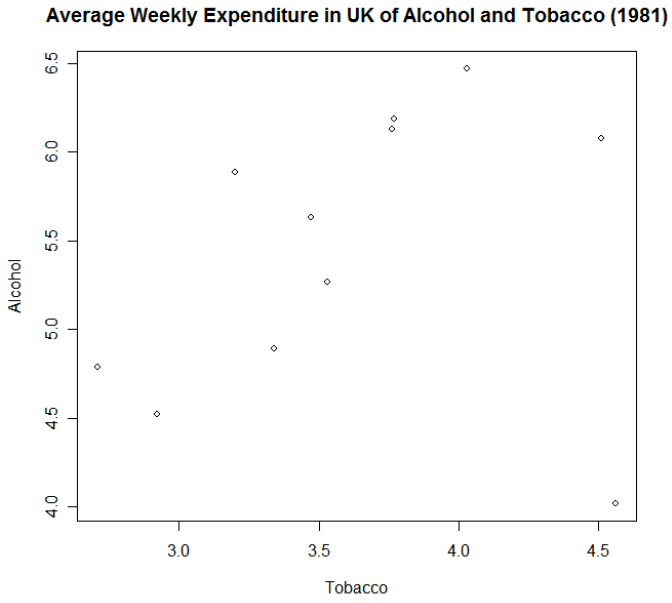
Chapter 10

Key Ideas
Correlation, Correlation Coefficient (*r*),

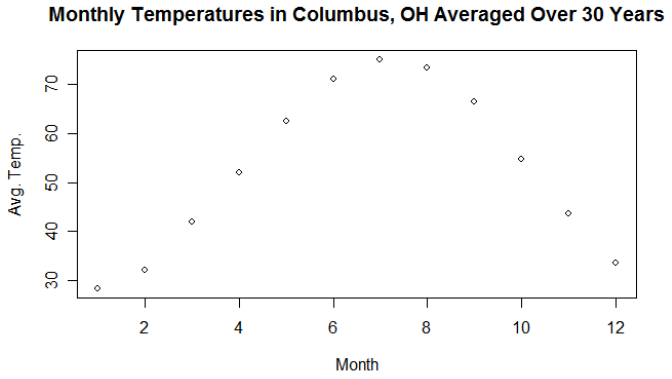
Section 10-1: Overview
We have already explored the basics of describing single variable data sets. However, when two quantitative variables are present, we can examine the relationship *between* these two variables as well. The easiest and most common way to do this is called *simple linear regression*. If we call one variable *x* and the other *y*, linear regression allows us to use the available data to create an equation that can be used to predict future values of *y* given a value of *x*. The equation will be for a line, which is why it is called "linear" regression. Before we get to regression, however, we need to look at a basic measure of the linear relationship between *x* and *y*.

Section 10-2: Correlation
When two variables *x* and *y* have an association (or relationship), we say there exists a **correlation** between them. Alternatively, we could say *x* and *y* are *correlated*. To find such an association, we usually look at a scatterplot and try to find a pattern.

Example 1: Average Alcohol and Tobacco Weekly Expenditure in the UK (1981)
We see here that there is a general trend in alcohol and tobacco sales in the UK. From looking at the plot, it seems that the more tobacco someone buys each week, the more likely they are to buy more alcohol as well. Therefore, Alcohol and Tobacco sales are *correlated*. In this case, we can say they are **positively correlated**, since when one increases, so does the other one.



Example 2: Average Monthly Temperature in Columbus, OH.
Here we see the average temperature each month for Columbus, averaged over a 30 year period. There is a definite trend here between temperature and month of the year. As we would expect, the temperature gets higher as we get to the summer months, then declines as winter moves in. From this plot, we see that month and temperature are correlated.



As can be expected, simply looking at a scatterplot and making a decision about correlation is highly subjective. To create a more rigorous way to explain the association between two variables, we will focus on the *linear relationship* between the two variables. In other words, we will focus on relationships like the first example, and not so much on those like the second.

