Homework 03 Solutions

2022-02-16

```
library(arm)  ## includes lme4, MASS, Matrix
library(ggplot2); theme_set(theme_bw())
library(gridExtra) ## to arrange ggplots...
library(HLMdiag)
```

```
## Warning: package 'HLMdiag' was built under R version 4.1.2
```

```
library(cAIC4)
```

```
## Warning: package 'cAIC4' was built under R version 4.1.2
```

Problem 1.

Return to the CD4 data from HW02, and consider the model

```
lmer.1 <- lmer(sqrt.CD4PCT ~ 1 + VISIT * treatmnt + (1 + VISIT | newpid),
               data=cd4)
```

Bring together all the residuals (except random effects residuals) that we have talked about, using library(HLMdiag):

```
cd4 <- read.csv("allvar.csv",header=T)

## It's worth doing a little exploration of the data here,
## especially to look at missing values.

apply(cd4,2,function(x) mean(is.na(x)))
```

```
##       VISIT      newpid       VDATE      CD4PCT         arv      visage
## 0.000000000 0.000000000 0.000000000 0.142743222 0.103668262 0.109250399
##    treatmnt      CD4CNT     baseage
## 0.000000000 0.146730463 0.007177033
```

```
## we only need the variables newpid, VISIT, visage, CD4PCT and treatmnt
## so let's get rid of some (but not all) missing data problems
## by deleting the other variables

cd4 <- with(cd4,data.frame(newpid=newpid,
                           VISIT=VISIT,
                           visage=visage,
                           CD4PCT=CD4PCT,
                           treatmnt=treatmnt))

## and since we still have some missing data, we will just
## delete the rows that continue to have NA's (not a great
## practice in general, but good enough for this exercise...)
cd4 <- cd4[!apply(cd4,1,function(x) any(is.na(x))),]
```

```
## set up the sqrt of the response variable...
cd4$sqrt.CD4PCT <- sqrt(cd4$CD4PCT)

## and for coloring ggplot elements it will better if treatmnt is a factor...
cd4$treatmnt <- as.factor(cd4$treatmnt)

## select the first 12 kids...
first.12 <- (cd4$newpid <= 12)

## Fit the base model for this exercise...
display(lmer.1 <- lmer(sqrt.CD4PCT ~ 1 + VISIT * treatmnt + (1 + VISIT | newpid),
               data=cd4))
```

```
## lmer(formula = sqrt.CD4PCT ~ 1 + VISIT * treatmnt + (1 + VISIT |
##     newpid), data = cd4)
##                 coef.est coef.se
## (Intercept)      4.71     0.13
## VISIT           -0.03     0.01
## treatmnt2        0.14     0.19
## VISIT:treatmnt2  0.01     0.01
##
## Error terms:
##  Groups   Name        Std.Dev. Corr
##  newpid   (Intercept) 1.40
##           VISIT       0.05     -0.10
##  Residual             0.72
## ---
## number of obs: 1075, groups: newpid, 251
## AIC = 3149.4, DIC = 3093
## deviance = 3113.2
```

```
r.1 <- hlm_resid(lmer.1,level=1,include.ls=F)
r.1s <- hlm_resid(lmer.1,level=1,include.ls=F,standardize=T)
r.2 <- hlm_resid(lmer.1,level="newpid",include.ls=F)
r.2s <- hlm_resid(lmer.1,level="newpid",include.ls=F,standardize=T)
names(r.1)
```

```
## [1] "id"          "sqrt.CD4PCT" "VISIT"       "treatmnt"    "newpid"
## [6] ".resid"      ".fitted"     ".mar.resid"  ".mar.fitted"
```

```
names(r.1s)
```

```
## [1] "id"              "sqrt.CD4PCT"     "VISIT"           "treatmnt"
## [5] "newpid"          ".std.resid"      ".fitted"         ".chol.mar.resid"
## [9] ".mar.fitted"
```

```
names(r.2)
```

```
## [1] "newpid"           ".ranef.intercept" ".ranef.visit"
```

