**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  and . These values are moderately negative due to the large sample size of 500 and the corresponding relatively small standard error of skewness of .109. Given that the *NELS* data set contains approximately 45 percent females, we know that GENDER is not skewed, and, is in fact, reasonably symmetric.





The scatterplot of the scale variables and the correlation values of each of the independent variables with ACHMAT12 (*r* = .42 with UNITMATH and *r* = -.20 with GENDER) are both statistically significant (*p* < .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; and, 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 independent variables (the correlation between UNITMATH and GENDER is *r* = -.07, *p* = .13), there is little overlap in their proportion of shared variance and, therefore, both variables should contribute uniquely to the model.

Exercise 15.6 utilizes regression diagnostics to determine aspects of the model fit and whether underlying assumptions appear to be met.

1. According to the ANOVA summary table, the regression model with both first-order variables included in the equation is statistically significant, *F*(2, 497) = 65.35, *p* < .005.
2. According to the value of *R2*, approximately 20.8% of the variance in twelfth grade math achievement can be explained by gender and the NAEP units of math taken in high school. Because the ratio of sample size to independent variables is so large (500 to 2), the *R2* and adjusted *R2* values are quite similar.
3. Predicted ACHMAT12 = 43.793 – 2.719(GENDER) + 4.013(UNITMATH).
4. Predicted ACHMAT12 = 43.793 – 2.719(0) + 4.013(4) = 59.85.
5. Based on the relative magnitudes of the beta weights, UNITMATH is the more important variable in the equation. An analysis of the respective unique proportions of variance accounted for by each independent variable yields the same result. In particular, with UNITMATH in the equation by itself, *R2* is .179 and when GENDER is added to that equation, *R2* increases to .208, an increase of only .029. Alternatively, with GENDER in the equation by itself, R-squared is .04 and when UNITMATH is added to that equation, *R2* increases to .208, an increase of .204. Accordingly, the unique proportion of variance accounted for by GENDER is only 2.9% while for UNITMATH it is 20.4%, indicating that UNITMATH makes a greater unique contribution to explaining ACHMAT12 variance and is therefore the more important variable in the equation.
6. It would not be appropriate to interpret the value of the y-intercept in this case as the smallest number of NAEP 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.
7. Holding gender constant, each additional NAEP unit of math taken in high school corresponds to a 4.01 point increase in twelfth grade math achievement, on average.
8. Holding NAEP units of math taken in high school constant, females perform 2.72 points lower in twelfth grade math achievement, on average, than males.
9. Yes. The coefficient or slope associated with gender is statistically significant, *t*(497) = -4.30, *p* < .0005.
10. Boxplots of the three variables indicate that all variables are negatively skewed with skewness values of -.423 for ACHMAT12, -.247 for UNITMATH, and -.384 for SLFCNC12. Because of the large sample size of 500, skewness ratios all exceed two standard deviations away from a skewness value of zero.

 

An investigation of the bivariate scatterplots between the dependent variable and each of the independent variables and between the two independent variables using a matrix scatterplot, 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 independent and dependent variables and between the two independent variables are as follows:

****

All correlations are statistically significant.

1. According to the ANOVA summary table, the regression model is statistically significant, *F*(2, 497) = 6.800, *p* < .005 with both independent variables in the equation.
2. 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.
3. Only ACHMAT12 is statistically significant with *t*= 2.287, *p* < .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.

****

1. According to the equation, the predicted twelfth grade self concept score of a individual who takes zero units of math in high school and who scores zero on twelfth grade math achievement is 23.020, 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.
2. Holding the number of math units taken in high school constant, each additional one-point increase in twelfth grade math achievement corresponds to a .1 point increase in twelfth grade self-concept, on average.
3. Boxplots suggest that MATHCOMP is fairly symmetric. The *Learndis* data set has approximately 63 percent in resource room placement; accordingly, the dichotomous variable PLACEMEN is also reasonably symmetric. READCOMP is less symmetric with a skewness value of -.956 and a skewness ratio of -3.46.



