LECTURE 5

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.

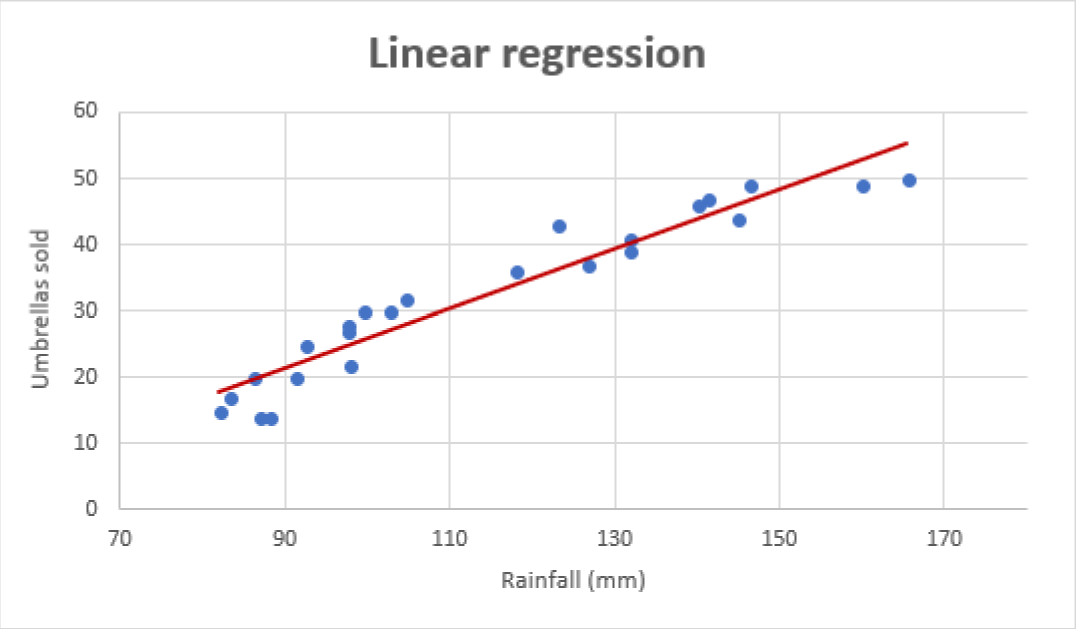
Independent variables (aka *explanatory* variables, or *predictors*) are the factors that might influence the dependent variable.

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 + ε

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

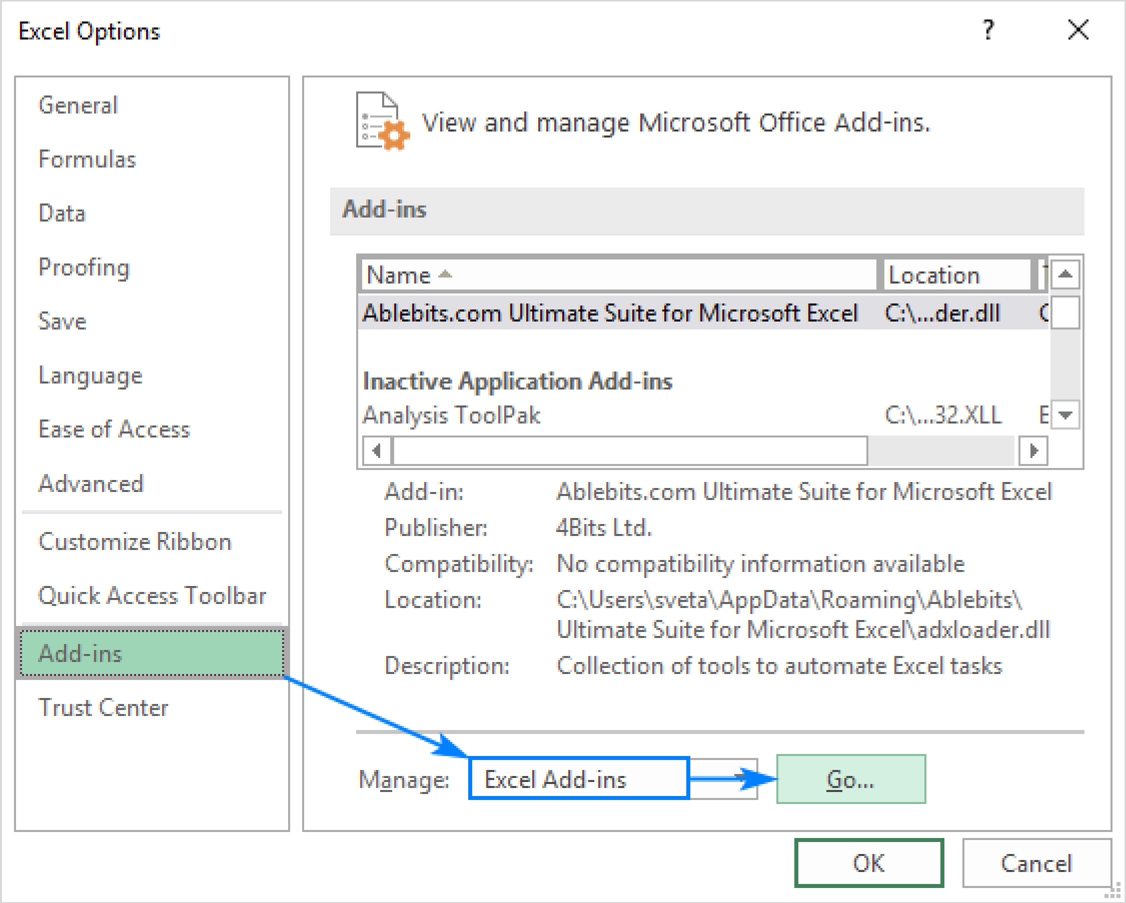
This example shows how to run regression in Excel by using a special tool included with the Analysis ToolPak add-in.

