Table of contents

# Simple Regression

## Overview

This example will examine whether average change in CBCL externalizing scores from Basline (T0) to the 1-Year follow-up (T1) differs significantly between boys and girls taking part in the ABCD Study. This analysis is conducted in two primary steps: 1) computing a difference score (DiffScore = CBCL EXT T2 - CBCL EXT T1); 2) conducting a simple regression on the difference score. Our primary aim is to determine whether group membership (boys, girls) predict the average difference value in CBCL externalizing scores from T1 to T2.

## Preliminary Setup

### Install Packages

**This code installs the r packages necessary for this example, if they are not already installed**

### Load Packages

**This code loads the r libraries necessary for this example**

#rm(list = ls())  
  
#Load packages  
library(lme4)  
library(lmerTest)  
library(tidyverse)  
library(arrow)  
library(arsenal)  
library(kableExtra)  
library(rstatix)  
library(ggpubr)

### Config Options

**This code configures knitr code chunk options**

knitr::opts\_chunk$set(echo = T, message=F, warning=F, error=F,   
 comment=NA, cache=T, code\_folding=T,  
 R.options=list(width=220), fig.align='center',   
 out.width='75%', fig.asp=.75)

## Descriptives Overview

### Read and View Data

**This code reads in and shows the data to be used in the current example**

## Read data  
df\_long<- read\_csv("/Users/shawes/Desktop/data/df\_long.csv")  
df\_wide <- read\_csv("/Users/shawes/Desktop/data/df\_wide.csv")  
str(df\_wide)

spc\_tbl\_ [99 × 12] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
 $ ids : num [1:99] 1 2 3 4 5 6 7 8 9 10 ...  
 $ site\_id : chr [1:99] "site06" "site06" "site06" "site10" ...  
 $ fam\_id : num [1:99] 8781 8781 8781 10210 10210 ...  
 $ sex : chr [1:99] "Female" "Female" "Female" "Female" ...  
 $ age : num [1:99] 9 10 9 9 9 10 10 10 10 9 ...  
 $ mature\_vg : num [1:99] 0 1 0 1 0 0 0 1 1 0 ...  
 $ vg\_total\_Year\_1 : num [1:99] 7 7 7 7 4 0 4 1 1 2 ...  
 $ vg\_total\_Baseline : num [1:99] 4 7 4 2 42 0 4 2 2 2 ...  
 $ vg\_total\_Year\_2 : num [1:99] 1 1 2 0 3 0 2 2 7 3 ...  
 $ cbcl\_extern\_Year\_1 : num [1:99] 7 7 7 7 2 0 4 1 1 2 ...  
 $ cbcl\_extern\_Baseline: num [1:99] 0 1 0 1 0 0 0 1 1 0 ...  
 $ cbcl\_extern\_Year\_2 : num [1:99] 7 7 7 7 4 6 4 4 2 2 ...  
 - attr(\*, "spec")=  
 .. cols(  
 .. ids = col\_double(),  
 .. site\_id = col\_character(),  
 .. fam\_id = col\_double(),  
 .. sex = col\_character(),  
 .. age = col\_double(),  
 .. mature\_vg = col\_double(),  
 .. vg\_total\_Year\_1 = col\_double(),  
 .. vg\_total\_Baseline = col\_double(),  
 .. vg\_total\_Year\_2 = col\_double(),  
 .. cbcl\_extern\_Year\_1 = col\_double(),  
 .. cbcl\_extern\_Baseline = col\_double(),  
 .. cbcl\_extern\_Year\_2 = col\_double()  
 .. )  
 - attr(\*, "problems")=<externalptr>

# set types  
df\_long$ids <- as.factor(df\_long$ids)  
df\_long$event <- as.factor(df\_long$event)  
df\_long$site\_id <- as.factor(df\_long$site\_id)  
df\_long$fam\_id <- as.factor(df\_long$fam\_id)  
df\_long$sex <- as.factor(df\_long$sex)  
df\_long$age <- as.numeric(df\_long$age)  
df\_long$mature\_vg <- as.numeric(df\_long$mature\_vg)  
df\_long$cbcl\_extern <- as.numeric(df\_long$cbcl\_extern)  
df\_long$vg\_total <- as.numeric(df\_long$vg\_total)

