HDFS 523: Strategies for Data Analysis in Developmental Research

Zachary F. Fisher

2023-01-23

# 1 About This Book

This book provides the course notes for HDFS 523. It is currently under development, so any feedback is appreciated (e.g., during class, via email, or the edit link in the header). This first chapter is just about how to use the book – the course content starts in Chapter 2.

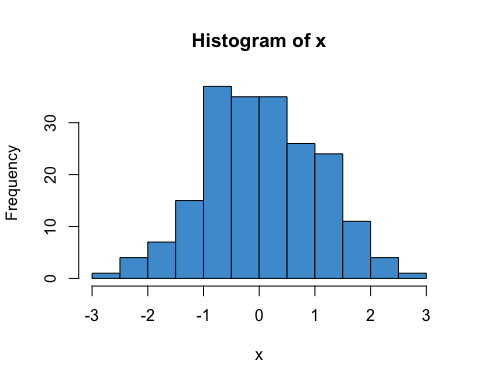
## 1.1 Why this book?

There are a few goals of moving from “textbook + slides + exercises” to an ebook. For now, the main goal is to update and integrate code contents from the course into one consistent format, rather than having multiple files to sort through on Canvas.

## 1.2 Code Folding

The book combines lecture slides and R coding examples. It is often convenient to hide code when introducing new material. This is accomplished using code folding. An example of code folding is given on this page. Below, a histogram integrated into the text. By clicking on the button called “Show Code” on the top of the page, the R code that produced the histogram will also be visible. Notice that you may need to scroll horizontally to see all of the text in the code window. Also notice that when you hover your mouse over the code window, an icon appears in the top right corner – this lets you copy the block of code with one click.

# Here is some R code. You don't have to look at it when reading the book, but it is here when you need it  
x <- rnorm(200)  
hist(x, col = "#4B9CD3")



## 1.3 Acknowledgements

Many people have contributed to the course materials for HDFS 523. Most importantly, many of the original R markdown files for the course were developed by Nilam Ram and Zita Oravecz.

# 2 Data Cleaning

In Chapter 2 we will work through some basic data cleaning operations useful in longitudinal data analysis. The basic idea is provide a set of scripts to use for exploring new repeated measures data sets.

## 2.1 Example Data

For Chapter 2 we will make use of the longitudinal Wechsler Intelligence Scale for Children [WISC; Wechsler, -Wechsler ([1949](#ref-wechsler1949))] dataset described by Osborne and Suddick ([1972](#ref-osborne1972)). These data have been detailed extensively in a number of papers ([McArdle and Epstein 1987](#ref-mcardle1987); [McArdle 1988](#ref-mcardle1988); [Mcardle and Aber 1990](#ref-mcardle1990); [McArdle and Nesselroade 1994](#ref-mcardle1994)) and are used here with with permission.

The WISC data contains repeated measures data from 204 children between the ages of 6 and 11 years old (during grades 6, 7, 9 and 11). Thee repeated measures include component scores for the verbal tests and performance subtests at all four occasions, along with verbal subtest scores for the information, comprehension, similarities, and vocabulary domains at the first and last measurement occasion. The demographics variables mother’s education (continuous in years) and mother graduated high school (dichotomous) are also included.

## 2.2 Reading in Repeated Measures Data

We can read in the WISC data directly from the [QuantDev](https://quantdev.ssri.psu.edu/) website.

filepath <- "https://quantdev.ssri.psu.edu/sites/qdev/files/wisc3raw.csv"  
wisc3raw <- read.csv(file=url(filepath), header=TRUE)

Additional details on importing different data types into R can be found here: <http://www.statmethods.net/input/importingdata.html>.

## 2.3 Familiarize Yourself with the Data

Let’s take an initial look at the structure of our data object using str()

str(wisc3raw)

## 'data.frame': 204 obs. of 20 variables:  
## $ id : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ verb1 : num 24.4 12.4 32.4 22.7 28.2 ...  
## $ verb2 : num 27 14.4 33.5 28.4 37.8 ...  
## $ verb4 : num 39.6 21.9 34.3 42.2 41.1 ...  
## $ verb6 : num 55.6 37.8 50.2 44.7 71 ...  
## $ perfo1 : num 19.8 5.9 27.6 33.2 27.6 ...  
## $ perfo2 : num 23 13.4 45 29.7 44.4 ...  
## $ perfo4 : num 43.9 18.3 47 46 65.5 ...  
## $ perfo6 : num 44.2 40.4 77.7 61.7 64.2 ...  
## $ info1 : num 31.3 13.8 35 24.8 25.3 ...  
## $ comp1 : num 25.6 14.8 34.7 31.4 30.3 ...  
## $ simu1 : num 22.93 7.58 28.05 8.21 15.98 ...  
## $ voca1 : num 22.2 15.4 26.8 20.2 35.4 ...  
## $ info6 : num 69.9 41.9 60.4 52.9 67.4 ...  
## $ comp6 : num 44.4 44.9 50.3 42.7 86.7 ...  
## $ simu6 : num 68 33.9 35.8 45.8 72.4 ...  
## $ voca6 : num 51.2 37.7 55.5 36 60.4 ...  
## $ momed : num 9.5 5.5 14 14 11.5 14 9.5 5.5 9.5 11.5 ...  
## $ grad : int 0 0 1 1 0 1 0 0 0 0 ...  
## $ constant: int 1 1 1 1 1 1 1 1 1 1 ...

