Question #1
Complete the following table:

Number of repeated measurements	1 2 3 4 5
Number of subjects	7 6 5 3 91

Interpret this table while mentioning the issue of missing data in this study. What is this table telling us in terms of the 21 subjects that dropped out of the study and were lost on the follow-up?
Note: The R command **Bone Wide[!complete.cases(BoneWide),]** will return only those cases with missing data, where **BoneWide** is the name of the data set in wide form

- Answer: In longitudinal data, there is a high probability there will be missing data, especially with human studies where there are repeated measures (as with the study presented in this problem). Thus, it is important to look into which participants are missing, consider potential missing data mechanisms, and analyze the data accordingly. After the baseline measurement, 7 subjects did not show up to the first follow up measurement (with visit index 2). Then, after the first follow up measurement (with visit index 2), 6 participants dropped out and did not show up to the second follow up measurement (with visit index 3). 5 more participants dropped out after the second follow up (with visit index 3) and did not attend the third follow up (with visit index 4). Near the end of the study, an additional 3 participants dropped out and did not come to the final, fourth follow up (with visit index 5). Overall, the largest number of dropouts occurred after the baseline measurement, and this number gradually decreased.

Question #2
What do we need to assume about the subjects who were lost to follow-up in order for a complete case analysis to be valid with these data? That is to say, what type of missing data mechanism would not cause any issues for us with inference.

- Answer: For a complete case analysis to be valid with these data, we would need to assume that the type of missing data mechanism would be missing completely at random (MCAR) or missing at random (MAR). For example, if the missing data mechanism were MCAR, the subject's data could have been accidentally deleted or destroyed; alternatively, the subject could have had an incident that was completely out of control for the study, such as a personal issue or a car accident, that prevented them from attending their appointment. In the case of MAR, missing data due to this missing data mechanism is due, in part, to their observed and covariate values. For example, if a subject is in the "Under" or "Over" BMI category (and if BMI category is included as a covariate in a model for this study), they might be more likely to drop out because they

might not want to continue hearing the researchers making comments about their BMI. It is important to note that complete case methods are biased on MAR, but this is not too big of a bias, so conducting a complete case analysis should still be acceptable.

Now, what type of missing data mechanism would cause issues for our analysis? What could be causing this type of missing data?

- Answer: The missing data mechanism that would cause issues with our analysis is: not missing at random (NMAR); this is when the missing responses are missing due to the specific values of the response that are to be obtained, which is undesirable because this could result in a biased sample. For example, maybe the subjects who were not experiencing an increase in bone density after taking the calcium supplementation or experienced side effects due to the calcium supplementation dropped out of the study.

<u>Question #3</u>

Create a table of summary statistics of the frequencies in each category for categorical variables.

```
> # Summary for Trt, BMICat
> table(Bonewide$Trt)

 P  C
57 55
> table(Bonewide$BMICat)

Normal   Over  Under
    65     31     16
```

How many subjects are there in the control group (Placebo) and how many are there in the treatment group (Calcium).

- Answer: There are 57 subjects in the control group (Placebo) and 55 in the treatment group (Calcium).

Describe the distribution of the BMI (body mass index) categories counts.

- Answer: Most of the subjects have normal BMI (65 subjects); fewer have BMI in the over category (31 subjects), and even fewer in the under category (16 subjects).

Question #4

Create a table reporting the mean and standard deviation for age, mean and standard deviation for bone mineral density, and number of missing observations for each visit index by treatment.

