

## TOPIC: SIMPLE LINEAR REGRESSION

Using regression analysis, we can derive an equation by which the dependent variable (Y) is expressed (and estimated) in terms of its relationship with the independent variable (X).

In *simple* regression, there is only one independent variable (X) and one dependent variable (Y). The dependent variable is the one we are trying to predict.

In *multiple* regression, there are several independent variables ( $X_1, X_2, \dots$ ), and still only one dependent variable, Y. We are trying to use the X variables to predict the Y variable.

For example,

X (hours)	Y (Grade on quiz)
1	40
2	50
3	60
4	70
5	80

If you want to plot this line, what would it look like?

If  $X=6$ , then  $Y=?$

Note that for this straight line,

As X changes by 1, Y changes by 10

That's the slope  $b_1 = \frac{\Delta Y}{\Delta X} = 10$ .

$b_0$  is the Y-intercept, or the value of Y when  $X=0$ .

$$b_0 = 30$$

The following equation is the plot of the above data:

$$\hat{Y} = 30 + 10X$$

Note that we have a perfect relationship between X and Y and all the points are on the line ( $r=1, R^2=100\%$ ).

In general,

$$\hat{Y}_i = b_0 + b_1 X_i$$

This is the simple linear regression equation.

Why do we need regression in addition to correlation?

1- to predict a Y for a new value of X

2- to answer questions regarding the slope. E.g., for an additional amount of shelf space (X), what effect will there be on sales (Y). Example: if we raise prices by X%, will it cause sales to drop? This measures elasticity.

3- it makes the scatter plot a better display (graph) of the data if we can plot a line through it. It presents much more information on the diagram.

In correlation, on the other hand, we just want to know if two variables are *related*. This is used a lot in social science research.

The regression equation  $\hat{Y}_i = b_0 + b_1X_i$  is a sample estimator of the ‘true’ population regression equation:

$$Y_i = \beta_0 + \beta_1X_i + \varepsilon_i$$

where,

$\beta_0$  = true Y intercept for the population

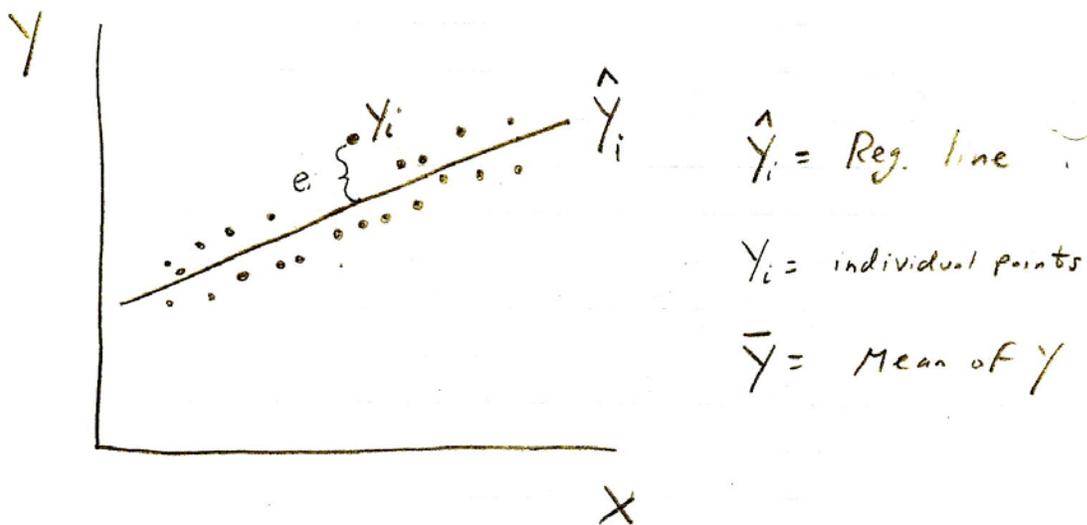
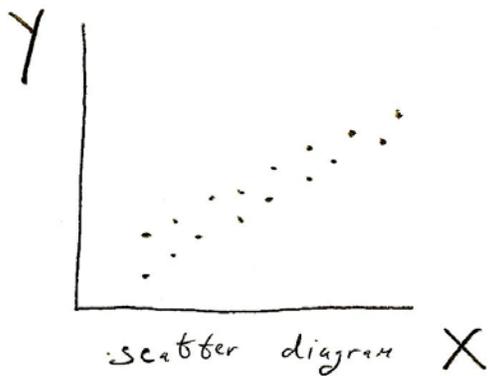
$\beta_1$  = true slope for the population

$\varepsilon_i$  = random error in Y for observation i

Our *estimator* of the above true population regression model, using the sample data, is:

$$\hat{Y}_i = b_0 + b_1X_i$$

There is a true regression line for the population. The  $b_0$  and  $b_1$  coefficients are estimates of the population coefficients,  $\beta_0$  and  $\beta_1$ .



