Problem Set 3

Ella Karagulyan

Applied Stats/Quant Methods 1

Question 1: Economics

In this question, we use the `prestige` dataset in the `car` library. First, we run the following commands to load the data and prepare the environment:

```
# set wd for current folder
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

# loading the necessary packages and data
# install.packages("car")
# install.packages("stargazer")
library(car)
library(stargazer)
data(Prestige)
help(Prestige)
summary(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`).

```
# Creating a new variable
Prestige$professional <- ifelse(Prestige$type == "prof", 1, 0)

# Checking the results
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 $\times$ dummy interaction.)

```r
# Running a regression model with interaction
model1 <- lm(prestige ~ income +
                        professional +
                        income*professional, data = Prestige)
summary(model1)

# Creating a table of the results
stargazer(model1, type = "latex", title = "Linear Regression Results",
          out = "regression_table_model1.tex")
```

The table below shows the results of the regression model with `prestige` as an outcome and `income`, `professional`, and the interaction of the two as predictors.

Table 1: Linear Regression Results

	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.

The multivariate regression equation based on the above model is as follows:

$$Prestige = 21.142 + 0.003 \times (\text{income}) + 37.781 \times (\text{professional}) - 0.002 \times (\text{income} \times \text{professional}))$$

where:

Prestige - Pineo-Porter prestige score for occupation, from a social survey conducted in the mid-1960s.

Income - average income of incumbents in dollars.

Professional - type of occupation, a dummy variable indicating whether the incumbent is a professional or blue/white collar.

(d) Interpret the coefficient for `income`.

With one dollar increase in the income of incumbents the prestige of the profession is on average by 0.003 points higher controlling for the type of profession.

(e) Interpret the coefficient for `professional`.

Compared to blue and white collars professional occupations have by 37.781 points higher prestige score controling fot the income of the incumbents.

(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).

To calculate the marginal effect of \$1,000 increase in income on prestige score for the professionals, let's first calculate the prestige when income is zero.
$Income = 0$

$$Prestige(professional) = 21.142 + 0.003 \times (0) + 37.781 \times (1) - 0.002 \times (0 \times 1))$$

$$Prestige(professional) = 21.142 + 37.781 + 0 = 58.923$$

Now we need to calculate the prestige score when income equals \$1000.
$Income = \$1000$

$$Prestige(professional) = 21.142 + 0.003 \times (1000) + 37.781 \times (1) - 0.002 \times (1000 \times 1))$$

$$\text{Prestige}(professional) = 21.142 + 3 + 37.781 - 2 = 59.923$$

The marginal effect of \$1000 income on the pretige of occupation in case its type is professional equals to:

$$\Delta\text{Prestige(professional)} = 59.923 - 58.923 = 1$$

(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).

Let's first calculate the prestige score when the variable `income` takes the value of $6,000$ and the occupation type is blue/white collar (non-professional).
$Profession = 0$

$$Prestige(income\$6000) = 21.142 + 0.003 \times (6000) + 37.781 \times (0) + (-0.002 \times (6000 \times 0))$$

$$Prestige(income\$6000) = 21.142 + 18 + 0 + 0 = 39.142$$

Now we need to calculate the prestige score when occupation type is professional.
$Profession = 1$

$$Prestige(income\$6000) = 21.142 + 0.003 \times (6000) + 37.781 \times (1) + (-0.002 \times (6000 \times 1))$$

$$\text{Prestige(income} = 6000) = 21.142 + 18 + 37.781 - 12 = 64.923$$

The marginal effect of occupation type when changing from non-professional to professional in case the income of incumbents is \$6000 equals to:

$$\Delta\text{Prestige(income\$6000)} = 64.923 - 39.142 = 25.781$$

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 $\alpha = .05$).

To test whether the regression coefficient is significant, we need to conduct a two-tailed t-test. First, we set up our hypotheses.

H_0: Assigning lawn signs has no effect on the proportion of votes `(voteshare)`. In other words, the effect of precinct assigned lawn size (0.042) on voteshare is not significant.

