Linear Regression Definition

The most basic linear model is a ***simple linear regression***, which looks like a basic *y* = *mx* + *b* best-fit line.

Independent and Dependent Variable

Independent variables are also known as: predictor variables, input variables, explanatory variables, features. The values on the x-axis.

Dependent variables are also known as: outcome variables, target variables, response variables. The value on the y-axis

An Example of independent and Dependent Variables

If we are looking at how profit is affected by sales. We will have sales as an independent variable and profit as the dependent variable.

In another situation where we are looking at how sales are affected by advertisements, Sales will be the dependent and advertisements is the independent variable.

Statistical Model Definition

A statistical model can be thought of as some kind of a transformation that helps us express dependent variables as a function of one or more independent variables.

A statistical model defines a **relationship** between a dependent and an independent variable.

The goal of statistical modeling is to find the best fitting line that represents the relationship between the dependent and independent variable allowing predictions or insights into how changes in the predictors (independent variables) affect the response (dependent variables).

We can define and **fit** such a straight line to our data following a straight line equation:

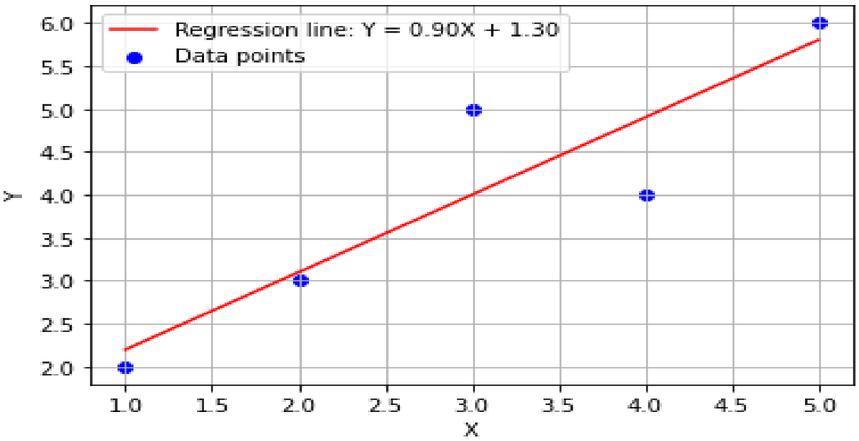
y=mx+c

x: the dependent variable

m: is the slope of the line

c: is where the line intercepts the y-axis

C = y_mean - m*x_mean

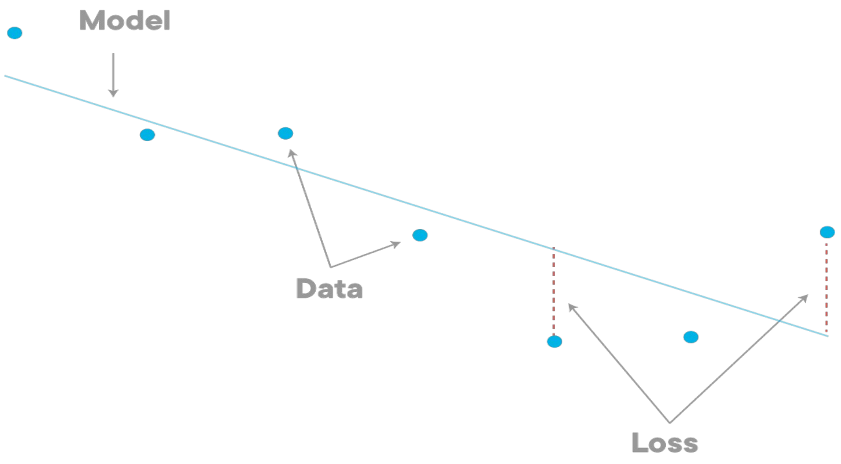


Model Loss

A loss function evaluates how well your model represents the relationship between data variables.

If the model is unable to identify the underlying relationship between the independent and dependent variable(s), the loss function will output a very high number.

These individual losses, which is essentially the vertical distance between the individual data points and the line are taken into account to calculate the overall model loss.



Linear Regression and T-Test

For example, the null hypothesis might be that μ1=μ2 while the alternative hypothesis might be μ1≠μ2.

In the context of linear regression, the t-test that we use is for whether a given **coefficient** is equal to **zero**.

The null hypothesis is that **β=0** and alternative hypothesis is that **β≠0**.

