Marginal Logistic Regression Model and GEE

> Marginal models are suitable to estimate population average parameters

For example, in the Indonesian Study, a marginal model can be used to address questions such as:

- what is the prevalence of respiratory infection in children as function of age?
- is the prevalence of respiratory infection greater in the sub-population of children with vitamin A deficiency?
- how does the association of vitamin A deficiency and respiratory infection change with age?

the scientific objective is to characterize and contrast populations of children.

Marginal Models for Binary Responses: Logistic Regression

- Model for the mean

$$E(Y_{ij}) = P(Y_{ij} = 1) = \mu_{ij}$$
$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 x_{ij}$$

- Model for Association

Marginal odds ratio

$$\gamma_{ijk} = \frac{P(Y_{ij} = 1, Y_{ik} = 1) Pr(Y_{ij} = 0, Y_{ik} = 0)}{P(Y_{ij} = 1, Y_{ik} = 0) Pr(Y_{ij} = 0, Y_{ik} = 1)}$$

Marginal odds ratio

$\gamma_{ijk} > 0$ a greater value indicates positive associations

1. $\gamma_{ijk} = \alpha$

   the degree of association is the same for all pairs of observations from the same subject

2. $\gamma_{ijk} = \alpha_0 + \alpha_1 \mid t_{ij} - t_{ik} \mid^{-1}$

   the degree of association is inversely proportional to the time between observations

Parameter Interpretation in Logistic Regression: ICHS study

- **Marginal Logistic Regression:**

$$\begin{array}{lcl} \text{logit} P(Y_{ij} = 1) & = & \beta_0 + \beta_1 x_{ij} \\ \text{corr}(Y_{ij}, Y_{ik}) & = & \text{AR(1), Uniform, unstructured} \\ e^{\beta_1} & = & \frac{P(Y_{ij}=1|x_{ij}=1)/P(Y_{ij}=0|x_{ij}=1)}{P(Y_{ij}=1|x_{ij}=0)/P(Y_{ij}=0|x_{ij}=0)} \end{array}$$

e^{β_1}*: odds of infection among vitamin A deficient children divided by the odds of infection among children replete with vitamin A.*

population-average *probability of infection:*

$$P(Y_{ij} = 1 \mid x_{ij} = 0) = \exp(\beta_0)/(1 + \exp(\beta_0))$$

the **population-average** *odds of infection is multiplied by* $\exp(\beta_1)$ *for the sub-population that is vitamin A deficient.*

$$\frac{P(Y_{ij} = 1 \mid x_{ij} = 1)}{P(Y_{ij} = 0 \mid x_{ij} = 1)} = e^{\beta_1} \frac{P(Y_{ij} = 1 \mid x_{ij} = 0}{P(Y_{ij} = 0 \mid x_{ij} = 0)}$$

- **Logistic Regression with Random Effects:**

$$\begin{array}{lcl} \text{logit} P(Y_{ij} = 1) & = & \beta_0^* + U_i + \beta_1^* x_{ij} \\ U_i & \sim & N(0, v^2) \\ e^{\beta_1^*} & = & \frac{P(Y_{ij}=1|U_i,x_{ij}=1)/P(Y_{ij}=0|U_i,x_{ij}=1)}{P(Y_{ij}=1|U_i,x_{ij}=0)/P(Y_{ij}=0|U_i,x_{ij}=0)} \end{array}$$

$e^{\beta_1^*}$*: odds of infection* **for a child with random effect** U_i *when he/she is vitamin A deficient relative to when the same child is not*

each child has his/her baseline probability of infection:

$$P(Y_{ij} = 1 \mid U_i, x_{ij} = 0) = \exp(\beta_0^* + U_i)/(1 + \exp(\beta_0^* + U_i))$$

and that a child's odds of infection is multiplied by $\exp(\beta_1^*)$ *if he/she become vitamin A deficient.*

$$\frac{P(Y_{ij}=1 \mid U_i, x_{ij}=1)}{P(Y_{ij}=0 \mid U_i, x_{ij}=1)} = e^{\beta_1^*}\frac{P(Y_{ij}=1 \mid U_i, x_{ij}=0)}{P(Y_{ij}=0 \mid U_i, x_{ij}=0)}$$

v^2: *degree of heterogeneity in the propensity of disease not attributable to* x

