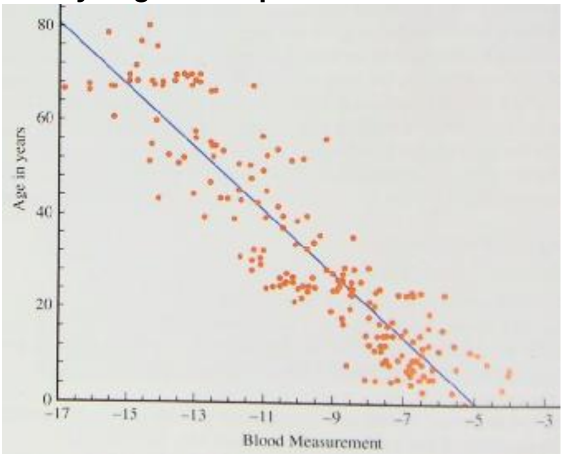
AP statistics Chapter 3 Notes
Describing Relationships

Name ______________________
Per ___ Date ________________

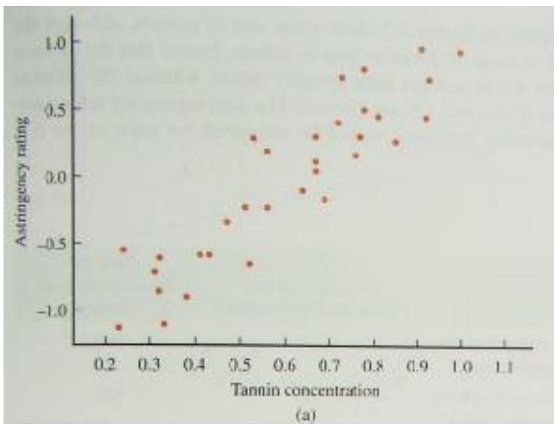
3.1A	**<u>Response variable</u>** A response variable measures the _________________ of a study.
3.1A	**<u>Explanatory variable</u>** An explanatory variable may help ____________________________ changes in a response variable.
3.1A	**Identify the explanatory and response variables in each setting.** 1. How does drinking beer affect the level of alcohol in our blood? The legal limit for drinking in all states is 0.08%. In a study, adult volunteers drank different numbers of cans of beer. Thirty minutes later, a police officer measured their blood alcohol levels. *explanatory variable* *response variable* 2. The National Student Loan Survey provides data on the amount of debt for recent college graduates, their current income, and how stressed they feel about college debt. A sociologist looks at the data with the goal of using amount of debt and income to explain the stress caused by college debt. *explanatory variable* *response variable* 3. Julie wants to know if she can predict a student's weight from his or her height because information about height is easier to obtain than information about weight! *explanatory variable* *response variable*
3.1A	**<u>Scatterplots</u>** Scatterplots are graphs used to display the relationship between two __________________ variables. (For categorical variables, we used two-way tables) • Explanatory variable should be graphed on the _____________ • Response variable should be graphed on the _______________ • ALWAYS label your axes!
3.1A	**<u>Interpreting scatterplots</u>** • **Direction –** a.) **Positive association –** "_______________" as explanatory variable increases __________ the response variable. b.) **negative association –** "_______________" as explanatory variable increases the response variable _______________. • **Form** _____________________________? • **Strength** How close do the points come to forming a _______________ or forming a continuous ______? • **Outliers** Outliers are points that are outside the general pattern of the scatterplot. Usually really far ______________________________ the grouping of points. **Influential points** Another type of outlier that lies outside (____________________) the grouping of points.

