FINAL PROJECT

NAME: Chaitanya Rajeev

COURSE: ST 537 - Applied Multivariate and Longitudinal Data Analysis

DUE DATE: 5/1/20

Contents

1. Introduction

1.1 Background

Exposure to lead can produce a variety of adverse health effects in infants and children, including hyperactivity, hearing or memory loss, learning disabilities, and damage to the nervous system. Although the use of lead as a gasoline additive has been discontinued in the US, so that airborne lead levels have been reduced dramatically, a small percentage of children continue to be exposed to lead at levels that can produce such health problems. Much of this exposure is due to deteriorating lead-based paint that may be chipping and peeling in older homes. Lead-based paint in housing was banned in the US in 1978; however, many older homes (built pre-1978) do contain lead-based paint, and chips and dust can be ingested by young children living in these homes during normal teething and hand-to-mouth behavior. This is especially a problem among children in deteriorating, inner-city housing. The US Centers for Disease Control and Prevention (CDC) has determined that children with blood levels above 10 micrograms/deciliter (μg/dL) of whole blood are at risk of adverse health effects.

1.2 Study

Researchers were interested in evaluating the effectiveness of a lead chelating treatment Succimer, on 120 identified exposed children aged 12 (36 months with confirmed blood lead levels of *between* 15μg/dL and 40μg/dL) in a large, inner-city housing project. 3 random groups each of 40 exposed children were given the following treatments for the study:

- Placebo: an inactive agent with no lead-lowering properties
- Low dose of Succimer
- High dose of Succimer

Blood lead levels were measured at a clinic for each child at baseline (time 0), prior to initiation of the assigned treatments. Then, then assigned treatment was started, and ideally, each child was to return to the clinic at weeks 2, 4, 6, and 8. At each visit, blood lead level was measured for each child.

1.3 Dataset

The data collected from the study are presented in the form of one data record per observation. The columns of the data set are described below:

a) **id**: Child id (1,2,3…)
b) **ind.age**: Indicator of age (= 0 if <= 24 months; = 1 if > 24 months)
c) **sex**: Gender indicator (= 0 if female, = 1 if male)
d) **week**: time of visit (week 0, 2, 4, 6 or 8)
e) **blood**: Blood lead level (µg/dL)
f) **trt**: Treatment indicator (= 1 if placebo, = 2 if low dose, = 3 if higher dose)

1.4 Questions of Interest

From analyzing the dataset, this report aims to find answers to a few questions regarding the effectiveness of the treatment and the significance of other factors:

a) Does *gender* have any association with blood lead level?
b) Does *age* have any association with blood lead level?
c) Is the Succimer treatment successful in reducing blood lead levels of the subject?
d) Is there a meaningful difference in performance between a low dose and high dose of the Succimer treatment?

2. Executive Summary

Exposure to lead can produce a variety of adverse health effects in infants and children, including hyperactivity, hearing or memory loss, learning disabilities, and damage to the nervous system. In the interest of developing an effective treatment for lead exposure in children, a study was carried out to ascertain the efficacy of one such chelating lead treatment known as Succimer by dividing a group of 120 children into three groups each of 40 subjects. Each group was administered one of three treatments: Placebo, Low Dosage of Succimer and High Dosage of Succimer. The Placebo treatment involved injecting an inactive agent with no lead-lowering properties. The subjects had lead level measurements taken at 0, 2, 4, 6 and 8 weeks post treatment at the clinic.

From the analysis of data, it was found that both Succimer treatments showed a significant improvement in lowering lead levels over time as compared to the Placebo. There was however no statistical difference in lead level trend between the Low Dosage and High Dosage version of the treatment.

The sex of the subject was found to have no significant association with the trend of lead level response. The two age groups within the subjects had significantly different lead concentration levels at the beginning of the study. But both treatments reduced lead levels at equal rates for both age groups with the passage of time.

3. Methods

3.1 Exploratory Data Analysis (Part A)

The dataset consists of 4 variables age, sex, time and treatment, and the response measured is the blood lead level. There are a total of 120 subjects out of which 3 groups of 40 each are administered 3 different treatments. The response is measured at fixed times (Weeks 0, 2, 4, 6 and 8). Since responses are measured at different instances of time for every individual, the dataset is essentially longitudinal in nature. The dataset is also a balanced one as all measurements have been taken at fixed points in time. Firstly, we shall explore whether all subjects have measurements taken at every instance of time.

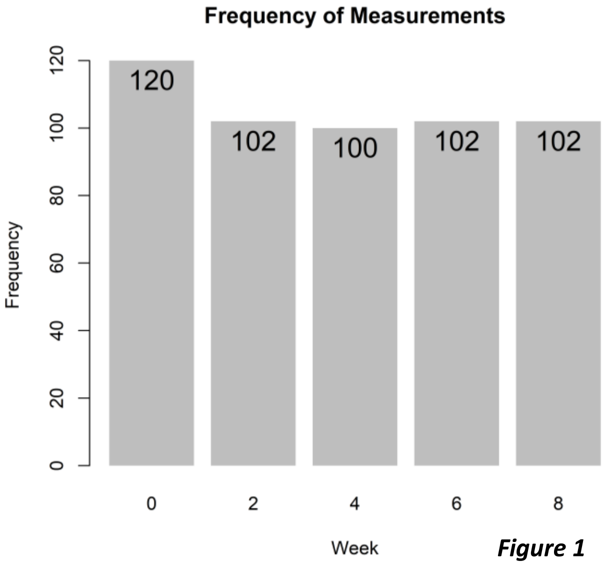


Figure 1

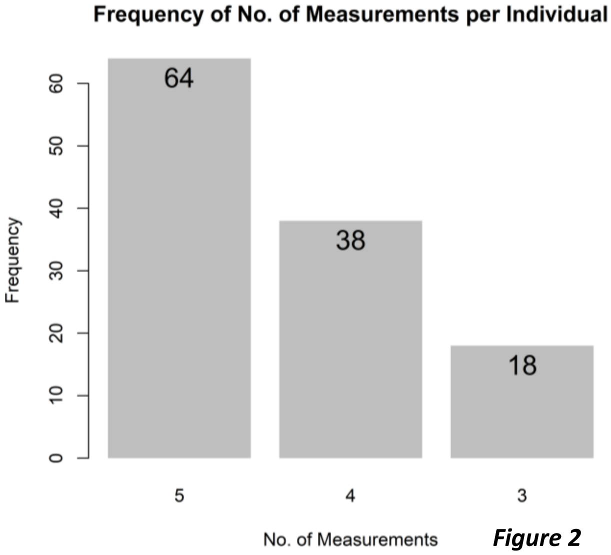


Figure 2

Figure 1 shows the frequency of measurements at Weeks 0, 2, 4, 6 and 8. From the bar chart, it is evident that apart from Week 0, all other weeks have missing time points. Figure 2 tells us that out of 120 subjects, 64 had given all measurements at all 5 time points, 38 had given measurements at 4 time points and 18 had given measurements at only 3 time points.

Since there is missing data with respect to timepoints, we shall investigate whether the missing data follows any consistent pattern or not. If there is a consistent pattern, reasons for the pattern can be further investigated.

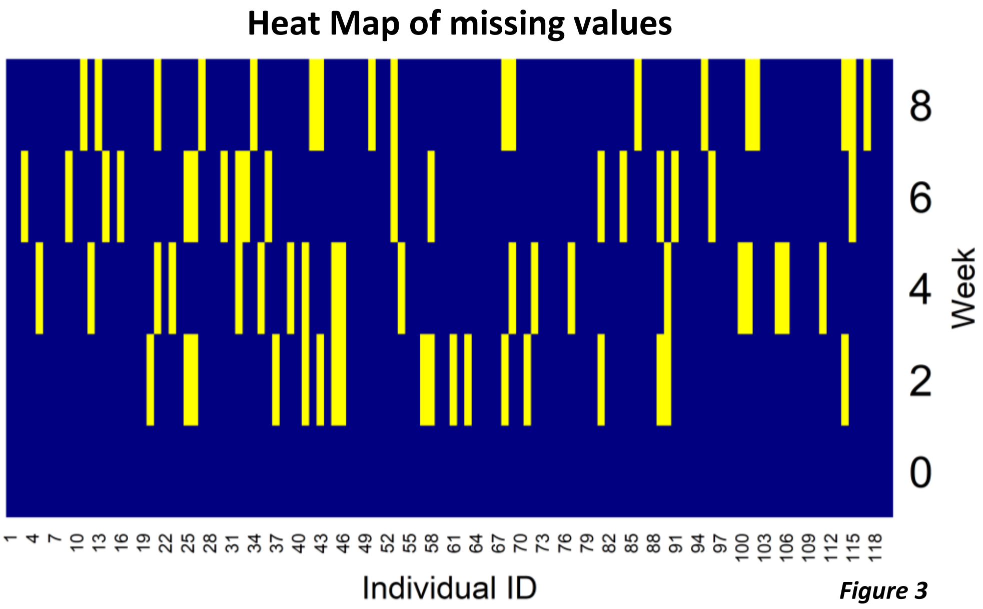


Figure 3

Figure 3 shows a heat map of the missing values. Time points are depicted on the Y-axis and individuals on the X-Axis. A yellow cell indicates a missing value whereas a blue cell indicates an existent value. From the map, no consistent pattern can be discerned. Hence, one can assume reasonably that values are missing at random.

Next, we compare the 3 different treatments by plotting their individual profile plots and then overlaying their mean trajectories in a single plot. The three treatments in question are Placebo treatment, Low Dose of Succimer treatment and a High Dose of Succimer treatment. The Placebo treatment consists of injecting an inactive agent with no lead-lowering properties. Hence one should not expect a large reduction in response with the Placebo treatment.

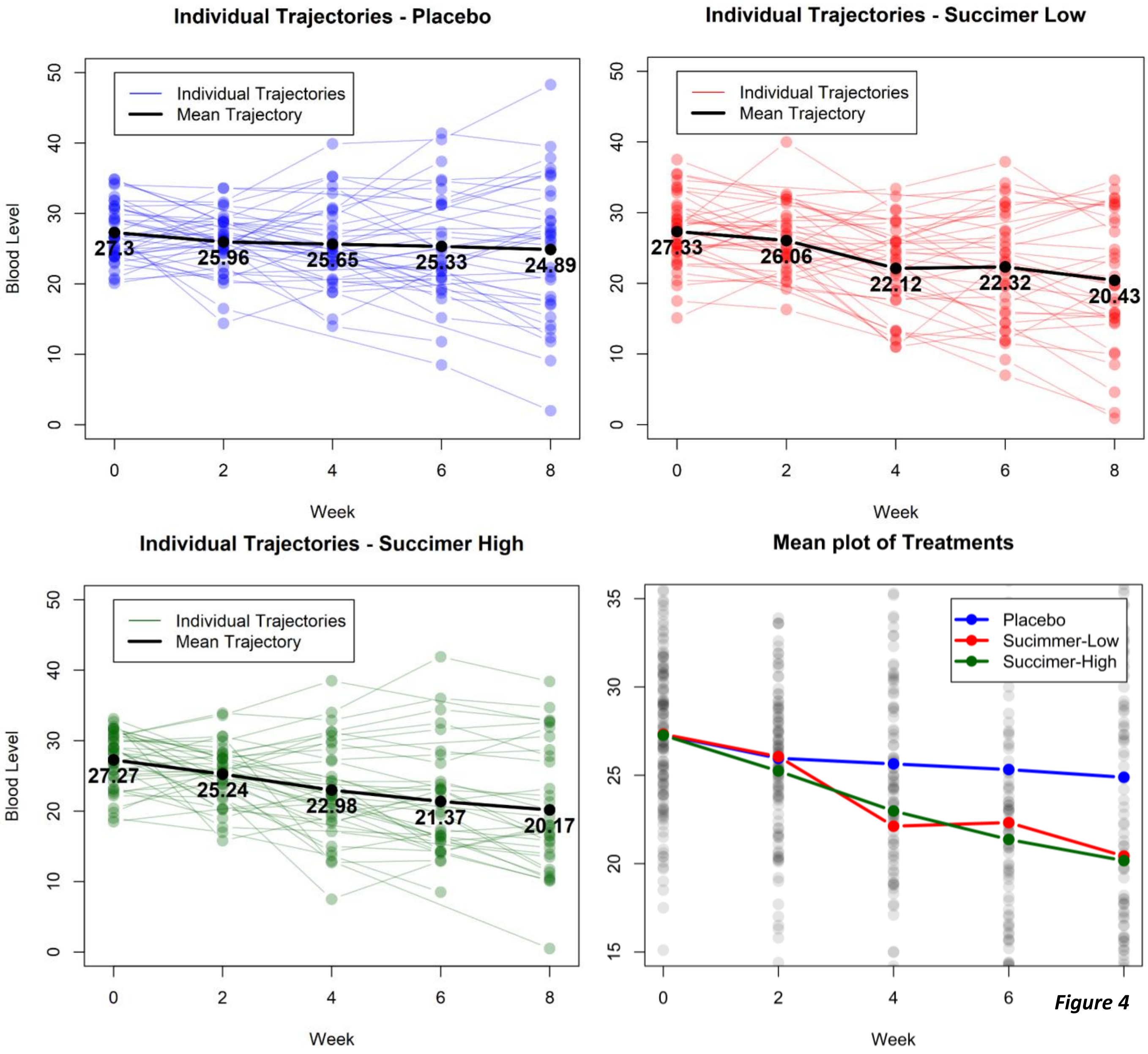


Figure 4

Figure 4 consists of four different plots. The first three depict the individual and mean profile of each of the three treatments and the fourth plot depicts the mean profile of the 3 treatments in one plot for a comparison. From the individual treatment plots, it is suggestive that both the Succimer treatments reduce mean lead levels significantly more than the Placebo treatment. At Week 8, the Placebo treatment yields a mean response of 24.89 whereas the Succimer Low and