- **Transition Logistic Regression Model:**

$$\begin{array}{lcl} \text{logit}P(Y_{ij}=1) & = & \beta_0^{**}+\beta_1^{**}x_{ij}+\alpha y_{ij-1} \\ e^{\beta_1^{**}} & = & \frac{P(Y_{ij}=1|y_{ij-1},x_{ij}=1)/P(Y_{ij}=0|y_{ij-1},x_{ij}=1)}{P(Y_{ij}=1|y_{ij-1},x_{ij}=0)/P(Y_{ij}=0|y_{ij-1},x_{ij}=0)} \end{array}$$

$e^{\beta_1^{**}}$: *odds of infection* **for a child with outcome at the previous visit** y_{ij-1} *when he/she is vitamin A deficient relative to when the same child is not*

each child has his/her baseline probability of infection

$$P(Y_{ij}=1 \mid y_{i,j-1}, x_{ij}=0) = \exp(\beta_0^{**}+\alpha y_{ij-1})/(1+\exp(\beta_0^{**}+\alpha y_{ij-1}))$$

and that a child's odds of infection is multiplied by $\exp(\beta_1^{**})$ *if he/she become vitamin A deficient.* In summary:

$$\begin{array}{lcl} P(Y_{ij}=1 \mid x_{ij}=0, y_{ij-1}=0) & = & \frac{\exp(\beta_0^{**})}{1+\exp(\beta_0^{**})} \\ P(Y_{ij}=1 \mid x_{ij}=0, y_{ij-1}=1) & = & \frac{\exp(\beta_0^{**}+\alpha)}{1+\exp(\beta_0^{**}+\alpha)} \\ P(Y_{ij}=1 \mid x_{ij}=1, y_{ij-1}=0) & = & \frac{\exp(\beta_0^{**}+\beta_1^{**})}{1+\exp(\beta_0^{**}+\beta_1^{**})} \\ P(Y_{ij}=1 \mid x_{ij}=1, y_{ij-1}=1) & = & \frac{\exp(\beta_0^{**}+\beta_1^{**}+\alpha)}{1+\exp(\beta_0^{**}+\beta_1^{**}+\alpha)} \end{array}$$

Maximum Likelihood Estimation of β in GLM Cross Sectional Data

- If Y_i is binary or a count, we specify the Likelihood function and estimate the parameters of interest using Maximum Likelihood Estimation.
- For example, if Y_i is binary:

$$\begin{array}{lcl} Y_i & \sim & \text{Bernoulli}(\mu_i) \\ \text{logit}P(Y_i=1) & = & \text{logit}\mu_i = \beta_0+\beta_1 x_i \\ \ell(\beta_0,\beta_1) & \propto & \prod_i \mu_i^{y_i}(1-\mu_i)^{1-y_i} \end{array}$$

  we estimate β_0 and β_1 by maximizing $\ell(\beta_0,\beta_1)$.

- For example, if Y_i is count:

$$\begin{array}{lcl} Y_i & \sim & \text{Poisson}(\mu_i) \\ \log\mu_i & = & \beta_0+\beta_1 x_i \\ \ell(\beta_0,\beta_1) & \propto & \prod_i e^{-\mu_i}\mu_i^{y_i} \end{array}$$

  we estimate β_0 and β_1 by maximizing $\ell(\beta_0,\beta_1)$.

