MODELLING CORRELATED NON-NORMAL DATA

- Interested in modelling correlated non-normal response data
- That arise from either
  - longitudinal studies, in which multiple measurements are taken on the same subject (or unit) at different points in time
  - clustering, where measurements are taken on subjects (sub-units) that share a common characteristic that leads to correlation
• Examples

  Clinical trials on repeated episodes of asthmatic attacks, or repeated episodes of epileptic seizures, or respiratory illness

  Toxicology/teratology studies on litters of mice

  GP practice studies or family member studies

• The correlation must be taken into account for these types of studies

• Otherwise…

  may lead to incorrect inferences being made concerning regression parameters

  may lead to inefficient parameter estimates being obtained
GENERALIZED LINEAR MODELS (GLMs)

• Standard approach to fitting regression models to univariate response data, assumed to have come from an exponential family distribution

• Examples

  Linear regression - continuous (normal) data
  Logistic regression - binary data or proportions
  Poisson regression - count data
• Exponential family distributions all take the form

\[ f(y_i | \theta_i, \phi) = \exp\left\{ \frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right\} \]

• We can show that

\[ E(Y_i) = \mu_i = b'(\theta_i) \]
\[ \text{var}(Y_i) = \phi b''(\theta_i) = \phi V(\mu_i) \]

• In regression setting, we also have

\[ g(\mu_i) = \beta^T x_i \]
• Examples, assuming the canonical form, are

  Binary logistic regression where

  \[ b(\theta_i) = \log(1 + e^{\theta_i}) \]

  \[ g(\mu_i) = \log\left(\frac{\mu_i}{1 - \mu_i}\right) \]

  \[ \text{var}(Y_i) = \mu_i(1 - \mu_i) \text{ and } \phi = 1 \]
Poisson regression where
\[ b(\theta_i) = e^{\theta_i} \]
\[ g(\mu_i) = \log \mu_i \]
\[ \text{var}(Y_i) = \mu_i \quad \text{and} \quad \phi = 1 \]

Linear regression where
\[ b(\theta_i) = \frac{1}{2} \theta_i^2 \]
\[ g(\mu_i) = \mu_i \]
\[ \text{var}(Y_i) = \phi = \sigma^2 \]

(Observe that mean and variance are separately parameterised)
• Estimation of $\beta$

$$U(\beta) = \sum_{i=1}^{n} \left( \frac{\partial \mu_i}{\partial \beta} \right)^T V(\mu_i)^{-1} \{y_i - \mu_i(\beta)\} = 0$$

• Under certain regularity conditions, the mle

$$\hat{\beta} \sim \text{asym} \ N(\beta_0, \Sigma(\beta_0)),$$

• Covariance matrix estimated by

$$\hat{\Sigma}(\hat{\beta}) = \phi \left( \sum_{i=1}^{n} \left( \frac{\partial \mu_i}{\partial \beta} \right)^T V_i^{-1} \left( \frac{\partial \mu_i}{\partial \beta} \right) \right)^{-1} \bigg|_{\beta=\hat{\beta}} = \phi \left( \sum_{i=1}^{n} D_i^T V_i^{-1} D_i \right)^{-1} \bigg|_{\beta=\hat{\beta}}$$
• Score equations require only the first two moments of Y
• Has led Wedderburn to recommend using these score equations to estimate $\beta$ for any choice of link and variance function
• Even when the integral does not correspond to any known parametric distribution
• He named this approach to estimation - quasi-likelihood
• Computationally simple to fit and only requires the first two moments of Y
• Parameter estimates are consistent
• Asymptotic normality of estimates follow under certain regularity conditions
• Covariance matrix for the mle of $\beta$ shown earlier is valid if variance structure for $Y$ is correct

• If this is not the case, replace with

$$\hat{\Sigma}(\hat{\beta})\hat{\Sigma}_U(\hat{\beta})\hat{\Sigma}(\hat{\beta})$$

