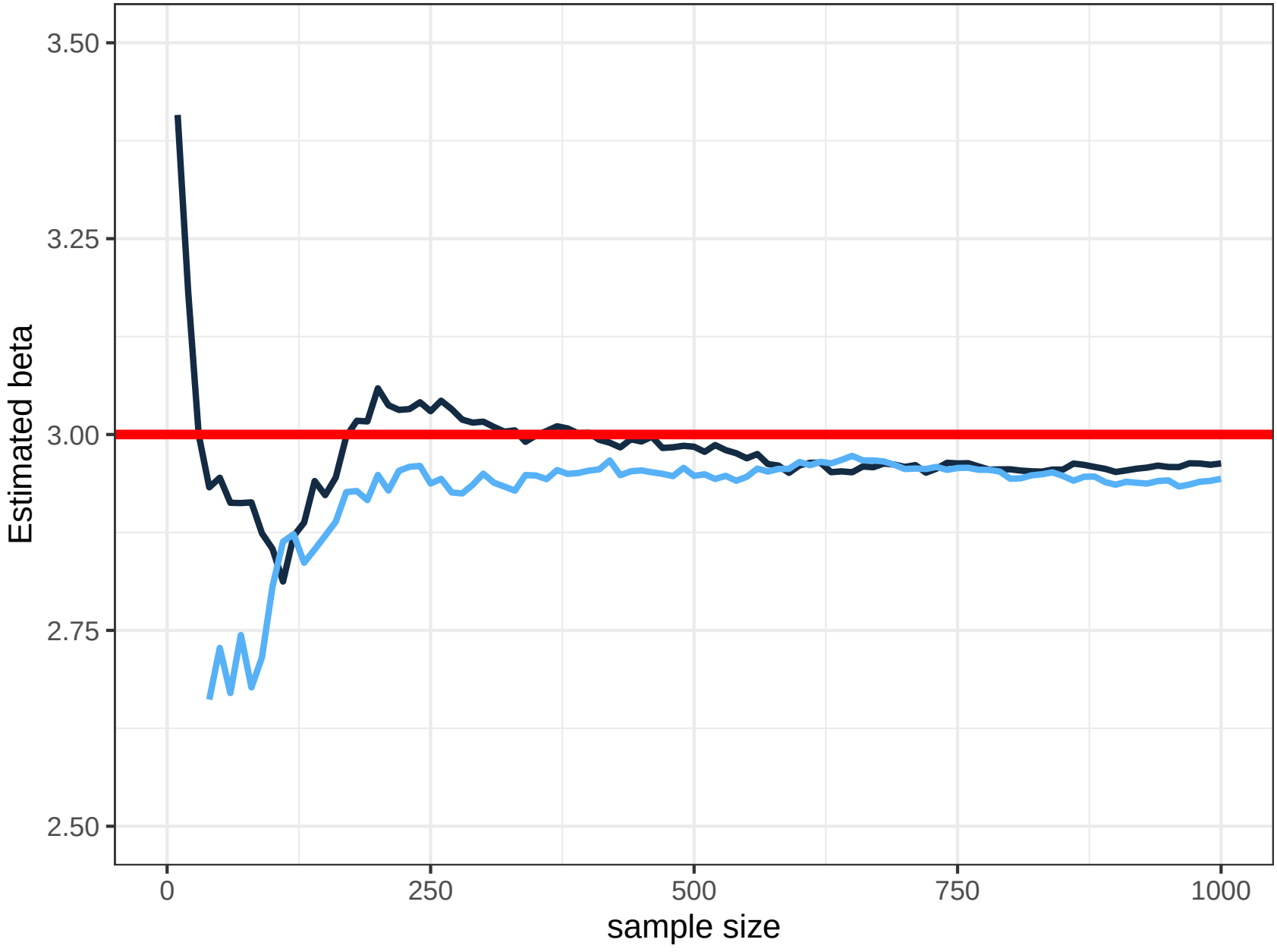
Thought experiment



Thought experiment

- There is a second econometrician
- The second econometrician does not know the population, he collects 10 observations and estimates an OLS model with x and y
- The second econometrician is not satisfied and gathers 10 additional observations and estimates the model again
- Thereafter, the second econometrician is not satisfied and...
- Thereafter, the second econometrician is...
- Thereafter,...

