Clinical trails hw4

Jiaying Liu

December 4, 2018

Homework Set 4

Question 1

This project relates to the analysis of the data in the file GWU-A1C-Graphs.csv. This is a longitudinal trial that observed levels of HbA1c from patients with type 1 diabetes at various visits including a baseline visit (Visit.nmber=1). The objective of the trial is to demonstrate that the treatment would lower patients' HbA1c more than the control. We will evaluate the treatment effective based on visits 1-9, so you will need to delete data collected at visit 10 and later visits. Labels of the column variables are given below. Visit.number: This gives the number for each visit the patient paid to the clinic. 1 is the baseline visit prior to the start of intervention, thus 2 is the first visit after intervention. HBA1C.level: This is the primary endpoint of the trial. HbA1c is a biomarker for diabetes which is interpreted as the three-month average of blood sugar level in the body. This column gives the level of HbA1c at each visit. PATID: This gives the unique identication number of each patient. The same number is repeated because of multiple visits. Control.treatment: Tells you which arm the patients were assigned to. Age: Age of patients at enrollment to the study.

(1)

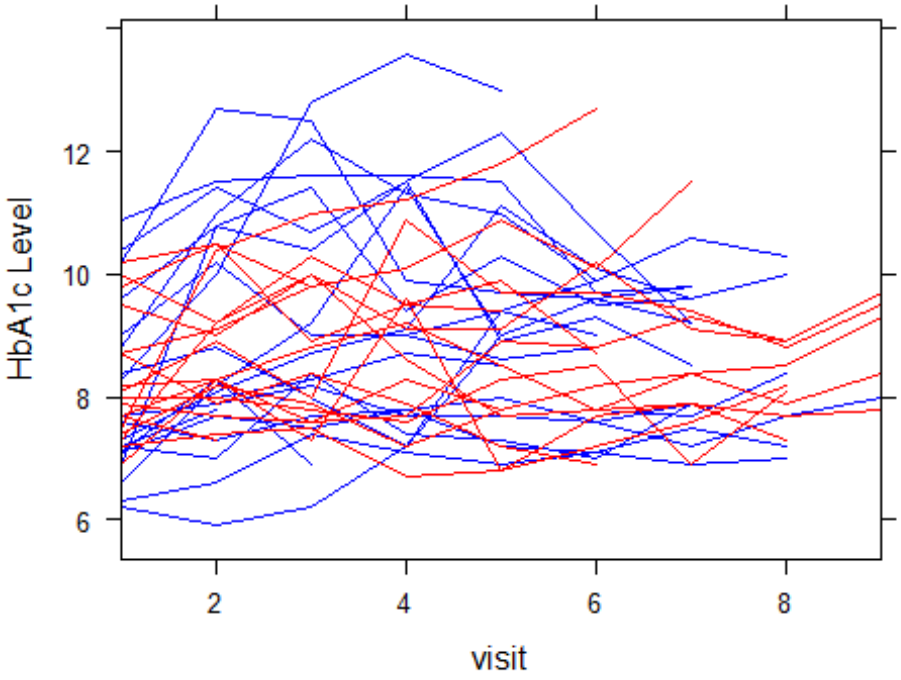
Randomly select 20 patients from each treatment arm. For each arm, make a spaghetti plot of the HbA1c levels of the 20 patients against their visits. Put the two plots in the same figure with the same scales. Use red color for treatment arm and blue for control arm. What can you comment on the comparison of the two arms?

```
#After deleting the visit greater than 9, and read in dataset
MyData=read.csv('d:/GWU-A1C-Graphs.csv')
#Get variable names
Var.list = names(MyData)
#Get partial data by variables
Partial.Data1 = MyData[,c("Visit.number", "HBA1C.level", "PATID")]
Partial.Data2 = MyData[MyData[, "Visit.number"]==1, c("Visit.number",
"HBA1C.level", "PATID")]
Partial.Data3 = MyData[MyData[, "Control.treatment"]=="Treatment",
c("Visit.number", "HBA1C.level", "PATID")]
```

```
Trt.Data = MyData[MyData[,"Control.treatment"]=="Treatment",]
Ctr.Data = MyData[MyData[,"Control.treatment"]=="Control",]
#Get Patient IDs
Unique.ids = unique(MyData[,"PATID"])
#Get Patient IDs by treatment
Trt.ids = unique(Trt.Data[,"PATID"])
Ctr.ids = unique(Ctr.Data[,"PATID"])
```

