Chapter 2 - Relationships

Read Chapter 2 first, then read the notes and try the WEBCT assignment questions. If you need more practice, try the practice questions with answers available on the web.

Introduction to Chapter 2

Scientists often focus on more than one variable in their research. Researchers often want to know how two variables are related. We start this chapter with graphs and then move to numerical summaries of more than one variable. Straight line relationships are introduced along with regression and correlation methods.

View the Video – Describing Relationships

Relationships Between Two Variables

We have two variables X and Y in a data set of n observations:

$X_1 \quad Y_1$

$X_2 \quad Y_2$

. .

. .

$X_n \quad Y_n$

Question:

Is there a relationship between X &Y? If the data suggests there is - What Kind? Listed below are several examples of two variable problems. The rest of the chapter will show you how to investigate the relationship between these types of variables.

Example 1: Is there a relationship between smoking (a categorical variable) and life length (a quantitative variable)? We code the X variable as follows:

X = 1 (smoker) 0 (non-smoker)	**categorical**
Y = life of length in years	**quantitative**

Example 2: Is there a relationship between exposure to high tension wires and incidence of childhood leukaemia?

Here both X and Y are coded as follows:

X = 1	if one had lived within a mile of high voltage wires during ages of 0-5 for more than 1 month
= 0	otherwise
Y = 1	if contracted leukemia before age 5
= 0	otherwise

In this case, X and Y are **both categorical**.

Example 3: Is there a relationship between endurance and consumption caffeine?

X = quantity of caffeine ingested before test in ml **(quantitative)**

Y = length of time subject could continue on treadmill run at fixed speed after ingestion **(quantitative)**

In this example, both X and Y are **quantitative variables**. To see if there is a relationship between X and Y, we plot the data.

Scatterplots are for pairs of quantitative variables (i.e., for X quantitative AND for Y quantitative).

You plot ordered pairs (X_1,Y_1), (X_2,Y_2) ... (X_n,Y_n) from the data set.

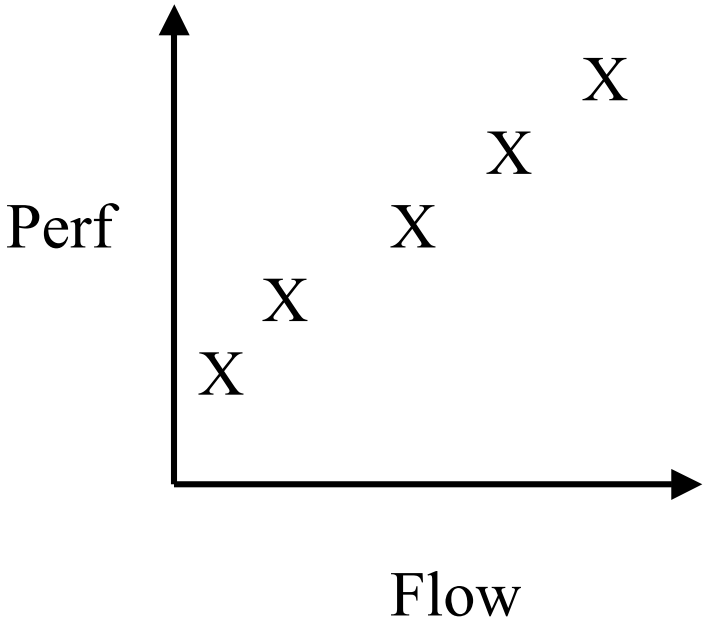
Example: Let's examine the performance of animals when given stimulant, this is common in neuro-psychology.

The two variables are:
Flow = flow rate of the drug in mililiters (X)
Perf = performance measure in ekg (Y) an electrical measure of neuron activity.

The data are given below for 5 cases. For example, case one (.31, .82) etc.

Flow	Perf
.31	.82
.85	1.95
1.26	2.18
2.47	3.01
3.75	6.07

Next, plot Flow vs Perf to see if there is a relationship.



It seems clear there is a relationship that as Flow increases so does Performance. The graph of the points looks like a positive, linear trend. The researcher wants to see how Flow (X) explains Perf (Y) the response variable.

Another example that examines how the age of a child in months is related to its height in centimeters is given below.

TABLE 2.7 Mean height of Kalama children

Age x in months	Height y in centimeters
18	76.1
19	77.0
20	78.1
21	78.2
22	78.8
23	79.7
24	79.9
25	81.1
26	81.2
27	81.8
28	82.8
29	83.5

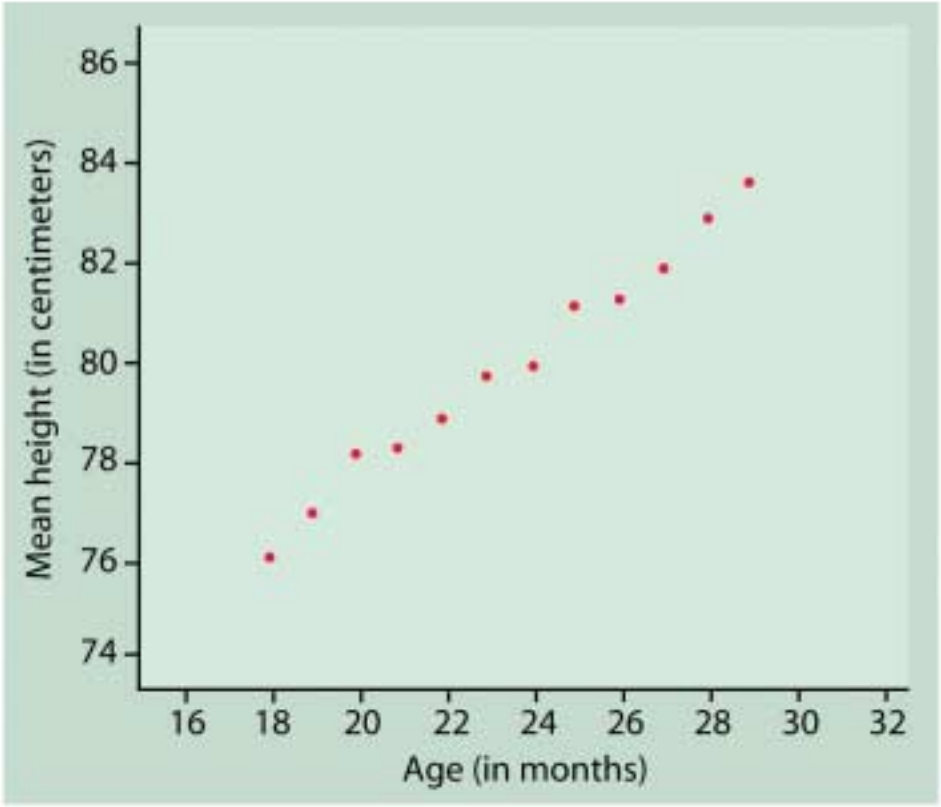


Figure 2.11 Mean height of children in Kalama, Egypt, plotted against age from 18 to 29 months, from table 2.7.

X - Flow rate plays the role of **independent** variable or **explanatory** variable or **predictor** variable.(age in example 2)

Y - Performance measure plays role of **dependent** variable or **response** variable.(height in example 2)

* Note, it is not always clear which is which, see below.

Example: X = wife's years of education
Y= husband's years of education

Does a wife's years of education explain the husband's or the husband's explain the wife's?

