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`).

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

(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.)

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

```
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
(Four observations were deleted because they do not exist)
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.

Based on the provided linear model results, the prediction equation can be written as:

$$\text{Prestige} = 21.1423 + 0.0032 \times \text{Income} + 37.7813 \times \text{Professional} - 0.0023 \times \text{Income} \times \text{Professional}$$

Explain the various terms in the equation:
(1)Prestige is the predicted outcome variable, Income is the continuous variable, and Professional is the dummy variable (1 for professionals and 0 for blue collar and white-collar workers).
(2)21.1423 is the intercept term, which represents the predicted reputation value when Income=0 and Professional=0 (i.e., blue collar or white-collar workers with zero income).
(3)0.0032 × Income is the effect of income, indicating that for every unit increase in income, reputation increases by 0.0032, while controlling for occupational type.
(4)37.7813 x Professional is the effect of professionalism, which means that under income control, professionals have a reputation increase of 37.7813 compared to blue collar or white-collar workers.
(5)0.0023 x Income x Professional is the interaction effect between income and profession, indicating that for every unit increase in income of a professional, their reputation decreases by 0.0023, while controlling for other variables.

(d) Interpret the coefficient for `income`.

In the linear regression model, the coefficient of income is 0.0031709. The explanation for this coefficient is as follows:

For non professionals: for every unit increase in income, the expected increase in prestige is 0.0031709 units. This is because when profession is 0, the interaction term is 0, so the total effect of income is its main effect.

For professionals (professional=1): For every unit increase in income, the increase in reputation is the main effect of income minus the interaction effect. Therefore, the increase is 0.0032-0.0023=0.0009, which means that for professionals, for every unit increase in income, the expected reputation increases by 0.0009 units.

In summary, the income coefficient of 0.0031709 indicates that for non professionals, for every unit increase in income, reputation increases by approximately 0.0032 units. For professionals, due to the positive interaction effect between income and professionals, for every unit increase in income, reputation increases by approximately 0.0009 units. This indicates that the positive impact of income on reputation is more significant among non professionals.

(e) Interpret the coefficient for `professional`.

In the linear regression model, the professional coefficient is 37.7812800. The explanation for this coefficient is as follows:
For professionals: This coefficient indicates that, while controlling for constant income, professionals have an average reputation of about 37.7813 units higher than non professionals (i.e. blue collar and white-collar workers, professional=0). This coefficient reflects the average difference in reputation between professionals and non professionals.

However, it should be noted that this coefficient is obtained assuming that income remains constant, as the model includes income variables. This means that the difference of 37.7813 is an estimate at a specific income level. If income changes, the difference in reputation between professionals and non professionals may vary, which is why the model includes income and professional interaction terms.

(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 professionals, the impact of a $1000 increase in income on reputation scores can be determined by calculating the change in the income term in the prediction equation. According to the given prediction equation:

Prestige=21.1423+0.0032 x Income+37.7813 x Professional
-0.0023 x Income x Professional

When the value of professional is 1, the equation becomes:

Prestige=58.9236+0.0009×Income

In this equation, the coefficient of income is 0.0009, which means that for every $1 increase in income, the expected reputation score of professionals increases by 0.0009 units. Therefore, for an increase in income of $1000, the expected increase in reputation score is: y^=0.0009 × 1000=0.9

So, for professionals, an increase of $1000 in income is expected to increase their reputation score by 0.9 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).

To calculate the impact of an income of \$6000 on reputation scores when transitioning from non professionals to professionals, it is necessary to consider the coefficients of professional variables and their interaction with income. According to the given prediction equation: Prestige=21.1423+0.0032 x Income+37.7813 x Professional -0.0023 x Income x Professional

Firstly, calculate the reputation score for non professionals:
Prestige non-pro=40.3423
Then calculate the reputation score of professionals:
Prestige pro=74.3236
Finally, calculate the change in reputation score when transitioning from a non professional to a professional:
Prestige^=33.9813

Therefore, when the income reaches \$6000, the expected increase in reputation score from non professionals to professionals is about 33.98 units. This change includes the direct effect of professional variables (37.7813) and the interaction effect with income (-13.8).

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 $\alpha = .05$).

(a) Null hypothesis (H_0): The presence of lawn signs does not affect vote share.

(b) Alternative hypothesis (H_1): The presence of lawn signs affects vote share.

The t-values for the coefficients are calculated as follows:

$$t = \frac{\text{Coefficient}}{\text{Standard Error}}$$

For Precinct assigned lawn signs:

$$t1 = \frac{0.042}{0.016} = 2.625$$

[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.

For Precinct adjacent to lawn signs:

$$t2 = \frac{0.042}{0.013} = 3.231$$

For alpha=0.05 and double tailed test, the degrees of freedom are N-K-1=131-2-1=128. By using a t-distribution table or calculator, we can find the critical t-value. For a two tailed test with 128 degrees of freedom and alpha=0.05, the critical t-value is approximately 1.98.

Conclusion
For the constituency with assigned lawn signs (t1 is 2.625): Since 2.625 is greater than 1.98, we reject the null hypothesis, indicating that lawn signs have a significant impact on voting shares in the constituency with assigned lawn signs.

For the constituency adjacent to the one assigned the lawn sign (t2 is 3.231): Since 3.231 is greater than1.98, we also reject the null hypothesis, indicating that the lawn sign also has a significant impact on the voting share in the constituency adjacent to the one assigned the lawn sign.

(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 $\alpha = .05$).

(a) Null hypothesis (H_0): Being next to precincts with lawn signs does not affect vote share, i.e., $\beta_2 = 0$.

(b) Alternative hypothesis (H_1): Being next to precincts with lawn signs affects vote share, i.e., $\beta_2 \neq 0$.

The t-values for the coefficients are calculated as follows:

$$t = \frac{\text{Coefficient}}{\text{Standard Error}}$$

For precincts adjacent to lawn signs:

$$t = \frac{0.042}{0.013} = 3.231$$

For a two-tailed test with $\alpha = 0.05$ and degrees of freedom $df = N-k-1 = 131-2-1 = 128$, the critical t-value is approximately 1.98.

