Problem Set 4

Applied Stats/Quant Methods 1

Due: December 4, 2022

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 4, 2022. 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`).

I created this new variable "professional" with the following code:

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

I then added this variable as a column to the Prestige dataset, creating a new dataset entitled: "Prestige_plus_prof". The code for how I did this is below:

```
Prestige_plus_prof <- cbind(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.)

I ran this linear model with the below code:

```
interact_reg <- lm(prestige ~ income + professional + income:professional
    ,
                   data = Prestige_plus_prof)
```

The summary of this regression output is:

```
Call:lm(formula = prestige ~ income + professional + income:professional,
 data = Prestige_plus_prof)

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 ***
---
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
```

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

The prediction equation based on this result is:

```
(prestige) = (21.142) + (0.003)(income) +
         (37.781)(professional)  + (-0.002)(income*professional)
```

(d) Interpret the coefficient for `income`.

The coefficient of 0.003 for income means that for every one dollar increase in income, there is an average increase in prestige score of 0.003, holding the professional variable and the incteraction effect between professional and income variables constant.

(e) Interpret the coefficient for `professional`.

The coefficient of 37.781 for professional means that for every one unit increase in professional variable (moving from 0(non-professional) to 1(professional), there is on average an increase in prestige score of 37.781, holding the other predictor variables (income, and the interaction between income and professional) 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 calculated this change in y hat with the following code:

```
y_1f1 <- 21.142 + (0.003*0) + (37.781*1) - (0.002*0*1)
y_1f2 <- 21.142 + (0.003*1000) + (37.781*1) - (0.002*1000*1)
Answer_1f <- y_1f2 - y_1f1
```

The marginal effect if a 1000 dollar increase in income on prestige when the variable of of professional is 1 is 1. This means that when someone with a professional occupation's income increases by \$1000, their prestige score is predicted to be 1 point higher 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).

I calculated this effect with the following code:

```
y_1g1 <- 21.142 + (0.003*6000) + (37.781*0) - (0.002*6000*0)
y_1g2 <- 21.142 + (0.003*6000) + (37.781*1) - (0.002*6000*1)
Answer_1g <- y_1g2 - y_1g1
```

When income is \$6000, the effect of changing occupations from non-professional (0) to professional (1) is an on average increase in prestige score from 39.142 to 64.923. In other words, an on average increase in prestige score of 25.781 occurs when income remains at \$6000, but an occupation changes from non-professional to professional.

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 $\alpha = .05$).

I conducted a t-test because I want to test if there is a relationship between the outcome (voteshare) and one of the variables in this model (x1, or actually having signs in a precinct).

First, I identify my Hypotheses:

H0: (beta of x1) = 0; Ha: (beta of x1) =/= 0

Second, I calculate the Test Statistic (t_x1) which equals b of x1 / se of x1

I calculated t_x1 with the following code in R:

```
t_x1 <- 0.042/0.016
```

t_x1 = 2.625

Third, I calculated the P value (pval_x1) with the following code in R:

```
pval_x1 <- 2*pt(abs(t_x1), 131-2-1, lower.tail = F)
```

pval_x1 = 0.0097

Alpha = 0.05, and pval_x1 = 0.0097; meaning pval_x1 < alpha, so we can reject H0. We can reject the null hypothesis that voteshare is not affected by sign presence in a precinct. Instead, this indicates there is some degree of linear relationship between this explanatory variable and the outcome variable (voteshare).

(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 $\alpha = .05$).

Again, conduct t-test but on 2nd variable now (x2): First, Hypotheses. H0: beta x2 = 0 ; Ha: beta x2 =/= 0

Second, Calculate Test Stat (t_x2) = b of x2 / se of x2

I calculate the Test Statistic with the following code in R:

```
t_x2 <- 0.042/0.013
```

t_x2 = 3.231

Third, I calculate the p-value with the following code:

```
pval_x2 <- 2*pt(abs(t_x2), 131-2-1, lower.tail = F)
```

pval_x2 = 0.0016

Conclusion: pval is 0.0016 and alpha is 0.05, so pval_x2 < alpha, so again can reject H0. The effect is significant. We can reject the null hypothesis that being a precinct beside one with signs does not have impact on voteshare. Instead, it does seem that being a precinct adjacent to one with signs has some degree of a linear relationship with the outcome variable, voteshare.

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

The coefficient for the constant term is 0.302. This is the value on the y-axes (voteshare) when the value of both predictor/explanatory variables is 0. Basically, the proportion of votes that go to Ken C. is 0.302 in precincts with no signs posted that are also not adjacent to any precincts with signs posted.

(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?

First, I conducted an overall F-test just to see generally if the regression model is at all useful in explaining variation in the outcome variable. We somewhat already know there is some relationship between the independent variables and the outcome through our t-test results, but this is an additional check. When we reject the overall f test, it indicates that at least one of our predictors does have some kind of linear relationship with the outcome.

For this F test, I first listed my hypotheses. H0 = the sum of all coefficient slopes is 0. Ha = at least one of the coefficient slopes $=/= 0$.

Second, I calculated the F Test Statistic with the following code:

```
F_Stat <- ((0.094/(2-1))/((1-0.094)/(131-2)))
```

F_Stat = 13.384

Third, I calculated the F p-value with the following code:

```
df1 <- 2-1
df2 <- 131-2-1

F_pval <- df(F_Stat, df1, df2)
```

F_pval = 0.00018

This is very close to 0, which is an indication that we can reject our H0, and it is likely that at least one of our explanatory variables is related to our outcome variable.

In conclusion however, the R squared value is very small (0.094). While both variables do seem to have a linear relationship with the outcome (according to our t- and F-tests), the proportion of variance in outcome (voteshare) that can be explained by this model is quite low (this is what is explained with R squared); thus other variables should be added to the model as they may have more explanatory power / may give the model better fit.