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`).

The structure of the syntax `ifelse` is:

```
ifelse(test, yes, no)
```

In this structure, we have:

- test: is a logical condition. If it evaluates to TRUE, the function returns the corresponding value from yes; otherwise, it returns the value from no.
- yes: if "test" is TRUE, then operate it.
- no: if "test" is FALSE, then operate it.

So I choose to write my R code in this style:

```r
####### (a) #######

# Create variable professional
Prestige$professional <- ifelse(Prestige$type == "prof", 1,
  ifelse(Prestige$type %in% c("wc", "bc"), 0, NA))
summary(Prestige$professional); table(Prestige$professional)
```

In this code, I make a judgment in this logic:

(a) If the type in Prestige is "prof", then recode as 1.

(b) If the type in Prestige is not "prof", then judge (in another `ifelse` structure):

  i. if the type in Prestige included "wc" or "bc", then recode as 0.

  ii. If else, then recode as NA.

And we can see the output and check our answers in R:

```
> summary(Prestige$professional); table(Prestige$professional)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
0.0000  0.0000  0.0000  0.3163  1.0000  1.0000     4

0   1
67  31
```

So we know that there are 67 people were coded as 0, 31 people were coded as 1, and 4 people were coded as 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.)

The basic format of this interaction regression model is:
$prestige = \alpha + \beta_1 income_i + \beta_2 professional_i + \beta_3 income_i * professional_i + \epsilon_i$

So I choose to write my R code in this style:

```
######## (b) ########

# Run a linear regression model, in this model:

# - prestige: the outcome variable
# - income: dependent variable
# - professional: dependent variable
# - incprof: the interaction of income and professional

q1b_reg <- lm(prestige ~ income + professional + income * professional,
              data = Prestige)
summary(q1b_reg)
```

Table 1: **Outcome is prestige, predictors are income, professional and interaction**

	prestige
income	0.003***
	(0.0005)
professional	37.781***
	(4.248)
income:professional	−0.002***
	(0.001)
Constant	21.142***
	(2.804)
Observations	98
R^2	0.787
Adjusted R^2	0.780
Residual Std. Error	8.012 (df = 94)
F Statistic	115.878*** (df = 3; 94)
Note:	*p<0.1; **p<0.05; ***p<0.01

There is a positive and statistically reliable relationship between the income and prestige, the professional and prestige. But there is a negative and statistically reliable relationship between the prestige and the interaction of income and professional.

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

$prestige = 21.142 + 0.003 * income + 37.781 * professional - 0.002 * income * professional$

(d) Interpret the coefficient for `income`.

Because the p-value of the coefficient for `income` is less than 0.001 and is significant (***), so we have enough evidence to say the coefficient is not equal to 0.

- **For non professional people**: Such a 1 unit increase in the income is associated with an average increase of 0.003 in prestige.
- **For professional people**: Such a 1 unit increase in the income is associated with an average increase of $(0.003 - 0.002 = -0.001)$ in prestige.

(e) Interpret the coefficient for `professional`.

Because the p-value of the coefficient for `professional` is less than 0.001 and is significant (***), so we have enough evidence to say the coefficient is not equal to 0.

- **For non professional people**: The prestige is $21.142 + 0.003 * income$
- **For professional people**: Compared with those who are not professional, professional people is associated with an average increase of $(37.781 - 0.002 = 37.779)$ in prestige.

(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).

According to the question, we know: `professional` = 1, `income` increase = 1000
So we can have the formula:
$prestige_1 = 21.142 + 0.003 * income + 37.781 * professional - 0.002 * income * professional$ .....(1)
$prestige_2 = 21.142 + 0.003 * (income + 1000) + 37.781 * professional - 0.002 * (income + 1000) * professional$ .....(2)

$\hat{y} = (2) - (1) = prestige_2 - prestige_1 =$
$0.003 * (income + 1000) - 0.003 * income - [-0.002 * (income + 1000) * professional - (-0.002 * income * professional)] =$
$3 - 2 * professional = 3 - 2 = 1$

We can conclude that: For `professional` people, each \$1,000 `income` increase will raise 1 prestige in average.

(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).

