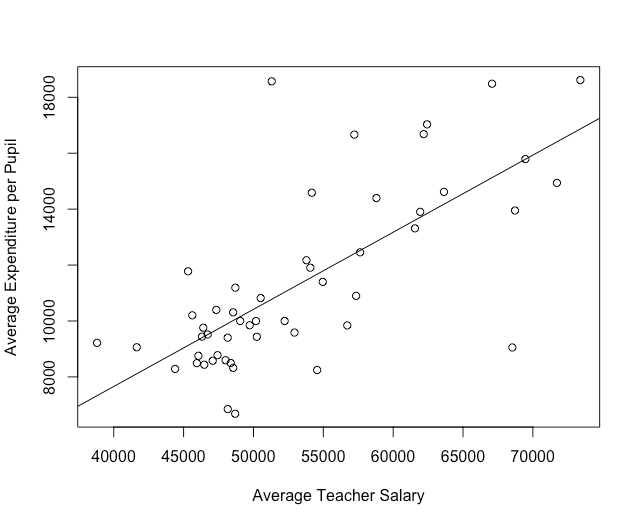
**CHAPTER 6 SOLUTIONS**

1. Because educexpe and teachpay are both scale- or ratio-leveled variables, their level of measurement is appropriate. Furthermore, the scatterplot indicates that the pattern of points is effectively approximated by a line. The R commands to generate the scatterplot with regression line superimposed are:

**plot(States$educexpe~States$teachpay, xlab = "Average Teacher Salary", ylab = "Average Expenditure per Pupil")**

**abline(lm(States$educexpe~States$teachpay))**

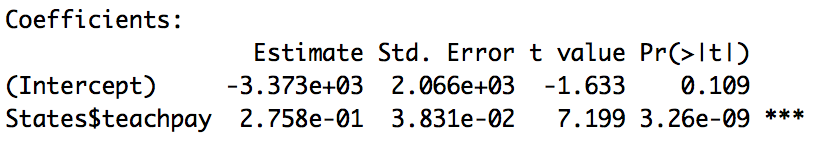


1. = – 3372.988 + .276(teachpay). The R commands to conduct the regression analysis and display the results are:

**model = lm(States$educexpe~States$teachpay)**

**summary(model)**

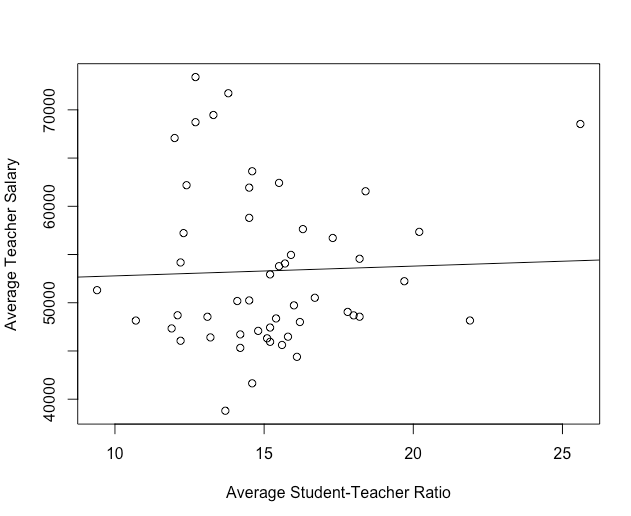
The part of the resulting output that is used to obtain the regression equation is displayed below.

****

1. A one dollar increase in the average annual salary for public school teachers corresponds to a .276 dollar (almost 28 cents) increase in the expenditure per pupil for the state, on average.
2. In this case the intercept is not meaningful because there are no states for which the average annual salary for teachers is close to $0.
3. = – 3372.988 + .276(40000) = $7,667.01.
4. No. The value 80,000 is well beyond the largest value of the data from which the model was created.
5. The output indicates that *R*-squared = .514. Taking the square root, we have *R* = .72, indicating that the correlation between the actual and predicted educational expenditures is strong.
6. Looking at scatterplots of teachpay with each of the three variables, although the first two scatterplots do not depict a strong linear relationship, there is not a simple curve that provides a better fit than the regression line. A line gives a good fit to the third scatterplot. In all three cases, regression is appropriate. The first scatterplot was obtained using the R commands:

