In-Class Lab 5

ECON 4223 (Prof. Tyler Ransom, U of Oklahoma)

February 3, 2022

The purpose of this in-class lab is to better understand omitted variable bias and multicollinearity. The lab should be completed in your group. To get credit, upload your .R script to the appropriate place on Canvas.

For starters

Open up a new R script (named `ICL5_XYZ.R`, where `XYZ` are your initials) and add the usual "preamble" to the top:

```
library(tidyverse)
library(broom)
library(wooldridge)
library(modelsummary)
```

Also install the package `car` by typing **in the console**:

```
install.packages("car", repos='http://cran.us.r-project.org')
```

and then add to the preamble of your script

```
library(car)
```

The `car` package allows us to easily compute useful statistics for diagnosing multicollinearity.

Load the data

We'll use a new data set on wages, called `wage2`.

```
df <- as_tibble(wage2)
```

Check out what's in the data by typing

```
glimpse(df)
# or, equivalently
datasummary_df(df)
```

We can also look at summary statistics in the data by typing

```
datasummary_skim(df,histogram=FALSE)
```

Properties of Omitted Variables

Think of the following regression model:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 IQ + u$$

where *wage* is a person's hourly wage rate (in cents, not dollars).

We want to verify the property in Wooldridge (2015) that $\widetilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \widetilde{\delta}_1$, where $\widetilde{\beta}_1$ comes from a regression of $log(wage)$ on $educ$, $\widetilde{\delta}_1$ comes from a regression of IQ on $educ$, and the $\hat{\beta}$'s come from the full regression (in the equation above).

First, run a regression of `IQ` on `educ` to obtain $\widetilde{\delta}_1$:

```
est1 <- lm(IQ ~ educ, data=df)
tidy(est1)
```

Now run a regression of `log wage` on `educ` to obtain $\widetilde{\beta}_1$. **Note: You'll need to create the log wage variable first. If you can't remember how to do that, refer back to previous labs.**

```
est2 <- lm(logwage ~ educ, data=df)
tidy(est2)
```

Now run the full regression of `log wage` on `educ` and `IQ` to obtain $\hat{\beta}_1$ and $\hat{\beta}_2$. Verify that $\widetilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \widetilde{\delta}_1$.

```
est3 <- lm(logwage ~ educ + IQ, data=df)
tidy(est3)
est2$coefficients["educ"] == est3$coefficients["educ"] +
    est3$coefficients["IQ"]*est1$coefficients["educ"]
```

(The last line returns `TRUE` if the equality holds and `FALSE` if it doesn't hold.)

We can also look at the output with `modelsummary()`:

```
modelsummary(
             list(est1,est2,est3)
            )
```

Is $\widetilde{\beta}_1$ larger or smaller than $\hat{\beta}_1$? What does this mean in terms of omitted variable bias?

Multicollinearity

Now let's see how to compute diagnostics of multicollinearity. Recall from Wooldridge (2015) that multicollinearity can better be thought of as "a problem with small sample sizes." Let's use the `meapsingle` data set from the `wooldridge` package. We are interested in the variable `pctsgle` which gives the percentage of single-parent families residing in the same ZIP code as the school. The outcome variable is `math4`, which is the percentage of students at the school who passed the 4th grade state test in math.

Load the data and run a regression of `math4` on `pctsgle`. (I won't include the code, since this should be old hat by now.) Interpret the slope coefficient of this regression. Does the effect seem large?

Now consider the same model, but with `lmedinc` and `free` as additional regressors. `lmedinc` is the log median household income of the ZIP code, and `free` is the percent of students who qualify for free or reduced-price lunch. Do you think there might be a strong correlation between `lmedinc` and `free`? Compute the correlation. Does it have the sign you would expect? Do you think it's close enough to 1 that it would violate the "no perfect collinearity" assumption?

```
cor(df$lmedinc,df$free)
```

Now run the model with `pctsgle`, `lmedinc`, and `free` as regressors. (Again, I won't include the code here.)

Comment on the value of the `pctsgle` coefficient, compared to the first regression you ran. What can you say about `lmedinc` and `free` as confounding variables?

Computing variance inflation factors (VIF)

A commonly used diagnostic of multicollinearity is the VIF. We can use the `vif()` function from the `car` package to do this. Let's compute the VIF from our estimates in the previous equation:

```
vif(est)
```

VIFs of 10 or more are typically thought to be problematic, because $VIF = \frac{1}{1-R_j^2}$, meaning $R_j^2 > 0.9$. See p. 86 of Wooldridge (2015).

Is multicollinearity a problem?

Multicollinearity is typically only a problem in data sets of small sample size. As sample size increases, R_j^2 might decrease. Also, the total variation in x_j (SST_j) increases with sample size. So multicollinearity is typically not a problem we worry about much.

References

Wooldridge, Jeffrey M. 2015. *Introductory Econometrics: A Modern Approach*. 6th ed. Cengage Learning.