Chapter 7 Scatterplots, Association, and Correlation

"Correlation", "association", "relationship" between two sets of numerical data is often discussed. It's believed that there is a relationship between amount of smoked cigarettes and likelihood (in percent) to get a lung cancer; between the number of cold days in winter and number of babies born next fall; even the values of Dow Jones Industrial Average and the length of fashionable skirts show an association! (For more of surprising relations see "crazy correlations" http://tylervigen.com/spurious-cor)

Questions to ask about paired data:

1. Is there a relationship?
2. Can I find an equation that describes it?
3. How good my find is? Can I use it to make predictions?

A way to observe such relationships is constructing a scatter plot.

A scatter diagram (scatter plot) is a graph that displays a relationship between two **quantitative** variables. Each point of the graph is plotted with a pair of two related data: x and y. Each individual (case or subject) in the data set is represented by a point in the scatter diagram.
In a scatter plot a variable assigned to x-axis is called **explanatory (or predictor)**, and a variable assigned to y-axis a **response variable.** Often a response variable is a variable that we want to predict.
The explanatory variable is plotted on the horizontal axis, and the response variable is plotted on the vertical axis.

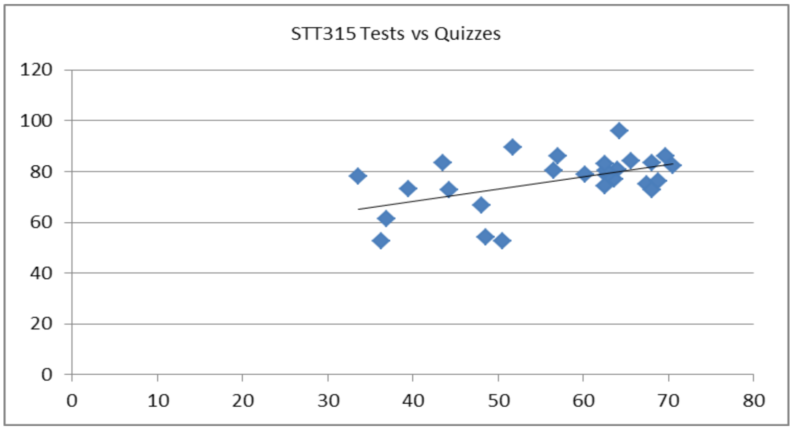
Things to look at:

- Direction (negative or positive)
- Strength (no, moderate, strong)
- Form (linear or not)
- Clusters, subgroups and outliers

Example: The results recorded in Summer 16 section of a Stat course are collected in two columns: "Quizzes" represents average grade for MML homework quizzes. The second column represents averaged grade for Tests. Twenty seven students took the tests. The predictor is average homework quizzes grade, and response is a test grade.

Homework	TESTS
44.3	**72.9**
69.7	**86.3**
64.1	**80.7**
70.6	**82.3**
65.6	**84.2**
48.6	**54.1**

67.5	74.9
63.7	76.9
60.2	78.8
33.6	78.1
64.3	95.9
36.9	61.2
62.8	78.5
39.5	73.0
57.1	86.2
50.5	52.7
43.6	83.4
62.7	80.2
56.5	80.4
68.1	83.3
68.8	76.3
62.6	74.2
51.8	89.3
48.1	66.6
68.1	72.8
62.6	83.1
36.3	52.5



Another Example:

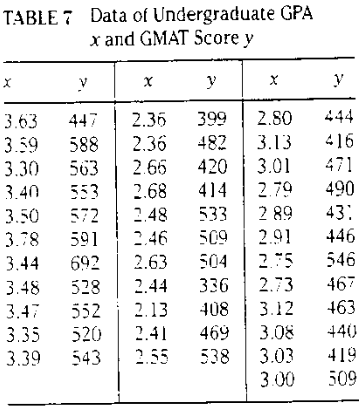
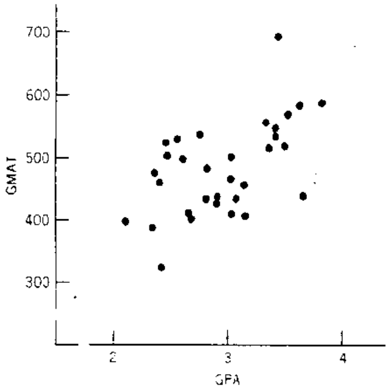
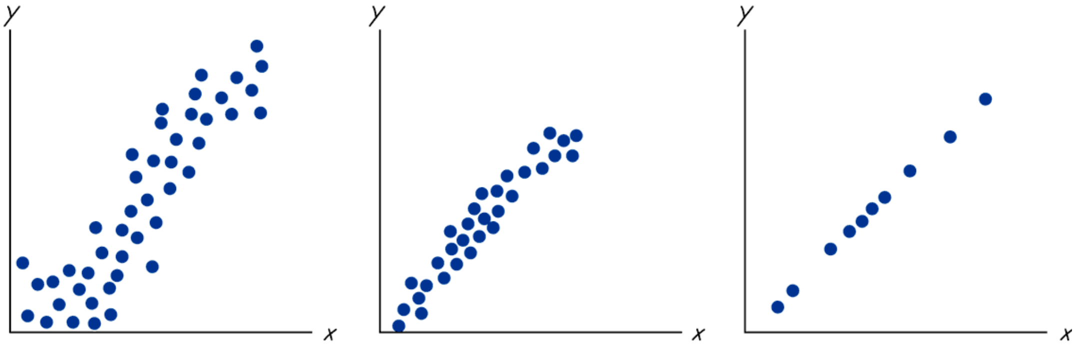


TABLE 7 Data of Undergraduate GPA x and GMAT Score y

x	y	x	y	x	y
3.63	447	2.36	399	2.80	444
3.59	588	2.36	482	3.13	416
3.30	563	2.66	420	3.01	471
3.40	553	2.68	414	2.79	490
3.50	572	2.48	533	2.89	431
3.78	591	2.46	509	2.91	446
3.44	692	2.63	504	2.75	546
3.48	528	2.44	336	2.73	467
3.47	552	2.13	408	3.12	463
3.35	520	2.41	469	3.08	440
3.39	543	2.55	538	3.03	419
				3.00	509



