Answers to Application Activities in Chapter 7

7.2.4 Application activity: Graphs for Understanding Complex Relationships (Answers only for R)

Coplots

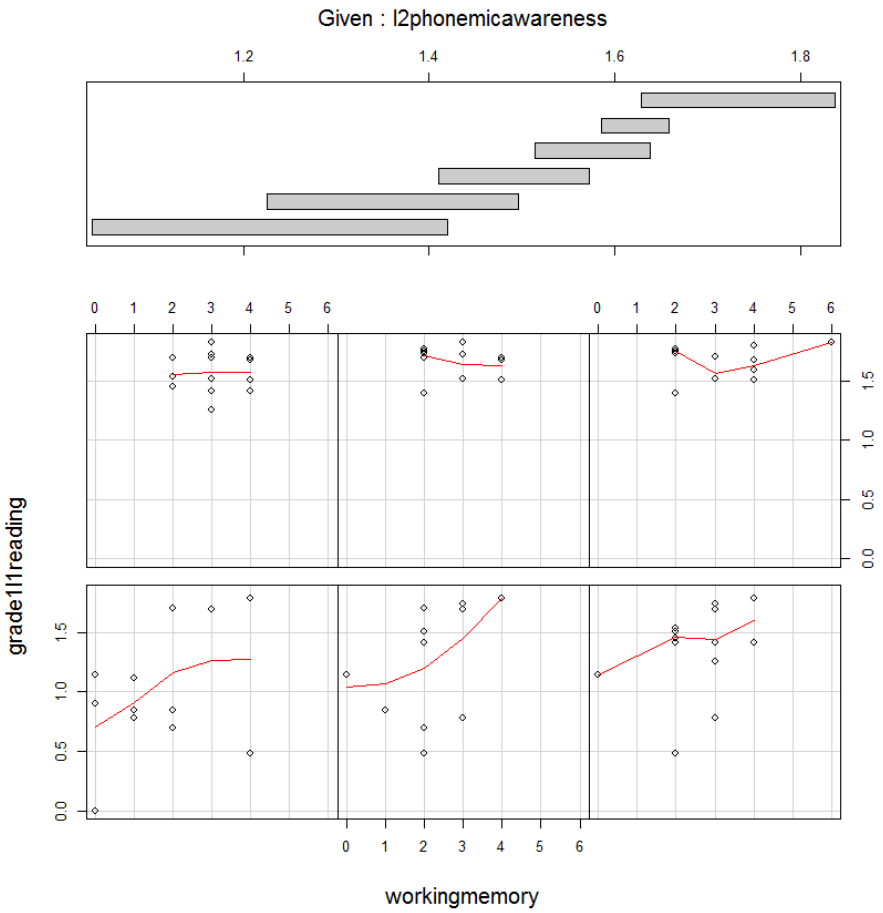
1 LaFrance & Gottardo (2005)

Use the data file LafranceGottardo.sav and import into R as lafrance.

Here is the R code:

```
names(lafrance) #helps me remember exactly how to spell names in coplot( ) function
coplot(grade1l1reading~workingmemory|l2phonemicawareness,
panel= function(x,y,...)
panel.smooth(x,y,span=.8,...),data=lafrance)
```

Trying it in this order I get the figure reprinted below. It looks like at lower levels of phonemic awareness working memory and grade 1 L1 reading measures are more strongly correlated, but that there is less variability in working memory scores and reading scores as phonemic awareness is higher, resulting in lines which show little correlation. The reading scores are especially concentrated in the very top of the graphs in the top row (which is read last remember; start reading from left to right, bottom to top).



2. LaFrance & Gottardo (2005)

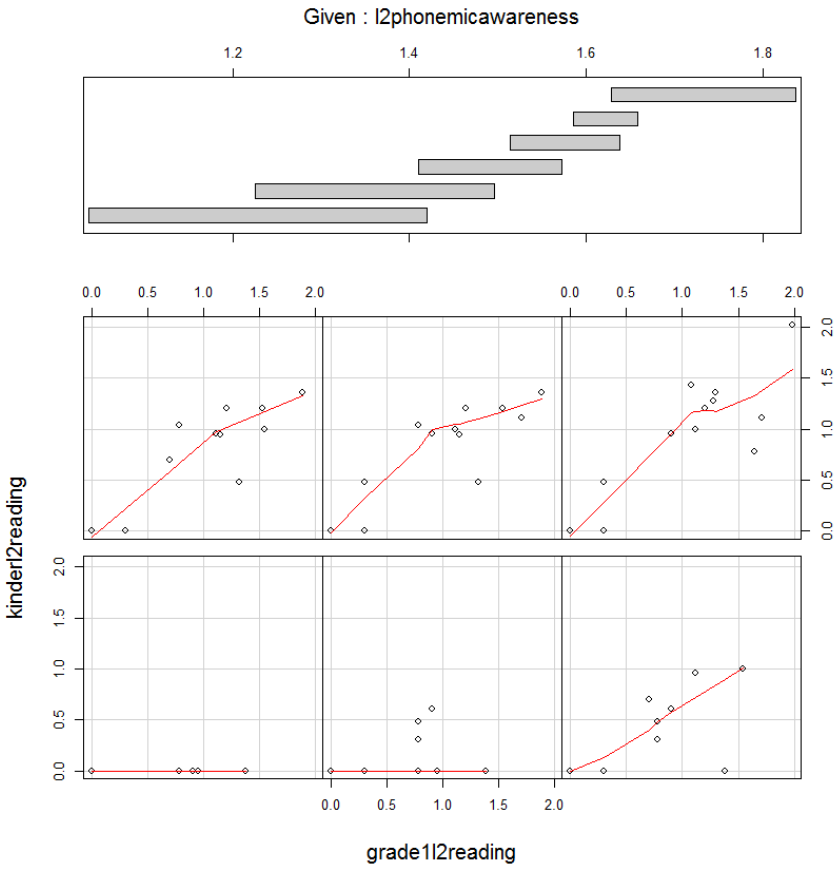
Use the same lafrance file as in Exercise #1.

Here is the R code:

```
coplot(kinderl2reading~grade1l2reading|l2phonemicawareness,
panel= function(x,y,...)
```

panel.smooth(x,y,span=.8,...),data=lafrance)

Trying it in this order I get the figure reprinted below. It looks like at lower levels of phonemic awareness reading measures in L2 from Kindergarten and First Grade are not correlated at all. In other words, children with lower levels of phonemic awareness scored at very low levels on the Kindergarten reading measures, but had more variable scores in First Grade. Once the phonemic awareness scores are higher we see a stronger correlation between Kindergarten and First Grade reading scores, with those children who scored lower in Kindergarten also scoring lower in First Grade, and those scoring higher in Kindergarten also scoring higher in First Grade.