Succimer High treatment yields responses of 20.43 and 20.17 respectively. The fourth plot also suggests that both treatments are successful in reducing lead levels better than the Placebo. But there also seems to be no significant difference between the Low Dose and High Dose treatments. This shall be investigated statistically in the later sections.

Figure 5 shows the Standard Deviation Profile of the 3 treatments. From the graph, it is evident that response variance increases with the passage of time across all three treatments.

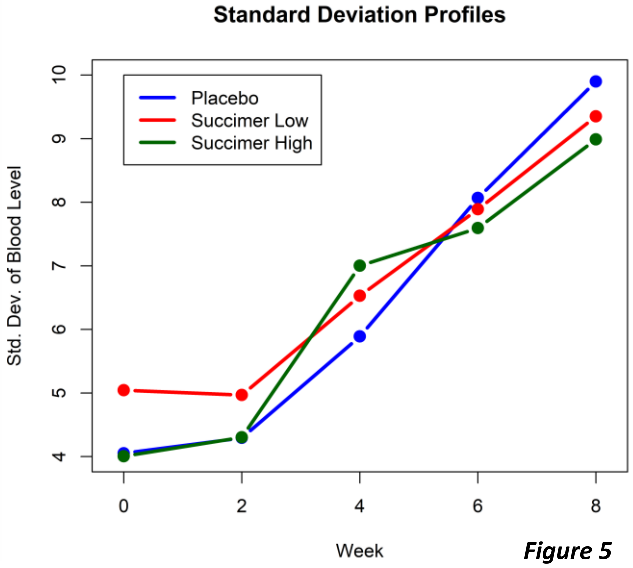


Figure 5

This maybe because some children continue with their lead-ingesting habits whereas parents of other children might be actively monitoring child habits after becoming aware of the harm. From the treatment perspective, some children might have followed the treatment regimen regularly whereas other children might have faltered with the routine.

Next, we shall investigate the profile plots of subjects with different levels of age and sex. Age and sex are two of the categorical variables given in the dataset. Hence, exploratory visualization of the mean response profile at different levels of the two variables may yield useful insights.

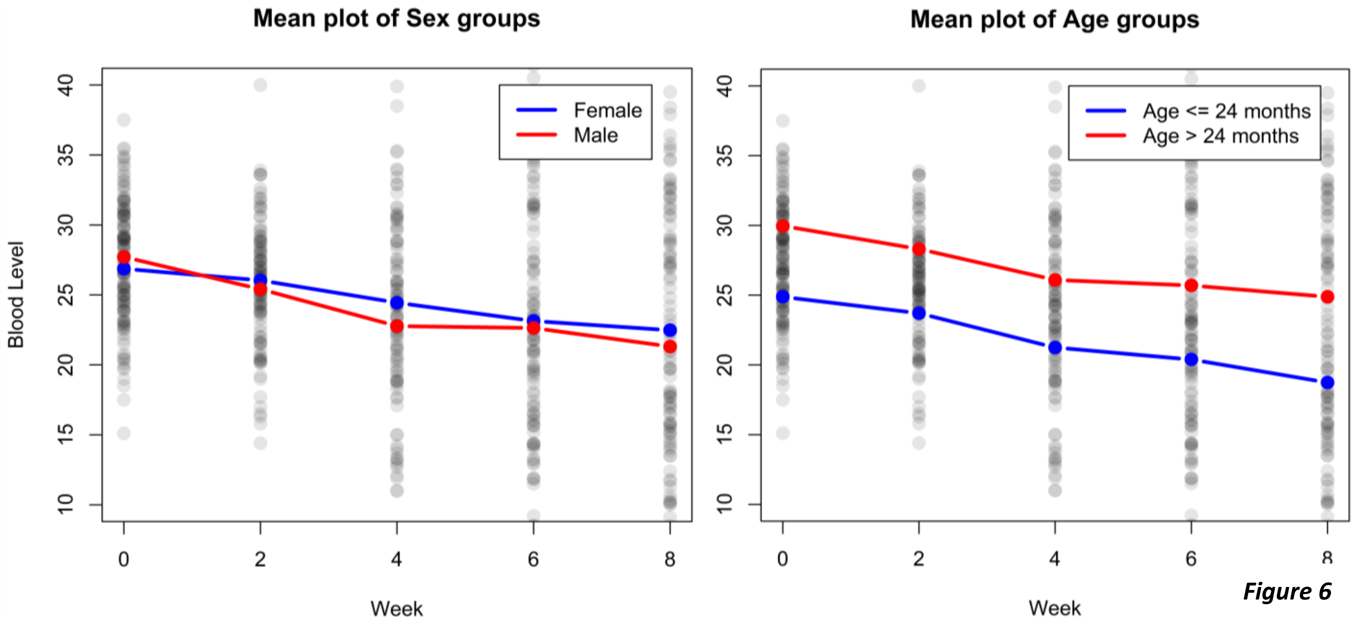
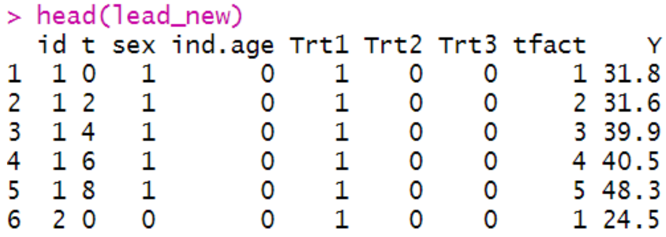


Figure 6

Figure 6 shows profile plots of the mean response with different levels of Sex variable (first plot) and Age variable (second plot). The Age levels plot shows that there is a significant near-constant difference in response at all time points. This could be due to a difference in lead-ingesting habits between children of the two age groups. It also tells us that the lead levels reduce at the same rate in both the age groups. The sex levels plot does not show any significant difference in response levels at any time point. Hence, one can assume that both sexes have similar lead levels at the start and that the lead levels wane equally for both sex groups. This means that the sex variable may have an insignificant effect on the response. For more detail, we can plot separate profiles for all four combinations of age and sex levels for each treatment. All detailed plots are provided in **Appendix 5.1** for reference.

3.2 Model Fitting (Part B)

To model the data, we shall first import it, perform a few data transformations and bring it to a form which is easy for modelling. We shall use techniques like one-hot encoding and create a new data frame called lead_new for modelling. Relevant code is presented in **Appendix 5.4**. The first few observations in the lead_new dataset are also presented below for reference.



```
> head(lead_new)
  id t sex ind.age Trt1 Trt2 Trt3 tfact    Y
1  1 0   1       0    1    0    0     1 31.8
2  1 2   1       0    1    0    0     2 31.6
3  1 4   1       0    1    0    0     3 39.9
4  1 6   1       0    1    0    0     4 40.5
5  1 8   1       0    1    0    0     5 48.3
6  2 0   0       0    1    0    0     1 24.5
```

$Trt1 = 1\ if\ Placebo, 0\ otherwise$
$Trt2 = 1\ if\ Succimer\ Low, 0\ otherwise$
$Trt3 = 1\ if\ Succimer\ High, 0\ otherwise$
$Y \rightarrow blood\ lead\ level\ (response)$
$t \rightarrow week$

In order to model the response column 'Y', we write a model equation considering all main effects and interactions upto 3rd order interactions and common random effects for Intercept and 't'.

$$
\begin{aligned}
Y_{ij} = {} & Trt1_i\big(\beta_{0,1} + \beta_{1,1}sex_i + \beta_{2,1}ind.age_i + \beta_{3,1}t_{ij} + \beta_{4,1}(sex_i \times ind.age_i) + \\
& \quad \beta_{5,1}(ind.age_i \times t_{ij}) + \beta_{6,1}(sex_i \times t_{ij}) + \beta_{7,1}(sex_i \times ind.age_i \times t_{ij})\big) \\
& + Trt2_i\big(\beta_{0,2} + \beta_{1,2}sex_i + \beta_{2,2}ind.age_i + \beta_{3,2}t_{ij} + \beta_{4,2}(sex_i \times ind.age_i) + \\
& \quad \beta_{5,2}(ind.age_i \times t_{ij}) + \beta_{6,2}(sex_i \times t_{ij}) + \beta_{7,2}(sex_i \times ind.age_i \times t_{ij})\big) \\
& + Trt3_i\big(\beta_{0,3} + \beta_{1,3}sex_i + \beta_{2,3}ind.age_i + \beta_{3,3}t_{ij} + \beta_{4,3}(sex_i \times ind.age_i) + \\
& \quad \beta_{5,3}(ind.age_i \times t_{ij}) + \beta_{6,3}(sex_i \times t_{ij}) + \beta_{7,3}(sex_i \times ind.age_i \times t_{ij})\big) + b_0 + b_1 t_{ij} + e_{ij}
\end{aligned}
$$

$i(Individual\ ID) \rightarrow (1,2,3 \ldots 120); \quad j(time\ point) \rightarrow (0,2,4,6,8)$

$$
b_i = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \sim N\left[0, D = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix}\right]
$$

$\beta \rightarrow Fixed\ effect\ coefficient$
$b \rightarrow Random\ effect\ coefficient$

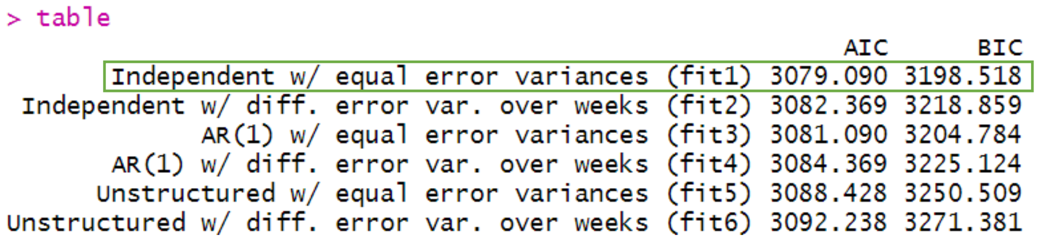
The equivalent R formula is also given as follows:

$$
\begin{aligned}
& meanform <- \ Y \sim -1 + Trt1 + Trt1{:}sex + Trt1{:}ind.age + Trt1{:}t + Trt1{:}sex{:}ind.age + \\
& Trt1{:}ind.age{:}t + Trt1{:}sex{:}t_{ij} + Trt1{:}sex{:}ind.age{:}t + Trt2 + Trt2{:}sex + Trt2{:}ind.age + \\
& Trt2{:}t + Trt2{:}sex{:}ind.age + Trt2{:}ind.age{:}t + Trt2{:}sex{:}t_{ij} + Trt2{:}sex{:}ind.age{:}t + Trt3 + \\
& Trt3{:}sex + Trt3{:}ind.age + Trt3{:}t + Trt3{:}sex{:}ind.age + Trt3{:}ind.age{:}t + Trt3{:}sex{:}t_{ij} + \\
& Trt3{:}sex{:}ind.age{:}t
\end{aligned}
$$

