

Chapter 3: Examining Relationships

Most statistical studies involve more than one variable. Often in the AP Statistics exam, you will be asked to compare two data sets by using side by side boxplots or histograms etc. However, there are times where we want to examine relationships among several variables for the same group of data. For example, is the temperature outside related to the amount of ice cream is sold. Or does the amount of TV watched affect IQ? When you examine the relationship between two variables you need to start with a scatterplot.

Section 3.1: Scatterplots and Correlation

A scatterplot shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as a point in the plot fixed by the values of both variables for that individual.

A **response variable** measures the outcome of a study. An **explanatory variable** helps explain or influences change in a response variable.

You will often find explanatory variables called *independent variables*, and response variable called *dependent variables*. The idea behind this language is that the response variable *depends* on the explanatory variable. Because the words independent and dependent have other, unrelated meanings in statistics, we won't use them here.

Example:

The costs of mountain bikes are given in the following table, along with overall rating scores given to each bike by *Consumer Reports* magazine. Graph the data using a scatterplot. Identify the explanatory and response variable.

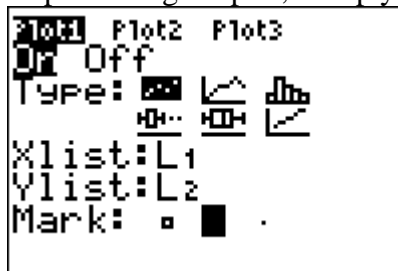
Model	Price	Rating	Model	Price	Rating
Trex VRX200	\$1000	80	Raleigh M-40	\$320	60
GT XCR-4000	940	80	Schwinn Frontier GS	285	60
Specialized FSR	1100	79	Specialized Hardrock	300	60
Cannondale V400	1100	78	Giant Rincon	290	58
Trek 6500	700	80	Trek 820	330	58
Schwinn Moab 3	680	80	Diamondback Response	300	56
Specialized Rockhopper	600	79	GT Outpost	330	56
Haro Escape A7.1	440	75	Mongoose SX4.3	330	48
Giant Yukon SE	450	72	Fuji Odessa	210	45
Mongoose SX6.5	550	68	Specialized Globe A1	380	60
Diamondback Sorrento	340	63	Schwinn Sierra GS	400	60
Motiv Rockridge	180	40	Raleigh SC-40	300	55
Huffy Anorak 36789	140	22			

Source: *Consumer Reports*, June 1999.

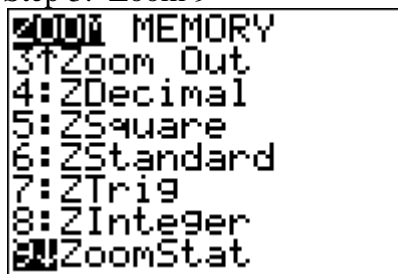
Step 1: Load the data into your lists.

L1	L2	L3	1
40000	80		-----
940	80		
1100	79		
1100	78		
700	80		
680	80		
600	79		
L1(1)=1000			

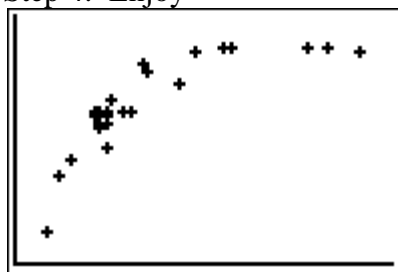
Step 2: using statplot, set up your scatterplot



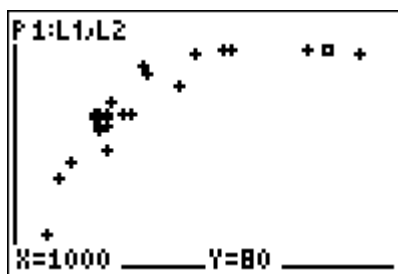
Step 3: Zoom 9



Step 4: Enjoy



Notes: Using the trace button, you can find the coordinates of each point on the scatterplot.



