GEE vs. GLMM comparison with binary outcomes

2020.06.25

The following document shows how to run a logistic regression model using generalized estimating equations (GEEs) and comparing results using a generalized linear mixed model (GLMM or hierarchical GLM; HGLM) with binary outcomes. Shows the difference between population averaged (PA) and cluster specific (CS) results.

Load in packages

```
library(geepack) #for GEE
library(GLMMadaptive) #for GLMM
library(haven) #for importing SPSS datafile
```

Read and inspect data

Patterned after a cluster randomized control trial (CRCT). - Variable of interest: `y`. - Treatment variable (`tr`) at level 2 is of interest. - Can include other covariates: `w1` (L2), `x1` (L1), `x2` (L1) in the model.

```
dat <-
haven::read_sav('https://github.com/flh3/jebsgee/blob/main/01_data/log_CRCT.s
av?raw=true')
summary(dat)
```

```
##        y              school            tr                w1
##  Min.   :0.0000   Min.   : 1.00   Min.   :0.0000   Min.   :-1.5677
##  1st Qu.:0.0000   1st Qu.: 8.00   1st Qu.:0.0000   1st Qu.:-0.6982
##  Median :1.0000   Median :15.00   Median :1.0000   Median :-0.1933
##  Mean   :0.6954   Mean   :15.56   Mean   :0.5077   Mean   :-0.1505
##  3rd Qu.:1.0000   3rd Qu.:23.00   3rd Qu.:1.0000   3rd Qu.: 0.4883
##  Max.   :1.0000   Max.   :30.00   Max.   :1.0000   Max.   : 1.8428
##        x1                 x2
##  Min.   :-3.04098   Min.   :0.0000
##  1st Qu.:-0.61150   1st Qu.:0.0000
##  Median : 0.04691   Median :1.0000
##  Mean   : 0.04680   Mean   :0.5154
##  3rd Qu.: 0.70126   3rd Qu.:1.0000
##  Max.   : 2.96546   Max.   :1.0000
```

```
head(dat)

## # A tibble: 6 x 6
##       y school    tr     w1     x1    x2
##   <dbl>  <dbl> <dbl>  <dbl>  <dbl> <dbl>
## 1     1      1     1 -0.593  0.639     0
## 2     0      1     1 -0.593  1.57      0
## 3     1      1     1 -0.593 -0.776     0
## 4     1      1     1 -0.593 -0.850     1
## 5     0      1     1 -0.593  0.360     0
## 6     1      1     1 -0.593  0.983     0

length(unique(dat$school)) #how many clusters

## [1] 30

xtabs(~y + tr, data = dat)

##    tr
## y     0   1
##   0 111  87
##   1 209 243
```

Run the models

GEE estimation using geepack. Note: `id`, `corstr`, and `family` are specified. For continuous outcomes, no need to specify the `family` option. NOTE: output is in log odds units. Need to exponentiate to get the odds ratios.

- Indicate the cluster variable in the `id` option
- Indicate the exchangeable correlation structure

NOTE: At times, gee functions in R may require the datasets to be sorted by the clustering variable to work properly (other software such as HLM may require this as well). To sort the dataset using Base R:

```
dat <- dat[order(dat$school), ]
```

or (if using dplyr)

```
dat <- dplyr::arrange(dat, school)
```

```
m1 <- geeglm(y ~ tr + w1 + x1 + x2,
             id = school,
             data = dat,
             corstr = 'exchangeable',
             family = binomial)
summary(m1)

##
## Call:
## geeglm(formula = y ~ tr + w1 + x1 + x2, family = binomial, data = dat,
##     id = school, corstr = "exchangeable")
##
##  Coefficients:
##             Estimate  Std.err   Wald Pr(>|W|)
## (Intercept)  0.92330  0.17079 29.225 6.44e-08 ***
## tr           0.57760  0.27465  4.423 0.035464 *
## w1           0.31734  0.16624  3.644 0.056268 .
## x1           0.50987  0.08131 39.321 3.60e-10 ***
## x2          -0.52802  0.15300 11.911 0.000558 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)    1.016  0.1616
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha  0.04847 0.02849
## Number of clusters:   30  Maximum cluster size: 24
```

NOTE: the alpha shown is the conditional intraclass correlation as estimated using GEEs.

