Simple Linear Regression

The Regression Line

Linear regression is one of the most commonly used statistical methods. In many situations, the relationship between two quantitative variables can be summarized by a straight line. The regression line is a model that describes the relationship. Here we'll describe how to find the line that provides the best fit for our observed data.

We can visualize the relationship between two quantitative variables with a scatterplot. Our interest will be in whether or not the relationship in a scatterplot can be summarized by a straight line. The line is an example of a model for the data. We can use the line to make predictions. For example, there is a positive relationship with the size of a house and it's selling price. So, if we know a house has a certain size, we can predict what its selling price will be. Or we can use the model to investigate whether there is a relationship between the variables. For example, is consuming more fibre associated with lower blood pressure?

A line can be completely described by its slope and its intercept. In statistics, the data won't fall perfectly on the line, and we use the line to model the expected or mean value of y. The intercept on the vertical or y-axis, gives the expected value of y when x is equal to 0. The slope tells us how y changes with x on average. A positive slope indicates that as x gets larger, y also increases. A negative slope indicates that an increase in x is, on average, associated with a decrease in y. See Figure 1.

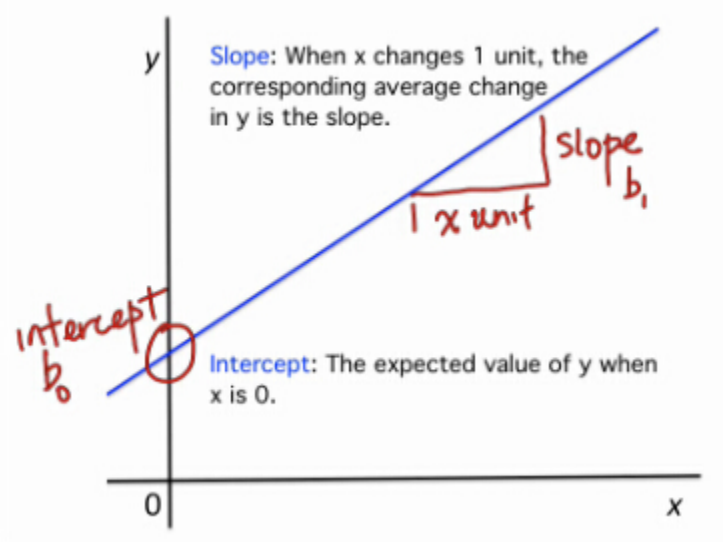

Figure 1: The linear regression line.

In statistics, the regression intercept that we calculate from the data is usually denoted by

b_0, and the slope estimated from the data is denoted by b_1. So the equation of the line is

$$y = b_0 + b_1 x.$$

This is often written as $\hat{y} = b_0 + b_1 x$, where $\hat{y}$ indicates that this is the value of y for the corresponding value of x that is predicted by the line that was fit to the data.

Example 1
A study published in 1993 was interested in how the conditions babies experience affect the development of their motor skills. In particular, is a baby more likely to begin crawling earlier or later, depending on the season in which he or she was born? 414 babies were studied. A scatterplot of the data is given in Figure 2. The scatterplot includes a point for each of the 12 months of the year. On the horizontal or x axis is the average temperature for the month when the babies were 6 months old, which is a typical age when babies first try to crawl, and on the vertical or y axis we have the average age (in weeks) for when the babies could competently crawl. The plot shows a negative relationship. Babies born in warmer months crawled, on average, at younger ages, perhaps because they tended to be bundled in less restrictive clothing.

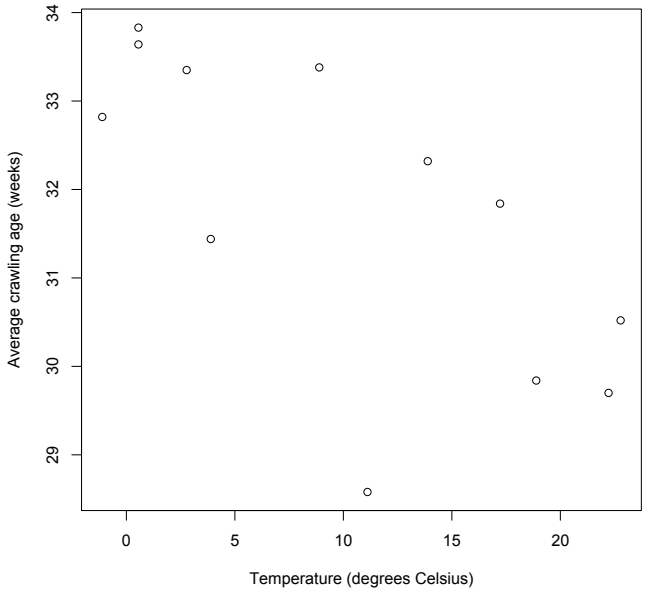

Figure 2: Scatterplot for Example 1 showing average crawling age (in weeks) versus average monthly temperature (degrees Celsius).

