

Regression

Regression is a **method to determine the statistical relationship between a dependent variable and one or more independent variables**. The change independent variable is associated with the change in the independent variables. This can be broadly classified into two major types.

Regression models describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line, while logistic and nonlinear regression models use a curved line. Regression allows you to estimate how a **dependent variable** changes as the independent variable(s) change.

Simple linear regression is used to estimate the relationship between two **quantitative variables**. You can use simple linear regression when you want to know:

1. How strong the relationship is between two variables (e.g. the relationship between rainfall and soil erosion).
2. The value of the dependent variable at a certain value of the independent variable (e.g. the amount of soil erosion at a certain level of rainfall).

Example You are a social researcher interested in the relationship between income and happiness. You survey 500 people whose incomes range from 15k to 75k and ask them to rank their happiness on a scale from 1 to 10.

Your independent variable (income) and dependent variable (happiness) are both quantitative, so you can do a regression analysis to see if there is a linear relationship between them.

If you have more than one independent variable, use **multiple linear regression** instead.

Table of contents

Assumptions of simple linear regression

Simple linear regression is a **parametric test**, meaning that it makes certain assumptions about the data. These assumptions are:

1. Homogeneity of variance (homoscedasticity): the size of the error in our prediction doesn't change significantly across the values of the independent variable.
2. Independence of observations: the observations in the dataset were collected using **statistically valid sampling methods**, and there are no hidden relationships among observations.
3. Normality: The data follows a **normal distribution**.

Linear regression makes one additional assumption:

4. The relationship between the independent and dependent variable is linear: the line of best fit through the data points is a straight line (rather than a curve or some sort of grouping factor).

If your data do not meet the assumptions of homoscedasticity or normality, you may be able to use a [nonparametric test](#) instead, such as the Spearman rank test.

Example: Data that doesn't meet the assumptions. You think there is a linear relationship between cured meat consumption and the incidence of colorectal cancer in the U.S. However, you find that much more data has been collected at high rates of meat consumption than at low rates of meat consumption, with the result that there is much more variation in the estimate of cancer rates at the low range than at the high range. Because the data violate the assumption of homoscedasticity, it doesn't work for regression, but you perform a Spearman rank test instead.

If your data violate the assumption of independence of observations (e.g. if observations are repeated over time), you may be able to perform a linear mixed-effects model that accounts for the additional structure in the data.

How to perform a simple linear regression

Simple linear regression formula

The formula for a simple linear regression is:

$$y = \beta_0 + \beta_1 X + \epsilon$$

- **y** is the predicted value of the dependent variable (**y**) for any given value of the independent variable (**x**).
- **B₀** is the **intercept**, the predicted value of **y** when the **x** is 0.
- **B₁** is the regression coefficient – how much we expect **y** to change as **x** increases.
- **x** is the independent variable (the variable we expect is influencing **y**).
- **e** is the **error** of the estimate, or how much variation there is in our estimate of the regression coefficient.

Linear regression finds the line of best fit line through your data by searching for the regression coefficient (**B₁**) that minimizes the total error (**e**) of the model.

While you can perform a linear regression [by hand](#), this is a tedious process, so most people use statistical programs to help them quickly analyze the data.

Application of Simple Linear Regression

Regression analysis is performed to predict the continuous variable. The regression analysis has a wide variety of applications.

Some examples are as follows:

- Predictive Analytics.
- Effectiveness of marketing.
- pricing of any listing.

- promotion prediction for a product.

Multiple Linear Regression:

Regression models are used to describe relationships between variables by fitting a line to the observed data. Regression allows you to estimate how a [dependent variable](#) changes as the independent variable(s) change.

Multiple linear regression is used to estimate the relationship between **two or more independent variables** and **one dependent variable**. You can use multiple linear regression when you want to know:

1. How strong the relationship is between two or more independent variables and one dependent variable (e.g. how rainfall, temperature, and amount of fertilizer added affect crop growth).
2. The value of the dependent variable at a certain value of the independent variables (e.g. the expected yield of a crop at certain levels of rainfall, temperature, and fertilizer addition).

Example You are a public health researcher interested in social factors that influence heart disease. You survey 500 towns and gather data on the percentage of people in each town who smoke, the percentage of people in each town who bike to work, and the percentage of people in each town who have heart disease.

Because you have two independent variables and one dependent variable, and all your variables are quantitative, you can use multiple linear regression to analyze the relationship between them.

Assumptions of multiple linear regression

Multiple linear regression makes all of the same assumptions as [simple linear regression](#):

Homogeneity of variance (homoscedasticity): the size of the error in our prediction doesn't change significantly across the values of the independent variable.

Independence of observations: the observations in the dataset were collected using statistically valid methods, and there are no hidden relationships among variables.

In multiple linear regression, it is possible that some of the independent variables are actually correlated with one another, so it is important to check these before developing the regression model. If two independent variables are too highly correlated ($r^2 > \sim 0.6$), then only one of them should be used in the regression model.

Normality: The data follows a [normal distribution](#).

Linearity: the line of best fit through the data points is a straight line, rather than a curve or some sort of grouping factor.

How to perform a multiple linear regression

Multiple linear regression formula

The formula for a multiple linear regression is:

$$y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n + \epsilon$$

- y = the predicted value of the dependent variable
- B_0 = the y-intercept (value of y when all other parameters are set to 0)
- $B_1 X_1$ = the regression coefficient (B_1) of the first independent variable (X_1) (a.k.a. the effect that increasing the value of the independent variable has on the predicted y value)
- \dots = do the same for however many independent variables you are testing
- $B_n X_n$ = the regression coefficient of the last independent variable
- ϵ = model error (a.k.a. how much variation there is in our estimate of y)

To find the best-fit line for each independent variable, multiple linear regression calculates three things:

- The regression coefficients that lead to the smallest overall model error.
- The t -statistic of the overall model.
- The associated [*p*-value](#) (how likely it is that the t -statistic would have occurred by chance if the [*null hypothesis*](#) of no relationship between the independent and dependent variables was true).

It then calculates the t -statistic and p -value for each regression coefficient in the model.

Multiple linear regression in R

While it is possible to do multiple linear regression by hand, it is much more commonly done via statistical software. We are going to use R for our examples because it is free, powerful, and widely available. Download the sample dataset to try it yourself.

Dataset for multiple linear regression (.csv)

Load the heart.data dataset into your R environment and run the following code:

R code for multiple linear regression
`heart.disease.lm<-lm(heart.disease ~ biking + smoking, data = heart.data)`

This code takes the data set `heart.data` and calculates the effect that the independent variables `biking` and `smoking` have on the dependent variable `heart disease` using the equation for the linear model: `lm()`.

The summary first prints out the formula ('Call'), then the model residuals ('Residuals'). If the residuals are roughly centered around zero and with similar spread on either side, as these do (median 0.03, and min and max around -2 and 2) then the model probably fits the assumption of heteroscedasticity.

Next are the regression coefficients of the model ('Coefficients'). Row 1 of the coefficients table is labeled (Intercept) – this is the y-intercept of the regression equation. It's helpful to

know the estimated intercept in order to plug it into the regression equation and predict values of the dependent variable:

$$\text{heart disease} = 15 + (-0.2 * \text{biking}) + (0.178 * \text{smoking}) \pm e$$

The most important things to note in this output table are the next two tables – the estimates for the independent variables.

The Estimate column is the estimated **effect**, also called the **regression coefficient** or r^2 value. The estimates in the table tell us that for every one percent increase in biking to work there is an associated 0.2 percent decrease in heart disease, and that for every one percent increase in smoking there is an associated .17 percent increase in heart disease.

The Std.error column displays the **standard error** of the estimate. This number shows how much variation there is around the estimates of the regression coefficient.

The t value column displays the **test statistic**. Unless otherwise specified, the test statistic used in linear regression is the t -value from a two-sided **t-test**. The larger the test statistic, the less likely it is that the results occurred by chance.

The $\text{Pr}(> | t |)$ column shows the **p-value**. This shows how likely the calculated t -value would have occurred by chance if the null hypothesis of no effect of the parameter were true.

Because these values are so low ($p < 0.001$ in both cases), we can **reject the null hypothesis** and conclude that both biking to work and smoking both likely influence rates of heart disease.