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*.



View and manage Microsoft Office Add-ins.

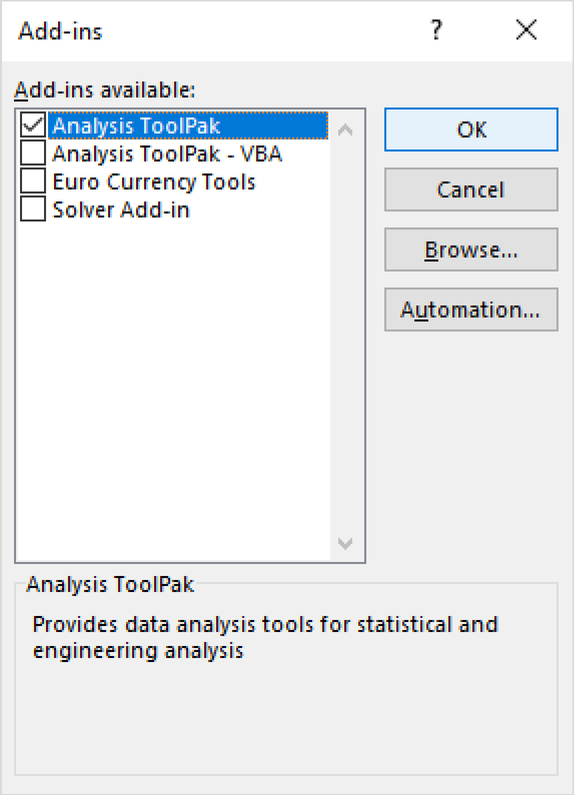
Add-ins

Name	Location	T
Ablebits.com Ultimate Suite for Microsoft Excel	C:\...der.dll	C
Inactive Application Add-ins		
Analysis ToolPak	C:\...32.XLL	E

Add-in: Ablebits.com Ultimate Suite for Microsoft Excel
Publisher: 4Bits Ltd.
Compatibility: No compatibility information available
Location: C:\Users\sveta\AppData\Roaming\Ablebits\Ultimate Suite for Microsoft Excel\adxloader.dll
Description: Collection of tools to automate Excel tasks

Manage: Excel Add-ins

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



Add-ins available:

- ☑ Analysis ToolPak
- □ Analysis ToolPak - VBA
- □ Euro Currency Tools
- □ Solver Add-in

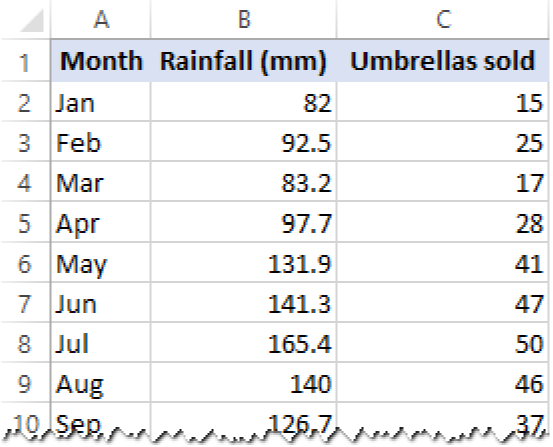
Analysis ToolPak

Provides data analysis tools for statistical and engineering analysis

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

Run regression analysis

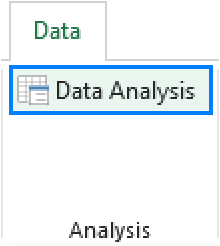
In this example, we are going to do a simple linear regression in Excel. What we have is a list of average monthly rainfall for the last 24 months in column B, which is our independent variable (predictor), and the number of umbrellas sold in column C, which is the dependent variable. Of course, there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



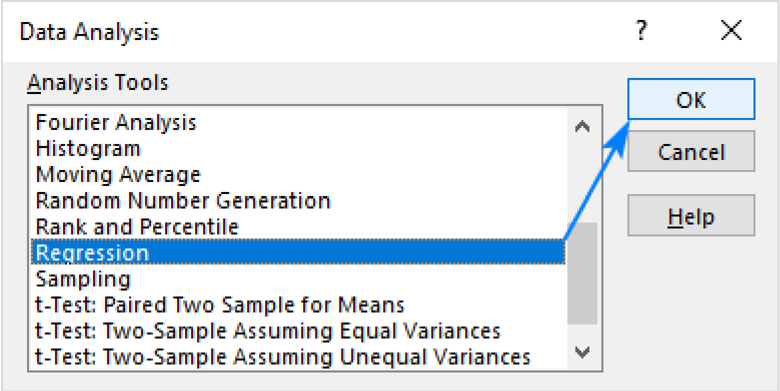
	A	B	C
1	Month	Rainfall (mm)	Umbrellas sold
2	Jan	82	15
3	Feb	92.5	25
4	Mar	83.2	17
5	Apr	97.7	28
6	May	131.9	41
7	Jun	141.3	47
8	Jul	165.4	50
9	Aug	140	46
10	Sep	126.7	37

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*.

