CHAPTER 2
CURE MODELS: A PARTIAL REVIEW WITH AN APPLICATION TO RECURRENT EVENT COUNT DATA

2.1 Introduction

In a community based study on human epidemiology, the study population is generally heterogeneous. It eventually divides into two groups, one group consisting of those patients who respond favourably to the treatment and subsequently become immune or insusceptible to the disease. The other group consists of those patients who do not respond to the treatment. The former group is said to be cured. This cured proportion became an important and a very useful measure in obtaining the trends in the survival of patients, especially those suffering from cancer. As mentioned in the previous chapter, the widely used model in survival analysis is the Cox’s (1972) proportional hazards (PH) model which is based on the assumption that every individual in the population under study is susceptible to the adverse event of interest such as a disease. This assumption cannot be used in modern health-related studies. The reason for this is that, after a sufficient follow-up, it has been observed that a large group of patients are being cured of cancer. As a result, a cure model was developed for estimating the cured proportion among the patients suffering from cancer.

Cure models are survival models basically developed to estimate the proportion of patients cured of cancer. These models also estimate the probability of survival of the uncured patients up to a given point of time. A cure model was first developed by Boag (1949) to estimate the proportion of patients being cured among those who were receiving treatment for cancer of mouth and throat, cervix, uterus, and breast. This model was called the mixture cure model since it could estimate the proportion of patients being cured and also, the survival function of the uncured. Boag (1949) modeled the survival function of the uncured group as a product of the survival functions of a log-normal distribution and some background distribution for the normal population. The model developed by Boag (1949) was later modified by Berkson and Gage in 1952 and was called the standard cure model. Yakovlev, et al. (1993) observed that the mixture cure model had a few limitations, especially when a...
set of covariates were present and thus, developed an alternative to the standard cure model called the bounded cumulative hazard (BCH) model.

In this chapter, a brief review of literature on cure models is given. Section 2.2 discusses the mixture cure model, the BCH model and also a model based on Box-Cox (1964) transformation. An extension of a univariate model to a bivariate model and also to a multivariate set up is possible through a function known as copulas. Therefore section 2.3 provides a brief introduction to copulas and discusses some of its basic properties. A brief review on multivariate cure models is given in section 2.4. Section 2.5 proposes a new cure model for recurrent event count data. The proposed model takes into account the number of times a disease recurs in an individual, given that he/she is cured. The chapter concludes with section 2.6.

### 2.2 Univariate Cure Models

#### 2.2.1 Mixture Cure Model

As mentioned in the previous section, it was Boag in 1949 who proposed a cure model for estimating the proportion of patients being cured of cancer. Boag's model was later modified by Berkson and Gage in 1952 and was called the mixture cure model. The model works on the assumption that there exists a group of patients who will be cured of a disease under study. The derivation of the model is as follows.

Suppose a group of patients enter a clinical trial. Let $p$ be the proportion of patients being cured on treatment and $1 - p$ be the proportion uncured. The survivor function $S(t)$ for the entire population of patients entering the clinical trial is given by the model

$$S(t) = p + (1 - p)S_u(t)$$

(2.2.1)

where $S_u(t)$ is the survivor function of the uncured patients with corresponding probability density function $f_u(t)$. The model (2.2.1) is known as the mixture cure model or the standard cure model. The probability density function corresponding to (2.2.1) is $f(t) = (1 - p)f_u(t)$

(2.2.2)

and the hazard function is

$$h(t) = \frac{(1 - p)f_u(t)}{p + (1 - p)f_u(t)}$$

(2.2.3)
In long term follow up studies, the patients may be lost to follow up for several reasons such as migration and death. In such cases, the data is said to be censored. Taking into consideration, the assumption of the existence of a cured proportion and censoring, the likelihood function based on a random sample of \( n \) individuals entering a study is given by

\[
L = \prod_{i=1}^{n} \left\{ \left( 1 - p \right) f_u(t_i) \right\}^{\delta_i} \left\{ p + \left( 1 - p \right) S_u(t_i) \right\}^{1 - \delta_i}
\]  
(2.2.4)

where \( t_i, i = 1, 2, \ldots, n \), is the observed survival time for the \( i^{th} \) patient and \( \delta_i \) is a censoring indicator defined such that

\[
\delta_i = \begin{cases} 
1, & \text{if } t_i \text{ is censored} \\
0, & \text{otherwise}
\end{cases}
\]  
(2.2.5)

The mixture model can be parametric or non-parametric depending on whether \( f_u(t) \) is specified or not. The commonly used standard parametric probability distributions for \( f_u(t) \) are the exponential and the Weibull distributions. The maximum likelihood estimates (MLE) of the parameters of the model can be obtained by maximizing the likelihood equation (2.2.4). Any numerical methods of iterative solutions such as the Newton-Raphson’s method may be used if the maximum likelihood equations do not give a solution in a closed form.

