Chapter 4: Describing the Relationship Between Two Variables

In Chapter 3, we looked at numerically summarizing data from one variable (**univariate data**), but newspaper articles and studies frequently describe the relationship between *two* variables (**bivariate data**). It's this second class that we'll be focusing on in Chapter 4.

There are plenty of variables which seem to be related. The links below are articles from various news sources, all discussing relationships between two variables.

> Do SAT Scores Really Predict Success?
> Range of Variables Affect How SAT Correlates to College GPA
>
> Proximity to highways affects newborns' health: study
>
> Study: Weight-loss surgery cuts cancer risk in women

Our goal in this chapter will be to find ways to describe relationships like the one between a student's SAT score and his/her GPA, and to describe the *strength* of that relationship.

Section 4.1: Scatter Diagrams and Correlation

Objectives

By the end of this lesson, you will be able to...

1. draw and interpret scatter diagrams
2. describe the properties of the linear correlation coefficient (LCC)
3. estimate the LCC based on a scatter diagram
4. compute and interpret the LCC
5. explain the difference between correlation and causation

In Chapter 3, we looked at numerically summarizing data from one variable (**univariate data**), but newspaper articles and studies frequently describe the relationship between *two* variables (**bivariate data**). It's this second class that we'll be focusing on in Chapter 4.

There are plenty of variables which seem to be related. The links below are articles from various news sources, all discussing relationships between two variables.

> Do SAT Scores Really Predict Success?
> Range of Variables Affect How SAT Correlates to College GPA
>
> Proximity to highways affects newborns' health: study
>
> Study: Weight-loss surgery cuts cancer risk in women

In each case, there's a **response variable** (GPA, newborn's health, cancer levels) whose value can be explained at least in part by a **predictor variable** (SAT score, proximity to highways, weight-loss pill consumption).

> Remember, unless we perform a designed experiment, we can only claim an *association* between the predictor and response variables, not a *causation*.

Our goal in this chapter will be to find ways to describe relationships like the one between a student's SAT score and his/her GPA, and to describe the *strength* of that relationship.

First, we need a new type of graph.

Scatter Diagrams

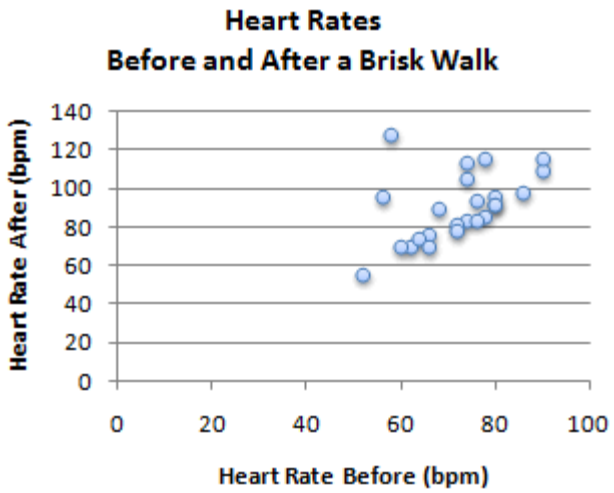
Scatter diagrams are the easiest way to graphically represent the relationship between two quantitative variables. They're just x-y plots, with the predictor variable as the x and the response variable as the y.

Example 1

The data below are heart rates of students from a Statistics I class at ECC

during the Spring semester of 2008. Students measured their heart rates (in beats per minute), then took a brisk walk and measured their heart rates again.

before	after	before	after	before	after
86	98	58	128	60	70
62	70	64	74	80	92
52	56	74	106	66	70
90	110	76	84	80	92
66	76	56	96	78	116
80	96	72	82	74	114
78	86	72	78	90	116
74	84	68	90	76	94



We can see that the heart rate before going on the walk is the predictor (x), and the heart rate after the walk is the response (y).

Here's an excellent video showing a scatter diagram on steroids created by the BBC:

Technology

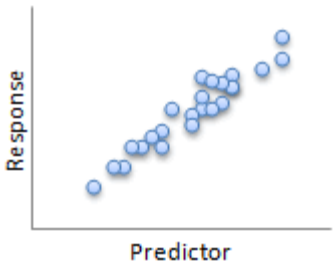
Here's a quick overview of the steps for creating scatter diagrams in StatCrunch.

1. Select **Graphics** > **Scatter plot**
2. Select quantitative variables for the X & Y axes.

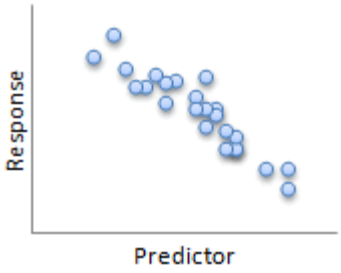
You can also go to the video page for links to see videos in either Quicktime or iPod format.

Types of Relationships

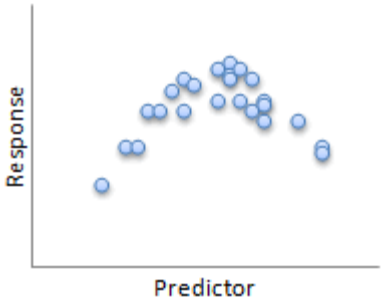
Not all relationships have to be linear, like the before/after heart rate data. The images below show some of the possibilities for the relationship (or lack thereof) between two variables.



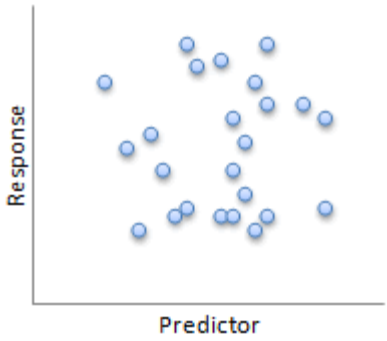
Linear



Linear



Nonlinear



No relation

The price of a manufactured item and the profit the company gains from it, for example, do not have a linear relationship. When prices are low, sales are high, but profit is still low since very little is made from each sale. As prices increase, profits increase, but at some point, sales will start to drop, until eventually too steep of a price will drive sales down so far as to not be profitable. This might be represented by the third, "Nonlinear" image.

