# LAB 5 INSTRUCTIONS

# LINEAR REGRESSION AND CORRELATION

In this lab you will learn how to use Excel to display the relationship between two quantitative variables, 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, you will learn how to run regression in Excel and interpret the computer output.

## 1. Motivation Example

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 (X) $^{o}$F | Peak Power Load (Y) in megawatts |
|---|---|
| 95 | 214 |
| 82 | 152 |
| 90 | 156 |
| 81 | 129 |
| 99 | 254 |
| 100 | 266 |
| 93 | 210 |
| 95 | 204 |
| 93 | 213 |
| 87 | 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).

## 2. Scatterplots

Scatterplots have already been discussed in *Lab 1 Instructions*. For your convenience we will cover here the basic commands necessary to obtain and edit scatterplots in Excel. In Excel a scatterplot is called an *XY(Scatter)* chart and it can be created and edited using *ChartWizard.*

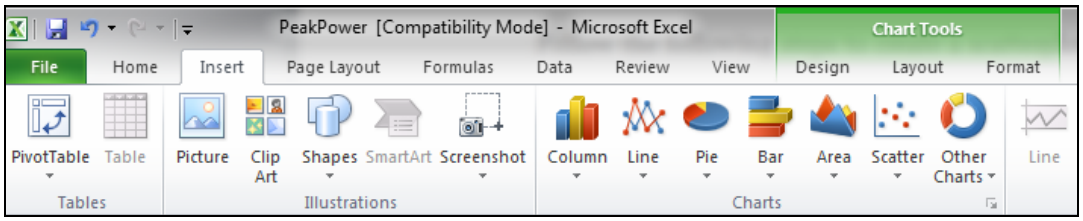
### 2.1 Creating a scatterplot

Follow the following steps to create a scatterplot:

1. Arrange the data in adjacent columns on a worksheet with the x variable (for the horizontal axis) on the left and the y variable (for the vertical axis) on the right.

2. Highlight the columns that contain the data you want to represent in the scatter plot. In this example, those columns are **Maximum Temperature** and **Peak Power Load**.

3. Open the **Insert** tab on the Excel ribbon. Click on **Scatter** in the **Charts** section to expand the chart options box. Select the first item, **Scatter with only Markers**, from this box.




4. After making this selection, the initial scatter plot will be created in the same worksheet. You can resize this chart window and drag it to any other part of the worksheet.

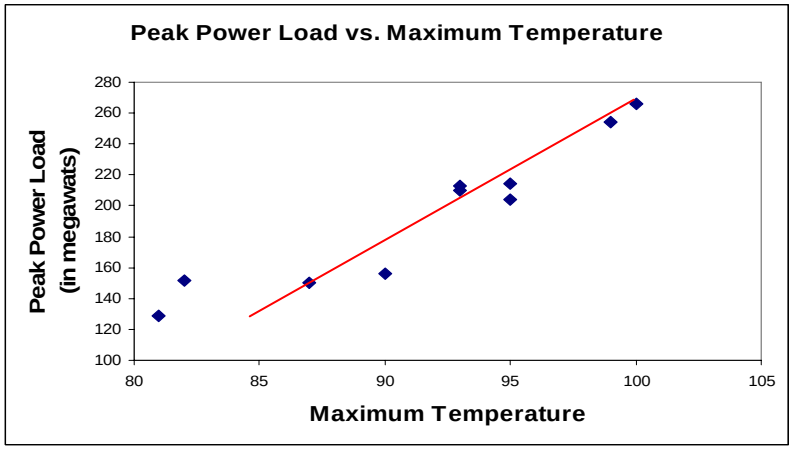
5. Make any formatting or design changes you wish in the **Design**, **Layout**, and **Format** tabs located under **Chart Tools** on the Excel ribbon.

   For example, to label the horizontal axis, select the **Layout** tab under **Chart Tools**, click on **Axis Titles** in the **Labels** section, then choose **Primary Horizontal Axis** and finally select **Title Below Axis**. A text box with the default wording **Axis Title** will appear on the chart. Click anywhere in that text box and edit the information.

   In order to change the chart title, click on the title to open the text box that contains it and edit it with your new description.

