PHP2516: Applied Longitudinal Data Analysis

Homework 1

Antonella Basso

Part I

Question 1: Chapter 2 - Problem 2.1

The Treatment of Lead-Exposed Children (TLC) trial was a placebo-controlled, randomized study of succimer (a chelating agent) in children with blood lead levels of 20 to 44 micrograms/dL. Recall that the data consists of four repeated measurements of blood lead levels obtained at baseline (or week 0), week 1, week 4, and week 6 on 100 children who were randomly assigned to chelation treatment with succimer or to placebo. For this problem set we focus only on the 50 children assigned to chelation treatment with succimer.

a) Read the data from the external file and calculate the sample means, standard deviations, and variances of the blood lead levels at each occasion.

b) Construct a time plot of the mean blood lead levels (BLL) versus time (in weeks). Describe the general characteristics of the time trend.

c) Calculate the 4 x 4 covariance and correlation matrices for the four repeated measures of blood lead levels.

d) Verify that the diagonal elements of the covariance matrix are the variances by comparing to the descriptive statistics obtained in part (a).

e) Verify that the correlation between blood lead levels at baseline (week 0) and week 1 is equal to the covariance between blood lead levels at baseline and week 1, divided by the product of the standard deviations of the blood lead levels at baseline and week 1.

```
#importing "TLC" data with N=100 (treatment and placebo groups)
lead <- read.table("/Users/antonellabasso/Desktop/PHP2516/DATA/lead.txt")
TLC_100 <- rename(lead, ID=V1, treatment=V2, week0=V3, week1=V4, week4=V5, week6=V6)

#"TLC" data with N=50, only treatment group
#subsetting data and removing treatment column
TLC <- subset(TLC_100, treatment!="P", select=-treatment)
rownames(TLC) <- 1:nrow(TLC) #updating row names
TLC["ID"] <- 1:nrow(TLC) #updating ID values

head(TLC)
```

```
##   ID week0 week1 week4 week6
## 1  1  26.5  14.8  19.5  21.0
## 2  2  25.8  23.0  19.1  23.2
## 3  3  20.4   2.8   3.2   9.4
## 4  4  20.4   5.4   4.5  11.9
## 5  5  24.8  23.1  24.6  30.9
## 6  6  27.9   6.3  18.5  16.3
```

a) Summary Statistics

- BLL means by week/occasion
- BLL standard deviations by week/occasion
- BLL variances by week/occasion

```r
#Summary Statistics

#of all subject BLL by week/occasion
#summary(TLC[2:5])

#mean of all subject BLL by week/occasion
apply(TLC[2:5], 2, mean)
```

```
##  week0  week1  week4  week6
## 26.540 13.522 15.514 20.762
```

```r
#standard deviation of all subject BLL by week/occasion
apply(TLC[2:5], 2, sd)
```

```
##    week0    week1    week4    week6
## 5.020936 7.672487 7.852207 9.246332
```

```r
#variance of all subject BLL by week/occasion
apply(TLC[2:5], 2, var)
```

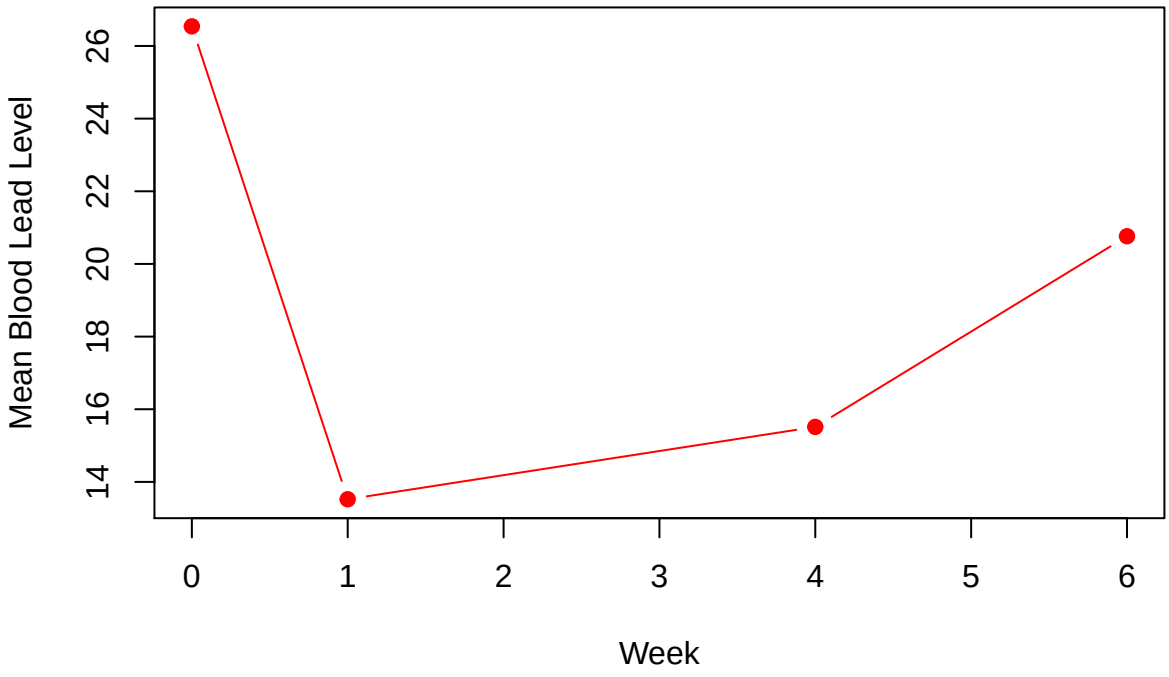
```
##    week0    week1    week4    week6
## 25.20980 58.86706 61.65715 85.49465
```

b) Time Plot

The time plot below shows the mean blood lead levels of subjects at each time point (in weeks). Here, we see that before the treatment was introduced (baseline - week 0), subjects demonstrated the highest average blood lead level of the whole the study (26.54 micrograms/dL). Moreover, we notice a significant drop in BLL's within the first week of receiving treatment, bringing the average down to 13.522 micrograms/dL, the lowest of recorded averages. With each subsequent visit however, we start to notice a gradual increase in mean BLL that reaches 20.76 micrograms/dL by the end of the study (week 6). Based on this time plot of computed averages from the data, an immediate effect of treatment on BLL can be speculated, as well as an increasing trend in mean with time.

```r
#Time Plot

week <- c(0, 1, 4, 6)
avg_BLL <- c(mean(TLC$week0), mean(TLC$week1), mean(TLC$week4), mean(TLC$week6))

plot(week, avg_BLL, type="b", pch = 19, col = "red",
     xlab = "Week", ylab = "Mean Blood Lead Level", main="Figure 1: Mean Blood Lead Levels by Week")
```

