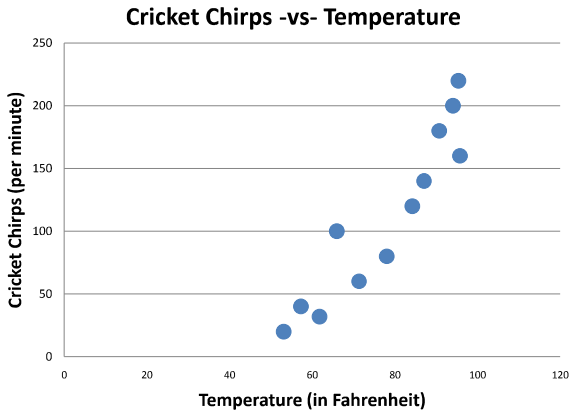
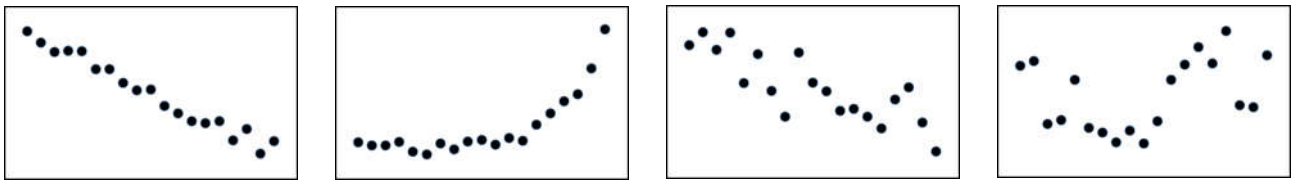
10 Correlation and Regression

10.1 Correlation

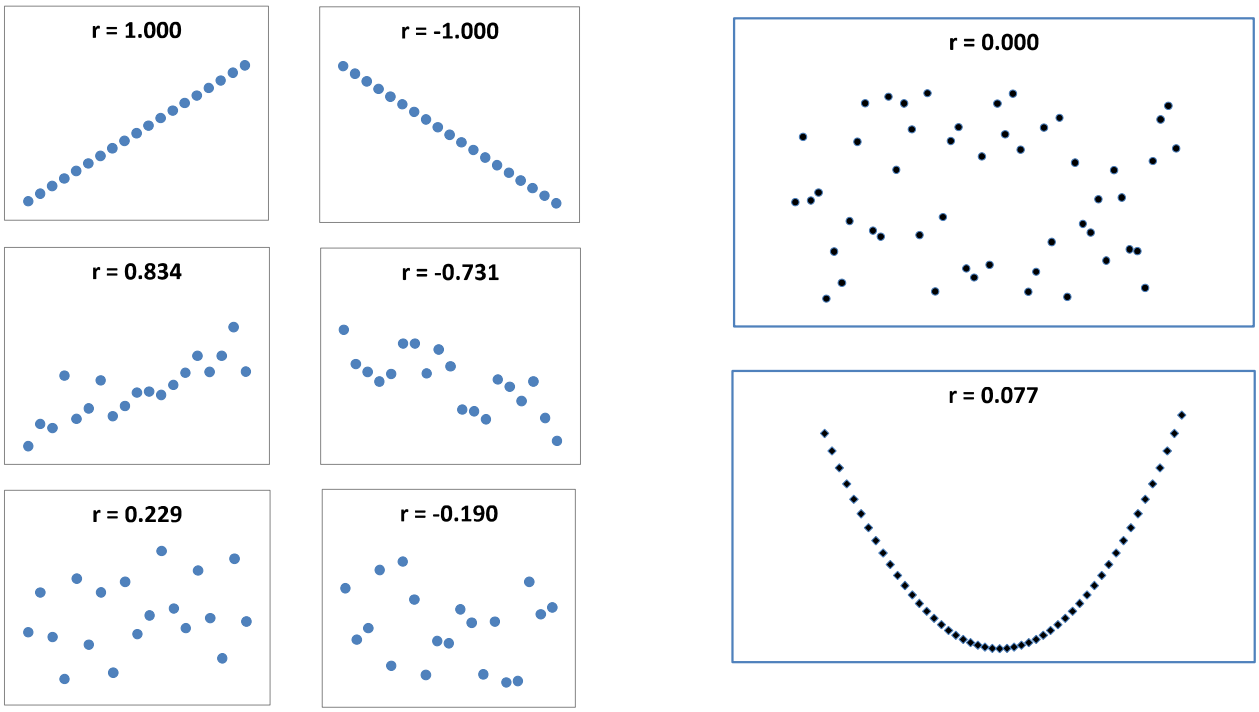
- In Chapter 3 we discussed scatter plots which can be used to help determine if there is an **association** or **correlation** between two variables.



- **Terminology**
  - A **strong relationship** results in a scatter plot that tightly follows a line or curve.
  - A **positive relationship** results in a scatter plot that goes **up** from left to right.
  - A **negative relationship** results in a scatter plot that goes **down** from left to right.
  - Two variables are **linearly related** if the scatter plot reveals a pattern that follows a straight line.
  - An **association** exists between two variables when they are related in some way.
  - A **correlation** exists between two variables when they are linearly related.
- **Your Turn:** Here are some scatter plots. Describe each relationship as a weak/strong, positive/negative, linear/nonlinear, association/correlation.



- The **linear correlation coefficient** r measures the direction and strength of the linear relationship between paired $x-$ and $y-$quantitative values in a sample. It is formally called Pearson's correlation coefficient. It has the following properties:
  - $-1 \leq r \leq 1$.
  - The closer r is to 1, the stronger the positive linear relationship.
  - The closer r is to -1, the stronger the negative linear relationship.
  - The closer R is to 0, the weaker the linear relationship.
  - If $r = 0$ there is no linear relationship.
  - The value of r does not change when variables are converted to a different scale.
  - The value of r is not affected by the choice of x or y.



- **How is r calculated?**

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

Here, $\bar{x}$ is the mean of the x-values, $\bar{y}$ is the mean of the y-values, s_x is the standard deviation of the x-values, s_y is the standard deviation of the y-values, z_x and z_y are the z-scores associated with each x and y respectively, and n is the number of data pairs.

You should avoid calculating r by hand. All statistical software packages have functions for creating scatterplots and calculating r. Examples can be found at the textbook website: www.StevensStats.com

- **When is a correlation significant?**
  - **Using Table 4 (page 292)**
    If the absolute value of the correlation coefficient is larger than the critical values presented in the table for your sample size, then the correlation is significant. If your sample size is not listed, use the closest **smaller** value.
  - **Using Software**
    Software packages will generally give a P-value for the correlation coefficient. The smaller the P-value, the more significant the correlation. Usually, a P-value less than 0.05 is considered significant. Examples can be found at www.StevensStats.com .