   To **move** an embedded scatterplotplot, select it by clicking anywhere in the chart and then drag it where you want it. To **resize** an embedded plot, select it, and then drag one of the eight handles in the desired direction indicated by the pointer. To **delete** it, select the plot and then press *Del.*

   The scatterplot for our example is displayed below. Notice that the axes have been rescaled to display only the observed values:




   As you can see the points in the plot follow approximately 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.

## 3. Correlation

Correlation measures the strength of the linear relationship between two quantitative variables. It is denoted by r.

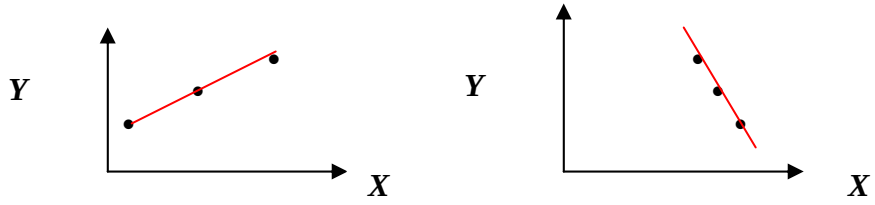
**Properties of correlation:**

1. $-1 \leq r \leq 1$,

2. $r > 0$ indicates a positive linear relationship between X and Y;

   $r < 0$ indicates a negative linear relationship between X and Y.

3. The correlation $r = -1$ or $r = +1$ occur only when the points in a scatterplot lie exactly along a straight line.

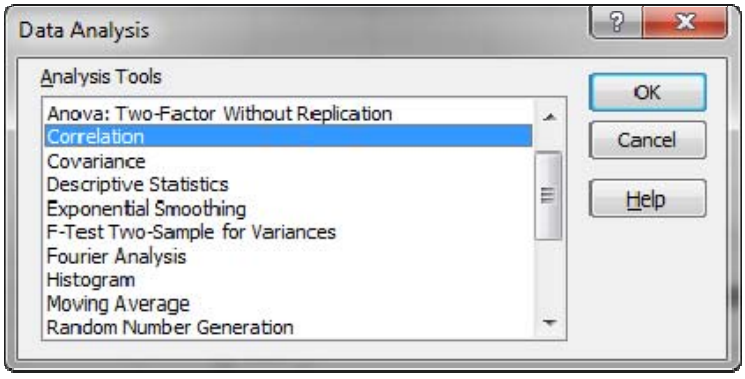



4. The larger the absolute value of |r| of r, the stronger the linear relationship between X and Y. Thus the value of r close to -1 or +1 indicates a strong linear relationship between X and Y, whereas the value of r close to 0 indicates a very weal linear relationship.

## 4. Correlation in Excel

To obtain the correlation coefficient in Excel follow the following steps:

1. Enter the x and y values in a worksheet. Open the *Data* tab on the Excel ribbon and choose *Data Analysis*. In the *Data Analysis* dialog box, select *Correlation* and press OK.




2. In the *Input* box, specify the location of the data, including the labels. Verify that the data are grouped in columns and be sure the *Labels* box is checked. In the *Output* section, click the *Output Range* button, and specify the upper-left cell where the correlation output will be located. Click OK.

The output is a matrix of pairwise correlations. 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.

Alternatively, you could use the *CORREL* function in the *Insert Function* (select *Statistical* as the *Function Category*) or enter the formula =CORREL(*Range of x values*, *Range of y values*). 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 the feature.

## 5. Linear Regression Model

Examine again the scatterplot of peak power load versus maximum temperature for our data. Notice that though the points lie close to the straight line, 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). Let us consider a model that accounts for this random error:

$$Y= \beta_0 +\beta_1 \cdot X+ \varepsilon.$$

In general, Y is a random variable and X is a non-random explanatory variable. We assume that $\varepsilon$ (random error) follows a normal distribution with mean zero and unknown variance $\sigma^2$.

We will call the above model the **simple linear regression model**, it is simple because it has only one independent (explanatory) variable. The line $Y= \beta_0 +\beta_1 \cdot X$ is called a regression line.

