EPsy 8251 Notes

Andrew Zieffler

2018-12-02

Contents

Chapter 1

Prerequisites

This is a *sample* book written in **Markdown**. You can use anything that Pandoc's Markdown supports, e.g., a math equation $a^2 + b^2 = c^2$.

The **bookdown** package can be installed from CRAN or Github:

```r
install.packages("bookdown")
# or the development version
# devtools::install_github("rstudio/bookdown")
```

Remember each Rmd file contains one and only one chapter, and a chapter is defined by the first-level heading `#`.

To compile this example to PDF, you need XeLaTeX. You are recommended to install TinyTeX (which includes XeLaTeX): https://yihui.name/tinytex/.

Chapter 2

Simple Linear Regression—Description

In this set of notes, you will begin your foray into regression analysis. To do so, we will use the *riverside.csv* data to examine whether education level is related to income. The data come from **?** and contain five attributes collected from a random sample of $n = 32$ employees working for the city of Riverview, a hyopothetical midwestern city. The attributes include:

- `education`: Years of formal education
- `income`: Annual income (in thousands of U.S. dollars)
- `seniority`: Years of seniority
- `gender`: Employee's gender
- `male`: Dummy coded gender variable (0 = Female, 1 = Male)
- `party`: Political party affiliation

2.1 Preparation

```
# Load libraries
library(corrr)
library(dplyr)
library(ggplot2)
library(readr)
library(sm)

# Read in data
city = read_csv(file = "~/Documents/github/epsy-8251/data/riverside.csv")
head(city)
```

```
## # A tibble: 6 x 6
##   education income seniority gender  male party
##       <int>  <dbl>     <int> <chr>  <int> <chr>
## 1         8   37.4         7 male       1 Democrat
## 2         8   26.4         9 female     0 Independent
## 3        10   47.0        14 male       1 Democrat
## 4        10   34.2        16 female     0 Independent
```

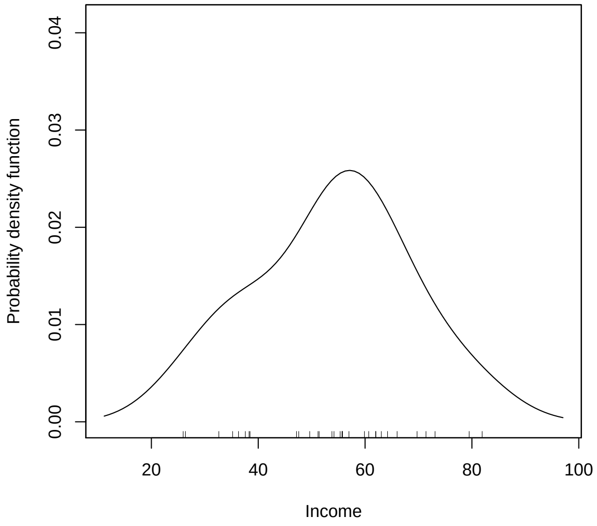

Figure 2.1: Density plot of employee incomes.

```
## 5          10   25.5          1 female     0 Republican
## 6          12   46.5         11 female     0 Democrat
```

2.2 Data Exploration

Any analysis should start with an initial exploration of the data. During this exploration, you should examine each of the variables that you will be including in the regression analysis. This will help you understand results you get in later analyses, and will also help foreshadow potential problems with the analysis. For additional detail, this blog post describes initial ideas of data exploration reasonably well. You could also refer to almost any introductory statistics text.

It is typical to begin by exploring the distribution of each variable separately. These distributions are referred to as *marginal distributions*. After that, it is appropriate to explore the relationships between the variables.

2.2.1 Income

To begin this exploration, we will examine the marginal distribution of employees' incomes. We can plot a marginal distribution using the `sm.density()` function from the **sm** package.

```
sm.density(city$income, xlab = "Income")
```

This plot suggests that employees' incomes are unimodal with most incomes between roughly $50,000 and $70,000. The rug at the bottom of the plot (the small vertical line segments) show the 32 incomes from our sample. The smallest income in the sample is about $25,000 and the largest income is over $80,000. (We could find the exact values using the `summary()` function.) This suggests there is a fair amount of variation in the data.

To further summarize the distribution, it is typical to compute and report summary statistics such as the mean and standard deviation. One way to compute these values is to use functions from the **dplyr** library.