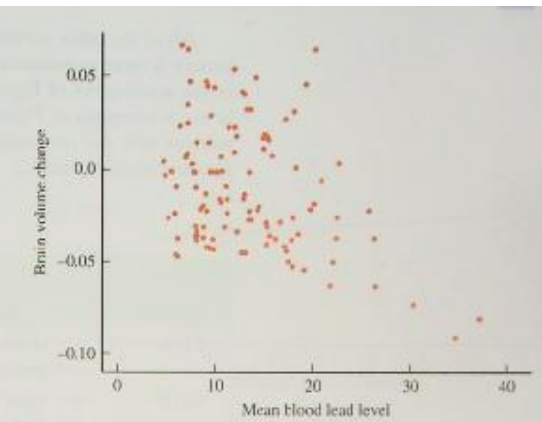
3.1A

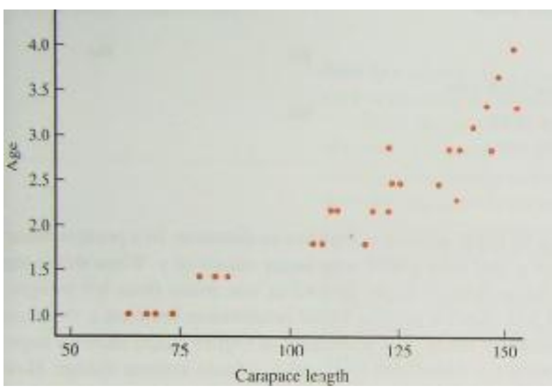
Analyzing Scatterplots

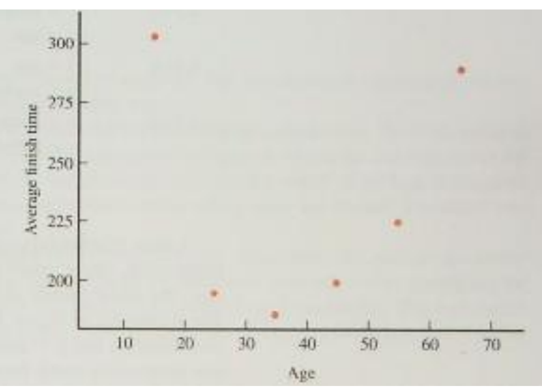


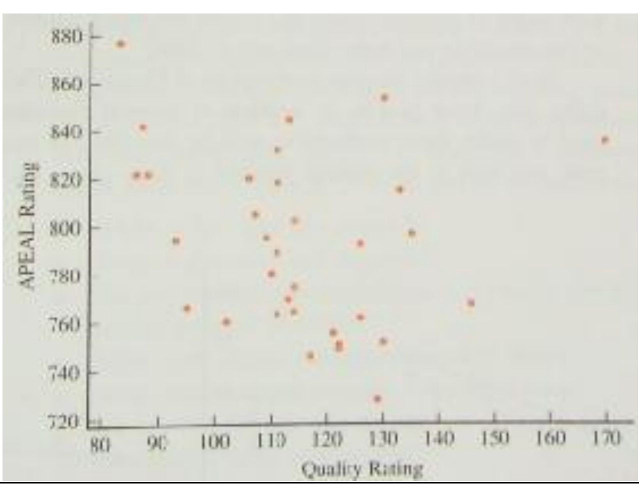
Can you tell a person's age from a blood test? Analyze the scatterplot.











3.1B

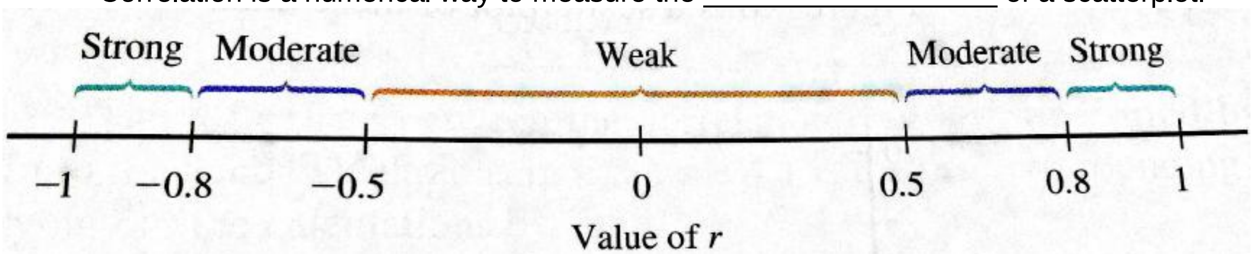
<u>Correlation – “r” value</u>

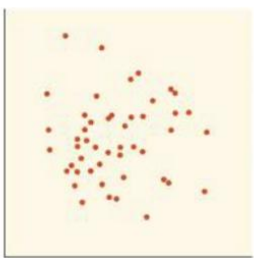
The correlation “r” measures the ______________ of the ______________ relationship between two quantitative variables.