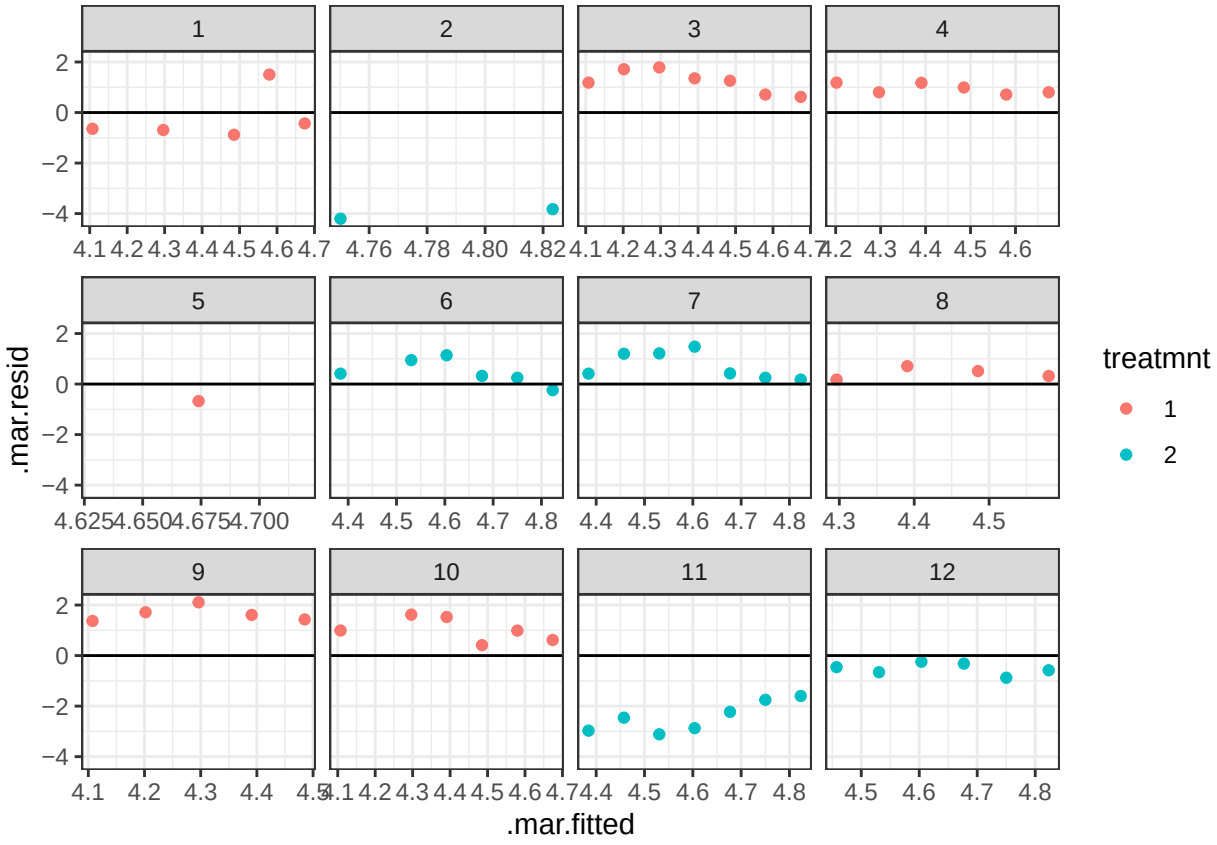
```
names(r.2s)
```

```
## [1] "newpid"               ".std.ranef.intercept" ".std.ranef.visit"
```

Problem 1(a).

Make a facets plot of the marginal residuals, as a function of the marginal fitted values (use `scales="free_x"` if needed to make the plot legible). Explain in a sentence or two why a facets plot is not very useful for assessing model fit for this problem, whether we look at the first 12 children, or all 251 children).

```
ggplot(r.1[first.12,],aes(x=.mar.fitted,y=.mar.resid)) +
  facet_wrap( ~ newpid,scales="free_x") +
  geom_point(aes(color=treatmnt)) +
  geom_abline(intercept=0,slope=0)
```