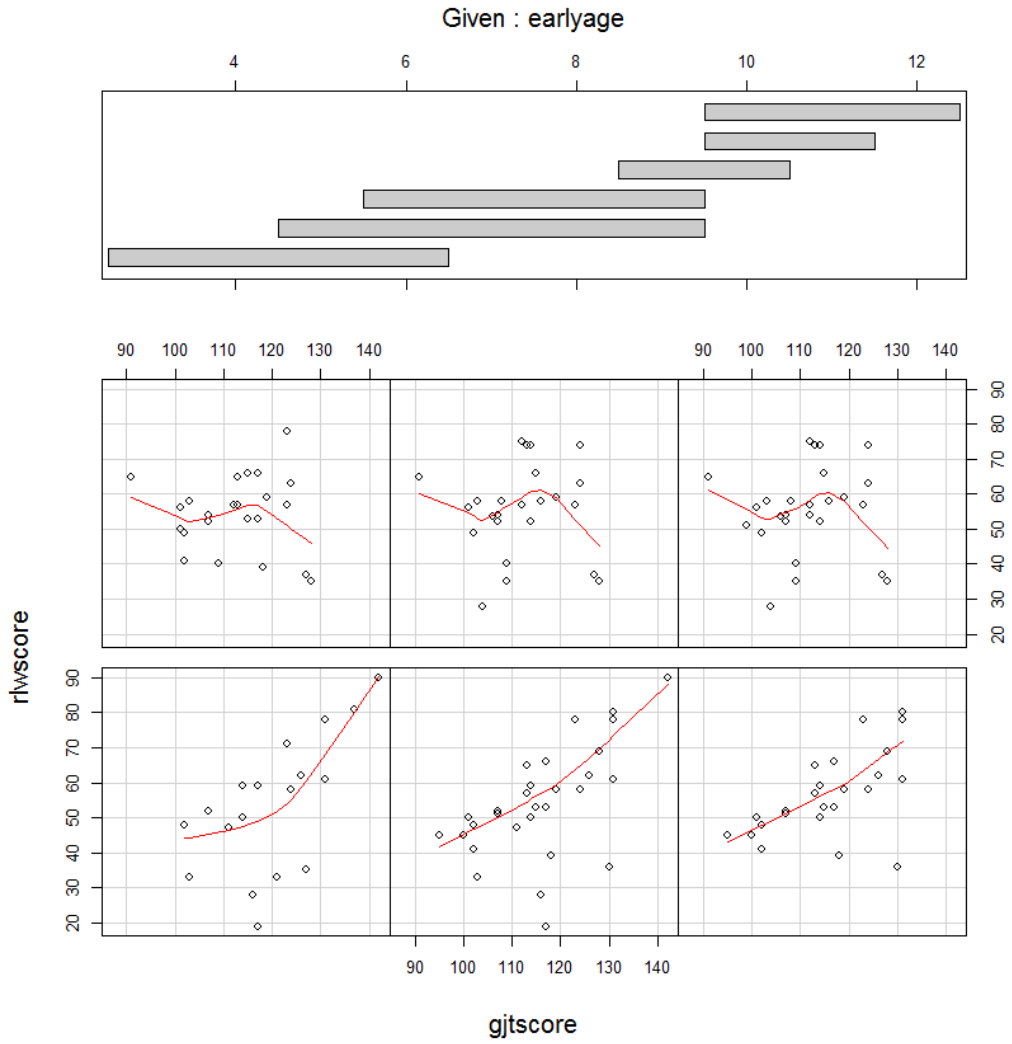
3 Larson-Hall (2008)

Use the data file LarsonHall2008.sav and import into R as lh2008.

Here is the R code:

```
names(llh2008) #helps me remember exactly how to spell names in coplot( ) function
coplot(rlwscore~gjtscore|earlyage,
panel= function(x,y,...)
panel.smooth(x,y,span=.8,...),data=lh2008)
```

Note that not all of the participants in the study began studying English at an early age, so after entering the code you will get a message about missing rows of data. Trying the code in the order I wrote above, I get the figure reprinted below. It looks like scores between the grammar and pronunciation tests were more highly correlated at earlier ages of starting English (up to about 8 or 9) but that after that the Loess line doesn't show any definite trend and the regression line would probably be a straight line with very little correlation.



3D Plots

4 LaFrance & Gottardo (2005)

Use the data file LafranceGottardo.sav imported into R as lafrance.

We can use R Commander to open this up. To open R Commander type:

```
library(Rcmdr)
```

at the prompt line. R Commander should open up.

In R Commander, choose GRAPHS > 3D GRAPH > 3DSCATTERPLOT.

For the explanatory variables choose kinderl2reading and l1phonemicawareness (hold down the CTRL button and use the mouse to click both of these), and for the response variable choose grade1l1reading. Press OK.

It wouldn't make too much sense to show you what this looked like as it is a dynamic picture. Note that you can pull the window of the 3D scatterplot to make it bigger. I will leave it to the reader to describe what trends they can discern from this figure.

Tree Models

5 LaFrance & Gottardo (2005)

Use the data file LafranceGottardo.sav imported into R as lafrance.

This graph is only available using the R console. The code is:

```
library(tree)
model=tree(grade1l2reading~., data=lafrance)
plot(model)
text(model)
```

The first run of the tree model shows that explanatory variables for Grade 1 L2 reading are Kinder L2 reading, grade 1 L1 reading, L1 phonemic awareness and L2 phonemic awareness, with Kinder L2 reading being the variable with the most influence (it is the first one that splits the data).

The second run will cut out Kinder L2 reading and grade 1 L1 reading because these are the measures of reading performance which were different in the first run. The R code is:

```
model=tree(grade1l2reading-kinderl2reading-grade1l1reading~., data=lafrance)
```

The variables that are explanatory in this run are L2 phonemic awareness, Kinder L1 reading, and naming speed, with L2 phonemic awareness being the variable with the most influence because it is the one with the first split.

6 Dewaele and Pavlenko (2001–3)

Use the data file BEQ.Swear.sav and import it into R as beq.swear.

```
library(tree) #if you haven't opened this library yet
model=tree(swear2~., data=beq.swear)
plot(model)
text(model)
```

The explanatory variables are frequency of using the L2 (l2freq, by far the most important variable), the weight swearing carries in the L2 (weight 2), and L2 speaking proficiency (self-rated, found in l2speak).

For looking at the weight swearing carries in the L2, we change the response variable:

```
model=tree(weight2~., data=beq.swear)
plot(model)
text(model)
```

There are many more explanatory variables for this model than for the first one. The explanatory variables are frequency of swearing in L2 (swear2), L2 listening proficiency (l2_comp), the weight swearing carries in the L1 (weight 1), L2 reading proficiency (l2_read), frequency of using the L2 (l2freq), agesec (the age they learned the L2 at), chronological age (age), and frequency of swearing in L1 (swear1).

7.4.5 Application activity: Multiple Regression (Answers for both SPSS and R given for each item)

1 LaFrance & Gottardo (2005)

Use the data file Lafrance5.sav imported into R as `lafrance5`. It has 6 variables in it.

Answers to #1 are not outlined here since you will just follow the steps outlined in the book for SPSS and you should end up with results that look like Table 7.6. For R, use the models for sequential regression listed in section 7.4.5 that begin with this model:

```
model1=lm(grade1l1readingperformance~nonverbalreasoning, data=lafrance5)
```

Then call for a summary of each model in order to get the R, R2 and unstandardized regression coefficients.

```
summary(model1)
```

Here is a summary of the R2 from those models (I will list the Intercept unstandardized regression coefficient only for each model):
Model 1: R2=0.12, Intercept=0.32 ; Model 2: R2=0.32, Intercept=0.78; Model 3: R2=0.60, Intercept=10.01; Model 4: R2=0.47, Intercept=3.87; Model 5: R2=0.67, Intercept=-7.01.

Use the following command to get confidence intervals:

```
confint(model1)
```

2 LaFrance & Gottardo (2005)

Use the data file LafranceGottardo.sav (imported into R as lafrance). It has 9 variables in it (in this file I have imputed values for a few missing cases under Naming Speed).

SPSS Instructions:

To look at linearity, first look at a multiple scatterplot (GRAPHS > LEGACY DIALOGS > SCATTER/DOT, then choose the "Matrix Scatter" box and press "Define"). Put the following variables into the "Matrix Variables" box: G1L2READING , NONVERBALREASONING, WORKINGMEMORY, NAMINGSPEED, KINDERL2READING, L2PHONEMICAWARENESS. Click OK.

For standard regression, go to ANALYZE > REGRESSION > LINEAR. Put Grade 1 L2 Reading (G1L2READING) in the "Dependent" box. Put the following variables into the "Independent" box: NONVERBALREASONING, WORKINGMEMORY, NAMINGSPEED, L2PHONEMICAWARENESS, KINDERL2READING. Leave the Method as "Enter." Open the STATISTICS button and tick "confidence intervals," "casewise diagnostics," "descriptive," "part and partial correlations" and "collinearity diagnostics," besides those which are already ticked. Press "Continue." Open the PLOTS button and put SRESID in the "Y" axis box and ZPRED in the "X" axis box and tick "Normal probability plot." Press "Continue." Open the Save button and check Mahalanobis and Cook's under the "Distances" box. Press OK and run the regression.