It measures “**How close the data comes to forming a straight line”**

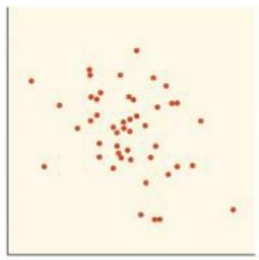
Facts about correlation:

- Correlation (r) is always between and
- A correlation of zero means there is _____ pattern whatsoever
- The closer the number gets to 1, the closer the dots come to forming a straight line with __________ slope (____________ association)
- The closer the number gets to -1, the closer the dots come to forming a straight line with __________ slope (_____________ association)
- Correlation is a numerical way to measure the __________________ of a scatterplot.

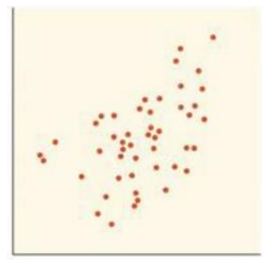




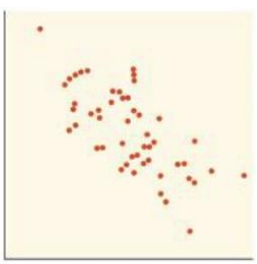
r =



r =

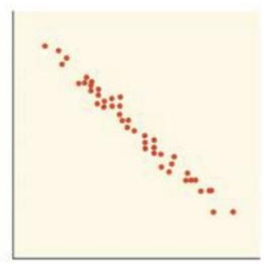


r =



r =

r =



r =

SAMPLE: A recent study discovered that the correlation between the age at which an infant first speaks and the child’s score on an IQ test upon entering elementary school is -0.68. A scatterplot of the data show a linear form. Which of the following statements about this finding is correct?

(a) Infants who speak at very early ages will have higher IQ scores by the beginning of elementary school than those who speak later.

(b) 68 % of the variation in IQ test scores is explained by the least squares regression of age at first word spoken and IQ score.

(c) Encouraging infants to speak before they are ready can have a detrimental effect later in life, as evidenced by their lower IQ scores.

(d) There is a moderately strong, negative linear association between age at first spoken word and later IQ test score for the individuals in this study.

3.1B

Calculating correlation

The correlation coefficient “r” is calculated using the following formula:

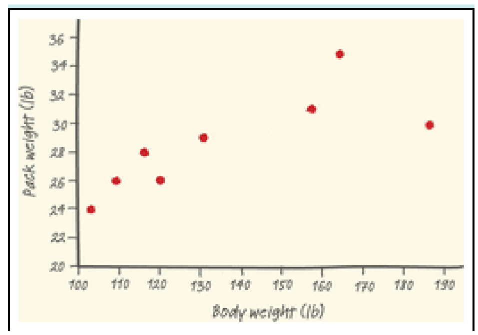
$$r = \frac{\sum z_x z_y}{n-1}$$

$\sum z_x z_y$ is found by multiplying z_x and z_y for each observation in the data set and then adding the $z_x z_y$ values.

3.1B

Calculate the correlation for the following data set

Body weight (lb):	120	187	109	103	131	165	158	116
Backpack weight (lb):	26	30	26	24	29	35	31	28



- Put the values for body weight into list 1, backpack weight into list 2.
- First calculate the mean and standard deviation for each list using 1-variable stats:
  Body weight
  mean = 136.125, standard deviation = 30.3
  Backpack weight
  mean = 28.625, standard deviation = 3.46
- Calculate the z-score for each
- Multiply the z-scores

