Using Linear Regression to Explore Environmental Factors Affecting Vector-borne Diseases

STUDENT GUIDE FOR STATISTICS

Excel Instructions

The first step is to open Microsoft Excel. Please do so.

Select a Blank Workbook. Save it now. Call it "Analysis of Disease XXXXX" or some other appropriate file name.

We are going to use a small sample data for illustration. For a selection of vehicles, the weight of the vehicle (in pounds) was measured and the efficiency of that same vehicle – measured in miles per gallon (mpg) – was captured under a testing situation.

Now, we will put the data onto the spreadsheet. Look at the following table screenshot for the suggested format.

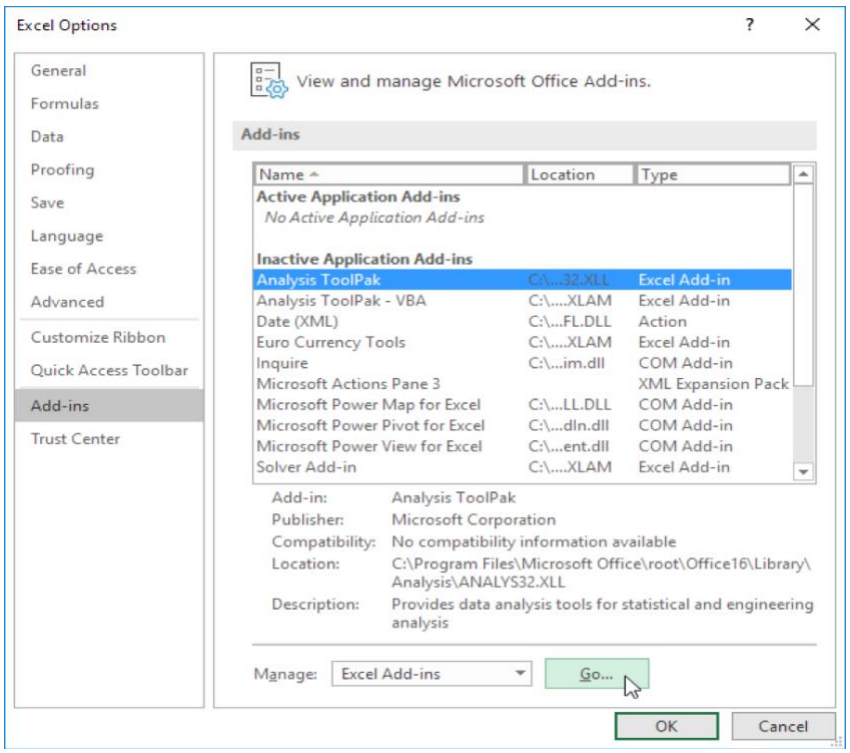
Weight (X)	MPG (Y)	Predict MPG	Residual
2900	31		
3500	27		
2800	29		
4400	25		
2500	31		
3400	29		
3000	28		
3300	28		
2800	28		
2400	33		

Create column headings so you won't get lost. The first column is your independent variable which, in this case, is vehicle weight. In many texts, this variable is also called "explanatory." The second column is your dependent variable, the vehicle MPG. Likewise, other statistical references also refer to this Y as the "response." (Note that in Excel the first column is always the independent (X) and the second column is always the dependent (Y). Please be careful.) Skip a line under the headings for better organization. Create two more columns that will be filled in later. These are the predicted values of the MPG that will be generated by the regression equation, and the residuals. These residuals are the set of differences between "what you observe" and "what you would predict" using the model you construct.