From the output, we can also see that the data frame consists of 204 observations (rows) and 20 variables (columns). Each variable’s name and data type is also listed. Methods like the ones above can be an effective way to initially familiarize yourself with the main features of a dataset.

## 2.4 Look for Duplicated IDs

It is always worth looking for non-unique ID numbers when ID labels are included in a dataset. Here we have an id variable indicating the subject number. Since our data is in a long format (more on that later) duplicate IDs may indicate a potential problem with the data source or clues on how the data is structured.

any(duplicated(wisc3raw$id))

## [1] FALSE

## 2.5 Using table() to Spot Irregularities

When a variable takes on a limited range of values it is often useful to screen for irregularities or invalid values. This is common across all variable types and can occur for character strings, numeric, integer and factor types. For example, we would expect the grad variable to only take the values of zero or one. We can use the table() function to quickly confirm this.

By default table() simply omits any values coded as NA. To include a count of the NA values use the useNA argument of table() as follows:

table(wisc3raw$grad, useNA = "always")

##   
## 0 1 <NA>   
## 158 46 0

## 2.6 Missing Data

Dealing with missing data in a consistent manner is one of the most important aspects of data cleaning. When data are imported into R it is common to discover missing values are coded according to a variety of conventions.

Often a first step in handling missing data involves recoding missing values as NA. Writing bespoke code to handle the different types of missing data one might encounter is tedious and unnecessary.

naniar ([Tierney et al. 2021](#ref-naniar)) is a useful package with many convenience functions for managing missing data in R. Here we demonstrate some of this functionality.

### 2.6.1 Generating Example Data

Since the WISC data does not contain missing values it is helpful to generate a synthetic dataset containing some commonly encountered missing data codes.

set.seed(123)  
wisc\_miss <- wisc3raw  
wisc\_miss$verb1[sample(nrow(wisc\_miss),100)] <- -99  
wisc\_miss$comp1[sample(nrow(wisc\_miss),75)] <- "N/A"  
wisc\_miss$info1[sample(nrow(wisc\_miss),50)] <- "NA"

### 2.6.2 Recoding Values with NA

Now that we have a dataset with missing values we can use naniar to recode these values to NA.

na\_strings <- c("NA", "N/A", -99)  
   
wisc\_miss <- naniar::replace\_with\_na\_all(  
 wisc\_miss, condition = ~.x %in% na\_strings  
)

See the [naniar vignette on recoding NA values](https://cran.r-project.org/web/packages/naniar/vignettes/replace-with-na.html) for more detailed information on the package functionality.

### 2.6.3 Missing Data Visualization

Once we have recoded our data in a consistent manner we can use visualizations to explore the missing data. The vis\_miss() function from naniar is a good starting point for visualizing the amount of missing data in our dataset. The plots shows the missing values in black and non-missing values in gray. In addition, percentages of missing data in both the dataset and individual variables are provided.

naniar::vis\_miss(wisc\_miss)

## Warning: `gather\_()` was deprecated in tidyr 1.2.0.  
## ℹ Please use `gather()` instead.  
## ℹ The deprecated feature was likely used in the visdat package.  
## Please report the issue at <]8;;https://github.com/ropensci/visdat/issueshttps://github.com/ropensci/visdat/issues]8;;>.



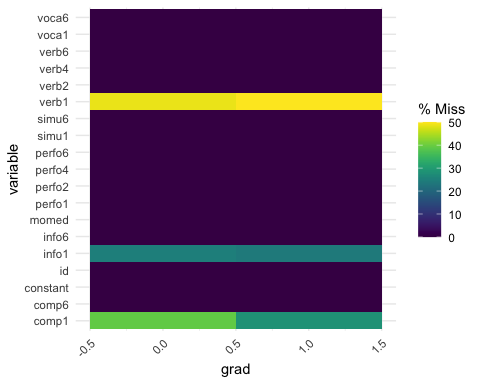
It is often useful to look at combinations of missingness among different variables.

naniar::gg\_miss\_upset(wisc\_miss)



We can also look at the percentage of missing data across a factor variable.

naniar::gg\_miss\_fct(x = wisc\_miss, fct = grad)



Many missing data visualizations are described in the [naniar vignette on missing data visualization](https://cran.r-project.org/web/packages/naniar/vignettes/naniar-visualisation.html) including plots for exploring missing data mechanisms.

## 2.7 Exporting Data

Depending on work-flow, you may need to export your dataset for use in another statistical software program. The write.csv() function is a convenient method for outputting comma delimited files.

write.csv(wisc3raw, file = "wisc3raw.csv", row.names = FALSE, na = "-99")

Note that by default the write.csv() function will include an extra column of row numbers and will notate missing data with an NA. More information on exporting data is available at <http://www.statmethods.net/input/exportingdata.html>.

## 2.8 Reshaping Repeated Measures Data

Behavioral science tends to use relational data structures - in basic form, spreadsheets. Typically, the data are stored in a data frame (a “fancy” matrix) with multiple rows and columns. Two common schemata used to accommodate repeated measures data are *wide format* and *long format*. Different analysis and plotting functions require different kinds of data input. Thus, it is imperative that one can convert the data back and forth between wide and long formats.

There are lots of ways to do this. We illustrate one way.

