LINEAR REGRESSION

Linear Regression in Machine Learning

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Mathematically, we can represent a linear regression as:

y= a0+a1x+ ε

**Here,**

Y= Dependent Variable (Target Variable)  
X= Independent Variable (predictor Variable)  
a0= intercept of the line (Gives an additional degree of freedom)  
a1 = Linear regression coefficient (scale factor to each input value).  
ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

Types of Linear Regression

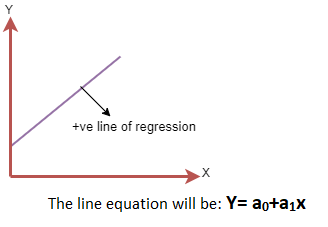
Linear regression can be further divided into two types of the algorithm:

* **Simple Linear Regression:**  
  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
* **Multiple Linear regression:**  
  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

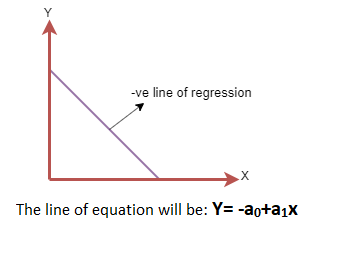
Linear Regression Line

A linear line showing the relationship between the dependent and independent variables is called a **regression line**. A regression line can show two types of relationship:

* **Positive Linear Relationship:**  
  If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.



* **Negative Linear Relationship:**  
  If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.

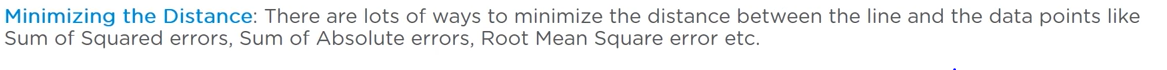


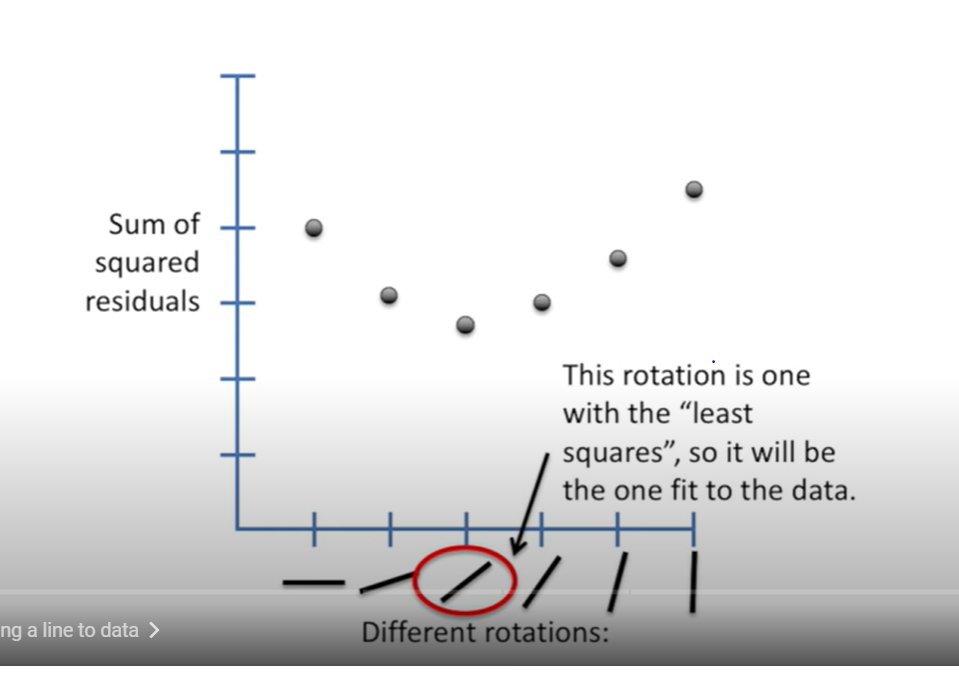
Finding the best fit line:

When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

**Residuals:** The distance between the actual value and predicted values is called residual. If the observed points are far from the regression line, then the residual will be high, and so cost function will high. If the scatter points are close to the regression line, then the residual will be small and hence the cost function.

**The Best Fit line will have the least value for the “sum of squares of the Residuals” among all the other possible lines (see fig below).**





The different values for weights or the coefficient of lines (a0, a1) gives a different line of regression, so we need to calculate the best values for a0 and a1 to find the best fit line, so to calculate this we use cost function.