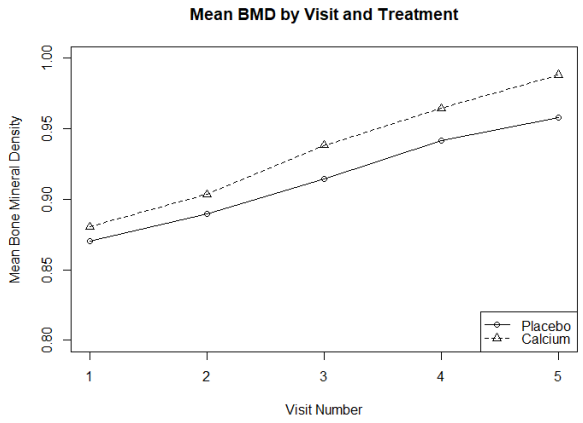
```
> # Mean, SD, and number missing by visit for Age and BMD:
>
> # Summary statistics for age:
> tapply(BoneLong$Age, list(BoneLong$Trt,BoneLong$visit), mean, na.rm=TRUE)
        1        2        3        4        5
P 11.05614 11.57170 12.02157 12.55417 13.03617
C 11.05455 11.56346 12.02500 12.55000 13.03864
> tapply(BoneLong$Age, list(BoneLong$Trt,BoneLong$visit), sd, na.rm=TRUE)
          1         2         3         4         5
P 0.1018008 0.10261459 0.09233358 0.1110076 0.1030524
C 0.1015038 0.09707253 0.10619132 0.1295291 0.1125103
> # Summary statistics for BMD:
> tapply(BoneLong$BMD, list(BoneLong$Trt,BoneLong$visit), mean, na.rm=TRUE)
          1         2         3         4         5
P 0.8700702 0.8896792 0.9141569 0.9416458 0.9582340
C 0.8804545 0.9032692 0.9382500 0.9645000 0.9881591
> tapply(BoneLong$BMD, list(BoneLong$Trt,BoneLong$visit), sd, na.rm=TRUE)
           1          2          3          4          5
P 0.06577012 0.07509936 0.07802830 0.07530519 0.07363931
C 0.05968895 0.05932731 0.05878214 0.06510120 0.06287194
> # Number missing:
> tapply(BoneLong$BMD, list(BoneLong$Trt,BoneLong$visit), na.count)
  1 2 3 4  5
P 0 4 6 9 10
C 0 3 7 9 11
```

Does it seem like the missingness pattern along the visits is similar for the Placebo and Calcium groups?

- Answer: Yes, the missingness pattern along the visits is similar for the Placebo and Calcium groups.

Question #5

Produce a plot of mean bone mineral density on the y-axis and visit number on the x-axis by treatment. Use different point characters and line types for the two treatments and include a legend.

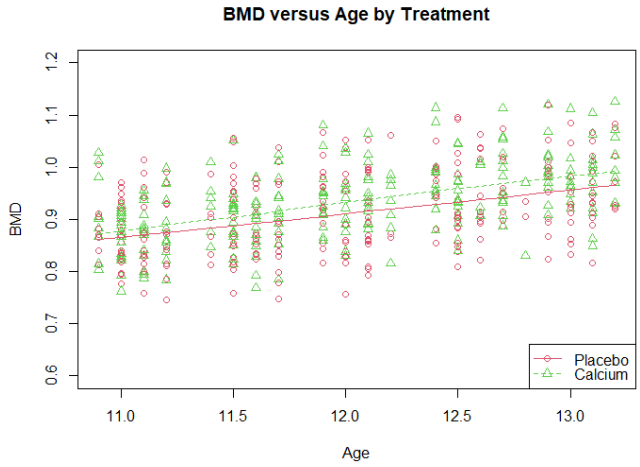


Describe what you see in this plot. Does there appear to be any differences in BMD between the Calcium and Placebo group across the visits?

- Answer: Both groups have an increase in mean bone mineral density across the visits. In addition, the Calcium group has slightly higher mean bone mineral density than the Placebo group across all visits.

Question #6

Produce a scatterplot of bone mineral density vs. age using different point characters for the two treatments. Add a smoother for each treatment.