Positive and Negative Association

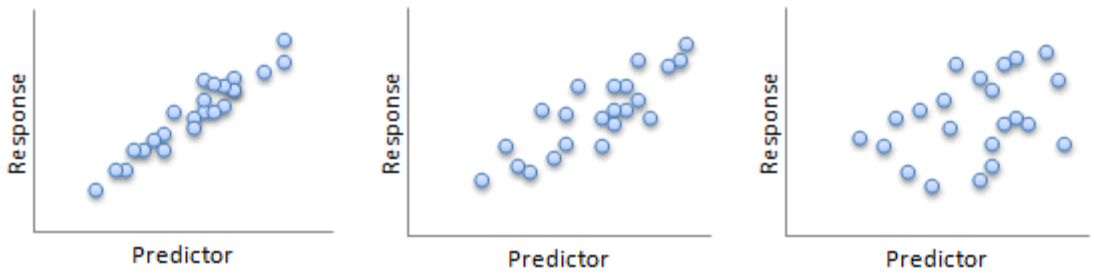
The next thing we to do is somehow quantify the strength and direction of the relationship between two variables.

Here's how we'll describe the direction:

In general, we say two linearly related variables are **positively associated** if an increase in one causes an increase in the other (first "Linear" image). We say two linearly related variables are **negatively associated** if an increase in one causes a decrease in the other (second "Linear" image).

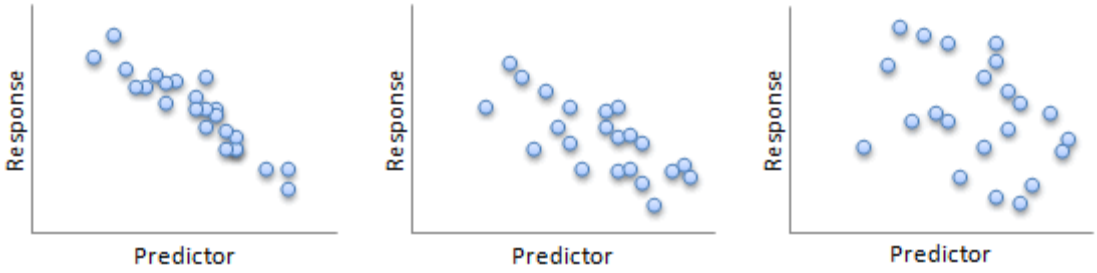
The images below show some examples of what scatter plots might look like for two positively associated variables.

positively associated



And these are some examples of what scatter plots might look like for two negatively associated variables.

negatively associated



The Linear Correlation Coefficient

As we can see from these examples, knowing the directions isn't enough - we need to quantify the strength of the relationship as well. What we'll use to do that is a new statistic called the **linear correlation coefficient**. (In this class, we'll be dealing solely with linear relationships, so we usually just call it the *correlation*.)

The **linear correlation coefficient** is a measure of the strength of the linear relationship between two variables.

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

where $\bar{x}$ is the sample mean of the predictor variable
s_x is the sample standard deviation of the predictor variable
$\bar{y}$ is the sample mean of the response variable
s_y is the sample standard deviation of the response variable
n is the sample size

I know that's quite a mouthful, but we'll be using technology to calculate it. Here's a quick summary of some of the properties of the linear correlation coefficient, as described in your text.

Properties of the Linear Correlation Coefficient

1. The linear correlation coefficient is always between -1 and 1.
2. If r = +1, there is a perfect positive linear relation between the two variables.
3. If r = -1, there is a perfect negative linear relation between the two variables.
4. The closer r is to +1, the stronger is the evidence of positive association between the two variables.
5. The closer r is to -1, the stronger is the evidence of negative association between the two variables.
6. If r is close to 0, there is little or no evidence of a linear relation between the two variables - this does not mean there is *no* relation, only that there is no *linear* relation.

Next, I'd like you to visit two web sites that offer Java applets. These will help you interact with data to get a sense of the linear correlation coefficient.

Example 2

This first applet was created for use with another textbook, Introduction to the Practice of Statistics, by David S. Moore and George P. McCabe.

The applet is designed to allow you to add your own points and watch it calculate the linear correlation coefficient for you. (There are other capabilities as well, but we'll get to those in the next section.)

Applet: Correlation and Regression

Example 3

This second applet was designed as part of the Rossman/Chance Applet Collection at California Polytechnic State University.

This applet generates scatter plots for you and asks you to guess the correlation for each. Click on "New Sample" to start, enter your answer, and then "Enter" to see if you're correct.

Applet: Guess the Correlation

Example 4

Let's try to calculate a correlation ourselves. To make our data set a bit more manageable, let's use the before/after data from Example 1 in Section 4.1, but let's just use the first 8 as our sample.

before	after	$\frac{x_i-\bar{x}}{s_x}$	$\frac{y_i-\bar{y}}{s_y}$	$\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$
86	98	0.97865	0.78657	0.76978

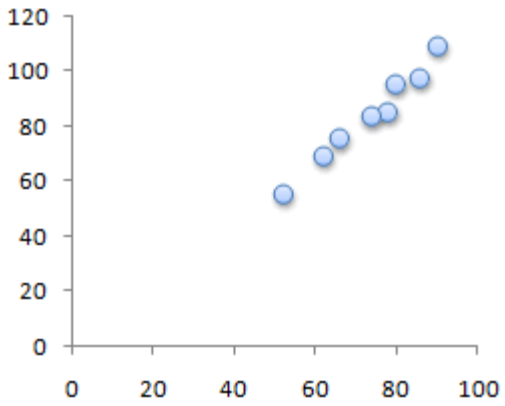
62	70	-0.90036	-0.84484	0.76065
52	56	-1.68327	-1.66054	2.79514
90	110	1.29181	1.48575	1.91931
66	76	-0.58719	-0.49525	0.29080
80	96	0.50890	0.67004	0.34098
78	86	0.35231	0.08740	0.03079
74	84	0.03915	-0.02913	-0.00114
				6.90632

Using computer software, we find the following values:

$\bar{x} = 73.5$
$s_x \approx 12.77274$
$\bar{y} = 84.5$
$s_y \approx 17.16308$

Note: We don't want to round these values here, since they'll be used in the calculation for the correlation coefficient - only round at the very last step.

Since we have a sample size of 8, we divide the sum by 7 and get a correlation factor of **0.99**. That seems fairly high, but looking at the scatter plot (below), we can see why it's so strong.



Technology

Here's a quick overview of the formulas for finding the linear correlation coefficient in StatCrunch.

1. Select **Stat** > **Regression** > **Simple Linear**
2. Select the predictor variable for X & the response variable for Y
3. Select **Calculate**

You can also go to the video page for links to see videos in either Quicktime or iPod format.

Here's one for you to try.

Example 5

Researchers at General Motors collected data on 60 U.S. Standard Metropolitan Statistical Areas (SMSA's) in a study of whether air pollution contributes to mortality. The dependent variable for analysis is age adjusted mortality (called "Mortality").

The data below show the age adjusted mortality rate (deaths per 100,000) and the sulfur dioxide polution potential. Use StatCrunch to calculate the linear correlation coefficient. Round your answer to three digits.

