CHAPTER 6

Linear Model with One Explanatory Variable

6.1. Model specification

A linear model with one explanatory variable (or univariate linear model) is any specification of the form

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{6.1}$$

where i is an observation, y is the continuous outcome variable, x is the explanatory variable, β_0 is the intercept, β_1 is the slope, and ε is the error term.

In the linear model you deal with obtaining estimates for the model parameters, which are the intercept ($\hat{\beta_0}$) and the slope ($\hat{\beta_1}$), and allow you to link changes in x to changes in y. The predictor variable (x) and the outcome variable (y) are observable and not random, and the error term (ε) is not observable and random.

A linear model can be used to explain a the dependent variable (y) in terms of the independent variable (x), and also to predict the value of the outcome variable for a given value of the predictor variable.

It it usual to focus on the slope because it is the parameter that explains the relationship between the predictor and the outcome variable, and it is used for theory testing.

When we have a relationship such as $F = 32 + \frac{9C}{5}$ to convert Celcius to Fahrenheit, we have a linear relationship. Finding $\hat{\beta_0}$ and $\hat{\beta_1}$ is equivalent to finding the values of 32 and 9/5 in the temperature conversion equation by means of an experiment.

I have recorded the temperature in Celcius and Fahrenheit for 5 days outside Sid Smith at noon with two thermometers, on in Celcius and one in Fahrenheit, which are actually clock-thermometer-calendar digital devices from the office and I doubt these are calibrated measurement devices. These were the results:

Day	Temperature (Celsius)	Temperature (Fahrenheit)
1	24	75
2	25	77
3	26	79
4	27	81
5	24	75

Table 6.1: Temperature in Celcius and Fahrenheit

A simple plot of the data shows that, while the relationship is not perfect, mostly because my termomethers are not highly precise but are still informative, there is a linear relationship between the two variables, and I can see my estimates for $F = \hat{\beta}_0 + \hat{\beta}_1 C$.

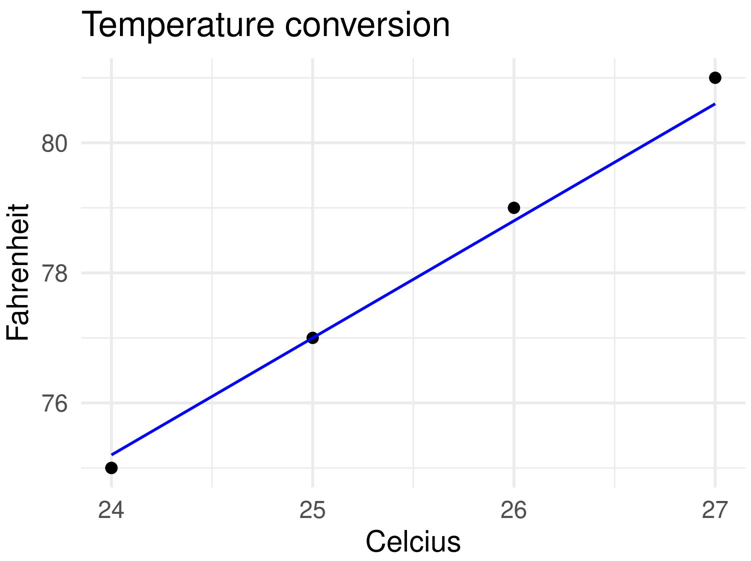

Figure 6.1: Temperature conversion from Celcius to Fahrenheit.

In this case, I cannot find a line that contains all the points (x_i, y_i) from the observations in the experiment. The added red line, which is the exact celcius to fahrenheit relation, was added to show that my experiment allows me to approximate the true relationship.

When you want to explore the relationship between x and y by means of a linear model, the inclusion of an error term accounts for the fact that the outcome variable is not perfectly explained by the predictor variable, and what you do is to fit a trend.

In the next plot created with simulated data around the line $y = 2 + 3x$, something totally arbitrary for the example. The blue line is what you will find by obtaining the values for $\hat{\beta}_0$ and $\hat{\beta}_1$.

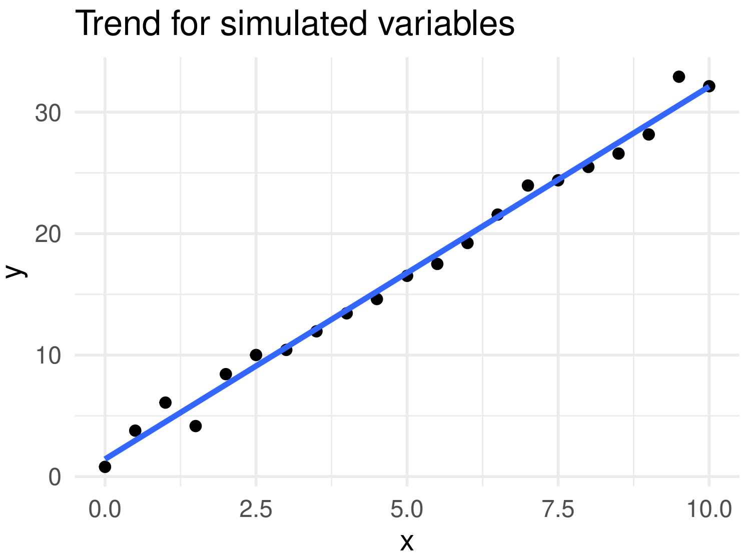

Figure 6.2: Trend for simulated variables.