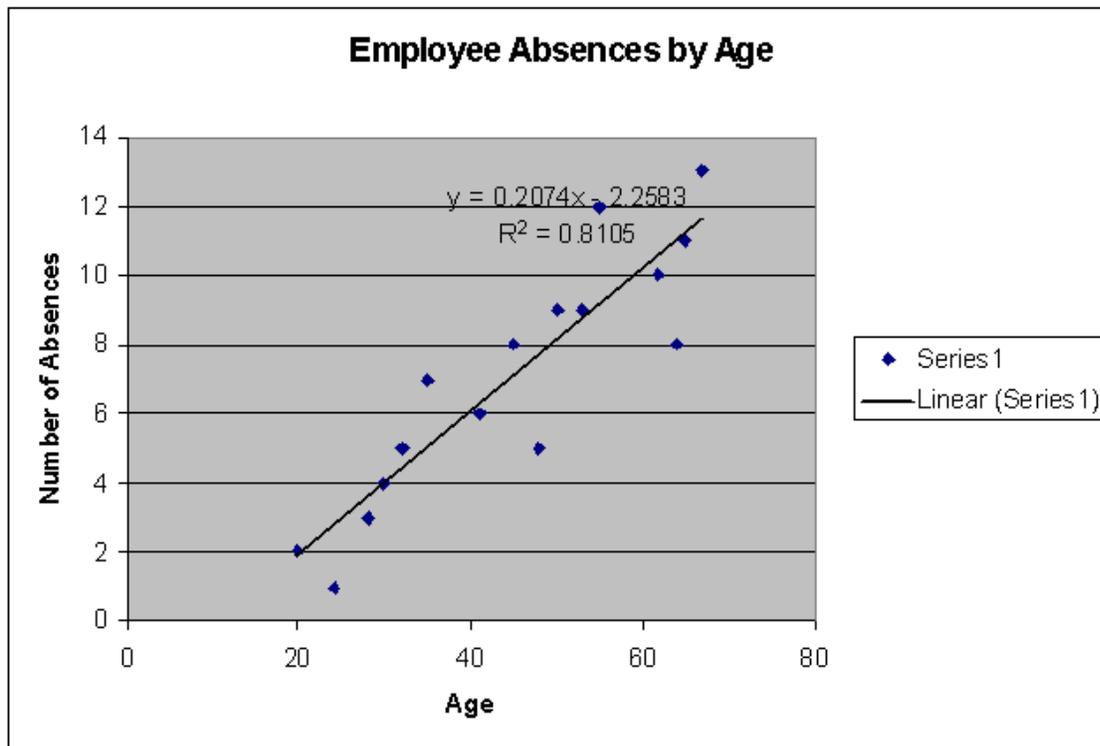
In regression, the levels of  $X$  are fixed.  $Y$  is a random variable.

The deviations of the individual observations (the points) from the regression line,  $(Y_i - \hat{Y}_i)$ , the *residuals*, are denoted by  $e_i$  where  $e_i = (Y_i - \hat{Y}_i)$ . Some deviations are positive (the points are above the line); some are negative (the points are below the line). If a point is on the line, its deviation = 0. Note that the  $\sum e_i = 0$ .

Mathematically, the regression line minimizes  $\sum e_i^2$  (this is SSE)  
 $= \sum (Y_i - \hat{Y}_i)^2 = \sum [Y_i - (b_0 + b_1 X_i)]^2$

-----  
 Taking partial derivatives, we get the “normal equations” that are used to solve for  $b_0$  and  $b_1$ .  
 -----

This is why the regression line is called the *least squares* line. It is the line that minimizes the sum of squared residuals. In the example below (employee absences by age), we can see the dependent variable (this is the data you entered in the computer) in blue and the regression line as a black straight line. Most of the points are either above the line or below the line. Only about 5 points are actually on the line or touching it.



## Steps in Regression:

1- For  $X_i$  (*independent variable*) and  $Y_i$  (*dependent variable*),

Calculate:

$$\begin{aligned} &\Sigma Y_i \\ &\Sigma X_i \\ &\Sigma X_i Y_i \\ &\Sigma X_i^2 \\ &\Sigma Y_i^2 \end{aligned}$$

2- Calculate the *correlation coefficient*,  $r$ :

$$r = \frac{n\Sigma X_i Y_i - (\Sigma X_i)(\Sigma Y_i)}{\sqrt{[n\Sigma X_i^2 - (\Sigma X_i)^2][n\Sigma Y_i^2 - (\Sigma Y_i)^2]}}$$

$$-1 \leq r \leq 1$$

[This can be tested for significance.  $H_0: \rho=0$ . If the correlation is not significant, then X and Y are not related. You really should not be doing this regression!]

3- Calculate the *coefficient of determination*:  $r^2 = (r)^2$

$$0 \leq r^2 \leq 1$$

This is the proportion of the variation in the dependent variable ( $Y_i$ ) explained by the independent variable ( $X_i$ )

4- Calculate the *regression coefficient*  $b_1$  (the slope):

$$b_1 = \frac{n\Sigma X_i Y_i - (\Sigma X_i)(\Sigma Y_i)}{n\Sigma X_i^2 - (\Sigma X_i)^2}$$

Note that you have already calculated the numerator and the denominator for parts of  $r$ . Other than a single division operation, no new calculations are required. BTW,  $r$  and  $b_1$  are related. If a correlation is negative, the slope term must be negative; a positive slope means a positive correlation.

5- Calculate the regression coefficient  $b_0$  (the Y-intercept, or constant):

$$b_0 = \bar{Y} - b_1 \bar{X}$$

The Y-intercept ( $b_0$ ) is the predicted value of Y when  $X = 0$ .

6- The regression equation (a straight line) is:

$$\hat{Y}_i = b_0 + b_1 X_i$$

7- [OPTIONAL] Then we can test the regression for statistical significance.

There are 3 ways to do this in simple regression:

(a) t-test for correlation:

$$H_0: \rho=0$$

$$H_1: \rho \neq 0$$

$$t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

(b) t-test for slope term

$$H_0: \beta_1=0$$

$$H_1: \beta_1 \neq 0$$

(c) F-test – we can do it in MS Excel

$$F = \frac{MSE_{\text{Explained}}}{MS_{\text{Unexplained}}} \quad F = \frac{MS_{\text{Regression}}}{MS_{\text{Residual}}}$$

where numerator is Mean Square (variation) Explained by the regression equation, and the denominator is Mean Square (variation) unexplained by the regression.

