Regression analysis in Excel - the basics

After this practice, you should be able to answer the following questions:

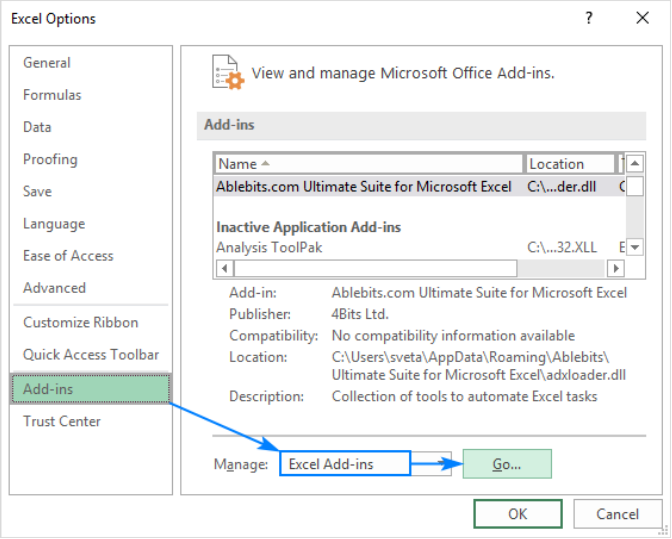
Question 1: For **April 2023**, the Bureau of Meteorology has predicted the average amount of monthly rainfall will be 95mm. Based on this, can you predict the number of umbrellas the shop will sell?

Question 2: For **December 2023**, Bureau of Meteorology has predicted the average amount of monthly rainfall will be 40 mm which is the lowest in last 24 months. Based on this, can you predict the number of umbrellas the shop will sell? How much confident you are with this prediction? Discuss your answer.

Step 1: Enable the Analysis ToolPak add-in

In your Excel, click File > Options.

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.

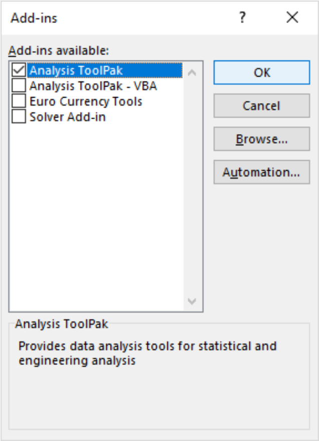
Add-ins

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

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

1. 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.

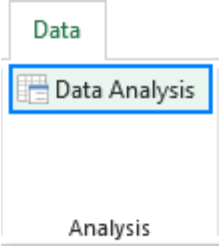
Step 2: Run regression analysis

In this example, we are going to do a simple linear regression in Excel. Please download the **excel-regression-analysis.xlsx** from your Moodle account. In the file, there are three columns.

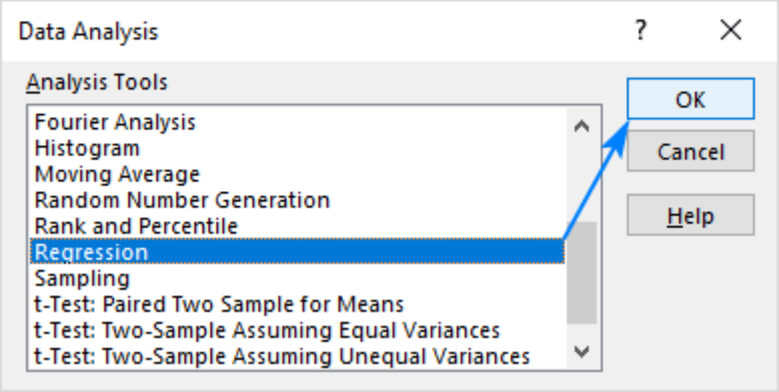
- Column A: Name of the Month
- Column B: Average monthly rainfall for the last 24 months (independent variable/predictor)
- Column C: The number of umbrellas sold (dependent variable)

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*.



