# ACRM 2018 Longitudinal Data Analysis Workshop

Keith Lohse1, and Al Kozlowski2, 2018-09-07

1 University of Utah; 2 Michigan State University, Mary Free Bed Rehabilitation Hospital

## Practical Session 3: The Effects of Missing Data

This handout is designed to accompany the script you will be working with in the practical session. A copy of the script file, the data, set, and this handout can be found at: <https://github.com/keithlohse/LMER_Clinical_Science>.

The R code is interspersed with explanations below. All R code is highlighted in grey and color coded to show different functions, arguments, and comments in the code.

First, you will need to open the five packages we will be using for this session using the library function:

# Loading the essential libraries.

library**(**"ggplot2"**)**; library**(**"lme4"**)**; library**(**"car"**)**; library**(**"dplyr"**)**; library**(**"lmerTest"**)**;

If you haven’t already installed these packages, you will need to use the install.packages() function first. This can take some time and will require an internet connect.

# If these packages are not installed already, run the following code:

install.packages**(**"ggplot2"**)**; install.packages**(**"lme4"**)**; install.packages**(**"car"**)**; install.packages**(**"dplyr"**)**; install.packages**(**"lmerTest"**)**;

Next, as we have done in the past, we will need to set our working directory to on LMER project folder and import the dataset we saved at the end of Session 2.

## Setting the Directory ----------------------------------------------------

getwd**()**

# Note that you need to use forward slashes. R will give you an error if you use forward slashes.

setwd**(**"C:/Users/Folder/SubFolder/SubSubFolder/"**)**

list.files**()**

# Make sure that the file data\_session2.csv is saved in your working directory.

# Import the .csv file into R.

# We will save this file in the R environment as an object called "DATA".

DATA**<-**read.csv**(**"./data\_session3.csv", header **=** **TRUE**, sep**=**",",

na.strings**=**c**(**"NA","NaN"," ",""**))**

# Use the head() function to check the structure of the data file.

head**(**DATA**)**

# Alternately you can also download the data file from the web here:

# DATA <- read.csv("https://raw.githubusercontent.com/keithlohse/LMER\_Clinical\_Science/master/data/data\_session2.csv")

# head(DATA)

## 3.1 Visualizing Missing Data

Up to this point, we have been working with a much idealized dataset in which every participant has an equal and complete set of observations. When doing clinical science “in the wild” this is almost never going to happen. We will have participants observed for differing lengths of time (e.g., “inpatient stay” might be two weeks for some but four weeks for others). We are also likely to have participants with different amounts of missing data (e.g., participants may withdraw from a study, move away, or even die, all of which result in missing data).

“Missingness” is an interesting phenomenon and the reasons for missing data interact with our design (e.g., the density and number of time-points) to ultimately shape our data. Many traditional methods of dealing with missingness are problematic. For instance, in a repeated measures ANOVA you need complete data for all participants. This means that participants with missing data either need to be excluded (which reduces or sample size) or the missing data needs to be imputed in some way (which biases the estimate of the variance).

One of the strengths of the LMER approach (compared to RM ANOVA) is that it can handle missing data very flexibly, estimating slopes and intercepts for individual participants that are weighted proportionally to the number of observations (i.e., the slope for participant with 18 observations has a smaller standard error than a participant with 8 observations). However, this does not mean that we can ignore the effects of missing data when using LMER. In the exercises below, we consider three different cases: ***Missing at Random*** (MAR) in which data from any time-point are equally likely to be missing; ***Missing Not at Random*** (MNAR) in which data from later time-points are more likely to be missing; and ***Last Observation Carried Forward*** (LOCF), which is a common method of imputation in which the last observation is used in place of missing values.

## ------------------- Visualizing Missing Data -----------------------------

## FIM scores with complete data --------------------------------------------

g1**<-**ggplot**(**DATA, aes**(**x **=** month, y **=** rasch\_FIM**))** **+**

geom\_point**(**aes**(**fill**=**as.factor**(**subID**))**, pch**=**21, size**=**2, stroke**=**1.25**)** **+**

geom\_line**(**aes**(**group**=**subID**))** **+**

facet\_wrap**(~**AIS\_grade**)**

g2**<-**g1**+**scale\_x\_continuous**(**name **=** "Time from Admission (Months)", breaks**=**c**(**0**:**18**))** **+**

scale\_y\_continuous**(**name **=** "Rasch-Scaled FIM Score (0-100)",limits**=**c**(**0,100**))**

