CHAPTER 6

Relationships between two quantitative variables

correlation analysis with elements of linear regression modelling

What you will learn from this chapter:

Will your knowledge of statistics improve as you read more and more books on the subject? Is there a relationship between the length of a word and its frequency? Does grammatical proficiency of children depend on the number of lexical items which they have mastered? Does the number of phonemes in a language depend on the number of speakers? All these questions involve correlation between two variables. This chapter explains the principles of correlation analysis and demonstrates how it can be carried out using popular parametric and non-parametric tests. You will also learn how to produce correlograms and scatter plots with a regression line. Some fundamental notions of regression analysis, such as residuals, homo- and heteroscedasticity, will be introduced. The case studies investigate the relationship between word frequency and mean reaction time in a lexical decision task and the correlation between vocabulary size and grammatical proficiency in first language acquisition.

6.1 What is correlation?

The previous chapters dealt with descriptive and inferential statistics that describe a single variable (mean, median, standard deviation, etc.) or pinpoint the differences between scores of two groups on the same variable (the *t*-test and analogous non-parametric tests). In this chapter, you will learn how to investigate the relationship between two quantitative variables. More specifically, you will learn to measure and test **correlations**. A correlation is called **positive** if the values of both variable X and variable Y increase and decrease together: if X increases, Y increases, and if X decreases, Y decreases, as well. For instance, suppose you are interested whether there is a relationship between the size of a child's vocabulary and his or her grammatical proficiency (see Section 6.3). You will probably expect to find a positive correlation: the more words a child knows, the higher his or her

grammatical proficiency will be. A **negative**, or inverse, correlation is observed when the values of X and Y change in opposite directions: if X increases, Y decreases, or if X decreases, Y increases. An example of a negative correlation would be the relationship between word frequency and word length: according to Zipf's Law of Abbreviation (see Chapter 5), the more frequent a word, the shorter it is, i.e. the fewer syllables or phonemes it contains.

The strength of such relationships is usually measured with the help of a **correlation coefficient**. It normally ranges from –1 (perfect negative correlation) to 1 (perfect positive correlation). 0 indicates a lack of relationship. In this chapter we will discuss three different correlation coefficients: Pearson's product-moment coefficient r, Spearman's ρ ('rho') and Kendall's τ ('tau'). The Pearson r is applied to interval- or ratio-scaled variables, whereas the Spearman and Kendall coefficients deal with ordinal data (ranks), as well as interval or ratio-scaled variables transformed into ranks.

After a researcher has computed a correlation coefficient, he or she also needs to know whether this relationship will be observed if one takes another sample from the same population. In other words, one has to test if the coefficient is statistically significant. As in the previous chapter, the choice of an appropriate correlation test depends on a number of conditions. While the significance test of the Pearson r is based on the assumption of normality, the Spearman ρ and Kendall τ are non-parametric tests, which do not assume a distribution of a particular shape.

The remaining part of the chapter is organized as follows. Section 6.2 discusses the Pearson product-moment correlation coefficient in a case study that investigates the relationship between word length and mean reaction times in a lexical decision task. Section 6.3 explores the relationship between vocabulary size and grammatical proficiency in first language acquisition using the Spearman ρ and Kendall τ. Section 6.4 shows how one can visualize correlations between more than two variables in a correlogram. Section 6.5 summarizes the main ideas of the chapter.

6.2 Word length and word recognition: The Pearson product-moment correlation coefficient

6.2.1 The data and hypothesis

For this case study you will need data and functions from the following add-on packages that should be installed and loaded, unless you have done so already:

```
> install.packages(c("ggplot2", "energy", "car"))
> library(Rling); library(ggplot2); library(energy); library(car)
```

This case study will investigate if there is a correlation between the length of a word, on the one hand, and how fast it is recognized by speakers in a lexical decision experiment, on the other hand. The data can be found in the data frame `ldt` in `Rling`. This dataset was introduced in Chapter 3. The variables of interest are *Length* (word length in letters) and *Mean_RT* (average reaction times in a lexical decision task). This time, the dataset will be attached, so that the variables can be easily accessed without specifying the name of the data frame.

```
> data(ldt)
> attach(ldt)
> summary(Length)
   Min.   1st Qu.  Median   Mean   3rd Qu.  Max.
   3.00   6.00     8.00     8.23   10.00    15.00
> summary(Mean_RT)
   Min.   1st Qu.  Median   Mean   3rd Qu.  Max.
   564.2  713.1    784.9    808.3  905.2    1459.0
```

The distributional characteristics of these variables should already be familiar to the reader from Chapter 3, where the variables were analysed in detail with the help of various descriptive statistics and graphs.

The alternative hypothesis of this case study is as follows: the longer a word, the longer the time that is needed to recognize it. The hypothesis is directional. The null hypothesis is that there is no correlation between word length and reaction time.

6.2.2 Descriptive statistics and visualizations

To visualize the relationship between two quantitative variables, one can create a scatter plot with the help of the `plot()` function. The first argument specifies the coordinates of the points on the horizontal axis, and the second argument provides the values on the vertical axis:

```
> plot(Length, Mean_RT, main = "Scatter plot of word length and mean
reaction times")
```

Alternatively, you can use an expression with a tilde. The variable on the left from the tilde will be plotted on the vertical axis, and the variable on the right will be plotted on the horizontal axis:

```
> plot(Mean_RT ~ Length, main = "Scatter plot of word length and
mean reaction times")
```

The plot is shown in Figure 6.1. The plot also displays a regression line. Roughly speaking, a regression line shows a general trend in the data. A more detailed explanation will be provided below. To add the line, use the following code:

```
> m <- lm(Mean_RT ~ Length)
> abline(m)
```

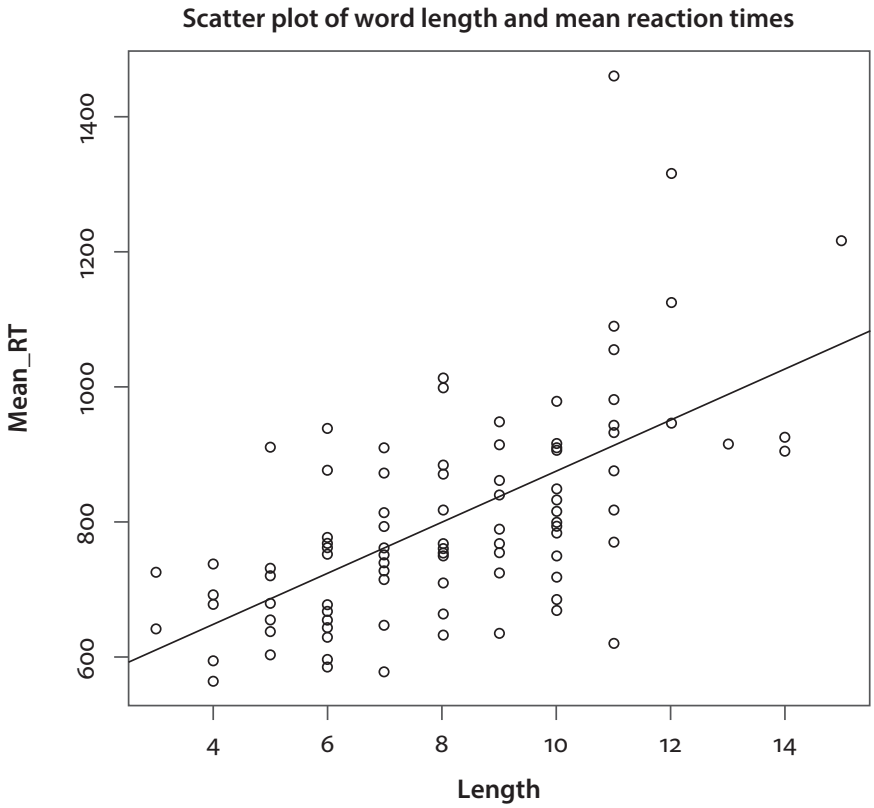
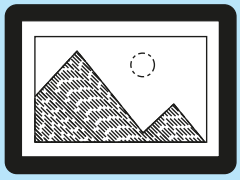


