**CHAPTER 15 SOLUTIONS**

1. Let **the correlation between reading comprehension and intellectual ability in the population. Then H0: ** 0 and H1: ** 0
2. There is a statistically significant, moderate, positive linear relationship between reading comprehension and intellectual ability among public school children in the urban area who have been diagnosed with learning disabilities, *r*(*N* = 76) = .29, *p* = .01. That is, such children who have relatively low intecllectual ability tend to have relatively low reading comprehension and those with relatively high intellectual ability tend to have relatively high reading comprehension. The strength of the linear relationship is moderate, according to Cohen’s rule of thumb guidelines.
3. There is a statistically significant moderate negative linear relationship between reading comprehension grade level among public school children in the urban area who have been diagnosed with learning disabilities, *r*(*N* = 76) = -.322, *p* = .004. That is, such children who have relatively low reading comprehension tend to be in the higher grades, while those with relatively high reading comprehension tend to be in the lower grades. Because the reading comprehension scores are relative to all students in the urban area, not just those with learning disabilities, these results indicate that the students with learning disabilities fall further behind their peers as they get older. The strength of the linear relationship is moderate, according to Cohen’s rule of thumb guidelines.
4. There is a statistically significant moderate correlation between reading comprehension and classroom placement among public school children in the urban area who have been diagnosed with learning disabilities, *r*(*N* = 76) = -.44, *p* < .0005. Students with the part-time resource placement have higher reading comprehension scores, on average, than those with the self-contained classroom placement. The strength of the linear relationship is moderate, according to Cohen’s rule of thumb guidelines.
5. The elimination of the 29 students without reading comprehension scores could bias obtained results if their missing data were related systematically to reading comprehension or any of the variables studied in relation to reading comprehension. For example, if the 29 students who did not have reading comprehension scores missed the reading comprehension test because they were doing especially poorly in reading and were being tutored at the time the reading test was given, the obtained correlations of reading comprehension with intellectual ability, grade level, and class placement would likely be biased.
6. (0.056, 0.435)
7. (0.067, 0.56)
8. Yes. Because 0 is not contained in the 95 percent confidence interval for *b*, it is not a plausible value for *b*. Therefore, the coefficient is statistically significant.
9. Because the range of values for both the confidence interval of ** and the population *b* are positive, the linear relationship between math comprehension and intellectual ability is positive. That is, public school children in the urban area who have been diagnosed with learning disabilities who have low math comprehension tend also to have low intellectual ability while those with high math comprehension tend also to have high reading comprehension. Because the slope in the population is likely to be between 0.067 and 0.56, a one point increase in intellectual ability is likely to be associated with a 0.067 to 0.56 point increase in the predicted math comprehension in the population.
10. Because the value of ** is thought to be between .06 and .44, we conclude that the strength of the correlation is anywhere from very weak to moderate.
11. Looking at scatterplot the relationship between these two variables appears to be linear, so a correlation analysis is appropriate in this case.



1. Using the R command **cor.test(NELS$famsize[NELS$urban == "Urban"], NELS$slfcnc12[NELS$urban == "Urban"])**, we see that there is a significant correlation between family size and 12th grade self-concept

(*r*(*N* = 123) = -0.20, *p* = 0.02). The negative correlation indicates that a larger family size tends to associate with a lower self-concept among 12th graders in urban settings.

