P8157 HW2

Shihui Zhu

Q1.

1. Consider a marginal model for the log odds of moderate or severe onycholysis. Using GEE, set up a suitable model assuming linear trends. Use month as the time variable. Assume "exchangeable" correlation for the association among the repeated binary responses.

```
toenail <- fread("toenail.txt")
colnames(toenail) <- c("id", "response", "treatment", "month", "visit")
toenail$treatment <- as.factor(toenail$treatment)
toenail$id <- as.factor(toenail$id)
```

Explore the proportion of moderate/severe onycholysis with time:

```
summary1 <- toenail[,j=list(prop_severe = mean(response,na.rm=TRUE)*100),
                    by = c("treatment","visit")]
dcast(summary1, treatment ~ visit, value = "prop_severe")
```

```
## Using 'prop_severe' as value column. Use 'value.var' to override
```

```
##    treatment        1        2        3        4         5        6         7
## 1:         0 36.98630 34.75177 31.88406 21.96970 10.769231 8.547009 10.526316
## 2:         1 37.16216 32.65306 27.58621 20.71429  6.015038 6.299213  4.580153
```

There seems to be a linear trend for both treatments with increasing in time. Group with oral treatment A seems to have a lower raet of moderate/severe onycholysis.

Set up GEEE model with linear trend with months and interaction with treatments:

```
gee1 <- geeglm(response ~ treatment*month, id = id, data = toenail,
               family = binomial(link = "logit"), corstr = "exchangeable")
summary(gee1)
```

```
##
## Call:
## geeglm(formula = response ~ treatment * month, family = binomial(link = "logit"),
##     data = toenail, id = id, corstr = "exchangeable")
##
##  Coefficients:
##                  Estimate  Std.err   Wald Pr(>|W|)
## (Intercept)      -0.58192  0.17206 11.439 0.000719 ***
## treatment1        0.00718  0.25949  0.001 0.977924
## month            -0.17128  0.03000 32.596 1.13e-08 ***
```

```
## treatment1:month -0.07773  0.05411  2.064 0.150862
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)    1.088  0.5013
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.4218  0.2119
## Number of clusters:   294  Maximum cluster size: 7
```

We see that the coeffciients for treatment and month interaction is not significant. We then test if the treatment and months interaction is necessary:

```
L <- matrix(0,ncol=4,nrow=1) # ncol = number of coeffcients in the model, nrow = number of tests
L[1,c(4)]  <- c(1)
L
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    0    0    0    1
```

```
esticon(gee1,L=L,joint.test = TRUE)
```

```
##   X2.stat DF Pr(>|X^2|)
## 1   2.064  1     0.1509
```

We have p-value of $0.151 > 0.05$, therefore we conclude that the profiles of change do not differ between treatment A and B.

Set up GEEE model with linear trend (without month and treatment interaction):

```
gee2 <- geeglm(response ~ treatment + month, id = id, data = toenail,
               family = binomial(link = "logit"), corstr = "exchangeable")
summary(gee2)
```

```
##
## Call:
## geeglm(formula = response ~ treatment + month, family = binomial(link = "logit"),
##     data = toenail, id = id, corstr = "exchangeable")
##
##  Coefficients:
##             Estimate Std.err  Wald Pr(>|W|)
## (Intercept)  -0.6104  0.1777 11.80  0.00059 ***
## treatment1    0.0402  0.2532  0.03  0.87388
## month        -0.2051  0.0259 62.66  2.4e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     1.09   0.423
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.424   0.182
## Number of clusters:   294  Maximum cluster size: 7
```

2. Provide Interpretations for the coefficients in your model.

Let $\boldsymbol{y}_i$ denote the vector of onycholysis outcome for subject id i, and $E(\boldsymbol{y}_i) = \boldsymbol{\mu}_i$.

The model is therefore $logit(\boldsymbol{\mu}_i) = -0.6104 - 0.2051\boldsymbol{month}_i + 0.0402(\boldsymbol{treatment = oral\ treatment\ A})_i$. However, the coefficient for treatment group A is not significant.

