BIS 628 HW 4

Joanna Chen

3/25/2020

Q1-Q2: Background for the data set: In a randomized, double-blind, parallel-group, multicenter study comparing two oral anti-fungal treatments (200 mg/day Itraconazole and 250 mg/day Terbinafine) for toenail infection (De Backer et al., 1998; also see Lesaffre and Spiessons, 2001), patients were evaluated for the degree of onycholysis (the degree of separation of the nail plate from the nail-bed) at baseline (week 0) and at weeks 4,8,12,24,36 and 48 thereafter. The onycholysis outcome variable is binary (none or mild versus moderate or severe). The binary outcome was evaluated on 294 patients comprising a total of 1908 measurements. The main objective of the analyses is to compare the effects of the two oral anti-fungal treatments (Itraconazole and Terbinafine) on changes in the probability of the binary onycholysis outcome over the duration of the study.
Dataset: toenail-data.csv or toenail-data.xlsx
The dataset is already in a long format, i.e. with repeated observations within a person. The variables are (in column order):
ID: participant ID number
Y (binary outcome): 0=none/mild onycholysis, 1=moderate/severe onycholysis
Treatment: 0=Itraconazole, 1=Terbinafine
Month: exact month of evaluation
Visit: visit sequence number (1=0 weeks (baseline), 2=4 weeks, 3=8 weeks, 4=12 weeks, 5=24 weeks, 6=36 weeks, 7=48 weeks)

Question 1 (20 points):

(a) (5 points) Read the data file into SAS or R (depending upon your preference). The data already have a header.

```
library(readr)
library(data.table)
library(ggplot2)
library(lme4)
```

```
## Loading required package: Matrix
```

```
dt <- read_csv("toenail-data.csv")
```

```
## Parsed with column specification:
## cols(
##   ID = col_double(),
##   Y = col_double(),
##   Treatment = col_double(),
##   Month = col_double(),
##   Visit = col_double()
## )
```

```
setDT(dt)
```

(b) (5 points) Create a table: obtain percent missing outcome across weeks of follow-up by intervention group and put your results in a table format. Summarize your observations in a few sentences.

```
dt$Weeks = dt[,ifelse(Visit==1,0,ifelse(Visit==2,4,ifelse(Visit==3,8,ifelse(Visit==4,12,ifelse(Visit==5
table1 = dt[,.N,by = c("Weeks","Treatment")]
dt[,uniqueN(ID),by = Treatment] #148 when trt = 1, 146 when trt = 0
```

```
##    Treatment  V1
## 1:         1 148
## 2:         0 146
```

