

Marginal Methods

In many biomedical applications outcomes are binary, ordinal or a count. In such cases we consider extension of generalized linear models for analyzing discrete longitudinal data. These non-linear models require that a linear transformation of the mean response can be modeled in a regression setting. The non-linearity raises issues with the interpretation of the regression coefficients.

We let Y_i denote the response variable for the i^{th} subject, and:

$$X_i = \begin{pmatrix} X_{i1} \\ \vdots \\ X_{ip} \end{pmatrix}$$

is a $p \times 1$ vector of covariates. A generalized linear model for Y_i needs the following three-part specification:

1. A Distributional Assumption

Generalized linear models assume that the response variable has a probability distribution belonging to the exponential family (normal, bernoulli, binomial or Poisson). A feature of the exponential family is the variance can be expressed as:

$$\text{Var}(Y_i) = \phi v(\mu_i)$$

Where ϕ is a dispersion parameter and $v(\mu_i)$ is the variance function. For example:

- Variance function of normal distribution: $v(\mu) = 1$
- Variance function of Bernoulli: $v(\mu) = \mu(1 - \mu)$

2. A Link Function

The link function $g(\cdot)$ applies to the mean and then links the covariates to the transformed mean η such that:

$$g(\mu_i) = \eta_i$$

For example, the canonical link functions for some common distributions are:

Normal \rightarrow Identity: $\mu = \eta$

Bernoulli \rightarrow Logit: $\log\left(\frac{\mu}{1-\mu}\right) = \eta$

3. A Systematic Component

The systematic component specifies the effects of the covariates X_i on the mean of Y_i can be expressed in terms of the following linear predictor:

$$\eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}$$

Note that the term 'linear' refers to the regression parameters.

Binary response

Let Y_i denote a binary response variable with two categories such as presence or absence of a disease. The probability distribution is Bernoulli with $\Pr(Y_i = 1) = \mu_i$ and $\Pr(Y_i = 0) = (1 - \mu_i)$. Using the logit as the link function we have:

$$\log\left(\frac{\mu_i}{1 - \mu_i}\right) = \text{logit}(\mu_i) = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}$$

Where $\mu_i / (1 - \mu_i)$ are the odds of success

A unit change of X_{ik} changes the odds of success *multiplicatively* by a factor of $\exp(\beta_k)$.

The logistic regression model can be derived from the notion of a latent variable model. Suppose that L_i is a latent continuous variable which follows a standard logistic distribution $(0, \pi^2/3)$ and that a positive response is observed only when L_i exceeds some threshold τ , such that:

$$Y_i = 1 \text{ if } L_i > \tau$$

$$Y_i = 0 \text{ if } L_i \leq \tau$$

It can be shown that:

$$\begin{aligned} \Pr(Y_i = 1) &= \Pr(L_i > \tau) \\ &= \int_{\tau}^{\infty} \frac{\exp(u)}{\{1 + \exp(u)\}^2} du \\ &= \frac{\exp(-\tau)}{1 + \exp(-\tau)} \end{aligned}$$

Marginal Models

Suppose $Y'_i = (Y_{i1}, Y_{i2}, \dots, Y_{ip})$, a vector of correlated responses from the i^{th} subject. To analyze such correlated data we must specify or make assumptions about the multivariate or joint distribution, and to do so we will consider these main extensions of GLM: Marginal Models and Mixed Effects Models (next lecture)

A marginal model for longitudinal data has the following three-part specification:

1. $E(Y_{ij} | X_{ij}) = \mu_{ij}$ is assumed to depend on the covariates through a known link function:

$$g(\mu_{ij}) = \eta_{ij} = X'_{ij}\beta$$

2. We also identify a known variance function:

$$\text{Var}(Y_{ij} | X_{ij}) = \phi v(\mu_{ij})$$

Where ϕ is a scale parameter that may be constant or may vary at each measurement occasion.

3. The conditional within-subject association among the response measurements, given the covariates, is assumed to be a function of an additional set of association parameters, α .

Note that the systematic component is the key building block of a marginal model and specifies the model for the mean response at each occasion, $E(Y_{ij} | X_{ij})$, and its dependence on the covariates. Marginal responses also assume that the conditional mean of the j^{th} response given X_{i1}, \dots, X_{in} depends only on X_{ij} :

$$E(Y_{ij}|X_{i1}, \dots, X_{in}) = E(Y_{ij}|X_{ij})$$

With time-invariant or fixed time-varying covariates this assumption holds. It does not hold when a time-varying covariate varies randomly over time.

