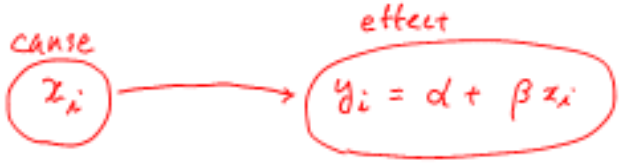
Simple linear regression model.



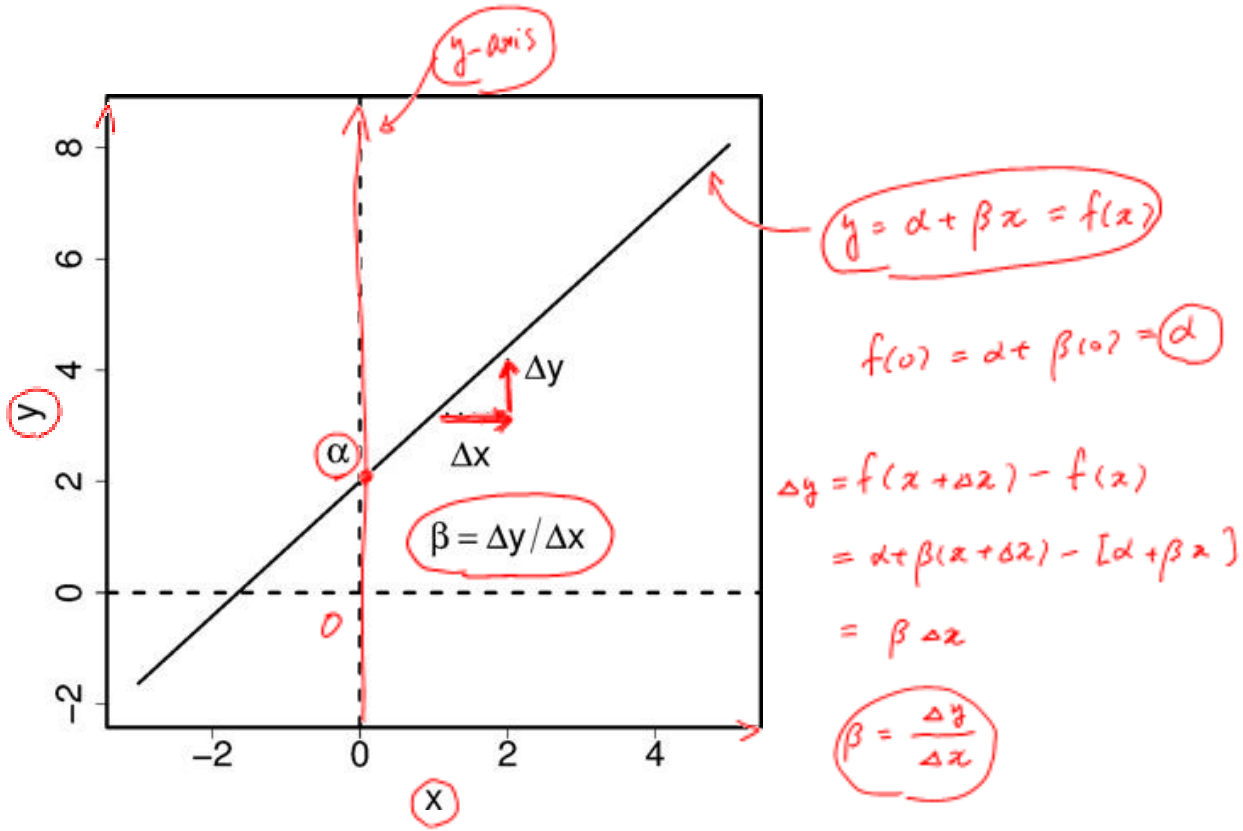
Simple linear regression attempts to model the relationship between two quantitative variables, x and y, by fitting a linear equation to the observed data.

$$y = \alpha + \beta x = f(x) \qquad (4.1)$$

Linear function

where α (also called the **intercept**) is the value of y when $x = 0$ and β is the **slope** (i.e., the change in y for each unit change in x). When we want to model the relationship between two variables we assume that one variable is the **dependent variable** (y in the linear equation (4.1)) while the other is an **explanatory variable** x in (4.1).

(independent variable)



We want to model y as a linear function of x in the hope that information about x will give us some information about the value of y. Therefore, it will "explain" the value of y, at least partly by the value of x.

Example: Stearic acid and digestibility.

