HDFS 523: Strategies for Data Analysis in Developmental Research

Zachary F. Fisher

2024-02-12

# 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.

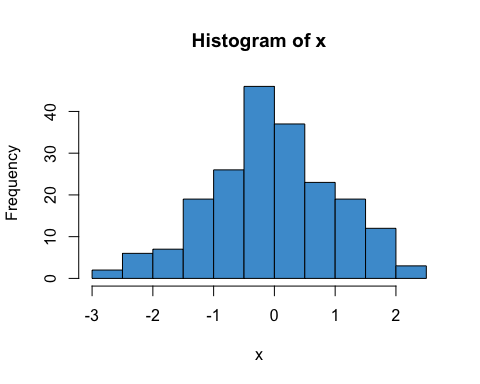
## 1.1 Why this book?

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.

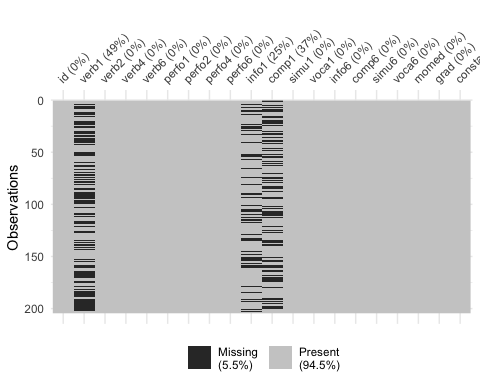
na\_strings <- c("NA", "N/A", -99)  
   
wisc\_miss <- naniar::replace\_with\_na\_all(  
 wisc\_miss, condition = ~.x %in% na\_strings  
)

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)



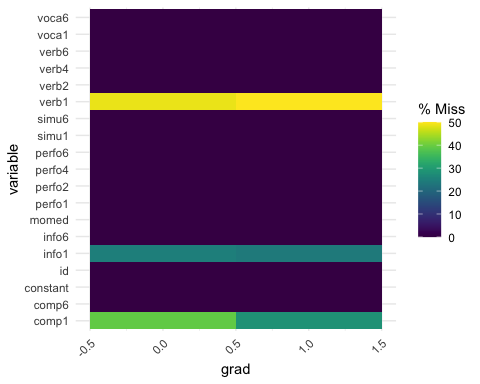
It is often useful to look at combinations of missingness among different variables.

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.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



### 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,

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:

### 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,

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

Then,

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

## 4.3 References

# 5 Ordinary Least Squares

In Chapter 5 we will briefly review the assumptions and properties of Ordinary Least Squares (OLS) regression, a cornerstone method that supports many of the other methods we will consider. We will present the regression model in both scalar and matrix forms to facilitate the material to follow.

## 5.1 Linear Regression Model

Generally, the regression model is written as

where

* is the value of the outcome variable for individual
* is an *intercept parameter*, the expected value of when the predictor variables are all
* is a regression parameter indicating the relation between and the outcome variable,
* are *errors or disturbances*

## 5.2 Ordinary Least Squares (OLS)

The Ordinary Least Squares (OLS) is one of the most common estimators of the linear regression model. What assumptions do we make with OLS? Why should we care?

How well me meet the assumptions of OLS determines (1) the accuracy of OLS coefficients, and (2) the accuracy of our inferences and substantive hypothesis tests.

## 5.3 Assumptions of OLS

The assumptions of OLS are as follows:

1. for all (homoscedasticity)
2. for all
3. No perfect collinearity among variables
4. for all and

Let’s discuss each assumption in more detail.

### 5.3.1 Assumption 1.

Note that is the expectation operator. The expected value is an “average” of whatever is inside the parentheses. This assumption states that, on average, the error for the observation is zero. Here “for all ” means the same is true for all cases.

### 5.3.2 Assumption 2. Homoscedasticity

In statistics, a vector of random variables is heteroscedastic if the variability of the random disturbance is different across elements of the vector, here our s. The errors or disturbances in our model are homoskedastic if the variance of is a constant (e.g. ), otherwise, they are heteroskedastic.

**Graphical Depiction of Homoskedasticity and Heteroskedasticity**



<https://en.wikipedia.org/wiki/Homoscedasticity>

### 5.3.3 3.

Assumption 3 is sometimes referred to as the *autocorrelation assumption*. This assumption states that the error terms of different observations should not be correlated with each other. For example, when we have time series data and use lagged variables we may want to examine residuals for the possibility of autocorrelation.

**Graphical Depiction of Positive and Negative Autocorrelation**



### 5.3.4 4. No Perfect Collinearity

Perfect collinearity occurs when one variable is a perfect linear function of any other explanatory variable. If perfect collinearity is found among the s then has no inverse and OLS estimation fails. Perfect collinearity is unlikely except for programming mistakes such as dummy coding all the values in a nominal variable.

### 5.3.5 5.

Note that is the covariance operator. Assumption five states that that the error of our equation is uncorrelated with all the s. This is often referred to as an *endogeneity assumption*.

This can be a confusing assumption because by definition the residuals are uncorrelated with the s. Here, however, we are concerned with the true errors . Unfortunately, there are a variety of conditions that lead to in applied contexts.

**Graphical Depiction of Sources of Endogeneity**



Figure 2 from Bollen (2012)

If we meet these assumptions what large sample properties can we expect?

## 5.4 Properties of the OLS Estimator

If assumptions (1) to (5) hold, then the OLS estimator is:

1. A consistent estimator of
2. Asymptotically normally distributed
3. Having a variance of

*Notice that we did not assume normality of or .*

Let’s discuss each of these properties in a little bit more detail.

### 5.4.1 1. Consistentcy of

is the OLS estimator of . A consistent estimator is one for which, as the sample size () increases, the estimate converges in probability to the value that the estimator is designed to estimate. This is often stated as . Stated differently, as the sample size grows, the OLS coefficients converge to the true coefficients.

### 5.4.2 2. Asymptotic Normality

Asymptotic normality is another property of the OLS estimator when all assumptions are met. “Asymptotic” refers to how an estimator behaves as the sample size tends to infinity. “Normality” refers to the normal distribution, so an estimator that is asymptotically normal will have an approximately normal distribution as the sample size gets larger.

### 5.4.3 Variance of

Having a variance of is another property of the OLS estimator when the previously stated assumptions are met. This means, for example, we can estimate the standard errors from the main diagonal of and perform significance testing based on this variance.

## 5.5 Failure to Meet Assumptions

It is worth thinking about the consequences of not meeting these assumptions.

### 5.5.1 Failure of Assumption 1.

Assumption 1 states . This assumption states that, on average, the error for the observation is zero. If instead, and , and all other assumptions hold, then only the intercept term is biased. Other coefficients OK.

### 5.5.2 Failure of Assumption 2 or 3.

Assumptions 2 and 3 are the *homoskedasticity* and *no autocorrelation* assumption, respectively. If we violate (2) or (3), but all other assumptions hold, (1) variance of is no longer dependable, (2) SEs possibly inaccurate, and (3) significance tests are possibly inaccurate. However, importantly, is still an unbiased and consistent estimator.

### 5.5.3 Failure of Assumption 5.

Assumption (5) states that that the error of our equation is uncorrelated with all the s. If this assumption fails, while others hold, OLS is no longer a consistent estimator.

## 5.6 Regression and Matrix Notation

Now that we have reviewed the assumptions of OLS, let’s return to the linear regression model and translate it into a matrix form.

### 5.6.1 An Intercept-Only Model

First, let’s take a simpler form of the model, an intercept-only model where

Note that we have made the “silent” 1 explicit. This will become important later (e.g., when fitting growth models). It is worthwhile to look at regression model without predictors to understand what it can tell us about the nature of the intercept (or constant).

So here we have no predictors, what is ?

Here, is the mean of the response variable, and we can show this with some algebra,

where (Assumption 1).

### 5.6.2 Intercept-Only Model in Matrix Form

Translating into matrix form, can be written as an x 1 matrix (a column vector). More specifically, for to individuals,

.

(Remember, matrices are often designated as bold capital letters)

Doing the same for all the other parts of the model, we get

Note that we have taken care that each matrix is of an order that will allow for matrix multiplication.

### 5.6.3 Simple Regression in Matrix Form

Now, let’s expand our regression model by adding a predictor . Our model becomes

Written out explicitly in matrix form, the model is

### 5.6.4 Multiple Regression in Matrix Form

Finally, extending the model to the general case with predictor variables, we have

which is written out in matrix form as

Where we have the following elements:

Observe the order of the matrices/vectors. On the right hand side you are matrix multiplying a matrix with a vector. This yields an vector, to which another vector is added, and this is equal to our outcome vector which is also .

When we implement this model in R, it will be important to know the portions of the model that are in our data frame, and , and to have them structured properly. This will become clear in the examples below.

Now that we have the model written out explicitly as matrices, we can easily simplify the notation.

In compact matrix notation, the regression model then can be written as

## 5.7 Solving the Regression Equation

In practice, we would like to know the contents of (i.e., solve for) .

Assuming the model is correct, the expected value of is 0, therefore,

Then we just need to solve for . We can think back about some of the matrix operations we discussed earlier.

### 5.7.1 Matrix Multiplication and Transpose

Our goal is to isolate . One initial idea might be to multiple each side of the equation by in an attempt to remove from the right hand side, and isolate . Why won’t this work?

Instead, let’s pre-multiply each side of the equation by . This would give us

This gets us a quantity, , a square matrix containing information about the relations among the s.

### 5.7.2 Matrix Inverse

Now, since is a square matrix and presumabely has an inverse (e.g. no perfect collinearity), we can premultiply both sides by , to obtain

Remembering our assumptions that a matrix multiplied by its inverse equals the identity matrix, the equation simplifies to

or more succinctly

We’ve now isolated the unknowns, onto one side of the equation and figured out how to use matrix algebra to obtain the regression coefficients. Quite literally, this algebra is what allows for estimation of the parameters when fitting a regression model to data.

We will now work through some practical examples - staying aware that this kind of matrix algebra is being done in the background.

## 5.8 The Linear Probability Model

While we are discussing the assumptions of OLS it is worth pausing to consider a model for dichotomous outcomes: the *linear probability model* (LPM).

In the LPM we don’t do anything fancy with a binary outcome variable. Instead, we simply apply OLS as we would with a continuous out come variable. Since we aren’t considering the normality of our outcome you might be curious how our assumptions would hold.

Remembering the assumptions of OLS:

1. for all (homoscedasticity)
2. for all
3. No perfect collinearity among variables
4. for all and

**Which assumptions are needed for consistency and asymptotic unbiasedness?**

1. No perfect collinearity among variables
2. Errors uncorrelated with all variables.

In regard to (4) having a dependent variable valued at does not cause any problems. In regard to (5), again no, nothing about a dichotomous outcome violates this assumption.

Therefore, in this model is still consistent and asymptotically unbiased.

**What about the remaining assumptions?**

1. Homoscedasticity

Here, *a dichotomous outcome does inherently violate the assumption of homoskedasticity.*

Why is this case? It can be shown that now directly depends on the value of that is taken. If the person has one set of values for the variables, and another individual has another set, the estimates of and will differ. This can be seem by looking at the variance of a Bernoulli random variable.

This means estimate of variance of is no longer reliable, SEs and significance tests possibly inaccurate. However, this could be addressed using robust standard errors.

It is common to look at plots of predicted values vs residuals to diagnose heteroskedasticity. Generally one would like to see a random blob of points without any discernible pattern. Here is an example of what that plot might look like for an LPM model. Each line represents a different outcome, , or .



### 5.8.1 Advantages of the LPM

1. Simplicity.
2. Regression coefficients give impact of on .
3. Effect same regardless of value of or values of other s.
4. Can extend with traditional methods easily (interactions, quadratic terms).

### 5.8.2 Disadvantages of the LPM

1. Functional form unlikely accurate at extreme lows & highs
2. Binary dependent variable creates heteroscedasticity
3. Some regression diagnostics assume homoscedastic error
4. Predicted probabilities not restricted to 0 to 1 range



# 6 Statistical Control

In Chapter 6 we will discuss statistical control.

* Introduction to DAGs
* Statistical Control Done Right
* Statistical Control Gone Wrong

Much of this presentation was drawn directly from Rohrer (2018) (<https://journals.sagepub.com/doi/full/10.1177/2515245917745629>).

