**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.



1. = .276(TEACHPAY) – 3372.988.



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. = .276(40000) – 3372.988 = $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. R = .72 indicating that the correlation between the actual and predicted educational expenditures is strong.
6. 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.





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.



1. *=* .368(SES) + 50.123.
2. Each one point increase in SES is associated with a .368 point increase in twelfth grade math achievement, on average.
3. A person with SES = 0 is predicted to score 50.123 in twelfth grade math achievement.
4. 57.483.
5. 59.69.
6. 58.59.
7. 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.



1. The value of the correlation, *r* = -.447, can be found using a correlation analysis, or using a simple regression analysis as the value of beta.
2. *R* = .447. That represents a moderate to strong goodness of fit.
3. = -2.488(SUPPORC) + 164.911.
4. Each one percent 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) and the value of the intercept is not meaningful.
6. = -2.488(50) + 164.911 = 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, 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 data set, it is inappropriate to extrapolate the model to that extreme.

 

1. The relationship between body mass index and diastolic blood pressure will have the higher Pearsons 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 init increase in BMI is associated with a greater increase in diastolic blood pressure.



* 1. Select cases so that PBMEDS1 = 0. Then, split the file by SEX and run the regression. The two regression equations are:

For men = .346(AGE1) + 112.431

For women = 1.388(AGE1) + 60.841

****

* 1. For men, a one year increase in age is associated with a .346 point increase in systolic blood pressure, on average. For women, a one year increase in age is associated with a 1.39 point increase in systolic blood pressure, on average. Blood pressure is increasing faster with age for women than it is for men.
	2. For men, R = .17, while for women, R = .52, indicating that the goodness of fit for the regression line for women is better than it is for men.
	3. For men = .346(50) + 112.431 = 129.73.

For women = 1.388(50) + 60.841 = 130.24.

1. $7125.70.
2. $6300.
3. $-825.70.
4. Over-predicts.
5. The residual with the largest magnitude corresponds to California with value -2945.79.
6. The largest positive residual is $2700.70 and corresponds to Vermont. The model under-predicts the educational expenditure for Vermont.
7. It is larger in magnitude for the first. The value of the residual for the first is -825.70 and for the second is 605.74.



* 1. The data does follow an approximately linear trend, albeit not a strong one. There is a bivariate outlier (case 205) and its effects on the model should be investigated.
	2. Approximately 220 mg/dL.
	3. Approximately130 mg/dL.
	4. Approximately -90 mg/dL.
	5. The most unusual person in the data set has ID 205. For this person, the total cholesterol is a lot higher than one would expect given the body mass index. Looking at the data view, we see that this person is female and is 42 years old at the start of the study.
	6. 1.82
	7. When omitting ID 205 the slope of the regression line increases slightly to 1.93.
1. = -3.345(GRADE) + 86.743.
2. = -3.345(AGE) + 103.468.
3. 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 reflection.

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. = -.113(SCHATTRT) + 31.816

****

****

1. = -.11.3(SCHATTRT) + 31.816
2. According to the boxplots, the distribution of socio-economic status is fairly symmetric while that of expected income at age 30 is severely positively skewed.



This visual impression is corroborated by the skewness ratios which indicate that the distribution of SES is fairly symmetric (-1.02) and that of expected income is severely positively skewed (95.91).

The scatterplot shows that a linear model may not be most appropriate due to the presence of several outliers.



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. 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.

****

1. According to the correlation analysis, the strongest correlation is with the square root transformed variable.

****

1.  OR 

****

1. 





1.

