SISCER 2022: Longitudinal Data Analysis, Toenail Data (partial solution)

Katie Wilson, Anna Plantinga (Instructors) Yiqun Chen (TA)

Contents

This file provides an analysis in `R` of longitudinal data collected to understand the treatment effect of compounds on toenail infection. This document is the "exercise" version where you get to practice analyzing the dataset (the designed exercises are indicated by the phrase **fill in your answers**).

Data Overview

This data comes from a randomized, double-blind, parallel group, multicenter study for the comparison of two oral treatments for toenail dermatophyte onychomycosis (TDO) (De Backer et al. 1998, Lesaffre et al. 2001). TDO is a common toenail infection, difficult to treat, affecting more than 2 out of 100 persons. Antifungal compounds, classically used for treatment of TDO, need to be taken until the whole nail has grown out healthy. The development of new compounds, however, has reduced the treatment duration to 3 months.

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.

Variables

- Subject ID
- Response (0=none or mild, 1=moderate or severe)
- Treatment (0=Itraconazole, 1=Terbinafine)
- Month: the exact timing of measurements in months (**for simplicity we will ignore this and use Visit instead**)
- Visit: denotes the visit number (visit numbers 1-7 correspond to scheduled visits at 0, 4, 8, 12, 24, 36, and 48 weeks).

Scientific Question

Has the percentage of severe infections decreased over time, and is that evolution different for the two treatment groups?

Set-up

First we will start by considering what the model could look like to answer the research question.

We are interested in analyzing severity of infection. This is a binary indicator, so a natural choice is to use a logistic regression model. Let Y_{ij} be 1 if subject i at time j has moderate/severe response and 0 if none/mild.

Marginal model: A marginal model (that we could fit using the GEE approach) would look as follows:

$$\text{logit}(P(Y_{ij} = 1|X_{ij})) = \beta_0 + \beta_1 \text{txt}_i + \beta_2 t_{ij} + \beta_3 \text{txt}_i \times t_{ij},$$

where t_{ij} is the time of ith subject's jth measurement, and txt_i is 1 if subject i is assigned to the Terbinafine group and 0 otherwise.

- Why is there an interaction term?
  - **Fill in your answers**
- What are the interpretations of β_1, β_2, and β_3?
  - β_1: **Fill in your answers**
  - β_2: **Fill in your answers**
  - β_3: **Fill in your answers**

Random effects model: A random effects model could look as follows:

$$\text{logit}(P(Y_{ij} = 1|X_{ij}, b_{0i})) = \beta_0 + b_{0i} + \beta_1 \text{txt}_i + \beta_2 t_{ij} + \beta_3 \text{txt}_i \times t_{ij}.$$

Here, we have included random intercepts b_{0i} for each subject i.

- What are the interpretations of β_1, β_2, and β_3 in this model?
  - β_1: **Fill in your answers**
  - β_2: **Fill in your answers**
  - β_3: **Fill in your answers**

As we covered in lectures (and see above), marginal and random effects model have **different** interpretations — **which one(s) do you think answer our scientific question of interest better**?

Exploratory Data Analysis

Packages and Data

We first load the packages we will need for this analysis.

```
library(tidyverse)
library(mice) # needed for toenail data
library(VIM)
library(ggplot2)
library(dplyr)
library(multcomp)
library(geepack)
library(lme4)
library(broom.mixed)
library(reshape)
```

If you need to install any of the listed packages (e.g., if you do not have the `geepack` package installed, you might get an error message "Error in library(geepack) : there is no package called `geepack`". In this case, you can install the missing package by

```
install.packages(c('geepack'), repos='https://cloud.r-project.org')
```

Next, we will read in the data (which comes with the mice package) and set parameters for plotting.

```
pal2use <- c("#E69F00", "#0072B2")  # Color palette (optional)
data(toenail)
toenail <- toenail[,c("ID", "outcome", "treatment", "visit")]
head(toenail)
```

```
##   ID outcome treatment visit
## 1  1       1         1     1
## 2  1       1         1     2
## 3  1       1         1     3
## 4  1       0         1     4
## 5  1       0         1     5
## 6  1       0         1     6
```