For modelling the error covariance structure, we fit 6 different models with the same fixed effects and random effects described above but different error covariance structures. The structures we would like to investigate are:

a) Independent, where error variance does not change over weeks (fit1)
b) Independent, where error variance changes over weeks (fit2)
c) AR(1) correlation structure, where error variance does not change over weeks (fit3)
d) AR(1) correlation structure, where error variance changes over weeks (fit4)
e) Unstructured, where error variance does not change over weeks (fit5)
f) Unstructured, where error variance changes over weeks (fit6)

Individual code for fitting, and obtaining AIC and BIC values for each of the six models above are presented in **Appendix 5.5**. From the above models, we choose the model that has lowest AIC and BIC values as our best fit model. The values are presented in the table below for comparison:



```
> table
                                                      AIC      BIC
     Independent w/ equal error variances (fit1) 3079.090 3198.518
Independent w/ diff. error var. over weeks (fit2) 3082.369 3218.859
            AR(1) w/ equal error variances (fit3) 3081.090 3204.784
       AR(1) w/ diff. error var. over weeks (fit4) 3084.369 3225.124
     Unstructured w/ equal error variances (fit5) 3088.428 3250.509
Unstructured w/ diff. error var. over weeks (fit6) 3092.238 3271.381
```

From the table, fit1 i.e. the model with independent error covariance structure with equal error variances has the lowest AIC and BIC value. Hence it is chosen as the best model for further analysis. The fit1 model will now be renamed as ‘bestfit’ throughout the rest of the report.

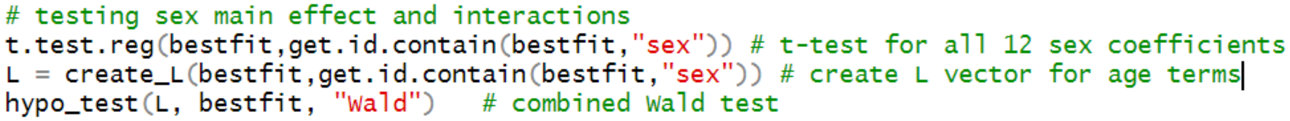
3.3 Model Reduction, Inference and Diagnostics (Part C)

3.3.1 Does Age and Sex have significant association with the response?

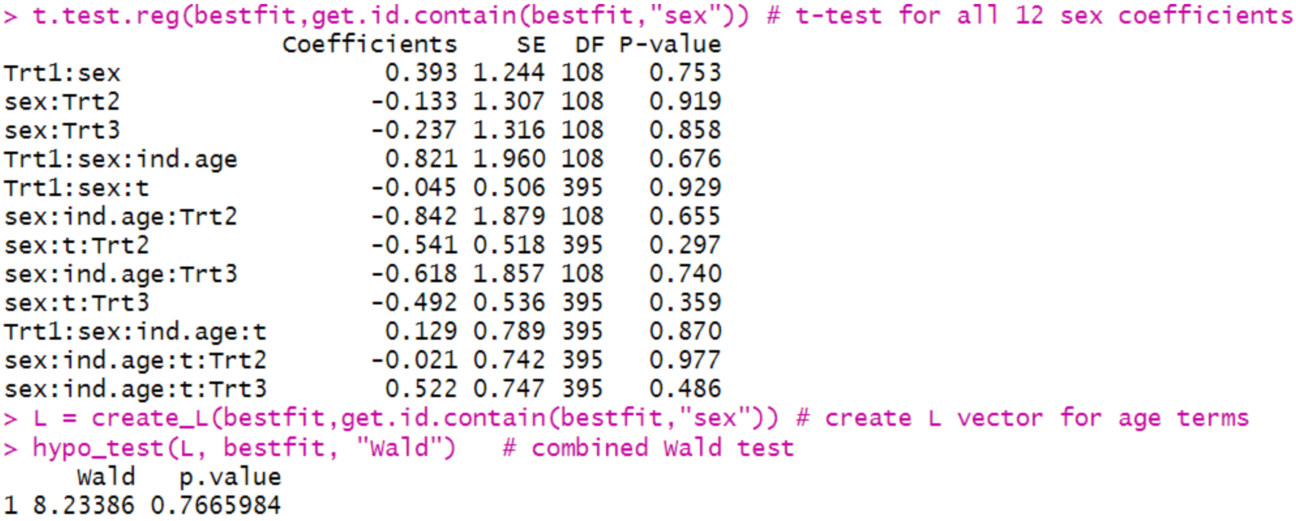
The first question we will investigate using our bestfit model is whether gender or age have any association with blood level. Let us look at the case of gender first. The sex variable in our dataset defines the gender of the subject. In our model we have 12 terms which contain interactions of the sex variable with one or more other variables. Hence, for the sex variable to have no association with the response, all the 12 coefficients corresponding to the terms should be equal to zero. Therefore, the null hypothesis we are trying to test is:

$$H_0: \beta_{1,1} = \beta_{4,1} = \beta_{6,1} = \beta_{7,1} = \beta_{1,2} = \beta_{4,2} = \beta_{6,2} = \beta_{7,2} = \beta_{1,3} = \beta_{4,3} = \beta_{6,3} = \beta_{7,3} = 0$$

To test this hypothesis, we can conduct individual t-tests for each coefficient or a combined Wald test for all coefficients. We shall look at results from both tests. Helper functions create_L(), get.id.contain(), hypo_test() and t.test.reg() have been coded to help with our analysis. Function definitions of each function are provided in **Appendix 5.2** for reference.



```
# testing sex main effect and interactions
t.test.reg(bestfit,get.id.contain(bestfit,"sex")) # t-test for all 12 sex coefficients
L = create_L(bestfit,get.id.contain(bestfit,"sex")) # create L vector for age terms
hypo_test(L, bestfit, "Wald")   # combined Wald test
```



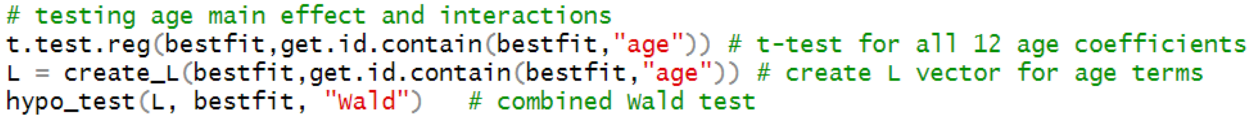
```
> t.test.reg(bestfit,get.id.contain(bestfit,"sex")) # t-test for all 12 sex coefficients
                     Coefficients    SE  DF P-value
Trt1:sex                    0.393 1.244 108   0.753
sex:Trt2                   -0.133 1.307 108   0.919
sex:Trt3                   -0.237 1.316 108   0.858
Trt1:sex:ind.age            0.821 1.960 108   0.676
Trt1:sex:t                 -0.045 0.506 395   0.929
sex:ind.age:Trt2           -0.842 1.879 108   0.655
sex:t:Trt2                 -0.541 0.518 395   0.297
sex:ind.age:Trt3           -0.618 1.857 108   0.740
sex:t:Trt3                 -0.492 0.536 395   0.359
Trt1:sex:ind.age:t          0.129 0.789 395   0.870
sex:ind.age:t:Trt2         -0.021 0.742 395   0.977
sex:ind.age:t:Trt3          0.522 0.747 395   0.486
> L = create_L(bestfit,get.id.contain(bestfit,"sex")) # create L vector for age terms
> hypo_test(L, bestfit, "Wald")   # combined Wald test
     Wald   p.value
1 8.23386 0.7665984
```

From the results of both individual t-tests and the combined Wald test shown above, it is evident that none of the sex term coefficients are statistically significant as all p-values are above 0.05 which is the chosen level of significance. Hence, our null hypothesis cannot be rejected, and we conclude that the sex variable does not have any statistically significant association with the blood level response in our model.

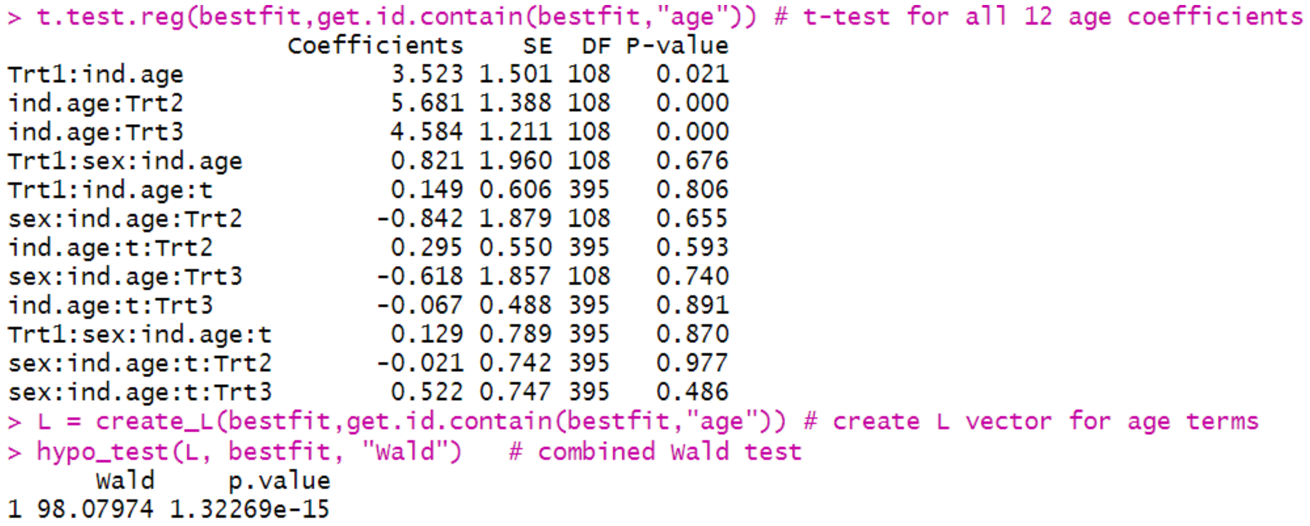
Next, we repeat a similar process for our age terms. Again, we find 12 such terms involving age in our model. The null hypothesis we will test is as follows:

$$H_0: \beta_{2,1} = \beta_{4,1} = \beta_{5,1} = \beta_{7,1} = \beta_{2,2} = \beta_{4,2} = \beta_{5,2} = \beta_{7,2} = \beta_{2,3} = \beta_{4,3} = \beta_{5,3} = \beta_{7,3} = 0$$

Relevant code and results are displayed below:



```
# testing age main effect and interactions
t.test.reg(bestfit,get.id.contain(bestfit,"age")) # t-test for all 12 age coefficients
L = create_L(bestfit,get.id.contain(bestfit,"age")) # create L vector for age terms
hypo_test(L, bestfit, "Wald")   # combined Wald test
```

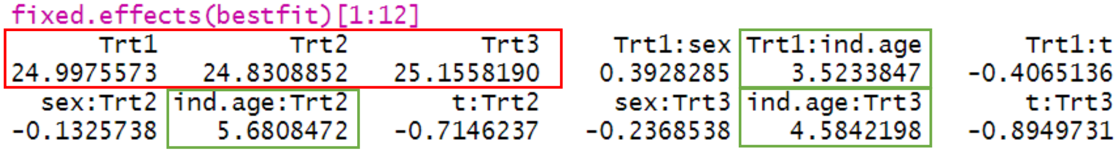


```
> t.test.reg(bestfit,get.id.contain(bestfit,"age")) # t-test for all 12 age coefficients
                   Coefficients    SE  DF P-value
Trt1:ind.age              3.523 1.501 108   0.021
ind.age:Trt2              5.681 1.388 108   0.000
ind.age:Trt3              4.584 1.211 108   0.000
Trt1:sex:ind.age          0.821 1.960 108   0.676
Trt1:ind.age:t            0.149 0.606 395   0.806
sex:ind.age:Trt2         -0.842 1.879 108   0.655
ind.age:t:Trt2            0.295 0.550 395   0.593
sex:ind.age:Trt3         -0.618 1.857 108   0.740
ind.age:t:Trt3           -0.067 0.488 395   0.891
Trt1:sex:ind.age:t        0.129 0.789 395   0.870
sex:ind.age:t:Trt2       -0.021 0.742 395   0.977
sex:ind.age:t:Trt3        0.522 0.747 395   0.486
> L = create_L(bestfit,get.id.contain(bestfit,"age")) # create L vector for age terms
> hypo_test(L, bestfit, "Wald")   # combined Wald test
      Wald      p.value
1 98.07974 1.32269e-15
```

From the results shown above, we can observe that Trt1:ind.age, Trt2:ind.age, Trt3:ind.age are significant terms. These terms are in fact the subject-level main effects for the ind.age variable. The Wald test result also shows that the combined test of all 12 ind.age terms are indeed significant owing to the small p-value (< 0.05). Hence, the null hypothesis is rejected, and we conclude that Age has a significant association with the blood level response in our model.

