Simple linear regression model. The linear regression has two parameters, β_0 and β_1, which determine the model

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \ldots, n,$$

where $e_1, \ldots, e_n$ are independent and $N(0, \sigma^2)$-distributed. Or, equivalently, the model assumes that $y_1, \ldots, y_n$ are independent. The parameters of the model are β_0, β_1, and σ. The slope parameter β_1 is the expected increment in y as x increases by one unit, whereas β_0 is the expected value of y when $x = 0$. The remainder terms $e_1, \ldots, e_n$ represent the vertical deviations from the straight line. The assumption of variance homogeneity means that the typical size of these deviations is the same across all values of x.

Residual standard error. In a linear model the residual

$$r_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \quad i = 1, \ldots, n,$$

measures the distance from the observed value y_i to the predicted value $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. The residual standard error $\hat{\sigma}$ is estimated by the average distance from the observations to the predicted. Thus, we can calculate it by

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}}$$

We can use it to describe the effectiveness of our prediction. If the residual standard deviation is small then the observations are generally closer to the predicted line, and they are further away if the residual standard deviation is large.

Residual analysis. The residuals are standardized with their common standard error $\hat{\sigma}$. Thus, the standardized residuals

$$\frac{y_1 - \hat{y}_1}{\hat{\sigma}}, \ldots, \frac{y_n - \hat{y}_n}{\hat{\sigma}} \tag{7.1}$$

are standardized such that they resemble the normal distribution with mean zero and standard deviation one if the model assumptions hold. Hence, we check these properties on the standardized residuals. The normality assumption is usually checked with a QQ-plot: Due to the standardization, the points should be scattered around the line with intercept zero and slope one. The assumption of homogeneous standard deviation is usually validated with a **residual plot**, where the standardized residuals are plotted against the predicted values.

Model validation based on residuals. A residual plot shows the standardized residuals (7.1) against the predicted values $\hat{y}_i$'s. The points should be spread randomly in the vertical direction, without any systematic patterns. In particular,

- points should be roughly equally distributed between positive and negative values in all parts of the plot (from left to right);
- there should be roughly the same variation in the vertical direction in all parts of the plot (from left to right);
- there should be no too extreme points.

Systematic deviations described above correspond to problems with the mean structure, the variance homogeneity, or the normal distribution, respectively.

Example: Stearic acid and digestibility. Researchers examined the digestibility of fat with different levels of stearic acid. The average digestibility percent was measured for nine different levels of stearic acid proportion. Data are shown in the table below, where x represents stearic acid and y is digestibility measured in percent.

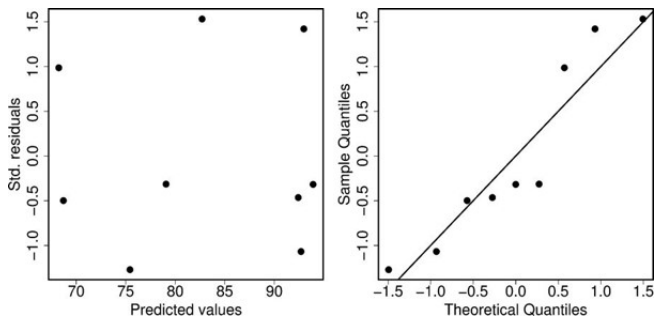
x	29.8	30.3	22.6	18.7	14.8	4.1	4.4	2.8	3.8
y	67.5	70.6	72.0	78.2	87.0	89.9	91.2	93.1	96.7

Then predicted value $\hat{y}_i$ corresponding to the original observation of x_i is obtained by

$$\hat{y}_i = 96.5334 - 0.9337x_i$$

where the estimated intercept and slope are calculated by the least squares method.

The residual analysis for the digestibility data is presented below. The left panel shows the residual plot, and QQ-plot (right) of the standardized residuals is also obtained. The straight line in the QQ-plot has intercept zero and slope one. In the residual plot we see that the points are spread out without any clearly visible pattern.



- There seem to be both positive and negative residuals in all parts of the plot (from left to right; for small, medium, as well as large predicted values). This indicates that the specification of the digestibility mean as a linear function of the stearic acid level is appropriate.
- There seems to be roughly the same vertical variation for small, medium, and large predicted values. This indicates that the standard deviation is the same for all observations (homoscedasticity, or homogeneity of variance).
- There are neither very small nor very large standardized residuals This indicates that there are no outliers and that it is not unreasonable to use the normal distribution.