Presenting the results

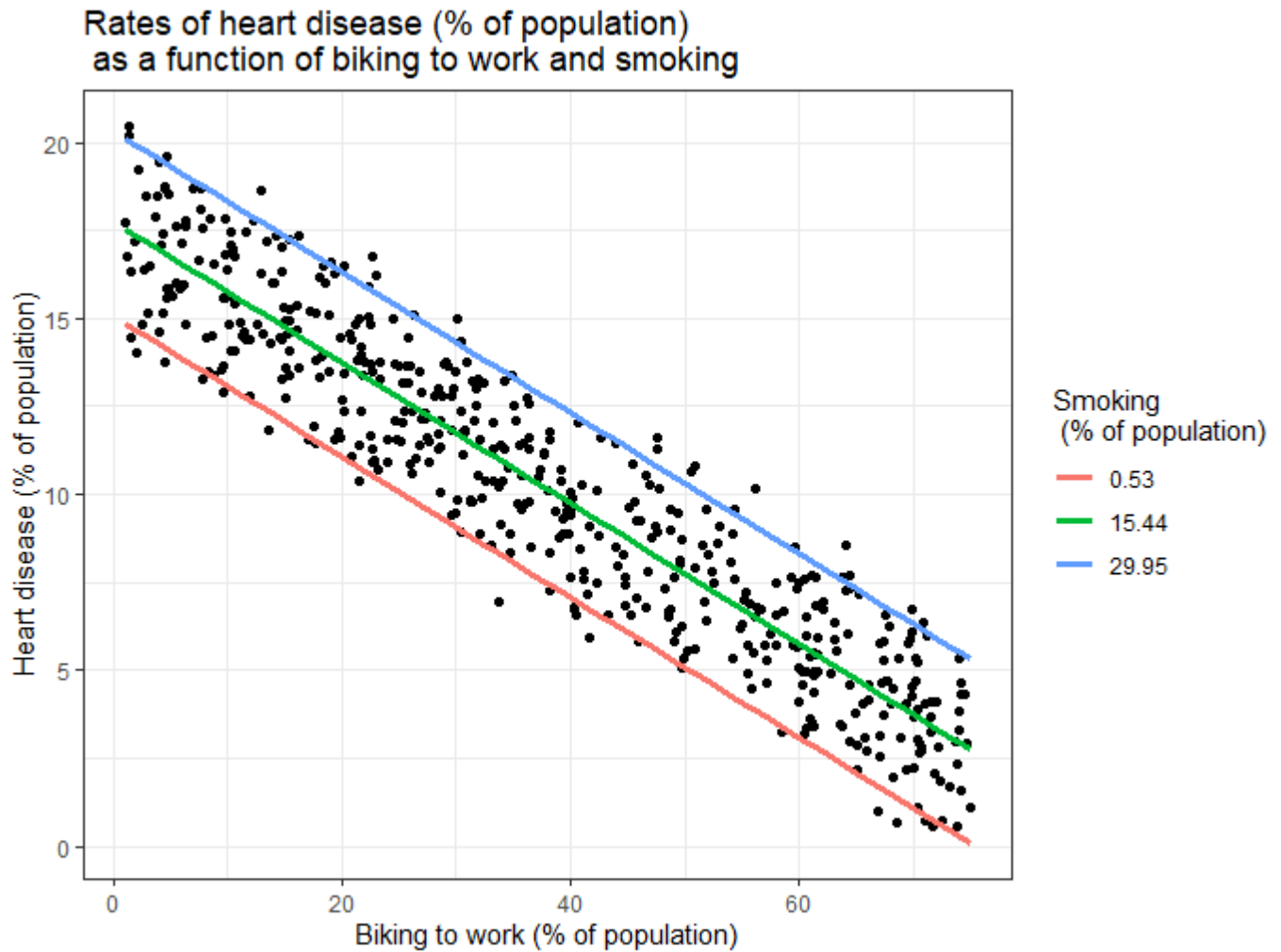
When reporting your results, include the estimated effect (i.e. the regression coefficient), the standard error of the estimate, and the p -value. You should also interpret your numbers to make it clear to your readers what the regression coefficient means.

In our survey of 500 towns, we found significant relationships between the frequency of biking to work and the frequency of heart disease and the frequency of smoking and frequency of heart disease ($p < 0.001$ for each). Specifically we found a 0.2% decrease (± 0.0014) in the frequency of heart disease for every 1% increase in biking, and a 0.178% increase (± 0.0035) in the frequency of heart disease for every 1% increase in smoking.

Visualizing the results in a graph

It can also be helpful to include a graph with your results. Multiple linear regression is somewhat more complicated than simple linear regression, because there are more parameters than will fit on a two-dimensional plot.

However, there are ways to display your results that include the effects of multiple independent variables on the dependent variable, even though only one independent variable can actually be plotted on the x-axis.



Here, we have calculated the predicted values of the dependent variable (heart disease) across the full range of observed values for the percentage of people biking to work.

To include the effect of smoking on the independent variable, we calculated these predicted values while holding smoking constant at the minimum, mean, and maximum observed rates of smoking.

Correlation vs. Regression:

Correlation	Regression
‘Correlation’ as the name says it determines the interconnection or a co-relationship between the variables.	‘Regression’ explains how an independent variable is numerically associated with the dependent variable.
In Correlation, both the independent and dependent values have no difference.	However, in Regression, both the dependent and independent variable are different.
The primary objective of Correlation is, to find out a quantitative/numerical value expressing the	When it comes to regression, its primary intent is, to reckon the values of a haphazard variable

association between the values.	based on the values of the fixed variable.
Correlation stipulates the degree to which both of the variables can move together.	However, regression specifies the effect of the change in the unit, in the known variable(p) on the evaluated variable (q).
Correlation helps to constitute the connection between the two variables.	Regression helps in estimating a variable's value based on another given value.

1. Sum of Squares Total (SST) – The sum of squared differences between individual data points (y_i) and the mean of the response variable (\bar{y}).

- $SST = \sum (y_i - \bar{y})^2$

2. Sum of Squares Regression (SSR) – The sum of squared differences between predicted data points (\hat{y}_i) and the mean of the response variable(\bar{y}).

- $SSR = \sum (\hat{y}_i - \bar{y})^2$

3. Sum of Squares Error (SSE) – The sum of squared differences between predicted data points (\hat{y}_i) and observed data points (y_i).

- $SSE = \sum (\hat{y}_i - y_i)^2$

The following relationship exists between these three measures:

$$SST = SSR + SSE$$

Thus, if we know two of these measures then we can use some simple algebra to calculate the third.

SSR, SST & R-Squared

R-squared, sometimes referred to as the coefficient of determination, is a measure of how well a linear regression model fits a dataset. It represents the proportion of the variance in the **response variable** that can be explained by the predictor variable.

The value for R-squared can range from 0 to 1. A value of 0 indicates that the response variable cannot be explained by the predictor variable at all. A value of 1 indicates that the response variable can be perfectly explained without error by the predictor variable.

Using SSR and SST, we can calculate R-squared as:

$$\mathbf{R\text{-}squared = SSR / SST}$$

For example, if the SSR for a given regression model is 137.5 and SST is 156 then we would calculate R-squared as:

$$R\text{-squared} = 137.5 / 156 = 0.8814$$

This tells us that 88.14% of the variation in the response variable can be explained by the predictor variable.

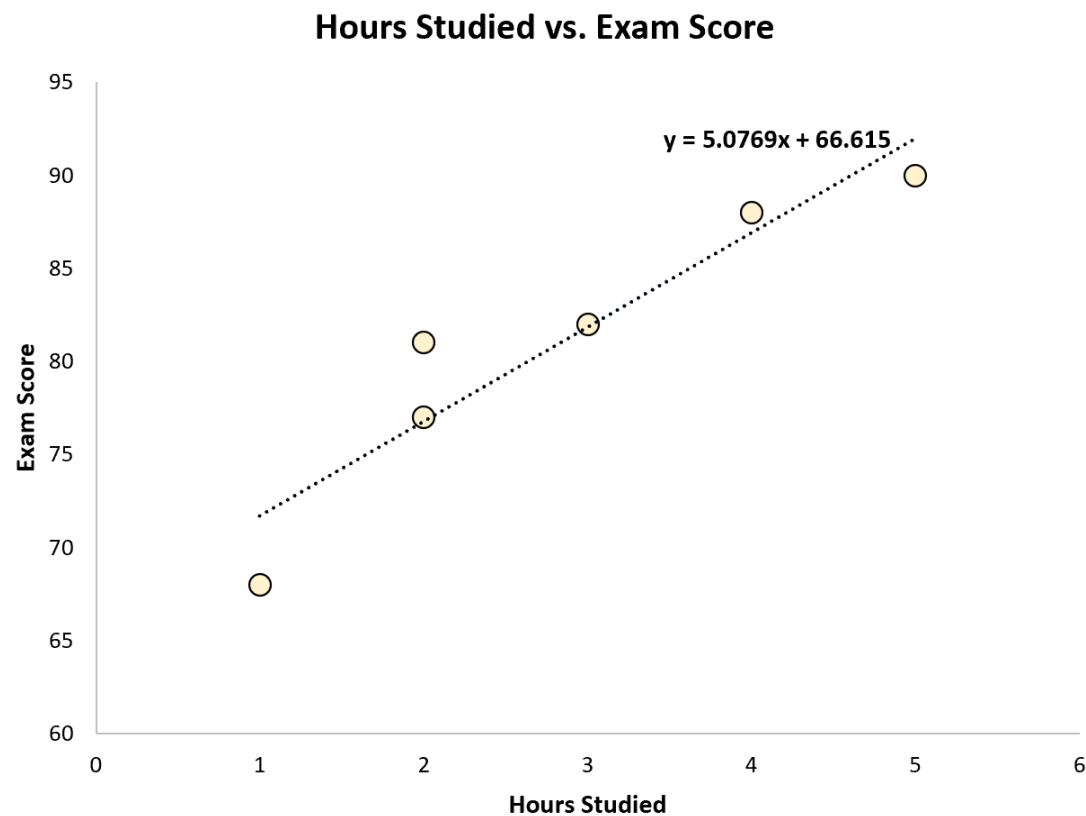
Calculate SST, SSR, SSE: Step-by-Step Example

Suppose we have the following dataset that shows the number of hours studied by six different students along with their final exam scores:

Hours Studied	Exam Score
1	68
2	77
2	81
3	82
4	88
5	90

Using some statistical software (like [R](#), [Excel](#), [Python](#)) or even [by hand](#), we can find that the line of best fit is:

$$\text{Score} = 66.615 + 5.0769 * (\text{Hours})$$



Once we know the line of best fit equation, we can use the following steps to calculate SST, SSR, and SSE:

Step 1: Calculate the mean of the response variable.

The mean of the response variable (\bar{y}) turns out to be **81**.

Hours Studied	Exam Score	\bar{y}
1	68	81
2	77	81
2	81	81
3	82	81
4	88	81
5	90	81

Step 2: Calculate the predicted value for each observation.

Next, we can use the line of best fit equation to calculate the predicted exam score (\hat{y}) for each student.