Figure 6.1. Scatter plot of word length in letters and mean reaction times



How to create a scatter plot with a regression line the help of `ggplot2`

To create a `ggplot2` scatter plot with a regression line, similar to the one shown in Figure 6.1, one can use the following code:

```
> ggplot(ldt, aes(x = Length, y = Mean_RT)) + geom_point(shape =
1, size = 3) + stat_smooth(method = lm)
```

Note that `stat_smooth()` also adds a 95% confidence region around the regression line. The result can be seen in Figure 6.1a.

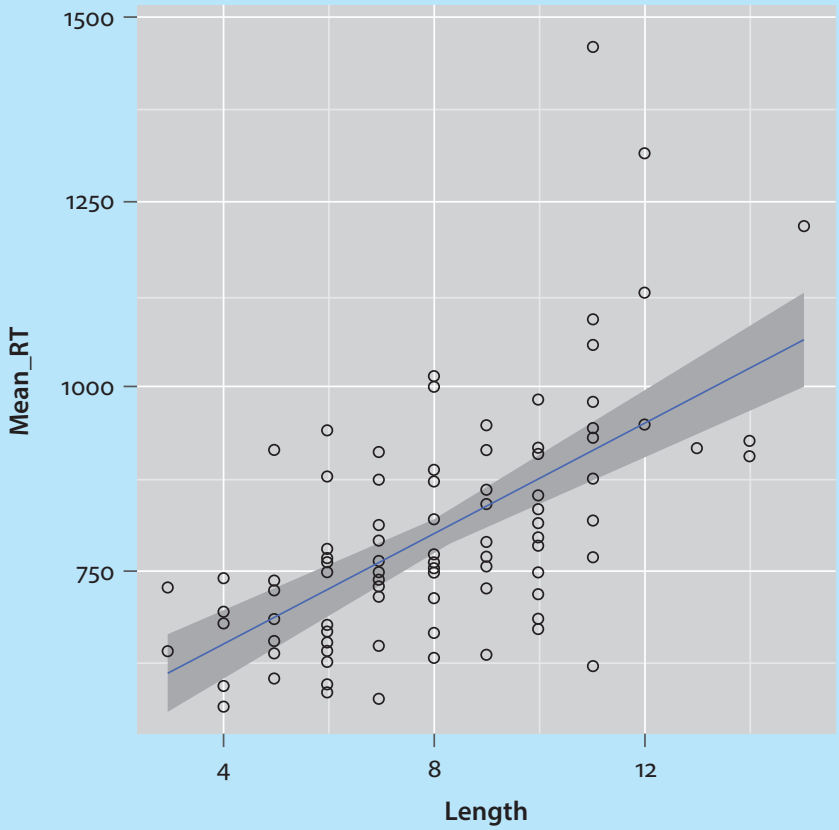


Figure 6.1a. A `ggplot2` version of the scatter plot with a regression line in Figure 6.1

The correlation seems to be positive. The longer a word, the more time it is needed to recognize it. But what is the strength of this correlation? To find this out, we need to compute the correlation coefficient. Let us begin with the Pearson product-moment correlation coefficient, which is probably the most widely used one. It can be computed as follows:

```
> cor(Mean_RT, Length) # equivalent to cor(…, use = "everything",
method = "pearson")
[1] 0.6147456
```

The coefficient is positive: $r = 0.615$. Some other possibilities are shown in Figure 6.2. When r is –1 or +1, all points fall on the regression line. The correlation is perfect. The closer r to zero, the more individual points deviate from the line and the weaker the correlation. As a very approximate rule of thumb, if r is equal to or greater than 0.7 or smaller than –0.7, the correlation is considered to be strong. If r is between 0.3 and 0.7 or between –0.3 and –0.7, it is considered to be moderate. If r is between 0 and 0.3 or 0 and –0.3, the correlation is considered to be weak. Note that a steep slope does not mean that the correlation is strong. It only shows the number of units by which y will change if x changes. The angle of the slope also depends on how R scales the axes.

