Chapter 13

Comparing Two Samples

13.1 Student Learning Objectives

The next 3 chapters deal with the statistical inference associated with the relation between two variables. The relation corresponds to the effect of one variable on the distribution of the other. The variable whose distribution is being investigated is called the *response*. The variable which may have an effect on the distribution of the response is called the *explanatory variable*.

In this section we consider the case where the explanatory variable is a factor with two levels. This factor splits the sample into two sub-samples. The statistical inference compares between the distributions of the response variable in the two sub-samples. The statistical inference involves point estimation, confidence intervals, and hypothesis testing. `R` functions may be used in order to carry out the statistical inference. By the end of this chapter, the student should be able to:

- Define estimators, confidence intervals, and tests for comparing the distribution of a numerical response between two sub-populations.
- Apply the function "`t.test`" in order to investigate the difference between the expectations of the response variable in the two sub-samples.
- Apply the function "`var.test`" in order to investigate the ratio between the variances of the response variable in the two sub-samples.

13.2 Comparing Two Distributions

Up until this point in the book we have been considering tools for the investigation of the characteristics of the distribution of a single measurement. In most applications, however, one is more interested in inference regarding the relationships between several measurements. In particular, one may want to understand how the outcome of one measurement effects the outcome of another measurement.

A common form of a mathematical relation between two variables is when one of the variables is a function of the other. When such a relation holds then

the value of the first variably is determined by the value of the second. However, in the statistical context relations between variables are more complex. Typically, a statistical relation between variables does not make one a direct function of the other. Instead, the *distribution* of values of one of the variables is affected by the value of the other variable. For a given value of the second variable the first variable may have one distribution, but for a different value of the second variable the distribution of the first variable may be different. In statistical terminology the second variable in this setting is called an *explanatory variable* and the first variable, with a distribution affected by the second variable, is called the *response*.

As an illustration of the relation between the response and the explanatory variable consider the following example. In a clinical trial, which is a precondition for the marketing of a new medical treatment, a group of patients is randomly divided into a *treatment* and a *control* sub-groups. The new treatment is anonymously administered to the treatment sub-group. At the same time, the patients in the control sub-group obtain the currently standard treatment. The new treatment passes the trial and is approved for marketing by the Health Authorities only if the response to the medical intervention is better for the treatment sub-group than it is for the control sub-group. This treatment-control experimental design, in which a response is measured under two experimental conditions, is used in many scientific and industrial settings.

In the example of a clinical trial one may identify two variables. One variable measures the response to the medical intervention for each patient that participated in the trial. This variable is the response variable, the distribution of which one seeks to investigate. The other variable indicates to which sub-group, treatment or control, each patient belongs. This variable is the explanatory variable. In the setting of a clinical trial the explanatory variable is a factor with two levels, "treatment" and "control", that splits the sample into two sub-samples. The statistical inference compares the distribution of the response variable among the patients in the treatment sub-sample to the distribution of the response among the patients in the control sub-group.

The analysis of experimental settings such as the treatment-control trial ia a special case that involves the investigation of the effect an explanatory variable may have on the response variable. In this special case the explanatory variable is a factor with two distinct levels. Each level of the factor is associate with a sub-sample, either treatment or control. The analysis seeks to compare the distribution of the response in one sub-sample with the distribution in the other sub-sample. If the response is a numeric measurement then the analysis may take the form of comparing the response's expectation in one sub-group to the expectation in the other. Alternatively, the analysis may involve comparing the variance. In a different case, if the response is the indicator of the occurrence of an event then the analysis may compare two probabilities, the probability of the event in the treatment group to the probability of the same event in the control group.

In this chapter we deal with statistical inference that corresponds to the comparison of the distribution of a numerical response variable between two sub-groups that are determined by a factor. The inference includes testing hypotheses, mainly the null hypothesis that the distribution of the response is the same in both subgroups versus the alternative hypothesis that the distribution is not the same. Another element in the inference is point estimation and

confidence intervals of appropriate parameters.

In the next chapter we will consider the case where the explanatory variable is numeric and in the subsequent chapter we describe the inference that is used in the case that the response is the indicator of the occurrence of an event.

13.3 Comparing the Sample Means

In this section we deal with the issue of statistical inference when comparing the expectation of the response variable in two sub-samples. The inference is used in order to address questions such as the equality of the two expectations to each other and, in the case they are not equal, the assessment of the difference between the expectations. For the first question one may use statistical hypothesis testing and for the assessment one may use point estimates and/or confidence intervals.

In the first subsection we provide an example of a test of the hypothesis that the expectations are equal. A confidence interval for the difference between expectations is given in the output of the report of the `R` function that applies the test. The second subsection considers the construction of the confidence interval and the third subsection deals with the theory behind the statistical test.

13.3.1 An Example of a Comparison of Means

In order to illustrate the statistical inference that compars two expectations let us return to an example that was considered in Chapter 12. The response of interest is the difference in miles-per-gallon between driving in highway conditions and driving in city conditions. This response is produced as the difference between the variable "`cars$highway.mpg`" and the variable "`cars$city.mpg`". It is stored in the object "`dif.mpg`", which is a numerical sequence:

```
> cars <- read.csv("cars.csv")
> dif.mpg <- cars$highway.mpg - cars$city.mpg
```

The object "`heavy`" was defined in the previous chapter as a sequence with logical components. A component had the value "`TRUE`" if the curb weight of the car type associated with this component was above the median level of 2,414 lb. The component obtained the value "`FALSE`" if the curb weight did not exceed that level. The logical sequence "`heavy`" was used in order to select the subsequences associated with each weight sub-group. Statistical inference was applied separately to each subsequence.

In the current analysis we want to examine directly the relation between the response variable "`dif.mpg`" and an explanatory factor variable "`heavy`". In order to do so we redefine the variable "`heavy`" to be a factor:

```
> heavy <- factor(cars$curb.weight > 2414)
```

The variable "`curb.weight`" is numeric and the expression "`cars$curb.weight > 2414`" produces a sequence with logical "`TRUE`" or "`FALSE`" components. This sequence is not a factor. In order to produce a factor we apply the function "`factor`" to the sequence. The function "`factor`" transforms its input into a factor. Specifically, the application of this function to a sequence with logical

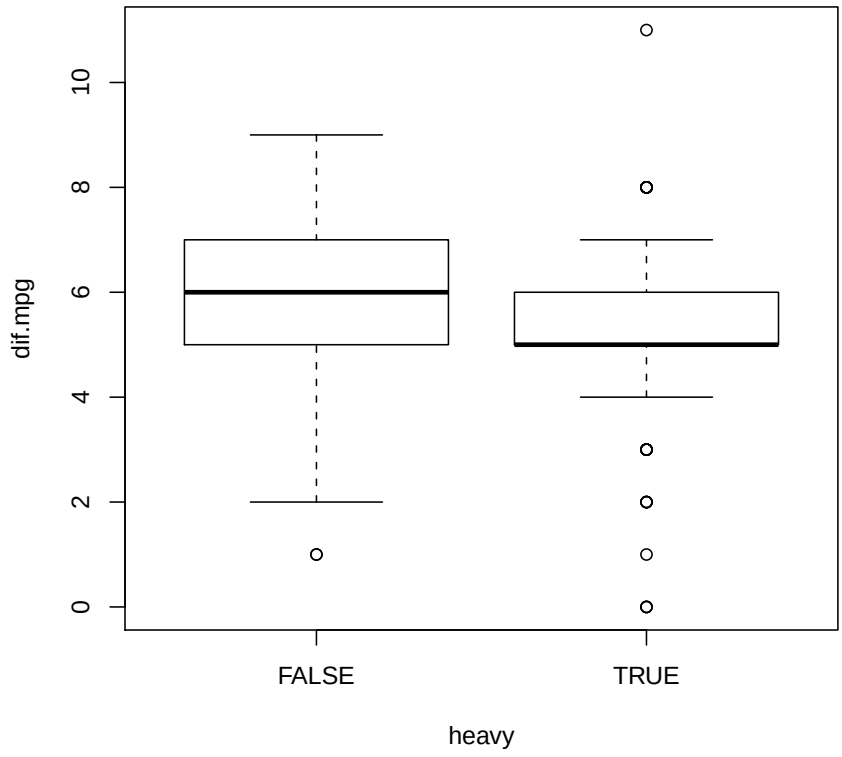

Figure 13.1: Distributions of Responses for the Weight Groups

components produces a factor with two levels that are given the names "`TRUE`" and "`FALSE`"[1].

We want to examine the relation between the response variable "`dif.mpg`" and the explanatory factor "`heavy`". Towards that end we produce a plot of the relation with the function "`plot`" and test for the equality of the expectations of the response with the function "`t.test`". First the plot:

```
> plot(dif.mpg~heavy)
```

The application of the function "`plot`" to the expression "`dif.mpg ~ heavy`" produces the plot that is given in Figure 13.1.

Observe that the figure contains two box plots, one associated with the level "`FALSE`" of the explanatory factor and the other with the level "`TRUE`" of that factor. The box plots describe the distribution of the response variable for each level of the explanatory factor. Overall, the distribution of the response for heavier cars (cars associated with the level "`TRUE`") tends to obtain smaller

[1]It should be noted that the redefined sequence "`heavy`" is no longer a sequence with logical components. It cannot be used, for example, as an index to another sequence in order to sellect the components that are associated with the "`TRUE`" logical value.

values than the distribution of the response for lighter cars (cars associated with the level "`FALSE`").

The input to the function "`plot`" is a *formula* expression of the form: "*response ~ explanatory.variable*". A formula identifies the role of variables. The variable to the left of the tilde character (`~`) in a formula is the response and the variable to the right is the explanatory variable. In the current case the variable "`dif.mpg`" is the response and the variable "`heavy`" is the explanatory variable.

Let us use a formal test in order to negate the hypothesis that the expectation of the response for the two weight groups is the same. The test is provided by the application of the function "`t.test`" to the formula "`dif.mpg~heavy`":

```
> t.test(dif.mpg~heavy)

        Welch Two Sample t-test

data:  dif.mpg by heavy
t = 2.4255, df = 191.561, p-value = 0.01621
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1029150 0.9989315
sample estimates:
mean in group FALSE  mean in group TRUE
           5.805825            5.254902
```

The function "`t.test`", when applied to a formula that describes the relation between a numeric response and a explanatory factor with two level, produces a special form of a t-test that is called the *Welch Two Sample t-test*. The statistical model associated with this test assumes the present of two independent sub-samples, each associated with a level of the explanatory variable. The relevant parameters for this model are the two expectations and the two variances associated with the sub-samples.

The hypotheses tested in the context of the Welch test are formulated in terms of the difference between the expectation of the first sub-sample and the expectation of the second sub-sample. In the default application of the test the null hypothesis is that the difference is equal to 0 (or, equivalently, that the expectations are equal to each other). The alternative is that the difference is not equal to 0 (hence, the expectations differ).

The test is conducted with the aid of a test statistic. The computed value of the test statistic in this example is "`t = 2.4255`". Under the null hypothesis the distribution of the test statistic is (approximately) equal to the t-distribution on "`df = 191.561`" degrees of freedom. The resulting p-value is "`p-value = 0.01621`". Since the computed p-value is less than 0.05 we reject the null hypothesis with a significance level of 5% and declare that the expectations are not equal to each other.

The bottom part of the report presents points estimates and a confidence interval. The point estimates of the two expectations are the sub-samples averages. The estimated value of the expected difference in miles-per-gallon for lighter cars is 5.805825, which is the average of the measurements associated with the level "`FALSE`". The estimated value of the expected difference for

heavier cars is 5.254902, the average of measurements associated with the level "`TRUE`".

The point estimate for the difference between the two expectations is the difference between the two sample averages: $5.805825 - 5.254902 = 0.550923$. A confidence interval for the *difference* between the expectations is reported under the title "`95 percent confidence interval:`". The computed value of the confidence interval is $[0.1029150, 0.9989315]$.

In the rest of this section we describe the theory behind the construction of the confidence interval and the statistical test.

13.3.2 Confidence Interval for the Difference

Consider the statistical model that is used for the construction of the confidence interval. The main issue is that the model actually deals with two populations rather than one population. In previous theoretical discussions we assumed the presence of a single population and a measurement taken for the members of this population. When the measurement was considered as a random variable it was denoted by a capital Latin letter such as X. Of concern were characteristics of the distribution of X such as $\mathrm{E}(X)$, the expectation of X, and $\mathrm{Var}(X)$, the variance.

In the current investigation two populations are considered. One population is the sub-population associated with the first level of the factor and the other population is associated with the second level. The measurement is taken for the members of both sub-populations. However, the measurement involves two random variables, one associated with the first sub-population and the other associated with the second sub-population. Moreover, the distribution of the measurement for one population may differ from the distribution for the other population. We denote the random variable associated with the first sub-population by X_a and the one associated with the other sub-population by X_b.

Consider the example in which the measurement is the difference in miles-per-gallon between highway and city driving conditions. In this example X_a is the measurement for cars with curb weight up to 2,414 lb and X_b is the same measurement for cars with curb weight above that threshold.

The random variables X_a and X_b may have different distributions. Consequently, the characteristics of their distributions may also vary. Denote by $\mathrm{E}(X_a)$ and $\mathrm{E}(X_b)$ the expectations of the first and second random variable, respectively. Likewise, $\mathrm{Var}(X_a)$ and $\mathrm{Var}(X_b)$ are the variances of the two random variables. These expectations and variances are subjects of the statistical inference.

The sample itself may also be divided into two sub-samples according to the sub-population each observation originated from. In the example, one sub-sample is associated with the lighter car types and the other sub-sample with the heavier ones. These sub-samples can be used in order to make inference with respect to the parameters of X_a and X_b, respectively. For example, the average of the observations from first sub-sample, $\bar{X}_a$, can serve as the estimator of the of the expectation $\mathrm{E}(X_a)$ and the second sub-sample's average $\bar{X}_b$ may be used in order to estimate $\mathrm{E}(X_b)$.

Our goal in this section is to construct a confidence interval for the difference in expectations $\mathrm{E}(X_a) - \mathrm{E}(X_b)$. A natural estimator for this difference in

expectations is the difference in averages $\bar{X}_a - \bar{X}_b$. The average difference will also serve as the basis for the construction of a confidence interval.

Recall that the construction of the confidence interval for a signal expectation was based on the sample average $\bar{X}$. We exploited the fact that the distribution of $Z = (\bar{X} - \mathrm{E}(X)/\sqrt{\mathrm{Var}(X)/n}$, the standardized sample average, is approximately standard Normal. From this Normal approximation we obtained an approximate 0.95 probability for the event

$$\left\{ -1.96 \cdot \sqrt{\mathrm{Var}(X)/n} \leq \bar{X} - \mathrm{E}(X) \leq 1.96 \cdot \sqrt{\mathrm{Var}(X)/n} \right\} ,$$

where $1.96 =$ `qnorm(0.975)` is the 0.975-percentile of the standard Normal distribution[2]. Substituting the estimator S for the unknown variance of the measurement and rewriting the event in a format that puts the expectation $\mathrm{E}(X)$ in the center, between two boundaries, produced the confidence interval:

$$\bar{X} \pm 1.96 \cdot S/\sqrt{n} .$$

Similar considerations can be used in the construction of a confidence interval for the difference between expectations on the basis of the difference between sub-sample averages. The deviation $\{\bar{X}_a - \bar{X}_b\} - \{\mathrm{E}(X_a) - \mathrm{E}(X_b)\}$ between the difference of the averages and the difference of the expectations that they estimate can be standardized. By the Central Limit Theorem one may obtain that the distribution of the standardized deviation is approximately standard Normal.

