Problem Set 4

Applied Stats/Quant Methods 1

Due: November 18, 2024

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 Monday November 18, 2024. 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`).

```
# for the "type" column, "prof" = 1, else = 0
Prestige$professional <- ifelse(Prestige$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.)

```
# For curiosity, look at the additive model first
lm(prestige ~ income + professional, data = Prestige)
library(scatterplot3d)
scatterplot3d(Prestige$income, Prestige$professional, Prestige$prestige)

# As well, plot as simple 2D projection (i.e. just bivariate baseR plot
    of income and prestige coloured by profession)
colors <- ifelse(Prestige$profession == 1, "red", "blue")
plot(Prestige$income, Prestige$prestige,
     col = colors,
     xlab = "Income",
     ylab = "Prestige")

# Back to question 1b:
model1 <- lm(prestige ~ income + professional + income:professional, data
    = Prestige)
```

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

```
# hard-coding the results of the regression from 1b, we get the following equation
Y1 <- 21.142259 + (0.003171*Prestige$income) + (37.781280*Prestige$professional) - (0.002326*(Prestige$income*Prestige$professional))
```

(d) Interpret the coefficient for `income`.
When `professional` = 0 for so-called "non-professional" jobs, for every 1 unit increase in `income` there is a 0.003171 increase in prestige units

(e) Interpret the coefficient for `professional`.
When income is 0 and `professional` changes from 0 to 1, there is a 37.781280 increase in prestige units

(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).

```
# When professional is set to 1, our full question...
Y1 <- 21.142259 + (0.003171*Prestige$income) + (37.781280*Prestige$
    professional) - (0.002326*(Prestige$income*Prestige$professional))

# ...simplifies to
Y2 <- 58.92354 + 0.000845*Prestige$income # again, also hard-coded

# For every $1000 increase in income, we have 0.000845*1000, or 0.845.
# Excluding the intercept value, this equates to a marginal effect of an
    increase in 0.845 prestige units for every $1000 increase
```

(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).

```
# First input 6000 for income and 1 for professional:
ans_1g1 <- 21.142259 + (0.003171*6000) + 37.781280 - (0.002326*6000*1)
# Second, calculate for non-professional jobs
ans_1g2 <- 21.142259 + 0.003171*6000
print(ans_1g1 - ans_1g2)
```

```
[1] 23.82528
```

In other words, a step value change in occupation from non-professional to professional, but keeping an income of $6000, is associated with a 23.82528 increase in prestige units.

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 $\alpha = .05$).

```
# From the results of the regression detailed in Question 2, we have
    enough information to calculate the t-statistics for both regression
    coefficients
# To calculate our test-statistic, divide the estimated coefficient by
    the standard error (code below is raw-values, not abstracted)
t_stat_assinged <- 0.042/0.016

# calculate the degrees of freedom
dfreedom <- 131 - 2 - 1
# Now take this value and calculate the p-value. Given N - k - 1, we have
    131 (number of observations) - 2 (number of variables) - 1 (number of
    statistics used)
```

[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.

```r
# Remember to use the formula for a two-sided t-test,
p_assigned <- 2*pt(abs(t_stat_assinged), df=dfreedom, lower.tail = FALSE)
    # or equivalently and more compactly 2*pt(-abs(2.625), df=128)
p_assigned < 0.05
```

This gives a result of

```
[1] 0.00972
```

Can reject a null hypothesis that the presence of yard signs in a precinct does not effect vote share, as 0.00972 is less than 0.05

(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 $\alpha = .05$).

```
# repeat procedure from (a) for the values provided for adjacent
    precincts
t_stat_adjacent <- 0.042/0.013
p_adjacent <- 2*pt(abs(t_stat_adjacent), df=dfreedom, lower.tail = FALSE)
p_adjacent < 0.05
```

This gives a result of

```
[1] 0.00156946
```

Can reject a null hypothesis that the presence of yard signs in a nearby precinct does not effect proportion of vote, as 0.00157 is less than 0.05

(c) Interpret the coefficient for the constant term substantively.
This is the vote share for candidate KC for precincts that neither had a yard sign against candidate TMcA, or were beside a precinct that did. In other words, controlling for lawn signs there was a 0.302 proportion of the vote for KC.

(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-squared, or variance in proportion of vote explained by our explanatory variables, has a value of only 0.094. This indicates that it is important not to omit or fail to consider other potential explanatory variables that could explain the variance in proportion of the vote.

As an aside, we also have enough information provided to calculate the F-statistic, so we can consider goodness of fit in conjunction with statistical significance of fit. The F-stat is lower than 0.05, but this is unsurprising as we have already conducted our single hypothesis tests above. Thus, even though the R-squared tells us yard signs only explain a small amount of the variation in voting, there may still be a minor effect, at least in this sample.