Linear Relationships

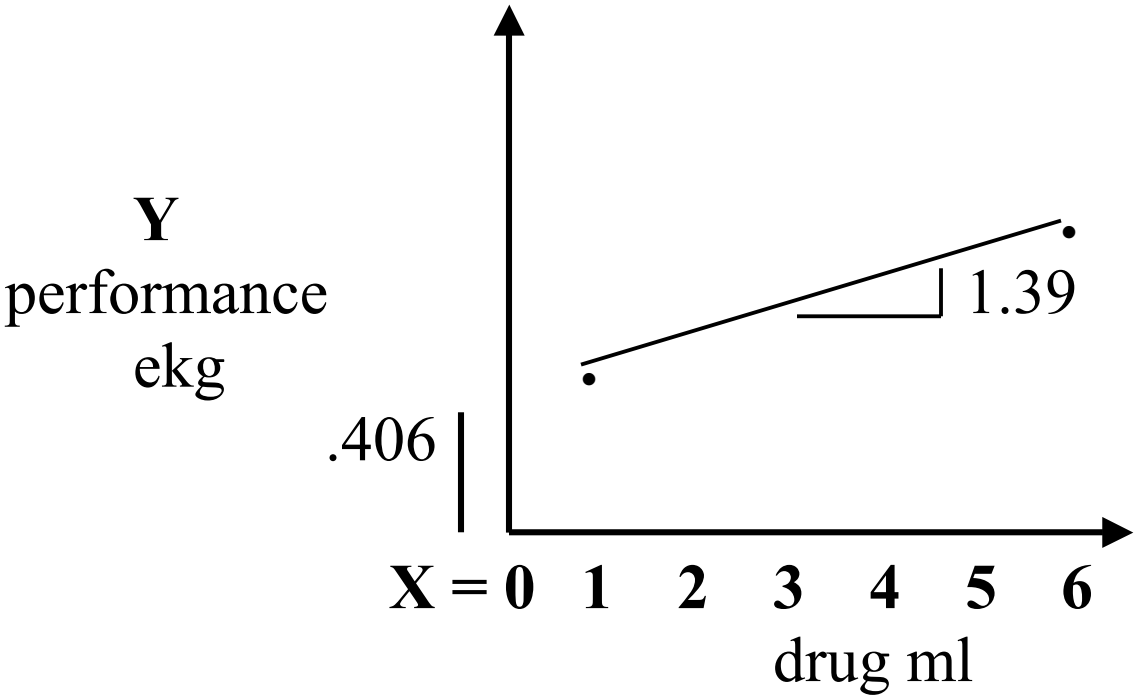
What kind of relationship is it that relates X to Y? The simplest is the linear relationship as given by the equation:

$$\mathbf{Y = a+bX}$$

… for some constants (numbers) a, b.
(you may have used y=mx+b in high school)

Example: Lets use the data from our previous example with Flow (X) and Perf (Y). Let {a=.406} then {b=1.39}.Our linear equation is: Y = .406 + 1.39X OR Perf. = .406 + 1.39Flow

We say “a” is the **y–intercept**. This is the point where the line crosses the y-axis when X=0. We call “b” the **slope** since this is the amount the performance measure (Y) increases when the amount of drug is increased 1 *ml*. The graph below helps you see what these numbers tell us about the linear relationship.



In summary: a = .406 (y-intercept)
b = 1.39 (slope)

In words, the slope of 1.39 means that performance increases by 1.39 ekg per 1 ml drug **OR** a 1 unit increase in Flow (X) gives a 1.39 unit increase in Perf (Y).

See another example below where age is used to help explain height, note for a given age you can predict height using the least squares line.

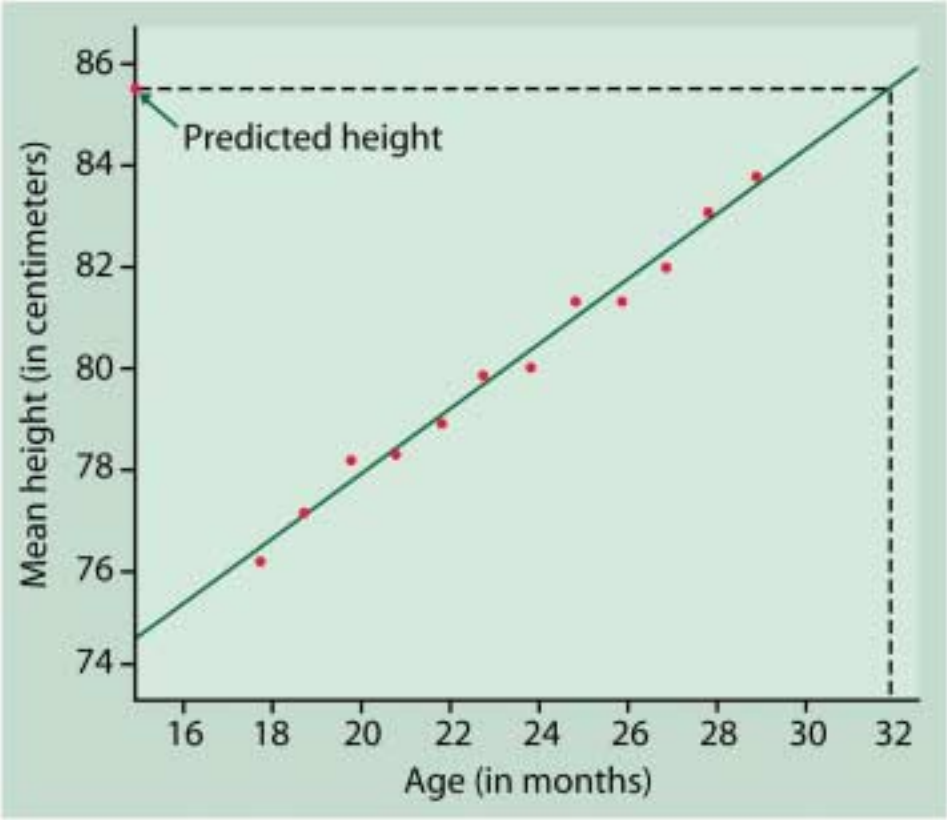
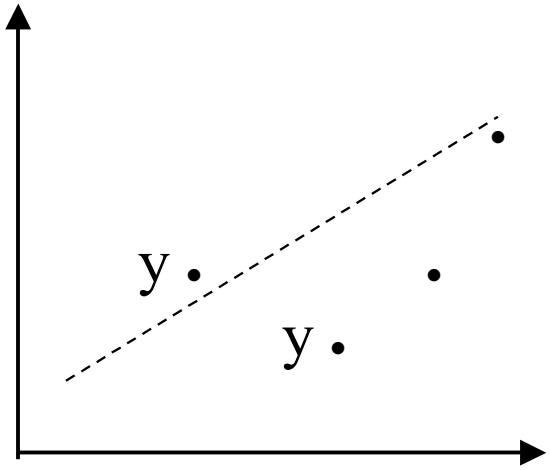


Figure 2.12 The regression line fitted to the Kalama data and used to predict height at age 32 months.

Can a relationship be described as approximately linear?

For data we have collected we want to find the best fitting line. How do we find "a" and "b" such that the line is 'best'? (i.e., **Y = a + bX**)



Scientist use the optimization procedure called **least squares**. It is a procedure where we find "a" and "b" that minimizes:

$$\sum_{i=1}^{n} (Y_i - (a + bx_i))^2$$

The key part of the formula is that:

$$\textbf{(observed Y – predicted Y)}^2$$

This vertical distance for all the data is minimized for the selected "a" and "b". We use a subscript ***i*** to keep track of all the data points. The sum of squared vertical distances between $(X_i \quad Y_i)$ and $(X_i, a+bX_i)$ is minimized using least squares. We define a **residual** (***e***) as the vertical distance between the data (y) and that predicted by the line.

$$\textbf{residual} = \text{observed Y} - \text{predicted Y}$$

$$e_i = Y - \hat{Y}$$

Remember: The vertical distance is called a residual (***e***). The a,b that minimize the residuals for all the data is given by the following formula:

Solutions

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\,\bar{x}$$

"a" and "b" are called **least squares estimates**

The **least squares line** or **least squares regression line** is given by:

$$\mathbf{Y = a + bX}$$

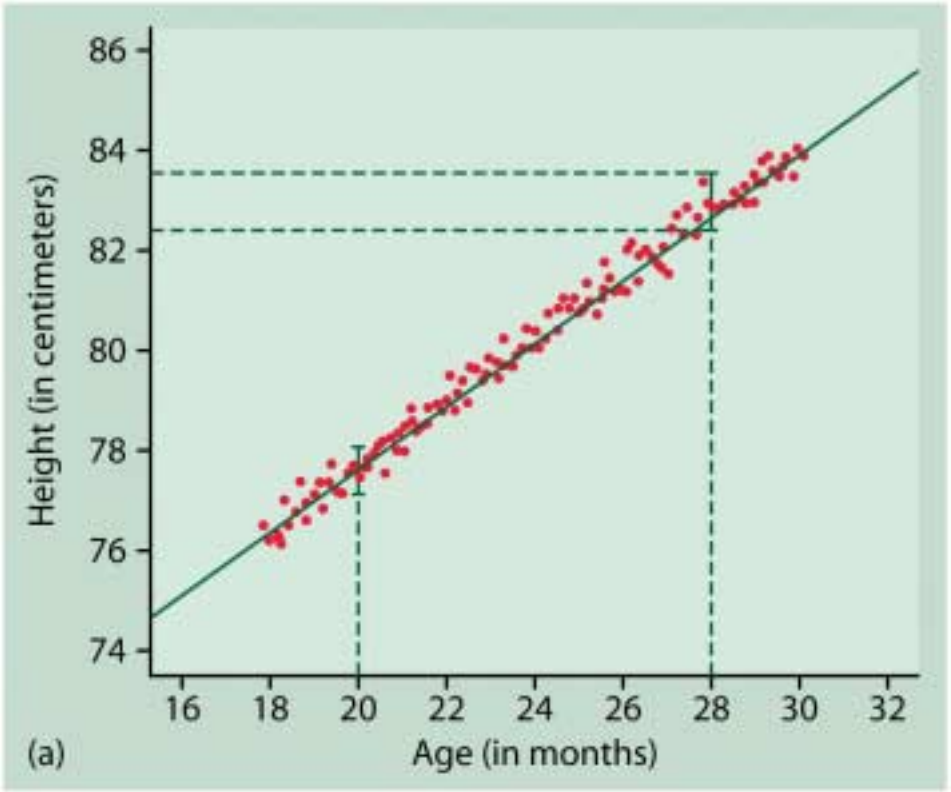
We use the subscript i to keep track of all the data points.