Standardization is obtained by dividing by the standard deviation of the estimator. In the current setting the estimator is the difference between the averages. The variance of the difference is given by

$$\mathrm{Var}(\bar{X}_a - \bar{X}_b) = \mathrm{Var}(\bar{X}_a) + \mathrm{Var}(\bar{X}_b) = \frac{\mathrm{Var}(X_a)}{n_a} + \frac{\mathrm{Var}(X_b)}{n_b} ,$$

where n_a is the size of the sub-sample that produces the sample average $\bar{X}_a$ and n_b is the size of the sub-sample that produces the sample average $\bar{X}_b$. Observe that both $\bar{X}_a$ and $\bar{X}_b$ contribute to the variability of the difference. The total variability is the sum of the two contributions[3]. Finally, we use the fact that the variance of the sample average is equal to he variance of a single measurement divided by the sample size. This fact is used for both averages in order to obtain a representation of the variance of the estimator in terms of the variances of the measurement in the two sub-population and the sizes of the two sub-samples.

The standardized deviation takes the form:

$$Z = \frac{\bar{X}_a - \bar{X}_b - \{\mathrm{E}(X_a) - \mathrm{E}(X_b)\}}{\sqrt{\mathrm{Var}(X_a)/n_a + \mathrm{Var}(X_b)/n_b}} .$$

When both sample sizes n_a and n_b are large then the distribution of Z is approximately standard Normal. As a corollary from the Normal approximation one gets that $\mathrm{P}(-1.96 \leq Z \leq 1.96) \approx 0.95$.

[2] In the case where the sample size is small and the observations are Normally distributed we used the t-distribution instead. The percentile that was used in that case was `qt(0.975,n-1)`, the 0.975 percentile of the t-distribution on $n-1$ degrees of freedom.

[3] It can be proved mathematically that the variance of a difference (or a sum) of two independent random variables is the sum of the variances. The situation is different when the two random variables are correlated.

The values of variances $\mathrm{Var}(X_a)$ and $\mathrm{Var}(X_b)$ that appear in the definition of Z are unknown. However, these values can be estimated using the sub-samples variances S_a^2 and S_b^2. When the size of both sub-samples is large then these estimators will produce good approximations of the unknown variances:

$$\mathrm{Var}(X_a) \approx S_a^2,\ \mathrm{Var}(X_b) \approx S_b^2 \quad \Longrightarrow \quad \frac{\mathrm{Var}(X_a)}{n_a} + \frac{\mathrm{Var}(X_b)}{n_b} \approx \frac{S_a^2}{n_a} + \frac{S_b^2}{n_b}\ .$$

The event $\{-1.96 \le Z \le 1.96\}$ may be approximated by the event:

$$\left\{ -1.96 \cdot \sqrt{\frac{S_a^2}{n_a} + \frac{S_b^2}{n_b}} \le \bar{X}_a - \bar{X}_b - \{\mathrm{E}(X_a) - \mathrm{E}(X_b)\} \le 1.96 \cdot \sqrt{\frac{S_a^2}{n_a} + \frac{S_b^2}{n_b}} \right\},$$

The approximation results from the use of the sub-sample variances as a substitute for the unknown variances of the measurement in the two sub-populations. When the two sample sizes n_a and n_b are large then the probability of the given event will also be approximately equal to 0.95.

Finally, reexpressing the least event in a format that puts the parameter $\mathrm{E}(X_a) - \mathrm{E}(X_b)$ in the center will produce the confidence interval with boundaries of the form:

$$\bar{X}_a - \bar{X}_b \pm 1.96 \cdot \sqrt{S_a^2/n_a + S_b^2/n_b}$$

In order to illustrate the computations that are involved in the construction of a confidence interval for the difference between two expectations let us return to the example of difference in miles-per-gallon for lighter and for heavier cars. Compute the two sample sizes, sample averages, and sample variances:

```
> table(heavy)
heavy
FALSE  TRUE
  103   102
> tapply(dif.mpg,heavy,mean)
   FALSE     TRUE
5.805825 5.254902
> tapply(dif.mpg,heavy,var)
   FALSE     TRUE
2.020750 3.261114
```

Observe that there 103 lighter cars and 102 heavier ones. These counts were obtained by the application of the function "`table`" to the factor "`heavy`". The lighter cars are associated with the level "`FALSE`" and heavier cars are associated with the level "`TRUE`".

The average difference in miles-per-gallon for lighter cars is 5.805825 and the variance is 2.020750. The average difference in miles-per-gallon for heavier cars is 5.254902 and the variance is 3.261114. These quantities were obtained by the application of the functions "`mean`" or "`var`" to the values of the variable "`dif.mpg`" that are associated with each level of the factor "`heavy`". The application was carried out using the function "`tapply`".

The computed values of the means are equal to the vales reported in the output of the application of the function "`t.test`" to the formula "`dif.mpg~heavy`". The difference between the averages is $\bar{x}_a - \bar{x}_b = 5.805825 - 5.254902 = 0.550923$.

This value is the center of the confidence interval. The estimate of the standard deviation of the difference in averages is:

$$\sqrt{s_a^2/n_a + s_b^2/n_b} = \sqrt{2.020750/103 + 3.261114/102} = 0.227135 \ .$$

Therefore, the confidence interval for the difference in expectations is

$$\bar{x}_a - \bar{x}_b \pm 1.96 \cdot \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}} = 0.550923 \pm 1.96 \cdot 0.227135 = [0.1057384, 0.9961076] \ ,$$

which is (essentially) the confidence interval that is presented in the report[4].

13.3.3 The t-Test for Two Means

The statistical model that involves two sub-populations may be considered also in the context of hypothesis testing. Hypotheses can be formulated regarding the relations between the parameters of the model. These hypotheses can be tested using the data. For example, in the current application of the t-test, the null hypothesis is $H_0 : \mathrm{E}(X_a) = \mathrm{E}(X_b)$ and the alternative hypothesis is $H_1 : \mathrm{E}(X_a) \neq \mathrm{E}(X_b)$. In this subsection we explain the theory behind this test.

Recall that the construction of a statistical test included the definition of a test statistic and the determination of a rejection region. The null hypothesis is rejected if, and only if, the test statistic obtains a value in the rejection region. The determination of the rejection region is based on the sampling distribution of the test statistic under the null hypothesis. The significance level of the test is the probability of rejecting the null hypothesis (i.e., the probability that the test statistic obtains a value in the rejection region) when the null hypothesis is correct (the distribution of the test statistic is the distribution under the null hypothesis). The significance level of the test is set at a given value, say 5%, thereby restricting the size of the rejection region.

In the previous chapter we consider the case where there is one population. For review, consider testing the hypothesis that the expectation of the measurement is equal to zero ($H_0 : \mathrm{E}(X) = 0$) against the alternative hypothesis that it is not ($H_1 : \mathrm{E}(X) \neq 0$). A sample of size n is obtained from this population. Based on the sample one may compute a test statistic:

$$T = \frac{\bar{X} - 0}{S/\sqrt{n}} = \frac{\bar{X}}{S/\sqrt{n}} \ ,$$

where $\bar{X}$ is the sample average and S is the sample standard deviation. The rejection region of this test is $\{|T| > \texttt{qt(0.975,n-1)}\}$, for "`qt(0.975,n-1)`" the 0.975-percentile of the t-distribution on $n-1$ degrees of freedom.

Alternatively, one may compute the p-value and reject the null hypothesis if the p-value is less than 0.05. The p-value in this case is equal to $\mathrm{P}(|T| > |t|)$,

[4]The confidence interval given in the output of the function "`t.test`" is $[0.1029150, 0.9989315]$, which is very similar, but not identical, to the confidence interval that we computed. The discrepancy stems from the selection of the percentile. We used the percentile of the normal distribution $1.96 =$ `qnorm(0.975)`. The function "`t.test`", on the other hand, uses the percentile of the t-distribution $1.972425 =$ `qt(0.975,191.561)`. Using this value instead would give $0.550923 \pm 1.972425 \cdot 0.227135$, which coincides with the interval reported by "`t.test`". For practical applications the difference between the two confidence intervals are not negligible.

where t is the computed value of the test statistic. The distribution of T is the t-distribution of $n-1$ degrees of freedom.

A similar approach can be used in the situation where two sub-population are involved and one wants to test the null hypothesis that the expectations are equal versus the alternative hypothesis that they are not. The null hypothesis can be written in the form $H_0 : \mathrm{E}(X_a) - \mathrm{E}(X_b) = 0$ with the alternative hypothesis given as $H_1 : \mathrm{E}(X_a) - \mathrm{E}(X_b) \neq 0$.

It is natural to base the test static on the difference between sub-samples averages $\bar{X}_a - \bar{X}_b$. The T statistic is the ratio between the deviation of the estimator from the null value of the parameter, divided by the (estimated) standard deviation of the estimator. In the current setting the estimator is difference in sub-samples averages $\bar{X}_a - \bar{X}_b$, the null value of the parameter, the difference between the expectations, is 0, and the (estimated) standard deviation of the estimator is $\sqrt{S_a^2/n_a + S_b^2/n_b}$. It turns out that the test statistic in the current setting is:

$$T = \frac{\bar{X}_a - \bar{X}_b - 0}{\sqrt{S_a^2/n_a + S_b^2/n_b}} = \frac{\bar{X}_a - \bar{X}_b}{\sqrt{S_a^2/n_a + S_b^2/n_b}} \ .$$

Consider as a measurement the difference in miles-per-gallon. Define the sub-population a to the lighter cars and the sub-population b to be the heavier cars. Recall that the sub-sample sizes are $n_a = 103$ and $n_b = 102$. Also, the sub-sample averages are $\bar{x}_a = 5.805825$ and $\bar{x}_b = 5.254902$, and the sub-sample variances are $s_a^2 = 2.020750$ and $s_b^2 = 5.254902$.

In order to calculate the observed value of the test statistic we use once more the fact that the difference between the averages is $\bar{x}_a - \bar{x}_b = 5.805825 - 5.254902 = 0.550923$ and the estimated value of the standard deviation of the sub-samples average difference is:

$$\sqrt{s_a^2/n_a + s_b^2/n_b} = \sqrt{2.020750/103 + 3.261114/102} = 0.227135 \ .$$

It follows that the observed value of the T statistic is

$$t = \frac{\bar{x}_a - \bar{x}_b}{\sqrt{s_a^2/n_a + s_b^2/n_b}} = \frac{0.550923}{0.227135} = 2.425531 \ ,$$

which, after rounding up, is equal to the value presented in the report that was produced by the function "`t.test`".

The p-value is computed as the probability of obtaining values of the test statistic more extreme than the value that was obtained in our data. The computation is carried out under the assumptions of the null hypothesis. The limit distribution of the T statistic, when both sub-sample sizes n_a and n_b are large, is standard Normal. In the case when the measurements are Normally distributed then a refined approximation of the distribution of the statistic is the t-distribution. Both the standard Normal and the t-distribution are symmetric about the origin.

The probability of obtaining a value in either tails for a symmetric distribution is equal to twice the probability of obtaining a value in the upper tail:

$$\mathrm{P}(|T| > 2.4255) = 2 \times \mathrm{P}(T > 2.4255) = 2 \times \left[1 - \mathrm{P}(T \leq 2.4255)\right] \ .$$

The function "`t.test`" computes the p-value using the t-distribution. For the current data, the number of degrees of freedom that are used in this approximation[5] is `df` $= 191.561$. When we apply the function "`pt`" for the computation of the cumulative probability of the t-distribution we get:

```
> 2*(1-pt(2.4255,191.561))
[1] 0.01621458
```

which (after rounding) is equal to the reported p-value of 0.01621. This p-value is less than 0.05, hence the null hypothesis is rejected in favor of the alternative hypothesis that assumes an effect of the weight on the expectation.

13.4 Comparing Sample Variances

In the previous section we discussed inference associated with the comparison of the expectations of a numerical measurement between two sub-population. Inference included the construction of a confidence interval for the difference between expectations and the testing of the hypothesis that the expectations are equal to each other.

In this section we consider a comparisons between variances of the measurement in the two sub-populations. For this inference we consider the ratio between estimators of the variances and introduce a new distribution, the F-distribution, that is associated with this ratio.

Assume, again, the presence of two sub-populations, denoted a and b. A numerical measurement is taken over a sample. The sample can be divided into two sub-samples according to the sub-population of origin. In the previous section we were interested in inference regarding the relation between the expectations of the measurement in the two sub-populations. Here we are concerned with the comparison of the variances.

Specifically, let X_a be the measurement at the first sub-population and let X_b be the measurement at the second sub-population. We want to compare $\mathrm{Var}(X_a)$, the variance in the first sub-population, to $\mathrm{Var}(X_b)$, the variance in the second sub-population. As the basis for the comparison we may use S_a^2 and S_b^2, the sub-samples variances, which are computed from the observations in the first and the second sub-sample, respectively.

Consider the confidence interval for the ratio of the variances. In Chapter 11 we discussed the construction of the confidence interval for the variance in a single sample. The derivation was based on the sample variance S^2 that serves as an estimator of the population variance $\mathrm{Var}(X)$. In particular, the distribution of the random variable $(n-1)S^2/\mathrm{Var}(X)$ was identified as the chi-square distribution on $n-1$ degrees of freedom[6]. A confidence interval for the variance was obtained as a result of the identification of a central region in the chi-square distribution that contains a pre-subscribed probability[7].

[5] The Weltch t-test for the comparison of two means uses the t-distribution as an approximation of the null distribution of the T test statistic. The number of degrees of freedom is computed by the formula: `df` $= (v_a + v_b)^2/\{v_a^2/(n_a-1) + v_b^2/(n_b-1)\}$, where $v_a = s_a^2/n_a$ and $v_b = s_b^2/n_b$.

[6] This statement holds when the distribution of the measurement is Normal.

[7] Use P(`qchisq(0.025,n-1)` $\leq (n-1)S^2/\mathrm{Var}(X) \leq$ `qchisq(0.975,n-1)`) $= 0.95$ and rewrite the event in a format that puts the parameter in the center. The resulting 95% confidence interval is $[(n-1)S^2/$`qchisq(0.975,n-1)`$, (n-1)S^2/$`qchisq(0.025,n-1)`$]$.

In order to construct a confidence interval for the ratio of the variances we consider the random variable that is obtained as a ratio of the estimators of the variances:

$$\frac{S_a^2/\mathrm{Var}(X_a)}{S_b^2/\mathrm{Var}(X_b)} \sim F_{(n_a-1,n_b-1)} \ .$$

The distribution of this random variable is denoted the F-distribution[8]. This distribution is characterized by the number of degrees of freedom associated with the estimator of the variance at the numerator and by the number of degrees of freedom associated with the estimator of the variance at the denominator. The number of degrees of freedom associated with the estimation of each variance is the number of observation used for the computation of the estimator, minus 1. In the current setting the numbers of degrees of freedom are $n_a - 1$ and $n_b - 1$, respectively.

The percentiles of the F-distribution can be computed in `R` using the function "`qf`". For example, the 0.025-percentile of the distribution for the ratio between sample variances of the response for two sub-samples is computed by the expression "`qf(0.025,dfa,dfb)`", where `dfa` $= n_a - 1$ and `dfb` $= n_b - 1$. Likewise, the 0.975-percentile is computed by the expression "`qf(0.975,dfa,dfb)`". Between these two numbers lie 95% of the given F-distribution. Consequently, the probability that the random variable $\{S_a^2/\mathrm{Var}(X_a)\}/\{S_b^2/\mathrm{Var}(X_b)\}$ obtains its values between these two percentiles is equal to 0.95:

$$\frac{S_a^2/\mathrm{Var}(X_a)}{S_b^2/\mathrm{Var}(X_b)} \sim F_{(n_a-1,n_b-1)} \quad \Longrightarrow$$

$$\mathrm{P}\big(\texttt{qf(0.025,dfa,dfb)} \le \tfrac{S_a^2/\mathrm{Var}(X_a)}{S_b^2/\mathrm{Var}(X_b)} \le \texttt{qf(0.975,dfa,dfb)}\big) = 0.95 \ .$$

A confidence interval for the ratio between $\mathrm{Var}(X_a)$ and $\mathrm{Var}(X_b)$ is obtained by reformulation of the last event. In the reformulation, the ratio of the variances is placed in the center:

$$\left\{ \frac{S_a^2/S_b^2}{\texttt{qf(0.975,dfa,fdb)}} \le \frac{\mathrm{Var}(X_a)}{\mathrm{Var}(X_b)} \le \frac{S_a^2/S_b^2}{\texttt{qf(0.025,dfa,dfb)}} \right\} .$$

This confidence interval has a significance level of 95%.

