



Published on *Plant Breeding E-Learning in Africa* (<https://pbea.agron.iastate.edu>)

[Home](#) > [Course Materials](#) > [Quantitative Methods](#) > Linear Correlation, Regression and Prediction

Linear Correlation, Regression and Prediction



By Ron Mowers, Dennis Today, Kendra Meade, William Beavis, Laura Merrick (ISU)



Except otherwise noted, this work is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/).

Introduction

Determining the relationship between two continuous variables can help us to understand a response to an associated action. The concept of linear correlation can illustrate a possible relationship between two variables. For instance, the ideas that more rainfall and more fertilizer available to a crop produce greater yield are very plausible. To determine whether a relationship exists statistically employs the use of linear models. Once a relationship is established, methods of linear regression can be used to quantify the amount of response and strength of the relationship, such as finding that 5 cm of additional precipitation produces a 30 kg ha⁻¹ yield increase or applying 10 kg ha⁻¹ less N reduces yields by 40 kg ha⁻¹. For students of plant breeding, the concepts of regression and prediction will be fundamental to understanding Quantitative Genetics and Breeding Values.

Objectives

- The proper use of and differences between correlation and regression
- How to estimate a correlation relationship from a scatter plot
- How to establish a linear relationship between a dependent variable and an independent variable using regression methods

Correlation

Correlation Coefficient

Correlation is a measure of the strength and direction of linear relationship.

The Pearson Correlation Coefficient (r), or correlation coefficient for short, is a measure of the degree of linear relationship between two variables. The measure determines how close to linear is the change in one variable with respect to the other. The emphasis is on the degree to which they vary linearly. Later in this lesson we will discuss regression where the interest is in rate of change, how one variable is predicted by the other. In correlation the strength of the relationship is of interest.

The correlation coefficient may take any value between 1 and -1 .

The sign of the correlation coefficient (+, -) defines the direction of the relationship, either positive or negative. A positive relationship means that a positive change in one variable is related to a corresponding positive change in the other (e.g. more fertilizer produces more yield), while a negative relationship produces a negative result (e.g. increasing numbers of black cutworms decreases yields).

The absolute value of the correlation coefficient describes the strength of the relationship. A correlation coefficient of 0.50 indicates a stronger degree of linear relationship than one of $r = 0.40$. Likewise a correlation coefficient of $r = -0.50$ indicates a greater degree of relationship than one of $r = -0.40$. Thus a correlation coefficient of $r = 0.0$ indicates the absence of a linear relationship; correlation coefficients of $r = +1.0$ and $r = -1.0$ indicate perfect linear relationships.

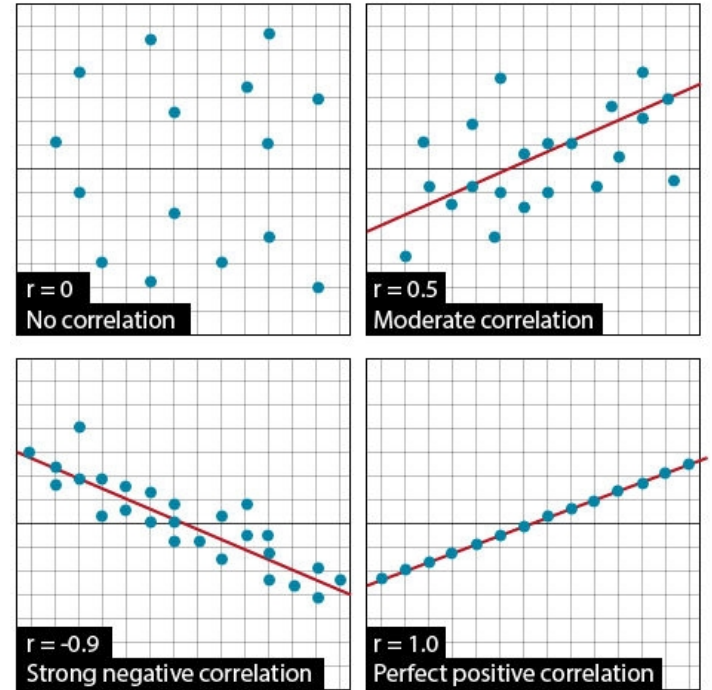


Fig. 1 Example scatter plots and the r -values associated with them.

Scatter Plots

A straightforward and necessary way to visualize correlations is through the use of scatter plots. Usually, the dependent variable is plotted on the vertical axis of the plot while the other variable is plotted on the horizontal axis. Such a plot can provide evidence of a linear relationship between the variables. An example is shown below in Fig. 2.

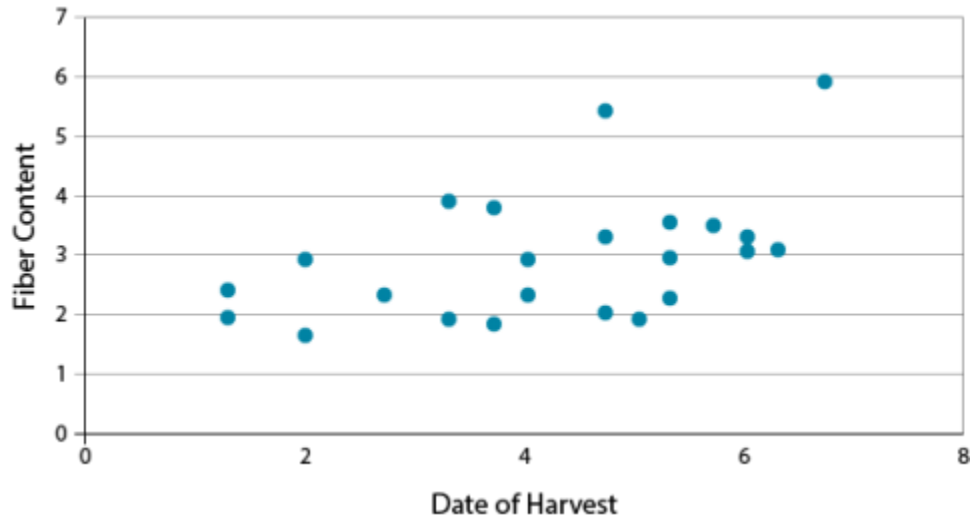


Fig. 2 Fiber content plotted as compared to the date of harvest. The fiber content seems to increase linearly as the date of harvest increases.

Try the correlation exercise in the next screen!

Try This: Correlation

How well related are these two measurements? What is their correlation coefficient?

Applet: Try estimating some correlations here using randomly generated data sets.

Data	
COLLECT	
Regression line	
MSE of your guesses:	
Calculated by least squares:	
SHOW	

Study Question 1

A paper in Science found a correlation of 0.9 between the length of the sunspot cycle and global temperature change. What can we interpret from that data?

- ☐ The length of the sunspot cycle is changing global temperatures.
- ☐ We can't say anything conclusive because we don't know the physical relationship between the two.
- ☐ There is no relationship between sunspots and global temperatures.
- ☐ We are not sure if the results are significant.

✓ Check

Correlation: Calculating r

Correlation is an often misused concept and statistic. When two things are correlated, what does that really mean? Misconceptions about correlations between variables are common. Correlations can also be totally spurious. For example, a positive relationship between the number of sheep in the United States and the number of golf courses does not mean that sheep numbers have increased because there are more golf courses. Both variables are likely to be related to an underlying trend of increasing population in the U.S. Many things can be correlated, but it is the physical or biological relationship that gives a correlation relevance. Correlation only states the degree of linear association (not cause and effect) between the two variables.

Calculation of r involves estimating the co-variance of two variables, or how much they vary together. The correlation is defined in the equation below, where one of the variables is represented by x and the other by y .

$$r = \frac{S_{xy}}{\sqrt{S_{xx}} \times \sqrt{S_{yy}}}$$

Equation 1

where:

$$S_{xy} = \text{sum of products} = \sum xy - \frac{\sum x \sum y}{n}$$

$$S_{xx} = \text{sum of squares of } x = \sum x^2 - \frac{\sum (x)^2}{n}$$

$$S_{yy} = \text{sum of squares of } y = \sum y^2 - \frac{\sum (y)^2}{n}$$

The correlation equation may seem monstrous at first. Do not panic! Actually, the concept behind the equation is closely related to the z-scores we calculated earlier.

The numerator, the sum of squares of xy (S_{xy}) measures the combined distances of all points from the center of the plot (\bar{x}, \bar{y}) . The more closely X and Y are related, the greater this value will be.

The denominator is the product of the square roots of the sums of squares of X and Y . The product of these two roots quantifies how much X and Y vary independently of each other.

Thus, r is the ratio of the amount that X and Y vary together to the amount X and Y vary total. The more X and Y vary together, the greater the ratio will be. The maximum possible values (1 or -1) occur when all variation in X and Y is related.

How individual variables vary is of interest. If large Y 's are associated with large X 's, it would stand to reason that there would be a positive correlation between the variables.

Correlation Example

Some measurements were taken on the amount of flow, Y (m^3/s), in a normally dry drainage ditch, next to a field. These were measured from run-off after 30 minute rainfalls. The total rainfalls were designated as X and measured in millimeters (mm). Hydrologists wanted to know how much water ran off under different conditions and how closely the two measurements were related in this field. The relevant sums are included along with the computational form of the r calculation.

Table 1

x	x^2	y	y^2	sy
2.6	6.76	0.1	0.01	0.26
12.2	148.84	1.3	1.69	15.86
14.1	198.81	2.5	6.25	35.25
14.6	213.04	3.5	12.25	51.1
15.2	231.04	9.1	82.81	138.32
15.6	243.36	9.3	86.49	145.08
15.9	252.81	12.2	148.84	193.98
17.4	302.76	13.2	174.24	229.68
18.8	353.44	15.9	252.81	298.92
19.0	361.0	19.3	372.49	366.7
Σx 145.4	Σx^2 2311.98	Σy 86.4	Σy^2 1137.88	Σxy 1475.15

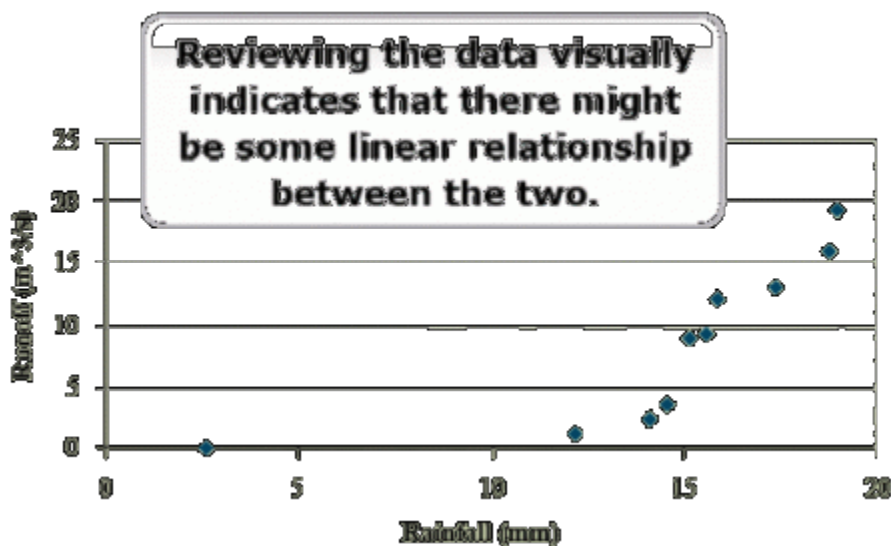


Fig. 3 Runoff from a field as a function of rainfall.

