Analyzing Incomplete Longitudinal Data

Alexander McLain

The data are from a longitudinal clinical trial of contracepting women. In this trial women received an injection of either 100 mg or 150 mg of depot-medroxyprogesterone acetate (DMPA) on the day of randomization and three additional injections at 90-day intervals. There was a final follow-up visit 90 days after the fourth injection, i.e., one year after the first injection.

Throughout the study each woman completed a menstrual diary that recorded any vaginal bleeding pattern disturbances. The diary data were used to determine whether a women experienced amenorrhea, the absence of menstrual bleeding for a specified number of days.

A total of 1151 women completed the menstrual diaries and the diary data were used to generate a binary sequence for each woman according to whether or not she had experienced amenorrhea in the four successive three month intervals.

Reference: Machin D, Farley T, Busca B, Campbell M and d'Arcangues C. (1988). *Assessing changes in vaginal bleeding patterns in contracepting women.* **Contraception**, 38, 165-179.

```
library(sas7bdat)
library(tidyverse)
library(mice)
library(lattice)
amenor <- read.sas7bdat("amenorrhea.sas7bdat")
head(amenor)
```

```
##   ID TRT TIME   Y Ctime prevy
## 1  1   0    1   0     1   NaN
## 2  1   0    2 NaN     2     0
## 3  1   0    3 NaN     3   NaN
## 4  1   0    4 NaN     4   NaN
## 5  2   0    1   0     1   NaN
## 6  2   0    2 NaN     2     0
```

```
summary(amenor)
```

```
##        ID            TRT              TIME            Y               Ctime
##  Min.   :   1   Min.   :0.0000   Min.   :1.00   Min.   :0.0000   Min.   :1.00
##  1st Qu.: 288   1st Qu.:0.0000   1st Qu.:1.75   1st Qu.:0.0000   1st Qu.:1.75
##  Median : 576   Median :0.0000   Median :2.50   Median :0.0000   Median :2.50
##  Mean   : 576   Mean   :0.4996   Mean   :2.50   Mean   :0.3404   Mean   :2.50
##  3rd Qu.: 864   3rd Qu.:1.0000   3rd Qu.:3.25   3rd Qu.:1.0000   3rd Qu.:3.25
##  Max.   :1151   Max.   :1.0000   Max.   :4.00   Max.   :1.0000   Max.   :4.00
##                                                 NA's   :988
##      prevy
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.3402
##  3rd Qu.:1.0000
```

```
##  Max.   :1.0000
##  NA's   :989
```

Notice that this data already has the previous Y as a variable. We can use that variable to perform multiple imputation. If we wanted to go further, we could make the data set wide and use all the variables to predict each other. That would ignore the time component or any other time-varying variables. Alternatively, we could create further lag variables (lag2 and lag3) and use those in the imputation.

