

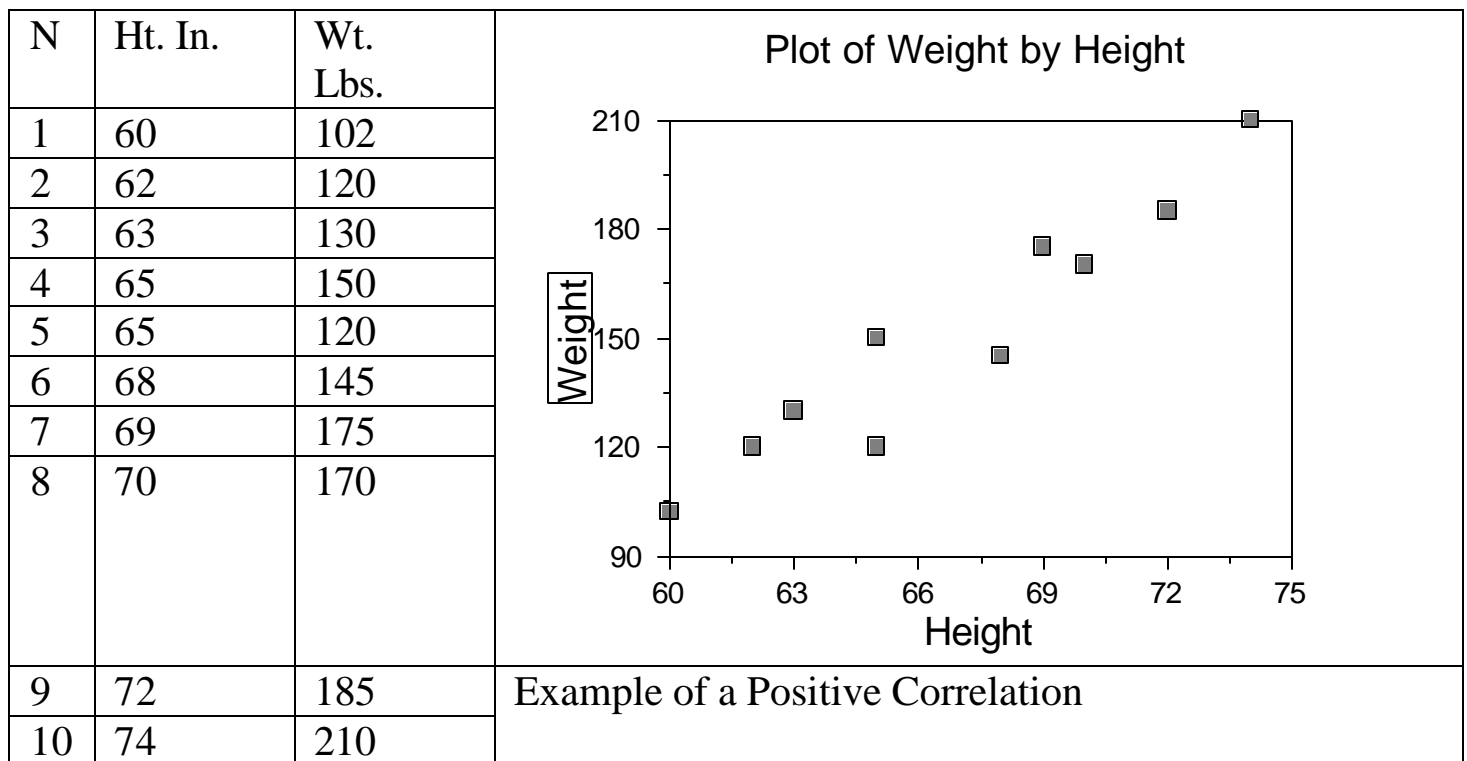
Correlation and Regression

There are two *major* different types of data analysis, depending on the type of IV