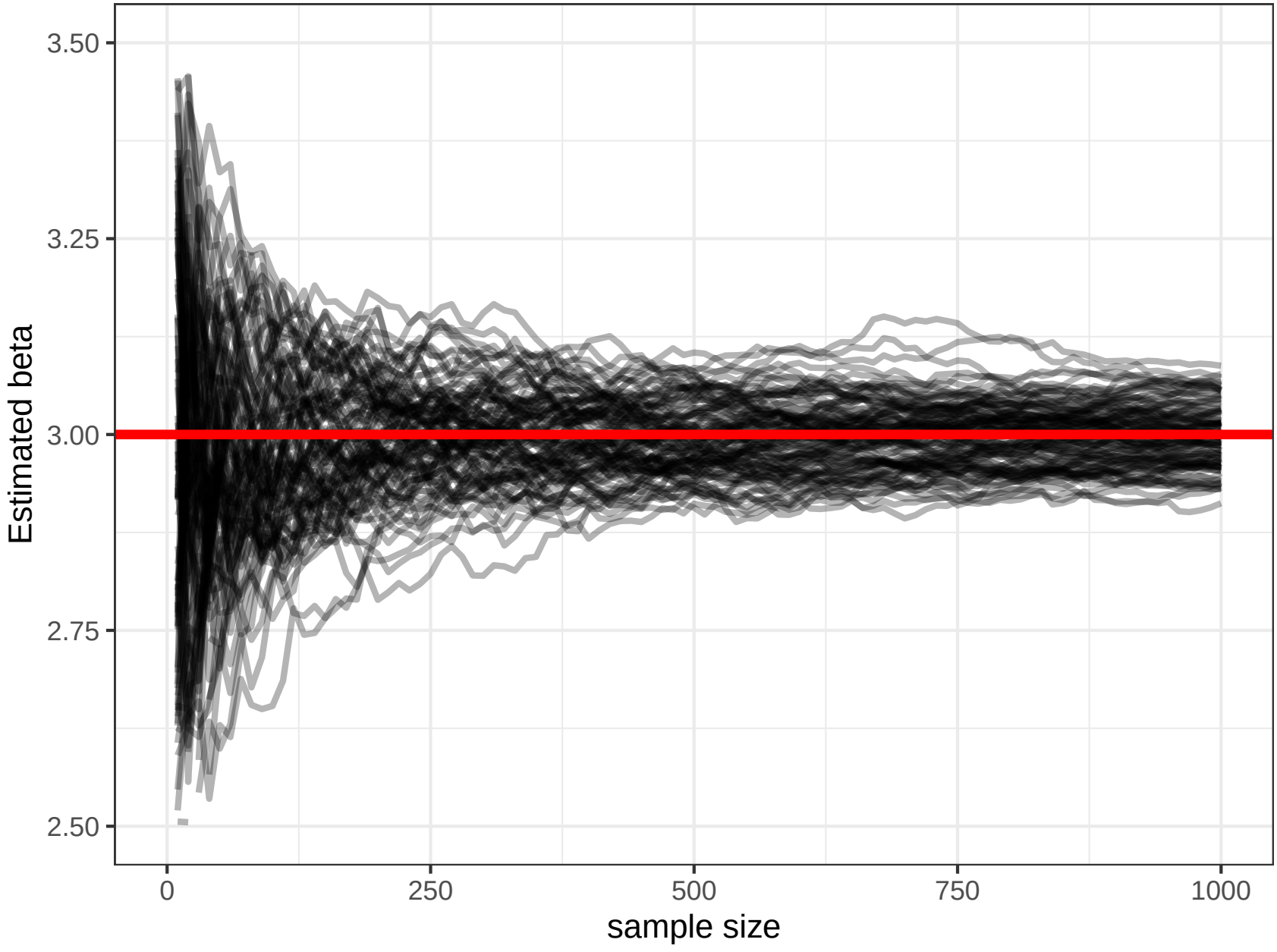
Thought experiment



Thought experiment

- There are a 100 econometricians
- The 100 econometricians do not know the population, they collect 10 observations each and estimate an OLS model with x and y each
- The 100 econometricians are not satisfied and gathers 10 additional observations and estimates the model again
- Thereafter, the 100 econometricians are not satisfied and...
- Thereafter, the 100 econometricians are...
- Thereafter,...