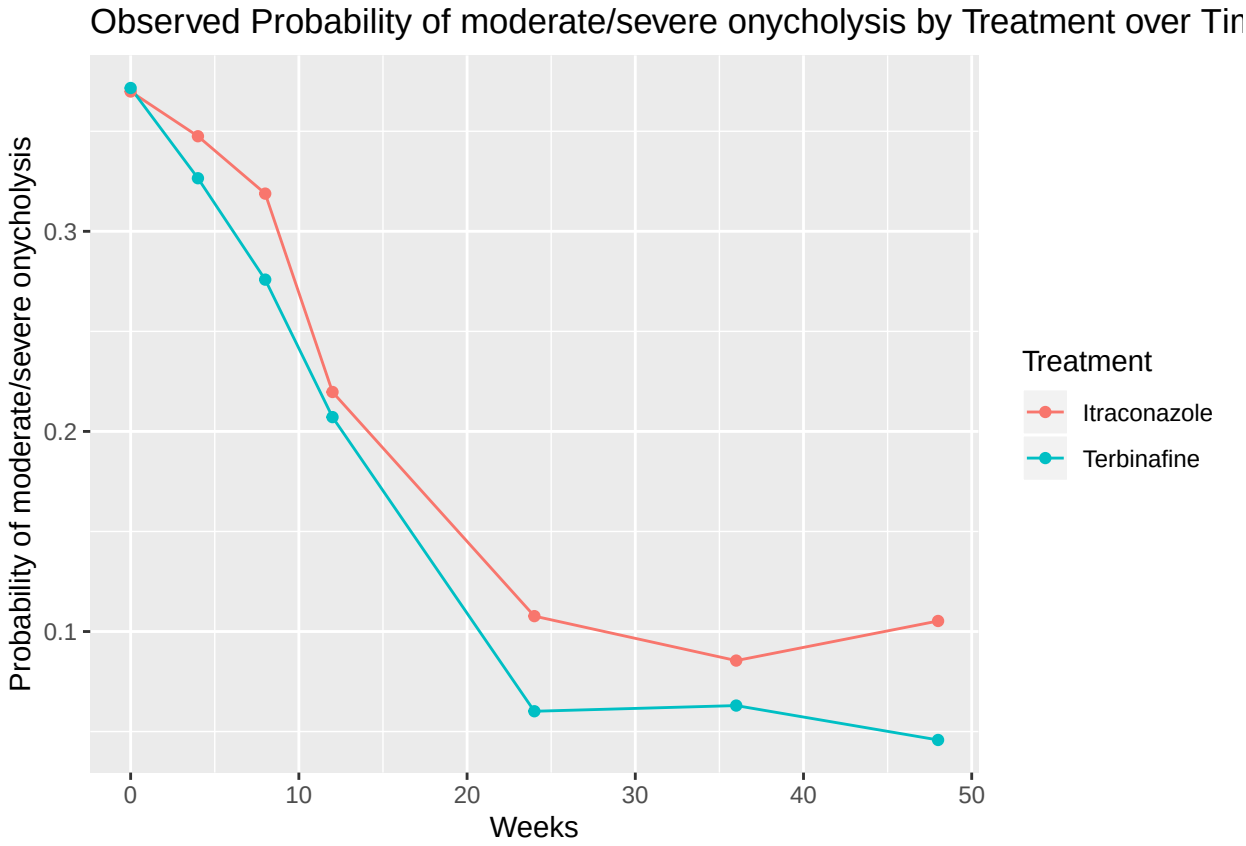
```
table1$total = table1[,ifelse(Treatment==1,148,146)]
table1$missing = 1- table1$N/table1$total

table1
```

```
##     Weeks Treatment   N total     missing
##  1:     0         1 148   148 0.000000000
##  2:     4         1 147   148 0.006756757
##  3:     8         1 145   148 0.020270270
##  4:    12         1 140   148 0.054054054
##  5:    24         1 133   148 0.101351351
##  6:    36         1 127   148 0.141891892
##  7:    48         1 131   148 0.114864865
##  8:     0         0 146   146 0.000000000
##  9:     4         0 141   146 0.034246575
## 10:     8         0 138   146 0.054794521
## 11:    12         0 132   146 0.095890411
## 12:    24         0 130   146 0.109589041
## 13:    36         0 117   146 0.198630137
## 14:    48         0 133   146 0.089041096
```

No one is missing during the first visit (week 0) for both treatment groups. Week 36 has the most missing rate acorss all weeks for both groups.

(c) (5 points) Figure 1: on a single graph, construct a time plot of the observed probability of moderate/severe onycholysis over weeks of follow up, stratified by the treatment group. Describe in a few sentences wheat you observe.