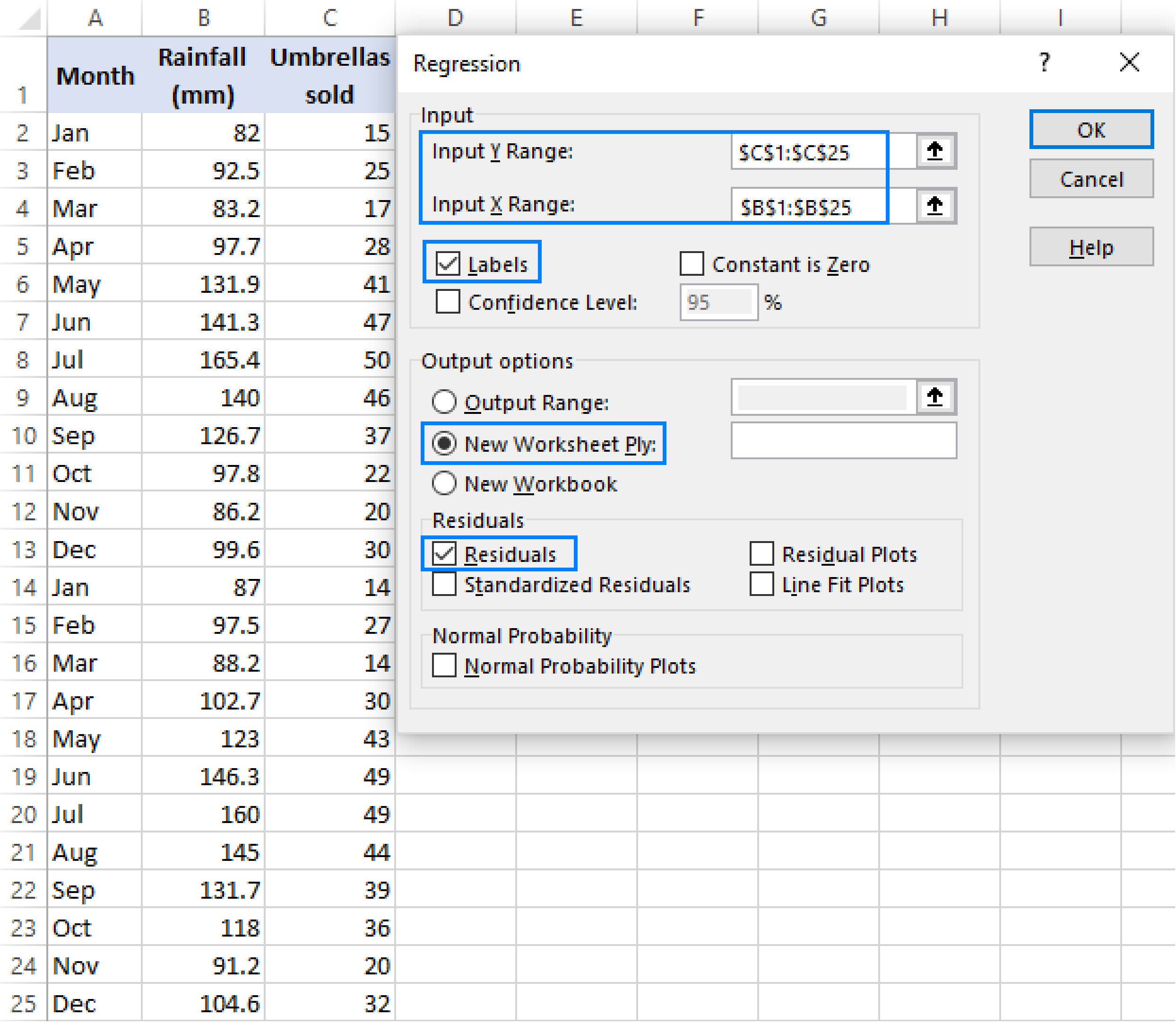


3. In the *Regression* dialog box, configure the following settings:
   - Select the *Input Y Range*, which is your **dependent variable**. In our case, it's umbrella sales (C1:C25).
   - Select the *Input X Range*, i.e. your **independent variable**. In this example, it's the average monthly rainfall (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: $C$1:$C$25

Input X Range: $B$1:$B$25

☑ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

☑ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

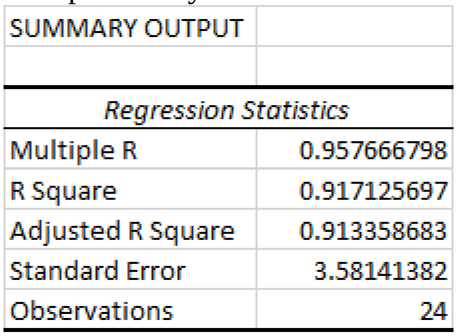
4. Click *OK* and observe the regression analysis output created by Excel.

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 equation fits your source data.



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

Here's what each piece of information means:

Multiple R. It is the C*orrelation 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 fairy 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.

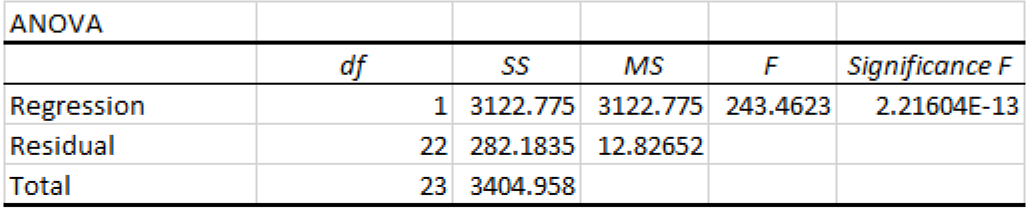
Adjusted R Square. It is the *R square* adjusted for the number of independent variable in the model. You will want to use this value instead of *R square* for multiple regression analysis.

