Problem Set 4 Answers

Applied Stats/Quant Methods 1

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`).

- First, I imported the data:

```
install.packages(car)
library(car)
data(Prestige)
help(Prestige)
```

- Then I created a binary professional variables from the variable 'type', using professionals as 1 and blue and white collar workers as 0:

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

Table 1: Table of professional and type

	0	1
bc	44	0
prof	0	31
wc	23	0

- We can see from the table above that this was correctly coded, as anything coded as 1 corresponds to "professional" in the type variable, and anything coded as 0 corresponds to "bc" or "wc" in the type variable.

(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.)

- Then I fit the linear regression model using prestige as the outcome variable, and income, professional, and the interaction between income and professional as the predictor variables

```
model_int <- lm(prestige~income +
                    professional +
                    income*professional, data=Prestige)
```

- The results of the linear regression are as follows:

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.

- The prediction equation is as follows:

$$Prestige = 21.142 + 0.003 * Income + 37.781 * Professional - 0.002 * Income * Professional$$

(d) Interpret the coefficient for `income`.

- **For those in a blue or white collar profession, with every 1 unit increase in income, prestige score increases, on average, by 0.003 units, while holding all other variables in the model constant.**
- Furthermore, the p-value is 7.55×10^{-9}, which means that we can reject the null hypothesis that there is no relationship between income and prestige.

(e) Interpret the coefficient for `professional`.

- **For poor individuals in a professional occupation, the average prestige score is 37.781 points higher than for poor individuals in a blue or white collar profession.**
- Furthermore, the p-value is 4.14e-14, which means that we can reject the null hypothesis that there is no relationship between type of profession and prestige.

(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).

- First I calculated the prestige score for a professional income at $0
- ```
  zerodol <- 21.1422589 + 0.0031709*0 + 37.7812800*1 - 0.0023257*0*1
  ```
- Then I calculated the prestige score for a professional income at $1000
- ```
  thousdol <- 21.1422589 + 0.0031709*1000 + 37.7812800*1 - 0.0023257*
      1000*1
  ```
- Then I subtracted the calculation for $0 from the calculation for $1000 to find the marginal effect of income for professional occupations.
- ```
  marginaleffectdol <- thousdol - zerodol
  ```
- **I found that the marginal effect of income when the variable *professional* takes the value of 1 is 0.8452**
  ```

(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 I calculated the prestige score for a non-professional income at \$6000
- ```
  nonprofmarg <- 21.1422589 + 0.0031709*6000 + 37.7812800*0 - 0.0023257
      *6000*0
  ```
- Then I calculated the prestige score for a professional income at \$6000
- ```
  profmarg <- 21.1422589 + 0.0031709*6000 + 37.7812800*1 - 0.0023257*
      6000*1
  ```
- Then I subtracted the calculation for non-professionals from the calculation for professionals to find the marginal effect of changing to a professional income when one's income is \$6000.
- ```
  margeffectprof <- profmarg- nonprofmarg
  ```
- **I found that the marginal effect of changing from a non-professional occupation to a professional occupation when one's income is \$6000 is 23.82708**
  ```

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 $\alpha = .05$).

$$\mathrm{t} = \frac{\beta_i}{se}$$

- Formula from pg. 338 of Agresti and Finlay
- I also used code from the week 10 lecture for using this to get the p-values
- My code is as follows:

[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
TS1 <- (0.042)/(0.016)
TS2 <- (0.042)/(0.013)
TSconst <- (0.302)/(0.011)

n = 131
k = 2

pval1 <- 2*pt(abs(TS1), n-k, lower.tail = F)
pval2 <- 2*pt(abs(TS2), n-k, lower.tail = F)
pvalconst <- 2*pt(abs(TSconst), n-k, lower.tail = F)
```

- The results are in the table below:

Table 3: Impact of Lawn Signs on Vote Share

	Coefficient	**Standard Error**	**p-value**
Precinct assigned lawn signs (n=30)	0.042	(0.016)	0.009711646
Precinct adjacent to lawn signs (n=76)	0.042	(0.013)	0.001566685
Constant	0.302	(0.011)	1.013866×10^{-55}

- The p-value for this β coefficient is 0.009711646, which is less than 0.05. Therefore, we can **reject the null hypothesis with $\alpha = 0.05$** which states that there is no relationship between having a yard sign in a precinct and the vote share of Ken Cuccinell. This β coefficient indicates that having a yard sign in the precinct compared to not having a yard sign in the precinct is, on average, associated with a 0.042 increase in vote share for 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 $\alpha = .05$).

- The p-value for this β coefficient is 0.001566685, which is less than 0.05. Therefore, we can **reject the null hypothesis with $\alpha = 0.05$** which states that there is no relationship between having a yard sign next to a precinct and the vote share of Ken Cuccinell. This β indicates that having a yard sign next to a precinct compared to not having a yard sign next to a precinct is, on average, associated with a 0.042 increase in vote share for Cuccinelli

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

- The coefficient for the constant term is 0.302. The p-value associated with this is 1.013866e-55. Therefore, the analysis for the constant term is as follows:
- **When there are no yard signs placed in or next to precincts, the vote share value is 0.302 units.** This value is significant at the 0.001 level, as the p-value is significantly small. We can therefore reject the null hypothesis which states that the constant term is 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?

- I calculated an overall F-statistic using the formula from lecture: $F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$ where n is the overall number of participants and k is the number of explanatory variables in the model.

```
n = 131
k = 2
```

```
R2=0.094

ftest <- (R2/(k-1)) / ((1 - R2)/(n-k))

df1 <- k-1
df2 <- n - k - 1

fpval <- df(ftest, df1, df2)
```

f-statistic	13.38411
p value	0.0001782264

Table 4: F-statistic and p-value

- Given that there is an extremely p-value for this F-value, we have strong evidence against $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$.
- **This overall suggests that at least one of the explanatory variables is related to vote share for Ken Cuccinelli. Also, the R^2 value is 0.094, meaning that 9.4% of variance in vote share for Cuccinelli is explained by lawn signs in or adjacent to precincts.**
- So, we can conclude that we obtain significantly better predictions of y using the multiple regression equation than by using $\bar{y}$.
- In other words, at least one of the variables in our model should have some explanatory power in regards to vote share.