Thought experiment



Properties of $\hat{\beta}_0$ and $\hat{\beta}_1$

- Unbiased
- Consistent
- Normally distributed

Unbiasedness

- $E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$
- The estimated coefficients may be smaller or larger, depending on the sample size
- However, on average they will be equal to the values that characterize the true relationship between y and x
- On average means if sampling was repeated
- In a given sample, estimates may differ considerably from true values

Consistency

- $plim(\hat{\beta}_0) = \beta_0$ and $plim(\hat{\beta}_1) = \beta_1$
- Consistency means that the probability that the estimate is arbitrarily close to the true population value can be made arbitrarily high by increasing the sample size
- Consistency is a minimum requirement for sensible estimators

Normality

- $\hat{\beta}_1 \sim^a N(\beta_1, Var(\hat{\beta}_1))$
- If n is large enough, we can use critical values from a normal distribution for inference purpose

Additional assumptions in multiple regressions

Least squares assumption - Multiple regression

- LS.1 - LS.3
- LS.4 - No perfect multicollinearity

LS.4 - No perfect multicollinearity

Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors:

- including a variable twice in the regression
- dummy variable trap

Dummy variable trap

- Regressing test scores on a constant, small and large class size dummy
  - $small_i = 1$ if class size less than 20, and 0 otherwise
  - $large_i = 1$ if class size more or equal to 20, and 0 otherwise
- Solution
  - omit one group
  - omit the intercept
- What are the implications of the two solutions for interpretation of the coefficients?

LS.4 - No perfect multicollinearity

- Difference perfect and imperfect multicollinearity
- Perfect multicollinearity occurs if two grous or more regressors are perfectly correlated
- Imperfect multicollinearity occurs if two grous or more regressors are highly correlated
- Implies that one or more regression coefficients are estimated imprecisely
- Results in large standard errors for affected coefficients
- Volatility inflation factor can help to detect problem

Volatility inflation factor

$$y = \beta_0 + \beta_1 x + \beta_2 x_2 + ... + \beta_K x_K + u$$

VIF estimation:

- $x_1 = \gamma_0 + \gamma_2 x_2 + \gamma_3 x_3 + ... + \gamma_K x_K + e => VIF_{x_1} = 1/(1 - R^2_{x_1})$
- $x_2 = \gamma_0 + \gamma_1 x_1 + \gamma_3 x_3 + ... + \gamma_K x_K + e => VIF_{x_2} = 1/(1 - R^2_{x_2})$
- ...
- $x_k = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + ... + \gamma_{K-1} x_{K-1} + e =>$ $VIF_{x_K} = 1/(1 - R^2_{x_K})$
- Rule of thumb: $VIF_{x_j} > 10 =>$ Multicollinearity!
- $\sqrt{VIF_{x_j}}$ indicates how much larger the standard error is, compared with what it would be if that variable j were uncorrelated with the other predictor variables in the model.
- Solution for the problem: drop the variable j from the specification.