- Nominal IVs use t or ANOVA
- Continuous IVs use correlation or regression

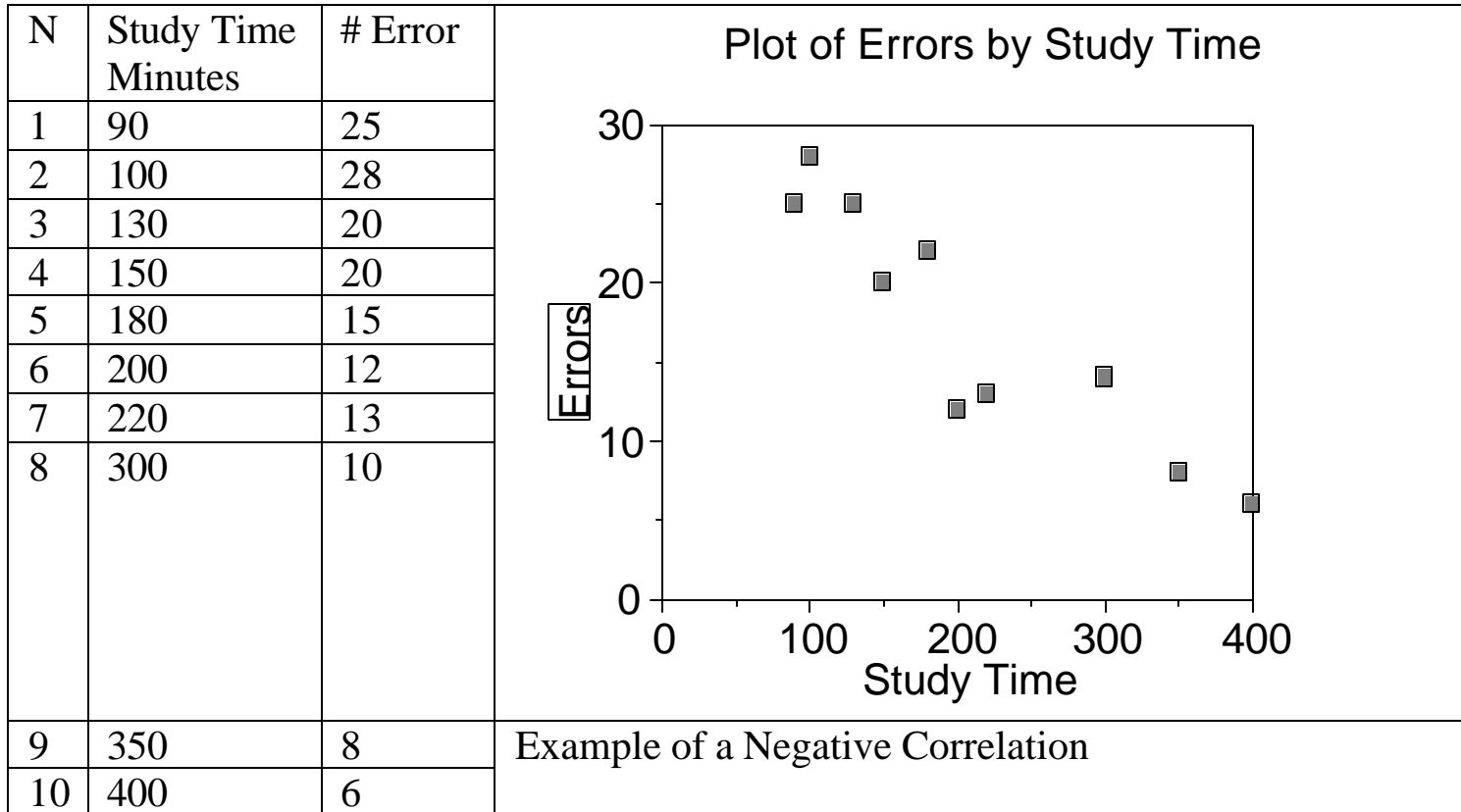
Linear correlation

- The **correlation coefficient** is a number that summarizes the direction and degree (closeness) of linear relations between two variables. The correlation coefficient is also known as the *Pearson Product-Moment Correlation Coefficient*.
- The sample value is called r , and the population value is called ρ (rho).
- The correlation coefficient can take values between -1 through 0 to +1.
- The sign (+ or -) of the correlation affects its interpretation. When the correlation is positive ($r > 0$), as the value of one variable increases, so does the other. For example, on average, as height in people increases, so does weight.