In this example, we're trying to see what effect the temperature has on the average crawling age. The average crawling age is our **response variable** or **dependent variable**, because we're investigating whether it depends on temperature. Temperature is called the **indepen-**

dent variable, or **explanatory variable** or **predictor variable**. The response variable is plotted on the y axis, and the explanatory variable is plotted on the x axis.

Our data points are a set of 12 points, (x_i, y_i), for the average temperature in the ith month, and the average crawling age for the babies who turned 6 months old in that month.

We can model the relationship between temperature and crawling age with a straight line. The linear regression line for these data is the line that best predicts the response variable, the crawling age, from the explanatory variable, the temperature. So we want to minimize how far the data points are away from the line in the vertical direction. See Figure 3. Sometimes these vertical distances are above the line so positive, and sometimes they are below, or negative. These deviations are called residuals. We can label these residuals

$$e_i = y_i - \hat{y}_i.$$

We want to minimize the total of these deviations, so we calculate $\sum_{i=1}^{n} e_i^2$. This is called the method of least squares. In the method of least squares, the regression line is found by finding the line that minimizes $\sum_{i=1}^{n} e_i^2$. That is, we solve for the slope and intercept by minimizing $\sum_{i=1}^{n} e_i^2$.

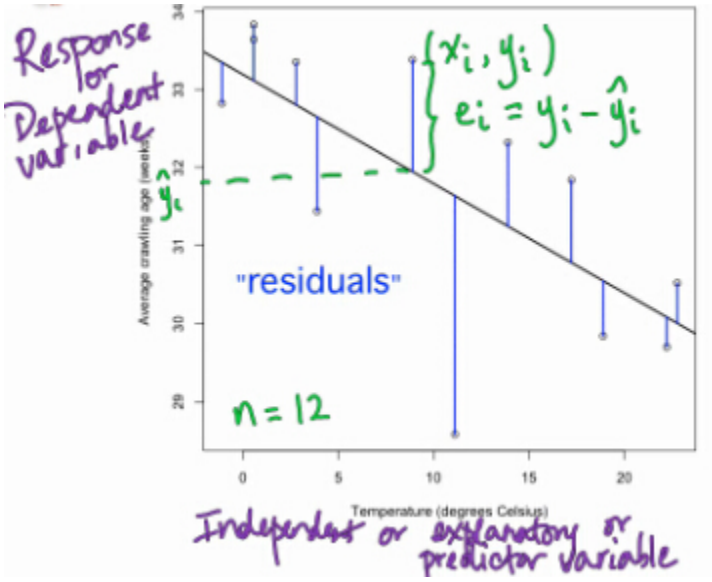

Figure 3: Thre regression line is found by minimizing the sum of the squares of the vertical distances between the line and the data points.

We can calculate the slope and intercept from a few summary statistics: the mean and standard deviations of the response and predictor variables, and the correlation r between the response and predictor variables. r is a value between -1 and 1 that measures the strength of the linear relationship between two quantitative variables. So it is also related to the regression line. For Example 1, these summary statistics are in Table 1.

	Temperature (°C) (x)	Average Crawling Age (weeks) (y)
mean	$\bar{x} = 10.14$	$\bar{y} = 31.77$
sd	$s_x = 8.80$	$s_y = 1.76$
correlation	r=-0.70	

Table 1: Summary statistics for Example 1.

The slope of a regression line is

$$b_1 = \frac{s_y}{s_x} r$$

A negative correlation will result in a negative slope, and a positive correlation will result in a positive slope.

The intercept of the regression line is

$$b_0 = \bar{y} - b_1 \bar{x}$$

Note that the regression line passes through $(\bar{x}, \bar{y})$.

Using these formulae with the summary statistics for the data in Example 1 (Table 1),

$$b_1 = \frac{1.76}{8.80} \times (-0.70) \doteq -0.14$$

$$b_0 = 31.77 - (-0.14) \times 10.14 \doteq 33.19$$

$$\widehat{age} = 33.19 - 0.14 \times temp.$$

From the intercept we know that when the average temperature is 0, we expect the crawling age to be just over 33 weeks. Since the slope is negative, our line estimates on average, for each increase in the temperature, we expect the crawling age to decrease by -0.14 weeks, or almost a day.

To predict the average crawling age for a baby who experiences an average temperature of 15 degrees Celsius, we can find the corresponding value of the response from the regression line,

$$\widehat{age} = 33.19 - 0.14 \times 15 \doteq 31.09.$$

Carrying out a reverse prediction, i.e., to predict the temperature, for, as an example, a baby who crawled at 32 weeks, is not as simple as rearranging our regression line to solve for the predictor variable given a value of the response. To find the regression line we minimized the sum of the vertical deviations, that is we minimized how far our predicted values of crawling age, on the line, were from our observed values of crawling age. However, a line of best fit for predicting temperature from crawling age would minimize the errors in the horizontal direction and we'd get a completely different regression line.

Modelling the relationship between two quantitative variables with a straight line, either for prediction or to better understand the relationship between the variables, involves identifying one of the variables as the response, and the other as its predictor, and using the method of least squares to find the regression line. However, taking a first look at the data with a scatterplot, to see if a straight line model is a reasonable choice, should always be the first step.