City	Mortality*	SO_2 potential**
Akron, OH	921.87	59
Albany, NY	997.87	39
Allentown, PA	962.35	33
Atlanta, GA	982.29	24
Baltimore, MD	1071.29	206
Birmingham, AL	1030.38	72
Boston, MA	934.7	62
Bridgeport, CT	899.53	4
Buffalo, NY	1001.9	37
Canton, OH	912.35	20
Chattanooga, TN	1017.61	27
Chicago, IL	1024.89	278
Cincinnati, OH	970.47	146
Cleveland, OH	985.95	64
Columbus, OH	958.84	15
Dallas, TX	860.1	1
Dayton, OH	936.23	16
Denver, CO	871.77	28
Detroit, MI	959.22	124
Flint, MI	941.18	11
Fort Worth, TX	891.71	1
Grand Rapids, MI	871.34	10
Greensboro, NC	971.12	5
Hartford, CT	887.47	10
Houston, TX	952.53	1
Indianapolis, IN	968.67	33
Kansas City, MO	919.73	4
Lancaster, PA	844.05	32
Los Angeles, CA	861.26	130
Louisville, KY	989.26	193
Memphis, TN	1006.49	34
Miami, FL	861.44	1
Milwaukee, WI	929.15	125
Minneapolis, MN	857.62	26
Nashville, TN	961.01	78
New Haven, CT	923.23	8
New Orleans, LA	1113.16	1

New York, NY	994.65	108
Philadelphia, PA	1015.02	161
Pittsburgh, PA	991.29	263
Portland, OR	893.99	44
Providence, RI	938.5	18
Reading, PA	946.19	89
Richmond, VA	1025.5	48
Rochester, NY	874.28	18
St. Louis, MO	953.56	68
San Diego, CA	839.71	20
San Francisco, CA	911.7	86
San Jose, CA	790.73	3
Seattle, WA	899.26	20
Springfield, MA	904.16	20
Syracuse, NY	950.67	25
Toledo, OH	972.46	25
Utica, NY	912.2	11
Washington, DC	967.8	102
Wichita, KS	823.76	1
Wilmington, DE	1003.5	42
Worcester, MA	895.7	8
York, PA	911.82	49
Youngstown, OH	954.44	39

* Age Adjusted Mortality (deaths per 100,000)
** Sulfer Dioxide pollution potential

Source: StatLib

[reveal answer]

Section 4.2: Least-Squares Regression

Objectives

By the end of this lesson, you will be able to...

1. find the least-squares regression (LSR) line
2. use the LSR line to make predictions
3. interpret the slope and y-intercept of the LSR line

Because we'll be talking about the *linear* relationship between two variables, we need to first do a quick review of lines.

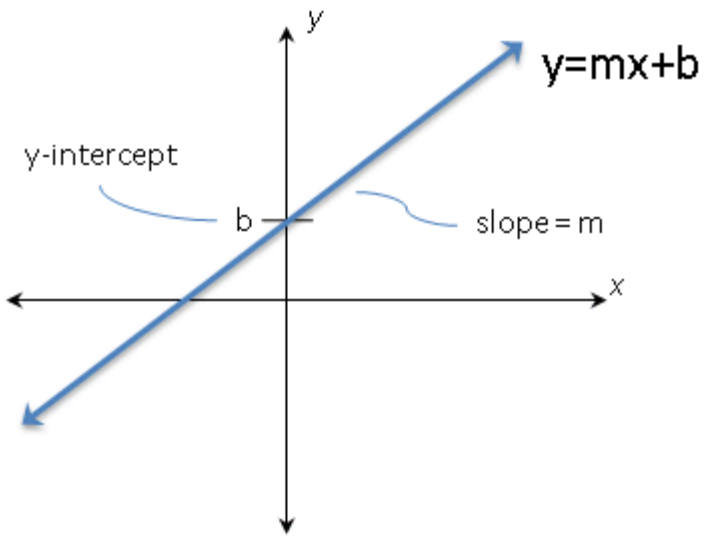
The Slope and Y-intercept

If there's one thing we all remember about lines, it's the **slope-intercept form** of a line:

The **slope-intercept form** of a line is

$$y = mx + b$$

where m is the slope of the line and b is the y-intercept.

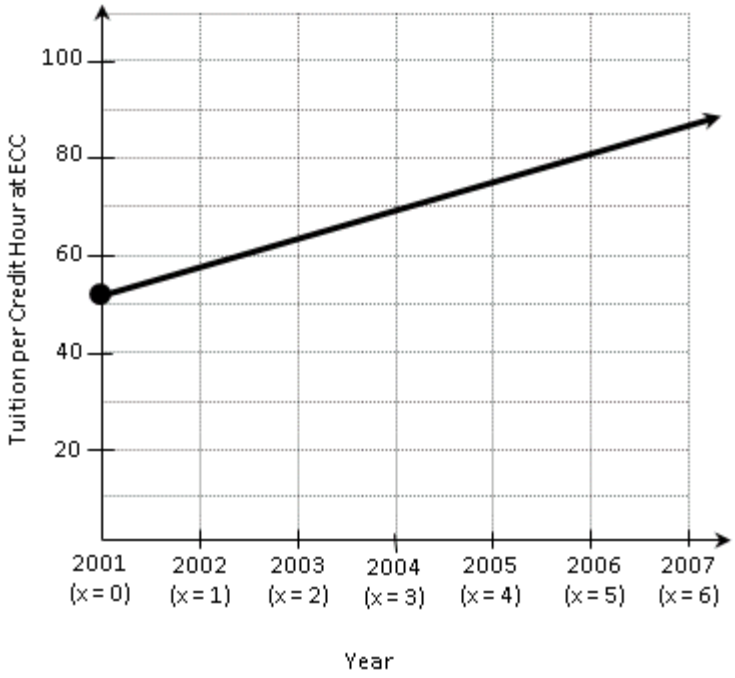


Knowing the form isn't enough, though. We also need to know what each part *means*. Let's start with the slope. Most of us remember the slope as "rise over run", but that only helps us graph lines. What we really need to know is what the slope represents in terms of the original two variables. Let's look at an example to see if we can get the idea.

Example 1

The equation T = 6x + 53 roughly approximates the tuition per credit at ECC since 2001. In this case, x represents the number of years since 2001 and T represents the tuition amount for that year.

The graph below illustrates the relationship.



In this example, we can see that both the 6 and the 53 have very specific meanings:

The 6 is the *increase per year*. In other words, for every additional year, the tuition increases $6.

The 53 represents the *initial tuition*, or the tuition per credit hour in 2001.

As we progress into the relationship between two variables, it's important to keep in mind these meanings behind the slope and y-intercept.