Body Weight L3 = (L1 – 136.125)/30.3	*Backpack weight* L4 = (L2 – 28.625)/3.46	• Multiply Z-scores (L5 = L3*L4)
$\left(\frac{120-136.125}{30.3}\right) = -0.532$	$\left(\frac{26-28.625}{3.46}\right) = -0.759$	0.404
$\left(\frac{187-136.125}{30.3}\right) = 1.679$	$\left(\frac{30-28.625}{3.46}\right) = 0.397$	0.667
$\left(\frac{109-136.125}{30.3}\right) = -0.895$	$\left(\frac{26-28.625}{3.46}\right) = -0.759$	0.679
$\left(\frac{103-136.125}{30.3}\right) = -1.093$	$\left(\frac{24-28.625}{3.46}\right) = -1.337$	1.461
$\left(\frac{131-136.125}{30.3}\right) = -0.169$	$\left(\frac{29-28.625}{3.46}\right) = 0.108$	-0.183
$\left(\frac{165-136.125}{30.3}\right) = 0.953$	$\left(\frac{35-28.625}{3.46}\right) = 1.843$	1.756
$\left(\frac{158-136.125}{30.3}\right) = 0.722$	$\left(\frac{31-28.625}{3.46}\right) = 0.686$	0.496
$\left(\frac{116-136.125}{30.3}\right) = -0.664$	$\left(\frac{28-28.625}{3.46}\right) = -0.181$	0.120

- Add the last column: (stat, calc, 1 var stats, L5) = $\sum x = 5.565$
- Divide the summation by n-1 = number of values – 1 = 8 – 1
- r = 5.565/7 = 0.795

3.1B

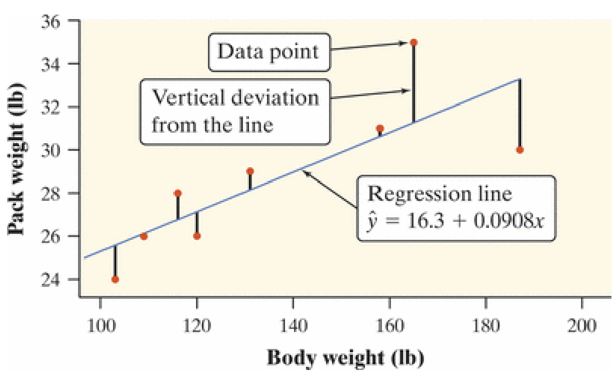
Facts about correlation

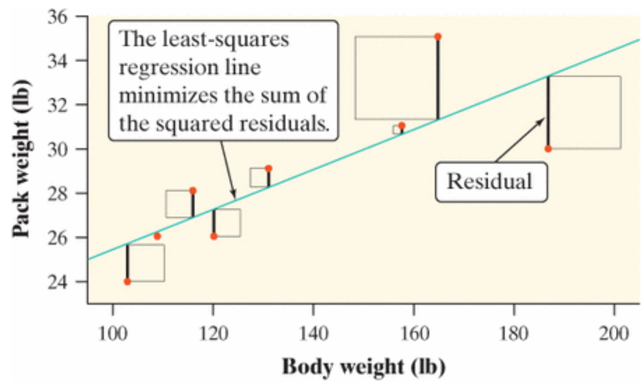
1. Both values MUST be ________________, correlation does not work for categorical data.
2. If you switch your response variable and explanatory variable, the correlation _______ change.
3. If your change the units of any variable, the correlation __________________ change.
4. Correlation has ______ unit of measure.
5. A correlation close to 1, does not guarantee that it is ________________________.
6. It is ______________ by outliers since the mean and standard deviation (which are used to calculate correlation) are affected by outliers.
7. Correlation only measures how close the data fits to a ___________, not how it fits to a curve.

3.2A

Trend Line

A trend line is a line that ___________ the data from a scatterplot. What happens if you can't tell which trend line best represents the data? We analyze the ___________________ between the actual points and the line. We ____________ those deviations (to get rid of the negative values) and add those values together. The line with the lowest (least sum) is called the **Least Squares Regression Line**.





We can use our calculator to calculate the regression line, which can also calculate a residual for us.

- Enter your x-values (explanatory) into List 1 and your y-values (response) into List 2
- Stat
- Calc
- Arrow down to LinReg (ax+b) OR **LinReg (a+bx)**
- Select L1 for x list, L2 for y list, leave FreqList blank, Store RegEQ (this is to graph the line with your scatterplot, if you want to graph it – hit VARS, Y-VARS, Function, Y1
- Hit enter
- It will give you the parts for your equation, r which is the correlation, and r^2 which we talk about later.

