Stats 112 Final

Christopher Hong

2023-06-12

Number of Repeated Measurements	1	2	3	4	5
Number of Subjects	7	6	5	3	91

1.

This table displays the number of subjects grouped by the amount of repeated measurements that were taken on them. 91 of the 112 subjects had all five measurements taken on them. The remaining 21 subjects dropped out of the study and were lost on the follow up, resulting in missing data.
The table shows that, of the 21 subjects that dropped out, 7 had only their first measurement taken, 6 had only the first two taken, 5 had the first three taken, and 3 had the first four taken. Respectively, this means that the 7 are missing 4 observations each, the 6 are missing 3 observations each, the 5 are missing 2 observations each, and the 3 are missing 1 observation each.

2. In order for complete case analysis to be valid with these data, we need to assume that the missing data is missing completely at random (MCAR). If our MCAR assumption is true, our parameter estimates will be unbiased and we should not experience any issues with inference due to the missing data.
   The types of missing data mechanisms that would cause issues for our analysis would be missing at random (MAR) and not missing at random (NMAR). In complete case analysis, either could lead to potential bias in our parameter estimates.
   In this study, the MAR case could potentially arise due to the parents of the subject being more private. They may not want their children to be participating in a study or having their biological factors being recorded, and decide that they no longer want to be a part of the study and drop out as a result.
   The NMAR case could potentially arise due to the parents of the subject not wanting their child to a part of the study because of their child's BMI being too high (overweight). They may have seen their child being categorized as being in the "Over" category, and decided to drop out because of that.

3. Subject Counts by Treatment Group:

```
##
##  P  C
## 57 55
```

P denotes the placebo (control) group and C denotes the Calcium (treatment) group. From the table, there are 57 subjects in the control group (placebo). There are 55 subjects in the treatment group (calcium).

Subject Counts by BMI Category:

```
##
## Normal   Over  Under
##     65     31     16
```

From the above table, we see that the majority of subjects in the study were categorized as having normal BMI. About half this number were categorized as being over the normal BMI, and about half of that number were categorized as being under the normal BMI. If we order the categories by increasing BMI (Under, Normal, Over), we see that the counts appear to be distributed left skew.

4. Let P denote the placebo group (control) and C denote the calcium group (treatment).

   Mean of Age at Each Visit for Each Treatment Group:

```
##          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
```

Standard Deviation of Age at Each Visit for Each Treatment Group:

```
##           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
```

Mean of Bone Mineral Density at Each Visit for Each Treatment Group:

```
##           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
```

Standard Deviation of Bone Mineral Density at Each Visit for Each Treatment Group:

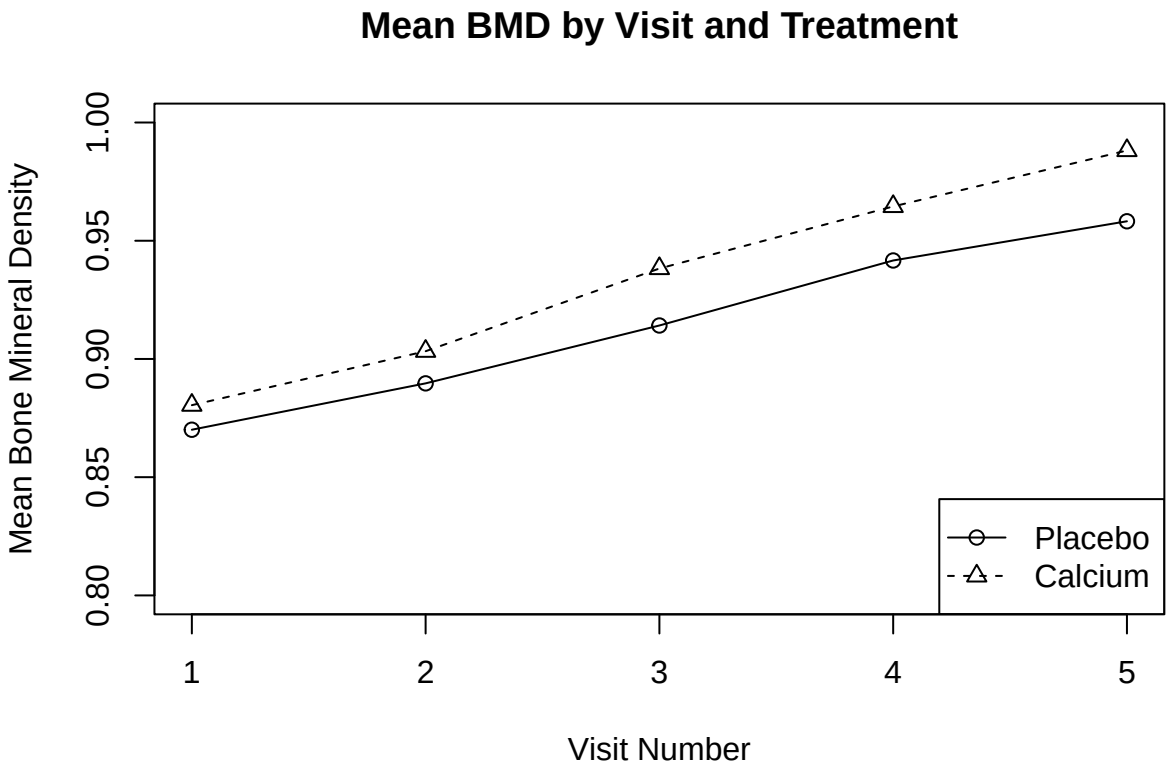
```
##            1          2          3          4          5
## P 0.06577012 0.07509936 0.07802830 0.07530519 0.07363931
## C 0.05968895 0.05932731 0.05878214 0.06510120 0.06287194
```