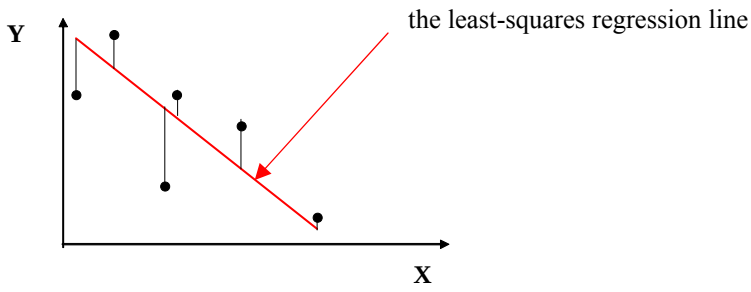
The value of $\sigma^2$ shows the size of the scatter around the regression line. The smaller $\sigma^2$ (relative to the average value of Y), the better the predictions about Y using the above linear regression model. As the mean of $\varepsilon$ is zero, thus for each fixed value x of X, the mean $E(Y|x)$ of Y is

$$E(Y|x) = E(\beta_0 +\beta_1 \cdot x+ \varepsilon) = \beta_0 +\beta_1 \cdot x + E(\varepsilon) = \beta_0 +\beta_1 \cdot x + 0 = \beta_0 +\beta_1 \cdot x.$$

In other words, the points on the regression line show the means of the response Y for each value x of X. Notice that $\beta_1$ shows the mean change in the variable Y as X increases by 1 unit. Moreover, if the slope $\beta_1$=0, the explanatory variable X is useless as the predictor of Y.

In order to estimate the unknown values of the slope $\beta_1$ and the y-intercept $\beta_0$ in the regression equation, a random sample of n pairs of observations $(X_1, Y_1)$, $(X_2, Y_2)$, …, $(X_n, Y_n)$ is taken. For the peak power load example, the sample of n=10 such pairs is provided in the table above.

For the sample data displayed in the scatterplot below, we can find a line that minimizes the sum of squares of the vertical deviations of the points from the line. The line obtained in this manner is called **the least-squares regression line**. Thus the least-squares line is as close as possible to all points in the scatterplot.




The least-squares regression line has the form

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X$$

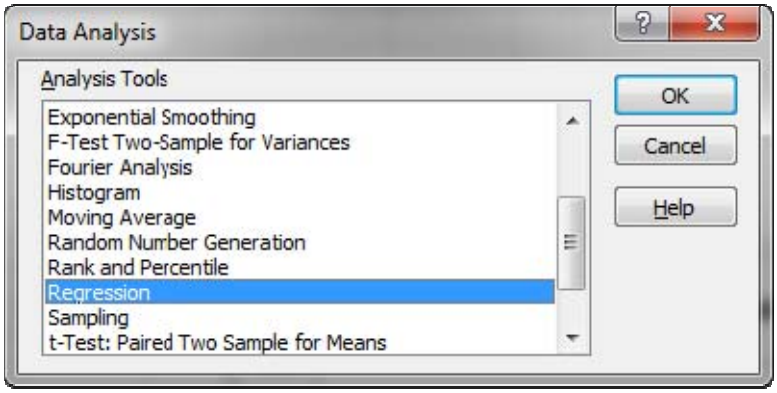
and its slope $\hat{\beta}_1$ and the y-intercept $\hat{\beta}_0$ are estimates of the unknown regression coefficients $\beta_1$ and $\beta_0$, respectively. The values of $\hat{\beta}_0$ and $\hat{\beta}_1$ can be obtained easily with Excel. The equation can be used to make predictions about the mean value of Y for each value x of X.

## 6. Linear Regression Analysis Input in Excel

If the points on the scatterplot fall approximately on a straight line, you can fit a straight line to your data with Excel by following the following instructions:

1. Arrange the data in adjacent columns with the x variables on one side and the y variable on the other side. Make space for the results of the regression analysis to the right of the data (at least sixteen columns).

2. Open the *Data* tab on the Excel ribbon and choose the *Data Analysis* command. Select *Regression*, and click *OK*. The Regression dialog box appears. In the Regression dialog box, move from box to box using the mouse or the tab key. Do not press *Enter* until all input data are entered.




3. Enter the range of the y and x values in the input boxes. Include the labels above the data. Notice that in case of several x variables, you will select a rectangular range of values. Now select the *Labels* box because the labels were included in those ranges. Select the box *Constant is Zero* only if you want to force the regression line to pass through the origin (0,0). Excel automatically includes 95% confidence intervals for the least-squares line coefficients. If you want different level than 95%, enter it in the *Confidence Level* box.

