Longitudinal Data Exploration and Visualization

Introduction

- Data on individuals followed over time with information collected at several time points.
- Clusters are the individuals who are followed over time.
- Repeated observations may or may not be taken at regular times (balanced, fixed occasions, do not differ between subjects).
- Our interest is in the change from baseline.

Datasets used in this course:

- Example data is taken from (Mallinckrodt and Lipkovich 2016). The authors generated data sets based on two nearly identically designed antidepressant clinical trials by randomly selecting subjects from the original data.
- Contain data on the continuous variable HAMD17 (Hamilton 17-item rating scale for depression).
- Two treatement arms are included: placebo (arm 1) vs. drug (arm 2).
- Assessments were taken at baseline and weeks 1, 2, 4, 6, and 8.

There are 3 data sets created from the original data:

- Data *all2* = Subsample of the large dataset with n=50, visits: weeks 2, 4, 8.
- Data *high2* = Large dataset with n=100, high dropout = 70% (drug), 60% (placebo).
- Data *low2* = Large dataset with n=100, low dropout = 18%.

We are mainly working with the *all2* data set in the following. There is one application on the *high2* data set. We are not considering the *low2* data set.

Data set all2

- Small data set with n=50 subjects.
- 1st version: complete data where all subjects adhered to the originally assigend study medication, variable *change*

- 2nd version = missing data: identical to the first except some data were missing (drop-out), variable *chgdrop*

Looking at the variables in the data set

```
head(all2)
```

```
# A tibble: 6 x 14
  subject  time chgdrop trt   basval change pgiimp gender chgrescue dropout_grp
  <fct>   <dbl>   <dbl> <chr>  <dbl>  <dbl>  <dbl> <chr>      <dbl> <chr>
1 1           1     -11 2         24    -11      3 F            -11 Week 2 Drop~
2 1           2      NA 2         24    -16      2 F            -26 Week 2 Drop~
3 1           3      NA 2         24    -24      2 F            -34 Week 2 Drop~
4 2           1      -6 1         20     -6      4 F             -6 Week 2 Drop~
5 2           2      NA 1         20     -8      4 F            -18 Week 2 Drop~
6 2           3      NA 1         20     -5      5 F            -15 Week 2 Drop~
# i 4 more variables: aval <dbl>, avisit <fct>, week <dbl>, group <fct>
```

Task 1 - Exploration of data set all2 - 15 minutes working time

Only consider the complete data, variable *change*

- Are the data balanced and equally spaced?
- Number of observations by week? - Summary statistics for HAMD17 (change from baseline) by week.
- Plot trajectories for each individual, different colors for each treatment group (or panels).
- Add mean to your plot or generate new plot with mean change from baseline by treatment group.
- Plot mean change from baseline for each treatment group stratified by sex. Comment on the plot.

Task 1 - Discussion, possible solution

Table: Summary statistics mean (SD) for HAMD17 by treatment and week in the all2 data set

```
all2 %>%
  select(change, group, avisit) %>%
  tbl_strata(strata=group,
             ~.x %>%
               tbl_summary(by = avisit,
```

```
                            statistic = list(
      all_continuous() ~ "{mean} ({sd})"),
      digits = all_continuous() ~ 2 ) %>%
      modify_header(label = "**Variable**")
)
```

Variable	**Week 2**, N = 25	**Week 4**, N = 25	**Week 8**, N = 25	**Week 2**, N = 25	**Week 4**, N = 25	**Week 8**, N = 25
change	-4.20 (3.66)	-6.80 (4.25)	-9.88 (4.85)	-5.24 (5.49)	-8.60 (5.39)	-13.24 (5.54)

Figure: individual trajectories stratified by treatment group

```
ggplot(data = all2, aes(x = week, y = change, group=subject)) +
  geom_point() + geom_line() + facet_grid(.~group) + ylab("Change from baseline HAMD17") +
  scale_x_continuous(name="Visit [week]", breaks=c(2,4,8))
```