g3 **<-** g2 **+** theme\_bw**()** **+**

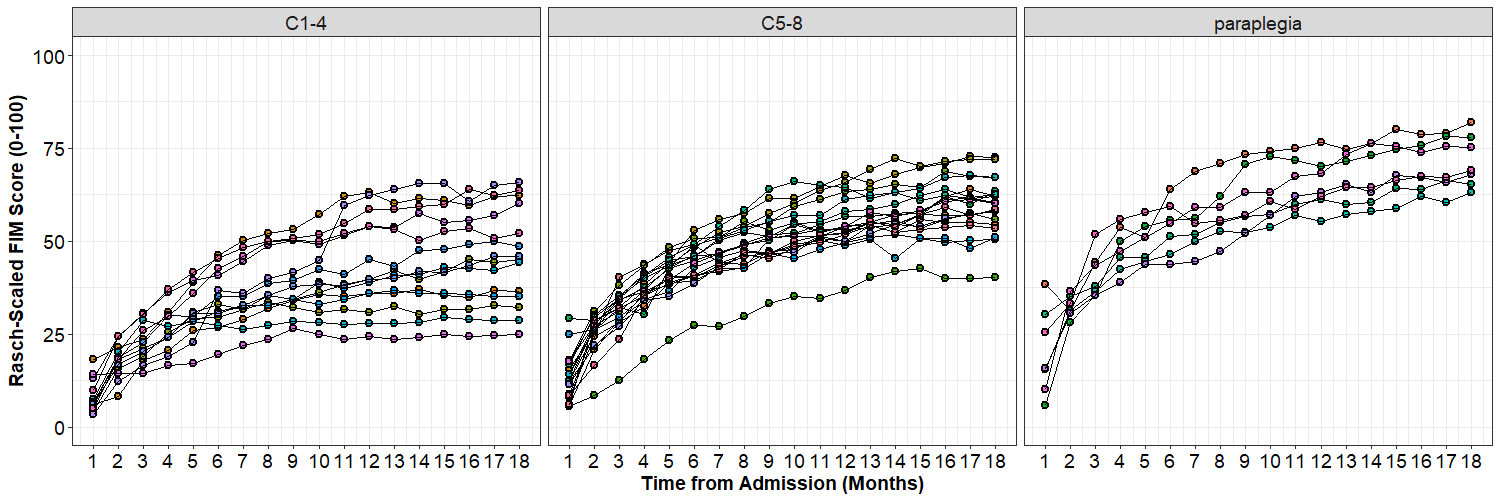
theme**(**axis.text**=**element\_text**(**size**=**14, colour**=**"black"**)**,

axis.title**=**element\_text**(**size**=**14,face**=**"bold"**))** **+**

theme**(**strip.text.x **=** element\_text**(**size **=** 14**))+**

theme**(**legend.position**=**"none"**)**

plot**(**g3**)**



## FIM scores with data missing random --------------------------------------

g1**<-**ggplot**(**DATA, aes**(**x **=** month, y **=** rasch\_FIM\_MAR**))** **+**

geom\_point**(**aes**(**fill**=**as.factor**(**subID**))**, pch**=**21, size**=**2, stroke**=**1.25**)** **+**

geom\_line**(**aes**(**group**=**subID**))** **+**

facet\_wrap**(~**AIS\_grade**)**

g2**<-**g1**+**scale\_x\_continuous**(**name **=** "Time from Admission (Months)", breaks**=**c**(**0**:**18**))** **+**

scale\_y\_continuous**(**name **=** "Rasch-Scaled FIM Score (0-100)",limits**=**c**(**0,100**))**