Random choose the dataset

```
set.seed(1234)
#Random selection of 20 patients form control/treatment
Ctr20 = sample(Ctr.ids, 20)
Trt20 = sample(Trt.ids, 20)
#Get data from selected patients
Ctr20.Data = Ctr.Data[Ctr.Data[, "PATID"] %in% Ctr20,]
Trt20.Data = Trt.Data[Trt.Data[, "PATID"] %in% Trt20,]
library(lattice)
library(latticeExtra)

## Loading required package: RColorBrewer

#Spaghetti plots
##To put on the same figure
##common scale
x.min=1
x.max=9
y.min=min(MyData[,"HBA1C.level"])
y.max=max(MyData[,"HBA1C.level"])
a=xyplot(HBA1C.level ~ Visit.number, groups = PATID, data =Ctr20.Data,
xlim=c(x.min,x.max), ylim=c(y.min,y.max),type ="l", xlab="visit",
ylab="HbA1c Level", col='blue')
b=xyplot(HBA1C.level ~ Visit.number, groups = PATID, data =Trt20.Data,
xlim=c(x.min,x.max), ylim=c(y.min,y.max),type ="l", xlab="visit",
ylab="HbA1c Level", col='red')
a + as.layer(b)
```



First set.seed=1234 to make sure everytime the sampling result is the same. Aftering sampling 20 patient in each group and put it in the same scale, we can see that the red line which represent the treatment group are generally lower than blue line which represent the control group. However, the difference is so small. So, we could not know if the treatment would lower patients' HbA1c more than the control.

(2)

Again using the HbA1c levels from (1), compute the sample means of HbA1c at each visit for each treatment arm. Make the mean plots by treatment group. Use red color for treatment arm and blue for control arm. What can you comment on the comparison of the two arms?