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					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3122.775	3122.775	243.4623	2.21604E-13
Residual	22	282.1835	12.82652		
Total	23	3404.958			

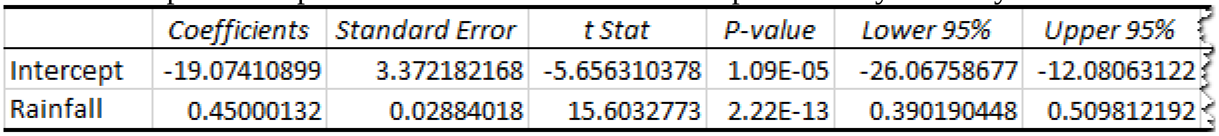
Basically, it splits the sum of squares into individual components that give information about the levels of variability within your regression model:

- *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:



	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
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 = Rainfall Coefficient * x + 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
```

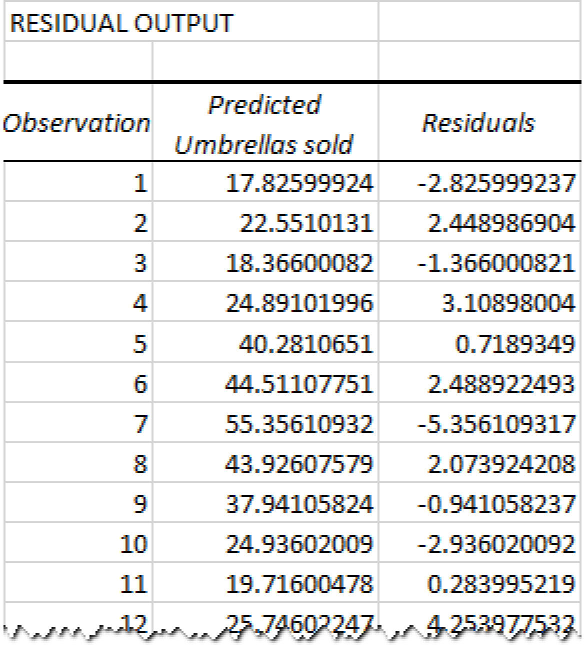
In a similar manner, you can find out how many umbrellas are going to be sold with any other monthly rainfall (x variable) you specify.

