GENERALIZED LINEAR MARGINAL MODELS FOR LONGITUDINAL DATA

Outline:

Generalized linear marginal models and GEE

- Generalized linear models for categorical outcomes
- Correlation structure for GLM
- GEE: Quasi-likelihood Estimation and Robust SE estimator
- (Review) Inferences in the Mean Model
- Criterion for Selecting GEE Models
- GEE may be biased with time-varying covariates

Generalized Linear Models

- In a generalized linear model (GLM), focus is on the **mean response** of Y_{ij} as a function of covariates $\boldsymbol{x}_{ij}$:

$$\mathrm{E}(Y_{ij}|\boldsymbol{x}_{ij}) = \mu_{ij}$$

- $\boldsymbol{\beta}$ represents cross-sectional association between outcomes Y and covariates X.

- The mean response has a **population average** interpretation:

  For a given covariate value $\boldsymbol{x}_{ij}$, choose a random subject i and a random response j from that subject (with covariate value $\boldsymbol{x}_{ij}$),

  then μ_{ij} is the average value of Y_{ij} among all those observations

- Special case: **linear model** for longitudinal data:

$$\mathrm{E}(Y_{ij}|\boldsymbol{x}_{ij}) = \mu_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta}$$

  $\rightarrow$ Recall this is the **marginal mean model**

- How about non-continuous data:
  - Binary Y? (logistic regression)
  - Count Y? (Poisson regression)
- Key components of a **generalized linear model** for a response Y_{ij}
  - **Linear predictor:** (same as for the linear model)

$$\eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta}$$

  - **Link function:** Mean is connected to linear predictor via link function $h(\cdot)$:

$$h(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta}$$

  - Special case: Linear model has identity link function $h(\mu) = \mu$
  - We can use other links (eg, log)
- The variance of Y_{ij} is expressed as a function of the mean:

$$\text{var}(Y_{ij}|\boldsymbol{x}_{ij}) = \phi v(\mu_{ij})$$

  - $v(\cdot)$ is called the **variance function**

  - ϕ is dispersion/scale parameter (some software report scale as $\sqrt{\phi}$)

- Note: We do not assume distribution for Y
- These components are cross-sectional data models
- For independent data: $Y_{ij}|\boldsymbol{X}_i$ is independent to $Y_{ij'}|\boldsymbol{X}_i$
- For longitudinal data: Correlation structure of Y needs to be accounted for

Example (Binary Response)
Logistic Regression Model

- Response $Y_{ij} = 1$ or 0 indicating presence or absence of some characteristic for subject i at time j:

  Example: $Y_{ij} = I(\text{breastfeeding}_{ij})$ in the data set on Nepalese children

- Note that

$$\mu_{ij} = \mathrm{E}(Y_{ij}|\boldsymbol{x}_{ij}) = \Pr(Y_{ij} = 1) \in (0, 1)$$

$$\eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} \in (-\infty, \infty)$$

- Link function is the "logit":

$$\text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = \text{"log odds"}$$

so that

$$h(\mu_{ij}) = \text{logit}(\mu_{ij}) = \log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta}$$

- This means that

$$\text{E}(Y_{ij}|\boldsymbol{x}_{ij}) = \mu_{ij} = \frac{1}{1+e^{-\eta_{ij}}}$$

and

$$\frac{\mu_{ij}}{1-\mu_{ij}} = e^{\boldsymbol{x}'_{ij}\boldsymbol{\beta}} = \text{“odds”}$$

Example: In a logistic regression of breastfeeding on age of the child (in years)

– $e^{\beta_{\text{age}}}$ is odds ratio of breastfeeding comparing two children who differ in age by one year

- If $Y \sim$ binomial distribution, the variance is
  $\text{var}(Y_{ij}|\boldsymbol{x}_{ij}) = \Pr(Y_{ij} = 1)\{1 - \Pr(Y_{ij} = 1)\} = \mu_{ij}(1 - \mu_{ij})$, so
  – $\phi = 1$
  – $v(\mu) = \mu(1 - \mu)$

- Such a model can be used to answer (**population average**) questions such as:
  1. What is the prevalence of breastfeeding in the population of Nepalese mothers and children?
  2. How does the prevalence of breastfeeding vary by age in this population?
  3. Does the prevalence of breastfeeding vary by the socio-economic status of the mother in this population?
- Longitudinal data can provide more statistically efficient answers for certain questions, such as 2
  – because longitudinal data are especially powerful for detecting changes over time
- However, we will need a correlation model in order to take advantages offered by longitudinal data.

Example (Count Data)
Log-linear Regression Model (Poisson Regression)

- Response Y_{ij} is a count for subject i at time j:
  1. number of seizures at each of a series of two week periods
  2. number of packs of cigarettes smoked/week, over a series of two-week periods

- Link function is the "log":

$$h(\mu_{ij}) = \log(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta}$$

- This means that

$$\mathrm{E}(Y_{ij}|\boldsymbol{x}_{ij}) = \mu_{ij} = e^{\eta_{ij}} = e^{\boldsymbol{x}'_{ij}\boldsymbol{\beta}} = \text{"risk" or "rate"}$$

  – β's are **log risk ratios** or **log rate ratios** relating the x_{ijl}'s to the response Y_{ij}

Example: In a log-linear regression model of the number of seizures on treatment (0 or 1):
– $e^{\beta_{\texttt{treat}}}$ is rate ratio for number of seizures per week for a subject on treatment versus a subject not on treatment

- The variance is

$$\text{var}(Y_{ij}|\boldsymbol{x}_{ij}) = \phi\mu_{ij}$$

– $v(\mu) = \mu$
– ϕ can be $\neq 1$

- This model is sometimes called "Poisson regression" model because:
If Y_{ij} is Poisson, we have
– mean μ_{ij}
– variance μ_{ij}
It is a special case with $\phi = 1$

- The model is more appropriately called a "log-linear" model because we allow $\phi \neq 1$ and estimate it (often $\phi > 1$):
  - this is called **overdispersion**
  - we can also call this model **Overdispersed Poisson Regression**
- This is a model for cross-sectional data. To apply it to longitudinal data, we will further require a correlation model

Correlation for Marginal Models

- Recall that the response is **longitudinal**:

$$\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$$

- In linear models, marginal models consisted of a model for the mean

$$\mathrm{E}(\boldsymbol{Y}_i|X_i) = \boldsymbol{\mu}_i \quad (\boldsymbol{\mu}_i \text{ is the mean vector})$$

and a model for the variance

$$\mathrm{var}(\boldsymbol{Y}_i|X_i) = V_i \quad (V_i \text{ is the variance-covariance matrix for subject } i)$$

- With marginal generalized linear models, variance is more closely linked to the mean function (via $v(\cdot)$) and $\boldsymbol{\beta}$

$$\mathrm{var}(Y_{ij}|\boldsymbol{x}_{ij}) = \phi v(\mu_{ij})$$

- What is left to specify is the **correlation** between two responses on the same subject:

$$\rho_{jk} = \text{corr}(Y_{ij}, Y_{ik})$$

- We specify the correlation using a vector of parameters $\boldsymbol{\alpha}$:

$$\text{corr}(Y_{ij}, Y_{ik}) = \rho(j, k, ; \boldsymbol{\alpha}) = \rho_{jk}(\boldsymbol{\alpha})$$

  – For simplicity, we use ρ_{jk}

- Then the covariance between two responses on the same subject is:

$$\text{cov}(Y_{ij}, Y_{ik}) = \text{corr}(Y_{ij}, Y_{ik})\sqrt{\text{var}(Y_{ij})\text{var}(Y_{ik})} = \phi\rho_{jk}\sqrt{v(\mu_{ij})v(\mu_{ik})}$$

- (Matrix form) For subject i, variance-covariance matrix of $\boldsymbol{Y}_i$ is

$$V_i = V_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi) = \mathrm{var}(\boldsymbol{Y}_i) = \phi A_i^{1/2} C_i A_i^{1/2}$$

where

$$A_i = A_i(\boldsymbol{\beta}) = \begin{pmatrix} v(\mu_{i1}) & 0 & \cdots & 0 \\ 0 & v(\mu_{i2}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & v(\mu_{in_i}) \end{pmatrix}$$

and

$$C_i = C_i(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1n_i} \\ \rho_{21} & 1 & \cdots & \rho_{2n_i} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n_i1} & \rho_{n_i2} & \cdots & 1 \end{pmatrix}$$