Correlation Example Calculations

$$r = \frac{S_{xy}}{\sqrt{S_{xx}} \times \sqrt{S_{yy}}}$$

Equation 1

where:

$$S_{xy} = \text{sum of squares of product} = \sum xy - \frac{\sum x \sum y}{n} = 1476.8 - \frac{(145.5)(86.5)}{10} = 218.2$$

$$S_{xx} = \text{sum of squares of } x = \sum x^2 - \frac{(\sum x)^2}{n} = 2313.4 - \frac{145.5^2}{10} = 196.4$$

$$S_{yy} = \text{sum of squares of } y = \sum y^2 - \frac{\sum y^2}{n} = 1139.4 - \frac{86.5^2}{10} = 391.2$$

$$r = \frac{218.2}{\sqrt{196.4} \times \sqrt{391.2}} = 0.79$$

The computed r value is 0.79. This is a moderately large correlation. How large the correlation is depends upon the variability of the data. Correlations can range well above 0.9 or below -0.9 in many cases. Physically, there would seem to be a cause and effect here. Heavier rainfall would produce more run-off, while light rainfall produces little or none. The best indicator of that can be seen at the heaviest rainfall rates, where magnitude of the run-off increases substantially. The one data point at lower rainfall levels is problematic. We assume it is real. Occasionally, a single outlier data point can slightly skew a relationship. Although not as linearly related to the other data, it does fit the plausible model: lighter rainfall, less run-off.

Table 2

X	X ²	Y	Y ²	XY
2.6	6.76	0.1	0.01	0.26
12.2	148.84	1.3	1.69	15.86
14.1	198.81	2.5	6.25	35.25
14.6	213.04	3.5	12.25	51.1

X	X²	Y	Y²	XY
15.2	231.04	9.1	82.81	138.32
15.6	243.36	9.3	86.49	145.08
15.9	252.81	12.2	148.84	193.98
17.4	302.76	13.2	174.24	229.68
18.8	353.44	15.9	252.81	298.92
19.0	361.0	19.3	372.49	366.7
145.4	2311.98	86.4	1137.88	1475.15
Σx	Σx^2	Σy	Σy^2	Σxy

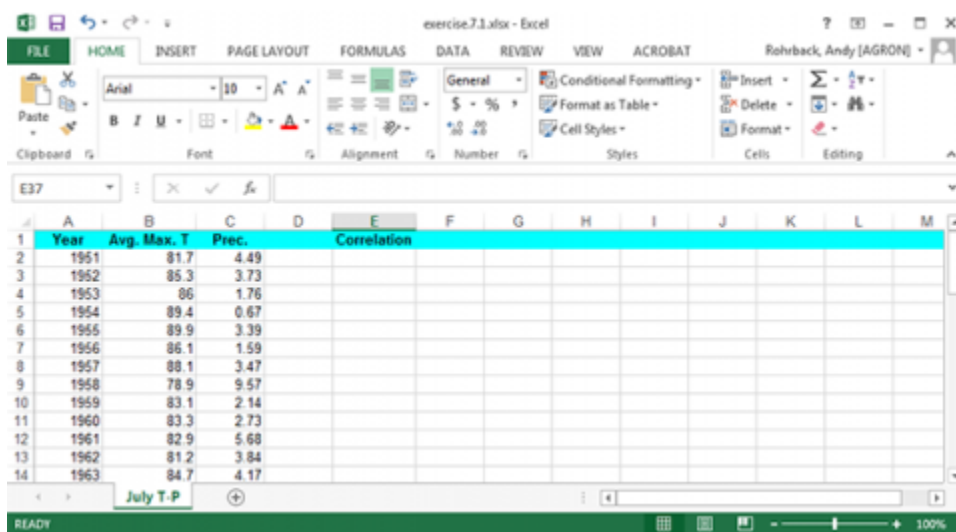
Try the Correlation Exercise in the Next Screens!

Ex. 1: Correlation Exercise (1)

Exercise 1: Calculating the Correlation in a Bivariate Set of Data

As shown in the examples displayed in the text, a good first guess in establishing a relationship between variables is to view the data on an X-Y scatter plot. Trends should start to appear. We will produce a scatterplot in Excel and determine the correlation.

- Open the [QM-mod7-ex1data.xls](#) workbook. It is July weather data with average temperature and precipitation data.
- We will begin by producing a scatter plot to view the data. For this analysis, the Temperature will be assigned to the x-axis and Precipitation to the y-axis.

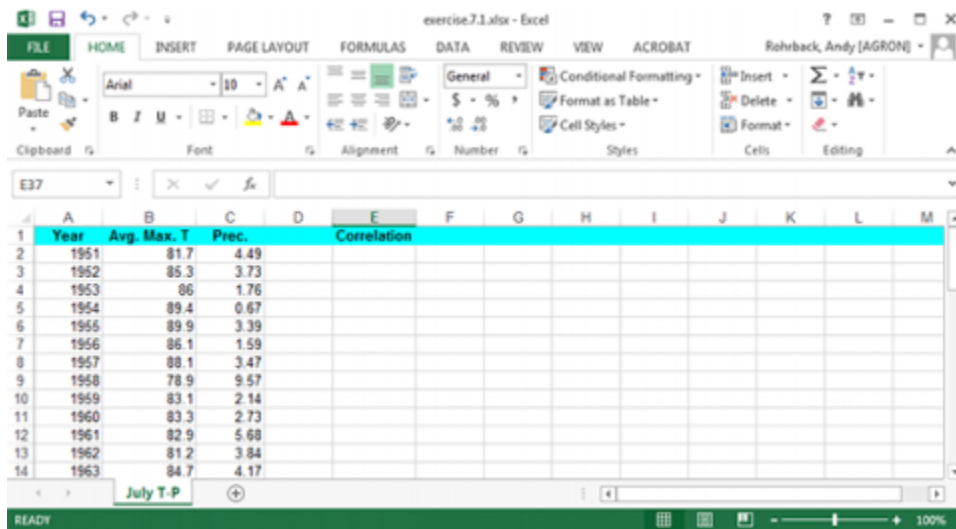


Year	Avg. Max. T	Prec.	Correlation
1951	81.7	4.49	
1952	85.3	3.73	
1953	86	1.76	
1954	89.4	0.67	
1955	89.9	3.39	
1956	86.1	1.59	
1957	88.1	3.47	
1958	78.9	9.57	
1959	83.1	2.14	
1960	83.3	2.73	
1961	82.9	5.68	
1962	81.2	3.84	
1963	84.7	4.17	

Fig. 4

Ex. 1: Bivariate Set of Data

We will begin by producing a scatter plot to view the data. For this analysis, the Temperature will be assigned to the x-axis and Precipitation to the y-axis.



The screenshot shows an Excel spreadsheet titled "exercise7.1.xlsx". The data is organized as follows:

Year	Avg. Max. T	Prec.	Correlation
1951	81.7	4.49	
1952	85.3	3.73	
1953	86	1.76	
1954	89.4	0.67	
1955	89.9	3.39	
1956	86.1	1.59	
1957	88.1	3.47	
1958	78.9	9.57	
1959	83.1	2.14	
1960	83.3	2.73	
1961	82.9	5.68	
1962	81.2	3.84	
1963	84.7	4.17	

Fig. 5

Ex. 1: Bivariate Set of Data (2)

Highlight the two columns with the precipitation and temperature data.

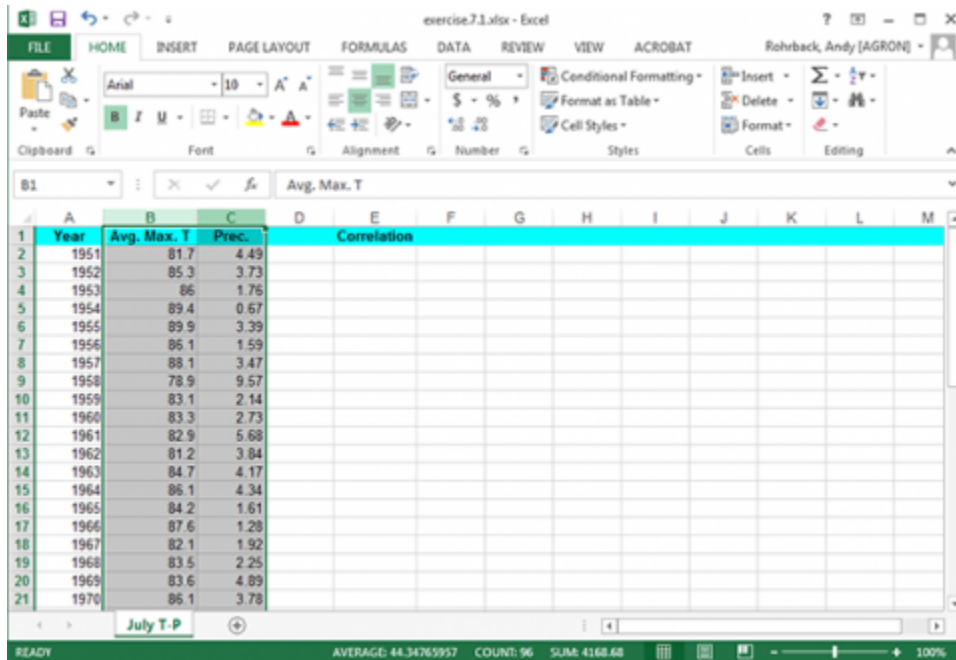


Fig. 6

Select the Insert tab and click the Scatter Plot tool. Select the first type.

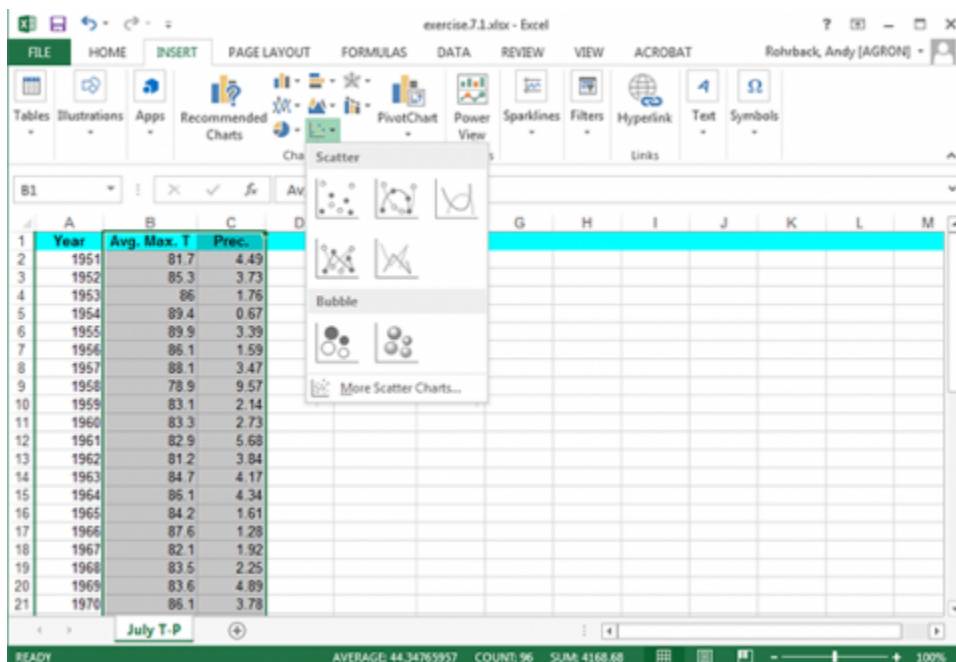


Fig. 7

Ex. 1: Bivariate Set of Data (3)

Change the x-axis label to Temperature, the y-axis label to Precipitation, and the plot title to Precipitation vs. Temperature. Click on each text box in the plot to change it.