**EXAMPLE:**

$n = 5$  pairs of X,Y observations

Independent variable (X) is amount of water (in gallons) used on crop; Dependent variable (Y) is yield (bushels of tomatoes).

$Y_i$	$X_i$	$X_i Y_i$	$X_i^2$	$Y_i^2$
2	1	2	1	4
5	2	10	4	25
8	3	24	9	64
10	4	40	16	100
15	5	75	25	225
40	15	151	55	418

Step 1-

$$\Sigma Y_i = 40$$

$$\Sigma X_i = 15$$

$$\Sigma X_i Y_i = 151$$

$$\Sigma X_i^2 = 55$$

$$\Sigma Y_i^2 = 418$$

$$\text{Step 2- } r = \frac{(5)(151) - (15)(40)}{\sqrt{((5)(55) - (15)^2)((5)(418) - (40)^2)}} = \frac{155}{\sqrt{(50)(490)}} = .9903$$

$$\text{Step 3- } r^2 = (.9903)^2 = 98.06\%$$

Step 4-  $b_1 = \frac{155}{50} = 3.1$  The slope is positive. There is a positive relationship between water and crop yield.

$$\text{Step 5- } b_0 = \left(\frac{40}{5}\right) - 3.1\left(\frac{15}{5}\right) = -1.3$$

Step 6- Thus,  $\hat{Y}_i = -1.3 + 3.1X_i$

$\hat{Y}_i$	=	-1.3	+	3.1	$X_i$
# bushels of tomatoes		Does no water result in a negative yield?		Every gallon adds 3.1 bushels of tomatoes	# gallons of water

$Y_i$	$X_i$	$\hat{Y}_i$	$e_i$	$e_i^2$
2	1	1.8	.2	.04
5	2	4.9	.1	.01
8	3	8.0	0	0
10	4	11.1	-1.1	1.21
15	5	14.2	.8	.64
			$\Sigma e_i = 0$	$\Sigma e_i^2 = 1.90$

$\Sigma e_i^2 = 1.90$ . This is a minimum, since regression minimizes  $\Sigma e_i^2$  (SSE)

Now we can answer a question like: How many bushels of tomatoes can we expect if we use 3.5 gallons of water?  $-1.3 + 3.1(3.5) = 9.55$  bushels.

Notice the danger of predicting outside the range of X. The more water, the greater the yield? No. Too much water can ruin the crop.

[See the class handout “Simple Regression Using MS Excel” on the Virtual Handouts page.]

	Amt of Water	Tomato Yield					
	1	2					
	2	5					
	3	8					
	4	10					
	5	15					
SUMMARY OUTPUT							
<b>Regression Statistics</b>							
Multiple R	0.990258676						
R Square	0.980612245						
Adjusted R Square	0.97414966						
Standard Error	0.795822426						
Observations	5						
<b>ANOVA</b>							
	df	SS	MS	F	Significance F		
Regression	1	96.1	96.1	151.7368421	0.001152458		
Residual	3	1.9	0.633333333				
Total	4	98					
<b>Coefficients</b>							
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	
Intercept	$b_0$	-1.3	0.834665602	-1.557509974	0.2172241	-3.956280952	1.356280952
X Variable 1	$b_1$	3.1	0.251661148	12.31815092	0.001152458	2.299101159	3.900898841
RESIDUAL OUTPUT							
	Observation	Predicted Y	Residuals				
	1	1.8	0.2				
	2	4.9	0.1				
	3	8	0				
	4	11.1	-1.1				
	5	14.2	0.8				

$n = 5$  observations.  
Normally, one would not do regression with only 5 points. Better if  $n \geq 10$ .

$\leftarrow r$   
 $\leftarrow r^2$

$1.9 = SS_{Residuals}$   
[This is what the reg. line is minimized]

$b_0 = -1.3$   
 $b_1 = 3.1$

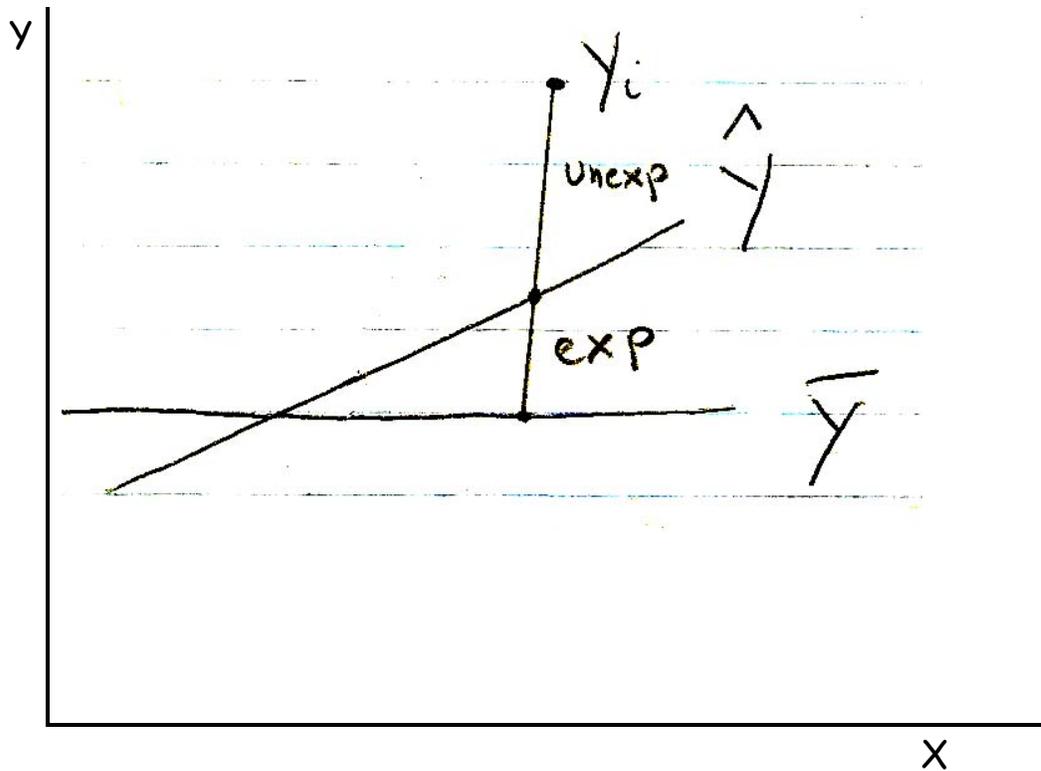
Test of  $b_1$  for sig.

