CHAPTER FOUR
REPEATED MEASURES IN TOXICITY TESTING

4.1 Introduction

The repeated measures study is a very commonly used experimental design in toxicity testing because it not only allows one to investigate the effects of the toxicants, but also enables one to look at how the effect changes over a time period. The repeated measures study in toxicology shares similarities with the split-plot design, which is very popular in agricultural research. It is carried out with at least two factors of treatments in a nested fashion where a whole plot is divided into several sub-plots. The first factor of treatments is randomly assigned to the whole-plots, then the second factor of treatments is randomly assigned to the sub-plots within each whole plot. When the subjects are measured repeatedly through time, each time slot can be thought of as a sub-plot within each subject, and each subject can be thought of as a whole plot in the experiment. Some authors call the whole-plots or subjects in the experiment clusters (Longford, 1993). However, the measurements within each subject are very likely to be correlated and it is impossible to randomization time. Hence, the analysis requires a model that includes an special covariance structure for the observations.

This type of repeated measures study is often referred to as a longitudinal study. Longitudinal studies have some specific advantages over classical cross-sectional studies (Diggle *et al*. 1994). The *Ceriodaphnia dubia* test mentioned in Chapter Three could be considered as a longitudinal study since the number of offspring is obtained on a

daily basis. The practical problem again is the existence of mortality and the mixing of zeroes in the data. Moreover, the repeated measurements within the same subject would likely be correlated. The model proposed in the Chapter three only deals with the daily egg counts or the total egg counts. It does not address the repeated measure effect or the time effect of the experiment. With repeated measurements through time, there are different approaches that can be used. The purpose of this chapter is to examine the existing approaches and modify their application in longitudinal toxicity studies to handle the problem of mixing zeroes due to mortalities.

4.2 Common approaches to Longitudinal Data

In general, there are three approaches to handle longitudinal data. The first one is the time-wise approach which is also the easiest way to handle longitudinal data. This approach treats each repeated measurement as an independent experiment and analyzes the data on a time by time basis. This is a rather straightforward approach, since any standard statistical method can be applied; for example, one can use a time by time ANOVA or a time by time regression analysis to examine the longitudinal data.

The second approach is the derived variable approach which summarizes the repeated values into one number and then analyzes the summary variable as a function of the treatments. This method is sometimes referred as the two-stage method which can be dated back to Wishart (1938). It works well when the covariates are independent of time. But, it is not very useful when the covariates change over time.

The last approach is based on the generalized estimating equations (GEE) by Liang and Zeger (1986). This method combines estimating function theory (Godambe 1960, 1976; Godambe and Heyde, 1987; Godambe and Thompson, 1989; Godambe and Kale, 1991; Schabenberger and Gregoire, 1995) with a multivariate extension of quasi-likelihood principle (Wedderburn, 1974; McCullagh, 1983; McCullagh and Nelder, 1989;

Nelder and Lee, 1992). By allowing a block–diagonal covariance matrix of the response, it estimates both the regression coefficients and the covariance structure at the same time. As mentioned earlier in this Chapter, the mortality problem in chronic toxicity testing further complicates the analyses in these approaches. The sections below will discuss the following:

- the issue of incorporating mortality zeroes,
- the issue of handling longitudinal data, and
- how to use various approaches (the time, derived variable and GEE approach) to incorporate mortality problems,
- the applications in the *Ceriodaphnia dubia* test example.

4.3 Time-wise approach

The time-wise approach is a rather simple and straightforward approach that treats the time-wise data as if they are independent. It applies standard methods, such as t-tests, ANOVA, or regression analyses at each time slot. The advantage of this approach is its simplicity, but there are two drawbacks in this approach. The first one is the inability of the time–wise approach to handle the longitudinal aspect of the data, i.e. the time effect and the change in the treatment effect over time. The second problem concerns with the dependence of the analyses at different time slots, there is no well-defined way to combine these analyses. Hence, by assuming independence of the observations of the same subject at different time slots, the p-values of the tests would be inappropriate.

In the *Ceriodaphnia dubia* test example, one can apply regression analysis, since the main interests are in the regression coefficients (β) and the inhibition concentrations (IC_x). With the idea of mixture distribution and the ZIP model developed in Chapter Three, the data obtained from the *Ceriodaphnia dubia* test example can be fitted to the ZIP model on a daily basis. As discussed in Chapter Three, the ZIP model will also adjust for the extra zeroes that are caused by mortality. Comparing the estimates of the

daily regression coefficients and inhibition concentrations can give us only a rough idea of the longitudinal effect of the data without any statistical conclusion. The idea of the ZIP approach is based on the independent assumption of the repeated measures which is not very reliable in most longitudinal studies. Nevertheless, since the individual analyses can be carried out using rather standard methods, this approach is simple and easy to use.

