Analyzing Longitudinal Data: Generalized Estimating Equations

Ashley I Naimi

Fall 2024

Contents

1 GEE for Correlated Data

Generalized estimating equations (GEEs) are another method we can use to adjust our generalized linear model to account for a lack of independence, and recover the interpretation of the p-values, confidence intervals, and standard errors of interest. This was the focus of the paper by Liang and Zeger (Liang and Zeger, 1986), who originally introduced the method. Their "extension" did just that: adjusted the GLM by incorporating information on the correlation structure within individual units. This extension generalized the GLM using a theory of estimating equations, which does not rely on parametric likelihood methods, and the new method was hence named generalized estimating equations.

A key distinction between deploying GEEs versus GLMs is that, in the former, we have to consider the structure of the correlation within units in our data. Several correlation structures exist, and include things like the independence, exchangeable, or unstructured correlation matrices.

Let's again use our BMI data as we did with the robust variance and clustered bootstrap above. These data can then be analyzed using the `geeglm` functions in the `geepack` library in R. Note the additional arguments in this function, beyond the standard arguments in the `glm` function: we have an `id`, `scale.fix`, and `corstr` argument in the GEE function.

The `id` argument takes as input the information on the clustering of the data. In our case, observations are clustered within `practice`.

The scale parameter in a GEE model is a parameter that allows us to deal with potential overdispersion. Setting the `scale.fix = T` causes the scale parameter to be fixed at a value of 1, rather than estimated. Because binary and gaussian data cannot be "overdispersed," leaving the `scale.fix` argument at it's default value of `FALSE` is typically not appropriate for these distributions. When the outcome data are assumed generated from a Poisson distribution we could let this argument take its default value of F so that a scale parameter is estimated.

Deeper Dive: Overdispersion

Overdispersion describes a situation where the variance in an outcome variable is greater than the variance that can be captured by the model of that outcome variable. This concept is most commonly encountered with a Poisson random variable. The variance of a Poisson distribution is equal to the mean. If we let μ denote the mean of an outcome variable Y, and we let σ^2 denote it's variance, then modeling Y with a Poisson regression model (GLM or GEE) implies that $\mu = \sigma^2$. However, if we let $\bar{y}$ denote the sample mean and s^2 denote the sample variance, we may find that: $\bar{y} < s^2$.

In this case, we would say that the outcome is "overdispersed". One solution to overdispersion is to estimate a scale parameter in a GLM or GEE model. In a Poisson regression context, adding a scale parameter re-defines the variance of the outcome to be:

$$V(Y) = \phi\lambda,$$

where λ is the variance of the Poisson distribution (which is equal to the mean), and ϕ is a scale that multiplies this variance accordingly. In a regression modeling context, we can set the scale parameter ϕ to be 1 if there is no overdispersion. Or we can estimate it, and scale the standard errors from the regression model, thus accounting for overdispersion in the outcome.

Other distributions can be affected by overdispersion (e.g., binomial distribution with $N > 1$). However, the Gaussian distribution and Binomial distribution with $N = 1$ (i.e., a Bernoulli random variable) are not typically affected by overdispersion. In this cases, it is common to set the scale parameter to 1 instead of estimate it. Other techniques are also available to handle overdispersion (e.g., using the Negative Binomal distribution instead, which allows the variance to exceed the mean for count outcome data).

