# 3.4 — Multivariate OLS Estimators: Bias, Precision, and Fit — R Practice

ECON 480 — Fall 2021

Thursday, October 28, 2021

## Getting Set Up

Before we begin, start a new file with `File` → `New File` → `R Script`. As you work through this sheet in the console in `R`, also add (copy/paste) your commands that work into this new file. At the end, save it, and run to execute all of your commands at once.

First things first, load `tidyverse`.

```r
library("tidyverse")
```

1. Download and read in (`read_csv`) the data below.

- `speeding_tickets.csv`

This data comes from a paper by Makowsky and Strattman (2009) that we will examine later. Even though state law sets a formula for tickets based on how fast a person was driving, police officers in practice often deviate from that formula. This dataset includes information on all traffic stops. An amount for the fine is given only for observations in which the police officer decided to assess a fine. There are a number of variables in this dataset, but the one's we'll look at are:

| Variable | Description |
|---|---|
| `Amount` | Amount of fine (in dollars) assessed for speeding |
| `Age` | Age of speeding driver (in years) |
| `MPHover` | Miles per hour over the speed limit |

We want to explore who gets fines, and how much. We'll come back to the other variables (which are categorical) in this dataset in later lessons.

2. *How does the age of a driver affect the amount of the fine*? Make a scatterplot of the `Amount` of the fine (`y`) and the driver's `Age` (`x`) along with a regression line.

3. Next, we'll want to find the correlation between `Amount` and `Age`. Do this first.

Then notice that it won't work. This is because there are a lot of `NA`s (missing data) for `Amount` (if tried to get the `mean()` of `Amount`, it would do the same thing.

You can verify the `NA`s with:

```r
data %>% # use your named dataframe!
  select(Amount) %>%
  summary()

# OR
# data %>% count(Amount) # but this has a lot of rows!
```

In order to run a correlation, we need to drop or ignore all of the `NA`s. You could `filter()` the data:

```
# this would OVERWRITE data
data <- data %>%
  filter(!is.na(Amount)) # remove all NAs
```

Or, if you don't want to change your data, the `cor()` command allows you to set `use = "pairwise.complete.obs"` as an argument.

4. We want to estimate the following model:

$$\widehat{\text{Amount}}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{Age}_i$$

Run a regression, and save it as an object. Then get a `summary()` of it.

a. Write out the estimated regression equation.

b. What is $\hat{\beta}_0$ for this model? What does it mean in the context of our question?

c. What is $\hat{\beta}_1$ for this model? What does it mean in the context of our question?

d. What is the marginal effect of `Age` on `Amount`?

5. Redo question 4 with the `broom` package. Try out `tidy()` and `glance()`. This is just to keep you versatile!

6. How big would the difference in expected fine be for two drivers, one 18 years old and one 40 years old?

7. Now run the regression again, controlling for speed (`MPHover`).

a. Write the new regression equation.

b. What is the marginal effect of `Age` on `Amount`? What happened to it, compared to Question 4D?

c. What is the marginal effect of `MPHover` on `Amount`?

d. What is $\hat{\beta}_0$ for our model, and what does it mean in the context of our question?

e. What is the adjusted $\bar{R}^2$? What does it mean?

8. Now suppose both the 18 year old and the 40 year old each went 10 MPH over the speed limit. How big would the difference in expected fine be for the two drivers?

9. What is the difference in expected fine between two 18 year-olds, one who went 10 MPH over, and one who went 30 MPH over?

10. Use the `huxtable` package's `huxreg()` command to make a regression table of your two regressions: the one from question 4, and the one from question 7.

11. Are our two independent variables multicollinear? Do younger people tend to drive faster?

a. Get the correlation between `Age` and `MPHover`.

b. Make a scatterplot of `MPHover` (`y`) on `Age` (`x`).

c. Run an auxiliary regression of `MPHover` on `Age`.

d. Interpret the coefficient on `Age`.

e. Look at your regression table in question 10. What happened to the standard error on `Age`? Why (consider the formula for variance of $\hat{\beta}_1$)?

f. Calculate the Variance Inflation Factor (VIF) using the `car` package's `vif()` command. Run it on your regression object saved from Question 7.

g. Calculate the VIF manually, using what you learned in this question.

12. Let's now think about the omitted variable bias. Suppose the "true" model is the one we ran from Question 7.

a. Do you suppose that `MPHover` fits the two criteria for omitted variable bias?

b. Look at the regression we ran in Question 4. Consider this the "omitted" regression, where we left out `MPHover`. Does our estimate of the marginal effect of `Age` on `Amount` overstate or understate the *true* marginal effect?

c. Use the "true" model (Question 7), the "omitted" regression (Question 4), and our "auxiliary" regression (Question 11) to identify each of the following parameters that describe our biased estimate of the marginal effect of `Age` on `Amount`:

$$\alpha_1 = \beta_1 + \beta_2 \delta_1$$

See the notation I used in class.

d. From your answer in part C, how large is the omitted variable bias from leaving out `MPHover`?

13. Make a coefficient plot of your coefficients from the regression in Question 7. The package `modelsummary` (which you will need to install and load) has a great command `modelplot()` to do this on your regression object.