H_1: Assigning lawn signs has an effect on the proportion of the votes `(voteshare)`. In other words, the effect of precinct assigned lawn size (0.042) on voteshare is significant.

[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.

$$H_0 : \beta_1 = 0 \text{ versus } H_1 : \beta_1 \neq 0$$

Next, we calculate the t-test statistics using the following formula.

$$t = \frac{\beta_1 - 0}{SE(\beta_1)}$$

```
# Calculating the t-statistics for the coefficient beta_1
beta_1 <- 0.042
se_1 <- 0.016
t_statistics_1 <- (beta_1 - 0) / se_1
print(round(t_statistics_1, 4))
```

`The t-test statistics equals to 2.625.`

Next, we calculate the p-value and determine the significance level.

```
# Calculate the p-value based on the t-statistics beta_1
n <- 131
k <- 2
p_value_1 <- 2 * pt(abs(t_statistics_1), df = n - k, lower.tail = FALSE)
print(round(p_value_1, 4))
```

`The p-value equals to 0.0097.`

$H_0 : \beta_1 = 0$ can be rejected since $p < \alpha = 0.05$. Hence, we have enough evidence to reject the null hypothesis, that assigning lawn signs has no effect on the proportion of votes.

(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 $\alpha = .05$).

To test whether the regression coefficient is significant, we need to conduct a two-tailed t-test. First, we set up our hypotheses.

H_0: Being adjascent to lawn signs has no effect on the proportion of votes (`voteshare`). In other words, the effect of being adjacsent to lawn signs (0.042) on voteshare is not significant.

H_1: Being adjascent to lawn signs has an effect on the proportion of the votes (`voteshare`). In other words, the effect of being adjascent to lawn signs (0.042) on voteshare is significant.

$$H_0 : \beta_2 = 0 \text{ versus } H_1 : \beta_2 \neq 0$$

Next, we calculate the t-test statistics using the following formula.

$$t = \frac{\beta_2 - 0}{SE(\beta_2)}$$

```r
# Calculating the t-statistics for the coefficient beta_2
beta_2 <- 0.042
se_2 <- 0.013
t_statistics_2 <- (beta_2 - 0) / se_2
print(round(t_statistics_2, 4))
```

`The t-test statistics equals to 3.2308.`

Next, we calculate the p-value and determine the significance level.

```r
# Calculate the p-value based on the t-statistics beta_2
n <- 131
k <- 2
p_value_2 <- 2 * pt(abs(t_statistics_2), df = n - k, lower.tail = FALSE)
print(round(p_value_2, 4))
```

`The p-value equals to 0.0016.`

$H_0 : \beta_2 = 0$ can be rejected because $p < \alpha = 0.05$. Hence, we have enough evidence to reject the null hypothesis, that being adjascent to lawn signs has no effect on the proportion of votes.

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

$$\begin{aligned}\text{Voteshare} = 0.302 + 0.042 \times (\text{Precinct assigned lawn signs}) \\ +0.042 \times (\text{Precinct adjacent to lawn signs})\end{aligned}$$

The intercept represents the predicted value of `(voteshare)` given that both `precincts assigned lawn signs` and `adjascent to lawn signs` equal zero. In practical terms, the coefficient of the constant term represents the proportion of the vote (0.302) that went to Ken Cuccinelli in precincts other than the 131 ones included in the experiment. Then, being from the precincts that were assigned the sign or those adjascent increase the voteshare by the same percent, hence, it can be concluded that exposure to lawn signs effectively increases the voteshare.

(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 model fit for the regression indicates an R-squared (R^2) of 0.094. This means that roughly 9% of the variance in the `voteshare` is predicted by the two input variables in the model, `precincts assigned lawn signs` and `adjascent to lawn signs`. Although this is a relatively low model fit, and most of the variance is explained by other factors not included in the model, the aim of the analysis was to test the results of the experiment and the effect of exposure to signs on the vote share. Therefore, the model still gives valuable insights.