For example: $\hat{\mathbf{Y}}_i = \mathbf{a + bX}_i$ is the ***i*th predicted value**.

$e_i = Y_i - \hat{Y}_i$ is the ***i*th residual**

We speak of the data being composed of two pieces: (1) the Fit or Model **AND**
(2) the Residual or Error.

Examine the two plots below one shows a 'good' model(small error) and the other a 'bad' model(large error).



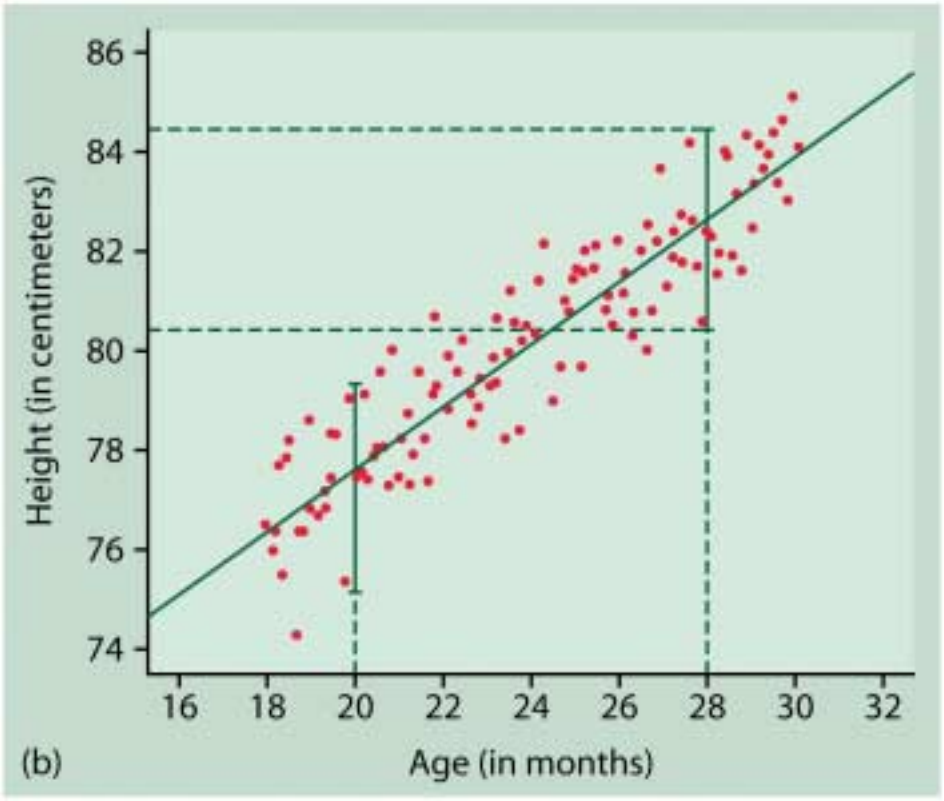


Figure 2.16 Explained versus unexplained variation. In (a), almost all of the variation in height is explained by the linear relationship between height and age. The remaining variation (spread of heights when x = 28 months, for example) is small. In (b), the linear relationship explains a smaller part of the variation in height. The remaining variation (illustrated again for x = 28) is larger.

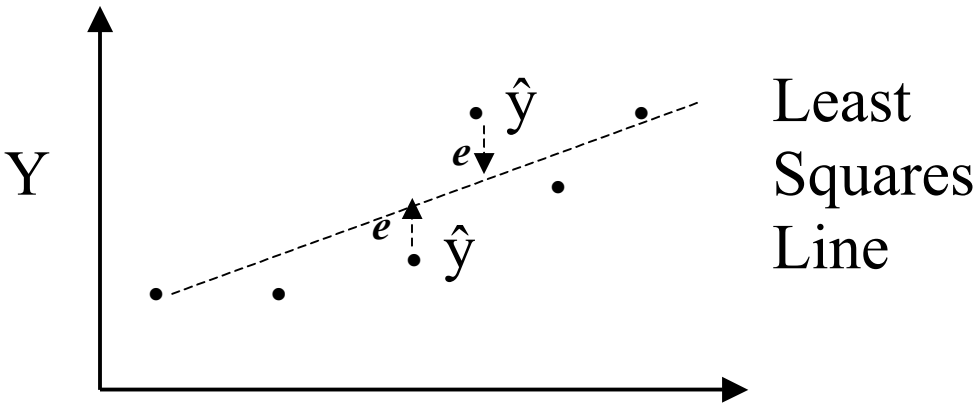
NOTE: If the line predicts the data well then the residuals are small.

$$Y_i = \hat{Y}_i + e_i$$

Data = Fit + Residual

OR

Data = Model + Error



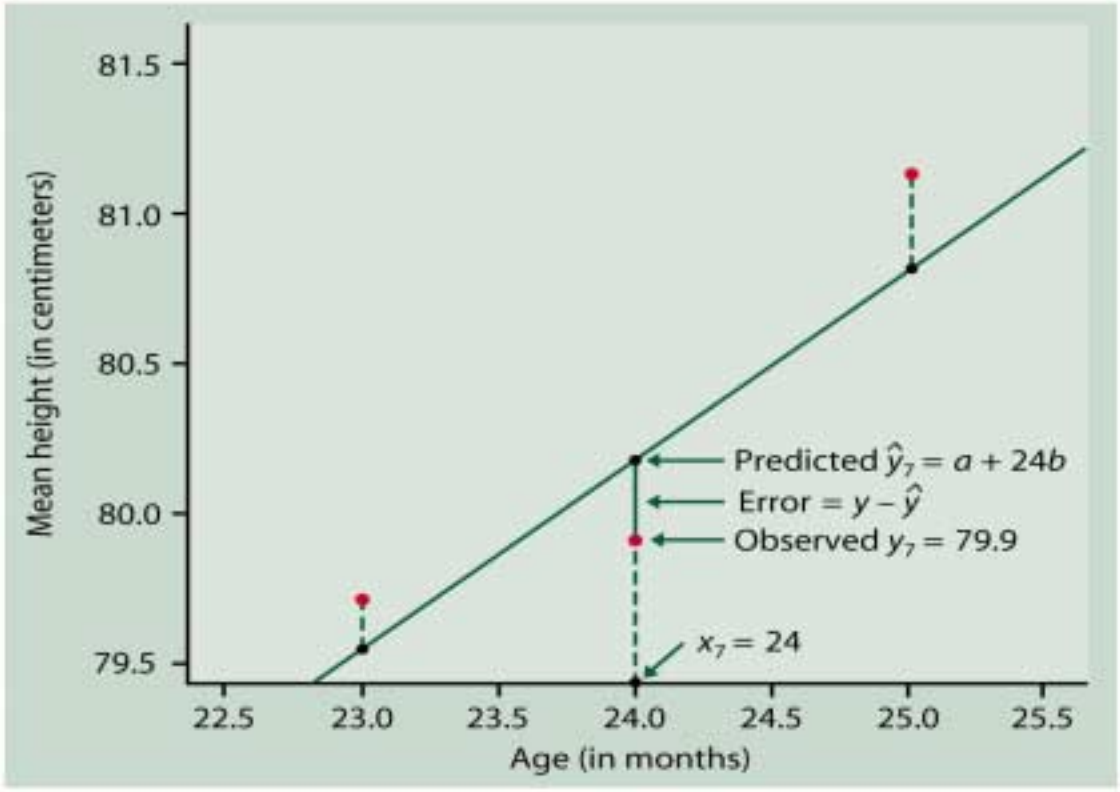


Figure 2.13 The least squares idea: make the errors in predicting y as small as possible by minimizing the sum of their squares.

