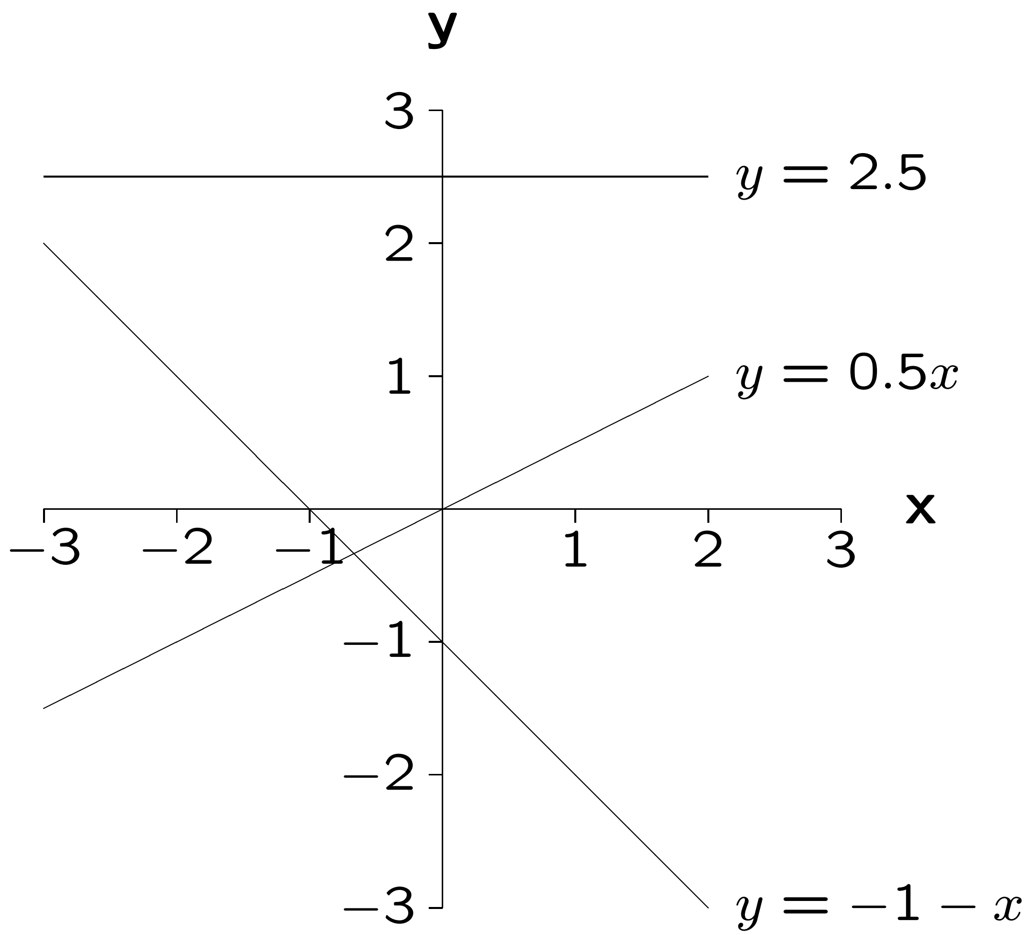
Chapter 13: The Correlation Coefficient and the Regression Line

We begin with a some useful facts about straight lines.

Recall the x, y coordinate system, as pictured below.



We say that y is a linear function of x if

$$y = a + bx, \text{ for some numbers } a \text{ and } b.$$

If $y = a + bx$ then the graph of the function is a straight line with y-intercept equal to a and slope equal to b. The line is horizontal if, and only if, $b = 0$; o.w. it 'slopes up' if $b > 0$ and slopes down if $b < 0$. The only lines not covered by the above are the vertical lines, e.g. $x = 6$. Vertical lines are not interesting in Statistics.

In math class we learn that lines extend forever. In statistical applications, as we will see, they never extend forever. This distinction is very important.

In fact, it would be more accurate to say that statisticians study line segments, not lines, but everybody says lines.

It will be very important for you to understand lines in two ways, what I call **visually** and **analytically**.

Here is what I mean. Consider the line $y = 5 + 2x$. We will want to substitute (plug-in) values for x to learn what we get for y. For example, $x = 3$. We do this analytically by substituting in the equation:
$y = 5 + 2(3) = 11$.

But we can also do this visually, by graphing the function. Walk along the x axis until we get to $x = 3$ and then climb up a rope (slide down a pole) until we hit the line. Our height when we hit the line is $y = 11$. (Draw picture on board.)

The Scatterplot

We are interested in situations in which we obtain two numbers per subject. For example, if the subjects are college students, the numbers could be:
$X =$ height and $Y =$ weight.
$X =$ score on ACT and $Y =$ first year GPA.

X = number of AP credits and Y = first year GPA.
Law schools are interested in:
X = LSAT score and Y = first year law school GPA.
and so on. In each of these examples, the Y is considered more important by the researcher and is called the **response**. The X is important b/c its value might help us understand Y better and it is called the **predictor**.

For some studies, reasonable people can disagree on which variable to call Y. Here are two examples:
–The subjects are married couples and the variables are: wife's IQ and husband's IQ.
–The subjects are identical twins and the variables are: first born's IQ and second born's IQ.

We study two big topics in Chapter 13. For the first of these, the correlation coefficient, it does not matter which variable is called Y.