• The information sandwich estimator, where

$$\hat{\Sigma}_U(\hat{\beta}) = \frac{1}{\phi^2} \sum_{i=1}^{n} D_i^T V_i^{-1} (y_i - \mu_i(\beta))(y_i - \mu_i(\beta))^T V_i^{-1} D_i$$

• In Poisson and logistic regression, scale parameter, $\phi$, is 1
OVER-DISPERSION

• Observed (or empirical variance) exceeds the nominal variance under the presumed model

• Under Poisson regression, $\text{var}(Y) = \text{E}(Y)$

• This relationship between the mean and variance may be too restrictive

• A more flexible relationship, such as, $\text{var}(Y) = \phi \text{E}(Y)$ ($\phi > 1$) may be required

• Fit using quasi-likelihood

• An alternative mean-variance relationship is

$$\text{var}(Y) = \text{E}(Y) + \tau \text{E}(Y)^2$$
• Reasons for over-dispersion are

1. Important covariates are excluded or ignored

2. Heterogeneity across subjects

3. The sub-units that contribute to a unit’s response are correlated

   This correlation may be due to heterogeneity/frailty or,
   when measurements are made over time for the same subject it
   may be due to serial dependence or,
   from dependence of the present observation on past
   observations
   or combination of all these

• Third point relevant to modelling clustered or longitudinal data (data
  that can be disaggregated)
LONGITUDINAL/CLUSTERED NON-NORMAL DATA

• Using Poisson or logistic regression models for these types of data would require us to assume that

  1. Independence across subjects (or units) - reasonable

  2. Independence across sub-units for a particular unit - not reasonable

• Approaches used before analysed univariate (aggregated) data

• Account for the correlation between sub-units implicitly by specifying a more “flexible” mean-variance relationship

• Correlation was treated as a nuisance

• Interested in the effects of baseline variables
• Problem of adopting this univariate GLM/Quasi-likelihood approach when data can be disaggregated

  1. Cannot investigate the over-dispersion in any meaningful way

  2. Not able to assess the effects of time-dependent covariates or sub-unit specific covariates on response

• For example,

  If interest lies in the evolution of the rates over time

• Therefore sensible, when possible, to disaggregate aggregated data first before proceeding to model the data

• We do not discuss how to disaggregate the data here
• Interested instead on answering the question:
  How might standard GLMs for independent responses be extended to analyse correlated responses?

• Three approaches will be described
  Marginal approach using GEE
  Random effects approach
  Transition (Markov) model approach

• Discussion of these three approaches will necessarily be brief

• Refer to text by Diggle, Heagerty, Liang and Zeger (2002) for a more comprehensive account
NOTATION

\( Y_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m \)

jth outcome for the ith subject/unit

- Observations are independent across subjects/units, but correlated within units
- \( \text{E}(Y_{ij} | x_{ij}) = \mu_{ij} \) – marginal expectation
- \( \text{E}(Y_{ij} | x_{ij}, b, d_{ij}) = \eta_{ij} \) – conditional expectation
- Let \( g(.) \) be the link function
MARGINAL GEE MODELS

• We model separately
  the marginal expectation as a function of the covariates
  the within-unit/subject correlation
• Interested in making inferences at the population-averaged level
• That is we are interested in the marginal expectation or average
  response over subjects that share the same covariate values
• Assumptions of the marginal approach are similar to GLM, but in
  addition we specify a within-unit correlation structure

1. \[ g(\mu_{ij}) = \beta^T x_{ij} \]
2. \[ \text{var}(Y_{ij}) = \phi V(\mu_{ij}) \]
3. \[ \text{corr}(Y_{ij}, Y_{ik}) = \rho(\mu_{ij}, \mu_{ik}; \alpha) \]
• This approach appears to be the most natural generalisation of GLMs to dependent data

• The three assumptions by themselves constitute a semi-parametric approach to analysing correlated data

• $\beta$ regression parameters are interpreted in the same way as those obtained from fitting GLMs to independent, cross-sectional data

• Liang and Zeger developed the GEE method for estimating the parameters from a marginal model

