P8157 HW 4

Sabrina Lin stl2137

11/21/2020

Question 1

```
### Import Data

toenail_dat <- read.delim("/Users/SabrinaLin/Documents/Fall_2020_Longitudinal/HW2/toenail.txt", header =
  janitor::clean_names() %>%
  mutate(
    treatment = as.factor(treatment),
    visit = as.factor(visit)
  )
```

```
## Warning in read.table(file = file, header = header, sep = sep, quote = quote, :
## header and 'col.names' are of different lengths
```

```
#making into data.table

toenail_tab_dat <- data.table(toenail_dat)
```

Part 1

Consider a first order transition model for the log odds of moderate or severe onycholysis. Set up a suitable model assuming linear trends. Use month as the time variable.

```
### add response at lag 1
toenail_tab_dat[, y_1 := shift(y, n = 1, type = "lag", fill = NA), by = "id"]

### transition probabilities
tab1 <- table(toenail_tab_dat$y, toenail_tab_dat$y_1)
round(prop.table(tab1,margin = 1),2)
```

```
##
##        0    1
##   0 0.91 0.09
##   1 0.09 0.91
```

```
### model w/ interaction term

toenail_mod_lag_1 <- gee(y ~ treatment*month + treatment*y_1, corstr = "independence", family = binomial
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
## running glm to get initial regression estimate
```

```
##      (Intercept)       treatment1            month              y_1
##      -2.91754387      -0.58731398      -0.09042707       4.20408170
## treatment1:month   treatment1:y_1
##      -0.05921630       0.69205101
```

```
round(summary(toenail_mod_lag_1)$coeff,2)
```

```
##                  Estimate Naive S.E. Naive z Robust S.E. Robust z
## (Intercept)         -2.92       0.32   -9.05        0.30    -9.58
## treatment1          -0.59       0.53   -1.10        0.48    -1.21
## month               -0.09       0.04   -2.24        0.04    -2.31
## y_1                  4.20       0.31   13.40        0.33    12.57
## treatment1:month    -0.06       0.07   -0.91        0.07    -0.81
## treatment1:y_1       0.69       0.52    1.33        0.49     1.40
```

- Since the interaction term between treatment and month is insignificant with a naive z-score of -0.91 and a robust z-score of -0.81, it will be taken out of the model.
- Since the interaction term between treatment and lag 1 (represented by `y_1`) is insignificant with a naive z-score of 1.33 and a robust z-score of 1.40, it will be taken out of the model.

```
### Model w/o interaction term

toenail_mod_lag_1b <- gee(y ~ treatment + month + y_1, corstr = "independence", family = binomial("logi
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
## running glm to get initial regression estimate
```

```
## (Intercept)  treatment1       month         y_1
##  -3.0094361  -0.3090397  -0.1152287   4.4906918
```

```
toenail_mod_summary <- round(summary(toenail_mod_lag_1b)$coeff, 2)
toenail_mod_summary
```

```
##             Estimate Naive S.E. Naive z Robust S.E. Robust z
## (Intercept)    -3.01       0.27  -11.24        0.25   -12.01
## treatment1     -0.31       0.21   -1.46        0.18    -1.74
## month          -0.12       0.03   -3.73        0.03    -3.41
## y_1             4.49       0.24   18.64        0.25    18.26
```

Part 2

Repeat the model using a second order transition model. Is there a justification for a second order transition model?

```
### add response at lag 2
toenail_tab_dat[, y_2 := shift(y, n = 2, type = "lag", fill = NA), by = "id"]

### transition probabilities
tab2 <- table(toenail_tab_dat$y, toenail_tab_dat$y_2)
round(prop.table(tab2, margin = 1), 2)
```

```
##
##        0    1
##   0 0.83 0.17
##   1 0.16 0.84
```

```
### model w/ interaction term

toenail_mod_lag_2 <- gee(y ~ treatment + month + y_1 + treatment*y_2, corstr = "independence", family =
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27

## running glm to get initial regression estimate

##    (Intercept)     treatment1          month            y_1            y_2
##    -3.06249558    -0.39979024    -0.08284208     3.99128908     0.21405909
## treatment1:y_2
##     0.13003179
```

```
round(summary(toenail_mod_lag_2)$coeff,2)
```

```
##                Estimate Naive S.E. Naive z Robust S.E. Robust z
## (Intercept)       -3.06       0.34   -8.99        0.34    -9.04
## treatment1        -0.40       0.41   -0.97        0.39    -1.02
## month             -0.08       0.03   -2.52        0.04    -2.23
## y_1                3.99       0.40   10.09        0.38    10.57
## y_2                0.21       0.44    0.49        0.39     0.55
## treatment1:y_2     0.13       0.49    0.27        0.48     0.27
```

- It looks like you do not need the second order transition model, as the naive z-score and the robust z-score for lag 2 are respectively 0.49 and 0.55. Thus the model from here on now will only include the first lag.