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 independent variable with READCOMP are both statistically significant. The correlation with MATHCOMP is *r* = .49, *p* < .0005 and the correlation with PLACEMEN is *r* = -.44, *p* < .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* = -.34, *p* = .001.

****

In Exercise 15.9, we investigate the fit of the model and the appropriateness of underlying assumptions using regression diagnostics.

1. According to the ANOVA summary table, the regression model is statistically significant, *F*(2, 71) = 16.65, *p* < .0005. According to the coefficients table, PLACEMEN (*t*(71) = -2.79, *p* =.007) and MATHCOMP (*t*(71) = 3.37, *p* =.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* (.319) and *R2adjusted* (.30) are similar because the ratio of the number of subjects (*N*=76) to independent 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. The unique contribution of the variable MATHCOMP is statistically significant, (*t*(71) = 3.37, *p* =.001).
7. Holding PLACEMEN constant, a one point increase in MATHCOMP is associated with a .329 point increase in READCOMP, on average.
8. Yes. 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) score statistically significantly lower, on average, than those in a resource room for part of the day (coded as 0). The unique contribution of the variable PLACEMEN is statistically significant, (*t*(71) = -2.79, *p* =.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) score 8.533 points lower, on average, than those in a resource room for part of the day (coded as 0).
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. The correlation between READCOMP and IQ is statistically significant, *r* = .29, *p* = .01. Approximately 8.2 percent of the variance in READCOMP can be explained by IQ (*R2* = .2862 = .082).
13. The correlation between READCOMP and MATHCOMP is statistically significant, *r* = .49, *p* < .0005. Approximately 49.4 percent of the variance in READCOMP can be explained by MATHCOMP (*R2* = .4942 = .244).
14. The correlation between IQ and MATHCOMP is statistically significant (*r* = .26, *p* = .01).
15. Approximately 17.7 percent of the variance in READCOMP can be explained by MATHCOMP after controlling for intellectual ability (*R2Change* = .177), a statistically significant amount, *FChange*(1, 71) = 17.33, *p* < .0005.

****

1. Approximately 3.1 percent of the variance in READCOMP can be explained by IQ controlling for MATHCOMP (*R2Change* = .031), which is not a statistically significant amount, *FChange*(1, 71) = 2.97, *p* = .09.

****



1. Using simultaneous multiple regression and placing both independent variables in the equation at the same time, we see that the proportion of variance in READCOMP that can be explained by IQ after controlling for MATHCOMP is not statistically significant *t*(71) = 1.73, *p* = .09.

****

1. Approximately 8.2 percent of the variance in READCOMP can be explained by IQ, but only about 3.1 percent of the variance in READCOMP can be explained by IQ after controlling for MATHCOMP, a much smaller percentage. This is because MATHCOMP and IQ are intercorrelated, *r* = .26, *p* = .01.
2. *R2* for the model is .274 which is less than *r2*Readng Comprehension,IQ + *r2*Reading Comprehension, Math Comprehension = .082 + .244 = .326. That is because *R2* for the model only counts the overlap of approximately .06 between IQ and Math Comprehension (that is, the squared intercorrelation between IQ and Math Comprehension) once. See the Venn diagrams given in the answer to part (d) of this problem.
3. 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 6.77, 9.45, and 3.63, respectively, indicating that the variables are significantly positively skewed.

 

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



In particular, the zero-order correlations with TOTCHOL3 are given in the following table.

****

The bivariate correlations of SYSBP1 and DIABP1 with TOTCHOL3 are statistically significant. The intercorrelations between the independent variables is very strong (*r* = .78, *p* < .0005).

1. The regression model is statistically significant, *F*(2, 283) = 4.20, *p* = .02. However, none of the independent variables is statistically significant in the model.

