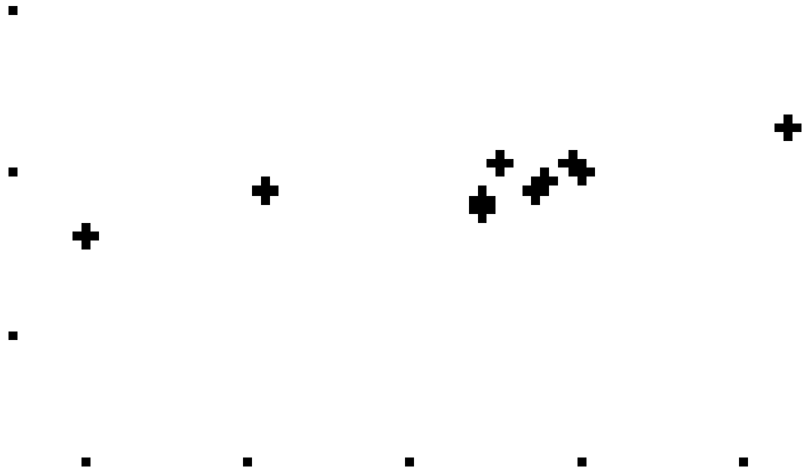
Least Squares Regression Line
Part 1

NY Yankees 1995-2005

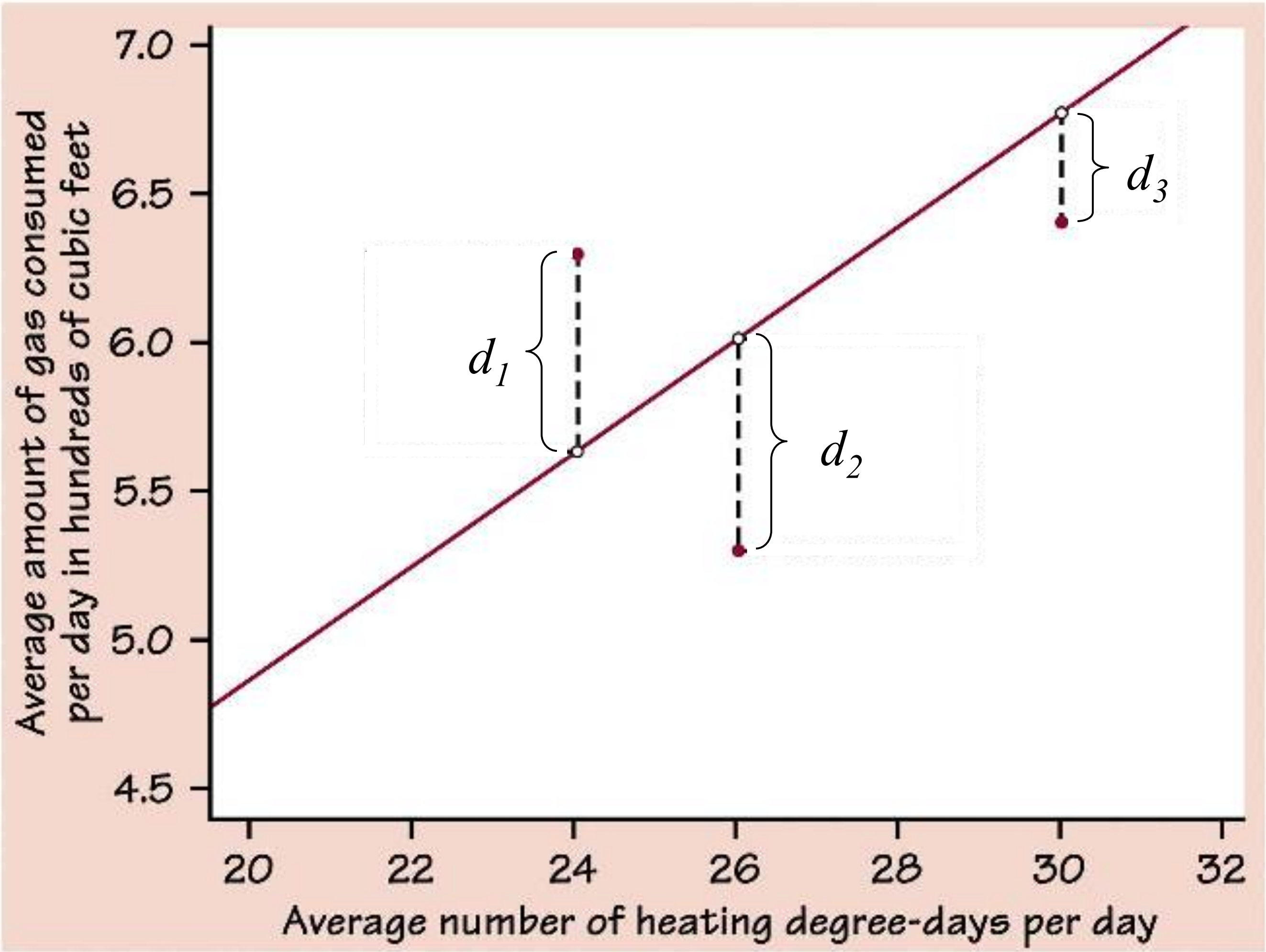
Here is the scatterplot of runs scored vs wins. We are now going to draw a “best-fit line”.