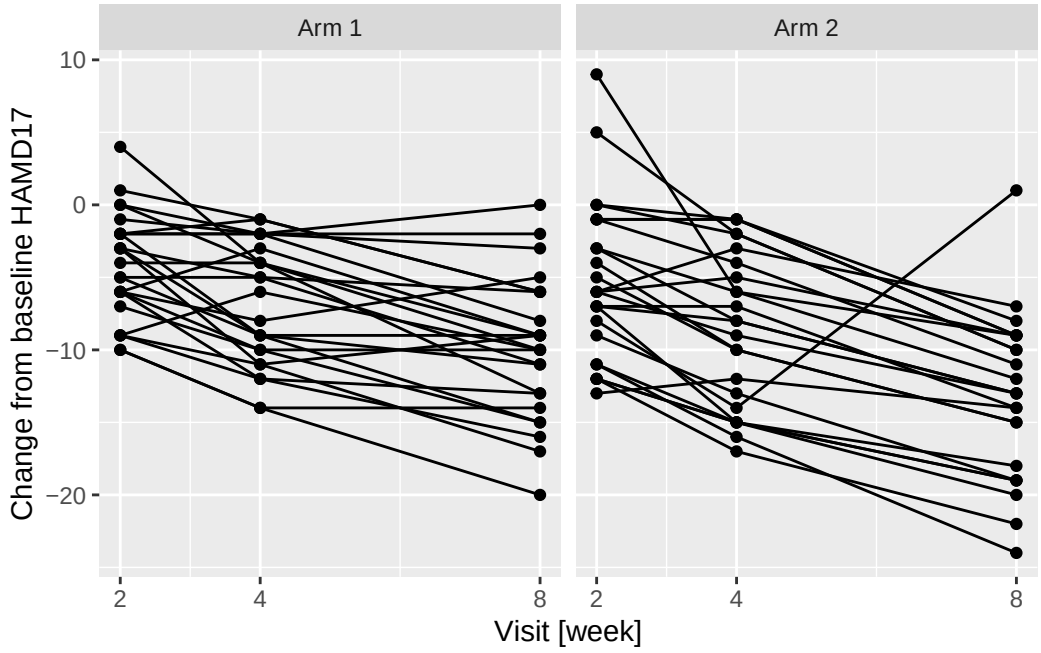

Figure 1: Individual trajectories of HAMD17 by treatment group

Figure: Mean change from baseline for each treatment group

```
ggplot(data = all2, aes(x = week, y = change)) +
  geom_point(aes(colour=factor(group))) + ylab("Change from baseline HAMD17") +
  scale_x_continuous(name="Visit [week]", breaks=c(2,4,8)) +
  stat_summary(aes(group = group, colour=factor(group)), geom = "line", fun.y = mean,
               size = 1) +
  stat_summary(aes(group = group, colour=factor(group)), geom = "point", fun.y = mean,
               shape=17,size = 2)
```

