Examining Scatterplots

Chapter 6, Lab 1

OpenIntro Biostatistics

Topics

- Fitting and interpreting a line
- Checking assumptions

The relationship between two numerical variables can be visualized using a scatterplot in the xy-plane. The *predictor* variable is plotted on the horizontal axis, while the *response* variable is plotted on the vertical axis. This lab introduces the idea of using a straight line, $y = b_0 + b_1 x$, where b_0 is the y-intercept and b_1 is the slope, to summarize data that exhibit an approximately linear relationship.

The material in this lab corresponds to Section 6.1 of *OpenIntro Biostatistics*.

Background information

This lab uses data from the Prevention of REnal and Vascular END-stage Disease (PREVEND) study, which took place between 2003 and 2006 in the Netherlands. Clinical and demographic data for 4,095 individuals are stored in the `prevend` dataset in the `oibiostat` package.

As adults age, cognitive function declines over time; this is largely due to various cerebrovascular and neurodegenerative changes.

The Ruff Figural Fluency Test (RFFT) is one measure of cognitive function that provides information about cognitive abilities such as planning and the ability to switch between different tasks. Scores on the RFFT range from 0 to 175 points, where higher scores are indicative of better cognitive function.

The goal of this lab is to begin exploring the relationship between age and RFFT score in the `prevend` dataset.

Fitting and interpreting a line

The questions in this lab will be based around data from a random sample of $n = 500$ individuals from the prevend dataset.

1. Run the following code chunk to load the prevend data and store a random sample of 500 individuals as prevend.sample.

```r
#load the data
library(oibiostat)
data("prevend")

#take a random sample
set.seed(5011)
sample.size = 500
sample.rows = sample(1:nrow(prevend), sample.size, replace = FALSE)
prevend.sample = prevend[sample.rows, ]
```

2. Create a scatterplot of RFFT score (RFFT) and age in years (Age) in prevend.sample, then add a line of best fit.

   a) What are the slope and intercept values of the line of best fit?

   b) Interpret the slope and intercept values in the context of the data; i.e., explain the linear model in terms that a non-statistician would understand. Comment on whether the intercept value has any interpretive meaning in this setting.

   c) Based on the linear model, how much does RFFT score differ, on average, between an individual who is 60 years old versus an individual who is 50 years old?

   d) Write the equation of the least squares line in the form $\hat{y} = b_0 + b_1 x$. According to the linear model, what is the average RFFT score for an individual who is 70 years old?

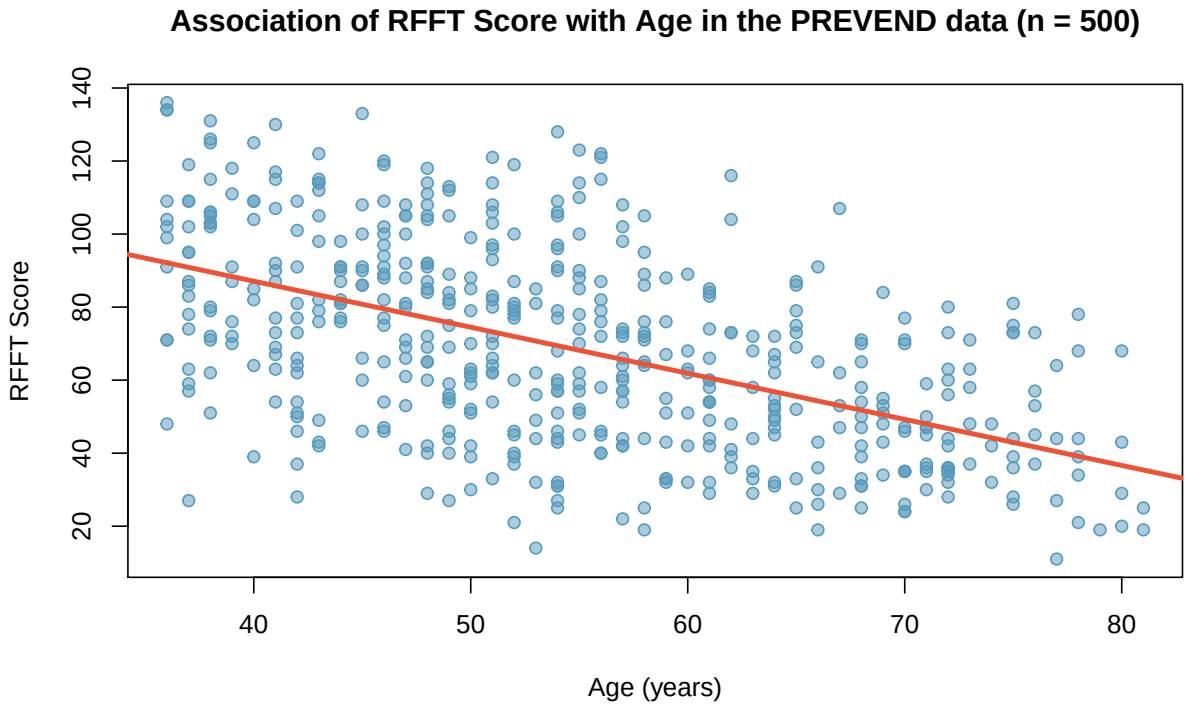
   e) Is it valid to use the linear model to estimate RFFT score for an individual who is 20 years old? Explain your answer.

Checking assumptions

There are four assumptions that should be true for a line to be considered a reasonable approximation for a relationship shown in a scatterplot, and for methods of inference to be applied to the linear model:

1. *Linearity*: The data show a linear trend.
2. *Constant variability*: The variability of the response variable about the line remains roughly constant as the predictor variable changes.
3. *Independent observations*: The (x, y) pairs are independent.
4. *Approximate normality of residuals*: This condition will be explained in the next lab.

The next lab discusses specialized plots for checking assumptions. For now, the following question addresses how to do an informal check of assumptions based on simply assessing the scatterplot of the data with a line of best fit.



3. Perform an informal check of assumptions 1 - 3.
   a) Do the data seem to show a linear trend for the response variable y as a function of the predictor variable x?
   b) Does the variability of the response variable around the line remain roughly constant for all values of the predictor variable?
   c) From the description of the study, does it seem reasonable to assume that the observations are independent, such that the values of one (x, y) pair provide no information about any other (x, y) pair?