### Continuous Outcomes

**This code creates a table of descriptive information for continuous outcomes**

## Create a descriptives table of study variables by measurement occasion   
tab\_descriptives\_1 <- tableby.control(test=FALSE, total=FALSE,  
numeric.test="kwt", cat.test="chisq",  
numeric.stats=c("N", "meansd", "median", "range"   
), #"Nmiss2"  
cat.stats=c("countpct"), #"Nmiss2"  
stats.labels=list(N='Count', meansd="Mean (SD)", median  
='Median', range='Min - Max'  
)) # , Nmiss2 ='Missing'  
  
my\_cont\_labels <- list(  
 age = "Age",  
 vg\_total = "Weekly # of Video Gaming Hrs",  
 cbcl\_extern = "CBCL Externalizing Scale"  
)  
  
tab\_descriptives\_1 <- tableby(event ~ age + vg\_total +   
 cbcl\_extern,  
 data=df\_long, control=tab\_descriptives\_1)  
   
#summary(tab\_descriptives\_1, labelTranslations = my\_cont\_labels , text=TRUE, title = #"Continuous Outcomes", term.name = TRUE)  
  
# Push table object through kable and kable\_styling  
tab\_descriptives\_1 %>%  
 summary(text=TRUE, digits.pct=1, digits=1) %>%  
 kable(caption = "Continuous Outcomes") %>%  
 kable\_styling(bootstrap\_options = "striped", full\_width = FALSE, html\_font = "Cambria",  
 font\_size = 15,  
 position = "center", fixed\_thead = T) %>%  
 row\_spec(2:3, bold = F, extra\_css = 'vertical-align: middle !important;') %>%  
 column\_spec(1, width = "20em", background = "light grey", bold = T, border\_right = T) %>%  
 column\_spec(2, width = "20em", border\_right = T) %>%  
 column\_spec(3, width = "20em", border\_right = T) %>%  
 footnote(general = "Here is a general comments of the table. ") %>%  
 scroll\_box(width = "75%", height = "500px")

Continuous Outcomes

|  | Baseline (N=35) | Year\_1 (N=34) | Year\_2 (N=30) |
| --- | --- | --- | --- |
| age |  |  |  |
| - Count | 35 | 34 | 30 |
| - Mean (SD) | 9.3 (0.5) | 9.5 (0.5) | 9.6 (0.5) |
| - Median | 9.0 | 9.5 | 10.0 |
| - Min - Max | 9.0 - 10.0 | 9.0 - 10.0 | 9.0 - 10.0 |
| vg\_total |  |  |  |
| - Count | 35 | 34 | 30 |
| - Mean (SD) | 8.7 (8.4) | 8.7 (8.1) | 6.8 (4.3) |
| - Median | 6.0 | 7.0 | 7.0 |
| - Min - Max | 0.0 - 28.0 | 0.0 - 28.0 | 1.0 - 16.0 |
| cbcl\_extern |  |  |  |
| - Count | 35 | 34 | 30 |
| - Mean (SD) | 4.8 (5.7) | 5.5 (6.1) | 6.9 (6.9) |
| - Median | 3.0 | 4.0 | 4.5 |
| - Min - Max | 0.0 - 21.0 | 0.0 - 25.0 | 0.0 - 25.0 |
| Note: |  |  |  |
| Here is a general comments of the table. |  |  |  |

### Categorical Outcomes

**This code creates a descriptives table for categorical outcomes**

## Create a descriptives table of study variables by measurement occasion   
tab\_descriptives\_2 <- tableby.control(test=FALSE, total=FALSE,  
numeric.test="kwt", cat.test="chisq",  
numeric.stats=c("N", "meansd", "median", "range"  
), # "Nmiss2"  
cat.stats=c("countpct"), # "Nmiss2"  
stats.labels=list(N='Count', meansd="Mean (SD)", median  
='Median', range='Min - Max'  
)) # , Nmiss2 ='Missing'  
  
my\_cat\_labels <- list(  
 event = "Year",  
 sex = "Sex",  
 mature\_vg = "Mature Video Games"  
 )  
   
tab\_descriptives\_2 <- tableby(event ~ sex + mature\_vg, data=df\_long, control=tab\_descriptives\_2)  
  
