


Answers to Application Activities in Chapter 6

6.3.3 Application Activities with Scatterplots (Answers for both SPSS and R given for each item)

1 DeKeyser (2000)

Use the data file Dekeyser2000.sav.

SPSS Instructions:

Choose **GRAPHS > LEGACY DIALOGS > SCATTER/DOT**, then **SIMPLE SCATTER**. Click the **DEFINE** button. Enter **GJTSCORE** into the “Y Axis” box and **AGE** into the “X Axis” box. Click **OK** and a graph appears (if you have any other output already, you may have to scroll down to the bottom to see your new output). To insert regression line, click the graph twice so that **Chart Editor** appears. Then choose **ELEMENTS > FIT LINE AT TOTAL** or click the button that looks like this:  on the lowest line of the menu bar. I don't like the label for the regression line so I recommend ticking off the box at the bottom of the dialogue box that says “Add label to line”, but you can see whether you like it yourself or not. Close the Properties dialogue box (Linear is already chosen). You have added the regression line. To add the Loess line, repeat the procedure but this time tick the choice Loess in the Properties dialogue box (it may already be chosen).

To change the look of the two lines, while still in the **Chart Editor**, click slowly on the line until it is highlighted in yellow, then quickly double-click it and the Properties dialogue box will open. Click the **Lines** tab and you can change the weight, style, and color of the line. Click **Apply** to apply the changes, and if you are happy with the changes click **Close**.

To change additional dimensions, you can change the x- and y-axis labels in Chart Editor by clicking once on the label until a blue box appears around it, then clicking again to be able to type. You can change the y-axis limits by double-clicking on a number along the y-axis, then going to the NUMBER FORMAT tab. Put zero in the Decimal Places box to get rid of decimals. You can also go to the Labels & Ticks (previously called Ticks & Grids in version 15.0) tab and choose to display major or minor ticks Inside instead of Outside as is done by default.

R Instructions:

Open the R program. Type:

```
library(Rcmdr)
```

at the prompt line. R Commander should open up.

Either import the DeKeyser2000.sav SPSS file as `dekeyser`, or click the Dataset button to open the DeKeyser file that you have already imported and make it the active dataset.

Choose the menu sequence GRAPHS > SCATTERPLOT.

In the “Data” tab, choose AGE for one variable and GJTSCORE for the other.

In the “Options” tab, four boxes are ticked under Plot Options: “Marginal boxplots,” “Least-squares line,” “Smooth line” and “Show spread.” You can tick off them or leave on—the Least-

squares line will give you the regression line and the Smooth line will give you the Loess line, so leave those ticked at least.

Below the “Plot Options” you will see the “Identify Points” option. It is set to identify extreme points automatically, and the number of points to identify is 2.

If you would like to add axis labels or a Graph title, the right-hand side of the “Options” box gives you the place to do this.

R code is:

```
scatterplot(gjtscore~age, reg.line=lm, smooth=TRUE, spread=TRUE, id.method='mahal',  
id.n=2, boxplots='xy', span=0.5, data=dekeyser)
```


Evaluation of Graphic:

The Loess line does not follow the regression line exactly but the data seem mostly linear so a correlational analysis seems warranted for this data. However, the Loess line may hint that scores would not continue to go down indefinitely, as it goes flat in the 20–30 age range where most of the adult data is, and there’s very little data to say much between 14 and 20, as DeKeyser points out in his article. Point 48 is clearly an outlier in the data.

2 DeKeyser (2000)

Use the data file Dekeyser2000.sav.

SPSS Instructions:

Choose **GRAPHS > LEGACY DIALOGS > SCATTER/DOT**, then **SIMPLE SCATTER**. Click the **DEFINE** button. Enter **GJTSCORE** into the “Y Axis” box and **AGE** into the “X Axis” box (if you have done Exercise #1 these will already be entered). Additionally, move the **Status** variable into the **Set Markers by** box. Click **OK** and a graph appears (if you have any other output already, you may have to scroll down to the bottom to see your new output). To insert regression line, click the graph twice so that **Chart Editor** appears. Then choose **ELEMENTS > FIT LINE AT SUBGROUPS** or click the button that looks like this:  on the lowest line of the menu bar. Close the **Properties** dialogue box (**Linear** is already chosen). You have added the regression line. To add the **Loess** line, repeat the procedure but this time tick the choice **Loess** in the **Properties** dialogue box (it may already be chosen). The **Loess** line with 50% smoothing may be too detailed in these smaller groups; you might want to set the smoothing to a larger dimension, such as 70% or 80% to get a smoother line.

To change the look of the two lines, while still in the **Chart Editor**, click slowly on the line until it is highlighted in yellow, then quickly double-click it and the **Properties** dialogue box will open. Click the **Lines** tab and you can change the weight, style, and color of the line. Click **Apply** to apply the changes, and if you are happy with the changes click **Close**.

R Instructions:

Open the R program. Type:

```
library(Rcmdr)
```

at the prompt line. R Commander should open up.

With the DeKeyser data as the active dataset open in R Commander (see Exercise #1 if you have not done this yet), choose the menu sequence **GRAPHS > SCATTERPLOT**.