We will first investigate number and percentage of patients with severe toenail infection, by treatment arm:

```
summary_tab <- toenail %>%
  group_by(treatment, visit) %>%
  summarise(n_obs = n(),
            n_severe = sum(outcome),
            percent_severe = mean(outcome))
summary_tab
```

```
## # A tibble: 14 x 5
## # Groups:   treatment [2]
##    treatment visit n_obs n_severe percent_severe
##        <int> <int> <int>    <int>          <dbl>
##  1         0     1   146       54         0.370
##  2         0     2   141       49         0.348
##  3         0     3   138       44         0.319
##  4         0     4   132       29         0.220
##  5         0     5   130       14         0.108
##  6         0     6   117       10         0.0855
##  7         0     7   133       14         0.105
##  8         1     1   148       55         0.372
##  9         1     2   147       48         0.327
## 10         1     3   145       40         0.276
## 11         1     4   140       29         0.207
## 12         1     5   133        8         0.0602
## 13         1     6   127        8         0.0630
## 14         1     7   131        6         0.0458
```

Number of available repeated measurements per subject, by treatment arm:

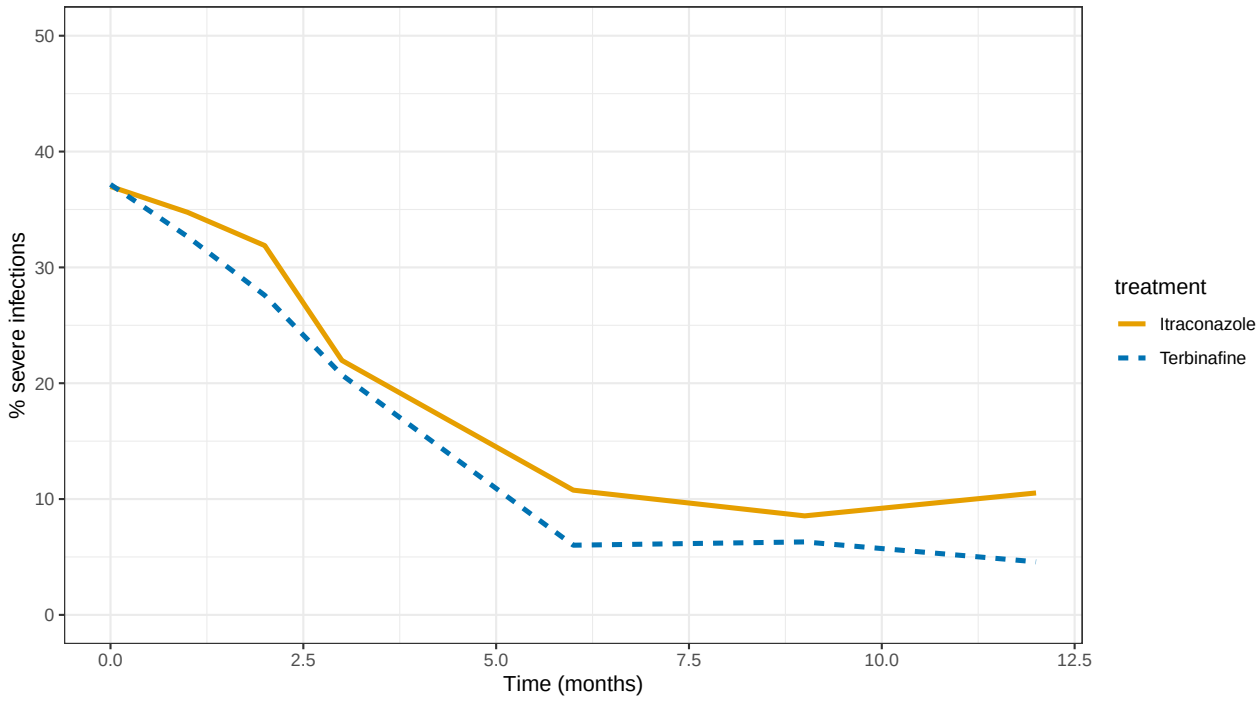
```
(toenail %>% group_by(treatment, ID) %>%
   summarise(nvisit = n())) %>%
  group_by(treatment, nvisit) %>%
  summarise(N = n())
```

```
## # A tibble: 14 x 3
## # Groups:   treatment [2]
##    treatment nvisit     N
##        <int>  <int> <int>
##  1         0      1     4
##  2         0      2     2
##  3         0      3     4
##  4         0      4     2
```

```
##  5          0       5     2
##  6          0       6    25
##  7          0       7   107
##  8          1       1     1
##  9          1       2     1
## 10          1       3     3
## 11          1       4     4
## 12          1       5     8
## 13          1       6    14
## 14          1       7   117
```

