BIOST 540 Homework 1

Hantong Hu, Ivy Zhang

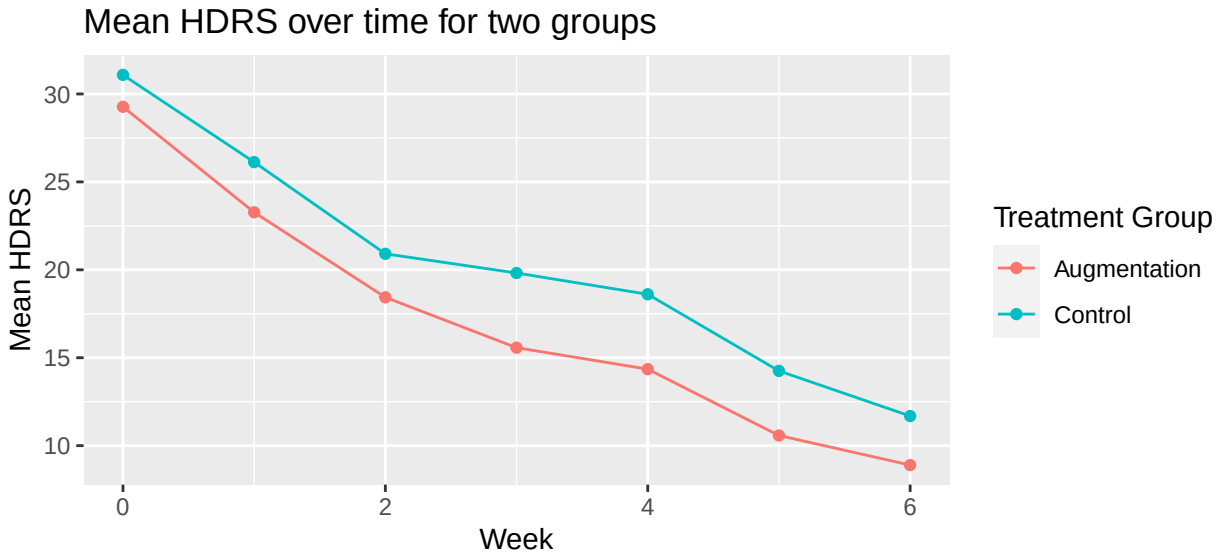
04/09/2021

Responses

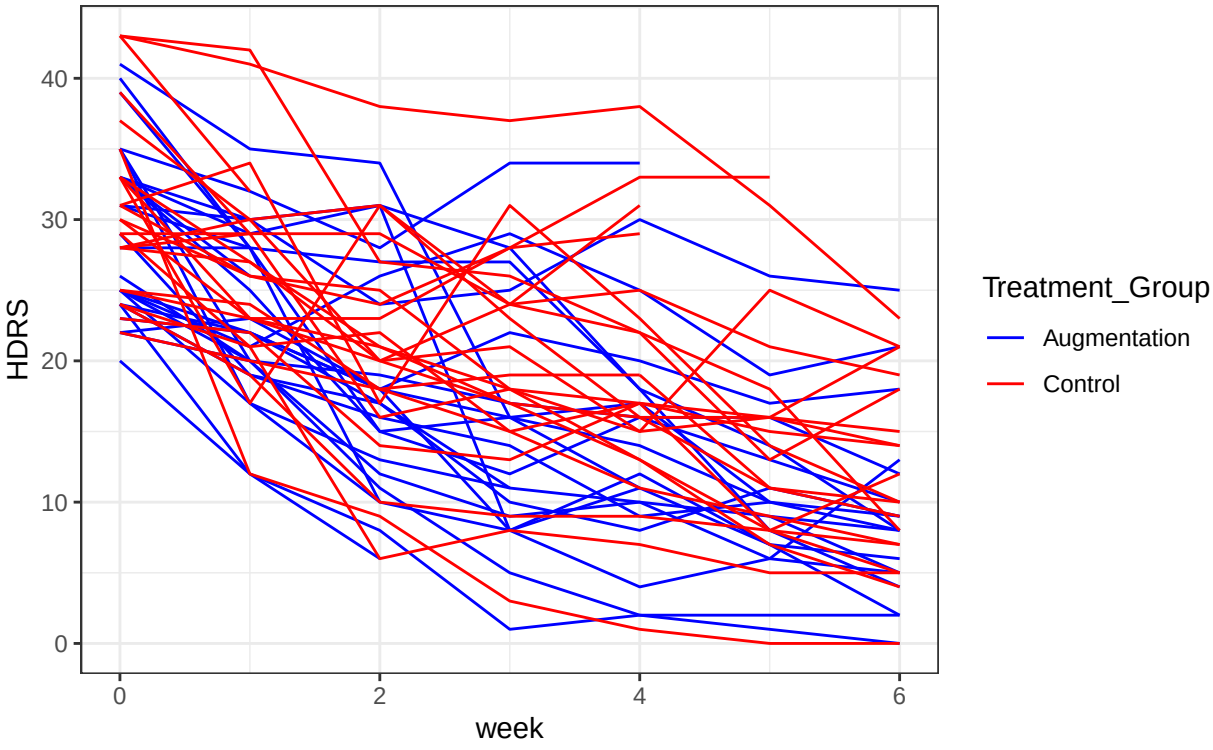
Part A

1. The summary of distribution of HDRS at baseline and week 1-6 in the two groups are shown below. The figure below the summary is the mean HDRS over time for the two groups. In the figure, we can see that the mean HDRS for both groups decreases over time, and the control group has a higher mean HDRS than the treatment group at all times.

	N	Msng	Mean	Std Dev	Min	25%	Mdn	75%	Max
Augmentation Baseline	26	0	29.269	6.024	20	24.25	28.5	33.00	41
Augmentation Week 1	26	0	23.269	5.862	12	20.00	22.0	28.00	35
Augmentation Week 2	26	3	18.435	7.913	6	12.50	17.0	25.00	34
Augmentation Week 3	26	5	15.571	8.846	1	9.00	14.0	22.00	34
Augmentation Week 4	26	6	14.350	8.468	2	9.75	13.0	18.00	34
Augmentation Week 5	26	7	10.579	5.975	1	7.00	10.0	13.50	26
Augmentation Week 6	26	7	8.895	6.582	0	4.50	8.0	11.00	25
Control Baseline	24	0	31.083	6.296	22	27.25	30.0	33.50	43
Control Week 1	24	0	26.125	6.918	12	21.75	26.0	29.25	42
Control Week 2	24	2	20.909	7.621	6	17.25	20.5	24.75	38
Control Week 3	24	2	19.818	7.914	3	15.50	18.5	24.00	37
Control Week 4	24	1	18.609	8.617	1	14.00	17.0	22.50	38
Control Week 5	24	4	14.250	8.353	0	8.00	13.5	16.50	33
Control Week 6	24	5	11.684	6.549	0	7.00	10.0	16.50	23