For the second of these, the regression line, changing the assignment of Y and X will change the answer. Thus, if you are uncertain on the assignment, you might choose to do the regression line analysis twice, once for each assignment.

The material in Chapter 13 differs substantially from what we have done in this class. In Chapter 13, we impose fairly strict **structure** on how we view the data. This structure allows researchers to obtain very elaborate answers from a small amount of data. Perhaps surprisingly, these answers have a history of working very well in science.

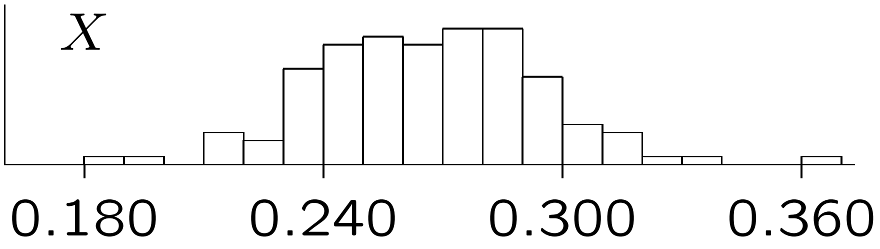
But it will be important to have a healthy skepticism about the answers we get and to examine the data carefully to decide whether the imposed structure seems reasonable.

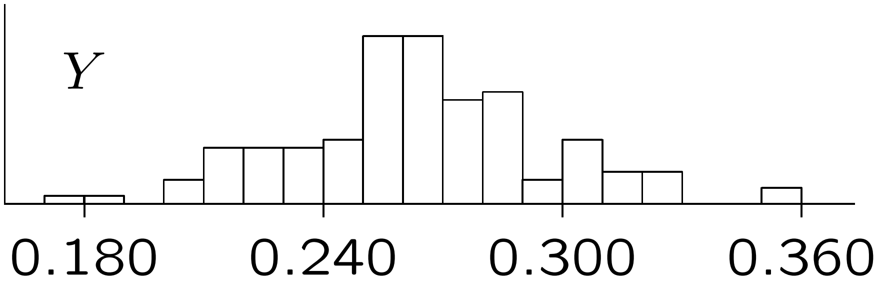
We begin with an example with $n = 124$ subjects, a very large number of subjects for these problems. As we will see, often n is 10 or smaller.

The subjects are 124 men who played major league baseball in both 1985 and 1986. This set contains every man who had at least 200 official at-bats in the American League in both years. The variables are:
$Y = 1986$ Batting Average (BA) and
$X = 1985$ BA.
The idea is that, as a baseball executive, you might be interested in learning how effectively offensive performance one year (1985) can predict offensive performance the next year (1986).

In case you are not a baseball fan, here is all you need to know about this example.
–BA is a measure of offensive performance, with larger values better.
–BA is not really an average; it is a proportion: BA equals number of hits divided by number of official at-bats.
–BA is always reported to three digits of precision and a BA of, say, 0.300 is referred to as 'hitting 300.' BTW, 300 is the threshold for good hitting and 200 is the threshold for really bad hitting.

The names and data for the 124 men are on pp. 442–3. Behaving like the MITK, we first study the variables individually, following the ideas of Chapter 12.