Finding the Equation for a Line

Another very important skill is finding the equation for a line. In particular, it's important for us to know how to find the equation when we're given two points.

A very useful equation to know is the point-slope form for a line.

The **point-slope form** of a line is

$$y - y_1 = m(x - x_1)$$

where m is the slope of the line and (x_1, y_1) is a point on the line.

Let's practice using this form to find an equation for the line.

Example 2

In Example 1 from section 4.1, we talked about the relationship between student heart rates (in beats per minute) before and after a brisk walk.

before	after	before	after	before	after
86	98	58	128	60	70
62	70	64	74	80	92
52	56	74	106	66	70
90	110	76	84	80	92
66	76	56	96	78	116
80	96	72	82	74	114
78	86	72	78	90	116
74	84	68	90	76	94

Let's highlight a pair of points on that plot and use those two points to find an equation for a line that might fit the scatter diagram.

Using the points (52, 56) and (90, 116), we get a slope of

$$m = \frac{116-56}{90-52} = \frac{60}{38} \approx 1.58$$

So an equation for the line would be:

$$y - y_1 = m(x - x_1)$$

$$y - 56 = 1.58(x - 52)$$

$$y - 56 = 1.58x - 82.16$$

$$\mathbf{y = 1.58x - 26.16}$$

It's interesting to note the meanings behind the slope and y-intercept for this example. A slope of 1.58 means that for every additional beat per minute before the brisk walk, the heart rate after the walk was 1.58 faster.

The y-intercept, on the other hand, doesn't apply in this case. A y-intercept of -26.16 means that if you have 0 beats per minute before the walk, you'll have -26.16 beats per minute after the walk. ?!?!

> This brings up a very important point - models have limitations. In this case, we say that the y-intercept is **outside the scope of the model**.

Now that we know how to find an equation that sort of fits the data, we need a strategy to find the best line. Let's work our way up to it.

Residuals

Unless the data line up perfectly, any line we use to model the relationship will have an error. We call this error the *residual*.

The **residual** is the difference between the observed and predicted values for y:

residual = observed y - predicted y

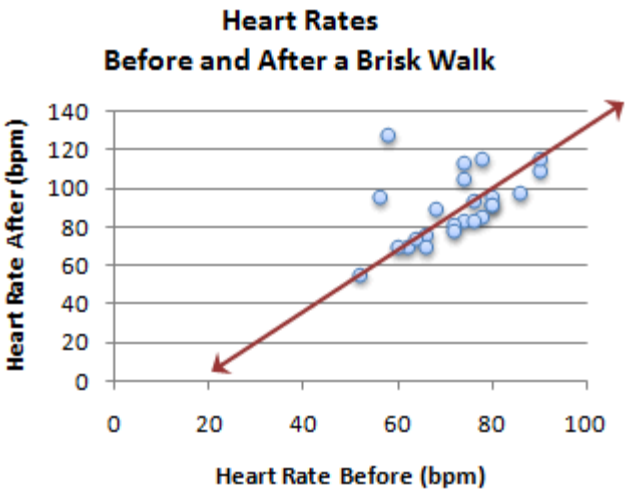
residual = $y - \hat{y}$

Notice here that we used the symbol $\hat{y}$ (read "y-hat") for the predicted. This is standard notation in statistics, using the "hat" symbol over a variable to note that it is a predicted value.

Example 3

Let's again use the data from Example 1 from section 4.1. In Example 2 from earlier this section, we found the model:

$\hat{y}$ = 1.58x - 30.16



Let's use this model to predict the "after" heart rate for a particular students, the one whose "before" heart rate was 86 beats per minute.

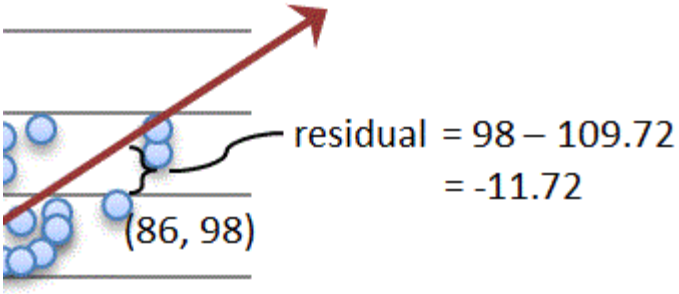
The predicted heart rate, using the model above, is:

$\hat{y}$ = 1.58(86) - 26.16 = 109.72

Using that predicted heart rate, the residual is then:

residual = $y - \hat{y}$ = 98 - 109.72 = -11.72

Here's that residual if we zoom in on that particular student:



Notice here that the residual is negative, since the predicted value was more than the actual observed "after" heart rate.

The Least-Squares Regression (LSR) line

So how do we determine which line is "best"? The most popular technique is to make the sum of the squares of the residuals as small as possible. (We use the squares for much the same reason we did when we defined the variance in Section 3.2.) The method is called the *method of least squares*, for obvious reasons!

The Equation for the Least-Squares Regression line

The equation of the least-squares is given by

$$\hat{y} = b_1 x + b_0$$

where

$b_1 = r \cdot \frac{s_y}{s_x}$ is the **slope** of the least-squares regression line

and

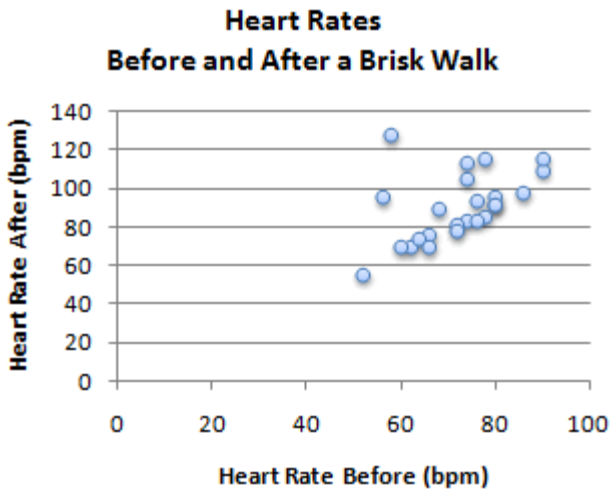
$b_0 = \bar{y} - b_1 \bar{x}$ is the **y-intercept** of the least squares regression line

Let's try an example.