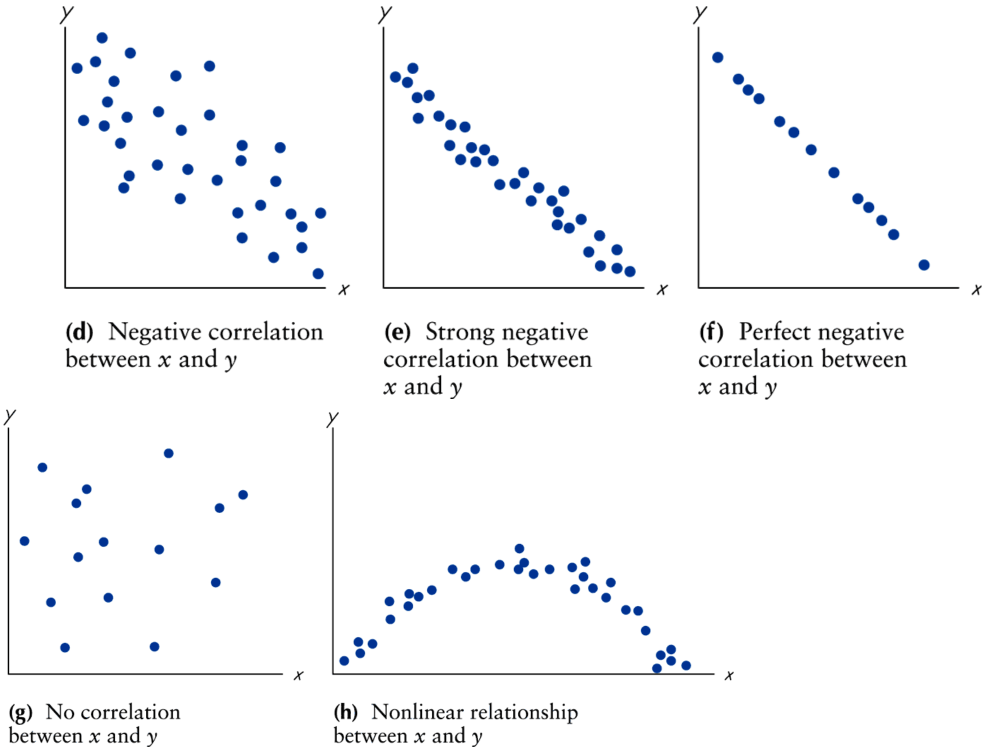
Correlation: linear relationship between two quantitative variables



(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 relationship between *x* and *y*

Correlation Coefficient r is a measure of the strength of the linear association between two quantitative variables.

Properties

1. The sign gives direction
2. *r* is always between –1 and 1; 1 is a perfect positive correlation and -1 is a perfect negative correlation
3. *r* has no units
4. Correlation is not affected by shifting or re-scaling either variable.
5. Correlation of x and y is the same as of y and x
6. *r= 0* indicates lack of linear association (but **could be strong non-linear association**)
7. Existence of strong correlation does not mean that the association is **causal**, that is change of one variable is caused by the change of the other (it may be third factor that causes both variables change in the same direction)

Before you use correlation, you must check several conditions:

- Quantitative Variables Condition
- Straight Enough Condition
- Outlier Condition

If you notice an outlier then it is a good idea to report the correlations with and without that point.

Question: HOW BIG (or how small) the correlation coefficient must be to consider the **significant correlation** between the explanatory and response variables?

Answer: It depends on the size of the sample**.** The farther from zero is r, the stronger correlation. For instance, for n=10, a significant correlation starts with |r|>0.68, for n=50 |r|>0.35, but for n=100 you just need |r|>0.25 to call the correlation significant. In our case “Test grades vs Homework Quizzes grades” (example 1) for n=27 students observed coefficient is 0.514. We observe a possibly moderate linear relationship between quiz grades and test grades.

Next pictures and example comes from http://www.mathsisfun.com/data/correlation.html

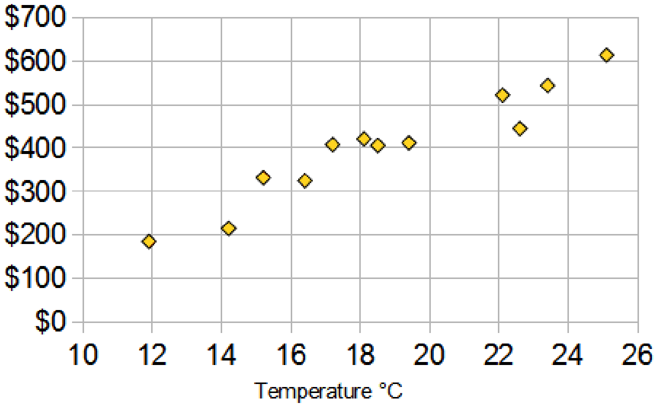
Example: Ice Cream Sales

The local ice cream shop keeps track of how much ice cream they sell versus the temperature on that day, here are their figures for the last 12 days:

Ice Cream Sales vs Temperature

Temperature °C	Ice Cream Sales
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408

And here are the same data on a Scatter Plot:

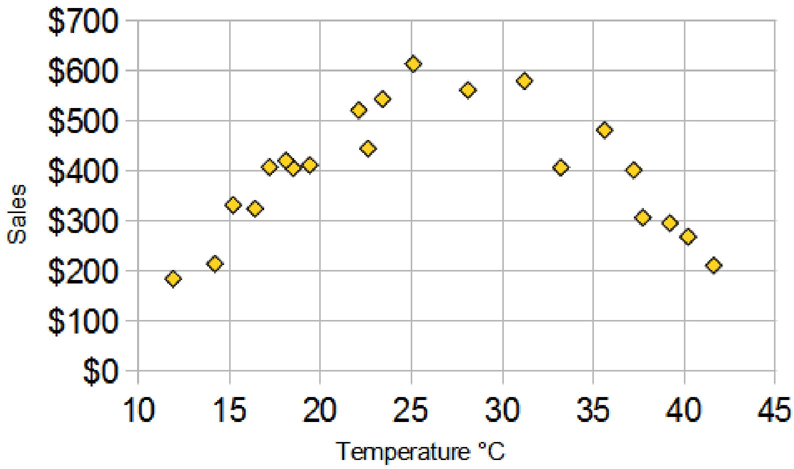


We can easily see that warmer weather leads to more sales, the relationship is good but not perfect.
In fact the correlation is VERY strong: 0.9575! (this was easily computed with EXCEL)

WATCH OUT: **Correlation Is Not Good at Curves**
The correlation calculation only works well for relationships that follow a straight line.

