Omitted Variables[1]

STA305 Fall 2016

[1]See last slide for copyright information.

The fixed x regression model

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

Think of the model as *conditional* given $\mathbf{X}_i = \mathbf{x}_i$.

Independence of ϵ_i and $\mathbf{X}_i$

- The statement $\epsilon_i \sim N(0, \sigma^2)$ is a statement about the *conditional* distribution of ϵ_i given $\mathbf{X}_i$.
- It says the density of ϵ_i given $\mathbf{X}_i$ does not depend on $\mathbf{X}_i$.
- For convenience, assume $\mathbf{X}_i$ has a density.

$$
\begin{aligned}
& f_{\epsilon|\mathbf{x}}(\epsilon|\mathbf{X}) = f_\epsilon(\epsilon) \\
\Rightarrow \quad & \frac{f_{\epsilon,\mathbf{x}}(\epsilon, \mathbf{x})}{f_{\mathbf{x}}(\mathbf{X})} = f_\epsilon(\epsilon) \\
\Rightarrow \quad & f_{\epsilon,\mathbf{x}}(\epsilon, \mathbf{X}) = f_{\mathbf{x}}(\mathbf{X}) f_\epsilon(\epsilon)
\end{aligned}
$$

Independence!

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 $\mathbf{x}_i$, this model implies independence of ϵ_i and $\mathbf{x}_i$, because the conditional distribution of ϵ_i given $\mathbf{x}_i$ does not depend on $\mathbf{x}_i$.
- What is ϵ_i? *Everything else* that affects y_i.
- So the usual model says that if the independent varables are random, they have *zero covaiance* with all other variables that are related to y_i, but are not included in the model.
- 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 dependent variable are generated as follows:

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

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

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

where ϵ_i is statistically 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.

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_2 X_{i,3} + \epsilon_i \\
&= (\beta_0 + \beta_2\mu_2 + \beta_3\mu_3) + \beta_1 X_{i,1} + (\beta_2 X_{i,2} + \beta_3 X_{i,3} - \beta_2\mu_2 - \beta_3\mu_3 + \epsilon_i) \\
&= \beta_0^* + \beta_1 X_{i,1} + \epsilon_i^*.
\end{aligned}
$$

And,

$$Cov(X_{i,1}, \epsilon_i^*) = \beta_2\phi_{12} + \beta_3\phi_{13} \neq 0$$

The "True" Model

Almost always closer to the truth than the usual model, 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 β_1 as usual

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{X})^2} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2} \\
&= \frac{\widehat{\sigma}_{xy}}{\widehat{\sigma}_x^2} \\
&\stackrel{a.s.}{\to} \frac{\sigma_{xy}}{\sigma_x^2} \\
&= \frac{\beta_1\sigma_x^2 + c}{\sigma_x^2} \\
&= \beta_1 + \frac{c}{\sigma_x^2}
\end{aligned}$$

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

- $\widehat{\beta}_1$ is biased (Exercise)
- It's inconsistent.
- It could be almost anything, depending on the value of c, the covariance between x_i and ϵ_i.
- The only time $\widehat{\beta}_1$ behaves properly is when $c = 0$.
- Test $H_0 : \beta_1 = 0$: Probability of Type I error goes almost surely to one.
- What if $\beta_1 < 0$ but $\beta_1 + \frac{c}{\sigma_x^2} > 0$, and you test $H_0 : \beta_1 = 0$?

All this applies to multiple regression

Of course

When a regression model fails to include all the independent variables that contribute to the dependent variable, and those omitted independent 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 ϵ have zero covariance.
- But random assignment is not always possible.
- Most applications of regression to observational data provide very poor information about the regression coefficients.
- Is bad information better than no information at all?
- Ultimately the solution is better data – *different* data.

Copyright Information