- Note that

$$A_i^{1/2} = A_i^{1/2}(\boldsymbol{\beta}) = \begin{pmatrix} \sqrt{v(\mu_{i1})} & 0 & \cdots & 0 \\ 0 & \sqrt{v(\mu_{i2})} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{v(\mu_{in_i})} \end{pmatrix}$$

Correlation Model Examples

- Exchangeable / Compound Symmetry:

$$\rho_{jk} = \alpha$$

so that

$$C_i = \begin{pmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \cdots & 1 \end{pmatrix}$$

- AR-1 / Exponential

$$\rho_{jk} = \alpha^{|j-k|}$$

so that

$$C_i = \begin{pmatrix} 1 & \alpha & \cdots & \alpha^{n_i-1} \\ \alpha & 1 & \cdots & \alpha^{n_i-2} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha^{n_i-1} & \alpha^{n_i-2} & \cdots & 1 \end{pmatrix}$$

– A more general AR with continuous time: $\rho_{jk} = \alpha^{|t_{ij}-t_{ik}|}$ (C_i will be different for each subject i)

- M-dependent (1):

$$C_i = \begin{pmatrix} 1 & \alpha & 0 & \cdots & 0 \\ \alpha & 1 & \alpha & \cdots & 0 \\ 0 & \alpha & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

Review ML Estimation for GLM with Independent Data

- Assume independent Y_i $(i = 1, \ldots, m)$ follows a scaled exponential distribution

$$f(y; \theta, \phi) = exp\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\}$$

  – includes Gaussian, Binomial, Poisson, Gamma, Inverse Gaussian distributions as special cases.

- We have

$$\mathrm{E}(Y) = b'(\theta) = \mu$$

$$\mathrm{var}(Y) = a(\phi)b''(\theta) = a(\phi)v(\mu)$$

  with link function

$$h(\mu) = \eta = \boldsymbol{x}'\boldsymbol{\beta}$$

- Log-likelihood function for Y_i:

$$l(\boldsymbol{\beta}, \theta) = \sum_i \left\{ \frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i, \phi) \right\}$$

- Score equation is

$$\begin{aligned} \frac{\partial l}{\partial \boldsymbol{\beta}} &= \frac{\partial l}{\partial \theta}\frac{\partial \theta}{\partial \mu}\frac{\partial \mu}{\partial \boldsymbol{\beta}} \\ &= \sum_i \frac{y_i - \mu_i}{a(\phi)}\frac{1}{v(\mu_i)}\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = 0 \end{aligned}$$

- So estimating equation for $\boldsymbol{\beta}$ is

$$S(\boldsymbol{\beta}) = \sum_i \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} v(\mu_i)^{-1}(y_i - \mu_i) = 0$$

- (Special case) For linear model:

$$S(\boldsymbol{\beta}) = \sum_i \boldsymbol{x}_i' v(\mu_i)^{-1}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}) = 0$$

  $\rightarrow$ OLS or WLS/GLS

- Distribution assumption is actually not necessary:
  We only need $\mu_i = \mathrm{E}(y_i)$ and $v(\mu_i) = \mathrm{var}(y_i)/a(\phi)$ in

$$S(\boldsymbol{\beta}) = \sum_i \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} v(\mu_i)^{-1}(y_i - \mu_i) = 0$$

- We have $\mathrm{E}(y_i - \mu_i) = 0$
  $\Rightarrow \mathrm{E}(S(\boldsymbol{\beta})) = 0$
  $\Rightarrow$ Consistent estimator for $\boldsymbol{\beta}$ without knowing actual distribution
  (with regularity conditions)

- Thus, we call it Quasi-likelihood (QL) estimation

Quasi-likelihood Estimation for Correlated Data

- The marginal model we have specified consists of models for:
  - Mean $\mathrm{E}(\boldsymbol{Y}_i)$ $\leftarrow$ link and linear predictor
  - Variance $\mathrm{var}(\boldsymbol{Y}_i)$ $\leftarrow$ variance function, dispersion parameter ϕ
  - Correlation model
- We have not specified the "full likelihood"
  - We do not specify the **distribution function** of $\boldsymbol{Y}_i$
- To do maximum likelihood: need assume a distribution (eg, Binomial and Poisson)
- However, **Quasilikelihood** is an excellent alternative.

- QL estimation works by solving the QL equations

$$S(\boldsymbol{\beta}) = \sum_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = 0$$

  for $\boldsymbol{\beta}$, where

$$V_i = V_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi) \quad \text{and} \quad \boldsymbol{\mu}_i = \boldsymbol{\mu}_i(\boldsymbol{\beta})$$

  No need to specify distribution!

- **Example (Linear Models)**:

$$\boldsymbol{\mu}_i = X_i\boldsymbol{\beta}$$

so

$$\frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} = X_i'$$

and the QL equations are

$$\sum_i X_i' V_i^{-1}(\boldsymbol{y}_i - X_i\boldsymbol{\beta}) = 0$$

which is equal to (because V is block diagonal)

$$X'V^{-1}(\boldsymbol{y} - X\boldsymbol{\beta}) = 0,$$

The solution $\widehat{\boldsymbol{\beta}}$ is then:

$$(X'V^{-1}X)^{-1}X'V^{-1}\boldsymbol{y} = \widehat{\boldsymbol{\beta}}$$

which is the **weighted least squares estimator**

- Key differences for GLM here:
  - $\boldsymbol{\mu}_i$ is **not** linear in $\boldsymbol{\beta}$
  - so that $\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$ is not simply X_i and depends on $\boldsymbol{\beta}$
  - V_i depends on $\boldsymbol{\beta}$ (as well as $\boldsymbol{\alpha}$ — we'll get to that)

$\Rightarrow$ No closed solution. Solving the QL equations requires iteration

- Why does QL estimation work (i.e., is approximately unbiased)?
  - The main reason is that the QL equation is **unbiased**:

$$\mathrm{E}\left\{\sum_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)|X_i\right\} = \sum_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1}\{\mathrm{E}(\boldsymbol{Y}_i|X_i) - \boldsymbol{\mu}_i\} = 0$$

  - There are some other technical conditions which must be satisfied as well
  - As a result of the unbiasedness of the QL equation, the solution $\boldsymbol{\beta}$ is **consistent** and asymptotically Normal for the true value $\boldsymbol{\beta}$ as number of subjects $m \to \infty$
    * Consistency replaces the unbiasedness of $\widehat{\beta}$ that we had in the linear models
    * Consistency means that $\widehat{\beta}$ gets closer and closer to the true value β as m gets large

Comments:

- For any chosen weight matrix W_i, the estimating equation

$$S(\boldsymbol{\beta}) = \sum_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} W_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = 0$$

  will give rise to a consistent estimator for $\boldsymbol{\beta}$

- If $W_i = V_i$, the estimator will be the most efficient.

- Models describe how the marginal mean of Y_{ij} at the jth occasion varies as a function of covariates, but don't describe how the means of one occasion depends on past values of the outcome
  – Eg, asthma rates of children at a given time t may depend on if the children experience an asthma episode at time $t-1$.

- Covariates X_{it} may include covariates recorded before time t, their difference, average, etc.

Variance Estimation for $\widehat{\boldsymbol{\beta}}$ under QL

- For the QL estimator $\widehat{\boldsymbol{\beta}}$,

$$\text{var}(\widehat{\boldsymbol{\beta}}) = (\sum_i \widetilde{D}_i' V_i^{-1} \widetilde{D}_i)^{-1}$$

  where

$$\widetilde{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \quad \text{(replaces } X_i \text{ in linear models)}$$

  This is sometimes called the **model-based** variance estimator (note: $\widetilde{D}_i$ here is partial derivative, **not** the same as D_i from random effects models!)

