Exam 4

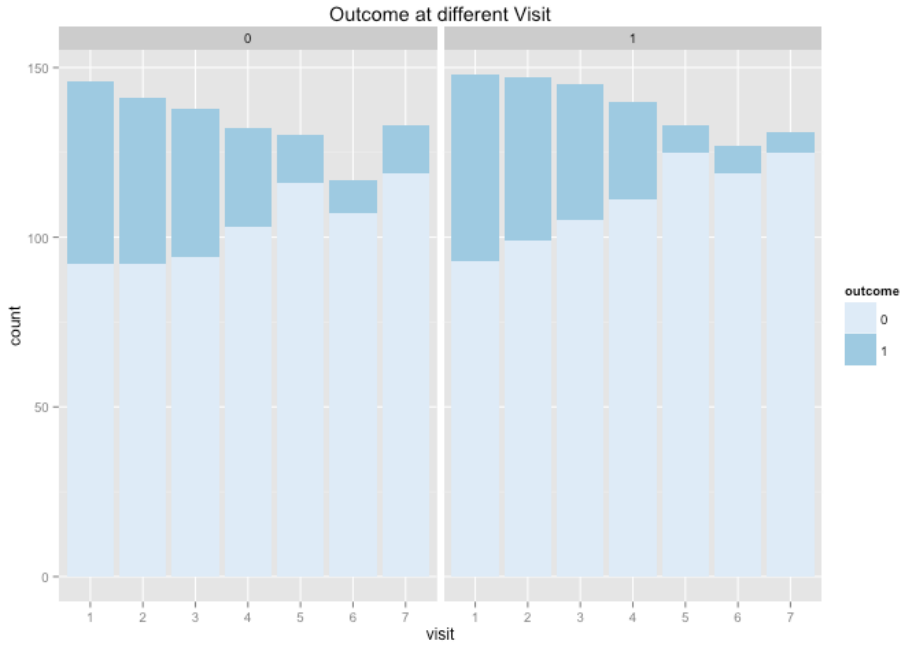
STAT 426

Yilin Zhao

NetID: zhao100

1. Introduction

In this project we'll analyze the *toenail* dataset in Faraway package. The data come from a Multicenter study comparing two oral treatments for toenail infection. Here we are interested in the degree of onycholysis which express the degree of separation of the nail plate from the nail-bed. Patients were randomized into two treatments and were evaluated at 7 visits, i.e on week 0, 4, 8, 12, 24,36, and 48. The secondary end point was scored in 4 levels and was evaluated on 294 patients comprising 1908 measurements.



Plot 1. Data Description Plot

We will explore the effect of treatment on the outcome of the toenail infection while regarding the time of the test as well as treatment*month. In this model fomulation, the cover treatment represents the effect of treatment at the base-line. We will perform this analysis using following methods: Generalized linear mixed models, Generalized estimating equations, and Transition models. And we will compare the results of each model.