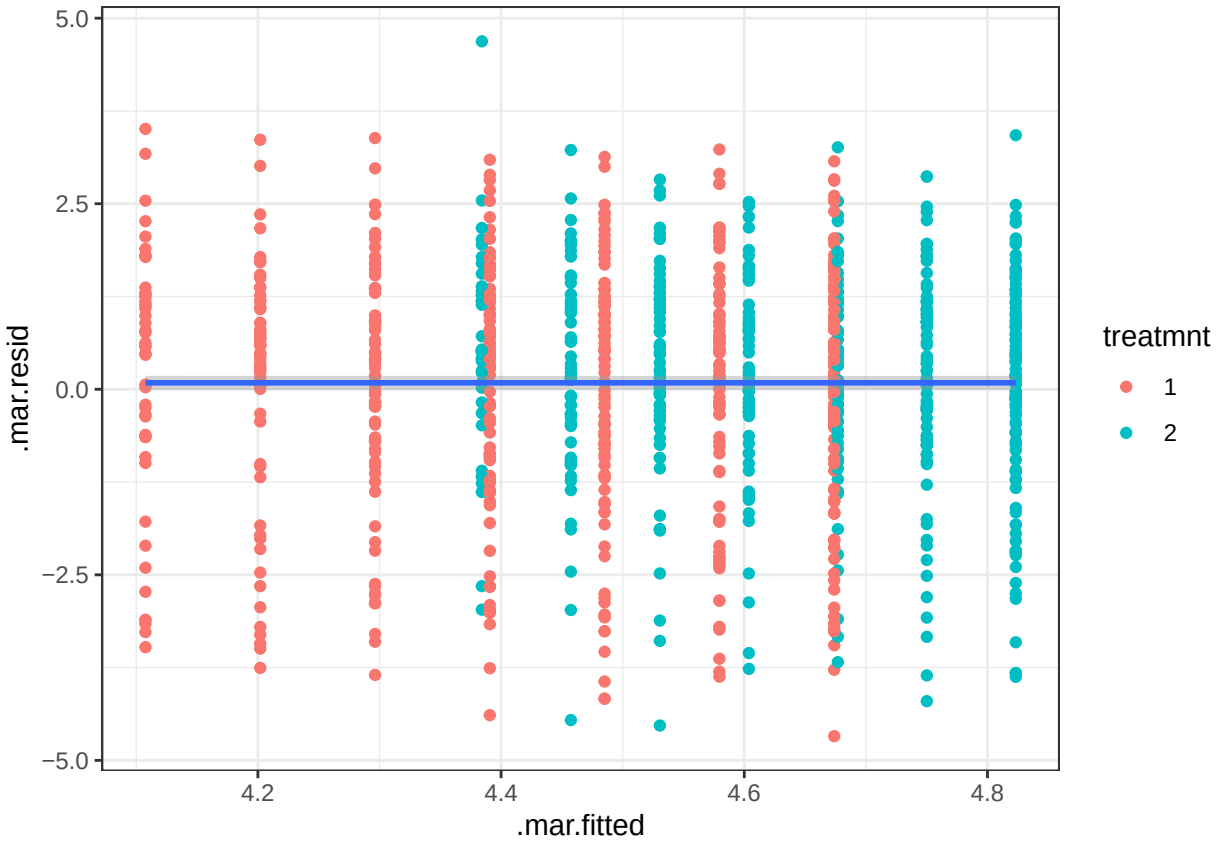


There are very few data points per child, and so the residual plot within facet for each child contains almost no information about fit. (Having said that, there are some hints of misfit, to the extent that several individual children's residuals are entirely above, or entirely below, the zero line.)

Problem 1(b)

Make an ungrouped (that is, no facets) scatter plot of marginal residuals as a function of marginal fitted values, using the full data set (not just the first 12 children). Color the points for `treatmnt=1` kids and treatmnt=2 kids with different colors. Overlay a smooth fit (`geom_smooth` is the easist to use here).

```
ggplot(r.1,aes(x=.mar.fitted,y=.mar.resid)) +
  geom_point(aes(color=treatmnt)) +
  geom_smooth()
```



Explain, in a couple of sentences (optionally with some math):

- What is causing the dominant structure in this plot, and why that dominant structure is essentially irrelevant for checking the relationship between `sqrt.CD4PCT` and `VISIT`;

  The dominant structure in the data is the way the residuals are grouped horizontally into 14 groups—7 groups for `treatmnt=1` and 7 for `treatmnt=2`. They are caused by the fact that our main predictor variable is `VISIT`, which only takes on 7 values. The spacing of the groups is not the same between `treatmnt=1` and `treatmnt=2` because the model contains a `treatmnt` × `VISIT` interaction, so that the slope on `VISIT` is different for children in the `treatmnt=1` and `treatmnt=2` groups.

  This does not affect our assessment of the fit of the model, since to check the model fit we are interested in vertical patterns (trends, curves, outliers, changing variance, etc.) in the residuals, not horizontal (grouping) patterns.

