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`.)

```
# Recode professionals as 1, and blue and white as 0
Prestige$professional<- ifelse(Prestige$type=="prof", 1, 0)
Prestige$professional
```

(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.)

```
# Use the lm function to run a linear model with prestige as the outcome
    variable (y), and income, professional, and their interactions as the
    explanatory variables (x)
Prestige_regress <- lm(prestige ~ income + professional+ income:
    professional, data = Prestige)
Prestige_regress
summary(Prestige_regress)

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 ***
---

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 =  beta0 + beta1 *x1 +beta2 *x2 + beta3*x1*x2
# beta0 = intercept
# beta1 = slope of the relationship between income and prestige
# x1 = income
# beta2 = slope of the relationship between professional and prestige
# x2 = professional
# beta3 = interaction between income and professional
# SOLUTION: Y = 21.1422589 + 0.0031709*x1 + 37.7812800*x2 + -0.0023257*x1
    *x2
```

(d) Interpret the coefficient for `income`.

```
# If all other variables are held constant, with every $1 increase in
    income, there is an average increase in the prestige score of
    0.0031709.
```

(e) Interpret the coefficient for `professional`.

```
# If all other variables are held constant, being a professional rather
    than a white collar or blue collar worker increases the presteige
    score by 37.7812800 on average.
```

(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).

```
# Calculate the prestige score when x1 = 0 and x2 = 1
(21.1422589) + (0.0031709*0) + (37.7812800*1) + (-0.0023257*0*1)=
    58.92354
# Calculate the prestige score when x1 = 1000 and x2 = 1
(21.1422589) + (0.0031709*1000) + (37.7812800*1) + (-0.0023257*1000*1) =
    59.76874
# Calculate the difference between the outcomes
59.76874-58.92354 = 0.8452
# SOLUTION: When professional is held constant at 1, a $1000 increase in
    income increases the prestige score by 0.8452 on average.
```

(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).

```
# Calculate the prestige score when x1=6000 and x2=0
(21.1422589) + (0.0031709*6000) + (37.7812800*0) + (-0.0023257*6000*0) =
    40.16766
# Calculate the prestige score when x1=6000 and x2=1
(21.1422589) + (0.0031709*6000) + (37.7812800*1) + (-0.0023257*6000*1) =
    63.99474
# Calculate the difference between the outcomes
63.99474- 40.16766 = 23.82708
# SOLUTION: When income is held constant at $6000, being a professional
    rather than a white collar or blue collar worker increases the
    presteige score by 23.82708 on average.
```

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 $\alpha = .05$).

```
# Ho: There is no relationship between having yard signs and vote share (
    beta = 0)
# Ha: There is a relationship between having yard signs and vote share (
    beta does not equal 0)
# Calculate the test statistic using the equation: (beta-0)/standard
    error
test_statistic<- (0.042-0)/0.016
# test_statistic = 2.625
# Calculate the p value using the test_statistic = 2.625 and degrees of
    freedom = n-1 (n=30)
p_value<- 2*pt(2.625,df=29, lower.tail = F)
p_value = 0.01368397
```

[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.

```
# Given a significance level of 0.05, reject the null hypothesis that there is no relationship between having yard signs and vote share (0.01368397<0.05).
```

(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 $\alpha = .05$).

```
# Ho: There is no relationship between being adjacent to yard signs and vote share (beta = 0)
# Ha: There is a relationship between being adjacent to yard signs and vote share (beta = 0)
# Calculate the test statistic using the equation: (beta-0)/standard error
test_statistic_b <- (0.042-0)/0.013
test_statistic_b = 3.230769
# Calculate the p value using the test_statistic = 3.230769 and degrees of freedom = n-1 (n=76)
p_value_b<- 2*pt(3.230769,df=75,lower.tail = F)
p_value_b = 0.001834303
# Given a significance level of 0.05, reject the null hypothesis that there is no relationship between being adjacent to yard signs and vote share (0.001834303<0.05).
```

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

```
#c) Given that there are no yard signs in the precint or adjacent to the precint, the proportion of the vote that went to Cuccinelli is 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?

```
# The R squared value is 0.094. This means that the model explains 9.4% of the variation in vote share. This suggests that other variables might have a stronger prediction of the variability  within vote share than the presence of signs at/adjacent to the presinct.
```