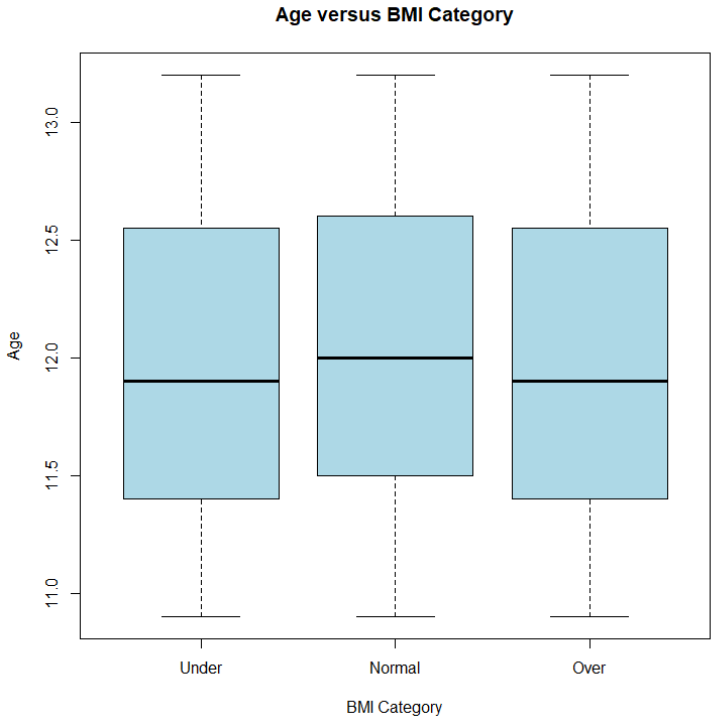
Question #7

Using the plots and summary statistics from parts 4 to 6., describe any trends in bone mineral density with age, differences between treatments, and variability among individuals that are suggested by the plots.

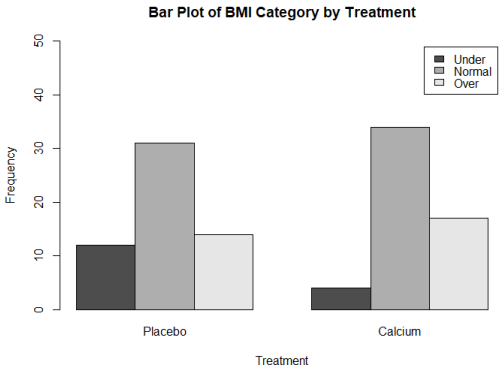
- Answer: According to the table in Question #4, the mean and standard deviation age of the Placebo and Calcium groups were very similar. On the other hand, the mean of the BMD of the Calcium group is slightly higher than that of the Placebo group across all visits, which is shown in the table in Question #4 as well as the plot in Question #5. However, it is not clear whether this difference in the means are significant, since there is quite a bit of overlap in the BMD of the Placebo and Calcium group across the visits in the scatterplot in Question #6. In addition, according to the table in Question #4, the standard deviation of the Placebo group is consistently higher than that of the Calcium group, suggesting slightly greater variability among individuals in the Placebo group.

Question #8

Create two (2) more well-labeled plots that are informative to the context of this study. Write a few sentences describing each plot.



- Plot Description: This is a boxplot that allows us to compare the Age of each BMI category in the study. The interquartile ranges of the boxplots for each of the BMI categories have quite a bit of overlap, and there does not appear to be a significant difference in Age between BMI categories.



- Plot Description: This is a bar plot showing the frequency of BMI categories within each of the treatment groups (Placebo and Calcium). The frequency of participants in each of the BMI categories appears to be fairly similar in both the Placebo and Calcium group. However, the Calcium group appears to have a lower frequency of participants in the 'Under' category for BMI.

Question #9

Using only the complete cases, fit the generalized least squares model

$$\text{BMD} \sim \text{Trt}+\text{I(Visit-1)}+\text{Trt*I(Visit-1)}$$

assuming a **compound symmetry covariance structure with constant variance**.
Note: We do I(Visit-1) to set our first visit (Visit-1) to a baseline Visit 0. *Also note*: I(.) in R is not the indicator function, it is the **identity** function. What are we doing is taking Visits and subtracting 1 from them (so Visit=1 is the baseline Visit 0 now).

