**CHAPTER 16 SOLUTIONS**

1. Boxplots and skewness ratios of the two scale variables, achmat12 and unitmath, indicate that these variables are negatively skewed with skewness values less than -1.00 in both cases. The skewness ratios for achmat12 and unitmath are respectively -3.86 and -2.26. These values are moderately negative due to the large sample size of 500 and the corresponding relatively small standard error of skewness of 0.109. Given that the NELS dataset contains approximately 55% females, we know that gender is not skewed, and, is in fact, reasonably symmetric.

The R commands used to generate the univariate statistics and plots are:

**boxplot(NELS$achmat12, main = "Twelfth-Grade Math Achievement")**

**boxplot(NELS$unitmath, main = "Units of Math Taken")**

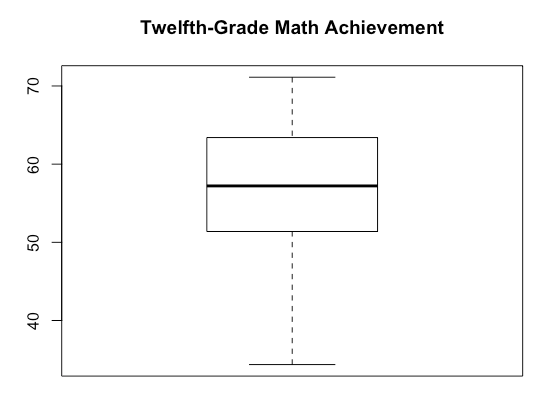
**skew.ratio(NELS$achmat12)**

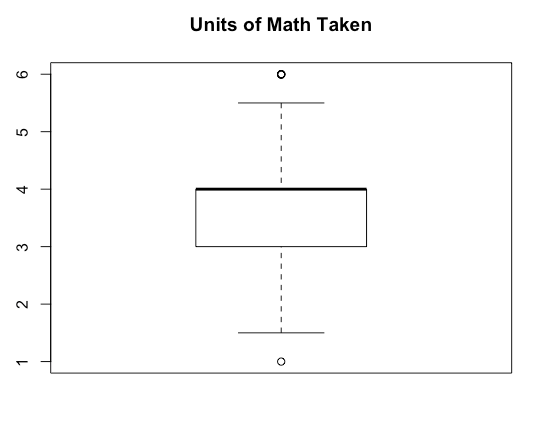
**skew.ratio(NELS$unitmath)**

**se.skew(NELS$achmat12)**

**se.skew(NELS$unitmath)**

**percent.table(NELS$gender)**

****



The scatterplot of the scale variables and the correlation values of each of the predictor variables with achmat12 (*r* = 0.42 with unitmath and *r* = -0.20 with gender) are both statistically significant (*p* < 0.0005) and suggest the appropriateness of fitting these data with a regression model. According to the zero-order correlations with achmat12, college-bound students who are always at grade level who take more math in high school tend to do better on twelfth-grade math achievement tests; among college-bound students who are always at grade level, males outperform females on twelfth-grade math achievement tests. Furthermore, because there is little or no relationship between the predictor variables (the correlation between unitmath and gender is *r* = -0.07, *p* = 0.13), there is little overlap in their proportion of shared variance and, therefore, both variables should contribute uniquely to the model.

The R commands used to generate the bivariate statistics and plots are:

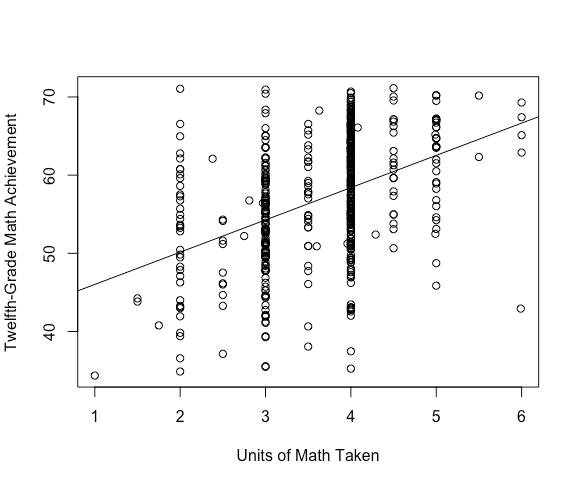
**plot(NELS$achmat12~NELS$unitmath, xlab = "Units of Math Taken", ylab = "Twelfth-Grade Math Achievement")**

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

**cor.test(NELS$achmat12,NELS$unitmath)**

**cor.test(NELS$achmat12,as.numeric(NELS$gender))**

**cor.test(NELS$unitmath,as.numeric(NELS$gender))**

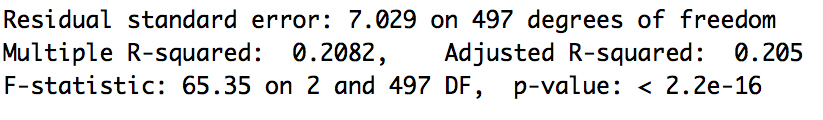


1. According to the R output, the regression model with both first-order variables included in the equation is statistically significant, *F*(2, 497) = 65.35, *p* < 0.005.

The R commands to generate the output is:

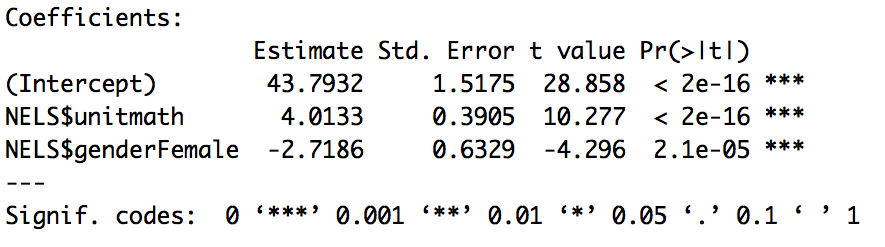
**results = lm(NELS$achmat12 ~ NELS$unitmath + NELS$gender)**

**summary(results)**



1. According to the value of *R2*, approximately 20.8% of the variance in twelfth-grade math achievement can be explained by gender and the number of units of math taken in high school. Because the ratio of sample size to predictor variables is so large (500 to 2), the *R2* and adjusted *R2* values are quite similar.
2. Predicted achmat12 = 43.793 – 2.719(gender) + 4.013(unitmath).

The relevant portion of the regression output is shown below.



1. Predicted achmat12 = 43.793 – 2.719(0) + 4.013(4) = 59.85.
2. Based on the relative magnitudes of the beta weights, unitmath is the more important variable in the equation.

The R commands to run the standardized regression are:

**results2 = lm(scale(NELS$achmat12) ~ scale(NELS$unitmath) + scale(as.numeric(NELS$gender)) -1)**

**summary(results2)**

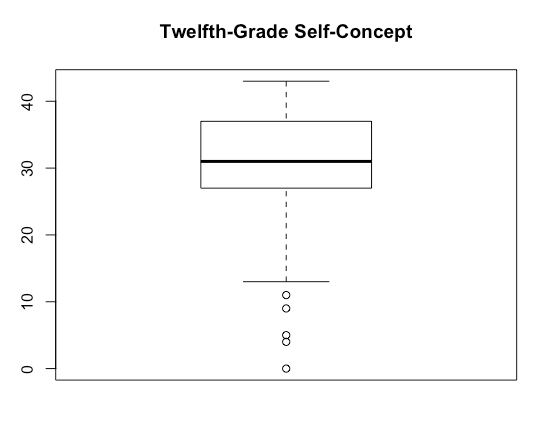
An analysis of the respective unique proportions of variance accounted for by each predictor variable yields the same result. The squared semipartial correlations indicate that the unique proportion of variance accounted for by gender is only 2.9%, while for unitmath it is 16.8%, indicating that unitmath makes a greater unique contribution to explaining achmat12 variance and is therefore the more important variable in the equation.