Properties of Residuals (e_i)

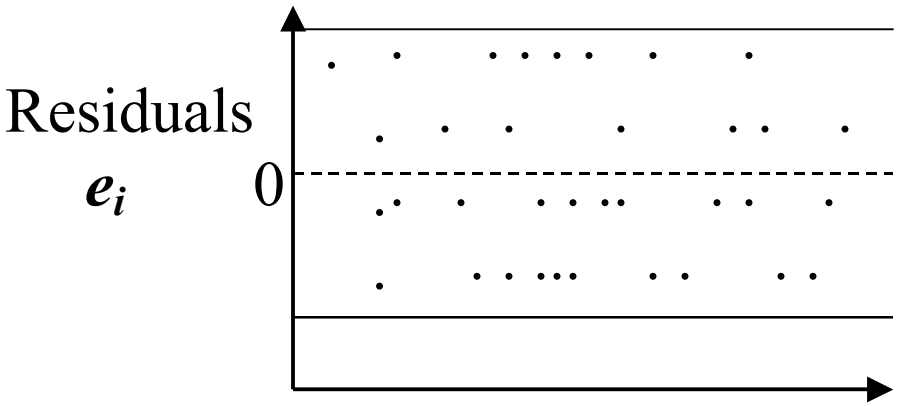
1) The mean of the residuals is zero.

2) The variance or spread of residuals about the line is minimized.

3) Some of the residuals will be positive and some will be negative since some of the data will lie above the line and some below the line.

Example: Below is a plot of X vs ***e*** (residuals) for a reasonable line model.

Note residuals have mean zero, are close to the mean of zero, and some are positive and some negative producing a "band-shaped" pattern.



Below are some patterns that are not band shaped but cyclical, curved and fan like.

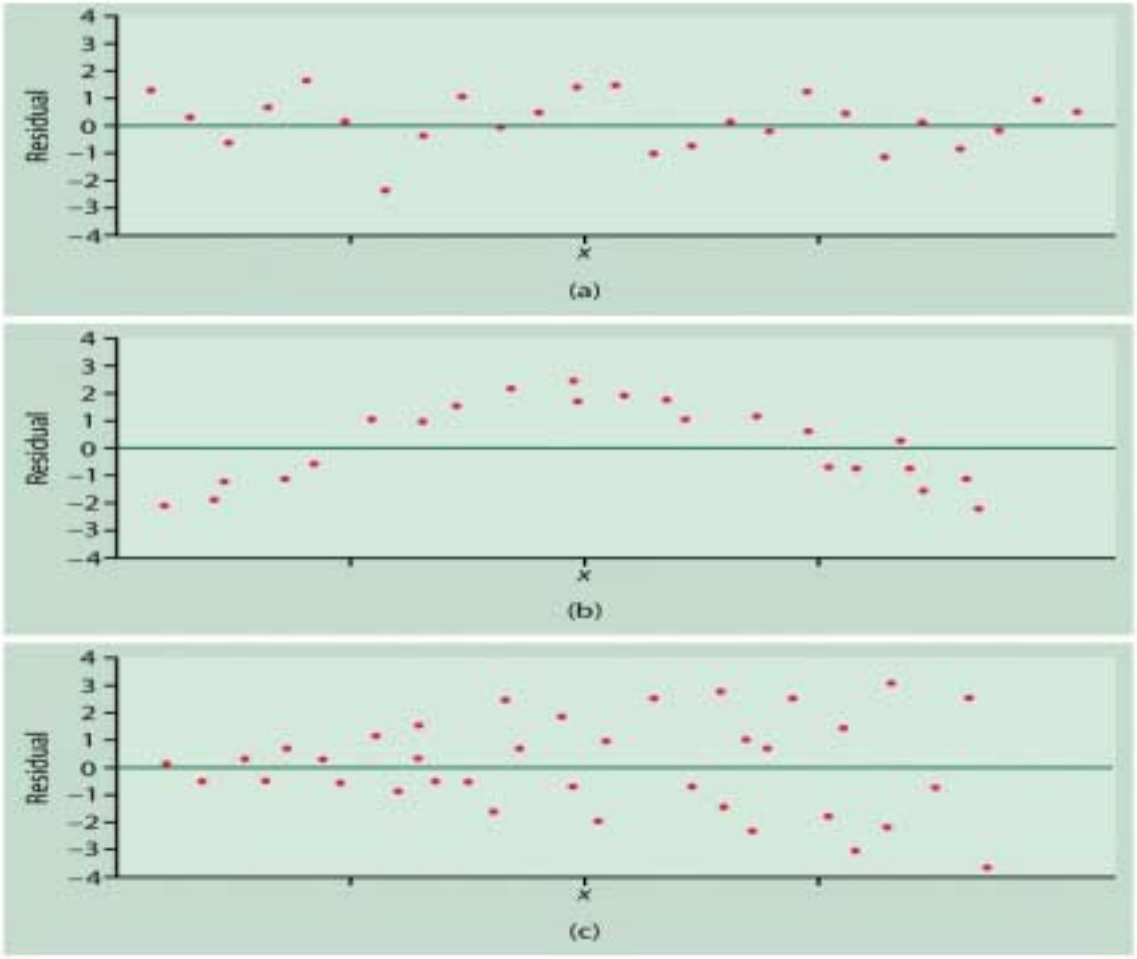


Figure 2.19 Simplified patterns in plots of least squares residuals.

Remember: X – Flow Y – Perf

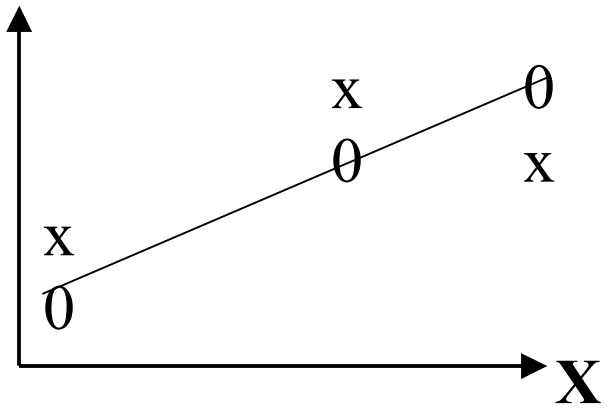
A zero residual indicates the data point is on the line. The least squares or regression equation is:

$$Y = 0.406 + 1.39\,X$$

a b

The predicted value is $\hat{Y} = .406 + 1.39 * X$
The residual is $\boldsymbol{e} = Y - \hat{Y}$

The graph below shows how the observed and the predicted values form a residual.



The ovals (0) are connected by hand.

A = $\hat{y}$ vs X o B = Y vs X x

The observed data Y are graphed as crosses and the predicted Y (from the least squares line) are graphed as zeros. The vertical distance between the two is the residual. As stated previously the plot of X vs. ***e*** should be a 'band' shape with some residuals above zero and some below.

