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`).

To do so I used the ifelse() function in R, subsetting the type ”prof” and setting this as my: if ”prof” then 1, else 0.

I also include this as a new column in the Prestige dataset.

```
> Prestige$professional<-ifelse(Prestige$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 × dummy interaction.)

As I already dummy coded the variable professional in (a), I use income as my second explanatory variable, with income as my first, and the interaction of income and professional as my third.

```
> prestige_interact <- lm(prestige ~ income + professional + income:professional,
> summary(prestige_interact)

Call:
lm(formula = prestige ~ income + professional + 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
(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.

Y = 21.142 + 0.003 * x_1 + 37.781 * x_2 - 0.002 * x_3
prestige = 21.142 + 0.003 * income + 37.781 * professional - 0.002 * income * professional

(d) Interpret the coefficient for `income`.

If income increases by 1 unit ($), prestige increases by 0.0031709 units, provided the individual considered is considered a blue or white collar worker (our reference category 0).

(e) Interpret the coefficient for `professional`.

If the assigned type changes from 0 (blue and white collar workers) to 1 (professionals) we would expect prestige to increase by 37.7812800 units, provided the income equals 0 (so the interaction term = 0). For other income values, the effect of work type on prestige increases by 0.002 units for each 1-unit ($) increase in income due to the interaction.

(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 calculate this effect, I have decided to calculate the difference (subtract) between the initial, regular prediction equation written in (c) and a new equation considering income + 1000. I have also decided to round the estimations to three decimal points, which I recognize changes the result of this calculation slightly.

Original equation:
prestige_initial = 21.142 + 0.003 * income + 37.781 * 1 - 0.002 * (income*1) =
21.142 + 37.781 + 0.003 * income - 0.002 * income

Equation with income+1000:
prestige_new = 21.142 + 0.003 * (income+1000) + 37.781 * professional - 0.002 * (income+1000)*1 =
21.142 + 0.003 * income + 3 + 37.781 - 0.002*income - 2 =
21.142 + 37.781 - 2 + 3 + 0.003 * income - 0.002 * income

Subtracting the simplified formulas: 21.142 + 37.781 + 0.003 * income - 0.002 * income - (21.142 + 37.781 - 2 + 3 + 0.003 * income - 0.002 * income) =

-2+3=1

So, in summary when income increases by $1000 prestige is expected to increase by 1 unit (after rounding to three decimal points) for somebody considered a professional.

(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 this change I once again subtract. This time I subtract the prediction equation for prestige for the new occupation professional, from the prediction equation for the old occupation non-professional.

Equation for professional occupations:
prestige = 21.142 + 0.003 * 6000 + 37.781 * 1 - 0.002 * 6000*1
Equation for non-professional occupations:
prestige = 21.142 + 0.003 * 6000 + 37.781 * 0 - 0.002 * 6000*0

Simplified, subtracted: 21.142 + 0.003 * 6000 + 37.781 - 0.002 * 6000 - (21.142 + 0.003 * 6000) =
37.781+12 = 25.781

So, in summary, we expect prestige to increase by 25.781 units if income is 6000 and the occupation then shifts from non-professional to professional.

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 $\alpha = .05$).

To answer this question I first formulate the following hypotheses:
H0 = having these yard signs in a precinct does not affect vote share
H1 = having these yard signs in a precinct affects vote share

The definition of H1 means we have to conduct a two-tailed hypothesis test.

To conduct the test I first calculate my t-value:

```
> #t*_yes = coefficient/SE = 0.042/0.016
> t_yes <- 0.042/0.016
```

[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.

I use the coefficient for precincts assigned lawn sign from the regression (0.042) and divide it by the standard error already calculated in the regression right below the coefficient (0.016).

Then my degrees of freedom:

```
> #degrees of freedom
> # n-2-1
> df <- 131-2-1
```

I calculate my degrees of freedom with the formula n-k-1. Using the entire sample N=131 and the number of conditions/variable=2 (for precincts assigned and precincts adjacent) for k.

Lastly, I am able to calculate my p-value:

```
> #p-values
> #yard signs
> 2*pt(abs(t_yes), df = df, lower.tail=F)
[1] 0.00972002
```

I calculate this with the usual formula for two-sided hypothesis testing, with my previously calculate t-value (t_yes) and my degrees of freedom (df).

Since the result is 0.00972002 which is less than $\alpha = .05$ we reject our null hypothesis that having these yard signs in a precinct does not affect 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 $\alpha = .05$).

To answer this question I once again first formulate the following hypotheses:
H0 = being from a precinct adjacent to one with these yard signs does not affect vote share
H1 = being from a precinct adjacent to one with these yard signs affects vote share

The definition of H1 means we have to conduct a two-tailed hypothesis test.

To conduct the test I first calculate my t-value:

```
> #t*_no = coefficient/SE = 0.042/0.013
> t_no <- 0.042/0.013
```

I use the coefficient for precincts adjacent to lawn signs from the regression (0.042) and divide it by the standard error already calculated in the regression right below the coefficient (0.013).

I can use the same degrees of freedom as in (a).

So then, I am able to calculate my p-value:

```
> #no yard signs
> 2*pt(abs(t_no), df = df, lower.tail=F)
[1] 0.00156946
```

I calculate this with the usual formula for two-sided hypothesis testing, with my previously calculate t-value (t_yes) and my degrees of freedom (df).

Since the result is 0.00156946 which is less than $\alpha = .05$ we reject our null hypothesis that being near precinct with those yard signs does not affect vote share.

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

The constant is the intercept, representing the expected value of the outcome variable, if the other variables = 0. In this case, this number represents the expected vote proportion for the opponent Ken Cuccinelli in a precinct that is neither assigned yard signs nor adjacent to one. This value shows the baseline vote share in precincts unaffected by the yard sign intervention and could be seen as a reference for evaluating the impact of yard signs on 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?

The R^2 value of 0.094 indicates that approximately 9.4% of the variance in the proportion of the vote that went to McAuliff's opponent Ken Cuccinelli across the considered 131 precincts is explained by the precinct's being assigned or being adjacent to yard signs. This value is relatively low, which suggests that while yard signs may have a statistically measurable impact, the majority of variation in voting behavior is influenced by other factors not included in this model.

This means that while yard signs might contribute to the voting outcomes, their effect is rather small when compared to other factors not represented in this model. (other factors might be demographic differences, history of the area, other advertisements).