Problem Set 4

Applied Stats/Quant Methods 1

Due: November 18, 2024

Name: Ombeline Mussat
Student Number: 24346050

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.

(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 can create a new variable in R called `professional` which is equal to 1 for professionals (which is 'prof' in our data) and 0 for blue and white collar workers (so all the other workers).

```
1 Prestige$professional <- ifelse(Prestige$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.)

Let's run a regression in R with prestige as an outcome and income, professional, and the interaction of the two as predictors. We display the results below.

```
1 regression <- lm(prestige ~ income + professional + income:professional,
    data = Prestige)
2 stargazer(regression, type = "latex")
```

Table 1:

	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. The prediction equation is:

$$\text{prestige} = 21.142 + 0.003 \times \text{income} + 37.781 \times \text{professional} - 0.002 \times (\text{income} \times \text{professional})$$

(d) Interpret the coefficient for `income`.

The coefficient for `income` means that a one-unit increase in income is associated with a 0.003 increase in prestige for blue-collar and white-collar workers (when professional = 0).

(e) Interpret the coefficient for `professional`.

The coefficient for `professional` means that professionals have, on average, a prestige that is 37.7813 units higher than blue-collar and white-collar workers, holding income constant.

(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).

We are interested in the effect of a \$1,000 increase in income on the prestige score for professional occupations (i.e., when the variable `professional` equals 1). We need to calculate the change in $\hat{y}$ (the predicted prestige score) associated with a \$1,000 increase in income, based on the regression results. Given that `professional = 1`, the regression equation for prestige is:

$$\text{prestige} = (21.142 + 37.781) + (0.003 - 0.002) \times \text{income}$$

This equation simplifies to:

$$\text{prestige} = 58.923 + 0.001 \times \text{income}$$

Thus, the marginal effect of income on prestige for professionals is 0.001 per 1-unit increase in income. This means that for each 1-unit increase in income, the prestige score for professional occupations increases by 0.001. Now, to calculate the change in $\hat{y}$ associated with a \$1,000 increase in income, we multiply the marginal effect by 1,000.

$$\text{Change in } \hat{y} = 0.001 \times 1000 = 1$$

Therefore, for professionals, an increase of \$1,000 in income results in a predicted increase of 1 unit in the 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).

We want to find the effect of changing one's occupation from non-professional to professional when her income is $6,000. So we need to calculate the marginal effect of changing the occupation from non-professional to professional when the variable `income` is equal to 6,000.

We can calculate the change in predicted prestige score $(\hat{y})$ when the occupation changes from non-professional (`professional` $= 0$) to professional (`professional` $= 1$) while `income` $= 6,000$. We can replace from our previous equation (in question c) the variable for income (6000) and for professional (either 0 or 1) to find the predicted prestige score.

Let's first calculate the prestige score predicted by the model for non-professional occupation (so when `professional` $= 0$). We will have the following equation:

$$\hat{y}_{\text{non-professional}} = 21.142 + 0.003 \times 6000 + 37.781 \times 0 - 0.002 \times (6000 \times 0)$$

$$\hat{y}_{\text{non-professional}} = 21.142 + 18$$

$$\hat{y}_{\text{non-professional}} = 39.142$$

We also have to calculate the predicted prestige score for professional occupation (so when `professional` $= 1$). We will have the following equation:

$$\hat{y}_{\text{professional}} = 21.142 + 0.003 \times 6000 + 37.781 \times 1 - 0.002 \times (6000 \times 1)$$

$$\hat{y}_{\text{professional}} = 21.142 + 18 + 37.781 - 12$$

$$\hat{y}_{\text{professional}} = 65.923$$

We can then calculate the change in prestige score from non professional to professional:

$$\text{Change in } \hat{y} = \hat{y}_{\text{professional}} - \hat{y}_{\text{non-professional}}$$

$$\text{Change in } \hat{y} = 65.923 - 39.142 = 26.781$$

$$\text{Change in } \hat{y} = 26.781$$

When someone with an income of $6,000 changes her occupation from non-professional to professional, her prestige score increases by 26.781 points.

Question 2: Political Science

Researchers are interested in learning the effect of all of those yard signs on voting prefer-ences.[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 use the results from the regression to test whether having yard signs in a precinct affects the vote share for Cuccinelli. We will conduct a hypothesis test with significance level $\alpha = 0.05$. Since we are testing the impact of a single predictor variable, we will conduct an individual t-test, as we are interested in the average effect of this covariate.

Step 1: Assumptions
For our hypothesis test, we make the following assumptions:

- There is an approximately linear relationship between predictor variables and the outcome variable.

- The covariates are independent of each other.

- The errors are normally distributed with constant variance.

- The errors are independent of the covariates.

Step 2: Hypotheses

Let β_1 represent the coefficient for the precincts assigned lawn signs.

- Null Hypothesis (H_0): $\beta_1 = 0$. This means that precincts with lawn signs have no effect on the vote share for Cuccinelli.

- Alternative Hypothesis (H_A): $\beta_1 \neq 0$. This means that precincts with lawn signs have some effect on the vote share for Cuccinelli.

Step 3: Test Statistic

We use a t-test for this individual coefficient. The test statistic is calculated as follows:

$$ t = \frac{\hat{\beta}_1}{\text{Standard Error}(\hat{\beta}_1)} $$

Given the results from the regression, we can calculate the t-test as follows:

```
1  t_statistics_1 <- 0.042/0.016
2  t_statistics_1
```

We have a test statistic of 2.625.

Step 4: P-value

The degrees of freedom for this test is calculated as:

```
1  df <- 131-2-1
2  df
```

We have a degree of freedom equal to 128.

Using this t-value and df $= 128$, we calculate the two-tailed p-value:

```
1  p_value_1 <- 2*pt(abs(t_statistics_1) , df, lower.tail = F)
2  p_value_1
```

We have a p-value of approximately 0.0097.

Step 5: Conclusion

Since the p-value is less than $\alpha = 0.05$, we reject the null hypothesis. We have sufficient evidence to conclude that precincts with lawn signs have some effect on the vote share for Cuccinelli.

(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 use the results from the regression to test whether being adjacent to precincts with yard signs affects the vote share for Cuccinelli. We will conduct a hypothesis test with significance level $\alpha = 0.05$. Since we are testing the impact of a single predictor variable, we will conduct an individual t-test, as we are interested in the average effect of this covariate.

Step 1: Assumptions
For our hypothesis test, we make the following assumptions:

- There is an approximately linear relationship between predictor variables and the outcome variable.
- The covariates are independent of each other.
- The errors are normally distributed with constant variance.
- The errors are independent of the covariates.

Step 2: Hypotheses
Let β_1 represent the coefficient for the precincts assigned lawn signs.

- Null Hypothesis (H_0): $\beta_2 = 0$. This means that precincts adjacent to lawn signs have no effect on the vote share for Cuccinelli.
- Alternative Hypothesis (H_A): $\beta_2 \neq 0$. This means that precincts adjacent to lawn signs have some effect on the vote share for Cuccinelli.

Step 3: Test Statistic
We use a t-test for this individual coefficient. The test statistic is calculated as follows:

$$t = \frac{\hat{\beta}_2}{\text{Standard Error}(\hat{\beta}_2)}$$

Given the results from the regression, we can calculate the t-test as follows:

```
1  t_statistics_2 <- 0.042/0.013
2  t_statistics_2
```

We have a test statistic of approximately 3.231.

Step 4: P-value
The degrees of freedom for this test is calculated as:

```
1  df <- 131−2−1
2  df
```

We have a degree of freedom equal to 128.

Using this t-value and df $= 128$, we calculate the two-tailed p-value:

```
1 p_value_2 <- 2*pt(abs(t_statistics_2) , df, lower.tail = F)
2 p_value_2
```

We have a p-value of approximately 0.0016.

Step 5: Conclusion
Since the p-value is less than $\alpha = 0.05$, we reject the nul hypothesis. We have sufficient evidence to conclude that precincts adjacent to lawn signs have some effect on the vote share for Cuccinelli.

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

This coefficient represents the expected value when all the predictors are equal to 0, it is the intercept of the regression line with the y axis. In this example, it would mean that it is a a precinct where there are no lawn signs within the precinct or in adjacent precincts. The constant term of 0.302 means that, in precincts without any assigned lawn signs or adjacent lawn signs, the predicted vote share for Cuccinelli is approximately 30.2%. This could act as a reference level for Cuccinelli's vote share in precincts which are unaffected by the lawn sign intervention.

(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 indicates that the predictors included in the model explain only 9.4% of the variance in the vote share for Cuccinelli. This low R^2 value suggests that the model has limited explanatory power. It means that lawn signs alone may not play a major role in determining the vote share. It is likely that other factors not included in the model have a bigger influence on the outcome.