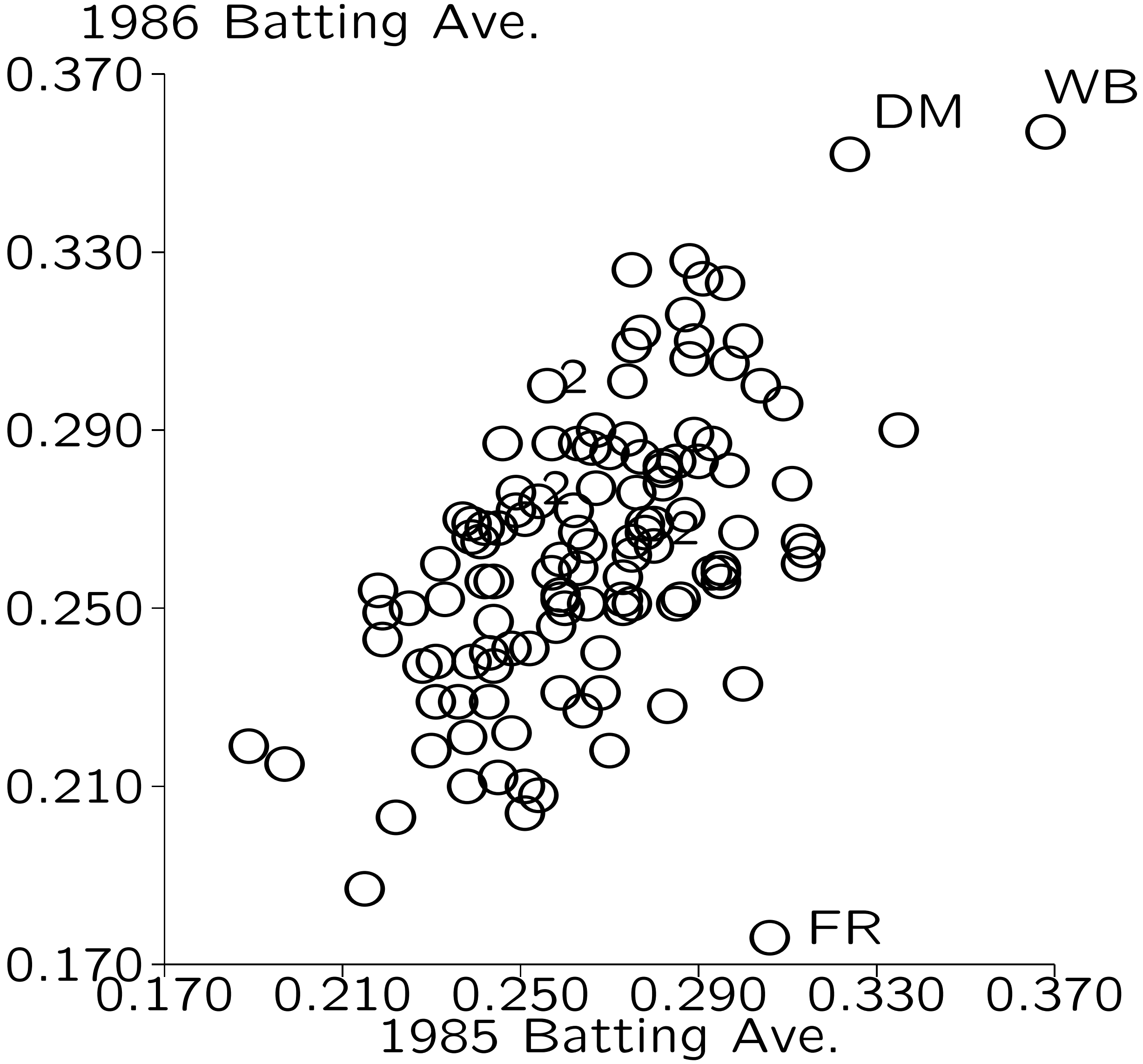




These histograms suggest small and large outliers both years. In addition, both histograms are close to symmetry and bell-shape. Also, the means and sd's changed little from X to Y.

Year	Mean	St.Dev.
1985	0.2664	0.0280
1986	0.2636	0.0320

Below is the scatterplot of these BA data. The first thing we look for are **isolated cases (IC)**. I see two, possibly three, IC identified by initials below: WB, DM and FR.



Now, ignore the outliers and look for a 'pattern' in the remaining data. For the BA data, the data describe an ellipse that is tilted upwards (lower to the left, higher to the right). This is an example of a linear relationship between X and Y; i.e. as X grows larger (sweep your eyes from left to right in the picture), the Y values tend to increase (become higher).

In Chapter 13, we limit attention to data sets that reveal a linear relationship between X and Y. If your data do not follow a linear relationship, you should not use the methods of Chapter 13. Thus, your analysis should **always** begin with a scatterplot to investigate whether a linear relationship is reasonable.

Page 447 of the text presents five hypothetical scatterplots: one reveals an increasing linear pattern; one reveals a decreasing linear pattern; and the remaining three show various curved relationships between X and Y. Thus, to reiterate; if your scatterplot is curved, do not use the methods of Chapter 13.

Page 448 of the text presents four scatterplots for data sets for small values of n (the n's are 9, 6, 12 and 13, typical sizes in practice). The subjects are spiders and the four scatterplots correspond to four categories of spiders. For each spider, Y is heart rate and X is weight.

Above each scatterplot is the numerical value of r, the **correlation coefficient** of the data. At this time, it suffices to note that $r > 0$ indicates (reflects?) an increasing linear relationship and $r < 0$ indicates a decreasing linear relationship between Y and X.

There are two important ideas revealed by these scatterplots. First, for small n it can be difficult to decide whether a case is isolated; whenever possible, use your scientific knowledge to help with this decision.

Second, especially for a small n, the presence of one or two isolated cases can drastically change our view of the data. For example, consider the $n = 9$ small hunters.

The two spiders in the lower left of the scatterplot might be labeled isolated. Including these cases, the text states that $r > 0$, but if they are deleted from the data set (which could be a deliberate action by the researcher, or perhaps these guys were stepped on during their commute to the lab) then $r < 0$. Scientists typically get very excited about whether r is positive or negative, so it is noteworthy that its sign can change so easily.

Thus far, we have been quite casual about looking at scatterplots. We say, "The pattern is linear and looks increasing (decreasing, flat)." It will remain (in this course) the job of our eyes and brain to decide on linearity, but the matter of increasing or decreasing will be usurped by the statisticians. Furthermore, using my eyes and brain, I can say that the pattern is decreasing for tarantulas and for web weavers (r agrees with me), and I can say that the linear pattern is **stronger** for the tarantulas.

The correlation coefficient agrees with me on the issue of strength and has the further benefit of quantifying the notion of stronger in a manner that is useful to scientists.

I am not very good at motivating the formula for the correlation coefficient. In addition, the end of the semester is near, so time is limited. The interested student is referred to pp. 450–3 of the text for a (partial) explanation of the formula.

Here is the briefest of presentations of the formula. Each subject has an x and a y. We standardize these values into x' and y':

$$x' = (x - \bar{x})/s_X; \quad y' = (y - \bar{y})/s_Y.$$

We then form the product $z' = x'y'$.

The idea is that $z' > 0$ provides evidence of an increasing relationship and $z' < 0$ provides evidence of a decreasing relationship.

