Applied Statistical Analysis - ProblemSet 4

Eoghan O'Sullivan

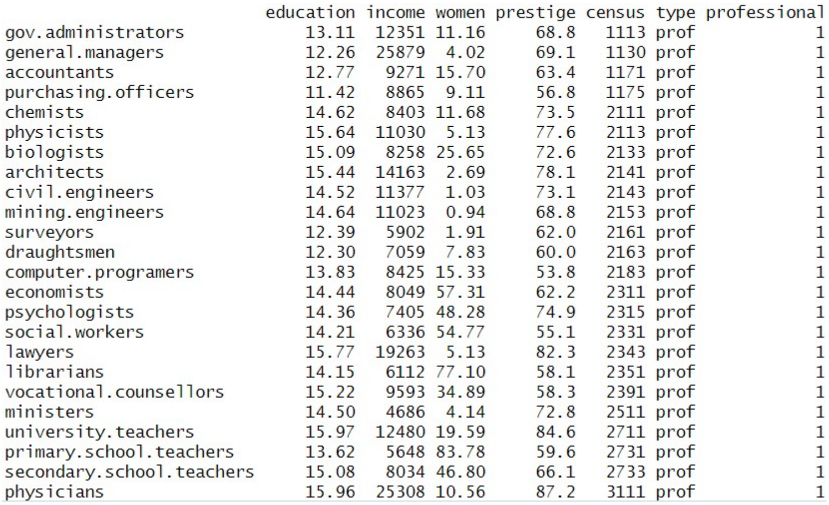
2024-11-17

1 Question 1

I read in the data from its library and made it a data frame called "df" using R statistical software. From there, I wrangled the data and below are the results I produced using the object df. This was done to study whether individuals with higher levels of income have more prestigious jobs than blue and white collar workers, in the context of Canadian occupations.

1.1 Part a

Create a new variable professional by recoding the variable type so that professionals are coded as 1, and the blue and white collar workers are coded as 0. (Hint: ifelse). The method I used was ifelse applied to the type variable with contained missing data, prof, bc and wc. I ignored the missing data from being coded as 1/0. The bc and wc were 0 and prof accordingly, 1. Figure 1 shows these data.

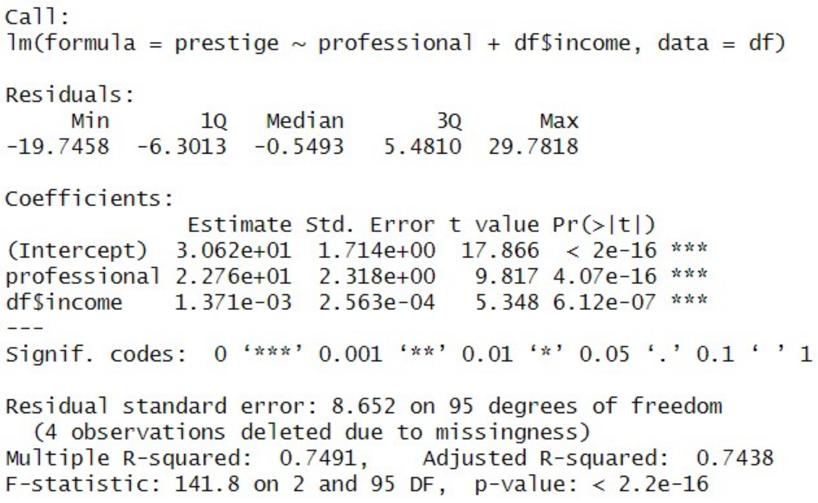


```
                          education income women prestige census type professional
gov.administrators            13.11  12351 11.16     68.8   1113 prof            1
general.managers              12.26  25879  4.02     69.1   1130 prof            1
accountants                   12.77   9271 15.70     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
```

Figure 1: Data

1.2 Part 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 x dummy interaction.) The results of this are shown in the Figure 2.



```
Call:
lm(formula = prestige ~ professional + df$income, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-19.7458  -6.3013  -0.5493   5.4810  29.7818 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.062e+01  1.714e+00  17.866  < 2e-16 ***
professional 2.276e+01  2.318e+00   9.817 4.07e-16 ***
df$income    1.371e-03  2.563e-04   5.348 6.12e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.652 on 95 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.7491,	Adjusted R-squared:  0.7438 
F-statistic: 141.8 on 2 and 95 DF,  p-value: < 2.2e-16
```

Figure 2: Regression statistics.

