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`).

```
# Question 1: Economics

# remove objects
rm(list=ls())

# set wd for current folder
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
library(ggplot2)
library(tidyr)
install.packages("car")
library("car")
data(Prestige)
help(Prestige)
View(Prestige)

# (a)
# Create a new variable professional by using ifelse
Prestige$professional <- ifelse(Prestige$type == "prof", 1,
                                ifelse(Prestige$type %in% c("bc","wc"),0,
    NA))
```

(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.)

```
# (b)
# The regression equation is
# y_prestige = alpha + beta_1*income + beta_2*professional + beta_3*
    income*professional

# Run regression model with interaction term and the lm function can
    automatically cleans NA
lm <- lm(prestige ~ income + professional + income * professional, data =
     Prestige)

# Print the results
summary(lm)

# Locate the coefficients
coefficients <- coefficients(lm)
alpha <- coefficients[1]
beta_1 <- coefficients[2]
beta_2 <- coefficients[3]
beta_3 <- coefficients[4]

```

```r
# Print the coefficients
print(alpha)
print(beta_1)
print(beta_2)
print(beta_3)
```

```
Call:
lm(formula = prestige ~ income + professional + 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
```

Interpretation:
Check the significance of coefficients: from the summary of the regression model, the P-values of the coefficients are both less than 0.05 , so the coefficients are statistically significant.

So, The regression model is:
y-hat-prestige = 21.14226 + 0.0032*income + 37.7812*professional - 0.0023*income*professional

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

Interpretation:
Prediction equation for professionals (professional = 1):

y-hat = 21.1422 + 0.0032*income + 37.7812*professional - 0.0023*income*professional
y-hat-1 = 21.1422 + 0.0032*income + 37.7812*1 - 0.0023*income*1

So, the prediction equation for professionals is :
y-hat-1 = 58.9234 + 0.0009*income

Prediction equation for non-professionals/Blue Collar and white Collar(professional = 0):

y-hat = 21.1422 + 0.0032*income + 37.7812*professional - 0.0023*income*professional
y-hat-0 = 21.1422 + 0.0032*income + 37.7812*0 - 0.0023*income*0

So, the prediction equation for non-professionals is :
y-hat-0 = 21.1422 + 0.0032*income

(d) Interpret the coefficient for `income`.

Interpretation:
Income effect for professionals:
For professionals, with every additional 1 USD of income, the average increase of Prestige is 0.0009.

Income effect for non-professionals:
For non-professionals, with every additional 1 USD of income, the average increase of Prestige is 0.0032.

(e) Interpret the coefficient for `professional`.

Interpretation:
Rearrange the regression model:
y-hat = 21.1422 + 37.7812*professional +(0.0032-0.0023*professional)*income

When the income of professionals and non-professionals is both zero, the average increase in prestige for professionals is 37.7812 higher than 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).

```r
# (f)
# Calculate the marginal effect by using the prediction equation for professionals
# And do not need to add the intercept term
income_1000 <- 1000
delta_y_hat <- 0.0009*income_1000

# Print the result
print(delta_y_hat)

# delta_y_hat = 0.9
```

Interpretation:
So, the marginal effect in y-hat associated with a 1,000 dollars increase in income is almost 0.9.

(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).

```r
# (g)
# Calculate the y_hat when income is 6000 by using the prediction equation for non_professionals
income_6000 <- 6000
y_hat_0_6000 <- 21.1422 + 0.0031*income_6000
print(y_hat_0_6000)
# y_hat_0_6000 = 39.7422

# Calculate the y_hat when income is 6000 by using the prediction equation for professionals
y_hat_1_6000 <- 58.9234 + 0.0009*income_6000
print(y_hat_1_6000)
# y_hat_1_6000 = 64.3234

# Calculate the gap
y_hat_gap <- y_hat_1_6000 - y_hat_0_6000

# Print the result
print(y_hat_gap)
```

```
18 # y_hat_gap = 24.5812
```

Interpretation:
So, the marginal effect in y-hat is almost 24.5812.

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 $\alpha = .05$).

From the table above, the regression equation is:
y-voteshare = 0.302(alpha) + 0.042(beta-1)*assigned + 0.042(beta-2)*adjacent

Then, Conduct T-test to test the significance of the coefficient of Precinct assigned lawn signs(beta-1)

```
# Save coefficients
alpha <- 0.0302
beta_1 <- 0.042
beta_2 <- 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.

```r
se_1 <- 0.016
se_2 <- 0.013

# Step1: State a null and alternative (two-tailed) hypothesis:
# H0: beta_1 = 0 VS. HA: beta_1 != 0

# Step2: Check the standard error
se_1 <- 0.016

# Step3: calculate Test statistic:
t_statistic <- (beta_1-0)/se_1
print(t_statistic)

# Step4: calculate degree of freedom
df <- 131-2-1
print(df)

# Step5:calculate p-value when two tailed
p_value <- 2 * (1 - pt(abs(t_statistic), df))
print(p_value)
```

Interpretation:
Step6: draw the conclusion:
P value is almost 0.01 which is less than 0.05, so, we have sufficient evidence to reject H0(beta-1 = 0). So, beta-1 is statistically significant, we can conclude that having these yard signs in a precinct affects voteshare.

(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 $\alpha = .05$).

```r
# (b)
# Conduct T-test to test the significance of the coefficient of Precinct
    adjacent to lawn signs(beta_2)

# Step1: State a null and alternative (two-tailed) hypothesis.
# H0: beta_2 = 0 VS. HA: beta_2 != 0

# Step2: Check the standard error
se_2 <- 0.013

# Step3: calculate Test statistic:
t_statistic_2 <- (beta_2-0)/se_2
print(t_statistic_2)

# Step4: calculate degree of freedom
df <- 131-2-1
print(df)

# Step5:calculate p-value when two tailed
p_value_2 <- 2 * (1 - pt(abs(t_statistic_2), df))
print(p_value_2)
```

Interpretation:
Step6: draw the conclusion:
P value is almost 0.00 which is less than 0.05, so, we have sufficient evidence to reject H0(beta-2 = 0).So, beta-2 is statistically significant, We can conclude that being next to precincts with these yard signs affects voteshare.

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

Interpretation:
The coefficient for the constant term means: the baseline of the impact on voteshare is 0.302 when the precinct is NOT assigned lawn signs and NOT adjacent to lawn 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?

Interpretation:
We can evaluate the model fit for the regression by R squared. The closer the R-squared is to 1, the better the model fits the data; if it is closer to 0, the model is less

able to explain the data.

From the table, the R squared is 0.094 which is almost zero.

That means approximately 9.4 percent of the variation in the dependent variable can be explained by the independent variables in the model. The remaining 90.6 percent of the variation may be caused by other factors not considered in the model.