#summary(tab\_descriptives\_2, labelTranslations = my\_cat\_labels , text=TRUE, title = #"Categorical Outcomes", term.name = TRUE)  
  
# Push table object through kable and kable\_styling  
tab\_descriptives\_2 %>%  
 summary(text=TRUE, digits.pct=1, digits=1) %>%  
 kable(caption = "Categorical Outcomes") %>%  
 kable\_styling(bootstrap\_options = "striped", full\_width = FALSE, font\_size = 15,  
 position = "center", fixed\_thead = T) %>%  
 row\_spec(2:3, bold = F, extra\_css = 'vertical-align: middle !important;') %>%  
 column\_spec(1, width = "20em", background = "light grey", bold = T, border\_right = T) %>%  
 column\_spec(2, width = "20em", border\_right = T) %>%  
 #column\_spec(3, width = "20em", border\_right = T) %>%  
 footnote(general = "Here is a general comments of the table. ") %>%  
 scroll\_box(width = "75%", height = "500px")

Categorical Outcomes

|  | Baseline (N=35) | Year\_1 (N=34) | Year\_2 (N=30) |
| --- | --- | --- | --- |
| sex |  |  |  |
| - Female | 19 (54.3%) | 20 (58.8%) | 17 (56.7%) |
| - Male | 16 (45.7%) | 14 (41.2%) | 13 (43.3%) |
| mature\_vg |  |  |  |
| - Count | 35 | 34 | 30 |
| - Mean (SD) | 0.6 (0.9) | 1.7 (6.3) | 0.6 (0.9) |
| - Median | 0.0 | 0.0 | 0.0 |
| - Min - Max | 0.0 - 3.0 | 0.0 - 37.0 | 0.0 - 3.0 |
| Note: |  |  |  |
| Here is a general comments of the table. |  |  |  |

:::

## Results

::: panel-tabset ### Build Model {.tabset .tabset-fade .tabset-pills}

The code snippet below tells R to compute a difference score by subtracting each participant’s Externalizing score at T2 from their Externalizing score at T1. Next, a simple regression analyses is conducted to examine whether a grouping variable (participant sex) significantly predicts the difference score value (indicating significant group differences in the average difference score). The simple regression function (lm) is provided by the r ‘stats’ package.

xxx First, compute the difference () between a score on some measure () assessed at baseline () and follow-up (). The result of this formula (= - ) can be included as the outcome variable in a regression model that analyzes the role of the grouping variable (e.g., 0 = boy; 1 = girl) on changes in scores on the measure across follow-ups. xxx

**STEP 1: Compute Difference Score**

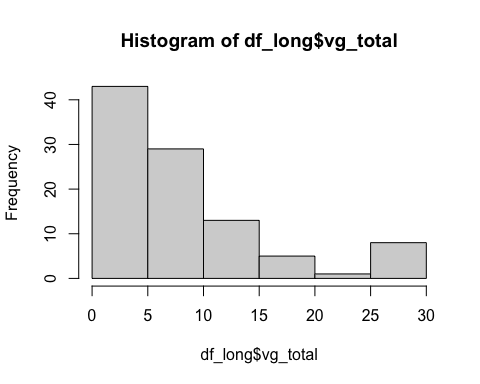
# Compute difference score based on CBCL Externalizing subscale scores at baseline (t1) and 1-Year Follow-up (t2)  
df\_wide$diffscore <- (df\_wide$cbcl\_extern\_Year\_1 - df\_wide$cbcl\_extern\_Baseline)

# Compute statistical summaries for the difference score variable  
summary(df\_wide$diffscore)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
-27.000 1.000 5.000 5.576 8.000 33.000

This summary of the difference score variable indicates xxxxx.

hist(df\_long$vg\_total)



## Summary statistics  
summary<-df\_long %>%  
group\_by(event) %>%  
get\_summary\_stats(cbcl\_extern, type = "mean\_sd")  
data.frame(summary)

event variable n mean sd  
1 Baseline cbcl\_extern 35 4.771 5.678  
2 Year\_1 cbcl\_extern 34 5.529 6.091  
3 Year\_2 cbcl\_extern 30 6.867 6.897