g3 **<-** g2 **+** theme\_bw**()** **+**

theme**(**axis.text**=**element\_text**(**size**=**14, colour**=**"black"**)**,

axis.title**=**element\_text**(**size**=**14,face**=**"bold"**))** **+**

theme**(**strip.text.x **=** element\_text**(**size **=** 14**))+**

theme**(**legend.position**=**"none"**)**

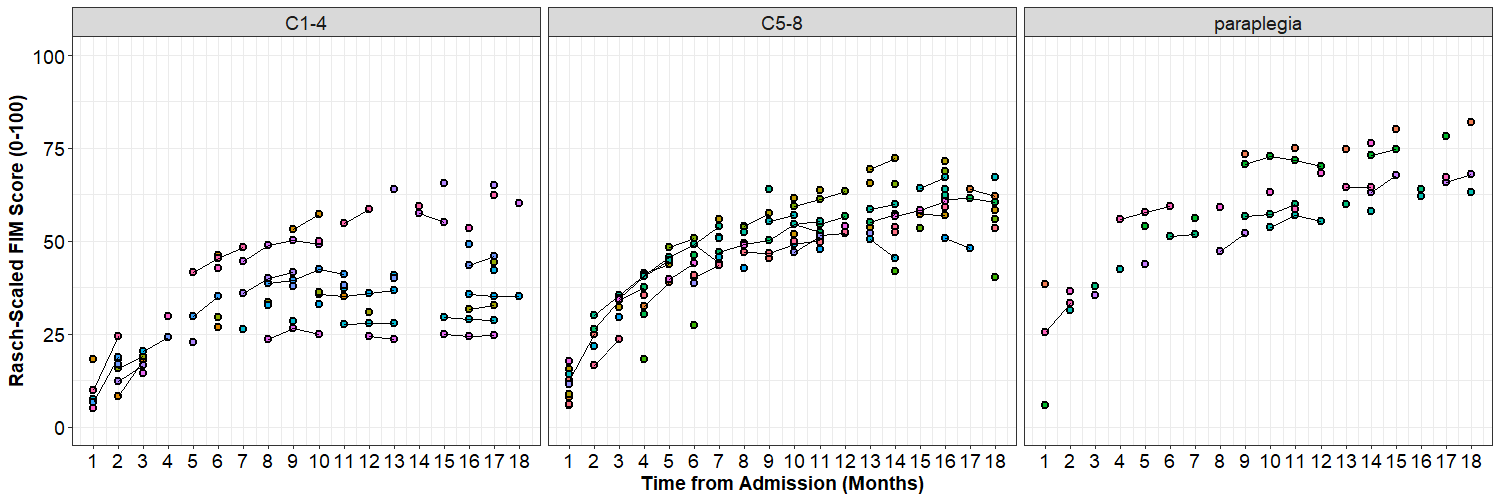
plot**(**g3**)**

Warning messages:

1: Removed 437 rows containing missing values (geom\_point).

2: Removed 126 rows containing missing values (geom\_path).

# Note we get a warning message due to the missing values.



## FIM scores with data missing not at random -------------------------------

g1**<-**ggplot**(**DATA, aes**(**x **=** month, y **=** rasch\_FIM\_MNAR**))** **+**

geom\_point**(**aes**(**fill**=**as.factor**(**subID**))**, pch**=**21, size**=**2, stroke**=**1.25**)** **+**

geom\_line**(**aes**(**group**=**subID**))** **+**

facet\_wrap**(~**AIS\_grade**)**

g2**<-**g1**+**scale\_x\_continuous**(**name **=** "Time from Admission (Months)", breaks**=**c**(**0**:**18**))** **+**

scale\_y\_continuous**(**name **=** "Rasch-Scaled FIM Score (0-100)",limits**=**c**(**0,100**))**

g3 **<-** g2 **+** theme\_bw**()** **+**

theme**(**axis.text**=**element\_text**(**size**=**14, colour**=**"black"**)**,

axis.title**=**element\_text**(**size**=**14,face**=**"bold"**))** **+**

theme**(**strip.text.x **=** element\_text**(**size **=** 14**))+**

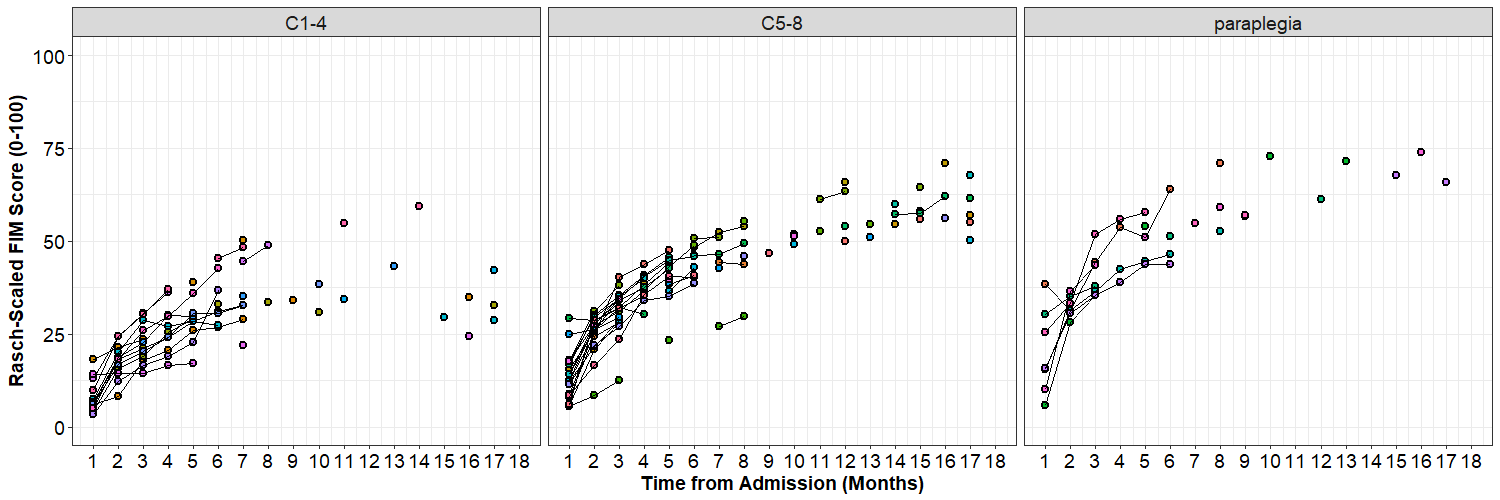
theme**(**legend.position**=**"none"**)**

