Chapter 10: Regression and Correlation

The previous chapter looked at comparing populations to see if there is a difference between the two. That involved two random variables that are similar measures. This chapter will look at two random variables that are not similar measures, and see if there is a relationship between the two variables. To do this, you look at regression, which finds the linear relationship, and correlation, which measures the strength of a linear relationship.

Please note: there are many other types of relationships besides linear that can be found for the data. This book will only explore linear, but realize that there are other relationships that can be used to describe data.

Section 10.1: Regression

When comparing two different variables, two questions come to mind: "Is there a relationship between two variables?" and "How strong is that relationship?" These questions can be answered using **regression** and **correlation**. Regression answers whether there is a relationship (again this book will explore linear only) and correlation answers how strong the linear relationship is. To introduce both of these concepts, it is easier to look at a set of data.

Example #10.1.1: Determining If There Is a Relationship

Is there a relationship between the alcohol content and the number of calories in 12-ounce beer? To determine if there is one a random sample was taken of beer's alcohol content and calories ("Calories in beer,," 2011), and the data is in table #10.1.1.

Table #10.1.1: Alcohol and Calorie Content in Beer

Brand	Brewery	Alcohol Content	Calories in 12 oz
Big Sky Scape Goat Pale Ale	Big Sky Brewing	4.70%	163
Sierra Nevada Harvest Ale	Sierra Nevada	6.70%	215
Steel Reserve	MillerCoors	8.10%	222
O'Doul's	Anheuser Busch	0.40%	70
Coors Light	MillerCoors	4.15%	104
Genesee Cream Ale	High Falls Brewing	5.10%	162
Sierra Nevada Summerfest Beer	Sierra Nevada	5.00%	158
Michelob Beer	Anheuser Busch	5.00%	155
Flying Dog Doggie Style	Flying Dog Brewery	4.70%	158
Big Sky I.P.A.	Big Sky Brewing	6.20%	195

Solution:

To aid in figuring out if there is a relationship, it helps to draw a scatter plot of the data. It is helpful to state the random variables, and since in an algebra class the

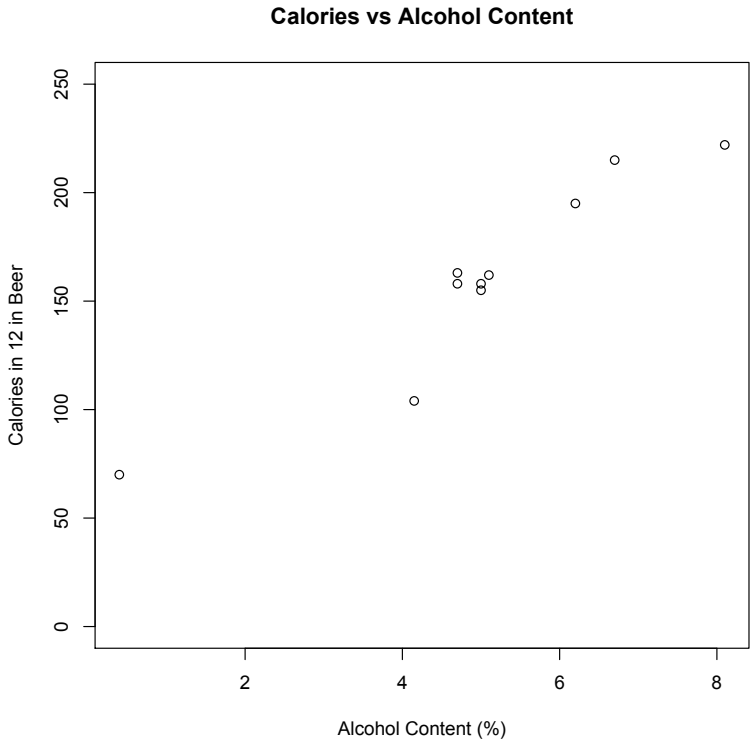
variables are represented as x and y, those labels will be used here. It helps to state which variable is x and which is y.

State random variables

x = alcohol content in the beer

y = calories in 12 ounce beer

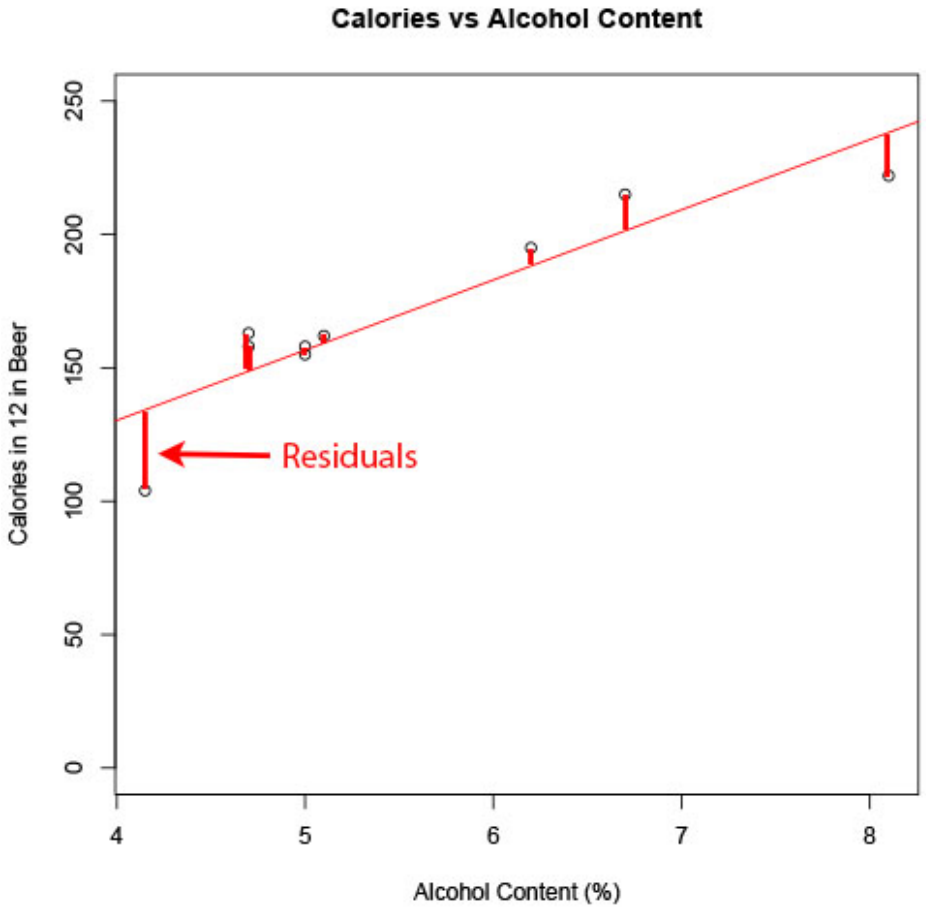
Figure #10.1.1: Scatter Plot of Beer Data



This scatter plot looks fairly linear. However, notice that there is one beer in the list that is actually considered a non-alcoholic beer. That value is probably an outlier since it is a non-alcoholic beer. The rest of the analysis will not include O'Doul's. You cannot just remove data points, but in this case it makes more sense to, since all the other beers have a fairly large alcohol content.

To find the equation for the linear relationship, the process of regression is used to find the line that best fits the data (sometimes called the best fitting line). The process is to draw the line through the data and then find the distances from a point to the line, which are called the residuals. The regression line is the line that makes the square of the residuals as small as possible, so the regression line is also sometimes called the least squares line. The regression line and the residuals are displayed in figure #10.1.2.

Figure #10.1.2: Scatter Plot of Beer Data with Regression Line and Residuals



The find the regression equation (also known as best fitting line or least squares line)

Given a collection of paired sample data, the regression equation is

$$\hat{y} = a + bx$$

where the slope = $b = \dfrac{SS_{xy}}{SS_x}$ and y-intercept = $a = \overline{y} - b\overline{x}$

The **residuals** are the difference between the actual values and the estimated values.

$$\text{residual} = y - \hat{y}$$

SS stands for sum of squares. So you are summing up squares. With the subscript *xy*, you aren't really summing squares, but you can think of it that way in a weird sense.

$$SS_{xy} = \sum (x - \overline{x})(y - \overline{y})$$

$$SS_x = \sum (x - \overline{x})^2$$

$$SS_y = \sum (y - \overline{y})^2$$

Note: the easiest way to find the regression equation is to use the technology.

The **independent variable**, also called the **explanatory variable** or **predictor variable**, is the *x*-value in the equation. The independent variable is the one that you use to predict what the other variable is. The **dependent variable** depends on what independent value you pick. It also responds to the explanatory variable and is sometimes called the **response variable**. In the alcohol content and calorie example, it makes slightly more sense to say that you would use the alcohol content on a beer to predict the number of calories in the beer.

The **population equation** looks like:

$y = \beta_o + \beta_1 x$

β_o = slope

β_1 = y-intercept

$\hat{y}$ is used to predict *y*.

Assumptions of the regression line:

a. The set (x, y) of ordered pairs is a random sample from the population of all such possible (x, y) pairs.

b. For each fixed value of *x*, the *y*-values have a normal distribution. All of the *y* distributions have the same variance, and for a given *x*-value, the distribution of *y*-values has a mean that lies on the least squares line. You also assume that for a fixed *y*, each *x* has its own normal distribution. This is difficult to figure out, so you can use the following to determine if you have a normal distribution.
   i. Look to see if the scatter plot has a linear pattern.
   ii. Examine the residuals to see if there is randomness in the residuals. If there is a pattern to the residuals, then there is an issue in the data.

Example #10.1.2: Find the Equation of the Regression Line

a.) Is there a positive relationship between the alcohol content and the number of calories in 12-ounce beer? To determine if there is a positive linear relationship, a random sample was taken of beer's alcohol content and calories for several different beers ("Calories in beer,," 2011), and the data are in table #10.1.2.

Table #10.1.2: Alcohol and Calorie Content in Beer without Outlier

Brand	Brewery	Alcohol Content	Calories in 12 oz
Big Sky Scape Goat Pale Ale	Big Sky Brewing	4.70%	163
Sierra Nevada Harvest Ale	Sierra Nevada	6.70%	215
Steel Reserve	MillerCoors	8.10%	222
Coors Light	MillerCoors	4.15%	104
Genesee Cream Ale	High Falls Brewing	5.10%	162
Sierra Nevada Summerfest Beer	Sierra Nevada	5.00%	158
Michelob Beer	Anheuser Busch	5.00%	155
Flying Dog Doggie Style	Flying Dog Brewery	4.70%	158
Big Sky I.P.A.	Big Sky Brewing	6.20%	195

Solution:

State random variables

x = alcohol content in the beer

y = calories in 12 ounce beer

Assumptions check:

a. A random sample was taken as stated in the problem.

b. The distribution for each calorie value is normally distributed for every value of alcohol content in the beer.

   i. From Example #10.2.1, the scatter plot looks fairly linear.

   ii. The residual versus the x-values plot looks fairly random. (See figure #10.1.5.)