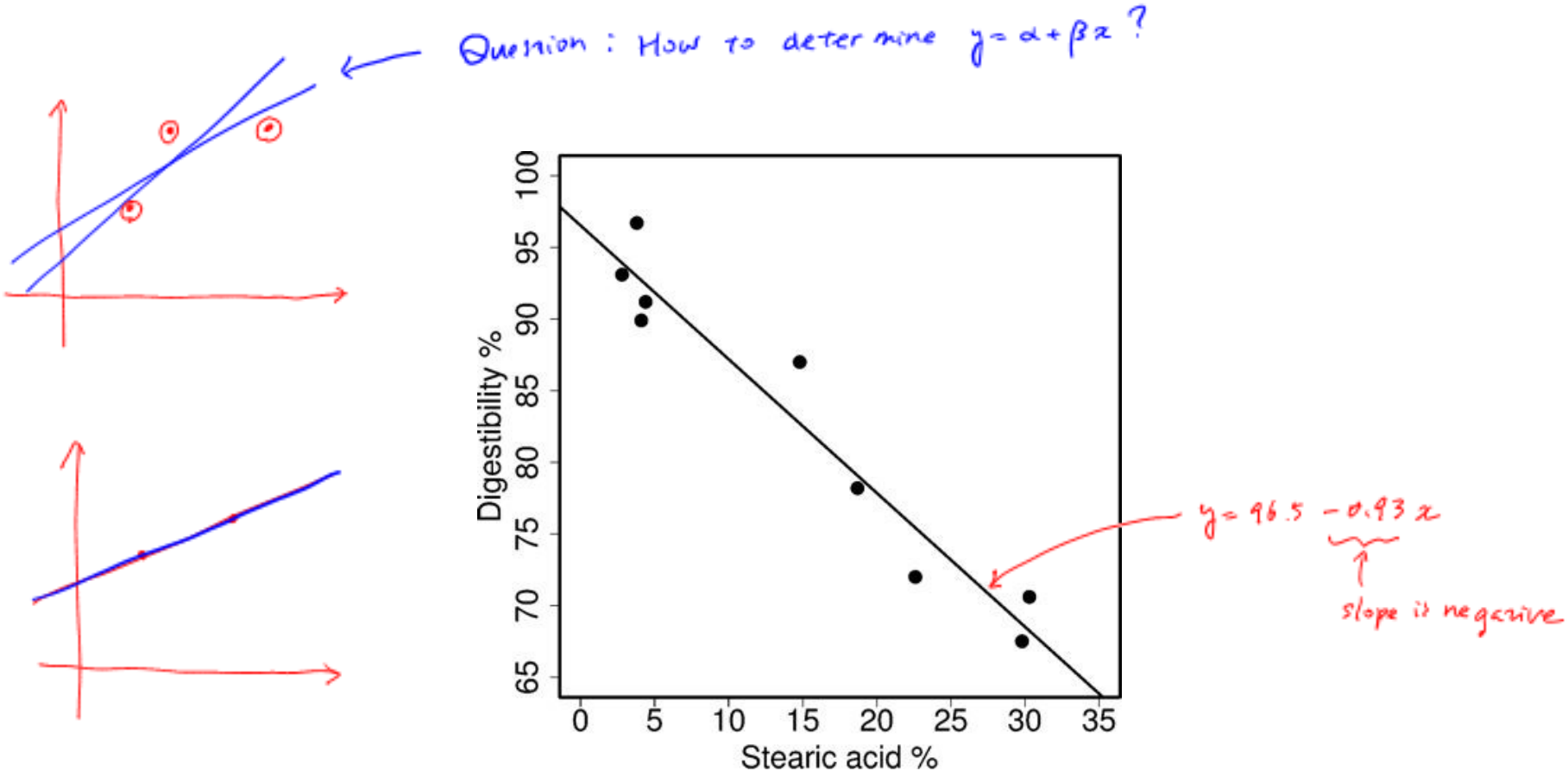
Researchers examined the digestibility of fat with different levels of stearic acid. The average digestibility percent was measured for nine different levels of stearic acid proportion. Data are shown in the table below, where x represents stearic acid and y is digestibility measured in percent.

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

The scatter plot can be obtained together with the straight line defined by

$$y = 96.5334 - 0.9337x$$

It will become clear why these values are used for the parameters in the model.



It shows that the relationship between stearic acid and digestibility appears to scatter around a straight line and that the line plotted in the figure seems to capture the general trend of the data.

Estimates for a regression line.

- Fitting a regression line means identifying the "best" line; i.e., the optimal parameters to describe the observed data.
- Let (x_i, y_i), $i = 1, \ldots, n$, denote our n pairs of observations and assume that we somehow "guess" the two parameters, α and β, from a linear equation.