4.4 Derived variable approach

The derived variable approach, which is like the time-wise approach, attempts to simplify the data and applies standard analysis method. There are two steps involve in this approach. It starts with summarizing the repeated values into one number and then analyzing the summary variable as a function of the covariates. This method is sometimes referred as the two-stage method and can be dated back to Wishart (1938). It works well when the covariates remain constant over time.

In *Ceriodaphnia dubia* reproduction toxicity tests, estimation and inferences are traditionally done based on the mean number of the offspring produced per adult. In this case, the derived variable can simply be the mean (average) number of the eggs produced by each treatment within a seven-day period. The regression analysis can then be carried out in the second stage to investigate the effluent toxicant effect. However, when mortality exists, the calculation of mean is complicated. In order to adjust for the mortality effect, Hamilton (1986) discusses two different ways to define the mean when mortality occurs. They are the mean overall (MOA) and the mean ignoring mortality (MIM). The reasoning behind Hamilton's mean ignoring mortality is based on the mixture model idea which separates mortality effect from fecundity. This is similar to the idea behind the zero inflated Poisson model discussed in Chapter Three.

4.4.1 The Mean Overall (MOA) and Mean Ignoring Mortality (MIM)

A sample data is shown in Table 4.1 which is obtained from a 7 days *Ceriodaphnia dubia* test. Table 4.1 also demonstrates and compares how the estimates of MOA and MIM can be found using simple calculation. The mean overall is calculated by averaging the number of offsprings produced by all the animals that take part in the test. If an animal dies before reproducing the next generation, then zero reproduction is recorded for that animal. If an animal dies after reproducing the next generation, then the total number of offspring produced before death is used. This MOA is just our usual mean if mortality does not exist in the test. For MIM, only the surviving animals are used in mean. Since the animals do not mature and reproduce on the first three days, only four days of data are recorded. Table 4.1 shows only two treatments (A and B), and each treatment is replicated 10 times. The mean reproduction for each replication is calculated on a daily basis. The daily means of each of the four days are then summed together to obtain MOA and MIM.

Table 4.1. Sample data of *Ceriodaphnia* test

Treatment		Replication number										Mean	
		1	2	3	4	5	6	7	8	9	10	Per 10 animals	Per surviving animals
A	Day4	5	0	5	5	5	4	4	0	2	3	3.30	3.30
	Day5	6	7	6	7	6	0	8	6	5	0	5.10	5.10
	Day6	0	10	0	0	0	4	0	10	7	2	3.30	3.30
	Day7	12	0	10	13	10	5	11	0	0	4	6.50	6.50
	Total	23	19	21	25	21	13	23	20	15	9	18.20 (MOA)	18.20 (MIM)
B	Day4	0	0	0	0	0	0	2	0	0	0	0.20	0.20
	Day5	0	5	0	0	0	0	0	0	0	0	0.50	0.50
	Day6	0	0	0	0	0	0	0	0	0	0	0.00	0.00
	Day7	0	0	0	5	-	9	0	0	0	0	1.40	1.56
	Total	0	5	0	5	0	9	2	0	0	0	2.10 (MOA)	2.26 (MIM)

A negative sign "-" indicates the death of the animal.

Notice that MOA and MIM are the same under treatment A because mortality does not exist. On the other hand, when mortality occurs under treatment B, the two means are different. In order to illustrate the mixture concept, a MIM hypothetical model is examined in the following section.

Mathematical derivation of MOA and MIM

Let the variable X_i be the number of offsprings reproduced by a randomly chosen animal at time i (i = 1,2, …, t) and Z be the survival indicator variable, where

$Z_i = 0$, if the animal dies at time i;
$Z_i = 1$, if the animal survives until time i.

At time i, the observed number of offspring can be denoted as Y_i and $Y_i = X_i * Z_i$. If the animal dies at time i, then $X_i, X_{i+1}, \ldots, X_t$ will be unobservable. Hence, the total observed number of offspring produced by the animal is denoted as:

$$Y = \sum_{i=1}^{t} Y_i = \sum_{i=1}^{t} X_i Z_i$$

and the total number of offspring produced by the animal (assuming the animal survives till time t) is denoted as:

$$X = \sum_{i=1}^{t} X_i .$$

MIM and MOA are the average of X and the average of Y over all animals respectively. With the assumption that X_i and Z_I are independent, the expected value of Y_I, $E(Y_i) = E(X_i)E(Z_i)$. This is exactly the same mixture idea as for the ZIP where:

- $E(X_I)$ is the $\hat{I}$ modeled by a Poisson regression, and
- $E(Z_I)$ is the $(1-\hat{p})$ modeled by a logistic regression.

