Generalized Estimating Equations

Kerby Shedden

Department of Statistics, University of Michigan

December 6, 2021

Score equations

Suppose we have multivariate Gaussian data with mean structure $E[y|X] = X\beta$ and covariance structure $\Sigma \in \mathcal{R}^{n \times n}$.

Using generalized least squares, we estimate β by minimizing

$$(y - X\beta)'\Sigma^{-1}(y - X\beta).$$

The minimizer solves the following system of equations for β:

$$X'\Sigma^{-1}(y - X\beta) = 0$$

These are the score equations $\ell'(\beta; y, X) = 0$ for the Gaussian log likelihood. They also imply orthogonality between the residuals and $\mathrm{col}(X)$ in the metric of Σ^{-1}.

Score equations for clustered data

Now suppose the sample can be partitioned into m clusters of sizes $n_1, \ldots, n_m$ such that there is no dependence between clusters. Let $y_i \in \mathcal{R}^{n_i}$ and $X_i \in \mathcal{R}^{n_i \times p}$ denote the components of y and X containing the data for the i^{th} cluster, and let $\Sigma_i = \text{Cov}[y_i|X_i]$.

Then

$$(y - X\beta)'\Sigma^{-1}(y - X\beta) = \sum_i (y_i - X_i\beta)'\Sigma_i^{-1}(y_i - X_i\beta),$$

and the score equations can be written

$$\sum_i X_i'\Sigma_i^{-1}(y_i - X_i\beta) = 0_{p+1}.$$

Can we obtain score equations of this form for dependent data in a GLM framework?

Score equations for non-Gaussian data

Recall that the GLM score function in terms of the canonical parameter θ is $y - \gamma'(\theta) \in \mathcal{R}^n$.

Write $\eta_i = X_i\beta \in \mathcal{R}^{n_i}$ for the linear predictor for cluster i, and $\theta_i \in \mathcal{R}^{n_i}$ for the canonical parameter in cluster i. Then use the chain rule to differentiate with respect to β:

$$\sum_i \frac{\partial \theta_i}{\partial \beta}(y_i - \gamma'(\theta_i)) = 0_{p+1}.$$

which in turn implies that

$$\sum_i \frac{\partial \eta_i}{\partial \beta}\frac{\partial \theta_i}{\partial \eta_i}(y_i - \gamma'(\theta_i)) = \sum_i X_i^T \frac{\partial \theta_i}{\partial \eta_i}(y_i - \gamma'(\theta_i)) = 0_{p+1}.$$

Score equations for non-Gaussian data

Using GLM notation, $\theta_i = g(\eta_i)$, so $\partial\theta_i/\partial\eta_i = g'(\eta_i)$.

Since the mean satisfies $\mu_i = \gamma'(g(\eta_i))$, it follows that

$$\begin{aligned} \frac{\partial \mu_i}{\partial \beta} &= \frac{\partial \eta_i}{\partial \beta} \cdot \gamma''(g(\eta_i))g'(\eta_i) \\ &= X_i^T V_i(\eta_i)g'(\eta_i)/\phi \end{aligned}$$

where $V_i(\theta) = \mathrm{Var}[y_i|X_i] \in \mathcal{R}^{n_i \times n_i}$ is the variance function.

Putting everything together yields

$$\sum_i \frac{\partial \mu_i}{\partial \beta} V_i^{-1}(y_i - \mu_i) = 0_{p+1}.$$

Score equations for non-Gaussian data

These score equations are a system of $p+1$ equations in $p+1$ variables, so if they are not degenerate, they should uniquely determine β.

In the Gaussian case, $\partial \mu_i / \partial \beta = X_i^T$, and $V_i \propto 1$, thus the score equations are equivalent to requiring that the residuals be orthogonal to every column of X.

These score estimating equations are a generalization to this condition for non-linear models.

Score equations for dependent non-Gaussian data

Suppose the data are dependent within clusters, and we block the data so that $y_i \in \mathcal{R}^{n_i}$ is the response data for cluster i. By analogy with GLM's, we can propose the following score equations:

$$\sum_i D_i' \Sigma_i^{-1} (y_i - \mu_i) = 0.$$

where $D_i = \partial E[y_i|X_i]/\partial \beta' = X_i^T \partial \mu_i / \partial \eta_i$, where $\mu_i = E[y_i|X_i]$ and $\partial \mu_i / \partial \eta_i$ is a diagonal matrix.

The variance matrix can be modeled as

$$\Sigma_i = V_i^{1/2} R_i(\alpha) V_i^{1/2}$$

where V_i is the diagonal matrix determined by a GLM variance function, and $R_i(\alpha)$ is a correlation model determined by a separate parameter α.