The Correlation Coefficient

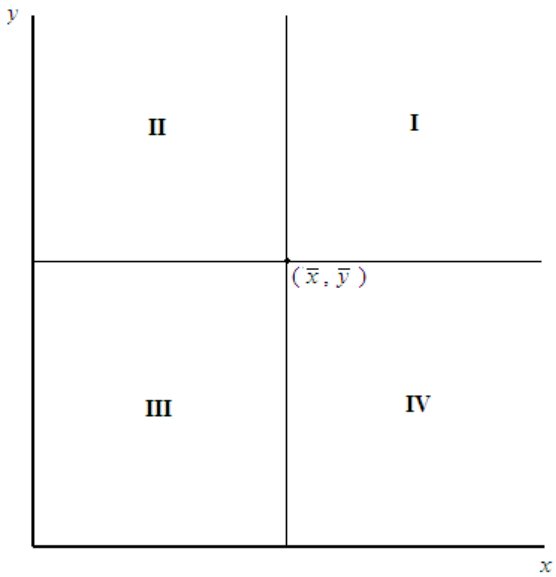
The linear **correlation coefficient** (denoted *r*) measures the strength of the linear relationship between two variables *x* and *y* in a sample. It is sometimes called a Pearson Correlation Coefficient, after a famous statistician named Karl Pearson. The formula for this statistic is as follows:

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

Here, $\bar{x}$ and $\bar{y}$ are the means of *x* and *y*. Likewise, s_x and s_y are the standard deviations of each variable. An alternate formula on page 520 of the text can be used for hand calculations.

Interpreting the Formula

Looking at the formula above, the first thing to note is that for a given observation x_i, the quantity $(x_i - \bar{x})$ will be negative if the observation is below the mean, and positive if it is above the mean. This is also true for the *y* observations. Therefore, the numerator $(x_i - \bar{x})(y_i - \bar{y})$ will be positive when both x_i and y_i are below their means (the negatives cancel) or when they are both above their means. So if *r* turns out to be positive, this means many of the observations are falling in regions I and III in the picture below (this corresponds to positive correlation). Likewise, if *r* is negative, many of the observations will fall in regions II and IV (this would be negative correlation).



Therefore, if *r* is positive and large, then most of the points will be in regions I and III, which means they are arranged in a linear fashion from bottom left to upper right. Similarly, if *r* is negative and large, most points are in regions II and IV and are arranged from top left to bottom right in a linear fashion. If *r* is very small, this means that the number of points in each region was close to equal, and the dots would be spread out all over the place in no discernable pattern.

The standard deviations in the denominator of the sum are there to standardize the distances between each observation and its mean (since *x* and *y* probably do not have the same units, this will put them on the same scale). Finally, we divide by $n - 1$ to average the standardized distances from the mean. (The reason we use $n - 1$ and not n is the same as it was when we calculated standard deviation).

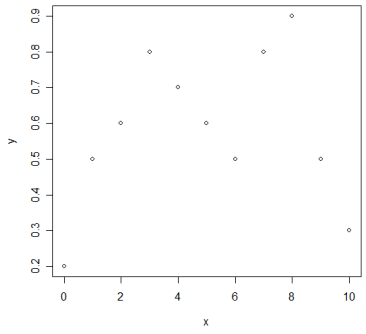
Properties of the Correlation Coefficient

Due to the standardization that takes place in the formula, there are a couple of interesting properties of *r* :

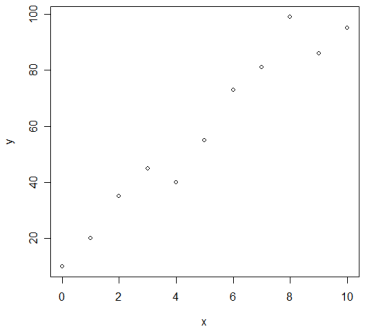
1. $-1 \le r \le 1$
2. If the values of either variable are converted to a different scale, *r* will be the same.
3. If the variables *x* and *y* are interchanged, *r* will be the same.
4. The correlation coefficient *r* will only measure the strength of a *linear* relationship. It says nothing about other kinds of relationships, like the temperature data on the previous page.

