Exercises

7.1 (*mammals* data). The `mammals` data set in the `MASS` package records brain size and body size for 62 different mammals. Fit a regression model to describe the relation between brain size and body size. Display a residual plot using the plot method for the result of the `lm` function. Which observation (which mammal) has the largest residual in your fitted model?

7.2 (*mammals*, continued). Refer to the `mammals` data in package `MASS`. Display a scatterplot of `log(brain)` vs `log(body)`. Fit a simple linear regression model to the transformed data. What is the equation of the fitted model? Display a fitted line plot and comment on the fit. Compare your results with results of Exercise 7.1.

7.3 (*mammals* residuals). Refer to Exercise 7.2. Display a plot of residuals vs fitted values and a normal-QQ plot of residuals. Do the residuals appear to be approximately normally distributed with constant variance?

7.4 (*mammals* summary statistics). Refer to Exercise 7.2. Use the `summary` function on the result of `lm` to display the summary statistics for the model. What is the estimate of the error variance? Find the coefficient of determination (R^2) and compare it to the square of the correlation between the response and predictor. Interpret the value of (R^2) as a measure of fit.

7.5 (Hubble's Law). In 1929 Edwin Hubble investigated the relationship between distance and velocity of celestial objects. Knowledge of this relationship might give clues as to how the universe was formed and what may happen in the future. Hubble's Law is is

$$\text{Recession Velocity} = H_0 \times \text{Distance},$$

where H_0 is Hubble's constant. This model is a straight line through the origin with slope H_0. Data that Hubble used to estimate the constant H_0 are given on the DASL web at `http://lib.stat.cmu.edu/DASL/Datafiles/Hubble.html`. Use the data to estimate Hubble's constant by simple linear regression.

7.6 (*peanuts* data). The data file "peanuts.txt" (Hand et al. [21]) records levels of a toxin in batches of peanuts. The data are the average level of aflatoxin X in parts per billion, in 120 pounds of peanuts, and percentage of non-contaminated peanuts Y in the batch. Use a simple linear regression model to predict Y from X. Display a fitted line plot. Plot residuals, and comment on the adequacy of the model. Obtain a prediction of percentage of non-contaminated peanuts at levels 20, 40, 60, and 80 of aflatoxin.

7.7 (*cars* data). For the `cars` data in Example 7.1, compare the coefficient of determination R^2 for the two models (with and without intercept term in the model). Hint: Save the fitted model as `L` and use `summary(L)` to display R^2. Interpret the value of R^2 as a measure of the fit.

7.8 (*cars* data, continued). Refer to the `cars` data in Example 7.1. Create a new variable `speed2` equal to the square of `speed`. Then use `lm` to fit a quadratic model

$$dist = \beta_0 + \beta_1 speed + \beta_2 (speed)^2 + \varepsilon.$$

The corresponding model formula would be `dist ~ speed + speed2`. Use `curve` to add the estimated quadratic curve to the scatterplot of the data and comment on the fit. How does the fit of the model compare with the simple linear regression model of Example 7.1 and Exercise 7.7?

7.9 (Cherry Tree data, quadratic regression model). Refer to the Cherry Tree data in Example 7.3. Fit and analyze a quadratic regression model $y = b_0 + b_1 x + b_2 x^2$ for predicting volume y given diameter x. Check the residual plots and summarize the results.

7.10 (*lunatics* data). Refer to the "lunatics" data in Example 7.8. Repeat the analysis, after deleting the two counties that are offshore islands, `NANTUCKET` and `DUKES` counties. Compare the estimates of slope and intercept with those obtained in Example 7.8. Construct the plots and analyze the residuals as in Example 7.8.

7.11 (*twins* data). Import the data file "twins.txt" using `read.table`. (The commands to read this data file are shown in the twins example in Section 3.3, page 85.) The variable `DLHRWAGE` is the difference (twin 1 minus twin 2) in the logarithm of hourly wage, given in dollars. The variable `HRWAGEL` is the hourly wage of twin 1. Fit and analyze a simple linear regression model to predict the difference `DLHRWAGE` given the *logarithm* of the hourly wage of twin 1.