Note on association: We avoid using the term correlation. This is because 1) correlation is not a natural measure of within-subject association for discrete responses, and 2) the joint distribution of discrete responses is not often well specified or not easily tractable.

Generalized Estimating Equation

Since there is no convenient or natural specification of the joint multivariate distribution of Y_i for marginal models when the responses are discrete, we need an alternative to the Maximum Likelihood estimation. In 1986 Liang and Zeger proposed such a method based on the concept of 'estimating equations' which provides a general and unified approach for analyzing discrete and continuous responses with marginal models. For linear models Generalized Least Squares is a special case of estimating equations, for non-linear models the approach is called Generalized Estimating Equations (GEE).

Given a model for the pairwise correlations, the covariance matrix can be expressed as:

$$V_i = \phi A_i^{\frac{1}{2}} R(\alpha) A_i^{\frac{1}{2}}$$

$$A_i = \begin{bmatrix} v(\mu_{i1}) & 0 & \dots & 0 \\ 0 & v(\mu_{i2}) & \dots & 0 \\ \vdots & \dots & \ddots & \vdots \\ 0 & 0 & \dots & v(\mu_{ip}) \end{bmatrix}$$

And $R(\alpha)$ is the correlation matrix. V_i is known as a working covariance matrix to distinguish it from the true underlying covariance matrix.

Generalized Estimating Equations

An estimate of β can be obtained as the solution of the following generalized estimating equations:

$$\sum_{i=1}^n D_i' V_i^{-1} (Y_i - \mu_i) = 0$$

$$D_i = \begin{bmatrix} \partial \mu_{i1} / \partial \beta_1 & \partial \mu_{i1} / \partial \beta_2 & \dots & \partial \mu_{i1} / \partial \beta_k \\ \partial \mu_{i2} / \partial \beta_1 & \partial \mu_{i2} / \partial \beta_2 & \dots & \partial \mu_{i2} / \partial \beta_k \\ \vdots & \dots & \ddots & \vdots \\ \partial \mu_{ip} / \partial \beta_1 & \partial \mu_{ip} / \partial \beta_2 & \dots & \partial \mu_{ip} / \partial \beta_k \end{bmatrix}$$

D_i can be thought of as a matrix that transforms from the original units of μ_{ij} to the units of $g(\mu_{ij})$

The generalized estimating equations are functions of both β and α and in general have no closed-form solution. In this case, the following two-stage estimation procedure is required:

1. Given current estimates of α and ϕ , V_i is estimated and an updated estimate of β is obtained as the solution of GEE
2. Given the current estimate of β updated estimates of α and ϕ are obtained based on the standardized residuals:

$$e_{ij} = (Y_{ij} - \hat{\mu}_{ij}) / \sqrt{v(\hat{\mu}_{ij})}$$

For example, ϕ can be estimated by:

$$\hat{\phi} = \frac{\sum_{i=1}^N \sum_{j=1}^{n_i} e_{ij}^2}{\sum_{i=1}^N n_i}$$

α can be estimated in a similar way:

$$\hat{\alpha}_{jk} = \left(\frac{1}{\hat{\phi}N} \right) \sum_{i=1}^N e_{ij} e_{ik}$$

And steps 1 and 2 are repeated until convergence

Properties of GEE Estimators

β_{hat} is a consistent estimate of β , with very high probability and sufficiently large N then $\beta_{\text{hat}} \sim \beta$. The consistency of β_{hat} depends on the correct specification of the mean, but it still holds even if the covariance of Y_i has been misspecified.

$\hat{\beta}$ is $\hat{\beta} \sim N(\beta, B^{-1}MB^{-1})$ where

$$B = \sum_{i=1}^N D_i' V_i^{-1} D_i$$

$$M = \sum_{i=1}^N D_i' V_i^{-1} \Sigma_i V_i^{-1} D_i$$

By replacing D_i , V_i , and Σ_i by their estimates, we get the empirical or sandwich estimator.