*Sidebar*: The dput() function provides a convenient method to get the variable names (or any R object) into a format that can be read back into R. For example, this can be helpful when working with a long vector of strings.

dput(colnames(wisc3raw))

## c("id", "verb1", "verb2", "verb4", "verb6", "perfo1", "perfo2",   
## "perfo4", "perfo6", "info1", "comp1", "simu1", "voca1", "info6",   
## "comp6", "simu6", "voca6", "momed", "grad", "constant")

First, let’s subset our data to only include the variables we need for this analysis.

var\_names\_sub <- c(  
 "id", "verb1", "verb2", "verb4", "verb6",  
 "perfo1", "perfo2", "perfo4", "perfo6",  
 "momed", "grad"  
)  
  
wiscraw <- wisc3raw[,var\_names\_sub]  
head(wiscraw)

## id verb1 verb2 verb4 verb6 perfo1 perfo2 perfo4 perfo6 momed grad  
## 1 1 24.42 26.98 39.61 55.64 19.84 22.97 43.90 44.19 9.5 0  
## 2 2 12.44 14.38 21.92 37.81 5.90 13.44 18.29 40.38 5.5 0  
## 3 3 32.43 33.51 34.30 50.18 27.64 45.02 46.99 77.72 14.0 1  
## 4 4 22.69 28.39 42.16 44.72 33.16 29.68 45.97 61.66 14.0 1  
## 5 5 28.23 37.81 41.06 70.95 27.64 44.42 65.48 64.22 11.5 0  
## 6 6 16.06 20.12 38.02 39.94 8.45 15.78 26.99 39.08 14.0 1

### 2.8.1 Reshape Wide to Long

One way to go *from wide to long* is using the reshape() function from base R.

Notice, the varying argument contains the repeated measures columns we want to stack and the timevar is a new variable containing the grade level information previosuly appended at the end of the colnames listed in varying.

# reshape data from wide to long  
wisclong <- reshape(  
 data = wiscraw,  
 varying = c("verb1", "verb2", "verb4","verb6", "perfo1","perfo2","perfo4","perfo6"),  
 timevar = c("grade"),   
 idvar = c("id"),  
 direction = "long",   
 sep = ""  
)  
  
# reorder by id and day   
wisclong <- wisclong[ order(wisclong$id, wisclong$grade), ]  
  
head(wisclong, 8)

## id momed grad grade verb perfo  
## 1.1 1 9.5 0 1 24.42 19.84  
## 1.2 1 9.5 0 2 26.98 22.97  
## 1.4 1 9.5 0 4 39.61 43.90  
## 1.6 1 9.5 0 6 55.64 44.19  
## 2.1 2 5.5 0 1 12.44 5.90  
## 2.2 2 5.5 0 2 14.38 13.44  
## 2.4 2 5.5 0 4 21.92 18.29  
## 2.6 2 5.5 0 6 37.81 40.38

Again, notice how reshape automatically split verb1, verb2, etc. into a string name and a grade variable.

### 2.8.2 Reshape Long to Wide

Now we go *from long to wide*, again using the reshape() function. The v.names argument specifies the variables to be expanded column wise based on the repeated measure specified in timevar.

#reshaping long to wide  
wiscwide <- reshape(  
 data = wisclong,   
 timevar = c("grade"),   
 idvar = c("id"),  
 v.names = c("verb","perfo"),  
 direction = "wide",   
 sep = ""  
)  
  
# reordering columns   
wiscwide <- wiscwide[, c(  
 "id", "verb1", "verb2", "verb4", "verb6",  
 "perfo1", "perfo2", "perfo4", "perfo6",  
 "momed","grad"   
)]  
  
head(wiscwide)

## id verb1 verb2 verb4 verb6 perfo1 perfo2 perfo4 perfo6 momed grad  
## 1.1 1 24.42 26.98 39.61 55.64 19.84 22.97 43.90 44.19 9.5 0  
## 2.1 2 12.44 14.38 21.92 37.81 5.90 13.44 18.29 40.38 5.5 0  
## 3.1 3 32.43 33.51 34.30 50.18 27.64 45.02 46.99 77.72 14.0 1  
## 4.1 4 22.69 28.39 42.16 44.72 33.16 29.68 45.97 61.66 14.0 1  
## 5.1 5 28.23 37.81 41.06 70.95 27.64 44.42 65.48 64.22 11.5 0  
## 6.1 6 16.06 20.12 38.02 39.94 8.45 15.78 26.99 39.08 14.0 1

Using functions included in base R can be useful in a number of situations. One example is package development where one may wants to limit dependencies.

That said, many people find reshape to be unnecessarily complicated. A similar, and potentially more convenient, set of functions have been developed for reshaping data in the tidyr ([Wickham 2021](#ref-tidyr)) package. For those interested take a look at the pivot\_longer() and pivot\_wider() functions.

For examples using tidyr to reshape data see the [tidyr vignette on pivoting](https://cran.r-project.org/web/packages/tidyr/vignettes/pivot.html).

# 3 Describing Longitudinal Data

In Chapter 3 we will look at some option for describing and visualizing longitudinal data.