   It appears that the distribution for calories is a normal distribution.

To find the regression equation on the TI-83/84 calculator, put the x's in L1 and the y's in L2. Then go to STAT, over to TESTS, and choose LinRegTTest. The setup is in figure #10.1.3. The reason that >0 was chosen is because the question was asked if there was a positive relationship. If you are asked if there is a negative relationship, then pick <0. If you are just asked if there is a relationship, then pick $\neq 0$. Right now the choice will not make a different, but it will be important later.

Figure #10.1.3: Setup for Linear Regression Test on TI-83/84

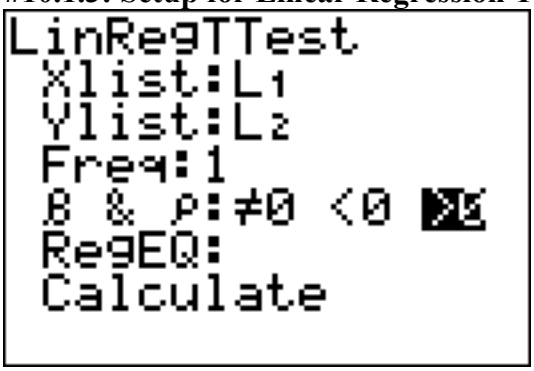
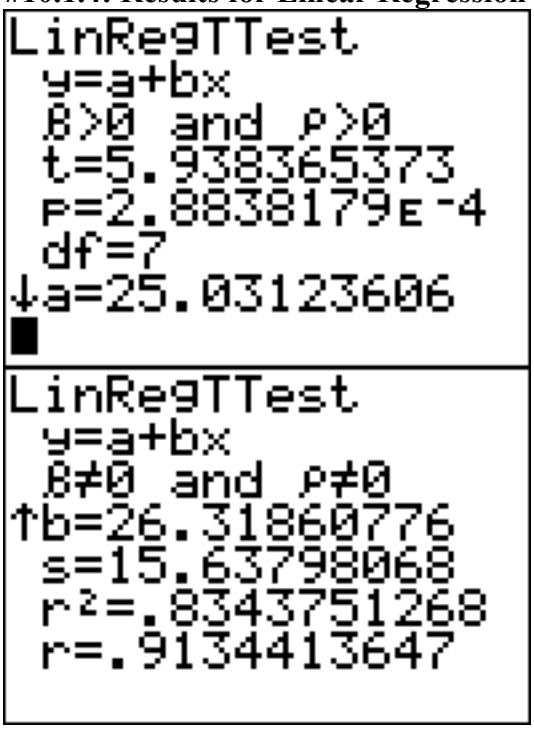


Figure #10.1.4: Results for Linear Regression Test on TI-83/84



From this you can see that

$$\hat{y} = 25.0 + 26.3x$$

To find the regression equation using R, the command is lm(dependent variable ~ independent variable), where ~ is the tilde symbol located on the upper left of most keyboards. So for this example, the command would be lm(calories ~ alcohol), and the output would be

Call:
lm(formula = calories ~ alcohol)

Coefficients:

(Intercept)	alcohol
25.03	26.32

From this you can see that the y-intercept is 25.03 and the slope is 26.32. So the regression equation is $\hat{y} = 25.03 + 26.32x$.

Remember, this is an estimate for the true regression. A different random sample would produce a different estimate.

b.) Use the regression equation to find the number of calories when the alcohol content is 6.50%.

Solution:

$x_o = 6.50$

$\hat{y} = 25.0 + 26.3(6.50) = 196$ calories

If you are drinking a beer that is 6.50% alcohol content, then it is probably close to 196 calories. Notice, the mean number of calories is 170 calories. This value of 196 seems like a better estimate than the mean when looking at the original data. The regression equation is a better estimate than just the mean.

c.) Use the regression equation to find the number of calories when the alcohol content is 2.00%.

Solution:

$x_o = 2.00$

$\hat{y} = 25.0 + 26.3(2.00) = 78$ calories

If you are drinking a beer that is 2.00% alcohol content, then it has probably close to 78 calories. This doesn't seem like a very good estimate. This estimate is what is called extrapolation. It is not a good idea to predict values that are far outside the range of the original data. This is because you can never be sure that the regression equation is valid for data outside the original data.

d.) Find the residuals and then plot the residuals versus the *x*-values.

Solution:

To find the residuals, find $\hat{y}$ for each *x*-value. Then subtract each $\hat{y}$ from the given *y* value to find the residuals. Realize that these are sample residuals since they are calculated from sample values. It is best to do this in a spreadsheet.

Table #10.1.3: Residuals for Beer Calories

x	y	$\hat{y} = 25.0 + 26.3x$	$y - \hat{y}$
4.70	163	148.61	14.390
6.70	215	201.21	13.790
8.10	222	238.03	-16.030
4.15	104	134.145	-30.145
5.10	162	159.13	2.870
5.00	158	156.5	1.500
5.00	155	156.5	-1.500
4.70	158	148.61	9.390
6.20	195	188.06	6.940

Notice the residuals add up to close to 0. They don't add up to exactly 0 in this example because of rounding error. Normally the residuals add up to 0.

You can use R to get the residuals. The command is
lm.out = lm(dependent variable ~ independent variable) – this defines the linear model with a name so you can use it later. Then
residual(lm.out) – produces the residuals.

For this example, the command would be

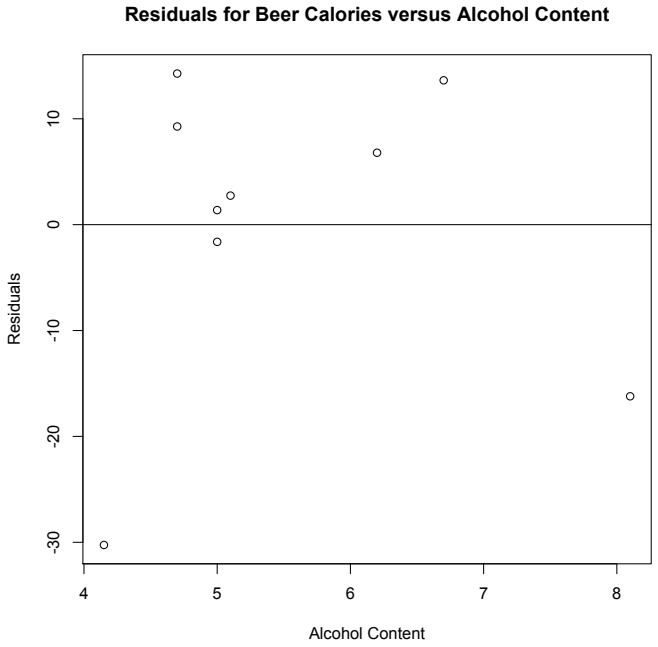
```
lm(calories~alcohol)

Call:
lm(formula = calories ~ alcohol)

Coefficients:
(Intercept)      alcohol
     25.03        26.32

> residuals(lm.out)
       1           2           3           4           5           6
14.271307  13.634092  -16.211959  -30.253458   2.743864   1.375725  -
       7           8           9
 1.624275   9.271307   6.793396
```

So the first residual is 14.271307 and it belongs to the first x value. The residual 13.634092 belongs to the second x value, and so forth.

You can then graph the residuals versus the independent variable using the plot command. For this example, the command would be plot(alcohol, residuals(lm.out), main="Residuals for Beer Calories versus Alcohol Content", xlab="Alcohol Content", ylab="Residuals"). Sometimes it is useful to see the x-axis on the graph, so after creating the plot, type the command abline(0,0).

The graph of the residuals versus the *x*-values is in figure #10.1.5. They appear to be somewhat random.

Figure #10.1.5: Residuals of Beer Calories versus Content



Notice, that the 6.50% value falls into the range of the original x-values. The processes of predicting values using an x within the range of original x-values is called **interpolating**. The 2.00% value is outside the range of original x-values. Using an x-value that is outside the range of the original x-values is called **extrapolating**. When predicting values using interpolation, you can usually feel pretty confident that that value will be close to the true value. When you extrapolate, you are not really sure that the predicted value is close to the true value. This is because when you interpolate, you know the equation that predicts, but when you extrapolate, you are not really sure that your relationship is still valid. The relationship could in fact change for different x-values.

An example of this is when you use regression to come up with an equation to predict the growth of a city, like Flagstaff, AZ. Based on analysis it was determined that the population of Flagstaff would be well over 50,000 by 1995. However, when a census was undertaken in 1995, the population was less than 50,000. This is because they extrapolated and the growth factor they were using had obviously changed from the early 1990's. Growth factors can change for many reasons, such as employment growth, employment stagnation, disease, articles saying great place to live, etc. Realize that when you extrapolate, your predicted value may not be anywhere close to the actual value that you observe.

What does the slope mean in the context of this problem?

$$m = \frac{\Delta y}{\Delta x} = \frac{\Delta \text{ calories}}{\Delta \text{ alcohol content}} = \frac{26.3 \text{ calories}}{1\%}$$

The calories increase 26.3 calories for every 1% increase in alcohol content.

The y-intercept in many cases is meaningless. In this case, it means that if a drink has 0 alcohol content, then it would have 25.0 calories. This may be reasonable, but remember this value is an extrapolation so it may be wrong.

Consider the residuals again. According to the data, a beer with 6.7% alcohol has 215 calories. The predicted value is 201 calories.

$$\begin{aligned}\text{Residual} &= \text{actual} - \text{predicted} \\ &= 215 - 201 \\ &= 14\end{aligned}$$

This deviation means that the actual value was 14 above the predicted value. That isn't that far off. Some of the actual values differ by a large amount from the predicted value. This is due to variability in the dependent variable. The larger the residuals the less the model explains the variability in the dependent variable. There needs to be a way to calculate how well the model explains the variability in the dependent variable. This will be explored in the next section.

The following example demonstrates the process to go through when using the formulas for finding the regression equation, though it is better to use technology. This is because if the linear model doesn't fit the data well, then you could try some of the other models that are available through technology.

Example #10.1.3: Calculating the Regression Equation with the Formula

Is there a relationship between the alcohol content and the number of calories in 12-ounce beer? To determine if there is one a random sample was taken of beer's alcohol content and calories ("Calories in beer,," 2011), and the data are in table #10.1.2. Find the regression equation from the formula.

