Introduction to linear regression

Flora Green

Summary

Linear regression is a statistical model which uses one or more variables to predict the behaviour of another. It is widely used in fields such as business, economics and social sciences.

Before reading this guide, it is strongly recommended that you read [Guide: Introduction to straight lines] and Guide: Expected value, variance, standard deviation.

What is linear regression?

One of the key purposes of statistics is the interpretation of data, which often involves considering the relationships between variables. Typically, this information is split into two categories: 'the thing which is observed' and 'the thing(s) which affect this observation', and a statistician collects information in both categories to build a full picture of the scenario. Then, the statistician tries to explain how and why a certain outcome is reached - which is incredibly useful for predicting which outcome may be reached in the future!

For instance, if the seaside branch of Cantor's Confectionery had 150 ice cream sales in one day, and only 30 in another, they may wish to understand what it was that motivated this change, so that they can have the appropriate number of ice creams in stock. It could be that one of these days was a warm, sunny, summer's Saturday, whilst the other was a cold, rainy winter's Monday. If you take a sample of the weather, season and day of the week and the corresponding number of ice cream sales across multiple days, you can build a model which allows you to predict the number of ice cream sales on a particular day.

One of the models which describes the relationship between variables is known as **linear regression**. Linear regression finds a line of best fit for your data, known as a **regression line**, which can be used to predict outcomes - like the number of ice cream sales.

This guide will introduce you to linear regression. It will explain how to apply a simple linear regression model to a sample of data by finding the regression line. It will also explore the practical applications of linear regression, including in the construction of confidence intervals which encapsulate the uncertainty about how closely your sample matches the real-world scenario, and in testing hypotheses about whether some factors impact the overall outcome.

Simple linear regression

Definition of simple linear regression

Simple linear regression is a linear regression model used when there is only one explanatory variable, x. The **regression line** predicts the value of the response variable, which is expressed as $\mathbb{E}(Y)$.
For more on expected values, please see Guide: Expected value, variance, standard deviation. For more about explanatory and response variables, please see Overview: Statistical Notation.

To find the equation of the regression line for a population of data, you will need to find the y-intercept of the line, which will be denoted by α, and its gradient, which will be described as β.

For more about the equation of a straight line, please read [Guide: Introduction to straight lines].

Therefore the equation of the regression line is given by:

$$\mathbb{E}(Y) = \alpha + \beta x$$

In practice, it is impractical to work with an entire population of data - instead, statisticians analyze samples. This means that you will need to obtain **estimates** for the values of α and β from your sample of data, and these will be denoted by $\hat{\alpha}$ and $\hat{\beta}$.

Tip

$\hat{\alpha}$ and $\hat{\beta}$ are **unbiased** estimators of the population **regression parameters** α and β, which means that their long-term averages are the same.
For more about biased estimators, please read [Guide: Biased and unbiased estimators].

Example 1

The seaside branch of Cantor's Confectionery wishes to see the relationship between temperature (x) and ice cream sales (y). To do this, they track the number of ice cream sales in a 10 day period alongside the peak temperature that day.

Day number	Peak temperature (°C)	Number of sales
1	22	150
2	20	100
3	19	110
4	21	210
5	24	260
6	25	280
7	27	310
8	26	350
9	28	360
10	25	270

You can define Y to be the random variable associated with observations x. Then, you can plot the graph of the simple linear regression model:

$$\mathbb{E}(Y) = \alpha + \beta x$$

The graphs below illustrate how changing the values of the **regression parameters** affects how closely the **simple linear regression model** fits the observed data.