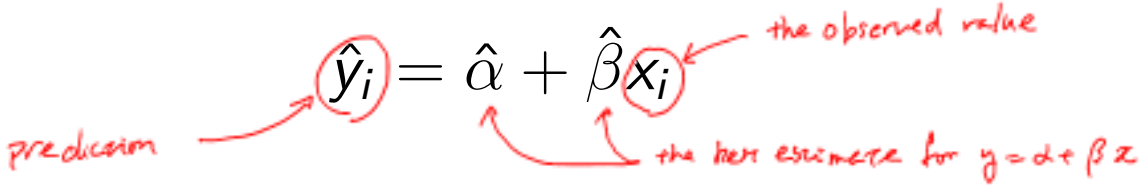
$$y = \hat{\alpha} + \hat{\beta}x$$

- They are used to model the relationship between the x's and the y's. Notice how we placed "hats" over α and β to indicate that the values are not necessarily the true (but unknown) values of α and β but **estimates**.

Residuals in a regression line.

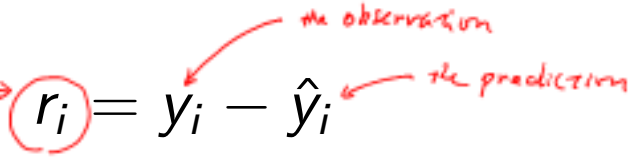
- For any x, we can use this model to predict the corresponding y-value. In particular, we can do so for each of our original observations, $x_1, \ldots, x_n$, to find the predicted values

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$



- We can use these predicted values to evaluate how well the model fits to the actual observed values. This is achieved by looking at the **residuals**, which are defined as follows:

$$r_i = y_i - \hat{y}_i$$



- The residuals measure how far away each of our actual observations (y_i's) are from the expected value given a specific model (the straight line in this case).

Example: Stearic acid and digestibility.

Let us for now assume that we have eyeballed the data and have found that a line defined by the parameters

$$\hat{\alpha} = 96.5334 \quad \hat{\beta} = -0.9337$$

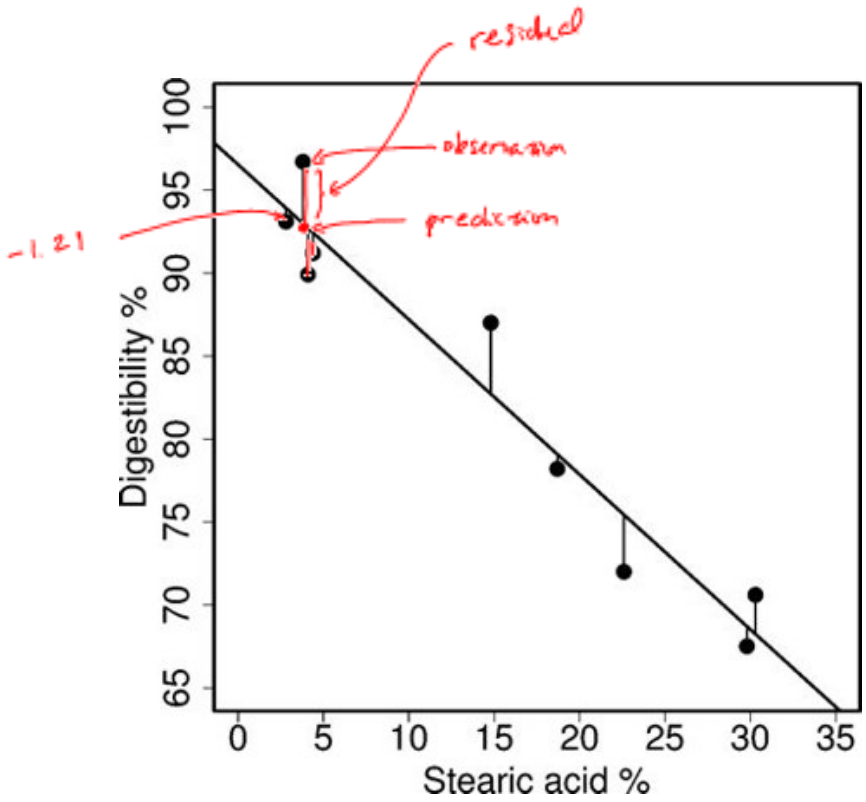
We can then calculate the predicted value for each observed x; e.g., for $x_1 = 29.8$ by

$$\hat{y}_1 = 96.5334 - (0.9337)(29.8) = 68.709$$

The best prediction

This value is slightly higher than the observed value of 67.5, and the residual for the first observation is

$$r_1 = 67.5 - 68.709 = -1.209$$



The figure above shows a graphical representation of the residuals for all nine levels of stearic acid. The vertical lines between the model (the straight line) and the observations are the residuals.

Least squares estimation.

