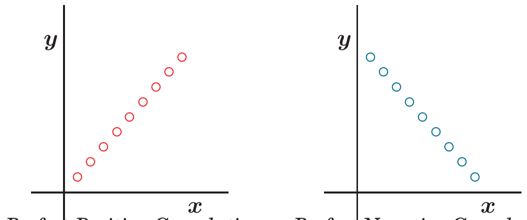
21. Correlation

Sometimes two measurements on a single individual will appear to be related:

Years of Education	income
height	weight
blood pressure	cholesterol

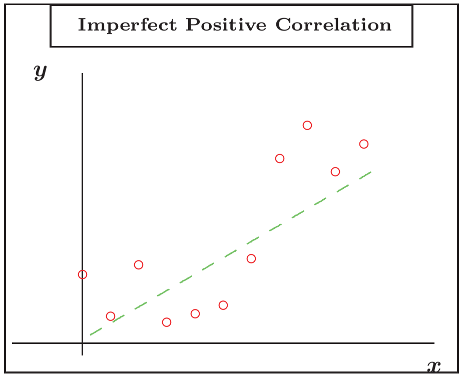
- In each case, we are taking two measurements on a single individual.
- Experience (or reasoning) suggests that the two measurements are not independent: whenever we observe a change in one we will also observe a change in the other.

- “Correlation” is a measure of straight-line relationships:



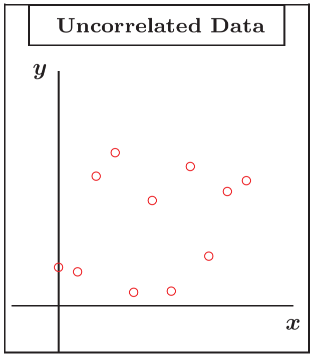
Perfect Positive Correlation *Perfect Negative Correlation*

The circles (○) represent actual observations. In perfect positive correlation, as x increases, so does y. In perfect negative correlation, as x increases, y decreases.
In the real world, you never get perfect correlation of either type; at the least, there will be observational errors which cause some of the data points to slightly miss the straight line.

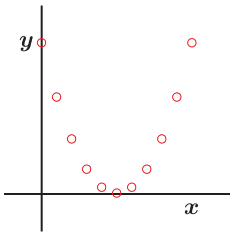


- Other times the observations may turn out to be completely unrelated or *uncorrelated* – we would expect that shoe size and income are

uncorrelated, for example.



- Correlations only measure straight line relationships:



The data at left are *uncorrelated* even though they are obviously related (parabolically).

Remark If we were to graph x against the logarithm of y ($\ln(y)$) we would get a straight-line relationship. This kind of transformation is often done to linearize nonlinear relationships.

Given a set of paired data points

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \cdots, (x_n, y_n)$$

it is possible to compute a number, called the *correlation coefficient* ρ, which measures how closely the data fall on a straight line.

- The symbol ρ is used when census data are the basis for the calculation.
- The symbol r is used when sample data are the basis for the calculation.

The correlation coefficient ρ has the following properties:

- $-1 \leq \rho \leq +1$ and $-1 \leq r \leq +1$
- If $\rho = 0$ then the data are uncorrelated.
- If $\rho = +1$ then the data have a perfect positive correlation.
- If $\rho = -1$ then the data have a perfect negative correlation.

One of the things we can do is test the hypotheses:

$$H_0 : \ \rho = 0 \text{ against}$$
$$H_A : \ \rho > 0 \text{ or}$$
$$H_A : \ \rho < 0$$

to decide if the two variables are really correlated.

Suppose that the two variables really are correlated. Think of x as the input (or independent) variable and y as the output (or dependent) variable. Some of the variability in y is due the influence of x. Some of the variability in y is due to residual factors not measured by x (such as sampling error, measurement error, other things that might influence y). For example, we might gather data on each subject's income and educational achievement. We'd expect to see a positive correlation between the two: as education goes up, income goes up. We might even test the hypothesis that this is true, i.e., that

$$H_A : \ \rho > 0.$$

A second thing we might be interested in is whether or not education predicts income, i.e., whether there is a reliable formula that connects the two with reasonably small error. This is a slightly different hypothesis, since it is saying that education is a primary source of the variablity in income, something that is probably untrue.