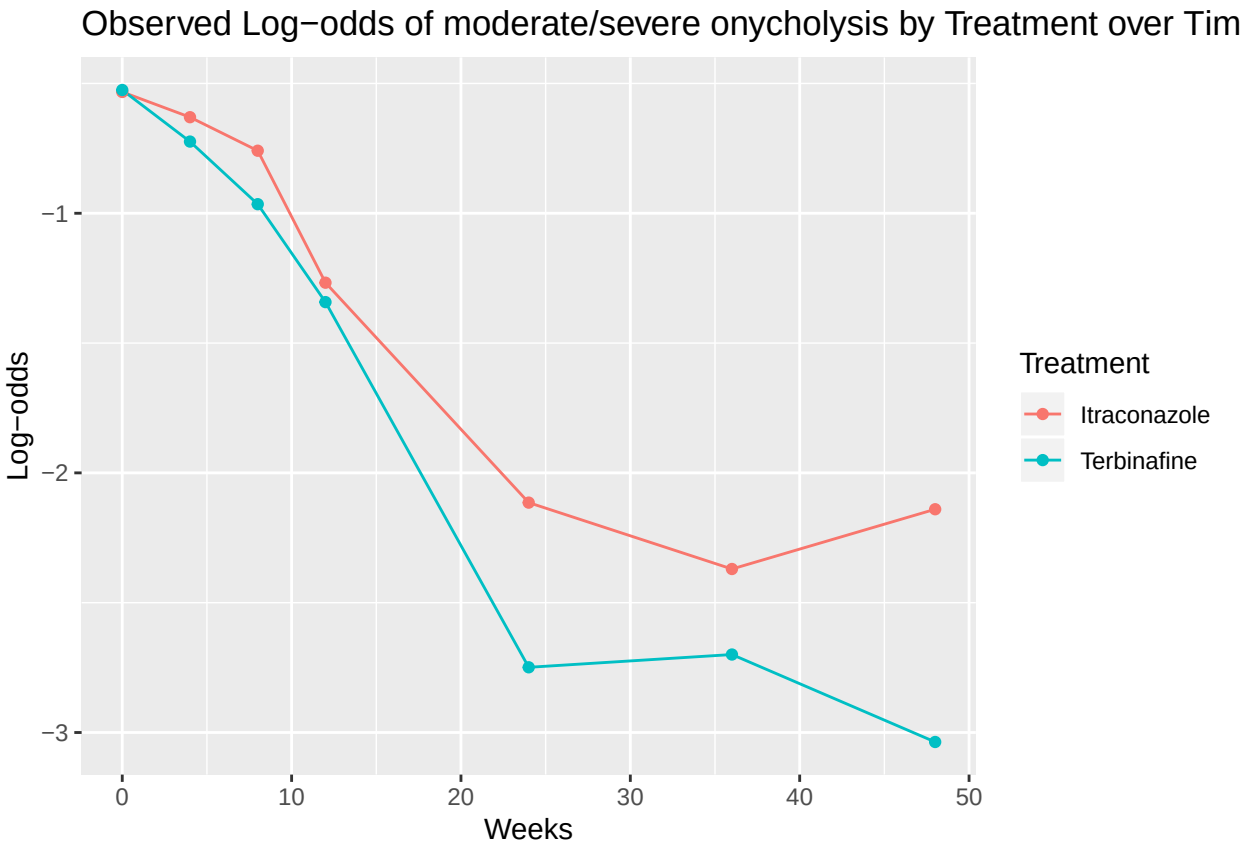
```
table2 <- aggregate(dt$Y, list(dt$Weeks,dt$Treatment), mean, na.rm=T)
names(table2) <- c("Weeks","Treatment","p")

#Create a factor variable for Treatment
table2$cTreatment <- factor(table2$Treatment, levels=c(0,1), labels=c("Itraconazole","Terbinafine"))

#Plot observed probability by treatment group over time
p1 <- ggplot(table2, aes(x=Weeks,y=p,group=cTreatment))
p1 + geom_line(aes(color=cTreatment)) + geom_point(aes(color=cTreatment)) +
  labs(title="Observed Probability of moderate/severe onycholysis by Treatment over Time",
       color="Treatment", y="Probability of moderate/severe onycholysis")
```



The observed probability of moderate/severe oncholysis is decreasing with over weeks of follow up for both treatment group before week 25. After that, the plot converge then diverge. The Terbinafine group probability is lower than the Itraconazole group.

(d) (5 points) Figure 2: on a single graph, construct a time plot of the observed log- odds of moderate/severe onycholysis over weeks of follow up, stratified by the treatment group. Describe in a few sentences what you observe.

```
table2$odds <- table2$p/(1-table2$p)
table2$logit <- log(table2$odds)

#Plot observed log-odds by treatment group over time
p2 <- ggplot(table2, aes(x=Weeks,y=logit,group=cTreatment))
p2 + geom_line(aes(color=cTreatment)) + geom_point(aes(color=cTreatment)) +
  labs(title="Observed Log-odds of moderate/severe onycholysis by Treatment over Time",
       color="Treatment", y="Log-odds")
```



The observed log-odds of moderate/severe oncholysis is decreasing over weeks of follow up for both treatment group before week 25. The plot looks similar as the previous one. The Terbinafine group log odds is lower than the Itraconazole group overall.

Question 2 (60 points):

Fit a generalized linear mixed effects model (GLMM) for the subject-specific probability of moderate/severe onycholysis with the following covariates: week of follow-up (continuous variable, starting with baseline week=0) and intervention by week of follow-up interaction, and a random intercept for each subject that is assumed to be generated from a Normal distribution with a mean zero and a variance parameter.