3.2A

Interpreting a regression line

Regression lines are usually written in the form:

Where $\mathbf{b_0}$ is your y-intercept

Interpret y-intercept: (response value) when your (explanatory value) is zero

Where $\mathbf{b_1}$ your slope/rate of change

Interpret slope:

(response variable) changes by ____ for every increase of 1 in (explanatory variable)

Sample Problem: Used Hondas

The following data show the number of miles driven and advertised price for 11 used Honda CR-Vs from the 2002-2006 model years (www.carmax.com).

Thousand Miles Driven	Cost (dollars)
22	17998
29	16450
35	14998
39	13998
45	14599
49	14988
55	13599
56	14599
69	11998
70	14450
86	10998

a) Use the calculator to find the equation of the regression line.

What is the correlation?

What is the equation? In context?

b.) Interpret the slope in context:

c.) Interpret the y-intercept in context:

3.2A

<u>Predicting from a regression line</u>

The purpose of having a regression equation is to able to ___________ what the response value MIGHT be with a certain explanatory value. That is why we use the symbol $\left(y \right)$ instead y.

- "***y***" is the ______________ from the scatterplot
- y (_____________) is the ____________________ based on the regression line.

The difference (_______________________) also $(y - y)$ is called the ______________ and is the same as the vertical deviation used for the least squares regression line.

Sample: Using the previous problem, predict the price for a car with 49,000 miles. Compare that to the actual price.

3.2A

<u>Extrapolation</u>

Extrapolation occurs when you use the regression line to predict for a value ____________ the data's domain (x-values). If you only have data for the explanatory variable from 10 to 50, you CANNOT predict a value lower than 10, or higher than 50. Since we _________________ the behavior of the data outside this domain, we take a huge risk trying to predictions outside those values.

Sample: Using the previous problem, should we predict the asking price for a used 2002-2006 Honda CR-V with 250,000 miles? Explain.

3.2B

<u>Residuals of the least-squares regression line</u>

A **residual** is the difference between an observed (Actual) value of the response variable and the value Predicted by the regression line.

- A negative residual means we ____________________ the response value
- A positive residual means we ____________________ the response value

Sample: Back to the Track

Sprint Time (s)	5.41	5.05	9.49	8.09	7.01	7.17	6.83	6.73	8.01	5.68	5.78	6.31	6.04
Long Jump Distance (in)	171	184	48	151	90	65	94	78	71	130	173	143	141

The equation of the least-squares regression line for the sprint time (x) and long-jump distance (y) data is $y = 304.56 - 27.63x$. Find and interpret the residual for the student who had a sprint time of 8.09 seconds.

3.2B

Calculating regression with means and standard deviation

We have used technology to find the least-squares regression line, but we can also find it using means, standard deviations, and their correlation.

If we know the mean ($\bar{x}$) and standard deviation (S_x) of our explanatory variable, mean ($\bar{y}$) and standard deviation (S_y) of our response variable, and their correlation (r) then the equation of the least-squares regression line

- With
- Slope

  y-intercept

- All least-squares regression lines will run through the point

Sample: Used Hondas The number of miles (in thousands) for the 11 used Hondas has a mean of 50.5 and a standard deviation of 19.3. The asking prices had a mean of $14,425 and a standard deviation of $1,899. The correlation for these variables is *r* = -0.874.

a) Find the equation of the least-squares regression line

b) Explain what change in price we would expect for each additional 19.3 thousand miles.

What happens if we standardize both variables?

We can standardize by changing all values to ____________________ which will have a mean of 0 and a standard deviation of 1