3.3.2 Model Reduction

From the previous analysis, it is evident that a lot of terms in our model are statistically close to zero. In other words, they have a statistically insignificant bearing on the response we are trying to predict. Hence, in the interest of simplifying the model, one can omit these terms from the bestfit model for a negligible sacrifice in goodness of fit. The terms which shall be omitted are all the terms (interactions and main effects) containing the sex variable and the $ind.age \times t$ interaction for all three treatment groups. A few of the full model coefficients are shown below:

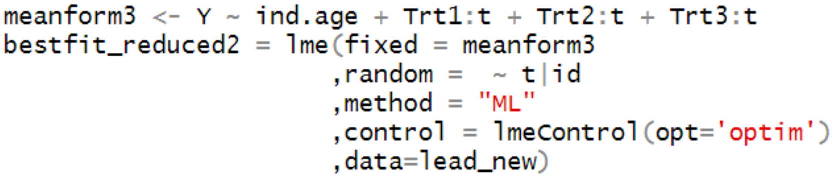


```
fixed.effects(bestfit)[1:12]
       Trt1         Trt2         Trt3     Trt1:sex Trt1:ind.age       Trt1:t
24.9975573   24.8308852   25.1558190    0.3928285    3.5233847   -0.4065136
  sex:Trt2 ind.age:Trt2       t:Trt2     sex:Trt3 ind.age:Trt3       t:Trt3
-0.1325738    5.6808472   -0.7146237   -0.2368538    4.5842198   -0.8949731
```

The subject-wise Intercept coefficients (in red box) and ind.age coefficients (in green box) seem to be similar. If they are indeed statistically similar, then we can replace these 6 terms with a common intercept and ind.age term. This hypothesis is tested using the anova.lme function and results are provided in **Appendix 5.3**. From the results, we can say that the coefficients for the two sets of terms in each marked box are similar to each other statistically. Hence, they are also replaced as described previously. The newly reduced model equation is presented below:

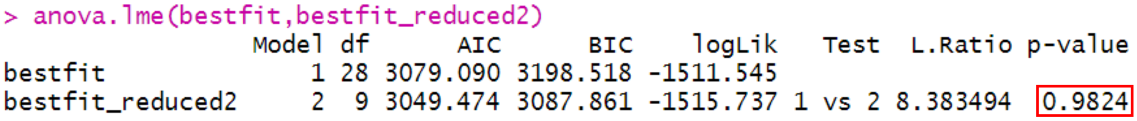
$$Y_{ij} = \beta_0 + \beta_1(Trt1_i \times t_{ij}) + \beta_2(Trt2_i \times t_{ij}) + \beta_3(Trt3_i \times t_{ij}) + \beta_4 ind.age_i$$

The code shown below fits the reduced model:



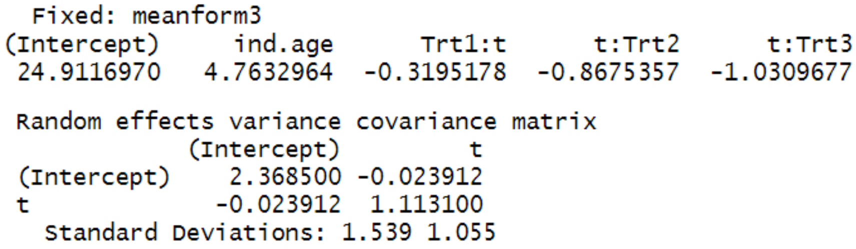
```
meanform3 <- Y ~ ind.age + Trt1:t + Trt2:t + Trt3:t
bestfit_reduced2 = lme(fixed = meanform3
                      ,random =   ~ t|id
                      ,method = "ML"
                      ,control = lmeControl(opt='optim')
                      ,data=lead_new)
```

To check whether the reduced model is sufficient, we use the anova.lme function to compare the two models:



```
> anova.lme(bestfit,bestfit_reduced2)
                 Model df      AIC      BIC    logLik   Test  L.Ratio p-value
bestfit              1 28 3079.090 3198.518 -1511.545
bestfit_reduced2     2  9 3049.474 3087.861 -1515.737 1 vs 2 8.383494  0.9824
```

The insignificant p-value of 0.9824 tells us that the reduced model is indeed sufficient. Also, the reduced model is better in terms of its AIC and BIC values, which are lesser than those of the full model. The AIC and BIC penalizes the addition of covariates to the model if it does not improve the goodness of fit of the model by a significant amount. Hence, our reduced model has the better balance between goodness of fit and model complexity. The fixed effect estimates and the random effects covariance matrix of the reduced model are presented below:



```
  Fixed: meanform3
(Intercept)     ind.age      Trt1:t      t:Trt2      t:Trt3 
 24.9116970   4.7632964  -0.3195178  -0.8675357  -1.0309677 

 Random effects variance covariance matrix
            (Intercept)         t
 (Intercept)    2.368500 -0.023912
 t             -0.023912  1.113100
   Standard Deviations: 1.539 1.055 
```

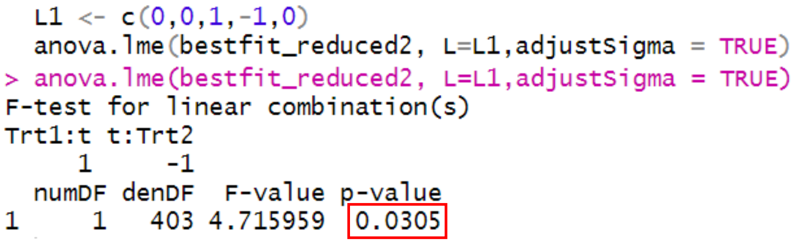
3.3.3 Mean Trend Analysis

The next question the report investigates is whether the mean trends of blood level are the same for all three treatments. This question can be answered using a pairwise comparison of coefficients for each of the three trends. For each pairwise comparison, we shall have 3 individual hypotheses testing the equality of corresponding coefficients for the two treatments in comparison.

<u>Comparing Placebo (Trt1) and Succimer Low Dose (Trt2):</u>

The null hypothesis is $H_0 : \beta_1 = \beta_2$

We test the hypothesis using the anova.lme function and an appropriate L vector. Relevant code and results are displayed below:



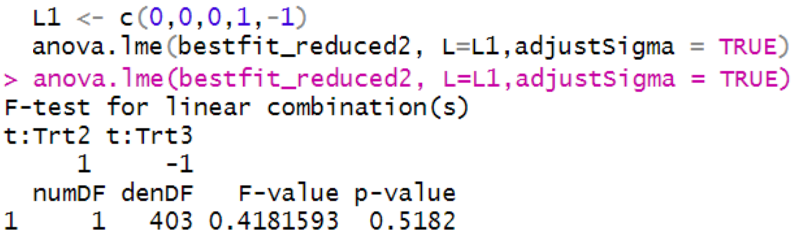
```
  L1 <- c(0,0,1,-1,0)
  anova.lme(bestfit_reduced2, L=L1,adjustSigma = TRUE)
> anova.lme(bestfit_reduced2, L=L1,adjustSigma = TRUE)
F-test for linear combination(s)
Trt1:t t:Trt2 
     1     -1 
  numDF denDF  F-value p-value
1     1   403 4.715959  0.0305
```

The p-value for $H_0 : \beta_1 = \beta_2$ is 0.0305. Hence, there is a significant difference in response as time passes by. In other words the slopes of the t:Trt1 and t:Trt2 interactions are significantly different. <u>Hence the mean trends for Placebo and Succimer Low Dose treatments are not the same.</u>

<u>Comparing Succimer Low Dose (Trt2) and Succimer High Dose (Trt3):</u>

The null hypothesis is $H_0 : \beta_2 = \beta_3$

Relevant code and results are presented below:



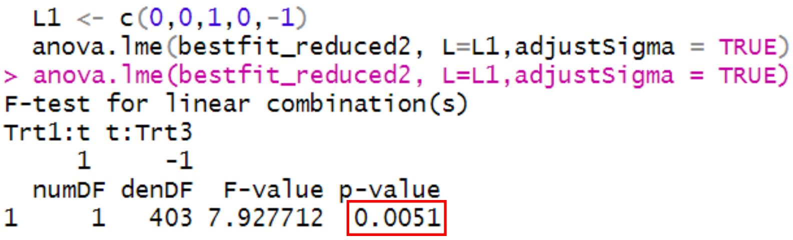
```
  L1 <- c(0,0,0,1,-1)
  anova.lme(bestfit_reduced2, L=L1,adjustSigma = TRUE)
> anova.lme(bestfit_reduced2, L=L1,adjustSigma = TRUE)
F-test for linear combination(s)
t:Trt2 t:Trt3
     1     -1
  numDF denDF   F-value p-value
1     1   403 0.4181593  0.5182
```

From the results, the p-value in comparison is insignificant (> 0.05). Hence the null hypotheses cannot be rejected. <u>Hence the mean trends for Succimer Low Dose and Succimer High Doses are statistically the same</u>.

<u>Comparing Placebo (Trt1) and Succimer High Dose (Trt3):</u>

The null hypothesis is $H_0 : \beta_1 = \beta_3$

Relevant code and results are presented below:



```
  L1 <- c(0,0,1,0,-1)
  anova.lme(bestfit_reduced2, L=L1,adjustSigma = TRUE)
> anova.lme(bestfit_reduced2, L=L1,adjustSigma = TRUE)
F-test for linear combination(s)
Trt1:t t:Trt3
     1     -1
  numDF denDF  F-value p-value
1     1   403 7.927712  0.0051
```

From the results we can see that the slopes of the Trt1:t and Trt3:t variables are statistically different from one another as the corresponding p-value is less than 0.05. Hence, we conclude that the mean trends for Placebo and Succimer High Dose treatments are not the same.

From the pairwise comparison and testing, we can conclude that the mean trends of both Succimer Doses (Low and High) are significantly different from the Placebo treatment but aren't very different from each other. Hence while the Succimer treatment is effective in general, a High Dose of the treatment may not be necessary.

