# p8157_hw2

Anna Ma

2022-10-16

## Question 1

Randomized, double-blind, parallel-group, multicenter study comparing two oral treatments (denoted A and B) for toe-nail infection, 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.

The main objective of the analyses is to compare the effects of oral treatments A and B on changes in the probability of the binary onycholysis outcome over the duration of the study.

- The *binary onycholysis outcome variable* $Y$ is coded 0 = none or mild, 1 = moderate or severe.
- The categorical variable Treatment is coded 1=oral treatment A, 0=oral treatment B.
- The variable Month denotes the exact timing of measurements in months.
- The variable Visit denotes the visit number (visit numbers 1-7 correspond to scheduled visits at 0, 4, 8, 12, 24, 36, and 48 weeks).

```
toenail_df = read_delim(file = "toenail.txt", delim = " ", col_names = c("id", "response", "treatment",
```

**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.**

- Model setup: $\mu_{ij} = E[Y_{ij}] = E[Y_{ij}|X_{ij}]$
- Link function: logit link function: $log(\frac{\mu_{ij}}{1-\mu ij}) = \eta_{ij} = \sum_{k=1}^{p} X_{ij}\beta_k$
- Under binomial assumption: $Var(Y_{ij}) = \phi v(\mu_{ij}) = \mu_{ij}(1-\mu_{ij})$
- correlation structure: exchangeable

The model can be written as:

$$\eta_{ij} = \beta_0 + \beta_1 month_{ij} + \beta_2 treatment_i + \beta_3 (month_{ij} * treatment_i)$$

```
gee1 = geeglm(response ~ month * treatment, data = toenail_df, id = id, family = binomial(link = "logit
summary(gee1)
```

```
##
## Call:
## geeglm(formula = response ~ month * treatment, family = binomial(link = "logit"),
##     data = toenail_df, id = id, corstr = "exchangeable")
##
##  Coefficients:
```

```
##                 Estimate  Std.err   Wald Pr(>|W|)
## (Intercept)     -0.58192  0.17206 11.439 0.000719 ***
## month           -0.17128  0.03000 32.596 1.13e-08 ***
## treatment        0.00718  0.25949  0.001 0.977924
## month:treatment -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
```

Test whether the month and treatment interaction is needed: $H_0 : \beta_3 = 0$

```
L = matrix(0, ncol = 4, nrow = 1) #ncol = number of coefficients in the model, nrow = number of tests
L[1,4] = 1
esticon(gee1,L=L,joint.test = FALSE)
```

```
##      estimate std.error statistic p.value   beta0 df
## [1,]  -0.0777    0.0541    2.0635  0.1509  0.0000  1
```

Since the p-value $= 0.1509 > 0.05$, therefore, at significance level at $\alpha = 0.05$, we fail to reject the null hypothesis and conclude that the interaction term can be removed. After removing the interaction, we have the model as follows:

$$\eta_{ij} = \beta_0 + \beta_1 month_{ij} + \beta_2 treatment_i$$

```
gee2 = geeglm(response ~ month + treatment, data = toenail_df, id = id, family = binomial(link = "logit"
summary(gee2)
```

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

```
summary(gee2)$coefficients %>% knitr::kable()
```

**2. Provide Interpretations for the coefficients in your model.**

| | Estimate | Std.err | Wald | Pr(>\|W\|) |
|---|---|---|---|---|
| (Intercept) | -0.610 | 0.178 | 11.796 | 0.001 |
| month | -0.205 | 0.026 | 62.656 | 0.000 |
| treatment | 0.040 | 0.253 | 0.025 | 0.874 |

The estimate by the GEE is shown as above. Plug the estimate into our model, than we have

$$\eta_{ij} = -0.6104 - 0.2051 * month_{ij} + 0.0402 * treatment_i$$

From this model, we can interpret the estimates as follows:

- The log odds of moderate or severe onycholysis for oral treatment B at baseline is -0.610, that is, at baseline, patients taking oral treatment B are 0.543 times more likely to develop a moderate or severe onycholysis.
- The log odds ratio of moderate or severe onycholysis for 1 unit increase in month among treatment B is -0.2051. That is, for every 1 unit increase in month, patients taking oral treatment B are 0.815 times more likely to develop a moderate or severe onycholysis.
- The log odds ratio of moderate or severe onycholysis comparing treatment A to treatment B holding month constant is 0.0402. That is, in a fixed month, patients taking oral treatment A are 1.041 times more likely to develop a moderate or severe onycholysis.
- Combining the result above and the fact that the treatment covariate had a p-value greater than 0.05, we can conclude that treatment is insignificant in the development of onycholysis status.