• Their approach is a multivariate generalisation of GLM and Quasi-likelihood

• The rationale behind GEE is that increased efficiency in estimating $\beta$ can be obtained if correlation structure is specified
• Difficult to specify the correlation structure correctly

• Led Liang and Zeger to propose using a working correlation structure, $R(\alpha)$, instead

• $R(\alpha)$ does not depend on $\beta$

• If $R(\alpha)$ is correctly specified then GEE leads not only to consistent estimates of $\beta$ but also consistent standard errors

• Unlikely to specify $R(\alpha)$ correctly in practice

• Thus we would need to calculate robust standard errors instead
• More mathematically, the GEE approach requires solving for $\beta$ the estimating equations

$$U(\beta, \alpha) = \sum_{i=1}^{n} \left( \frac{\partial \mu_i}{\partial \beta} \right)^T V_i^{-1} \{ y_i - \mu_i(\beta) \} = 0$$

• $V_i$ is the variance covariance matrix for the $i$th unit

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

• $A_i$ is the diagonal matrix of variance functions – elements of which depend on $\mu_i$
• Choices of $R_i(\alpha)$

Independence: $R_i(\alpha) = I$
  - No within-unit correlation
  - Equivalent to standard GLM, except still estimate a scale parameter

Exchangeable: $R_i(\alpha) = \alpha$
  - Observations covary equally within units
  - Estimate a single parameter

Auto-regressive: $R_i(\alpha) = \alpha^{|j-k|}$
  - AR(1)
  - Correlation decays over time

Unstructured: $R_i(\alpha) = \alpha_{jk}$
  - $\alpha$ is a $m \times m$ matrix
  - Estimate $m(m-1)/2$ unique pairwise correlations
  - Very flexible; often hard to estimate
• If \( m \) is small and data are balanced and complete, then an unstructured matrix may be suitable

• If observations are mistimed, then suggest using a structure that accounts for correlation as a function of time (M-stationary or AR)

• If observations are clustered (with no logical ordering), then try exchangeable

• If number of clusters (i.e. \( n \)) is small, independence may be best

• See Diggle, Heagerty, Liang and Zeger (2002)
• Notice that the generalized estimating equations are functions of $\alpha$ and $\beta$, not just $\beta$ alone

• Thus need to replace $\alpha$ with a consistent estimate that is in terms of $\beta$ and $\phi$

• Need to replace $\phi$ with a consistent estimate that is in terms of $\beta$

• Do this using moment estimators involving Pearson’s residuals or by developing a new set of estimating equations for $\alpha$

• Asymptotic normality of the estimate of $\beta$ follow under regularity conditions
\[ \hat{\beta} \text{ asymp } \sim N(\beta_0, \Sigma) \]

\[
\Sigma = \left\{ \sum_{i=1}^{n} \left( \frac{\partial \mu_i}{\partial \beta} \right)^T V_i^{-1} \left( \frac{\partial \mu_i}{\partial \beta} \right) \right\}^{-1} \left\{ \sum_{i=1}^{n} \left( \frac{\partial \mu_i}{\partial \beta} \right)^T V_i^{-1} \text{cov}(Y_i) V_i^{-1} \left( \frac{\partial \mu_i}{\partial \beta} \right) \right\} \left\{ \sum_{i=1}^{n} \left( \frac{\partial \mu_i}{\partial \beta} \right)^T V_i^{-1} \left( \frac{\partial \mu_i}{\partial \beta} \right) \right\}^{-1}
\]

Substitute

\[
\text{cov}(Y_i) = \left( y_i - \mu_i(\hat{\beta}) \right) \left( y_i - \mu_i(\hat{\beta}) \right)^T
\]

If \( \text{cov}(Y_i) = V_i \), then

\[
\Sigma = \left\{ \sum_{i=1}^{n} \left( \frac{\partial \mu_i}{\partial \beta} \right)^T V_i^{-1} \left( \frac{\partial \mu_i}{\partial \beta} \right) \right\}^{-1}
\]