The R commands to obtain the semipartial correlations are:

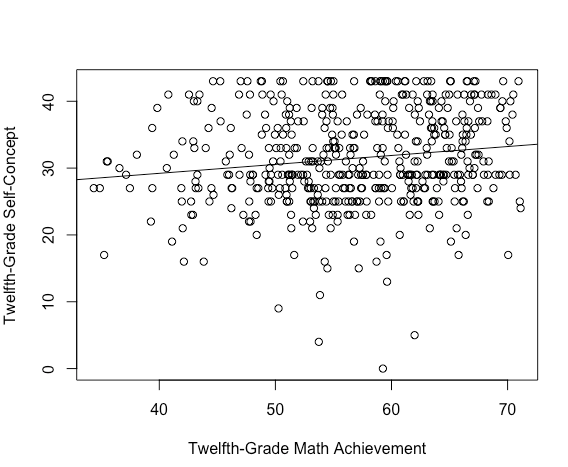
**spcor.test(NELS$achmat12,NELS$unitmath,as.numeric(NELS$gender))**

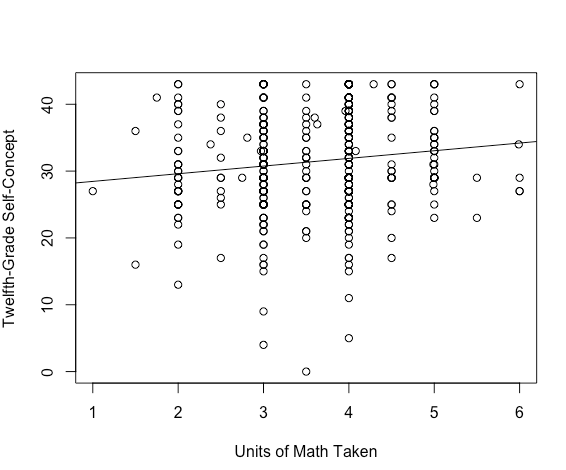
**spcor.test(NELS$achmat12,as.numeric(NELS$gender),NELS$unitmath)**

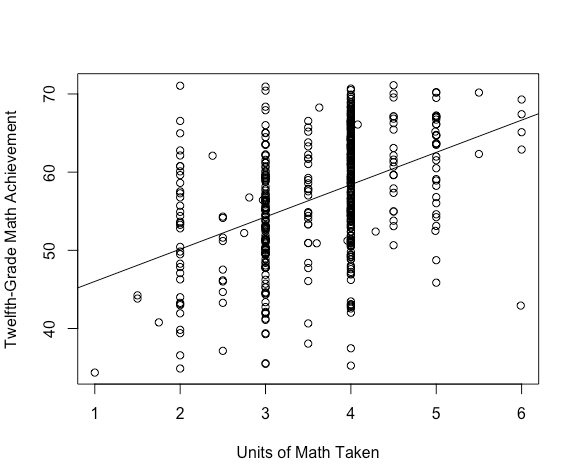
1. It would not be appropriate to interpret the value of the *y*-intercept in this case as the smallest number of units of math taken in high school is 1 by those in our sample. We have no reason to believe that the model would generalize to values beyond those in our sample: namely, unitmath = 0.
2. Holding gender constant, each additional unit of math taken in high school corresponds to a 4.01-point increase in the predicted twelfth-grade math achievement, on average.
3. Holding units of math taken in high school constant, females perform 2.72 points lower in twelfth-grade math achievement, on average, than males.
4. Yes. From the regression output, we see the coefficient or slope associated with gender is statistically significant, *t*(497) = -4.30, *p* < .0005.
5. Boxplots of the three variables indicate that all variables are negatively skewed. Summary statistics indicate that the skewness values are -0.422 for achmat12, -0.247 for unitmath, and -0.383 for slfcnc12. Because of the large sample size of 500, skewness ratios all exceed two standard deviations away from a skewness value of zero. The boxplots for achmat12 and unitmath are provided in the solution to 16.1(a), and the boxplot for slfcnc12 is shown below.



An investigation of the bivariate scatterplots between the outcome variable and each of the predictor variables and between the two predictor variables indicates that even though at least one of the variables may be considered moderately negatively skewed, the relationships appear to be linear as opposed to curvilinear.

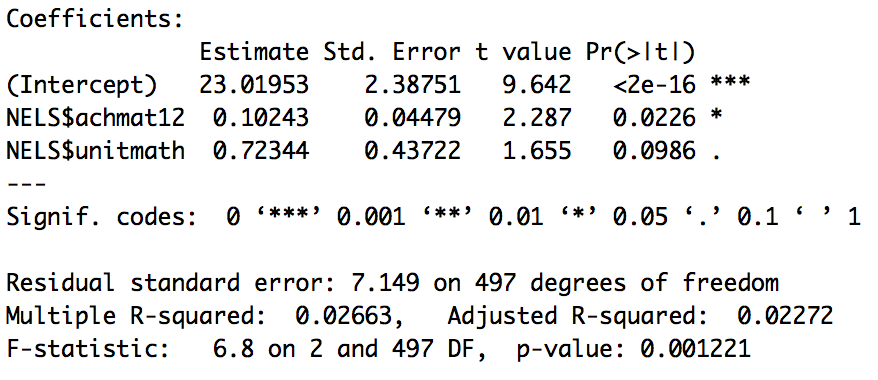






The bivariate correlations between pairs of predictor and outcome variables and between the two predictor variables are as follows: unitmath and slfcnc12, *r* = .128; achmat12 and slfcnc12, *r* = .146; unitmath and achmat12, *r* = .423. All correlations are statistically significant.

1. According to the R output, the regression model is statistically significant, *F*(2, 497) = 6.800, *p* < 0.005 with both predictor variables in the equation.