Imperfect multicollinearity simulation example

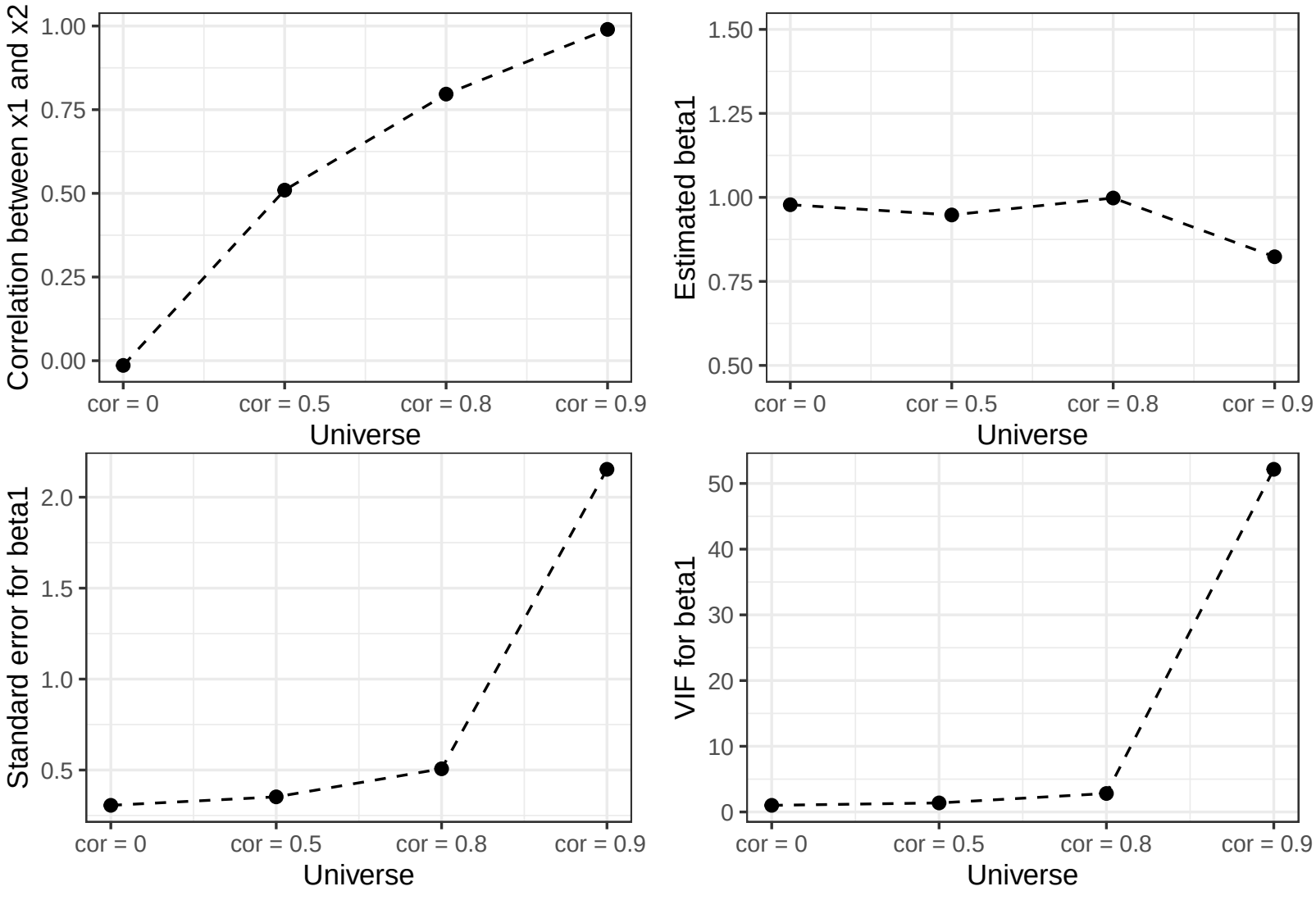
- Let's assume the population model is as follows