Example 4

Let's again use the data from Example 1 in Section 4.1, but instead of just using two points to get a line, we'll use the method of least squares to find the Least-Squares Regression line.

before	after	before	after	before	after
86	98	58	128	60	70
62	70	64	74	80	92
52	56	74	106	66	70
90	110	76	84	80	92
66	76	56	96	78	116
80	96	72	82	74	114
78	86	72	78	90	116
74	84	68	90	76	94



Using computer software, we find the following values:

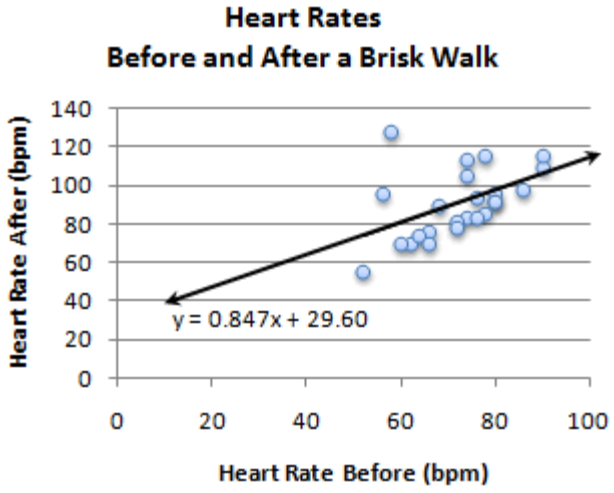
$\bar{x} \approx 72.16667$
$s_x \approx 10.21366$
$\bar{y} = 90.75$
$s_y \approx 17.78922$
$r \approx 0.48649$

Note: We don't want to round these values here, since they'll be used in the calculation for the correlation coefficient - only round at the very last step.

Using the formulas for the LSR line, we have

$$\hat{y} = 0.8473x + 29.6018$$

(A good general guideline is to use 4 decimal places for the slope and y-intercept, though there is no strict rule.)



One thought that may come to mind here is that this doesn't really seem to fit the data as well as the one we did by picking two points! Actually, it does do a much better job fitting ALL of the data as well as possible - the previous line we did ourselves did not address most of the points that were above the main cluster. In the next section, we'll talk more about how outliers like the (58, 128) point far above the rest can affect a model like this one.

Technology

Here's a quick overview of how to find the Least-Squares Regression line in StatCrunch.

1. Select **Stat** > **Regression** > **Simple Linear**
2. Select the predictor variable for X & the response variable for Y
3. Select **Calculate**

The fourth line shows the equation of the regression line. Note that it will not have x and y shown, but rather the names that you've given for x and y. For example:

Avg. Final Grade = 88.73273 - 2.8272727 Num. Absences


You can also go to the video page for links to see videos in either Quicktime or iPod format.

Section 4.3: Diagnostics on the Least-Squares Regression Line

Objectives

By the end of this lesson, you will be able to...

1. compute and interpret the coefficient of determination
2. perform residual analysis on a regression model
3. determine if a linear regression model is appropriate
4. identify influential observations

Before we can go on, we need to first determine two things:

1. Does our model do a good job of predicting the results?
2. Is a linear model appropriate?

We'll have two key factors to help us answer these questions. The first is called the coefficient of determination.

The Coefficient of Determination

The **coefficient of determination**, R^2, is the percent of the variation in the response variable (y) that can be explained by the least-squares regression line.

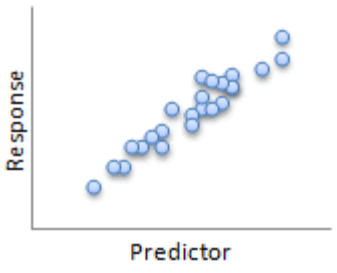
Looking at the definition, we can see that a higher R^2 is better - the LSR line does a *better* job of explaining the variation in the response variable.

Your textbook has a relatively detailed explanation of how R^2 is calculated, so I won't repeat it here. If you'd like to, you can also view the derivation on this page from Wikipedia. (Not that Wikipedia is a perfect source, but this particular page is accurate.)

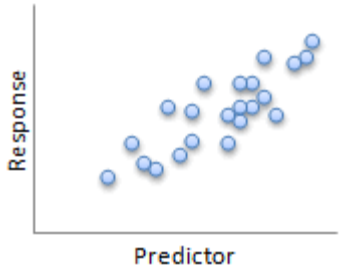
Here's the end result, which shouldn't come as too much of a surprise:

$$R^2 = r^2$$

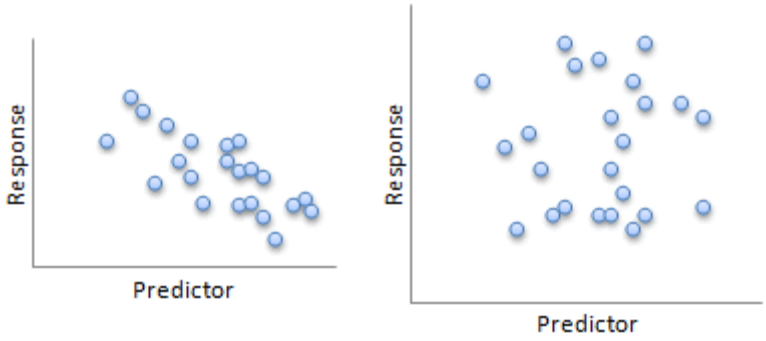
Here are some examples:



$R^2 = 92.1\%$

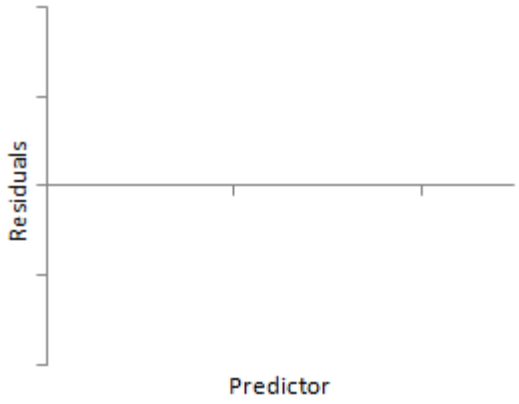


$R^2 = 70.0\%$



R^2 = 53.4% No R^2 = 13.5%

The second step in residual analysis is using the residuals to determine if a linear model is appropriate. We do this by creating **residual plots**. A residual plot is a scatter diagram with the predictor as the x and the corresponding residual as the y.