The covariance matrix is $\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2}\boldsymbol{R}_i\boldsymbol{A}_i^{1/2} \cdot \phi$, where $\phi = 1.09$ is the dispersion parameter, and $\alpha = 0.422$ for the working correlation matrix. Therefore, $R(\alpha)_i$ is a working correlation matrix of

$$R(\alpha)_i = \begin{bmatrix} 1 & 0.424 & ... & 0.424 \\ 0.424 & 1 & ... & 0.424 \\ ... & & & \\ ... & 0.424 & ... & 1 \end{bmatrix}_{7\times 7}$$

```
se.txtB <- summary(gee1)$coefficients["(Intercept)","Std.err"]
se.txtA <- summary(gee1)$coefficients["treatment1","Std.err"]
se.time <- summary(gee1)$coefficients["month","Std.err"]
txtB <- exp(coef(gee1)["(Intercept)"] + c(0, -1, 1) * se.txtB * qnorm(0.975))
txtA <- exp(coef(gee1)["treatment1"] + c(0, -1, 1) * se.txtA * qnorm(0.975))
time <- exp(coef(gee1)["month"] + c(0, -1, 1) * se.time * qnorm(0.975))

out.logit <- rbind(txtB, txtA, time)
colnames(out.logit)=c('Estimate of OR','95% CI lower','95% CI upper')
out.logit %>% knitr::kable(digits = 3)
```

	Estimate of OR	95% CI lower	95% CI upper
txtB	0.559	0.399	0.783
txtA	1.007	0.606	1.675
time	0.843	0.794	0.894

- $\beta_{i0} = -0.6104$: Holding the month post randomization at constant, the odds of having a moderate/severe onycholysis with the oral treatments B is about 0.559 (95% CI [0.399,0.783]) times the corresponding odds from the oral treatments A.
- $\beta_{i1} = 0.0402$: Holding the month post randomization at constant, the odds of having a moderate/severe onycholysis with the oral treatments A is about 1.007 (95% CI [0.606,1.675]) times the corresponding odds from the oral treatments B.
- $\beta_{i2} = -0.2051$: Holding the treatment group at constant, the odds of having a moderate/severe onycholysis in the current month is 0.843 (95% CI [0.794,0.894]) of the previous month.

3. From the results of your analysis what conclusions do you draw about the effect of treatment on changes in the severity of onycholysis over time? Provide results that support your conclusions.

From the above result, we see that the profiles of change do not differ between treatment A and B over time i.e. no interaction between time and treatments. Group treated by oral treatment B tends to have a lower odds of getting moderate/severe onycholysis. And as the treatment time increases, the severity of onycholysis decreases among both treatment groups.

4. Try Different correlation structures. Is the analysis and inference sensitive to this choice?

Set up GEE with AR(1) correlation:

```
gee3 <- geeglm(response ~ treatment + month, id = id, data = toenail,
               family = binomial(link = "logit"), corstr = "ar1")
summary(gee3)
```

```
##
## Call:
## geeglm(formula = response ~ treatment + month, family = binomial(link = "logit"),
##     data = toenail, id = id, corstr = "ar1")
##
##  Coefficients:
##             Estimate Std.err  Wald Pr(>|W|)
## (Intercept)  -0.5725  0.1596 12.86  0.00033 ***
## treatment1   -0.0989  0.2156  0.21  0.64638
## month        -0.1778  0.0241 54.57  1.5e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = ar1
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)    0.987   0.224
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha     0.69  0.0947
## Number of clusters:   294  Maximum cluster size: 7
```

