A REGRESSION ANALYSIS OF NCHS LONGITUDINAL DATA

Jai W. Choi
National Center for Health Statistics, 6525 Belcrest Rd. Hyattsville, MD 20782

Linear model, Complex survey data, Correlation, Multiplicative model, covariance

SUMMARY

The errors may be accumulated through the process of obtaining data, including time correlations. We define a multiplicative effect modeland, and obtain a variance of data with time correlation, sampling and subject errors. This variance accounts are applied to longitudinal data.

INTRODUCTION

NCHS collects data from each sampled person for a number of years in follow-up studies, and repeated responses or measurements form one person may be correlated. Other example is that, since deaths occur in a same population, death rates may be correlated during consecutive years. Sampling errors may arise from NCHS sample data.

Liang and Scott (1986) corrected longitudinal data for time correlation. Thall and Vail (1990), Morton (1987) and Firth and Harris (1991) used multiplicative model to correct time correlation or other errors.

We assume that each y_i has the density of

$$f(y_i)=\exp\{(y_i\theta_i-b(\theta_i))/a(\sigma^2) + c(y_i,a(\sigma^2)\}.$$

The first two moments of y_i are given by $E(y_i)=b'(\theta_i)$ and $var(y_i)=b''(\theta_i)a(\sigma^2)$, which is the product of two function, the variance function $b''(\theta_i)$ and the dispersion function $a(\sigma^2)$. $a(\sigma^2)$ is σ^2 for normal density.

The quasi-likelihood may be applied to derive regression parameter β from data $y_1, \ldots, y_i, \ldots, y_M$ in a model defined by $E(y_i) = \mu_i(\beta)$, and $cov(y_i) = V_i(\mu_i)$, where the mean μ_i and variance V_i are known. The mean μ_i is often determined by known independent variables $x_{i1}, \ldots, x_{ip}$, possibly by a model

$$g(\mu_i) = x_{i1}\beta_1 + , \ldots, + x_{ip}\beta_p$$

in which the link function g is known. An estimate of β is obtained by solving quasi-likelihood equations for each y.
Let $\partial\mu/\partial\beta = D$.

$$(1.1)\quad U(\mu)= D^T V^{-1}(y-\mu) = 0.$$

When V is a diagonal matrix, (1.1) is the same as the weighted least square case, but the variance-covariance matrix V is rarely a diagonal matrix.

In Section 2, a multiplicative effect model is defined, and the covariance V is obtained from this model. The point estimation is discussed in Section 3 when log link function for $\mu(\beta)$ is assumed. In Section 4, we analyze NCHS sample rate based on this $V(\mu)$.

2 MULTIPLICATIVE MODEL

Let z_{3ij}, and z_{2ik} indicate the effects of subject, and sampling, respectively, and z_{1ikt} repeated measurements or responses ($1 \le i \le M$), ($1 \le j \le N_i$), ($1 \le k \le D_i$), and ($1 \le t \le T$). M is the number of data sets, and the M monthly data are assumed to be independent. N_i is the units in the i-th data set, D_i the number of subjects with symptom among N_i units, and T the number of time points for repeated measurements.

The variables z_{3ij}, z_{2ik}, and z_{1ikt} may be continuous or discrete. The subscript numbers indicate respective level. The level 1 is the lowest level and the level 3 the highest. Let weight w_i be the known numbers. Define a multiplicative model

$$(2.1)\quad y_{ijkt} = w_i z_{3ij} z_{2ik} z_{1ikt}.$$

Suppose that z_{3ij}, z_{2ik}, and z_{1ikt} have mean μ_3, μ_2, and μ_{1t}, and variance ϵ_{3j}, ϵ_{2k}, and $\epsilon_{1kt't'}$, respectively. The (jkt, j'k't')-th element of covariance matrix for y_{ijkt} is