Interpreting a Scatterplot

- In any graph of data, look for **overall pattern** and for **striking deviations** from that pattern.
- You can describe the overall pattern of a scatterplot by the **direction, form, and strength** of the relationship.
- An important kind of deviation is an **outlier**, an individual value that falls outside the overall pattern of the relationship.

Interpret the scatterplot that you just constructed on your calculator.

There are two ways of determining whether two variables are related:

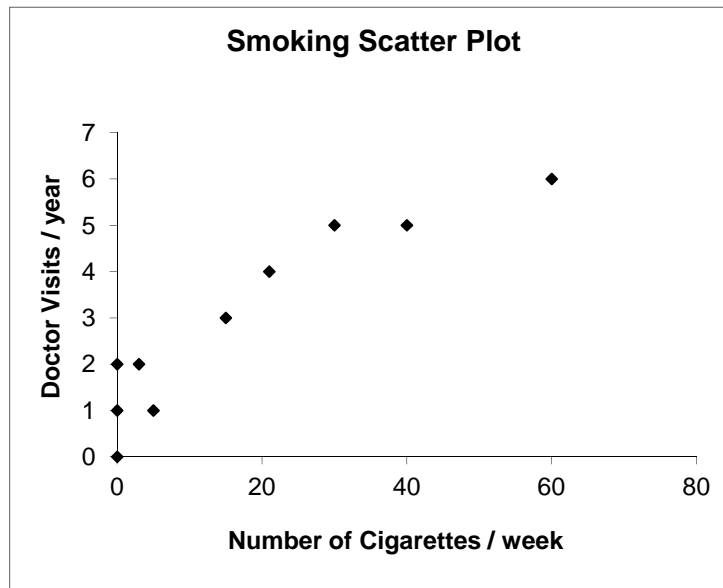
- a) By looking at a scatter plot (graphical approach)
- b) By calculating a “correlation coefficient” (mathematical approach)

Graphical Approach:

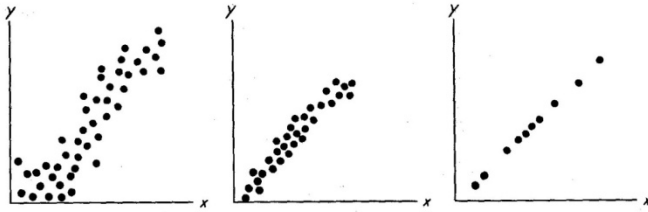
Suppose we hypothesize that the number of doctor visits a person has can be explained by the amount of cigarettes they smoke. So we want to see if there is a relationship between the number of cigarettes one smokes a week (the explanatory or independent variable) and the number of times per year one visits a doctor (response or dependant variable). We ask 10 random people and get the following information:

# of Cigarettes Per Week	0	3	21	15	30	5	40	60	0	0
Number of doctor visits per year	1	2	4	3	5	1	5	6	2	0

If we were to plot the following data point on an X-Y coordinate plane, we would get a scatter plot that looks like this:



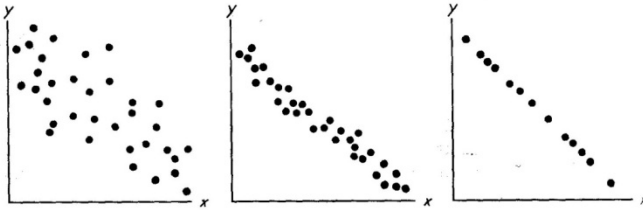
Note that the graph shows us that as you smoke more cigarettes / week, you also tend to go to the doctor more often. This result demonstrates that there is a positive relationship between the two variables. The following scatter plots show other types of relationships that might occur between two variables:



(a) Positive correlation between x and y

(b) Strong positive correlation between x and y

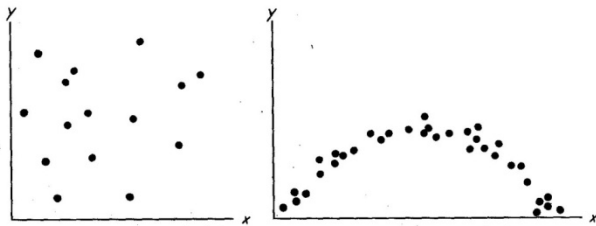
(c) Perfect positive correlation between x and y