Figure depicting change in percentage of severe toenail infections in the two treatment groups separately:

```
visit_mo <- c(0, 4, 8, 12, 24, 36, 48)/4
summary_tab$visit_mo <- visit_mo[summary_tab$visit]
toenail$visit_mo <- visit_mo[toenail$visit]
summary_tab$treatment <- factor(summary_tab$treatment)
levels(summary_tab$treatment) <- c("Itraconazole", "Terbinafine")
ggplot(summary_tab,
       aes(x=visit_mo,
           y=percent_severe*100,
           group=treatment)) +
  geom_line(aes(color=treatment, linetype=treatment), size=1.2) +
  theme_bw() + ylab("% severe infections") + ylim(c(0, 50)) +
  xlab("Time (months)") + scale_color_manual(values = pal2use)
```



We see that at baseline the two groups have very similar proportion of severe infections (question: is this a coincidence?); there is some evidence the group treated with Terbinafine has a smaller proportion of severe toenail infections over time.

We will investigate the proportion of missing data in the two different groups next. We will first get the data in the so-called "wide format", because missed visits are not explicitly included in the original ("long format") dataset.

```
toenail_wide <- reshape(toenail, timevar = c("visit"), idvar = "ID",
                        v.names = c("outcome"),
                        direction = "wide")
head(toenail_wide, 5)
```

```
##    ID treatment visit_mo outcome.1 outcome.2 outcome.3 outcome.4 outcome.5
## 1   1         1        0         1         1         1         0         0
## 8   2         0        0         0         0         1         1         0
## 14  3         0        0         0         0         0         0         0
## 21  4         0        0         1         0         0         0         0
## 28  6         1        0         1         1         1         0         0
##    outcome.6 outcome.7
## 1          0         0
## 8          0        NA
## 14         0         1
## 21         0         0
## 28         0         0
```

Now if we convert the wide format back to the long format, we see that the missing visits are included!

```
toenail_long_with_missing <- reshape(toenail_wide,
                                     timevar = c("visit"),
                                     idvar = "ID",
                        v.names = colnames(toenail_wide)[4:10],
                        direction = "long") %>%
  arrange(ID, visit_mo)
# put back visit month
visit_mo <- c(0, 4, 8, 12, 24, 36, 48)/4
toenail_long_with_missing$visit_mo <- visit_mo[toenail_long_with_missing$visit]
toenail_long_with_missing[8:14, ] # see Subj 2 to observe missing
```

```
##     ID treatment visit_mo visit outcome
## 2.1  2         0        0     1       0
## 2.2  2         0        1     2       0
## 2.3  2         0        2     3       1
## 2.4  2         0        3     4       1
## 2.5  2         0        6     5       0
## 2.6  2         0        9     6       0
## 2.7  2         0       12     7      NA
```

We can obtain summary statistics for the proportion of missing data by treatment and by visit:

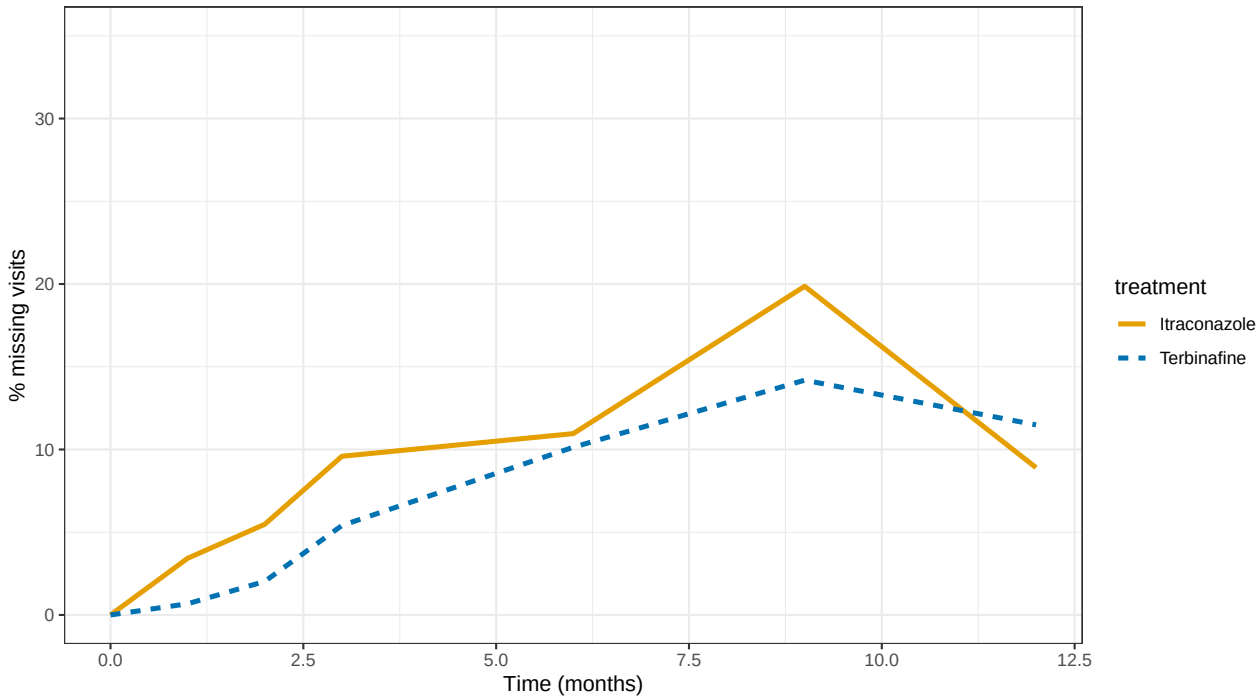
```
miss_summary <- toenail_long_with_missing %>%
  group_by(treatment, visit_mo) %>%
  summarise(n_missing = sum(is.na(outcome)),
            n_total = n(),
            prop_missing = n_missing/n_total)
head(miss_summary, 5)
```

```
## # A tibble: 5 x 5
## # Groups:   treatment [1]
##   treatment visit_mo n_missing n_total prop_missing
##       <int>    <dbl>     <int>   <int>        <dbl>
## 1         0        0         0     146       0
## 2         0        1         5     146       0.0342
## 3         0        2         8     146       0.0548
```

```
## 4         0        3        14      146       0.0959
## 5         0        6        16      146       0.110
```

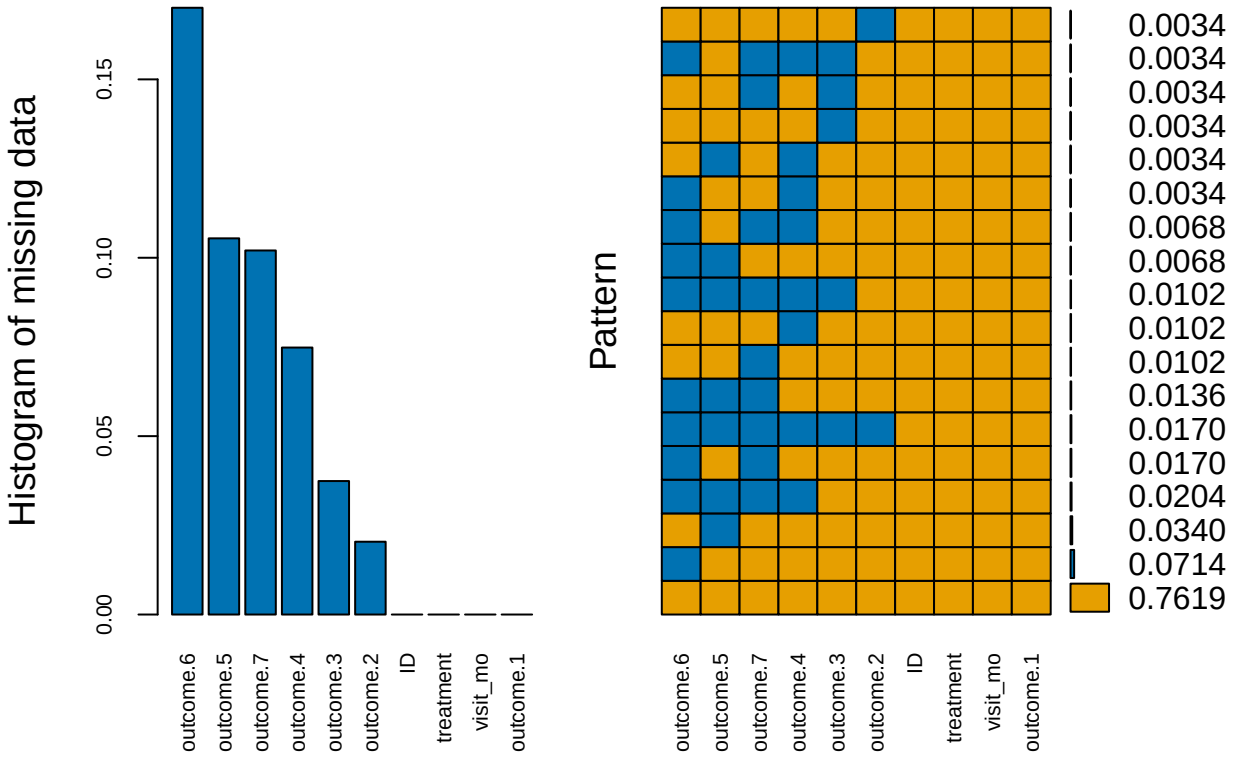
We can also visualize the proportion of missing data using the code below; it seems like two groups have comparable amount of missing data, with the Itraconazole group having a slightly higher proportion over time.

```
miss_summary$treatment <- as.factor(miss_summary$treatment)
levels(miss_summary$treatment) <- c("Itraconazole", "Terbinafine")
ggplot(miss_summary,
       aes(x=visit_mo,
           y=prop_missing*100,
           group=treatment)) +
  geom_line(aes(color=treatment, linetype=treatment), size=1.2) +
  theme_bw() + ylab("% missing visits") + ylim(c(0, 35)) +
  xlab("Time (months)") + scale_color_manual(values = pal2use)
```