  and

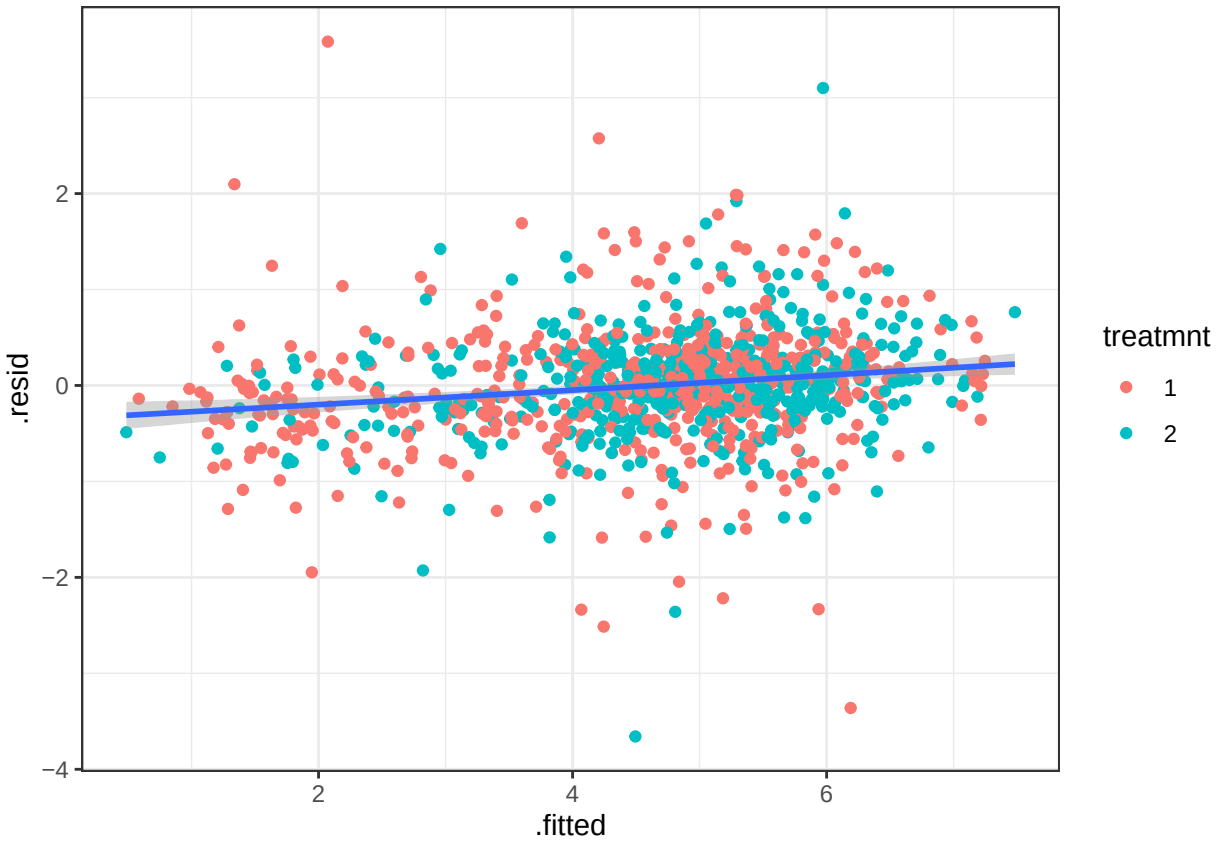
- What in this plot makes you happy or unhappy about having a linear relationship between `sqrt.CD4PCT` and `VISIT` in the model.

  The plot actually looks pretty wonderful. The variance of the residuals looks pretty much constant, as a function of the fitted values, and the trend modeled with the smooth really does look like a horizontal line at 0.

Problem 1(c)

Make an ungrouped (that is, no facets) scatter plot of conditional residuals as a function of conditional fitted values, using the full data set (not just the first 12 children). Color the points for `treatmnt=1` kids and `treatmnt=2` kids with different colors. Overlay a smooth fit (`geom_smooth` is the easist to use here).

```
ggplot(r.1,aes(x=.fitted,y=.resid)) +
  geom_point(aes(color=treatmnt)) +
  geom_smooth()
```



Explain, in a couple of sentences (optionally with some math):

- Why the dominant structure in the marginal residuals is not also present in this plot of conditional residuals

  The marginal residuals plot shows

  $$\begin{aligned} \text{.mar.resid}_i &= \text{sqrt.CD4PCT}_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot \text{VISIT}_i - \hat{\beta}_3 \cdot \text{treatmnt}_{j[i]} - \hat{\beta}_4 \cdot \text{VISIT}_i \cdot \text{treatmnt}_{j[i]} \\ &\text{vs.} \\ \text{.mar.fitted}_i &= \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{VISIT}_i + \hat{\beta}_3 \cdot \text{treatmnt}_{j[i]} + \hat{\beta}_4 \cdot \text{VISIT}_i \cdot \text{treatmnt}_{j[i]} \end{aligned}$$

  and because ***VISIT*** *and* ***treatmnt*** *are both discrete, the fitted values will only take on a few discrete values, causing the horizontal grouping in the marginal plot.*