The value of ρ^2 measures the *proportion* of variability in y that is due to the influence of x.

$$\rho^2 \approx \begin{cases} \text{proportion of variability in } y \\ \text{due to the influence of } x. \end{cases}$$

While we will develop tests for how well the independent variable x predicts the dependent variable y, there is a rough standard that generally applies.

- *Accepted Practice:* you should not use x to predict y unless ρ^2 is at least 0.16 (or the absolute value of ρ ($|\rho|$) is at least 0.4).

There are several equivalent formulas for calculating the correlation coefficient. The simplest to write down is probably

$$r = \frac{\text{average of the products} - \mu_x\mu_y}{\sigma_x\sigma_y}$$

Conceptually, the products in the numerator measure the interaction between x and y. Dividing by the product of the standard deviations eliminates the "scale" or units from the number–in effect, standardizes it.

A mathematically equivalent formula that is computationally less sensitive to rounding and hence more frequently used in textbooks and other settings is

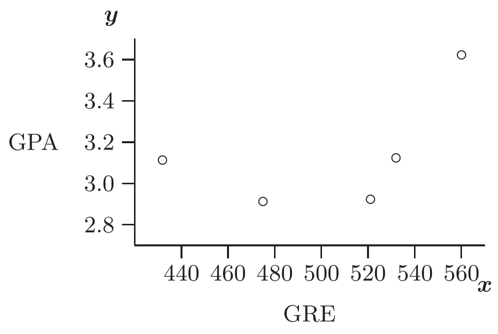
$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{(n\sum x^2 - (\sum x)^2)\,((n\sum y^2 - (\sum y)^2)}}$$

21.1. Example.

Consider the following data:

GRE Scores	1st year GPA
475	2.91
521	2.92
532	3.12
560	3.62
432	3.11

Find the correlation coefficient for the above data.
Note that the data have the following graph:



Solution. You can find the correlation coefficient r with the formula:

$$r = \frac{\text{average of the products} - \mu_x\mu_y}{\sigma_x\sigma_y}$$

where the “average of the products” is the average of the products of the pair of observations (x, y) that you have for each subject. (When using this formula, it is essential that you use the key on your calculator which finds the "*population*" standard deviation (σ_n).)

Step 1. To use the formula, make a "products" column with your table of data:

GRE Scores	1st year GPA	xy
475	2.91	1382.25
521	2.92	1521.32
532	3.12	1659.84
560	3.62	2027.20
432	3.11	1343.52

Step 2. Find the means for all three columns and the standard deviations for the first two:

	GRE Scores	*1st year GPA*	xy
means	μ_x = 504	μ_y = 3.14	μ_{xy} = 1585.83
st. dev	σ_x = 45.24	σ_y = 0.26	–

Step 3. Apply the formula to find the correlation coefficient:

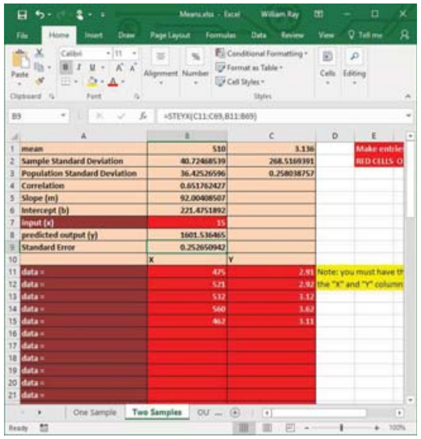
$$
\begin{aligned}
r &= \frac{\text{average of the products} - \mu_x\mu_y}{\sigma_x\sigma_y} \\
&= \frac{1585.83 - (504)(3.14)}{(45.24)(0.26)} \\
&= 0.538
\end{aligned}
$$

(If you use the rounded values for the various calculated values reported in the table, you get $r = 0.34$. Again, these calculations are sensitive round-off error.)

In this problem, $r^2 = 0.289$, so about 28% of the variability in first year GPA can be predicted by the GRE. The remaining variability is due to residual factors (such as motivation, persistence, support, etc.)

Remark 1. The above formula is fairly simple, but is *extremely* sensitive to round-off error. For this reason, most books will use the more complex but equivalent formula given above.