(a) Describe the specification we are enforcing on the correlation of BMD across the visits when we use the compound symmetric covariance structure. What if we were using an AR1 (auto regressive order 1) covariance structure?

(i) When we use a compound symmetric covariance structure, we are assuming that the variances of the BMD and the correlation of the BMD across the visits are the same. If we were using an AR1 covariance structure, we would also assume that the variances are the same, but the correlation of the BMD across the visits decreases as the "spacing" between the visits increases; furthermore, two equally spaced visits have the same correlation. For example, the correlation between the BMD taken at visit 1 and visit 3 is the same as the correlation between the BMD taken at visit 2 and visit 4.

(b) Report the equation for the fitted mean bone mineral density. Use appropriate statistical notation and clearly define any variables used.

```
>
> mod1 = gls(BMD ~ Trt+I(visit-1)+Trt*I(visit-1), correlation=corCompSymm(, form=~visit | ID), method="REML", data=BoneLong.comp)
> summary(mod1)
Generalized least squares fit by REML
  Model: BMD ~ Trt + I(visit - 1) + Trt * I(visit - 1)
  Data: BoneLong.comp
       AIC       BIC   logLik
  -2046.5 -2021.831 1029.25

Correlation Structure: Compound symmetry
 Formula: ~visit | ID
 Parameter estimate(s):
       Rho
0.9480044

Coefficients:
                      Value   Std.Error  t-value p-value
(Intercept)       0.8699447 0.009751713 89.20942  0.0000
TrtC              0.0123462 0.014024109  0.88036  0.3791
I(visit - 1)      0.0226894 0.000710605 31.92965  0.0000
TrtC:I(visit - 1) 0.0043243 0.001021933  4.23146  0.0000

 Correlation:
                  (Intr) TrtC   I(v-1)
TrtC              -0.695
I(visit - 1)      -0.146  0.101
TrtC:I(visit - 1)  0.101 -0.146 -0.695

Standardized residuals:
        Min         Q1        Med         Q3        Max
-2.1418111 -0.7471617 -0.1227180  0.6908222  2.4032626

Residual standard error: 0.06756064
Degrees of freedom: 455 total; 451 residual
```

(i)

(ii) $\widehat{Y_{ij}} = 0.8699447 + 0.0123462Trt_i + 0.0226894I(v_{ij} - 1) + 0.0043243Trt_i * I(v_{ij} - 1)$

(1) Trt_i is the treatment group the i[th] individual is assigned to. If the participant is assigned to Calcium, $Trt_i = 1$. If the participant is assigned to Placebo, $Trt_i = 0$

(2) v_{ij} is the visit number of the i[th] individual on the j[th] visit (1, 2, 3, 4, 5)

(c) For each of the estimated β coefficients, write a sentence interpreting the value in context of the problem

(i) $\widehat{\beta}_0$: The estimated baseline BMD of a subject in the Placebo Group is 0.8699447 g/cm^3.

(ii) $\widehat{\beta}_1$: The estimated BMD for the Calcium group is 0.0123462 g/cm^3 higher than that of the placebo group.

(iii) $\widehat{\beta}_2$: For every increase in visits by 1 visit, the predicted value of BMD increases by 0.0226894 g/cm^3.

(iv) $\widehat{\beta}_3$: The estimated BMD for the Calcium group increases by an additional 0.0043243 g/cm^3 as the number of visits increases by 1 visit.

(d) Use the fitted model to calculate the estimated difference in mean bone mineral density between the two treatments on the first visit, and the estimated difference on the last visit.