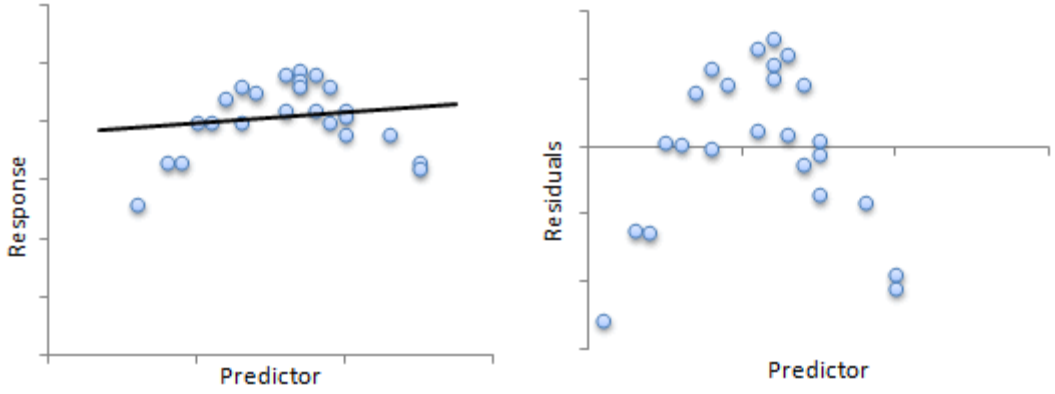


In general, there are three things to watch out for in a residual plot:

1. a pattern in the residuals
2. increasing or decreasing spread
3. influential observations

Patterned Residuals

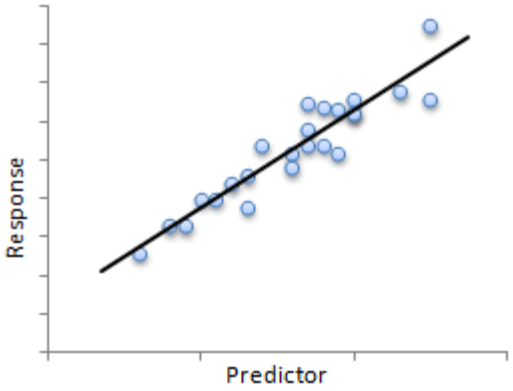
If a residual plot shows a discernable pattern (like a curve), then the predictor and response variables may not be linearly related.

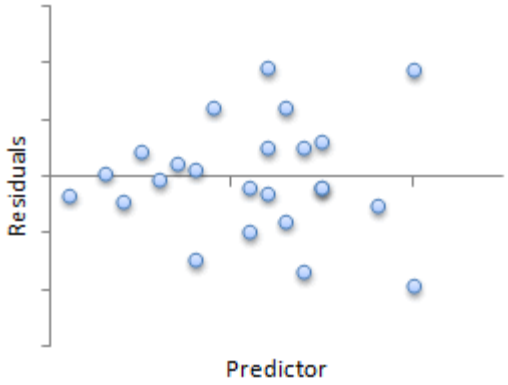


The LSR line clearly does not fit. The residuals show an obvious pattern.

Increasing or Decreasing Spread

If a residual plot shows the spread increasing or decreasing, then a strict requirement of the linear model is related. (This strict requirement is called *constant error variance* - the error must be evenly spread.)



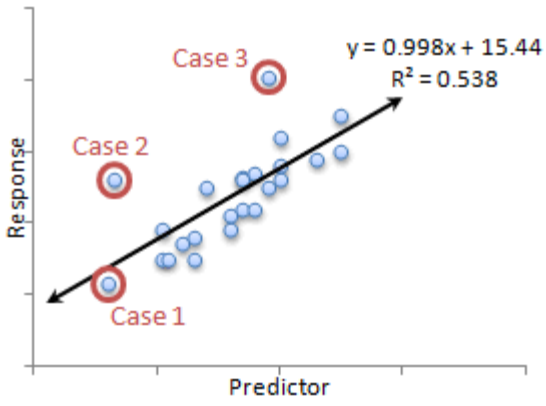


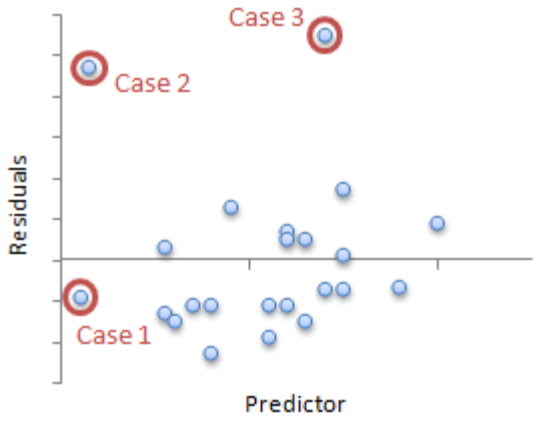
The LSR line seems to be a great fit. The residuals start very small, but increase as the predictor variable increases - this model does not have constant error variance.

Outliers and Influential Observations

The next point we need to consider is the existence of outliers and influential observations. We can think of an **outlier** as an observation that doesn't seem to fit the rest of the data. **Influential observations** are similar, but with the added quality that their existence significantly affects the slope and/or y-intercept of the line.

Consider the scatter diagram shown below, along with its corresponding residual plot:

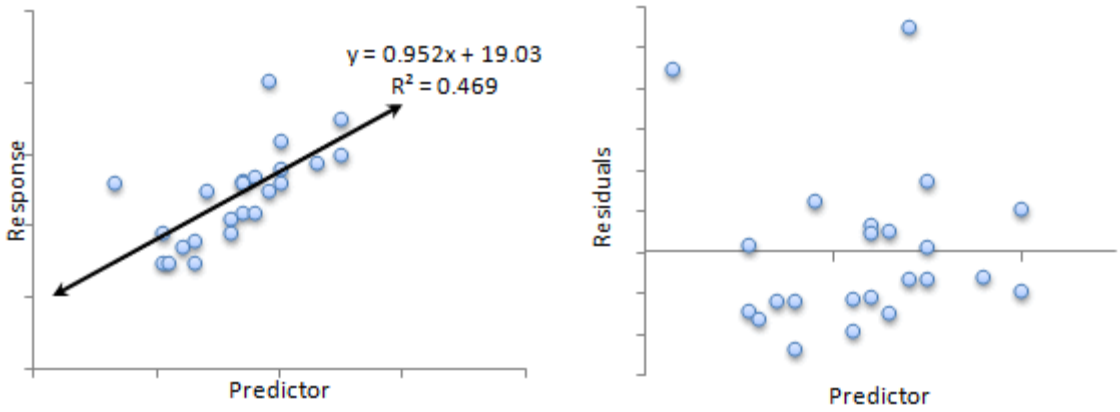




Let's consider the three cases indicated.