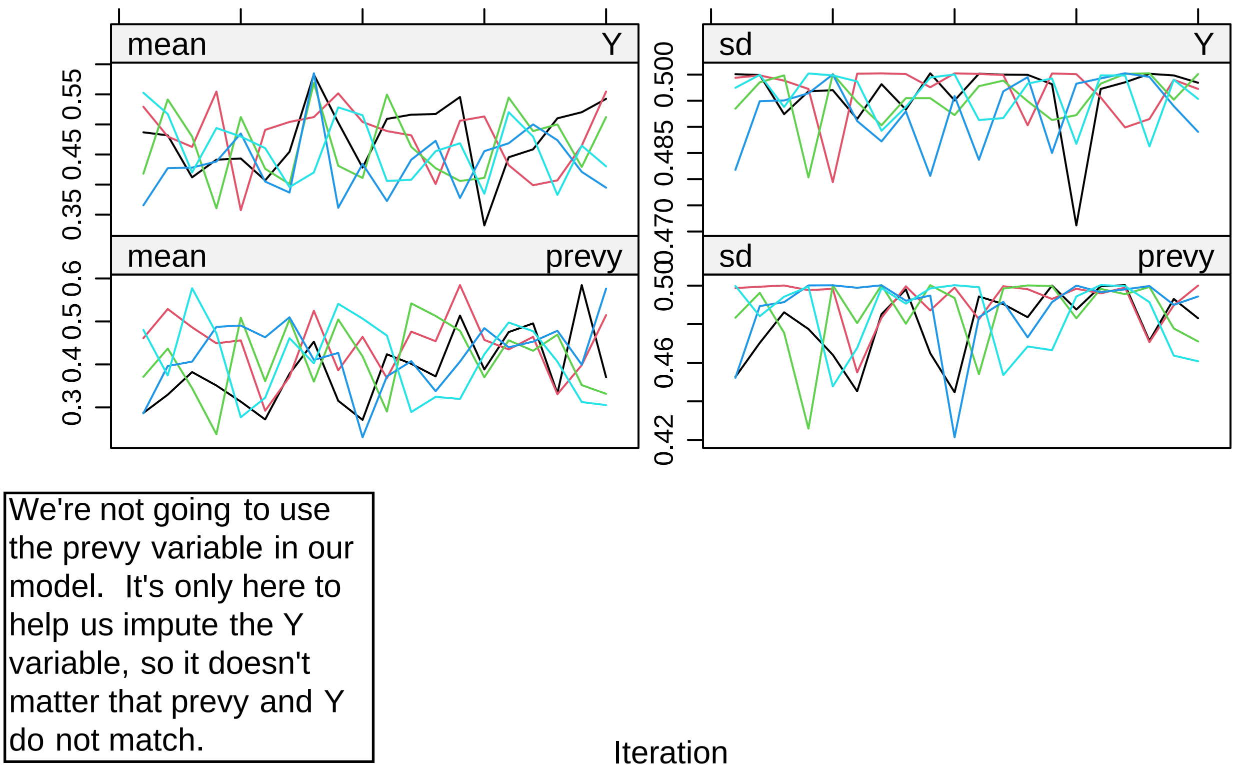
```
amen_imp1 <- mice(amenor, maxit = 0)
pred <- amen_imp1$pred
pred
```

```
##       ID TRT TIME Y Ctime prevy
## ID     0   1    1 1     0     1
## TRT    1   0    1 1     0     1
## TIME   1   1    0 1     0     1
## Y      1   1    1 0     0     1
## Ctime  1   1    1 1     0     1
## prevy  1   1    1 1     0     0
```

```
pred[ , c(1)] <- 0
amen_imp <- mice(amenor, pred = pred, maxit = 20, print = FALSE, seed = 123)
amen_imp
```

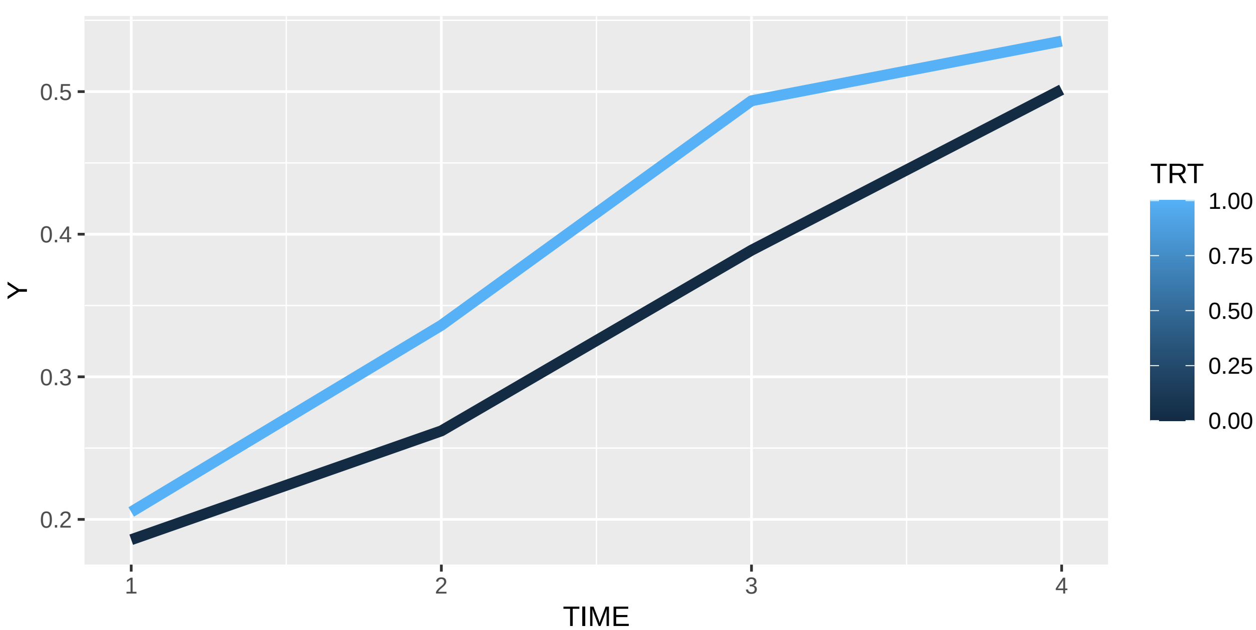
```
## Class: mids
## Number of multiple imputations:  5
## Imputation methods:
##    ID   TRT  TIME     Y Ctime prevy
##    ""    ""    "" "pmm"    "" "pmm"
## PredictorMatrix:
##       ID TRT TIME Y Ctime prevy
## ID     0   1    1 1     0     1
## TRT    0   0    1 1     0     1
## TIME   0   1    0 1     0     1
## Y      0   1    1 0     0     1
## Ctime  0   1    1 1     0     1
## prevy  0   1    1 1     0     0
```

```
plot(amen_imp)
```



