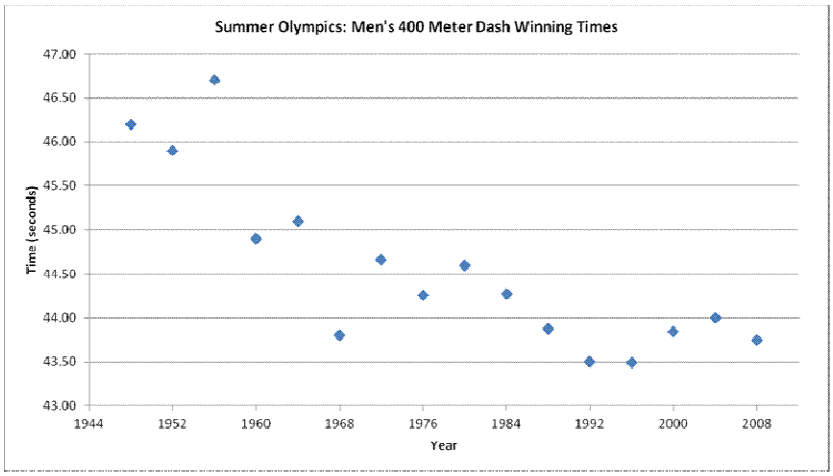
Scatterplots, Linear Regression, and Correlation

When we have a set of data, often we would like to develop a model that fits the data.

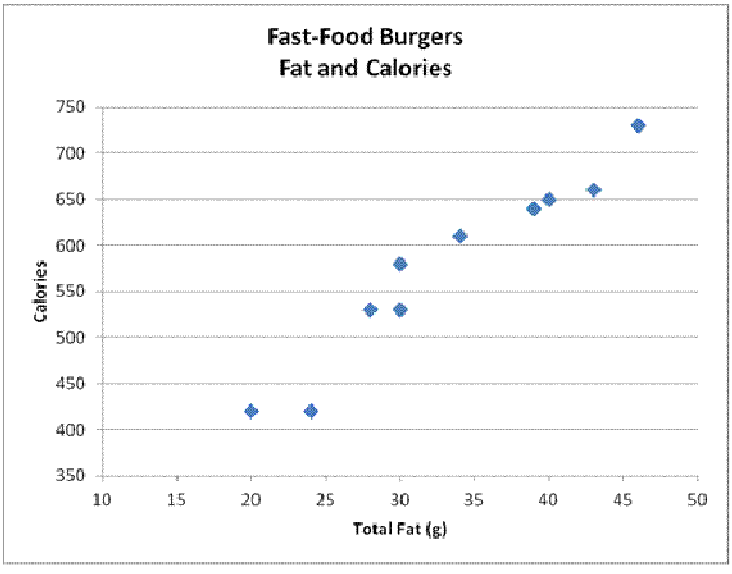
First we graph the data points (*x*, *y*) to get a **scatterplot.** Take the data, determine an appropriate scale on the horizontal axis and the vertical axis, and plot the points, carefully labeling the scale and axes.

Summer Olympics: Men's 400 Meter Dash Winning Times	
Year (x)	**Time(y) (seconds)**
1948	46.20
1952	45.90
1956	46.70
1960	44.90
1964	45.10
1968	43.80
1972	44.66
1976	44.26
1980	44.60
1984	44.27
1988	43.87
1992	43.50
1996	43.49
2000	43.84
2004	44.00
2008	43.75



Burger	Fat (x) (grams)	Calories (y)
Wendy's Single	20	420
BK Whopper Jr.	24	420
McDonald's Big Mac	28	530
Wendy's Big Bacon Classic	30	580
Hardee's The Works	30	530
McDonald's Arch Deluxe	34	610
BK King Double Cheeseburger	39	640
Jack in the Box Jumbo Jack	40	650
BK Big King	43	660
BK King Whopper	46	730

Data from 1997



If the scatterplot shows a relatively linear trend, we try to fit a linear model, to find a line of best fit.

We could pick two arbitrary data points and find the line through them, but that would not necessarily provide a good linear model representative of all the data points.

A mathematical procedure that finds a line of "best fit" is called **linear regression**. This procedure is also called the method of least squares, as it minimizes the sum of the squares of the deviations of the points from the line. In MATH 107, we use software to find the regression line. (We can use Microsoft Excel, or Open Office, or a hand-held calculator or an online calculator --- more on this in the Technology Tips topic.)

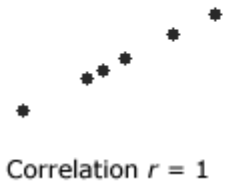
Linear regression software also typically reports parameters denoted by r or r^2.

The real number ***r*** is called the **correlation coefficient** and provides a measure of the strength of the linear relationship.

r is a real number between -1 and 1.

$r = 1$ indicates perfect positive correlation --- the regression line has positive slope and all of the data points are on the line.

$r = -1$ indicates perfect negative correlation --- the regression line has negative slope and all of the data points are on the line



Correlation $r = 1$

Correlation $r = -1$

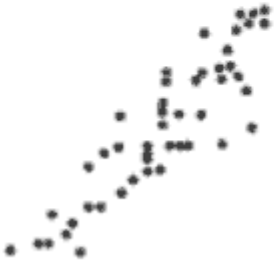
The closer $|r|$ is to 1, the stronger the linear correlation. If $r = 0$, there is no correlation at all. The following examples provide a sense of what an r value indicates.

Correlation $r = 0$

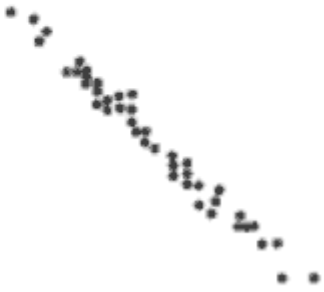
Correlation $r = - 0.3$

Correlation $r = 0.5$

Correlation $r = - 0.7$



Correlation $r = 0.9$



Correlation $r = - 0.99$

Source: *The Basic Practice of Statistics*, David S. Moore, page 108.

Notice that a positive r value is associated with an increasing trend and a negative r value is associated with a decreasing trend. The strongest linear models have r values close to 1 or close to -1.

The nonnegative real number $\boldsymbol{r^2}$ is called the **coefficient of determination** and is the square of the correlation coefficient r.

Since $0 \le |r| \le 1$, multiplying through by $|r|$, we have $0 \le |r|^2 \le |r|$ and we know that $-1 \le r \le 1$.
So, $0 \le r^2 \le 1$. The closer r^2 is to 1, the stronger the indication of a linear relationship.
Some software packages (such as Excel) report r^2, and so to get r, take the square root of r^2 and determine the sign of r by observing the trend (+ for increasing, – for decreasing).