Intro to Data Science
Linear Regression
Lab 1
Basics

1 Demonstrating Simple Linear Equation

File to use: `Linear Equation Demo`

Linear regression is based on finding a best-fitting line that is a linear equation that represents the relationship between two or more variables
A basic intro to linear equations for two variables: x and y

The `simple Linear Equation` worksheet demonstrates a basic linear equation for two variables: x and y, plotted as a scatter plot. Create a copy of it and play around with changing the key parameters of this equation (the value of b0 constant - the y-intercept and the value of b1 - the slope coefficient).

2 Demonstrating best fit line through Ordinary Least Squared Error (OLSE) method for linear regression

File to use: `Best-Fit-Demo`

The `Best-Fit-Demo-initial` worksheet shows the initial dataset and corresponding graph, whereas `Best-Fit-Demo-final` shows the final generated regression table

The dataset provides initial datapoints for the independent variable, Avg Temp, (values plotted on the X axis) and dependent variable, Sales (values plotted on the Y axis).

Try setting various values for the Y-intercept and Line Slope (using only the arrow buttons to increase / decrease the values), and view the linear equation line generated for these values (select on Show Fit Line box). For each particular combination of values, you can see the sum of Squared error, which sums up the square of the distance from each point on the linear equation line (the regression model) to the actual datapoint (select on Show Error Lines).

The intention is to try and obtain a combination of values for the Y-intercept and Line Slope, so that the Sum of Squared Errors (SSE) is the smallest possible. Try playing around with setting various values for the Y-intercept and Line Slope to get an intuition for what we are trying to achieve.

Finally, click on the Show Trend Line (uses the underlying Excel Trend Line functionality). This is the regression model line whose values for the Y-intercept and Line Slope will result in the SMALLEST possible value of SSE for all possible combinations of values for the Y-intercept and Line Slope (this should be around 330). This optimization is done by Excel's TREND statistical function in the background.

The trend line is a linear equation that is the simple linear regression model that best approximates the relationship between these two variables. The regression algorithm essentially creates this model when it is provided the table of dependent and independent variables as its input.

Try to set the values for the Y-intercept and Line Slope to be as close to these values as possible (using only the arrow buttons to increase / decrease the values).

Finally, to get predictions for this model (which will just simply be Y-values corresponding to the existing X-values using the regression model linear equation), simply enter the linear equation in the Prediction column and drag and drop.
In Cell E4, this should look like:
`=$H$2+($H$3*C4)`
Drag and drop down to fill up the remaining cells in the Prediction column with the same formula.

Now uncheck the Show Fit Line and click on the Show Forecast, which generates a linear graph line from the data in the newly populated Prediction column, and you will see this completely aligns with dotted Trend line we generated earlier, which is expected since both have nearly identical Y-Intercept and Line Slope values.

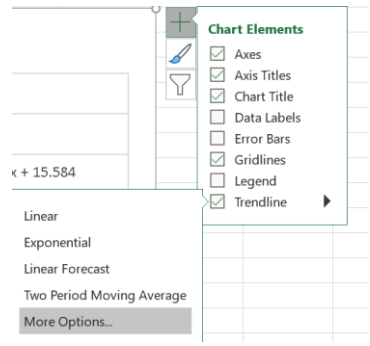
Finally, we can plug in new values for Avg Daily Temp box (the independent variable) and get predicted forecasts for the sales (the dependent variable) using the regression model linear equation.

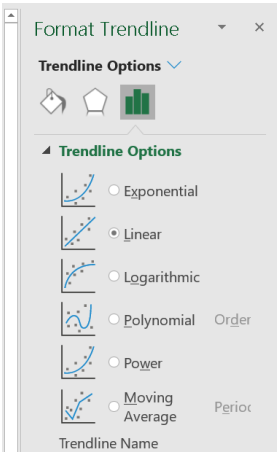
This essentially demonstrates how a regression model linear equation is generated for a simple case of a single independent variable and dependent variable.

The `Demo Different TrendLines` worksheet demonstrates simulating a series of values for a dependent (y-axis) variable by adding / subtracting some random numbers from the results of a linear

equation involving the independent (x-axis) variable. We can plot a scatter plot involving the new dependent (y-axis) variable and same independent (x-axis) variable, and generate a trendline for the scatter plot. Once again, the trend line will be the linear equation of the simple linear regression model that is the best approximate of the relationship between the dependent and independent variable.

Excel graphs offer multiple trendline options, here we are using the Linear Trendline for a linear regression model.





You can change the values of the upper limit and lower limit for the random number generated (which will also change every time the sheet is refreshed or saved) to view how different trend lines are generated for different distributions of random dependent variable values.

3 Sample Case Study for Multiple Linear Regression: Estimating and interpreting model coefficients

There is a Sales manager of a toys retail company which sells various kinds of toys in the local market. This sales manager would like to make some kind of projections / forecast about the number of monthly units that the retail company will be able to sell of this particular toy in the future quarter. In the past she has been making such projections based on her gut feeling and now wishes to utilize a data driven approach to decision making.