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

**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.**

```
##      estimate std.error statistic p.value   beta0 df
## [1,]  -0.0777    0.0541    2.0635  0.1509  0.0000  1
```

In the first part of the question, we conducted a hypothesis test regarding the significance of the interaction term of treatment and month. From our test result shown above, we can see that the p-value $= 0.1509 > 0.05$, therefore, at significance level at $\alpha = 0.05$, we fail to reject the null hypothesis and conclude that the coefficient for the interaction term is 0. This means that **the effect of the treatment does not change over time**.

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

```
gee3 = geeglm(response ~ month * treatment, data = toenail_df, id = id, family = binomial(link = "logit
summary(gee3)
```

```
##
## Call:
## geeglm(formula = response ~ month * treatment, family = binomial(link = "logit"),
##     data = toenail_df, id = id, corstr = "unstructured")
##
##  Coefficients:
##                 Estimate Std.err  Wald Pr(>|W|)
## (Intercept)      -0.7396  0.1664 19.75  8.8e-06 ***
## month            -0.1319  0.0263 25.11  5.4e-07 ***
## treatment         0.0373  0.2469  0.02    0.880
## month:treatment  -0.0896  0.0482  3.46    0.063 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = unstructured
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     1.04   0.334
##   Link = identity
##
## Estimated Correlation Parameters:
##           Estimate Std.err
## alpha.1:2    0.904  0.2983
## alpha.1:3    0.712  0.2414
## alpha.1:4    0.512  0.1885
## alpha.1:5    0.247  0.1180
## alpha.1:6    0.157  0.0927
## alpha.1:7    0.131  0.0945
## alpha.2:3    0.824  0.2737
## alpha.2:4    0.608  0.2178
## alpha.2:5    0.272  0.1248
## alpha.2:6    0.238  0.1152
## alpha.2:7    0.157  0.1027
## alpha.3:4    0.789  0.2723
## alpha.3:5    0.284  0.1278
## alpha.3:6    0.215  0.1107
## alpha.3:7    0.192  0.1135
## alpha.4:5    0.368  0.1525
## alpha.4:6    0.282  0.1297
## alpha.4:7    0.250  0.1302
## alpha.5:6    0.498  0.1989
## alpha.5:7    0.475  0.2034
## alpha.6:7    0.706  0.2607
## Number of clusters:   294  Maximum cluster size: 7
```

Test for interaction term:

```
esticon(gee3,L=L,joint.test = FALSE)
```

```
##      estimate std.error statistic p.value   beta0 df
## [1,]  -0.0896    0.0482    3.4563  0.0630  0.0000  1
```

Since the p-value $= 0.063 > 0.05$, we fail to reject the null hypothesis and conclude that the coefficient for the interaction term is 0 and can be removed from the model. Therefore, the inference of parameters are not sensitive to this choice.

```
gee3_1 = geeglm(response ~ month + treatment, data = toenail_df, id = id, family = binomial(link = "logi
summary(gee3_1)
```

```
##
## Call:
## geeglm(formula = response ~ month + treatment, family = binomial(link = "logit"),
##     data = toenail_df, id = id, corstr = "unstructured")
##
##  Coefficients:
##             Estimate Std.err  Wald Pr(>|W|)
## (Intercept)  -0.6458  0.1587 16.56  4.7e-05 ***
## month        -0.1705  0.0227 56.25  6.4e-14 ***
## treatment    -0.1429  0.2157  0.44     0.51
## ---
## 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
```

**AR(1)**

```
gee4 = geeglm(response ~ month * treatment, data = toenail_df, id = id, family = binomial(link = "logit
summary(gee4)
```

```
##
## Call:
## geeglm(formula = response ~ month * treatment, family = binomial(link = "logit"),
##     data = toenail_df, id = id, corstr = "ar1")
##
##  Coefficients:
##                 Estimate Std.err  Wald Pr(>|W|)
## (Intercept)      -0.6441  0.1684 14.64  0.00013 ***
## month            -0.1376  0.0274 25.27    5e-07 ***
## treatment         0.0691  0.2520  0.08  0.78409
## month:treatment  -0.0968  0.0517  3.51  0.06105 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = ar1
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     1.01   0.362
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.687   0.135
## Number of clusters:   294  Maximum cluster size: 7
```

Test for interaction term:

```
esticon(gee4,L=L,joint.test = FALSE)
```

```
##      estimate std.error statistic p.value   beta0 df
## [1,]  -0.0968    0.0517    3.5086  0.0610  0.0000  1
```

Since the p-value $= 0.061 > 0.05$, we fail to reject the null hypothesis and conclude that the coefficient for the interaction term is 0 and can be removed from the model. Therefore, the inference of parameters are not sensitive to this choice.