(d) Negative correlation between x and y

(e) Strong negative correlation between x and y

(f) Perfect negative correlation between x and y

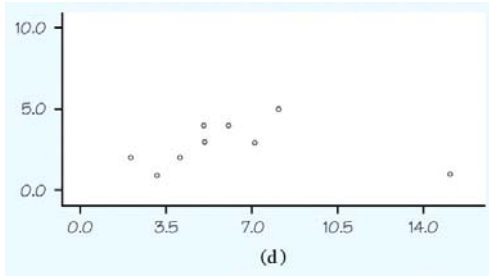


(g) No correlation between x and y

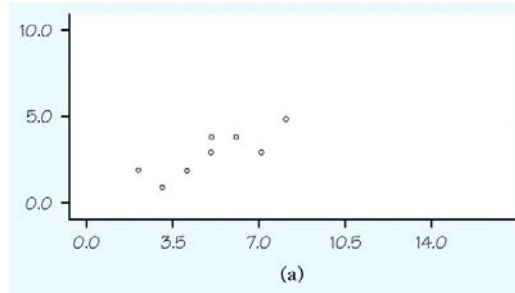
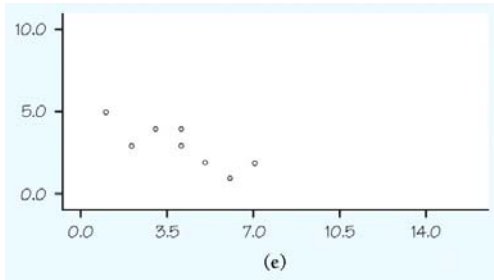
(h) Nonlinear correlation between x and y

Things to look for in a scatterplot:

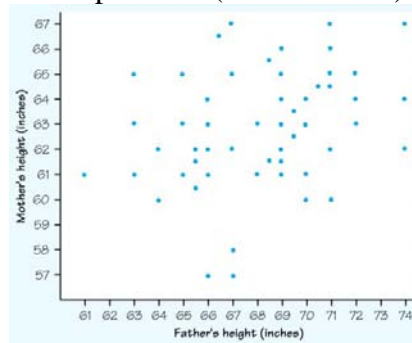
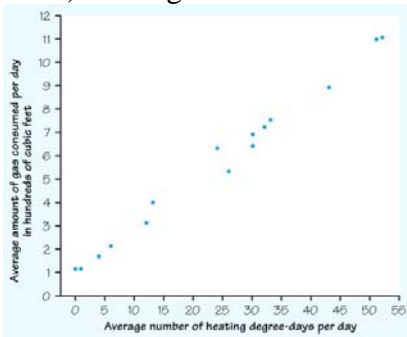
- 1) Overall pattern (linear, exponential, etc.) or deviations from the pattern (outliers)



- 2) Direction: Positive or negative slope

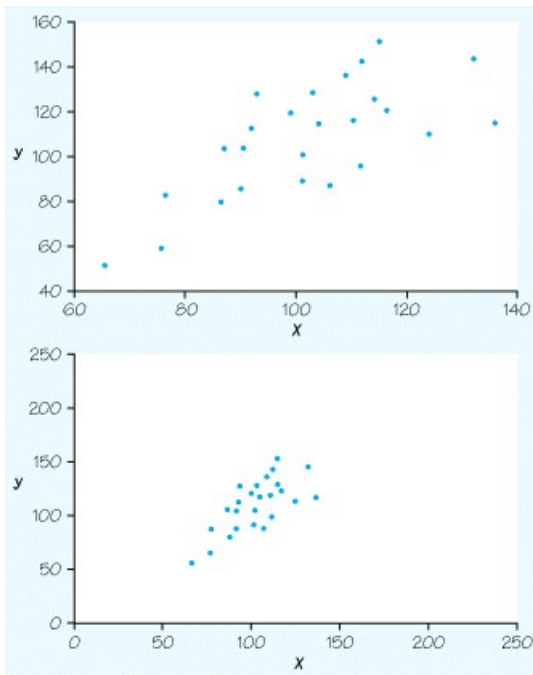


- 3) Strength: How close do the points lie to a simple form (such as a line)



