LAB 5 INSTRUCTIONS

CORRELATION AND LINEAR REGRESSION

In the lab instructions, we will explore how to use Excel to display the relationship between two quantitative variables, to measure the strength and direction of the relationship, and to make predictions about one variable in terms of another variable using a linear regression. In particular, we will learn how to run regression in Excel and interpret the computer output.

For **Activating the Data Analysis Add-In** or **Inserting Excel Output into a Word Document**, see the **Lab 1 instructions**.

1. Scatterplots

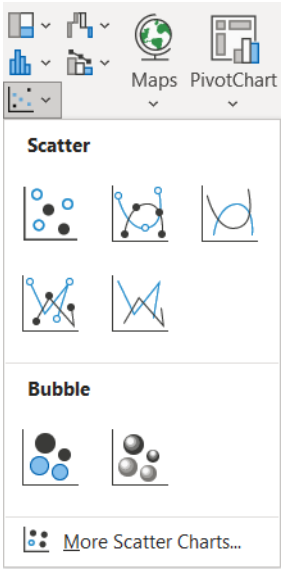
A problem facing every power plant is the estimation of the daily peak power load. Suppose we wanted to model the peak power load (y) as a function of the maximum temperature (x) for the day. The data for 10 days are provided in the table below.

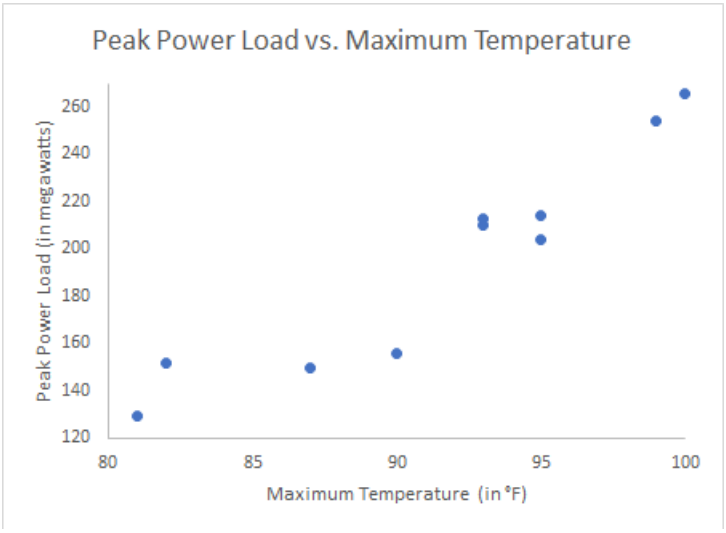
Maximum Temperature (in °F)	95	82	90	81	99	100	93	95	93	87
Peak Power Load (in megawatts)	214	152	156	129	254	266	210	204	213	150

The first step in examining the relationship between peak power load (y) and maximum temperature (x) is to obtain a scatterplot of the response variable (y) versus the explanatory variable (x).

Similar to the creation of boxplots in the **Lab 1 Instructions**, a scatterplot can be created from the **Insert** tab. The following will provide a brief example with preferred formatting, yet more can be determined from this Microsoft link that focuses on the Windows operating system.

After inserting the data above into Excel (you may wish to tranpose the data from rows to columns first), highlight the entire data (including the labels) and then access the **Insert** tab in Excel and clicking the drop-down menu for the scatterplot (diagram on left). This leads to a scatterplot that can be reformatted to become the scatterplot on the right. Note the formatting in terms of the title, the name of the y-axis, the range of the y-axis, the name of the x-axis, and the range of the x-axis.





As you can see the points in the plot approximately follow a linear pattern. We can say that there is a linear relationship between peak power load and maximum temperature. The closer the points follow the linear pattern, the stronger the linear relationship between peak power load and maximum temperature.

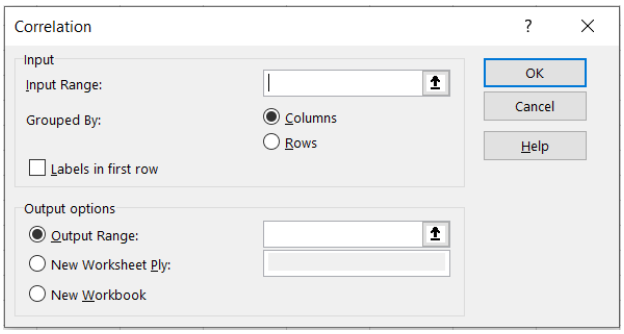
2. Correlation

Correlation measures the strength of the linear relationship between two quantitative variables. It is denoted by *r*. There are several properties regarding correlation that can be reviewed in an instructor's course notes.