$$\hat{Y} = -1.3 + 3.1X$$

$r = ?$        $r^2 = ?$        $\hat{Y}_i = ?$

Is this regression significant?  
 $H_0$ : No regression  
 $H_1$ : Yes regression  
 F-statistic = 151.7368  
 $P(\text{sample evidence} \mid H_0 \text{ is true}) = .00115$

## Measures of Variation in Regression



If we did not have a significant regression (i.e.,  $X$  does not predict  $Y$ ) we would use  $\hat{Y}_i = \bar{Y}$  as our regression equation.

$$(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

$$\underbrace{\sum (Y_i - \bar{Y})^2}_{\substack{\text{Total} \\ \text{Variation} \\ \text{in } Y}} = \underbrace{\sum (\hat{Y}_i - \bar{Y})^2}_{\substack{\text{Explained} \\ \text{Variation}}} + \underbrace{\sum (Y_i - \hat{Y}_i)^2}_{\substack{\text{Unexplained} \\ \text{Variation}}}$$

$$\text{Total Variation } \Sigma(Y_i - \bar{Y})^2 = \Sigma Y_i^2 - \frac{(\Sigma Y)^2}{n}$$

$$\text{Explained Variation } \Sigma(\hat{Y}_i - \bar{Y})^2 = b_0 \Sigma Y_i + b_1 \Sigma X_i Y_i - \frac{(\Sigma Y_i)^2}{n}$$

Unexplained Variation

$$\Sigma(Y_i - \hat{Y}_i)^2 = \Sigma Y_i^2 - b_0 \Sigma Y_i - b_1 \Sigma X_i Y_i$$

From previous problem:

$$\text{Total variation in } Y = 418 - (40)^2/5 = 98$$

$$\text{Explained variation (explained by } X) = -1.3(40) + 3.1(151) - (40)^2/5 = 96.10$$

$$\text{Unexplained variation} = 418 - -1.3(40) - 3.1(151) = 1.90$$

The coefficient of determination,  $r^2$ , is the proportion of  $Y$  explained by  $X$ .

$$r^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{96.10}{98} = .98$$

In other words, 98% of the total variation in crop yield is explained by the linear relationship of yield with amount of water used on the crop.

**EXERCISE:**

Let's examine the relationship between hours studied and grade on a quiz.  $n=7$  pairs of data. Highest grade on quiz is a 15.

$X \equiv$  hours studied;  $Y \equiv$  grade on quiz.

$X_i$	$Y_i$	$X_i Y_i$	$X_i^2$	$Y_i^2$
1	5	5	1	25
2	8	16	4	64
3	9	27	9	81
4	10	40	16	100
5	11	55	25	121
6	12	72	36	144
7	14	98	49	196
$\Sigma X = 28$	$\Sigma Y = 69$	$\Sigma XY = 313$	$\Sigma X^2 = 140$	$\Sigma Y^2 = 731$

Calculate the correlation coefficient,  $r$ :

$$r = \frac{7(313) - (28)(69)}{\sqrt{(7(140) - (28)^2)(7(731) - (69)^2)}} = \frac{259}{\sqrt{(196)(356)}} = \frac{259}{264.2} = 0.98$$

[slope will be positive]

Calculate the coefficient of determination,  $r^2$ :

$$r^2 = (0.98)^2 = .9604$$

Calculate the *regression coefficient*  $b_1$  (the slope):

$$b_1 = \frac{7(313) - (28)(69)}{7(140) - (28)^2} = \frac{259}{196} = 1.32$$

Calculate the regression coefficient  $b_0$  (the Y-intercept, or constant):

$$b_0 = \left(\frac{69}{7}\right) - 1.32\left(\frac{28}{7}\right) = 4.58$$

The regression equation:

$$\hat{Y}_i = 4.58 + 1.32X_i$$

Q: Explain the meaning of the regression coefficients.

Q: If someone studies 3.5 hours, what would we predict his/her quiz score to be?

## SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.980498039
R Square	0.961376404
Adjusted R Square	0.953651685
Standard Error	0.626783171
Observations	7

<i>ANOVA</i>					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	48.89285714	48.89285714	124.4545455	0.000100948
Residual	5	1.964285714	0.392857143		
Total	6	50.85714286			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>
Intercept	4.571428571	0.529728463	8.62975824	0.000344965	3.209718206
X Variable 1	1.321428571	0.118450885	11.15591975	0.000100948	1.016940877

## REGRESSION – USING EXCEL

Before using MS Excel, you should know the following:

df is degrees of freedom

SS is sum of squares

MS is mean square (SS divided by its degrees of freedom)

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	SSR	MSR	MSR/MSE	
Residual (Error)	n-2	SSE	MSE		
Total	n-1	SST			

Sum of Squares Total (SST) = Sum of Squares Regression (SSR) + Sum of Squares Error (SSE)

SSE is the sum of the squared residuals. Please note that some textbooks use the term Residuals and others use Error. They are the same thing and deal with the unexplained variation, i.e., the deviations. This is the number that is minimized by the least squares (regression) line.

$$SST = SSR + SSE$$

Total variation in Y = Explained Variation (Explained by the X-variable) + Unexplained Variation

SSR/SST is the proportion of the variation in the Y-variable explained by the X-variable. This is the R-Square,  $r^2$ , the coefficient of determination.

The F-ratio is the 
$$\frac{(\text{SS Regression} / \text{degrees of freedom})}{(\text{SS Residual} / \text{degrees of freedom})} = \frac{\text{MS Regression}}{\text{MS Residual}}$$

In simple regression, the degrees of freedom of the SS Regression is 1 (the number of independent variables). The number of degrees of freedom for the SS Residual is  $(n - 2)$ . Please note that SS Residual is the SSE.

If X is not related to Y, you should get an F-ratio of around 1. In fact, if the explained (regression) variation is 0, then the F-ratio is 0. F-ratios between 0 and 1 will not be statistically significant.

On the other hand, if all the points are on a line, then the unexplained variation (residual variation) is 0. This results in an F-ratio of infinity.

An F-value of, say, 30 means that the explained variation is 30 times greater than the unexplained variation. This is not likely to be chance and the F-value will be significant.

The following are some examples of simple regression using MS Excel.

Example 1: A researcher is interested in determining whether there is a relationship between years of education and income.

Education (X)	Income ('000s) (Y)
9	20
10	22
11	24
11	23
12	30
14	35
14	30
16	29
17	50
19	45
20	43
20	70

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.860811139
R Square	0.740995817
Adjusted R Square	0.715095399
Standard Error	7.816452413
Observations	12

<i>ANOVA</i>					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	1747.947383	1747.947383	28.60941509	0.000324168
Residual	10	610.9692833	61.09692833		
Total	11	2358.916667			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	-11.02047782	8.909954058	-1.236872575	0.244393811	-30.87309606	8.832140427
X Variable 1	3.197952218	0.597884757	5.348776972	0.000324168	1.865781732	4.530122704

This regression is very significant; the F-value is 28.61. If the X-variable explains very little of the Y-variable, you should get an F-value that is 1 or less. In this case, the explained variation (due to regression = explained by the X-variable) is 28.61 times greater than the unexplained (residual) variation. The probability of getting the sample evidence or even a stronger relationship if the X and Y are unrelated (Ho is that X does not predict Y) is .000324168. In other words, it is almost impossible to get this kind of data as a result of chance.

The regression equation is:  
 Income = -11.02 + 3.20 (years of education).

In theory, an individual with 0 years of education would make a negative income of \$11,020 (i.e., public assistance). Every year of education will increase income by \$3,200.

The correlation coefficient is .86 which is quite strong.

The coefficient of determination,  $r^2$ , is 74%. This indicates that the unexplained variation is 26%.

One way to calculate  $r^2$  is to take the ratio of the sum of squares regression/ sum of squares total.  
 $SSREG/SST = 1747.947383/ 2358.916667 = .741$

The Mean Square Error (or using Excel terminology, MS Residual) is 61.0969. The square root of this number 7.816 45 is the standard error of estimate and is used for confidence intervals.

The mean square (MS) is the sum of squares (SS) divided by its degrees of freedom.

Another way to test the regression for significance is to test the  $b_1$  term (slope term which shows the effect of X on Y). This is done via a t-test. The t-value is 5.348776972 and this is very, very significant. The probability of getting a  $b_1$  of this magnitude if  $H_0$  is true (the null hypothesis for this test is that  $B_1 = 0$ , i.e., the X variable has no effect on Y), or one indicating an even stronger relationship, is 0.000324168. Note that this is the same sig. level we got before for the F-test. Indeed, the two tests give exactly the same results. Testing the  $b_1$  term in simple regression is equivalent to testing the entire regression. After all, there is only one X variable in simple regression. In multiple regression we will see tests for the individual  $b_i$  terms and an F-test for the overall regression.

Prediction: According to the regression equation, how much income would you predict for an individual with 18 years of education?

Income =  $-11.02 + 3.20(18)$ . Answer = 46.58 in thousands which is \$46,580 Please note that there is sampling error so the answer has a margin of error. This is beyond the scope of this course so we will not learn it.

Example 2: A researcher is interested in knowing whether there is a relationship between the number of D or F grades a student gets and number of absences.