Data Analysis ? ×

Analysis Tools

- Fourier Analysis
- Histogram
- Moving Average
- Random Number Generation
- Rank and Percentile
- Regression
- Sampling
- t-Test: Paired Two Sample for Means
- t-Test: Two-Sample Assuming Equal Variances
- t-Test: Two-Sample Assuming Unequal Variances

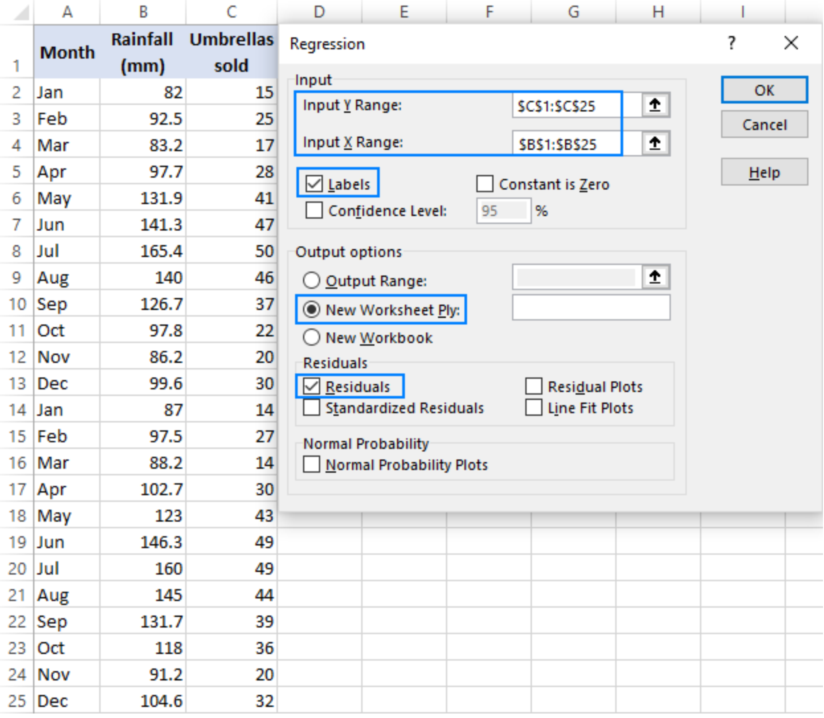
3. In the *Regression* dialog box, configure the following settings:

- 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

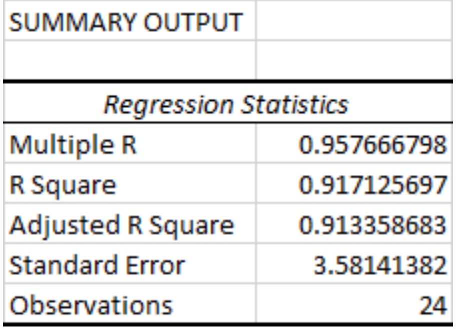
□ Normal Probability Plots

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

Step 3: Interpret 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.

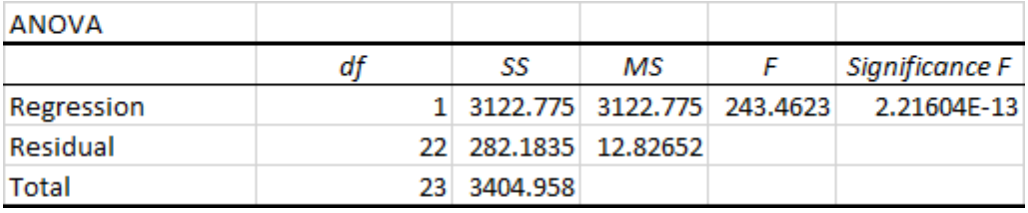
Adjusted R Square. It is the *R square* adjusted for the number of independent variable in the model. You will want to use this value instead of *R square* for multiple regression analysis.

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			

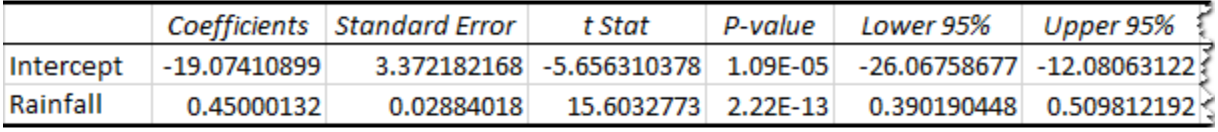
Basically, it splits the sum of squares into individual components that give information about the levels of variability within your regression model:

- *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

Share your result with the class.