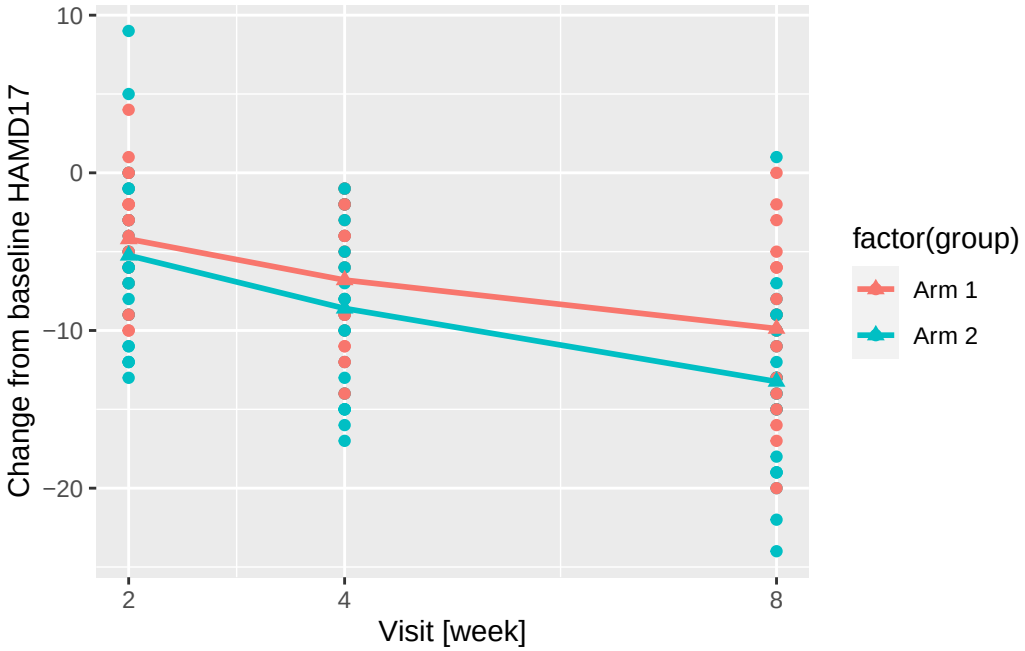

Figure 2: Mean HAMD17 change from baseline by treatment group

Frequency for sex per treatment group

```
all2 %>% filter(time==1) %>%
  tbl_summary(
    include = c(gender),
    by = group
  )
```

Characteristic	**Arm 1**, N = 25	**Arm 2**, N = 25
PATIENT SEX		
F	10 (40%)	19 (76%)

Characteristic	Arm 1, N = 25	Arm 2, N = 25
M	15 (60%)	6 (24%)

Figure: Mean change from baseline stratified by sex

```
ggplot(data = all2, aes(x = week, y = change)) +  facet_grid(.~gender) +
  geom_point(aes(colour=factor(group))) + ylab("Change from baseline HAMD17") +
  scale_x_continuous(name="Visit [week]", breaks=c(2,4,8)) +
  stat_summary(aes(group = group, colour=factor(group)), geom = "line", fun.y = mean,
               size = 1) +
  stat_summary(aes(group = group, colour=factor(group)), geom = "point", fun.y = mean,
               shape=17,size = 2)
```