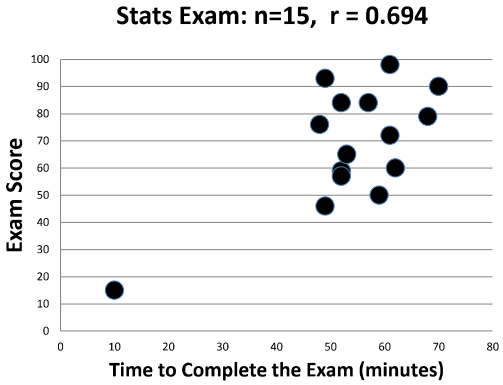
- **Examples:** Use Table 4 to determine if the given correlation is significant.
  1. When correlating car weight and fuel economy the correlation coefficient from a sample of seven cars was r = -0.944.
     According to Table 4, the critical value of r, when $n = 7$, is 0.754. Since the absolute value of correlation coefficient is greater than 0.754, we conclude the correlation is significant.
  2. **Your Turn:** When correlating mother's heights and daughter's heights the correlation coefficient from a sample of 8 pairs resulted in a correlation coefficient of 0.693. What if there had been 20 pairs in the sample?

- **Outliers** can make or break a correlation. If they are known to be in error, they should be eliminated. If not, you should investigate further.

Outlier Makes Correlation:

Here, $n = 15$, and $r = 0.694$
This is a significant correlation
because $0.694 > 0.514$.

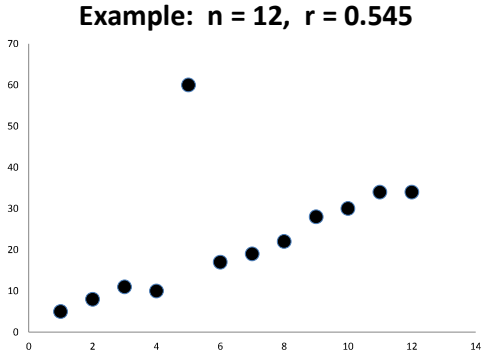
The correlation is due to the outlier.
If this point is removed you can
see there would be no correlation.



Outlier Breaks Correlation:

Here, $n = 12$, and $r = 0.545$
This is not a significant correlation
because $0.545 < 0.576$.

It is wrecked by the outlier.
If this point was removed you can see
a pretty strong correlation.

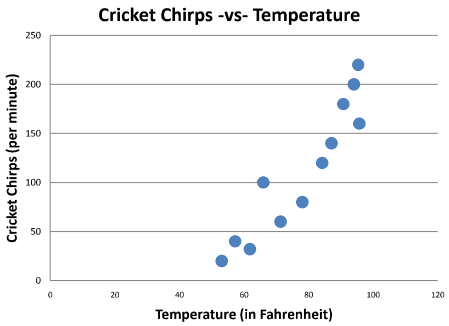


- **Interpreting r - Explained Variation:**

  The value of r^2 represents the proportion of variation in y that is explained by the linear relationship between x and y. *

- **Examples - Explained Variation:**

  1. If the correlation coefficient between temperature and rate of cricket chirps is 0.936, then about 87.6% of the variation in rate of cricket chirps can be explained by the linear relationship to temperature. (because $0.936^2 = .876$). Try switching the implication here.



  2. **Your Turn:** When correlating car weight and fuel economy the correlation coefficient from a sample of seven cars was r = -0.944. Estimate the proportion of a car's fuel economy that can be attributed to the linear relationship to a car's weight?

  3. **Your Turn:** In a sample of eleven US cities, the correlation coefficient between population and murder rate was 0.727. Make a statement about the dependence of the murder rate on population.

- **Issues with Correlation, Causation, and Lurking Variables:**

  - Don't say correlation when you mean association. The word correlation indicates the strength of a linear relationship. The term association is deliberately vague.

  - **Correlation $-vs-$ Causation**.

    Scatter plots and correlations never prove causation by themselves. While some relationships are indeed causal, the nature and direction of the causation may be very hard to establish. Always be on the lookout for lurking variables. A **lurking variable** is one that is not included in the scatterplot but may be causing the two variables to rise or drop together.

    While a correlation doesn't prove causation between the two variables, it does provide evidence that there is some type of causal relationship happening. It might not be the one you initially expect and it might not even exist between the two variables you are studying. However, people will often disregard a correlation that doesn't appeal to them and claim *correlation doesn't prove causation*. This is not good practice either. In Chapter 10.4 there is an introduction on how to control for variables via multiple linear regression which helps clarify *cause and effect*.

*This is actually a strict mathematical statement that is often misused rhetorically. I may be guilty of this from time to time and apologize up-front. It's just so tempting.

- **Examples:**

1. **Smoking and Lung Cancer:** Despite a very strong correlation between smoking and lung cancer it took a long time (about 100 years) to **prove** that smoking was causing lung cancer. The tobacco industry-funded *Tobacco Research Council* found no compelling evidence of a cause and effect relationship and proposed various lurking variables such as air pollution. They even suggested that maybe early-stage cancer was causing a propensity to smoke. Controlled experiments and gene studies finally closed the case against the tobacco industry in the 1980's and 90's.

2. **Global Warming:** In the scientific community, there is a strong consensus that global temperatures are rising and this rise is primarily caused by the increase in greenhouse gases produced by human endeavors. However, disputes still exist in popular media about the cause and effect relationship and its extent. As the *Tobacco Research Council* demonstrated, it is possible to delay a verdict of cause and effect for many years even in the face of over-whelming evidence. Unfortunately, controlled experiments on global warming are impossible because there is only one earth.

3. **Books and Grades:** There is a strong correlation between the number of books in a child's home and his/her performance in elementary school. Does this mean putting books in your house will cause your child to do better in school? What are the lurking variables?

   Answer: Of course not. The lurking variable is probably the home atmosphere and the emphasis/appreciation of reading and education in general.

Your Turn:

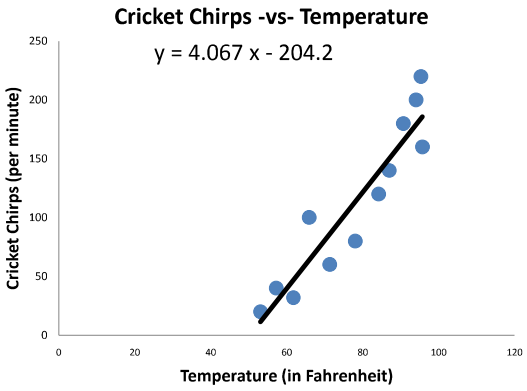
1. There is a significant correlation between the price of rum and the salaries of statistics professors. Does one cause the other? What is the lurking variable?

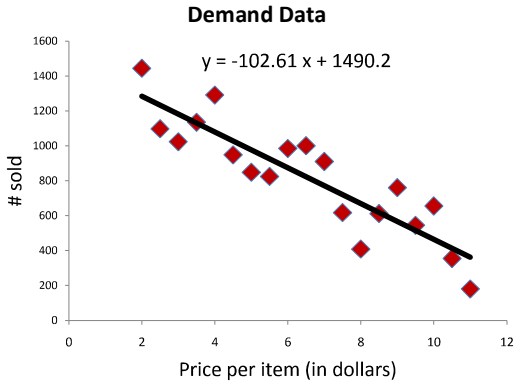
2. There is a strong correlation between ice-cream sales and deaths by drowning. Does ice-cream cause drownings? What is the lurking variable?