## 3.1 Example Data

Again we will make use of the WISC data described in Chapter 2. The following commands recreate the wide and long data we will use throughout this chapter.

filepath <- "https://quantdev.ssri.psu.edu/sites/qdev/files/wisc3raw.csv"  
  
wisc3raw <- read.csv(file=url(filepath),header=TRUE)  
  
var\_names\_sub <- c(  
 "id", "verb1", "verb2", "verb4", "verb6",  
 "perfo1", "perfo2", "perfo4", "perfo6",  
 "momed", "grad"  
)  
  
wiscraw <- wisc3raw[,var\_names\_sub]  
  
# reshaping wide to long  
wisclong <- reshape(  
 data = wiscraw,  
 varying = c("verb1", "verb2", "verb4","verb6", "perfo1","perfo2","perfo4","perfo6"),  
 timevar = c("grade"),   
 idvar = c("id"),  
 direction = "long",   
 sep = ""  
)  
  
# reorder by id and day   
wisclong <- wisclong[ order(wisclong$id, wisclong$grade), ]  
  
#reshaping long to wide  
wiscwide <- reshape(  
 data = wisclong,   
 timevar = c("grade"),   
 idvar = c("id"),  
 v.names = c("verb","perfo"),  
 direction = "wide",   
 sep = ""  
)  
  
# reordering columns   
wiscwide <- wiscwide[, c(  
 "id", "verb1", "verb2", "verb4", "verb6",  
 "perfo1", "perfo2", "perfo4", "perfo6",  
 "momed","grad"   
)]

## 3.2 Describing Means and Variances

Once the wide and long data sets are in place, we can begin describing and plotting the data. Descriptive statistics and visualization are one of the most important aspects of data analysis.

Descriptives and plots will be produced from wide data and long data to show the information that can be gleaned from each construction. Having both in place facilitates learning about the data. Continually keep in mind what portions of the data-box are being described (e.g., persons, variables, occasions).

We can do a quick look at descriptives using the describe() function from the psych ([Revelle 2021](#ref-psych)) package. Note the n in both outputs.

psych::describe(wiscwide)

## vars n mean sd median trimmed mad min max range skew  
## id 1 204 102.50 59.03 102.50 102.50 75.61 1.00 204.00 203.00 0.00  
## verb1 2 204 19.59 5.81 19.34 19.50 5.41 3.33 35.15 31.82 0.13  
## verb2 3 204 25.42 6.11 25.98 25.40 6.57 5.95 39.85 33.90 -0.06  
## verb4 4 204 32.61 7.32 32.82 32.42 7.18 12.60 52.84 40.24 0.23  
## verb6 5 204 43.75 10.67 42.55 43.46 11.30 17.35 72.59 55.24 0.24  
## perfo1 6 204 17.98 8.35 17.66 17.69 8.30 0.00 46.58 46.58 0.35  
## perfo2 7 204 27.69 9.99 26.57 27.34 10.51 7.83 59.58 51.75 0.39  
## perfo4 8 204 39.36 10.27 39.09 39.28 10.04 7.81 75.61 67.80 0.15  
## perfo6 9 204 50.93 12.48 51.76 51.07 13.27 10.26 89.01 78.75 -0.06  
## momed 10 204 10.81 2.70 11.50 11.00 2.97 5.50 18.00 12.50 -0.36  
## grad 11 204 0.23 0.42 0.00 0.16 0.00 0.00 1.00 1.00 1.30  
## kurtosis se  
## id -1.22 4.13  
## verb1 -0.05 0.41  
## verb2 -0.34 0.43  
## verb4 -0.08 0.51  
## verb6 -0.36 0.75  
## perfo1 -0.11 0.58  
## perfo2 -0.21 0.70  
## perfo4 0.59 0.72  
## perfo6 0.18 0.87  
## momed 0.01 0.19  
## grad -0.30 0.03

psych::describe(wisclong)

## vars n mean sd median trimmed mad min max range skew  
## id 1 816 102.50 58.93 102.50 102.50 75.61 1.00 204.00 203.00 0.00  
## momed 2 816 10.81 2.69 11.50 11.00 2.97 5.50 18.00 12.50 -0.36  
## grad 3 816 0.23 0.42 0.00 0.16 0.00 0.00 1.00 1.00 1.31  
## grade 4 816 3.25 1.92 3.00 3.19 2.22 1.00 6.00 5.00 0.28  
## verb 5 816 30.34 11.86 28.46 29.39 11.33 3.33 72.59 69.26 0.71  
## perfo 6 816 33.99 16.14 33.14 33.34 18.14 0.00 89.01 89.01 0.34  
## kurtosis se  
## id -1.20 2.06  
## momed 0.03 0.09  
## grad -0.28 0.01  
## grade -1.43 0.07  
## verb 0.33 0.42  
## perfo -0.43 0.56

### 3.2.1 Verbal Ability (All Persons and Occasions)

Let’s focus on the repeated measures of verbal ability. This step is useful to get a general view of what verbal ability scores look like across persons and occasions, but note that we are *ignoring Time*. In doing so we are not considering how the repeated measures are nested within individuals.

psych::describe(wisclong$verb)

## vars n mean sd median trimmed mad min max range skew kurtosis  
## X1 1 816 30.34 11.86 28.46 29.39 11.33 3.33 72.59 69.26 0.71 0.33  
## se  
## X1 0.42

In addition to the descriptive statistics we can look at a boxplot of verbal ability scores across persons and occasions. Here we will start to use the ggplot2 ([Wickham 2016](#ref-ggplot2)) package.

library("ggplot2")  
ggplot(data = wisclong, aes(x=verb, y=..density..)) +  
 geom\_histogram(binwidth=2.5, fill = "white", color = "black") +   
 geom\_density(color = "red") +  
 ggtitle("Verbal Ability Score (across persons and occasions)") +  
 xlab("Verbal Ability (Grade 1 to 6)") +  
 ylab("Density") +  
 theme\_bw() +  
 theme(  
 panel.grid.major = element\_blank(),  
 panel.grid.minor = element\_blank()  
 )

## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.  
## ℹ Please use `after\_stat(density)` instead.