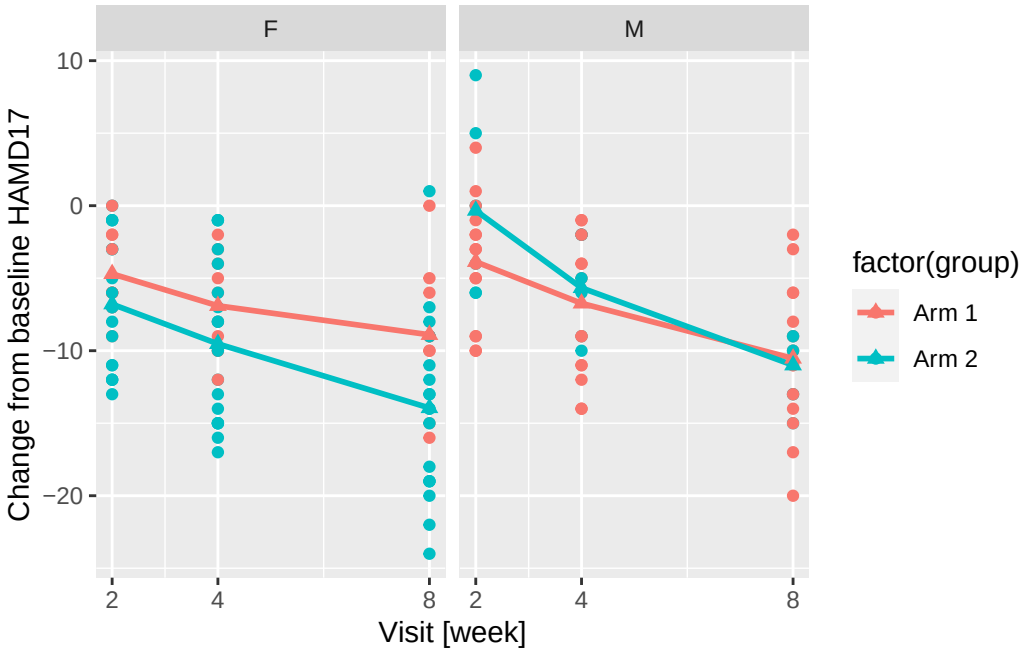

Figure 3: Mean HAMD17 change from baseline by treatment group stratified by sex

Data set all2 with drop-out

- 2nd version = missing data: identical to the first except some data were missing (drop-out), variable *chgdrop*
- This version is later relevant when considering missing data. Thus, have a short look at the data.

Table: Summary statistics for HAMD17 by treatment and week in the all2 data set with drop-outs

```
all2 %>%
  select(chgdrop, group, avisit) %>%
  tbl_strata(strata=group,
             ~.x %>%
               tbl_summary(by = avisit,
                           statistic = list(
      all_continuous() ~ "{mean} ({sd})"),
      digits = all_continuous() ~ 2 ) %>%
      modify_header(label = "**Variable**")
)
```

Variable	**Week 2**, N = 25	**Week 4**, N = 25	**Week 8**, N = 25	**Week 2**, N = 25	**Week 4**, N = 25	**Week 8**, N = 25
chgdrop	-4.20 (3.66)	-6.80 (4.63)	-10.17 (4.88)	-5.24 (5.49)	-8.14 (5.27)	-13.11 (5.44)
Unknown	0	5	7	0	3	6

Figure: Mean change from baseline for each treatment group in the all2 data set with drop-outs

```
ggplot(data = all2, aes(x = week, y = chgdrop)) +
  geom_point(aes(colour=factor(group))) + ylab("Change from baseline HAMD17") +
  scale_x_continuous(name="Visit [week]", breaks=c(2,4,8)) +
  stat_summary(aes(group = group, colour=factor(group)), geom = "line", fun.y = mean,
               size = 1) +
  stat_summary(aes(group = group, colour=factor(group)), geom = "point", fun.y = mean,
               shape=17,size = 2)
```