The least squares method estimates the unknown parameters of a model by minimizing the sum of the squared deviations between the data and the model. Thus for a linear regression model we seek to identify the parameters α and β such that

$$\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$$

observed; $(\alpha + \beta x_i)$ predicted; residual

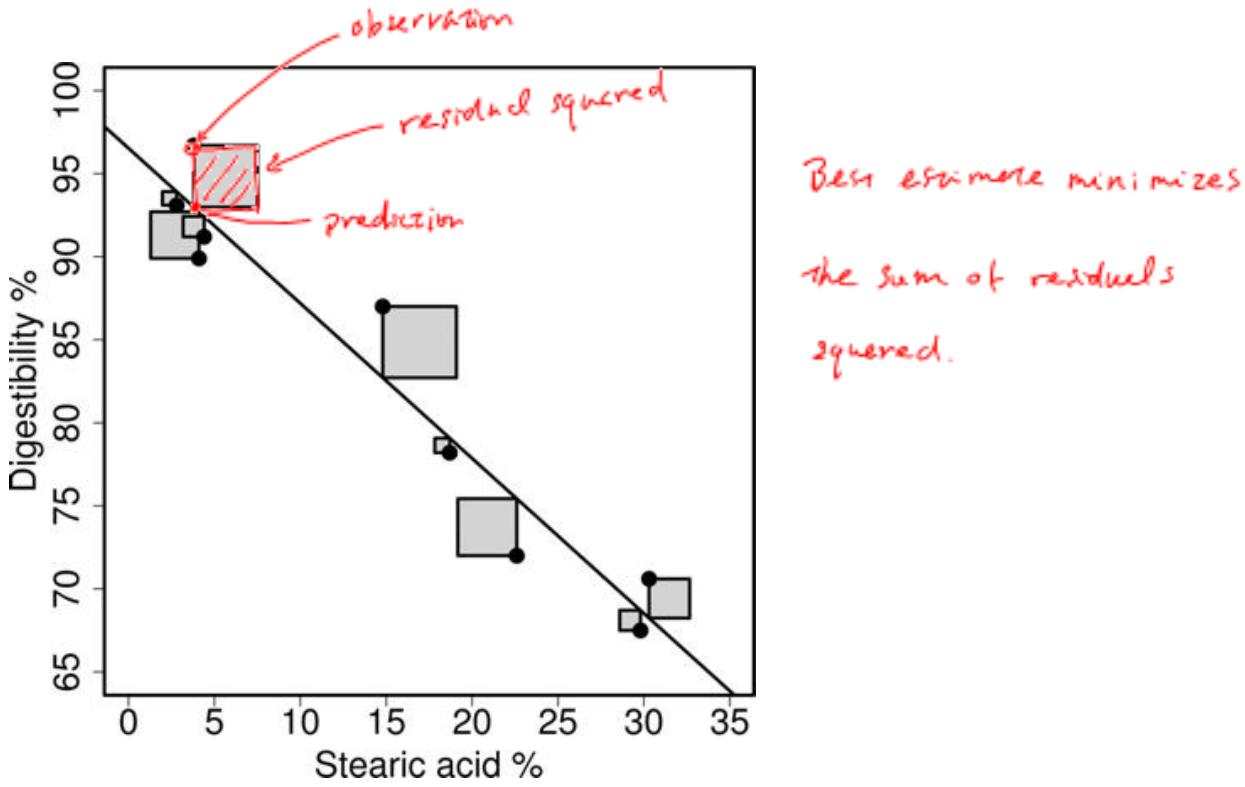
becomes as small as possible. The line that best fits the data has slope and intercept given by

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

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

observation

where $\bar{x} = \sum_{i=1}^{n} x_i/n$ and $\bar{y} = \sum_{i=1}^{n} y_i/n$ denote the mean values.



The figure above visualizes squared residuals for the dataset on digestibility and stearic acid. Gray areas represent the squared residuals for the proposed regression line.

Example: Stearic acid and digestibility.

The mean values of x and y are

$$\bar{x} = \frac{131.3}{9} = 14.5888 \quad \bar{y} = \frac{746.2}{9} = 82.9111$$

Once we have $\bar{x}$ and $\bar{y}$ we can obtain $(x_i - \bar{x})^2$ and $(x_i - \bar{x})(y_i - \bar{y})$ and finally calculate the estimated slope and intercept:

$$\hat{\beta} = \frac{-960.40}{1028.549} = -0.9337$$
$$\hat{\alpha} = 82.9111 - (-0.9337)(14.5888) = 96.5334$$

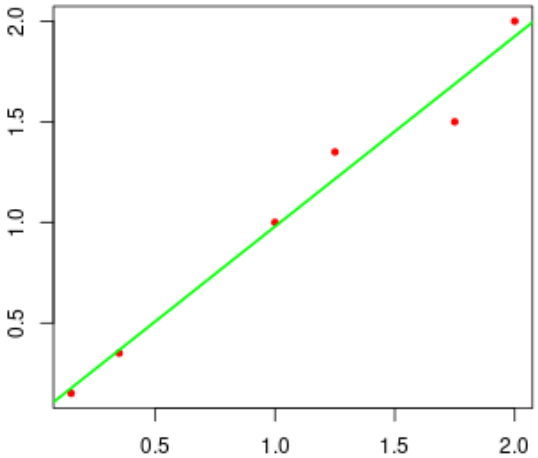
The regression line allows us to provide statements about the change in digestibility: "If we increase the stearic acid level by 10 percentage points we expect the digestibility to decrease by 9.33 percentage points."

