

A QUICK INTRODUCTION TO SOME STATISTICAL CONCEPTS

Mean (average)	page 2
Variance	page 7
Standard deviation	page 11
Bivariate data	page 13
Covariance	page 14
Correlation	page 19
Linear regression	page 21
Residuals	page 24
Additional formulas	page 25
Regression effect	page 27

This document was prepared by Professor Gary Simon with the advice (and consent) of Professor William Silber, Stern School of Business. If you have comments or suggestions, please send them to either gsimon@stern.nyu.edu or silber@stern.nyu.edu

Release date 15 JULY 2002

THE MEAN, OR AVERAGE

This document will introduce a number of statistical concepts, and perhaps some of these may be very new to you. Statistical topics can be confusing because identical subject matter can be described in very different terms.

The very same statistical concept can be described in several ways:

1. Data, as numbers.
2. Data, represented algebraically.
- 2*m*. Data, represented algebraically, allowing multiple values.
3. Conceptually as random variables with probabilities.

A list of data, as in points 1 through 3, might be described as a *variable*. Each column of a data spreadsheet could be called a variable. In item 3, we do not necessarily have data, and we conceptualize the idea as a *random* variable.

Let's illustrate the notions first for the concept of the *average* of a set of values. The *average* is also called the *mean*.

1. Consider the list of values 48, 46, 54, 51, 53. The average (or mean) of this list is found as $\frac{48 + 46 + 54 + 51 + 53}{5} = \frac{252}{5} = 50.4$.
2. Let n be the number of items in a list. Represent the list as x_1, x_2, \dots, x_n . The three dots simply indicate that we're omitting some of the values.

Note that we're using x as the symbol for the list items. You'll frequently see x_i as the generic i^{th} element of the list. You can think of the symbol i as a counter or as an index. Some people will describe the list as $\{x_i; i = 1, \dots, n\}$. You can read this as "x-sub-i, as i runs from 1 to n ."

The average of this list is $\frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n}(x_1 + x_2 + \dots + x_n)$.

Since we'll be adding lists of numbers rather frequently, it helps to create a simple notation for this concept. We use the summation notation $\sum_{i=1}^n x_i$ to represent the sum of the x 's, using the index i to enumerate from the starting value $i = 1$ to the ending value $i = n$.

Then we have $\sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_n$ and the average can be

written as $\frac{1}{n} \sum_{i=1}^n x_i = \frac{\sum_{i=1}^n x_i}{n}$. The symbol i is nothing but a

counting convenience. You should note that $\sum_{i=1}^n x_i =$

$$\sum_{j=1}^n x_j = \sum_{u=1}^n x_u .$$

In nearly every case you'll encounter, the entire list of n values will be added, and it's burdensome to keep the notation above and below the Σ sign. You can then use $\sum x_i$ as a simpler notation for $x_1 + x_2 + \dots + x_n$. Again, the symbol i is a mere counting convenience, and so $\sum x_i = \sum x_j = \sum x_u$.

The symbol \bar{x} , which we read as "x bar," is the most common notation for the average of the x 's. Thus $\bar{x} = \frac{1}{n} \sum x_i$.

- 2m. It can happen that the list of value x_1, x_2, \dots, x_n will involve duplications. Suppose that there are k different values and that we name them as v_1, v_2, \dots, v_k . Let's say that v_1 occurs n_1 times, v_2 occurs n_2 times, and so on. The data could then be reorganized to look like this:

Value	Multiplicity
v_1	n_1
v_2	n_2
v_3	n_3
.	.
.	.
v_k	n_k
TOTAL	n

Now $\bar{x} = \frac{1}{n} \sum n_i v_i$. You will also see this as $\bar{x} = \frac{\sum n_i v_i}{\sum n_i}$.

The formulas in item 2 above are still correct, even if the list involves duplications.

3. There are times in which we consider problems hypothetically, rather than with numbers (as in item 1) or with algebra symbols (as in items 2 and 2m). In the hypothetical form, we'll consider X as the phenomenon under discussion, and we'll give X the technical name *random variable*. In this style of thinking, X is endowed with randomness. We should write

X in upper case. We may have no data yet, but we can still discuss the possible values for X . Let's suppose that x_i is a possible value, and that its associated probability is p_i . In such a case, the average of X is $\sum p_i x_i$. When dealing with random variables (rather than with numbers or with algebra symbols), the mean is generally denoted μ or perhaps μ_X . We also say that μ is the *expected value* of X .

The table below concerns the numbers of employees of an office calling in sick on any particular morning.

$x :$	0	1	2	3	4
$p :$	0.40	0.30	0.20	0.08	0.02

The probabilities are likely based on past observations, but they could as well be someone's hypothetical conjectures. We can apply these probabilities to *tomorrow's* sick calls, for which we certainly do not yet have data. We can use X to represent the phenomenon, and in this case we'd calculate the expected value of X as

$$\begin{aligned} \mu &= 0.40 \times 0 + 0.30 \times 1 + 0.20 \times 2 + 0.08 \times 3 + 0.02 \times 4 \\ &= 0 + 0.30 + 0.40 + 0.24 + 0.08 \\ &= 1.02 \end{aligned}$$

That is, we expect 1.02 employees to call in sick tomorrow.

