Chapter 2

Linear Regression

2.1 Introduction

This chapter is based on the book [Searle, 1971] and on the course notes [Bühlmann and Mächler, 2016]. Regression analysis is used to assess relationships between a given variable and other measurements or observations on the same animal. The relationships between the variable and the other characteristics of the animal are estimated based on observed data.

2.2 Example

A classical example of a regression analysis in animal science is the relationship between body weight and breast circumference in cattle. This example has a practical application because the results of the regression analysis of body weight on breast circumference can be used for a measuring band. With such a measuring band the breast circumference of an animal is measured. On the back side of the band, the estimated body weight can directly be determined.

At this point the question is how is it possible to determine the relationship between the values of breast circumference in centimeters and body weight in kilograms. The answer to this question can be given by a regression analysis. The most important pre-requisites for doing a regression analysis is to have a dataset. For our example, Table 2.1 shows such a dataset which can be used for a regression analysis.

Table 2.1: Breast Circumference and Body Weight for 10 Animals

Animal	Breast Circumference	Body Weight
1	176	471

2	177	463
3	178	481
4	179	470
5	179	496
6	180	491
7	181	518
8	182	511
9	183	510
10	184	541

The dataset in Table 2.1 contains measurements of body weight and breast circumference for 10 animals. Figure 2.1 is a graphical representation of our example dataset.

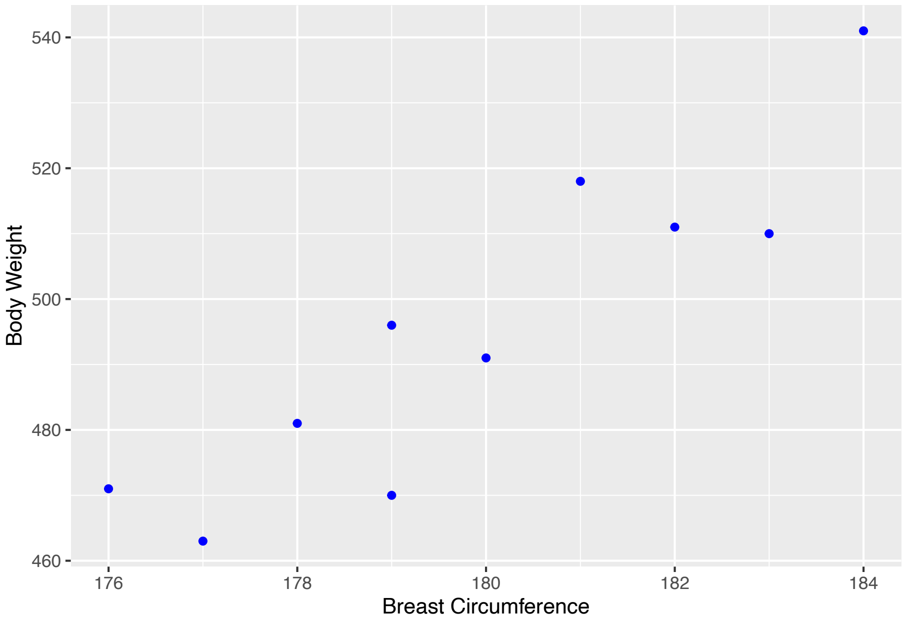

Figure 2.1: Breast Circumference and Body Weight

The diagram in Figure 2.1 shows on the x-axis the breast circumference in cm and on the y-axis the body weight in kg. Each of the blue dots correspond to an observation of one animal in the dataset. A diagram like the one shown in Figure 2.1 is also called a *dot plot*. From a first visual inspection of the dot plot for our dataset, we can see that there is a tendency that larger values of breast circumference of animals are related to heavier animals. The relationship is not deterministic that means there are exceptions which do not follow the rule of this relationship. One example of an exception are animals 1 and 2.

Animal 2 has a larger breast circumference value compared to animal 1, but animal 2 has a lower body weight compared to animal 1. But despite such exceptions, we can still observe that on average there is a relationship between breast circumference and body weight. Furthermore, the apparent relationship between breast circumference and body weight seams to be the same for low and high values of breast circumference. Based on this last fact, we can say that the relationship between breast circumference and body weight is *linear*.