Our Ice Cream Example: there has been a heat wave!
It gets so hot that people aren't going near the shop, and sales start dropping.

Here is the latest graph:

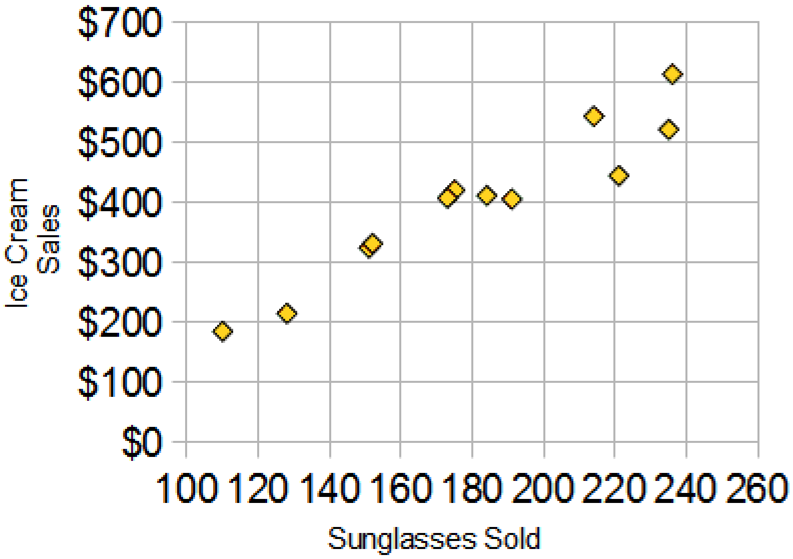


The calculated value of correlation is 0 which says there is "no correlation".
But we can see the data follows a nice curve that reaches a peak around 25° C. But the correlation calculation is not "smart" enough to see this.

More of this: click HERE.

A strong correlation does not mean that one thing causes the other (there could be other reasons the data has a good correlation).

Example: Sunglasses vs Ice Cream
Our Ice Cream shop finds how many sunglasses were sold by a big store for each day and compares them to their ice cream sales:



The correlation between Sunglasses and Ice Cream sales is high
Does this mean that sunglasses make people want ice cream? That eating ice-cream makes people want to buy sunglasses? Or is there another variable as the weather which causes grow of both numbers?

REMEMBER!

Correlation measures the strength of the *linear* association between two *quantitative* variables.
Correlation does not explain causation!

Before you use correlation, you must check if:

- Is your variable quantitative?
- Is your scatterplot looking strong? Straight?
- Any outliers?

Strength of correlation is measured with Correlation Coefficient r and Coefficient of Determination r^2.

Correlation Coefficient on TI-83.

Correlation coefficient is not displayed automatically. We have to set the diagnostic display mode. First go to the CATALOG: *press 2nd, 0, press D (the key with x^{-1})* and scroll down the screen to point to *Diagnostic On,* then press *ENTER twice*. The calculator responds with "Done". You need to do this only once for all computations. Then:

1. Enter x-values in L1 and y-values in L2.
2. Press STAT, select CALC, select LinReg(ax+b), ENTER.
3. The display says *LinReg(ax+b)*. Enter L1,L2 and then ENTER.

   The display will look like this:

   y = ax+b
   a = …
   b =**…**
   r^2 =… (← this is the coefficient of determination)
   r =…. (← this is the correlation coefficient)

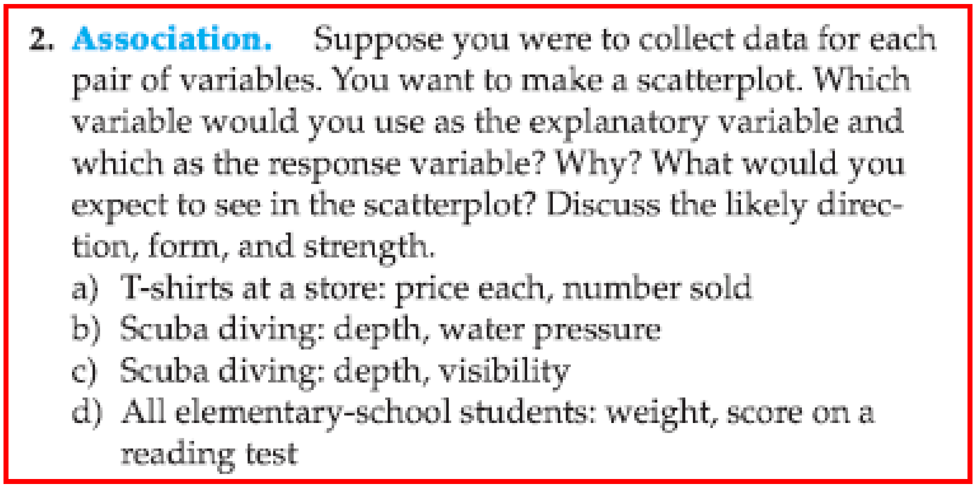
Common Errors Involving Correlation

1. Causation: It is wrong to conclude that correlation implies causality.

2. Averages: Averages suppress individual variation and may affect the correlation coefficient.

3. Linearity: There may be some nonlinear relationship between x and y even when there is no significant linear correlation.

