Problem Set 4

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)
```

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`).

The dataset is imported as follows:

```
data(Prestige )
df<-Prestige
```

and professional variable encoded into a binary one with:

```
professional <- ifelse(df$type=="prof", 1,
                       ifelse(df$type=="bc",0,
                       ifelse(df$type=="wc",0,NA)))
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 × dummy interaction.)

The folowing line do the linear regression:

```
lm1 <- lm(formula = df$prestige ~ df$income +professional + df$income:professional)
```

```
> summary(lm1)

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

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 ***
df$income               0.0031709  0.0004993   6.351 7.55e-09 ***
professional           37.7812800  4.2482744   8.893 4.14e-14 ***
df$income:professional -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 effacées parce que manquantes)
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.

Considering $\hat{Y}$ the predictor of `prestige`,
$\hat{\beta}_2$ the slope along `income` ,
$\hat{\beta}_1$ the slope along `professional` ,
$\hat{\beta_{1,2}}$ the slope of calculated interaction ,
$\hat{\beta}_0$ the calculated intercept,
X_2 the `income` variable
X_1 the `professional` variable
and the margin of error $\epsilon \rightsquigarrow \mathcal{N}(\mu, \sigma), (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$,
the linear regression model gives:

$$\begin{aligned} \hat{Y} &= \hat{\beta}_2 X_2 + \hat{\beta}_1 X_1 + \hat{\beta_{1,2}} X_1 X_2 + \hat{\beta}_0 + \epsilon \\ &= 0.003171 X_2 + 37.781280 X_1 - 0.002326 X_1 X_2 + 21.142259 + \epsilon \end{aligned}$$

(d) Interpret the coefficient for `income`.

We have $\hat{\beta}_2 \ll \hat{\beta}_1$ (by 4 orders of magnitude). So we can interpret this as $\hat{\beta}_2$ as neglectable.
Although it cannot be rigorously considered as 0, since $\sigma_{\hat{\beta}_2} = 0.0004993 < 0.1\hat{\beta}_2$ and the associated p-value is 6.351 7.55e-09. I.e. `income` seems having an influence on prestige, even if this one is small enough regarding other variations for being negected.

(e) Interpret the coefficient for `professional`.

We have$\hat{\beta}_1 = 37.781280$ and $\sigma_{\hat{\beta}_1} = 4.2482744$ with a p-value of 4.14e-14.
This means that `professional` has a important impact on `prestige`.

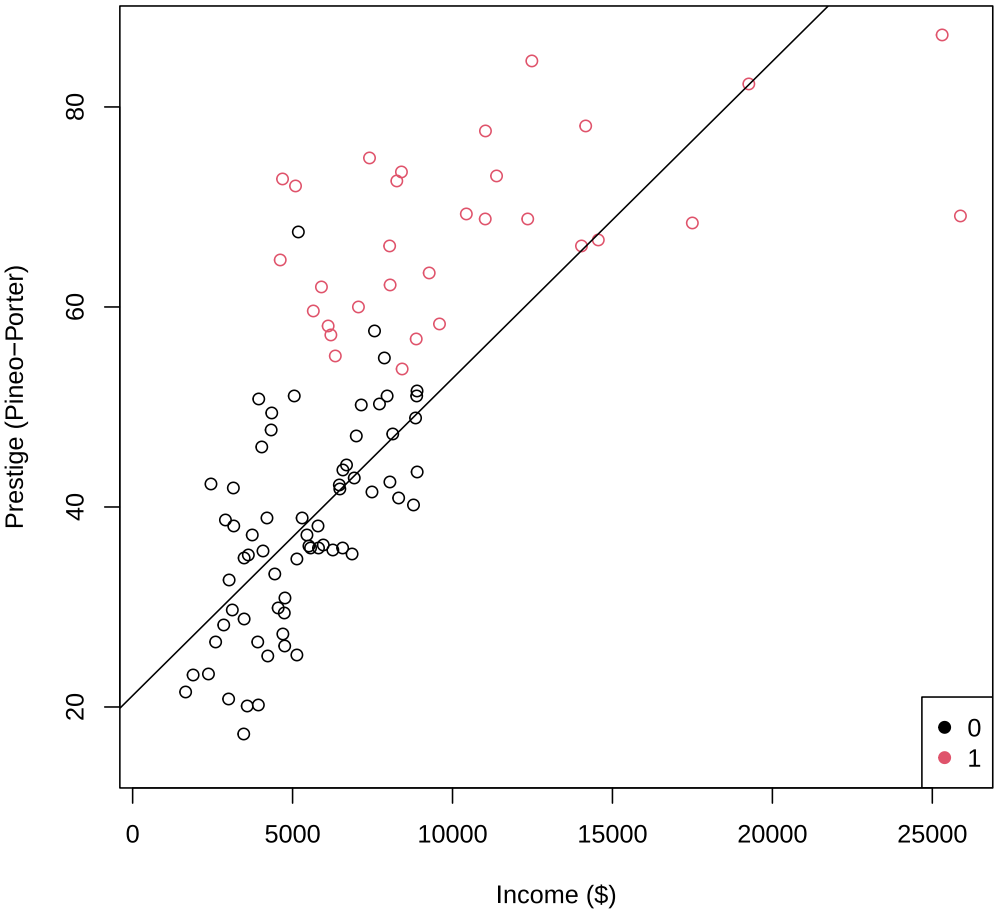
(f) What is the effect of a \$1,000 increase in income on prestige score for professional occupations? In other words, we are interested in the marginal effect of income when the variable `professional` takes the value of 1. Calculate the change in $\hat{y}$ associated with a \$1,000 increase in income based on your answer for (c).

With the previous model plotted in Fig. 1., if professional is 1, the latter equation is simplified to:

$$\begin{aligned} \hat{Y} &= \hat{\beta}_2 X_2 + \hat{\beta}_1 + \hat{\beta_{1,2}} X_2 + \hat{\beta}_0 + \epsilon \\ &= 0.003171 X_2 + 37.781280 - 0.002326 X_2 + 21.142259 + \epsilon \quad = 0.000845 X_2 + 58,923539 + \epsilon \end{aligned}$$

We had at marginally for `income`: $\frac{\partial \hat{Y}}{\partial X_2} = \hat{\beta}_1 + \hat{\beta_{1,2}} = 0.000845$.
Wich means, for a variation of income $\delta = \$1000$, the variation of prestige is

Figure 1: Regression between `prestige` and `income`



$\delta . \frac{\partial \hat{Y}}{\partial X_2} = \delta(\hat{\beta}_1 + \hat{\beta_{1,2}}) = 1000 \times 0.000845 = 0.0845$ points on Pineo-Porter prestige score.

Note that this gives the same result if we consider from the beginning the bivariate regression between `prestige` and `income` for sample where `professional` is equal to 1, as follows:

```
summary(lm2)

Call:
lm(formula = ifelse(professional == 1, df$prestige, NA) ~ ifelse(professional ==
    1, df$income, NA))
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-12.2443  -5.1013  -0.5626   6.9908  15.1284 

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              5.892e+01  2.985e+00   19.74  < 2e-16 **
ifelse(professional == 1, df$income, NA) 8.452e-04  2.523e-04    3.35  0.00226 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.494 on 29 degrees of freedom
  (71 observations effacées parce que manquantes)
Multiple R-squared:  0.279, Adjusted R-squared:  0.2541 
F-statistic: 11.22 on 1 and 29 DF,  p-value: 0.002255
```

with this code:

```
lm2 <- lm(formula =  ifelse(professional==1,df$prestige,NA) ~ ifelse(
    professional==1,df$income,NA))
```

which plotted gives Fig. 2

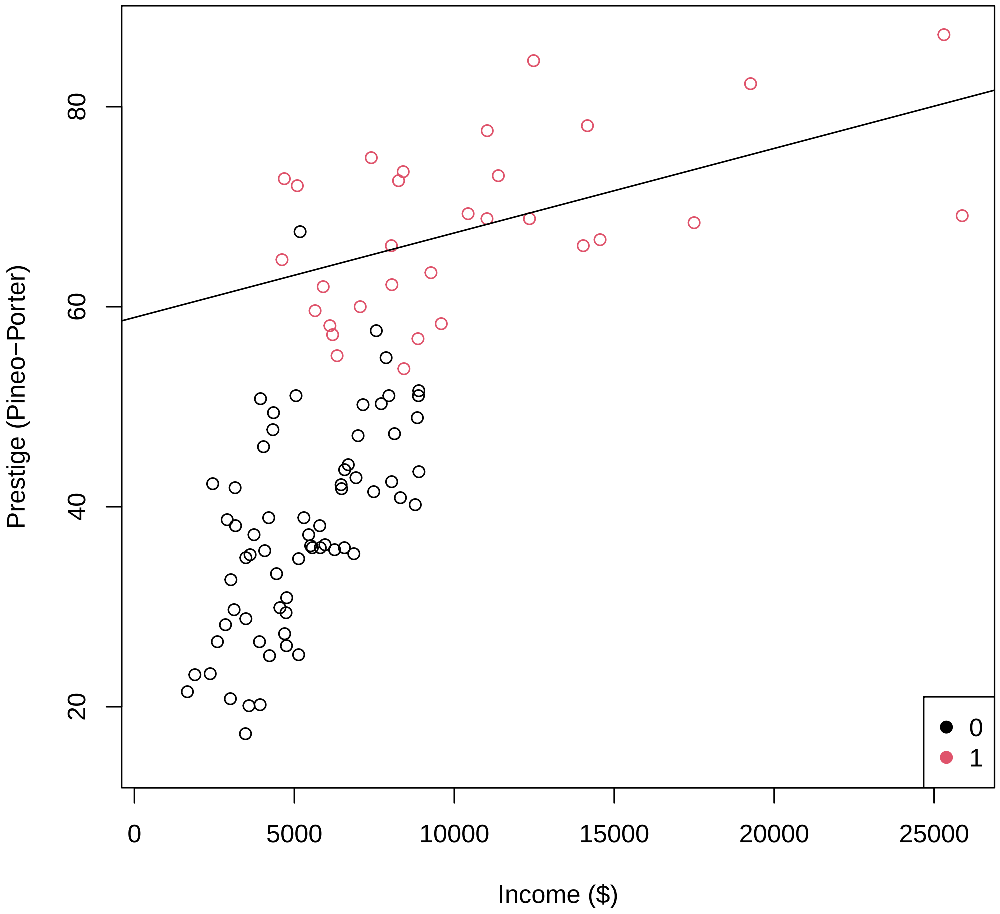
This means that, using the same notations (which is accurate here since supposing $X_1 = 1$ leads to the simplified equation of the model above), we have:
$\delta.\frac{\partial \hat{Y}}{\partial X_2} = \delta\hat{\beta}_1 = 1000 \times 8.452e - 04 = 0.8452$ points on Pineo-Porter (PPpt) prestige score with this second regression model.
This is exactly the value found above. Note that the dispersion is smaller with this model (the standard deviation is 2.5%) compared to the first model (5% with the sum of the two concerned coefficent standard deviation). The p-value is still reasonnable in this model (2.3%) even if its stronger with the first model (10^{-8} and 10^{-4}).

(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).

For professional jobs, we have the folowing regression (cf. f.) :

$$\begin{aligned}\hat{Y} &= (\hat{\beta}_2 + \widehat{\beta_{1,2}})X_2 + (\hat{\beta}_1 + \hat{\beta}_0) + \epsilon \\ &= 0.000845X_2 + 58,923539 + \epsilon\end{aligned}$$

Figure 2: Regression between `prestige` and `income`



For non-professional occupations, we consider $X_1 = 0$, so

$$\begin{aligned}\hat{Y} &= \hat{\beta}_2 X_2 + \hat{\beta}_1 X_1 + \hat{\beta}_0 + \epsilon \\ &= 0.003171 X_2 + 21.142259 + \epsilon\end{aligned}$$