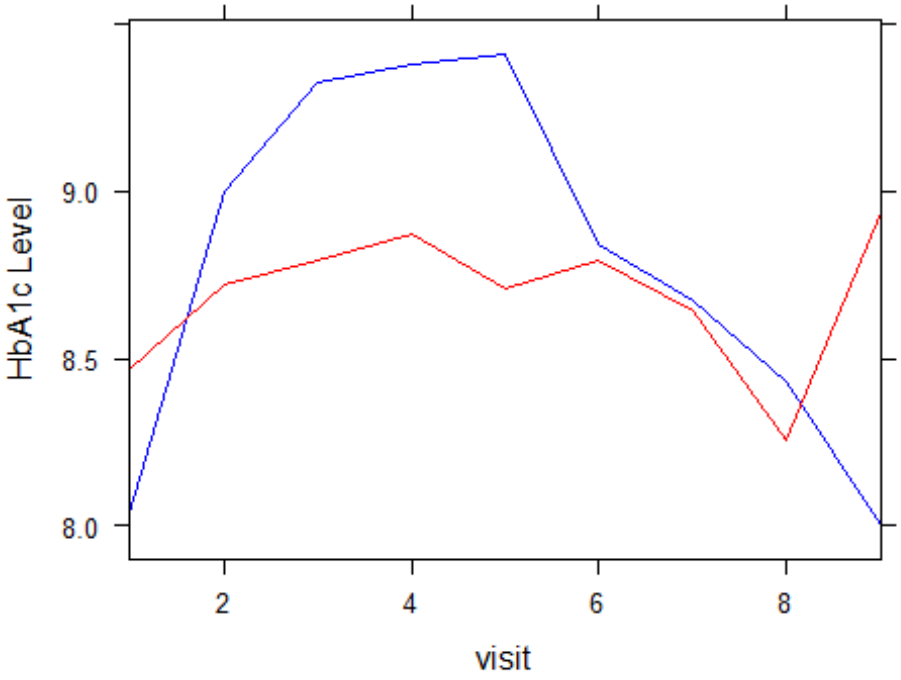
```
##Get column mean by visit and treatment
Mean.A1c = aggregate(HBA1C.level ~ Visit.number + Control.treatment,
data=Ctr20.Data, mean)
Mean.A1c

##   Visit.number Control.treatment HBA1C.level
## 1            1           Control    8.040000
## 2            2           Control    9.000000
## 3            3           Control    9.321053
## 4            4           Control    9.378947
## 5            5           Control    9.410526
## 6            6           Control    8.841176
```

```
## 7            7           Control    8.671429
## 8            8           Control    8.433333
## 9            9           Control    8.000000
```

```
Mean.A1t = aggregate(HBA1C.level ~ Visit.number + Control.treatment,
data=Trt20.Data, mean)
Mean.A1t
```

```
##   Visit.number Control.treatment HBA1C.level
## 1            1         Treatment    8.465000
## 2            2         Treatment    8.717647
## 3            3         Treatment    8.793750
## 4            4         Treatment    8.873333
## 5            5         Treatment    8.706667
## 6            6         Treatment    8.792308
## 7            7         Treatment    8.640000
## 8            8         Treatment    8.255556
## 9            9         Treatment    8.940000
```

```
#get mean plot
x.min=1
x.max=9
y.min=min(MyData[,"HBA1C.level"])
y.max=max(MyData[,"HBA1C.level"])
a=xyplot(HBA1C.level ~ Visit.number, data =Mean.A1c,
xlim=c(x.min,x.max), ylim=c(y.min,y.max),type ="l", xlab="visit",
ylab="HbA1c Level", col=4)
b=xyplot(HBA1C.level ~ Visit.number, data =Mean.A1t,
xlim=c(x.min,x.max), ylim=c(y.min,y.max),type ="l", xlab="visit",
ylab="HbA1c Level", col=2)
a + as.layer(b)
```

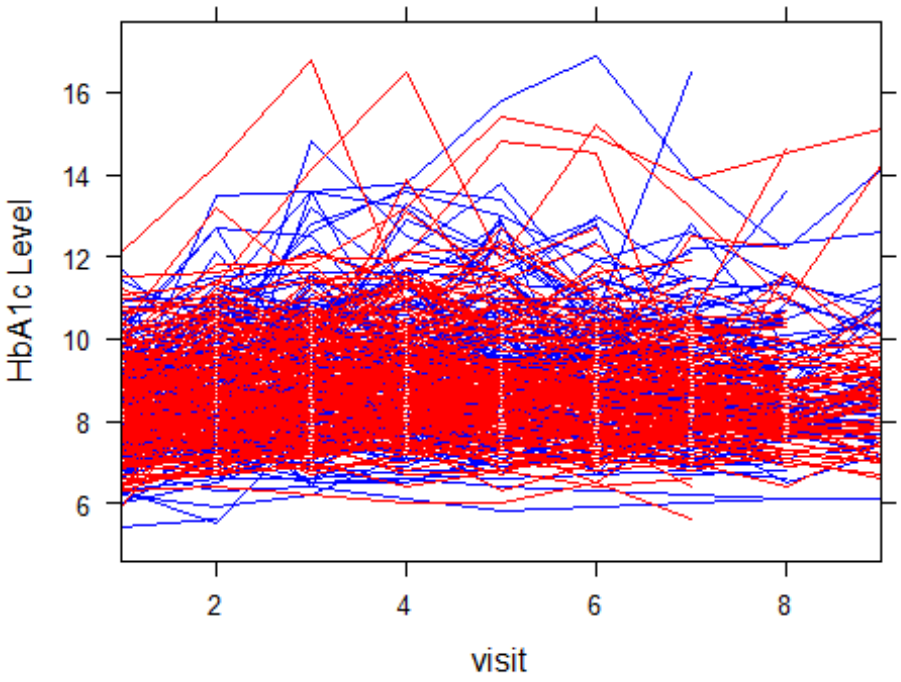


From the above plot for the mean of treatment and control group. We can see that at the baseline level(visit=1), the level of HbAlc is greater than control group. In the following visit, it is obvious that the level of HbAlc in treatement is lower than the control. But when the visist come to 6 or more, the preformance of treatment is not good. However the mean is not that representive because it contain outlier.