(a) (4 points) Write out the equation for this GLMM.

$$\text{logit}\left\{E\left(Y_{ij}|b_i\right)\right\} = \left(\beta_1 + b_i\right) + \beta_2 \text{ Time}_{ij} + \beta_3 \text{ Treatment}_i \times \text{ Time }_i$$

Note that here Time is the Weeks variable in our data.

(b) (2 points) What important assumption is being used for this GLMM?

Assume b_i, Y_{ij} have a Bernoulli distribution and

$$b_i \sim N\left(0, \sigma_b^2\right)$$

for i $= 1, \ldots, 294$, j $= 1, \ldots, n_i$.
For GLMM, there's an conditional independence assumption that conditional on the random effects, b_i, Y_{ij} are independent observations from an exponential family distribution. Random

(c) (2 points) What is the link function?

The link function is the logit function. Specifically, for conditional mean response μ_{ij}^*,

$$g\left(\mu_{ij}^*\right) = logit\left(\mu_{ij}^*\right) = \log\left\{\mu_{ij}^* / \left(1 - \mu_{ij}^*\right)\right\}$$

(d) (2 points) What is the linear predictor?

$$\eta_{ij} = (\beta_1 + b_i) + \beta_2 \text{ Time }_{ij} + \beta_3 \text{ Treatment}_i \times \text{Time}_i$$

Note that here Time is the Weeks variable in our data.

(e) Use SAS or R to fit this GLMM:

```
#M1 - Mixed effects logistic regression model with random intercept
m1 <- glmer(Y ~ Weeks + Weeks:Treatment + (1|ID), data=dt,
            family=binomial, control=glmerControl(optimizer="bobyqa"),
            nAGQ=20, na.action=na.omit)
summary(m1)
```

```
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 20) [glmerMod]
##  Family: binomial  ( logit )
## Formula: Y ~ Weeks + Weeks:Treatment + (1 | ID)
##    Data: dt
## Control: glmerControl(optimizer = "bobyqa")
##
##      AIC      BIC   logLik deviance df.resid
##   1255.8   1278.0   -623.9   1247.8     1904
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -2.934 -0.191 -0.089 -0.007 38.570
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  ID     (Intercept) 16.1     4.012
## Number of obs: 1908, groups:  ID, 294
##
## Fixed effects:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.68676    0.32916  -5.124 2.98e-07 ***
## Weeks           -0.10057    0.01121  -8.969  < 2e-16 ***
## Weeks:Treatment -0.04138    0.01714  -2.414   0.0158 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Weeks
## Weeks       -0.044
## Weks:Trtmnt  0.003 -0.527
```

i. (10 points) Test whether there is a significant effect of intervention group on the change in subject-specific probability for moderate/severe onycholysis. Write out your null and alternative hypotheses. Show which test you are using and interpret your findings in words. Please, do not just say 'we reject the null' or 'we have insufficient evidence to reject the null', also state your conclusion in a sentence relative to the research question raised here(i).

Hypothesis: $H_0 : \beta_3 = 0, H_1 : \beta_3 \neq 0$. Since our two models will be nested, we conduct a likelihood ratio test.

```
m2 <- glmer(Y ~ Weeks + (1|ID), data=dt,
            family=binomial, control=glmerControl(optimizer="bobyqa"),
            nAGQ=20, na.action=na.omit)
anova(m1,m2)
```

```
## Data: dt
## Models:
## m2: Y ~ Weeks + (1 | ID)
## m1: Y ~ Weeks + Weeks:Treatment + (1 | ID)
##    Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## m2  3 1259.9 1276.5 -626.94   1253.9
## m1  4 1255.8 1278.0 -623.90   1247.8 6.0738      1    0.01372 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result gives $p = 0.01372 < 0.05$ which is statistically significant. Therefore, we reject the null hypothesis, conclude that $\beta_3 \neq 0$ and there is a significant effect of intervention group on the change in subject-specific probability for moderate/severe onycholysis.

ii. (10 points) Interpret the estimate for the intercept of this model and use two statistical approaches to describe the heterogeneity among subjects in their underlying propensity to have moderate/severe onycholysis prior to randomization.

The intercept -1.68676 is estimated subject-specific log-odds of moderate/severe onycholysis at baseline; subject-specific risk of moderate/severe onycholysis at baseline in both groups: $\mu^*_{ij} = \exp(-1.68676)/(1 + \exp(-1.68676)) = 15.6\%$.
Two statistical approaches:
(1) 95% CI for intercept