Number of Missing Observations for Each Visit for Each Treatment Group:

```
##   1 2 3 4  5
## P 0 4 6 9 10
## C 0 3 7 9 11
```

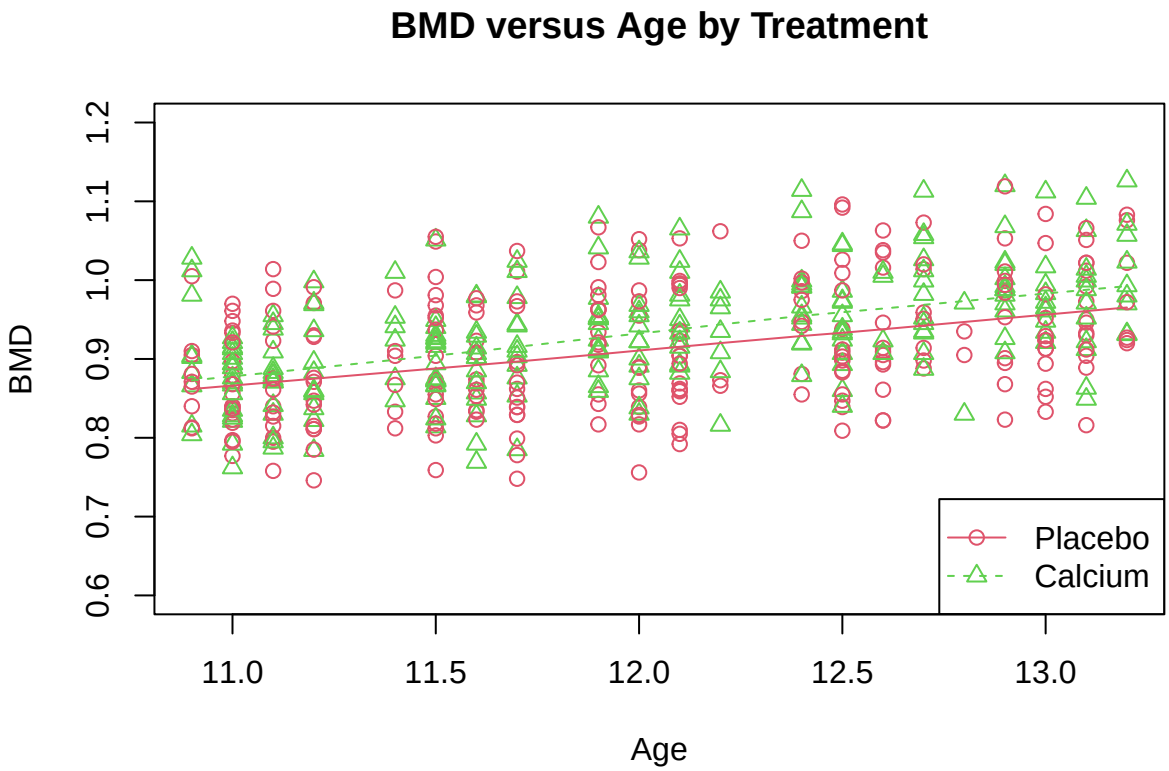
It seems like the missingness pattern along the visits is similar for both the Placebo and Calcium groups, as the number of missing observations for each visit between treatment groups differs by at most 1.