Furthermore, the expected value of X_I can be written as $E(X_i) = \frac{E(Y_i)}{E(Z_i)}$, and

MIM can be calculated as $\sum_{i=1}^{t} E(X_i) = \sum_{i=1}^{t} \frac{E(Y_i)}{E(Z_i)}$.

By using MIM in the second stage of the analysis, one is able to investigate the effluent toxicant effect with the adjustment of mortality. However, in order to calculate MIM, mortality information is required. In addition to the requirement of the mortality information, MIM also has the drawback of not addressing the longitudinal feature (time effect) of the data. But this may still be a sensible method to use if the time effect is not the interest of the experiment.

4.5 Generalized estimating equations approach

Zeger and Liang (1986) propose using generalized estimating equations (GEE) to analyze longitudinal data. This method combines the estimating function theory (Godambe, 1960 and 1976; Godambe and Heyde, 1987; Godambe and Thompson, 1989; Godambe and Kale, 1991; Schabenberger and Gregoire, 1995) with an extension of the quasi-likelihood principle (Wedderburn, 1974; McCullagh, 1983; McCullagh and Nelder, 1989; Nelder and Lee, 1992). By allowing a block–diagonal covariance matrix of the response, one is able to estimate both the regression coefficients and the covariance structure at the same time. Diggle et al. (1994) discuss how GEE applies to three different modeling approaches. They are the marginal model, the random effect model and the transition model. The complete details of the generalized estimating equation model and its theory can be found in Zeger and Liang (1986), Liang and Zeger (1986), Zeger et al.

(1988), and Zeger and Liang (1992). In the example of toxicity testing, the focus is on the population-average, and we will concentrate on the marginal model because it can be used to address questions such as:

- How does the toxicant affect the reproduction of the water flea?
- How does the fecundity of the *Ceriodaphnia* change over time?

The marginal GEE model can be viewed as a multivariate extension of the generalized linear model (GLM) with correlated observations among independent subjects. For example, suppose t_i observations are obtained from n subjects and let the response be Y_{ij} (the j measurement of the i subject) with i =1,…,n, j=1,…,t_i, and covariate x_{ij}. By using the GLM notation, we have $E(Y_{ij}) = \mu_{ij} = \beta' x_{ij}$, where β is the regression coefficient, $Var(Y_{ij}) = \varphi V(\mu_{ij})$, and the link function $g(\mu_{ij}) = \eta_{ij}$ which link the linear predictor ($\beta' x_{ij}$) to the mean. The covariance structure between the repeated measurement is modeled as $Cov(Y_{ij'}, Y_{ij}) = c(\mu_{ij'}, \mu_{ij}; \alpha)$ where $1 \le j' \le j \le t_i$, for some α. Under mild regularity conditions, Liang and Zeger (1986) prove that the GEE estimator of β is given by solving the following estimating equation:

$$U = \sum_{i=1}^{n} \mathbf{D}_i' \mathrm{Var}(Y_i)^{-1} (Y_i - \mu) = 0;$$

where $\mathbf{D}_i = \frac{\partial \mu}{\partial \beta}$, $\mu = (\mu_1, \ldots, \mu_{n_i})'$, and $\mathrm{Var}(Y_i) = A_i^{1/2} R_i(\alpha) A_i^{1/2}$,

where A_i is a $n_i \times n_i$ diagonal matrix with $a_{ii} = \mathrm{Var}(Y_{ij})$, $j = 1, \ldots, n_i$, and $R_i(\alpha)$ a $n_i \times n_i$ "working" correlation matrix. The estimates of the regression coefficients and the correlation can then be obtained iteratively. The term "working" implies that the estimate of the correlation is updated at each iteration. Both $\hat{\beta}$ and the estimate of $\mathrm{Var}(\hat{\beta})$ are

consistent even if $R_i(\alpha)$ is not correctly specified (Zeger and Liang, 1986). The GEE is actually a multivariate version of quasi-likelihood score equation (Wedderburn, 1974):

$$\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta_j} \mathrm{var}(y_i)^{-1}(y_i - \mu_i) \text{ where } j = 1,\ldots,p.$$

With this GEE setup, different correlation structures ($R_i(\alpha)$) can be specified to reflect different time dependence within each subject. Some common correlation structures are shown in Table 4.2. In marginal models, the subjects are assumed to be independent, the main interest is in the population-average information, and hence the focus of the model is on the regression coefficients. The regular regression models have a similar focus on the regression coefficients. Since the method of GEE is still in its early stage, the goodness of fit criteria is not yet well developed (Schabenberger, 1995). In the following section, a GEE based approach using the mixture idea is presented to incorporate the extra zeroes due to mortality, and the focus of the discussion will be on the inference of the regression coefficients.

