

How to Easily Perform Simple Linear Regression in Excel

Authored by
stats writer

March 10, 2026

RECOMMENDED CITATION

stats writer (2026). *How to Easily Perform Simple Linear Regression in Excel*.
PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=134952>

Understanding the Fundamentals of Simple Linear Regression

Simple Linear Regression is a foundational **statistical method** used to quantify and model the relationship between two **continuous variables**. In a typical analysis, one variable is designated as the **independent variable** (also known as the explanatory or **predictor variable**), while the other is the **dependent variable** (often referred to as the **response variable**). The primary objective of this technique is to determine the extent to which changes in the independent variable can predict changes in the dependent variable by fitting a **linear equation** to the observed data. This mathematical approach allows researchers and analysts to identify trends, make predictions, and understand the strength of the association between data points.

In the context of **predictive modeling**, **Simple Linear Regression** assumes that the relationship between the two variables can be approximated by a straight line. This line, known as the **least squares regression line**, minimizes the sum of the squares of the vertical deviations between each data point and the line itself. By establishing this relationship, analysts can derive a **regression equation** that serves as a tool for forecasting future outcomes based on historical patterns. This method is widely applied across various fields, including **economics**, **social sciences**, and **business analytics**, where understanding the impact of one factor on another is critical for **decision-making**.

To successfully perform a **regression analysis**, certain assumptions must be met to ensure the validity of the results. These include **linearity**, where the relationship between the variables is indeed linear; **independence**, meaning the observations are independent of each other; **homoscedasticity**, where the **variance of residual errors** is constant across all levels of the independent variable; and **normality**, suggesting that the residuals of the model are **normally distributed**. When these conditions are satisfied, **Simple Linear Regression** provides a powerful and reliable means of extracting **actionable insights** from **raw data**.

Preparing Your Dataset for Analysis in Microsoft Excel

The first practical step in conducting **Simple Linear Regression** within **Microsoft Excel** involves the meticulous organization of your **dataset**. For the analysis to be effective, your data should be structured in a **tabular format**, typically consisting of two adjacent columns. One column must contain the values for the **independent variable** (X), while the parallel column contains the values for the **dependent variable** (Y). Ensuring that each row represents a single, paired observation is vital for maintaining **data integrity** and ensuring the **statistical software** interprets the relationship correctly.

Consider a scenario where an educator wishes to investigate the correlation between the number of hours a student spends studying and their subsequent performance on an examination. In this

instance, the "hours studied" serves as the **explanatory variable**, while the "exam score" functions as the **response variable**. Before inputting this information into **Excel**, it is advisable to perform **data cleaning** to remove any **outliers** or erroneous entries that might skew the **regression results**. A clean dataset is the cornerstone of any robust **statistical analysis**, as it prevents the introduction of **bias** into the model.

Once the data is refined, you should enter the values into an **Excel worksheet**. It is a best practice to include descriptive **column headers** at the top of each column, such as "Hours Studied" and "Exam Score." These labels not only make the spreadsheet easier to navigate but also allow **Excel** to automatically name the variables in the final **output report**. This organizational step simplifies the interpretation phase, as the generated **coefficients** and **statistics** will be clearly linked to their respective real-world concepts.

Step 1: Enter the data.

Enter the following data for the number of hours studied and the exam score received for 20 students:

	A	B	C	D	E
1	hours	score			
2	1	76			
3	2	78			
4	2	85			
5	4	88			
6	2	72			
7	1	69			
8	5	94			
9	4	94			
10	2	88			
11	4	92			
12	4	90			
13	3	75			
14	6	96			
15	5	90			
16	3	82			
17	4	85			
18	6	99			
19	2	83			
20	1	62			
21	2	76			
22					
23					
24					