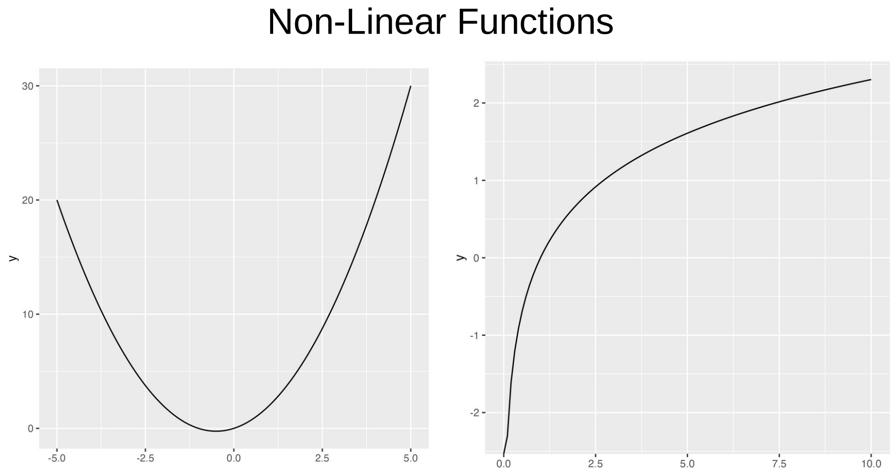

Figure 2.2: Examples of Non-Linear Functions

Figure 2.2 shows two examples of non-linear functions. Both examples show that the relationship between the shown variables x and y is not the same over the shown range of values. Hence by inspecting the diagram of the two variables breast circumference and body weight of our example in Figure 2.1, we can say that the relationship between the variables of our example dataset show a linear relationship.

2.3 Regression Model

With the regression model we want to find a mathematical formulation to describe and to quantify the relationship between the two variables from our example, breast circumference and body weight. One possibility to find this relationship is to take an animal with x cm of breast circumference. Then the question is what would be the expected value for its body weight y in kg. Under the assumption of a linear relationship between the variables from our example, the expected value $E(y)$ for the body weight y can be written as

$$E(y) = b_0 + b_1 * x \qquad (2.1)$$

The above reasoning on how the variables are related is often referred to as *model building*. The model shown in (2.1) is called a linear model, because the expected body weight ($E(y)$) is a linear function of the unknown *parameters* b_0 and b_1. The number of possible models between two variables x and y is infinite. And therefore the process of model building is difficult and requires some experience. But in general, we can say that a simpler model with fewer unknown parameters is always preferable over a more complex model as long as the simpler model is able to capture the most important aspects of a relationship between two variables.

2.4 Observations

For an animal with a breast circumference of x cm, the body weight (y) will not exactly be $b_0 + b_1 * x$. It has to be noted here that the values for b_0 and b_1 will be the same for all animals. The fact of the discrepancy between the recorded body weights (y) and the output of the model is taken into account by writing $E(y)$ instead of y in the model shown in equation (2.1). For a given observed body weight y_i of animal i with a breast circumference of x_i, we can write

$$E(y_i) = b_0 + b_1 * x_i \tag{2.2}$$

where $E(y_i)$ is not the same as y_i. The difference $y_i - E(y_i)$ represents the difference between the observed body weight from its expected value $E(y_i)$ and is written as

$$e_i = y_i - E(y_i) = y_i - b_0 - b_1 * x_i \tag{2.3}$$

Hence for the body weight y_i of animal i, we can write

$$y_i = b_0 + b_1 * x_i + e_i \tag{2.4}$$

Equations (2.2), (2.3) and (2.4) apply to all observations $y_1, y_2, \ldots, y_{10}$, of our example dataset shown in Table 2.1. The e_i terms for all observations might take many different values. They include potential measurement errors or deficiencies of the model itself. Due to the described properties of e_i's, they are considered to be random variables and are usually called *random errors* or *random residuals*.

To complete the description of our model in terms of equation (2.4), further characteristics of the random errors (e_i) must be specified. These characteristics consist of

- the *expected value* $E(e_i)$ of e_i and
- the variance $var(e_i)$ of e_i.