In the “Data” tab, choose AGE for one variable and GJTSCORE for the other, and then open the “Plot by groups . . .” button. You should pick the STATUS variable and press OK (leave the box that says “Plot lines by group” checked!). If you don’t see the STATUS variable when you open the “Plot by groups . . .” button, change the STATUS variable to “character” instead of “numeric” with this command:

```
dekeyser$STATUS=factor(dekeyser$STATUS)
```

In the Options tab tick off “Marginal boxplots” and “Show spread.” The Loess line with 50% smoothing may be too detailed in these smaller groups; you might want to change the slider in “Span for smooth” to a larger dimension, such as 70% or 80% to get a smoother line. Press OK and the graph appears.

The R code for this is:

```
scatterplot(gjtscore~age | status, reg.line=lm, smooth=TRUE, spread=FALSE,  
id.method='mahal', id.n=2, boxplots=FALSE, span=0.7, by.groups=TRUE,  
data=dekeyser)
```

Evaluation:

What kind of relationship do age and test scores have for each of these groups?

Regression lines on each group of variables show that there is a slightly negative slope on the group of participants age 15 or less, but the regression line on the older group appears to be flat, indicating no decline on GJT scores with age over 15. As with the entire group above, we note that the Loess line does not follow the regression line exactly but the data seem mostly linear (especially with a wider smoother) so a correlational analysis seems warranted for this data. Point 48 is clearly an outlier.

3 Flege, Yeni-Komshian, and Liu (1999)

Use the FlegeYeniKomshianLiu.sav file. Import into R as **fyl** (Note: This is different from the FYL file you imported in Chapter 2!).

SPSS Instructions:

In the SIMPLE SCATTERPLOT dialogue box (see the answer to Exercise #1 for how to get to it), put the variable PRONENG in the Y Axis box and LOR in the X Axis box, and click OK. To insert the Loess line, open the Chart Editor and choose ELEMENTS > FIT LINE AT TOTAL. In the Properties dialogue box, choose the Loess line (click off “Attach label to line” if, like me, you don’t like this label on your line) and click APPLY, then CLOSE.

R Instructions:

Open the R program. Type:

```
library(Rcmdr)
```

at the prompt line. R Commander should open up.

Make sure `fyl` is the active dataset in R Commander (see Exercise #1 if you have not done this yet), then choose the menu sequence **GRAPHS > SCATTERPLOT**. In the “Data” tab, choose `LOR` for the x-variable and `PRONENG` for the y-variable. In the “Options” tab tick off “Marginal boxplots,” “Least-squares line” and “Show spread” so only the “Smooth line” is ticked (this is the Loess line).

The R code for this is:

```
scatterplot(proneng~lor, reg.line=FALSE, smooth=TRUE, spread=FALSE,  
id.method='mahal', id.n=4, boxplots=FALSE, span=0.5, data=fyl)
```


Evaluation: The Loess line shows an upward correlation among the data until about 15 or 16 years of residence, then becomes basically flat. There may be some outliers in the data (R will mark 4 of the most extreme points automatically).

4 Larson-Hall (2008)

Use the `LarsonHall2008.sav` file. Import into R as `lh2008`.

SPSS Instructions:

In the SIMPLE SCATTERPLOT dialogue box (see Exercise #1 answers for how to get it), put the variable GJTSCORE in the Y Axis box and TOTALHRS in the X Axis box, and click OK. To insert regression line, click the graph twice so that Chart Editor appears. Then choose ELEMENTS > FIT

LINE AT TOTAL or click the button that looks like this:  on the lowest line of the menu bar.

Close the Properties dialogue box (Linear is already chosen). You have added the regression line.

To add the Loess line, repeat the procedure but this time tick the choice Loess in the Properties dialogue box (it may already be chosen).

To change the type or color of the lines (which you would want to do if you were publishing this graphic, in order to distinguish the two lines), see instructions in the answer to Exercise #1. SPSS does not contain any algorithm for systematically identifying outliers in the scatterplot, but with your naked eye you might suspect the two points to the farthest right (at about 4000 hours and 5000 hours) could be outliers.

R Instructions:

Open the R program. Type:

```
library(Rcmdr)
```

at the prompt line. R Commander should open up.

With the LarsonHall2008.sav file imported as **lh2008** and the active dataset in R Commander (see Exercise #1 if you have not done this yet), choose the menu sequence GRAPHS > SCATTERPLOT.

Pick **totalhrs** for the x-variable, and **gjtscore** for the y-variable. In the “Options” tab tick off “Marginal boxplots” and “Show spread” so only the “Least-squares line” and “Smooth line” are ticked.

I asked you to identify outliers so think about the “Identify Points” area on the “Options” tab. We want to let the computer tell us which points are outliers according to a mathematical algorithm, so leave the tick on “Automatically,” but let’s choose to identify 4 outliers, so click up to 4 where it says “Number of points to identify”. Press OK.

The code for this is:

```
scatterplot(gjtscore~totalhrs, reg.line=lm, smooth=TRUE, spread=FALSE,  
id.method='mahal', id.n=4, boxplots='xy', span=0.5, data=lh2008)
```

Evaluation:

The data points seem fairly scattered about and do not closely congregate around a line. However, the Loess line and regression line are fairly close together and do trend slightly upwards, so it could be concluded there is a slight linear relationship. From the SPSS scatterplot we can see that $R^2 = .03$, meaning there is little covariance between the variables, and the regression line is essentially flat, even though it does look like it trends upward on the graph. Such a small amount of R^2 means there really is no relationship between the total number of hours of English studied and scores on the GJT test.