Note that if we model V_i correctly then $V_i = \Sigma_i$ and $\text{Var}(\beta) = B^{-1}$

Daignostics

We can easily calculate residuals:

$$r_{ij} = Y_{ij} - g^{-1}(X_{ij}'\hat{\beta})$$

Since $\text{Var}(r_{ij}) = f(\mu_{ij})$ it is preferable to use studentized residuals:

$$e_{ij} = \frac{Y_{ij} - g^{-1}(X_{ij}'\hat{\beta})}{\sqrt{\phi v(\hat{\mu}_{ij})(1 - h_{ij})}}$$

where h_{ij} is the leverage of the j_{th} observation on the i^{th} individual and describes the influence each observation has on its own predicted value

We can use diagnostic plots similar to the continuous case to find outlying observations; such as the Mahalanobis-type statistic:

$$d_i = r_i' \hat{V}_i^{-1} r_i$$

which has a chi-squared distribution if the model for the mean is correctly specified and $V_i \sim \Sigma_i$

SAS Code

* No repeated measures, logistic regression, uses maximum likelihood;

```
proc logistic data=s857.BPD plots=EFFECT descending;
class BPD;
model BPD=Weight;
title1 'Simple Logistic Regression using PROC LOGISTIC';
run;
```

* Use GEE, can use if no repeated measures, interpret estimate, never use genmod in non-repeated measures, just here as intro;

```
proc genmod data=s857.BPD descending;
class BPD;
model BPD=Weight/DIST=binomial LINK=LOGit;
title1 'Simple Logistic Regression using PROC GENMOD';
run;
```

```
proc logistic data=s857.BPD plots=EFFECT descending;
class BPD;
model BPD=Weight Gest_Age Toxemia;
title1 'Multiple Logistic Regression using PROC LOGISTIC';
run;
```

*Repeated measures, y = obesity;

```
proc genmod data=s857.BPD descending;
class BPD;
model BPD=Weight Gest_Age Toxemia/DIST=BINOMIAL LINK=LOGIT;
title1 'Multiple Logistic Regression using PROC GENMOD';
run;
title1 'Marginal Logistic Regression Model for Obesity';
title2 'Muscatine Coronary Risk Factor Study';
```

*have increase in prevalence in obesity from 8 to 12 then becoems constant for both genders;

*repeated measures since each person has 4 binary measures (obesity status at 4 ages);

```
proc freq data=s857.muscatine;
where occasion=2;
by gender;
tables age*y;
run;
```

*apply quad or cubic model since trend not linear;

* 12 is pop mean, dont want multicollinearity when using quad or cubic in model;

* may use these in the model;

data muscatine;

set s857.muscatine;

cage=age - 12;

cage2=cage*cage;

cage3=cage2*cage;

run;

*quad model with respect to age;

* use repeated statement by id (ppl have mult measurements), that variable should also be in class statement;

*type = compound symmetry, so $R(\alpha) = [1 \text{ } \rho \text{ } \rho \dots \rho, \rho \text{ } 1 \text{ } \rho \text{ } \rho, \rho \text{ } \rho \text{ } 1 \text{ } \rho \dots]$;

/*using a compound symmetry structure working correlation*/

proc genmod data=muscatine ;

class id occasion;

model y(event='1')=gender cage cage2 / dist=bin link=logit

type3 wald;

repeated subject=id / type=cs corrw covb;

run;

*look at Analysis of GEE Parameter Estimates table, need to exp estimate if want OR;

*corr parameter is 0.54 from output, no clear interpretation;

/*Using Log OR correlation structure*/

*use logor structure, fullclust says go and find all the associations btwn all the different measures (subj have at most 3 diff measurements) 1&3,1&2,2&3;

proc genmod data=muscatine ;

class id occasion;

model y(event='1')=gender cage cage2 / dist=bin link=logit

type3 wald;

repeated subject=id / withinsubject=occasion logor=fullclust;

run;

*look at Analysis of GEE Parameter Estimates table;

*three alphas;

* $\alpha_1 = \log OR(Y_{i1}, Y_{i2})$, $OR(Y_{i1}, Y_{i2}) = P(Y_{i1} = 1, Y_{i2} = 1) * P(Y_{i1} = 0, Y_{i2} = 0) / (P(Y_{i1} = 1, Y_{i2} = 0) * P(Y_{i1} = 0, Y_{i2} = 1))$;

* to find OR exp estimate, is OR of having same outcome in occasion 1 and occasion 2 vs diff outcome;

* $\alpha_2 = \log OR(Y_{i1}, Y_{i2})$;

* $\alpha_3 = \log OR(Y_{i2}, Y_{i3})$;