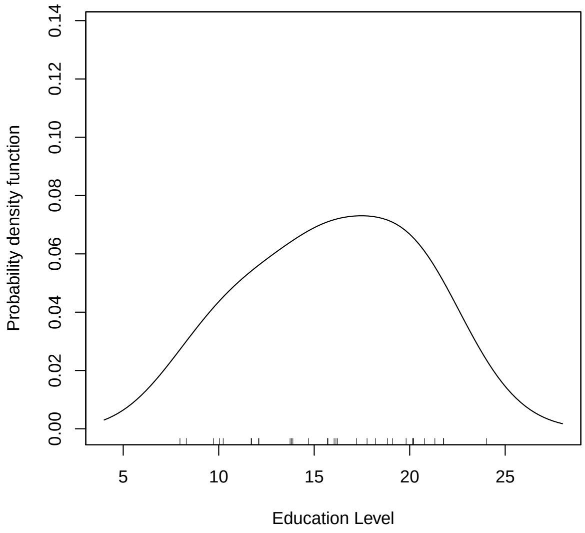

Figure 2.2: Density plot of employee education levels.

```
city %>%
  summarize(
    M = mean(income),
    SD = sd(income)
    )
```

```
## # A tibble: 1 x 2
##        M    SD
##    <dbl> <dbl>
## 1  53.7  14.6
```

Describing this variable we might write,

> The marginal distribution of income is unimodal with a mean of 53.74 thousand dollars. There is variation in employees' salaries (SD = 14.55 thousand dollars).

2.2.2 Education Level

We will also examine the distribution of the education level variable.

```
# Plot
sm.density(city$education, xlab = "Education Level")
```

Computing the mean and standard deviation:

```
# Summary statistics
city %>%
  summarize(
    M = mean(education),
    SD = sd(education)
    )
```

```
## # A tibble: 1 x 2
##       M    SD
##   <dbl> <dbl>
## 1    16  4.36
```

Again, we might write,

> The marginal distribution of education is unimodal with a mean of 16 years. There is variation in employees' level of education (SD = 4.4).

2.3 Relationship Between Variables

Although examining the marginal distributions is an important first step in the analysis, those descriptions do not help us directly answer our research question. To better understand the relationship between income and education level we need to explore the distribution of income (Y) as a function of education (X). To do this, we will create a scatterplot of incomes versus education.

2.3.1 Scatterplot

Below, we use `ggplot()` to create a scatterplot.

```
ggplot(data = city, aes(x = education, y = income)) +
  geom_point() +
  theme_bw() +
  xlab("Education (in years)") +
  ylab("Income (in thousands of U.S. dollars)")
```

The plot suggests a relationship (at least for these employees) between level of education and income. When describing the relationship we want to touch on four characteristics of the relationship:

- Functional form (structure) of the relationship
- Direction
- Strength
- Observations that do not fit the trend (outliers)

2.3.2 Correlation

To numerically summarize relationships between variables, we typically compute correlation coefficients. The correlation coefficient is a quantification of the direction and strength of the relationship. (It is important to note that the correlation coefficient is only an appropriate summarization of the relationship if the functional form of the relationship is linear.)

To compute the correlation coefficient, we use the `correlate()` function from the **corrr** package. We can use the dplyr-type syntax to select the variables we want correlations between, and then pipe that into the `correlate()` function. Typically the response (or outcome) variable is the first variable provided in the `select()` function, followed by the predictor.

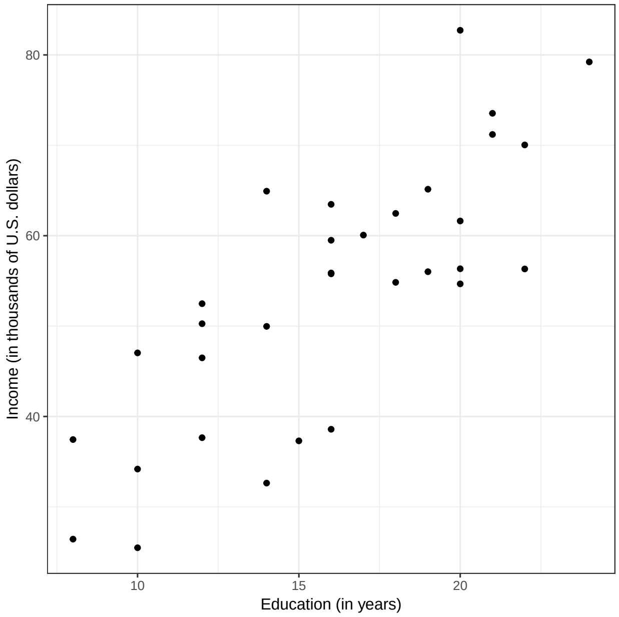

Figure 2.3: Scatterplot displaying the relationship between employee education levels and incomes.

```
# Load corrr package
library(corrr)

# Compute correlation between income and education level
city %>%
  select(income, education) %>%
  correlate()
```

```
##
## Correlation method: 'pearson'
## Missing treated using: 'pairwise.complete.obs'
```

```
## # A tibble: 2 x 3
##   rowname   income education
##   <chr>      <dbl>     <dbl>
## 1 income    NA         0.795
## 2 education  0.795    NA
```

When reporting the correlation coefficient is is conventional to use a lower-case r and report the value to two decimal places. Subscripts are also generally used to indicate the variables. For example,

$$r_{\text{education, income}} = 0.79$$

Combining the information culled from the scatterplot with that of the correlation analysis, we could summarize the relationship between education level and income as,

> There is a strong, positive, linear relationship between education level and income ($r = .79$). This suggests that city employees with lower education levels tend to have lower incomes, on average, than employees with higher education levels.