Below, bordered by ***, is a technical note which you can skip at first reading.

 This description works perfectly for random variables whose possible values can be identified and isolated; such random variables are called *discrete*. Other random variables are obtained by measurement processes and have uncountably many values. These random variables are called *continuous* and a special mathematical framework is needed to deal with them.

We will not discuss every single idea in all four forms noted above. However you should be aware that it's possible to describe statistical notions in different styles.

EXAMPLE:

Over ten business days, the number of shares traded of Miraco were these:

400 200 500 300 800 400 700 200 600 200

Miraco is, of course, a very small company. The average number of shares per day is $\bar{x} = 430$. This uses $400 + 200 + \dots + 200 = 4,300$.

There are some duplications of values in this list. The value 200 occurs three times an 400 occurs twice. You can use these duplications to rearrange your computation if you wish. This rearranged form would be

$$3 \times 200 + 300 + 2 \times 400 + 500 + 600 + 700 + 800 = 4,300$$

For this situation, it's simple enough to add the original list of ten values without searching for duplicates.

EXAMPLE:

The number of medical emergency calls coming per day to the Eastside Ambulance Service is random, and we know (either from past experience or as a hypothetical suggestion) that

- The probability of 0 calls is 0.10.
- The probability of 1 call is 0.15.
- The probability of 2 calls is 0.25.
- The probability of 3 calls is 0.35.
- The probability of 4 calls is 0.10.
- The probability of 5 calls is 0.05.

The mean number of calls per day is

$$\begin{aligned} & 0.10 \times 0 + 0.15 \times 1 + 0.25 \times 2 + 0.35 \times 3 + 0.10 \times 4 + 0.05 \times 5 \\ & = 0.00 + 0.15 + 0.50 + 1.05 + 0.40 + 0.25 = 2.35 \end{aligned}$$

We can conceptualize the number of calls on any single day as a random variable X . This calculation shows that $\mu_X = 2.35$.

EXAMPLE:

A sample was taken of 20 suburban families, and each sampled family was asked how many cars it owned. The data were these:

2 2 1 2 3 2 2 1 2 2 2 1 2 2 1 0 2 2 1 2

You can get \bar{x} by simply adding these numbers and then dividing by 20. However, the work is easier if you organize it like so:

- 1 family owned no cars.
- 5 families owned one car.
- 13 families owned two cars.
- 1 family owned three cars.

The total value is then

$$1 \times 0 + 5 \times 1 + 13 \times 2 + 1 \times 3 = 0 + 5 + 26 + 3 = 34$$

$$\text{Then } \bar{x} = \frac{34}{20} = 1.7.$$

THE VARIANCE

Another useful statistical summary is the variance. We will use the variance as an intermediary to get to the calculation known as the standard deviation.

The standard deviation is simply the square root of the variance, so the connection between the variance and the standard deviation is very simple.

If you are dealing with the variance concept for the first time, you'll be distressed by the fact that there are several different formulas that lead to the same result.

If you are using a computer to do your arithmetic, you will not generally be concerned about the details. We strongly recommend the use of a computer, because arithmetic with a calculator is error-prone. On the other hand, just to make sure the computer is doing it right, as well as to make sure you understand the formula, you should do a sample calculation by hand.

With a modest amount of data, the variance can be calculated without a computer. However, the choice of formula requires a judgment call based on the appearance of the information. A good choice will lead to the answer more quickly, and with lower probability of error, than a bad computational choice.

In formula terms, we can describe the process of finding the variance of the list x_1, x_2, \dots, x_n .

1. Start by finding \bar{x} , the average.
2. Find the n differences $x_1 - \bar{x}$, $x_2 - \bar{x}$, $x_3 - \bar{x}$, \dots , $x_n - \bar{x}$. The differences are also called deviations. As a computational check, this list must sum to zero.
3. Square these n differences to produce the values $(x_1 - \bar{x})^2$, $(x_2 - \bar{x})^2$, $(x_3 - \bar{x})^2$, \dots , $(x_n - \bar{x})^2$.
4. Sum these values to get $(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \dots + (x_n - \bar{x})^2$.
5. Divide this sum by $n - 1$, the sample size, less 1.

The usual symbol for the variance is s^2 when the computation starts from data. (When we work from random variables, the symbol is σ^2 .) The variance can be done through the formula which summarizes the steps above.

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

It's a little puzzling that this is divided by $n - 1$, rather than by n . There is a perfectly good explanation for this choice, but we're not yet ready for it.

EXAMPLE:

Consider the list of values 48, 46, 54, 51, 53. This was considered previously, and we found the average $\bar{x} = 50.4$. The list of differences from the average is this:

$$-2.4, -4.4, 3.6, 0.6, 2.6$$

The value -2.4 was obtained as $48 - 50.4$. This list of five values sums to zero, as it must.

Next we square the values to produce the following:

$$5.76, 19.36, 12.96, 0.36, 6.76$$

The value 5.76 was obtained as $(-2.4)^2 = (-2.4) \times (-2.4)$.