```
exp(-1.68676-1.96*4.012)/(1+exp(-1.68676-1.96*4.012))
```

```
## [1] 7.117626e-05
```

```
exp(-1.68676+1.96*4.012)/(1+exp(-1.68676+1.96*4.012))
```

```
## [1] 0.9979272
```

So 95% CI [7.117626e-05 0.9979272]. The CI is wide. That means subjects have a lot of difference. There is a substantial variability in the propensity of experiencing moderate/severe onycholysis: 95% of subjects have a baseline risk of moderate/severe onycholysis that varies from 0% to 99.8%.

(2) ICC

```
icc <- 16.1/(16.1+3.289868)
icc
```

```
## [1] 0.8303306
```

16.1 is the σ^2_b which shows that between subject variance is relatively large for moderate or severe onycholysis at baseline. The marginal correlation, the ICC is 0.8303306, i.e. average over the distribution of random

effects.

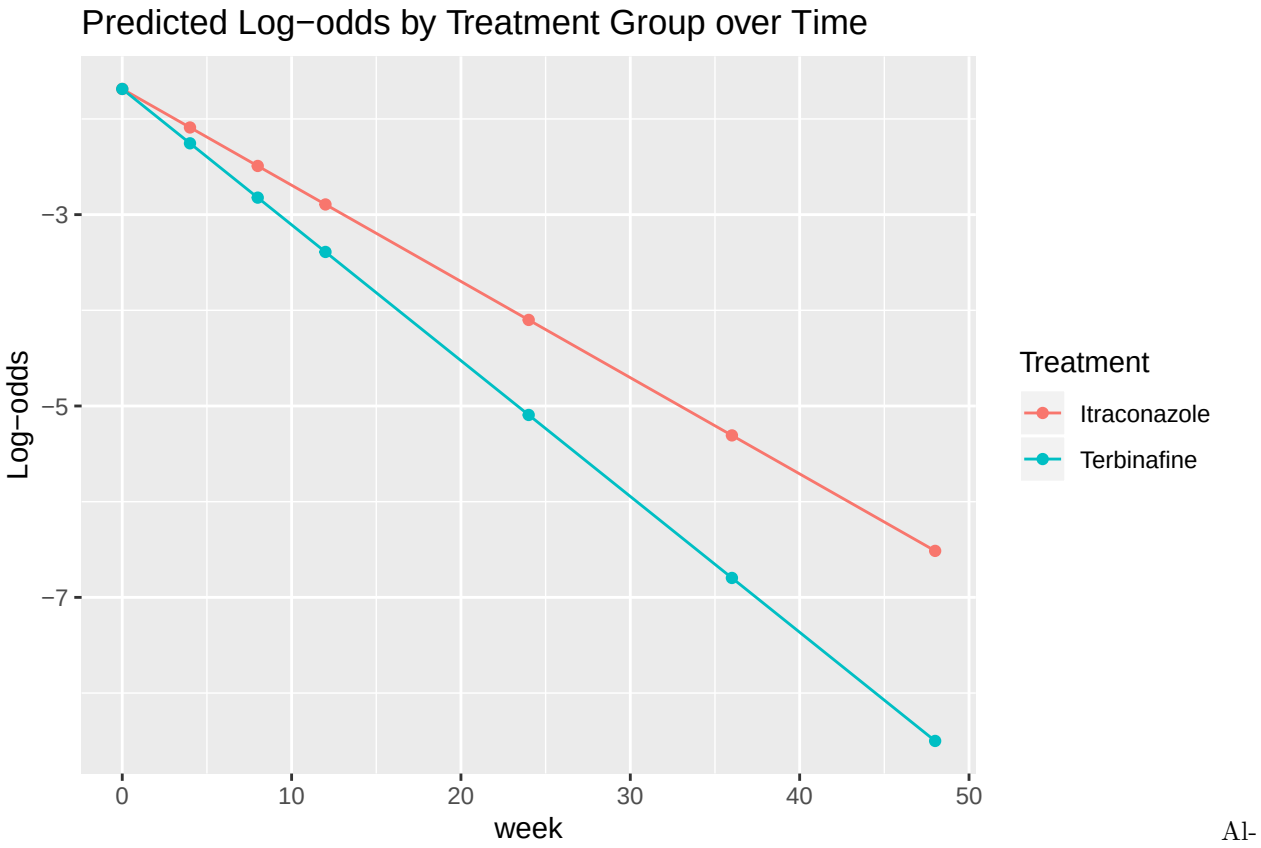
iii. (10 points) Interpret in words the estimated beta coefficients for the variables 'week' and 'week by treatment interaction'.

$\hat{\beta}_2$ -0.10057, the coefficient of week, is an estimate of the change in log odds of moderate or severe onycholysis for each 1 week increase in time for an individual in the Itraconazole group.
$\hat{\beta}_3$ -0.04138, the coefficient of treatment interaction, is an estimate of the difference in the change in log odds of moderate or severe onycholysis at 1 week between an individual assigned Terbinafine and an individual assigned Itraconazole where both individuals have the same baseline value for a moderate or severe onycholysis.

iv. (5points) Figure 3: Estimate subject-specific log-odds of moderate/severe onycholysis in each of the treatment groups across time (weeks 0, 4,8,12,24,36, 48) and plot them on the same figure. Describe what you observe in a few sentences. Compare what you see here to your findings in Q1(d).