We shall now estimate the individual mean trends of the following combinations of age, sex and treatment levels.

a) Male/female *with age* < 24 receiving placebo:

For this combination, the fixed effects in our reduced model get the following values:

Fixed Effect	Intercept	ind.age	Trt1:t	Trt2:t	Trt3:t
Coefficient Value	24.91	4.76	-0.32	-0.87	-1.03
Variable Value	1	0	t	0	0

Hence the estimated mean trend will be $y = 24.91 - 0.32t$ where ‘t’ can take any value in (0, 2, 4, 6, 8). Note that since the sex variable has been removed, the mean trends will stay the same for both males and females.

b) Male/female *with age* > 24 receiving placebo:

Here, the ind.age term additionally gets a value of 1 since ind.age = 1 and Trt1 = 1. Hence, the estimated mean trend will be $y = 24.91 + 4.76 - 0.32t = 29.67 - 0.32t$

c) Male/female *with age* < 24 receiving Succimer Low Dose:

Here, the Intercept and Trt2:t effects get a value of 1 and t respectively. The remaining effects variable values are zero. Hence the estimated mean trend will be $y = 24.91 - 0.87t$

d) Male/female *with age* > 24 receiving Succimer Low Dose:

Here, the Intercept, ind.age and Trt2:t effects get a value of 1, 1 and t respectively. The remaining effect variable values are zero. Hence the estimated mean trend will be $y = 24.91 + 4.76 - 0.88t = 29.67 - 0.88t$

e) Male/female *with age* < 24 receiving Succimer High Dose:

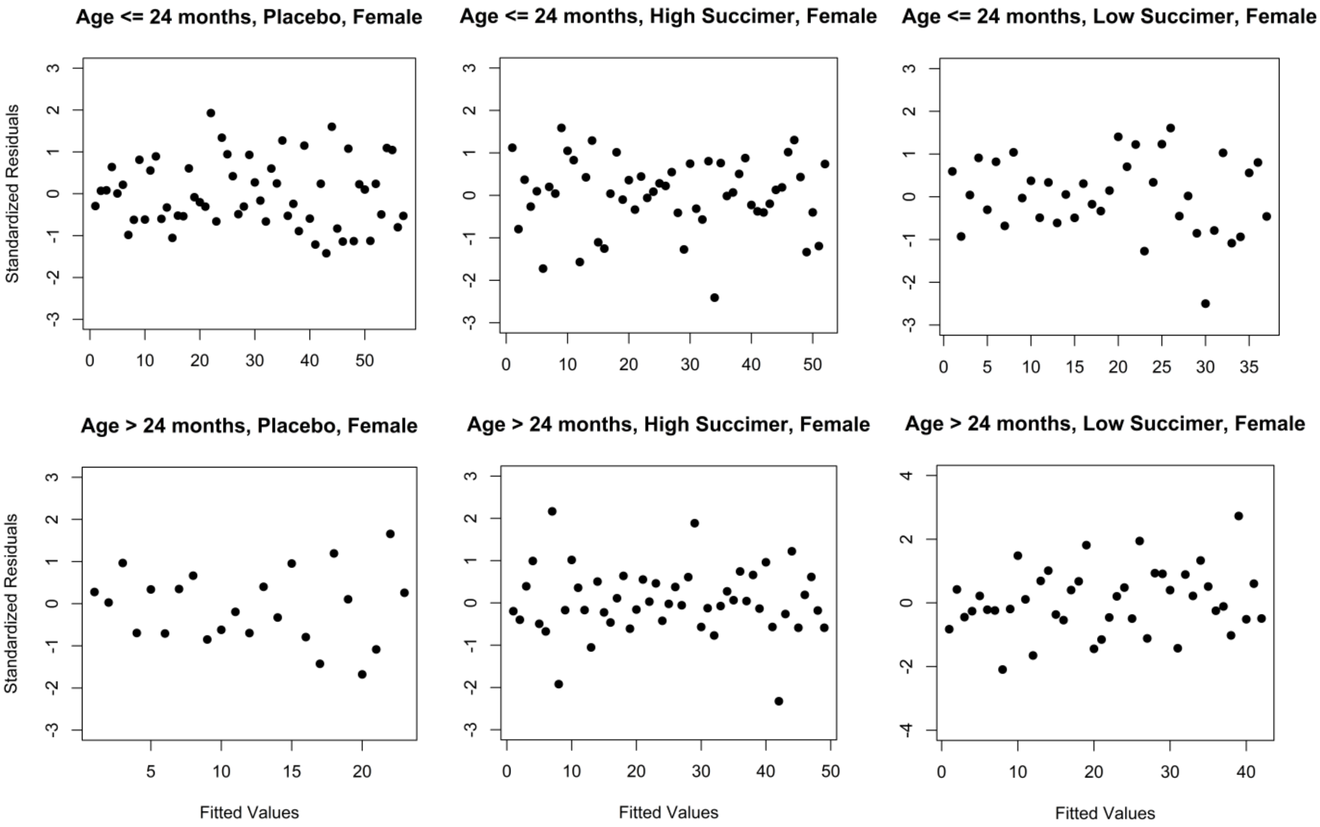
Here, the Intercept and Trt3:t effects get a value of 1 and t respectively. The remaining effect variable values are zero. Hence the estimated mean trend will be $y = 24.91 - 1.03t$

f) <u>Male/female *with age* > 24 receiving Succimer High Dose:</u>

Here, the Intercept, ind.age and Trt3:t effects get a value of 1, 1 and t respectively. The remaining effect variable values are zero. Hence the estimated mean trend will be $y = 24.91 + 4.76 - 1.03t = 29.67 - 1.02t$

3.3.4 Model Diagnostics

As per our chosen reduced model, the assumptions made were that residual errors were normally distributed, independent, have a mean value of zero and have constant variance across treatments, age and sex levels $(cov(e_{ij}) = \sigma^2 I)$. We can visually verify few assumptions by constructing residual plots of the Person Residuals for every combination of variable levels.