As for outliers, the R analysis that was asked to identify 4 extreme points showed the two points to the farther right (cases 200 and 83) as outliers, along with case 199 (above and to the left of the two right-most points) and case 98 (the one point above 140 on the gjtscore).

5 Larson-Hall (2008)

SPSS Instructions:

In the SIMPLE SCATTERPLOT dialogue box (see Exercise #1 answers for how to get it), put the variable GJTSCORE in the Y Axis box, TOTALHRS in the X Axis box, and ERLYEXP in the Set Markers By box, then click OK. Insert regression line and Loess line (see earlier problems for explanation of how to do this).

R Instructions:

Use the LarsonHall2008.sav file. Import as **lh2008** into R.

Same procedure as in #4, but this time click the “Plot by Groups” button and choose **erlyexp**.

R code for this is:

```
scatterplot(gjtscore~totalhrs|erlyexp, reg.line=lm, smooth=TRUE, spread=FALSE,  
id.method='mahal', id.n=2, boxplots=FALSE, span=0.5, by.grups=TRUE, data=lh2008)
```

Evaluation:

Data still appear to be randomly distributed among each group (it is hard to separate these groups out graphically). The Loess and regression lines seem to have some significant differences from each other. Specifically, the data for the early learners seem to have large disgressions from linearity, and may be better modeled by a different kind of function. The data for the later learners seems close enough to linear to model it linearly.

Also, the line for the early learners' group is steeper than the line for the later learners' group (the SPSS outputs lists the R squared for the early learners as .07 and for the later learners as .02, meaning that there is more covariance between total hours and score for those who started early than for those who started later).

6 Dewaele and Pavlenko (2001–2003)

Use the BEQ.Swear.sav file. Import into R as **beq.swear**.

SPSS Instructions:

In the SIMPLE SCATTERPLOT dialogue box (see Exercise #1 answers for how to get it), put the variable L2SPEAK in the Y Axis box and AGESEC in the X axis box. Because the L2SPEAK variable is on a 1 to 5 point scale only, the points look a little strange lined up in rows and it is hard to see trends. Insert regression line (see previous answers for an explanation of how to do this).

R Instructions:

With the BEQ.Swear.sav file imported as **beq.swear** and the active dataset in R Commander choose the menu sequence GRAPHS > SCATTERPLOT. Pick **l2speak** for the x-variable, and **agesec** for the y-variable. In the “Options” tab tick off “Marginal boxplots” and “Show spread” so only the “Least-squares line” and “Smooth line” are ticked. Press OK.

The R code is:

```
scatterplot(l2speak~agesec, reg.line=lm, smooth=TRUE, spread=FALSE,  
id.method='mahal', id.n=2, boxplots=FALSE, span=0.5, data=beq.swear)
```

Evaluation:

This scatterplot looks odd because the answers are so discrete for the L2 speaking rating (there are only 5 choices). The Loess line does not fit the regression line exactly but may be close enough to assume a linear relationship between the variables. The regression line has a negative slope, meaning there is a negative correlation between age and estimated ability in speaking a second language. In other words, the older the person was when they started learning their second language, the lower they estimated their speaking ability.

7 Abrahamsson & Hyltenstam (2009)

Use the Abrahamsson&Hyltenstam.NoNS.sav file. Import into R as **ahNoNS**.

SPSS Instructions:

In the SIMPLE SCATTERPLOT dialogue box (see Exercise #1 answers for how to get it), put the variable PERCNATIVE in the Y Axis box and AGEONSET in the X axis box. Press OK. Insert regression line and Loess line (see previous exercises for an explanation of how to do this).

R Instructions:

With the Abrahamsson&Hyltenstam.NoNS.sav file imported as **ahNoNS** and the active dataset in R Commander, choose the menu sequence GRAPHS > SCATTERPLOT. Pick **ageonset** for the x-variable, and **percnative** for the y-variable. In the “Options” tab tick off “Marginal boxplots” and “Show spread” so only the “Least-squares line” and “Smooth line” are ticked. Press OK.

The R code is:

```
scatterplot(ageonset~percnative, reg.line=lm, smooth=TRUE, spread=FALSE,  
id.method='mahal', id.n=2, boxplots=FALSE, span=0.5, data=ahNoNS)
```

Evaluation:

The Loess line and regression line look like they are similar enough to consider this a linear relationship between variables. The relationship is negative, so the older the age of onset, the lower the score as a native speaker of Swedish.

6.4.4 Application Activities for Correlations (Answers for both SPSS and R given for each item)

1 DeKeyser (2000) data

Use the data file Dekeyser2000.sav.

a. Check Assumptions

We have already checked this data for linearity in Section 6.3.3, Exercise #1. We assume that the variables were independently gathered. Look at histograms for everyone all together and for the data divided into groups depending on STATUS and also look at skewness and kurtosis numbers.

SPSS Instructions:

To look at normality graphically and numerically over the entire dataset, choose ANALYZE > DESCRIPTIVE STATISTICS > FREQUENCIES. Move AGE and GJTSCORE into the “Variable” box. Click on the “Statistics” button and tick on the “Skewness” and “Kurtosis” boxes under “Distribution.” Click off “Display frequency tables.” Press “Continue.” Click on the “Charts” button to choose a Histogram with a normal curve superimposed. Press OK to run the analysis.