Runs Scored	Wins
886	95
897	101
877	101
897	103
804	95
871	87
900	98
965	114
891	96
871	92
749	79



Such a line is also called a ***regression line***.

A regression line is based on finding a line in which the sum of the vertical distances from each of the points to the line is as small as possible.



But d_1 is positive while d_2 and d_3 are negative because the points are below the line. For this and other reasons not shown here, it is best to find the least total *squares* of the vertical distances in the scatterplot.

→

The regression line equation is given by: $\hat{y} = a + bx$

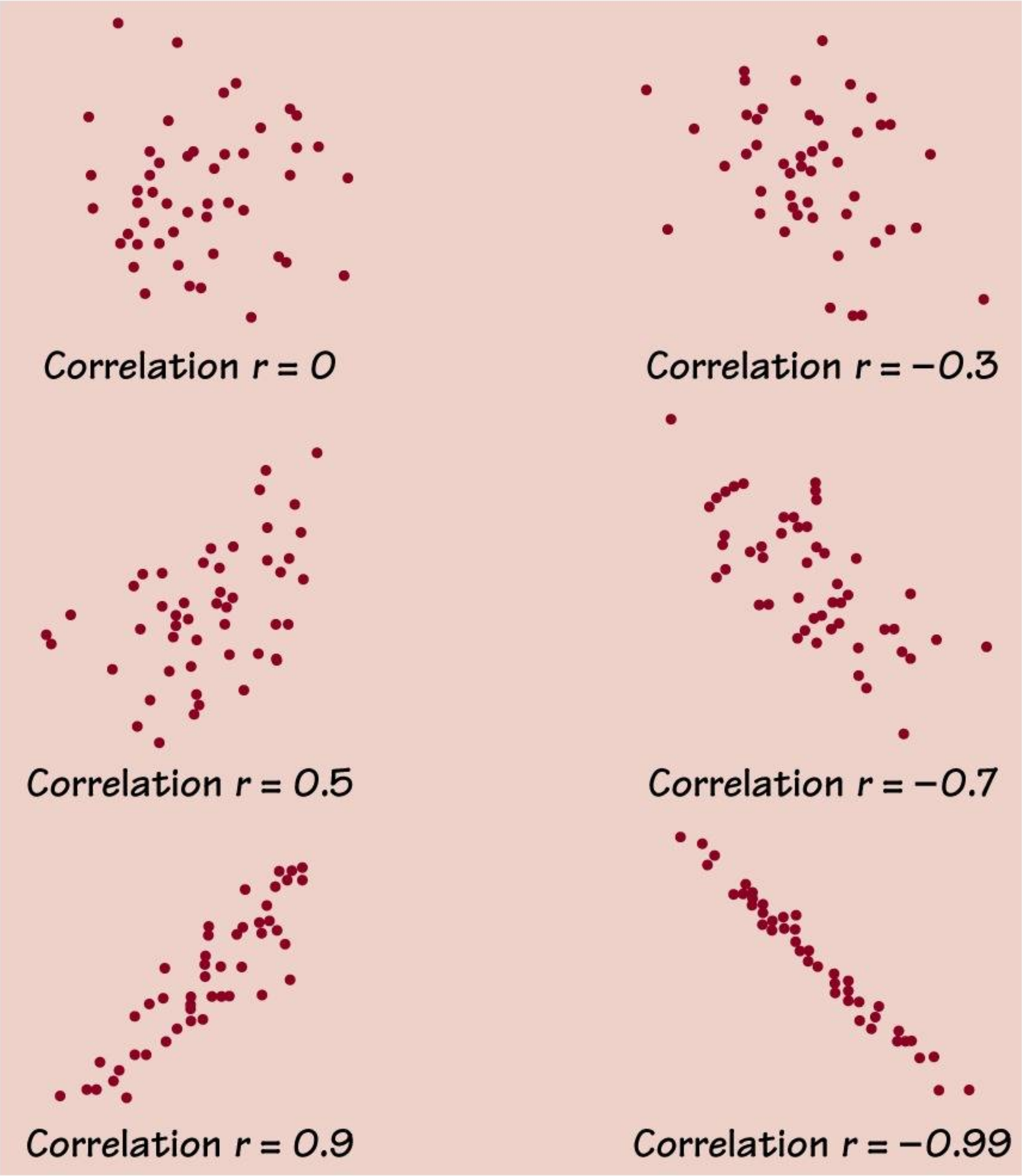
Where a and b can be found in this way: $b = r\frac{s_y}{s_x}$ $a = \bar{y} - b\bar{x}$

