**Week 2 Excel**

### **Microsoft Excel**

Go to the course website and download the files:

1. Week2-Excel-Instructions.pdf
2. Week2-Excel-Data.xlsx

**2.2 Excel: Data Analysis Add-In**

The statistics options are available as an add-in to Excel. The steps to add it are:

1. In Excel, click on the File tab, then Options
2. Click on Add-Ins
3. Click Analysis ToolPak Add-in, then Go
4. Check the box for Analysis ToolPak, then OK

**2.3 Excel: Descriptive Statistics**

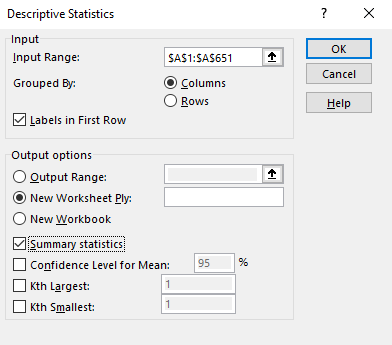
Use the **DescriptiveStatistics** spreadsheet tab for this exercise.

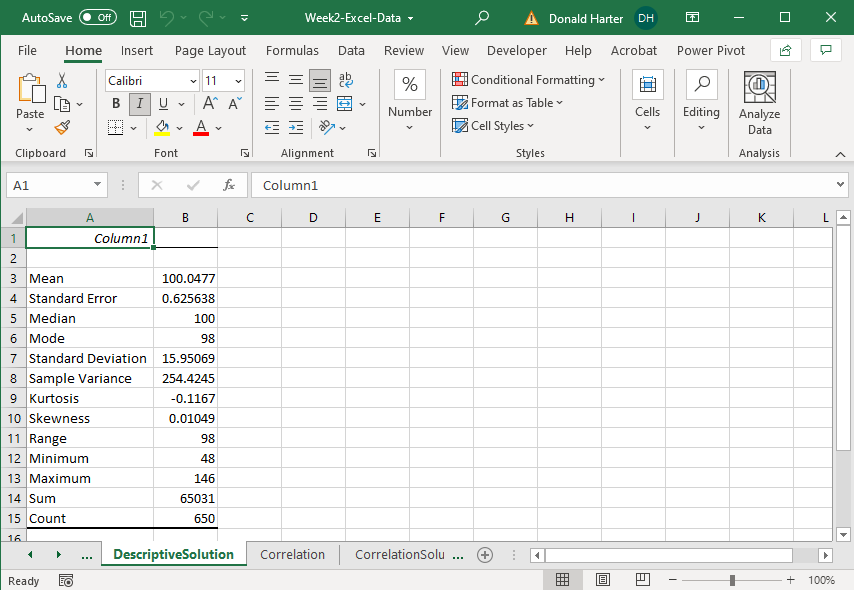
There are several descriptive statistics that can be automatically generated, including:

* Mean: arithmetic average
* Median: middle point in distribution
* Mode: most common value (highest frequency of occurrence)
* Kurtosis: is the data peaked higher or lower than normal?
* Skewness; is the peak shifted left or right?
* Standard deviation: measure of spread
* Range: highest value minus lowest value

To calculate the descriptive statistics:

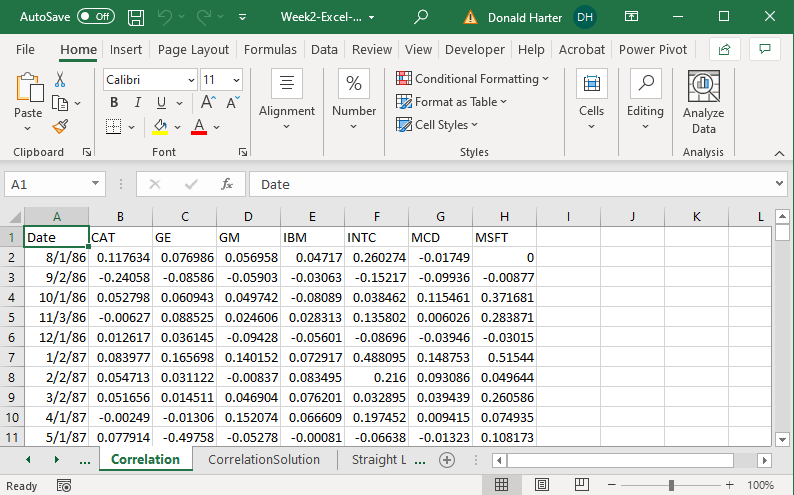
1. Click on the data tab, then data analysis, descriptive statistics, and OK
2. Enter the input range for the IQ data; if you include the header, click on Labels in first row
3. Check Summary Statistics, then OK





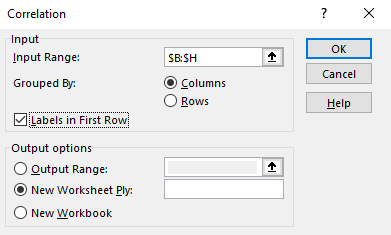
**2.4 Excel: Correlations**

Correlation analysis identifies how two or more variables are related. For this exercise, use the **Correlation** spreadsheet. This spreadsheet records the upward or downward movement of stock by month.

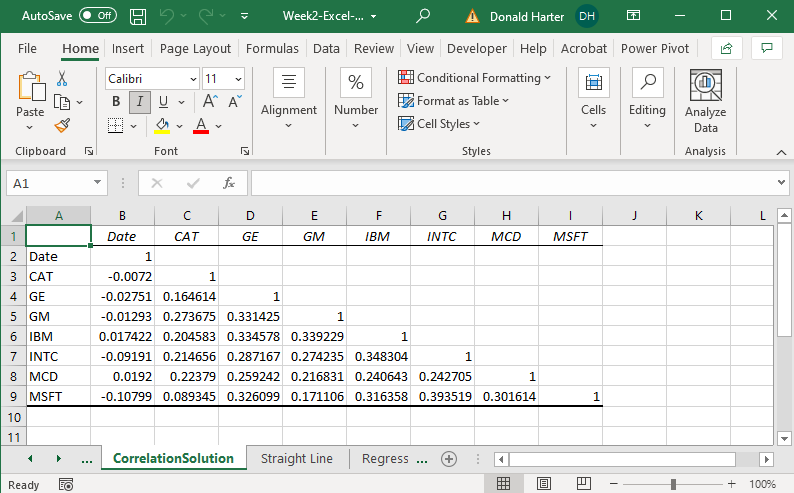


The stocks listed are Caterpillar, General Electric, General Motors, IBM, Intel, McDonalds, and Microsoft.

1. Click on the data tab, data analysis, correlation, then OK
2. Highlight columns B through H (don’t include the date column), group by columns, check Labels in First Row, then OK



The result is shown below. A positive correlation means that when one variable increases, the other increases. A negative correlation means that when one increases, the other decreases.



