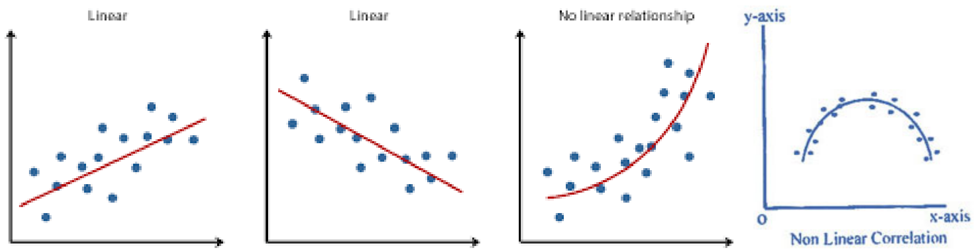
Chapter 4 –Describing the Relation Between Two Variables

Regression and Correlation

Section 4.1, 4.2, and 4.3: In this section we show how the **least square method** can be used to develop a linear equation, $Y = aX + b$, relating two variables, Y and X. The variable that is being predicted is called the **Dependent(Y)** or **Response** variable and the variable that is being used to predict the value of the dependent variable is called the **Independent(X)** or **Explanatory** variable. We generally use Y to denote the dependent variable and use X to denote the independent variable.(Common types of relationships-see images below)



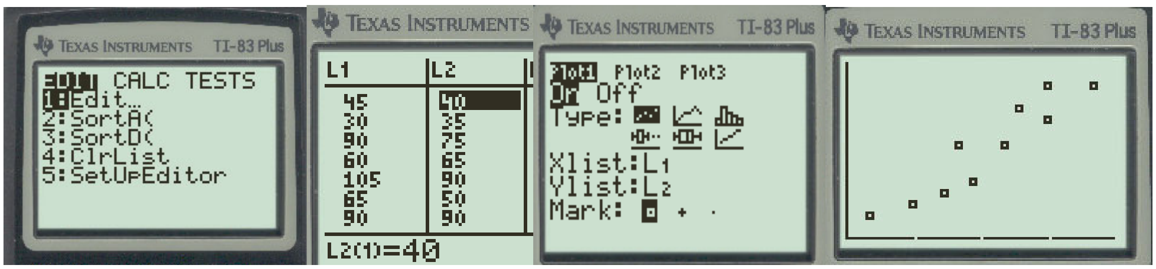
Example 1: The instructor in a freshman computer science course is interested in the relationship between the time using the computer system (X) and the final exam score (Y). Data collected for a sample of 10 students who took the course last semester are presented below. Draw the scatter plot.

In regression, the regular plot of Y vs X is called “scatter Plot”.

X= Hours Using Computer System	Y= Final Exam Score
45	40
30	35.
90	75
60	65
105	90
65	50
90	90
80	80
55	45
75	65

Use the TI 83/84 to do a Scatter Plot using the data in the above table.
First enter the values in the calculator

Using TI-83/84: stat, 1:Edit and enter X in L1 and Y in L2.
After entering the data, do 2nd and Y to access the statplot . Choose 1 for statplot 1 and turn it on, Type:1st one, Xlist:L1, Ylist:L2, Mark: Choose anyone of the three symbols, ZOOM #9. Now you should be able to see the scater plot. For "Mark", I chose the square symbol to represent the points on the scatter plot.(see pictures below)



Looking at the scatter plot, the relationship between the two variables can be approximated by a straight line. Clearly, there are many straight lines that could represent the relationship between x and y. The question is, which of the straight lines that could be drawn "best" represents the relationship?

The **least square method** is a procedure that is used to find the line that provides the best approximation for the relationship between X and Y. We refer to this equation of the line developed using the least square method as the **regression line**. Use the TI 83/84 calculator to find the regression line. First let us make sure your calculator is setup correctly. Perform the following sequence of commands: 2nd and 0 to access the catalog, scroll down until you see "DiagnosticOn". Put your cursor next to DiagnosticOn and hit enter twice. (see pictures below)

Regression Line: $\hat{Y} = aX + b$ **where**

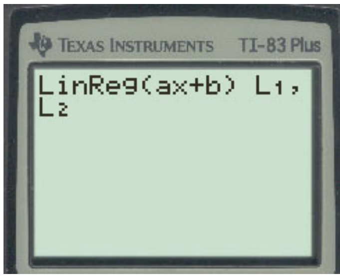
a = slope of the line

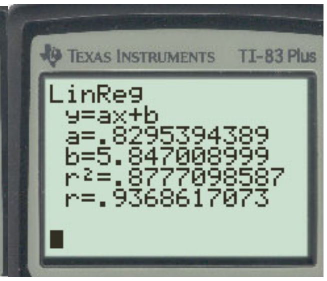
b = y-intercept of the line

$\hat{Y}$ = Predicted value of Y

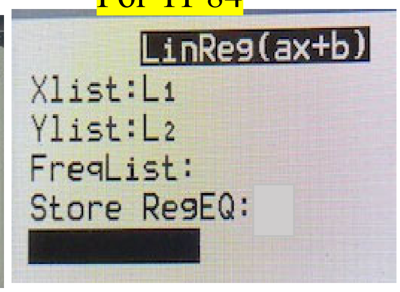
Now use the calculator to find the regression line.