Age <= 24 months, Placebo, Female

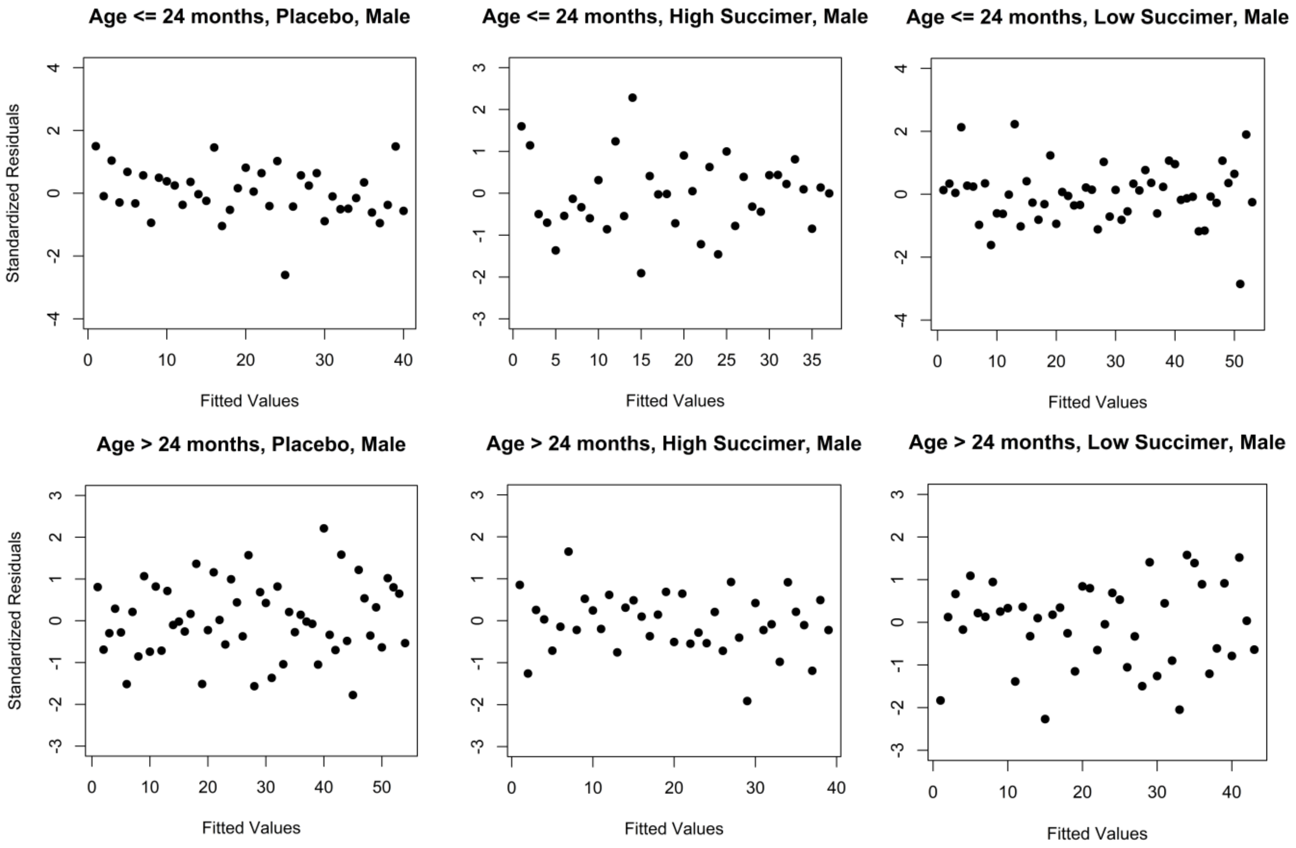
Age <= 24 months, High Succimer, Female

Age <= 24 months, Low Succimer, Female

Age > 24 months, Placebo, Female

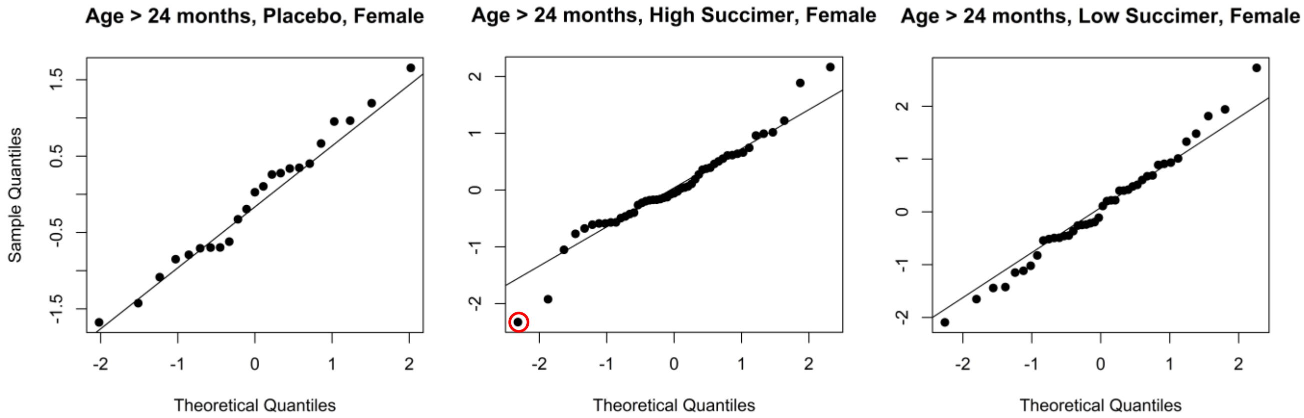
Age > 24 months, High Succimer, Female

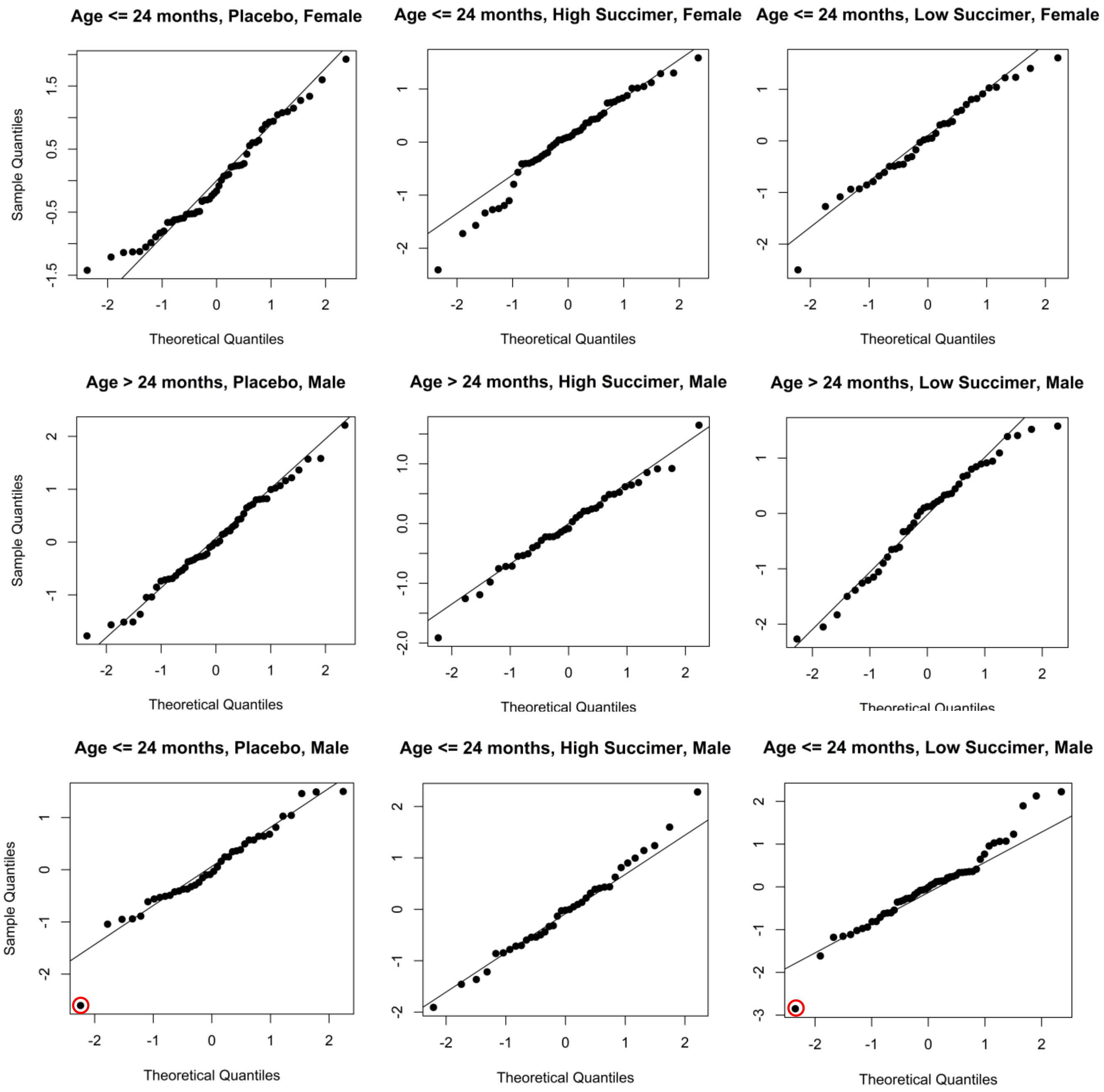
Age > 24 months, Low Succimer, Female



From analyzing the residual plots of all 12 possible combinations above, the assumptions that errors have a zero mean and have constant variances across all level combinations are plausible.

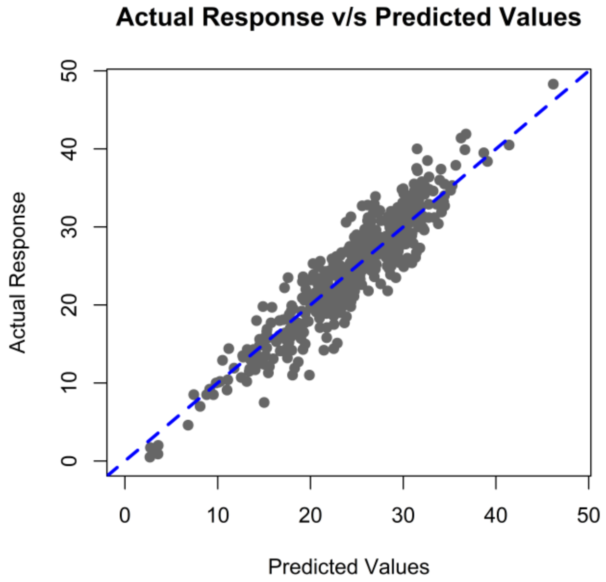
To verify the normality of errors, one can visually inspect the normal quantile-quantile plots of the normalized residuals. If observations fall approximately on the normal straight line, then the normality assumption is viable. The normal Q-Q plots of the residuals for all group-level combinations are presented below:





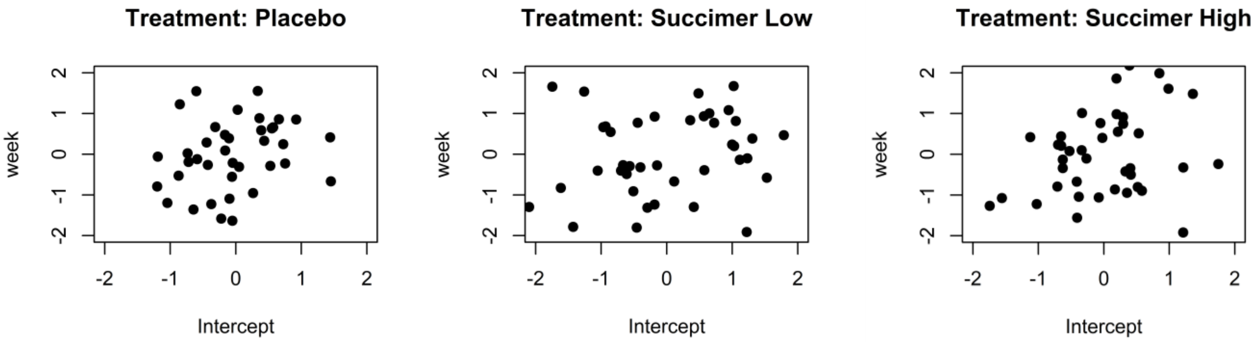
From the Q-Q plots, the observations mostly seem to follow the normal straight line. There are a few outliers to this assumption which have been circled in red. Apart from these outliers, the overall assumption of normality of residuals is also plausible.

We can also visualize the quality of fit by plotting observed responses vs. subject-level fitted values.

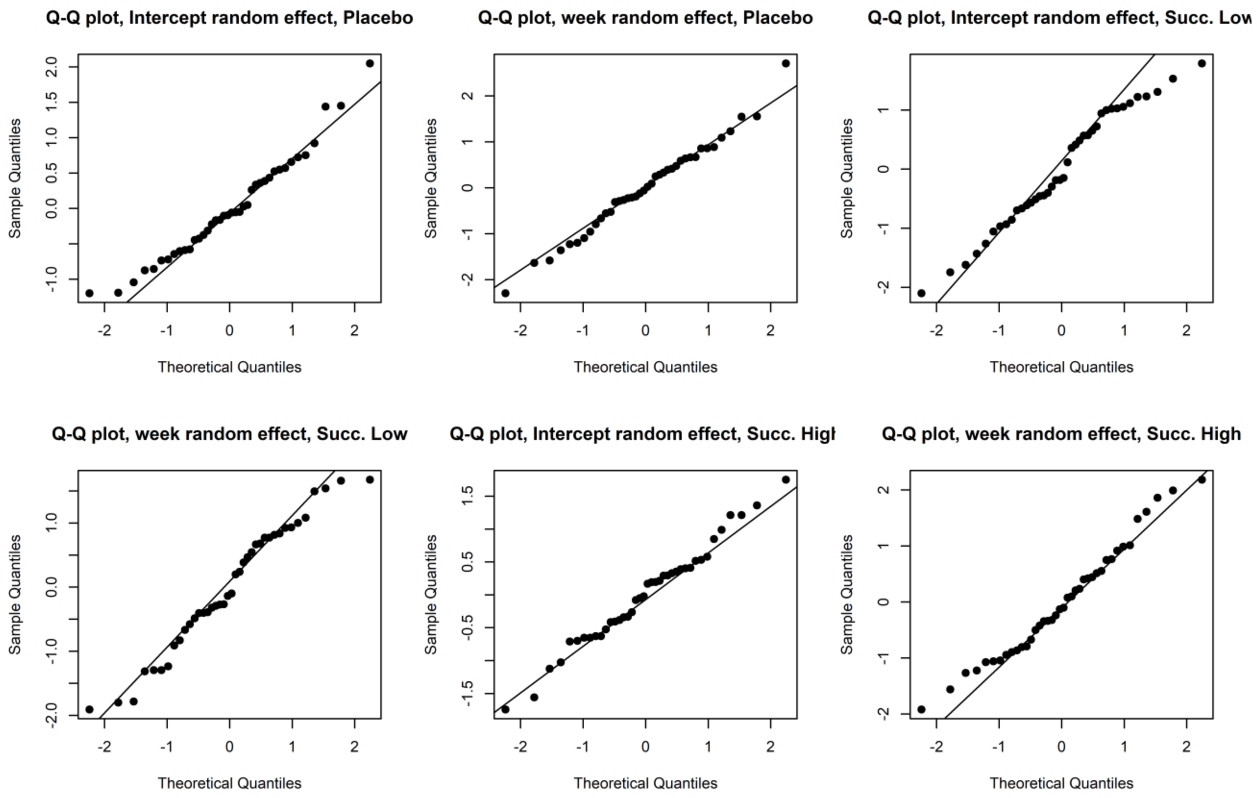


The Actual Response vs. Predicted values are indeed close as they approximately follow the bisection line (blue dotted).

We can also verify the normality and zero mean assumptions of our random effects. The procedure is similar to the one followed previously where we investigated the residual plots and normal Q-Q plots. The plots shown below are of the random effects where the intercept random effect is plotted on the X-Axis and the slope of week random effect is plotted on the Y-Axis. Three plots one for each treatment are presented to keep the report succinct.



The plots above show that random effects for the three treatments have similar ranges and are meaned at zero. The shrinkage effect around point (0,0) is also discernable from the plots.



A normal Q-Q plot of the random effects can also be plotted to verify the normality assumption. From the above random effect Q-Q plots, the normality assumptions of the random effects are reasonable.

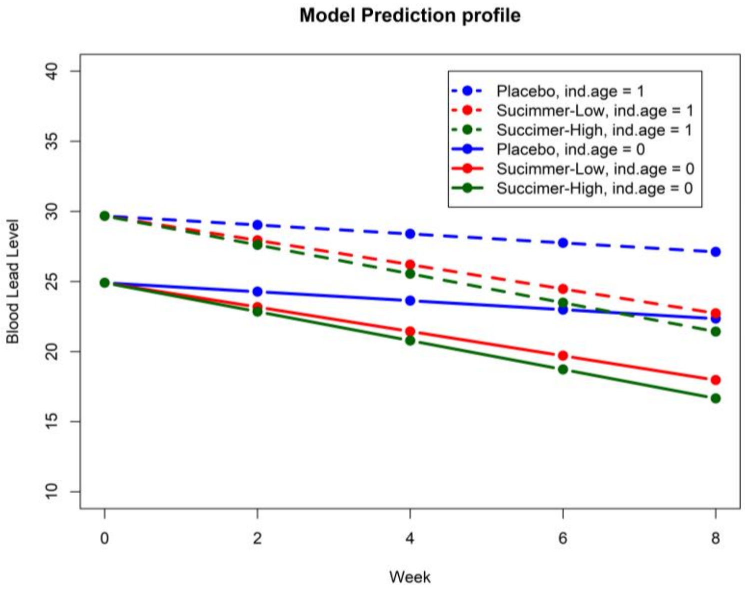
To conclude, all model assumptions have been investigated visually and have been found to be reasonably obeyed and appropriate.

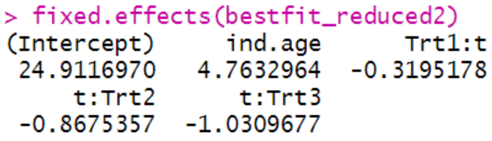
4. Conclusions

The aim of the study was to model the blood lead level response as a function of the covariates in the dataset and answer a few fundamental questions regarding the Succimer chelating treatment and it's effectiveness in reducing blood levels which were presented in section 1.4. The following conclusions can be made post analysis.