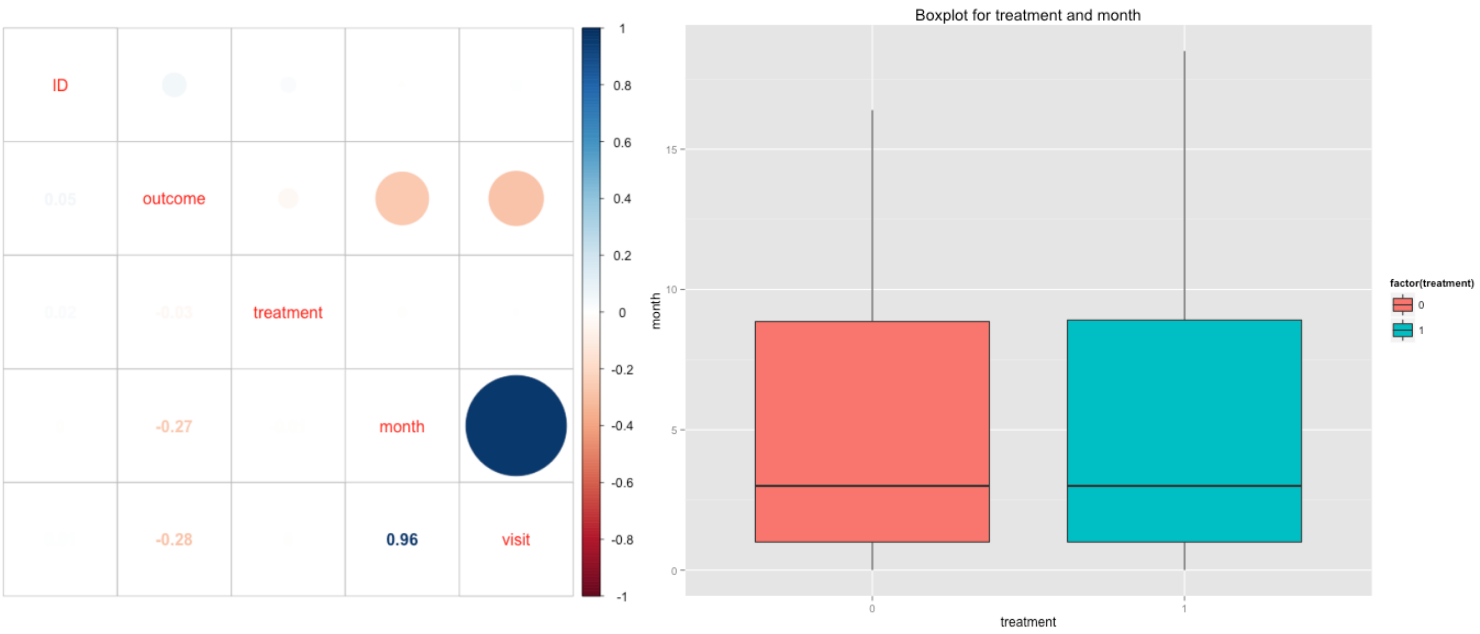
2. Data Description

There are 5 variables in toenail dataset. I transformed some of them into factors. Here is the description table.

Varible	Type	Description
ID	Integer	ID of patient
outcome	Factor	Outcome of the toenail infection. 0=none of mild seperation, 1=moderate or severe
treatment	Factor	The treatment A or B
month	Numeric	Time of the visit (not exactly monthly intervals hence not round numbers)
visit	Factor	The number of the visit

Table.1 Data Description Table

Next I'll check the correlation between each pair of predictors. In this part I just treat the factor variables as integer and perform Pearson chi-square. Here is the ellipse correlation plot and box plot for treatment and month.



Plot 2. Relationship between each pair of predictors

We can see that *visit* and *month* have strong positive correlation. So we'll rule out *visit* when build models and use *treatment, month and ID*(random effect) as predictors.

3. Analysis

- **Generalized Linear Mixed Model**

Package *lme4* in R can help us build GLMM model. First I build model with both main effects and their interaction. After that I build a model only use the main effects. After comprising these two models I find the model with interaction is better.

