Data Structures and Longitudinal Analysis

Andrew S. Zieffler

Wide Format

- Response variable occupies multiple columns
  - Single row for each subject (*subjects-by-variables*)
- Format used for RM-ANOVA and RM-MANOVA

Each subjects' data contained in a single row

	subid	read.5	read.6	read.7	read.8	risk	gen	eth	ell	sped	att
1	1	172	185	179	194	HHM	F	Afr	0	N	0.94
2	2	200	210	209	-99	HHM	F	Afr	0	N	0.91
3	3	191	199	203	215	HHM	M	Afr	0	N	0.97
4	4	200	195	194	-99	HHM	F	Afr	0	N	0.88
5	5	207	213	212	213	HHM	F	Afr	0	N	0.85
6	6	191	189	206	195	HHM	M	Afr	0	N	0.90
7	7	199	208	213	218	POV	M	Afr	1	N	0.97
8	8	191	194	194	-99	POV	F	His	1	Y	0.97
9	9	149	154	174	177	POV	F	Afr	0	Y	0.97
10	10	200	212	213	-99	POV	F	Afr	0	N	0.96
11	11	218	231	233	239	POV	F	Whi	0	N	0.98
12	12	228	232	248	246	POV	F	Whi	0	Y	0.96
13	13	228	236	228	239	ADV	F	Whi	0	N	0.99
14	14	199	210	225	235	ADV	M	Afr	1	N	0.99
15	15	218	223	236	-99	ADV	F	Whi	0	N	1.00
16	16	228	226	234	227	ADV	M	Whi	0	N	0.97
17	17	201	210	208	219	ADV	M	Whi	0	N	0.98
18	18	218	220	217	221	ADV	M	Whi	0	N	1.00
19	19	215	216	221	-99	ADV	F	Whi	0	N	0.96
20	20	204	215	219	214	ADV	F	Afr	0	N	0.95
21	21	237	241	243	-99	ADV	M	Whi	0	N	0.98
22	22	219	233	236	-99	ADV	F	Afr	0	N	0.96

Response Measure

	subid	read.5	read.6	read.7	read.8	risk	gen	eth	ell	sped	att
1	1	172	185	179	194	HHM	F	Afr	0	N	0.94
2	2	200	210	209	-99	HHM	F	Afr	0	N	0.91
3	3	191	199	203	215	HHM	M	Afr	0	N	0.97
4	4	200	195	194	-99	HHM	F	Afr	0	N	0.88
5	5	207	213	212	213	HHM	F	Afr	0	N	0.85
6	6	191	189	206	195	HHM	M	Afr	0	N	0.90
7	7	199	208	213	218	POV	M	Afr	1	N	0.97
8	8	191	194	194	-99	POV	F	His	1	Y	0.97
9	9	149	154	174	177	POV	F	Afr	0	Y	0.97
10	10	200	212	213	-99	POV	F	Afr	0	N	0.96
11	11	218	231	233	239	POV	F	Whi	0	N	0.98
12	12	228	232	248	246	POV	F	Whi	0	Y	0.96
13	13	228	236	228	239	ADV	F	Whi	0	N	0.99
14	14	199	210	225	235	ADV	M	Afr	1	N	0.99
15	15	218	223	236	-99	ADV	F	Whi	0	N	1.00
16	16	228	226	234	227	ADV	M	Whi	0	N	0.97
17	17	201	210	208	219	ADV	M	Whi	0	N	0.98
18	18	218	220	217	221	ADV	M	Whi	0	N	1.00
19	19	215	216	221	-99	ADV	F	Whi	0	N	0.96
20	20	204	215	219	214	ADV	F	Afr	0	N	0.95
21	21	237	241	243	-99	ADV	M	Whi	0	N	0.98
22	22	219	233	236	-99	ADV	F	Afr	0	N	0.96