1. According to the value of *R2*, approximately only 2.7% of the variance in twelfth-grade self-concept is explained by both achievement in math in twelfth grade and units of math taken in high school.
2. Only achmat12 is statistically significant with *t* = 2.287, *p* < 0.025. That is, after controlling for units of math taken in high school, achievement in math in twelfth grade accounts for a statistically significant (albeit small) proportion of twelfth-grade self-concept variance. Notice that unitmath is not statistically significant in the equation although it was statistically significantly related to slfcnc12 in the bivariate relationship. That it is not statistically significant in the equation suggests that once achmat12 is controlled, that part of unitmath that remains does not correlate significantly with slfcnc12. That is, the part correlation between unitmath and slfcnc12, after removing from unitmath that part related to achmat12, is not statistically significantly related to slfcnc12.
3. According to the equation, the predicted twelfth grade self-concept score of an individual who takes zero units of math in high school and who scores zero on twelfth grade math achievement is 23.02, on average. However, since no one in the sample has taken zero units of math (the minimum number of units taken is 1), or scored zero on the math achievement test, we cannot interpret the y-intercept meaningfully.
4. Holding the number of math units taken in high school constant, each additional one-point increase in twelfth-grade math achievement corresponds to a 0.1-point increase in the predicted twelfth-grade self-concept, on average.
5. The correlation between achrdg12 and ses is statistically significant, *r* = 0.34, *p* < 0.0005. Approximately 11.6% of the variance in achrdg12 can be explained by ses (*R2* = 0.3412 = 0.116).
6. The correlation between achrdg12 and achmat12 is statistically significant, *r* = 0.64, *p* < 0.0005. Approximately 40.5% of the variance in achrdg12 can be explained by achmat12 (*R2* = 0.6362 = 0.405).
7. The correlation between ses and achmat12 is statistically significant (*r* = 0.32, *p* < 0.0005).
8. Approximately 30.9% of the variance in achrdg12 can be explained by achmat12 after controlling for ses (*R2Change* = *R2ses, achmat12* – *R2ses* = 0.4254 – 0.1164 = 0.3090), a statistically significant amount, *FChange*(1, 497) = 267.19, *p* < 0.0005.

The R commands to obtain the values of *R*2 and test the statistical significance of the change in *R*2 are shown below.

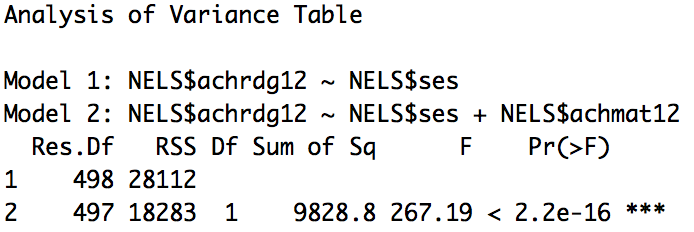
**results = lm(NELS$achrdg12 ~ NELS$ses)**

**results2 = lm(NELS$achrdg12 ~ NELS$ses + NELS$achmat12)**

**summary(results)**

**summary(results2)**

**anova(results, results2)**



1. Approximately 2.1% of the variance in achrdg12 can be explained by ses controlling for achmat12 (*R2Change* = *R2achmat12, ses* – *R2achmat12* = 0.4254 – 0.4048 = 0.0206), which, despite being small, is a statistically significant amount, *FChange*(1, 497) = 17.75, *p* < 0.0005.

The R commands to obtain the values of *R*2 and test the statistical significance of the change in *R*2 are shown below.

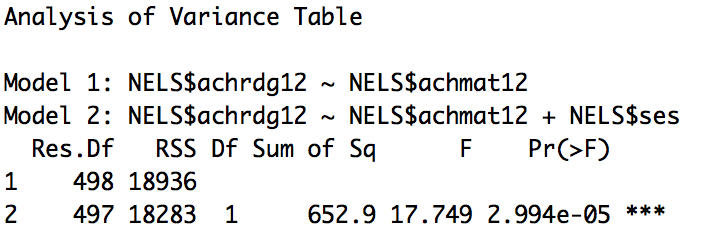
**results = lm(NELS$achrdg12 ~ NELS$achmat12)**

**results2 = lm(NELS$achrdg12 ~ NELS$achmat12 + NELS$ses)**

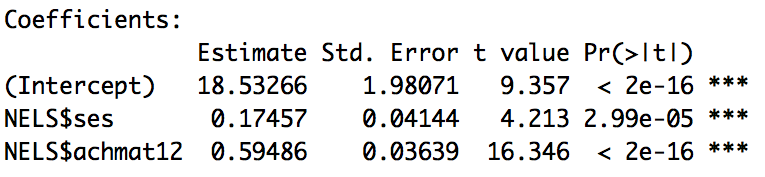
**summary(results)**

**summary(results2)**

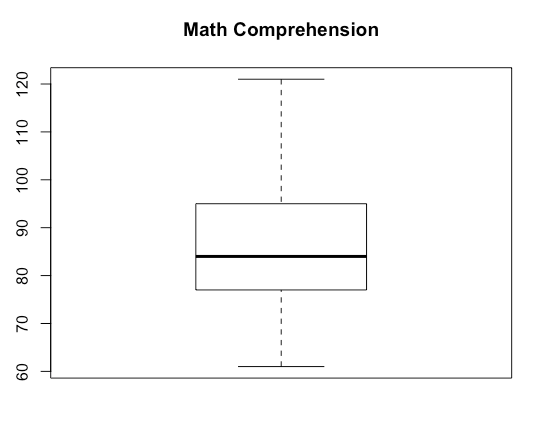
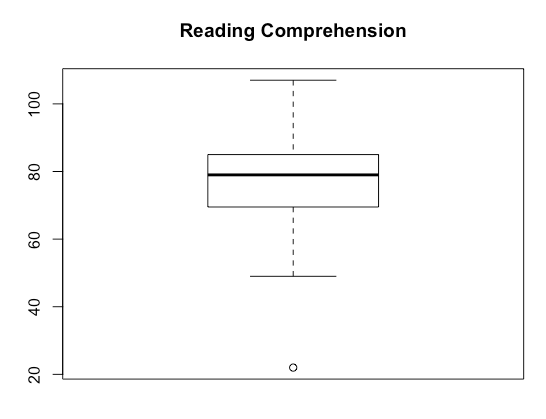
**anova(results, results2)**



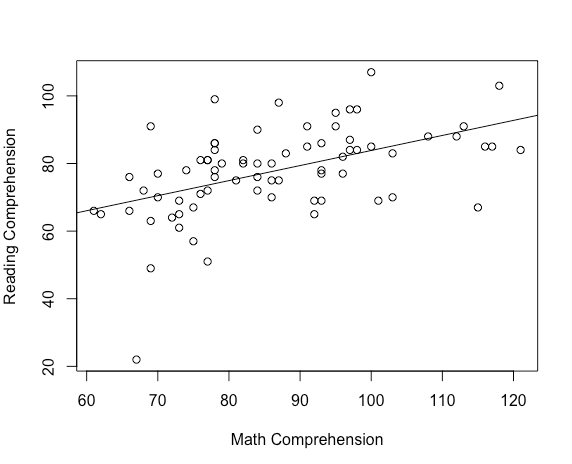
1. Using simultaneous multiple regression and placing both predictor variables in the equation at the same time, we see that the proportion of variance in achrdg12 that can be explained by ses after controlling for achmat12 is statistically significant *t*(497) = 4.213, *p* < 0.0005.

****

1. Approximately 11.6% of the variance in achrdg12 can be explained by ses, but only about 2.1% of the variance in readcomp can be explained by ses after controlling for achmat12, a much smaller percentage. This is because achmat12 and ses are intercorrelated, *r* = 0.32, *p* < 0.0005.
2. *R2* for the model is 0.425 which is less than *r2*achrdg12, ses + *r2*achrdg12, achmat12 = 0.116 + 0.405 = 0.521. That is because *R2* for the model only counts the overlap of approximately 0.104 between ses and achmat12 (that is, the squared intercorrelation between ses and achmat12) once.
3. Boxplots suggest that mathcomp is fairly symmetric. The Learndis dataset has approximately 63 percent in resource room placement; accordingly, the dichotomous variable placemen is also reasonably symmetric. Reading comprehension is less symmetric with a skewness value of -0.937 and a skewness ratio of -3.40.



