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`.)

```
#Create new variable
Prestige$professional <- ifelse(Prestige$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.)

```
#run linear regression
Reg1 <- lm(prestige ~ income + professional + income:professional, data =
    Prestige)
summary(Reg1)
```

R output:

```
Call:
lm(formula = prestige ~ income + professional + income:professional,
data = Prestige)

Residuals:
Min      1Q  Median      3Q     Max
-14.852  -5.332  -1.272   4.658  29.932

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)         21.1422589  2.8044261   7.539 2.93e-11 ***
income               0.0031709  0.0004993   6.351 7.55e-09 ***
professional        37.7812800  4.2482744   8.893 4.14e-14 ***
income:professional -0.0023257  0.0005675  -4.098 8.83e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.012 on 94 degrees of freedom
(4 observations deleted due to missingness)
Multiple R-squared:  0.7872,Adjusted R-squared:  0.7804
F-statistic: 115.9 on 3 and 94 DF,  p-value: < 2.2e-16
```

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

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 X_i D_i + \epsilon_i$$

Where $\epsilon \sim N(0, \bar{\sigma}^2)$

We get from results in Question 2:

$$Y = 21.142 + 0.003X_1 + 37.781D_1 - 0.002X_1D_1 + \epsilon_1$$

Where $\epsilon \sim N(0, 8.012^2)$

(d) Interpret the coefficient for `income`.

Given that all other variables remain constant, for each unit increase in income, there would be 0.003 unit increase in the prestige for jobs.

(e) Interpret the coefficient for `professional`.

For professionals, the whole regression line would shift up for 37.781 unit in general, indicating that regardless of income, prestiges for professional job are 37.781 higher.

For white and blue collar workers, there would be no effect.

(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).

From $Y_1 - Y_2 = 21.142 + 0.003(X_1 + 1000) + 37.781 - 0.002(X_1 + 1000) + \epsilon_1$
$- 21.142 - 0.003X_1 - 37.781 + 0.002X_1 - \epsilon_1$

We get the marginal effect to be
$0.003(X_1 + 1000 - X_1) - 0.002(X_1 + 1000 - X_1) = 1$

```
#Calculate marginal effect of a $1000 increase in income
effect1 <- 1000*0.003 -1000*0.002
effect1# 1
```

(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).

From $Y_1 - Y_2 = 21.142 + 0.003 * 6000 + 37.781 - 0.002 * 6000 + \epsilon_1$
$- 21.142 - 0.003 * 6000 - \epsilon_1$

We get the marginal effect to be 25.781

```
##Calculate marginal effect for change in the dummy variable
effect2 <- 37.781 - 0.002*6000
effect2#25.781
```

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 $\alpha = .05$).

```
#Testing beta1 = 0 or not
TS1 <- 0.042/0.016
n <- 131
k <- 3
P_value1 <- 2*(pt(abs(TS1), n-k, lower.tail = F))
P_value1# 0.00972002 < 0.05, affected
```

(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 $\alpha = .05$).

[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.

```r
#Testing beta2 = 0 or not
TS2 <- 0.042/0.013
P_value2 <- 2*(pt(abs(TS2), n-k, lower.tail = F))
P_value2# 0.00156946 < 0.05, affected
```

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

When there is no lawn signs in any precinct, the number of precinct assigned lawn signs and number of precinct adjacent to lawn signs are zero, so the proportion of the vote that went to Ken Cuccinelli would be 0.302.

(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?

```r
#F-test for the overal model fit
F.test <- ((0.094/k)/((1-0.094)/(n-k-1)))
df1 <- k
df2 <- n-k-1
F.pvalue <- df(F.test, df1, df2)
F.pvalue #0.007141957 < 0.05
#We can reject the null hypothesis and conclude that there's at least one
    coinficient is significant.
```

This tells us that yard signs does affect vote shares, however we do not know about other factors that are not modeled; since our R-squared is pretty low, we can't assume that this is the only factor leading to the outcome of the votes.