2.1 Correlation using Data Analyis feature

Using the **Correlation** feature within the **Data Analysis** feature and clicking on OK will present a dialog box exactly the same or similar to the following one. The **Input Range** should be the range presented by the cells containing the two variables used above. The 'up' arrow to the right of each box can be clicked to permit the mouse to select the data directly from the spreadsheet. If the data was transposed (as recommended above for the scatterplot), the default selection of **Columns** is apt, yet if the data was not transposed, then switching to **Rows** is likely needed. If including the cells that include the variable names, check the **Labels** box.

For the **Output Options**, it is preferred to select *Output Range* and choose a single cell to present the upper-left corner of all the corresponding output. Then click **OK**.



The output below a matrix of pairwise correlations, with the column widths expanded to fit the full variable description. The diagonal values are 1, indicating that each variable has perfect positive correlation with itself. The upper-right section is blank, because its values would be the same as those in the lower-left section.

Note: This method of finding correlation allows for more than two variables. For example, if there were three variables in the dataset rather than two (perhaps minimum temperature could also predict peak power load), the 2x2 matrix below would be expanded to 3x3.

	Maximum Temperature (in °F)	*Peak Power Load (in megawatts)*
Maximum Temperature (in °F)	1	
Peak Power Load (in megawatts)	0.944093	1

2.2 Correlation using Excel functions

Alternatively, the CORREL function can be used. Using the CORREL function enables you to see the changes in the value of the correlation coefficient as you change the data in the worksheet. The Correlation tool in the Data Analysis does not have this feature. The function takes two arguments as described below or via Microsoft.

CORREL(array1,array2)

The CORREL function syntax has the following arguments:

array1 A range of cell values.
array2 A second range of cell values.

The arguments in the CORREL function must satisfy the following conditions: **array1** and **array2** must have the same number of values; **array1** and **array2** must not be empty or have their respective standard deviations equal zero; if any cell contains text, logical values, or empty cells, those values are ignored, though, cells with zero values are included

Example:

Using the variables above (maximum temperature and peak power load), determine the correlation coefficient. Assuming the data are present in columns in the top left corner of the worksheet, the function could be used as follows.

r = CORREL(A1:A11, B2:B11) = 0.944093.

Note: This method of finding correlation is only valid for two variables (two cell ranges).

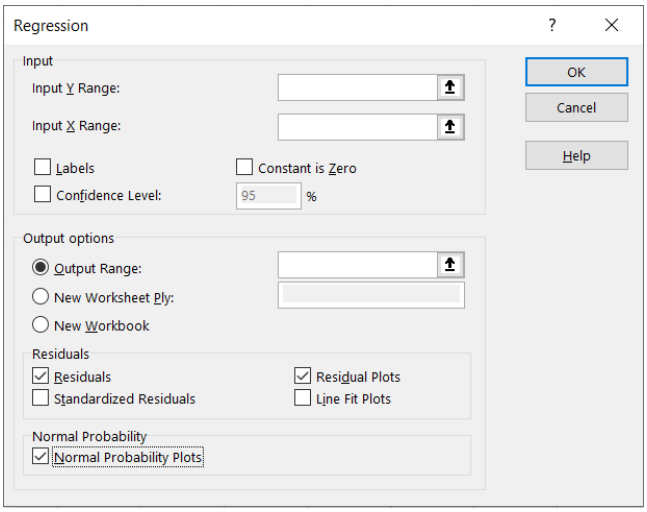
3. Simple Linear Regression

The pattern in the scatterplot above could be described by a straight line, yet there is some unexplained variation in the plot that cannot be explained by a linear relationship between peak power load (y) and maximum temperature (x). Consider the following models that account for this random error, the model relating to the individual value of y displayed on the left while the model relating to the mean value of y displayed on the right.

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ $\qquad\qquad$ $E(Y_i \mid x_i) = \beta_0 + \beta_1 x_i$

Using the **Regression** feature within the **Data Analysis** feature and clicking on OK will present a dialog box exactly the same or similar to the following one. The **Input Y Range** should be the range presented by the cells containing peak power load. The **Input X Range** should be the range presented by the cells containing maximum temperature. The 'up' arrow to the right of each box can be clicked to permit the mouse to select the data directly from the spreadsheet. If including the cells that include the variable names, check the **Labels** box. If the intercept (β_0) should be set to zero, check the **Constant is Zero** box. If a confidence interval for the slope and intercept are needed, check **Confidence Level** and add the desired confidence level in the corresponding box (the default seen below is 95%).