Remark 2. As usual, there is a built-in Excel function to calculate the correlation coefficient. So, in this class, it's never necessary to do this calculation by hand. See the spreadsheet `Means` at right.



=STEYX(C11:C69,B11:B69)

	A	B	C	D	E
1	mean	510	3.136		Make entrie
2	Sample Standard Deviation	40.72468539	268.5169391		RED CELLS O
3	Population Standard Deviation	36.42526596	0.258038757		
4	Correlation	0.651762427			
5	Slope (m)	92.00408507			
6	Intercept (b)	221.4751892			
7	input (x)	15			
8	predicted output (y)	1601.536465			
9	Standard Error	0.252650942			
10		X	Y		
11	data =	475	2.91	Note: you must have th	
12	data =	521	2.92	the "X" and "Y" column	
13	data =	532	3.12		
14	data =	560	3.62		
15	data =	462	3.11		
16	data =				
17	data =				
18	data =				
19	data =				
20	data =				
21	data =				

It is essential that you not confuse the notions of "correlation" and "cause and effect."
High correlations do NOT necessarily imply a cause-and-effect relationship!!

21.2. Example.

Ice cream sales and deaths by drowning are correlated at $r = 0.83$.

- *Does this mean that ice cream sales cause deaths by drowning (for example, by causing stomach cramps)?*
- *Maybe it means that people get upset by seeing a drowning and so they assuage their grief by eating ice cream?*

21.3. Example.

Among elementary school children, math scores on a standardized test and weight are correlated at $r = 0.72$.

- *Should we encourage overweight children in order to improve math scores?*
- *Maybe kids who are good at math are unfit and fat?*

NOTE: High correlations DO NOT IMPLY that cause-effect relationships exist!!!!

22. Linear Regression

If observations are correlated (if the absolute value $|r|$ of the correlation coefficient is at least 0.4), we can use the input x to predict the output y.

22.1. Example.

Fat content of the body is of great medical and physiological importance, influencing for example death rates, the effectiveness of drugs and anesthetics. Fat content is generally calculated from body density, with higher fat values corresponding to lower body density. Body density is not easy to calculate; the most accurate method requires the subject to be submerged under water. Another method involves averaging certain skinfold measurements.

In a sample of 16 males aged 20-29 both skinfold and body density measurements were taken; the following data were gathered:

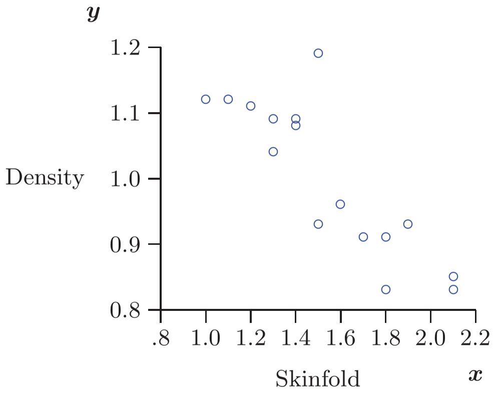
Skinfold	*Density*
1.0	1.12
1.1	1.12
1.2	1.11
1.3	1.09
1.3	1.04
1.4	1.09
1.4	1.08
1.5	1.19
1.5	0.93
1.6	0.96
1.7	0.91
1.8	0.83
1.8	0.91
1.9	0.93
2.1	0.83
2.1	0.85

The average skinfold was found to be 1.55 with a standard deviation of 0.33.

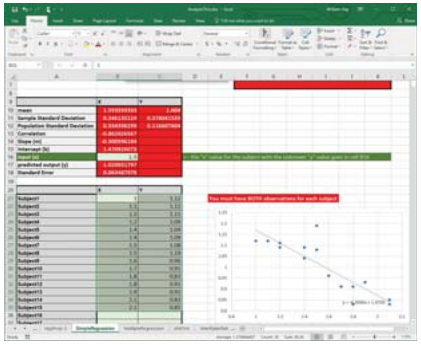
The average body density was found to be 1.00 with a standard deviation of 0.11. The variables "skinfold" and "body density" were shown to be correlated at $r = -0.86$.

You are confronted with a 23 year old male whose skinfold measurement is 1.5. Estimate his body density.