****

1. Because the independent variables are highly intercorrelated and only moderately correlated with the dependent variable, each one does not make a unique contribution to the model. Because of the reasonably strong intercorrelation between independent 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 TOTCHOL3variance.
2. With two highly overlapping independent variables, it would be better to eliminate one independent variable from the equation since they are tapping the same dependent 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 percent of the variance in TOTCHOL3 is explained by DIABP1 alone and 2.9 percent 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.



1. There are 22 cases that are bivariate outliers. Their case numbers and the values of their standardized residuals are given in the Case Summaries table. 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.

****

A person would have a positive residual if his or her 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.

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. It is appropriate as defined in Exercise 15.1.
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 SLFNC12 and ACHMAT12 or between SLFNC12 and UNITMATH does not appear to exist.



1. There are 12 cases that are bivariate outliers. Their case numbers and the values of their standardized residuals are given in the Case Summaries table.



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. It is appropriate as defined in Exercise 15.2.
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).

GENDER is reasonably symmetric with approximately 45 percent females and 55 percent males.

 

The outliers may also be observed clearly on the bivariate scatterplot between EXPINC30 and SES.



The bivariate correlations of SES with EXPINC30 (*r* = .16, *p* = .001) and GENDER with EXPINC30 (*r* = -.15, *p* = .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 independent variables SES and GENDER (*r* = -.09, *p* = .048).

b) 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.

****

c) 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 with SES is *r* = .22, *p* < .0005 and with GENDER it is *r* = -.20, *p* < .0005.

d) According to the ANOVA summary table, the regression model is statistically significant, *F*(2, 458) = 20.13, *p* < .0005.

e) 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 dependent variable in this case.

f) SQRT(EXPINC30 + 1) = 186.936 + 2.263(SES) – 26.899(GENDER).

g) 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. Given that the residuals for low values of math comprehension are largely negative, those with middle values of math comprehension are largely positive, and those with high values of math comprehension are largely negative, a curvilinear relationship is suggested between the residual READCOMP variance and MATHCOMP. In particular, the model would appear to be improved by the addition of the square of math comprehension, after math comprehension is centered.



1. MATHCOMPCTRDSQD = (MATHCOMP – 86.28)\*\*2
2. 
3. The model is statistically significant, *F*(3, 70) = 13.03, *p* < .0005. The squared term is statistically significant *tMATHCOMPCTRDSQD*(70) = -2.07, *p* =.043.
4. According to the value of *R2Change*, approximately 3.9 percent of the variance in READCOMP is explained uniquely by the square term.
5. Given that the set of points in the residual scatterplot appears to be randomly scattered and to have a rectangular shape around the studentized residual value of zero, the squared term appears to have captured the curvilinearassociation between READCOMP and MATHCOMP.



The boxplots 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. Including religious and non-religious: 130.
2. The lines are not parallel, so there appears to be an interaction, although tests of inference are needed to determine whether this appearance of an interaction is real or due to chance. According to the graph, students who took advanced math in eighth grade tended to do better than those who did not. Students who attended private schools tended to do better than those who did not. But, the interaction suggested by the non-parallel lines implies that the achievement gap between those who took and those who did not take advanced math in eighth grade is smaller for those in private schools.



1. .0949
2. According to the ANOVA summary table, the model including the two independent variables and their interaction is statistically significant, *F*(3, 487) = 20.32, *p* < .0005.

****

1. There are two ways to tell if a single interaction term is statistically significant. According to the value of *FChange*(1, 487) = 4.31, *p* = .04, the interaction is statistically significant. According to the *t*-test of the slope, *b* = -3.29, *p* = .04, the interaction is statistically significant.

