Scatter Plots and Correlation

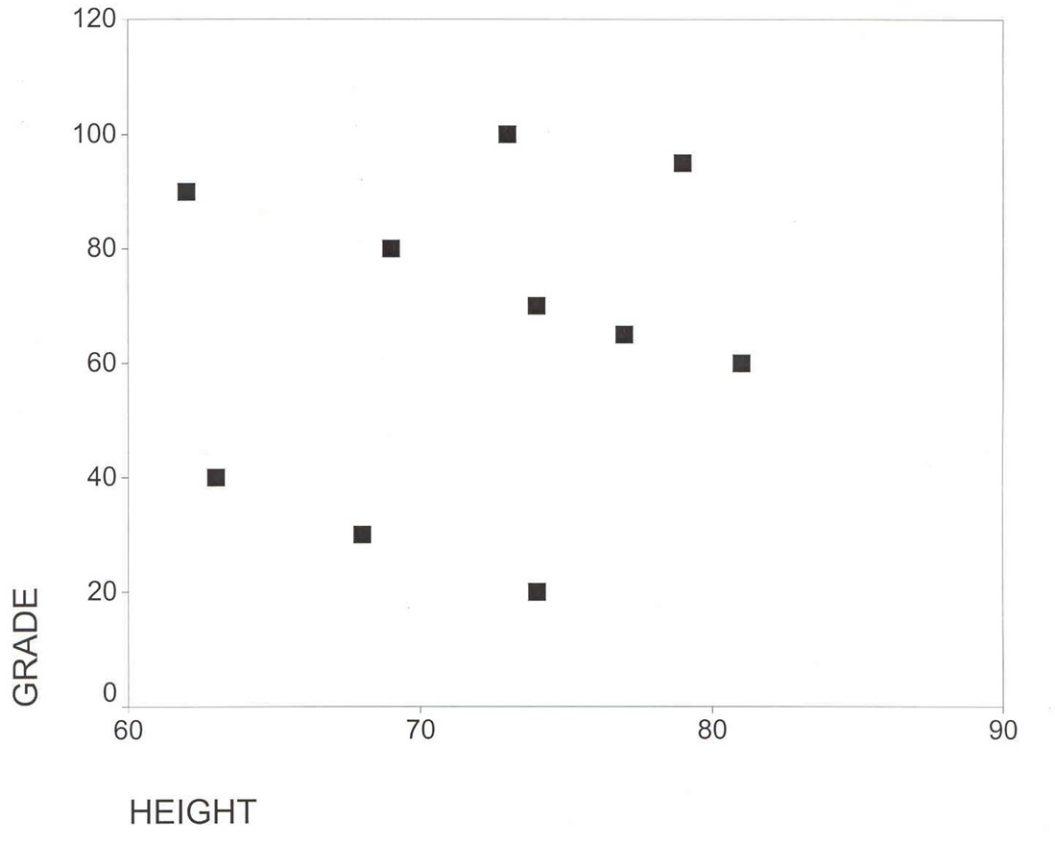
One way to see whether two variables are related is to graph them. For instance, a researcher wishes to determine whether there is a relationship between grades and height. A scatter plot will help us see whether the two variables are related. If you check the handouts, you will see how to use Excel to do a scatter plot.

Scatter Plot: Example 1

Example:

Y (Grade)	100	95	90	80	70	65	60	40	30	20
X (Height)	73	79	62	69	74	77	81	63	68	74

Height is in inches



($r = .12$; $R^2 = .01$; we will learn about r and R-squared later. A correlation coefficient, ***r***, of .12 is very weak. In this case we will ultimately find out that it is not significant, i.e., we have no evidence to reject the null hypothesis that the population correlation coefficient is 0.)