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

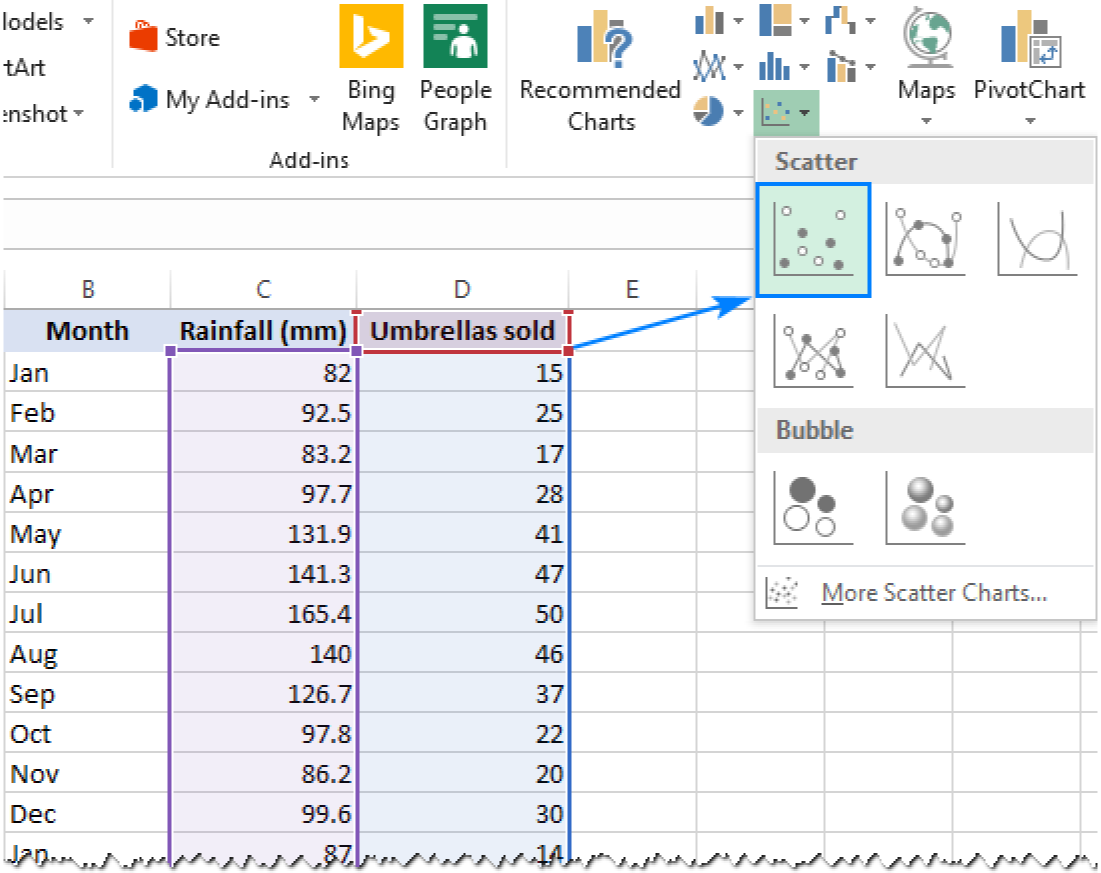
For the first data point (rainfall of 82 mm), the residual is approximately -2.8. So, we add this number to the predicted value, and get the actual value: 17.8 - 2.8 = 15.

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:

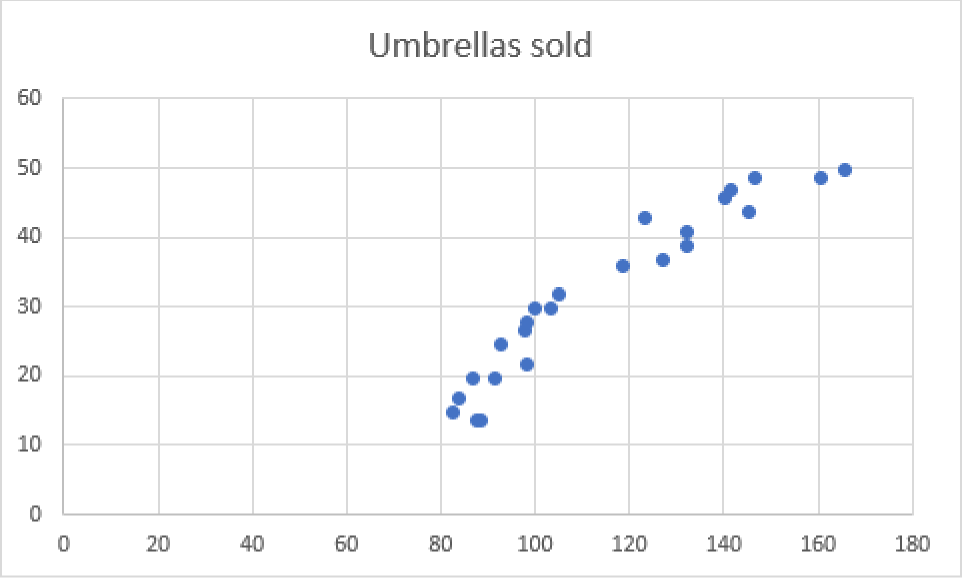
1. Select the two columns with your data, including headers.

2. On the *Inset* tab, in the *Chats* group, click the *Scatter chart* icon, and select the **Scatter** thumbnail (the first one):

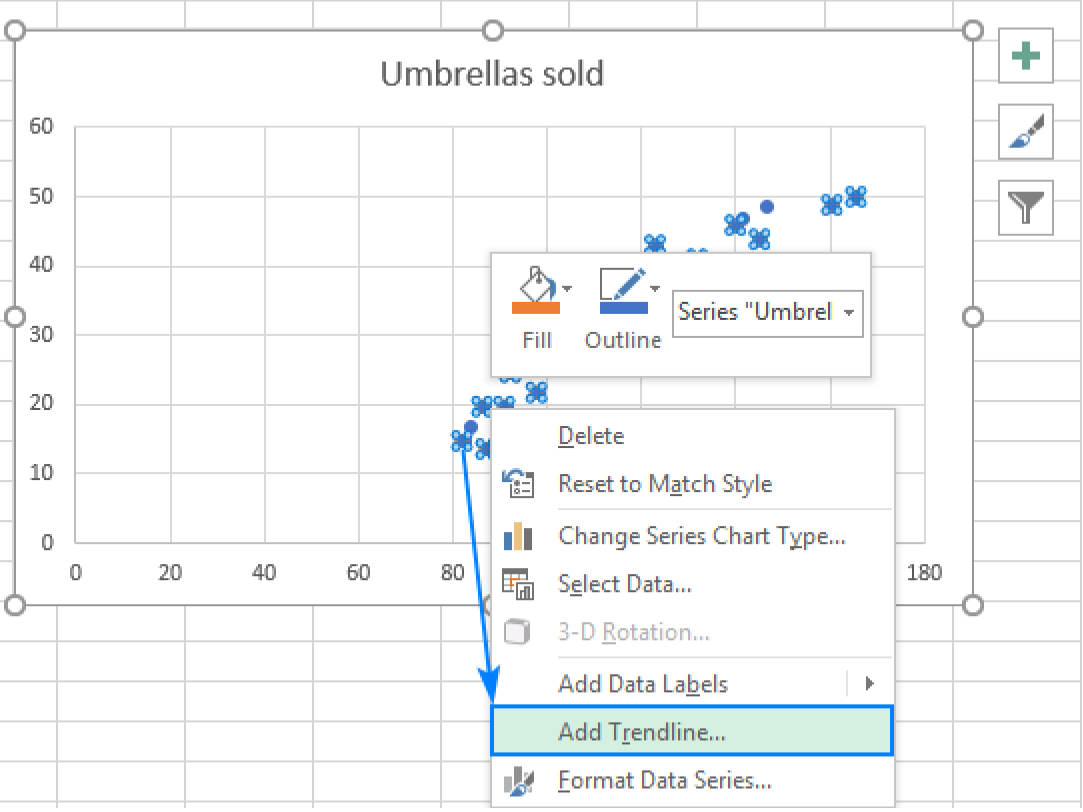


Month	Rainfall (mm)	Umbrellas sold
Jan	82	15
Feb	92.5	25
Mar	83.2	17
Apr	97.7	28
May	131.9	41
Jun	141.3	47
Jul	165.4	50
Aug	140	46
Sep	126.7	37