If r is close to ±1, then the points form a strong linear pattern, meaning that the sum of the squares of all of the distances is very small because all of the points are so close to the line.

This means that the equation of such a line could be used to "predict" the response (*y*) variable from the explanatory (*x*) variable. To draw the regression line, we need a formula. We also need to see how this line is interpreted.

So if we were to look for a regression line for our previous scatterplot of Runs Scored and Wins, we would use the data found on mean, standard deviation, and the correlation coefficient.

→



Correlation $r = 0$

Correlation $r = -0.3$

Correlation $r = 0.5$

Correlation $r = -0.7$

Correlation $r = 0.9$

Correlation $r = -0.99$

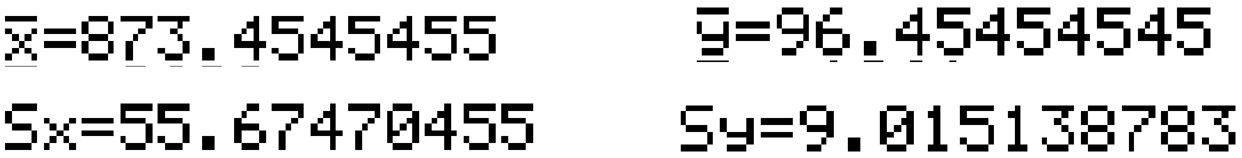
The regression line equation is given by: $\hat{y} = a + bx$

Where a and b can be found in this way: $b = r\frac{s_y}{s_x}$ $a = \bar{y} - b\bar{x}$

If r is close to ±1, then the points form a strong linear pattern, meaning that the sum of the squares of all of the distances is very small because all of the points are so close to the line.

This means that the equation of such a line could be used to "predict" the response (*y*) variable from the explanatory (*x*) variable. To draw the regression line, we need a formula. We also need to see how this line is interpreted.

So if we were to look for a regression line for our previous scatterplot of Runs Scored and Wins, we would use the data found on mean, standard deviation, and the correlation coefficient.



```
x̄=873.4545455        ȳ=96.45454545
Sx=55.67470455       Sy=9.015138783
```

r = 0.838

→

$$b = 0.838\frac{9.015}{55.675} = $$ 0.136

$$\hat{y} = a + bx$$

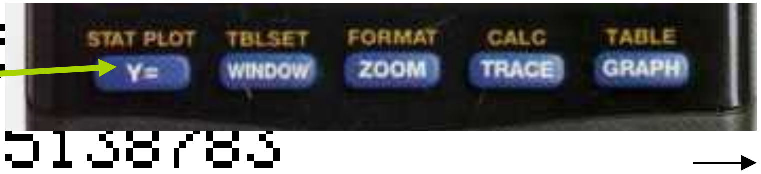
$$a = 96.455 - (0.136)(873.455) = $$ –22.335