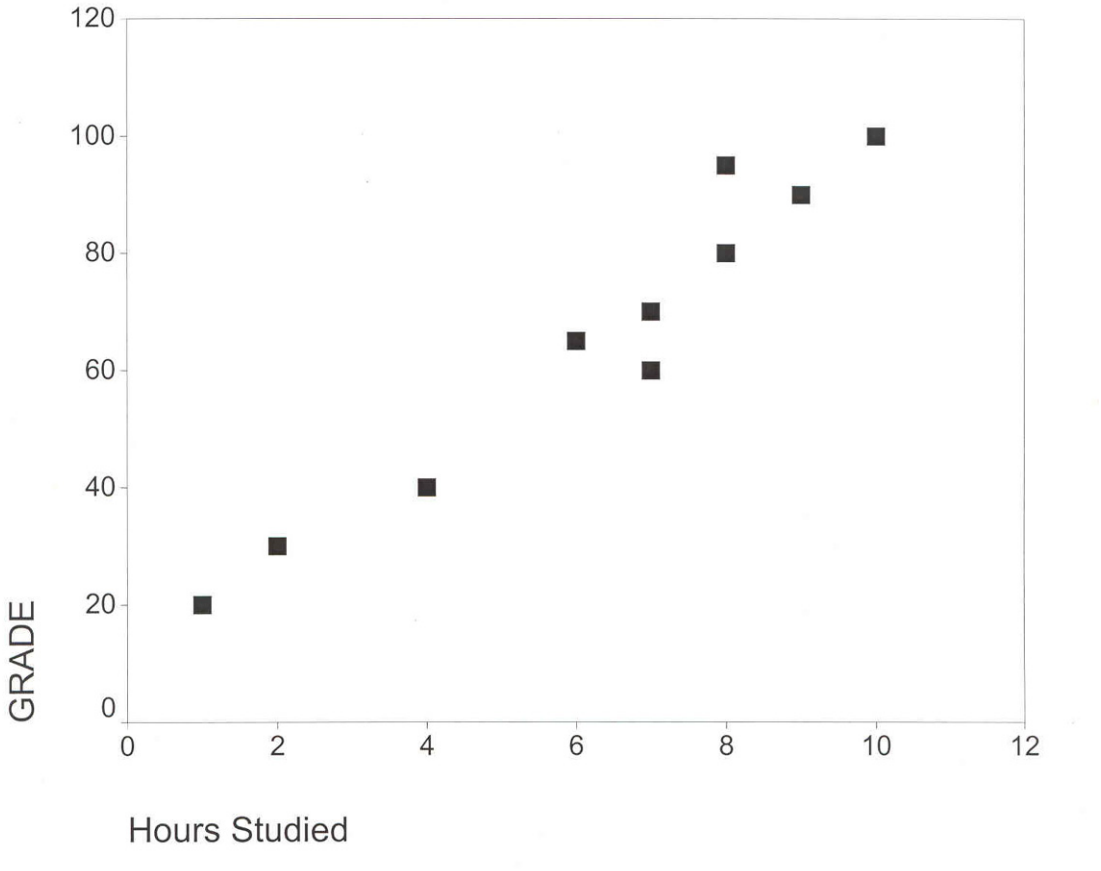
Note that the two variables do not appear to be related. Later, we will learn how to use the correlation coefficient to give us a measure to determine how weakly or strongly two variables are related.

Scatter Plot: Example two – this one's a little better. From the scatter plot below, we see that there appears to be a positive linear relationship between hours studied and grades. In other words, the more one studies the higher the grade (I am sure that this is a big surprise).