Solution:

State random variables

x = alcohol content in the beer

y = calories in 12 ounce beer

Table #10.1.4: Calculations for Regression Equation

Alcohol Content	Calories	$x-\overline{x}$	$y-\overline{y}$	$(x-\overline{x})^2$	$(y-\overline{y})^2$	$(x-\overline{x})(y-\overline{y})$
4.70	163	-0.8167	-7.2222	0.6669	52.1605	5.8981
6.70	215	1.1833	44.7778	1.4003	2005.0494	52.9870
8.10	222	2.5833	51.7778	6.6736	2680.9383	133.7593
4.15	104	-1.3667	-66.2222	1.8678	4385.3827	90.5037
5.10	162	-0.4167	-8.2222	0.1736	67.6049	3.4259
5.00	158	-0.5167	-12.2222	0.2669	149.3827	6.3148
5.00	155	-0.5167	-15.2222	0.2669	231.7160	7.8648
4.70	158	-0.8167	-12.2222	0.6669	149.3827	9.9815
6.20	195	0.6833	24.7778	0.4669	613.9383	16.9315
5.516667 $=\overline{x}$	170.2222 $=\overline{y}$			12.45 $=SS_x$	10335.5556 $=SS_y$	327.6667 $=SS_{xy}$

slope: $b=\dfrac{SS_{xy}}{SS_x}=\dfrac{327.6667}{12.45}\approx 26.3$

y-intercept: $a=\overline{y}-b\overline{x}=170.222-26.3(5.516667)\approx 25.0$

Regression equation: $\hat{y}=25.0+26.3x$

Section 10.1: Homework

For each problem, state the random variables. Also, look to see if there are any outliers that need to be removed. Do the regression analysis with and without the suspected outlier points to determine if their removal affects the regression. The data sets in this section are used in the homework for sections 10.2 and 10.3 also.

1.) When an anthropologist finds skeletal remains, they need to figure out the height of the person. The height of a person (in cm) and the length of their metacarpal bone 1 (in cm) were collected and are in table #10.1.5 ("Prediction of height," 2013). Create a scatter plot and find a regression equation between the height of a person and the length of their metacarpal. Then use the regression equation to find the height of a person for a metacarpal length of 44 cm and for a metacarpal length of 55 cm. Which height that you calculated do you think is closer to the true height of the person? Why?

Table #10.1.5: Data of Metacarpal versus Height

Length of Metacarpal (cm)	Height of Person (cm)
45	171
51	178
39	157
41	163
48	172
49	183
46	173
43	175
47	173

2.) Table #10.1.6 contains the value of the house and the amount of rental income in a year that the house brings in ("Capital and rental," 2013). Create a scatter plot and find a regression equation between house value and rental income. Then use the regression equation to find the rental income a house worth $230,000 and for a house worth $400,000. Which rental income that you calculated do you think is closer to the true rental income? Why?

Table #10.1.6: Data of House Value versus Rental

Value	Rental	Value	Rental	Value	Rental	Value	Rental
81000	6656	77000	4576	75000	7280	67500	6864
95000	7904	94000	8736	90000	6240	85000	7072
121000	12064	115000	7904	110000	7072	104000	7904
135000	8320	130000	9776	126000	6240	125000	7904
145000	8320	140000	9568	140000	9152	135000	7488
165000	13312	165000	8528	155000	7488	148000	8320
178000	11856	174000	10400	170000	9568	170000	12688
200000	12272	200000	10608	194000	11232	190000	8320
214000	8528	208000	10400	200000	10400	200000	8320
240000	10192	240000	12064	240000	11648	225000	12480
289000	11648	270000	12896	262000	10192	244500	11232
325000	12480	310000	12480	303000	12272	300000	12480

3.) The World Bank collects information on the life expectancy of a person in each country ("Life expectancy at," 2013) and the fertility rate per woman in the country ("Fertility rate," 2013). The data for 24 randomly selected countries for the year 2011 are in table #10.1.7. Create a scatter plot of the data and find a linear regression equation between fertility rate and life expectancy. Then use the regression equation to find the life expectancy for a country that has a fertility rate of 2.7 and for a country with fertility rate of 8.1. Which life expectancy that you calculated do you think is closer to the true life expectancy? Why?

Table #10.1.7: Data of Fertility Rates versus Life Expectancy

Fertility Rate	Life Expectancy
1.7	77.2
5.8	55.4
2.2	69.9
2.1	76.4
1.8	75.0
2.0	78.2
2.6	73.0
2.8	70.8
1.4	82.6
2.6	68.9
1.5	81.0
6.9	54.2
2.4	67.1
1.5	73.3
2.5	74.2
1.4	80.7
2.9	72.1
2.1	78.3
4.7	62.9
6.8	54.4
5.2	55.9
4.2	66.0
1.5	76.0
3.9	72.3

4.) The World Bank collected data on the percentage of GDP that a country spends on health expenditures ("Health expenditure," 2013) and also the percentage of women receiving prenatal care ("Pregnant woman receiving," 2013). The data for the countries where this information are available for the year 2011 is in table #10.1.8. Create a scatter plot of the data and find a regression equation between percentage spent on health expenditure and the percentage of women receiving prenatal care. Then use the regression equation to find the percent of women receiving prenatal care for a country that spends 5.0% of GDP on health expenditure and for a country that spends 12.0% of GDP. Which prenatal care percentage that you calculated do you think is closer to the true percentage? Why?

Table #10.1.8: Data of Health Expenditure versus Prenatal Care

Health Expenditure (% of GDP)	Prenatal Care (%)
9.6	47.9
3.7	54.6
5.2	93.7
5.2	84.7
10.0	100.0
4.7	42.5
4.8	96.4
6.0	77.1
5.4	58.3
4.8	95.4
4.1	78.0
6.0	93.3
9.5	93.3
6.8	93.7
6.1	89.8

5.) The height and weight of baseball players are in table #10.1.9 ("MLB heightsweights," 2013). Create a scatter plot and find a regression equation between height and weight of baseball players. Then use the regression equation to find the weight of a baseball player that is 75 inches tall and for a baseball player that is 68 inches tall. Which weight that you calculated do you think is closer to the true weight? Why?

Table #10.1.9: Heights and Weights of Baseball Players

Height (inches)	Weight (pounds)
76	212
76	224
72	180
74	210
75	215
71	200
77	235
78	235
77	194
76	185
72	180
72	170
75	220
74	228
73	210
72	180
70	185
73	190
71	186
74	200
74	200
75	210
78	240
72	208
75	180

6.) Different species have different body weights and brain weights are in table #10.1.10. ("Brain2bodyweight," 2013). Create a scatter plot and find a regression equation between body weights and brain weights. Then use the regression equation to find the brain weight for a species that has a body weight of 62 kg and for a species that has a body weight of 180,000 kg. Which brain weight that you calculated do you think is closer to the true brain weight? Why?

Table #10.1.10: Body Weights and Brain Weights of Species

Species	Body Weight (kg)	Brain Weight (kg)
Newborn Human	3.20	0.37
Adult Human	73.00	1.35
Pithecanthropus Man	70.00	0.93
Squirrel	0.80	0.01
Hamster	0.15	0.00
Chimpanzee	50.00	0.42
Rabbit	1.40	0.01
Dog_(Beagle)	10.00	0.07
Cat	4.50	0.03
Rat	0.40	0.00
Bottle-Nosed Dolphin	400.00	1.50
Beaver	24.00	0.04
Gorilla	320.00	0.50
Tiger	170.00	0.26
Owl	1.50	0.00
Camel	550.00	0.76
Elephant	4600.00	6.00
Lion	187.00	0.24
Sheep	120.00	0.14
Walrus	800.00	0.93
Horse	450.00	0.50
Cow	700.00	0.44
Giraffe	950.00	0.53
Green Lizard	0.20	0.00
Sperm Whale	35000.00	7.80
Turtle	3.00	0.00
Alligator	270.00	0.01

7.) A random sample of beef hotdogs was taken and the amount of sodium (in mg) and calories were measured. ("Data hotdogs," 2013) The data are in table #10.1.11. Create a scatter plot and find a regression equation between amount of calories and amount of sodium. Then use the regression equation to find the amount of sodium a beef hotdog has if it is 170 calories and if it is 120 calories. Which sodium level that you calculated do you think is closer to the true sodium level? Why?

Table #10.1.11: Calories and Sodium Levels in Beef Hotdogs

Calories	Sodium
186	495
181	477
176	425
149	322
184	482
190	587
158	370
139	322
175	479
148	375
152	330
111	300
141	386
153	401
190	645
157	440
131	317
149	319
135	298
132	253

8.) Per capita income in 1960 dollars for European countries and the percent of the labor force that works in agriculture in 1960 are in table #10.1.12 ("OECD economic development," 2013). Create a scatter plot and find a regression equation between percent of labor force in agriculture and per capita income. Then use the regression equation to find the per capita income in a country that has 21 percent of labor in agriculture and in a country that has 2 percent of labor in agriculture. Which per capita income that you calculated do you think is closer to the true income? Why?

Table #10.1.12: Percent of Labor in Agriculture and Per Capita Income for European Countries

Country	Percent in Agriculture	Per capita income
Sweden	14	1644
Switzerland	11	1361
Luxembourg	15	1242
U. Kingdom	4	1105
Denmark	18	1049
W. Germany	15	1035
France	20	1013
Belgium	6	1005
Norway	20	977
Iceland	25	839
Netherlands	11	810
Austria	23	681
Ireland	36	529
Italy	27	504
Greece	56	324
Spain	42	290
Portugal	44	238
Turkey	79	177

9.) Cigarette smoking and cancer have been linked. The number of deaths per one hundred thousand from bladder cancer and the number of cigarettes sold per capita in 1960 are in table #10.1.13 ("Smoking and cancer," 2013). Create a scatter plot and find a regression equation between cigarette smoking and deaths of bladder cancer. Then use the regression equation to find the number of deaths from bladder cancer when the cigarette sales were 20 per capita and when the cigarette sales were 6 per capita. Which number of deaths that you calculated do you think is closer to the true number? Why?

Table #10.1.13: Number of Cigarettes and Number of Bladder Cancer Deaths in 1960

Cigarette Sales (per Capita)	Bladder Cancer Deaths (per 100 Thousand)	Cigarette Sales (per Capita)	Bladder Cancer Deaths (per 100 Thousand)
18.20	2.90	42.40	6.54
25.82	3.52	28.64	5.98
18.24	2.99	21.16	2.90
28.60	4.46	29.14	5.30
31.10	5.11	19.96	2.89
33.60	4.78	26.38	4.47
40.46	5.60	23.44	2.93
28.27	4.46	23.78	4.89
20.10	3.08	29.18	4.99
27.91	4.75	18.06	3.25
26.18	4.09	20.94	3.64
22.12	4.23	20.08	2.94
21.84	2.91	22.57	3.21
23.44	2.86	14.00	3.31
21.58	4.65	25.89	4.63
28.92	4.79	21.17	4.04
25.91	5.21	21.25	5.14
26.92	4.69	22.86	4.78
24.96	5.27	28.04	3.20
22.06	3.72	30.34	3.46
16.08	3.06	23.75	3.95
27.56	4.04	23.32	3.72

10.) The weight of a car can influence the mileage that the car can obtain. A random sample of cars' weights and mileage was collected and are in table #10.1.14 ("Passenger car mileage," 2013). Create a scatter plot and find a regression equation between weight of cars and mileage. Then use the regression equation to find the mileage on a car that weighs 3800 pounds and on a car that weighs 2000 pounds. Which mileage that you calculated do you think is closer to the true mileage? Why?

