Section I

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

This table highlights the participation attrition trend in a longitudinal bone density study. Of the 112 total subjects, 91 completed all five measurements, while 21 subjects dropped out before the final measure, with fewer participants leaving after each successive measurement. The higher participant retention toward the later stages of the study may be a result of participants who faced more challenges (e.g., scheduling conflicts, long commute times, or anxiety) or a mismatch of expectations for the study, exiting earlier compared to those who were willing to continue. Longitudinal studies, especially human studies with repeated measures, have a high probability of missing data; the participant attrition in this study exemplifies this and suggests the potential for attrition bias in this study.

Section II

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.

For a complete case analysis to be valid, we would assume that the missing data mechanism is either:

- **missing completely at random (MCAR)**
- or **missing at random (MAR)**

MCAR would be the most ideal scenario, since this is less likely to result in biased analysis. In the case of MCAR, data would be missing due incidents unrelated to the measured variables, such as data being accidentally being deleted or destroyed, or circumstances unrelated to the measured variables that prevented the subject from joining the study.

In the case of **MAR**, the missing data is due, in part, to their observed and covariate values. However, the data available could still provide valid inferences if the missingness is accounted and explained for in the analysis. For example, if the likelihood of participants continuing in the

study is related to characteristics such as gender and age, but is unrelated to their bone mineral density, this would be MAR.

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?

The missing data mechanism that would cause issues with our analysis is: **missing not at random (MNAR)**, when the missing responses are due to events or variables that are not measured by the researcher. For example, if subjects who were not experiencing an increase in bone density are more likely to drop out of the study early. This is undesirable because this could result in a biased sample.

Section III

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).

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

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

Most of the subjects have a normal BMI (65 subjects); fewer are classified as overweight (31 subjects), and the number of subjects classified as underweight was the smallest (16 subjects).

Section IV

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.