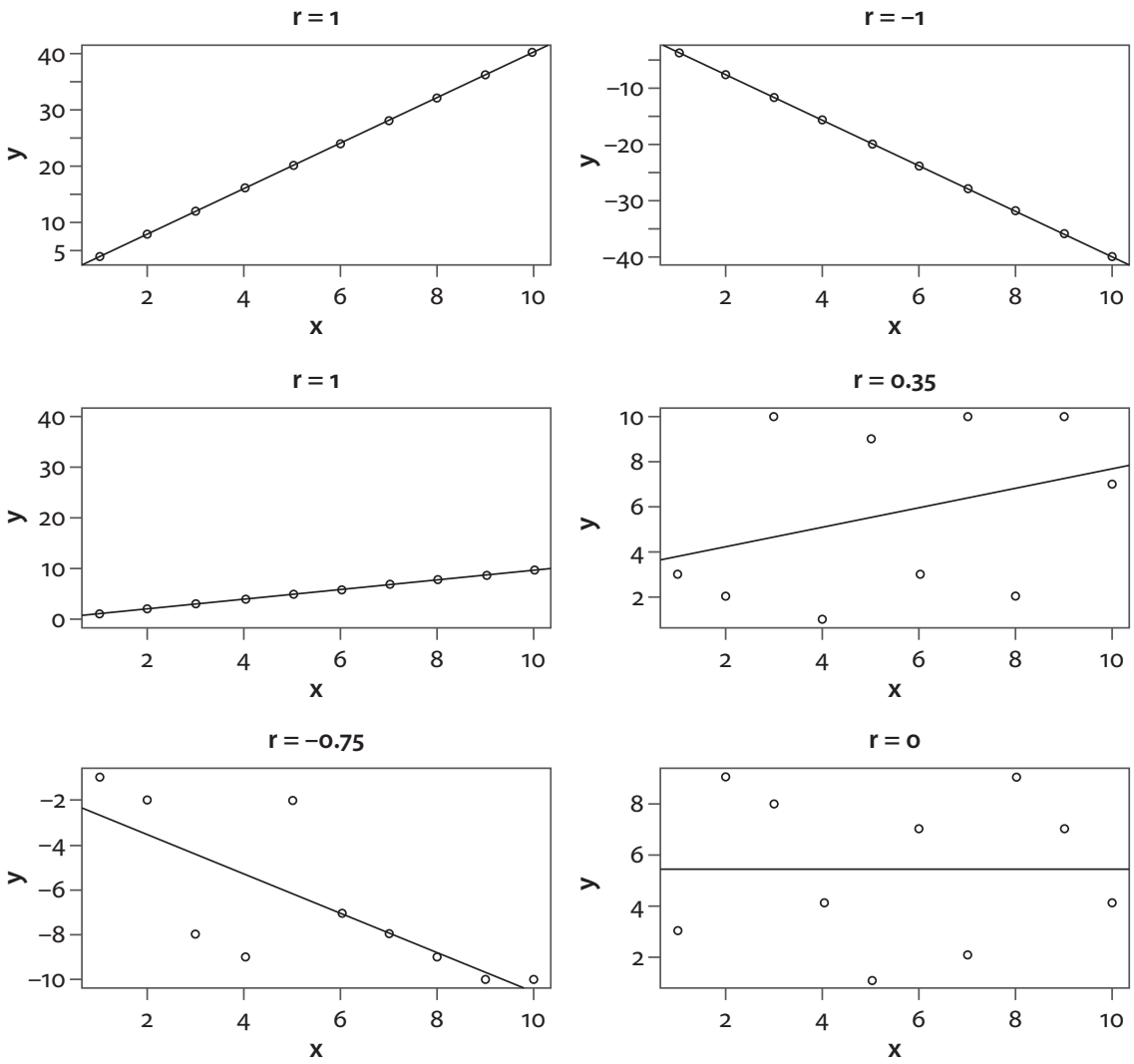


Figure 6.2. Several possible values of Pearson product-moment correlation coefficient *r*

Correlation can be discussed in terms of regression analysis, which models the relationship between one response (dependent) variable and one or several explanatory (independent) variables. In our case, we model the relationship between mean reaction time and word length. In regression analysis, it is crucial to distinguish between **observed** and **fitted values** of the response variable. The observed *y*-scores are as follows (only six first numbers are shown):

```
> head(Mean_RT)
[1] 819.19 977.63 908.22 766.30 1125.42 948.33
```

The corresponding fitted values are as follows:

```
> head(fitted(m))
1         2         3         4         5         6
799.5952  874.8831  761.9512  724.3072  950.1711  950.1711
```

Another crucial concept is **residuals**, which are the differences between the observed and fitted values of the response variable:

```
> head(residuals(m))
1           2           3           4           5
19.594813  102.746875  146.268782  41.992751  175.248936
6
-1.841064
```

Consider the scatter plot in Figure 6.3. Observed values are the actual values of the points on the y-axis. Fitted values correspond to the *y*-values of the projections of the observed values on the regression line. Finally, the value of a residual is equal to the height of the vertical line that can be drawn from a point to the regression line. Most commonly, the regression line is drawn in such a way as to approach all points as closely as possible and thus to minimize the sum of squared residuals (this is called the least squares method). The smaller the residuals (relative to the total variation of *y*), the stronger the correlation. More on this will follow in the next chapter.

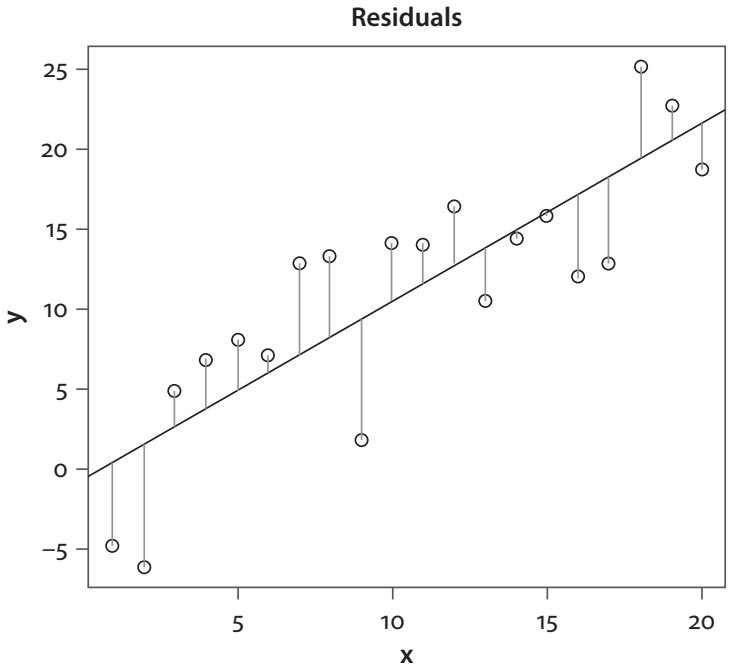


Figure 6.3. Plot with a cloud of points, a regression line (the diagonal) and residuals (vertical lines)

Correlation analysis and paired *t*-test: superficial similarity

Both correlation analysis and the paired *t*-test deal with paired observations represented by vectors *X* and *Y*. However, the similarity is only superficial. The paired *t*-test

(*Continued*)

tests the difference between the mean values of X and Y, whereas correlation analysis measures the strength of relationship between X and Y. Correlation analysis does not tell you whether the mean value of X is greater or smaller than the mean value of Y.

There are a few caveats to keep in mind when using the Pearson correlation coefficient. First, it makes sense only if the relationship between the variables is monotonic and linear. Second, it is very sensitive to the presence of outliers.

A relationship between variables X and Y is called **monotonic** when an increase in X is followed by an increase in Y, or a decrease in X is followed by a decrease in Y. Relationships are linear when Y decreases or increases at the same rate as X does, and vice versa. Consider the illustrations in Figure 6.4. The left graph shows a monotonic linear relationship. The central plot displays a monotonic, but non-linear relationship. Finally, a non-monotonic relationship is shown in the right graph. As an example, consider the relationship between age and weight. A baby's weight increases by three times during its first year, but then the growth slows down. Thus, the relationship between age and weight is non-linear: if one draws a graph with age as the x-axis, and weight as the y-axis, it will look like a curve, not like a straight line. Sometimes a person's weight even decreases in the old age, so the relationship can be described as both non-linear and non-monotonic.

It does not make sense to use the Pearson correlation coefficient if the relationship is non-monotonic and/or non-linear. Consider the right graph of Figure 6.4, which illustrates a non-monotonic relationship. The correlation coefficient computed by R is in fact almost zero: $r = -0.02$. However, this does not mean that there is no relationship: on the contrary, the relationship between x and y is quite strong, but it is a quadratic one, $y = x^2$. In such cases, the Pearson r gives a wrong idea of the relationship between the variables. However, if we transform the data by squaring x and measure the correlation between y and x^2, the correlation coefficient will be 0.61, which represents the relationship more correctly.

In the central plot of Figure 6.4, the underlying relationship is a logarithmic one: $y = log(x)$. The correlation is positive and strong, $r = 0.8$, if one measures the correlation between x and y, but it becomes somewhat higher, $r = 0.85$, if the logarithmic nature of the relationship is taken into account, $log(x)$ vs. y. These and other transformations will become very useful when you learn to fit linear regression models in the next chapter.