Next, consider testing hypotheses regarding the relation between the variances. Of particular interest is testing the equality of the variances. One may formulate the null hypothesis as $H_0 : \mathrm{Var}(X_a)/\mathrm{Var}(X_b) = 1$ and test it against the alternative hypothesis $H_1 : \mathrm{Var}(X_a)/\mathrm{Var}(X_b) \neq 1$.

The statistic $F = S_a^2/S_b^2$ can used in order to test the given null hypothesis. Values of this statistic that are either much larger or much smaller than 1 are evidence against the null hypothesis and in favor of the alternative hypothesis. The sampling distribution, under that null hypothesis, of this statistic is the $F_{(n_a-1,n_b-1)}$ distribution. Consequently, the null hypothesis is rejected either if $F <$ `qf(0.025,dfa,dfb)` or if $F >$ `qf(0.975,dfa,dfb)`, where `dfa` $= n_a - 1$ and `dfb` $= n_b - 1$. The significance level of this test is 5%.

Given an observed value of the statistic, the p-value is computed as the significance level of the test which uses the observed value as the threshold. If

[8]The F distribution is obtained when the measurement has a Normal distribution. When the distribution of the measurement is not Normal then the distribution of the given random variable will not be the F-distribution.

the observed value f is less than 1 then the p-value is twice the probability of the lower tail: $2 \cdot \mathrm{P}(F < f)$. On the other hand, if f is larger than 1 one takes twice the upper tail as the p-value: $2 \cdot \mathrm{P}(F > f) = 2 \cdot [1 - \mathrm{P}(F \leq f)]$. The null hypothesis is rejected with a significance level of 5% if the p-value is less than 0.05.

In order to illustrate the inference that compares variances let us return to the variable "`dif.mpg`" and compare the variances associated with the two levels of the factor "`heavy`". The analysis will include testing the hypothesis that the two variances are equal and an estimate and a confidence interval for their ratio.

The function "`var.test`" may be used in order to carry out the required tasks. The input to the function is a formula such "`dif.mpg ~ heavy`", with a numeric variable on the left and a factor with two levels on the right. The default application of the function to the formula produces the desired test and confidence interval:

```
> var.test(dif.mpg~heavy)

        F test to compare two variances

data:  dif.mpg by heavy
F = 0.6197, num df = 102, denom df = 101, p-value = 0.01663
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4189200 0.9162126
sample estimates:
ratio of variances
         0.6196502
```

Consider the report produced by the function. The observed value of the test statistic is "`F = 0.6197`", and it is associated with the F-distribution on "`num df = 102`" and "`denom df = 101`" degrees of freedom. The test statistic can be used in order to test the null hypothesis $H_0 : \mathrm{Var}(X_a)/\mathrm{Var}(X_b) = 1$, that states that the two variance are equal, against the alternative hypothesis that they are not. The p-value for this test is "`p-value = 0.01663`", which is less than 0.05. Consequently, the null hypothesis is rejected and the conclusion is that the two variances are significantly different from each other. The estimated ratio of variances, given at the bottom of the report, is 0.6196502. The confidence interval for the ratio is reported also and is equal to $[0.4189200, 0.9162126]$.

In order to relate the report to the theoretical discussion above let us recall that the sub-samples variances are $s_a^2 = 2.020750$ and $s_b^2 = 3.261114$. The sub-samples sizes are $n_a = 103$ and $n_b = 102$, respectively. The observed value of the statistic is the ratio $s_a^2/s_b^2 = 2.020750/3.261114 = 0.6196502$, which is the value that appears in the report. Notice that this is the estimate of the ration between the variances that is given at the bottom of the report.

The p-value of the two-sided test is equal to twice the probability of the tail that is associated with the observed value of the test statistic as a threshold. The number of degrees of freedom is `dfa` $= n_a - 1 = 102$ and `dfb` $= n_b - 1 = 101$. The observed value of the ratio test statistic is $f = 0.6196502$. This value is less than one. Consequently, the probability $\mathrm{P}(F < 0.6196502)$ enters into the computation of the p-value, which equals twice this probability:

```
> 2*pf(0.6196502,102,101)
[1] 0.01662612
```

Compare this value to the p-value that appears in the report and see that, after rounding up, the two are the same.

For the confidence interval of the ratio compute the percentiles of the F distribution:

```
> qf(0.025,102,101)
[1] 0.676317
> qf(0.975,102,101)
[1] 1.479161
```

The confidence interval is equal to:

$$\left[\frac{s_a^2/s_b^2}{\texttt{qf(0.975,102,101)}}, \frac{s_a^2/s_b^2}{\texttt{qf(0.025,102,101)}}\right] = \left[\frac{0.6196502}{1.479161}, \frac{0.6196502}{0.676317}\right] = [0.4189200, 0.9162127] ,$$

which coincides with the reported interval.

13.5 Solved Exercises

Question 13.1. In this exercise we would like to analyze the results of the trial that involves magnets as a treatment for pain. The trial is described in Question 9.1. The results of the trial are provided in the file "`magnets.csv`".

Patients in this trail where randomly assigned to a treatment or to a control. The responses relevant for this analysis are either the variable "`change`", which measures the difference in the score of pain reported by the patients before and after the treatment, or the variable "`score1`", which measures the score of pain before a device is applied. The explanatory variable is the factor "`active`". This factor has two levels, level "`1`" to indicate the application of an active magnet and level "`2`" to indicate the application of an inactive placebo.

In the following questions you are required to carry out tests of hypotheses. All tests should conducted at the 5% significance level:

1. Is there a significance difference between the treatment and the control groups in the expectation of the reported score of pain before the application of the device?

2. Is there a significance difference between the treatment and the control groups in the variance of the reported score of pain before the application of the device?

3. Is there a significance difference between the treatment and the control groups in the expectation of the change in score that resulted from the application of the device?

4. Is there a significance difference between the treatment and the control groups in the variance of the change in score that resulted from the application of the device?

Solution (to Question 13.1.1): The score of pain before the application of the device is measured in the variable "`score1`". This variable is used as the response. We apply the function "`t.test`" in order to test the equality of the expectation of the response in the two groups. First we read in the data from the file into a data frame and then we apply the test:

```
> magnets <- read.csv("magnets.csv")
> t.test(magnets$score1 ~ magnets$active)

        Welch Two Sample t-test

data:  magnets$score1 by magnets$active
t = 0.4148, df = 38.273, p-value = 0.6806
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3757896  0.5695498
sample estimates:
mean in group "1" mean in group "2"
          9.62069           9.52381
```

The computed p-value is 0.6806, which is above 0.05. Consequently, we do not reject the null hypothesis that the expectations in the two groups are equal. This should not come as a surprise, since patients were assigned to the groups randomly and without knowledge to which group they belong. Prior to the application of the device, the two groups should look alike.

Solution (to Question 13.1.2): Again, we use the variable "`score1`" as the response. Now apply the function "`var.test`" in order to test the equality of the variances of the response in the two groups:

```
> var.test(magnets$score1 ~ magnets$active)

        F test to compare two variances

data:  magnets$score1 by magnets$active
F = 0.695, num df = 28, denom df = 20, p-value = 0.3687
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2938038 1.5516218
sample estimates:
ratio of variances
         0.6950431
```

The computed p-value is 0.3687, which is once more above 0.05. Consequently, we do not reject the null hypothesis that the variances in the two groups are equal. This fact is reassuring. Indeed, prior to the application of the device, the two groups have the same characteristics. Therefore, any subsequent difference between the two groups can be attributed to the difference in the treatment.

Solution (to Question 13.1.3): The difference in score between the treatment and the control groups is measured in the variable "`change`". This variable is

used as the response for the current analysis. We apply the function "`t.test`" in order to test the equality of the expectation of the response in the two groups:

```
> t.test(magnets$change ~ magnets$active)

        Welch Two Sample t-test

data:  magnets$change by magnets$active
t = 5.9856, df = 42.926, p-value = 3.86e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.749137 5.543145
sample estimates:
mean in group "1" mean in group "2"
         5.241379          1.095238
```

The computed p-value is 3.86×10^{-7}, which is much below 0.05. Consequently, we reject the null hypothesis that the expectations in the two groups are equal. The conclusion is that, according to this trial, magnets do have an effect on the expectation of the response[9].

Solution (to Question 13.1.4): Once more we consider the variable "`change`" as the response. We apply the function "`var.test`" in order to test the equality of the variances of the response in the two groups:

```
> var.test(magnets$change ~ magnets$active)

        F test to compare two variances

data:  magnets$change by magnets$active
F = 4.2062, num df = 28, denom df = 20, p-value = 0.001535
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.778003 9.389902
sample estimates:
ratio of variances
          4.206171
```

The computed p-value is 0.001535, which is much below 0.05. Consequently, we reject the null hypothesis that the variances in the two groups are equal. Hence, magnets also affect the variance of the response.

Question 13.2. It is assumed, when constructing the F-test for equality of variances, that the measurements are Normally distributed. In this exercise we what to examine the robustness of the test to divergence from the assumption. You are required to compute the significance level of a two-sided F-test of $H_0 : \mathrm{Var}(X_a) = \mathrm{Var}(X_b)$ versus $H_1 : \mathrm{Var}(X_a) \neq \mathrm{Var}(X_b)$. Assume there are $n_a = 29$ observations in one group and $n_b = 21$ observations in the other group. Use an F-test with a nominal 5% significance level.

[9] The evaluation of magnets as a treatment for pain produced mixed results and there is a debate regarding their effectiveness. More information can be found in the NIH NCCAM site.

1. Consider the case where $X \sim \text{Normal}(4, 4^2)$.
2. Consider the case where $X \sim \text{Exponential}(1/4)$.

Solution (to Question 13.2.1): We simulate the sampling distribution of the sample standard deviation for two samples, one sample of size $n_a = 29$ and the other of size $n_b = 21$. Both samples are simulated from the given Normal distribution:

```
> mu <- 4
> sig <- 4
> n.a <- 29
> n.b <- 21
> S.a <- rep(0,10^5)
> S.b <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+    X.a <- rnorm(n.a,mu,sig)
+    X.b <- rnorm(n.b,mu,sig)
+    S.a[i] <- sd(X.a)
+    S.b[i] <- sd(X.b)
+ }
> F <- S.a^2/S.b^2
> mean((F < qf(0.025,n.a-1,n.b-1))|(F > qf(0.975,n.a-1,n.b-1)))
[1] 0.05074
```

We compute the test statistic "`F`" as the ratio of the two sample standard deviations "`S.a`" and "`S.b`". The last expression computes the probability that the test statistic is either less than "`qt(0.025,n.a-1,n.b-1)`", or it is larger than "`qt(0.975,n.a-1,n.b-1)`". The term "`qt(0.025,n.a-1,n.b-1)`" is the 0.025-percentile of the F-distribution on 28 and 20 degrees of freedom and the term "`qt(0.975,n.a-1,n.b-1)`" is the 0.975-percentile of the same F-distribution. The result of the last expression is the actual significance level of the test.

We obtain that the actual significance level of the test when the measurements are Normally distributed is 0.05074, which in agreement with the nominal significance level of 5%. Indeed, the nominal significance level is computed under the assumption that the distribution of the measurement is Normal.

Solution (to Question 13.2.2): We repeat essentially the same simulations as before. We only change the distribution of the samples from the Normal to the Exponential distribution:

```
> lam <- 1/4
> n.a <- 29
> n.b <- 21
> S.a <- rep(0,10^5)
> S.b <- rep(0,10^5)
> for(i in 1:10^5)
+ {
+    X.a <- rexp(n.a,lam)
+    X.b <- rexp(n.b,lam)
+    S.a[i] <- sd(X.a)
```

```
+    S.b[i] <- sd(X.b)
+ }
> F <- S.a^2/S.b^2
> mean((F < qf(0.025,n.a-1,n.b-1))|(F > qf(0.975,n.a-1,n.b-1)))
[1] 0.27596
```

The actual significance level of the test is 0.27596, which is much higher than the nominal significance level of 5%.

Through this experiment we may see that the F-test is not robust to the divergence from the assumed Normal distribution of the measurement. If the distribution of the measurement is skewed (the Exponential distribution is an example of such skewed distribution) then the application of the test to the data may produce unreliable conclusions.

Question 13.3. The sample average in one sub-sample is $\bar{x}_a = 124.3$ and the sample standard deviation is $s_a = 13.4$. The sample average in the second sub-sample is $\bar{x}_b = 80.5$ and the sample standard deviation is $s_b = 16.7$. The size of the first sub-sample is $n_a = 15$ and this is also the size of the second sub-sample. We are interested in the estimation of the ratio of variances $\mathrm{Var}(X_a)/\mathrm{Var}(X_b)$.

1. Compute the estimate of parameter of interest.
2. Construct a confidence interval, with a confidence level of 95%, to the value of the parameter of interest.
3. It is discovered that the size of each of the sub-samples is actually equal to 150, and no to 15 (but the values of the other quantities are unchanged). What is the corrected estimate? What is the corrected confidence interval?

Solution (to Question 13.3.1): We input the data to `R` and then compute the estimate:

```
> s.a <- 13.4
> s.b <- 16.7
> s.a^2/s.b^2
[1] 0.6438381
```

The estimate is equal to the ratio of the sample variances s_a^2/s_b^2. It obtains the value 0.6438381. Notice that the information regarding the sample averages and the sizes of the sub-samples is not relevant for the point estimation of the parameter.

Solution (to Question 13.3.2): We use the formula:

$$\left[(s_a^2/s_b^2)/\texttt{qf(0.975,14,14)}, (s_a^2/s_b^2)/\texttt{qf(0.025,14,14)}\right]$$

in order to compute the confidence interval:

```
> n.a <- 15
> n.b <- 15
> (s.a^2/s.b^2)/qf(0.975,n.a-1,n.b-1)
[1] 0.2161555
> (s.a^2/s.b^2)/qf(0.025,n.a-1,n.b-1)
[1] 1.917728
```

The confidence interval we obtain is $[0.2161555, 1.917728]$.

Solution (to Question 13.3.3): The estimate of the parameter is not affected by the change in the sample sizes and it is still equal to 0.6438381. For the confidence interval we use now the formula:

$$\left[(s_a^2/s_b^2)/\texttt{qf(0.975,149,149)}, (s_a^2/s_b^2)/\texttt{qf(0.025,149,149)}\right] :$$

```
> n.a <- 150
> n.b <- 150
> (s.a^2/s.b^2)/qf(0.975,n.a-1,n.b-1)
[1] 0.466418
> (s.a^2/s.b^2)/qf(0.025,n.a-1,n.b-1)
[1] 0.8887467
```

The corrected confidence interval is $[0.466418, 0.8887467]$.

13.6 Summary

Glossary

Response: The variable who's distribution one seeks to investigate.

Explanatory Variable: A variable that may affect the distribution of the response.

Discuss in the forum

Statistics has an important role in the analysis of data. However, some claim that the more important role of statistics is in the design stage when one decides how to collect the data. Good design may improve the chances that the eventual inference of the data will lead to a meaningful and trustworthy conclusion.