plot**(**g3**)**

Warning messages:

1: Removed 445 rows containing missing values (geom\_point).

2: Removed 228 rows containing missing values (geom\_path).



## FIM scores with Last Observation Carried Forward -------------------------

g1**<-**ggplot**(**DATA, aes**(**x **=** month, y **=** rasch\_FIM\_LOCF**))** **+**

geom\_point**(**aes**(**fill**=**as.factor**(**subID**))**, pch**=**21, size**=**2, stroke**=**1.25**)** **+**

geom\_line**(**aes**(**group**=**subID**))** **+**

facet\_wrap**(~**AIS\_grade**)**

g2**<-**g1**+**scale\_x\_continuous**(**name **=** "Time from Admission (Months)", breaks**=**c**(**0**:**18**))** **+**

scale\_y\_continuous**(**name **=** "Rasch-Scaled FIM Score (0-100)",limits**=**c**(**0,100**))**

g3 **<-** g2 **+** theme\_bw**()** **+**

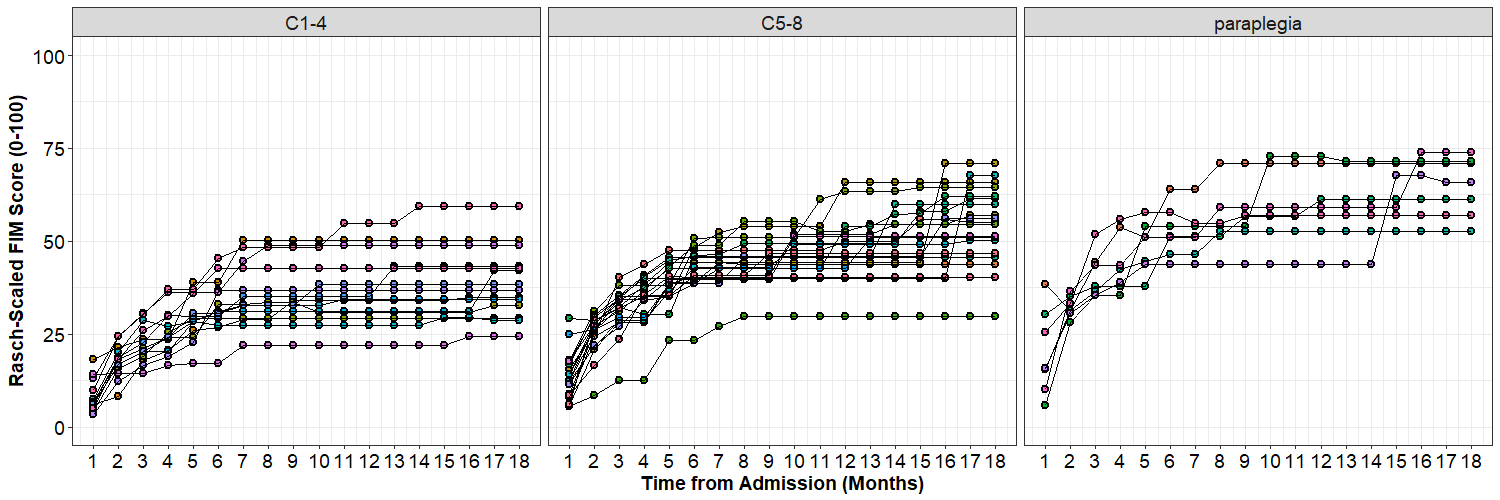
theme**(**axis.text**=**element\_text**(**size**=**14, colour**=**"black"**)**,

axis.title**=**element\_text**(**size**=**14,face**=**"bold"**))** **+**

theme**(**strip.text.x **=** element\_text**(**size **=** 14**))+**

theme**(**legend.position**=**"none"**)**

plot**(**g3**)**



## 2.2 Identifying Missing Data

To really see the difference between the MAR and MNAR data, lets plot the proportion of missing values at each time-point. In order to plot these frequencies, we will use the is.na() function in R. This function returns the value of “TRUE” if the value is missing and “FALSE” and a value is present. We then use the as.numeric() function to convert these logical TRUE/FALSE statements into 1s and 0s respectively.