Part 3

Provide Interpretations for the parameters in your model.

```
toenail_mod_summary
```

```
##             Estimate Naive S.E. Naive z Robust S.E. Robust z
## (Intercept)    -3.01       0.27  -11.24        0.25   -12.01
## treatment1     -0.31       0.21   -1.46        0.18    -1.74
## month          -0.12       0.03   -3.73        0.03    -3.41
## y_1             4.49       0.24   18.64        0.25    18.26
```

- -3.01 is the log odds of moderate or severe onycholysis for those who did not receive treatment and did not have moderate or severe onycholysis in the previous month.
- -0.31 is the log odds ratio of moderate or severe onycholysis comparing those with or without treatment who had an identical onycholysis status in the previous month.
- -0.12 is the log odds ratio of moderate or severe onycholysis for every one month increase for those with an identical onycholysis status and treatment.
- 4.49 is the log odds ratio of moderate or severe onycholysis comparing those with and without treatment in the previous month for those who have an identical onycholysis status.

Part 4

How are the interpretations different from the models in HW2 and HW3.

- The interpretations here are different from the models in HW2 and HW3 because the previous months are accounted for on the onycholysis status.

Question 2

```
### importing given code

toenail <- fread("/Users/SabrinaLin/Documents/Fall_2020_Longitudinal/HW2/toenail.txt")
colnames(toenail) <- c("id","response","treatment","month","visit")
toenail2 <- tidyr::complete(toenail, id, visit) %>%
              tidyr::fill(treatment)

toenail2 <- as.data.table(toenail2)
```

Part 1

Perform a complete case analysis considering a GEE model for the log odds of moderate or severe onycholysis. Set up a suitable model assuming linear trends. Use visit as the time variable.

```
# complete case analysis
count <- toenail2[,j = list(n=sum(!is.na(response))), by = "id"]
table(count$n)
```

```
##
##   1   2   3   4   5   6   7
##   5   3   7   6  10  39 224
```

```
count <- count[n==7]
toenail_1 <- toenail2[id %in% count$id]
table(toenail_1$response,useNA = "always")
```

```
##
##    0    1 <NA>
## 1266  302    0
```

```
table(toenail_1$visit,toenail_1$response, useNA = "always")
```

```
##
##          0   1 <NA>
##   1    144  80    0
##   2    152  72    0
##   3    161  63    0
##   4    180  44    0
##   5    207  17    0
##   6    211  13    0
##   7    211  13    0
##   <NA>   0   0    0
```

```
gee1 <- geeglm(response ~ treatment + visit, id = id, data = toenail_1, family = binomial(link = "logit
summary(gee1)
```

```
##
## Call:
## geeglm(formula = response ~ treatment + visit, family = binomial(link = "logit"),
##     data = toenail_1, id = id, corstr = "unstructured")
##
##  Coefficients:
##               Estimate    Std.err   Wald Pr(>|W|)
```

```
## (Intercept)  1.513e+14  3.576e+14  0.179    0.672
## treatment    1.042e+15  1.499e+14 48.344 3.58e-12 ***
## visit       -3.574e+14  6.106e+13 34.256 4.83e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = unstructured
## Estimated Scale Parameters:
##
##              Estimate   Std.err
## (Intercept) 1.232e+15 6.869e+36
##   Link = identity
##
## Estimated Correlation Parameters:
##           Estimate   Std.err
## alpha.1:2  1.76224 9.824e+21
## alpha.1:3  1.59907 8.914e+21
## alpha.1:4  0.24476 1.364e+21
## alpha.1:5  0.14685 8.186e+20
## alpha.1:6  0.04895 2.729e+20
## alpha.1:7  0.03263 1.819e+20
## alpha.2:3  1.71329 9.551e+21
## alpha.2:4  0.26107 1.455e+21
## alpha.2:5  0.16317 9.096e+20
## alpha.2:6  0.06527 3.638e+20
## alpha.2:7  0.03263 1.819e+20
## alpha.3:4  0.31002 1.728e+21
## alpha.3:5  0.14685 8.186e+20
## alpha.3:6  0.04895 2.729e+20
## alpha.3:7  0.03263 1.819e+20
## alpha.4:5  0.24476 1.364e+21
## alpha.4:6  0.14685 8.186e+20
## alpha.4:7  0.13054 7.277e+20
## alpha.5:6  0.17949 1.001e+21
## alpha.5:7  0.14685 8.186e+20
## alpha.6:7  0.16317 9.096e+20
## Number of clusters:   224  Maximum cluster size: 7
```

- Treatment and visit are both significant, with p-values of 3.6e-12 and 4.8e-09 respectively.
- The beta estimates for the intercept, treatment, and visit are unreasonably large.

