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`). **The R script for regression is:**

```
install.packages("car", dependencies = TRUE)
library(car)
data("Prestige")
help("Prestige")
Prestige$professional <- ifelse(Prestige$type == "prof", 1, 0)
head(Prestige)
tail(Prestige)
```

(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.)

The R script for regression is:

```
model <- lm(prestige ~ income + professional + income:professional,
            data = Prestige)
summary(model)
```

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

prestige = B0 + B1 × income + B2 × professional + B3 × (income × professional) + e

(d) Interpret the coefficient for `income`.

B1 is the coefficient of income variable. The summary statistics show that there is a positive relationship between both the income (exploratory variable) and prestige (response variable). Although they both are statistically significant as evident by the p-value, the low coefficient value suggests that the impact will be modest.Therefore an expected increase of one unit in income will result in a slight increase of prestige.

(e) Interpret the coefficient for `professional`.

B2 is the coefficient of professional variable. The summary statistics show that there is a positive relationship between both the professional (exploratory variable) and prestige (response variable). The p value suggests that they both are statistically significant and the prestige of individuals who are classified as professional (coded by 1) is approx. 38 units higher than individuals who are not professional (coded by 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).

An increase of \$ 1000 in income has a very low effect on prestige of professional individuals which can be calculated using B3 or the coefficient of interaction while a similar increase is different for non professional individuals as the interaction term is not applied on them. Therefore, the effect of a \$1000 increase in income is less for professionals because of the negative interaction value as compared to non professionals.

(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).

Based on the prediction equation, putting all the values of coefficients in the equation and inserting \$6000 as the value of income, it can be calculated that the difference between prestige of professional and prestige of non professional comes roughly around 38 percent which means that individuals who join professional category are expected to experience an increase of 38 percent in their prestige.

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 $\alpha = .05$). **The R script for regression is:**

```
coeff_signs <- 0.042
se_signs <- 0.016
df <- 128
t_stat <- coeff_signs / se_signs
# similarly, p value may be derived through:
p_value <- 2 * pt(-abs(t_stat), df)
summary(p_value)
```

[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.

The hypotheses may be set up to carry out hypothesis test:

H0: There is no effect of lawn signs on vote share.

H1: There is a significant effect of lawn signs on vote share.

The coefficients of precincts who are assigned lawn signs and standard error are given as 0.042 and 0.016 respectively. The degrees of freedom may be calculated using the formula: df= n - k -1 (df= 131 - 2 -1) which results in 128. The lower p value of 0.00972 which is less than the benchmark value of 0.05 indicates that the null hypothesis may be rejected as there is a significant effect of yard signs 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 $\alpha = .05$).

The R script for regression is:

```
coeff_signs <- 0.042
se_signs <- 0.013
df <- 128
t_stat <- coeff_signs / se_signs
# similarly, p value may be derived through:
p_value <- 2 * pt(-abs(t_stat), df)
summary(p_value)
```

The hypotheses may be set up to carry out hypothesis test:

H0: There is no effect of adjacent lawn signs on vote share.

H1: There is a significant effect of adjacent lawn signs on vote share.

The coefficients of precincts who are assigned lawn signs and standard error are given as 0.042 and 0.013 respectively. The degrees of freedom may be calculated using the formula: df= n - k -1 (df= 131 - 2 -1) which results in 128. The lower p value of 0.001569 which is less than the benchmark value of 0.05 indicates that the null hypothesis may be rejected as there is a significant effect of adjacent yard signs on vote share.

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

The coefficient constant is the baseline or reference level of the response variable when all the value of all predictor variables are zero. In this specific regression, the coefficent constant (B0) is 0.302 having a standard error of 0.011 and this coefficient constant value represents the baseline level of vote share variable without any presence or adjacency lawn signs. The coefficient of lawn signs and adjacent lawn signs represents the effect of these on the response variable.

(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 key indicator in this scenario to evaluate the model fit is the value of R square which is 0.094 or 9.4 percent. This R square basically represents the variance in response variable as explained by predictor variables. In this case, the modest value of R square of 9.4% suggests that the model has a limited ability to explain the variability in response variable that is vote share. Consequently,the majority of variablity may be explained by other factors that are not modeled and the yard signs alone are insufficient to predict vote share.