Removing interaction term:

```
gee4_1 = geeglm(response ~ month + treatment, data = toenail_df, id = id, family = binomial(link = "log
summary(gee4_1)
```

```
##
## Call:
## geeglm(formula = response ~ month + treatment, family = binomial(link = "logit"),
##     data = toenail_df, id = id, corstr = "ar1")
##
##  Coefficients:
##             Estimate Std.err  Wald Pr(>|W|)
## (Intercept)  -0.5725  0.1596 12.86  0.00033 ***
## month        -0.1778  0.0241 54.57  1.5e-13 ***
## treatment    -0.0989  0.2156  0.21  0.64638
## ---
## 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
```

- Scale parameter: In the three correlation structures, the scale parameter $\phi$ is estimated to be: **Exchangeable: 1.088**, **Unstructured: 1.039**, **AR(1): 1.008**. All of those are close to 1, which is in line with our assumption of using the logit link function.
- Estimates:

1) Exchangeable

| | Estimate | Std.err | Wald | Pr(>\|W\|) |
|---|---|---|---|---|
| (Intercept) | -0.610 | 0.178 | 11.796 | 0.001 |
| month | -0.205 | 0.026 | 62.656 | 0.000 |
| treatment | 0.040 | 0.253 | 0.025 | 0.874 |

2) Unstructured

| | Estimate | Std.err | Wald | Pr(>\|W\|) |
|---|---|---|---|---|
| (Intercept) | -0.646 | 0.159 | 16.555 | 0.000 |
| month | -0.170 | 0.023 | 56.251 | 0.000 |
| treatment | -0.143 | 0.216 | 0.439 | 0.508 |

3) AR(1)

| | Estimate | Std.err | Wald | Pr(>\|W\|) |
|---|---|---|---|---|
| (Intercept) | -0.572 | 0.160 | 12.86 | 0.000 |
| month | -0.178 | 0.024 | 54.57 | 0.000 |
| treatment | -0.099 | 0.216 | 0.21 | 0.646 |

From the table we can see that the estimated coefficients, especially coefficient for treatment does change depending on the correlation structure we choose. However, the coefficient of treatment has a p-value greater than 0.05 in all three models, indicating that the treatment is not significant. Therefore, we can say that the significant variables of **insensitive** to the choice of correlation structure.

## Question 2

The Skin Cancer Prevention Study was a randomized, double-blind, placebo-controlled clinical trial of beta carotene to prevent non-melanoma skin cancer in high-risk subjects.A total of 1805 subjects were randomized to either placebo or 50 mg of beta carotene per day for 5 years.

The main objective of the analyses is to compare the effects of beta carotene on skin cancer rates.

- The outcome variable Y is a count of the of the number of new skin cancers per year.
- The categorical variable Treatment is coded 1=beta carotene, 0=placebo.
- The variable Year denotes the year of follow-up.
- The categorical variable Gender is coded 1 male, 0 female.
- The categorical variable Skin denotes skin type and is coded 1 = burns, 0 otherwise.
- The variable Exposure is a count of the number of previous skin cancers.
- The variable Age is the age (in years) of each subject at randomization.

```
skin_df = read.table(file = "skin.txt", header = FALSE,col.names = c("id", "center", "age", "skin", "ge
```

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

- Model set up $\mu_{ij} = E[Y_{ij}]$
- Under poisson assumption:
  - Link function: $log(\mu_{ij}) = \eta_{ij} = \sum_{k=1}^{p} X_{ij}\beta_k$
  - $V(\mu_{ij}) = \mu_{ij}$
- correlation structure: unstructured