Generally, in health related studies, the study variable is said to be influenced by a number of factors called covariates (Ghitany, et. al (1994), Kannan, et. al. (2010)). Suppose, \( Y_1, Y_2, \ldots, Y_n \) denote the study variable for the \( n \) individuals who are on follow up for a sufficiently long period. Let \( X_i = (x_{i0}, x_{i1}, x_{i2}, \ldots, x_{ik}) \) with \( x_{i0} = 1 \), denote the values of the baseline covariates associated with the \( i^{th} \) individual. Then the dependency of the probability \( p_i \) of the \( i^{th} \) individual being cured, on the covariates \( X_i \), can be modeled using a logistic function \( p_i = \frac{1}{1 + \exp(X_i \beta)} \). Here, \( \beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k) \) denotes the vector of regression coefficients. The maximum likelihood estimates (MLE) can be obtained by maximizing the likelihood equation (2.2.4). If the maximum likelihood equations do not give a solution in closed form, then any numerical method of iterative solutions may be used. Kannan, et al (2010)
have used the expectation-maximization (EM) algorithm for the estimation of parameters in their paper on generalized exponential cure model with covariates.

Suppose \( f_u(t) \) is not specified. Then we have a non-parametric mixture cure model. Mallor and Zhou (1992) and Sposto, Sather and Baker (1992) have suggested a non-parametric estimator for the cured proportion \( p \) as the maximum observed value of the Kaplan-Meier (KM) (1958) estimator given by

\[
\hat{S}(t) = \prod_{k} \left( \frac{1 - d_k}{r_k} \right)
\]  \hspace{1cm} (2.2.6)

This estimator has been obtained by considering the \( m \) distinct ordered failure times \( t_1 < t_2 < \ldots < t_m \) in a group of \( n \) patients. In (2.2.6), \( d_k \) represents the number of failures at time \( t_k \), and \( r_k \) is the number of patients at risk just before \( t_k \). The KM estimator is also known as the product-limit (PL) estimator. In the presence of covariates, Peng and Dear (2000) have suggested that the Cox proportional hazards (PH) model given by

\[
h_u(t|X) = h_0(t) \exp(X'\beta)
\]  \hspace{1cm} (2.2.7)

may be used, where \( h_u(t|X) \) is the hazard function for the uncured patients and \( h_0(t) \) is the baseline hazard which can be an arbitrary unspecified hazard function independent of \( X \). The estimation technique based on expectation-maximization (EM) algorithm has also been discussed by the authors.

2.2.2 Bounded Cumulative Hazard (BCH) Model

The second type of cure model is the bounded cumulative hazard (BCH) model developed by Yakovlev, et al (1993) This model was developed as an alternative to the standard cure model given by (2.2.1) It was developed by considering the patients suffering from cancer entering a clinical trial and was based on the assumption that after treatment, a patient is left with \( N \) cancer cells capable of metastasizing. The process in which the cancer cells move to various parts of the human body, grow rapidly and replace the normal tissue, is known as metastasis. Here, the variable \( N \) is assumed to follow a Poisson distribution with mean \( \theta \). If \( t_k, k = 1, 2, \ldots, N, \) denotes the time for the \( k \textsuperscript{th} \) metastatic tumor cell to produce detectable cancer, then the time to relapse of cancer in the patient is given by \( T = \min \{ t_1, t_2, \ldots, t_N \} \). The variable \( T \) is observable. Conditional on \( N \), it is assumed that \( t_k \)'s are independent and identically distributed (i.i.d.) with cumulative distribution function \( F_u(t) \), survival function \( S_u(t) \) and probability density function \( f_u(t) \), where the suffix \( u \) denotes the uncured group. The survival function \( S(t) \) for the entire population of patients entering the clinical trial is given by

\[
S(t) = P[\text{There is no detectable cancer by time } t] = P[N = 0] + P[t_1 > t, t_2 > t, \ldots, t_N > t \mid N \geq 1]
\]

\[
= \exp(-\theta) + \sum_{N=1}^{\infty} \left( S_u(t) \right)^N \frac{\exp(-\theta)\theta^N}{N!}
\]

\[
= \exp(-\theta F_u(t)) \tag{2.2.8}
\]

It has to be noted that the population survival function \( S(t) \) is improper. The cured proportion is given by \( p = \lim_{t \to \infty} S(t) = \exp(-\theta) \).

