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`).

```
# Load the necessary libraries and data sets
install.packages("car")
library(car)
data(Prestige)
help(Prestige)

# inspect data through summary
summary(Prestige)

#     education          income          women           prestige
    census       type
#  Min.   : 6.380   Min.   :  611   Min.   : 0.000   Min.   :14.80   Min.
    :1113   bc  :44
#  1st Qu.: 8.445   1st Qu.: 4106   1st Qu.: 3.592   1st Qu.:35.23   1st
   Qu.:3120   prof:31
#  Median :10.540   Median : 5930   Median :13.600   Median :43.60
   Median :5135   wc  :23
#  Mean   :10.738   Mean   : 6798   Mean   :28.979   Mean   :46.83   Mean
    :5402   NA's: 4
#  3rd Qu.:12.648   3rd Qu.: 8187   3rd Qu.:52.203   3rd Qu.:59.27   3rd
   Qu.:8312
#  Max.   :15.970   Max.   :25879   Max.   :97.510   Max.   :87.20   Max.
    :9517


# View the first few lines of the Prestige datasets
head(Prestige)

#                     education income women prestige census type
#  gov.administrators      13.11  12351 11.16     68.8   1113 prof
#  general.managers        12.26  25879  4.02     69.1   1130 prof
#  accountants             12.77   9271 15.70     63.4   1171 prof
#  purchasing.officers     11.42   8865  9.11     56.8   1175 prof
#  chemists                14.62   8403 11.68     73.5   2111 prof
#  physicists              15.64  11030  5.13     77.6   2113 prof


# Create a new variable professional by recoding the variable type,
# professionals are coded as 1, blue and white collar workers are coded
   as 0.
Prestige$professional <- ifelse(Prestige$type == "prof", 1, 0)
Prestige$professional

#  [1]  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
    1  1  1  1  1  0  1  1  0  1  0 NA  0  0
# [37]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 NA  0  0  0  0  0
    0  0  0  0 NA  0  0  0 NA  0  0  0  0  0
# [73]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
    0  1  0  0  0  0  0  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.)

```
# Running linear model
model <- lm(prestige ~ income * professional, data = Prestige)
# inspect data through summary
summary(model)
# get summary of model with (coefficient estimates).

# 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.

```
# Prediction equation
# prestige = intercept + 1 * income + 2 * professional + 3 * (income * professional)

# Model Coefficients:
# Intercept: 21.1422589
#  1 (income coefficient): 0.0031709
#  2 (professional coefficient): 37.7812800
#  3 (income * professional coefficient): -0.0023257

# Interpretation:
# The prestige of an entity is predicted by the sum of the intercept and the product of the respective coefficients
# and predictor variables (income and professional). The equation is as follows:
prestige = 21.1422589 + (0.0031709 * income) + (37.7812800 * professional) + (-0.0023257 * (income * professional))

# Coefficient Explanations:
# Intercept: Represents the baseline prestige when both income and professional are zero.
#  1 (income coefficient): Indicates the change in prestige for a one-unit increase in income, holding other variables constant.
#  2 (professional coefficient): Represents the change in prestige for a one-unit increase in professional status, holding other variables constant.
#  3 (income * professional coefficient): Signifies the interaction effect of income and professional status on prestige.
```

For the question about whether individuals with higher income have more prestigious jobs:

Answer: Yes, there is a significant positive relationship between individual income levels and their prestige. According to the predictive equation, holding other factors constant, for each additional unit of income, the average prestige increases by 0.0031709 units. This suggests that individuals with higher incomes are more likely to have more prestigious jobs.

For the question about whether professionals have more prestigious jobs than blue and white-collar workers:

Answer: Yes, professionals have a higher level of prestige compared to non-professionals. According to the predictive equation, this difference persists even when controlling for individual income and the interaction term. On average, professionals have a prestige level that is 37.78 units higher than non-professionals.

(d) Interpret the coefficient for `income`.

The coefficient of income is 0.0031709, which means that when other variables remain unchanged, for each unit of income increase, prestige's estimate increases by 0.0031709.