Y (Grade)	100	95	90	80	70	65	60	40	30	20
X (Hours Studied)	10	8	9	8	7	6	7	4	2	1

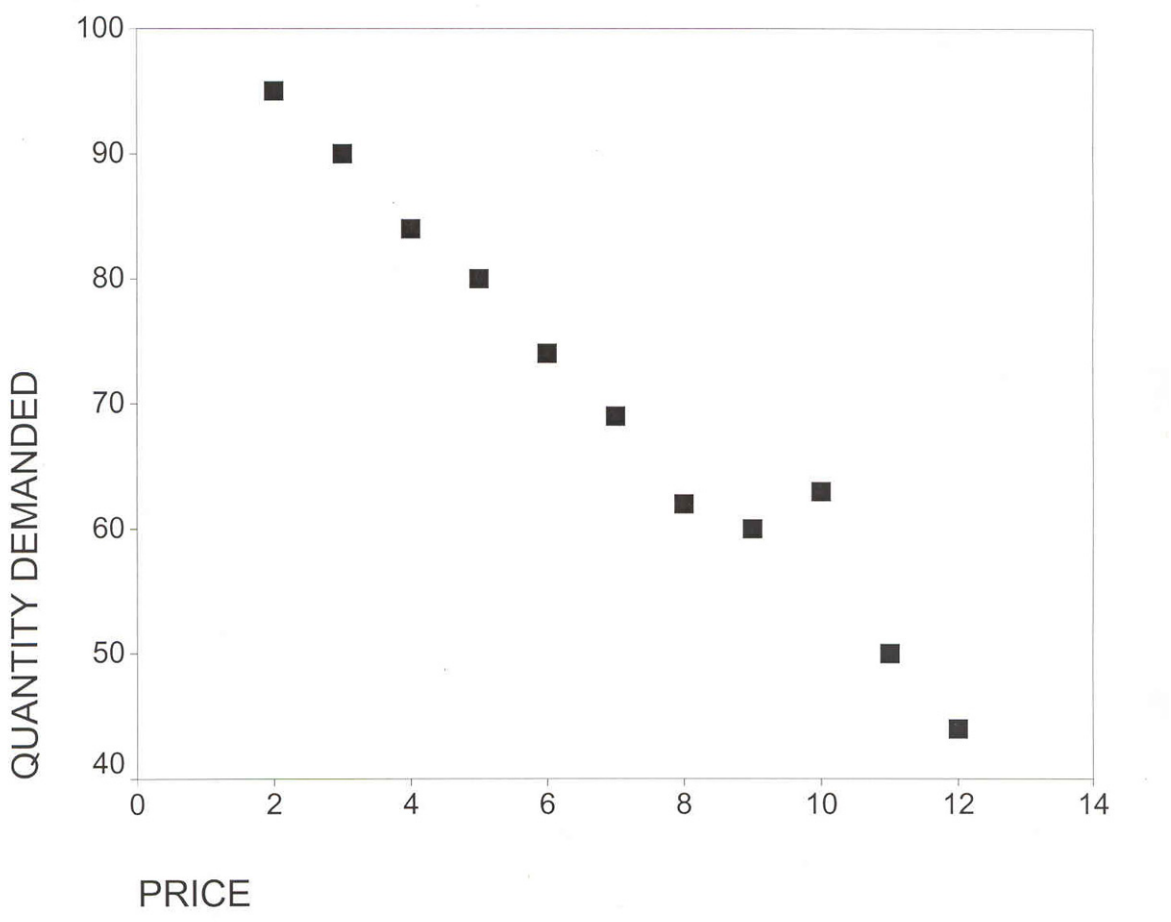
r = .97 We did not learn this yet but a correlation coefficient of .97 is very strong. The coefficient of determination, $r^2 = .94$. We will learn about this later.

$\hat{Y} = 8.92 + 9.05X$
This is the regression equation and we will also learn about this later.



Scatter Plot –Example 3

Price (X)	Quantity Demanded (Y)
$2	95
3	90
4	84
5	80
6	74
7	69
8	62
9	60
10	63
11	50
12	44



This is an example of an inverse relationship (negative correlation). When price goes up, quantity demanded goes down.

($r = -.99$; $r^2 = 98.01\%$; $\hat{Y} = 104 - 4.82X$. We will learn about this soon.)

Measuring Correlation

In correlation analysis, one assumes that both the x and y variables are random variables. We are only interested in the ***strength*** of the relationship between x and y.

Correlation represents the strength of the association between two variables.

Covariance is the starting point for measuring correlation. The covariance is used to determine the relationship between two data sets.

The unbiased formula (for a sample) is:

$$\text{COV(X,Y)} = \frac{\sum_{i=1}^{N}(x_i - \overline{X})(y_i - \overline{Y})}{n-1}$$

Where:

$\overline{X}$ = the mean of dataset X
$\overline{Y}$ = the mean of dataset Y
N = the population size
n = the sample size

NOTE: The formula for the population is:

$$\text{COV(X,Y)} = \frac{\sum_{i=1}^{N}(x_i - \overline{X})(y_i - \overline{Y})}{N}$$

To compute the Correlation Coefficient, r:

$$r_{xy} = \frac{COV(X,Y)}{s_x s_y}$$

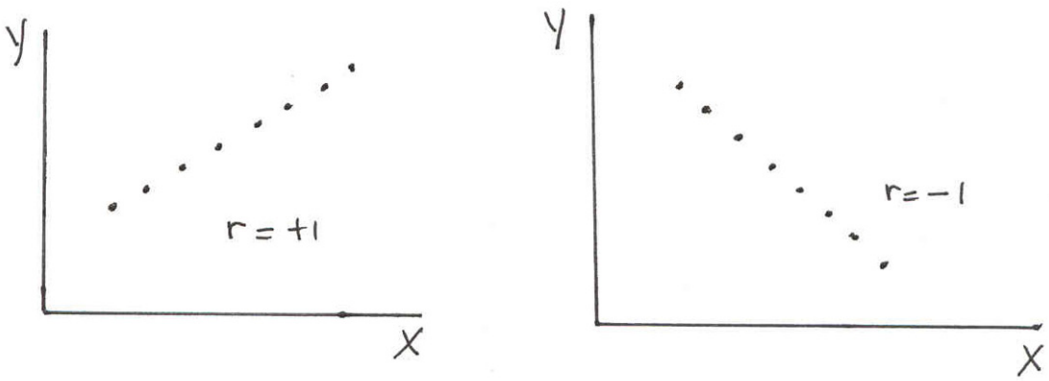
$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{\left[n\sum X^2 - \left(\sum X\right)^2\right]\left[n\sum Y^2 - \left(\sum Y\right)^2\right]}}$$

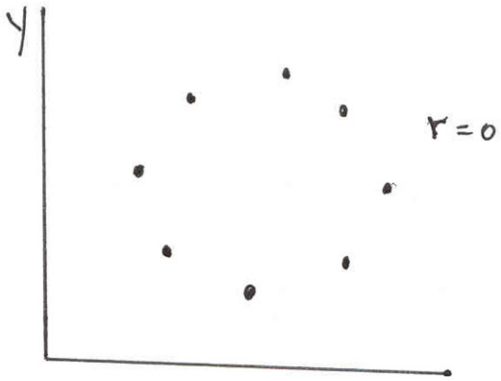
where n = the number of PAIRS of observations

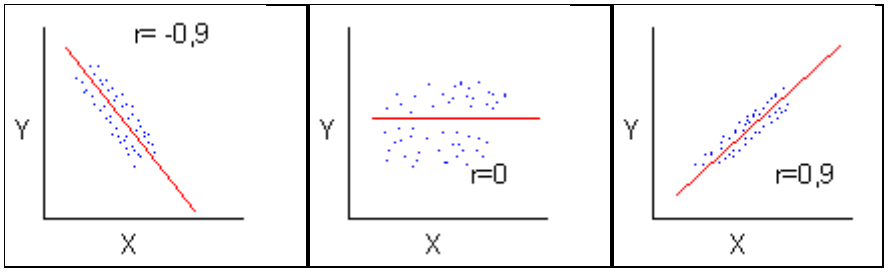
r is the correlation coefficient and ranges from -1 to +1. A correlation of coefficient of +1 indicates a perfect positive linear relationship between the variables X and Y. In fact, if we did a scatter plot, all the points would be on the line. This indicates that X can be used to predict Y perfectly. Of course, in real life, one almost never encounters perfect relationships between variables. For instance, it is certainly true that there is a very strong positive relationship between hours studied and grades. However, there are other variables that affect grades. Two students can spend 20 hours studying for an exam and one will get a 100 on the exam and the other will get an 80. This indicates that there is also random variation and/or other variables that explain performance on a test (e.g., IQ, previous knowledge, etc.).

A correlation of -1 indicates a perfect negative linear relationship (i.e., an inverse relationship). In fact, if we did a scatter plot, all the points are on the line. This indicates that X can be used to predict Y perfectly.

A correlation of 0 indicates absolutely no relationship between X and Y. In real life, correlations of 0 are very rare. You might get a correlation of .10 and it will not be significant, i.e., it is not statistically different from 0. (We will learn how to test correlations for significance.)