Cost function-

* The different values for weights or coefficient of lines (a0, a1) gives the different line of regression, and the cost function is used to estimate the values of the coefficient for the best fit line.
* Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing.
* We can use the cost function to find the accuracy of the **mapping function**, which maps the input variable to the output variable. This mapping function is also known as **Hypothesis function**.

For Linear Regression, we use the **Mean Squared Error (MSE)** cost function, which is the average of squared error occurred between the predicted values and actual values. It can be written as:

For the above linear equation, MSE can be calculated as:

Linear Regression in Machine Learning

**Where,**

N=Total number of observation  
Yi = Actual value  
(a1xi+a0) = Predicted value.

**Residuals:** The distance between the actual value and predicted values is called residual. If the observed points are far from the regression line, then the residual will be high, and so cost function will high. If the scatter points are close to the regression line, then the residual will be small and hence the cost function.

Gradient Descent:

* Gradient descent is used to minimize the MSE by calculating the gradient of the cost function.
* A regression model uses gradient descent to update the coefficients of the line by reducing the cost function.
* It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

Model Performance:

The Goodness of fit determines how the line of regression fits the set of observations. The process of finding the best model out of various models is called **optimization**. It can be achieved by below method:

**1. R-squared method:**

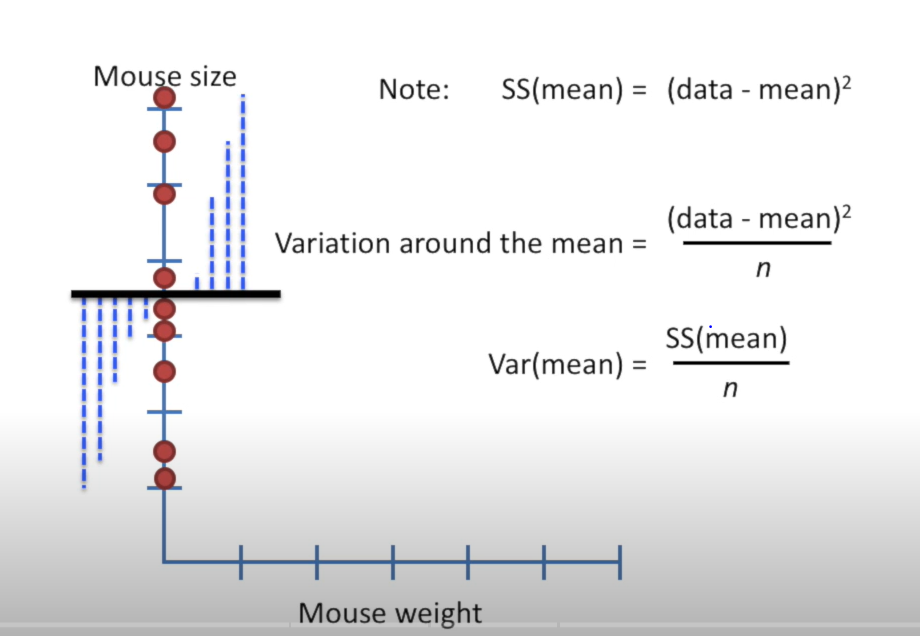
* R-squared is a statistical method that determines the goodness of fit.
* It measures the strength of the relationship between the dependent and independent variables on a scale of 0-100%.
* **The high value of R-square determines the less difference between the predicted values and actual values and hence represents a good model.**
* It is also called a **coefficient of determination,** or **coefficient of multiple determination** for multiple regression.
* It can be calculated from the below formula:

Linear Regression in Machine Learning

**Terminology :**

1. SS(mean) = Sum of squares around the mean
2. Variance (see fig below).

**NOTE**: We can see that Variation(mean) is directly proportional to SS (mean)



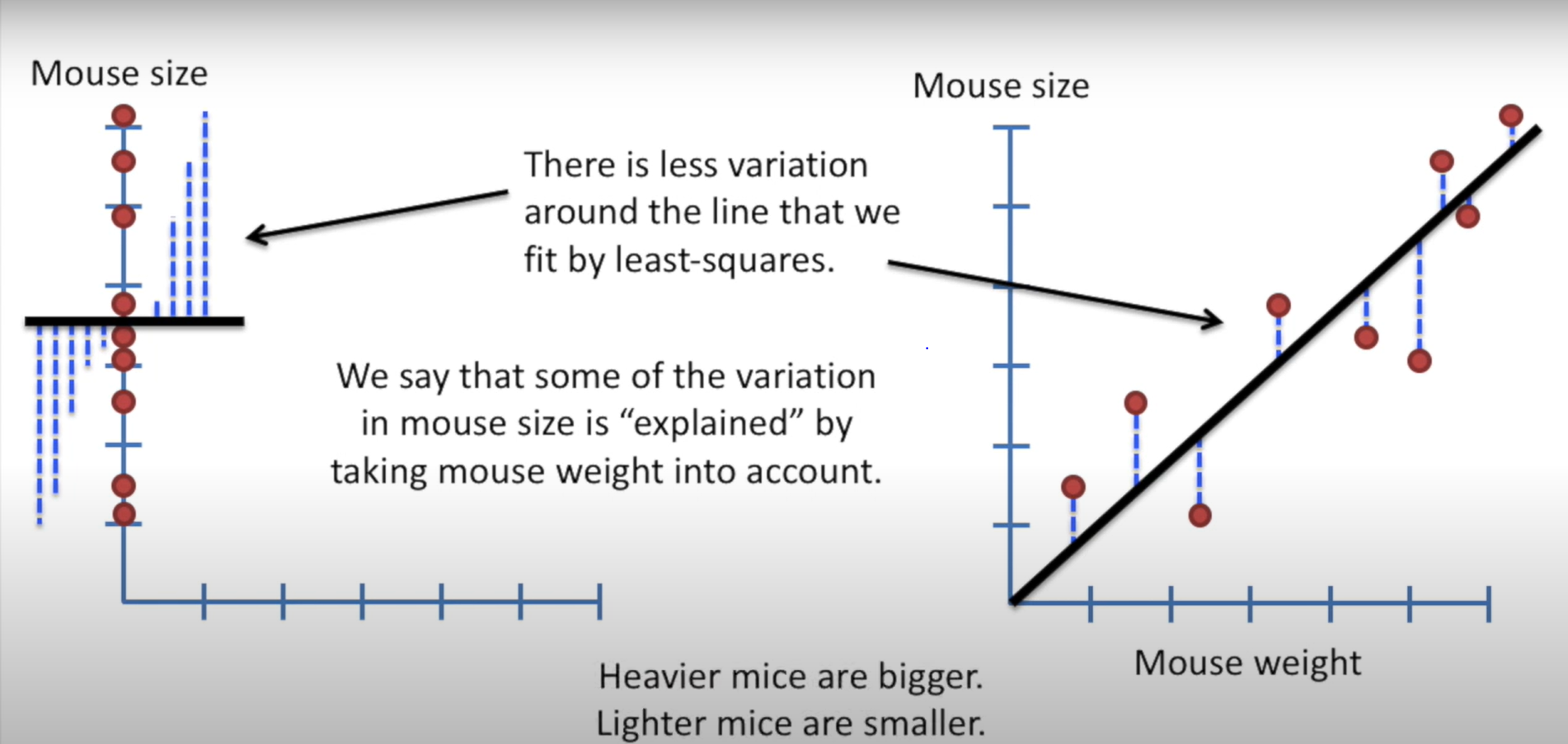
**Consider an example of mice, where we have data for two features given as**

