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
library(car)

# Load the Prestige dataset
data(Prestige)


# Recode 'type' variable to create 'professional' variable
Prestige$professional <- ifelse(Prestige$type == "prof", 1, 0)

# Check the new 'professional' variable
str(Prestige)
```

```
## 'data.frame':    102 obs. of  7 variables:
##  $ education   : num  13.1 12.3 12.8 11.4 14.6 ...
##  $ income      : int  12351 25879 9271 8865 8403 11030 8258 14163 11377 11023 ...
##  $ women       : num  11.16 4.02 15.7 9.11 11.68 ...
##  $ prestige    : num  68.8 69.1 63.4 56.8 73.5 77.6 72.6 78.1 73.1 68.8 ...
##  $ census      : int  1113 1130 1171 1175 2111 2113 2133 2141 2143 2153 ...
##  $ type        : Factor w/ 3 levels "bc","prof","wc": 2 2 2 2 2 2 2 2 2 2 ...
##  $ professional: num  1 1 1 1 1 1 1 1 1 1 ...
```

(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
#use the function 'lm' to run a linear model

model <- lm(prestige ~ income * professional, data = Prestige)

#use the function 'summary' to check the main character of the linear
    model

summary(model)
```

```
##
## 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
##   (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.

```
#use the information we got from the summary to write the prediction
    equatation
\[ \text{Prestige} = 21.14 + 0.0032 \times \text{Income} + 37.78 \times \
    text{Professional} - 0.0023 \times \text{Income} \times \text{
    Professional} \]

#explaination of the variables
Where:

- \( \text{Prestige} \) is the predicted prestige level.
- \( \text{Income} \) represents the individual's income.
- \( \text{Professional} \) is a binary variable (1 if professional, 0
    otherwise).
```

$$\text{Prestige} = 21.14 + 0.0032 \times \text{Income} + 37.78 \times \text{Professional} - 0.0023 \times \text{Income} \times \text{Professional}$$

(d) Interpret the coefficient for `income`

```
The coefficient for income in the context of this regression model is \(0.0032\). This coefficient indicates the expected change in the prestige level for a one-unit increase in income, holding other variables constant.

Specifically, for every one-unit increase in income, there is an estimated increase of \(0.0032\) units in the prestige level, assuming all other variables (including professional status and the interaction term) remain constant.

As income increases, the level of personal professional reputation often increases accordingly
```

(e) Interpret the coefficient for `professional`.

```
In this regression model, the coefficient for the variable 'professional' is \(37.78\).

This coefficient signifies the average difference in the prestige level between individuals categorized as professionals (coded as 1) and those categorized as non-professionals (coded as 0), while holding other variables constant.

on average, individuals classified as professionals have a prestige level approximately \(37.78\) units higher than individuals categorized as non-professionals.
```

(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).

```
From the regression equation:
```

$$\text{Prestige} = 21.14 + 0.0032 \times \text{Income} + 37.78 \times \text{Professional} - 0.0023 \times \text{Income} \times \text{Professional}$$

```
The effect of a $1,000 increase in income specifically for professional
    occupations (where 'Professional' = 1) can be found by considering the
     coefficient of the income-professional interaction term, which is
    \(-0.0023\).

The change in the predicted prestige score (\( \hat{y} \)) associated
    with a $1,000 increase in income for professional occupations is
    calculated by multiplying the income change by the interaction
    coefficient:
```

$$\text{Change in } \hat{y} = \text{Change in Income} \times \text{Interaction Coefficient}$$

```
Given a $1,000 increase in income, the change in \( \hat{y} \) for
    professional occupations would be:
```

$$\text{Change in } \hat{y} = \$1,000 \times (-0.0023) = -\$2.3$$

```
Therefore, based on the model, for professional occupations, a \$1,000
    increase in income is associated with a decrease of approximately \$
    2.30 in the predicted prestige score.
```

(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).

```
Given the regression equation:
```

$$\text{Prestige} = 21.14 + 0.0032 \times \text{Income} + 37.78 \times \text{Professional} - 0.0023 \times \text{Income} \times \text{Professional}$$

```
We're interested in the marginal effect of professional jobs when income is $6,000. So, let's compute the change in \( \hat{y} \) associated with the transition from non-professional (Professional = 0) to professional (Professional = 1) specifically at an income level of $6,000.