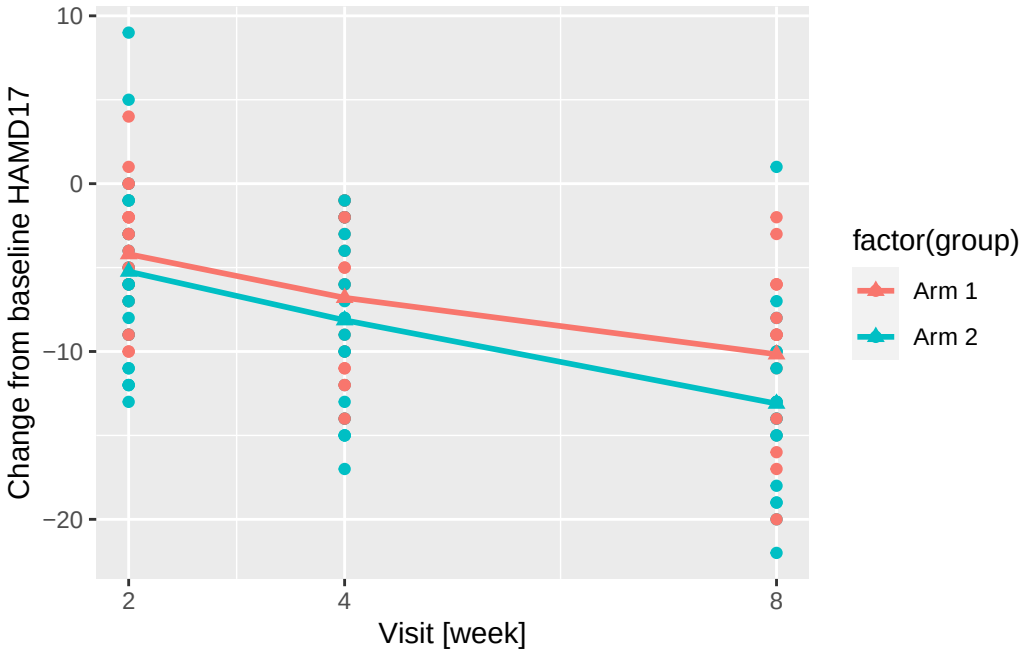

Figure 4: Mean HAMD17 change from baseline by treatment group

Data set high2

- Large data set with n=100 subjects.
- Note that we have no intermittent missing values but drop-outs.

Looking at the variables in the data set.

```
head(high2)
```

```
# A tibble: 6 x 16
# Groups:   patient [2]
  patient trt   poolinv basval  week change pgiimp   age gender  drop .groups
    <dbl> <chr> <chr>    <dbl> <dbl>  <dbl>  <dbl> <dbl> <chr>  <dbl> <chr>
1    1401 1     005         19     1     -7      3  44.5 F          2 drop
2    1401 1     005         19     2     -4      3  44.5 F          2 drop
3    1411 2     005         17     1      0      3  35.7 F          8 drop
4    1411 2     005         17     2     -2      3  35.7 F          8 drop
5    1411 2     005         17     4      2      3  35.7 F          8 drop
6    1411 2     005         17     6     -3      2  35.7 F          8 drop
```

```
# i 5 more variables: aval <dbl>, group <fct>, avisit <fct>, dropout_grp <fct>,
#   subject <fct>
```

Task 2 - Exploration of data set high2 - 15 minutes working time

- Explore the drop-outs e.g. number of observations by week.
- Summary statistics for HAMD17 change.
- Generate and interpret the group-wise boxplots of the change from baseline.
- Mean change from baseline for different drop-out groups (by treatment). Comment on the plot.

Task 2 Discussion, possible solution

Table: Summary statistics for HAMD17 by treatment and week in the high2 data set

```
high2 %>% ungroup() %>%
  select(change, group, avisit) %>%
  tbl_strata(strata=group,
             ~.x %>%
               tbl_summary(by = avisit,
                           statistic = list(
                             all_continuous() ~ "{mean} ({sd})"),
                           digits = all_continuous() ~ 2 ) %>%
              modify_header(label = "**Variable**")
  )
```

Variable	**Week 1**, N = 100	**Week 2**, N = 92	**Week 4**, N = 85	**Week 6**, N = 73	**Week 8**, N = 60	**Week 1**, N = 100	**Week 2**, N = 90	**Week 4**, N = 85	**Week 6**, N = 75	**Week 8**, N = 70
change	-1.49 (3.91)	-3.16 (5.69)	-4.51 (6.23)	-5.51 (6.16)	-6.58 (5.99)	-1.84 (5.58)	-4.30 (6.82)	-6.47 (6.84)	-8.29 (6.96)	-8.99 (7.04)

Figure: Distribution of HAMD17 change from baseline

```
ggplot(data = high2, aes(x = avisit, y = change, fill=group)) +
  geom_boxplot() + ylab("Change from baseline HAMD17") + xlab("Visit")
```

