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:
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
# Install and load the car package
install.packages("car")
library(car)

# Load the Prestige dataset
data(Prestige)

# Create a new dummy variable professional
Prestige$professional <- ifelse(Prestige$type == "prof", 1, 0)

# Check the data frame
head(Prestige[c('professional','type')])
```

```
                    professional type
gov.administrators             1 prof
general.managers               1 prof
accountants                    1 prof
purchasing.officers            1 prof
chemists                       1 prof
physicists                     1 prof
```

(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
# Run a linear model with interaction
# Linear model: prestige ~ income + professional + income:professional
model <- lm(prestige ~ income * professional, data = Prestige)
summary(model)
```

1.
```
Call:
lm(formula = prestige ~ income + professionals + income:professionals,
    data = data)

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 ***
professionals        37.7812800  4.2482744   8.893 4.14e-14 ***
income:professionals -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
  (4 observations deleted due to missingness)
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.

This is the general formula

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 D_1 + \beta_3 x_1 D_1$$

Based on the regression above, we got this prediction equation

$$\begin{aligned}\widehat{\text{prestige}} = {} & 21.1422589 + 0.0031709 \times \text{income} \\ & + 37.7812800 \times \text{professional} \\ & - 0.0023257 \times \text{income} \times \text{professional}\end{aligned} \tag{1}$$

(d) Interpret the coefficient for `income`.
The coefficient for income can be interpreted that for non-professionals (professional = 0) when there is an unit increase in income there will be on average an increase of 0.0031709 unit in prestige.

(e) Interpret the coefficient for `professional`.
The coefficient for professional 37.7812800 represents the average difference in prestige between professional and non-professional jobs when income = 0.

(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).

Marginal effect = $\beta_{\text{income}} + \beta_{\text{income:professional}}$
Marginal effect = 0.0031709 + -0.0023257
Marginal effect = 0.0008452 *1000
change in prestige = 0.8452

A 1,000 dollars increase in income increases the prestige score by 0.8452 for professionals.

(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).

Marginal effect = $\beta_{\text{professional}} + \beta_{\text{income:professional}} \times 6000$

Marginal effect = 37.7812800 + -0.0023257 * 6000
Marginal effect = 23.82708

The prestige score increases by 23.83 points when moving to a professional job with an income of 6,000 dollars.

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 $\alpha = .05$).

H0: $\beta_1 = 0$
$H1$:$\beta_1 \neq 0$

β_1 is the effect of being assigned to lawn signs with alpha 0.05

This is the formula for t-statistic for hypotheses testing

$$t - statistic = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

[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.

```r
# Hypothesis test: Effect of yard signs on vote share
# Coefficients and standard errors
coef_lawn <- 0.042
se_lawn <- 0.016

# t-statistic
t_lawn <- coef_lawn / se_lawn

# p-value
p_lawn <- 2 * (1 - pt(abs(t_lawn), df = 131 - 3))  # df = n - predictors
  - 1
p_lawn
```

The p-value is 0.00972002 and we can reject the null-hypothesis that $\beta_1 = 0$.
Therefore the assigned yard signs on precinct has an effect the votes hare.

(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 $\alpha = .05$).

H0: $\beta_2 = 0$
$H1 : \beta_2 \neq 0$

β_2 is the effect of being assigned to lawn signs with alpha 0.05

This is the formula for t-statistic for hypotheses testing

$$t - statistic = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)}$$

```
# Hypothesis test: Effect of adjacency to treatment precincts

# Coefficients and standard errors
coef_adj <- 0.042
se_adj <- 0.013

# t-statistic
t_adj <- coef_adj / se_adj

# p-value
p_adj <- 2 * (1 - pt(abs(t_adj), df = 131 - 3))
p_adj
```

The p value is 0.00156946
The p-value is 0.00156946 and we can reject the null-hypothesis that $\beta_1 = 0$, therefore we can conclude that adjacent vote share on Precinct has an effect on the voteshare.

(c) Interpret the coefficient for the constant term substantively.
The constant (β_0) represents the proportion of votes for Cuccinelli in precincts without signs and not adjacent to treated precincts.

(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?
$R^2 = 0.094$ implies that only 9.4% of the variance in vote share is explained by the model.

Yard signs may have a small effect, but other unmodeled factors likely explain most of the variation.