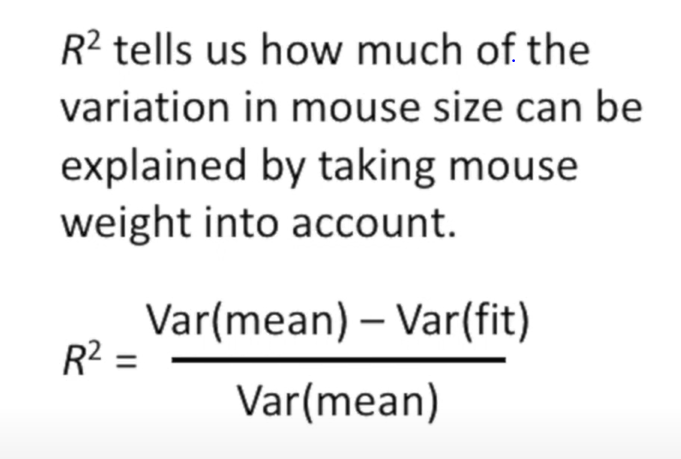
**I: Mouse size**

**II: Mouse weight**

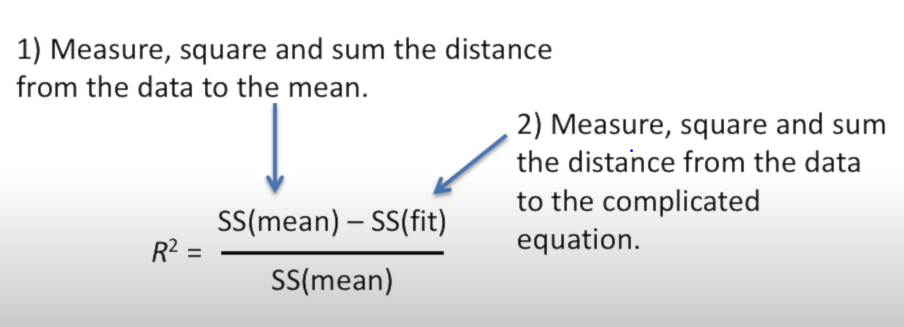
We plot the best fit line (using least squares method) to get the RHS figure below. And we calculate its Variance using actual points (see red dots on RHS figure) and corresponding value from the best fit line, we get Var(fit) = Var(along the best fit line) = 4.4

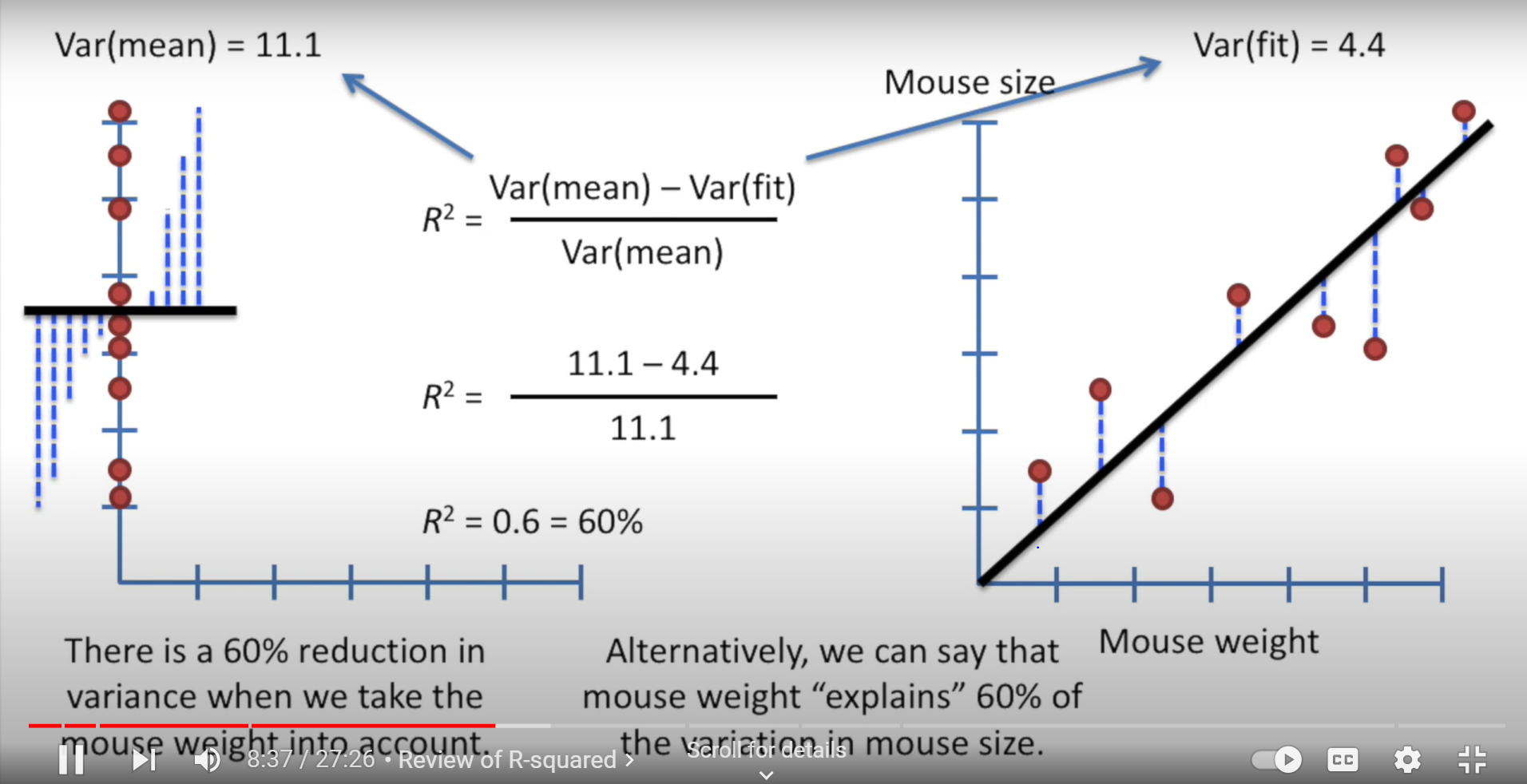
Now, we plot the values of Mouse Size(only) and plot it on a vertical axis (see LHS figure) and get its Variance using these values and the mean of these value (shown in black line in LHS figure) , we get Var(mean) = Var( along the mean for Size values) = 11.1