Useful specifications are that the expected value $E(e_i)$ is zero and all covariances between any pair of e_i and e_j with $i \neq j$ are also zero. Then the variance $var(e_i)$ is assumed to be a constant for all i and is represented by the symbol σ^2. Summarizing all the proposed properties in a mathematical notation, we obtain

$$E(e_i) = 0 \tag{2.5}$$

which is obtained from the definition of e_i given in (2.3). The variance $var(e_i)$

$$var(e_i) = E\left[e_i - E(e_i)\right]^2 = E(e_i^2) = \sigma^2 \tag{2.6}$$

and

$$cov(e_i, e_j) = E\left[e_i - E(e_i)\right]\left[e_j - E(e_j)\right] = E(e_i e_j) = 0 \tag{2.7}$$

Equations (2.2) - (2.7) give a constitutional description of the linear model that we have designed so far for our example dataset. These properties form the basis for the procedure used to estimated the unknown parameters b_0 and b_1.

2.5 Parameter Estimation

There are several methods to estimate the unknown parameters b_0 and b_1 of the proposed linear model. The most frequently used method which is also implemented in the R-function `lm()` is called *least squares*.

Least squares estimation is based on the idea of minimizing the sum of the squared deviations of the observations y_i from their expected values. This sum can be written as

$$\mathbf{e}^T\mathbf{e} = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left[y_i - E(e_i)\right]^2 = \sum_{i=1}^{N} \left[y_i - b_0 - b_1 * x_i\right]^2 \tag{2.8}$$

where $\mathbf{e}^T = \left[\begin{array}{cccc} e_1 & e_2 & ... & e_N \end{array}\right]$ is the transpose of the vector $\mathbf{e}$ of length N containing all the e_i values and N stands for the number of observations.

Although b_0 and b_1 are fixed (but unknown) values, we treat them for a moment like mathematical variables. The reason for this changed view is that we want to find the values for b_0 and b_1 that minimize the expression in (2.8). The resulting values from the minimization of (2.8) will be represented by the symbols $\hat{b}_0$ and $\hat{b}_1$ and they will be called the least squares estimators of b_0 and b_1.

Minimization of (2.8) is done by taking partial derivatives with respect to both unknowns b_0 and b_1 and setting both derivatives to zero[1]. This yields two equations from which solutions called $\hat{b}_0$ and $\hat{b}_1$ can be computed.

$$\begin{aligned}\frac{\partial \mathbf{e}^T\mathbf{e}}{\partial b_0} &= -2\sum_{i=1}^{N}\left[y_i - b_0 - b_1 x_i\right] \\ &= -2\left[\sum_{i=1}^{N} y_i - N b_0 - b_1 \sum_{i=1}^{N} x_i\right] \end{aligned} \tag{2.9}$$

$$\begin{aligned}\frac{\partial \mathbf{e}^T\mathbf{e}}{\partial b_1} &= -2\sum_{i=1}^{N} x_i \left[y_i - b_0 - b_1 x_i\right] \\ &= -2\left[\sum_{i=1}^{N} x_i y_i - b_0 \sum_{i=1}^{N} x_i - b_1 \sum_{i=1}^{N} x_i^2\right] \end{aligned} \tag{2.10}$$

Setting both expression in (2.9) and (2.10) to zero and writing them in terms of $\hat{b}_0$ and $\hat{b}_1$ gives

$$N\hat{b}_0 + \hat{b}_1 x. = y. \tag{2.11}$$

and

$$\hat{b}_0 x. + \hat{b}_1 (x^2). = (xy). \tag{2.12}$$

using the dot notation for the following sums $x. = \sum_{i=1}^{N} x_i$, $y. = \sum_{i=1}^{N} y_i$, $(x^2). = \sum_{i=1}^{N} x_i^2$ and $(xy). = \sum_{i=1}^{N} x_i y_i$. With the bar notation for the means, we can further write

$$\bar{x}. = \frac{x.}{N} \tag{2.13}$$

and

$$\bar{y}. = \frac{y.}{N} \tag{2.14}$$

The solutions in (2.15) and (2.16) can then be written as

[1] The verification of higher order derivative to confirm that the obtained extreme value is a minimum is not done here.

$$\hat{b}_0 = \bar{y}. - \hat{b}_1 \bar{x}. \tag{2.15}$$

and

$$\hat{b}_1 = \frac{(xy). - N\bar{x}.\bar{y}.}{(x^2). - N\bar{x}.^2} \tag{2.16}$$

2.6 Estimates for Example Dataset

The estimates as shown in (2.15) and (2.16) are computed for our example dataset. We start by computing all the components of the formulas for the estimators, then we plug those components in and get the results.