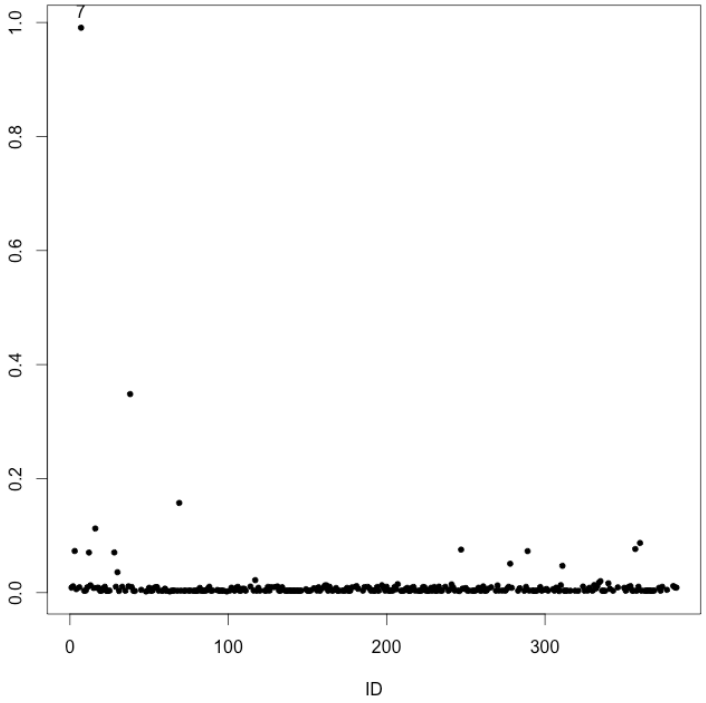
```
> anova(mod.1,mod.1.temp)
Data:
Models:
mod.1.temp: outcome ~ treatment + month + (1 | ID)
mod.1: outcome ~ treatment * month + (1 | ID)
           Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
mod.1.temp  4 1268 1290   -630     1260
mod.1       5 1266 1293   -628     1256  4.02      1      0.045 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

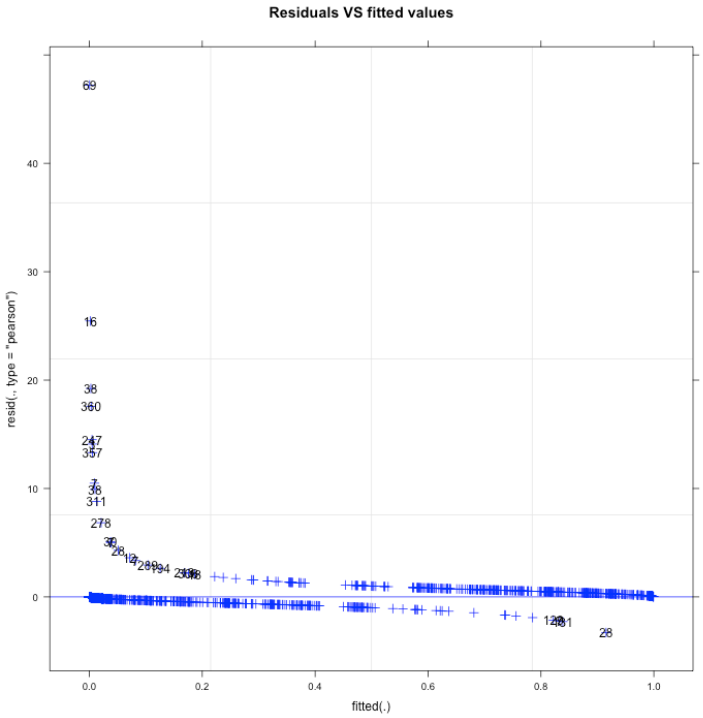
Plot 3. ANOVA for two models

mod.1.temp has larger AIC and deviance. So my final GLMM model is

$$outcome \sim treatment * month + (1 | ID)$$

Now lets check the influential points and outliers. Function influence() can help us calculate the cook distance. We can also take a look at the residuals plot and find outliers.





From the plots above, we can notice that there is no ID group with cooks.distance greater than 1. So there is no influential points. Regarding outliers, we may notice that there are several groups with absolute value of residuals greater than 2. These groups can be treated as outliers that need to remove out.

After removing the outliers, we got a new dataset I called toenail.fix. And I rebuild the GLMM model with this new dataset to get a final result. Here is the summary of the final GLMM model in R:

	Estimate	LL	UL	P.value
(Intercept)	-9.2763	-10.9757	-7.5769	0.0000
treatment1	0.0589	-1.9730	2.0908	0.9547
month	-2.0532	-2.5991	-1.5073	0.0000
treatment1:month	-0.8491	-1.5557	-0.1426	0.0185

Table 2. Summary of GLMM

Interpretation: We can notice that the estimation of *month* and the interaction are significant. The estimate of *month* is -2.05, which means in overall when there is one unit increase in month, the odds ratio of outcome will be exp(-2.05)=0.129. So we'll see 88% decrease in odds ratio. As time goes by, the severity of toenail infection will be less. The interaction has an estimate of -0.850, which means compare to *treatment*=0, when there is one unit increase in month, the odds ratio of outcome will decrease by 1-0.427=60%. So when *treatment*=1 *month* will have more obvious positive effect on alleviating severity of toenail infection compare to treatment=0.

	data_oucome.1	data_outcome.0
mod_outcome=1	362	10
mod_outcome=0	10	1396

Table 3. Predict Values VS Observed Values

Sensitivity=99.29%

Specificity=97.31%

97.31% of those whose observed outcome=1 will be correctly classified in the final model. 99.29% of those whose observed outcome=0 will be correctly classified. The result is satisfied.

- **Generalized Estimating Equations**

Package *gee* in R can help us build GEE model. First I build model with both main effects and their interaction. Now we need to determine an appropriate covariance structure.

QIC is a good way to select the best covariance structure for quasi-likelihood based model. Package *MuMIn* can help us achieve QIC selection. Here is the result in R:

```
> model.sel(mod.2,mod.2.1,mod.2.2, rank = QIC)
Model selection table
        (Int)  mnt    trt mnt:trt corstr qLik QIC  delta weight
mod.2   -0.699 -0.141 +   +       unstrc -909 1836 0.00  0.590
mod.2.2 -0.557 -0.170 +   +       indpnd -908 1838 1.68  0.255
mod.2.1 -0.582 -0.171 +   +       exchng -908 1839 2.66  0.156
Abbreviations:
corstr: exchng = 'exchangeable', indpnd = 'independence',
        unstrc = 'unstructured'
Models ranked by QIC(x)
```

So we will obtain our GEE model which works on unstructured correlation matrix. My GEE model is: outcome~treatment*month. Here is the analysis result of GEE.