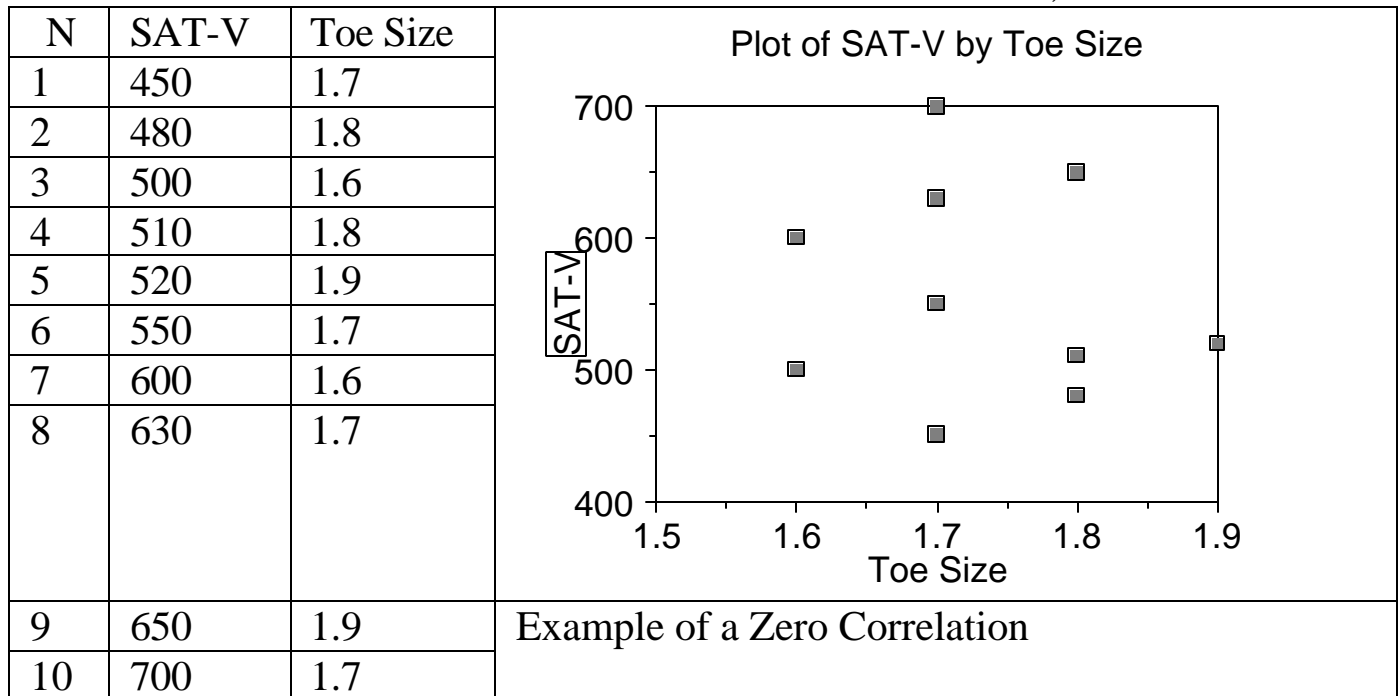
- If the correlation is positive, when one variable increases, so does the other.

- If a correlation is negative, when one variable **increases**, the other variable **decreases**.



- If the correlation is negative, when one variable increases, the other decreases.

- If there is no relationship between the two variables, then as one variable increases, the other variable neither increases nor decreases. In this case, the correlation is zero.