****

****

1. According to the value of *R*2, approximately 11.1 percent of the variance in twelfth grade math achievement can be explained by the type of school, whether or not the student took advanced math in eighth grade and their interaction. According to the value of *R2Change*, approximately .8 percent of the variance in twelfth grade math achievement can be explained uniquely by the interaction.
2. 
3. For those who did not take advanced math in eighth grade, the equation is:



For those who did take advanced math in eighth grade, the equation is:



1. Based on the interpretation of the slopes of these equations, among those who did not take advanced math in eighth grade, those in private school score approximately 4.694 points higher in twelfth grade math achievement, on average, than those in public school. Among those who did take advanced math in eighth grade, those in private school score approximately 1.404 points higher in twelfth grade math achievement, on average, than those in public school. The difference in math achievement between public and private school students is less pronounced among those who do not take advanced math in eighth grade.
2. The Syntax file should include the following:

DATASET ACTIVATE DataSet1.

RECODE schtyp8 (1=0) (2=1) (3=1) (MISSING=SYSMIS) INTO SCHTYPDI.

VARIABLE LABELS SCHTYPDI 'Whether the student attended public or private school'.

EXECUTE.

FREQUENCIES VARIABLES=SCHTYPDI

 /ORDER=ANALYSIS.

GRAPH

 /LINE(MULTIPLE)=MEAN(achmat12) BY SCHTYPDI BY advmath8.

COMPUTE INTSCHADV=SCHTYPDI \* advmath8.

EXECUTE.

DESCRIPTIVES VARIABLES=INTSCHADV

 /STATISTICS=MEAN STDDEV MIN MAX.

REGRESSION

 /DESCRIPTIVES MEAN STDDEV CORR SIG N

 /MISSING LISTWISE

 /STATISTICS COEFF OUTS R ANOVA CHANGE

 /CRITERIA=PIN(.05) POUT(.10)

 /NOORIGIN

 /DEPENDENT achmat12

 /METHOD=ENTER SCHTYPDI advmath8

 /METHOD=ENTER INTSCHADV

* 1. The mean of the centered variable is 0.
	2. -0.0272
	3. Although the regression model with all three variables is statistically significant, *F*(3, 496) = 44.39, *p* < .0005, the interaction term is not statistically significant, *b* = 1.146, *p* = .14.or *FChange*(1, 496) = 2.16, *p* = .14.
	4. Because the interaction is not statistically significant, we may conclude that the relationship between twelfth grade math achievement and the number of NAEP credits taken in math is not different for males and females.
1. Because the two slopes appear to be approximately equal, there appears to be no statistically significant interaction between GENDER and SES.



1. The interaction is not statistically significant. According to the value of *FChange*(1, 496) = .13, *p* = .72, the interaction is not statistically significant. According to the *t*-test of the slope, *b* = -.005, *p* = .72, the interaction is not statistically significant.
2. According to the histograms, EXPINC30 is severely positively skewed, SLFCNC08 is negatively skewed, and SES is symmetric.



According to the scatterplots, there are several bivariate outliers.

 

According to the first order correlations, both eighth grade self concept and socio-economic status are significantly related to expected income at age 30.

****

1. According to both the *Fchange* statistic and the b-weight of the interaction term, the interaction is statistically significant, *Fchange*(1, 455) = 4.451, *p* = .04, *t*(455) = 2.11, *p* = .04.

****

1. = 51123.759 + 902.28(SLFCNC08 – 21.1) + 1271.079(SES – 18.4) + 141.473(SLFCNC08 – 20.94)(SES – 18.52).

OR

= 51123.76 + 902.28(SLFC8C) +1271.079(SESC) +141.473(PRODUCT).

****

1. According to the descriptive statistics available through the regression procedure, the standard deviation of SLFC8C is 5.8423 and of SESC is 6.8321. Thus, we call low SLFC8C = –5.8423, moderate SLFC8C = 0, and high SLFC8C = 5.8423. We call low SESC = -6.8321, moderate SESC = 0 and high SESC = 6.8321. Substituting each combination of low and high values for these variables into the regression equation produces the following 2 x 2 table of cell means that gives (predicted expected income at age 30) values at each of the four pairs of temperature and humidity values.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | SLFC8C |  |
|  |  | Low | High |
| SESC | Low | $42,815.15 | $42,064.09 |
|  | High | $48,889.59 | $70,726.21 |



1. From the line graph, we understand the nature of the interaction. For students with low socio-economic status, there is little or no relationship between expected income at age 30 and eighth grade self concept whereas for students with high socio-economic status, there is a positive relationship betweent these two variables; i.e., for students with high socio-economic status, as eighth grade self concept increases, so does expected income at age 30.
	1. The full regression model is statistically significant, *F*(3, 416) = 33.07, *p* < .0005. The interaction term is statistically significant, *t*(416) = 2.34, *p* = .02.

