

# Omitted Variables<sup>1</sup>

STA442/2101 Fall 2013

---

<sup>1</sup>See last slide for copyright information.

## The fixed $x$ regression model

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,p-1} + \epsilon_i, \text{ with } \epsilon_i \sim N(0, \sigma^2)$$

- If viewed as conditional on  $\mathbf{X}_i = \mathbf{x}_i$ , this model implies independence of  $\epsilon_i$  and  $\mathbf{X}_i$  (homework).
- What is  $\epsilon_i$ ? *Everything else* that affects  $Y_i$ .
- So the usual model says that if the explanatory variables are random, they have *zero covariance* with all other variables that are related to  $Y_i$ , but are not included.
- For observational data, this assumption is almost always violated.
- Does it matter?

## Example

Suppose that the variables  $X_2$  and  $X_3$  have an impact on  $Y$  and are correlated with  $X_1$ , but they are not part of the data set. The values of the response variable are generated as follows:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \epsilon_i,$$

independently for  $i = 1, \dots, n$ , where  $\epsilon_i \sim N(0, \sigma^2)$ . The explanatory variables are random, with expected value and variance-covariance matrix

$$E \begin{bmatrix} X_{i,1} \\ X_{i,2} \\ X_{i,3} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} \quad \text{and} \quad V \begin{bmatrix} X_{i,1} \\ X_{i,2} \\ X_{i,3} \end{bmatrix} = \begin{bmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ & \phi_{22} & \phi_{23} \\ & & \phi_{33} \end{bmatrix},$$

where  $\epsilon_i$  is independent of  $X_{i,1}$ ,  $X_{i,2}$  and  $X_{i,3}$ .

## Absorb $X_2$ and $X_3$

Since  $X_2$  and  $X_3$  are not observed, they are absorbed by the intercept and error term.

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \epsilon_i \\ &= (\beta_0 + \beta_2 \mu_2 + \beta_3 \mu_3) + \beta_1 X_{i,1} + (\beta_2 X_{i,2} + \beta_3 X_{i,3} - \beta_2 \mu_2 - \beta_3 \mu_3 + \epsilon_i) \\ &= \beta'_0 + \beta_1 X_{i,1} + \epsilon'_i. \end{aligned}$$

And,

$$\text{Cov}(X_{i,1}, \epsilon'_i) = \beta_2 \phi_{12} + \beta_3 \phi_{13} \neq 0$$

# The “True” Model

Almost always true for observational data

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

where  $E(X_i) = \mu_x$ ,  $Var(X_i) = \sigma_x^2$ ,  $E(\epsilon_i) = 0$ ,  $Var(\epsilon_i) = \sigma_\epsilon^2$ , and  $Cov(X_i, \epsilon_i) = c$ .

Under this model,

$$\sigma_{xy} = Cov(X_i, Y_i) = Cov(X_i, \beta_0 + \beta_1 X_i + \epsilon_i) = \beta_1 \sigma_x^2 + c$$

# Estimate $\beta_1$ as usual

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2} \\ &\xrightarrow{\text{a.s.}} \frac{\sigma_{xy}}{\sigma_x^2} \\ &= \frac{\beta_1 \sigma_x^2 + c}{\sigma_x^2} \\ &= \beta_1 + \frac{c}{\sigma_x^2}\end{aligned}$$

$$\widehat{\beta}_1 \xrightarrow{a.s.} \beta_1 + \frac{c}{\sigma_x^2}$$

- $\widehat{\beta}_1$  is biased, even as  $n \rightarrow \infty$ .
- It's inconsistent.
- It could be almost anything, depending on the value of  $c$ , the covariance between  $X_i$  and  $\epsilon_i$ .
- The only time  $\widehat{\beta}_1$  behaves properly is when  $c = 0$ .
- Probability of Type I error goes almost surely to one. item  
What if  $\beta_1 < 0$  but  $\beta_1 + \frac{c}{\sigma_x^2} > 0$ ?

# All this applies to multiple regression

Of course

*When a regression model fails to include all the explanatory variables that contribute to the response variable, and those omitted explanatory variables have non-zero covariance with variables that are in the model, the regression coefficients are biased and inconsistent.*



# Correlation-Causation

- The problem of omitted variables is the technical version of the correlation-causation issue.
- The omitted variables are “confounding” variables.
- With random assignment and good procedure,  $x$  and  $\epsilon$  have zero covariance.
- But random assignment is not always possible.

This slide show was prepared by **Jerry Brunner**, Department of Statistics, University of Toronto. It is licensed under a **Creative Commons Attribution - ShareAlike 3.0 Unported License**. Use any part of it as you like and share the result freely. The  $\text{\LaTeX}$  source code is available from the course website:  
<http://www.utstat.toronto.edu/~brunner/oldclass/appliedf13>