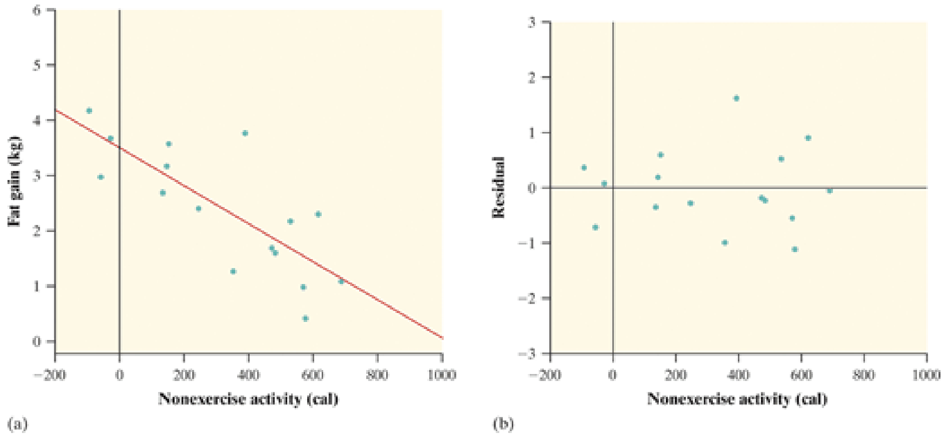
A) slope will become

B)

3.2C

<u>Is a line the best choice for our data? – residual plots</u>

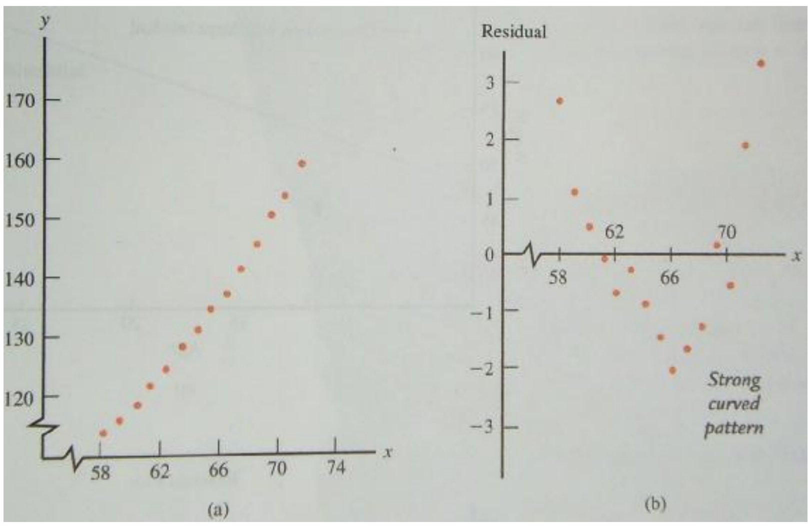
A careful look at the residuals can reveal many potential ______________. To assess the appropriateness of the regression line, a ___________________ is a good place to start



A **residual plot** is a scatterplot of the

pairs.

Important: The sum of the residuals is ALWAYS _____!



Isolated points or a pattern of points in the residual indicate ____________ problems

A desirable residual plot is one that exhibits ___ particular pattern, such as _______________.

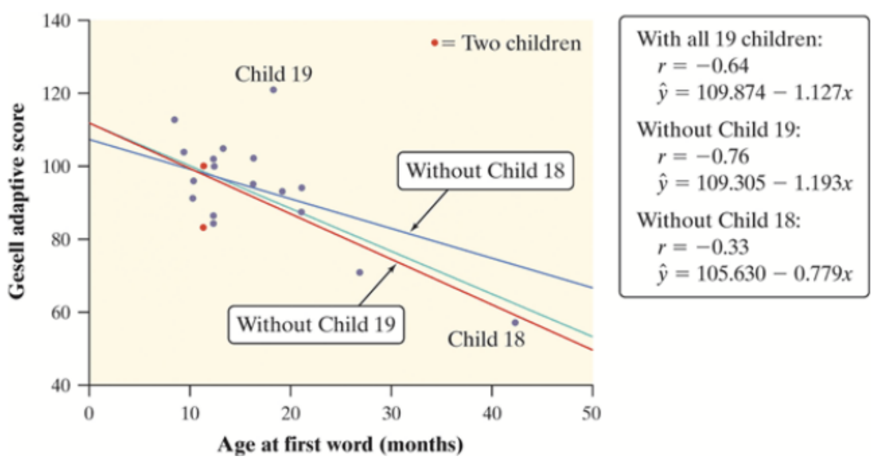
Sometimes it easier to detect ________________ in a residual plot than in a scatterplot.

