Introduction to Least Squares Regression

Chapter 6, Lab 1

OpenIntro Biostatistics

Topics

- Fitting and interpreting a line
- Calculating a least squares line
- Checking assumptions with residual plots

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 statistical model for least squares regression is also formally introduced, along with the residual plots used to assess the assumptions for linear regression.

The material in this lab corresponds to Section 6.1, 6.2, and 6.3.1 of *OpenIntro Biostatistics*.

Introduction

Least squares regression

The vertical distance between a point in the scatterplot and the predicted value on the regression line is the **residual** for the point. For an observation (x_i, y_i), where $\hat{y}_i$ is the predicted value according to the line $\hat{y} = b_0 + b_1 x$, the residual is the value $e_i = y_i - \hat{y}_i$.

The **least squares regression line** is the line which minimizes the sum of the squared residuals (SSE) for all the points in the plot; i.e., the regression line is the line that minimizes $e_1^2 + e_2^2 + ... + e_n^2$ for the n pairs of points in the dataset.[1]

For a general population of ordered pairs (x, y), the population regression model is

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where ϵ is a normally distributed 'error term' with mean 0 and standard deviation σ.

The terms β_0 and β_1 are parameters with estimates b_0 and b_1. These estimates can be calculated from summary statistics: the sample means of x and y ($\overline{x}$ and $\overline{y}$), the sample standard deviations of x and y (s_x, s_y), and the correlation between x and y (r).

$$b_1 = r\frac{s_y}{s_x} \qquad b_0 = \overline{y} - b_1\overline{x}$$

[1] SSE stands for "sum of squared errors" and refers to the sum of squared residuals.

Plots for checking assumptions

There are a variety of **residual plots** used to check the fit of a least squares line. The ones used in this textbook are scatterplots in which predicted values are on the x-axis and residual values on the y-axis. Residual plots are useful for checking the assumptions of linearity and constant variability.

To assess the normality of residuals, **normal probability plots** are used. These plots are also known as quantile-quantile plots, or Q-Q plots.

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; the sample is stored as `prevend.samp` in the `oibiostat` package.

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

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

2. Create a scatterplot of RFFT score (`RFFT`) and age in years (`Age`) in `prevend.samp`.

3. Examine the plot and consider possible lines that are a reasonable approximation for the relationship in the plot.

   a) Consider the line $\hat{y} = -20 + 2x$.

      i. Using the code provided in the template, add the line to the plot. Does the line appear to be a good fit to the data?

      ii. Calculate the SSE, the sum of the squared residuals, for this line. Do you expect this SSE to be relatively low or relatively high? Explain your answer.

   b) From a visual inspection, determine a line that you think is a good fit to the data and add the line to the plot. Calculate the SSE and compare it to the SSE from the line in part a).

c) Consider the line $\hat{y} = 137.55 - 1.261x$. Add this line to the plot. Calculate the SSE and compare it to the SSE from the line in part b).

4. Create a scatterplot of RFFT score (`RFFT`) and age in years (`Age`) in `prevend.samp`, then add a least squares line of best fit.

   a) What are the slope and intercept values of the least squares 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 with residual plots

There are four assumptions that must be met for a linear model to be considered reasonable: linearity, constant variability, independent observations, and normally distributed residuals.

Even though linearity and constant variability can be assessed from the scatterplot of y versus x, it is helpful to consult residual plots for a clearer view. Normality of residuals is best assessed through a normal probability plot; although skew can be visible from a histogram of the residuals, deviations from normality are more obvious on a normal probability plot.

RFFT and age in the `prevend` *data*

5. Run the following chunk to create a residual plot where the residual values are plotted on the y-axis against predicted values from the model on the x-axis, using data in `prevend.samp`.

