Problem Set 4/

Quant Methods 1/Due: November 18, 2024

Zengyuan Zhao / zhaoze@tcd.ie

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`).

```
install.packages("car")
library(car)
data(Prestige)
help(Prestige)
#Q1.1
#check data
paste('There is :',sum(is.na(Prestige)),'na value in dataset')
#Add a professional column and assign a value
Prestige$professional <- ifelse(Prestige$type == "prof", 1, 0)
head(Prestige,n = 30)
```

To better inspect the changed dataframe, I set the number of rows to 30.

Figure 1: The first 30 rows of the data frame

	education	income	women	prestige	census	type	professional
gov.administrators	13.11	12351	11.16	68.8	1113	prof	1
general.managers	12.26	25879	4.02	69.1	1130	prof	1
accountants	12.77	9271	15.70	63.4	1171	prof	1
purchasing.officers	11.42	8865	9.11	56.8	1175	prof	1
chemists	14.62	8403	11.68	73.5	2111	prof	1
physicists	15.64	11030	5.13	77.6	2113	prof	1
biologists	15.09	8258	25.65	72.6	2133	prof	1
architects	15.44	14163	2.69	78.1	2141	prof	1
civil.engineers	14.52	11377	1.03	73.1	2143	prof	1
mining.engineers	14.64	11023	0.94	68.8	2153	prof	1
surveyors	12.39	5902	1.91	62.0	2161	prof	1
draughtsmen	12.30	7059	7.83	60.0	2163	prof	1
computer.programers	13.83	8425	15.33	53.8	2183	prof	1
economists	14.44	8049	57.31	62.2	2311	prof	1
psychologists	14.36	7405	48.28	74.9	2315	prof	1
social.workers	14.21	6336	54.77	55.1	2331	prof	1
lawyers	15.77	19263	5.13	82.3	2343	prof	1
librarians	14.15	6112	77.10	58.1	2351	prof	1
vocational.counsellors	15.22	9593	34.89	58.3	2391	prof	1
ministers	14.50	4686	4.14	72.8	2511	prof	1
university.teachers	15.97	12480	19.59	84.6	2711	prof	1
primary.school.teachers	13.62	5648	83.78	59.6	2731	prof	1
secondary.school.teachers	15.08	8034	46.80	66.1	2733	prof	1
physicians	15.96	25308	10.56	87.2	3111	prof	1
veterinarians	15.94	14558	4.32	66.7	3115	prof	1
osteopaths.chiropractors	14.71	17498	6.91	68.4	3117	prof	1
nurses	12.46	4614	96.12	64.7	3131	prof	1
nursing.aides	9.45	3485	76.14	34.9	3135	bc	0
physio.therapsts	13.62	5092	82.66	72.1	3137	prof	1
pharmacists	15.21	10432	24.71	69.3	3151	prof	1

(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.)

```r
#Q1.2
#Linear model between prestige,income,professional and intercation.
model_interaction <- lm(prestige ~ income*professional, data = Prestige)
#check the model
summary(model_interaction)
```

Figure 2: The linear model between prestige,income, professional, and the interaction

Call:
lm(formula = prestige ~ 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.

```r
#Q1.3
#Get the coefficient and the intercept
Intercept <- coef(model_interaction)[1]
income_coef <- coef(model_interaction)[2]
professional_coef <- coef(model_interaction)[3]
income_professional_coef <- coef(model_interaction)[4]
#Write the prediction equation
paste('prestige = ',Intercept,'+',income_coef,'* income','+',
      professional_coef,'* professional',income_professional_coef,
      '* income * professional')
```

The code output is:

"prestige = 21.1422588538203 + 0.00317090909728508 * income + 37.7812799549884 * professional -0.00232570911767063 * income * professional"

Keep the equation to 5 decimal places:

prestige = 21.14226 + 0.00317 * income + 37.78128 * professional -0.00233 * income * professional

(d) Interpret the coefficient for `income`.

Ignoring the interaction effect and the coefficient of professional(coefficient=0), the coefficient of income reflects the direct impact of income on prestige. Specifically, when income increases by one unit, prestige is expected to increase by 0.0031709 units, which is the marginal effect under the premise that the level of professional remains unchanged.