Some say that the quantity of data that is collected is most important. Other say that the quality of the data is more important than the quantity. What is your opinion?

When formulating your answer it may be useful to come up with an example from your past experience where the quantity of data was not sufficient. Else, you can describe a case where the quality of the data was less than satisfactory. How did these deficiencies affected the validity of the conclusions of the analysis of the data?

For illustration consider the surveys. Conducting the survey by the telephone may be a fast way to reach a large number of responses. However, the quality of the response may be less that the response obtained by face-to-face interviews.

Formulas:

- Test statistic for equality of expectations: $t = (\bar{x}_a - \bar{x}_b)/\sqrt{s_a^2/n_a + s_b^2/n_b}$.
- Confidence interval: $(\bar{x}_a - \bar{x}_b) \pm \texttt{qnorm(0.975)}\sqrt{s_a^2/n_a + s_b^2/n_b}$.
- Test statistic for equality of variances: $f = s_a^2/s_b^2$.

- Confidence interval:

$$\left[(s_a^2/s_b^2)/\texttt{qf(0.975,dfa,dfb)}, (s_a^2/s_b^2)/\texttt{qf(0.025,dfa,dfb)}\right].$$

Chapter 14

Linear Regression

14.1 Student Learning Objectives

In the previous chapter we examined the situation where the response is numeric and the explanatory variable is a factor with two levels. This chapter deals with the case where both the response and the explanatory variables are numeric. The method that is used in order to describe the relations between the two variables is *regression*. Here we apply *linear regression* to deal with a linear relation between two numeric variables. This type of regression fits a line to the data. The line summarizes the effect of the explanatory variable on the distribution of the response.

Statistical inference can be conducted in the context of regression. Specifically, one may fit the regression model to the data. This corresponds to the point estimation of the parameters of the model. Also, one may produce confidence intervals for the parameters and carry out hypotheses testing. Another issue that is considered is the assessment of the percentage of variability of the response that is explained by the regression model.

By the end of this chapter, the student should be able to:

- Produce scatter plots of the response and the explanatory variable.
- Explain the relation between a line and the parameters of a linear equation. Add lines to a scatter plot.
- Fit the linear regression to data using the function "`lm`" and conduct statistical inference on the fitted model.
- Explain the relations among R^2, the percentage of response variability explained by the regression model, the variability of the regression residuals, and the variance of the response.

14.2 Points and Lines

In this section we consider the graphical representation of the response and the explanatory variables on the same plot. The data associated with both variables is plotted as points in a two-dimensional plane. Linear equations

can be represented as lines on the same two-dimensional plane. This section prepares the background for the discussion of the linear regression model. The actual model of linear regression is introduced in the next section.

14.2.1 The Scatter Plot

Consider two numeric variables. A scatter plot can be used in order to display the data in these two variables. The scatter plot is a graph in which each observation is represented as a point. Examination of the scatter plot may revile relations between the two variables.

Consider an example. A marine biologist measured the length (in millimeters) and the weight (in grams) of 10 fish that where collected in one of her expeditions. The results are summarized in a data frame that is presented in Table 14.2.1. Notice that the data frame contains 10 observations. The variable x corresponds to the length of the fish and the variable y corresponds to the weight.

Observation	x	y
1	4.5	9.5
2	3.7	8.2
3	1.8	4.9
4	1.3	6.7
5	3.2	12.9
6	3.8	14.1
7	2.5	5.6
8	4.5	8.0
9	4.1	12.6
10	1.1	7.2

Table 14.1: Data

Let us display this data in a scatter plot. Towards that end, let us read the length data into an object by the name "`x`" and the weight data into an object by the name "`y`". Finally, let us apply the function "`plot`" to the formula that relates the response "`y`" to the explanatory variable "`x`":

```
> x <- c(4.5,3.7,1.8,1.3,3.2,3.8,2.5,4.5,4.1,1.1)
> y <- c(9.5,8.2,4.9,6.7,12.9,14.1,5.6,8.0,12.6,7.2)
> plot(y~x)
```

The scatter plot that is produced by the last expression is presented in Figure 14.1.

A scatter plot is a graph that displays jointly the data of two numerical variables. The variables ("`x`" and "`y`" in this case) are represented by the x-axis and the y-axis, respectively. The x-axis is associated with the explanatory variable and the y-axis is associated with the response.

Each observation is represented by a point. The x-value of the point corresponds to the value of the explanatory variable for the observation and the y-value corresponds to the value of the response. For example, the first observation is represented by the point $(x = 4.5, y = 9.5)$. The two rightmost points have an x value of 4.5. The higher of the two has a y value of 9.5 and is therefore

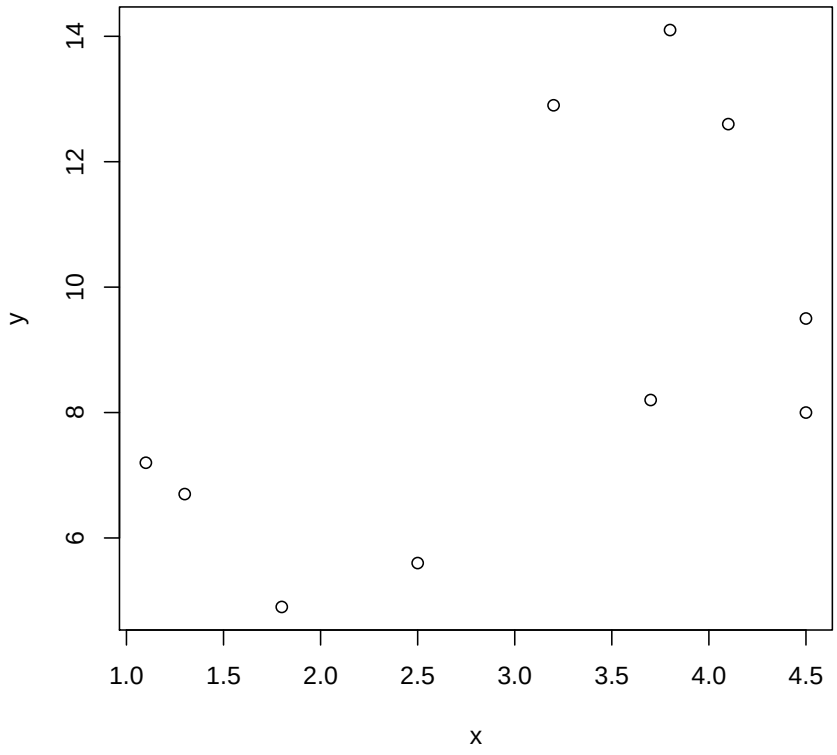

Figure 14.1: A Scatter Plot

point associated with the first observation. The lower of the two has a y value of 8.0, and is thus associated with the 8th observation. Altogether there are 10 points in the plot, corresponding to the 10 observations in the data frame.

Let us consider another example of a scatter plot. The file "`cars.csv`" contains data regarding characteristics of cars. Among the variables in this data frame are the variables "`horsepower`" and the variable "`engine.size`". Both variables are numeric.

The variable "`engine.size`" describes the volume, in cubic inches, that is swept by all the pistons inside the cylinders. The variable "`horsepower`" measures the power of the engine in units of horsepower. Let us examine the relation between these two variables with a scatter plot:

```
> cars <- read.csv("cars.csv")
> plot(horsepower ~ engine.size, data=cars)
```

In the first line of code we read the data from the file into an `R` data frame that is given the name "`cars`". In the second line we produce the scatter plot with "`horsepower`" as the response and "`engine.size`" as the explanatory variable.

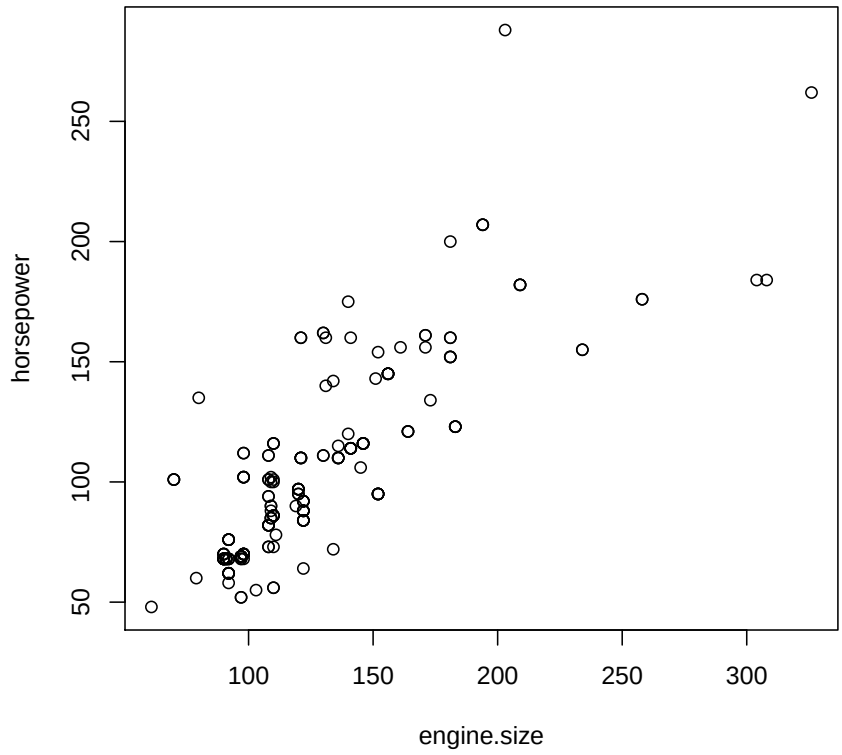

Figure 14.2: The Scatter Plot of Power versus Engine Size

Both variables are taken from the data frame "`cars`". The plot that is produced by the last expression is presented in Figure 14.2.

Consider the expression "`plot(horsepower~engine.size, data=cars)`". Both the response variable and the explanatory variables that are given in this expression do not exist in the computer's memory as independent objects, but only as variables within the object "`cars`". In some cases, however, one may refer to these variables directly within the function, provided that the argument "`data=`*data.frame.name*" is added to the function. This argument informs the function in which data frame the variables can be found, where *data.frame.name* is the name of the data frame. In the current example, the variables are located in the data frame "`cars`".

Examine the scatter plot in Figure 14.2. One may see that the values of the response (`horsepower`) tend to increase with the increase in the values of the explanatory variable (`engine.size`). Overall, the increase tends to follow a linear trend, a straight line, although the data points are not located exactly on a single line. The role of linear regression, which will be discussed in the subsequent sections, is to describe and assess this linear trend.

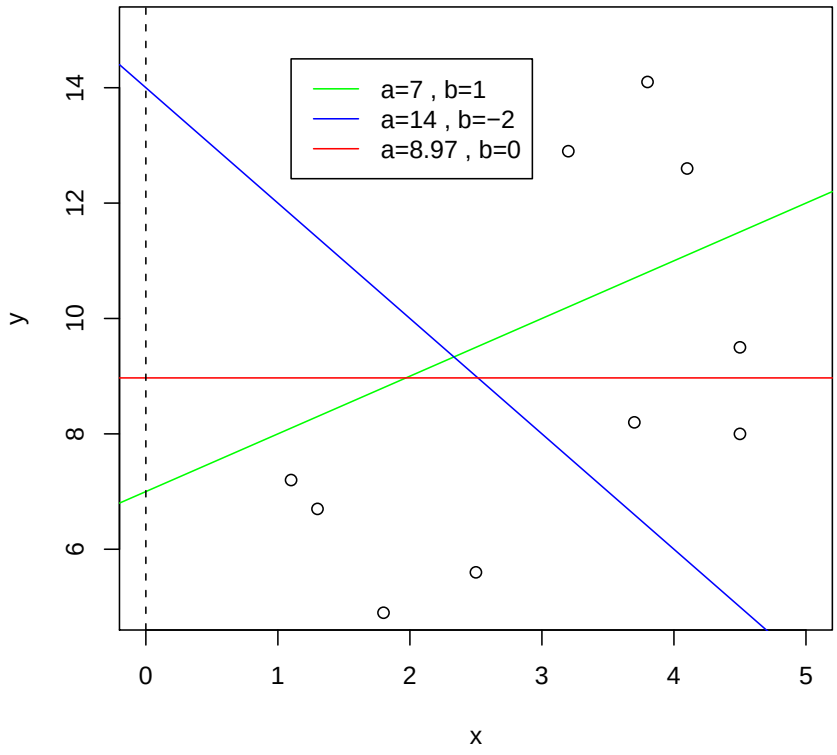

Figure 14.3: Lines

14.2.2 Linear Equation

Linear regression describes linear trends in the relation between a response and an explanatory variable. Linear trends may be specified with the aid of linear equations. In this subsection we discuss the relation between a linear equation and a linear trend (a straight line).

A linear equation is an equation of the form:

$$y = a + b \cdot x \ ,$$

where y and x are variables and a and b are the coefficients of the equation. The coefficient a is called the *intercept* and the coefficient b is called the *slope*.

A linear equation can be used in order to plot a line on a graph. With each value on the x-axis one may associate a value on the y-axis: the value that satisfies the linear equation. The collection of all such pairs of points, all possible x values and their associated y values, produces a straight line in the two-dimensional plane.

As an illustration consider the three lines in Figure 14.3. The *green* line is produced via the equation $y = 7+x$, the intercept of the line is 7 and the slope is

1. The *blue* is a result of the equation $y = 14 - 2x$. For this line the intercept is 14 and the slope is -2. Finally, the *red* line is produced by the equation $y = 8.97$. The intercept of the line is 8.97 and the slope is equal to 0.

The intercept describes the value of y when the line crosses the y-axis. Equivalently, it is the result of the application of the linear equation for the value $x = 0$. Observe in Figure 14.3 that the *green* line crosses the y-axis at the level $y = 7$. Likewise, the *blue* line crosses the y-axis at the level $y = 14$. The *red* line stays constantly at the level $y = 8.97$, and this is also the level at which it crosses the y-axis.

The slope is the change in the value of y for each unit change in the value of x. Consider the *green* line. When $x = 0$ the value of y is $y = 7$. When x changes to $x = 1$ then the value of y changes to $y = 8$. A change of one unit in x corresponds to an *increase* in one unit in y. Indeed, the slope for this line is $b = 1$. As for the *blue* line, when x changes from 0 to 1 the value of y changes from $y = 14$ to $y = 12$; a *decrease* of two units. This decrease is associated with the slope $b = -2$. Lastly, for the constant *red* line there is no change in the value of y when x changes its value from $x = 0$ to $x = 1$. Therefore, the slope is $b = 0$. A positive slope is associated with an increasing line, a negative slope is associated with a decreasing line and a zero slope is associated with a constant line.

Lines can be considered in the context of scatter plots. Figure 14.3 contains the scatter plot of the data on the relation between the length of fish and their weight. A regression line is the line that best describes the linear trend of the relation between the explanatory variable and the response. Neither of the lines in the figure is the regression line, although the *green* line is a better description of the trend than the *blue* line. The regression line is the best description of the linear trend.

The *red* line is a fixed line that is constructed at a level equal to the average value[1] of the variable y. This line partly reflects the information in the data. The regression line, which we fit in the next section, reflects more of the information by including a description of the trend in the data.

Lastly, let us see how one can add lines to a plot in `R`. Functions to produce plots in `R` can be divided into two categories: high level and low level plotting functions. High level functions produce an entire plot, including the axes and the labels of the plot. The plotting functions that we encountered in the past such as "`plot`", "`hist`", "`boxplot`" and the like are all high level plotting functions. Low level functions, on the other hand, add features to an existing plot.

An example of a low level plotting function is the function "`abline`". This function adds a straight line to an existing plot. The first argument to the function is the intercept of the line and the second argument is the slope of the line. Other arguments may be used in order to specify the characteristics of the line. For example, the argument "`col=`*color.name*" may be used in order to change the color of the line from its default black color. A plot that is very similar to plot in Figure 14.3 may be produced with the following code[2]:

```
> plot(y~x)
```

[1] Run the expression "`mean(y)`" to obtain $\bar{y} = 8.97$ as the value of the sample average.

