

Examining Scatterplots

Chapter 6, Lab 1: Solutions

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, $y = b_0 + b_1x$, 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

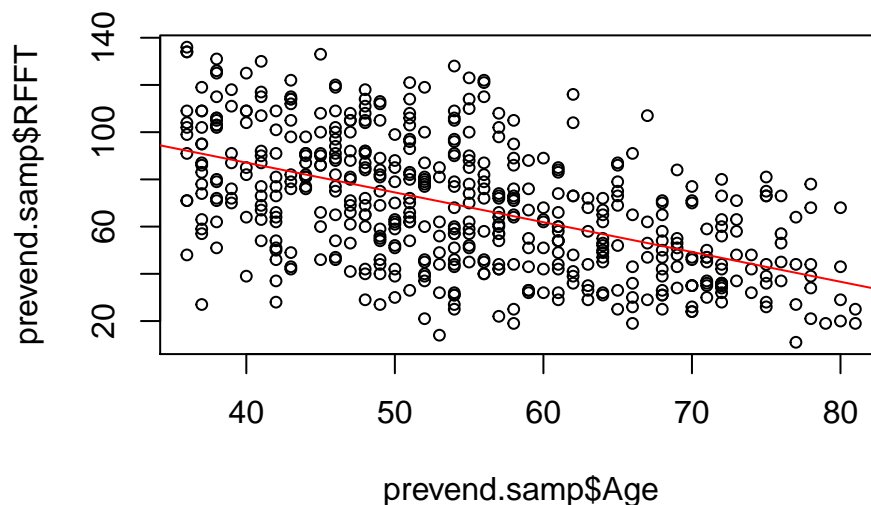
The questions in this lab will be based around data from a random sample of $n = 500$ individuals from the prevent dataset; the sample is stored as `prevent.samp` in the `oibiostat` package.

1. Run the following code chunk to load the `prevent.samp` dataset.

```
#load the data
library(oibiostat)
data("prevent.samp")
```

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

```
#create a plot of RFFT versus age with line of best fit
plot(prevent.samp$RFFT ~ prevent.samp$Age, cex = 0.75)
abline(lm(prevent.samp$RFFT ~ prevent.samp$Age), col = "red")
```



```
#print the equation of the line
lm(prevent.samp$RFFT ~ prevent.samp$Age)
```

```
##
## Call:
## lm(formula = prevent.samp$RFFT ~ prevent.samp$Age)
##
## Coefficients:
##      (Intercept)  prevent.samp$Age
##           137.550           -1.261
```

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

The slope is -1.261 and the intercept is 137.55.

- 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.

According to the model slope, an increase in age of 1 year is associated with an average decrease in RFFT score by 1.3 points. The model intercept suggests that the average RFFT score for an individual of age 0 is 137.55. The intercept value does not have interpretive meaning in this setting, since it is not reasonable to assess a newborn's cognitive function with a test like the RFFT.

- 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?

Based on the linear model, RFFT score differs, on average, by $1.261(10) = 12.61$ points for individuals that have an age difference of 10 years; the younger individual is expected to have a higher RFFT score.

- d) Write the equation of the least squares line in the form $\hat{y} = b_0 + b_1x$. According to the linear model, what is the average RFFT score for an individual who is 70 years old?

The equation of the line is $\widehat{RFFT} = 137.55 - (1.261)(age)$. According to the model, the average RFFT score for an individual who is 70 years old is $137.55 - (1.261)(70) = 49.28$ points.

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

No, it is not valid to use the linear model to estimate RFFT score for an individual who is 20 years old. The data in the sample only extend from ages 36 to 81 and should not be used to predict scores for individuals outside that age range. It may well be that the relationship between RFFT and age is different for individuals in a different age group.

```
summary(prevend.samp$Age)
```

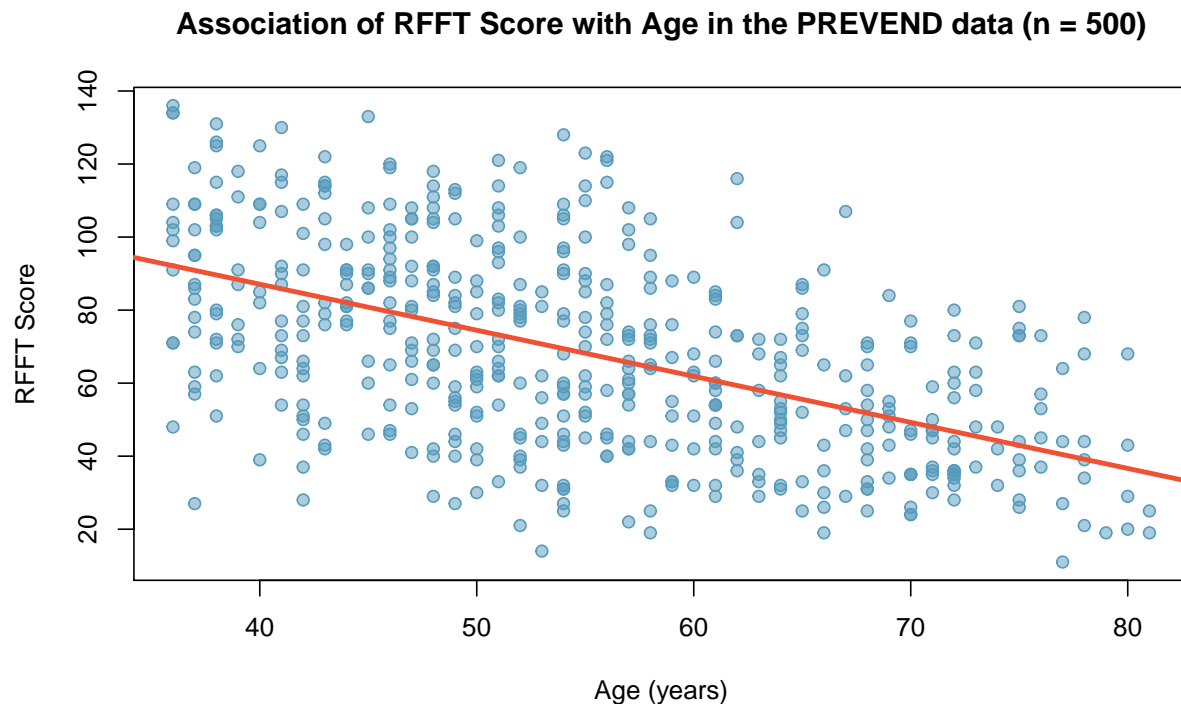
##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	36.00	46.00	54.00	54.82	64.00	81.00

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 ?

Yes, the data show a linear trend, with RFFT scores tending to decrease as age increases.

- b) Does the variability of the response variable around the line remain roughly constant for all values of the predictor variable?

Yes, the variability around the line seems roughly constant across values of age, although

observations at higher ages do seem somewhat less variable than those at younger ages. This decrease in variability may simply be from having fewer observations on older individuals.

- 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?

The data are from individuals enrolled in a large study. It seems reasonable to assume that the observations were independent; one pair of age and score values does not provide information about other pairs. Independence could be violated, if, for example, many of the individuals were related, because certain conditions associated with cognitive decline have genetic risk factors.