Static Predictor(s)

	subid	read.5	read.6	read.7	read.8	risk	gen	eth	ell	sped	att
1	1	172	185	179	194	HHM	F	Afr	0	N	0.94
2	2	200	210	209	-99	HHM	F	Afr	0	N	0.91
3	3	191	199	203	215	HHM	M	Afr	0	N	0.97
4	4	200	195	194	-99	HHM	F	Afr	0	N	0.88
5	5	207	213	212	213	HHM	F	Afr	0	N	0.85
6	6	191	189	206	195	HHM	M	Afr	0	N	0.90
7	7	199	208	213	218	POV	M	Afr	1	N	0.97
8	8	191	194	194	-99	POV	F	His	1	Y	0.97
9	9	149	154	174	177	POV	F	Afr	0	Y	0.97
10	10	200	212	213	-99	POV	F	Afr	0	N	0.96
11	11	218	231	233	239	POV	F	Whi	0	N	0.98
12	12	228	232	248	246	POV	F	Whi	0	Y	0.96
13	13	228	236	228	239	ADV	F	Whi	0	N	0.99
14	14	199	210	225	235	ADV	M	Afr	1	N	0.99
15	15	218	223	236	-99	ADV	F	Whi	0	N	1.00
16	16	228	226	234	227	ADV	M	Whi	0	N	0.97
17	17	201	210	208	219	ADV	M	Whi	0	N	0.98
18	18	218	220	217	221	ADV	M	Whi	0	N	1.00
19	19	215	216	221	-99	ADV	F	Whi	0	N	0.96
20	20	204	215	219	214	ADV	F	Afr	0	N	0.95
21	21	237	241	243	-99	ADV	M	Whi	0	N	0.98
22	22	219	233	236	-99	ADV	F	Afr	0	N	0.96

Long Format

- Response variable occupies single column
  - Multiple rows for each subject
- Static predictors are constant among rows
- Dynamic predictors vary among rows
- Format used for mixed-effects models

Each subjects' data contained in a multiple rows

	subid	risk	gen	eth	ell	sped	att	grade	read
1	1	HHM	F	Afr	0	N	0.94	5	172
2	1	HHM	F	Afr	0	N	0.94	6	185
3	1	HHM	F	Afr	0	N	0.94	7	179
4	1	HHM	F	Afr	0	N	0.94	8	194
5	2	HHM	F	Afr	0	N	0.91	5	200
6	2	HHM	F	Afr	0	N	0.91	6	210
7	2	HHM	F	Afr	0	N	0.91	7	209
8	2	HHM	F	Afr	0	N	0.91	8	-99
9	3	HHM	M	Afr	0	N	0.97	5	191
10	3	HHM	M	Afr	0	N	0.97	6	199
11	3	HHM	M	Afr	0	N	0.97	7	203
12	3	HHM	M	Afr	0	N	0.97	8	215
13	4	HHM	F	Afr	0	N	0.88	5	200
14	4	HHM	F	Afr	0	N	0.88	6	195
15	4	HHM	F	Afr	0	N	0.88	7	194
16	4	HHM	F	Afr	0	N	0.88	8	-99
17	5	HHM	F	Afr	0	N	0.85	5	207
18	5	HHM	F	Afr	0	N	0.85	6	213
19	5	HHM	F	Afr	0	N	0.85	7	212
20	5	HHM	F	Afr	0	N	0.85	8	213
21	6	HHM	M	Afr	0	N	0.90	5	191
22	6	HHM	M	Afr	0	N	0.90	6	189

Response Measure

	subid	risk	gen	eth	ell	sped	att	grade	read
1	1	HHM	F	Afr	0	N	0.94	5	172
2	1	HHM	F	Afr	0	N	0.94	6	185
3	1	HHM	F	Afr	0	N	0.94	7	179
4	1	HHM	F	Afr	0	N	0.94	8	194
5	2	HHM	F	Afr	0	N	0.91	5	200
6	2	HHM	F	Afr	0	N	0.91	6	210
7	2	HHM	F	Afr	0	N	0.91	7	209
8	2	HHM	F	Afr	0	N	0.91	8	-99
9	3	HHM	M	Afr	0	N	0.97	5	191
10	3	HHM	M	Afr	0	N	0.97	6	199
11	3	HHM	M	Afr	0	N	0.97	7	203
12	3	HHM	M	Afr	0	N	0.97	8	215
13	4	HHM	F	Afr	0	N	0.88	5	200
14	4	HHM	F	Afr	0	N	0.88	6	195
15	4	HHM	F	Afr	0	N	0.88	7	194
16	4	HHM	F	Afr	0	N	0.88	8	-99
17	5	HHM	F	Afr	0	N	0.85	5	207
18	5	HHM	F	Afr	0	N	0.85	6	213
19	5	HHM	F	Afr	0	N	0.85	7	212
20	5	HHM	F	Afr	0	N	0.85	8	213
21	6	HHM	M	Afr	0	N	0.90	5	191
22	6	HHM	M	Afr	0	N	0.90	6	189