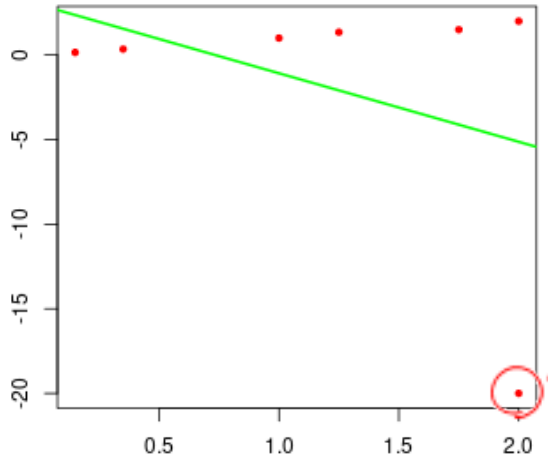
When is linear regression appropriate?

- **Paired Quantitative variables.** Linear regression applies only to a pair (x, y) of two quantitative variables. Both variables are quantitative before a linear regression is used to model the relationship between x and y. ⇒ You should be able to produce a scatter plot.
- **Does the relationship appear to be linear?** Is it reasonable to model the relationship between x and y as a straight line? We should always start our analysis by plotting the data and checking the overall relationship between x and y in a graph: a curvilinear relationship between x and y makes a linear regression inappropriate.
- **Influential points.** Influential points are data points with extreme values that greatly affect the slope of the regression line. Influential points are often outliers in the y-direction.

Example of an influential point.

The scatter plot located to the left shows the regression line. If we include this additional pair of data: $(x, y) = (2.00, -20.00)$, the corresponding plot is located to the right.





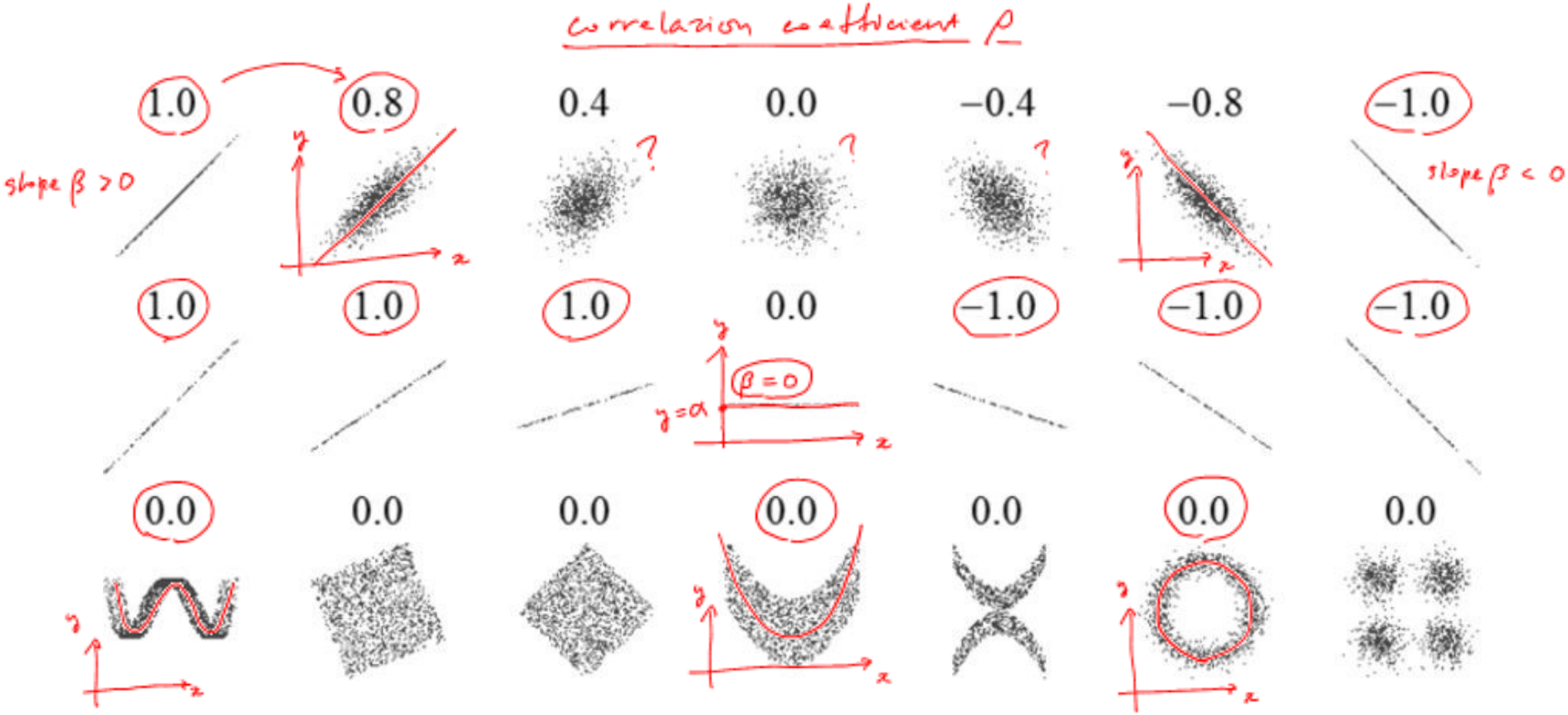
An additional point would be an influential point because the graph of the regression line would change considerably, as shown by the regression line located to the right.