As one could see in the previous exploratory plots, the relationship between word length and mean reaction time seems to be linear. More precise diagnostic tools will follow in the next chapter.

The second potential problem is outliers. Consider two situations shown in Figure 6.5. The left graph (without the outlier) shows no relationship, $r = 0$. The right part shows a strong positive correlation between x and y when the outlier is present, $r = 0.87$. Obvi-

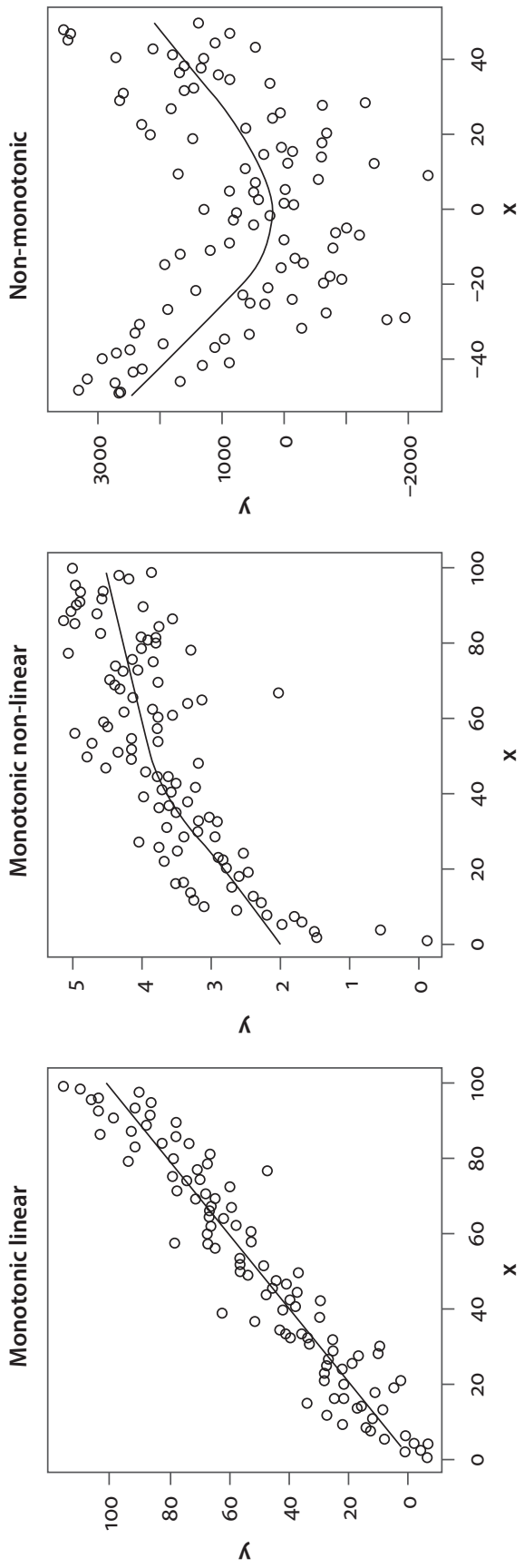


Figure 6.4. Monotonic and non-monotonic relationships between variables x and y

ously, $r = 0$ describes the general trend (or, rather, a lack of any trend) more correctly. Such observations are called **leverage** points because they can 'pull' the regression line in some direction.

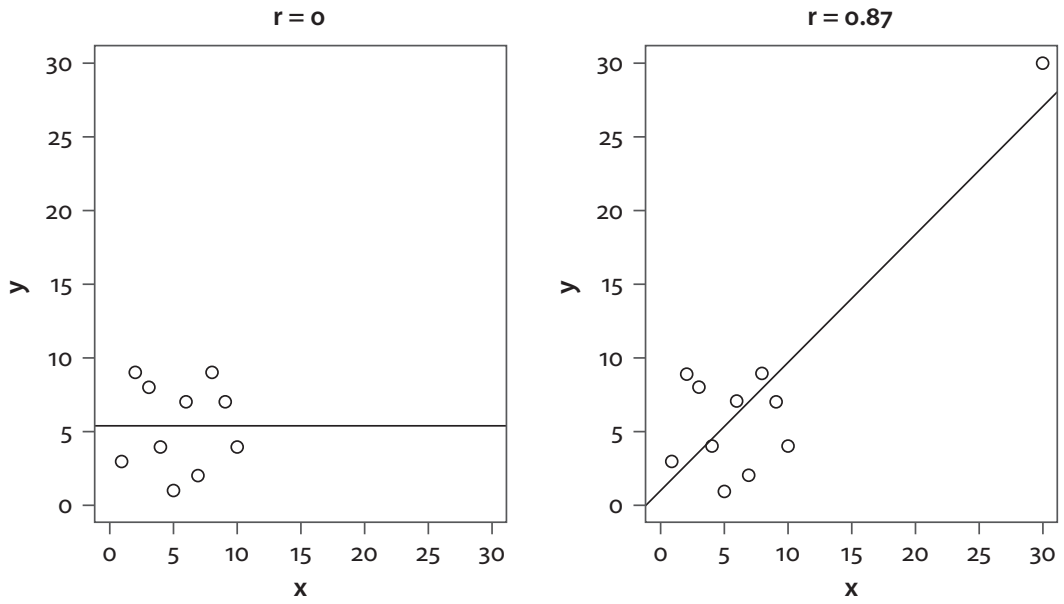


Figure 6.5. Impact of an outlier on the value of the Pearson *r*

Recall that we identified three outliers in the mean reaction times in Chapter 3 with the help of different diagnostic techniques. Those were unusually long reaction times. If you look at Figure 6.1 again, you will see a few points above $y = 1200$, which do not fit the pattern well, especially the one with the score above 1400. They are located in the top left corner. Let us try to exclude these points and see what happens:

```
> Mean_RT_1 <- Mean_RT[Mean_RT < 1200]
> length(Mean_RT_1)
[1] 97
```

The corresponding values in the frequency vector should be excluded, as well:

```
> Length_1 <- Length[Mean_RT < 1200]
> length(Length_1)
[1] 97
```

Now we are left with 97 observations out of the initial 100. What will change? Let us add a new regression line to the scatter plot in Figure 6.1. If you have already closed the graphics window with the plot, you will need to create the plot again before adding the line. This new line will be based on the data without these three points. The new trend is represented with a dashed line (`lty = 2`):

```
> m1 <- lm(Mean_RT_1 ~ Length_1)
> abline(m1, lty = 2)
```

The result is shown in Figure 6.6.

