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.

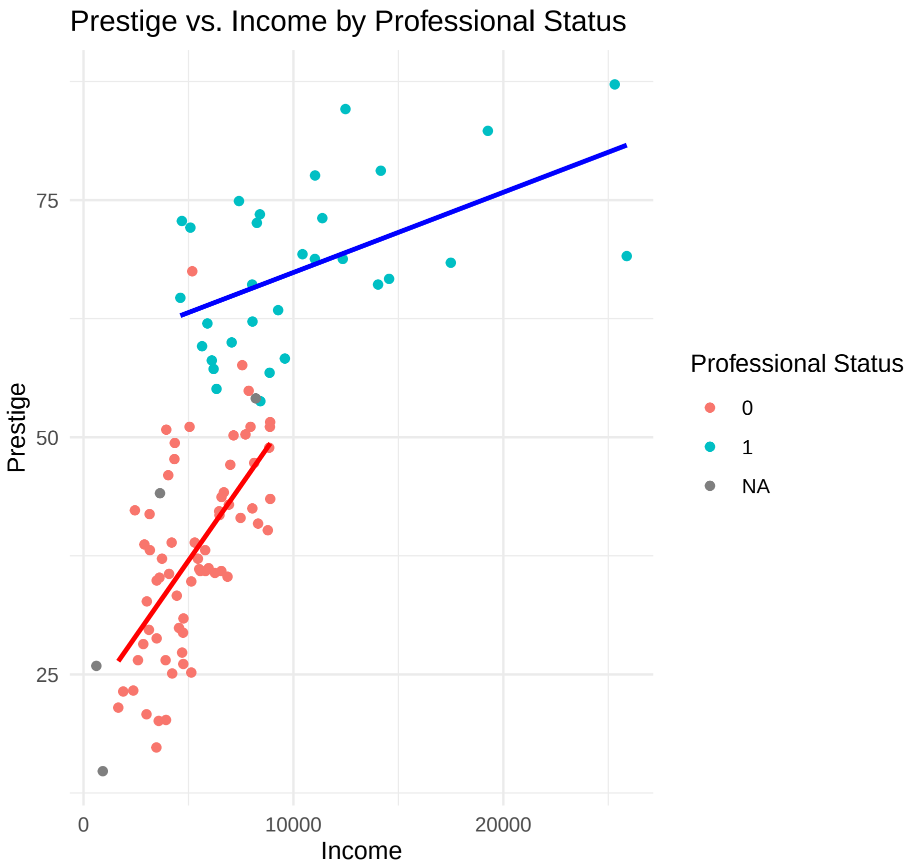
(a) Create a new variable `professional` by recoding the variable `type` so that professionals are coded as 1, and blue and white collar workers are coded as 0 (Hint: `ifelse`).

```r
install.packages("car")
library(car)
data(Prestige)
help(Prestige)
```

```r
Prestige$professional = ifelse(Prestige$type == "prof", 1, 0)
#I use the ifelse function to assign 1 to people coded as professionals under
#the type column and #0 for anyone else (white collars and blue collars)
#I didn't touch the NAs since they seem to be irrelevant per the instructions.
```

(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
l_model = lm(prestige ~ income + professional + income:professional,
             data = Prestige)
#I prepared the model.income:professional shows the interaction term between the two.
summary(l_model)
#Since, the model includes three variables, I research a bit and use ggplot
# to plot it. I ran into errors and used ChatGPT to fix the code.
# https://www.geeksforgeeks.org/multiple-linear-regression-using-ggplot2-in-r/
library(ggplot2)
ggplot(Prestige, aes(x = income, y = prestige)) +
  geom_point(aes(color = factor(professional))) +  # Add points with color based on professional status
  geom_smooth(data = subset(Prestige, professional == 1),
              method = "lm",
              formula = y ~ x,
              se = FALSE,
              color = "blue") +
  geom_smooth(data = subset(Prestige, professional == 0),
              method = "lm",
              formula = y ~ x,
              se = FALSE,
              color = "red") +
  labs(x = "Income",
       y = "Prestige",
       color = "Professional Status",
       title = "Prestige vs. Income by Professional Status") +
  theme_minimal()
```



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

```r
#first I need the coefficients
coef_l_model <- coef(l_model)
#I extract the intercept and slopes for income, professional, and their interaction
intercept <- coef_l_model[1]
slope_income <- coef_l_model[2]
slope_professional <- coef_l_model[3]
slope_interaction <- coef_l_model[4]
#I define the prediction equation with support from ChatGPT as I encountered many errors:
prediction_equation <- function(income, professional) {
  prestige_prediction = intercept + slope_income * income + slope_professional * professional + slope_interaction * income * professional
  return(prestige_prediction)
}
```

(d) Interpret the coefficient for `income`.

