P8157 HW2 yz4184

Yunlin Zhou

2022-10-17

```
# import dataset for question 1
toenail <- fread("toenail.txt")
colnames(toenail) <- c("id", "y", "treatment", "month", "visit")
toenail$id <- as.factor(toenail$id)
toenail$treatment <- as.factor(toenail$treatment)

# import dataset for question 2
skin <- fread("skin.txt")
colnames(skin) <- c("id","center","age","skin","gender","exposure", "y", "treatment", "year")
skin$id <- as.factor(skin$id)
skin$treatment <- as.factor(skin$treatment)
skin$gender <- as.factor(skin$gender)
skin$skin <- as.factor(skin$skin)
```

Question 1

1.

First, set a model with month effect and treatment interaction.

```
gee1 <- geeglm(y ~ treatment * (month + I(month^2)), id = id, data = toenail, family = binomial(link =
summary(gee1)
```

```
##
## Call:
## geeglm(formula = y ~ treatment * (month + I(month^2)), family = binomial(link = "logit"),
##     data = toenail, id = id, corstr = "exchangeable")
##
##  Coefficients:
##                        Estimate   Std.err   Wald Pr(>|W|)
## (Intercept)           -0.378812  0.176363  4.614  0.03172 *
## treatment1            -0.053047  0.251016  0.045  0.83263
## month                 -0.308201  0.053739 32.892 9.74e-09 ***
## I(month^2)             0.012364  0.004076  9.202  0.00242 **
## treatment1:month      -0.029879  0.081520  0.134  0.71398
## treatment1:I(month^2) -0.003161  0.006998  0.204  0.65145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)   0.9988  0.2733
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.4391  0.1405
## Number of clusters:   294  Maximum cluster size: 7
```

Then test if treatment interaction term is required.

```
L <- matrix(0,ncol=6,nrow=2)
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(gee1,L=L,joint.test = TRUE)
```

```
##   X2.stat DF Pr(>|X^2|)
## 1   1.885  2     0.3896
```

As shown above, the p-value is 0.39. We fail to reject the null hypothesis at 5% level of significance.The treatment interaction term is not significantly associated with outcome.

Finally, we build up a model without treatment interaction.

```
gee2 <- geeglm(y ~ treatment + (month + I(month^2)), id = id, data = toenail, family = binomial(link =
summary(gee2)
```

```
##
## Call:
## geeglm(formula = y ~ treatment + (month + I(month^2)), family = binomial(link = "logit"),
##     data = toenail, id = id, corstr = "exchangeable")
##
##  Coefficients:
##             Estimate  Std.err  Wald Pr(>|W|)
## (Intercept) -0.39889  0.17545  5.17  0.02300 *
## treatment1  -0.00653  0.25168  0.00  0.97929
## month       -0.32603  0.04039 65.17  6.7e-16 ***
## I(month^2)   0.01151  0.00326 12.43  0.00042 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
```

```
##
##             Estimate Std.err
## (Intercept)    0.992   0.205
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.442   0.113
## Number of clusters:   294  Maximum cluster size: 7
```

Since the P-values of month term and month^2 are smaller than 0.05, we conclude that we need the month term and month^2 terms. The final model is gee2.

2.

- beta0 = -0.39889

beta0 is the baseline log odds ratio between having moderate or severe onycholysis in population, holding all other variables constant.

- beta1 = -0.00653

Treatment is not a significant predictor.

For those subjects receiving treatment A, expected log odds ratio of having severe onycholysis in population decreases by a factor of -0.00653.

- beta2 = -0.32603

Month is a significant predictor (p-value < 0.001).

With each unit of increase in month, expected log odds ratio of having severe onycholysis in population decreases by a factor of -0.32603.

- beta3 = 0.01151

Month^2 is a significant predictor (p-value < 0.001).

With each unit of increase in month^2, expected log odds ratio of having severe onycholysis in population increases by a factor of 0.01151.

3.

As we can see from gee2 model, the coefficient of treatment (beta1) is negative but not significant (p-value = 0.97929).The coefficients of month (beta2 and beta3) are significant.

We can conclude that the treatment 1 might have negative effect on onycholysis but the effect is not significant. However, as time goes by, the severity of onycholysis might be affected.