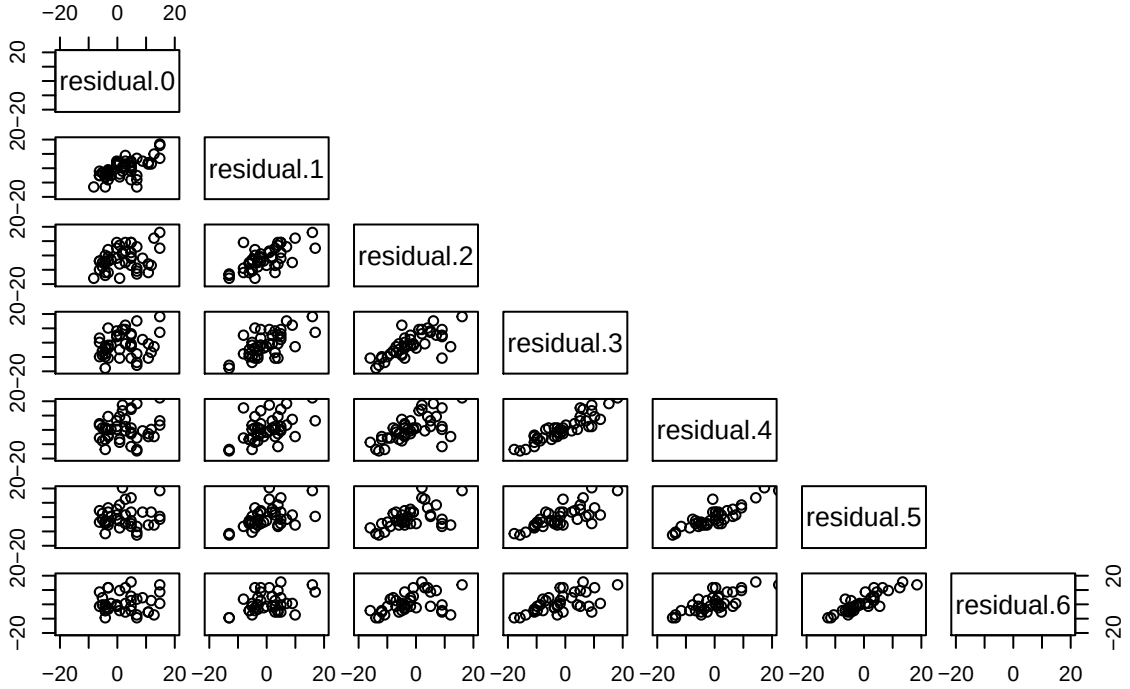


2. In the plot below, we can see that the HDRS for almost everyone decreases over time. It is hard to say which group has higher HDRS because the lines overlap a lot, but in the general sense, there are more red lines (control groups) above the mean HDRS for all subjects, meaning the control group has a higher HDRS over time than the treatment group.



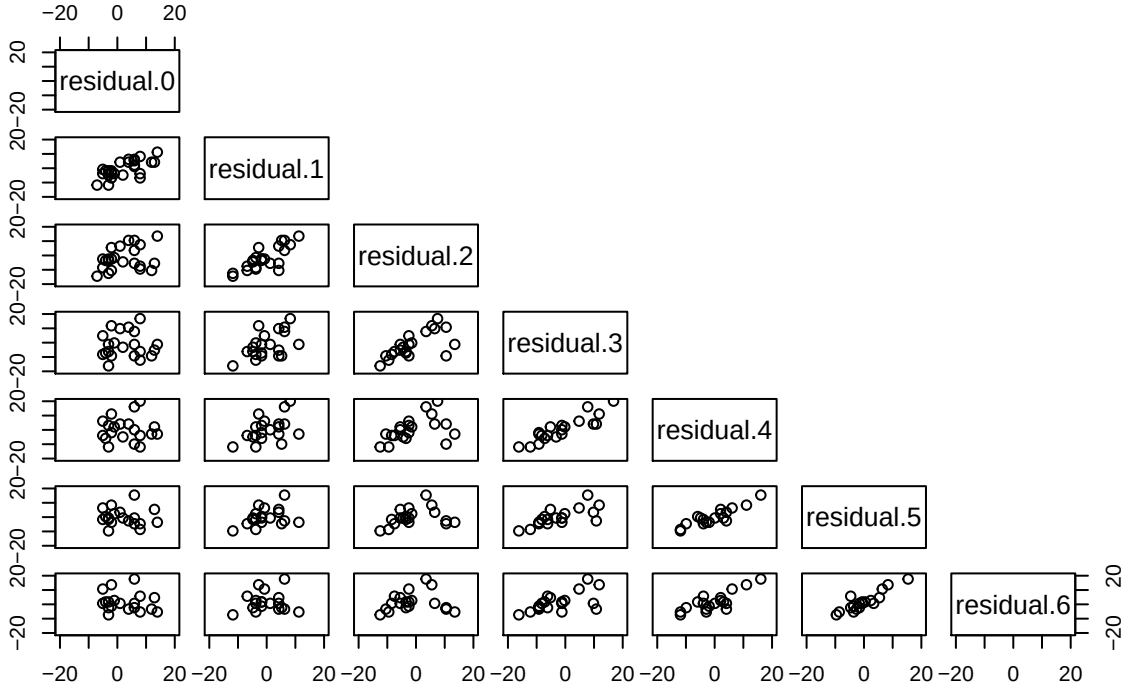
3. For the correlation among the HDRS, both overall and separately, the correlations are mostly positive with a few exceptions in the augmentation group. Observations with closer time points tend to display higher positive correlation values, for example the correlation is higher in week 5 vs week 6 than week 4 vs week 6. The variability in responses are different at different time points, increasing from baseline to week 3 and decreasing from week 3 to week 6. In general, control group has higher correlation value than augmentation group.