Based on her experience (domain knowledge is important in data science !), the manager figures out that the monthly unit sales (dependent variable) depends on 3 important variables (independent variables):

- the price at which the toy is sold (Price)
- the monthly amount that the company spends on advertising the toy (AdExp)
- the monthly amount spent on promotions for the toy (PromExp)

File to use: `Toy-Sales-pt1`

The `initial` worksheet shows the initial table, whereas `final-regression` shows the final generated regression table

You can make a copy from the initial worksheet and use that as a temporary worksheet and compare your final results with `final-regression`

We generate the regression model table using Excel's Analysis ToolPak:

Data ribbon tab -> select Data Analysis in the Analyze group
Highlight Regression from the Data Analysis dialog box, select ok
Complete the required info in the Regression Dialog box
Check Labels if you are going to select columns in the table with their headers included (recommended)
Select Input Y Range, then select the dependent / target / y variable (1 column)
Select Input X Range, then select the independent / feature / x variable (multiple columns)
Note: The X variable columns that you intend to use in your regression model must be stacked contiguously next to each other, if they are not you will have to rearrange the columns manually to make it so
Note: You can use shortcut Ctrl-Shift-Down arrow key to select an entire column / series of columns in large table when the first couple of row are already highlighted.
For output options, you can choose to generate the regression model in either the current worksheet in an empty location or a in a new worksheet. For here, we will generate in the current worksheet.
Select Output Range and select on a suitable empty cell that will become the upper left hand corner of the regression table to be generated.
Finally select Ok.

The required coefficients / parameters are in the 2nd column of the table

	Coefficients
Intercept	-25096.83292
Price ($)	-5055.269866
Adexp ('000$	648.6121403
Promexp ('0(	1802.610956

There is a variety of other important info generated in the table such as R-Squared, confidence intervals, etc that we will explore in later labs.

Interpretation for the current model:

For every 1 unit increase in price ($ 1), the unit sales drops by 5055 units, with all other variables remaining at the same level. Conversely, for every 1 unit decrease in price, the unit sales increase by 5055 units.

For every 1 unit increase in ad expenditure (which in our example is $1,000 as indicated in column header), **the unit sales increase by 648 units**, with all other variables remaining at the same level.
This interpretation scales linearly, which means that if ad expenditure increases by $10,000 (10 units), then we would expect the unit sales to increase by 6,486.12 units with all other variables remaining at the same level.

For every 1 unit increase in promotion expenditure (which in our example is $1,000 as indicated in column header), the **unit sales increase by 1802** units, with all other variables remaining at the same level.

Lets assume that we are trying to maximize the unit sales.

In the simplest case, we locate the coefficient that results in the greatest unit of unit sales.

Based on the 3 coefficients we have obtained so far, we can conclude that decreasing the price by $ 1 results in the greatest increase in unit sales (5055), compared to increasing ad expenditure by $1,000 (increases sales by 648) or increasing promotion expenditure by $1,000 (increases sales by 1802)

Note that in a real world situation, there may be practical considerations that come into play which prevent the application of this interpretation. For e.g. prices of products after often times fixed due to pricing strategies and operational concerns. So in that situation, you may have to resort to increasing promotion expenditure as the 2nd most effective way to increase overall sales.

When interpreting the beta zero coefficient, the estimate of the beta zero coefficient is the value of the y variable when all x variable values are zero. So, in this case, it implies that the value of unit sales would be -25096.83, when all the price is zero, the ad expenditure is zero, and the promotional expenditure is also zero. This technical interpretation of beta zero which does not have any managerial / operational relevance. This is because we are considering a situation where you're selling a toy for free (price is 0), which would never make sense in the real world. Also negative unit sales also has no meaning when translated into real world situations.

In this situation, we say that the beta zero coefficient is a technical interpretation that is required merely to fit the model to data but has **no real life managerial / operational significance**. There could be some situations where the intercept also could have a managerial interpretation. And we'll look at those situations later on in some lab sessions.

4 Generating predictions from the model

File to use: Toy-Sales-pt2

The `initial` worksheet shows the initial table, whereas `final-prediction` shows the final table with generated predictions

We generate predictions from the model by constructing the regression linear equation from scratch and then substituting values for the various independent variables for the different scenarios that we wish to obtain a predicted value for.

This then helps us to determine which scenario maximizes (or minimizes) the dependent variable that we are interested in.

5 Exercise: Interpreting model coefficients and generating predictions

File to use: `Home_Prices-exercise.xlsx`

This sample study involves dataset pertaining to house prices in various regions in a country. Each row in this dataset contains data on the following properties:

a) average house price in a specific region (Price)
b) average number of rooms in all the houses in that region (Rooms)

c) average income of the house owners in that region (Income ($))
d) average property tax of all the houses (Tax_Rate)
e) percentage of property in that area that is used for commercial purposes (%_Commercial)

The first property (Price) is the dependent variable of interest that we want to generate predictions for, while the remaining 4 properties are the selected independent variables that are deemed to most likely contribute to the dependent variable.

Generate a regression model from this dataset, interpret the coefficients correctly, and generate predictions for the average house price in 3 possible scenarios.