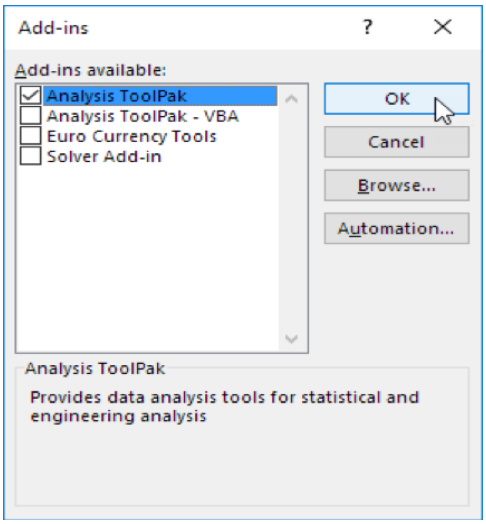
To perform some of the analysis, and to obtain some of the common statistics, you will need the Analysis ToolPak. It is a collection of operations within EXCEL – you just don't see it. It must be called to the front to put in motion.

To load the Analysis ToolPak add-in, please execute the following steps :

- ➔ On the **File** tab, click **Options**.
- ➔ Under **Add-Ins**, select "Analysis ToolPak," and click on the **Go** button.
- ➔ Check "Analysis ToolPak" and click on **OK**.



➔ Within the "Add-In" window, click on **OK**.



The Analysis ToolPak is now ready to assist you in the analysis. We will come back to it after we get a picture of the data. That picture will be a "scatterplot." It is a good idea to get a visual display first before getting into the weeds. The plot will provide insight into examining a relationship between two variables. The plot may indicate that there is a reasonable or even strong linear – or straight line -- relationship. That relationship may be direct or inverse. Or, on the downside, the plot may infer that one variable has nothing to do with the other. That being the case, further numerical analysis may be futile.

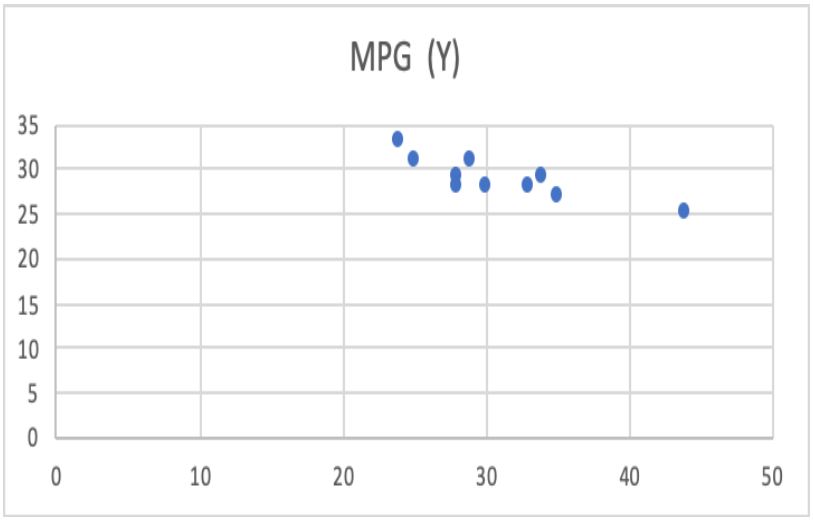
Here are the steps for getting a scatterplot:

➔ Highlight the two columns of data.

Weight (X)	MPG (Y)	Predict MPG	Residual
2900	31		
3500	27		
2800	29		
4400	25		
2500	31		
3400	29		
3000	28		
3300	28		
2800	28		
2400	33		

➔ From the menu, click on **Insert**.
➔ Then click on **Recommended Charts**.
➔ The select **Scatter** (meaning scatterplot)

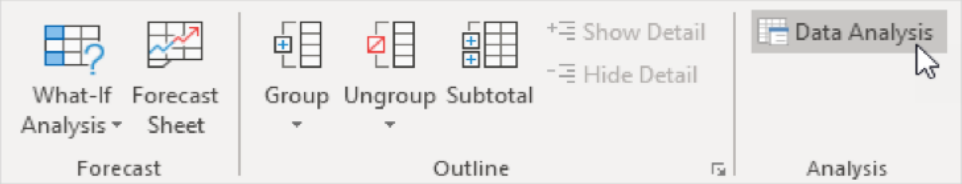
This plot is produced:



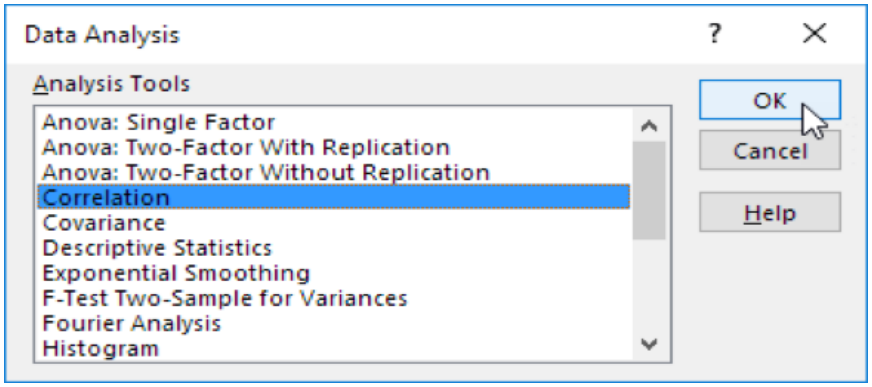
There seems to be a strong relationship between the miles per gallon achieved by a vehicle and its weight. "Strong" means that, subjectively, you can see that as X increases, Y decreases in a fairly straight-line fashion. The possible relationship portrayed in the plot is a motivation to go to the next step. That step would be acquiring the correlation coefficient. The picture indicates that this statistic should be a high negative number, perhaps fairly close to -1.00.

To go further, we will use the Analysis ToolPak.

On the "Data" tab, in the "Analysis" group, you can now click on **Data Analysis**. It is on the far right of the screen.

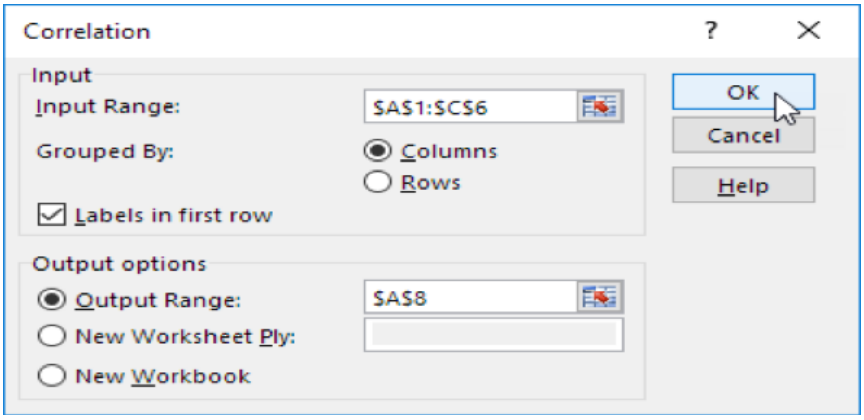


The following dialog box below appears. Click on **Correlation**.



Follow these steps :

- ➔ Click on **OK**.
- ➔ With the cursor in **Input Range**, highlight the numbers in both columns. Exclude the column headings.
- ➔ For **Output Range** , choose a cell location far from the table to avoid overwriting the data. For example, **$G$1**.



Note that the input range in the figure above does not represent the analysis we are now working on. The important item is the figure itself, not the content within it.

The designated cell (G1) is the top left corner of the correlation table that is produced.

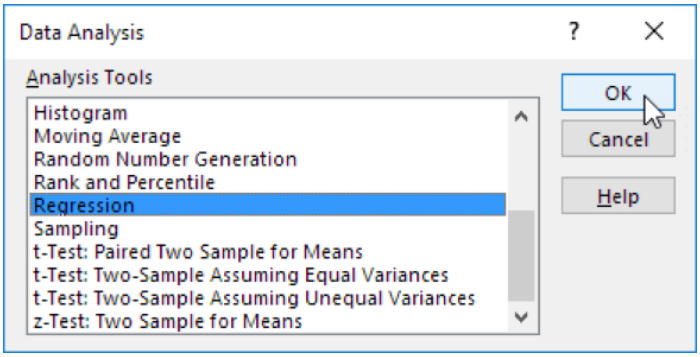
	Column 1	*Column 2*
Column 1	1	
Column 2	-0.85144	1

The coefficient of weight and efficiency is **-.85** , indicating a strong inverse relationship (as depicted by the scatterplot).