Table 4.2. Common Correlation Structures
(Assuming $t_i = 4$)

Variance-Covariance structure	$R_i(\alpha)$ (symmetric)
Independent	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ & 1 & 0 & 0 \\ & & 1 & 0 \\ & & & 1 \end{bmatrix}$
Auto-regressive	$\begin{bmatrix} 1 & a & a^2 & a^3 \\ & 1 & a & a^2 \\ & & 1 & a \\ & & & 1 \end{bmatrix}$
Compound Symmetric (Exchangeable)	$\begin{bmatrix} 1 & a & a & a \\ & 1 & a & a \\ & & 1 & a \\ & & & 1 \end{bmatrix}$
Unstructured	$\begin{bmatrix} 1 & a_1 & a_2 & a_3 \\ & 1 & a_4 & a_5 \\ & & 1 & a_6 \\ & & & 1 \end{bmatrix}$

The GEE model discussed above is able to model longitudinal data with a pre-assumed correlation structure. Using the zero inflated Poisson idea in Chapter Three, a GEE based approach is developed to handle longitudinal toxicity data. This approach further extends the zero inflated Poisson model in a multivariate sense and makes use of the GEE to estimate the coefficients. The model can be interpreted as a regression model,

and the estimates can be obtained by using the iterated re-weighted least squares procedure. Furthermore, many standard programs have been developed to fit the GEE model by simply specifying the distribution and the link function. For example, SAS, S-plus, Sudaan, and Stata all have standard procedures to handle GEE.

Based on the distribution of the ZIP model, and the assumption of the independence of the subjects over time, a likelihood function similar to (3.2) can be obtained. Even though the assumption of independence is not realistic in this case, it is made to simplify the following development. Here, the matrix **G** contains the covariates of the reproduction part of the GEE model, and the matrix **B** contains the covariates of the mortality part of the GEE model. The data is measured from n subjects with t repeated measures on each subject. There are a total of $n \times t$ observations. By assuming all the observations are independent (both over time and subjects, i.e. $n \times t$ independent observations), the likelihood functions can be written as follows:

$$L(\gamma, \beta; y, z) = \sum_{i=1}^{n\times t}(z_i \mathbf{G}_i\gamma - \log(1+e^{\mathbf{G}i\gamma})) + \sum_{i=1}^{n\times t}(1-z_i)(y_i\mathbf{B}_i\beta - e^{\mathbf{B}_i\beta}) - \sum_{i=1}^{n\times t}(1-z_i)\log(y_i!) \tag{4.1}$$

where $\mathbf{G}_i$ and $\mathbf{B}_i$ are the i^{th} row of the covariate matrix **G** and **B**, $Z_i = 0$ if Y_i is from a surviving animal and $Z_i = 1$ if Y_i is from a dead animal. Again the two components which include β and γ in the likelihood can be re-written as:

$$L(\gamma; y, z) = \sum_{i=1}^{n\times t}(z_i \mathbf{G}_i\gamma - \log(1+e^{\mathbf{G}_t\gamma})) \tag{4.2}$$

$$L(\beta; y, z) = \sum_{i=1}^{n\times t} (1 - z_i)(y_i \mathbf{B}_i \beta - e^{\mathbf{B}_i \beta}) \qquad (4.3)$$

The estimation of β and γ would again involve an iterative E-M type procedure by alternating between $Z_{i,}$ β and γ. The $(k+1)^{th}$ iteration of the procedure is as follows:

1. Estimate Z_I

As in Chapter 3, given the most recent estimates of β and γ, $\beta^{(k)}$ and $\gamma^{(k)}$, Z_i can be found by using the Bayes rule .

Here $Z_i^{(k)} = Prob(\text{mortality} \mid y_i, \gamma^{(k)}, \beta^{(k)})$

$$= \frac{Prob(y_i \mid mortality)Prob(mortality)}{Prob(y_i \mid mortality)Prob(mortality) + Prob(y_i \mid survival)Prob(survival)}$$

$$= (1 + e^{-G_i\gamma^{(k)} - \exp(B_i\beta^{(k)})})^{-1} \qquad \text{if } y_i = 0$$

$$= 0 \qquad \text{if } y_i > 0 \qquad (4.4)$$

2. Estimate γ

Use the function (4.2) to obtain the estimate of $\gamma^{(k+1)}$.