The correlation coefficient.

The sample **correlation coefficient** describes the linear association between x and y and is defined as

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \bar{x})^2)(\sum_{i=1}^{n}(y_i - \bar{y})^2)}} \tag{4.2}$$

The sign of the correlation coefficient is identical to the sign of the regression slope. Another point worth noting is that x and y enter (4.2) symmetrically, so the correlation of x and y is identical to the correlation between y and x. The correlation is a measure of the strength of the linear relationship between the two variables and it can be shown that it is always between -1 and 1. In some situations we may expect an association between x and y but it may not be reasonable to say that x is the cause of y or vice versa.



The graphs above show correlation coefficients for different datasets. The second row of graphs illustrates that the slope estimate $\hat{\beta}$ has no influence on the correlation coefficient $\hat{\rho}$. The last row of graphs suggests that the correlation may be zero even though the data are structured.

- The correlation coefficient of unity occurs when the observations lie exactly on a straight line with some positive or some negative slope, and the value of negative one corresponds to the situation where the observations are exactly on a straight line with negative slope.
- The correlation coefficient is zero when the best-fitting straight line of y on x does not depend on the observed value of the x's.
- It is vital to remember that a correlation, even a very strong one, does not mean we can make any conclusion about causation. Moreover, there may be a strong non-linear relationship between x and y even though the correlation coefficient is zero.

When is correlation coefficient relevant?

- **Quantitative variables.** The sample correlation coefficient applies only to two quantitative variables in pairs. Make sure that both variables are quantitative before the correlation coefficient between x and y is calculated.
- **Visualize linear association by a scatter plot.** It is important to plot the data before the correlation coefficient is calculated—the association may be highly structured, but if the relationship is non-linear the correlation coefficient may still be zero.

Example: Stearic acid and digestibility.

The data file "stearicacid.txt" 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. To read this type of data set, we need to use the following command:

```
Data <- read.table(file.choose(), header=T)
```

Once it is read, we have to attach the data frame "Data" and find the names of variable with summary statistics:

```
attach(Data)
summary(Data)
```

In the call to `lm()` we specify the statistical model "`digest ~ stearic.acid`" which can be interpreted in the following way: The response "`digest`" is modeled as a linear function of the explanatory variable "`stearic.acid`." By default, R interprets numerical vectors, which is one of the requirements for a linear regression model.

```
model <- lm(digest ~ stearic.acid)
model
```

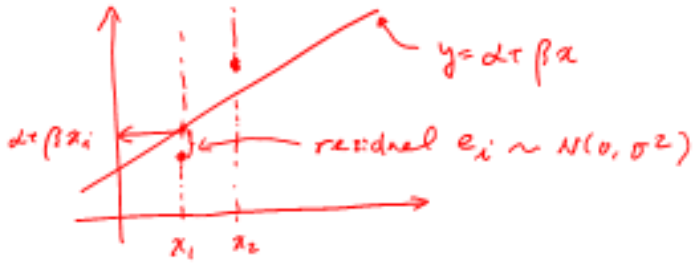
The output from `lm()` shows the estimated parameters from the model. The estimated intercept is found under Intercept to be 96.5334 and the slope, -0.9337, is listed under "`stearic.acid`."

In R, the correlation between two quantitative variables is calculated with the `cor()` function.

```
cor(digest, stearic.acid)
```

The correlation is a measure of the strength of the linear relationship between the two variables and it can be shown that it is always between -1 and 1. The value -0.9672452 corresponds to the situation where the observations should be very close to a straight line with negative slope.

Linear regression model.



