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`.)

```
Prestige$type.Dummy<-ifelse(Prestige$type=="prof",1, ifelse(Prestige$type == "b

#type.Dummy is to be renamed as prof


DONE
```

(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 $\times$ dummy interaction.)

```
regression_model_1 <- lm(prestige ~ income*prof, data=Prestige)
summary(regression_model_1)
Call:
lm(formula = prestige ~ income *prof, data = Prestige)


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



Coefficients:
Estimate Std. Error t value

(Intercept)       21.1422589  2.8044261   7.539
income             0.0031709  0.0004993   6.351
prof        37.7812800  4.2482744   8.893
income:prof -0.0023257  0.0005675  -4.098
Pr(>|t|)
(Intercept)       2.93e-11 ***
income            7.55e-09 ***
prof        4.14e-14 ***
income:prof 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


summary(Prestige)
education          income          women
Min.   : 6.380   Min.   :  611   Min.   : 0.000
1st Qu.: 8.445   1st Qu.: 4106   1st Qu.: 3.592
Median :10.540   Median : 5930   Median :13.600
Mean   :10.738   Mean   : 6798   Mean   :28.979
3rd Qu.:12.648   3rd Qu.: 8187   3rd Qu.:52.203
Max.   :15.970   Max.   :25879   Max.   :97.510
prestige         census       type
Min.   :14.80   Min.   :1113   bc  :44
1st Qu.:35.23   1st Qu.:3120   prof:31
Median :43.60   Median :5135   wc  :23
Mean   :46.83   Mean   :5402   NA's: 4
3rd Qu.:59.27   3rd Qu.:8312
Max.   :87.20   Max.   :9517

confint(regression_model_1)
2.5 %
(Intercept)       15.574005075
income             0.002179562
prof         29.346231606
income:prof  -0.003452474
97.5 %
(Intercept)       26.710512633
income             0.004162257
prof        46.216328304
income:prof  -0.001198945


DONE
```

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

$\hat{Y}_i = \beta_0 + \beta_1 x_i + \beta_2 x_i D_i$

The predicted value of outcome variable = regression coefficient + regressioncoefficient multiplied by independant variable + regressioncoefficient multiplied by interaction

between two independant variables

The predicted value of Prestige = 21.142 + 0.003*income variable + 37.781*prof + -0.002*income:prof

Wondering what numeric values to slot into the prediction equation in the place of the variables?

(d) Interpret the coefficient for `income`. Regression coefficients represent the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant.

```
Predicted value of Prestige is not influenced by income.
(as the coefficient value is 0.0) or very little depending on the decimal points yo
```

(e) Interpret the coefficient for `professional`. Regression coefficients represent the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant.

Professional or prof is the dummy variable. Predicted value of prestige increases in mean change 4.3 for every unit of professional while holding other predictors in the model 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).

A $1000 increase in income on prestige score for professional occupations$

The predicted value of Prestige = 21.142 + 0.003*1000 + 37.781*1 + -0.002*1000*1

Predicted value = 59.923

Adding an interaction term to a model drastically changes the interpretation of all the coefficients.

(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).

```
The predicted value of Prestige = 21.142 +  0.003*6000   +   37.781*0
+ -0.002*6000*0
Predicted value = 76.923  -37.781 =

The predicted value of Prestige = 21.142 +  0.003*6000   +   37.781*1
+ -0.002*6000*1

No idea
```

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 $\alpha = .05$).

Conducting a hypothesis test: There are five steps Step 1: Make assumptions about the data and where it came from Step 2: Proof by contradiction, set up a null hypothesis Step 3: Calculate a test statistic, usually a Z- or t- statistic Step 4: Calculate a P-Value or a measure of surprise i.e. how likely is it that we would observe a test-statistic this extremem or more? Step 5: Draw a conclusion, based on test statistic, Step 1: Type of data: appears to be continuous Population distribution:View histogram to observe whether the data is normally distributed. No dataset is supplied for this question so it is assumed that the data is normally distributed. Sample size: sample size is greater than 30 Sampling method: Randomised field experiments Step 2: The null hypothesis is that the p value will NOT be smaller or equal to the alpha value of 0.5. The null

[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.

hypothesis is that having yard signs in a precinct DOES NOT affect vote share. Step 3: Calculate a t-test or either

Both tail as direction is not implied in the question

```
#get tstat and pvalues

TS < (betas0)/SEs

p_values < 2*pt(abs(TS), nk, lower.tail = F)

#get tstat and pvalues
conduct individual t-test
TS < (betas0)/SEs

p_values < 2*pt(abs(TS), nk, lower.tail = F)
#get tstat and pvalues

TS < (betas0)/SEs
Step 4:
p_values < 2*pt(abs(0.042/ 0.016), 1313, lower.tail = F)

p_values= 0.0097200119
P_values= 0.010
p_values= 0.01

DONE
```

$\alpha = .05$) 0.05 is greater than 0.01 Step 5: Therefore we can reject the Null hypothesis as there is less than 5In this case, we reject the hypothesis that having yard signs in a precinct DOES NOT affect vote share.

DONE

(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 $\alpha = .05$).

Conducting a hypothesis test: Step 1: Type of data: appears to be continuous Population distribution:View histogram to observe whether the data is normally distributed. No dataset is supplied for this question so it is assumed that the data is normally distributed. Sample size: sample size is greater than 30 Sampling method: Randomised field experiments Step 2: The null hypothesis is that the p value will NOT be smaller or equal to the alpha value of 0.5. The null hypothesis is that being next to precincts with these yard signs DOES affect vote share. Step 3: Calculate a t-test or either

```
#get tstat and pvalues

TS < (betas0)/SEs

p_values < 2*pt(abs(0.042/0.013), 1313, lower.tail = F)
Step 4: p_values = 0.00156946
p_values = 0.002
0.05 is greater than 0.002
```

Step 5: Therefore we can reject the Null hypothesis as there is less than 5In this case, we reject the null hypothesis that being next to precincts with these yard signs DOES affect vote share.

DONE

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

Remember interpretting the constant term can be tricky and may be un interpretable. If a large negative value it may be outside the range of the model and hence uninterpretable. Could graph and add line for constant and see where it falls? The constant is a predicted outcome when all the other variables equals 0. The constant is the predicted outcome when the other variables have no affect on the predicted outcome? The constant term is not meaningless when interpretted as a part of the overall equation. Regression coefficients represent the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant.

The constant in this model has a larger value than the other two variables, meaning that the mean change in the voteshare for one unit of change in the constant is much larger than the mean change in voteshare for one unit of change in any other of the two other variables. This indicates that the other two variables do not influence voteshare greatly.

DONE

(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?

Evaluating the model fit means graphing doesn't it? no there are three quantities for goodness-of-fit: 1. Residual Standard Error(RSE) 2. R-squared and adjusted R-squared 3. F-statistic

```
2. R-squared = 0.094
Rounded to 0.09
This is not very close to 1,
so this model does not explain the variance very well.
This value tells us that the data is explaining 9% of the variance.
```

This indicates to us that yard signs is not a very important predictor in this model versus other factors that are not modeled.

DONE