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 recommended that you read Guide: Expected value, variance, standard deviation and Factsheet: Overview of statistical notation.

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 and 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.

By collecting information about all of these factors, it is possible to build a model which predicts how many ice cream sales there will be on a particular day, based on any combination of factors. This relationship between the factors which affect an outcome, and the outcome itself, is known as regression, and in this guide, you will see how one type of regression - namely linear regression - works.

Simple linear regression

Definition of simple linear regression

Simple linear regression is a form of linear regression. The function

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

relates the explanatory variable, x, to the response variable, Y. α and β are known as **regression parameters**.

For more on expected values, please see Guide: Expected value, variance, standard deviation. For a reminder about explanatory and response variables, please see Overview: Statistical Notation.

Tip

When the population regression line is plotted on a set of axes, it forms a straight line. This is where the name **linear** regression comes from. α represents the y-intercept and β represents the gradient.

Each individual point is described as (x_i, y_i), where $i \in [1, n]$.

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 average temperature that day.

Day number	Average 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 plot the graph fit the simple linear regression model

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

Use the figures below to observe how changing the values of the **regression parameters** affects how closely the **simple linear regression model** fits the observed data.

Setting up the linear model

It is usually impractical to work with the entire population of data, so statisticians typically take samples of data to work with. For more on sampling data, please see [Factsheet: Sampling data].

With each x_i in your sample of observed data, you can associate the random variable Y_i. Then, the linear regression line becomes:

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

Definition of the error term

The **error term** describes the difference between your **estimated** value of Y_i and the **actual** value of Y_i.
Each **error term** is denoted ϵ_i, where $i \in [1, n]$. These error terms have sample mean $= 0$ and variance $= \sigma^2$.

The random variable Y_i can therefore be expressed as:

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

This is called a **linear model**.

To fit your **linear model** to your data, you will need to estimate the **regression parameters**.

Since you have a collection of random variables, you can assume that each $Y_1, ..., Y_n$ follow a distribution - which will allow you to estimate the **regression parameters**.

Warning

Before you can assume that your data follows a distribution, you need to check that your data follows the assumptions which underpin the distribution.

You want your linear model to match your observed data as closely as possible, so you want the difference between the **observed** data and that **estimated** by the model to be as small as possible.

Definition of residuals

A **residual** is the vertical difference between the observed value and the estimated value.
By rearranging the equation for the linear model, you can see that for $i \in [1, n]$, the residual is:

$$\epsilon_i = y_i - (\alpha + \beta x_i)$$

Least squares estimation

Least squares estimation is used to minimize the **sum of the squares of the residuals**.
Considering the sum of each ϵ_i^2, you will find:

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

The **method of least squares** minimizes this function to find estimates for the **regression parameters**, which are denoted by $\hat{\alpha}$ and $\hat{\beta}$.

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

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

As seen in Factsheet: Overview of statistical notation, $\bar{y}$ denotes the **sample mean** of y.

Example 2

The figures below explore the effects of changing the values of the **regression parameters** on **least squares estimation** in more detail.

Example 3

Find estimates for regression parameters mathematically.

Normal linear regression

One of the distributions which your data could follow is the **normal distribution**. For a reminder about the normal distribution, please see Factsheet: The normal distribution.

Suppose that the variables $Y_1, ..., Y_n$ are normally distributed with mean $\alpha + \beta x_i$ and variance σ^2. In other words:

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2)$$

Assumptions for normality

You must check that your data can be modelled using the normal distribution before you can use this model.

a) **Independence** - $Y_1, ..., Y_n$ are independent random variables.

b) **Normality** - $Y_1, ..., Y_n$ are normally distributed.

c) **Linearity** - the mean of Y_i is a linear function of x_i, or equivalently, $\mathbb{E}(Y) = \alpha + \beta x_i$

d) **Constant variance** - $Y_1, ..., Y_n$ have the same variance.

Example 4

Cantor's Confectionery wants to predict the number of ice cream sales each day y_i based on the daily temperature x_i using a normal linear regression model. Based on the data below, does the assumption of normality seem acceptable?

a) Since each data point comes from a different day, the assumption of independence is reasonable.

b)

c) By plotting this on a graph, you can see that the line of best fit is roughly linear.

d)

Fitting the normal linear regression model

Likelihood function

To estimate the values of α and β, you can use the log-likelihood function:

$$l(\alpha, \beta, \sigma^2; (x_1, y_1), \dots, (x_n, y_n)) = -\frac{n}{2}\log(2\pi\sigma^2) \; - \; \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2$$

If you maximize this with respect to the regression parameters α and β, this is equivalent to minimizing $S(\alpha, \beta)$.

For more on likelihood functions, please go to [Study guide: Likelihood functions].

For a derivation of this equation, please go to Proof sheet: Linear regression - log-likelihood.

Confidence intervals

It can be useful to find $1-\alpha\%$ confidence intervals for the regression parameters α and β.

Confidence intervals of regression parameters

The confidence interval for α is given by:

$$\hat{\alpha} \pm t_{n-2;1-\frac{\alpha}{2}} \sqrt{S^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SS_{XX}} \right)}$$

The confidence interval for β is given by:

$$\hat{\beta} \pm t_{n-2;1-\frac{\alpha}{2}} \sqrt{\frac{S^2}{SS_{XX}}}$$

For more on confidence intervals, please read Proof sheet: Confidence intervals

Example 5

find a confidence intervals

Quick check problems

1. Which of these statements must be true for you to use use a linear regression model on your data?

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

(b) All x_i in the data set are independent.

(c) σ^2 is small.

2. Name the type of linear regression with one explanatory variable.

3. Which type of distribution is used when finding confidence intervals?

4. Are the following statements true or false?

(a) ϵ is used to describe an error term.

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

(c) The regression parameter α must not be zero.

Further reading

Version history

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