Next we sum the list of squares to get the total 45.20. Finally, we compute $s^2 = \frac{45.20}{5-1} = \frac{45.20}{4} = 11.30$.

You might find it easier to lay out the arithmetic in a table.

Values x_i	Deviations $x_i - \bar{x}$	Squares of Deviations
48	-2.4	5.76
46	-4.4	19.63
54	3.6	12.96
51	0.6	0.36
53	2.6	6.76
TOTAL	0.0	45.20

This example should make it clear why we recommend that variances be calculated with a computer!

There are other computational formulas for the variance, but the technique given here will suffice for hand calculations with short lists.

Variances can also be found for random variables with probabilities. For example, if the value x_i has associated probability p_i , then the expected value is $\sum p_i x_i$, and it is usually denoted by μ . This was discussed above. The variance would then be defined as

$$\sigma^2 = \sum p_i (x_i - \mu)^2 = \sum p_i x_i^2 - \mu^2$$

You might note that a different symbol, σ^2 rather than s^2 , is used for this situation.

Also, the forms $\sum p_i (x_i - \mu)^2$ and $\sum p_i x_i^2 - \mu^2$ represent two different computational strategies.

As an example of the variance of a random variable, consider this situation regarding values for the price at which a certain home will sell. The prices are in thousands of dollars.

Price	Probability
200	0.20
225	0.40
250	0.30
275	0.10

The expected value here is

$$\begin{aligned} \mu &= 0.20 \times 200 + 0.40 \times 225 + 0.30 \times 250 + 0.10 \times 275 \\ &= 40 + 90 + 75 + 27.5 = 232.5 \end{aligned}$$

This example illustrates well what we mean by a random variable. There is no data (yet). The home will sell only one time, and the entire process will come down to a single number. Before the home sells, however, we conceptualize the random process which creates the price. This has been done by suggesting possible prices and corresponding probabilities. (This is of course a *hypothetical* probability mechanism, perhaps based on a suggestion from a real estate expert.) The random variable will get some symbol, likely X , and we making statements of the form “the probability that X will take the value 200 is 0.20.” The mean of this random variable is $\mu = 232.5$. This is not among the listed possible values (200, 225, 250, 275), but the statistical analyst is not concerned.

We can next find the variance of the random variable X though this calculation:

$$\begin{aligned} \sigma^2 &= 0.20 \times (200-232.5)^2 + 0.40 \times (225-232.5)^2 \\ &\quad + 0.30 \times (250-232.5)^2 + 0.10 \times (275-232.5)^2 \\ &= 211.25 + 22.5 \\ &\quad + 91.875 + 180.625 \\ &= 506.25 \end{aligned}$$

We could also work through $\sum p_i x_i^2$, finding

$$\begin{aligned}\sum p_i x_i^2 &= 0.20 \times 200^2 + 0.40 \times 225^2 + 0.30 \times 250^2 + 0.10 \times 275^2 \\ &= 8,000 + 20,250 + 18,750 + 7,562.50 \\ &= 54,562.50\end{aligned}$$

This would allow us to compute

$$\sigma^2 = \sum p_i x_i^2 - \mu^2 = 54,562.50 - 232.5^2 = 54,562.50 - 54,056.25 = 506.25$$

The formulas $\sum p_i (x_i - \mu)^2$ and $\sum p_i x_i^2 - \mu^2$ will produce the same answer.

THE STANDARD DEVIATION

If the data values have units of measurements, perhaps dollars, then the variance will have units that are the *squares* of the units of the data. Thus, if x_1, x_2, \dots, x_n gives a list of values in dollars, then the variance s^2 will have units of dollars². This would be read as “dollars, squared” or as “square dollars.” This is all very reasonable from a mathematical or physical point of view, but most users are not comfortable with the concept of square dollars. Accordingly, it is common to take the square root to convert back to the original units.

The standard deviation is the square root of the variance, and it will have the same units of measurements as the original values. Thus, the standard deviation of a list of dollar values will also be in dollar units.

If you’ve calculated a variance, then you get the standard deviation simply as the square root. There are no other added computational complexities associated with the standard deviation. For data the standard deviation is usually written as s , and for random variables the standard deviation is usually written as σ .

It’s fairly easy to make quick judgments about standard deviations because of the following *empirical rule* that frequently applied to observed data. We’ll state it twice, once for data and once for random variables.

If x_1, x_2, \dots, x_n is a set of data values (such as annual rates of return on n mutual funds), and if \bar{x} is the average and s the standard deviation, then

About $\frac{2}{3}$ of the values are in the interval from $\bar{x} - s$ to $\bar{x} + s$.

About 95% of the values are in the interval from $\bar{x} - 2s$ to $\bar{x} + 2s$.

If X denotes a random variable with mean μ and with standard deviation σ , then

The probability is about $\frac{2}{3}$ that X will take a value in the interval from $\mu - \sigma$ to $\mu + \sigma$.

The probability is about 0.95 that X will take a value in the interval from $\mu - 2\sigma$ to $\mu + 2\sigma$.

EXAMPLE:

Consider all American males between the ages of 21 and 30. What would be the standard deviation of their weights?