3. There is a significant correlation between the rate of cricket chirps and temperature. Does this **prove** that increased temperature causes crickets to chirp faster? Is it possible the faster chirping causes the temperature to increase? Could there be a lurking variable?

4. There is a positive correlation between the amount of rat poison in a house and the number of rats in the area. Does more rat poison increase rat populations?

10.2 Linear Regression





- The **regression equation** $\hat{y} = m\,x + b$ gives the equation of the line that **best fits** the data given in a scatterplot. Here m is the slope of the line, b is the y-intercept, and $\hat{y}$ is the value of y predicted from the regression equation.

Regression Equation:	$\hat{y} = m\,x + b$

- **How do you find this line?**

  The regression line is chosen to minimize the sum of the squares of the residuals. Residual is another name for error in the prediction. For each x value you compare the actual y value and the predicted y-value denoted $\hat{y} = m\,x + b$. The residual is then $y - \hat{y}$. Then the regression equation (line defined by m and b) minimizes the sum of these things squared:

$$\sum (y - \hat{y})^2$$

  The regression equation is often called the least-squares line.

  Fortunately, we don't have to use trial and error or even calculus. We will let software do it[†]
  See www.StevensStats.com .

- **Common Misconception:**

  Many students initially believe that the regression equation should produce the observed value of y for every x from the data set. This is not true. For some values of x, the predicted values of y will be far from those observed in the data.

- **Predictor and Response Variables:**

  While a significant correlation does not guarantee that one variable causes a change in the other. We, as humans, tend to infer such a cause and effect relationship. If we decide to take this risk, there is some conventional structure and terminology involved.

  - Let x (the horizontal axis) be the **predictor** variable.
    This is also called the **explanatory** or **independent** variable.
  - Let y (the vertical axis) be the **response** variable.
    This is also called the **dependent** variable.

[†]Some texts will give formulas for finding m and b. It is tedious work.

- **Your Turn - Discussion:** Suppose you were to collect data for each pair of variables. You want to make a scatter plot. Which variable would you use as the predictor variable, and which as the response variable? Why? What would you expect to see in the scatter plot? Discuss the likely direction, form, and strength.

  1. Test Scores: scores from Test 2 and Scores on the final exam.

  2. Students: Height and weight.

  3. Students: Height in inches and height in centimeters.

  4. Students: Shoe size and GPA.

  5. Gasoline: Number of miles driven and gasoline used.

- **Interpreting the Slope:**

  The slope (m) of the regression equation predicts the change in y given a unit increase in x. In economics, the slope is called **marginal change.**

- **Interpreting the y-intercept:**

  The y-intercept (b) of the regression equation represents the predicted value of y when $x = 0$. Sometimes it has meaning and sometimes it just acts as an upper or lower bound on an expected value.

- **Examples:** Each problem presents a regression equation for the defined variables. Assume the correlation between the variables is significant.

  1. Let x be the cost of a pair of a Nike Air CB34 shoes (in dollars) at the Foot Locker and let y be the number of shoes sold at that price in one week. Regression Equation: $\hat{y} = -2.5\,x + 400$

     (a) If The Foot Locker charges \$150 dollars for a pair of these shoes, how many would they expect to sell?
     Here we put $x = 150$ into the regression equation
     $\tilde{y} = -2.5(150) + 400 = 25$. So they can expect to sell about 25 pairs of CB34's.

     (b) Interpret the slope:
     The slope (-2.5) denotes the expected change in sales when the price is increased by \$1. So, for every increase of \$1 in price they can expect to sell 2.5 fewer pairs of shoes. Remember, this is an approximation so 2.5 pairs of shoes is a valid result.

     (c) Interpret the intercept:
     The intercept (400) represents the number of pairs they can expect to *sell* if the price was set at \$0. While this would never happen, it does provide an upper limit on the number of shoes they could ever expect to sell provided the linear relationship remains valid near $x = 0$.

2. **Your Turn:** Let x be the scores on test 2 and y be the scores on the final exam.
   Regression equation: $\hat{y} = 0.4\ x + 52$
   (a) If you scored a 78 on test 2, what would you expect to score on the final?

   (b) Interpret the slope:

   (c) Interpret the intercept:

3. **Your Turn:** Let x be the birth weight of a male baby (in pounds), and y be their weight at age 20 (in pounds). Regression equation: $\hat{y} = 11.2\ x + 92$.
   (a) If you have a 10 pound baby boy, what is the expected weight at age 20?

   (b) Interpret the slope:

   (c) Interpret the intercept:

- **When to use the regression equation.**
  - The idea of regression is to make predictions about y for a given value of x.
    We call this predicted value $\hat{y}$.
  - You should only use the regression equation to make predictions when the correlation between the two variables is significant.
  - To emphasize this point we will use the following conventions.
    (a) **If the correlation is significant**
    In this case, we want to use the regression equation. Take your value of x and plug it into the regression equation.

    $$\hat{y} = mx + b$$

    (b) **If the correlation is not significant**
    In this case we should not use the regression equation. Instead, use the average of all the y values as the best prediction for any value of x.

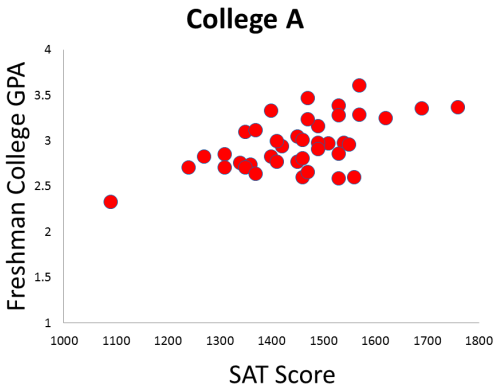
    $$\hat{y} = \overline{y}$$

Examples: Are SAT scores correlated with college GPA's? Here we look at a sample of 65 students who went to one of two colleges. Let x be the student's SAT score and y be the GPA after his/her first year of college. Notice what happens when we combine schools. This is another example of *Simpson's Paradox*.

1. **College A:** Based on the data given, if a student goes to College A with a 1500 SAT score, what is the best prediction for that student's Freshman GPA?