Mathematical Approach:

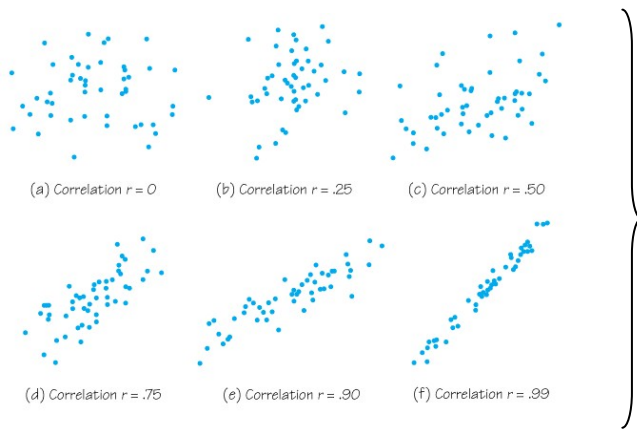
In order to strengthen the analysis when comparing two variables, we can attach a number, called the correlation coefficient (r), to describe the linear relationship between two variables. This number helps remove any subjectivity in reading a linear scatter plot and helps us avoid being fooled by axis manipulation (see example below):



The correlation measures the strength and direction of the **linear** relationship between two quantitative variables:

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

Essentially, the correlation coefficient, r , finds the average of the product of the standardized scores.



Examples of scatter plots and their corresponding correlation coefficients (in order to give you a feeling for the “strength” or r)

Example: Lets calculate the correlation between the amount of cigarettes smoked per week and the number of doctor visits per year:

# of Cigarettes Per Week	0	3	21	15	30	5	40	60	0	0
Number of doctor visits per year	1	2	4	3	5	1	5	6	2	0

$$\bar{x} =$$

$$s_x =$$

$$\bar{y} =$$

$$s_y =$$

x	$\frac{x - \bar{x}}{s_x}$	y	$\frac{y - \bar{y}}{s_y}$	$\frac{x - \bar{x}}{s_x} \cdot \frac{y - \bar{y}}{s_y}$
0		1		
3		2		
21		4		
15		3		
30		5		
5		1		
40		5		
60		6		
0		2		
0		0		

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

So sum up the last column and multiply it by $\frac{1}{9}$

$$r = \underline{\hspace{2cm}}$$

Now check your answer by doing it on the calculator.

Interpreting Correlation Coefficients:

- 1) The value of r is always between -1 and 1 .

- 2) A correlation of -1 implies two variables are perfectly negatively correlated: As one goes up the other goes down and as one goes down the other goes up.

- 3) A correlation of 1 implies that there is perfect positive correlation between the two variables. As one goes up, so does the other.

- 4) A correlation of 0 implies that there is no correlation between the two variables. In other words, there is no relationship between them. As one goes up, we know nothing about whether the other will go up or down.

- 5) Positive correlations between 0 and 1 have varying strengths, with the strongest positive correlations being closer to 1 .

- 6) Negative correlation between 0 and -1 are also of varying strength with the strongest negative correlation being closer to -1 .

- 7) r does not have units. Changing the units on your data will not affect the correlation.

- 8) Correlation describes only **linear** relationships between two variables.

- 9) r is very strongly affected by outliers. Use r with caution when outliers appear in your scatter plot. Don't rely on r alone to determine the linear strength between two variables – graph a scatter plot first.

- 10.) Correlation makes no distinction between explanatory and response variables. It makes no difference which variable you call x and which you call y when calculating the correlation.

Section 3.2: Least-Squares Regression

Least Squares Regression A.K.A. linear regression allows you to fit a line to a scatter diagram in order to be able to predict what the value of one variable will be based on the value of another variable.