3.2C

<u>Outliers and influential points with the regression line</u>

An **outlier** is an observation that lies outside the overall pattern of the other observations. Points that are outliers in scatterplots have _____________ therefore lie outside the pattern in the y-values (too high, too low).

An observation is **influential** if it is an outlier in the ________________. Removing an influential point would markedly change the result of the correlation and regression line.



**We delete students ONLY to investigate influence. In practice, we cannot delete without VERY GOOD justification.

3.2C

Coefficient of determination – r^2

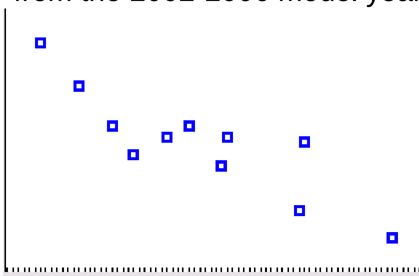
There is another numerical quantity that tells us ______________ the least-squares regression line predicts values of the response *y*. It also happens to be the correlation _______________.

We interpret this value

“The regression line accounts for __% of the variation in the (response variable).”

The following data show the number of miles driven and advertised price for 11 used Honda CR-Vs from the 2002-2006 model years (www.carmax.com).

Thousand Miles Driven	Cost (dollars)
22	17998
29	16450
35	14998
39	13998
45	14599
49	14988
55	13599
56	14599
69	11998
70	14450
86	10998



a) Enter the values into your lists and graph the scatterplot on your calculator.

b.) Use the calculator to find the equation of the regression line.

What is the equation?

What is the correlation?

What is the coefficient of determination?

Interpret the coefficient of determination:

Practice:

1. For the least squares regression of fat gain from Non-Exercise Activity, $r^2 = 0.606$. Which of the following gives the correct interpretation in context?
   (a) 60.6% of the points lie on the least squares regression line.
   (b) 60.6% of the fat gain values are accounted for by the least squares line.
   (c) 60.6% of the variation in fat gain is accounted for by the least squares line.
   (d) 60.6% of the variation in Non-exercise Activity is accounted for by the least squares line

3.2C

Standard deviation of the residuals – **“S”**

Standard deviation of the residuals gives us the ________________________________

$$s = \sqrt{\frac{\sum residuals^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n-2}}$$

Interpreting “s” –

"the average error in predicting (response variable) is ___ (s) using the least squares regression line"

Sample: Used Hondas

Thousand Miles Driven	Cost (dollars)
22	17998
29	16450
35	14998
39	13998
45	14599
49	14988
55	13599
56	14599
69	11998
70	14450
86	10998

The following data show the number of miles driven and advertised price for 11 used Honda CR-Vs from the 2002-2006 model years (www.carmax.com).

A) Enter the values into your lists

B) Calculate the linear regression

C) Back in your lists, highlight L3, 2^{nd} stat (list), arrow down to RESID, enter, enter

D) Your L3 now has all of your residuals

F) To calculate s - the standard deviation of the residuals

Calculate the one-variable statistics of L3 (or just RESID)

(stat, calc, 1-var stat)

G) Interpret the s value

3.2D	**Computer output of regression line**

Computer output of regression line

Many times you will be presented with numerical information from different sources: graphing calculators, fathom, mini-tab, JMP. Let's look at several examples to see if you can find information.

Example: Body Weight and Pack Weight

Minitab Output:

```
Predictor        Coef  SE Coef     T      P
Constant       16.265    3.937  4.13  0.006
Body Weight   0.09080  0.02831  3.21  0.018


S = 2.26954   R-Sq = 63.2%   R-Sq(adj) = 57.0%
```

1. What is the equation for the regression line?
2. What is the typical prediction error?
3. What is the coefficient of determination?
4. What is the correlation?

Alternate Example: Used Hondas

Minitab Output:

```
Predictor     Coef  SE Coef      T      P
Constant   18773.3    856.2  21.93  0.000
Miles       -86.18    15.95  -5.40  0.000


S = 971.647   R-Sq = 76.4%   R-Sq(adj) = 73.8%
```

1. What is the equation for the regression line?
2. What is the typical prediction error?
3. What is the coefficient of determination?
4. What is the correlation?