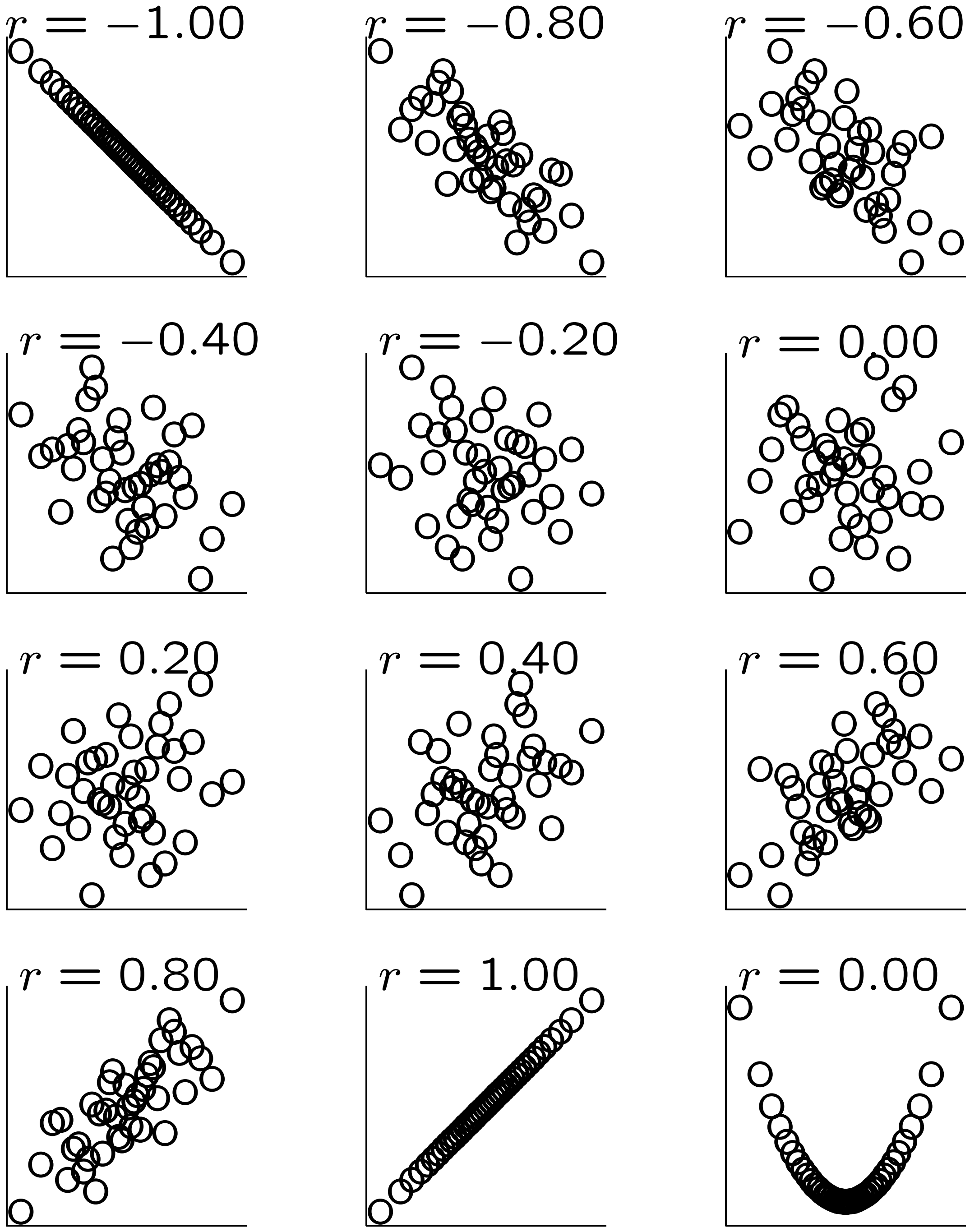
(The product is positive if both terms are positive or both are negative. Both positive means a large x is matched with a large y; both negative means a small x is matched with a small y.)

The correlation coefficient, r, combines the z''s by almost computing their mean:

$$r = \frac{\sum z'}{n-1}.$$

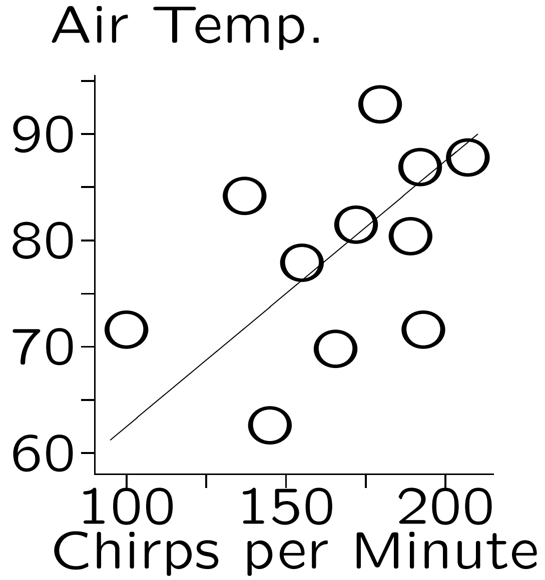
The next slide presents 12 prototypes of the correspondence between a scatterplot and its correlation coefficient.

These 12 scatterplots illustrate six important facts about correlation coefficients. These six facts appear on pages 454 and 456 of the text and will not be reprinted here.