Set up GEE with unstructured correlation:

```
gee4 <- geeglm(response ~ treatment + month, id = id, data = toenail,
               family = binomial(link = "logit"), corstr = "unstructured")
summary(gee4)
```

```
##
## Call:
## geeglm(formula = response ~ treatment + month, family = binomial(link = "logit"),
##     data = toenail, id = id, corstr = "unstructured")
##
```

```
##  Coefficients:
##             Estimate Std.err  Wald Pr(>|W|)
## (Intercept)  -0.6458  0.1587 16.56  4.7e-05 ***
## treatment1   -0.1429  0.2157  0.44     0.51
## month        -0.1705  0.0227 56.25  6.4e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = unstructured
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     1.02    0.22
##   Link = identity
##
## Estimated Correlation Parameters:
##           Estimate Std.err
## alpha.1:2    0.923  0.2062
## alpha.1:3    0.724  0.1728
## alpha.1:4    0.521  0.1422
## alpha.1:5    0.259  0.0997
## alpha.1:6    0.156  0.0832
## alpha.1:7    0.136  0.0853
## alpha.2:3    0.837  0.1933
## alpha.2:4    0.618  0.1614
## alpha.2:5    0.286  0.1057
## alpha.2:6    0.249  0.1001
## alpha.2:7    0.163  0.0916
## alpha.3:4    0.794  0.1946
## alpha.3:5    0.299  0.1076
## alpha.3:6    0.216  0.0966
## alpha.3:7    0.191  0.0986
## alpha.4:5    0.391  0.1258
## alpha.4:6    0.286  0.1091
## alpha.4:7    0.248  0.1100
## alpha.5:6    0.488  0.1532
## alpha.5:7    0.440  0.1560
## alpha.6:7    0.616  0.1951
## Number of clusters:   294  Maximum cluster size: 7
```

The α values for both correlation matrix are quite different from 0.424. Therefore analysis and inference is sensitive to the choice.

Q2

1. Set up a suitable GEE model for rate of skin cancers with Treatment and Year as covariates.

```r
skin <- fread("skin.txt")
colnames(skin) <- c("id", "center", "age", "skin", "gender", "exposure", "y", "treatment", "year")
skin$year <- as.numeric(skin$year)
skin$treatment <- as.factor(skin$treatment)
skin$gender <- as.factor(skin$gender)
```

```
skin$skin <- as.factor(skin$skin)
skin$id <- as.factor(skin$id)
```

Explore the trend of year v.s. outcome:

```
itp <- interaction(skin$treatment, skin$year)
tapply(skin$y, itp, mean) # crude check of group mean by year*treatment
```

```
##   0.1   1.1   0.2   1.2   0.3   1.3   0.4   1.4   0.5   1.5
## 0.271 0.298 0.240 0.261 0.247 0.286 0.233 0.315 0.272 0.298
```

We see there is a quadratic relationship between year and count of new skin cancers for both treatment groups.

Set up GEE with Treatment and Year as covariates (quadratic trend with year, interaction with treatment), with exchangeable correlation:

```
gee5 <- geeglm(y ~ treatment * (year + I(year^2)), data = skin,
               family = "poisson", id = id, corstr = "exchangeable")
summary(gee5)
```

```
##
## Call:
## geeglm(formula = y ~ treatment * (year + I(year^2)), family = "poisson",
##     data = skin, id = id, corstr = "exchangeable")
##
##  Coefficients:
##                      Estimate Std.err  Wald Pr(>|W|)
## (Intercept)           -1.1466  0.1962 34.15  5.1e-09 ***
## treatment1            -0.0369  0.2922  0.02     0.90
## year                  -0.1978  0.1376  2.07     0.15
## I(year^2)              0.0347  0.0229  2.30     0.13
## treatment1:year        0.1212  0.2268  0.29     0.59
## treatment1:I(year^2)  -0.0156  0.0373  0.18     0.68
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     2.64   0.359
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.378   0.111
## Number of clusters:   1683  Maximum cluster size: 5
```

Check if the interaction term is necessary:

```
L <- matrix(0,ncol=6,nrow=2) # ncol = number of coeffcients in the model, nrow = number of tests
L[1,c(5)]  <- c(1)
L[2,c(6)]  <- c(1)
L
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    0    0    0    0    1    0
## [2,]    0    0    0    0    0    1
```

```
esticon(gee5,L=L,joint.test = TRUE)
```

```
##   X2.stat DF Pr(>|X^2|)
## 1   0.552  2      0.759
```

We got a p-value of $0.759 > 0.05$, therefore the interaction is not necessary. Then we set up the GEE with Treatment and Year as covariates (quadratic trend with year), with exchangeable correlation:

```
gee6 <- geeglm(y ~ treatment + year + I(year^2), data = skin,
               family = "poisson", id = id, corstr = "exchangeable")
summary(gee6)
```

```
##
## Call:
## geeglm(formula = y ~ treatment + year + I(year^2), family = "poisson",
##     data = skin, id = id, corstr = "exchangeable")
##
##  Coefficients:
##             Estimate Std.err  Wald Pr(>|W|)
## (Intercept)  -1.2456  0.1738 51.37  7.6e-13 ***
## treatment1    0.1469  0.1089  1.82     0.18
## year         -0.1322  0.1165  1.29     0.26
## I(year^2)     0.0262  0.0191  1.89     0.17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     2.65   0.375
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.377   0.111
## Number of clusters:   1683  Maximum cluster size: 5
```