Using TI-83/84: stat, CALC, 4:LinReg(ax+b), 2nd and 1 for L1, 2nd and 2 for L2, and enter. (see pictures below)





For TI-84



Please note: the slope: a= 0.8295 and y-intercept: b= 5.847
linear correlation: r = 0.93686 and R-Squared: $100(r^2)\%=$ $100(0.8777)\%=87.77\%$

Example 2: Find the regression line for the data given in **Example 1**. Use the regression line to estimate y when x = 80. (Use TI-83/84)

a = 0.8295 and b = 5.847 (From TI-83/84)

Thus the regression line is : $\hat{Y} = (0.8295)X + 5.847$

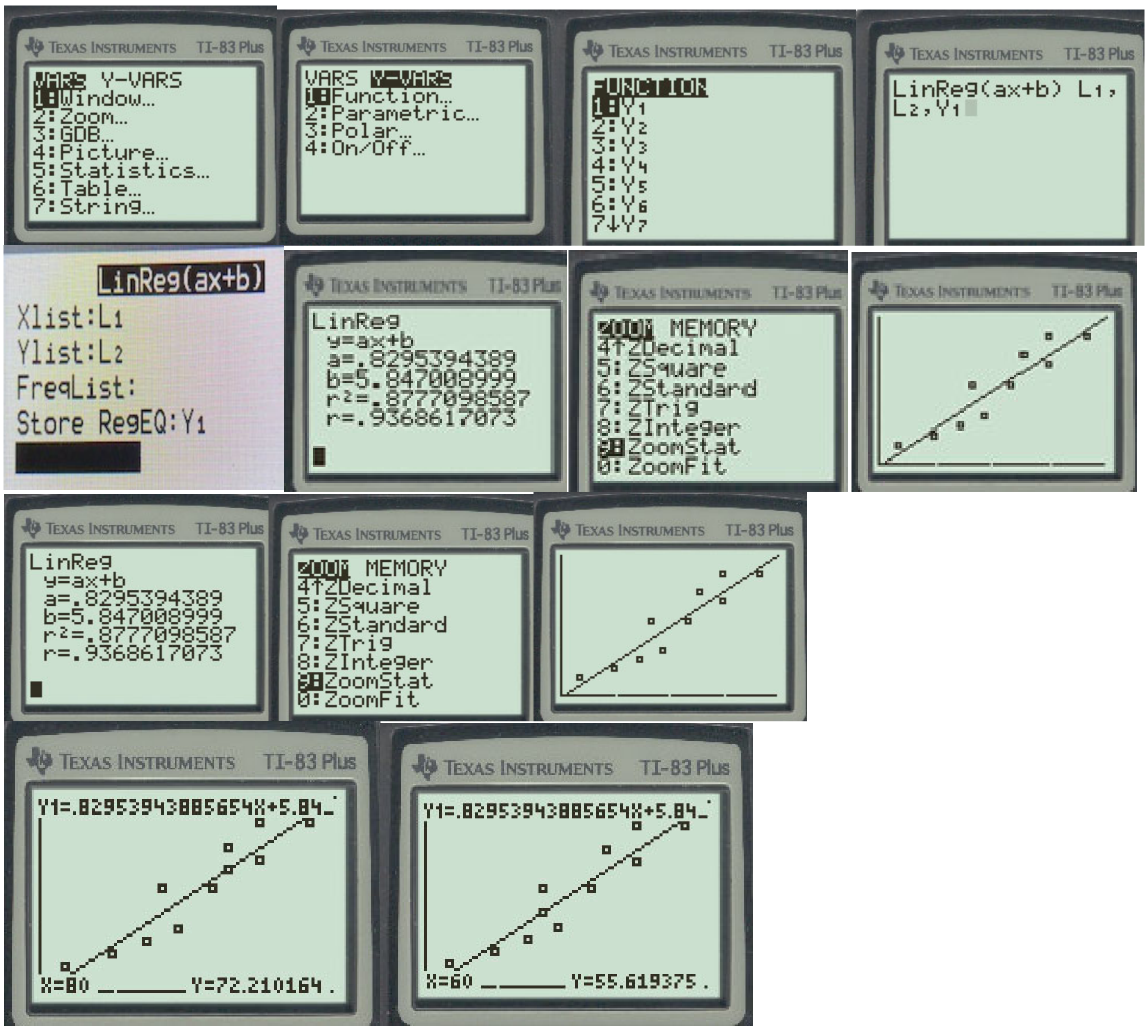
The linear relationship between X and Y is r=0.93686 or 93.686%.