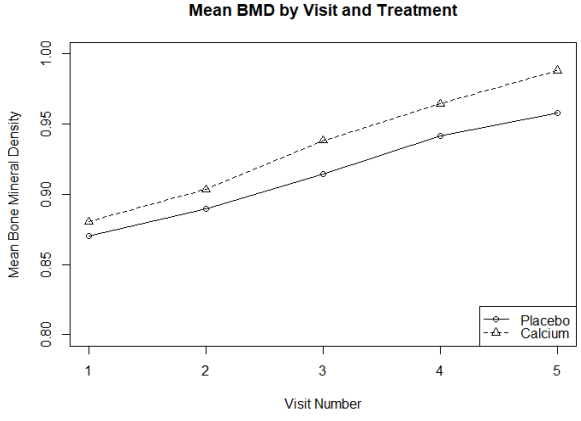
```
> # Mean, SD, and number missing by visit for Age and BMD:
>
> # Summary statistics for age:
> tapply(BoneLong$Age, list(BoneLong$Trt,BoneLong$Visit), mean, na.rm=TRUE)
        1        2        3        4        5
P 11.05614 11.57170 12.02157 12.55417 13.03617
C 11.05455 11.56346 12.02500 12.55000 13.03864
> tapply(BoneLong$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?

Yes, it seems like there is a nondifferential missingness pattern between the Placebo and Calcium groups. In both conditions, the number of missing responses increases monotonically after each successive measure. The similar pattern suggests the missingness data mechanism for this study is **MCAR** or **MAR**.

Section V

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.

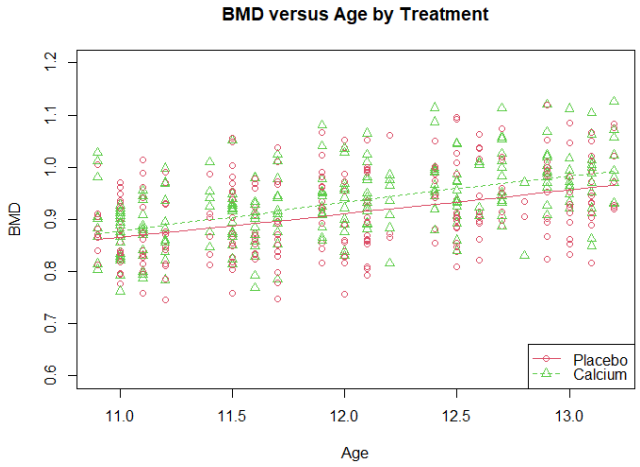


Describe what you see in this plot. Does there appear to be any differences in BMD between the Calcium and Placebo group across the visits?

Both groups have an increase in mean bone mineral density across all visits. However, the Calcium group has a higher average bone mineral density than the Placebo group across all visits.

Section VI

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



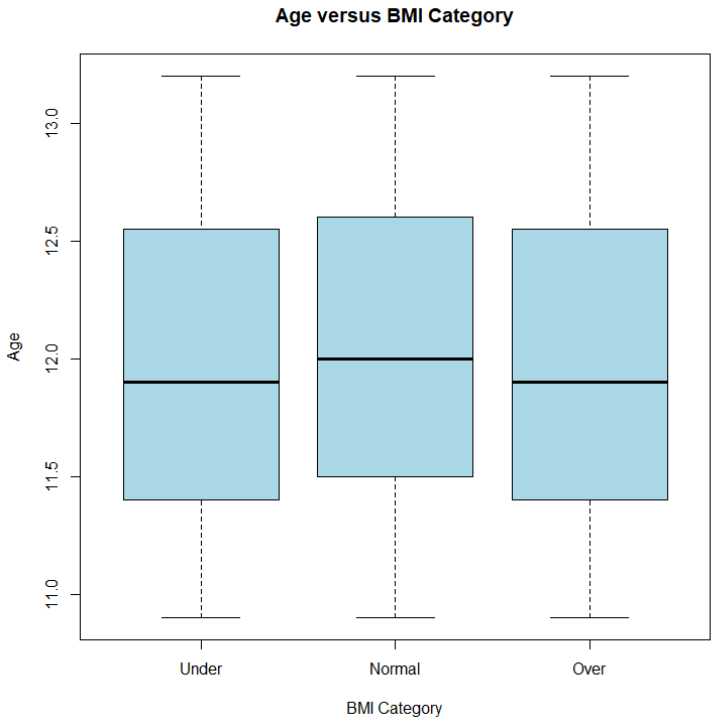
Section VII

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.

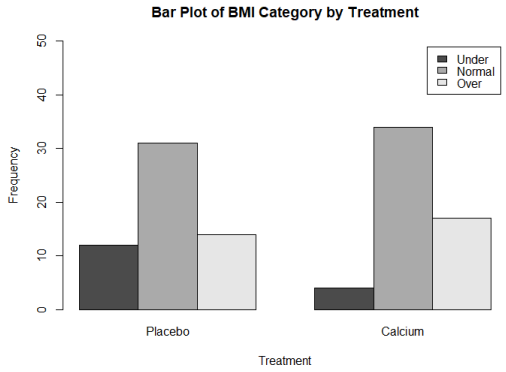
The mean and standard deviation of age were very similar for the Placebo and Calcium groups, suggesting there is less potential for a confounding effect due to age. For our variable of interest, bone mineral density (BMD), the Calcium group had a higher mean BMD than the Placebo group across all visits (Section VI) and across all ages (Section VII). However, the statistical significance of this difference should be assessed, especially given the overlapping BMD values between the groups across visits in the Section VI scatterplot. In addition, the standard deviation of the BMD was consistently higher for the Placebo group, indicating greater variability among individuals compared to the Calcium group. This suggests a potential stabilizing effect of calcium supplementation on BMD. However, further analysis is needed to determine whether this difference is statistically significant.

Section VIII

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



This side-by-side boxplot graph allows us to compare the distribution of **Age** across the three BMI categories. The interquartile ranges of the boxplots overlap, and the medians are similar across categories, suggesting that the age distributions are similar across the three different BMI categories.



The bar plot shows the frequency of BMI categories within each of the treatment groups (Placebo and Calcium). Overall, the distribution of participants across BMI categories appears similar between the Placebo and Calcium group. However, the Calcium group has a lower frequency of participants classified in the ‘Under’ BMI category and slightly higher frequencies in

the ‘Normal’ and ‘Over’ BMI categories. If BMI was measured when the study concluded, this may reflect changes in BMI over time, potentially due to calcium supplementation.

Section IX

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

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

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).

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?

When we use a compound symmetric covariance structure, we assume that the variances of BMD are the same for all visits and the correlations between the BMD measurements for any pair of visits are identical. In contrast, for an AR1 covariance structure, we also assume equal variances across visits, but the correlation between BMD measurements decreases as the number of visits apart (i.e., the “space” between the visits) increases. Furthermore, pairs of visits that are equally spaced have the same correlation. For example, the correlation between BMD measured at visits 1 and 3 is the same as that between visits 2 and 4.

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
Trt               0.0123462 0.014024109  0.88036  0.3791
I(Visit - 1)      0.0226894 0.000710605 31.92965  0.0000
Trt:I(Visit - 1) 0.0043243 0.001021933  4.23146  0.0000

 Correlation:
                  (Intr) Trt    I(V-1)
Trt               -0.695
I(Visit - 1)      -0.146  0.101
Trt: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
```

Fitted mean model

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

Treatment groups

Trt_i is the treatment group the i[th] individual is assigned to.