**plot(States$teachpay~States$stuteach, xlab = "Average Student-Teacher Ratio", ylab = "Average Teacher Salary")**

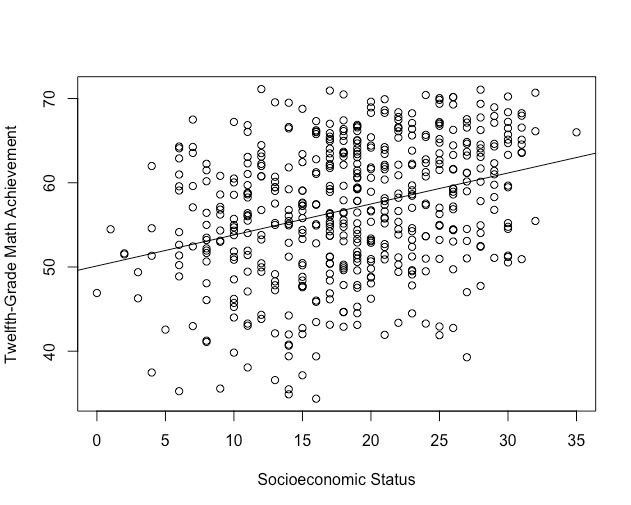
**abline(lm(States$teachpay~States$stuteach))**



1. Educational expenditure per pupil. It is most strongly correlated with teacher salary.
2. Because region is nominal with more than two categories, there is no inherent ordering of the variable, as a linear model requires.
3. Because both variables are interval, the level of measure is appropriate. Furthermore, the pattern of points in the scatterplot is sufficiently modeled by a line. The R commands to generate the scatterplot with regression line superimposed are:

**plot(NELS$achmat12~NELS$ses, xlab = "Socioeconomic Status", ylab = "Twelfth-Grade Math Achievement")**

**abline(lm(NELS$achmat12~NELS$ses))**

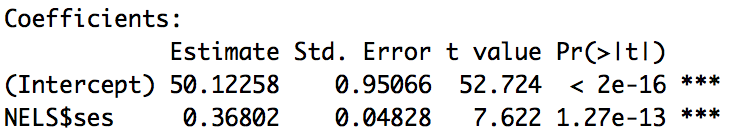


1. *=* 50.123 + .368(ses). The R commands to conduct the regression analysis and display the results are:

**model = lm(NELS$achmat12~NELS$ses)**

**summary(model)**

The part of the resulting output that is used to obtain the regression equation is displayed below.

**

1. Each one point increase in ses is associated with a .368-point increase in twelfth-grade math achievement, on average.
2. A person with ses = 0 is predicted to score 50.123 in twelfth-grade math achievement.
3. 57.483.
4. 59.69. The command to obtain the achmat12 score for the first student is **NELS$achmat12[NELS$id==1]**.
5. 58.59. To obtain this value, run the following R commands.

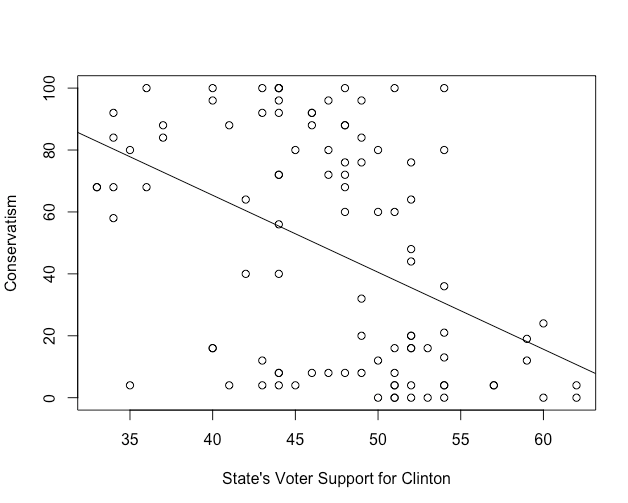
**yhat = predict(model)**

**yhat[NELS$id==1]**

1. Although the shape of the data is not exactly linear, there is no other simple curve that does a better job of approximating the data, so a linear model is appropriate. The R commands to generate the scatterplot with regression line superimposed are:

**plot(Impeach$conserva~Impeach$supportc, xlab = "State's Voter Support for Clinton", ylab = "Conservatism")**

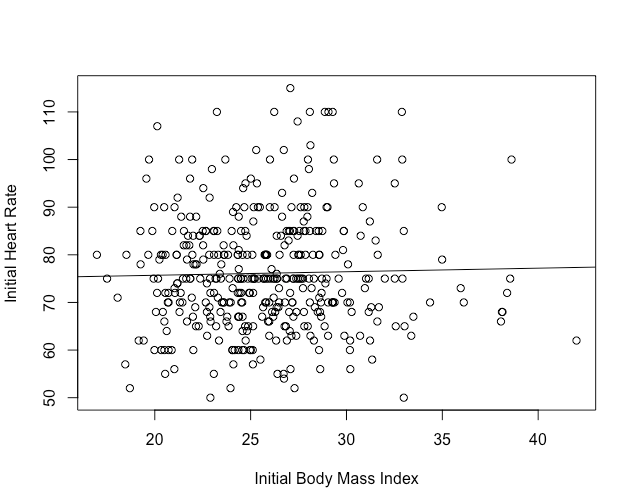
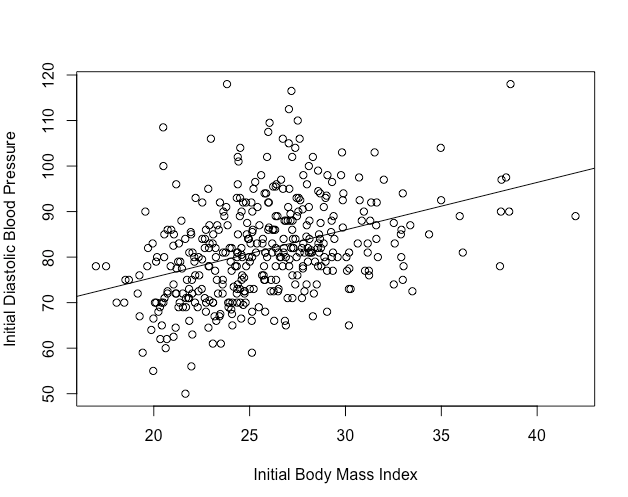
**abline(lm(Impeach$conserva~Impeach$supportc))**



1. The value of the correlation, *r* = -.447, can be found using a correlation analysis with the command **cor(Impeach$conserva,Impeach$supportc)**.
2. *R* = .447. That represents a moderate to strong goodness of fit. In simple linear regression, *R* = |*r*|, but we can also obtain this value by taking the positive square root of the *R*-squared value given in the regression output as “Multiple R-squared”: 0.1995.
3. = 164.911 - 2.488*(X)*.
4. Each one-percentage increase in the state voter support for Clinton is associated with a 2.488-point decrease in that state’s senator’s conservatism rating, on average.
5. Because there were no states with no voter support for Clinton, there was no data collected near supportc = 0 (the lowest support level was larger than 30), so the value of the intercept is not meaningful.
6. = 164.911 - 2.488(50) = 40.51.
7. Yes. The variables are both ratio-leveled and the data in the scatterplot are effectively modeled by a line. There are a few outliers that warrant further investigation, however.
8. The slope of the regression line is positive, indicating that the correlation is positive as well. Thus, adults with relatively high BMI tend also to have relatively high diastolic blood pressure.
9. Approximately 75 mmHg.
10. Because 50 is above the highest value of BMI measured in the dataset, it is inappropriate to extrapolate the model to that extreme.
11. The R commands used to generate the scatterplot for diastolic blood pressure is given below, along with both scatterplots.