With a scatterplot that portrays a strong linear relationship, and with a correlation coefficient value that brings home the point that weight and efficiency are very much related, the next logical step is to "quantify" that relationship with a model – specifically a linear regression model. This model will provide the "best fit" of the data by minimizing the sum of the squared differences, each difference resulting from subtracting the predicted value for MPG from the observed value for MPG for a given value of vehicle weight. The theory behind this is contained in any number of statistics texts.

Here is the series of steps for regression :

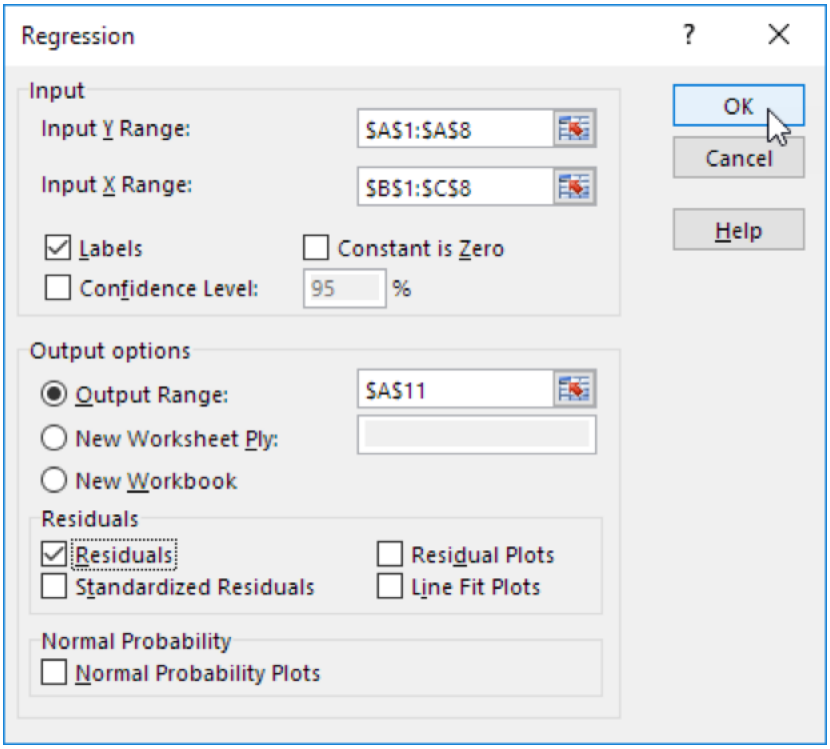
- ➔ Click on the **Data** tab.
- ➔ Then **Data Analysis**.
- ➔ Then **Regression.**
- ➔ Then **OK.**



In the table, select the Select the **Y– range** by highlighting the column of Y-values which, in our case, is MPG. In this example, the range for Y will be reflected as $B$3:$B$12. Again, just highlight the values only.

Then, in the table, select the **X – range** by highlighting the column of X-values which, in our case is weight. In this example, the range for X will be reflected as $A$3:$A$12.

As was the case for correlation, the ranges given for both X and Y in the figure illustration do not represent the current analysis of MPG and weight.



- ➔ For the **Output Range**, select a location away from the data just as you did when finding the correlation coefficient. For example, **$I$1**.
- ➔ Check on **Residuals**. You may also check on Residual Plots, Normal Probability Plots, and other items that you may be interested in or which may contribute to your analysis.
- ➔ Click on **OK** only after you have made all your selections.

