Omitted Variables[1]

STA431 Winter/Spring 2017

[1] See last slide for copyright information.

Overview

A Practical Data Analysis Problem

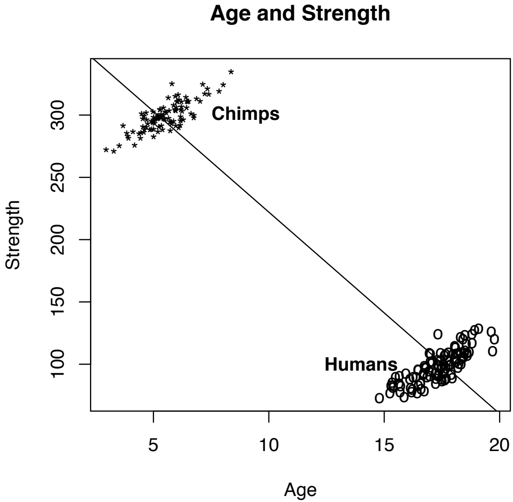
When more explanatory variables are added to a regression model and these additional explanatory variables are correlated with explanatory variables already in the model (as they usually are in an observational study),

- Statistical significance can appear when it was not present originally.
- Statistical significance that was originally present can disappear.
- Even the signs of the $\widehat{\beta}$s can change, reversing the interpretation of how their variables are related to the response variable.

An extreme, artificial example

To make a point

Suppose that in a certain population, the correlation between age and strength is $r = -0.93$.



The fixed x regression model

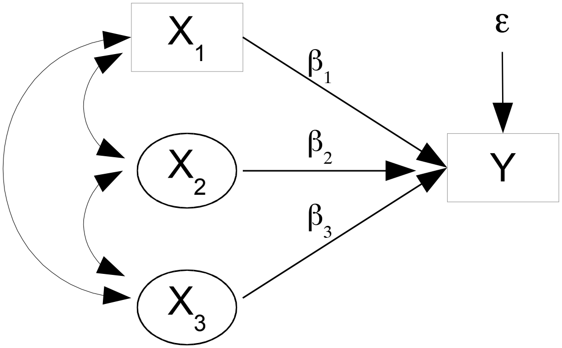
$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,p-1} + \epsilon_i, \text{ with } \epsilon_i \sim N(0, \sigma^2)$$

- If viewed as conditional on $\mathbf{X}_i = \mathbf{x}_i$, this model implies independence of ϵ_i and $\mathbf{X}_i$, because the conditional distribution of ϵ_i given $\mathbf{X}_i = \mathbf{x}_i$ 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 explanatory 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: $Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \epsilon_i$

As usual, the explanatory variables are random.

Suppose that the variables X_2 and X_3 affect Y and are correlated with X_1, but they are not part of the data set.



Statement of the model

The explanatory variables X_2 and X_3 affect 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 $i = 1, \ldots, n$, where $\epsilon_i \sim N(0, \sigma^2)$. The explanatory variables are random, with expected value and variance-covariance matrix

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

where ϵ_i is independent of $X_{i,1}$, $X_{i,2}$ and $X_{i,3}$. Values of the variables $X_{i,2}$ and $X_{i,3}$ are latent, and are not included in the data set.

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,

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

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 with least squares

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

$$
\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2} \\
&= \frac{\widehat{\sigma}_{xy}}{\widehat{\sigma}_x^2} \\
&\xrightarrow{p} \frac{\sigma_{xy}}{\sigma_x^2} \\
&= \frac{\beta_1 \sigma_x^2 + c}{\sigma_x^2} \\
&= \beta_1 + \frac{c}{\sigma_x^2}
\end{aligned}
$$

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

It converges to the wrong thing.

- $\widehat{\beta}_1$ is inconsistent.
- For large samples it could be almost anything, depending on the value of c, the covariance between X_i and ϵ_i.
- Small sample estimates could be accurate, but only by chance.
- The only time $\widehat{\beta}_1$ behaves properly is when $c = 0$.
- Test $H_0 : \beta_1 = 0$: Probability of making a Type I error goes to one as $n \to \infty$.

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 inconsistent. Estimation and inference are almost guaranteed to be misleading, especially for large samples.

Correlation-Causation

- The problem of omitted variables is a technical aspect 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 observational data have the disease.

