Problem Set 4

Applied Stats/Quant Methods 1

Due: November 18, 2024

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 explore variable `type` and we become aware that there are some NA's.

```
bc prof   wc
44   31   23
```

First, we get rid of lost values, NAs. Next, we create a variable, named `professional`: in this variable, observations that are "prof" in variable `type` take the value "1"; observations that are not "prof" (i.e., blue and white collars) take values "0". Lastly, we convert our new variable into a factor one. As a result we have:

```
0  1
67 31
```

```
help(Prestige)
summary(Prestige$type)
Prestige$professional <- ifelse(!is.na(Prestige$type) & Prestige$type ==
    "prof", 1,
                                ifelse(!is.na(Prestige$type), 0, NA))
Prestige <- Prestige[!is.na(Prestige$professional), ]
summary(Prestige$professional)
is(Prestige$professional)
Prestige$professional <- factor(Prestige$professional)
summary(Prestige$professional)
```

(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 $\times$ dummy interaction.)

Table 1:

	Dependent variable:
	prestige
income	0.003^{***}
	(0.0005)
professional1	37.781^{***}
	(4.248)
income:professional1	-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.

Since there is a binary variable involved, the prediction equations are slightly different:

For non-professionals (`professional` = 0):

$$\text{prestige} = 21.142 + 0.003 \times \text{income}$$

Here, we remove the effect of "professional = 0", because mutliplying the values by 0 is 0 as a result.

For professionals (`professional` = 1):

$$\text{prestige} = 21.142 + 0.003 \times \text{income} + 37.781 \times 1 - 0.002 \times \text{income} \times 1$$

(d) Interpret the coefficient for `income`.

Since we have a dummy variable and an interaction, the way we interpret coefficients from the previous table slightly differs. The coefficient for `income` is read as follows: considering that all other variables are 0, for each unit increase in variable `income`, the prestige for non-professionals (`professional` = 0) experiments an increase, on average, of 0.003 units. This coefficient is statistically relevant at the 1% level.

(e) Interpret the coefficient for `professional`.

Considering that the reference category is `professional = 0`, i.e., non-professional workers (blue and white collars), we interpret the coefficient for `professional` as follows: considering that all other variables are 0, i.e., when `income` is 0, professional workers (`professional = 1`) have an average prestige score of 37,781 units compared to non-professional workers (`professional = 0`). That is, professional workers would have jobs that have, on average, 21,142 + 37,781 = 58,923 units of prestige, whilst non-professional workers' job prestige is situated, on average, at 21,142 units.

(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 determine the marginal effect of income on prestige when `professional` = 1, we use the formula:

$$\frac{\partial \hat{y}}{\partial \text{income}} = \beta_{\text{income}} + \beta_{\text{income*professional1}}$$

The marginal effect of income for professionals is:

$$\frac{\partial \hat{y}}{\partial \text{income}} = 0.003 + (-0.002) = 0.001$$

We are interested in the effect of a \$1,000 increase in income on the prestige score for professionals (`professional1` = 1). From the prediction equation for professionals:

We simplify the initial equation we provided in question c. We are left with the following equation:

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

The coefficient of `income` for professionals is 0.001. Thus, the effect of a \$1,000 increase in income is:

$$\Delta\hat{y} = 0.001 \times 1,000 = 1$$

For professional occupations (`professional1` = 1), a \$1,000 increase in income is associated with an average 1-unit increase 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).

To calculate the effect of changing one's occupation from non-professional to professional when income is \$6,000, we need to evaluate the change in the prediction equation when `professional1` changes from 0 to 1, while keeping `income` = 6,000.

For Professionals (`professional1` = 1):

$$\hat{y} = 21.142 + 0.003 \times \text{income} + 37.781 - 0.002 \times \text{income}$$

Difference in Prestige ($\Delta\hat{y}$):

The effect of changing from non-professional to professional is the difference between these two equations:

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

Or a simplified version:

$$\Delta\hat{y} = 37.781 - 0.002 \times \text{income}$$

Now we substitute income = 6,000:

$$\Delta\hat{y} = 37.781 - 0.002 \times 6,000$$

$$\Delta\hat{y} = 37.781 - 12 = 25.781$$

Changing from a non-professional to a professional occupation when income is \$6,000 is associated with an average 25.781-unit increase in the 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 first determine the critical value:

- With a $\alpha = .05$, and considering that we have a large sample size (131 observations), we assume we have a normal distribution (Z-distribution). The critical value for a two-tailed test at $\alpha = .05$ is 1.96.

- Our degrees of freedom are: 131 - 3 = 128, because we have 3 parameters.

Our hypothesis are:

- Null hypothesis: The coefficient for "Precinct assigned lawn signs" is 0, i.e., the yard signs have no effect on the vote share.

[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.

$$H_0 : \beta_{\text{lawn signs}} = 0$$

- Alternative hypothesis: The coefficient for "Precinct assigned lawn signs" is not 0, i.e., the yard signs do affect the vote share.

$$H_1 : \beta_{\text{lawn signs}} \neq 0$$

We calculate the t-test:

```
coefficient <- 0.042
se <- 0.016
t_statistic <- coefficient / se
t_statistic
```

The result is: 2.625

Now we calculate the p-value:

```
df <- 128
p_value <- 2 * pt(-abs(t_statistic), df)
p_value
```

The result is 0.00972002.

Our p-value is less than the significance level of 0.05, so we can reject the null hypothesis. That is, we have approximately 0.09% chance of observing a t-statistic as extreme as 2.625 under the null hypothesis. We can reject that the coefficient for "Precinct assigned lawn signs" has no effect on the 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 first determine the critical value:

- With a $\alpha = .05$, and considering that we have a large sample size (131 observations), we assume we have a normal distribution (Z-distribution). The critical value for a two-tailed test at $\alpha = .05$ is 1.96.

- Our degrees of freedom are: 131 - 3 = 128, because we have 3 parameters.

Our hypothesis are:

- Null hypothesis: The coefficient for "Precinct next to lawn signs" is 0, i.e., being next to precincts with these yard signs have no effect on the vote share.

$$H_0 : \beta_{\text{next lawn signs}} = 0$$

- Alternative hypothesis: The coefficient for "Precinct next to lawn signs" is not 0, i.e., being next to precincts with these yard signs have no effect on the vote share.

$$H_1 : \beta_{\text{next lawn signs}} \neq 0$$

We calculate the t-test:

```
se2 <- 0.013
t_statistic2 <- coefficient / se2
t_statistic2
```

The result is: 3.230769

Now we calculate the p-value:

The result is 0.00972002.

Our p-value is less than the significance level of 0.05, so we can reject the null hypothesis. That is, we have approximately 0.09% chance of observing a t-statistic as extreme as 3.230769 under the null hypothesis. We can reject that the coefficient for "Being next to lawn signs" has no effect on the vote share.

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

The constant coefficient, 0.302, represents the value of vote share received by Ken Cuccinelli when all other variables are 0. That is, in those precincts without lawn signs and that are also not adjacent to precincts with lawn signs, Ken Cuccinelli would receive, on average, 30.2% of the vote share.

(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?

To be able to assess the good fit for this regression, we must focus on our R-squared. In this case, the R-squared is 0.094. That means that about 9.4% of the variation we observe in variable `vote share` is explained by our explanatory variables. Despite the fact that the coefficients of the explanatory variables are statistically significant, the R-squared is relatively small and might mean that there are other factors not included in the model that could also be affecting our outcome variable, such as the candidate characteristics, voter demographic, etc.