Oct	97.8	22
Nov	86.2	20
Dec	99.6	30
Jan	87	14

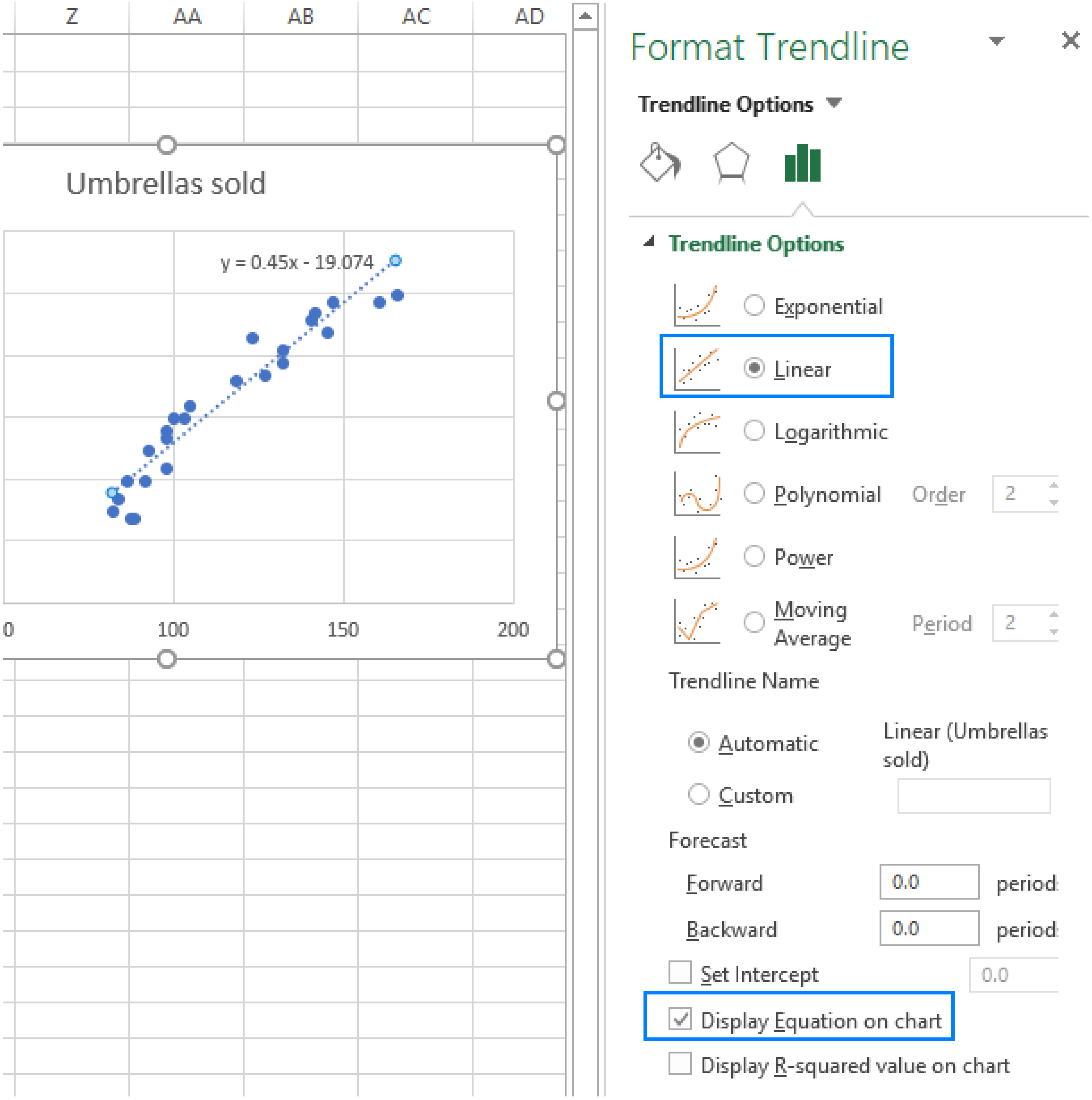
This will insert a scatter plot in your worksheet, which will resemble this one:



3. 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:



Format Trendline

Trendline Options

Trendline Options

○ Exponential

● Linear

○ Logarithmic

○ Polynomial Order 2

○ Power

○ Moving Average Period 2

Trendline Name

● Automatic Linear (Umbrellas sold)

○ Custom

Forecast

Forward 0.0 period

Backward 0.0 period

□ Set Intercept 0.0

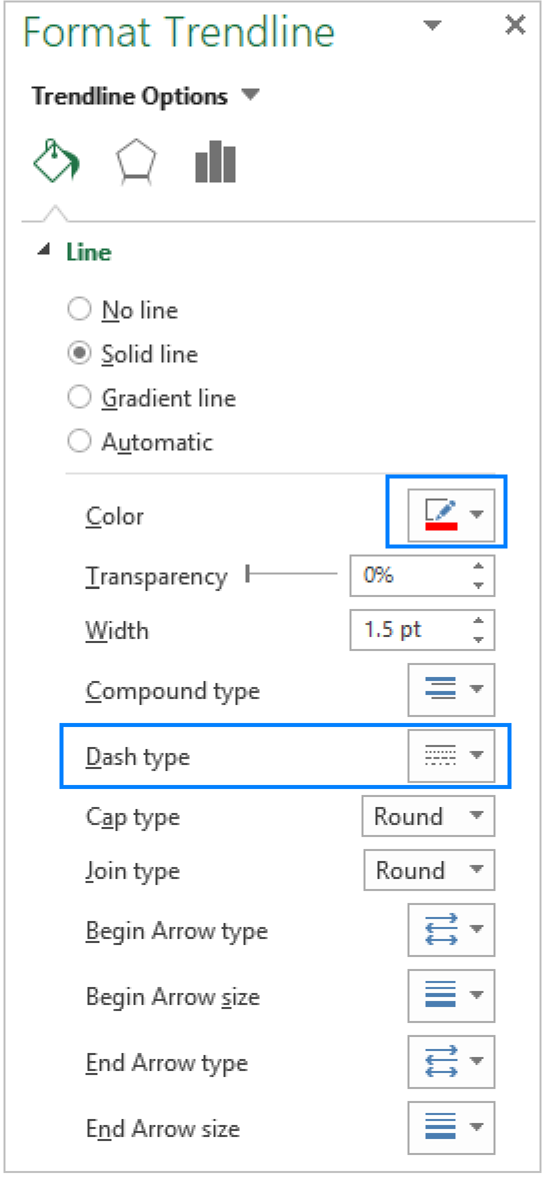
☑ Display Equation on chart

□ Display R-squared value on chart

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):



Format Trendline

Trendline Options

Line

○ No line
● Solid line
○ Gradient line
○ Automatic

Color