(2.2) $w_i^2\{\epsilon_{1kt},E(z_{2ik}z_{2ik'})E(z_{3ij}z_{3ij'})$

$+\mu_{1t}\mu_{1t'}\epsilon_{2kk'}E(z_{3ij}z_{3ij'})+\mu_{1t}\mu_{1t'}\mu_2^2\epsilon_{3jj'}\}$,

where $\epsilon_{1kt'} = Cov(z_{1ikt}\, z_{1ikt'})$ for the two repeated observations on the k-th subject. If z_{3ij} and z_{2ik} are all independent with common variance, ϵ_3 and ϵ_2 respectively, (2.2) can be written as

(2.2a) $w_i^2\{\epsilon_{1kt'}(\epsilon_2+\mu_2^2)(\epsilon_3+\mu_3^2)$

$+\mu_{1t}\mu_{1t'}\epsilon_2(\epsilon_3+\mu_3^2)+\mu_{1t}\mu_{1t'}\mu_2^2\epsilon_3\}$.

It is interesting to compare above results with the variances of additive model,

$y_{ijkt} = \mu_{ijkt}+z_{3ij}+z_{2ik}+z_{1ijkt}$,

would be the same as variance (2.2a), replacing $w_i^2\,\epsilon_{1kt'}(\epsilon_2+\mu_2^2)(\epsilon_3+\mu_3^2)$ with $var(z_{3ij})$, $w_i^2\,\mu_{1t}\mu_{1t'}\,\epsilon_2(\epsilon_{3j}+\mu_{3j}^2)$ with $var(z_{2ik})$, and $w_i^2\,\mu_{1t}\mu_{1t'}\mu_2^2\,\epsilon_3$ with $var(z_{1ijkt})$. The mean of additive y_{ijkt} is $(\mu_{ijkt}+\mu_{1t}+\mu_2+\mu_3)$, while the mean of multiplicative model is μ_{ijkt}.

Denote the relative variances $C_3 = \epsilon_3/\mu_3^2$, $C_2 = \epsilon_2/\mu_2^2$, and relative covariance $C_{1kt'} = \epsilon_{1kt'}/(\mu_{1t}\,\mu_{1t'})$. Let $E(y_{ijkt})=w_i\mu_{1t}\mu_2\mu_3 = \mu_{ijkt}(\beta)$. (2.2a) can be expressed as

(2.3) $\mu_{ijkt}\mu_{ijkt'}\{C_{1kt'}(C_2+1)(C_3+1)+C_2(C_3+1)+C_3\}$.

We may use variable $y_{ikt} = \Sigma_{(j=1,Ni)}\, w_i z_{3ij} z_{2ik} z_{1ikt}$ to simplify covariance structure. $\mu_{ikt} = N_i\, w_i\mu_1\mu_2\mu_{3t}$, and the (kt, kt')-th element of V_i is

(2.4) $\mu_{ikt}\mu_{ikt'}\{C_{1kt'}(C_2+1)(C_3+1)+C_2(C_3+1)+C_3\}/N$

The inverse of (2.4) can be easily obtained as $C_{1kt'}$ is specified. For instance, if z_{1ikt} follow the multinomial distribution, then $C_{1tt'} = (1/\mu_{1t} - 1)$ for $t = t'$ and $C_{1tt'} = -1$ for $t \neq t'$.

$$V_i = D_i + c_i\, u\, u'$$

where D_i is the diagonal, ikt-th element $\mu_{ikt}^2\{(C_2+1)(C_3+1)/\mu_{1t} + C_2(C_3+1)+C_3\}/N_i$ and the vector μ, ikt-th element μ_{ikt}, $c_i=(1/N_i)\{-(C_2+1)(C_3+1)+ C_2(C_3+1) +C_3\}$.

$$V_i^{-1} = D_i^{-1} + d_i\, u^*\, u^{*\prime}$$

$d_i=c_i[1+c_iN_i\Sigma_t((C_2+1)(C_3+1)/\mu_{1t} + C_2(C_3+1)+C_3]^{-1}$, and u^* is the vector, t-th element of $N_i[(C_2+1)(C_3+1)/\mu_{1t}+ C_2(C_3+1) + C_3]^{-1}$.