4.

```r
gee3 <- geeglm(y ~ treatment + (month + I(month^2)), id = id, data = toenail, family = binomial(link =
summary(gee3)
```

```
##
## Call:
## geeglm(formula = y ~ treatment + (month + I(month^2)), family = binomial(link = "logit"),
##     data = toenail, id = id, corstr = "unstructured")
##
##  Coefficients:
##              Estimate   Std.err   Wald Pr(>|W|)
## (Intercept) -1.53e+16  2.88e+14 2801.0  < 2e-16 ***
## treatment1  -1.25e+15  1.66e+14   56.3  6.2e-14 ***
## month        2.86e+15  8.11e+13 1244.9  < 2e-16 ***
## I(month^2)  -1.29e+14  5.90e+12  476.5  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = unstructured
## Estimated Scale Parameters:
##
##             Estimate  Std.err
## (Intercept) 1.38e+15 1.72e+37
##   Link = identity
##
## Estimated Correlation Parameters:
##           Estimate  Std.err
## alpha.1:2   1.0532 1.31e+22
## alpha.1:3   0.8468 1.06e+22
## alpha.1:4   0.5982 7.56e+21
## alpha.1:5   0.1918 2.39e+21
## alpha.1:6  -0.3609 4.49e+21
## alpha.1:7  -0.3653 4.56e+21
## alpha.2:3   0.8697 1.09e+22
## alpha.2:4   0.6217 7.85e+21
## alpha.2:5   0.2038 2.54e+21
## alpha.2:6  -0.3111 3.87e+21
## alpha.2:7  -0.3360 4.19e+21
## alpha.3:4   0.6804 8.58e+21
## alpha.3:5   0.1798 2.24e+21
## alpha.3:6  -0.2738 3.40e+21
## alpha.3:7  -0.2484 3.10e+21
## alpha.4:5   0.2038 2.54e+21
## alpha.4:6  -0.1742 2.17e+21
## alpha.4:7  -0.1607 2.01e+21
## alpha.5:6   0.0498 6.19e+20
## alpha.5:7  -0.0146 1.82e+20
## alpha.6:7   1.1834 1.48e+22
## Number of clusters:   294  Maximum cluster size: 7
```

The result of unstructured correlation structure is different from that using exchangeable correlation structure. In this model we can see that every coefficient is significant, but they are also very small.

```r
gee4 <- geeglm(y ~ treatment + (month + I(month^2)), id = id, data = toenail, family = binomial(link =
summary(gee4)
```

```
##
## Call:
## geeglm(formula = y ~ treatment + (month + I(month^2)), family = binomial(link = "logit"),
##     data = toenail, id = id, corstr = "ar1")
##
##  Coefficients:
##             Estimate  Std.err  Wald Pr(>|W|)
## (Intercept) -0.41343  0.16234  6.49    0.011 *
## treatment1  -0.12275  0.21801  0.32    0.573
## month       -0.32645  0.04054 64.85  7.8e-16 ***
## I(month^2)   0.01321  0.00312 17.94  2.3e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = ar1
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)    0.975   0.145
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.699  0.0703
## Number of clusters:   294  Maximum cluster size: 7
```

The result of ar1 correlation structure is similar to that using exchangeable correlation structure.

Question 2

1.

First, set a model with year effect and treatment interaction.

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

```
##
## Call:
## geeglm(formula = y ~ treatment * (year + I(year^2)), family = poisson(link = "log"),
##     data = skin, id = id, corstr = "unstructured")
##
##  Coefficients:
##                        Estimate Std.err  Wald Pr(>|W|)
## (Intercept)             -1.1590  0.1968 34.67  3.9e-09 ***
## treatment1              -0.0129  0.2939  0.00     0.96
## year                    -0.1755  0.1406  1.56     0.21
```

```
## I(year^2)               0.0288  0.0239  1.45     0.23
## treatment1:year         0.0847  0.2308  0.13     0.71
## treatment1:I(year^2)  -0.0086  0.0389  0.05     0.83
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = unstructured
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     2.68   0.387
##   Link = identity
##
## Estimated Correlation Parameters:
##           Estimate Std.err
## alpha.1:2    0.289  0.0842
## alpha.1:3    0.327  0.1115
## alpha.1:4    0.360  0.1258
## alpha.1:5    0.394  0.2113
## alpha.2:3    0.251  0.0598
## alpha.2:4    0.237  0.0661
## alpha.2:5    0.237  0.1086
## alpha.3:4    0.762  0.4120
## alpha.3:5    0.514  0.2039
## alpha.4:5    0.498  0.2247
## Number of clusters:   1683  Maximum cluster size: 5
```

Then test if treatment interaction term is required.

```
L2 <- matrix(0,ncol=6,nrow=2)
L2[1,c(5)]  <- c(1)
L2[2,c(6)]  <- c(1)
L2
```

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

```
esticon(gee5,L=L2,joint.test = TRUE)
```

```
##   X2.stat DF Pr(>|X^2|)
## 1   0.575  2       0.75
```

As shown above, the p-value is 0.75. We fail to reject the null hypothesis at 5% level of significance.The treatment interaction term is not significantly associated with outcome.

