Problem Set 4

Applied Stats/Quant Methods 1

Maiia Skrypnyk 23371609

Due: 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`).

```
Prestige$professional <- ifelse(Prestige$type == "prof", 1, 0)
Prestige$professional <- factor(Prestige$professional)
head(Prestige, 30)
```

Table 1: 'Prestige' Dataset Header

Occupation	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 Programmers	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
Physiotherapists	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.)

```

m1 <- lm(prestige ~ income + professional + income:professional, data =
    Prestige)
```

Exploring the model's summary and using 'stargazer' function to format results:

```
summary(m1)
stargazer(m1)
```

Table 2: Linear Model 1

	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.

A general formula for such equation would be:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 X_i D_i + \varepsilon_i$$

- Y_i = predicted value of the response variable
- β_0 = estimated intercept (the expected value of Y when X=0 and D=0)
- β_1 = estimated slope coefficient (the change in Y when X increases by one unit, when D=0)
- β_2 = estimated slope coefficient (the change in Y when D increases by one unit, when X=0)
- β_3 = estimated slope coefficient (the interaction term of X and D)
- X_i = (any chosen) value of the explanatory variable
- D_i = (any chosen) value of the moderator variable (= dummy variable, in our case)

Therefore, a prediction equation for Model 1 will look like this:

$$\hat{\boldsymbol{Prestige}}_i = \hat{\beta}_0 + \hat{\beta}_1 Income_i + \hat{\beta}_2 Professional_i + \hat{\beta}_3 Income_i Professional_i + \varepsilon_i$$

$$\hat{\boldsymbol{Prestige}}_i = 21.142 + (0.003 \cdot Income_i) + (37.781 \cdot Professional_i) - (0.002 \cdot Income_i \cdot Professional_i) + \varepsilon_i$$

(d) Interpret the coefficient for `income`.

1) Holding 'professionalism' (= the '*Professional*' variable) constant, a one unit (= one US dollar) increase in income is, on average, associated with a 0.003 scale point increase in prestige.

2) For non-professionals (blue and white collar workers; *Professional* = 0), a one unit (= one US dollar) increase in income is, on average, associated with a 0.003 scale point increase in prestige.

This estimated coefficient is statistically differentiable from 0 (= significant) at the α level = 0.01.

(e) Interpret the coefficient for `professional`.

1) Holding income constant, being a professional (*Professional* = 1) is associated with having 37.781 scale points of prestige higher as compared to a non-professional (*Professional* = 0).

2) Holding income constant, a change from being a non-professional to becoming a professional (*Professional* goes from 0 to 1) is associated with, on average, 37.781 scale points increase of prestige.

3) For professionals with zero income, prestige is estimated as being, on average, by 37.781 scale points higher in comparison with non-professionals with zero income.

This estimated coefficient is statistically differentiable from 0 (= significant) at the α level = 0.01.

(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).

1) Given Y = Prestige, D = Professional = 1, X = Income = 0:

$$\hat{y}_0 = \hat{\beta}_0 + (\hat{\beta}_1 \cdot 0) + (\hat{\beta}_2 \cdot 1) + (\hat{\beta}_3 \cdot 0 \cdot 1)$$

$$\hat{y}_0 = 21.142 + (0.003 \cdot 0) + (37.781 \cdot 1) + (-0.002 \cdot 0 \cdot 1)$$

$$\hat{y}_0 = 21.142 + 37.781 = 58.923$$

2) Given Y = Prestige, D = Professional = 1, X = Income = 1000:

$$\hat{y}_1 = \hat{\beta}_0 + (\hat{\beta}_1 \cdot 0) + (\hat{\beta}_2 \cdot 1) + (\hat{\beta}_3 \cdot 0 \cdot 1)$$

$$\hat{y}_1 = 21.142 + (0.003 \cdot 1000) + (37.781 \cdot 1) + (-0.002 \cdot 1000 \cdot 1)$$

$$\hat{y}_1 = 21.142 + 3 + 37.781 - 2 = 59.923$$

3) Therefore, we can calculate the change in $\hat{y}$ as the difference between the two:

$\hat{y}_1 - \hat{y}_0$ = 59.923 - 58.823 = 1

4) The other way to calculate this value would be:

Change in $\hat{y} = (\hat{\beta}_1 + \hat{\beta}_3) \cdot \Delta X$ (where ΔX, or change in income, is \$1,000)

Change in $\hat{y} = (0.003 - 0.002) \cdot 1000 = 1$

So, for professional occupations, on average, the change in $\hat{y}$ associated with a \$1,000 increase in income equals 1 scale point of prestige.

(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).

