Problem Set 4

Applied Stats/Quant Methods 1

Due: November 26, 2021

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 GitHub in `.pdf` form.
- This problem set is due before class on Friday November 26, 2021. No late assignments will be accepted.
- Total available points for this homework is 80.

Question 1: 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`.)

```
Professional <- ifelse(Prestige$type == 'prof', 1, 0)
Blue_Collar <- ifelse(Prestige$type == 'bc', 0, 0) #to be sure I created both WC and BC as 0 values
White_Collar <- ifelse(Prestige$type == 'wc', 0, 0) #as it did not cause problems further on as WC/BC aren't called further on

Prestige_1 <- data.frame(education = Prestige$education,
                         income = Prestige$income,
                         women = Prestige$women,
                         prestige = Prestige$prestige,
                         census = Prestige$census,
                         Professional =Professional,
                         Blue_Collar = Blue_Collar,
                         White_Collar = White_Collar)

#I am unsure what to do with the NAs as I assume they will not alter the findings and won't be counted.
```

(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.)

```
#Q1 Part B
lm_mod1 = lm(prestige ~ income + Professional, data = Prestige_1) #used the + function, per the ones I saw in class

summary(lm_mod1)
```

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

```
#Prediction Equation =

#Prestige = 3.062 + 1.371*(Income) + 2.276*(Professional)  #this is for
    the + model lm_mod1
```

I omitted an error term as this could not be quantified for the equation.

(d) Interpret the coefficient for `income`.

The output, b1=1.371 implies that prestiege will be expected to increase by 1.371 units for an every additional unit of income. In addition, the hypothesis that the prestige rating is linearly related to the income level with other predictors being constant.

Ho : b1 is equal to 0 (no linear relationship)

Ha : b1 is not equal to 0 (significant linear relationship)

So, for the test statistic t = 5.348 and p-value for the test statistic(t=5.348) is less than 6.12*10(-7). Which means that the probability of getting test statistic 5.348 by chance under the assumption of b1 =0 is extremely rare. So we reject the null hypothesis b1=0 and it shows the evidence of a positive linear relationship between Income and prestige rating level.

(e) Interpret the coefficient for `professional`.

The output, b1=2.267 implies that prestiege will be expected to increase by 2.267 units if professional. In addition, the hypothesis that the prestige rating is linearly related to professional with other predictors being constant.

Ho : b1 is equal to 0 (no linear relationship)

Ha : b1 is not equal to 0 (significant linear relationship)

So, for the test statistic t = 9.817 and p-value for the test statistic(t=9.817) is less than 4.07*10(-16). Which means that the probability of getting test statistic 9.817 by chance under the assumption of b1 =0 is extremely rare. So we reject the null

hypothesis b1=0 and it shows the evidence of a positive linear relationship between profession and prestige rating level.

(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).

```
#Prestige = 3.062 + 1.371*(Income) + 2.276*(Professional)
#Prestige = 3.062 + 1.371*(1,000) + 2.276*(0)
#Prestige = 3.062 + 1,371 + 0
#Prestige = 1,374.62
```

A 1,000 increase in salary corresponds to .374 unit increase in prestige.

(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).

```
#Prestige = 3.062 + 1.371*(Income) + 2.276*(Professional)
#Prestige = 3.062 + 1.371*(0) + 2.276*(6000)
#Prestige = 3.062 + 0 + 8226
#Prestige = 8229.062
```

there is a 2,226 positive difference in the salary of professionals when all other factors are controlled for.

Question 2: 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 from a linear regression to determine whether having these yard signs in a precinct affects vote share (e.g., conduct a hypothesis test with $\alpha = .05$).

```
qt(p=.05/2, df=29, lower.tail=FALSE) # 2 sided as it can be positive or
    negative

2*pt(q=1.699127, df=29, lower.tail =FALSE)

#H0:B1 = 0
#vs
#Ha:B1 =(not equal) 0

#As the P Vlaue = 0.01 and is less than or equal to our Alpha it
    indicates strong evidence against the null hypothesis, so we reject it
    .
```

[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 $\alpha = .05$).

```
qt(p=.05/2, df=75, lower.tail=FALSE) # 2 sided as it can be positive or
    negative

2*pt(q=1.9921, df=75, lower.tail =FALSE)

#H0:B2 = 0
#vs
#Ha:B2 =(not equal) 0

#As our P value is 0.05000024 and is slightly above our alpha of 0.05 it
    indicates weak evidence against the null hypothesis, so we fail to
    reject it.
```

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

```
qt(p=.05/2, df=106, lower.tail=FALSE)
```

```
2*pt(q=1.982597, df=106, lower.tail =FALSE)
```

The constant, b1=0.302 implies that voteshare will be expected to increase by 0.302 units for an every additional unit of b1. In addition, the hypothesis that voteshare is linearly related to b1 with other predictors being constant.The constant is more significant a factor than eith of the two other varaibles regarding signs.

(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-Squared in this model states that 9.4 percent of the variability in vote share is explained by lawn signs. The R-squared of the regression is the fraction of the variation in your dependent variable that is accounted for your independent variables. This level of explanation is relativly low and leaves approximalty 90 percent of the factors that drive vote share as unexplained.