   Sample Size = 40
   $\bar{x} = 1451.3$ and $\bar{y} = 2.964$.
   correlation coefficient $r = 0.584$.
   regression equation: $\hat{y} = 0.0014\,x + 0.98$

   Is there a significant linear correlation?
   Yes, 0.584 is greater then the critical value of 0.312
   If $x = 1500$ the best predicted y-value =
   $\hat{y} = 0.0014(1500) + 0.98 =$ **3.08**

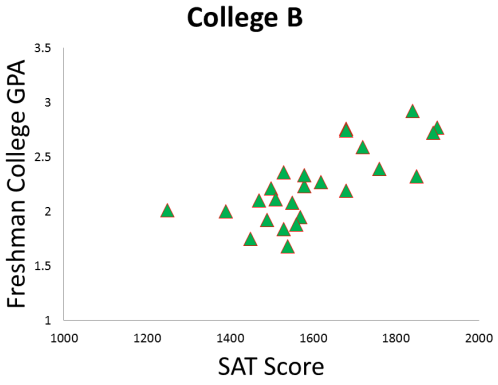


2. **Your Turn - College B:** Based on the data given, if a student goes to College B with a 1500 SAT score, what is the best prediction for that student's Freshman GPA?

   Sample Size = 25
   $\bar{x} = 1604.8$ and $\bar{y} = 2.245$.
   correlation coefficient $r = 0.741$.
   regression equation: $\hat{y} = 0.0016\,x - 0.34$

   Is there a significant linear correlation?

   If $x = 1500$ the best predicted y-value =

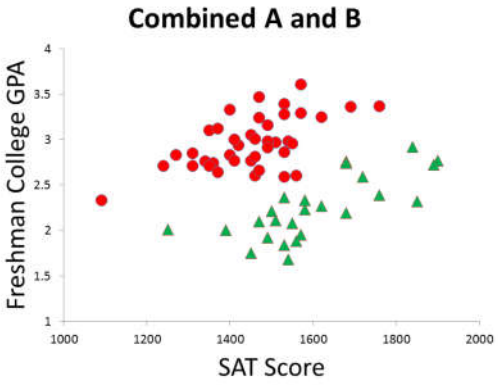


3. **Your Turn - Colleges A and B combined:** Based on the data given, if a student goes to one of the two colleges with a 1500 SAT score, what is the best prediction for that student's Freshman GPA?

   Sample Size = 65
   $\bar{x} = 1510.3$ and $\bar{y} = 2.687$.
   correlation coefficient $r = 0.018$.
   regression equation: $\hat{y} = 0.00005\,x + 2.61$

   Is there a significant linear correlation?

   If $x = 1500$ the best predicted y-value =



10.3 The Hypothesis Test Behind the Scenes

- Table 4 covers more than it seems.

  There is actually a hypothesis test going on here. Suppose ρ (*rho*) is the correlation coefficient between all pairs of variables in a population and r is the correlation coefficient between the variables in a sample. The claim being tested is whether or not ρ is significantly different from zero based on the sample data.

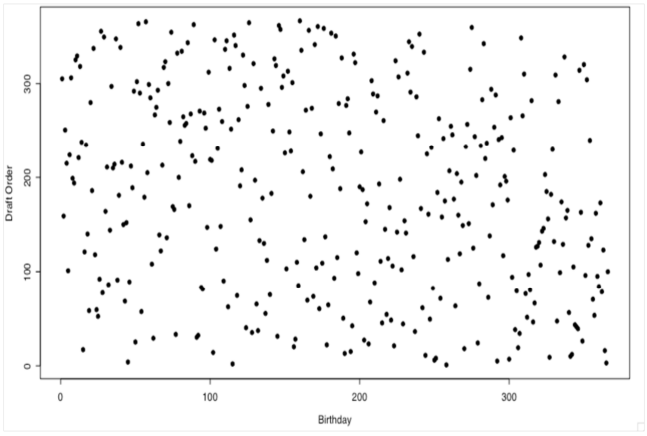
  The null hypothesis is $H_o : \rho = 0$. If we get a sample correlation coefficient (r) whose absolute value is greater than that in Table 4, then we reject the null hypothesis and the data supports the alternate hypothesis that $\rho \neq 0$. In Table 4, we are conducting the test at the 0.05 significance level. As with any hypothesis test there are some assumptions. Here they are:

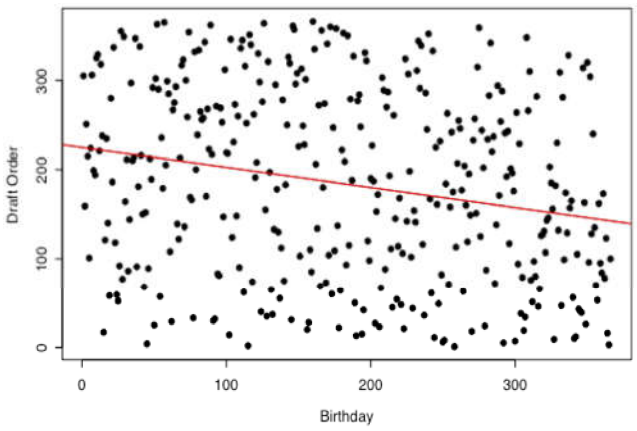
  1. The sample of paired (x,y) data is a random sample of independent quantitative data.
  2. The pairs of (x,y) data must have a **bivariate normal distribution**[‡]

  While the first of these should be met, the second is difficult to check so we just check to make sure that the scatter plot displays a linear pattern and there are no erroneous outliers.

- A significant linear relationship does not necessarily mean the scatter plot displays an obvious correlation.

  When we say there is a significant linear relationship we are really just saying that the linear correlation coefficient of the population is probably not zero. With large sample sizes, there can be a significant linear relationship without any obvious pattern in the scatter plot. This happened during the draft process for the Vietnam war:





  $n = 365$, $r = -.22$, regression line: draft order = 224.9 - 0.226 birthday

  Conclusion: Later birthdays have a lower draft order!

  Was this a result of a biased process or just an act of randomness? Who knows, but people spent a lot of time arguing about it.

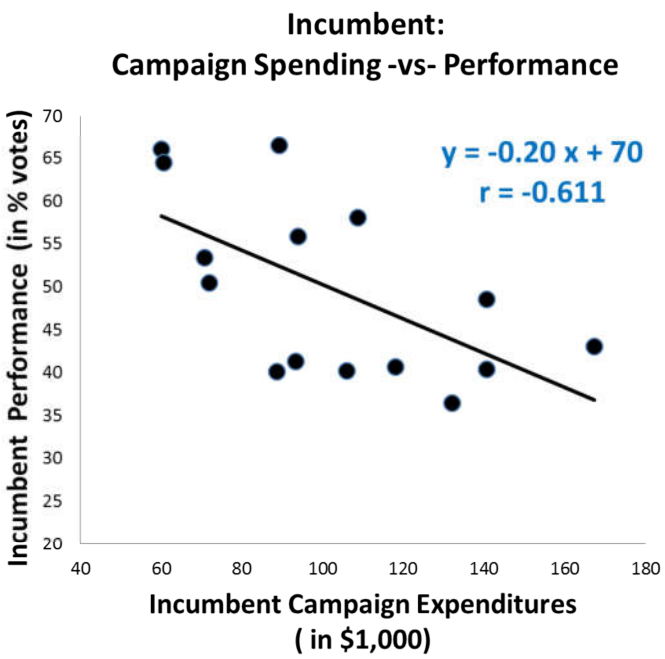
[‡]This means that for any x, associated y values must be normally distributed and vice-versa.