Class Exercises: Ch 7: 2, 6, 10, 12, 26



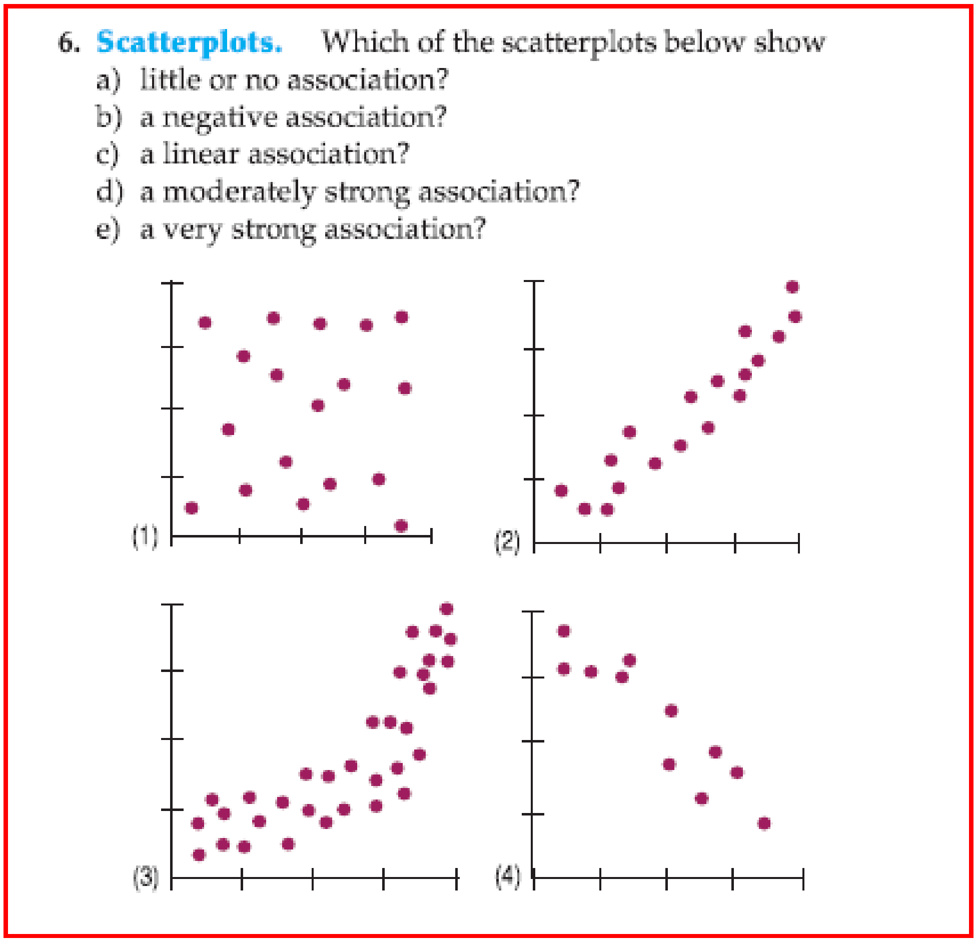
2. Association. Suppose you were to collect data for each pair of variables. You want to make a scatterplot. Which variable would you use as the explanatory variable and which as the response variable? Why? What would you expect to see in the scatterplot? Discuss the likely direction, form, and strength.

a) T-shirts at a store: price each, number sold

b) Scuba diving: depth, water pressure

c) Scuba diving: depth, visibility

d) All elementary-school students: weight, score on a reading test



6. Scatterplots. Which of the scatterplots below show

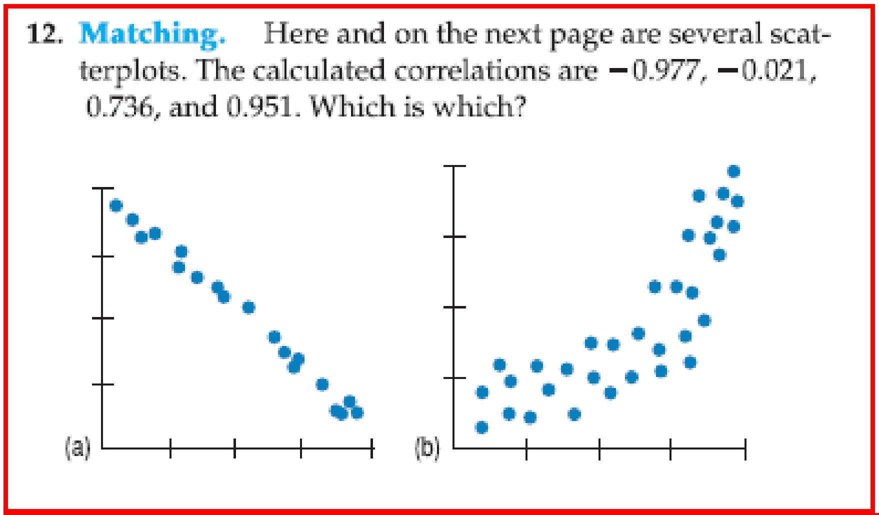
a) little or no association?

b) a negative association?

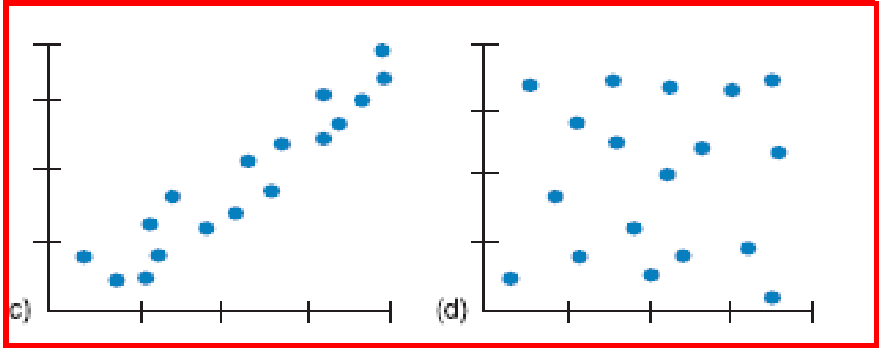
c) a linear association?

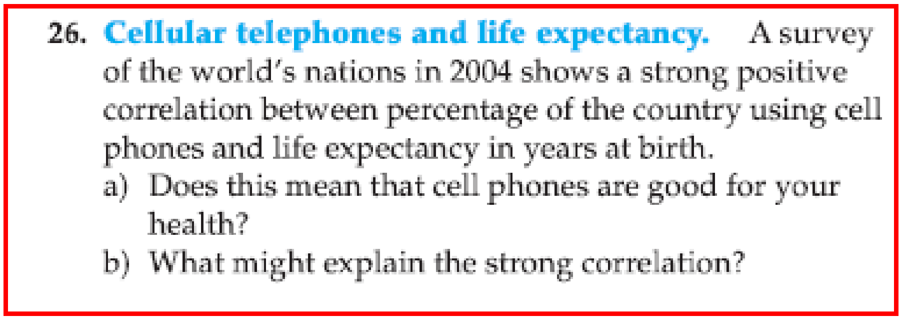
d) a moderately strong association?

e) a very strong association?