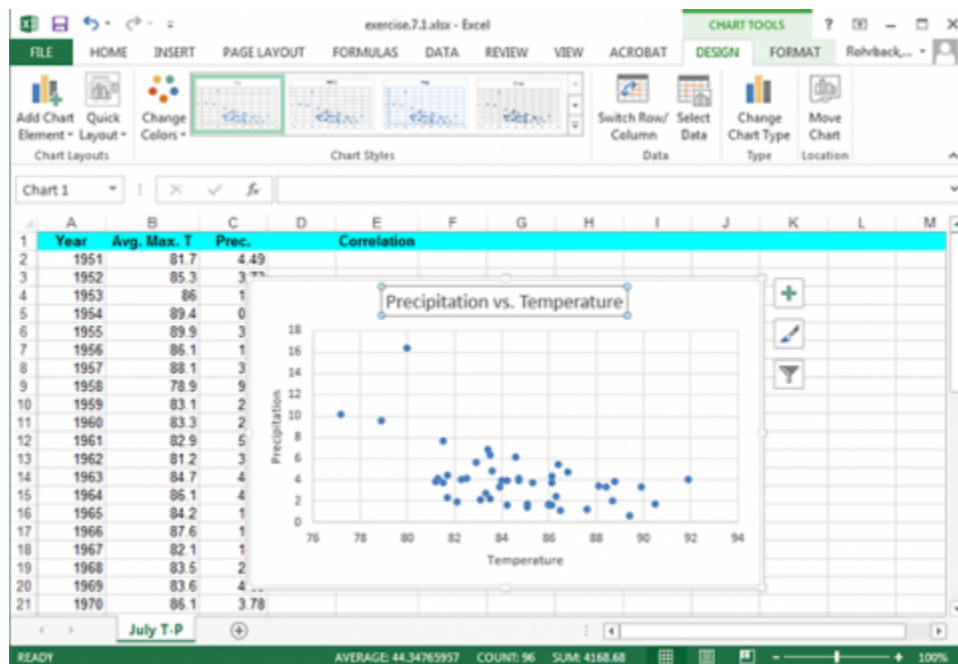


Fig. 8

Ex. 1: Bivariate Set of Data (4)

The correlation between precipitation and correlation can be easily calculated. Label a fourth column "Correlation" and enter the formula "`=Correl(B2:B48, C2:C48)`".

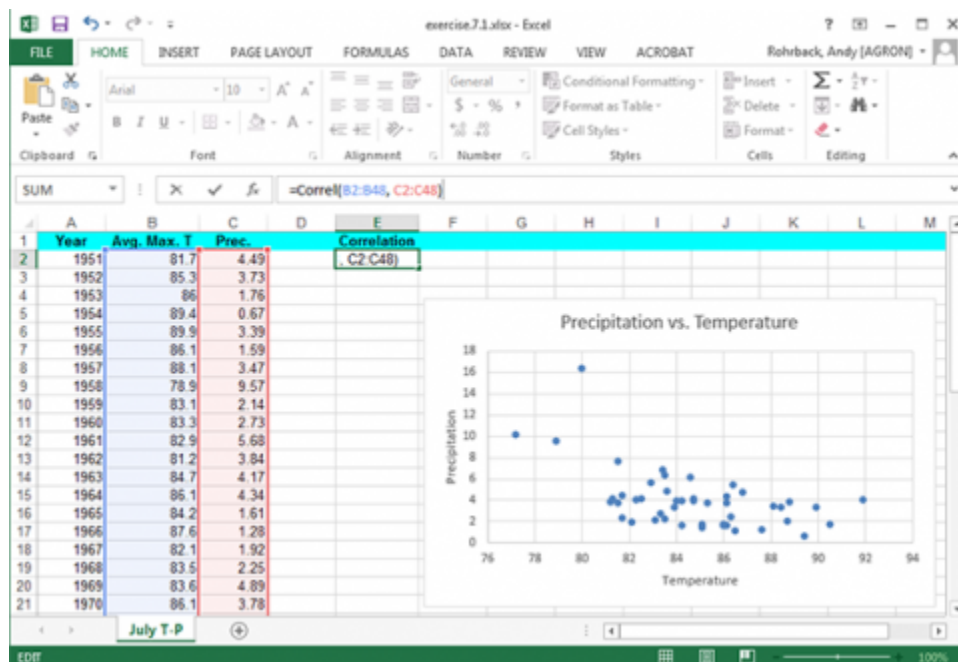


Fig. 9

Ex. 1: Bivariate Set of Data (5)

The correlation is moderately large (but not exceptionally high), with an r value of -0.534 . You may see the effects of an outlier here. The correlation between July temperatures and precipitation makes sense. This relationship follows a meteorological pattern.

More precipitation wets the soil surface, causing more latent heating and less warming of the air by sensible heating. More precipitation generally means more clouds. Both are associated with lower temperatures. The single data value at the top may skew your view of the correlation while having a rather small effect on the total correlation. There does seem to be a qualitative relationship without a very strong correlation.



Fig. 10 Rain falls on crop fields. Photo by Malene, Wikimedia Commons.

Ex. 1: Bivariate Set of Data (6)

Hold the cursor over the possible outlier in the Excel scatterplot.

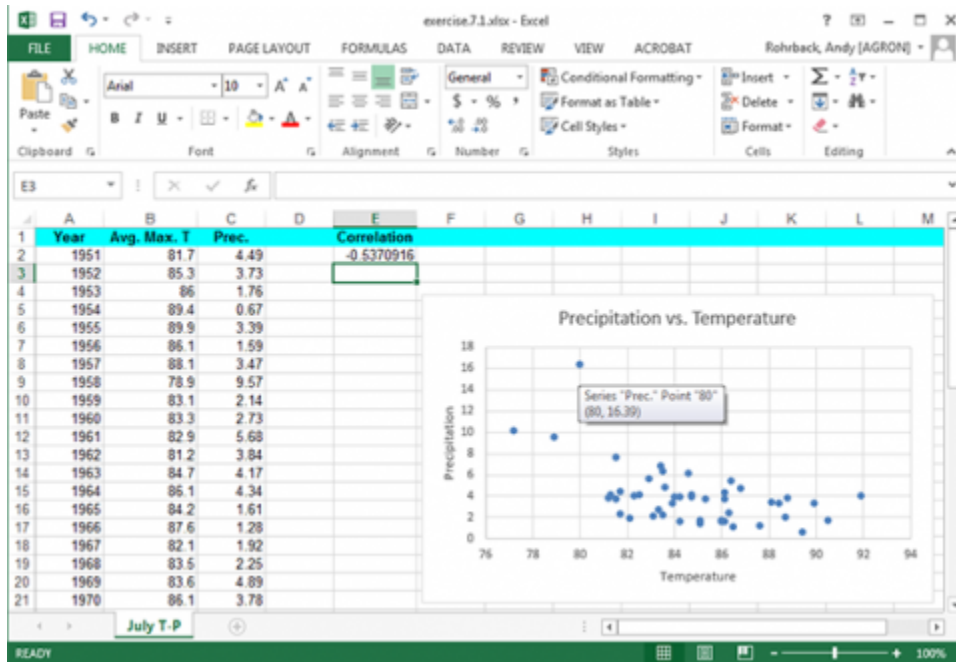


Fig. 11

Select the row with 1993 data in the original data and delete it.

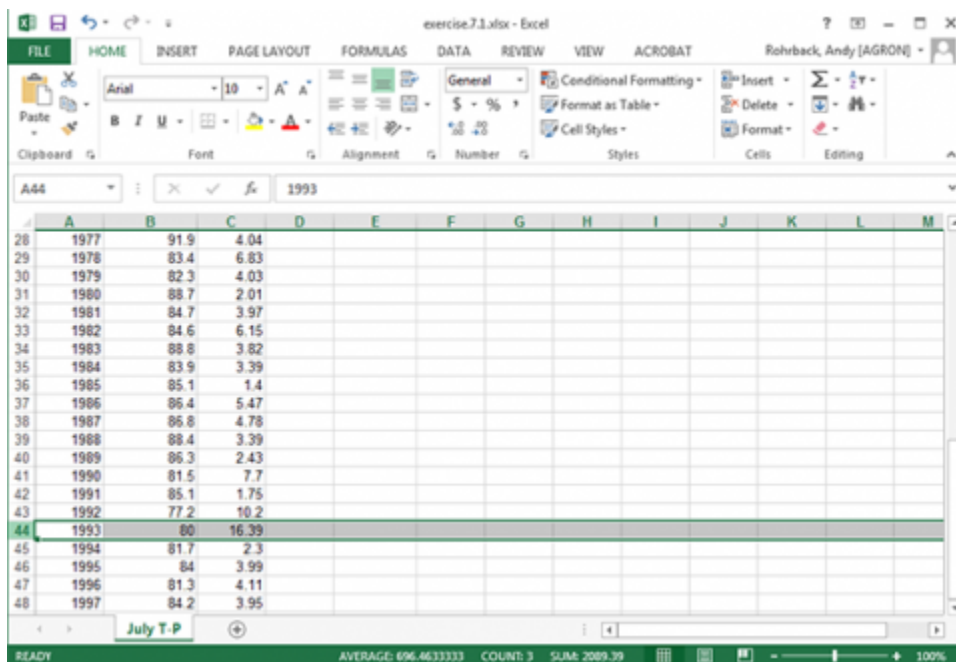


Fig. 12

The plot and correlation will change. However, be careful when discarding what appears to be an outlier. It is a good idea to try and determine if there was an obvious experimental error that can be found. If there is not, it may not be a true outlier.

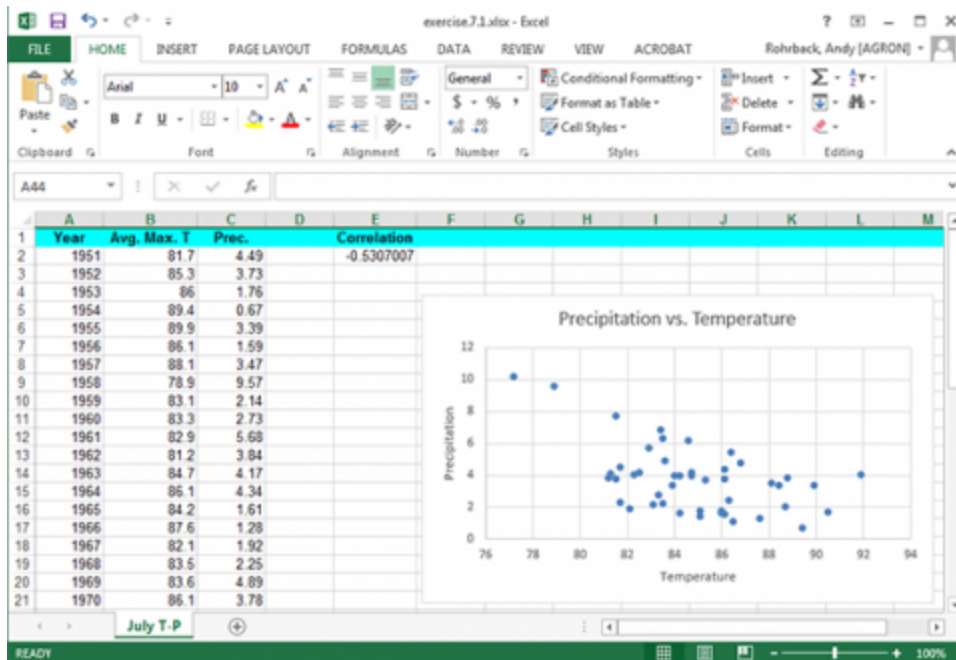


Fig. 13

Discussion: Correlation

How much did the correlation change by removing the 1993 data? What do you think about the results of this?

Linear Regression

Definition

Linear Regression establishes a predictive relationship between two variables.

While correlation attempts to establish a linear relationship between two variables, regression techniques try to determine a predictive relationship between the two. Translated, "Can values of one variable be used to predict values of the other?"

When someone wishes to apply fertilizer, the expected amount of yield gained for the amount of fertilizer applied is needed. Sound economic and ecological choices may be based on regression and relationships between variables. Understanding the regression relationship allows the producer to use the amount of fertilizer that can give the best yield or financial return for the money invested.