## 6.1 Statistical Control

**Randomized experiments often considered the gold standard in scientific research.However, it is often infeasible, difficult or impossible to manipulate the putative effect of interest**:

* cannot randomly resettle individuals into different strata of society
* cannot assign couples to stay married or get a divorce
* cannot randomize children to different levels of adversity

**Researchers have employed a number of different strategies in response to the limitations of observational data**:

* surrogate interventions
  + item the real-life cause of interest cannot be manipulated, often a proxy can be randomized in the lab.
* avoiding causal language
  + item write about associations and relationships despite substantive question about causal effects
* statistical control
  + include control variables in analysis but choice of variables often difficult and statistical control itself insufficient

## 6.2 Directed Acyclic Graphs (DAGs)

**Directed acyclic graphs (DAGs) are another approach that can be used to examine causal inferences from observational data.**

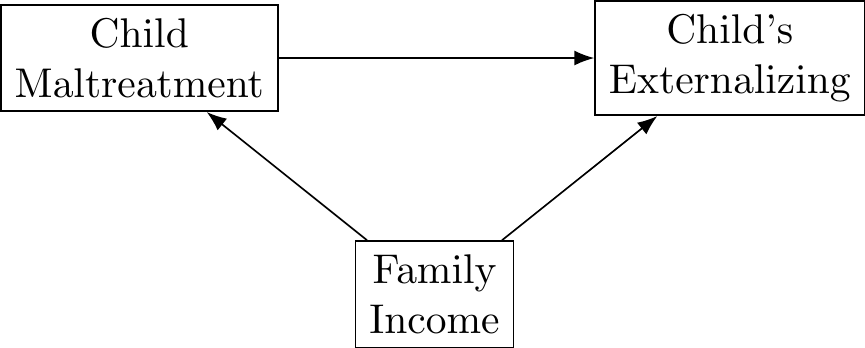
* They were developed primarily by the computer scientist Judea Pearl
* DAGs provide a visual representation of causal assumptions.
* Some overlap with structural equation models (SEMs).

**Importantly, DAGs can provide insights on**

* What variables should be controlled for?
* What variables should not be controlled for?
* In what situations might control worsen causal inference?

### 6.2.1 Introduction to DAGs

**Below is a simple DAG depicting a model in which the relationship between maltreatment externalizing is confounded by a common cause, income.**



DAGs consist of **nodes** (variables) and **arrows** (edges) between these nodes, which reflect causal relationships.

It is assumed that manipulation of a variable at which an arrow begins (e.g., a manipulation of child maltreatment with income held constant) would change the variable at the end of the arrow (e.g., externalizing).

### 6.2.2 Introduction to DAGs: Paths

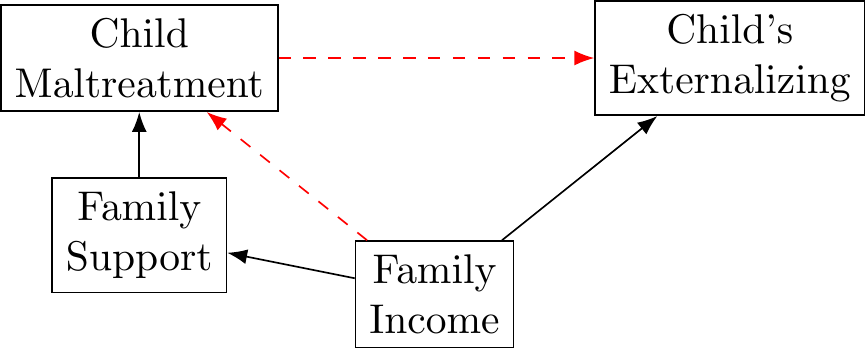
From these two simple building blocks—**nodes** and **arrows**—one can visualize more complex situations and trace paths from variable to variable:



**Paths**

* A simple path leads just from one node to another (income → stress).
* Paths can also contain multiple nodes:
  + income → stress → child maltreatment
* Paths can also travel against the direction indicated by arrows
  + child maltreatment ← stress ← income → externalizing

### 6.2.3 Introduction to DAGs: Chains



**Chains**

* Chains have the structure A → B → C.
* Chains can transmit an association between the beginning and end nodes.
  + If income causally affects child maltreatment, and child maltreatment causally affects externalizing, then income and externalizing can be correlated.

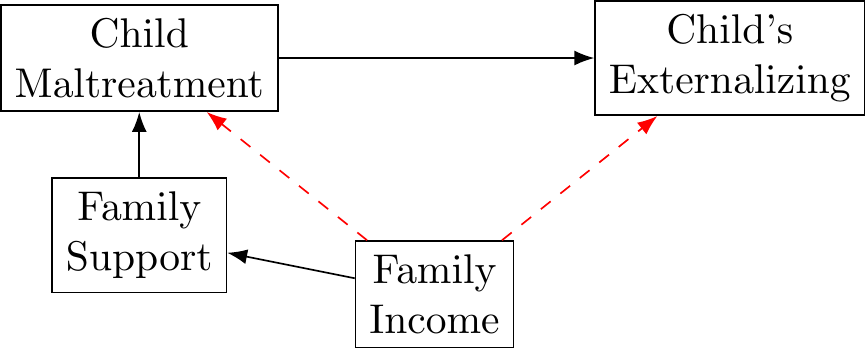
### 6.2.4 Introduction to DAGs: Descendants and Ancestors



**Chains: Descendants and Ancestors**

* Along a chain, variables that are directly or indirectly causally affected by a certain variable are called its descendants
  + externalizing is a descendant of child maltreatment
* Variables that directly or indirectly affect a certain variable are considered its ancestors.
  + income is an ancestor of support, child maltreatment and internalizing

### 6.2.5 Introduction to DAGs: Forks



**Forks**

* Forks have the structure A ← B → C.
* A fork can transmit an association, but it is not causal.
  + In isolation, this fork indicates that child maltreatment and externalizing may be correlated because they share a common cause, income.
* Forks are the causal structure most relevant for the phenomenon of confounding.

### 6.2.6 Introduction to DAGs: Inverted Forks



**Inverted Forks**

* Inverted forks have the structure A → B ← C.
* An inverted fork does not transmit an association.
  + In isolation, If child maltreatment and income both affect externalizing, this does not imply that they are in any way correlated.
* Inverted forks are relevant to the problem of collider bias.

### 6.2.7 Introduction to DAGs: Acyclicity



**Acyclicity**

* DAGs are acyclic because they do not allow for cyclic paths in which variables become their own ancestors.
  + a variable cannot causally affect itself
* Developmental systems often contain feedback loops and reciprocal relationships.
  + Often feedback loops can be modeled in a DAG by taking the temporal order into account and adding nodes for repeated measures.

## 6.3 Statistical Control Done Right

**The central problem of observational data is confounding**:

* the presence of a common cause that lurks behind the potential cause of interest and the outcome of interest.

**A confounding influence can introduce what is often called a spurious correlation, which ought not to be confused with a causal effect.**

* The extraordinarily influence of randomized experiments in testing causal inferences is based on the simple fact that if the independent variable is randomly assigned—for example, by the flip of a coin—by design it cannot share a common cause with the outcome.

**How can a DAG be used to figure out how to remove all such noncausal associations so that only the true causal effect remains?**

### 6.3.1 Building a DAG

**To derive a valid causal conclusion, one must ensure the DAG includes everything that is relevant to the causal effect of interest. What is missing**



If we want to derive a valid causal conclusion, we need to build a causal DAG that is complete because it includes all common causes of all pairs of variables that are already included in the DAG (Spirtes, Glymour, & Scheines, 2000).

That is, any additional variable that either directly or indirectly causally affects at least two variables already included in the DAG should be included.

### 6.3.2 Building a DAG: Back-Door Paths

**After a DAG is built, back-door paths can be discerned.**

Back-door paths are all paths that start with an arrow pointing to the independent variable and end with an arrow pointing to the dependent variable.



If we plan to investigate the causal relationship between child maltreatment and internalizing what are the back-door paths in our example DAG?

**Back-Door Paths**

* child maltreatment ← support ← income → externalizing
* child maltreatment ← income → externalizing

**Back-Door Problems**

* Back-door paths are problematic whenever they transmit an association.
* In this case, both backdoor paths consist of only chains and forks, thus, these two back-door paths are open, and they can transmit a spurious association.

**Back-Door Solutions**

* The zero-order correlation between child maltreatment and externalizing is a mix of the true causal effect (child maltreatment → externalizing) of interest plus any noncausal association transmitted by the two back-door paths.
* To remove the undesirable noncausal association, we must block the two back-door paths.

**Blocking Back-Door Paths**

**The purpose of third-variable control is to block open back-door paths.**

* If all back-door paths between the independent and dependent variables can be blocked, then the causal effect connecting the independent and dependent variables can be identified.
* Such a causal effect would be considered identifiable, always under the assumption that the DAG captures the true underlying causal web.
* A back-door path can be blocked by “cutting” the transmission of association at any point in the path by statistically controlling a node.

**What variables would we want to control for to identify the causal effect of child maltreatment on externalizing?**



## 6.4 Statistical Control Gone Wrong

In certain fields, it has become common practice to include as many covariates as possible.

Unfortunately, it is not true that simply adding more covariates will improve the estimate of a causal effect.

**There are two types of variables that researchers should not control for without taking into account potential negative side effects: colliders and mediators.**

Whereas confounders causally affect the independent variable of interest, colliders and mediators are causally affected by the independent variable.

### 6.4.1 Collider Bias

A collider for a certain pair of variables is any variable that is causally influenced by both of them.

**Controlling for, or conditioning analysis on, such a variable (or any of its descendants) can introduce a spurious (i.e., noncausal) association between its causes.**

In DAG terminology, a collider is the variable in the middle of an inverted fork, for example, variable B in A → B ← C.

The collider variable normally blocks the path, but when one controls for it, a spurious association between A and C can arise.

### 6.4.2 Conditioning on a Collider

**Thought Experiment**

Imagine we are interested in the effect of child maltreatment on personality features, such as extroversion.

For this thought experiment, let us assume that there is actually no causal effect of child maltreatment on extroversion.

To investigate the association, we look at all individuals with substantiated claims of maltreatment via CPS.

**We find a sizable negative association: those who experienced more maltreatment show less extroversion and vice versa.**

Suppose we then realize bias might be an issue and conduct a follow-up study on individuals who self-report experiencing maltreatment but do not have CPS involvement.

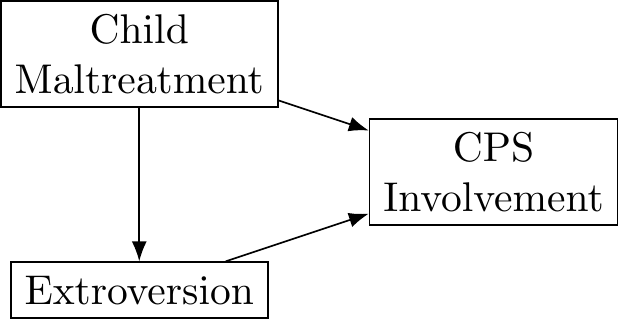
**Again we find a sizable negative association.**

By assessing substantiated and unsubstantiated cases separately, we have stratified, or conditioned, our analyses by CPS involvement.

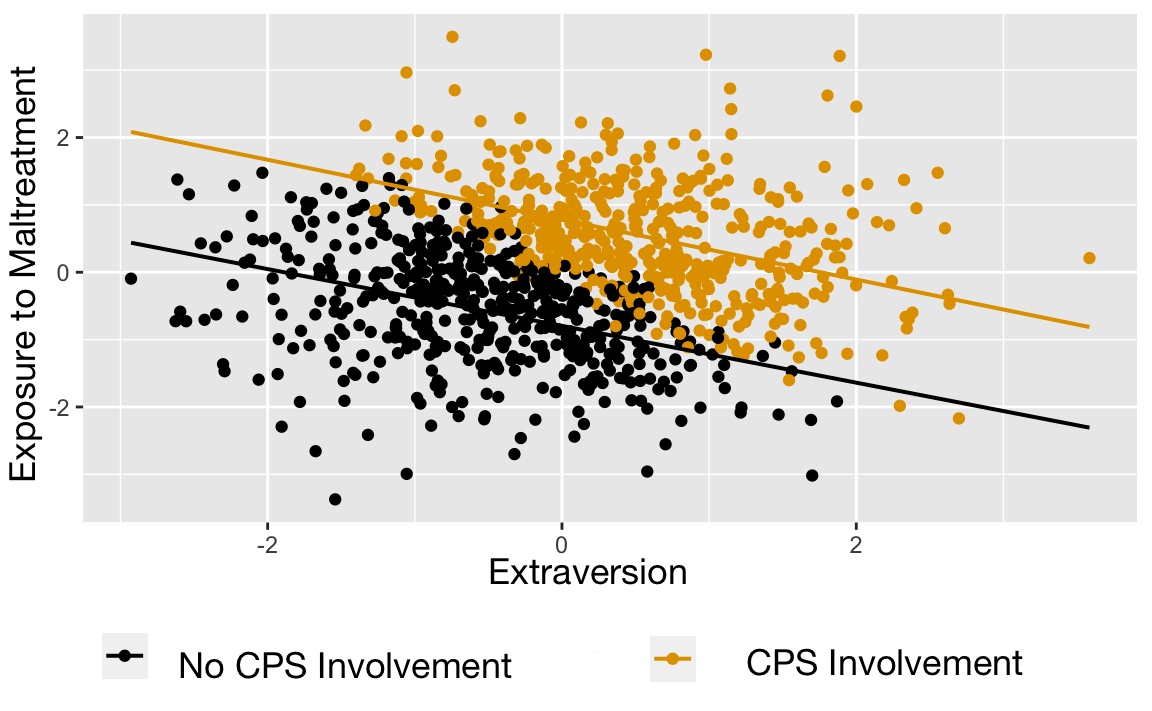
However, let’s assume that exposure to child maltreatment and extroversion are likely to be associated with CPS contact.