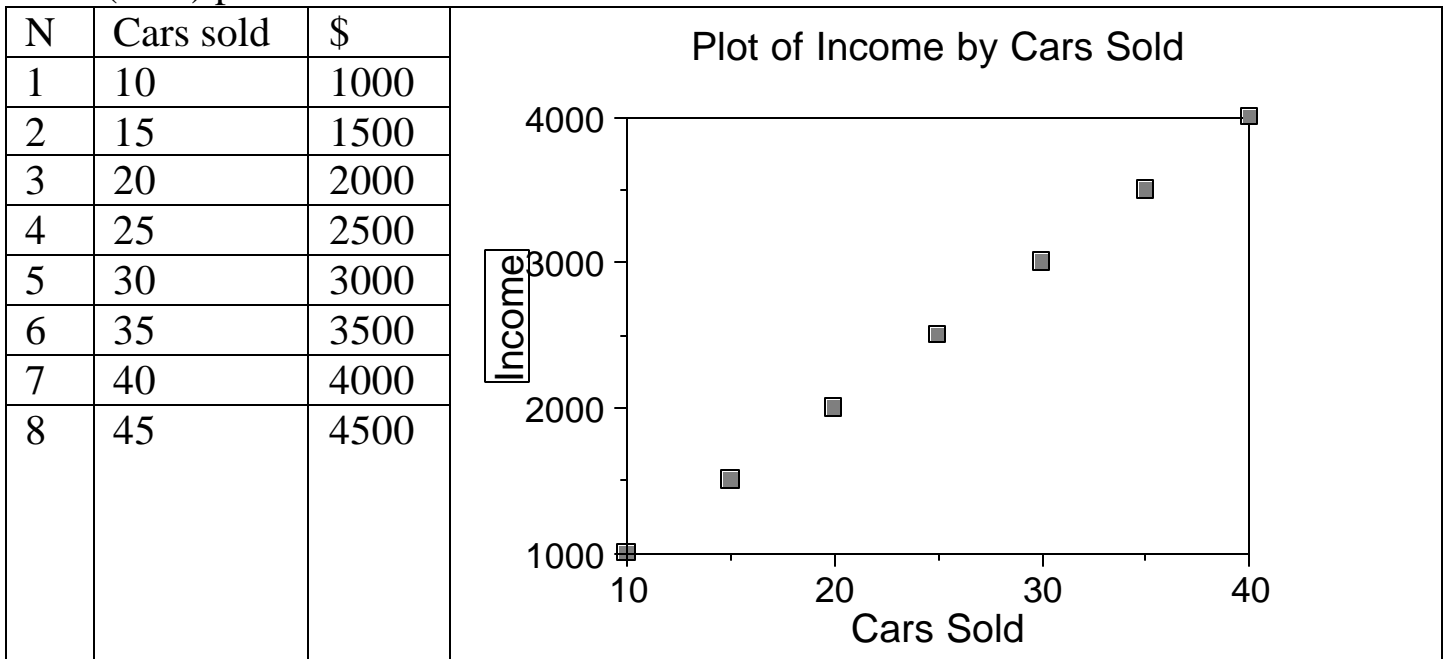
Note that as either toe size or SAT increases, the other variable stays the same on average.

- The sign of the correlation indicates the direction of the relationship
- The absolute value of the correlation indicates the strength of the relationship

Some other examples of positive, negative, and zero correlations:

Variable X	Variable Y	Correlation
Salary	Taxes paid	Positive
Shyness	N of people greeted at party	Negative
Price of car	Prestige of car	Positive
Price of tennis shoe	Foot support	Zero
Time of use of flashlight	Battery life	Negative
Weight in lbs.	Average daily caloric intake	Positive
Price of quartz watch	Accuracy of time kept	Zero
Salary of sales people	Number of cars sold	Positive
Instructor knowledge of subject matter	Clarity of presentation	Zero? (I don't really know)

Perfect (1.00) positive correlation



This example shows a perfect positive correlation. The value of correlation coefficient that corresponds to the example is $r = 1.00$. For a perfect negative relationship, $r = -1.00$, and for no relations, $r = 0.00$.

The conceptual (definitional) formula of the correlation coefficient is:

$$r = \frac{\sum z_x z_y}{N}$$

where z_x is X in z-score form, z_y is Y in z-score form, And Σ and N have their customary meaning. This says that r is the average cross-product of z-scores.