R Instructions:

For this case we will use R in the same way as SPSS to do a linear regression (see Section 7.4.5).

Examine linearity by looking at a scatterplot matrix. The fastest way is probably just to use R Commander (make sure lafrance is the active dataset): GRAPHS > SCATTERPLOT MATRIX. Choose these variables: grade1l2reading, kinderl2reading, l2phonemicawareness, naming

speed, and workingmemory (hold down the CTRL key while choosing with the mouse). Leave options alone.

Examine multicollinearity by using R Commander to choose STATISTICS > SUMMARIES > CORRELATION MATRIX. Choose the same variables as above for the scatterplot.

Make the model for the standard regression and get the summary results:

```
model=lm(grade1l2reading~nonverbalreasoning+kinderl2reading
+namingspeed+workingmemory+l2phonemicawareness, data=lafrance)
summary(model)
```

Create standardized regression coefficients:

```
model.standard=lm(scale(grade1l1reading)~scale(nonverbalreasoning) +
scale(workingmemory) + scale(namingspeed) + scale(kinderl2reading) +
scale(l2phonemicawareness), data=lafrance)
(summary(model.standard))
```

For relative importance metrics:

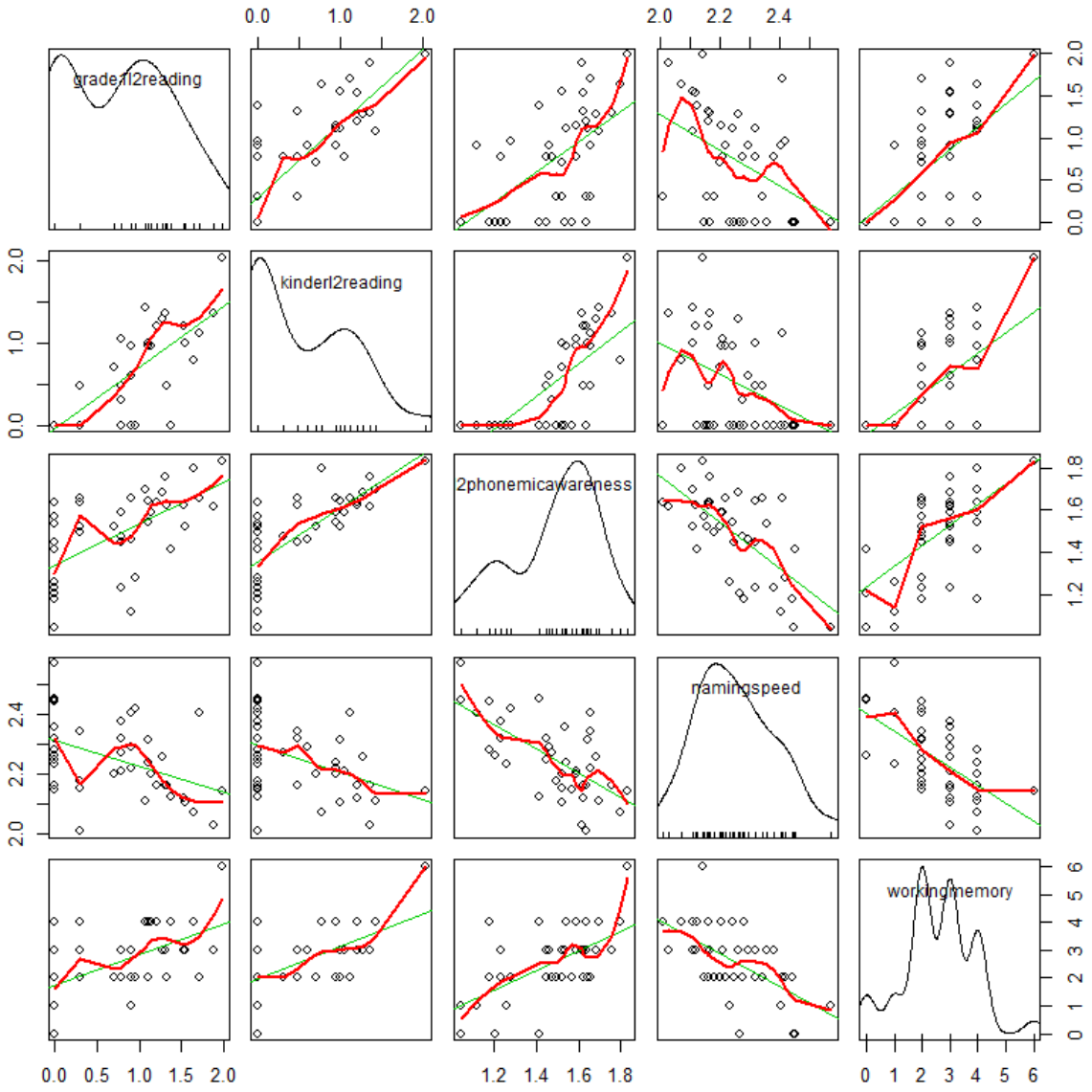
```
library(relaimpo)
calc.relimp(model)
```

For examining assumptions, use these commands:

```
plot(model,cex=1.5)
vif(model)
```

Evaluation:

Linearity: The SPSS multiple scatterplot is too small and hard to see. The R scatterplot, which contains the regression and Loess line, shows more clearly that although the data are not perfect, it is probably safe to assume linearity between the variables (R's scatterplot printed below).



Multicollinearity: Looking at the relations between the response variable (G1L2READING) and the explanatory variables in the Correlations output box. the correlation between KINDERL2READING and G1L2READING is high (r = .807) and indicates multicollinearity. The correlation between KINDERL2READING and L2PHONEMICAWARENESS is also high and

above .70 (r = .712). This indicates that these variables should be taken out but because I am recreating an analysis done by the authors I will leave both terms in the regression.

The regression model: The $R^2 = .672$, which is quite high. Of the individual terms of this equation (G1L2READING ~ NONVERBALREASONING + WORKINGMEMORY + NAMINGSPEED + L2PHONEMICAWARENESS + KINDERL2READING), only KINDERL2READING is statistical (t = 5.21, p < .0005). In their paper, Lafrance and Gottardo (2005) report the standardized coefficients of the 5 terms in Table 3, the 4th section which involves Grade 1 Word Reading L2 and uses the L2 Kindergarten reading performance and L2 phonological awareness (the authors used a hierarchical analysis but I just look at their final model which has all 5 terms). Here they are (and the ones in SPSS are just a bit different from the paper because of the imputed values, while the standardized coefficients in R are quite different):

β = -.05 for non-verbal reasoning, β = .06 for working memory, β = -.16 for naming speed, β = -.06 for L2 phonemic awareness, and β = .79 for KINDERL2READING. For R the standardized coefficients are: β = -.04 for non-verbal reasoning, β = .12 for working memory, β = .03 for naming speed, β = .67 for L2 phonemic awareness, and β = .04 for KINDERL2READING.

Relative importance metrics (sr2): For SPSS, report the squared semipartial correlations (in the box labeled “Coefficients” and the column labeled “Part” : Kinder L2 reading is highest at 51%, next is working memory at 9%, L2 phonemic awareness at 5%, next naming speed at 7%, next L2 phonemic awareness at 5%, and last non-verbal reasoning at 4%. For the R results, the

importance of Kinder L2 reading is highest at 39%, next is L2 phonemic awareness at 11%, next working memory at 9%, next naming speed at 5%, and last non-verbal reasoning at 3%.

Assumptions: For SPSS, at the end of the Coefficients output you will find the VIF column. Here no values are over 5, so presumably this indicates that there is no problem with multicollinearity. The Residuals Statistics output box does not indicate problems with outliers (standardized residuals are not beyond ±3, Cook's distance values are not over 1 and Mahalanobis values over about 15), but the residuals vs. predicted values plot could indicate some heteroscedasticity (values on the right side of the plot are more constrained than values on the left). The P–P plot does show variance away from a straight line, indicating that data may not be normally distributed. For R, the Residuals vs. Fitted plot and Scale-Location plot seem to show the data randomly scattered. For the Normal Q-Q plot there is definitely deviation away from a straight line at the ends of the distribution, indicating a heavy-tailed residual distribution and non-normality, and lastly the Residuals vs. Leverage plot does not seem to show outliers. The VIF shows no values over 5 which indicates that there is no problem with multicollinearity.

