

Problem Set 4 Clare Zureich

Applied Stats/Quant Methods 1

Due: November 18, 2024

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)
```

```
1 install.packages("car")
2 library("car")
3 data(Prestige)
4 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`).

```
1 #Part A New Variable
2 Prestige$professional <- as.factor(ifelse(Prestige$type=="prof", 1,
3     ifelse(Prestige$type=="bc" | Prestige$type=="wc", 0, NA)))
4
5 levels(Prestige$professional)
6 summary(Prestige$professional)
```

- (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.)

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

Figure 1: Income and Professional Interaction Regression.

Call:

```
lm(formula = prestige ~ income + professional + 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 ***
professional1	37.7812800	4.2482744	8.893	4.14e-14 ***
income:professional1	-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.

$$E(y) = \beta_0 + \beta_1 \times \text{Income} + \beta_2 \times \text{Professional} + \beta_3 \times (\text{Income} \times \text{Professional})$$

$$E(y) = 21.142 + .003 \times \text{Income} + 37.781 \times \text{Professional} - .002 \times (\text{Income} \times \text{Professional})$$

- (d) Interpret the coefficient for **income**.

When someone does not have a professional career, a one unit increase in income (1 dollar), will on average result in a .003 increase in prestige

- (e) Interpret the coefficient for **professional**.

When income is 0, the change from not-professional (blue or white collar workers) to professional will, on average, result in a 37.781 increase in prestige. This is the difference in prestige between the groups (professional vs white/blue collar worker) when income is 0.

- (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).

$$E(y_1) = 21.142 + (.003 \times 1000) + (37.781 \times 1) - (.002 \times 1000 \times 1) = 59.923$$

$$E(y_0) = 21.142 + (.003 \times 0) + (37.781 \times 1) - (.002 \times 0 \times 1) = 58.923$$

$$E(y_1) - E(y_0) = 1$$

There is a 1 unit increase in prestige when income increases from \$0 to \$1000 for individuals with professional careers. Professional individuals with an income of \$1,000 have a prestige of 59.923. Professional individuals with an income of \$0 have a prestige of 58.923.

```
1 #Prediction equation: prestige = 21.1422 + .003*income + 37.781*
  professional - .002(income*professional)
2 income_change <- 21.1422 + (.003*1000) + (37.781*1) - (.002*1000*1)
3 income_change
4 income_change2 <- 21.1422 + (.003*0) + (37.781*1) - (.002*0*1)
5 income_change2
6 prestige_change = income_change - income_change2
7 prestige_change
```

- (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).

$$E(y_1) = 21.142 + (.003 \times 6000) + (37.781 \times 1) - (.002 \times 6000 \times 1) = 64.923$$

$$E(y_0) = 21.142 + (.003 \times 6000) + (37.781 \times 0) - (.002 \times 6000 \times 0) = 39.142$$

$$E(y_1) - E(y_0) = 25.781$$

With an income of \$6,000, a change from non-professional to professional will increase prestige by 25.781, on average.

```
1 professional_change <- 21.1422 + (.003*6000) + (37.781*1) - (.002*6000*1)
2 professional_change
3 professional_change2 <- 21.1422 + (.003*6000) + (37.781*0) - (.002*6000*
  0)
4 professional_change2
5 prestige_change2 <- professional_change - professional_change2
6 prestige_change2
```

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 Virginia 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 McAuliffe’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 $\alpha = .05$).

Assumptions for multiple regression:

1. Samples from the population are independent and random samples
2. Population distribution of the response variable for each group is normal
3. Standard deviation of the population distribution is constant

¹Donald 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.

Null hypothesis: H_0 : The yard signs have no affect on vote share ($\beta_1 = 0$)
Alternative hypothesis: H_a : The yard signs have an affect on vote share ($\beta_1 \neq 0$)

Test statistic = Estimate/SE = $.042/.016 = 2.625$

P-value = 0.00972002

```
1 #Find the test statistic
2 t_stat <- .042/.016
3 t_stat
4
5 #Find the p-value
6 df <- 131-2-1
7 p_value <- 2*pt(t_stat, df, lower.tail = FALSE)
8 p_value
```

Conclusion: Based on the p-value of 0.00972002 and a .05 significance level, there is sufficient evidence to reject the null hypothesis that the yard signs have no affect on vote share. Therefore, a one unit increase in yard signs will, on average, increase vote share by .042.

- (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 $\alpha = .05$).

Assumptions for multiple regression:

1. Samples from the population are independant and random samples
2. Population distribution of the response variable for each group is normal
3. Standard deviation of the population distribution is constant

Null hypothesis: H_0 : The yard signs in adjacent precincts have no affect on vote share ($\beta_2 = 0$)

Alternative hypothesis: H_a : The yard signs in adjacent precincts have an affect on vote share ($\beta_2 \neq 0$)

Test statistic = Estimate/SE = $.042/.013 = 3.230769$

P-value = 0.00156946

```
1 t_stat_adjacent <- .042/.013
2 t_stat_adjacent
3 p_value_adjacent <- 2*pt(t_stat_adjacent, df, lower.tail = FALSE)
4 p_value_adjacent
```

Conclusion: Based on the p-value of 0.00156946 and a .05 significance level, there is sufficient evidence to reject the null hypothesis that the adjacent precincts with yard signs have no affect on vote share. Therefore, a one unit increase in yard signs in adjacent precincts will, on average, increase vote share by .042.

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

The baseline group, a district with no assigned lawn signs and not adjacent to a lawn sign precinct, is expected to have a vote share of .302 per Cuccinelli. The t-stat for the constant is 27.45, which is large and significant when compared to the test statistics calculated for the precinct assigned lawn signs and precinct adjacent to lawn signs coefficients.

```
1 t_stat_constant <- .302/.011
```

- (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 R^2 for this regression is .094. Given that R^2 will always fall between 0 and 1, this is a very small R^2 . Only 9.4% the variation in vote share is explained by the yard signs in or adjacent to precincts. There are other factors that need to be accounted for when trying to explain vote share besides just yard sign placement.