Example 1

Let the top level 3 is the subject effect $z_{3ij} = 1$ if the j-th person has a condition and $z_{3ij} = 0$ otherwise. Its mean is μ_3 and variance ϵ_3. The level 2 is the sampling of d_i out of D_i persons with a condition. Let $z_{2ik} = 1$ if the k-th person is sampled among the subjects with symptom, and $z_{2ik} = 0$ otherwise. Its mean is μ_2 and variance ϵ_2. Depending on the type of sampling method, the mean and variance are determined.

The level 1 is the time effect. Each of d_i sample units is measured or observed over all time points except random missing, and these observations are correlated. Let $z_{1ikt} = 1$ if the k-th sampled unit is still have symptom at the t-th time with mean μ_{1t} and variance ϵ_{1kt}, and $z_{1ikt} = 0$ otherwise.

The variable z_{1ikt} of time effect arise for the original units only when the previous two random events happened. For instance, the time effects arise only if the subject is sampled after this subject found to have a symptom.

The second and third level variables are all independent, and the first level variables of repeated measurements for each subject are correlated. When the distribution is not specified for any of these variables, V_i is already given in (2.2).

However, if the second and third level variables were binomial, then the (kt, kt')-th element of V_i would be

(2.5) $N_iw_i^2\{\epsilon_{1tt'}\mu_2\mu_3+\mu_{1t}\mu_{1t'}\epsilon_2\mu_3+\mu_{1t}\mu_{1t'}\mu_2^2\epsilon_3\}$

$=N_iw_i^2\mu_2^2\mu_3^2\mu_{1t}\mu_{1t'}\{C_{1tt'}/(\mu_2\mu_3)+C_2/\mu_3+C_3\}$.

Furthermore, if the first level variables z_{1ikt} are known to be multinomial, V_i can be expressed as

(2.6) $V_i = D_i + c\,\mu_i\mu_i^T$,

where D_i is the diagonal matrix with (kt)-th

element $w_i\ \mu_{ikt}$, where $\mu_{ikt} = N_i w_i \mu_{1t}\mu_2\mu_3$, $c_i = (-1/\mu_2\mu_3 + C_2/\mu_3 + C_3)/N_i$, and μ_i is column vector with t-th element μ_{ikt}.

$$(2.7) \qquad V_i^{-1} = D_i^{-1} + d_i\ \nu_i\ \nu_i^T,$$

where $d_i = -c_i/(1 + c_i\ N_i\mu_1 + \mu_2\mu_3)$, $\nu_i^T = 1/w_i\ (1,\ldots,1)$.

Note that V_i depends not only on the distribution of z_{3ij}, z_{2ik}, and z_{1ikt} in each stage but also on the link function. For instance, if link g is log link function, $\mu_{ijkt}(\beta) = \exp(X'\beta)$, and the variance is now the function of $\exp(X'\beta)$. Defining $D_i = \partial\mu/\partial\beta$ be the matrix of partial differentiations of this mean with respect to β, and using V_i^{-1} of (2.7), the quasi-score equation for the a regression parameter β_r is now derived from (1.1):

$$(2.8)\ [\Sigma_{ikt}\ X_{ikt}(y_{ikt}-\mu_{ikt}) + d_i\Sigma_{ikt}X_{ikt}\mu_{ikt} + \Sigma_{ikt}(y_{ikt}-\mu_{ikt})]\ /w_i = 0$$

Example 2

Liang and Scott (1989) adjusted original V for the correlation of repeated responses from a same subject. The correlation of repeated responses, expressed in a correlation matrix R, is included as

$$V^* = V_i^{1/2}\ R\ V_i^{1/2}.$$

They used V* instead of V for point estimation. V_i is now the covariance matrix for two level model $\Sigma_j\ w_i\ z_{3ij}\ z_{2jk}$.

We may assume that z_{3ij} and z_{2jk} are binomial variables in order to compare this result with the covariance (2.6) where we used three level model is used. For the comparison of two covariances, V_i of (2.6) and V_i^*, we further assume that the third level variables are also independent, and that R_i is diagonal matrix, the kt-th element $\mu_{1t}(1 - \mu_{1t})$.

