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`).

```
# read in data
data(Prestige)
help(Prestige)
View(Prestige)

# set wd for current folder
setwd("/Users/poisson/Documents/GitHub/Fork_Statsl Fall2023")
getwd()

###### Q1 #####

#(a)
# check if value is na
Prestige$type[is.na(Prestige$type)]
Prestige$prestige[is.na(Prestige$prestige)]
Prestige$income[is.na(Prestige$income)]

# ignore missing values
Prestige <- Prestige[!is.na(Prestige[, 'type']), ]

# create dummy variables for type
unique(Prestige$type)
Prestige$professional <- ifelse(Prestige$type == 'prof', 1, 0)
Prestige$professional <- as.factor(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.)

```
#(b) run a model
model <- lm(prestige ~ income + professional + income : professional,
    data = Prestige)
summary(model)
texreg(list(model), digits=3)
```

	Model
(Intercept)	21.142***
	(2.804)
income	0.003***
	(0.000)
professional	37.781***
	(4.248)
income:professional	−0.002***
	(0.001)
R^2	0.787
Adj. R^2	0.780
Num. obs.	98

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1: Fitted regression model with an interation

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

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{X} + \hat{\beta}_2 \times \text{D} + \hat{\beta}_3 \times \text{X} \times \text{D}$$

Professional : when D = 1

$$\begin{aligned} \textit{prestige} &= 21.142 + 0.003 \times \text{income} + 37.781 \times 1 - 0.002 \times \text{income} \times 1 \\ \textit{prestige} &= 58.924 + 0.001 \times \text{income} \end{aligned}$$

Non-professional (Blue and white collar workers): when D= 0

$$\begin{aligned} \textit{prestige} &= 21.142 + 0.003 \times \text{income} + 37.781 \times 0 - 0.002 \times \text{income} \times 0 \\ \textit{prestige} &= 21.142 + 0.003 \times \text{income} \end{aligned}$$

(d) Interpret the coefficient for `income`.

There is a positive and statistically reliable relationship between the explanatory variable `income` and the response variable `prestige`. Given that the influence of the interaction term is statistically significant, the average effect on prestige of a one-unit increase in income depends on the type of occupations. In comparison to blue or white collar workers, one unit increase in the income of professional is associated with an average increase of 0.001 in the prestige score. In contrast, one unit increase in the income of non-professional is associated with an average increase of 0.003 in the prestige score.

(e) Interpret the coefficient for `professional`.

There is a poistive and statistically reliable relationship between the explanatory variable `professional` and the response variable `prestige`. The coefficient of `professional` represents the average change of prestige score when one's occupations from non-professional to professional. In this case, as D changes from 0 to 1, the prestige score increase 37.781 on average. In other words, professionals have a 37.781 scale point higher score on prestige compared to non-professional on average, under control the income variable.

(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).

```
#(f) marginal effect of income when professional = 1
y_hat1 = 21.1422589 + 0.0031709 * 1000 + 37.7812800 * 1 - 0.0023257 *
    1000 * 1
y_hat2 = 21.1422589 + 0.0031709 * 0 + 37.7812800 * 1 - 0.0023257 * 0 * 1
print(y_hat1-y_hat2)
```

Given the type of one's occupation is professional (D takes 1 as value), each \$1000 increase in his or her income increase the score of prestige by 0.8452 units 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).

```
#(g) marginal effect of income when income = 6000
y_hat3 = 21.1422589 + 0.0031709 * 6000 + 37.7812800 * 1 - 0.0023257 *
    6000 * 1
y_hat4 = 21.1422589 + 0.0031709 * 6000 + 37.7812800 * 0 - 0.0023257 *
    6000 * 0
print(y_hat3-y_hat4)
```

Given someone changes his or her job from non-professional to professional and the income remains \$6000 (X takes 6000 as value) , the prestige score increase by 23.827 units on average.

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 $\alpha = .05$).

Null: Having these yard signs in a precinct does not affect vote share.

Alternative: Having these yard signs in a precinct has an effect on vote share.

$$H_0 : \beta_1 = 0$$

$$H_A : \beta \neq 0$$

```
#(a) Hypotheis test for coefficient of 'precinct assigned lawn signs'
b1 <- 0.042
se1 <- 0.016
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.

```r
k <- 2
p_b1 <- 2*pt((b1-0)/se1, n-k, lower.tail = F) # p_b1 = 0.009711646
print(p_b1 < 0.05) # True
```

The estimated coefficient of 'precinct assigned lawn signs' is statistically differentiable from 0 at the a = 0.05 level because the p-value < 0.05 ($\approx .01$). We have sufficient evidence to reject the null hypothesis.

(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 $\alpha = .05$).

Null: Being next to precincts with these yard signs does not affect vote share.

Alternative: Being next to precincts with these yard signs has an effect on vote share.

$$H_0 : \beta_2 = 0$$

$$H_A : \beta \neq 0$$

```r
#(b) Hypotheis test for coefficient of 'precinct adjacent to lawn signs'
b2 <- 0.042
se2 <- 0.013
n <- 131
k <- 2
p_b2 <- 2*pt((b2-0)/se2, n-k, lower.tail = F) # p_b2 = 0.001566685
print(p_b2 < 0.05) # True
```

The estimated coefficient of 'precinct adjacent to lawn signs' is statistically differentiable from 0 at the a = 0.05 level because the p-value < 0.05 ($\approx .002$) . We have sufficient evidence to reject the null hypothesis.

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

$Constant = 0.0302$

It can be interpreted as that the predicated proportion of votes for Ken Cuccinelli among remaining 25 control groups. In other words, in those where were not posted the yard signs and not adjacent to a precinct in treatment groups(Precinct assigned lawn signs = 0 and Precinct adjacent to lawn signs = 0) the predicated Ken Cuccinelli's vote average is about 30.2%.

(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?

We can evaluate the model fit using determined coefficient R^2=0.094. It represents the liner regression model is able to explain about 9.4% of the variation in the dependent variable yard signs, but there is still a large portion of the variation in vote share that is not explained by the model. This tells us that the yard signs have a certain degree of impact on changes in vote share based on our statistical results, but if we are interested what factors cause changes in vote share, we need to continue to investigate and add other variables into this model.