## -------------------- Identifying Missing Data ----------------------------

# Missing at Random

DATA**$**MAR\_missing**<-**as.numeric**(**is.na**(**DATA**$**rasch\_FIM\_MAR**))**

summary**(**DATA**$**MAR\_missing**)**

xtabs**(**MAR\_missing **~** month, DATA**)**

x**<-**as.data.frame**(**xtabs**(**MAR\_missing **~** month, DATA**))**

x

g1**<-**ggplot**(**x, aes**(**x **=** month, y**=**Freq**))** **+**

geom\_col**(**fill**=** "light grey", color**=**"black"**)**

g2**<-**g1**+**scale\_x\_discrete**(**name **=** "Time from Admission (Months)"**)** **+**

scale\_y\_continuous**(**name **=** "Count of Missing Data", limits**=**c**(**0,40**))**

g3 **<-** g2 **+** theme\_bw**()** **+**

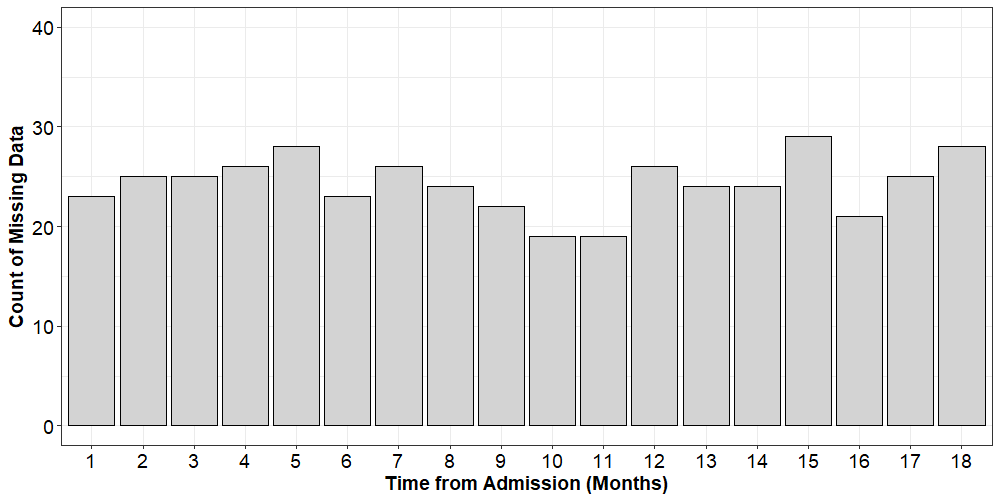
theme**(**axis.text**=**element\_text**(**size**=**14, colour**=**"black"**)**,

axis.title**=**element\_text**(**size**=**14,face**=**"bold"**))** **+**

theme**(**strip.text.x **=** element\_text**(**size **=** 14**))+**

theme**(**legend.position**=**"none"**)**

plot**(**g3**)**



# Missing Not at Random

DATA**$**MNAR\_missing**<-**as.numeric**(**is.na**(**DATA**$**rasch\_FIM\_MNAR**))**

summary**(**DATA**$**MNAR\_missing**)**

xtabs**(**MNAR\_missing **~** month, DATA**)**

y**<-**as.data.frame**(**xtabs**(**MNAR\_missing **~** month, DATA**))**

y

g1**<-**ggplot**(**y, aes**(**x **=** month, y**=**Freq**))** **+**

geom\_col**(**fill**=** "light grey", color**=**"black"**)**

g2**<-**g1**+**scale\_x\_discrete**(**name **=** "Time from Admission (Months)"**)** **+**

scale\_y\_continuous**(**name **=** "Count of Missing Data", limits**=**c**(**0,40**))**