Dynamic Predictor

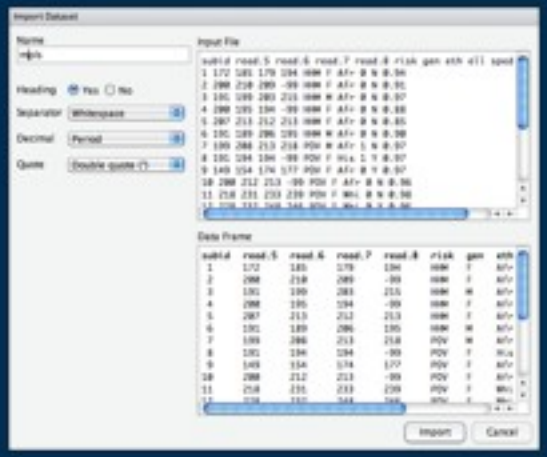
	subid	risk	gen	eth	ell	sped	att	grade	read
1	1	HHM	F	Afr	0	N	0.94	5	172
2	1	HHM	F	Afr	0	N	0.94	6	185
3	1	HHM	F	Afr	0	N	0.94	7	179
4	1	HHM	F	Afr	0	N	0.94	8	194
5	2	HHM	F	Afr	0	N	0.91	5	200
6	2	HHM	F	Afr	0	N	0.91	6	210
7	2	HHM	F	Afr	0	N	0.91	7	209
8	2	HHM	F	Afr	0	N	0.91	8	-99
9	3	HHM	M	Afr	0	N	0.97	5	191
10	3	HHM	M	Afr	0	N	0.97	6	199
11	3	HHM	M	Afr	0	N	0.97	7	203
12	3	HHM	M	Afr	0	N	0.97	8	215
13	4	HHM	F	Afr	0	N	0.88	5	200
14	4	HHM	F	Afr	0	N	0.88	6	195
15	4	HHM	F	Afr	0	N	0.88	7	194
16	4	HHM	F	Afr	0	N	0.88	8	-99
17	5	HHM	F	Afr	0	N	0.85	5	207
18	5	HHM	F	Afr	0	N	0.85	6	213
19	5	HHM	F	Afr	0	N	0.85	7	212
20	5	HHM	F	Afr	0	N	0.85	8	213
21	6	HHM	M	Afr	0	N	0.90	5	191
22	6	HHM	M	Afr	0	N	0.90	6	189

Static Predictor(s)

	subid	risk	gen	eth	ell	sped	att	grade	read
1	1	HHM	F	Afr	0	N	0.94	5	172
2	1	HHM	F	Afr	0	N	0.94	6	185
3	1	HHM	F	Afr	0	N	0.94	7	179
4	1	HHM	F	Afr	0	N	0.94	8	194
5	2	HHM	F	Afr	0	N	0.91	5	200
6	2	HHM	F	Afr	0	N	0.91	6	210
7	2	HHM	F	Afr	0	N	0.91	7	209
8	2	HHM	F	Afr	0	N	0.91	8	-99
9	3	HHM	M	Afr	0	N	0.97	5	191
10	3	HHM	M	Afr	0	N	0.97	6	199
11	3	HHM	M	Afr	0	N	0.97	7	203
12	3	HHM	M	Afr	0	N	0.97	8	215
13	4	HHM	F	Afr	0	N	0.88	5	200
14	4	HHM	F	Afr	0	N	0.88	6	195
15	4	HHM	F	Afr	0	N	0.88	7	194
16	4	HHM	F	Afr	0	N	0.88	8	-99
17	5	HHM	F	Afr	0	N	0.85	5	207
18	5	HHM	F	Afr	0	N	0.85	6	213
19	5	HHM	F	Afr	0	N	0.85	7	212
20	5	HHM	F	Afr	0	N	0.85	8	213
21	6	HHM	M	Afr	0	N	0.90	5	191
22	6	HHM	M	Afr	0	N	0.90	6	189

