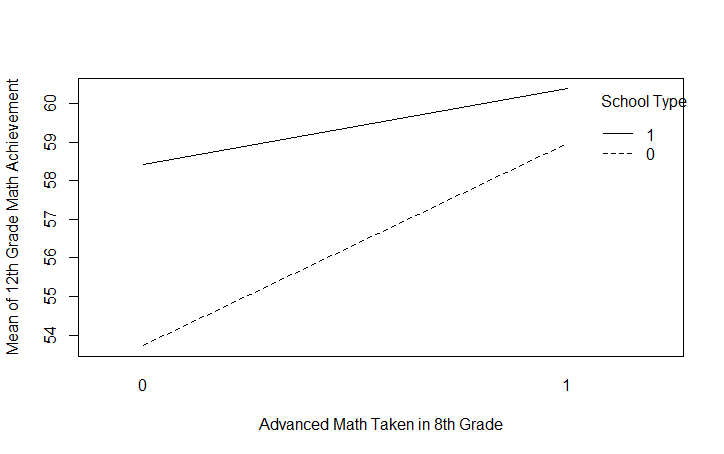
**CHAPTER 17 SOLUTIONS**

1. Including religious and non-religious: 130.

The R commands are given in part k).

1. 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. 0.0949
2. According to the summary table, the model including the two predictor variables and their interaction is statistically significant, *F*(3, 487) = 20.32, *p* < 0.0005.

Estimate Std. Error t value Pr(>|t|)

(Intercept) 53.7247 0.5459 98.423 < 2e-16 \*\*\*

schtypdi 4.6939 1.0061 4.665 3.99e-06 \*\*\*

advmath8 5.2491 0.7805 6.725 4.93e-11 \*\*\*

schtypdi:advmath8 -3.2901 1.5840 -2.077 0.0383 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.464 on 487 degrees of freedom

(9 observations deleted due to missingness)

Multiple R-squared: 0.1112, Adjusted R-squared: 0.1058

F-statistic: 20.32 on 3 and 487 DF, p-value: 2.018e-12

1. There are two ways to tell if a single interaction term is statistically significant. When comparing nested models we get the value of *FChange*(1, 487) = 4.31, *p* = 0.04, and can determine that the interaction is statistically significant. According to the *t*-test of the slope given in part (d), *b* = -3.29, *p* = 0.04, we can determine that the interaction is statistically significant.

Analysis of Variance Table

Model 1: achmat12 ~ schtypdi + advmath8

Model 2: achmat12 ~ schtypdi \* advmath8

Res.Df RSS Df Sum of Sq F Pr(>F)

1 488 27375

2 487 27135 1 240.38 4.3142 0.03832 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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 0.8 percent of the variance in twelfth grade math achievement can be explained uniquely by the interaction.
2. If model 1 is the model with main effects and model 2 is the model with main and interaction effects, then the following table summarizes the statistics for model selection.

Adjusted *R2* and AIC and BIC indicators for the two models.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | AIC | BIC | Adjusted *R2* |
| 1 | 3375.68 | 3392.47 | 0.0997 |
| 2 | 3373.35 | 3394.33 | 0.1058 |

Although the AIC decreased and the BIC increased, they did so by a minimal amount. The slightly higher adjusted *R2* for model 2 provides a rationale for selecting model 2 for this example.

1. 
2. 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 R file for the first few parts of this question should include the following:

**NELS$schtypdi = ifelse(NELS$schtyp8=="Public", 0, 1)**

**NELS$advmath8 = ifelse(NELS$advmath8=="No", 0, 1)**

**interaction.plot(NELS$advmath8, NELS$schtypdi, NELS$achmat12,   
trace.label = "School Type",**

**xlab = "Advanced Math Taken in 8th Grade",**

**ylab = "Mean of 12th Grade Math Achievement")**

**NELS$product = NELS$schtypdi\*NELS$advmath8**

**summary(NELS$product2)**

**results = lm(achmat12~schtypdi\*advmath8, data = NELS)**

**anova(results)**

**summary(results)**

**model1 = lm(achmat12~schtypdi+advmath8, data = NELS)**

**model2 = lm(achmat12~schtypdi\*advmath8, data = NELS)**

**anova(model1, model2)**

**(summary(model2)$r.squared) - (summary(model1)$r.squared)**

The R Commands to find the adjusted *R2* and AIC and BIC indicators are:

**AIC(model1)**

**BIC(model1)**

**summary(model1)$adj.r.squared**

**AIC(model2)**

**BIC(model2)**

**summary(model2)$adj.r.squared**

* 1. The mean of the centered variable is 0.
  2. -0.0268
  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* = 0.14.or *FChange*(1, 496) = 2.16, *p* = 0.14.
  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.

The R commands used to generate the graph are:

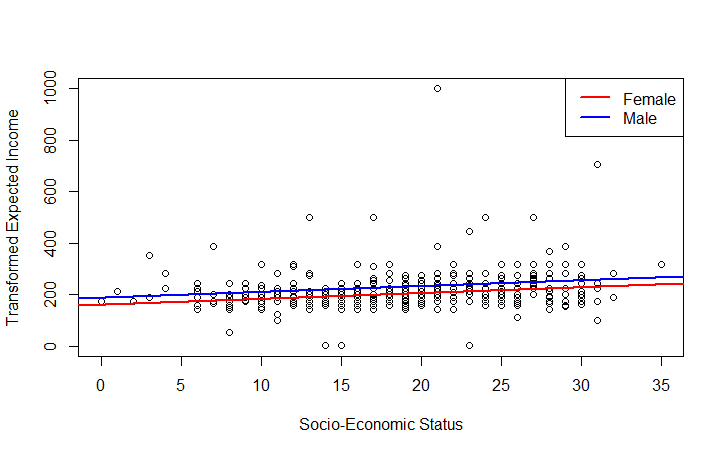
**plot(expinc30v2 ~ ses, data = NELS, xlab = "Socio-Economic Status",**

**ylab = "Transformed Expected Income" )**

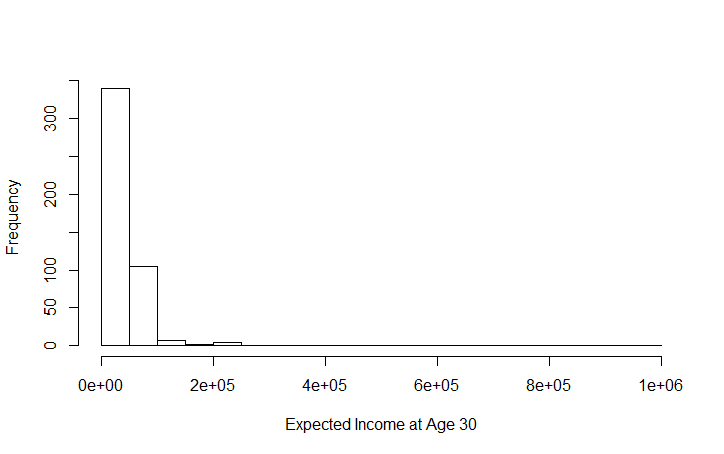
**abline(lm(expinc30v2 ~ ses, data = NELS[NELS$gender=="Female",]), col="red", lty=1, lwd = 2)**