10.4 Multiple Linear Regression: Controlling for Variables - An Introduction

Earlier in this chapter it was noted that a correlation does not prove cause and effect. However, isn't that what we really want to know? Controlling for outside variables is critical when trying to demonstrate cause and effect. I present this little cliff-hanger in hopes that, if you have made it this far, you may be inspired to take another course in statistics in the future.

- **Preliminary Example: Incumbent Campaign Spending**

	Incumbent Campaign Expenditures (in $1,000)	Incumbent Performance (% of votes)
1	70.67	53.44
2	132.00	36.44
3	89.33	66.49
4	88.67	40.08
5	106.00	40.23
6	60.00	66.05
7	108.67	58.05
8	118.00	40.65
9	140.67	40.43
10	140.67	48.51
11	167.33	43.09
12	94.00	55.83
13	72.00	50.45
14	60.67	64.44
15	93.33	41.35

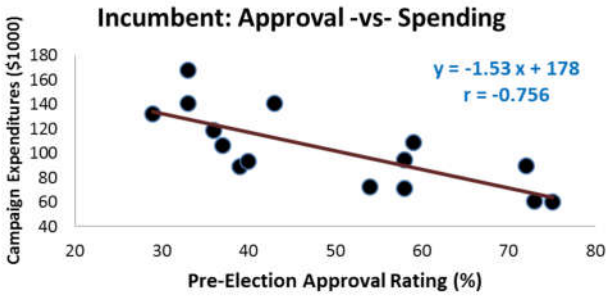


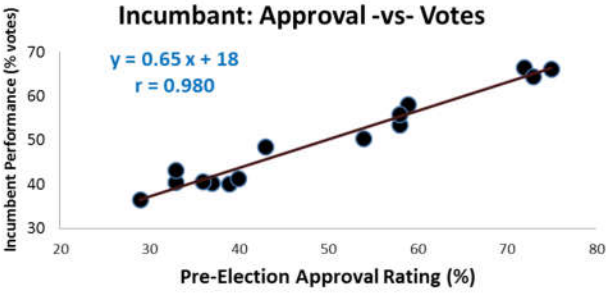
The data presented here is fictitious but demonstrates a well-known relationship. The table gives the campaign expenditures and eventual performance of 15 incumbents in 15 different elections. The scatter plot depicts the negative correlation between these two variables.

- **Observations:**
  - There is a significant negative correlation between incumbent campaign spending and success in the election, $r = -0.611$.
  - The slope of the regression equation (-.20) suggests that that for every extra $1000 spent on campaigning an incumbent can expect to lose about 0.20 percentage points in the election.
  - Is the extra spending **causing** the incumbent to do worse in the election ?
  - If so, this would suggest that incumbents should spend as little as possible for re-election campaigns.
  - Could there be a lurking variable?
  - Can we *control* for that variable?
  - On the next page we *control* for pre-election approval ratings by including it in the model.

- **Controlling for Pre-Election Approval Rating**

	Incumbent Pre-Election Approval (%)	Incumbent Campaign Expenditures (in $1,000)	Incumbent Performance (% of votes)
1	58	70.67	53.44
2	29	132.00	36.44
3	72	89.33	66.49
4	39	88.67	40.08
5	37	106.00	40.23
6	75	60.00	66.05
7	59	108.67	58.05
8	36	118.00	40.65
9	33	140.67	40.43
10	43	140.67	48.51
11	33	167.33	43.09
12	58	94.00	55.83
13	54	72.00	50.45
14	73	60.67	64.44
15	40	93.33	41.35





Multi-Variable Linear Regression Equation (software required)

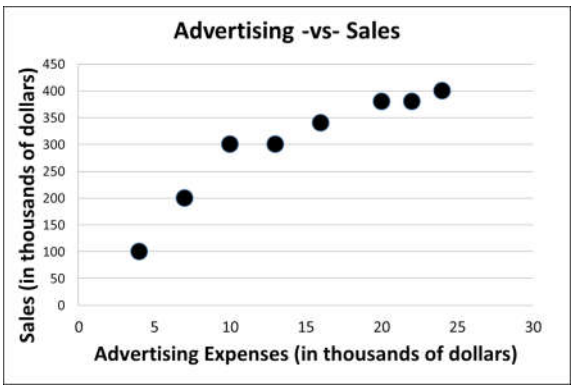
$$\%\ \text{Votes} = 0.80 \cdot (\%\ \text{approval}) + 0.10 \cdot (\text{spending}) + 0.14$$

- **Observations**
  - The first scatter plot tells us there is a significant negative correlation between pre-election approval ratings and campaign expenditures. Does this seem right to you?
  - The second scatter plot tells us there is a significant positive relationship between pre-election approval ratings and election results. Does this seem right to you?
  - We have *controlled* for the pre-election approval rating by making it part of our model.
  - The multi-variable linear regression equation has two slopes.
    - The 0.80 tells us that for every percentage point increase in pre-election approval an incumbent can expect an increase of 0.80 percentage points on election day.
    - The 0.10 tells us that for every $1000 dollars spent on campaigning, an incumbent can expect an increase of 0.10 percentage points on election day.
  - Notice: According to this model, campaign spending helps the incumbent win the election. This is a direct contradiction to our previous conclusion but does make a lot more sense.
  - What is more important; pre-election approval or campaign spending?
  - This type of multi-variable regression allows us to clarify *cause and effect.*
  - If you like this, you might consider taking another course in statistics.

Chapter 10: Summary Worksheet

Alright marketing majors, your job is on the line. Answer the following questions using the given information for monthly sales and monthly advertising expenditures for 8 different months given below.

Advertising (thousands)	Sales (thousands)
22	380
10	300
4	100
13	300
20	380
16	340
7	200
24	400

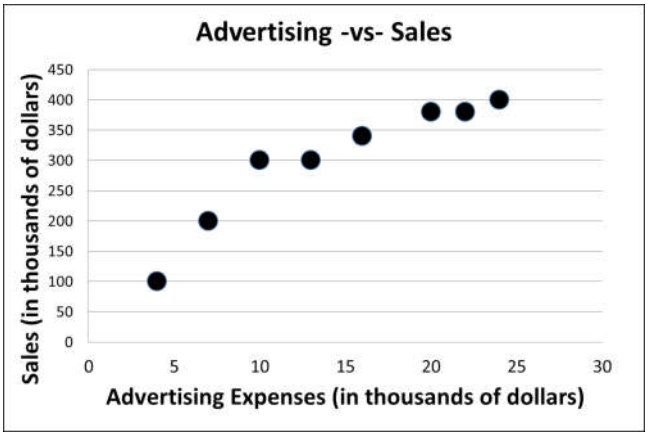


correlation coefficient:
$\boxed{r = 0.94}$

regression equation:
$\boxed{\hat{y} = 13.3\ x + 107}$

1. Is the linear correlation significant? Anything suspicious here?
2. What percentage of the variation in sales can be explained by the linear relation to advertising?
3. Sketch an approximation to the least squares line on the scatterplot.
4. How much in sales would you expect if you spent \$20,000 for advertising?
5. How much in sales would you expect if you spent \$30,000 for advertising?
   Is this a risky prediction? Why or why not?
6. What does the slope of the regression equation represent?
7. What does the y-intercept represent? Is it meaningful?
8. What is the natural choice for the causative variable and the response variable.
9. Can we say that an increase in advertising expenditures causes an increase in sales?
10. You convince the boss to spend extra money on advertising, including a little extra for yourself. On month 9 you spend \$30,000 on advertising and sales are \$400,010.