In the simplest case, both have a positive effect:

* With increasing child maltreatment, the likelihood of CPS involvement increases.
* With increasing extroversion, the likelihood of CPS involvement increases.



**In our thought experiment, there is no association between child maltreatment and extroversion if all individuals—with and without CPS involvement—are considered simultaneously without statistical control (of the collider or any descendants)**



Collider Bias

The spurious negative correlation emerges only when the joint outcome of the two variables of interest is controlled for.

This observation generalizes to similar situations in which selection into a group is based on multiple desirable features:

* Group membership is a collider variable, and conditioning analysis on it will introduce or exaggerate trade-offs between desirable features.

### 6.4.3 Avoiding Collider Bias

Avoiding collider bias requires two steps.

* One must be aware of the collider variable, and this may entail using a DAG to identify colliders that exist between and independent variable and outcome.
* One must be able to run analyses that are not conditional on the collider. This entails not controlling for the collider when examining the main effect of interest.

**Thought Experiment**

* In our thought experiment, we must include individuals involved and not involved with CPS
* Outside of thought experiments, one might often be unaware of collider variables or collect data in such a way that collider bias is built in.

### 6.4.4 Variations on Collider Bias: Nonresponse Bias

Nonresponse bias occurs if, for example, a researcher analyzes only completed questionnaires, and the variables of interest are associated with questionnaire completion.

Assume that we are interested in the association between grit and intelligence, and our assessment ends up being very burdensome.

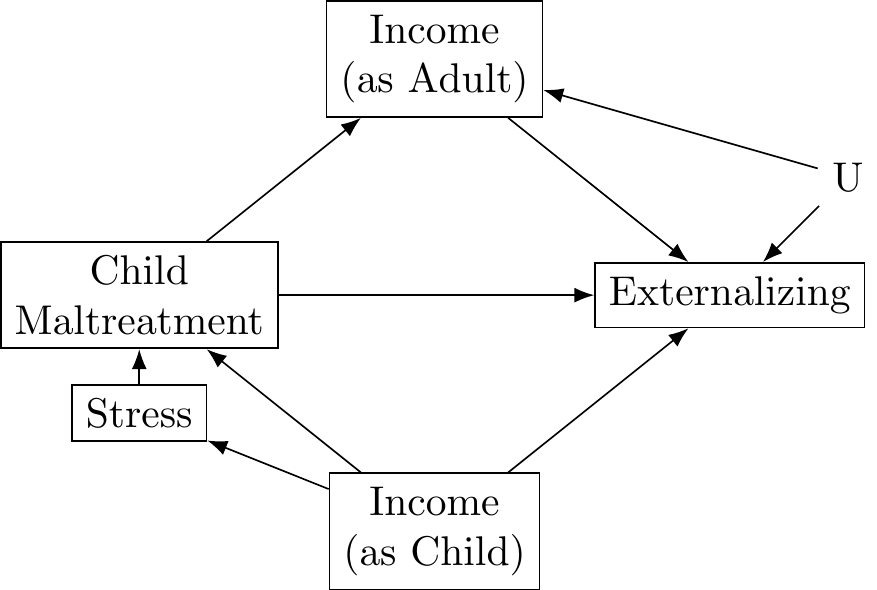
* Both grit and intelligence make it easier for respondents to push through and complete the assessment.
* Questionnaire completion is thus a collider between grit and intelligence.
* Although there might be no association between grit and intelligence in the population, we might find a spurious negative association if we analyze only completed questionnaires.
  + completers low on intelligence and high levels of grit
  + completers low on grit and high on intellgence
  + noncompleter low on both variables less likely to finish

### 6.4.5 Controlling for Mediators

Overcontrol bias is another example of statistical control hurting instead of helping:

* If mediating variables are controlled for, the very processes of interest are controlled away.

Consider our previous example, now slightly modified:



# 7 Linear Regression

In Chapter 7 we will demonstrate how to estimate the linear regression model in R with an eye towards the longitudinal modeling to follow.

## 7.1 Example Data

Chapter 6 make use of the same WISC data used in Chapter 3. Here we again read in, subset, and provide descriptives for the WISC data. We will also add a simulated variable childgrad indicating whether the student graduated highschool.

filepath <- "https://quantdev.ssri.psu.edu/sites/qdev/files/wisc3raw.csv"  
  
wisc3raw <- read.csv(file=url(filepath),header=TRUE)  
  
colnames(wisc3raw) <- c(  
 "id", "verb1", "verb2", "verb4", "verb6", "perfo1", "perfo2",   
 "perfo4", "perfo6", "info1", "comp1", "simu1", "voca1", "info6",   
 "comp6", "simu6", "voca6", "daded", "grad", "constant"  
)  
  
var\_names\_sub <- c(  
 "id", "verb1", "verb2", "verb4", "verb6",  
 "perfo1", "perfo2", "perfo4", "perfo6",  
 "daded", "grad"  
)  
  
wiscsub <- wisc3raw[,var\_names\_sub]  
  
set.seed(1234)  
wiscsub$childgrad <- sample(c(0,1), replace=TRUE, size=nrow(wiscsub))  
  
psych::describe(wiscsub)

## 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  
## daded 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  
## childgrad 12 204 0.55 0.50 1.00 0.56 0.00 0.00 1.00 1.00 -0.20  
## 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  
## daded 0.01 0.19  
## grad -0.30 0.03  
## childgrad -1.97 0.03

## 7.2 Intercept-Only Model

For our first example, we focus on verbal ability at Grade 2 as an outcome (verb2 in the data frame wiscsub). Examining the distribution for ‘verb2’.

library("ggplot2")  
  
psych::describe(wiscsub$verb2)

## 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

ggplot(data=wiscsub, aes(x=verb2)) +   
 geom\_histogram(binwidth=2.5, fill="white", color="black", boundary=0) +  
 xlab("Verbal Ability Grade 2") + ylab("Count") +  
 xlim(0,50) +  
 theme\_classic()



### 7.2.1 Intercept-Only Equation

The simplest model is an intercept only model. In this case, we would fit the model

Written out explicitly with the “silent” 1 in it, we get

This is helpful for explicit translation into the R code, specifically the formula within the lm() function.

### 7.2.2 Intercept-Only Model in R

We fit the model using the following code. Note that the code has the ‘1’ predictor variable stated explicitly.

model1 <- lm(formula = verb2 ~ 1,  
 data = wiscsub,  
 na.action = na.exclude)  
summary(model1)

##   
## Call:  
## lm(formula = verb2 ~ 1, data = wiscsub, na.action = na.exclude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.4653 -4.6403 0.5647 4.2822 14.4347   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 25.4153 0.4275 59.45 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.106 on 203 degrees of freedom

Note that we used na.exclude instead of na.omit (default); practically speaking na.omit deletes missing data entries while na.exclude just excludes from the analysis.Therefore with na.exclude, in the residuals and fitted values, NA will show up where there were missing values.

The output indicates that = 25.4153, and its standard error = 0.4275.

The intercept reflects the expected value of the outcome variable when all of the predictor variables (i.e. ) = 0.

So, in the absence of any additional information other than the descriptive statistics of , what is our best guess for a person’s score?

It is the mean of . The regression above confirms this notion; regressing the outcome on a vector of 1s allows us to ‘recover’ the mean.

### 7.2.3 Intercept as Mean of Outcome

Notice we can confirm this finding using matrix algebra, as well.

From the properties of expectation, we have .

Another property of expectation relates to taking the expectation of a constant, , thus , implying

Remembering Assumption 1, , we have

We can confirm this by looking at Verbal Scores at Wave 2.

mean(wiscsub$verb2)

## [1] 25.41534

### 7.2.4 Intercept-Only Model

Yes - we recovered the *mean*, but we did not attempt to explain any of the *variance*. Let’s take a look at the variance explained for Model 1.

summary(model1)$r.squared

## [1] 0

It thus makes sense that we get 0 as the R-square. From the properties of variance, we know that . There is no variability due to the regression model because there are no predictors, only a constant.

## 7.3 Simple Linear Regression

Let’s build up the model further. For example, we could attempt to explain some of the *between-person variance* in the Grade 2 verbal score from the Grade 1 verbal scores. But, before we do, let’s examine the distribution of the *between-person differences* in the Grade 1 verbal scores.

ggplot(wiscsub, aes(x=verb1)) +   
 geom\_histogram(binwidth=2.5, fill="white", color="black", boundary=0) +  
 xlab("Verbal Ability Grade 1") +   
 ylab("Count") +  
 xlim(0,50) +  
 theme\_classic()

 And the relation between the Grade 2 and Grade 1 verbal ability scores.

ggplot(wiscsub, aes(x=verb1, y = verb2)) +   
 geom\_point() +  
 stat\_ellipse(color="blue", alpha=.7) +  
 xlab("Verbal Ability Grade 1") +   
 ylab("Verbal Ability Grade 2") +  
 ylim(0,45) +   
 xlim(0,45) +  
 theme\_classic()



### 7.3.1 Regression Equation and Model Fitting

Our regression model becomes

model2 <- lm(verb2 ~ 1 + verb1,  
 data = wiscsub,  
 na.action = na.exclude)  
summary(model2)

##   
## Call:  
## lm(formula = verb2 ~ 1 + verb1, data = wiscsub, na.action = na.exclude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.5305 -3.0362 0.2526 2.7147 12.5020   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.62965 1.05164 10.11 <2e-16 \*\*\*  
## verb1 0.75495 0.05149 14.66 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.261 on 202 degrees of freedom  
## Multiple R-squared: 0.5156, Adjusted R-squared: 0.5132   
## F-statistic: 215 on 1 and 202 DF, p-value: < 2.2e-16

### 7.3.2 Path Diagram

We might also be interested in a graphical depiction of our model. This can be accomplished with the semPaths package.

semPlot::semPaths(model2, what = "paths")



### 7.3.3 Interpreting Model Parameters

How do we interpret the parameters here?

*The intercept, , is the expected value for the outcome variable when all of the predictor variables equal zero.* So, we would expect a child to have a Grade 2 verbal score of 10.62965 *if* they have a Grade 1 verbal score of 0.

*The slope, is the expected difference in the outcome variable for each 1-unit difference in the predictor variable*. So, *across children*, for each 1-point difference in a child’s Grade 1 verbal score, we would expect a 0.75 point difference in the Grade 2 verbal score.

### 7.3.4 Plotting Regression Line

We can plot the relation between ‘verb1’ and ‘verb2’, and include the predicted line from the analysis.

ggplot(data=wiscsub, aes(x=verb1,y=verb2)) +  
 geom\_point(size = 2, shape=19) +  
 geom\_smooth(method=lm,se=TRUE,fullrange=TRUE,colour="red", size=2) +  
 labs(x= "Verbal Ability Grade 1", y= "Verbal Ability Grade 2") +  
 xlim(0,50) +  
 ylim(0,50) +  
 theme\_bw() +  
 theme(  
 plot.background = element\_blank(),  
 panel.grid.major = element\_blank(),  
 panel.grid.minor = element\_blank(),  
 panel.border = element\_blank()  
 ) +  
 #draws x and y axis line  
 theme(axis.line = element\_line(color = 'black')) +  
 #set size of axis labels and titles  
 theme(axis.text = element\_text(size=12),  
 axis.title = element\_text(size=14))

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.

## `geom\_smooth()` using formula = 'y ~ x'



## 7.4 Mean Centering Predictors

In this case, and in many other cases, the intercept does not have a ‘useful’ interpretation for the empirical example. This is because no students had a Grade 1 verbal score equal to 0.

Therefore, if we want to make the intercept more meaningful, we need to make a Grade 1 verbal score with a more meaningful 0 point. Typically we center the *predictor* variables in regression analysis.

For example, we create a centered variable, by subtracting the sample mean, from each observation,

Our model becomes

We can sample-mean center in R as follows

#calculate the mean centered variable  
wiscsub$verb1\_star <- wiscsub$verb1 - mean(wiscsub$verb1, na.rm = TRUE)

Then we can fit a new model using , such that

model3 <- lm(verb2 ~ 1 + verb1\_star,  
 data = wiscsub,  
 na.action = na.exclude)  
summary(model3)

##   
## Call:  
## lm(formula = verb2 ~ 1 + verb1\_star, data = wiscsub, na.action = na.exclude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.5305 -3.0362 0.2526 2.7147 12.5020   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 25.41534 0.29831 85.20 <2e-16 \*\*\*  
## verb1\_star 0.75495 0.05149 14.66 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.261 on 202 degrees of freedom  
## Multiple R-squared: 0.5156, Adjusted R-squared: 0.5132   
## F-statistic: 215 on 1 and 202 DF, p-value: < 2.2e-16

**Note**: Mean centering should be used to aid interpretation. Historically, it has been suggested that mean centering will reduce multicollinearity, however this is not the case. See ![Olvera & Kroc (2018)](data:text/html; charset=UTF-8;base64,) for more information.