Since the coefficient is positive, it can be explained that the increase of income is positively correlated with the increase of prestige. That is, as an individual's income increases, their professional prestige also tends to increase. Indicates that an individuals income is positively correlated with his professional reputation.

(e) Interpret the coefficient for `professional`.

In this model, professional is a dummy variable, a value of 1 indicates that the individual belongs to a professional occupation, and a value of 0 indicates that the individual belongs to a non-professional occupation.

The coefficient 37.7812800 indicates that, relative to non-professional occupation, when an individual changes from non-professional occupation to professional occupation, average reputation increased by 37.7812800 units. The average effect when other explanatory variables are held constant.

In this model, professional occupations are positively associated with higher average prestige relative to non-professional occupations.

(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).

We are concerned with the marginal effect of income on prestige scores for professional occupations. The marginal effect can be obtained by solving the prediction equation.

```
# Our prediction equation is:
prestige = 21.1422589 + 0.0031709 * income + 37.7812800 * professional +
    (-0.0023257) * (income * professional)

prestige(professional=1) =
  21.1422589 + 0.0031709 * income + 37.7812800 * 1 + (-0.0023257) * (
    income)

the prestige changes(for example: income from 0$ to 1000$) with each
    increase in income of $1,000:
 y 0 prestige (when income is 0)   =
  21.1422589 + 0.0031709 * 0 + 37.7812800 * 1 + (-0.0023257) * 0
  = 58.92354

 y 1 0 0 0 prestige (when income increase 1000$)   =
  21.1422589 + 0.0031709 * 1000 + 37.7812800 * 1 + (-0.0023257) * 1000
  = 59.76874

y (change: y1000 - y0) =  59.76874 - 58.92354 = 0.8452
```

The effect of a $1,000 increase in income on the prestige score for professional occupations is that, when income increases by $1,000, the prestige score is expected to increase by 0.8452 units.

In other words, the marginal effect of income indicates that the prestige score of professional occupations is expected to increase by approximately 0.8452 units when income increases by $1,000.

This suggests that for every $1,000 increase in income, the expected increase in prestige for professional occupations is around 0.8452 units.

(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).

Our prediction equation is:

```
prestige = 21.1422589 + 0.0031709 * income + 37.7812800 * professional +
    (-0.0023257) * (income * professional)
prestige(income = 6000) = 21.1422589 + 0.0031709 * income + 37.7812800 *
    professional + (-0.0023257) * (6000 * professional)

y (from non-professional to professional) = prestige(professional=1) -
    prestige(professional=0)

y  = (21.1422589 + 0.0031709 * 6000 + 37.7812800 * 1 + (-0.0023257) *
    (6000 * 1)) - (21.1422589 + 0.0031709 * 6000)
   = 37.7812800 + (-0.0023257) * 6000
   = 23.8270800
```

The effect of transitioning from a non-professional to a professional occupation, with an income of \$6,000, is an expected increase in the prestige score by approximately 23.8270800 units.

The marginal effect of professional occupations, with the income variable set at \$6,000, is expected to result in an increase in prestige by around 23.8270800 units.

This suggests that transitioning from a non-professional to a professional occupation, with an income of \$6,000, is expected to have the effect of increasing the prestige for professional occupations by around 23.8270800 units.

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

[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.

(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 $\alpha = .05$).

Steps 1, Establish null hypothesis and alternative hypothesis
H0: these yard signs have no a precinct affects vote sharethe coefficient "Precinct assigned lawn signs" is 0.
H1: these yard signs in a precinct affects vote sharethe coefficient "Precinct assigned lawn signs" is not equal 0.

```
# Steps 2   Its    a two-tailed test

# Steps 3  Determine   significance  level  = .05
alpha <- 0.05

# Steps 4   calculate   t statistic
# Extract the coefficient and standard error from the output
coefficient <- 0.042
standard_error <- 0.016
# Calculate the t statistic
t_statistic <- coefficient / standard_error
t_statistic # 2.625

# Steps 5   calculate   df
# parameters are 3   "Precinct assigned lawn signs"   "Precinct adjacent
    to lawn signs"    Constant
N <- 131
k <- 3
df <- N - k
df # 131 - 3 = 128

# Steps 6   calculate   critical_value
critical_value <- qt(1 - alpha/2, df)
critical_value # 1.978671

# Steps 7  compare   t statistic and critical_value
t_statistic > critical_value # true


# double check answer, use the p-value
# Get the p-value from t-distribution
p_value <- 2 * (1 - pt(abs(t_statistic), df))
p_value # 0.00972002
p_value < alpha # true
```