12. Matching. Here and on the next page are several scatterplots. The calculated correlations are −0.977, −0.021, 0.736, and 0.951. Which is which?





26. Cellular telephones and life expectancy. A survey of the world's nations in 2004 shows a strong positive correlation between percentage of the country using cell phones and life expectancy in years at birth.

a) Does this mean that cell phones are good for your health?

b) What might explain the strong correlation?

Chapter 8 – Linear Regression

Linear regression model is a model of a relationship between two variables x, and y

Response = linear function of the predictor + Error

$$y = b_o + b_1 x + Error$$

b_o and b_1 are parameters of the model.

Goal: Estimate b_o and b_1 and the regression line $\hat{y} = b_o + b_1 x$

Method: Least squares regression line – the line that minimizes the sum of squared vertical distances between points on the scatterplot and the line (regression line, best fit line, prediction line).

Review from algebra: the equation describing a straight line is $y = b + mx$

where ***b=y-intercept, and m=the slope of the line***

We'll denote the line as $\hat{y} = b_0 + b_1 x$

The coefficient b_1 *is the slope, which tells us* ***the average change of y is if x changes one unit.***
The coefficient b_0 *is the intercept, which tells where* ***the line crosses (intercepts) the y-axis.***

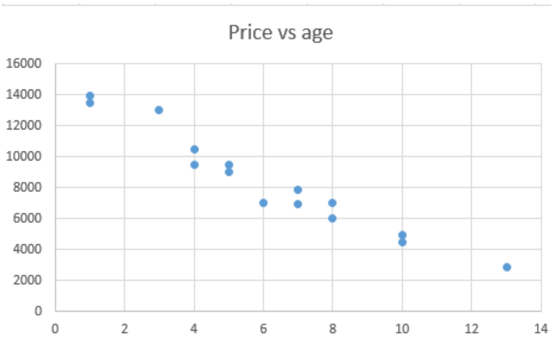
To interpret the *y*-intercept, we must first ask two questions:

1. Is 0 a reasonable value for the explanatory variable?
2. Do any observations near *x* = 0 exist in the data set?

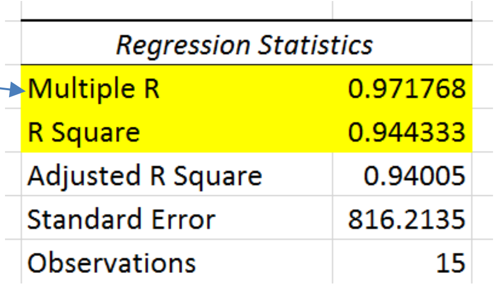
A linear model of two variables can be easily found on the calculator: put your data into L1 and L2, then go STAT, CALC, 4 (or 8)

Example: Used cars lot

Age (yr)	Price Advertised ($)
1	13990
1	13495
3	12999
4	9500
4	10495
5	8995
5	9495
6	6999
7	6950
7	7850
8	6999
8	5995
10	4950
10	4495
13	2850



Correlation coefficient

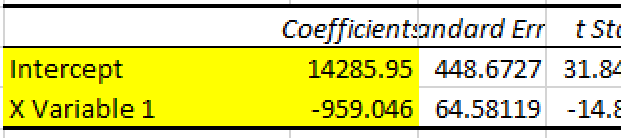


Regression Statistics	
Multiple R	0.971768
R Square	0.944333
Adjusted R Square	0.94005
Standard Error	816.2135
Observations	15

Predictor: x= age; **Response:** advertised price
The model: $\hat{y}$**=b₀+b₁x;** that is,

<$\widehat{price}$> = y-intercept + slope * <age>

Using software to find b_0 and b_1, the line is



	Coefficients	tandard Err	t St
Intercept	14285.95	448.6727	31.84
X Variable 1	-959.046	64.58119	-14.8

<price> = 14 285.95 – 959.046 <age>

Interpret the slope: On average, each additional year of age takes away $959.05 of the price.

Interpret the intercept: new car, on average, would be advertised for $14,285.95

This value does not always make sense, but is useful as a starting point for the graph.

With the existing model we can **make prediction** for the price of a 10 year old car

$\widehat{price} = 14285.95 - 959.046 * 10$ =$4695.49

If one of the observed values for x=10 was y=4950, we can compute the difference between the observed and predicted value of fat:

$y - \hat{y}$ =*4950 – 4695.49 =254.51*

This is called **the residual. Interpretation:** This 10-year-old car is advertised above the predicted value by $*254.51*

Check and interpret the difference for the other 10 year old car on the list.

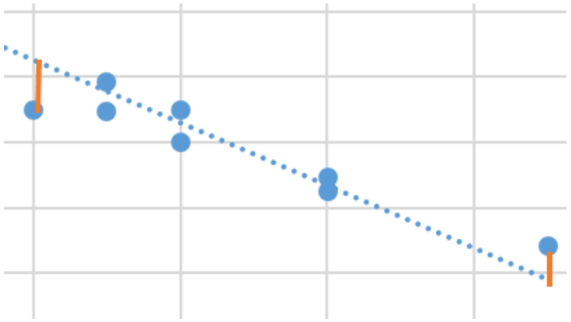
Question: can we use the line to predict the price for a 20 year old car?

....

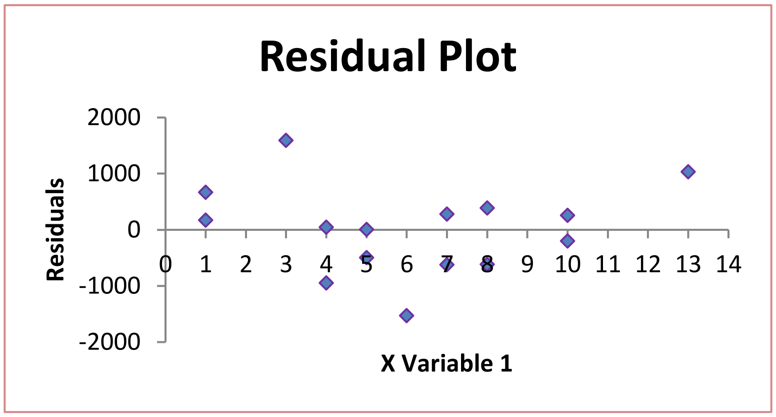
Residuals: $e = observed - predicted = y - \hat{y}$

- if e **is positive** then model **underestimates** the actual data value
- if e is **negative** then model **overestimates** the actual data value

- Some residuals are positive, others are negative, and, on average, they cancel each other out.



