Problem Set 4

Applied Stats/Quant Methods 1

Zhuo Zhang/23346227

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`).

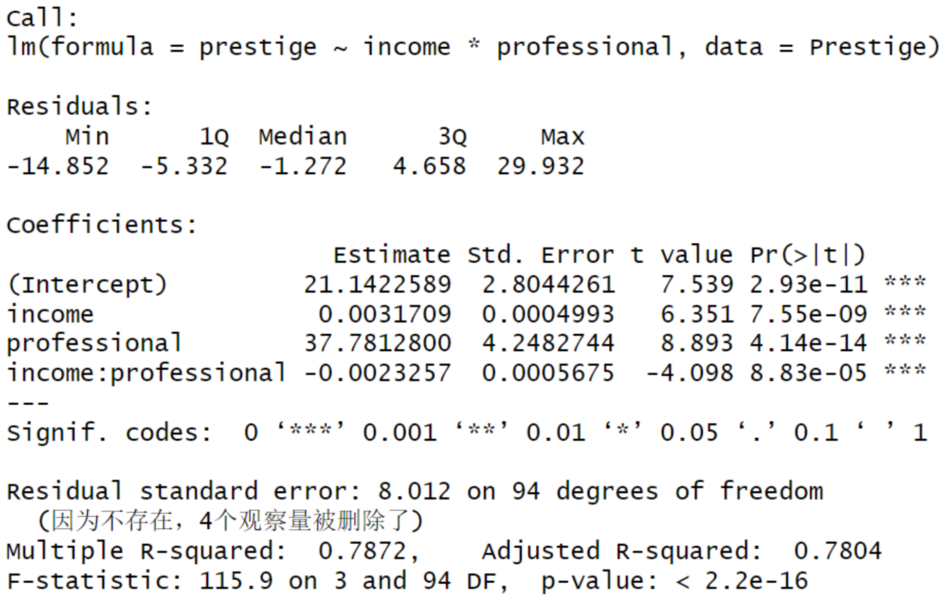
```r
# Create a new variable 'professional'
Prestige$professional <- ifelse(Prestige$type == "prof", 1, 0)
```

(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.)

```r
# Run the linear model
model <- lm(prestige ~ income * professional, data = Prestige)
summary(model)  # View the model summary
pdf("Q1_b.pdf")
# Draw graphics: preview vs. income, and differentiate with different colors based on occupation
plot(Prestige$income, Prestige$prestige,
     col = ifelse(Prestige$professional == 1, "blue", "red"),
     pch = 16,  # Set Scatter Chart Style
     ylab = "prestige", xlab = "income")

# Add regression line
abline(model, col = "green", lty = 2)
legend("topright", legend = c("Non-Professional", "Professional", "Regression Line"),
        col = c("red", "blue", "green"), pch = c(16, 16, NA))
dev.off()
```

Result:

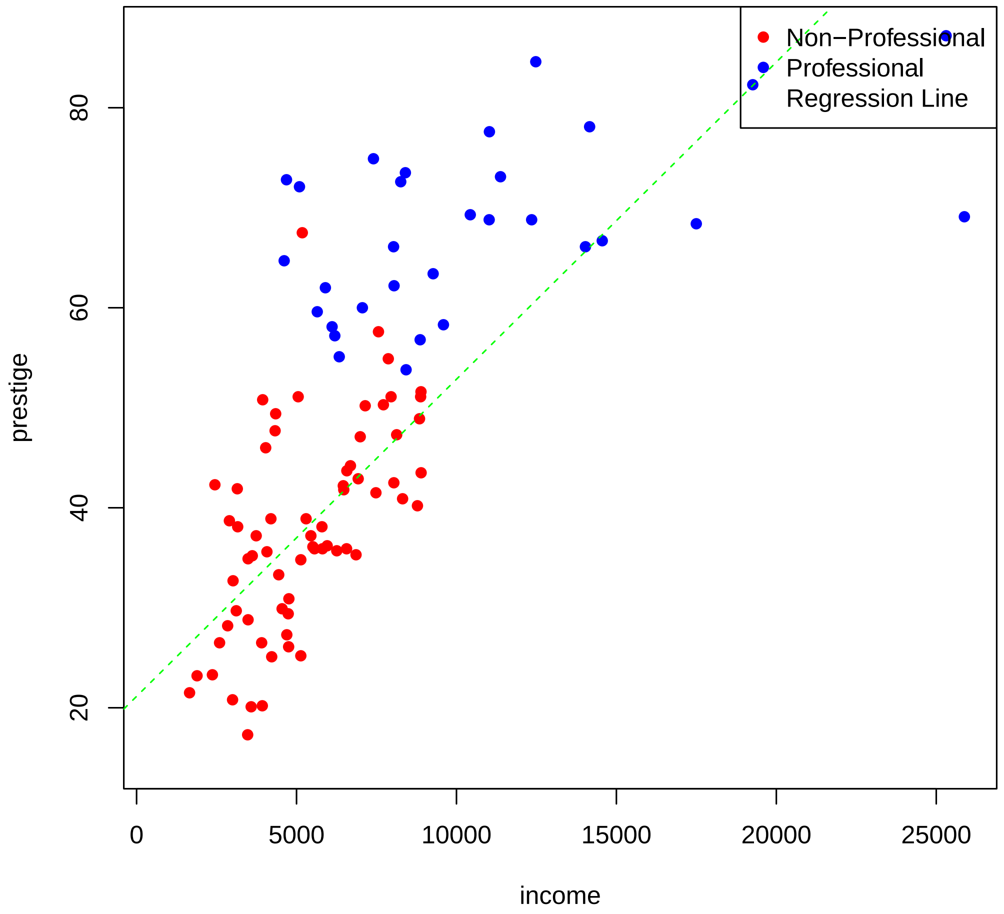


```
Call:
lm(formula = prestige ~ income * professional, data = Prestige)

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
  (因为不存在，4个观察量被删除了)
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.