3 French and O'Brien (2008)

Use the French & O'Brien Grammar.sav file (imported into R as French). It has 26 variables in it.

SPSS Instructions:

To look at linearity, first look at a multiple scatterplot (GRAPHS > LEGACY DIALOGS > SCATTER/DOT, then choose the "Matrix Scatter" box and press "Define." Put the following

variables into the “Matrix Variables” box: GRAM_1, GRAM2 , INTELLIG , ANWR_1, ENWR_1, L2CONTA. Click OK.

For hierarchical (sequential) regression, go to ANALYZE > REGRESSION > LINEAR. Put Time 2 grammar (GRAM_2) in the “Dependent” box. Put Time 1 grammar (GRAM_1) into the “Independent” box and change the Method to “Stepwise”. Press the NEXT button to indicate that you will enter that variable in the first step. Now put intelligence test scores (INTELLIG) into the “Independent” box and press NEXT. The third step should enter L2CONTA, the fourth ANWR_1, and the last ENWR_1.

Open the STATISTICS button and tick “confidence intervals,” “casewise diagnostics,” “R squared change,” “descriptives” and “collinearity diagnostics.” Press “Continue.” Open the PLOTS button and put SRESID in the “Y” axis box and ZPRED in the “X” axis box and tick “Normal probability plot.” Press “Continue.” Click the SAVE button and check Mahalanobis and Cook’s under the “Distances” box. Press “Continue.” Press OK when back to the LINEAR REGRESSION dialog box and run the regression.

R Instructions:

For this case we will use R in the same way as SPSS to do a linear regression (see Section 7.4.5).

Examine linearity by looking at a scatterplot matrix. The fastest way is probably just to use R Commander (make sure French is the active dataset): GRAPHS > SCATTERPLOT MATRIX. Choose

these variables: anwr_1, enwr_1, gram_1, gram_2, intellig and l2_conta (hold down the CTRL key while choosing with the mouse). Leave options alone and press OK.

Examine multicollinearity by using R Commander to choose STATISTICS > SUMMARIES > CORRELATION MATRIX. Press OK to continue.

Make the model for the hierarchical regression by starting with a model with just one and get the summary results:

```
model1=lm(gram_2~gram_1, data=French)
summary(model1)
model2=lm(gram_2~gram_1+intellig, data=French)
summary(model2)
model3=lm(gram_2~gram_1+intellig+l2_conta, data=French)
summary(model3)
model4=lm(gram_2~gram_1+intellig+l2_conta+anwr_1, data=French)
summary(model4)
model5=lm(gram_2~gram_1+intellig+ l2_conta+anwr_1+enwr_1, data=French)
summary(model5)
```

(By the way, an easier way to write these models especially as they get longer, which is not found in the book chapter but can be seen in the online document “Finding the Best Fit in Multiple Regression” would be this:

```
model2=update(model1,~.+intellig, data=French)
```

The sequence “comma tilde dot minus” after the name of the original model means to include everything in the previous model.)

To get confidence intervals for the unstandardized regression coefficients in model5, use this code:

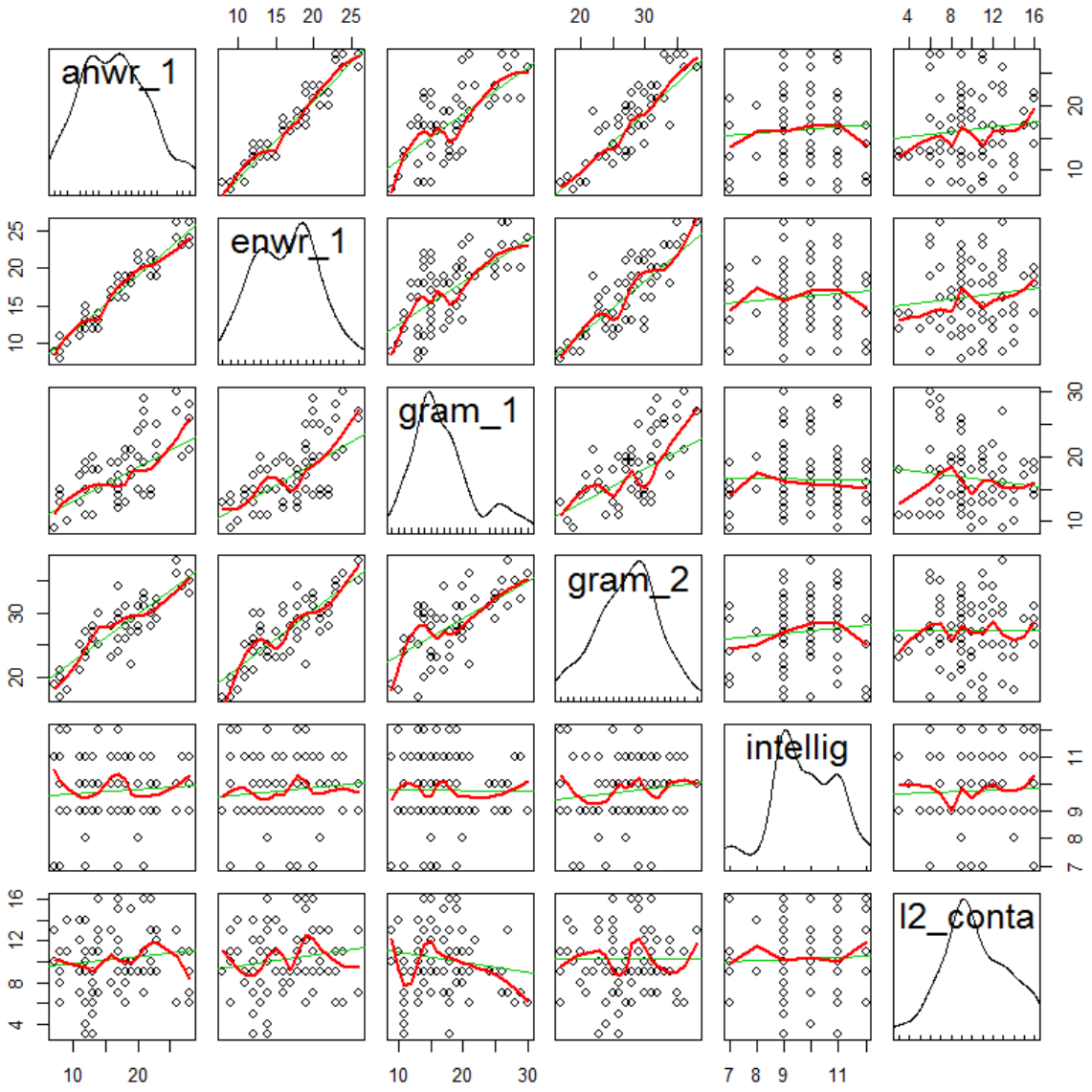
```
confint(model5)
```

For examining assumptions, use these commands with the final model:

```
plot(model5,cex=1.5)
vif(model5)
```

Evaluation:

Linearity: The SPSS multiple scatterplot is too small and hard to see. The R scatterplot, which contains the regression and Loess line, shows more clearly that although the data are not perfect, it is probably safe to assume linearity between the variables (R’s scatterplot printed below).



Multicollinearity: Looking at the relations between the response variable (GRAM_1) and the explanatory variables in the Correlations output box, no correlation are above above .70. For correlations between other variables, the correlation between GRAM_2 and ANWR_1 is high (r = .82) as well as the correlation between GRAM_2 and ENWR_1 (r = .80). Additionally, the correlation between ANWR_1 and ENWR_1 is almost total (r=.95) and indicates that these two

measures are measuring the same thing and one should be taken out of the regression. Because I am recreating an analysis done by French and O'Brien (2008), however, I will leave both terms in the regression.