### 3.2.2 Verbal Ability (Across Time)

Note that our variable is actually “multivariate” because we have repeated measures. We should really consider the time-dependence when we are looking at descriptive statistics and plots.

Let’s now look at **verbal ability scores across time collapsed across individuals**. This can be done using either the describe() function and the wide data or the describeBy() function and the long data.

Let’s look at descriptives using the wide data.

psych::describe(wiscwide[,c("verb1","verb2","verb4","verb6")])

## vars n mean sd median trimmed mad min max range skew  
## verb1 1 204 19.59 5.81 19.34 19.50 5.41 3.33 35.15 31.82 0.13  
## verb2 2 204 25.42 6.11 25.98 25.40 6.57 5.95 39.85 33.90 -0.06  
## verb4 3 204 32.61 7.32 32.82 32.42 7.18 12.60 52.84 40.24 0.23  
## verb6 4 204 43.75 10.67 42.55 43.46 11.30 17.35 72.59 55.24 0.24  
## kurtosis se  
## verb1 -0.05 0.41  
## verb2 -0.34 0.43  
## verb4 -0.08 0.51  
## verb6 -0.36 0.75

Identical results can be obtained using the long data.

psych::describeBy(wisclong[,c("verb")], group = wisclong$grade)

##   
## Descriptive statistics by group   
## group: 1  
## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 204 19.59 5.81 19.34 19.5 5.41 3.33 35.15 31.82 0.13 -0.05 0.41  
## ------------------------------------------------------------   
## group: 2  
## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 204 25.42 6.11 25.98 25.4 6.57 5.95 39.85 33.9 -0.06 -0.34 0.43  
## ------------------------------------------------------------   
## group: 4  
## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 204 32.61 7.32 32.82 32.42 7.18 12.6 52.84 40.24 0.23 -0.08 0.51  
## ------------------------------------------------------------   
## group: 6  
## vars n mean sd median trimmed mad min max range skew kurtosis  
## X1 1 204 43.75 10.67 42.55 43.46 11.3 17.35 72.59 55.24 0.24 -0.36  
## se  
## X1 0.75

We can visualize the distribution of verbal scores across grades in a number of different ways. Here we have a histogram.

ggplot(data=wisclong, aes(x=verb)) +  
 geom\_histogram(binwidth=5, pad = TRUE, fill="white", color="black") +   
 facet\_grid(grade ~ .) +   
 ggtitle("Verbal Ability Score (across grades 1, 2, 4, 6)") +  
 xlab("Verbal Ability Score") +  
 ylab("Density") +  
 theme\_bw() +  
 theme(  
 panel.grid.major = element\_blank(),  
 panel.grid.minor = element\_blank(),  
 strip.background = element\_blank()  
 )

## Warning: Duplicated aesthetics after name standardisation: pad



We can also create notched boxplots of the within-grade distributions (across individuals).

From [Wikipedia](https://en.wikipedia.org/wiki/Box_plot): *Notched box plots apply a notch or narrowing of the box around the median. Notches are useful in offering a rough guide of the significance of the difference of medians; if the notches of two boxes do not overlap, this can provide evidence of a statistically significant difference between the medians.*

Adding the mean value to the plot gives us additonal information about central tendency and skew of the distribution.

#boxplot by grade  
ggplot(data=wisclong, aes(x=factor(grade), y=verb)) +   
 geom\_boxplot(notch = TRUE) +  
 stat\_summary(fun="mean", geom="point", shape=23, size=3, fill="white") +  
 ggtitle("Verbal Ability Score (across grades 1, 2, 4, 6)") +  
 ylab("Verbal Ability Score") +  
 xlab("Grade") +  
 theme\_bw() +  
 theme(  
 panel.grid.major = element\_blank(),  
 panel.grid.minor = element\_blank(),  
 strip.background = element\_blank()  
 )



Finally, we can view overlapping densities of the within-grade distributions of verbal ability scores.

ggplot(data=wisclong, aes(x=verb)) +   
 geom\_density(aes(group=factor(grade), colour=factor(grade), fill=factor(grade)), alpha=0.3) +  
 guides(colour="none", fill=guide\_legend(title="Grade")) +  
 ggtitle("Verbal Ability Score (across grades 1, 2, 4, 6)") +  
 ylab("Density") +  
 xlab("Verbal Ability Score") +  
 theme\_bw() +  
 theme(  
 panel.grid.major = element\_blank(),  
 panel.grid.minor = element\_blank(),  
 strip.background = element\_blank()  
 )



Notice in these plots how much “change” there is at the sample level across grades. Is that expected?

