Problem Set 4

Applied Stats/Quant Methods 1: 16327268, Marcus Ó Faoláin

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 create a new variable `professional` in which professionals are coded as 1 and blue and white collar workers are coded as 0 using `ifelse`.

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

```

We make sure the new variable has been added to the `Prestige` dataset by calling the `head()` function.

```
head(Prestige)
```

```
> head(Prestige)
                    education income women prestige census type professional
gov.administrators      13.11  12351 11.16     68.8   1113 prof            1
general.managers        12.26  25879  4.02     69.1   1130 prof            1
accountants             12.77   9271 15.70     63.4   1171 prof            1
purchasing.officers     11.42   8865  9.11     56.8   1175 prof            1
chemists                14.62   8403 11.68     73.5   2111 prof            1
physicists              15.64  11030  5.13     77.6   2113 prof            1
```

Based on the output from the screenshot above, the code appears to have successfully added the new binary `professional` column.

(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 run a linear model with `prestige` as an outcome and `income`, `professional` and the interaction of the two as predictor variables. `Professional`, with it's binary values of 0 and 1, acts as a dummy variable in this case. We format the model using a * multiplication symbol instead of an + addition symbol to show that it is an interactive rather than additive model.

```
prestige.lm <- lm(prestige ~ income * professional, data = Prestige)
```

Running a summary call on the model yields output in the console, which is shown through the following image file

```
summary(prestige.lm)
```

```
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.

The formula for the prediction eqaution for a model with an interaction term is as follows:

$$Y_i = \beta_o + \beta_1 X_i + \beta_2 D_i + \beta_3 X_i D_i + \varepsilon_i$$

In our case, $Y_i$is prestige, β_o is the intercept, β_1 is the beta coefficient of `income`, X_i is the `income` variable, β_2 is the beta coefficient of the `professional` variable, D_i is whether someone is `professional` or not, β_3 is the coefficient of the interaction between the terms and ϵ_i is the error term.

From our results, the intercept β_0 has a value of 21.1422589, our coefficient of `income` β_1 has a value of 0.0031709, our `professional` coefficient β_2 has a value of 37.7812800 and the coefficient of our interaction between `income` and `professional` β_3 is -0.0023257.

Therefore, our generalised linear equation for `prestige` is as follows:

$$Y_i = 21.142 + 0.0031709 X_i + 37.781 D_i - 0.0023257 X_i D_i$$

For professionals, D_i takes a value of 1, therefore for professionals the linear equation for `prestige` is as follows:

$$Y_i = 21.142 + 0.0031709 X_i + 37.781 - 0.0023257 X_i$$

Simplified, the linear equation for `prestige` for professionals is:

$$Y_i = 58.923 + 0.0008452 X_i$$

In the linear equation for non-professionals, D_i takes a value of 0, so the linear equation for `prestige` for non professionals is as follows:

$$Y_i = 21.142 + 0.0031709 X_i$$

(d) Interpret the coefficient for `income`.

The coefficient for income is the amount by which our outcome variable, `prestige` will increase by, if our income is increased by one unit, if all other variables are kept constant.

In practice, this means an increase in `income` of 1 will result in a 0.0031709 increase in `prestige`, if all other variables are kept constant.

(e) Interpret the coefficient for `professional`.

The coefficient for `professional` is the amount by which `prestige` changes on average depending on whether one is professional or non-professional, keeping all other variables constant.

In practice this means that, keeping all other variables constant, being a professional results in a 37.781 increase in `prestige` on average versus not being a professional.

(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).

To answer this question, we must use the first formula for `prestige` for professionals from part (c), namely:

$$Y_i = 58.923 + 0.0008452X_i$$

To calculate the marginal effect on `prestige` of a \$1000 increase in income for professionals, we can calculate the value of `prestige` at two different income values \$1000 apart.

We shall take these `prestige` values at \$0 and at \$1000.