    (a) Place the new point on the scatterplot.
    (b) Is this data point an *outlier*?
    (c) What happens to r?
    (d) What happens to the regression line?
    (e) What happened?
    (f) What argument can you make to save your job?

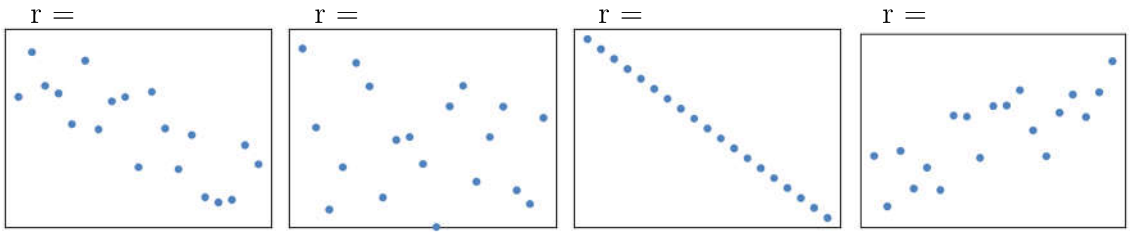


Chapter 10: Problem Set

Numbers with an asterisk* have solutions in the back of the book.

1. Match each scatterplot to one of the correlation coefficients below.

r = 1.00 r = -0.224 r = -1.00 r = 0.763 r = -0.785



2.* **Business/Economics:** Consider the Demand Data demonstrated earlier.§

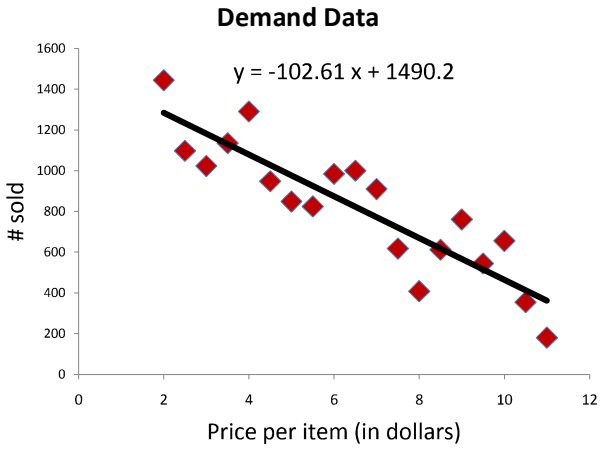
correlation coefficient:
r = -0.89

regression equation:
$\hat{y} = -102.61\ x + 1490.2$

sample size = 19

$\bar{x}$ = \$6.50

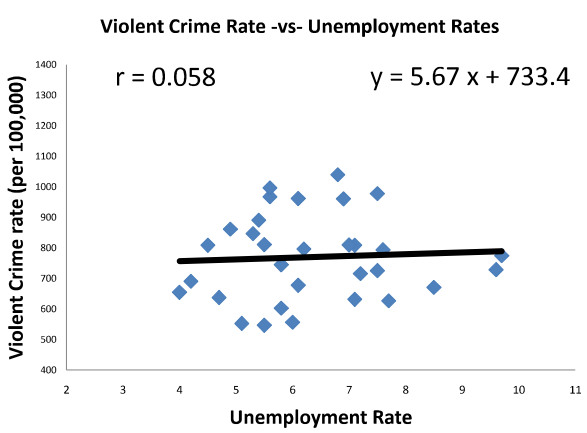
$\bar{y}$ = 823.3



(a) Is there a significant linear correlation between demand (# sold) and price?
(b) What percentage of the variation in demand can be explained by the linear relation to price?
(c) How many items would you expect to sell if the price was set at \$8.00?
(d) How many items would you expect to sell if the price was set at \$15.00?
(e) What does the slope of the regression equation represent?
(f) What does the y-intercept represent? Is it meaningful?
(g) **Marginal demand** is a term in economics that refers to the change in demand for a unit increase in price. What is the marginal demand for this item?
(h) What is the natural choice for the causative variable and the response variable.
(i) According to convention, the demand curve is drawn with price on the vertical (y) axis and quantity sold on the horizontal (x) axis. If we did this, would it change the direction or the strength of the correlation? Would it change the regression equation?
(j) According to this scatter-plot, as demand increases, price goes down. What if there was a fixed supply?

§In Economics texts, demand is usually presented on the x-axis and price on the y-axis.

3. **Sociology/Criminology/Economics:** Records comparing unemployment rates, violent crime rates (per 100,000) and property crime rates (per 100,000) were gathered in the state of Illinois for the years 1975 - 2005 ($n = 31$). The correlation coefficients and regression equations are given in the scatter plots below.



(a) Is there a significant linear correlation between violent crime rates and unemployment rates?

(b) Is there a significant linear correlation between property crime rates and unemployment rates?

(c) With respect to property crime rates, what does the slope of the regression equation represent?

(d) With respect to property crime rates, what does the y-intercept represent? Is it meaningful?

(e) The average unemployment rate in 2008 for the state of Illinois was 6.4%. Use this value and the regression equation to predict the property crime rate of Illinois for 2008.

(f) It turns out that the property crime rate in 2008 for the state of Illinois was 2,932.6 property crimes per 100,000 people. How well does that fit with your prediction from the previous question?

(g) The average unemployment rate in 2009 for the state of Illinois was 10.1 %. Use this value to predict the property crime rate of Illinois for 2009. Is this a risky prediction?

(h) What percentage of the variation in property crime rate can be explained by the linear relation to unemployment?

(i) What is the natural choice for the causative variable and the response variable.

4. **Used Cars - Speculation:** There are 50 used cars at a local dealer. Speculate on the sign and strength of the correlation between the given variables of Mileage (the number of miles it has been driven), Model Year (the year it was made), and Price. Would the correlation be weak, strong, or very strong? Would it be positive or negative?

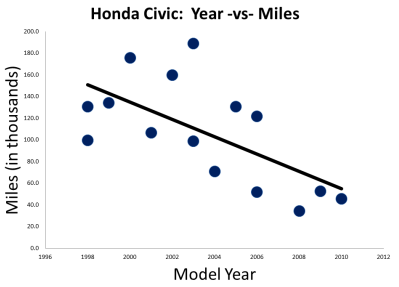
   (a) Model Year & Mileage.

   (b) Mileage & Price.