The regression model: The steps of the hierarchical model are clear in R, but in SPSS, in looking at your output, first look at the box labeled "Variables Entered/Removed" and make sure everything was done in steps the way you wanted (there should be 5 models with one variable entered in each step). SPSS output will give a compact "Model Summary" box for R2 change while for R you will have to search for the information in the "Multiple R-squared" line for each model (and subtract the R2 of the previous model from the former model; for example, for Model 1 R2=0.3029, while for Model 2 R2=.3156, so the R2 change is .0127). The overall R^2 for the model with all 5 variables entered was $R^2 = .688$, adjusted $R^2 = .672$. This explains quite a lot of what is going on! I will give a table with the results for the change in R2 (found in the "Model Summary" box), the unstandardized coefficients for each of the variables in the last model (found in SPSS below the "Model Summary" box in the "Coefficients" box and in R under the "Estimate" column in the summary of Model 5). Notice that in some cases the 95% CIs are quite large, indicating we don't have a very precise estimate for these coefficients!

	R^2 change	Unstandardized coefficients in Model 5	95% CIs for unstandardized coefficients
Time 2 grammar			
Model 1: +Time 1	.303	.045	[-0.11, 0.20]

grammar			
Model 2: +Intelligence	.013	.186	[-0.25, 0.62]
Model 3: +L2 contact	.006	-.132	[-0.31, 0.04]
Model 4: +ANWR1	.363	.546	[0.23, 0.86]
Model 5: +ENWR1	.004	.213	[-0.19, 0.62]

We can compare the strength of the variables by looking at the R^2 change. It is clear that at least entered in this order, Time 1 grammar adds a lot of explanatory value to Time 2 grammar scores, but of even more value is scores on the Arabic non-word test (its R2 change is even higher than that of the Time 1 grammar). The *t*-test shows that the ANWR is the only constituent that is statistical (by the way, French and O'Brien tried reversing the order of the ENWR and ANWR and found that in that case the ENWR received most of the R^2 change (.328) and the ANWR just a little (.038). So it is clear that a measure of phonological memory was the big predictor, and which one it was was not so important, but we could have predicted this from the fact that ENWR and ANWR were highly correlated).

Relative importance metrics (sr2): For a hierarchical (sequential) regression, the sr^2 is the R square change for each variable, so we have already examined it.

Assumptions: In examining regression assumptions, the VIF shows that in the model with all 5 variables, both of the phonological memory tests received VIF values of a little over 10, indicating a problem with multicollinearity, but we already knew that. Given what I said above about reversing the order of the two tests, in order to find the optimal model it would be best to

choose one or the other of the phonological memory tests. In the Residuals Statistics box SPSS output, no standardized residuals are above 3 (or below -3), no Cook's distance scores are above 1, and for Mahalanobis' distance no scores are above 15, so we do not seem to have any problems with outliers. For normality, looking at the P–P plot, there appears to be a very good fit of the data to the line, indicating the residuals are normally distributed. For looking at the homoscedasticity requirement, the scatterplot of residuals vs. predicted values does not show any evidence of data being more constricted on one side over another. This is quite a clean dataset that satisfies all of the assumptions of regression, except for the multicollinearity! For the diagnostic plots in R, the Residuals vs. Fitted plot and Scale-Location plot seem to show the data randomly scattered. For the Normal Q-Q plot there is slight deviation away from a straight line at one of the distribution, but it appears to be a fairly good fit to the line, indicating a normal distribution of residuals. Lastly the Residuals vs. Leverage plot does not seem to show outliers.

4 Howell (2002) data

Use the HowellChp15Data.sav file (imported into R as howell). It has 6 variables in it.

SPSS Instructions:

To look at linearity, first look at a multiple scatterplot (GRAPHS > LEGACY DIALOGS > SCATTER/DOT, choose the "Matrix Scatter" box and press "Define"). Put the all 6 variables into the "Matrix Variables" box.

For a standard regression go to ANALYZE > REGRESSION > LINEAR. Put OVERALL in the "Dependent" box and all of the other variables in the "Independent" box. Leave the Method as

"Enter." We won't examine regression assumptions yet on this step, so just press OK and run the regression.

R Instructions:

Examine linearity by looking at a scatterplot matrix. Use R Commander (make sure howell is the active dataset): GRAPHS > SCATTERPLOT MATRIX. Choose all of the variables (the fastest way is to click the first variable with your mouse then hold down the SHIFT key while choosing the last variable with the mouse). Leave options alone and press OK.

Examine multicollinearity by using R Commander to choose STATISTICS > SUMMARIES > CORRELATION MATRIX. Choose all of the variables (as explained in the previous paragraph, this is faster using the SHIFT key than the CTRL key).

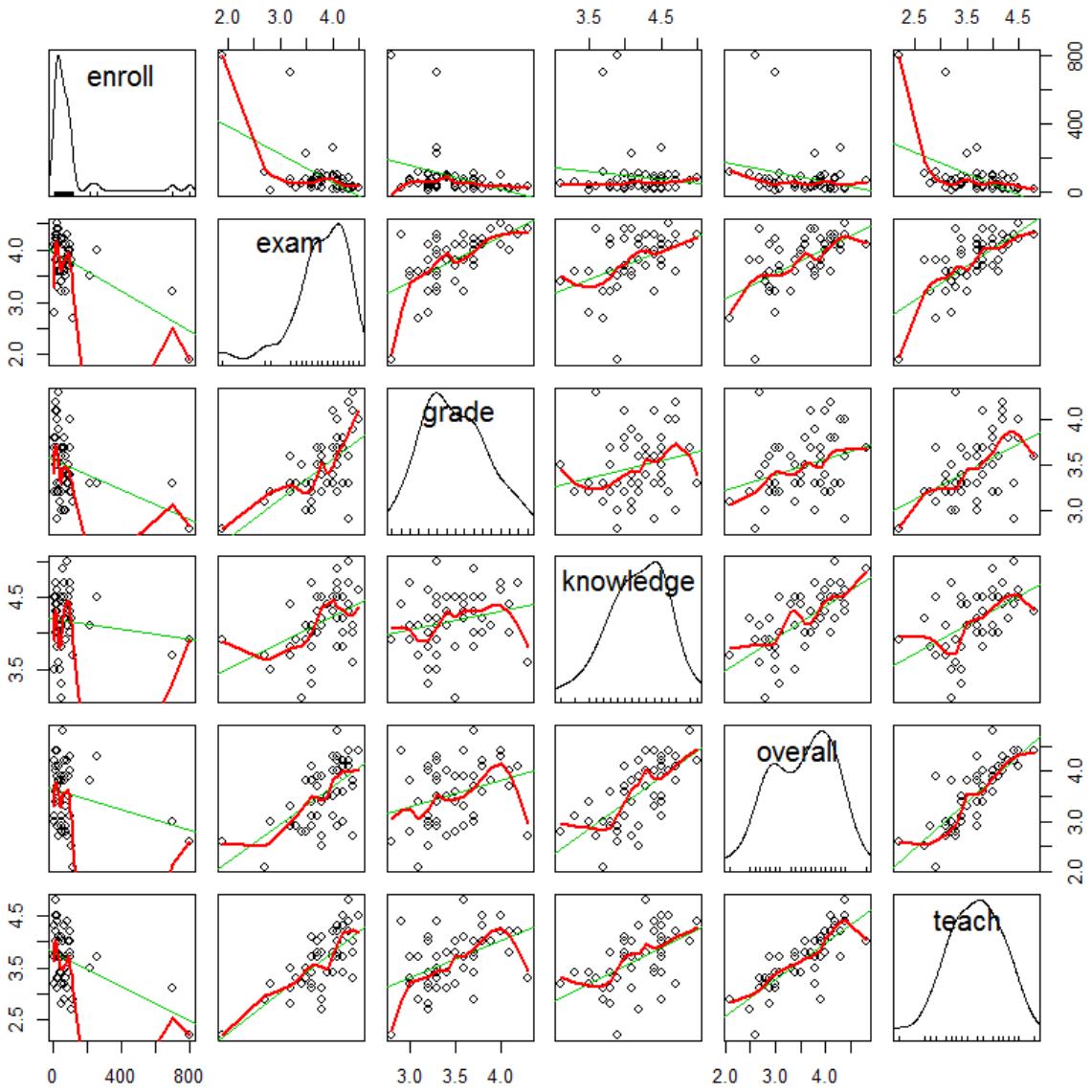
Make the model for the standard regression and get the summary results:

```
model=lm(overall~enroll+exam+grade+knowledge+ teach,data=howell)
summary(model)
```

We won't examine regression assumptions yet on this step, so just press OK and run the regression.