Now let's use the imputed data to analyze the outcome of interest. First, let's do a little exploratory analysis with the complete data:

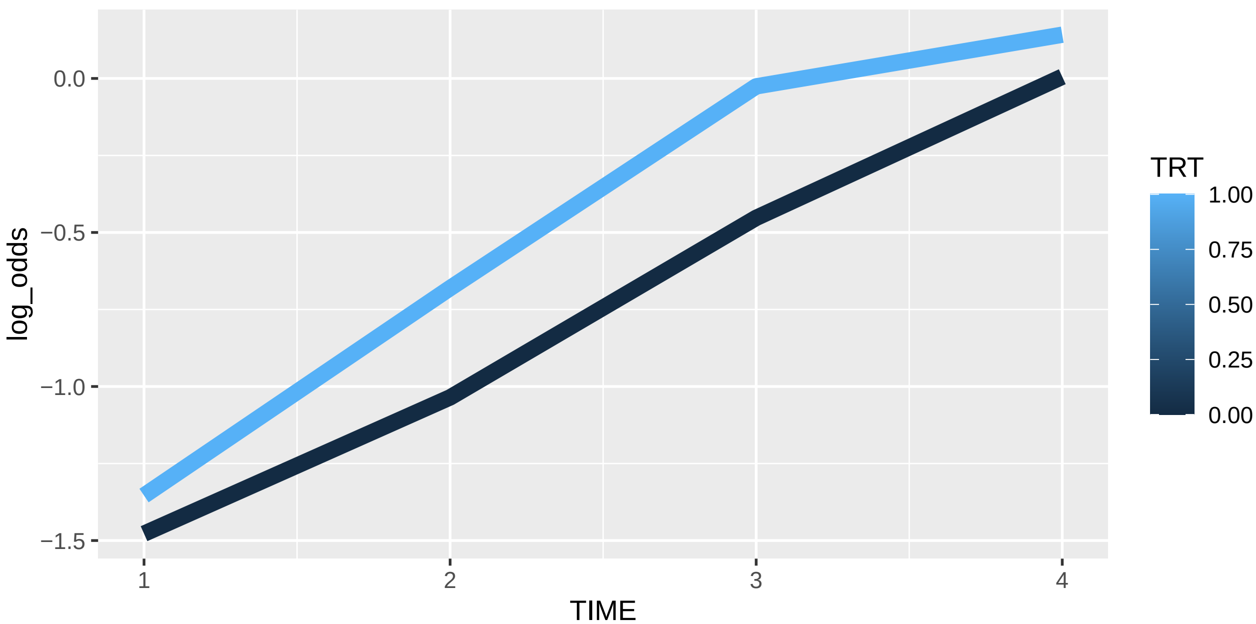
```
p <- ggplot(data = amenor, aes(x = TIME, y = Y, group = ID))
p + stat_summary(aes(group = TRT, col = TRT), geom = "line", fun = mean,
               size = 2)
```



However, this tells us what's happening to the probability. We're modeling the log odds. So let's look at those.

```
amenor_stats <- amenor %>%  group_by(TRT,TIME) %>%
  summarize(mean_prob = mean(Y, na.rm = TRUE)) %>%
  mutate(log_odds = log( mean_prob / (1 - mean_prob) ) )


p <- ggplot(data = amenor_stats, aes(x = TIME, y = log_odds, group = TRT,
                                     col = TRT))
p + geom_line(lwd = 3)
```