Figure 1: Mean Blood Lead Levels by Week



c) Covariance and Correlation Matrices

```
#Variance-Covariance Matrix

cov(TLC[2:5])
```

```
##          week0    week1    week4    week6
## week0 25.20980 15.46543 15.13800 22.98543
## week1 15.46543 58.86706 44.02907 35.96596
## week4 15.13800 44.02907 61.65715 33.02197
## week6 22.98543 35.96596 33.02197 85.49465
```

```
#Correlation Matrix

cor(TLC[2:5])
```

```
##           week0     week1     week4     week6
## week0 1.0000000 0.4014589 0.3839654 0.4951063
## week1 0.4014589 1.0000000 0.7308221 0.5069743
## week4 0.3839654 0.7308221 1.0000000 0.4548224
## week6 0.4951063 0.5069743 0.4548224 1.0000000
```

d) Verification I

Verifying that the diagonal elements of the covariance matrix are equal to the weekly variances.

```
apply(TLC[2:5], 2, var) #weekly variances
```

```
##    week0    week1    week4    week6
## 25.20980 58.86706 61.65715 85.49465
```

```r
diag(cov(TLC[2:5])) #same-week covariances (diagonal of covariance matrix)
```

```
##    week0    week1    week4    week6
## 25.20980 58.86706 61.65715 85.49465
```

```r
setequal(apply(TLC[2:5], 2, var), diag(cov(TLC[2:5]))) #equal
```

```
## [1] TRUE
```

d) Verification II

Verifying that the correlation between BLL's at baseline (week 0) and week 1 is equal to the covariance BLL's at baseline and week 1, divided by the product of the standard deviations of the BLL's at baseline and week 1.

```r
#correlation between baseline and week 1
cor(TLC[2:5])[1, 2]
```

```
## [1] 0.4014589
```

```r
#covariance between baseline and week 1 / baseline sd * week 1 sd
cov(TLC[2:5])[1, 2]/(sd(TLC$week0)*sd(TLC$week1))
```

```
## [1] 0.4014589
```

```r
setequal(apply(TLC[2:5], 2, var), diag(cov(TLC[2:5]))) #equal
```

```
## [1] TRUE
```

Part II

The following questions (2-4) are based on the 'calcium_all' dataset:

The 'calcium_all' data is a subset of the 'calcium' dataset that can be found in the 'lava' R package. This dataset contains information collected from an RCT aimed at comparing calcium vs placebo with respect to changes in bone mineral density measures (g/cm2) over time. For this purpose BMD measurements were taken on girls at approximately every 6th months in 3 years. The "calcium_all" dataset has information on the following variables:

- **bmd**: bone mineral density (BMD) in g/cm2
- **group**: treatment group ('C'=calcium, 'P'=placebo)
- **person**: person (girl) ID
- **visit**: visit number (time point)
- **age**: age at each visit in years

Question 2: Exploratory Data Analysis (EDA)

The variables in this dataset are as follows:

- Outcome Y_{ij}: bone mineral density (BMD) in g/cm2 for the i^{th} individual on the j^{th} visit/occasion:
  - $i = 1, 2, ..., 112$ and $j = 1, 2, 3, 4, 5$
- Covariate X_1: treatment group ('C'=treatment, 'P'=placebo)

The primary outcome of interest is a continuous random variable with ratio scale, while the predictor variable (covariate) is binary (categorical with nominal scale).

This EDA consists of:

- Descriptive Statistics
- Smooth Line Plots
- Boxplots

- Individual Trajectory Plots
- Variance-Covariance Matrix
- Correlation Matrix

NOTE: Given that the starting age of subjects in the study was (relatively) constant, age can be interpreted here as a continuous measure of time (unlike visit which is discrete) and not a predictor of BMD. That being the case, the primary objective is to conduct a bivariate analysis of the data to assess changes in the primary outcome over time.

```
#importing "calcium" data
calcium <- read.csv("/Users/antonellabasso/Desktop/PHP2516/DATA/calcium_all.csv")
head(calcium)
```

```
##     bmd group person visit      age
## 1 0.815     C    101     1 10.91307
## 2 0.875     C    101     2 11.42779
## 3 0.911     C    101     3 11.89322
## 4 0.952     C    101     4 12.41068
## 5 0.970     C    101     5 12.90897
## 6 0.813     P    102     1 10.91307
```

```
#number of individuals in study
length(unique(calcium$person))
```

```
## [1] 112
```

```
#visits per individual
unique(calcium$visit)
```

```
## [1] 1 2 3 4 5
```

```
#age range
range(calcium$age)
```

```
## [1] 10.91307 13.24846
```

```
# Descriptive Statistics

#by visit
calcium_by_visit <- calcium %>%
  group_by(visit)
calcium_by_visit_sum <- calcium_by_visit %>%
  summarise(mean=mean(bmd),
            sd=sd(bmd),
            var=var(bmd),
            min_age=min(age),
            max_age=max(age))
calcium_by_visit_sum
```