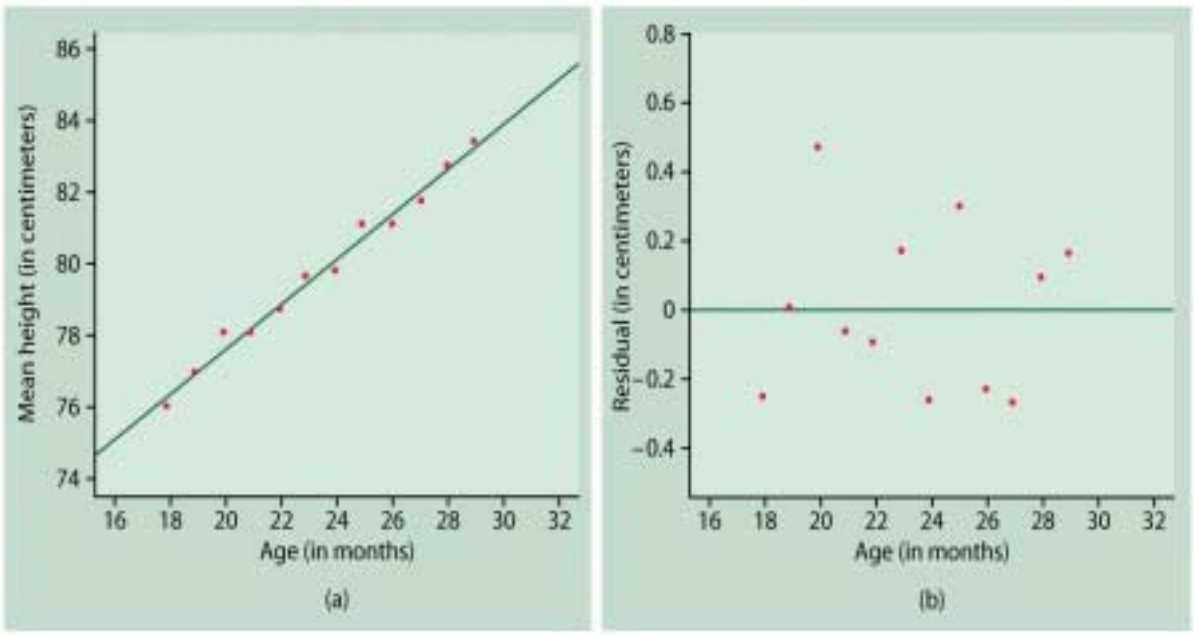
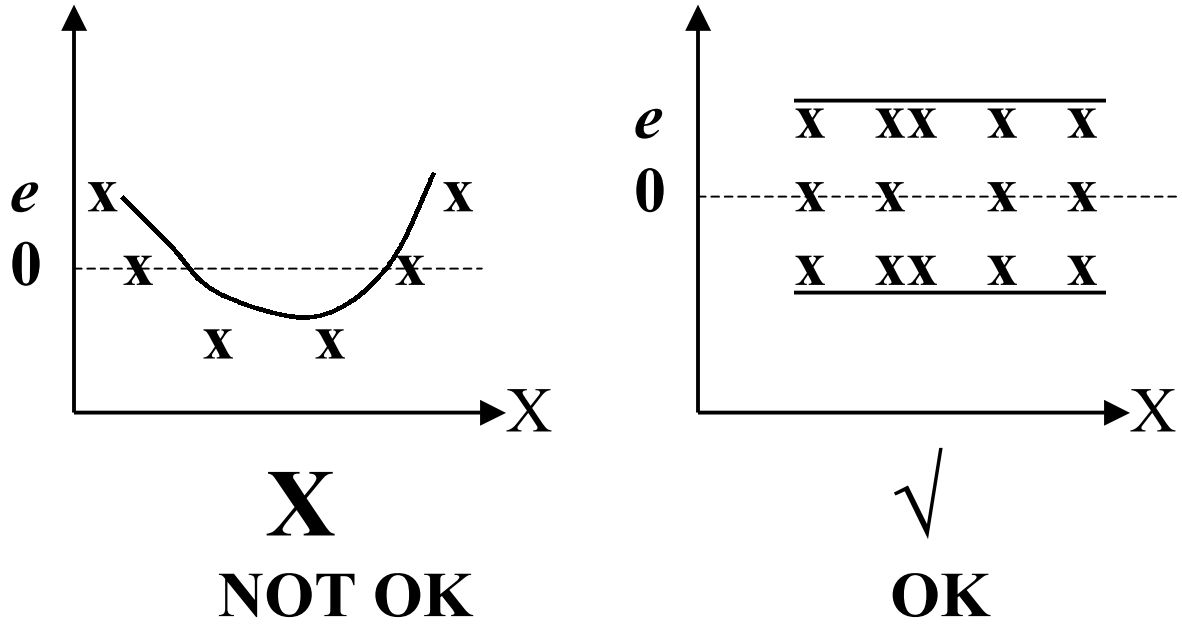


Figure 2.18 (a) Kalama growth data with least-squares line. (b) Plot of the residuals from the regression in (a) against the explanatory variable.

If fitting a line to the data was OK then the residual plot should be pattern less or a band as mentioned previously.

Note the curved pattern in the X vs. e plot on the left. This indicates a line was not a good choice of model for this data. The plot on the right displays the band shaped pattern that is typical of a model or line that fits well.



X
NOT OK

√
OK

The above left plot shows a clear curve pattern indicating a problem with the line model.

Prediction

Scientists fit lines to data in order to make predictions about Y for a given X.

If we predict Y (performance) at X = 2.5 this is called **interpolation** since it is within the range of the data collected. However if we predict Y when X = 8.0 this is called **extrapolation** since it is outside the range of the data collected.

If we substitute the above X values into our equation we obtain:

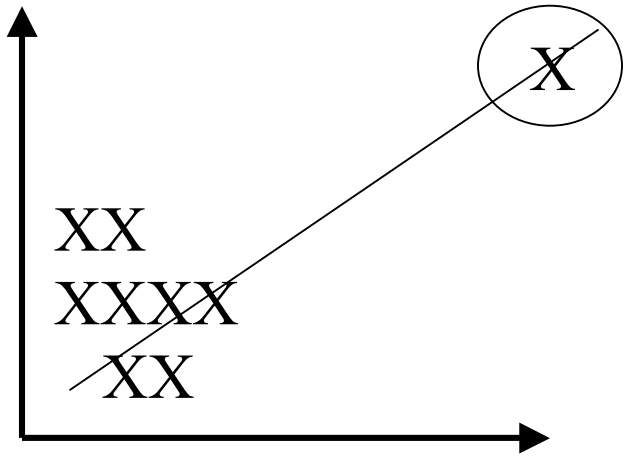
$$Y1 = .406 + 1.39 * 2.5$$
$$Y2 = .406 + 1.39 * 8.0$$

Where: $Y1 = 7.535$ $Y2 = 15.18$

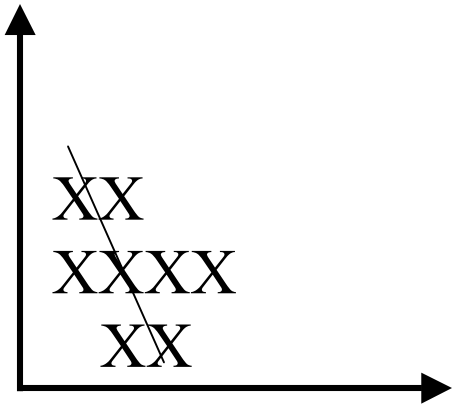
There is a danger in extrapolating since we have no data collected at these points.

Influential Observation

Look at the graph below:



Note one observation is far from the central mass of the data. If we remove this observation, the line changes dramatically. This is called an influential observation. The residuals are **small** for influential observations.
graph below:



Without the data point circled previously the line changes dramatically.

Remember outliers are observations with large residuals (e_i). From the graph below can you find the outlier, the influential observation?