How to Interpret *r*

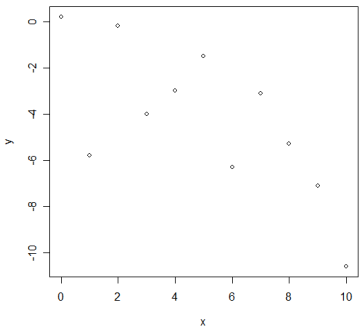
Perhaps the easiest way to understand how to interpret a value of *r* is to look at some examples:



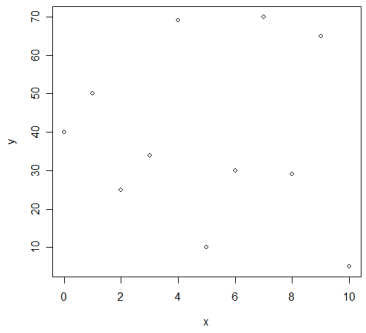
r = 0.169



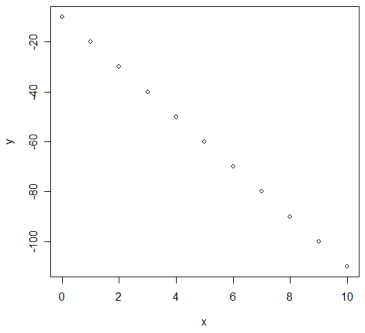
r = 0.971



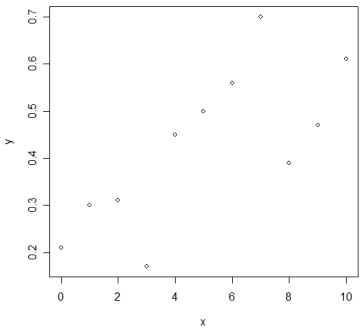
r = -0.714



r = -0.094



r = -1.000



r = 0.740

From these plots, we see that the closer the observations are to being in a line, the closer *r* gets to 1 or –1. The sign of the correlation coefficient tells whether the line is sloping upward (positive) or downward (negative). When *r* is close to 0, there seems to be no linear pattern at all. Here is a general guide for how to interpret the value of *r* :

Range of Values	Strength of Linear Relationship	Direction of Linear Relationship
-1 to -0.8	Strong	Negative
-0.8 to -0.6	Moderate	Negative
-0.6 to -0.3	Weak	Negative
-0.3 to 0.3	None	None
0.3 to 0.6	Weak	Positive
0.6 to 0.8	Moderate	Positive
0.8 to 1	Strong	Positive

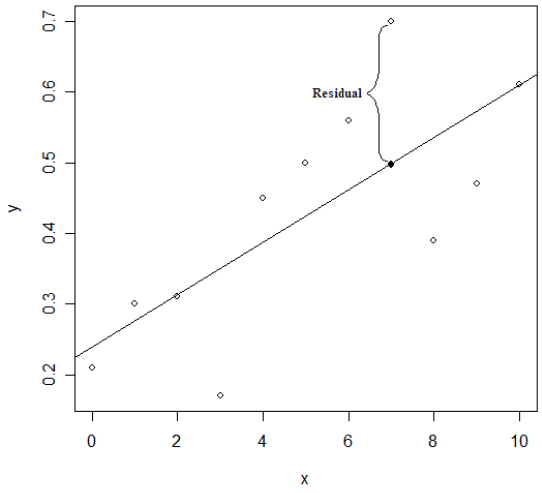
Section 10-3: Regression

Correlation sheds some light on a linear relationship between two variables, but it only tells us how strong such a relationship is. Linear Regression allows us to go one step further and actually specify a *linear equation* that relates the two variables.