Overall correlation



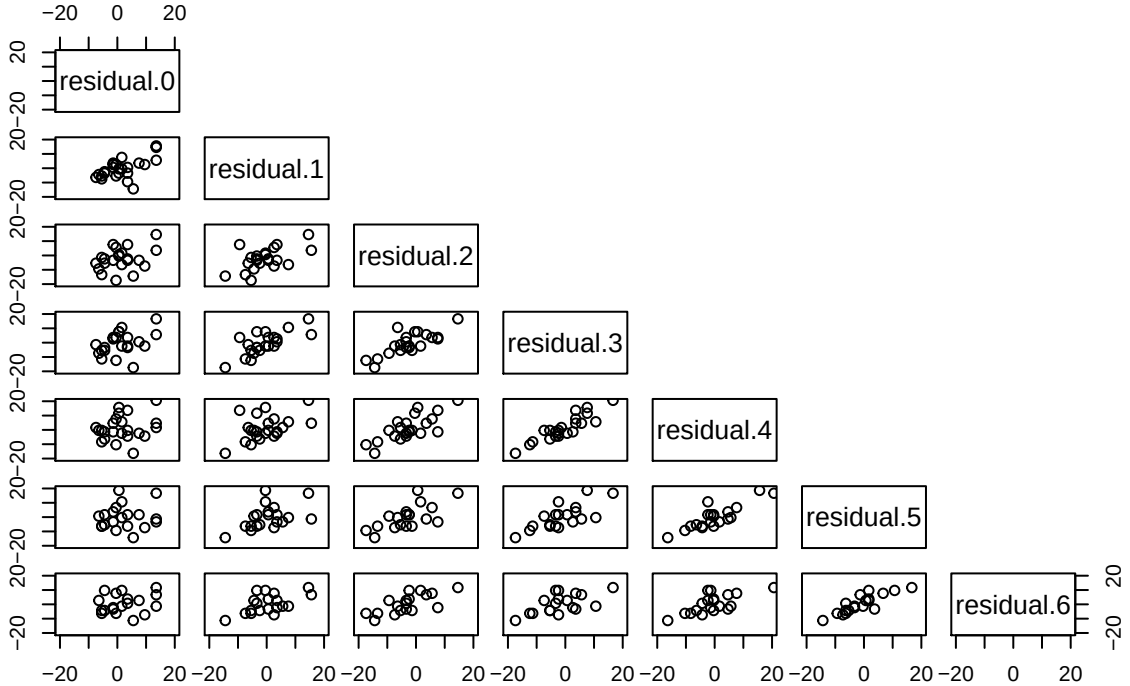
	residual.0	residual.1	residual.2	residual.3	residual.4	residual.5	residual.6
residual.0	6.21	NA	NA	NA	NA	NA	NA
residual.1	0.61	6.66	NA	NA	NA	NA	NA
residual.2	0.33	0.74	7.74	NA	NA	NA	NA
residual.3	0.24	0.74	0.70	8.38	NA	NA	NA
residual.4	0.17	0.64	0.60	0.90	7.83	NA	NA
residual.5	0.12	0.54	0.53	0.73	0.87	6.97	NA
residual.6	0.09	0.39	0.41	0.61	0.76	0.88	6.76

Augmentation/Treatment correlation



	residual.0	residual.1	residual.2	residual.3	residual.4	residual.5	residual.6
residual.0	6.19	NA	NA	NA	NA	NA	NA
residual.1	0.57	5.66	NA	NA	NA	NA	NA
residual.2	0.29	0.81	7.70	NA	NA	NA	NA
residual.3	-0.03	0.49	0.62	8.45	NA	NA	NA
residual.4	-0.07	0.37	0.38	0.87	7.62	NA	NA
residual.5	-0.07	0.36	0.27	0.71	0.91	6.32	NA
residual.6	-0.13	0.08	0.07	0.58	0.82	0.90	6.78

Control correlation



	residual.0	residual.1	residual.2	residual.3	residual.4	residual.5	residual.6
residual.0	6.29	NA	NA	NA	NA	NA	NA
residual.1	0.63	7.36	NA	NA	NA	NA	NA
residual.2	0.37	0.72	8.00	NA	NA	NA	NA
residual.3	0.46	0.91	0.80	8.19	NA	NA	NA
residual.4	0.34	0.81	0.80	0.92	7.99	NA	NA
residual.5	0.23	0.62	0.74	0.74	0.83	7.48	NA
residual.6	0.24	0.58	0.72	0.60	0.70	0.86	6.73

4. Missing data start to occur at week 2 and continue to get more frequently. Augmentation group has overall more missing value (both in number and percentage) than control group. For subjects with missing data, their baseline HDRS are higher in mean and larger in range compared to subjects without missing data.