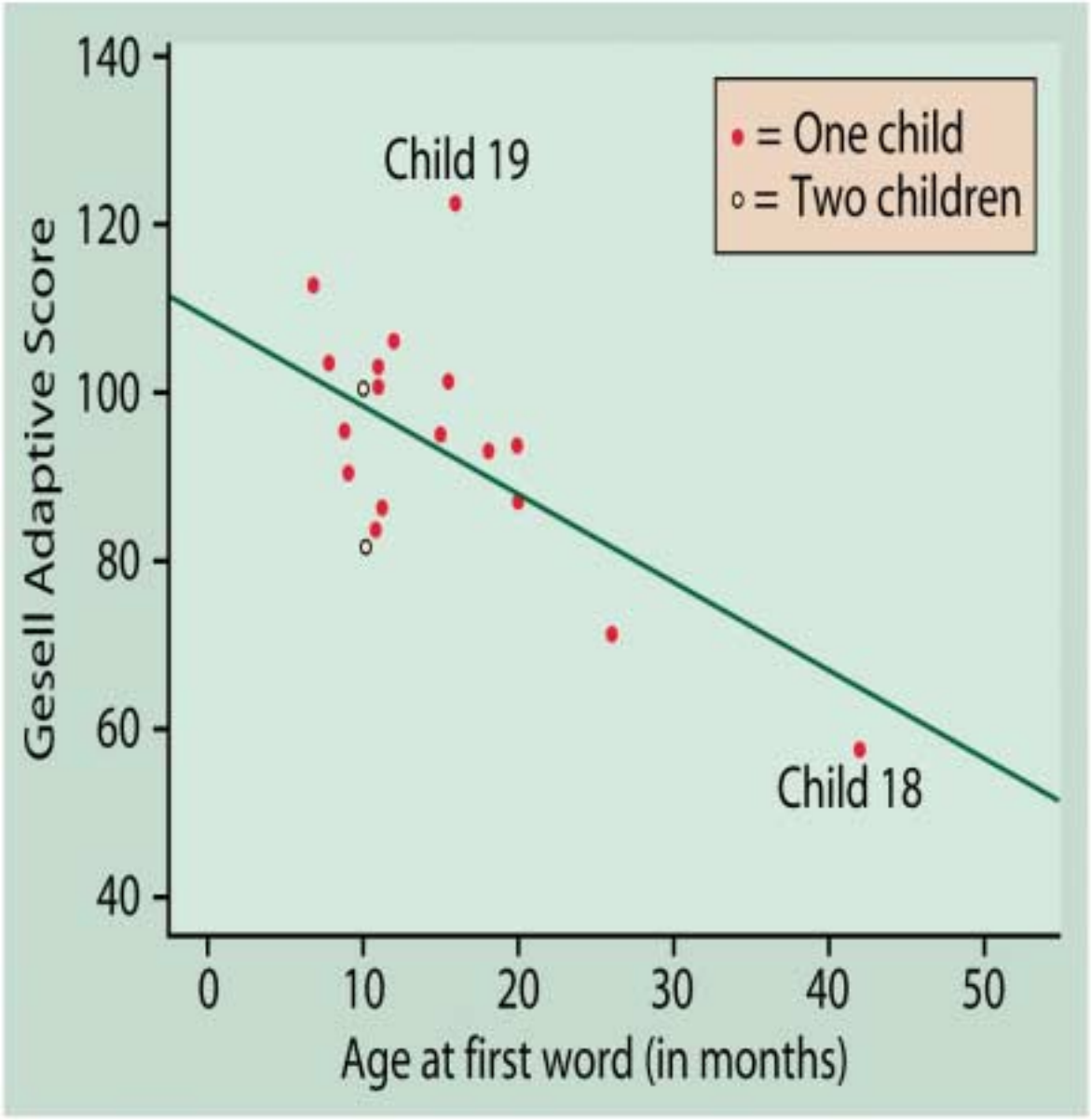


Figure 2.23 Regression of Gessel Adaptive Score on age at first word for 21 children, from Table 2.9.

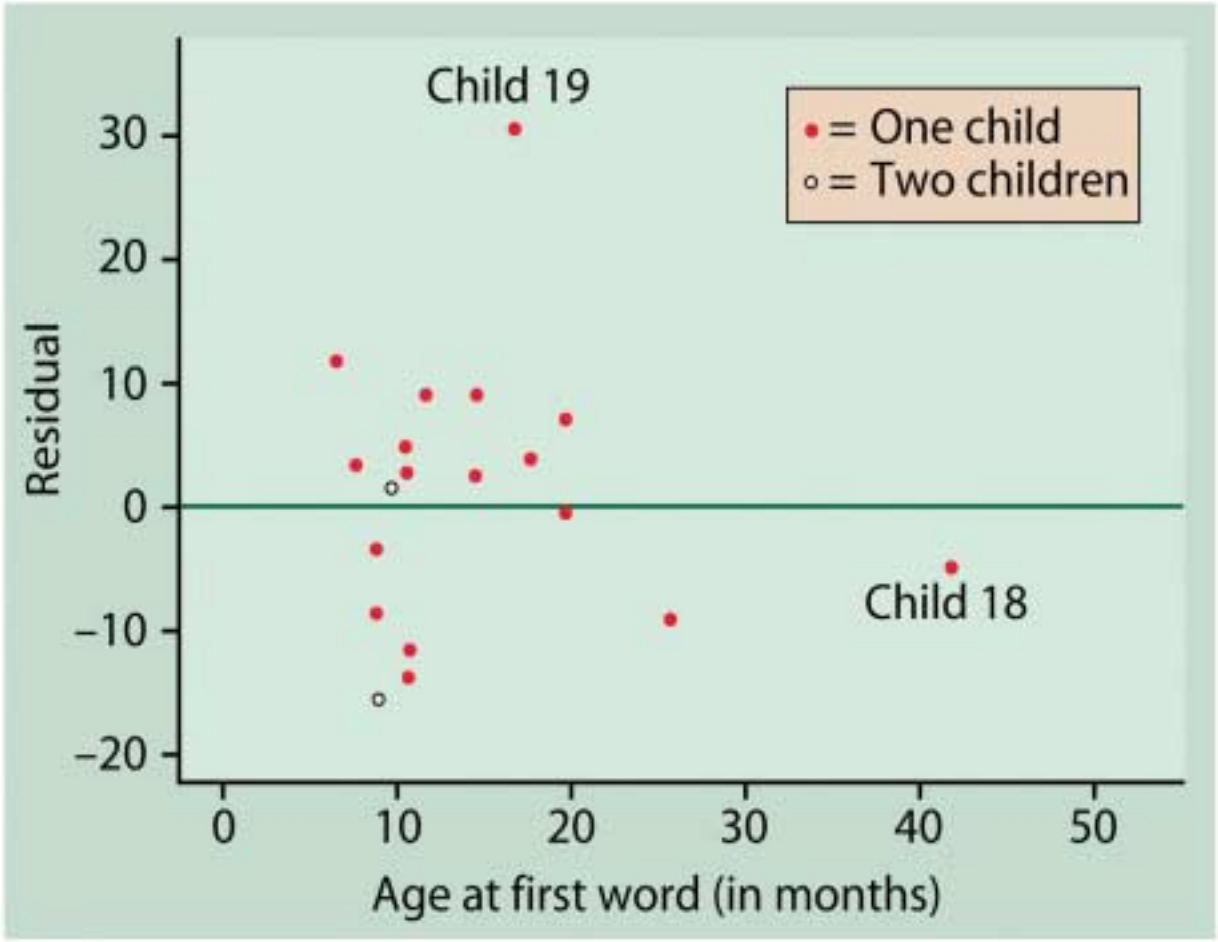


Figure 2.24 Residual plot for the regression of Gessel score on age at first word. Child 19 is an outlier in y. Child 18 is an influential case that does not have a large residual.

What do you do when you conclude the fit is poor?

You might consider a simple transformation of the data. For example if the data has a skewed distribution a logarithmic transformation of the data may be helpful.

Click below for an example of how to fit a line to data.

SPSS Example One

View the Video-Correlation

Correlation Coefficient

How strong is the linear relationship between quantitative variables X and Y? Researchers report the correlation coefficient to summarize the strength of association between two variables. The correlation is defined as follows:

$$r_{xy} = \textit{correlation coefficient}$$

$$= \frac{s_x}{s_y} b_{xy} \qquad b_{xy} \neq b_{yx}$$

$$= \frac{s_{xy}}{s_x s_y}$$

where

$$s_x = \sqrt{\frac{1}{(n-1)} \sum (x - \bar{x})^2}$$

$$s_y = \sqrt{\frac{1}{(n-1)} \sum (y - \bar{y})^2}$$

$$s_{xy} = \sqrt{\frac{1}{(n-1)} \sum (x - \bar{x})^2 \sum (y - \bar{y})^2}$$

b_{xy} = slope of least square line for (x_i, y_i)
i=1,2 ... n

The correlation is a messy calculation. It is best calculated by computer. If the standard deviation of X and Y is equal to one then the correlation is equal to the slope of the line.

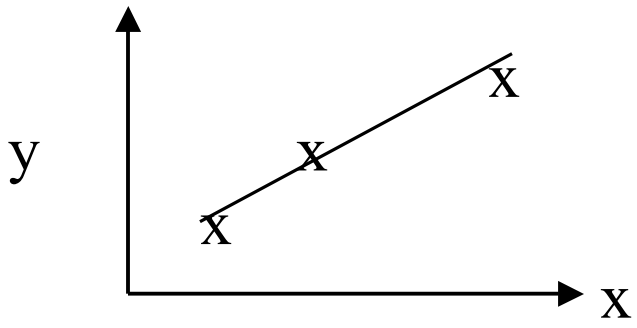
Facts

1) $r_{xy} = r_{yx}$ (the correlation of X and Y is equal to the correlation of Y and X)

2) $\mathbf{-1 \leq r_{xy} \leq 1}$
(the correlation lies between -1 and +1)

3) $r_{xy} = 1$ **if and only if** $Y_i = a + b\,X_i$
$i = 1, 2 \ldots n \quad b > 0$

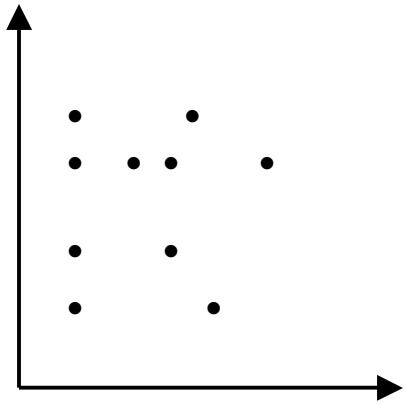
In other words the correlation is 1 only if all of the points lie on a line whose slope is greater than 0.