(3)

Repeat the analysis (1) and (2), but this time using all patients. Comment on your findings.Are the findings similar to those from (1) and (2)?

```
#all data
x.min=1
x.max=9
y.min=min(MyData[,"HBA1C.level"])
y.max=max(MyData[,"HBA1C.level"])
a=xyplot(HBA1C.level ~ Visit.number, groups = PATID, data =Ctr.Data,
xlim=c(x.min,x.max), ylim=c(y.min,y.max),type ="l", xlab="visit",
ylab="HbA1c Level", col='blue')
b=xyplot(HBA1C.level ~ Visit.number, groups = PATID, data =Trt.Data,
xlim=c(x.min,x.max), ylim=c(y.min,y.max),type ="l", xlab="visit",
ylab="HbA1c Level", col='red')
a + as.layer(b)
```

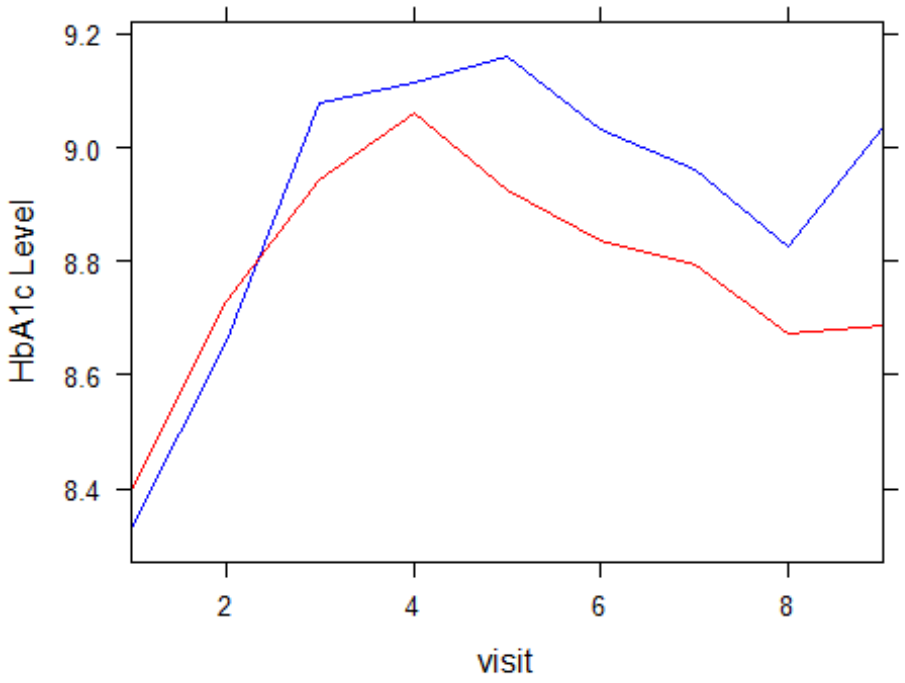


According to the plot above, most patients in the treatment arm(red) have a lower level of HbA1c than the control arm(blue). So, we may conclude that the treatment would lower patients' HbA1c more than the control.

```
#all data
Mean.A1c = aggregate(HBA1C.level ~ Visit.number + Control.treatment,
data=Ctr.Data, mean)
Mean.A1c

##   Visit.number Control.treatment HBA1C.level
## 1            1           Control    8.329032
## 2            2           Control    8.657923
## 3            3           Control    9.077473
## 4            4           Control    9.115642
## 5            5           Control    9.162069
## 6            6           Control    9.032716
## 7            7           Control    8.961806
## 8            8           Control    8.826852
## 9            9           Control    9.037500

Mean.A1t = aggregate(HBA1C.level ~ Visit.number + Control.treatment,
data=Trt.Data, mean)
Mean.A1t

##   Visit.number Control.treatment HBA1C.level
## 1            1         Treatment    8.393939
## 2            2         Treatment    8.730526
## 3            3         Treatment    8.943243
```

```
## 4            4         Treatment    9.062222
## 5            5         Treatment    8.926857
## 6            6         Treatment    8.836646
## 7            7         Treatment    8.792908
## 8            8         Treatment    8.671845
## 9            9         Treatment    8.687500
```

```
x.min=1
x.max=9
y.min=min(MyData[,"HBA1C.level"])
y.max=max(MyData[,"HBA1C.level"])
a=xyplot(HBA1C.level ~ Visit.number, data =Mean.A1c,
xlim=c(x.min,x.max), ylim=c(y.min,y.max),type ="l", xlab="visit",
ylab="HbA1c Level", col=4)