Evaluation:

Linearity: The R scatterplot, printed below, shows that all data may have a linear relationship with OVERALL except for ENROLL, which seems to be a vertical line with a few outliers. ENROLL may not be a good variable to include in the regression because of its strange distribution.



Multicollinearity: Looking at the relations between the response variable (OVERALL) and the explanatory variables in the Correlations output box, the correlation between OVERALL and TEACH is high (r = .80) and indicates multicollinearity. The correlation between OVERALL and KNOWLEDGE is not over .70 but is close (r = .68). Other high intercorrelations are found between TEACH and EXAM (r = .72). In our subsequent step we'll get rid of some of these variables that are highly intercorrelated.

The regression model: This first model with all of the variables explains $R^2 = .76$ of the variance in overall scores, a large amount; however, the statistical factors were Teach and Knowledge only (these are the only ones with a *t*-value probability under *p*=.05).

Second Regression:

Run another regression with just Teach and Knowledge as the two explanatory factors.

SPSS Instructions:

Same steps as above, but just enter the two explanatory variables. Now open up the additional buttons to look at regression assumptions: Open the STATISTICS button and tick "confidence intervals," "casewise diagnostics," "descriptive," "part and partial correlations" and "collinearity diagnostics," besides those that are already ticked. Press "Continue." Open the PLOTS button and put SRESID in the "Y" axis box and ZPRED in the "X" axis box and tick "Normal probability plot." Press "Continue." Open the Save button and check Mahalanobis and Cook's under the "Distances" box. Press "Continue." Press OK and run the regression.

R Instructions:

Repeat the steps for linearity and multicollinearity as above with only the two explanatory variables.

```
model=lm(overall~knowledge+ teach,data=howell)
summary(model)
```

Get information on sr2:

```
library(relaimpo)
calc.relimp(model)
```

Examine regression assumptions:

```
plot(model,cex=1.5)
vif(model)
```

Evaluation:

Linearity: The multiple scatterplot shows that the data are linear.

Multicollinearity: The correlation between OVERALL and TEACH is still problematic (r = .80) and indicates multicollinearity.

The regression model: The $R^2 = .74$, which means that 74% of the variability in course ratings depends on teaching skills of the instructor and the instructor's knowledge of the subject matter (hm. . . this sounds very nice but I'm assuming it's a made-up dataset as this is not the usual finding! There's more evidence that factors such as class size, time of day, and the grading reputation of the professor are more explanatory than teaching skills or instructor's knowledge!). All factors in this regression equation are now statistically explanatory.

The regression equation is:

Overall course rating = -1.30 + .54 (instructor's knowledge of subject matter) + .71(teaching skills of the instructor)

This model can be obtained by looking at the constant and the unstandardized coefficients in the "Coefficients" box of the SPSS output or at the "Estimate" column under "Coefficients" in the R output.

Relative importance metrics (sr2): For SPSS, report the squared semipartial correlations (in the box labled "Coefficients" and the column labeled "Part": the importance of Teaching skills of the instructor is highest at 52%, followed by instructor's knowledge at 30%. For the R results from the reliampo package, the importance of Teaching skills of the instructor is highest at 46%, followed by instructor's knowledge at 28%.

Assumptions: Overall, the VIF does not indicate a problem with multicollinearity, residuals statistics, Cook's and Mahalanobis do not indicate a problem with outliers or influence points, the P–P plot looks good indicating a normal distribution, and there is no clear heteroscedasticity in the residuals vs. predicted fit plot. Overall, this model seems to satisfy regression assumptions quite well.

5 Dewaele and Pavlenko (2001–2003) data

Use the BEQ.Swear file, imported into R as beq.swear.

SPSS Instructions:

To look at linearity, first look at a multiple scatterplot (GRAPHS > LEGACY DIALOGS > SCATTER/DOT, choose the "Matrix Scatter" box and press "Define"). Put the following variables into the "Matrix Variables" box: SWEAR2, L2FREQ, WEIGHT2, L2_COMP and L2SPEAK.

For this plot, because it looks like there is no pattern to the data because the points are so discrete, add regression lines to the data (open the Chart Editor, go to menu choice ELEMENTS > ADD FIT LINE AT TOTAL button and then CLOSE).

For a standard regression go to ANALYZE > REGRESSION > LINEAR. Put SWEAR2 in the "Dependent" box and L2_COMP, L2FREQ, L2SPEAK, SWEAR2, WEIGHT2, in the "Independent" box. Leave the Method as "Enter." We won't examine regression assumptions yet on this step, so just press OK and run the regression.

R Instructions:

Examine linearity by looking at a scatterplot matrix. Use R Commander (make sure beq.swear is the active dataset): GRAPHS > SCATTERPLOT MATRIX. Choose L2_COMP, L2FREQ, L2SPEAK, SWEAR2 and WEIGHT2 (hold down the CTRL key while choosing with the mouse). Leave options alone and press OK.

Examine multicollinearity by using R Commander to choose STATISTICS > SUMMARIES > CORRELATION MATRIX. Choose L2_COMP, L2FREQ, L2SPEAK, SWEAR2 and WEIGHT2 (hold down the CTRL key while choosing with the mouse). Press OK.

Make the model for the standard regression and get the summary results:

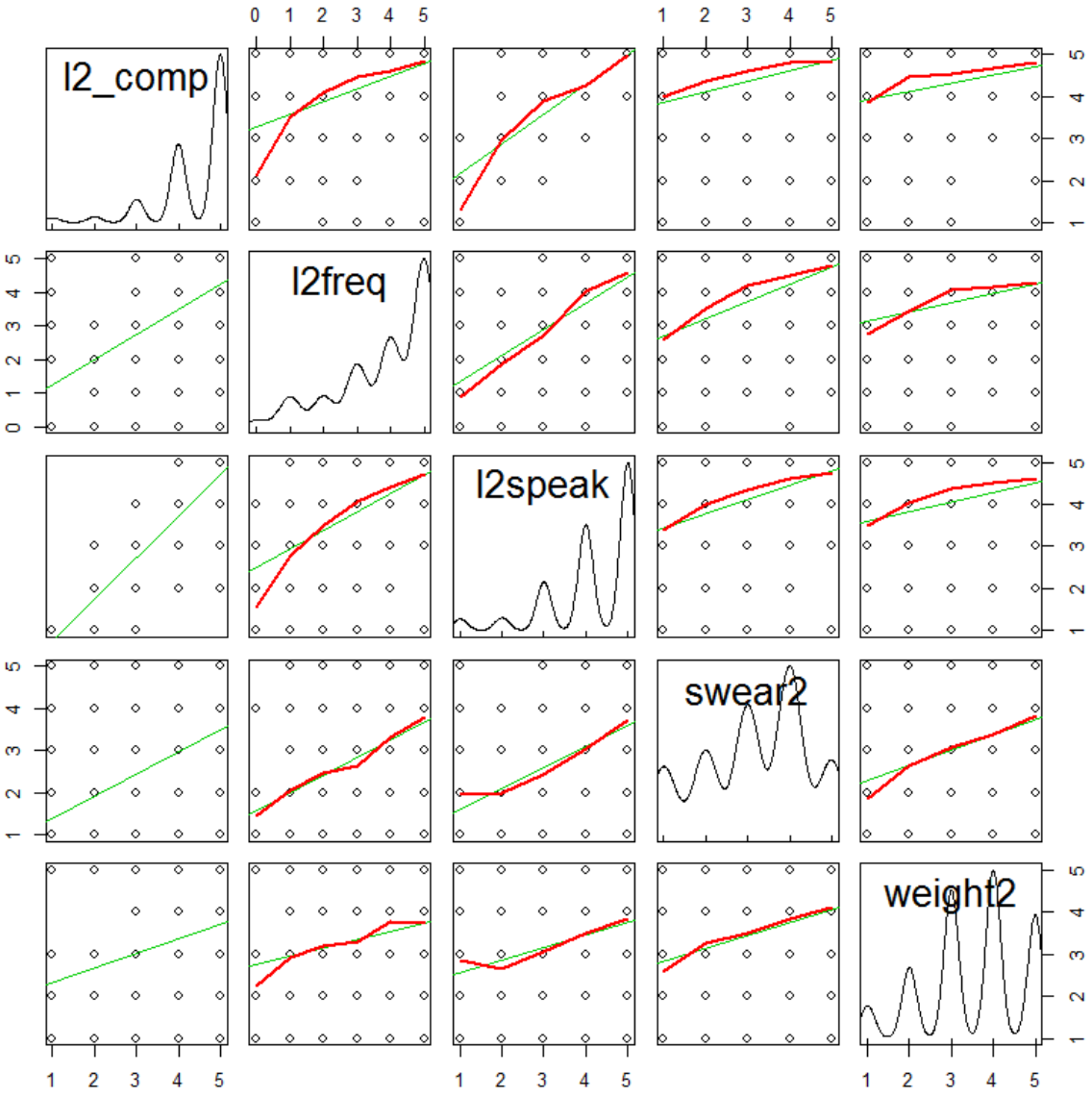
```
model=lm(swear2~l2_comp+l2freq+weight2+l2speak,data=beq.swear)
summary(model)
```