Example: Height and Weight Revisited

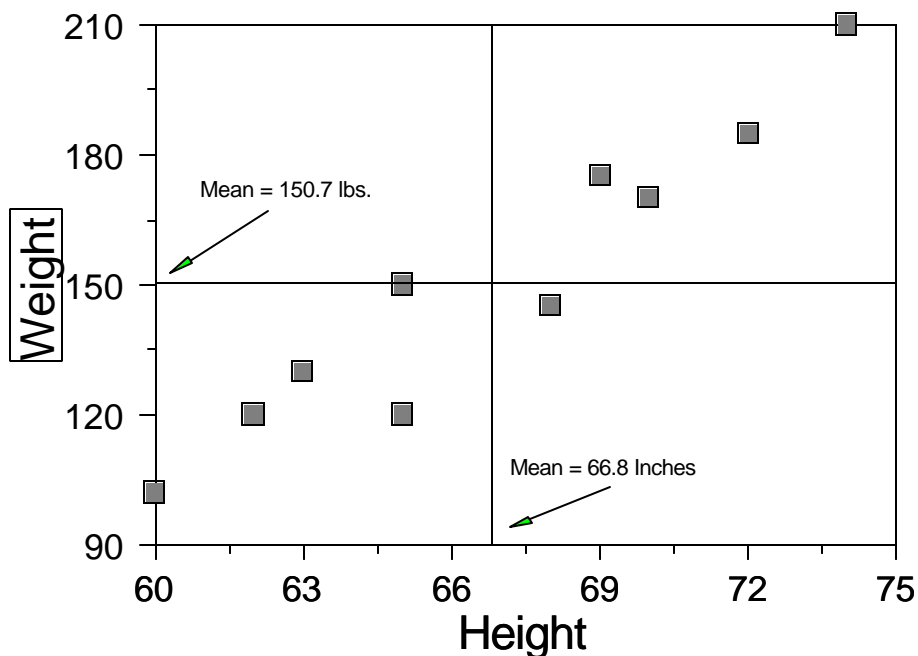
N	Ht	Wt	Z_{ht}	Z_{wt}	$Z_h * Z_w$
1	60	102	-1.58	-1.51	2.39
2	62	120	-1.11	-0.95	1.06
3	63	130	-0.88	-0.64	0.57
4	65	150	-0.42	-0.02	0.01
5	65	120	-0.42	-0.95	0.40
6	68	145	0.28	-0.18	-0.05
7	69	175	0.51	0.75	0.39
8	70	170	0.74	0.60	0.45
9	72	185	1.21	1.06	1.29

Points to notice: The mean height is 66.8 inches, the sample SD (divided by N rather than N-1) is 4.31 inches. The first height is 60 inches, which is 1.58 standard deviations below the mean, or a z-score of -1.58. The first weight is 102 pounds, which is 1.51 standard deviations below the mean weight $z = (102-150.7)/32.2 = -1.51$. The product of the two z-scores is 2.39 $(-1.58 \times -1.51 = 2.39)$. If we average the products, we get .96, which is the correlation coefficient.

- Why does the correlation coefficient have a maximum of 1, and a min of -1? Why is the correlation positive when both increase together?

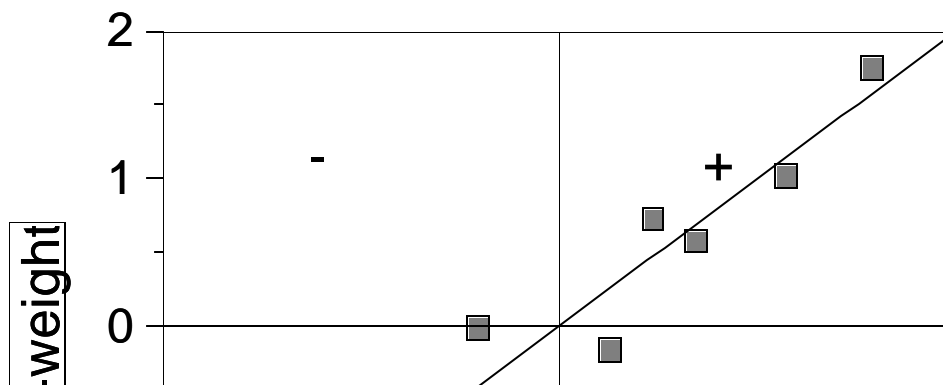
Let's look at graphs of height and weight. First in raw scores:

Plot of Weight by Height



Now in z-scores

Plot of Weight by Height in Z-scores



Points to notice:

1. The pattern or plot of scores is the same in the new and original graphs, but the new mean of height and weight are now zero, and of course the new units are standard deviations.
2. Each score is represented by a point, and the points fall into four quadrants defined by the two means. Points in the upper right quadrant are above the mean on both x and y. Points in the lower right are above the mean on x and below on y. Points in the upper left are below the mean on x and above the mean on y. Points in the lower left are below the mean on both x and y.
3. The products of the z-scores will be positive if both z_x and z_y are positive or if both z_x and z_y are negative. Positive cross-products will correspond to points found in the upper right and lower left quadrants. Products will be negative if one or the other, but not both, x and y are negative. Negative cross products correspond to points in the upper left and lower right quadrants.
4. The correlation coefficient is the average cross-product of z-scores. If most of the points are in the positive quadrants, the correlation will be positive. If most of the points are in the negative quadrants, the correlation coefficient will be negative. If the points are equally spaced in all four of the quadrants, the correlation coefficient will be zero.
5. The maximum cross-products occur when $x=y$, that is, when the points fall on the 45 degree straight line passing through 2 quadrants (either both positive or both negative). This happens because then $z_x * z_y = z_x^2$. For example, if z_x is 1, the maximum product obtains when z_y is 1 because $1 * 1 = 1$. Any other value less than 1 results in a smaller product, e.g., $1 * .9 = .9$. Of course, if z_y is greater than 1, then the product is larger than 1, e.g., $1 * 2 = 2$, but this is smaller than if z_x had been 2, the same value as z_y ($2 * 2$ is 4, which is greater than $1 * 2 = 2$).

The maximum value of the correlation coefficient (+ or - 1) occurs when the values of x and y fall on a straight line, that is when the co-relation is perfect.