(Vertical distances between points and the line represent the residuals on this gaph)



When is a linear model reasonable?

1. **Linear relationship shown on scatterplot**
2. **Large R^2**
3. **Residuals should be evenly distributed around zero.**

R^2 is called the **coefficient of determination**, and tells the percentage of the total variation of y explained by x

Interpretation of *R^2*:

R^2=0.69 means that 94.4% of the variation in advertised prices is explained for by variation in cars' ages.

Regression Assumptions and Conditions

Quantitative Variables Condition:

Regression can only be done on two **quantitative** variables

Straight Enough Condition:

The linear model assumes that the relationship between the variables is linear. A scatterplot will let you check that the assumption is reasonable.

Residuals

Their scatterplot should not show any pattern and the points should be evenly spread around zero. Simply, we want to see "nothing special" in a plot of the residuals

Outliers

You should also check for outliers, which could change the regression. Even a single outlier can dominate the correlation value and change regression line

Example:

1. Use the calculator or computer to make a scatter plot and to find correlation coefficient and coefficient of determination for the following data

x	y
0	0
0	1
1	0
2	0
0	2
1	1
1	2
2	2
2	1

2. Add one more point: (10, 10) and compute the correlation coefficient again. Any change?

What can go wrong?

- Don't fit a straight line to a nonlinear relationship.
- Beware of extraordinary points (*y*-values that stand off from the linear pattern or extreme *x*-values).
- Don't invert the regression. To swap the predictor-response roles of the variables, we must fit a new regression equation.
- Don't extrapolate beyond the data—the linear model may no longer hold outside of the range of the data.
- Don't infer that *x* causes *y* just because there is a good linear model for their relationship—association is *not* causation.
- Don't choose a model based on R^2 alone, scatterplot alone, or residual plot alone. Use all three.

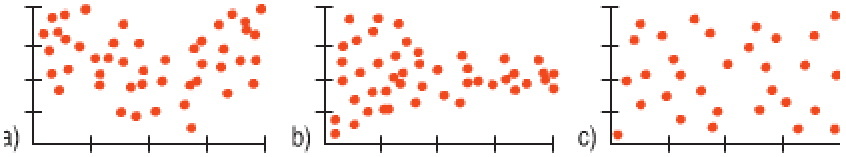
2. **Horsepower.** In Chapter 7's Exercise 35 we examined the relationship between the fuel economy (mpg) and horsepower for 15 models of cars. Further analysis produces the regression model $\widehat{mpg} = 46.87 - 0.084HP$. If the car you are thinking of buying has a 200-horsepower engine, what does this model suggest your gas mileage would be?

4. **Horsepower, again.** Exercise 2 describes a regression model that uses a car's horsepower to estimate its fuel economy. In this context, what does it mean to say that a certain car has a positive residual?

6. **More horsepower.** In Exercise 2, the regression model $\widehat{mpg} = 46.87 - 0.084HP$ relates cars' horsepower to their fuel economy (in mpg). Explain what the slope means.

8. **Another car.** The correlation between a car's horsepower and its fuel economy (in mpg) is $r = -0.869$. What fraction of the variability in fuel economy is accounted for by the horsepower?

14. **Residuals.** Tell what each of the residual plots below indicates about the appropriateness of the linear model that was fit to the data.



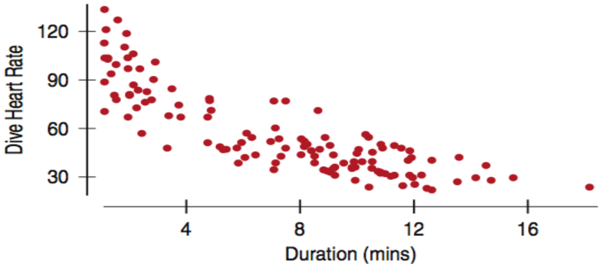
Chapter 9 Regression Wisdom (Skip Chapter 10)

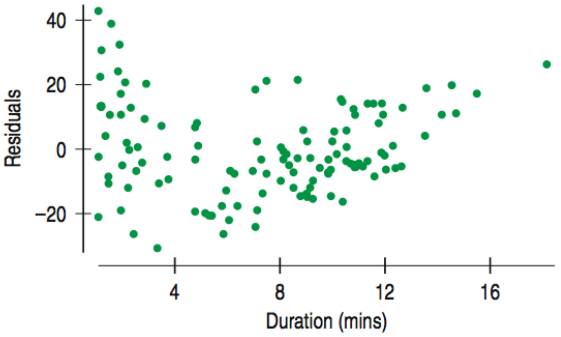
Residuals - should be "evenly" distributed around zero; their scatterplot has no visible pattern, and their histogram should be bell-shaped, centered at zero

Check the **scatterplot of the residuals** for bends that you might have overlooked in the original scatterplot.

Residuals often show patterns that were not clear from a scatterplot of the original data.

Scatterplot: Duration vs. heart rate Scatterplot for the Residuals:



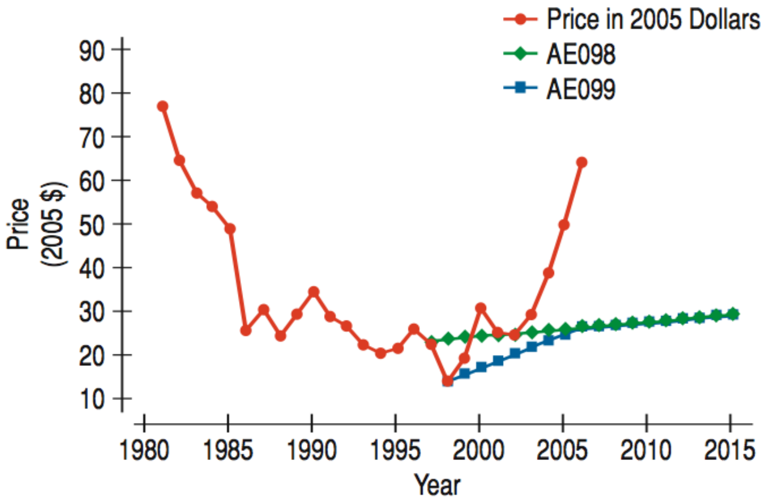


(Scatterplot for residuals shows a pattern which indicates that linear model is not appropriate here)

Predictions:
Linear models let us predict the value of y for each case x in the data.