- **Conclusion:** If we can specify the mean and the variance-covariance model for the response $\boldsymbol{Y}_i$:
  – Enough to make QL inferences for $\boldsymbol{\beta}$
  – No need to specify distribution

- **Problem:** It might be hard to get the variance-covariance model correct, and this is especially true with categorical data:
  - the variance function $v(\mu_{ij})$ might be wrong,
    e.g., we specify
    $$\text{var}(Y_{ij}) = \phi\mu_{ij}$$
    but really
    $$\text{var}(Y_{ij}) = \phi\sqrt{\mu_{ij}}$$
  - ϕ might not be constant across all X_i
  - the correlation model (structure) used for C_i might not be right,
    e.g., we specify exchangeable, but it is really unstructured
  - the correlation structure is correct, but the value of parameter $\boldsymbol{\alpha}$ varies by X_i,
    e.g., for subjects with large X_i, C_i is exponential with $\alpha = 0.2$, while for subjects with small X_i, $\alpha = 0.4$

- So, now we are "at risk" of getting V_i wrong.
- Is QL estimator still good?
- (Under some regularity conditions) If the model for the mean $\boldsymbol{\mu}_i$ is correct,

$$\sum_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$$

  is **still unbiased** even if V_i is misspecified

  – Analogous to WLS in linear models: Unbiased $\widehat{\beta}$ even if the variance-covariance model is wrong.

- Conclusion: $\widehat{\boldsymbol{\beta}}$ is **still consistent** for $\boldsymbol{\beta}$ even if we get the correlation or variance models wrong!

- What about $\mathrm{var}(\widehat{\boldsymbol{\beta}})$?

  – If V_i is wrong, model-based $\widehat{\mathrm{var}}(\widehat{\beta}) = (\sum_i \widetilde{D}_i' V_i^{-1} \widetilde{D}_i)^{-1}$ is also wrong

- We might call V_i a "working variance" and C_i a "working correlation"

- Instead we have **sandwich** form:

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\sum_i \widetilde{D}_i' V_i^{-1} \widetilde{D}_i)^{-1} (\sum_i \widetilde{D}_i' V_i^{-1} \mathrm{var}(\boldsymbol{Y}_i) V_i^{-1} \widetilde{D}_i) (\sum_i \widetilde{D}_i' V_i^{-1} \widetilde{D}_i)^{-1}$$

  (same form as for linear models, but with $\widetilde{D}_i$ replacing X_i)

- For generalized linear models, $\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}})$ can be obtained via the Huber-White method too

- define

$$\boldsymbol{\epsilon}_i = \boldsymbol{y}_i - \boldsymbol{\mu}_i$$

and note that

$$\mathrm{var}(\boldsymbol{\epsilon}_i) = \mathrm{var}(\boldsymbol{Y}_i|X_i) = \mathrm{E}(\boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i')$$

(note: sometimes the notation $\boldsymbol{r}_i = \boldsymbol{\epsilon}_i$ is used for this quantity)

- now suppose we had estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$ and $\hat{\phi}$
- then we could compute estimates $\widehat{\boldsymbol{\mu}}_i$, $\widehat{V}_i$, and

$$\widehat{\boldsymbol{\epsilon}}_i = \boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i$$

- We could then plug all of this into the sandwich formula to obtain

$$\widehat{\text{var}}(\widehat{\boldsymbol{\beta}}) = \left(\sum_i \widehat{\widetilde{D}}_i' \widehat{V}_i^{-1} \widehat{\widetilde{D}}_i\right)^{-1} \left(\sum_i \widehat{\widetilde{D}}_i' \widehat{V}_i^{-1} (\widehat{\boldsymbol{\epsilon}}_i \widehat{\boldsymbol{\epsilon}}_i') \widehat{V}_i^{-1} \widehat{\widetilde{D}}_i\right) \left(\sum_i \widehat{\widetilde{D}}_i' \widehat{V}_i^{-1} \widehat{\widetilde{D}}_i\right)^{-1}$$

  which is the **robust variance estimator** for the QL estimator $\widehat{\boldsymbol{\beta}}$

- All of this is an extension of our development for linear models

Generalized Estimating Equations (GEE) for Marginal Models

GEE combines quasi-likelihood estimation with robust variance estimation to estimate generalized linear marginal models for longituindal data:

1. Specify a marginal model parameterized by $\boldsymbol{\beta}$ for $\mathrm{E}(\boldsymbol{Y}_i) = \boldsymbol{\mu}_i$
   – linear predictor $\eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta}$
   – link function $h(\mu_{ij}) = \eta_{ij} \Rightarrow h(\mu_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta}$
2. Specify a model for $\mathrm{var}(Y_{ij}) = \phi v(\mu_{ij})$ via variance function $v(\mu_{ij})$ ("working" variance function)
3. Specify a "working" correlation model C_i parameterized by $\boldsymbol{\alpha}$
4. 2. and 3. give a "working" variance model

$$\mathrm{var}(\boldsymbol{Y}_i) = V_i = \phi A_i^{1/2} C_i A_i^{1/2}$$

5. For given $\widehat{\boldsymbol{\alpha}}$, estimate $\boldsymbol{\beta}$ via QL with working $\widehat{V}_i$ (see above)
6. For given $\widehat{\boldsymbol{\beta}}$, estimate $\boldsymbol{\alpha}$ (see below)
7. Iterate 5. and 6. until convergence of $\widehat{\beta}$
8. The resulting $\widehat{\boldsymbol{\beta}}$ is the GEE estimator
9. Do Huber-White robust variance estimation for $\mathrm{var}(\widehat{\boldsymbol{\beta}})$

Estimation of Correlation Parameter α (and ϕ) with given $\widehat{\beta}$

- For a given estimate $\widehat{\boldsymbol{\beta}}$, compute the **standardized residuals**

$$\hat{r}_{ij} = \frac{y_{ij} - \widehat{\mu}_{ij}}{\sqrt{v(\widehat{\mu}_{ij})}} = \frac{\widehat{\epsilon}_{ij}}{\sqrt{v(\widehat{\mu}_{ij})}} \qquad \text{(new!)}$$

- **Rest** of this has been covered when we did GEE for linear models

- Estimate ϕ (or set equal to 1 if Y_{ij} is binary):
  (Analagous to estimation of σ^2 in linear models)

$$\hat{\phi} = \frac{1}{N-p} \sum_{ij} \hat{r}_{ij}^2$$

- Now estimate $\boldsymbol{\alpha}$ using the $\hat{r}_{ij}$'s and $\hat{\phi}$
- Exchangeable (compound symmetry) correlation:

$$\hat{\alpha} = \frac{1}{\hat{\phi}} \frac{1}{\sum_i n_i(n_i-1)/2 - p} \sum_i \sum_{j<k} \hat{r}_{ij}\hat{r}_{ik}$$

  (average product of all pairs of residuals on same subject)

- Exponential (AR-1) correlation (for simplicity, assume all observations are equally-spaced in time):

$$\hat{\alpha} = \frac{1}{\hat{\phi}} \frac{1}{\sum_i (n_i - 1) - p} \sum_i \sum_{j=2}^{n_i} \hat{r}_{ij} \hat{r}_{i,j-1}$$

(average product of all pairs of residuals on same subject and separated by 1 unit of time)

- Unstructured correlation (assuming all observations equally-spaced in time):

$$\hat{\alpha}_{jk} = \frac{1}{\hat{\phi}} \frac{1}{\sum_i n_i - p} \sum_i \hat{r}_{ij} \hat{r}_{i,k}$$

Properties of GEE

- GEE treats the variance-covariance parameters $(\phi, \boldsymbol{\alpha})$ as **nuisances**, focusing on the marginal mean model parameter $\boldsymbol{\beta}$
- $\widehat{\boldsymbol{\beta}}$ estimated by GEE is **consistent** for $\boldsymbol{\beta}$ under mild regularity conditions
- $\widehat{\boldsymbol{\beta}}$ estimated by GEE is **nearly efficient** relative to maximum likelihood, provided that the working model for $\mathrm{var}(\boldsymbol{Y}_i)$ is a reasonably-good approximation of the true variance model

(Review) Inferences in the Mean Model

- With $\widehat{\boldsymbol{\beta}}$ estimated by GEE and $\widehat{\text{var}}(\widehat{\boldsymbol{\beta}})$
  $\Rightarrow$ Inferences in the mean model are the same with linear model (Note 5)
- Use χ^2-test for vector-valued parameter
- Eg, with $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)'$ we would test

$$H_0 : \beta_1 = \beta_2 = \beta_3 \quad \text{vs.} \quad H_A : \beta_1 \neq \beta_2 \text{ or } \beta_1 \neq \beta_3$$

  Equivalently

$$H_0 : \beta_1 - \beta_2 = \beta_1 - \beta_3 = 0 \quad \text{vs.} \quad H_A : \beta_1 - \beta_2 \neq 0 \text{ or } \beta_1 - \beta_3 \neq 0$$

- In general, we would test

$$H_0 : L\boldsymbol{\beta} = 0 \quad \text{vs.} \quad H_A : L\boldsymbol{\beta} \neq 0$$

- For this example,

$$L = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \end{pmatrix}$$

- Then we compute $L\widehat{\boldsymbol{\beta}}$ and

$$\widehat{\text{var}}(L\widehat{\boldsymbol{\beta}}) = L\widehat{\text{var}}(\widehat{\boldsymbol{\beta}})L' = L\widehat{C}L'$$

and finally

$$\chi^2 = (L\widehat{\boldsymbol{\beta}})'(L\widehat{C}L')^{-1}(L\widehat{\boldsymbol{\beta}})$$

where $\widehat{C} = \widehat{\text{var}}(\widehat{\boldsymbol{\beta}})$.

