Problem Set 4

QTM 200: Applied Regression Analysis

Due: February 24, 2020

Instructions

- Please show your work! You may lose points by simply writing in the answer. If the problem requires you to execute commands in `R`, please include the code you used to get your answers. Please also include the `.R` file that contains your code. If you are not sure if work needs to be shown for a particular problem, please ask.
- Your homework should be submitted electronically on the course GitHub page in `.pdf` form.
- This problem set is due at the beginning of class on Monday, February 24, 2020. No late assignments will be accepted.
- Total available points for this homework is 100.

Question 1 (50 points): Economics

In this question, use the `prestige` dataset in the `car` library. First, run the following commands:

```
install.packages(car)
library(car)
data(Prestige)
help(Prestige)
```

We would like to study whether individuals with higher levels of income have more prestigious jobs. Moreover, we would like to study whether professionals have more prestigious jobs than blue and white collar workers.

(a) Create a new variable `professional` by recoding the variable `type` so that professionals are coded as 1, and blue and white collar workers are coded as 0 (Hint: `ifelse`.)

```
Prestige$type
professional <- ifelse(Prestige$type=="prof", 1, 0)
```

(b) Run a linear model with prestige as an outcome and `income`, `professional`, and the interaction of the two as predictors (Note: this is a continuous × dummy interaction.)

```
regression1<-lm(prestige ~ income + professional + income:professional,
    data = Prestige)
regression1
```

(c) Write the prediction equation based on the result.

$\hat{Y}$=21.142259 + 0.003171(income) + 37.781280(professional) -0.002326 (income:professional)

(d) Interpret the coefficient for `income`.

Holding all other variables in the equation constant, each unit increase in income is associated with a 0.003171 unit increase in prestige score.

(e) Interpret the coefficient for `professional`.

Holding all other variables in the equation constant, being a professional is associated with a 37.781280 unit increase in prestige score.

(f) What is the effect of a $1,000 increase in income on prestige score for professional occupations? In other words, we are interested in the marginal effect of income when the variable `professional` takes the value of 1. Calculate the change in $\hat{y}$ associated with a $1,000 increase in income based on your answer for (c).

```
regression1
21.142259+0.003171*(0)+37.781280*(1)-0.002326
21.142259+0.003171*(1000)+37.781280*(1)-0.002326
(21.142259+0.003171*(1000)+37.781280*(1)-0.002326)-(21.142259+0.003171*
    (0)+37.781280*(1)-0.002326)
```

The change in $\hat{y}$ is 3.171 units per $1,000 increase in income.

(g) What is the effect of changing one's occupations from non-professional to professional when her income is \$6,000? We are interested in the marginal effect of professional jobs when the variable `income` takes the value of $6,000$. Calculate the change in $\hat{y}$ based on your answer for (c).

```
regression1
21.142259+0.003171*(6000)+37.781280*(1) -0.002326
21.142259+0.003171*(6000)+37.781280*(0) -0.002326*(0)
(21.142259+0.003171*(6000)+37.781280*(1) -0.002326) -(21.142259+0.003171*
    (6000)+37.781280*(0) -0.002326*(0))
```

The change in $\hat{y}$ is 37.77895 units by having a professional job.

Question 2 (50 points): Political Science

Researchers are interested in learning the effect of all of those yard signs on voting preferences.[1] Working with a campaign in Fairfax County, Virginia, 131 precincts were randomly divided into a treatment and control group. In 30 precincts, signs were posted around the precinct that read, "For Sale: Terry McAuliffe. Don't Sellout Virgina on November 5."

Below is the result of a regression with two variables and a constant. The dependent variable is the proportion of the vote that went to McAuliff's opponent Ken Cuccinelli. The first variable indicates whether a precinct was randomly assigned to have the sign against McAuliffe posted. The second variable indicates a precinct that was adjacent to a precinct in the treatment group (since people in those precincts might be exposed to the signs).

Impact of lawn signs on vote share

Precinct assigned lawn signs (n=30)	0.042
	(0.016)
Precinct adjacent to lawn signs (n=76)	0.042
	(0.013)
Constant	0.302
	(0.011)

Notes: R^2=0.094, N=131

(a) Use the results to determine whether having these yard signs in a precinct affects vote share (e.g., conduct a hypothesis test with $\alpha = .05$).

```
# H0: B is equal to 0, HA: B is not equal to 0
na <- 30
(0.042-0)/0.016
# Test statistic is 2.625
2*pt(-abs(2.625), df=na-1)
# P-value is 0.01368397
```

After conducting a hypothesis test, I calculated a p-value of 0.01368397 which is less than $\alpha = .05$. We reject our null hypothesis that having these yard sings in a precinct has no effect on vote share.

[1] Donald P. Green, Jonathan S. Krasno, Alexander Coppock, Benjamin D. Farrer, Brandon Lenoir, Joshua N. Zingher. 2016. "The effects of lawn signs on vote outcomes: Results from four randomized field experiments." Electoral Studies 41: 143-150.

(b) Use the results to determine whether being next to precincts with these yard signs affects vote share (e.g., conduct a hypothesis test with $\alpha = .05$).

```
# H0: B is equal to 0, HA: B is not equal to 0
nb <- 76
(0.042-0)/0.013
# Test statistic is 3.230769
2*pt(-abs(3.230769), df=nb-1)
# P-value is 0.001834303
```

After conducting a hypothesis test, I calculated a p-value of 0.001834303 which is less than $\alpha = .05$. We reject our null hypothesis that being next to precincts with these yard signs has no effect on vote share.

(c) Interpret the coefficient for the constant term substantively.

A precinct that was not assigned lawn signs nor was adjacent to another precinct with lawn signs is expected to have 30.2% of votes go to Ken Cuccinelli, on average.

(d) Evaluate the model fit for this regression. What does this tell us about the importance of yard signs versus other factors that are not modeled?

According to the model, only 9.4% of the variance in votes is explained by the selected variables. This is a low number and there must be other variables to better explain the variance.