The density function corresponding to the entire population is

\[
f(t) = \theta f_u(t) \exp(-\theta F_u(t)) \tag{2.2.9}
\]

and the hazard function, on simplification, is

\[
h(t) = \frac{f_u(t)}{S(t)} = \theta f_u(t) \tag{2.2.10}
\]

The hazard function \( h(t) \) given by (2.2.10) is multiplicative in \( \theta \) and \( f_u(t) \) and thus has a proportional hazards (PH) structure which is a desirable property in survival
analysis whereas, the hazard function given by (2.2.3) for the mixture cure model does not possess the PH structure

The BCH model, which is also known as the promotion time cure model, can be written as a mixture model, i.e., equation (2.2.8) can be written as

\[ S(t) = \exp(-\theta) + \exp(-\theta)(\exp(S_u(t)) - 1) \]

\[ = \exp(-\theta) + (1 - \exp(-\theta)) \frac{\exp(-\theta) - \exp(-\theta)}{(1 - \exp(-\theta))} \]  \hspace{1cm} (2.2.11)

where the survival function for the uncured population is given by

\[ S_u(t) = P[T > t | N \geq 1] = \frac{\exp(-\theta) - \exp(-\theta)}{(1 - \exp(-\theta))} \]  \hspace{1cm} (2.2.12)

Thus every model defined by the BCH model (2.2.8) can be expressed as a mixture model with cured proportion \( p = \exp(-\theta) \). In the presence of covariates, the canonical link function \( \theta = \exp(X'\beta) \) can be used

The literature available on BCH model is only in the Bayesian context. The parametric and semi-parametric methods of estimation using a BCH model are discussed elaborately by Ibrahim, Chen and Sinha (2001). Yin and Ibrahim (2005) developed a general class of cure models through Box-Cox (1964) transformation on the population survival function. The model is given by

\[ \frac{S(\theta | Z, X_i)^a - 1}{a} = -\theta(a, Z_i)F(\theta | X_i), \quad a \in [0,1], \quad 0 \leq a \theta(a, Z_i) \leq 1 \]  \hspace{1cm} (2.2.13)

where \( X_i \) and \( Z_i \) are the covariate vectors corresponding to the \( i \)th individual, the parameter \( a \) is the transformation parameter, \( S(\theta | X_i) \) is the survival function of the population and \( F(\theta | X_i) \) is the cumulative distribution function of the failure time. A discrete uniform prior is taken for the parameter \( a \). The method of estimation has been discussed by the authors. They have observed that when \( a = 1 \), the model reduces to a mixture cure model and, when \( a = 0 \), it becomes a BCH model.

### 2.3 Copulas

The interest in modeling multivariate survival data is increasing rapidly. For example, in a study involving diabetic patients, the interest may be in studying the times to primary and secondary complications of diabetes. The two commonly used
approaches in modeling multivariate survival data are the random effects (frailty) approach and the marginal approach. In the random effects approach, independence of observations is assumed conditional on a scalar non-negative random variable known as frailty which multiplies the hazard and, when mixed over the distribution, produces dependence. Here, the estimation of the dependence structure is of primary concern. The marginal distributions are treated as nuisance functions. The marginal approach considers the marginal distributions to be modeled first and then imposes a dependence structure. The major concern here is the consistency of the estimators of the parameters and therefore, the association among dependent failure times is treated as nuisance. Modeling correlated multivariate survival data using the marginal approach can be done using the copulas.