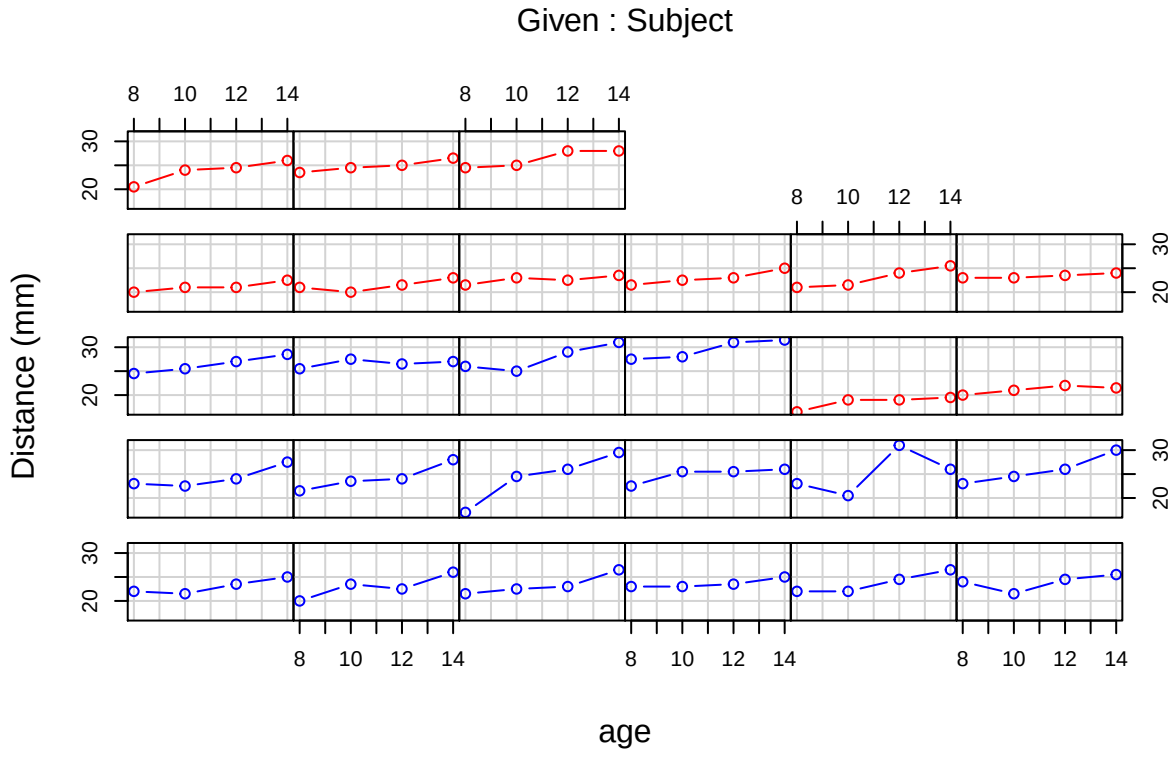
	N	Msng	Pct_Missing
Augmentation Baseline	26	0	0
Augmentation Week 1	26	0	0
Augmentation Week 2	26	3	12
Augmentation Week 3	26	5	19
Augmentation Week 4	26	6	23
Augmentation Week 5	26	7	27
Augmentation Week 6	26	7	27
Control Baseline	24	0	0
Control Week 1	24	0	0
Control Week 2	24	2	8
Control Week 3	24	2	8
Control Week 4	24	1	4
Control Week 5	24	4	17
Control Week 6	24	5	21

Part B

1. From the following table, we have the basic summary statistics for the distances at baseline and at the end of the study for males and females. We can see females tend to have shorter distances than males at the same time, and the distances at the endpoint tend to be longer than the distance at the baseline in the same sex.