Score equations for non-Gaussian data

We can easily calculate D_i for any of the familiar generalized linear model link functions. For example, in the logistic model

$$\mu_i = 1/(1 + \exp(-X_i\beta)),$$

$$\partial\mu_i(j)/\partial\beta' = \frac{\exp(X_i(j,:)\beta)}{(1 + \exp(-X_i(j,:))\beta)^2} X_i(j,:) = \mu_i(j)(1-\mu_i(j))X_i(j,:).$$

Mahalanobis distance minimization

Now we will essentially start over and show that we can obtain the same estimating equations (for dependent data and nonlinear models), using a different approach.

Suppose we model the mean structure as $E[y_i|X_i] = \mu_i = \mu_i(\beta)$.

The Mahalanobis distance between the data and the fitted means is

$$\sum_i (y_i - \mu_i)^T \Sigma_i^{-1} (y_i - \mu_i).$$

The gradient of this funtion is

$$-2 \sum_i \frac{\partial \mu_i}{\partial \beta} \Sigma_i^{-1} (y_i - \mu_i)$$

Solving the score equations

Note that in general D_i depends on β, and Σ_i may also depend on β.

To solve the score equations, linearize the mean structure

$$\mu_i \approx \mu_i^{(0)} + D_i^{(0)}(\beta - \beta^{(0)})$$

Using Gauss-Seidel iterations, we substitute the linearized mean structure into the score equations, updating $\hat{\beta}$ for the current iteration using D_i, Σ_i, and μ_i calculated from the value of $\hat{\beta}$ at the previous iteration.

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + (\sum_i D_i' \Sigma_i^{-1} D_i)^{-1} \sum_i D_i \Sigma_i^{-1} (y_i - \mu_i)$$

Working covariance structures

If the Σ_i are misspecified, $\hat{\beta}$ is still $\sqrt{n}$-consistent (under similar regularity conditions as when Σ_i is known).

We can therefore choose a working correlation $R_i(\alpha)$ when fitting the model.

We then set $\Sigma_i = V_i^{1/2} R_i(\alpha) V_i^{1/2}$, where V_i is the diagonal matrix containing the variance values.

Working covariance structures

Using a "sandwich covariance estimate" gives correct inferences even when the working correlation structure is wrong.

First, define

$$A \equiv \sum_j D_i' \Sigma_i^{-1} D_i$$

$$B \equiv \sum_i D_i' \Sigma_i^{-1} r_i r_i' \Sigma_i^{-1} D_i$$

where $r_i = y_i - \hat{\mu}_i$ are the residuals for cluster i.

Working covariance structures

Note that when the working covariance structure is correct, $E[r_i r_i'] = \Sigma_i$, and B becomes

$$B = \sum_i D_i' \Sigma_i^{-1} D_i.$$

The following covariance approximation holds even when the working correlation matrices R_i are misspecified:

$$\mathrm{Cov}[\hat{\beta}] \approx A^{-1} B A^{-1}.$$

Working covariance structures

Why does it work?

We'll focus on the linear case where

$$\begin{aligned}\hat{\beta} &= (\sum_i X_i' \Sigma_i^{-1} X_i)^{-1} \cdot \sum_i X_i' \Sigma_i^{-1} y_i \\ &= B^{-1} \sum_i X_i' \Sigma_i^{-1} y_i\end{aligned}$$

where in the case of a linear model,

$$B = \sum_i D_i' \Sigma_i^{-1} D_i = \sum_i X_i' \Sigma_i^{-1} X_i.$$

Working covariance structures (why does it work?)

$$\mathrm{Cov}[\hat{\beta}] = B^{-1}\left(\sum_i X_i'\Sigma_i^{-1}\mathrm{Cov}[y_i|X_i]\Sigma^{-1}X_i\right)B^{-1}.$$

Since $E[r_i r_i'] \approx \mathrm{Cov}[y_i|X_i]$, the "sandwich expression" is a plug-in estimate of the covariance matrix given above.

Working correlation structures

The working covariance is usually defined in terms of a working correlation structure. Examples include:

- Independence: $R_i = I$
- Exchangeable: $R_i(j, j) = 1$, $R_i(j, j') = r$ if $j \neq j'$
- Autoregressive: $R_i(j, j') = r^{|j-j'|}$.
- Many more...

We then have as the working covariance $\Sigma_i = V_i^{1/2} R_i(\alpha) V_i^{1/2}$, where $V_i = \mathrm{diag}(\mathrm{Var}[y_i|X_i])$.

Working correlation structures

These variances are often specified in terms of the mean via a mean/variance relationship, e.g. for a Poisson model we have $\mathrm{Var}[y_i|X_i] = E[y_i|X_i]$.

To implement this, we need to devise a way to update R_i based on the residuals $r_i = y_i - \hat{\mu}_i$. Usually these updates are based on the method of moments. These updates alternate with the Gauss-Seidel iterations until convergence.