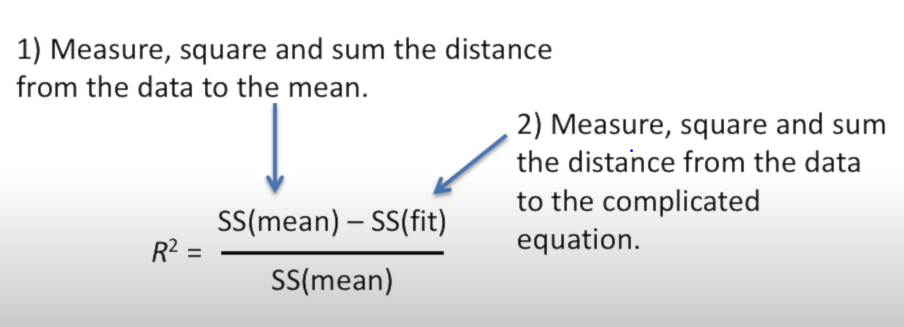




NOTE:

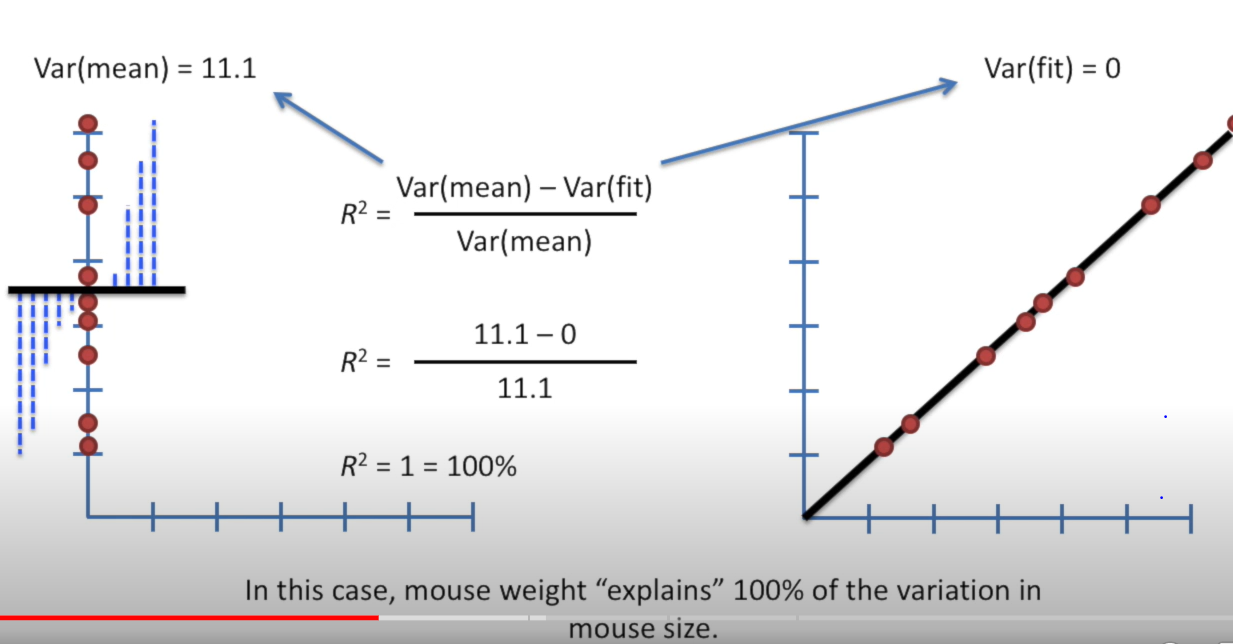
1. We can use SS in place of Var and modify the above formula, because , Variation is directly proportional to SS . 