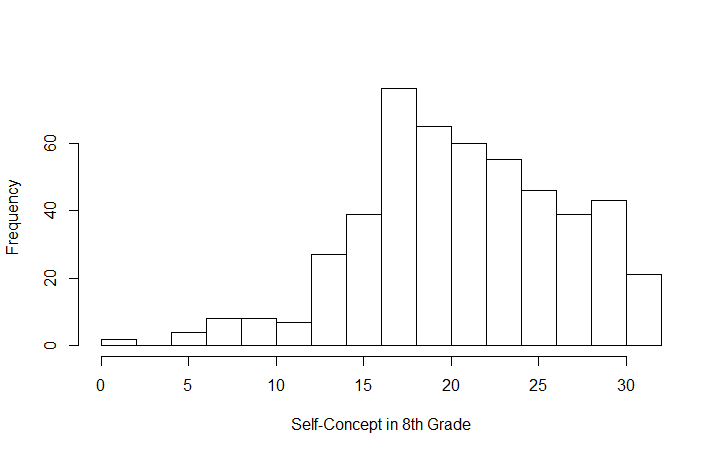
**abline(lm(expinc30v2 ~ ses, data = NELS[NELS$gender=="Male",]), col="blue", lty=1, lwd = 2)**

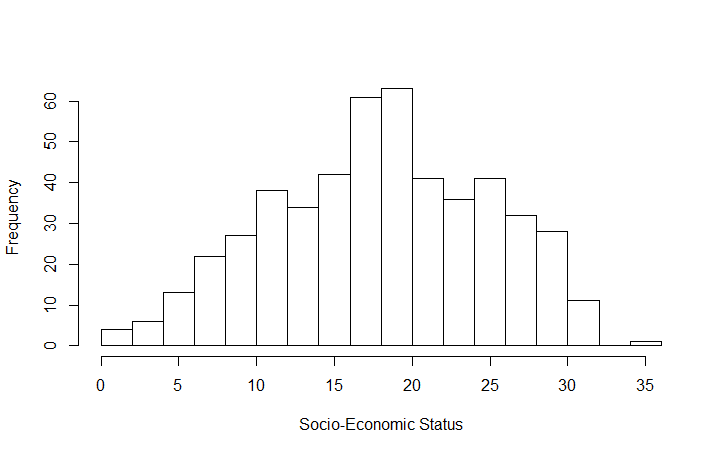
**legend("topright", legend=c("Female", "Male"), col=c("red", "blue"), lty=1, lwd = 2)**



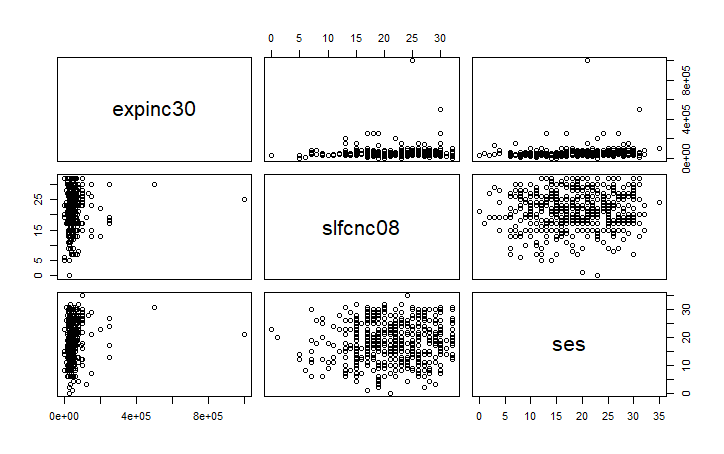
1. The interaction is not statistically significant. According to the value of *FChange*(1, 455) = 0.0013, *p* = 0.96, the interaction is not statistically significant. According to the *t*-test of the slope, *b* = -0.054, *p* = 0.96, the interaction is not statistically significant.
2. According to the histograms, expinc30 is severely positively skewed, slfcnc08 is negatively skewed, and ses is symmetric. The R command for generating the first histogram is: **hist(NELS$expinc30, breaks = 20, xlab = "Expected Income at Age 30", main = "")**







According to the scatterplots, there are several bivariate outliers. This matrix scatterplot was created using the R command: **pairs(NELS[,c("expinc30", "slfcnc08", "ses")])**



The correlations were generated using the R command: **cor(NELS[,c("expinc30", "slfcnc08", "ses")], use = "complete.obs")**

expinc30 slfcnc08 ses

expinc30 1.00000000 0.09690464 0.15657459

slfcnc08 0.09690464 1.00000000 0.08032435

ses 0.15657459 0.08032435 1.00000000

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.