2. Provide Interpretations for the coefficients in your model.

Let $\boldsymbol{y}_i$ denote the vector of onycholysis outcome for subject id i, and $E(\boldsymbol{y}_i) = \boldsymbol{\mu}_i$.

The model is therefore $logit(\boldsymbol{\mu}_i) = -1.2456 + 0.1469(\boldsymbol{treatment = beta\ carotene})_i - 0.1322\boldsymbol{year}_i + 0.0262\boldsymbol{year}_i^2$. However, the coefficient for year, year^2 and treatment group beta carotene is not significant.

The covariance matrix is $\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2}\boldsymbol{R}_i\boldsymbol{A}_i^{1/2} \cdot \phi$, where $\phi = 2.65$ is the dispersion parameter, and $\alpha = 0.377$ for the working correlation matrix. Therefore, $R(\alpha)_i$ is a working correlation matrix of

$$R(\alpha)_i = \begin{bmatrix} 1 & 0.377 & ... & 0.377 \\ 0.377 & 1 & ... & 0.377 \\ ... & & & \\ ... & 0.377 & ... & 1 \end{bmatrix}_{5\times 5}$$

```
se.txtP <- summary(gee6)$coefficients["(Intercept)","Std.err"]
se.txt1 <- summary(gee6)$coefficients["treatment1","Std.err"]
se.year <- summary(gee6)$coefficients["year","Std.err"]
se.yearSquare <- summary(gee6)$coefficients["I(year^2)","Std.err"]
txtP <- exp(coef(gee6)["(Intercept)"] + c(0, -1, 1) * se.txtP * qnorm(0.975))
txt1 <- exp(coef(gee6)["treatment1"] + c(0, -1, 1) * se.txt1 * qnorm(0.975))
year <- exp(coef(gee6)["year"] + c(0, -1, 1) * se.year * qnorm(0.975))
yearSquare <- exp(coef(gee6)["I(year^2)"] + c(0, -1, 1) * se.yearSquare * qnorm(0.975))

out.logit <- rbind(txtP, txt1, year, yearSquare)
colnames(out.logit)=c('Estimate of OR','95% CI lower','95% CI upper')
out.logit %>% knitr::kable(digits = 3)
```

	Estimate of OR	95% CI lower	95% CI upper
txtP	0.288	0.205	0.405
txt1	1.158	0.936	1.434
year	0.876	0.697	1.101
yearSquare	1.027	0.989	1.066

- $\beta_{i0} = -1.2456$: Holding the year post randomization at constant, the rate of skin cancers with the placebo treatment is about 0.288 (95% CI [0.205,0.405]) times the corresponding rate from the beta carotene treatment. This is statistically significant.
- $\beta_{i1} = 0.1469$: Holding the year post randomization at constant, the rate of skin cancers with beta carotene treatment is about 1.158 (95% CI [0.936,1.434]) times the corresponding rate from the placebo. This is not statistically significant.
- $\beta_{i2} = -0.1322$: Holding the treatment group at constant, the rate of skin cancers in the current year is 0.876 (95% CI [0.697,1.101]) of the previous year. This is not statistically significant.
- $\beta_{i3} = 0.0262$: Holding the treatment group at constant, the rate of skin cancers in the square of current year is 1.027 (95% CI [0.989,1.066]) of the previous year squared. This is not statistically significant.

3. From the results of your analysis what conclusions do you draw about the effect of beta carotene on the rate of skin cancers? Provide results that support your conclusions.

The beta carotene doesn't prevent the rate of skin cancers from growing in each year because placebo group has a lower rate of new skin cancers compared to the group treated by beta carotene.