  The conditional residuals plot shows

  $$\begin{aligned} \text{.resid}_i &= \text{sqrt.CD4PCT}_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot \text{VISIT}_i - \hat{\beta}_3 \cdot \text{treatmnt}_{j[i]} - \hat{\beta}_4 \cdot \text{VISIT}_i \cdot \text{treatmnt}_{j[i]} \\ &\quad -\eta_{0j[i]} - \eta_{1j[i]} \cdot \text{VISIT}_i \\ &\text{vs.} \\ \text{.fitted}_i &= \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{VISIT}_i + \hat{\beta}_3 \cdot \text{treatmnt}_{j[i]} + \hat{\beta}_4 \cdot \text{VISIT}_i \cdot \text{treatmnt}_{j[i]} \\ &\quad +\eta_{0j[i]} + \eta_{1j[i]} \cdot \text{VISIT}_i \end{aligned}$$

  The additional continuous (because the η's are continuous) terms $\eta_{0j[i]} + \eta_{1j[i]} \cdot VISIT_i$ essentially "jitter" the fitted values so they no longer show the horizonal grouping structure of the marginal residuals.

- What might be causing the trend you see in this plot to be different from the trend in the plot of the marginal residuals.

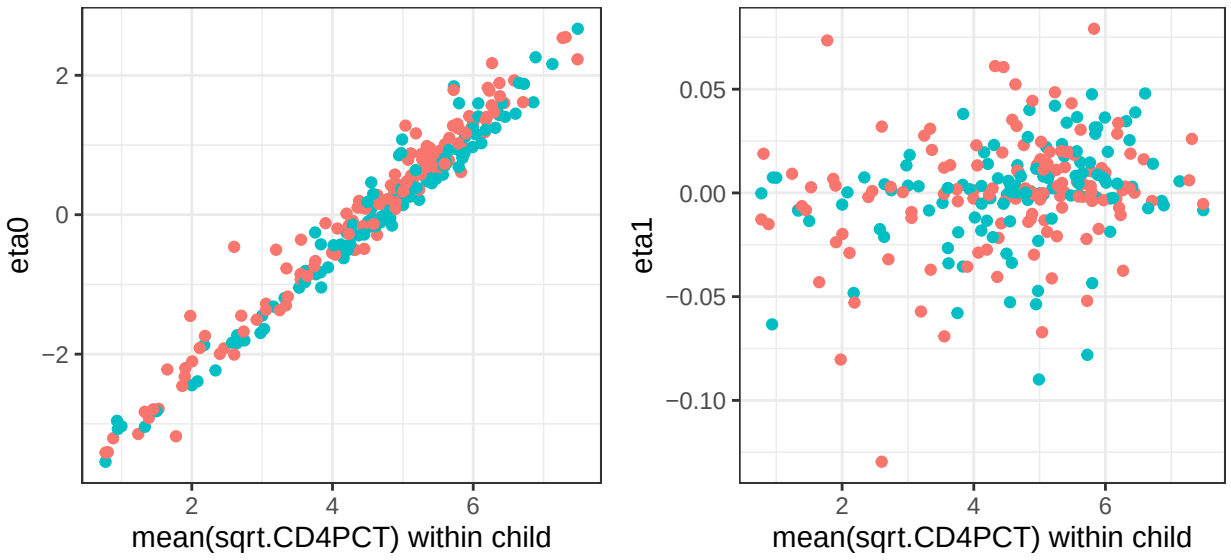
  The trend shows that when the conditional fitted value $\hat{y}_{cond}$ is larger, it tends to underpredict y = sqrt.CD4PCT (positive residual), and when $\hat{y}_{cond}$ is smaller, it tends to overpredict y (negative residual).

  The only difference between the two plots is the presence of the $\eta_{0j[i]} + \eta_{1j[i]} \cdot VISIT_i$ terms.

  In ordinary regression,we know that residuals and fitted values will be uncorrelated: that is why we don't see overall increasing or overall decreasing trends in the marginal residual plots (as far as the math is concerned, marginal residuals and marginal fitted values behave exactly like ordinary regression). The fact that conditional residuals and conditional fitted values are correlated is due to the values we use for the η's: $\hat{\eta} = E[\eta|\text{the data}]$, and "the data" includes y. Since, therefore, the conditional fitted values depend on y, it's not surprising that the residuals and fitted values are correlated.