- The Succimer treatment in both Low and High Doses was found to be effective in reducing lead levels owing to their significantly different slopes from the Placebo treatment.
- No significant difference in performance was found between the Low Dosage and High Dosage version of the treatment.
- The sex of the subject and it's interactions with other variables was found to have no significant association with the blood lead level.
- The Age group of the subject had significantly different lead levels at the start of the treatment. The group of subjects with age > 24 months showed significantly higher concentrations of lead at the beginning.
- Although the age groups had different lead concentrations at the start, both treatments reduced blood lead levels at equal rates with respect to time for both groups owing to their statistically similar slopes (coefficients of Trt2:t and Trt3:t)

The final model mean prediction profile and fixed effects are presented below for reference:





```
> fixed.effects(bestfit_reduced2)
(Intercept)      ind.age      Trt1:t
 24.9116970    4.7632964  -0.3195178
     t:Trt2       t:Trt3
 -0.8675357   -1.0309677
```

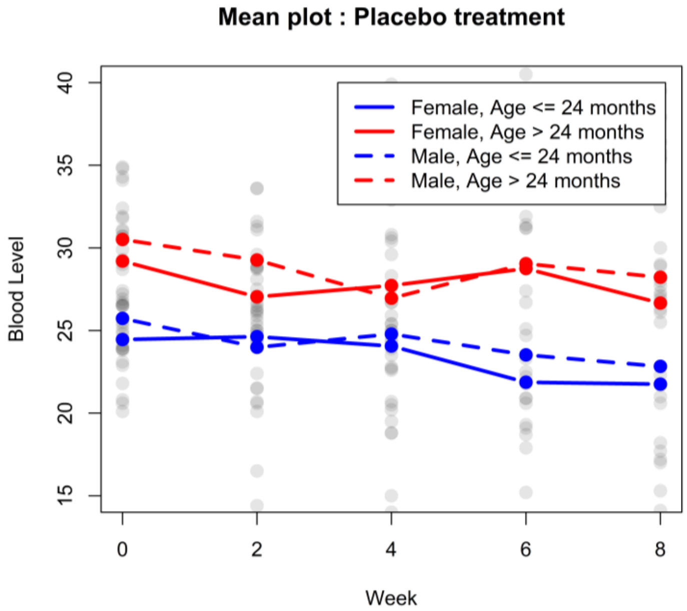
The final mean prediction equation is:

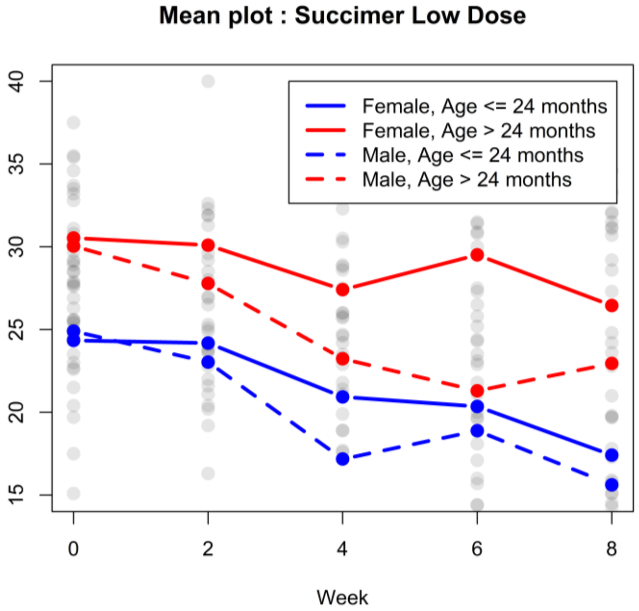
$$\hat{Y}_{ij} = 24.91 + 4.76 ind.age - 0.32(Trt1 \times t)$$
$$-0.87(Trt2 \times t) - 1.03(Trt3 \times t)$$

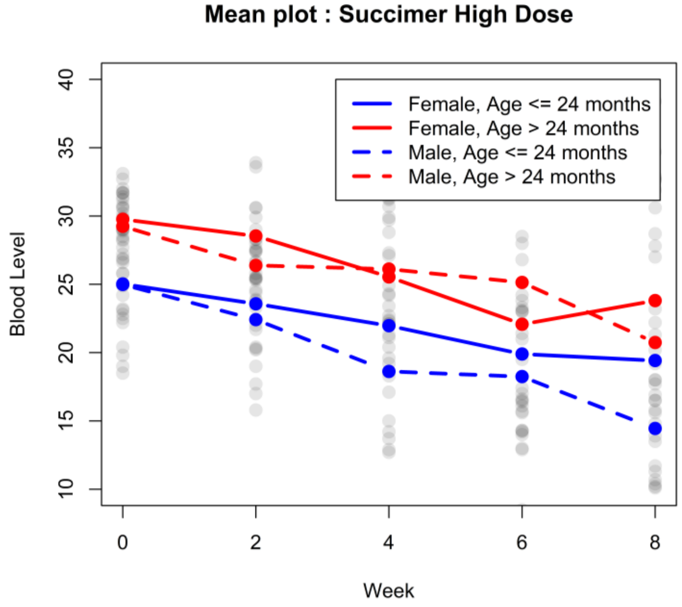
5. Appendix

5.1 Mean profiles for level combinations

Age and Sex variable combination profiles (Treatment-wise)

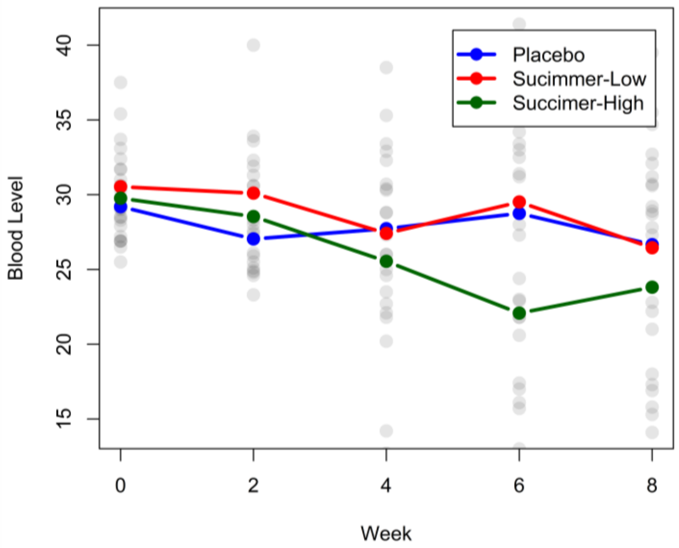




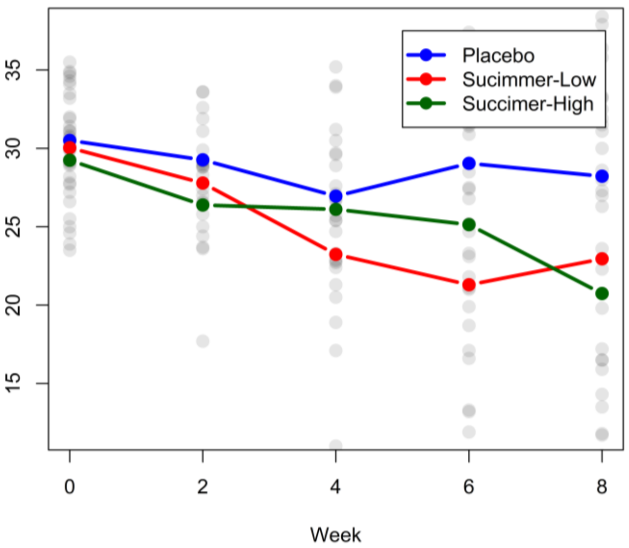


Treatment profiles (Age and Sex variable combination wise)

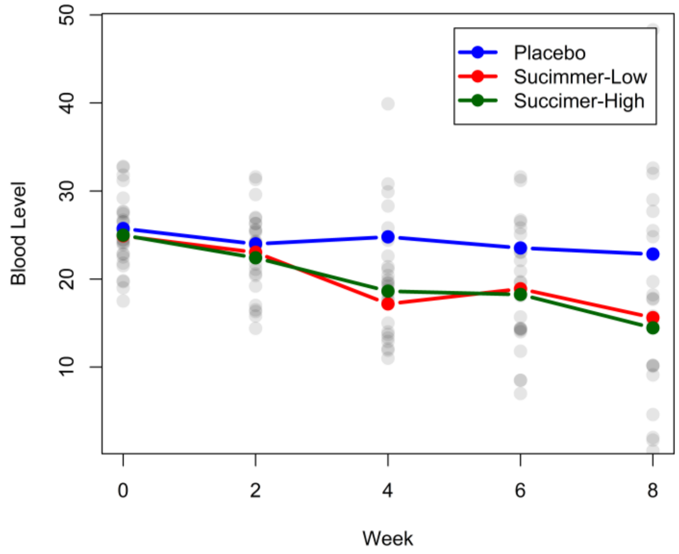
Mean plot : Age > 24 months, Female



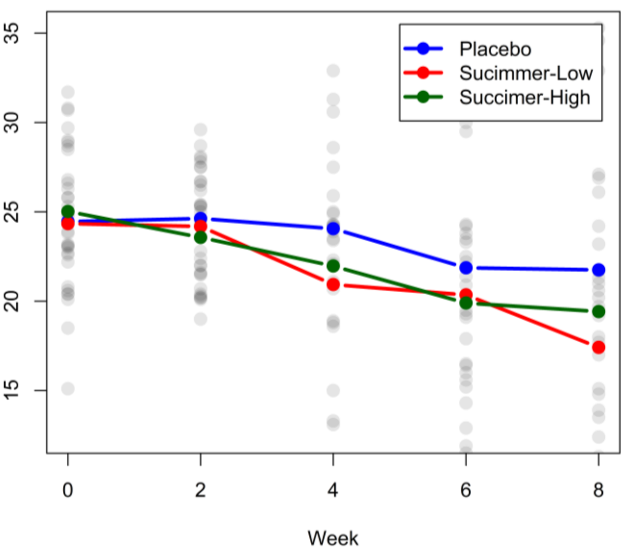
Mean plot : Age > 24 months, Male



Mean plot : Age <= 24 months, Male



Mean plot : Age <= 24 months, Female



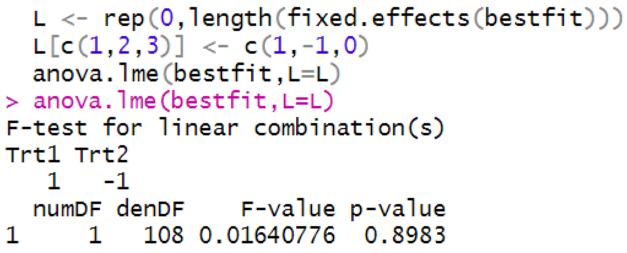
5.2 Custom function definitions

```
create_L <- function(lme_object,vector){
  L = matrix(0L, nrow = length(vector), ncol = length(fixed.effects(lme_object)))
  for (i in 1:length(vector)){
    L[i,vector[i]] = 1
  }
  return(L)
}

get.id.contain <- function(fitobject,string){
  return (grep(string,names(fixed.effects(fitobject))))
}

t.test.reg <- function(fitobject,L){
  df <- fitobject$fixDF$X
  df_L <- df[L]
  betahat_L <- fixed.effects(fitobject)[L]
  SE_L <- sqrt( diag(fitobject$varFix)[L] )
  t.stat <- betahat_L/SE_L
  p.value <- round( 2*pt(q = abs(t.stat), df = df_L, lower.tail = FALSE), 4 )
  out_sex <- data.frame(betahat_L, SE_L, df_L,p.value)
  colnames(out_sex) <- c("Coefficients","SE","DF","P-value")
  return (round(out_sex, 3))
}

hypo_test <- function(contrast,lme_object,test) {
  betahat <- fixed.effects(lme_object)
  V.robust <- vcovCR(lme_object, type = "CR0")
  cc <- nrow(L)
  df <- length(eval(lhs(eval(lme_object$call$fixed)))) - length(betahat)
  # estimate and covariance matrix of L\beta
  est <- contrast %*% betahat
  SE <- contrast %*% V.robust %*% t(contrast)
  varmat <- L %*% V.robust %*% t(L)
  if (test == "Wald") {
    # Wald test
    Wald <- c( t(est) %*% solve(varmat) %*% (est) )
    p.value <- pchisq(q = Wald, df = cc, lower.tail=FALSE)
    return(data.frame(Wald, p.value))
  }
  if (test == "F-test") {
    # F-test
    Fstat <- c( t(est) %*% solve(varmat) %*% (est) ) / cc
    p.value <- pf(q = Fstat, df1 = cc, df2 = df, lower.tail=FALSE)
    return(data.frame(Fstat, p.value))
  }
}
```

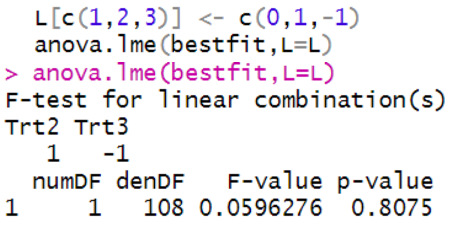
5.3 ANOVA Contrast test for the full model

In this section we test the following hypotheses for the full unreduced model (bestfit).