According to the question, we know: `income` = 6000
So we can have the formula:
$prestige_1 = 21.142 + 0.003 * income + 37.781 * professional - 0.002 * income * professional$ .....(1)
$prestige_2 = 21.142 + 0.003 * income + 37.781 * professional - 0.002 * income * professional$ .....(2)

And for formula (1), `professional` = 0, for formula (2), `professional` = 1
$prestige_1 = 21.142 + 0.003 * 6000 = 39.142$
$prestige_2 = 21.142 + 0.003 * 6000 + 37.781 * 1 - 0.002 * 6000 * 1 = 64.923$

$\hat{y} = (2) - (1) = prestige_2 - prestige_1 = 64.923 - 39.142 = 25.781$

We can conclude that: If someone `income` is \$6000, and change from non-professional to professional, this person will have a 25.781 increase in prestige in average.

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 $\alpha = .05$).

According to the question, we need to detect the coefficient of `Precinct assigned lawn signs`, which has a coefficient $\beta_1 = 0.042$, and the standard error of this coefficient is 0.016. So let's make hypothesis:

H_0: Have yard signs has no affect on vote share. ($\beta_1 = 0$)
H_α: Have yard signs has a significant affect on vote share. ($\beta_1 \neq 0$)

We can calculate t-value in this formula: $t - value = \frac{CoefficientEst.}{Std.Er.} = \frac{0.042}{0.016} = 2.625$
Get p-value in R, and we can get R output:

```
> q2a_pvalue <- 1 - pt(2.625, df = 128); print(q2a_pvalue)
[1] 0.00486001
```

We can see that the p-value (≈ 0.005) is below the $\alpha = 0.05$ threshold, so we would say we have found sufficient evidence to reject the null hypothesis that have yard signs has no affect on vote share and accept alternative hypothesis.($\beta_1 \neq 0$).

(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 $\alpha = .05$).

According to the question, we need to detect the coefficient of `Precinct assigned lawn signs`, which has a coefficient $\beta_1 = 0.042$, and the standard error of this coefficient is 0.013. So let's make hypothesis:

H_0: Being next to precincts with yard signs has no affect on vote share. ($\beta_2 = 0$)
H_α: Being next to precincts with yard signs has a significant affect on vote share. ($\beta_2 \neq 0$)

We can calculate t-value in this formula: $t - value = \frac{CoefficientEst.}{Std.Er.} = \frac{0.042}{0.013} \approx 3.231$
Get p-value in R, and we can get R output:

```
> q2b_pvalue <- 1 - pt(3.231, df = 128); print(q2b_pvalue)
[1] 0.0007841451
```

We can see that the p-value (≈ 0.001) is below the $\alpha = 0.05$ threshold, so we would say we have found sufficient evidence to reject the null hypothesis that have yard signs has no affect on vote share and accept alternative hypothesis.($\beta_2 \neq 0$).

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

In a multiple regression, the basic formula is:
$\hat{y} = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon_i$
In this case, the model is:
$voteshare = constant + \beta_1 * assignedsigns + \beta_2 * adjacentsigns$
So we can write the formula:
$voteshare = 0.302 + 0.042 * assignedsigns + 0.042 * adjacentsigns$

The constant means when independent variables observation is 0, the dependent variable value will be the constant. In tables:

Table 2: **Constant meaning explanation**

		If assigned lawn signs, voteshare = yes	If assigned lawn signs, voteshare = no
If adjacent to lawn signs, voteshare =	yes	$constant + \beta_1 + \beta_2 = 0.386$	$constant + \beta_2 = 0.344$
	no	$constant + \beta_1 = 0.344$	$constant = 0.302$

In words:

- **When people not assigned or adjacent to lawn signs**: The voteshare value will be *constant* itself (0.302).
- **When people was assigned but not adjacent to lawn signs**: The voteshare value will be $constant + \beta_1 = 0.302 + 0.042 = 0.344$.
- **When people was not assigned but adjacent to lawn signs**: The voteshare value will be $constant + \beta_2 = 0.302 + 0.042 = 0.344$.
- **When people was assigned and adjacent to lawn signs**: The voteshare value will be $constant + \beta_1 + \beta_2 = 0.302 + 0.042 + 0.042 = 0.386$

(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?

According to the model output table, the $R^2 = 0.094$. This means independent variables in the model can explain approximately 9.4% of the variability in the dependent variable. And This suggests a relatively low proportion of variance explained. I think this results suggests us the factors beyond yard signs and adjacency may contribute to voting preferences. We can consider to add other relevant variables to improve the model.