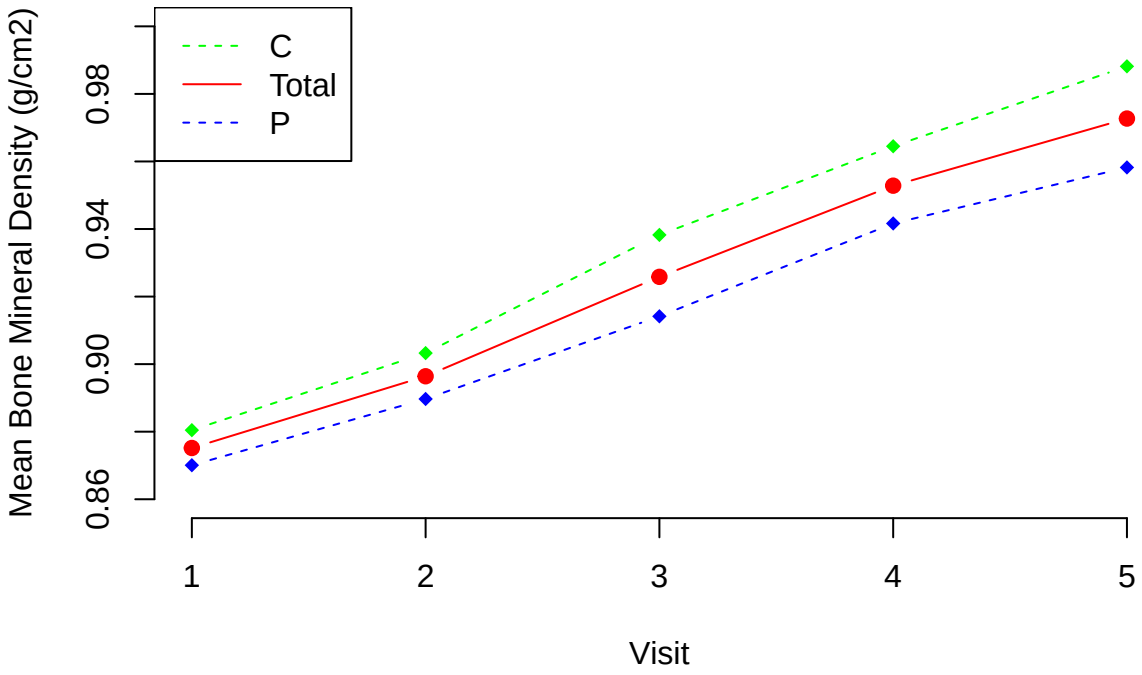
```
## # A tibble: 5 x 6
##   visit  mean     sd     var min_age max_age
##   <int> <dbl>  <dbl>   <dbl>   <dbl>   <dbl>
## 1     1 0.875 0.0628 0.00394    10.9    11.2
## 2     2 0.896 0.0678 0.00459    11.4    11.7
## 3     3 0.926 0.0701 0.00491    11.9    12.2
## 4     4 0.953 0.0711 0.00505    12.4    12.8
## 5     5 0.973 0.0699 0.00489    12.9    13.2
```

```r
#by treatment group
calcium_by_group <- calcium %>%
  group_by(group)
calcium_by_group_sum <- calcium_by_group %>%
  summarise(mean=mean(bmd),
            sd=sd(bmd),
            var=var(bmd))
calcium_by_group_sum
```

```
## # A tibble: 2 x 4
##   group  mean     sd     var
##   <chr> <dbl>  <dbl>   <dbl>
## 1 C     0.932 0.0722 0.00522
## 2 P     0.913 0.0798 0.00637
```

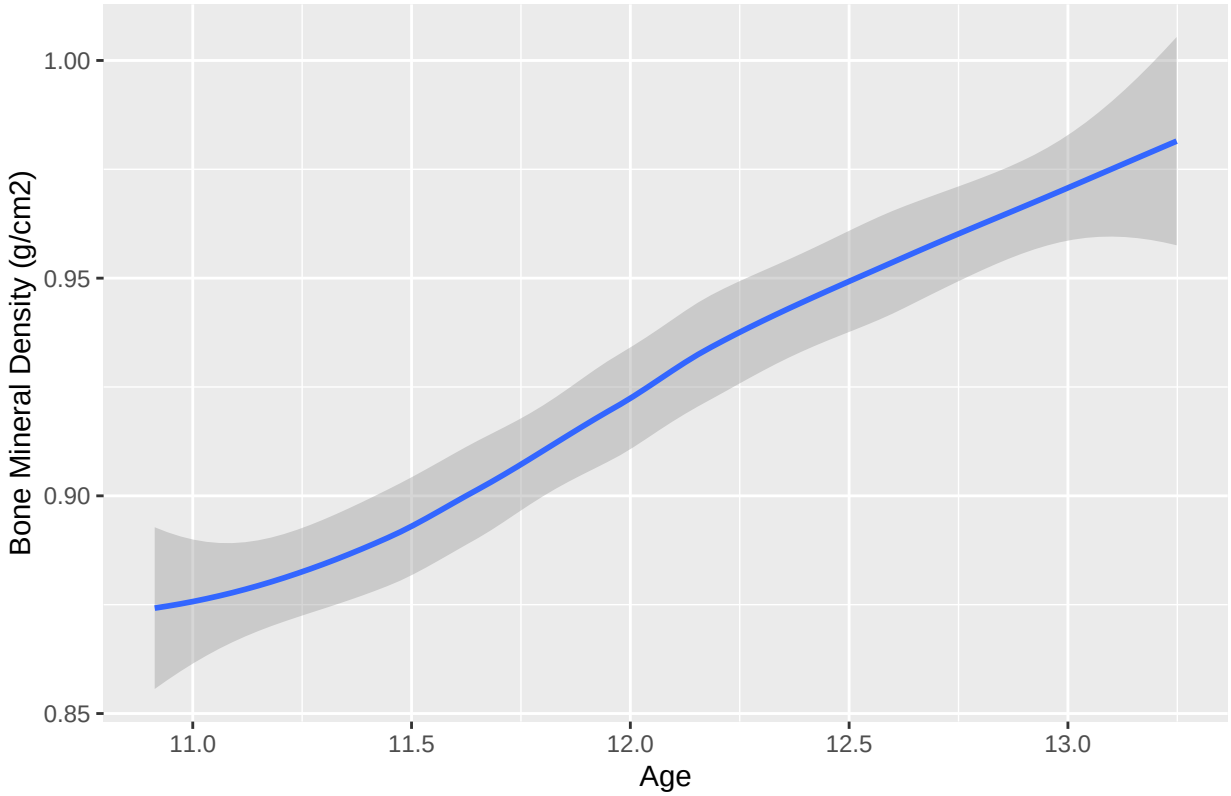
```r
#Line Plots

#by-group and total means over time (visit)
calcium_C <- subset(calcium, group=="C") %>%
  group_by(visit) %>%
  summarise(mean=mean(bmd)) #means for treatment group
calcium_P <- subset(calcium, group=="P") %>%
  group_by(visit) %>%
  summarise(mean=mean(bmd)) #means for placebo group

plot(calcium_by_visit_sum$visit, calcium_by_visit_sum$mean,
     pch=19, col="red", type="b", lty=1, frame=FALSE, ylim=c(0.86, 1),
     xlab="Visit", ylab="Mean Bone Mineral Density (g/cm2)",
     main="Figure 2: Mean BMD by Visit")
lines(calcium_by_visit_sum$visit, calcium_C$mean,
      pch=18, col="green", type="b", lty=2)
lines(calcium_by_visit_sum$visit, calcium_P$mean,
      pch=18, col="blue", type= "b", lty=2)
legend("topleft", legend=c("C", "Total", "P"),
       col=c("green", "red", "blue"), lty = 2:1)
```