### 7.4.1 Interpreting Model Parameters

Note that the estimate for the slope stays the same, but the estimate for the intercept is different. This is because the variable ‘verb1\_star’ equals 0 when a child has an average 1st grade verbal score. Therefore the expected value for the 2nd grade verbal score, for *a child with an average 1st grade verbal score*, is 25.41534.

### 7.4.2 Plotting Regression Line

ggplot(data=wiscsub, aes(x=verb1\_star,y=verb2)) +  
 geom\_point(size = 2, shape=19) +  
 geom\_smooth(method=lm,se=TRUE,fullrange=TRUE,colour="red", size=2) +  
 labs(x= "Sample-Centered Verbal Ability Grade 1", y= "Verbal Ability Grade 2") +  
 xlim(-20,20) +  
 ylim(0,50) +  
 #theme with white background  
 theme\_bw() +  
 #eliminate background, gridlines, and chart border  
 theme(  
 plot.background = element\_blank()  
 ,panel.grid.major = element\_blank()  
 ,panel.grid.minor = element\_blank()  
 ,panel.border = element\_blank()  
 ) +  
 #draws x and y axis line  
 theme(axis.line = element\_line(color = 'black')) +  
 #set size of axis labels and titles  
 theme(axis.text = element\_text(size=12),  
 axis.title = element\_text(size=14))

## `geom\_smooth()` using formula = 'y ~ x'



Note the change of scale on the x-axis.

## 7.5 Multiple Linear Regression

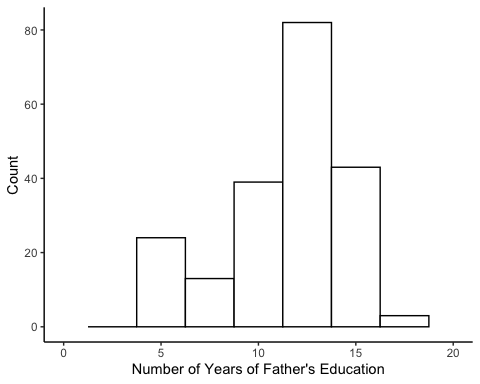
Now, let’s include a second predictor. We have information on the number of years of education for the children’s fathers, variable daded. The values in daded indicate the number of years of education each father completed. First, let’s take a look at the distribution of this new predictor variable.

psych::describe(wiscsub$daded)

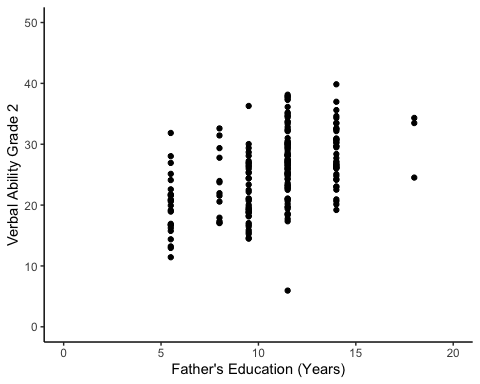
## vars n mean sd median trimmed mad min max range skew kurtosis se  
## X1 1 204 10.81 2.7 11.5 11 2.97 5.5 18 12.5 -0.36 0.01 0.19

ggplot(data=wiscsub, aes(x=daded)) +   
 geom\_histogram(binwidth=2.5, fill="white", color="black") +  
 xlim(0,20) +  
 xlab("Number of Years of Father's Education") +   
 ylab("Count") +  
 theme\_classic()

## Warning: Removed 2 rows containing missing values (`geom\_bar()`).

 And the relation between Grade 2 verbal scores and daded.

ggplot(data=wiscsub, aes(x=daded, y = verb2)) +   
 geom\_point() +  
 xlim(0,20) +   
 ylim(0,50) +  
 xlab("Father's Education (Years)") + ylab("Verbal Ability Grade 2") +  
 theme\_classic()



### 7.5.1 Regression Equation

Our model now becomes

where is the sample-centered version of , and is the sample-centered version of .

The slope, is the expected difference in grade 2 verbal score for each 1 year difference in father’s education.

We can also center the daded variable.

# Calculate mean-centered version of father's education variable  
wiscsub$daded\_star <- wiscsub$daded - mean(wiscsub$daded)

### 7.5.2 Fit Model in R

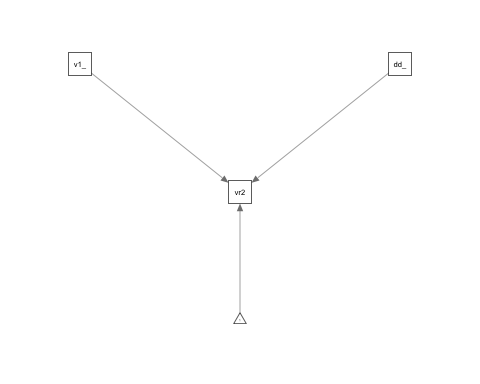
model4 <- lm(verb2 ~ 1 + verb1\_star + daded\_star,  
 data = wiscsub,  
 na.action = na.exclude)  
summary(model4)

##   
## Call:  
## lm(formula = verb2 ~ 1 + verb1\_star + daded\_star, data = wiscsub,   
## na.action = na.exclude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.4354 -2.9189 -0.1542 2.3746 11.1678   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 25.41534 0.29069 87.430 < 2e-16 \*\*\*  
## verb1\_star 0.66786 0.05626 11.872 < 2e-16 \*\*\*  
## daded\_star 0.41454 0.12108 3.424 0.000749 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.152 on 201 degrees of freedom  
## Multiple R-squared: 0.5422, Adjusted R-squared: 0.5377   
## F-statistic: 119.1 on 2 and 201 DF, p-value: < 2.2e-16

Now we have an intercept and two slopes.

### 7.5.3 Path Diagram

semPlot::semPaths(model4, what = "paths")



#### 7.5.3.1 Interpreting Model Parameters

is the expected value of the outcome variable when all other variables are 0. Therefore, in this case, is the expected Grade 2 verbal score for a child with an average Grade 1 verbal score (i.e.  = 0) *and* whose father had an average education (i.e.  = 0, = 10.81 years of education.

is the expected difference in the outcome for a 1-unit difference in . In this example (i.e. ‘model4’), is the expected difference in Grade 2 verbal score (outcome variable, = ) for a 1 point difference in the Grade 1 verbal score ( = ), holding constant the level of father’s education.

is the expected difference in the outcome for a 1-unit difference in . For this example (i.e. ‘model4’), is the expected difference in Grade 2 verbal score (outcome variable, = ) for each year difference in father’s education ( = ), holding constant in Grade 1 verbal score.

### 7.5.4 A Note on Interpretation

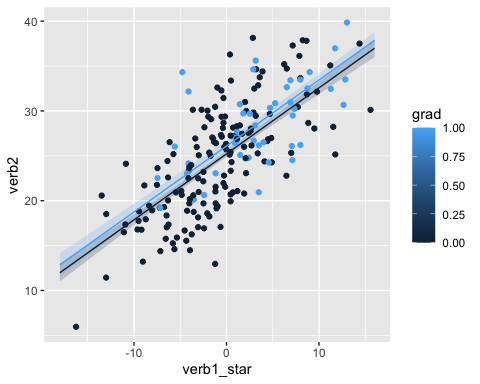
The distinguishing feature for interpretation in linear models without interactions or higher-order terms is that **the effect of a given change in an independent variable is the same regardless of the value of that variable (at the start of its change) and regardless of the level of the other variables in the model.**

Interpretation only needs to specify which variable is changing, by how much, and that other variables are being held constant.

In regard to our last example, years of father’s education does impact the relationship between Grade 1 and 2 verbal scores. Likewise, the effect of father’s education on Grade 1 verbal scores does not depend on Grade 1 verbal scores. Said in a different way, no matter what a child’s Grade 1 verbal score was, the effect of father’s education on Grade 2 verbal scores is the same.

Consider another model with categorical and continuous predictors, grad and verb1\_star, respectively.

library("ggiraphExtra")  
model4b <- lm(verb2 ~ 1 + verb1\_star + grad,  
 data = wiscsub,  
 na.action = na.exclude)  
ggPredict(model4b,se=TRUE,interactive=FALSE)



Notice in the plot, whether a student’s father graduated HS does not impact the relationship between Grade 1 and 2 verbal scores. Differences in Grade 1 and 2 verbal scores are not dependent on whether or not the father graduated HS.

**Most importantly, using the coefficients themselves we can easily interpret the model parameters.**

## 7.6 Categorical Variable Interaction

Ok, let’s move on to the topic of an *interaction* which uses the product of two predictor variables as a new predictor.

Working up a slightly different example with the ‘grad’ variable (whether dad graduated high school),

Where is the mean-centered version of , and is a dummy coded variable that equals 0 if the child’s father *did not* graduate high school, and equals 1 if the child’s father *did* graduate high school.

We did not sample-mean center in this example because a value of 0 already has substantive meaning for the current example (i.e. when equals 0, the father *did not* graduate high school).

### 7.6.1 Interaction as Moderation

Often, we describe phenomena in terms of *moderation*; or that the relation between two variables (i.e.  and ) is *moderated* by a third variable (i.e. ). For example, the relation between Grade 1 and Grade 2 verbal scores may be *moderated* by father’s graduation status. More specifically, the relation between 1st and 2nd grade verbal score may be different for children whose fathers’ did not or did graduate from high school.

The inclusion of product terms (i.e. interactions) allows for a direct investigation of a *moderation* hypothesis.

#### 7.6.1.1 Choosing a Moderator

When we use a product term, we should define one of the variables as the moderator and one of the variables as the predictor of interest. Let’s call the predictor of interest, and the moderator.

### 7.6.2 Moderation by Categorical Variable

When the moderator is a dummy variable then the form of the moderation becomes fairly simple; we will have one equation for , and a second equation for .

#### 7.6.2.1 Rewriting Equation

To illustrate the notion of two equations, let’s rewrite the regression equation

as two separate regression equations, one for fathers who graduated from highschool and one for fathers that did not. We can accomplish this by plugging in and into the regression equation and rearranging some of the terms. Doing so we get

**Equation for Students whose father Graduated Highschool**

**Equation for Students whose father Did Not Graduate from Highschool**

### 7.6.3 Interpretation

Without an interaction, our linear regression model assumes that the only difference between the regression line for each group (graduate HS vs not) is the intercept. That is, it assumes that the relationship between verbal scores at Grades 1 and 2 is the same for both groups.

*Children Whose father’s Did Not Graduate HS*

The expected Grade 2 verbal score for a child whose father did not graduate high school *and* who had an average Grade 1 verbal score is . Also, for a child whose father did not graduate high school, is the expected difference in their Grade 2 verbal score for a one-point difference in their Grade 1 verbal score.

*Children Whose father’s Did Not Graduate HS*

The parameter estimates and maintain their interpretation from before. But now each of them is *moderated* (i.e. shifted or altered) by or .

Specifically, the expected Grade 2 verbal score for a child whose father did graduate high school *and* who earned an average Grade 1 verbal score is .

And, for a child whose father did graduate high school, is the expected difference in their Grade 2 verbal score for a one-point change in their Grade 1 verbal score.

### 7.6.4 Fit Regression Model in R

OK - let’s fit the model! Note that within this model we use the code I(verb1\_star \* grad). This produces the interaction term within the model. The wrapper function I() indicates to R to perform this data computation as-is, otherwise we would need to perform this computation (i.e. the multiplication of verb1\_star by grad) outside of the function lm().

model5 <- lm(verb2 ~ 1 + verb1\_star + grad + I(verb1\_star\*grad),  
 data = wiscsub,  
 na.action = na.exclude)  
summary(model5)

##   
## Call:  
## lm(formula = verb2 ~ 1 + verb1\_star + grad + I(verb1\_star \* grad),   
## data = wiscsub, na.action = na.exclude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.3433 -3.0761 -0.0825 2.5689 10.7289   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 25.2663 0.3416 73.956 <2e-16 \*\*\*  
## verb1\_star 0.7861 0.0604 13.015 <2e-16 \*\*\*  
## grad 1.4632 0.8107 1.805 0.0726 .   
## I(verb1\_star \* grad) -0.2430 0.1324 -1.836 0.0678 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.232 on 200 degrees of freedom  
## Multiple R-squared: 0.5268, Adjusted R-squared: 0.5197   
## F-statistic: 74.22 on 3 and 200 DF, p-value: < 2.2e-16

### 7.6.5 Path Diagram

semPlot::semPaths(model5, what = "paths")



The parameter estimates from this model indicate that, for children whose father did not graduate high school, the expected Grade 2 verbal score for a child that earned an average 1st grade verbal score equals 25.2663 ().