Examining records of 14 students: Number of absences in an academic year and number of D or F grades

#absences (X)	D or F grade (Y)
0	0
0	2
1	0
2	1
4	0
5	1
6	2
7	3
10	8
12	12
13	1
18	9
19	0
28	10

SUMMARY  
OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.609912681
R Square	0.371993478
Adjusted R Square	0.319659601
Standard Error	3.525520635
Observations	14

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	88.34845106	88.34845106	7.108081816	0.020558444
Residual	12	149.1515489	12.42929574		
Total	13	237.5			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	0.697778132	1.41156929	0.494327935	0.629999773	-2.377767094	3.773323358
X Variable 1	0.313848849	0.117718395	2.666098613	0.020558444	0.057362505	0.570335194

df is degrees of freedom; SS is sum of squares; MS is mean square (the MS is the SS divided by its degrees of freedom). ANOVA stands for analysis of variance. We are breaking down the total variation in Y (SS Total) into two parts: (1) the explained variation – the variation in Y explained by X. This is SS Regression and (2) the unexplained variation –the variation in Y that is not explained by X. The residuals indicate that there is unexplained variation. This variation is the SS Residual. Thus, SS Total = SS Regression + SS Residual.

The F-ratio is the 
$$\frac{(\text{SS Regression} / \text{degrees of freedom})}{(\text{SS Residual} / \text{degrees of freedom})} = \frac{\text{MS Regression}}{\text{MS Residual}}$$

In simple regression, the degrees of freedom of the Regression SS is 1 (the number of independent variables). The number of degrees of freedom for the Residual SS is  $(n - 2)$ .

This regression is significant; the F-value is 7.108. If the X-variable explains very little of the Y-variable, you should get an F-value that is 1 or less. In this case, the explained variation (due to regression = explained by the X-variable) is 7.108 times greater than the unexplained (residual) variation. The probability of getting the sample evidence (or data indicating an even stronger relationship between X and Y) if the X and Y are unrelated ( $H_0$  is that X does not predict Y, i.e., there is no regression) is .02056.

The regression equation is:

Number of DsFs = .698 + .314 (absences).

In theory, an individual with 0 absences would have .698 Ds and Fs for the academic year.

Every absence will increase the number of Ds and Fs by .314.

The correlation coefficient is .61 which is reasonably strong.

The coefficient of determination,  $r^2$ , is .372 or 37.2%.

One way to calculate  $r^2$  is to take the ratio of the sum of squares regression/ sum of squares total.  
 $SSREG/SST = 88.35/ 237.5 = .372$

The standard error is 3.525520635 . This is the square root of the Mean Square Residual (also known as the MSE or Mean Square Error) which is 12.42929574.

Prediction: According to the regression equation, how many Ds or Fs would you predict for an individual with 15 absences?

Number of DsFs = .698 + .314 (15). = 5.408

Example 3: A researcher is interested in determining whether there is a relationship between number of packs of cigarettes smoked per day and longevity (in years).

packs of cigarettes smoked (X)	Longevity (Y)
0	80
0	70
1	72
1	70
2	68
2	65
3	69
3	60
4	58
4	55

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.875178878
R Square	0.765938069
Adjusted R Square	0.736680328
Standard Error	3.802137557
Observations	10

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	378.45	378.45	26.17898833	0.000911066
Residual	8	115.65	14.45625		
Total	9	494.1			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	75.4	2.082516507	36.20619561	3.71058E-10	70.59770522	80.20229478
X Variable 1	-4.35	0.850183804	-5.11654066	0.000911066	-6.310528635	-2.389471365

This regression is significant; the F-value is 26.18. If the X-variable explains very little of the Y-variable, you should get an F-value that is 1 or less. In this case, the explained variation (due to regression = explained by the X-variable) is 26.18 times greater than the unexplained (residual) variation. The probability of getting the sample evidence (or data indicating an even stronger relationship) if the X and Y are unrelated (Ho is that X does not predict Y, i.e., the regression is not significant) is .000911066.

The regression equation is:  
 longevity = 75.4 – 4.35 (packs).

In theory, an individual who does not smoke (0 packs) absences would live to the age of 75.4 years. Every pack of cigarettes will reduce one’s lifetime by 4.35 years.

The correlation coefficient is  $-0.875$  which is quite strong. Note that MS Excel does not indicate that the correlation is negative. If the  $b_1$  term is negative, the correlation is negative.

The coefficient of determination,  $r^2$ , is  $.76594$  or  $76.6\%$ .

One way to calculate  $r^2$  is to take the ratio of the sum of squares regression/ sum of squares total.  
 $SSREG/SST = 378.45/494.10 = 76.6\%$ .

The MS Residual (also known as MSE or Mean Square Error) =  $14.45625$ . The square root of this, is the standard error of estimate =  $3.802$ .

Prediction: According to the regression equation, how long will one live who smokes 2.5 packs per day?

longevity =  $75.4 - 4.35(2.5) = 64.525$     Answer  $64.525$  years

Example 4: A researcher is interested in determining whether there is a relationship between the amount of vitamin C an individual takes and the number of colds.

mgs. of vitamin C (X)	#colds –year (Y)
985	7
112	1
830	0
900	3
900	1
170	1
230	5
50	2
420	2
280	2
200	3
200	4
80	5
50	7

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.100098669
R Square	0.010019744
Adjusted R Square	-0.072478611
Standard Error	2.314411441
Observations	14

<i>ANOVA</i>					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	0.650567634	0.650567634	0.121453859	0.733500842
Residual	12	64.27800379	5.356500316		
Total	13	64.92857143			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	3.315318136	0.934001032	3.549587232	0.00399968	1.280304741	5.350331532
X Variable 1	-0.000631488	0.001812004	-0.348502308	0.733500842	-0.004579506	0.00331653

This regression is not significant; the F-value is .12145. If the X-variable explains very little of the Y-variable, you should get an F-value that is 1 or less. The probability of getting the sample evidence (or sample evidence indicating a stronger relationship) if the X and Y are unrelated (Ho is that X does not predict Y, i.e., the regression is not significant) is .7335. We do not have any evidence to reject the null hypothesis.