We certainly cannot answer this question precisely without data, but it's easy to make a plausible guess at the standard deviation s . Perhaps these men have an average weight of 165 pounds. It would be believable that about $\frac{2}{3}$ of these men have weights between $165 - 25 = 140$ pounds and $165 + 25 = 190$ pounds, so that 25 pounds is a plausible guess at the standard deviation. We'd also be reasonably happy with the statement that about 95% have weights between $165 - 50 = 115$ pounds and $165 + 50 = 215$ pounds.

EXAMPLE:

Consider data on daily orders for bagged refined flour from Carlborg Mills in Brainerd, Minnesota. What would be a reasonable value for the standard deviation of the daily amounts ordered?

The ability to guess standard deviations depends on some familiarity with the concept. It's easy to formulate a guess for the weights of American males. Since we have no information about this flour mill and no experience with flour data, we should probably not make a guess here.

BIVARIATE DATA

The mean (or average) and the standard deviation are very common statistical summaries. These two simple quantities can tell us quite a lot about a set of data or about a random variable. The notions get even more interesting when we describe two sets of data (or two random variables) at the same time. The word *bivariate* suggests that we are dealing with two variables.

Consider this set of data on ten recently-sold homes:

Area	Price
1800	182400
1362	172800
1819	190000
1594	167600
1605	192500
2741	243800
2190	184400
2393	230500
1654	171600
2209	223100

All homes were located in the same suburban subdivision. The first column gives the indoor area in square feet, and the second column gives the price in dollars.

Spreadsheets are now the preferred way to exhibit data like this. You might however also see this in coordinate form:

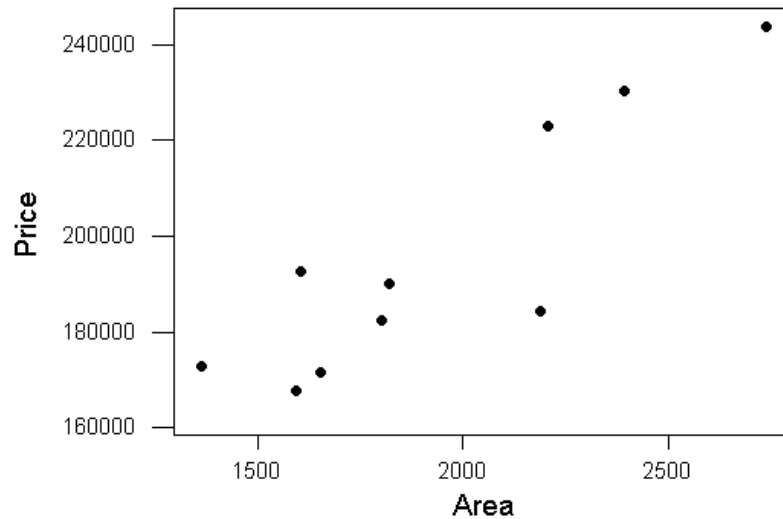
(1800, 182400) , (1362, 172800) , (1819, 190000) , (1594, 167600) ,
(1605, 192500) , (2741, 243800) , (2190, 184400) , (2393, 230500) ,
(1654, 171600) , (2209, 223100)

We can find the following information:

Variable	Average	Standard deviation
Area	1,937	430
Price	195,870	26,910

This small table gives us quite a good impression about the data. We might, however, like to see something that tells us how the variables act *together*. Do the larger homes have higher prices?

A simple first step consists of making a scatterplot:



This picture has ten points, one for each of the homes. Yes, there is definitely the impression that the larger homes have higher prices.

COVARIANCE

There are several quantities that quantify the relationship between the two variables. Let's use y as a symbol for Price and let's use x as a symbol for Area. The quantity known as the *sample covariance* between x and y is given as

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

For the data on home prices and floor areas, this calculation is structured as

$$s_{xy} = \frac{1}{10-1} \{ \begin{aligned} &(1,800-1,937)(182,400-195,870) \\ &+ (1,362-1,937)(172,800-195,870) \\ &+ \dots \\ &+ (2,209-1,937)(223,100-195,870) \end{aligned} \}$$

$$= \frac{1}{9} \{ \begin{array}{l} (-137) \times (-13,470) \\ + (-575) \times (-23,070) \\ + \dots \\ + (272) \times (27,230) \end{array} \}$$

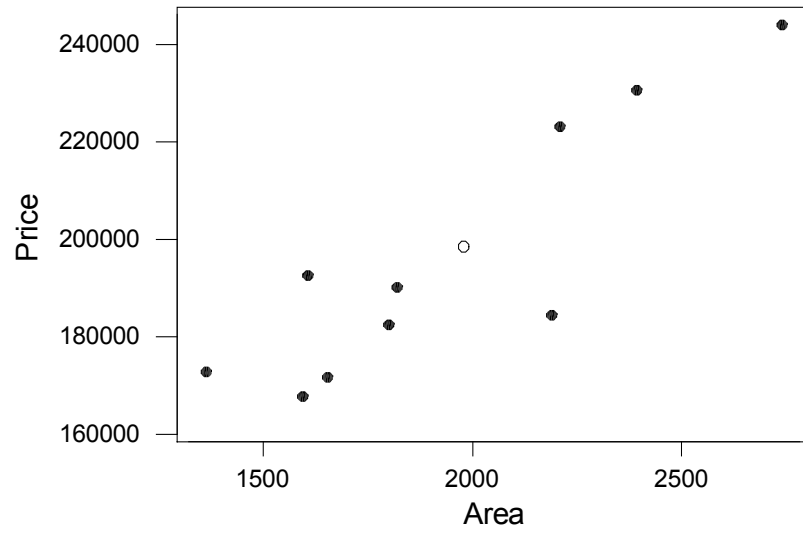
Although you should do a sample calculation by hand to reinforce your understanding of the concept, actual calculations of this type should certainly be done by computer! The value above is 10,257,646. This calculation has the form