For example, the predicted exam score for the student who studied one hours is:

$$\text{Score} = 66.615 + 5.0769*(1) = \mathbf{71.69}.$$

We can use the same approach to find the predicted score for each student:

Hours Studied	Exam Score	\bar{y}	\hat{y}
1	68	81	71.69
2	77	81	76.77
2	81	81	76.77
3	82	81	81.85
4	88	81	86.92
5	90	81	92.00

Step 3: Calculate the sum of squares total (SST).

Next, we can calculate the sum of squares total.

For example, the sum of squares total for the first student is:

$$(y_i - \bar{y})^2 = (68 - 81)^2 = \mathbf{169}.$$

We can use the same approach to find the sum of squares total for each student:

Hours Studied	Exam Score	\bar{y}	\hat{y}	$(y_i - \bar{y})^2$
1	68	81	71.69	169
2	77	81	76.77	16
2	81	81	76.77	0
3	82	81	81.85	1
4	88	81	86.92	49
5	90	81	92.00	81
				316

SST

The sum of squares total turns out to be **316**.

Step 4: Calculate the sum of squares regression (SSR).

Next, we can calculate the sum of squares regression.

For example, the sum of squares regression for the first student is:

$$(\hat{y}_i - \bar{y})^2 = (71.69 - 81)^2 = \mathbf{86.64}.$$

We can use the same approach to find the sum of squares regression for each student:

Hours Studied	Exam Score	\bar{y}	\hat{y}	$(y_i - \bar{y})^2$	$(\hat{y}_i - \bar{y})^2$
1	68	81	71.69	169	86.64
2	77	81	76.77	16	17.90
2	81	81	76.77	0	17.90
3	82	81	81.85	1	0.72
4	88	81	86.92	49	35.08
5	90	81	92.00	81	120.99
				316	279.23

SST

SSR

The sum of squares regression turns out to be **279.23**.

Step 5: Calculate the sum of squares error (SSE).

Next, we can calculate the sum of squares error.

For example, the sum of squares error for the first student is:

$$(\hat{y}_i - y_i)^2 = (71.69 - 68)^2 = \mathbf{13.63}.$$

We can use the same approach to find the sum of squares error for each student:

Hours Studied	Exam Score	\bar{y}	\hat{y}	$(y_i - \bar{y})^2$	$(\hat{y}_i - \bar{y})^2$	$(\hat{y}_i - y_i)^2$
1	68	81	71.69	169	86.64	13.63
2	77	81	76.77	16	17.90	0.05
2	81	81	76.77	0	17.90	17.90
3	82	81	81.85	1	0.72	0.02
4	88	81	86.92	49	35.08	1.16
5	90	81	92.00	81	120.99	4.00
				316	279.23	36.77
				SST	SSR	SSE

We can verify that $SST = SSR + SSE$

- $SST = SSR + SSE$
- $316 = 279.23 + 36.77$

We can also calculate the R-squared of the regression model by using the following equation:

- $R\text{-squared} = SSR / SST$
- $R\text{-squared} = 279.23 / 316$
- $R\text{-squared} = 0.8836$
- **Multiple Linear Regression:**
- Multiple Linear Regression is an extension of Simple Linear regression as it takes more than one predictor variable to predict the response variable. We can define it as:
 - *Multiple Linear Regression is one of the important regression algorithms which models the linear relationship between a single dependent continuous variable and more than one independent variable.*

Example:

Prediction of CO₂ emission based on engine size and number of cylinders in a car.

Some key points about MLR:

- For MLR, the dependent or target variable(Y) must be the continuous/real, but the predictor or independent variable may be of continuous or categorical form.
- Each feature variable must model the linear relationship with the dependent variable.
- MLR tries to fit a regression line through a multidimensional space of data-points.

MLR equation:

In Multiple Linear Regression, the target variable(Y) is a linear combination of multiple predictor variables $x_1, x_2, x_3, \dots, x_n$. Since it is an enhancement of Simple Linear Regression, so the same is applied for the multiple linear regression equation, the equation becomes:

$$1. Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n \quad \dots\dots\dots (a)$$

Where,

Y= Output/Response variable

$b_0, b_1, b_2, b_3, b_n, \dots$ = Coefficients of the model.

$x_1, x_2, x_3, x_4, \dots$ = Various Independent/feature variable

Assumptions for Multiple Linear Regression:

- A **linear relationship** should exist between the Target and predictor variables.
- The regression residuals must be **normally distributed**.
- MLR assumes little or **no multicollinearity** (correlation between the independent variable) in data.

Implementation of Multiple Linear Regression model using Python:

To implement MLR using Python, we have below problem:

Problem Description:

We have a dataset of **50 start-up companies**. This dataset contains five main information: **R&D Spend, Administration Spend, Marketing Spend, State, and Profit for a financial year**. Our goal is to create a model that can easily determine which company has a maximum profit, and which is the most affecting factor for the profit of a company.

Since we need to find the Profit, so it is the dependent variable, and the other four variables are independent variables. Below are the main steps of deploying the MLR model:

1. Data Pre-processing Steps

2. **Fitting the MLR model to the training set**
3. **Predicting the result of the test set**

Step-1: Data Pre-processing Step:

The very first step is [data pre-processing](#), which we have already discussed in this tutorial. This process contains the below steps:

- **Importing libraries:** Firstly we will import the library which will help in building the model. Below is the code for it:

1. # importing libraries
2. **import** numpy as nm
3. **import** matplotlib.pyplot as mtp
4. **import** pandas as pd

- **Importing dataset:** Now we will import the dataset(50_CompList), which contains all the variables. Below is the code for it:

1. #importing datasets
2. data_set= pd.read_csv('50_CompList.csv')

Output: We will get the dataset as:

data_set - DataFrame

Index	R&D Spend	Administration	Marketing Spend	State	Profit
0	165349	136898	471784	New York	192262
1	162598	151378	443899	California	191792
2	153442	101146	407935	Florida	191050
3	144372	118672	383200	New York	182902
4	142107	91391.8	366168	Florida	166188
5	131877	99814.7	362861	New York	156991
6	134615	147199	127717	California	156123
7	130298	145530	323877	Florida	155753
8	120543	148719	311613	New York	152212
9	123335	108679	304982	California	149760
10	101913	110594	229161	Florida	146122
11	100672	91790.6	249745	California	144259
12	93863.8	127320	249839	Florida	141586
13	91992.4	135495	252665	California	134307

Format Resize ☒ Background color ☒ Column min/max Save and Close Close

In above output, we can clearly see that there are five variables, in which four variables are continuous and one is categorical variable.

- **Extracting dependent and independent Variables:**

1. #Extracting Independent and dependent Variable
2. `x= data_set.iloc[:, :-1].values`
3. `y= data_set.iloc[:, 4].values`

Encoding Dummy Variables:

As we have one categorical variable (State), which cannot be directly applied to the model, so we will encode it. To encode the categorical variable into numbers, we will use the **LabelEncoder** class. But it is not sufficient because it still has some relational order, which may create a wrong model. So in order to remove this problem, we will use **OneHotEncoder**, which will create the dummy variables. Below is code for it:

1. #Categorical data
2. from sklearn.preprocessing **import** LabelEncoder, OneHotEncoder
3. labelencoder_x= LabelEncoder()
4. x[:, 3]= labelencoder_x.fit_transform(x[:,3])
5. onehotencoder= OneHotEncoder(categorical_features= [3])
6. x= onehotencoder.fit_transform(x).toarray()

Here we are only encoding one independent variable, which is state as other variables are continuous.

Output:

x - NumPy array

	0	1	2	3	4	5
0	0	0	1	165349	136898	471784
1	1	0	0	162598	151378	443899
2	0	1	0	153442	101146	407935
3	0	0	1	144372	118672	383200
4	0	1	0	142107	91391.8	366168
5	0	0	1	131877	99814.7	362861
6	1	0	0	134615	147199	127717
7	0	1	0	130298	145530	323877
8	0	0	1	120543	148719	311613
9	1	0	0	123335	108679	304982
10	0	1	0	101913	110594	229161
11	1	0	0	100672	91790.6	249745
12	0	1	0	93863.8	127320	249839
13	1	0	0	91992.4	135495	252665

Format Resize ☒ Background color

As we can see in the above output, the state column has been converted into dummy variables (0 and 1). **Here each dummy variable column is corresponding to the one State.** We can check by comparing it with the original dataset. The first column corresponds to the **California State**, the second column corresponds to the **Florida State**, and the third column corresponds to the **New York State**.

Regression analysis in Excel - the basics

In statistical modeling, **regression analysis** is used to estimate the relationships between two or more variables:

Dependent variable (aka *criterion* variable) is the main factor you are trying to understand and predict.