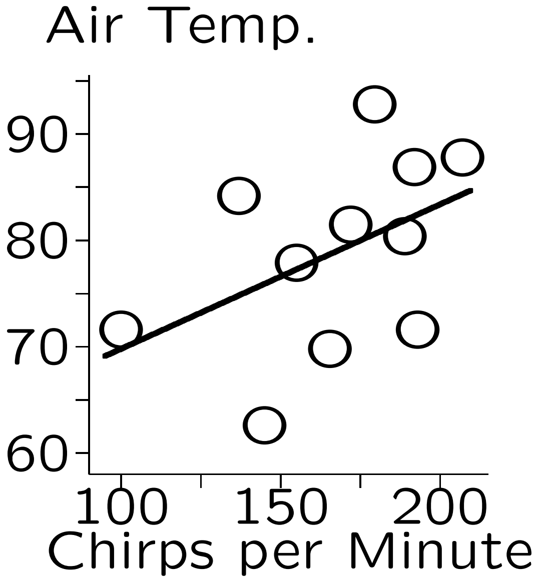

$r = -1.00$
$r = -0.80$
$r = -0.60$
$r = -0.40$
$r = -0.20$
$r = 0.00$
$r = 0.20$
$r = 0.40$
$r = 0.60$
$r = 0.80$
$r = 1.00$
$r = 0.00$

13.3: The regression line.

$\dot{y} = 37.5 + 0.25x$



$\hat{y} = 56.2 + 0.136x$



x	y	$\dot{y}$	$y - \dot{y}$	$(y - \dot{y})^2$
145.0	62.6	73.75	−11.15	124.32
172.0	81.5	80.50	1.00	1.00
155.0	77.9	76.25	1.65	2.72
137.0	84.2	71.75	12.45	155.00
179.5	92.8	82.37	10.43	108.68
192.0	86.9	85.50	1.40	1.96
207.0	87.8	89.25	−1.45	2.10
165.5	69.8	78.87	−9.07	82.36
193.0	71.6	85.75	−14.15	200.22
100.0	71.6	62.50	9.10	82.81
189.0	80.4	84.75	−4.35	18.92

$\text{SSE}(\dot{y}) = 780.10$

x	y	$\widehat{y}$	$y-\widehat{y}$	$(y-\widehat{y})^2$
145.0	62.6	75.92	−13.32	177.42
172.0	81.5	79.59	1.91	3.64
155.0	77.9	77.28	0.62	0.38
137.0	84.2	74.83	9.37	87.76
179.5	92.8	80.61	12.19	148.55
192.0	86.9	82.31	4.59	21.05
207.0	87.8	84.35	3.45	11.89
165.5	69.8	78.71	−8.91	79.35
193.0	71.6	82.45	−10.85	117.68
100.0	71.6	69.80	1.80	3.24
189.0	80.4	81.90	−1.50	2.26
			$\text{SSE}(\widehat{y}) =$	653.23

On $n = 11$ occasions, Susan Robords determined two values for different crickets: Y is the air temperature and X is the cricket's chirp rate in chirps per minute. In her campcraft class, she was told that one can 'calculate' the air temperature with the following equation:

$\dot{y} = 37.5 + 0.25x$.

Above, we have a scatterplot of Susan's data with this line. The most obvious fact is that 'calculate' was way too optimistic!

As Yogi Berra once said, "You can observe a lot by just watching." Let's follow his advice and examine the scatterplot and table above.

We see that on some occasions, $\dot{y}$ provides an accurate prediction of y. Visually, this is represented by circles that are on, touching, or nearly touching the line. But on many other occasions, the predictions are poor: the line is either far lower than the circle (the prediction is too small) or the line is far higher than the circle (the prediction is too large).

Next, we do something very strange. We change perspective and instead of saying that the prediction is too small (large) we say that the observation is too large (small). Egocentric? Yes, but there are two reasons.

First, look at the scatterplot and line again. It is easier to focus on the line and see how the points deviate from it, than it is to focus on all the points (n could be large) and see how the line deviates.

Second, we plan to compare $\dot{y}$ and y by subtraction. We could use $\dot{y} - y$ or $y - \dot{y}$. The former takes y as the 'standard' and the latter reverses the roles. For circles below the line, I want this 'error' to be a negative number; to get that I must subtract in the order $y - \dot{y}$; that is, I take the prediction as the standard and the observation 'errs' by not agreeing.

Look at the table again. The ideal for the error $y - \dot{y}$ is 0. As the error moves away from 0, in either direction, the inadequacy of the prediction becomes more and more serious. For math reasons (and often it makes sense scientifically; at least approximately) we consider an error of, say, -5 to be exactly as serious as an error of $+5$. As in Chapter 12, we might be tempted to achieve this by taking the absolute value of each error, but, again, we get much better math results by squaring the errors.

Finally, we sum all of the squared errors to obtain: $SSE(\dot{y}) = 780.10$.

Ideally, $SSE = 0$ and the larger it is, the worse the prediction.

You are probably thinking that we need to adjust SSE to account for sample size, but we won't bother with that.

Instead, we pose the following question: Can we improve on Susan's line? Or: Can we find another prediction line which has an SSE that is smaller than Susan's 780.10?

I suggest the line $\widehat{y} = 56.2 + 0.136x$. From the table, we see that $SSE(\widehat{y}) = 653.23$. Thus, according to **The Principle of Least Squares** $\widehat{y}$ is superior to $\dot{y}$.

Can we do better than my $\widehat{y}$? No.