The fitted line is called the *line of best fit*, *linear regression line*, or *least squares regression line*, (LSRL) and has the form $\hat{y} = a + bx$ where:

a: y intercept

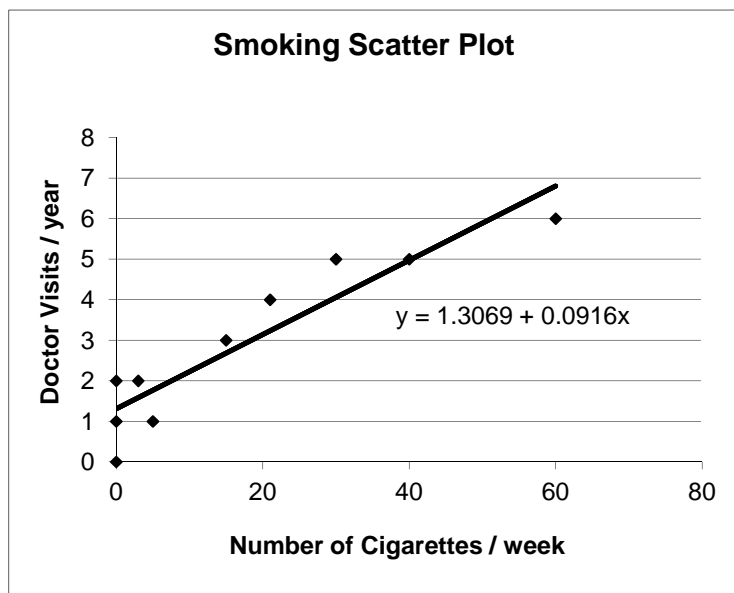
b: slope of the line

The way the line is fitted to the data is through a process called the *method of least squares*. The main idea behind this method is that the square of the **vertical** distance between each data point and the line is minimized.

The least squares regression line is a mathematical model for the data that helps us **predict** values of the response (dependant) variable from the explanatory (independent) variable. Therefore, with regression, unlike with correlation, we must specify which is the response and which is the explanatory variable.

Example:

# of Cigarettes Per Week	0	3	21	15	30	5	40	60	0	0
Number of doctor visits per year	1	2	4	3	5	1	5	6	2	0



Formula for finding the slope and y-intercept in a linear regression line:

Slope:

$$b = r \frac{s_y}{s_x}$$

The slope of the regression line is important in the sense that it gives us the rate of change of \hat{y} with respect to x . In other words, it gives us the amount of change in \hat{y} when x increases by 1.

In our example above, the slope is 0.0916 which means that for every cigarette smoked per week the number of doctor visits per year increases by 0.0916.

Intercept:

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

The intercept is statistically meaningful only when x can actually take values close to zero. In our example the intercept is 1.3069 which would mean that even if a person didn't smoke any cigarettes we would predict that they would make an average of 1.3069 doctor visits per year.

So for the example on the previous page the process for finding the LSRL is:

- 1) We found the correlation coefficient is $r = 0.9257$
- 2) We can find s_y and s_x using the calculator
- 3) Therefore:

$b =$

- 4) We can find \bar{y} and \bar{x} from the calculator as well.

- 5) Therefore:

$a =$

The equation of the regression line is:

Because the LSRL enables us to make predictions about the outcome of the y-variable based on the value of the x-variable, we can use our line to predict how many doctor visits per year a person makes based on the number of cigarettes they smoke per week. For example, if I tell you that a person smokes 10 cigarettes a week, you can plug 10 for x into your regression equation and predict how many doctor visits (\hat{y}) that person makes per year:

$$\hat{y} = 1.3069 + .0916(10) = 2.2229$$

How good is our prediction?

The strength of a prediction which uses the LSRL depends on how close the data points are to the regression line. The mathematical approach to describing this strength is via the **coefficient of determination**. The coefficient of determination gives us *the proportion of variation in the values of y that is explained by least-squares regression of y on x*. The coefficient of determination turns out to be r^2 (correlation coefficient squared).