[2] The actual plot in Figure 14.3 is produced by a slightly modified code. First an empty plot is produced with the expression "`plot(c(0,5),c(5,15),type="n",xlab="x",ylab="y")`" and then the points are added with the expression "`points(y~x)`". The lines are added as in the text. Finally, a legend is added with the function "`legend`".

```
> abline(7,1,col="green")
> abline(14,-2,col="blue")
> abline(mean(y),0,col="red")
```

Initially, the scatter plot is created and the lines are added to the plot one after the other. Observe that color of the first line that is added is green, it has an intercept of 7 and a slope of 1. The second line is blue, with a intercept of 14 and a negative slope of -2. The last line is red, and its constant value is the average of the variable y.

In the next section we discuss the computation of the regression line, the line that describes the linear trend in the data. This line will be added to scatter plots with the aid of the function "`abline`".

14.3 Linear Regression

Data that describes the joint distribution of two numeric variables can be represented with a scatter plot. The y-axis in this plot corresponds to the response and the x-axis corresponds to the explanatory variable. The regression line describes the linear trend of the response as a function of the explanatory variable. This line is characterized by a linear equation with an intercept and a slope that are computed from the data.

In the first subsection we present the computation of the regression linear equation from the data. The second subsection discusses regression as a statistical model. Statistical inference can be carried out on the basis of this model. In the context of the statistical model, one may consider the intercept and the slope of the regression model that is fitted to the data as point estimates of the model's parameter. Based on these estimates, one may test hypotheses regarding the regression model and construct confidence intervals for parameters.

14.3.1 Fitting the Regression Line

The `R` function that fits the regression line to data is called "`lm`", an acronym for *Linear Model*. The input to the function is a formula, with the response variable to the left of the tilde character and the explanatory variable to the right of it. The output of the function is the fitted linear regression model.

Let us apply the linear regression function to the data on the weight and the length of fish. The output of the function is saved by us in a object called "`fit`". Subsequently, the content of the object "`fit`" is displayed:

```
> fit <- lm(y~x)
> fit

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
      4.616        1.427
```

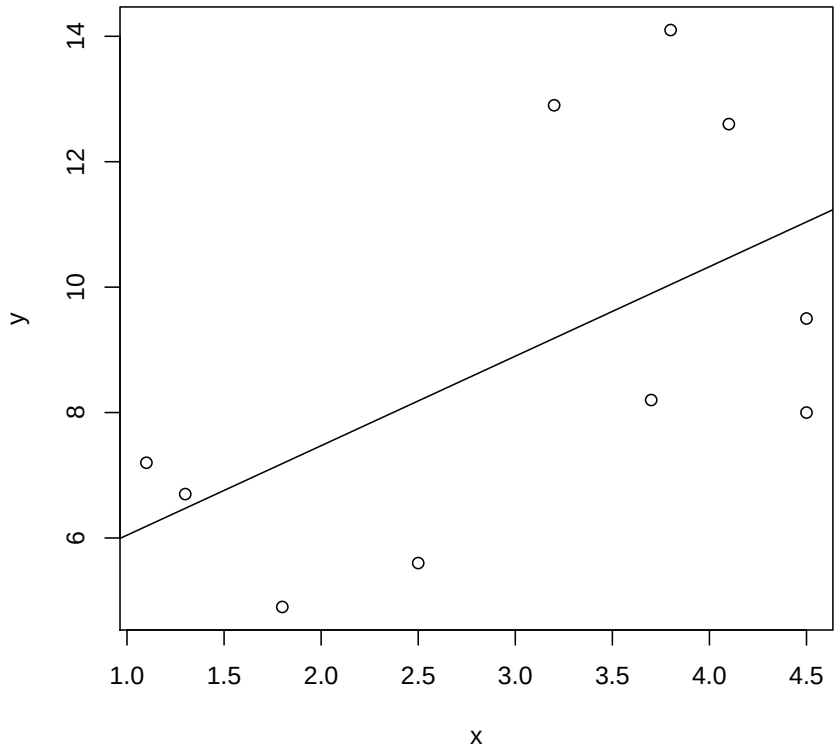

Figure 14.4: A Fitted Regression Line

When displayed, the output of the function "`lm`" shows the formula that was used by the function and provides the coefficients of the regression linear equation. Observe that the intercept of the line is equal to 4.616. The slope of the line, the coefficient that multiplies "`x`" in linear equation, is equal to 1.427.

One may add the regression line to the scatter plot with the aid of the function "`abline`":

```
> plot(y~x)
> abline(fit)
```

The first expression produces the scatter plot of the data on fish. The second expression adds the regression line to the scatter plot. When the input to the graphical function "`abline`" is the output of the function "`lm`" that fits the regression line, then the result is the addition of the regression line to the existing plot. The line that is added is the line characterized by the coefficients that are computed by the function "`lm`". The coefficients in the current setting are 4.616 for the intercept and 1.427 for the slope.

The scatter plot and the added regression line are displayed in Figure 14.4. Observe that line passes through the points, balancing between the points that

are above the line and the points that are below. The line captures the linear trend in the data.

Examine the line in Figure 14.4. When $x = 1$ then the y value of the line is slightly above 6. When the value of x is equal to 2, a change of one unit, then value of y is below 8, and is approximately equal to 7.5. This observation is consistent with the fact that the slop of the line is 1.427. The value of x is decreased by 1 when changing from $x = 1$ to $x = 0$. Consequently, the value of y when $x = 0$ should decrease by 1.427 in comparison to its value when $x = 1$. The value at $x = 1$ is approximately 6. Therefore, the value at $x = 0$ should be approximately 4.6. Indeed, we do get that the intercept is equal to 4.616.

The coefficients of the regression line are computed from the data and are hence statistics. Specifically, the slope of the regression line is computed as the ratio between the *covariance* of the response and the explanatory variable, divided by the variance of the explanatory variable. The intercept of the regression line is computed using the sample averages of both variables and the computed slope.

Start with the slope. The main ingredient in the formula for the slope, the numerator in the ratio, is the covariance between the two variables. The covariance measures the joint variability of two variables. Recall that the formula for the sample variance of the variable x is equal to::

$$s^2 = \frac{\text{Sum of the squares of the deviations}}{\text{Number of values in the sample} - 1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \ .$$

The formula of the sample covariance between x and y replaces the square of the deviations by the product of deviations. The product is between an y deviation and the parallel x deviation:

$$\text{covariance} = \frac{\text{Sum of products of the deviations}}{\text{Number of values in the sample} - 1} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{n-1} \ .$$

The function "`cov`" computes the sample covariance between two numeric variables. The two variables enter as arguments to the function and the sample covariance is the output. Let us demonstrate the computation by first applying the given function to the data on fish and then repeating the computations without the aid of the function:

```
> cov(y,x)
[1] 2.386111
> sum((y-mean(y))*(x-mean(x)))/9
[1] 2.386111
```

In both cases we obtained the same result. Notice that the sum of products of deviations in the second expression was divided by 9, which is the number of observations, minus 1.

The slope of the regression line is the ratio between the covariance and the variance of the explanatory variable.

The regression line passes through the point $(\bar{x}, \bar{y})$, a point that is determined by the means of the both the explanatory variable and the response. It follows that the intercept should obey the equation:

$$\bar{y} = a + b \cdot \bar{x} \quad \Longrightarrow \quad a = \bar{y} - b \cdot \bar{x} \ ,$$

The left-hand-side equation corresponds to the statement that the value of the regression line at the average $\bar{x}$ is equal to the average of the response $\bar{y}$. The right-hand-side equation is the solution to the left-hand-side equation.

One may compute the coefficients of the regression model manually by computing first the slope as a ratio between the covariance and the variance of explanatory variable. The intercept can then be obtained by the equation that uses the computed slope and the averages of both variables:

```
> b <- cov(x,y)/var(x)
> a <- mean(y) - b*mean(x)
> a
[1] 4.616477
> b
[1] 1.427385
```

Applying the manual method we obtain, after rounding up, the same coefficients that were produced by the application of the function "`lm`" to the data.

As an exercise, let us fit the regression model to the data on the relation between the response "`horsepower`" and the explanatory variable "`engine.size`". Apply the function "`lm`" to the data and present the results:

```
> fit.power <- lm(horsepower ~ engine.size, data=cars)
> fit.power

Call:
lm(formula = horsepower ~ engine.size, data = cars)

Coefficients:
(Intercept)  engine.size
     6.6414       0.7695
```

The fitted regression model is stored in an object called "`fit.power`". The intercept in the current setting is equal to 6.6414 and the slope is equal to 0.7695.

Observe that one may refer to variables that belong to a data frame, provided that the name of the data frame is entered as the value of the argument "`data`" in the function "`lm`". Here we refer to variables that belong to the data frame "`cars`".

Next we plot the scatter plot of the data and add the regression line:

```
> plot(horsepower ~ engine.size, data=cars)
> abline(fit.power)
```

The output of the plotting functions is presented in Figure 14.5. Again, the regression line describes the general linear trend in the data. Overall, with the increase in engine size one observes increase in the power of the engine.

14.3.2 Inference

Up to this point we have been considering the regression model in the context of descriptive statistics. The aim in fitting the regression line to the data was to characterize the linear trend observed in the data. Our next goal is to deal with

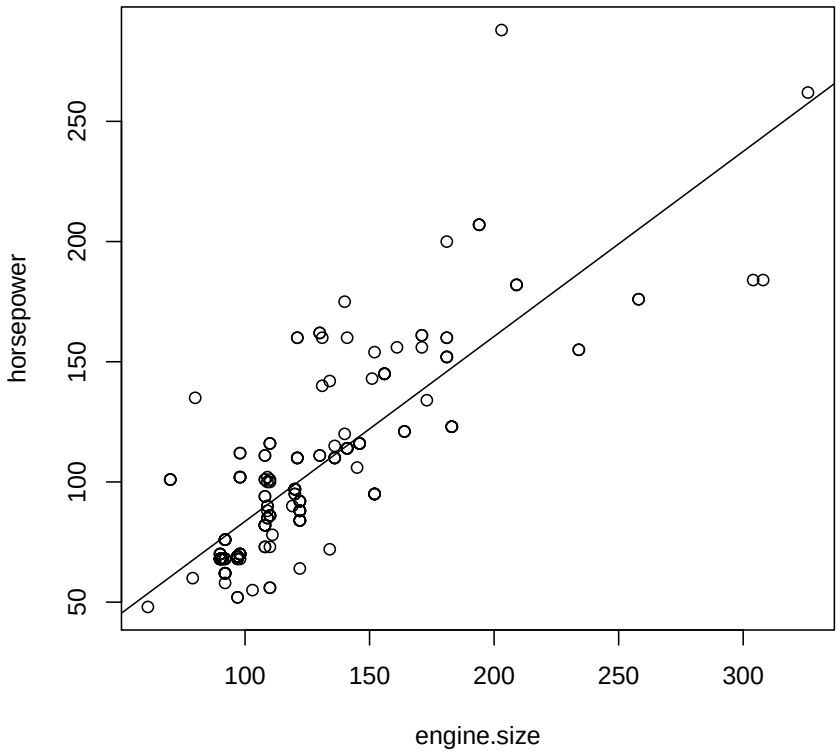

Figure 14.5: A Regression Model of Power versus Engine Size

regression in the context of inferential statistics. The goal here is to produce statements on characteristics of an entire population on the basis of the data contained in the sample.

The foundation for statistical inference in a given setting is a statistical model that produces the sampling distribution in that setting. The sampling distribution is the frame of reference for the analysis. In this context, the observed sample is a single realization of the sampling distribution, one realization among infinitely many potential realizations that never take place. The setting of regression involves a response and an explanatory variable. We provide a description of the statistical model for this setting.

The relation between the response and the explanatory variable is such that the value of the later affects the distribution of the former. Still, the value of the response is not uniquely defined by the value of the explanatory variable. This principle also hold for the regression model of the relation between the response Y and the explanatory variable X. According to the model of linear regression the value of the *expectation* of the response for observation i, $\mathrm{E}(Y_i)$, is a linear function of the value of the explanatory variable for the same observation. Hence, there exist and intercept a and a slope b, common for all observations,

such that if $X_i = x_i$ then

$$\mathrm{E}(Y_i) = a + b \cdot x_i \, .$$

The regression line can thus be interpreted as the average trend of the response in the population. This average trend is a linear function of the explanatory variable.

The intercept a and the slope b of the statistical model are parameters of the sampling distribution. One may test hypotheses and construct confidence intervals for these parameters based on the observed data and in relation to the sampling distribution.

Consider testing hypothesis. A natural null hypothesis to consider is the hypothesis that the slope is equal to zero. This hypothesis corresponds to statement that the expected value of the response is constant for all values of the explanatory variable. In other words, the hypothesis is that the explanatory variable does not affect the distribution of the response[3]. One may formulate this null hypothesis as $H_0 : b = 0$ and test it against the alternative $H_1 : b \neq 0$ that states that the explanatory variable does affect the distribution of the response.

A test of the given hypotheses can be carried out by the application of the function "`summary`" to the output of the function "`lm`". Recall that the function "`lm`" was used in order to fit the linear regression to the data. In particular, this function was applied to the data on the relation between the size of the engine and the power that the engine produces. The function fitted a regression line that describes the linear trend of the data. The output of the function was saved in an object by the name "`fit.power`". We apply the function "`summary`" to this object:

```
> summary(fit.power)

Call:
lm(formula = horsepower ~ engine.size, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-59.643 -12.282  -5.515  10.251 125.153

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.64138    5.23318   1.269    0.206
engine.size  0.76949    0.03919  19.637   <2e-16 ***
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 23.31 on 201 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.6574,     Adjusted R-squared: 0.6556
F-statistic: 385.6 on 1 and 201 DF,  p-value: < 2.2e-16
```

[3] According to the model of linear regression, the only effect of the explanatory variable on the distribution of the response is via the expectation. If such an effect, according to the null hypothesis, is also excluded then the so called explanatory variable is not effecting at all the distribution of the response.

The output produced by the application of the function "`summary`" is long and detailed. We will discuss this output in the next section. Here we concentrate on the table that goes under the title "`Coefficients:`". The said table is made of 2 rows and 4 columns. It contains information for testing, for each of the coefficients, the null hypothesis that the value of the given coefficient is equal to zero. In particular, the second row may be used is order to test this hypothesis for the slope of the regression line, the coefficient that multiplies the explanatory variable.

Consider the second row. The first value on this row is 0.76949, which is equal (after rounding up) to the slope of the line that was fitted to the data in the previous subsection. However, in the context of statistical inference this value is the *estimate* of the slope of the population regression coefficient, the realization of the estimator of the slope[4].

The second value is 0.03919. This is an estimate of the standard deviation of the estimator of the slope. The third value is the test statistic. This statistic is the ratio between the deviation of the sample estimate of the parameter (0.76949) from the value of the parameter under the null hypothesis (0), divided by the estimated standard deviation (0.03919): $(0.76949 - 0)/0.03919 = 0.76949/0.03919 = 19.63486$, which is essentially the value given in the report[5].

The last value is the computed p-value for the test. It can be shown that the sampling distribution of the given test statistic, under the null distribution which assumes no slope, is asymptotically the standard Normal distribution. If the distribution of the response itself is Normal then the distribution of the statistic is the t-distribution on $n-2$ degrees of freedom. In the current situation this corresponds to 201 degrees of freedom[6]. The computed p-value is extremely small, practically eliminating the possibility that the slope is equal to zero.