The model can be written as:

$$\eta_{ij} = \beta_0 + \beta_1 treatment_i + \beta_2 year_{ij} + \beta_3 treatment_i * year_{ij}$$

```
q2_gee1 = geeglm(y ~ treatment*year, data = skin_df, family = "poisson"(link = "log"), id = id, corstr =
summary(q2_gee1)
```

```
##
## Call:
## geeglm(formula = y ~ treatment * year, family = poisson(link = "log"),
##     data = skin_df, id = id, corstr = "exchangeable")
##
##  Coefficients:
##                 Estimate   Std.err   Wald Pr(>|W|)
## (Intercept)    -1.365665  0.117549 134.97   <2e-16 ***
## treatment       0.060628  0.157957   0.15      0.7
## year            0.000147  0.030806   0.00      1.0
## treatment:year  0.032294  0.048372   0.45      0.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     2.64   0.364
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.378   0.111
## Number of clusters:   1683  Maximum cluster size: 5
```

Again, we will test if the interaction term is required: $H_0 : \beta_3 = 0$

```
L_q2 = matrix(0, ncol = 4, nrow = 1)
L_q2[1,4] = 1
esticon(q2_gee1, L = L_q2, joint.test = FALSE)
```

```
##      estimate std.error statistic p.value  beta0 df
## [1,]   0.0323    0.0484    0.4457  0.5044 0.0000  1
```

Since p-value = 0.504, which is greater than the significance level 0.05, we fail to reject the null and conclude that the coefficient for the interaction term is 0 and thus can be removed from our model.That is, the model will be written as follows:

$$\eta_{ij} = \beta_0 + \beta_1 treatment_i + \beta_2 year_{ij}$$

The new model without the interaction term will be fitted as:

```
q2_gee2 = geeglm(y ~ treatment+year, data = skin_df, family = "poisson"(link = "log"), id = id, corstr =
summary(q2_gee2)
```

```
##
## Call:
## geeglm(formula = y ~ treatment + year, family = poisson(link = "log"),
##     data = skin_df, id = id, corstr = "exchangeable")
##
##  Coefficients:
##             Estimate Std.err   Wald Pr(>|W|)
## (Intercept)  -1.4123  0.1080 171.10   <2e-16 ***
## treatment     0.1478  0.1094   1.83     0.18
## year          0.0173  0.0247   0.49     0.48
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     2.65   0.374
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.378   0.111
## Number of clusters:   1683  Maximum cluster size: 5
```

```
summary(q2_gee2)$coefficients %>% knitr::kable()
```

**2. Provide Interpretations for the coefficients in your model.**

| | Estimate | Std.err | Wald | Pr(>\|W\|) |
|---|---|---|---|---|
| (Intercept) | -1.412 | 0.108 | 171.095 | 0.000 |
| treatment | 0.148 | 0.109 | 1.825 | 0.177 |
| year | 0.017 | 0.025 | 0.492 | 0.483 |

From the estimate, we can see that the p-value for all the covariates are larger than 0.05, indicating that the covariates are insignificant. However, if we were still to interpret the result, we can say that:

- The log rate of having non-melanoma skin cancers in placebo group at baseline is -1.412.That is, the rate of having skin cancer in the placebo group at baseline is 0.244
- The log rate ratio of having non-melanoma skin cancer between treatment group and placebo group holding year constant is 0.148. That is, the rate of having skin cancer for the treatment group is 1.159 times that of the placebo group, holding follow up years constant.
- The log rate ratio of having non-melanoma skin cancer with one unit increase in years of follow-up for the placebo group is 0.017. That is, the rate of having skin cancer for the placebo group is 1.018 times that of the treatment group with every unit increase in follow up year.

Based on these results, the treatment group had higher rate of having new skin cancer compared to the placebo group when followup years are the same and the two group's rate ratio is close to one with prolonged increasing follow-up year. Therefore, it appears that beta-carotene is not effective in reducing skin cancer rates.

```
esticon(q2_gee1,L=L_q2,joint.test = FALSE)
```

**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.**

```
##      estimate std.error statistic p.value  beta0 df
## [1,]   0.0323    0.0484    0.4457  0.5044 0.0000  1
```

In the first part of the question, we conducted a hypothesis test regarding the significance of the interaction term of treatment and month. From our test result shown above, we can see that the p-value $= 0.504 > 0.05$, therefore, at significance level at $\alpha = 0.05$, we fail to reject the null hypothesis and conclude that the coefficient for the interaction term is 0. This means that **the effect of the treatment does not change over time.**

**4. Repeat the above analysis adjusting for skin type, age, and the count of the number of previous skincancers. What conclusions do you draw about the effect of beta carotene on the adjusted rate of skin cancers? 1) Model fitting**