Visualizing Relationships with Scatter Plots

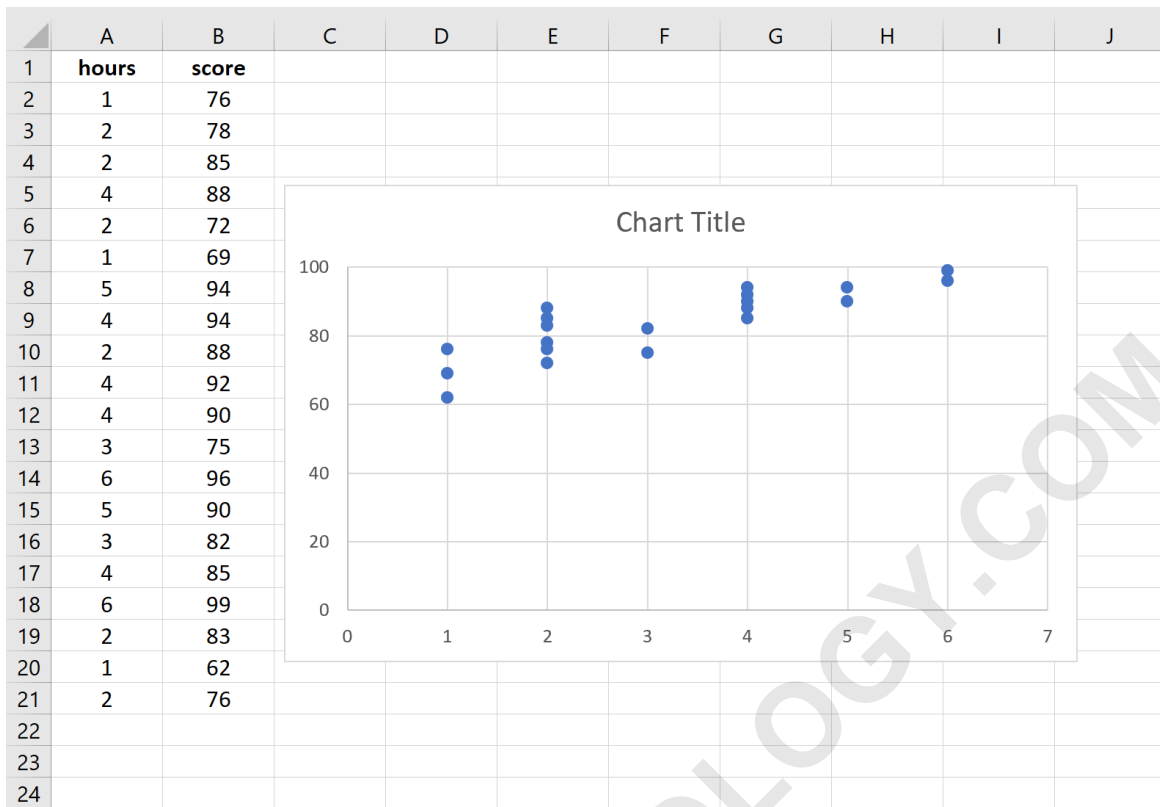
Before executing the formal **regression analysis**, it is highly beneficial to create a **scatter plot** to visually inspect the data. Visualization acts as a preliminary diagnostic tool, allowing you to determine if a **linear relationship** is plausible. If the data points appear to follow a general straight-line path, either upward or downward, then **Simple Linear Regression** is likely an appropriate model. Conversely, if the points form a curve or show no discernible pattern, more complex **non-linear models** or different **statistical techniques** may be required.

To generate this visual representation in **Excel**, you must first highlight the range of data in your columns. Navigating to the **Insert** tab on the **Excel ribbon**, you will find the **Charts** group, where you can select the **Scatter (X, Y)** chart type. Selecting the basic scatter option will instantly produce a graph where the **independent variable** is plotted along the **horizontal axis** (x-axis) and the **dependent variable** is plotted along the **vertical axis** (y-axis). This visual layout provides an intuitive sense of the **correlation** between the variables before any numbers are crunched.

Step 2: Visualize the data.

Before we perform simple linear regression, it's helpful to create a of the data to make sure there actually exists a linear relationship between hours studied and exam score.

Highlight the data in columns A and B. Along the top ribbon in Excel go to the **Insert** tab. Within the **Charts** group, click **Insert Scatter (X, Y)** and click on the first option titled **Scatter**. This will automatically produce the following scatterplot:



The number of hours studied is shown on the x-axis and the exam scores are shown on the y-axis. We can see that there is a linear relationship between the two variables - more hours studied is associated with higher exam scores. This positive **slope** indicates that as students invest more time in preparation, their scores tend to improve proportionally. To quantify the specific strength and mathematical nature of this relationship, we proceed to the formal **regression** calculation.