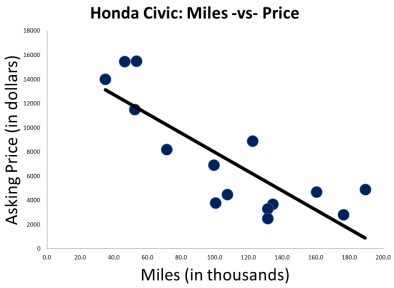
   (c) Model Year & Price.

5.* **Used Cars - Actual Data:**

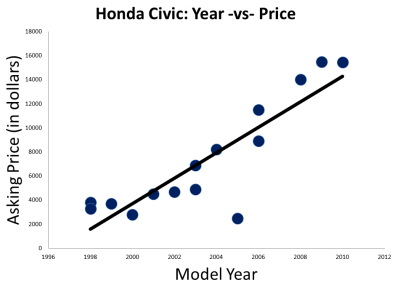
Below are the scatterplots for the data on 15 different Honda Civics found on craigslist in May 2012. These scatterplots should confirm your expectations from the previous problem. The correlation coefficients and regression equations are found below each scatterplot.
Note: The miles are given in thousands.



Model Year (x) -vs- Miles (y)
$r = -0.639$
$\hat{y} = -7.987\ x + 16,109$



Miles (x) -vs- Price (y)
$r = -0.821$
$\hat{y} = -79.2\ x + 15,853$



Model Year (x) -vs- Price (y)
$r = 0.877$
$\hat{y} = 1056.2\ x - 2,108,670$

   (a) Suppose you see a 2002 Honda Civic on craiglist and it has 143 thousand miles on it. Is that more than you would expect on a car from 2002 based on the craigslist data above?

   (b) Suppose you see a Honda Civic with 84 thousand miles on it, but the owner does not give the year it was made. The asking price is $6000. Is this a good price for a Civic with this many miles?

   (c) Suppose you see a 2004 Honda Civic with 140 thousand miles. Estimate a reasonable price for this car via the following methods.

      i. Estimate the price using the model year.

      ii. Estimate the price using the mileage.

      iii. Use the multiple regression equation below where x_1 = model year, x_2 = mileage, and $\hat{y}$ = the expected price.

$$\hat{y} = 716.9\ x_1 - 42.5\ x_2 - 1,424,349$$

   (d) Comment on which of the estimations from part (c) is the best.

6.* **Psychology:** In the 1940's - 60's Sir Cyril Burt conducted famous studies involving identical twins and IQ. Much of his work during this time has been discredited due to accusations of falsifying data. His research became suspect after it was noted that his correlation coefficients remained surprisingly constant despite increased sample sizes. However, much work in this area has continued and the table below gives an approximate summary of the correlation coefficients for IQ scores between groups based on numerous studies of different sizes. These studies tend to support Burt's original conclusions.

Correlation Between	Correlation Coefficient (r)
the same person taking the test twice	0.87
identical twins raised together	0.86
identical twins raised apart	0.76
non-identical twins raised together	0.58

(a) With respect to identical twins raised together, what percentage of one siblings IQ can be attributed to the linear correlation to the other sibling's IQ?

(b) With respect to identical twins raised apart, what percentage of one siblings IQ can be attributed to the linear correlation to the other sibling's IQ?

(c) With respect to non-identical twins raised together, what percentage of one siblings IQ can be attributed to the linear correlation to the other sibling's IQ?

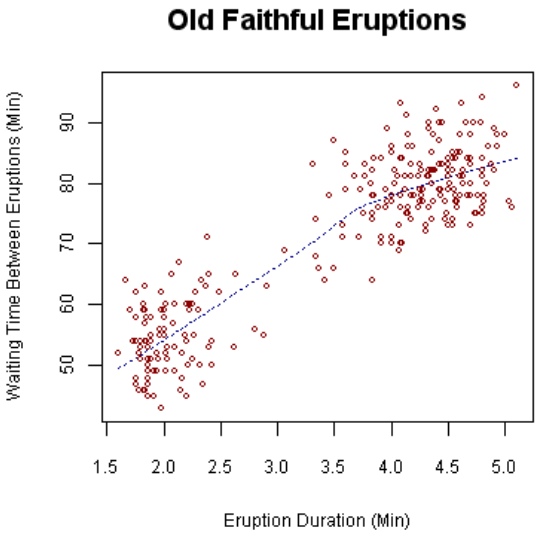
(d) If you were to make generalizations (very risky business) regarding how IQ is inherited, what could you say? This is the age-old argument of nature-vs-nurture in IQ.

(e) What critical piece of information is missing from the data given in the table?

7. **Old Faithful - Clustering Resolved with Piecewise Regression:** At Yellowstone National Park they want to predict the waiting time after one eruption of Old Faithful to the beginning of the next. They have prior data regarding duration of eruptions and the associated waiting time after each eruption.

Here is a scatter plot of eruption duration -vs- waiting time. They have broken it into two different linear regression lines: one for short eruptions and one for long eruptions. The regression equation is given in two parts.

$$\hat{y} = \begin{cases} 13\,x + 28.00 & \text{if} \quad x < 3.75 \\ 8\,x + 46.75 & \text{if} \quad x \geq 3.75 \end{cases}$$



(a) If an eruption lasts 2 minutes, what is the predicted wait time until the next eruption?

(b) If an eruption lasts 4.5 minutes, what is the predicted wait time until the next eruption?

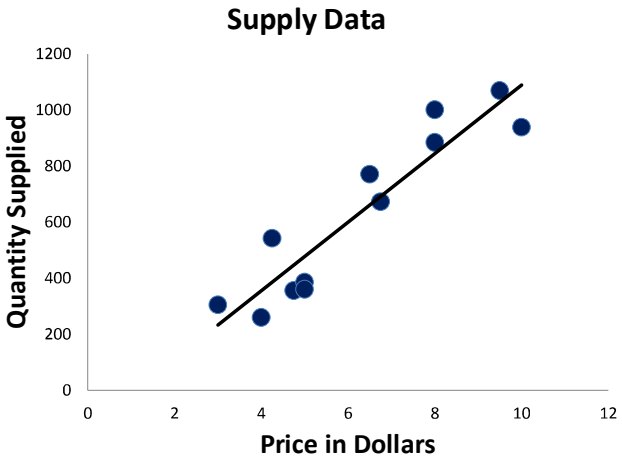
(c) If an eruption lasts 3.75 minutes, what is the predicted wait time until the next eruption?

Software Required.

In these problems, you are given the raw data and asked to answer correlation/regression questions. The data sets have been kept relatively small so that you can enter them by hand without hogging up too much time.

8.* **Law of Supply:** The *Law of Supply* states that an increase in price will result in an increase in the quantity supplied (assuming all other factors remain unchanged). Consider the price and supply data presented below.¶

x=price	y = Quantity Supplied
3.00	304
4.00	259
4.25	542
4.75	355
5.00	385
5.00	360
6.50	770
6.75	672
8.00	884
8.00	1000
9.50	1069
10.00	938



(a) Use software to generate the scatterplot for price vs supply. It should look like the one pictured above.