BUT:

We cannot assume that a linear relationship in the data exists beyond the range of the data. Such a prediction is called an extrapolation.



Above is a time-plot of the Energy Information Administration (EIA) predictions and actual prices.

Prediction - reasonable only in the range of x-values or **slightly beyond**

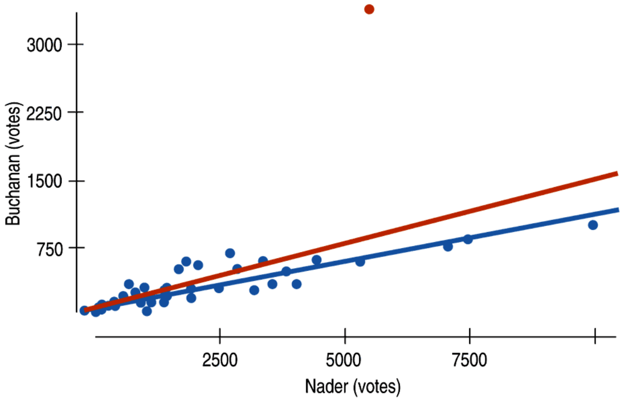
- Extrapolation: outside data range - not reliable
- Interpolation: inside data range

Lurking variables. Strong association does not mean that one variable **causes** the change of other. The fact that both variables *x* and *y* change simultaneously could be due to another, so called **lurking,** variable.

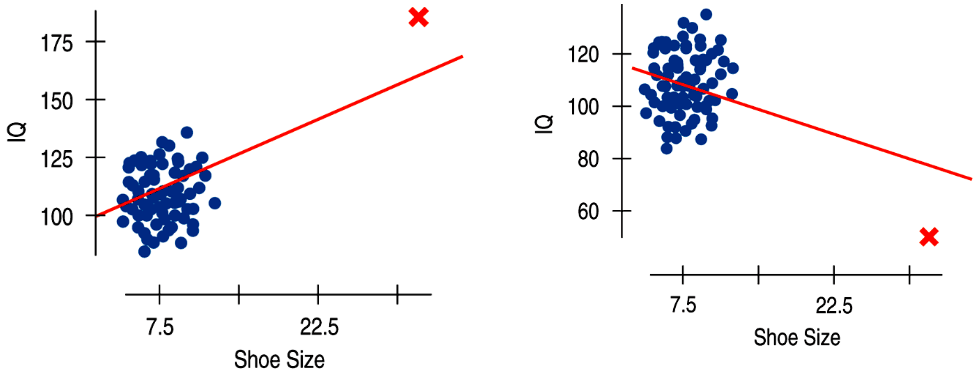
Outliers, Leverage, and Influence

Outliers - points that stands away from others. If there is an outlier you should build 2 models, with and without the outlier, and compare them.

The red line shows the effects that one unusual point can have on a regression:



We say that a point is influential if omitting it from the analysis gives a very different model.
The extraordinarily large shoe size gives the data point high leverage. Wherever the IQ is, the line will follow!



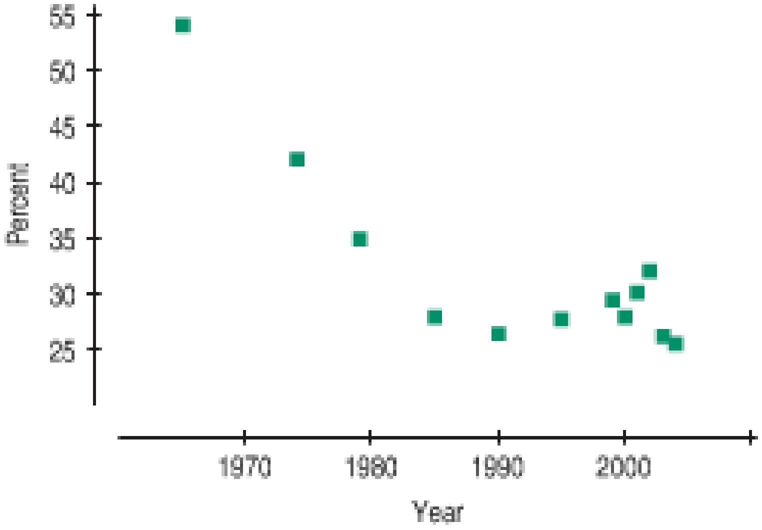
You cannot simply delete unusual points from the data. You can fit a model with and without these points, and examine the two regression models to understand how they differ.

Lurking Variables and Causation

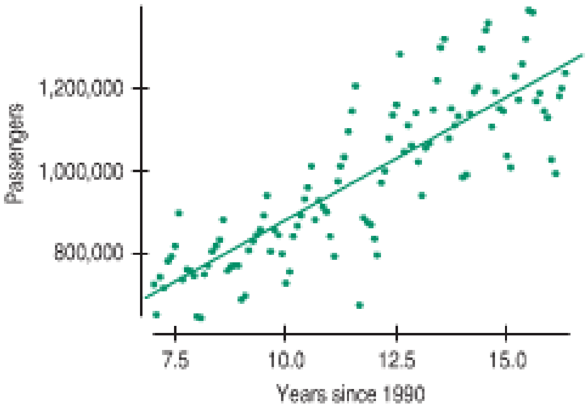
No matter how strong the association, no matter how large the R^2 value, no matter how straight the line, there is no way to conclude from a regression alone that one variable *causes* the other from an observational study.