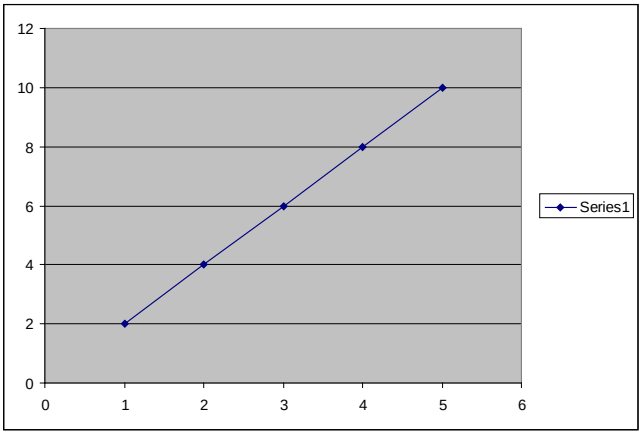

y
r = +1
X
y
r = −1
X


y
r = 0


r= -0,9
Y
X
Y
r=0
X
Y
r=0,9
X

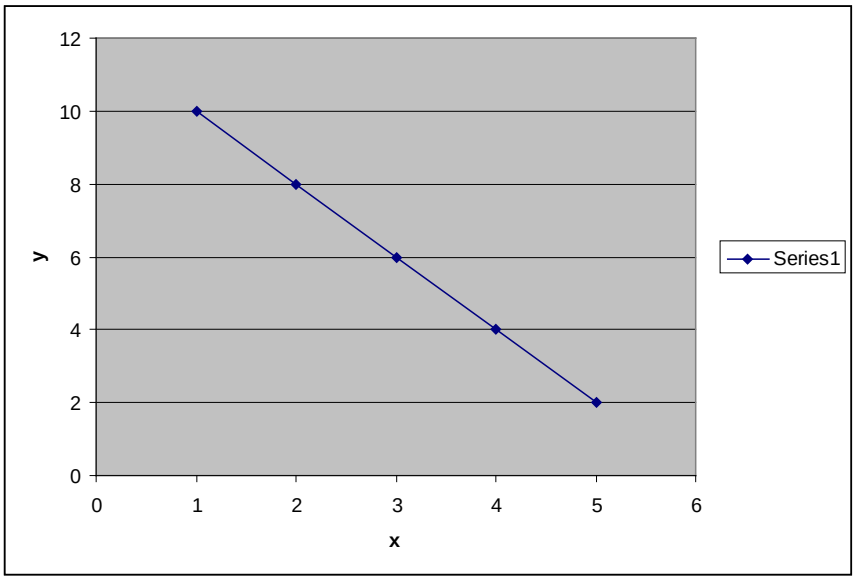
If there is a perfect relationship between X and Y then

X	Y
1	2
2	4
3	6
4	8
5	10
correlation	1



Or …

X	y
1	10
2	8
3	6
4	4
5	2
correlation	-1



Correlation does NOT imply causality:
4 possible explanations for a significant correlation:

X causes Y
Y causes X
Z causes both X and Y
Spurious correlation (a fluke)

Examples:

- Poverty and crime are correlated. Which is the cause?
- ADHD and hours TV watched by child under age 2. Study claimed that TV caused ADHD. Do you agree?
- 3% of older singles suffer from chronic depression; does being single cause depression?
- Cities with more cops also have more murders. Does 'more cops' cause 'more murders'? If so, get rid of the cops!
- There is a strong inverse correlation between the amount of clothing people wear and the weather; people wear more clothing when the temperature is low and less clothing when it is high. Therefore, a good way to make the temperature go up during a winter cold spell is for everyone to wear very little clothing and go outside.
- There is a strong correlation between the number of umbrellas people are carrying and the amount of rain. Thus, the way to make it rain is for all of us to go outside carrying umbrellas!

The correlation coefficient, r, ranges from -1 to +1. The coefficient of determination, r^2 (in Excel, it is called R-squared) is also an important measure. It ranges from 0% to 100% and measures the proportion of the variation in Y explained by X. If all the points are on the line, r = 1 (or -1 if there is an inverse relationship), then r^2 is 100%. This means that all of the variation in Y is explained by (variations) X. This indicates that X does a perfect job in explaining Y and there is no unexplained variation.

Thus, if r = .30 (or -.30), then r^2 = 9%. Only 9% of the variation in Y is explained by X and 91% is unexplained. This is why a correlation coefficient of .30 is considered weak—even if it is significant.

If r = .50 (or -.50), then r^2 = 25%. 25% of the variation in Y is explained by X and 75% is unexplained. This is why a correlation coefficient of .50 is considered moderate.

If r = .80 (or -.80), then r^2 = 64%. 64% of the variation in Y is explained by X and 36% is unexplained. This is why a correlation coefficient of .8 is considered strong.

If r = .90 (or -.90), then r^2 = 81%. 81% of the variation in Y is explained by X and 19% is unexplained. This is why a correlation coefficient of .90 is considered very strong.