The kt-th element of V_i^* and V_i are respectively:

$$N_i\ w_i^2\ \mu_2\ \mu_3\ \mu_{1t}(1-\mu_2\mu_3)(1-\mu_{1t}),$$

$$N_i\ w_i^2\ \mu_2\ \mu_3\ \mu_{1t}(1-\mu_2\mu_3\mu_{1t}).$$

V_i^* and V_i are not likely the same even in this simple case.

One reason of the difference is that the covariance depends on R_i for V_i^*, the number of levels, and distribution of variables. Thus the estimation would hardly be the same under these two diverse models even if the objective is the same. Further empirical study might be needed to compare these two models.

Example 3

If the third level is Poisson process instead in Example 1, and the second level is simple random sample, and the first level follows multinomial, V_i is a diagonal matrix, the (ktt')-th element

$$(2.9)\ N_i\ w_i^2\ \{\epsilon_{1tt'}\ \mu_2\ (\mu_3+\mu_3^2) + \mu_{1t}\mu_{1t'}\ \epsilon_2\ (\mu_3+\mu_3^2) + \mu_{1t}\mu_{1t'}\ \mu_2^2\ \mu_3\}$$

$$= N_i\ w_i^2\mu_2^2\mu_3^2\mu_{1t}\mu_{1t'}\ \{C_{1tt'}(1+\mu_3)/\mu_2\mu_3 + C_2(1+\mu_3)/\mu_2\mu_3 + 1/\mu_3\}$$

where for multinomial case, the matrix form and its inverse of (2.9) is the same as (2.6) and (2.7) with new

$$c_i = [-(1+\mu_3)/\mu_2\mu_3 + C_2(1+\mu_3)/\mu_3 + 1/\mu_3]/N_i,$$

Example 4

Note that the change of levels would give different variance. For instance, z_{1ij}, z_{2ik}, and z_{3ikt}, reversing the ordering of levels, the level 1 being z_{1ij}, and the level 3 z_{3ikt}, and the z_{2ik} remaining at the same level, we obtain a different variance from (2.2). The variables z_{1ij} and z_{2ik} are all independent. Let z_{1ij} be the 1st level with Poisson expectation,

$$E_1(y_{ijkt}/z_{1ij}z_{2ik}z_{3ikt}) = w_i\ \mu_1 z_{2ik}z_{3ikt}.$$

$Cov_1(y_{ijkt}\ y_{ijkt'}) = 0$, and define Kronecker delta function $\delta_{1tt'}$, the (jkt,jkt')-th element of V_i is

$$(2.10)\ w_i^2\{\mu_1\mu_2\mu_{3t}\delta_{1tt'} + \mu_1^2\epsilon_2(\epsilon_{3ktt'}+\mu_{3t}\mu_{3t'})$$

$$+ \mu_1^2 \mu_2^2 \epsilon_{3ktt'}\}$$

$$= w_i^2\{\mu_1\mu_2\mu_{3t}\delta_{1tt'} + \mu_1^2\mu_2^2\mu_{3t}\mu_{3t'}[C_2(C_{3ktt'}+1) + C_{3ktt'}]\}$$

$$= w_i^2 \mu_1\mu_2\mu_{3t}\delta_{1tt'} + b_{ktt'} w_i^2 \mu_1^2 \mu_2^2 \mu_{3t}\mu_{3t'},$$

where $b_{ktt'} = [C_2 + (C_2+1)C_{3ktt'}]$.
If z_{3kt} were multinomial variables,
$\mu_{ikt} = N_i w_i \mu_1\mu_2\mu_{3t}$, the covariance matrix of y_{i+kt} is

$$(2.11) \qquad V_i = D_i + c_i \mu_i \mu_i',$$

D_i is diagonal matrix kt-th element of $\mu_{ikt}\{w_i + (C_2+1)\mu_{ikt}/N_i\mu_{3t}\}$, and the vector $\mu_i = (\mu_{i11}, \ldots, \mu_{ikt}, \ldots, \mu_{idT})$, $c_i = 1/N_i$.

Example 5

The NCHS publications is the estimate for annualized cause specific rate of deaths per 100,000 persons for the i-th month. The weight w_i are known.

Let N_i be the number for the population of month i, D_i the number of deaths outm of N_i, and d_i the number of sample out of D_i deaths.