Transparency 0%

Width 1.5 pt

Compound type

Dash type

Cap type Round

Join type Round

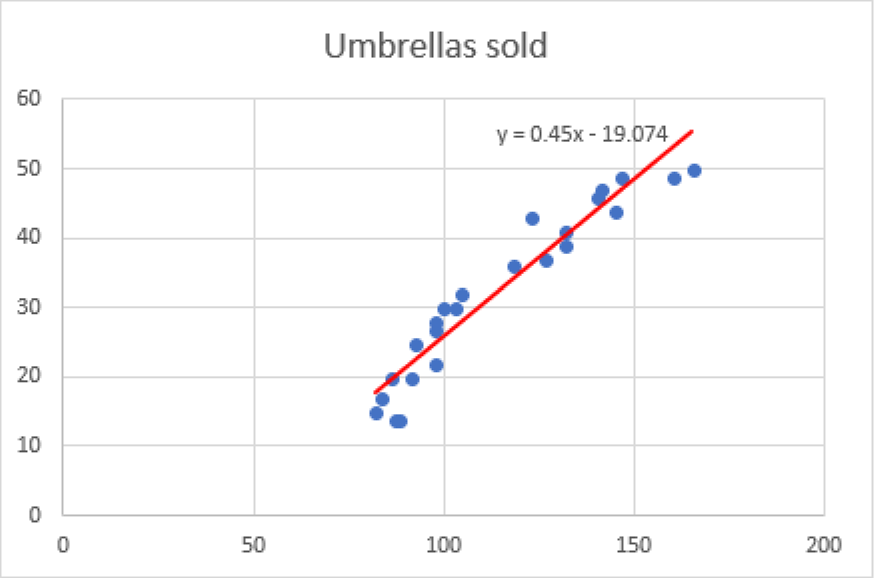
Begin Arrow type

Begin Arrow size

End Arrow type

End Arrow size

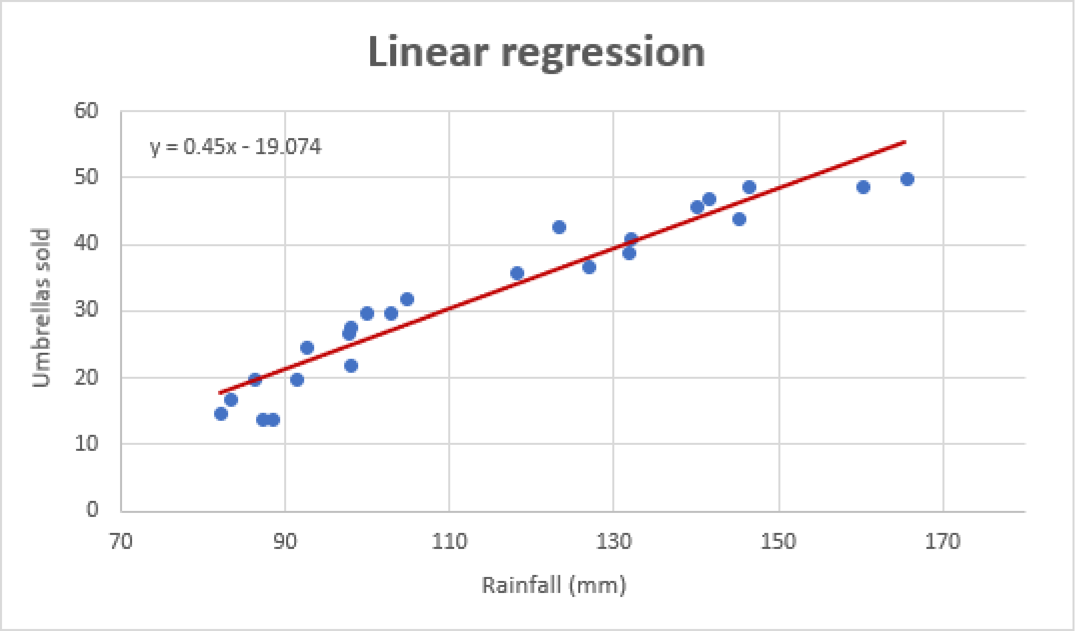
At this point, your chart already looks like a decent regression graph:



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, INTERCPET, and CORREL.

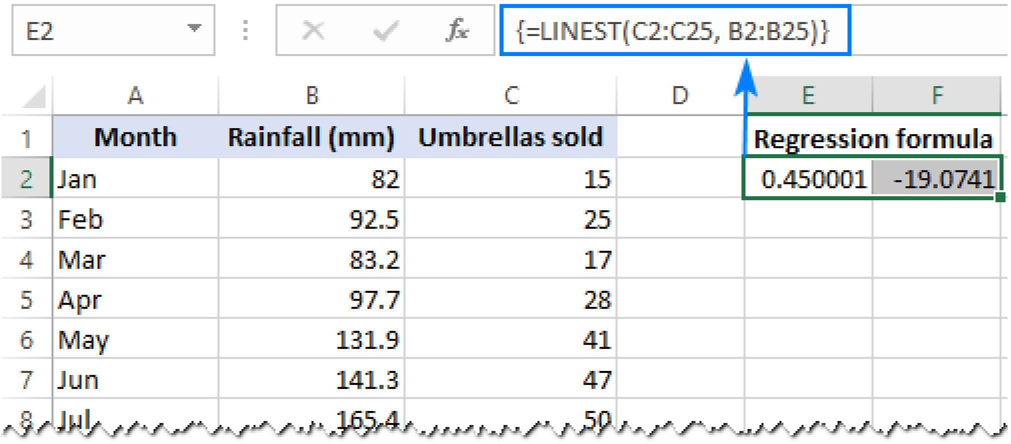
The LINEST function uses the least squares regression method to calculate a straight line that best explains the relationship between your variables and returns an array describing that line. You can find the detailed explanation of the function's syntax in this tutorial. For now, let's just make a formula 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:

```
y = bx + a
```



E2

{=LINEST(C2:C25, B2:B25)}

	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			

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)
```

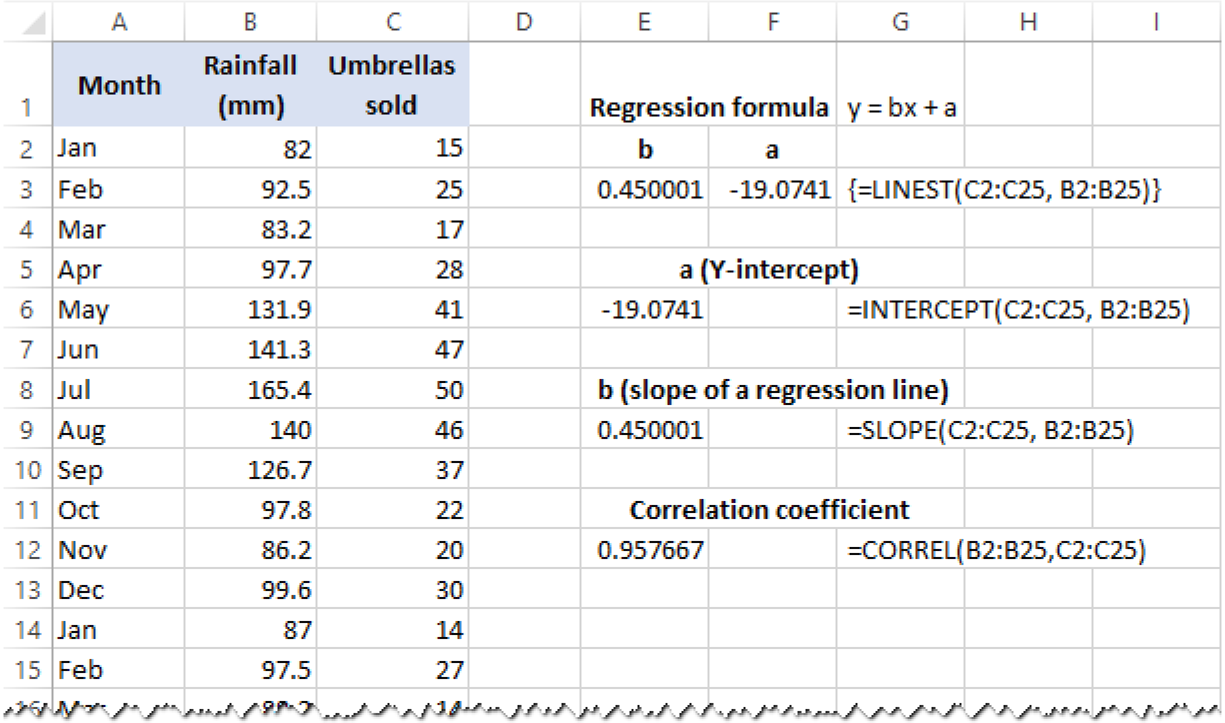
Get the slope (b):

```
=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						

Tip. If you'd like to get additional statistics for your regression analysis, use the LINEST function with the s*tats* parameter set to TRUE as shown in this example.

That's how you do linear regression in Excel. That said, please keep in mind that Microsoft Excel is not a statistical program. If you need to perform regression analysis at the professional level, you may want to use targeted software such as XLSTAT, RegressIt, etc.