The R commands to generate the related output are:

**NELS$slfc8c = NELS$slfcnc08 - mean(NELS$slfcnc08)**

**NELS$sesc = NELS$ses - mean(NELS$ses)**

**model1 = lm(expinc30v2~slfc8c+sesc, data = NELS)**

**model2 = lm(expinc30v2~slfc8c\*sesc, data = NELS)**

**anova(model1, model2)**

**summary(model2)**

Analysis of Variance Table

Model 1: expinc30 ~ slfc8c + sesc

Model 2: expinc30 ~ slfc8c \* sesc

Res.Df RSS Df Sum of Sq F Pr(>F)

1 456 1.5056e+12

2 455 1.4910e+12 1 1.4585e+10 4.4506 0.03543 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 51123.04 2681.16 19.068 < 2e-16 \*\*\*

slfc8c 890.11 459.80 1.936 0.05350 .

sesc 1288.34 392.82 3.280 0.00112 \*\*

slfc8c:sesc 141.47 67.06 2.110 0.03543 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 57250 on 455 degrees of freedom

(41 observations deleted due to missingness)

Multiple R-squared: 0.04105, Adjusted R-squared: 0.03473

F-statistic: 6.493 on 3 and 455 DF, p-value: 0.0002614

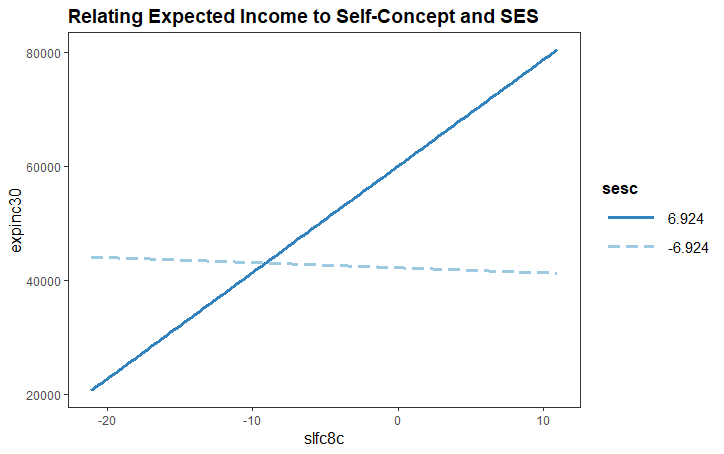
1. = 51123.04 + 890.11(slfc8c) +1288.34(sesc) +141.47(slfc8c\*sesc).
2. The standard deviation of slfc8c is 5.971 and of sesc is 6.924. Thus, we call low slfc8c = –5.971 and high slfc8c = 5.971. We call low sesc = -6.924 and high sesc = 6.924. Substituting each combination of low and high values for 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 self-concept and ses values.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | SLFC8C |  |
|  |  | Low | High |
| SESC | Low | $42,736.55 | $41,668.60 |
|  | High | $48,879.84 | $71207.18 |

1. The R command for generating the line graph is:

**interact\_plot(model2, pred = "slfc8c", modx = "sesc", modx.values = c(-6.924, 6.924),**

**ylab = "Expected Income at Age 30", main = "Relating Expected Income to Self-Concept and SES")**



1. From the line graph, we understand the nature of the interaction. For students with low ses, 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 between these two variables; i.e., for students with high socio-economic status, those with higher eighth grade self-concept tended also to have higher expected income at age 30.
   1. The R commands to generate the univariate and bivariate statistics are:

**NELS$nurseryn <- as.numeric(NELS$nursery) -1**

**cor(NELS[,c("ses", "unitmath", "nurseryn")], use = "complete.obs")**

**plot(ses~nurseryn, data = NELS, xlab = "Nursery School Attended?", ylab = "SES")**

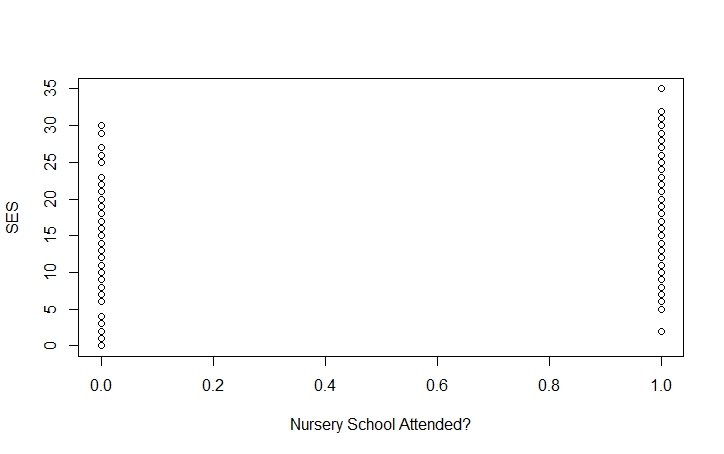
**plot(ses~unitmath, data = NELS, xlab = "Units in Math (NAEP)", ylab = "SES")**

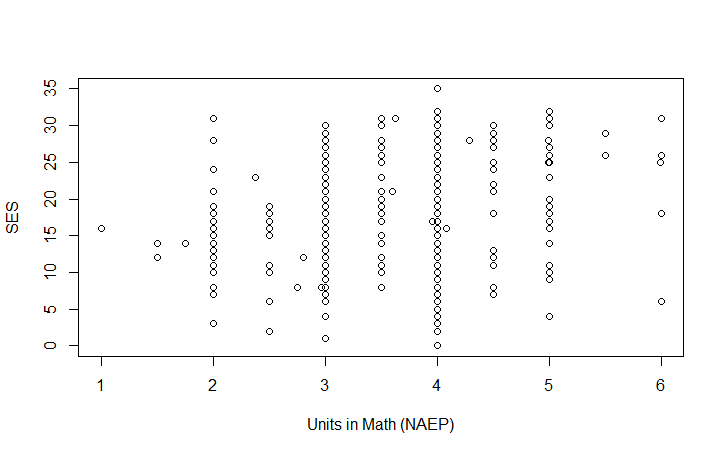
ses unitmath nurseryn

ses 1.0000000 0.20945061 0.38430294

unitmath 0.2094506 1.00000000 0.06461218

nurseryn 0.3843029 0.06461218 1.00000000





* 1. The following R commands is used to center unitmath.

**NELS$unitmathc = NELS$unitmath - mean(NELS$unitmath, na.rm = TRUE)**

The full regression model is statistically significant, *F*(3, 416) = 33.07, *p* < 0.0005. The interaction term is statistically significant, *t*(416) = 2.34, *p* = 0.02.

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 15.0951 0.5325 28.349 < 2e-16 \*\*\*

unitmathc 0.4053 0.6503 0.623 0.5334

nurseryn 5.5235 0.6509 8.486 3.81e-16 \*\*\*

unitmathc:nurseryn 1.8993 0.8102 2.344 0.0195 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.262 on 416 degrees of freedom

(80 observations deleted due to missingness)

Multiple R-squared: 0.1926, Adjusted R-squared: 0.1868

F-statistic: 33.07 on 3 and 416 DF, p-value: < 2.2e-16

* 1. = 15.095 + .405(unitmathc) – 5.52 (nursery) + 1.899(unitmathc)(nursery)
  2. = 9.575 + 2.304(unitmathc)
  3. = 15.095 + .405(unitmathc)
  4. The slopes are significantly different because the interaction is statistically significantly different.
  5. The slope for predicting ses from units of mathematics taken for students who did not attend nursery school (.41) is not statistically significantly different from 0 whereas the slope for students who did attend nursery school (2.3) is statistically significantly different from 0, indicating that students who took more units of mathematics in high school tend to have higher ses.