Figure 2: Mean BMD by Visit



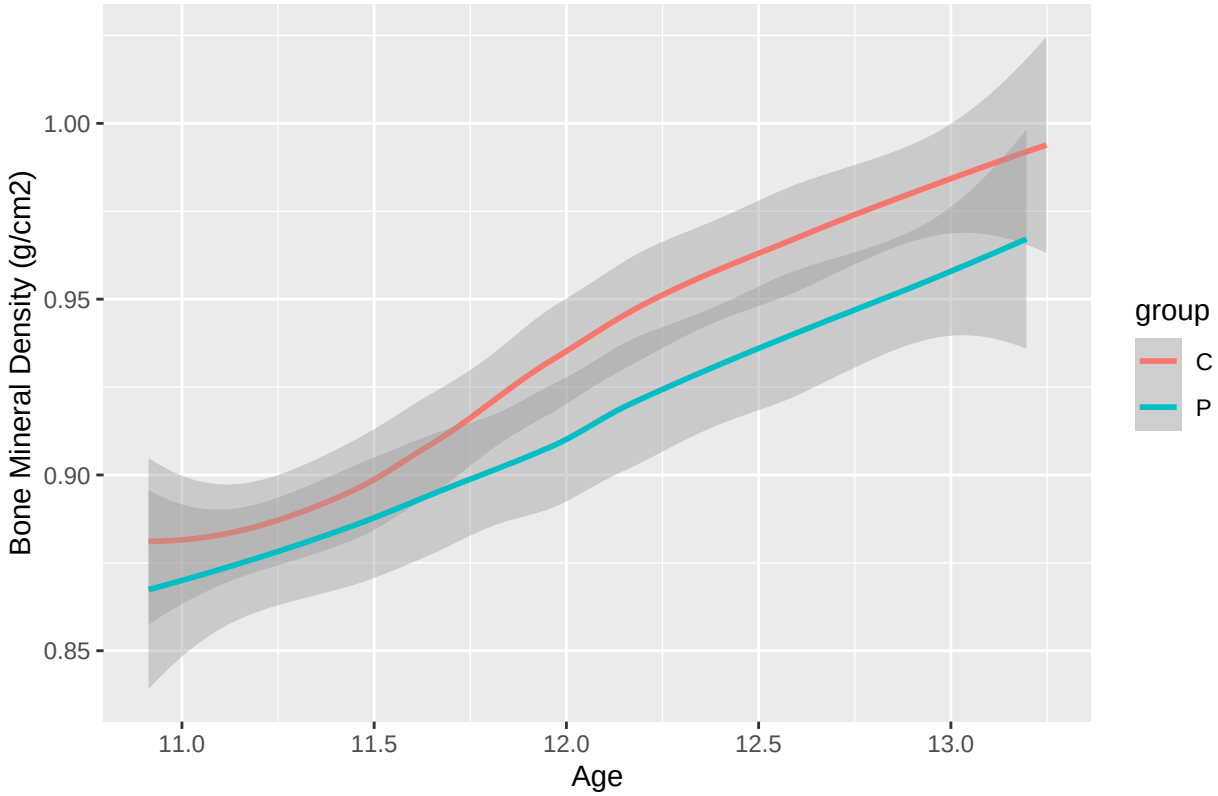

```
#Smooth Line Plots

#BMD agaisnt age (continuous measure of time)
ggplot(calcium, aes(x=age, y=bmd)) +
    geom_smooth() +
    labs(x="Age", y="Bone Mineral Density (g/cm2)", title="Figure 3: BMD by Age")
```

Figure 3: BMD by Age



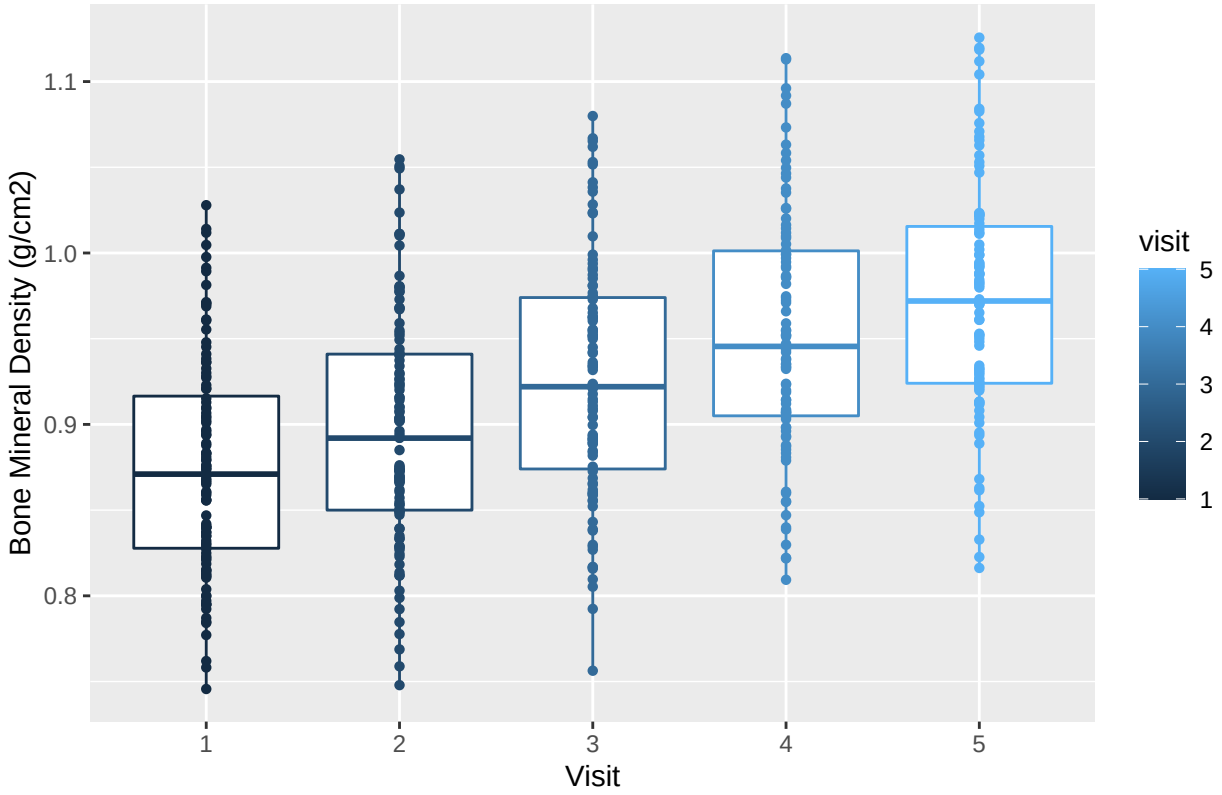

```
#by-group BMD agaisnt age (continuous measure of time)
ggplot(calcium, aes(x=age, y=bmd, color=group)) +
    geom_smooth() +
    labs(x="Age", y="Bone Mineral Density (g/cm2)", title="Figure 4: BMD per Group by Age")
```