1. The 95 percent confidence interval for ** is (-0.37, -0.03). Because 0 is not contained in the confidence interval, we know that the correlation between family size and self-concept in eighth grade is statistically significant.
2. In this case, the sample correlation value is *r* = -.20, which is a small to moderate effect size. The confidence interval shows that the correlation in the population could be anything from a moderate negative correlation to a small negative correlation.
3. The regression model and coefficient are both statistically significant. The results of the ANOVA *F*(1,498) = 58.10, *p* < .0005 and the test of the significance of the *b*-weight, *t*(498) = 7.62, *p* < .0005.
4.  = 50.123 + 0.368 (X)
5. The value of the slope of the regression equation indicates that for each one point increase is socioeconomic status is associated, on average, with a .368 point increase in the predicted twelfth grade math achievement.
6. Given that an ses value of 0 is meaningful on this scale used to measure ses in the NELS study, the value of the y-intercept indicates that a person with an ses score of 0 is predicted to have a twelfth grade math achievement score of 50.123.
7. The value of *R*2 = adjusted *R*2 = .10, which is an effect size measure indicating the proportion of twelfth grade math achievement variance that can be explained by socioeconomic status, a moderate effect.
8. The result of the significance test associated with the *b*-weight, *t(103)* = -0.11, *p* = .92 indicates that the regression model is not statistically significant.
9. Because the b-weight is not statistically significant, gender is not useful for predicting intellectual ability in the population.
10. The b-weight is statistically significant. The result of the significance test associated with the *b*-weight, *t*(74) = -4.27, *p* < .0005.
11. = 81.434 - 12.564(*X*)
12. The slope is the difference between the average reading comprehension score of students in the full-time self-contained placement (placemen = 1) and those in the resource room part-time (placemen = 0). That is, according to this analysis, those in the self-contained classroom are predicted to score 12.56 points lower, on average, than those in the resource room.
13. Because the value of 0 on the placement variable is meaningful (placemen = 0 represents students in the part-time resource room), we may interpret the *y*-intercept as the average reading comprehension score of resource room students
14. The predicted reading comprehension score for a student with a resource room placement is = -12.564(0) + 81.434 = 81.434. The predicted value is equal to the mean reading comprehension score for all students with a resource room placement.
15. The predicted reading comprehension score for a student with a self-contained classroom placement is = -12.564(1) + 81.434 = 68.87.
16. A good approximation is given by *R*2. Thus the proportion of the variance in reading comprehension scores that is explained by placement type is 0.20, which may be considered a moderate effect.
17. Equivalently, an independent samples *t*-test or one-way ANOVA could have been used to determine whether reading comprehension may be predicted from type of placement.
18. The regression model and b-weight are both statistically significant. The results of the ANOVA *F*(1,498) = 15.79, *p* < .0005 and the test of the significance of the *b*-weight, *t*(498) = 3.97, *p* < .0005.
19.  = 55.596 + 2.765(*X*)
20. Students whose families owned a computer when they were in eighth grade are predicted to score 2.765 points higher, on average, in twelfth grade math achievement, than those whose families did not own a computer.
21. The value of the y-intercept is obtained when the variable, computer ownership, equals zero, which represents those students whose families did not own a computer in eighth grade. Accordingly, we may interpret the y-intercept of 55.596 as the average twelfth grade math achievement of those students whose families did not own a computer in eighth grade.
22. The value *R*2 = .03 equals the proportion of twelfth grade math achievement scores variance that can be explained by computer ownership, which may be described as a small to moderate effect. Because the relationship between twelfth grade math achievement and computer ownership is a comparison of those who did and did not own computers in eighth grade in terms of average twelfth grade math achievement, Cohen’s *d* may be calculated as another measure of effect size. To perform the calculation, we need means and standard deviations of twelfth grade math achievement for the two computer ownership groups. According to this calculation, students who owned a computer in eighth grade performed approximately .36 standard deviations higher in twelfth grade math achievement than those who did not own a computer in eighth grade, which may be considered a small to moderate effect.



1. There is a statistically significant, moderate, positive correlation between intelligence and brain size, *r*(*N* = 40) = .36, *p* = .02. That is, among students with extremely high or extremely low intelligence, those with relatively low intelligence tend to have relatively small brain size and those with relatively high intelligence tend to have relatively large brain size.
2. The relationship between intelligence and gender among students with extremely high or extremely low intelligence, *r*(*N* = 40) = -.07, *p* = .69 is not statistically significantly different from zero.
3. Recall that in Chapter 5 it was mentioned that a correlation (calculated on a sample in which, at least, one of the variables contains only extreme values) tends to be inflated relative to the correlation value calculated on the entire distribution of values (i.e., including the middle values as well). Accordingly, one must take care not to generalize the finding in part (a) to all college students (those with low, middle, and high intelligence) as the correlation in that group is likely to be weaker than what was calculated in part (a).
4. The R commands used to generate the scatterplot are:

**plot(Brainsz$MRI ~ Brainsz$FSIQ, ylab = "Brain Size", xlab = "IQ", pch = 16)**

**abline(lm(Brainsz$MRI~Brainsz$FSIQ))**



1. The two clouds of points represent the two extreme groups that form this sample of data, those with extremely low intelligence and those with extremely high intelligence. There are no points in the middle because there is no one in this sample with moderate intelligence.
2. The regression model, = 5.168 + .0001192(*X*), is statistically significant. The results of the ANOVA *F*(1,38) = 5.57, *p* = .02 and the test of the significance of the *b*-weight, *t*(38) = 2.36, *p* = .02. The model fit as measured by *R*2 equals .13, indicating that 13 percent of Full Scale IQ variance is explained by brain size.
3. *R*2single model = .13. *R*2low = .28. *R*2high = .30

The separate regression models provide a better fit to the data overall than the single regression equation fit to the entire sample. In short, we were able to take advantage of the existence of the two separate clouds of points in the scatterplot to produce two models which fit the data set better overall.

* 1. The R commands to generate the scatterplot are:

**plot(c(0), type = "n", xlim = range(Learndis$grade), ylim = range(na.omit(Learndis$readcomp)), xlab = "Grade", ylab = "Reading Comprehension")**

