Correlation does not imply causation

Campbell Miller

Due by Noon on Friday, November 5th

1. Signing Omitted Variable Bias

a. Suppose people start drinking diet soda when they find out they have diabetes. How will this bias estimates of the effect of diet soda consumption on health?

The people who are drinking diet sode are then proportionally more unhealthy people who switch to diet after hearing they have diabetes. This will lower the average health of diet soda drinkers compared to normal soda drinkers and so this bias may show that diet soda is more unhealthy than it actually is.

b. Suppose a researcher does not control for parent education in a regression of earnings on years of education. Will the resulting bias likely be positive or negative?

The resulting bias will likely be positive as parent education is likely a factor that would increase a childs earnings so this increase will thus be shown as a positive effect of years of education while it is actually from parents education.

c. Suppose doctor's advise individuals with severe anxiety or depression to buy a dog. Would failing to control for this create positive or negative omitted variable bias?

If you were looking at how owning a dog lowers depression, failing to control for this fact would be a negative bias. The doctor suggestion would make a larger proportion of depressed people get a dog so the effect of a dog on lowering depression would be much smaller than the actual effect since it is starting out with a higher depressed proportion.

d. What has to be true for an omitted variable to bias the returns to schooling downward? Can you think of an example of a variable meeting these criteria? Why is it difficult to think of one?

To bias the returns to school downward, the omitted variable would have to be correlated with a downward trend of schooling and an explanatory variable. It could be possible to tell if an effect made the effects of schooling downward or upward as many effects could do both such as higher wealth allowing you tob uy more things to distract you but also hire tutors or other help for schooling.

2. Testing the Conditional Independence Assumption

For this exercise, we will simulate the following data generating process:

$$Y_i = 2 + 3T_i + \varepsilon_i$$

Where:

$$\varepsilon_i = -5\mu_i + e_i$$

$$P(T_i = 1) = 0.5\mu_i + u_i$$

$$\begin{gathered}\mu_i \sim Uniform(0.4, 0.6)\\ u_i \sim Uniform(0, 0.3)\\ e_i \sim \mathcal{N}(0, 1)\end{gathered}$$

a. Generate a dataset with 1,000 observations from this data generating process. Hint: You can draw from the uniform and normal distirbutions using the .mono[rnorm] and .mono[runif] functions.

Answer

```
#X <- sample(c(0,1), 1000, replace=TRUE)

#Pt <- .5 + runif(1000, min = 0.40, max = 0.60) + runif(1000, min = 0, max = 0.30)
#summary(Pt)

#T <- rbinom(1000,1,Pt)
set.seed(123123)
e <- rnorm(1000, mean = 0, sd = 1)

u <- runif(1000, 0, 0.30)
mu <- runif(1000, 0.40, 0.60)

pt <- (0.5 * mu) + u

epsilon = -5 * mu + e

T <- rbinom(1000, 1, pt)

Y = 2 + 3*T + epsilon
```

b. Use the `cov()` command to estimate

$$\hat{Cov}(T_i, \varepsilon_i)$$

in your simulated data.

Answer

Note you can put R code in text like: .

```
cov(T, epsilon)
```

```
## [1] -0.02828357
```

c. Now, regress Y on T. Report the results in a nice table. Briefly interpret the results.

Answer

```
reg1 <- lm(Y ~ T)

library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
stargazer(reg1, keep.stat = c("n", "rsq"),
          covariate.labels = c(),
          dep.var.labels = c(),
          title = "Regression of Y on T",
          type = "text", style = "qje")
```

```
##
## Regression of Y on T
## ===============================================
##                             Y
## -----------------------------------------------
## T                       2.883***
##                         (0.067)
##
## Constant               -0.434***
##                         (0.043)
##
## N                        1,000
## R2                       0.652
## ===============================================
## Notes:   ***Significant at the 1 percent level.
##           **Significant at the 5 percent level.
##           *Significant at the 10 percent level.
```

We know that the true value of Y should be 3 (when T=1 (3*1)).This regression shows an estimated value of Y as 2.883 which is very close to the expected value.

d. We saw in class that we can interpet the coefficient from a regression as the causal effect of T on Y if the conditional independence assumption is satisifed, i.e. if $Cov(\varepsilon_i, T_i) = 0$.

Calculate the residuals from your regression of Y on T using the fitted values saved in your regression output (e.g. `m1$fitted`). What is the covariance between the estimated residuals and T_i? How does this compare to the true covariance between T_i and ε_i? Why does this make it difficult to test the conditional independence assumption?

Answer

```
estResid <- resid(reg1)

cov(T, estResid)
```

```
## [1] 9.298189e-17
```

This cov shows a very very small number while the true cov is a bit bigger at -.028. Our cov with estimated residuals is essentially 0 so it would say we basically satisfy the conditional independence assumption. However we know there is noise in our equation through epsilon so the value we get should not be so close to the expected value of 3 so this cov of ~0 would tell us that the conditional independence assumption was met and this would be false. We cannot properly measure the bias.

3. Evaluating causal claims

Find an article (whereever is convenient to look - causal claims are everywhere!) that makes a causal claim. Assess whether the implied causal effect may suffer from omitted variable bias.

https://www.psychologytoday.com/us/blog/media-spotlight/201809/video-games-school-success-and-your-child

This article looks at how gaming affects school performance but there are a number areas that could be omitted variable bias that the article does not say were controlled for. Gender is correlated with school performance as generally girls and boy excel at various subjects and this study based school performance on different scores of tests. Gender also is correlated with gaming as generally boy will game more. Another possible omitted vairable bias is parent wealth. More wealthy families could affor tutoring and extra help to prepare students which would affect the scores they get. Wealth would also affect gaming as being able to purchase the newest gear would encourage a child to play longer and more.