As in Chapter 3, since $Z_i^{(k)} = 0$ whenever $y_i > 0$, (4.2) can be expressed as:

$$L(\gamma; y, z) = \sum_{y_i=0} Z_i G_i \gamma - \sum_{y_i=0} Z_i \log(1 + e^{G_i\gamma}) - \sum_{i=1}^{n\times t} (1 - Z_i) \log(1 + e^{G_i\gamma}) \qquad (4.5)$$

Now suppose there are n_0 animals who recorded zero at least on one out of the four days. Let their responses be $y_l = (y_{l_1 1}, \ldots, y_{l_{n_0\times 4}})$, then we define the following:

$$y_*^{'} = (y_1,\ldots,y_{n\times 4}, y_{l_1 1},\ldots,y_{l_{n_0\times 4}}),$$

$$\mathbf{G}_*^{'} = (\mathbf{G}_1^{'},\ldots,\mathbf{G}_{n\times 4}^{'},\mathbf{G}_{l_1 1}^{'},\ldots,\mathbf{G}_{l_{n_0\times 4}}^{'}),$$

$$\mathbf{P}_*^{'} = (p_{11},\ldots,p_{n\times 4}, p_{l_1},\ldots,p_{l_{n_0\times 4}}).$$

Then a diagonal matrix $\mathbf{W}^{(k)}$ with diagonal element is defined as:

$$w^{(k)} = (1-Z_1^{(k)},\ldots,1-Z_{n\times 4}^{(k)}, V_{l_1}^{(k)},\ldots,V_{l_{n_0\times 4}}^{(k)})$$

where $V_{l_i}^{(k)} = Z_{l_i}^{(k)}$ when $y_{l_i} = 0$, and $V_{l_i}^{(k)} = 0$ when $y_{l_i} > 0$.

$\mathbf{W}^{(k)}$ has a similar weighting scheme as that in Chapter 3, except:

- when there is at least a zero observation from an animal, then all the responses from that animal will be added to the y_*', and
- the nonzero observations of these animals are forced to become zero with the weights $V = 0$.

Now the function (4.2) becomes:

$$L(\gamma; y, Z^{(k)}) = \sum_{i=1}^{n\times t+n_0\times t} y_{*i} w_i^{(k)} G_{*i}\gamma - \sum_{i=1}^{n\times t+n_0\times t} w_i^{(k)} \log(1+e^{G_{*i}\gamma}) \qquad (4.6)$$

Following the discussion in Chapter three, consider the score function of weighted logistic regression. The function (4.6) will again give rise to the score function for (4.2) which can also be written as:

$$\mathbf{G}_*^{'}\mathbf{W}^{(k)}(\mathbf{y}_* - \mathbf{P}_*) = 0$$

as in Chapter three, and the negative information matrix is:

$$\mathbf{G}_*^{'}\mathbf{W}^{(k)}\mathbf{Q}_*\mathbf{G}_*$$

where $\mathbf{Q}_*$ is the diagonal matrix with $\mathbf{P}_*$ (1-$\mathbf{P}_*$) on the diagonal.

The score function is the same as for the univariate case for the logistic GEE with $\mathbf{D}_i \rightarrow \frac{\partial G_{*i}}{\partial}Var(y_{*i}) = G_{*i}(Var(y_{*i}))$, E($Y_i$) = $\mu_i \rightarrow p_i$, and except with an extra weight called $\mathbf{W}^{(k)}$. The variances in the equations cancel each other. This can be extended to a multivariate score function using the underlying development in the GEE model with $Var(y_{*i})$ as a matrix. Hence, this leads to a weighted GEE equation $\sum \mathbf{D}_{*i}^{'}\mathbf{W}_i^{(k)}\mathrm{Var}(Y_{*i})^{-1}(Y_{*i} - \mu_{*i}) = 0$ with responses Y_{*i} and weights $\mathbf{W}_i^{(k)}$. The estimate of γ for the current iteration can be obtained.

3. Estimate β

Use the function L(β; y, $Z^{(k)}$) to obtain the estimate of $\beta^{(k+1)}$. In Chapter three, this is achieved by using a weighted log-linear Poisson regression with weights 1- $\mathbf{Z}^{(k)}$ (McCullagh and Nelder, 1989). When the independent assumption is relaxed, this can be extended to a weighted GEE model using the Poisson distribution with weights (1-$\mathbf{Z}^{(k)}$). Again, by pre-assuming a correlation structure ($R_i(\alpha)$) among the repeated measures, the estimates of the regression coefficients can be obtained.

The estimates of β and γ can be obtained by this E-M type procedure. The weighted GEE model program is available through standard software which applies an

iterative re-weighted least squares procedure together with a pre-assumed correlation structure ($R_i(\alpha)$) among the repeated measures. Some possible correlation structures are shown in Table 4.2. In the following example, three different correlation structure will be used. They are the independent, auto-regressive, and compound symmetric correlation structure.

4.6 Examples for GEE Approach

In the *Ceriodaphnia dubia* test example, there are 660 animals. Each animal is observed for a seven-day period. Since the animals do not reproduce in the first three days, only four days of data are used in the analysis (in equation 4.6, $n = 660$ and $t_i = t = 4$). The approach is based on the zero inflated Poisson density. There are two parts in the model, one is modeling reproduction (Poisson data), and the other is modeling mortality (binary data). It can be viewed as a reproduction model with an adjustment due to mortality. In order to examine time and toxicant effects, both are used as the covariates in the model, however, only one toxicant is used to keep the discussion simple. An S-plus computer program is written to fit the GEE-ZIP model (see Appendix C). The model has six regression coefficients and the results are shown in Table 4.3 to Table 4.5.