Bonus: If you have experience with missing data before, you might have seen the `aggr` plot below, which visualizes (i) how much missing data is there (note the table created we created before); and (ii) what are the patterns of missing data:

```
# aggr() is in package VIM
aggr_plot <- aggr(toenail_wide,
                  col=pal2use,
                  numbers=TRUE,
                  sortVars=TRUE,
                  labels=names(toenail_wide),
                  cex.axis=.7,
                  gap=3,
                  ylab=c("Histogram of missing data","Pattern"))
```




```
##
##  Variables sorted by number of missings:
##   Variable      Count
##  outcome.6 0.17006803
##  outcome.5 0.10544218
##  outcome.7 0.10204082
##  outcome.4 0.07482993
##  outcome.3 0.03741497
##  outcome.2 0.02040816
##         ID 0.00000000
##  treatment 0.00000000
##   visit_mo 0.00000000
##  outcome.1 0.00000000
```

In particular, we see that the 76.2% of participants have all seven visits; the most common missingness pattern is missing the toenail infection status at month 9 (17.0%), followed by missing the outcome at month 6 (10.5%), and month 12 (10.2%).

Question: What else do you notice about the missing data? Are there any other figures/numbers that might be useful?

Inferential Analysis

Marginal model with GEE

We will first fit a marginal model (fit using GEE, with working independence covariance matrix).

In order to answer the scientific question of interest, we could present the estimated ratio of odds of severe infection over time for each group:

- Itraconazole: $\exp(\beta_2)$

  - $H_0 : \beta_2 = 0$ the odds ratio of severe infection is constant over time among those assigned to Itraconazole
  - $\beta_2 < 0$: odds of severe infection is decreasing over time
- Terbinafine: $\exp(\beta_2 + \beta_3)$
  - $H_0 : \beta_2 + \beta_3 = 0$ the odds ratio of severe infection is constant over time among those assigned to Terbinafine
  - $\beta_2 + \beta_3 < 0$: odds of severe infection is decreasing over time

To determine if the evolution is different for the two treatment groups we would want to test the null hypothesis: **Fill in your answers**. From above, if $\beta_3 = 0$ then $\exp(\beta_2) = \exp(\beta_2 + 0)$; or in words the two treatment groups have the same odds ratio over time.

One approach we could use is an available data analysis where we exclude all missing observations (but include subjects even if they have missing data). For GEE, this approach is valid if the data is missing completely at random.

```
mod_available <- geeglm(outcome ~ treatment*visit_mo, id = ID,
                        data=toenail,
                        family=binomial(link="logit"))
summary(mod_available)
```

```
##
## Call:
## geeglm(formula = outcome ~ treatment * visit_mo, family = binomial(link = "logit"),
##     data = toenail, id = ID)
##
##  Coefficients:
##                    Estimate  Std.err   Wald Pr(>|W|)
## (Intercept)        -0.55706  0.17134 10.570  0.00115 **
## treatment           0.02402  0.25061  0.009  0.92363
## visit_mo           -0.17693  0.03017 34.394  4.5e-09 ***
## treatment:visit_mo -0.07833  0.05461  2.057  0.15147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = independence
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)    1.046  0.4169
## Number of clusters:   294  Maximum cluster size: 7
```