$$s_{xy} = \frac{1}{\text{sample size} - 1} \sum (x\text{-difference from mean})(y\text{-difference from mean})$$

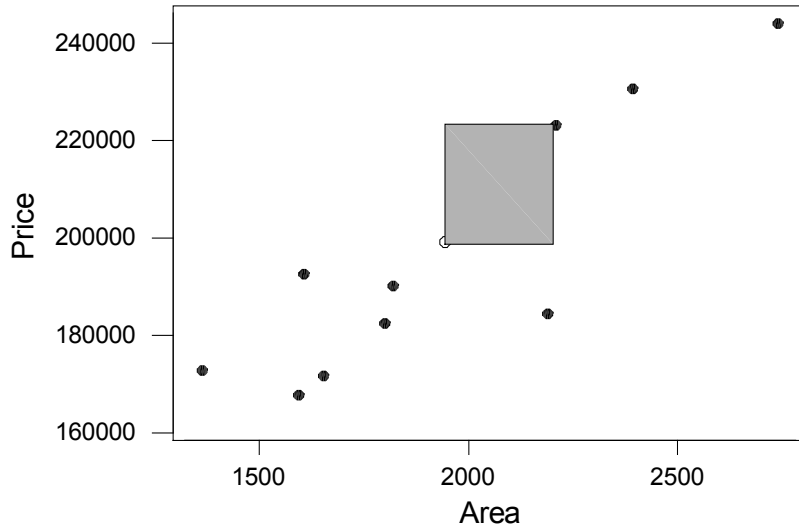
so that the units must come from the products. Since x is measured in ft^2 (square feet) and y is measured in dollars, the covariance will have units of $\text{ft}^2\text{-\$}$. We cannot even provide good advice about how to pronounce such a thing, but you could try “square-foot dollars” or “dollars-square-foot” or maybe “foot-foot-dollars.” The units are annoying, and fortunately there is something we will do about this – soon.

Each summand of the covariance s_{xy} is a product of two factors. Either or both of the factors can be negative, so some of the summands are positive, and some are negative. A picture can show what this means.

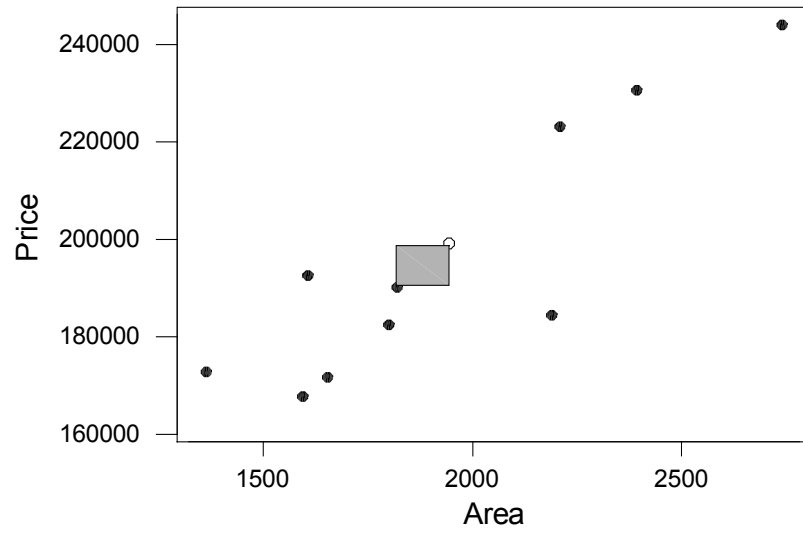
First of all, let's note the point of averages (1937, 195870) and put it on our plot. It's shown here as the open circle.