To divide the data into groups, DATA > SPLIT FILE. Click COMPARE GROUPS and move STATUS variable over to box, click OK. Run the exploratory analysis again.

R Instructions:

Import the DeKeyser2000.sav SPSS file as **dekeyser**.

We can use histograms check for normal distribution. In R Commander, make sure **dekeyser** is the active dataset, and choose GRAPHS > HISTOGRAM and first choose **age** and later come back and choose **gjtscore**. Press OK.

To explore histograms of the separate groups, the easiest way is probably to take the code for the histogram generated for the entire dataset and add line numbers that split the data into groups.

For example, calling for a histogram for the entire group for age yields this code from R

Commander:

```
Hist(dekeyser$age, scale="frequency", breaks="Sturges", col="darkgray")
```

If you look at the data (use the “View dataset” button below the pull-down menus) you can see that the “Under 15” group runs from line 1–15 and the “Over 15” from 16–57. In the R console run this code:

```
Hist(dekeyser$age[1:15], scale="frequency", breaks="Sturges", col="darkgray")
```

Do the same for the other group ([16:57]) and the other variable (`gjtscore`).

To look at normality the fBasics library provides a lot of numbers.

```
library(fBasics)
```

```
basicStats(dekeyser$gjtscore[1:15])
```

```
basicStats(dekeyser$age[1:15])
```

```
basicStats(dekeyser$age[16:57])
```

```
basicStats(dekeyser$gjtscore[16:57])
```

Evaluation:

For all data together, the histogram for AGE shows a gap between about 13 and 20 where we would expect more data. Doing the same thing for GJTSCORE shows a highly negatively skewed distribution of test scores (more people scored highly than we would expect). Over the entire dataset, the assumption of normality does not seem highly accurate. For the groups separated by STATUS, for age, we could consider both groups normally distributed. On the GJT score, for the Over 15 group the distribution looks fairly normal; for the Under 15 it is still highly negatively skewed. For numeric results, the skewness for all data is -0.5 (-.4 if using R) for Age and -0.3 for GJTscore, both under 1. Kurtosis is also low. For both groups separated, the skewness for the Under 15 group for Age is -0.4 (-0.3 for R), but for GJT score is much larger at -1.5 (-1.2 for R), and the kurtosis is also higher than it is for the Over 15 group. For the Over 15 group the skewness for Age is .8 (.7 for R) and for GJT score is -.2. Because both the graphical and numerical results show non-normality, we might want to consider robust options of correlational analysis especially for the “Under 15” group.

b. Run the Correlation

You might be interested in running the correlation over the entire dataset first and then separately for each STATUS group.

SPSS Instructions:

Choose ANALYZE > CORRELATE > BIVARIATE. Move AGE and GJTSCORE to the right-hand side.

Leave default boxes for Pearson's correlation and Flag significant correlations checked. Open the "Bootstrap" button and tick the "Perform bootstrapping" box. Change radio button under "Confidence Intervals" to BCa. Press "Continue", then press OK.

To run the correlation over the data divided into groups: DATA > SPLIT FILE. Click COMPARE GROUPS and move Status to the box. Run the same commands as before.

R Instructions:

To obtain the correlation in R Commander, make sure **dekeyser** is the active dataset, then choose STATISTICS > SUMMARIES > CORRELATION MATRIX.

Choose "Age" and "GJTScore" as variables (use the CTRL button to choose both); leave Pearson button marked and click "Pairwise-complete observations." Tick the "Pairwise p-values box." Press OK.

The R code for this was:

```
rcorr.adjust(dekeyser[,c("age", "gjtscore")], type="pearson", use="pairwise.complete")
```

To repeat this with for the groups, first create a new subset of the data:

```
dekeyserUnder<-dekeyser[1:15,]  
rcorr.adjust(dekeyserUnder[,c("age", "gjtscore")], type="pearson",  
use="pairwise.complete")
```

```
dekeyserOver<-dekeyser[16:57,]  
rcorr.adjust(dekeyserOver[,c("age","gjtsscore")], type="pearson",  
use="pairwise.complete")
```

To get confidence intervals, follow the directions for bootstrapping in Section 6.4.2.

```
library(boot)  
f <- function(data,indices){  
  obs<-data[indices,]  
  return(cor(obs$age, obs$gjtsscore))}  
bootcorr<-boot(dekeyser,f, R=1000)  
bootcorr  
boot.ci(bootcorr)
```

To do the same with split groups, just change `obs$age` into `obs$age[1:15]` and `obs$gjtsscore` into `obs$gjtsscore[1:15]` for the Under 15 group, and then change the row numbers to [16:57] for the Over 15 group.

Evaluation:

The correlation between age of arrival and scores over the entire dataset is Pearson's $r = -.62$, $p < .0005$ (it says .000 in the SPSS & R output, so we know it is less than .0005), $N = 57$. This is a strong effect size ($R^2 = .38$). The 95% BCa confidence interval I got was [-0.8, -0.5] in SPSS and

[-0.7, -0.4] in R, but remember, bootstrapped numbers change every time since there is randomization involved. Anyway, these numbers are similar.