- $Trt_i = 1$ if the participant is assigned to the **Calcium** treatment group.
- $Trt_i = 0$ if the participant is assigned to the **Placebo** treatment group.

Visit Number

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

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

Interpretation of coefficients

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

- $\widehat{\beta}_1$: For visit 1, the estimated BMD of the Calcium group is 0.0123462 g/cm^3 higher than the Placebo group's estimated BMD.
- $\widehat{\beta}_2$: The estimated BMD of Placebo subjects increases by 0.0226894 g/cm^3 per visit.
- $\widehat{\beta}_3$: The Calcium group's estimated BMD increases by an additional 0.0043243 g/cm^3 relative to the Placebo group.

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.

Estimated mean BMD of the Placebo group:

Mean BMD for first visit**:**

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

Mean BMD for last visit:

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

Estimated difference:

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

Estimated mean BMD of the Calcium group:

Mean BMD for first visit:

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

Mean BMD for last visit:

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

Estimated difference:

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

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.

- $H_0: \beta_1 = 0$
- $H_a: \beta_1 \neq 0$
- p-value = 0.3791

At a 5% significance level, we fail to reject the null hypothesis and conclude that $\beta_1 = 0$. There is no statistically significant evidence that mean bone mineral density across visits differs between the Calcium and Placebo treatments.

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.

```
> 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
```

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.

```
> 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
```

- H_0: The reduced model with constant variance across visits is sufficient.
- H_a: The full model with nonconstant variance is sufficient.
- p-value = 8×10^{-4}

At a 5% significance level, we reject the null hypothesis and conclude evidence for the alternative hypothesis that the model with nonconstant variance is sufficient.

Section X

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)**

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

Fitted model

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

Model assumptions and distributional specifications

- The ε_{ij} is independently distributed from a normal distribution with mean 0 and variance σ^2.
- The random intercept (b_{0i}) follows a normal distribution with mean 0 and variance σ_0^2.
- The ε_{ij} and b_{0i} are independent.

Fit the LME model described above and report the output from R

```
> 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
```

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

Since the estimated standard deviation of the random intercept is 0.06653545, this indicates that $\sigma_0^2 > 0$, which suggests that the subjects each have a unique random intercept (i.e., distinct baseline BMD).

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.

```
> 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
```

- H_0: The reduced model with just a random intercept is sufficient.
- H_a: The full model with a random intercept and random slope on Age is sufficient.
- p-value < 0.0001

At a 5% significance level, we reject the null and conclude evidence for the alternative that the full model with subject-specific age effects is sufficient.

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.

```
> 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
```

- H_0: The reduced model without the BMI category as a fixed effect is sufficient.
- H_a: The full model with the BMI category as a fixed effect is sufficient.
- p-value < 0.0001

At a 5% significance level, we reject the null and conclude evidence for the alternative that the full model with the BMI category as a fixed effect is sufficient.

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

Residuals can be used to check for any systematic departures mean responses have from the model. For longitudinal data, residuals are usually correlated within subjects, and the data may have patterns related to within-subject correlation. Thus, it is a good practice to transform residuals beforehand to account for this within-subject correlation.

Section XI

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.

Since this study examined the effect of calcium supplementation on bone mineral density in young children under the age of thirteen, for future studies, male participants could be recruited in addition to female participants to account for potential gender effects. Additionally, because most participants were from middle-class neighborhoods, sampling from a broader range of socioeconomic backgrounds would improve the generalizability of findings. Furthermore, since most of the participants were close to or 11 years old, recruiting a broader age range of participants would better represent the target population of young children and allow potential age effects to be considered.

Section XII

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):

Fitted LME model

To examine the effect of calcium bone supplementation on bone mineral density (BMD) while accounting for within-subject variability in Age and BMI, this LME model was fitted with Age as a random effect and BMI as a fixed effect.

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1 a_{ij} + b_{1i} a_{ij} + \beta_2 Trt_i + \beta_3 a_{ij} Trt_i + \beta_4 BO_i + \beta_5 BU_i + \varepsilon_{ij}$$

```
> 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
```

After fitting this model in R (adjusting for Age and BMI):

- The estimated BMD for the Calcium group is 0.0975534 g/cm^3 lower than that of the Placebo group.
- However, the BMD in the calcium group increased by an additional 0.0088406 g/cm^3 per one-year increase in age relative to the Placebo group.

Based on these estimates, the net effect of calcium supplementation on adolescent girls is positive approximately at the age of eleven. This suggests that calcium supplementation may be more effective at increasing BMD for female participants around this age, while less effective for younger female adolescents.