- Statistic χ^2 has an approximate χ^2-distribution
  – DF $=$ rank of L ($= 2$ for this example)

- Note: Since there is no parametric assumption for GEE model, we cannot use a likelihood ratio test to compare nested models.

Example
Logistic Regression Model for Breastfeeding in Nepalese Children

- Suppose we wish to examine whether and how breastfeeding behavior varies with age and/or sex of child
- We use a binary outcome variable of "any" versus "no" breastfeeding
- We can estimate this model using GEE: Need to specify a **linear predictor**, a **link function**, a **variance function**, and a **working correlation model**

- First get data, generate an observation number representing "equally-spaced" times
- R code:

```
# read data
data=read.csv("nepal.csv")
data$obs=rep(1:5,200)
  #if data is not sorted, need to first sort it by age within each id
```

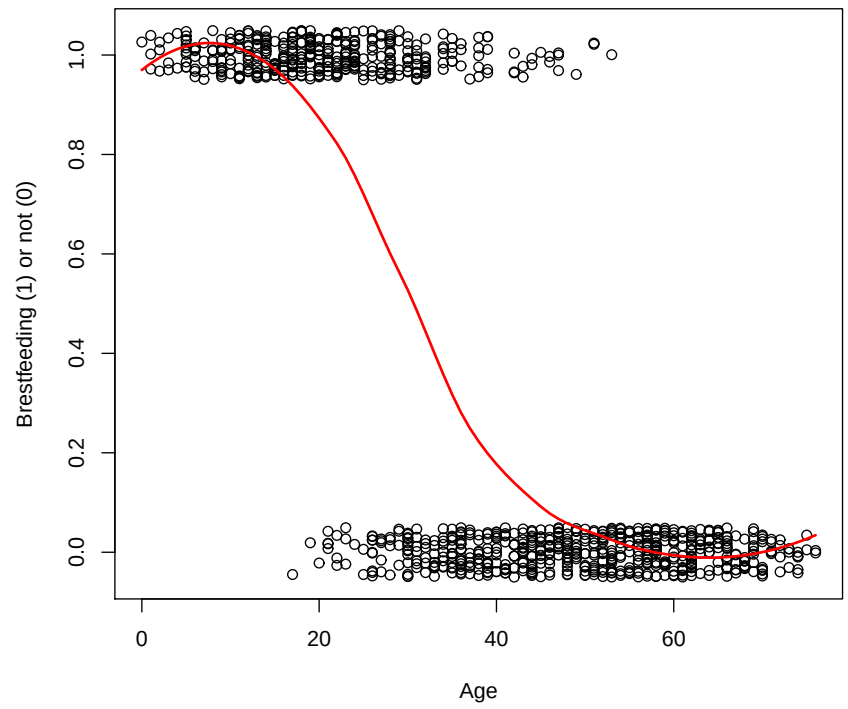
- Generate binary breastfeeding variable, where original variable is 3-level:
  - 0: none
  - 1: $<$10 times/day
  - 2: $\geq$10 times/day

```
> table(data$bf)

  0   1   2
564 151 232
```

```
> data$bfbin=(data$bf>0)
> table(data$bf,data$bfbin)

    FALSE TRUE
  0   564    0
  1     0  151
  2     0  232
```

- Exploratory data analysis for mean model:
  We use smooth curve to explore mean model with respect to age

```
model.lowess=loess(bfbin~age,data=data,span = 0.7)
pred.lowess=predict(model.lowess)

# Add small error to easily see observations on plots
data$bfbinjg = data$bfbin + 0.1*(runif(dim(data)[1])-.5)

# plot observed outcome and smooth curve
plot(data$age,data$bfbinjg,xlab='Age',ylab='Brestfeeding (1) or not (0)')
ord=order(data$age[!is.na(data$bf)]) # sort by age
lines(data$age[!is.na(data$bf)][ord],pred.lowess[ord],lwd=2,col=2)
```



Note: It appears as if the logistic function will be appropriate for modelling the effects of child's age on prevalence of breastfeeding

- Thus, define

$$Y_{ij} = I(\text{any breastfeeding for child } i, \text{ time } j)$$

and

$$\mu_{ij} = \Pr(Y_{ij} = 1 | \text{age}_{ij}, \text{sex}_i)$$

Our preliminary model is:

Mean Model:

$$\text{logit}(\mu_{ij}) = \frac{\mu_{ij}}{1-\mu_{ij}} = \beta_0 + \beta_1 \texttt{age}_{ij} + \beta_2 \texttt{sex}_i$$

Variance Model:

$$\text{var}(Y_{ij}|\boldsymbol{x}_{ij}) = \mu_{ij}(1-\mu_{ij}), \quad \phi = 1$$

- Now, in order to do GEE, we require some knowledge of the correlation structure
- Exploratory data analysis for correlation structure:
  – Fit a (preliminary and **flexible**) model with age, sex and their **interaction**
  – Obtain standardized residuals from that model
  – Examine autocorrelation function of these standardized residuals
- R code:

```
> # Explore correlation structure in breastfeeding,
> # removing effects of age and sex
>
> logit.model=glm(bfbin~age*sex,data=data,family="binomial")
> summary(logit.model)

<snip>
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.1577627  1.3266837   4.641 3.46e-06 ***
```

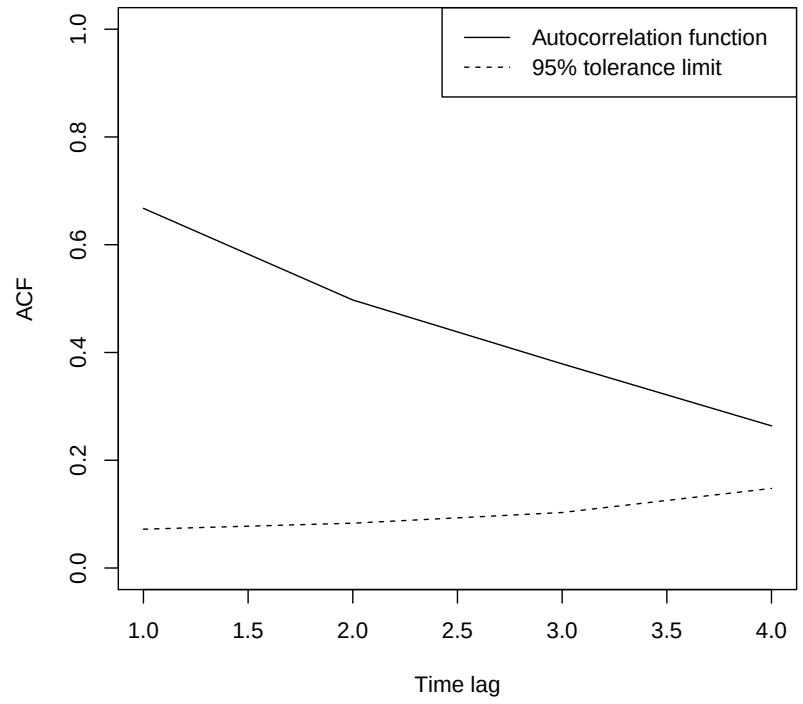
```
age         -0.1773359  0.0399157  -4.443 8.88e-06 ***
sex          0.0003587  0.8512807   0.000    1.000
age:sex     -0.0127863  0.0263824  -0.485    0.628
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

> # Calculate standardized residuals
> data$muhat=NA
> data$muhat[!is.na(data$bfbin)]=predict(logit.model,type="response")
> data$r = (data$bfbin - data$muhat)/sqrt(data$muhat*(1-data$muhat))
> summary(data$r)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
-4.31899 -0.22649 -0.04097  0.02540  0.25463  8.12485       53
```