  The plots below explore the association between the $\hat{\eta}$'s and y. We can see that, although there's not much association between $\hat{\eta}_{1j}$ and y, there is quite a strong association between $\hat{\eta}_{0j}$ and y.

```
y.grouped <- with(cd4,sapply(split(sqrt.CD4PCT,newpid),mean))
tx.grouped <- with(cd4,sapply(split(treatmnt,newpid),function(x) x[1]))

g1 <- ggplot(r.2,aes(x=y.grouped,y=.ranef.intercept)) +
  geom_point(aes(color=tx.grouped)) +
  xlab("mean(sqrt.CD4PCT) within child") +
  ylab("eta0") +
  theme(legend.position="none")

g2 <- ggplot(r.2,aes(x=y.grouped,y=.ranef.visit)) +
  geom_point(aes(color=tx.grouped)) +
  xlab("mean(sqrt.CD4PCT) within child") +
  ylab("eta1") +
  theme(legend.position="none")

grid.arrange(g1,g2,ncol=2)
```

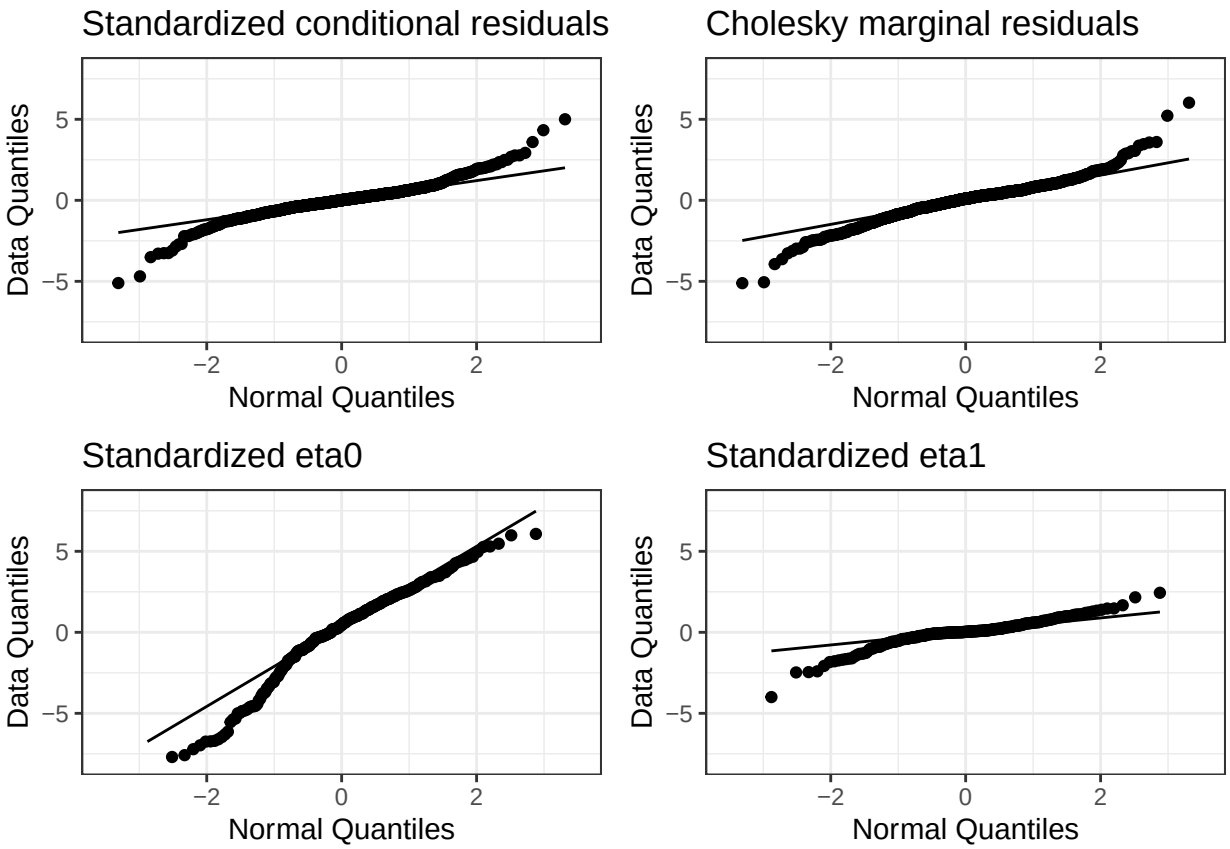


Problem 1(d)