**ANOTHER EXAMPLE**:

Say in above mouse example, we have data with Variance (for best fit line) = 0 and with Variance (along the mean of size) = 11.1



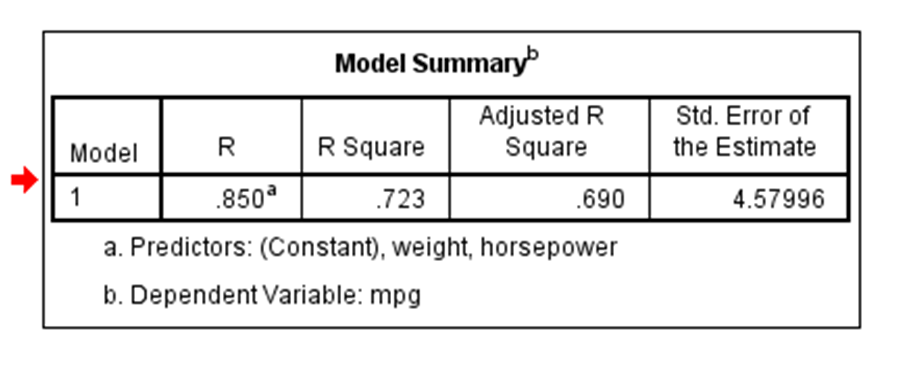
In this case, Mouse weight explains 100% of Variation in Mouse size.

**What is the difference between coefficient of determination (R squared), and coefficient of correlation(R)?**

Coefficient of correlation is “R” value which is given in the summary table in the Regression output. R square is also called coefficient of determination. Multiply R times R to get the R square value. In other words, Coefficient of Determination is the square of Coefficeint of Correlation.

R square or coeff. of determination shows percentage variation in y which is explained by all the x variables together. Higher the better. It is always between 0 and 1. It can never be negative – since it is a squared value.

 It is easy to explain the R square in terms of regression. It is not so easy to explain the R in terms of regression.

[](https://blog.uwgb.edu/bansalg/files/2013/11/Regression-model-summary-table.png)Coefficient of Correlation is the R value i.e. .850 (or 85%). Coefficient of Determination is the R square value i.e. .723 (or 72.3%). R square is simply square of R i.e. R times R.

 Coefficient of Correlation: is the degree of relationship between two variables say x and y. It can go between -1 and 1.  1 indicates that the two variables are moving in unison. They rise and fall together and have perfect correlation. -1 means that the two variables are in perfect opposites. One goes up and other goes down, in perfect negative way. Any two variables in this universe can be argued to have a correlation value. If they are not correlated then the correlation value can still be computed which would be 0. The correlation value always lies between -1 and 1 (going thru 0 – which means no correlation at all – perfectly not related). Correlation can be rightfully explalined for simple linear regression – because you only have one x and one y variable. For multiple linear regression R is computed, but then it is difficult to explain because we have multiple variables invovled here. That’s why R square is a better term. You can explain R square for both simple linear regressions and also for multiple linear regressions.

Assumptions of Linear Regression

Below are some important assumptions of Linear Regression. These are some formal checks while building a Linear Regression model, which ensures to get the best possible result from the given dataset.

* **Linear relationship between the features and target:**  
  Linear regression assumes the linear relationship between the dependent and independent variables.
* **Small or no multicollinearity between the features:**  
  Multicollinearity means high-correlation between the independent variables. Due to multicollinearity, it may difficult to find the true relationship between the predictors and target variables. Or we can say, it is difficult to determine which predictor variable is affecting the target variable and which is not. So, the model assumes either little or no multicollinearity between the features or independent variables.
* **Homoscedasticity Assumption:**  
  Homoscedasticity is a situation when the error term is the same for all the values of independent variables. With homoscedasticity, there should be no clear pattern distribution of data in the scatter plot.
* **Normal distribution of error terms:**  
  Linear regression assumes that the error term should follow the normal distribution pattern. If error terms are not normally distributed, then confidence intervals will become either too wide or too narrow, which may cause difficulties in finding coefficients.  
  It can be checked using the **q-q plot**. If the plot shows a straight line without any deviation, which means the error is normally distributed.
* **No autocorrelations:**  
  The linear regression model assumes no autocorrelation in error terms. If there will be any correlation in the error term, then it will drastically reduce the accuracy of the model. Autocorrelation usually occurs if there is a dependency between residual errors.