At \$0

$$Y_1 = 58.923 + 0.0008453(0)$$
$$Y_1 = 58.923$$

At \$1000

$$Y_2 = 58.923 + 0.0008452(1000)$$
$$Y_2 = 58.923 + 0.8452$$
$$Y_2 = 59.7682$$

To calculate the change in $\hat{y}$, we subtract Y_1 from Y_2.

$$Y_2 - Y_1$$
$$= 59.7682 - 58.923$$
$$= 0.8452$$

Therefore, the marginal effect of income on `prestige` on for professionals is 0.8452. This means a \$1000 increase in income for professionals results in a 0.8452 increase in `prestige`.

(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).

We can calculate the effect of changing one's occupation from non-professional to professional when income is \$6,000 using the two formulae for professional and non-professional discovered in (c).

For professionals:

$$Y_i = 58.923 + 0.0008452 X_i$$

For non-professionals:

$$Y_i = 21.142 + 0.0031709 X_i$$

We set the value of X_i to \$6,000 for each linear equations.

For professionals:

$$\begin{gathered} Y_p = 58.923 + 0.0008452(6,000) \\ Y_p = 58.923 + 5.0712 \\ Y_p = 63.9942 \end{gathered}$$

For non-professionals:

$$\begin{gathered} Y_n = 21.142 + 0.0031709(6,000) \\ Y_n = 21.142 + 19.0254 \\ Y_n = 40.1674 \end{gathered}$$

To calculate the effect of changing one's occupation from non-professional to professional when income is \$6,000 (ΔY), we subtract Y_n from Y_p.

$$\begin{gathered} \Delta Y = Y_p - Y_n \\ \Delta Y = (63.9942) - (40.1674) \\ \Delta Y = 23.8268 \end{gathered}$$

This means that a change from a non-professional role to a professional role results in a 23.8268 increase in the outcome variable `prestige`.

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 $\alpha = .05$).

We can use the results of the linear regression above to create a linear equation for the impact of lawn signs on vote share. The formula for linear relationships is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

where y is the proportion of vote that went to Cucinelli, β_0 is the constant term, β_1 is the coefficient of the variable for a precinct having a lawn sign assigned, X_1 is the variable for whether the precint has the lawn sign assigned, β_2 is the coefficient of the variable for whether the precinct is adjacent to a precint with a lawn sign, X_2 is the variable for whether a precinct has a sign adjacent to it and ε is the error term.

By putting in the results from the table, we can get the following equation:

$$Y = 0.302 + (0.042)X_1 + (0.042)X_2$$

The question above asks us to determine whether having these yard signs in a precinct affects vote share. This means we have to see whether having a yard sign in one's precinct has an effect on vote share. To see if it has an effect, we must investigate whether β_1 (0.042), the coefficient for X_1, is equal to zero or not

Our null hypothesis, H_0, states that the coefficient of the variable for precincts assigned with lawn signs, β_1, is equal to 0.

Our alternative hypothesis, H_α, states that the coefficient of the variable for precincts assigned with lawn signs, β_1, is not equal to 0.

$$H_0 : \beta_1 = 0$$
$$H_\alpha : \beta_1 \neq 0$$

In order to perform a hypothesis test, we must calculate a t-statistic for beta 1. We do this using the formula:

$$t = \frac{\hat{\beta}_0 - \beta_0}{se}$$

Substituting in the values and calculating values through code in R, we get:

```
t.statistic.assigned <- (assigned.b1) / (assigned.se)
t.statistic.assigned
```

$$t = \frac{0.042 - 0}{0.016}$$
$$t = 2.65$$

We must also calculate the critical values of a two tailed t distribution at a significance level of $\alpha = 0.05$.

If the value of our "assigned" t-statistic is greater than the critical value of the t-distribution, we can reject the null hypothesis and conclude that $\beta_1 \neq 0$.

This would mean that a precinct having these yard signs does affect vote share.

We can calculate the critical value with the `qt()` function in R.

We use a p-value of 0.05/2 since it is a two tailed t-test, one used simply measure if there is an effect rather than how much that effect is.