Several other physical variables obviously are involved in translating the fertilizer into a yield result, such as rainfall, soil fertility, pest populations, etc.



Fig. 14 Fertilizer application on a field. Photo by Iowa State University.

Regression Lines

Referring to the scatter plot diagrams in the web exercise, one can estimate the magnitude of a correlation. When establishing a regression relationship, a single line delineating the relationship is necessary. One could use several methods to estimate the linear relationship that best fits the data.

Connecting the two end points in the data or eye balling a resultant line are two examples. These will usually provide a qualitative result that lacks precision and accuracy.

In the applet, we drew a line of "best fit" by eye, and observed the least squares regression line was different. The regression line is that line which minimizes the sum of squared vertical distances of points on the line. If you mentally determined the line minimizing the perpendicular distances, it would not be the same as the least-squares regression line. The regression line is "best" in the sense of least error for the line with fixed x-values.

Applet: Try estimating some correlations here using randomly generated data sets.

Data	
COLLECT	
Regression line	
MSE of your guesses:	
Calculated by least squares:	
SHOW	

Sources of Variation

Estimating Regression Line

The preferred method to estimate a regression line is to use the data to numerically calculate the line which minimizes the error or the scatter of the points around the line. This is done using the least-squares method.

The result of a linear regression is an equation of the form ($\hat{Y} = \hat{\alpha} + \hat{\beta}x$). The hats over Y, α , and β indicate these are estimates, not the actual regression line. This equation determines the relationship between the x and a predicted \hat{Y} based on the estimated slope of the regression line and the vertical intercept ⚠ Invalid Equation.

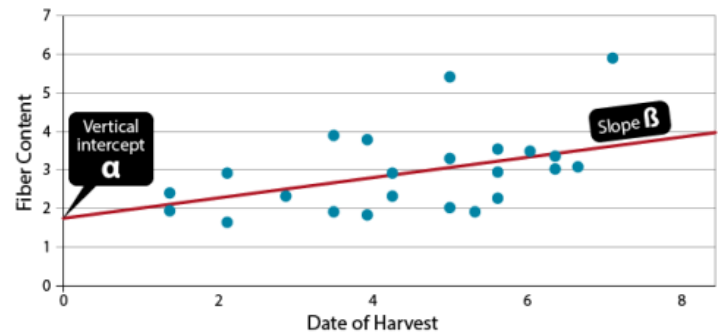


Fig. 16 Regression line statistics.

As we have discussed before, we do not know the actual relationship between the two variables.

Therefore, we estimate it, based on gathered data. We assume there is a true regression line: $y = \alpha + \beta x + \epsilon$, and we estimate intercept with α and slope with ⚠ Invalid Equation.

Point-Slope Formula

The point-slope formula to create the line can be found using sums of squares as calculated in the previous section. The slope of the line is determined using this equation.

$$\beta = \frac{S_{xy}}{S_{xx}}$$

Equation 2

where:

$$\mathbf{S_{xy}} = \text{sum of products} = \sum xy - \frac{\sum x \sum y}{n}$$

$$\mathbf{S_{xx}} = \text{sum of squares of x} = \sum x^2 - \frac{\sum (x)^2}{n}$$

Note the similarities to and distinct differences from the calculation of r. There are an infinite number of lines which can be described with this slope, thus another piece of information to describe a line is necessary.

Y-Intercept Formula

A specific point on the line (usually the vertical-intercept) along with the slope fixes a single line to the data. The Y-intercept of the line is determined by the equation below.

$$\hat{\alpha} = \frac{\sum y - \hat{\beta} \sum x}{n}$$

OR

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

Equation 3

Another point which the line intercepts is the point (\bar{X}, \bar{Y}) . Knowing a point on the line and its slope completely describes the regression line through the data.

Example Calculation: Slope

Using the data from the previous section, we can calculate the regression slope using a hand computational formula in Equation 2.

$$\beta = \frac{S_{xy}}{S_{xx}}$$

Equation 2

$$\mathbf{S_{xy}} \text{ (sum of squares of product)} = \sum xy - \frac{\sum x \sum y}{n} = 1476.8 - \frac{(145.5)(86.5)}{10} = 218.2$$

$$\mathbf{S_{xx}} \text{ (sum of squares of x)} = \sum x^2 - \frac{(\sum x)^2}{n} = 2313.4 - \frac{145.5^2}{10} = 196.4$$

$$\beta = \frac{218.2}{196.4} = 1.11$$

The Y-intercept of the data can be calculated similarly.

$$\hat{\alpha} = \frac{\sum y - \beta \sum x}{n}$$

$$\hat{\alpha} = \frac{86.5 - 1.1(145.5)}{10}$$

$$\hat{\alpha} = -7.4$$

Example Calculation: Interpretation

This line indicates that according to the measured data, the run-off will increase by 1.11 m³/s for each additional mm of rainfall. The line created is the "best" in describing the linear response of run-off to the associated rainfalls (Fig. 17).

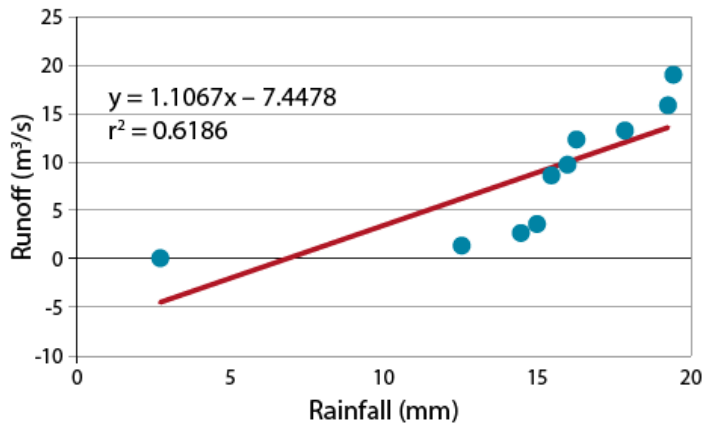


Fig. 17 The best fit line for the rainfall runoff data. The equation of the line, and r^2 value of the data are included.

The strength of the relationship is r^2 , i.e., the correlation coefficient squared. Note that the line fitted to the data intercepts the vertical axis at a negative value. An initial interpretation of "negative runoff" is clearly nonsense, but a little reflection on the nature of the problem suggests that up to a certain level of rainfall the water will infiltrate the soil before there is runoff. Thus the negative value can be interpreted as the "infiltration potential" of the soil. You may also notice that there is some bias in the way the data deviate from the regression line. The line overestimates the run-off for rainfalls from 10-15 mm and underestimates above 16 mm.

This hints that a linear relationship may not be the best choice for this relationship.

Try: Estimating Regression in the Next Screen!

Ex. 2: Estimate Regression (1)

Exercise 2: Plotting Data to Estimate Regression

We will now use data to calculate a regression line. We could have calculated a regression line for the previous data, but since it is not obvious which is the cause and which the effect for July temperatures and precipitation, using one to predict the other is somewhat questionable.

Here we will use the relationship between water stress and corn yield reduction in Iowa.

Water is the independent variable with yield being the dependent (or predicted) variable. In this case, the researcher controlled the amount of water applied and measured the yield.



Fig. 18 Predictive analysis of corn yields are key to farm economies. Photo by Iowa State University.

Ex. 2: Plotting Data

Download and open the Excel file [QM-mod7-ex2data.xls](#). Select the “Water stress” worksheet.

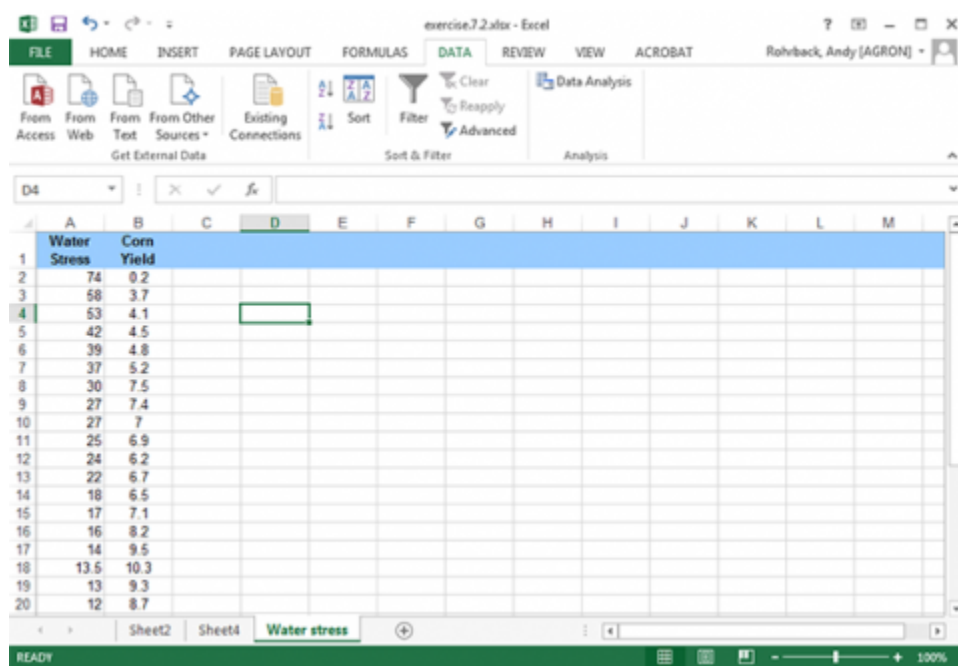


Fig. 19

Ex. 2: Plotting Data (2)

Select the Data tab and click on the Data Analysis tool.

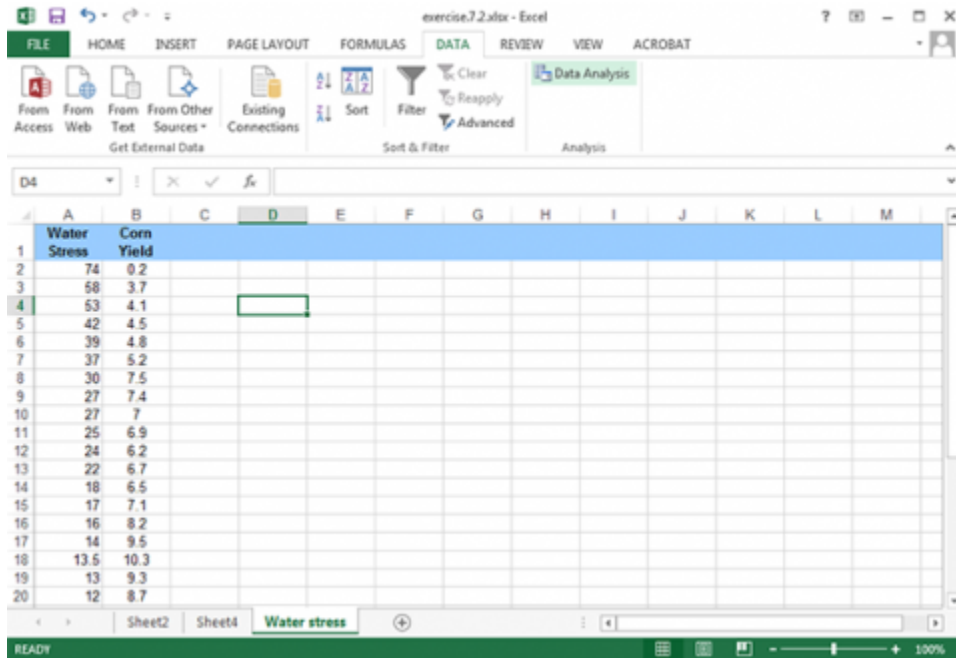


Fig. 20

Scroll down to Regression; highlight it and click OK.