The linear regression has two parameters, α and β, which determine the model

$$y_i = \alpha + \beta x_i + e_i, \quad i = 1, \ldots, n,$$

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 α, β, and σ. The slope parameter β is the expected increment in y as x increases by one unit, whereas α 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.

Standard errors for linear regression.

In the linear regression model, the estimate of parameters are given by

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

In order to simplify the above formula we can define

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

as the denominator. The **standard errors** are given by

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

where the estimate s is obtained from the **mean square error**

$$s^2 = \mathrm{MS}_e = \frac{\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2}{n-2}$$

Example: Stearic acid and digestibility.

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

$$\hat{\beta} = -0.9337 \quad \hat{\alpha} = 96.5334$$

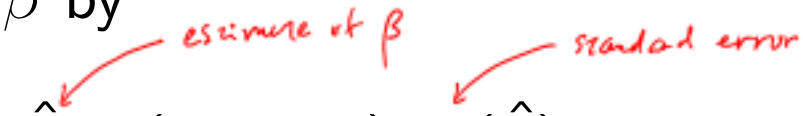
The denominator value $\mathrm{SS}_x = 1028.549$ in the definition for $\hat{\beta}$ and the mean square error $\mathrm{MS}_e = 8.8234$ are used to calculate

$$s = \sqrt{8.8234} = 2.970, \quad \mathrm{SE}(\hat{\beta}) = \frac{2.970}{\sqrt{1028.549}} = 0.0926$$

$$\mathrm{SE}(\hat{\alpha}) = (2.970)\sqrt{\frac{1}{9} + \frac{(14.5889)^2}{1028.549}} = 1.6752$$

Example: Stearic acid and digestibility.

We need to use a critical value $t_{\alpha/2,n-2}$ from the t-distribution with $n-2$ degrees of freedom, and obtain $(1-\alpha)\%$ confidence interval for the slope parameter β by



$$\hat{\beta} \pm (t_{\alpha/2,n-2})\mathrm{SE}(\hat{\beta})$$

There are $n = 9$ observations and two parameters. Since $t_{0.05,7} = 1.895$ and $t_{0.025,7} = 2.365$, we compute the 90% and the 95% confidence interval

$$-0.9337 \pm (1.895)(0.0926) = (-1.11, -0.76)$$
$$-0.9337 \pm (2.365)(0.0926) = (-1.15, -0.71)$$

Hence, decrements between 0.76 and 1.11 percentage points of the digestibility per unit increment of stearic acid level are in agreement with the data on the 90% confidence level.

Standard error for the predicted value.

In regression analysis we often seek to estimate the prediction for a particular value of x. Let x_0 be such an x-value of interest. The predicted value of the response is denoted y_0; that is, $y_0 = \alpha + \beta x_0$. It is estimated by

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

The predicted value has the standard error

$$\mathrm{SE}(\hat{y}_0) = s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SS}_x}}$$

The $(1-\alpha)\%$ confidence interval for the predicted value y_0 is given by

$$\hat{y}_0 \pm (t_{\alpha/2, n-2})\mathrm{SE}(\hat{y}_0)$$

Example: Stearic acid and digestibility.

If we consider a stearic acid level of $x_0 = 20\%$ then we will expect a digestibility percentage of

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

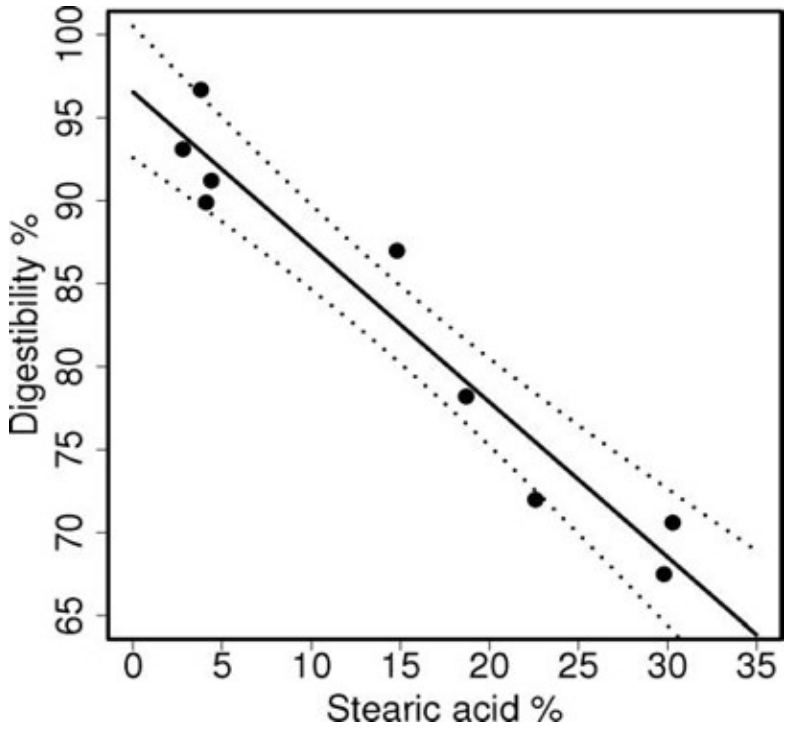
which has standard error

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

The 95% confidence interval for y_0 is given by

$$77.859 \pm (2.365)(1.1096) = (75.235, 80.483)$$

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.