- Here:
  - `muhat` is prelimary fitted mean: $\hat{\mu}_{ij}$
  - `r` is preliminary standardized residual: $r_{ij} = (y_{ij} - \hat{\mu}_{ij})/\sqrt{\hat{\mu}_{ij}(1-\hat{\mu}_{ij})}$

```
> # into wide format
> wide<-reshape(data[,c(1,15,18)],v.names="r",idvar="id",timevar="obs",
    direction="wide")
>
> # correlation matrix
> cor(wide[,-1],use="pairwise.complete.obs")
          r.1       r.2       r.3       r.4       r.5
r.1 1.0000000 0.4666275 0.4032542 0.3426585 0.2638496
r.2 0.4666275 1.0000000 0.8482971 0.6287798 0.4491978
r.3 0.4032542 0.8482971 1.0000000 0.7274633 0.5230721
r.4 0.3426585 0.6287798 0.7274633 1.0000000 0.6969190
r.5 0.2638496 0.4491978 0.5230721 0.6969190 1.0000000
```

```
## Autocorrelation function and correlogram

# extract all pairs within each subject
data$id=as.factor(data$id)
id.list=levels(data$id)
data.pairs=list()
for(i in 1:length(id.list))
{
subject.i=data[data$id==id.list[i],]
data.pairs.i=gtools::combinations(dim(subject.i)[1], 2,repeats=FALSE)
```

```
#index of all possible pairs
data.pairs.add=data.frame(id=rep(i,dim(data.pairs.i)[1]),
obs1=subject.i$obs[data.pairs.i[,1]],obs2=subject.i$obs[data.pairs.i[,2]],
r1=subject.i$r[data.pairs.i[,1]],r2=subject.i$r[data.pairs.i[,2]]
) # all pairs within subject i
data.pairs=rbind(data.pairs,data.pairs.add)
}
data.pairs$lag=abs(data.pairs$obs2-data.pairs$obs1)
data.pairs[1:10,]

# remove missing observations
# otherwise N in TL calculation will include missing observations
data.pairs=data.pairs[!(is.na(data.pairs$r1) | is.na(data.pairs$r2)),]

# Autocorrelation function and tolerance limit
lag.list=1:4
ACF=TL=rep(NA,length(lag.list))
for(lag in 1:length(lag.list))
{
subset=data.pairs[which(data.pairs$lag==lag.list[lag]),4:5]
ACF[lag]=cor(subset,use="pairwise.complete.obs")[1,2]
TL[lag]=1.96/sqrt(dim(subset)[1])
}
```



- The correlation matrix and autocorrelation function
  - Suggest correlation decays with time lag
  - Suggests "exponential" / "AR(1)" **working** correlation model

- We fit a GEE logistic model
  – With AR(1) working correlation
  – Mean model including age and sex

- SAS code:

```
data nepal;
set nepal;
if(bf eq 0)then bfbin=0;
if(bf eq 1)or(bf eq 2)then bfbin=1;
run;

proc sort data=nepal;
by id age;
run;

*GEE;
proc genmod data=nepal DESCENDING;
class id;
model bfbin=age sex / dist=bin noscale;
repeated subject=id / type=AR(1) covb corrw modelse ;
run;
```

- Note for SAS options:
  - `DESCENDING` will model Pr(Y=1), instead of Pr(Y=0) (default)
  - `modelse` requests model-based standard errors. By default, only empirical standard errors is displayed.
  - `corrw` Displays the estimated working correlation matrix
  - `covb` Displays the estimated covariance matrix

- Some results:

GEE Model Information

Correlation Structure	AR(1)
Subject Effect	id (200 levels)
Number of Clusters	200
Clusters With Missing Values	30

Working Correlation Matrix

	Col1	Col2	Col3	Col4	Col5
Row1	1.0000	0.6025	0.3630	0.2187	0.1318
Row2	0.6025	1.0000	0.6025	0.3630	0.2187
Row3	0.3630	0.6025	1.0000	0.6025	0.3630
Row4	0.2187	0.3630	0.6025	1.0000	0.6025
Row5	0.1318	0.2187	0.3630	0.6025	1.0000

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	5.8088	0.7343	4.3697	7.2480	7.91	<.0001
age	-0.1793	0.0165	-0.2116	-0.1470	-10.88	<.0001
sex	-0.1907	0.3442	-0.8653	0.4839	-0.55	0.5795

Analysis Of GEE Parameter Estimates
Model-Based Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	5.8088	0.7684	4.3029	7.3148	7.56	<.0001
age	-0.1793	0.0169	-0.2124	-0.1462	-10.62	<.0001
sex	-0.1907	0.3314	-0.8402	0.4588	-0.58	0.5650
Scale	1.0000	.	.	.	.	.

NOTE: The scale parameter was held fixed.

- Working correlation parameter $\alpha = 0.6025$:
  Within same child, (working) correlation between two outcomes separated by 1-unit time interval (ie, 4 month) is 0.6025.
- Note: you can fit this model in Stata by
  `xtgee bfbin age sex, family(binomial) nmp robust corr(ar 1) force eform`

- We remove sex from model and refit:

```
proc genmod data=nepal DESCENDING;
class id;
model bfbin=age / dist=bin noscale;
repeated subject=id / type=AR(1) covb corrw modelse ;
output out=nepal.predict pred=predict;
run;
```

Working Correlation Matrix

	Col1	Col2	Col3	Col4	Col5
Row1	1.0000	0.6089	0.3708	0.2257	0.1375
Row3	0.3708	0.6089	1.0000	0.6089	0.3708
Row4	0.2257	0.3708	0.6089	1.0000	0.6089
Row5	0.1375	0.2257	0.3708	0.6089	1.0000

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	5.4959	0.4824	4.5504	6.4414	11.39	<.0001
age	-0.1782	0.0163	-0.2101	-0.1463	-10.94	<.0001

Analysis Of GEE Parameter Estimates
Model-Based Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	5.4959	0.5483	4.4214	6.5705	10.02	<.0001
age	-0.1782	0.0168	-0.2111	-0.1453	-10.62	<.0001
Scale	1.0000	.	.	.	.	.

NOTE: The scale parameter was held fixed.

Interpretation and conclusion:

- Sex of child is not a significant predictor of breastfeeding behavior.
- Age is a significant predictor of breastfeeding behavior.
- (Cross-sectional/marginal interpretation) The estimated odds ratio for breastfeeding for two children who differ by one month in age is $\exp(-0.1782) = 0.84$
  95% CI: [$\exp(-0.2111)$, $\exp(-0.1453)$]=[0.81, 0.86].

- Finally, compare our results to those obtained under an independence working correlation model (based on robust SE):
- SAS code:

```
* Comparison using ind model;
proc genmod data=nepal DESCENDING;
class id;
model bfbin=age / dist=bin noscale;
repeated subject=id / type=ind covb corrw modelse ;
run;
```

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	6.0796	0.5749	4.9527	7.2064	10.57	<.0001
age	-0.1932	0.0197	-0.2319	-0.1546	-9.80	<.0001

- Results are qualitatively similar
- However, standard errors are larger with the independence model, reflecting a loss in efficiency of using a clearly wrong correlation model
- Note: in Stata, you can fit this model with independence correlation by