g3 **<-** g2 **+** theme\_bw**()** **+**

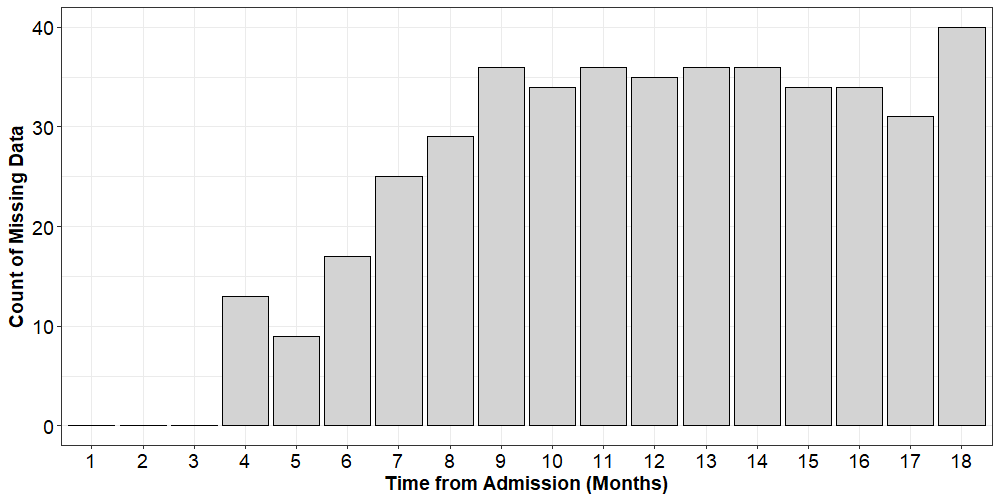
theme**(**axis.text**=**element\_text**(**size**=**14, colour**=**"black"**)**,

axis.title**=**element\_text**(**size**=**14,face**=**"bold"**))** **+**

theme**(**strip.text.x **=** element\_text**(**size **=** 14**))+**

theme**(**legend.position**=**"none"**)**

plot**(**g3**)**



## 2.3 Comparing Different Effects of Time

To understand the effects that different kinds of missingness (MAR or MNAR) or imputation (LOCF) have on our models, let’s take our cubic model from the previous session and apply it to the three different types of dependent variable.

## -------------- The Effects of Missingness on Time ------------------------

# Cubic model with complete data

complete**<-**lmer**(**rasch\_FIM**~**

# Fixed-effects

1**+**year.0**+**year.0\_sq**+**year.0\_cu**+**

# Random-effects

**(**1**+**year.0**+**year.0\_sq**+**year.0\_cu**|**subID**)**, data**=**DATA, REML**=FALSE)**

# Cubic model with data Missing at Random

MAR**<-**lmer**(**rasch\_FIM\_MAR**~**

# Fixed-effects

1**+**year.0**+**year.0\_sq**+**year.0\_cu**+**

# Random-effects

**(**1**+**year.0**+**year.0\_sq**+**year.0\_cu**|**subID**)**, data**=**DATA, REML**=FALSE)**

# Cubic model with Missing Not at Random

MNAR**<-**lmer**(**rasch\_FIM\_MNAR**~**

# Fixed-effects

1**+**year.0**+**year.0\_sq**+**year.0\_cu**+**

# Random-effects

**(**1**+**year.0**+**year.0\_sq**+**year.0\_cu**|**subID**)**, data**=**DATA, REML**=FALSE)**

# Cubic model with Last Observation Carried Forward

LOCF**<-**lmer**(**rasch\_FIM\_LOCF**~**

# Fixed-effects

1**+**year.0**+**year.0\_sq**+**year.0\_cu**+**

# Random-effects

**(**1**+**year.0**+**year.0\_sq**+**year.0\_cu**|**subID**)**, data**=**DATA, REML**=FALSE)**

summary**(**complete**)**

Linear mixed model fit by maximum likelihood t-tests use Satterthwaite approximations to degrees of freedom [lmerMod]

Formula: rasch\_FIM ~ 1 + year.0 + year.0\_sq + year.0\_cu + (1 + year.0 +

year.0\_sq + year.0\_cu | subID)

Data: DATA

**AIC BIC logLik deviance df.resid**

3741.4 3810.1 -1855.7 3711.4 705

Scaled residuals:

Min 1Q Median 3Q Max

-4.1697 -0.5265 -0.0002 0.5186 3.5018

**Random effects:**

**Groups Name Variance Std.Dev. Corr**

subID (Intercept) 42.779 6.541

year.0 635.959 25.218 0.06

year.0\_sq 1179.868 34.349 -0.07 -0.88

year.0\_cu 213.005 14.595 0.10 0.77 -0.98

Residual 5.481 2.341

Number of obs: 720, groups: subID, 40

**Fixed effects:**

**Estimate Std. Error df t value Pr(>|t|)**

(Intercept) 14.977 1.073 40.000 13.954 < 2e-16 \*\*\*

year.0 95.033 4.378 40.000 21.706 < 2e-16 \*\*\*

year.0\_sq -83.232 6.213 40.000 -13.397 2.22e-16 \*\*\*

year.0\_cu 26.483 2.698 40.000 9.815 3.30e-12 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

(Intr) year.0 yr.0\_s

year.0 -0.041

year.0\_sq 0.027 -0.891

year.0\_cu 0.001 0.797 -0.981

summary**(**MAR**)**

Linear mixed model fit by maximum likelihood t-tests use Satterthwaite approximations to degrees of freedom [lmerMod]

Formula: rasch\_FIM\_MAR ~ 1 + year.0 + year.0\_sq + year.0\_cu + (1 + year.0 +

year.0\_sq + year.0\_cu | subID)

Data: DATA

**AIC BIC logLik deviance df.resid**

1578.7 1633.4 -774.4 1548.7 268

Scaled residuals:

Min 1Q Median 3Q Max

-3.1629 -0.5279 -0.0332 0.4599 2.8911

**Random effects:**

**Groups Name Variance Std.Dev. Corr**

subID (Intercept) 69.001 8.307

year.0 1070.150 32.713 -0.21

year.0\_sq 2259.949 47.539 0.17 -0.94

year.0\_cu 387.773 19.692 -0.12 0.90 -0.99

Residual 3.774 1.943

Number of obs: 283, groups: subID, 40

**Fixed effects:**

**Estimate Std. Error df t value Pr(>|t|)**

(Intercept) 14.454 1.440 36.330 10.039 5.06e-12 \*\*\*

year.0 96.908 6.144 39.880 15.774 < 2e-16 \*\*\*

year.0\_sq -86.756 9.197 41.260 -9.433 7.43e-12 \*\*\*

year.0\_cu 28.363 3.933 40.750 7.211 8.60e-09 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

(Intr) year.0 yr.0\_s

year.0 -0.350

year.0\_sq 0.291 -0.945

year.0\_cu -0.237 0.898 -0.990

summary**(**MNAR**)**

Linear mixed model fit by maximum likelihood t-tests use Satterthwaite approximations to degrees of freedom [lmerMod]

Formula: rasch\_FIM\_MNAR ~ 1 + year.0 + year.0\_sq + year.0\_cu + (1 + year.0 +

year.0\_sq + year.0\_cu | subID)

Data: DATA

**AIC BIC logLik deviance df.resid**

1651.8 1706.0 -810.9 1621.8 260

Scaled residuals:

Min 1Q Median 3Q Max

-2.51404 -0.49176 -0.03217 0.50924 2.64296

**Random effects:**

**Groups Name Variance Std.Dev. Corr**

subID (Intercept) 42.712 6.535

year.0 1216.743 34.882 -0.07

year.0\_sq 3433.180 58.593 0.08 -0.95

year.0\_cu 787.349 28.060 -0.01 0.93 -1.00

Residual 8.101 2.846

Number of obs: 275, groups: subID, 40

**Fixed effects:**

**Estimate Std. Error df t value Pr(>|t|)**

(Intercept) 13.998 1.102 39.890 12.705 1.33e-15 \*\*\*

year.0 109.302 6.540 36.790 16.713 < 2e-16 \*\*\*

year.0\_sq -114.288 12.300 32.430 -9.292 1.17e-10 \*\*\*

year.0\_cu 42.066 6.339 32.420 6.636 1.64e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

(Intr) year.0 yr.0\_s

year.0 -0.196

year.0\_sq 0.188 -0.939

year.0\_cu -0.130 0.886 -0.987

summary**(**LOCF**)**

Linear mixed model fit by maximum likelihood t-tests use Satterthwaite approximations to degrees of freedom [lmerMod]

Formula: rasch\_FIM\_LOCF ~ 1 + year.0 + year.0\_sq + year.0\_cu + (1 + year.0 +

year.0\_sq + year.0\_cu | subID)

Data: DATA

**AIC BIC logLik deviance df.resid**

4107.0 4175.7 -2038.5 4077.0 705

Scaled residuals:

Min 1Q Median 3Q Max

-3.2159 -0.4227 0.0041 0.4549 4.5406

**Random effects:**

**Groups Name Variance Std.Dev. Corr**

subID (Intercept) 43.598 6.603

year.0 586.840 24.225 -0.04

year.0\_sq 1739.954 41.713 0.06 -0.86

year.0\_cu 458.504 21.413 -0.02 0.73 -0.97

Residual 9.326 3.054

Number of obs: 720, groups: subID, 40

**Fixed effects:**

**Estimate Std. Error df t value Pr(>|t|)**

(Intercept) 14.892 1.109 40.000 13.426 2.22e-16 \*\*\*

year.0 96.284 4.498 40.020 21.405 < 2e-16 \*\*\*

year.0\_sq -101.753 7.680 40.030 -13.249 4.44e-16 \*\*\*

year.0\_cu 37.040 3.846 40.020 9.632 5.60e-12 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

(Intr) year.0 yr.0\_s

year.0 -0.176

year.0\_sq 0.166 -0.887

year.0\_cu -0.107 0.770 -0.972

First, let’s look at the amount of data available in each model. In the MAR data, about 61% of the data are missing, but recall these data are missing from all different times. In the MNAR data, 62% of the data are missing, but these data are missing largely from the later time-points. We also include the deviance and the AIC from these models so that you can see how these values depend on the number of observations. This should reinforce the idea that we can only compare the deviance and AIC from datasets of the same size!

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Complete** | **MAR** | **MNAR** | **LOCF** |
| # of Subjects | 40 | 40 | 40 | 40 |
| # of Observations | 720 | 283 | 275 | 720 |
| Deviance | 3711.4 | 1548.7 | 1621.8 | 4077.0 |
| AIC | 3741.4 | 1478.7 | 1651.8 | 4107.0 |

Next, let’s look at the effect that missing data has on the fixed-effects of our cubic time model. Each model contains a point-estimate for the effect (β) and a standard error (SE). Broadly speaking, we expect the SE to be inversely related to the amount of data (i.e., more missing data should mean bigger SE). Similarly, we expect the point estimates to change as a function of the “shape” of the data. The complete dataset is our archtypal shape. The MAR dataset will mimic this shape, albeit with fewer data points. The MNAR dataset departs from this shape because most of the data is only available early in time, when the rate of change is at its hightest. Similarly, the LOCF dataset departs from this shape because it looses the fine resolution of change in later time-points. (Look back at the LOCF data above to see this “stair-case” effect in the data.)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Complete** | | **MAR** | | **MNAR** | | **LOCF** | |
| **Effect** | **β** | **SE** | **β** | **SE** | **β** | **SE** | **β** | **SE** |
| Time (linear) | 95.03 | 4.38 | 96.91 | 6.14 | 109.30 | 6.54 | 96.28 | 4.50 |
| Time (quadratic) | -83.23 | 6.21 | -86.76 | 9.20 | -114.28 | 12.30 | -101.75 | 7.68 |
| Time (cubic) | 26.48 | 2.70 | 28.36 | 3.93 | 42.07 | 6.34 | 37.04 | 3.85 |

We can more easily show these effects by expressing each β and SE as a proportion of the β/SE from the complete data set.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **MAR** | | **MNAR** | | **LOCF** | |
| **Effect** | **β** | **SE** | **β** | **SE** | **β** | **SE** |
| Time (linear) | 1.02 | 1.40 | 1.15 | 1.49 | 1.01 | 1.03 |
| Time (quadratic) | 1.04 | 1.48 | 1.37 | 1.98 | 1.22 | 1.23 |
| Time (cubic) | 1.07 | 1.46 | 1.59 | 2.35 | 1.40 | 1.42 |

If the missingness or our imputation is not having an effect, these values should be close to one. You can see that MAR does well in terms of the estimate of β, but the SE’s are larger by a factor of 40 to 50%. This isn’t too bad; it means our estimates are less certain, but not biased. In contrast, look at the MNAR where the β’s are biased to overestimate. However, the SE’s are also biased upward, so at least as our estimate gets less precise our confidence is correspondingly going down. The LOCF estimates, in our opinion, pose the biggest threat to decision making. Note the β’s can overestimate the effect by as much as 40%, but the SE’s are somewhere between the MAR and the complete dataset. As such, by carrying the last observation forward, we might have a spurious confidence in our model’s estimates.

These general effects on the β’s and the SE’s will be true, on average, across different types on data sets. In the current dataset, however, the distortions might not seem too bad. Keep in mind, however, that we have an idealized dataset for two reasons. First, our simulated data were generated from mathematical functions with relatively little noise. Second, we have a lot of time-points, 18, in our study. In most studies, this level of resolution in time is probably not feasible.