The first row presents information regarding the intercept. The estimated intercept is 6.64138 with an estimated standard deviation of 5.23318. The value of the test statistic is 1.269 and the p-value for testing the null hypothesis that the intercept is equal to zero against the two sided alternative is 0.206. In this case the null hypothesis is not rejected since the p-value is larger than 0.05.

The report contains an inference for the intercept. However, one is advised to take this inference in the current case with a grain of salt. Indeed, the intercept is the expected value of the response when the explanatory variable is equal to zero. Here the explanatory variable is the size of the engine and the response is the power of that engine. The power of an engine of size zero is a quantity that has no physical meaning! In general, unless the intercept is in the range of observations (i.e. the value 0 is in the range of the observed explanatory variable) one should treat the inference on the intercept cautiously. Such inference requires extrapolation and is sensitive to the miss-specification of the regression model.

Apart from testing hypotheses one may also construct confidence intervals for the parameters. A crude confidence interval may be obtained by taking

[4]The estimator of the slope is obtained via the application of the formula for the computation of the slope to the sample: $\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})/\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

[5]Our computation involves rounding up errors, hence the small discrepancy between the value we computed and the value in the report.

[6]Notice that the "`horsepower`" measurement is missing for two observation. These observations are deleted for the analysis, leaving a total of $n = 203$ observations. The number of degrees of freedom is $n - 2 = 203 - 2 = 201$.

1.96 standard deviations on each side of the estimate of the parameter. Hence, a confidence interval for the slope is approximately equal to $0.76949 \pm 1.96 \times 0.03919 = [0.6926776, 0.8463024]$. In a similar way one may obtain a confidence interval for the slope[7]: $6.64138 \pm 1.96 \times 5.23318 = [-3.615653, 16.89841]$.

Alternatively, one may compute confidence intervals for the parameters of the linear regression model using the function "`confint`". The input to this function is the fitted model and the output is a confidence interval for each of the parameters:

```
> confint(fit.power)
                 2.5 %     97.5 %
(Intercept) -3.6775989 16.9603564
engine.size  0.6922181  0.8467537
```

Observe the similarity between the confidence intervals that are computed by the function and the crude confidence intervals that were produced by us. The small discrepancies that do exist between the intervals result from the fact that the function "`confint`" uses the t-distribution whereas we used the Normal approximation.

14.4 R-squared and the Variance of Residuals

In this section we discuss the residuals between the values of the response and their estimated expected value according to the regression model. These residuals are the regression model equivalence of the deviations between the observations and the sample average. We use these residuals in order compute the variability that is not accounted for by the regression model. Indeed, the ratio between the total variability of the residuals and the total variability of the deviations from the average serves as a measure of the variability that is not explained by the explanatory variable. R-squared, which is equal to 1 minus this ratio, is interpreted as the fraction of the variability of the response that is explained by the regression model.

We start with the definition of residuals. Let us return to the artificial example that compared length of fish to their weight. The data for this example was given in Table 14.2.1 and was saved in the objects "`x`" and "`y`". The regression model was fitted to this data by the application of the function "`lm`" to the formula "`y~x`" and the fitted model was saved in an object called "`fit`". Let us apply the function "`summary`" to the fitted model:

```
> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
```

[7]The warning message that was made in the context of testing hypotheses on the intercept should be applied also to the construction of confidence intervals. If the value 0 is not in the range of the explanatory variable then one should be careful when interpreting a confidence interval for the intercept.

```
-3.0397 -2.1388 -0.6559  1.8518  4.0595

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.6165     2.3653   1.952   0.0868 .
x             1.4274     0.7195   1.984   0.0826 .
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 2.791 on 8 degrees of freedom
Multiple R-squared: 0.3297,     Adjusted R-squared: 0.246
F-statistic: 3.936 on 1 and 8 DF,  p-value: 0.08255
```

The given report contains a table with estimates of the regression coefficients and information for conducting hypothesis testing. The report contains other information that is associated mainly with the notion of the residuals from regression line. Our current goal is to understand what is that other information.

The residual from regression for each observation is the difference between the value of the response for the observation and the estimated expectation of the response under the regression model[8]. An observation is a pair (x_i, y_i), with y_i being the value of the response. The expectation of the response according to the regression model is $a + b \cdot x_i$, where a and b are the coefficients of the model. The estimated expectation is obtained by using, in the formula for the expectation, the coefficients that are estimated from the data. The residual is the difference between y_i and $a + b \cdot x_i$.

Consider an example. The first observation on the fish is $(4.5, 9.5)$, where $x_1 = 4.5$ and $y_1 = 9.5$. The estimated intercept is 4.6165 and the estimated slope is 1.4274. The estimated expectation of the response for the first variable is equal to

$$4.6165 + 1.4274 \cdot x_1 = 4.6165 + 1.4274 \cdot 4.5 = 11.0398 \, .$$

The residual is the difference between the observes response and this value:

$$y_1 - (4.6165 + 1.4274 \cdot x_1) = 9.5 - 11.0398 = -1.5398 \, .$$

The residuals for the other observations are computed in the same manner. The values of the intercept and the slope are kept the same but the values of the explanatory variable and the response are changed.

Consult the upper plot in Figure 14.6. This is a scatter plot of the data, together with the regression line in *black* and the line of the average in *red*. A vertical arrow extends from each data point to the regression line. The point where each arrow hits the regression line is associated with the estimated value of the expectation for that point. The residual is the difference between the value of the response at the origin of the arrow and the value of the response at the tip of its head. Notice that there are as many residuals as there are observations.

The function "`residuals`" computes the residuals. The input to the function is the fitted regression model and the output is the sequence of residuals. When we apply the function to the object "`fit`", which contains the fitted regression model for the fish data, we get the residuals:

[8] The estimated expectation of the response is also called *the predicted response*.

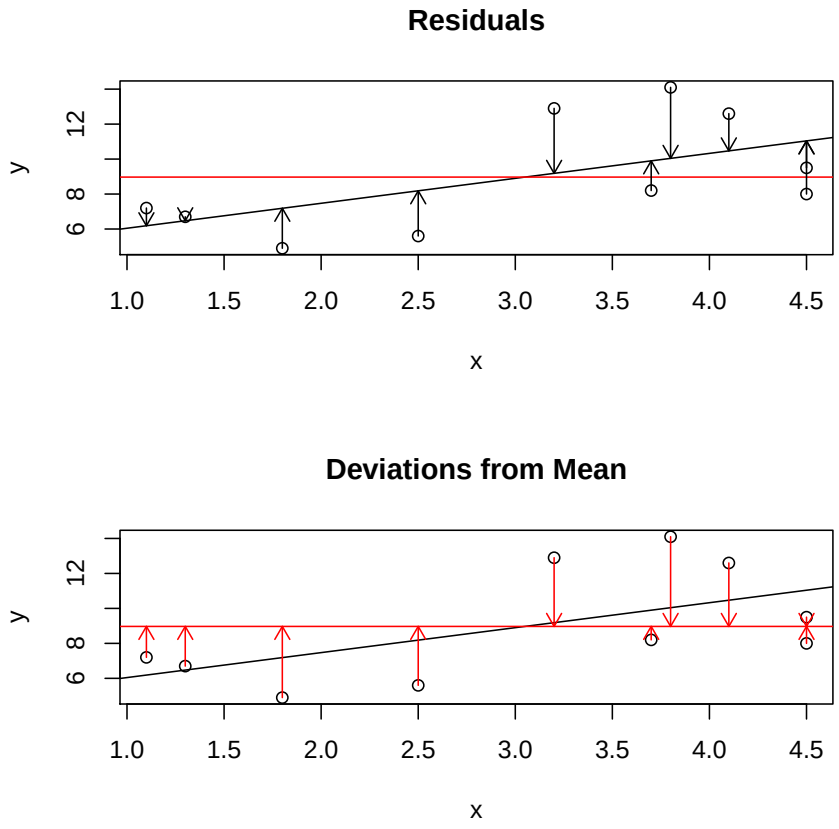

Figure 14.6: Residuals and Deviations from the Mean

```
> residuals(fit)
         1          2          3          4          5
-1.5397075 -1.6977999 -2.2857694  0.2279229  3.7158923
         6          7          8          9         10
 4.0594616 -2.5849385-3.0397075  2.1312463  1.0133998
```

Indeed, 10 residuals are produced, one for each observation. In particular, the residual for the first observation is -1.5397075, which is essentially the value that we obtained[9].

Return to the report produced by the application of the function "`summary`" to the fitted regression model. The first component in the report is the formula that identifies the response and the explanatory variable. The second component, the component that comes under the title "`Residuals:`", gives a summary of the distribution of the residuals. This summary includes the smallest and the largest values in the sequence of residuals, as well as the first and third quartiles

[9]The discrepancy between the value that we computed and the value computed by the function results from rounding up errors. We used the vales of the coefficients that appear in the report. These values are rounded up. The function "`residuals`" uses the coefficients without rounding.

and the median. The average is not reported since the average of the residuals from the regression line is always equal to 0.

The table that contains information on the coefficients was discussed in the previous section. Let us consider the last 3 lines of the report.

The first of the three lines contains the estimated value of the standard deviation of the response from the regression model. If the expectations of the measurements of the response are located on the regression line then the variability of the response corresponds to the variability about this line. The resulting variance is estimated by the sum of squares of the residuals from the regression line, divided by the number of observations minus 2. A division by the number of observation minus 2 produces an unbiased estimator of the variance of the response about the regression model. Taking the square root of the estimated variance produces an estimate of the standard deviation:

```
> sqrt(sum(residuals(fit)^2)/8)
[1] 2.790787
```

The last computation is a manual computation of the estimated standard deviation. It involves squaring the residuals and summing the squares. This sum is divided by the number of observations minus 2 ($10 - 2 = 8$). Taking the square root produces estimate. The value that we get for the estimated standard deviation is 2.790787, which coincides with the value that appears in the first of the last 3 lines of the report.

The second of these lines reports the R-squared of the linear fit. In order to explain the meaning of R-squared let us consider Figure 14.6 once again. The two plots in the figure present the scatter plot of the data together with the regression line and the line of the average. Vertical *black* arrows that represent the residuals from the regression are added to the upper plot. The lower plot contains vertical *red* arrows that extend from the data points to the line of the average. These arrows represent the deviations of the response from the average.

Consider two forms of variation. One form is the variation of the response from its average value. This variation is summarized by the sample variance, the sum of the squared lengths of the *red* arrows divided by the number of observations minus 1. The other form of variation is the variation of the response from the fitted regression line. This variation is summarized by the sample variation of the residuals, the sum of squared lengths of the *black* arrows divided by the number of observations minus 1. The ratio between these two quantities gives the relative variability of the response that remains after fitting the regression line to the data.

The line of the average is a straight line. The deviations of the observations from this straight line can be thought of as residuals from that line. The variability of these residuals, the sum of squares of the deviations from the average divided by the number of observations minus 1, is equal to the sample variance.

The regression line is the unique straight line that minimizes the variability of its residuals. Consequently, the variability of the residuals from the regression, the sum of squares of the residuals from the regression divided by the number of observations minus 1, is the smallest residual variability produced by any straight line. It follows that the sample variance of the regression residuals is less than the sample variance of the response. Therefore, the ratio between the variance of the residuals and the variance of the response is less than 1.

R-squared is the difference between 1 and the ratio of the variances. Its value is between 0 and 1 and it represents the fraction of the variability of the response that is *explained* by the regression line. The closer the points are to the regression line the larger the value of R-squared becomes. On the other hand, the less there is a linear trend in the data the closer to 0 is the value of R-squared. In the extreme case of R-squared equal to 1 all the data point are positioned exactly on a single straight line. In the other extreme, a value of 0 for R-squared implies no linear trend in the data.

Let us compte manually the difference between 1 and the ratio between the variance of the residuals and the variance of the response:

```
> 1-var(residuals(fit))/var(y)
[1] 0.3297413
```

Observe that the computed value of R-squared is the same as the value "`Multiple R-squared:  0.3297`" that is given in the report.

The report provides another value of R-squared, titled *Adjusted R-squared*. The difference between the adjusted and unadjusted quantities is that in the former the sample variance of the residuals from the regression is replaced by an unbiased estimate of the variability of the response about the regression line. The sum of squares in the unbiased estimator is divided by the number of observations minus 2. Indeed, when we re-compute the ratio using the unbiased estimate, the sum of squared residuals divided by $10 - 2 = 8$, we get:

```
> 1-(sum(residuals(fit)^2)/8)/var(y)
[1] 0.245959
```

The value of this adjusted quantity is equal to the value "`Adjusted R-squared: 0.246`" in the report.

Which value of R-squared to use is a matter of personal taste. In any case, for a larger number of observations the difference between the two values becomes negligible.

The last line in the report produces an overall goodness of fit test for the regression model. In the current application of linear regression this test reduces to a test of the slope being equal to zero, the same test that is reported in the second row of the table of coefficients[10]. The F statistic is simply the square of the t value that is given in the second row of the table. The sampling distribution of this statistic under the null hypothesis is the F-distribution on 1 and $n - 2$ degrees of freedom, which is the sampling distribution of the square of the test statistic for the slope. The computed p-value, "`p-value:  0.08255`" is the identical (after rounding up) to the p-value given in the second line of the table.

Return to the R-squared coefficient. This coefficient is a convenient measure of the goodness of fit of the regression model to the data. Let us demonstrate this point with the aid of the "`cars`" data. In Subsection 14.3.2 we fitted a regression model to the power of the engine as a response and the size of the engine as an explanatory variable. The fitted model was saved in the object called "`fit.power`". A report of this fit, the output of the expression "`summary(fit.power)`" was also presented. The null hypothesis of zero slope

[10] In more complex applications of linear regression, applications that are not considered in this book, the test in the last line of the report and the tests of coefficients do not coincide.

was clearly rejected. The value of R-squared for this fit was 0.6574. Consequently, about 2/3 of the variability in the power of the engine is explained by the size of the engine.

Consider trying to fit a different regression model for the power of the engine as a response. The variable "`length`" describes the length of the car (in inches). How well would the length explain the power of the car? We may examine this question using linear regression:

```
> summary(lm(horsepower ~ length, data=cars))

Call:
lm(formula = horsepower ~ length, data = cars)

Residuals:
   Min     1Q Median     3Q    Max
-53.57 -20.35  -6.69  14.45 180.72

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -205.3971    32.8185  -6.259 2.30e-09 ***
length         1.7796     0.1881   9.459  < 2e-16 ***
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 33.12 on 201 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.308,      Adjusted R-squared: 0.3046
F-statistic: 89.47 on 1 and 201 DF,  p-value: < 2.2e-16
```

We used one expression to fit the regression model to the data and to summarize the outcome of the fit.

A scatter plot of the two variables together with the regression line is presented in Figure 14.7. This plot may be produced using the code:

```
> plot(horsepower ~ length, data=cars)
> abline(lm(horsepower ~ length, data=cars))
```

From the examination of the figure we may see that indeed there is a linear trend in the relation between the length and the power of the car. Longer cars tend to have more power. Testing the null hypothesis that the slope is equal to zero produces a very small p-value and leads to the rejection of the null hypothesis.

The length of the car and the size of the engine are both statistically significant in their relation to the response. However, which of the two explanatory variables produces a better fit?

An answer to this question may be provided by the examination of values of R-squared, the ratio of the variance of the response explained by each of the explanatory variable. The R-squared for the size of the engine as an explanatory variable is 0.6574, which is approximately equal to 2/3. The value of R-squared for the length of the car as an explanatory variable is 0.308, less than 1/3. It follows that the size of the engine explains twice as much of the variability of the power of the engine than the size of car and is a better explanatory variable.