We find sufficient evidence to reject the null hypothesis (H0) that these yard signs have no effect on precinct vote share, as the coefficient for 'Precinct assigned lawn signs' is statistically different from 0.
Therefore, we support the alternative hypothesis (H1) that these yard signs in a precinct do affect vote share, with the coefficient for 'Precinct assigned lawn signs' being significantly non-zero.

(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 $\alpha = .05$).

Steps 1, Establish null hypothesis and alternative hypothesis
H0: The coefficient of "Precinct adjacent to lawn signs" is equal to 0, indicating that these adjacent signs have no effect on vote share.
H1: The coefficient of "Precinct adjacent to lawn signs" is not equal to 0, indicating that these adjacent signs have an effect on vote share.

```
# Steps 2   Its    a two-tailed test

# Steps 3  Determine   significance  level  = .05
alpha <- 0.05

# Steps 4   calculate    t statistic
# Calculate the t statistic
coefficient <- 0.042
standard_error <- 0.013
t_statistic <- coefficient / standard_error
# Extract the coefficient and standard error from the output
t_statistic  # 3.230769

# Steps 5   calculate    df
# parameters are 3   "Precinct assigned lawn signs"   "Precinct adjacent
    to lawn signs"    Constant
N <- 131
k <- 3
df <- N - k
df # 131 - 3 = 128

# Steps 6   calculate    critical_value
critical_value <- qt(1 - alpha/2, df)
critical_value # 1.978671

# Steps 7  compare   t statistic and critical_value
t_statistic > critical_value # true


# double check answer, use the p-value
# Get the p-value from t-distribution
p_value <- 2 * (1 - pt(abs(t_statistic), df))
p_value # 0.00156946
p_value < alpha # true
```

We find sufficient evidence to reject the null hypothesis (H0) that the coefficient of 'Precinct adjacent to lawn signs' is equal to 0, suggesting that these adjacent signs have no effect on vote share. Therefore, we support the alternative hypothesis (H1) that the coefficient of 'Precinct adjacent to lawn signs' is not equal to 0, indicating that these adjacent signs have a significant effect on vote share.

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

Constant Coefficient (0.302): When all other independent variables (Precinct assigned lawn signs and Precinct adjacent to lawn signs) are zero, the predicted average value of vote share is 0.302.

In this context, the constant term represents the intercept or baseline level of vote share when the impact of other variables is not considered.

(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?

```
vote share = 0.302 + 0.042    Precinct assigned lawn signs + 0.042
   Precinct adjacent to lawn signs

This is a multiple linear regression model:
  "vote share" is the dependent variable
  "Precinct assigned lawn signs"     "Precinct adjacent to lawn signs"
    are independent variables
  The intercept is 0.302
  A coefficient of 0.042 indicates the impact of each unit change on the
    "vote share"
```

1. An R-squared value of 0.094 indicates that the model is able to explain approximately 9.4This implies that yard signs, as predictive factors, contribute to explaining a certain degree of variability in vote share.

2. Without a specific dataset, the calculation of the F-statistic for an overall model significance test is not possible. Therefore, I believe that when evaluating the model, the magnitude of R-squared alone cannot determine the model's goodness of fit. A smaller R-squared does not necessarily imply model ineffectiveness, and a larger R-squared does not guarantee a well-fitting model.

3. Within the given precincts, yard signs significantly impact vote share, whether placed within a precinct or in an adjacent precinct.

4. The importance of yard signs relative to unmodeled factors needs to be considered, and a more comprehensive assessment may require the inclusion of other predictor variables or comparisons with alternative models. The model's integrity and explanatory power may be influenced by unaccounted factors.