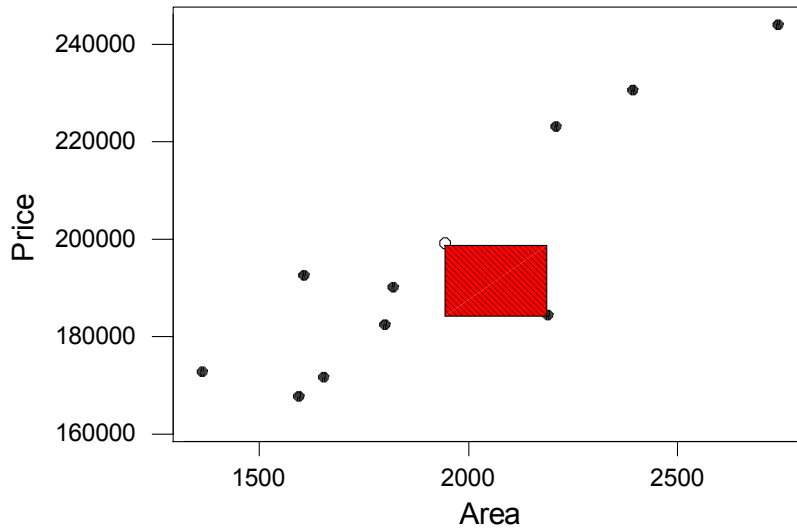
The contribution to the covariance of any single data point is the area of a rectangle with one corner at the point of averages and the diagonally opposite corner at the data point. The illustration here is of a *positive* contribution to the covariance, as both the area and the price are above average.



Here is another positive contribution to the covariance, as both the area and the price are *below* average:



Contributions to the covariance are negative when one of the variables is above average while the other is below average. For the home shown here, the area is above average, but the price is below average.



CORRELATION

Data analysts frequently need to deal with the relationship between two variables. In the example above, the covariance served this purpose, but it had units that are hard to interpret. We can make this much easier to understand by scaling the covariance by dividing by the standard deviations of the two variables. The resulting calculation is called the *correlation*, or sometimes the *correlation coefficient*. The data-based correlation is noted as r , and it's calculated as

$$r = \frac{s_{xy}}{s_x \times s_y} = \frac{\text{Covariance of } x \text{ and } y}{(\text{Standard deviation of } x) \times (\text{Standard deviation of } y)}$$

When a correlation is used to describe two random variables (as opposed to its calculation from data), it is denoted as ρ , the Greek letter rho.

For our data on house sizes and prices, this is

$$r = \frac{10,257,646}{26,910 \times 430} \approx 0.8865$$

The calculation of r must produce a value between -1 and +1; values outside this interval are arithmetic errors.

With some algebraic work, we can show that

$$r = \frac{1}{n-1} \sum \left(\frac{x-\bar{x}}{s_x} \right) \left(\frac{y-\bar{y}}{s_y} \right)$$

so that r is (almost) an average of the product of scaled differences from the means.

These correspond to the scaled areas of the rectangles associated with the covariances.

Correlations that are close to +1 represent strong positive relationships. The correlation found above as 0.8865 tells us that large homes tend to have high prices, and small homes tend to have low prices.

A correlation will be +1 exactly if and only if the data points all lie on a line with positive slope. Positive slope refers to a line that rises from the left end of the scatter plot to the right end. This can happen only with a perfect accounting relationship between x and y , and it thus rarely encountered with real data.

Correlations that are close to -1 represent strong negative relationships. For instance, the prices of used 1998 Toyota Corollas have a strong negative relationship with the mileage on the odometer. That is, high prices are generally found on cars with low mileage, and low prices are found on cars with high mileage.

A correlation will be -1 exactly if and only if the data points all lie on a line with negative slope, referring to a line that drops from the left end of the scatter plot to the right end. This requires a perfect accounting relationship between x and y .

Correlations that are close to zero represent weak relationships.

We've used the words "close to," "strong," and "weak" very loosely. Our feelings about correlations will vary according to the problem. Sometimes we work with real estate prices, sometimes with bond interest rates, and sometimes with corporate profits.

REGRESSION

In most statistical work with two variables, we move beyond the correlation concept to explore the related notion of *regression*. In the regression context we think of one of the variables as a possible influence on the other. The correlation concept treats the two variables symmetrically, but regression does not.

In the regression context, the variable that is doing the (potential) influencing is called the *independent variable*. The variable that is (potentially) influenced is the *dependent variable*. Thus, the independent variable is a (potential) influence on the dependent variable.

We are being very, very careful with the word “potential.” Consider a data base of manufacturing firms in which we are concerned with year 2000 R&D (research and development) spending and year 2001 profits. Certainly we will consider year 2000 R&D as the independent variable and year 2001 profits as the dependent variable, as we expect that year 2000 R&D will influence year 2001 profits. The data, however, might not support our expectations. This is exactly why we will think of year 2000 R&D as a *potential* influence.

We have many reasons for doing the regression. We will certainly ask whether the potential influence turns out to be an actual influence. We will want to quantify the strength of the relationship. A very important use of regression will be in prediction. If we decide that the relationship between year 2000 R&D and year 2001 profits is strong, we would use that relationship in predicting future profits based on prior R&D spending.

The word *cause* has been scrupulously avoided. We can describe something as a cause only if the data has been collected as part of a controlled randomized experiment. Economic and financial data are observational and not collected in an experimental framework. As much as we might like to say that spending on R&D causes future profit, we simply do not have the logical basis for such a claim.

In the scatterplot that accompanies regression work, it is customary to place the dependent variable on the vertical axis and the independent variable on the horizontal axis. In the pictures above regarding home price and floor area, the price was the vertical axis. We certainly want to think that area is a potential influence on price.

If the variable labels are x and y , it is customary to use y for the dependent variable and place it on the vertical axis.