$$N = 10,\ \bar{x}. = 179.9,\ \bar{y}. = 495.2,\ (xy). = 891393,\ (x^2). = 323701$$

$$\hat{b}_1 = \frac{891393 - 10 * 179.9 * 495.2}{323701 - 10 * 179.9^2} = 8.673 \tag{2.17}$$

$$\hat{b}_0 = 495.2 - 8.6732348 * 179.9 = -1065.115 \tag{2.18}$$

2.7 Obtain Parameter Estimates in R

As already mentioned, the same type of computation as shown in Section 2.6 is also implemented in the R-function `lm()`. In what follows, we show how the estimates of the linear regression model are obtained using `lm()`. An important pre-requisite for using the function `lm()` is that the dataset is assigned to a dataframe. Here we assume that we have a dataframe named `tbl_reg` that contains our dataset. Then the parameter estimates are obtained using the following statements in R.

```r
lm_bw_bc <- lm(`Body Weight` ~ `Breast Circumference`, data = tbl_reg)
summary(lm_bw_bc)
```

```
##
## Call:
## lm(formula = `Body Weight` ~ `Breast Circumference`, data = tbl_reg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.3941  -6.5525  -0.0673   9.3707  13.2594
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1065.115    255.483  -4.169 0.003126 **
## `Breast Circumference`     8.673      1.420   6.108 0.000287 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.08 on 8 degrees of freedom
## Multiple R-squared:  0.8234, Adjusted R-squared:  0.8014
## F-statistic: 37.31 on 1 and 8 DF,  p-value: 0.000287
```

The output of the summary function shows a lot of information about the data and the model. Under the section `Coefficients` there are two entries

- `(Intercept)` corresponding to our b_0 and
- `Breast Circumference` corresponding to our b_1.

The values in the first column entitled `Estimate` correspond to the values that we have computed in the previous section.

2.8 The General Case

Suppose that in the study on body weight and breast circumference, we have an additional observation for each animal consisting of the height of the animal. The new extended data set is shown in Table 2.2.

Table 2.2: Extended Dataset of Body Weight for 10 Animals

Animal	Breast Circumference	Body Weight	Height
1	176	471	161
2	177	463	121
3	178	481	157
4	179	470	165
5	179	496	136
6	180	491	123
7	181	518	163
8	182	511	149
9	183	510	143
10	184	541	130

The model developed so far is not extended to be

$$E(y) = b_0 + b_1x_1 + b_2x_2 \tag{2.19}$$

where x_1 represents the breast circumference in cm, and x_2 stands for the height of the animal in cm. Thus for animal i with a breast circumference of x_{i1} cm and a height of x_{i2} the body weight y_i can be written as

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + e_i \tag{2.20}$$

As the number of x variables increase, the equations such as shown in (2.20) are getting longer and they are getting tedious to handle. This problem is solved by a change of notation. Let us define the matrix $\mathbf{X}$ and the vectors $\mathbf{y}$, $\mathbf{e}$ and $\mathbf{b}$ as follows.

$$\mathbf{X} = \left[\begin{array}{ccc} x_{10} & x_{11} & x_{12} \\ x_{20} & x_{21} & x_{22} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ x_{N0} & x_{N1} & x_{N2} \end{array}\right], \ \mathbf{y} = \left[\begin{array}{c} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_N \end{array}\right], \ \mathbf{e} = \left[\begin{array}{c} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_N \end{array}\right] \text{ and } \mathbf{b} = \left[\begin{array}{c} b_0 \\ b_1 \\ b_2 \end{array}\right]$$

Because all equations contain the term b_0, the first column of matrix $\mathbf{X}$ consists of all ones. Hence, $x_{10} = x_{20} = ... = x_{N0} = 1$. The complete set of equations for all animals in our extended dataset represented by (2.20) is

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}, \text{ with } E(\mathbf{y}) = \mathbf{X}\mathbf{b} \tag{2.21}$$

The big advantage of the matrix-vector notation is that equation (2.21) is invariant to the number of x-variables. That means no matter how many x-variables we include in our dataset, the linear regression model can always be written as shown in equation (2.21). The only thing that changes are the definitions of $\mathbf{X}$ and $\mathbf{b}$. In the general case with k variables

$$\mathbf{X} = \left[\begin{array}{cccc} x_{10} & x_{11} & \cdot & x_{1k} \\ x_{20} & x_{21} & \cdot & x_{2k} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ x_{N0} & x_{N1} & \cdot & x_{Nk} \end{array}\right], \ \mathbf{b} = \left[\begin{array}{c} b_0 \\ b_1 \\ \cdot \\ \cdot \\ b_k \end{array}\right]$$

The topic of parameter estimation can also be shown using matrix-vector notation. The only restriction that we are currently imposing is that the number of x-variables k is smaller than the number of observations N.