1 Using `glmer` with imputed data.

Now let's fit the model to the imputed data using the `with` and `pool` functions. Note that when using the `with` function we need to specify the formula within the function (not outside of the function):

```
library(lme4)
```

```
## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
system.time(MI_GLMM <- with(amen_imp, glmer( Y ~ TIME + TRT + TIME*TRT + (1|ID),
                                              family = binomial, nAGQ = 5)))
```

```
##    user  system elapsed
##   6.630   0.022   6.662
```

```
MI_GLMM
```

```
## call :
## with.mids(data = amen_imp, expr = glmer(Y ~ TIME + TRT + TIME *
```

```
##     TRT + (1 | ID), family = binomial, nAGQ = 5))
##
## call1 :
## mice(data = amenor, predictorMatrix = pred, maxit = 20, printFlag = FALSE,
##     seed = 123)
##
## nmis :
##    ID   TRT  TIME     Y Ctime prevy
##     0     0     0   988     0   989
##
## analyses :
## [[1]]
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 5) [glmerMod]
##  Family: binomial  ( logit )
## Formula: Y ~ TIME + TRT + TIME * TRT + (1 | ID)
##       AIC       BIC    logLik  deviance  df.resid
##  5247.042  5279.216 -2618.521  5237.042      4599
## Random effects:
##  Groups Name        Std.Dev.
##  ID     (Intercept) 1.624
## Number of obs: 4604, groups:  ID, 1151
## Fixed Effects:
## (Intercept)         TIME          TRT     TIME:TRT
##    -3.22528      0.92104      0.53209     -0.08545
##
## [[2]]
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 5) [glmerMod]
##  Family: binomial  ( logit )
## Formula: Y ~ TIME + TRT + TIME * TRT + (1 | ID)
##       AIC       BIC    logLik  deviance  df.resid
##  5253.760  5285.933 -2621.880  5243.760      4599
## Random effects:
##  Groups Name        Std.Dev.
##  ID     (Intercept) 1.705
## Number of obs: 4604, groups:  ID, 1151
## Fixed Effects:
## (Intercept)         TIME          TRT     TIME:TRT
##     -3.0959       0.8739       0.4161      -0.0455
##
## [[3]]
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 5) [glmerMod]
##  Family: binomial  ( logit )
## Formula: Y ~ TIME + TRT + TIME * TRT + (1 | ID)
##       AIC       BIC    logLik  deviance  df.resid
##  5202.069  5234.243 -2596.035  5192.069      4599
## Random effects:
##  Groups Name        Std.Dev.
##  ID     (Intercept) 1.657
## Number of obs: 4604, groups:  ID, 1151
## Fixed Effects:
## (Intercept)         TIME          TRT     TIME:TRT
```

```
##    -3.10137      0.86499      0.09521      0.06033
##
## [[4]]
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 5) [glmerMod]
##  Family: binomial  ( logit )
## Formula: Y ~ TIME + TRT + TIME * TRT + (1 | ID)
##       AIC       BIC    logLik  deviance  df.resid
##  5360.373  5392.547 -2675.187  5350.373      4599
## Random effects:
##  Groups Name        Std.Dev.
##  ID     (Intercept) 1.494
## Number of obs: 4604, groups:  ID, 1151
## Fixed Effects:
## (Intercept)         TIME          TRT     TIME:TRT
##     -2.7290       0.6726       0.5961      -0.1254
##
## [[5]]
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 5) [glmerMod]
##  Family: binomial  ( logit )
## Formula: Y ~ TIME + TRT + TIME * TRT + (1 | ID)
##       AIC       BIC    logLik  deviance  df.resid
##  5394.492  5426.665 -2692.246  5384.492      4599
## Random effects:
##  Groups Name        Std.Dev.
##  ID     (Intercept) 1.5
## Number of obs: 4604, groups:  ID, 1151
## Fixed Effects:
## (Intercept)         TIME          TRT     TIME:TRT
##   -2.499147     0.603385     0.260887     0.003798
```

To pool the estimates we need to install another package `broom.mixed`.

```r
library(broom.mixed)
summary(est <- pool(MI_GLMM)) #pool my results
```

```
##          term    estimate std.error  statistic        df      p.value
## 1 (Intercept) -2.93014491 0.3762891 -7.7869506  6.463052 0.0001631151
## 2        TIME  0.78718373 0.1625751  4.8419709  4.999183 0.0047088563
## 3         TRT  0.38006946 0.3192293  1.1905845 16.547922 0.2506163411
## 4    TIME:TRT -0.03844842 0.1075942 -0.3573464 12.965280 0.7265820140
```

This does not provide an estimate of the random effect variance. We can work around this by getting one manually.

```r
# Extract the variances from each model
vars_est <- lapply( 1:5, function(x){as.data.frame(VarCorr(MI_GLMM$analyses[[x]]))$vcov})

# Take the mean of those
mean( unlist(vars_est) )
```

```
## [1] 2.553859
```