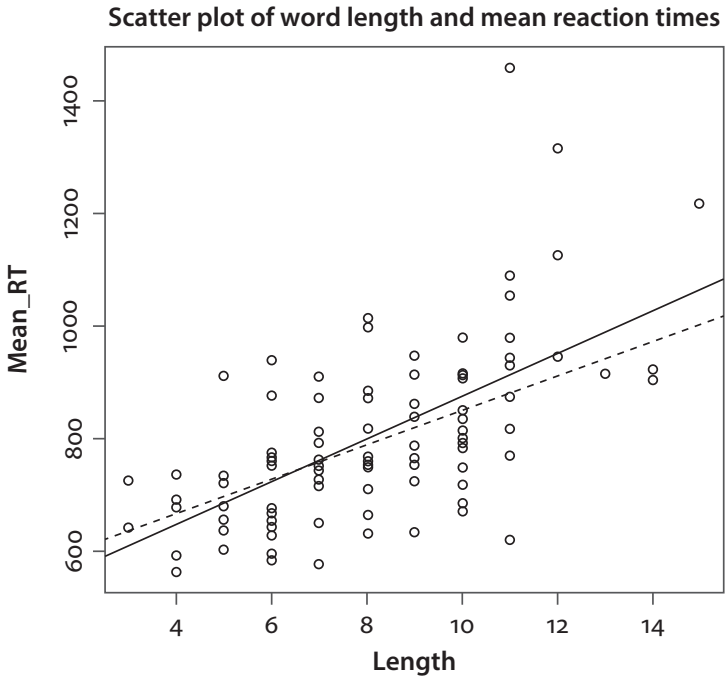


Figure 6.6. Scatter plot of word length and mean reaction time, with two regression lines. The solid line is based on the full dataset; the dashed line is based on the dataset without three outliers

The slope has changed slightly. This is because the line is no longer 'pulled' up by the outliers in the top right corner. Has the correlation coefficient changed?

```
> cor(Mean_RT_1, Length_1)
[1] 0.5886011
```

The correlation coefficient has become more moderate.

Handling missing data

If your data contain missing scores in at least one variable, R will return 'NA' instead of the correlation coefficient. Consider two vectors, *x* and *y*. The former contains no missing data, whereas the latter has one missing value:

```
> x <- 1:10
> x
[1] 1 2 3 4 5 6 7 8 9 10

> y <- x*2
> y[5] <- NA
```

(*Continued*)

```
> y
[1]  2  4  6  8 NA 12 14 16 18 20
```

If you try computing a correlation coefficient, R will return 'NA':

```
> cor(x, y)
[1] NA
```

You may want to tell R to consider only those cases where both scores are non-missing by adding `use = "complete"` or `use = "pairwise"`:

```
> cor(x, y, use = "complete")
[1] 1
```

6.2.3 Testing the significance of the correlation coefficient

It is not enough to compute the correlation coefficient and interpret it. One should also test whether the observed correlation is statistically significant. That is, if one takes another sample and measures the correlation, will the results be similar?

If you want to test whether the Pearson correlation coefficient is statistically significant, a few assumptions should be met:

- *The sample is randomly selected from the population it represents.* In our case, this means that the words should have been selected randomly, which is the case.
- *Both variables are at least interval-scaled.*
- *Both variables come from a* ***bivariate*** *normal distribution* and/or the sample size is large (30 and more observations). A bivariate normal distribution means that both variables and their linear combination are normally distributed. In other words, for any given value of variable X, the scores on variable Y will be normally distributed, and vice versa.[1] We will discuss a test that can help you detect violations of this assumption.
- *The residual (error) variance is* ***homoscedastic*** (*homo* is 'same' and *scedastic* comes from 'scatter'). That is, the relationship between the variables should be of equal strength across the entire range of both variables.
- *The residuals are independent*. This means that there should be no **autocorrelation** between residuals. One speaks about autocorrelation when the value of a variable depends on its previous or next value. Consider temperature: it increases gradually in the summer and decreases in the winter. It is very unlikely to have $+35^0$ C one day and -20^0 next day. Another example is economic cycles. Economic indicators, such as GDP, in one year tend to depend on their values in the previous year. As linguistic examples one can mention within-subject priming and syntactic persistence in language production. Autocorrelation plays a central role in time series analysis.

1. In practice, however, it is common to test only univariate normality of each variable.

Let us check whether the reaction times data without the outliers meet these assumptions. The observations were sampled randomly. Both variables are ratio-scaled. The sample size is large enough, so we do not have to worry about normality. To test the assumption of a bivariate normal distribution, one can use the function `mvnorm.etest()` from the package `energy`. The function requires a matrix or a data frame, so we will combine the vectors as columns in a matrix:

```
> mvnorm.etest(cbind(Length_1, Mean_RT_1))

  Energy test of multivariate normality: estimated parameters

data: x, sample size 97, dimension 2, replicates 999
E-statistic = 0.485, p-value = 0.8969
```

Since the test is implemented by bootstrap, which involves drawing random samples from the data (999 replicates by default) and re-computing the test statistic, your results will be slightly different from what is shown in the output. The *p*-value is much greater than 0.05. This means that we cannot reject the null hypothesis of normality. In other words, we can consider that the assumption of bivariate normality is met. This function can be used to test not only bivariate, but also multivariate and univariate normality (see examples on the help page of the function).

Now it is time to discuss the assumption of homoscedasticity. Consider Figure 6.7, where the left panel shows a homoscedastic pattern, and the right panel displays a violation of homoscedasticity (in other words, it displays heteroscedasticity).

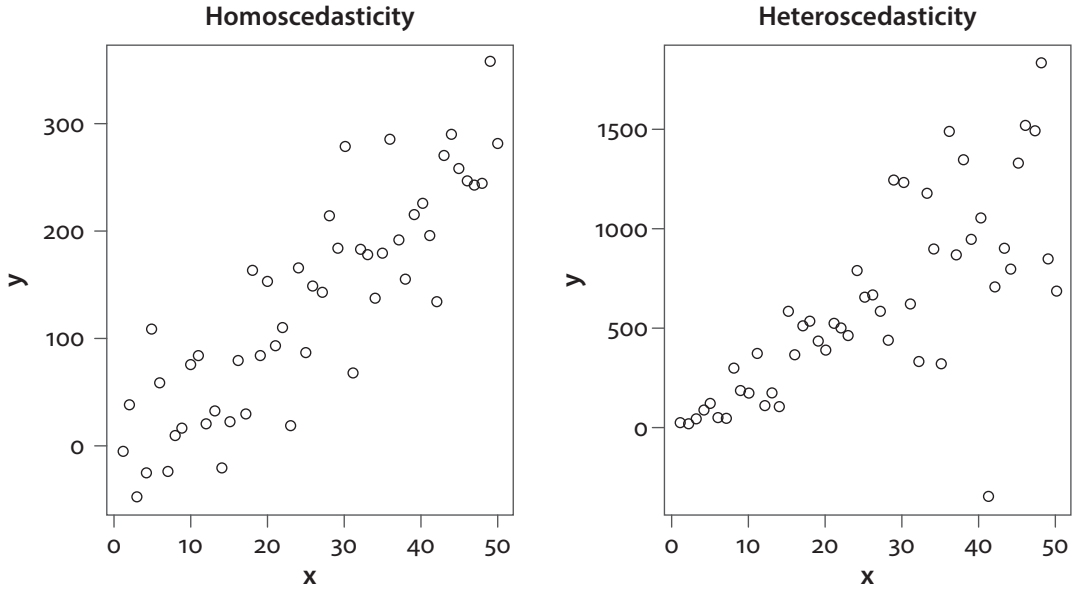


Figure 6.7. Homoscedasticity (left) and heteroscedasticity (right)

We do not find clear indications of heteroscedasticity in Figure 6.6 (three outliers should be disregarded). To perform a more formal test, one can use a function in the

package `car` created by John Fox. It is called `ncvTest()`. The abbreviation stands for 'non-constant variance test'. The main argument is a fitted linear regression model.

```
> ncvTest(m1)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.243717       Df = 1 p = 0.2647559
```

The null hypothesis of the test is that the error variance is homoscedastic. Since the *p*-value is greater than 0.05, the null hypothesis cannot be rejected. Thus, we do not have to worry about heteroscedasticity.