Table #10.1.14: Weights and Mileages of Cars

Weight (100 pounds)	Mileage (mpg)
22.5	53.3
22.5	41.1
22.5	38.9
25.0	40.9
27.5	46.9
27.5	36.3
30.0	32.2
30.0	32.2
30.0	31.5
30.0	31.4
30.0	31.4
35.0	32.6
35.0	31.3
35.0	31.3
35.0	28.0
35.0	28.0
35.0	28.0
40.0	23.6
40.0	23.6
40.0	23.4
40.0	23.1
45.0	19.5
45.0	17.2
45.0	17.0
55.0	13.2

Section 10.2: Correlation

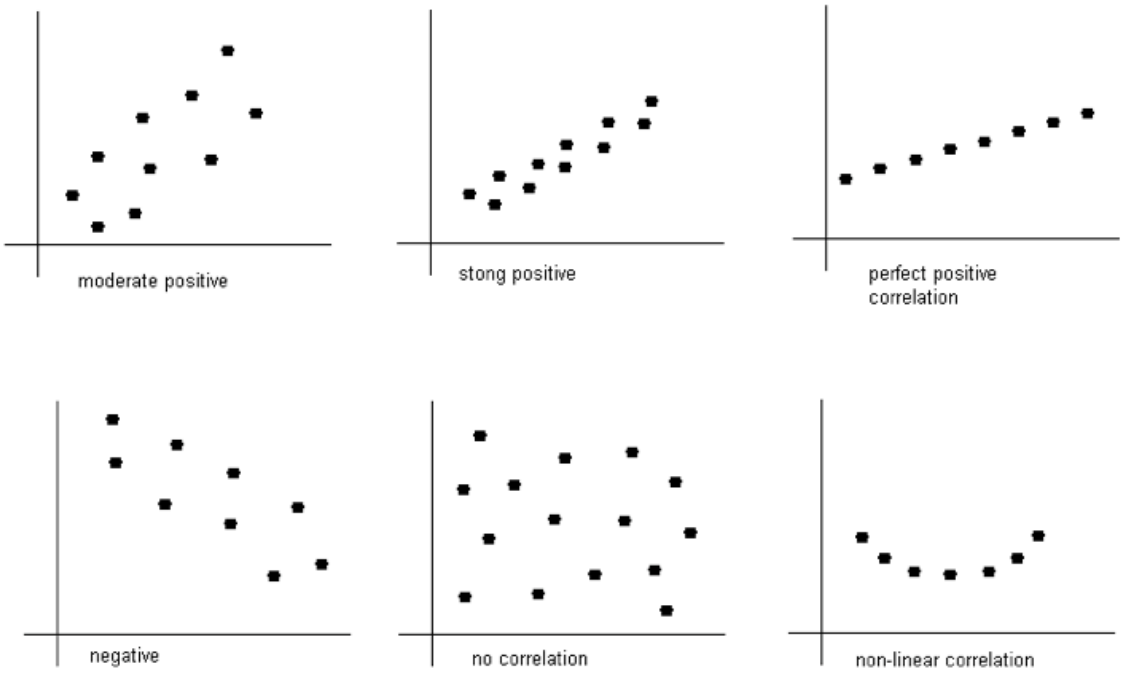
A **correlation** exists between two variables when the values of one variable are somehow associated with the values of the other variable.

When you see a pattern in the data you say there is a correlation in the data. Though this book is only dealing with linear patterns, patterns can be exponential, logarithmic, or periodic. To see this pattern, you can draw a scatter plot of the data.

Remember to read graphs from left to right, the same as you read words. If the graph goes up the correlation is positive and if the graph goes down the correlation is negative.

The words " weak", "moderate", and "strong" are used to describe the strength of the relationship between the two variables.

Figure 10.2.1: Correlation Graphs



The **linear correlation coefficient** is a number that describes the strength of the linear relationship between the two variables. It is also called the Pearson correlation coefficient after Karl Pearson who developed it. The symbol for the sample linear correlation coefficient is *r*. The symbol for the population correlation coefficient is ρ (Greek letter rho).

The formula for r is

$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

Where

$$SS_x = \sum (x - \bar{x})^2$$
$$SS_y = \sum (y - \bar{y})^2$$
$$SS_{xy} = \sum (x - \bar{x})(y - \bar{y})$$

Assumptions of linear correlation are the same as the assumptions for the regression line:

a. The set (x, y) of ordered pairs is a random sample from the population of all such possible (x, y) pairs.
b. For each fixed value of x, the y-values have a normal distribution. All of the y distributions have the same variance, and for a given x-value, the distribution of y-values has a mean that lies on the least squares line. You also assume that for a fixed y, each x has its own normal distribution. This is difficult to figure out, so you can use the following to determine if you have a normal distribution.
   i. Look to see if the scatter plot has a linear pattern.
   ii. Examine the residuals to see if there is randomness in the residuals. If there is a pattern to the residuals, then there is an issue in the data.

Interpretation of the correlation coefficient
r is always between -1 and 1. $r = -1$ means there is a perfect negative linear correlation and $r = 1$ means there is a perfect positive correlation. The closer r is to 1 or -1, the stronger the correlation. The closer r is to 0, the weaker the correlation. CAREFUL: $r = 0$ does not mean there is no correlation. It just means there is **no linear correlation.** There might be a very strong curved pattern.

Example #10.2.1: Calculating the Linear Correlation Coefficient, *r*

How strong is the positive relationship between the alcohol content and the number of calories in 12-ounce beer? To determine if there is a positive linear correlation, a random sample was taken of beer's alcohol content and calories for several different beers ("Calories in beer,," 2011), and the data are in table #10.2.1. Find the correlation coefficient and interpret that value.

Table #10.2.1: Alcohol and Calorie Content in Beer without Outlier

Brand	Brewery	Alcohol Content	Calories in 12 oz
Big Sky Scape Goat Pale Ale	Big Sky Brewing	4.70%	163
Sierra Nevada Harvest Ale	Sierra Nevada	6.70%	215
Steel Reserve	MillerCoors	8.10%	222
Coors Light	MillerCoors	4.15%	104
Genesee Cream Ale	High Falls Brewing	5.10%	162
Sierra Nevada Summerfest Beer	Sierra Nevada	5.00%	158
Michelob Beer	Anheuser Busch	5.00%	155
Flying Dog Doggie Style	Flying Dog Brewery	4.70%	158
Big Sky I.P.A.	Big Sky Brewing	6.20%	195

Solution:

State random variables

x = alcohol content in the beer

y = calories in 12 ounce beer

Assumptions check:

From example #10.1.2, the assumptions have been met.

To compute the correlation coefficient using the TI-83/84 calculator, use the LinRegTTest in the STAT menu. The setup is in figure 10.2.2. The reason that >0 was chosen is because the question was asked if there was a positive correlation. If you are asked if there is a negative correlation, then pick <0. If you are just asked if there is a correlation, then pick $\neq 0$. Right now the choice will not make a different, but it will be important later.

Figure #10.2.2: Setup for Linear Regression Test on TI-83/84

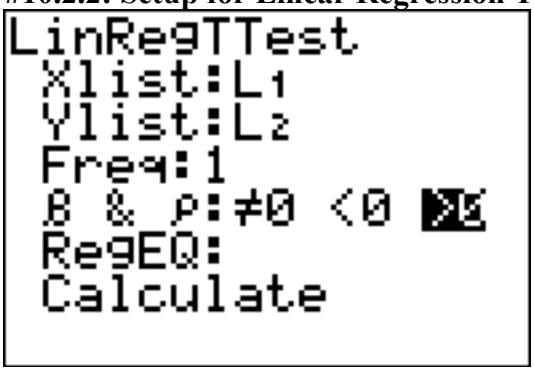
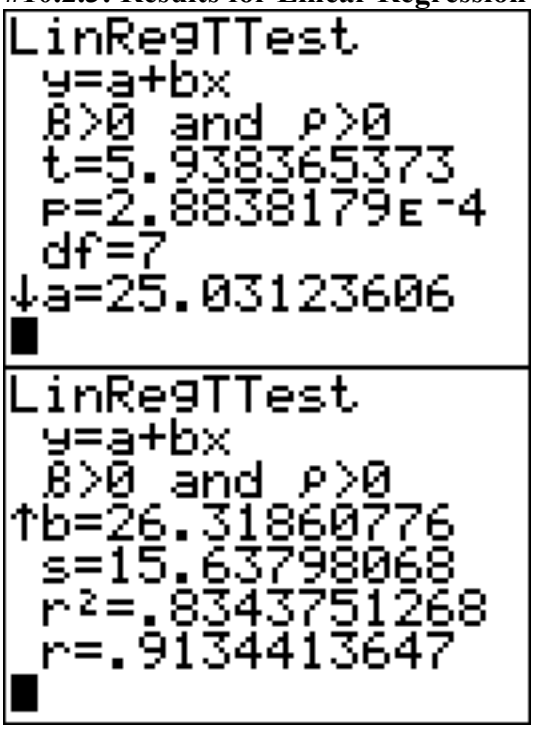


Figure #10.2.3: Results for Linear Regression Test on TI-83/84



To compute the correlation coefficient in R, the command is cor(independent variable, dependent variable). So for this example the command would be cor(alcohol, calories). The output is

[1] 0.9134414

The correlation coefficient is $r = 0.913$. This is close to 1, so it looks like there is a strong, positive correlation.

Causation

One common mistake people make is to assume that because there is a correlation, then one variable causes the other. This is usually not the case. That would be like saying the amount of alcohol in the beer causes it to have a certain number of calories. However, fermentation of sugars is what causes the alcohol content. The more sugars you have, the more alcohol can be made, and the more sugar, the higher the calories. It is actually the amount of sugar that causes both. Do not confuse the idea of correlation with the concept of causation. Just because two variables are correlated does not mean one causes the other to happen.

Example #10.2.2: Correlation Versus Causation

a.) A study showed a strong linear correlation between per capita beer consumption and teacher's salaries. Does giving a teacher a raise cause people to buy more beer? Does buying more beer cause teachers to get a raise?

Solution:
There is probably some other factor causing both of them to increase at the same time. Think about this: In a town where people have little extra money, they won't have money for beer and they won't give teachers raises. In another town where people have more extra money to spend it will be easier for them to buy more beer and they would be more willing to give teachers raises.

b.) A study shows that there is a correlation between people who have had a root canal and those that have cancer. Does that mean having a root canal causes cancer?

Solution:
Just because there is positive correlation doesn't mean that one caused the other. It turns out that there is a positive correlation between eating carrots and cancer, but that doesn't mean that eating carrots causes cancer. In other words, there are lots of relationships you can find between two variables, but that doesn't mean that one caused the other.

Remember a correlation only means a pattern exists. It does not mean that one variable causes the other variable to change.