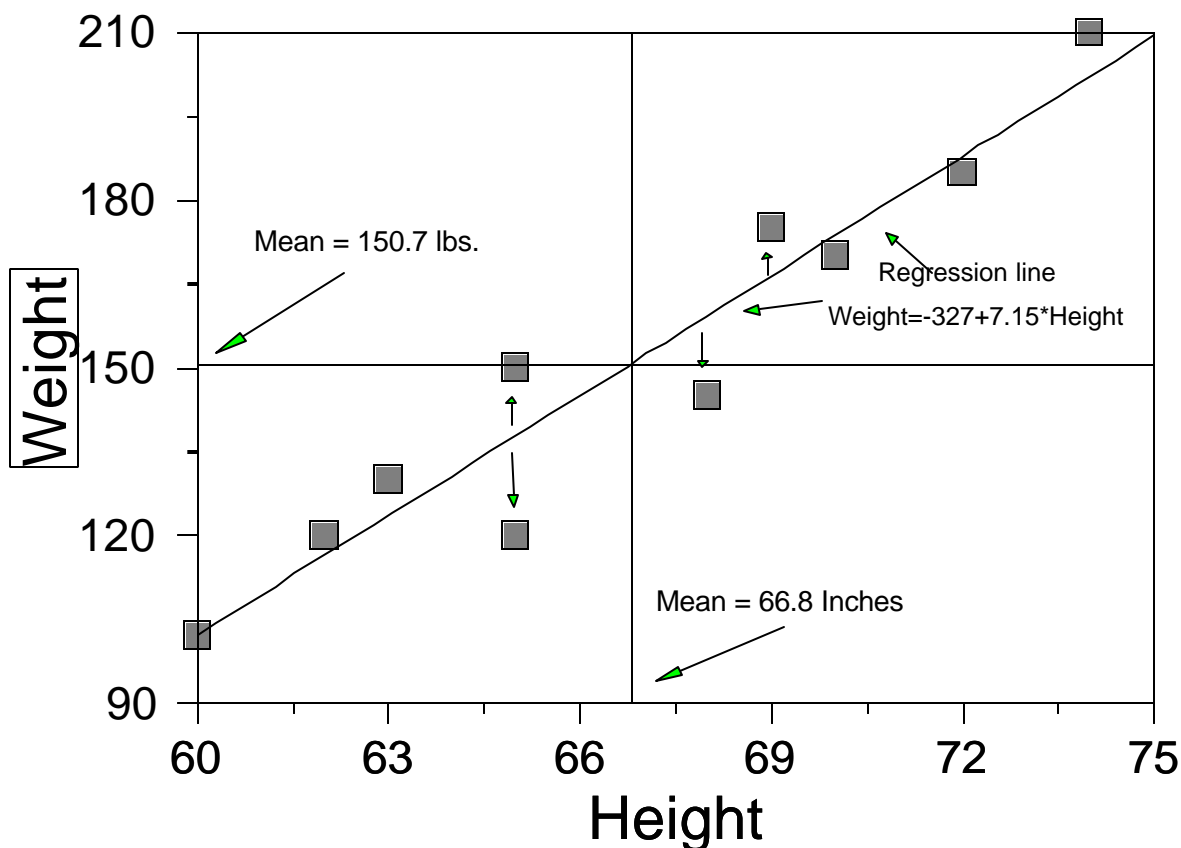
Regression

Correlation and regression are closely related mathematically and in use.

- We use the correlation coefficient to summarize the relations between 2 variables.
- Regression is used to predict values of one variable from the values of another variable. If we draw a line to predict the values of Y from the values of X, we have a regression line. For example,

Plot of Weight by Height

Second Title



- Note that the line passes through the means of both height and weight.
- The line passes close to the points (height-weight pairs), but it does not and cannot touch all of them unless the correlation is perfect. The line is called a regression line. Like any line, it can be described by an equation. You might recall from geometry or algebra the equation

$Y = mx + b$, which describes a line.

- The standard formula used in linear regression is

$Y = a + bX$, Where

X and Y are the independent and dependent variables (X is height and Y is weight in our example), a is an intercept, and b is a slope (also known as the regression weight).

The slope tells us how much Y changes when X changes 1 unit. In our example, line has a slope of 7.15. This means that as height increases by 1 inch, weight increases by 7.15 pounds.

The intercept (or Y intercept) is the value where the line crosses the Y-axis (where x is zero). It is found by

$$a = \bar{Y} - b\bar{X}$$

The actual result is not very meaningful in our example. It says essentially that if a you were zero inches tall, we expect you to weigh -327 lbs. The intercept has the function of moving the entire line up or down the Y-axis so that the line falls in the middle of the points.

The correlation coefficient and regression coefficient are closely related.

- The correlation coefficient is in fact the slope of the regression line if both X and Y are measured in z-scores (so that their means are zero and standard deviations are one). The correlation coefficient says how many standard score units Y changes when X changes 1 unit (absolute value from zero to 1). Recall that for our example, the correlation between height and weight was .96, so that when height in standard scores increase by 1, we can expect weight to increase .96 in standard scores. Remember that the slope indicates rise over run, and the correlation indicates standardized rise over standardized run. We use the regression coefficient for raw scores, so we need to move from standardized scores to raw scores. We can do this by multiplying the correlation coefficient by the ratio of the standard deviations of the Y and X variables, thus:

$$b = r \frac{S_y}{S_x}$$

in our example, the regression weight can be found by

$$b = .96*(33.95/4.54) = .96*7.45 = 7.15.$$

To use regression to make predictions, we just plug numbers into the equation or use the graph. For example, If someone were 68 inches tall, we would predict their weight to be $68*7.15-327$ or 159.2

Compare to graph.

Both correlation and regression show the linear relations between 2 variables.

Correlation shows this information in standard scores. regression shows it in raw scores.