Using RStudio to Read in Data

- Click Import Dataset (workspace)




```
> head(mpls)
> str(mpls)
> summary(mpls)
```

Missing Data

- summary() suggests problems
- Codebook indicates –99 is missing
- R treats –99 as a value
- –99 needs to be replaced with NA

```
> mpls == -99
```

Note double equal (==) sign

```
> mpls[mpls == -99] <- NA
> head(mpls)
> summary(mpls)
```

Means and Variances

- `mean()` for mean
- `sd()` for standard deviation
- If there are NAs, then the argument `na.rm=TRUE` needs to be added

```
> mean( mpls[ , 2] )
> mean( mpls[ , 3] )
> mean( mpls[ , 4] )
> mean( mpls[ , 5] )
```

Error

```
> mean( mpls[ , 5], na.rm = TRUE )
```

```
> mean( mpls[ , 2:5], na.rm = TRUE )
```

Faster

```
> apply( mpls[ , 2:5], 2, mean, na.rm = TRUE )

  read.5   read.6   read.7   read.8
205.1364 211.4545 215.6818 218.0000
```

means increase over time...

but, difference between first two means is greater than difference between last two means...

indicating deceleration in growth

```
> apply( mpls[ , 2:5], 2, sd, na.rm = TRUE )

  read.5   read.6   read.7   read.8
19.99356 20.06116 19.44562 19.37881
```

standard deviations show very slight decrease over time

Correlations

- `cor()` for correlation
- If there are NAs, then the argument `use="complete.obs"` needs to be added (listwise deletion)

```
> cor( mpls[ , 2:5], use = "complete.obs" )

          read.5    read.6    read.7    read.8
read.5 1.0000000 0.9756825 0.9283549 0.8825860
read.6 0.9756825 1.0000000 0.9130901 0.9287729
read.7 0.9283549 0.9130901 1.0000000 0.9227732
read.8 0.8825860 0.9287729 0.9227732 1.0000000
```

decay pattern with more closely spaced time points having higher correlation than time points which are farther away

Conditioning on Static Predictors

- `tapply()` for conditioning

```
> tapply( mpls[ , 2], mpls[ , 6], mean, na.rm = TRUE )

  ADV   HHM   POV
216.7 193.5 197.5
```

```
> tapply( mpls[ , 2], mpls[ , 11], mean, na.rm = TRUE )
```

too many conditioning values for descriptive work...

Splitting Variables

- Only for descriptive work...not inference
- Can use `ifelse()` function for dichotomies
- Use `recode()` from the **car** package for polychotomies
- Use `cut()` from the **car** package for splitting quantitative variables

```
> median( mpls[ , 11] )

[1] 0.97
```

if att is ≤ 0.97 then 0
if att is > 0.97 then 1

```
> ifelse( mpls[ , 11] <= 0.97, 1, 0)

> mpls$att2 <- ifelse( mpls[ , 11] <= 0.97, 1, 0)

> tapply( mpls[ , 2], mpls[ , 12], mean, na.rm = TRUE )

    0     1
217.0 199.6
```

if att ≤ 0.95 then 1
if 0.95 < att ≤ 0.97 then 2
if 0.97 < att ≤ 0.98 then 3
if 0.98 < att ≤ 1 then 4

```
> library( car )
```

```
> cut( mpls[ ,11], c(0, 0.95, 0.97, 0.98, 1) )
```

```
> tapply( mpls[ , 2], cut( mpls[ , 11], c(0, 0.95, 0.97, 0.98, 1) ),
    mean, na.rm = TRUE )

   (0,0.95] (0.95,0.97] (0.97,0.98]    (0.98,1]
   195.6667    202.2222    218.6667    215.7500
```

