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`.)**

To create the new variable `professional`, the ifelse function was used. This was saying that if the type of job was called ”prof” it would be coded as 1 and otherwise (else) it would be coded as 0 (for wc and bc)

The following code was used:

```
Prestige$professional <- ifelse(Prestige$type == "prof", 1, 0)
```

This created a new variable in the Prestige data set that could then be used in the Regression.

(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.)**

To run the linear model which included the interaction the following code was used:

```
interaction_reg <- lm(data = Prestige, prestige ~ income + professional +
    income:professional)
summary(interaction_reg)
```

When the summary function was used the following was 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 0.0000000000292686 ***
income               0.0031709  0.0004993   6.351 0.0000000075482422 ***
professional        37.7812800  4.2482744   8.893 0.0000000000000414 ***
income:professional -0.0023257  0.0005675  -4.098 0.0000882872162594 ***
---
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: < 0.00000000000000022
```

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

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

Based on the above results we can produce the following prediction equation:

```
y = 21.142259 + 0.003171income + 37.781280professional  -0.002326(income*
    professional)
```

This means that for Professional workers the equations becomes:

Inserting 1 for professional

```
y = 21.142259 + 0.003171income + 37.781280  -0.002326income
```

And for non-professional workers (white collar or blue collar) the equations becomes:

Inserting 0 for professional

```
y = 21.142259 + 0.003171income
```

(d) **Interpret the coefficient for `income`.**

In multivariate regression the coefficient explains how much the outcome variable is expected to increase when the explanatory variable increases by one, while holding all other variables constant.

In this case out explanatory variable is income and our outcome variable is prestige. So while holding all other variables constant, prestige store increases by 0.003171 on average, for every one dollar increase in income.

(e) **Interpret the coefficient for `professional`.**

The coefficient for professional is 37.781280. This means that for a professional this is 37.781280 x 1, but for a non professional job (wc or bc) this is multiplied by 0 and therefore not included in the equation. So this means that people with professional jobs, on average have an increase of 37.781280 in prestige score than those in white collar or blue collar jobs, while holding all other variables constant.

(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).**

I input $0 and $1,000 into the income variable. This is allows exploration of a 1000 dollar increase at any pay scale, as this is a linear relationship.

I input "1" for professional as we are interested in professional jobs, (not wc or bc which I would input as "0".)

When income is 0:

```
prestige0 <-  21.142259 + 0.003171*0 + 37.781280*1 + (-0.002326*0*1)
prestige0 = 58.92354
```

When income is 1000:

```
prestige1000 <-  21.142259 + 0.003171*1000 + 37.781280*1 + (-0.002326*1000*1)
prestige1000 = 59.76854
```

To find the marginal effect I looked at the difference between this 1000 unit increase in income:

```
increase <- prestige1000 - prestige0
increase = 0.845
```

On average, prestige score increases by 0.845 in professional jobs, for a $1,000 increase in income.

(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).**

To understand what the effect of changing one's occupation for non-professional to professional we can input the income into the prediction equation, while changing the binary profession variable.

For non-professional:

```
nonprofes6000 = 21.142259 + 0.003171*6000 + 37.781280*0 + (-0.002326*6000*0)
nonprofes6000 = 40.16826
```

For professional:

```
profes6000 = 21.142259 + 0.003171*6000 + 37.781280*1 + (-0.002326*6000*1)
profes6000 = 63.99354
```

Difference = 63.993539 - 40.168259 = 23.8252

The prestige score difference between a $6,000 income in a professional job is 23.82528 more prestige points than a non-professional job.

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 $\alpha = .05$).**

To conduct a hypothesis test to see if yard sign usage affects voter preference we first need to determine the null and alternative hypothesis.

The null hypothesis is that slope of the assigned == 0.

The alternative hypothesis is that slope of the assigned != 0

```
Ho == 0
Ha != 0
```

From here I calculated the t statistic using the following formula:

t statistic = coeff assigned / standard error of coeff assigned

[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.

```
t stat = 0.042 / 0.016
t stat = 2.625
```

Now we can find the p value using this t statistic. Firstly we find the degrees of freedom. In a multiple regression the df equals N-k-1, (where k is the number of variables so in this case the df is n-2-1, so 131-3 = 128.

P-value = 2 x pr(t128 > 2.625). Using a two tailed test as we are interested in negative or positive answers.

```
2*pt(2.625, df = 128, lower.tail =FALSE)
```

This gives a p value of 0.00972002.

We can conclude that this is less than the $\alpha = .05$, so we reject the null hypothesis that the slope is 0.

(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 $\alpha = .05$).**

To conduct a hypothesis test to see if yard sign usage affects voter preference we first need to determine the null and alternative hypothesis.

The null hypothesis is that slope of adjacent == 0

The alternative hypothesis is that slope of adjacent != 0

```
Ho == 0
Ha != 0
```

From here we can then find the t statistic using the following formula:

t stat = coeff adjacent / standard error of coeff adjacent

```
t stat = 0.042/ 0.013
t stat = 3.23076923
```

Now we can find the p value using this t statistic. Firstly we find the degrees of freedom. In a multiple regression the df equals N-k-1, (where k is the number of variables so in this case the df is n-2-1, so 131-3 = 128

P-value = 2 x pr(t128 > 3.2308). Using a two tailed test as this hypothesis test is interested in negative or positive answers.

Using the following code we can find the p value:

```
2*pt(3.23076923, df = 128, lower.tail =FALSE)
```

This gives a p value of 0.00156946.

We can conclude that this is less than the $\alpha = .05$, so we reject the null hypothesis that the slope is not different to 0.

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

The coefficient for the constant term is 0.302. This is where the linear model crosses the y axis, so the value of the outcome variable ”vote share” is 0.0302 when the explanatory variables are 0.

To apply this, the regression is showing that when there are no adjacent nor assigned signs in a precinct that are negatively attacked Cuccinellis opponent McAuliffe, Cuccinelli would gain 30.2% of the vote share.

In this particular study, they did not have a control whereby there were no signs in a precinct, or adjacent, so it is difficult to to say if this is an accurate figure. However mathematically shows what the vote share is when the explanatory variables are 0.

(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^2 in a regression model shows us how much variance is being accounted for. In this instance, R^2 is 0.094, portraying that the model is accounting for 9.4% of variance.

The R^2 being 0.094 tells us that there is 1 - 0.094 = .906 of the variance not accounted for.

This suggests that yard sign usage is not accounting for that much of the variance in vote share. However, seeing at we are looking at voteshare, and other factors such a political opinion, the policies of Cuccinelli and McAuiffe and previous voting history will probably play a large role in peoples voting decision, the fact that 9.4% of the variance can be explained with whether negative signs against the opponent are in a precinct or adjacent, this is actually quite a good finding.