There's an interesting exception to this, and you should be aware of it. Economists frequently use P versus Q (price versus quantity) graphs, and they place P on the vertical axis. This layout would agree with custom perhaps for things like agricultural commodities, where Q precedes, and is a potential influence on, P . In other contexts, economists consider suppliers who set P , so that price is a potential influence on Q , the quantity that sells; in such a case they have the dependent variable on the horizontal axis.

In its simplest form, regression is executed as a straight-line relationship, and we call the process *linear regression*. If the dependent variable is y and the independent variable is x , then the fitted regression line will have the form

$$y = b_0 + b_1 x$$

In this form b_0 (the intercept) and b_1 (the slope) will be numbers computed from the data. In our exploration of the real estate prices, the fitted line will be

$$\text{Price} = b_0 + b_1 \text{Area}$$

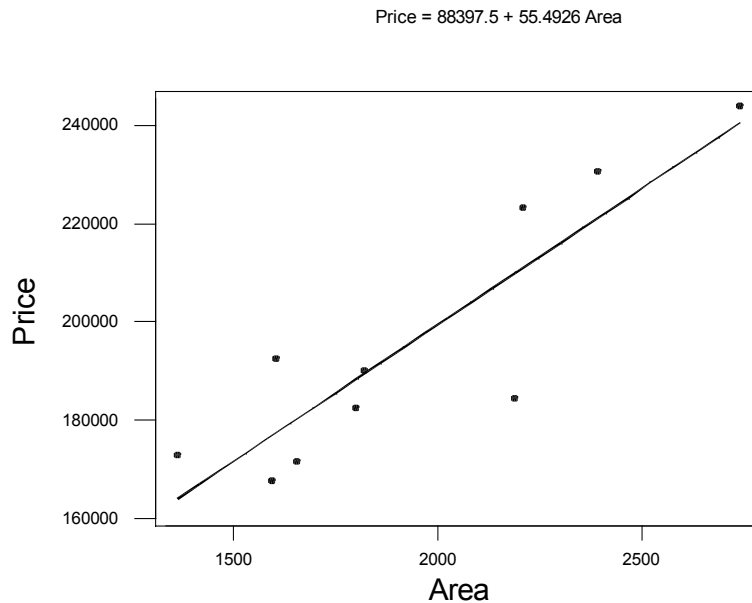
The numeric version, obtained with the help of a computer, is

$$\text{Price} = 88,397.5 + 55.4926 \text{Area}$$

You might feel that the precision is a bit pretentious, and perhaps you'd like to report

$$\text{Price} = 88,400 + 55.49 \text{Area}$$

Here is the scatterplot, this time with the fitted regression line superimposed:



We have a strong interest in the estimated regression slope, 55.49. This represents the average movement in price per unit movement in area. Specifically, this is to be interpreted as \$55.49 per square foot of area. This is the estimated marginal effect on price of one additional square foot of area.

The estimated intercept, here 88,400, can be interpreted in some problems, but probably not here. Some might try to say that this is “the value of a house with no area,” meaning perhaps an empty building lot. However, the data base included no building lots, and we should refrain from such a speculation.

An immediate use of this regression is for prediction. If a home with 2,500 square feet of area comes onto this market, we would predict that it would sell for

$$88,400 + 55.49 \times 2,500 = 88,400 + 138,725 = 227,125$$

The home might sell for more than \$227,125 or it might sell for less. We are, after all, just making a prediction.

RESIDUALS

Suppose that we use this prediction method for a home that was part of our original data set. The home listed as the first data point had an area of 1,800 square feet. The fitted price would be

$$88,400 + 55.49 \times 1,800 = 88,400 + 99,882 = 188,282$$

We are going to call this a “fitted price” rather than a “predicted price” because this home is in our data set and we know the price at which it sold. Indeed, this price was \$182,400.

The difference between the actual observed price (here \$182,400) and the fitted price (\$188,282) is called the *residual*. For this home the residual is

$$182,400 - 188,282 = -5,882$$

Relative to this regression, this particular home sold for \$5,882 less than it should have. This is not a mistake, it is not an error, it is not the result of ineffective negotiation. The data points do not lie neatly on a perfect line, so that some homes will have positive residuals and some will have negative residuals.

The table below lists the residuals for all the homes in the data set. The values were obtained by computer; the difference between our -5,882 and the computer’s -5,884.2 is related only to round-off issues.

Area	Price	Residual
1800	182400	-5884.2
1362	172800	8821.6
1819	190000	661.5
1594	167600	-9252.7
1605	192500	15036.9
2741	243800	3297.3
2190	184400	-25526.3
2393	230500	9308.7
1654	171600	-8582.2
2209	223100	12119.4

The residuals will sum to zero; this is a consequence of the formulas used to calculate the regression. We can see that the fifth home sold for about \$15,000 more than the fitted line would suggest. The seventh home sold for about \$25,000 less than the fitted line would suggest.

The residuals are non-zero simply because the points do not all fall on a straight line. It is tempting to say that the fifth home was over-valued by the purchasers, but we should

resist this interpretation. After all, the regression work utilized only floor area, and there are many other factors determining price.

