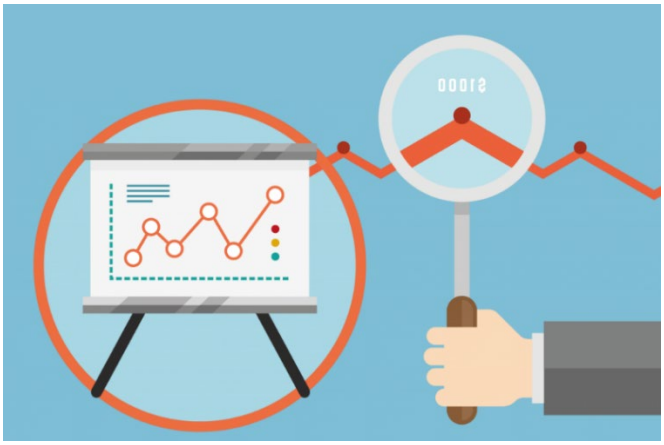
Activity – Regression Model



In this activity, you will learn to build a regression model for prediction. This activity requires the Microsoft Excel software with the Analysis Toolpak installed.

To install the Analysis Tookpak, follow the steps below:

1. Click the **Data** menu to check if you see an "Data Analysis" ribbon at the top right-hand corner of the menu. If not, go to step 2.
2. Click "**File**" → "**Options**", select "**Add-ins**". At the bottom of the dialogue window, check that "**Excel Add-ins**" is selected in the field "**Manage:**"
3. Highlight "**Analysis ToolPak**" and click "**OK.**"

(Sometimes, the 'Data Analysis' ribbon may 'disappear' in Excel. If it happens, uncheck the 'Analysis ToolPak' and click 'OK.' Then reenable the 'Analysis ToolPak' again by following steps one to three above.)

A. Background

Imagine that you and a friend own a small ice cream shop. You discuss how many kilograms (kg) of ice cream to produce each day, and you both agree that the hotter the weather is, the more ice cream will be sold. You add that this is not the only factor to consider, but other variables can also affect the number of sales. You decide to run a small experiment by recording the **mean temperature** during the shop opening hours and the **amount of ice cream sold**.

B. Download the dataset from the web.

1. Open Microsoft Excel with an empty worksheet.
2. In the menu, choose 'Data,' followed by 'From Web.' Enter the url, https://raw.githubusercontent.com/yongweefoo/intro_ai_ml/main/ice-cream.csv, into the field URL field and click 'OK.'

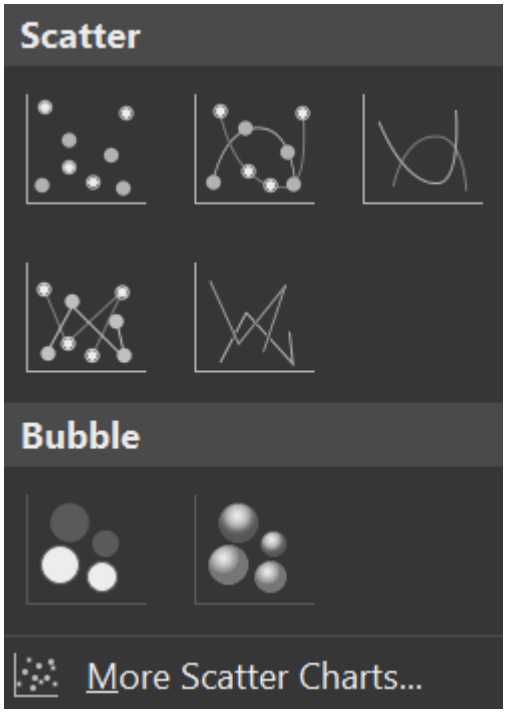
You should see a table that looks like this:

	A	B
	Mean temperature (°C)	Ice cream sold (kg)
	26	45
	23	42.5
	29	53.5
	23	35.5
	15	32.5
	19	34.5
	21	33.5
	18	35
	15	32.5
	25	40.5
	25	39.5
	16	32
	23	44.5
	23	39.5
	20	33
	17	26.5
	21	37.5
	29	49.5
	25	40.5
	24	44