Use standardized residuals and standardized random effects estimates to assess the normality of ϵ_i, η_{0j} and η_{1j} in the fitted model, and to check for any outliers. Include qq plots for each, and accompany each plot with a sentence or two describing what is good or bad in that plot.

```
g1 <- ggplot(r.1s,aes(sample=.std.resid)) +
  geom_qq() +
  geom_qq_line() +
  xlim(-3.5,3.5) +
  ylim(-8,8) +
  xlab("Normal Quantiles") +
  ylab("Data Quantiles") +
  ggtitle("Standardized conditional residuals")

g2 <- ggplot(r.1s,aes(sample=.chol.mar.resid)) +
  geom_qq() +
  geom_qq_line() +
  xlim(-3.5,3.5) +
  ylim(-8,8) +
  xlab("Normal Quantiles") +
  ylab("Data Quantiles") +
  ggtitle("Cholesky marginal residuals")

g3 <- ggplot(r.2s,aes(sample=.std.ranef.intercept)) +
  geom_qq() +
  geom_qq_line() +
  xlim(-3.5,3.5) +
  ylim(-8,8) +
  xlab("Normal Quantiles") +
  ylab("Data Quantiles") +
  ggtitle("Standardized eta0")

g4 <- ggplot(r.2s,aes(sample=.std.ranef.visit)) +
  geom_qq() +
  geom_qq_line() +
  xlim(-3.5,3.5) +
  ylim(-8,8) +
  xlab("Normal Quantiles") +
  ylab("Data Quantiles") +
  ggtitle("Standardized eta1")

grid.arrange(g1,g2,g3,g4,ncol=2)
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



The "middle" part of all four plots seem to follow the normal distribution fairly well, but the tails fail in various ways:

- *The standardized conditional and marginal residuals both have longer tails than the Normal distribution. There seem to be two low outliers and two high outliers among these residuals.*
- *The Cholesky marginal residual seem a little more extreme that the standardized conditional residuals, which make sense since $X\beta + Z\eta$ should do a better job predicting y than $X\beta$ alone. There seem to be two low outliers and two high outliers among these residuals. It would be useful to know if these are the same data points as in the first QQ plot.*
- *The standardized $\hat{\eta}_0$'s have a long left tail and slighltly short right tail. In fact the left tail for η_0 is the most extreme of all four plots. There don't appear to be any clear outliers among the $\hat{\eta}_0$'s.*
- *The distribution of $\hat{\eta}_1$ is closest to Normal, with a right tail that is only a little longer than the Normal, and a left tail that is the least extreme of all four plots. There may be one low outlier among the $\hat{\eta}_1$'s.*

Problem 2.

Continuing with the CD4 data...

Problem 2(a)

Make a table giving values of AIC, BIC, DIC, and cAIC (you compared two of these on the last assignment,using just AIC, BIC and DIC):

```
lmer.1 <- lmer(sqrt.CD4PCT ~ 1 + visage +           (1+visage|newpid),data=cd4,REML=F)
lmer.2 <- lmer(sqrt.CD4PCT ~ 1 + visage + treatmnt + (1+visage|newpid),data=cd4,REML=F)
lmer.3 <- lmer(sqrt.CD4PCT ~ 1 + visage * treatmnt + (1+visage|newpid),data=cd4,REML=F)
```

```
lmer.4 <- lmer(sqrt.CD4PCT ~ 1 + VISIT  +           (1+VISIT|newpid),data=cd4,REML=F)
lmer.5 <- lmer(sqrt.CD4PCT ~ 1 + VISIT  + treatmnt + (1+VISIT|newpid),data=cd4,REML=F)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00584617 (tol = 0.002, component 1)
```

```
lmer.6 <- lmer(sqrt.CD4PCT ~ 1 + VISIT  * treatmnt + (1+VISIT|newpid),data=cd4,REML=F)

DIC <- function(M) {
         if(class(M)=="lm") { x <- AIC(M) }
         else { x <- unname(extractDIC(M)) }
         return(x)
       }

CAIC <- function(M) {cAIC(M)$caic}

IC <- function(M) {
    x <-c(AIC=AIC(M),BIC=BIC(M),DIC=DIC(M),cAIC=CAIC(M))
    return(x)
}

IC.table <- function(...) {
    Mlist <- list(...)
    x <- suppressWarnings(lapply(Mlist,IC))
    x <- data.frame(matrix(unlist(x),ncol=4,byrow=T))
    names(x) <- c("AIC","BIC","DIC","cAIC")
    models <- sapply(Mlist,formula)
    models <- abbreviate(substr(models,14,200),30)
    rownames(x) <- models
    return(x)
}