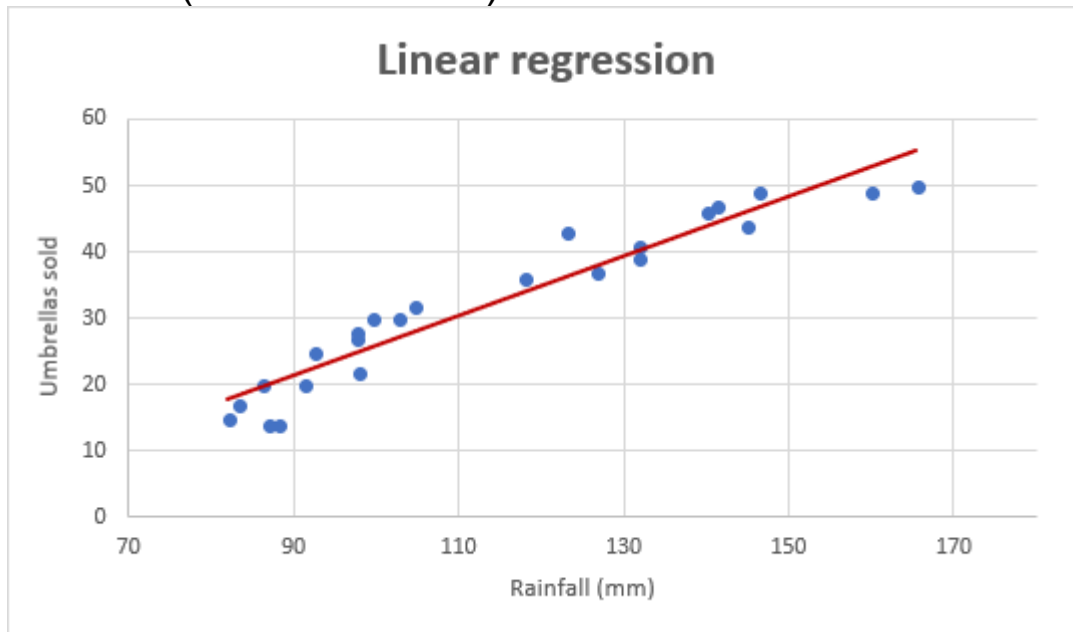
Independent variables (aka *explanatory* variables, or *predictors*) are the factors that might influence the dependent variable.

Regression analysis helps you understand how the dependent variable changes when one of the independent variables varies and allows to mathematically determine which of those variables really has an impact.

Technically, a regression analysis model is based on the **sum of squares**, which is a mathematical way to find the dispersion of data points. The goal of a model is to get the smallest possible sum of squares and draw a line that comes closest to the data.

In statistics, they differentiate between a simple and multiple linear regression. **Simple linear regression** models the relationship between a dependent variable and one independent variables using a linear function. If you use two or more explanatory variables to predict the dependent variable, you deal with **multiple linear regression**. If the dependent variable is modeled as a non-linear function because the data relationships do not follow a straight line, use **nonlinear regression** instead. The focus of this tutorial will be on a simple linear regression.

As an example, let's take sales numbers for umbrellas for the last 24 months and find out the average monthly rainfall for the same period. Plot this information on a chart, and the regression line will demonstrate the relationship between the independent variable (rainfall) and dependent variable (umbrella sales):



Linear regression equation

Mathematically, a linear regression is defined by this equation:

$$y = bx + a + \epsilon$$

Where:

- x is an independent variable.
- y is a dependent variable.
- a is the *Y-intercept*, which is the expected mean value of y when all x variables are equal to 0. On a regression graph, it's the point where the line crosses the Y axis.
- b is the *slope* of a regression line, which is the rate of change for y as x changes.

- ε is the random error term, which is the difference between the actual value of a dependent variable and its predicted value.

The linear regression equation always has an error term because, in real life, predictors are never perfectly precise. However, some programs, including Excel, do the error term calculation behind the scenes. So, in Excel, you do linear regression using the **least squares** method and seek coefficients a and b such that:

$$y = bx + a$$

For our example, the linear regression equation takes the following shape:

```
Umbrellas sold = b * rainfall + a
```

There exist a handful of different ways to find a and b . The three main methods to perform linear regression analysis in Excel are:

- Regression tool included with Analysis ToolPak
- Scatter chart with a trendline
- Linear regression formula

Below you will find the detailed instructions on using each method.

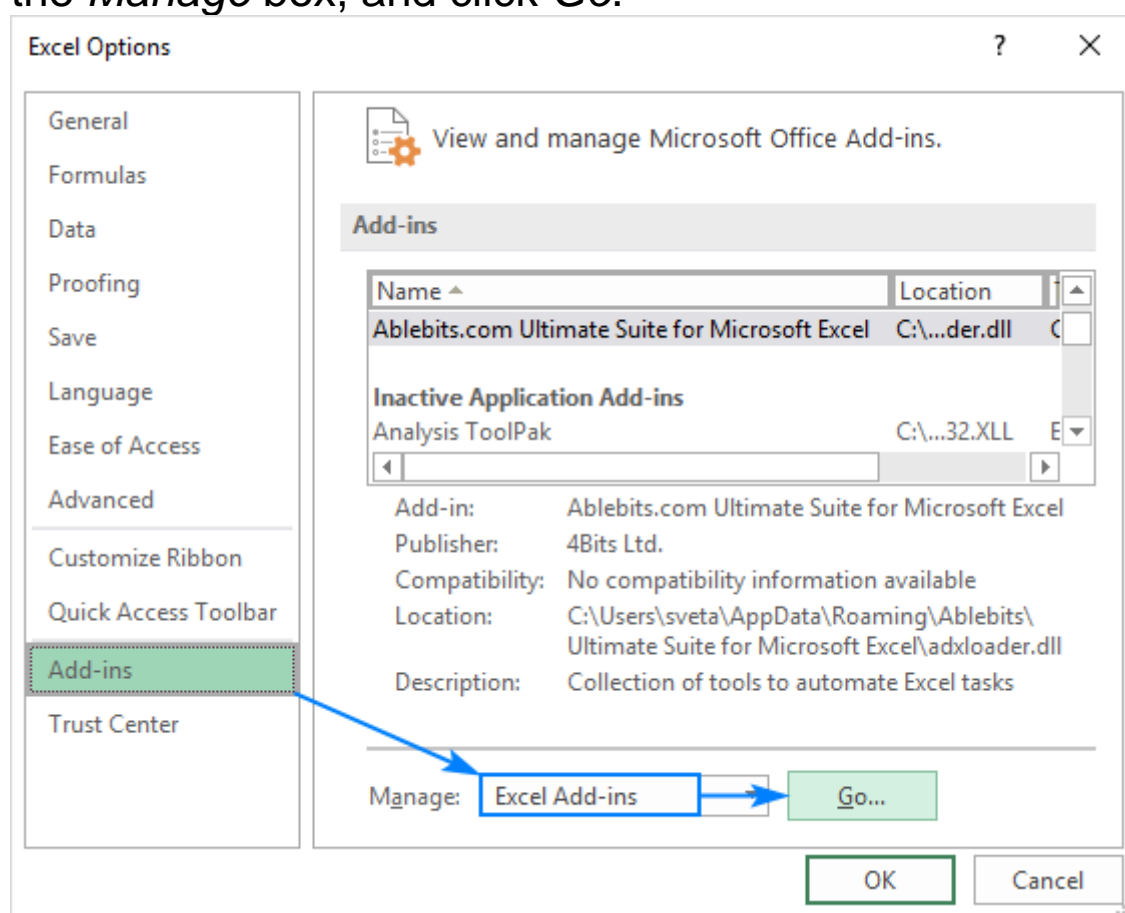
How to do linear regression in Excel with Analysis ToolPak

This example shows how to run regression in Excel by using a special tool included with the Analysis ToolPak add-in.

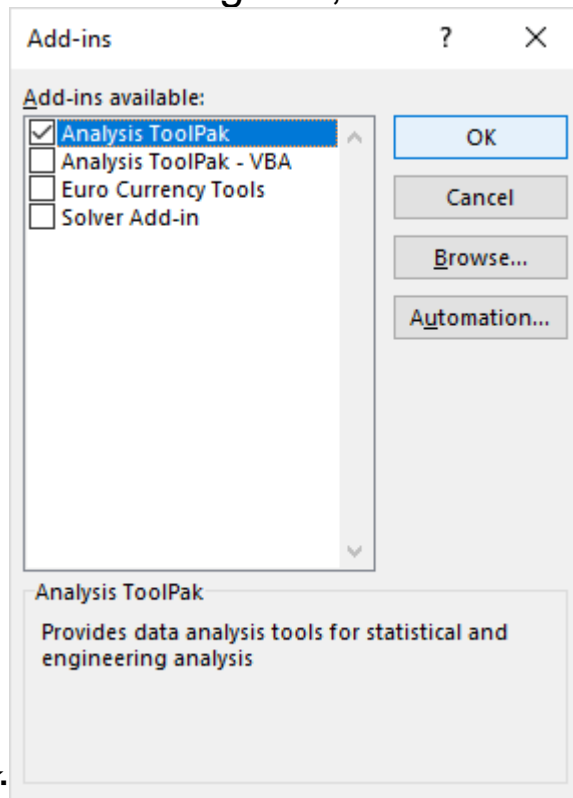
Enable the Analysis ToolPak add-in

Analysis ToolPak is available in all versions of Excel 2019 to 2003 but is not enabled by default. So, you need to turn it on manually. Here's how:

1. In your Excel, click *File > Options*.
2. In the *Excel Options* dialog box, select **Add-ins** on the left sidebar, make sure **Excel Add-ins** is selected in the *Manage* box, and click *Go*.



3. In the *Add-ins* dialog box, tick off **Analysis Toolpak**, and



click *OK*:

This will add the **Data Analysis** tools to the *Data* tab of your Excel ribbon.

