Omitted Variables and Instrumental Variables[1]

STA305 Fall 2020

[1] See last slide for copyright information.

Overview

Omitted Variables: A Practical Issue

- If you fit a regression model and then fit another model with additional x variables, anything can happen.
- $\widehat{\beta}_j$ values will change, and can even reverse sign.
- Tests that were significant can become non-significant.
- Tests that were non-significant can become significant.
- Tests that were significant in one direction can become significant in the other direction.
- This happens when the additional variables are related to y and also to the x variables that are already in the model.
- If your only interest is in prediction, who cares?
- If you are interested in the *meaning* of the results, it's a serious issue.
- Now we will examine this on a technical level.

The fixed x regression model

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

- $x_{i,j}$ fixed constants is unrealistic.
- Think of the model as *conditional* given the random vector $\mathcal{X}_i = \mathbf{x}_i$.
- All the expected values and probabilities in this course so far are conditional expected values and conditional probabilities.

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 (joint) 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 covariance* with all other variables that are related to y_i, but are not included in the model.
- For observational data (no random assignment), this assumption is almost always violated.
- Does it matter?

Example

Suppose that the explanatory 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:

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

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

$$E \left(\begin{array}{c} x_{i,1} \\ x_{i,2} \\ x_{i,3} \end{array} \right) = \left(\begin{array}{c} \mu_1 \\ \mu_2 \\ \mu_3 \end{array} \right) \quad \text{and} \quad cov \left(\begin{array}{c} x_{i,1} \\ x_{i,2} \\ x_{i,3} \end{array} \right) = \left(\begin{array}{ccc} \phi_{11} & \phi_{12} & \phi_{13} \\ & \phi_{22} & \phi_{23} \\ & & \phi_{33} \end{array} \right),$$

and 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

Recalling $Cov(x_i, y_i) = \beta_1 \sigma_x^2 + c$

$$
\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} \\
&= \frac{\frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2} \\
&= \frac{\widehat{\sigma}_{xy}}{\widehat{\sigma}_x^2} \\
&\stackrel{p}{\rightarrow} \frac{\sigma_{xy}}{\sigma_x^2} \quad \text{as } n \rightarrow \infty \\
&= \frac{\beta_1 \sigma_x^2 + c}{\sigma_x^2} \\
&= \beta_1 + \frac{c}{\sigma_x^2} \neq \beta_1 \text{ unless } c = 0.
\end{aligned}
$$

$\widehat{\beta}_1 \xrightarrow{p} \beta_1 + \frac{c}{\sigma_x^2}$

- $\widehat{\beta}_1$ is inconsistent, meaning it approaches the wrong target as $n \to \infty$.
- 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 $H_0 : \beta_1 = 0$, and the probability of Type I error goes to one as $n \to \infty$.
- 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 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 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?

How about another estimation method?
Other than ordinary least squares

- Can *any* other method be successful?
- This is a very practical question, because almost all regressions with observed (as opposed to manipulated) independent variables have the disease.

For simplicity, assume normality

$y_i = \beta_0 + \beta_1 y_i + \epsilon_i$

- Assume (x_i, ϵ_i) are bivariate normal.
- This makes (x_i, y_i) bivariate normal.
- $(x_1, y_1), \ldots, (x_n, y_n) \stackrel{i.i.d.}{\sim} N_2(\mathbf{m}, \mathbf{V})$, where

$$\mathbf{m} = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} = \begin{pmatrix} \mu_x \\ \beta_0 + \beta_1 \mu_x \end{pmatrix}$$

and

$$\mathbf{V} = \begin{pmatrix} v_{11} & v_{12} \\ & v_{22} \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & \beta_1 \sigma_x^2 + c \\ & \beta_1^2 \sigma_x^2 + 2\beta_1 c + \sigma_\epsilon^2 \end{pmatrix}.$$

- All you can ever learn from the data are the approximate values of $\mathbf{m}$ and $\mathbf{V}$.
- Even if you knew $\mathbf{m}$ and $\mathbf{V}$ exactly, could you know β_1?

Five equations in six unknowns

The parameter is $\theta = (\mu_x, \sigma^2_x, \sigma^2_\epsilon, c, \beta_0, \beta_1)$. The distribution of the data is determined by

$$\begin{pmatrix} m_1 \\ m_2 \end{pmatrix} = \begin{pmatrix} \mu_x \\ \beta_0 + \beta_1\mu_x \end{pmatrix} \text{ and } \begin{pmatrix} v_{11} & v_{12} \\ & v_{22} \end{pmatrix} = \begin{pmatrix} \sigma^2_x & \beta_1\sigma^2_x + c \\ & \beta_1^2\sigma^2_x + 2\beta_1 c + \sigma^2_\epsilon \end{pmatrix}$$

- $\mu_x = m_1$ and $\sigma^2_x = v_{11}$.
- The remaining 3 equations in 4 unknowns have infinitely many solutions.
- So infinitely many sets of parameter values yield the *same probability distribution of the sample data.*
- How could you decide which one is correct based on the sample data?
- The problem is fatal, if all you have is this data set.
- Ultimately the solution is better data – *different* data.