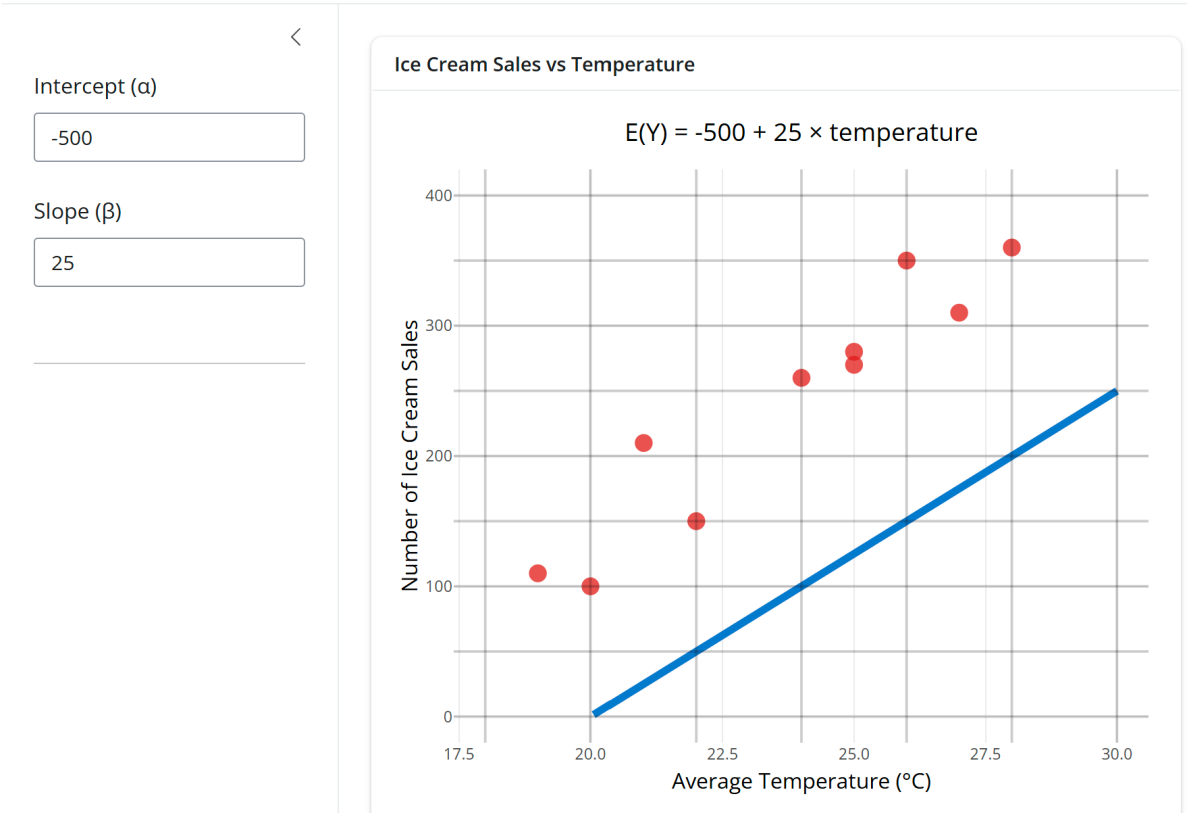

Figure 1: A simple linear regression model for ice cream sales vs. temperature, with parameters $\alpha = -500$ and $\beta = 25$