b=xyplot(HBA1C.level ~ Visit.number, data =Mean.A1t,
xlim=c(x.min,x.max), ylim=c(y.min,y.max),type ="l", xlab="visit",
ylab="HbA1c Level", col=2)
a + as.layer(b)
```



According to the mean plot above, the treatment arm(red) tends to have a lower HbA1c level than the control arm(blue) as the increasing of visit.Plus, the different of two group in the baseline level is become bigger and bigger since the visit change. So, we may conclude that the treatment would lower patients' HbA1c more than the control.

(4)

For the data in (3), use a linear mixed effect model to test whether the two arms have the same rate of change.

For the mixed linear model: let Y=HbAlc T=visit x: treatment=1; control= 0 So we have $E(Y) = \beta_0 + \beta_1 X + \beta_2 T + \beta X * T$ when X=0 control group $E(Y) = \beta_0 + \beta_2 T$ When X=1 treatment group $E(Y) = (\beta_0 + \beta_1) + (\beta_2 + \beta)T$ So in order to test the effect of treatment group, we actually is going to test $ H_0: =0$

```
library(lme4)

## Loading required package: Matrix

m2=lmer(HBA1C.level~Control.treatment+Visit.number+Control.treatment*Vi
sit.number+(Visit.number | PATID),data=MyData)
summary(m2)

## Linear mixed model fit by REML ['lmerMod']
## Formula:
## HBA1C.level ~ Control.treatment + Visit.number + Control.treatment *
##     Visit.number + (Visit.number | PATID)
##    Data: MyData
##
## REML criterion at convergence: 7928.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.5345 -0.4771 -0.0494  0.4109  5.8248
##
## Random effects:
##  Groups   Name         Variance Std.Dev. Corr
##  PATID    (Intercept)  1.16988  1.0816
##           Visit.number 0.02398  0.1549   -0.20
##  Residual              0.64166  0.8010
## Number of obs: 2755, groups:  PATID, 388
##
## Fixed effects:
##                                           Estimate Std. Error t value
## (Intercept)                                8.51649    0.09218  92.386
## Control.treatmentTreatment                 0.11409    0.12868   0.887
## Visit.number                               0.09619    0.01550   6.205
## Control.treatmentTreatment:Visit.number -0.04174    0.02188  -1.908
##
## Correlation of Fixed Effects:
##             (Intr) Cntr.T Vst.nm
## Cntrl.trtmT -0.716
## Visit.numbr -0.432  0.310
## Cntrl.tT:V.  0.306 -0.432 -0.709
```