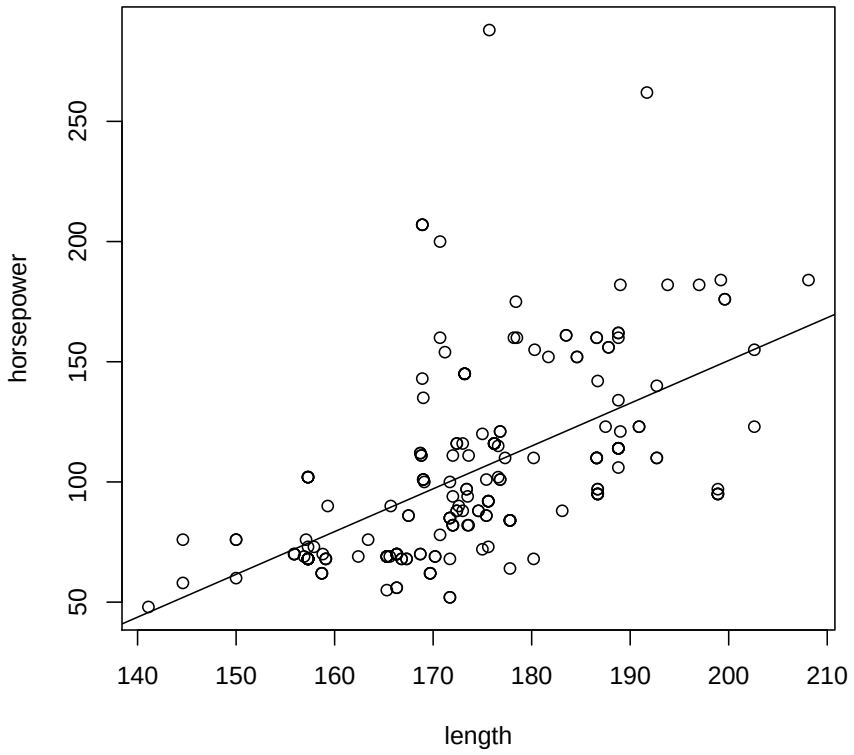

Figure 14.7: A Regression Model of Power versus Length

14.5 Solved Exercises

Question 14.1. Figure 14.8 presents 10 points and three lines. One of the lines is colored *red* and one of the points is marked as a *red triangle*. The points in the plot refer to the data frame in Table 14.1 and the three lines refer to the linear equations:

1. $y = 4$
2. $y = 5 - 2x$
3. $y = x$

You are asked to match the marked line to the appropriate linear equation and match the marked point to the appropriate observation:

1. Which of the three equations, 1, 2 or 3, describes the line marked in *red*?
2. The poind marked with a *red triangle* represents which of the observations. (Identify the observation number.)

Observation	x	y
1	2.3	-3.0
2	-1.9	9.8
3	1.6	4.3
4	-1.6	8.2
5	0.8	5.9
6	-1.0	4.3
7	-0.2	2.0
8	2.4	-4.7
9	1.8	1.8
10	1.4	-1.1

Table 14.2: Points

Solution (to Question 14.1.1): The line marked in *red* is increasing and at $x = 0$ it seems to obtain the value $y = 0$. An increasing line is associated with a linear equation with a positive slope coefficient ($b > 0$). The only equation with that property is Equation 3, for which $b = 1$. Notice that the intercept of this equation is $a = 0$, which agrees with the fact that the line passes through the origin $(x, y) = (0, 0)$. If the x-axis and the y-axis were on the same scale then one would expect the line to be tilted in the 45 degrees angle. However, here the axes are not on the same scale, so the tilting is different.

Solution (to Question 14.1.2): The x-value of the line marked with a *red triangle* is $x = -1$. The y-value is below 5. The observation that has an x-value of -1 is Observation 6. The y-value of this observation is $y = 4.3$. Notice that there is another observation with the same y-value, Observation 3. However, the x-value of that observation is $x = 1.6$. Hence it is the point that is on the same level as the marked point, but it is placed to the right of it.

Question 14.2. Assume a regression model that describes the relation between the expectation of the response and the value of the explanatory variable in the form:

$$\mathrm{E}(Y_i) = 2.13 \cdot x_i - 3.60 \, .$$

1. What is the value of the intercept and what is the value of the slope in the linear equation that describes the model?

2. Assume the $x_1 = 5.5$, $x_2 = 12.13$, $x_3 = 4.2$, and $x_4 = 6.7$. What is the expected value of the response of the 3rd observation?

Solution (to Question 14.2.1): The intercept is equal to $a = -3.60$ and the slope is equal to $b = 2.13$. Notice that the lope is the coefficient that multiplies the explanatory variable and the intercept is the coefficient that does not multiply the explanatory variable.

Solution (to Question 14.2.2): The value of the explanatory variable for the 3rd observation is $x_3 = 4.2$. When we use this value in the regression formula we obtain that:

$$\mathrm{E}(Y_3) = 2.13 \cdot x_3 - 3.60 = 2.13 \cdot 4.2 - 3.60 = 5.346 \, .$$

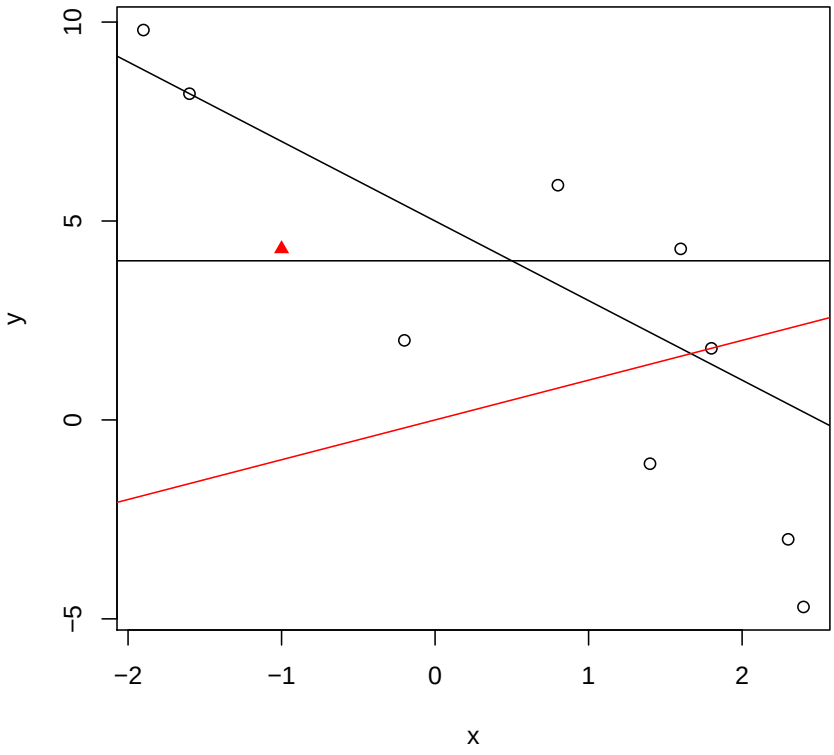

Figure 14.8: Lines and Points

In words, the expectation of the response of the 3rd observation is equal to 5.346

Question 14.3. The file "`aids.csv`" contains data on the number of diagnosed cases of Aids and the number of deaths associated with Aids among adults and adolescents in the United States between 1981 and 2002[11]. The file can be found on the internet at `http://pluto.huji.ac.il/~msby/StatThink/Datasets/aids.csv`.

The file contains 3 variables: The variable "`year`" that tells the relevant year, the variable "`diagnosed`" that reports the number of Aids cases that were diagnosed in each year, and the variable "`deaths`" that reports the number of Aids related deaths in each year. The following questions refer to the data in the file:

1. Consider the variable "`deaths`" as response and the variable "`diagnosed`"

[11]The data is taken from Table 1 in section "Practice in Linear Regression" of the online Textbook "Collaborative Statistics" (Connexions. March 22, 2010. http://cnx.org/content/col10522/1.38/) by Barbara Illowsky and Susan Dean.

as an explanatory variable. What is the slope of the regression line? Produce a point estimate and a confidence interval. Is it statistically significant (namely, significantly different than 0)?

2. Plot the scatter plot that is produced by these two variables and add the regression line to the plot. Does the regression line provided a good description of the trend in the data?

3. Consider the variable "`diagnosed`" as the response and the variable "`year`" as the explanatory variable. What is the slope of the regression line? Produce a point estimate and a confidence interval. Is the slope in this case statistically significant?

4. Plot the scatter plot that is produced by the later pair of variables and add the regression line to the plot. Does the regression line provided a good description of the trend in the data?

Solution (to Question 14.3.1): After saving the file "`aids.csv`" in the working directory of `R` we read it's content into a data frame by the name "`aids`". We then produce a summary of the fit of the linear regression model of "`deaths`" as a response and "`diagnosed`" as the explanatory variable:

```
> aids <- read.csv("aids.csv")
> fit.deaths <- lm(deaths~diagnosed,data=aids)
> summary(fit.deaths)

Call:
lm(formula = deaths ~ diagnosed, data = aids)

Residuals:
     Min       1Q   Median       3Q      Max
-7988.73  -680.86    23.94  1731.32  7760.67

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   88.7161  1370.7191   0.065    0.949
diagnosed      0.6073     0.0312  19.468 1.81e-14 ***
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 3589 on 20 degrees of freedom
Multiple R-squared: 0.9499,     Adjusted R-squared: 0.9474
F-statistic:   379 on 1 and 20 DF,  p-value: 1.805e-14
```

The estimated value of the slope 0.6073. The computed p-value associated with this slope is 1.81×10^{-14}, which is much smaller than the 5% threshold. Consequently, the slope is statistically significant.

For confidence intervals apply the function "`confint`" to the fitted model:

```
> confint(fit.deaths)
                    2.5 %      97.5 %
(Intercept) -2770.5538947 2947.986092
diagnosed       0.5422759    0.672427
```

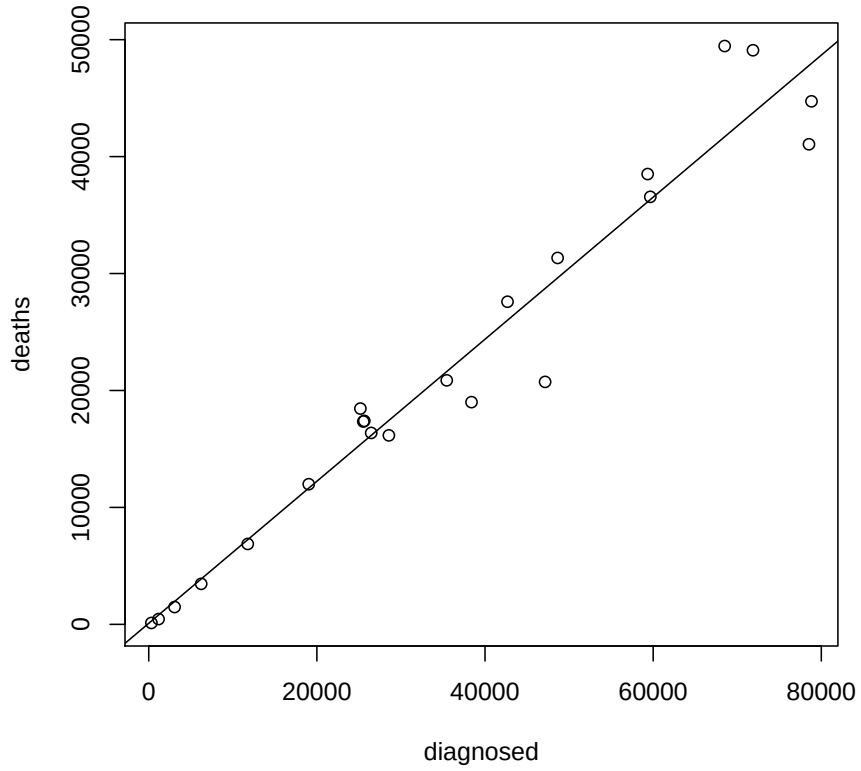

Figure 14.9: Aids Related Deaths versus Diagnosed Cases of Aids

We get that the confidence interval for the slope is $[0.5422759, 0.672427]$.

Solution (to Question 14.3.2): A scatter plot of the two variables is produced by the application of the function "`plot`" to the formula that involves these two variables. The regression line is added to the plot using the function "`abline`" with the fitted model as an input:

```
> plot(deaths~diagnosed,data=aids)
> abline(fit.deaths)
```

The plot that is produced is given in Figure 14.9. Observe that the points are nicely placed very to a line that characterizes the linear trend of the regression.

Solution (to Question 14.3.3): We fit a linear regression model of "`diagnosed`" as a response and "`year`" as the explanatory variable and save the fit in the object "`fit.diagnosed`". A summary of the model is produced by the application of the function "`summary`" to the fit:

```
> fit.diagnosed <- lm(diagnosed~year,data=aids)
```

```
> summary(fit.diagnosed)

Call:
lm(formula = diagnosed ~ year, data = aids)

Residuals:
   Min     1Q Median     3Q    Max
-28364 -18321  -3909  14964  41199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3448225.0  1535037.3  -2.246   0.0361 *
year            1749.8      770.8   2.270   0.0344 *
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 22940 on 20 degrees of freedom
Multiple R-squared: 0.2049,     Adjusted R-squared: 0.1651
F-statistic: 5.153 on 1 and 20 DF,  p-value: 0.03441
```

The estimated value of the slope 1749.8. The computed p-value associated with this slope is 0.0344, which is less than the 0.05. Consequently, one may declare the slope to be statistically significant. Confidence intervals are produced using the function "`confint`":

```
> confint(fit.diagnosed)
                    2.5 %      97.5 %
(Intercept) -6650256.6654 -246193.429
year              141.9360    3357.618
```

We get that the 95% confidence interval for the slope is $[141.9360, 3357.618]$.

Solution (to Question 14.3.4): A scatter plot of the two variables is produced by the application of the function "`plot`" and the regression line is added with function "`abline`":

```
> plot(diagnosed~year,data=aids)
> abline(fit.diagnosed)
```

The plot is given in Figure 14.10. Observe that the points do not follow a linear trend. It seems that the number of diagnosed cases increased in an exponential rate during the first years after Aids was discovered. The trend changed in the mid 90's with a big drop in the number of diagnosed Aids patients. This drop may be associated with the administration of therapies such as AZT to HIV infected subjects that reduced the number of such subjects that developed Aids. In the late 90's there seams to be yet again a change in the trend and the possible increase in numbers. The line of linear regression misses all these changes and is a poor representation of the historical development.

The take home message from this exercise is to not use models blindly. A good advise is to plot the data. An examination of the plot provides a warning that the linear model is probably not a good model for the given problem.

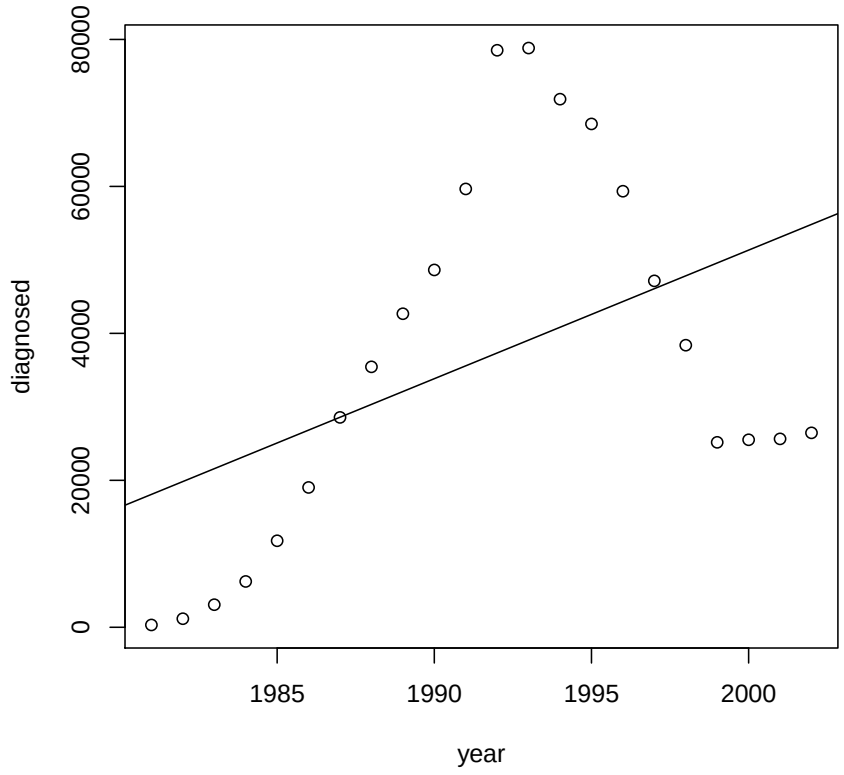

Figure 14.10: Diagnosed Cases of Aids versus Year of Report

Question 14.4. Below are the percents of the U.S. labor force (excluding self-employed and unemployed) that are members of a labor union[12]. We use this data in order to practice the computation of the regression coefficients.