Case 1:

This case is considered an outlier because it's x-value is much lower than all but one of the other observations. To determine if it's an influential observation, we'll need to recalculate the LSR line without that observation included. Here are the results:

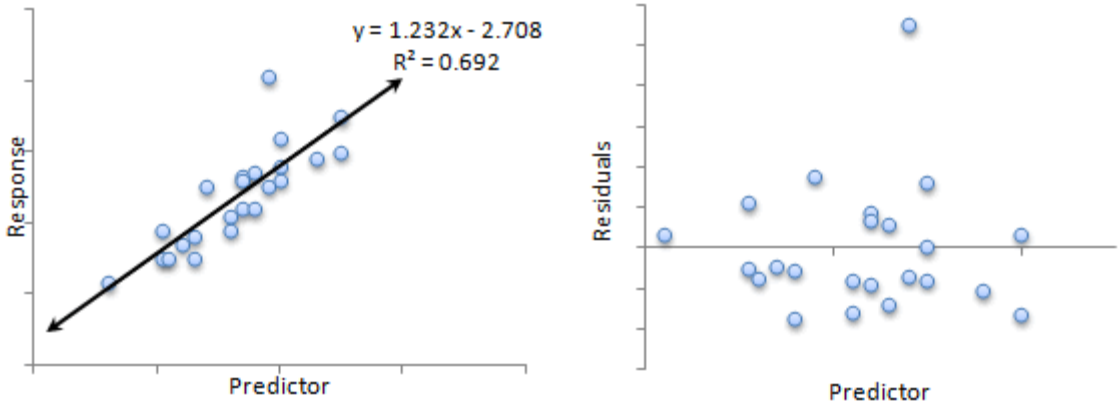


We can see that while there are some changes in the slope and y-intercept, both are reasonably similar to what they were with Case 1 included. In this case, we would describe Case 1 as an outlier, but not an influential observation.

An interesting point to note, though, is the decreased R^2 value. The implication is that Case 1 actually *strengthened* the correlation. Think of that point pulling the line "tighter".

Case 2:

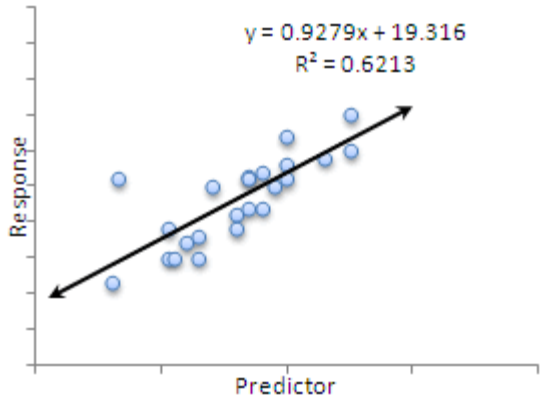
Looking back at the original diagram, it seems as though Case 2 should be influential, because there are not many values near it to minimize its effect on the LSR line. Here's the output from computer software:

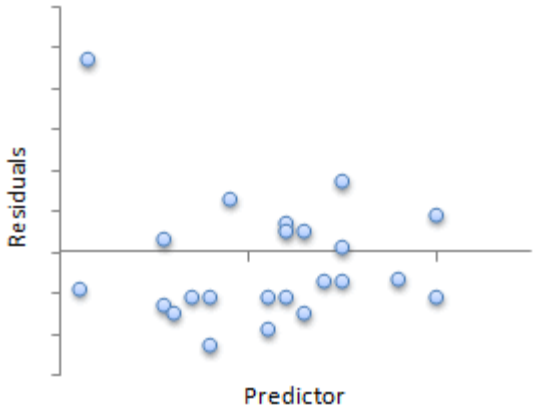


Here we can clearly see that both the slope and y-intercept (as well as R^2) are significantly different, so we would definitely characterize Case 2 as an influential observation.

Case 3:

Unlike in Case 2, this particular observation has others near it to minimize its effect, so it most likely will not be influential. Here's the output from computer software:



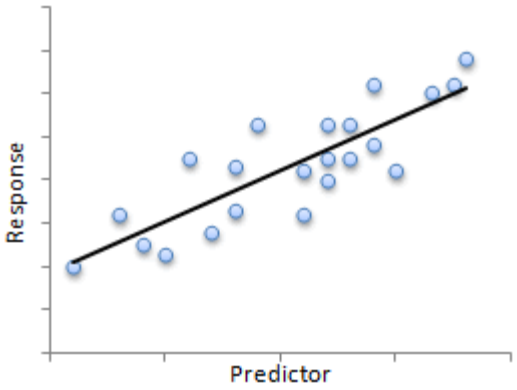


Comparing those to the original values, we can indeed see that the slope and y-intercept are both relatively similar. So while this value is an outlier (as seen very clearly on the earlier residual plot), it is not influential.

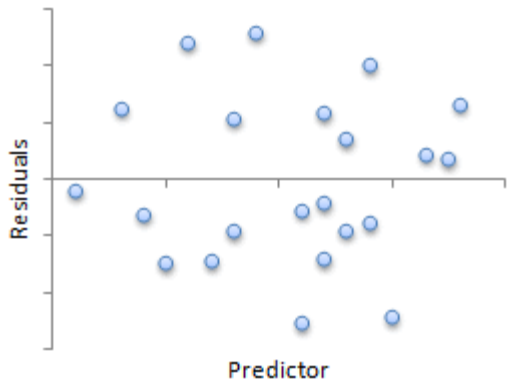
> If you find your data contain outliers or influential observations, and those observations cannot be removed (because they are due to data entry errors or similar) you have only a couple options. The primary option is to **collect more data to minimize their impact**. The second is to use analysis methods that minimize the effect of outliers. Unfortunately, those techniques are fairly advanced and outside the scope of this course.

When a Linear Model is Appropriate

Sometimes it can be difficult to determine if any of the three above conditions have been violated, but here's a good example of a situation where a linear model does seem appropriate.



The LSR line seems to fit the data.



The residuals are evenly spread above and below zero, there is no discernable pattern, and there are no outliers.

Technology

Here's a quick overview of how to create a residual plot in StatCrunch.

1. Select **Stat > Regression > Simple Linear**
2. Set the X-Variable and Y-Variable and press **Next**.
3. Select **Save residuals** (optional) and press **Next**.
4. Select the options you want - make sure to select "Residuals vs. X-values" is the residual plot.
5. Press **Calculate**.

6. The output will show your regression analysis. On the bottom, press **Next** to see any graphics.

You can also go to the video page for links to see videos in either Quicktime or iPod format.

Section 4.4: Contingency Tables and Association

Objectives

By the end of this lesson, you will be able to...

1. compute the marginal distribution of a variable
2. construct a conditional distribution of a variable
3. use the conditional distribution to identify association between categorical data

In sections 4.1-4.3, we studied relationships between two quantitative variables. We learned that we could quantify the strength of the linear relationship with the correlation.

What about qualitative (categorical) variables, though? For example, suppose we consider a survey given to 82 students in a Basic Algebra course at ECC, with the following responses to the statement "I enjoy math."

