Chapter 3: Fitting linear models to data

1. Load the dataset `heights.txt` into R as a dataframe called `heights_data`. Assume that `heights.txt` has been downloaded into your current working directory. Use the following command:

```
heights_data <- read.table("heights.txt", header = TRUE)
```

   The dataset contains two columns: `Mother` and `Daughter`, representing heights in inches.

   - Convert the heights from inches to centimetres (1 inch = 2.54 cm). Add two new columns to `heights_data` named `Mother_cm` and `Daughter_cm`, containing the heights in centimetres.

2. Create a scatter plot of daughters' heights against mothers' heights using the measurements in centimetres. Label the axes appropriately.

   - Describe any patterns or relationships you observe in the scatter plot. Does the relationship appear to be linear?

3. Using the measurements in centimetres, fit a linear regression model to predict daughters' heights based on mothers' heights. Store the model in an object called `height_model`.

```
height_model <- lm(Daughter_cm ~ Mother_cm, data = heights_data)
```

   - Display the summary of the linear model using the `summary()` function.
   - Interpret the slope and intercept of the model. What does the slope tell you about the relationship between mothers' and daughters' heights? Are the results statistically significant?
   - Using the coefficients from the model (which you can extract with `coef(height_model)`), calculate by hand the predicted height of a daughter whose mother is 170 cm tall.
   - Verify your calculation by using the `predict()` function:

```
predict(height_model, newdata = data.frame(Mother_cm = 170))
```

4. Add the best-fit line from your linear model to the scatter plot from Exercise 2.

   - Extract the residuals from your model using the `residuals()` function.
   - Plot a histogram of the residuals. Comment on whether the distribution of residuals appears approximately normal.
   - Create a scatter plot of residuals as a function of the predicted values. What do you observe in this plot?

5. Standardize `Mother_cm` and `Daughter_cm` by subtracting their mean and dividing by their standard deviation. Add two new columns, `Mother_std` and `Daughter_std`, to `heights_data`.

   - Fit a new linear regression model predicting `Daughter_std` from `Mother_std`. Store this model in an object called `height_model_std`.

- Display the summary of the standardized model. What is the slope coefficient? How does it relate to the Pearson correlation coefficient between `Mother_cm` and `Daughter_cm`?
- Calculate the Pearson correlation coefficient using the `cor()` function, and also perform a correlation test using the `cor.test()` function. Extract the correlation coefficient and p-value.
- Comment on the strength and direction of the relationship. Is the correlation statistically significant? How does the correlation coefficient relate to the slope in the standardized regression model?

6. Suppose the results of a visual attention experiment suggest that we can predict the average time (in seconds) it takes participants to find a visual target among a set of distractors. The data indicate the following:

   - when there are 5 distractors, the predicted response time is 0.4 seconds
   - for each additional distractor, the predicted response time increases by approximately 35 milliseconds (i.e., 0.035 seconds)
   - for approximately 95 per cent of participants, the average response time falls within ±0.1 seconds of the predicted value.

   Based on this information:

   - Write down the equation of the regression line, where the response time is predicted from the number of distractors.
   - What is the residual standard deviation of the regression model (i.e., the standard deviation of the residuals)?

7. Load the dataset `wagespeed.csv` into R as a dataframe called `wagespeed_data`, assuming it has been downloaded into your working directory:

```r
wagespeed_data <- read.csv("wagespeed.csv")
```

   This dataset contains measurements of the average walking speed (`wspeed`) of passers-by in major cities around the world, and a normalized hourly wage (`wage`) based on New York City (wage = 100). The theory is that people walk faster when the cost of their time is higher.

   - Fit a linear regression model predicting walking speed (`wspeed`) from wage. Store the model in an object called `wage_model`.
   - Create a scatter plot of walking speed against wage, and add the regression line.
   - Using the coefficients of the model, compute the expected walking speed in a hypothetical city where the average wage is two-thirds that of New York City (i.e., wage = 66.67).
   - Create a histogram of the residuals from the model. Save the residuals into a variable called `wspeed_residuals`:

     ```r
     wspeed_residuals <- residuals(wage_model)
     ```

   - Use a quantile-quantile (QQ) plot to assess whether the residuals are approximately normally distributed:

```r
qqnorm(wspeed_residuals)
qqline(wspeed_residuals)
```

Recall from Chapter 2 that a quantile divides data into intervals of equal probability. In a normal QQ plot, if the points lie along the line, this suggests that the distribution is approximately normal (i.e. the quantiles are approximately at the same distance from the mean as in a normal distribution). Deviations from the line – particularly at the edges (the most extreme quantiles) – can indicate departures from normality. Look especially at the points far from the centre of the distribution: do they tend to fall above or below the line? Do both ends behave similarly? Or is there more deviation on one side than the other?

8. Write a brief report summarizing your findings from the regression analyses on both the `heights_data` and `wagespeed_data` datasets. Follow APA style guidelines for reporting statistical results.
    - Include the scatter plots with best-fit lines as figures in your report. Reference the figures appropriately in your text.
    - Summarize and interpret the regression coefficients, confidence intervals, and key findings for each model.
    - Reflect on the relationship between wage and walking speed. Can you think of any other factors – besides the economic cost of time – that might influence how fast people walk in different cities?