Appendix A: Estimating a generalized linear model (GLM) using GEEs with SPSS

The following guide shows how to use GEEs for a binary outcome using clustered data. The dataset is available at: [http://francish.netlify.app/docs/cont_binary.sav](http://francish.netlify.app/docs/cont_binary.sav).

1. With a dataset already open, select **Analyze → Generalized Linear Models → Generalized Estimating Equations**

2. Under the **Repeated** tabs, select the clustering variable and place it in the **Subject variables** field. In this example, `district` is the clustering variable.

3. In the same window, choose the **Working Correlation Matrix** in the dropdown list which indicates **Structure**. As this example focuses on analyzing a cross-sectional clustered dataset, choose the **Exchangeable** option.

4. Click on the **Type of Model** tab. For continuous outcomes, no change is necessary on this screen. For this example, a logistic regression model will be run, select the **Binary logistic** option.

5. Click on the **Response** tab. Select the outcome variable (i.e., `cuse` in this example) and place it in the **Dependent Variable** field.
6. Since a logistic regression will be run, users should make sure that the reference group is correctly specified. In this case, the outcome is a 0 or a 1. As we would like to model the 1s in comparison to the 0s, click on Reference Category… and select First (lowest value). If this is not done, the model may estimate the likelihood of getting a 0 compared to getting a 1 (and the resulting coefficients may be in the opposite direction as to what was expected). Click Continue.

7. Select Predictors to include in the model. As all the variables are already numeric (dummy coded as a 1 or a 0), select the variables of interest and place them in the Covariates: section. [in this example, nochild is not included as this is the reference category]
8. Click on the **Model** tab. Select the variables of interest on the left and click on ![button](image) to transfer them to the **Model** box on the right.

![Image](image)

9. At this point, users can click **OK** to run the model. However, as this is a logistic regression model, it may be easier to interpret the exponentiated log odds which are odds ratios (**ORs**). To output the **ORs**, click on the **Statistics** tab. Click on **Include exponential parameter estimates**. Click **OK**. **Do not do this if running a linear model.**

![Image](image)
10. The following is a portion of the sample output showing the regression coefficients, standard errors, and statistical significance of the estimates—and also the odds ratios under the heading \( \text{Exp}(B) \).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>B</th>
<th>Std Error</th>
<th>95% Wald Confidence Interval Lower</th>
<th>95% Wald Confidence Interval Upper</th>
<th>Wald Ch-Square</th>
<th>df</th>
<th>Sig</th>
<th>Exp(B)</th>
<th>95% Wald Confidence Interval for Exp(B) Lower</th>
<th>95% Wald Confidence Interval for Exp(B) Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>(intercept)</td>
<td>-0.965</td>
<td>0.004</td>
<td>-1.378</td>
<td>-0.593</td>
<td>24.179</td>
<td>1</td>
<td>0.000</td>
<td>0.373</td>
<td>0.252</td>
<td>0.553</td>
</tr>
<tr>
<td>age</td>
<td>0.004</td>
<td>0.0007</td>
<td>-0.013</td>
<td>0.021</td>
<td>1.97</td>
<td>1</td>
<td>0.657</td>
<td>1.004</td>
<td>0.987</td>
<td>1.021</td>
</tr>
<tr>
<td>age2</td>
<td>-0.004</td>
<td>0.0007</td>
<td>-0.006</td>
<td>-0.003</td>
<td>41.059</td>
<td>1</td>
<td>0.000</td>
<td>0.996</td>
<td>0.994</td>
<td>0.997</td>
</tr>
<tr>
<td>child1</td>
<td>0.778</td>
<td>0.1881</td>
<td>0.409</td>
<td>1.147</td>
<td>17.095</td>
<td>1</td>
<td>0.000</td>
<td>2.177</td>
<td>1.506</td>
<td>3.148</td>
</tr>
<tr>
<td>child2</td>
<td>0.668</td>
<td>0.1603</td>
<td>0.534</td>
<td>1.192</td>
<td>29.319</td>
<td>1</td>
<td>0.000</td>
<td>2.383</td>
<td>1.740</td>
<td>3.262</td>
</tr>
<tr>
<td>child3plus</td>
<td>0.870</td>
<td>0.2022</td>
<td>0.474</td>
<td>1.266</td>
<td>18.518</td>
<td>1</td>
<td>0.000</td>
<td>2.387</td>
<td>1.606</td>
<td>3.548</td>
</tr>
<tr>
<td>urb</td>
<td>0.644</td>
<td>0.1532</td>
<td>0.343</td>
<td>0.944</td>
<td>17.654</td>
<td>1</td>
<td>0.000</td>
<td>1.904</td>
<td>1.410</td>
<td>2.571</td>
</tr>
</tbody>
</table>

Dependent Variable: cuse
Model: (intercept), age, age2, child1, child2, child3plus, urb
Appendix B: Small sample correction

Francis L. Huang, PhD / huangf@missouri.edu

2020.06.25

1. Load in the libraries

- sampling is used to facilitate the random selection of 20 groups (J = 20).
- geesmv (Wang et al., 2016) has eight different standard error corrections for GEEs. It is used to estimate the standard error adjustments.
- geepack is still used to estimate the GLM using GEE.

```r
library(mlmRev)  #has the Hsb82 dataset
library(sampling) #to randomly select 20 schools
library(geesmv)  #for small sample correction
library(geepack) #for geeglm

data(Hsb82)
set.seed(123)

#randomly select 20 schools
sel <- cluster(Hsb82, "school", 20, method = 'srswor')
J20 <- getdata(Hsb82, sel)  #create the J20 dataset
J20$school <- droplevels(J20$school) #remove unused school factor levels
length(table(J20$school))  #how many schools
```