```r
#I check the income coefficient:
income_coefficient <- coef_l_model["income"]
print(income_coefficient)
#It is 0.003170909, meaning for each additional unit of income, the prestige score
#is expected to increase by approximately 0.003170909 units, holding all other
#factors constant in the model.
```

(e) Interpret the coefficient for `professional`.

```r
#I do the same for professional.
professional_coefficient <- coef_l_model["professional"]
print(professional_coefficient)
#The coefficient is 37.78128, meaning being in a professional occupation as opposed
#to being blue-collar or white-collar s associated with an increase of approximately
#37.78 units in the prestige score, holding all other factors constant in the model.
```

(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).

```
#In order to use my equation from c, I will calculate two separate predictions
#that have a $1000 increase between them for professionals. Then I'd subtract them
#to find the change in prestige:
predicted_prestige1 = prediction_equation(5000, 1)
predicted_prestige2 = prediction_equation(6000, 1)
change_in_prestige = predicted_prestige2 - predicted_prestige1
print(change_in_prestige)
#an income increase of $1,000 for professional occupations increase the prestige
#score by 0.8452
```

(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).

```
#I need to calculate the predicted prestige for both scenarios (non-professional
#and professional) at the specified income level and then find the difference.
# I calculate the predicted prestige for a non-professional occupation at $6,000
predicted_prestige_nonpf = prediction_equation(6000, 0)

# I do the same for professionals at $6,000
predicted_prestige_pf = prediction_equation(6000, 1)

# I calculate the change in prestige due to changing occupation from nonpf to pf
change_in_prestige2 = predicted_prestige_pf - predicted_prestige_nonpf

# Print the change in prestige
print(change_in_prestige2)
#It is 23.82701. In other words, changing one's occupation from nonpf to pf when
#their income is $6000 increases that person"s prestige points by 23.82703
```

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 $\alpha = .05$).

```r
#I start by adding the coefficient and standard error for lawn signs from the table
#because I'd need to performa t-test to see if the coefficient is significantly
#different from zero to conduct the hypothesis test. This way I can tell if the
#signs have an effect. My null hypothesis is the true coefficient equals to zero,
#meaning the signs do not have an effect on voting. My alternative hypothesis is
#that it #doesn't equal to zero and they do not lack an effect.
coef_lawn_signs <- 0.042
```

[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.

```
se_lawn_signs <- 0.016
#t value is the coefficients divided by standard error
t_value_lawn_signs <- coef_lawn_signs / se_lawn_signs
#I calculate the degrees of freedom
df <- 131 - 3 #total parameters are 131 and we have 3 parameters
#Since I do not have a specific direction of effect regarding sign's effectiveness,
#(we do not know if it affects positively or negatively), I make a two-tailed
#t test with 95% confidence. I use 0.975 because it is a two-tailed test.
critical_t <- qt(0.975, df)
#I check if the absolute t-value is greater than the critical t-value
abs(t_value_lawn_signs) > critical_t
#Yes, it is greater than the critical t-value, meaning I can reject the null
#hypothesis. In other words, the signs do not lack an effect on voting. Having a
#lawn sign have statistically significant effect on the vote share for Cuccinelli.
```

(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 $\alpha = .05$).

```
#I do the same as a for b since the only difference is being adjacent to a lawn
#sign as opposed to having it on your own lawn.
coef_adjacent_signs <- 0.042
se_adjacent_signs <- 0.013
#I calculate the t-value
t_value_adjacent_signs <- coef_adjacent_signs / se_adjacent_signs
#again, I do a two-tailed test and I already have the degrees of freedom
critical_t_adjacent <- qt(0.975, df)
abs(t_value_adjacent_signs) > critical_t_adjacent
#It is TRUE, meaning I can reject the null hypothesis. In other words, being
#adjacent to a lawn sign have statistically significant effect on the vote share
#for Cuccinelli.
```

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

The coefficient is 0.302. It represents the expected baseline support for Cuccinelli when there are no lawn signs (when all independent variables are zero) In other words Cucccinelli is expected to have 30.2any lawn signs per this model.

(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?

$\mathbf{R^2}$ is 0.094. It means that 9.4 percent of the variability in Cuccinelli's vote share is explained by the model(yard signs and adjacency of yard signs). Yard signs have a statistically significant effect but their impact on the overall vote share is limited when compared to all factors that could potentially influence the outcome. There are likely many other factors like campaigns, spending, candidate performance, conditions related to issues associated with candidates, etc. The statistical significance of the yard signs does not necessarily mean a significant effect on the wider vote share.