Figure 4: BMD per Group by Age



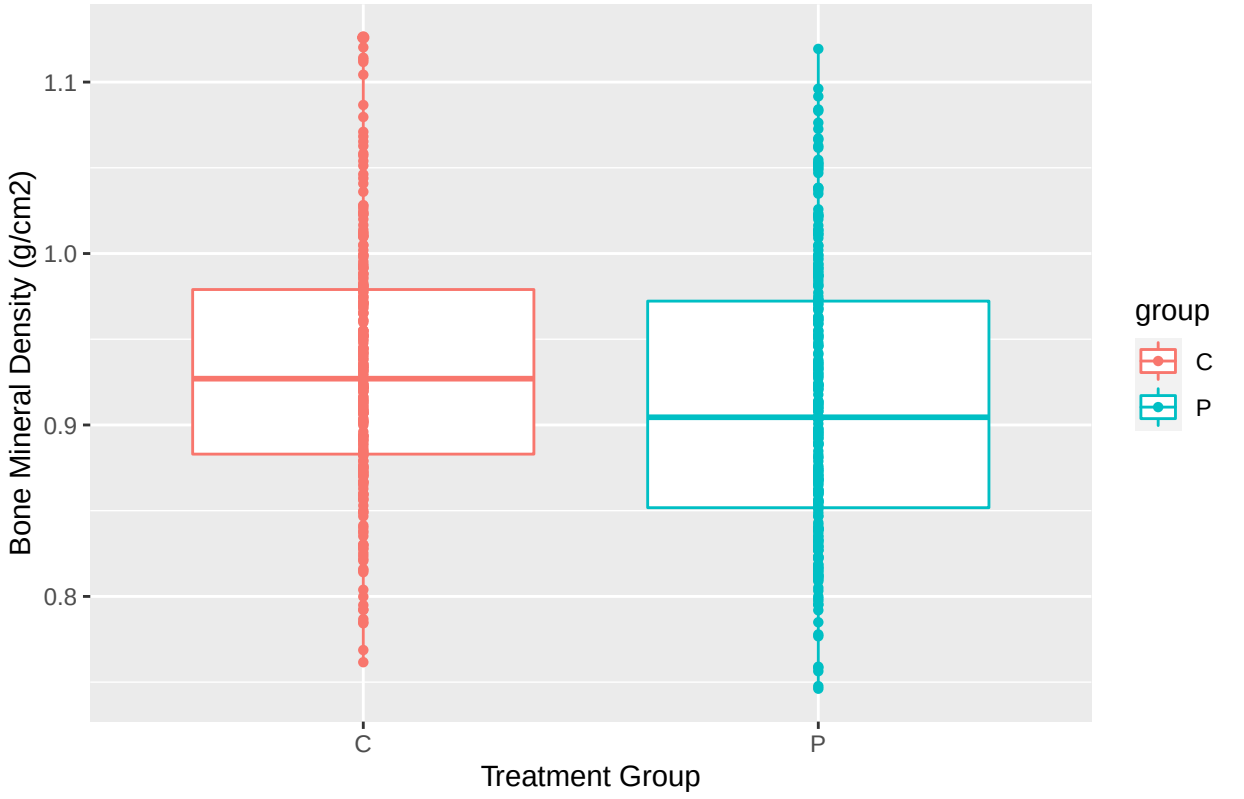

```
#Boxplots

#spread of subject bmd by visit
ggplot(calcium, aes(x=as.character(visit), y=bmd, color=visit)) +
    geom_boxplot() +
    geom_jitter(shape=16, position=position_jitter(0)) +
    labs(x="Visit", y="Bone Mineral Density (g/cm2)", title="Figure 5: Spread of BMD by Visit")
```