```
Conclusion
Since the calculated t-value (3.231) is greater than the critical t-value
(1.98), we reject the null hypothesis. This indicates that there is a
statistically significant effect of being next to precincts with yard
signs on vote share.
```

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

Explanation of the constant term: The constant term represents the value of the dependent variable (i.e. the voting ratio of McAuliffe to Cucinelli) predicted by the model when the values of all independent variables (i.e. the lawn marker variable) are zero. In this specific research context, this means that if we consider a constituency that is neither assigned a lawn sign nor adjacent to these constituencies (i.e., both professional and adjacent variables are 0), the predicted voting ratio for Cucinelli is 0.302.
Practical significance: This constant term can be seen as the average expected voting percentage of Cucinelli in these constituencies without any influence from lawn signs. It provides a baseline for comparing the impact of lawn signs on voting proportions.
In short, the constant term 0.302 indicates that in the absence of any lawn signs, the expected voting percentage for Cucinelli is 30.2%. This value can help understand the basic situation in the absence of external influencing factors and serve as a starting point for evaluating the effectiveness of lawn signage.

(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-squared value is a very important indicator when evaluating linear regression models. The R-squared value measures the proportion of dependent variable variation explained by the model to the total variation, ranging from 0 to 1. The closer the R-squared value is to 1, the higher the goodness of fit of the model, and the better the model can explain the variation of the dependent variable
For the linear regression model of this question, the R-squared value is 0.094. This means that the lawn logo factor in the model can only explain 9.4% of the variation in voting shares. That is to say, although lawn signs may have a certain impact on voting shares, their importance is relatively low compared to other factors not included in the model, so there are other factors that have not been considered that may have a greater impact on voting shares
Therefore, although the influence of lawn signs is statistically significant (as shown in the t-test results), their overall explanatory power for voting shares is limited from the R-squared values. This may mean that voters' voting preferences are influenced by multiple factors, including but not limited to lawn signs.