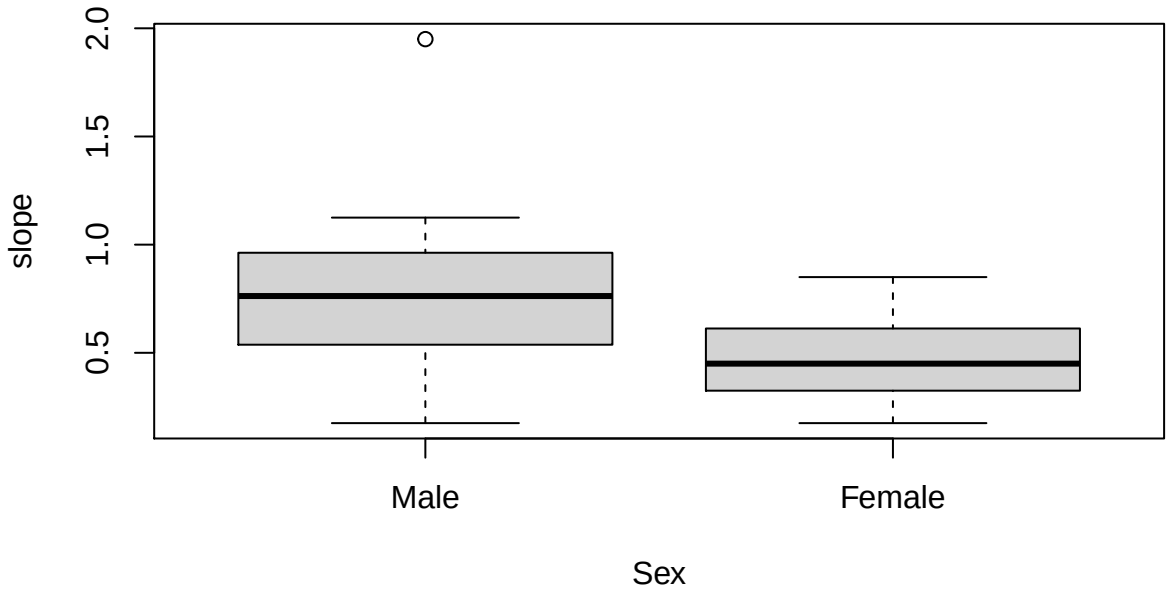
	N	Msng	Mean	Std Dev	Min	25%	Median	75%	Max
Male Baseline	16	0	22.875	2.453	17.0	21.875	23.00	24.125	27.5
Female Baseline	11	0	21.182	2.125	16.5	20.250	21.00	22.250	24.5
Male Endpoint	16	0	27.469	2.085	25.0	26.000	26.75	28.750	31.5
Female Endpoint	11	0	24.091	2.437	19.5	22.750	24.00	25.750	28.0

2. From the following table and plot, we can clearly see the slopes are all positive, which means the distance is all positively associated with age. Most of the observations have a slope between 0 and 1, meaning for the population that differs one year in age, the average distance difference is lower than 1 mm, with the elder group having a long distance. From the plot, we also can see female tends to have smaller slope compared to male.



N	Msng	Mean	Std Dev	Min	25%	Median	75%	Max
27	0	0.66	0.37	0.175	0.375	0.675	0.825	1.95

3. From the following box plot, we can see that the male population tends to have larger slopes than the female population. The difference between the two sex groups appears to be obvious. Therefore, there may be an association between sex and the rate of dental growth.



4. From the following test result, we estimated the mean dental growth rate for males is 0.784 mm per year, and the mean dental growth rate for the female is 0.480 mm per year. We estimate the difference in mean dental growth rate between two sex groups is 0.305 mm per year (95% CI: 0.055, 0.555). Our null hypothesis is the true difference in mean dental growth rate between two sex groups is zero, and our alternative hypothesis is the actual difference in mean dental growth rate is not zero. We have a p-value equals to 0.019, meaning if the null hypothesis is true, the possibility we will get this result or result is more extreme than that is 0.01906. Therefore, at the significance level of 0.05, we find a statistically significant association between sex and the dental growth rate. Thus, we have strong evidence for dental growth rate is associated with sex.

5. The following table is the estimated slope of the three analyses model. We interpret that:

Post: We estimated at the age of 14, the average distance of the male group will be 3.378 mm longer than the average distance of the female group.

Change: We estimated that the average difference between age eight and age 14 in the male group would be 1.685 mm longer than the average difference in the female group.

ANCOVA: We estimated that comparing male and female with the same distance at the age of eight, male population are estimated to have a distance that is 2.53 mm longer than the female at the age of 14.

In my opinion, I think the change model will be most useful in answering the question of whether there is a difference in growth rates between males and females. The post model is not helpful since the two sex groups have differences in the distance at the age of eight. ANCOVA can be useful when we have two observations in two sex but have the same distance at the age of eight. The change model can help us evaluate the difference in the mean dental growth rate in the two sex groups.