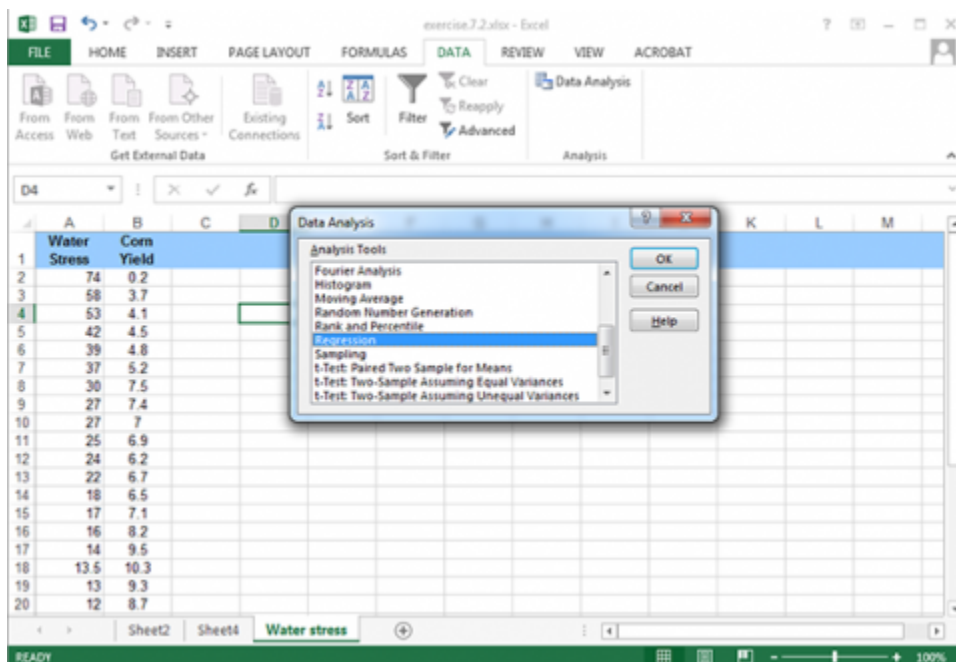


Fig. 21

Ex. 2: Plotting Data (3)

Fill in the options as shown:

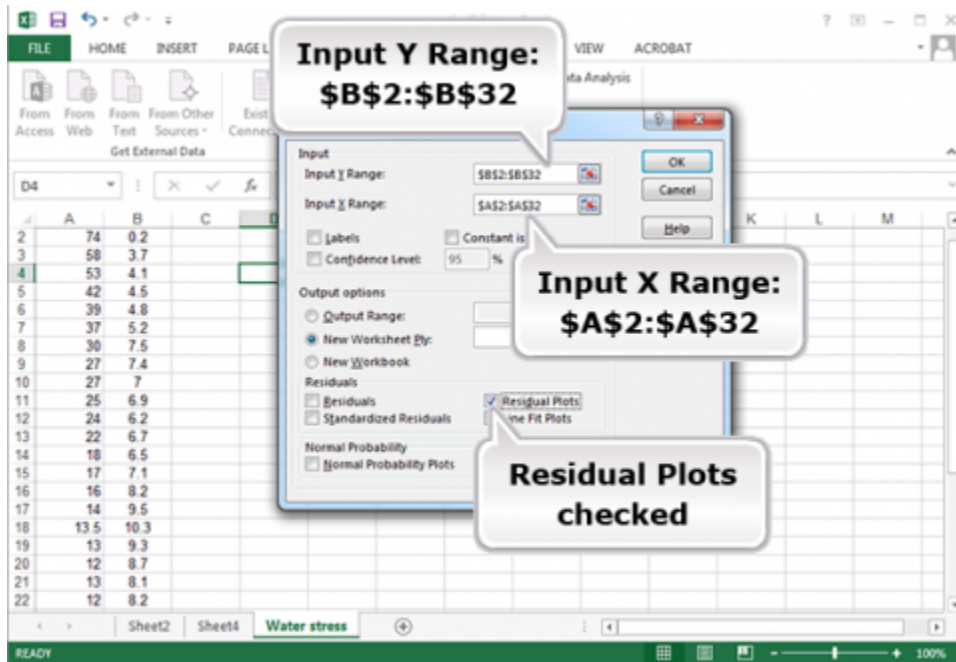


Fig. 22

Ex. 2: Plotting Data (4)

This gives the Linear Fit for the least squares regression line and an ANOVA for Regression. It also gives the residual plot.

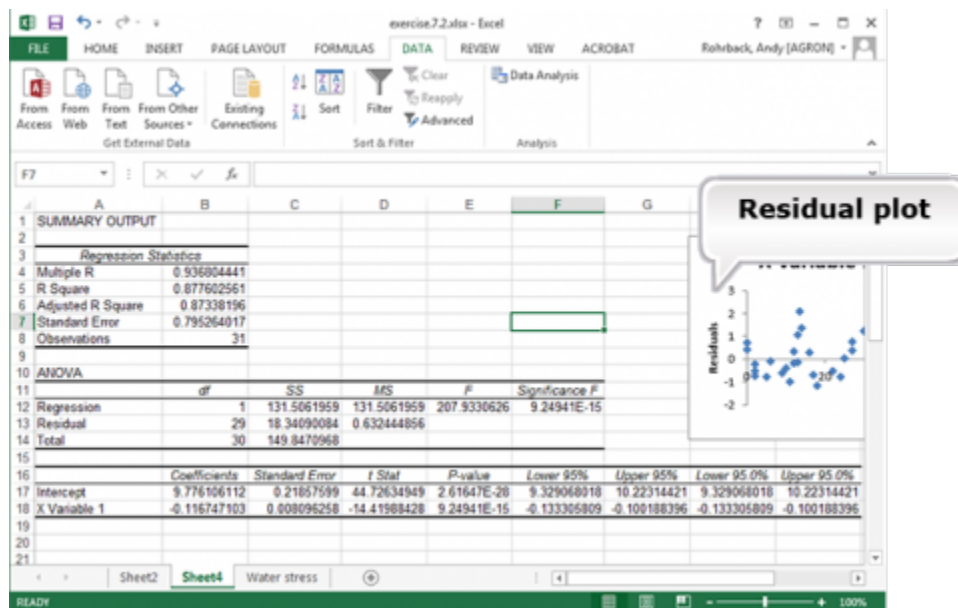


Fig. 23

Notice that the prediction equation is : $E(\text{Yield (in 1000 kg/ha)}) = 9.78 - 0.117 (\text{Water Stress})$. This can be determined from the coefficients for Intercept and X Variable 1. The regression equation is $Y = 9.78 - 0.117(\text{WS}) + \text{error}$.

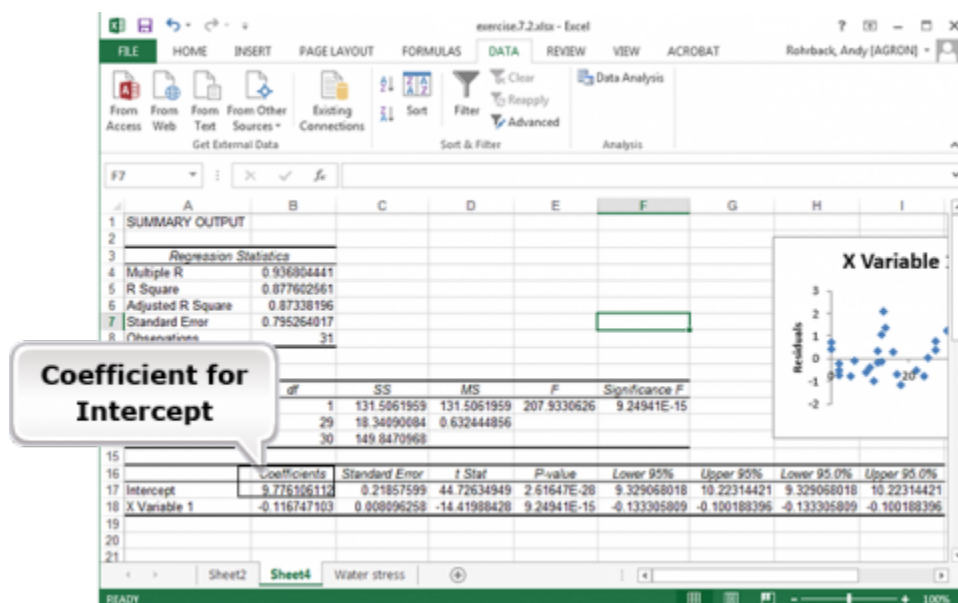


Fig. 24

Ex. 2: Plotting Data (5)

Save this analysis for Exercise 3.

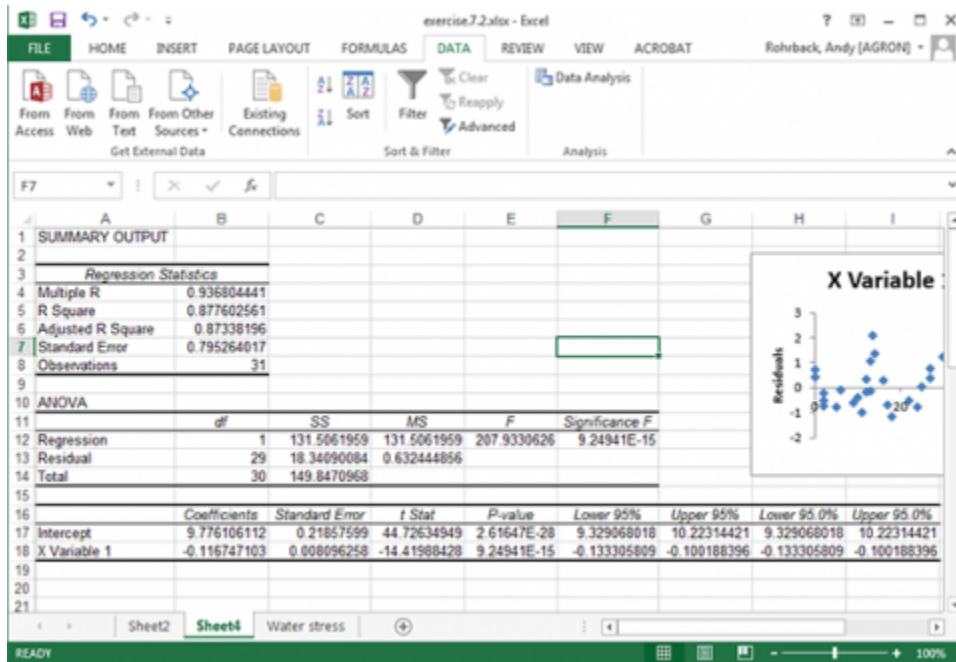


Fig. 25

Estimation Formula

Sources of Variation include the line and deviation from the line.

The line produced in linear regression is calculated to minimize the average distance of the Y-values from the line. Thus, summing the deviations of the actual Y-values from the regression-predicted values will equal 0. Measurement of observed data always has some variability associated with it due to the nature of error in experimental data. This variability can be accounted for and partitioned into its sources with an Analysis of Variance (ANOVA). Some variability of the Y's occurs because of their relationship with X. This is quantified by the squared correlation coefficient (r^2), the proportion of variance in Y that can be accounted for by linear association with X. The rest of the variability around the line cannot be accounted for (at least in its relationship with the X variable). This is attributed to error. The linear model that accounts for this is depicted in this equation.

$$Y = \alpha + \beta x + \epsilon$$

Equation 4

where:

α = estimated Y intercept

β = estimated slope

ϵ = deviation of Y value from line (error)

Errors

The true error is assumed to be in the measurement of the Y values only. It is assumed that the X's are fixed or that their measurement error is very small. The situation where the X's have error is termed a bivariate normal distribution, in which case the assumptions for regression are not valid. We saw the effect of measurement errors in the X-variable in an earlier "Try This". Some other assumptions are necessary for regression:

- for any value of X there is a normal distribution of errors
- the variances must be the same for all Y values
- the Y-values are randomly obtained and independent of each other
- the mean of the Y-values at a given X is on the regression line

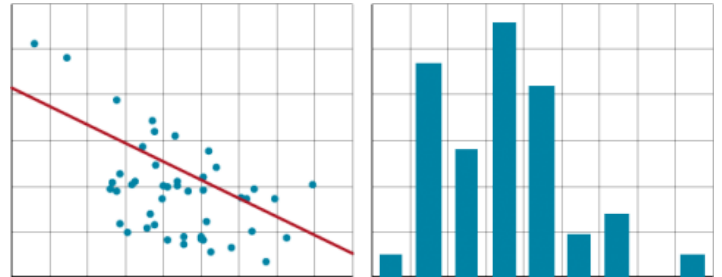


Fig. 26 Scatter plots and histograms are visual representations of variation in data sets.