Copula is a word taken from Latin. It means connecting or joining together. It is a function which connects the multivariate probability distribution to a univariate probability distribution. If \( T_1, T_2, \ldots, T_n \) are random variables with a joint distribution function \( F(t_1, t_2, \ldots, t_n) \) and marginal distribution functions \( F_1(t_1), F_2(t_2), \ldots, F_n(t_n) \), then we have

\[
F(t_1, t_2, \ldots, t_n) = C(F_1(t_1), F_2(t_2), \ldots, F_n(t_n)) \tag{2.3.1}
\]

where \( C \) denotes a copula function which generates an \( n \)-variate distribution function from an arbitrary set of \( n \) univariate distributions. The copula function is a multivariate cumulative distribution function defined on the \( n \)-dimensional unit space \([0,1]^n\) such that every marginal distribution is uniform on \([0,1]\). The following properties are satisfied by a copula function \( C \) when a bivariate set-up is considered:

\[
C(u, 0) = C(0, v) = 0, \\
C(u, 1) = u, \\
C(1, v) = v.
\]

The Frechet's (1951) bounds for all the copulas is given by \( M(x, y), W(x, y) \), where \( M(x, y) = \max(0, x + y - 1) \) represents perfect negative correlation between the random variables, and \( W(x, y) = \min(x, y) \) represents perfect positive correlation between the random variables. The extension of these properties to \( n \)-dimensional copulas and a brief review of the various properties, results and also the different families of copulas is found in Kolev, Anjos and Mendes (2006).
The family of copulas that has been used in multivariate survival analysis is the Archimedean class. Consider a bivariate cumulative distribution function $F(x, y)$ for two random variables $X$ and $Y$. The function $F(x, y)$ is said to belong to an Archimedean class of copulas if it can be written in the form

$$F(x, y) = \phi^{-1}[\phi(F_1(x)) + \phi(F_2(y))]$$

(2.3.2)

where $\phi$ is the generator function which satisfies the following properties:

- $\phi(0) = 0$,
- $\lim_{x \to 0} \phi(x) = \infty$,
- $\phi'(x) < 0$,
- $\phi''(x) > 0$.

The following three models belong to the Archimedean family of copulas:

   
   $$C_\theta(u, v) = \begin{cases} 
   (u^{1-\theta} + v^{1-\theta} - 1)^{1/(1-\theta)}, & \theta > 1 \\
   uv, & \theta = 1
   \end{cases}$$

2. Frank’s (1979) model.
   
   $$C_\kappa(u, v) = \exp[\log\left(1 - \frac{(1 - \kappa^u)(1 - \kappa^v)}{1 - \kappa} \right)/\log\kappa), \kappa > 1$$

3. Positive stable model: (Hougaard, 1986a)
   
   $$C_\omega(u, v) = \begin{cases} 
   \exp[\{(-\log u)^{1/\omega} + (-\log v)^{1/\omega}\}]^{\omega}, & 0 < \omega < 1 \\
   uv, & \omega = 1
   \end{cases}$$

In the above three models, the parameters $\theta, \kappa$ and $\omega$ are called the copula parameters. These parameters measure the degree of association between the random variables $U$ and $V$ taking values $u$ and $v$ respectively. The range of the above parameters account only for the positive association of the random variables.

The models developed by Clayton (1978) and Frank (1979) consider negative correlation when $\theta < 1$ and $\kappa > 1$ respectively, whereas the positive stable model considers positive correlation only. Some important work on copulas can be found in...
2.4 Multivariate Cure Models

The multivariate models are used when the interest is in jointly modeling several types of failure time random variables such as times to relapse of a disease and death, times to detectability of cancer at two or more organs, times to primary and secondary complications of a disease and familial association between various genetic diseases such as breast cancer, diabetes and heart diseases. The literature available on multivariate cure models in the classical framework are that of Chatterjee and Shih (2001) and Price and Manatunga (2001). Both these papers were based on standard cure models. In the Bayesian context, the work by Chen, Ibrahim and Sinha (2002) and Yin (2005) are noteworthy and are based on the BCH model.

The model proposed by Chatterjee and Shih (2001) is an extension of the univariate cure mixture model to a bivariate setting. The model was developed to analyze the correlated survival data when there exists a cured proportion in the study population. Correlated survival data are those which consider the familial association for diseases like breast cancer, diabetes and heart diseases. Suppose two members from each family are involved in the study. Let $Y_j$ and $Y_2$ be the two random variables taking values either 1 or 0 depending on whether the $j^{th}$ individual is susceptible to the disease or not, $j=1,2$. Let $T_j$ be the age at onset of the disease when the $j^{th}$ individual is susceptible to the disease. Then the marginal distributions of $Y_j$ and the failure time $T_j$ for the susceptible individuals are given by