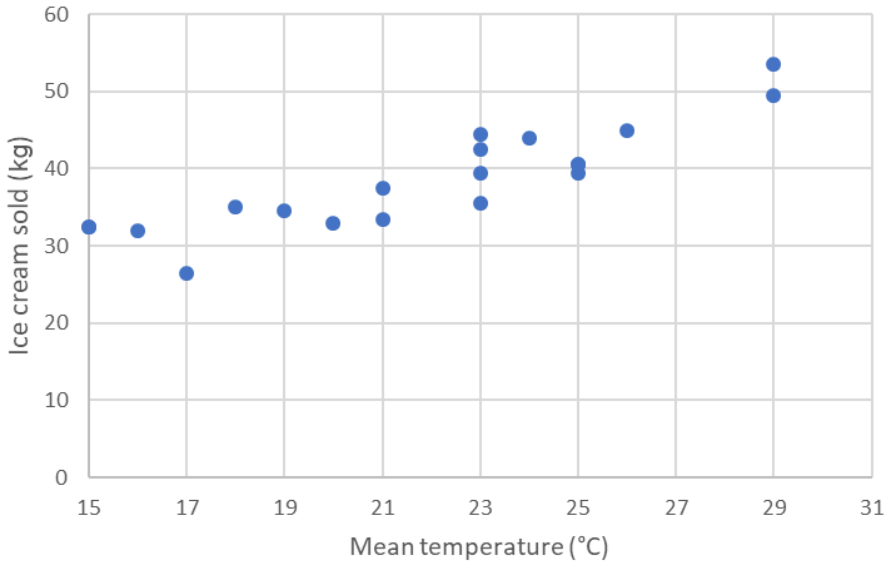
C. Visualize the dataset

1. Select the full range of cells containing the table, click on Insert menu, and select Charts:

2. Now, click on Scatter, as follows:



3. Name the axis titles, you should get a chart that looks like the one below.



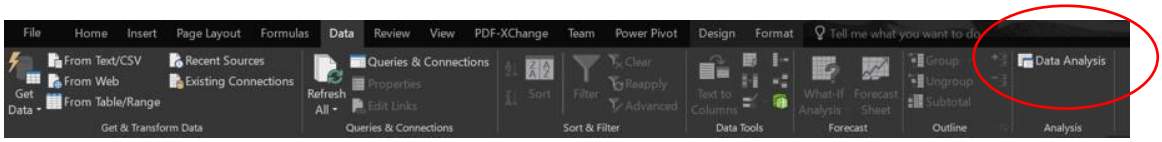
4. You can see that there is a linear correlation and that it is positive (the larger the temperature value, the more ice cream you sell). You can then represent the model using a linear equation, as follows:

$$IC = a * T + b \qquad (1)$$

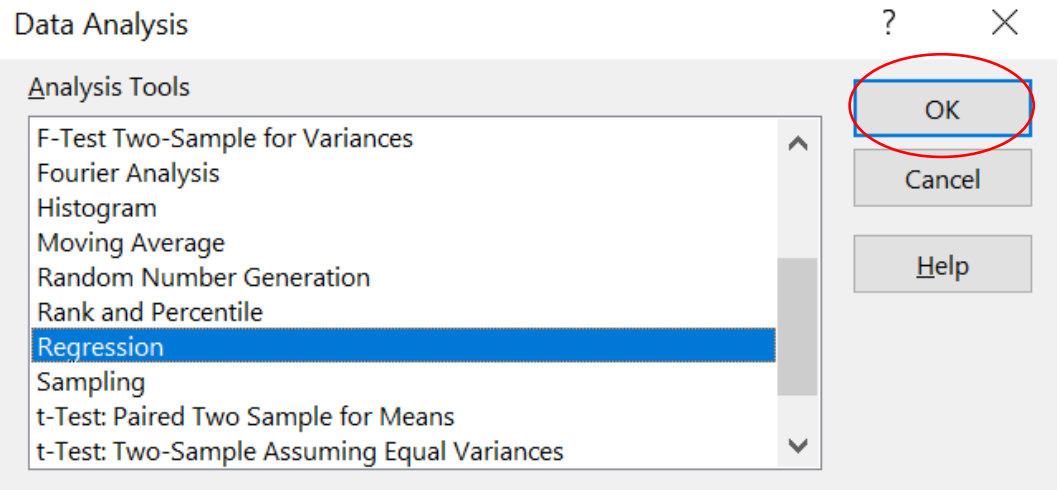
where, IC is the amount of ice cream sold, T is the mean temperature, and a and b are constant values to be calculated by a linear regression.y

To obtain the values of *a* and *b*, yuo can use Excel's Analysis ToolPak data analysis add-in.