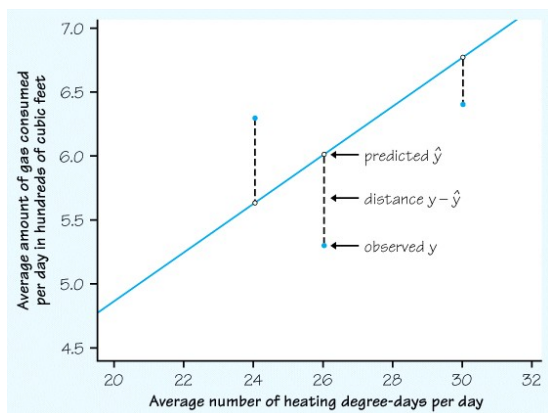
Whenever you use the regression line for prediction, also include r^2 as a measure of how successful the regression is in explaining the response.

In our example $r^2 = 0.8569$. This means that over 85% of the variation in doctor visits per year can be explained by the linear relationship it has with cigarettes smoked per week.

Residuals:

Since the LSRL minimized the vertical distance between the data values and a trend line we have a special name for these vertical distances. They are called **residuals**. A residual is simply the difference between the observed y and the predicted y.

Residual = $y - \hat{y}$



In our example, if we want to find the residual for the person in our sample who smoked 3 cigarettes per week and had 2 doctor visits, we would need to use the regression equation to find \hat{y} .

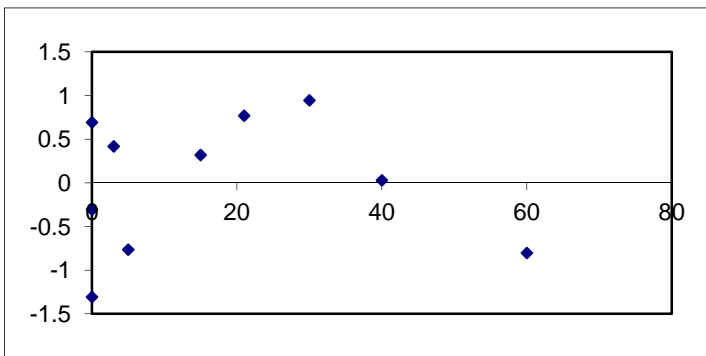
Residual =

Residual Plots:

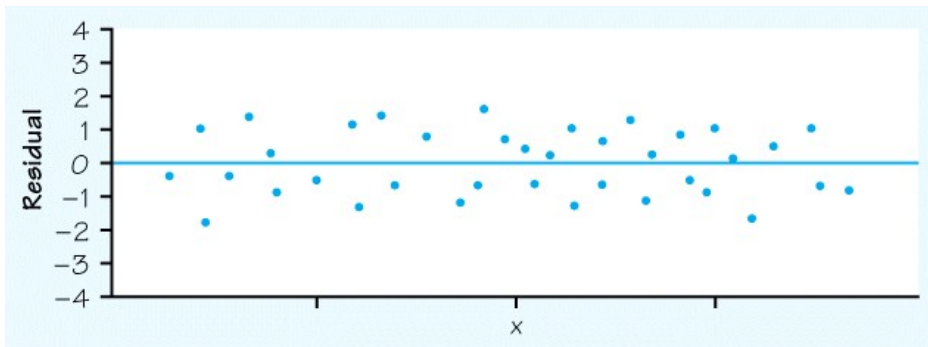
Residuals help us determine how well our data can be modeled by a straight line, by enabling us to construct a residual plot. A residual plot is a scatter diagram that plots the residuals on the y-axis and their corresponding x values on the x-axis. So for our example, the residual plot would be plotted as follows:

x	y	y hat	Residuals
0	1	1.3069	-0.3069
3	2	1.5817	0.4183
21	4	3.2305	0.7695
15	3	2.6809	0.3191
30	5	4.0549	0.9451
5	1	1.7649	-0.7649
40	5	4.9709	0.0291
60	6	6.8029	-0.8029
0	2	1.3069	0.6931
0	0	1.3069	-1.3069
			-0.0074