(b) Use software to calculate the correlation coefficient between price and supply. Is there a significant correlation.

(c) Use software to find the regression equation. What is the slope? What is the y-intercept?

(d) If the price is set at \$5.00, what is the predicted quantity supplied? Round your answer to the nearest whole number.

(e) If the price is set at \$1.00, what is the predicted quantity supplied? Does your answer make sense?

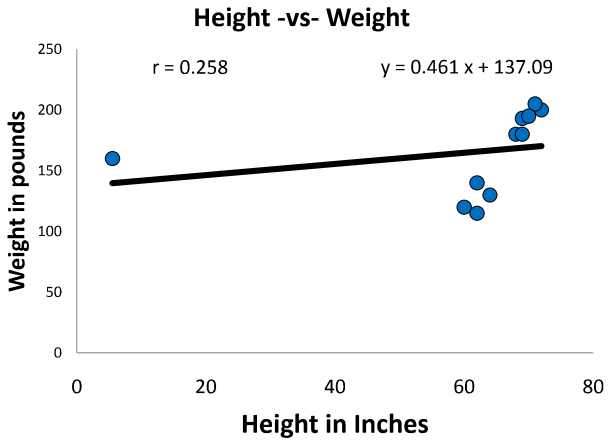
(f) With respect to the variables involved, interpret the slope of the regression equation.

(g) Interpret the y-intercept. Is it meaningful?

¶Again, in Economics texts, supply is usually put on the x-axis and price on the y-axis.

9. **Height vs Weight - Erroneous Data:** As mentioned earlier in this chapter, sometimes an outlier can make or break a correlation. Data from 11 people regarding height and weight is given in the table below and the associated scatter plot is given with the correlation coefficient and regression equation in the graph.

x=height (inches)	y = weight (pounds)
60	120
72	200
64	130
71	205
68	180
69	180
69	193
70	195
62	115
62	140
5.5	**160**



(a) According to this data, is there a significant correlation between height and weight?

(b) As can be seen from the scatter plot, there seems to be something odd going on here. What is it and how should you remedy the situation?

(c) If you exclude the last data point corresponding to a height of 5.5 inches, create the scatter plot that results. Include the regression line and equation in your plot.

(d) After excluding this last data point, is there a significant linear correlation between height and weight?

(e) What did this one data point do to the correlation?

(f) Using the new regression equation with the last data point excluded, what is the expected weight of a person who is 62 inches tall.

10.* **Test Time vs Score - Outlier:** Below is the time it took each of 15 students to complete a Stats test and the score that each student got on the exam. Notice, the last data pair is somewhat unusual.

x = Test Time (min)	59	49	61	52	61	52	48	53	68	57	49	70	62	52	**10**
y = Score (out of 100)	50	93	72	59	98	84	76	65	79	84	46	90	60	57	**15**

(a) Using all 15 data pairs, use software to create the scatterplot and least squares line. Also, calculate the correlation coefficient and regression equation. Is there a significant correlation?

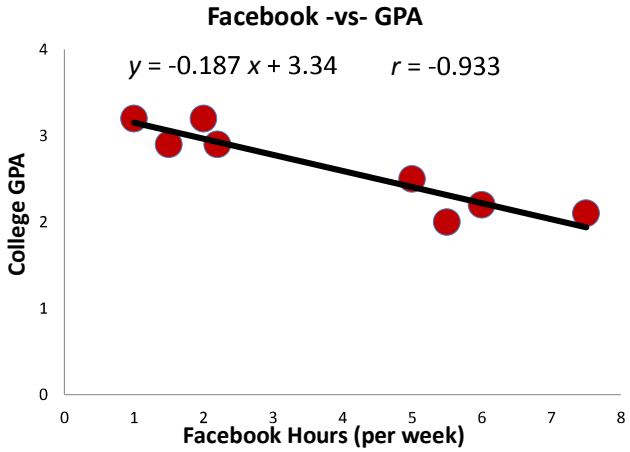
(b) Using only the first 14 data pairs, use software to create the scatterplot and least squares line. Also, calculate the correlation coefficient and regression equation.

(c) What did the outlier do to the correlation coefficient and regression equation?

(d) What false conclusion might be drawn from the original analysis with all 15 data pairs?

11. **Facebook vs GPA - Clustering:** (This data is based on a real report but is not the actual data). Eight college students are surveyed for the number of hours per week they spend on Facebook. This number is paired with the students GPA. The data is in the table below and presented in the scatter plot with the correlation coefficient, regression equation, and regression line.

x=facebook hours per week	y = GPA
1.0	3.2
1.5	2.9
2.0	3.2
2.2	2.9
5.0	2.5
5.5	2.0
6.0	2.2
7.5	2.1



(a) Based on the information given, is there a linear correlation between weekly hours spent on Facebook and GPA?

(b) Comment on the clustering of data here.

(c) Find the correlation coefficient for the 4 data points in the higher GPA cluster (these are the first four in the data set). Is there a significant linear correlation for these four?

(d) Find the correlation coefficient for the 4 data points in the lower GPA cluster (these are the last four in the data set). Is there a significant linear correlation for these four?

(e) Tell the story here. What may be the *lurking* variables in this correlation?

12.* **Facebook Friends - Nonlinear:** There is some discussion as to whether *virtual* newtworking and friendships has a positive or negative effect on personal networks and friendships. In a survey of 10 people over the age of 30, each person was requested to report on the number of facebook friends they have and the number of real-world personal friends they have. The table below gives these reported values.

x = Facebook Friends	22	232	78	168	122	153	97	195	230	51
y = Real Friends	38	36	27	22	20	17	22	24	33	28

(a) Create a scatter-plot for this data. Does there appear to be an association between the two variables?

(b) What is the linear correlation coefficient? Is there a significant linear correlation?

(c) Why is there no point in determining the regression equation?

13. **IQ vs Shoe Size - Outlier:** As mentioned earlier in this chapter, outliers can make a correlation when one doesn't really exist. Here is the example of clown's shoe sizes versus IQ. Below is a list of the data from 10 different clowns. The last entry is for Bozo the clown who was unusually intelligent and wore very large shoes. ‖

x = Shoe Size	9	8	10	11	7.5	9.5	9	10.5	10	**18**
y = IQ	95	85	110	85	115	95	107	102	90	**155**

(a) Create a scatter plot from this data. Do one scatter-plot of all the data and one scatter plot which excludes Bozo. Include the regression line and the regression equation on the graphs.

(b) What happens to the regression line when Bozo is excluded?

(c) What is the correlation coefficient including Bozo? Is there a significant linear correlation?

(d) What is the correlation coefficient if you exclude Bozo? Is there a significant linear correlation?

(e) Summarize what Bozo did to the correlation between shoe size and IQ.

‖Courtesy of Dick DeVeaux, Paul Velleman, & David Bock, Intro Stats (3rd Ed), Pearson Addison Wesley, 2011