Explained Variation
As stated before, there is some variability in the dependent variable values, such as calories. Some of the variation in calories is due to alcohol content and some is due to other factors. How much of the variation in the calories is due to alcohol content?

When considering this question, you want to look at how much of the variation in calories is explained by alcohol content and how much is explained by other variables. Realize that some of the changes in calories have to do with other ingredients. You can have two beers at the same alcohol content, but beer one has higher calories because of the other ingredients. Some variability is explained by the model and some variability is not explained. Together, both of these give the total variability. This is

$$(\text{total variation}) = (\text{explained variation}) + (\text{unexplained variation})$$

$$\sum (y-\overline{y})^2 \quad = \quad \sum (\hat{y}-\overline{y})^2 \quad + \quad \sum (y-\hat{y})^2$$

The proportion of the variation that is explained by the model is

$$r^2 = \frac{\text{explained variation}}{\text{total variation}}$$

This is known as the **coefficient of determination**.

To find the coefficient of determination, you square the correlation coefficient. In addition, r^2 is part of the calculator results.

Example #10.2.3: Finding the Coefficient of Determination

Find the coefficient of variation in calories that is explained by the linear relationship between alcohol content and calories and interpret the value.

Solution:

From the calculator results,

$$r^2 = 0.8344$$

Using R, you can do (cor(independent variable, dependent variable))^2. So that would be (cor(alcohol, calories))^2, and the output would be

[1] 0.8343751

Or you can just use a calculator and square the correlation value.

Thus, 83.44% of the variation in calories is explained to the linear relationship between alcohol content and calories. The other 16.56% of the variation is due to other factors. A really good coefficient of determination has a very small, unexplained part.

Example #10.2.4: Using the Formula to Calculate *r* and r^2

How strong is the relationship between the alcohol content and the number of calories in 12-ounce beer? To determine if there is a positive linear correlation, a random sample was taken of beer's alcohol content and calories for several different beers ("Calories in beer,," 2011), and the data are in table #10.2.1. Find the correlation coefficient and the coefficient of determination using the formula.

Solution:

From example #10.1.2, $SS_x = 12.45$, $SS_y = 10335.5556$, $SS_{xy} = 327.6667$

Correlation coefficient:

$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{327.6667}{\sqrt{12.45 * 10335.5556}} \approx 0.913$$

Coefficient of determination:

$$r^2 = (r)^2 = (0.913)^2 \approx 0.834$$

Now that you have a correlation coefficient, how can you tell if it is significant or not? This will be answered in the next section.

Section 10.2: Homework

For each problem, state the random variables. Also, look to see if there are any outliers that need to be removed. Do the correlation analysis with and without the suspected outlier points to determine if their removal affects the correlation. The data sets in this section are in section 10.1 and will be used in section 10.3.

1.) When an anthropologist finds skeletal remains, they need to figure out the height of the person. The height of a person (in cm) and the length of their metacarpal bone 1 (in cm) were collected and are in table #10.1.5 ("Prediction of height," 2013). Find the correlation coefficient and coefficient of determination and then interpret both.

2.) Table #10.1.6 contains the value of the house and the amount of rental income in a year that the house brings in ("Capital and rental," 2013). Find the correlation coefficient and coefficient of determination and then interpret both.

3.) The World Bank collects information on the life expectancy of a person in each country ("Life expectancy at," 2013) and the fertility rate per woman in the country ("Fertility rate," 2013). The data for 24 randomly selected countries for the year 2011 are in table #10.1.7. Find the correlation coefficient and coefficient of determination and then interpret both.

4.) The World Bank collected data on the percentage of GDP that a country spends on health expenditures ("Health expenditure," 2013) and also the percentage of women receiving prenatal care ("Pregnant woman receiving," 2013). The data for the countries where this information is available for the year 2011 are in table #10.1.8. Find the correlation coefficient and coefficient of determination and then interpret both.

5.) The height and weight of baseball players are in table #10.1.9 ("MLB heightsweights," 2013). Find the correlation coefficient and coefficient of determination and then interpret both.

6.) Different species have different body weights and brain weights are in table #10.1.10. ("Brain2bodyweight," 2013). Find the correlation coefficient and coefficient of determination and then interpret both.

7.) A random sample of beef hotdogs was taken and the amount of sodium (in mg) and calories were measured. ("Data hotdogs," 2013) The data are in table #10.1.11. Find the correlation coefficient and coefficient of determination and then interpret both.

8.) Per capita income in 1960 dollars for European countries and the percent of the labor force that works in agriculture in 1960 are in table #10.1.12 ("OECD economic development," 2013). Find the correlation coefficient and coefficient of determination and then interpret both.

9.) Cigarette smoking and cancer have been linked. The number of deaths per one hundred thousand from bladder cancer and the number of cigarettes sold per capita in 1960 are in table #10.1.13 ("Smoking and cancer," 2013). Find the correlation coefficient and coefficient of determination and then interpret both.

10.) The weight of a car can influence the mileage that the car can obtain. A random sample of cars weights and mileage was collected and are in table #10.1.14 ("Passenger car mileage," 2013). Find the correlation coefficient and coefficient of determination and then interpret both.

11.) There is a negative correlation between police expenditure and crime rate. Does this mean that spending more money on police causes the crime rate to decrease? Explain your answer.

12.) There is a positive correlation between tobacco sales and alcohol sales. Does that mean that using tobacco causes a person to also drink alcohol? Explain your answer.

13.) There is a positive correlation between the average temperature in a location and the morality rate from breast cancer. Does that mean that higher temperatures cause more women to die of breast cancer? Explain your answer.

14.) There is a positive correlation between the length of time a tableware company polishes a dish and the price of the dish. Does that mean that the time a plate is polished determines the price of the dish? Explain your answer.

Section 10.3: Inference for Regression and Correlation

How do you really say you have a correlation? Can you test to see if there really is a correlation? Of course, the answer is yes. The hypothesis test for correlation is as follows:

Hypothesis Test for Correlation:

1. State the random variables in words.
   x = independent variable
   y = dependent variable

2. State the null and alternative hypotheses and the level of significance
   $H_o : \rho = 0$ (There is no correlation)
   $H_A : \rho \neq 0$ (There is a correlation)
   or
   $H_A : \rho < 0$ (There is a negative correlation)
   or
   $H_A : \rho > 0$ (There is a postive correlation)
   Also, state your α level here.

3. State and check the assumptions for the hypothesis test
   The assumptions for the hypothesis test are the same assumptions for regression and correlation.

4. Find the test statistic and p-value
   $$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$
   with degrees of freedom = $df = n - 2$
   p-value:
   Using the TI-83/84: $\text{tcdf}(\text{lower limit}, \text{upper limit}, df)$
   (Note: if $H_A : \rho < 0$, then lower limit is $-1E99$ and upper limit is your test statistic. If $H_A : \rho > 0$, then lower limit is your test statistic and the upper limit is $1E99$. If $H_A : \rho \neq 0$, then find the p-value for $H_A : \rho < 0$, and multiply by 2.)
   Using R: $\text{pt}(t, df)$
   (Note: if $H_A : \rho < 0$, then use $\text{pt}(t, df)$, If $H_A : \rho > 0$, then use $1 - \text{pt}(t, df)$. If $H_A : \rho \neq 0$, then find the p-value for $H_A : \rho < 0$, and multiply by 2.)

5. Conclusion

This is where you write reject H_o or fail to reject H_o. The rule is: if the p-value $< \alpha$, then reject H_o. If the p-value $\geq \alpha$, then fail to reject H_o

6. Interpretation
   This is where you interpret in real world terms the conclusion to the test. The conclusion for a hypothesis test is that you either have enough evidence to show H_A is true, or you do not have enough evidence to show H_A is true.

Note: the TI-83/84 calculator results give you the test statistic and the p-value. In R, the command for getting the test statistic and p-value is cor.test(independent variable, dependent variable, alternative = "less" or "greater"). Use less for $H_A : \rho < 0$, use greater for $H_A : \rho > 0$, and leave off this command for $H_A : \rho \neq 0$.

Example #10.3.1: Testing the Claim of a Linear Correlation

Is there a positive correlation between beer's alcohol content and calories? To determine if there is a positive linear correlation, a random sample was taken of beer's alcohol content and calories for several different beers ("Calories in beer,," 2011), and the data is in table #10.2.1. Test at the 5% level.

Solution:

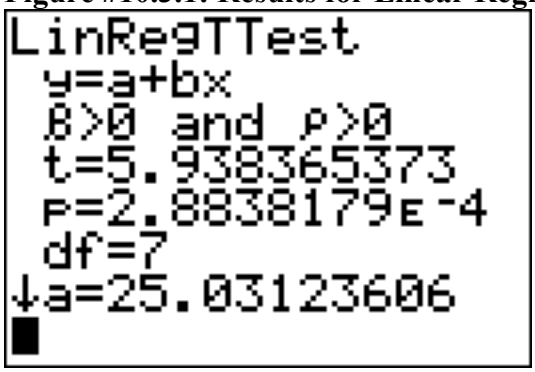
1. State the random variables in words.
   x = alcohol content in the beer
   y = calories in 12 ounce beer

2. State the null and alternative hypotheses and the level of significance
   Since you are asked if there is a positive correlation, $\rho > 0$.
   $H_o : \rho = 0$
   $H_A : \rho > 0$
   $\alpha = 0.05$

3. State and check the assumptions for the hypothesis test
   The assumptions for the hypothesis test were already checked in example #10.1.2.

4. Find the test statistic and p-value
   The results from the TI-83/84 calculator are in figure #10.3.1.

Figure #10.3.1: Results for Linear Regression Test on TI-83/84



Test statistic: $t \approx 5.938$ and p-value: $p \approx 2.884 \times 10^{-4}$

The results from R are

cor.test(alcohol, calories, alternative = "greater")

Pearson's product-moment correlation

data: alcohol and calories
t = 5.9384, df = 7, p-value = 0.0002884
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
0.7046161 1.0000000
sample estimates:
cor
0.9134414

Test statistic: $t \approx 5.9384$ and p-value: $p \approx 0.0002884$

5. Conclusion
   Reject H_o since the p-value is less than 0.05.

6. Interpretation
   There is enough evidence to show that there is a positive correlation between alcohol content and number of calories in a 12-ounce bottle of beer.

Prediction Interval

Using the regression equation you can predict the number of calories from the alcohol content. However, you only find one value. The problem is that beers vary a bit in calories even if they have the same alcohol content. It would be nice to have a range instead of a single value. The range is called a prediction interval. To find this, you need to figure out how much error is in the estimate from the regression equation. This is known as the **standard error of the estimate**.