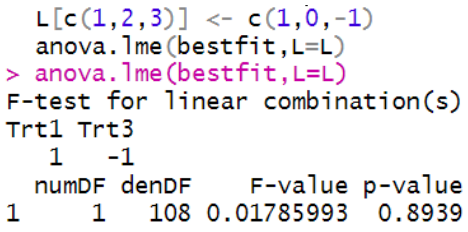
a) $H_0\colon \beta_{0,1} = \beta_{0,2} = \beta_{0,3}$ (beta_Trt1 = beta_Trt2 = beta_Trt3)



```
  L <- rep(0,length(fixed.effects(bestfit)))
  L[c(1,2,3)] <- c(1,-1,0)
  anova.lme(bestfit,L=L)
> anova.lme(bestfit,L=L)
F-test for linear combination(s)
Trt1 Trt2 
   1   -1 
  numDF denDF    F-value p-value
1     1   108 0.01640776  0.8983
```



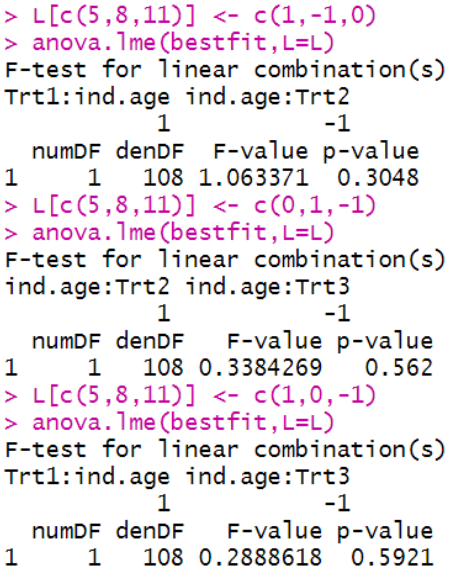
```
  L[c(1,2,3)] <- c(0,1,-1)
  anova.lme(bestfit,L=L)
> anova.lme(bestfit,L=L)
F-test for linear combination(s)
Trt2 Trt3 
   1   -1 
  numDF denDF   F-value p-value
1     1   108 0.0596276  0.8075
```



```
  L[c(1,2,3)] <- c(1,0,-1)
  anova.lme(bestfit,L=L)
> anova.lme(bestfit,L=L)
F-test for linear combination(s)
Trt1 Trt3 
   1   -1 
  numDF denDF    F-value p-value
1     1   108 0.01785993  0.8939
```

All three pairwise p-values are not significant. Hence the null hypothesis cannot be rejected.

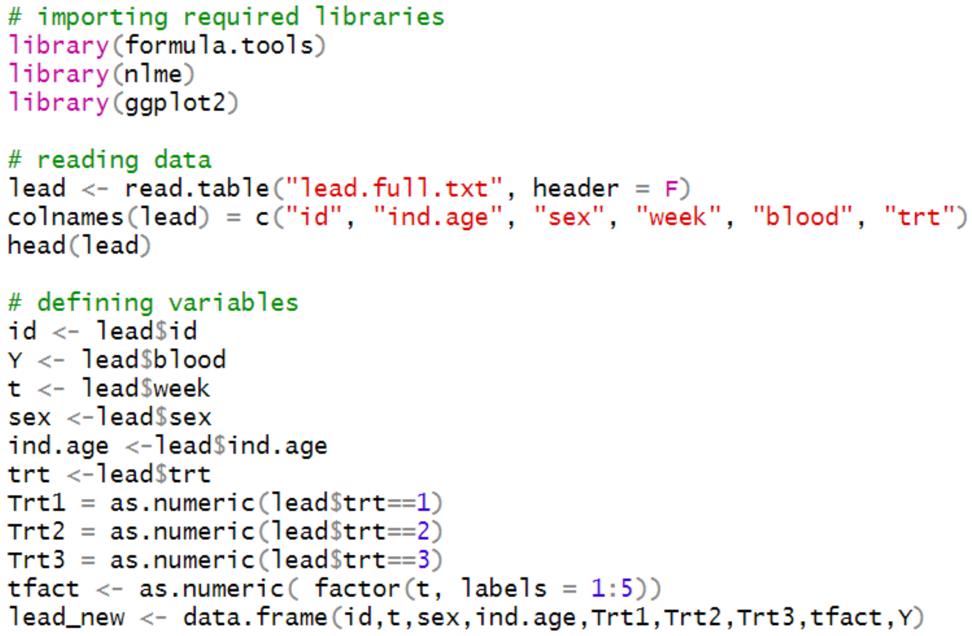
b) $H_0\colon \beta_{2,1} = \beta_{2,2} = \beta_{2,3}$ (beta_ind.ageXTrt1 = beta_ ind.ageXTrt2 = beta_ ind.ageXTrt3)



```
> L[c(5,8,11)] <- c(1,-1,0)
> anova.lme(bestfit,L=L)
F-test for linear combination(s)
Trt1:ind.age ind.age:Trt2 
           1           -1 
  numDF denDF  F-value p-value
1     1   108 1.063371  0.3048
> L[c(5,8,11)] <- c(0,1,-1)
> anova.lme(bestfit,L=L)
F-test for linear combination(s)
ind.age:Trt2 ind.age:Trt3 
           1           -1 
  numDF denDF   F-value p-value
1     1   108 0.3384269   0.562
> L[c(5,8,11)] <- c(1,0,-1)
> anova.lme(bestfit,L=L)
F-test for linear combination(s)
Trt1:ind.age ind.age:Trt3 
           1           -1 
  numDF denDF   F-value p-value
1     1   108 0.2888618  0.5921
```

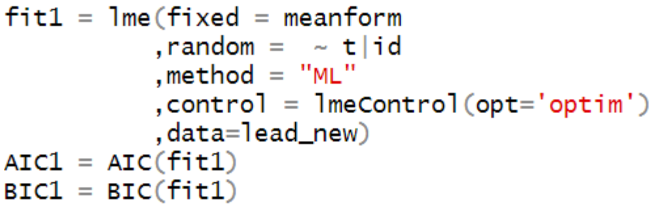
All three pairwise p-values are not significant. Hence the null hypothesis cannot be rejected.

5.4 Data Preparation model fitting code

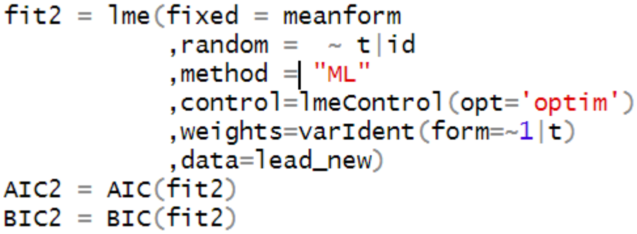


```
# importing required libraries
library(formula.tools)
library(nlme)
library(ggplot2)

# reading data
lead <- read.table("lead.full.txt", header = F)
colnames(lead) = c("id", "ind.age", "sex", "week", "blood", "trt")
head(lead)

# defining variables
id <- lead$id
Y <- lead$blood
t <- lead$week
sex <-lead$sex
ind.age <-lead$ind.age
trt <-lead$trt
Trt1 = as.numeric(lead$trt==1)
Trt2 = as.numeric(lead$trt==2)
Trt3 = as.numeric(lead$trt==3)
tfact <- as.numeric( factor(t, labels = 1:5))
lead_new <- data.frame(id,t,sex,ind.age,Trt1,Trt2,Trt3,tfact,Y)
```

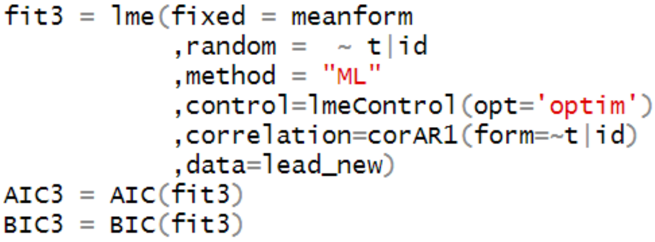
5.5 Model fitting code (for different error covariance structures)



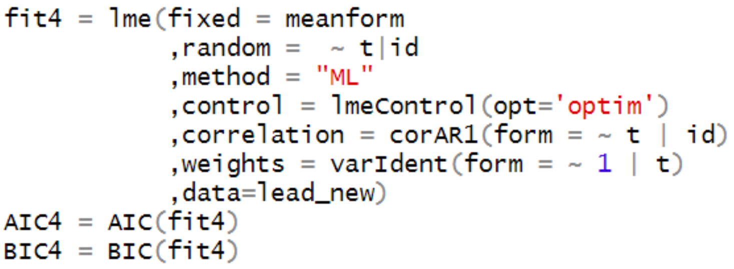
```
fit1 = lme(fixed = meanform
        ,random =  ~ t|id
        ,method = "ML"
        ,control = lmeControl(opt='optim')
        ,data=lead_new)
AIC1 = AIC(fit1)
BIC1 = BIC(fit1)
```



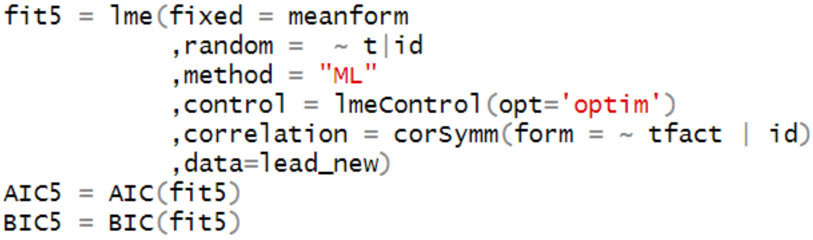
```
fit2 = lme(fixed = meanform
         ,random =  ~ t|id
         ,method = "ML"
         ,control=lmeControl(opt='optim')
         ,weights=varIdent(form=~1|t)
         ,data=lead_new)
AIC2 = AIC(fit2)
BIC2 = BIC(fit2)
```



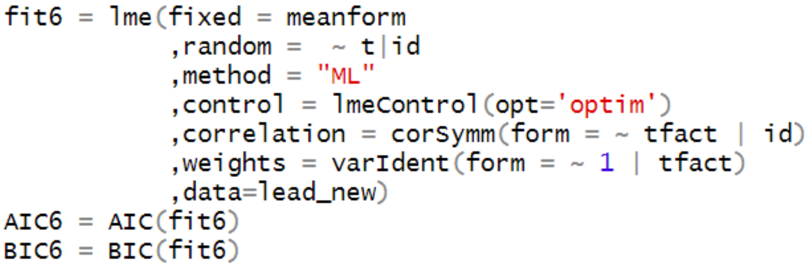
```
fit3 = lme(fixed = meanform
         ,random =  ~ t|id
         ,method = "ML"
         ,control=lmeControl(opt='optim')
         ,correlation=corAR1(form=~t|id)
         ,data=lead_new)
AIC3 = AIC(fit3)
BIC3 = BIC(fit3)
```



```
fit4 = lme(fixed = meanform
           ,random =  ~ t|id
           ,method = "ML"
           ,control = lmeControl(opt='optim')
           ,correlation = corAR1(form = ~ t | id)
           ,weights = varIdent(form = ~ 1 | t)
           ,data=lead_new)
AIC4 = AIC(fit4)
BIC4 = BIC(fit4)
```



```
fit5 = lme(fixed = meanform
           ,random =  ~ t|id
           ,method = "ML"
           ,control = lmeControl(opt='optim')
           ,correlation = corSymm(form = ~ tfact | id)
           ,data=lead_new)
AIC5 = AIC(fit5)
BIC5 = BIC(fit5)
```



```
fit6 = lme(fixed = meanform
           ,random =  ~ t|id
           ,method = "ML"
           ,control = lmeControl(opt='optim')
           ,correlation = corSymm(form = ~ tfact | id)
           ,weights = varIdent(form = ~ 1 | tfact)
           ,data=lead_new)
AIC6 = AIC(fit6)
BIC6 = BIC(fit6)
```

6. References

1) Longitudinal Data Analysis: Mixed Effects Models, Dr. Arnab Maity