5.



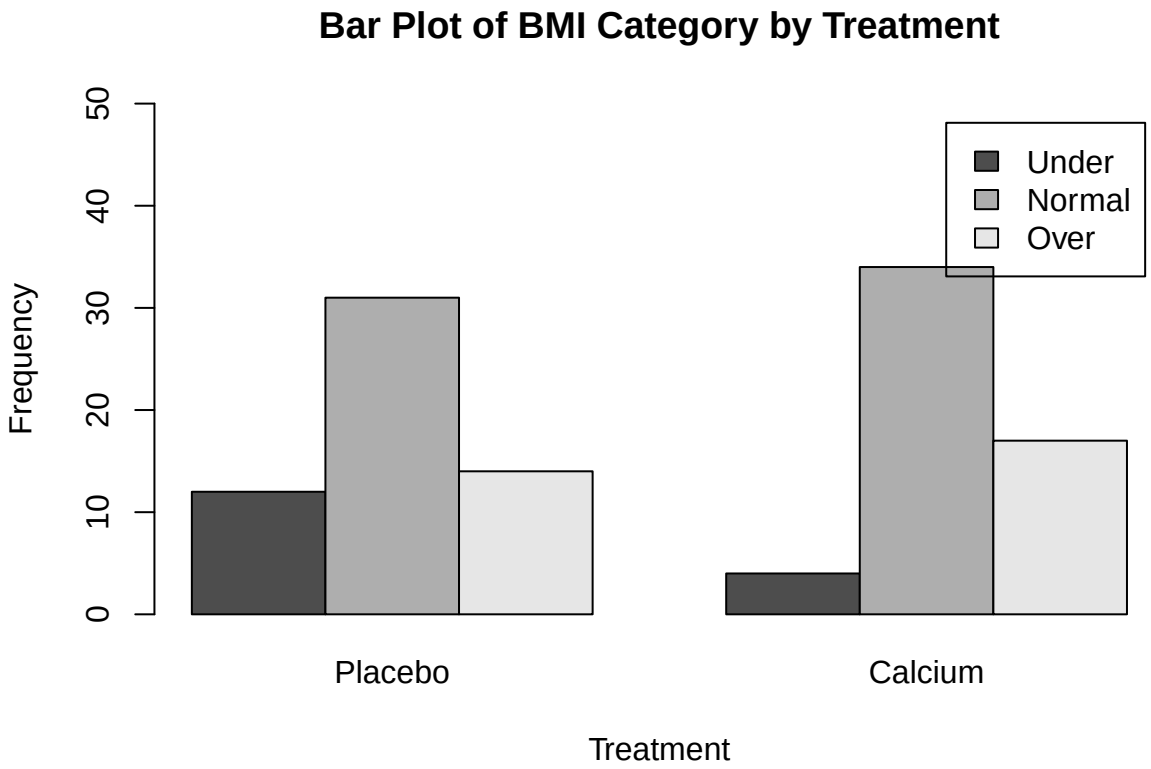
The plot above shows that as visit number increases, the mean bone mineral density for both groups also increases. For all visits, the calcium (treatment) group has higher mean bone mineral density than the placebo (control) group. So, across the visits, it appears that the BMD for the Calcium group is higher than the Placebo group.

6.