Part 2

Perform an available case analysis considering a GEE model for the log odds of moderate or severe onycholysis. Set up a suitable model assuming linear trends. Use visit as the time variable.

```
toenail_2 <- toenail2
table(toenail_2$response, useNA = "always")
```

```
##
##    0    1 <NA>
## 1500  408  150
```

```r
table(toenail_2$visit, toenail_2$response, useNA = "always")
```

```
##
##          0   1 <NA>
##   1    185 109    0
##   2    191  97    6
##   3    199  84   11
##   4    214  58   22
##   5    241  22   31
##   6    226  18   50
##   7    244  20   30
##   <NA>   0   0    0
```

```r
gee2 <- geeglm(response ~ treatment + visit, id = id, data = toenail_2, family = binomial(link = "logit
summary(gee2)
```

```
##
## Call:
## geeglm(formula = response ~ treatment + visit, family = binomial(link = "logit"),
##     data = toenail_2, id = id, corstr = "unstructured")
##
##  Coefficients:
##              Estimate   Std.err   Wald Pr(>|W|)
## (Intercept)  3.81e+15  4.11e+14  85.76   <2e-16 ***
## treatment    1.62e+14  1.44e+14   1.26     0.26
## visit       -9.27e+14  8.13e+13 130.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = unstructured
## Estimated Scale Parameters:
##
##             Estimate  Std.err
## (Intercept)    2e+15 2.55e+37
##   Link = identity
##
## Estimated Correlation Parameters:
##           Estimate  Std.err
## alpha.1:2   1.3842 1.77e+22
## alpha.1:3   1.3201 1.67e+22
## alpha.1:4   1.2807 1.63e+22
## alpha.1:5  -0.0247 3.12e+20
## alpha.1:6  -0.0598 7.55e+20
## alpha.1:7  -0.0502 6.35e+20
## alpha.2:3   1.4144 1.79e+22
## alpha.2:4   1.3774 1.75e+22
## alpha.2:5  -0.0247 3.12e+20
## alpha.2:6  -0.0427 5.39e+20
## alpha.2:7  -0.0502 6.35e+20
## alpha.3:4   1.5143 1.93e+22
## alpha.3:5  -0.0329 4.16e+20
## alpha.3:6  -0.0598 7.55e+20
## alpha.3:7  -0.0502 6.35e+20
## alpha.4:5  -0.0247 3.12e+20
```

```
## alpha.4:6  -0.0513 6.47e+20
## alpha.4:7  -0.0502 6.35e+20
## alpha.5:6   0.1196 1.51e+21
## alpha.5:7   0.0903 1.14e+21
## alpha.6:7   0.1003 1.27e+21
## Number of clusters:   294  Maximum cluster size: 7
```

- Treatment is no longer significant, as the p-value is now 0.26. Visit remains significant with a p-value of <2e-16.
- The beta estimates for the intercept, treatment, and visit are unreasonably large.