## 3.3 Describing Covariances

In the previous section we looked at the means and variances. Because these are repeated measures, we can also look at covariances and correlations over time. A simple covariance and correlation matrix of the verbal scores across grades can be produced using the cov() and cor() function.

cov(wiscwide[,c("verb1","verb2","verb4","verb6")], use="complete.obs")

## verb1 verb2 verb4 verb6  
## verb1 33.72932 25.46388 30.88886 40.51478  
## verb2 25.46388 37.28784 33.81957 47.40488  
## verb4 30.88886 33.81957 53.58070 62.25489  
## verb6 40.51478 47.40488 62.25489 113.74332

cor(wiscwide[,c("verb1","verb2","verb4","verb6")], use="complete.obs")

## verb1 verb2 verb4 verb6  
## verb1 1.0000000 0.7180209 0.7265974 0.6541040  
## verb2 0.7180209 1.0000000 0.7566242 0.7279080  
## verb4 0.7265974 0.7566242 1.0000000 0.7974552  
## verb6 0.6541040 0.7279080 0.7974552 1.0000000

A plot corresponding to the correlation matrix can be obtained in a number of different ways. First, using the pairs() function from base R.

pairs(wiscwide[,c("verb1","verb2","verb4","verb6")])



There is also a pairs.panel() function in the psych package. Here we see a LOESS smoothed fit line in red.

psych::pairs.panels(wiscwide[,c("verb1","verb2","verb4","verb6")])



Finally, thescatterplotMatrix() from the car ([Fox and Weisberg 2019](#ref-car)) package can be used to create scatterplot matrices with confidence bands around the line of best fit.

car::scatterplotMatrix(~ verb1 + verb2 + verb4 + verb6, data=wiscwide)



Each of these functions can be customized with additional features. Those interested in specifics should consult the help documentation for each function (e.g. ?car::scatterplotMatrix). It is also worth noting the default behavior of these functions is to provide automatic, data-based ranges for each pair of variables separately.

## 3.4 Individual-Level Descriptives

Note that our interest is often in *individual* development, rather than sample development. We need to consider how each individual is changing over time. Thus, we are interested in verbal ability across Time for each individual person. Visualization is typically our best tool for synthesizing the large amounts of information in individual-level data.

ggplot(data = wisclong, aes(x = grade, y = verb, group = id)) +  
 geom\_point() +   
 geom\_line() +  
 scale\_x\_continuous(breaks=seq(1,6,by=1)) +  
 ylim(0,80) +  
 ggtitle("Verbal Ability Score (across grades 1, 2, 4, 6)") +  
 xlab("Grade") +  
 ylab("Verbal Ability Score") +  
 theme\_bw() +  
 theme(  
 panel.grid.major = element\_blank(),  
 panel.grid.minor = element\_blank(),  
 strip.background = element\_blank()  
 )



Sometimes the “blob” gets too dense. This can be fixed by selecting a subset of persons to visualize.

ggplot(subset(wisclong, id < 30), aes(x = grade, y = verb, group = id)) +  
 geom\_point() +   
 geom\_line() +  
 scale\_x\_continuous(breaks=seq(1,6,by=1)) +  
 ylim(0,80) +  
 ggtitle("Verbal Ability Score (across grades 1, 2, 4, 6)") +  
 xlab("Grade") +  
 ylab("Verbal Ability Score") +  
 theme\_bw() +  
 theme(  
 panel.grid.major = element\_blank(),  
 panel.grid.minor = element\_blank(),  
 strip.background = element\_blank()  
 )



We can add some color to our plot using the color argument and treating id as a factor.

ggplot(subset(wisclong, id < 30), aes(x = grade, y = verb, group = id, color = factor(id))) +  
 geom\_point() +   
 geom\_line() +  
 scale\_x\_continuous(breaks=seq(1,6,by=1)) +  
 ylim(0,80) +  
 ggtitle("Verbal Ability Score (across grades 1, 2, 4, 6)") +  
 xlab("Grade") +  
 ylab("Verbal Ability Score") +  
 theme\_bw() +  
 theme(  
 panel.grid.major = element\_blank(),  
 panel.grid.minor = element\_blank(),  
 strip.background = element\_blank(),  
 legend.position = "none"  
 )



We can also get a gradient of colors by treatingid as continuous.

ggplot(subset(wisclong, id < 30), aes(x = grade, y = verb, group = id, color = id)) +  
 geom\_point() +   
 geom\_line() +  
 scale\_x\_continuous(breaks=seq(1,6,by=1)) +  
 ylim(0,80) +  
 ggtitle("Verbal Ability Score (across grades 1, 2, 4, 6)") +  
 xlab("Grade") +  
 ylab("Verbal Ability Score") +  
 theme\_bw() +  
 theme(  
 panel.grid.major = element\_blank(),  
 panel.grid.minor = element\_blank(),  
 strip.background = element\_blank(),  
 legend.position = "none"  
 )



It is also sometimes useful to look at the collection of individual-level plots.

ggplot(subset(wisclong, id <= 20), aes(x = grade, y = verb)) +  
 geom\_point() +   
 geom\_line() +  
 scale\_x\_continuous(breaks=seq(1,6,by=1)) +  
 ylim(0,80) +  
 ggtitle("Verbal Ability Score (across grades 1, 2, 4, 6)") +  
 xlab("Grade") +  
 ylab("Verbal Ability Score") +  
 theme\_bw() +  
 facet\_wrap( ~ id) +  
 theme(  
 panel.grid.major = element\_blank(),  
 panel.grid.minor = element\_blank(),  
 strip.background = element\_blank(),  
 legend.position = "none"  
 )



Some other aesthetics to get to the formal APA style.