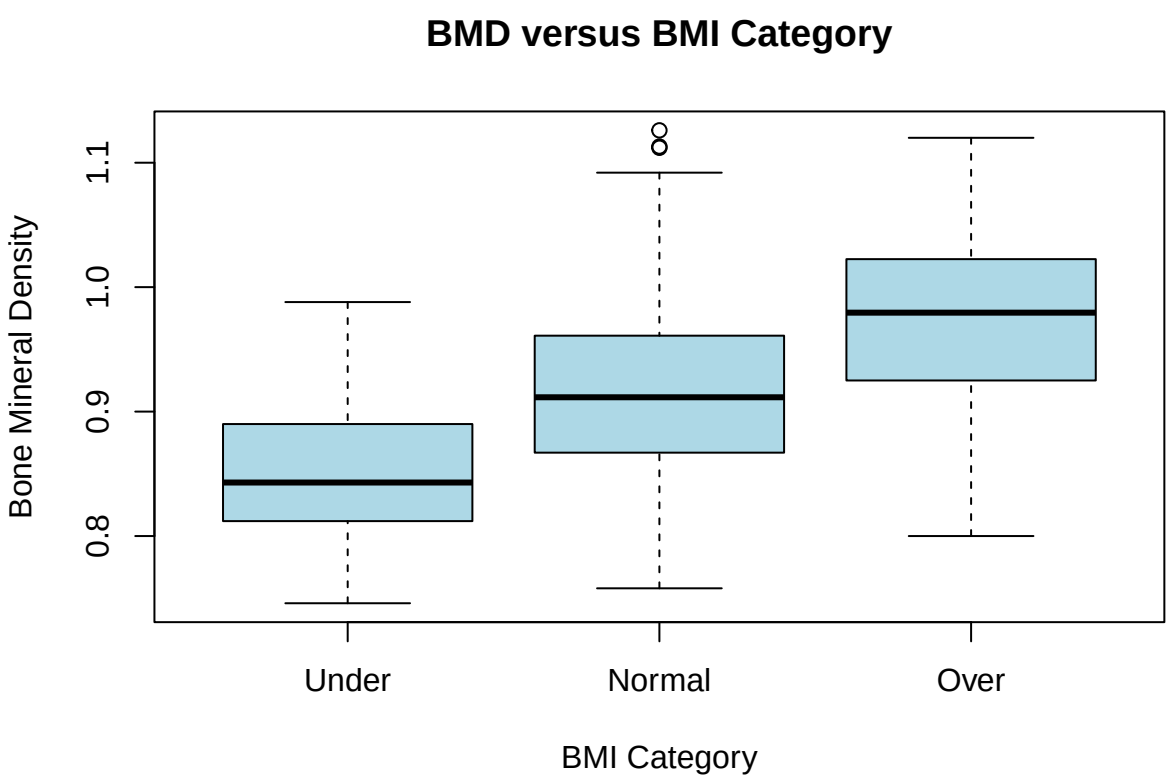


7. The table of mean BMD by visits from part 4, the plot of Mean BMD by Visit and Treatment from part 5, and the scatter plot of BMD versus Age by Treatment from part 6 all suggest that bone mineral density increases with age. Part 4 and 5 appear to show that mean BMD increases with visit number for both treatment groups, and subjects increase in age between visits. Part 6 suggest a positive trend between age and BMD for both treatment groups.
The table of mean BMD by visit, the plot of Mean BMD by Visit and Treatment, and the scatter plot of BMD versus Age by Treatment also suggest that the BMD for subjects in the Calcium (treatment) group is higher than for subjects in the Placebo (control) group. These all show that the mean BMD of the treatment group is higher than that of the control group at each visit/controlling for age.
The table of the Standard Deviations for BMD by Visit and Treatment Group appear to show that the variability of BMD in the placebo group is higher than that of the calcium group, as its standard deviations are greater than those of the treatment group at a given visit. The greater variability of the placebo group can also be seen in the scatter plot from part 6, as the calcium group's BMD measurements appear to be closer grouped together accounting for age.

8.



The plot above shows the counts of subjects in each BMI category by treatment. We see that in the placebo group, there is a much greater number of subjects that fall into the underweight category than in the calcium group. In the placebo group, there are slightly fewer counts of subjects in both the normal and overweight categories than in the calcium group, but these differences are not as large as the difference in counts of underweight subjects between the calcium and placebo groups. This may be a problem, as BMI categorization could be a potential confounder if not controlled for, since the proportion of counts of each BMI category differ between the two groups.




```
## Normal   Over  Under
## 0.9115 0.9795 0.8430
```

This box plot depicts the distribution of BMD for each BMI category, not controlling for treatment group or visit. It shows that more than three quarters of underweight individuals have BMD lower than the median BMD of the normal weight group. It also shows that three quarters of overweight individuals have BMD higher than than the median BMD of the normal weight group. This suggests that there may be a positive relationship between BMI and bone mineral density, where a higher BMI is correlated with a higher BMD. This may be an issue with the study, as BMI category could be a potential confounder if not controlled for.