Reshaping Data

- Wide format to long format
- Use the reshape() function

`data = mpls`	name of data frame
`idvar = "subid"`	subject ID (character string)
`varying = 2:5`	columns with response measures
`v.names = "read"`	new name for response
`times = 5:8`	columns with time predictors
`timevar = "grade"`	new name for time predictor
`direction = "long"`	direction for new data frame

```
> mpls.l <- reshape( data = mpls, idvar = "subid", varying = 2:5,
    v.names = "read", times = 5:8, timevar = "grade",
    direction = "long" )


> head( mpls.l, n = 10 )

      subid risk gen eth ell sped  att att2 grade read
1.5       1  HHM   F Afr   0    N 0.94    1     5  172
2.5       2  HHM   F Afr   0    N 0.91    1     5  200
3.5       3  HHM   M Afr   0    N 0.97    1     5  191
4.5       4  HHM   F Afr   0    N 0.88    1     5  200
5.5       5  HHM   F Afr   0    N 0.85    1     5  207
6.5       6  HHM   M Afr   0    N 0.90    1     5  191
7.5       7  POV   M Afr   1    N 0.97    1     5  199
8.5       8  POV   F His   1    Y 0.97    1     5  191
9.5       9  POV   F Afr   0    Y 0.97    1     5  149
10.5     10  POV   F Afr   0    N 0.96    1     5  200
```

Sort by Subject ID

- Use the `arrange()` function from the **plyr** library

```
> library( plyr )

> arrange( mpls.l, subid )


> mpls.l <- arrange( mpls.l, subid )

> head( mpls.l, n = 10 )

   subid risk gen eth ell sped  att att2 grade read
1      1  HHN   F Afr   0    N 0.94    1     5  172
2      1  HHN   F Afr   0    N 0.94    1     6  185
3      1  HHN   F Afr   0    N 0.94    1     7  179
4      1  HHN   F Afr   0    N 0.94    1     8  194
5      2  HHN   F Afr   0    N 0.91    1     5  200
6      2  HHN   F Afr   0    N 0.91    1     6  210
7      2  HHN   F Afr   0    N 0.91    1     7  209
8      2  HHN   F Afr   0    N 0.91    1     8   NA
9      3  HHN   M Afr   0    N 0.97    1     5  191
10     3  HHN   M Afr   0    N 0.97    1     6  199
```

Missing Data in LMER

- Missing data will be ignored
- Any row in long format having NA will be omitted

Subject with no missing data

All rows included

	subid	risk	gen	eth	ell	sped	att	att2	grade	read
1	1	HHN	F	Afr	0	N	0.94	1	5	172
2	1	HHN	F	Afr	0	N	0.94	1	6	185
3	1	HHN	F	Afr	0	N	0.94	1	7	179
4	1	HHN	F	Afr	0	N	0.94	1	8	194

Subject with missing response

Some rows included

	subid	risk	gen	eth	ell	sped	att	att2	grade	read
5	2	HHN	F	Afr	0	N	0.91	1	5	200
6	2	HHN	F	Afr	0	N	0.91	1	6	210
7	2	HHN	F	Afr	0	N	0.91	1	7	209
8	2	HHN	F	Afr	0	N	0.91	1	8	NA

Subject with missing covariate

No rows included

	subid	risk	gen	eth	ell	sped	att	att2	grade	read
1	1	HHN	NA	Afr	0	N	0.94	1	5	172
2	1	HHN	NA	Afr	0	N	0.94	1	6	185
3	1	HHN	NA	Afr	0	N	0.94	1	7	179
4	1	HHN	NA	Afr	0	N	0.94	1	8	194

Not a problem if gender is not included in analysis

Missing Data in LMER

- Time will be unbalanced (not every subject has same number of rows)
- If subjects are missing data on covariates number of subjects will vary across different analyses
- Having a different sample sizes makes comparisons difficult (maybe even invalid)
- Use `na.omit()` to remove rows with NAs