(i) Placebo group:

(1) mean BMD for first visit:

$0.8699447 + 0.0123462(0) + 0.0226894(1-1) + 0.0043243(0) * (1-1)$
$= 0.8699447\ g/cm^3$

(2) mean BMD for last visit:

$0.8699447 + 0.0123462(0) + 0.0226894(5-1) + 0.0043243(0) * (5-1)$
$= 0.8699447 + 0.0226894(4) = 0.9607023\ g/cm^3$

(3) estimated difference:

$(mean\ BMD\ for\ last\ visit) - (mean\ BMD\ for\ first\ visit)$
$= 0.9607023 - 0.8699447 = 0.0907576\ g/cm^3$

(ii) Calcium group:

(1) mean BMD for first visit:

a) $0.8699447 + 0.0123462(1) + 0.0226894(1-1) + 0.0043243(1) * (1-1)$
$= 0.8699447 + 0.0123462(1) = 0.8822909\ g/cm^3$

(2) mean BMD for last visit:

a) $0.8699447 + 0.0123462(1) + 0.0226894(5-1) + 0.0043243(1) * (5-1)$
$= 0.8699447 + 0.0123462(1) + 0.0226894(4) + 0.0043243(1) * (4)$
$= 0.9903457\ g/cm^3$

(3) estimated difference:

a) $(mean\ BMD\ for\ last\ visit) - (mean\ BMD\ for\ first\ visit)$
$0.9903457 - 0.8822909 = 0.1080548\ g/cm^3$

(e) Is there any evidence that the change in mean bone mineral density across visits differs between the two treatments? If so, how does it differ? Support your answer with appropriate statistical inference/tests.

(i) $H_0: \beta_1 = 0$

(ii) $H_a: \beta_1 \neq 0$

(iii) p-value = 0.3791

(iv) Conclusion: At a 5% significance level, we fail to reject the null and conclude that $\beta_1 = 0$. Thus, there is no evidence that mean bone mineral density across visits differs between the two treatments.

(f) Report the estimated marginal variance-covariance matrix of the responses, that is $cov(Y_{ij}, Y_{ik})$ for j, k = 1, 2, 3, 4, 5.

(i)

```
> getVarCov(mod1)
Marginal variance covariance matrix
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.0045644 0.0043271 0.0043271 0.0043271 0.0043271
[2,] 0.0043271 0.0045644 0.0043271 0.0043271 0.0043271
[3,] 0.0043271 0.0043271 0.0045644 0.0043271 0.0043271
[4,] 0.0043271 0.0043271 0.0043271 0.0045644 0.0043271
[5,] 0.0043271 0.0043271 0.0043271 0.0043271 0.0045644
  Standard Deviations: 0.067561 0.067561 0.067561 0.067561 0.067561
```

(g) Perform a likelihood ratio test to determine if the variance in bone mineral density differs across visits or we can use constant variance across visits. Write the null and alternative hypothesis, state the p-value, and make a conclusion assuming a 5% significance level.

(i)

```
> anova(mod1, mod2)
     Model df       AIC       BIC   logLik   Test  L.Ratio p-value
mod1     1  6 -2046.500 -2021.831 1029.250
mod2     2 10 -2057.456 -2016.341 1038.728 1 vs 2 18.95576   8e-04
```

(1) H_0: the reduced model (model with constant variance across visits) is good enough

(2) H_a: the full model (model with nonconstant variance) is good enough

(3) p-value = 8e-04

(4) Conclusion: At a 5% significance level, we reject the null and conclude evidence for the alternative that the model with nonconstant variance is good enough.

Question #10

Using all available data (use argument **na.action = na.omit** in the **lme** function), we now will fit a linear mixed effects model (LME) where:

- the marginal mean of BMD varies with age,
- each treatment can have a different marginal mean BMD trajectory both in intercept and in slope (with respect to age), so that is to say treatment has a fixed effect and an interaction effect with age,
- the model is **not** adjusted for BMI category (that is to say BMI is not in the model in any way), and
- subject-specific mean response trajectories can differ in intercept but not in slope (random intercept only)

(a) Write out the theoretical notational form of the (population) model we assume above (you may use variable definitions from part 9.). State all model assumptions and distributional specifications

(i) $Y_{ij} = \beta_0 + b_{0i} + \beta_1 a_{ij} + \beta_2 Trt_i + \beta_3 a_{ij} Trt_i + \varepsilon_{ij}$