Standard Error of the Estimate
This is the sum of squares of the residuals

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

This formula is hard to work with, so there is an easier formula. You can also find the value from technology, such as the calculator.

$$s_e = \sqrt{\frac{SS_y - b * SS_{xy}}{n-2}}$$

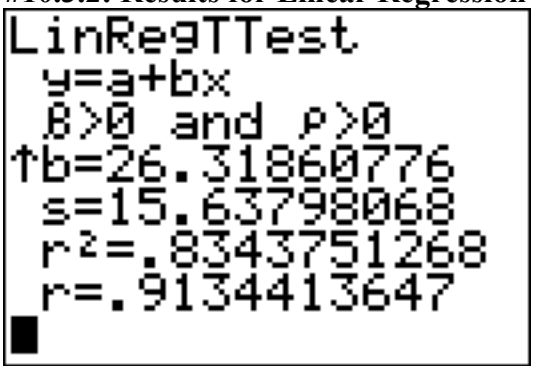
Example #10.3.2: Finding the Standard Error of the Estimate

Find the standard error of the estimate for the beer data. To determine if there is a positive linear correlation, a random sample was taken of beer's alcohol content and calories for several different beers ("Calories in beer,," 2011), and the data are in table #10.2.1.

Solution:
x = alcohol content in the beer
y = calories in 12 ounce beer

Using the TI-83/84, the results are in figure #10.3.2.

Figure #10.3.2: Results for Linear Regression Test on TI-83/84



The s in the results is the standard error of the estimate. So $s_e \approx 15.64$.

To find the standard error of the estimate in R, the commands are
lm.out = lm(dependent variable ~ independent variable) – this defines the linear model with a name so you can use it later. Then
summary(lm.out) – this will produce most of the information you need for a regression and correlation analysis. In fact, the only thing R doesn't produce with this command is the correlation coefficient. Otherwise, you can use the command to find the regression equation, coefficient of determination, test statistic, p-value for a two-tailed test, and standard error of the estimate.

The results from R are

```
lm.out=lm(calories~alcohol)
summary(lm.out)

Call:
lm(formula = calories ~ alcohol)

Residuals:
   Min     1Q  Median     3Q    Max
-30.253  -1.624   2.744   9.271  14.271

Coefficients:
            Estimate Std.  Error    t value   Pr(>|t|)
(Intercept)   25.031     24.999   1.001   0.350038
alcohol       26.319      4.432    5.938   0.000577 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.64 on 7 degrees of freedom
Multiple R-squared:  0.8344,  Adjusted R-squared:  0.8107
F-statistic: 35.26 on 1 and 7 DF,  p-value: 0.0005768
```

From this output, you can find the y-intercept is 25.031, the slope is 26.319, the test statistic is t = 5.938, the p-value for the two-tailed test is 0.000577. If you want the p-value for a one-tailed test, divide this number by 2. The standard error of the estimate is the residual standard error and is 15.64. There is some information in this output that you do not need.

If you want to know how to calculate the standard error of the estimate from the formula, refer to example# 10.3.3.

Example #10.3.3: Finding the Standard Error of the Estimate from the Formula

Find the standard error of the estimate for the beer data using the formula. To determine if there is a positive linear correlation, a random sample was taken of beer’s alcohol content and calories for several different beers ("Calories in beer,," 2011), and the data are in table #10.2.1.

Solution:

x = alcohol content in the beer
y = calories in 12 ounce beer

From Example #10.1.3:

$$SS_y = \sum (y - \overline{y})^2 = 10335.56$$

$$SS_{xy} = \sum (x - \overline{x})(y - \overline{y}) = 327.6666$$

$$n = 9$$

$$b = 26.3$$

The standard error of the estimate is

$$s_e = \sqrt{\frac{SS_y - b * SS_{xy}}{n-2}}$$
$$= \sqrt{\frac{10335.56 - 26.3(327.6666)}{9-2}}$$
$$= 15.67$$

Prediction Interval for an Individual *y*

Given the fixed value x_o, the prediction interval for an individual *y* is

$$\hat{y} - E < y < \hat{y} + E$$

where

$$\hat{y} = a + bx$$

$$E = t_c s_e \sqrt{1 + \frac{1}{n} + \frac{(x_o - \overline{x})^2}{SS_x}}$$

$$df = n - 2$$

Note: to find

$$SS_x = \sum (x - \overline{x})^2$$

remember, the standard devation formula from chapter 3

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

So,

$$s_x = \sqrt{\frac{SS_x}{n-1}}$$

Now solve for SS_x

$$SS_x = s_x^2 (n-1)$$

You can get the standard deviation from technology.

R will produce the prediction interval for you. The commands are (Note you probably already did the lm.out command. You do not need to do it again.)

lm.out = lm(dependent variable ~ independent variable) – calculates the linear model

predict(lm.out, newdata=list(independent variable = value), interval="prediction", level=C) – will compute a prediction interval for the independent variable set to a particular value (put that value in place of the word value), at a particular C level (given as a decimal)

Example #10.3.4: Find the Prediction Interval

Find a 95% prediction interval for the number of calories when the alcohol content is 6.5% using a random sample taken of beer's alcohol content and calories ("Calories in beer,," 2011). The data are in table #10.2.1.

Solution:

x = alcohol content in the beer
y = calories in 12 ounce beer

Computing the prediction interval using the TI-83/84 calculator:
From Example #10.1.2

$\hat{y} = 25.0 + 26.3x$

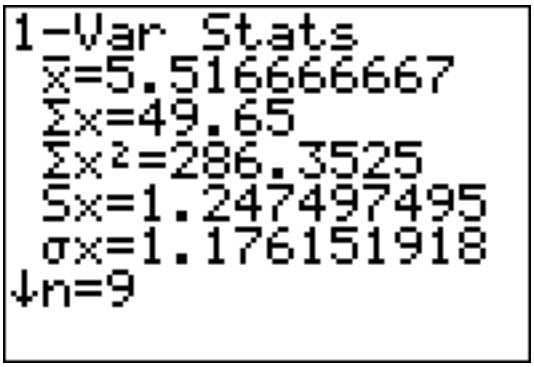
$x_o = 6.50$

$\hat{y} = 25.0 + 26.3(6.50) = 196$ calories

From Example #10.3.2

$s_e \approx 15.64$

Figure #10.3.3: Results of 1-Var Stats on TI-83/84



$\overline{x} = 5.517$

$s_x = 1.247497495$

$n = 9$

Now you can find

$SS_x = s_x^2(n-1)$

$= (1.247497495)^2(9-1)$

$= 12.45$

$df = n - 2 = 9 - 2 = 7$

Now look in table A.2. Go down the first column to 7, then over to the column headed with 95%.

$t_c = 2.365$

$$E = t_c s_e \sqrt{1 + \frac{1}{n} + \frac{(x_o - \bar{x})^2}{SS_x}}$$

$$= 2.365(15.64)\sqrt{1 + \frac{1}{9} + \frac{(6.50 - 5.517)^2}{12.45}}$$

$$= 40.3$$

Prediction interval is

$$\hat{y} - E < y < \hat{y} + E$$

$$196 - 40.3 < y < 196 + 40.3$$

$$155.7 < y < 236.3$$

Computing the prediction interval using R:

```
predict(lm.out, newdata=list(alcohol=6.5), interval = "prediction", level=0.95)
      fit       lwr       upr
1 196.1022 155.7847 236.4196
```

fit = $\hat{y}$ when x = 6.5%. lwr = lower limit of prediction interval. upr = upper limit of prediction interval. So the prediction interval is $155.8 < y < 236.4$.

Statistical interpretation: There is a 95% chance that the interval $155.8 < y < 236.4$ contains the true value for the calories when the alcohol content is 6.5%.

Real world interpretation: If a beer has an alcohol content of 6.50% then it has between 156 and 236 calories.

Example #10.3.5: Doing a Correlation and Regression Analysis Using the TI-83/84
Table #10.3.1 contains randomly selected high temperatures at various cities on a single day and the elevation of the city.

Table #10.3.1: Temperatures and Elevation of Cities on a Given Day

Elevation (in feet)	7000	4000	6000	3000	7000	4500	5000
Temperature (°F)	50	60	48	70	55	55	60

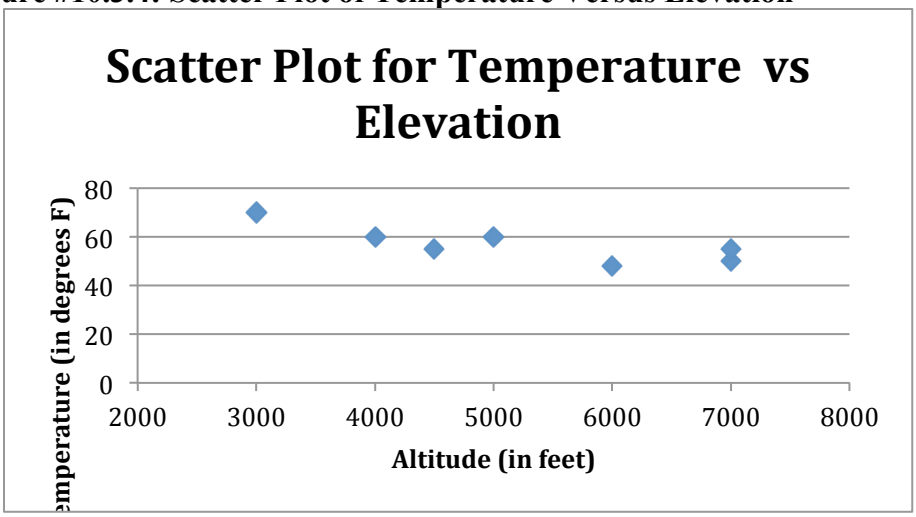
a.) State the random variables.
Solution:
x = elevation
y = high temperature

b.) Find a regression equation for elevation and high temperature on a given day.

Solution:
a. A random sample was taken as stated in the problem.

b. The distribution for each high temperature value is normally distributed for every value of elevation.
   i. Look at the scatter plot of high temperature versus elevation.

Figure #10.3.4: Scatter Plot of Temperature Versus Elevation



The scatter plot looks fairly linear.

   ii. There are no points that appear to be outliers.
   iii. The residual plot for temperature versus elevation appears to be fairly random. (See figure #10.3.7.)

It appears that the high temperature is normally distributed.

All calculations computed using the TI-83/84 calculator.

