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`).

```
Prestige$Professional <- ifelse(Prestige$type == "prof", 1, 0)
Prestige <- subset(Prestige, select = -type)
head(Prestige)
tail(Prestige)
```

```
                  education  income  women  prestige  census  Professional
gov.administrators     13.11  12351  11.16      68.8    1113             1
general.managers       12.26  25879   4.02      69.1    1130             1
accountants            12.77   9271  15.70      63.4    1171             1
purchasing.officers    11.42   8865   9.11      56.8    1175             1
chemists               14.62   8403  11.68      73.5    2111             1
physicists             15.64  11030   5.13      77.6    2113             1
                  education  income  women  prestige  census  Professional
train.engineers         8.49   8845   0.00      48.9    9131             0
bus.drivers             7.58   5562   9.47      35.9    9171             0
taxi.drivers            7.93   4224   3.59      25.1    9173             0
longshoremen            8.37   4753   0.00      26.1    9313             0
typesetters            10.00   6462  13.58      42.2    9511             0
bookbinders             8.55   3617  70.87      35.2    9517             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 $\times$ dummy interaction.)

```
model <- lm(prestige ~ income + Professional + income*Professional, data
   = Prestige)
summary(model)
```

```
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 ***

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.

$$Y_i = b_0 + b_1 * x_i + b_2 * d_i + b_3 * x_i * d_i \tag{1}$$

$$prestige = 21.14 + 0.003 * income + 37.78 * Professional + -0.002 * income * Professional \tag{2}$$

(d) Interpret the coefficient for `income`.

The coefficient for income of 0.003 means that a 1 unit increase in income is associated with a 0.003 increase in Prestige, holding other variables constant.

(e) Interpret the coefficient for `professional`.

The coefficient for Professional suggests that the professional status (1 instead of zero, professionals not blue or white collar) is asociated with a 37.78 increase in prestige holding 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).

$$prestige = 21.14 + 0.003 * 1 + 37.78 * 1 + (0.002 * 1 * 1) \tag{3}$$

$$prestige = 58.92 \tag{4}$$

$$prestige = 21.14 + 0.003 * 1000 + 37.78 * 1 + (-0.002 * 1000 * 1) \tag{5}$$

$$prestige = 61.92 \tag{6}$$

`For the sake of comparison, first I assumed that the base value of income is 1, because we want to know the effect of a $1000 increase. So with an assumed base income, prestige should be 58.92, considering the variable Professional to be 1, which is what we want to compare. An increase in $1000 is associated with an expected increase of 3 units in average prestige, holding Professional constant and (^y) according to the prediction equation in (c)`

(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).

$$prestige_np = 21.14 + 0.003 * 6000 + 37.78 * 0 + (-0.002 * 6000 * 0) \tag{7}$$

$$prestige_np = 39.14 \tag{8}$$

$$prestige_pro = 21.14 + 0.003 * 6000 + 37.78 * 1 + (-0.002 * 6000 * 1) \tag{9}$$

$$prestige_pro = 64.92 \tag{10}$$

`The marginal effect of changing someone's occupation from non_professional to professional is an expected increase 25.78 units in average prestige, when income is held constant.`

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 $\alpha = .05$).

With the coefficient as 0.042 and the standard error as 0.016 we can do two tailed t-test where:
t = 0.042/0.016 = 2.62
With the degree of freedom as df = N - k - 1 = 128
Using a t-test table the critical value is 1.98. The t-test value is greater than the critical value.
The results of the linear regression suggest that the presence of the yard signs affects Ken Cuccinelli's vote share and the result of the t-test suggests that the effect is statistically significant and we can reject the null hyphotesis.

[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.

The null hyphothesis would be that the yard signs have no effect on Ken Cuccinelli's vote share. The t-test is two tailed because the question states that we want to learn "the effect" of the yard signs over vote share.

(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 $\alpha = .05$).

With the coefficient as 0.042 and the standard error as 0.013. we can do a two tailed t-test where:
t = 0.042/0.013 = 3.23
With the degree of freedom as df = 128
Using a t-test table the critical value is 1.98. The t-test value is greater than the critical value.
The results of the linear regression suggest that the presence of adjacent yard signs affects Ken Cuccinelli's vote share and the result of the t-test suggests that the effect is statistically significant and we can reject the null hyphotesis.

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

The coefficient fo the constant term means that on average, the baseline of Ken C."s vote share is 0.302 when two variables are at their reference levels.

(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 Rsquared value 0.094 indicates that 9.4% of the variation in Ken C's vote share is explained by the yard signs. The other 90.6% of the variation is not explained by the model. We don't have enough information to infer the importance of yard signs against other models.