Also, for children whose father did not graduate high school, a 1-point difference in their Grade 1 verbal score is expected to correspond with a 0.7861 () point difference in the Grade 2 verbal score.

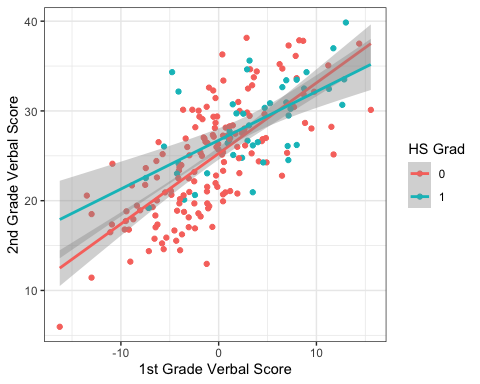
Moreover, the parameter estimates indicate that, for children whose father did graduate high school, the expected Grade 2 verbal score for a child that earned an average Grade 1 verbal score is 25.2663 + 1.4632 = 26.7295 ().

Also, for children whose father graduated high school, a 1-point difference in their Grade 1 verbal score is expected to correspond with a () = 0.7861 - 0.2430 = 0.5431 point difference in the Grade 2 verbal score.

Even though the interaction is not significant, we can plot it for illustrating the moderation effect:

#plot of moderation  
ggplot(data=wiscsub,   
 aes(y=verb2,x=verb1\_star, color = factor(grad))) +  
 geom\_jitter() +  
 stat\_smooth(method='lm', se=TRUE, fullrange=TRUE) +  
 xlab("1st Grade Verbal Score") +   
 ylab("2nd Grade Verbal Score") +   
 guides(color=guide\_legend(title="HS Grad")) +  
 theme\_bw()

## `geom\_smooth()` using formula = 'y ~ x'



The example from ‘model5’ contained an interaction using a dummy variable (i.e., ). Interactions may also occur between two continuous variables (i.e., and ). We will not cover here, but note that it is still very useful to consider and communicate those interactions as *moderation*. There are many resources on interactions of two (or more) continuous variables.

# 8 Logistic Regression

In Chapter 8 we will introduce the logistic regression model through the lens of generalized linear models (GLMs). The GLM is an essential tool for modeling discretely distributed dependent variables.

## 8.1 Categorical Data in the Social Sciences

Linear regression is a workhorse procedure of modern statistics. Our introduction to regression in this class was framed around the idea of a continuous dependent (outcome) variable. However, categorical data is extremely common in many health, behavioral and social science applications.

### 8.1.1 Examples of Categorical Data

* Binary Variables have two categories and are often used to indicate that an event has occurred or a characteristic is present. Are you sick? Did you vote in the last election? Are you married?
* Ordinal variables have categories that can be ranked. Surveys often ask respondents to indicate their agreement to a statement, how frequently then engage in a behavior, or even educational attainment.
* Nominal variables occur when there are multiple outcomes that cannot be ordered. For example, left or right handedness or occupation.
* Censored variables occur when the value of a variable is unknown over some range of the variable. For example, measuring hourly wages might be restricted on the lower end by minimum wage laws.
* Counts indicate the number of times that some event has occurred. How many drinks last week? How many people living in a house? How many years of education? *Censored and count variables are often lumped in with more traditional categorical variables under the umbrella of limited dependent variables.*

## 8.2 Introduction to GLMs

Earlier we noted linear regression is typically applied to continuous variables. The ubiquity of categorical data leads us to a modeling framework better suited to handling a wide range of categorical outcomes: the Generalized Linear Model (GLM).

In the GLM, the response variable is assumed to follow an exponential distribution with mean , which itself is a nonlinear function of . We can think about as the mean of a conditional response distribution at a given point in the covariate space.

There are three important components to the GLM:

1. *A random component*: The random component of the GLM contains the response variable and its probability distribution (e.g. the binomial distribution of in the binary regression model).
2. *A Linear Predictor*: The linear predictor typically takes the form of where is an matrix of observations and is an column vector.
3. *Link Function*: The link function, typically specified as , is used to relate each component of to the linear predictor, .

### 8.2.1 Linear Regression as GLM

Linear regression can be formulated in the GLM framework as follows:

1. *A random component*: We can make specify .
2. *A Linear Predictor*: are the continuous or discrete explanatory variables. The way we think about the structural component here doesn’t really differ from how we think about it with standard linear models; in fact, that’s one of the nice advantages of the GLM.
3. *Link Function*: For linear regression we use the *identity link* (e.g. ).

### 8.2.2 Logistic Regression as GLM

Let’s also take a look at binary logistic regression formulated as GLM.

1. *A random component*: The distribution of is assumed to be binomial with success probability .
2. *A Linear Predictor*: are the continuous or discrete explanatory variables.
3. *Link Function*: For logistic regression we use the *log-odds (or logit) link* (e.g. ), where is the transformed outcome.

### 8.2.3 Poisson Regression as GLM

Poisson regression can also be formulated as a GLM:

1. *A random component*: The distribution of is assumed to be Poisson with mean .
2. *A Linear Predictor*: are the continuous or discrete explanatory variables.
3. *Link Function*: For Poisson regression the *log link* is used.

### 8.2.4 Additional Remarks

When the outcome data are not normally distributed, we can always do transformation to change its scale. These are typically done via *link functions* denoted as - so we get . If we denote the transformed outcome as , then we can denote it as:

From a conceptual point of view, the link function *transforms* into a normal outcome. Note that we are simplifying notation somewhat: while we are modeling some expectation of , not exactly , we will keep on using . This is to say the link is applied to the parameter governing the response distribution, not the actual response data. We use *link functions* to formalize that the conditional expectation for (conditional because it is the expected value of depending on the level of the predictors and the chosen link).

Each link function also has an inverse, , which allows us to define

The inverse of a link function back-converts the linear combination of predictors into the original outcome.

## 8.3 Binary Logistic Regression

### 8.3.1 Overcoming LPM

To avoid the problems of the LPM we’d like a model where

is forced to be within the range of to . One way to do this is to transform the probability above into the odds metric,

which has a range of to . So, we are halfway there. Indeed, by taking the log of the odds (or logit) we extend the support of to have a range of to . This maps probability ranging between and to log odds ranging from negative infinity to positive infinity.

This is one example of why the logit link is used for logistic regression. Nowe we can seamlessly model the probability of an event occurring, giving the explanatory variables, .

We denote this probability as , or equivalently, .

Often times you will simply see for convenience, but it is important to remember this probability is conditional on the explanatory variables in the model.

### 8.3.2 Model

The binary logistic regression model is expressed as

Where is the odds of an event occurring and is the natural logarithm. Therefore, the parameter estimates from a generalized linear regression using the logistic link function are scaled in *log-odds* or *logit* units.

We can also rewrite the model above, solving for , as

This is also called the inverse function for the logit link function, or the *logistic* link, . In practice, this transformation is what is used for solving the regression equation, and it is called logistic regression:

## 8.4 Example Data

Chapter 7

In Chapter 7 we use data from Dunn, Aknin, and Norton (2007), who examined the relationship between spending habits and happiness using OLS regression. To capture spending habits self-reported monthly spending was categorized as being either personal or prosocial, and then summed to create a category-specific total. Measures of happiness were obtained using a 5-item ordinal scale.

The dependent variable in this analysis (GeneralHappiness) was self-reported general happiness. Participants rated their general happiness by answering the question “Do you feel happy, in general?” by selecting from five possible response options (no, rarely, sometimes, most of the time, and yes), which were then scored from 1-5. Here, higher numbers were indicative of greater happiness. For the purpose of our analysis we will dichotomize GeneralHappiness depending on whether a person was not happy vs happy.

The sample was selected to be a nationally representative sample of 632 Americans (287 males and 345 females). Participants responded as part of a larger, online survey, in return for points that could be redeemed for prizes (Dunn, 2008). No further details on sampling were available.

### 8.4.1 Variables

* **Happy**: Dichotomous variable indicating whether the subject responded “yes” when asked if they felt happy, in general.
* **PersonalSpending**: Self-reported dollars spent per month on (a) bills and expenses, and (b) gifts for themselves.
* **ProsocialSpending**: Self-reported dollars spent per month on (a) gifts for others, and (b) donations to charity.
* **PersonalIncome**: Participants selected their personal income category from 6 options: less than $20,000, $20,000-$35,000, $35001-$50,000, $50,001-$65000, $65,001-$80,000, $80,001+.

library("ggplot2")  
dunn2008 <- read.csv("data/DUNN2008.csv")  
dunn2008$Happy <- ifelse(dunn2008$GeneralHappiness == "yes", 1, 0)  
dunn2008$PersonalSpending <- dunn2008$PersonalSpending/100  
dunn2008$ProsocialSpending <- dunn2008$ProsocialSpending/100  
dunn2008$PersonalSpending\_star <- as.numeric(scale(dunn2008$PersonalSpending, scale = FALSE))  
dunn2008$ProsocialSpending\_star <- as.numeric(scale(dunn2008$ProsocialSpending, scale = FALSE))  
dunn2008$Income <- dplyr::recode(dunn2008$PersonalIncome,   
 "20001-35000" = "20-35K",   
 "35001-50000" = "35-50K",   
 "50001-65000" = "50-65K",   
 "65001-80000" = "65-80K",   
 "80001andup" = "> 80K",   
 "less20000" = "< 20K")  
inc\_lev\_order <- c("< 20K","20-35K", "35-50K","50-65K","65-80K","> 80K")  
dunn2008$Income <- factor(dunn2008$Income, levels=inc\_lev\_order)

## 8.5 Intercept-Only Model

In logistic regression, we are interested in how various predictors are related to the probability of a specific outcome . In this example we are interested in the probability an individual reports being happy, in a general sense.

Making use of the logit link function, the general equation for logistic regression is

Which after back transformation gives us …

### 8.5.1 Intercept-Only Model in R

In our example the variable Happy indicates whether a subject reporting being happy. Let’s start with the simplest model for predicting Happy, the intercept-only model. More specifically, we have $ logit(\_i) = b\_0(1\_i)$where .

We can use the glm() function to fit the model to the data

model9 <- glm(Happy ~ 1,   
 family = "binomial",   
 data = dunn2008,   
 na.action = na.exclude)  
summary(model9)

##   
## Call:  
## glm(formula = Happy ~ 1, family = "binomial", data = dunn2008,   
## na.action = na.exclude)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.69077 0.08435 -8.19 2.62e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 805.02 on 631 degrees of freedom  
## Residual deviance: 805.02 on 631 degrees of freedom  
## AIC: 807.02  
##   
## Number of Fisher Scoring iterations: 4

Without wanting to get to detailed we don’t need to specify the logit link here because it is the canonical link function for the binomial distribution. This essentially means there is a direct correspondence between the predicted mean and the distribution’s canonical location parameter.

### 8.5.2 Interpretation

#### 8.5.2.1 Intercept Parameter

In the intercept-only model, the intercept, , reflects

1. The expected log-odds () of an individual reporting they were happy.
2. The odds of someone reporting being happy .

exp(-0.69077)

## [1] 0.50119

1. The expected probability () of the a subject reported being happy in general.

or, equivalently, in R

exp(-0.69077)/(1 + exp(-0.69077))

## [1] 0.3338618

We can also confirm that the backward transformed parameter from this intercept-only logistic regression matches the expectation we get from the descriptives of the raw data.

mean(dunn2008$Happy)

## [1] 0.3338608

*Note*: If then , indicating a positive relationship between and the probability of the event occurring. If , the opposite relationship holds.

## 8.6 Single Predictor Model

OK, let’s include a predictor in our logistic regression model. Let’s start with PersonalSpending such that

where . Here, is the mean-centered amount of money one spends on themselves in a month (in units of dollars).

Let’s fit the model in R.

model10 <- glm(Happy ~ 1 + PersonalSpending\_star,   
 family = "binomial",   
 data = dunn2008,   
 na.action = na.exclude)  
summary(model10)

##   
## Call:  
## glm(formula = Happy ~ 1 + PersonalSpending\_star, family = "binomial",   
## data = dunn2008, na.action = na.exclude)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.690890 0.084358 -8.190 2.61e-16 \*\*\*  
## PersonalSpending\_star -0.001355 0.004605 -0.294 0.769   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 805.02 on 631 degrees of freedom  
## Residual deviance: 804.93 on 630 degrees of freedom  
## AIC: 808.93  
##   
## Number of Fisher Scoring iterations: 4

### 8.6.1 Overdispersion

A quick digression. In the binary logistic regression model *overdispersion* occurs when the observed variance is larger than what the binomial distribution would predict. For example, if , the mean is $\u\_{i}=n\_{i}\pi\_{i}$ and the variance is . Since both of these moments rely on , it can be overly restrictive, and if overdispersion is present inferences can become distorted. We will talk about this more later.