In the figure above we calculated the confidence interval for the expected digestibility percentage for other values of the stearic acid level. The lower and upper limits are shown in the dotted lines. The width of the confidence band is smallest close to $\bar{x}$ and becomes larger as x_0 moves away from $\bar{x}$.

When is it appropriate to predict?

- **x on y or y on x?** The regression of x on y is different from the regression of y on x, and we have to fit a new model with digestibility as the explanatory variable and stearic acid as the response variable if we wish to predict stearic acid levels from digestibility. In some experiments it is clear which of the variables should take the role of the response variable, y, and which variable should take the role of the explanatory variable.
- **Interpolation** is making a prediction within the range of observed values for the explanatory variable x. It enables us to predict values of y for values of x that do not exist in the sample.
- **Extrapolation** concerns the situations where predictions are made outside the range of values used to estimate the model parameters. The prediction becomes increasingly uncertain when the distance to the observed range of values is increased as there is no way to check that the relationship continues to be linear outside the observation range.

Hypothesis test in regression model.

Recall the linear regression model for the digestibility data

$$y_i = \alpha + \beta x_i + e_i, \quad i = 1, \ldots, n,$$

where $e_1, \ldots, e_n$ are residuals. The null hypothesis

$$H_0 : \beta = 0$$

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

$$T_{\text{obs}} = \frac{\hat{\beta} - 0}{SE(\hat{\beta})} = \frac{-0.9337 - 0}{0.0926} = -10.08$$

where the estimate -0.9337 and its standard error 0.0926 are obtained from the data. The value of T_{obs} should be compared to the t_7-distribution.

For the alternative hypothesis "$H_A : \beta \neq 0$" we get

$$p\text{-value} = P(|T| \geq 10.08) = 0.00002$$

and we conclude that there is strong evidence of an association between digestibility and the stearic acid level. That is, 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. When we fail to reject H_0 we should take the following consequences into consideration:

- The hypothesis test for the slope coefficient β is used to validate the linear regression model by rejecting the null hypothesis.
- If the linear regression is not validated then the best predicted value $\hat{y}$ is simply the mean $\bar{y}$ of the y_i's

Example: Stearic acid and digestibility.

Let us first study the output corresponding to a linear regression model and run an analysis of the digestibility data. First the linear regression model is fitted and stored in the object model.

```
model <- lm(digest ~ 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.

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, level=0.90) # 90% confidence intervals
confint(model) # Default is 95%
```

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, interval="confidence", level=0.95)
```

First, a new data frame with the new values of the explanatory variables is constructed. It has three observations for the 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 level of the confidence intervals may be changed with the level option.

Example: Song of Insects.

`ChirpsData` contains the numbers of chirps in one minute and the corresponding temperatures in Fahrenheit.

```
ChirpsData <- read.csv(file.choose(), header=T)
ChirpsData
attach(ChirpsData)
```

The correlation between two quantitative variables is calculated with the `cor()` function.

```
cor(Temp, Chirps)
```

Here the numbers of chirps in one minutes is used to predict temperatures outside. By default, R interprets numerical vectors—in our case, both Chirps and Temp—as quantitative variables, which is one of the requirements for a linear regression model.

```
outcome <- lm(Temp ~ Chirps)
summary(outcome)
```

The variable "outcome" contains everything we need to know about the model. Here we have two parameters: the intercept and the slope. The estimated intercept is found at the intercept and the slope is listed at Chirps. The respective p-values test the null hypothesis "$H_0 : \alpha = 0$" (intercept) or "$H_0 : \beta = 0$" (slope). By rejecting H_0 for the slope parameter β we find evidence of linear relationship between two quantitative variables, the numbers of chirps and the temperatures.

We use a simple scatter plot to illustrate the relationship between two quantitative variables. The `plot()` function is used to produce a scatter plot, and we can add additional information to the plot by specifying the labels for the x-axis and the y-axis with the xlab and ylab options to `plot()`.

```
plot(Chirps, Temp, xlab="Chirps per Minute",
ylab="Temperature")
```

The `abline()` function adds a straight line to an existing plot, and we can use abline() to illustrate the estimated linear relationship between the two variables.

```
abline(outcome, col="red")
```