Run a mixed model using `GLMMadaptive`. Can also use `glmer` in `lme4` but when using `mixed_model`, can easily marginalize results (i.e., convert from CS -> PA).

```
m2 <- mixed_model(y ~ tr + w1 + x1 + x2,
                  random = ~1|school,
                  data = dat, family = binomial)
summary(m2)

##
## Call:
## mixed_model(fixed = y ~ tr + w1 + x1 + x2, random = ~1 | school,
##     data = dat, family = binomial)
##
## Data Descriptives:
## Number of Observations: 650
## Number of Groups: 30
##
## Model:
##  family: binomial
##  link: logit
##
## Fit statistics:
##  log.Lik   AIC   BIC
##   -364.8 741.7 750.1
##
## Random effects covariance matrix:
##             StdDev
## (Intercept) 0.5065
##
## Fixed effects:
##             Estimate Std.Err z-value p-value
## (Intercept)   0.9776  0.2160   4.526  <1e-04
## tr            0.5759  0.2680   2.149  0.0316
## w1            0.3296  0.1548   2.128  0.0333
## x1            0.5372  0.0961   5.589  <1e-04
## x2           -0.5548  0.1887  -2.940  0.0033
##
## Integration:
## method: adaptive Gauss-Hermite quadrature rule
## quadrature points: 11
##
## Optimization:
## method: EM
## converged: TRUE
```

Can marginalize afterwards:

```
marginal_coefs(m2) #marginalize using GLMMadaptive

## (Intercept)          tr          w1          x1          x2
##      0.9295      0.5482      0.3152      0.5124     -0.5287

# can include robust standard errors too
marginal_coefs(m2, std_errors = T, sandwich = T)

##             Estimate Std.Err z-value p-value
## (Intercept)   0.9295  0.1678   5.540 < 1e-04
## tr            0.5482  0.2723   2.013 0.04409
## w1            0.3152  0.1542   2.044 0.04092
## x1            0.5124  0.0851   6.025 < 1e-04
## x2           -0.5287  0.1434  -3.687 0.00023
```

Not necesssary but inspecting the intraclass correlation coefficient:

```
(Tau <- summary(m2)$D) #variance at level2

##             (Intercept)
## (Intercept)      0.2566

(Tau / (Tau + 3.29)) #ICC

##             (Intercept)
## (Intercept)     0.07235

performance::icc(m2) #automatic checking

## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.072
##   Conditional ICC: 0.063
```

Can manually convert CS to PA results. Can compute the correction factor using the variance at level 2.

```
adj <- sqrt((.346 * Tau) + 1) #Allison, 2009, p. 66
```

Create a data frame with all the results– all quite similar.

```
outp <- data.frame(MLM_CS = fixef(m2), GEE_PA = coef(m1), margin =
marginal_coefs(m2)$betas, manual = fixef(m2) / adj)
outp
```

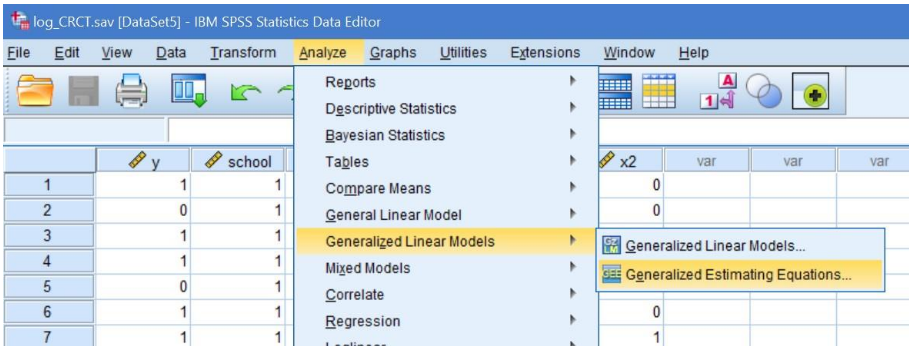
```
##              MLM_CS  GEE_PA  margin  manual
## (Intercept)  0.9776  0.9233  0.9295  0.9369
## tr           0.5759  0.5776  0.5482  0.5520
## w1           0.3296  0.3173  0.3152  0.3158
## x1           0.5372  0.5099  0.5124  0.5148
## x2          -0.5548 -0.5280 -0.5287 -0.5317
```