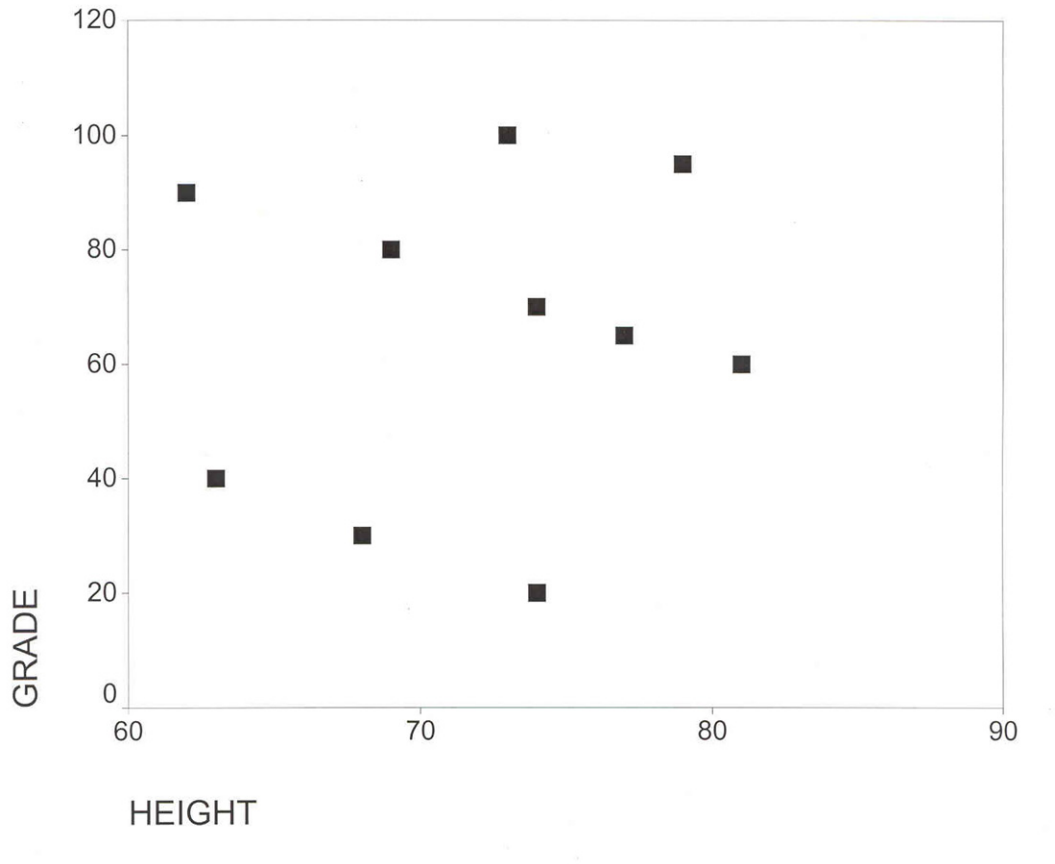
What would you say about a correlation coefficient of .20?

[Answer: even if it turns out to be significant, it will be of little practical importance. R-squared is 4% and 96% of the variation in Y is still unexplained.]

Example 1 (from above):

Y (Grade)	100	95	90	80	70	65	60	40	30	20
X (Height)	73	79	62	69	74	77	81	63	68	74

Height is in inches



Note that the two variables do not appear to be related. Let's get a statistical measure of this. We need:

$\sum X_i = 720$
$\sum Y_i = 650$
$\sum X_i Y_i = 46{,}990$
$\sum X_i^2 = 52{,}210$
$\sum Y_i^2 = 49{,}150$

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{\left[n\sum X^2 - \left(\sum X\right)^2\right]\left[n\sum Y^2 - \left(\sum Y\right)^2\right]}}$$

$$r = \frac{10(46{,}990) - 720(650)}{\sqrt{\left[10(52{,}210) - (720)^2\right]\left[10(49{,}150) - (650)^2\right]}}$$