At \( \text{Income} = $6,000 \) and transitioning from non-professional to professional (0 to 1), the change in \( \hat{y} \) is computed by adding the coefficient of the 'Professional' variable to the interaction term's coefficient:
```

$$\text{Change in } \hat{y} = \text{Professional Coefficient} + \text{Interaction Coefficient} \times \text{Income}$$

```
Given:
- Professional Coefficient: \(37.78\)
- Interaction Coefficient: \(-0.0023\)
- Income: \( $6,000 \)

The change in \( \hat{y} \) associated with this transition:
```

$$\text{Change in } \hat{y} = 37.78 + (-0.0023) \times 6000$$

$$\text{Change in } \hat{y} = 37.78 - 13.8$$

$$\text{Change in } \hat{y} = 23.98$$

```
Therefore, transitioning from a non-professional occupation to a professional one at an income level of $6,000 is associated with an increase of approximately 23.98 units in the predicted prestige score (\( \hat{y} \)).
```

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 null hypothesis (\(H_0\)) would be that having yard signs (either
    assigned or adjacent) does not affect the vote share. The alternative
    hypothesis (\(H_1\)) would be that having yard signs does have an
    effect on vote share.

For each coefficient (assigned lawn signs and adjacent to lawn signs), we
    can perform a hypothesis test by checking if the coefficient is
    significantly different from zero at a significance level of \( \alpha
    = 0.05 \).

The hypotheses can be formulated as:
```

[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.

- Assigned lawn signs:
  - H_0: Coefficient for assigned lawn signs (β_{assigned}) = 0
  - H_1: Coefficient for assigned lawn signs (β_{assigned}) ≠ 0
- Adjacent to lawn signs:
  - H_0: Coefficient for adjacent to lawn signs (β_{adjacent}) = 0
  - H_1: Coefficient for adjacent to lawn signs (β_{adjacent}) ≠ 0

$$t = \frac{\text{Coefficient}}{\text{Standard Error}}$$

```
Then compare the obtained t-values with critical values from a t-
    distribution with \(N-3\) degrees of freedom (where \(N\) is the
    sample size) at the \( \alpha = 0.05 \) significance level to
    determine if the coefficients are statistically significant.
```

```
## [1] 0.00972002
```

```
## [1] 0.00156946
```

```
coefficient_assigned <- 0.042
se_assigned <- 0.016

coefficient_adjacent <- 0.042
se_adjacent <- 0.013

N <- 131

t_assigned <- coefficient_assigned / se_assigned
t_adjacent <- coefficient_adjacent / se_adjacent

df <- N - 3

p_value_assigned <- 2 * pt(abs(t_assigned), df, lower.tail = FALSE)
p_value_assigned

p_value_adjacent <- 2 * pt(abs(t_adjacent), df, lower.tail = FALSE)
p_value_adjacent
```

```
Both p-values are less than the significance level of \( \alpha = 0.05 \)
    . Both having assigned lawn signs and being adjacent to precincts with
     signs have a statistically significant effect on vote share in the
    precincts studied.
```

(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 $\alpha = .05$).

```
affects vote share (e.g., conduct a hypothesis test with $\alpha = 0.05$



- Coefficient for "Adjacent to lawn signs": \( p = 0.0016 \) (less than \( \alpha = 0.05 \))

Since the p-value for the coefficient representing being adjacent to precincts with yard signs is less than the significance level (\( \alpha = 0.05 \)), we reject the null hypothesis. Therefore, being next to precincts with these yard signs does have a statistically significant effect on vote share in the studied precincts.
```

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

```
In regression analysis, the constant term (also known as the intercept) represents the predicted value of the dependent variable when all independent variables in the model are zero.

In this specific context:

- Constant coefficient: \(0.302\)


When both the variables "Assigned lawn signs" and "Adjacent to lawn signs" are zero (meaning no signs were assigned or adjacent), the predicted proportion of the vote that went to McAuliffe's opponent Ken Cuccinelli 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 can be assessed using the \( R^2 \) value, which measures the proportion of variance in the dependent variable (vote share in this case) explained by the independent variables in the model.

In this context:

- \( R^2 = 0.094 \)

This \( R^2 \) value of \(0.094\) indicates that approximately \(9.4\%\) of the variation in the vote share for McAuliffe's opponent Ken Cuccinelli is explained by the variables included in the model (assigned lawn signs, adjacent to lawn signs).

Interpreting this \( R^2 \) value in the context of yard signs and other unmodeled factors:

- The model, with only information about assigned and adjacent yard signs, explains a relatively small proportion (\(9.4\%\)) of the variability in vote share.
- This suggests that factors beyond the presence or adjacency to yard signs significantly influence voting preferences. Unmodeled factors such as candidate characteristics, campaign messaging, voter demographics, and other contextual elements likely play substantial roles in determining voting preferences.
- The importance of yard signs, as indicated by the model, seems relatively modest in explaining the variation in vote share when compared to other unaccounted factors.
```