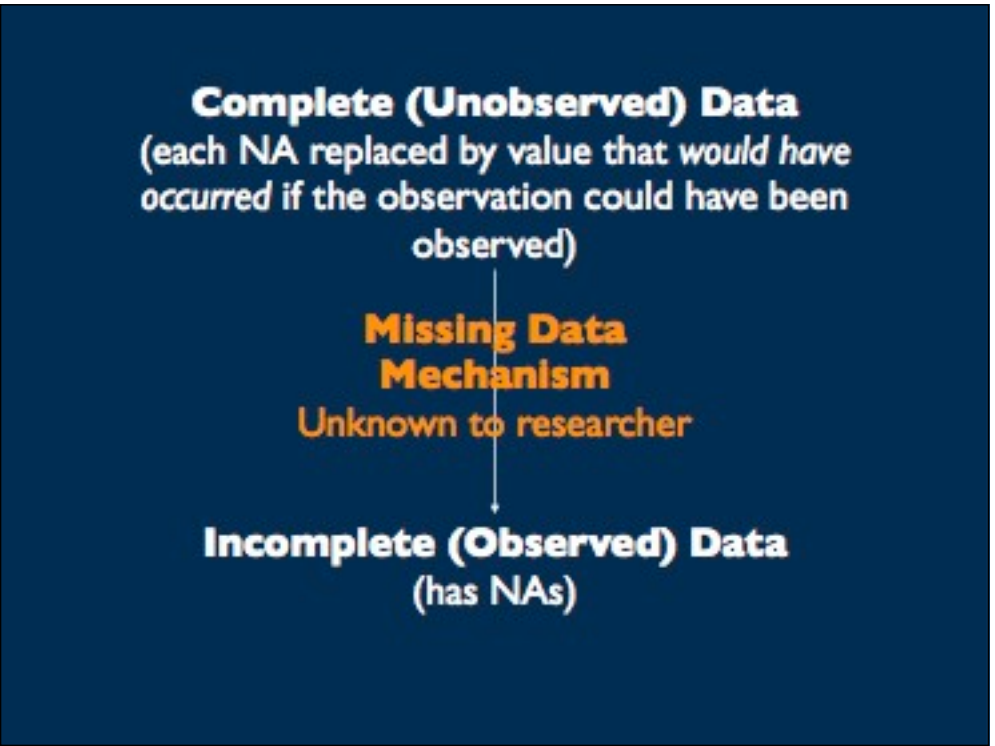
```
> mpls.l <- na.omit( mpls.l )

> head( mpls.l, n = 10 )
```

	subid	risk	gen	eth	ell	sped	att	att2	grade	read
1	1	HHN	F	Afr	0	N	0.94	1	5	172
2	1	HHN	F	Afr	0	N	0.94	1	6	185
3	1	HHN	F	Afr	0	N	0.94	1	7	179
4	1	HHN	F	Afr	0	N	0.94	1	8	194
5	2	HHN	F	Afr	0	N	0.91	1	5	200
6	2	HHN	F	Afr	0	N	0.91	1	6	210
7	2	HHN	F	Afr	0	N	0.91	1	7	209
9	3	HHN	M	Afr	0	N	0.97	1	5	191
10	3	HHN	M	Afr	0	N	0.97	1	6	199

Missing Data Mechanism

- Justification to omit rows with NA is based on assumptions about *missing data mechanism* (process responsible for missing data)
- Three processes considered
  - Missing completely at random (MCAR)
  - Missing at random (MAR)
  - Not missing at random (NMAR)



MCAR

- Incomplete observed data is a random sample of the complete unobserved data
- NA values are randomly assigned
- Missingness due to sickness can be considered random

MAR

- Incomplete observed data is a *conditional* random sample of the complete unobserved data
- NA values are randomly assigned conditional on some attribute
- Missingness due to sickness is higher for families with higher levels of poverty. However, within levels of poverty, illness can be considered random

NMAR

- Incomplete observed data is *not* a random sample of the complete unobserved data
- NA values are not random in any way
- Missingness due to students not tested because students skipped (self-selection)

Results on LMER

- Results when missing data is MCAR are valid (generally).
- Results when missing data is MAR are valid (generally) when the attribute that missingness is conditioned on is included in the analysis.
- Results when missing data is NMAR are not valid (generally). The results will be biased to an unknown extent.