For the **Output Options**, it is preferred to select *Output Range* and choose a single cell to present the upper-left corner of all the corresponding output. To obtain residuals (observed y minus predicted $\hat{y}$) and the predicted y values, check the **Residuals** box. Check the **Residual Plots** box to obtain plot of residuals versus the x variable. Check the **Normal Probability Plots** box for a normality plot of the residuals. Then click **OK**.



The regression output on the next page is divided into four sections. Analysis follows after the output.

The first section (*Regression Statistics*) of the output shows summary statistics of the regression. We will discuss only those components of the output that are discussed in the course.

The *R Square* (coefficient of determination) measures the proportion of variation in the response variable y that is explained by the least-squares regression of y on the predictors. In other words, *R Square* = *SSR*/*SST*. The proportion must be a number between zero and one, and it is often expressed as a percentage.

The standard error s estimates σ, which measures the variation of y about the population regression line. The smallest value that s can assume is zero, which occurs when all the points fall on the least-squares regression line.

The second section in the regression output is the *ANOVA* (Analysis of Variance) table for regression. It analyzes the variation in the data by breaking it into two parts: the first due to the regression (model), and the second due to the residuals (or error). ANOVA gives the degrees of freedom, sum of squares, and mean squares for the regression and residuals. Moreover, it includes the value of the *F*-statistic and *p*-value (*Significance F*) to test the null hypothesis that at least one slope of the population regression equation is not zero; in other words, at least one explanatory variable is a useful predictor of *y*.

The third section of the output includes the statistics concerning the regression coefficients. The intercept (b_0) and the slope (b_1) of the least-squares regression line are in the column labelled *Coefficients*. The *t Stat* column contains the respective *t*-statistic value to test the hypothesis that the respective coefficient is equal to zero. The *p*-values are provided for the two-sided alternatives in the *P-value* column. The *p*-values for one-sided alternatives can be obtained with brief calculations. Also, the lower and upper bounds for confidence intervals for the slope and intercept are also provided.

On the left, the fourth section contains predicted values ($\hat{y}$) using the estimated regression line for each value of *x* in the dataset and residuals that compare the corresponding observed value of *y* in the dataset to the predicted value ($\hat{y}$). On the right are values calculated to create the normal probability plot on the next page.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.944093
R Square	0.891312
Adjusted R Square	0.877726
Standard Error	16.17764
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	17169.87	17169.87	65.60495	3.99E-05
Residual	8	2093.729	261.7161		
Total	9	19263.6			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-419.849	76.05778	-5.52013	0.00056	-595.239	-244.46
MaxTemp	6.717477	0.82935	8.099688	3.99E-05	4.804992	8.629962

RESIDUAL OUTPUT

Observation	*Predicted Peak Power Load*	*Residuals*
1	218.3112	-4.31117
2	130.984	21.01603
3	184.7238	-28.7238
4	124.2665	4.733509
5	245.1811	8.818922
6	251.8986	14.10145
7	204.8762	5.123784
8	218.3112	-14.3112
9	204.8762	8.123784
10	164.5714	-14.5714

PROBABILITY OUTPUT

Percentile	*Peak Power Load*
5	129
15	150
25	152
35	156
45	204
55	210
65	213
75	214
85	254
95	266

The equation of the least-squares can be displayed as follows, where using (abbreviated) variable names as seen on the right is preferred. Note that *PeakPowerLoad* is peak power load and *MaxTemp* is maximum temperature.

$\hat{y}$ = -419.849 + 6.617477*x* OR *PeakPowerLoad* = -419.849 + 6.617477**MaxTemp*

Thus, to interpret the slope, as the maximum temperature increases by 1°F, the mean peak power load increases approximately by 6.617477 megawatts.

The estimate of the model standard deviation (σ) is 16.17764 and 89.1312% of the variation in peak power load is explained by maximum temperature.

Notice that the two-sided *p*-value of 3.99E-05 (or 3.99×10^{-5} in scientific notation) for the *t*-test for the slope (equivalent to the ANOVA *F*-test) indicates that linear regression on maximum temperature is very useful in explaining peak power load. The output also shows that the 95% confidence interval for the slope indicates that the mean peak power load increases between 4.804992 and 8.629962 megawatts as temperature increase by 1°F.

After adjusting the axes of each plot, the reformatted residual plot (left) and normal probability plot (right) are seen below. In the residual plot, the variances are not quite constant as *x* increases, so it is debatable if the constant variance assumption holds. (It might also be possible to view the parabolic pattern.) In the normal probability plot, the points are reasonably close to a straight line, indicating normality, yet there are some deviations, so it is also debatable if the normality assumption holds.