From Example 3, the first level z_{1ij}'s (i = 1, .., M, j = 1, ..., N_i) of random deaths are independent and distributed as Poisson with mean $E(z_{1ij}) = D_i/N_i = \mu_1$.

The second level effect z_{2ik} (k = 1, ..., D_i) of sampling of selecting d_i deaths by simple random sample. $E(z_{2ik}) = \mu_2 = d_i/D_i$, and $\epsilon_2 = V(z_{2ik}\, z_{2ik'}) = \mu_2(1-\mu_2)$ if k = k' for a large D_i, and = 0 otherwise.

The third level $z_{3ikt} = 1$ if the k-th death falls in the year t, and zero otherwise with $E(z_{3ikt}) = \mu_{3t}$ and $var(z_{3ikt}) = \epsilon_{3tt'}$. We may assume that the number of deaths in the consecutive years are correlated, and the third level also follows multinomial distribution. $var(z_{3ikt}) = \mu_{3t}(1 - \mu_{3t})$.

Now the impacts of these three random errors, death process, sampling, time correlation on the variance V_i are investigated.

For these assumptions, then matrix V_i is the same as the variance (2.13) except ϵ_2 replaced with $\mu_2(1 - \mu_2)$.

Example 6

In Example 3, the correlation $\alpha_{tt'}$ between t-th and t'-th time for the k-th subject may be defined

$$cov(z_{3ikt}\, z_{3ikt'})/\{var(z_{3ikt})var(z_{3ikt'})\}^{1/2},$$

$$= C_{3ktt'}\, \mu_{3t}\, \mu_{3t'}/[\epsilon_{3t}, \epsilon_{3t'}]^{1/2}.$$

or $C_{3ktt'} = \alpha_{tt'}\,(C_{3t}\, C_{3t'})^{1/2}$. We may replace $C_{3ktt'}$ in (2.10) with this definition. Define

$$a_{ktt'} = \{[\alpha_{tt'}(C_{3t}\, C_{3t'})^{1/2}](C_2+1)+C_2\}.$$

$\mu_{ikt} = N_i w_i \mu_1\mu_2\mu_{3t}$. The (kt, kt')-th element of V_i for y_{ikt} is,

$$(2.13) \qquad w_i\, \mu_{ikt} + \mu_{ikt}\, \mu_{ikt'}\, a_{ktt'}/N_i,$$

3. ESTIMATION

Let y_i be the (1 x d) vector of dependent variables for month i, mean $u_i = (\mu_{i11},\ldots,\mu_{ikt},\ldots, \mu_{idT})'$ and covariance V_i. Let $X_i = (x_{i111}, \ldots, x_{iktr}, \ldots, x_{idTp})'$ be the p x (dT) matrix of covariates.

Denote linear predictor $\eta_i = (\eta_{i11},\ldots,\eta_{ikt}, \ldots,\eta_{idT})'$. A link function g relates the predictor η_i to the expected value, $g(u_i) = \eta_i$. If the link function is wrong and $\eta_i = X_i\beta$ is not a correct predictor, the variance would not be correct as the variance V_i depends on $u_i(\beta)$ through link.

Let $S_i = y_i - u_i$ with $E(S_i) = 0$, and $V_i(\mu_i)$ be the covariance of $var(y_i)$, and $D_i = \partial u_i/\partial\beta$. The estimating equation is given as (1.1),

$$(3.1) \qquad U(\beta) = \Sigma_i\, D_i^T\, V_i^{-1}\, S_i = 0.$$

β is the solution of $U(\beta) = 0$, in which the variance V_i is not only a function of β and σ, but of u_{1t}, u_2, and u_3 as well. Assuming that σ are known, the equation (3.1) may be expressed as

$$(3.2) \qquad \Sigma_i\, U_i(\beta, \hat{u}_{1t}(\beta), \hat{u}_2(\beta)) = 0$$

where $\hat{u}_2(\beta) = \hat{u}_2(\beta, \hat{u}_3(\beta))$, and $\hat{u}_{1t}(\beta) = \hat{u}_{1t}(\beta, \hat{u}_3(\beta), \hat{u}_2(\beta))$. $\hat{\beta}$ is now defined to be the solution of this equation.