Maximum Likelihood Estimation of β in GLM Cross Sectional Data

In general:

$$\begin{array}{lcl} L(\boldsymbol{\beta};\boldsymbol{y}) & = & \prod_{i=1}^m f(y_i,\boldsymbol{\beta}) \\ U(\boldsymbol{\beta}) & = & \frac{\partial \log L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^m \frac{\partial \mu_i'}{\partial \boldsymbol{\beta}} v_i^{-1}(y_i-\mu_i(\boldsymbol{\beta})) \end{array}$$

find $\hat{\boldsymbol{\beta}}$ such that $L(\hat{\boldsymbol{\beta}}) \geq L(\boldsymbol{\beta})$

or find $\hat{\boldsymbol{\beta}}$ such that $U(\hat{\boldsymbol{\beta}})=0$

where $\mu_i = EY_i,\ v_i = var(Y_i) = v(\mu_i)$.

- $U(\boldsymbol{\beta})=0$ is called *score equation*. Solve the score equation is equivalent to maximize a likelihood function.
- solutions of $U(\boldsymbol{\beta})=0$ are not available in closed form, and require and iterative procedure called *Iterative weighted least squares (IWLS) algorithm*.

Main ideas of IWLS:

1. $\mu_i = EY_i,\ v_i = var(Y_i) = v(\mu_i)$
2. Choose $\hat{\boldsymbol{\beta}}$ to make $\mu_i(\hat{\boldsymbol{\beta}})$ close to y_i on average
3. weight y_i by v_i^{-1}

GEE Estimation of β in GLM Longitudinal Data

In the case of a linear regression model with the assumption of normality, the extension from ordinary linear regression to longitudinal problem was facilitated by thinking about a multivariate normal distribution.

By specifying a model for the mean $E[\boldsymbol{Y}_i]$ and the model for the covariance matrix V_i, we can fully specify the multivariate normal distribution:

$$\boldsymbol{Y}_i \sim MVN(X_i\boldsymbol{\beta}_i, V_i)$$

and use MLE.

Unfortunately if the elements of $\boldsymbol{Y}_i$ are counts or binary response, we cannot naturally extend the Bernoulli or Poisson distributions to take into account of correlation. Multivariate extensions of these distributions are quite complex (*except for biostats students!*).

The main impediments with binary and count data are:

1. there are not multivariate generalizations of the necessary probability distributions
2. population-average and subject-specific approaches do not lead to the same model for the mean response

GEE

- Under a GEE approach, we forget about trying to specify a model for the whole multivariate distribution of a data vector. Instead the idea is to just model the mean response $E[\boldsymbol{Y}_i]$ and the covariance matrix V_i of a data vector as in the normal case.
- In absence of a convenient likelihood to work with, it is sensible to estimate $\boldsymbol{\beta}$ by solving the following multivariate equation:

$$\boxed{S(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{m} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = 0}$$

- where $V_i = V_i(\boldsymbol{\alpha}, \boldsymbol{\beta})$
- The method of *generalized estimating equations* provides consistent estimates for the mean parameter when a model for the correlation may not be reliably specified.
- $S(\boldsymbol{\beta}, \boldsymbol{\alpha}) = 0$ is a multivariate generalization of the score equation $U(\boldsymbol{\beta}) = 0$ used to maximize the likelihood function under a GLM

GLM for Longitudinal Data (GEE)

In summary, for GEE models, we specify:

- A GLM for the mean response

$$\begin{aligned} h(E[\boldsymbol{Y}]) &= X\boldsymbol{\beta},\ h() \text{is the link function} \\ V_i &= V_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) \end{aligned}$$

- $V_i(\boldsymbol{\alpha}, \boldsymbol{\beta})$ independence, completely unstructured
- the estimate of $\boldsymbol{\beta}$ and their standard error will be consistent (i.e. unbiased for large sample size),
- if the specification of V_i is correct then the GEE solution is the maximum likelihood estimate

GEE

- One important property of the GLM family is that the score function $S(\boldsymbol{\beta}, \boldsymbol{\alpha})$ depends only on the mean and variance of Y_i. Therefore the estimating equation:

$$\boxed{S(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{m} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = 0}$$

can be used to estimate the regression coefficients for any choices of link and variance functions, whether or no they correspond to a particular member of the exponential family.
$S(\boldsymbol{\beta}, \boldsymbol{\alpha}) = 0$ is the generalized estimating equation.