Hypothesis test. In the linear regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \ldots, n,$$

the slope parameter β_1 measures the linear relationship between x_i and y_i, and the assumption of no association between x_i and y_i is expressed by the null hypothesis

$$H_0 : \beta_1 = 0$$

Under the null hypothesis (that is, if H_0 is true), all y_i's are independently determined without x_i, which implies a simplification of the original linear regression model. Usually the alternative hypothesis is simply the complement of the null hypothesis

$$H_A : \beta_1 \neq 0$$

in the linear regression model.

Example: Stearic acid and digestibility. The hypothesis

$$H_0 : \beta_1 = 0$$

that there is no relationship between the level of stearic acid and digestibility is tested by the test statistic

$$T_{obs} = \frac{\hat{\beta}_1}{\mathrm{SE}(\hat{\beta}_1)} = \frac{-0.9337}{0.0926} = -10.08$$

where the estimate $\hat{\beta}_1 = -0.9337$ and its standard error

$$\mathrm{SE}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\mathrm{SS}_x}} = 0.0926$$

are obtained from

$$\mathrm{SS}_x = \sum_{i=1}^{n} (x_i - \bar{x})^2 = 1028.55$$

The value $T_{obs} = -10.08$ of test statistic should be compared to a random variable T from the t_7-distribution. For the alternative hypothesis

$$H_A : \beta_1 \neq 0$$

we get the p-value

$$P(|T| \geq 10.08) = 0.00002$$

and we conclude that there is strong evidence of an association between digestibility and the stearic acid level, and that the slope is significantly different from zero. Since the estimate -0.9337 of slope is negative, we conclude that the digestibility percentage decreases as the percentage of stearic acid increases.

R coding: Stearic acid and digestibility. The variable `model` contains everything we need to know about the linear regression. Here we have two parameters: the intercept and the slope. The estimated intercept is found under (Intercept) and the variable `stearic.acid`.

```
summary(model)
```

The most important part of the output from summary() is the coefficients part (roughly in the middle). Four values are listed for each parameter: (i) The estimate; (ii) the standard error

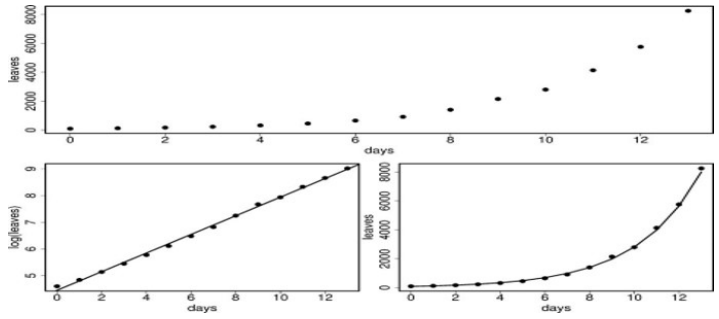
for the estimate; (iii) a t-test statistic for the hypothesis that the corresponding parameter is equal to zero; and (iv) the corresponding p-value. After the coefficients part we find the residual standard error, the residual degrees of freedom, and a few things that are not of interest at the moment.

Transformation. We may find the usefulness of the linear regression model in situations where there appears to be a non-linear relationship between two variables x_i and y_i. In those situations a direct application of linear regression model is inappropriate. In some cases, however, we may be able to remedy the situation by transforming the response variable y_i in such a way that the transformed data shows a linear relationship with the explanatory variable x_i. Let (x_i, y_i), $i = 1, \ldots, n$, denote our n pairs of observations and assume that a straight line does not reasonably describe the relationship between x_i and y_i. By transformation we seek a function $f(y)$ such that the transformed variables, $z_i = f(y_i)$, can be modeled as a linear function of the x_i. That is,

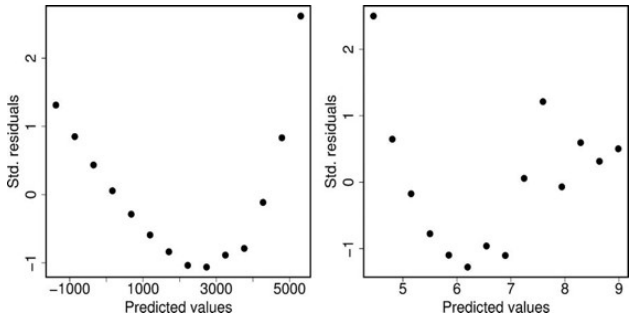
$$z_i = \beta_0 + \beta_1 x_i$$

This is the case in the following example.

Example: Growth of duckweed. Top panel shows the original duckweed data. Bottom left shows the data and fitted regression line after logarithmic transformation and bottom right shows the fitted line transformed back to the original scale.