6 R-Squared for quantifying model prediction accuracy

File to use: `R-Squared Scenarios`

The `initial` worksheet demonstrates simulating a series of values for a dependent (y-axis) variable by adding / subtracting some random numbers from the results of a linear equation involving the independent (x-axis) variable.

Change the values for the upper and lower limit of the random number that is added to the Model Y values. The larger these values, the bigger the deviation of the simulated actual dependent values from the trend line that represents the regression model. This essentially represents a regression model that is less likely to generate accurate predictions.
Use these values for the upper / lower limit : 0, +- 20, +- 40, +- 60, +- 100

For each value that you use, generate a new regression table for that value (using the Y dependent and X independent value).

Notice that for the case of 0 for upper / lower limit, the simulated actual dependent values align perfectly with the trend line that represents the regression model. This represents a regression model that will always make perfectly accurate predictions, which in turn corresponds to a R-squared value of 1. Also notice the coefficients for the intercept and X correspond to the actual original linear equation (**y = 6x + 10**)

For higher values of the upper / lower limit, the R-squared value of the regression model decreases correspondingly, reflecting a lower accuracy level of predictions generated from that particular regression model.

7 Effects of increasing / decreasing independent variables on R-Squared

File to use: `Toy-Sales-pt4`

One of the key reasons for a low R-squared (which is due to high errors / residuals) is that important independent variables that contribute significantly to the dependent variable have been left out of the regression model / equation.

The `initial` worksheet shows the initial table. Create a copy from this.

Generate 3 regression models from this table where the dependent variable is unit sales and the independent variable is

1. Price only
2. Price and AdExp
3. Price, AdExp and PromExp

The final result should be similar to that in `final-incremental-v1`. Here we can see that incrementally adding independent variables to the regression model / equation increases the R-squared. Also, we can see incremental increase in R-Square due to the addition of each new variable is different for each variable. The bigger the incremental increase, the more significant that newly added variable is in the overall regression equation.

This provides another alternative way to determine the significance of each particular variable in influencing the dependent variable. We had seen earlier the simplest way to do this was to check the magnitude of the coefficient associated with that dependent variable (Section 3).

We can now repeat the previous process of generating 3 different regression models from incremental addition of the independent variables, but with a reversed sequence of adding these incremental variables:

1. PromExp
2. PromExp and AdExp
3. PromExp, AdExp and Price

The final result should be similar to that in `final-incremental-v2`.

By comparing the increment in R-Squared for these two processes, we can deduce the individual contribution of each independent variable to the overall R-Squared:

a) Price (0.61 - 0.65)
b) AdExp (0.04 - 0.06)
c) PromExp (0.10 - 0.12)

We can see that Price here is the most significant independent variable by a far margin compared to the two other independent variables.

This observation is also double confirmed by the value of the coefficient for Price (the highest among the 3 coefficients), which we determined earlier in session 3.

8 R-Squared and Adjusted-R Squared

File to use: `Toy-Sales-pt4`

The easiest way to increase R-Squared is to add extra relevant independent variables to the regression model. However, there is a subtle problem here: R-Squared will either increase or remain the same when new independent variables are added to an existing model **REGARDLESS** of the relevance of these variables to the dependent variable.

This implies that simply adding independent variable values that have no or very minimal contribution to the overall regression model will still increase R-Squared. The end result of this is if that you add

enough meaningless or non-significant independent variables, you will manage to artificially increase your R-Squared to a high enough value that gives a false / incorrect impression of an accurate regression model. This phenomenon is known as overfitting in machine learning.

We can continue from the spreadsheet in the previous lab. Make a copy of `final-incremental-v1`and then add 3 new columns with random values (generated via the RANDBETWEEN function). You can also copy the values for these 3 new columns directly from the result worksheet `final-incremental-dummy`.
These 3 new columns are essentially a simulation of independent variables whose values have no or very minimal contribution to the overall regression model. For e.g. you may assume that Dummy Random 1, 2, 3 could be average rainfall, number of traffic violations, average temperature, etc which you would not expect to directly or significantly impact the unit sales of a particular product (the dependent variable).

Create another regression model involving the 3 new dummy variable columns.
You will notice that:

- R-Square will still increase from the previous model, incorrectly implying that these 3 new columns actually contain data that meaningfully contribute to the dependent variable (when in fact they don't)
- However, the incremental increase in R-Squared is very small (compared to the incremental increase of the 3 original variables Price, AdExp, PromExp)
- The coefficients associated with these 3 new dummy variables is also extremely small compared to that of the 3 original variables (Price, AdExp, PromExp)
- Adjusted R-Square on the other hand will decrease, indicating correctly that the addition of the 3 new columns actually reduce the overall prediction accuracy of the regression model. You can contrast this with the increase of Adjusted R-Square when the 3 original variables (Price, AdExp, PromExp) had been added incrementally.

Point to note: If using the RANDBETWEEN function to generate the random numbers, you will need to select the entire range of cells first and then specify the RANDBETWEEN function and press Ctrl + Enter

9 P-Values to interpret significance of independent variable

File to use: `Toy-Sales-pt4`

Another way to determine how significant a newly introduced independent variable is to the overall regression model is to check whether its P-value is smaller than the α for the regression model.