**2.5 Excel: Regression**

**2.5.1 Regression Overview**

Regression is a technique that attempts to measure the relationship between and outcome variable (dependent) and explanatory variables (independent). To use linear regression, there are three key assumptions:

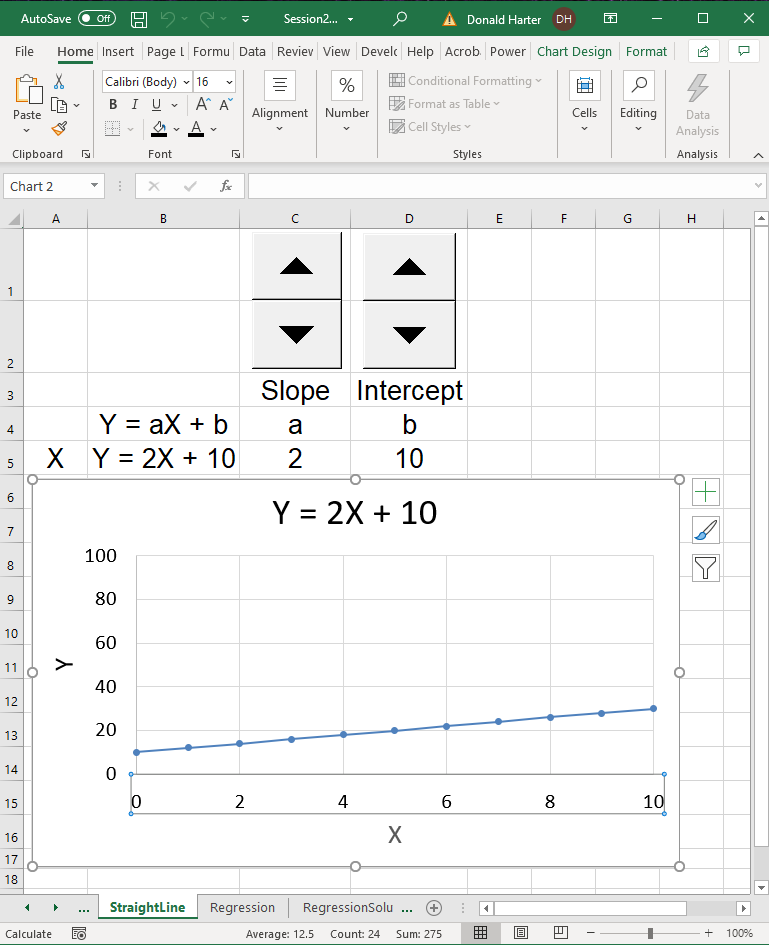
1. Relationship between x and y is linear
2. The x’s are fixed numbers, not random variables (non-stochastic), not related to each other, i.e., independent: Corr(Xi,Xj)=0
3. The error terms:
   1. have zero mean and constant variance: E(εi) = 0, V(εi) = σ2
   2. the error terms are independent: Cov(εi,εj) = 0
   3. the error terms are normally distributed ~N(0,σ2)

Violation of these assumptions requires the use of more sophisticated techniques.

Before performing a linear regression, let us first review the equation of a straight line. A line is represented by the equation:

Y = aX + b

Where a is the slope of X and b is the intercept. We will use the example in the **Straight Line** spreadsheet tab to demonstrate what happens when the slope (coefficient of X) or the intercept change.

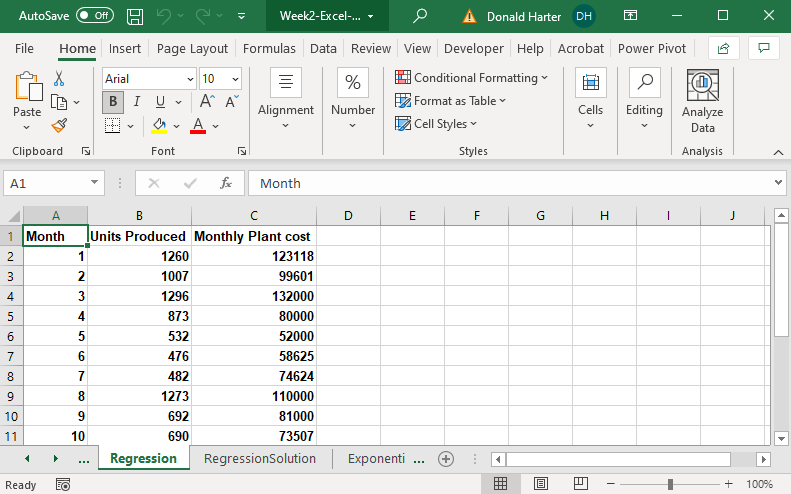


**2.5.3 Excel: Univariate Linear Regression Demo**

Next, we will perform a linear regression on factory costs (Y) explained by the number of units produced in a factory.

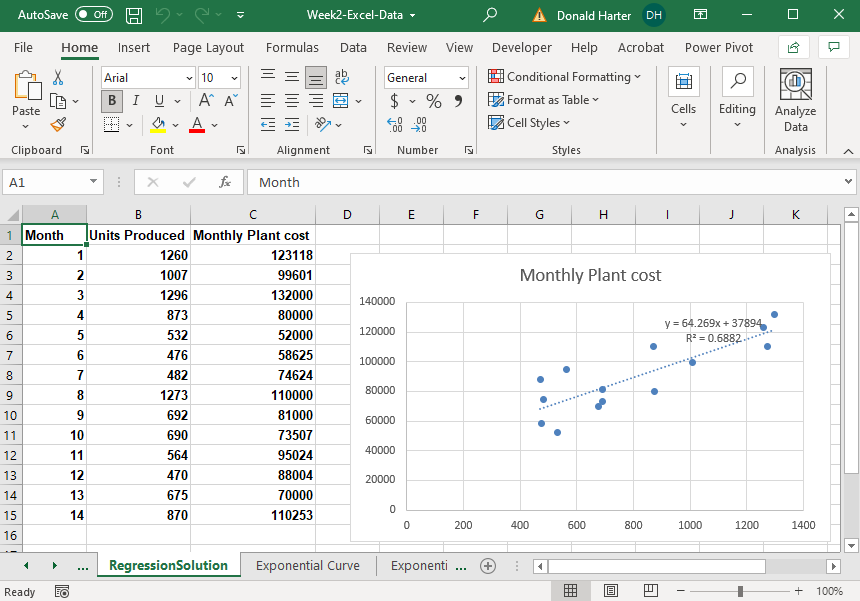
When you want to determine if there is a straight-line relationship in statistics, you can run a regression. Excel has the ability to perform regression analysis. For example, if you wanted to model the relationship between items produced and factory costs, you could estimate the linear relationship. Units produced would be called the independent variable; production costs would be the dependent variable. The output, costs, depends on the input, number of units produced.