Table 4.3. GEE-ZIP model for the ceriodaphnia data; S.E. in ()
(Independent correlation structure)

	Chromium	Copper	Mercury	Zinc
β_0 (int.)	1.093 (0.0606)	1.084 (0.5695)	1.115 (0.0543)	1.112 (0.0625)
β_1 (time)	0.281 (0.0128)	0.282 (0.0128)	0.282 (0.127)	0.281 (0.0129)
β_2 (toxic)	-0.050 (0.0270)	-0.019 (0.0083)	-0.043 (0.0087)	-0.004 (0.0016)
γ_0 (int.)	0.691 (0.0984)	0.699 (0.0921)	0.721 (0.0932)	1.045 (0.1035)
γ_1 (time)	-0.455 (0.0320)	-0.481 (0.0339)	-0.476 (0.0331)	-0.438 (0.0312)
γ_2 (toxic)	0.338 (0.0417)	0.128 (0.0125)	0.157 (0.0132)	0.003 (0.0028)
Sum squared Residual	29317.57	27354.99	25926.82	30788.25

Table 4.4. GEE-ZIP model for the ceriodaphnia data; S.E. in ()
(Compound Symmetric correlation structure)

	Chromium	Copper	Mercury	Zinc
β_0 (int.)	1.086 (0.0487)	1.078 (0.0467)	1.117 (0.0453)	1.096 (0.0499)
β_1 (time)	0.293 (0.0113)	0.292 (0.0113)	0.291 (0.0113)	0.293 (0.0114)
β_2 (toxic)	-0.054 (0.0198)	-0.017 (0.0058)	-0.042 (0.0064)	-0.003 (0.0012)
γ_0 (int.)	0.555 (0.0958)	0.584 (0.0897)	0.597 (0.0906)	0.821 (0.1001)
γ_1 (time)	-0.366 (0.0294)	-0.404 (0.0311)	-0.392 (0.0303)	-0.347 (0.0288)
γ_2 (toxic)	0.275 (0.0394)	0.109 (0.0107)	0.131 (0.0124)	0.002 (0.0027)
Sum squared Residual	29374.26	27441.42	26013.00	30844.54

Table 4.5. GEE-ZIP model for the ceriodaphnia data; S.E. in ()
(Auto-regressive correlation structure)

	Chromium	Copper	Mercury	Zinc
β_0 (int.)	1.114 (0.0546)	1.109 (0.0519)	1.135 (0.0500)	1.125 (0.0558)
β_1 (time)	0.288 (0.0131)	0.289 (0.0130)	0.288 (0.0128)	0.289 (0.0132)
β_2 (toxic)	-0.057 (0.0214)	-0.022 (0.0067)	-0.043 (0.0070)	-0.004 (0.0013)
γ_0 (int.)	0.649 (0.0995)	0.674 (0.0928)	0.694 (0.0939)	0.961 (0.1047)
γ_1 (time)	-0.434 (0.0319)	-0.469 (0.0337)	-0.462 (0.0329)	-0.408 (0.0311)
γ_2 (toxic)	0.322 (0.0410)	0.125 (0.0122)	0.153 (0.0130)	0.003 (0.0028)
Sum squared Residual	29319.52	27305.96	25895.80	30796.90

Different correlation structures can be embedded in the two parts to model the two different mechanisms. However, in this example, the same correlation structure is used in both the reproduction and mortality part. The estimates of the correlation structures for different models are shown in Table 4.6. Based on the sum of squared residuals, the mercury and copper model are found to be the better model among the four single toxicant models. The day effect is consistently significant in all the models, and also all four models have very similar time effect in both parts of the model (0.282 to 0.281 in reproduction, –0.438 to –0.481 in mortality part). The zinc model has the lowest toxicant effects among the four models while the chromium model has the highest. In addition, the models with auto-regressive correlation structure have the lowest sum of squared residuals (25895 for Mercury and 27305 for Copper). The auto-regressive correlation structure implies that the further apart the data points are, the lesser the correlation is between them. In the mortality parts of the copper and mercury model, the off-diagonal terms are very close to zero. It implies the use of the assumption of independent correlation structure in the model would be acceptable.