Finally, we build up a model without treatment interaction.

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

```
##
```

```
## Call:
## geeglm(formula = y ~ treatment + (year + I(year^2)), family = poisson(link = "log"),
##     data = skin, id = id, corstr = "unstructured")
##
##  Coefficients:
##             Estimate Std.err  Wald Pr(>|W|)
## (Intercept)  -1.2346  0.1705 52.41  4.5e-13 ***
## treatment1    0.1284  0.1048  1.50     0.22
## year         -0.1301  0.1179  1.22     0.27
## I(year^2)     0.0241  0.0198  1.48     0.22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = unstructured
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     2.69   0.402
##   Link = identity
##
## Estimated Correlation Parameters:
##           Estimate Std.err
## alpha.1:2    0.291  0.0852
## alpha.1:3    0.327  0.1120
## alpha.1:4    0.359  0.1256
## alpha.1:5    0.393  0.2106
## alpha.2:3    0.250  0.0596
## alpha.2:4    0.235  0.0652
## alpha.2:5    0.234  0.1065
## alpha.3:4    0.766  0.4218
## alpha.3:5    0.510  0.2039
## alpha.4:5    0.495  0.2262
## Number of clusters:   1683  Maximum cluster size: 5
```

2.

- beta0 = -1.2346

beta0 is the baseline log rate ratio for the count of the number of new skin cancers in population, holding all other variables constant.

- beta1 = 0.1284

Treatment is not a significant predictor.

On average, the count of the number of new skin cancers per year for the patients receiving beta carotene is 0.1284 times the number for the patients receiving placebo, holding all other variables constant.

- beta2 = -0.1301

Year is not a significant predictor.

On average, one unit increase in the year is associated with 0.1301 decrease in the number of new skin cancers, holding all other variables constant.

- beta3 = 0.0241

Year^2 is not a significant predictor.

On average, one unit increase in the year^2 is associated with 0.0241 increase in the number of new skin cancers, holding all other variables constant.

3.

As we can see from gee6 model, the coefficient of treatment (beta1) is positive but not significant (p-value = 0.22).The coefficients of year (beta2 and beta3) are also not significant.

We can conclude that beta carotene has positive effect on the rate of skin cancers, but the effect is not significant. Also, the time doesn't have significant effect on the rate of skin cancers.

4.

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

```
##
## Call:
## geeglm(formula = y ~ treatment + year + I(year^2) + skin + age +
##     exposure, family = poisson(link = "log"), data = skin, id = id,
##     corstr = "unstructured")
##
##  Coefficients:
##             Estimate  Std.err   Wald Pr(>|W|)
## (Intercept) -2.92093  0.34330  72.39   <2e-16 ***
## treatment1   0.11518  0.09755   1.39   0.2377
## year        -0.10969  0.11971   0.84   0.3595
## I(year^2)    0.02182  0.01969   1.23   0.2678
## skin1        0.18671  0.10789   2.99   0.0835 .
## age          0.01523  0.00511   8.89   0.0029 **
## exposure     0.13782  0.01012 185.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = unstructured
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     1.64  0.0772
##   Link = identity
##
## Estimated Correlation Parameters:
##           Estimate Std.err
## alpha.1:2    0.164  0.0351
## alpha.1:3    0.178  0.0367
## alpha.1:4    0.197  0.0564
## alpha.1:5    0.178  0.0479
## alpha.2:3    0.202  0.0490
```

```
## alpha.2:4    0.184  0.0445
## alpha.2:5    0.149  0.0445
## alpha.3:4    0.325  0.0893
## alpha.3:5    0.310  0.0764
## alpha.4:5    0.240  0.0669
## Number of clusters:   1683  Maximum cluster size: 5
```

After adjusting for skin type, age, and the count of the number of previous skin cancers, the coefficient of treatment (beta1) is still positive and not significant (p-value = 0.2377).The coefficients of age and exposure are significant.