Major Result: There is always a unique line that minimizes SSE over all possible lines. The equation of the line is given as

$$\widehat{y} = b_0 + b_1 x,$$

where $b_1 = r(s_Y/s_X)$ and $b_0 = \bar{y} - b_1\bar{x}$.

For the cricket data, for example, it can be shown that $\bar{x} = 166.8, s_X = 31.0, \bar{y} = 78.83$, $s_Y = 9.11$, and $r = 0.461$. Substituting these values into the above yields

$$b_1 = 0.461(9.11/31.0) = 0.1355, \text{ and}$$

$$b_0 = 78.83 - 0.1355(166.8) = 56.23.$$

Thus, the equation of the best prediction line is

$$\hat{y} = 56.23 + 0.1355x,$$

which I rounded in my earlier presentation of it.

The means and sd's of the BA data were given on slide 343 and it has $r = 0.554$. Thus,

$$b_1 = 0.554(0.032/0.028) = 0.633, \text{ and}$$

$$b_0 = 0.2636 - 0.633(0.2664) = 0.095.$$

Thus, the equation of the regression line is

$$\hat{y} = 0.095 + 0.633x.$$

This line appears on page 471 of the text.

We have seen that it is easy to calculate $\widehat{y}$ and it is the best line possible (based on the principle of least squares), but is it any good? (Is Sylvester Stallone's best performance any good? Is there a reason he has never done Shakespeare?)

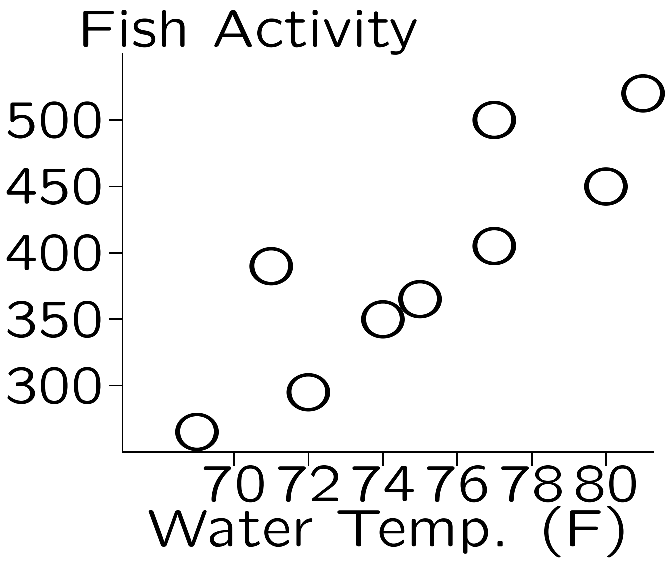
First, note that we can see why r is so important. We need five numbers to calculate $\widehat{y}$: two numbers that tell us about x only; two numbers that tell us about y only; and one number (r) that tells us how x and y relate to each other. In other words, r tells us all we need to know about the association between x and y.

We obtain the regression line by calculating two numbers: b_0 and b_1. Thus, obviously, this pair of numbers is important. Also, b_1, the slope, is important by itself; it tells us how a change in x affects our prediction $\widehat{y}$.

Unlike mathematics, however, the intercept, b_0, alone usually is **not** of interest. Now in math, the intercept is interpreted as the value of y when $x = 0$. Consider our examples. For the Cricket study, $x = 0$ gives us $\hat{y} = 56.2$. But we have no data at or near $x = 0$; thus, we really don't know what it means for x to equal 0. (Discuss.)

Similarly, for the BA study, $x = 0$ predicts a 1986 BA of 0.095. But nobody batted at or near 0.000 in 1985. In fact, I conjecture that in the history of baseball there has never been a position player with at least 200 at-bats who batted 0.000.

Consider the following scatterplot of fish activity versus water temperature for fish in an aquarium. (Should we use these data to predict fish activity for $x = 32$? For $x = 212$?)



The above considerations has resulted in some statisticians advocating a second way to write the equation for $\widehat{y}$:

$$\widehat{y} = \bar{y} + b_1(x - \bar{x}).$$

For the cricket study:

$$\widehat{y} = 78.83 + 0.461(\frac{9.11}{31.0})(x - 166.8) =$$

$$78.83 + 0.1355(x - 166.8).$$

This second formula contains three numbers and they all have meaning: the mean of the predictor; the mean of the response and the slope. For better or worse, this formulation has not become popular and you are not responsible for it on the final.

It does, however, give us an easy proof of one of the most important features of the regression line, something I like to call: The law of preservation of mediocrity!

Suppose that a subject is mediocre on x; that is, the subject's $x = \bar{x}$. What is the predicted response for this subject? Plugging $x = \bar{x}$ into

$$\widehat{y} = \bar{y} + b_1(x - \bar{x})$$

we get

$$\widehat{y} = \bar{y} + b_1(\bar{x} - \bar{x}) =$$

$$\widehat{y} = \bar{y} + b_1(0) = \bar{y}.$$

Visually, the law of preservation of mediocrity means that the regression line passes thru the point $(\bar{x}, \bar{y})$.

Let us return to the BA study. Recall that the regression line is: $\widehat{y} = 0.095 + 0.633x$. Recall that $\bar{x} = 0.266$ and $\bar{y} = 0.264$ are close and the two sd's are similar. Thus, for the entire sample there was not much change in center or spread from 1985 to 1986.