- model based variance-covariance matrix
SUMMARY

• GEE approach avoids the need for specifying the multivariate distribution by only assuming a particular form for the marginal distribution

• The covariance (or correlation) structure of the responses need not be specified correctly

• The GEE estimates of $\beta$ are robust to misspecification of $R_i(\alpha)$. That is, they are consistent and asymptotically normal even with misspecification

• The variance-covariance matrix is not

• Replace with the Information sandwich estimator

• Population-averaged interpretation of parameters
<table>
<thead>
<tr>
<th>Variable</th>
<th>Independence</th>
<th>Exchangeable</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Naïve se</td>
</tr>
<tr>
<td>Intercept</td>
<td>-2.370</td>
<td>0.876</td>
</tr>
<tr>
<td>log(base/4)</td>
<td>1.225</td>
<td>0.071</td>
</tr>
<tr>
<td>log(age)</td>
<td>0.588</td>
<td>0.239</td>
</tr>
<tr>
<td>trt</td>
<td>-0.018</td>
<td>0.105</td>
</tr>
<tr>
<td>V4 (fourth visit)</td>
<td>-0.161</td>
<td>0.118</td>
</tr>
<tr>
<td>φ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>α</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

On average the rate of seizures in the progabide group is 0.98 times the rate of seizures in the placebo group.
RANDOM EFFECTS MODELS

• Essential assumption underlying random effects models for longitudinal or clustered data is that there is natural heterogeneity across units in a subset of their regression parameters
  • For example, their intercepts and slopes
  • This heterogeneity can be explained by an appropriate probability distribution

• More formally, the general framework is as follows:
1. The conditional distribution of $Y_{ij}$ given the random effects $b_i$ follows an exponential family distribution.

2. Given $b_i$’s the repeated or clustered measurements from the same units are independent.

3. The $b_i$’s are iid with some probability distribution $f_{RE}(.)$.

4. Covariates are related to the conditional mean by

$$g(\eta_{ij}) = \beta^T x_{ij} + b_i^T d_{ij}$$

5. Conditional variance is given by

$$\text{var}(Y_{ij} \mid b_i) = \phi V(\eta_{ij})$$
• The interpretation of the regression coefficients $\beta$ is at a subject-specific level

• Therefore represents the effects of explanatory variables on a particular individual’s response

• To estimate $\beta$ we need to integrate out the random effects to form the marginal likelihood

• Numerical integration or the E-M algorithm may be required

• From the marginal likelihood the mle’s can be found
# RANDOM EFFECTS MODELS

<table>
<thead>
<tr>
<th>Variable</th>
<th>Gaussian random effects</th>
<th></th>
<th>Gamma random effects</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Standard error</td>
<td>Estimate</td>
<td>Standard error</td>
</tr>
<tr>
<td>Intercept</td>
<td>-1.104</td>
<td>1.193</td>
<td>-0.979</td>
<td>1.220</td>
</tr>
<tr>
<td>log(base/4)</td>
<td>1.030</td>
<td>0.101</td>
<td>1.038</td>
<td>0.095</td>
</tr>
<tr>
<td>log(age)</td>
<td>0.337</td>
<td>0.341</td>
<td>0.326</td>
<td>0.352</td>
</tr>
<tr>
<td>trt</td>
<td>-0.320</td>
<td>0.150</td>
<td>-0.263</td>
<td>0.151</td>
</tr>
<tr>
<td>V4 (fourth visit)</td>
<td>-0.158</td>
<td>0.055</td>
<td>-0.161</td>
<td>0.055</td>
</tr>
<tr>
<td>( \sigma_{RE} )</td>
<td>0.51</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \tau )</td>
<td></td>
<td></td>
<td>0.27</td>
<td></td>
</tr>
</tbody>
</table>
• The interpretation of the treatment effect corresponds to a particular subject (actually any subject with same value for the random effect) who is on placebo having a 27% (RR=0.73) reduced risk of seizures if (s)he switches to progabide.