2 Estimating with a weighted GEE.

To fit a weighted gee we're going to use the `wgeesel` package. This package allows for easy specification and estimation of weighted GEE models. This package works for linear, logistic and poisson regression models.

```r
library(wgeesel)
args(wgee)
```

```
## function (model, data, id, family, corstr, scale = NULL, mismodel = NULL,
##     maxit = 200, tol = 0.001)
## NULL
```

The `model`, `family` and `corstr` are the same as we had before. What's new with this is the `mismodel` arguement, which will be a symbolic description of the missingness model to be fitted. The first thing we need to do is to create a variable that indicates an observation is missing

```r
amenor <- amenor %>% mutate( R = ifelse( is.na(Y), 0, 1))
head(amenor)
```

```
##   ID TRT TIME   Y Ctime prevy R
## 1  1   0    1   0     1   NaN 1
## 2  1   0    2 NaN     2     0 0
## 3  1   0    3 NaN     3   NaN 0
## 4  1   0    4 NaN     4   NaN 0
## 5  2   0    1   0     1   NaN 1
## 6  2   0    2 NaN     2     0 0
```

```r
# Note that id = amenor$ID not id = ID
fit <- wgee( Y ~ TIME + TRT + TIME*TRT, data = amenor, id = amenor$ID, family="binomial",
            corstr="exchangeable", scale = NULL,
            mismodel= R ~ TIME + TRT + TIME*TRT + prevy)
```

First, let's look at the missingness model

```r
summary(fit$mis_fit)
```

```
##
## Call:
## glm(formula = mismodel, family = binomial(), data = data[adjusted_idx,
##     ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2135   0.4445   0.5572   0.6054   0.7819
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.1198     0.2698   4.151 3.31e-05 ***
## TIME          0.2863     0.0942   3.039  0.00237 **
## TRT          -0.2719     0.3801  -0.715  0.47444
## prevy        -0.5753     0.1124  -5.121 3.04e-07 ***
## TIME:TRT      0.0916     0.1323   0.692  0.48880
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2459.3  on 2901  degrees of freedom
```

$e^{-0.57} = 0.562$ $\quad 1 - 0.562 = 0.437$

If at time t-1 Y = 1, the odds of observing her outcome at time t are decreased by 43.7% versus if Y = 0 at time t-1, after ...

Test of MCAR vs MAR

```
## Residual deviance: 2417.9  on 2897  degrees of freedom
## AIC: 2427.9
##
## Number of Fisher Scoring iterations: 4
```

Now, let's look at the estimates from the actual model

Valid under MAR

```r
summary(fit)
```

```
## Call:
## wgee(model = Y ~ TIME + TRT + TIME * TRT, data = amenor, id = amenor$ID,
##     family = "binomial", corstr = "exchangeable", scale = NULL,
##     mismodel = R ~ TIME + TRT + TIME * TRT + prevy)
##
##               Estimates  Robust SE z value Pr(>|z|)
## (Intercept) -2.025e+00  1.301e-01 -15.557   <2e-16 ***
## TIME         5.334e-01  4.312e-02  12.373   <2e-16 ***
## TRT          2.781e-01  1.723e-01   1.613    0.107
## TIME:TRT     3.275e-05  5.955e-02   0.001    1.000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Estimated Scale Parameter:  0.9966
##
##  Estimated Correlation:  0.3823
```

TRT = 0.27

Let's compare this with a non-weighted GEE

```r
library(geepack)
summary(geeglm(Y ~ TIME + TRT + TIME*TRT, data = amenor, id = ID,
               family = "binomial", corstr = "exchangeable"))
```

Assumes MCAR

```
##
## Call:
## geeglm(formula = Y ~ TIME + TRT + TIME * TRT, family = "binomial",
##     data = amenor, id = ID, corstr = "exchangeable")
##
##  Coefficients:
##             Estimate  Std.err    Wald Pr(>|W|)
## (Intercept) -2.00929  0.12941 241.061   <2e-16 ***
## TIME         0.51676  0.04303 144.254   <2e-16 ***
## TRT          0.20773  0.17143   1.468    0.226
## TIME:TRT     0.04132  0.05882   0.493    0.482
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)   0.9959 0.02447
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.3637 0.02341
```

0.208

35% diff.

```
## Number of clusters:   1151  Maximum cluster size: 4
```