**plot(Framingham$DIABP1~Framingham$BMI1, xlab="Initial Body Mass Index", ylab="Initial Diastolic Blood Pressure")**

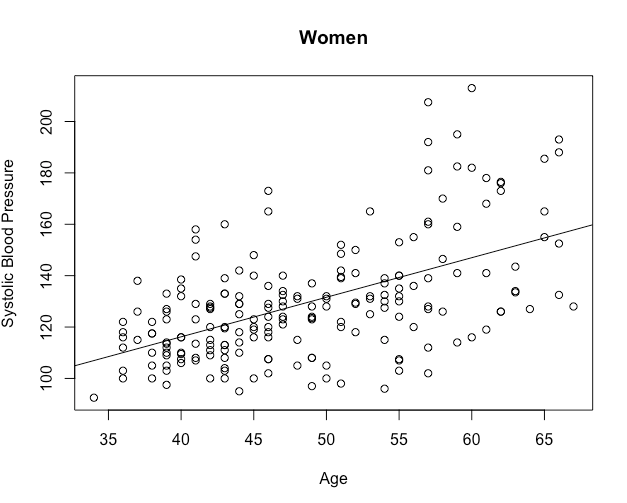
**abline(lm(Framingham$DIABP1~Framingham$BMI1))**

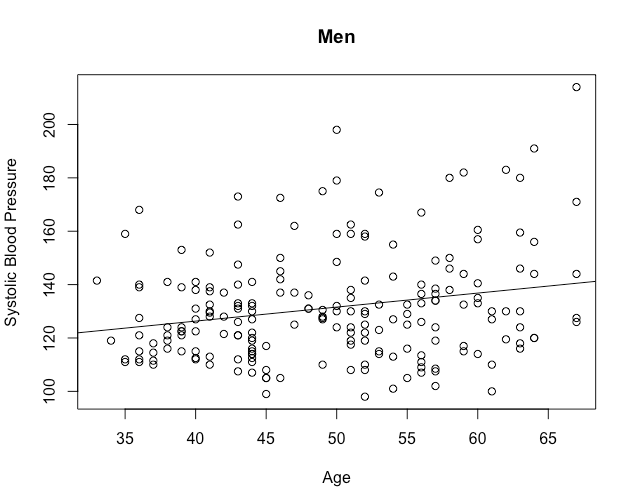


1. The relationship between body mass index and diastolic blood pressure will have the higher Pearson’s *r* value because the data points, overall, conform more closely to the regression line.
2. Because the slope of the regression line between BMI and blood pressure is steeper than that between BMI and heart rate, a one-unit increase in BMI is associated with a greater increase in diastolic blood pressure.
   1. The R commands below may be used to generate the scatterplot for women.

**plot(Framingham$SYSBP1[Framingham$SEX=="Women"]~Framingham$AGE1[Framingham$SEX=="Women"], xlab = "Age", ylab = "Systolic Blood Pressure", main = "Women")**

**abline(lm(Framingham$SYSBP1[Framingham$SEX=="Women"]~Framingham$AGE1[Framingham$SEX=="Women"]))**

****

****

* 1. The two regression equations are:

For women: = 54.587 + 1.540(AGE1)

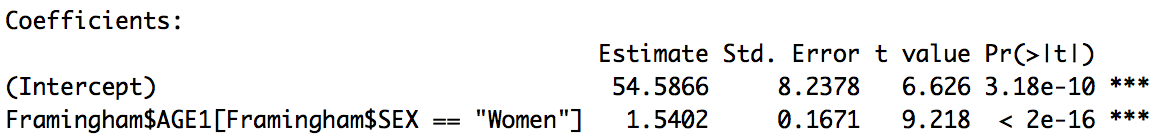
For men: = 105.301 + 0.526(AGE1)

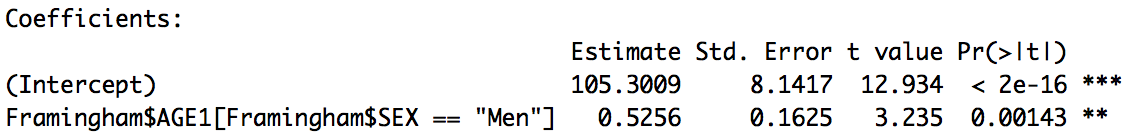
The R commands to obtain the regression output for women is:

**w\_model = lm(Framingham$SYSBP1[Framingham$SEX=="Women"] ~ Framingham$AGE1[Framingham$SEX=="Women"])**

**summary(w\_model)**

The relevant portions of the resulting outputs for each sex are shown below.

