Problem Set 4

Student: Shekhar Kedia (23351315)
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)
```

`R` code to load the required dataset for use:

```r
# Reading in data
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`).

`R` code to create the new variable:

```r
# Recoding the 'type' variable to create 'professional' variable
Prestige$professional <- ifelse(Prestige$type %in% c("prof"), 1, 0)
# Removing the missing values (same as 'type' variable)
Prestige$professional[is.na(Prestige$type)] <- 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.)

`R` code to run the linear regression model:

```r
model_1 <- lm(prestige ~ income + professional + income:professional,
    data = Prestige)
stargazer(model_1) #Reporting Regression results
```

Table 1: Regression model result

	Dependent variable: 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

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

$\textit{prestige} = 21.142 + 0.003 * \textit{income} + 37.781 * \textit{professional} - 0.002 * \textit{income} * \textit{professional}$
or, $\textit{prestige} = 21.142 + 37.781 * \textit{professional} + (0.003 - 0.002 * \textit{professional}) * \textit{income}$

(d) Interpret the coefficient for `income`.

The coefficient for `income` is as follows:

- For non-professional people: 0.003
- For professional people: $(0.003 - 0.002) = 0.001$

There is a positive and statistically reliable relationship between income and prestige score, such that a one dollar increase in income is associated with an average increase of 0.003 units in prestige score for non-professionals and an average increase of 0.001 units in prestige score for professionals.

(e) Interpret the coefficient for `professional`.

For professionals with no income the prestige score is on average 37.781 units higher in comparison to non-professionals with 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).

The prestige score predicted equation for professional occupations is:
$\textit{prestige} = 21.1422589 + 0.0031709 * \textit{income} + 37.7812800 * \textit{professional} - 0.0023257 * \textit{income} * \textit{professional}$
or, $\textit{prestige} = 21.1422589 + 0.0031709 * \textit{income} + 37.7812800 * 1 - 0.0023257 * \textit{income} * 1$
or, $\textit{prestige} = 58.92354 + 0.0008452 * \textit{income}$

When the income increases by \$1,000 for professional occupations, the change in prestige score is:
$\Delta\textit{prestige} = (58.92354 + 0.0008452 * 1000) - (58.92354 + 0.0008452 * 0)$
or, $\Delta\textit{prestige} = 0.8452$

Interpretation: We can say that, the effect of a \$1,000 increase in income for professional is associated with an average increase in 0.845 units in prestige score.

(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).

The prestige score predicted equation for professional occupations with \$6,000 income is:
$\textit{prestige} = 21.1422589 + 0.0031709 * \textit{income} + 37.7812800 * \textit{professional} - 0.0023257 * \textit{income} * \textit{professional}$
or, $\textit{prestige} = 21.1422589 + 0.0031709 * 6000 + 37.7812800 * 1 - 0.0023257 * 6000 * 1$
or, $\textit{prestige} = 63.99474$

The prestige score predicted equation for non-professional occupations with \$6,000 income is:
$\textit{prestige} = 21.1422589 + 0.0031709 * \textit{income} + 37.7812800 * \textit{professional} - 0.0023257 * \textit{income} * \textit{professional}$
or, $\textit{prestige} = 21.1422589 + 0.0031709 * 6000 + 37.7812800 * 0 - 0.0023257 * 6000 * 0$
or, $\textit{prestige} = 40.16766$

When the occupation changes from non-professional to professional and income is $6,000$, the change in prestige score is:
$\Delta\textit{prestige} = 63.99474 - 40.16766$
or, $\Delta\textit{prestige} = 23.82708$

Interpretation: We can say that, the effect of a changing one's occupation from non-professional to professional when income is \$6,000 is associated with an average increase in 23.827 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 $\alpha = .05$).

To understand if having yards signs in a precinct affects vote share, we can conduct a hypothesis test to see if $\beta_1 = 0$:

H_0: Having yard signs in a precinct does not affect vote share. ($\beta_1 = 0$)
H_α: Having yard signs in a precinct affects vote share. ($\beta_1 \neq 0$)

The t-statistics is calculated using the formula:
$t - stat = \frac{Coefficient Estimate}{Standard Error} = \frac{0.042}{0.016} = 2.625$

[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.

Next, to calculate the p-value, we can run the following R code:

```
t1 <- 0.042/0.016
pvalue_1 <- 2 * pt(t1, df = (131-3), lower.tail = F) #For two-tail
print(pvalue_1) #Printing the output
```

We see that the p-value is 0.0097 which is below the given $\alpha = 0.05$ threshold. This indicates, we have found sufficient evidence to reject the null hypothesis that having yard signs in a precinct does not 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 $\alpha = .05$).

To understand if being next to precincts with yards signs affects vote share, we can conduct a hypothesis test to see if $\beta_2 = 0$:

H_0: Being next to precincts with yard signs does not affect vote share. ($\beta_2 = 0$)
H_α: Being next to precincts with yard signs affects vote share. ($\beta_2 \neq 0$)

The t-statistics is calculated using the formula:
$t - stat = \frac{CoefficientEstimate}{StandardError} = \frac{0.042}{0.013} \approx 3.231$

Next, to calculate the p-value, we can run the following R code:

```
t2 <- 0.042/0.013
pvalue_2 <- 2 * pt(t2, df = (131-3), lower.tail = F) #For two-tail
print(pvalue_2) #Printing the output
```

We see that the p-value is 0.0015 which is below the given $\alpha = 0.05$ threshold. This indicates, we have found sufficient evidence to reject the null hypothesis that being next to precincts with yard signs does not affect vote share.

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

The coefficient for the constant term is 0.302 which is the y-intercept of the predicted regression line. Meaning, the value of vote share when both the predictors i.e. precinct assigned lawn signs (say X1) and precinct adjacent to lawn signs (say X2) take value as zero.

We can interpret the result as, if the precinct was not assigned the lawn signs and was not adjacent to precincts having lawn signs, the proportion of vote share that went to McAuliff's opponent Ken Cuccinelli is 0.302 or 30.2%.

(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 model fit for the regression output i.e., the $R^2 = 0.094$. This shows that the predictors or the independent variables (precinct assigned lawn signs and precinct adjacent to lawn signs) explain approximately 9.4% of the variability in the dependent variable i.e., vote share.

This also shows that there is a huge amount of unexplained variance in the model which might be explained by introducing some other factors apart from yard signs.