We get our degrees of freedom using the formula $df = n - 3$, where n is the total number of observations and 3 is the number of estimated coefficients.

```
df.b1 <- 131 - 3
qt(p = 0.05/2, df = df.b1, lower.tail = FALSE)
```

This gives us a critical value of 1.978671.

Since this critical value of 1.978671 is less than the t-statistic of 2.65, we can reject the null hypothesis that having yard signs in a precinct has no effect on voteshare. This means that β_1 is not equal to zero. ($\beta_1 \neq 0$)

We can therefore conclude that having these yard signs in a precinct does in fact affect 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 $\alpha = .05$).

Part (b) asks us to perform a hypothesis test on the second coefficient in the equation, β_2. This is the coefficient for the variable for precincts adjacent to lawn signs, X_2. We can perform a similar hypothesis test to the one in part (a).

To see if having a lawn sign in an adjacent precinct has an effect on voteshare or not, we must investigate whether β_2, the coefficient for X_1 is equal to zero or not.

Our null hypothesis, H_0, states that the coefficient of the variable for precincts adjacent to those with lawn signs, is equal to zero.

Our alternative hypothesis, H_α, states that the coefficient of the variable for precincts adjacent to those with lawn signs is not equal to zero.

$$H_0 : \beta_2 = 0$$
$$H_\alpha : \beta_2 \neq 0$$

In order to perform a hypothesis test for β_2, we must calculate a t-statistic. We do this using the following formula:

$$t = \frac{\hat{\beta}_0 - \beta_0}{se}$$

Substituting in values and calculating it in R, we get the following:

```
t.statistic.adjacent <- (adjacent.b2) / (adjacent.se)
t.statistic.adjacent
```

$$t = \frac{0.042 - 0}{0.013}$$
$$t = 3.230769$$

We must also calculate the critical values of a two tailed t distribution at a significance level of $\alpha = 0.05$.

If the value of our "adjacent" t-statistic is greater than the the critical value of the t-distribution, we can reject the null hypothesis and conclude that $\beta_2 \neq 0$.

This would mean that being adjacent to a precinct with the anti-McAuliffe yard signs does affect voter share.

We can calculate the critical value with the `qt()` function in R.

We use a p=value of 0.05/ 2 since it is a two tailed t-test.

We get out degrees of freedom using the formula df = n-3, where n is the total number of observations and 3 is the number of estimated coefficients.

```r
df.b2 = 131-3
qt(p = 0.05/2, df = df.b2, lower.tail = FALSE)
```

The critical value is 1.978671, the same as in the part (a).

Since the critical value of 1.978671 is less than the t-statistic of 3.230769, we can reject the null hypothesis that having yard signs in a precinct has no effect on vote share. This means that β_2 is not equal 0. ($\beta_2 \neq 0$).

We can therefore conclude that these yard signs being present in an adjacent precinct does in fact affect vote share.

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

The coefficient β_0 for the constant term is the point at which the regression line crosses the y-axis. This means it is the value of y when X_1 is equal to zero and X_2 is equal to zero.

What this means in practice is that it is the proportion of vote share that is predicted to go to Cuccinelli in precincts that neither have the anti-McAuliffe yard signs nor are adjacent to precincts with the anti-McAuliffe yard signs.

This means that we would expect Cuccinelli to receive around 30.2% of the vote share in precincts that neither have the lawn signs nor are adjacent to those with the 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?

The R^2 value of 0.094 tells us the proportion of total variance accounted for by the regression model. This means that around 9.4% of the variance of the model is accounted for by the model, which means around 90.6% of the variance of the outcome variable is not accounted for by it.

This means that a very large proportion, over 90%, of the variance of the outcome variable is not accounted for by the model. This suggests to us that other factors combined are far more important in explaining the variance of the outcome variable than the yard signs.

While the yard signs do have some effect on that variance of the vote share, it would be important to consider it as only one small part of all the factors that affected vote share in this election.