1. Produce the scatter plot of the data and add the regression line. Is the regression model reasonable for this data?

2. Compute the sample averages and the sample standard deviations of both variables. Compute the covariance between the two variables.

3. Using the summaries you have just computed, recompute the coefficients of the regression model.

Solution (to Question 14.4.1): We read the data in the table into `R`. The variable "`year`" is the explanatory variable and the variable "`percent`" is the

[12]Taken from Homework section in the chapter on linear regression of the online Textbook "Collaborative Statistics" (Connexions. March 22, 2010. http://cnx.org/content/col10522/1.38/) by Barbara Illowsky and Susan Dean.

year	percent
1945	35.5
1950	31.5
1960	31.4
1970	27.3
1980	21.9
1986	17.5
1993	15.8

Table 14.3: Percent of Union Members

response. The scatter plot is produced using the function "`plot`" and the regression line, fitted to the data with the function "`lm`", is added to the plot using the function "`abline`":

```
> year <- c(1945,1950,1960,1970,1980,1986,1993)
> percent <- c(35.5,31.5,31.4,27.3,21.9,17.5,15.8)
> plot(percent~year)
> abline(lm(percent~year))
```

The scatter plot and regression line are presented in Figure 14.11. Observe that a linear trend is a reasonable description of the reduction in the percentage of workers that belong to labor unions in the post World War II period.

Solution (to Question 14.4.2): We compute the averages, standard deviations and the covariance:

```
> mean.x <- mean(year)
> mean.y <- mean(percent)
> sd.x <- sd(year)
> sd.y <- sd(percent)
> cov.x.y <- cov(year,percent)
> mean.x
[1] 1969.143
> mean.y
[1] 25.84286
> sd.x
[1] 18.27957
> sd.y
[1] 7.574927
> cov.x.y
[1] -135.6738
```

The average of the variable "`year`" is 1969.143 and the standard deviation is 18.27957. The average of the variable "`percent`" is 25.84286 and the standard deviation is 7.574927. The covariance between the two variables is -135.6738.

Solution (to Question 14.4.3): The slope of the regression line is the ratio between the covariance and the variance of the explanatory variable. The intercept is the solution of the equation that states that the value of regression line at the average of the explanatory variable is the average of the response:

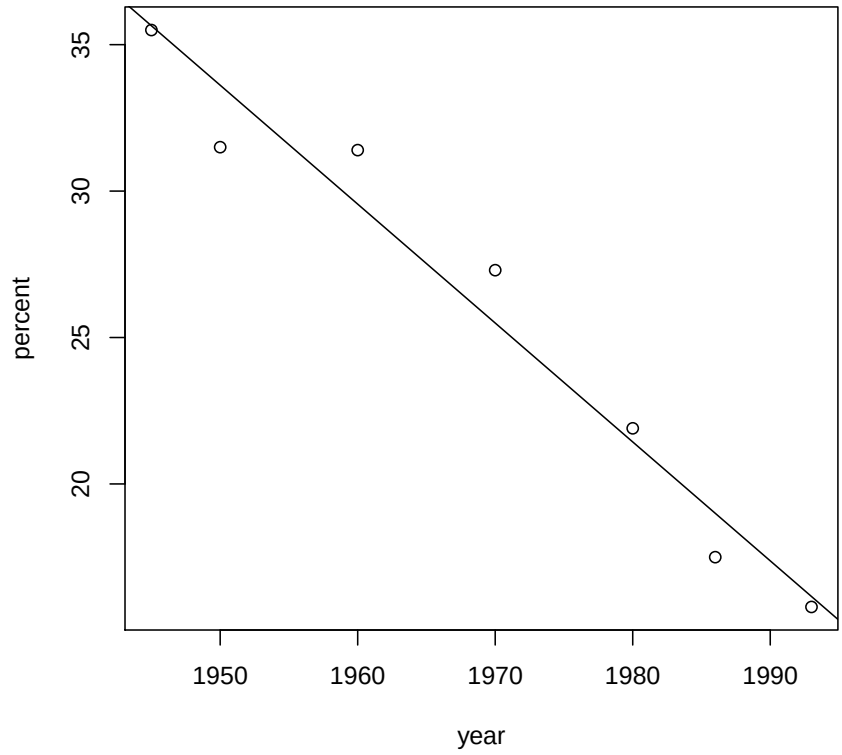

Figure 14.11: Percent of Union Workers

```
> b <- cov.x.y/sd.x^2
> a <- mean.y - b*mean.x
> a
[1] 825.3845
> b
[1] -0.4060353
```

We get that the intercept is equal to 825.3845 and the slope is equal to -0.4060353.

In order to validate these figures, let us apply the function "`lm`" to the data:

```
> lm(percent~year)

Call:
lm(formula = percent ~ year)

Coefficients:
(Intercept)         year
    825.384       -0.406
```

Indeed, we get the same numbers that we got from the manual computation.

Question 14.5. Assume a regression model was fit to some data that describes the relation between the explanatory variable x and the response y. Assume that the coefficients of the fitted model are $a = 2.5$ and $b = -1.13$, for the intercept and the slope, respectively. The first 4 observations in the data are $(x_1, y_1) = (5.5, 3.22)$, $(x_2, y_2) = (12.13, -6.02)$, $(x_3, y_3) = (4.2, -8.3)$, and $(x_4, y_4) = (6.7, 0.17)$.

1. What is the estimated expectation of the response for the 4th observation?
2. What is the residual from the regression line for the 4th observation?

Solution (to Question 14.5.1): The estimate for the expected value for the ith observation is obtained by the evaluation of the expression $a + b \cdot x_i$, where a and b are the coefficients of the fitted regression model and x_i is the value of the explanatory variable for the ith observation. In our case $i = 4$ and $x_4 = 6.7$:

$$a + b \cdot x_4 = 2.5 - 1.13 \cdot x_4 = 2.5 - 1.13 \cdot 6.7 = -5.071 \ .$$

Therefore, the estimate expectation of the response is -5.071.

Solution (to Question 14.5.2): The residual from the regression line is the difference between the observed value of the response and the estimated expectation of the response. For the 4th observation we have that the observed value of the response is $y_4 = 0.17$. The estimated expectation was computed in the previous question. Therefore, the residual from the regression line for the 4th observation is:

$$y_4 - (a + b \cdot x_4) = 0.17 - (-5.071) = 5.241 \ .$$

Question 14.6. In Chapter 13 we analyzed an example that involved the difference between fuel consumption in highway and city driving conditions as the response[13]. The explanatory variable was a factor that was produced by splitting the cars into two weight groups. In this exercise we would like to revisit this example. Here we use the weight of the car directly as an explanatory variable. We also consider the size of the engine as an alternative explanatory variable and compare between the two regression models.

1. Fit the regression model that uses the variable "`curb.weight`" as an explanatory variable. Is the slope significantly different than 0? What fraction of the standard deviation of the response is explained by a regression model involving this variable?
2. Fit the regression model that uses the variable "`engine.size`" as an explanatory variable. Is the slope significantly different than 0? What fraction of the standard deviation of the response is explained by a regression model involving this variable?
3. Which of the two models fits the data better?

[13]The response was computed using the expression "`cars$highway.mpg - cars$city.mpg`"

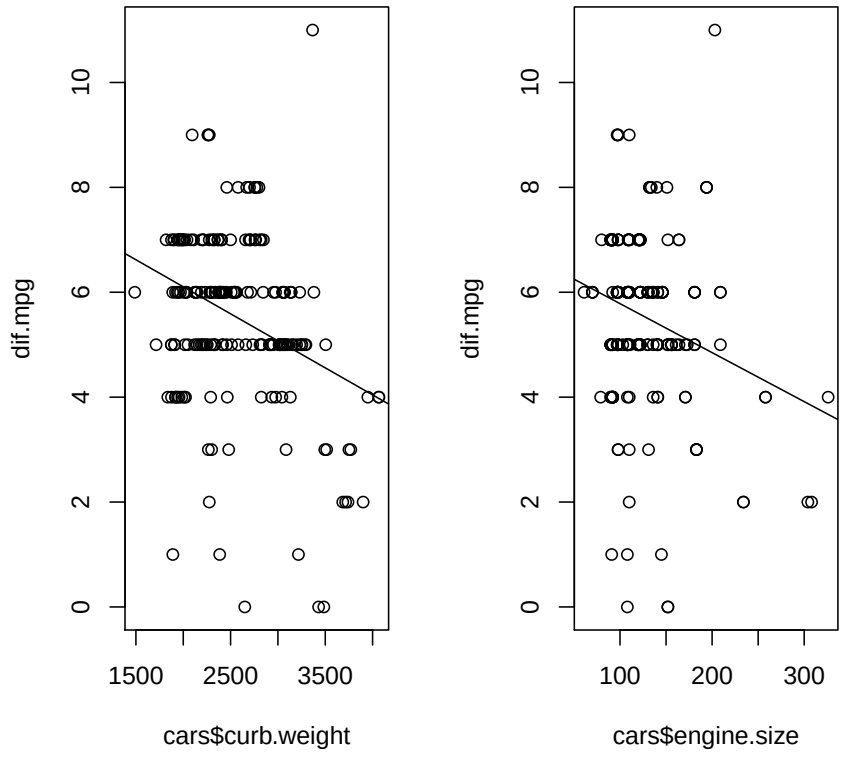

Figure 14.12: Percent of Union Workers

Solution (to Question 14.6.1): We create the response and then fit a model and apply the summarizing function to the model:

```
> dif.mpg <- cars$highway.mpg - cars$city.mpg
> summary(lm(dif.mpg ~ cars$curb.weight))

Call:
lm(formula = dif.mpg ~ cars$curb.weight)

Residuals:
    Min      1Q  Median      3Q     Max
-5.4344 -0.7755  0.1633  0.8844  6.3035

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       8.1653491  0.5460856  14.953  < 2e-16 ***
cars$curb.weight -0.0010306  0.0002094  -4.921 1.77e-06 ***
---
```

```
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 1.557 on 203 degrees of freedom
Multiple R-squared: 0.1066,     Adjusted R-squared: 0.1022
F-statistic: 24.22 on 1 and 203 DF,  p-value: 1.775e-06
```

The p-value associated with the slope, 1.77×10^{-6}, is much smaller than the 5% threshold proposing a significant (negative) trend. The value of R-squared, the fraction of the variability of the response that is explained by a regression model, is 0.1066.

The standard deviation is the square root of the variance. It follows that the fraction of the standard deviation of the response that is explained by the regression is $\sqrt{0.1066} = 0.3265$.

Following our own advice, we plot the data and the regression model:

```
> plot(dif.mpg ~ cars$curb.weight)
> abline(lm(dif.mpg ~ cars$curb.weight))
```

The resulting plot is presented on the left-hand side of Figure 14.12. One may observe that although there seems to be an overall downward trend, there is still allot of variability about the line of regression.

Solution (to Question 14.6.2): We now fit and summarize the regression model with the size of engine as the explanatory variable:

```
> summary(lm(dif.mpg ~ cars$engine.size))

Call:
lm(formula = dif.mpg ~ cars$engine.size)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7083 -0.7083  0.1889  1.1235  6.1792

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.717304   0.359385  18.691  < 2e-16 ***
cars$engine.size -0.009342   0.002691  -3.471 0.000633 ***
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 1.601 on 203 degrees of freedom
Multiple R-squared: 0.05603,    Adjusted R-squared: 0.05138
F-statistic: 12.05 on 1 and 203 DF,  p-value: 0.0006329
```

The regression slope is negative. The p-value is 0.000633, which is statistically significant. The value of R-squared is 0.05603. Consequently, the fraction of the standard deviation of the response that is explained by the current regression model is $\sqrt{0.05603} = 0.2367$.

Produce the scatter plot with the line of regression:

```
> plot(dif.mpg ~ cars$engine.size)
> abline(lm(dif.mpg ~ cars$engine.size))
```

The plot is given on the right-hand side of Figure 14.12. Again, there is variability about the line of regression.

Solution (to Question 14.6.3): Of the two models, the model that uses the curb weigh as the explanatory variable explains a larger portion of the variability in the response. Hence, unless other criteria tells us otherwise, we will prefer this model over the model that uses the size of engine as an explanatory variable.

14.6 Summary

Glossary

Regression: Relates different variables that are measured on the same sample. Regression models are used to describe the effect of one of the variables on the distribution of the other one. The former is called the explanatory variable and the later is called the response.

Linear Regression: The effect of a numeric explanatory variable on the distribution of a numeric response is described in terms of a linear trend.

Scatter Plot: A plot that presents the data in a pair of numeric variables. The axes represents the variables and each point represents an observation.

Intercept: A coefficient of a linear equation. Equals the value of y when the line crosses the y-axis.

Slope: A coefficient of a linear equation. The change in the value of y for each unit change in the value of x. A positive slope corresponds to an increasing line and a negative slope corresponds to a decreasing line.

Covariance: A measures the joint variability of two numeric variables. It is equal to the sum of the product of the deviations from the mean, divided by the number of observations minus 1.

Residuals from Regression: The residual differences between the values of the response for the observation and the estimated expectations of the response under the regression model (the predicted response).

R-Square: is the difference between 1 and the ratio between the variance of the residuals from the regression and the variance of the response. Its value is between 0 and 1 and it represents the fraction of the variability of the response that is *explained* by the regression line.

Discuss in the Forum

The topic for discussion in the Forum of Chapter 6 was mathematical models and how good they should fit reality. In this Forum we would like to return to the same topic subject, but consider it specifically in the context of statistical models.

Some statisticians prefer complex models, models that try to fit the data as closely as one can. Others prefer a simple model. They claim that although

simpler models are more remote from the data yet they are easier to interpret and thus provide more insight. What do you think? Which type of model is better to use?

When formulating your answer to this question you may thing of a situation that involves inference based on data conducted by yourself for the sack of others. What would be the best way to report your findings and explain them to the others?

Formulas:

- A Linear Equation: $y = a + b \cdot x$.
- Covariance: $\frac{\text{Sum of products of the deviations}}{\text{Number of values in the sample}-1} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{n-1}$.
- Regression Slope: $b = \text{Covariance}(x, y)/\text{Var}(x)$.
- Regression Intercept: $a = \bar{y} - b\bar{x}$.
- The Regression Model: $\text{E}(Y_i) = a + b \cdot x_i$, a and b population parameters.
- Residuals: $y_i - (a + bx_i)$, a and b estimated from the data.
- Estimate of Residual Variance: $\sum_{i=1}^{n}(y_i - (a + bx_i))^2/(n - 2)$, a and b estimated from the data.
- R-Squared: $1 - \sum_{i=1}^{n}(y_i - (a + bx_i))^2 / \sum_{i=1}^{n}(y_i - \bar{y})^2$, a and b estimated from the data.