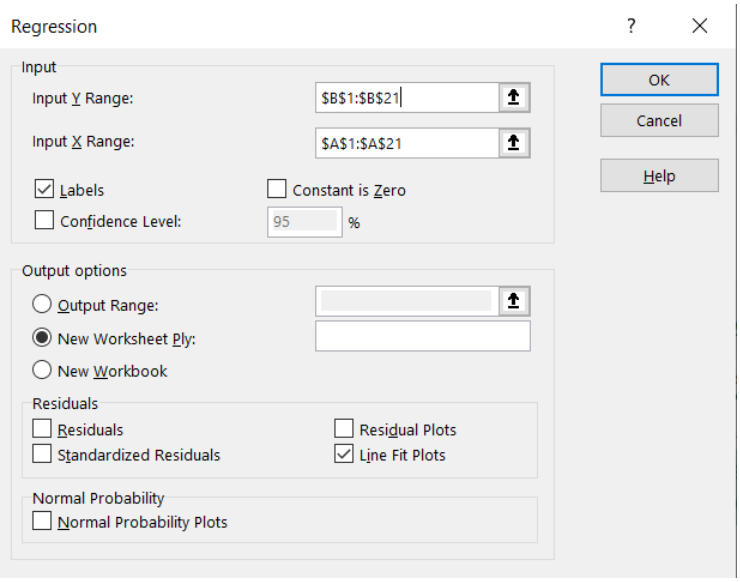
5. Go to 'Data' in the main menu and then select Data Analysis:



6. In the pop-up menu, select Regression and click on OK:



7. Ensure that the x and y ranges are correct (x is temperature and y is ice-cream amount). Check "Label" if your data includes the header. Select Line Fit Plots to see the regression line on top of the data points in a new diagram:



8. Analyze the results.

Q1 What is the Coefficient for the y-intercept?

_____

Q2 What is the Coefficient of the x-variables?

_____

Q3 How do you explain R square?

_____

Q4 How do you explain the *p-values*?

_____

Q5 See if you can come up with the predictive model with the regression formula.

_____

9. At the output, you can see that the line that best fits the data can be written as follows:

$$IC = 1.38 * T + 8.36 \qquad (2)$$

This means that for every one-degree rise in the temperature, there is a high probability that 1.38 kg more ice cream will get sold. If tomorrow's average temperature is going to be 28 degrees Celsius, how many kg of ice cream will be sold?

Regression Statistics	
Multiple R	0.882157
R Square	0.7782
Adjusted R Square	0.765878
Standard Error	3.202152
Observations	20

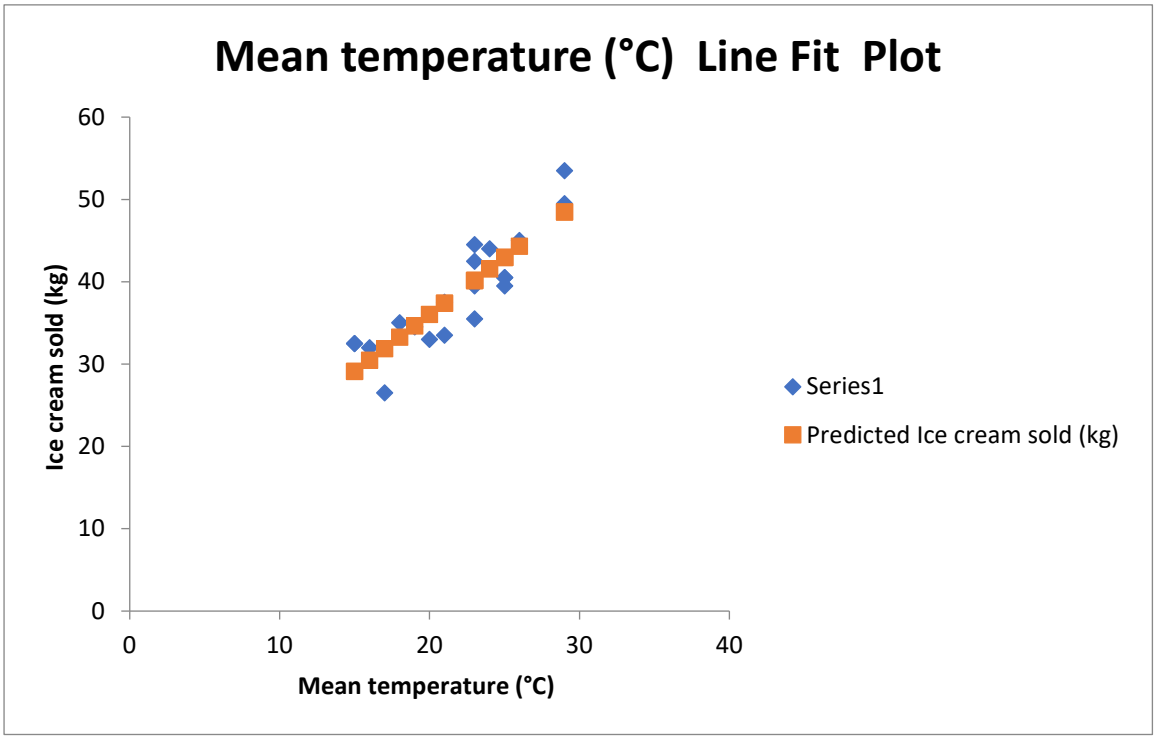
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	647.5695	647.5695	63.15423	2.69E-07
Residual	18	184.568	10.25378		
Total	19	832.1375			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	8.355782	3.869437	2.159431	0.044558	0.226396	16.48517	0.226396	16.48517
Mean temperature (°C)	1.383031	0.174033	7.946964	2.69E-07	1.017402	1.748659	1.017402	1.748659