It will not surprise you to learn that regression models are used to relate stock prices to sets of independent variables. Companies with negative residuals might be described as undervalued, and therefore attractive purchases, but we have to be very careful about such judgments. After all, the regression work might be missing variables that are relevant to the values of the companies. Of course, these companies might *really* be undervalued, and certainly any company that produces a large negative residual should be examined closely!

The work above has been described as a *fitted* regression line, based on data. There is also a true regression line, involving a model equation with random variables and a number of assumptions. We will not here go into the formalism of the model equation, but it's important to realize that our work with data produces a *fitted* (or estimated) line and not the guaranteed truth.

Suppose, hypothetically, that all the residuals turned out to be zero. This would indicate that the dependent variable y could be predicted perfectly from the independent variable x . This means, of course, that there is a perfect accounting relationship between x and y and further that the correlation between x and y would be +1 or -1 (depending on the slope); see the indented comments on page XX.

Indeed, the residuals are related to the correlation coefficient. This is a complicated story, as this is the equation which relates them:

$$r^2 = 1 - \frac{\sum(\text{residual})^2}{(n-1) \text{Variance}(y)}$$

The tells us that a set of large residuals produces a correlation coefficient close to zero.

ADDITIONAL FORMULAS

The calculations to produce a fitted regression line are intense, and we recommend that computer software be used, although, once again, a sample hand calculation might improve your understanding of what is going on. There are nonetheless some interesting quantitative relationships that we can explore. Here's one:

$$b_1 = \text{estimated regression slope} = \frac{\text{Sample covariance of } x \text{ and } y}{\text{Sample variance of } x} = \frac{s_{xy}}{s_x^2}$$

$$= \frac{\frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1} \sum (x_i - \bar{x})^2} = r \frac{s_y}{s_x}$$

$$= \text{Correlation} \times \frac{\text{Standard deviation of } x}{\text{Standard deviation of } y}$$

There is a little bit of algebra work behind this. Since $r = \frac{s_{xy}}{s_x s_y}$, it

must happen that $s_{xy} = r s_x s_y$. If we substitute this into $b_1 = \frac{s_{xy}}{s_x^2}$,

$$\text{we get } b_1 = \frac{r s_x s_y}{s_x^2} = r \frac{s_y}{s_x}.$$

$$b_0 = \text{estimated regression intercept} = \bar{y} - b_1 \bar{x}$$

There's one further additional interesting way to present a regression line. The fitted line is

$$y = b_0 + b_1 x$$

We can make the substitutions $b_0 = \bar{y} - b_1 \bar{x}$ and $b_1 = r \frac{s_y}{s_x}$ to get this:

$$\frac{y - \bar{y}}{s_y} = r \frac{x - \bar{x}}{s_x}$$

In this expression, x and y represent the variables (meaning the names of the horizontal and vertical axes), while the remaining items (\bar{x} , s_x , \bar{y} , s_y , r) are quantities computed from the data.

THE REGRESSION EFFECT

Let's consider the above form of the fitted regression equation with regard to a prediction. We showed previously that a home of 2,500 square feet would be predicted to sell for \$227,125. We obtained this value as

$$88,400 + 55.49 \times 2,500 = 88,400 + 138,725 = 227,125$$

but we could also obtain it by solving for y in

$$\frac{y - 195,870}{26,910} = 0.8865 \times \frac{2,500 - 1,937}{430}$$

Observe that on the right side the fraction $\frac{2,500 - 1,937}{430} \approx 1.31$ shows that this home is 1.31 standard deviations above average in size. One would think that the predicted price would then be 1.31 standard deviation above average. However, the predicted price is only $\frac{227,125 - 195,870}{26,910} \approx 1.16$ standard deviations above average in price. Indeed, we

see that within rounding $1.16 \approx 0.8865 \times 1.31$. The correlation coefficient in this last result makes the predicted price relatively closer to average than was the floor area used to make the prediction! The floor area is somewhat high, and we predict the price to be high, *but not quite as high as the floor area*.

This particular phenomenon is known as the *regression effect*. It is a fact of statistical life, and it tends to turn up over and over. It is not always recognized.

Suppose that a man is 6' 9" tall, and suppose that he has a son. What height would you predict that this son would grow to? Most people would predict that he would be tall, but it would be quite unusual for him to be as tall as his father. Perhaps a reasonable prediction is 6' 5". We are expecting the son to "regress" back to mediocrity, meaning that we expect him to revert to average values.

This is of course a prediction on average. The son could well be even taller than his father, but this is unlikely.

This regression effect applies also at the other end of the height scale. If the father is 5' 1", then his son is likely to be short also, but *the son will probably be taller than his father*.

The regression effect is strongest when far away from the center. For fathers of average height, the probability is about 50% that the sons will be taller and about 50% that the sons will be shorter.

The regression effect is everywhere.

People found to have high blood pressure at a mass screening (say at an employer's health fair) will have, on average, lower blood pressures the next time a reading is taken.

Mutual fund managers who have exceptionally good performances in year T will be found to have not-so-great performances, on average, in year $T + 1$.

Professional athletes who have great performances in year T will have less great performances, on average, in year $T + 1$.