GEE properties

- $\hat{\beta}$ is nearly efficient relative to the maximum likelihood estimate of $\boldsymbol{\beta}$ - provided that var$(\boldsymbol{Y}_i)$ has been reasonable approximated
- GEE is the maximum likelihood score equation for multivariate Gaussian data, and for binary data when var$(\boldsymbol{Y}_i)$ is correctly specified
- $\hat{\beta}$ is consistent from $m \to \infty$, even when var$(\boldsymbol{Y}_i)$ is incorrectly specified

Using GEE

1. Specify a model for the mean

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta}$$

   $g()$ is a link function: linear, logit, log

2. Specify model for $Cov(Y_{ij}, Y_{ik})$

$$Cov(Y_{ij}, Y_{ik}) = (VarY_{ij}VarY_{ik})^{1/2}Corr(Y_{ij}, Y_{ik})$$

   (a) Choose variance from
   Gaussian, Poisson, binomial
   $v(\mu) = 1,\ \mu,\ \mu(1-\mu)$

   (b) Choose correlation from
   Exchangeable, AR-1, unstructured

Bottom Line

If the scientific focus are the regression coefficients β:

1. focus on modeling the mean structure
2. use a reasonable approximation of the covariance structure
3. check the inferences for β by comparing βs robust std with respect different covariance assumptions
4. if the β's std differ substantially, a more careful treatment of the covariance model might be necessary.

Maximum Likelihood estimation for binary data

- Y_i is the # of patients responding negatively to treatment i
- n_i is the total # of patients
- p_i is the probability of negative response to treatment
- we consider two treatments $i = 1, 2$
- the likelihood function for p_1 and p_2 is

$$l(p_1, p_2) \propto p_1^{y_1}(1-p_1)^{n_1-y_1} \times p_2^{y_2}(1-p_2)^{n_2-y_2}$$

  which we can write as function of θ_1 and θ_2 so defined
- $\theta_1 = \log \frac{p_1(1-p_2)}{p_2(1-p_1)}$ and $\theta_2 = \log \frac{p_2}{1-p_2}$
- it turns out that the mle of θ_1 and θ_2 are

$$\begin{aligned}\hat{\theta}_1 &= \log \frac{y_1(n_2-y_2)}{y_2(n_1-y_1)} \\ \hat{\theta}_2 &= \log \frac{y_2}{n_2-y_2} \\ \hat{V}(\hat{\theta}_1) &= \frac{1}{y_1} + \frac{1}{n_1-y_1} + \frac{1}{y_2} + \frac{1}{n_2-y_2}\end{aligned}$$

Data from the 2×2 crossover trial on cerebrovascular deficiency adapted from Jones and Kenward, where treatment A and B are active drug and placebo, respectively; the outcome indicates whether an electrocardiogram was judged abnormal (0) or normal (1).

Group	(1,1)	(0,1)	(1,0)	(0,0)	Total	1	2
AB	22	0	6	6	34	28	22
BA	18	4	2	9	33	20	22

Example: a 2×2 crossover trial

- Goal: to compare the effect of an active drug (A) and a placebo (B) on cerebrovascular deficiency
- 34 patients received A followed by B
- 33 patients received B followed by A
- $Y_{ij} = 1$ if normal electrocardiogram reading
- At period 1, $\hat{p}_1 = y_1/n_1 = 28/34 = 82\%$ of patients receiving drug A were normal
- At period 1, $\hat{p}_2 = y_2/n_2 = 20/33 = 61\%$ of patients receiving placebo B were normal
- Odds ratio of the chance of being normal for the active drug versus the placebo is

$$\begin{aligned}\hat{\theta}_1 &= \log\left(\frac{y_1(n_2-y_2)}{y_2(n_1-y_1)}\right) = (13 \times 28)/(20 \times 6) = 3 \\ \sqrt{V(\hat{\theta}_1)} &= \left(\frac{1}{y_1} + \frac{1}{n_1-y_1} + \frac{1}{y_2} + \frac{1}{n_2-y_2}\right)^{1/2} \\ &= (28^{-1} + 6^{-1} + 20^{-1} + 13^{-1})^{1/2} = 0.57\end{aligned}$$

  this estimate is larger than 1 and therefore indicates that the active drug produces higher proportion of normal reading

