Analysis of Women income and education of Prestige

Jitong Jiang

2026-01-16

Contents

Download data 1
Clean data 1
Modeling 2
Visualization 3

Download data

The Prestige dataset, from the carData package, contains information on 102 occupations and 6 variables derived from Canadian census data and social surveys conducted in the 1960s and early 1970s. The variables include:

- education: Average years of education of occupational incumbents in 1971
- income: Average income (in dollars) of incumbents in 1971
- women: Percentage of incumbents who are women
- prestige: Pineo–Porter prestige score based on a mid-1960s social survey
- census: Canadian census occupational code
- type: Occupational category (blue collar, white collar, or professional)

This analysis focuses on professional occupations only. The research question is: Among **professional** occupations, how are the percentage of women in an occupation and the average years of education related to average income?

```r
if(!file.exists((here::here("data", "raw.RData")))){
  source(here("codes", "01_data_download.R"))
}else{
  load(here::here("data", "raw.RData"))
}
```

Clean data

The Prestige dataset was loaded and restricted to professional occupations only. Occupation names were converted from row names to a variable, and only the variables relevant to the analysis (education, income, and percentage of women) were retained. Observations with missing values were removed to ensure complete data for all included occupations.

```r
if(!file.exists((here::here("data", "clean.Rdata")))){
  source(here("codes", "02_data_cleaning.R"))
}else{
  load(here::here("data", "clean.Rdata"))
}
Prestige
```

```
##                      occupation education income women
## 1           gov.administrators     13.11  12351 11.16
## 2             general.managers     12.26  25879  4.02
## 3                  accountants     12.77   9271 15.70
## 4            purchasing.officers     11.42   8865  9.11
## 5                     chemists     14.62   8403 11.68
## 6                   physicists     15.64  11030  5.13
## 7                   biologists     15.09   8258 25.65
## 8                   architects     15.44  14163  2.69
## 9              civil.engineers     14.52  11377  1.03
## 10            mining.engineers     14.64  11023  0.94
## 11                   surveyors     12.39   5902  1.91
## 12                 draughtsmen     12.30   7059  7.83
## 13         computer.programers     13.83   8425 15.33
## 14                  economists     14.44   8049 57.31
## 15               psychologists     14.36   7405 48.28
## 16              social.workers     14.21   6336 54.77
## 17                     lawyers     15.77  19263  5.13
## 18                  librarians     14.15   6112 77.10
## 19      vocational.counsellors     15.22   9593 34.89
## 20                   ministers     14.50   4686  4.14
## 21         university.teachers     15.97  12480 19.59
## 22     primary.school.teachers     13.62   5648 83.78
## 23   secondary.school.teachers     15.08   8034 46.80
## 24                  physicians     15.96  25308 10.56
## 25               veterinarians     15.94  14558  4.32
## 26    osteopaths.chiropractors     14.71  17498  6.91
## 27                      nurses     12.46   4614 96.12
## 28            physio.therapsts     13.62   5092 82.66
## 29                 pharmacists     15.21  10432 24.71
## 30          commercial.artists     11.09   6197 21.03
## 31                      pilots     12.27  14032  0.58
```

Modeling

Let Y_i denote the average income for occupation i, Education$_i$ the average years of education, and Women$_i$ the percentage of women in the occupation. The model is

$$Y_i = \beta_0 + \beta_1 \,\text{Education}_i + \beta_2 \,\text{Women}_i + \varepsilon_i,$$

where ε_i are independent error terms with mean zero and constant variance.

```
if(!file.exists((here::here("data", "analysis.Rdata")))){
  source(here("codes", "03_data_analysis.R"))
}else{
  load(here::here("data", "analysis.Rdata"))
}

summary(income_lm)
```

```
##
```

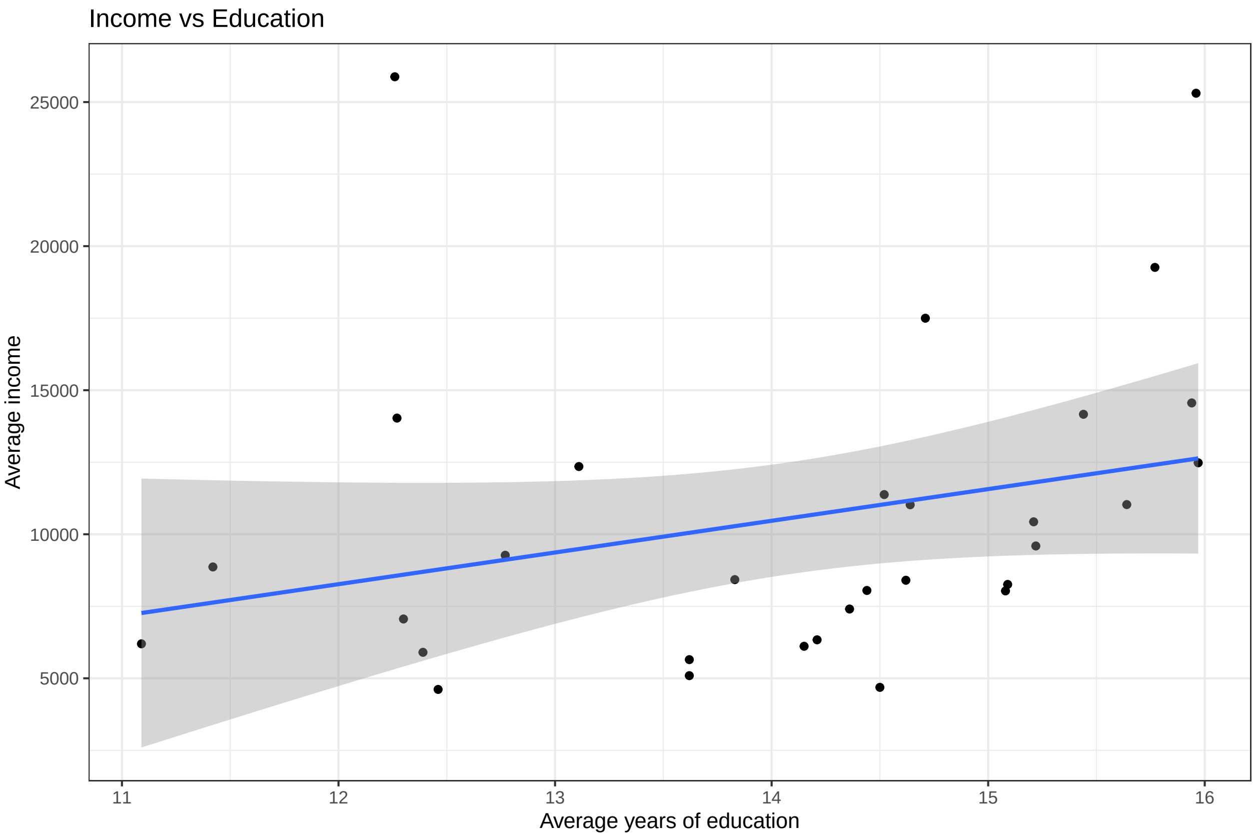
```
## Call:
## lm(formula = model, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8277.5 -2168.6  -974.2   681.6 15014.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -307.32    8697.47  -0.035  0.97206
## education     942.11     607.88   1.550  0.13241
## women         -94.16      29.87  -3.152  0.00384 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4626 on 28 degrees of freedom
## Multiple R-squared:  0.3209, Adjusted R-squared:  0.2724
## F-statistic: 6.616 on 2 and 28 DF,  p-value: 0.004436
```