1) Given Y = Prestige, D = Professional = 0, X = Income = 6000:

$$\hat{y}_0 = \hat{\beta}_0 + (\hat{\beta}_1 \cdot 6000) + (\hat{\beta}_2 \cdot 0) + (\hat{\beta}_3 \cdot 6000 \cdot 0)$$

$$\hat{y}_0 = 21.142 + (0.003 \cdot 6000) + (37.781 \cdot 0) + (-0.002 \cdot 6000 \cdot 0)$$

$$\hat{y}_0 = 21.142 + 18 = 39.142$$

2) Given Y = Prestige, D = Professional = 1, X = Income = 6000:

$$\hat{y}_1 = \hat{\beta}_0 + (\hat{\beta}_1 \cdot 0) + (\hat{\beta}_2 \cdot 1) + (\hat{\beta}_3 \cdot 0 \cdot 1)$$

$$\hat{y}_1 = 21.142 + (0.003 \cdot 6000) + (37.781 \cdot 1) + (-0.002 \cdot 6000 \cdot 1)$$

$$\hat{y}_1 = 21.142 + 18 + 37.781 - 12 = 64.923$$

3) Therefore, we can calculate the change in $\hat{y}$ as the difference between the two:

$\hat{y}_1 - \hat{y}_0$ = 64.923 - 39.142 = 25.781

4) The other way to calculate this value would be:

Change in $\hat{y} = \hat{\beta}_2 + \hat{\beta}_3 \cdot Income_i$ (where $Income_i$ is \$6,000)

Change in $\hat{y} = 37.781 - 0.002 \cdot 6000 = 37.781 - 12 = 25.781$

So, holding income constant at \$6,000, a change from being a non-professional to becoming a professional (*Professional* goes from 0 to 1) is associated with, on average, 25.781 scale points increase 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

[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.

Before we start analysing the model, let's annotate the coefficients:

- $\hat{\beta}_0$ (constant/intercept) = 0.302 (SE = 0.011)
- $\hat{\beta}_1$ (estimated slope coefficient) = 0.042 (SE = 0.016)
- $\hat{\beta}_2$ (estimated slope coefficient) = 0.042 (SE = 0.013)

(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 $\alpha = .05$).

As we need to determine the partial effect of individual regression coefficients, I am using a Student's t-test to conduct a hypothesis test in both (a) and (b).

1) **Assumptions:** categorical data, randomly obtained, normally distributed.

2) **Null and Alternative Hypotheses:**

- $H_0 : \hat{\beta}_1 = 0$ (There is NO association between the proportion of vote share that went to Cuccinelli and whether a precinct was randomly assigned to display signs against McAuliffe);
- $H_a : \hat{\beta}_1 \neq 0$ (There IS an association...).

3) **T-statistic** is given by:

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_0}{SE}$$

$$t = \frac{0.042 - 0}{0.016} = 2.625$$

4) **P-value:**

- two-tailed probability from t-distribution;
- DF = n - k - p = 131 - 2 - 1 = 128;
- $p = 2 \times Pr(t_{128} > |2.625|) = 0.0097$

```
t1 <- 2.625
df <- 128
alpha <- 0.05

#Calculating the critical t-value for a two-tailed test
critical_t1 <- qt(1 - alpha/2, df)

#Calculating the p-value
p_value1 <- 2 * (1 - pt(abs(t1), df))
```

Interpretation: Since the p-value of 0.0097 is less than the significance level of 0.05, we have found statistically significant evidence to **reject** the null hypothesis that there is no association between the proportion of vote share that went to Cuccinelli and whether a precinct was randomly assigned to display signs against McAuliffe (after controlling for the effects of the other predictor variable in the model).

(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) **Assumptions:** categorical data, randomly obtained, normally distributed.

2) **Null and Alternative Hypotheses:**

- $H_0 : \hat{\beta}_2 = 0$ (There is NO association between the proportion of vote share that went to Cuccinelli and whether a precinct was adjacent to the precincts that were randomly assigned to display signs against McAuliffe);
- $H_a : \hat{\beta}_2 \neq 0$ (There IS an association...).

3) **T-statistic** is given by:

$$t = \frac{\hat{\beta}_2 - \hat{\beta}_0}{SE}$$

$$t = \frac{0.042 - 0}{0.013} \approx 3.23$$

4) **P-value:**

- two-tailed probability from t-distribution;
- DF = n - k - p = 131 - 2 - 1 = 128;
- $p = 2 \times Pr(t_{128} > |3.23|) = 0.00157 \approx 0.0016$

```r
t2 <- 3.23

#Calculating the critical t-value for a two-tailed test
critical_t2 <- qt(1 - alpha/2, df)

#Calculating the p-value
p_value2 <- 2 * (1 - pt(abs(t2), df))
```

Interpretation: Since the p-value of ≈ 0.0016 is less than the significance level of 0.05, we have found statistically significant evidence to **reject** the null hypothesis that there is no association between the proportion of vote share that went to Cuccinelli and whether a precinct was adjacent to the precincts that were randomly assigned to display signs against McAuliffe (after controlling for the effects of the other predictor variable in the model)..

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

The coeffiecient for the constant term (intercept) represents the expected proportion of vote share (0.302) that went to Cuccinelli in precincts that neither had yard signs nor were adjacent to the ones that had yard signs (control group with no influence of external factors); 'baseline' 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?

To evaluate the model fit, we should pay attention to the correlation coefficient $R^2 = 0.094$. It's value is positive though quite low (compared to max = 1), which suggests that only 9.4% of variability in vote share for Cuccinelli is explained by precincts having or being by implication exposed to the yard signs. While the effect of signs on vote share is statistically significant, there are still a lot of other factors that should have been accounted for in the model to be fully comprehensive. Such factors might be gender, race, age, religiosity, income, occupation, etc., etc...