**Hypothesis Testing and p-value test**

p-values are used in [hypothesis testing](https://www.scribbr.com/statistics/hypothesis-testing/) to help decide whether to reject the null hypothesis. The smaller the p-value, the more likely you are to reject the null hypothesis.

**Typically, if p-value is less than 0.05 (p-value<0.05) ---> then, we Reject the null Hypothesis (Ho) and accept the Alternate Hypothesis (H1).**

## What is a null hypothesis?

All statistical tests have a [null hypothesis.](https://www.scribbr.com/statistics/hypothesis-testing/)For most tests, the null hypothesis is that there is no relationship between your variables of interest or that there is no difference among groups .

**Example:** in Regression Model, p-value tests the Null Hypothesis that whether each term’s coefficient is 0 . Say for the Regression Eqn : **Y = m(Xi) + c**; m and c are coeff.

i.e. **Ho: Coefficient of independent variable (i.e Xi’s Coeff. = m) = 0** . and p-value test is used to test this hypothesis and then reject it or to accept it .

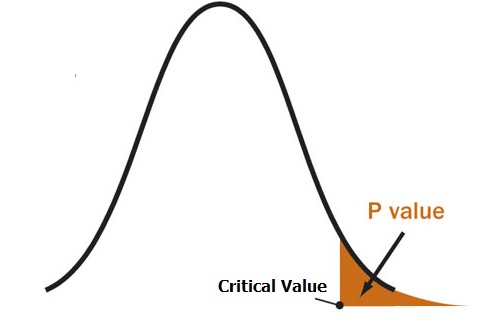
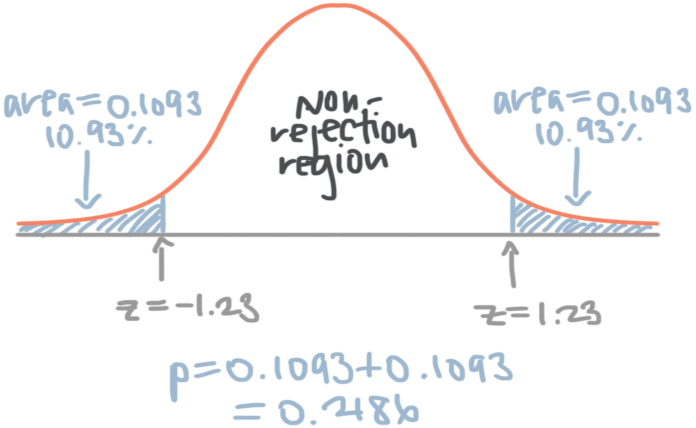
**If,** p-value is low (p-value < 0.05) -- > Reject the Ho -- > Coeff is NOT 0 and must include it (Xi) in our Linear Regression Equation as Coeff.\*Xi .

**Slly if**, p-value is high (p-value > 0.05) -- > Accept the Ho -- > Coeff is 0 and then we can Exclude this (Xi) from our Linear Regression Eqn as 0\*Xi = 0 .

NOTE: Similarly we would need to do Hypothesis testing for intercept value (i.e. c) and get p-value .

**In Other Words,** Predictors (like Xi) which have low p-value (usually<0.05) is likely to be a meaningful addition to your model and which have high p-value(>0.05) suggest that changes in predictor variable is not associated with the changes in the output variable and therefore can be rejected .

Graphically, the p value is the area in the **tail** of a [probability distribution](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/probability-distribution/). It’s calculated when you run hypothesis test and is the area to the right of the [test statistic](https://www.statisticshowto.com/test-statistic/) (if you’re running a [two-tailed test](https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/one-tailed-test-or-two/), it’s the area to the left *and* to the right).

[](https://www.statisticshowto.com/wp-content/uploads/2014/01/p-value1.jpg) 

**p-Value for R-squared (R-squared = Coeff of Determination)**

Recall our Mouse Weight v/s Mouse Size Problem where we got R-squared as 60% i.e. 60% variation in mouse size was explained by mouse weight . This relationship was found using R-sqared test.

BUT, we also need to do a p-value test to find out how reliable this relationship is ? This p-value needs to be small to have a good fit .