Real life data is not always that perfect. For example, you can have a situation like the next plot, where there is no linear relationship between the variables.

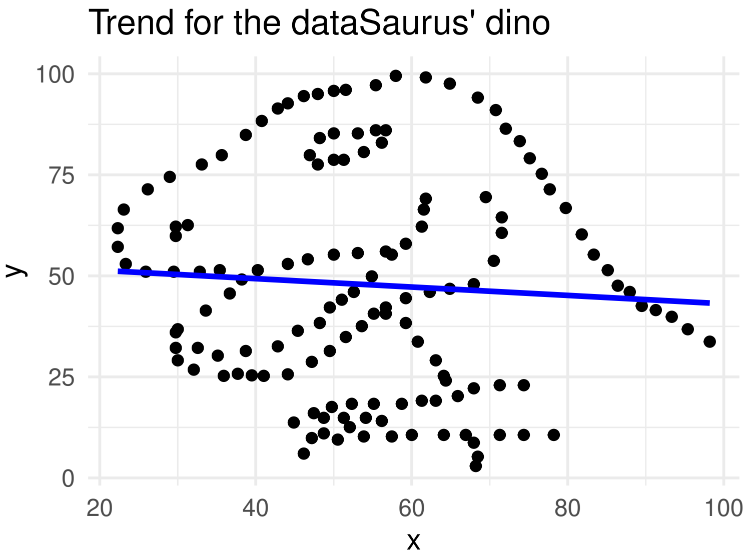

Figure 6.3: Trend for the dataSaurus. Source: Adapted from Cairo (2013) and Davies, Locke, and D'Agostino McGowan (2022).

In some situations you can have influential points, which are observations that are far away from the rest of the data, and that distort the estimation as in the next plots.

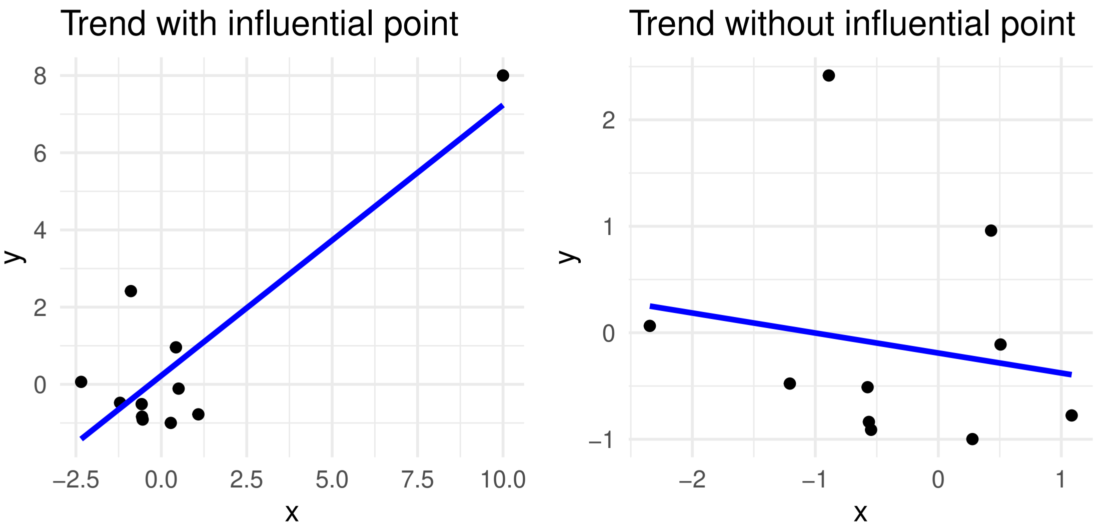

Figure 6.4: Trend with and without influential point.

This is why you should explore your datasets, be aware of their issues and dedicate time to conduct an exploratory data analysis and create plots before fitting a model.

Galton (1886) discovered that, with one predictor variable, you can estimate the model parameters by:

$$\hat{\beta}_1 = \mathrm{cor}(y, x)\frac{\mathrm{sd}(y)}{\mathrm{sd}(x)} \text{ and } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}. \tag{6.2}$$

The first equation is the slope of the line, and the second equation is the intercept.

Back to the temperature experiment, you can estimate the parameters with a short code.

```
celcius <- c(24, 25, 25, 27, 24)
fahrenheit <- c(75, 77, 77, 81, 76)

beta1 <- cor(fahrenheit, celcius) * sd(fahrenheit) / sd(celcius)
beta0 <- mean(fahrenheit) - beta1 * mean(celcius)

c(beta0, beta1)
```

```
[1] 31.366667  1.833333
```

You can check that the estimated values are close to the real values:

```
c(beta0, beta1) - c(32, 9 / 5)
```

```
[1] -0.63333333  0.03333333
```