4. Repeat the above analysis adjusting for skin type, age, and the count of the number of previous skin cancers. What conclusions do you draw about the effect of beta carotene on the adjusted rate of skin cancers?

```
gee7 <- geeglm(y ~ treatment + skin + exposure + year + I(year^2),
               data = skin, family = "poisson", id = id, corstr = "exchangeable")
summary(gee7)
```

```
##
## Call:
## geeglm(formula = y ~ treatment + skin + exposure + year + I(year^2),
##     family = "poisson", data = skin, id = id, corstr = "exchangeable")
##
##  Coefficients:
##             Estimate Std.err   Wald Pr(>|W|)
## (Intercept)  -1.9377  0.1829 112.27   <2e-16 ***
## treatment1    0.1155  0.0996   1.34     0.25
## skin1         0.1602  0.1104   2.11     0.15
## exposure      0.1404  0.0103 185.07   <2e-16 ***
## year         -0.1167  0.1177   0.98     0.32
## I(year^2)     0.0234  0.0193   1.47     0.22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     1.64  0.0762
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.215  0.0275
## Number of clusters:   1683  Maximum cluster size: 5
```

The more previous exposures are, the larger of the rate of skin cancers. And the effect of beta carotene on the adjusted rate of skin cancers is still negative i.e. group treated with beta carotene has larger adjusted rate of skin cancer.

5. Try Different correlation structures. Is the analysis and inference sensitive to this choice?

Try GEE with unstructured correlation:

```
gee8 <- geeglm(y ~ treatment + skin + exposure + year + I(year^2),
               data = skin, family = "poisson", id = id, corstr = "unstructured")
summary(gee8)
```

```
##
## Call:
## geeglm(formula = y ~ treatment + skin + exposure + year + I(year^2),
##     family = "poisson", data = skin, id = id, corstr = "unstructured")
##
##  Coefficients:
##             Estimate  Std.err   Wald Pr(>|W|)
## (Intercept) -1.94415  0.18133 114.95   <2e-16 ***
```

```
## treatment1   0.10719  0.09786   1.20    0.273
## skin1        0.18203  0.10784   2.85    0.091 .
## exposure     0.13942  0.00998 195.34   <2e-16 ***
## year        -0.11134  0.11896   0.88    0.349
## I(year^2)    0.02201  0.01955   1.27    0.260
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = unstructured
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     1.64  0.0759
##   Link = identity
##
## Estimated Correlation Parameters:
##           Estimate Std.err
## alpha.1:2    0.177  0.0375
## alpha.1:3    0.187  0.0396
## alpha.1:4    0.214  0.0611
## alpha.1:5    0.183  0.0488
## alpha.2:3    0.193  0.0445
## alpha.2:4    0.190  0.0463
## alpha.2:5    0.155  0.0418
## alpha.3:4    0.332  0.0821
## alpha.3:5    0.292  0.0687
## alpha.4:5    0.248  0.0691
## Number of clusters:   1683  Maximum cluster size: 5
```

Try GEE with AR(1):

```
gee9 <- geeglm(y ~ treatment + skin + exposure + year + I(year^2),
               data = skin, family = "poisson", id = id, corstr = "ar1")
summary(gee9)
```

```
##
## Call:
## geeglm(formula = y ~ treatment + skin + exposure + year + I(year^2),
##     family = "poisson", data = skin, id = id, corstr = "ar1")
##
##  Coefficients:
##             Estimate Std.err   Wald Pr(>|W|)
## (Intercept)  -1.9195  0.1793 114.63   <2e-16 ***
## treatment1    0.1201  0.1011   1.41     0.23
## skin1         0.1501  0.1121   1.79     0.18
## exposure      0.1405  0.0105 180.45   <2e-16 ***
## year         -0.1234  0.1172   1.11     0.29
## I(year^2)     0.0233  0.0195   1.43     0.23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = ar1
## Estimated Scale Parameters:
```

```
##
##             Estimate Std.err
## (Intercept)     1.64  0.0762
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.304  0.0336
## Number of clusters:   1683  Maximum cluster size: 5
```

We observe that the value for dispersion parameter and α does not differ much with different correlation structures. Therefore we conclude that the analysis and inference are insensitive to this choice.

6. Do you need to account for overdisperion. Comment.

Yes. We have $\phi = 1.64 > 1$, therefore we need to account for overdispersion.