Some applications

- Clustered data: Cases are families, classrooms, coworkers, etc., possibly multiply nested; may use exchangeable working correlation.
- Longitudinal data: Mean structure + serial dependence structure (e.g. autoregressive, stationary autocovariance).
- Spatial data: May use any spatial covariance function; may have only one cluster.

Limitations/criticisms of GEE

- Estimates the marginal mean structure parameters
- Does not follow the likelihood principle
- Details are not very intuitive
- May not be fully efficient
- No formal inference for variance parameters (seldom-used GEE2 allows this)
- May be inconsistent in some longitudinal settings with time-dependent covariates, or in certain settings with missing data
- Does not propagate uncertainty of dependence parameters to uncertainty in regression coefficient estimates.

Limitations/criticisms of GEE (continued)

- Some familiar likelihood-based tools like the LRT and AIC are not available (but modified versions like QIC exist)
- Iterations sometimes may not converge, or multiple solutions to score equations may exist; not clear if convergence is necessary, or how many iterations shoud be performed
- Difficult to simulate data that exactly matches the model

Positive aspects of GEE analysis

- Focuses on the mean structure, where the main interest usually lies
- "Robust" to misspecification of things other that the mean structure
- Builds on familiar GLM analysis, results are consistent with studies that have an independent subjects design.
- Can be quite fast to fit (depends on details of covariance structure update)

Comparison to multilevel (random effects) models

The generalized linear mixed effects (GLIMIX) model for single-level clustered data is

$$\gamma_1, \ldots, \gamma_m \overset{iid}{\sim} N(0, \Phi)$$

where $\gamma_i \in \mathcal{R}^q$ and $\Phi \in \mathcal{R}^{q \times q}$.

The observed responses $y_{ij}|\gamma_i$ are independent over i and j, and follow a GLM with linear predictor

$$x_{ij}\beta + z_{ij}\gamma_i$$

where $x_{ij} \in \mathcal{R}^p$ and $z_{ij} \in \mathcal{R}^q$.

Comparison to multilevel (random effects) models

In the case of a linear model, we can average over γ to obtain the marginal model:

$$E[y_{ij}|\gamma_i] = x_{ij}\beta + z_{ij}\gamma_i,$$

and

$$E[y_{ij}] = x_{ij}\beta.$$

Thus, the coefficients for x in the multilevel model are the same as the coefficients for x in the marginal model.

In the linear case, the multilevel and marginal models are both linear, and Gaussian, and β has the same meaning for both representations.

Comparison to multilevel (random effects) models

For any nonlinear GLM, the marginal model is not a GLM.

For example, the marginal form for a binomial mixed GLM with logit link is:

$$\int \prod_j \frac{\exp(y_{ij} \cdot (x_{ij}\beta + z_{ij}\gamma_i))}{1 + \exp(x_{ij}\beta + z_{ij}\gamma_i)} \phi(\gamma_i) d\gamma_i$$

This integral cannot be solved analytically, and is not equal to a binomial GLM.

Comparison to multilevel (random effects) models

In a GEE analysis, the marginal distributions are GLM's and the joint distribution is not specified.

In a multilevel model, the conditional distributions factor as independent GLM's. The marginalization is not tractable, and the marginal distributions are not GLM's.

Comparison to multilevel (random effects) models

There is an important difference between the interpretation of the regression coefficients in a GEE (marginal GLM) and in a multilevel model (conditional GLM).

Let's focus on the interpretation of a coefficient β_1 in a binomial model, where the corresponding covariate x_{ij1} is either 0 or 1.

In the multilevel (conditional) model, β_1 is the log odds ratio between y_{ij} and x_{ij1}, conditioned on all other covariates, and also conditioned on the random effect γ_i.

An alternative way to get approximately the same log odds ratio β_1 would be to calculate the sample log odds ratio within each cluster, then average the corresponding odds ratios over the clusters. This procedure is known as the Mantel-Haensel estimate of the common odds ratio.

Comparison to multilevel (random effects) models

In the GEE (marginal model), β_1 is the log odds ratio between y_{ij} and x_{ij1}, pooling over all clusters (i.e. ignoring the clusters).

In general, the marginal odds ratio (given by GEE) is closer to 1 than the conditional odds ratio given by multilevel modeling.

The marginal parameters estimated by GEE are sometimes referred to as population averaged effects, while the conditional parameters estimated in multilevel modeling are referred to as individual effects.

Comparison to linear mixed effects models

Some comments about the relationship between these two frameworks:

- The marginal (unconditional) mean structure $E[y_{ij}|x_{ij}]$ usually cannot be expressed in closed form, and is not a GLM or exponential family in general.
- Variance parameters (Φ) also affect the mean structure.
- Fitting requires numerical integration, sometimes over a high dimensional domain.
- Difficult to assess fit of random effects distribution, almost always taken to be Gaussian in practice
- Estimates and inference for mean structure parameters are in general not robust to misspecification of the variance structure parameters.
- Can test variance parameters just like any other parameters.