The linear regression model fitted with income as the response variable and education and the percentage of women as predictors shows that the women percantion in professional occupation is significantly negatively associated with income, while education time shows a positive but statistically insignificant association.

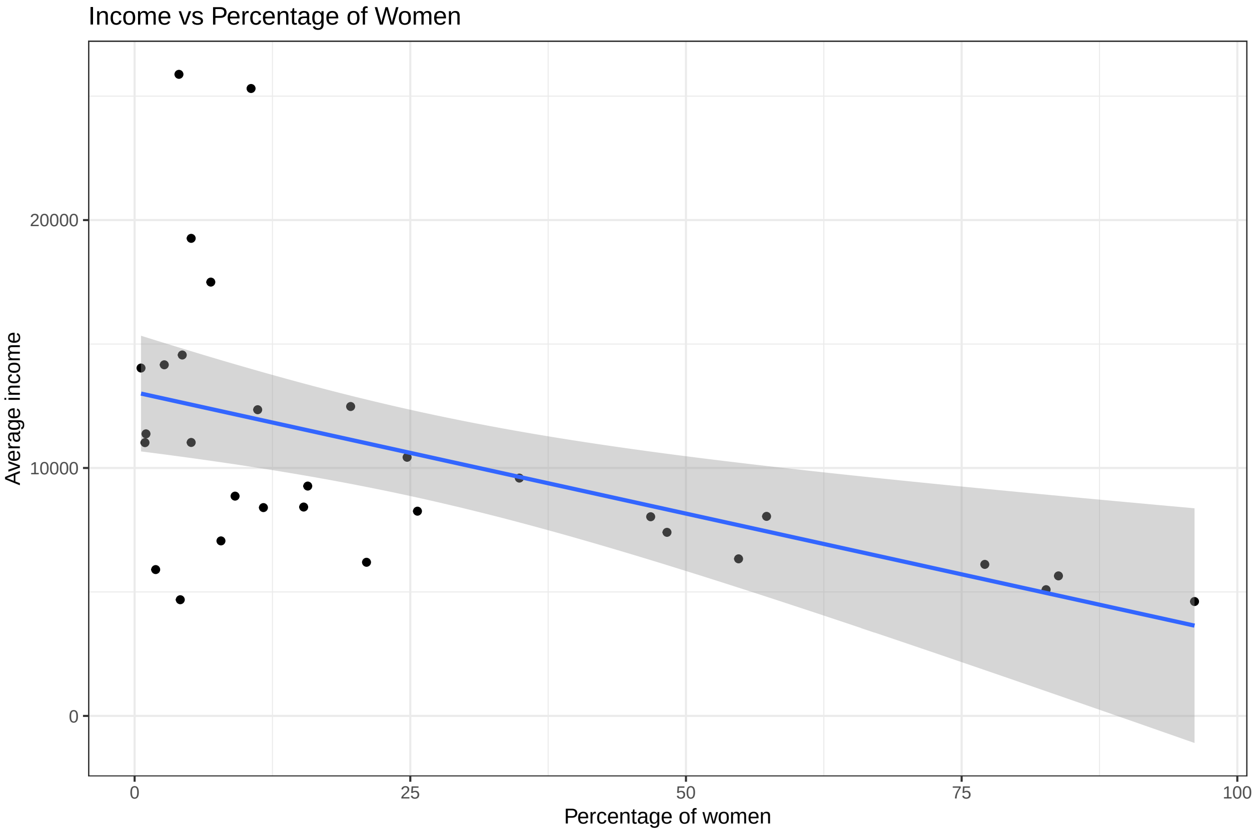
Visualization

```
source(here("codes", "04_data_visualization.R"))
```

```
plot1
```



```
plot2
```



The scatter plots with fitted regression lines show a positive association between education and income, and a negative association between the percentage of women in and income. These are consist with the results in the previous linear model.