Point estimate and confidence interval for the ratio of odds of severe infection per month in the Itraconazole group ($\exp(\beta_2)$) can be found:

```
i_avail <- glht(mod_available, "visit_mo = 0")
summary(i_avail)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: geeglm(formula = outcome ~ treatment * visit_mo, family = binomial(link = "logit"),
##     data = toenail, id = ID)
##
## Linear Hypotheses:
##               Estimate Std. Error z value Pr(>|z|)
## visit_mo == 0  -0.1769     0.0302   -5.86  4.5e-09 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

```
# CI for difference in log odds ratio
confint(i_avail)
```

```
##
##   Simultaneous Confidence Intervals
##
## Fit: geeglm(formula = outcome ~ treatment * visit_mo, family = binomial(link = "logit"),
##     data = toenail, id = ID)
##
## Quantile = 1.96
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##               Estimate lwr    upr
## visit_mo == 0 -0.177   -0.236 -0.118
```

```
# CI for ratio of odds
exp(confint(i_avail)$confint)
```

```
##          Estimate    lwr    upr
## visit_mo   0.8378 0.7897 0.8889
## attr(,"conf.level")
## [1] 0.95
## attr(,"calpha")
## [1] 1.96
```

Point estimate and confidence interval for the ratio of odds of severe infection per month in the Terbinafine group ($\exp(\beta_2 + \beta_3)$) can be found:

```
t_avail <- glht(mod_available, "visit_mo + treatment:visit_mo = 0")
summary(t_avail)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: geeglm(formula = outcome ~ treatment * visit_mo, family = binomial(link = "logit"),
##     data = toenail, id = ID)
##
## Linear Hypotheses:
##                                    Estimate Std. Error z value Pr(>|z|)
## visit_mo + treatment:visit_mo == 0  -0.2553     0.0455   -5.61    2e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

```
# CI for difference in log odds ratio
confint(t_avail)
```

```
##
##   Simultaneous Confidence Intervals
##
## Fit: geeglm(formula = outcome ~ treatment * visit_mo, family = binomial(link = "logit"),
##     data = toenail, id = ID)
```

```
##
## Quantile = 1.96
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##                                    Estimate lwr    upr
## visit_mo + treatment:visit_mo == 0 -0.255   -0.344 -0.166
```

```
# CI for ratio of odds
exp(confint(t_avail)$confint)
```

```
##                                Estimate    lwr   upr
## visit_mo + treatment:visit_mo   0.7747 0.7086 0.847
## attr(,"conf.level")
## [1] 0.95
## attr(,"calpha")
## [1] 1.96
```

We summarize the inferential results using GEE below:

- Among those assigned to Itraconazole, the odds of severe infection is estimated to be **Fill in your answers** lower (95% CI: 21% lower to 11% lower) per month.
- Among those assigned to Terbinafine, the odds of severe infection is estimated to be **Fill in your answers** lower (95% CI: 29% lower to 15% lower) per month.
- We do not have evidence that the ratio of odds of severe infection over time differs between the two groups (p = **Fill in your answers**).

Random effects model

Next, we will fit the random effects model with only random intercepts (note: the interpretations of the coefficients will be different!) using the `glmer` function. As in the marginal model case, we will adopt an available data analysis approach. For random effects model, this approach is valid if the data is missing at random and the likelihood is correctly specified.

```
# nAGQ is an argument that determines the accuracy of the solution
# (the higher the better accuracy but also longer run time) -- defaults to 1,
# not applicable for random slopes.
glmm.fit <- glmer(outcome ~ treatment*visit_mo + (1|ID),
                  data=toenail,
                  family=binomial(link="logit"),
                  nAGQ=25)
summary(glmm.fit)
```

```
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 25) [glmerMod]
##  Family: binomial  ( logit )
## Formula: outcome ~ treatment * visit_mo + (1 | ID)
##    Data: toenail
##
##      AIC      BIC   logLik deviance df.resid
##   1257.9   1285.6   -623.9   1247.9     1903
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
##  -2.95  -0.19  -0.09  -0.01  38.17
```

```
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  ID     (Intercept) 16.1     4.01
## Number of obs: 1908, groups:  ID, 294
##
## Fixed effects:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.6270     0.4341   -3.75  0.00018 ***
## treatment           -0.1140     0.5843   -0.20  0.84530
## visit_mo            -0.4041     0.0460   -8.79  < 2e-16 ***
## treatment:visit_mo  -0.1613     0.0718   -2.25  0.02474 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) trtmnt vist_m
## treatment   -0.652
## visit_mo    -0.171  0.211
## trtmnt:vst_  0.201 -0.299 -0.554
```