Click the *Output Range* button and type an address for the top left corner of a range sixteen columns wide where the summary output and charts should appear.

To obtain **residuals** (observed Y minus predicted $\hat{Y}$) and the predicted y values, select the *Residuals* box. Select the *Residual Plots* to obtain plot of residuals versus the x variable. After selecting all needed options, click OK. The summary output and charts appear. Notice that the plot of residuals versus predicted values has to be obtained using the values provided in the output, it is not included in the output.

## 7. Regression Analysis Output Interpretation

The regression output displayed below is for only one explanatory variable. It is divided into three sections.

| ***Regression Statistics*** | | | | | | |
|---|---|---|---|---|---|---|
| **Multiple R** | \|r\| | | | | | |
| **R Square** | $R^2$ | | | | | |
| **Adjusted R Square** | | | | | | |
| **Standard Error** | s | | | | | |
| **Observations** | | | | | | |
| | | | | | | |
| **ANOVA** | | | | | | |
| | ***df*** | ***SS*** | ***MS*** | ***F*** | ***Significance F*** | |
| **Regression** | 1 | SSR | SSR/1 | | | |
| **Residual** | n-2 | SSE | SSE /(n-2) | | | |
| **Total** | n-1 | SST | | | | |
| | | | | | | |
| | ***Coefficients*** | ***Standard Error*** | ***t Stat*** | ***P-value*** | ***Lower 95%*** | ***Upper 95%*** |
| **Intercept** | $\hat{\beta}_0$ | $s.e.(\hat{\beta}_0)$ | $\hat{\beta}_0 / s.e.(\hat{\beta}_0)$ | | | |
| **X Variable 1** | $\hat{\beta}_1$ | $s.e.(\hat{\beta}_1)$ | $\hat{\beta}_1 / s.e.(\hat{\beta}_1)$ | | | |

Fraction of variation in y values explained by regression

Estimate of model standard deviation σ

Regression Model Sums of Squares

P-value of F-test

Error Sum of Squares

P-values of the t-tests (two-sided) about the slope and the y-intercept of the population regression line

Lower and Upper Bounds of Confidence Interval for the coefficients $\beta_0$, $\beta_1$

The *Regression Statistics* section 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 is the ratio 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 (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 to test the null hypothesis that at least one slope of the population regression equation is not zero, i.e. 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 lower-left part labelled *Coefficients*. The *t Stat* column contains the values of the t statistic used to test the hypothesis that the coefficients are equal to zero. The P-values are provided for the two-sided alternatives. The P-value for a one-sided alternative can be obtained by dividing the corresponding P-value for the two-sided test by two. Moreover, the lower and upper bounds for confidence intervals for the slope and y-intercept are also provided.

## 8. Peak Power Load Example

The simple linear regression output for the example is given below:

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.936231687 |
| R Square | 0.876529771 |
| Adjusted R Square | 0.861095992 |
| Standard Error | 17.10175737 |
| Observations | 10 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1 | 16610.23916 | 16610.24 | 56.79295 | 6.69531E-05 |
| Residual | 8 | 2339.760841 | 292.4701 | | |
| Total | 9 | 18950 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
|---|---|---|---|---|---|---|
| Intercept | -412.5492773 | 80.40242616 | -5.13106 | 0.000895 | -597.9577244 | -227.14083 |
| Temperature | 6.607095926 | 0.876725039 | 7.53611 | 6.7E-05 | 4.585363054 | 8.6288288 |

The equation of the least-squares is Y = -412.5493 + 6.6071X. Thus as the temperature increases by 1 degree, the mean peak power load increases approximately by 6.6 megawatts. The estimate of the model standard deviation σ is 17.102 and 87.65% of the variation in peak power load is explained by maximum temperature. Notice that the p-value of 6.7E-05 for the t-test about the slope (equivalently F-test for significance) indicates that linear regression on maximum temperature is very useful in explaining peak power load. The 95% confidence interval for the slope shows that the mean peak power load increases by a number between 4.5854 and 8.6288 as temperature increase by 1 degree.