For this example, use the **Regression** spreadsheet.



Let’s first draw a scatterplot to see what the data looks like.

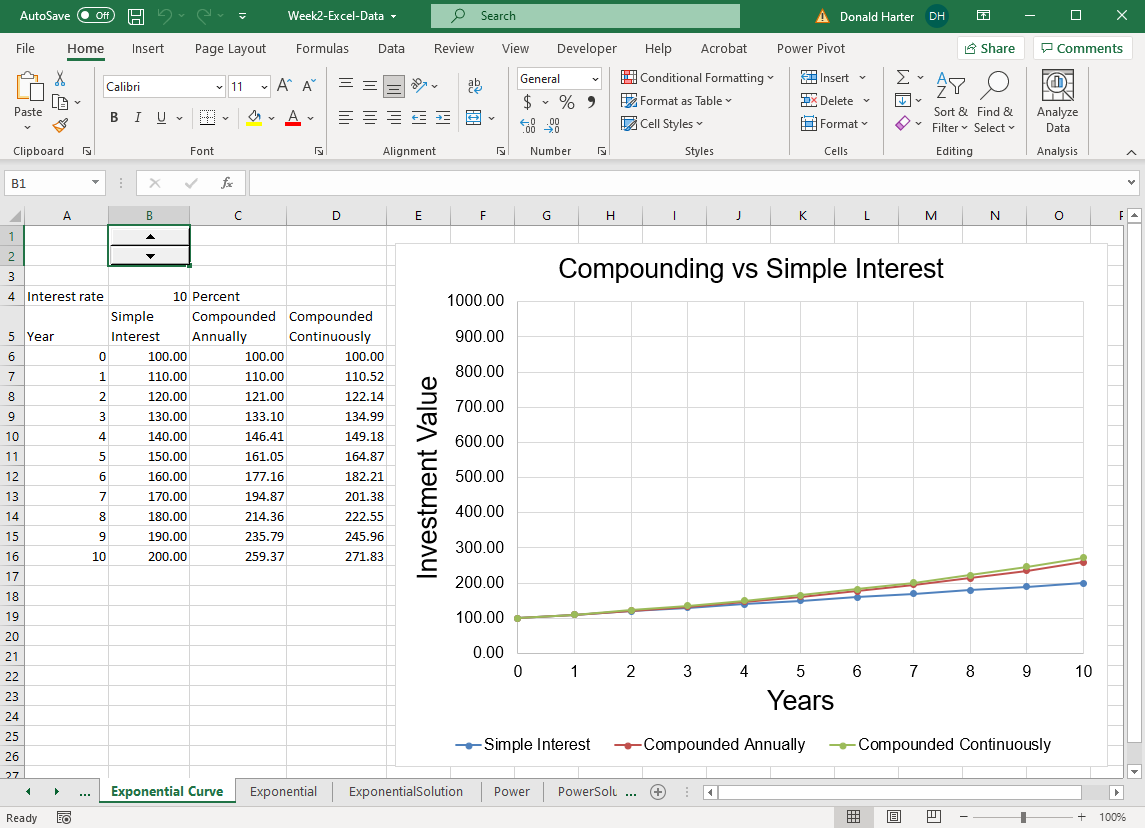
1. Click on the Insert tab.
2. Highlight the cells b1:c15.
3. Click on Scatter in the charts group.



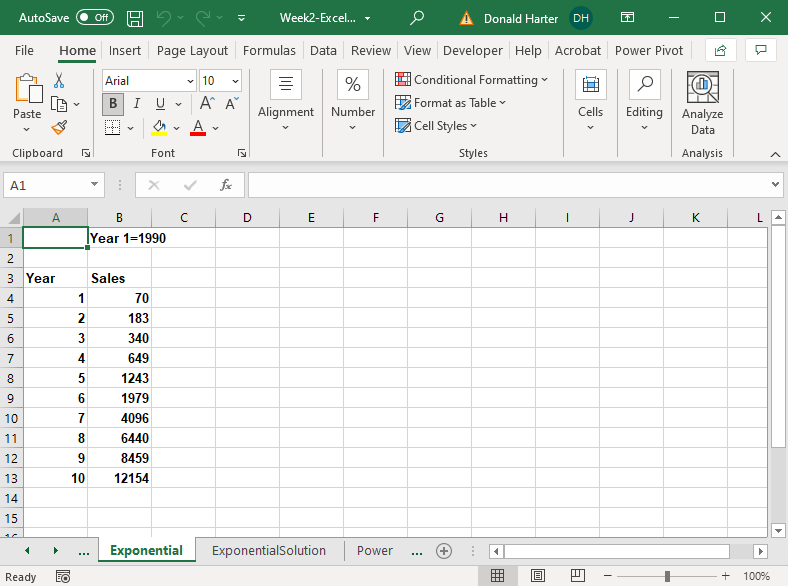
1. Reviewing the chart, it appears that there is a linear relationship. We will therefore perform a linear regression. Click on any data point, right click, then add trendline.
2. In the Format Trendline, Trendline Options, select Linear, then check the boxes for Display Equation and Display R-squared value.
3. In the picture above, the coefficient on x is approximately 64. This means that as unit production increases by one, costs increase by $64. What does the number 37,894 represent? What does the R2 = 0.6882 mean?

**2.5.4 Excel: Exponential Regression Demo**

Before performing an exponential regression, we will examine what an exponential curve looks like. For this demonstration, use the Exponential Curve spreadsheet tab.

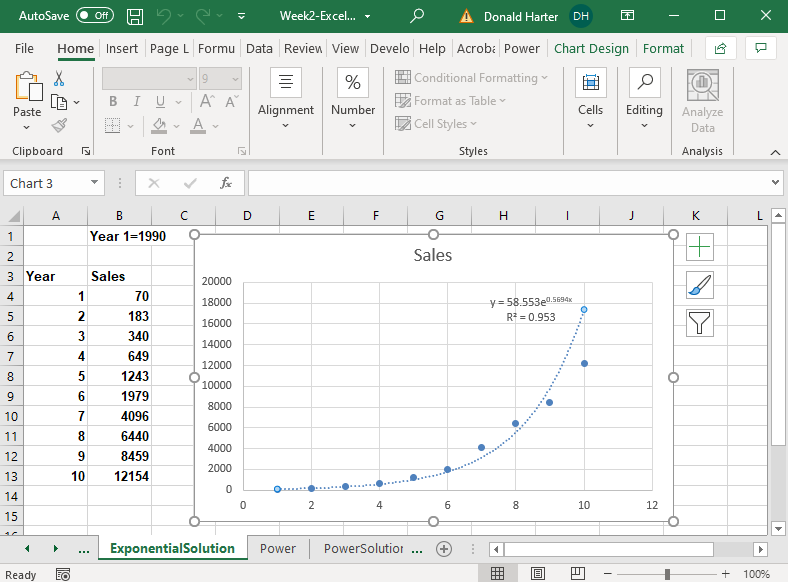


Some data relationships are not linear, but grow at an increasing rate. These curves often follow the exponential growth curve. An exponential growth curve will have the same percentage growth per period compounded over time. Use the Exponential spreadsheet.