Exercises Ch 9

2. **Smoking 2004.** The Centers for Disease Control and Prevention track cigarette smoking in the United States. How has the percentage of people who smoke changed since the danger became clear during the last half of the 20th century? The scatterplot shows percentages of smokers among men 18–24 years of age, as estimated by surveys, from 1965 through 2004. (www.cdc.gov/nchs/)



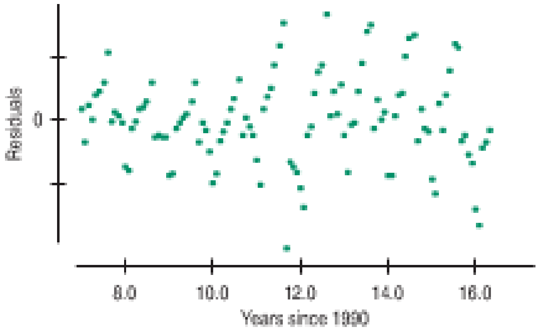
9. Oakland passengers. The scatterplot below shows the number of passengers departing from Oakland (CA) airport month by month since the start of 1997. Time is shown as years since 1990, with fractional years used to represent each month. (Thus, June of 1997 is 7.5—halfway through the 7th year after 1990.) www.oaklandairport.com



Here's a regression and the residuals plotted against *Years since 1990*:

```
Dependent variable is: Passengers
R-squared = 71.1%    s = 104330

Variable         Coefficient
Constant         282584
Year-1990        59704.4
```



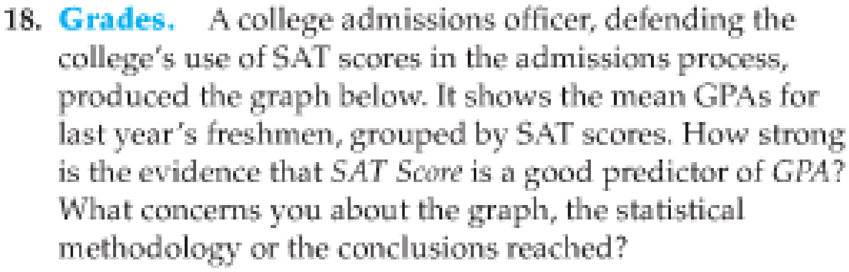
a) Interpret the slope and intercept of the regression model.

b) What does the value of R^2 say about how successful the model is?

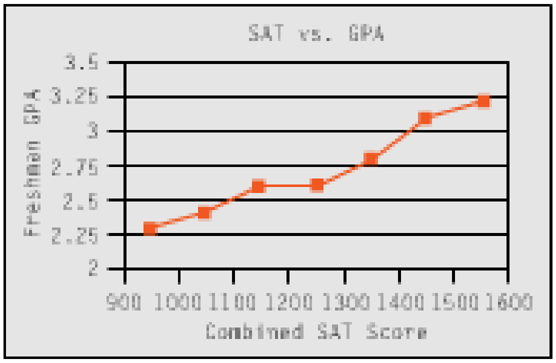
c) Interpret s_e in this context.

d) Would you use this model to predict the numbers of passengers in 2010 (*YearsSince1990* = 20)? Explain.

e) There's a point near the middle of this time span with a large negative residual. Can you explain this outlier?



18. Grades. A college admissions officer, defending the college's use of SAT scores in the admissions process, produced the graph below. It shows the mean GPAs for last year's freshmen, grouped by SAT scores. How strong is the evidence that *SAT Score* is a good predictor of *GPA*? What concerns you about the graph, the statistical methodology or the conclusions reached?



What Can Go Wrong?

- Don't say "correlation" when you mean "association."
  - The word "correlation" should be reserved for measuring the strength and direction of the linear relationship between two quantitative variables.
- Don't confuse "correlation" with "causation."
  - Scatterplots and correlations never demonstrate causation.
- There may be a strong association between two variables that have a nonlinear association

What (else) can go wrong?

- Extrapolation far from the mean can lead to silly and useless predictions.

- An R^2 value near 100% doesn't indicate that there is a causal relationship between x and y.

- Watch out for lurking variables.

- Watch out for regressions based on *summaries* of the data sets.
  These regressions tend to look stronger than the regression on the original data.

What have we learned?

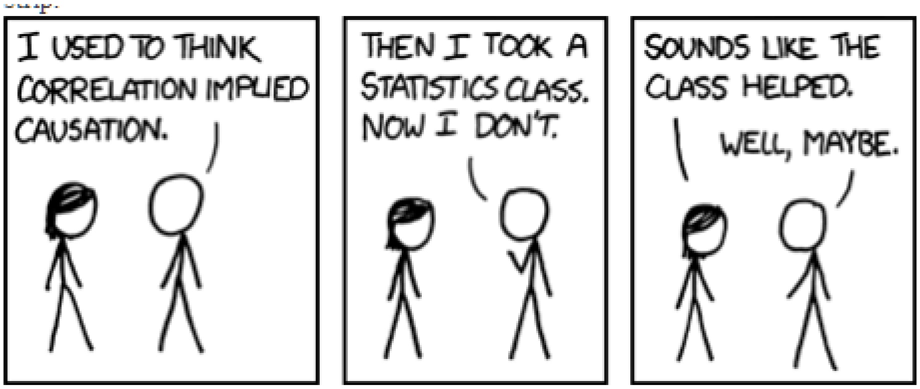
- We examine scatterplots for *direction*, *form*, *strength*, and *unusual features*.

- Although not every relationship is linear, when the scatterplot is straight enough, the *correlation coefficient* is a useful numerical summary.
  - The sign of the correlation tells us the direction of the association.
  - The magnitude of the correlation tells us the *strength* of a linear association.
  - Correlation has no units, so shifting or scaling the data, standardizing, or swapping the variables has no effect on the numerical value.
- The residuals also reveal how well the model works.
  - If a plot of the residuals against predicted values shows a pattern, we should re-examine the data to see why.
  - The standard deviation of the residuals s_e quantifies the amount of scatter around the line. The more scatter – the larger s_e.

Bonus +1: here is what I found in the article "Should SAT be used to predict student success in college?" Source: http://abcnews.go.com/Technology/WhosCounting/story?id=98373&page=1
"Most studies find that the correlation between SAT scores and first-year college grades is not overwhelming, and that only 10 percent to 20 percent of the variation in first-year GPA is explained by SAT scores." Use this statement to find the correlation coefficient discussed it the article.

BONUS +1: Go to http://www.marketskeptics.com/2009/06/amazing-correlation-between-us.html
What is wrong with their interpretation of correlation coefficient?
(you don't have to read any far to get an idea)

Finally…☺


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.