For the results divided into groups:

For the Under 15, $r = -.26$, $p = .35$, $N = 15$; bootstrapped 95% Bca CI [-6.4, .13]

For the Over 15, $r = -.03$, $p = .86$, $N = 42$; bootstrapped 95% Bca CI [-.33, .28]

The correlation is statistical over the entire group, but when the groups are separated the correlation is not statistical.

2 Flege, Yeni-Komshian, and Liu (1999) data

a. Check Assumptions

We have already checked this data for linearity and found that it could be considered linear although it might best be considered a third-order function (see Figure 6.6). We assume that the variables were independently gathered. Use histograms as well as numerical values for skewness and kurtosis to check on normality assumptions.

SPSS Instructions:

To look at normality graphically and numerically over the entire dataset, choose ANALYZE > DESCRIPTIVE STATISTICS > FREQUENCIES. Move AOA and PRONKOR into the “Variable” box. Click on the “Statistics” button and tick on the “Skewness” and “Kurtosis” boxes under “Distribution.” Click off “Display frequency tables”. Press “Continue.” Click on the “Charts” button to choose a Histogram with a normal curve superimposed. Press OK to run the analysis.

To look at a histogram of the data only up to age 20, you'll need to select cases (DATA > SELECT CASES). Choose the variable AOA, then the radio button "If." Push the "If" button that is available after you click the radio button, and move the AOA variable to the right, then add <= 20 and click "Continue." Keep the default for Output ("Filter out unselected cases"). Press OK, then use the same sequence as above to examine normality.

R Instructions:

Import the FlegeYeniKomshianLiu.sav file into R as `fyl`. Check a histogram for the normality assumption. In R Commander, choose GRAPHS > HISTOGRAM, and then choose each variable (`aoa`, `pronkor`) separately. To explore histograms of the data through age 20 only, the easiest way is probably to take the code for the histogram generated for the entire dataset and add line numbers for AOA up to age 20. If you look at the data (use the "View dataset" button below the pull-down menus) you can see that the data up to AOA 18.5 (the highest AOA that is 20 or below) runs from line 1-216. In the R console run this code:

```
Hist(fyl$aoa [1:216], scale="frequency", breaks="Sturges", col="darkgray")
```

Do the same for the other variable (`pronkor`).

To look at normality the fBasics library provides a lot of numbers.

```
library(fBasics)
```

```
basicStats(fyl$pronkor)
```

```
basicStats(fyl$aoa)
```

```
basicStats(fyl$pronkor[1:216])
```

```
basicStats(fyl$aoa[1:216])
```

Evaluation:

For the entire dataset, skewness is very small for AOA (.05 in SPSS, .04 in R) and only slightly larger for English pronunciation (-.7). Kurtosis is -1.1(1.2 in R) for AOA and -.6 for pronunciation. Skewness is not a problem but kurtosis is not regular for AOA.

For the AOA data up to AOA 20, skewness is negligible (-0.002), while it is still quite small for English pronunciation (-0.56). The kurtosis is -1.2 for AOA and -.8 for pronunciation. Skewness is not a problem but kurtosis is not normal, especially for AOA.

The histograms (for both the entire dataset and for the data only up to age 20) show that AOA is not normally distributed, since the researchers manipulated this variable to have approximately equal numbers of participants in each AOA band. The histogram of pronunciation negatively skewed, with more participants scoring highly on this measurement than would be expected given a normal distribution.

b. Run the Correlation

SPSS Instructions:

To calculate the correlation between AOA and pronunciation scores, use ANALYZE > CORRELATE > BIVARIATE. Move AOA and PRONKOR to the right-hand side. Leave default boxes for

Pearson's correlation and Flag significant correlations checked. Open the "Bootstrap" button and tick the "Perform bootstrapping" box. Change radio button under "Confidence Intervals" to BCa. Press "Continue," then press OK.

R Instructions:

In R Commander choose STATISTICS > SUMMARIES > CORRELATION MATRIX. Choose AOA and PRONKOR (Use the CTRL button to choose both); leave Pearson button marked and click "Pairwise-complete observations." Tick the "Pairwise p-values box." Press OK.

The R code for this was:

```
rcorr.adjust(fyl[,c("aoa","pronkor")], type="pearson", use="pairwise.complete")
```

To repeat this with only a subset of the data (up to AOA 20), first create a new subset of the data:

```
fyl20<-fyl[1:216,]
```

```
rcorr.adjust(fyl20[,c("aoa","pronkor")], type="pearson", use="pairwise.complete")
```

To get confidence intervals, follow the directions for bootstrapping in Section 6.4.2.

```
library(boot)
```

```
f <- function(data,indices){
```

```
  obs<-data[indices,]
```

```
return(cor(obs$aao, obs$proneng)))}
```

```
bootcorr<-boot(fyl,f, R=1000)
```

```
bootcorr
```

```
boot.ci(bootcorr)
```

To do the same with split groups, just change `obs$aao` into `obs$aao [1:216]` and `obs$proneng` into `obs$proneng [1:216]`.