Run regression analysis

In this example, we are going to do a simple linear regression in Excel. What we have is a list of average monthly rainfall for the last 24 months in column B, which is our independent variable (predictor), and the number of umbrellas sold in column C, which is the dependent variable. Of course, there are many other factors that can

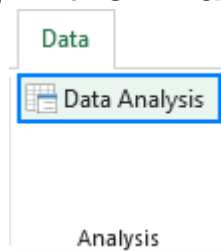
affect sales, but for now we focus only on these two

	A	B	C
1	Month	Rainfall (mm)	Umbrellas sold
2	Jan	82	15
3	Feb	92.5	25
4	Mar	83.2	17
5	Apr	97.7	28
6	May	131.9	41
7	Jun	141.3	47
8	Jul	165.4	50
9	Aug	140	46
10	Sep	126.7	37

variables:

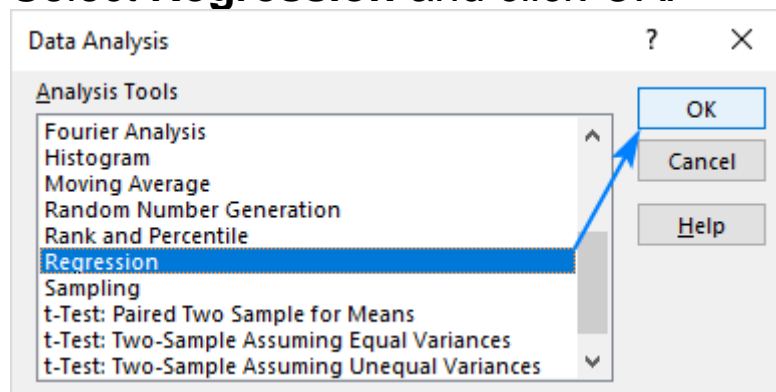
With Analysis Toolpak added enabled, carry out these steps to perform regression analysis in Excel:

1. On the *Data* tab, in the *Analysis* group, click the **Data**



Analysis button.

2. Select **Regression** and click *OK*.



3. In the *Regression* dialog box, configure the following settings:

- Select the *Input Y Range*, which is your **dependent variable**. In our case, it's umbrella sales (C1:C25).

- Select the *Input X Range*, i.e. your **independent variable**. In this example, it's the average monthly rainfall (B1:B25).

If you are building a multiple regression model, select two or more adjacent columns with different independent variables.

- Check the **Labels box** if there are headers at the top of your X and Y ranges.
- Choose your preferred **Output option**, a new worksheet in our case.
- Optionally, select the **Residuals** checkbox to get the difference between the predicted and actual values.

	A	B	C
1	Month	Rainfall (mm)	Umbrellas sold
2	Jan	82	15
3	Feb	92.5	25
4	Mar	83.2	17
5	Apr	97.7	28
6	May	131.9	41
7	Jun	141.3	47
8	Jul	165.4	50
9	Aug	140	46
10	Sep	126.7	37
11	Oct	97.8	22
12	Nov	86.2	20
13	Dec	99.6	30
14	Jan	87	14
15	Feb	97.5	27
16	Mar	88.2	14
17	Apr	102.7	30
18	May	123	43
19	Jun	146.3	49
20	Jul	160	49
21	Aug	145	44
22	Sep	131.7	39
23	Oct	118	36
24	Nov	91.2	20
25	Dec	104.6	32

Regression

Input

Input Y Range:

Input X Range:

☒ Labels ☐ Constant is Zero

☐ Confidence Level: %

Output options

☐ Output Range:

☒ New Worksheet Ply:

☐ New Workbook

Residuals

☒ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

OK Cancel Help

4. Click *OK* and observe the regression analysis output created by Excel.

Interpret regression analysis output

As you have just seen, running regression in Excel is easy because all calculations are preformed automatically. The interpretation of the results is a bit trickier because you need to know what is behind each number. Below you will find a breakdown of 4 major parts of the regression analysis output.

Regression analysis output: Summary Output

This part tells you how well the calculated linear regression

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.957666798
R Square	0.917125697
Adjusted R Square	0.913358683
Standard Error	3.58141382
Observations	24

equation fits your source data.

Here's what each piece of information means:

Multiple R. It is the *Correlation Coefficient* that measures the strength of a linear relationship between two variables. The correlation coefficient can be any value between -1 and 1, and its [absolute value](#) indicates the relationship strength. The larger the absolute value, the stronger the relationship:

- 1 means a strong positive relationship
- -1 means a strong negative relationship
- 0 means no relationship at all

R Square. It is the *Coefficient of Determination*, which is used as an indicator of the goodness of fit. It shows how

many points fall on the regression line. The R^2 value is calculated from the total sum of squares, more precisely, it is the sum of the squared deviations of the original data from the mean.

In our example, R^2 is 0.91 (rounded to 2 digits), which is fairly good. It means that 91% of our values fit the regression analysis model. In other words, 91% of the dependent variables (y-values) are explained by the independent variables (x-values). Generally, R Squared of 95% or more is considered a good fit.

Adjusted R Square. It is the *R square* adjusted for the number of independent variable in the model. You will want to use this value instead of *R square* for multiple regression analysis.

Standard Error. It is another goodness-of-fit measure that shows the precision of your regression analysis - the smaller the number, the more certain you can be about your regression equation. While R^2 represents the percentage of the dependent variables variance that is explained by the model, Standard Error is an absolute measure that shows the average distance that the data points fall from the regression line.

Observations. It is simply the number of observations in your model.

Regression analysis output: ANOVA

The second part of the output is Analysis of Variance (ANOVA):

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	3122.775	3122.775	243.4623	2.21604E-13
Residual	22	282.1835	12.82652		
Total	23	3404.958			

Basically, it splits the sum of squares into individual components that give information about the levels of variability within your regression model:

- *df* is the number of the degrees of freedom associated with the sources of variance.
- *SS* is the sum of squares. The smaller the Residual *SS* compared with the Total *SS*, the better your model fits the data.
- *MS* is the mean square.
- *F* is the *F* statistic, or *F*-test for the null hypothesis. It is used to test the overall significance of the model.
- *Significance F* is the *P*-value of *F*.

The ANOVA part is rarely used for a simple linear regression analysis in Excel, but you should definitely have a close look at the last component. The **Significance F** value gives an idea of how reliable (statistically significant) your results are. If Significance *F* is less than 0.05 (5%), your model is OK. If it is greater than 0.05, you'd probably better choose another independent variable.

Regression analysis output: coefficients

This section provides specific information about the components of your analysis:

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	-19.07410899	3.372182168	-5.656310378	1.09E-05	-26.06758677	-12.08063122
Rainfall	0.45000132	0.02884018	15.6032773	2.22E-13	0.390190448	0.509812192

The most useful component in this section is **Coefficients**. It enables you to build a [linear regression equation](#) in Excel:

$$y = bx + a$$

For our data set, where y is the number of umbrellas sold and x is an average monthly rainfall, our linear regression formula goes as follows:

$$Y = \text{Rainfall Coefficient} * x + \text{Intercept}$$

Equipped with a and b values rounded to three decimal places, it turns into:

$$Y = 0.45 * x - 19.074$$

For example, with the average monthly rainfall equal to 82 mm, the umbrella sales would be approximately 17.8:

$$0.45 * 82 - 19.074 = 17.8$$

In a similar manner, you can find out how many umbrellas are going to be sold with any other monthly rainfall (x variable) you specify.

Regression analysis output: residuals

If you compare the estimated and actual number of sold umbrellas corresponding to the monthly rainfall of 82 mm, you will see that these numbers are slightly different:

- Estimated: 17.8 (calculated above)

- Actual: 15 (row 2 of the source data)

Why's the difference? Because independent variables are never perfect predictors of the dependent variables. And the residuals can help you understand how far away the actual values are from the predicted values:

RESIDUAL OUTPUT		
Observation	Predicted Umbrellas sold	Residuals
1	17.82599924	-2.825999237
2	22.5510131	2.448986904
3	18.36600082	-1.366000821
4	24.89101996	3.10898004
5	40.2810651	0.7189349
6	44.51107751	2.488922493
7	55.35610932	-5.356109317
8	43.92607579	2.073924208
9	37.94105824	-0.941058237
10	24.93602009	-2.936020092
11	19.71600478	0.283995219
12	25.74602247	4.253977532

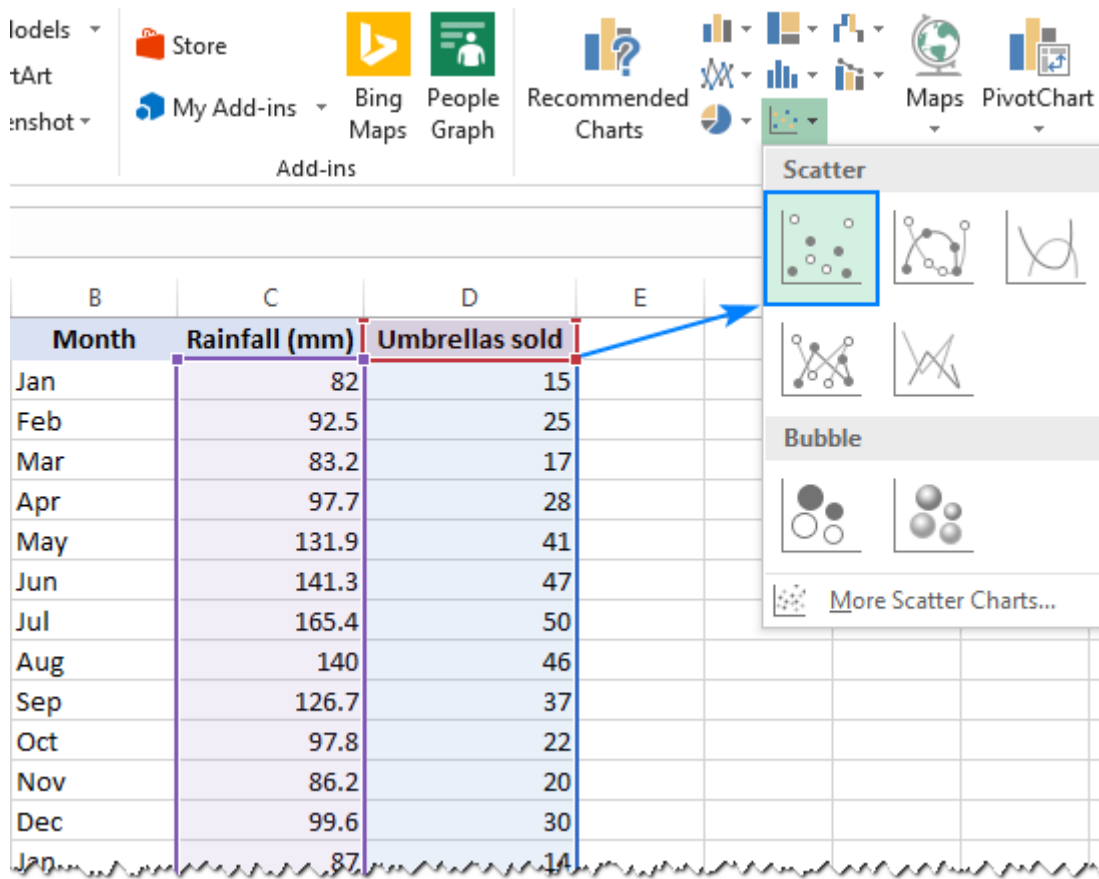
For the first data point (rainfall of 82 mm), the residual is approximately -2.8. So, we add this number to the predicted value, and get the actual value: $17.8 - 2.8 = 15$.