#ggplot version .. see also http://ggplot.yhathq.com/docs/index.html  
ggplot(subset(wisclong, id <= 20), aes(x = grade, y = verb, group = id)) +  
 geom\_point() +   
 geom\_line() +  
 xlab("Grade") +   
 ylab("WISC Verbal Score") +   
 ylim(0,100) +  
 scale\_x\_continuous(breaks=seq(1,6,by=1)) +  
 ggtitle("Intraindividual Change in Verbal Ability") +  
 theme\_classic() +   
 #increase font size of axis and point labels  
 theme(axis.title = element\_text(size = rel(1.5)),  
 axis.text = element\_text(size = rel(1.2)),  
 legend.position = "none")



Saving the plot file. See also [outputting plots to a file](http://www.cookbook-r.com/Graphs/Output_to_a_file/).

ggsave(filename = "wiscverbal.png", width = 5, height = 5, dpi=300)

Now we have a good set of strategies to apply when looking at new longitudinal data.

## 3.5 References

# 4 Matrix Algebra

In Chapter 4 we will briefly review some basic algebra results useful for this course. Those needing a reliable reference for basic results in matrix algebra should consult the *The Matrix Cookbook* at <https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf>.

## 4.1 Types of matrices

Remember that matrices are defined by rows (the first dimension) and columns (the second dimension):

You can refer to a specific element in matrix using a subscript of the row and column index (e.g. ). For our purposes there are a few *special* matrices worth mentioning,

### 4.1.1 Square

A square matrix has the same number of rows and columns. Covariance and correlation matrices are square.

### 4.1.2 Symmetric

A symmetric matrix is a square matrix that equals its transpose. This means that corresponding entries on either side of the main diagonal are equal.

**Matrix Transpose**

The transpose of a matrix is an operator which flips a matrix over its diagonal. That is, it switches the row and column indices of the matrix by producing another matrix, often denoted by (or ).

**Graphical Depiction of a Matrix Transpose**



<https://leetcode.com/problems/transpose-matrix/>

### 4.1.3 Diagonal

A diagonal matrix is a special case of a square symmetric matrix in which there are values along the diagonal, but zeros elsewhere:

### 4.1.4 Identity

An identity matrix is a special case of a diagonal matrix in which the elements of the diagonal are all 1:

Any matrix multiplied by an identity matrix is unchanged.

## 4.2 Operations on Matrices

### 4.2.1 Matrix Transpose

As stated earlier the transpose of a matrix is an operator which flips a matrix over its diagonal. That is, it switches the row and column indices of the matrix by producing another matrix, often denoted by (or ). Some useful properties of the matrix transpose include:

$$
(\mathbf{A + B})' = \mathbf{A' + B'}\\
(c\mathbf{A'}) = c(\mathbf{A'}) = (\mathbf{A'})c \\
(\mathbf{A'B}) = \mathbf{B'A}\\
(\mathbf{AB})' = \mathbf{B'A'}\\
(\mathbf{A'})' = \mathbf{A}
$$

**Graphical Depiction of a Matrix Transpose**



<https://leetcode.com/problems/transpose-matrix/>

### 4.2.2 Matrix Trace

The *trace* of a square matrix is the sum of elements along the diagonal. The trace is only defined for a square matrix. For an matrix the trace is defined as follows:

**Graphical Depiction of a Matrix Trace**



Some useful properties of the matrix trace include:

$$
tr(\mathbf{A + B}) = tr(\mathbf{A}) + tr(\mathbf{B})\\
tr(c\mathbf{A}) = c(tr(\mathbf{A})) \\
tr(\mathbf{A}) = tr(\mathbf{A'})\\
tr(\mathbf{AB}) = tr(\mathbf{BA})\\
tr(\mathbf{ABC}) = tr(\mathbf{CAB})=tr(\mathbf{BCA})
$$

### 4.2.3 Addition

For addition, matrices must be of the same order. Addition of two matrices is accomplished by adding corresponding elements,

$$
\mathbf{A} = \begin{bmatrix}
10 & 5 \\
9 & 1
\end{bmatrix}
, \enspace
\mathbf{B} = \begin{bmatrix}
2 & 1 \\
20 & 0
\end{bmatrix}, \enspace
\textrm{then }
\mathbf{A}+\mathbf{B}=
\begin{bmatrix}
12 & 6 \\
29 & 1
\end{bmatrix}
$$

Matrix addition is *commutative* (gives the same result whatever the order of the quantities involved),

and *associative* (gives the same result whatever grouping their is, as long as order remains the same),

and

### 4.2.4 Subtraction

Like addition, subtraction requires matrices of the same order. Elements in the difference matrix are given by the algebraic difference between corresponding elements in matrices being subtracted:

$$
\mathbf{A} = \begin{bmatrix}
10 & 5 \\
9 & 1
\end{bmatrix}
, \enspace
\mathbf{B} = \begin{bmatrix}
2 & 1 \\
20 & 0
\end{bmatrix}, \enspace
\textrm{then }
\mathbf{A}-\mathbf{B}=
\begin{bmatrix}
8 & 4 \\
-11 & 1
\end{bmatrix}
$$

### 4.2.5 Matrix Multiplication

Three useful *rules* to keep in mind regarding matrix multiplication:

1. Only matrices of the form are conformable for multiplication. The number of columns in the premultiplier must equal the number of rows in the post multiplier.
2. The product matrix will have the following order: .

**Graphical Depiction of Rules 1 and 2**



1. The element in the product matrix is the result of multiplying row of the premultiplier matrix, and row of the post multiplier matrix (e.g. ()).