Theorem 1. Under basic conditions for Taylor expansion, and

(i) $\hat{u}_{1t}$, $\hat{u}_2$, and $\hat{u}_3$ are the $M^{1/2}$-consistent estimates of u_{1t}, u_2, and u_3, respectively given β and σ, that is, $o_p(1)$;

(ii) $(\partial/\partial u_2)\ \hat{u}_3(\beta,\hat{u}_2(\beta))$, $(\partial/\partial u_2)\ \hat{u}_{1t}(\beta,\hat{u}_2(\beta),\hat{u}_3(\beta))$, and $(\partial/\partial u_3)\ \hat{u}_{1t}(\beta,\hat{u}_2(\beta),\hat{u}_3(\beta))$ are bounded in probability, $O_p(1)$;

then $M^{1/2}(\hat{\beta} - \beta)$ has asymptotically normal with mean 0 and variance

(3.3) $(\Sigma_{i\epsilon M}(D_i^T V_i^{-1} D_i))^{-1}$

$[\Sigma_{i\epsilon M}\ D_i^T V_i^{-1} \mathrm{cov}(y_i) V_i^{-1} D_i](\Sigma_{i\epsilon M}(D_i^T V_i^{-1} D_i))^{-1}$.

If $\mathrm{cov}(y_i) = V_i$, (3.3) reduces to $(\Sigma_{i\epsilon M}(D_i^T V_i^{-1} D_i))^{-1}$. When a link function is specified, we can obtain the explicit form of V_i. Variance of $\hat{\beta}$ may be correctly estimated by replacing $\mathrm{cov}(y_i)$ with $S_i^T S_i$. $S_i^T S_i$ may be more efficient then V_i when the model or link used for the derivation of V_i is not correct.

3.2. ITERATION

We may begin iteration with $\hat{\beta}^0$ substantially close to $\hat{\beta}$. The sequence of parameter estimates are generated by iterative reweighted least square method, dropping the subscripts,

$$(3.4)\quad \hat{\beta}^{r+1} = \hat{\beta}^r + (\hat{D}^T \hat{V}^{-1} \hat{D})^{-1} (\hat{D}^T \hat{V}^{-1} \hat{S})$$

The estimate $\hat{\beta}$ may be obtained by iterating until it converges. We may start the iteration with ordinary least square estimate of β. Convergence criterion is to stop the iteration at $(r + 1)$ step when $MAX|(\hat{\beta}^{r+1} - \hat{\beta}^r)/\hat{\beta}^r| \le 10^{-5}$.

Provided that the eigenvalues of $\hat{D}^T \hat{V}^{-1} \hat{D}$ are sufficiently large, the second term of (3.4) is negligible. Then, we may take the first round approximation $\hat{\beta}^1 = \hat{\beta}$, even when $\hat{\beta}^1$ is not a computable statistics. When V_i is set equal to one of the common densities, existing GLIM software provides the estimates of the parameters.

4. FOLLOW-UP STUDY

The longitudinal study of aging follows a cohort of older individuals over time, and provided information on changes of functioning, living arrangement, health care, and death. A subsample of the 1986 sample was taken among those 70 years of age and over in 1984 HIS base survey. The original number of HIS sample remained at 7,527 eligible for subsampling. The persons in the subsample were interviewed during 1986, 1988, and 1990. There are several variables of interest, which show the deterioration as time passes. However the data tape was not available to me, and could not apply our method to the follow-up data.

REFERENCES

D. Firth and I.R. Harris (1991) Quasi-likelihood for multiplicative random effects. Biometrika 78, 3, pp. 545-55.

Kung-Hee Liang and Scott L. Zeger (1986). Longitudinal data analysis using generalized linear models, Biometrika 73. 1. pp 13-22.

Richard Morton (1987). A generalized linear model with nested strata of extra-Poisson variation. Biometrika, 74, 2, 247-57.

Peter F. Thall and Stephen C. Vail (1990). Some Covariance Models for Longitudinal Count Data with Overdispersion. Biometrics 46. 657-671.