```
Call:
gee(formula = outcome ~ treatment * month, id = ID, data = toenail,
    family = binomial, corstr = "unstructured", scale.fix = TRUE)

Summary of Residuals:
    Min      1Q  Median      3Q     Max
-0.3404 -0.2524 -0.1222 -0.0333  0.9744


Coefficients:
                Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)      -0.6993      0.167  -4.182       0.167   -4.187
treatment1        0.0376      0.237   0.159       0.244    0.154
month            -0.1414      0.026  -5.426       0.027   -5.235
treatment1:month -0.0828      0.042  -1.971       0.048   -1.726

Estimated Scale Parameter:  1
Number of Iterations:  6

Working Correlation
      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]
[1,] 1.000 0.885 0.690 0.486 0.235 0.144 0.102
[2,] 0.885 1.000 0.799 0.577 0.258 0.218 0.123
[3,] 0.690 0.799 1.000 0.749 0.269 0.196 0.149
[4,] 0.486 0.577 0.749 1.000 0.348 0.257 0.194
[5,] 0.235 0.258 0.269 0.348 1.000 0.453 0.368
[6,] 0.144 0.218 0.196 0.257 0.453 1.000 0.548
[7,] 0.102 0.123 0.149 0.194 0.368 0.548 1.000
```

	Estimate	LL	UL	P.value(Robust Z)
(Intercept)	-0.6993	-1.0266	-0.3720	0.0000
treatment1	0.0376	-0.4403	0.5156	1.1226
month	-0.1414	-0.1943	-0.0884	0.0000
treatment1:month	-0.0828	-0.1769	0.0112	0.0843

Table 4. Summary of GEE(Robust Z)

Interpretation: We can notice that the estimation of *month* is significant. The estimate of *month* is -0.14, which means in overall when there is one unit increase in month, the odds ratio of outcome will decrease by exp(-0.14)=0.869. In other words, we will se 13% decrease in the odds ratio. As time goes by, the severity of toenail infection will be less. The estimates of treatment and its interaction are insignificant. The interaction has an estimate of -0.08, which means compare to *treatment*=0, when there is one unit increase in month, the odds ratio of outcome will decrease by 1-0.92=8%. So when *treatment*=1 *month* will have more obvious positive effect on alleviating severity of toenail infection compare to treatment=0.

	data_outcome.1	data_outcome.0
mod_outcome=1	55	353
mod_outcome=0	93	1407

Table 5. Predict Values VS Observed Values

Sensitivity=79.94%

Specificity=37.16%

37.16% of those whose observed outcome=1 will be correctly classified in the final model. It will miss around 40% of all observed outcome=1 cases. 79.94% of those whose observed outcome=0 will be correctly classified. This model will miss lots of correct results. It will miss around 20% of all observed outcome=0 cases.

• Transition Model

Now I am building transition model. First I build model with both main effects and their interaction. I use stepAIC help me to select the model. And the backward selection shows that the model without interaction is better.

So my final transition model is:

outcome ~ treatment + month + preoutcome

Here is the result of backward selection.

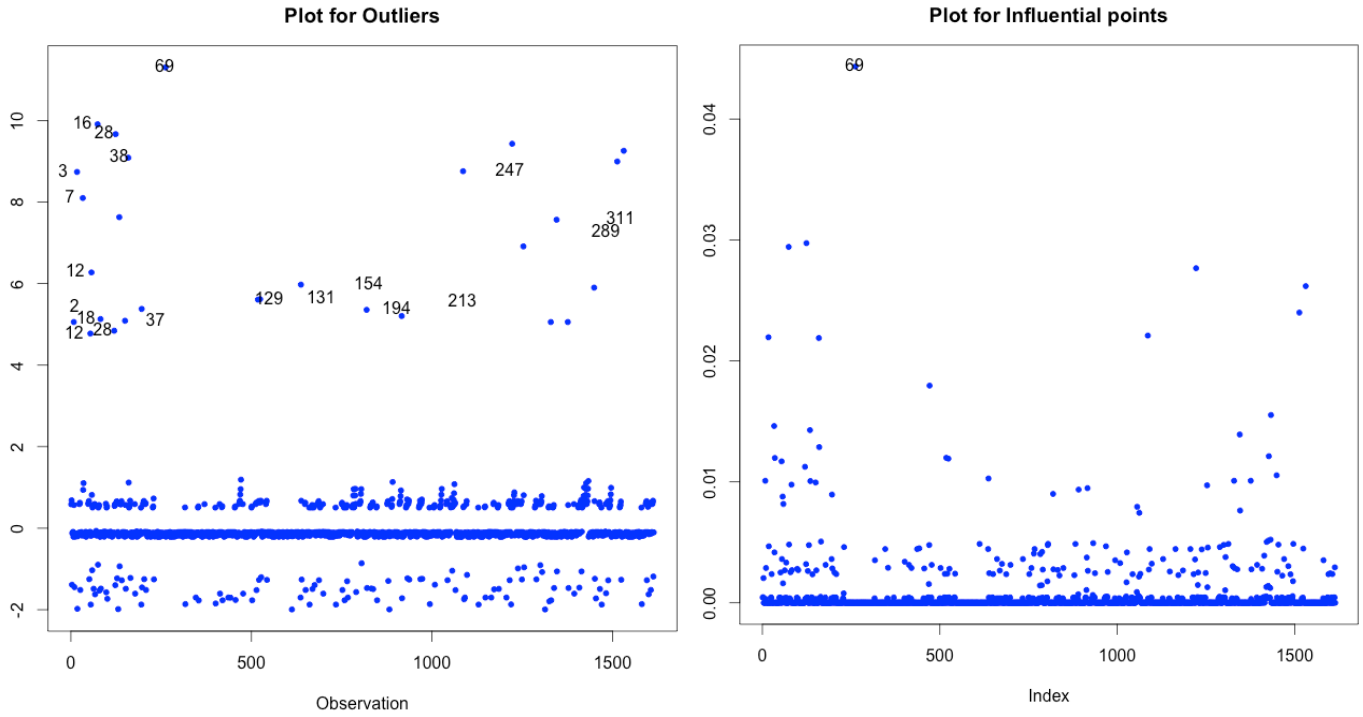
```
> mod.3.fix <- stepAIC(mod.3)
Start:  AIC=720
outcome ~ treatment * month + preoutcome