Notice that these assumptions – independence of Y-value, normal distribution, constant variance, and adequacy of the model – will be essential throughout the remainder of the course.

Sum of Squares

It can be shown that the total sum of squares for Y is the sum of that associated with the regression line and that from the errors, or sum of squares not related to the relationship with X.

The correlation coefficient squared, r^2 , from the previous section describes the amount of variation attributable to the regression equation below. For example, if $r^2 = 0.75$, then 75% of the total variation in Y is accounted for by the linear regression.

The total Sums of Squared (SS) deviations in the response variable (Y) is given by

$$\sum_{i=1}^n (Y_i - \bar{Y})^2$$

Equation 5

This represents the total variability about the average response variable and is used extensively throughout this and future units.

Regression and Total SS

Because we have to estimate one parameter (the average response) in the the total SS, we need to recognize that there are n-1 degrees of freedom (df) associated with this calculation.

$$\begin{aligned}\text{Regression SS} &= \beta \times S_{xy} = \beta \times \left[\sum xy - \frac{\sum x \sum y}{n} \right] \\ \text{Total SS} &= S_{yy}\end{aligned}$$

Equation 6

where:

$$\text{Syy} = \text{sum of squares of } y = \sum y^2 - \frac{(\sum y)^2}{n}$$

or

$$\text{Syy} = \text{each observation} = \sum (y_i - \bar{y})^2$$

Partitioning Variation

The total sum of square deviations can be partitioned into regression and residual sums of squares based on these formulae.

$$\text{Regression SS} = r^2 S_{yy} \quad (\text{Residual (Error) SS} = (1 - r^2) S_{yy})$$

Equation 7

This captures the essential relationship between the correlation coefficient, the variance of the Y values, and the partition of variation into that associated with the model (Regression SS) and that which is unexplained (residual). As the residual becomes large relative to the total variance, the correlation coefficient becomes smaller. Thus, the correlation coefficient is a function of both the residual and the total variance of Y.

Statistical Significance

F-tests

Statistical Significance of the regression relationship can be tested with an F-test

The calculated regression slope is based on gathered data, from a sample. Calculations based on these data estimate the actual regression relationship. Even though we have calculated a regression coefficient, it is merely an estimate of how the variables are related. The true relationship may be slightly different from the one calculated. This could result in stating that there is a relationship when none exists. Testing the significance of the slope of the regression is done in a manner similar to other types of hypothesis testing.

The null and alternative hypotheses for this test:

$$H_0 : \beta = 0) (H_a : \beta \neq 0$$

Equation 8

Formula for F

The test is used to determine if the slope of the regression line is different from 0. Two related statistical tests may be used to test this hypothesis. The first, shown below, uses the sums of squares to determine if the regression coefficient captures enough of the variance in the data using the F test.

$$F = \frac{\text{Regression MS}}{\text{Residual MS}} \left\{ \text{with 1 and } n - 2 \text{ degrees of freedom} \right.$$

Equation 9

If the slope explains a significant proportion of the variability in the regression, then the slope is considered different from 0. If not enough variation is explained at some level of significance, often 0.05, then the slope cannot be considered different from 0.

ANOVA Table

As discussed, the variability can be partitioned. The linear regression sum of squares is calculated as shown in the Analysis of Variance (ANOVA) table or using the equation from the previous section. The ANOVA table for linear regression is shown in Table 3.

Table 3 The ANOVA table for linear regression.

Source of Variaton	Sum of Squares	Df	Mean Square	F
Regression	$\hat{\beta}S_{xy}$	1	$\frac{\text{Regression SS}}{\text{Regression Df}}$	
Residuals	$S_{yy} - \text{regression SS}$	n -2	$\frac{\text{Residual SS}}{\text{Residual Df}}$	

Notice the sums of squares and degrees of freedom for regression. The regression SS is written as a formula involving a ratio. This equates to the proportion R^2 of the Total SS of Y (Equation 6). The regression relationship has 1 df, because the test is for the slope being zero. In an ANOVA, mean squares are SS divided by df.

Example: ANOVA

Let's test the regression calculated from the previous data set. Calculating the sum of squares for Y gives a value of 391.3. Knowing the r^2 value of 0.62, we can fill in the following ANOVA table:

Table 4

Source of Variation	Sum of Squares	Df	Mean Square	F	P < F
Regression	242.1	1	242.1	12.95	0.007
Residuals	149.2	8	18.7		

The F-test is used to compare the equality of variances. In this case, we are testing whether the variance associated with the estimated slope, $\hat{\beta}$, is greater than the residual variance.

A calculated F value greater than the critical F value indicates the slope is significantly different from zero. Alternatively, most statistical software will calculate the probability of a given F value.

Ex. 3: Calculating a Regression Line and Testing the Slope (1)

In this exercise we will use Excel to find the regression equation. Calculations from the raw data are possible, but equations can be calculated easily using statistical software.

Return to the water stress computer output from the last exercise or re-run that analysis.

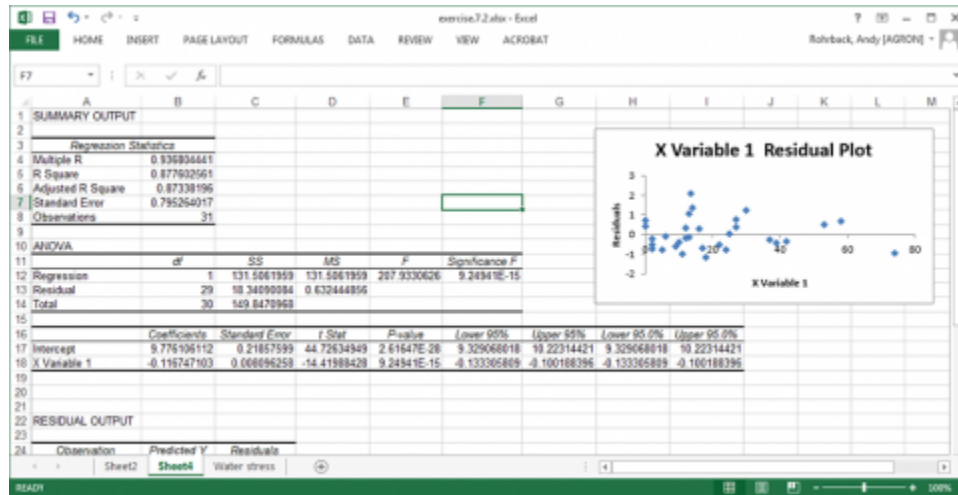


Fig. 27

Ex. 3: Calculating a Regression Line and Testing the Slope (2)

A great deal of information is available from the Summary Output and Analysis of Variance.

1. R (correlation between yield and water stress)

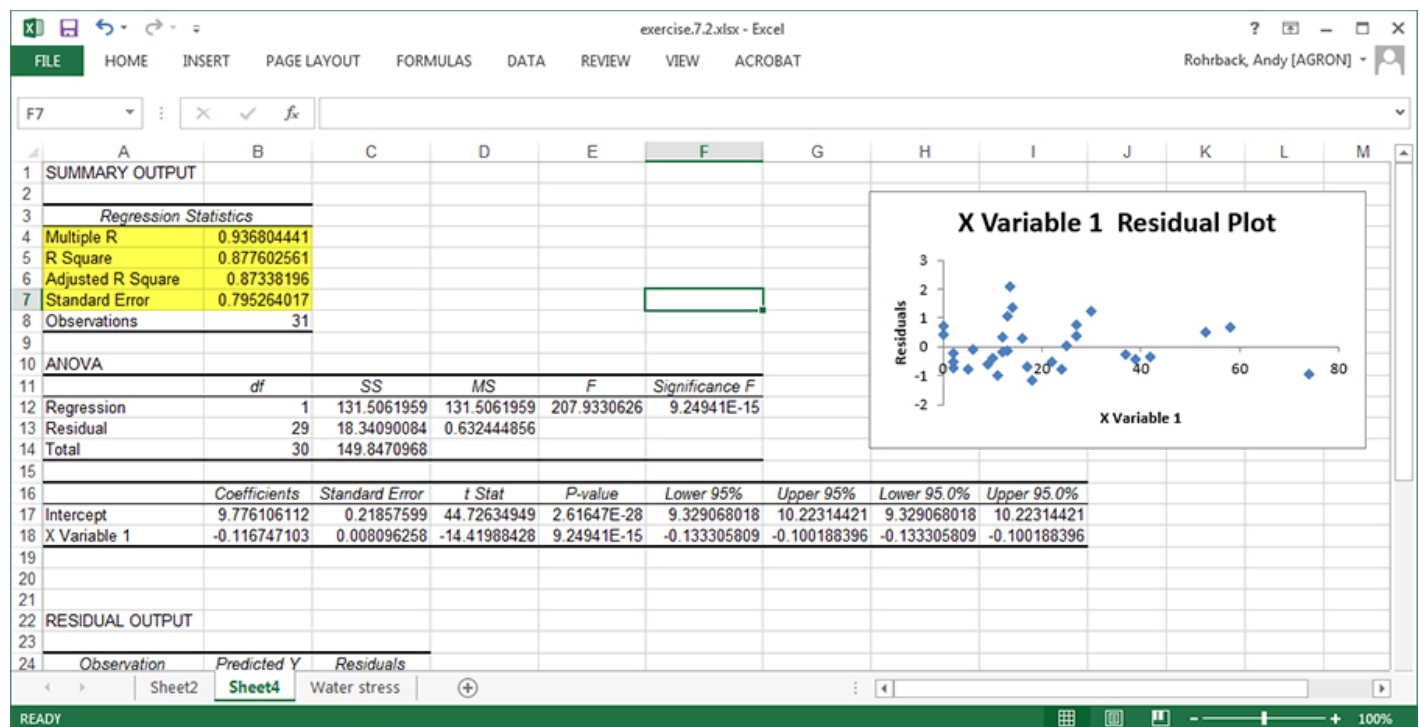
2. R-Squared (R^2)

3. Adjusted R Square

1-(Residual MS/Total MS)

A better measure for "goodness of fit" in multiple regression and comparing regression lines with different numbers of replication than is R-squared.

4. Standard Error (square root of the residual mean square)



Again notice the regression equation (Linear Fit). $E(Y) = 9.78 - 0.117x$ Is the slope statistically significantly different from zero?

The ANOVA table, which has the F-test for slope based on the residual mean square (0.632), supplies the answer.

The Prob > F, which tests the null hypothesis of no linear regression relationship (i.e., slope = 0), implies to reject H_0 because the probability is < .0001.

There is another test for the significance of water stress, as a t-ratio for water stress in the table beneath the ANOVA.

The regression slope, estimated by -0.117 , is significantly different from zero.

Ex. 3: Calculating a Regression Line and Testing the Slope (2)

We can get some information on how well the line fits the data by examining the Residual plot.