How to make a linear regression graph in Excel

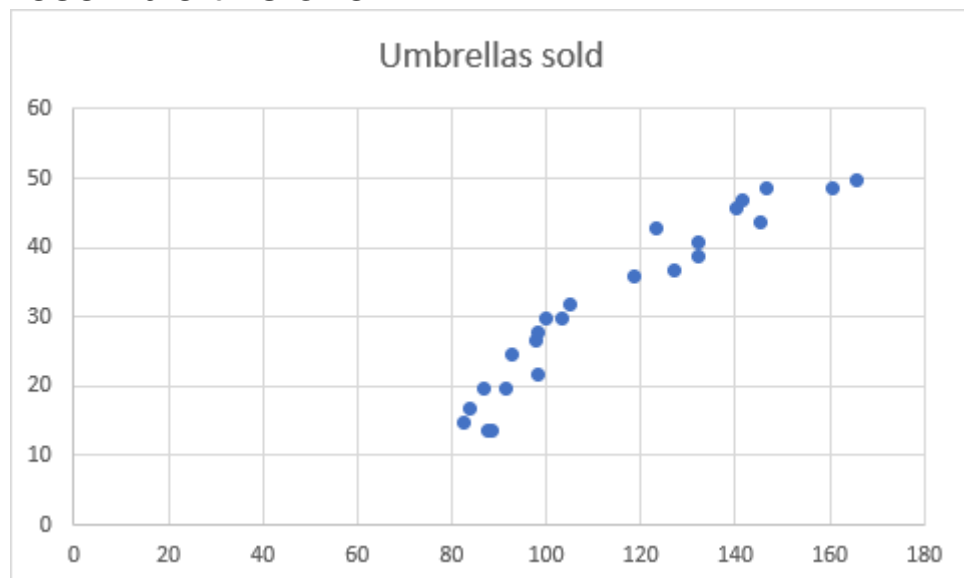
If you need to quickly visualize the relationship between the two variables, draw a linear regression chart. That's very easy! Here's how:

1. Select the two columns with your data, including headers.
2. On the *Insert* tab, in the *Charts* group, click the *Scatter chart* icon, and select the **Scatter** thumbnail (the first

one):

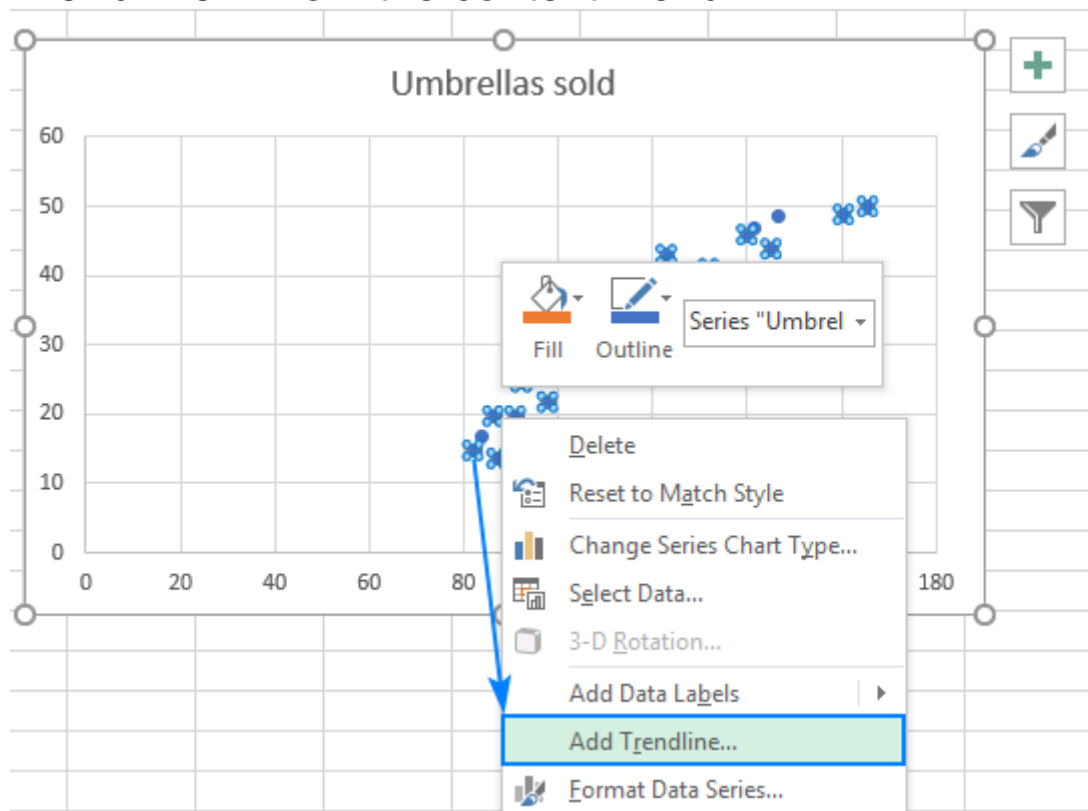


This will insert a [scatter plot](#) in your worksheet, which will resemble this one:



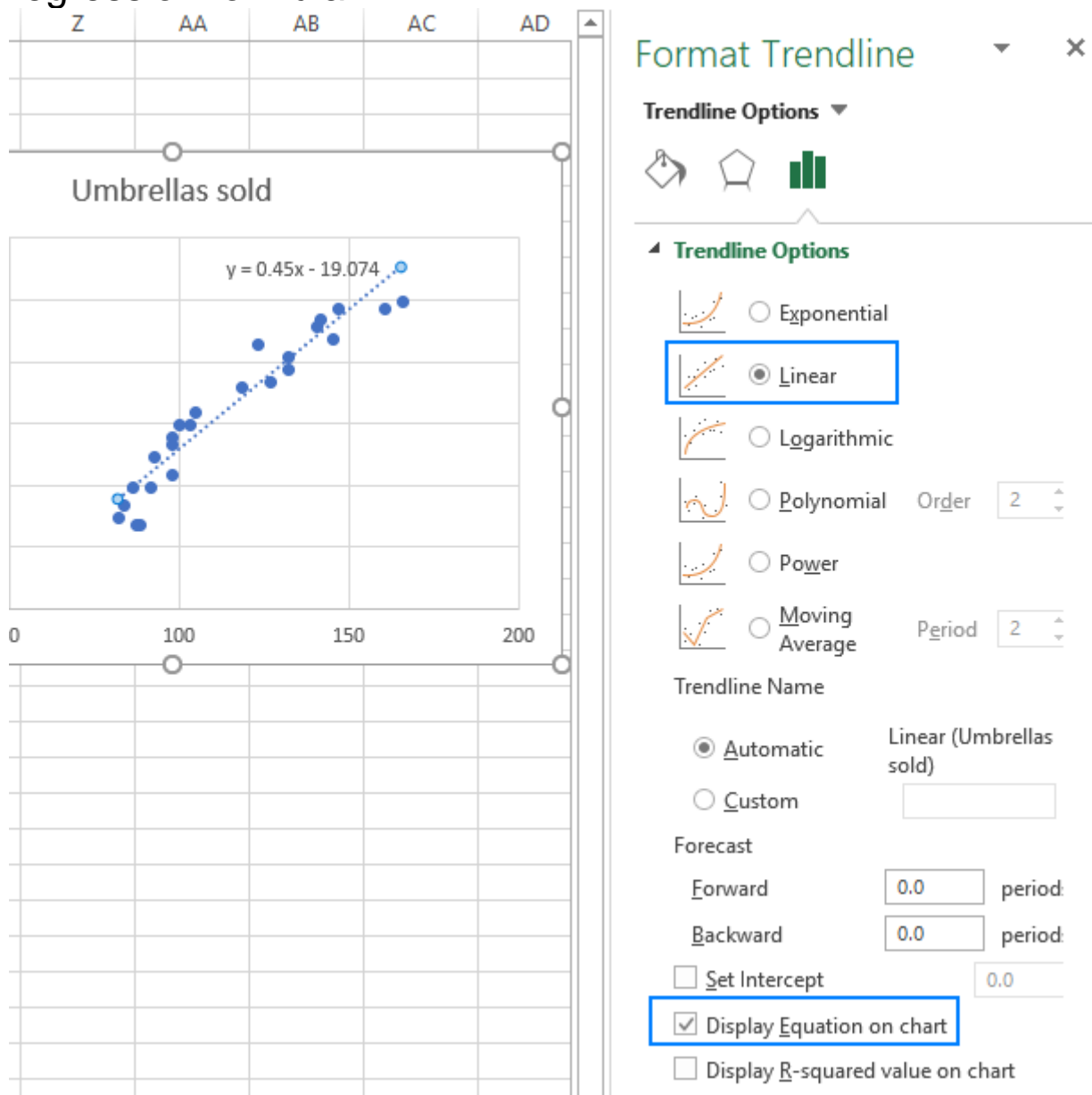
- Now, we need to draw the least squares regression line. To have it done, right click on any point and choose **Add**

Trendline... from the context menu.