Activating and Utilizing the Data Analysis ToolPak

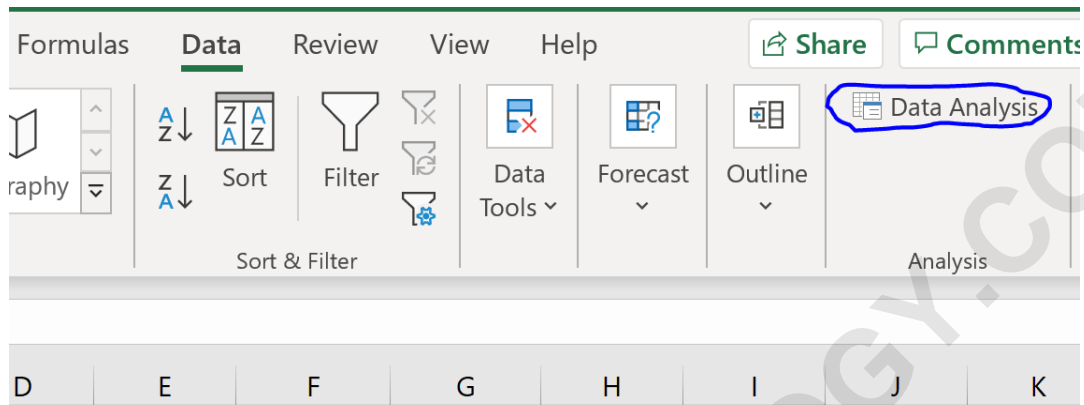
To perform advanced **statistical modeling** like **Simple Linear Regression**, **Excel** users typically rely on the **Analysis ToolPak**. This **add-in** provides a suite of **data analysis tools** that extend beyond the standard **Excel functions**. If the **Data Analysis** button is not visible under the **Data** tab of your ribbon, you must enable it through the **Excel Options** menu. This process involves navigating to "Add-ins," selecting "Excel Add-ins" from the Manage box, and checking the box for "Analysis ToolPak."

Once the **ToolPak** is activated, the **Data Analysis** command becomes a central hub for various **statistical tests**, including **ANOVA**, **correlation**, and **regression**. By clicking this button, a **dialog box** appears, presenting a list of analysis tools. Selecting **Regression** from this list initiates the process of defining the parameters for your specific **linear model**. This interface is designed to be user-friendly, guiding the analyst through the necessary inputs required to generate a

comprehensive **summary output**.

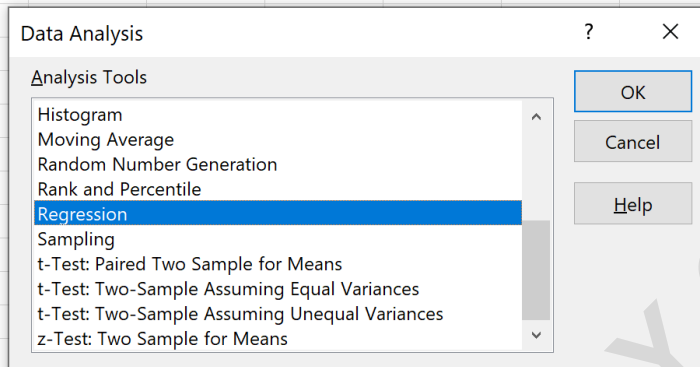
Step 3: Perform simple linear regression.

Along the top ribbon in Excel, go to the **Data** tab and click on **Data Analysis**. If you don't see this option, then you need to first load the **Analysis ToolPak** as described above.



Once you click on **Data Analysis**, a new window will pop up. Select **Regression** and click OK. This selection will open the primary configuration menu where you will define the **data ranges** and **confidence levels** for your study.

	A	B	C	D	E	F	G	H
1	hours	score						
2	1	76						
3	2	78						
4	2	85						
5	4	88						
6	2	72						
7	1	69						
8	5	94						
9	4	94						
10	2	88						
11	4	92						
12	4	90						
13	3	75						
14	6	96						
15	5	90						
16	3	82						
17	4	85						
18	6	99						
19	2	83						
20	1	62						
21	2	76						
22								
23								
24								



Configuring the Regression Analysis Parameters

In the **Regression dialog box**, the configuration of **input ranges** is the most critical step for accuracy. The **Input Y Range** field requires the selection of the cells containing your **dependent variable** (the outcomes you are trying to predict), while the **Input X Range** requires the selection of the cells containing the **independent variable**. It is essential to ensure that both ranges contain the same number of rows; otherwise, **Excel** will return an error indicating that the **input ranges** are mismatched.