JMP Output: Age and reading scores

Summary of Fit

RSquare	0.409971
RSquare Adj	0.378917
Root Mean Square Error	11.02291
Mean of Response	93.66667
Observations (or Sum Wgts)	21

Parameter Estimates

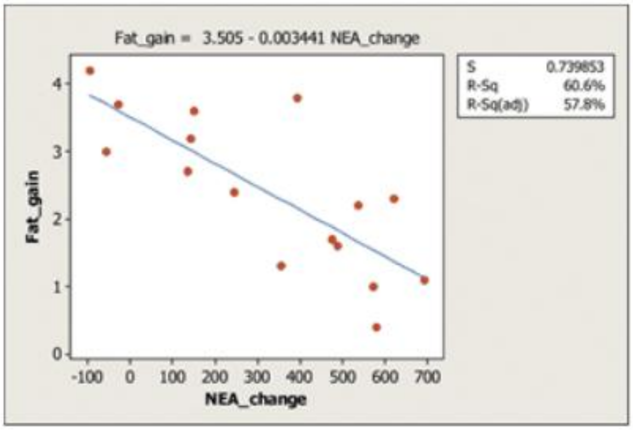
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	109.87384	5.067802	21.68	<.0001
Age	-1.126989	0.310172	-3.63	0.0018

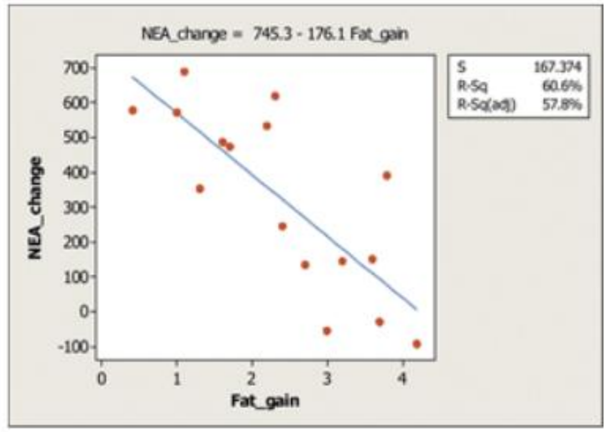
1. What is the equation for the regression line?
2. What is the typical prediction error?
3. What is the coefficient of determination?
4. What is the correlation?

3.2D	**<u>Correlation and regression wisdom</u>**

Correlation and regression are powerful tools for describing the relationship between two variables. When you use these tools, be aware of their ______________________.

1. The distinction between explanatory and response variables is important in regression.



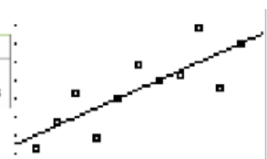


2. Correlation and regression lines describe only _________________ relationships.

Be careful about calculating without _________________ the data! All of the data sets below have a regression equation of $y = 3 + 0.5x$

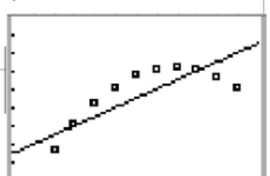
Data Set A

x:	10	8	13	9	11	14	6	4	12	7	5
y:	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68



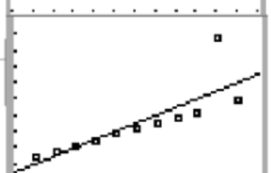
Data Set B

x:	10	8	13	9	11	14	6	4	12	7	5
y:	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74



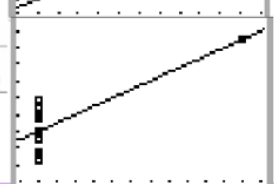
Data Set C

x:	10	8	13	9	11	14	6	4	12	7	5
y:	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73



Data Set D

x:	8	8	8	8	8	8	8	8	8	8	19
y:	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50



3. Correlation and least-squares regression lines are ___________________________!

3.2D	**<u>Association vs causation</u>** Association does not imply CAUSATION! An association between an explanatory variable x and a response variable y, even if it is very strong, is not by itself good evidence that changes in x actually cause changes in y.  A serious study once found that people with two cars live longer than people who only own one car. Owning three cars is even better, and so on. There is a substantial positive correlation between number of cars *x* and length of life *y*. Why?