4. On the right pane, select the **Linear** trendline shape and, optionally, check **Display Equation on Chart** to get your

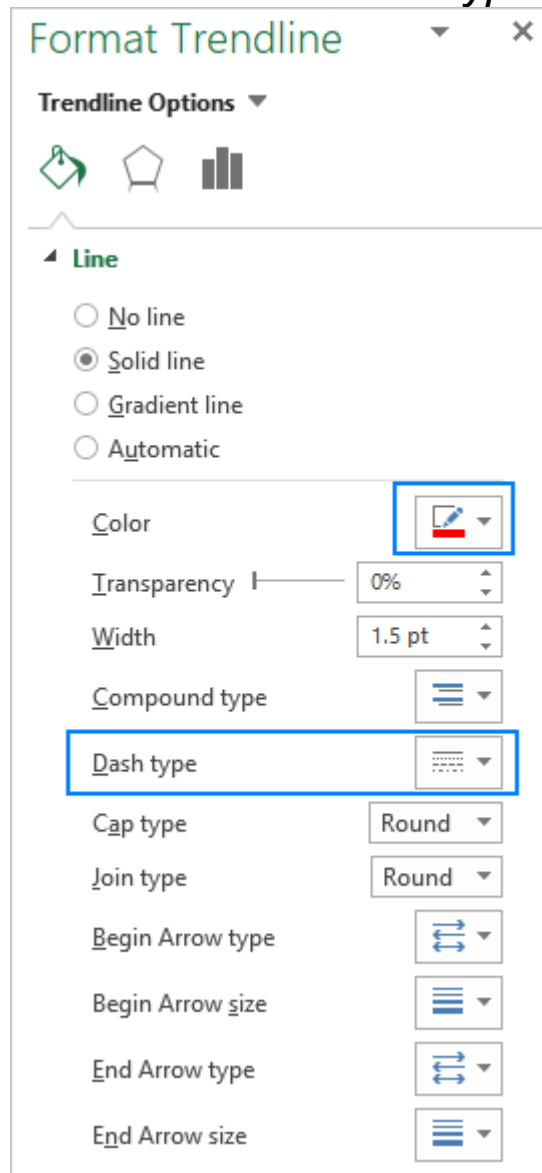
regression formula:



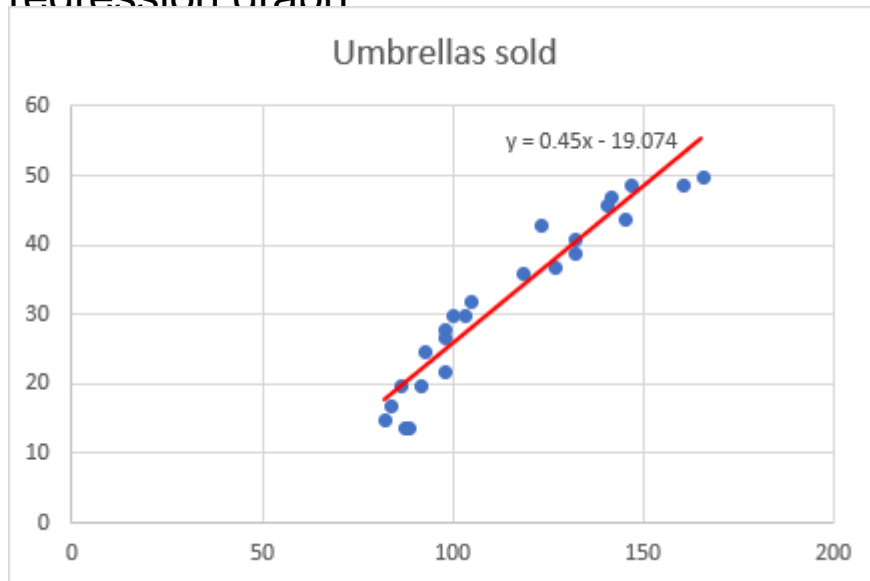
As you may notice, the regression equation Excel has created for us is the same as the linear regression formula we built based on the [Coefficients output](#).

5. Switch to the *Fill & Line* tab and customize the line to your liking. For example, you can choose a different line color and use a solid line instead of a dashed line (select

Solid line in the *Dash type* box):



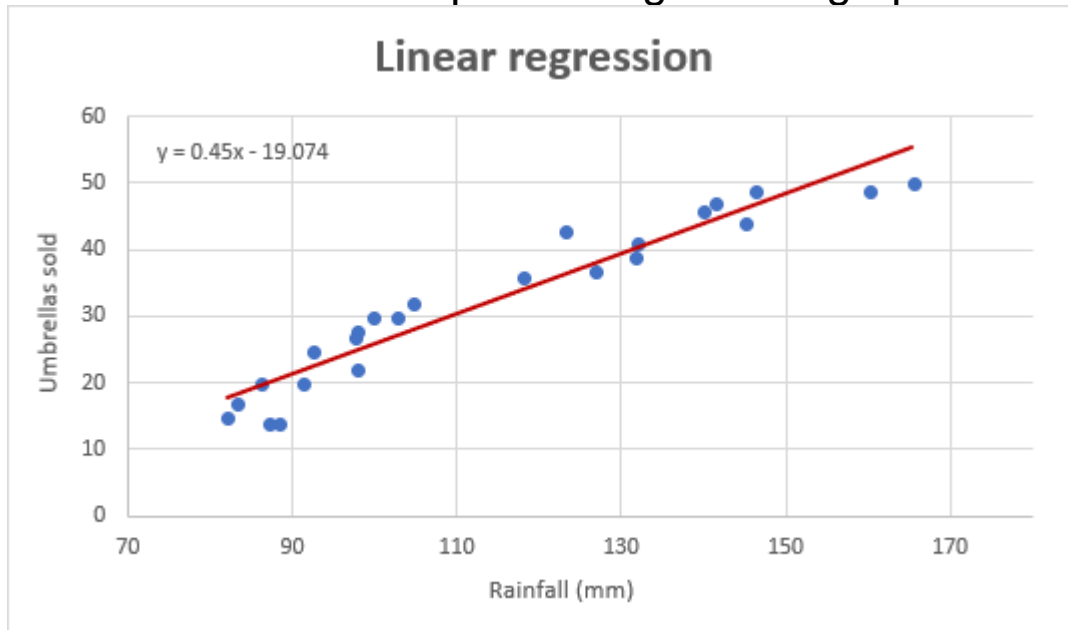
At this point, your chart already looks like a decent regression graph:



Still, you may want to make a few more improvements:

- Drag the equation wherever you see fit.
- Add axes titles (*Chart Elements* button > *Axis Titles*).
- If your data points start in the middle of the horizontal and/or vertical axis like in this example, you may want to get rid of the excessive white space. The following tip explains how to do this: [Scale the chart axes to reduce white space](#).

And this is how our improved regression graph looks like:



Important note! In the regression graph, the independent variable should always be on the X axis and the dependent variable on the Y axis. If your graph is plotted in the reverse order, swap the columns in your worksheet, and then draw the chart anew. If you are not allowed to rearrange the source data, then you can [switch the X and Y axes](#) directly in a chart.

How to do regression in Excel using formulas

Microsoft Excel has a few statistical functions that can help you to do linear regression analysis such as LINEST, SLOPE, INTERCEPT, and CORREL.

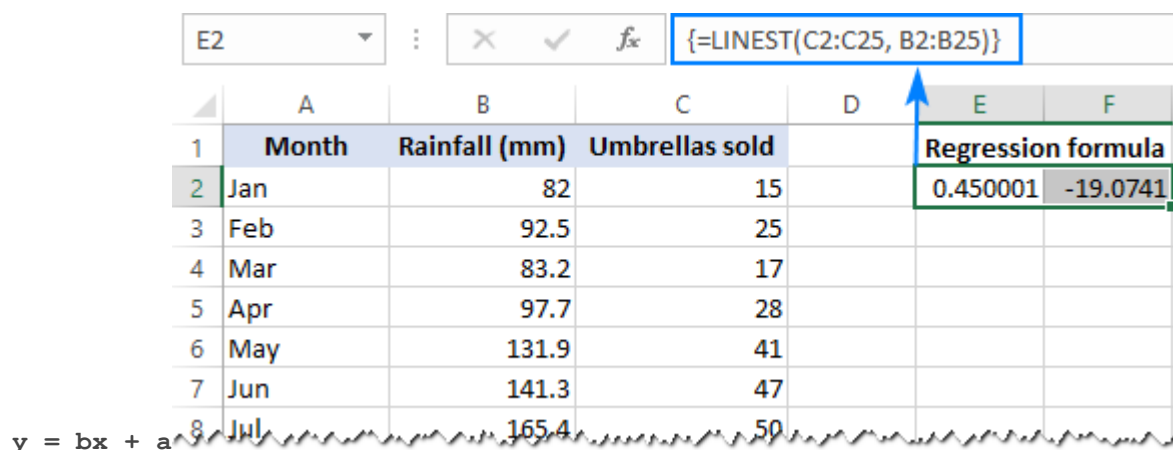
The [LINEST function](#) uses the least squares regression method to calculate a straight line that best explains the relationship between your variables and returns an array describing that line. You can find the detailed explanation of

the function's syntax in [this tutorial](#). For now, let's just make a formula for our sample dataset:

```
=LINEST(C2:C25, B2:B25)
```

Because the LINEST function returns an array of values, you must enter it as an [array formula](#). Select two adjacent cells in the same row, E2:F2 in our case, type the formula, and press **Ctrl + Shift + Enter** to complete it.