The advantage of GEE-ZIP approach is the flexibility of allowing different correlation structures for the models. Even though the ZIP model in Chapter Two is in fact a univariate case of the GEE-ZIP model, and time can be treated as a covariate in the ZIP model, the estimate of the ZIP model would still be inappropriate or biased. This is because the ZIP model does not consider the possible correlation between the repeated measures. Unlike the derived variable approach, the GEE-ZIP model is able to handle covariates that change over time. By investigating the correlation structure together with the time effects, a better understanding of the underlying mechanism can be achieved.

Table 4.6. The correlation estimates for different models (symmetric)

	Chromium	Copper	Mercury	Zinc
Compound Symmetric				
Reproduction	$\begin{bmatrix} 1 & 0.60 & 0.60 & 0.60 \\ & 1 & 0.60 & 0.60 \\ & & 1 & 0.60 \\ & & & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.59 & 0.59 & 0.59 \\ & 1 & 0.59 & 0.59 \\ & & 1 & 0.59 \\ & & & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.58 & 0.58 & 0.58 \\ & 1 & 0.58 & 0.58 \\ & & 1 & 0.58 \\ & & & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.60 & 0.60 & 0.60 \\ & 1 & 0.60 & 0.60 \\ & & 1 & 0.60 \\ & & & 1 \end{bmatrix}$
Mortality	$\begin{bmatrix} 1 & 0.20 & 0.20 & 0.20 \\ & 1 & 0.20 & 0.20 \\ & & 1 & 0.20 \\ & & & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.15 & 0.15 & 0.15 \\ & 1 & 0.15 & 0.15 \\ & & 1 & 0.15 \\ & & & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.17 & 0.17 & 0.17 \\ & 1 & 0.17 & 0.17 \\ & & 1 & 0.17 \\ & & & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.22 & 0.22 & 0.22 \\ & 1 & 0.22 & 0.22 \\ & & 1 & 0.22 \\ & & & 1 \end{bmatrix}$
Auto-regressive				
Reproduction	$\begin{bmatrix} 1 & 0.61 & 0.37 & 0.23 \\ & 1 & 0.61 & 0.37 \\ & & 1 & 0.61 \\ & & & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.60 & 0.35 & 0.21 \\ & 1 & 0.60 & 0.35 \\ & & 1 & 0.60 \\ & & & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.58 & 0.34 & 0.19 \\ & 1 & 0.58 & 0.34 \\ & & 1 & 0.58 \\ & & & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.62 & 0.38 & 0.24 \\ & 1 & 0.62 & 0.38 \\ & & 1 & 0.62 \\ & & & 1 \end{bmatrix}$
Mortality	$\begin{bmatrix} 1 & 0.08 & 0.01 & 0.00 \\ & 1 & 0.08 & 0.01 \\ & & 1 & 0.08 \\ & & & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.04 & 0.00 & 0.00 \\ & 1 & 0.04 & 0.00 \\ & & 1 & 0.04 \\ & & & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.05 & 0.00 & 0.00 \\ & 1 & 0.05 & 0.00 \\ & & 1 & 0.05 \\ & & & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.11 & 0.01 & 0 \\ & 1 & 0.11 & 0.01 \\ & & 1 & 0.11 \\ & & & 1 \end{bmatrix}$

4.6.1 Inhibition Concentration

In order to interpret the results for the GEE approaches, an overall prediction and the inhibition concentration (IC) can be defined as those in Chapter three. By using the same definition in Chapter three, the overall mean can be expressed as $\hat{\mu}(1-\hat{p})$, where $\hat{\mu}$ is the predicted mean of the Poisson GEE and $\hat{p}$ is the predicted mean of the logistic GEE. This can be viewed as the reproduction weighted by mortality probability. Since there are two independent variables (time and toxicant concentration), the mean response is a surface and the IC levels are a set of contour lines. Since the auto-regressive correlation models have the lowest sum of squared residuals, their contours of the IC levels are plotted in Figure 4.1. The IC levels of the chromium model are the lowest, which means that chromium has a relatively stronger toxic effect than the other effluents. However, the slopes of the IC lines are relatively flat when compared with other models. This means the toxic effect of chromium does not change as fast as the other effluents.

Figure 4.1. IC levels of the four single toxicant models with time effect
(PDF, 4K, Fig4-1.pdf)