Figure 5: Spread of BMD by Visit



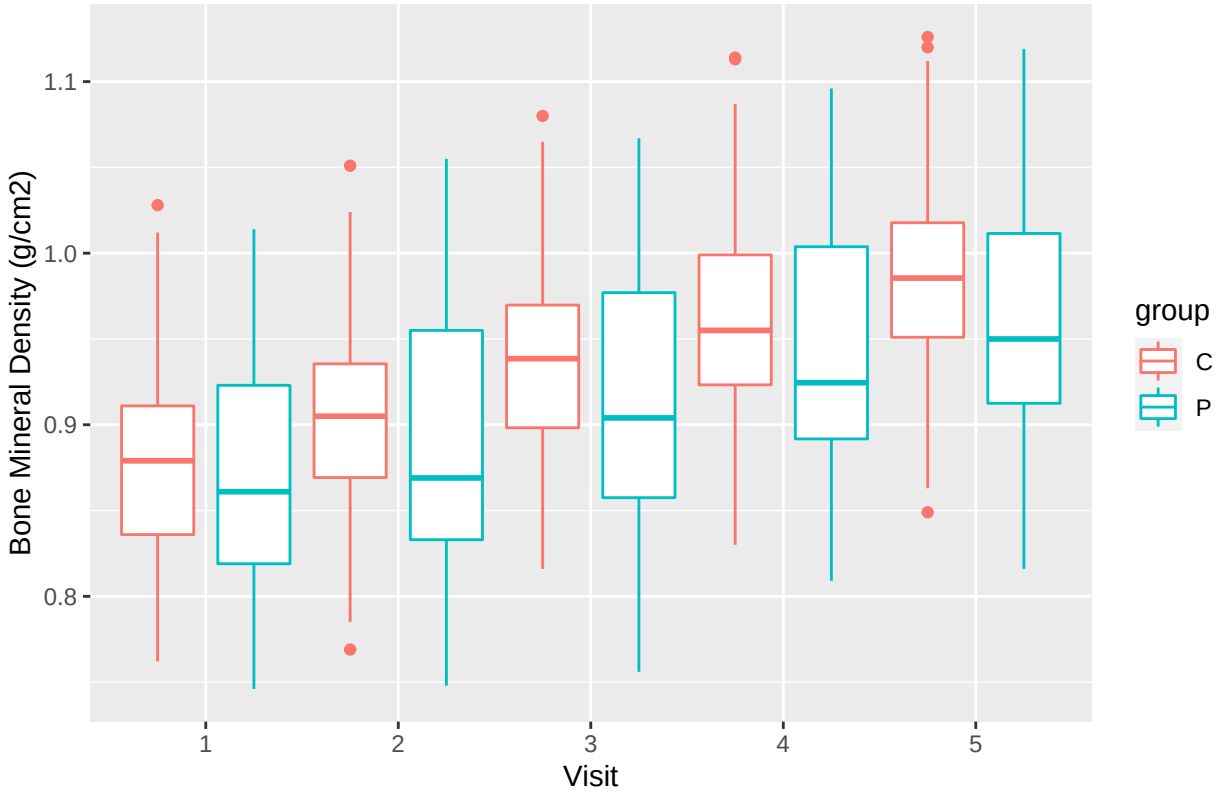

```
#spread of subject bmd by group
ggplot(calcium, aes(x=group, y=bmd, color=group)) +
    geom_boxplot() +
    geom_jitter(shape=16, position=position_jitter(0)) +
    labs(x="Treatment Group", y="Bone Mineral Density (g/cm2)", title="Figure 6: Spread of BMD by Treatr
```

Figure 6: Spread of BMD by Treatment Group



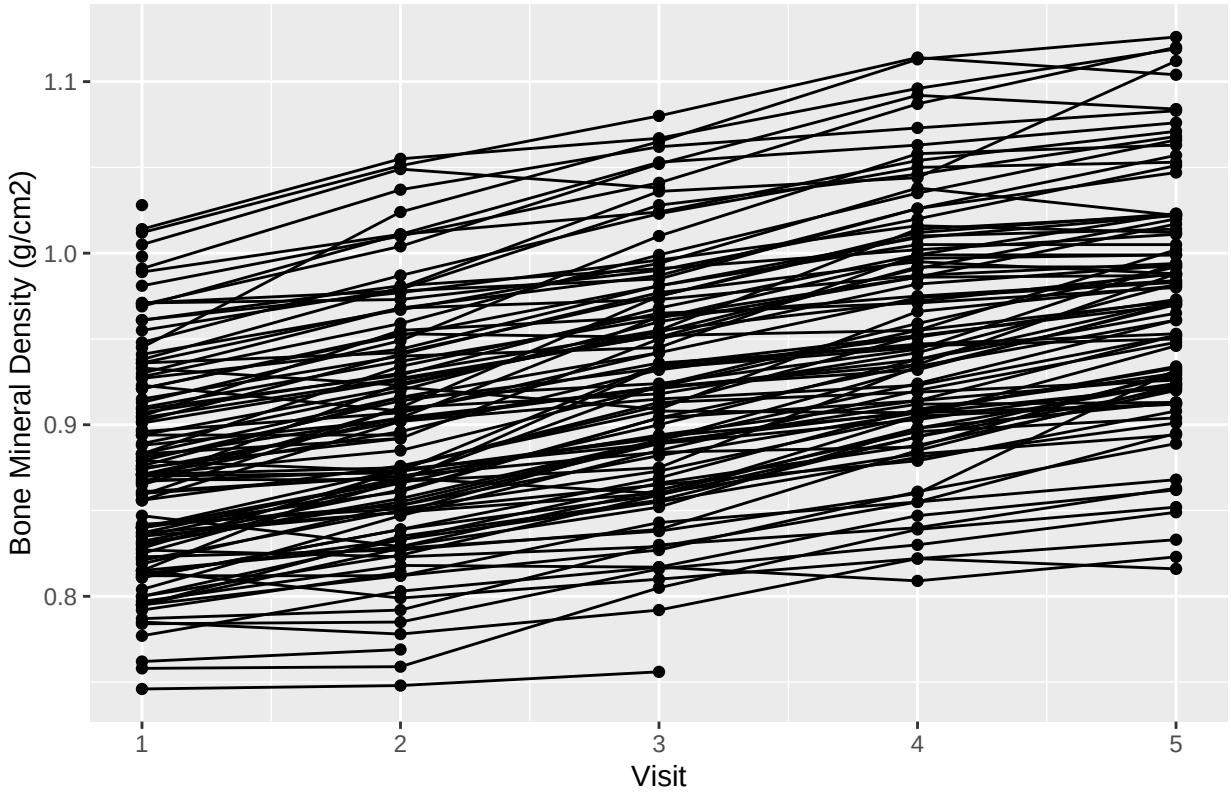

```
#spread of subject bmd by visit and treatment
ggplot(calcium, aes(x=as.character(visit), y=bmd, color=group)) +
    geom_boxplot(position=position_dodge(1)) +
    labs(x="Visit", y="Bone Mineral Density (g/cm2)", title="Figure 7: Spread of BMD by Treatment Group
```