****

1. = 4.83(COMPUTER) + 16.144

****

1. Students in the *NELS* data set 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 *NELS* data set 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) + 16.144 = 20.97
4. = 4.83(0) + 16.144 = 16.14
5. One possibility is to use Means, Compare Means. That gives the following output:

****

1. = 4.83(COMP1) + 11.31
2. = -4.83(COMP2) + 20.97
3. Yes. The intercept is meaningful when one of the two levels of the dichotomous variable is coded as a 0. If neither is coded as 0, then the intercept is not meaningful.
4. Positive, same as it is in the regression equation.
5. Females.
6. = .573(2) + 4.059 = 5.21
7. = .573(1) + 4.059 = 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

To see this by example, use the *States* data set with X = STUTEACH and Y = EDUCEXPE. The correlation between X and Y is -.46. The scatterplot follows:



The scatterplot of X and X follows:



The pattern is a perfect line with positive slope, indicating that the correlation is 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, , we obtain *=* .

 Substituting Formula 6.4 for *a* in, we obtain .

 Subtracting we obtain *=* , as desired.

* 1. According to Formula 6.3, .

Substituting the expression for r given in Formula 5.5, we have

.

That expression is equivalent to.

* 1. The mean of the predicted values is given by . Thus, we need to show that = . Substituting Formula 6.2 for , we have

= . Substituting Formula 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.  *= b*X + *a*.

When r is positive, b is also positive, and = *b*X + *a* 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 = *b*X + *a* 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*).

a) In order to use Equation 6.1, we need to find the residuals, or *di*’s.

To use SPSS to find the residuals we use the Regression procedure. Go to **Statistics** on the main menu bar, **Regression**, **Linear**. Put CALORIES in the box for the Dependent variable and FAT in the box for the Independent variable. Click Save and in the box labeled Residuals, click the box next to Unstandardized. Click **Continue** and **OK**. In the data window, you will see that a new variable, RES\_1 has been created which gives, for each hamburger the corresponding residual.

These values are provided in Table 6.2.

 Table 6.2 Residuals for the McDonald’s example.

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Fat | Calories | Res\_1 |
| Hamburger | 10 | 270 | -.67513 |
| Cheeseburger | 14 | 320 | -5.51471 |
| Quarter Pounder™ | 21 | 430 | 8.51604 |
| Quarter Pounder with Cheese™ | 30 | 530 | -14.87299 |
| Big Mac™ | 28 | 530 | 12.54679 |

To calculate D using Equation 6.1, we may use one of two procedures:

1) Use the Descriptives procedure to calculate D as the standard deviation of the residuals. The value of *SDRES* obtained through SPSS is 10.98. Recall that this value results from division by 4 instead of 5. We approximate D by 

2) Use the SPSS Compute procedure to find the values of , the squared residuals.

To use the Compute procedure to find the squared residuals, go to the SPSS Data Editor. Go to **Transform**, **Compute**. Type squares in the Target Variable box. Type res\_1\*\*2 in the Numeric Expression box. (\*\* means “to the power”.) Click **OK**.

To find the sum of these squared residuals, use Options in the Descriptives procedure to obtain the value 482.02. Finally, to obtain D from the sum of squared residuals, use a calculator to divide the sum of squared residuals by N and take the square root. You should obtain the value D = 9.82.

b) To use formula 6.5 to calculate D, we use the SPSS Descriptives and Correlation. From Descriptives we obtain that the variance of and from Correlation, we obtain that rXY = .996. Substituting these values into formula 6.5 we obtain 

c) To find D directly on the SPSS Regression output, perform the regression as in Example 6.1. The relevant SPSS output is reproduced below:



The value of D is labeled “Std. Error of the Estimate” and equals 12.68. Again, for reasons that will become clear when we discuss inferential statistics, SPSS uses a different denominator. In particular, SPSS uses *N*-2 instead of *N*. To adjust the value of the standard error of estimate to conform to our own descriptive measure, we take .

d) All obtained descriptive measures of D are equal, regardless of how they were obtained. Although, in this small *N* example, the adjustments changed the value of D considerably, when *N* is large, the adjustments can be expected to make only minor changes in the value of D. Hence, when *N* is large, all methods can be expected to produce similar results without replacement.