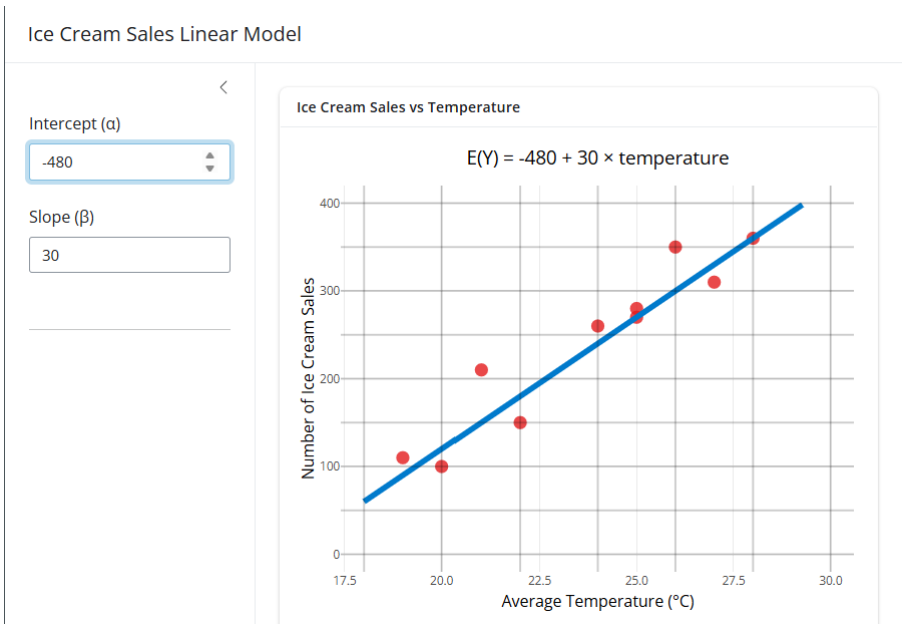

Figure 2: A simple linear regression model for ice cream sales vs. temperature, with parameters $\alpha = -480$ and $\beta = 30$

You can see that the line in Figure 2 is much closer to the data points than the line in Figure 1, meaning the equation in Figure 2 is a better fitted linear model.

Tip

There are $n \in \mathbb{N}$ observations in any sample of data.
If you consider $i \in \mathbb{Z}$ such that $1 \leq i \leq n$, the i^{th} data point on the graph of your data is given by (x_i, y_i) and its corresponding estimated value is given by:

$$\mathbb{E}(Y_i) = \alpha + \beta x_i$$

In practice, it is not always possible to plot the data points on a set of axes and manually tweak the values of the regression parameters until they appear to match the data - especially when you are working with a large sample of data. A quantitative method is a far more practical, so you will now learn how to use **residuals** in the **method of least squares estimation** to find the optimal regression line for your data.

Least squares estimation

Definition of residuals

The **residual** describes the difference between the **observed** value y_i and the **estimated** value $\mathbb{E}(Y_i)$.
Each **residual** is denoted e_i, where $i \in \mathbb{Z}$ such that $1 \leq i \leq n$.
So, using the equation for $\mathbb{E}(Y_i)$, you can define the residuals by:

$$e_i = y_i - E(Y_i) = y_i - (\alpha + \beta x_i)$$

Tip

When plotted on a graph, the residual can be interpreted as the vertical difference between the line of best fit and the plotted data point.

You want $\mathbb{E}(Y_i)$ to be as close to your observed data as closely as possible, so you want the difference between your **observed** data and the data **estimated** by the model to be as small as possible. To find the values of the regression parameters which achieve this, you can use the method of least squares estimation.

Least squares estimation

Least squares estimation is used to minimize the **sum of the squares of the residuals**, which is defined as $S(\alpha, \beta)$.
You consider the squares of the residuals (e_i^2) to ensure that the summed values are always positive.
So, considering the sum of each e_i^2, you will find:

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

The **method of least squares** minimizes this function to find **estimates** (sometimes known as **point estimates**) for the **regression parameters**. These are denoted by $\hat{\alpha}$ and $\hat{\beta}$, and these estimates generate a line of best fit where the average difference between the **observed** data and that **estimated** by the model is as small as possible.

Important

To find estimates for the regression parameters, you need to understand the following terminology:
$\bar{x}$ and $\bar{y}$ denote the **sample means** of x and y.
The sum of the squared differences between each observation x_i and the sample mean $\bar{x}$ is given by:

$$SS_{XX} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

The sum of the differences between each observation x_i and the sample mean $\bar{x}$, multiplied by the differences between each observation y_i and the sample mean $\bar{y}$ is described by:

$$SS_{XY} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

For more detail about these terms, please read Overview: Statistical notation.

Finding estimates of the regression parameters

You can differentiate $S(\alpha, \beta)$ with respect to the parameters α and β and set these equal to 0 to find the point at which there is no change in the value of the parameters. These are the estimates for the parameters. For more about derivatives, please read Guide: Introduction to differentiation.