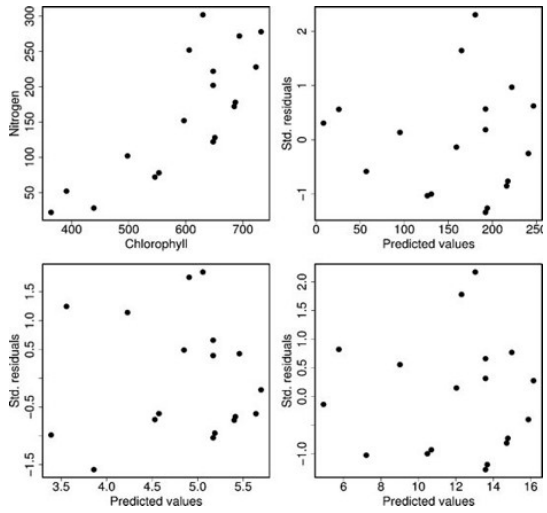
The population size y_i of a species can often be described by an exponential growth model. This corresponds to a linear regression model with $\log(y_i)$ as response variable.



Residual plots for the duckweed data. Left panel: linear regression with the leaf counts as response. Right panel: linear regression with the logarithmic leaf counts as response.

Example: Chlorophyll concentration. An experiment with winter wheat was carried out in order to investigate if the concentration of nitrogen in the soil can be predicted from the

concentration of chlorophyll in the plants. This could improve the adjustment of nitrogen supply. The chlorophyll concentration in the leaves as well as the nitrogen concentration in the soil were measured for 18 plants. The nitrogen concentration N is plotted against the chlorophyll concentration C (upper left panel below). The other three panels show residual plots for three different models, the only difference being the choice of response: Residual plots for the regression of nitrogen concentration N predicted by chlorophyll content C in the plants (upper right), for the regression of $\log(N)$ on C (lower left), and for the regression of the square root $\sqrt{N}$ on C (lower right).



For the regression with N as response (the upper right plot) there is an indication of a **trumpet shape**: the variation seems to be larger for large predicted values compared to small predicted values. This is quite often the case for biological data, and it can often be remedied by transformation. The logarithmic transformation $\log(N)$ has the property that it squeezes large values and in that way diminishes the variation for large values. For this particular dataset, however, it seems like the log-transformation has been too powerful (the lower left plot): there seems to be larger variation for small predicted values. The square root transformation $\sqrt{N}$ is sometimes a useful compromise between no transformation and the logarithmic transformation, and except for two large positive residuals, the variation seems to be constant across the different values of predicted values (the lower right plot).

Standard errors in linear regression. Consider the linear regression model. As already derived, the least squares estimates for slope and intercept are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}; \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Define SS_x and s respectively by

$$\mathrm{SS}_x = \sum_{i=1}^n (x_i - \bar{x})^2; \quad s = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}}$$

The standard errors for the estimates are

$$\mathrm{SE}(\hat{\beta}_1) = \frac{s}{\sqrt{\mathrm{SS}_x}}; \quad \mathrm{SE}(\hat{\beta}_0) = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\mathrm{SS}_x}}$$

Confidence interval in linear regression. In general we denote the confidence level $(1-\alpha)$, such that 95% and 90% confidence intervals corresponds to $\alpha = 0.05$ and $\alpha = 0.10$, respectively. The relevant critical value corresponds to $\alpha/2$, assigning probability $\alpha/2$ to the right. Then $(1-\alpha)$-confidence interval for parameter β_j is of the form

$$\hat{\beta}_j \pm t_{\alpha/2,n-2}\mathrm{SE}(\hat{\beta}_j), \quad j = 0, 1,$$

where $n-2$ is the degree of freedom. The $(1-\alpha)$-confidence interval includes the values of β_j for which it is reasonable, at confidence degree $(1-\alpha)$, to believe that they could have generated the data. If we repeated the experiment many times then a fraction $1-\alpha$ of the corresponding confidence intervals would include the true value β_j.

Example: Stearic acid and digestibility. Consider the linear regression model which describes the association between the level of stearic acid and digestibility.

$$\hat{\beta}_1 = -0.9337; \quad \hat{\beta}_0 = 96.5334$$

$\mathrm{SS}_x = 1028.549$ and $s = 2.970$ are used to calculate

$$\mathrm{SE}(\hat{\beta}_1) = \frac{2.970}{\sqrt{1028.549}} = 0.0926;$$
$$\mathrm{SE}(\hat{\beta}_0) = (2.970)\sqrt{\frac{1}{9} + \frac{(14.5889)^2}{1028.549}} = 1.6752$$

There are $n = 9$ observations and two parameters. Thus, we need critical values from the t-distribution with $(n-2) = 7$ degrees of freedom. Since $t_{0.025,7} = 2.365$, we compute the 95% confidence interval by

$$-0.9337 \pm (2.365)(0.0926) = (-1.15, -0.71);$$
$$96.5334 \pm (2.365)(1.6752) = (92.57, 100.49),$$

for the slope and the intercept parameter β_1 and β_0. Hence, decrements between 0.71 and 1.15 percentage points of the digestibility per unit increment of stearic acid level are in agreement with the data on the 95% confidence level.