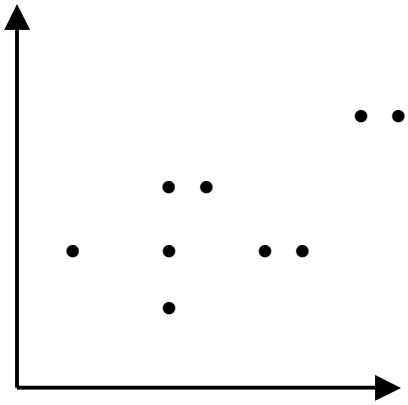
4) $r_{xy} = -1$ if and only if $Y_i = a + b\,X_i$
$i = 1,2 \ldots n \quad b<0$
The correlation is -1 only if all of the points lie on a line whose slope is less than 0.

5) r^2_{xy} = proportion of observed variation in Y explained by a linear dependence on X.

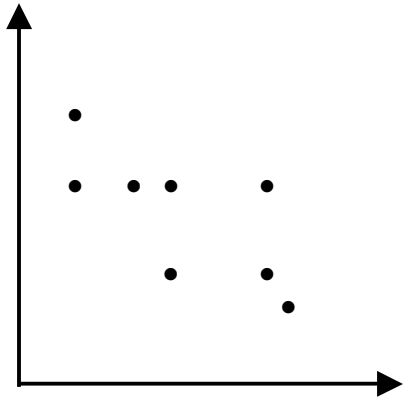
Some examples are given below of data and the corresponding correlation. Note how the points cluster closer to a line as the correlation moves away from 0 and closer to ± 1.



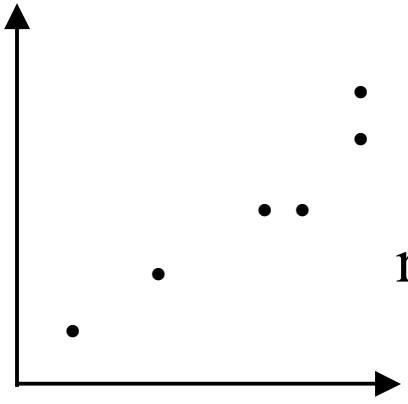
r=.02



r=.30



r= - .70



r=.90

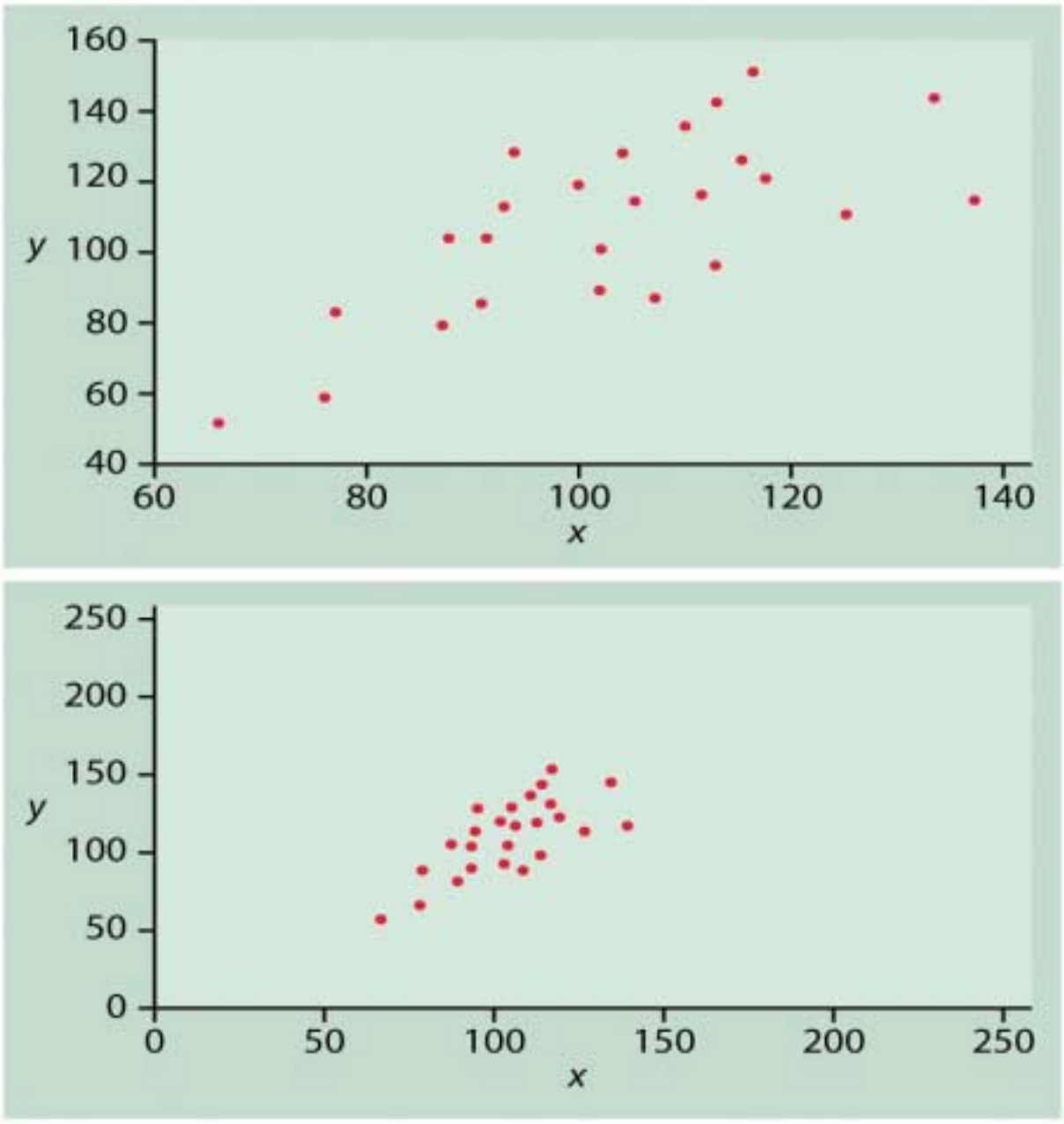


Figure 2.9 Two scatterplots of the same data; the linear pattern in the lower plot appears stronger because of the surrounding white space.