The correlation coefficient is a very weak .10 and is not statistically significant. It may be 0 (in the population) and we are simply looking at sampling error.

If the regression is not significant, we do not look at the regression equation. There is nothing to look at as it all may reflect sampling error.

Example 5: A researcher is interested in determining whether there is a relationship between crime and the number of police.

n=12 districts	
X	Y
# police	crimes
4	49
6	42
8	38
9	31
10	24
12	24
12	28
13	23
15	21
20	19
26	12
28	14

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.886344142
R Square	0.785605937
Adjusted R Square	0.764166531
Standard Error	5.429309071
Observations	12

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	1080.142697	1080.1426	36.64308	0.0001230
Residual	10	294.7739699	29.477396		
Total	11	1374.916667			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	44.94145886	3.340608522	13.453075	9.903E-08	37.4981179	52.384799
X Variable 1	-1.314708628	0.217186842	-6.05335301	0.00012306	-1.79863115	-0.83078610

This regression is significant; the F-value is 36.64. If the X-variable explains very little of the Y-variable, you should get an F-value that is 1 or less. In this case, the explained variation (due to regression = explained by the X-variable) is 36.64 times greater than the unexplained (residual) variation. The probability of getting the sample evidence (or sample data indicating an even stronger relationship) if the X and Y are unrelated (Ho is that X does not predict Y, i.e., the regression is not significant) is .00012306.

The regression equation is:  
 Crimes = 44.94 – 1.31 (police officers).

In theory, a district with no police officers will have 44.94 crimes. Every police officer reduces crimes by 1.3147.

The correlation coefficient is  $-.886$  which is quite strong. Note that MS Excel does not indicate that the correlation is negative. If the  $b_1$  term is negative, the correlation is negative.

The coefficient of determination,  $r^2$ , is  $.7856$  or  $78.56\%$ .

The MS Residual (also known as MSE or Mean Square Error) =  $29.477$ . The square root of this, is the standard error of estimate =  $5.429$ .

Prediction: According to the regression equation, how many crimes will an area have that has 34 police officers

$$\text{Crimes} = 44.94 - 1.31 (34).$$

Answer .40 crimes

Example 6: A researcher is interested in determining whether there is a relationship between advertising and sales for her firm.

n=11 areas	
X	Y
advertising in \$thousands	Sales in millions
1	0
1	1
2	4
4	3
5	5
6	4
6	7
6	8
7	9
10	9
10	7

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.85091766
R Square	0.72406087
Adjusted R Square	0.69340096
Standard Error	1.71236726
Observations	11

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	69.2465488	69.24654882	23.61588908	0.000896307
Residual	9	26.3898148	2.932201646		
Total	10	95.6363636			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	0.75370370	1.04731113	0.719655963	0.49000172	-1.6154804	3.12288789
X Variable 1	0.83981481	0.17281497	4.859618203	0.00089630	0.44887987	1.23074975

This regression is significant; the F-value is 23.615. If the X-variable explains very little of the Y-variable, you should get an F-value that is 1 or less. In this case, the explained variation (due to regression = explained by the X-variable) is 23.615 times greater than the unexplained (residual) variation. The probability of getting the sample evidence (or sample data indicating an even stronger relationship) if the X and Y are unrelated (Ho is that X does not predict Y, i.e., the regression is not significant) is .000896307.

The regression equation is:  
 Sales (in millions) = .753704 + .8398 (advertising in thousands).

In theory, an area with no advertising will produce sales of \$753,704. Every \$one thousand of advertising increases sales by \$839,800.

The correlation coefficient is .85 which is quite strong.  
The coefficient of determination,  $r^2$ , is .7241 or 72.41%.

The MS Residual (also known as MSE or Mean Square Error) = 2.9322. The square root of this, is the standard error of estimate = 1.712.

Prediction: According to the regression equation, what would you predict sales to be in districts where the firm spends \$9,000 on advertising?

Sales (in millions) = .753704 + .8398 (9). Answer = 8.3119 or \$8,311,900

Example 7: A researcher is interested in constructing a linear trend line for sales of her firm. 1997 is coded as 0, 1998 is 1, 1999 is 2, 2000 is 3, ..., 2011 is 14. Sales are in millions.

TIME (X)	SALES (Y)
0	10
1	12
2	15
3	18
4	18
5	16
6	19
7	22
8	25
9	30
10	35
11	32
12	31
13	35
14	40

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.968105308
R Square	0.937227887
Adjusted R Square	0.932399263
Standard Error	2.440744647
Observations	15

<i>ANOVA</i>						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	1156.289286	1156.289286	194.0983352	3.42188E-09	
Residual	13	77.44404762	5.957234432			
Total	14	1233.733333				

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	9.641666667	1.199860403	8.035657014	2.12982E-06	7.049526359	12.23380697
X Variable 1	2.032142857	0.145862392	13.93191786	3.42188E-09	1.717026379	2.347259335

This (time series) regression is significant; the F-value is 194.098. If the X-variable explains very little of the Y-variable, you should get an F-value that is 1 or less. The probability of getting the sample evidence if the X and Y are unrelated (Ho is that X does not predict Y, i.e., the regression is not significant) is .00000000342.

The regression equation is:

$$\text{Sales (in millions)} = 9.641667 + 2.032143 (\text{Time}).$$

According to the trend line, sales increase by \$2,032,143 per year.

Prediction: What are expected sales for 2016? Note 2016 is 19.  
Sales (in millions) =  $9.641667 + 2.032143 (19)$ . Answer \$48,252,384

Example 8: A researcher is interested in determining whether there is a relationship between the high school average and GPA in Partytime College .