Instrumental Variables (Wright, 1928)

A partial solution

- An instrumental variable is a variable that is correlated with an explanatory variable, but is not correlated with any error terms and has no direct effect on the response variable.
- Usually, the instrumental variable *influences* the explanatory variable.
- An instrumental variable is often not the main focus of attention; it's just a tool.

A Simple Example

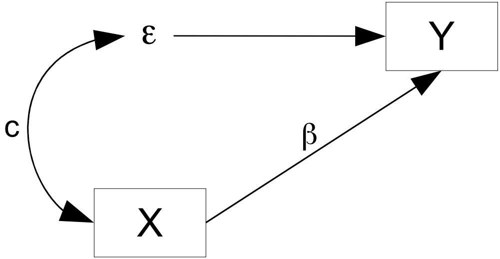
What is the contribution of income to credit card debt?

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

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

A path diagram

Again, $y_i = \alpha + \beta x_i + \epsilon_i$, where $E(x_i) = \mu$, $Var(x_i) = \sigma^2_x$, $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2_\epsilon$, and $Cov(x_i, \epsilon_i) = c$.



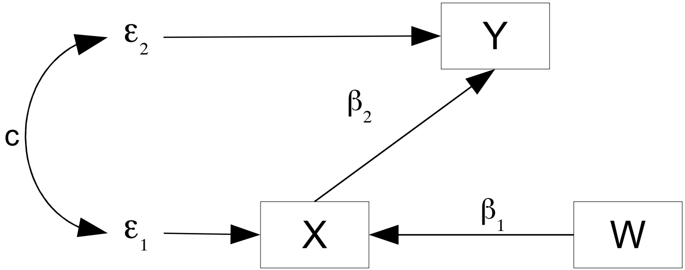
Least squares estimate of β is inconsistent, and so is every other possible estimate. If the data are normal.

Add an instrumental variable

x is income, y is credit card debt.

Focus the study on real estate agents in many cities. Include median price of resale home w_i.

$$x_i = \alpha_1 + \beta_1 w_i + \epsilon_{i1}$$
$$y_i = \alpha_2 + \beta_2 x_i + \epsilon_{i2}$$



Main interest is in β_2.

Base estimation and inference on the covariance matrix of (w_i, x_i, y_i): Call it $V = [v_{ij}]$

From $x_i = \alpha_1 + \beta_1 w_i + \epsilon_{i1}$ and $y_i = \alpha_2 + \beta_2 x_i + \epsilon_{i2}$,

$\mathbf{V} =$

	w	x	y
w	σ_w^2	$\beta_1\sigma_w^2$	$\beta_1\beta_2\sigma_w^2$
x		$\beta_1^2\sigma_w^2 + \sigma_1^2$	$\beta_2(\beta_1^2\sigma_w^2 + \sigma_1^2) + c$
y			$\beta_1^2\beta_2^2\sigma_w^2 + \beta_2^2\sigma_1^2 + 2\beta_2 c + \sigma_2^2$

$$\beta_2 = \frac{v_{13}}{v_{12}}$$

The remaining 5 equations in 5 unknowns have unique solutions too.

A close look

The v_{ij} are elements of the covariance matrix of the observable data.

$$\beta_2 = \frac{v_{13}}{v_{12}} = \frac{\beta_1 \beta_2 \sigma_w^2}{\beta_1 \sigma_w^2} = \frac{Cov(W,Y)}{Cov(W,X)}$$

- $\widehat{v}_{ij}$ are sample variances and covariances.
- $\widehat{v}_{ij} \stackrel{p}{\rightarrow} v_{ij}$.
- It is safe to assume $\beta_1 \neq 0$.
- Because it's the connection between real estate prices and the income of real estate agents.
- By continuous mapping, $\frac{\widehat{v}_{13}}{\widehat{v}_{12}} \stackrel{p}{\rightarrow} \frac{v_{13}}{v_{12}} = \beta_2$.
- That is, $\frac{\widehat{v}_{13}}{\widehat{v}_{12}}$ is a consistent estimate of β_2.
- $H_0 : \beta_2 = 0$ is true if and only if $v_{13} = 0$.
- Test $H_0 : v_{13} = 0$ by standard methods. `help(cor.test)`

Comments

- Good instrumental variables are not easy to find.
- They will not just happen to be in the data set, except by a miracle.
- They really have to come from another universe, but still have a strong and clear connection to the explanatory variables.
- Wright's original example was tax policy for cooking oil.
- Econometricians are good at this.
- Time series applications are common.
- Instrumental variables can help with measurement error in the explanatory variables too.
- The usual advice is at least one instrumental variable for each explanatory variable.

Copyright Information