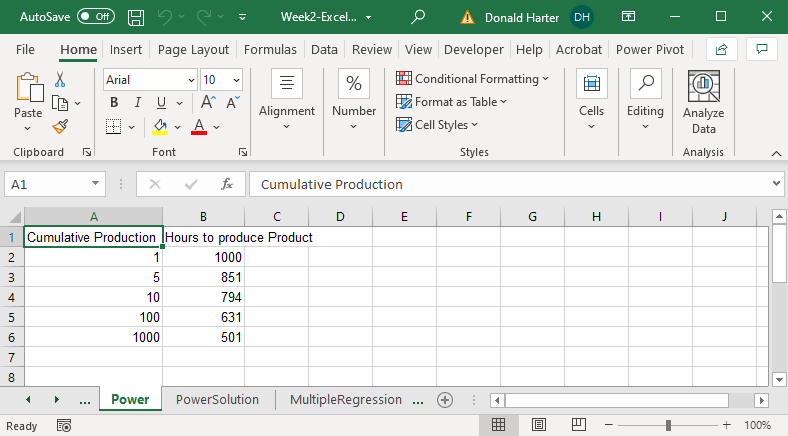
Let’s first draw a scatterplot to see what the data looks like.

1. Click on the Insert tab.
2. Highlight the cells a3:b13.
3. Click on Scatter in the charts group.
4. This data definitely does not look linear. So let’s use the exponential curve. Click on any data point, right click, then add trendline. Select exponential, display equation, and display R-squared, then Close.



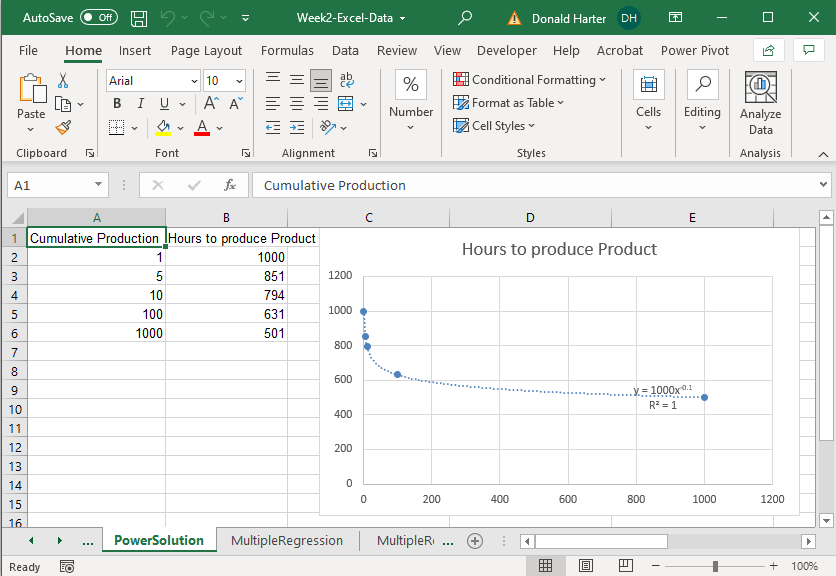
**2.5.5 Excel: Power Regression Demo**

The power curve allows you to examine economies of scale and diseconomies of scale. Economies of scale means that you become more efficient as volume increases. Diseconomies of scale means that you become less efficient as volume increases. Use the **Power** spreadsheet.



Let’s graph as before.

1. Click on the Insert tab
2. Highlight the cells a1:b6
3. Click on Scatter in the charts group
4. Now click on any data point in the graph, right click, add trendline
5. Click on Power, display equation, display R-squared



**2.5.6 Excel: Time Series and Moving Average Demo**

Time series problems have data where one data point is dependent on the previous data point. For example, the closing price of Microsoft stock can be tracked day by day. Today’s price is dependent on yesterday’s price. This dependency from one day to the next, or one time period to the next, is a characteristic of time series data.

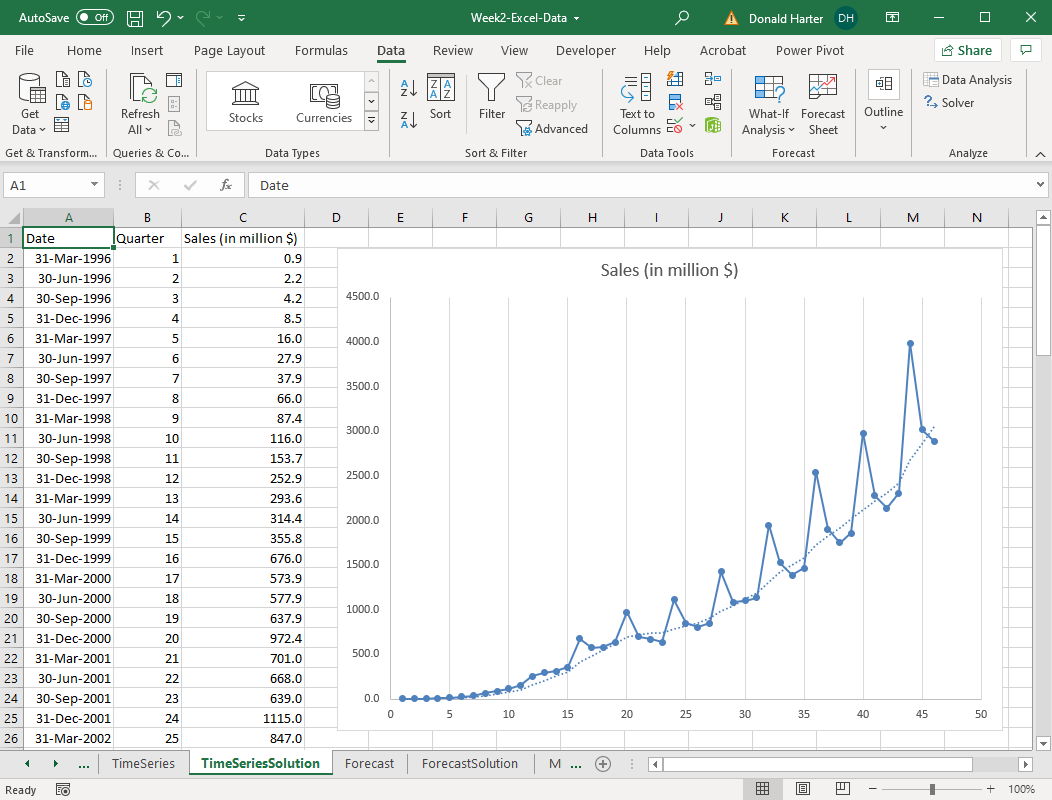
Often there is so much variation in time series data that it’s hard to see trends. Seasonality also masks a trend. Seasonality is variations in data due to high or low points that occur at regular intervals. Create a scatter plot for the Amazon data using the **Time Series** spreadsheet. For this scatterplot, select the option to connect the dots.