To generate the relevant output, we used the following R command

**sim\_slopes(model2, pred = "unitmathc", modx = "nurseryn")**

**SIMPLE SLOPES ANALYSIS**

*Slope of unitmathc when nurseryn = 1.00 (1):*

Est. S.E. t val. p

2.30 0.48 4.77 0.00

*Slope of unitmathc when nurseryn = 0.00 (0):*

Est. S.E. t val. p

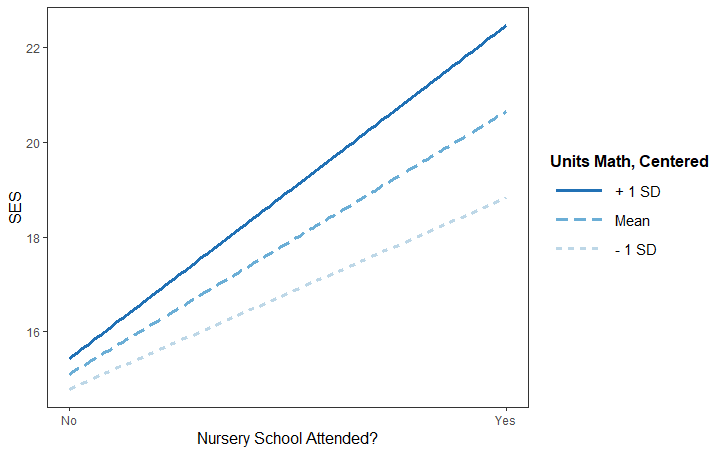
0.41 0.65 0.62 0.53

* 1. The R command to create the graph is:

**interact\_plot(model2, pred = "nurseryn", modx = "unitmathc",**

**x.label = "Nursery School Attended?", y.label = "SES",**

**legend.main = "Units Math, Centered", pred.labels = c("No", "Yes"))**



* 1. 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). 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.
  2. The main effect of units of mathematics taken in high school is not statistically significant, indicating that, controlling for nursery school attendance, there is no relationship between units of mathematics taken and socioeconomic status. The interaction is statistically significant because the relationship between socioeconomic status and units of mathematics taken varies depending on whether the student attended nursery school.

a) 77.63

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.

d) The interaction is not statistically significant, *Fchange*(1, 70) = 0.56, *p* = 0.46, confirming our graphical impression.

e) *=*  78.293+ 0.412(mathcompc) – 2.804(gradec)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 78.29257 1.29665 60.381 < 2e-16 \*\*\*

mathcompc 0.41192 0.08915 4.621 1.67e-05 \*\*\*

gradec -2.80383 1.02266 -2.742 0.00773 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.01 on 71 degrees of freedom

(31 observations deleted due to missingness)

Multiple R-squared: 0.3168, Adjusted R-squared: 0.2975

F-statistic: 16.46 on 2 and 71 DF, p-value: 1.339e-06

Both main effects are statistically significant. The regression equation with main 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.

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

g) Holding math comprehension constant, each 1-year increase in grade level is associated with a 2.804-point decrease in predicted reading comprehension score, on average.

h) According to the value of *R*2, approximately 32 percent of the variance in reading comprehension score is explained by the two predictor variables. A more accurate estimate, given by the adjusted *R*2 is approximately 30 percent of the variance is explained.

* 1. According to the results of a hierarchical multiple regression analysis using the centered variables, the interaction term is statistically significant, *t*(26) = -2.59, *p* = 0.016. 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 17.1 based on the non-centered variables.