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u$$

- The population parameter β_1 equals 1

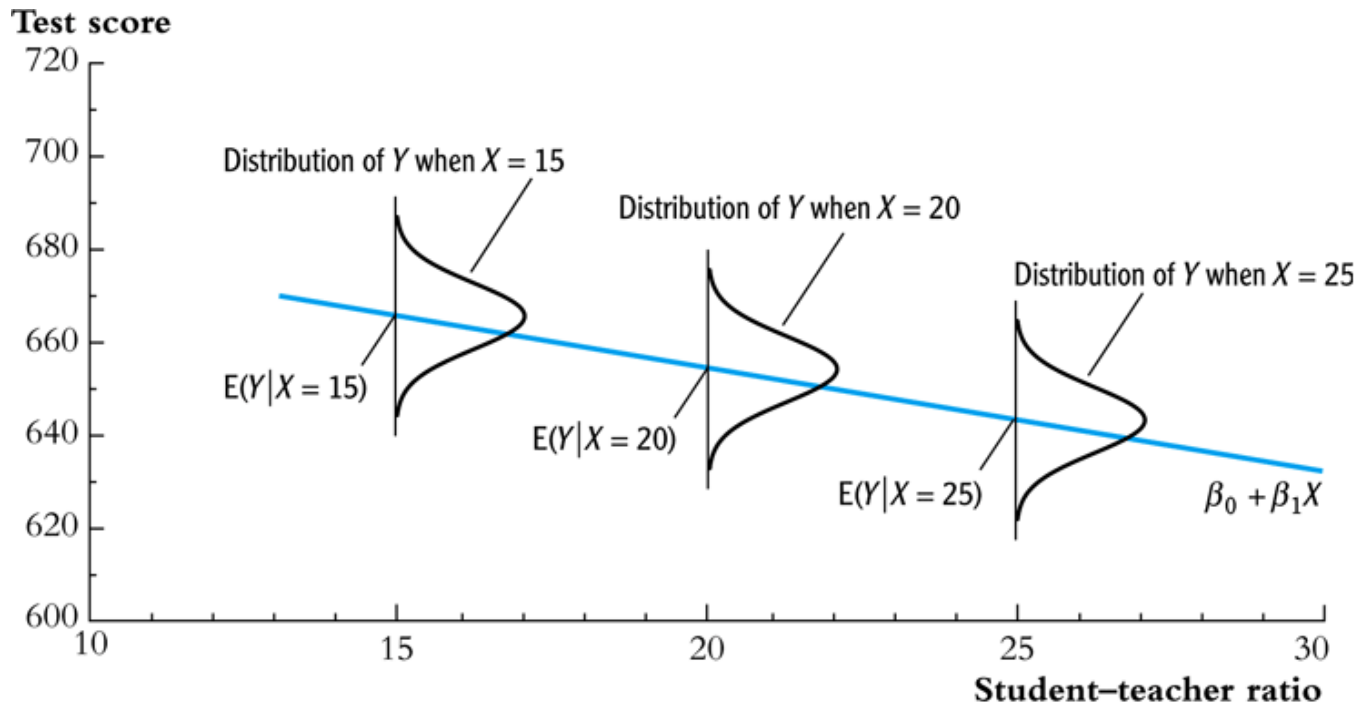
- There are four parallel universe, in which the correlation between x_1 and x_2 equals 0, 0.5, 0.8, & 0.99
- In each universe, there are 100 econometricians that gather 100 observations each and estimate a regression in order to learn β_1
- For each universe, we report the average $\bar{\hat{\beta}}_1$, the average standard error $\bar{SE}(\beta)$, and the average volatility inflation factor $\bar{VIF}$ for β_1