The **least squares estimate** for the regression parameter β is defined as:

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

The **least squares estimate** for the regression parameter α is defined as:

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

For proofs of these results, please read [Proof sheet: Deriving estimates for regression parameters].

Tip

When working with sums, you use programming languages such as R to save time manually typing each value into a calculator!

Example 2

The graphs below demonstrate how changing the values of the **regression parameters** affects how close the estimated values $\mathbb{E}(Y_i)$ are to the observed values y_i. You can see that the residuals are the vertical differences between the observed and expected values.

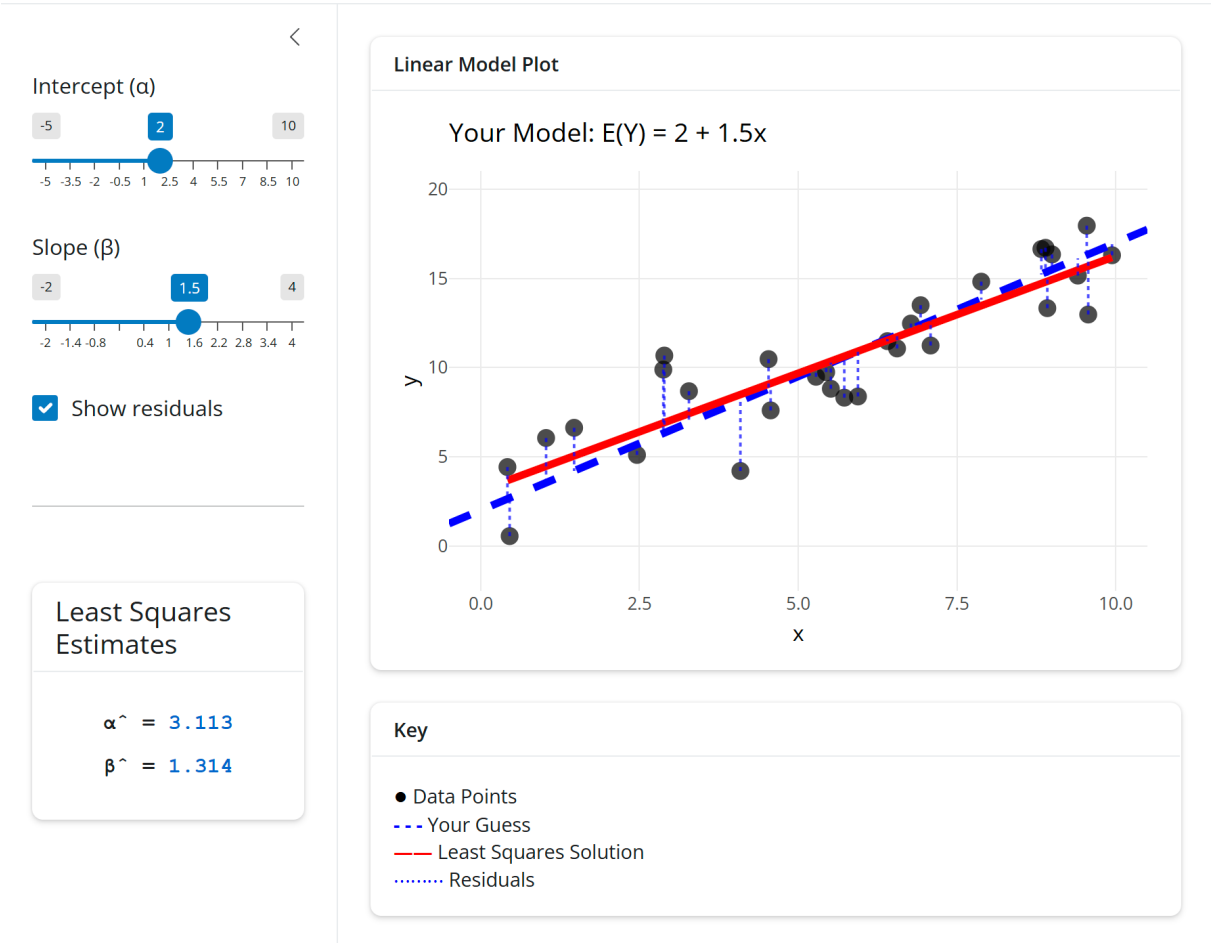

Figure 3: A regression line with parameters $\alpha = 2$ and $\beta = 1.5$ plotted alongside the regression line with parameters $\hat{\alpha}$ and $\hat{\beta}$.