Part 3

Perform an LOCF analysis considering a GEE model for the log odds of moderate or severe onycholysis. Set up a suitable model assuming linear trends. Use visit as the time variable.

```
toenail_3 <- lapply(unique(toenail2$id), function(z){tidyr::fill(toenail2[id == z], response)})
toenail_3 <- rbindlist((toenail_3))
table(toenail_3$visit, toenail_3$response, useNA = "always")
```

```
##
##          0   1 <NA>
##   1    185 109    0
##   2    195  99    0
##   3    207  87    0
##   4    228  66    0
##   5    261  33    0
##   6    269  25    0
##   7    269  25    0
##   <NA>   0   0    0
```

```
gee3 <- geeglm(response ~ treatment + visit, id = id, data = toenail_3, family = binomial(link = "logit
summary(gee3)
```

```
##
## Call:
## geeglm(formula = response ~ treatment + visit, family = binomial(link = "logit"),
##     data = toenail_3, id = id, corstr = "unstructured")
##
##  Coefficients:
##             Estimate Std.err  Wald Pr(>|W|)
## (Intercept)  -0.2213  0.1922  1.32     0.25
## treatment    -0.1592  0.2383  0.45     0.50
## visit        -0.2903  0.0315 85.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = unstructured
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     0.99   0.139
##   Link = identity
##
## Estimated Correlation Parameters:
```

```
##           Estimate Std.err
## alpha.1:2    0.923  0.1258
## alpha.1:3    0.761  0.1196
## alpha.1:4    0.582  0.1002
## alpha.1:5    0.351  0.0845
## alpha.1:6    0.234  0.0732
## alpha.1:7    0.188  0.0703
## alpha.2:3    0.898  0.1315
## alpha.2:4    0.700  0.1111
## alpha.2:5    0.406  0.0922
## alpha.2:6    0.259  0.0792
## alpha.2:7    0.244  0.0785
## alpha.3:4    0.912  0.1274
## alpha.3:5    0.439  0.0958
## alpha.3:6    0.275  0.0832
## alpha.3:7    0.262  0.0836
## alpha.4:5    0.567  0.1127
## alpha.4:6    0.354  0.0961
## alpha.4:7    0.340  0.0962
## alpha.5:6    0.554  0.1202
## alpha.5:7    0.476  0.1167
## alpha.6:7    0.596  0.1336
## Number of clusters:   294  Maximum cluster size: 7
```

- Treatment is not significant, as the p-value is now 0.5. Visit remains significant with a p-value of <2e-16.
- The beta estimates are now much more reasonable compared to the prior 2 models utilizing complete and available cases.

Part 4

Perform an multiple imputation based analysis considering a GEE model for the log odds of moderate or severe onycholysis. Set up a suitable model assuming linear trends. Use visit as the time variable.

```
# MI
toenail_4 <- toenail2[,-5] # need to take out `month` b/c also has missing values and we're not imputin
pred <- make.predictorMatrix(toenail_4)
pred
```

```
##           id visit response treatment
## id         0     1        1         1
## visit      1     0        1         1
## response   1     1        0         1
## treatment  1     1        1         0
```

```
pred["response", "id"] <- -2
pred
```

```
##           id visit response treatment
## id         0     1        1         1
## visit      1     0        1         1
## response  -2     1        0         1
## treatment  1     1        1         0
```

```
pred <- pred["response",,drop = FALSE]
pred
```

```
##          id visit response treatment
## response -2     1        0         1
```

```
toenail_4$id <- as.integer(toenail_4$id)
imp <- mice(toenail_4, method = "2l.bin", pred = pred, seed = 1234, maxit = 1, m = 5, print = FALSE, bl
table(mice::complete(imp)$response, useNA = "always")
```

```
##
##    0    1 <NA>
## 1639  419    0
```

```
### GEE
implist <- mids2mitml.list(imp)
gee4 <- with(implist, geeglm(response ~ treatment + visit, id=id,family = binomial, corstr = "unstructu
testEstimates(gee4)
```

```
##
## Call:
##
## testEstimates(model = gee4)
##
## Final parameter estimates and inferences obtained from 5 imputed data sets.
##
##                          Estimate            Std.Error              t.value                   df
## (Intercept)  147777690530970.625  402000789275369.188                0.368                6.085
## treatment   -329786962662731.625  815449038785300.875               -0.404                4.153
## visit        -44672696627274.328  112621334008614.047               -0.397                4.488
##
## Unadjusted hypothesis test as appropriate in larger samples.
```

- After imputation, the intercept, treatment, and visit in the model are not significiant with respective p-values of 0.726, 0.706, and 0.710.
- The beta estimates are back to being unreasonably large.

Part 5

Perform an multiple imputation based analysis considering a mixed effects model for the log odds of moderate or severe onycholysis. Set up a suitable model assuming linear trends. Use visit as the time variable.

```
lme1 <- mice::complete(imp, "all") %>%
  purrr::map(lme4::glmer,
             formula = response ~ treatment + as.numeric(visit) + (1 | id),
             family = binomial,
             control = glmerControl(optimizer = "bobyqa",
                                    optCtrl = list(maxfun=2e5))) %>%
  pool()

summary(lme1)
```

```
##               term estimate std.error statistic   df p.value
## 1      (Intercept)   -0.374    0.4650    -0.804 1311   0.422
## 2        treatment   -0.583    0.5503    -1.059  894   0.290
```

```
## 3 as.numeric(visit)   -0.839    0.0654   -12.822  186   0.000
```

- The intercept and treatment are insignificant with respective p-values of 0.422 and 0.290, but visit is significant with a p-value of 0.000.
- The beta estimates are more reasonable in this mixed effects model compared to the GEE.