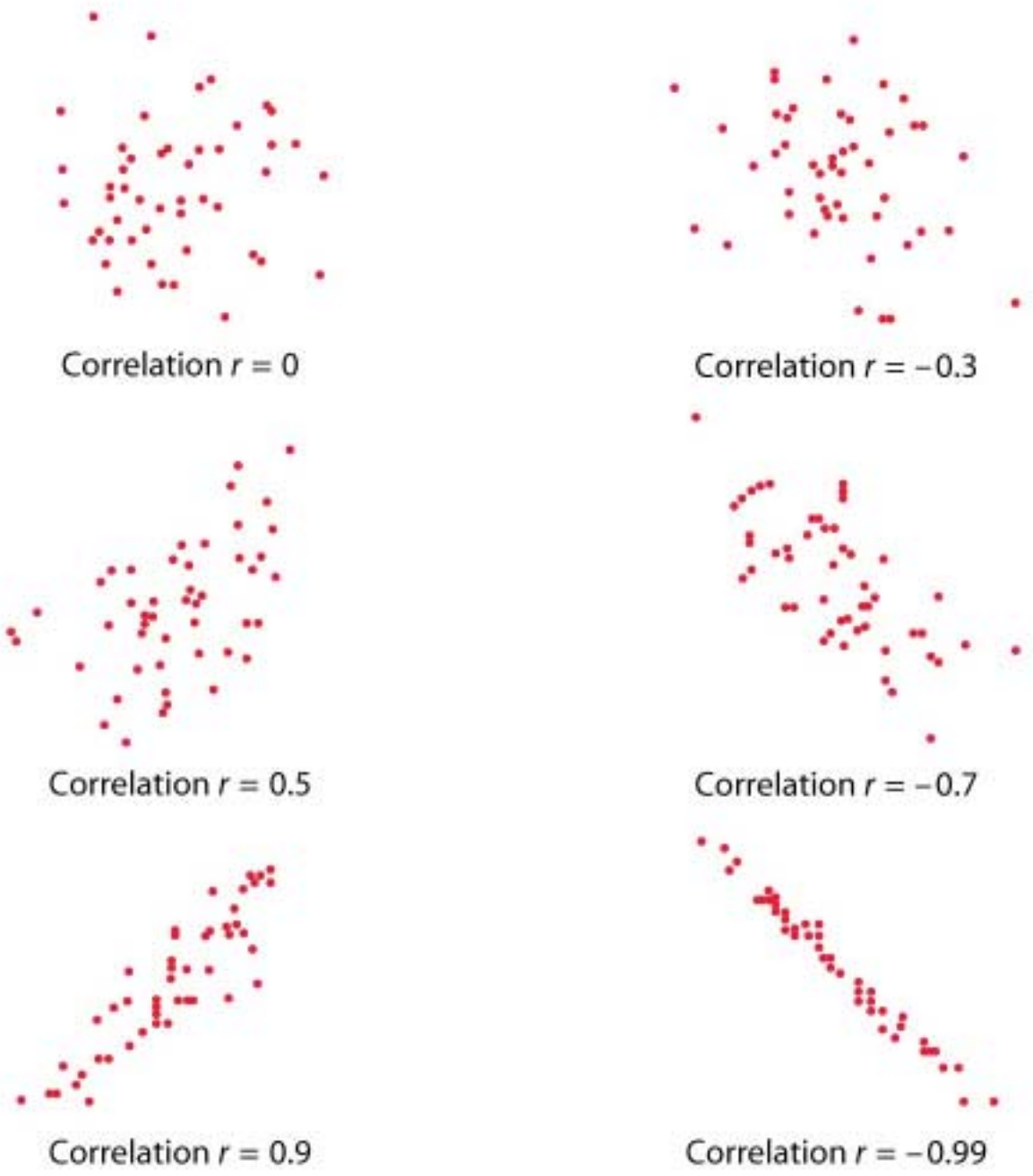


Figure 2.10 How the correlation *r* measures the direction and strength of linear association.

Example (cont): Flow Rate is (X)
Performance is (Y)

The **correlation** between X and Y is:

$r = .967$

This is a very strong, positive relationship:

$r^2 = (.967)^2 =$ **.94 of the variation in Y is explained by X.**

R squared is .94 or we can say that 94% of the variance in performance is explained by the drug flow rate.

Almost all of the variance in Y (Perf) is explained by X (Flow).

In other words there is very little variance in Y (6%) that X does not explain.

SPSS Example Two

Click above for an example of how to calculate the correlation by computer using SPSS.

Relationships Amongst Categorical Variables

Suppose X, Y are categorical variables. The table below gives the number of admissions to a university broken down by gender.

Sex Bias in Admissions

		Y = admission		
		Y	N	
X = sex	M	490	210	700
	F	280	220	500
		770	430	1,200

Question: Is there a bias in the admissions policy? In other words, is it more likely for a male to be admitted? To answer this, compare the conditional distributions of Y given X = sex.

M	Y	490/700	= 7/10	= .70
	N	210/700	= 3/10	=.30
F	Y	280/500	= 14/25	= .56
	N	220/500	= 11/25	= .44

:. since .70 > .56 there appears to be evidence of a sex bias. But, break the data down further by the two schools at the university. (Business and Law) and their admissions and calculate the conditional distributions.

Business Admissions

Sex	Y	N	
M	480	120	600

Law Admissions

Sex	Y	N	
M	10	90	100

F	180	20	200	F	100	200	300
	660	140	800		110	290	400

Conditional distributions given sex are below:

M	Y	480/600	= .8	M	Y	10/100	= .1
	N		= .2		N		=.9
F	Y	180/200	= .9	F	Y	100/300	= .3
	N		= .1		N		= .7

In both the Business School and Law School there appears to be a bias in favor of females but overall in our first table we conclude a bias in favor of males! How is this possible?

Simpsons Paradox

Why? A third variable (**a l**urking variable) is varying with X and Y!!!

Note the acceptance rates for the two schools:

Business $\frac{660}{800} = .825$ Law $\frac{110}{400} = .275$

Of 500 females $\frac{200}{500} = .4$ apply to business

$\frac{300}{500} = .6$ apply to law

Thus, more women apply to law but law has a much lower acceptance rate.

View the Video-A Question of Causation

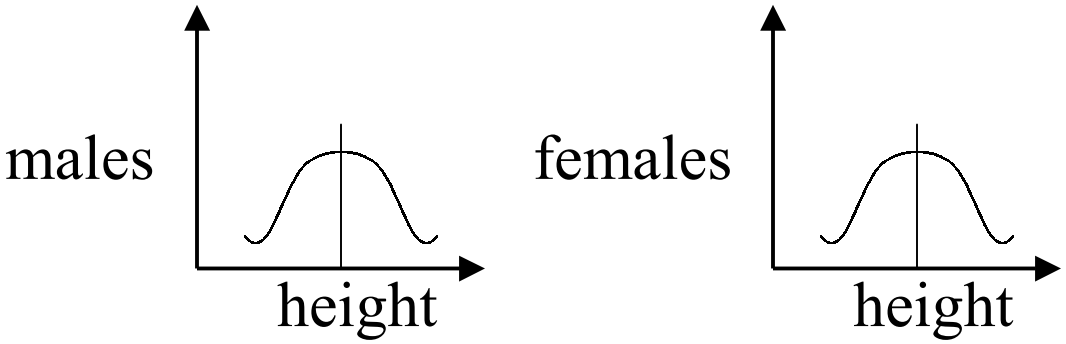
Cause – Effect Relationships

X = predictor variable Y = response variable

Consider repeated observations of Y for X fixed. This leads to a distribution of Y values for each value of X. We call this the conditional distribution of Y given X.

Example: Y = height in cm
X = sex (M or F)

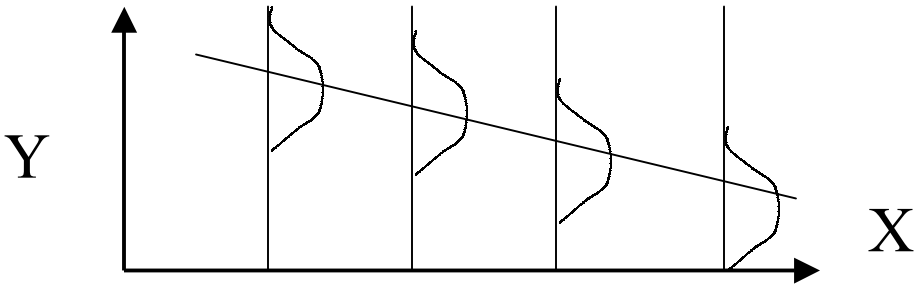
The conditional distribution of Y given X is graphed below.



A relationship between X and Y exists if the conditional distribution of Y given X changes as X changes.

When Y is quantitative we sometimes describe the relationship by Y = a+bX, but this could be any function. Now suppose a relationship exists between X and Y. Does the change in X cause the change in the conditional distribution of Y?

Example: Y = life length of a rat
X = quantity of tobacco smoke exposed to each day



If we know that the conditions under which the observations of Y were taken were identical for each value of X, then the only thing that could cause the conditional distributions of Y to change are the changes in X. Thus, we say a **cause-effect** relationship exists between X & Y.

How can we ensure these conditions are met?

Answer: Conduct an experiment!

Alas, this is not always possible.

Example: Y = life length of a human
X = quantity of tobacco smoke exposed to each day

We cannot conduct an experiment in the above case and strictly speaking cannot establish cause and effect. However, there is a vast amount of animal and correlational research that indicates tobacco smoke does decrease the life length of a person. Given this abundance of evidence we do speak of a cause effect relationship even though an experiment has not been conducted with humans.

Remember try the links to the virtual statistics lab and the freeman quizzes to help you understand the material.

Chapter Two Summary

The least **squares regression line** is a straight line that describes how a dependent variable Y is related to an explanatory variable X. **Least squares** is a method of fitting a line that minimizes the vertical distance of the observed Y values to the line. The vertical distance is called a **residual**.

The **slope** b of the line is the rate y changes per unit change in X. The **intercept** of the line is the value Y when X = 0. The **correlation (r*)*** is the

slope of the line when both X and Y are in standardized(z) units. R-squared ($\mathbf{r^2}$) is the proportion of variance in Y accounted for by X.

Influential observations are points that have a large impact on the line. **Outliers** have large residuals. A **two-way table** of counts organizes the data into **rows** and **columns**.

To find the **conditional distribution** of the row variable for a given column variable, consider only that one column and find each entry in the column as a percent of the column total. When data are aggregated sometimes conclusions change, this is called **Simpsons Paradox**.

Remember: Tips for Success

1) Read the text
2) Read the notes
3) Try the assignments
4) If needed, try the exercise questions

5) Try the simulations and view the videos if you need more help with a concept
6) Try the self tests for practice on each chapter of the text at www.whfreeman.com/ips
7) Steady Work = SUCCESS !!!