Problem Set 4

Applied Stats/Quant Methods 1

Due: December 3, 2023

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.
- This problem set is due before 23:59 on Sunday December 3, 2023. No late assignments will be accepted.

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`).

```
# a) New variable "professional" based on "type"
# I am using mutate() from dplyr, along with ifelse()
# What I am telling R is that the new variable professional
# will take the value 1 if type is equal to "prof" and 0
# if type is anything else.

Prestige <- Prestige %>% mutate(
  professional = ifelse(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.)

```
# b) Running the first model:
# prestige ~ income + professional + interaction

m1 <- lm(prestige~income+professional+income*professional, Prestige)
summary(m1)
```

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

$$\text{Prestige} = 21.14 + 0.0032 \cdot \text{Income} + 37.78 \cdot \text{Professional} - 0.0023 \cdot \text{Income} \cdot \text{Professional}$$

where *Prestige*: Prestige score for occupation, *Income*: Average income of incumbents, and *Professional*: 1 indicates professionals, 0 indices white or blue collar.

And it is also useful to have the prediction equations particular to both values of Professional:

$$\text{PrestigeNonProf} = 21.14 + 0.0032 \cdot \text{Income}$$

$$\text{PrestigeProf} = 58.92 + 0.0009 \cdot \text{Income}$$

These equations simply come from substituting the values 0 and then 1 for Professional, and then in the second equation factoring Income out. + 0.0009 is simply the result of subtracting the 0.0023 from the 0.0032 after factoring Income out.

(d) Interpret the coefficient for `income`.

Since we are including an interaction term in our model, the coefficient for income, on its own, speaks to the association between income and prestige, for white and blue collar workers (professional = 0). So, for white and blue collar workers, a one-unit increase in income is associated, on average, with an increase of 0.0032 in prestige.

If we wanted to know about the association between income and prestige for those who are professionals, we would add up the income coefficient and the coefficient of the interaction. This gives 0.0009. So, for professionals, a one-unit increase in income is associated, on average, with an increase of 0.0009 in prestige. The effect of income on prestige is smaller for professionals than for white and blue collar workers.

(e) Interpret the coefficient for `professional`.

Since we are including an interaction term in our model, the coefficient for professional, on its own, speaks to the association between income and professional, when income = 0. A one-unit increase in professional (or, in other words, going from being white/blue collar to being professional) is associated, on average, with an increase of 37.78 in prestige, when income is 0.

(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).

As it was explained above, for professionals, a one-unit increase in income is associated, on average, with an increase of 0.0009 in prestige. So, a 1000 dollar increase is associated, on average, with an increase of 0.9 in prestige. Considering the range for the prestige variable, this change is quite small. We had already seen that the effect of income on prestige was smaller for those who are professionals.

Code for the calculations:

```
# Increasing income by 1000 for prof = 1:

# One way:
pres1.2 = 58.92 + 0.0009*0
pres2.2 = 58.92 + 0.0009*1000
change_c = pres2.2 - pres1.2
change_c

# Another way:
change_d = 0.0009*1000
change_d
```

(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).

Here, we can substitute 6000 in both of our prediction equations, and then compare the associated value of Prestige that we expect for both, or we can use the coefficients of Professional and of the interaction term from our second prediction equation, to calculate the change. At an income of \$6000, the effect of changing one's occupations from non-professional to professional is associated with an increase in prestige of 23.98 units.

Code for the calculations:

```
# Going from prof = 0 to prof = 1 for income = 6000

# One way:
pres1.1 = 21.14 + 0.0032 * 6000
pres2.1 = 58.92 + 0.0009 * 6000
change_a = pres2.1 - pres1.1
change_a

# Another way:
change_b = 37.78 - 0.0023 * 6000
change_b
```

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

[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.

(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 $\alpha = .05$).

Here, our null hypothesis would be that the coefficient for yard signs, which we may denote β_1, is equal to zero. We are given the point estimate for β_1, as well as the standard error. So, we can do the following:

```
# a) Hypothesis test for signs coefficient:

# The general form for our test statistic is:
# (observed value - expected if null is true) / SE

# We will also need our sample size and degrees of
# freedom handy:

n = 131
df = n - 2

# Substituting with the given results:

t1 = (0.042 - 0) / 0.016
t1

# And calculating the p-value using pt():

p1 = 2 * pt (abs(t1), df, lower.tail=F)
p1
```

And we see that the associated p-value is 0.0098, which is smaller than 0.05, so we can reject the null hypothesis that $\beta_1 = 0$ at the 95% confidence level. In other words, we can reject the null hypothesis that putting up signs is not associated to a larger vote share for Cuccinelli, or that the slope of the regression line between the two variables is zero. Moreover, we may interpret the point estimate of 0.042 as follows: the vote share for Cuccinelli would be, on average, 0.042 points higher in precincts with a sign, as compared to those without a sign.

(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 $\alpha = .05$).

For this, we need to repeat the same process as above, but this time for the coefficient of the adjacent variable, which we can denote as β_2. Our null hypothesis in this case is that β_2 is equal to zero.

```r
# b) Hypothesis test for adjacent coefficient:

# Substituting with the given results:

t2 = (0.042 - 0) / 0.013
t2

# And calculating the p-value using pt():

p2 = 2 * pt (abs(t2), df, lower.tail=F)
p2
```

And we see that the associated p-value is 0.0016, which is smaller than 0.05, so we can reject the null hypothesis that $\beta_2 = 0$ at the 95% confidence level. In other words, we can reject the null hypothesis that putting up signs, even in adjacent precincts, is not associated to a larger vote share for Cuccinelli in the precinct adjacent to where the sign was placed, or that the slope of the regression line between the two variables (being adjacent to a precinct with a sign, and vote share) is zero. Moreover, we may interpret the point estimate of 0.042 as follows: the vote share for Cuccinelli would be, on average, 0.042 points higher in precincts that are adjacent to a precinct where a sign was put, as compared to those that are not.

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

The constant is 0.302. This means that when both our 'sign' and 'adjacent' variables are zero, Cuccinelli gets, on average, a vote share of 0.302. In other words, in precincts that neither got a sign not are adjacent to a precinct that got a sign, Cuccinelli is expected to get a vote share of 0.302 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?

There are several ways to evaluate the model fit. Given the information we have, we can first look at the R squared. In this case, the reported R squared is 0.094. This means that 9.4% of the variation in our outcome ()Cuccinelli's vote share) is explained or accounted for by the variables that we included in our model (yard signs and being adjacent to a precinct where a yard sign was placed). The R squared value may seem low, but given the range of vote share variables, the importance of these signs in determining the vote share is quite large: 9.4 percentage points can make the difference between winning or losing an election.

We can also carry out an overall f-test to see if the model is at all useful at explaining variation in the vote share:

```
# d) Overall f test

# First we need our R squared value handy:

r2 = 0.094

# I will also now define k:

k = 2

# We already have our sample size and degrees of freedom
# And we can use the test statistic from slide 17, week 10,
# as well as the code in slide 19, week 10:

f = ((r2/(k-1))) / ((1-r2)/(df))
f

df1 = k - 1
df2 = n - k - 1

pf = df(f, df1, df2)
pf
```

The resulting F statistic's value is 13.38, and the corresponding p-value is around 0.0002. Thus, we can reject the null hypothesis that all of the predictor coefficients are zero, meaning at least one of the coefficients is different to zero. In other words, the results of this test indicate that at least one of the predictors (getting a sign or being adjacent to a precinct that got a sign) is useful at explaining variation in vote share.