$$= \frac{1900}{\sqrt{[3{,}700][69{,}000]}} = .1189$$

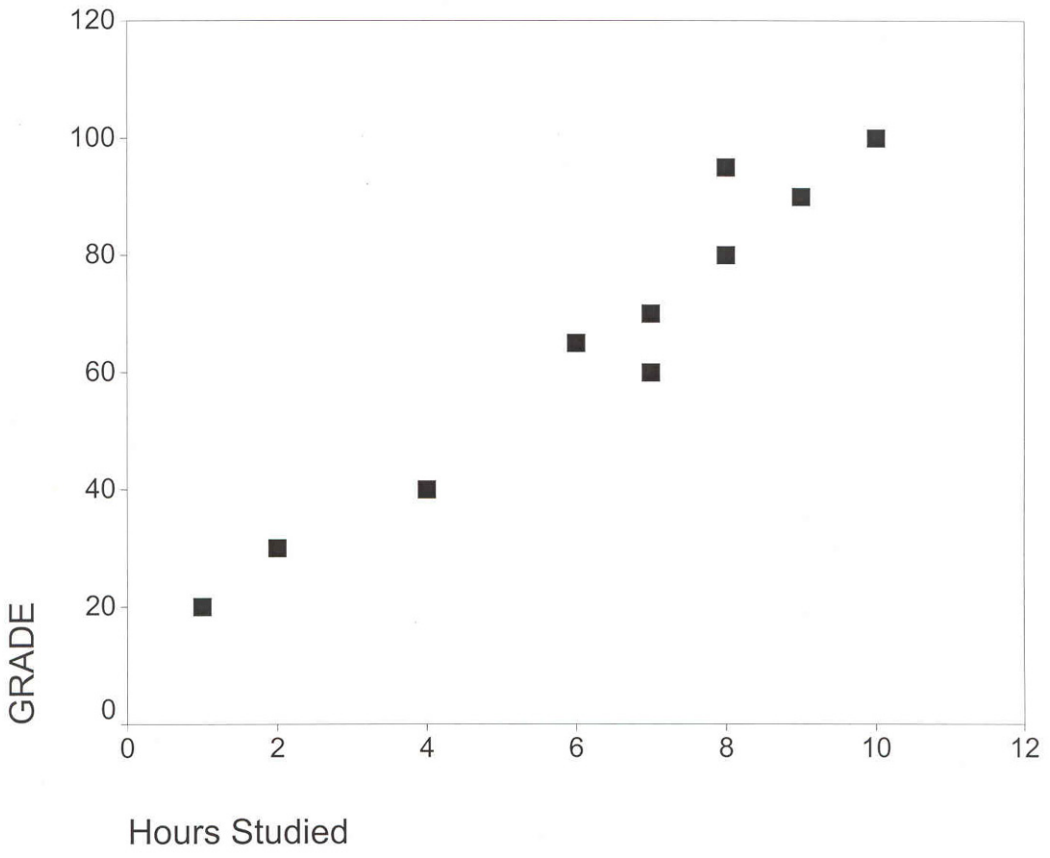
$r^2 = 1.4\%$

[To test the significance of the correlation coefficient, a t-test can be done. We will learn how to use Excel to test for significance.]

The correlation coefficient is not significant (you have to trust me on this). A correlation coefficient of .1189 is not significantly different from 0. Thus, there is no relationship between height and grades. Correlation coefficients of less than .30 are generally considered very weak and of little practical importance even if they turn out to be significant.

Example 2 (from above):

Y (Grade)	100	95	90	80	70	65	60	40	30	20
X (Hours Studied)	10	8	9	8	7	6	7	4	2	1



$\sum X_i = 62$
$\sum Y_i = 650$
$\sum X_i Y_i = 4{,}750$
$\sum X_i^2 = 464$
$\sum Y_i^2 = 49{,}150$

$$r = \frac{10(4{,}750) - 650(62)}{\sqrt{\left[10(464) - (62)^2\right]\left[10(49{,}150) - (650)^2\right]}}$$