• Notice that treatment is a between-subjects factor

• Consistency of estimates follow if mean structure is specified correctly

• Variance-covariance matrix of estimates is model-based and depends on the covariance structure induced for the responses by the random effects
TRANSITION (MARKOV) MODELS

• Extension of GLMs for describing the conditional distribution of $Y_{ij}$ on past responses and covariates

• The correlation amongst responses arises because of this dependence between present and past responses

$$g(\mu_{ij}^c) = \beta^T x_{ij} + \sum_{k=1}^{s} f_k(y_{i1}, \ldots, y_{ij-1}; \alpha)$$

$$\text{var}(Y_{ij}^c) = \phi V(\mu_{ij}^c),$$

• Fit in the same way as GLMs

• These models can really only be applied for longitudinal data
• We fit the simple transition model to the respiratory status data

\[
\frac{P(Y_{ij} = 1)}{1 - P(Y_{ij} = 1)} = \beta_1 + \beta_2 \text{treatment}_i + \alpha Y_{ij-1} \quad j = 1, \ldots, 4
\]

\[
\text{var}(Y_{ij} | Y_{ij-1}) = P(Y_{ij} = 1)(1 - P(Y_{ij} = 1))
\]

### Transition Model

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate (OR)</th>
<th>Model based-standard error</th>
<th>Robust standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.369 (0.254)</td>
<td>0.194</td>
<td>0.221</td>
</tr>
<tr>
<td>Treatment (Active v Placebo)</td>
<td>0.814 (2.257)</td>
<td>0.231</td>
<td>0.240</td>
</tr>
<tr>
<td>(Y_{i-1})</td>
<td>2.363 (10.623)</td>
<td>0.231</td>
<td>0.279</td>
</tr>
</tbody>
</table>
INTERPRETATION

• The parameters in a transition model have marginal interpretations.

• For example

The odds of having a good respiratory status amongst those in the active group is 2.3 times the odds of having a good respiratory status amongst those in the placebo group, given that they had the same respiratory status at their previous visit.

• A simple check of the Markov assumption is to compare the model-based standard errors to the robust standard errors
COMPARISON OF THREE METHODS

• Developed from different philosophical standpoints
  - Marginal models developed to focus on differences in population averaged responses
  - Random effects models focus on changes in response at an individual level
  - Transition models developed to take account of lagged responses
• Thus different interpretations for parameters
• When should each approach be used?

  - If interest lies at a population-averaged level (e.g. public health implications for the population), then use marginal models. That is, interested in the effects of explanatory variables on the population averaged response

  - If interest lies in the scientific investigation of the individual-level process or in individual-level prediction then a random effects approach should be used. Note however that this approach can be used to get population-averaged estimates as well.

  - If there is a reason to believe that there is a “causal” relationship between lagged responses and the present response, then use a transition model approach.
• Note that there are situations where the marginal interpretation of the parameters may turn out to be the same for random effects and marginal models (see notes)

• Marginal models rarely correspond to a probability model for the process under investigation, unlike random effects and transition models. Marginal models (based on GEE) are not likelihood based

• Cannot use likelihood ratio test for GEE models

• Advantage of GEE is the robust variance-covariance matrix for the estimates
MISSING DATA

• We have considered up to now only balanced response data
• Almost inevitable that missing data will arise in some studies
  - Missed appointments
  - Varied appointment times
  - Withdrawal from study
  - Varying number of kittens born between litters
• Validity of the approaches described depend on the missing data assumptions
• Consistency results for GEE are valid when data is MCAR

  - MCAR: missingness is independent of both the observed responses and those responses that would have been available had they not been missing

• For the likelihood-based approaches (RE and TM), the less strict assumption of MAR is required

  - MAR: missingness can depend on a subject’s previously observed responses, but given this information and also possible covariate information, the missingness is conditionally independent of the unobserved (missing) response values

• Note there are situations (e.g. when missingness is due to death) where we may be interested in modelling the response conditional on subjects being alive and the survival time distribution, rather than modelling the responses alone, treating death as simply missing.