END

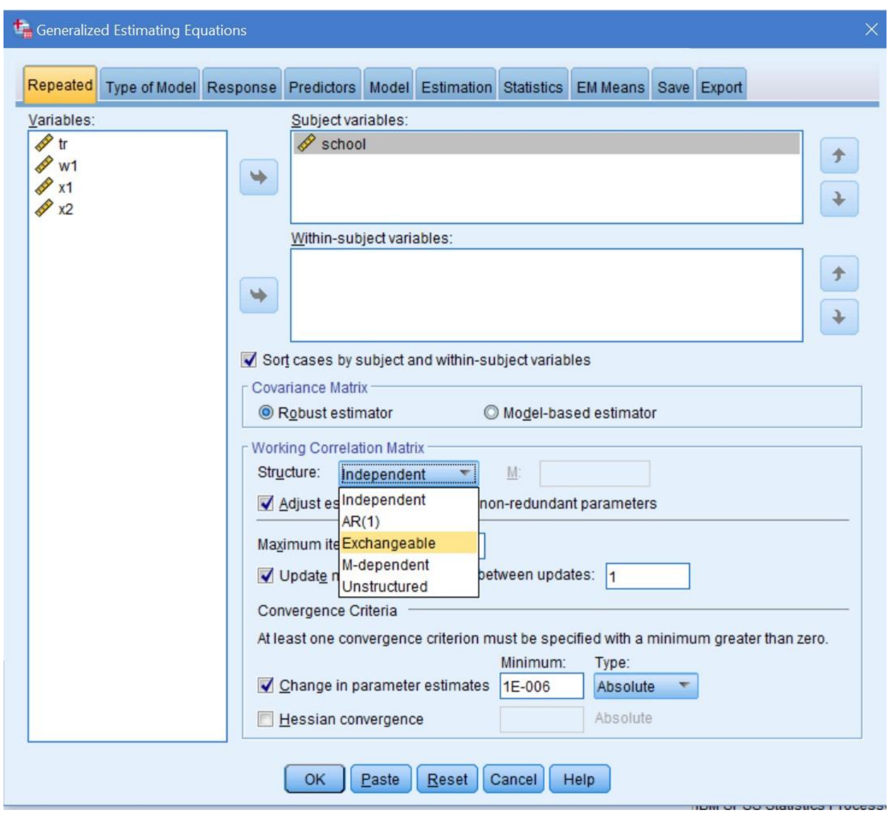
Appendix B: Estimating a generalized linear model (GLM) using GEEs with SPSS

The following guide shows how to use GEEs for a binary outcome using clustered data. The dataset is available at:
https://github.com/flh3/jebsgee/blob/main/01_data/log_CRCT.sav?raw=true