```
weeks= c(0,4,8,12,24,36,48)
#treatment = c(0,1)

log_odds = NULL
prob = NULL
trt = NULL
wk = NULL
for(i in 1:length(weeks)) {
  j = 0
  log_odds[i] = -1.68676-0.10057*weeks[i] -0.04138*j*weeks[i]
  prob[i] = exp(log_odds[i])/(1+exp(log_odds[i]))
  trt[i]=j
  wk[i] = i
  }

log_odds1 = NULL
prob1 = NULL
trt1 = NULL
wk1 = NULL
for(i in 1:length(weeks)) {
  j = 1
  log_odds1[i] = -1.68676-0.10057*weeks[i] -0.04138*j*weeks[i]
  prob1[i] = exp(log_odds1[i])/(1+exp(log_odds1[i]))
  trt1[i]=j
  wk1[i] = i
}

log_odds = c(log_odds,log_odds1)
prob = c(prob,prob1)
trt = c(0,0,0,0,0,0,0,1,1,1,1,1,1,1)
wk = c(wk,wk1)
plot_table = cbind(log_odds,prob)
plot_table = as.data.frame(plot_table)
plot_table$treatment = c(0,0,0,0,0,0,0,1,1,1,1,1,1,1)
plot_table$week = c(0,4,8,12,24,36,48,0,4,8,12,24,36,48)

#Create a factor variable for tx
```

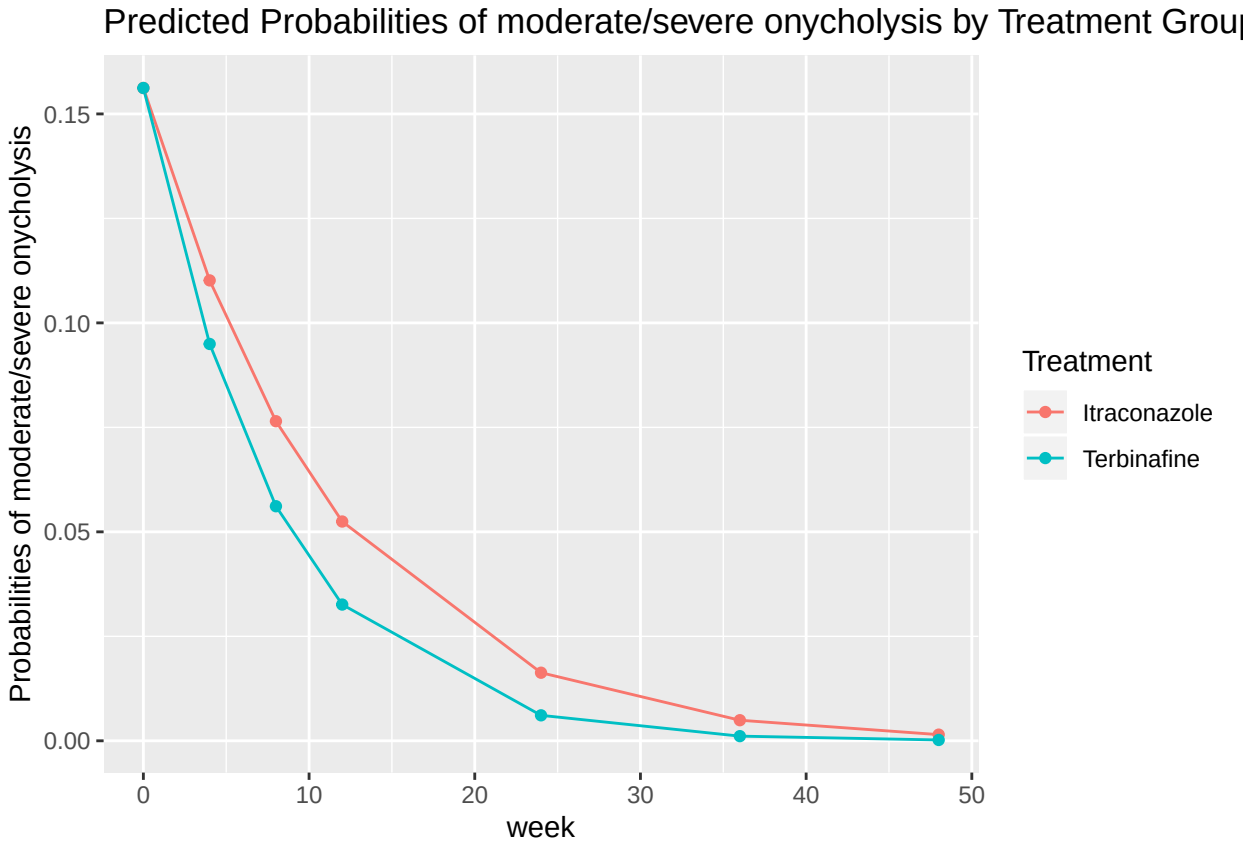
```
plot_table$cTreatment <- factor(plot_table$treatment, levels=c(0,1), labels=c("Itraconazole","Terbinafi