                   Df Deviance  AIC
- treatment:month  1      712  720
<none>                     710  720
- preoutcome       1     1428 1436

Step:  AIC=720
outcome ~ treatment + month + preoutcome

             Df Deviance  AIC
<none>              712  720
- treatment   1      714  720
- month       1      728  734
- preoutcome  1     1431 1437
```

Now let's do some diagnostic to check outliers and influential points. To check the outliers I calculate the Student Residuals. There are many points with Student Residuals greater than 2. Regarding influential points, cooks distance will work. There is no point with Cooks distance greater than 1. Here are my diagnostic plots.



Plot 4. Diagnostic for Transition Model

After removing the outliers, we got a new dataset. And I rebuild the transition model with this new dataset to get a final result. Here is the summary of the final transition model in R:

	Estimate	LL	UL	P.value
(Intercept)	-20.6094	-1629.0627	1587.8439	0.9800
treatment1	-0.1584	-0.6163	0.2994	0.4976
month	-0.1716	-0.2406	-0.1027	0.0000
preoutcome1	22.2430	-1586.2102	1630.6963	0.9784

Table 6. Summary of Transition Model

Interpretation: We can notice that only the estimate of *month* is significant. The estimate of *month* is -0.171, which means in overall when there is one unit increase in month, the odds ratio of outcome will be exp(-0.171)=0.843. So we'll see 16% decrease in odds ratio. As time goes by, the severity of toenail infection will be less. The estimate of *preoutcome* is 22.24, which means compare to have no infection before, *preoutcome=1* will highly increase the odds ratio.

	data_outcome.1	data_outcome.0
mod_outcome=1	250	21
mod_outcome=0	101	1214

Table 7. Predict Values VS Observed Values

Sensitivity=98.30%

Specificity=71.23%

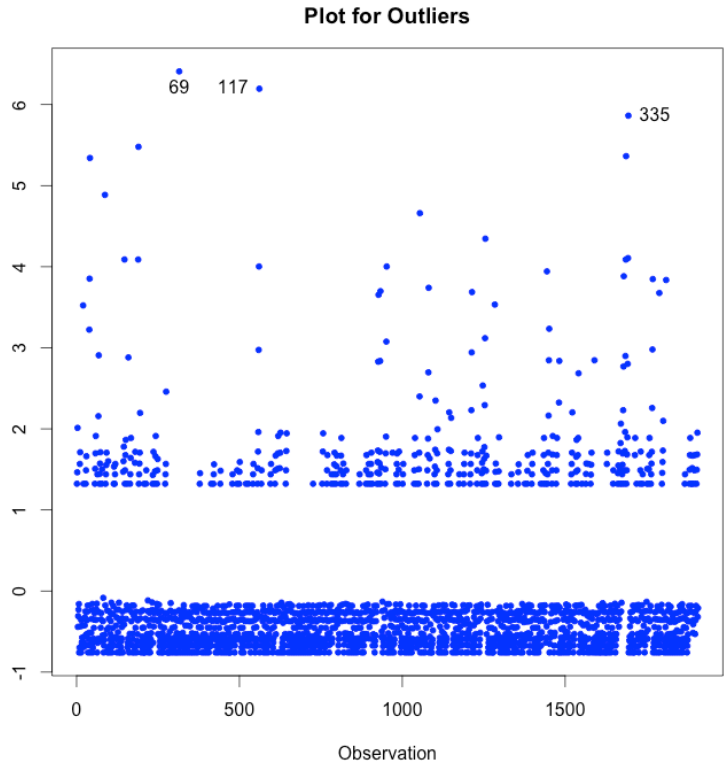
71.23% of those whose observed outcome=1 will be correctly classified in the final model. It will miss around 30% of all observed outcome=1 cases. 98.30% of those whose observed outcome=0 will be correctly classified.

• Naive Model

Now I am building GLM model with independent assumption. First I build model with both main effects and their interaction. Then I use stepAIC help me to select the model. And the backward selection shows that the model with interaction has lowest AIC. So my model in this part is:

outcome~treatment*month

After checking the Cooks Distance, I find there is no point with cooks distance greater than 1. But there are many points with Student Residuals greater than 2. Here is the Student Residuals plot.



Plot 5. Diagnostic for Naive Model

After removing the outliers, we got a new dataset. And I rebuild the Naive model with this new dataset to get a final result. Here is the summary of the final transition model in R:

```
Call:
glm(formula = outcome ~ treatment * month, family = binomial,
    data = toenail.fix)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1010  -0.7274  -0.2785  -0.0636   2.0514

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.1824     0.1243   -1.47     0.14
treatment1       -0.0755     0.1769   -0.43     0.67
month            -0.4632     0.0526   -8.81   <2e-16 ***
treatment1:month -0.0322     0.0773   -0.42     0.68
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1774.6  on 1843  degrees of freedom
Residual deviance: 1413.3  on 1840  degrees of freedom
AIC: 1421
```

Interpretation: We can notice that only the estimate of *month* is significant. The estimate of *month* is -0.46, which means in overall when there is one unit increase in month, the odds ratio of outcome will decrease by exp(-0.46)=0.63. In other words, we will se 37% decrease in the odds ratio. As time goes by, the severity of toenail infection will be less. The estimates of treatment and its interaction are insignificant. The interaction has an estimate of -0.03, which means compare to *treatment*=0, when there is one unit increase in month, the odds ratio of outcome will decrease by 1-0.97=3%. So when *treatment*=1 *month* will have more obvious positive effect on alleviating the severity of toenail infection compare to treatment=0.

	data_outcome.1	data_outcome.0
mod_outcome=1	0	344
mod_outcome=0	0	1500

Table 8. Predict Values VS Observed Values

Sensitivity=81.34%

Specificity=NA

It will miss all observed outcome=1 cases while 81.34% of those whose observed outcome=0 will be correctly classified.

- **Summary**

I already interpret the models in the previous section. Here I'll compare these four models and find out their similarities and differences.

Both GEE and GLMM will consider the random effect(here is group=ID). But GEE is based on quasi-likelihood while others are likelihood-based. Transition model depends on the previous values in the process. Take a look at the results of these models, we'll notice that *month* is always significant with the similar decrease effect on odds ratio while *treatment* is always insignificant.

The GLMM and Naive model may consider the interaction of month and treatment but in GEE and Transition model, the model selection procedure removes out the interaction.

Take a look at the specificity and sensitivity, we'll notice that GLMM is the most accurate model among these 4 models while Naive model is the most inaccurate. This is reasonable since Naive model is based on the independence assumption.

We can conclude that month has significant effect on alleviating the severity of toenail infection. The longer the time, the less severe the infection is. But treatment has no significant effect on toenail infection.

- Appendix

```
require(faraway)
require(corrplot)
require(lme4)
require(car)
require(MASS)
require(MESS)
require(gee)
require(ggplot2)
require(xlsx)
require(MuMIn)
setwd('~/Documents/stat426/exam4')
View(toenail)
str(toenail)
## Check Correlation
corr <- cor(data.matrix(toenail))
corrplot.mixed(corr)
ggplot(toenail,aes(factor(treatment), y =
month, fill=factor(treatment))) +
geom_boxplot()+ggtitle("Boxplot for
treatment and month")+labs(x='treatment')
## Transform some variables into factor
toenail[,c(2,3,5)] <- lapply(toenail[,c(2,3,5)],
as.factor)
str(toenail)
## GLMM model
toenail <- data.frame(toenail)
attach(toenail)
mod.1 <- glmer(outcome ~
treatment*month + (1|ID),family =
binomial)
summary(mod.1)
mod.1.temp <- glmer(outcome ~ treatment
+month + (1|ID),family = binomial)
summary(mod.1.temp)
anova(mod.1,mod.1.temp)
## outliers and influential points
inf <- influence(mod.1, group='ID')
coods.d <- cooks.distance(inf)
cooks.d <- data.frame(coods.d)
dimnames(cooks.d)
par(mar=c(5,3,3,3))
ggplot(aes(x=dimnames(cooks.d)
[[1]],y=coods.d))+geom_point()
plot(x=row.names(cooks.d),cooks.d[,
1],pch=20,main='Cooks Distance for each
ID group',xlab='ID',ylab='Cooks.distance')
identify(x=row.names(cooks.d),cooks.d[,
1],row.names(coods.d),tolerance=0.5)
plot(mod.
1,id=0.05,idLables=~.ID,pch=3,col='blue',ma
in='Residuals VS fitted values')
remove <-
toenail[which(abs(residuals(object = mod.
1, type='pearson'))>2),]$ID
toenail.fix <- toenail[which(!toenail$ID %in
% remove),]
mod.1.fix <- glmer(outcome ~
treatment*month + (1|
ID),data=toenail.fix,family = binomial)
summary(mod.1.fix)