Point estimate and confidence interval for β_2 and $\exp(\beta_2)$:

```
i_glmm <- glht(glmm.fit, "visit_mo = 0")
summary(i_glmm)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: glmer(formula = outcome ~ treatment * visit_mo + (1 | ID), data = toenail,
##     family = binomial(link = "logit"), nAGQ = 25)
##
## Linear Hypotheses:
##               Estimate Std. Error z value Pr(>|z|)
## visit_mo == 0   -0.404      0.046   -8.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

```
# \beta_2
confint(i_glmm)
```

```
##
##   Simultaneous Confidence Intervals
##
## Fit: glmer(formula = outcome ~ treatment * visit_mo + (1 | ID), data = toenail,
##     family = binomial(link = "logit"), nAGQ = 25)
##
## Quantile = 1.96
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##               Estimate lwr    upr
## visit_mo == 0 -0.404   -0.494 -0.314
```

```
# exp(\beta_2)
exp(confint(i_glmm)$confint)
```

```
##          Estimate  lwr    upr
## visit_mo   0.6675 0.61 0.7305
## attr(,"conf.level")
## [1] 0.95
## attr(,"calpha")
## [1] 1.96
```

Point estimate and confidence interval for the ratio of odds of severe infection per month in the Terbinafine group ($\exp(\beta_2 + \beta_3)$) can be found:

```
t_glmm <- glht(glmm.fit, "visit_mo + treatment:visit_mo = 0")
summary(t_glmm)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: glmer(formula = outcome ~ treatment * visit_mo + (1 | ID), data = toenail,
##     family = binomial(link = "logit"), nAGQ = 25)
##
## Linear Hypotheses:
##                                     Estimate Std. Error z value Pr(>|z|)
## visit_mo + treatment:visit_mo == 0  -0.5654     0.0601   -9.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

```
# (\beta_2 + \beta_3)
confint(t_glmm)
```

```
##
##   Simultaneous Confidence Intervals
##
## Fit: glmer(formula = outcome ~ treatment * visit_mo + (1 | ID), data = toenail,
##     family = binomial(link = "logit"), nAGQ = 25)
##
## Quantile = 1.96
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##                                     Estimate lwr    upr
## visit_mo + treatment:visit_mo == 0 -0.565   -0.683 -0.448
```

```
# exp(\beta_2 + \beta_3)
exp(confint(t_glmm)$confint)
```

```
##                               Estimate   lwr    upr
## visit_mo + treatment:visit_mo   0.5681 0.505 0.6391
## attr(,"conf.level")
## [1] 0.95
## attr(,"calpha")
## [1] 1.96
```

We summarize the inferential results using GLMM below:

- For a typical subject in the Itraconazole group, the odds of severe infection is estimated to be **Fill in your answers** lower (95% CI: 27% lower to 39% lower) per month.
- For a typical subject in the Terbinafine group, the odds of severe infection over month is estimated to be **Fill in your answers** lower (95% CI: 36% lower to 50% lower) per month.
- We find **evidence/no evidence** (**Fill in your answers**) that for two subjects who happen to have the same underlying risk of experiencing the infection but who are in different treatment groups, their odds of severe infection per month are different (p=**Fill in your answers**).

References:

De Backer, M., De Vroey, C., Lesaffre, E., Scheys, I., and De Keyser, P. (1998), "Twelve weeks of continuous oral therapy for toenail onychomycosis caused by dermatophytes: a double-blind comparative trial of terbinafine 250 mg/day versus itraconazole 200 mg/day," Journal of the American Academy of Dermatology, 38, S57–63. https://doi.org/10.1016/s0190-9622(98)70486-4.

Lesaffre, E., and Spiessens, B. (2001), "On the effect of the number of quadrature points in a logistic random effects model: an example," Journal of the Royal Statistical Society. Series C, Applied statistics, Wiley, 50, 325–335. https://doi.org/10.1111/1467-9876.00237.