### 8.6.2 Coefficients

Again, There are essentially three ways to interpret coefficients from a logistic regression model:

1. The log-odds (or logit)
2. The Odds
3. Probabilities

#### 8.6.2.1 Log-Odds

The parameter estimate reflects the expected log-odds () of being happy for an individual with an average amount of personal spending.

The estimate for indicates the expected difference of the log-odds of being happy for a dollar difference in personal spending. Therefore, we expect a difference in the log-odds of being happy for a dollar difference in personal spending.

#### 8.6.2.2 Odds

Parameter estimates from a logistic regression are often reported in terms of *odds* rather than *log-odds*. To obtain parameters in odds units, we simply exponentiate the coefficients. Note that this is just one of the steps of the inverse link function (which would take us all the way to probability units).

exp(cbind(OR = coef(model10), confint(model10)))

## Waiting for profiling to be done...

## OR 2.5 % 97.5 %  
## (Intercept) 0.5011299 0.4240750 0.5903901  
## PersonalSpending\_star 0.9986459 0.9886908 1.0072747

In other words, the odds of being happy when personal spending is at average levels is .

In regard to the slope coefficient, for a dollar difference in monthly personal spending, we expect to see about decrease in the odds of being happy. This decrease does not depend on the value that personal spending is held at. Note this is not significant and we would not report this interpretation in practice. Essentially, if the odds ratio is equal to one, the predictor did not have an impact on the outcome.

#### 8.6.2.3 Probability

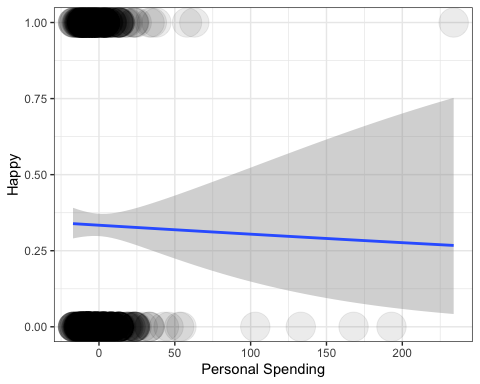
Remember, probabilities range from , whereas log-odds (the output from the raw logistic regression equation) can range from , and odds and odds ratios can range from . Due to the bounded range of probabilities, probabilities are non-linear, but log-odds can be linear.

For example, as personal spending goes up by constant increments, the probability of happiness will increase (decrease) by varying amounts, but the log-odds will increase (decrease) by a constant amount, and the odds will increase (decrease) by a constant multiplicative factor.

For this reason it is not so simple to interpret probabilities in logistic regression from the coefficient directly. Often it is much simpler to plot the probabilities across a range of the predictor variables.

ggplot(data=dunn2008,  
 aes(x=PersonalSpending\_star,y=Happy)) +  
 geom\_point(alpha = .08, size = 10) +  
 xlab("Personal Spending") +  
 ylab("Happy") +  
 theme\_bw() +  
 stat\_smooth(method = 'glm', method.args = list(family = "binomial"), se = TRUE)

## `geom\_smooth()` using formula = 'y ~ x'



Notice how the density of the observations is visualized by manipulating the transparency (alpha) level of the data points. The predicted curve based on our model has of course a non-linear shape (however, if we were to plot the relationship between the variables with using the logit link, it would be a straight line).

## 8.7 Marginal Effects

So far we have considered two possibilities for interpreting logistic regression results:

* Interpreting the log-odds directly
* Transforming the log-odds into odds
* A probability metric (for a single explanatory variable)

However, as we include more covariates in our model, interpretation becomes more difficult. We can only think about “holding other variable constant” in the log-odds and odds scale. For nonlinear model **marginal effects** provide us with an intuitive and easy to interpret method for understanding and communicating results.

### 8.7.1 A Definition of Marginal Effects

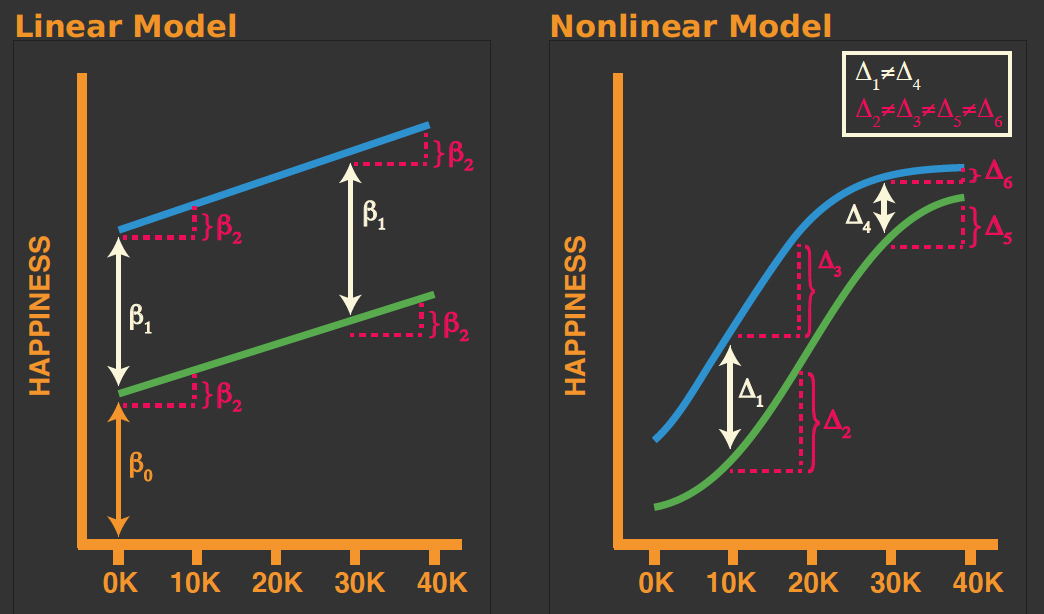
**Marginal effects are partial derivatives of the regression equation with respect to each variable in the model for each unit in the data.**

Put differently, the marginal effect measures the association between a change in an explanatory variable and a change in the response. The marginal effect is the slope of the prediction function, measured at a specific value of the explanatory variable.

In linear models the effect of a given change in an independent variable is the same regardless of (1) the value of that variable at the start of its change, and (2) the level of the other variables in the model.

In nonlinear models the effect of a given change in an independent variable (1) depends on the values of other variables in the model, and (2) is no longer equal to the parameter itself.

Consider a linear and nonlinear model for happiness as a function of personal spending and a dummy variable indicating whether someone is rich.



### 8.7.2 A Few Observations

**For the linear model**: - Whether one is rich or poor does no impact the relationship between happiness and personal spending. - Differences in happiness levels between rich and poor are not dependent on the amount of money one spends.

**From the nonlinear model**: - Whether one is rich or poor does impact the relationship between happiness and personal spending. - Differences in happiness levels between rich and poor are dependent on the amount of money one spends.

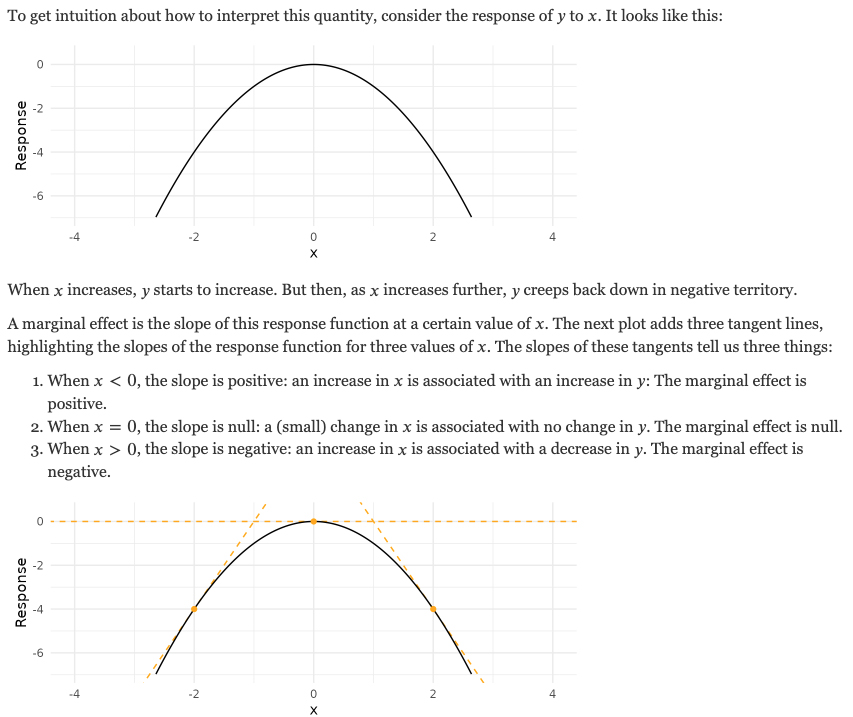
#### 8.7.2.1 Another Nonlinear Example

A helpful example is provided in the [marginaleffects vignette](https://vincentarelbundock.github.io/marginaleffects/articles/mfx.html).

Consider a simple quadratic

$$
y = -x^2 \\
$$

with partial derivative of with respect to



<https://vincentarelbundock.github.io/marginaleffects/articles/mfx.html>

### 8.7.3 Types of Marginal Effects

There are generally three types of *marginal effects* people consider:

* Marginal Effects at the Means (MEM)
* Average Marginal Effects (AME)
* Marginal Effects at Representative Values (MEM)

We will focus on marginal effects at representative values as this is the most powerful option.

### 8.7.4 Example Model

Let’s fit a more complicated model. To look at marginal effects we will use the marginaleffects package.

library("marginaleffects")  
model11 <- glm(Happy ~ 1 + PersonalSpending + ProsocialSpending + Income,   
 family = "binomial",   
 data = dunn2008,   
 na.action = na.exclude)  
summary(model11)

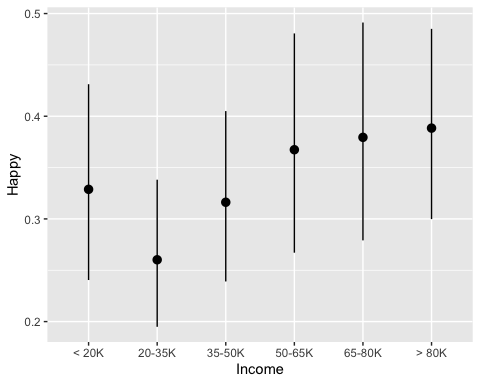
##   
## Call:  
## glm(formula = Happy ~ 1 + PersonalSpending + ProsocialSpending +   
## Income, family = "binomial", data = dunn2008, na.action = na.exclude)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.636366 0.223151 -2.852 0.00435 \*\*  
## PersonalSpending -0.011635 0.006643 -1.752 0.07985 .   
## ProsocialSpending 0.083741 0.032975 2.540 0.01110 \*   
## Income20-35K -0.331327 0.288006 -1.150 0.24997   
## Income35-50K -0.057632 0.297299 -0.194 0.84629   
## Income50-65K 0.170117 0.326317 0.521 0.60214   
## Income65-80K 0.221637 0.323873 0.684 0.49376   
## Income> 80K 0.259641 0.307140 0.845 0.39791   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 805.02 on 631 degrees of freedom  
## Residual deviance: 791.06 on 624 degrees of freedom  
## AIC: 807.06  
##   
## Number of Fisher Scoring iterations: 4

### 8.7.5 Marginal Effects at Representative Values (MER)

For example, let’s look at the impact of Income on the probability of being happy.

marginaleffects::plot\_cap(model11, condition = c("Income"), conf.int = TRUE)

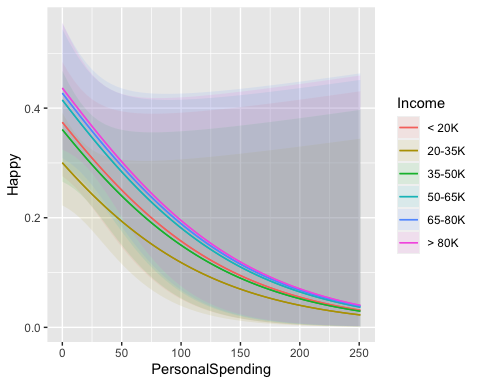
## Warning: This function has been renamed to `plot\_predictions()`. The `plot\_cap()`  
## alias will be removed in the near future.



What if we were interested in the relationship between Income and PersonalSpending on the probability of being happy.

marginaleffects::plot\_cap(model11, condition = c("PersonalSpending","Income"))

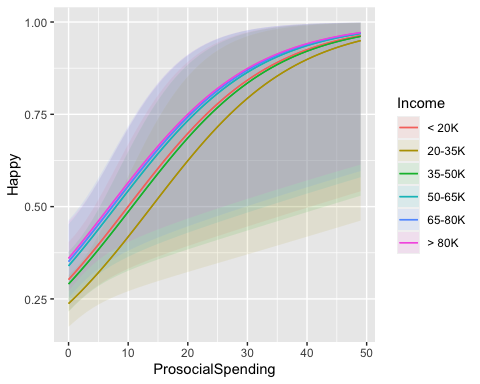
## Warning: This function has been renamed to `plot\_predictions()`. The `plot\_cap()`  
## alias will be removed in the near future.