Here is the output produced for this example:

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.851438016
R Square	0.724946695
Adjusted R Square	0.690565032
Standard Error	1.26984251
Observations	10

ANALYSIS OF VARIANCE (ANOVA) TABLE					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	34.00	34.00000	21.08527	0.00177
Residual	8	12.90	1.61250		
Total	9	46.90			

RESIDUAL OUTPUT		
Observation	*Predicted Y*	*Residuals*
1	29.56667	1.43333
2	27.56667	-0.56667
3	29.90000	-0.90000
4	24.56667	0.43333
5	30.90000	0.10000
6	27.90000	1.10000
7	29.23333	-1.23333
8	28.23333	-0.23333
9	29.90000	-1.90000
10	31.23333	1.76667

There are many quantitative values to review, and each one has an interpretation. The most important table (for our purposes) that Excel produces is the Model Coefficients Table, depicted in the following screenshot.

MODEL COEFFICIENTS								
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	39.23333	2.28590	17.16318	0.00000	33.96204	44.50463	33.96204	44.50463
X Variable 1	-0.33333	0.07259	-4.59187	0.00177	-0.50073	-0.16594	-0.50073	-0.16594

Two important quantities are in this table. The slope of the best-fit line is calculated to be -.3333 (entitled “X-Variable” in the table). The y-intercept is 39.2333. The “best fit” line for the data then becomes Y (MPG) = 39.23333 - .33333* X (Weight).

The Coefficient of Determination -- the R-Squared value -- is .7249. This is also – for a one-variable linear relationship – the square of the correlation coefficient. This value implies that 72% of the variability in the measured miles per gallon for the set if vehicles can be accounted for by the weights of those vehicles. The remaining 28% of the variability in MPGs would be due to other variables sources as types of terrain driven, highway versus urban driving, driver behavior, fuel types used, and many other factors to include simple error.

To complete the table we started with, we can finish the other two columns – the predicted values and the residuals. In Cell C3, enter “ = 39.2333-.33333*A3 “ and drag it down to the other cells in that column. In Cell D3, enter “ = B3-C3,” and drag it down through the column.

Weight (X)	MPG (Y)	Predict MPG	Residual
29	31	29.5667	1.4333
35	27	27.5668	-0.5667
28	29	29.9001	-0.9001
44	25	24.5668	0.4332
25	31	30.9001	0.0999
34	29	27.9001	1.0999
30	28	29.2334	-1.2334
33	28	28.2334	-0.2334
28	28	29.9001	-1.9001
24	33	31.2334	1.7666

A discussion of the other values in the regression output, and the deliberation of aspects of a particular disease to include other possible explanatory factors, are dependent upon sophistication of the students, judgment of the faculty member, interest and motivation of the class, and time available.

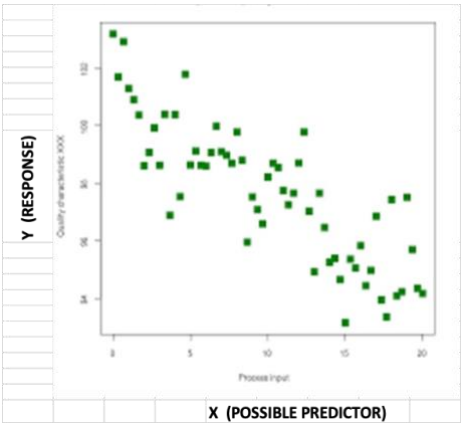
MATHEMATICAL BACKGROUND

STATISTICAL TERMS RELATED TO ANALYSIS

The "Y" (Dependent Variable) This is the entity that you want to analyze and say something about. Your goal is to discover other quantities – independent variables, sometimes called explanatory variables – that may contribute to different values of Y. Your interest in Y – also termed a response variable in many statistical references -- prompts you to gather one or more of these possible "X s," each of which may have a relationship with Y. The end result is a predictive model. An example for an independent variable would be the miles per gallon (mpg) attained by a vehicle.

The "X" (Independent Variable) This is the entity you are going to explore to see if a strong relationship exists with Y. If the relationship is evident, you may develop a predictive model to estimate values for Y given specified values of X. An example of an independent variable is the weight of a vehicle (in pounds or kilograms) that possibly affects the mpg achieved.