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

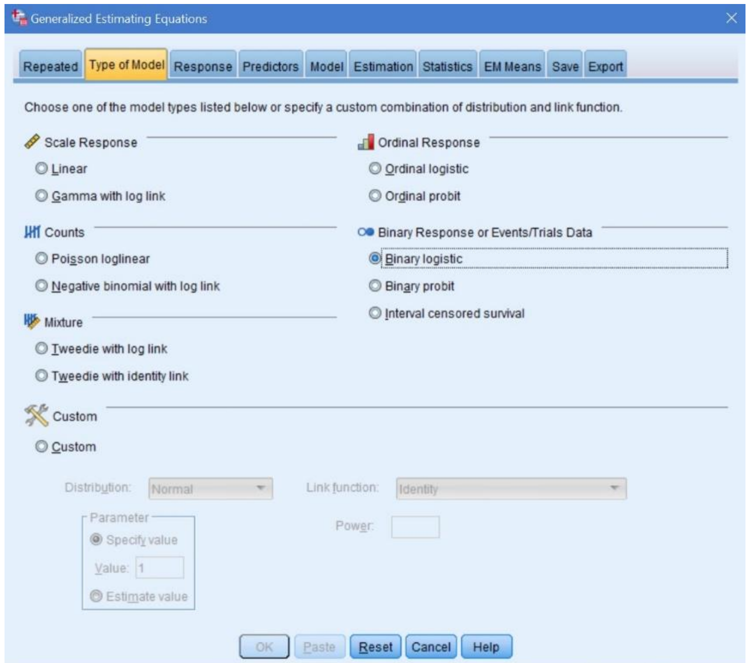


2. Under the **Repeated** tabs, select the clustering variable and place it in the **Subject variables** field. In this example, **school** is the clustering variable.
3. In the same window, choose the **Working Correlation Matrix** in the dropdown list which indicates **Structure**. As this example focuses on a *cluster randomized control trial*, choose the **Exchangeable** option.

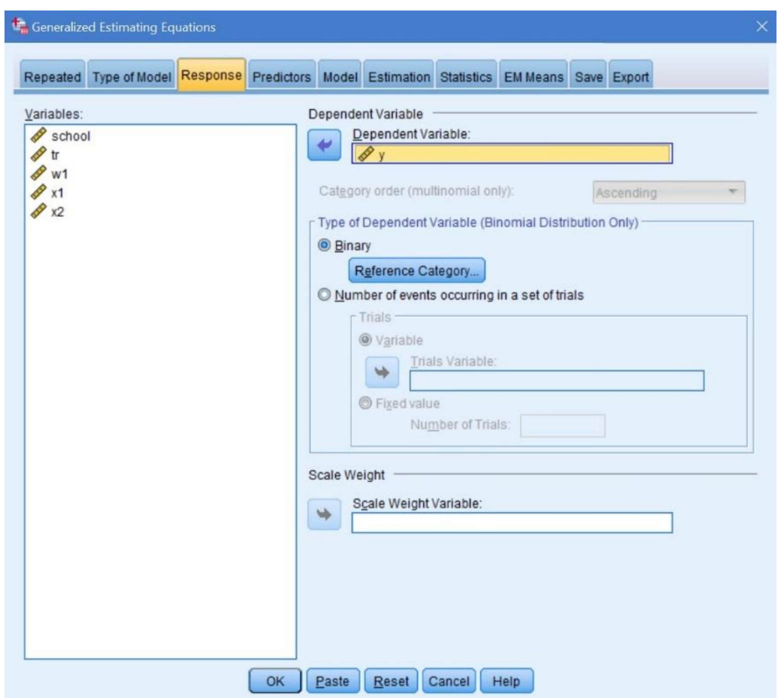


From: Huang, F. L. (2022). Analyzing cross-sectionally clustered data using generalized estimating equations. *Journal of Educational and Behavioral Statistics, 47,* 101-125. doi: 10.3102/10769986211017480

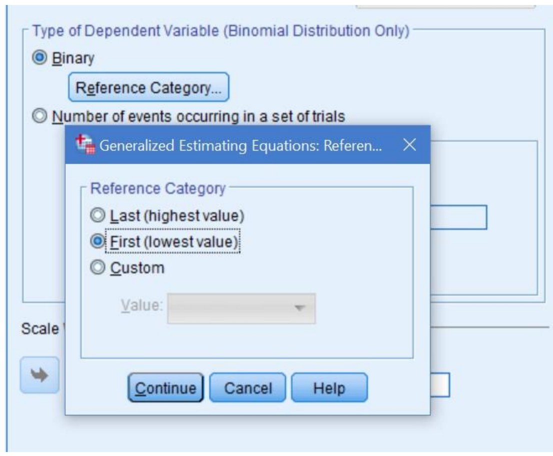
4. Click on the **Type of Model** tab. For continuous outcomes, no change is necessary on this screen. For this example, a logistic regression model will be run, select the **Binary logistic** option.