With adjusting for skin type, age, and the count of the number of previous skincancers, the fitted model can be written as:

$$\eta_{ij} = \beta_0 + \beta_1 treatment + \beta_2 year + \beta_3 age + \beta_4 skin_1 + \beta_5 exposure$$

```
q2_gee3 = geeglm(y ~ treatment + year + age + skin + exposure, data = skin_df, family = "poisson"(link =
summary(q2_gee3)
```

```
##
## Call:
## geeglm(formula = y ~ treatment + year + age + skin + exposure,
##     family = poisson(link = "log"), data = skin_df, id = id,
##     corstr = "exchangeable")
##
##  Coefficients:
##             Estimate  Std.err   Wald Pr(>|W|)
## (Intercept) -3.04458  0.33263  83.78   <2e-16 ***
## treatment    0.12357  0.09941   1.55   0.2139
## year         0.01759  0.02521   0.49   0.4854
## age          0.01496  0.00525   8.12   0.0044 **
## skin         0.16191  0.11079   2.14   0.1439
```

```
## exposure     0.13899  0.01055 173.42   <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  0.0769
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.209  0.0262
## Number of clusters:   1683  Maximum cluster size: 5
```

**2) Test for beta carotene efficacy**

To test for the efficacy of beta carotene, we set the null hypothesis to be such that $H_0 : \beta_1 = 0$

```
L_q2_2 = matrix(0, ncol = 6, nrow = 1)
L_q2_2[1,2] = 1
esticon(q2_gee3, L = L_q2_2, joint.test = FALSE)
```

```
##      estimate std.error statistic p.value  beta0 df
## [1,]   0.1236    0.0994    1.5451  0.2139 0.0000  1
```

From the test result, we can see that the p-value = 0.214, which is greater than the significant level 0.05. Thus, we fail to reject the null hypothesis and conclude that the coefficient for treatment is 0. That is, the effect of beta carotene on skin cancer is insignificant. Instead, age and the count of the number of previous skin cancers do significantly effect the outcome on the rate of skin cancers however, with respective p-values of 0.0044 and <2e-16.

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

```
q2_gee4 = geeglm(y ~ treatment + year + age + skin + exposure, data = skin_df, family = "poisson"(link =
summary(q2_gee4)
```

```
##
## Call:
## geeglm(formula = y ~ treatment + year + age + skin + exposure,
##     family = poisson(link = "log"), data = skin_df, id = id,
##     corstr = "unstructured")
##
##  Coefficients:
##             Estimate  Std.err   Wald Pr(>|W|)
## (Intercept) -3.06545  0.32970  86.45   <2e-16 ***
## treatment    0.11595  0.09772   1.41   0.2354
## year         0.01637  0.02469   0.44   0.5072
## age          0.01527  0.00513   8.88   0.0029 **
## skin         0.18398  0.10808   2.90   0.0887 .
## exposure     0.13806  0.01016 184.49   <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  0.0776
##   Link = identity
##
## Estimated Correlation Parameters:
##           Estimate Std.err
## alpha.1:2    0.164  0.0353
## alpha.1:3    0.178  0.0365
## alpha.1:4    0.199  0.0572
## alpha.1:5    0.186  0.0513
## alpha.2:3    0.197  0.0479
## alpha.2:4    0.181  0.0436
## alpha.2:5    0.150  0.0457
## alpha.3:4    0.317  0.0884
## alpha.3:5    0.312  0.0773
## alpha.4:5    0.245  0.0686
## Number of clusters:   1683  Maximum cluster size: 5
```

**AR(1)**

```
q2_gee5 = geeglm(y ~ treatment + year + age + skin + exposure, data = skin_df, family = "poisson"(link =
summary(q2_gee5)
```

```
##
## Call:
## geeglm(formula = y ~ treatment + year + age + skin + exposure,
##     family = poisson(link = "log"), data = skin_df, id = id,
##     corstr = "ar1")
##
##  Coefficients:
##             Estimate  Std.err   Wald Pr(>|W|)
## (Intercept) -3.02093  0.32857  84.53   <2e-16 ***
## treatment    0.12808  0.10083   1.61   0.2040
## year         0.01056  0.02508   0.18   0.6737
## age          0.01494  0.00511   8.53   0.0035 **
## skin         0.15284  0.11232   1.85   0.1736
## exposure     0.13915  0.01065 170.79   <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  0.0788
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.294  0.0328
## Number of clusters:   1683  Maximum cluster size: 5
```

- Scale parameter: In the three correlation structures, the scale parameter $\phi$ is estimated to be: **Ex-**

**changeable: 1.638**, **Unstructured: 1.642**, **AR(1): 1.636**. All of the three correlation structure had the identical estimation for the scale parameter.