1. Click on the Insert tab
2. Highlight b1:c47
3. Click on Scatter in the charts group
4. Select the option to connect the dots

Notice that there is seasonality in the data. Amazon sales tend to peak during the fourth quarter of each year due to holiday sales. However, this seasonality masks the true trend. A moving average helps to see the trend.

To add a moving average line, follow these steps:

1. Right click on a data point
2. Click on Add Trendline
3. Click the checkbox for moving average. Since we have quarterly data, let’s identify the number of periods as four

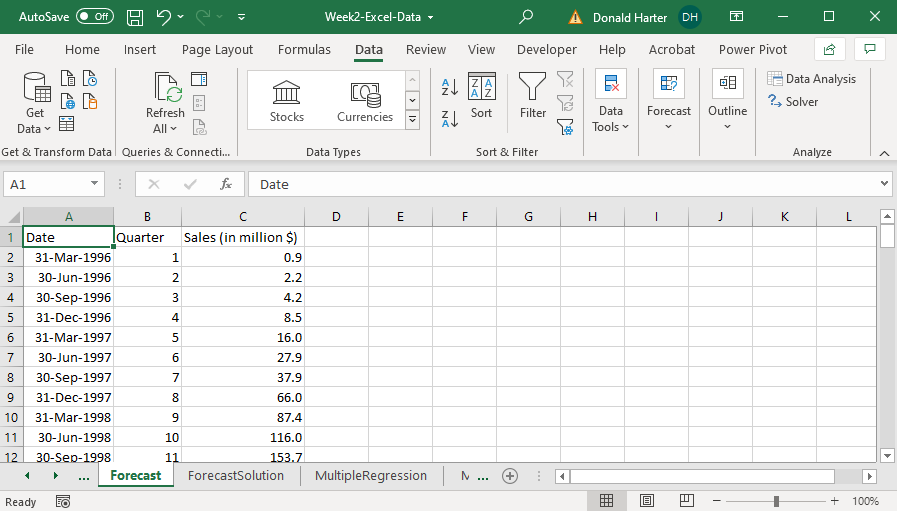


The moving average line is superimposed on the graph. It’s now very clear what the trend looks like when a moving average accounts for seasonality.

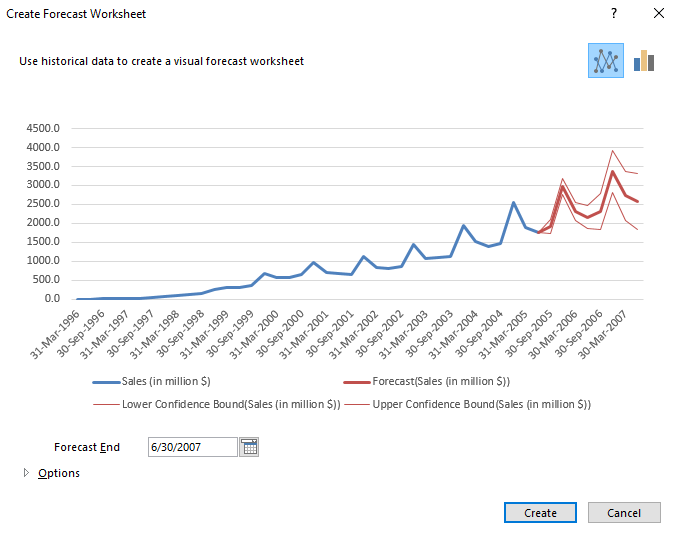
**2.6 Excel: Forecasting**

Regression models describe how changes in explanatory variables affect the outcome over the period of the data. If you want to forecast a trend into the future, then the new forecast feature initially offered in Excel 2016 is available. Forecasts use one date variable and one outcome variable. Note that forecasting outcomes over time can be risky because there is no guarantee that the trend will continue.

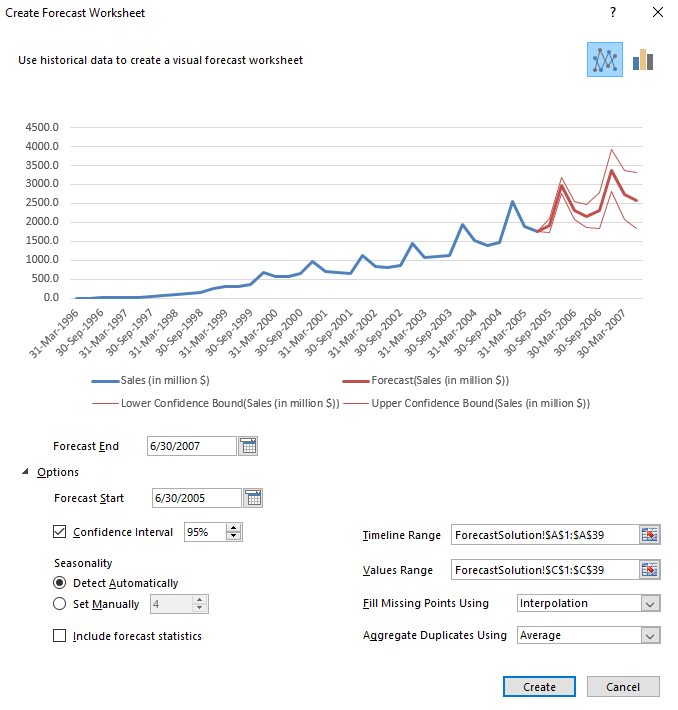
1. Use the **Forecast** spreadsheet for this exercise. The data below is the quarterly sales data for Amazon.com. Recall that the data displayed seasonality.



1. Click on the Data tab, then Forecast Sheet. In this example, Excel automatically identifies the relevant date and trend data.