<u>Regression Evaluation</u>

<u>Intercept Model</u>

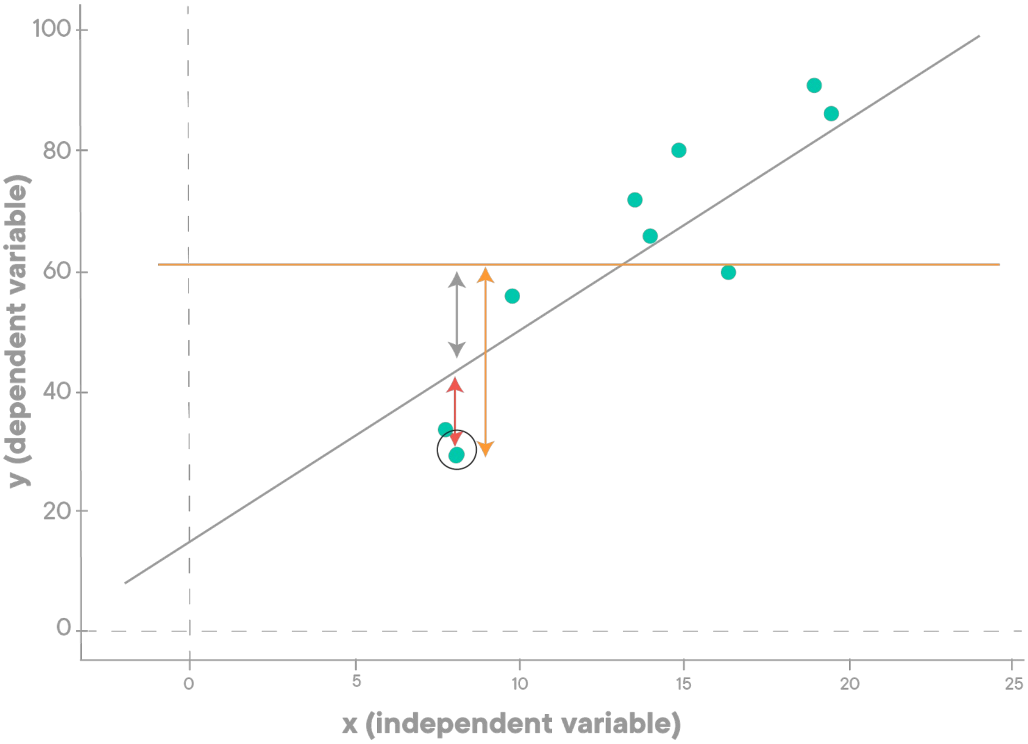
There is no universal, objective standard for evaluating models, because every set of x and y values is different.

Therefore we measure the fit of our models against a baseline model known as the ***intercept-only model***.

Instead, it uses the mean of the observed responses of the dependent variable y and always predicts this mean as the value of y for any value of x.

For all x values the intercept model has y_values = np.mean(y_values)

We will determine this answer in terms of both statistical significance and goodness-of-fit by comparing the the ***errors*** made by our model (red vertical arrow) to the errors made by the intercept-only model (orange vertical arrow) and determining the difference between them (gray vertical arrow).



F-Test for Statistical Significance

So, is our overall model statistically significant? Let's frame this in terms of a null and alternative hypothesis:

- H0 (null hypothesis): the intercept-only model fits the data just as well as (or better than) our model

- Ha (alternative hypothesis): our model fits the data better than the intercept-only model

F-statistic measures whether the regression model provides a better fit than a model with no predictors (just the mean).

A high F-statistic (and low p-value, typically < 0.05) indicates that the model is significant.

The p-value helps determine statistical significance. If $p<0.05$, we reject the null hypothesis that all coefficients are zero.

We use the stats.model library to compute the F-statistic

Measuring Goodness of Fit with R-Squared

Now that we know that our model is statistically significant, we can go one step further and quantify how much of the variation in the dependent variable is explained by our model. This measure is called the R2 or ***coefficient of determination***.

For a least-squares regression model, R-Squared can take a value between 0 and 1 where values closer to 0 represent a poor fit and values closer to 1 represent an (almost) perfect fit

An R-Squared of 1 means that you are explaining 100% of the variation. This is very unlikely to see with linear regression on real-world data, so you probably want to double-check that you have set up your variables properly.