Note: Linear Regression is *only* useful if there is evidence of a linear relationship between x and y. If the correlation is very low, then the line will fit horribly and will be useless. In this situation, use the mean $\bar{y}$ to estimate y.

Assuming two variables have a linear relationship, regression is the technique of finding the equation of the line that fits as closely as possible to the observed data values. In a nutshell, this is done by minimizing the *residuals*, which are the differences between the observed values and the values predicted by the equation of the line (see picture below). The formula for calculating a residual is:

$$\text{Residual} = \text{Actual Value} - \text{Predicted Value} = y - \hat{y}$$



It turns out that the best fitting line can always be calculated the same way from the observed data.

The form of the equation is: $\hat{y} = b_0 + b_1 x$, where $b_1 = r\dfrac{s_y}{s_x}$ and $b_0 = \bar{y} - b_1\bar{x}$.

Here, $\bar{x}, s_x$ and $\bar{y}, s_y$ are the means and standard deviations of x and y. The value r is the correlation coefficient from the previous section. Also, b_0 is the y-intercept of the equation and b_1 is the slope. Finally, we put a little hat on the y in the equation to indicate that these are not exact values of y, but rather they are *estimates*.

An alternate formulation for b_1 is given on page 542 in the text, as well.

Interpreting the Coefficients

Now that we can find the best fitting line, it would be helpful to know what it says about the relationship between the two variables x and y.

The y-intercept b_0: From basic high school algebra, we know that the y-intercept of a line is the value y takes when $x = 0$. The interpretation is the same in regression, only now y and x stand for variables. Sometimes this interpretation makes sense, and sometimes it does not. If $x = 0$ is possible (e.g. if x stands for the number of children in a family) then b_0 represents the *estimated* value of y when $x = 0$. However, in many cases $x = 0$ doesn't make any sense (e.g. if x stands for the diameter of a tree, you can't have a tree with 0 diameter). In this case, it is just there as a part of the equation, with no interpretation.

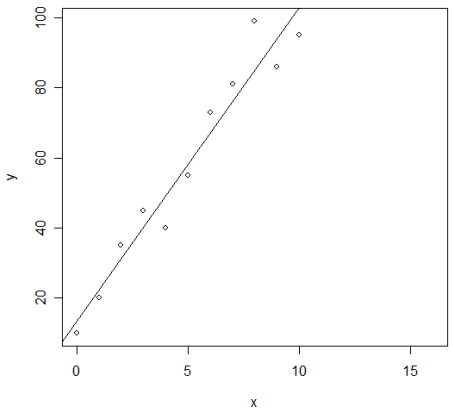
The slope b_1: The slope will always have an interpretation. To illustrate it, consider the equation of the best fit line: $\hat{y} = b_0 + b_1 x$. Here, we are saying that the estimated value of y is given by $b_0 + b_1 x$, for a known value of x. What would happen if we increased x by 1 unit?

$$\begin{aligned}\hat{y} &= b_0 + b_1(x+1)\\ &= (b_0 + b_1 x) + b_1\\ &= (\text{old value of } \hat{y}) + b_1\end{aligned}$$

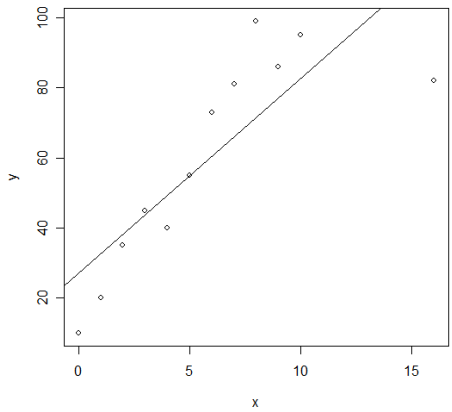
So when x increased by 1, y increased by b_1.
Therefore, we interpret the slope by saying it is the average change in y when x is increased by 1 unit.