Figure #10.3.5: Setup for Linear Regression on TI-83/84 Calculator

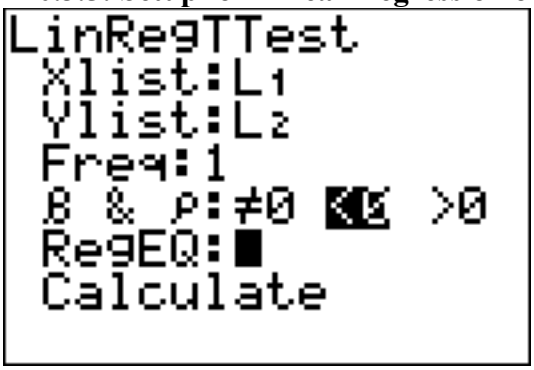
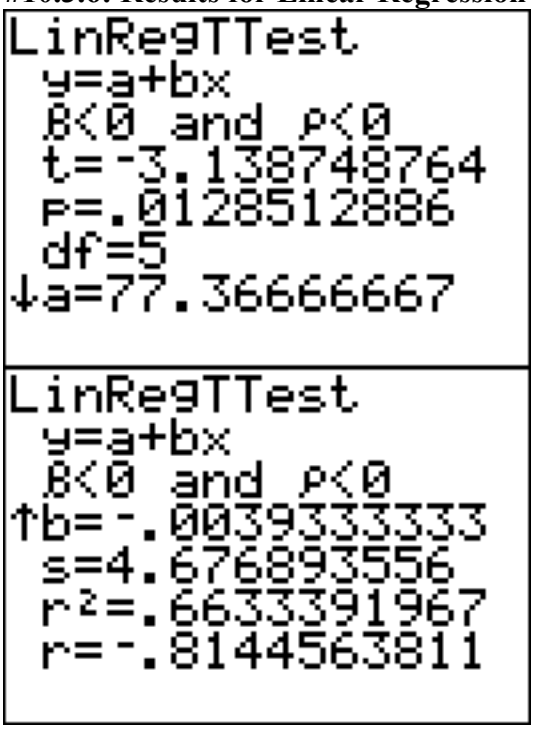


Figure #10.3.6: Results for Linear Regression on TI-83/84 Calculator



$\hat{y} = 77.4 - 0.0039x$

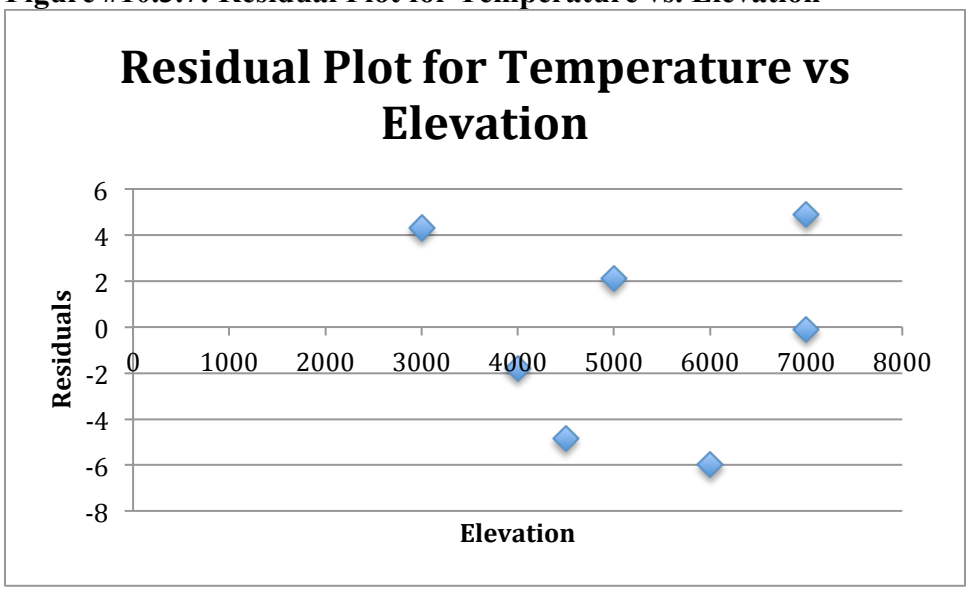
c.) Find the residuals and create a residual plot.

Solution:

Table #10.3.2: Residuals for Elevation vs. Temperature Data

x	y	$\hat{y}$	$y - \hat{y}$
7000	50	50.1	-0.1
4000	60	61.8	-1.8
6000	48	54.0	-6.0
3000	70	65.7	4.3
7000	55	50.1	4.9
4500	55	59.85	-4.85
5000	60	57.9	2.1

Figure #10.3.7: Residual Plot for Temperature vs. Elevation



The residuals appear to be fairly random.

d.) Use the regression equation to estimate the high temperature on that day at an elevation of 5500 ft.

Solution:

$x_o = 5500$

$\hat{y} = 77.4 - 0.0039(5500) = 55.95^\circ F$

e.) Use the regression equation to estimate the high temperature on that day at an elevation of 8000 ft.

Solution:

$x_o = 8000$

$\hat{y} = 77.4 - 0.0039(8000) = 46.2^\circ F$

f.) Between the answers to parts d and e, which estimate is probably more accurate and why?

Solution:

Part d is more accurate, since it is interpolation and part e is extrapolation.

g.) Find the correlation coefficient and coefficient of determination and interpret both.

Solution:

From figure #10.3.6, the correlation coefficient is

$r \approx -0.814$, which is moderate to strong negative correlation.

From figure #10.3.6, the coefficient of determination is

$r^2 \approx 0.663$, which means that 66.3% of the variability in high temperature is explained by the linear model. The other 33.7% is explained by other variables such as local weather conditions.

h.) Is there enough evidence to show a negative correlation between elevation and high temperature? Test at the 5% level.

Solution:

1. State the random variables in words.
   x = elevation
   y = high temperature

2. State the null and alternative hypotheses and the level of significance
   $H_o : \rho = 0$
   $H_A : \rho < 0$
   $\alpha = 0.05$

3. State and check the assumptions for the hypothesis test
   The assumptions for the hypothesis test were already checked part b.
4. Find the test statistic and p-value
   From figure #10.3.6,
   Test statistic:
   $t \approx -3.139$
   p-value:
   $p \approx 0.0129$

5. Conclusion
   Reject H_o since the p-value is less than 0.05.

6. Interpretation
   There is enough evidence to show that there is a negative correlation between elevation and high temperatures.

i.) Find the standard error of the estimate.

Solution:

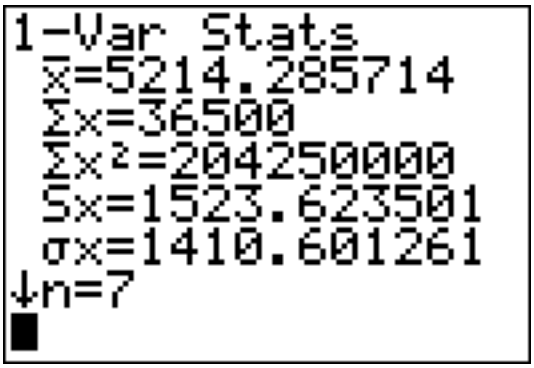
From figure #10.3.6,
$s_e \approx 4.677$

j.) Using a 95% prediction interval, find a range for high temperature for an elevation of 6500 feet.

Solution:

$\hat{y} = 77.4 - 0.0039(6500) \approx 52.1^\circ F$

Figure #10.3.8: Results of 1-Var Stats on TI-83/84



$\overline{x} = 5214.29$

$s_x = 1523.624$

$n = 7$

Now you can find

$$SS_x = s_x^{\,2}(n-1)$$
$$= (1523.623501)^2 (7-1)$$
$$= 13928571.43$$

$$df = n - 2 = 7 - 2 = 5$$

Now look in table A.2. Go down the first column to 5, then over to the column headed with 95%.

$t_c = 2.571$

So

$$E = t_c s_e \sqrt{1 + \frac{1}{n} + \frac{(x_o - \overline{x})^2}{SS_x}}$$
$$= 2.571(4.677)\sqrt{1 + \frac{1}{7} + \frac{(6500 - 5214.29)^2}{13928571.43}}$$
$$= 13.5$$

Prediction interval is

$\hat{y} - E < y < \hat{y} + E$

$52.1 - 13.5 < y < 52.1 + 13.5$

$38.6 < y < 65.6$

Statistical interpretation: There is a 95% chance that the interval $38.6 < y < 65.6$ contains the true value for the temperature at an elevation of 6500 feet.
Real world interpretation: A city of 6500 feet will have a high temperature between 38.6°F and 65.6°F. Though this interval is fairly wide, at least the interval tells you that the temperature isn't that warm.

Example #10.3.6: Doing a Correlation and Regression Analysis Using R
Table #10.3.1 contains randomly selected high temperatures at various cities on a single day and the elevation of the city.

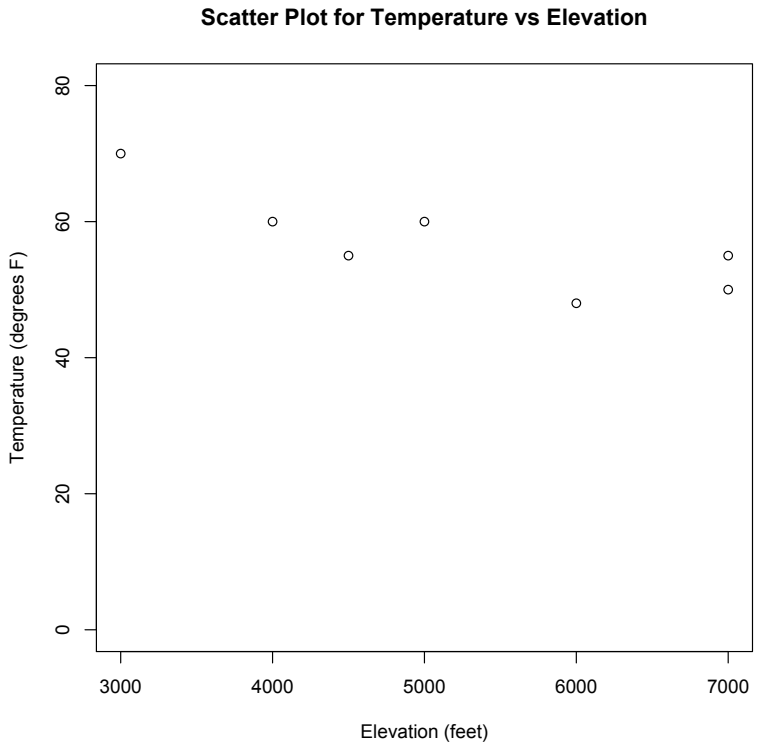
a.) State the random variables.
Solution:
x = elevation
y = high temperature

b.) Find a regression equation for elevation and high temperature on a given day.

Solution:
a. A random sample was taken as stated in the problem.
b. The distribution for each high temperature value is normally distributed for every value of elevation.
   i. Look at the scatter plot of high temperature versus elevation.

R command: plot(elevation, temperature, main="Scatter Plot for Temperature vs Elevation", xlab="Elevation (feet)", ylab="Temperature (degrees F)", ylim=c(0,80))

Figure #10.3.9: Scatter Plot of Temperature Versus Elevation



The scatter plot looks fairly linear.
   ii. The residual plot for temperature versus elevation appears to be fairly random. (See figure #10.3.10.)

It appears that the high temperature is normally distributed.