*1,2 and 2,3 are only two years apart;
 * maybe want $\alpha_1 = \log OR(Y_{i1}, Y_{i2}) = \log OR(Y_{i2}, Y_{i3})$;

/*Equivalent model*/

* three columns, three alphas;
 *first row: for 1 2, this is only α_1 ;
 *second row: for 1 3, this is only α_2 ;
 *third row: for 2 3, this is only α_3 ;
 proc genmod data=muscatine descending;
 class id occasion;
 model y=gender cage cage2 / dist=bin link=logit
 type3 wald;
 repeated subject=id / withinsubject=occasion
 logor=zrep((1 2) 1 0 0,
 (1 3) 0 1 0,
 (2 3) 0 0 1) ;
 run;

/*The same association between any two occasions */

* compound symmetry but with logOR presentation;
 * only one alpha that is same for (1 2) (1 3) (2 3);
 * this parametrization works if same measurement on same subjects at same time (i.e. age 8, 10, 12 for all)
 like clinical trial design;
 proc genmod data=muscatine descending;
 class id occasion;
 model y=gender cage cage2 / dist=bin link=logit
 type3 wald;
 repeated subject=id / withinsubject=occasion
 logor=zrep((1 2) 1 ,
 (1 3) 1 ,
 (2 3) 1);
 run;

/*Testing whether the association between occasions 1-2
 and 2-3 are the same */

* three columns, three alphas;
 *first row: for 1 2, this is only α_1 ;
 *second row: for 1 3, this is only α_2 ;
 *third row: for 2 3, this is $\alpha_1 + \alpha_3$;
 * if α_3 is zero, then (1 2) is same as (2 3), get rid and only have 2 alphas;
 proc genmod data=muscatine descending;

```

class id occasion;
model y=gender cage cage2 / dist=bin link=logit
type3 wald;
repeated subject=id / withinsubject=occasion
    logor=zrep( (1 2) 1 0 0,
                (1 3) 0 1 0,
                (2 3) 1 0 1) ;

run;

/*Fitting a model with the association between occasions 1-2
and 2-3 are the same */
* store p1 = store residuals;
proc genmod data=muscatine descending;
    class id occasion;
    model y=gender cage cage2 / dist=bin link=logit
    type3 wald;
    repeated subject=id / withinsubject=occasion
        logor=zrep( (1 2) 1 0 ,
                    (1 3) 0 1,
                    (2 3) 1 0);

store p1;
run;

* plot predicted values for prevalence for diff genders at diff age for quad model;
ods html style=journal;
proc plm source=p1;
    score data = muscatine out=pred /ilink;
run;
proc sort data = pred;
    by gender age;
run;
proc sgplot data = pred;
    series x = age y = predicted /group=gender;
run;

/*Cubic Model*/
proc genmod data=muscatine descending;
    class id occasion;
    model y=gender cage cage2 cage3/ dist=bin link=logit
    type3 wald;
    repeated subject=id / withinsubject=occasion

```



```

logor= zrep( (1 2) 1 0,
            (1 3) 0 1,
            (2 3) 1 0);

store p1;

run;

ods html style=journal;
proc plm source=p1;
  score data = muscatine out=pred /ilink;
run;
proc sort data = pred;
  by gender age;
run;
proc sgplot data = pred;
  series x = age y = predicted /group=gender;
run;

/* Example of a quadratic model with gender age interaction */
proc genmod data=muscatine descending;
  class id occasion gender ;
  model y=gender cage cage2 gender*cage gender*cage2 / dist=bin link=logit
  type3 wald;
  contrast 'Age X Gender Interaction' gender*cage 1 -1, gender*cage2 1 -1 /wald;
  repeated subject=id / withinsubject=occasion logor=fullclust;
run;

/*****
Diagnostics(Optional)
*****/

*assess statement = uses bootstrap to look at sum of residuals, should be centered around 0;
proc genmod data=muscatine descending;
  class id occasion;
  model y=gender cage cage2 cage3/ dist=bin link=logit
  type3 wald;
  repeated subject=id / withinsubject=occasion
  logor=zrep( (1 2) 1 0,
            (1 3) 0 1,
            (2 3) 1 0);
  assess var=(cage)/resamples=10000 seed=7435865;
run;
ods rtf close;

```

Revision #5

Created 23 February 2023 19:11:51 by Elkip

Updated 23 February 2023 21:40:32 by Elkip