X HS Average	Y GPA
60	2.4
65	3.2
66	3.1
70	2.7
74	3.1
80	3.3
83	2.9
85	3.2
88	2.3
90	2.6
92	2.8
95	2.9
96	3.9
98	3.5
99	3.3

## SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.335962172
R Square	0.112870581
Adjusted R Square	0.044629857
Standard Error	0.412819375
Observations	15

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	0.281875465	0.281875	1.654006199	0.220849316
Residual	13	2.215457868	0.17042		
Total	14	2.497333333			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	2.107800193	0.712124579	2.959876	0.011059958	0.569348868	3.646252
X Variable 1	0.010945203	0.008510504	1.286082	0.220849316	-0.007440619	0.029331

This regression is not significant; the F-value is 1.654. If the X-variable explains very little of the Y-variable, you should get an F-value that is 1 or less. The probability of getting the sample evidence (or data indicating an even stronger relationship) if the X and Y are unrelated (Ho is that X does not predict Y, i.e., the regression is not significant) is .2208. We do not have any evidence to reject the null hypothesis.

The correlation coefficient is a weak .336 and is not statistically significant. It may be 0 (in the population) and we are simply looking at sampling error.

If the regression is not significant, we do not look at the regression equation. There is nothing to look at as it all may reflect sampling error.

Example 9: A researcher is interested in computing the beta of a stock. The beta of a stock measures the volatility of a stock relative to the stock market as a whole. Thus, a stock with a beta of 1 is just as volatile (risky) as the stock market as a whole. A stock with a beta of two is twice as volatile as the stock market as a whole. The Standard & Poor 500 is typically used as a surrogate for the entire stock market.

Returns (Y) Stock ABQ	Returns (X) S&P 500
0.11	0.20
0.06	0.18
-0.08	-0.14
0.12	0.18
0.07	0.13
0.08	0.12
-0.10	-0.20
0.09	0.14
0.06	0.13
-0.08	-0.17
0.04	0.04
0.11	0.14

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.973281463
R Square	0.947276806
Adjusted R Square	0.942004487
Standard Error	0.019265806
Observations	12

<i>ANOVA</i>					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	0.066688287	0.066688287	179.6698442	1.02536E-07
Residual	10	0.003711713	0.000371171		
Total	11	0.0704			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	0.006735691	0.006090118	1.106003315	0.294622245	-0.00683394	0.020305322
X Variable 1	0.532228948	0.039706435	13.40409804	1.02536E-07	0.443757482	0.620700413

This regression is significant; the F-value is 179.67. If the X-variable explains very little of the Y-variable, you should get an F-value that is 1 or less. The probability of getting the sample evidence (the X and Y input data) if the X and Y are unrelated (Ho is that X does not predict Y, i.e., the regression is not significant) is .0000001.

The regression equation is:

Returns Stock ABQ = .0067 + .5322 (Returns S&P 500).

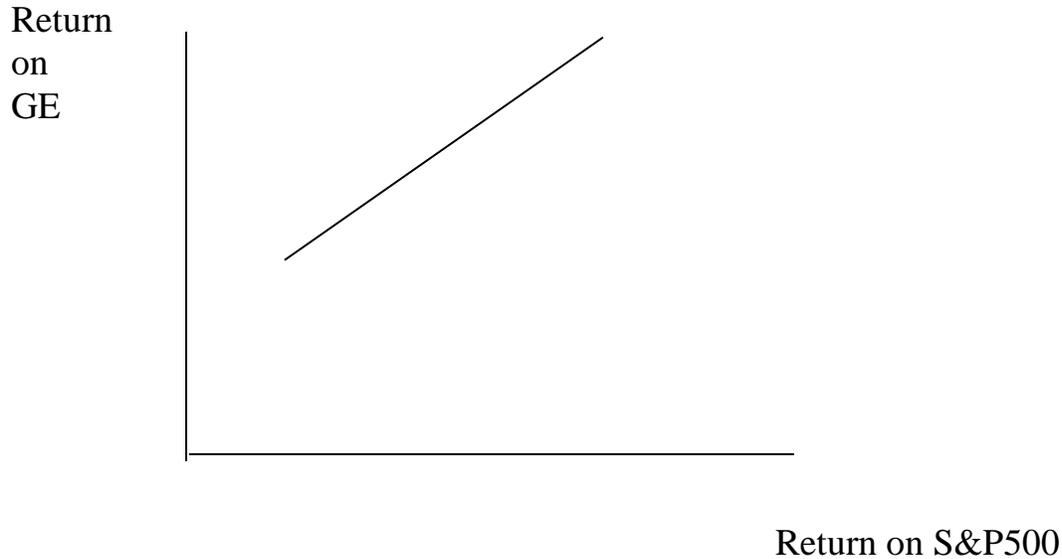
The beta of ABQ stock is .5322. It is less volatile than the market as a whole.

## Measuring a Stock's Beta:

Dependent variable: Quarterly returns on a specific stock, say GE.

Independent variable: Quarterly returns on the S&P500 which is a surrogate for the entire stock market.

[Return = difference in Price + Dividend]



$$\hat{Y} = b_0 + b_1X$$

$b_1$  = the slope of the line = the beta of the stock

if the beta = 1, GE is just as volatile as the S&P500

if the beta = 2, GE is 2 times as volatile as the S&P500

We have two rates of change and  $\frac{\Delta GE}{\Delta S \& P}$ . Do they change together (say, beta of 1.0) or differently?