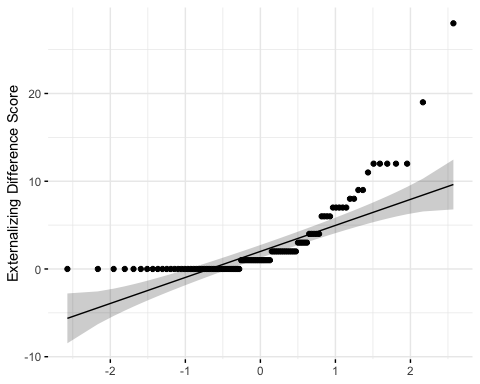
This histogram indicates xxxxxx. The summary command shows xxxxxx.

### Shapiro-Wilk test and normality (Q-Q) plot (visualization of correlation between a given sample and the normal distribution)  
shapiro.test(df\_wide$cbcl\_extern\_Baseline[0:5000])

Shapiro-Wilk normality test  
  
data: df\_wide$cbcl\_extern\_Baseline[0:5000]  
W = 0.67151, p-value = 1.452e-13

This Shapiro-wilkes test shows xxxxxx.

qqplot <- ggqqplot(df\_wide$cbcl\_extern\_Baseline,  
 ylab = "Externalizing Difference Score", xlab = FALSE,  
 ggtheme = theme\_minimal()  
)  
  
suppressWarnings(print(qqplot))



This qqplot shows xxxxxx.

**STEP 2: Conduct simple regression on Difference Score**

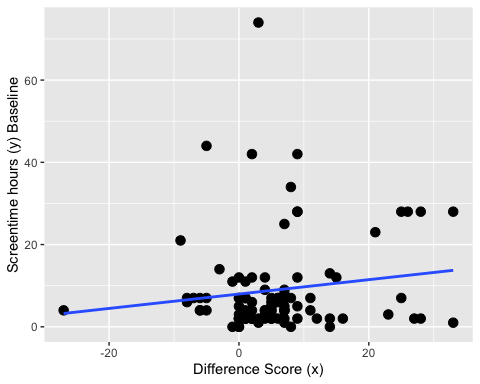
#Simple linear regression is used to examine the relationship between a continuous predictor variable and the average difference score.  
  
#Regression model  
result <- lm(df\_wide$diffscore ~ vg\_total\_Baseline, data = df\_wide)  
#summarize the reults  
summary(result)

Call:  
lm(formula = df\_wide$diffscore ~ vg\_total\_Baseline, data = df\_wide)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-32.022 -4.685 -0.387 2.091 28.315   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 4.57285 1.18272 3.866 0.0002 \*\*\*  
vg\_total\_Baseline 0.11219 0.08052 1.393 0.1667   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 9.337 on 97 degrees of freedom  
Multiple R-squared: 0.01962, Adjusted R-squared: 0.009511   
F-statistic: 1.941 on 1 and 97 DF, p-value: 0.1667

The output from our model provides: i. a parameter estimate; ii. standard error; iii. p-value. In our example p = .001 which is less than .05 indicating significant differences in predicted cbcl externalizing scores between boys and girls in the ABCD Study. An examination of the regression coeefficient for sex (female = 0; male = 1) is b = .xx, indicating the difference in cbcl externalizing scores from Time 1 to Time 2 is significantly greater in boys taking part in the ABCD Study relative to girls. Then we conclude that participant sex (does/does not) predict change in cbcl externalizing scores from t1 to t2.

### Model Plots

#Scatterplot to visualize relationship between a continuous predictor and a difference score  
scatterplot <- ggplot (df\_wide, aes(x = diffscore, y = vg\_total\_Baseline)) + geom\_point(size = 3) + geom\_smooth(method = lm, se = F) +  
xlab("Difference Score (x)") +  
ylab("Screentime hours (y) Baseline")  
  
suppressWarnings(print(scatterplot))



Examination of this scatterplot indicates xxxxx.

This output shows xxxxxxxx. Briefly walk through each metric

## Wrapping Up

### Write-up

A simple regression analysis was conducted to examine whether participant’s sex predicted change in cbcl externalzing scores measured across two timepoints. Findings showed a significant positive effect such that boys demonstrated increases in cbcl externalzing scores relative to girls (b = .xx; SE = .xx; p = .xx)