IC.table(lmer.1,lmer.2,lmer.3,lmer.4,lmer.5,lmer.6)
```

```
##                                      AIC      BIC      DIC     cAIC
## 1+visage+(1+visage|newpid)     3142.447 3172.327 3130.447 2697.868
## 1+visag+treatmnt+(1+visg|nwpd) 3142.755 3177.616 3128.755 2698.827
## 1+visag*treatmnt+(1+visg|nwpd) 3144.755 3184.595 3128.755 2699.562
## 1+VISIT+(1+VISIT|newpid)       3126.563 3156.443 3114.563 2642.798
## 1+VISIT+tretmnt+(1+VISIT|nwpd) 3127.625 3162.486 3113.625 2643.307
## 1+VISIT*tretmnt+(1+VISIT|nwpd) 3129.215 3169.056 3113.215 2644.733
```

Comment briefly on any similarities or differences in how the different criteria choose fixed effects.

Here's an overall sum of the best (minimizing the criterion) models for each criterion, with second and third place winners as well:

AIC: *Best model is lmer.4, with lmer.5 a close second.*

BIC: *Best model is lmer.4, with lmer.5 a substantially less close second.*

DIC: *Best model is lmer.6, with lmer.4 and lmer.5 almost indistinguishably close.*

cAIC: *Best model is lmer.4, with lmer.5 and lmer.5 almost indistinguishably close.*

All the criteria prefer having `VISIT` *as a predictor rather than* `visage`*. (This kind of surprises me because visage has more physical meaning, but so be it...) Consistent with our findings on earlier hw that treatmnt did not seem like an important predictor,* `lmer.4` *(without* `treatmnt` *in the model) is first or second with all*

the criteria, with `lmer.5` (main effect for `treatmt`, but no interaction) a strong second or third with all the criteria.

Problem 2(b)

Make a table giving values of AIC, BIC, DIC, and cAIC for the following models:

```
lm.7 <- lm(sqrt.CD4PCT ~ 1 + VISIT + treatmnt,data=cd4)
lmer.8 <- lmer(sqrt.CD4PCT ~ 1 + VISIT + treatmnt + (1|newpid),data=cd4,REML=F)
lmer.9 <- lmer(sqrt.CD4PCT ~ 1 + VISIT + treatmnt + (0+VISIT|newpid),data=cd4,REML=F)
lmer.10 <- lmer(sqrt.CD4PCT ~ 1 + VISIT + treatmnt + (1+VISIT|newpid),data=cd4,REML=F)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00584617 (tol = 0.002, component 1)
```

```
IC.table(lm.7,lmer.8,lmer.9,lmer.10)
```

```
##                                      AIC      BIC      DIC     cAIC
## 1 + VISIT + treatmnt            4009.810 4029.730 4009.810 4009.810
## 1+VISIT+treatmnt+(1|newpid)     3153.879 3178.779 3143.879 2728.147
## 1+VISIT+tretmnt+(0+VISIT|nwpd) 3688.619 3713.520 3678.619 3490.226
## 1+VISIT+tretmnt+(1+VISIT|nwpd) 3127.625 3162.486 3113.625 2643.307
```

You'll fit the first model with lm(), and the others with lmer(). Comment briefly on any similarities or differences in how the different criteria choose random effects. (Note that only BIC and cAIC have a strong theoretical justification here).

There's a convergence warning for one of the models, which you can (and should) pursue using the methods in the R notes from lecture.

All four information criteria strongly favor at least a random intercept (`lm.7` and `lmer.9` are very strongly disfavored). Among the remaining two models, the model with random slope and random intercept is strongly favored by all the criteria.

Problem 2(c)

Repeat part (b) but with the interaction `VISIT * treatmnt` in each model instead of just the main effects `VISIT + treatmnt`.

```
lm.11 <- lm(sqrt.CD4PCT ~ 1 + VISIT * treatmnt,data=cd4)
lmer.12 <- lmer(sqrt.CD4PCT ~ 1 + VISIT * treatmnt + (1|newpid),data=cd4,REML=F)
lmer.13 <- lmer(sqrt.CD4PCT ~ 1 + VISIT * treatmnt + (0+VISIT|newpid),data=cd4,REML=F)
lmer.14 <- lmer(sqrt.CD4PCT ~ 1 + VISIT * treatmnt + (1+VISIT|newpid),data=cd4,REML=F)

IC.table(lm.11,lmer.12,lmer.13,lmer.14)
```

```
##                                      AIC      BIC      DIC     cAIC
## 1 + VISIT * treatmnt            4010.343 4035.244 4010.343 4010.343
## 1+VISIT*treatmnt+(1|newpid)     3155.181 3185.062 3143.181 2729.893
## 1+VISIT*tretmnt+(0+VISIT|nwpd) 3690.150 3720.030 3678.150 3491.763
## 1+VISIT*tretmnt+(1+VISIT|nwpd) 3129.215 3169.056 3113.215 2644.733
```

The story here is essentially the same as in part (b): The model with random slope and random intercept is strongly favored by all four criteria.

Not part of the required answer here, but, comparing parts (b) and (c) it seems like the cross-level interaction is not needed.