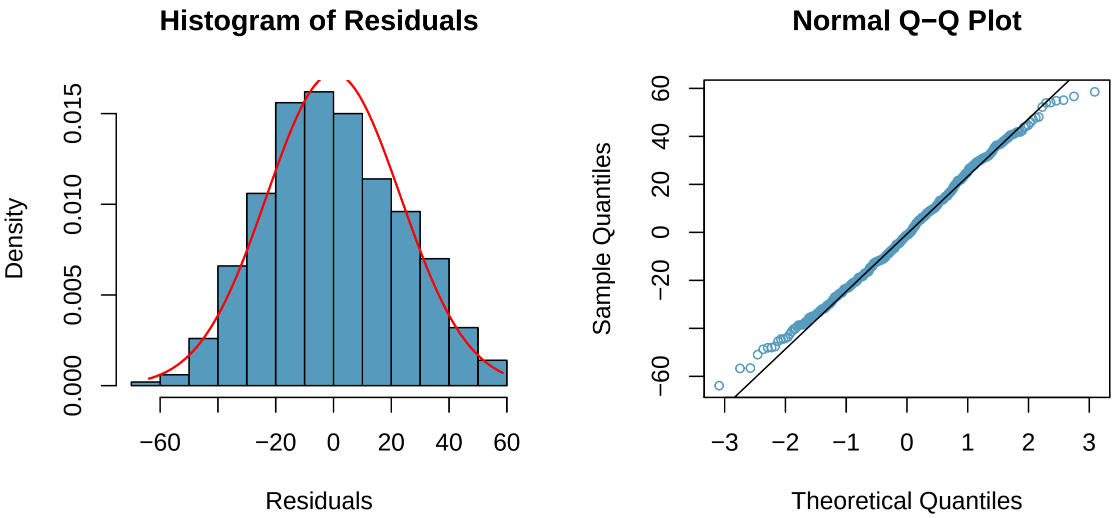
```r
#store the residuals from the linear model
prevend.residuals = resid(lm(RFFT ~ Age, data = prevend.samp))

#store the predicted RFFT scores from the linear model
prevend.predicted = predict(lm(RFFT ~ Age, data = prevend.samp))

#create residual plot
plot(prevend.residuals ~ prevend.predicted)
abline(h = 0, col = "red", lty = 2)
```

   a) When a linear model is a good fit for the data, the residuals should scatter around the horizontal line $y = 0$ with no apparent pattern. Does a linear model seem appropriate for these data?

   b) Does the variability of the residuals seem constant across the range of predicted RFFT scores?

6. Run the code chunk shown in the template to create a normal probability plot of the residuals. For comparison purposes, the following figure shows a histogram of the residual values overlaid with a normal curve and the normal probability plot.

   Do the residuals appear to be normally distributed?



7. Overall, does it seem that a least squares regression line is an appropriate model for estimating the relationship between cognitive function (as measured by RFFT score) and age?

Clutch volume and body size in the `frog` data

The `frog` dataset in the `oibiostat` package contains observations from a study conducted on a frog species endemic to the Tibetan Plateau. Researchers collected measurements on egg clutches and female frogs found at breeding ponds across five study sites.

Previous research suggests that larger body size allows females to produce egg clutches with larger volumes. Frog embryos are surrounded by a gelatinous matrix that may protect developing embryos from temperature fluctuation or ultraviolet radiation; a larger matrix volume provides added protection. In the data, clutch volume (`clutch.volume`) is recorded in cubic millimeters and female body size (`body.size`) is measured as length in centimeters.

The following questions step through examining whether a linear regression model is appropriate for the relationship between female body size and clutch volume.

8. Create a scatterplot of clutch volume versus female body size and plot the least squares line.
9. Create a residual plot where the residual values are plotted on the y-axis against predicted values from the model on the x-axis.
   a) Does the linearity assumption seem to be satisfied?
   b) Is the variability of the residuals constant across the range of predicted clutch volumes?
10. Assess whether it can be reasonably assumed that the observations are independent.

11. Create a Q-Q plot and assess whether the residuals appear to be normally distributed.

12. Evaluate whether a least squares regression line is an appropriate model for estimating the relationship between female body size and clutch volume.