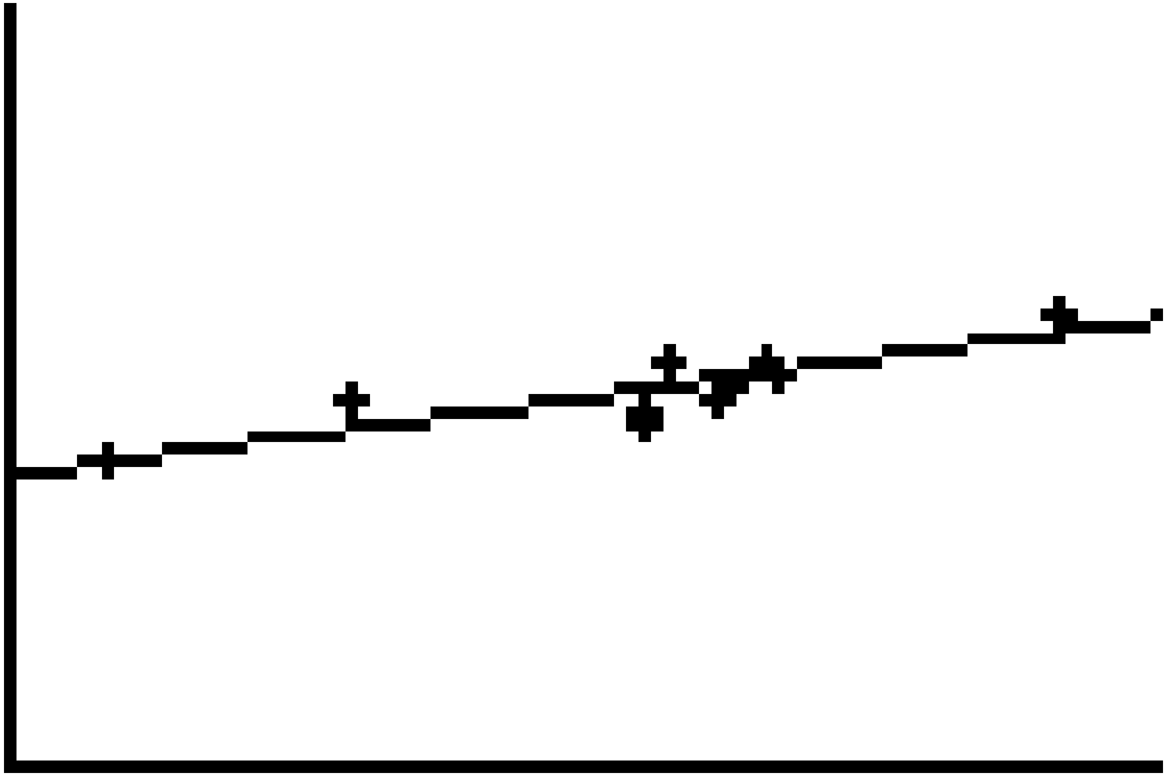
$$b = r\frac{s_y}{s_x} \qquad a = \bar{y} - b\bar{x}$$

$$\hat{y} = -22.335 + 0.136x$$

Now insert this equation into your calculator entering it into Y_1 which you can find by hitting this button.

```
x̄=873.4545455          ȳ=96.45
Sx=55.67470455         Sy=9.015138783
```





Notice how close each point is to the line. The fact that r = 0.838…

. indicates that the points are all close to the line and the graph shows the same

Is it strong or weak?

Strong

Positive or negative?

Positive because of the positive slope of graph

Use process demonstrated in the power point on finding *r* to find the regression line for Total Runs Allowed vs. Wins. The data you need are shown here.

Total Runs Allowed	Wins
789	95
808	101
716	101
697	103
713	95
814	87
731	98
656	114
688	96
787	92
688	79