****

****

* 1. = 13.621 + .405(UNITMATH) – 1.385(NURSERY) + 1.899(UNITMATH)(NURSERY)
	2. = 12.236 + 2.304(UNITMATH)
	3. = 13.621 + .405(UNITMATH)
	4. The slopes are significantly different because the interaction is statistically significantly different.
	5. For students who attended nursery school, taking one more unit of math is associated with a much larger increase in SES (2.3 points) than it is for students who did not attend nursery school (.41 points).



* 1. It means that although there is not a difference in SES between those who did and did not attend nursery school after controlling for units of math taken in high school and there is not a relationship between SES and the units of math taken in high school after controlling for nursery school attendance, the relationship between SES and the number of units of math taken in high school differs for those who did and those who did not attend nursery school.

a) 77.78

b) No. Centering is a translation in which the mean of the distribution is subtracted from every score. This type of linear transformation does not change the standard deviation.

 c) Because the plane looks flat, as opposed to warped, there does not appear to

 be an interaction.

c) The interaction is not statistically significant, *Fchange*(1, 70) = .56, *p* = .46, confirming our graphical impression.

d) *=*  77.925 + .396(CENMATH) – 2.696(CENGRADE) + .05251 (PRODUCT).

Both main effects are statistically significant. The regression equation with mean effects only is a better model because it is more parsimonious; it contains only the variables that are contributing to the model and no more.

****

e) Holding grade constant, each 1 point increase in math comprehension score is associated with a .412 point increase in reading comprehension score, on average.

f) Holding math comprehension constant, each 1 year increase in grade level is associated with a 2.696 point decrease in reading comprehension score, on average.

g) According to the value of *R*2, approximately 32 percent of the variance in reading comprehension score is explained by the two independent variables. A more accurate estimate, given by the adjusted *R*2 is approximately 29 percent of the variance is explained.

h) .227

i) .095

k) According to the change in *R*2, if math comprehension were dropped from the equation, the proportion of variance explained would be only .322 - .227 or .095, whereas, if math comprehension were dropped from the equation, the proportion of variance explained would be .322 - .095 or .227, indicating that math comprehension is more important. Furthermore, the beta weight for math (.46) is larger in magnitude than for grade (–.27).

l) The *F*-value of 11.09 is based on the proportion of dependent variable variance explained by the full model including the two main effects and their interaction. The *F*-value of .556 is based on the proportion of dependent variable variance explained by the interaction only.

1. By analyzing these data in SPSS, we see that there are 10 observations in the sample (*N* =10). Descriptives for Weight Loss are *M* = 7.5, SD = 3.308 and for Food Intake are *M* = 5, SD = 2.16. The correlation between Weight Loss and Food Intake, is *rWF* = .047 (*p* = .90).

The Syntax file should contain the following:

REGRESSION

 /DESCRIPTIVES MEAN STDDEV CORR SIG N

 /MISSING LISTWISE

 /STATISTICS COEFF OUTS R ANOVA

 /CRITERIA=PIN(.05) POUT(.10)

 /NOORIGIN

 /DEPENDENT WeightLoss

 /METHOD=ENTER FoodIntake.

1. The regression equation for predicting Weight Loss (W) from Food Intake (F) is

$\hat{W}=7.143+.071F$. The *b*-weight in this regression model is *bF* = .071 (*t*(8) = .13, *p* = .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; that the slope of the regression equation would be negative; that for every unit increase in food intake weight loss decreases by X 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.