Finally, the `corstr` argument allows us to specify the form of the working correlation matrix. In our BMI data, we can fit the following GEE models:

```r
cluster_trial <- read_csv(here("data","cluster_trial_data_bmi.csv"))

#install.packages("geepack")
library(geepack)

## use GEE
mod1_ind <- geeglm(BMI ~ treatment,
                   family = gaussian(link = "identity"),
                   id = practice,
                   data=cluster_trial,
                   scale.fix = T,
                   corstr="independence")

summary(mod1_ind)
```

```
##
## Call:
## geeglm(formula = BMI ~ treatment, family = gaussian(link = "identity"),
##     data = cluster_trial, id = practice, corstr = "independence",
##     scale.fix = T)
##
##  Coefficients:
##              Estimate Std.err    Wald Pr(>|W|)
## (Intercept)   28.390   2.324 149.250   <2e-16 ***
## treatment      0.420   2.474   0.029    0.865
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = independence
## Scale is fixed.
##
## Number of clusters:   10  Maximum cluster size: 3
```

```
mod1_exch <- geeglm(BMI ~ treatment,
                    family = gaussian(link = "identity"),
                    id = practice,
                    data=cluster_trial,
                    scale.fix = T,
                    corstr="exchangeable")

summary(mod1_exch)
```

```
##
## Call:
```

```
## geeglm(formula = BMI ~ treatment, family = gaussian(link = "identity"),
##     data = cluster_trial, id = practice, corstr = "exchangeable",
##     scale.fix = T)
##
##  Coefficients:
##             Estimate Std.err   Wald Pr(>|W|)
## (Intercept)   28.390   2.324 149.25   <2e-16 ***
## treatment      0.386   2.459   0.02     0.88
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Correlation structure = exchangeable
## Scale is fixed.
##
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.797   0.436
## Number of clusters:   10  Maximum cluster size: 3
```

```r
mod1_unstr <- geeglm(BMI ~ treatment,
                   family = gaussian(link = "identity"),
                   id = practice,
                   data=cluster_trial,
                   scale.fix = T,
                   corstr="unstructured")

summary(mod1_unstr)
```

```
##
## Call:
## geeglm(formula = BMI ~ treatment, family = gaussian(link = "identity"),
##     data = cluster_trial, id = practice, corstr = "unstructured",
##     scale.fix = T)
```

```
##
##  Coefficients:
##             Estimate Std.err   Wald Pr(>|W|)
## (Intercept)   28.390   2.324 149.25   <2e-16 ***
## treatment      0.922   5.158   0.03     0.86
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Correlation structure = unstructured
## Scale is fixed.
##
##   Link = identity
##
## Estimated Correlation Parameters:
##           Estimate Std.err
## alpha.1:2   0.9484   0.582
## alpha.1:3  -0.0875   0.642
## alpha.2:3   0.2882   1.263
## Number of clusters:   10  Maximum cluster size: 3
```

We’ll explain each of these in turn.

2 Working Correlation Matrices

As we can see above, generalized estimating equations are an extension of GLMs that require you specify a working correlation structure. This working correlation structure represents our assumption about how we think observations are correlated within each cluster. In our BMI example, this can be translated to “the way in which patients are correlated within each practice”, but this can be generalized to other correlated outcome scenarios (such as when repeated measures are taken on a single individual).

Here are some example correlation structures for an example that has a total of **three** measurements in each cluster (this can represent a measurement per week for four weeks on each individual in the sample, or four individuals within a clustering unit such as medical practices).

2.1 Independent Working Correlation Matrix

First, we have an **independent** working correlation structure. This structure assumes that there the correlation between any two people in a cluster is zero. Note that this is not the same as assuming that individuals in our data are independent. Instead, this states that observations within each cluster (in our case, individuals nested in distinct practices) share a **common variance**, and that variance can be different across clusters (e.g., practices). However, within each cluster, the independent working correlation matrix assumes that the correlation (covariance) is zero between observations.

$$\begin{pmatrix} 1 & 0 & 0 \\ & 1 & 0 \\ & & 1 \end{pmatrix}$$

The independent working correlation cluster is the simplest working correlation matrix. Notice, also, that if we compare the results we obtained from our GEE with an independent working correlation matrix, we get exactly the same results we obtained using the robust variance estimator:

```
library(lmtest)
library(sandwich)

mod1 <- glm(BMI ~ treatment,
            data = cluster_trial,
            family = gaussian(link = "identity"))

coeftest(mod1, vcov=vcovCL(mod1,
                           type = "HC0",
                           cadjust = F,
                           cluster = cluster_trial$practice))
```

```
##
## z test of coefficients:
##
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    28.39       2.32   12.22   <2e-16 ***
## treatment       0.42       2.48    0.17     0.87
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

This is important to note, since in it’s most fundamental representation, one can conceptualize a GEE as a generalized linear model with adjusted standard errors, where the adjustment is made via the robust variance estimator.

> **Deeper Dive**: Robust Variance and Generalized Estimating Equations
>
> There are a number of different versions of robust variance estimators, and also a number of different GEE versions. Between these sets, there is some overlap. That is, there is a version of the robust variance estimator that is equivalent to a version of GEE. In R, this overlap occurs when the cluster robust variance estimator is used (i.e., `vcovCL`) with the heteroscedastic consistent variance estimator `type = "HC0"` and no small sample adjustment is used `cadjust = F` AND when the GEE is fit with a fixed scale parameter (`scale.fix = T`) and an independent working correlation matrix (`corstr="independence"`).
>
> However, beyond these scenarios, the robust variance approach and GEE can differ in important and fundamental ways. In simple terms, the robust variance approach is relies on a simple *post hoc* variance correction strategy. That is, one fits a regression model, and then one adjusts the variance of the regression model using one of many “robust variance” adjustments. These adjustments typically rely on the distributions of outcome model residuals in the data (and are thus sometimes referred to as empircal variance estimators). The “*post hoc*” nature of this approach stems from the fact that the model fitting procedure is separate from the variance adjustment procedure (which happens after we fit the model).
>
> In contrast, the GEE approach relies on a similar *post hoc* variance correction strategy. But the estimation procedure itself (quasi-likelihood estimation), which is **not post hoc** also takes into consideration the assumed structure of the correlation within units. Because of this, parameter estimates (e.g., exposure effect estimates) from a GEE approach can be more efficient than those from a simple robust variance appraoch, especially if the working correlation within units is correct. The robust variance method can then be applied to the variance-covariance matrix after it is estimated using quasi-likelihood that takes into account the correlation structure (more below). However, there are tradeoffs to specifying a working correlation matrix other than independent. We cover these below.

2.2 Exchangeable or Compound Symmetric Working Correlation Matrix

Next is an example of an **exchangeable** or **compound symmetry** working correlation matrix. This structure assumes that the correlation between each and every individual in a cluster is exactly the same.

$$\begin{pmatrix} 1 & \rho & \rho \\ & 1 & \rho \\ & & 1 \end{pmatrix}$$

Note that this matrix requires that we estimate a single parameter. In our BMI data, this parameter is estimated to be:

```r
summary(mod1_exch)$corr$Estimate
```

```
## [1] 0.797
```

This tells us that the correlation matrix representing the relationship in BMI between all individuals in each practice in our data is:

$$\begin{pmatrix} 1 & 0.8 & 0.8 \\ & 1 & 0.8 \\ & & 1 \end{pmatrix}$$

In words, within all practices in our cluster trial, the correlation in BMI between individual 1 and 2, 1 and 3, and 2 and 3 is 0.8.

Important Note!: Recall the structure of our data: eight practices had two individuals in them, one practice had three individuals, and one practice had one individual. This is important to note. Our data are unbalanced, and thus the parameter of this working correlation matrix is being estimated for a 3x3 matrix where some observations only have one data point, and most only have two.

2.3 Unstructured Working Correlation Matrix

Finally, we have an example of an **unstructured** working correlation matrix. This allows the correlation between any two people in a cluster to be anything.

$$\begin{pmatrix} 1 & \rho_{1,2} & \rho_{1,3} \\ & 1 & \rho_{2,3} \\ & & 1 \end{pmatrix}$$

However, in this case, we estimate a unique parameter for every pair of units in the cluster. In our BMI data, these parameters are estimated to be:

```r
summary(mod1_unstr)$corr$Estimate
```

```
## [1]  0.9484 -0.0875  0.2882
```

This tells us that the correlation matrix representing the relationship in BMI between all individuals in each practice in our data is:

$$\begin{pmatrix} 1 & 0.95 & -0.1 \\ & 1 & 0.3 \\ & & 1 \end{pmatrix}$$

In words, within all practices in our cluster trial, the correlation in BMI between individual 1 and 2 is 0.9, 1 and 3 is -0.1, and 2 and 3 is 0.3. This is considerable variability in the correlation between BMIs across individuals within practices.

3 Selecting a Working Correlation Matrix

There is also considerable variability in the estimated correlations across working correlation structures. There is also some variability in the point estimates and standard errors from the regression model output, particularly between the model fit with the unstructured correlation matrix and the independent versus exchangeable:

```
summary(mod1_ind)$coefficients
```

```
##             Estimate Std.err     Wald Pr(>|W|)
## (Intercept)    28.39    2.32 149.2501    0.000
## treatment       0.42    2.47   0.0288    0.865
```

```
summary(mod1_exch)$coefficients
```

```
##             Estimate Std.err     Wald Pr(>|W|)
## (Intercept)   28.390    2.32 149.2501    0.000
## treatment      0.386    2.46   0.0247    0.875
```