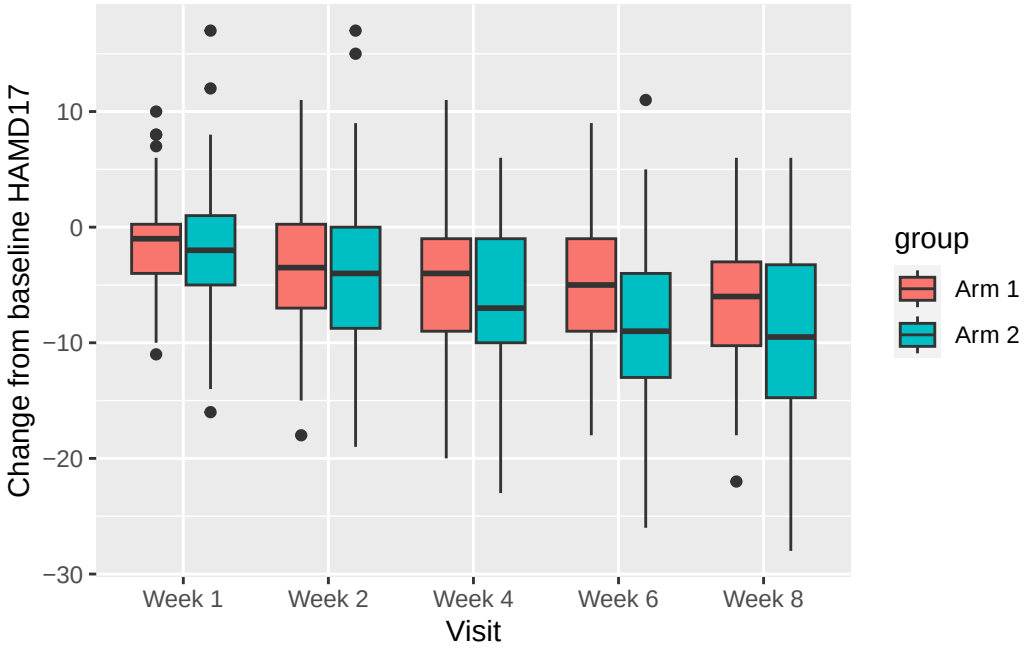

Figure 5: Distribution of HAMD17 change from baseline by treatment group at each visit

Figure: Mean HAMD17 changes by drop-out group

```
ggplot(data = high2, aes(x = week, y = change, group=patient)) +
  geom_point(col="lightgray") + geom_line(col="lightgray") + facet_grid(.~group) +
  ylab("Change from baseline HAMD17") + scale_x_continuous(name="Visit [week]", breaks=c(1
  stat_summary(aes(group = dropout_grp, colour=factor(dropout_grp)), geom = "line", fun.y
               size = 1) +
  stat_summary(aes(group = dropout_grp, colour=factor(dropout_grp)), geom = "point", fun.y
               shape=17,size = 2)
```