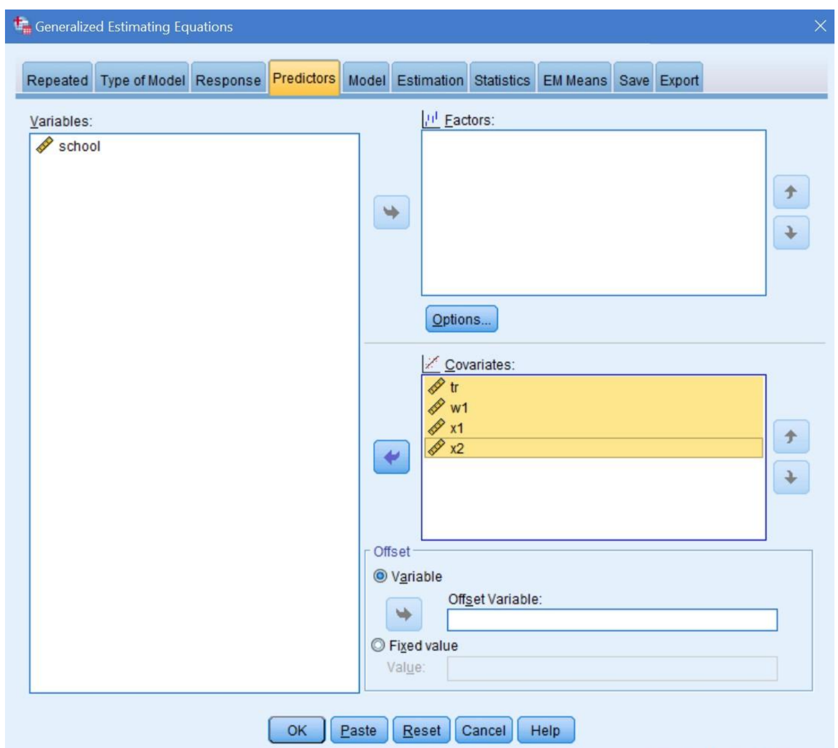
5. Click on the **Response** tab. Select the outcome variable and place it in the **Dependent Variable** field.



6. As the model is a logistic regression, 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 will estimate the likelihood of getting a 0 compared to getting a 1 (and the coefficients may be in the opposite direction as to what was expected). Click **Continue**.

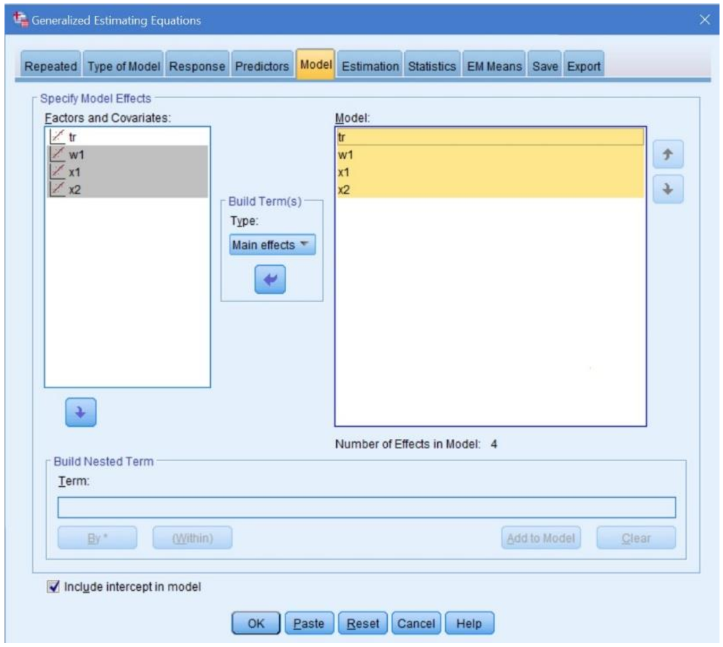


7. Select **Predictors** to include in the model. As all the variables are numeric (dummy coded as a 1 or a 0 for tr and also for x2), select the variables of interest and place them in the **Covariates:** section.