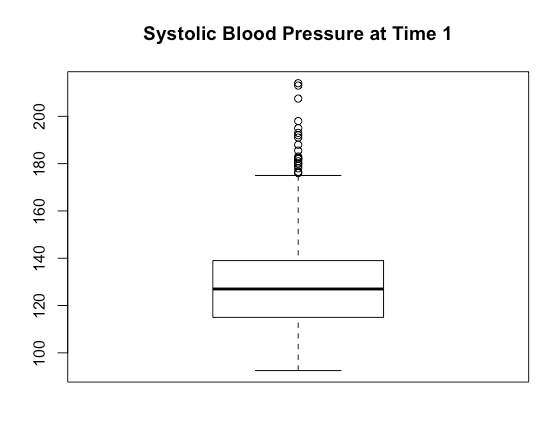
An investigation of the bivariate scatterplot between the scale variables indicates the presence of at least one bivariate outlier.

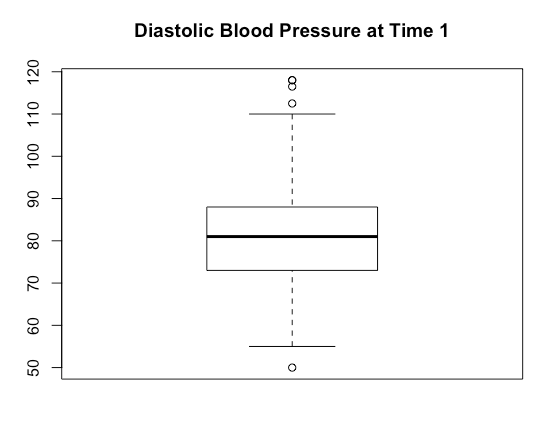


The two zero-order (bivariate) correlations of each predictor variable with readcomp are both statistically significant. The correlation with mathcomp is *r* = 0.49, *p* < 0.0005 and the correlation with placemen is *r* = -0.44, *p* < 0.0005. Among children attending public school in the urban area who have been diagnosed with learning disabilities, higher reading comprehension is associated with higher math comprehension, and placement in a resource room. The intercorrelation between placemen and mathcomp is *r* = -0.34, *p* = 0.001.

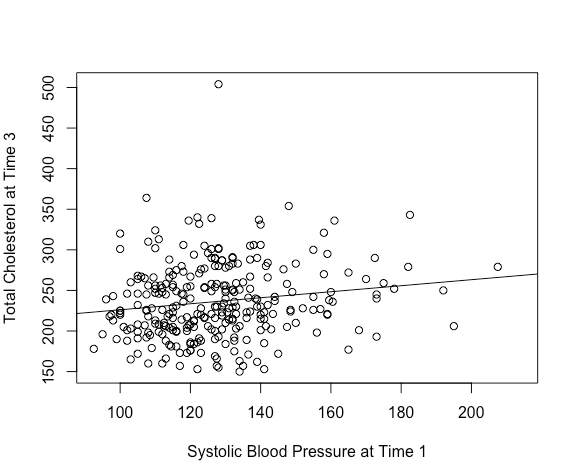
1. According to the R output, the regression model is statistically significant, *F*(2, 71) = 16.65, *p* < 0.0005. According to the coefficients table, placemen (*t*(71) = -2.79, *p* = 0.007) and mathcomp (*t*(71) = 3.37, *p* = 0.001) both make a statistically significant unique contribution to the model.
2. According to the value of adjusted *R2*, approximately 30 percent of the variance in reading comprehension can be explained by type of placement and math comprehension.
3. *R2* (0.319) and *R2adjusted* (0.30) are similar because the ratio of the number of subjects (*N* = 76) to predictor variables (*k* = 2) is relatively large.
4. Predicted readcomp = 52.043 + .329(mathcomp) – 8.533(placemen).
5. Because mathcomp does not take on scores near 0, it would not be meaningful to interpret the value of the intercept.
6. Controlling for placemen, children attending public school in the urban area who have been diagnosed with learning disabilities who have higher math comprehension scores tend also to have higher reading comprehension scores on average. The unique contribution of the variable mathcomp is statistically significant and positive, (*t*(71) = 3.37, *p* = 0.001).
7. Holding placemen constant, a one-point increase in mathcomp is associated with a 0.329-point increase in the predicted readcomp, on average.
8. Yes. Controlling for math comprehension, children attending public school in the urban area who have been diagnosed with learning disabilities who are full time in a self-contained classroom (coded as 1 in the regression equation) score statistically significantly lower, on average, than those in a resource room for part of the day (coded as 0 in the regression equation). The unique contribution of the variable placemen is statistically significant, (*t*(71) = -2.79, *p* = 0.007)
9. Holding mathcomp constant, children attending public school in the urban area who have been diagnosed with learning disabilities who are full time in a self-contained classroom (coded as 1 in the regression equation) are predicted to score 8.533 points lower, on average, than those in a resource room for part of the day (coded as 0 in the regression equation).
10. Predicted readcomp = 52.043 + .329(84) – 8.533(0) = 79.68
11. Because data have not been collected on students with mathcomp scores near 40, it would not be meaningful to make a prediction for those students based on this regression model.
12. An investigation of the univariate distributions indicates that all of the variables are significantly positively skewed. Boxplots of the variables all have outliers. The skewness ratio for TOTCHOL3, SYSBP1, and DIABP1 are 6.73, 9.41, and 3.61, respectively, indicating that the variables are significantly positively skewed.

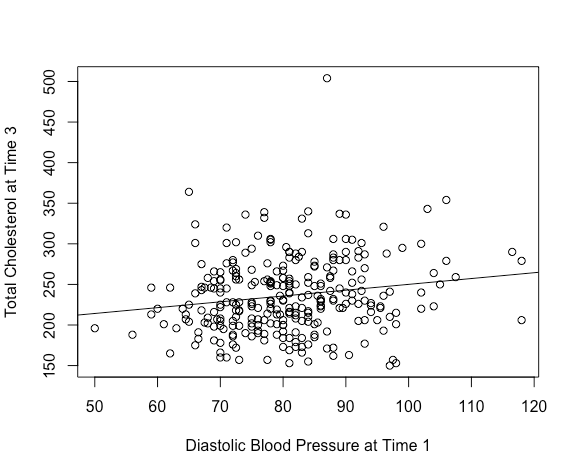


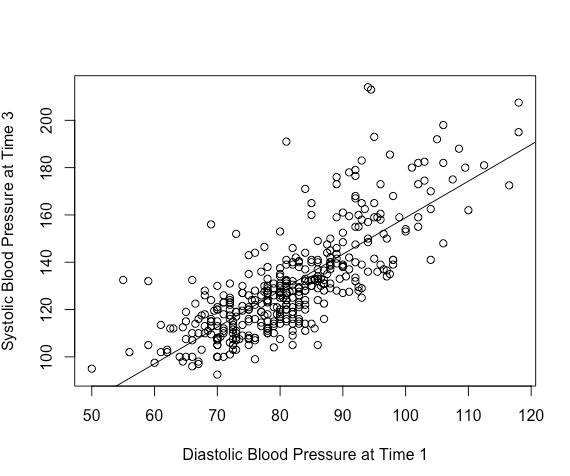




The bivariate scatterplots between the variables suggest that the relationships of the predictor variables with the outcome variable are weak, and the intercorrelation between the two predictor variables is strong.

