# Problem Set 4

Applied Stats/Quant Methods 1
Minh Trinh (Student ID:24350478)

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`).

First we run the package, extract and examine the data. Next we create the dummy variable `professional` in a new column.

```
#Question 1
install.packages(car)
library(car)
data(Prestige)
help(Prestige)
df <- Prestige
#create new dummy variable
df$professional <- ifelse(df$type == 'prof',1,0)
#check data
```

Check first several rows:

```
                    professional type
gov.administrators             1 prof
general.managers               1 prof
accountants                    1 prof
purchasing.officers            1 prof
chemists                       1 prof
physicists                     1 prof
```

(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.)

We need to declare our assumption:

- Each observation is independent
- Observations are randomly generated
- There is a linear relationship between mean of the dependent variable and the value of independent variables
- The error term is normally distributed and has constant variable

Run the regression in R and get the model summary:

```
#run regression and check model
model <- lm(prestige ~ income   + professional + income:professional,
   data = df)
```

```
Call:
lm(formula = prestige ~ income + professional + income:professional,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-14.852  -5.332  -1.272   4.658  29.932

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         21.1422589  2.8044261   7.539 2.93e-11 ***
income               0.0031709  0.0004993   6.351 7.55e-09 ***
professional        37.7812800  4.2482744   8.893 4.14e-14 ***
income:professional -0.0023257  0.0005675  -4.098 8.83e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.012 on 94 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.7872, Adjusted R-squared:  0.7804
F-statistic: 115.9 on 3 and 94 DF,  p-value: < 2.2e-16
```

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

General form for prediction equation in this case is:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 D_1 + \beta_3 x_1 D_1$$

Based on the coefficients of the regression on section b we have the following prediction equation:

$$\begin{aligned}\widehat{\text{prestige}} = {} & 21.1422589 + 0.0031709 \times \text{income} \\ & + 37.7812800 \times \text{professional} \\ & - 0.0023257 \times \text{income} \times \text{professional}\end{aligned} \tag{1}$$

With $\widehat{prestige}$ as the predicted value for outcome variable and income as the explanatory variable, professional as the dummy variable and income × professional as the interaction term

(d) Interpret the coefficient for `income`.

The coefficient for `income` means that for people who are not professional (the dummy variable professional = 0), for every 1 unit increases in income, prestige will on average increase by 0.0031709 unit

(e) Interpret the coefficient for `professional`.

The coefficient for `professional` can be interpreted that when `income` variable = 0, people who are in professional job will have on average 37.7812800 more unit in `prestige` compared to people who are in blue collar or white collar job.

(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).

When professional dummy coefficient = 1 we can rearrange formular (1) :

$$\widehat{\text{prestige}} = (21.1422589 + 37.7812800) + (0.0031709 - 0.0023257) \times \text{income} \quad (2)$$

And finally get:

$$\widehat{\text{prestige}} = 58.9235389 + 0.0008452 \times \text{income} \quad (3)$$

So now we know that for professional, \$1,000 increase in income increase the prestige score on average by 0.8452 (0.0008452x1000)

(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 professional dummy coefficient = 0 we can rearrange formular (1) :

$$\widehat{\text{prestige}} = 21.1422589 + 0.0031709 \times \text{income} \quad (4)$$

When one changing from non-professional to professional when her income = \$6000, we will plug in value of income into (3) and (4) and take the subtraction to get the effect.

Prestige score for non-professional is $21.1422589 + 0.0031709 \times 6000 = 40.1676589$
Prestige score for professional is $58.9235389 + 0.0008452 \times 6000 = 63.9947389$

So the marginal effect of professional jobs when the variable income takes the value of 6000 is on average an increase of 23.82708 (63.9947389 - 40.1676589) in prestige score

# 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 $\alpha = .05$).

Lets do hypothesis testing for the effect of having a lawn signs($\beta_1$) with $\alpha = .05$. Here the number of observations is large (n = 131) and there are only 2 independent variable so the degree of freedom is also large. The t-distribution would be very similar to normal distribution. We can do a two-tails t-test with hypothesis:

$H_0$: $\beta_1 = 0$
$H_a$: $\beta_1 \neq 0$

[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.

We then calculate the t value and then p value in R. The degree of freedom for the t distribution is (n - 1 - number of variables) = 131 - 3. According to the result table, we have estimated $\beta_1 = 0.042$ and SE = 0.016

```
#Calculate the t-value for assigned precinct
t_1 = (0.042-0)/0.016
#Calculate the p value for assigned precinct
p_val_1 <- pt(t_1,lower.tail = FALSE,df = 131-3)*2
```

The p value is 0.0097 which is smaller than $\alpha$. We can reject the null hypothesis that $\beta_1 = 0$. We have sufficient evidence to say that having the yard sign in a precinct does have a statistically 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 $\alpha = .05$).

Lets do hypothesis testing for effect being adjacent to precinct having lawn signs($\beta_2$) with $\alpha = .05$. Similarly with part (a) we can do a two-tails t-test with hypothesis:

$H_0$: $\beta_2 = 0$
$H_a$: $\beta_2 \neq 0$

We then calculate the t value and then p value in R. According to the result table, we have estimated $\beta_2 = 0.042$ and SE = 0.013. We also have calculated the degree of freedom for t-distribution above:

```
#Calculate the t-value for adjacent precinct
t_2 = (0.042-0)/0.013
#Calculate the p value for adjacent precinct
p_val_2 <- pt(t_2,lower.tail = FALSE,df = 131-3)*2
```

The p value is 0.0016 which is smaller than $\alpha$. We can reject the null hypothesis that $\beta_2 = 0$. We have sufficient evidence to say that being next to precinct that has lawn sign does have a statistically significant effect on vote share.

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

The constant term is the baseline effect on vote share when variable for precinct assigned lawn signs = 0 and variable for precinct adjacent to lawn signs = 0. We can interpret the constant term as: in Fairfax, Virginia, when a precinct is not assigned a lawn sign and not next to one that has lawn sign, their average value for vote share 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?

The model fit for the regression can be measured by $R^2$. In this model the $R^2 = 0.094$ which means that only 9.4% of the variation in vote share is explained by the variables assigned lawn sign and being adjacent to assigned lawn sign precinct. This

shows that those variables only explain a small proportion of variation in vote share and there should be other factors that can help explains the variation of vote share much better. For example, demographics information of the precinct like average age, income, level of educations would be much more important factors that explain vote share than having a lawn sign.