Question: Are the predictions of Y using X good(reliable)? Yes, if $R-Squared = 100(r^2)$ is 65% or more, then the prediction is acceptable.

Prediction: When x = 80, $\hat{Y} = 0.8295(80) + 5.847 = 72.21$ (or use TI-83/84). Is this a good prediction and why? Yes, since R-Squared=87.77% is greater than 65% .

Using the calculator to make a prediction. First plot the scatter plot and regression line the same time.

Using TI-83/84: stat, CALC, 4:LinReg(ax+b), 2nd and 1 for L1, 2nd and 2 for L2, VARS, Y-VARS, 1: Function, 1: Y1, and Enter, ZOOM #9, TRACE, move cursor UP(arrow up) on the regression line, Type 80, Enter. This is a prediction for X=80, Y=72.21. Now, Type 60, and Enter. This is another prediction for X=60, Y=55.619. (see pictures below)



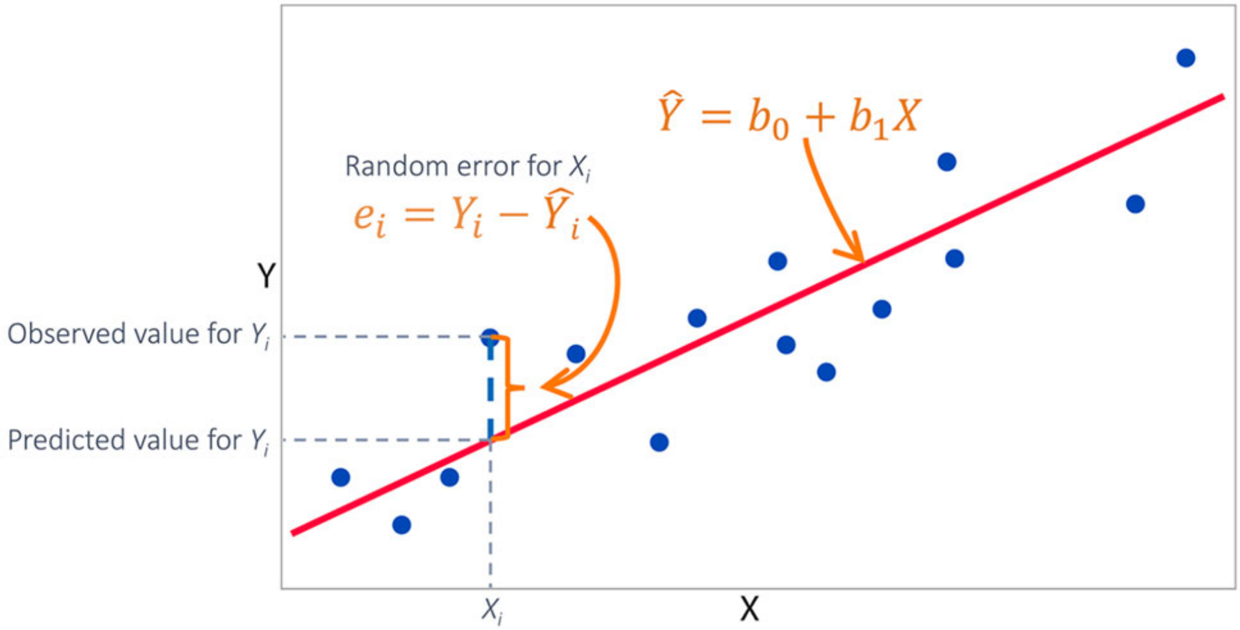
Least Square Method (Finding slope-a and y-intercept-b by hand)

The values of b and a can be computed using the following equations.

$$a = \frac{\sum xy - n\bar{X}\bar{Y}}{\sum x^2 - n(\bar{X})^2} \quad \text{and} \quad b = \bar{Y} - a\bar{X}$$

where $\bar{X} = \frac{\sum x}{n}$, $\bar{Y} = \frac{\sum y}{n}$, and n = total number of observations.

Residual is the difference between the actual value of Y and the predicted value $\hat{Y}$, $Y-\hat{Y}$**.** The Residual is denoted by e**,** $e_i = Y_i - \hat{Y}_i$ **.** If the residual is negative, Y is below $\hat{Y}$ (Y is overestimated by $\hat{Y}$). If the residual is positive, Y is above $\hat{Y}$ (Y is underestimated by $\hat{Y}$). Please note that the residual is the error of your estimate.