Note that the data have the following graph:



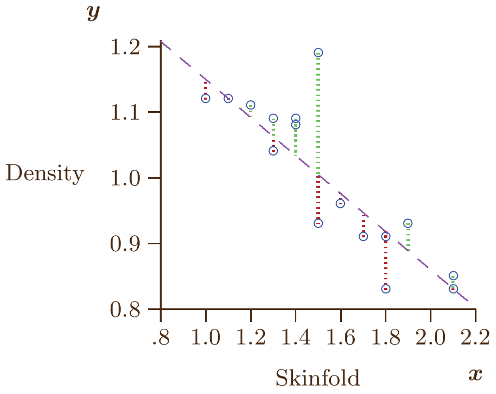
Remark 1. Since this problem gives you the actual data, you could use the `Simple Regression` tab of the `AnalyzeThis` spreadsheet to obtain the answer. Note that the spreadsheet even gives you the above scatterplot. However, since the problem also gives you the summary statistics, you can use the `Formulas` spreadsheet and avoid entering the data.



	X	Y
mean	[illegible]	[illegible]
Sample Standard Deviation	[illegible]	[illegible]
Population Standard Deviation	[illegible]	[illegible]
Correlation	[illegible]	
Slope (m)	[illegible]	
Intercept (b)	[illegible]	
input (x)	[illegible]	
predicted output (y)	[illegible]	
Standard Error	[illegible]	

Solution. The idea is to find a formula for a line which best approximates the data. Since the correlation is not perfect, some of the data

points will miss the line which best fits the data:



The vertical (dotted) lines represent the “error” between the actual value of y (the data point ∘) and the predicted value of y which is on the

(dashed) line. The *regression line* is the straight line which results in the mean squared error (average of the squares of the error distances) being minimized. (For this reason the regression line is sometimes called the least squares line.)
The general formula for a straight line is

$$y = mx + b.$$

The method involves first finding the parameters m and b and then using the particular observation (the 1.5 skinfold for our 23 year old male) to predict body density.
Step 1. The first step is to make a list of the data. The hardest step is to decide which measurement will be x and which will be y.

The value you are going to *predict* will be y.

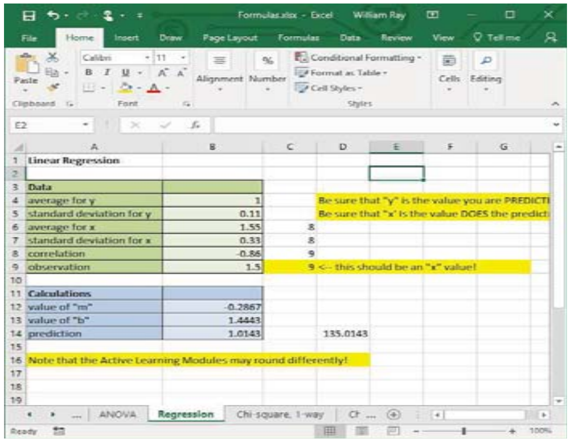
The value that *does the predicting* will be x.

In this problem we are going to try to predict body density; since this is what we are trying to predict, body density must be y. We will be using skinfold measurements to do the prediction, so x must be skinfold measurements. This will be the summary data that we put in the spreadsheet, which will calculate values for m and b and then use those to do predictions based on skinfold measurements.

$\bar{x}$	1.55
s_x	0.33
$\bar{y}$	1.00
s_y	0.11
r	-0.86
x_0	1.5

The x_0 represents the particular observation given in the problem.

Step 2. Read the predicted results from the spreadsheet.



	A	B	C	D
1	Linear Regression			
2				
3	Data			
4	average for y	1		Be sure that "y" is the value you are PREDICTI
5	standard deviation for y	0.11		Be sure that "x" is the value DOES the predicti
6	average for x	1.55	8	
7	standard deviation for x	0.33	8	
8	correlation	-0.86	9	
9	observation	1.5	9	<-- this should be an "x" value!
10				
11	Calculations			
12	value of "m"	-0.2867		
13	value of "b"	1.4443		
14	prediction	1.0143		135.0143
15				
16	Note that the Active Learning Modules may round differently!			

Question: why does this answer not agree with either of our observations for 1.5 skinfold? Why is it even different from the *average* of the two 1.5 observations?