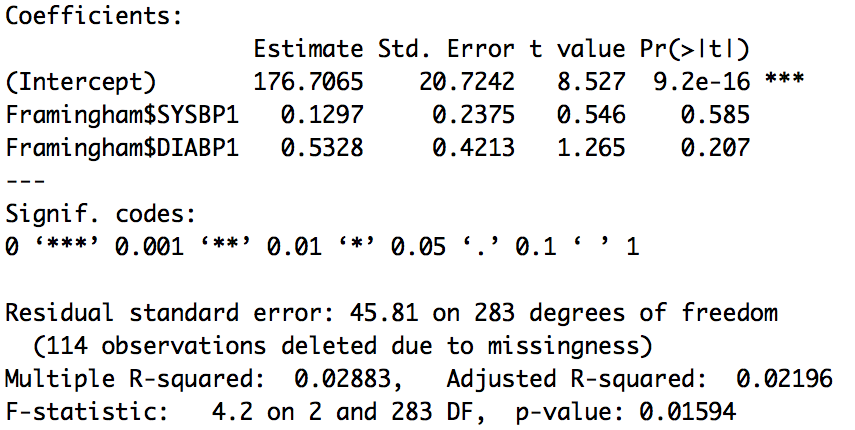






In particular, the zero-order correlations with TOTCHOL3 are *r* = 0.15 for SYSBP1 and *r* = 0.17 for DIABP1. The bivariate correlations of SYSBP1 and DIABP1 with TOTCHOL3 are statistically significant. The intercorrelation between the predictor variables is very strong (*r* = 0.78, *p* < 0.0005).

1. The regression model is statistically significant, *F*(2, 283) = 4.20, *p* = 0.02. However, none of the predictor variables is statistically significant in the model.

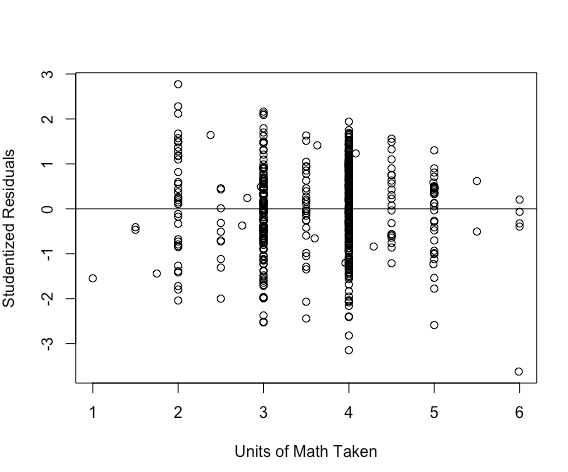


1. Because the predictor variables are highly intercorrelated and only moderately correlated with the outcome variable, each one does not make a unique contribution to the model. Because of the reasonably strong intercorrelation between predictor variables in this case, the two variables may be considered to be multicollinear. While individually each variable is not statistically significant, as a set of two taken together, they explain a statistically significant amount of TOTCHOL3 variance.
2. With two highly overlapping predictor variables, it would be better to eliminate one predictor variable from the equation since they are tapping the same outcome variable variance and in that sense, may be considered somewhat redundant of each other. In this case, the bivariate correlations indicate that DIAPB1 is a slightly better predictor of TOTCHOL3, so an appropriate regression equation is Predicted TOTCHOL3 = 178.33 + .717(DIABP1), which is still statistically significant.
3. According to the values of *R2*, approximately 2.8% of the variance in TOTCHOL3 is explained by DIABP1 alone and 2.9% is explained by the combination of DIABP1 and SYSBP1. SYSBP1 does not contribute much explanatory power to the model over and above that explained by DIABP1.
4. Given that the set of points in this residual scatterplot appears to be randomly scattered and to have a rectangular shape around the studentized residual value of zero, there does not appear to be a curvilinearassociation between mathematics achievement in twelfth grade and the regressor unitmath. The R commands to save the Studentized residuals and create the scatterplot are given below.

**NELS$rstu = rstudent(results)**

**plot(NELS$rstu~NELS$unitmath, xlab = "Units of Math Taken", ylab = "Studentized Residuals")**

**abline(h=0)**

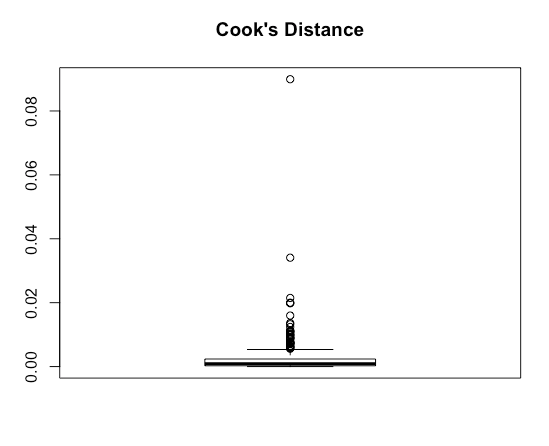


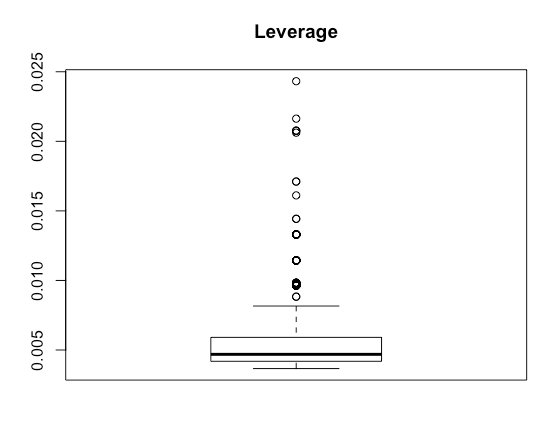
1. There are 23 cases that are bivariate outliers. Any two of these with positive values and any two with negative values may be cited as examples of what is being asked by this question, including id values 36, 192, and 317 with high positive residuals, as well as id values 12, 39, and 45 with low negative residuals. A person would have a positive residual if their twelfth-grade math achievement were under-predicted by the model. That could occur, for example, when the person’s math achievement was much better than would be expected based on the number of math classes taken. Likewise, a person would have a negative residual if his or her twelfth-grade math achievement were over-predicted by the model. The R commands to obtain large-magnitude Studentized residuals are:

**NELS$id[NELS$rstu > 2]**

**NELS$id[NELS$rstu < -2]**

1. The graphs suggest that no point or set of points is unduly influencing the results of the analysis and distorting the results obtained as all distance and leverage values cluster near zero.