**Graphical Depiction of Rule 3**



<https://code.kx.com/q/ref/mmu/>

Matrix multiplication is associative (i.e. rearranging the parentheses in an expression will not change the result). That is,

and is distributive with respect to addition,

$$
\mathbf{A(B+C)} = \mathbf{AB + AC} \\
\mathbf{(B+C)A} = \mathbf{BA + CA} \\
$$

If is a scalar, then

or equivalently,

$$
\mathbf{A} = \begin{bmatrix}
10 & 5 \\
9 & 1
\end{bmatrix}, \enspace
k=2, \enspace
k\mathbf{A} = \begin{bmatrix}
20 & 10 \\
18 & 2
\end{bmatrix}.
$$

In general, matrices that can be multiplied are called ‘compatible’ or ‘comformable.’ Matrices in which the *inner dimensions* (i.e., columns of , rows of ) do not match are called ‘incompatible’ or ‘non-conformable.’ These cannot be multiplied.

### 4.2.6 Matrix Division

Division is not defined for matrix operations, but may be accomplished by multiplication by the inverse matrix. In algebra, the reciprocal of a scalar is, by definition, the scalar raised to the minus one power (e.g. ), and equations may be solved by multiplication by reciprocals.

For example:

$$
5^{-1} = 1/5\\
5x=35\\
5^{-1}(5x)=5^{-1}(35)\\
x = 7
$$

Now consider the following equation where the vector is unknown,

Each element in the column vector is unknown and the solution involves solving a set of simultaneous equations for the unknown element of ,

$$
a\_{11}x\_{1} + a\_{12}x\_{2} + \dots + a\_{1p}x\_{p} = b1 \\
a\_{21}x\_{1} + a\_{22}x\_{2} + \dots + a\_{2p}x\_{p} = b2 \\
\vdots \\
a\_{p1}x\_{1} + a\_{p2}x\_{2} + \dots + a\_{pp}x\_{p} = bp
$$

A solution analogous to the scalar equations above would give the following solution for the elements of the vector :

$$
\mathbf{A}\_{p \times p} \mathbf{x}\_{p \times 1} = \mathbf{b}\_{p \times 1} \\
\mathbf{A}^{-1}\_{p \times p}\mathbf{A}\_{p \times p} \mathbf{x}\_{p \times 1} = \mathbf{A}^{-1}\_{p \times p}\mathbf{b}\_{p \times 1} \\
\mathbf{I}\_{p \times p}\mathbf{x}\_{p \times 1} = \mathbf{A}^{-1}\_{p \times p}\mathbf{b}\_{p \times 1} \\
\mathbf{x}\_{p \times 1} = \mathbf{A}^{-1}\_{p \times p}\mathbf{b}\_{p \times 1}
$$

The inverse of a matrix must satisfy the following properties:

where is the identity matrix with 1’s along the diagonal and 0’s elsewhere.

So, why is division undefined for matrices. Here is a quick example. Suppose, is a matrix and is the inverse of , such that

Now, let

$$
\mathbf{A} = \begin{bmatrix}
1 & 2 \\
1 & 2
\end{bmatrix}
, \enspace
\mathbf{B} = \begin{bmatrix}
a & b \\
c & d
\end{bmatrix},
$$

Then,

This means that and , which is a contradiction, suggesting A does not have an inverse.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>.

McArdle, J. J. 1988. “Dynamic but Structural Equation Modeling of Repeated Measures Data.” In *Handbook of Multivariate Experimental Psychology, 2nd Ed*, 561–614. Perspectives on Individual Differences. New York, NY, US: Plenum Press. <https://doi.org/10.1007/978-1-4613-0893-5_17>.

Mcardle, J. J., and MARK S. Aber. 1990. “Chapter 5 - Patterns of Change Within Latent Variable Structural Equation Models.” In *Statistical Methods in Longitudinal Research*, edited by Alexander von Eye, 1:151–224. Statistical Modeling and Decision Science. San Diego: Academic Press. <https://doi.org/10.1016/B978-0-12-724960-5.50010-X>.

McArdle, J. J., and David Epstein. 1987. “Latent Growth Curves Within Developmental Structural Equation Models.” *Child Development* 58 (1): 110–33. <https://doi.org/10.2307/1130295>.

McArdle, J. J., and John R. Nesselroade. 1994. “Using Multivariate Data to Structure Developmental Change.” In *Life-Span Developmental Psychology: Methodological Contributions*, 223–67. The West Virginia University Conferences on Life-Span Developmental Psychology. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Osborne, R. T., and D. E. Suddick. 1972. “A Longitudinal Investigation of the Intellectual Differentiation Hypothesis.” *The Journal of Genetic Psychology: Research and Theory on Human Development* 121 (1): 83–89. <https://doi.org/10.1080/00221325.1972.10533131>.

Revelle, William. 2021. *Psych: Procedures for Psychological, Psychometric, and Personality Research*. Evanston, Illinois: Northwestern University. <https://CRAN.R-project.org/package=psych>.

Tierney, Nicholas, Di Cook, Miles McBain, and Colin Fay. 2021. *Naniar: Data Structures, Summaries, and Visualisations for Missing Data*. <https://CRAN.R-project.org/package=naniar>.

Wechsler, David. 1949. *Wechsler Intelligence Scale for Children*. Wechsler Intelligence Scale for Children. San Antonio, TX, US: Psychological Corporation.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>.

———. 2021. *Tidyr: Tidy Messy Data*. <https://CRAN.R-project.org/package=tidyr>.