Finally, the residuals have to be tested for autocorrelation. We do not expect to find autocorrelation in the data, but it might be useful to show how one's data can be tested for the presence of autocorrelation. Again, we will use a function from the package `car`. The procedure is called the Durbin-Watson test.

```
> durbinWatsonTest(m1)
lag   Autocorrelation   D-W Statistic   p-value
1     0.03234626        1.923466        0.698
Alternative hypothesis: rho != 0
```

The test returns the *D–W* test statistic, which ranges between 0 and 4. The closer it is to 2, the smaller the chance of positive or negative autocorrelation. The *p*-value is very high. This means that we cannot reject the null hypothesis of no autocorrelation. Note that the test provides bootstrapped *p*-values based on many random resamples from the data, so the results will differ every time you run the test (see more in Sections 7.2.8 and 7.2.9 of Chapter 7). They also depend on the order of observations (rows) in your data.

Since all assumptions have been met, we can now use `cor.test()`. As in the *t*-test, one has to decide in advance whether we need a one- or two-tailed test. Since the alternative hypothesis is directional, a one-tailed test should be preferred. We expect the correlation between the mean reaction times and the word lengths to be positive. In other words, the alternative hypothesis of the test is that the correlation coefficient is greater than zero. Therefore, one has to add `alternative = "greater"`. If one tests a negative correlation, the alternative hypothesis would be that the correlation coefficient is less than zero, and one should add `alternative = "less"`.

```
> cor.test(Length_1, Mean_RT_1, alternative = "greater")

  Pearson's product-moment correlation

data: Length_1 and Mean_RT_1
t = 7.0965, df = 95, p-value = 1.145e-10
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
```

```
0.4667205 1.0000000
sample estimates:
cor
0.5886011
```

The function returns a lot of useful information, including the effect size (correlation coefficient 0.589), as well as the test statistic t, degrees of freedom, the p-value, and the 95% confidence interval. The p-value is very small, so the null hypothesis of no correlation can be rejected.

To summarize, the initial prediction has been borne out: the longer a word, the slower it is recognized by speakers. However, as was mentioned in Chapter 1, one should be careful with interpretation of correlations. In this situation, one can think of other factors that may influence the reaction times. For instance, longer words tend to be also less frequent, according to Zipf's Law of Abbreviation (see Chapter 5), and less frequent words are less familiar and therefore more difficult to recognize. Thus, one should take into account other potentially relevant explanatory variables, as well. This is a task for multiple linear regression, which will be introduced in the next chapter. For the present moment, we finish the case study and detach the dataset:

```
> detach(ldt)
```

Effect size versus statistical significance

It is crucial to understand the difference between effect size and statistical significance. Effect size shows how strongly different variables are related/associated, or how greatly groups of observations differ from one another. The correlation coefficient r is a good example of effect size. Statistical significance, which is associated with the p-value, does not show the strength of a relationship or the magnitude of a difference. It only shows how confident one can be that the observed relationship or difference are not due to chance alone. A strong effect does not automatically entail significance, and vice versa. Crucially, if the same effect size is observed in a smaller sample and a larger sample, the p-value will be smaller in the latter. Consider an example with two variables x and y and ten observations:

(*Continued*)

```
# do not run; the code in this box provided as an example
> x
[1]  1    2   3    4   5   6   7   8   9   10
> y
[1]  -10   5   14   4   6   7   7   9   4   10
```

The correlation coefficient is positive and moderate (r = 0.462), but not statistically significant (p = 0.179). Let us simply double the number of observations by repeating the values of x and y.

```
> x1
[1]  1   2  3   4  5  6  7  8  9  10 1   2  3   4  5  6  7  8  9  10
> y1
[1]  -10  5  14  4  6  7  7  9  4  10 -10 5  14  4  6
[16] 7   7  9   4  10
```

The correlation coefficient remains the same (r = 0.462), but the p-value becomes smaller: p = 0.04. If there is a correlation in the population, the chances of detecting it in the data increase with the sample size. Of course, one should not increase the number of observations just to be able to report some significant p-values. This would be similar to 'p-hacking', or manipulating the data or tests in order to achieve statistical significance, – an unethical but not uncommon practice in science. Moreover, it is important to remember that not every significant correlation is meaningful.

6.3 Emergence of grammar from lexicon: Spearman's ρ and Kendall's τ

6.3.1 The data and hypothesis

To perform this case study, you will need no additional packages. The R objects will be constructed from scratch as will be shown below.

Language acquisition has been a battleground for empiricists and nativists for a long time. Is grammar innate or is it learnt by children from the input? Is it autonomous or does it depend on the knowledge of lexicon? To answer these questions, Bates and Goodman (1997) investigated relationships between vocabulary size and grammatical development of young children during the period from 16 to 30 months. The vocabulary size was measured as the number of words produced by the children, and grammatical development was operationalized as the total number of selected target constructions acquired by a child (from 0 to 37). Bates and Goodman found a very high correlation between the levels of lexical and grammatical development. Their results strongly support the empiricist view of language as one dynamic system, which does not consist of separate domain-specific neural modules.

This case study will reproduce their findings with some simulation data from ten imaginary children from 16 to 30 months old. Let us create two numeric vectors: `lex` with total numbers of lexical units acquired by each child, and `gram` with grammatical complexity scores.

```
> lex <- c(47, 89, 131, 186, 245, 284, 362, 444, 553, 627)
> gram <- c(0, 2, 1, 3, 5, 9, 7, 16, 25, 34)
```

The alternative hypothesis of this study is that there is a positive correlation between the size of productive vocabulary and the complexity of grammatical structures. The hypothesis is directional. The null hypothesis states that there is no correlation between these two variables.

6.3.2 Exploring the data and computing correlation coefficients

As in the previous case study, let us begin with plotting the variables one against the other. The vocabulary size will be plotted on the *x*-axis, and the grammatical complexity scores on the *y*-axis.

```
> plot(gram ~ lex, main = "Vocabulary size and grammatical complexity",
xlab = "Productive vocabulary size", ylab = "Grammatical complexity
score")
```

Obviously, the relationship is not linear. We will also add a curved (polynomial) regression line that describes the general trend:

```
> lines(lowess(gram ~ lex))
```

The result can be seen in Figure 6.8.