Some Important Facts and Cautionary Notes About Regression Lines

1. Using a regression line makes the implicit assumption that there is a strong linear relationship between x and y. If their correlation is weak, you can still find a regression line but it will not predict very well at all. Therefore, you should only use regression if the value of the correlation coefficient r indicates a linear relationship is present.
2. A regression line only fits well to data that you have collected. For this reason, prediction with the line should only be done over the range of observations that were used to produce the line. Predicting outside the range of data values is called *extrapolation*, and it is dangerous to do, since you don't know what happens outside those values. Prediction is best at the point $(\bar{x}, \bar{y})$, which is called the *centroid*.
3. If a regression line is based on old data, it may not always be accurate for the present. Try to use lines for prediction that are based on current data.
4. Do not use a line to predict for a population that was not represented in the sample. For example, let's say you use a sample of American college students to make a regression line to predict credit card debt (y) using average weekly earnings (x). This line should not be used to predict credit card debt for anyone other than American college students. So, for instance, the relationship does not necessarily hold true for retired senior citizens.
5. If an observation is an outlier in the x direction (i.e. on the horizontal scale it is far away from the other points), then it is called an *influential point*. This is because the line's slope will change dramatically to adjust to that point being in the data set. (see plots below)



A regression line without the influential point.



The new line with the influential point added.

6. If an observation is an outlier in the y direction (i.e. on the vertical scale it is far away from the other points), then it will usually have a large residual. This observation should be examined to see if there was a measurement error, human error, or unusual circumstances that led to its unnaturally high (or low) value.

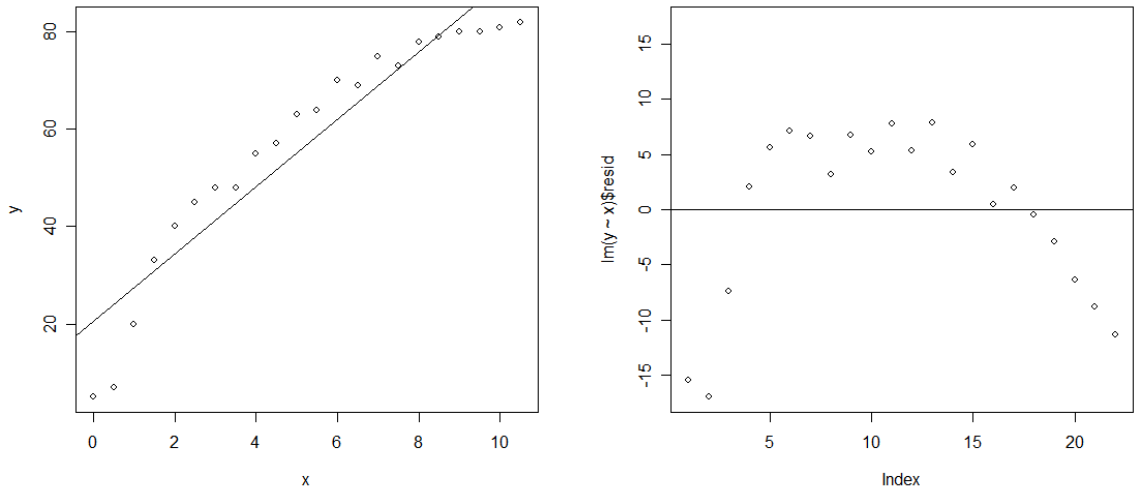
Residual Plots

A major diagnostic to use in making sure that using a regression line is appropriate is a *residual plot*. This plot displays the residual for each observation centered at zero. Ideally, what we would like is a line that fits the data well, and in this case the residuals should be small. Also, if the relationship between x and y is truly linear, we should expect to see positive values as much as negative values, and in a random pattern. For this reason, the ideal residual plot is one in which the points are randomly scattered in a horizontal band centered at zero (see next page for an example). Any patterns in the residual plots indicate a potentially poor model fit.