Notice that the sum of the residuals is zero (any error is due to rounding)



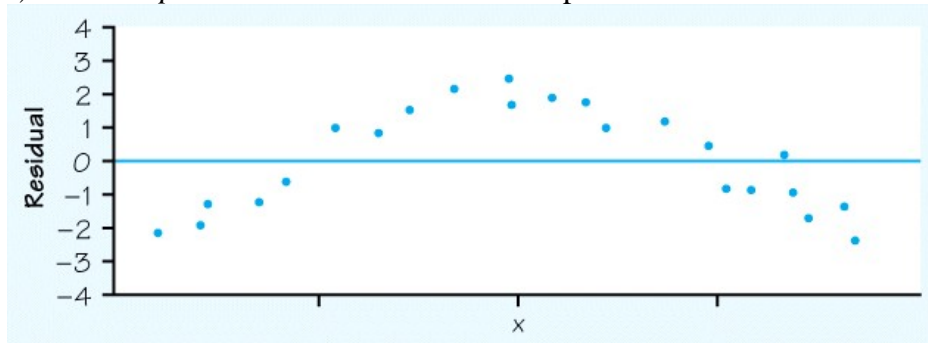
How does a residual plot help us? The pattern of the plot is the indicator of whether or not the data can be modeled linearly. The ideal plot is the one below:



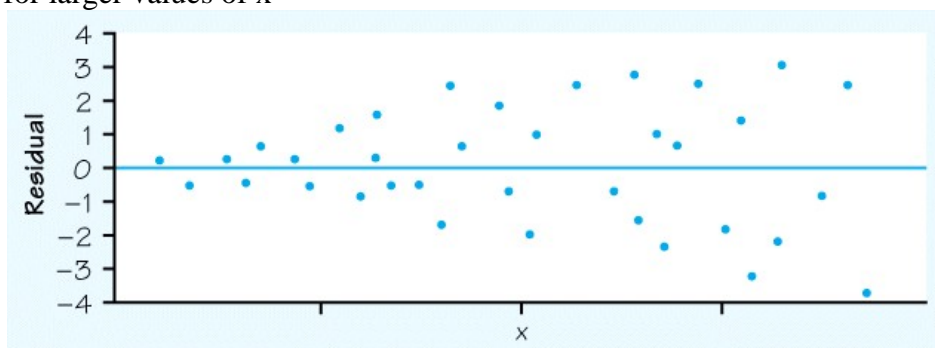
The plot shows a uniform scatter of the points above and below the fitted line with no unusual individual observations.

Here are a few things to watch out for when examining the pattern of a residual plot:

1) *A curved pattern:* Shows that the overall pattern of the data is not linear



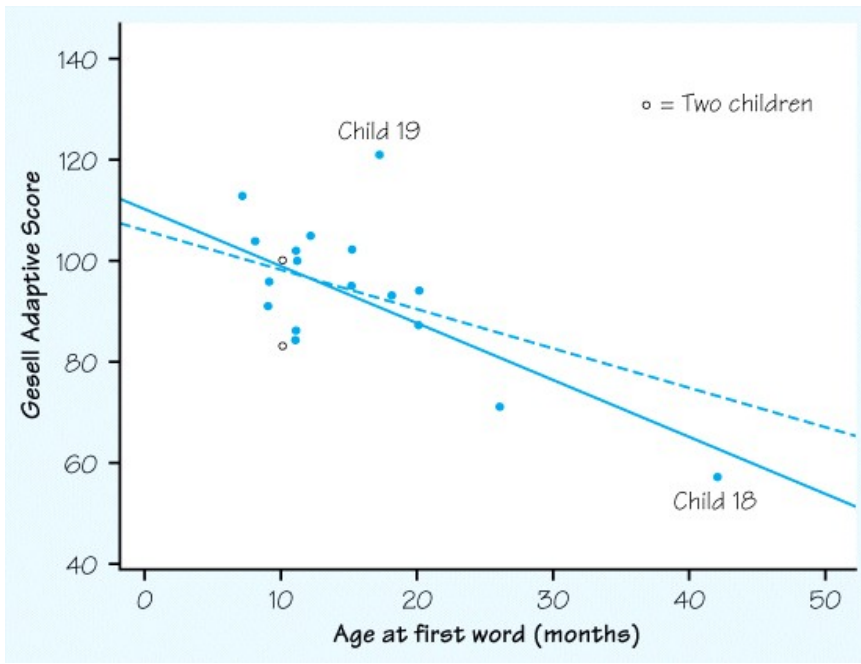
2) *Increasing or decreasing spread about the line as x increases:* Predictions for y will be less accurate for larger values of x