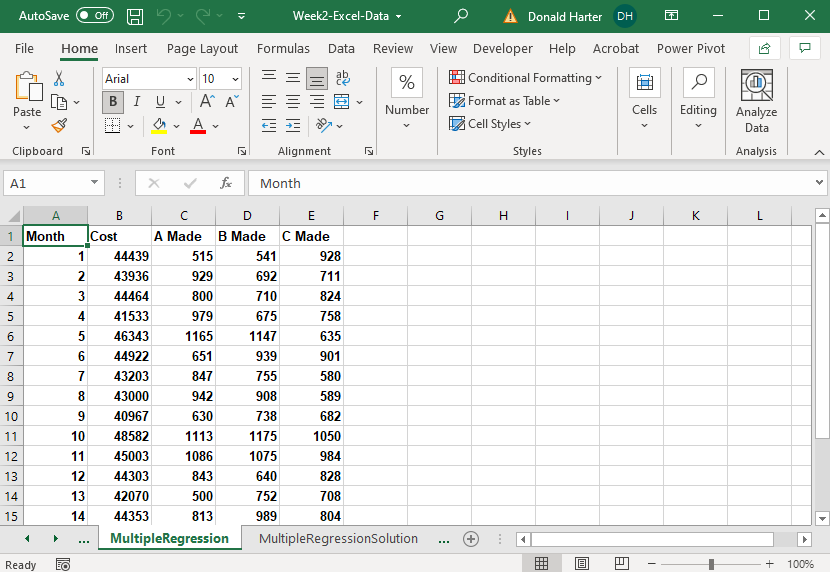
1. Click on the Options drop down arrow in the lower left corner.



1. You can set the Forecast Start and Forecast End.
2. Uncheck the box on Confidence Interval. This removes the confidence interval.
3. Change the Confidence Interval to 50%. What happens?
4. Excel Forecast is generally good at identifying Seasonality. If it has difficulty, you can click on Set Manually and set the seasonality parameter (4 for quarters, 12 for months, 52 for weeks, etc.).
5. If your data has missing data points, you can select Fill Missing Points Using: Interpolation or Zeros. Interpolation is usually better.
6. If there are duplicates in the data, set Aggregate Duplicates Using: Average.
7. In the upper right corner is the option for line versus bar chart. Click each.
8. Click on Create to generate the forecast. New columns with forecasted data are created.

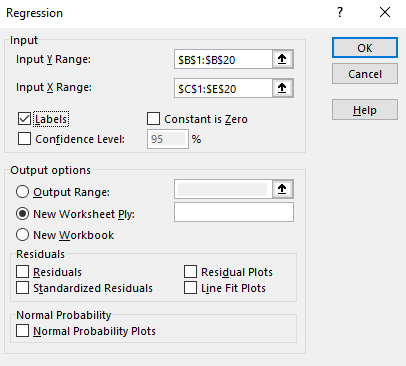
**2.7 Excel: Multivariate Regression**

When we reviewed linear regression earlier, we only had one independent variable. Multiple regression includes several independent variables. Use the **Multiple Regression** spreadsheet.

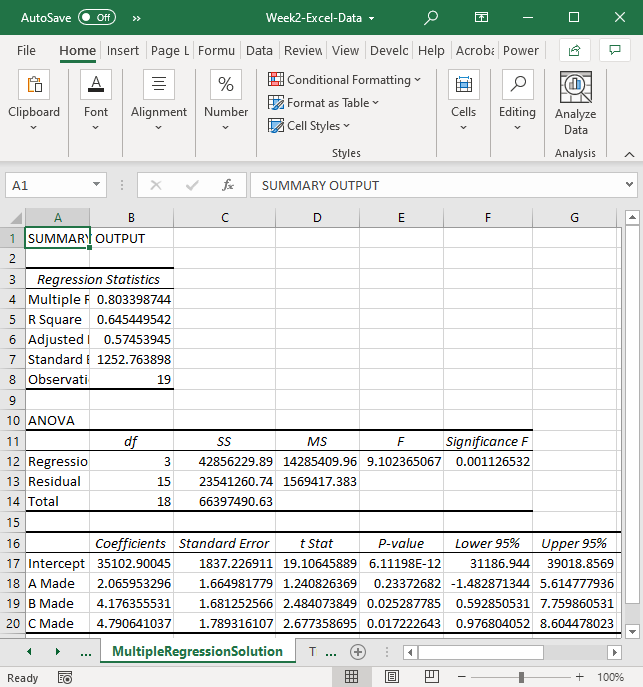


To run a multiple regression:

1. Click on the data tab, data analysis, regression, then OK.
2. For the Y-range, highlight the values in the B column for cost.
3. For the X-range, highlight the values in the C, D, and E columns.
4. If you included the headings at the top of the columns, click labels.
5. Click OK.

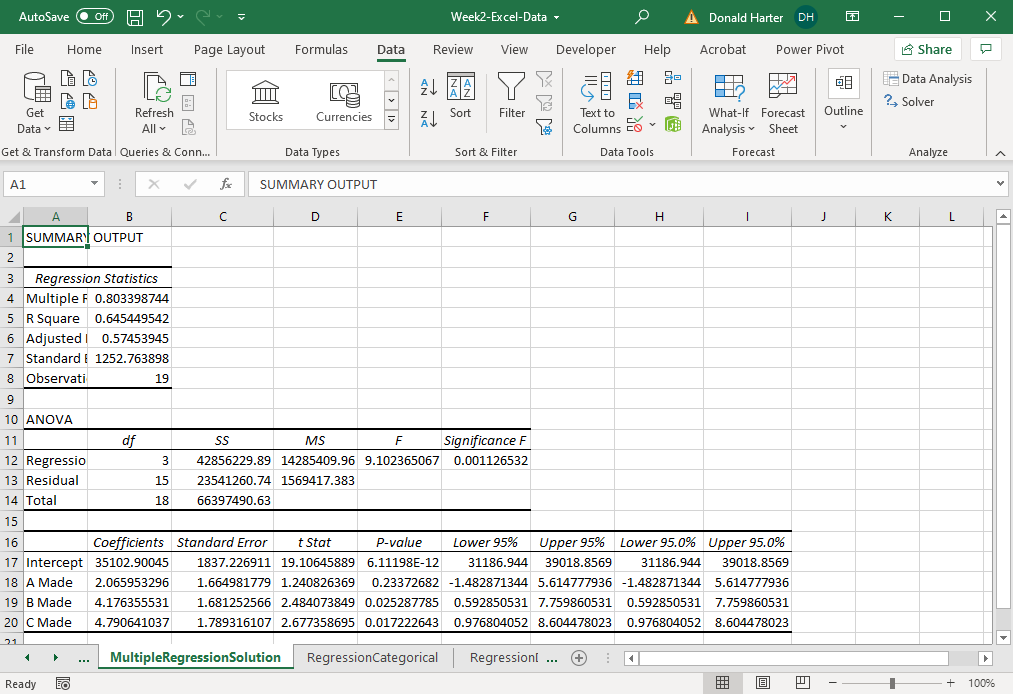


1. The first test is whether the equation is good. If the Significance of F is < 0.05, then the equation is significant. If the equation is not significant, throw it out and do not proceed any further.
2. Second, the R-Square measures the percent change in Y explained by the change in X. In this example, 64.5% of the change in factory cost is explained by the change in the number of A, B, and C produced.
3. Third, the *p*-value of each coefficient determines if it is significant and can be interpreted. If the *p*-value is < 0.05, then it is significant. If the *p*-value is > 0.05, the coefficient is not significant and you cannot say anything about it.