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
Men	9	13	5	2	1
Women	12	18	11	6	5

How do we study this relationship? Is there a way to tell if gender and whether the student enjoys math? In fact, there is! Like usual, though, we need a bit of background work first.

Contingency Tables

A **contingency table** relates two categories of data. In the example above, the relationship is between the gender of the student and his/her response to the question.

A **marginal distribution** of a variable is a frequency or relative frequency distribution of either the row or column variable in the contingency table.

Example 1

If we consider the previous example:

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
Men	9	13	5	2	1
Women	12	18	11	6	5

The entire table is referred to as the **contingency table**.

The **marginal distribution** for gender removes the effect of whether or not the student enjoys math:

	Strongly				Strongly

	Agree	Agree	Neutral	Disagree	Disagree	**Total**
Men	9	13	5	2	1	**30**
Women	12	18	11	6	5	**52**

Whereas, the **marginal distribution** for whether or not the student enjoys math removes the effect of gender:

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
Men	9	13	5	2	1
Women	12	18	11	6	5
Total	**21**	**31**	**16**	**8**	**6**

We can also create a **relative frequency marginal distribution**, which, as expected, is simply relative frequencies rather than frequencies.

Example 2

The combined **relative frequency marginal distributions** would look like this:

	SA	A	N	D	SD	Total
Men	9	13	5	2	1	30/82 ≈ 0.37
Women	12	18	11	6	5	52/82 ≈ 0.63
Total	21/82 ≈ 0.26	31/82 ≈ 0.39	16/82 ≈ 0.20	8/82 ≈ 0.10	6/82 = 0.07	1

Let's consider the frequency marginal distributions from Example 2.

Example 3

	SA	A	N	D	SD	Total
Men	9	13	5	2	1	30
Women	12	18	11	6	5	52
Total	21	31	16	8	6	82

We might now be interested in comparing the two variables. For example:

a. What proportion of women strongly agreed with the statement "I enjoy math"?
b. What proportion of women disagreed?
c. What proportion of men were neutral?
d. What proportion of men strongly agreed?

Solution:

a. There were 12 women who strongly agreed, and 52 women in all, so 12/52 ≈ 0.23
b. Similarly, there were 6 women who disagreed, and 52 overall, so 6/52 ≈ 0.12
c. 5/30 ≈ 0.17
d. 9/30 ≈ 0.30

If we completed the table in this fashion, we get something called a **conditional distribution**.

> A **conditional distribution** lists the relative frequency of each category of variable, given a specific value of the other variable in the contingency table.

For another explanation of marginal and conditional distributions, watch this YouTube video:

Example 4

The conditional distribution of how the students feel about math by gender would be as follows:

	SA	A	N	D	SD	Total
Men	9/30 ≈ 0.30	13/30 ≈ 0.43	5/30 ≈ 0.17	2/30 ≈ 0.07	1/30 ≈ 0.03	30/30 = 1
Women	12/52 ≈ 0.23	18/52 ≈ 0.35	11/52 ≈ 0.21	6/52 ≈ 0.12	5/52 ≈ 0.10	52/52 = 1

Note: The row totals sometimes do not add up to 1 due to rounding.

Another way to think of this distribution is that it's the distribution of how students feel *for each gender*. That's what the "by gender" indicates.

Example 5

The conditional distribution of gender by how the student feels would be:

	SA	A	N	D	SD
Men	9/21 ≈ 0.43	13/31 ≈ 0.42	5/16 ≈ 0.31	2/8 = 0.25	1/6 ≈ 0.17
Women	12/21 ≈ 0.57	18/31 ≈ 0.58	11/16 ≈ 0.69	6/8 = 0.75	5/6 ≈ 0.83
Total	21/21 = 1	31/31 = 1	16/16 = 1	8/8 = 1	6/6 = 1

Using Conditional Distributions to Identify Association

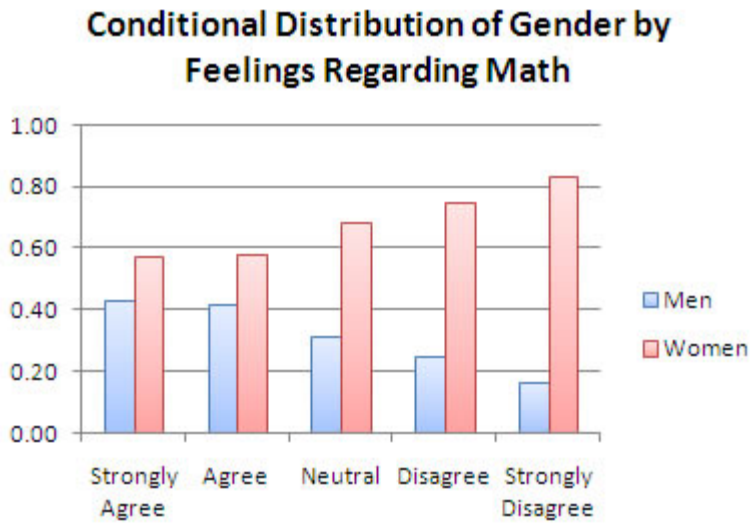
One thing we can use conditional distributions for is to identify an association between qualitative variables. The best way to do this is a side-by-side bar graph. We'll illustrate with the same data we've been using.

Example 6

In Example 5, we found the conditional distribution of gender by how the student feels regarding math:

	SA	A	N	D	SD
Men	9/21 ≈ 0.43	13/31 ≈ 0.42	5/16 ≈ 0.31	2/8 = 0.25	1/6 ≈ 0.17
Women	12/21 ≈ 0.57	18/31 ≈ 0.58	11/16 ≈ 0.69	6/8 = 0.75	5/6 ≈ 0.83
Total	21/21 = 1	31/31 = 1	16/16 = 1	8/8 = 1	6/6 = 1

Since it's difficult to gain much from this table alone, a good way to analyze this would be to make a side-by-side bar graph.



From the graph, we can see that there definitely appear to be some differences between the different responses. The proportions of responses were similar for both "Strongly Agree" and "Agree", but very different for "Neutral" and "Disagree". As we go down the scale, the proportion of the responses that are by women increases.

We might conclude, then, that men tend to enjoy math more than women.

> One thing we *can't* conclude is that their gender *caused* them to not enjoy math. We've only done an observational study, so we can only claim *association*, not *causation*.

One question you might have as a result of this is, "How do we know when it's different enough from equal to say

that there might be a relationship?" It's a very good question. In order to draw a fine line, we'll need a hypothesis test, which we won't see until we get to Chapter 12.

<< previous section | next section >>

1 2 3 4 5 6 7 8 9 10 11 12 13