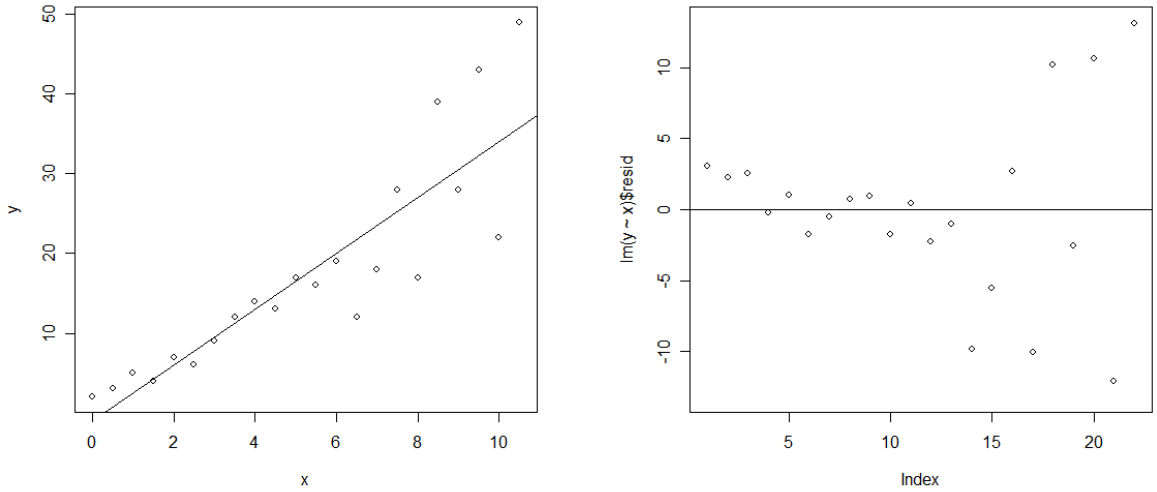
Examples of Patterns in the Residual Plot

1. If there seem to be a lot of runs of many positive or negative residuals in a row, the line may not be good for prediction because observations are *dependent* upon previous data values.
2. If there is a curved pattern in the residuals, then x and y may not have a linear relationship, but rather a quadratic or higher order relationship (see next page for an example).
3. If the residuals fan out (i.e. start small then get bigger, or vice versa), then the line will predict poorly where residuals are larger. This indicates a variability issue with the model that is difficult to resolve (see next page for an example).

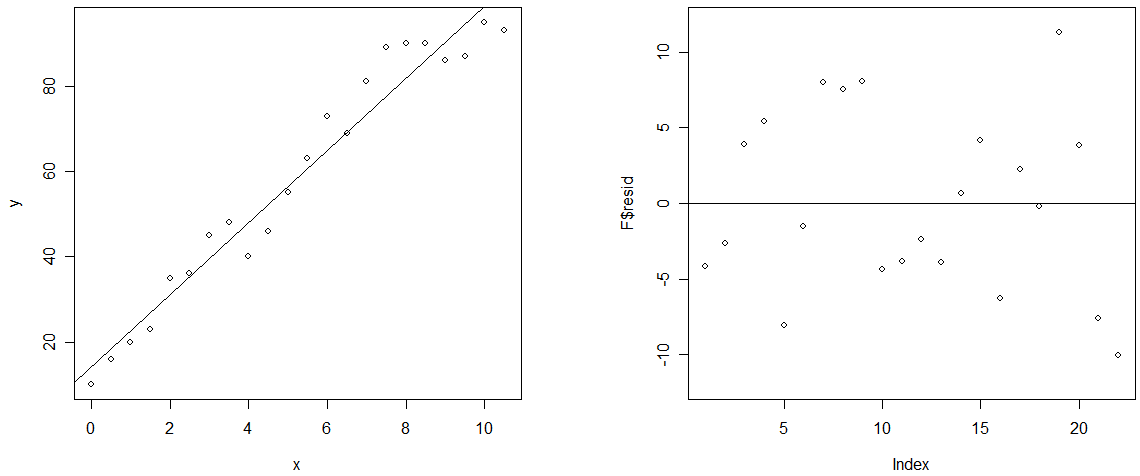
The next page shows some plots of data sets (with their regression lines) and the associated residual plot.