Evaluation:

For the entire dataset, Pearson's $r = .74$, $p < .0005$, and $N = 240$. This is an extremely large effect size ($R^2 = .55$), although smaller than the size of the effect of English pronunciation and age of arrival. There correlation is strong and positive, meaning that pronunciation in Korean is better the later the participant arrived in the US. The 95% BCa CI is $[.68, .79]$ (or something close to this), a quite small confidence interval that means we can be very confident the true correlation coefficient is very large!

For the data only up to age 20, Pearson's $r = .76$, $p < .0005$, and $N = 216$. The 95% BCa CI is $[.7, .8]$ (or something close to that). There is no practical difference between the data only up to this point and the entire dataset.

3 Larson-Hall (2008) data

Use the LarsonHall2008.sav file. Import as `lh2008` into R.

a. Check Assumptions

Use histograms as well as numerical values for skewness and kurtosis to check on normality.

SPSS Instructions:

We assume that variables were independently gathered. We need to check this data for linearity:

GRAPHS > LEGACY DIALOGS > SCATTERPLOT (MATRIX), enter USEENG, LIKEENG, and GJTSCORE. This gives us a matrix of the variables.

Next, check for normality of distribution of the variables: GRAPHS > LEGACY DIALOGS > HISTOGRAM.

R Instructions:

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

We can use a multiple scatterplot to check both scatterplots and histograms at the same time. In R Commander, choose GRAPHS > SCATTERPLOT MATRIX. Hold down the Ctrl button and use the mouse to choose the 3 variables of **gjtscore**, **likeeng**, and **useeng**. Click the Options tab and click on the “Histograms” radio button to put these on the diagonal. Leave both boxes “Least-squares lines” and “Smooth lines” ticked. Press OK.

The R code for this was:

```
scatterplotMatrix(~gjtscore+likeeng+useeng, reg.line=lm, smooth=TRUE,
```



```
spread=FALSE, span=0.5, id.n=0, diagonal='histogram', data=lh2008)
```

Evaluation:

Linearity (look at scatterplots): The regression line seems to match the Loess line fairly well except in the case of USEENG and GJTSCORE where the Loess line is curved. We can also note that in the case of USEENG and LIKEENG the data are not randomly scattered. There is a fan-like effect, which means the data are heteroscedastic (variances are not equal over the entire area). This combination would probably not satisfy the parametric assumptions. There could be some outliers in the LIKEENG ~ USEENG combination, and also in the USEENG ~ GJTSCORE combination, which would also not satisfy parametric assumptions.

Normality (look at histograms): Use of English is highly positively skewed. Most Japanese learners of English don't use very much English. The degree to which people enjoy studying English (LIKEENG) is more normally distributed, although there seems to be too much data on the far right end of the distribution to be a normal distribution. The GJT scores seem fairly normally distributed.

b. Run the Correlation**SPSS Instructions:**

Run the correlation on all of the variables: ANALYZE > CORRELATE > BIVARIATE, enter variables. Leave default boxes for Pearson's correlation and Flag significant correlations checked. Open the "Bootstrap" button and tick the "Perform bootstrapping" box. Change radio button under "Confidence Intervals" to BCa. Press "Continue," then press OK.

R Instructions:

In R Commander choose STATISTICS > SUMMARIES > CORRELATION MATRIX. Choose the same variables as for the scatterplots; leave Pearson button marked and click “Pairwise-complete observations.” Tick the “Pairwise p-values box.” Press OK.

The R code for this was:

```
rcorr.adjust(lh2008[,c("gjtsscore", "likeeng", "useeng")],  
type="pearson", use="pairwise.complete")
```

Previous experience would lead us to believe we could use the code given below to get confidence intervals (from Section 6.4.2). A little experimentation with having all three variables at once seems to show that you cannot use more than two variables at a time, so we'll start by looking at GJTSCORE with LIKEENG (a warning comes up when I try all 3 at once that says something about only the first element will be used and also something about an invalid “use” argument, so I'm assuming this is so).

```
library(boot)  
f <- function(data, indices){  
  obs<-data[indices,]  
  return(cor(obs$gjtsscore, obs$likeeng))}  
bootcorr<-boot(lh2008,f, R=1000)
```

```
bootcorr
```

```
boot.ci(bootcorr)
```

In this case, the ordinary bootstrap shows that the bias is NA, and then the `boot.ci` warning says that “w' is infinite.” My suspicion is that there is something wrong with the data, so I check and find there are NAs in the LIKEENG and USEENG variables. So I went back to Section 1.5.3 and imputed the data using the `mice` library.

```
library(mice)
```

```
imp<-mice(lh2008, 5)
```

```
implh2008<-complete(imp)
```

If I now try the line with my new variable:

```
bootcorr<-boot(implh2008,f, R=1000)
```

```
boot.ci(bootcorr)
```

Things now seem to turn out as they should. I think at this point I'll go back and check correlation sizes with my new imputed dataset as well:

```
rcorr.adjust(implh2008[,c("gjtscore", "likeeng", "useeng")],
```

```
type="pearson", use="pairwise.complete")
```

Evaluation: All correlations are statistical, although the 95% CI shows that the possible effect size varies quite a bit (the CI is fairly wide). Note that 95% CIs will almost certainly not be the

same as mine because of the randomization process, and also note that SPSS may differ quite a bit since I used an imputed dataset for the R numbers I am providing here.