We won't examine regression assumptions yet on this step, so just press OK and run the regression.

Evaluation:

Linearity: The scatterplot matrix of the intersection of SWEAR2 seems to show a random scattering of the variables pretty much over the entire graph, which would violate the assumption of linearity. However, since the points are discrete and not jittered so we can see their frequency, it could be that there are indeed linear trends that are not apparent in the scatterplot. In other

words, there may be many *more* points along a linear line in the plot, but because we can only see 25 discrete points on the scatterplot, we cannot tell how often each point is chosen. The Loess lines (which in R, shown below, appear on all plots except those that are combined with the variable of L2 comprehension) show linear trends, so perhaps it is OK to proceed!



In the regression, put SWEAR2 in the "Dependent" box and the other variables in the "Independent" box. Leave the Method as "Enter". Open the same buttons and tick the same boxes as described for #2.

Multicollinearity: Looking at the relations between the response variable (SWEAR2) and the explanatory variables in the Correlations output box, the correlations seem to be of acceptable effect size, but not too high so as to pose a problem. For other variables, the correlation between L2SPEAK and L2_COMP is high (r = .84) and indicates multicollinearity.

The regression model: The $R^2 = .29$, which is fairly low. The statistical factors in the model were weight given to swearing in L2 (WEIGHT2), L2 speaking ability (L2SPEAK) and L2 frequency usage (L2FREQ) (these are the only ones with a *t*-value probability under $p = .05$).

Second Regression:

Run another regression but take out the variable of L2_Comp as an explanatory factor.

SPSS Instructions:

Same steps as above, without L2_Comp. Now open up the additional buttons to look at regression assumptions: Open the STATISTICS button and tick "confidence intervals," "casewise

diagnostics," "descriptive," "part and partial correlations" and "collinearity diagnostics," besides those which are already ticked. Press "Continue." Open the PLOTS button and put SRESID in the "Y" axis box and ZPRED in the "X" axis box and tick "Normal probability plot." Press "Continue." Open the Save button and check Mahalanobis and Cook's under the "Distances" box. Press "Continue." Press OK and run the regression.

R Instructions:

Repeat the steps for linearity and multicollinearity as above without L2_Comp.

```
model=lm(swear2~l2freq+weight2+l2speak,data=beq.swear)
summary(model)
confint(model)
```

Get information on sr2:

```
library(relaimpo)
calc.relimp(model)
```

Examine regression assumptions:

```
plot(model,cex=1.5)
vif(model)
```

Evaluation:

Linearity: The multiple scatterplot is still quite strange-looking but we will proceed with the analysis.

Multicollinearity: Without the variable of L2_COMP there are no problems of multicollinearity.

The regression model: The $R^2 = .29$, which means the model explains 29% of the variance in swearing frequency which is a goodly amount but there is room for more explanation. All variables in the regression model are now statistically significant.

The regression equation is:

Swearing frequency = .41 + .23(Weight given to swearing in L2) + .21(L2 speaking ability) + .29 (L2 frequency of use).

This model can be obtained by looking at the constant and the unstandardized coefficients in the "Coefficients" box of the SPSS output or at the "Estimate" column under "Coefficients" in the R output. The 95% CIs for these coefficients are:

L2 frequency [0.22, 0.35]

Weight in L2 [0.17, 0.30]

L2 speaking [0.12, 0.29]

These CIs are not too wide, probably because of the very large sample size.

Relative importance metrics (sr2): For SPSS, report the squared semipartial correlations (in the box labeled “Coefficients” and the column labeled “Part”): the importance of the frequency with which a person uses their L2 is highest at 25%, followed by the weight a person gives to swearing in the L2 at 24%, and last in importance is L2 speaking skills at 16%. For the R results from the reliampo package, the importance of the frequency with which a person uses their L2 is highest at 13%, followed by L2 speaking skills at 9%, and last in importance is the weight a person gives to swearing in the L2 at 7%.

Assumptions: The VIF does not indicate a problem with multicollinearity. For SPSS, the Residuals Statistics output box could indicate problems with outliers (standardized residuals are not beyond ±3, Cook’s distance values are not over 1 but Mahalanobis values are quite high). Additionally, the residuals vs. predicted values plot indicates heteroscedasticity (there seem to be constraints on both sides of the graph). The P–P plot shows some slight variance away from a straight line in the middle of its distribution, but it is still fairly close to the line, so it is probably OK.

For R, the Residuals vs. Fitted plot and Scale-Location plot show the data is constrained in a certain way on both sides of the graph, indicating heteroscedasticity. For the Normal Q-Q plot there is a lot of data but it seems to mostly follow the straight line. Lastly the Residuals vs. Leverage plot does not seem to show outliers.

6 Larson-Hall (2008)

Use the data file LarsonHall2008.sav, imported into R as lh2008.

SPSS Instructions:

To look at linearity, first look at a multiple scatterplot (To look at linearity, first look at a multiple scatterplot (GRAPHS > LEGACY DIALOGS > SCATTER/DOT, choose the "Matrix Scatter" box and press "Define"). Put the following variables into the "Matrix Variables" box: TOTALHRS, RLWSCORE, GJTSCORE, APTSCORE.

For hierarchical (sequential) regression, go to ANALYZE > REGRESSION > LINEAR. Put GJTSCORE in the "Dependent" box. We want to try out at last three different permutations of the explanatory variables. I indicated in the instructions to put each variable first once, and methods could vary, but just make sure to change the Method to "Stepwise" and then enter each variable into the "Independent" box and press the NEXT button to indicate that you will enter that variable in the first step, then put each variable in until you have done 3 blocks.

Open the STATISTICS button and tick "confidence intervals," "casewise diagnostics," "R squared change," "descriptives" and "collinearity diagnostics." Press "Continue." Open the PLOTS button and put SRESID in the "Y" axis box and ZPRED in the "X" axis box and tick "Normal probability plot." Press "Continue." Click the SAVE button and check Mahalanobis and Cook's under the "Distances" box. Press "Continue." Press OK when back to the LINEAR REGRESSION dialog box and run the regression.

To repeat this with different variables first, open up the regression dialog box. You could redo the regression by pressing the RESET button, but then you would have to open up all the sub-dialog boxes as well and tick everything again. It's probably easiest to just trace back your steps and move each variable out from the 3 blocks you created, then move them back in again in a different order.

