Problem Set 4

Applied Stats/Quant Methods 1

Athena Rodrigues

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 Monday November 18, 2024. 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.

```r
#initial setup
library(car)
data(Prestige) #prestige of Canadian occupations
help(Prestige)
occupations <- Prestige #change name for clarity
```

(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
#part a professional = 1, else = 0
occupations$professional <- as.factor(ifelse(occupations$type == "prof",
    1, 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.)

```r
#part b use * because continuous X dummy
parta_regression <- lm(occupations$prestige ~ occupations$income*
    occupations$professional)
summary(parta_regression)
```

Figure 1: Linear Regression

```
Call:
lm(formula = occupations$prestige ~ occupations$income * occupations$professional)

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 ***
occupations$income                            0.0031709  0.0004993   6.351 7.55e-09 ***
occupations$professional1                    37.7812800  4.2482744   8.893 4.14e-14 ***
occupations$income:occupations$professional1 -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.

```
#part c
intercept <- round(parta_regression$coefficients[1],3)
income_coef <- round(parta_regression$coefficients[2],3)
professional_coef <- round(parta_regression$coefficients[3],3)
interaction_coef <- round(parta_regression$coefficients[4],3)
cat(intercept, "+", income_coef, "* income", "+", professional_coef, "*
    professional", "+", interaction_coef, "*income*professional")
```

E[Y] = 21.142 + 0.003 * income + 37.781 * professional + -0.002*income*professional

(d) Interpret the coefficient for `income`.

For every one-unit increase in income, a 0.003 increase in prestige score will occur for those who are not considered part of the professional category (professional = 0)

(e) Interpret the coefficient for `professional`.

There is a 37.781 increase in prestige score when there is a change from non-professional to professional occupation (professional changes from 0 to 1), when there is no income.

(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
#part f
intercept <- 21.142
income_coef <- 0.003
income <- 1000
no_income <- 0
prof_coef <- 37.781
professional <- 1
interaction_coef <- -0.002
income_partf <- intercept + income_coef * income + prof_coef *
    professional + interaction_coef * income * professional #59.923
noincome_partf <- intercept + income_coef * no_income + prof_coef *
    professional + interaction_coef * no_income * professional #58.923
change_in_income <- income_partf - noincome_partf #59.923 - 58.923
print(change_in_income)
```

$$\text{E[Y]} = 21.142 + 0.003 * 1000 + 37.781 * 1 + \text{-}0.002*1000*1 = 59.923$$

$$\text{E[Y]} = 21.142 + 0.003 * 0 + 37.781 * 1 + \text{-}0.002*0*1 = 58.923$$

$$\Delta E[y] = 59.923 \text{ - } 58.923 = 1.000$$

According to the predicted equation, if someone in a professional occupation (professional = 1) has an income increase of 1000 their prestige score is predicted to be 59.923. However, if one is considered a professional with no income their prestige score is 58.923. This shows there is only a 1 prestige point difference in a professional job going from an income of 0 to 1000 dollars.

(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
#part g
intercept <- 21.142
income_coef <- 0.003
income <- 6000
prof_coef <- 37.781
professional <- 1
notprofessional <- 0
interaction_coef <- -0.002
professional_partg <- intercept + income_coef * income + prof_coef *
    professional + interaction_coef * income * professional #64.923
notprofessional_partg <- intercept + income_coef * income + prof_coef *
    notprofessional + interaction_coef * income * notprofessional #39.142
change_in_occupation <- professional_partg - notprofessional_partg #
    64.923 - 39.142
print(change_in_occupation)
```

$$\begin{aligned} \text{E[Y]} &= 21.142 + 0.003 * 6000 + 37.781 * 1 + \text{-}0.002*6000*1 \\ &= 64.923 \end{aligned}$$

$$\begin{aligned} \text{E[Y]} &= 21.142 + 0.003 * 6000 + 37.781 * 0 + \text{-}0.002*6000*0 \\ &= 39.142 \end{aligned}$$

$$\Delta E[y] = 64.923 \text{ - } 39.142 = 25.781$$

When a professional occupation has an income of 6000, their prestige score is 64.923 according to the predicted equation, While a non-professional occupation with an income of 6000 only has a 39.142 predicted prestige score. This means that if one is to change from non-professional to professional, holding income constant, there would be a predicted 25.781 increase in prestige score

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 McAuliffe'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 $\alpha = .05$).

Assumptions:

- Population distribution of y is normal for each combination of values
- Standard deviation of the conditional distribution of responses on y is the same at each combination
- Random Sampling

Hypothesis:

Null: lawn signs in precincts do not impact Cuccinelli's vote share

$$H_0 : \beta = 0$$

Alternate: lawn signs in precincts have an impact on Cuccinelli's vote share

$$H_A : \beta \neq 0$$

Test-Statistic:

$$test - stat = \beta / se$$

2.625 = 0.042 / 0.016

```
# finding t-statistic
assigned_coef <- 0.042
assigned_se <- 0.016
assigned_statistic <- assigned_coef / assigned_se #2.625
```

P-Value:

```
# finding p-value
degrees <- 131 - 2 - 1  #N - predictors - 1
assigned_p_value <- 2* pt(t_statistic, df = degrees, lower.tail = FALSE)
    #0.00972
```

Conclusion: Reject the null hypothesis as 0.00972 is less than 0.05; a one-unit increase in lawn signs in assigned precincts does impact Cuccinelli's vote share according to this regression.

(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 $\alpha = .05$). Assumptions:

- Population distribution of y is normal for each combination of values
- Standard deviation of the conditional distribution of responses on y is the same at each combination
- Random Sampling

Hypothesis:

Null: lawn signs in adjacent precincts do not impact Cuccinelli's vote share

$$H_0 : \beta = 0$$

Alternate: lawn signs in adjacent precincts have an impact on Cuccinelli's vote share

$$H_A : \beta \neq 0$$

Test-Statistic:

$$test - stat = \beta / se$$

$$3.231 = 0.042 \,/\, 0.013$$

```
# finding t-statistic
adjacent_coef <- 0.042
adjacent_se <- 0.013
adjacent_t_statistic <- adjacent_coef / adjacent_se #3.231
```

P-Value:

```
# finding p-value
degrees <- 131 - 2 - 1  #N - predictors - 1
adjacent_p_value <- 2* pt(adjacent_t_statistic, df = degrees, lower.tail = FALSE) #0.00157
```

Conclusion: Reject the null hypothesis as 0.00157 is less than 0.05; a one-unit increase in lawn signs in adjacent precincts does impact Cuccinelli's vote share according to this regression.

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

The coefficient for the constant term is 0.302, with a standard error of 0.011. This means that controlling for the two variables, no lawn signs in assigned or adjacent precincts, is expected to result in an average vote share of 0.302 for Cuccinelli.

(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 R^2 value for this regression is very small at 0.094. Yard signs result in an impact that accounts for a little more than 9 percent of the vote share. This means that yard signs do not carry as significant an impact on vote share as other factors may.