**2.8 Excel: Prediction, Sensitivity Analysis, and Conditional Formatting**

Multiple regression includes several independent variables. Use the **MultipleRegressionSolution** spreadsheet.



1. Is the equation any good? How do you know?
2. What does the R-square mean in this example in business terms?
3. Which coefficients are significant?
4. What do the coefficients mean in business terms?

**Prediction Models**

A prediction model allows you to enter values for each of the inputs (independent variables or X variables) and make a prediction of the outcome (Y variable or dependent variable).

The general form of the equation is:

Y = β0 + β1\*X1 + β2\*X2 + β3\*X3 + ….

For the regression results above, we have:

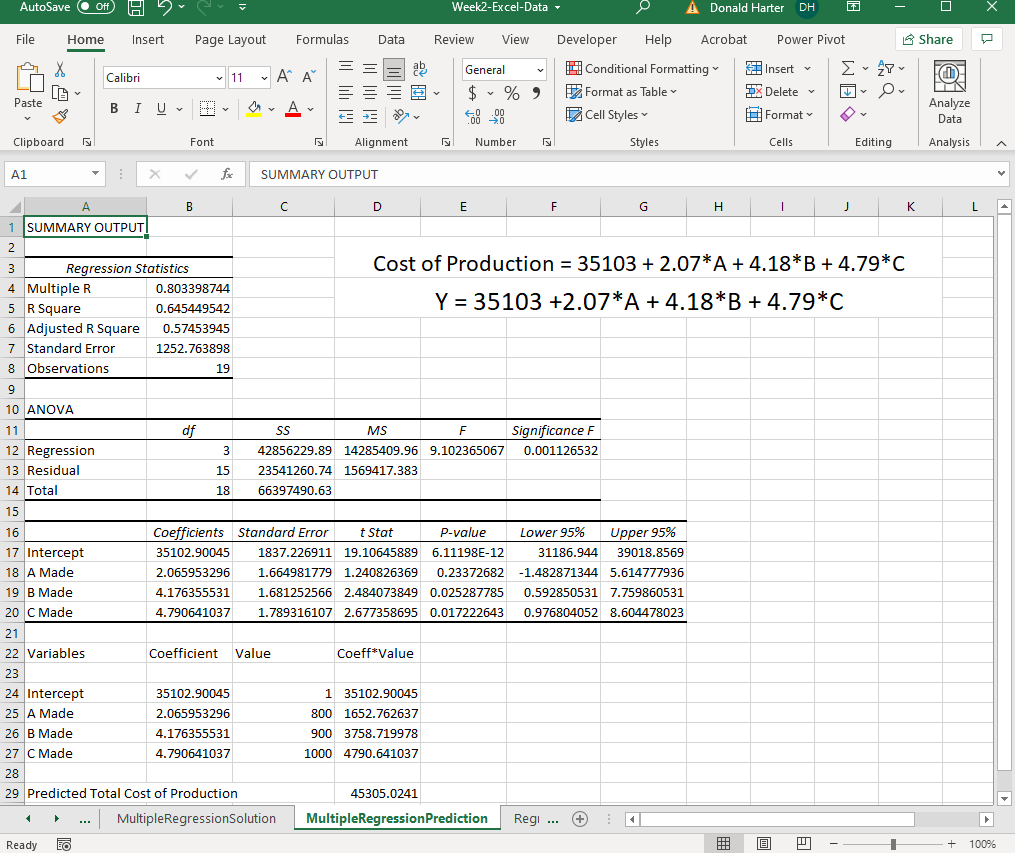
Y = 35103 + 2.07\*X1 + 4.18\*X2 + 4.79\*X3

Or:

Cost of Production = 35103 + 2.07\*A + 4.18\*B + 4.79\*C

Let’s now build the prediction model:

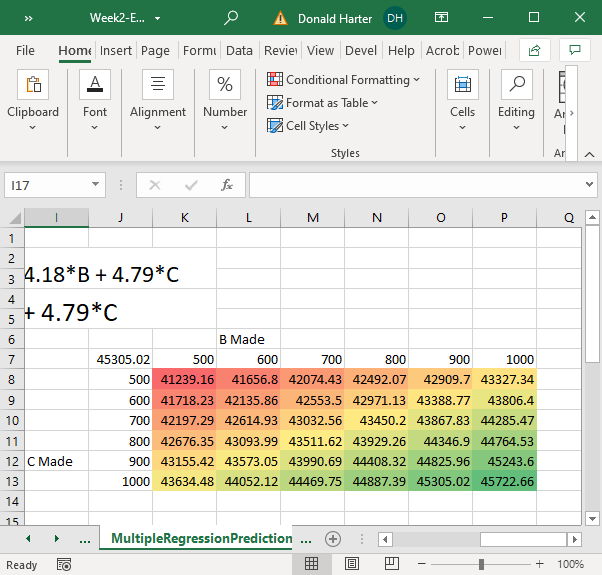
1. In cell A22, enter Variables
2. In cells A24:A27 enter Intercept, A Made, B Made, C Made
3. In cells B22, enter Coefficient
4. In cells B24:B27, copy the coefficients from the regression
5. In cell C22, enter Values
6. In cells C24:C27, enter 1, 800, 900, 1000. Note that the value for the intercept should be one; the values for the X variables must be in the range of the original data
7. In cell D22, enter Coeff\*Value
8. In cell D24, enter the formula =B24\*C24
9. Copy the formula from D24 to D25:D27
10. In cell A29, enter Predicted Total Cost of Production
11. In cell D29, enter the formula =sum(D24:D27)



**Sensitivity Analysis of Regression Results**

After you build the prediction model, you can create a sensitivity analysis of the regression results.

1. In cell L6, enter B Made
2. In cells K7:R7, enter 600, 700, …, 1000
3. In cell I12, enter C Made
4. In cells J8:J15, enter 600, 700, …, 1000
5. In cell J7, enter a formula that points to Predicted Cost, =D29
6. Highlight the data, cells J7:R15
7. Click on the Data tab, What if analysis, Data table
8. For Row input cell, enter C26, the value for B Made
9. For Column input cell, enter C27, the value for C Made
10. Click OK
11. To add conditional formatting, highlight the production costs, cells K8:R15
12. Click on the Home tab
13. Click on Conditional Formatting, Color Scales, your choice of color



**2.9 Excel: Categorical variables**

Often variables in a data set will not be numeric but will be categorical text variables. For this exercise, use the **RegressionCategorical** spreadsheet.

The following university salary data is from:

Weisberg, S. *Applied Linear Regression*, 2nd ed. New York: John Wiley and Sons, (1985), 194.