```
summary(mod1_unstr)$coefficients
```

```
##             Estimate Std.err    Wald Pr(>|W|)
## (Intercept)   28.390    2.32 149.250    0.000
## treatment      0.922    5.16   0.032    0.858
```

So a natural question that arises is which one should we use? This is a difficult question to answer, because there is no definitive test we can use to tell us which is “best.” However, there are a few tools that we can use to evaluate how well different methods are fitting the data.

Generally, the validity of these results depends on correct model specification, specifically of the linear model when using GEE. Because our data were generated from a cluster randomized trial, this assumption is easier, but not trivial (in fact, because of the clustering), to justify.

3.1 Naive Variance GEE and Robust Variance GEE Comparison

The first tool relies on a comparison of:

(1) the variance-covariance matrix for the parameters in the GEE model that are obtained from only relying on the chosen working correlation structure to

(2) the variance-covariance matrix for the parameters that rely on both the chosen working correlation structure AND the *post hoc* robust variance correction.

For example, we can explore the variance-covariance matrix for the parameters from the GEE model assuming ONLY an independent working correlation matrix:

```
mod1_ind[["geese"]]$vbeta.naiv
```

```
##       [,1]  [,2]
## [1,]  1.63 -1.63
## [2,] -1.63  3.26
```

We can also explore the variance-covariance matrix for the parameters from the GEE model assuming BOTH an independent working correlation matrix, AND adjusting these variance-covariance entries for the parameters using the cluster robust variance estimator:

```
mod1_ind[["geese"]]$vbeta
```

```
##      [,1]  [,2]
## [1,]  5.4 -5.40
## [2,] -5.4  6.12
```

These two matrices are very different, which suggests that the independent working correlation matrix structure is not capturing a lot of the variability present in the data after the linear model portion is fit, and thus the robust variance correction is doing a lot of work here.

In contrast, if we look at the corresponding output for the exchangeable working correlation matrix, we see something different:

```
mod1_exch[["geese"]]$vbeta.naiv
```

```
##       [,1]  [,2]
## [1,]  2.93 -2.93
## [2,] -2.93  5.90
```

```
mod1_exch[["geese"]]$vbeta
```

```
##      [,1]  [,2]
## [1,]  5.4 -5.40
## [2,] -5.4  6.05
```

These variance covariance matrices are more similar to each other, suggesting that the exchangeable structure is more appropriate to our data.

We can systematize this comparison using a function in R. The following function takes the sum of the absolute difference between each entry in the variance covariance matrix of the parameters, giving us a single number representing how “close” the naive and robust corrected matrices are to each other:

```
# adapted from: https://unc.live/4aDjF0I

var.comp <- function(model_fit){
  sum(
    abs(model_fit[["geese"]]$vbeta.naiv -
            model_fit[["geese"]]$vbeta)
    )
}
```

```
sapply(list(mod1_ind, mod1_exch, mod1_unstr),
       var.comp)
```

```
## [1] 14.16  7.55 24.61
```

And this tells us that the exchangeable correlation matrix is fitting best.

3.2 Quasi Information Criterion

In standard regression settings, we may elect to choose a model based on a likelihood ratio test, or by comparing the Akaike Information Criterion (AIC) or the Bayesian Information Criterion. However, all these methods require that the model is built using likelihood methods.

Generalized estimating equations are not based on likelihood methods, but rather use what's referred to as a quasi-likelihood approach. However, this prevents us from using techniques such as likelihood ratio tests, AIC, or BIC for model selection.

In the early 2000s, Wei Pan generalized the AIC to the GEE setting (Pan, 2001). He called this generalization the "quasi-likelihood under the independence model criterion," or quasi information criterion (QIC). The QIC is meant to provide information on both the best correlation structure and the best subset of explanatory variables.

There are several versions of the QIC we can obtain in R. Using the QIC function on each of our three models, we obtain the following output:

```
QIC(mod1_ind, mod1_exch, mod1_unstr)
```

```
##            QIC QICu Quasi Lik   CIC params QICC
## mod1_ind   334  330      -163  3.75      2  336
## mod1_exch  334  330      -163  3.71      2  338
## mod1_unstr 362  333      -164 16.30      2  377
```

Generally, models with smaller values of QIC, CIC, QICu, or QICC are to be preferred. In our case, the independent and exchangeable working correlation matrices are not much different. Given the evidence that the exchangeable models are fitting better, we might read these criterion as suggesting the model with an exchangeable working correlation matrix is a good choice.

4 GEE: Final Conisderations

Generalized estimating equations are a very useful tool for estimating population average or marginal effects in epidemiologic data. One of the main strengths of the GEE approach is that the linear model and correlation structures are handled separately in the estimation process. As a result, in theory, you can choose the wrong correlation structure, but still get correct answers for the parameters of interest in the main model (e.g., the exposure effect). This has important consequences that are often highlighted in the technical literature on GEE. For example:

> " . . . even if an incorrect structure is used for the correlation matrix, . . . only the efficiency of our estimated β is affected." Hardin and Hilbe (2003), page 85.

> "An attractive property of the GEE is that one can use some working correlation structure that may be wrong, but the resulting regression coefficient estimate is still consistent and asymptotically normal." Pan and Connett (2002), abstract

> "GEE has two important robustness properties. First, the estimated regression coefficients obtained using GEE are broadly valid estimates that approach the correct value with increasing sample size regardless of the choice of correlation model. . . . Second, the correlation choice is used to obtain model-based standard errors and these do require that the correlation model choice is correct in order to use the standard errors for inference. A standard feature of GEE is the additional reporting of *empirical standard errors* which provide valid estimates of the uncertainty in the coefficients even if the correlation model is not correct. Therefore, the correlation model can be any model, including one that assumes observations are independent, and proper large sample standard errors obtained using the empirical estimator." van Belle et al. (2004) page 754-5

However, there are two important pieces of context that need to be clarified with respect to these statements.

First, these are describing an *asymptotic* property, and not something that we expect to materialize in finite samples. As a case in point, the point estimates for the coefficient in our model fit with an independent, exchangeable, and unstructured working correlation matrix were different. If choosing the working correlation matrix didn't matter here, we'd expect to see no change in the point estimate. Certainly, our limited sample size is playing an important role in estimating different treatment effects under different correlation structures. But variation in point estimates across correlation structures is not an uncommon situation, even in moderate to large datasests.

Second, and more importantly, these statements assume a *correctly specified mean model*. If the linear model is not correctly specified, then all bets are off. Thus, the “robustness” property of GEE that is often raised as a major strength of the approach is a second order robustness property. You should not dupe yourself into thinking that the point estimates you obtain from GEE are “robust” because you used GEE. The robustness features in the GEE framework apply to the standard errors, and depend on correct specification of the linear model.

5 BMI Cluster Trial Data

We’ve looked at different ways of analyzing our BMI cluster trial data that included standard regression (ignoring the clustering), standard regression with the robust variance estimator, standard regression with the clustered bootstrap, and three versions of GEE. The results from all of these analyses are as follows:

Version	Estimate	Std.Err	LCL	UCL
Uncorrected	0.42	1.90	-3.31	4.15
Cluster Robust	0.42	2.47	-4.43	5.27
Cluster Bootstrap	0.42	2.68	-4.83	5.67
GEE Independent	0.42	2.47	-4.43	5.27
GEE Exchangeable	0.39	2.46	-4.43	5.21
GEE Unstructured	0.92	5.16	-9.19	11.03

Which method would be the ideal method in this situation? It’s always difficult to answer this question in any setting, but we can make the following judgements on the basis of our analysis:

1) The uncorrected analysis is unsuitable because it assumes independence which is clearly wrong. This information comes from our understanding of the study design, as well as our estimates of the ICC coefficients in the data.

2) The unstructured correlation, independent correlation and (by equivalence) the robust variance correction to the standard model can be deemed less suitable as the result of our comparison of naive and corrected GEE variances and (less so) our QIC analysis. Furthermore, for sample size reasons,

the unstructured model is unlikely to be a good idea, since we are estimating many more parameters, and it's hard to justify this correlation structure in our setting.

3) This leaves us with the GEE under an exchangeable working correlation matrix, and the clustered bootstrap analysis. Which should we choose?

Well, we can present both, and this would be a good idea. We should also note that, in this case, it doesn't really matter. The two estimates, though numerically different, are statistically identical.

Finally, and perhaps most importantly, we should note that this study simply did not have enough information available for us to make conclusions about the treatment effect. Note that this is **not reason to interpret that their is no effect of patient centered care versus normal care on BMI**. There is simply not enough information in these data for us to be able to make a conclusion.

References

James W Hardin and Joseph M Hilbe. *Generalized estimating equations*. Chapman & Hall/CRC, Boca Raton, Fla., 2003.

Kung-Yee Liang Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

Wei Pan. Akaike's information criterion in generalized estimating equations. *Biometrics*, 57(1):120–125, 2001.

Wei Pan and John E. Connett. Selecting the working correlation structure in generalized estimating equations with application to the lung health study. *Statistica Sinica*, 12(2):475–490, 2002.

G. van Belle, L. D. Fisher, P. J. Heagerty, and T. Lumley. *Biostatistics: A Methodology For the Health Sciences*. Wiley Series in Probability and Statistics. Wiley, 2004.