Problem Set 4

Linette Lim

Due: December 4, 2022

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 4, 2022. 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`).

We use the formula:

```
Prestige$prof <- ifelse(Prestige$type == "prof", 1, 0)
```

From the summary() function, we know that there are 31 coded as professionals. Using the table() function, we check the output to see if it worked as expected:

```
> table(type = Prestige$type,
+       Professional = Prestige$prof)
      Professional
type    0  1
  bc   44  0
  prof  0 31
  wc   23  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.)

We use the formula:

```
mod <- lm(prestige ~ income + prof + income:prof,
          data = Prestige)
```

We run summary() and get the following regression output:

```
> summary(mod)

Call:
lm(formula = prestige ~ income + prof + income:prof, 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 ***
prof        37.7812800  4.2482744   8.893 4.14e-14 ***
income:prof -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.

```
# Prestige= beta0 +beta1*income +beta2*prof + beta3 *income*prof
#         = 21.14 + 0.003*income + 37.78*prof - 0.002*income*prof

# Prestige(prof)= 21.14 + 0.003*income + 37.78*1 - 0.002*income*1
#               = 58.92 + 0.001*income

# Prestige(non-prof)= 21.14 + 0.003*income + 37.78*0 - 0.002*income*0
#                   = 21.14 + 0.003*income
```

Subbing in random values, it appears that if incomes are high (e.g. 100,000), the prestige score is higher for non-professionals (blue collared and white collared) than for professionals. However, at low levels of income (1000), the prestige score is higher for professionals than for non-professionals.

(d) Interpret the coefficient for `income`.

For every 1000 units (dollars) increase in income, prestige increases by 3.1709 units.

(e) Interpret the coefficient for `professional`.

The coefficient for professional is 37.78. A positive regression coefficient means that prestige is higher for the dummy variable 'professional' than for the reference group (bc and wc combined). The regression coefficient is statistically significant, which means the prestige discrepancy between the two groups is also statistically significant.

(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).

When we plug in income values of 2000 and 1000 to the prediction equation in 1c, we get:

```
Prestige(professionals)_2000 = 58.92 + 0.001*2000 = 60.92
Prestige(professionals)_1000 = 58.92 + 0.001*1000 = 59.92
Prestige(professionals)_2000 - Prestige(professionals)_1000 = 60.92-59.92 = 1
```

For professional occupations, a $1,000$ increase in income increases prestige score by 1 unit.

(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).

When we plug in income value of 6000 to the prediction equation in 1c, we get:

```
Prestige(professionals)     = 58.92 + 0.001*6000 = 64.92
Prestige(non-professionals) = 21.14 + 0.003*6000 = 39.14
Prestige(professionals) - Prestige(non-professionals) = 64.92 - 39.14 = 25.78
```

Holding income constant at 6000, a non-professional switching to a professional occupation will gain 25.78 units 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 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 $\alpha = .05$).

We write the null and alternative hypothesis as:

```
# H0: beta.yard = 0
# Ha: beta.yard /= 0
```

We will do a t.test, so we need coefficient estimate, test statistic, SE, and p-value.

```
T.test.yard <- 0.042/0.016
n <- 131
k <- 2
p.values.yard <- 2*pt(abs(T.test.yard) , n-k, lower.tail = F)
```

[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.

We get the following output:

```
> T.test.yard <- 0.042/0.016
> n <- 131
> k <- 2
> p.values.yard <- 2*pt(abs(T.test.yard) , n-k, lower.tail = F)
> p.values.yard
[1] 0.009711646
```

The p value, at 0.0097, is smaller than alpha = .05, so we reject the null hypothesis that there is no discernible linear relationship between the presence of yard signs and Cuccinelli's vote share.

(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 $\alpha = .05$).

We write the null and alternative hypothesis as: vspace.5cm

```
# H0: beta.adjacent = 0
# Ha: beta.adjacent /= 0
```

We will do a t.test, so we need coefficient estimate, test statistic, SE, and p-value.

```
T.test.adjacent <- 0.042/0.013
p.values.adjacent <- 2*pt(abs(T.test.adjacent) , n-k, lower.tail = F)
```

We get the following output:

```
> T.test.adjacent <- 0.042/0.013
> p.values.adjacent <- 2*pt(abs(T.test.adjacent) , n-k, lower.tail = F)
> p.values.adjacent
[1] 0.001566685
```

The p value, at 0.0015, is smaller than alpha = .05, so we reject the null hypothesis that there is no discernible linear relationship between the presence of yard signs in adjacent precincts and Cuccinelli's vote share.

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

The coefficient for the constant term indicates that if all the explanatory variables in the model (precincts with yard signs, precincts adjacent to yard signs) are zero, then the value of the dependent variable will be equal to the constant term. In other words, in the absence of the signs treatment, the proportion of the vote share for Cuccinelli stands at 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 R squared for the model is 0.094. Generally, the closer the R squared is to 1, the better the fit. The lower the R squared value, the smaller the proportion of the variance for a response variable (in this case, Cuccinelli's vote share) that can be explained by the explanatory variables (yard signs and precincts adjacent to yard signs) in the a regression model. This suggests that the predictive power of the model can benefit from the inclusion of other factors, for example, party identity and campaign spending.

However, we have to be careful not to overinterpret what a low R squared value means for a model's predictive power. This is because as more variables are added to the model, R squared cannot decrease. We can perform other checks, such as by plotting studentized residuals of the data, to ascertain if the low R squared value could be due to outliers or non-linearity.