If you want, you can verify the numbers in the following table.

		x	$\widehat{y}$
$\bar{x} - 2s_X$	$=$	0.210	0.229
$\bar{x} - s_X$	$=$	0.238	0.246
$\bar{x}$	$=$	0.266	0.264
$\bar{x} + s_X$	$=$	0.294	0.282
$\bar{x} + 2s_X$	$=$	0.322	0.299

In my experience, many people find this table surprising, if not wrong.

In particular, for small values of x the predictions seem (intuitively) to be too large, while for large values of x the predictions seem to be too small.

So, where is the error, in the regression line or in these people's intuition? Well, the quick answer is that the intuition is wrong. If you look at the data on page 443 of the text, you can verify the following facts:
–Of the 19 players with $x \geq 0.294$, 15 had their BA decline in 1986.
–Of the 18 players with $x \leq 0.238$, 11 had their BA increase in 1986.
This is not a quirk of the BA study. It always happens when we do regression. It is easiest to see when, as above, X and Y have similar means and similar sd's, as in the BA study.

This phenomenon is called **the regression effect**. (Discuss history.)

We can see the regression effect by rewriting the second version of the regression line as follows:

$$\frac{\widehat{y} - \bar{y}}{s_Y} = r(\frac{x - \bar{x}}{s_X}).$$

Ignore the r for a moment. Then the RHS of this equation is the standardized value of x, call it x'. The LHS is the value of $\widehat{y}$ 'standardized' using the mean and sd of the y's; call it $\widehat{y}'$. Thus, this equation becomes:

$$\widehat{y}' = rx'.$$

For ease of exposition, suppose that larger values of x and y are preferred to smaller values and that $r > 0$.

Consider a subject who has $x' = 1$. This is a talented subject; she is one sd better than the mean. Thus, we should predict that she will also be talented on y. The **intuitive prediction** is to ignore r and predict that $\widehat{y}' = 1$.

But as we saw in the BA study, this intuitive prediction is too large. We need to include r.

Now, if $r = 1$, then there is a perfect linear relationship between x and y and $x' = 1$ will yield $\widehat{y'} = 1$. (I conjecture that the reason people make intuitive predictions is that they tacitly assume $r = 1$; or, more likely, they don't have any feel for relationships that are not perfect.)

So, what does the regression line tell us? It tells us that we must pay attention to the r. Look again at the equation:

$$\widehat{y'} = rx'.$$

To make this argument precise, let's consider the BA study for which, recall, $r = 0.554$, which I will read as 55.4%.

We see that for a talented player, say $x' = 1$, the predicted value is $\widehat{y'} = 0.554(1) = 0.554$.

In words, for a player who is one sd above the mean on x, we predict that he will be 0.554 sd's above the mean on y. Thus, part of what we took for talent in x is transitory; dare I call it luck? And only part of the talent (r to be exact) is passed on to the $\hat{y}$. I really like this interpretation of r; it is the proportion of the advantage in x that is transmitted to $\hat{y}$.

A similar argument applies for $x' < 0$, the poor players. Only part (r again) of the poor performance in 1985 is predicted to carry over to 1986.

Example: Brett Favre isn't so great!

My subjects are the 32 NFL teams. $X =$ the number of victories in 2005 and $Y =$ the number of victories in 2006. Below are summary statistics:

$\bar{x} = \bar{y} = 8; s_X = 3.389; s_Y = 2.896; r = 0.286.$

The regression line is:

$$\hat{y} = 8 + 0.244(x - 8).$$

Thus, for $x = 13$, $\widehat{y} = 9.22$ and for $x = 4$, $\widehat{y} = 7.02$. Thus, using this equation, we would predict that the 2008 Jets would win 7.02 games and the 2008 Packers would win 9.22 games.

And the regression line does not take into account the incredibly easy schedule the Jets play (8 games against the pathetic western conferences) and the difficult schedule for the Packers. Moreover, if Crosby makes the last second field goal against the Vikings, the Packers are tied for first and would be the team in the playoffs. And this is not even allowing for the fact that the Packers lost to Tennessee b/c of a coin toss!

Finally, for $x = 16$, $\widehat{y} = 9.95$; so, even if Brady is not injured, the Patriots would be predicted to win about 10 games.

For the remainder of the course, we ponder one last question:

The regression line is the best line, but is it any good?

I will provide two answers to this question: one objective and one subjective. Not surprisingly, I greatly prefer the subjective answer.

The coefficient of determination

Recall that SSE measures how badly the regression line ‘fits’ the data, with $SSE = 0$ a perfect fit and larger values of SSE representing worse fits.

$$SSE = \sum(y - \widehat{y})^2.$$

Think of SSE as the best we can do using X to predict Y. (Implicit in this sentence is that we are predicting with a straight line.) It is natural to wonder: How well can we predict Y w/o using X? Let’s think about this. All we know for each subject is its x and y values. It seems unfair to predict y by y (always have the carnival-guy guess your weight *before* you get on the scale), and if we cannot use X, then all the subjects look the same to us;

that is, if we cannot use X then we must make the same prediction, call it c, for each subject.

Using c for each subject, the squared error for a subject is $(y - c)^2$ and the total of these is:

$$\sum(y - c)^2.$$

By using algebra or calculus, we can prove that this total is minimized by taking $c = \bar{y}$. Thus, the best prediction of Y w/o using X is $\bar{y}$ for every subject and the total of the squared errors is:

$$SSTO = \sum(y - \bar{y})^2.$$

Next, we note that for any set of data,

$SSE \leq SSTO$. Discuss.

