

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**).

```
1 # part 1
2 professional <- ifelse(Prestige$type == "prof", 1, 0)
3 # professional <- na.omit(professional) # 98 obs - ones included equal:
4 print(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 \times dummy interaction.)

Table 1: Regression Results

	<i>Dependent variable:</i>
	prestige
income	0.003*** (0.0005)
professional	37.781*** (4.248)
income:professional	-0.002*** (0.001)
Constant	21.142*** (2.804)
Observations	98
R ²	0.787
Adjusted R ²	0.780
Residual Std. Error	8.012 (df = 94)
F Statistic	115.878*** (df = 3; 94)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

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

$$Y_i = 21.142 + 0.003(X_i) + 37.781(D_i) + (-0.002(D_i)(X_i))$$

$$Y_i = 21.142 + 0.003(X_i) + 37.781(D_i) - 0.002(D_i)(X_i)$$

- (d) Interpret the coefficient for **income**.

The income coefficient β_1 is the prestige increase associated with a 1-dollar increase in income for non-professionals, on average 0.003 points of an increase.

- (e) Interpret the coefficient for **professional**.

The professional coefficient β_2 indicates part of the effect of being a professional on prestige. When considered along with the interaction term, professionals are expected to have, on average and holding income constant, 37.779 more prestige points than non-professionals at the same income amount.

- (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).

```

1 # part 6
2 # Yi = 21.142 + 0.003(Xi) + 37.781(Di) + (-0.002(Di)(Xi))
3 # for pros = 58.923 + 0.003(Xi) - 0.002(Xi) = 58.923 + 0.001(Xi)
4
5 # Yi+1000 where (Xi +1000) = 21.142 + 0.003(Xi+1000) + 37.781(Di) +
   (-0.002(Di)(Xi+1000))
6 # = 21.142 + 0.003(Xi+1000) + 37.781(1) - 0.002(1)(Xi+1000)
7 # = 21.142 + 0.003(Xi) + 0.003(1000) + 37.781 - 0.002(Xi) - 0.002(1000)
8 # = 21.142 + 0.003(Xi) + 3 + 37.781 - 0.002(Xi) - 2
9 -2+21.142+3+37.781 # 59.923
10 # = 59.923 + 0.003(Xi) - 0.002(Xi)
11 # = 59.923 + 0.001(Xi)
12
13 # Yi+1000 - Yi = [59.923 + 0.001(Xi)] - [58.923 + 0.001(Xi)] = 1
14
15 # around 1 prestige point
16
17 # test
18 21.142 + 0.003*(6000) + 37.781*(1) - 0.002*(1)*(6000) # 64.923
19 21.142 + 0.003*(7000) + 37.781*(1) - 0.002*(1)*(7000) # 65.923

```

- (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).

```
1 # part 7
2 #  $Y_i = 21.142 + 0.003(X_i) + 37.781(D_i) + (-0.002(D_i)(X_i))$ 
3 # for profs =  $58.923 + 0.001(X_i)$ 
4 # for non profs =  $21.142 + 0.003(X_i)$ 
5 # change in  $\hat{y}$  to prof =  $37.781 - 0.002(X_i)$ 
6  $37.781 - 0.002*(6000)$  # 25.781 point change in  $\hat{y}$ 
```

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 Virginia 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 McAuliffe’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 $\alpha = .05$).

```
1 # part 1 & 2
2
3 Estimates <- c(0.042,0.042,0.302)
4 Standard_Errors <- c(0.016, 0.013, 0.011)
5 coefficients <- c('lawn sign', 'adjacent', 'constant')
6 # tscores <- (Estimates-0)/(Standard_Errors)
7 # p_values <- 2*pt(abs(tscores), 131-3, lower.tail = F)
8 model_results <- data.frame(coefficients, Estimates, Standard_Errors)
9 print(model_results)
10
11 tscoresmod <- (model_results$Estimates-0)/(model_results$Standard_Errors)
12 p_valuesmod <- 2*pt(abs(tscoresmod), 131-3, lower.tail = F) # n = 131, k
    = 3
```

¹Donald 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.

```

13 model_results <- data.frame(model_results , p_valuesmod)
14
15
16 print(model_results)
17 #      precinct Estimates Standard_Errors  p_valuesmod
18 # 1 lawn sign      0.042           0.016 9.720020e-03
19 # 2 adjacent      0.042           0.013 1.569460e-03
20 # 3 constant      0.302           0.011 1.738775e-55

```

The first coefficient i.e. β_1 is the average increase in voteshare for McAuliffe's opponent for a change from a no sign precinct to a precinct with signs, i.e. a 4.2 percent average increase in voteshare. This increase is significant at the 98 percent confidence interval ($\alpha = 0.02$) as are the other coefficients.

- (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 $\alpha = .05$).

As above, the change from a precinct with no signs to a precinct which is adjacent to one with signs is significant when $\alpha = 0.02$, with a 4.2 percent on average increase in voteshare for the opponent. It is interesting that the effect is equal to having signs in one's own precinct, and this similarity is worth investigating further.

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

The constant coefficient is the y-intercept of the model, i.e. the value of voteshare for McAuliffe's opponent when not in a precinct with lawn signs. In this model, vote share for the opponent in such a precinct is around 30 percent.

- (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?

```

1 # part 4
2 # R-squared = 0.094, the coefficient of determination
3 # low R-squared. the importance of lawn signs in explaining the variation
4 # in voteshare is minimal. the above model has the ability to explain
5 # less than a tenth of the variation in voteshare in this state.
6
7 R.squared <- 0.094
8 F.test <- (R.squared/(3 - 1)) / ((1 - R.squared)/(131 - 3))
9 df1 <- 3 - 1
10 df2 <- 131 - 3
11 F.pvalue <- df(F.test , df1 , df2)
12 F.pvalue # 0.001634304

```

The R-Squared gives an indication that this model only very marginally explains variation in voteshare, i.e. below 10 percent. However, calculating the F-statistic which is significant when $\alpha = 0.02$, indicates that the regression model is at least somewhat useful in explaining that variation, i.e. not all of the coefficients are equal to zero. This tells us that while yard signs might explain a small amount of the opponent's share of the vote, other factors not measured are needed to reliably explain or predict the overall variation.