#Plot average predicted log-odds by treatment group over time
p1 <- ggplot(plot_table, aes(x=week,y=log_odds,group=cTreatment))
p1 + geom_line(aes(color=cTreatment)) + geom_point(aes(color=cTreatment)) +
  labs(title="Predicted Log-odds by Treatment Group over Time",
       color="Treatment", y="Log-odds")
```



Although Q1(d) plot is also decreasing, but it's kind of zig-zag. Here we observed two straight linear line. This makes sense because the logit and predictor has linear relationship. The observed and estimated are different. Moreover, we are using the subject specific beta here, not marginal beta which is being use in the observed case.

v. (5 points) Figure 4: Estimate subject-specific probability of moderate/severe onycholysis in each of the treatment groups across time (weeks 0, 4,8,12,24,36, 48) and plot them on the same figure. Describe what you observe in a few sentences. Compare what you see here to your findings in Q1 (c).

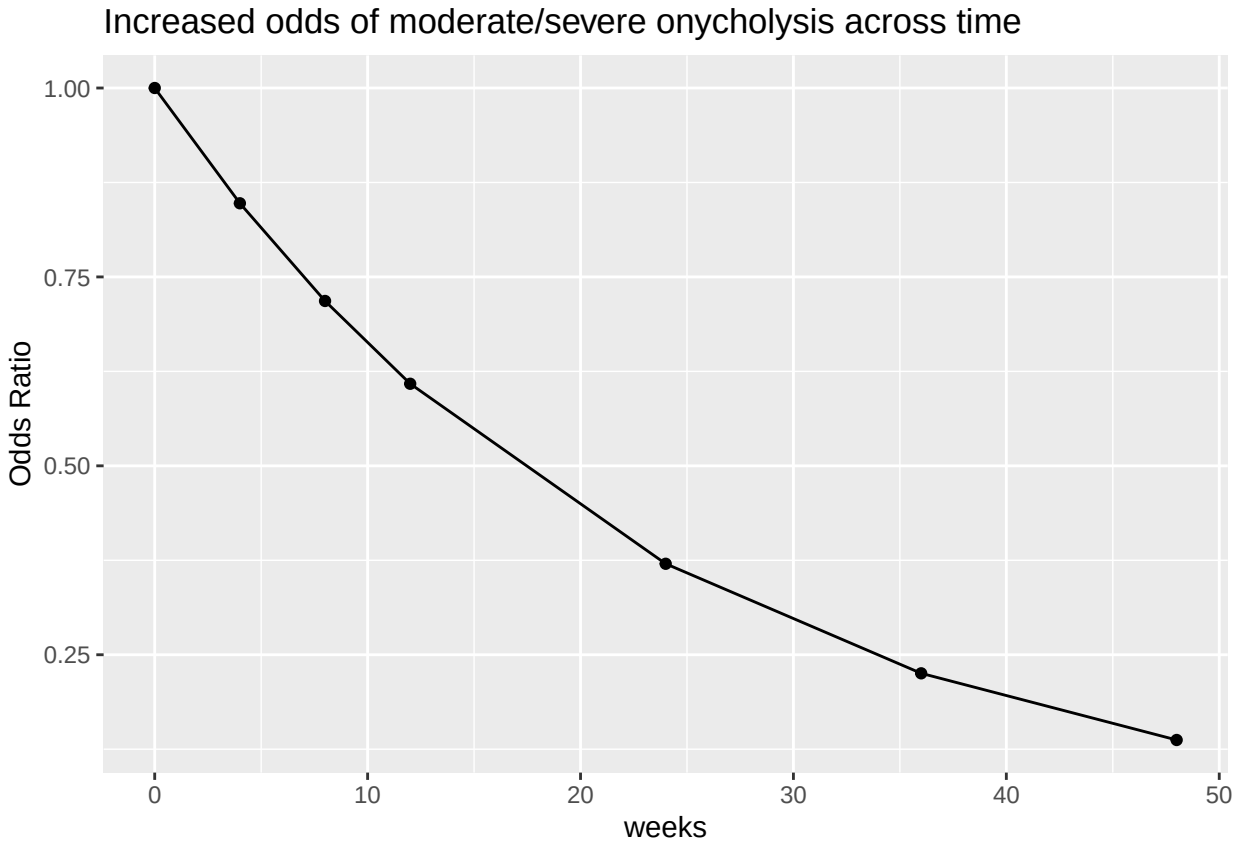
```
p2 <- ggplot(plot_table, aes(x=week,y=prob,group=cTreatment))
p2 + geom_line(aes(color=cTreatment)) + geom_point(aes(color=cTreatment)) +
  labs(title="Predicted Probabilities of moderate/severe onycholysis by Treatment Group over Time",
       color="Treatment", y="Probabilities of moderate/severe onycholysis")
```



Although Q1(c) plot is also decreasing, but it's kind not in a formal shape. Here we observed two quadratic line. This makes sense because once the logit is transformed as the probability, the straight line became a curve due to the effect of exponential. Comparing with Q1(c), the observed and estimated are different. Moreover, we are using the subject specific beta here, not marginal beta which is being use in the observed case.

vi. (5 points) Figure5: Estimate subject-specific ratios of increased odds(odds ratio) of moderate/severe onycholysis across time (weeks 0, 4,8,12,24,36, 48) in subjects randomized to Terbinafine relative to the subjects randomized to Itraconazole. Plot these odds ratios across time. Interpret the estimated odds ratio at 48 weeks of follow up.