Clearly, this is a problem. The variables do not have a linear relationship… it is curved. The curved pattern shows up on the residual plot, which tells us something is wrong.



Here, we see that the observations start out close together, then spread out. This is reflected in the residual plot as well, and lets us know that the line will be poor at predicting with higher values of *x*.



Here is a good example of a data set with *x* and *y* strongly correlated, with a linear relationship. The residual plot shows a random scattering of points between ±10, which is a good indication that nothing is wrong.

Section 10-4: Variation and Prediction Intervals

The Coefficient of Determination

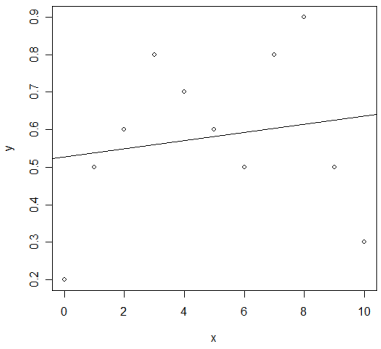
It is natural at this point to wonder whether there is a statistic that can tell you how well the regression line fits the data. Visual inspection is often useful, as is looking at the size of residuals. However, there is a nice statistic that can be calculated from the data that gives us a good way to gauge the effectiveness of a regression line. This is the **coefficient of determination**.

Calculation of the coefficient of determination is not difficult, because it is simply the square of the correlation coefficient, r. This statistic measures the percent of variation in y that is explained by the regression line. Since r is between –1 and 1, we can see that r^2 will be between 0 and 1, which is 0% and 100%. Thus if the observations are perfectly aligned in a linear pattern (i.e. $r = \pm 1$), then 100% of the variation in y will be explained by the regression line (since $r^2 = 1$). If no linear pattern is present (i.e. $r = 0$), then 0% of the variation in y will be explained. Due to this interpretation, the value of r^2 can be a good indicator of how well the regression line is predicting y given values of x.

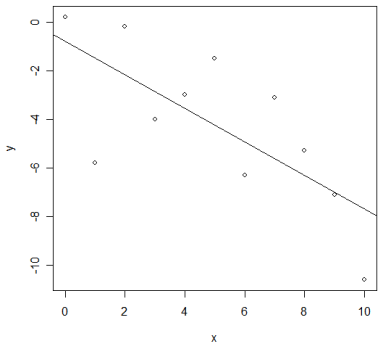
Note: On pages 557-559 of the text, a good explanation is given for why r^2 has this interpretation.

Examples

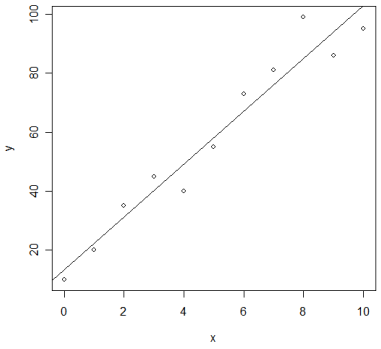
Here are the three regression lines, with their data sets and associated r^2 values:



$r^2 = 0.029$



$r^2 = 0.511$



$r^2 = 0.944$

Prediction Intervals

We will not discuss these at this point in the course, but you are encouraged to read about this topic if you are interested.