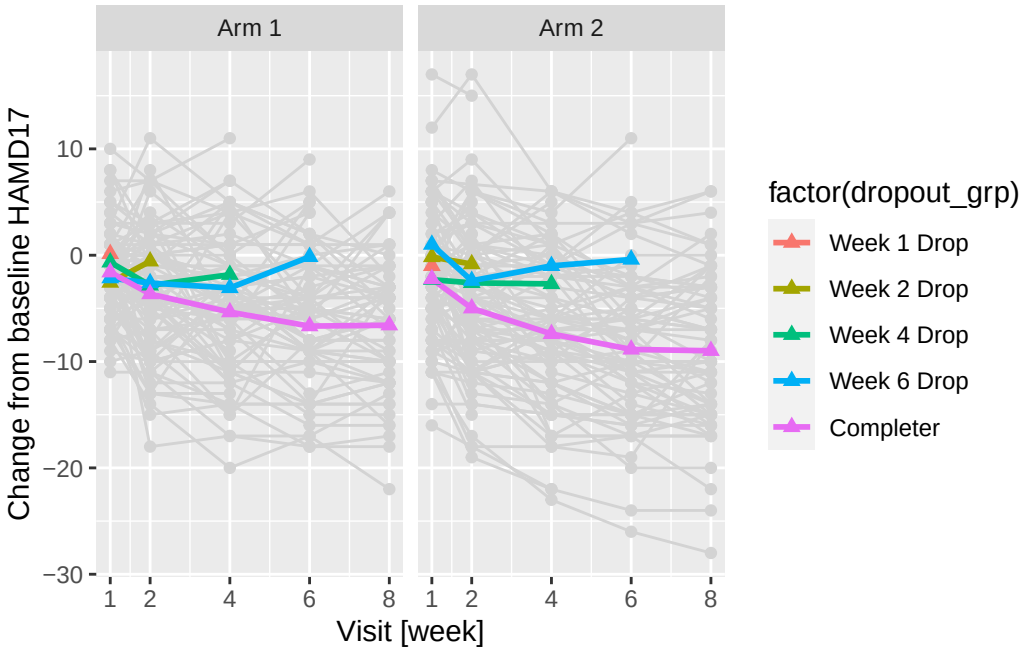

Figure 6: Visit-wise mean HAMD17 changes from baseline by treatment group and drop-out

Correlation structure, covariance matrices

- Longitudinal data allows to exploit the correlation between outcomes within subjects regardless of whether or not focus is on a single landmark time point.
- Model within-subject error correlation
- Different residual covariance structures can be implemented

Overview - different covariance matrices

- Variance components (VC) independence structure
- Compound symmetry (CS) also known as exchangeable
- Toeplitz (TOEP)
- First order auto regressive (AR(1))
- Unstructured (UN)

Selected covariance structures for data with three assessment times (t=3) are shown below. Note that with three assessment times, the number of parameters estimated for the various structures did not differ as much as would be the case with more assessment times. Thus, results from different covariance structures are more similar than would be the case with more assessment times.

Independence structure (VC)

Constant variance. It is assumed to be no correlation between assessments (residuals are independent across time).

$$R = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}$$

Compound symmetry (CS)

Constant variance and constant covariance across all assessments. Also known as exchangeable. It requires two parameter estimates. Most simplest repeated measures (i.e., correlated errors) structure.

$$R = \begin{bmatrix} \sigma^2 + \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma^2 + \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma^2 + \sigma_1 \end{bmatrix}$$

Unstructured (UN)

This is the most general (saturated) model. It has t + [t(t-1)/2] parameters to be estimated. Here it is 3 + 3 = 6 parameters.

$$R = \begin{bmatrix} \sigma_1^2 & \sigma_{21} & \sigma_{31} \\ \sigma_{21} & \sigma_2^2 & \sigma_{32} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix}$$

Toeplitz structure (TOEP)

Homogenous variances and heterogenous correlations. Same correlation value is used whenever the degree of adjacency is the same e.g. correlation between times 1 and 2 = correlation between times 2 and 3. Repeated measurements are assumed to be equally spaced. TOEP requries t parameter estimates so here we have t=3 parameter.

$$R = \begin{bmatrix} \sigma^2 & \sigma_1^2 & \sigma_2^2 \\ \sigma_1^2 & \sigma^2 & \sigma_1^2 \\ \sigma_2^2 & \sigma_1^2 & \sigma^2 \end{bmatrix}$$

Autoregressive structure (AR(1))

Correlation decreases as time between observations increases. Assumtpion of equal spacing between each repeated measurement must be reasonably applicable. This structure requires the estimation of two parameters.

$$R = \begin{bmatrix} \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho \\ \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 \end{bmatrix}$$

Spatial Power (SP)

Spatial covariance structures does not require equal spacing between measurements. Instead, as long as the distance between visits can be quantified in terms of time and/or other coordinates, the spatial covariance structure can be applied. Covariances are mathematical functions of Euclidean distances between observed measurements. Again, two parameters need to be estimated.