USEENG ~ LIKEENG, $r = .35$, $p < .005$, $N = 200$, 95% BCa CI using R's imputed dataset [0.22, 0.47]

USEENG ~ GJTSCORE, $r = .30$, $p < .005$, $N = 200$, 95% BCa CI using R's imputed dataset [0.15, 0.43]

GJTSCORE ~ LIKEENG, $r = .32$, $p < .005$, $N = 200$, 95% BCa CI using R's imputed dataset [0.19, 0.44]

Remember that with enough N , any statistical test can become “significant”! I have a very large N , so the question is, how large is the effect size? Use the Pearson's r to calculate R^2 , USEENG ~ LIKEENG, $R^2 = .12$, USEENG ~ GJTSCORE, $R^2 = .09$, GJTSCORE ~ LIKEENG, $R^2 = .09$. All of the effect sizes are medium. There appear to be moderate connections between how much Japanese learners of English like English and how much they use it, and between their scores on a grammar test and how much they like it and how much they use it.

4 Dewaele and Pavlenko (2001–2003) data

Use the BEQ.Swear.sav file. Import into R as `beq.swear`.

We assume that the variables were independently gathered. Check the data for linearity by looking at scatterplots (in Section 6.3.3 we already looked at the scatterplot between AgeSec and

L2Speak, and found it was roughly linear, and negatively correlated). Look at histograms to check for normality also look at skewness and kurtosis numbers.

a. Check Assumptions

SPSS Instructions:

To look at linearity, make a scatterplot matrix of the variables by choosing **GRAPHS > LEGACY DIALOGS > SCATTERPLOT**, then choose **MATRIX SCATTER**, enter variables AGESEC, L2SPEAK, and L2_COMP. Press OK. To make it easier to see trends, go into Chart Editor by double clicking on the graph, then choose **ELEMENTS > FIT LINE AT TOTAL**, and click CLOSE to see straight regression lines. Go in again and add Loess lines.

To look at normality graphically and numerically, choose **ANALYZE > DESCRIPTIVE STATISTICS > FREQUENCIES**. Move the three variables into the “Variable” box. Click on the “Statistics” button and tick on the “Skewness” and “Kurtosis” boxes under “Distribution.” Press “Continue.” Click on the “Charts” button to choose a Histogram with a normal curve superimposed. Click off “Display frequency tables.” Press OK to run the analysis.

R Instructions:

We can use a multiple scatterplot to check both scatterplots and histograms at the same time. In R Commander, choose **GRAPHS > SCATTERPLOT MATRIX**. Hold down the Ctrl button and use the mouse to choose the 3 variables of AGESEC, L2SPEAK, and L2_COMP. Click the Options tab and click on the “Histograms” radio button to put these on the diagonal. Leave both boxes “Least-squares lines” and “Smooth lines” ticked. Press OK.

The R code for this is:

```
scatterplotMatrix(~agesec+l2_comp+l2speak, reg.line=lm, smooth=TRUE,  
spread=FALSE, span=0.5, id.n=0, diagonal = 'histogram', data=beq.swear)
```

To look at normality use fBasics library:

```
basicStats(beq.swear$agesec)
```

```
basicStats(beq.swear$l2speak)
```

```
basicStats(beq.swear$l2_comp)
```

Evaluation:

Scatterplots: The fit of the regression and Loess lines is best for the correlation between L2 speaking and L2 comprehension. The other relationships show a negative correlation with age at first, but then seem to level off before going down at an older age again, so a linear relationship might not be the best model for this data (it might be bi-modal)

Histograms: Age at which a second language is learned is decidedly non-normally distributed and positively skewed, because there are a large number of participants who learned it at 0. This seems to be in contrast to a more normal-looking distribution after that age. The histograms of both L2 speaking and comprehension ability are highly negatively skewed, meaning more people

think they speak and comprehend well than would be expected in a normal distribution.

Parametric assumptions are violated for all three variables.

Numbers: Skewness is quite large for L1speak (-5.0) and well over 1 for L2_comp (-1.8).

Skewness is also quite large (27.9) for L1Speak. None of the variables appears to be completely normally distributed.

b. Run the Correlation

SPSS Instructions:

ANALYZE > CORRELATE > BIVARIATE. Move AGESEC, L2SPEAK, and L2_COMP to the right-hand side. Leave default boxes for Pearson's correlation and Flag significant correlations checked. Open the "Bootstrap" button and tick the "Perform bootstrapping" box. Change radio button under "Confidence Intervals" to BCa. Press "Continue", then press OK.

R Instructions:

In R Commander make sure `beq.swear` is the active dataset, then choose STATISTICS > SUMMARIES > CORRELATION matrix.

Choose AGESEC, L2_COMP and L2SPEAK as variables (use the CTRL button to choose all 3); leave Pearson button marked and click "Pairwise-complete observations." Tick the "Pairwise p-values box." Press OK.

The R code is:

```
rcorr.adjust(beq.swear[,c("agesec", "l2_comp", "l2speak")], type="pearson",
```

```
use="pairwise.complete")
```

As with the previous exercise in #3, in order to get bootstrapped CIs we'll have to get rid of the NAs, and my preferred method will be imputation. However, in trying to impute the entire file I get a warning that "comparison of these types is not implemented." This is a very large file with quite a number of variables. Let's make it a little easier and try to create a matrix with only the variables we are interested in. I used R Commander: DATA > ACTIVE DATASET > SUBSET ACTIVE DATASET, then I picked the 3 variables and named it `small.beqswear`. The R code for this is:

```
small.beqswear <- subset(beqswear, select=c(agesec,l2_comp,l2speak))
```

```
imp<mice(small.beqswear,5)# open mice package first if necessary
```

I still get the same error warning however. Since the file is so large, the easiest thing to do might be to just remove those cases that contain NAs. I'll go ahead and keep the smaller file though.