Figure 7: Spread of BMD by Treatment Group and Visit



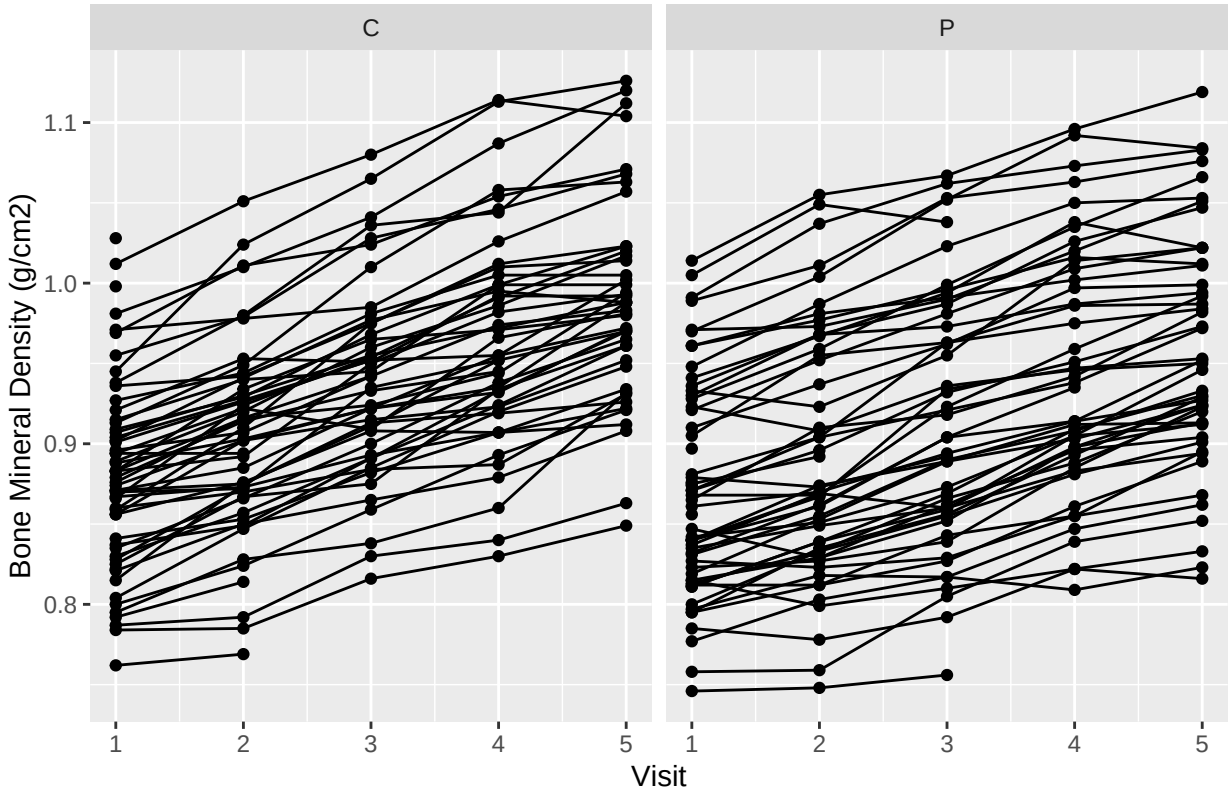

```
#Individual Trajectories

calcium_by_person <- calcium %>% group_by(visit, person) %>% summarise(avg_bmd=mean(bmd), group=group) %

## `summarise()` has grouped output by 'visit'. You can override using the `.groups` argument.
#ploting individual trajectories over time (per visit)
ggplot(as.data.frame(calcium_by_person), aes(x=visit, y=avg_bmd, group=person)) +
    geom_point() +
    geom_line() +
    labs(x="Visit", y="Bone Mineral Density (g/cm2)", title="Figure 8: Individual BMD Trajectories Over
```

Figure 8: Individual BMD Trajectories Over Time




```
#ploting individual trajectories by treatment group over time (per visit)
ggplot(as.data.frame(calcium_by_person), aes(x=visit, y=avg_bmd, group=person)) +
    geom_point() +
    geom_line() +
    facet_grid(. ~ group) +
    labs(x="Visit", y="Bone Mineral Density (g/cm2)", title="Figure 9: Individual BMD Trajectories by T
```

Figure 9: Individual BMD Trajectories by Treatment Group Over Time




```
#transforming overall and treatment group data frames to wide format (for matrices)
calcium_wide1 <- subset(calcium, select=-age)
calcium_wide2 <- as.data.frame(pivot_wider(calcium_wide1, names_from=visit, values_from=bmd))
calcium_wide_C <- subset(calcium_wide2, group=="C")
calcium_wide_P <- subset(calcium_wide2, group=="P")
```

```
#Variance-Covariance Matrices

cov(calcium_wide2[3:7], use="complete.obs") #overall
```

```
##             1           2           3           4           5
## 1 0.003738550 0.003955070 0.003999360 0.004063891 0.003799707
## 2 0.003955070 0.004478697 0.004537925 0.004629851 0.004383248
## 3 0.003999360 0.004537925 0.004855716 0.004932280 0.004662721
## 4 0.004063891 0.004629851 0.004932280 0.005212552 0.004915858
## 5 0.003799707 0.004383248 0.004662721 0.004915858 0.004886344
```

```
cov(calcium_wide_C[3:7], use="complete.obs") #treatment group
```

```
##             1           2           3           4           5
## 1 0.002977076 0.003079163 0.003072895 0.003220663 0.002949006
## 2 0.003079163 0.003472922 0.003485888 0.003720300 0.003443387
## 3 0.003072895 0.003485888 0.003771114 0.003973757 0.003645760
## 4 0.003220663 0.003720300 0.003973757 0.004422273 0.004024010
## 5 0.002949006 0.003443387 0.003645760 0.004024010 0.003952881
```

```
cov(calcium_wide_P[3:7], use="complete.obs") #placebo group
```

```
##             1           2           3           4           5
```

```
## 1 0.004472838 0.004761589 0.004831235 0.004814139 0.004516245
## 2 0.004761589 0.005352092 0.005417644 0.005369648 0.005087620
## 3 0.004831235 0.005417644 0.005725685 0.005675299 0.005382515
## 4 0.004814139 0.005369648 0.005675299 0.005793171 0.005509877
## 5 0.004516245 0.005087620 0.005382515 0.005509877 0.005422748
```

```
#use="complete.obs" handles missing values by casewise deletion (and if there are no complete cases, th
```

Question 3: EDA Results & Conclusions

Comment on the results from both descriptive statistics and plots. What are your conclusions regarding:

a) changes in variables over time?
b) variability?
c) association among measurements?
d) comparisons between the two treatment groups?