For spatial exponential, the covariance structure is defined as follows:

$$R = \begin{bmatrix} \sigma^2 & \sigma^2\rho_{12} & \sigma^2\rho_{13} \\ \sigma^2\rho_{21} & \sigma^2 & \sigma^2\rho_{23} \\ \sigma^2\rho_{31} & \sigma^2\rho_{32} & \sigma^2 \end{bmatrix}$$

with

$$\rho_{ij} = \rho^{d_{ij}}$$

where

$$d_{ij}$$

is the distance between time point i and time point j e.g. distance in weeks.

Selecting the covariance structure

There are a variety of considerations when selecting the covariance structure:

- number of parameters
- interpretation of the structure
- model fit

UN is the most flexible (complex) structure and can fail to run especially if one has many repeated measures. Choose a reasonable covaraiance structure which is the best compromise between model fit and complexity. E.g. use AIC as it penalises more complex models.

Task 3 - Exploration of correlation in the data

- Compute the empirical correlations between measurement timepoints in the all2 data set (e.g. correlation between baseline and post-baseline changes, variable *change*).
- Looking at these correlations + using your knowledge of the experiment (e.g., spacing of measurements), comment on the suitability of the correlation structures VC, CS, UN, AR(1).

Task 3 - Discussion and possible solution

Table: Correlation and covariance matrix

```
all2.w <- all2 %>%
  pivot_wider(id_cols=subject,names_from = time, values_from = c(basval,change)) %>%
  select(-c(basval_2,basval_3))

cor(all2.w[-1])
```

```
           basval_1   change_1   change_2    change_3
basval_1  1.00000000 -0.2636447 -0.3165711 -0.02915138
change_1 -0.26364471  1.0000000  0.7557078  0.51502724
change_2 -0.31657106  0.7557078  1.0000000  0.71298768
change_3 -0.02915138  0.5150272  0.7129877  1.00000000
```

```
cov(all2.w[-1])
```

```
          basval_1  change_1  change_2   change_3
basval_1 16.3330612 -4.955918 -6.253061 -0.6391837
change_1 -4.9559184 21.634286 17.179592 12.9967347
change_2 -6.2530612 17.179592 23.887755 18.9061224
change_3 -0.6391837 12.996735 18.906122 29.4351020
```

Taking a step back: Consequences of Ignoring Correlation among Longitudinal Data

This technical detour is motivated by (Fitzmaurice 2011). Let us assume we are only interested in the first two responses in a clinical study, say Visit 1 (Baseline) and Visit 2. Our interest lies in an assessment of mean changes over time (for the sake of simplicity in a single treatment group only), i.e. we wish to estimate

$$\hat{\delta} := \hat{\mu}_2 - \hat{\mu}_1 = \frac{1}{N}\sum_{i=1}^{N}(Y_{i2} - Y_{i1}),$$

where Y_{i1} and Y_{i2} are observations from subject i at Visit 1 and Visit 2, respectively. To obtain the standard error (SE) and get a notion of variability, we compute the variance of $\hat{\delta}$ and see that

$$\text{Var}(\hat{\delta}) = \text{Var}\left(\frac{1}{N}\sum_{i=1}^{N}(Y_{i2} - Y_{i1})\right) = \frac{1}{N}(\sigma_1^2 + \sigma_2^2 - 2\sigma_{12})\,.$$

The inclusion of the term $-2\sigma_{12}$ accounts for the correlation between responses at Visit 1 and Visit 2. As data from adjacent visits is usually positively correlated, the omission of the correlation term leads to an overestimation of the variance and thus the SE associated with the treatment effect.

Fitzmaurice, Laird, G. M. 2011. *Applied Longitudinal Analysis*. Vol. 2. USA: New York, Wiley. https://doi.org/10.1002/9781119513469.

Mallinckrodt, Craig, and Ilya Lipkovich. 2016. *Analyzing Longitudinal Clinical Trial Data*. Chapman; Hall/CRC. https://doi.org/10.1201/9781315186634.