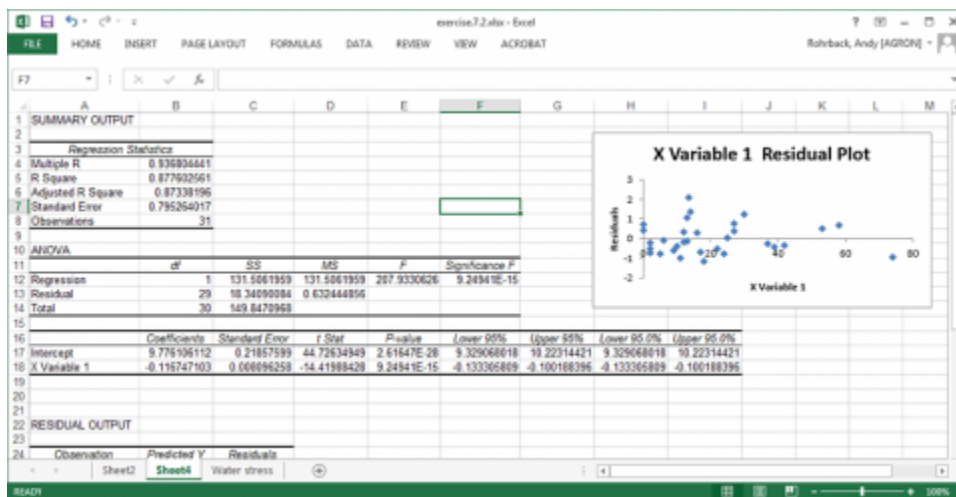


Fig. 28

The sum of the residuals should be 0 (or very small due to rounding errors of the computer and software). The plot of residuals displays how the actual Y values deviate from the regression-predicted Y values at each X. These should be scattered randomly along the X-axis. If there is any regularity to the residuals, the data may not be fit well by linear regression, or one of the assumptions of linear regression may have been violated. This is a small data set, so it would be easy to think that there is a pattern there. However, given the small size of the sample it does appear to be approximately randomly distributed around zero.

Study Question 2

Does the slope of the regression line in the rainfall and runoff study (Table 3) differ from 0?

☐ No

☐ Yes

☒ Check

Confidence Limits

Purpose

Confidence limits can be established for the regression slope.

When you have calculated a regression line and tested the slope for significance, you can be reasonably assured that the sampled regression line approximates the slope, b , of the model. Testing whether the slope describes a significant amount of the total variability is based on the F-test. From table 29 we note that the calculated F value of 12.95 is a really large value relative to an F distribution where the slope is equal to zero. Indeed, the probability of getting such a value or larger ($P > F$), given the null hypothesis, is 0.007. Thus the data do not support the null hypothesis of no linear response, although we might be wrong with such a statement about 7 times out of a thousand.

Another test related to the F-test is the t-test. The t-test can be used to test the significance of the regression line or more commonly it can be used to set error limits of the regression line. Typically, these are displayed as error bars encompassing some percentage of the data based on the estimated variance, s^2 . These limits come in three different types, error bars describing a confidence interval where the regression line occurs, error bars around the estimate of the average response, and bars encompassing an individual predicted value for a given X .

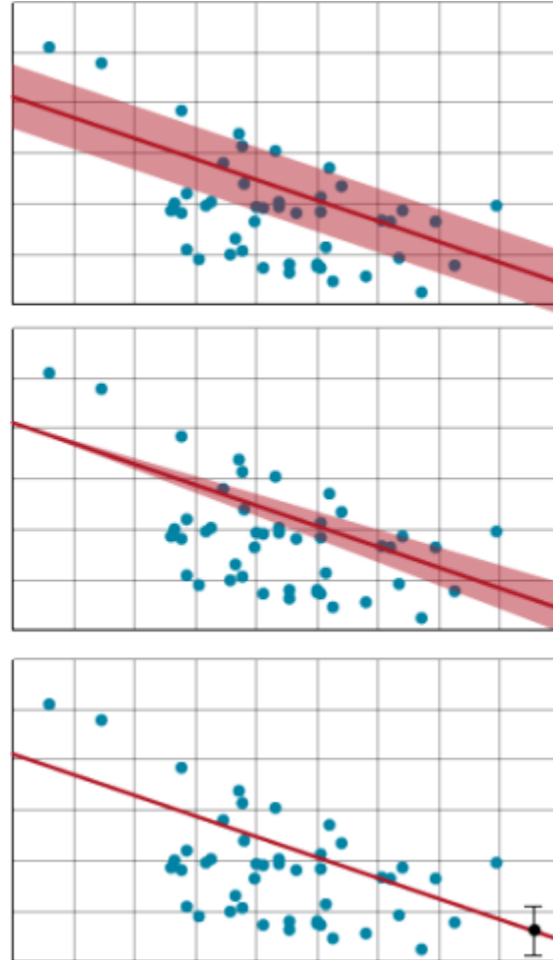


Fig. 29 Display methods for confidence limits on the regression line of a set of data and the predictions made from it.

Equation

Confidence limits are the lower and upper bounds of a confidence interval. In the context of regression line they can be used to test whether the slope of the regression line is different from 0. The null hypothesis is that the slope of the regression line is 0. The alternative hypothesis is that the slope of the line is not zero. If the confidence interval includes 0, the slope cannot be considered different from 0 at that level of significance. The error is merely that associated with the regression line. Confidence limits on a regression line are similar to those calculated for sample means.

$$CL = \hat{\beta} \pm t \times SE$$

Equation 10

where:

$\hat{\beta}$ = slope estimate

t = t - value for the given degrees of freedom and significance level

SE = standard error of $\hat{\beta} = \sqrt{\frac{\text{Residual Mean Square}}{S_{xx}}}$

Using t-Test

Restructuring this equation to solve for t allows us to use a t -test for testing whether the slope is different from 0. That test is equivalent to the F -test of regression in the ANOVA. The confidence limits bracket possible slopes of the regression line. A 95% confidence interval for the slope of a regression line means that this procedure will bracket the true regression slope 95% of the time.

$$CL = \hat{\beta} \pm t \times SE$$

Equation 10

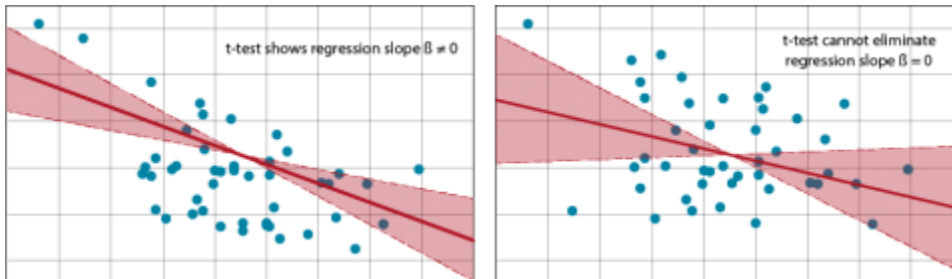


Fig. 30 When the standard error is large or the estimated slope of the regression line is small, a distribution will fail the t -test because a slope of 0 is possible under the given confidence level.

Limits on the estimates of a specific Y from the equation, correspondingly, will have a wider limit. The estimate of the mean \bar{Y} or predicted values include not only the variance of the regression line but also that of individual means or values at each X -value.

Try: Confidence Limits in the next screens

Ex. 4: Confidence Limits

Open the Excel water stress workbook you used earlier.

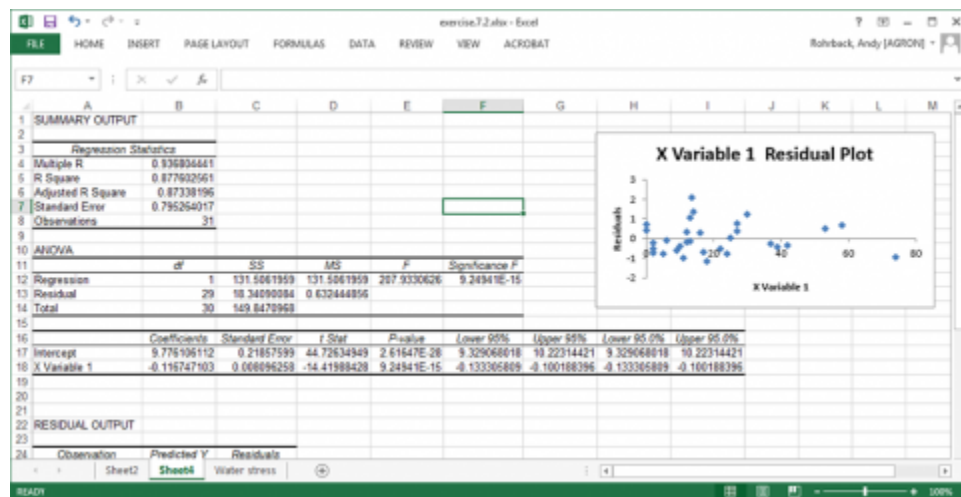


Fig. 28

Directly underneath the ANOVA table is a table with t-ratios and confidence intervals for the intercept and Water Stress coefficients.

A 95% confidence interval for β is between -0.133 and -0.100.

This is also easily computed from the formula $\hat{\beta} \pm t \times SE$, or $-0.117 \pm (2.045)(0.0081)$, where with alpha = 0.05 in two tails and 29 error df, the table t-value is 2.045 and the standard error for b is 0.008096.

Use a calculator to show that the confidence limits for β match those in the table.

Replicated Regression

Purpose

Regression in replicated data allows a goodness-of-fit test.

Agronomic experiments are usually replicated. In this situation, when the data are grouped, replication of Y values at X's will occur. Calculating a total regression includes the variability in Y replication at the X values.

Because of the replication at each X, we are able to separate the residual variation into two parts, that due to lack of fit of y-means from the regression line and a pure error. One of the assumptions of regression is that the regression line passes through the mean Y at each particular X-value. There may be some difference between the regression-predicted mean and the actual mean, called the lack of fit of the regression line. We wish to partition that error from the pure error in the data. The lack of fit SS is due to deviations from linearity.

The mean Y at each X could be used to create a regression line, but some information (on pure error) would be lost in the process.

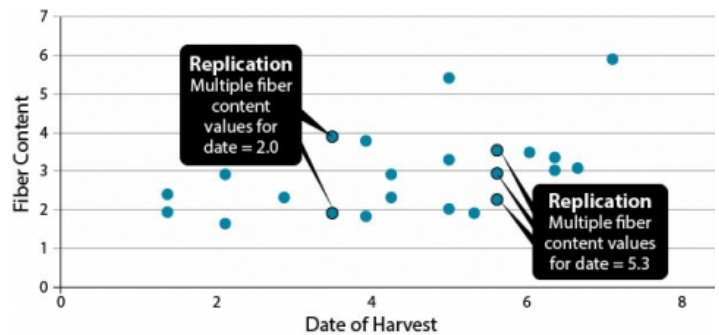


Fig. 31 This data on fiber content in corn kernels related to harvest date shows replication: some samples harvested on the same date have different fiber contents.

Example

An example compares the regression of percent of fiber content of corn by harvest date. Here you may have several data points for each date.

The initial ANOVA table provides the sum of squares and the test for the significance of the regression line. After the 1 df for regression 21 df remain. This variability can be partitioned into the two sources discussed, the lack of fit to the model and the pure error. The lack of fit variability comes from the difference between the actual means of the Y's at each X and the regression line predicted Y at each X. This value describes how much error is associated with the regression line. This value can be tested to determine if the regression lack of fit is different from 0. Error which is left over is termed the pure error.