The file `small.beqswear` is the active dataset, and I used R Commander again: DATA > ACTIVE DATASET > REMOVE CASES WITH MISSING DATA. I just kept all the variables and the same name and overwrote the previous file. The code for this action is:

```
small.beqswear <- na.omit(small.beqswear)
```

Now I will try the code for finding the 95% CI:

```
library(boot)
```



```
f <- function(data,indices){
  obs<-data[indices,]
  return(cor(obs$agesec, obs$l2_comp))}
bootcorr<-boot(small.beqswear,f, R=1000)
bootcorr
boot.ci(bootcorr)
```

The ordinary nonparametric bootstrap appears to be fine but the `boot.ci(bootcorr)` command does not return a result and says there is an error in `bca.ci`. So let's leave out the BCa and choose a different type of CI, since we know the problem is NOT missing data.

```
boot.ci(bootcorr, type= c("norm","basic", "stud", "perc"))
```

None of the CIs that I can call for look too different from one another, so I'm not so worried that I couldn't get the BCa interval.

Evaluation:

All correlations are statistical.

AGESEC~L2SPEAK, $r = -.20$, $p < .005$, $N = 1015$, 95% BCa CI from SPSS [-.25, -.14]

AGESEC~L2_COMP, $r = -.20$, $p < .005$, $N = 1015$, 95% BCa CI from SPSS [-.25, -.15]

L2SPEAK~L2COMP, $r = .85$, $p < .005$, $N = 1015$, 95% BCa CI from SPSS [.83, .87]

Again, with such a large N it is no miracle to find statistical associations. The important question is effect size. For the correlation between age of learning a second language and how well a person thinks they speak it, the effect size is $R^2 = .04$, a small effect size, which may be quite surprising. It is the same for the relationship between age of learning and comprehension. For the relationship between speaking and comprehension ability in an L2, however, the relationship is much larger: $R^2 = .72$, a very large effect size. The CIs are all rather narrow, a result of such large Ns, and confirm that the relationship between how well someone rates themselves as speaking their L2 is quite heavily tied to how well they rate themselves for comprehending that language as well.

5 Abrahamsson & Hyltenstam (2009) data

Use the Abrahamsson&Hyltenstam.NoNS.sav file. Import it into R as `ahNoNS`.

We want to use a robust correlation to examine the good part of the data, and first we'll just assume that all of the data from the NNS belong to the same group, although the authors split the NNS into group of age of onset up to 11 years, and 12 years or more.

Robust Analysis:

```
library(mvoutlier) #use install.packages("mvoutlier") first if you don't have this package
```

```
attach(ahNoNS)
```

```
corr.plot(ageonset,percnative)
```

Additionally, let's try some robust methods from Wilcox's WRS library. See Section 6.4.3 for the commands needed to install the required libraries in order to open the WRS library. Use the `scor()` function, which takes the overall structure of the data into account while eliminating outliers.

```
library(WRS)
```

```
scor(ahNoNS$ageonset, ahNoNS$percnative, plotit=T, xlab="Age of Onset",  
ylab="Ratings", STAND=T)
```

Last, let's try the percentile bootstrap with an OP correlation using Wilcox's WRS library:

```
corb(ahNoNS$ageonset, ahNoNS$percnative, corfun = scor, nboot = 599)
```

Non-robust analysis with bootstrapped 95% CI:

```
library(boot) #if necessary
```

```
f <- function(data,indices){
```

```
  obs<-data[indices,]
```

```
  return(cor(obs$ageonset, obs$percnative))}
```

```
bootcorr<-boot(ahNoNS,f, R=1000)
```

```
bootcorr
```

```
boot.ci(bootcorr)
```

Evaluation:

The `cor.plot()` command's result shows that the “good part” of the data is considered the data from age 0 to an oldest age of about 25, and also excludes all of the data from participants who received a score of zero. The robust correlation ($r=-0.46$) is smaller than the classical correlation ($r=-0.72$). Although this may be a mathematically reasonable way to analyze the data, theoretically we may not really want to exclude those who scored a zero, as we think there may be a reason they scored a zero!

The `scor()` function draws a polygon around most of the data except for a couple of outliers to the extreme far right of the distribution. The correlation remains the same as the classical correlation.

The percentile bootstrap with an OP correlation took a long time running and I eventually pressed the ESC button to stop it, so I have no results for this.

The robust correlations did not lead me to any idea of a better way to split the data file by age of onset. A different approach would be to look at correlations at different split points, but I won't do that here.

For the non-robust analysis I used the same code as I have been using throughout these application activity answers. The 95% BCa CI is $[-0.78, -0.65]$. Although this is not a really narrow CI, it is small enough to feel confident enough that the overall effect of age on the entire

group is quite strong (R^2 =at least $(.65 \times .65)=42\%$) and accounts for last least 42% of the variation in the scores.