Scatterplot This is the plot of points (X, Y), gathered through a sample, that may portray the existence of a relationship between the independent and dependent variables. However, in this exercise, the only relationship that would be pursued is one for which the scatterplot forms a straight line – a linear relation.



Y_0 This is an observed value of the dependent variable in a sample for an observed or controlled value of the independent variable X.

"r" (Correlation Coefficient) This is a measure of the strength and direction of the linear relationship between the two quantitative variables X and Y. This coefficient ranges from -1 to +1. Values close to -1 indicate a significant inverse relationship, meaning as X increases, Y will decrease. Values close to 1 indicate both variables are increasing or both are decreasing. Values close to zero signify little or no relation between the two entities.

"$Y_P = b_0 + b_1 X$" (Regression Line / Line of "Best Fit") This is the model that, for a linear relationship between the two variables, estimates an average value of Y for an input value of X. An example would be using the weight of a vehicle (X) to estimate the average miles per gallon (Y) achieved for a vehicle of that weight. "Y_P " is called the predicted value of the response from the model that has X as a predictor.

The quantity " b_0 " is the Y-intercept of the straight line. In most algebra courses, and in some statistics courses, the quantity is also designated as simply "b." Similarly the slope in an algebra course is usually represented by " m. " In statistics, owing to convention, this same quantity is usually given the label "b_1."

Therefore " Y = m X + b" and "$Y_P = b_0 + b_1 X$ " depict exactly the same relationship.

The values of the regression coefficients are calculated as

slope $b_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

and

vertical y intercept $b_0 = \bar{y} - b_1\bar{x},$

where $\bar{x}$ and $\bar{y}$ are the arithmetic means (or averages) of the x and y data, respectively.

EXAMPLE: Given the set of bivariate data

x	2	5	2	4	6
y	4	7	5	8	11

Find the best fit line using a least squares linear regression.

Note that $\bar{x} = 3.8$ and $\bar{y} = 7$. We calculate the denominator and the numerator for the slope.

$$\begin{aligned} denominator &= (2-3.8)^2 + (5-3.8)^2 + (2-3.8)^2 + (4-3.8)^2 + (6-3.8)^2 \\ &= 12.8 \\ numerator &= (2-3.8)(4-7) + (5-3.8)(7-7) + (2-3.8)(5-7) \\ &\quad +(4-3.8)(8-7) + (6-3.8)(11-7) \\ &= 18 \end{aligned}$$

We can calculate the slope

$$b_1 == \frac{18}{12.8} = 1.4$$

and then the y-intercept

$$b_O = 7 - (1.4)(3.8) = 1.7.$$

The equation for our best fit line using the least square linear regression method is

$$y = 1.4x + 1.7.$$

Resource: Bodine, E. N., Lenhart, S., & Gross, L. J. (2014). *Mathematics for the life sciences*. Princeton, NJ: Princeton University Press.

"e" (Residual) This is the difference between an observed value of Y obtained by the sample (Y_0) and the predicted value generated for a value of X (Y_P). Formula-wise, "$e = Y_0 - Y_P$." The sum of all the residuals is zero. The smaller in magnitude the set of residuals is, the better the linear model fits the data.

"R_2" (Coefficient of Determination) This percentage represents the amount of variability in the response (Y) (dependent) accounted for by the presence of a predictor (X) (independent). The remaining percent – " 100% - R_2 " – may be accounted for by one or more other explanatory variables or simply error. For a linear model, " $R_2 = r_2$."

Note a low coefficient of determination means a weak correlation between the two features being considered. In some ecological data, weak correlations can often occur, but even a weak correlation can be important. Sometimes a feature with a weak correlation can be combined in a regression with another feature and corresponding observed output to obtain a stronger correlation.

Co-Authors (Alphabetical): Andy Adams, Jessica Adams, John Bray, Tami Imbierowicz, Suzanne Lenhart, Breonna Martin. All co-authors contributed equally. This material is based upon work supported by the National Science Foundation under Grant No. 1919613. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.