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`.)

```
#Check the levels of the variable
levels(Prestige$type)
```

```
[1] "bc"   "prof" "wc"
```

```
#Store the edited varaible into a newe variable called "Prof"
Prof <- ifelse(Prestige$type=="prof",1,0)
str(Prof)
```

```
num [1:102] 1 1 1 1 1 1 1 1 1 1 ...
```

(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 $\times$ dummy interaction.)

```
lm(prestige~income+Prof+income:Prof, data=Prestige)
```

```
Coefficients:
(Intercept)       income         Prof  income:Prof
21.142259     0.003171    37.781280    -0.002326
```

(c) Write the prediction equation based on the result.
The simple linear prediction equation is $\hat{Y}_i = 21.14 + 0.003 * X_1(income) + 37.78 * X_2(Prof) - 0.002 * D(income : Prof)$.

(d) Interpret the coefficient for `income`.

The coefficient is a positive number, indicating that there is a positive relationship between income and prestige.
The coefficient of income is 0.003, which means, controlling other variables, a unit increase in average income will result in 0.3% increase in the score of prestige.

(e) Interpret the coefficient for `professional`.

The coefficient is a positive number, indicating that there is a positive relationship between professional and prestige.
The coefficient of professional is 37.78, which means, controlling income, a unit increase in type of occupation (i.e, from blue and white collar to professionals) will result in 3778% increase in the score of prestige.

(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).

Since we are interestd in the effect of income on profeessional occupations, $D(income : Prof)$ should equal to 1.
So we get $Yi = 21.14 + 0.003 * X1(income) + 37.78 * 1 - 0.002 * 1 * income = 58.92 + 0.001 * income$, where $0.001 * income$ is the marginal effect.
When income equals to 1000, the marginal effect would be $0.001 * 1000 = 1$. Therefore, the effect of a \$1000 increase in incomee will lead to 1 prestige score increase for professional occupations.

(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).

Recall the prediction equation: $Yi = 21.14 + 0.003 * X1(income) + 37.78 * X2(Prof) - 0.002 * D(income : Prof)$.
When income equals to \$6000, the prediction equation would be $Yi = 21.14 + 0.003 * 6000 + 37.78 * X2(Prof) - 0.002 * 6000 * Prof = 39.14 + 25.78 * Prof$.
When one's occupation is non-professional, $prof = 0$, $Yi = 39.14$.
When onee's occupation is professional, $prof = 1$, $Yi = 64.92$.
The difference between the two equations is 25.78, which means th effect of changing one's occupations from non-professional to professional, given her income is \$6000, is 25.78 unit increase in the prestige score.

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 $\alpha = .05$).

```
#Hypothesis: Ho:Beta2 = 0 vs. Ha:Beta2 is not 0.
#Test statistic:
t=(0.042-0)/0.016=2.625
#p-value:
#df= n-3 parameters = 131-3 = 128
2*pt(2.625, 128, lower.tail = F)
#p-value = 0.0097
```

p-value is 0.0097, which is smaller than the confidence level 0.05.
We have enough evidence to reject the null hypothesis that there is no relationship between the lawn signs and the 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.

In other words, we can say that having these yard signs does affect the vote share.

(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 $\alpha = .05$).

```
#Hypothesis: Ho:Beta3 = 0 vs. Ha:Beta3 is not 0.
#Test statistic:
t=(0.042-0)/0.013=3.230
#p-value:
#df= n-3 = 131-3 = 128
2*pt(3.23, 128, lower.tail = F)
#p-value = 0.0016
```

p-value is 0.0016, which is smaller than the confidence level 0.05.
We have enough evidence to reject the null hypothesis that there is no relationship between living next to lawn signs and the vote share.
In other words, we can say that adjacent to these yard signs does affect the vote share.

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

The coefficient of the constant equals to 0.302, which means when there is no lawn sign either in or near the precinct, the average proportion of the vote that went to Ken Cuccinelli is 30.2%.

(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?

The model does not work good enough to find out the influential factors that affect the vote share.
As we can learn from the value of R square, which is 0.094 in this case, only 9.4% of variation is explained by this model. 90.6% of variation cannot be explained by having lawn signs or not. So there should be other explanotary variables that have significant impact on the vote hare but is not included in this model. And these omitted variables might have greater explanotary power than yard signs.