From above imformation of estimation, we need to evalute the significance of which is the effect of control.treatment:visit.number. we find out that the t value is -1.908 and the absolute value of t is 1.908 which is smaller than 1.96 (the value of $t_{1-\alpha/2}$) so we can not reject the original hypothesis which mean the two arm have same range of change.

```
library(nlme)
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:lme4':
##
##     lmList
```

```
mydata=na.omit(MyData)
model_3=lme(HBA1C.level~Control.treatment+Visit.number+Control.treatmen
t*Visit.number,data=mydata,random=~Visit.number |PATID, method='ML')
summary(model_3)
```

```
## Linear mixed-effects model fit by maximum likelihood
##  Data: mydata
##        AIC      BIC    logLik
##   7924.876 7972.245 -3954.438
##
## Random effects:
##  Formula: ~Visit.number | PATID
##  Structure: General positive-definite, Log-Cholesky parametrization
##              StdDev    Corr
## (Intercept)  1.0775579 (Intr)
## Visit.number 0.1539418 -0.201
## Residual     0.8011060
##
## Fixed effects: HBA1C.level ~ Control.treatment + Visit.number +
Control.treatment *      Visit.number
##                                            Value  Std.Error   DF
t-value
## (Intercept)                                8.516613 0.09199513 2365
92.57678
## Control.treatmentTreatment                 0.114041 0.12841762  386
0.88805
## Visit.number                               0.096148 0.01546007 2365
6.21911
## Control.treatmentTreatment:Visit.number -0.041743 0.02181746 2365 -
1.91330
##                                           p-value
## (Intercept)                                0.0000
## Control.treatmentTreatment                 0.3751
## Visit.number                               0.0000
## Control.treatmentTreatment:Visit.number    0.0558
```

```
##  Correlation:
##                                             (Intr) Cntr.T Vst.nm
## Control.treatmentTreatment                  -0.716
## Visit.number                                -0.432  0.310
## Control.treatmentTreatment:Visit.number  0.306 -0.432 -0.709
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -4.53404508 -0.47770726 -0.04930967  0.41219470  5.82694258
##
## Number of Observations: 2755
## Number of Groups: 388
```

By the other statement lme which is also be used in mixed model. After deleting all the missing observation, we dit the model and find out that the p-value control.treatment:visitnumber is still greater than 0.05, so we can say the beta is equal to zero and two arm have same range of change.

Question 2

The sample size of 400 patients was based on comparison of the rate of change in HbA1c level between treatment and control. A compound symmetry correlation among repeated HbA1c was assumed with a common correlation coefficient of 0.5. A common standard deviation of 1.2. Suppose everything else remains the same, what would the sample size be, respectively, if the correlation coefficient is 0.4, 0.45, 0.55, 0.60, 0.65, 0.70 and the standard deviation is 1, 1.1, 1.4, 1.5? Describe the relationship between the correlation and the sample size

From the lecture note we have the formular $m = \frac{2(z_{1-\alpha/2}+z_{1-\beta})^2(1+(n-1)\rho)\sigma^2}{ns_x^2d^2}$ With this formular we can calculate the different between correlation and sample size

```
constant=400/(1-0.5)/1.2/1.2
corr=c(0.4,0.45,0.55,0.6,0.65,0.7)
sd=c(1,1.1,1.4,1.5)
result=c()
for (correlation in corr){
  for ( standard_deviation in sd){
    sample.size=constant*(1-
correlation)*standard_deviation*standard_deviation

result=rbind(result,data.frame(correlation,standard_deviation,sample.si
ze))

  }
}
result
```

```
##    correlation standard_deviation sample.size
## 1         0.40                1.0    333.3333
## 2         0.40                1.1    403.3333
## 3         0.40                1.4    653.3333
## 4         0.40                1.5    750.0000
## 5         0.45                1.0    305.5556
## 6         0.45                1.1    369.7222
## 7         0.45                1.4    598.8889
## 8         0.45                1.5    687.5000
## 9         0.55                1.0    250.0000
## 10        0.55                1.1    302.5000
## 11        0.55                1.4    490.0000
## 12        0.55                1.5    562.5000
## 13        0.60                1.0    222.2222
## 14        0.60                1.1    268.8889
## 15        0.60                1.4    435.5556
## 16        0.60                1.5    500.0000
## 17        0.65                1.0    194.4444
## 18        0.65                1.1    235.2778
## 19        0.65                1.4    381.1111
## 20        0.65                1.5    437.5000
## 21        0.70                1.0    166.6667
## 22        0.70                1.1    201.6667
## 23        0.70                1.4    326.6667
## 24        0.70                1.5    375.0000
```

From the above result, there is a negative relationship between the correlation coefficient and the sample size, and there is a positive relationship between the standard deviation(sd) and the sample size.