Problem Set 4

Rafaela Alves

November 24, 2024

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`).

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

	education	income	women	prestige	census	type	professional
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
medical.technicians	12.79	5180	76.04	67.5	3156	wc	0
commercial.artists	11.09	6197	21.03	57.2	3314	prof	1
radio.tv.announcers	12.71	7562	11.15	57.6	3337	wc	0

(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.)

In order to run a linear model with a dummy interaction as a predictor, I created an interaction term called `inter_1` (`income` * `professional`):

```
Prestige$inter_1 <- Prestige$income * Prestige$professional
```

	education	income	women	prestige	census	type	professional	inter_1
nurses	12.46	4614	96.12	64.7	3131	prof	1	4614
nursing.aides	9.45	3485	76.14	34.9	3135	bc	0	0
physio.therapsts	13.62	5092	82.66	72.1	3137	prof	1	5092
pharmacists	15.21	10432	24.71	69.3	3151	prof	1	10432
medical.technicians	12.79	5180	76.04	67.5	3156	wc	0	0
commercial.artists	11.09	6197	21.03	57.2	3314	prof	1	6197
radio.tv.announcers	12.71	7562	11.15	57.6	3337	wc	0	0

Running the linear model including the interaction term:

```
linear_model1 <- lm(prestige ~ income + professional + inter_1, data=
    Prestige)
summary(linear_model1)
```

```
Call:
lm(formula = prestige ~ income + professional + inter_1, 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 ***
inter_1      -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.

prestige = 21.142 + 0.003 * income + 37.781 * professional - 0.002 * inter_1

(d) Interpret the coefficient for `income`.

The coefficient for `income` (0.003) represents the effect of `income` on `prestige` when professional is = 0 (blue-collar or white-collar workers). For every 1-dollar-unit increase in `income`, the `prestige` increases 0.003 for blue-collar or white-collar.

But when `professional` = 1, the effect of `income` on `prestige` changes because of the interaction term coefficient (-0.002 for `inter_1`). In this case, the combined effect of `income` for `professionals` is: 0.003 - 0.002 = 0.001.

For 'bc' and 'wc' workers, each additional dollar in income increases prestige by 0.003. For professionals, each additional dollar in income increases prestige by 0.001.

(e) Interpret the coefficient for `professional`.

Being a `professional` (professional = 1) adds 37.781 points to the `prestige` score compared to blue-collar or white-collar (professional = 0), when holding `income` constant. In another words, when `income` = 0, `professionals` have 37.781 points higher `prestige` than blue-collar or white-collar workers.

(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).

As we've seeing before in letter (d), for `professionals` (when professional = 1) each additional 1-unit-dollar in `income` increases `prestige` by 0.001 points.

If we consider $1,000 dollars increase in `income`, we have: 1.000 * 0.001 = 1.

We can conclude that $1,000 dollars increase in `income` is associated with a 1 point increase in the `prestige` score.

(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).

Given the prediction equation for our linear model:

prestige = 21.142 + 0.003 * income + 37.781 * professional - 0.002 * inter_1

First, we need to calculate $\hat{y}$ for a non-professional (professional = 0). When professional = 0, the interaction term inter_1 is also zero:

prestige (non-professional) = 21.142 + 0.003 * 6.000 = 21.142 + 18 = 39.142

Now, we calculate $\hat{y}$ for a professional (professional = 1) When professional = 1, we include the interaction term inter_1:

prestige (professional) = 21.142 + 0.003 * 6.000 + 37.781 - 0.002 *(6.000) = 21.142 + 18 + 37.781 - 12 = 64.923

The marginal effect is the difference between the two prestige scores:

prestige (professional) - prestige (non-professional) = 64.923 - 39.142 = 25.781

Conclusion: if we change the occupation from a non-professional to a professional when `income` is \$6.000, it means an increase of 25.781 points in the `prestige` score.

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 $\alpha = .05$).

Setting the hypotheses:

Null Hypothesis: Yard signs have no effect on vote share

Alternative Hypothesis: Yard signs affect vote share

```
# coefficient and standard error for the variable "Precinct assigned lawn
    signs":
coef_assnglawn <- 0.042
se_assnglawn <- 0.016

```

[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
# calculating t-statistic
t_stat <- coef_assnglawn / se_assnglawn
cat("t-statistic:", t_stat, "\n")

# calculating two-tailed p-value using the t-distribution and degrees of freedom
n <- 131
predictors <- 2
df <- n - predictors - 1  # degrees of freedom

p_value <- 2 * (1 - pt(abs(t_stat), df = df))
cat("p-value:", p_value, "\n")
```

Conclusion: As the p-value is < 0.05, we can reject the Null Hypothesis and affirm that lawn signs have a significant effect on 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 $\alpha = .05$).

Setting the hypotheses:

Null Hypothesis: being next to precincts with yard signs have no effect on vote share

Alternative Hypothesis: being next to precincts with yard signs affect vote share

```r
# coefficient and standard error for the variable "Precinct adjacent to lawn signs":
coef_adjacentlawn <- 0.042
se_adjacentlawn <- 0.013

# calculating t-statistic
t_stat2 <- coef_adjacentlawn / se_adjacentlawn
cat("t-statistic:", t_stat2, "\n")

# calculating two-tailed p-value using the t-distribution and degrees of freedom
n <- 131
predictors <- 2
df <- n - predictors - 1  # degrees of freedom

p_value2 <- 2 * (1 - pt(abs(t_stat2), df = df))
cat("p-value:", p_value2, "\n")
```

Conclusion: As the p-value is < 0.05, we can reject the Null Hypothesis. There is sufficient evidence to conclude that being in a precinct adjacent to lawn signs significantly affects vote share at the 5% significance level.

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

The constant term (intercept) represents the predicted value of the dependent variable when all predictors are set to 0. Considering the regression:

vote share = 0.302 + 0.042 * assigned signs + 0.042 * adjacent signs, we conclude that the predicted vote share proportion for McAuliff's opponent, Ken Cuccinelli, in a precinct with no lawn signs (assigned or adjacent) is expected to be 30.2%.

(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?

Considering the data that was given in the 'Impact of lawn signs on vote share' table, we can use the R^2 value (0.094) to check if the model fits the regression. The $R^2 = 0.094$ means that only 9.4% of the variation in vote share is explained by the presence of lawn signs and adjacency to precincts with lawn signs. The remaining 90.6% (100% - 9.4%) of the variation in vote share can be explained by other factors not included in the model, such as campaign strategies, demographics issues, etc.

Even though lawn signs and adjacency significantly affect vote share as we saw before, their overall impact is small and they are not a major factor in determining vote share.