Solution Template

Step 1. Make a dictionary which assigns values to the variables. Before making your table, check to see if you are trying to do a prediction. If so, the quantity you are trying to predict must be y and the quantity used to do the prediction must be x.

average for y	$\bar{y}$
st. dev. for y	s_y
average for x	$\bar{x}$
st. dev. for x	s_x
correlation coeef.	r
indiv. x observation	x_0

The individual x_0 observation is the observation for a specific individual

which will be used to predict y.

The value you are going to *predict* will be y.

The value that *does the predicting* will be x.

Step 2. Now enter the list into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled `REGRESSION`. You can then read the predicted value out of the spreadhsheet, along with the calculated values for m and b.

End of Solution Template

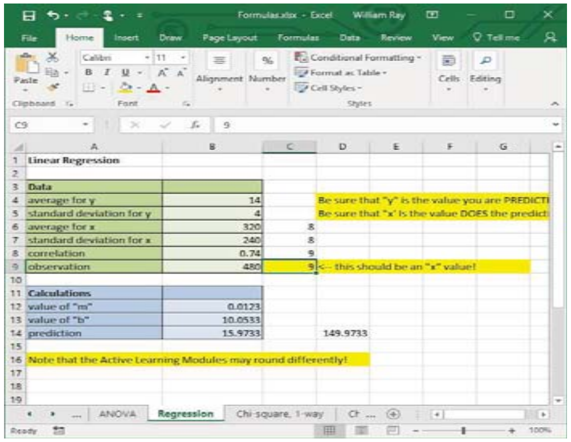
22.2. Example.

A researcher administered varying doses of caffeine to twenty-four randomly selected male subjects aged 24-30, following which their blood pressure levels were measured. The average caffeine dosage was 320 mgs (about the equivalent of four cups of coffee) with a standard deviation of 240 mgs. The average increase in systolic blood pressure was 14 mm HG with a standard deviation of 4 mm. In this study caffeine consumption and increase in systolic blood pressure were found to be correlated with a correlation coefficient of $r = 0.74$. You are confronted with a subject whose systolic blood pressure is 134 mm HG. If this subject drinks 6 cups of coffee (which will contain 480 mgs of caffeine), predict the subject's blood pressure.

Solution.
Step 1. We will begin by predicting the subject's increase in blood pressure. Thus, in this problem, we are trying to predict increase in blood pressure; thus $y =$"blood pressure increase." We are using caffeine dosage as our predictor, so $x =$"caffeine." Summarizing the data:

$\bar{y}$	14
s_y	4
$\bar{x}$	320
s_x	240
r	0.74
x_0	480

Step 2. Read the predicted results from the spreadsheet.



C9 = 9

	A	B	C	D	E	F	G
1	Linear Regression						
2							
3	Data						
4	average for y	14		Be sure that "y" is the value you are PREDICTI			
5	standard deviation for y	4		Be sure that "x' is the value DOES the predict			
6	average for x	320	8				
7	standard deviation for x	240	8				
8	correlation	0.74	9				
9	observation	480	9	<-- this should be an "x" value!			
10							
11	Calculations						
12	value of "m"	0.0123					
13	value of "b"	10.0533					
14	prediction	15.9733		149.9733			
15							
16	Note that the Active Learning Modules may round differently!						
17							
18							
19							

This gives the *increase* in blood pressure, but the problem asks for the blood pressure. Thus the answer

$$137 + 15.9733 = 149.9733.$$

■

This section dealt with simple regression because we used only one independent variable x. In the real world you will usually have several inputs contributing to a single output. For example, if the output is

$$y = \text{Income}$$

then some inputs might be

$$\begin{aligned} x_1 &= \text{age} \\ x_2 &= \text{education} \\ x_3 &= \text{gender} \\ x_4 &= \text{experience} \end{aligned}$$

and the regression line would be

$$\boldsymbol{y = m_1x_1 + m_2x_2 + m_3x_3 + m_4x_4 + b.}$$

This is called *multiple regression*; the concepts are the same but the computations are much more complex to do by hand. Spreadsheets and statistics programs make the calculations easy, of course. The spreadsheet AnalyzeThis includes tabs for both simple and multiple regression, with the latter permitting up to seven independent variables.