What if we were interested in the relationship between Income and ProsocialSpending on the probability of being happy.

marginaleffects::plot\_cap(model11, condition = c("ProsocialSpending","Income"))

## Warning: This function has been renamed to `plot\_predictions()`. The `plot\_cap()`  
## alias will be removed in the near future.

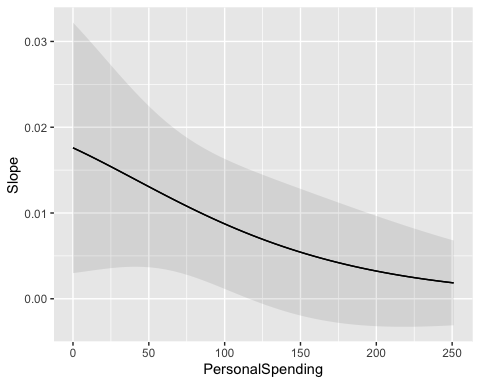


In nonlinear, the marginal effect of one variable is conditional on the value of the other variable. This function draws a plot of the marginal effect of the effect variable for different values of the condition variable.

Let’s look at the effect of PersonalSpending on the relationship between ProsocialSpending and Happy.

marginaleffects::plot\_cme(model11, variables = "ProsocialSpending", condition = "PersonalSpending")

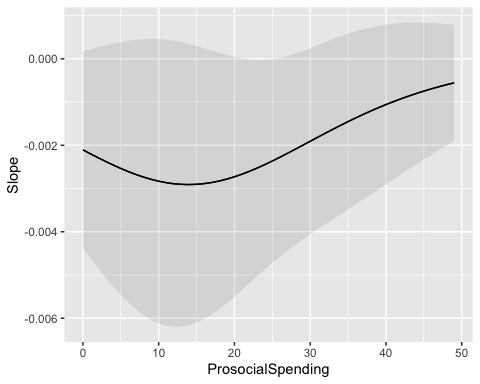
## Warning: This function has been renamed to `plot\_slopes()`. The `plot\_cme()`  
## alias will be removed in the near future.



In addition, we can look at the effect of PersonalSpending on the relationship between PersonalSpending and Happy.

marginaleffects::plot\_cme(model11, variables = "PersonalSpending", condition = "ProsocialSpending")

## Warning: This function has been renamed to `plot\_slopes()`. The `plot\_cme()`  
## alias will be removed in the near future.



# 9 Poisson Regression

In Chapter 9 we will round out our discussion of the GLM with **Poisson regression**.

* **Poisson regression can be a useful modeling approach for handling count dependent variables**
  + Counts typically describe nonnegative (or only positive) integer.
  + Examples of counts are number of drinks per day, children per household, etc.
* **In certain contexts, the Poisson distribution describes the number of events that occur in a given time period**
  + typically represents the mean number of events per period
  + In the Poisson distribution, the mean is also equal to the variance.
* **One important consideration when fitting Poisson regression models is overdispersion**
  + We will look at how one might assess overdispersion in Poisson regression and suggest some alternative procedures. A

## 9.1 Poisson Regression

### 9.1.1 Review of GLM

To review, there are three important components to the GLM:

1. *A random component*: The random component of the GLM contains the response variable and its probability distribution (e.g. the binomial distribution of in the binary regression model).
2. *A Linear Predictor*: The linear predictor typically takes the form of where is an matrix of observations and is an column vector.
3. *Link Function*: The link function, typically specified as , is used to relate each component of to the linear predictor, .

### 9.1.2 Poisson Regression as GLM

Poisson regression can also be formulated as a GLM:

or equivalently,

1. *A random component*: The distribution of is assumed to be Poisson, .
2. *A Linear Predictor*: The systematic component takes the form of .
3. *Link Function*: For Poisson regression the *log link* is used.

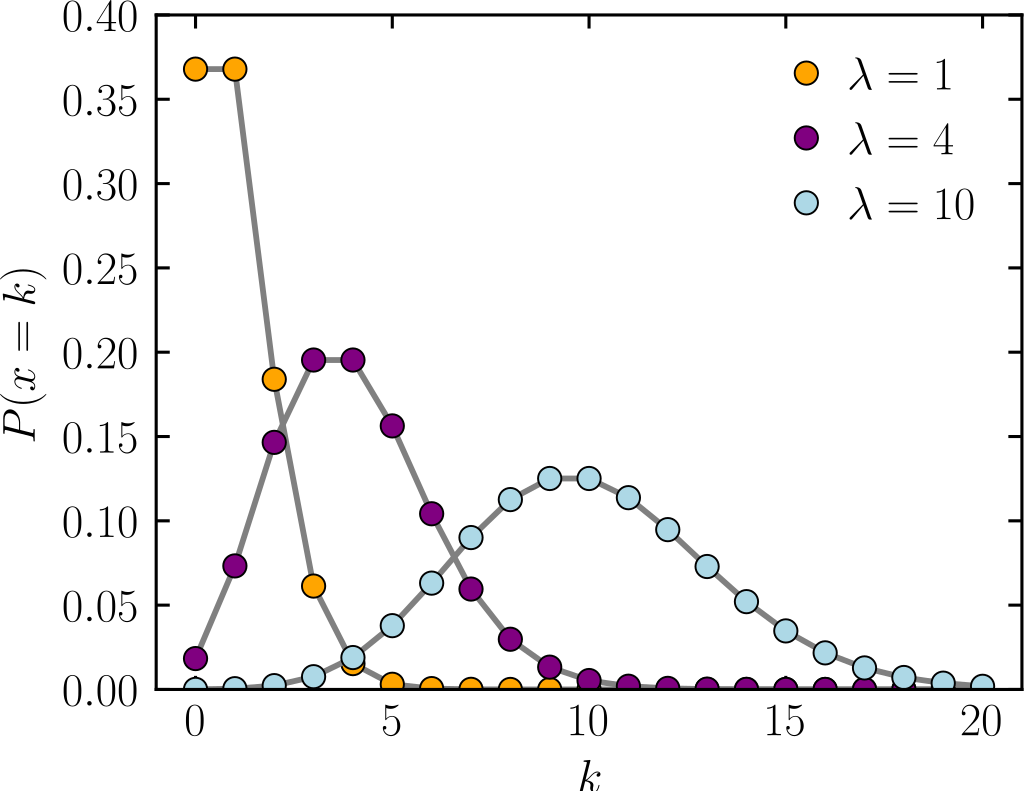
## 9.2 Poisson Distribution

To gain some intuition about the Poisson regression model consider the Poisson distribution

where

* is a random count variable
* is the expected number of times an event ocurrs
* is the factorical operator

The Poisson distribution relies on a single parameter, . Importantly, represents both the mean and the variance of the Poisson distribution (e.g. when is large both the mean and variance are large).



<https://en.wikipedia.org/wiki/Poisson_distribution>

**What Does This Mean In Practice**:

* As grows the center of the distribution shifts to right
* Departure of real count data from predictions from Poisson distribution
  + Variance frequently greater than mean (overdisperion)
  + Frequency of 0 counts exceed number predicted by Poisson

## 9.3 Notes on Interpretation

### 9.3.1 One Predictor Model

Consider a one-predictor Poisson regression,

$$
\mathrm{log}(\mu\_{i}) = \beta\_{0} + \beta\_{1}x\_{1i}\\
$$

where

### 9.3.2 Similarity to Logistic Regression

Interpretation of the Poisson regression coefficients is similar to logistic regression. For example,

* is the effect on the mean of when
* is the multiplicative effect on the mean of for each 1-unit difference in

### 9.3.3 Percent Change

We can also talk about these regression coefficients in terms of percent change as follows,

* **If is negative**:
  + All else being equal, we might expect to see a percent decrease in the expected count of , with each additional unit increase in , holding constant all other variables in the model.
* **If is positive**:
  + All else being equal, we might expect to see a percent increase in the expected count of , with each additional unit increase in , holding constant all other variables in the model.

The following relationships are helpful to keep in mind,

* If then
  + the expected count,
  + and are unrelated.
* If then
  + the expected count, , is times larger then when .
* If then
  + the expected count, , is times smaller then when .

**Note** that the parameter estimates in their original metric will describe the outcome variable in terms of *log* units. If we prefer to describe the phenomena in terms of the original *count* units we will need to use the inverse link function.

## 9.4 Example Data

Increasingly researchers are taking a life-course perspective to understanding how different life stages shape a variety of later in life outcomes.

In this case study, Ferraro, Schafer, and Wilkinson ([2016](#ref-ferraro2016)) examine the relationship between physical health in adulthood and multiple domains of childhood disadvantage using a count regression model.

Data are drawn from the National Survey of Midlife Development in the United States (MIDUS). MIDUS contains a battery of retrospective questions concerning childhood disadvantage, as well as extensive measures of adult risks and resources. Although the authors use data from both waves of MIDUS in the paper, here we focus only on their first model of adult health outcomes, which takes into account both childhood disadvantage and the mediating effects of later life resources and risk behaviors.

### 9.4.1 Dependent variable

The dependent variable for this analysis is health problems at Wave 1: morbidityw1.

Here, adult health problems are measured by the self-reported occurrence of 31 diseases or health conditions. For 29 of these items respondents were asked

“In the past 12 month have you experienced or been treated for any of the following?”

For the remaining 2, cancer and heart disease, respondents were asked if they had ever been diagnosed with the disease. Finally, morbidityw1 is the sum of these 31 items, where each is coded 1 for yes, and 0 for no.

### 9.4.2 Explanatory Variables

#### 9.4.2.1 Early Life Disadvantage

* *ses*: Childhood SES is a sum score based on standardized measures of (1) the education for the head of household, (2) financial strain and (3) receipt of welfare.
* *family*: Family composition is a sum score based on (1) the presence of a male in the household, (2) parental divorce, and (3) death of parent prior to age 16,
* *abuse\_rare*: Physical or Emotional Child abuse by parents is categorized by frequency of abuse. abuse\_rare indicates respondent rarely experience one or both types of abuse. The reference category is never having experienced emotional or physical abuse.
* *abuse\_frequency1*: abuse\_frequency1 indicates respondents frequently (sometimes or often) experienced one type of abuse during childhood. The reference category is never having experienced emotional or physical abuse.
* *abuse\_frequency2*: abuse\_frequency2 indicates respondents frequently experienced both types of abuse during childhood. The reference category is never having experienced emotional or physical abuse.
* *health*: Adolescent health problems are measured by self-rated physical and mental health at age 16.

#### 9.4.2.2 Adult Characteristics

* *age*: Age at time of Wave 1 interview.
* *nonwhite*: Race (white or nonwhite).
* *female*: Gender (female or male).
* *educate*: Number of years of completed education.
* *catincome*: Household income adjusted by household size and recoded into five percentile categories (< 21st \*percentile, 21st to 40th percentile41st to 60th percentile, 61st to 80th percentile, and > 80th percentile.).
* *a1sj6*: Financial strain during adulthood; responses range from 1 (no difficulty paying monthly bills) to 3 (very difficult to pay monthly bills).
* *smoke\_dose*: Lifetime smoking is calculated from information reported by respondents: age when started smoking, year stopped (for former smokers), and average number of cigarettes smoked daily. Using a yearly metric, lifetime smoking is the product of years smoked and annual number of cigarettes, divided by 10,000 (see Footnote 11, p. 130).
* *heavydr2*: The measurement of heavy drinking is sex differentiated and tapped respondents’ period of greatest lifetime consumption: five or more drinks per day for men and four or more drinks for women.
* *obese*: Obesity, dummy variable coded 1 if body mass index [kg/m2] > 30.
* *fampos*: Family support as measured by four items reflecting the presence of positive relationship characteristics.
* *friendpos*: Friend support as measured by four items reflecting the presence of positive relationship characteristics.
* *famneg*: Family strain as measured by four items reflecting the presence of difficult relationship characteristics.
* *friendneg*: Friend strain as measured by four items reflecting the presence of difficult relationship characteristics.
* *integration*: Social integration as measured by three 7-item Likert–type questions.
* *ever\_divor*: Ever divorced, a dummy variable coded 1 if the respondent reported having been divorced.
* *controlw1*: Average score for a 12-item index of the respondent’s feelings of personal control.

## 9.5 Single Predictor Model

### 9.5.1 Read in Data

We can read in the data and create a mean-centered version of income.

ferraro2016 <- read.csv("data/ferraro2016.csv")  
ferraro2016$income\_star <- as.numeric(scale(ferraro2016$catincome, scale = FALSE))

### 9.5.2 Single Predictor Model in GLM

Let’s fit a single predictor Poisson regression model for morbidity at Wave 1 using income bracket as our predictor. For now let’s treat income bracket as a continuous predictor.