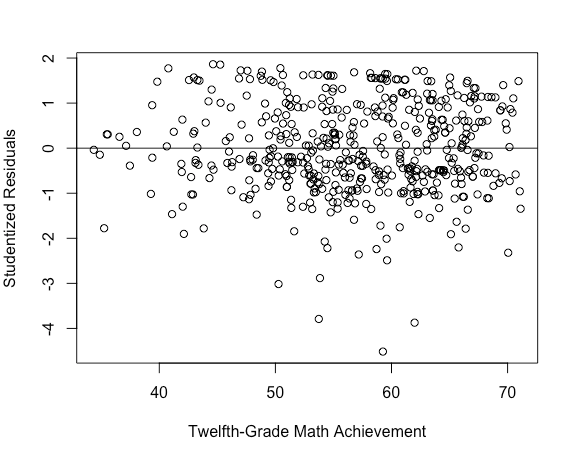


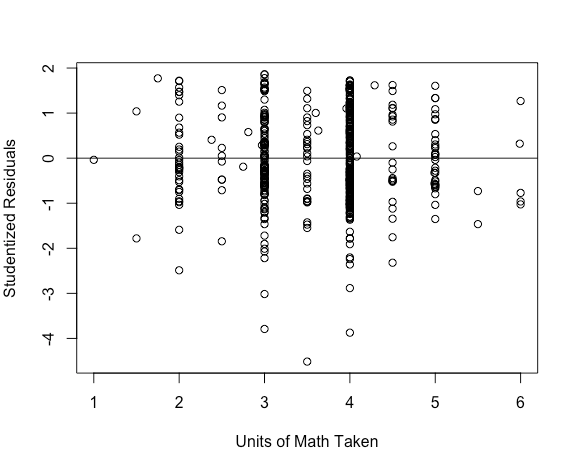


1. Since there is no concerning pattern to the Studentized residuals and no points with particularly large values of leverage or Cook’s distance, it is appropriate to use regression, as defined in Exercise 16.1.
2. Given that the sets of points in these residual scatterplots each appears to be randomly scattered and to have a reasonably rectangular shape around its respective studentized residual value of zero, a curvilinear association between slfcnc12 and achmat12 or between slfcnc12 and unitmath does not appear to exist. The R commands to create the plot for achmat12 are:

**plot(NELS$rstu~NELS$achmat12, xlab = "Twelfth-Grade Math Achievement", ylab = "Studentized Residuals")**

**abline(h=0)**

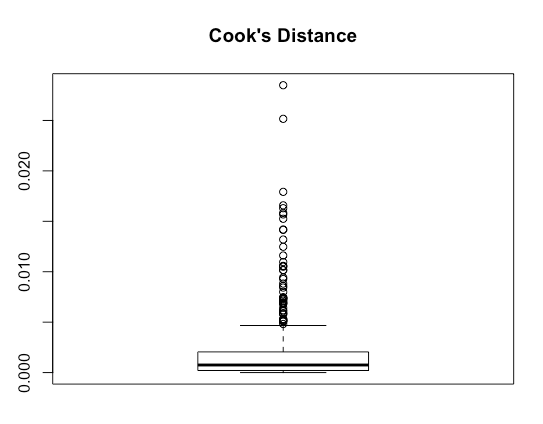


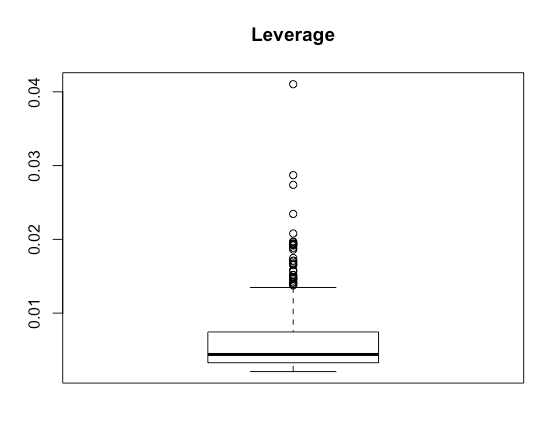


1. There are 13 cases that are bivariate outliers, all of which are negative. Their case numbers are shown in the output below, obtained using the R command **NELS$id[NELS$rstu < -2]**.

../../../Desktop/Screen%20Shot%202019-08-27%20at%209.56.17%20P

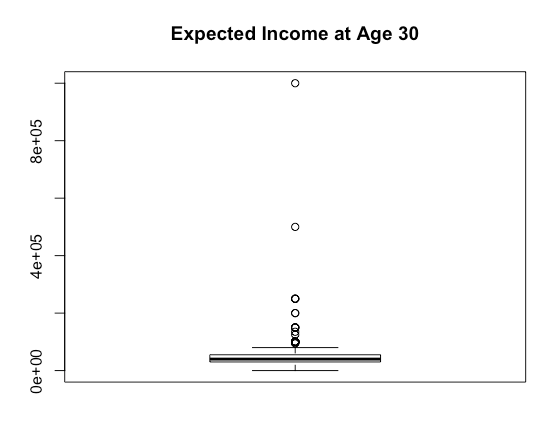
1. These graphs suggest that no point or set of points is unduly influencing the results of the analysis and distorting the results obtained as all distance and leverage values cluster near zero.

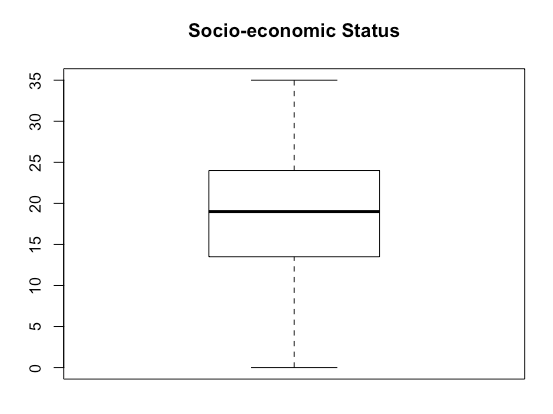




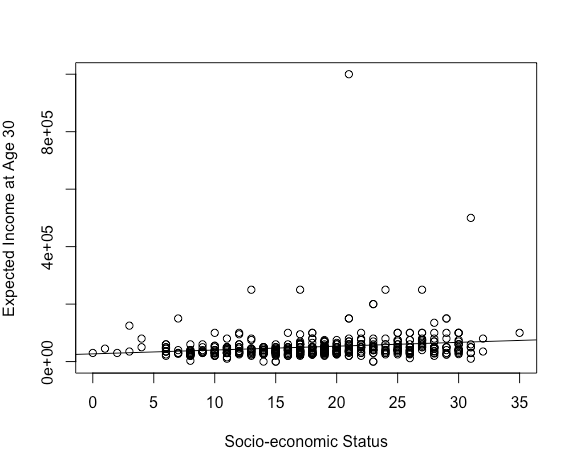
1. Since there is no concerning pattern to the Studentized residuals and no points with particularly large values of leverage or Cook’s distance, it is appropriate to use regression, as defined in Exercise 16.2.
2. Boxplots indicate that ses is fairly symmetric, but that expinc30 is severely positively skewed. The boxplot of expinc30 has many outliers which are creating the severe positive skew (the skewness ratio for expinc30 is 95.63).

The variable gender is reasonably symmetric with approximately 55% females and 45% males.





The outliers may also be observed clearly on the bivariate scatterplot between expinc30 and ses.