(ii) Model assumptions and distributional specifications:

(1) ε_{ij} is independently distributed from a normal distribution with mean 0 and variance σ^2

(2) The random intercept (b_{0i}) follows a normal distribution with mean 0 and variance σ_0^2

(3) ε_{ij} and b_{0i} are independent

(b) Fit the LME model described above and report the output from R

(i)

```
> mod3 = lme(BMD ~ Trt+Age+Trt*Age, random = ~ 1|ID, data=BoneLong, na.action=na.omit, method="ML")
> summary(mod3)
Linear mixed-effects model fit by maximum likelihood
  Data: BoneLong
       AIC       BIC   logLik
  -2278.122 -2252.822 1145.061

Random effects:
 Formula: ~1 | ID
        (Intercept)   Residual
StdDev:  0.06653545 0.01501359

Fixed effects:  BMD ~ Trt + Age + Trt * Age
                Value  Std.Error  DF  t-value p-value
(Intercept)  0.3631525 0.01874861 387 19.36956  0.0000
TrtC        -0.0863222 0.02683117 110 -3.21724  0.0017
Age          0.0457723 0.00138325 387 33.09041  0.0000
TrtC:Age     0.0088754 0.00198223 387  4.47751  0.0000
 Correlation:
         (Intr) TrtC   Age
TrtC     -0.699
Age      -0.880  0.615
TrtC:Age  0.614 -0.881 -0.698

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-3.36572939 -0.59950246  0.02745521  0.62033724  2.68297920

Number of Observations: 501
Number of Groups: 112
```

(c) Write a couple of sentences interpreting the estimated standard deviation of the random intercepts in context of the problem

(i) Since the estimated standard deviation of the random intercept is 0.06653545, this indicates that $\sigma_0^2 > 0$; in turn, this implies that the subjects in this study will each have a unique random intercept, since the variance of the random intercept term is not 0.

(d) Perform a likelihood ratio test to determine if a random slope on Age should be added to the model. Write the null and alternative hypothesis, state the p-value, and make a conclusion assuming a 5% significance level.

(i)

```
> anova(mod3, mod3s)
      Model df       AIC       BIC   logLik   Test  L.Ratio p-value
mod3      1  6 -2278.122 -2252.822 1145.061
mod3s     2  8 -2371.392 -2337.659 1193.696 1 vs 2 97.26974  <.0001
```

(1) H_0: the reduced model (model with just a random intercept) is good enough

(2) H_a: the full model (model with a random intercept and random slope on Age) is good enough

(3) p-value < 0.0001

(4) Conclusion: At a 5% significance level, we reject the null and conclude evidence for the alternative that the full model (the model with a random intercept and random slope on Age) is good enough.

(e) Using the model you decided is best from the previous part d, perform a likelihood ratio test to determine if we should adjust for BMI category in our analysis (fixed effect). Write the null and alternative hypothesis, state the p-value, and make a conclusion assuming a 5% significance level.

(i)

```
> mod4.0 = lme(BMD ~ Trt+Age+Trt*Age + BMICat, random = ~ 1 + Age|ID, data=BoneLong, na.action=na.omit, method="ML")
>
> mod4.1 = lme(BMD ~ Trt+Age+Trt*Age, random = ~ 1 + Age|ID, data=BoneLong, na.action=na.omit, method="ML")
>
> anova(mod4.0, mod4.1)
      Model df       AIC       BIC   logLik   Test L.Ratio p-value
mod4.0     1 10 -2408.607 -2366.441 1214.304
mod4.1     2  8 -2371.392 -2337.659 1193.696 1 vs 2 41.2159  <.0001
```

(ii) H_0: the reduced model (model without the BMI category as a fixed effect) is good enough

(iii) H_a: the full model (model with the BMI category as a fixed effect) is good enough

(iv) p-value < 0.0001

(v) Conclusion: At a 5% significance level, we reject the null and conclude evidence for the alternative that the full model (the model with the BMI category as a fixed effect) is good enough.