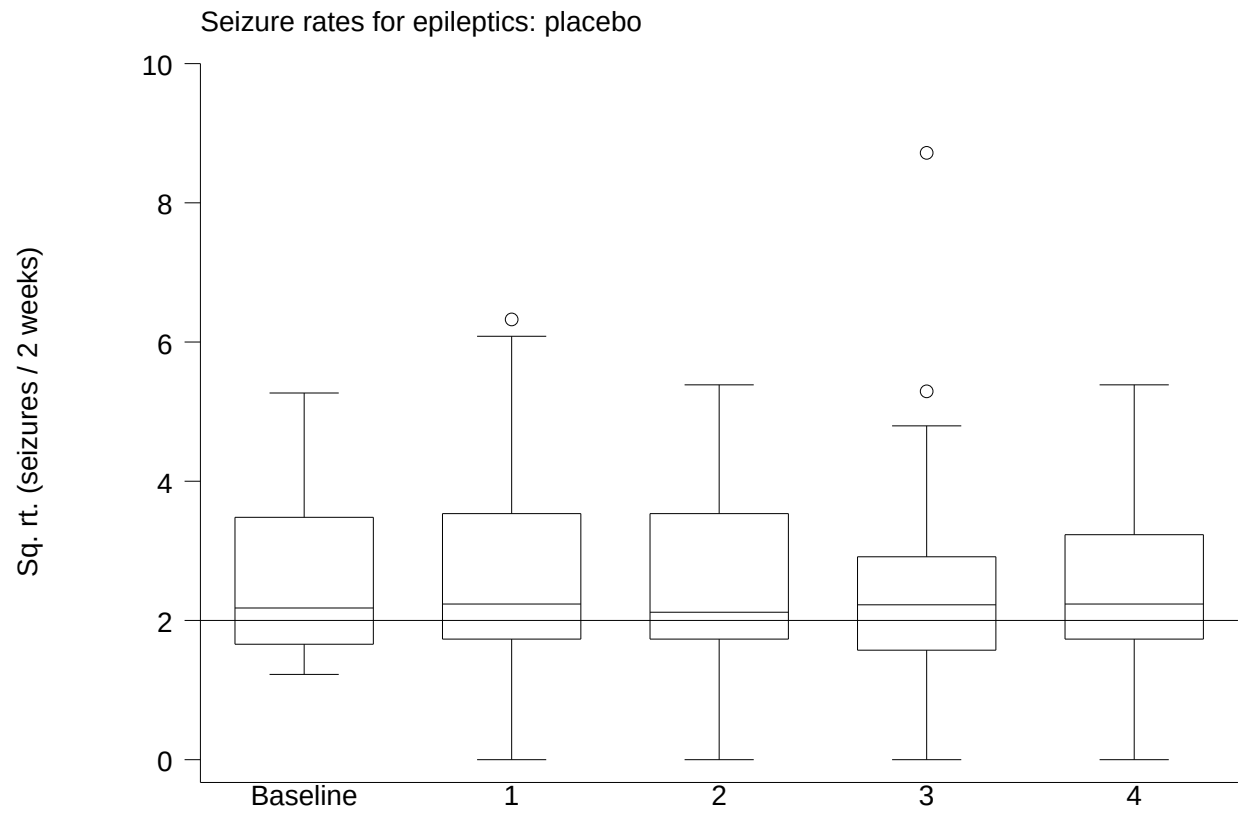
```
xtgee bfbin age , family(binomial) nmp robust
corr(ind) force eform
```

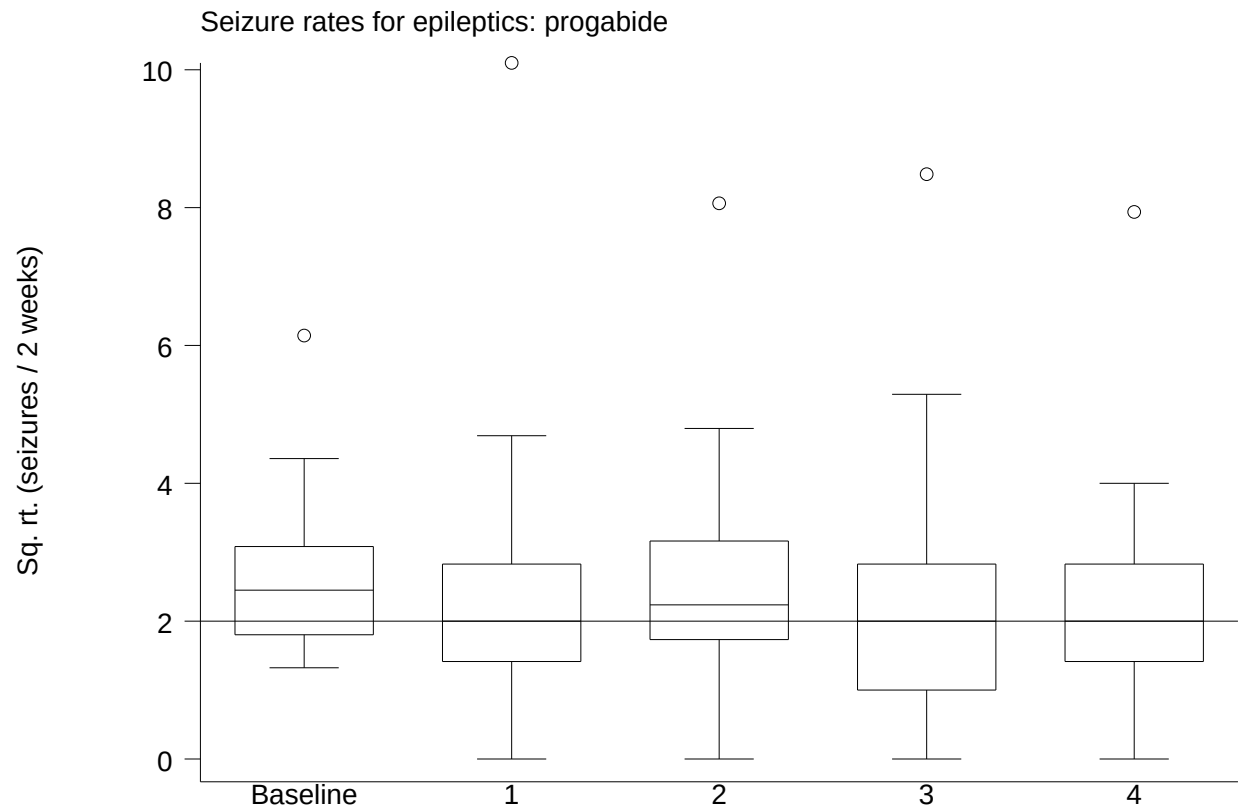
Poisson Regression Model for Number of Seizures in Progabide Study

- We wish to examine differential effects of treatment with progabide (versus placebo) on number of seizures
- Get data

```
> data=read.csv("seizure.csv")
> data[1:8,]
    id time age tx seiz length
1  101    0  18  1   76      8
2  101    1  18  1   11      2
3  101    2  18  1   14      2
4  101    3  18  1    9      2
5  101    4  18  1    8      2
6  102    0  32  1   38      8
7  102    1  32  1    8      2
8  102    2  32  1    7      2
```

- Length of baseline is 8 weeks, but 2 weeks for following observations


Seizure rates for epileptics: placebo
Sq. rt. (seizures / 2 weeks)
10
8
6
4
2
0
Baseline
1
2
3
4



- One subject might be outlier

- Now, compute mean for each combination of time and treatment, and then obtain standardized residuals

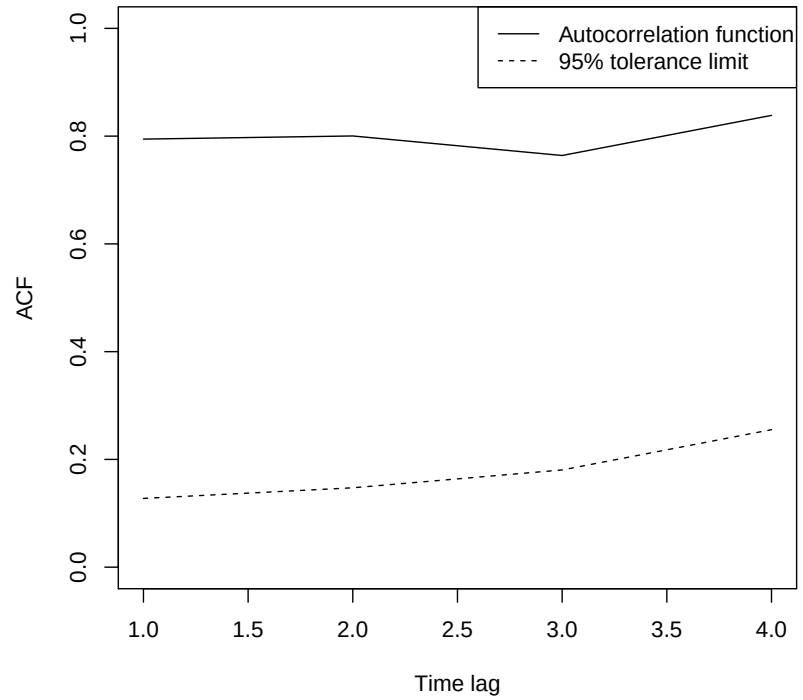
```
# Explore correlation structure for saturated mean model
# removing effects of time(as categorical), treatment, and ther interaction
model=lm(seiz~factor(time)*factor(tx),data=data)

# Calculate standardized residuals
data$seizmn=NA
data$seizmn=predict(model)
data$r = (data$seiz - data$seizmn)/sqrt(data$seizmn)
```