****

****

* 1. For men, a one-year increase in age is associated with a .526-point increase in systolic blood pressure, on average. For women, a one-year increase in age is associated with a 1.540-point increase in systolic blood pressure, on average. Blood pressure is increasing faster with age for women than it is for men.
  2. For women, *R* = .55, while for men, *R* = .22, indicating that the goodness of fit for the regression line for women is better than it is for men.
  3. For women, = = 54.587 + 1.540(50) = 131.587.

For men, = 105.301 + 0.526(50) = 131.601.

* 1. The following R code may be used to run the model and save the predicted and residual values to the dataset:

**model = lm(States$educexpe~States$teachpay)**

**States$yhat = predict(model)**

**States$res = model$residual**

1. $9,864.77, obtained with the commands **States$yhat[States$state=="Alabama"]** or **States$yhat[1]**
2. $8,597.00, obtained with the commands **States$educexpe[States$state=="Alabama"]** or **States$educexpe[1]**
3. $-1,267.77, by subtraction or with the command **States$res[States$state=="Alabama"]**
4. Over-predicts.
5. The R command to find the residual of the second state (Alaska) is **States$res[States$state=="Alaska"]** or **States$res[2]**. The larger residual (in magnitude) in the dataset is for the second state, not the first. The value of the residual for the first state is -1,267.77, while for the second, it is 3,190.09.
6. = 86.743 - 3.345(grade). Obtained using the R commands:

**model = lm(Learndis$readcomp~Learndis$grade)**

**summary(model)**

1. = 103.469 - 3.345(age).

Obtained using the R commands:

**Learndis$age = Learndis$grade + 5**

**model = lm(Learndis$readcomp~** **Learndis$age)**

**summary(model)**

1. The slope of the regression equation is given by 

 because the linear transformation from grade to age does not involve reflection. Also,  because the linear transformation from grade to age does not involve multiplication. Thus, the slope does not change.

1. The intercept of the regression equation with grade as the independent variable is given by



The mean of age =  + 5.

The intercept with age as the independent variable is



1. = 31.816 - .113(schattrt)
2. = 31.816 - 11.3(schattpp)

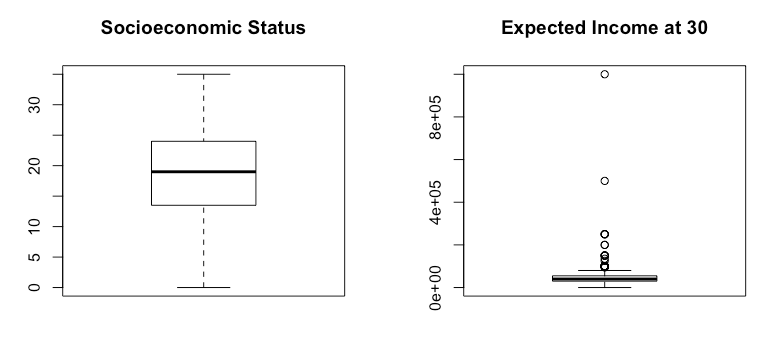
Obtained using the R commands:

**NELS$schattpp = NELS$schattrt/100**

**model = lm(NELS$slfcnc08~NELS$schattpp)**

**summary(model)**

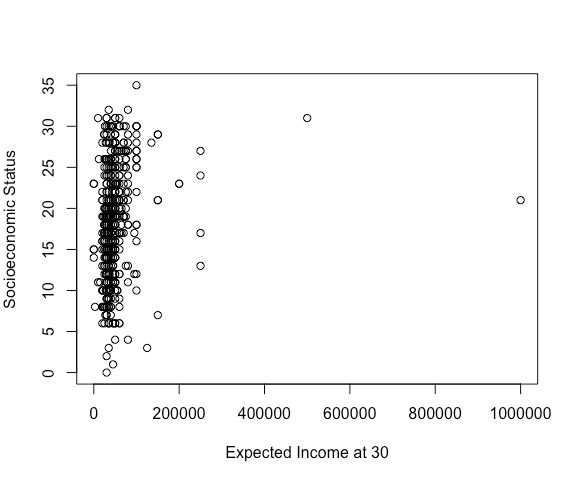
1. According to the boxplots, the distribution of socioeconomic status is fairly symmetric while that of expected income at age 30 is severely positively skewed. The R command used to generate the boxplot for ses is **boxplot(NELS$ses, main=”Socioeconomic Status”)**.



This visual impression is corroborated by the skewness ratios which indicate that the distribution of ses is fairly symmetric (-1.03) and that of expected income is severely positively skewed (95.63).

The R command used to calculate the skewness ratio for ses is **skew.ratio(NELS$ses)**.

The scatterplot shows that a linear model may not be most appropriate due to the presence of several outliers. The R command for creating the scatterplot is **plot(NELS$ses~NELS$expinc30, xlab="Expected Income at 30", ylab="Socioeconomic Status")**.



1. Prior to applying the non-linear tranformations, the variable was translated by adding 1 to all values to avoid taking a log of a zero value. The R commands for creating the transformed variables are:

**NELS$expinclg = log(NELS$expinc30 + 1)**

**NELS$expincsq = sqrt(NELS$expinc30 + 1)**

Although the transformed variable is still severely skewed, the square root transformation was the most effective at diminishing the severity of the positive skew. The log transformation overcorrected as it converted the positively skewed distribution to a negatively skewed distribution. The skewness statistics for each variable are given below.

expinc30:

skewness = 10.90; standard error of skewness = .11; skewness ratio = 95.63

expinclg:

skewness = -6.87; standard error of skewness = .11; skewness ratio = -60.26

expincsq:

skewness = 3.73; standard error of skewness = .11; skewness ratio = 32.71

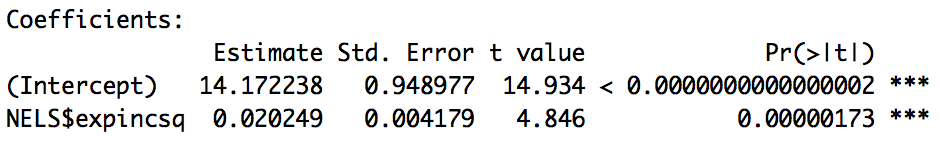
1. According to the results of the correlation analysis, the strongest correlation is with the square root transformed variable. The correlation of ses with expinc30 is obtained using the R command **cor(NELS$ses,NELS$expinc30, use="complete.obs")**.
2. Using the R commands

**model = lm(NELS$ses~NELS$expincsq)**

**summary(model)**

we obtain the following regression equation and results:

 OR 

****

1. 





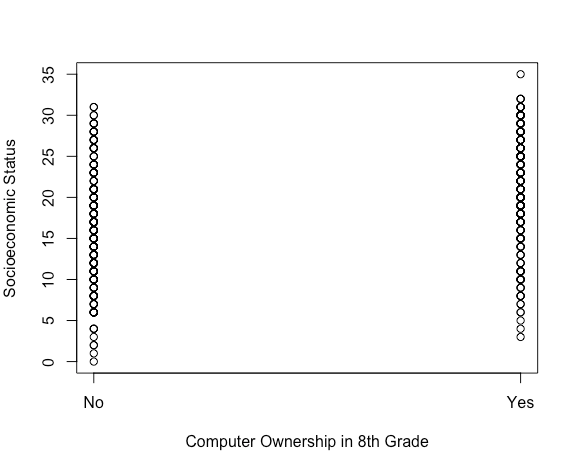
1. The R commands for creating the scatterplot are:

**plot(NELS$ses~as.numeric(NELS$computer), axes = F, xlab="Computer Ownership in 8th Grade", ylab="Socioeconomic Status")**

