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`).

```
# Create a new variable 'professional' by recoding the 'type' variable
Prestige$professional <- ifelse(Prestige$type %in% c("prof", "bc"), 1,
  0)

# View the updated dataset
head(Prestige)

```

```
                    education income women prestige census type professional
gov.administrators      13.11  12351 11.16     68.8   1113 prof            1
general.managers        12.26  25879  4.02     69.1   1130 prof            1
accountants             12.77   9271 15.70     63.4   1171 prof            1
purchasing.officers     11.42   8865  9.11     56.8   1175 prof            1
chemists                14.62   8403 11.68     73.5   2111 prof            1
physicists              15.64  11030  5.13     77.6   2113 prof            1
```

(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.)

```
# Run a linear model with prestige as the outcome
Regression_1 <- lm(prestige ~ income * professional, data = Prestige)

# Display the summary of the model
summary(Regression_1)

```

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.759e+01  5.649e+00   4.884 4.04e-06 ***
income               2.821e-03  1.070e-03   2.636  0.00976 **
professional        -5.594e-01  6.273e+00  -0.089  0.92913
income:professional  8.629e-05  1.114e-03   0.077  0.93844
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.21 on 98 degrees of freedom
Multiple R-squared:  0.5111,Adjusted R-squared:  0.4962
F-statistic: 34.15 on 3 and 98 DF,  p-value: 3.386e-15
```

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

$$Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + E \quad \text{(Multiple Linear Regression Model)}$$
$$\text{Prestige} = B_0 + B_1 \cdot \text{Income} + B_2 \cdot \text{Professional} + B_3 \cdot \text{Income} \cdot \text{Professional}$$
$$\text{Prestige} = 2.759 + 2.821 \cdot \text{Income} - 5.594 \cdot \text{professional} + 8.629 \cdot \text{income:professional}$$

(d) Interpret the coefficient for `income`.

For each one-unit increase in income, the Prestige is expected to increase by 2.821 units, assuming all other variables are held constant.

(e) Interpret the coefficient for `professional`.

The coefficient of -5.594 indicates that, on average, professionals have a Prestige score that is 5.594 units lower than that of non-professionals, assuming all other variables are held 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).

$$\text{Prestige} = 2.759 + 2.821 \cdot \text{Income} - 5.594 \cdot \text{professional} + 8.629 \cdot \text{income:professional}$$

If income = 1,000 then 2.821 * 1,000= 2,821

A 1,000 usd increase in income is associated with a 2,821 unit increase in the predicted Prestige score for professional occupations.

(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).

$$\text{Prestige} = 2.759 + 2.821 \cdot \text{Income} - 5.594 \cdot \text{professional} + 8.629 \cdot \text{income:professional}$$

$$\text{Prestige} = B_0 + B_1 \cdot \text{Income} + B_2 \cdot \text{Professional} + B_3 \cdot \text{Income} \cdot \text{Professional}$$

Change in Income = 6,000 usd

Change in Y = 2.821 * 6,000 + 8.629 * 6,000 * 1

Change in Y = 68,700

The estimate change in the predicted Prestige score associated with changing one's occupation from no-professional to professional when income is 6,000 is approximately 68,700 usd.

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

[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.

(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 $\alpha = .05$).

```
# Option A:
B1 <- 0.042
SEb1 <- 0.016
N <- 131

# Calculate the test statistic
 test_statistic <- B1 / SEb1

# Degrees of freedom
 df <- N - 1

# Two-tailed test, so multiply by 2
 p_value <- 2 * pt(-abs(test_statistic), df)

# Compare p-value to significance level (e.g., 0.05)
 if (p_value < 0.05) {
   print("Reject the null hypothesis: Having yard signs in a precinct affects vote share.")
 } else {
   print("Fail to reject the null hypothesis: No evidence that yard signs affect vote share.")
 }
```

H0: B1 equal to 0

H1: B1 non-zero

B1 = 0.042

SE(B1) = 0.016

t-statistic t1= B1/SE(B1) = 0.042/0.016 = 2.625

Significance level $\alpha = .05$ for a two-tailed test. With N=131 observations

Degrees of freedom DF = N - 3

DF = 131-3 = 128

As t1 (2.625) is greater than +-1.978, we reject the null hypothesis. The results suggest that the yard signs appear to influence the vote share in the precincts.

(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 $\alpha = .05$).

```
# Option B:
B2 <- 0.042
SEb2 <- 0.013

# Calculate the test statistic
test_statistic <- B2 / SEb2

# Degrees of freedom
df <- N - 1

# Two-tailed test, so multiply by 2
p_value <- 2 * pt(-abs(test_statistic), df)

# Compare p-value to significance level (e.g., 0.05)
if (p_value < 0.05) {
  print("Reject the null hypothesis: Being next to precincts with yard
    signs affects vote share.")
} else {
  print("Fail to reject the null hypothesis: No evidence that being
    adjacent to yard signs affects vote share.")
}
```

H0: B2 equal to 0

H1: B2 non-zero

B2 = 0.042

SE(B2) = 0.013

t-statistic t2= B2/SE(B2) = 0.042/0.013 = 3.23

Significance level $\alpha = .05$ for a two-tailed test. With N=131 observations

Degrees of freedom DF = N - 3

DF = 131-3 = 128

As t2 (3.23) is greater than +-1.978, we reject the null hypothesis.

This suggest that being next to precincts with yard signs has a statistically significance effect on vote share.

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

The constant term or B0 is the estimated value of the dependent variable (proportion of the vote that went to McAuliff's opponent Ken Cuccinelli) when all independent variables are set to zero.

As B0 = 0.302 The starting point for the vote share when there are no yard signs present is 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?

R2 = 0.094 represents the proportion of the variance in the dependent variable that is explained by the independent variables in the model.

In other words 9.4 percent of the variance in the vote share is explained by the variables included in the model (B1 Assigned yard signs and B2 Adjacent to yard signs). The other 90.6 percent is not explained by the variables in the model. This could mean that there are other factors (variables) that influence the vote share.