- Use these residuals to obtain correlation matrix and autocorrelation function

```
> wide<-reshape(data[,c(1,2,8)],v.names="r",idvar="id",timevar="time",
    direction="wide")
```

```
> cor(wide[,-1],use="pairwise.complete.obs")
          r.0       r.1       r.2       r.3       r.4
r.0 1.0000000 0.7916299 0.8314788 0.6761007 0.8385147
r.1 0.7916299 1.0000000 0.8701689 0.7450950 0.9013406
r.2 0.8314788 0.8701689 1.0000000 0.8084712 0.8990623
r.3 0.6761007 0.7450950 0.8084712 1.0000000 0.8288169
r.4 0.8385147 0.9013406 0.8990623 0.8288169 1.0000000

## Autocorrelation function and correlogram
# extract all pairs within each subject
data$id=as.factor(data$id)
id.list=levels(data$id)
data.pairs=list()
for(i in 1:length(id.list))
{
subject.i=data[data$id==id.list[i],]
data.pairs.i=gtools::combinations(dim(subject.i)[1], 2,repeats=FALSE)
#index of all possible pairs
data.pairs.add=data.frame(id=rep(i,dim(data.pairs.i)[1]),
obs1=subject.i$time[data.pairs.i[,1]],obs2=subject.i$time[data.pairs.i[,2]],
r1=subject.i$r[data.pairs.i[,1]],r2=subject.i$r[data.pairs.i[,2]]
) # all pairs within subject i
```

```
data.pairs=rbind(data.pairs,data.pairs.add)
}
data.pairs$lag=abs(data.pairs$obs2-data.pairs$obs1)
data.pairs[1:10,]

# remove missing observations
# otherwise N in TL calculation will include missing observations
data.pairs=data.pairs[!(is.na(data.pairs$r1) | is.na(data.pairs$r2)),]

# Autocorrelation function and tolerance limit
lag.list=1:4
ACF=TL=rep(NA,length(lag.list))
for(lag in 1:length(lag.list))
{
subset=data.pairs[which(data.pairs$lag==lag.list[lag]),4:5]
ACF[lag]=cor(subset,use="pairwise.complete.obs")[1,2]
TL[lag]=1.96/sqrt(dim(subset)[1])
}

# correlogram
plot(lag.list,ACF,"l",xlab="Time lag",xlim=c(1,4),ylim=c(0,1))
lines(lag.list,TL,lty=2)
legend("topright",c("Autocorrelation function","95% tolerance limit"),lty=1:2)
```



- An exchangeable correlation structure is a good "working" correlation model here

- Now fit GEE model with exchangeable correlation
  - Recall log-linear mean model:

$$\log(E(Y_{ij}|x_{ij})) = x'_{ij}\boldsymbol{\beta}$$

    ie,

$$E(Y_{ij}|x_{ij}) = \exp(x'_{ij}\boldsymbol{\beta})$$

  - **Issue**: length of baseline is 8 weeks, but 2 weeks for following observations

- To solve this:
  * Do not divide outcome by length
  * Instead consider the mean model

$$E(Y_{ij}|x_{ij}) = \texttt{length} \times \exp(x'_{ij}\boldsymbol{\beta})$$

  ie,

$$\log(E(Y_{ij}|x_{ij})) = \log(\texttt{length}) + x'_{ij}\boldsymbol{\beta}$$

    - Now interpretation of $\boldsymbol{\beta}$ is related to count of seizures **per week**
- Thus, we need add an offset log(length) for varying lengths of observation time (2 or 8 weeks).
- Since post-treatment are similar, recode time to be either "pre" or "post" treatment

– Our model is:

Mean Model:

$$\log(\mu_{ij}) = \log(\texttt{length}) + \beta_0 + \beta_1 \texttt{tx}_i + \beta_2 \texttt{post}_{ij} + \beta_3 \texttt{tx}_i \times \texttt{post}_{ij}$$

Variance Model:

$$\text{var}(Y_{ij}|\boldsymbol{x}_{ij}) = \phi\mu_{ij}$$

- SAS code:

```
data seizure;
set seizure;
post = 1*(time >= 1);
txtime = tx*post;
loglength=log(length);
run;

proc genmod data=seizure;
class id;
model seiz=tx post txtime /dist=poisson offset=loglength scale=Pearson;
repeated subject=id / type=exch covb corrb corrw modelse;
run;
```

- Note:
  – `offset=loglength` request offset term in model for varying lengths of observation time (2 or 8 weeks)
  – `scale=Pearson` request dispersion parameter to be estimated instead of 1 (SAS will report $\sqrt{\phi}$ as scale)
  – `type=exch` request working exchangeable correlation

- Some results:

```
Exchangeable Working
      Correlation

Correlation    0.7715879669
```

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	1.3476	0.1574	1.0392	1.6560	8.56	<.0001
tx	0.0265	0.2219	-0.4083	0.4613	0.12	0.9049
post	0.1108	0.1161	-0.1168	0.3383	0.95	0.3399
txtime	-0.1037	0.2136	-0.5223	0.3150	-0.49	0.6274

Analysis Of GEE Parameter Estimates
Model-Based Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	1.3476	0.1512	1.0513	1.6439	8.91	<.0001
tx	0.0265	0.2073	-0.3797	0.4328	0.13	0.8982
post	0.1108	0.1547	-0.1924	0.4140	0.72	0.4739
txtime	-0.1037	0.2199	-0.5348	0.3274	-0.47	0.6374
Scale	4.4388	.	.	.	.	.

- Note for Stata: you can fit the model by
  `xtgee seiz tx post txtime , family(poisson) nmp robust corr(exch) exposure(length) eform scale(x2)`
  - `exposure()` for varying lengths of observation time (2 or 8 weeks)
  - `scale()` to request printing of $\hat{\phi}$
- **Interpretation: Variance parameters**:
  - $\hat{\phi} = 4.4388^2 = 19.7$ indicates that the variance is 20 times greater than the mean number of seizures. This is very highly overdispersed relative to the Poisson distribution
  - $\hat{\rho} = \hat{\alpha} = 0.77$ is the correlation of any pair of observations on the same subject

- **Interpretation: Mean parameters**:
  - $\exp(\hat{\beta}_{\texttt{tx}}) = \exp(0.0265) = 1.03$ means that: at baseline, the mean of number of seizures per week in progabide group is 3% higher than placebo group
  - $\exp(\hat{\beta}_{\texttt{post}}) = \exp(0.1108) = 1.12$ means that, in placebo group, the number of seizures post treatment is 12% higher than pre-treatment
  - $\exp(\hat{\beta}_{\texttt{post}} + \hat{\beta}_{\texttt{txtime}}) = \exp(0.1108 - 0.1037) = 1.01$ means that: in progabide group, the number of seizures post treatment is 1% higher than pre-treatment
  - $\exp(\hat{\beta}_{\texttt{txtime}}) = \exp(-0.1037) = 0.90$ means that the post versus pre-treatment ratio in the progabide group is only 90% of that in the placebo group
  - However, all three coefficients are not significant

- In this analysis, one subject (id 207) is very different from the others
- Now refit model without patient 207

```
proc genmod data=seizure(where=(id ne 207));
class id;
model seiz=tx post txtime /dist=poisson offset=loglength scale=Pearson;
repeated subject=id / type=exch covb corrb corrw modelse;
run;
```

```
                Exchangeable Working
                     Correlation
           Correlation    0.5941485833
```

```
                Analysis Of GEE Parameter Estimates
                 Empirical Standard Error Estimates

                            Standard   95% Confidence
Parameter       Estimate       Error        Limits              Z Pr > |Z|

Intercept        1.3476      0.1574     1.0392    1.6560     8.56   <.0001
tx              -0.1080      0.1937    -0.4876    0.2716    -0.56   0.5770
post             0.1108      0.1161    -0.1168    0.3383     0.95   0.3399
txtime          -0.3016      0.1712    -0.6371    0.0339    -1.76   0.0781
```

Analysis Of GEE Parameter Estimates
Model-Based Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	1.3476	0.1106	1.1309	1.5644	12.19	<.0001
tx	-0.1080	0.1579	-0.4176	0.2015	-0.68	0.4940
post	0.1108	0.1233	-0.1308	0.3524	0.90	0.3687
txtime	-0.3016	0.1936	-0.6811	0.0779	-1.56	0.1193
Scale	3.2469	.	.	.	.	.