Imperfect multicollinearity simulation example

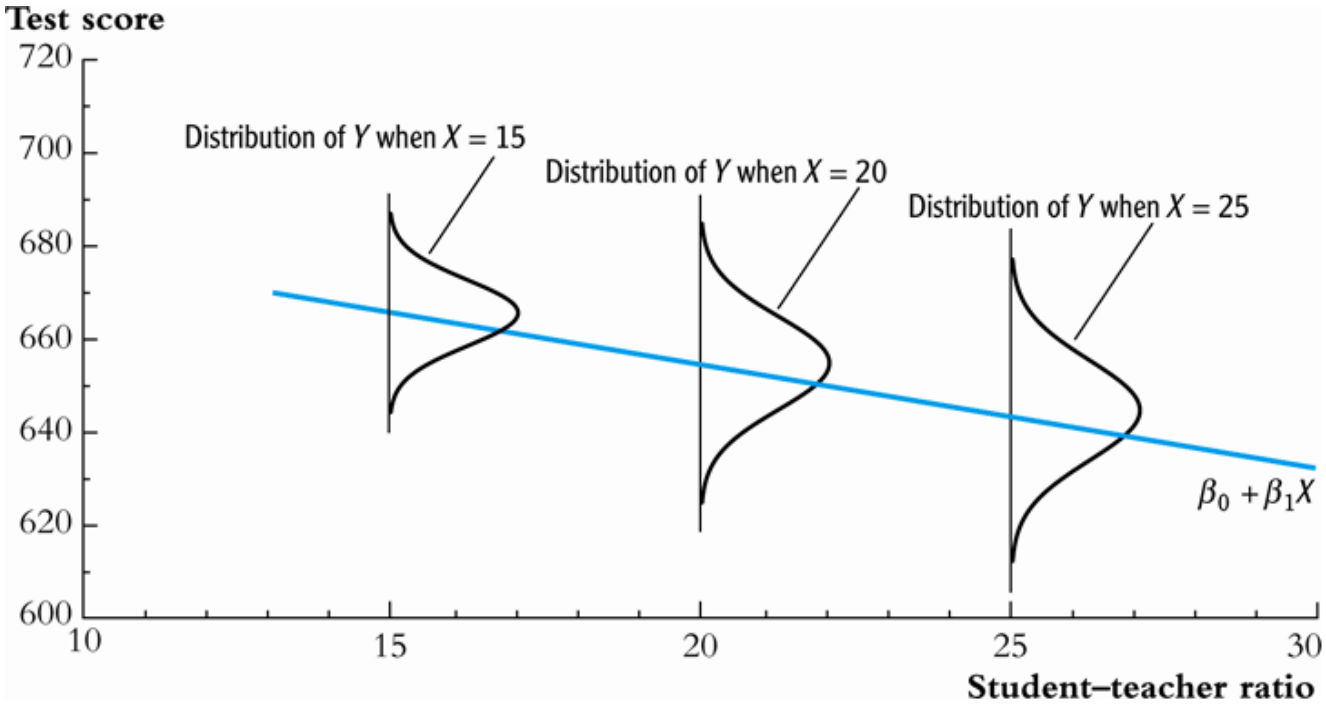


Standard error correction

Homeskedasticity



Heteroskedasticity



Problem and solution

- In the presence of heteroskedasticity
  - Coefficients are unbiased and consistent
  - Standard errors are biased
  - OLS t statistic does not follow a t distribution
  - (Fail to) reject H0 too often or not often enough
- Solution
  - Use heteroskedasticity robust standard errors
  - Prudent to assume errors are heteroskedasticity unless there is a compelling reason
  - Implementation see Lab example

Biases

Main take away

- Biases lead to violation of LS.1
- Hence, coefficient is biased
  - That means the estimate will not converge to the population estimate
- In contrast, heteroskedasticity and multicollinearity lead to biased standard errors, not biased coefficients

The usual suspects - biases

- Sample selection bias
- Omitted variable bias
- Simultaneity bias
- Measurement error in independent variable
- All biases imply a violation of LS.1 and biased coefficients
- Small trick to find what might bias your analysis: imagine the perfect experiment. How does the reality differ?

Omitted variable bias

Omitted variable bias

- Arises if we omit a variable that is a determinant of y and is correlated with x
- Thought experiment
  - RQ: Does NHH provide students with the skills that are rewarded in the labour market? Hypothetical example!

Thought experiment

- RQ: Does NHH provide students with the skills that are rewarded in the labour market?
- I hava a data set of individuals in Norway with two variables *salary* and *nhh*
  - *nhh* is a dummy variable indicating if someone attended NHH
  - *salary* is yearly salary in kNOK
- I estimate following regression:

$$salary_i = \hat{\alpha} + \hat{\beta} * nhh_i + u_i$$

- I get following estimates:

$$\widehat{salary}_i = 800 + 212 * nhh_i$$

- Great, but is there a problem? Are there some omitted variables?