The R commands for obtaining these results are:

**IceCream$tempc = IceCream$temp - mean(IceCream$temp, na.rm = TRUE)**

**IceCream$relhumidc = IceCream$relhumid - mean(IceCream$relhumid, na.rm = TRUE)**

**results = lm(barsold~tempc\*relhumidc, data = IceCream)**

**summary(results)**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 165.83887 0.99218 167.146 < 2e-16 \*\*\*

tempc 0.85236 0.11615 7.338 8.58e-08 \*\*\*

relhumidc 0.54236 0.11349 4.779 6.02e-05 \*\*\*

tempc:relhumidc -0.02724 0.01053 -2.586 0.0157 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.239 on 26 degrees of freedom

Multiple R-squared: 0.8864, Adjusted R-squared: 0.8732

F-statistic: 67.6 on 3 and 26 DF, p-value: 2.087e-12

* 1. On a day with average relative humidity (relhumidc = 0), a one-point increase in the temperature is associated with a 0.852 increase in predicted 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 0.542 increase in predicted 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 –1.74 and the largest is 2.02. Only one of the points in the data set has a Standardizedresidual that is greater than the absolute value of 2. The largest value for Cook’s distance is 0.521, which is within our acceptable range.

> summary(IceCream$stres)

Min. 1st Qu. Median Mean 3rd Qu. Max.

-1.74757 -0.78603 0.13568 0.01349 0.60619 2.01516

> summary(IceCream$cooks)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0000597 0.0045974 0.0170067 0.0600931 0.0476646 0.5210325

The R commands for performing this analysis and for saving the Standardized residuals and Cook’s distance are:

**IceCream$stres = rstandard(results)**

**IceCream$cooks = cooks.distance(results)**

**summary(IceCream$stres)**

**summary(IceCream$cooks)**

**table(IceCream$id[abs(IceCream$stres)>2])**

**table(IceCream$id[IceCream$cooks>1])**

1. In order to perform a sensitivity analysis, we perform the regression again while omitting the point with Studentized residual greater than 2. We use the R command **regress barsold tempc relhumidc prod, beta, if stres < 2**

For the original analysis, the results are:

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 165.83887 0.99218 167.146 < 2e-16 \*\*\*

tempc 0.85236 0.11615 7.338 8.58e-08 \*\*\*

relhumidc 0.54236 0.11349 4.779 6.02e-05 \*\*\*

tempc:relhumidc -0.02724 0.01053 -2.586 0.0157 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.239 on 26 degrees of freedom

Multiple R-squared: 0.8864, Adjusted R-squared: 0.8732

F-statistic: 67.6 on 3 and 26 DF, p-value: 2.087e-12

For the analysis with the outlier suppressed, the code is

**results2 = lm(barsold~tempc\*relhumidc, data = IceCream[IceCream$id!=1,])**

**summary(results2)**

and the results are:

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 165.484493 0.943948 175.311 < 2e-16 \*\*\*

tempc 0.824901 0.109555 7.530 6.97e-08 \*\*\*

relhumidc 0.573904 0.107323 5.347 1.52e-05 \*\*\*

tempc:relhumidc -0.026229 0.009878 -2.655 0.0136 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.971 on 25 degrees of freedom

Multiple R-squared: 0.9033, Adjusted R-squared: 0.8917

F-statistic: 77.84 on 3 and 25 DF, p-value: 8.151e-13

We see that there is a negligible change in the adjusted *R2*, from 0.87 with the outlier, to 0.89 without it. This results is consistent with the findings from Cook’s distance indicating that the outlier is not unduly influential on the model.

1. Because both of the residual scatterplot have points in a circular, cloud-like shape, none of the assumptions (normality, homoscedasticity, and linearity) underlying the multiple regression analysis appear to be seriously violated.

The R commands for obtaining the scatterplots are

**plot(IceCream$stres~IceCream$tempc, ylab = "Standardized Residuals")**

**plot(IceCream$stres~IceCream$relhumidc, ylab = "Standardized Residuals")**