Post	Change	ANCOVA
-3.378	-1.685	-2.53

Code Appendix

```
### Setting up the packages, options we'll need:
library(knitr)
knitr::opts_chunk$set(echo = TRUE)
### ------------------------------------------------------------
### Reading in the data for part A.
library(nlme)
library(dplyr)
library(reshape2)
library(ggplot2)
library(joineR)
library(MASS)
aug <- read.csv("~/Desktop/R hw/augmentation.csv")
aug <- aug[, c("id", "Treatment_Group", "HD_t0", "HD_t1", "HD_t2",
               "HD_t3", "HD_t4", "HD_t5", "HD_t6")]
aug_long <- melt(aug, id=c("id","Treatment_Group"))
aug_long$week <- as.numeric(gsub("HD_t","",aug_long$variable))
### Reading in the data for part B
library(nlme)
data(Orthodont)
library(magrittr)
library(knitr)
library(uwIntroStats)
ortho_wide <- reshape(Orthodont,
                      direction="wide",
                      idvar = c("Subject", "Sex"),
                      timevar="age")
### ------------------------------------------------------------
### Q1
library(uwIntroStats)
summary <- descrip(aug[,c(-1,-2)], strata=aug$Treatment_Group)
tab1 <- as.data.frame(summary[c(seq(2,20,3), seq(3,21,3)),c(1:9)])
tab1 <- round(tab1, 3)
rownames(tab1) = c("Augmentation Baseline", "Augmentation Week 1", "Augmentation Week 2",
                   "Augmentation Week 3", "Augmentation Week 4", "Augmentation Week 5",
                   "Augmentation Week 6", "Control Baseline", "Control Week 1", "Control Week 2",
                   "Control Week 3", "Control Week 4", "Control Week 5", "Control Week 6")
knitr::kable(tab1)


aug_mean <- aggregate(value~Treatment_Group+week, data = aug_long,
                      function(x) mean(x, na.rm = T))
ggplot(aug_mean, aes(x=week, y=value, col=Treatment_Group))+
  labs(x="Week", y="Mean HDRS", col="Treatment Group",
       title = "Mean HDRS over time for two groups") +
  geom_line() + geom_point()

### ------------------------------------------------------------
### Q2
aug_long_sort <- aug_long[order(aug_long$id),]

# matplot(t(matrix(aug_long_sort$week, ncol=7, byrow=T)),
```

```
#         t(aug[,3:9]), type="l", lty=1,
#         xlab="Week", ylab="HDRS",
#         col=ifelse(aug$Treatment_Group=="Augmentation", "blue", "red"))
# legend("topright", c("Augmentation", "Control"), col=c("blue", "red"),
#        lty=1, cex=0.8)

ggplot(data = aug_long, aes(x=week, y=value, group=id, col=Treatment_Group)) +
  geom_line() + scale_color_manual(values=c('blue','red')) +
  theme_bw() + ylab("HDRS")
### ------------------------------------------------------------
### Q3
# Overall
aug_long_nona <- aug_long[!is.na(aug_long$value),]

model <- lm(value ~ week, data=aug_long_nona)
aug_long_nona$residual <- model$residuals
aug_long_nona <- data.frame(aug_long_nona)
aug_long_nona_wide <- reshape(aug_long_nona[,c("week", "id", "residual")],
                              direction="wide",
                              idvar = c("id"),
                              timevar="week")
pairs(aug_long_nona_wide[,2:8], upper.panel = NULL, diag.panel = NULL,
      xlim=c(-20,20), ylim=c(-20,20), main="Overall correlation")

mat <- cor(aug_long_nona_wide[,2:8], use = "complete.obs")
mat[upper.tri(mat)] <- NA
diag(mat) <- sqrt(diag(cov(aug_long_nona_wide[,2:8], use = "complete.obs")))
knitr::kable(round(mat,2))

# Trt group
aug_trt <- aug_long_nona[aug_long_nona$Treatment_Group=="Augmentation",]
model_trt <- lm(value ~ week, data=aug_trt)
aug_trt$residual <- model_trt$residuals
aug_trt <- data.frame(aug_trt)
aug_trt_wide <- reshape(aug_trt[,c("week", "id", "residual")],
                              direction="wide",
                              idvar = c("id"),
                              timevar="week")