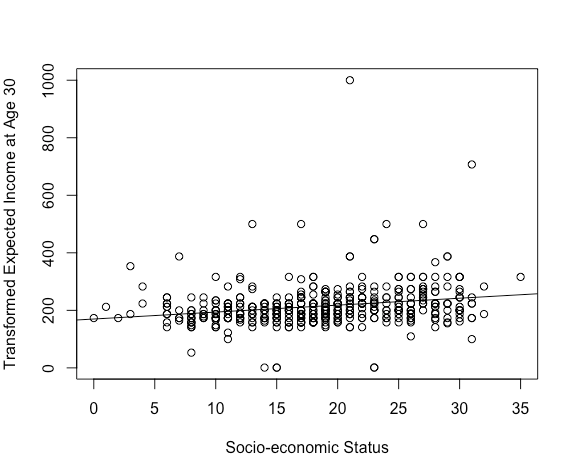
The bivariate correlations of ses with expinc30 (*r* = 0.16, *p* = 0.001) and gender with expinc30 (*r* = -0.15, *p* = 0.002) are both statistically significant. College-bound students who are always at grade level who have higher ses tend to have higher income expectations. Among college-bound students who are always at grade level, males have higher income expectations than females, on average. There is a weak correlation between the two predictor variables ses and gender (*r* = -0.09, *p* = 0.048).

1. To avoid taking a log of zero, we add a small value (+1) to each variable prior to transforming the variables. Although the resulting transformed variables remain skewed, the square root transformation is more effective in creating a more symmetric variable. The skewness ratios for the log transformed and square root transformed variables are -60.26 and 32.71, respectively. The R commands used to create these variables are:

**skew.ratio(NELS$expinclg)**

**skew.ratio(NELS$expincsq)**

1. With only few exceptions, the scatterplot of points between the transformed expinc30 and ses is more regularly shaped.



The bivariate correlations with the transformed expinc30 are each stronger than they were with the untransformed expinc30. The correlation of the square root of (expinc30+1) with ses is *r* = 0.22, *p* < 0.0005 and with gender it is *r* = -0.20, *p* < 0.0005.

1. According to the R output, the regression model is statistically significant, *F*(2, 456) = 20.13, *p* < 0.0005.
2. The percentage of variance in the untransformed expinc30 that is explained by ses and gender is 4.3%, but in the model using the transformed expinc30, it is 8.1%, suggesting that a better-fitting model is achieved through the transformation of the outcome variable in this case.
3. sqrt(expinc30 + 1) = 186.936 + 2.263(ses) – 26.899(gender).
4. sqrt(expinc30 + 1) = 186.936 + 2.263(15) – 26.899(0) = 186.936 + 33.945 =220.881. Squaring both sides of the equation we have expinc30 + 1 = 48,788.42 and determine that expinc30 = $48,787.42 for a male with an ses of 15.
   1. The regression equation is: Predicted WeightLoss = 7.143 + 0.071(FoodIntake)

There are 10 observations in the sample (*N* =10). The means for WeightLoss and FoodIntake are 7.5 and 5, respectively. The correlation between weight loss and food intake is *r* = 0.05, *p* = 0.90—a very weak relationship indicating that there is little to no relationship between the two variables.

The R commands and regression output are given below.

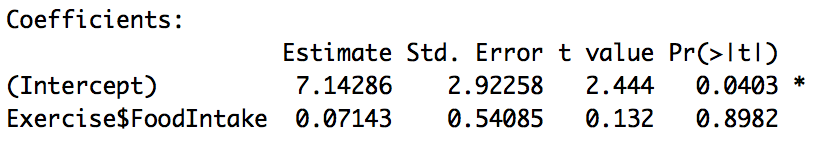
**mean(Exercise$WeightLoss)**

**mean(Exercise$FoodIntake)**

**cor.test(Exercise$WeightLoss,Exercise$FoodIntake)**

**results = lm(Exercise$WeightLoss ~ Exercise$FoodIntake)**

**summary(results)**

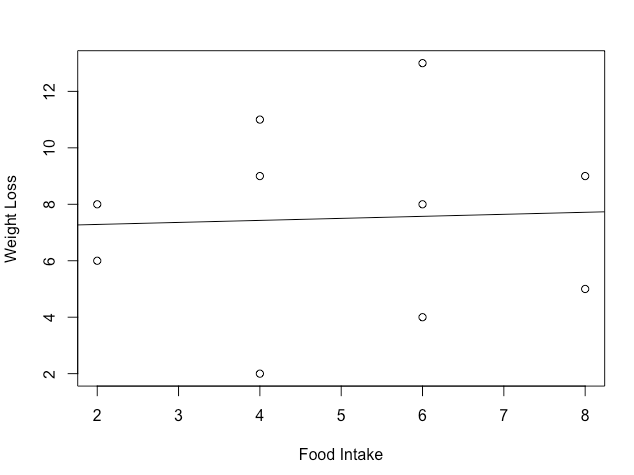


* 1. The *b*-weight in the regression model is *bF* = .071 (*t*(8) = .13, *p* = 0.90). The *b*-weight in this analysis is not statistically significant. To interpret the *b*-weight, we can say that for every 100 calorie increase in food intake (over and above a base of 1000 calories), weight loss increases on average by .071 pounds, an amount that is not statistically significantly different from 0. From this analysis, it would appear that food intake is not related to weight loss, a result that does not make sense. It makes more sense that food intake relates negatively to weight loss, which means that the slope of the regression equation would be negative and that for every unit increase in food intake, weight loss would *decrease* by some number of units. Looking at these data, we would argue that there is an omitted variable in this analysis. We know that there are multiple factors that affect weight loss, and we can posit that one measure of food intake does not provide an adequate measure of weight loss. Moreover, the variable food intake itself could be influenced by a variety of factors and we should explore how additional variables might help us better explain the relationship between food intake and weight loss. One might posit for example, that those who add an hour of exercise each day to their routine might be more likely to add additional calories to provide energy for their exercise, but that overall, they are more likely to lose weight than those who do not exercise but also eat less. To test that idea however, we would add a variable to this analysis to account for the omitted variable, Exercise.
  2. We may obtain the graph using the R commands given below.

**plot(Exercise$WeightLoss ~ Exercise$FoodIntake, xlab = "Food Intake", ylab = "Weight Loss")**

**abline(lm(Exercise$WeightLoss~Exercise$FoodIntake))**

Note that the regression line on the graph passes through (5, 7.5)



* 1. Using the R commands below we obtain the following graph:

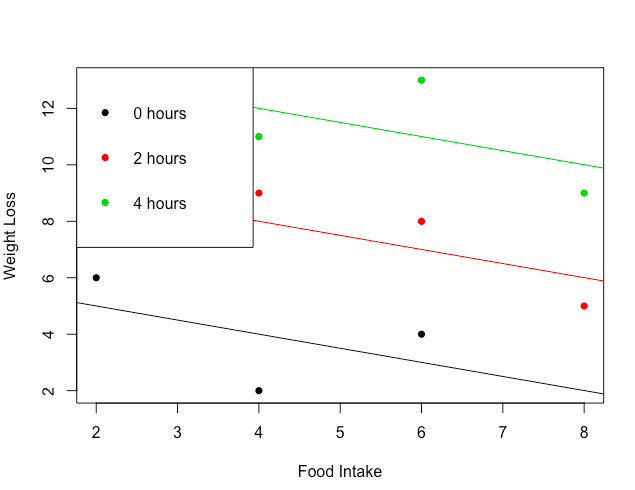
**plot(Exercise$WeightLoss ~ Exercise$FoodIntake, xlab = "Food Intake", ylab = "Weight Loss", pch = 16, col = factor(Exercise$Exercise))**