- Estimates:

1) Exchangeable

| | Estimate | Std.err | Wald | Pr(>\|W\|) |
|---|---|---|---|---|
| (Intercept) | -3.045 | 0.333 | 83.777 | 0.000 |
| treatment | 0.124 | 0.099 | 1.545 | 0.214 |
| year | 0.018 | 0.025 | 0.487 | 0.485 |
| age | 0.015 | 0.005 | 8.122 | 0.004 |
| skin | 0.162 | 0.111 | 2.136 | 0.144 |
| exposure | 0.139 | 0.011 | 173.419 | 0.000 |

2) Unstructured

| | Estimate | Std.err | Wald | Pr(>\|W\|) |
|---|---|---|---|---|
| (Intercept) | -3.065 | 0.330 | 86.45 | 0.000 |
| treatment | 0.116 | 0.098 | 1.41 | 0.235 |
| year | 0.016 | 0.025 | 0.44 | 0.507 |
| age | 0.015 | 0.005 | 8.88 | 0.003 |
| skin | 0.184 | 0.108 | 2.90 | 0.089 |
| exposure | 0.138 | 0.010 | 184.49 | 0.000 |

3) AR(1)

| | Estimate | Std.err | Wald | Pr(>\|W\|) |
|---|---|---|---|---|
| (Intercept) | -3.021 | 0.329 | 84.532 | 0.000 |
| treatment | 0.128 | 0.101 | 1.613 | 0.204 |
| year | 0.011 | 0.025 | 0.177 | 0.674 |
| age | 0.015 | 0.005 | 8.534 | 0.003 |
| skin | 0.153 | 0.112 | 1.852 | 0.174 |
| exposure | 0.139 | 0.011 | 170.786 | 0.000 |

From the table we can see that the estimated coefficients changes only slightly depending on the correlation structure we choose. Also, all of the three models had the same insignificant covariates. In all three models, treatment, year, and skin had p-values greater than 0.05, indicating that they are not significant in impacting the response y. On the other hand, age and the count of the number of previous skin cancers are covariates that identified by all three models to be factors that have significantly effect the outcome on the rate of skin cancers. Based on those results, we can say that the significant variables of *insensitive* to the choice of correlation structure.

**6. Do you need to account for overdisperion. Comment.** In part one, when we only included treatment and year as covariates with unstructured correlation, the scale parameter is estimated to be 2.647, 0.374. In part 4) and 5), we added skin type, age, and the count of the number of previous skin cancers in the model and tested different correlation structures. All three model gave the same scale parameter, $\phi$ is estimated to be: Unstructured: 1.642,Exchangeable: 1.638, AR(1): 1.636. In all cases, $\phi$ is greater than 1, which is the value we should have under our poisson assumption. This indicates that there may be overdispersion, and **we should account for overdispersion** when fitting the model. Similar conclusion cna be obtained from the dispersion tests below:

$H_0 : \phi = 1$, $H_a : \phi > 1$

```
library(AER)
```

1). For part 1 model with only treatment and year:

```
dispersiontest(q2_gee2)
```

```
##
##  Overdispersion test
##
## data:  q2_gee2
## z = 5, p-value = 1e-07
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##       2.66
```

2). For part 4 model with covariates treatment, year, skin type, age, and the count of the number of previous skin cancers:

```
dispersiontest(q2_gee3)
```

```
##
##  Overdispersion test
##
## data:  q2_gee3
## z = 9, p-value <2e-16
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##       1.72
```

3) For part 5 model with different correlation structure:

- unstructured

```
dispersiontest(q2_gee4)
```

```
##
##  Overdispersion test
##
## data:  q2_gee4
## z = 9, p-value <2e-16
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##       1.72
```

- AR(1)

```
dispersiontest(q2_gee5)
```

```
##
##  Overdispersion test
##
## data:  q2_gee5
## z = 9, p-value <2e-16
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
```

```
## dispersion
##       1.72
```

All three test showed p-value less than 0.05, we therefore reject the null hypothesis and conclude that there's no statistically significant evidence to show that the scale parameter is 1 in all those models, indicating that **overdispersion needs to be accounted for**.