4.7. MIM revisited

With the GEE-ZIP model established, we now can compare the GEE-ZIP estimate with the MIM estimate. The mixture idea suggests the observation Y is equal to the product X*Z, where X presents fecundity and Z indicates mortality. The GEE-ZIP model is then based on the two parts model:

$$\log(\underline{\lambda}) = \mathbf{B}\underline{\beta} \text{ and } \text{logit}(\underline{p}) = \log(\underline{p}/(1-\underline{p})) = \mathbf{G}\underline{\gamma},$$

where the Poisson part models fecundity (X), and the logistic part models mortality (Z). The prediction of the GEE-ZIP can be written as $\hat{\mu} = \hat{\lambda}(1-\hat{p})$, where $\hat{\mu}$ is just the expected value of the observation Y, $\hat{\lambda}$ is the expected value of X, and $(1-\hat{p})$ is the expected value of Z. This is exactly the idea in the MIM estimation

$$\sum_{i=1}^{t} E(X_i) = \sum_{i=1}^{t} \frac{E(Y_i)}{E(Z_i)},$$

where the fecundity is estimated with adjustment for the mortality effect. In the GEE-ZIP, the fecundity (X) is estimated (or predicted) by the Poisson part of the model and the adjustment for mortality is estimated by the logistic part (model for Z). However, the MIM estimation does not involve any covariate. The MIM estimation of the mean is then equivalent to the prediction of the intercept only GEE-ZIP model. An estimate for the variance of the X can then be developed based on the GEE-ZIP estimate. The quantity $\sum_{i=1}^{t} E(X_i) = \sum_{i=1}^{t} \hat{\lambda}_i$ can be expressed as a linear combination $c'\underline{\hat{\lambda}}$ where c is a $t\times 1$ vector of 1's, and $\underline{\hat{\lambda}}$ is a $t\times 1$ vector of $\hat{\lambda}_i\,(i = 1,\ldots,t)$. The variance of $\sum_{i=1}^{t} E(X_i)$ becomes the variance of $c'\underline{\hat{\lambda}}$, which is equal to $c'\hat{\Sigma}c$ where $\hat{\Sigma}$ is the estimated variance-covariance matrix of $\hat{\lambda}_i$. The value of $\hat{\Sigma}$ can be obtained from the estimated variance-covariance matrix of the GEE-ZIP model. Since the vector c is just a column of ones, $c'\hat{\Sigma}c$ is just the sum of all the terms in $\hat{\Sigma}$. For example, if the subject is measured four times (t=4), then c' is a 4×1 vector, and $\hat{\Sigma}$ is a 4×4 matrix with elements σ_{ij}^2 where $i, j = 1,2,3,4$. Then $c'\hat{\Sigma}c$ is just $\sum_{i=1}^{t}\sum_{j=1}^{t} c_i \sigma_{ij}^2 c_j$. This estimate of the variance of mean can then be compared with

Hamilton's (1986) bootstrap method (Table. 4.7) using Hamilton's site three data (Hamilton, 1986).

Table 4.7. MIM and GEE-ZIP mean estimates and their variance

	MIM	GEE (Auto-regressive)	GEE (Exchangeable)
mean	13.5	16.88	15.58
variance of mean	4.36	2.036	1.645

Two correlation structures are used in the GEE-ZIP model, and they are compared with the MIM method. The estimated mean reproduction for the exchangeable correlation model has the lowest variance among the three.

Hamilton uses a bootstrap method to obtain the variance without explicitly assuming any correlation structure in the data. The estimate obtained by the Hamilton method has a relatively high variance compared to that obtained by the GEE-ZIP model. The variance of the GEE-ZIP is obtained by using the Poisson part of the model and its estimated covariance matrix as discussed in the previous section. Since the calculation includes the covariance terms, the results of the GEE-ZIP model depend on the correlation assumption.

4.8. Conclusion

There are two major problems with exploring longitudinal toxicity data. One is the possible correlation structure among the repeated measures within each subject. The other is the problem of extra zeroes due to mortality.

In this chapter, three different approaches that handle longitudinal toxicity data are discussed. The first two approaches: the time-wise approach and the derived variables approach, are not able to simultaneously resolve the two problems that are listed above in one model. But they analyze the data using very simple straightforward methods. The third approach is based on the generalized estimating equations (GEE) technique and the zero inflated Poisson (ZIP) model which is the main focus of this chapter. This approach is able to model fecundity with the adjustment of extra zeroes due to mortality, and at the same time, estimate the correlation structure within each subject.

Since the GEE modeling technique is still in its early stage, there is no well-defined goodness-of-fit criterion. The example in this Chapter uses the sum of squared residuals each model as a goodness of fit criterion to decide which one is the best model among the four single toxicant GEE-ZIP models in Table 4.3 – Table 4.5. Copper model and mercury model are found to be the two better models among the four single toxicant models. The models also can be used to examine the time effect and look at the changes of the toxicant effect over time. The same definition of inhibition concentration (IC) can also be obtained as that in Chapter Three.

The GEE model discussed in this Chapter is a marginal model with the assumption that the subjects are independent of each another. The discussion is limited to single toxicant models. A straightforward extension to the GEE model is to include the time toxicant interaction in the model. In addition, by relaxing the assumption of independent subjects, more complicated models can be developed. For example, it may be possible to apply other types of GEE models like the random effect model or the transitional model to longitudinal toxicity data. However, more research is needed for models that are able to incorporate extra zeroes.