So we conclude that the effect of beta carotene on the adjusted rate of skin cancers didn't change much.

5.

```r
gee8 <- geeglm(y ~ treatment + year + I(year^2) + skin + age + exposure , id = id, data = skin, family =
summary(gee8)
```

```
##
## Call:
## geeglm(formula = y ~ treatment + year + I(year^2) + skin + age +
##     exposure, family = poisson(link = "log"), data = skin, id = id,
##     corstr = "ar1")
##
##  Coefficients:
##             Estimate  Std.err   Wald Pr(>|W|)
## (Intercept) -2.87671  0.34459  69.69   <2e-16 ***
## treatment1   0.12870  0.10071   1.63   0.2013
## year        -0.12477  0.11741   1.13   0.2879
## I(year^2)    0.02374  0.01952   1.48   0.2238
## skin1        0.15413  0.11222   1.89   0.1696
## age          0.01495  0.00511   8.56   0.0034 **
## exposure     0.13903  0.01062 171.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = ar1
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     1.64  0.0775
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.297  0.0329
## Number of clusters:   1683  Maximum cluster size: 5
```

```r
gee9 <- geeglm(y ~ treatment + year + I(year^2) + skin + age + exposure , id = id, data = skin, family =
summary(gee9)
```

```
##
## Call:
## geeglm(formula = y ~ treatment + year + I(year^2) + skin + age +
##     exposure, family = poisson(link = "log"), data = skin, id = id,
##     corstr = "exchangeable")
##
##  Coefficients:
##             Estimate  Std.err   Wald Pr(>|W|)
## (Intercept) -2.89301  0.34531  70.19   <2e-16 ***
## treatment1   0.12340  0.09922   1.55   0.2136
## year        -0.11740  0.11791   0.99   0.3194
## I(year^2)    0.02365  0.01930   1.50   0.2204
## skin1        0.16384  0.11055   2.20   0.1383
## age          0.01494  0.00523   8.15   0.0043 **
## exposure     0.13883  0.01051 174.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     1.64  0.0766
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.209  0.0261
## Number of clusters:   1683  Maximum cluster size: 5
```

The result of ar1 and exchangeable correlation structures are similar to that using unstructured correlation structure.

6.

```
# estimate over-dispersion parameter
res = residuals(gee7, type = "pearson")
G1=sum(res^2)
phi=G1/(gee7$df.residual)
phi
```

```
## [1] 1.64
```

we are certain that over dispersion exists since the over-dispersion parameter is estimated to be 1.64, which is larger than 1.

The model after adjusting for covariates has almost the same coefficient as the original model.

```
# fit model with constant over-dispersion
summary(gee7,dispersion=phi)
```

```
##
## Call:
## geeglm(formula = y ~ treatment + year + I(year^2) + skin + age +
##     exposure, family = poisson(link = "log"), data = skin, id = id,
##     corstr = "unstructured")
##
##  Coefficients:
##             Estimate  Std.err   Wald Pr(>|W|)
## (Intercept) -2.92093  0.34330  72.39   <2e-16 ***
## treatment1   0.11518  0.09755   1.39   0.2377
## year        -0.10969  0.11971   0.84   0.3595
## I(year^2)    0.02182  0.01969   1.23   0.2678
## skin1        0.18671  0.10789   2.99   0.0835 .
## age          0.01523  0.00511   8.89   0.0029 **
## exposure     0.13782  0.01012 185.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = unstructured
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     1.64  0.0772
##   Link = identity
##
## Estimated Correlation Parameters:
##           Estimate Std.err
## alpha.1:2    0.164  0.0351
## alpha.1:3    0.178  0.0367
## alpha.1:4    0.197  0.0564
## alpha.1:5    0.178  0.0479
## alpha.2:3    0.202  0.0490
## alpha.2:4    0.184  0.0445
## alpha.2:5    0.149  0.0445
## alpha.3:4    0.325  0.0893
## alpha.3:5    0.310  0.0764
## alpha.4:5    0.240  0.0669
## Number of clusters:   1683  Maximum cluster size: 5
```

```r
# goodness of fit
pval=1-pchisq(G1/phi,gee7$df.residual)
pval
```

```
## [1] 0.498
```

Using adjusted Pearson chi-squared statistic, we get p-value $0.498 > 0.05$. Hence we do not have enough evidence to show the model does not fit the data well.