Limitations

This approach has several limitations

1. Ignore the *carry-over effect*, i.e. the effect of the treatment at period 1 might influence the response at period 2 (treatment $\times$ period interaction)
2. two responses for the same subject are likely to be correlated
3. in fact the odds ratio:
$$\gamma = \frac{P(Y_{ij}=1, Y_{ik}=1)Pr(Y_{ij}=0, Y_{ik}=0)}{P(Y_{ij}=1, Y_{ik}=0)Pr(Y_{ij}=0, Y_{ik}=1)} =$$
is estimated to be
4. $(22 \times 6)/(6 \times 0.5) = 44$ for group AB
5. $(18 \times 9)/(4 \times 2) = 20.34$ for group BA

GEE Approach

- we combine the data from both periods
- We can analyze 2×2 crossover trail as a longitudinal study with $n_i = n = 2$ and $m = 67$
- $Y_{ij} = 1$ is subject i has a normal test at period j
- $x_1 = 1$ if period 2 or 0 if period 1
- $x_2 = 1$ if active drug (A) or 0 if placebo (B)

fit a logistic regression model:

$$\begin{array}{lcl} \text{logit } Pr(Y_{ij}=1) & = & \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij1} x_{ij2} \\ OR(Y_{ij}, Y_{ik}) & = & \gamma \end{array}$$

- $\hat{\beta}_3 = 1.02(+/-0.89)$ little support for a treatment by period interaction
- $\hat{\gamma} = \exp(3.54) = 34.5$ strong within-subject association, i.e. $\hat{\gamma}$ indicates that subjects with normal responses at the first visit have odds of normal reading at the next visit that are almost 35 times higher than those who first response was abnormal
- if we drop the interaction term, $\hat{\beta}_2 = 0.57+/-0.23$, $\exp(\hat{\beta}_2) = 1.77$, i.e. the odds of a normal electrocardiogram are 77% higher $(0.77 = exp(0.57) - 1)$ when using the active drug as compared to the placebo (GEE approach, robust standard error)
- if we assume that the repeated measurements within subjects are independent $(\gamma = 0)$ then $\hat{\beta}_2 = 0.57 \pm 0.38$ erroneously assessing that there is not a treatment effect (model based standard error)

In Summary

1. Model 1 includes the treatment$\times$period interaction (little support from the data), and estimates marginal odds ratio by GEE
2. Model 2 drops the period$\times$treatment interaction and estimate the marginal odds ratio by GEE
3. Model 3 assumes that the marginal odds ratio is zero, here $\hat{\beta}_2 = 0.56$ std 0.38 (much larger than under model 1 and 2)
4. if we fit model 3, but by using robust standard errors, than we obtain similar results to the GEE approach

Respiratory Infection in Indonesian preschool children (Sommer et al 1984)

- 275 children in Indonesia were examined for up to six consecutive quarters for the presence of respiratory infections ($i = 1, \ldots, m = 275$, $j = 1, \ldots, 6$ visits) Goals of the analysis:
  1. Whether the prevalence of respiratory infection is higher among children who suffer xerophthalmia (an ocular manifestation of chronic vitamin A deficiency)
  2. Estimate the change of respiratory infection with age
- seasonality as potential confounder

Cross - Sectional Analysis

Model 1: first visit only

- Look at only the data for the first visit
- Fit a logistic regression model of respiratory infection on xerophthalmia and age, adjusting for other covariates

$$\text{logit}P(Y_{i1}=1)=\beta_0+\beta_1\text{xer}_{i1}+\beta_2\text{age}_{i1}+\beta_3\text{age}_{i1}^2+\beta_4\text{gender}_i\ldots$$

- we find a strong non-linear cross-sectional age effect on the prevalence for respiratory infection
- cross-sectional analysis suggests that the prevalence for respiratory infection increases from age 12 months and reaches its peak at age 20 months before starting to decline.