## confidence interval
se <- summary(mod.1.fix)$coefficients[,2]
CI.1 <- cbind(Estimate = summary(mod.
1.fix)$coefficients[,1],
              LL = summary(mod.1.fix)
$coefficients[,1]-1.96*se,
              UL = summary(mod.1.fix)
$coefficients[,1]+1.96*se,
            P.value = summary(mod.1.fix)
$coefficients[,4])
write.xlsx(CI.1,file='CI_GLMM.xls')
## fitted.values
attach(toenail.fix)
fitted.values <- ifelse(predict(mod.
1.fix,type='response') > 0.5, 1, 0)
Table.1 <- matrix(0,2,2,dimnames =
list(c('mod_outcome=1',"mod_outcome=0
"),c('data_oucome=1','data_outcome=0')))
Table.1[1,1] <-
sum((fitted.values==1)*(outcome==1))
Table.1[1,2] <-
sum((fitted.values==0)*(outcome==1))
Table.1[2,1] <-
sum((fitted.values==1)*(outcome==0))
Table.1[2,2] <-
sum((fitted.values==0)*(outcome==0))
sensitivity=Table.1[1,1]/(Table.1[1,1]+Table.
1[2,1])
specificity=Table.1[2,2]/(Table.1[2,2]+Table.
1[1,2])
write.xlsx(Table.1,file='Table_GLMM.xls')

## GEE model
```

```
mod.2 <- gee(outcome ~
treatment*month,id = ID,

corstr="unstructured",data=toenail,scale.fix
=TRUE,family = binomial)
summary(mod.2)
mod.2.1 <- update(mod.2,
corstr='exchangeable')
mod.2.2 <- update(mod.2,
corstr='independence')
model.sel(mod.2,mod.2.1,mod.2.2, rank =
QIC)
QIC(mod.2)
## Confidence interval for GEE
se <- summary(mod.2)$coefficients[,4]
CI.2 <- cbind(Estimate = summary(mod.
2)$coefficients[,1],
          LL = summary(mod.
2)$coefficients[,1]-1.96*se,
          UL = summary(mod.
2)$coefficients[,1]+1.96*se,
          P.value =
round(2*pnorm(summary(mod.
2)$coefficients[,5],0,1),6))
write.xlsx(CI.2,file='CI_GEE.xls')


attach(toenail)
fitted.values <- ifelse(exp(predict(mod.2))
> 0.5, 1, 0)
Table.2 <- matrix(0,2,2,dimnames =
list(c('mod_outcome=1',"mod_outcome=0
"),c('data_outcome=1','data_outcome=0')))
Table.2[1,1] <-
sum((fitted.values==1)*(outcome==1))
Table.2[1,2] <-
sum((fitted.values==0)*(outcome==1))
Table.2[2,1] <-
sum((fitted.values==1)*(outcome==0))
Table.2[2,2] <-
sum((fitted.values==0)*(outcome==0))
sensitivity=Table.2[1,1]/(Table.2[1,1]+Table.
2[2,1])
specificity=Table.2[2,2]/(Table.2[2,2]+Table.
2[1,2])
write.xlsx(Table.2,file='Table_GEE.xls')

## Transition Model
IDfreq<-table(toenail[,1])
IDlength<-length(IDfreq)
newtoenail<-NULL
for(i in 1:IDlength)
{
  if(IDfreq[i]>1)
  {
    current_upper<-sum(IDfreq[1:i])
    if(i==1)
    {
      current_lower<-1
    }else {
      current_lower<-sum(IDfreq[1:(i-1)])+1
    }
    temp1<-toenail[(current_lower
+1):current_upper,]
    temp2<-toenail[current_lower:
(current_upper-1),2]
    temp<-cbind(temp1,temp2)
    newtoenail<-rbind(newtoenail,temp)
  }
}
colnames(newtoenail)<-
c(colnames(toenail),'preoutcome')