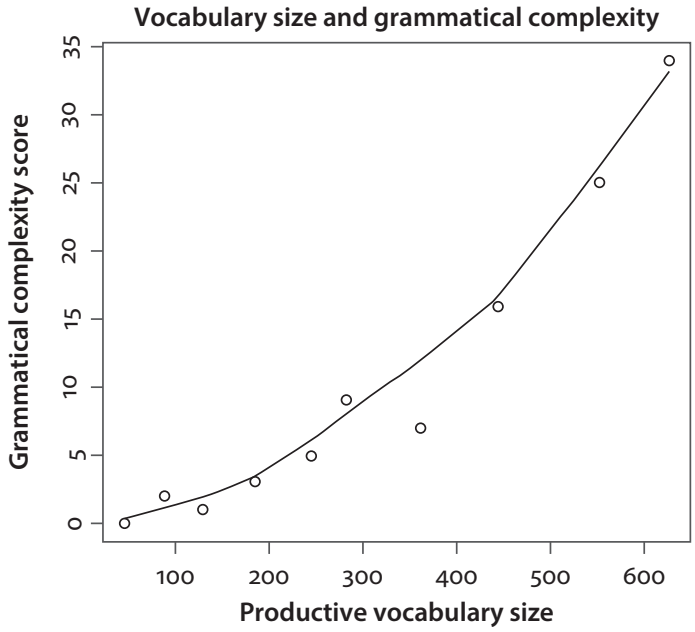


Figure 6.8. Relationship between vocabulary size and grammatical complexity scores

Although the relationship is not linear, it is clearly monotonic and positive: the more words a child uses, the more complex his or her grammatical structures are. The non-linear trend shows that one has to acquire quite a large vocabulary without any noticeable increase of complexity. However, after some critical amount of vocabulary has been learnt, complexity goes up very fast. Here we will perform a traditional correlation analysis, but see Chapter 7 on variable transformation in linear regression.

When the relationship is not linear but monotonic, one should use non-parametric correlation statistics, such as Spearman's ρ ('rho') and Kendall's τ ('tau'). To compute Spearman's ρ with the help of `cor()` or `cor.test()`, one should simply add `method = "spearman"`:

```
> cor(gram, lex, method = "spearman")
[1] 0.9757576
```

Spearman's statistic is identical with Pearson's r, when one ranks the original scores and computes r on the ranked data:

```
> cor(rank(gram), rank(lex))
[1] 0.9757576
```

The results are identical. Kendall's τ is also based on ranks, but the algorithm is different. To compute τ, one takes all pairs of ranks on the X variable and all pairs of ranks on Y. For each pair of observations, one looks at the difference in their ranks (positive or negative) on the X variable and on the Y variable. If both differences are positive, the pair of ranks is said to be concordant. If both differences are negative, the pair is considered to be concordant, as well. A pair is said to be discordant if one of the rank differences is positive, and the other is negative. The greater the proportion of concordant pairs, the higher Kendall's τ. For the same data, Kendall's τ is as follows:

```
> cor(gram, lex, method = "kendall")
[1] 0.9111111
```

Similar to Pearson's r, Spearman's ρ and Kendall's τ range from –1 (perfect negative correlation) to + 1 (perfect positive correlation). However, the scores are not identical. Normally, Kendall's τ yields less extreme values than the Spearman ρ, but this should not be a cause of concerns because this difference does not affect the statistical power of the test. That is, one has the same chances of finding a significant correlation, if it is there, as with the Spearman ρ.

Now we should check whether the correlation is statistically significant. The non-parametric tests of significance have only two main assumptions:

– *The sample is randomly drawn from the population.* This means that the subjects were selected randomly.

– *Both variables are on the ordinal scale of measurement.* If they are interval- or ratio-scaled, they will be transformed to ranks by R automatically.

Note that, as in the previous case, the relationship between *X* and *Y* should be monotonic, that is, it should not change its sign in different regions of *X* or *Y*. For non-monotonic relationships, the correlation coefficients simply do not make sense, unless you transform the data and make the relationship monotonic.

Since both assumptions are met, we can perform the significance tests. We will use the one-tailed version because our alternative hypothesis of positive correlation is directional:

```
> cor.test(gram, lex, method = "spearman", alternative = "greater")

  Spearman's rank correlation rho

data: gram and lex
S = 4, p-value < 2.2e-16
alternative hypothesis: true rho is greater than 0
sample estimates:
  rho
0.9757576

> cor.test(gram, lex, method = "kendall", alternative = "greater")

  Kendall's rank correlation tau

data: gram and lex
T = 43, p-value = 1.488e-05
alternative hypothesis: true tau is greater than 0
sample estimates:
  tau
0.9111111
```

Both tests yield statistically significant results, which is good for our research hypothesis, but which one should we report? This is often a matter of convention in a particular field of research. In addition, some statisticians say that the Kendall τ should be preferred if one has a small dataset and many ranks are tied (Field et al. 2012: 225). Ties are observed when two or more observations have identical scores and therefore identical ranks. If two scores are identical, they get an average of the ranks that they occupy. For example, 6.5 may correspond to two observations that share ranks 6 and 7.[2]

2. Of course, since correlation does not imply causation, one can also say that the correlation can be explained by a third factor, such as the children's age. However, there is evidence that the correlation remains strong even when age is partialled out (Bates and Goodman 1997: 519).

Why bother with parametric tests?

Why not simply use non-parametric tests and statistics in all situations, e.g. the Wilcoxon test instead of the *t*-test, or Spearman's ρ instead of Pearson' *r*? If parametric tests have so many assumptions to meet, why not use their non-parametric equivalents all the time?

The main reason is as follows. When one moves down the scale of measurements from interval- or ratio-scaled to ordinal data, one loses information about the actual differences between scores with different ranks. In addition, there are some exceptions to the general rule of thumb, which is to use non-parametric tests with non-normally distributed data. This depends on statistical power, that is, the probability of rejecting the false null hypothesis. Parametric tests are more powerful than their non-parametric 'colleagues' when the distributions have light (short) tails, even when the data are not normally distributed, e.g. discrete scores on a scale from 1 to 5 in an acceptability rating task. In contrast, non-parametric tests are more powerful in the presence of heavy (long) tails and outliers (Conover 1999: 116–117).

6.4 Visualization of correlations between more than two variables with the help of correlograms

To be able to reproduce the code in this case study, you will need the following add-on packages:

```
> install.packages("corrgram")
> library(Rling); library(corrgram)
```

When you have more than two quantitative variables, one can use investigate the relationships between them with the help of a correlation matrix. This is easy to do by using the function `cor()`. For example, one can create a correlation matrix of three variables in the `ldt` dataset, which was introduced in Chapter 3 and discussed in Section 6.2 of this chapter. The dataset contains 100 observations (words), which were used as stimuli in a lexical decision task. The three variables are word length in characters, word frequency in a corpus and mean reaction time of lexical decision.

```
> data(ldt)
> cor(ldt, method = "spearman")
```

```
           Length       Freq          Mean_RT
Length     1.0000000    -0.5157536    0.6266207
Freq       -0.5157536   1.0000000     -0.6733258
Mean_RT    0.6266207    -0.6733258    1.0000000
```

The correlation of each variable with itself is one. We also observe a positive correlation between word length and mean reaction time, a negative correlation between frequency and length, and a negative correlation between frequency and mean reaction time. This means, the more frequent a word, the shorter it is, and the faster it is recognized.