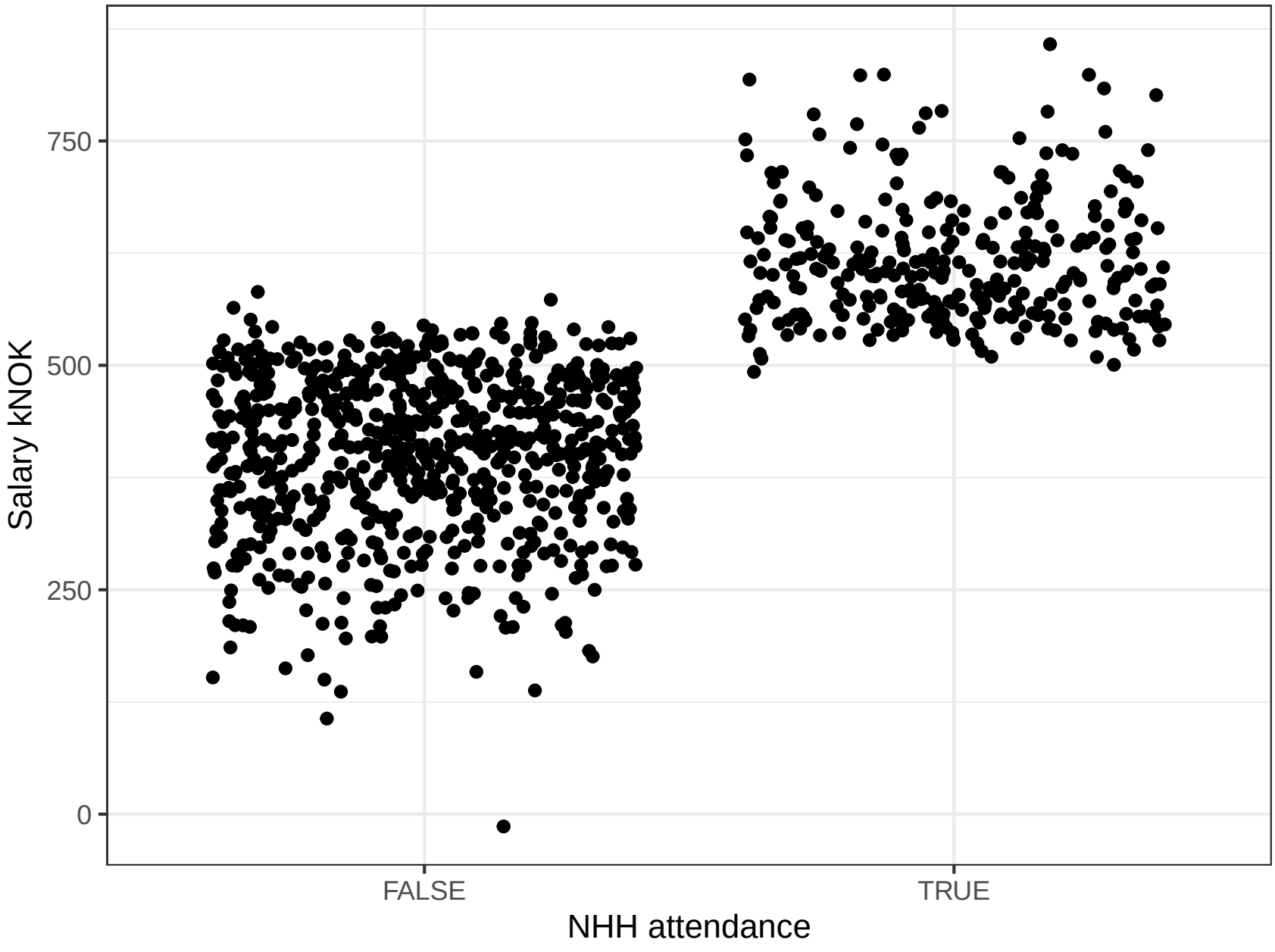
Population model

- Salary only depends on intelligence

$$salary = a + b_1 * nhh + b_2 * IQ + u$$

- with $a = 0$, $b_1 = 0$, and $b_2 = 4.25$
- NHH has no influence on the IQ of a person, but
- NHH selects students based on IQ
- or, NHH forces students to perform meaningless tasks that only students with an IQ over 125 pass (Michael Spence: Job market signaling (1973), Nobel prize laureate)

Plot



Derive omitted variable bias

- The population model:

$$y = \beta_0 + \beta_1 * x_i + \beta_2 * z_i + u_i$$

- But we estimate

$$y = \beta_0 + \beta_1 * x_i + \eta_i$$

- Hence, $\eta_i = \beta_2 * z_i + u_i$

Derive omitted variable bias

- Given Equation 4.25, the estimated coefficients $\hat{\beta}_1$ is defined by

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

- Similar to deriving LS.1,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 * z_i + u_i \;\Rightarrow\; y_i - \bar{y} = \beta_1(x_i - \bar{x}) + \beta_2(z_i - \bar{z}) + (u_i - \bar{u})$$

- We substitute back, to see if our $\hat{\beta}_1$ is an unbiased estimator of β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\left(\beta_1(x_i - \bar{x}) + \beta_2(z_i - \bar{z}) + (u_i - \bar{u})\right)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Derive omitted variable bias

- We simplify

$$\hat{\beta}_1 = \beta_1 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + \beta_2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + \frac{\sum_{i=1}^{n}(x_i - \bar{x})(u_i - \bar{u})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Taking expectations with respect to the data, using definition of covariance, and given $E(u|x) = 0$ in the population

$$\hat{\beta}_1 = \beta_1 + \frac{\text{cov}(x, z)}{\text{var}(x)} \beta_2$$

- As long as there is some correlation between the main dependent variable x and the omitted variable z, our estimate is biased