R Instructions:

Examine linearity by looking at a scatterplot matrix. Use R Commander (make sure lh2008 is the active dataset): GRAPHS > SCATTERPLOT MATRIX. Choose these variables: aptscore, gjtscore, rlwscore, and totalhrs (hold down the CTRL key while choosing with the mouse). Leave options alone and press OK.

Examine multicollinearity by using R Commander to choose STATISTICS > SUMMARIES > CORRELATION MATRIX.

For the regression models, in R it is quite easy to examine all possible permutations of the data in order to see which has the highest R2, so I will write down all of the possible models:

```
model1a=lm(gjtscore~rlwscore+totalhrs+aptscore, data=lh2008)
summary(model1a)
model1b=lm(gjtscore~rlwscore+ aptscore + totalhrs, data=lh2008)
summary(model1b)
model2a=lm(gjtscore~totalhrs+aptscore+ rlwscore, data=lh2008)
summary(model2a)
```

```
model2b=lm(gjtscore~totalhrs+ rlwscore+ aptscore, data=lh2008)
summary(model2b)
model3a=lm(gjtscore~ aptscore+ rlwscore+ totalhrs, data=lh2008)
summary(model3a)
model3b=lm(gjtscore~ aptscore+ totalhrs+ rlwscore, data=lh2008)
summary(model3b)
```

I also want to see which model will give the highest change in R2 for the three explanatory variables, so I can run the following code with each model:

```
calc.relimp(model)
```

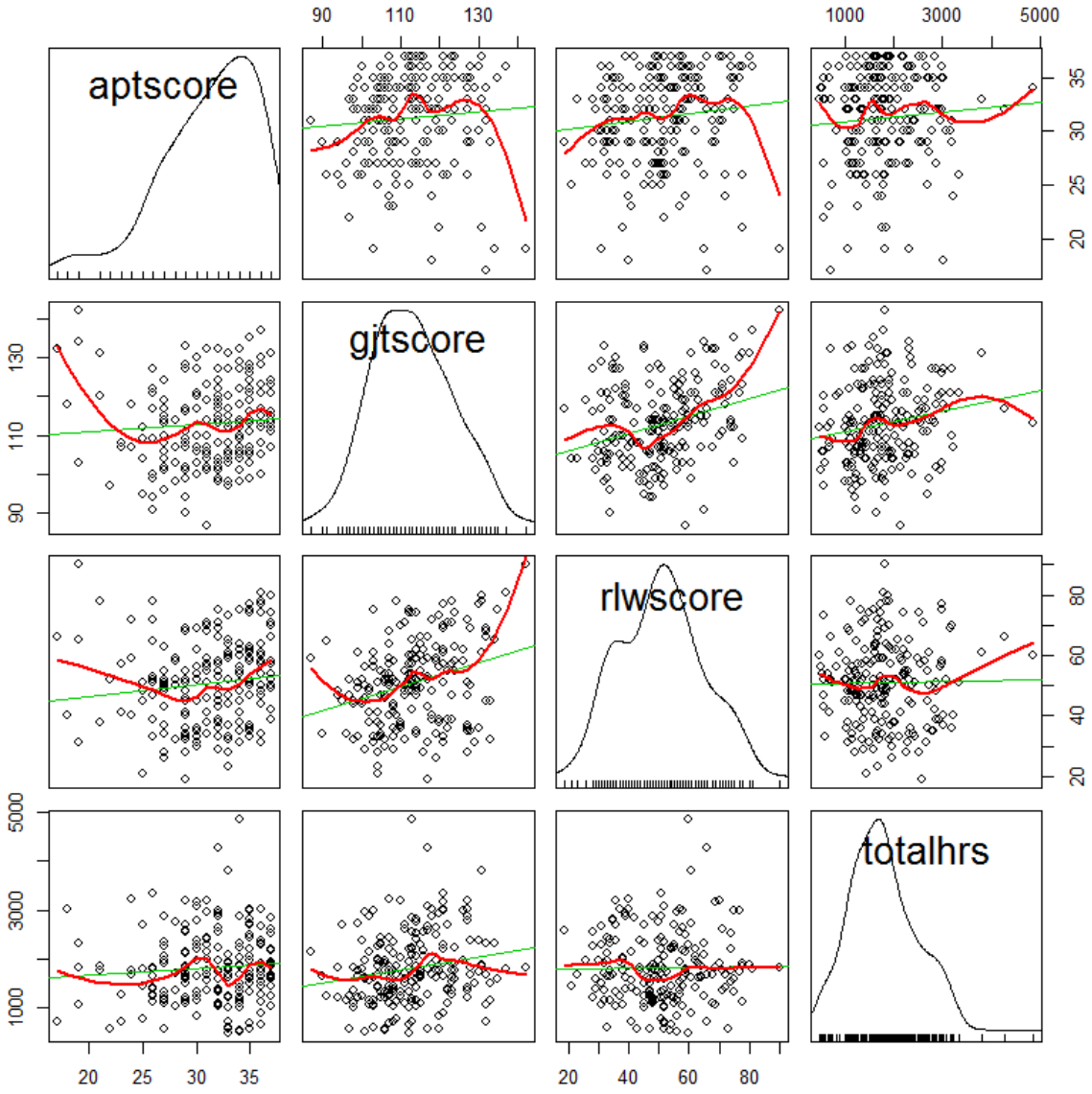
I don't think I need to examine all of the models for assumptions, just the one that seems the best. Because I'm doing hierarchical modeling I don't need to get the sr2 data. I'll look at the model assumptions with this syntax:

```
plot(model,cex=1.5)
vif(model)
```

Evaluation:

Linearity: The SPSS multiple scatterplot is too small and hard to see. The R scatterplot, which contains the regression and Loess line, shows more clearly that there are several significant departures from linearity in the intersections of variables *not* related to Total Hours (R's

scatterplot printed below). A curve may better describe the data than a line, but for now we will stick with a line.



Multicollinearity: Looking at the relations between the response variable (GJTSCORES) and the explanatory variables in the Correlations output box, correlations are rather low, and certainly

none are close to being too high. As for the correlations between other variables, the same applies. In fact, we may be worried that correlations are too low overall to find any interesting results.

The Regression Model:

SPSS results:

With this order (TOTALHRS, RLWSCORE, APTSCORE) the R^2 = .12 (fairly low). The R2 change is .034 for hours, .088 for RLW test, and .001 for aptitude.

With this order (RLWSCORE, APTSCORE, TOTALHRS) the R^2 = .12. The R2 change is .090 for RLW test, .002 for aptitude, and .031 for hours of input.

With this order (APTSCORE, RLWSCORE, TOTALHRS), the R^2 = .12. The R2 change is .034 for total hours and .088 for RLW test. Aptitude doesn't even get included when it is first!

The R^2 doesn't really change depending on the order, but the R^2 change does vary depending on the order it is entered. Aptitude gets very little R^2 change, but most when it is second after RLW. RLW is the strongest variable and it gets the most R^2 change when it comes first. RLW score gets the most when it is first.

R Results:

Trying different orders shows that the R^2 doesn't really change depending on the order. Trying the relaimpo package shows that the strength of each variable also does change depending on

the order that it's entered in (remember that the strength of the variable as determined by the relaimpo package is the sr^2 value). This is a difference between SPSS and R. The results using this method show that the strongest explanatory variable is the RLWscore, which explains 9% of the variance, followed by the total hours at 3% and aptitude explains nothing.

Assumptions:

For SPSS, most of the assumptions look good. The Mahalanobis values are over 15, which shows a problem with outliers, but Cook's distance does not show any problems. The P-P plot looks normal and the residuals vs. predicted values plot does not indicate any problem with heteroscedasticity. For R I just used model1a, since everything turned out to be the same, and did not notice any problems with any of the diagnostic graphics.

Bibliography

French, L. M., & O'Brien, I. (2008). Phonological memory and children's second language grammar learning. *Applied Psycholinguistics, 29*(3), 463–487.