**box()**

**axis(2)**

**axis(1, at = c(1,2), labels = c("No", "Yes"))**

****

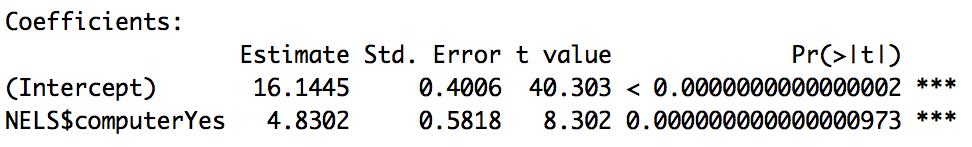
1. Using the R commands

**model = lm(NELS$ses~NELS$computer)**

**summary(model)**

we obtain the following regression equation and results:

= 16.14 + 4.83(computer)

****

1. Students in the NELS dataset who owned a computer in eighth grade have an average ses value that is 4.83 points higher than those who did not.
2. Students in the NELSdataset who did not own a computer in eighth grade are predicted by the model to have a mean ses of 16.14.
3. = 4.83(1) + 16.14 = 20.97
4. = 4.83(0) + 16.14 = 16.14
5. One possibility is to use the R commands:

**mean(NELS$ses[NELS$computer=="No"])**

**mean(NELS$ses[NELS$computer=="Yes"])**

The resulting outputs are the same as the results of parts (f) and (e), respectively.

1. Positive, same as it is in the regression equation.
2. Females.
3. = 4.059 + .573(2) = 5.21
4. = 4.059 + .573(1) = 4.63.
   1. b)
   2. a)
   3. d)
   4. b)
   5. a)
   6. b)
   7. a)
   8. a)
   9. Among adults, there is a positive correlation between weight and height. That is, adults that weigh more tend to be taller, on average. Because of this, an adult’s height may be predicted from his or her weight. However, increasing an adult’s weight will not cause his or her height to increase, as it would, if the relationship were causative.
   10. Group 1 = a

Group 2 = b

Group 3 = c

Group 4 = d

a) *rXX* = 1

In general, the correlation of a variable with itself is always 1.

b)   = .46

We have been calling, the correlation between the actual and predicted values for *Y*, *R*. We know that . This relationship holds because is a linear transformation of *X*, which involves reflection in this case because the correlation between *X* and *Y* is negative. By reflecting *X* (or, said differently, by multiplying *X* by -1), the correlation between *Y* and will retain the same magnitude as the correlation between *Y* and *X*, but will have a positive (as opposed to negative) sign.

c)   = -1

Because is a linear transformation of *X*, one that involves reflection in this case, the correlation between *X* and will have the same magnitude as the correlation between *X* and *X*, but with the opposite sign.

* 1. In order to show that the regression line always passes through the point (,), we show that when the value *X* =  is substituted into the regression equation given by Equation 6.2, , we obtain *=* .

Substituting Equation 6.4 for *a* and *X* =  in, we obtain .

By subtracting, we obtain *=* , as desired.

* 1. According to Equation 6.3, .

Substituting the expression for *r* given in Equation 5.3, we have

.

That expression is equivalent to .

* 1. The mean of the predicted values is given by . Thus, we need to show that = .

Substituting Equation 6.2 for , we have

= . Substituting Equation 6.4 for *a*, we have

= . Using the formula for the mean and the rules for

transforming the mean, we have

=  Simplifying yields the desired result:

=.

* 1.  *= a* + *bX*.

When *r* is positive, *b* is also positive, and = *a* + *bX* is a linear transformation of *X* that does not involve reflection. As we have seen, such a linear transformation preserves the sign and magnitude of the correlation. That is, the correlation between *X* and ** is the same as the correlation between *X* and *Y* (*R* = *r*).

On the other hand, when *r* is negative, *b* is also negative, and = *a* + *bX* is a linear transformation of *X* that does involve reflection. As we have seen, such a linear transformation changes the sign of the correlation. That is, the correlation between *X* and  is the negative of the correlation between *X* and *Y* (*R* = -*r*).