(f) Explain in a couple of sentences why you need to first transform the residuals from a LME model before doing any residual diagnostics

(i) Explanation: Residuals can be used to check for any systematic departures from the model for the mean responses. For longitudinal data, it is a good practice to transform residuals beforehand because residuals are correlated in longitudinal data. That being the case, if residuals are not transformed prior, it may appear that our model is failing because we are seeing a pattern in the residual plots. To truly see whether there is a trend in the residual plots, we need to transform the residuals first to normalize them.

Question #11

Discuss any limitations of the study and suggest a better study design than the completely randomized design used by the researchers. Consider a the topic of the demographics of the sample of subjects used and what our goal is with the study and what population we are trying to address.

- Discussion: Since the purpose of this study is to examine the effect of calcium supplementation on bone mineral density in young children (less than 13 years old), I would suggest that the researchers recruit not just female, but also male participants in local school districts. In addition, since the participants were from primarily white middle-class neighborhoods, to ensure that the results of this study are more generalizable, it may help to select participants from more racially and socioeconomically diverse neighborhoods. Furthermore, since most of the participants were close to or 11 years old, it may help to recruit participants with a broader age range, since the population we are trying to address is young children (less than 13 years old), rather than just 11-year-olds.

Question #12

Using the LME model you decide is best from part 10, write a few sentences summarizing what this study shows on the effect of calcium supplementation on bone mineral density among adolescent girls. Be clear in what LME (linear mixed effect) model you are using for this summary.

- Chosen LME Model (theoretical notation form of population model, which is to be estimated):
  - $Y_{ij} = \beta_0 + b_{0i} + \beta_1 a_{ij} + b_{1i} a_{ij} + \beta_2 Trt_i + \beta_3 a_{ij} Trt_i + \beta_4 BO_i + \beta_5 BU_i + \varepsilon_{ij}$
- Estimates for the chosen LME model after it is fitted using R:

```
> mod4.0 = lme(BMD ~ Trt+Age+Trt*Age + BMICat, random = ~ 1 + Age|ID, data=BoneLong, na.action=na.omit, method="ML")
> summary(mod4.0)
Linear mixed-effects model fit by maximum likelihood
  Data: BoneLong
        AIC      BIC    logLik
  -2408.607 -2366.441 1214.304

Random effects:
 Formula: ~1 + Age | ID
 Structure: General positive-definite, Log-Cholesky parametrization
             StdDev     Corr
(Intercept) 0.15204716 (Intr)
Age         0.01325883 -0.937
Residual    0.01111265

Fixed effects:  BMD ~ Trt + Age + Trt * Age + BMICat
                 Value  Std.Error  DF   t-value p-value
(Intercept)  0.3682059 0.02524893 387 14.583032  0.0000
TrtC        -0.0975534 0.03549086 108 -2.748690  0.0070
Age          0.0451775 0.00214456 387 21.066081  0.0000
BMICatOver   0.0573764 0.01188129 108  4.829138  0.0000
BMICatUnder -0.0584358 0.01545131 108 -3.781930  0.0003
TrtC:Age     0.0088406 0.00306681 387  2.882661  0.0042
 Correlation:
            (Intr) TrtC   Age    BMICtO BMICtU
TrtC        -0.692
Age         -0.939  0.668
BMICatOver  -0.147 -0.005  0.000
BMICatUnder -0.155  0.055 -0.002  0.244
TrtC:Age     0.657 -0.955 -0.699 -0.002  0.000

Standardized within-Group Residuals:
        Min          Q1         Med          Q3         Max
-2.08987618 -0.58418315 -0.01048876  0.52196540  2.31242873

Number of Observations: 501
Number of Groups: 112
```

- Summary: The estimated BMD for the Calcium group is 0.0975534 g/cm^3 lower than that of the Placebo group, while adjusting for Age and BMI category. In addition, the estimated BMD for the Calcium group increases an additional 0.0088406 g/cm^3 as the age of the participants increases by 1 year (while adjusting for Age and BMI category).