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`).

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

Table 1: Professional Occupation Data

Occupation	**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

(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
reg1 <- lm(prestige ~ income * professional, data = Prestige)
summary(reg1)
```

Table 2:

	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.

$$\hat{\text{prestige}} =$$

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

```
#prestige_hat = reg1$coefficients[1] +
                # reg1$coefficients[2]*Prestige$income +
                # reg1$coefficients[3]*Prestige$professional +
                # reg1$coefficients[4]*Prestige$income*Prestige$
    professional
```

(d) Interpret the coefficient for `income`.

The coefficient for `income` is 0.003170909.

This represents the relative change in prestige from `income` while holding the variable of `professional` constant. So while `professional` is constant a one-unit increase in income is associated with an increase in prestige by roughly a 0.003. Since the coefficient is small, this suggests a relatively minimal effect of income on prestige when professional status is constant.

(e) Interpret the coefficient for `professional`.

The coefficient for `professional` is 37.781279955.

This represents the relative change in prestige from `professional` while holding the variable of `income` constant (i.e. professional variable switches from 0 to 1). So while `income` is constant a one-unit increase in professional status is associated with an increase in prestige by roughly a 37.7. This indicates a substantial positive effect of being a professional on prestige, independent of 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).

For a professional the prediction equation becomes:
`(as professional = 1)`

$$\widehat{\text{prestige}} = 58.92354 + (0.003171 - 0.002326) \times \text{income}$$

where (0.003171 - 0.002326) is the marginal effect of income when professional = 1. Which gives us a value of: 0.0008452.

So to calculate the effect of a \$1000 increase in income on prestige score for professional occupations we have:

$$\begin{aligned} &1000 \times 0.0008452 \\ &= 0.8452 \end{aligned}$$

For professional occupations, a \$1,000 increase in income is associated with an increase in the prestige score by 0.8452 units.

(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).

For a non professional the prediction equation becomes:
`(as professional = 0)`

$$\widehat{\text{prestige}} = 21.14226 + 0.003171 \times \text{income}$$

where 0.003171 is the marginal effect of income when professional $= 0$.

So to calculate the effect of a \$6000 increase in income on prestige score for non professional occupations we have:

$$6000 \times 0.003171$$
$$= 19.02545$$

For non professional occupations, a \$6,000 increase in income is associated with an increase in the prestige score by 19.02545 units.

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 $\alpha = .05$).

Assumptions:

(a) **Random Assignment** The treatment (sign assignment) is assumed to be randomly assigned to the 30 precincts, which is a key assumption for ensuring that any difference in vote share can be attributed to the yard signs rather than other factors.

(b) **Normality of Errors** The error terms are assumed to be normally distributed.

Null Hypothesis ($\mathbf{H_0}$):
There is no effect of lawn signs on vote share. The coefficient of the "lawn signs" variable is equal to zero.

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

Alternative Hypothesis ($\mathbf{H_a}$):
There is an effect of lawn signs on vote share. The coefficient of the "lawn signs" variable is not equal to zero.

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

Calculating the Test Statistic:

```
# beta_0 / SE
t_stat <- 0.042/0.016
# = 2.625
```

Calculating the P-Value:

```
# calculate the degrees of freedom
df <- 131 - 2 -1
# = 128
```

```
# Compute the two-tailed p-value (2x for tails)
p_val <- pt(t_stat, df, lower.tail = FALSE,) * 2
```

This give us:

$$\text{P-Value} = 0.00972002$$

Conclusion:

Since the p-value is much lower than the significance level of 0.05, we reject the null hypothesis. This suggests that there is a significant effect of lawn signs on the vote share for Ken 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 $\alpha = .05$).

Assumptions:

(a) **Random Assignment** The treatment (sign assignment) is assumed to be randomly assigned to the 30 precincts, and the adjacent precincts are assumed to have been randomly selected. This ensures that any difference in vote share can be attributed to the proximity to precincts with lawn signs rather than other factors.

(b) **Normality of Errors** The error terms are assumed to be normally distributed.

Null Hypothesis ($\mathbf{H_0}$): There is no effect of being adjacent to precincts with lawn signs on vote share. The coefficient of the "adjacent to lawn signs" variable is equal to zero.

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

Alternative Hypothesis ($\mathbf{H_a}$): There is an effect of being adjacent to precincts with lawn signs on vote share. The coefficient of the "adjacent to lawn signs" variable is not equal to zero.

$$H_0 : \beta_{\text{adjacent to lawn signs}} \neq 0$$

Calculating the Test Statistic:

```
# beta_0 / SE
t_stat_2 <- 0.042/0.013
# = 3.230769
```

Calculating the P-Value:

```
# Compute the two-tailed p-value (2x for tails)
p_val_2 <- pt(t_stat_2, df,  lower.tail = FALSE,) * 2
```

This give us:

$$\text{P-Value} = 0.00156946$$

Conclusion:

Since the p-value is much lower than the significance level of 0.05, we reject the null hypothesis. This suggests that there is a significant effect of being adjacent to precincts with lawn signs on the vote share for Ken Cuccinelli.

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

The coefficient for the constant term (0.302) represents the estimated baseline proportion of votes for Ken Cuccinelli in precincts where neither of the explanatory variables is present (i.e. their constants = 0 in the prediction equation). In other words, it reflects the predicted vote share for Cuccinelli in precincts that:

- Were not assigned lawn signs (treatment = 0).
- Are not adjacent to precincts with lawn signs (adjacency = 0).

(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 evaluate the model fit, we examine the R^2 value, which represents the proportion of the variance in the dependent variable (vote share for Cuccinelli) explained by the independent variables (presence of lawn signs and adjacency to precincts with lawn signs).

$$R^2 = 0.094$$

This indicates that only 9.4% of the variance in Cuccinelli's vote share across precincts is explained by the two variables in the model.

The low R^2 value suggests that while yard signs have a statistically significant effect on vote share (as indicated by the hypothesis test), they only explain a small portion of the overall variation in vote outcomes. This suggests that other factors not included in the model likely have a larger influence on the vote share, such as demographic, socioeconomic, or political factors that may vary across precincts.