```
# Extract model coefficients
coefficients <- coef(model)

# Identify the various coefficients in the prediction equation
intercept <- coefficients["(Intercept)"]
income_coef <- coefficients["income"]
professional_coef <- coefficients["professional"]
income_professional_coef <- coefficients["income:professional"]

# Print prediction equation
cat("prediction equation prestige =", round(intercept, 2), "+", round(
    income_coef, 2), "* income +",
     round(professional_coef, 2), "* professional +", round(income_
    professional_coef, 2), "* income * professional")
```

Result:
prediction equation:prestige = 21.14 + 0 * income + 37.78 * professional + 0 * income * professional

(d) Interpret the coefficient for `income`.

Result:
The coefficient of the “income” variable is 0, indicating that “income” has no effect on the predicted prestige score in the model. The coefficient is 0 means that there is no linear relationship between “income” and “prestige”.

(e) Interpret the coefficient for `professional`.

Result:
The coefficient of the ’professional’ variable is 37.78, indicating that when the ’professional’ variable changes from 0 to 1 (indicating a professional profession compared to non professional), the predicted prestige score is expected to increase by 37.78 units. This coefficient indicates a positive linear relationship between professional occupation and prestige scores.

(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).

Result:
prediction equation:prestige = 21.14 + 0 * income + 37.78 * professional + 0 * income * professional
Since the coefficient of the “income” variable is 0, it indicates that “income” has no effect on the predicted prestige score in the model, so an increase of $1000 in income has no effect on the reputation of 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).

Result:
prediction equation:prestige = 21.14 + 0 * income + 37.78 * professional + 0 * income * professional
When professional = 1 and income = 6,000:
prestige = 21.14 + 0 * 6000 + 37.78 * 1 + 0 * 6000 * 1
prestige = 58.92
When professional =0 and income = 6,000:
prestige = 21.14 + 0 * 6000 + 37.78 * 0 + 0 * 6000 * 0

prestige = 21.14
58.92 - 21.14 = 37.78
The change in the predicted prestige score associated with changing one's occupation from non-professional to professional when income is $6,000$ is 37.78.

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 $\alpha = .05$).

Result:
Precinct assigned lawn signs:
Coefficient = 0.042
Standard error = 0.016
To determine whether having these yard signs in a precinct affects vote share:
Hypothesis Test:
Coefficient for "Precinct assigned lawn signs"
H0: Coefficient = 0 (no effect)
H1: Coefficient $\neq$ 0(have effect)
t = $\frac{Coefficient}{StandardError} = \frac{0.042}{0.016}$= 2.625

[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 critical value for a two-tailed test with $\alpha = .05 = 0.05$ is approximately 1.96.
For $2.625 > 1.96$,
I can reject the null hypothesis and conclude that the variable "Precinct assigned lawn signs" has a 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 $\alpha = .05$).

Result:
Precinct adjacent to lawn signs:
Coefficient = 0.042
Standard error = 0.013
To determine whether being next to precincts with these yard signs affects vote share:
Hypothesis Test:
Coefficient for "Precinct adjacent to lawn signs"
H0: Coefficient = 0 (no effect)
H1: Coefficient $\neq$ 0(have effect)
t = $\frac{Coefficient}{StandardError} = \frac{0.042}{0.013}$= 3.231
The critical value for a two-tailed test with $\alpha = .05 = 0.05$ is approximately 1.96.
For $3.231 > 1.96$,
I can reject the null hypothesis and conclude that the variable "Precinct adjacent to lawn signs" has a significant effect on vote share.

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

Result:
The coefficient for the constant term is 0.302.
It represents the expected vote share when all predictor variables (Precinct assigned lawn signs and precinct adjacent to lawn signs) are zero.
It is the expected vote share when there are no yard signs and no adjacency to yard signs.

(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?

Result:
R^2=0.094,
The correlation coefficient R^2 is a statistical indicator of the degree of correlation between two variables, which can reflect the strength of the correlation between two variables in the linear relationship. the larger the R^2, the stronger the relationship between the two variables, the more significant the linear relationship between the variables; on the contrary, the smaller the R^2, the weaker the relationship between the

two variables, the linear relationship between the variables is not significant.

R^2=0.094, the R^2 is low, it suggests that the model does not explain a large proportion of the variability in the dependent variable.

In summary, as included in the model, explain only a small portion of the variability in the vote share. Other unmodeled factors are likely influencing voting preferences to a greater extent. We need to consider additional other variables, as well as refine the model in order to find the important factors that I influence the outcome of the election.