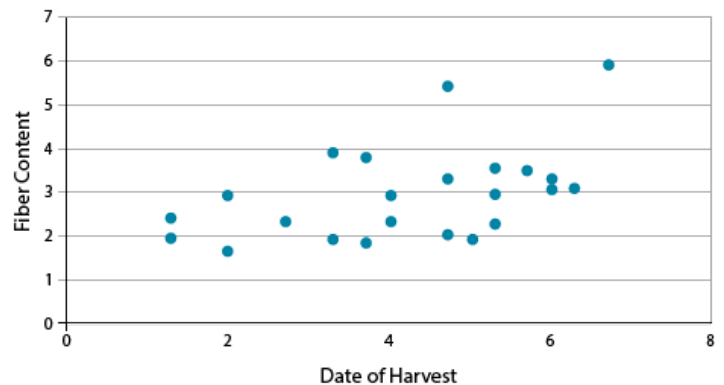


Fig. 32 Scatter plot showing fiber content of corn by harvest date

Table 5

Source	SS	Df	MS	F	P < F
Regression	6.32	1	6.32	6.26	<.05
Residual	21.19	21	1.01		

Error Calculation

Pure error is the deviation sum of squares of each individual Y from the mean Y at each X. The degrees of freedom are the sum of one less than the number of replicated Y's at each X. The pure error is calculated and subtracted from the residual to find the lack of fit.

$$\text{Total Error} = \sum_{i=1}^n \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2$$

Equation 11

where:

i = level of X

j = each replicated Y at a given X

Y_{ij} = each observation at a given level of the x (independent variable)

Y_i = mean for each level of the x variable

In effect, you are calculating a new sum of squares that estimates the total variance of observations around the mean for each value of x. This tells us how scattered were the data points we tried to fit with the regression line. The pure error degrees of freedom are calculated as:

$$\text{Pure Error df} = \sum (n_i - 1) \text{ at each X}$$

Equation 12

Example: ANOVA Table

These equations produce the ANOVA table (Table 6).

Table 6 Output from ANOVA of data

Source	SS	Df	MS	F	P < F
Regression	6.32	1	6.32	6.26	< .05
Regression	21.19	21	1.01		
Lack of fit	8.78	11	0.79	0.64	
Pure error	12.41	10	1.24		

The second F test compares the MS-lack of fit to the MS-pure error. This tests the linearity of the regression line. Since it is not significant, we assume the regression is linear and we do not have to try another model.

Ex. 5: ANOVA with Replicated Data

In this exercise we will calculate the ANOVA presented in the lesson for replicated measurements of fiber content over different harvest dates. This will require calculating the ANOVA and partitioning the degrees of freedom and sums of squares.

- Download the [QM-mod7-ex5data.xls](#) file and save it.
- Run the regression analysis covered in Exercise 2 on this data set, then go back to the worksheet with the original data set.
- The lack of fit test is not calculated automatically in the regression analysis, so we will have to do it step-by-step.

On the same page as the data set, add two columns. Label one Date and the other Mean at Date.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Date of Harvest	Fiber Content		Date	Mean at Date										
2	1.3	2.3													
3	1.3	1.8													
4	2.0	2.8													
5	2.0	1.5													
6	2.7	2.2													
7	3.3	3.8													
8	3.3	1.8													
9	3.7	3.7													
10	3.7	1.7													
11	4.0	2.8													
12	4.0	2.2													
13	4.7	3.4													
14	4.7	3.2													
15	4.7	1.9													
16	5.0	1.8													
17	5.3	3.5													
18	5.3	2.8													
19	5.3	2.1													
20	5.7	3.4													

Fig. 33

Ex. 5: ANOVA with Replicated Data (2)

Under date, copy each of the dates once.

Date of Harvest	Fiber Content	Date	Mean at Date
1.3	2.3	1.3	
1.3	1.8	2.0	
2.0	2.8	2.7	
2.0	1.5	3.3	
2.7	2.2	3.7	
3.3	3.8	4.0	
3.3	1.8	4.7	
3.7	3.7	5.0	
3.7	1.7	5.3	
4.0	2.8	5.7	
4.0	2.2	6.0	
4.7	3.4	6.3	
4.7	3.2	6.7	
4.7	1.9		
5.0	1.8		
5.3	3.5		
5.3	2.8		
5.3	2.1		
5.7	3.4		

Fig. 34

Under Mean at Date, calculate the average of the observations at the given date.

Date of Harvest	Fiber Content	Date	Mean at Date
1.3	2.3	1.3	2.05
1.3	1.8	2.0	2.13
2.0	2.8	2.7	2.20
2.0			
2.7			
3.3			
3.3			
3.7			
3.7			
4.0	3.0	3.7	3.10
4.0	2.2	5.0	3.00
4.7	3.4	5.3	5.90
4.7	3.2		
4.7	1.9		
5.0	1.8		
5.3	3.5		
5.3	2.8		
5.3	2.1		
5.7	3.4		

Fig. 35

Ex. 5: ANOVA with Replicated Data (3)

Now we will find the residuals associated with pure error using the means that were calculated.

1. Insert a new column next to the Fiber Content data.
2. Place the average for each date next to the associated date. Use this formula: =LOOKUP(A2,E\$2:E\$14, F\$2:F\$14)
3. Add another new column for the residuals.

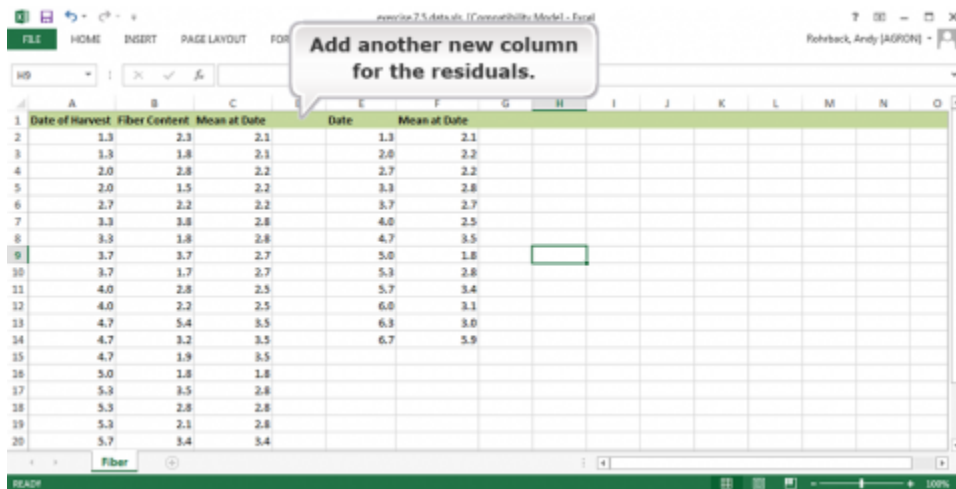


Fig. 36

4. The residuals can then be determined by subtracting the mean at each date from the observation at that date. =B2-C2
5. Add another column and insert the sums of squares (SS) associated with pure error by squaring each of the residuals. =POWER(D2,2)
6. Add another column and insert the sums of squares (SS) associated with pure error by squaring each of the residuals. =POWER(D2,2)
7. Then sum all the squares of residual pure error. =SUM(E2:E24)
8. The degrees of freedom for pure error are calculated by determining the number of observations at each date and subtracting one from the number of replications at each date. =IF(COUNTIF(A\$2:INDIRECT("A"&ROW()),A2)>1, "", COUNTIF(A\$2:A24,A2)-1)
9. These values are then summed. =SUM(F2:F24)
10. The mean squares for pure error are then calculated by dividing the pure error sums of squares by the degrees of freedom for pure error.
11. The lack of fit sums of squares and degrees of freedom can be found by subtracting the pure error sums of squares or degrees of freedom from the residual sums of squares.
12. The mean squares for lack of fit are found by dividing the lack of fit SS by the lack of fit DF.
13. Calculate the F-test for the lack of fit test by dividing the lack of fit MS by the pure error MS.
14. Finally, the p-value for the F-test can be found using the formula "=f.dist.rt(A,B,C)" where A is the F-statistic, B is the df for lack of fit, and C is the pure error DF.

Ex. 5: ANOVA with Replicated Data (4)

Fill in the values in the ANOVA table under the regression analysis.

Insert two rows for the “Lack of Fit” and “Pure Error” statistics.

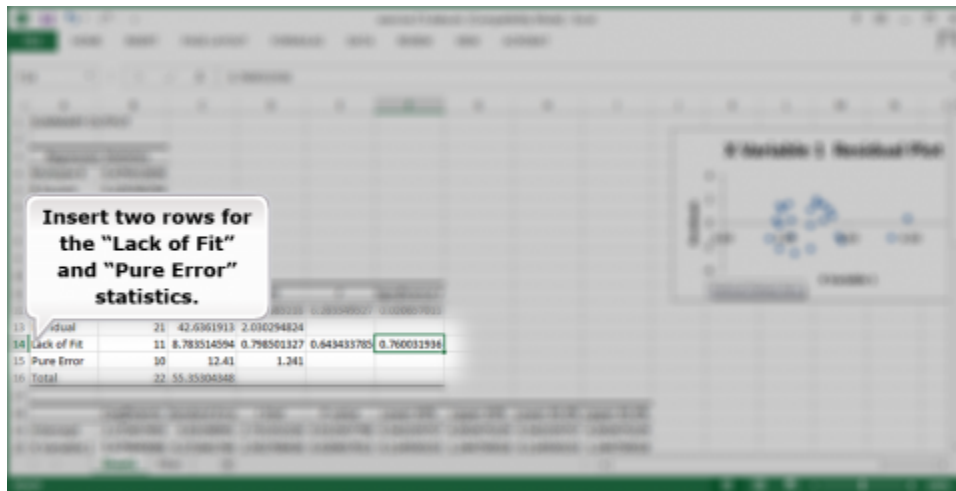


Fig. 37

Add the values you calculated to the ANOVA table under the regression analysis.

The completed analysis can be found in [qm-mod6-ex5solved.xlsx](#).

Summary

Correlation (r)

- Measures degree or strength of linear relationship
- Tells direction of linear relationship, positive implies x and y increase or decrease together; negative (y decreases as x increases or vice versa)
- Ranges between -1 and +1, with 0 being no linear relationship
- Scatter plot is important to help interpret

Linear regression

- Establishes a mathematical relationship between two variables
- Prediction equation is $Y = a + bx$
- Parameter estimates are intercept (a) and slope (b)
- Line of best fit minimizes squares of vertical deviations from line

ANOVA for Regression

- Has sources of variation for regression with 1 df and error with $(n - 2)$ df
- r^2 = (square of correlation) is proportion of variation attributed to linear regression
- Tests statistical significance of linear regression

Confidence Limits

- Can be established for regression slope
- $CL = b \pm t_{sb}$
- Can also be computed for mean of y-values or individual y given x.

Regression with Replicated Data

- Allows a goodness-of-fit test of the model

Reflection

The **Module Reflection** appears as the last "task" in each module. The purpose of the Reflection is to enhance your learning and information retention. The questions are designed to help you reflect on the module and obtain instructor feedback on your learning. Submit your answers to the following questions to your instructor.

1. In your own words, write a short summary (< 150 words) for this module.
2. What is the most valuable concept that you learned from the module? Why is this concept valuable to you?
3. What concepts in the module are still unclear/the least clear to you?

Acknowledgements

This module was developed as part of the Bill & Melinda Gates Foundation Contract No. 24576 for Plant Breeding E-Learning in Africa.

Quantitative Methods Linear Correlation, Regression and Prediction Author: Ron Mowers, Dennis Todey, Kendra Meade, William Beavis, and Laura Merrick (ISU)

Multimedia Developers: Gretchen Anderson, Todd Hartnell, and Andy Rohrback (ISU)

How to cite this module: Mowers, R., D. Todey, K. Meade, W. Beavis, and L. Merrick. 2016. Linear Correlation, Regression and Prediction. *In* Quantitative Methods, interactive e-learning courseware. Plant Breeding E-Learning in Africa. Retrieved from <https://pbea.agron.iastate.edu>.

Source URL: <https://pbea.agron.iastate.edu/course-materials/quantitative-methods/linear-correlation-regression-and-prediction-0?cover=1>