The model specification is only complete, if the properties of the vector $\mathbf{e}$ of random residuals is defined. In accordance with equations (2.5), (2.6) and (2.7), we can write

$$E(\mathbf{e}) = \mathbf{0} \tag{2.22}$$

where $E(\mathbf{e})$ stands for the vector of length N containing the expected values of all random residuals and $\mathbf{0}$ is a vector of N zeros and

$$var(\mathbf{e}) = E\left[\mathbf{e} - E(\mathbf{e})\right]\left[\mathbf{e} - E(\mathbf{e})\right]^T = E(\mathbf{e}\mathbf{e}^T) = \sigma^2 \mathbf{I}_N \tag{2.23}$$

where $var(\mathbf{e})$ is the variance-covariance matrix between all random residuals and $\mathbf{I}_N$ is the $N \times N$ identity matrix.

Derivation of the least squares estimator of $\mathbf{b}$ follows the same principles as shown in equations (2.9) - (2.16). The sum of squares of the deviations of the observations from their expected values using $E(\mathbf{e}) = \mathbf{0}$ and hence $E(\mathbf{y}) = \mathbf{X}\mathbf{b}$, is

$$\begin{aligned} \mathbf{e}^T\mathbf{e} &= \left[\mathbf{y} - E(\mathbf{y})\right]^T \left[\mathbf{y} - E(\mathbf{y})\right] \\ &= \left[\mathbf{y} - \mathbf{X}\mathbf{b}\right]^T \left[\mathbf{y} - \mathbf{X}\mathbf{b}\right] \\ &= \mathbf{y}^T\mathbf{y} - 2\mathbf{b}^T\mathbf{X}^T\mathbf{y} + \mathbf{b}^T\mathbf{X}^T\mathbf{X}\mathbf{b} \end{aligned} \tag{2.24}$$

The least squares estimator $\hat{\mathbf{b}}$ is found by minimizing $\mathbf{e}^T\mathbf{e}$ with respect to all elements of $\mathbf{b}$. This is corresponds to $\partial\mathbf{e}^T\mathbf{e}/\partial\mathbf{b}$ and is also called the gradient of $\mathbf{e}^T\mathbf{e}$ with respect to $\mathbf{b}$. Setting the gradient to zero leads to the following equations

$$\mathbf{X}^T\mathbf{X}\hat{\mathbf{b}} = \mathbf{X}^T\mathbf{y} \tag{2.25}$$

These equations are known as the least squares *normal equations*. Provided $(\mathbf{X}^T\mathbf{X})$ can be inverted, the unique solution for $\hat{\mathbf{b}}$ can be written as

$$\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{2.26}$$

2.9 Example Dataset Continued

We use the original dataset shown in Table 2.1 to illustrate the estimation of the least squares parameters using the matrix-vector notation. The matrix $\mathbf{X}$ and the vectors $\mathbf{y}$, $\mathbf{b}$ and $\mathbf{e}$ are defined as follows.

$$\mathbf{X} = \begin{bmatrix} 1 & 176 \\ 1 & 177 \\ . & . \\ . & . \\ . & . \\ 1 & 184 \end{bmatrix}, y = \begin{bmatrix} 471 \\ 463 \\ . \\ . \\ . \\ 541 \end{bmatrix}, b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}, e = \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_{10} \end{bmatrix}$$

The solution is obtained from equation (2.26) for this we also need

$$X^T X = \begin{bmatrix} 10 & 1799 \\ 1799 & 323701 \end{bmatrix}, \, (X^T X)^{-1} = \begin{bmatrix} 531.529 & -2.954 \\ -2.954 & 0.016 \end{bmatrix}, \, X^T y = \begin{bmatrix} 4952 \\ 891393 \end{bmatrix}$$

The solution vector $\hat{\mathbf{b}}$ contains the two components corresponding to the estimates of the two unknown parameters.

$$\hat{\mathbf{b}} = \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \end{bmatrix} = \begin{bmatrix} -1065.115 \\ 8.673 \end{bmatrix}$$

Comparing the above shown solutions to the results received earlier shows that for this example, the solution using the matrix-vector notation are the same.