3) *Individual points with large residuals:* indicate outliers from the overall pattern

Outlier: An observation that lies outside the overall pattern in the scatter plot (either in the x or y direction)

Influential point: A point is influential if removing it would markedly change the position of the regression line. Points that are outliers in the x direction are often influential.



In the scatter plot above child 19 is an outlier (its score lies far above the rest of the data) and child 18 is influential because if we were to fit the LSRL without child 18 (dashed line) it would look considerably different than with child 18 (solid line). Child 18, because it is an extreme in the x direction, has a disproportionate pull on the regression line and is therefore an influential point.

Important note: Influential points often have small residuals because they tend to pull the line towards themselves. Therefore, you might miss influential points if you only look at residuals.

Facts about least-squares regression

Fact 1. The distinction between explanatory and response variables is essential in regression. Least-squares regression looks at the distances of the data points from the line only in the y direction. If we reverse the roles of the two variables, we get different least-squares regression line.

Fact 2. There is a close connection between correlation and the slope of the LSRL. The slope is :

$$b = r \frac{s_y}{s_x}$$

This equation says that along the regression line, a change in one standard deviation in x corresponds to a change of r standard deviations in y.

Fact 3. The LSRL always passes through the point (\bar{x}, \bar{y})

Fact 4. The correlation r describes the strength of a straight-line relationship. In the regression setting, this description takes a specific form: the square of the correlation, r^2 , is the fraction of the variation in the values of y that is explained by the least-squares regression of y on x.

Example: An employer gives an aptitude test to his employees and compares it with their productivity at work. The results are given below:

X (Score on Aptitude Test)	Y (Productivity at Work)
6	30
9	49
3	18
8	42
7	39
5	25
8	70
20	40
4	15
11	45
10	52

a) Draw a scatter plot of the data. Label and scale the axes. Is there any pattern? Does the data appear to be positively correlated, negatively correlated, or neither? Which is your response and which is your explanatory variable?



b) Find the correlation coefficient, r ? Does it support your answer to a)?

c) Find the regression equation for this data and sketch it onto your scatter plot.

d) Discuss any outliers or influential points by visually looking at the scatter plot.

e) Remove any potential influential points from your data and recalculate your regression equation. Did they, in fact, make a significant difference to your regression line?

f) Construct a residual plot and discuss its implications



g) What proportion of the variation in productivity can be explained by the aptitude test?

i. Including the influential point

ii. Not including the influential point

h) A new job applicant scores a 9 on the aptitude test. The employer wants to predict his productivity level. Help the employer (use the regression equation without the influential point)

Section 3.3: Correlation and Regression Wisdom

Extrapolation – is the use of a regression line for prediction outside the domain of values of the explanatory variable x that you used to obtain the line. Such predictions cannot be trusted.

Example: The number of people living on American farms has steadily declined in the last 70 years

Year	1945	1950	1955	1960	1965	1970	1975	1980	1985	1990
Population	32.1	30.5	24.4	23.0	19.1	15.6	12.4	9.7	8.9	7.2

- Make a scatterplot of the data and find the LSRL.
- According to the regression equation, how much did the farm population decline each year on average during this period?
- What percent of the observed variation in farm population is accounted for by the linear change over time?
- Use the regression equation to predict the number of people living on farms in 2010. Is this a reasonable result? Why?