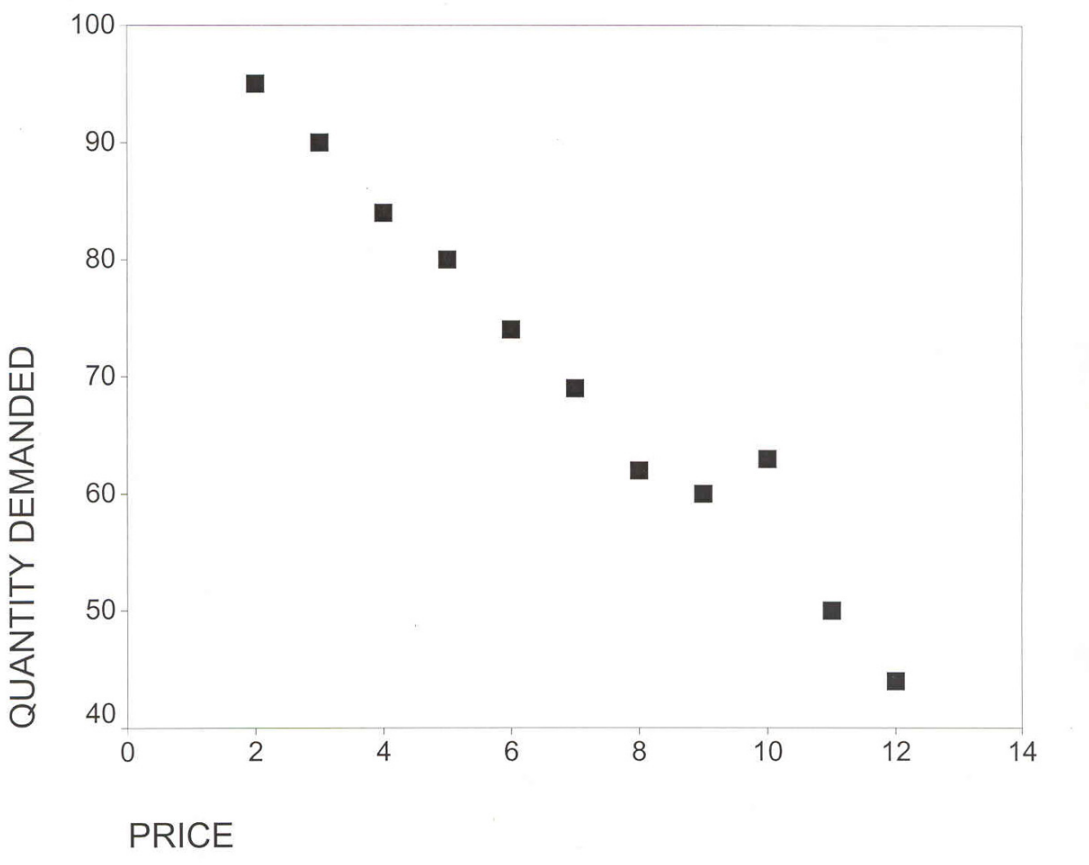
$$= \frac{7200}{\sqrt{[796][69{,}000]}} = .97$$

$r^2 = 94.09\%$

To test the significance of the correlation coefficient, a t-test can be done. We will learn how to use Excel to test for significance. The correlation coefficient is significant (again, you have to trust me on this). A correlation coefficient of .97 is almost perfect. Thus, there is a significant relationship between hours studied and grades. Correlation coefficients of more than .80 are generally considered very strong and of great practical importance.

Example 3 (from above):

X (price)	Y (Quantity Demanded)
$2	95
3	90
4	84
5	80
6	74
7	69
8	62
9	60
10	63
11	50
12	44



$\sum X_i = 77$
$\sum Y_i = 771$
$\sum X_i Y_i = 4{,}867$
$\sum X_i^2 = 649$
$\sum Y_i^2 = 56{,}667$

$$r = \frac{11(4867) - 77(771)}{\sqrt{\left[11(649) - (77)^2\right]\left[11(56{,}667) - (771)^2\right]}}$$

$$= \frac{-5{,}830}{\sqrt{[1210][28{,}896]}} = -.99$$

$r^2 = 98.01\%$

To test the significance of the correlation coefficient, a t-test can be done. We will learn how to use Excel to test for significance. The correlation coefficient is significant (again, you have to trust me on this). A correlation coefficient of -.99 is almost perfect. Thus, there is a significant and strong inverse relationship between price and quantity demanded.

Example 4: Does attractiveness affect an individual's salary?
Note: The more attractive the person, the higher the attractive score. Scale goes from 0 to 10.

X (attractiveness score)	Starting Salary (income in thousands)
0	20
1	24
2	25
3	26
4	20
5	30
6	32
7	38
8	34
9	40

$\sum X_i = 45$
$\sum Y_i = 289$
$\sum X_i Y_i = 1{,}472$
$\sum X_i^2 = 285$
$\sum Y_i^2 = 8{,}801$

$$r = \frac{10(1472) - 45(289)}{\sqrt{\left[10(285) - (45)^2\right]\left[10(8801) - (289)^2\right]}}$$

$$= \frac{1715}{\sqrt{[825][4489]}} = .891$$

$r^2 = 79.39\%$

To test the significance of the correlation coefficient, a t-test can be done. We will learn how to use Excel to test for significance. The correlation coefficient is significant (again, you have to trust me on this). A correlation coefficient of .891 is strong. Thus, there is a significant and strong relationship between attractiveness and starting salary.

[Was the original question answered?]