On page 476 of the text I have the values of SSE and $SSTO$ for several studies. With real data, these numbers tend to be 'messy;' thus, I will introduce my ideas with nice fake numbers.

Suppose that for a set of data $SSTO = 100$ and $SSE = 30$. This means that w/o using X our total squared error is 100, but by using X we can reduce it to 30. It seems obvious that we want to compare these numbers:

$$SSTO - SSE = 100 - 30 = 70.$$

The difficulty is: 70 what? So, we do one more comparison:

$$\frac{SSTO - SSE}{SSTO} = \frac{100 - 30}{100} = 70/100 = 0.70,$$

or 70%. We call this ratio the coefficient of determination and denote it by R^2. An R^2 of 70% tells us that by using X to predict Y 70% of the squared error in Y *disappears* or *is accounted for* (choose your favorite expression).

Note that R^2 is a perfectly objective measure: it tells us how much better we do using X than we do not using X. For example, using X is 70% better than not using X and if we get real happy about this we might overlook the fact that our predictions might still be bad.

(I have no doubt that Sylvester Stallone is a 99%–or more–better actor than I.)

Finally, it can be shown algebraically that $R^2 = r^2$. This equation has a good and a bad consequence. First, the good. This reinforces the fact that a relationship with, say, $r = -0.60$ is exactly as strong as a relationship with $r = 0.60$ b/c they both have $R^2 = 0.36$. Second, the bad. This equation has inspired many people to wonder which is better, R^2 or r, which I find annoying. Worse yet, most seem to reach the wrong conclusion: that R^2 is better. (Discuss.)

Section 13.4: The last section (time for massive rejoicing)

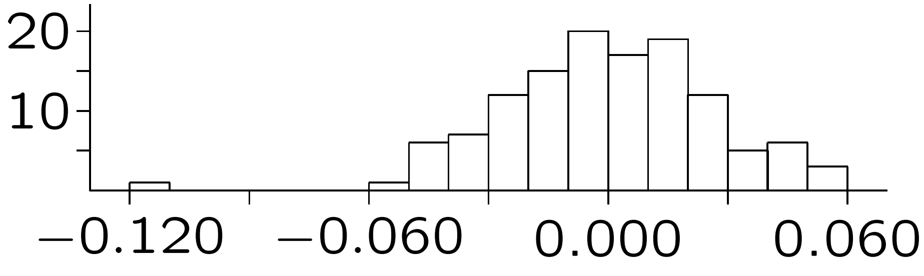
Each subject has an x and a y. Once we determine the regression line, each subject has a $\widehat{y}$ and thus an error $e = y - \widehat{y}$. Now things get 'funny.' Statisticians got tired of explaining to clients that we are not being judgmental when we talk about each of their cases having an 'error.'

So, we decided to start calling e the **residual**, not the error. But we cannot call these r b/c the letter r is ingrained as representing the correlation coefficient.

Thus, just remember e stands for residual!

If you study regression in the future, you will learn that the collection of residuals for a data set can be very useful. Here we will consider a few minor properties of them.

We have n numbers: $e_1, e_2, \ldots, e_n$; a residual for each subject. The MITK thinks, "This reminds me of Chapter 12." So, we draw a picture (dot plot or histogram) of the residuals. Below is a frequency histogram of the residuals for the BA study.



What do we see? Well, there is one small outlier (guess who?). Note that outlier, as in Chapter 12, is a term for one set of numbers, while isolated case is a term for two dimensional data. Also, we can see that every outlier is isolated, but a case can be isolated w/o having its residual be an outlier.

Next, we calculate the mean and sd of the residuals. **Fact: For any set of data** $\sum e = 0$. (Discuss.) Thus, $\bar{e} = 0$. The sd of the residuals is denoted by s. Note that there is no subscript on this s. In Chapter 13 we have three sd's of interest: s_X, s_Y and s. Statisticians think that the most important of these is the sd of the residuals; thus, it is unencumbered by a subscript.

For the BA data, $s = 0.027$, 27 points. B/c the distribution of the residuals is bell-shaped (curiously, it usually is for regression data), we can use the Empirical Rule of Chapter 12 to interpret $s = 0.027$.

In particular, approximately 68% of the residuals are between -0.027 and $+0.027$. In words, for approximately 68% of the players, $\widehat{y}$ is within 27 points of y. (And for the other 32% of the data $\widehat{y}$ and y differ by more than 27 points.)

Here is the subjective part. As a baseball fan, I opine that these are not very good predictions. To err by 27 points is a lot in baseball.

BTW, if FR is dropped from the data set, s is reduced to 0.025, which is better, but I still believe that the predictions are not very good.

As another example, for the ‘Favre data,’ $s =$ 2.821 which means that for approximately 32% of the teams $\widehat{y}$ misses y by 3 or more games. (By actual count, the number is 11 of 32, or 34%.) In my opinion this is a very bad prediction rule.

I will note that there is another restriction on the residuals (other than that they sum to 0). It is:

$$\sum(xe) = 0.$$

The interested reader can refer to the book to see one reason this is useful. (This fact is used a great deal in advanced work on regression.)

Finally, page 487 shows the (sometimes huge) effect a single isolated case can have on the value of r and, hence, the regression line. Page 487 and its consequences will not be on the final.