Problem Set 4

Applied Stats/Quant Methods 1

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`).

```
# Start setting the working directory
setwd("C:/Users/User/Documents/GitHub/StatsI_Fall2023/problemSets/PS04/
    template")

# Install and load the car library
install.packages("car")
library(car)

# Load the Prestige dataset
data(Prestige)

# Display the help documentation for Prestige dataset
help(Prestige)

# a)

# Create a new variable "professional" by recoding the "type" variable
Prestige$professional <- ifelse(Prestige$type == "prof", 1, 0)

# View the updated dataset
head(Prestige)
```

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

(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.)

```
# Now we will run a linear model
model <- lm(prestige ~ income * professional, data = Prestige)

# Now we can display the summary of the model
summary(model)
```

Variable	Estimate	Std. Error	t value	Pr(>—t—)
Intercept	21.1423	2.8044	7.539	2.93e-11 ***
Income	0.0032	0.0005	6.351	7.55e-09 ***
Professional	37.7813	4.2483	8.893	4.14e-14 ***
Income:Professional	-0.0023	0.0006	-4.098	8.83e-05 ***

Residual Standard Error	:	8.012 on 94 degrees of freedom
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.

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

(d) Interpret the coefficient for `income`.

The coefficient for "income" represents the estimated variation in the dependent variable ("prestige") for a one-unit increase in the independent variable ("income"), while keeping the other variables constant. In this case, since the coefficient is positive (0.0031709), it suggests a positive relationship between income and prestige. That is, as income increases, it is expected that prestige will also increase.

(e) Interpret the coefficient for `professional`.

The coefficient for "professional" represents the estimated difference in the dependent variable ("prestige") between the reference category and the category of interest, while holding other variables constant. Given that this coefficient is positive, it suggests that individuals classified as "professionals" have, on average, a prestige level approximately 37.78 units higher than those classified as "non-professionals".

(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).

```r
# We will extract coefficients from the model
coefficients <- coef(model)

# Then extract the coefficient for income and its interaction term
beta_income <- coefficients["income"]
beta_interaction <- coefficients["income:professional"]

# Now we can calculate the change in y associated with a $1,000 increase in income for professional occupations
change_in_y <- beta_income + beta_interaction * 1000

# Finally, print the result
cat("The change in prestige associated with a $1,000 increase in income for professional occupations is:", round(change_in_y, 2), "\n")
```

$$\text{Prestige} = 21.14 + 0\ \times\text{Income} + 37.78 \times \text{Professional} + 0 \times (\text{Income} \times \text{Professional})$$

The change in prestige associated with a $ 1,000 increase in income for professional occupations is: $-2.32.

(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).

```r
# To calculate it, we will use the coefficient for the professional variable and its interaction term with income from the linear model

# First off all, extract coefficients from the model
coefficients <- coef(model)

# Now we will extract the coefficient for the professional variable and its interaction term with income
beta_professional <- coefficients["professional"]
beta_interaction <- coefficients["income:professional"]

# Then set the income value to $6,000
income_value <- 6000

# Now we calculate the change in y associated with changing occupation to professional at $6,000 income
change_in_y <- beta_professional + beta_interaction * income_value

# Print the result
cat("The change in prestige associated with changing occupation to professional at $6,000 income is:", round(change_in_y, 2), "\n")
```

The change in prestige associated with changing occupation to professional at $6,000 income is: 23.83.

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 $\alpha = .05$).

```
# We will conduct hypothesis tests on the coefficients of the variables indicating the presence of the signs and we will use the t-statistics and p-values in the repression.


# The regression results
```

[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
assigned_signs_coef <- 0.042
assigned_signs_se <- 0.016

adjacent_signs_coef <- 0.042
adjacent_signs_se <- 0.013

# Degrees of freedom
df <- 131 - 3  # 3 coefficients: assigned_signs, adjacent_signs, constant

# The critical value for a two-tailed test at alpha = 0.05
critical_value <- qt(1 - 0.05 / 2, df)

# T-statistics
assigned_signs_t <- assigned_signs_coef / assigned_signs_se
adjacent_signs_t <- adjacent_signs_coef / adjacent_signs_se

# P-values
assigned_signs_p_value <- 2 * pt(-abs(assigned_signs_t), df)
adjacent_signs_p_value <- 2 * pt(-abs(adjacent_signs_t), df)

# Print results
cat("Assigned Signs:\n")
cat("  Coefficient:", assigned_signs_coef, "\n")
cat("  Standard Error:", assigned_signs_se, "\n")
cat("  t-Statistic:", assigned_signs_t, "\n")
cat("  P-Value:", assigned_signs_p_value, "\n")
cat("  Reject Null Hypothesis:", assigned_signs_p_value < 0.05, "\n")

cat("\nAdjacent Signs:\n")
cat("  Coefficient:", adjacent_signs_coef, "\n")
cat("  Standard Error:", adjacent_signs_se, "\n")
cat("  t-Statistic:", adjacent_signs_t, "\n")
cat("  P-Value:", adjacent_signs_p_value, "\n")
cat("  Reject Null Hypothesis:", adjacent_signs_p_value < 0.05, "\n")
```

Assigned Signs:

- Coefficient: 0.042
- Standard Error: 0.016
- t-Statistic: 2.625
- P-Value: 0.00972002
- Reject Null Hypothesis: **TRUE**

Adjacent Signs:

- Coefficient: 0.042
- Standard Error: 0.013
- t-Statistic: 3.230769

- P-Value: 0.00156946
- Reject Null Hypothesis: **TRUE**

(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 $\alpha = .05$).

```r
 # Similar to the exercise A, we will conduct a hypothesis test on the coefficient of the variable indicating adjacent signs and we will use the t-statistic and p-value reporter in the regresssion.


# The values from regression results
adjacent_signs_coef <- 0.042
adjacent_signs_se <- 0.013

# Degrees of freedom
df <- 131 - 3  # 3 coefficients: assigned_signs, adjacent_signs, constant

# Critical value for a two-tailed test at alpha = 0.05
critical_value <- qt(1 - 0.05 / 2, df)

# T-statistic
adjacent_signs_t <- adjacent_signs_coef / adjacent_signs_se

# P-value
adjacent_signs_p_value <- 2 * pt(-abs(adjacent_signs_t), df)

# Print results
cat("Adjacent Signs:\n")
cat("  Coefficient:", adjacent_signs_coef, "\n")
cat("  Standard Error:", adjacent_signs_se, "\n")
cat("  t-Statistic:", adjacent_signs_t, "\n")
cat("  P-Value:", adjacent_signs_p_value, "\n")
cat("  Reject Null Hypothesis:", adjacent_signs_p_value < 0.05, "\n")
```

Adjacent Signs:

- Coefficient: 0.042
- Standard Error: 0.013
- t-Statistic: 3.230769
- P-Value: 0.00156946
- Reject Null Hypothesis: **TRUE**

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

The constant term (0.302) represents the estimated proportion of votes that went to McAuliffe's opponent, Ken Cuccinelli, when all other variables in the model are set to zero. In other words, the estimated proportion of votes that went to Cuccinelli is approximately 0.302, with a standard error of 0.011.

(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 value of 0.094 in the regression results indicates that the model explains approximately 9.4% of the variability in the proportion of votes that went to McAuliffe's opponent, Ken Cuccinelli. The low R^2 value suggests that the included variables, such as yard signs, collectively account for only a small proportion of the variation in vote share. This implies that there are other unmodeled factors or variables that significantly contribute to the variability in vote share, and the current model does not capture a substantial portion of these factors.