However, the interaction between income and professional cannot be ignored. The coefficient of the interaction term -0.0023257 reveals how the impact of income on prestige changes with changes in the level of professional. Specifically, every time professional increases by one unit, the marginal effect of income on prestige will decrease by 0.0023257 units. This shows that as professional increases, the reinforcing effect of income on prestige gradually weakens.

(e) Interpret the coefficient for `professional`.

When the interaction between income and professional is ignored(coefficient=0), the coefficient of professional will directly reflect its marginal impact on prestige. Specifically, the coefficient of professional is 37.7812800, which means that when income remains unchanged, for every unit increase in professional, prestige will increase by an average of 37.7812800 units. This is a significant positive effect.

The interaction between income and professionalism is inevitable. The existence of the interaction term means that the impact of professional on prestige does not exist

in isolation, but will be moderated by income level. Specifically, the interaction term coefficient -0.0023257 reveals that as professionalism increases, the marginal effect of income on prestige will weaken.

(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).

The equation of question C is known to be:
prestige = 21.1423 + 0.00317 * income + 37.7812 * professional -0.00233 * income * professional
According to the question requirements, professional is 1, and the simplified equation is:
prestige=58.9235+0.000845 * income
Since the change value of income = 1000,
therefore the margin effect of prestige = (0.003171 - 0.002326) * 1000 = 0.845
According to the calculation process, we can conclude that when the professional variable value is 1:
Income marginal value = (income coefficient + interaction coefficient) * income change value
Run this formula in R:

```
#Q1.6
#income change value is 1000
income_increase <- 1000
income_margin_effect <- (income_coef + income_professional_coef) * 1000
paste("When professional is 1 and income increases", income_increase, ",
    the marginal effect of income is", income_margin_effect)
```

The code output is:

"When professional is 1 and income increases 1000 , the marginal effect of income is 0.845199979614447"

(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).

I need to calculate the prestige of professionals and non-professionals respectively, and calculate the marginal difference between them. That is, substitute income=6000, professional=1 and income=6000, professional=0 into the formula respectively, and finally calculate the difference between the prestige of professionals and non-professionals. The

calculation of this question will be completed in R.

```
#Q1.7
income <- 6000
#Calculate the prestige when people is professional
is_professional <- 1
is_professional_prestige <- Intercept + income_coef * income +
    professional_coef * is_professional + income_professional_coef *
    income * is_professional
#Calculate the prestige when people is not professional
no_professional <- 0
no_professional_prestige <- Intercept + income_coef * income +
    professional_coef * no_professional + income_professional_coef *
    income * no_professional
#Calculate the  change in y
margin_effect_prestige <- is_professional_prestige - no_professional_
    prestige
paste("When a person's income is",income,"switching from a
    nonprofessional to a professional occupation will result in an
    increase in her prestige of approximately",margin_effect_prestige, "
    units.")
```

The code output is:

"When a person's income is 6000 switching from a nonprofessional to a professional occupation will result in an increase in her prestige of approximately 23.8270252489646 units."

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 $\alpha = .05$).

According to the linear regression results, the coefficient of "Precinct assigned lawn signs (n=30)" is 0.042, and the standard error is 0.016. The coefficient of "Precinct adjacent to lawn signs (n=76)" is also 0.042, and the standard error is 0.013. The constant term (intercept) is 0.302, and its standard error is 0.011.

Setting hypothesis

Null hypothesis (H0): Placing campaign placards has no effect on voting proportion, that is, the coefficient of the Precinct assigned lawn signs variable is equal to 0.
Alternative hypothesis (H1): Placing campaign placards has an effect on voting proportion, that is, the coefficient of the Precinct assigned lawn signs variable is not equal 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.