Using R:
Commands:
lm.out=lm(temperature ~ elevation)
summary(lm.out)

Output:
Call:
lm(formula = temperature ~ elevation)

Residuals:

1	2	3	4	5	6	7
0.1667	-1.6333	-5.7667	4.4333	5.1667	-4.6667	2.3000

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	77.366667	6.769182	11.429	8.98e-05	***
elevation	-0.003933	0.001253	-3.139	0.0257	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 4.677 on 5 degrees of freedom
Multiple R-squared: 0.6633, Adjusted R-squared: 0.596
F-statistic: 9.852 on 1 and 5 DF, p-value: 0.0257

From the output you can see the slope = -0.0039 and the y-intercept = 77.4. So the regression equation is:

$$\hat{y} = 77.4 - 0.0039x$$

c.) Find the residuals and create a residual plot.

Solution:

R command: (notice these are also in the summary(lm.out) output, but if you have too many data points, then R only gives a numerical summary of the residuals.)

residuals(lm.out)

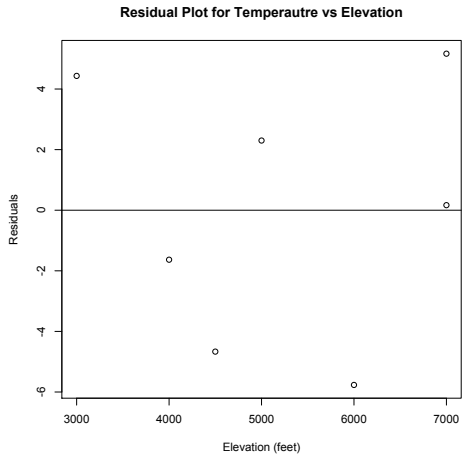
1	2	3	4	5	6	7
0.1666667	-1.6333333	-5.7666667	4.4333333	5.1666667	-4.6666667	2.3000000

So for the first x of 7000, the residual is approximately 0.1667. This means if you find the $\hat{y}$ for when x is 7000 and then subtract this answer from the y value of 50 that was measured, you would obtain 0.1667. Similar process is computed for the other residual values.

To plot the residuals, the R command is

```
plot(elevation, residuals(lm.out), main="Residual Plot for Temperautre vs
Elevation", xlab="Elevation (feet)", ylab="Residuals")
abline(0,0)
```

Figure #10.3.10: Residual Plot for Temperature vs. Elevation



The residuals appear to be fairly random.

d.) Use the regression equation to estimate the high temperature on that day at an elevation of 5500 ft.

Solution:

$x_o = 5500$

$\hat{y} = 77.4 - 0.0039(5500) = 55.95^\circ F$

e.) Use the regression equation to estimate the high temperature on that day at an elevation of 8000 ft.

Solution:

$x_o = 8000$

$\hat{y} = 77.4 - 0.0039(8000) = 46.2^\circ F$

f.) Between the answers to parts d and e, which estimate is probably more accurate and why?

Solution:

Part d is more accurate, since it is interpolation and part e is extrapolation.

g.) Find the correlation coefficient and coefficient of determination and interpret both.

Solution:
The R command for the correlation coefficient is
cor(elevation, temperature)
[1] -0.8144564

So, $r \approx -0.814$, which is moderate to strong negative correlation.
From summary(lm.out), the coefficient of determination is the Multiple R-squared. So $r^2 \approx 0.663$, which means that 66.3% of the variability in high temperature is explained by the linear model. The other 33.7% is explained by other variables such as local weather conditions.

h.) Is there enough evidence to show a negative correlation between elevation and high temperature? Test at the 5% level.

Solution:
1. State the random variables in words.
   x = elevation
   y = high temperature

2. State the null and alternative hypotheses and the level of significance
   $H_o : \rho = 0$
   $H_A : \rho < 0$
   $\alpha = 0.05$
3. State and check the assumptions for the hypothesis test
   The assumptions for the hypothesis test were already checked part b.

4. Find the test statistic and p-value
   The R command is cor.test(elevation, temperature, alternative = "less")

   Pearson's product-moment correlation

   data: elevation and temperature
   t = -3.1387, df = 5, p-value = 0.01285
   alternative hypothesis: true correlation is less than 0
   95 percent confidence interval:
   -1.0000000 -0.3074247
   sample estimates:
   cor
   -0.8144564

   Test statistic: $t \approx -3.1387$ and p-value: $p \approx 0.01285$

5. Conclusion
   Reject H_o since the p-value is less than 0.05.

6. Interpretation
   There is enough evidence to show that there is a negative correlation between elevation and high temperatures.

i.) Find the standard error of the estimate.

Solution:
From summary(lm.out), Residual standard error: 4.677.
So, $s_e \approx 4.677$

j.) Using a 95% prediction interval, find a range for high temperature for an elevation of 6500 feet.

Solution:
R command is predict(lm.out, newdata=list(elevation = 6500), interval = "prediction", level=0.95)

	fit	lwr	upr
1	51.8	38.29672	65.30328

So when x = 6500 feet, $\hat{y} = 51.8^\circ F$ and $38.29672 < y < 65.30328$.

Statistical interpretation: There is a 95% chance that the interval $38.3 < y < 65.3$ contains the true value for the temperature at an elevation of 6500 feet.
Real world interpretation: A city of 6500 feet will have a high temperature between 38.3°F and 65.3°F. Though this interval is fairly wide, at least the interval tells you that the temperature isn't that warm.

Section 10.3: Homework

For each problem, state the random variables. The data sets in this section are in the homework for section 10.1 and were also used in section 10.2. If you removed any data points as outliers in the other sections, remove them in this sections homework too.

1.) When an anthropologist finds skeletal remains, they need to figure out the height of the person. The height of a person (in cm) and the length of their metacarpal bone one (in cm) were collected and are in table #10.1.5 ("Prediction of height," 2013).
   a.) Test at the 1% level for a positive correlation between length of metacarpal bone one and height of a person.
   b.) Find the standard error of the estimate.
   c.) Compute a 99% prediction interval for height of a person with a metacarpal length of 44 cm.

2.) Table #10.1.6 contains the value of the house and the amount of rental income in a year that the house brings in ("Capital and rental," 2013).
   a.) Test at the 5% level for a positive correlation between house value and rental amount.
   b.) Find the standard error of the estimate.
   c.) Compute a 95% prediction interval for the rental income on a house worth $230,000.

3.) The World Bank collects information on the life expectancy of a person in each country ("Life expectancy at," 2013) and the fertility rate per woman in the country ("Fertility rate," 2013). The data for 24 randomly selected countries for the year 2011 are in table #10.1.7.
   a.) Test at the 1% level for a negative correlation between fertility rate and life expectancy.
   b.) Find the standard error of the estimate.
   c.) Compute a 99% prediction interval for the life expectancy for a country that has a fertility rate of 2.7.

4.) The World Bank collected data on the percentage of GDP that a country spends on health expenditures ("Health expenditure," 2013) and also the percentage of women receiving prenatal care ("Pregnant woman receiving," 2013). The data for the countries where this information is available for the year 2011 are in table #10.1.8.
   a.) Test at the 5% level for a correlation between percentage spent on health expenditure and the percentage of women receiving prenatal care.
   b.) Find the standard error of the estimate.
   c.) Compute a 95% prediction interval for the percentage of woman receiving prenatal care for a country that spends 5.0 % of GDP on health expenditure.

5.) The height and weight of baseball players are in table #10.1.9 ("MLB heightsweights," 2013).
   a.) Test at the 5% level for a positive correlation between height and weight of baseball players.
   b.) Find the standard error of the estimate.
   c.) Compute a 95% prediction interval for the weight of a baseball player that is 75 inches tall.

6.) Different species have different body weights and brain weights are in table #10.1.10. ("Brain2bodyweight," 2013).
   a.) Test at the 1% level for a positive correlation between body weights and brain weights.
   b.) Find the standard error of the estimate.
   c.) Compute a 99% prediction interval for the brain weight for a species that has a body weight of 62 kg.

7.) A random sample of beef hotdogs was taken and the amount of sodium (in mg) and calories were measured. ("Data hotdogs," 2013) The data are in table #10.1.11.
   a.) Test at the 5% level for a correlation between amount of calories and amount of sodium.
   b.) Find the standard error of the estimate.
   c.) Compute a 95% prediction interval for the amount of sodium a beef hotdog has if it is 170 calories.

8.) Per capita income in 1960 dollars for European countries and the percent of the labor force that works in agriculture in 1960 are in table #10.1.12 ("OECD economic development," 2013).
   a.) Test at the 5% level for a negative correlation between percent of labor force in agriculture and per capita income.
   b.) Find the standard error of the estimate.
   c.) Compute a 90% prediction interval for the per capita income in a country that has 21 percent of labor in agriculture.

9.) Cigarette smoking and cancer have been linked. The number of deaths per one hundred thousand from bladder cancer and the number of cigarettes sold per capita in 1960 are in table #10.1.13 ("Smoking and cancer," 2013).
   a.) Test at the 1% level for a positive correlation between cigarette smoking and deaths of bladder cancer.
   b.) Find the standard error of the estimate.
   c.) Compute a 99% prediction interval for the number of deaths from bladder cancer when the cigarette sales were 20 per capita.

10.) The weight of a car can influence the mileage that the car can obtain. A random sample of cars weights and mileage was collected and are in table #10.1.14 ("Passenger car mileage," 2013).
   a.) Test at the 5% level for a negative correlation between the weight of cars and mileage.
   b.) Find the standard error of the estimate.
   c.) Compute a 95% prediction interval for the mileage on a car that weighs 3800 pounds.

Data Source:
Brain2bodyweight. (2013, November 16). Retrieved from http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Brain2BodyWeight

Calories in beer, beer alcohol, beer carbohydrates. (2011, October 25). Retrieved from www.beer100.com/beercalories.htm

Capital and rental values of Auckland properties. (2013, September 26). Retrieved from http://www.statsci.org/data/oz/rentcap.html

Data hotdogs. (2013, November 16). Retrieved from http://wiki.stat.ucla.edu/socr/index.php/SOCR_012708_ID_Data_HotDogs

Fertility rate. (2013, October 14). Retrieved from http://data.worldbank.org/indicator/SP.DYN.TFRT.IN

Health expenditure. (2013, October 14). Retrieved from http://data.worldbank.org/indicator/SH.XPD.TOTL.ZS

Life expectancy at birth. (2013, October 14). Retrieved from http://data.worldbank.org/indicator/SP.DYN.LE00.IN

MLB heightsweights. (2013, November 16). Retrieved from http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_MLB_HeightsWeights

OECD economic development. (2013, December 04). Retrieved from http://lib.stat.cmu.edu/DASL/Datafiles/oecdat.html

Passenger car mileage. (2013, December 04). Retrieved from http://lib.stat.cmu.edu/DASL/Datafiles/carmpgdat.html

Prediction of height from metacarpal bone length. (2013, September 26). Retrieved from http://www.statsci.org/data/general/stature.html

Pregnant woman receiving prenatal care. (2013, October 14). Retrieved from http://data.worldbank.org/indicator/SH.STA.ANVC.ZS

Smoking and cancer. (2013, December 04). Retrieved from http://lib.stat.cmu.edu/DASL/Datafiles/cigcancerdat.html