1. The regression line on the total data set does indeed pass through , which in this case is (5, 7.5).





1. All three of the regression lines that regress Weight Loss on Food Intake controlling for Exercise have negative slopes. Thus, the partial relationship between Food Intake and Weight Loss, 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, he/she 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. Before running the regression to predict Weight Loss from both Food Intake and Exercise, we first look at the various correlations between the variables (f = Food Intake; w = Weight Loss; e = Exercise). The correlation rwf = 0.047 (p = .898); rwe = 0.864 (p = .001); and rfe = 0.378 (p = .282). Thus, Exercise and Weight Loss are highly correlated with one another, while Food Intake and Weight Loss are somewhat moderately correlated. As explained earlier, there is a correlation close to 0 for Weight Loss and Food Intake.

This equation fits the model quite well. Looking at the overall model fit, we find that the omnibus *F*-test indicates that the regression equation is statistically significant (*F*(2, 7) = 18.05, *p* = .002).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. The regression equation is $\hat{W}=6-.5F+2E$.

In the model, the *b*-weight for Food Intake is *bf*= -0.5 (*t*(7) = -1.984, *p* = .088), 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. The constant here is 6, which is the predicted weight loss value when Food Intake 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.

This finding is consistent with what we would have originally expected; controlling for exercise, people who eat more tend to lose less weight. .

We should note that the *b*-coefficient for Food Intake is not statistically significant (*p* > .05). However, given the small sample size (*N* = 10), and the fact that it is significant at *p* = .10 (*p* = .088), these results are promising and call for conducting a similar analysis with a larger sample size to increase the power of the analysis.

The bootstrap results given below provide the same substantive conclusions as those from the theoretically-derived results.



1. For this analysis, we first look at the *t*-test associated with the *b*-weight for Food Intake. We find that the *t*-test is not statistically significant, *bf*= -0.5 (*t*(7) = 1.98 , *p* =.09).

Next, we look at the *F*-test associated with the change in *R2*resulting from adding Food Exercise to the equation in Block 2. Similarly we find that the change in *R2*, 0.09, is not statistically significant (*F*(1, 7) = 3.94, *p* = .09).

From these analyses, we can conclude that the contribution of Food Intake over and above Exercise is not statistically significant. The *b*-coefficient, *bf*, and change in *R2* both have *p*-values of 0.09 (*p* > .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.sav* data set, a regression analysis was performed with RELHUMID as the dependent variable and TEMP as the independent variable. In this analysis, the unstandardized residuals were saved and the variable RES\_1 was renamed 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 independent variables, which was equal to .397) equals the b weight in the simple regression equation that predicts ice cream sales (Y) from that part of humidity unrelated to temperature (X21).

****

* 1. According to the results of the hierarchical multiple regression analysis using the centered variables, the interaction term is statistically significant, *t*(26) = -2.59, *p* = .016. The proportion of variance explained by the model including the interaction term is approximately 87 percent, according to the value of adjusted *R*2. These values are identical to those obtained in Example 15.4 based on the non-centered variables.

****

****

****

* 1. On a day with average relative humidity (RELHC = 0), a one point increase in the temperature is associated with a .852 increase in ice cream sales, on average.
	2. On a day with average temperature (TEMPC = 0), a one point increase in the relative humidity is associated with a .542 increase in ice cream sales, on average.
	3. Because although different values would be substituted into different regression equations, the resulting predicted number of ice cream bars sold would be the same.
1. The smallest standardized residual is –1.609 and the largest is 1.946. None of these are outside our acceptable range. The largest value for Cook’s distance is .521, which is also within our acceptable range.

****

1. Because both of the residual scatterplot have points in a circular, cloud-like shape, non of the assumptions (normality, homoscedasticity, and linearity) underlying the multiple regression analysis appear to be violated.

****

****

* 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.