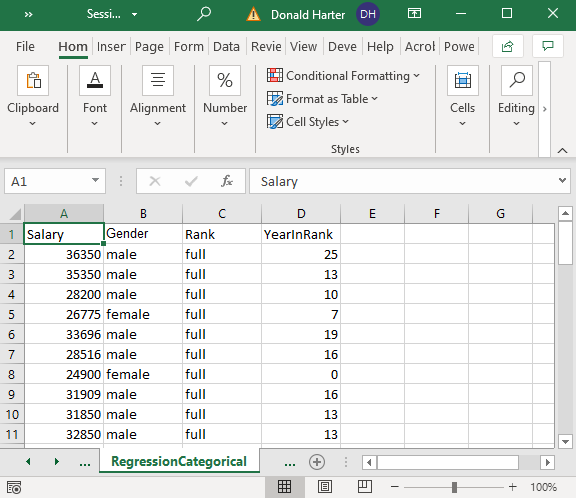
The data includes:

Salary dollars per year

Gender male or female

Rank full, associate, assistant

Years in Rank number of years in the current position



1. The Gender and Rank variables are categorical: they are text fields that put the individual in a group. To run a regression on categorical variables, we need to first convert the text values to dummy variables.
2. Dummy variables are variables that take on the values of zero or one. If a categorical field has two values, such as male or female, then you need one dummy variable, for example, Male = 1 for males, Male = 0 for females.
3. In this example there are three ranks for professor. We will create a dummy for full professor, Full = 1 if the person is a full professor, Full = 0 if not. We also create a dummy for associate professor, Associate = 1 if the person is an associate professor, Associate = 0 if not. In general, when you have n categories for variable, you need n-1 dummy variables for that variable.
4. Create the first dummy variable in column E and call it Male.
   1. Enter Male in E1
   2. An if statement tests a cell and returns a value. The format for an if statement is:

=if(cell=”condition”,1,0)

where “cell” is the cell address and condition is the value you are testing it against ; if the cell=condition, then it returns a 1, else it returns a 0

* 1. In E2, add the following if statement.

=if(B2=”male”,1,0)

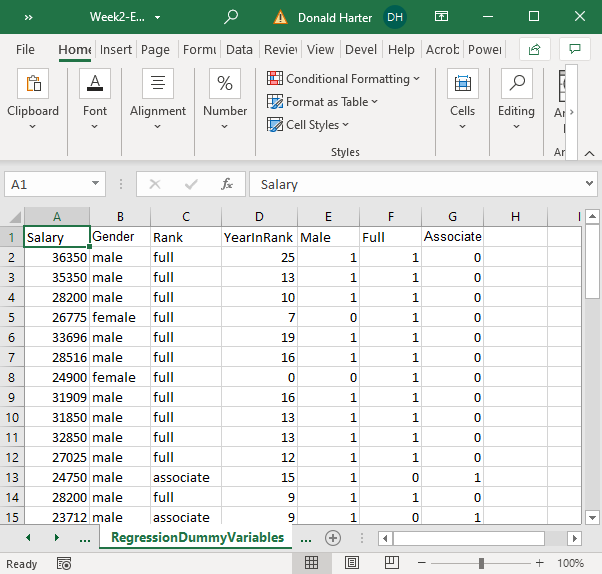
1. Create a dummy for Full Professor.
   1. Enter Full in F1
   2. Enter the following if condition in cell F2

=if(C2=”full”,1,0)

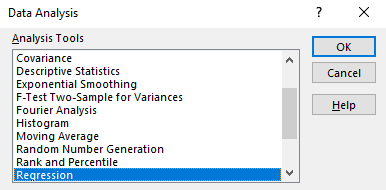
1. Create a dummy for Associate Professor
   1. Enter Associate in G1
   2. Enter the following if condition in cell G2

=if(C2=”associate”,1,0)

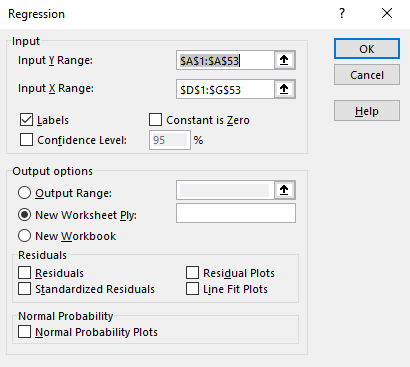
1. Copy cells E2, F2, and G2 down for all data rows



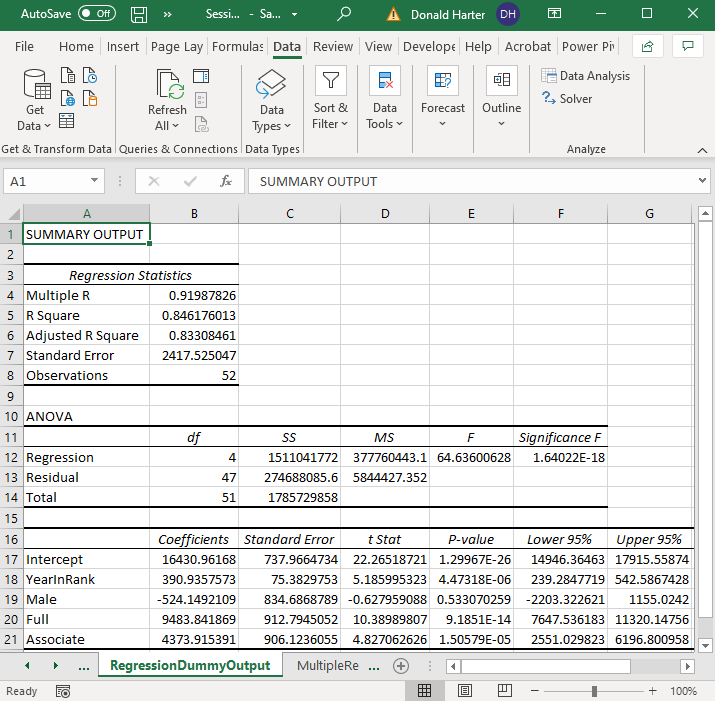
1. Run the regression of the dependent variable Salary on the independent (explanatory) variables YearInRank, Male, Full and Associate.
2. Click on Data at the top of the screen, then Data Analysis, then Regression.



1. In the Regression pop-up menu:
   1. For Input Y range, enter A1:A53
   2. For Input X range, enter D1:G53
   3. Check the box for Labels
   4. Click OK



1. Interpret the equation:
   1. Is the equation significant?
   2. What does the R-square mean?
   3. Which coefficients are significant?
   4. What do the coefficients mean?



1. The categorical variables that are not listed become part of the intercept and are called the base. For example, assistant professor is not listed; its intercept is 16,430.96.
2. The categorical variables that are listed are added to the base intercept. For example, the intercept for full professors is 16430.96 + 9483.84. The intercept for associate professors is 16430.96 + 4373.92.
3. Which level of professor makes more?
4. Which gender makes more?