For simplicity, assume normality
$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

- Assume (X_i, ϵ_i) are bivariate normal.
- This makes (X_i, Y_i) bivariate normal.
- $(X_1, Y_1), \ldots, (X_n, Y_n) \stackrel{i.i.d.}{\sim} N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \left(\begin{array}{c} \mu_1 \\ \mu_2 \end{array} \right) = \left(\begin{array}{c} \mu_x \\ \beta_0 + \beta_1 \mu_x \end{array} \right)$$

and

$$\boldsymbol{\Sigma} = \left(\begin{array}{cc} \sigma_{11} & \sigma_{12} \\ & \sigma_{22} \end{array} \right) = \left(\begin{array}{cc} \sigma_x^2 & \beta_1 \sigma_x^2 + c \\ & \beta_1^2 \sigma_x^2 + 2\beta_1 c + \sigma_\epsilon^2 \end{array} \right).$$

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

Five equations in six unknowns

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

$$\left(\begin{array}{c} \mu_1 \\ \mu_2 \end{array} \right) = \left(\begin{array}{c} \mu_x \\ \beta_0 + \beta_1\mu_x \end{array} \right) \text{ and } \left(\begin{array}{cc} \sigma_{11} & \sigma_{12} \\ & \sigma_{22} \end{array} \right) = \left(\begin{array}{cc} \sigma_x^2 & \beta_1\sigma_x^2 + c \\ & \beta_1^2\sigma_x^2 + 2\beta_1 c + \sigma_\epsilon^2 \end{array} \right)$$

- $\mu_x = \mu_1$ and $\sigma_x^2 = \sigma_{11}$.
- The remaining 3 equations in 4 unknowns have infinitely many solutions.
- So infinitely many sets of parameter values yield the *same distribution of the sample data*.
- This is serious trouble – lack of parameter identifiability.
- *Definition*: If a parameter is a function of the distribution of the observable data, it is said to be *identifiable*.

Showing identifiability

Definition: If a parameter is a function of the distribution of the observable data, it is said to be identifiable.

- How could a parameter be a function of a distribution?
- $D \sim F_\theta$ and $\theta = g(F_\theta)$
- Usually g is defined in terms of moments.
- Example: $F_\theta(x) = 1 - e^{-\theta x}$ and $f_\theta(x) = \theta e^{-\theta x}$.

$$\begin{aligned} f_\theta(x) &= \frac{d}{dx} F_\theta(x) \\ E(X) &= \int_0^\infty x f_\theta(x)\, dx = \frac{1}{\theta} \\ \theta &= \frac{1}{E(X)} \end{aligned}$$

Sometimes people use moment-generating functions or characteristic functions instead of just moments.

Showing identifiability is like Method of Moments Estimation

- The distribution of the data is always a function of the parameters.
- The moments are always a function of the distribution of the data.
- If the parameters can be expressed as a function of the moments,
  - Put hats on to obtain MOM estimates, or observe that
  - The parameter is a function of the distribution, and is identifiable.

Back to the five equations in six unknowns

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

$$\mathbf{D}_i = \left(\begin{array}{c} X_i \\ Y_i \end{array} \right) \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ where}$$

$$\begin{aligned}
\boldsymbol{\mu} &= \left(\begin{array}{c} \mu_1 \\ \mu_2 \end{array} \right) &= \left(\begin{array}{c} \mu_x \\ \beta_0 + \beta_1 \mu_x \end{array} \right) \\
\boldsymbol{\Sigma} &= \left(\begin{array}{cc} \sigma_{11} & \sigma_{12} \\ \cdot & \sigma_{22} \end{array} \right) &= \left(\begin{array}{cc} \sigma_x^2 & \beta_1 \sigma_x^2 + c \\ \cdot & \beta_1^2 \sigma_x^2 + 2\beta_1 c + \sigma_\epsilon^2 \end{array} \right)
\end{aligned}$$

We have expressed the moments in terms of the parameters, but we can't solve for $\theta = (\mu_x, \sigma_x^2, \sigma_\epsilon^2, c, \beta_0, \beta_1)$.

Skipping the High School algebra

$\theta = (\mu_x, \sigma_x^2, \sigma_\epsilon^2, c, \beta_0, \beta_1)$

- For *any* given $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, all the points in a one-dimensional subset of the 6-dimensional parameter space yield $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and hence the same distribution of the sample data.
- In that subset, values of β_1 range from $-\infty$ to $-\infty$, so $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ could have been produced by *any* value of β_1.
- There is no way to distinguish between the possible values of β_1 based on sample data.
- The problem is fatal, if all you can observe is X and Y.
- See text for details.

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 connection to the response variable.
- In Econometrics, the instrumental variable usually *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?