1.3 Part c

Write the prediction equation based on the result. I wrote the prediction equation as follows, having calculated the intercept as

$$\beta o \tag{1}$$

and the slopes, as

$$\beta_1, \beta_2 \tag{2}$$

The equation is:

$$\text{prestige} = \beta_0 + \beta_1 \cdot \text{professional} + \beta_2 \cdot \text{income} \tag{3}$$

1.4 Part d

Interpret the coefficient for income. I have calculated the coefficient which represents the change in the prestige variable for a one-one increase in income, holding the other variables constant. Given that it is positive, an increase in income is associated with an increase in prestige. The number however is very small, 0.001371 units, for a one-unit increase. There is statistical significance evident as seen by a tiny p-value shown in Figure 2.

1.5 Part e

Interpret the coefficient for professional. What is indicated is that being professional is associated with an increase of 22.76 units in prestige, holding other variables constant. The variance has been calculated as being reasonably precise. The p-value is extremely small - well below the common significance threshold of 0.05, so statistical significance is there. There is strong evidence that being professional has an effect on prestige.

1.6 Part f

What is the effect of a dollar amount increase of 1,000 on prestige in income score for professional occupation? 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 dollar increase in income based on your answer for part c. To do so for this task, I multiplied the income coefficient, 1.371e-03 by 1000, the change in income and printed the result from R. It is 1.371.

1.7 Part g

What is the effect of changing one's occupation 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 part c. It is simply the coefficient from before. Answer: 136560.

2 Question 2

Without data to analyze, I used the results of a regression with two variables and a constant as provided. The question asked whether the existence of yard signs affect voter share for two candidates and what follows is the conclusions, based on calculations in R.

2.1 Part a

Use the results from a linear regression to determine whether having these yard signs in a precint affects vote share (e.g., conduct a hypothesis test with a = .05). The hypothesis test was set up as follows. The null hypothesis is such that coefficient of yard signs is zero (i.e., yard signs do not affect vote share). The alternative hypothesis is that the coefficient of yard signs is not zero (i.e., yard signs affect vote share). Using the provided results of the linear regression, I extracted the coefficient and standard error and completed a t-test. The formula and equation are provided below.

$$t = \frac{Coefficient}{StandardError} = \frac{0.042}{0.016} = 2.625 \tag{4}$$

So, not zero. I then found the p-value by taking the negative of the absolute value function, abs(). Enclosed is the code. Given that the p-value is below the threshold, I am confident that having yard signs in a precinct affects vote share.

2.2 Part 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 a = .05). Null Hypothesis is that the coefficient of precinct adjacent to lawn signs is zero (i.e., being adjacent to precincts with yard signs does not affect vote share). Alternative Hypothesis is that the coefficient of precinct adjacent to lawn signs is not zero (i.e., being adjacent to precincts with yard signs affects vote share). I extracted the coefficient and standard error and plugged it into R. From it, I calculated the t-value, determined the p-value from it and the degrees of freedom. Comparing the p-value to the significance level (here 0.05) led to rejecting the null hypothesis. So, the coefficient of precinct adjacent to lawn signs is not zero and thus, being adjacent to precincts with yard signs affects vote share.

2.3 Part c

Interpret the coefficient for the constant term substantively. The intercept value of 0.302 represents the expected value of vote share when both independent variables (precinct assigned lawn signs and precinct adjacent to lawn signs) are zero. So, when there are no yard signs assigned or adjacent (i.e., the variables are zero), the predicted vote share is 30.2 percent. This 30.2 percent vote share serves as a baseline measurement. It suggests that, in precincts without the influence of yard signs (neither assigned nor adjacent), the average vote share is expected to be 30.2 percent. The standard error of 0.011 indicates the precision of the intercept estimate. A small standard error implies the intercept is estimated with high precision. This means there's relatively little variation in the baseline vote share estimate when the influence of yard signs is not present.

2.4 Part 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 t-value and p-value have been derived from the coefficients and based on the results: yard signs significantly impact vote share both directly and indirectly (adjacent precincts). This highlights their importance relative to other factors, though further data could provide a more comprehensive understanding.