Problem Set 4

Applied Stats/Quant Methods 1

Due: December 3, 2023

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 Sunday December 3, 2023. 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`).**

```
# Create dummy variables
df$professional <- ifelse(df$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 $\times$ dummy interaction.)**

```
# Run a linear model

Call:
lm(formula = df$prestige ~ df$income + df$professional, data = Prestige)

Residuals:
Min       1Q   Median       3Q      Max
-19.7458  -6.3013  -0.5493   5.4810  29.7818

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.062e+01  1.714e+00  17.866  < 2e-16 ***
df$income       1.371e-03  2.563e-04   5.348 6.12e-07 ***
df$professional 2.276e+01  2.318e+00   9.817 4.07e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.652 on 95 degrees of freedom
(4 observations deleted due to missingness)
Multiple R-squared:  0.7491,Adjusted R-squared:  0.7438
F-statistic: 141.8 on 2 and 95 DF,  p-value: < 2.2e-16

# check values
> sprintf("%.20f",3.062e+01)
[1] "30.62000000000000099476"
> sprintf("%.20f",1.371e-03)
[1] "0.00137100000000000004"
> sprintf("%.20f",2.276e+01)
```

```
[1] "22.760000000000156319"
```

(c) **Write the prediction equation based on the result.**

Prestige = 30.62 + 0.001 **Income* + 22.76 * *Professional*

(d) **Interpret the coefficient for `income`.**

The coefficient for income = 0.001. This is the slope associated with income, when controlling for professional.

(e) **Interpret the coefficient for `professional`.**

The coefficient for professional = 22.76. This is the effect associated with professional status when controlling for income.

Controlling for income, blue and white collar workers exhibit on average, a 22.76 unit drop in income compared to those with a higher professional status.

Controlling for income, those with professional status, exhibit, on average a 22.76 unit increase in income compared to blue and white collar workers.

(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).

Prestige = 30.62 + 0.001 * 1,000 + 22.76 * 1

A 1,000 dollar increase in income will increase the prestige score for professional occupations, by a unit value of 1 (0.001 * 1,000)

(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).

First, let's look at the outcome when non-professional status (professional set at 0):

$Prestige = 30.62 + 0.001 * 6{,}000 + 22.76 * 0$
$Prestige = 59.38$

If the individual changes their status to professional, the Prestige score will increase by 22.76 unit scale points.

$Prestige = 30.62 + 0.001 * 6{,}000 + 22.76 * 1$
$Prestige = 82.14$

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. **[(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 $\alpha = .05$).**

Category	D1	D2
Lawn Sign	1	0
Adjacent Sign	0	1

$$\hat{Y} = \beta_0 + \beta_1 D_1 + \beta_2 D_2$$

If the garden has a lawn sign, D1 equals to 1, and D2 equals to 0.

[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.

In that case...

CuccinelliVote = 0.302 + 0.042 (D1) + 0.042(D2)
CuccinelliVote = 0.302 + 0.042 (1) + 0.042(0)

Having a lawn sign seems to correspond to a 0.042 unit increase in the Cuccinelli Vote.

Null Hypothesis:

$$\beta_1 = 0$$

Alternative Hypothesis:

$$\beta_1 \neq 0$$

T-stat = 0.042/ 0.016 = 2.625

P-value = 0.00972

The p-value is statistically significant and we have evidence to reject the null hypothesis, and to support the alternative hypothesis that having a garden sign increases vote share by an on average unit scale increase of 0.042, controlling for gardens which are adjacent.

(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 $\alpha = .05$).

Category	D1	D2
Lawn Sign	1	0
Adjacent Sign	0	1

$$\hat{Y} = \beta_0 + \beta_1 D_1 + \beta_2 D_2$$

If the house is adjacent to a garden with a lawn sign, then D2 equals to 1, and D1 equals to 0. In that case...

CuccinelliVote = 0.302 + 0.042 (D1) + 0.042(D2)

CuccinelliVote = 0.302 + 0.042 (0) + 0.042(1)

Having an adjacent lawn sign seems to correspond to a 0.042 unit increase in the Cuccinelli Vote.

Null Hypothesis:

$$\beta_2 = 0$$

Alternative Hypothesis

$$\beta_2 \neq 0$$

T-stat = 0.042/ 0.013 = 3.23

P-value = 0.001573

The p-value is statistically significant and we have evidence to reject the null hypothesis, and to support the alternative hypothesis that having a garden adjacent to a garden with a lawn sign increases vote share by an on average unit scale increase of 0.042, controlling for gardens with lawn signs.

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

The coefficient for the constant term is equal to 0.302. This is the intercept, which is the predicted value Y value when both the variable for lawn sign (D1) = 0, and adjacent lawn sign (D2) = 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?

R^2=0.094 . This appears to indicate a poor fit, which may lead to inaccurate predictions. The R^2 value indicated that only approximately 9.5 per cent of the variation in Y is explained by the two variables used in the model.

Both the explanatory variables have the same coefficient indicating likely high correlation between the two.

Taken together, this tells us that the model could benefit from including other variables alongside yard signs (e.g. maybe variables such as income etc).