**abline(lm(Exercise$WeightLoss[Exercise$Exercise==0] ~ Exercise$FoodIntake[Exercise$Exercise==0]), col = 1)**

**abline(lm(Exercise$WeightLoss[Exercise$Exercise==2] ~ Exercise$FoodIntake[Exercise$Exercise==2]), col = 2)**

**abline(lm(Exercise$WeightLoss[Exercise$Exercise==4] ~ Exercise$FoodIntake[Exercise$Exercise==4]), col = 3)**

**legend("topleft", c("0 hours","2 hours","4 hours"), pch = 16, col = c(1,2,3))**

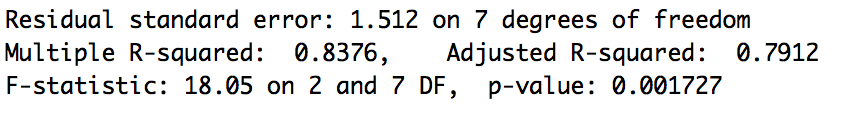


* 1. All three of the regression lines that regress WeightLoss on FoodIntake controlling for Exercise have negative slopes. Thus, the partial relationship between FoodIntake and WeightLoss, controlling for Exercise, is negative. Said differently, on average, the less that one eats above a bare minimum of 1000 calories per day, the more weight one will lose, controlling for amount of exercise. This is consistent with what we would expect in an analysis of weight loss. It makes sense that if one is exercising 4 hours per day, they will require substantially more calories than someone who is not exercising at all. However, even with that increase in food intake, because of the exercise, one is still likely to lose weight. Thus, once we include the variable Exercise in this analysis, to take into account the individuals’ amount of exercise, we obtain a more credible result and a more reasonable understanding of weight loss.
  2. The multiple regression including both food intake and exercise as predictors for weight loss is obtained using the following R commands:

**results2 = lm(Exercise$WeightLoss ~ Exercise$FoodIntake + Exercise$Exercise)**

**summary(results2)**

From the regression output we see that the regression model is statistically significant, *F*(2, 7) = 18.05, *p* = 0.002.



To understand this result, we may look at the various correlations between the variables (f = FoodIntake; w = WeightLoss; e = Exercise). The correlations are *rwf* = 0.05 (*p* = 0.90), *rwe* = 0.86 (*p* = 0.001), and *rfe* = 0.38 (*p* = 0.28). Thus, Exercise and WeightLoss are highly correlated with one another, while FoodIntake and Exercise are somewhat moderately correlated. As explained earlier, there is a correlation close to 0 for WeightLoss and FoodIntake.

This equation fits the model quite well. The *R2* for this equation is .838, indicating that exercise and food intake taken together can explain about 83.8% of the variation in weight loss for the sample.

* 1. From the output in part (f), we see that the regression equation is

Predicted WeightLoss = -0.5(FoodIntake) + 2(Exercise) + 6.

In the model, the *b*-weight for Food Intake is *bf*= -0.5 (*t*(7) = -1.98, *p* = 0.09), indicating that for every 100 calorie increase in food intake (above a base of 1000 calories), weight loss decreases on average by 0.5 pounds, controlling for exercise. That is consistent with the results from our graph. This finding also is consistent with what we would have originally expected: controlling for exercise, people who eat more tend to lose less weight. The constant here is 6, which is the predicted weight loss value when FoodIntake and Exercise are both equal to 0. In other words, someone who does 0 hours of exercise and eats 1000 calories per day is predicted to lose 6 pounds, on average.

We should note that the *b*-coefficient for FoodIntake is *not* statistically significant (*p* > 0.05). However, given the small sample size (*N* = 10), and the fact that it is significant at *p* = 0.10 (*p* = 0.088), these results are promising and call for conducting a similar analysis with a larger sample size to increase the power of the analysis.

* 1. According to the results in the coefficients table in part (f), we see that the contribution of FoodIntake over and above Exercise is not statistically significant, *t*(7) = -1.98, *p* = 0.09.

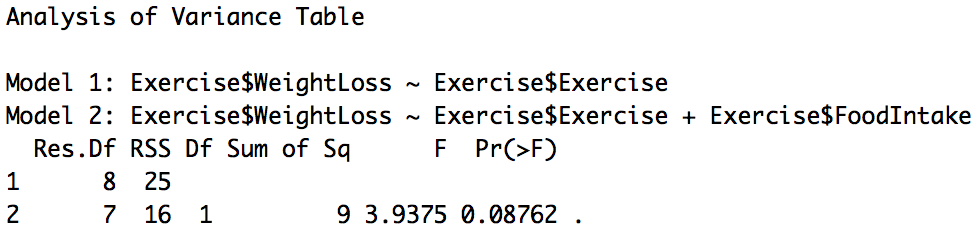
Alternatively, we could use the following R commands to obtain the results of the *F*-test associated with the change in *R2* resulting from adding FoodIntake to the equation.

**results = lm(Exercise$WeightLoss ~ Exercise$Exercise)**

**results2 = lm(Exercise$WeightLoss ~ Exercise$Exercise + Exercise$FoodIntake)**

**anova(results, results2)**

We see that the change in *R2* resulting from adding FoodIntake to the equation is not statistically significant, *F*(1, 7) = 3.94, *p* = 0.09. Note that the *p*-values for the two approaches are identical.



From these analyses, we can conclude that the contribution of FoodIntake over and above Exercise is not statistically significant. The *b*-coefficient, *bf*, and change in *R2* both have *p*-values of 0.09 (*p* > 0.05). As noted earlier, the non-significance may be due to the small sample size and resulting low power of the analysis.

* 1. As a first step, using the Icecream dataset, a regression analysis was performed with relhumid as the outcome variable and temp as the predictor variable. In this analysis, the unstandardized residuals were saved as x21. Then a regression analysis was used to show that *b*2 (or the coefficient of relhumid in the multiple regression analysis on barsold with relhumid and temp as predictor variables, which was equal to .397) equals the *b*-weight in the simple regression equation that predicts ice cream sales (barsold) from that part of humidity unrelated to temperature (x21).

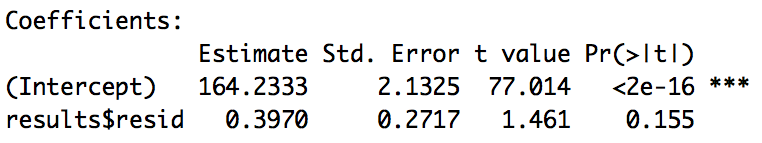
The R commands for completing these steps are:

**results = lm(IceCream$relhumid ~ IceCream$temp)**

**x21 = results$residuals**

**results2 = lm(IceCream$barsold ~ x21)**

**summary(results2)**

****

* 1. d)
  2. b)
  3. c)
  4. c)
  5. b)
  6. There are two possible circumstances. First, *X3* could be statistically significantly correlated with *Y* but uncorrelated with both *X1* and *X2*. Second, *X3* could be uncorrelated with *Y*.