If you included **column headers** during the data entry phase, you must check the **Labels** box. This tells the **software** to treat the first row of your selection as names rather than **numeric data**, which prevents **calculation errors** and ensures the final **regression table** is clearly labeled. Additionally, you have the option to set the **confidence level**, which defaults to 95%. This level determines the **confidence intervals** for your **coefficients**, providing a range within which the true population parameters are likely to fall.

Finally, you must specify where the results should be displayed. The **Output Range** option allows you to select a specific cell on the current worksheet for the report, while the "New Worksheet Ply" option will generate the results on a separate tab. For comprehensive analysis, you may also

choose to check boxes for **Residuals**, **Standardized Residuals**, or **Residual Plots**. These options provide deeper insights into the **model fit** and help identify whether the **linear regression assumptions** have been met. Once configured, clicking **OK** executes the **algorithm**.

For **Input Y Range**, fill in the array of values for the response variable. For **Input X Range**, fill in the array of values for the explanatory variable. Check the box next to **Labels** so Excel knows that we included the variable names in the input ranges. For **Output Range**, select a cell where you would like the output of the regression to appear. Then click **OK**.

	A	B	C	D	E	F	G	H	I
1	hours	score							
2	1	76							
3	2	78							
4	2	85							
5	4	88							
6	2	72							
7	1	69							
8	5	94							
9	4	94							
10	2	88							
11	4	92							
12	4	90							
13	3	75							
14	6	96							
15	5	90							
16	3	82							
17	4	85							
18	6	99							
19	2	83							
20	1	62							
21	2	76							
22									
23									
24									
25									
26									

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

The following output will automatically appear, containing several tables that detail the **statistical properties** of your regression model:

D	E	F	G	H	I	J	K	L
SUMMARY OUTPUT								
<i>Regression Statistics</i>								
Multiple R	0.8528							
R Square	0.7273							
Adjusted R Square	0.7121							
Standard Error	5.2805							
Observations	20							
ANOVA								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	1	1338.2906	1338.2906	47.9952	0.0000			
Residual	18	501.9094	27.8839					
Total	19	1840.2000						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	67.1617	2.6633	25.2178	0.0000	61.5664	72.7570	61.5664	72.7570
hours	5.2503	0.7578	6.9279	0.0000	3.6581	6.8424	3.6581	6.8424

Deciphering the Summary Output: Goodness of Fit

The **Summary Output** provided by **Excel** is divided into several sections, the first of which is the **Regression Statistics** table. This section evaluates the **goodness of fit** of the model. The **R Square** (or **coefficient of determination**) is a vital metric that indicates the proportion of the **variance** in the dependent variable that is predictable from the independent variable. An **R Square** value ranges from 0 to 1, with higher values indicating a stronger relationship where the model explains more of the data's variability.

In our student study example, the **R Square** value is **0.7273**. This implies that approximately 72.73% of the **variation** in exam scores can be attributed to the amount of time spent studying. While this is a high percentage, it also suggests that the remaining 27.27% of the variation is due to other factors not included in the **simple model**, such as prior knowledge, **intelligence quotient**, or testing conditions. Understanding this limitation is key to responsible **data interpretation**.

Another important figure in this table is the **Standard Error** of the regression. This value represents the average distance that the observed **data points** fall from the **fitted line**. A smaller **Standard Error** indicates that the data points are closer to the **regression line**, suggesting a more precise model. In this specific analysis, the **Standard Error** is **5.2805**, meaning that the predicted exam scores typically deviate from the actual scores by about 5.28 points. This provides a measure of the **model's accuracy** in its **predictions**.

Analyzing Statistical Significance and the ANOVA Table

The second major section of the output is the **ANOVA** (Analysis of Variance) table. This table tests the **statistical significance** of the overall **regression model**. The **F-statistic** (47.9952) is used to determine whether the observed relationship between variables is likely to have occurred by **random chance**. A high **F-value** typically suggests that the model provides a better fit than a model with no **predictor variables**.