Note for the Poisson regression we use family = poisson(link=log) argument.

model1 <- glm(  
 formula = morbidityw1 ~ 1 + income\_star,   
 family = poisson(link=log),   
 data = ferraro2016,  
 na.action = na.exclude  
)

### 9.5.3 Deviance and Goodness of Fit

summary(model1)

##   
## Call:  
## glm(formula = morbidityw1 ~ 1 + income\_star, family = poisson(link = log),   
## data = ferraro2016, na.action = na.exclude)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.007472 0.011047 91.195 < 2e-16 \*\*\*  
## income\_star -0.032445 0.007838 -4.139 3.48e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 8166.6 on 2994 degrees of freedom  
## Residual deviance: 8149.5 on 2993 degrees of freedom  
## (27 observations deleted due to missingness)  
## AIC: 14986  
##   
## Number of Fisher Scoring iterations: 5

We can think about the deviance as a measure of how well the model fits the data.

* If the model fits well, the observed values will be close to their predicted means , causing the deviance to be small.
* If this value greatly exceeds one, it may be indicative of overdispersion.

The rationale for this heuristic is based on the fact that the residual deviance is distributed with mean equal to the degrees of freedom (n-p). Instead of using this rule of thumb it is just as simple to formulate a goodness-of-fit test for our model as follows

1 - pchisq(summary(model1)$deviance, summary(model1)$df.residual )

## [1] 0

The test of the model’s deviance is a test of the model against the null model. The null hypothesis is the fitted model fits the data as well as the saturated model. In this case, rejecting the null hypothesis (e.g. ), indicates the Poisson model does not fit well. For more information see this wonderful post on stackexchange: <https://stats.stackexchange.com/a/531484>.

### 9.5.4 Overdispersion?

The deviance statistic suggests there may be a problem with overdispersion?

*Overdispersion indicates there is greater variability in the data than would be expected based on the model.*

Overdispersion is often encountered when fitting simple Poisson regression models. The Poisson distribution has one free parameter and does not allow for the variance to be adjusted independently of the mean.

If overdispersion is present the resultant model may yield biased parameter estimates and underestimated standard errors, possibly leading to invalid conclusions.

### 9.5.5 Interpretation of Single Predictor Model

**Intercept in Log Units**

This is the Poisson regression estimate when all variables in the model are evaluated at zero. For our model, we have centered the income variable. This means for an individual with an average income level the log of the expected count for health problems is units.

**Intercept in Count Units**

We can also exponentiate the intercept, indicating that at Wave 1 follow-up, an individual with an average income level is expected to have approximately health problems.

#### 9.5.5.1 Slope Coefficient

**Slope Coefficient in Log Units**

Within our single predictor model, is the difference in *log* number of health problems for a 1-level difference in income bracket.

Therefore, we expect a difference in the log-count of health problems for a 1-level difference in income brackets. Or, all else being equal, we might expect to see a percent decrease in the number of health problems with each additional unit-change in income level.

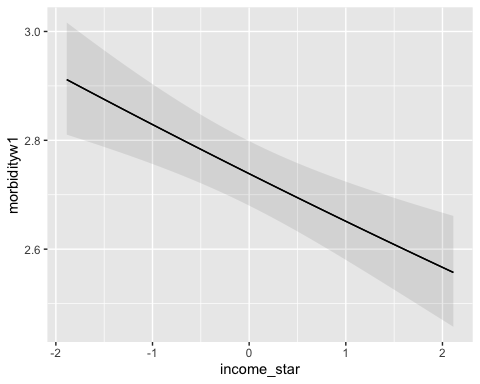
This relation may be difficult to conceptualize because the outcome variable is in terms of logarithm units, so a plot may be a more intuitive display of the results.

**Marginal Effects**

Let’s turn to the marginaleffects ([Arel-Bundock 2022](#ref-bundock)) package to look at the marginal effects of income on health.

marginaleffects::plot\_cap(model1, condition = c("income\_star"), conf.int = TRUE)

## Warning: This function has been renamed to `plot\_predictions()`. The `plot\_cap()`  
## alias will be removed in the near future.



## 9.6 Multiple Predictor Model

Let’s add another variable into the model. Specifically, the variable abuse\_rare, which equals if the child was rarely abused during early development, and if the child experienced abuse.

ferraro2016$abuse\_rare <- factor(ferraro2016$abuse\_rare)  
model2 <- glm(  
 formula = morbidityw1 ~ 1 + abuse\_rare + income\_star + abuse\_rare:income\_star,   
 family = poisson(link=log),   
 data = ferraro2016,  
 na.action = na.exclude  
)  
summary(model2)

##   
## Call:  
## glm(formula = morbidityw1 ~ 1 + abuse\_rare + income\_star + abuse\_rare:income\_star,   
## family = poisson(link = log), data = ferraro2016, na.action = na.exclude)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.0766511 0.0129647 83.045 < 2e-16 \*\*\*  
## abuse\_rare1 -0.2382614 0.0251265 -9.482 < 2e-16 \*\*\*  
## income\_star -0.0311089 0.0091684 -3.393 0.000691 \*\*\*  
## abuse\_rare1:income\_star -0.0000822 0.0179350 -0.005 0.996343   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 8068.4 on 2966 degrees of freedom  
## Residual deviance: 7956.7 on 2963 degrees of freedom  
## (55 observations deleted due to missingness)  
## AIC: 14737  
##   
## Number of Fisher Scoring iterations: 5

As the model becomes more complicated it can be helpful to write out the equation,

Again, when we have a dummy variable interaction we have essentially set up two different equations:

**Equation for Children Experiencing Maltreatment**

Coefficients and describe the relation between and for those who experienced childhood abuse .

**Equation for Children Rarely Experiencing Maltreatment**

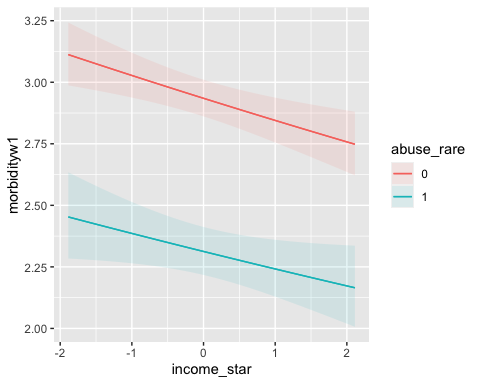
While coefficients and describe the relation for those who rarely experienced childhood abuse . Note that is not significantly different from zero, so we would not interpret the interaction between income and abuse directly.

**Marginal Effects**

Let’s again turn to the marginaleffects ([Arel-Bundock 2022](#ref-bundock)) package to look at the marginal effects of income and childhood abuse on health.

marginaleffects::plot\_cap(model2, condition = c("income\_star","abuse\_rare"), conf.int = TRUE)

## Warning: This function has been renamed to `plot\_predictions()`. The `plot\_cap()`  
## alias will be removed in the near future.



## 9.7 Revisisting Overdispersion

Let’s run our model fit test based on the deviance for model2.

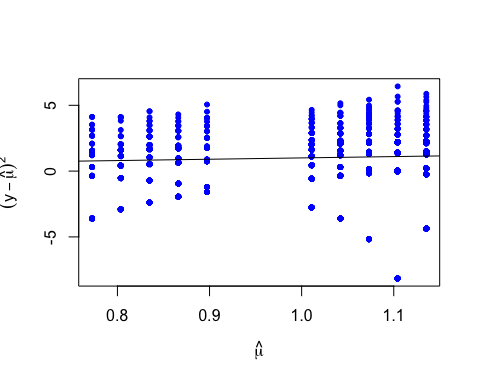
1 - pchisq(summary(model2)$deviance, summary(model2)$df.residual )

## [1] 0

Again, we reject the hypothesis of a close fit between model and data. To gain a little more insight we can plot estimates of the variance against the expected value, alongside a line with an intercept of zero and a slope of 1.

We expect the data points to fall somewhat evenly along that line. Here, it appears our variance is consistently larger than our mean, indicating the possibility of overdispersion.

plot(  
 log(fitted(model2)),  
 log((ferraro2016$morbidityw1-fitted(model2))^2),  
 xlab=expression(hat(mu)),  
 ylab=expression((y-hat(mu))^2),  
 pch=20,col="blue"  
)  
abline(0,1) ## 'varianc = mean' line



Two ways we might handle this overdispersion are **(1)** estimate the overdispersion parameter directly within the model, or **(2)** use a negative binomial model.

### 9.7.1 Quassi-Poisson Family

If we want to test and adjust for overdispersion we can add a scale parameter with the family=quasipoisson option. The estimated scale parameter will be labeled as Overdispersion parameter in the output.

model3 <- glm(  
 formula = morbidityw1 ~ 1 + abuse\_rare + income\_star + abuse\_rare:income\_star,   
 family = quasipoisson(link=log),   
 data = ferraro2016,  
 na.action = na.exclude  
)  
summary(model3)

##   
## Call:  
## glm(formula = morbidityw1 ~ 1 + abuse\_rare + income\_star + abuse\_rare:income\_star,   
## family = quasipoisson(link = log), data = ferraro2016, na.action = na.exclude)  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.0766511 0.0217425 49.518 < 2e-16 \*\*\*  
## abuse\_rare1 -0.2382614 0.0421387 -5.654 1.71e-08 \*\*\*  
## income\_star -0.0311089 0.0153759 -2.023 0.0431 \*   
## abuse\_rare1:income\_star -0.0000822 0.0300781 -0.003 0.9978   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for quasipoisson family taken to be 2.812535)  
##   
## Null deviance: 8068.4 on 2966 degrees of freedom  
## Residual deviance: 7956.7 on 2963 degrees of freedom  
## (55 observations deleted due to missingness)  
## AIC: NA  
##   
## Number of Fisher Scoring iterations: 5

The new standard errors (in comparison to the model without the overdispersion parameter), are larger, Thus, the Wald statistics will be smaller and less likely to be significant.

### 9.7.2 Negative Binomial Regression

We can also fit a negative binomial model, in which allows dispersion or variance in the outcome. The negative binomial distribution has one parameter more than the Poisson regression. This parameters adjusts the variance independently from the mean. In fact, the Poisson distribution is a special case of the negative binomial distribution.

Importantly, the Poisson and Negative Binomial have the same mean structure, so we can interpret coefficients in the same way. However, we will need to use the MASS package ([Venables and Ripley 2002](#ref-venables)) to fit the Negative Binomial regression.

model4 <- MASS::glm.nb(  
 formula = morbidityw1 ~ 1 + abuse\_rare + income\_star + abuse\_rare:income\_star,   
 data = ferraro2016,  
 na.action = na.exclude  
)  
summary(model4)

##   
## Call:  
## MASS::glm.nb(formula = morbidityw1 ~ 1 + abuse\_rare + income\_star +   
## abuse\_rare:income\_star, data = ferraro2016, na.action = na.exclude,   
## init.theta = 1.529056769, link = log)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.0766474 0.0221378 48.634 < 2e-16 \*\*\*  
## abuse\_rare1 -0.2382366 0.0406825 -5.856 4.74e-09 \*\*\*  
## income\_star -0.0311705 0.0156452 -1.992 0.0463 \*   
## abuse\_rare1:income\_star -0.0004478 0.0289957 -0.015 0.9877   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for Negative Binomial(1.5291) family taken to be 1)  
##   
## Null deviance: 3375.4 on 2966 degrees of freedom  
## Residual deviance: 3334.7 on 2963 degrees of freedom  
## (55 observations deleted due to missingness)  
## AIC: 12770  
##   
## Number of Fisher Scoring iterations: 1  
##   
##   
## Theta: 1.5291   
## Std. Err.: 0.0682   
##   
## 2 x log-likelihood: -12760.4610

We notice the smaller dispersion parameter, approximately , providing some support the negative binomial is a better fit for our data.

We can also test the hypothesis of overdispersion formally using a likelihood ratio test. The difference between the two models is captured by estimating a dispersion parameter that is held constant in a Poisson model. Thus, the Poisson model is actually nested in the negative binomial model. We can then use a likelihood ratio test to compare the two and test this model assumption.

pchisq(2 \* (logLik(model4) - logLik(model2)), df = 1, lower.tail = FALSE)

## 'log Lik.' 0 (df=5)

In this example the associated chi-squared value estimated from is with one degree of freedom. This strongly suggests the negative binomial model, estimating the dispersion parameter, is more appropriate than the Poisson model.

### 9.7.3 References

Arel-Bundock, Vincent. 2022. *Marginaleffects: Marginal Effects, Marginal Means, Predictions, and Contrasts*. <https://CRAN.R-project.org/package=marginaleffects>.

Ferraro, Kenneth F., Markus H. Schafer, and Lindsay R. Wilkinson. 2016. “Childhood Disadvantage and Health Problems in Middle and Later Life: Early Imprints on Physical Health?” *American Sociological Review* 81 (1): 107–33. <https://doi.org/10.1177/0003122415619617>.

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>.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s*. Fourth. New York: Springer. <https://www.stats.ox.ac.uk/pub/MASS4/>.

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>.