pairs(aug_trt_wide[,2:8], upper.panel = NULL, diag.panel = NULL,
      xlim=c(-20,20), ylim=c(-20,20), main="Augmentation/Treatment correlation")

mat_trt <- cor(aug_trt_wide[,2:8], use = "complete.obs")
mat_trt[upper.tri(mat_trt)] <- NA
diag(mat_trt) <- sqrt(diag(cov(aug_trt_wide[,2:8], use = "complete.obs")))
knitr::kable(round(mat_trt,2))

# Control group
aug_c <- aug_long_nona[aug_long_nona$Treatment_Group=="Control",]
model_c <- lm(value ~ week, data=aug_c)
aug_c$residual <- model_c$residuals
aug_c <- data.frame(aug_c)
aug_c_wide <- reshape(aug_c[,c("week", "id", "residual")],
                              direction="wide",
```

```
                           idvar = c("id"),
                           timevar="week")
pairs(aug_c_wide[,2:8], upper.panel = NULL, diag.panel = NULL,
      xlim=c(-20,20), ylim=c(-20,20), main="Control correlation")

mat_c <- cor(aug_c_wide[,2:8], use = "complete.obs")
mat_c[upper.tri(mat_c)] <- NA
diag(mat_c) <- sqrt(diag(cov(aug_c_wide[,2:8], use = "complete.obs")))
knitr::kable(round(mat_c,2))

### ------------------------------------------------------------
### Q4
# knitr::kable(aug_long_na %>% count(Treatment_Group, week))
tab1$Pct_Missing <- tab1$Msng/tab1$N*100
knitr::kable(round(tab1[,c(1,2,10),2]))

aug_long_na <- aug_long[is.na(aug_long$value),]
aug_na_obs <- sort(unique(aug_long_na$id))
aug$na_status <- "No Missing"
aug[aug_na_obs,]$na_status <- "Missing"

ggplot(aug, aes(x=na_status,y=HD_t0, color=na_status)) +
  geom_boxplot(outlier.shape=NA)+
  geom_point(aes(shape=Treatment_Group), position=position_jitterdodge()) +
  labs(x="Missing Status", y="Baseline HDRS", col=c("Missing Status", "Treatment"),
       title = "Baseline HDRS for subjects with and without missing data")

### ------------------------------------------------------------
### QB1
summary = by(ortho_wide[,c(3,6)], INDICES = ortho_wide$Sex, FUN = descrip)
tab1 = matrix(NA, nrow =4 , ncol = 9 )
tab1[c(1,3),] = summary[[1]][1:2,1:9]
tab1[c(2,4),] = summary[[2]][1:2,1:9]
tab1 = data.frame(tab1)
tab1 = round(tab1, 3)
colnames(tab1) = c("N","Msng","Mean","Std Dev","Min","25%","Median","75%", "Max")

rownames(tab1) = c("Male Baseline","Female Baseline",
                   "Male Endpoint","Female Endpoint")
knitr::kable(tab1)

### ------------------------------------------------------------
### QB2
coplot(distance ~ age | Subject, data = Orthodont,
       show.given = FALSE, ylab = "Distance (mm)", type = "b",
       col=ifelse(Orthodont$Sex=="Male", "blue", "red"))
slopes <- by(Orthodont, INDICES = Orthodont$Subject,
             FUN = function(data.set){
                lm(distance~age, data = data.set)$coef[2]})
tab2 = matrix(descrip(slopes)[1:9],nrow = 1)
colnames(tab2) = c("N","Msng","Mean","Std Dev","Min","25%","Median","75%", "Max")
tab2 = round(tab2,3)
knitr::kable(tab2)
```

```r
dat.slopes <- data.frame("Subject" = names(slopes),
                         "slope" = as.numeric(slopes))
dat.summary <- merge(dat.slopes, ortho_wide, by = "Subject")
### ------------------------------------------------------------
### QB3
boxplot(slope ~ Sex, data = dat.summary)
### ------------------------------------------------------------
### QB4
t.test(slope ~ Sex, data = dat.summary)
### ------------------------------------------------------------
### QB5
ortho_wide$diff = ortho_wide$distance.14-ortho_wide$distance.8
post = lm(distance.14~Sex, data = ortho_wide)$coef[2]
change = lm(diff~Sex, data = ortho_wide)$coef[2]
ancova = lm(distance.14~distance.8+Sex, data = ortho_wide)$coef[3]
tab3 = matrix(c(post, change, ancova),nrow = 1)
colnames(tab3) = c("Post", "Change", "ANCOVA")
tab3 = round(tab3,3)
knitr::kable(tab3)
```