Problem Set 4

Applied Stats/Quant Methods 1

Due: December 3, 2023///Wei Tang 23362496

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`).

```
Prestige$professional <- ifelse(Prestige$type == "prof", 1, 0)
print(View(Prestige))
```

(Create the variable `professional` and check the result.)

(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.)

(Because that lm() function will automatically drop out NA observations, so I don't need to clean the NA observations manually.)

```
lm1<-lm(prestige~income
                +professional
                +income*professional,data=Prestige)
```

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

```
#general equation for both professional and non professional
print(paste("prestige=",lm1$coefficients[[1]],"+",
            lm1$coefficients[[2]],"*", names(lm1$coefficients)[2],"+",
            lm1$coefficients[[3]],"*", names(lm1$coefficients)[3],"+",
            lm1$coefficients[[4]],"*", names(lm1$coefficients)[4]))
#equation for both professional
print(paste("prestige=",lm1$coefficients[[1]]+lm1$coefficients[[3]],"+",
            lm1$coefficients[[2]]+lm1$coefficients[[4]],"*", names(lm1$
    coefficients)[2]))
#equation for non-professional
print(paste("prestige=",lm1$coefficients[[1]],"+",
            lm1$coefficients[[2]],"*", names(lm1$coefficients)[2]))
```

General equation:
`prestige` = 21.1422588538203 + 0.00317090909728508 * `income` + 37.7812799549884 * `professional` − 0.00232570911767063 * `income` * `professional`

Equation for professional(`professional` = 1):
`prestige`= 58.9235388088087 + 0.000845199979614447 * `income`

Equation for non-professional(`professional` = 0):
`prestige`= 21.1422588538203 + 0.00317090909728508 * `income`

(d) Interpret the coefficient for `income`.

The coefficient for `income` is representing:
1.For professionals, with every additional 1 dollar of `income`, the estimated `prestige` increases by 0.000845199979614447 scale points.

2.For non-professionals, with every additional 1 dollar of `income`, the estimated `prestige` increases by 0.00317090909728508 scale points

(e) Interpret the coefficient for `professional`.

The coefficient for `professional` is representing:
When `income` = 0, professional has an `prestige` score of 37.7812799549884 scale points higher than non-professional.

(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).

```
delta_income<-1000
delta_prestige<-lm1$coefficients[[2]]*delta_income+lm1$coefficients[[4]]*
    delta_income*1
print(delta_prestige)
```

According to the meaning of the model, when the variable professional = 1, with every additional 1000 dollars of `income`, the estimated `prestige` increases by 0.845199979614447 scale points.
$\Delta prestige = 0.00317090909728508 * \Delta income - 0.00232570911767063 * \Delta income$
$\Delta prestige = 0.00317090909728508 * 1000 - 0.00232570911767063 * 1000$
$\Delta prestige = 3.17090909728508 - 2.32570911767063$
$\Delta prestige = 0.8452$

(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).

```
delta_professional<-1
income<-6000
delta_prestige2<-lm1$coefficients[[3]]*delta_professional+lm1$
    coefficients[[4]]*income*delta_professional
print(delta_prestige2)
```

According to the meaning of the model, when the variable `income` = 6000, the marginal effect of variable `professional` on `prestige` equals the coefficient of variable `professional` deduct the coefficient of interaction variable `income * professional`.

$\Delta prestige = 37.7812799549884 * \Delta professional - 0.00232570911767063 * income * \Delta professional$
$\Delta prestige = 37.7812799549884 * 1 - 0.00232570911767063 * 6000 * 1$
$\Delta prestige = 23.82703$

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 $\alpha = .05$).

```
t_value<-0.042/0.016
p_value <- 2 * (1 - pt(abs(t_value), df=131-3))
print(p_value)
```

Because the p-value is 0.00972002$< \alpha = .05$, we can REJECT the 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 $\alpha = .05$).

[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
t_value2<-0.042/0.013
p_value2<-2 * (1 - pt(abs(t_value2), df=131-3))
print(p_value2)
```

Because the p-value = 0.00156946$< \alpha = .05$, we can REJECT the null hypothesis that being next to precincts with these yard signs does NOT affect vote share.

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

The coefficient for the constant term represents that the baseline of the impact on vote share is 0.302 when the precinct did NOT have the sign against McAuliffe posted and was also NOT adjacent to a precinct in the treatment group.

(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 modelled?

R-squared is a statistical measure used to assess the goodness of fit of a regression model to the observed data. It represents the proportion of the variance in the dependent variable (the outcome variable) that is explained by the model. R-squared values range from 0 to 1, and the closer the r-squared is to 1, the better the model fits the data.
The R square here is 0.094, which means that the majority of the variance (approximately 90.6%) in the dependent variable is not accounted for by the model and may be attributed to other factors or errors, indicating that the model does NOT fit the data well enough.