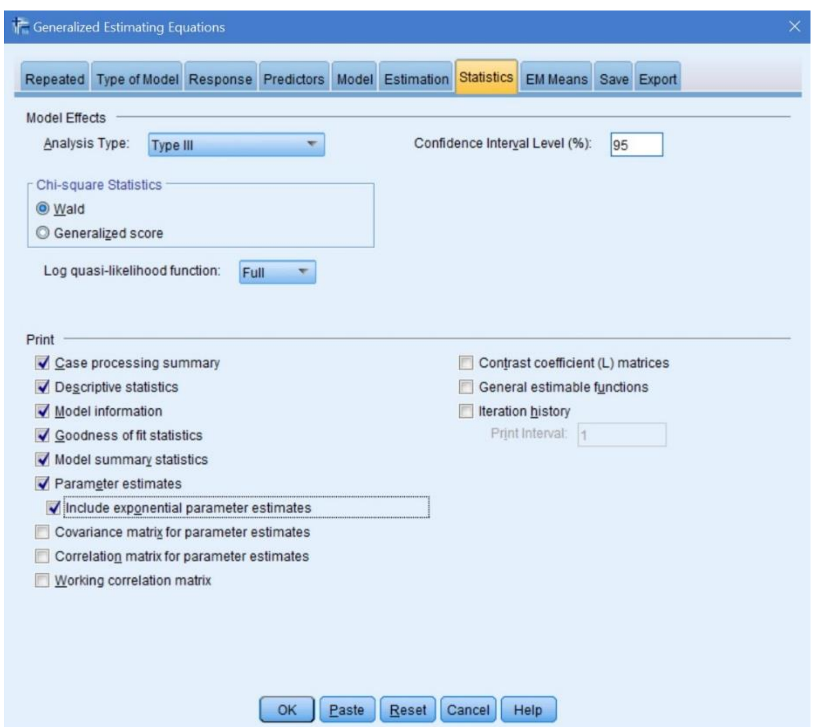


From: Huang, F. L. (2022). Analyzing cross-sectionally clustered data using generalized estimating equations. *Journal of Educational and Behavioral Statistics, 47,* 101-125. doi: 10.3102/10769986211017480

8. Click on the **Model** tab. Select the variables of interest on the left and transfer them to the **Model** field on the right.



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 (*OR*s). To output the *OR*s, click on the **Statistics** tab. Click on **Include exponential parameter estimates**. <u>Do not do this if running a linear model</u>.



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 **Exp(B)**.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval Lower	95% Wald Confidence Interval Upper	Hypothesis Test Wald Chi-Square	Hypothesis Test df	Hypothesis Test Sig.	Exp(B)	95% Wald Confidence Interval for Exp(B) Lower	95% Wald Confidence Interval for Exp(B) Upper
(Intercept)	.923	.1708	.589	1.258	29.228	1	.000	2.518	1.801	3.518
tr	.578	.2746	.039	1.116	4.422	1	.035	1.782	1.040	3.052
w1	.317	.1662	-.008	.643	3.647	1	.056	1.374	.992	1.903
x1	.510	.0813	.351	.669	39.328	1	.000	1.665	1.420	1.953
x2	-.528	.1530	-.828	-.228	11.912	1	.001	.590	.437	.796
(Scale)	1									

Dependent Variable: y
Model: (Intercept), tr, w1, x1, x2

11. Based on the above results, participants in the treatment group (tr = 1) had higher odds of passing (y = 1) by a factor of 1.78 (*OR* = 1.78, *p* = .04).