```
weeks= c(0,4,8,12,24,36,48)
OR = NULL
for (i in 1:length(weeks)){
  OR = exp(-0.04138*1*weeks)
}

plot_table2 = as.data.frame(cbind(weeks,OR))
p5 <- ggplot(plot_table2, aes(x=weeks,y=OR))
p5 + geom_line() + geom_point() +
  labs(title="Increased odds of moderate/severe onycholysis across time",
       y="Odds Ratio")
```




```
plot_table2
```

```
##   weeks        OR
## 1     0 1.0000000
## 2     4 0.8474529
## 3     8 0.7181764
## 4    12 0.6086207
## 5    24 0.3704192
## 6    36 0.2254448
## 7    48 0.1372104
```

Interpret the estimated odds ratio at 48 weeks of follow up: The ratio of the odds of moderate/severe onycholysis at 48 weeks for a subject assigned to Terbinafine vs. a subject with the same risk of moderate/severe onycholysis who was assigned to Itraconazole is exp(0.137) = 1.15. The subject specific odds in this case are 1.15 times higher for a subject randomied to Terbinafine comparing a subject with similar propensity for onycholysis baseline who's randomied to Itraconazole.

vii. (5 points) Going back to the missing outcome observations across time. What missing data mechanism assumption for the outcome was used in your GLMM to obtain parameter estimates?

(Book 14.5) assume data are missing at random (MAR) but not missing completely at random (MCAR). We can assume MCAR, but it's too strong. It's less common. GLMM can handle MAR so we assume MAR.