- Dispersion parameter ($\hat{\phi}$) is much lower now
- Treatment effect as measured by $\beta_{\texttt{txtime}}$ is stronger now

Criterion for Selecting GEE Models

- No parametric distribution is assumed for the GEE model (only the means and the variance/correlation structures are specified)
  $\rightarrow$ We cannot calculate log likelihood for a given model.
- Hence no model selection information such as the -2log likelihood, AIC, BIC, etc, are provided for GEE.
- Can use QIC (Quasi-likelihood under the Independence model Criterion) proposed by Pan (2001) and further discussed by Hardin and Hilbe (2003)
- QIC adds a penalty 2trace$(\widehat{\Sigma}_I^{-1}\widehat{V}_R)$ to the quasi-likelihood (Q)
  – $\widehat{\Sigma}_I^{-1}$ is variance estimator under independence correlation structure
  – $\widehat{V}_R$ is robust variance estimator under specified working correlation structure

- Another criteria, QICu, approximates QIC when the GEE model is correctly specified:
  Adds a penalty $2p$ to the quasi-likelihood (Q)
  – p is number of parameters
  – Recall AIC=-2LogL+$2p$

- QIC/QICu can be used to find an acceptable working correlation structure for a given model.

- Models do not need to be nested in order to use QIC or QICu to compare them.

- Another way to assess the specified correlation structure: compare both robust and model-based variance-covariance matrices $\widehat{\text{var}}(\beta)$
  – If these matrices are substantially different, then the working correlation structure is not a good choice.
  – For SAS: can use COVB option in the REPEATED statement to do this

GEE May Fail with time-varying Covariates

- Reference: Margaret Sullivan Pepe & Garnet L Anderson (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data, *Comm in Stat - Simulation and Computation*, 23:4, 939-951

- GEE method gives a consistent estimate for $\boldsymbol{\beta}$ under mild regularity conditions :
  If the model for the mean $\boldsymbol{\mu}_i$ (ie, $h(\mathrm{E}(Y_{ij}|x_{ij})) = x'_{ij}\boldsymbol{\beta}$) is correct,

$$S(\boldsymbol{\beta}) = \sum_i \frac{\partial \boldsymbol{\mu}'_i}{\partial \boldsymbol{\beta}} V_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = 0$$

  is **unbiased** even if V_i is misspecified

- GEE works for many types of longitudinal data

- However, GEE may be biased when your time-varying covariates are RANDOM (no control on covariates)

- Simulation for continuous outcome (h is identity function):
  Marginal mean model is $\mathrm{E}(Y_{ij}|x_{ij}) = x_{ij}\beta$ (one covariate)
- Models to generate data:
  - (Transition model) $Y_{i0} = 0$, $Y_{ij} = \alpha Y_{i,j-1} + \beta x_{ij} + \epsilon_{ij}$
    - $\mathrm{E}(x_{ij}) = \mathrm{E}(\epsilon_{ij}) = 0$
    - $Y_{i,j-1}$, x_{ij}, ϵ_{ij} are independent
    - $\Rightarrow \mathrm{E}(Y_{i1}|x_{i1}) = x_{i1}\beta$
    - $\Rightarrow \mathrm{E}(Y_{i2}|x_{i2}) = \alpha\mathrm{E}(x_{i1}\beta) + x_{i2}\boldsymbol{\beta} = x_{i2}\boldsymbol{\beta}$
  - (Random intercept model) $Y_{ij} = U_i + \beta x_{ij} + \epsilon_{ij}$
    - U_i is random effect with mean 0
    - U_i, X_{ij}, ϵ_{ij} are independent
- Above models have same marginal mean model $\mathrm{E}(Y_{ij}|x_{ij}) = x_{ij}\boldsymbol{\beta}$

- True $\beta = 0.5$ and replications=1000.

Average GEE estimates of $\hat{\beta}$ using different working correlation

Model Generating the Data	Working correlation			
	Identity	$C^{(1)}$	$C^{(2)}$	$C^{(3)}$
Transition	0.501	0.416	0.399	0.428
Random-intercept	0.500	0.495	0.498	0.500

- SE of the average is less then 0.005 in all cases.
- Random intercept model is unbiased
- Working identity correlation is unbiased
- Others are clearly biased

- Insights for transition model: $Y_{ij} = \alpha Y_{i,j-1} + \beta x_{ij} + \epsilon_{ij}$.
  – we have seen $\mathrm{E}(Y_{ij}|x_{ij}) = \beta x_{ij}$

- We have

$$\begin{aligned} Y_{ij} &= \alpha Y_{i,j-1} + \beta x_{ij} + \epsilon_{ij} \\ &= \alpha(\alpha Y_{i,j-2} + \beta x_{i,j-1} + \epsilon_{i,j-1}) + \beta x_{ij} + \epsilon_{ij} \\ &= \dots \\ &= \alpha^j Y_{i0} + \beta \sum_{s=1}^{j} \alpha^{j-s} x_{is} + \sum_{s=1}^{j} \alpha^{j-s} \epsilon_{is} \end{aligned}$$

$\Rightarrow \mathrm{E}(Y_{ij}|x_{is}, s = 1, .., n_i) = \beta \sum_{s=1}^{j} \alpha^{j-s} x_{is}$

- Y_{ij} depends on previous X

- **Important message** in the paper: For longitudinal data, we can get a consistent and asymptotically normal estimate for $\boldsymbol{\beta}$ if either of following holds (sufficient conditions):
  1. We validate the following equality:

$$\mathrm{E}(Y_{ij}|X_{ij}) = \mathrm{E}(Y_{ij}|X_{is}, s = 1, \ldots, n_i)$$

     – Satisfied if covariates are not time-varying, or pre-determined by design (eg, time)
     – Will be frequently violated with stochastic covariates.
  2. Or, we use independent working correlation matrix
- **Conclusion**: If time-varying covariates are RANDOM (no control on covariates), and you cannot validate

$$\mathrm{E}(Y_{ij}|X_{ij}) = E(Y_{ij}|X_{is}, s = 1, \ldots, n_i),$$

  You **should use working independence with GEE** to prevent bias

Summary and Recommendations

- If the focus of investigation is on the **marginal mean** of Y_{ij}, then GEE is a good choice. The estimates of $\boldsymbol{\beta}$ and standard errors are correct

- Make some attempt to use a good approximation of the covariance structure.
  – Exploratory tools can help with this, even for binary data

- If time-varying covariates are RANDOM (no control on covariates), should use working independence with GEE to prevent bias, unless you are sure that

$$\mathrm{E}(Y_{ij}|X_{ij}) = \mathrm{E}(Y_{ij}|X_{is}, s = 1, \ldots, n_i)$$

- When exploring correlation structure, use a fairly flexible mean model (relatively larger model)

- Use the empirical (robust) variance-covariance estimator of $\mathrm{var}(\widehat{\boldsymbol{\beta}})$ **unless** you are very confident of the correlation structure C_i and the variance model

  $$\mathrm{var}(Y_{ij}) = \phi v(\mu_{ij})$$

  This includes the facts that ϕ and C_i do not vary with X_i

- Do not "shop around" for a correlation model that makes your results significant. Better to decide on one model and then report the results. You can add "sensitivity analyses" if needed, comparing correlation models

- If the correlation and variance models are reasonable, and the sample size is large enough, the empirical and model-based standard errors should not differ greatly

- You should have at least $m = 100$ subjects in order to use the empirical variance-covariance estimator. $m = 200$ is much better, **especially** if Y_{ij} is **binary**

- If the sample size is smaller, analysis will require:
  – more exploratory work for the correlation and variance model
  – more assumptions about the correlation and variance model
  – use of the model-based variance-covariance estimator of $\mathrm{var}(\widehat{\boldsymbol{\beta}})$

  This may not be a bad thing

- Model-based variance estimator may also be much better than the robust estimator in cases where cluster sizes are especially large