Conclusions:

a) Given that there is only one primary outcome variable (BMD) and one predictor (treatment), the EDA above was primarily focused on observing and summarizing changes in BMD with and without respect to treatment group over time (both as a discrete (visit) and continuous (age) measurement). Generally, all graphs above show us that BMD tends to increase with time irrespective of treatment.
b) Looking at the diagonal entries of the covariance matrices (for each treatment group and the whole data) above, we notice that the variances of our observations increase over time (similarly) in both treatment groups, despite being slightly higher in the placebo group. This heterogeneity of variance can be observed in Figure 4 and is to be expected due to naturally-occuring person-person differences (which tend to become more prominent with the passage of time).
c) The correlation matrices (for each treatment group and the whole data) reveal strong positive association/correlation among measurements. Specifically, we see that no correlation coefficients go below 0.9 and 0.85 in the placebo and treatment groups respectively, and that these values tend to decrease slightly with increasing time separation (both groups showed smallest correlation coefficients between baseline and final visits). This demonstrates that there is strong dependence among repeated measures, which may (in part) be due to between-individual and within-individual heterogeneity/variability, both of which can be observed in Figures 8 and 9 above (time plot of individual trajectories).
d) It is evident that BMD levels are somewhat higher for the treatment group (consistently so) than they are for the placebo group. However, since this is true for observations at the start of the study as well (visit 1, which we assume displays measurements prior to treatment) and both groups display similar positive trends in BMD over time, we cannot be sure whether treatment could have a significant effect (if any) on BMD. We also see in Figures 2 and 4 that mean BMD's are slightly closer together at the start of the study and become more spread out as time elapses. This could be due to variances increasing with time, which we verify in the covariance matrices above, and is likely to stem from between-individual variability.

Question 4: Comparing BMD Change Over Time

Is the BMD change over time the same between the two treatment groups? Use appropriate statistical testing to support your answer. Comment on the results.

To assess the change in BMD over time for both treatment groups, two sets of paired t-tests were conducted to compare within-group differences in BMD means and a simple t-test was performed to compare between-group differences in total BMD mean change throughout the study. Specifically, the Holm p-value adjustment method was used in the paired t-tests to ensure a "uniformly" most powerful way of accounting for dependency within our observations. Looking at the results from these tests we notice that there is disagreement between groups only for mean comparisons of visits 2 and 3. While the treatment group ("C") gave statistically significant results (p-value of 0.01833), the placebo group (with a p-value of 0.2723) prevents us from rejecting the null hypothesis that BMD means are equal at these time points. This likely indicates a slightly larger

increase in BMD means from visit 2 to visit 3 in the treatment group, which may be observed in Figure 2, although both trajectories are rather similar. Aside from this conflicting output, all non-sequential visits had non-zero mean differences in BMD (non-diagonal entries) for both groups. In other words, all non-sequential visits show statistically significant differences in mean BMD's. Moreover, aside from the conflicting t-test mentioned, all sequential visit t-tests (diagonal entries) yielded p-values greater than 0.05 for both treatment groups. This indicates that we may not reject the null hypothesis of equal BMD means between these visits. Both of these observations may be supported by Figure 2 above. Noticeably, both treatment groups show similarly increasing trends in BMD means over time (visit). Thus, within-group means are closest between sequential visits (hence the high p-values along the diagonal) and get larger with increasing time separation (hence the low p-vlues elsewhere). Specifically, we obtain the smallest p-value in comparing baseline and final visits, indicating the largest difference in BMD means between these two time points for both treatment groups (this makes sense as BMD means are always increasing with time (as seen in Figure 2)).

In Figure 2, we also notice that although BMD trends are very similar between groups, means seem to be somewhat closer together in the beginning of the study than at the end. So, to see whether this difference (between baseline and final visits) is truly the same for both groups, a simple t-test was conducted on the mean differences (for these time points) for both groups. With a significant p-value of 0.0056, we reject the null hypothesis that the true difference in means is equal to 0 in favor of the alternative that both means are different. This is a valuable result in that it accounts for the fact that both treatment groups did not start out with the same mean BMD. Despite having different means, if the change had been the same over time, we would have obtained a p-value close to 1. Thus, despite the evident similarities in BMD trends over time (seen in the graphs above), this test confirms that we shouldn't reject the possibility of an effect of treatment on BMD.

```
#paired t-tests on within-group differences in bmd means (wrt time - visits)

#the data here are dependent
C_group <- subset(calcium, group=="C") #treatment group
P_group <- subset(calcium, group=="P") #placebo group

pairwise.t.test(C_group$bmd, C_group$visit) #treatment group paired t-tests
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  C_group$bmd and C_group$visit
##
##   1       2       3       4
## 2 0.11483 -       -       -
## 3 1.7e-05 0.01833 -       -
## 4 4.4e-10 9.6e-06 0.11483 -
## 5 4.7e-15 7.1e-10 0.00059 0.11483
##
## P value adjustment method: holm
```

```
pairwise.t.test(P_group$bmd, P_group$visit) #placebo group paired t-tests
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  P_group$bmd and P_group$visit
##
##   1       2       3      4
## 2 0.3266  -       -      -
## 3 0.0124  0.2723  -      -
## 4 1.1e-05 0.0032  0.2563 -
```

```
## 5 4.3e-08 4.2e-05 0.0165 0.3266
##
## P value adjustment method: holm
```

```
#paired t-test gives more accurate results (sensitive to dependent data)
#using Holm instead of Bonferroni p-value adjustment method as it is "uniformly" most powerful
```

```
#simple t-test on between-group differences in BMD mean difference (between baseline and visit 5)

#removing na values
calcium_wide_C2 <- na.omit(calcium_wide_C)
calcium_wide_P2 <- na.omit(calcium_wide_P)

#baseline-visit5 difference for both groups
C_diff <- as.vector(calcium_wide_C2[7]-calcium_wide_C2[3])
P_diff <- as.vector(calcium_wide_P2[7]-calcium_wide_P2[3])

#random sampling to have the same number of observations
set.seed(4)
diffs <- data.frame(C_diff[sample(nrow(C_diff),44),],
                    P_diff[sample(nrow(P_diff),44),])
colnames(diffs) <- c("C_diff", "P_diff")
head(diffs)
```

```
##   C_diff P_diff
## 1  0.124  0.109
## 2  0.098  0.089
## 3  0.098  0.087
## 4  0.071  0.057
## 5  0.113  0.050
## 6  0.144  0.083
```

```
t.test(diffs$C_diff, diffs$P_diff) #simple t-test
```

```
##
##  Welch Two Sample t-test
##
## data:  diffs$C_diff and diffs$P_diff
## t = 2.8415, df = 85.49, p-value = 0.005613
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.005631223 0.031868777
## sample estimates:
##  mean of x  mean of y
## 0.10690909 0.08815909
```