Next, I will use the t-test method in R to judge whether to accept the null hypothesis.

```
#Q2.1
# we can get the coefficient, standard error and a.
assigned_signs_coef <- 0.042
assigned_signs_se <- 0.016
alpha <- 0.05
# Calculate t value
assigned_signs_t <- assigned_signs_coef / assigned_signs_se
# Calculate critical value
critical_value <- qt(1 - alpha/2,df = 128)
# Comparing t-values to critical values
if (assigned_signs_t > critical_value) {
  paste("At a significance level of ", alpha, ", the null hypothesis is
    rejected.So, there is sufficient evidence that put these yard signs in
     a precinct has an effect on the vote share.")
} else {
  paste("At a significance level of ", alpha, ", the null hypothesis can'
    t be rejected.So, There is no sufficient evidence that There is
    sufficient evidence that put these yard signs in a precinct has an
    effect on the vote share has an effect on the vote share.")
}
```

The code output is:

"At a significance level of 0.05 , the null hypothesis is rejected. So, there is sufficient evidence that put these yard signs in a precinct has an effect on the vote share."

The t-test showed that the t-value(2.625) was greater than the critical value(1.978671), so the null hypothesis was rejected, indicating that there was sufficient evidence to show that placing signs in the precinct had an impact on the voting 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 $\alpha = .05$).

Setting hypothesis

Null Hypothesis (Ha): Being next to precincts with yard signs does not affect vote share.

Alternative Hypothesis(Hb): Being next to precincts with yard signs does affect vote share.

Next, I will use the t-test method in R to judge whether to accept the null hypothesis.

```
#Q2.2
adjacent_signs_coef <- 0.042
adjacent_signs_se <- 0.013
# Calculate t value
adjacent_signs_t <- adjacent_signs_coef / adjacent_signs_se
# Calculate critical value
critical_value <- qt(1 - alpha/2,df = 128)
# Comparing t-values to critical values
if (adjacent_signs_t > critical_value) {
  paste("At a significance level of ", alpha, ", the null hypothesis is
    rejected.So, there is sufficient evidence that being next to precincts
     with these yard signs has an effect on the vote share.")
} else {
  paste("At a significance level of ", alpha, ", the null hypothesis can'
    t be rejected.So, there is no sufficient evidence that  being next to
    precincts with these yard signs has an effect on the vote share.")
}
```

The code output is:

"At a significance level of 0.05 , the null hypothesis is rejected. So, there is sufficient evidence that being next to precincts with these yard signs has an effect on the vote share."

The t-test showed that the t-value (3.230769) was greater than the critical value (1.978671), so the null hypothesis was rejected, indicating that there was sufficient evidence to show that being near the precinct with signs had an impact on the voting share.

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

In the framework of linear regression analysis, the coefficient of the constant term (or intercept term) represents the conditional expected value of the dependent variable "vote share" when all independent variables take their baseline values (usually 0,

which is the case when the binary variables are not activated or do not exist). In this particular problem, the constant term coefficient (0.302) represents the expected level of vote share in the corresponding constituency when the two binary independent variables "constituency assigned with propaganda signs" and "constituency adjacent to propaganda signs" are both set to 0 (that is, when neither of these two conditions is met).

At the same time, the standard error of the constant term coefficient is 0.011, which is relatively small, indicating that the estimate of the constant term coefficient is relatively stable and reliable. The closer the standard error is to 0, it usually means that the estimation is more precise and the confidence interval is narrower, so we can be more confident in the estimate of the constant term coefficient.

(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, the coefficient of determination, is the core indicator for measuring the fitting effect of the regression model. It represents the proportion of variability in the dependent variable explained by the independent variables in the model. $R^2 = 0.094$: This shows that the independent variable in the model (i.e., constituencies assigned propaganda signs versus constituencies adjacent to propaganda signs) can only explain a small 9.4 of the total variation in vote share. Therefore, the model's ability to explain the data is weak.

The remaining 90.6 of the variability is attributed to various other factors not included in the model. Although the two variables of precinct assigned signs and precinct adjacent to signs have a statistically significant relationship with the opponent's vote share, they account for a relatively small proportion of the total variation in vote share explained, only 9.4. This suggests that in addition to these two variables, there are other more important factors affecting vote share.

Other factors may include the candidate's financial investment, the candidate's policy advocacy, the voter's age, education level, etc.