9.

a. By using a compound symmetry covariance structure, we are specifying that the correlation of BMD across the visits stays constant over increasing time separation. That is, the correlation of BMD across any two visits is the same no matter the number of visits between those observations.
If we were using a AR1 (auto regressive order 1) covariance structure, we would be specifying that the correlation of BMD across the visits declines over time as the distance between visits increases. Additionally, the correlation between two equally spaced visits would have the same correlation using AR1. For example, visit 2 and visit 5 would have the same correlation as visit 1 and visit 4.

b. $E[Y_{ij}] = \hat{Y}_{ij} = 0.8699447 + 0.0123462 * Trt_i + 0.0226894 * v_{ij} + 0.0043243 * Trt_i * v_{ij}$
Where $\hat{Y}_{ij}$ is the estimated mean bone mineral density for the i-th subject on the j-th visit,
Trt_i is the treatment group, being 1 if the i-th subject is part of the calcium group and 0 if the i-th subject is part of the placebo group, and
v_{ij} is the visit number of the i-th individual on the j-th visit, with 0 denoting the first visit.

c. $\hat{\beta}_0 = 0.8699447$ is the estimated value of BMD for individuals in the placebo group on the first visit.
$\hat{\beta}_1 = 0.0123462$ is the estimated difference in BMD between an individual in the placebo group and an

individual in the calcium group on the first visit. The estimated value of BMD for individuals in the calcium group on the first visit is $\hat{\beta}_0 + \hat{\beta}_1 = 0.8822909$.
$\hat{\beta}_2 = 0.0226894$ is the estimated difference in BMD of an individual in the placebo group between two adjacent visits. In other words, we estimate that a 1-unit increase in visit number would increase the expected BMD of an individual in the placebo group by 0.0226894.
$\hat{\beta}_3 = 0.0043243$ is the estimated interaction term between visit number and treatment group. That is, when comparing the calcium and placebo groups, a 1-unit increase in visit number for the calcium group increases the BMD by 0.0043243 more than a 1-unit increase in visit number for the placebo group. In other words, we estimate that a 1-unit increase in visit number would increase the expected BMD of an individual in the calcium group by $\hat{\beta}_2 + \hat{\beta}_3 = 0.0270137$.

d. The estimated difference in mean bone mineral density between the two treatments on the first visit is 0.0123462, where the calcium group is estimated to have the greater mean bone mineral density on the first visit.
The estimated difference in mean bone mineral density between the two treatments on the last visit is 0.0296434, where the calcium group is estimated to have the greater mean bone mineral density on the last visit.

e. Test for an interaction between v_{ij} and Trt_i
$H_0 : \beta_3 = 0$
$H_A : \beta_3 \neq 0$
The p-value for this test is approximately 0 (0.000).
Assuming a 5% significance level, since the p-value < 0.05, we reject the null hypothesis and conclude that there is evidence for the alternative. That is, there is evidence that the change in mean bone mineral density across visits differs between the two treatments.
$\hat{\beta}_3 = 0.0043243$ is the estimated interaction term between v_{ij} and Trt_i. That is to say, the estimated change in mean bone mineral density is 0.0043243 comparing the calcium group to the placebo group for a 1-unit increase in visit number.

f. Estimated marginal variance-covariance matrix:

```
## 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.

```
##           Model df       AIC       BIC   logLik   Test  L.Ratio p-value
## mod1.REML     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 using constant variance is good enough.
H_A : The full model using non-constant variance is good enough.
The p-value is $8 * 10^{-4}$.
Assuming a 5% significance level, since the p-value $= 8 * 10^{-4} < 0.05$, we reject the null and conclude that we have evidence for the alternative. That is, there is evidence that the variance in bone mineral density differs across visits.

10.

a. $Y_{ij} = \beta_0 + b_{0i} + \beta_1 * Trt_i + \beta_2 * a_{ij} + \beta_3 * Trt_i * a_{ij} + \varepsilon_{ij}$
Where Y_{ij} is the BMD of the i-th subject at the j-th visit number,
Trt_i is 1 if the i-th subject is part of the calcium group and 0 if the i-th subject is part of the placebo group,
a_{ij} is the age of the i-th subject at the j-th visit,
$\varepsilon_{ij} \sim_{iid} N(0, \sigma^2)$,
$b_{0i} \sim_{iid} N(0, \sigma_b^2)$,
ε_{ij} and b_{0i} are independent.