In other words, the marginal effect on income variable is :
$m_p := \frac{\partial \hat{Y}}{\partial X_2} = \hat{\beta}_1 = 8.452e - 04 PPpt/\$$ for professional jobs, were as
$m_{np} := \frac{\partial \hat{Y}}{\partial X_2} = \delta \hat{\beta}_1 = 0.003171 PPpt/\$$ for non-professional jobs.
i.e. the variation is reduced by $m_p - m_{np} = 8.452e - 04 - 0.003171 = -0,002326 PPpt/\$$.

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 $\alpha = .05$).
To test a hypothesis on the impact of lawn sign on vote share, a T-test on the associated coefficient has to be done. The null hypotesis is "This coefficent is equal to 0." The p-value resulting from this test is given and is equal to 1.6%.
I.e. at risk 5% ($> 1.6\%$), the null hypotesis is rejected,
i.e. yard signs in a precinct affects 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 $\alpha = .05$).

To test a hypothesis on the impact of lawn sign in an adjacent precinct on vote share,

[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.

a T-test on the associated coefficient has to be done. The null hypotesis is ”This coefficent is equal to 0.” The p-value resulting from this test is given and is equal to 1.3%. I.e. at risk 5% (>1.3%), the null hypotesis is rejected,
i.e. yard signs in an adjacent precinct affects vote share.

(c) Interpret the coefficient for the constant term substantively.
This coefficient represent the proportion of vote for McAuliff without lawn sign in a given precinct and its adjacent precincts (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?

It suffice to run an overall F-test on the proposed model.
I.e. the test statistic $f = \frac{R^2/k}{(1-R^2)/(N-k-1)} = \frac{0.094/2}{(1-0.094)/(131-2-1)} = 6.6$
which gives the p-value 6.0%.
This is done with the following code:

```
f_test <- 6.6
n<-131
k<- 2
f_pvalue <- df (f_test  , k-1 , n )
```

Considering a risk of 5% the null hypotesis would not have been rejected. In addition, the global $R^2 = 0.094$. I.e., les than 10% of the variation is represented by this regression. In other words, there is other factors wich explain the remaining 90% of the information.