The **Significance F** value, which is essentially the **p-value** for the entire model, is the most critical number in this section. In this case, the **p-value** is **0.0000** (or a value very close to zero). In **statistical hypothesis testing**, a p-value less than the standard **alpha level** of 0.05 indicates that the results are **statistically significant**. This means we can reject the **null hypothesis** that there is no relationship between study hours and exam performance, concluding instead that the time spent studying has a meaningful impact on scores.

By confirming that the **Significance F** is below 0.05, we gain confidence that the **independent variable** is indeed a useful predictor. Without this **statistical validation**, any patterns observed in the data might be dismissed as mere **coincidence**. The **ANOVA** table thus serves as a gatekeeper, ensuring that the model is robust enough to justify further interpretation of the **regression coefficients** and their practical implications.

Interpreting Regression Coefficients and the Intercept

The final section of the **regression output** provides the **coefficients** necessary to construct the **linear equation**. The **Intercept** value (67.16) represents the predicted value of the dependent variable when the independent variable is zero. In our example, this suggests that a student who studies for zero hours would be expected to receive an exam score of **67.16**. This serves as the starting point of the **regression line** on the y-axis.

The coefficient for the **independent variable** (Hours) is **5.2503**. This value represents the **slope** of the regression line. In **Simple Linear Regression**, the slope indicates the average change in the dependent variable for every one-unit increase in the independent variable. Here, for each additional hour a student studies, their exam score is expected to increase by **5.2503** points. This **quantitative measure** of impact is the primary output used for **forecasting** and **strategic planning**.

Together, these values allow us to formulate the **estimated regression equation: exam score = 67.16 + 5.2503 * (hours)**. This formula is a powerful tool for **extrapolation** and **interpolation**. By plugging in different values for "hours," we can estimate the "exam score" for any student within the range of our study. This mathematical representation simplifies complex **data relationships** into a single, actionable line of code or logic.

Practical Application: Making Predictions with the Regression Model

The ultimate goal of performing **Simple Linear Regression** is often to use the resulting model for **predictive analysis**. Once the **regression equation** is established, it can be applied to new data points to estimate **unobserved outcomes**. For instance, if a student plans to study for exactly three hours, we can substitute this value into our equation: $67.16 + 5.2503 * (3)$. The calculation yields an expected exam score of **82.91**, providing a clear **quantitative prediction**.

This predictive capability is invaluable in **academic advising**, **business forecasting**, and **scientific research**. However, it is important to remember that these predictions are **probabilistic** rather than deterministic. The model provides the "average" expected result, but individual results will vary due to **random error** and variables not included in the model. This is why the **Standard Error** and **confidence intervals** mentioned earlier are so important--they provide a "margin of error" for our **forecasts**.

In addition to point predictions, **Excel** users can use the regression results to understand the **marginal utility** of an action. In this case, the model clearly shows that studying has a positive **return on investment**. By understanding the **slope**, stakeholders can make informed decisions about resource allocation--such as deciding whether the benefit of an extra hour of study (approximately 5.25 points) outweighs the **opportunity cost** of that time.

Advanced Diagnostics and Next Steps in Statistical Modeling

While **Simple Linear Regression** is a robust tool, it is often just the beginning of a more thorough **statistical investigation**. To ensure the model is truly reliable, analysts frequently perform **diagnostic tests**. One common method is creating a **residual plot**, which graphs the differences between the observed and predicted values. A well-fitted model will show **residuals** that are randomly scattered around zero, whereas patterns in the residuals might suggest that a **non-linear model** or **multiple regression** would be more appropriate.

For those looking to deepen their analysis, **Excel** offers capabilities for more complex tasks. This includes calculating **prediction intervals**, which provide a range for an individual observation rather than just a mean, and generating **Q-Q plots** to verify the **normality assumption**. As datasets grow in complexity, you may also explore **Multiple Linear Regression**, which incorporates several independent variables to explain the variation in a single dependent variable, providing a more comprehensive view of the **causal factors** involved.

The following tutorials explain how to perform other common tasks in Excel to refine your **statistical analysis** and ensure the highest levels of **accuracy** and **validity** in your **data-driven insights**:

[How to Create a Residual Plot in Excel](#)

[How to Construct a Prediction Interval in Excel](#)

[How to Create a Q-Q Plot in Excel](#)

ARABPSYCHOLOGY.COM