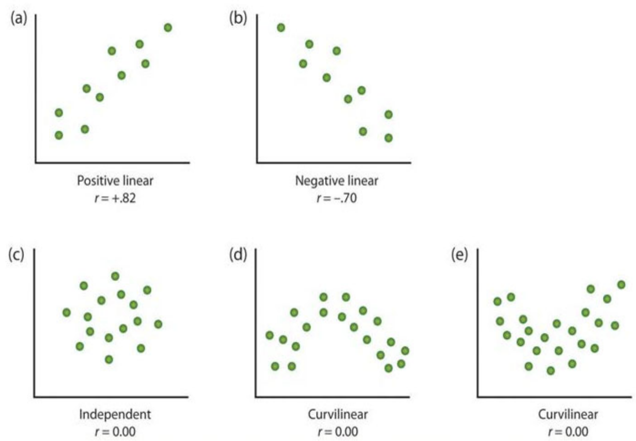
For X=80 the actual value for Y=80 from the data. In Example 2, the predicted value of Y, i.e. $\hat{Y}$=72.21. The **Residual(error) is** $e = Y - \hat{Y}$ **= 80-72.21 = 7.79.** The error is **7. 79.** The value of Y at X=80 is underestimated by 7.79 points.

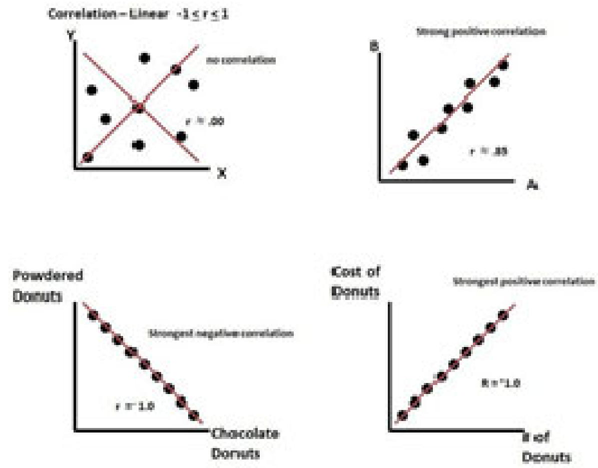
Linear Correlation(r)
The linear correlation coefficient, **r**, measures the strength of the linear association between two quantitative variables. You can get **r** from TI-83/84.

Rules for interpreting r:

a. The value of r always falls between –1 and 1. A positive value of r indicates positive correlation and a negative value of indicates negative correlation.
b. The closer r is to 1, the stronger the positive correlation and the closer r is to –1, the stronger the negative correlation. Values of r closer to zero indicate no linear association.
c. The larger the absolute value of r, the stronger the relationship between the two variables.
d. r measures only the strength of linear relationship between two variables.

Below are some images noting the degree of linear relationship(r)





The Coefficient of Determination ($R - Squared = 100(r^2)\%$ **)**

We define $R - Squared = 100(r^2)\%$ to be the coefficient of determination. You can get it from the TI-83/84.

Note: The coefficient of determination always lies between 0 and 1 and is a descriptive measure of the utility of the regression line for making prediction. Values of $R - Squared$ near to zero indicate that the regression equation is not very useful for making predictions, whereas values of r^2 near 1 or $R - Squared$ near 100% indicate that the regression equation is extremely useful for making predictions. If $R - Squared$ is 65% or more, then the prediction is acceptable.

Example 3: In example 1, are the predictions good?

From the TI-83, $R - Squared$ = **87.77%.**

Yes, using the regression line, $\widehat{Y} = (0.8295)X + 5.847$, the predictions are good.

Homework-Section 4.1, 4.2, and 4.3 Online - MyStatLab

Example: Real Life Application

Dr. XXXXX
I'm a VSU alumnus-class of '95. My sole proprietorship business (me) needs a solution to a math problem, and since my BS was in Psychology, I'm unqualified and was hoping you could help.

Much thanks in advance,
Mr. XXXXXXXXXXXX

Problem:
Below is a table. Left column is the length of an auger (it moves cement powder through a tube with a motorized "screw) . Right column is HP required to maintain a certain production "constant" at the respective length. I need to know the equation (if it exists) to obtain the HP given ANY length (eg. 12' or 28') . Length range is 10' - 40' . MS Excel showed me that the graph is a mild "S" shape, so I knew I was in trouble, given that anything beyond linear relationships is a nightmare for me.

Data	*Length*	*HP*
1	10	3.08
2	15	4.14
3	20	4.91
4	25	5.76
5	30	6.45
6	35	6.98
7	40	7.67