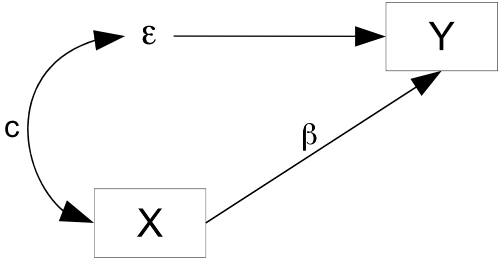
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

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

$$Y_i = \alpha + \beta X_i + \epsilon_i,$$

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



Least squares estimate of β is inconsistent, and so is every other possible estimate. This is strictly true if the data are normal.

Add an instrumental variable

Definition: An instrumental variable for an explanatory variable is another random variable that has non-zero covariance with the explanatory variable, and *no direct connection with any other variable in the model.*

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

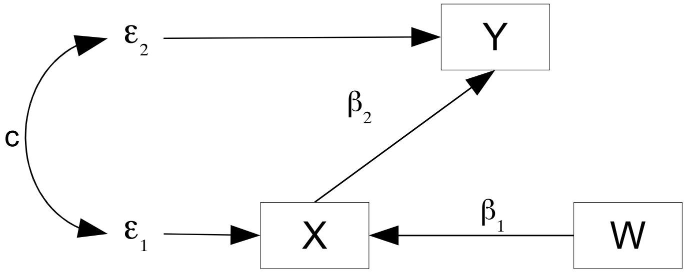
- X is income.
- Y is credit card debt.
- W is median price of resale home.

$$\begin{aligned} X_i &= \alpha_1 + \beta_1 W_i + \epsilon_{i1} \\ Y_i &= \alpha_2 + \beta_2 X_i + \epsilon_{i2} \end{aligned}$$

Picture

W_i is median price of resale home, X_i is income, Y_i is credit card debt.

$$\begin{aligned} X_i &= \alpha_1 + \beta_1 W_i + \epsilon_{i1} \\ Y_i &= \alpha_2 + \beta_2 X_i + \epsilon_{i2} \end{aligned}$$



Main interest is in β_2.

Calculate the covariance matrix

of the observable data (W_i, X_i, Y_i): Call it $\boldsymbol{\Sigma} = [\sigma_{ij}]$

From $X_i = \alpha_1 + \beta_1 W_i + \epsilon_{i1}$ and $Y_i = \alpha_2 + \beta_2 X_i + \epsilon_{i2}$,

$\boldsymbol{\Sigma} =$

	W	X	Y
W	σ_w^2	$\beta_1\sigma_w^2$	$\beta_1\beta_2\sigma_w^2$
X	$\cdot$	$\beta_1^2\sigma_w^2 + \sigma_1^2$	$\beta_2(\beta_1^2\sigma_w^2 + \sigma_1^2) + c$
Y	$\cdot$	$\cdot$	$\beta_1^2\beta_2^2\sigma_w^2 + \beta_2^2\sigma_1^2 + 2\beta_2 c + \sigma_2^2$

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

Parameter estimation

$X_i = \alpha_1 + \beta_1 W_i + \epsilon_{i1}$ and $Y_i = \alpha_2 + \beta_2 X_i + \epsilon_{i2}$

$\boldsymbol{\Sigma} =$

	W	X	Y
W	σ_w^2	$\beta_1\sigma_w^2$	$\beta_1\beta_2\sigma_w^2$
X	$\cdot$	$\beta_1^2\sigma_w^2 + \sigma_1^2$	$\beta_2(\beta_1^2\sigma_w^2 + \sigma_1^2) + c$
Y	$\cdot$	$\cdot$	$\beta_1^2\beta_2^2\sigma_w^2 + \beta_2^2\sigma_1^2 + 2\beta_2 c + \sigma_2^2$

$\beta_2 = \frac{\sigma_{13}}{\sigma_{12}}.$

- All the other parameters are identifiable too.
- The instrumental variable saved us.
- There are 9 model parameters, and 9 moments in $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.
- The invariance principle yields explicit formulas for the MLEs.
- If the data are normal, MLEs equal the Method of Moments estimates because they are both 1-1 with the moments.

More Comments

- Of course there is measurement error.
- Instrumental variables help with measurement error as well as with omitted variables.
- More later.
- 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 variable.
- Data collection has to be *planned*.
- Wright's original example was tax policy for cooking oil.
- Time series applications are common, but not in this course.

Copyright Information