For those who prefer visual ways of displaying information, an attractive option might be a correlogram,[3] which provides different colours or other intuitive symbols instead of numerical values in a correlation matrix. This may be particularly helpful for investigating a large set of variables. The relationships between different scores can be even more closely inspected with the help of the `corrgram()` function in the `corrgram` package. The `corrgram()` function can be applied to the original matrix with the scores (in that case, the Pearson correlation coefficient will be computed by default) or to a correlation matrix, which can contain any correlation coefficients.

For illustration, let us create a representation with shaded panels and pie charts based on the Spearman coefficient.

```
> corrgram(ldt, lower.panel = panel.shade, upper.panel = panel.pie,
cor.method = "spearman")
```

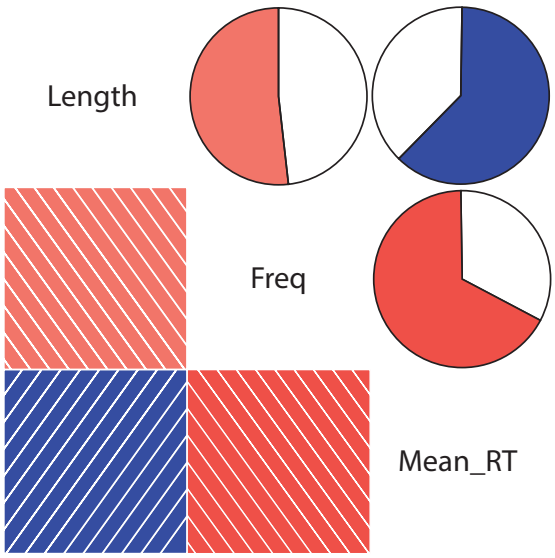


Figure 6.9. Correlogram of word length, corpus frequency and mean reaction times in a lexical decision task (ranked data): shaded panels and pies

3. Note that the term "correlogram" is also used in time series analysis to visualize autocorrelations.

Figure 6.9 displays the result. The strength of correlation is represented by the intensity of shading and also by the size of the coloured segments of the pie charts. The direction is represented by colours (blue for positive correlations, red for negative correlations), direction of shading lines in the panels (bottom left to top right for positive correlations, top left to bottom right for negative correlations), and the orientation of the coloured segments in the pie charts (clockwise for positive correlations, anticlockwise for negative correlations). When the number of variables is large, you may wish to add the argument `order = TRUE`, which orders the variables in such a way that one can see the groups of strongly correlated variables.

For the purposes of more precise diagnostics of the relationships and detection of outliers, it is also possible to visualize the observations as points plotted against the values of each variable compared, and ellipses that show the direction and strength of association. This is possible, however, only if we have original data with individual observations, rather than a matrix with correlation coefficients. Let us first create a new data frame that is identical to `ldt`, but the word frequencies are log-transformed (cf. a discussion of the purpose of this transformation in Chapter 3):

```
> ldt1 <- data.frame(Length = ldt$Length, LogFreq = log1p(ldt$Freq),
Mean_RT = ldt$Mean_RT)
> head(ldt1)
    Length   LogFreq    Mean_RT
1   8        4.882802   819.19
2   10       4.418841   977.63
3   7        0.000000   908.22
4   6        6.385194   766.30
5   12       1.098612   1125.42
6   12       2.302585   948.33
```

Now we can create the correlogram:

```
> corrgram(ldt1, lower.panel = panel.ellipse, upper.panel = panel.
pts)
```

The lower part of Figure 6.10 shows ellipses. The rounder the ellipse, the weaker the correlation. There are also the so-called LOESS smoothed curves, which show the general direction of relationships (the technique is closely related to `lowess()`, which was used in the previous section). The straighter the line, the more linear the relationship. The plots with points in the upper part of the correlogram are simple scatter plots, which can help one detect outliers. This kind of representation is a useful exploratory tool in correlation analysis and regression.

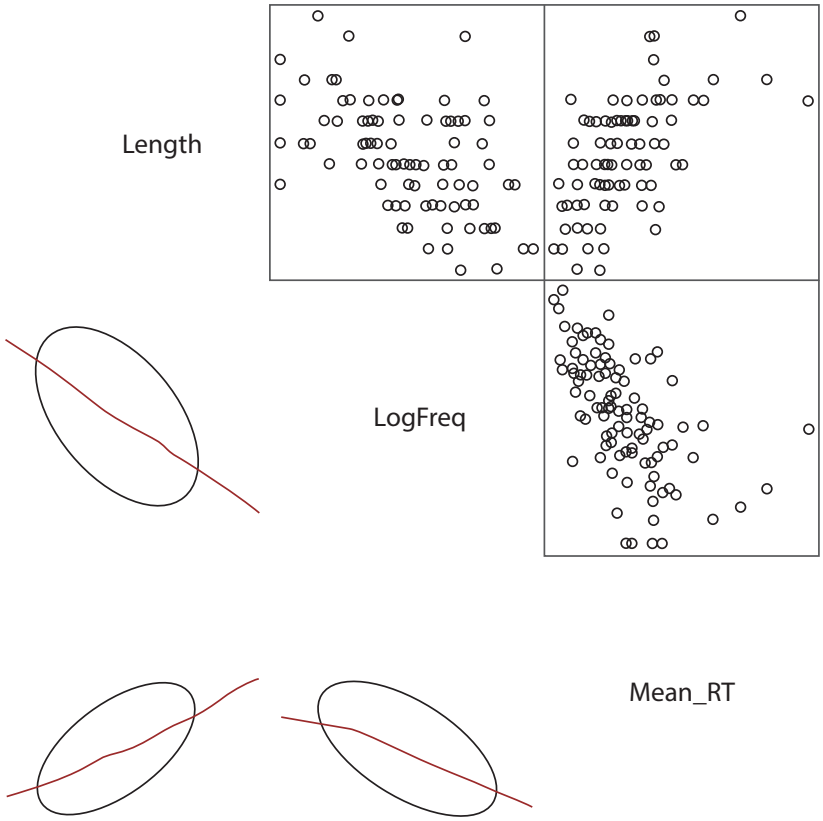


Figure 6.10. Correlogram of word length, log-transformed word frequency and mean reaction times in a lexical decision task: points (individual words) and ellipses with smoothed curves

6.5 Summary

This chapter has introduced parametric and non-parametric measures of correlation between two quantitative variables, as well as means of exploring relationships between more than two variables. Correlation analysis and linear regression are very closely related. This is why a few basic notions of linear regression modelling, such as residuals and homoscedasticity, have been discussed here, as well. The next chapter introduces linear regression analysis in greater depth. You will also learn how to perform multiple regression, which measures the relationships between a response variable and many explanatory variables simultaneously.

Writing up the results of a correlation analysis

To report the results of correlation analysis, you can use the following template, which describes the results of the first case study in this chapter: "The correlation between word length and average reaction time was positive, moderately strong and statistically significant, $r = 0.589$, $df = 95$, $p_{\text{one-tailed}} < 0.001$".