R Practice 3

Ryan Safner

11/13/2018

Download the `speeding_tickets.csv` dataset from Blackboard (under Data). 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.

- `Amount`: Amount of fine assessed for speeding
- `Age`: Age of speeder in years
- `MPHover`: Miles per hour over speed limit
- `Black`: = 1 if driver was black, = 0 if not
- `Hispanice`: = 1 if driver was Hispanic, = 0 if not
- `Female`: = 1 if driver was female, = 0 if not
- `OutTown`: = 1 if driver was not from local town, = 0 if not
- `OutState`: = 1 if driver was not from local state, = 0 if not
- `StatePol`: = 1 if driver was stopped by State Police, = 0 if stopped by other (local)

We want to explore who gets fines, and how much.

1. Load the data and inspect it briefly with `str()` and `head()`. We will have to do a little bit of cleaning to get the data in a more usable form.

a. What class of variable are `Black`, `Hispanic`, `Female`, `OutTown`, and `OutState`?

b. Notice that when importing the data from the `.csv` file, `R` interpretted these variables as `integer`, but we want them to be `factor` variables, to ensure `R` recognizes that there are two groups (categories), 0 and 1. Convert each of these variables into factors by reassigning it according to the format: `df$var.name<-as.factor(df$var.name)`, where

- `df` is the name of your data frame
- `var.name` is the name of the variable

c. Confirm they are each now factors by checking their class again.

2. Create a scatterplot between Amount and Female. Use `geom_jitter()` instead of `geom_point()` to plot the points, and play around with `width` settings inside `geom_jitter()`.

3. Check the distribution of `Amount` with `summary()`.

a. If you notice, `Amount` has a lot of missing values (for people that did not get fined). Let's keep only data for which `Amount` is a positive number. Use the `subset()` command and overwrite your data (or make a new object) with `df1<-subset(df,condition)`

b. Double check this worked by checking the `summary()` of `Amount` again.

4. Find the mean and standard deviation of Amount, for male drivers and for female drivers.

a. What is the difference between the average speed for Males and Females? (Calculate this manually)

b. Use `t.test` to check if this is a statistically significant difference. The syntax is similar for regression: `t.test(y~d, data=df)` where `y` is the variable we are testing (`Amount`) and `d` is the dummy variable (`Female`)

5. Now run the following regression to ensure we get the same result

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

a. Write out the estimated regression equation.

$$\text{Am}\hat{\text{o}}\text{unt}_i = 124.67 - 7.94\text{Female}_i$$

b. Use the regression coefficients to find (i) the average `Amount` for men, (ii) the average `Amount` for women, and (iii) the difference in average `Amount` between men and women

- Males get fined \$124.67 ($\hat{\beta}_0$)
- Females get fined $124.67 - 7.94 = 116.73$ ($\hat{\beta}_0 + \hat{\beta}_1$)
- The difference is $-\$7.94$ ($\hat{\beta}_3$)

6. Let's recode the sex variable. Make a new variable called `Male` and use the `ifelse()` function to define it as 1 when `df$Female==0` and 0 otherwise.

a. Run the same regression as in question 5, but use `Male` instead of `Female`.

b. Write out the estimated regression equation.

$$\widehat{Amount} = 116.73 + 7.94 Male$$

c. Use the regression coefficients to find (i) the average `Amount` for men, (ii) the average `Amount` for women, and (iii) the difference in average `Amount` between men and women

- Females get fined \$116.73 ($\hat{\beta}_0$)
- Males get fined $116.73 + 7.94 = \$124.67$ over ($\hat{\beta}_0 + \hat{\beta}_1$)
- The difference is 7.94 ($\hat{\beta}_3$)

7. Run a regression of `Amount` on `Male` and `Female`. What happens, and why?

Male and Female are perfectly multicollinear, as for every person i, Male_i+Female_i=1. We can confirm this by seeing the correlation between Male and Female is exactly -1. To run a regression, we must exclude one of the dummies, and as we've seen, it makes no difference which one we exclude.

8. Age probably has a lot to do with differences in fines, perhaps also age affects fines differences between males and females. Run a regression of Amount on Age and Female. How does the coefficient on Female change?

a. Now let's see if the difference in fine between men and women are different depending on the driver's age. Run the regression again, but add an interaction term between `Female` and `Age` interaction term.

b. Write out your estimated regression equation.

c. Interpret the interaction effect. Is it statistically significant?

d. Plugging in 0 or 1 as necessary, rewrite (on your paper) this regression as *two separate* equations, one for Males and one for Females.

f. Let's try to visualize this. Make a scatterplot of `Age` (X) and `Amount` (Y) and include a regression line.

g. Try adding to your base layer `aes()`, set `color=Female`. This will produce two lines and color the points by `Female`. Sometimes we may also need to remind `R` that `Female` is a factor with `as.factor(Female)`.

h. Add a facet layer to make two different scatterplots with an additional layer `+facet_grid(cols=vars(Female))`

9. Now let's look at the possible interaction between Sex (`Male` or `Female`) and whether a driver is from In-State or Out-of-State (`OutState`).

a. Use `R` to examine the data and find the mean for (i) Males In-State, (ii) Males Out-of-State, (iii) Females In-State, and (iv) Females Out-of-State. Hint: use `&` to join multiple conditions!

b. Now run a regression of the following model:

$$\text{Amount}_i = \hat{\beta}_0 + \hat{\beta}_1 Female_i + \hat{\beta}_2 OutState_i + \hat{\beta}_3 Female_i * OutState_i$$

c. Write out the estimated regression equation.

$$\widehat{Amount} = 123.68 - 8.88 Female + 4.29 OutState + 5.17 Female \times OutState$$

d. What does each coefficient mean?

- $\hat{\beta}_0 = \$123.68$; mean for in-state males
- $\hat{\beta}_1 = -\$8.88$: difference between in-state males and females
- $\hat{\beta}_2 = \$4.29$: difference between males in-state vs. out-of-state
- $\hat{\beta}_3 = \$5.17$: difference between effect of being in-state vs. out-of-state between males vs. females (or, equivalently, difference between effect of being male vs. female between in-state vs. out-of-state)

e. Using the regression equation, what are the means for (i) Males In-State, (ii) Males Out-of-State, (iii) Females In-State, and (iv) Females Out-of-State? Compare to your answers in part a.

$$\widehat{Amount} = 123.68 - 8.88 Female + 4.29 OutState + 5.17 Female * OutState$$

- Males In-State: $\hat{\beta}_0 = \$123.68$
- Males Out-of-State: $\hat{\beta}_0 + \hat{\beta}_2 = 123.68 + 4.29 = \$127.97$
- Females In-State: $\hat{\beta}_0 + \hat{\beta}_1 = 123.68 - 8.88 = \$114.80$
- Females Out-of-State: $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 = 123.68 - 8.88 + 4.29 + 5.17 = \$124.26$

10. Collect your regressions from questions 5, 6a, 8, 8a, and 9b and output them in a regression table with `stargazer`.