You can see that the regression line which uses the least squares estimates fits the data the closest.

> **Tip**
>
> When working with sums, you use programming languages such as R to save time manually typing each value into a calculator.

Example 3

In Example 1, you saw the following data from Cantor's Confectionery.

Day number	Peak temperature (°C)	Number of sales
1	22	150
2	20	100
3	19	110
4	21	210
5	24	260
6	25	280
7	27	310
8	26	350
9	28	360
10	25	270

You defined temperature as x and the number of ice cream sales as y.
You saw how changing the values of the **regression parameters** α and β affected how closely the **simple linear regression model** $\mathbb{E}(Y) = \alpha + \beta x$ fitted the observed data.
Now, you will use **least squares estimation** to estimate the values $\hat{\alpha}$ and $\hat{\beta}$ which best fit the data.
You can see that there are $n = 10$ entries in the table. You can use the definition of sample means - as seen in Overview: Statistical notation - to obtain values of $\bar{x}$ and $\bar{y}$.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}\sum_{i=1}^{n} x_i = 23.7$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{10}\sum_{i=1}^{n} y_i = 240$$

Use these sample means to calculate SS_{XX} and SS_{XY}:

$$SS_{XY} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 2460$$

$$SS_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = 84.1$$

As you saw earlier, the observations $Y_1, ..., Y_n$ are random variables and can therefore be assumed to follow a distribution. Since the observations are functions of the **regression parameters**, the least squares estimates for the regression parameters are also random variables.

For more information about this, please read [Guide: Maximum likelihood estimation].

Therefore, you can assume that the estimates follow a distribution.

For more information about the types of probability distributions which your data could follow, please read Overview: Probability distributions.

> **!** Important
>
> Before you can assume that your data follows a particular distribution, you must check whether your data aligns with the assumptions which underpin this distribution.

It is useful for your regression parameters to follow probability distributions because it allows you to construct confidence intervals and test hypotheses about them.

Constructing confidence intervals for the regression parameters enables you to construct a plausible range of values, within which it is likely that your estimation lies. This recognizes the uncertainty as to how closely your sample matches the real-life scenario.

For more information about confidence intervals, please read Guide: Confidence intervals.

To learn how to apply these ideas to data which follows a **normal distribution**, please read [Guide: Normal linear regression.]

Testing hypotheses is particularly useful when there are multiple explanatory variables, because you can test whether some of them are equal to zero - and if so, they can be discounted. This reduces the complexity of the linear regression model, meaning it is more transferable to new data.

For more on this, please read [Guide: Introduction to Multiple Linear Regression].

Quick check problems

1. Which of these is the equation of a linear regression model?

(a) $\mathbb{E}(Y) = \alpha + \beta x_i$

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

(c) $\hat{\beta} = \frac{SS_{XY}}{SS_{XX}}$

2. What is the name of the linear regression model which has one explanatory variable?

3. Are the following statements true or false?

(a) e_i is used to describe a residual.

(b) To use linear regression, your data must be normally distributed.

(c) The regression parameter α can be equal to zero.

Further reading

For more questions on the subject, please go to Questions: Introduction to linear regression.

To learn about normal linear regression, please read [Guide: Normal linear regression].

To extend linear regression to encompass multiple explanatory variables, please read [Guide: Multiple linear regression].

Version history

v1.0: initial version created 12/25 by Flora Green as part of a University of St Andrews VIP project.