Model 2: all visits plus controlling for seasonality

- Look at the data for all the visits

$$\begin{aligned}\text{logit}P(Y_{ij}=1) &= \beta_0+\beta_1\text{xer}_{ij}+\beta_2\text{age}_{ij}+\beta_3\text{age}_{ij}^2+\beta_4\text{gender}_i+\\ &+\ \textbf{harmonic terms}\\ OR(Y_{ij},Y_{ik}) &= \gamma\end{aligned}$$

- Fit a logistic regression model of respiratory infection on xerophthalmia and age, adjusting for other covariates
- we still find a strong non-linear cross-sectional age effect on the prevalence for respiratory infection
- The age coefficient in model 2 can be interpreted as weighted averages of the cross sectional age coefficients for each visit

Longitudinal Analysis

Here we want to distinguish the contributions of cross sectional and longitudinal information to the estimated relationship of respiratory infection and age.

Model 3: separate CS from LDA

- separate differences among sub-populations of children *at different ages and a fixed time (CS)* from changes in children *over time* (LD).
- $age_{ij}=age_{i1}+(age_{ij}-age_{i1})$
- age_{i1} is the age at entry (β_C)
- $age_{ij}-age_{i1}$ is the follow up time (β_L)

Model 3: separate CS from LDA

$$\begin{aligned}\text{logit}P(Y_{ij}=1) &= \beta_0+\beta_1\text{xer}_{ij}+\beta_2\text{gender}_i+\\ &+\ \beta_{C1}\text{age}_{i1}+\beta_{C2}\text{age}_{i1}^2+\\ &+\ \beta_{L1}(\text{age}_{ij}-\text{age}_{i1})+\beta_{L2}(\text{age}_{ij}-\text{age}_{i1})^2+\\ OR(Y_{ij},Y_{ik}) &= \gamma\end{aligned}$$

- β_C is the age effect on respiratory infection at the baseline age (*age at entry*)
- β_L change of the risk of respiratory infection as the children grow older (*follow up time*)

Results

- β_C suggests that the risk of RI climbs steadily in the first 20 months before declining
- β_L suggests that the risk of RI declines in the first 7-8 months of follow up before rising lately in life

Model 4: separate CS from LDA and adjusting for seasonality

$$\begin{aligned} \text{logit}P(Y_{ij}=1) &= \beta_0 + \beta_1 \text{xer}_{ij} + \beta_2 \text{gender}_i + \\ &+ \beta_{C1}\text{age}_{i1} + \beta_{C2}\text{age}_{i1}^2 + \\ &+ \beta_{L1}(\text{age}_{ij} - \text{age}_{i1}) \\ &+ +\beta_{L2}(\text{age}_{ij} - \text{age}_{i1})^2 + \textbf{harmonic terms} \\ OR(Y_{ij}, Y_{ik}) &= \gamma \end{aligned}$$

Summary of Results

- Pattern of convex relationship between age and the risk of respiratory infection appears to coincide with the pattern of seasonality
- If we include harmonic terms, then the longitudinal parameters (corresponding to the follow up in the table) are not statistically significant
- The longitudinal information (variation over time of respiratory infections versus variations over time of age) is highly confounded by seasonality
- Therefore, in presence of a strong seasonal signal, we can learn little about the effects of aging from data collected over 18 months period if we restrict our attention to longitudinal information
- However much can be learned by comparing children at different ages so long as we can assume that there are not cohort effects confounding the inferences about age
- *BE CAREFUL.. in longitudinal analysis always look for time-varying confounder*