## [1] 20

2. Run the model

```r
gee1 <- geeglm(mAch ~ sx + minrty + cses + sector + meanses + cses * sector, id = school, corstr = 'exchangeable', data = J20)
summary(gee1)
```

## Call:
## geeglm(formula = mAch ~ sx + minrty + cses + sector + meanses +
##          cses * sector, data = J20, id = school, corstr = "exchangeable")
##
## Coefficients:             Estimate  Std.err   Wald Pr(>|W|)
## (Intercept) 14.3923  0.8595 280.376  < 2e-16 ***
## sxFemale    -1.8082  0.3549  25.961  3.48e-07 ***
## minrtyYes   -2.0620  0.5752  12.853  0.000337 ***
## cses        3.1651  0.6432  24.215   8.62e-07 ***
## sectorCatholic  0.9306 1.4555   0.409  0.522568
## meanses     3.3461  1.7885   3.500  0.061353
## cses:sectorCatholic -1.7145  0.6805  6.347  0.0011760 *
## ---
## Signif. codes:  
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
## | Estimate | Std.err |
## |----------|---------|
## (Intercept) | 33.48   | 1.604  |
## Link = identity
##
## Estimated Correlation Parameters:
##
## | Estimate | Std.err |
## |----------|---------|
## alpha    | 0.07645 | 0.0246 |
##
## Number of clusters: 20  Maximum cluster size: 60

### 3. Compute adjusted standard errors

To get the corrected standard errors, use the `GEE.var.md` function. The specification is the same as with the `geeglm` function. Note though that a difference is that the cluster variable is surrounded in quotes. The `GEE.var.lz` function computes the standard Liang & Zeger (1986) robust standard errors.

```r
ggee.lz <- GEE.var.lz(mAch ~ sx + minrty + cses + sector + meanses + cses * sector, id = 'school', corstr = 'exchangeable', data = J20)
```

```r
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
## running glm to get initial regression estimate
##
## | (Intercept) | sxFemale | minrtyYes | cses |
## |------------|---------|-----------|------|
## | 14.474     | -2.388  | -2.053    | 3.143|
##
## gee.md <- GEE.var.md(mAch ~ sx + minrty + cses + sector + meanses + cses * sector, id = 'school', corstr = 'exchangeable', data = J20)
```

```r
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
## running glm to get initial regression estimate
##
## | (Intercept) | sxFemale | minrtyYes | cses |
## |------------|---------|-----------|------|
## | 14.474     | -2.388  | -2.053    | 3.143|
##
## lz.se <- sqrt(gee.lz$cov.beta)  #sqrt to get the SE
## md.se <- sqrt(gee.md$cov.beta)
```
A comparison between the two SEs shows that the MD SEs are higher (more conservative) vs. the LZ SEs. MD SEs can be 10 - 20% larger.

```r
data.frame(lz = lz.se, 
            md = md.se)

##                         lz     md
## (Intercept)         0.8595 1.0938
## sxFemale            0.3549 0.3784
## minrtyYes           0.5753 0.6620
## cses                0.6432 0.7698
## sectorCatholic      1.4554 2.1057
## meanses             1.7884 2.6795
## cses:sectorCatholic 0.6805 0.8051
```

To use the adjusted standard errors, need to use the `coeftest` function in the `lmtest` package. The `geesmv` function only provides the variances/standard errors. To use them, need to place the SEs in a diagonal matrix which is used in the `coeftest` function.

For comparison, compute the Liang and Zeger (LZ; 1986) standard errors and the Mancl and DeRouen (MD) standard errors.

```r
library(lmtest)
se1 <- diag(gee.lz$cov.beta)
coeftest(gee1, se1)
```

## z test of coefficients:

```r
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)            14.392     0.860  16.740 < 2e-16 ***
## sxFemale               -1.808     0.355  -5.108 3.5e-07 ***
## minrtyYes             -2.062     0.575  -3.580 0.00034 ***
## cses                  -3.165     0.643  -4.918 8.6e-07 ***
## sectorCatholic        -0.931     1.455  -0.644 0.52253
## meanses                3.346     1.788   1.866 0.06134 .
## cses:sectorCatholic   -1.714     0.681  -2.520 0.01176 *
```

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```r
summary(gee1) #should be the same
```

## Call:
```
# geeglm(formula = mAch ~ sx + minrty + cses + sector + meanses +
#         cses * sector, data = J20, id = school, corstr = "exchangeable")
```

## Coefficients:

```r
##                      Estimate Std.err   Wald Pr(>|W|)
## (Intercept)           14.392     0.860 280.38 < 2e-16 ***
## sxFemale               -1.808     0.355  25.96 3.5e-07 ***
```
## Supplementary Material

```
## minrtyYes    -2.062   0.575  12.85  0.00034 ***
## cses         3.165   0.643  24.21  8.6e-07 ***
## sectorCatholic 0.931  1.456   0.41  0.52257
## meanses      3.346   1.788   3.50  0.06135 .
## cses:sectorCatholic -1.714   0.681  6.35  0.01176 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##   (Intercept)     33.5     1.6
##   Link = identity
##
## Estimated Correlation Parameters:
##   alpha   0.0765  0.0246
## Number of clusters:   20  Maximum cluster size: 60
```

Compare this now to the Mancl and DeRouen (2001) correction (the standard errors are larger):

```r
se2 <- diag(gee.md$cov.beta) coeftest(gee1, se2)
```

## References