Derive omitted variable bias

- The zero conditional mean assumption is violated, as in our estimated model

$$E(\eta_i|X_1,\ldots,X_n) = E(\beta_2 * z_i + u_i|X_1,\ldots,X_n)$$
$$= \beta_2 * E(z_i|X_1,\ldots,X_n) + E(u_i|X_1,\ldots,X_n)$$

- If there is a non-zero covariance between x and z, then $E(\eta_i|X_1,\ldots,X_n) \neq 0$

Simultaneity bias

Simultaneity bias

- Arises when one or more of the independent variables are jointly determined with the dependent variable, typically through an equilibrium mechanism
- Examples
  - Quantity and price by demand and supply
  - Investment and productivity
  - Sales and advertizement
- Leads to violation of LS.1
- Hence coefficient is biased

Derive simultaneity bias

- Suppose we have a relationship expressed by following equations

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

$$x_i = \alpha_0 + \alpha_1 y_i + v_i$$

- By solving two equations, we have

$$y_i = \frac{\beta_0 + \beta_1 \alpha_0}{1 - \alpha_1 \beta_1} + \frac{\beta_1 v_i + u_1}{1 - \alpha_1 \beta_1}$$

$$x_i = \frac{\alpha_0 + \beta_0 \alpha_1}{1 - \alpha_1 \beta_1} + \frac{v_i + \alpha_1 u_1}{1 - \alpha_1 \beta_1}$$

Derive simultaneity bias

- Estimating following regression equation,

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

- Leeds to a bias, as $E(u|X_1, \ldots, X_n) \neq 0$, as

$$\text{cov}(x_i, u_i) = \text{cov}\left(\frac{v_i + \alpha_1 u_1}{1 - \alpha_1 \beta_1}, u_1\right) \neq 0$$

Sample selection bias

Sample selection bias

- Arises when a selection is influenced by a process related to the dependent variable
- Induces violation of LS.1
- Hence, OLS coefficient is biased
- For derivation see lecture on Difference in Difference estimation

Measurement error in independent variable

Measurement error in independent variable

- Data is often measured with error
  - reporting error
  - coding error
  - estimation error
- Violates LS.1
- Biases OLS estiamte towards zero (Attenuation bias)
- Measurement error in dependent variable is not problem

Derive bias due to measurement error

- Population regression model

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

- We observe x_i with an error e_i

$$\tilde{x}_i = x_i + e_i$$

- Assume that e_i is uncorrelated with x_i, $E(x_i e_i) = 0$
- Substitute into the regression equation

$$y_i = \beta_0 + \beta_1(\tilde{x}_i - e_i) + u_i = \beta_0 + \beta_1 \tilde{x}_i + \eta_i$$

- Where $\eta_i = u_i - \beta_1 e_i$

Derive bias due to measurement error

- Inserting η_i into equation 4.28

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(\tilde{x}_i - \bar{\tilde{x}})\eta_i}{\sum_{i=1}^{n}(\tilde{x}_i - \bar{\tilde{x}})^2} = \beta_1 + \frac{\sum_{i=1}^{n}(\tilde{x}_i - \bar{\tilde{x}})(u_i - \beta_1 e_i)}{\sum_{i=1}^{n}(\tilde{x}_i - \bar{\tilde{x}})^2}$$

- Re-arranging

$$\hat{\beta}_1 = \beta_1 - \beta_1 \frac{\sum_{i=1}^{n}(\tilde{x}_i - \bar{\tilde{x}})e_i}{\sum_{i=1}^{n}(\tilde{x}_i - \bar{\tilde{x}})^2} + \frac{\sum_{i=1}^{n}(\tilde{x}_i - \bar{\tilde{x}})u_i}{\sum_{i=1}^{n}(\tilde{x}_i - \bar{\tilde{x}})^2}$$

- Insterting x_i for $\tilde{x}_i$, simplifying, and taking conditional expectations with respect to the data,

$$E(\hat{\beta}_1 | X_1, \ldots, X_n) = \beta_1 - \beta_1 \frac{\text{var}(e)}{\text{var}(x) + \text{var}(e)}$$

- Given that variances are strictly positive, the estimate is smaller than the population parameter (Attenuation bias)

Summary

- Assumptions of single OLS regression
- Properties of OLS estimator
- Additional assumptions of multiple regression
- Standard error correction
- Biases

Textbook

- Own notes
- Chapter 1-3, 4, 5, Stock and Watson, Introduction to Econometrics, Global Edition, 4th edition