10. There is a standard error for *a* of ±0.17, and for *b* of ±3.87. The R^2 value is 0.78, which means that the fit is not very good, and only 78% of the variation in ice cream sales can be explained by the

mean temperature. Significance F value is less than 0.01, the p-value of the independent variable is less than 0.05. So, you and your friend were both right! The following diagram shows the fitted line:



11. It is clear that the line represents the data quite well, but some points are slightly off, showing that you need to consider other factors when predicting ice cream consumption. In any case, given the mean forecasted temperature for one day, you can use equation (2) to have a rough estimation of how much ice-cream to produce to cover the possible demand.

D. Focusing on model features

Choosing significant features that provide relevant information about the phenomenon that we try to explain or predict is of paramount importance. For supervised learning, the most important features are those that highly correlate with the target variable – that is, the value that we want to predict or explain.

The quality of the insights obtained from a machine learning model depends on the features used as input to the model. Feature selection and feature engineering are regularly-used techniques to improve a model's input. Feature selection is the process of selecting a subset of relevant features for use in any identified model construction. It can also be termed as variable selection or attribute selection.

While building any machine learning model, feature selection and data cleaning should be the first and most important steps. Feature engineering is defined as using the domain knowledge of the identified data to create features that make the machine learning algorithm(s) work. sIf this is done correctly, it will increase the predictive power of machine learning algorithms by creating features from new data fed into this model or system.

In our previous example, the model features are the mean temperature and the amount of ice cream sold. Since we have already proved that more variables are involved, we could add additional features to explain the daily ice cream consumption better.

For example, we could consider which day of the week we are recording data for and include this information as another feature. Additionally, any other relevant information can be represented, more or less accurately, into a feature. In supervised learning, it is customary to call the input variables *features* and the target or predicted variable *label.*

Features can be numerical (such as the temperature in our previous example), or categorical (such as the day of the week). Since everything in computers is represented as numerical data, categorical data should be converted into numerical form by assigning categories to numbers. One-hot encoding is a process by which categorical variables are converted into a numerical form (or *encoded*) so that they can be input into machine learning algorithms.

Following our example, we could translate the day-of-the-week into day number, as follows:

Day-of-the-week	**Day number**
Monday	1
Tuesday	2
Wednesday	3
Thursday	4
Friday	5
Saturday	6
Sunday	7

This encoding reflects the order of the days and reserves the highest values for the weekend.

If the feature includes whether a day is a holiday (IsHoliday:Yes or No), one way of translating to numbers is to create binary variables. This approach is known as one-hot encoding and looks like the following table:

IsHoliday=No	**IsHoliday=Yes**
0	1
1	0

Let's say that you want to be more specific and predict the amount of ice cream for each flavor that you sell. For ease, let's say that you produce four different flavors: chocolate, strawberry, lemon, and vanilla. Could you just assign one number to each flavor, in the same way that you did in the day-of-the-week encoding?

Flavor	Flavor number
Chocolate	1
Strawberry	2
Lemon	3
Vanilla	4

Using this encoding, we implicitly say that chocolate is closer to strawberry than to vanilla (1 unit versus 3 units), which is not a real property of the flavors. The right way of translating to numbers is to create binary variables that looks like the following table:

Flavor	Is it chocolate?	Is it strawberry?	Is it lemon?	Is it vanilla?
Chocolate	1	0	0	0
Strawberry	0	1	0	0
Lemon	0	0	1	0
Vanilla	0	0	0	1

This method creates some overhead since it increases the number of features by creating one binary variable for each possible value of the original variable. On the positive side, it correctly calculates the properties of the feature.

E. machine learning models in practice

The number of features required to represent a real-life problem correctly is often large. The feature engineering techniques previously mentioned are impossible to perform by hand, so automatic methods must be devised and applied.

It is far more important to assess the predictive power of a combination of input features than the significance of each individual one.

It is very unlikely that we will get a good result with the first model you apply. Testing and evaluating many different machine learning models im- plies repeating the same steps several times and usually requires automation as well.

The dataset should be large enough to use a percentage of the data for training purposes (usually 80%) and the rest for testing. Evaluating the accuracy of a model only on the training data is misleading. A model can be very precise at explaining and predicting the training dataset, but it can fail to generalize and deliver wrong results when presented with new, previously unseen data values.

Training and test data should be selected, usually at random, from the same full dataset. Trying to make a prediction based on input that lies far away from the training range is unlikely to give good results.

--- End of Activity ---