

# Problem Set 4

Applied Stats/Quant Methods 1

Dan Zhang 23335541

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**).

```
1 Prestige$professional<-ifelse (Prestige$type %in% c("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  $\times$  dummy interaction.)

```
1 # Run a linear model without interaction
2 model<-lm(prestige ~ income + professional + income:professional ,data =
  Prestige)
```

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

The results of the model as followed:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{income} + \hat{\beta}_2 \times \text{professional} + \hat{\beta}_3 \times \text{income} \times \text{professional}$$

$$\text{prestige} = 20.804 + 0.003 \times \text{income} + 38.120 \times \text{professional} - 0.002 \times \text{income} \times \text{professional}$$

Table 1: Outcome variable is **income ,professional** and the explanatory variable is **presige**.

	Model 1
(Intercept)	20.804*** (2.539)
income	0.003*** (0.000)
professional	38.120*** (4.080)
income:professional	-0.002*** (0.001)
R <sup>2</sup>	0.789
Adj. R <sup>2</sup>	0.783
Num. obs.	102

\*\*\* $p < 0.001$ ; \*\* $p < 0.01$ ; \* $p < 0.05$

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

There is a positive and statistically reliable relationship between income and prestige. For blue or white workers, with every additional one unit of income, the prestige increases by 0.003 scale points. For professionals, with every additional one unit of income, the prestige increases by 0.001 scale points. (0.003-0.002=0.001)

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

Holding all other variables constant, being a professional occupation is associated with an increase of 38.120 units in prestige compared to a blue or white worker.

- (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).

For linear regression models, the marginal effect can be calculated by taking the derivative of the model's coefficients.

$$\frac{\Delta \text{prestige}}{\Delta \text{income}} = 0.003 - 0.002 \times \text{professional}$$

$$\Delta \hat{y} = \Delta \text{prestige} = (0.003 - 0.002 \times \text{professional}) \times \Delta \text{income}$$

When **income** increases 1000, which means  $\Delta \text{income} = 1000$ , and **professional** = 1  
Then:

$$\begin{aligned}\Delta \hat{y} &= (0.003 - 0.002 \times 1) \times 1000 \\ &= (0.001) \times 1000 \\ &= 1\end{aligned}$$

So, a \$1000 increase in **income** for professional occupations, the changes in **prestige** is 1 scale point.

- (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).

When one's occupation changes from non-professional to professional, and her income is \$6000. The marginal effect is the amount change of  $\hat{y}$  when **professional** variable changes.

We need to consider the main effect and the interaction effect of **professional** on **prestige**.

$$\frac{\Delta_{\text{prestige}}}{\Delta_{\text{professional}}} = 38.120 - 0.002 \times \text{income}$$

$$\Delta \hat{y} = \Delta_{\text{prestige}} = (38.120 - 0.002 \times \text{income}) \times \Delta_{\text{professional}}$$

As we know above  $\Delta_{\text{professional}} = 1 - 0 = 1$  and  $\text{income} = 6000$  :

$$\begin{aligned} \Delta \hat{y} &= (38.120 - 0.002 \times 6000) \times 1 \\ &= (38.120 - 12) \\ &= 26.120 \end{aligned}$$

So, when one's income is 6000 and her occupation changes from non-professional to professional, the changes in **prestige** is 26.120 scale points.

## Question 2: Political Science

Researchers are interested in learning the effect of all of those yard signs on voting preferences.<sup>1</sup> 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  $\alpha = .05$ ).

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

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

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

$\beta_1$  is the coefficient of `Precinct assigned lawn signs` variable.

Run the following code in R, use two sides t-test to calculate the p-value. The degree of freedom:  $df = 131 - 3 = 128$ .

---

<sup>1</sup>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.

```

1 # Get the coefficients and standard error from the model results
2 coefficients<-c(0.042,0.042,0.302)
3 standard_error<-c(0.016,0.013,0.011)
4
5 # Calculate the t-statistic
6 t_stat_1<- coefficients[1]/standard_error[1]
7
8 # Calculate the two-sides p-value
9 p_value_1<-2*pt(-abs(t_stat_1),df=128)

```

Then we can check out the p-value=0.00972002 which is less than  $\alpha = 0.05$ . So we can reject null hypothesis. Which means that the precinct where yard signs are placed has a significant effect on 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  $\alpha = .05$ ).

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

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

$$H_0 : \beta_2 = 0$$

$$H_A : \beta_2 \neq 0$$

$\beta_2$  is the coefficient of Precinct adjacent lawn signs variable.

Run the following code in R, use two sides t-test to calculate the p-value. The degree of freedom:  $df = 131 - 3 = 128$ .

```

1 #Calculate the t-statistic
2 t_stat_2<-coefficients[2]/standard_error[2]
3
4 # Calculate the two-sides p-value
5 p_value_2<-2*pt(-abs(t_stat_2),df=128)

```

Then we can check out the p-value=0.00156946 which is less than  $\alpha = 0.05$ . So we can reject null hypothesis. Which means that being adjacent to a precinct where yard signs are placed has a significant impact on vote share.

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

The coefficient of the constant term is the estimate of the response variable (vote share) when all other variables are zero. In this model, the coefficient on the constant term is 0.302, indicating that, controlling for other variables, the average vote share in precincts that are not assigned yard signs or are not adjacent to precincts that are assigned yard signs is 0.302.

- (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?

Standard error: The standard errors for both **Precinct assigned lawn signs** and **Precinct adjacent lawn signs** are relatively small compared to the coefficients, which suggests that the estimates are precise.

$R^2$ : R-squared is used to evaluate the model fit.  $R^2$  reflects the proportion of the model's explanatory variables to the variation in the dependent variable. In this model,  $R^2=0.094$  which indicates that about 9.4% of the variability in the vote share can be explained by this model. This is relatively low, the model does provide some explanation into the relationship between lawn signs and vote share. However, a large portion of the variability in vote share is not explained due to factors not included in the model, such as other preferences or backgrounds of voters. The sample size seems big enough, but it's important to consider whether it is representative of the larger population.

In conclusion, this model suggests a certain positive effect of lawn signs on vote share, but it also highlights the importance of other unmodeled factors in the election process.