# 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`).

```
## 1a
Prestige$professional <- ifelse(Prestige$type == "prof", 1,0)
table(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.)

```
## 1b
model1 <- lm(prestige ~ income*professional, data = Prestige)
library(stargazer)
stargazer(model1)
```

Table 1:

| | *Dependent variable:* |
|---|---|
| | 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) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

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

prestige = 21.142 + 0.003income + 37.78professional - 0.002income*professional

(d) Interpret the coefficient for `income`.

A one unit increase in income is associated with a 0.003 unit increase in prestige when professional is equal to 0 - i.e. when the individual is not a professional (is either a blue or white collar worker).

(e) Interpret the coefficient for `professional`.

On average, moving from the baseline group (blue and white collar work) to a professional level is associated with a 37.78 unit increase in prestige score when income is equal to 0.

(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 Professionals:

E[Y|Income = 1000] = 21.142 + (0.003 * 1000) + 37.781 - (0.002 * 1000)

= 59.923

E[Y|Income = 0] = 21.142 + (0.003 * 0) + 37.781 - (0.002 * 0)

= 58.923

The expected prestige level for professionals with an income of \$1000 is 59.923. The marginal effect of going from an income of \$0 to an income of \$1000 for professionals is 1 prestige 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).

For Professionals:

E[Y|Income = 6000] = 21.142 + (0.003 * 6000) + 37.78 - (0.002 * 6000)

= 64.922

For Non-Professionals:

E[Y|Income = 6000] = 21.142 + (0.003 * 6000) + 37.78 * 0 - (0)

= 39.142

Marginal Effect = 226.922 - 201.142

= 25.78

On average, professionals with an income of \$6000 have a prestige level of 25.78 units higher than non-professionals that earn \$6000. Hence the marginal effect of professional jobs when income is \$6000 is 25.78 prestige points.

# 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 $\alpha = .05$).

```
1
2 # Step 1: Assumptions
3 # Quantitative data
4 # Data randomly sampled
5 # Errors are normally distributed with constant variance and 0 mean
6
7 # Step 2: Hypotheses
8 # H0: The yard signs have no effect on vote share
9 # H1: The yard signs have an effect on vote share
10
11 # Step 3: Test Statistic
12 t_stat <- 0.042/0.016
```

[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
# t-stat = 2.625

# Step 4: P-value
df <- 131-2-1
p_value <- 2 * (1 - pt(abs(t_stat), df))
p_value
```

T-stat = 2.265

P-value = 0.0097

Step 5: Conclusion

The p-value is less than 0.05, therefore we can reject the null hypothesis that the having the yard signs in a precinct doesn't affect vote share. There is evidence to suggest that there is a statistically reliable positive relationship between the number of yard signs in a precinct and the percentage of the vote that went to Ken Cuccinelli.

(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 $\alpha = .05$).

```r

# Step 1: Assumptions
# Same as previous

# Step 2: Hypotheses
# H0: Being next to precincts with yard signs has no effect on vote share
# H1: Being next to precincts with yard signs has an effect on vote share

# Step 3: Test Statistic
# t-stat = 3.23
t_stat2 <- 0.042/0.013

# Step 4: P-value
df2 <- 131-2-1
p_val2 <- 2 * (1 - pt(abs(t_stat2), df2))
```

T-stat = 3.23

P-value = 0.0015

Step 5: Conclusion The p-value is less than 0.05, therefore we can reject the null hypothesis that being next to precincts that have the yard sign does not affect vote share. The result indicates a statistically reliable positive relationship between being in the precinct next to yard signs and vote share.

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

The constant represents the proportion of the vote that went to Ken Cuccinelli when the precinct was assigned no signs against Mcauliffe and the precinct adjacent was exposed to no signs against Mcauliffe. Cucinelli's predicted vote share is 30.2% when no precint is assigned the signs against Mcauliffe and the precinct adjacent is exposed to no signs against Mcauliffe.

(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 value indicates that 9.4% of the variation in vote share can be explained by the variation in the precinct assigned lawn signs and the precinct adjacent to lawn signs. This means that a lot of variation in vote share is not captured by the model. The model could therefore help predict vote share more accurately if other independent variables that are associated with vote share were included.