## build transition Model
mod.3 <- glm(outcome ~ treatment*month
+ preoutcome,family =
binomial,data=newtoenail)
summary(mod.3)
mod.3.fix <- stepAIC(mod.3)
summary(mod.3.fix)
## outliers and influential points
par(mar=c(5,3,3,3))
plot(mod.3.fix)
Presiduals = residuals(object = mod.3.fix,
type='pearson')
h <- lm.influence(model=mod.3.fix)$h
Sresiduals <- Presiduals/sqrt(1-h)
plot(1:1614,pch=20,Sresiduals,
xlab="Observation",col='blue',
ylab="Standardized residuals",main='Plot
for Outliers')
identify(row.names(newtoenail),Sresiduals,
newtoenail$ID,tolerance=0.5)
remove <- which(abs(Sresiduals)>2)
outlierTest(mod.3.fix)
plot(cooks.distance(mod.
3.fix),pch=20,col='blue',ylab='Cooks.Distan
ce',main='Plot for Influential points')
identify(row.names(newtoenail),cooks.dist
ance(mod.3),newtoenail$ID,tolerance=0.5)
newtoenail.fix <- newtoenail[-remove,]

## rebuild the model
```

```
mod.3.fix <- glm(outcome ~ treatment +
month + preoutcome,family =
binomial,data=newtoenail.fix)
summary(mod.3.fix)

## confidence interval
se <- summary(mod.3.fix)$coefficients[,2]
CI.3 <- cbind(Estimate = summary(mod.
3.fix)$coefficients[,1],
        LL = summary(mod.3.fix)
$coefficients[,1]-1.96*se,
        UL = summary(mod.3.fix)
$coefficients[,1]+1.96*se,
        P.value = summary(mod.3.fix)
$coefficients[,4])
write.xlsx(CI.3,file='CI_TRN.xls')

## Fitted Values
fitted.values <- ifelse(predict(mod.
3.fix,type='response') > 0.5, 1, 0)
dim(newtoenail.fix)
length(fitted.values)
attach(newtoenail.fix)
Table.3 <- matrix(0,2,2,dimnames =
list(c('mod_outcome=1',"mod_outcome=0
"),c('data_outcome=1','data_outcome=0')))
Table.3[1,1] <-
sum((fitted.values==1)*(outcome==1))
Table.3[1,2] <-
sum((fitted.values==0)*(outcome==1))
Table.3[2,1] <-
sum((fitted.values==1)*(outcome==0))
Table.3[2,2] <-
sum((fitted.values==0)*(outcome==0))
sensitivity=Table.3[1,1]/(Table.3[1,1]+Table.
3[2,1])
specificity=Table.3[2,2]/(Table.3[2,2]+Table.
3[1,2])
write.xlsx(Table.3,file='Table_TRN.xls')

## naive model
mod.4 <- glm(outcome ~
treatment*month,data=toenail,family =
binomial)
summary(mod.4)

mod.4 <- stepAIC(mod.4)
plot(mod.4)
Presiduals = residuals(object = mod.4,
type='pearson')
h <- lm.influence(model=mod.4)$h
Sresiduals <- Presiduals/sqrt(1-h)
plot(1:1908,pch=20,Sresiduals,
xlab="Observation",col='blue',
ylab="Standardized residuals",main='Plot
for Outliers')
identify(row.names(toenail),Sresiduals,toen
ail$ID,tolerance=0.5)
outlierTest(mod.4)
remove <- which(abs(Sresiduals)>2)
toenail.fix <- toenail[-remove,]


mod.4.fix <- glm(outcome ~
treatment*month,family =
binomial,data=toenail.fix)
summary(mod.4.fix)
fitted.values <- ifelse(mod.4.fix
$fitted.values > 0.5, 1, 0)
dim(toenail.fix)
length(fitted.values)
attach(toenail.fix)
Table.3 <- matrix(0,2,2,dimnames =
list(c('mod_outcome=1',"mod_outcome=0
"),c('data_outcome=1','data_outcome=0')))
Table.3[1,1] <-
sum((fitted.values==1)*(outcome==1))
Table.3[1,2] <-
sum((fitted.values==0)*(outcome==1))
Table.3[2,1] <-
sum((fitted.values==1)*(outcome==0))
Table.3[2,2] <-
sum((fitted.values==0)*(outcome==0))
sensitivity=Table.3[1,1]/(Table.3[1,1]+Table.
3[2,1])
specificity=Table.3[2,2]/(Table.3[2,2]+Table.
3[1,2])

ggplot(toenail,aes(visit,fill=outcome))
+geom_bar()+facet_wrap(~treatment)
+scale_fill_brewer()+
  ggtitle('Outcome at different Visit')
```