**text(Learndis$readcomp ~ Learndis$grade, labels = row.names(Learndis), pch = 0)**

**abline(lm(Learndis$readcomp ~ Learndis$grade))**

****

1. With the exception perhaps of the homoscedasticity assumption, the scatterplot does not appear to suggest a violation of any of these underlying assumptions. Although there appears to be a non-constant variance of reading comprehension scores by grade (notice how tightly clustered the grade 1 distribution is relative to the grade 3 or grade 4 distribution, for example), the sample size is relatively small, and these variations simply may be due to chance. Nonetheless, reasons ought to be explored as to why grade 1 scores appear to cluster so much more tightly relative to the other grades. Once the regression model is developed, sensitivity analyses should be carried out to measure the extent to which the outliers in grade 4 influence results.
2. For case number 32 (in grade 4), the residual is negative and the case is over-predicted.
3. An extreme point in either the first or fifth grade would have relatively large leverage. One example is case number 39 in grade 5.  
   1. In order to evaluate the appropriateness of the fit of the simple linear regression model to the data and to explore possible violations of assumptions, a series of analyses was carried out using the unstandardized residuals. Given the difficulty of judging the normality of a outcome variable when the regressor is dichotomous (e.g., computer ownership groups) using a scatterplot (see below), boxplots of the residuals (unstandardized) were constructed. The boxplots of the residuals suggest that even with the outliers in the group that did not own a computer, both distributions appear to be only mildly negatively skewed. In addition to these residual analyses, an analysis of Cook’s influence was carried out. Note that although there are some outliers, all of the Cook’s scores are considerably less than one, suggesting that there were no individual points in the analysis that appeared to unduly influence the results and that the regression analysis was appropriate as reported.

The R commands to generate the graphs are:

**results = lm(NELS$achmat12 ~ NELS$computer)**

**plot(results, which = 1)**

**boxplot(results$residuals~NELS$computer)**

**boxplot(cooks.distance(results))**

****





1. Yes, the points of the scatterplot do not appear to violate the underlying assumptions of normality, homoscedasticity, and linearity.

The R commands used to generate the scatterplot are:

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

**plot(NELS$achmat12 ~ NELS$ses, xlab = "SES", ylab = "Math Achievement")**

**abline(results)**



1. The residual plot (ses values versus Studentized residuals) is a random scatter of points organized in a rectangular fashion about a residual value of zero for each value of SES and suggests an appropriate fit of the model to these data overall. In particular, since the variability of points for each value of SES is similar, the data appear to satisfy homoscedasticity and with no pattern of relationship between the residuals and SES, the assumption of linearity also appears to be met.

The R commands used to generate the following output are:

**plot(rstudent(results)~NELS$ses, xlab = "SES",ylab = "Studentized Residuals")**

**abline(h = 0)**



1. According to the boxplot of Cook’s influence scores, all of the values are substantially less than 1, suggesting that no single point unduly influenced the result the regression analysis and that the model is appropriate as obtained.

The R command used to generate the boxplot is:

**boxplot(cooks.distance(results))**



1. 58.59
2. 1.10
3. The value of the residual with the greatest magnitude is -21.65; it is associated with the person that has ID = 82. But, as shown in the plot of Cook’s influence scores, even this person’s data does not unduly influence the results of the analysis. That is, if this person’s data were removed from the analysis, and a new model were fit to the data (without this person), results would not change by very much.  
     
   The R command to obtain this information is:  
   **results$residuals[which.max(abs(results$residuals))]**
4. Using the R command **pwr.r.test(r = 0.1, sig.level = 0.05, power = 0.8, alternative = "greater")**, we determine that we need *N* = 617.
5. Using the R command **pwr.r.test(r = 0.1, sig.level = 0.05, power = 0.8, alternative = "two.sided")**, we determine that we need *N* = 782.
6. *R2* = .0196 corresponds to *f2* = .02?
7. Using the R command **pwr.f2.test(u = 1, f2 = 0.02, sig.level = 0.05, power = 0.8)**, we determine that we need N = 395.
8. The scatterplot was created with the case numbers included for each point. Because there are so few points, it is difficult to assess the normality and homoscedasticity assumptions. The data do appear linear, so the linearity assumption appears viable. Case number 8 is an outlier with a studentized residual value of -2.04. Cook’s influence scores were found for each point. The most influential point is case number 5. Although it is much more influential than the others, its Cook’s influence value is much less than 1 (.51), so that we do not need to be concerned about its undue influence on the analysis.

The R commands to generate the scatterplot are:

**plot(c(0), type = "n", xlim = range(data$mother), ylim = range(data$infant), xlab = "Mother Weight Gain", ylab = "Infant Weight")**

**text(data$infant ~ data$mother, labels = data$id, pch = 0)**

**abline(lm(data$infant ~ data$mother))**



1. The 95 percent CI for ** is (.72, .97). There is a statistically significant, strong, positive correlation between the weight gain of the mother and the birth weight of the infant. That is, such mothers with a relatively high weight gain tended to have infants with relatively high birth weight and mothers with a relatively low weight gain tended to have infants with relatively low birth weight.
2. =4.278 + .113(X)
3. The regression model is statistically significant. The results of the ANOVA *F*(1,13) = 54.96, *p* < .0005 and the test of the significance of the *b*-weight, *t*(13) = 11.24, *p* < .0005 both indicate equivalently that the model is significant.
4. 7.1lbs.
   1. c)
   2. a)
   3. e)
   4. b)
   5. c)
   6. c)
   7. a)
   8. b)
   9. c)
   10. a)
   11. b)
   12. c)
   13. e)
   14. e)
   15. b)

 = 