Problem Set 4

Applied Stats/Quant Methods 1

Due: November 26, 2021

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 in .pdf form.
- This problem set is due before class on Friday November 26, 2021. No late assignments will be accepted.
- Total available points for this homework is 80.

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`.)

Using the hint provided in the task and the wisdom of Google, I created a new variable professional and recoded the existing type variable in the dataset Prestige as a dummy. Category 'prof' was recoded to 1 and all else were recoded to 0.

```
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.)

I plugged in my new dummy variable in the model using the available variables for prestige as my response variable and income, dummy variable for professional category and the interaction of these variables.

The R code for the regression model looks like this:

```
mod1 <- lm(Prestige$prestige ~ Prestige$income + professional + Prestige$
    income:professional, data = Prestige)
options(scipen=999)
summary(mod1)
```

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

Based on the propositions of the regression the equation should look have the following form (I don't know how to remove the italic in x in the equation mode, but i put the variable X as an upper case):

$\hat{Y} = \hat{\beta}0 + \hat{\beta}1Xi + \hat{\beta}2x1 + \hat{\beta}3x1xXi$

Plugging in the coefficients, then the equation should take the following form:

$\hat{Y} = 21.1422589 + 0.0031709Xi + 37.7812800 + (-0.0023257)xXi$

which effectively translates into the following:

$\hat{(}Y) = 58.9235389 + (0.0008452)Xi$

(d) Interpret the coefficient for `income`.

The coefficient for income $\hat{\beta}1$ has the value of 0.0031709. Holding other terms in the equation at their observed mean, every additional 1 point jump on Pineo-Porter prestige scale is positively associated with additional \$0.0031709 of income (on average).

(e) Interpret the coefficient for `professional`.

The coefficient for professional $\hat{\beta}2$ has the value of 37.7812800. Holding other terms in the equation at their observed mean, every additional 1 point jump on Pineo-Porter prestige scale is positively associated with additional 37.7812800 score points for belonging to the category of professionals (on average).

(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).

Taking the shortened version of the interaction equation and the value of $1,000 (Xi) then the equation looks like this:

$\hat{(}Y) = 58.9235389 + (0.0008452) * 1000$ $\hat{(}Y) = 59.7687389$

Therefore the score of $\hat{(}Y)$ in case of professional categories is 59.7687389. To get a marginal effect of $1000 increase I will take the prestige score in case of $0 which is just the effect of the intercept and the belonging to the category of professionals, which is 58.9235389.

Therefore the marginal effect of additional $1000 on the prestige score equals to:

$MEF1 = 59.7687389 - 58.9235389$ $MEF1 = 0.8452$

Therefore the additional $1000 of income bring 0.8452 points on the prestige scale.

(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 marginal effect of a constant income, but the change of profession, I will plug in both values of my dummy variable in the equation, that will take the following form:

Professional: $\hat{(}Y) = 58.9235389{+}(0.0008452){*}6000 = 63.9947389$ Non-professional:$\hat{Y} = 21.1422589 + 0.0031709 * 6000 + 0 + 0 = 40.1676589$

In non-professional category, the dummy coefficient is multiplied with the value of dummy variable which takes 0 for all other categories, therefore excluding them from the equation.

The marginal effect of change of the category from non-professional to professional, while keeping the income constant is:

$MF2 = 63.9947389 - 40.1676589$ $MF2 = 23.82708$

The marginal effect tells us that the change of category from non-professional to professional is associated with 23.82708 increase on the scale of prestige.

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 determine whether having yard signs affects the vote share I will conduct a t using the regression coefficients.

If there is no impact of the yard signs then the coefficient of $\hat{\beta}2$ would have the value zero.

Therefore, the null hypothesis states that H0: $\hat{\beta}2 = 0$; on the other hand the alternative Ha:$\hat{\beta}2 \neq 0$

For that reason I am using the t test to calculate the t statistic.

$t = (\hat{\beta}2 - 0)/\hat{\sigma}(\hat{\beta}2)$

[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.

$t = (0.042 - 0)/0.016$

$t = 2.625$

With this value I will calculate the p-value and, using the n-p df, where p = to the number of parameters of the regression model (3), rendering .

The R code used to calculate the p-value:

```r
p1 <- 2*pt(2.625, 128, lower.tail = F)
p1
```

which renders the p-value = 0.00972002

If $\alpha = 0.05$, then the p value indicates there are evidence to support Ha, that is to say how having these yard signs in a precinct affects 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 $\alpha = .05$).

For this task, the logic and formulas are the same as for the previous one. H0: $\hat{\beta}3 = 0$; on the other hand the alternative Ha:$\hat{\beta}3 \neq 0$

$t = (\hat{\beta}2 - 0)/\hat{\sigma}(\hat{\beta}2)$

$t = (0.042 - 0)/0.013$

$t = 3.23076923077$

The R code used to calculate the p-value:

```r
p2 <- 2*pt(3.23076923077, 128, lower.tail = F)
p2
```

which renders the p-value = 0.00156946

If $\alpha = 0.05$, then the p value indicates there are evidence to support Ha, that is to say how having these yard signs in a precinct affects vote share.

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

If the we take that the equation of this regression has the following form:

$\hat{Y} = \hat{\beta}0 + \hat{\beta}1Xi + \hat{\beta}2Xs$

in which the response variable is the proportion of the vote that went to McAuli's opponent Ken Cuccinelli, and the response variables are Xi = a precinct was randomly assigned to have the sign against McAulie posted and Xs = precinct that was adjacent to a precinct in the treatment group. Therefore, the constant term in the equation represents the coefficient of intercept of the regression line. The $\beta 0$ coefficient tells us the proportion of votes that Ken Cuccinelli got if all other terms were zero.

I am not sure if the viable interpretation is also that for every vote that went to Cuccinelli, 0.302 votes went to McAuli, but I would say it is a good interpretation as well.

(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 measure of model fit for this model is R^2. From the table the value of $R^2 = 0.094$. The value of R^2 indicates that only a small portion (9.4%) of variance of the response variable is explained using these explanatory variables. The value of coefficients provided in the table, as well as their p-values indicate how these coefficients are valuable constitutes of explanation of Cuccinelli's votes, but the value of R^2 tells that even though the effect of lawn sign is 'statistically' significant (I tried to avoid using the term, so I used valuable in explaining the $\hat{Y}$), it represents only a small portion of explation of proportion of Cuccinelli's votes. Therefore, a lot of residual variance remained unexplained, *circa* 90%.