Confidence interval for expected value. The expected value of prediction at $x = x_0$ is obtained by the model with the estimates of intercept and the slope:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

It takes into account the estimation error and thus gives rise to the 95% confidence interval

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm (t_{0.025,n-2})s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SS}_x}} \tag{7.2}$$

for the expected value $y_0 = \beta_0 + \beta_1 x_0$.

Prediction interval. However, y_0 is subject to observation error. The observational error has standard deviation σ, and the prediction interval should take this source of variation into account, too. Intuitively, this corresponds to adding s to the residual standard error. Hence, the 95% prediction interval is computed as follows:

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm (t_{0.025,n-2}) s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SS}_x}}$$

The interpretation is that a (new) random observation at $x = x0$ will belong to this interval with probability 95%.

Example: Stearic acid and digestibility. If we consider a stearic acid level of $x_0 = 20\%$ then we will expect a digestibility percentage of

$$\hat{y}_0 = 96.5334 - (0.9337)(20) = 77.859$$

which has standard error

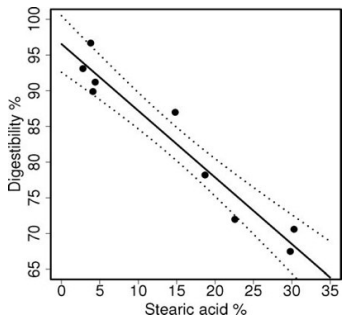
$$\mathrm{SE}(\hat{y}_0) = (2.970)\sqrt{\frac{1}{9} + \frac{(20 - 14.5889)^2}{1028.549}} = 1.1096$$

The 95% confidence interval for y_0 is given by

$$(77.859) \pm (2.365)(1.1096) = (75.26, 80.48)$$

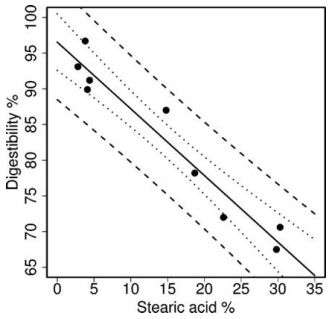
In conclusion, the predicted values of digestibility percentage corresponding to a stearic acid level of 20% between 75.2 and 80.5 are in accordance with the data on the 95% confidence level.

We can calculate the 95% confidence interval for the expected digestibility percentage for other values of the stearic acid level.



The lower and upper limits are indicated by dotted lines. The width of the confidence band is smallest when x_0 is close to $\bar{x}$ (interpolation) and becomes larger as x_0 moves away from $\bar{x}$ (extrapolation). This reflects that the confidence is higher within the range of data, as it is also clear from the formula (7.2).

The plots below display 95% prediction intervals (dashed lines), and 95% confidence intervals (dotted lines) for the digestibility data.



The prediction intervals are wider than the confidence intervals. Also notice that the confidence bands and the prediction bands are not straight lines. The closer x_0 is to the mean value, the more precise the prediction becomes (interpolation).

Confidence and prediction intervals.

- **Interpretation.** The confidence interval includes the expected values that are in accordance with the data (with a certain degree of confidence), whereas a new observation will be within the prediction interval with a certain probability.
- **Interval widths.** The prediction interval is wider than the corresponding confidence interval.
- **Dependence on sample size.** The confidence interval can be made as narrow as we want by increasing the sample size. This is not the case for the prediction interval.

R coding: Stearic acid and digestibility. By using the estimates and standard errors we can compute confidence intervals. Here we can make R do the computations for us and use the `confint()` function.

```
confint(model)
confint(model, level=0.90)
```

The output has one line per parameter, with the same names as in the output from `summary()`. Notice how we may specify the level of the confidence interval and that the 95% interval is computed if no level is specified.

The `predict()` function is applicable for computation of predicted values corresponding to the observations in a dataset. Consider the digestibility data and assume that we wish to predict the digestibility percent for levels 10, 20, and 25 of stearic acid. The predicted values are calculated in the linear regression model as follows:

```
new <- data.frame(stearic.acid=c(10, 20, 25))
new
predict(model, new, level=0.95, interval="confidence")
```

First, a new data frame with the new values of the explanatory variables is constructed. It has three observations and a single variable, stearic.acid. It is important that the variable has the same name as the explanatory variable in the original dataset. The `predict()` command asks R to calculate the predicted values. The type of the intervals may be changed with options.