b.

```
mod3.REML = lme(BMD ~ Trt+Age+Trt*Age, random = ~ 1|ID, data=BoneLong, na.action=na.omit, method="REML")
summary(mod3.REML)
```

```
## Linear mixed-effects model fit by REML
##   Data: BoneLong
##         AIC       BIC   logLik
##   -2240.336 -2215.085 1126.168
##
## Random effects:
##  Formula: ~1 | ID
##         (Intercept)   Residual
## StdDev:  0.06714559 0.01505221
##
## Fixed effects:  BMD ~ Trt + Age + Trt * Age
##                  Value   Std.Error  DF  t-value p-value
## (Intercept)  0.3631500 0.018749320 387 19.36870  0.0000
## TrtC        -0.0863231 0.026831968 110 -3.21717  0.0017
## Age          0.0457725 0.001381280 387 33.13777  0.0000
## TrtC:Age     0.0088756 0.001979407 387  4.48397  0.0000
##  Correlation:
##          (Intr) TrtC   Age
## TrtC     -0.699
## Age      -0.879  0.614
## TrtC:Age  0.613 -0.879 -0.698
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.35804878 -0.59825830  0.02763527  0.61860407  2.67699153
##
## Number of Observations: 501
## Number of Groups: 112
```

c. The estimated standard deviation of the random intercepts is equal to 0.06714559. This implies that the overall intercepts are not varying by a lot around estimated shared intercept of $\hat{\beta}_0 = 0.3631500$. Taking ± 2 standard deviations tells us that most of the individual intercepts are between 0.23 and 0.50, since taking ± 2 standard deviations tells us that most of the random intercepts are between ± 0.1342912.

d.

```
##            Model df       AIC       BIC   logLik   Test  L.Ratio p-value
## mod3.REML      1  6 -2240.336 -2215.085 1126.168
## mod3s.REML     2  8 -2335.238 -2301.569 1175.619 1 vs 2 98.90186  <.0001
```

$H_0 : \sigma_1^2 = 0$
$H_A : \sigma_1^2 > 0$

The p-value is <.0001
Assuming a 5% significance level, since the p-value <.0001 < 0.05, we reject the null and conclude that we have evidence for the alternative. That is, there is evidence that we should have a random slope on age.

e.

```
##        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
```

Consider the model:
$Y_{ij} = \beta_0 + b_{0i} + \beta_1 * Trt_i + \beta_2 * a_{ij} + \beta_3 * BMICatOver_i + \beta_4 * BMICatUnder_i + \beta_5 * Trt_i * a_{ij} + b_{1i} * a_{ij} + \varepsilon_{ij}$
where $BO_i = 1$ when the i-th subject's BMI is categorized as "Over" and 0 otherwise,
and $BU_i = 1$ when the i-th subject's BMI is categorized as "Under" and 0 otherwise.

To test whether we should adjust for BMICat in our model,
H_0 : Reduced model (not containing BMICat covariates) is good enough.
H_A : Full model (containing BMICat covariates) is good enough.
The p-value is <.0001
Assuming a 5% significance level, since the p-value <.0001 < 0.05, we reject the null and conclude that we have evidence for the alternative. That is, there is evidence that we need the full model and that we should include the BMICat covariate.

f. We need to first transform the residuals from a LME model before doing any residual diagnostics to de-correlate the residuals and have them be independent with constant variance. If we did not transform the residuals this way, they may be correlated with the covariates, leading to the appearance of a systematic trend in the scatter plot of the residuals against a specified covariate. In addition, with untransformed residuals, we could have non-constant variance, leading to the scatter plot of the residuals against predicted values not having a constant range. By transforming the residuals, we can make them similar to residuals from the standard linear regression so we can use the regular linear regression diagnostics, like QQ plots or plots of transformed residuals against predicted values.

11. This study is limited in that we are interested in the effect of calcium supplementation on bone mineral density in young children, but our sample consists of 112 adolescent girls from primarily white middle-class neighborhoods. Our sample isn't very representative of our target population, as we are not including males, lower-class, or upper-class neighborhoods. In addition, these neighborhoods are primarily white, which could cause our sample to consist of primarily white children and not be representative of the population of all young children.
    A better study design would make use of a random sample of children, both male and female, and from a range of different socioeconomic and ethnic backgrounds. This would allow for generalizability of the study's findings to a larger demographic.

12. From part 10, I decided that the best LME model of the ones tested would be the full model from part e, the linear mixed effect model involving covariates Age, Treatment, BMI category, an interaction between age and treatment, a random intercept, and a random slope on Age.
    Using this mode, the study suggests that calcium supplementation increases the effect of age on bone mineral density among adolescent girls. While the coefficient estimate corresponding to the calcium group is negative, the interaction term of treatment group and age has a positive coefficient estimate, and testing for the interaction term shows that it is significant and should be considered in the model. This tells us that calcium supplementation has a positive effect on how age affects the bone mineral density among adolescent girls.