The formula returns the *b* coefficient (E1) and the *a* constant (F1) for the already familiar linear regression equation:



	A	B	C	D	E	F
1	Month	Rainfall (mm)	Umbrellas sold		Regression formula	
2	Jan	82	15		0.450001	-19.0741
3	Feb	92.5	25			
4	Mar	83.2	17			
5	Apr	97.7	28			
6	May	131.9	41			
7	Jun	141.3	47			
8	Jul	165.4	50			

$y = bx + a$

If you avoid using array formulas in your worksheets, you can calculate *a* and *b* individually with regular formulas:

Get the Y-intercept (*a*):

```
=INTERCEPT(C2:C25, B2:B25)
```

Get the slope (*b*):

```
=SLOPE(C2:C25, B2:B25)
```

Additionally, you can find the **correlation coefficient** (*Multiple R* in the regression analysis [summary output](#)) that indicates how strongly the two variables are related to each other:

=CORREL (B2 : B25 , C2 : C25)

The following screenshot shows all these Excel regression formulas in action:

	A	B	C	D	E	F	G	H	I
1	Month	Rainfall (mm)	Umbrellas sold		Regression formula		y = bx + a		
2	Jan	82	15		b	a			
3	Feb	92.5	25		0.450001	-19.0741	{=LINEST(C2:C25, B2:B25)}		
4	Mar	83.2	17						
5	Apr	97.7	28		a (Y-intercept)				
6	May	131.9	41		-19.0741		=INTERCEPT(C2:C25, B2:B25)		
7	Jun	141.3	47						
8	Jul	165.4	50		b (slope of a regression line)				
9	Aug	140	46		0.450001		=SLOPE(C2:C25, B2:B25)		
10	Sep	126.7	37						
11	Oct	97.8	22		Correlation coefficient				
12	Nov	86.2	20		0.957667		=CORREL(B2:B25,C2:C25)		
13	Dec	99.6	30						
14	Jan	87	14						
15	Feb	97.5	27						

Significance of P-Value:

What Is P-Value?

In statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. The p-value serves as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.

How Is P-Value Calculated?

P-values are usually found using p-value tables or spreadsheets/statistical software. These calculations are based on the assumed or known [probability distribution](#) of the specific statistic tested. P-values are calculated from the deviation between the observed value and a chosen reference value, given the probability distribution of the statistic, with a greater difference between the two values corresponding to a lower p-value.

Mathematically, the p-value is calculated using integral calculus from the area under the probability distribution curve for all values of statistics that are at least as far from the

reference value as the observed value is, relative to the total area under the probability distribution curve.

The calculation for a p-value varies based on the type of test performed. The three test types describe the location on the probability distribution curve: lower-tailed test, upper-tailed test, or two-sided test.

In a nutshell, the greater the difference between two observed values, the less likely it is that the difference is due to simple random chance, and this is reflected by a lower p-value.

The P-Value Approach to Hypothesis Testing

The p-value approach to hypothesis testing uses the calculated probability to determine whether there is evidence to reject the null hypothesis. The null hypothesis, also known as the “conjecture,” is the initial claim about a population (or data-generating process). The alternative hypothesis states whether the population parameter differs from the value of the population parameter stated in the conjecture.

In practice, the significance level is stated in advance to determine how small the p-value must be in order to reject the null hypothesis. Because different researchers use different levels of significance when examining a question, a reader may sometimes have difficulty comparing results from two different tests. P-values provide a solution to this problem.

For example, suppose a study comparing returns from two particular [assets](#) was undertaken by different researchers who used the same data but different significance levels. The researchers might come to opposite conclusions regarding whether the assets differ.

If one researcher used a confidence level of 90% and the other required a confidence level of 95% to reject the null hypothesis and the p-value of the observed difference between the two returns was 0.08 (corresponding to a confidence level of 92%), then the first researcher would find that the two assets have a difference that is [statistically significant](#), while the second would find no statistically significant difference between the returns.

To avoid this problem, the researchers could report the p-value of the hypothesis test and allow readers to interpret the statistical significance themselves. This is called a p-value approach to hypothesis testing. Independent observers could note the p-value and decide for themselves whether that represents a statistically significant difference or not.

Example of P-Value

An investor claims that their investment [portfolio's](#) performance is equivalent to that of the [Standard & Poor's \(S&P\) 500 Index](#). To determine this, the investor conducts a two-tailed test.

The null hypothesis states that the portfolio's returns are equivalent to the S&P 500's returns over a specified period, while the alternative hypothesis states that the portfolio's returns and the S&P 500's returns are not equivalent—if the investor conducted a one-tailed test, the alternative hypothesis would state that the portfolio's returns are either less than or greater than the S&P 500's returns.

The p-value hypothesis test does not necessarily make use of a preselected confidence level at which the investor should reset the null hypothesis that the returns are equivalent. Instead, it provides a measure of how much evidence there is to reject the null hypothesis. The smaller the p-value, the greater the evidence against the null hypothesis.

Thus, if the investor finds that the p-value is 0.001, there is strong evidence against the null hypothesis, and the investor can confidently conclude the portfolio's returns and the S&P 500's returns are not equivalent.

Although this does not provide an exact threshold as to when the investor should accept or reject the null hypothesis, it does have another very practical advantage. P-value hypothesis testing offers a direct way to compare the relative confidence that the investor can have when choosing among multiple different types of investments or portfolios relative to a [benchmark](#) such as the S&P 500.

For example, for two portfolios, A and B, whose performance differs from the S&P 500 with p-values of 0.10 and 0.01, respectively, the investor can be much more confident that portfolio B, with a lower p-value, will actually show consistently different results.

Is a 0.05 P-Value Significant?

A p-value less than 0.05 is typically considered to be statistically significant, in which case the null hypothesis should be rejected. A p-value greater than 0.05 means that deviation from the null hypothesis is not statistically significant, and the null hypothesis is not rejected.

What Does a P-Value of 0.001 Mean?

A p-value of 0.001 indicates that if the null hypothesis tested were indeed true, there would be a one in 1,000 chance of observing results at least as extreme. This leads the observer to reject the null hypothesis because either a highly rare data result has been observed, or the null hypothesis is incorrect.

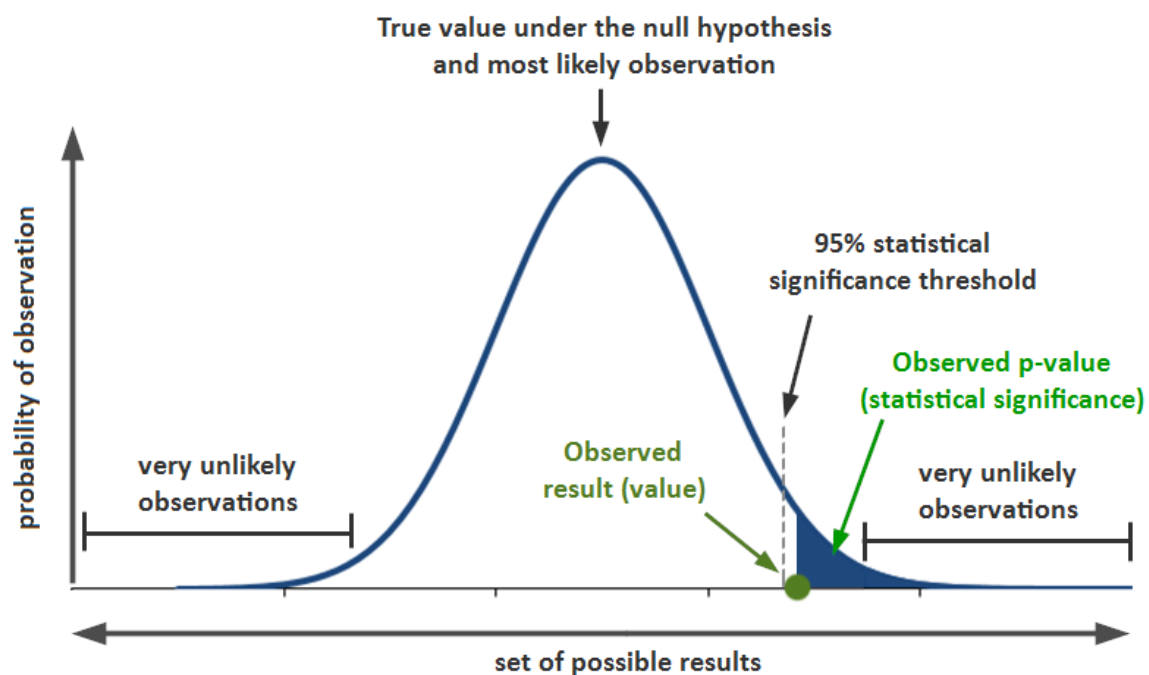
How Can You Use P-Value to Compare Two Different Results of a Hypothesis Test?

If you have two different results, one with a p-value of 0.04 and one with a p-value of 0.06, the result with a p-value of 0.04 will be considered more statistically significant than the p-value of 0.06. Beyond this simplified example, you could compare a 0.04 p-value to a 0.001 p-value. Both are statistically significant, but the 0.001 example provides an even stronger case against the null hypothesis than the 0.04.

The Bottom Line

The p-value is used to measure the significance of observational data. When researchers identify an apparent relationship between two variables, there is always a possibility that this correlation might be a coincidence. A p-value calculation helps determine if the observed relationship could arise as a result of chance.

Probability & Statistical Significance Explained



Significance Level		Specification
$p > 0.05$		not significant
$p \leq 0.05$	(5%)	significant
$p \leq 0.01$	(1%)	very significant
$p \leq 0.001$	(0.1%)	highly significant