$$\phi_j = P(Y_j = 1) \text{ and } S_j(t) = P(T_j \geq t \mid Y_j = 1)$$

respectively. A common marginal distribution for the two members of the family is assumed, i.e., $\phi_1 = \phi_2$ and $S_1(t) = S_2(t)$ for all $t$; $S_1(t)$ and $S_2(t)$ may be parametric or non-parametric. A dependence structure between the members of the pair is specified. The first type of association is between the susceptibility to the disease among the two individuals in the pair, that is, between $Y_1$ and $Y_2$. This association is given by the pair-wise odds ratio parameter $\gamma = \frac{P_{11} \times P_{00}}{P_{01} \times P_{10}}$, where $p_{00} = P(Y_1 = 1, Y_2 = 0), \ i, j = 0,1$.
The pair-wise odds ratio characterizes the dependence between the binary outcomes. The second type of association is between the failure times of the two susceptible members, i.e., between $T_1$ and $T_2$. This dependency structure between the failure times of the two susceptibles is specified using the copulas. Another assumption made in their approach was that the marginal distribution of the failure time of one susceptible is independent of the susceptibility status of another, $i.e., P(T_j \geq t_j | Y_j = 1, Y_i, i \neq j) = P(T_j \geq t_j | Y_j = 1), j = 1, 2$ (2.4.2)

Based on these assumptions, Chatterji and Shih (2001) have developed the model and have discussed it elaborately in their paper.

Wienke, et. al. (2003) proposed a mixture cure model to analyze bivariate time-to-event data. Correlated gamma frailty model was used to specify the dependency structure between the failure times of two susceptibles. As remarked by Chatterjee and Shih (2003), the model developed by Wienke, et al (2003) is a particular form of that proposed by Chatterjee and Shih (2001). The model proposed by Price and Manatunga (2001) used a frailty to account for the correlation between individuals.

Chen, Ibrahim and Sinha (1999) developed a bivariate cure model in the Bayesian context based on the BCH model. It was obtained as follows. Suppose $Y_1$ and $Y_2$ denote the time to relapse of cancer and time to death respectively, with cumulative distribution functions $F_1(y_1)$ and $F_2(y_2)$. Let $N_1$ and $N_2$ be the number of tumor cells capable of metastasizing corresponding to the variables $Y_1$ and $Y_2$ respectively. It is assumed that $N_1$ and $N_2$ are independent and follow Poisson distribution with mean $\theta_k \omega$, $k = 1, 2$. The component $\omega$ is a frailty component introduced into the model to induce correlation between $N_1$ and $N_2$. The survival function for the population is given by:

$$S(y_1, y_2 | \omega) = \exp[-\omega[\theta_1 F_1(y_1) + \theta_2 F_2(y_2)]]$$ (2.4.3)

This model was later extended by Chen, et al (2002) to a multivariate set up. A positive stable distribution was considered for the frailty component. Discussions on positive stable distribution are found in Hougaard (1986a, b and 1995), Manatunga (1989), Oakes (1994), Samorodnitsky and Taqqu (1994), Lam and Kuk (1997) and Qiou, Ravishankar and Dey (1999).
Yin (2005) proposed two forms of cure models with frailty components, based on the BCH model to analyze correlated or clustered failure time data in a multivariate set up. The first model was called the promotion time frailty cure model in which the population hazard was given by

\[ \lambda_{pop}(t | Z, W) = \lambda(t) W \exp(-\Lambda(t) W) \exp(\beta Z) \]  

(2.4.4)

The second model was called the gamma frailty cure model for which population hazard was

\[ \lambda_{pop}(t | Z, W) = f(t) W \exp(\beta Z) \]  

(2.4.5)

In both these models, \( Z \) denotes the covariate vector, \( W \) the frailty, \( \lambda(t) \) and \( f(t) \) the unknown baseline hazard and the density functions respectively and \( \Lambda(t) \) the cumulative hazard function. Details can be found in Yin (2005).

Techniques for estimation of cured proportion when there are partially observed or missing covariates have been discussed by Cho, Schenker, Taylor and Zhuang (2001) and Chen and Ibrahim (2001). There exists literature based on comparison of cured proportions in various groups. The work by Gray and Tsiatis (1989), Sposto, Sather and Baker (1992), Lee and Sather (1995) and Broet, et al (2001) are a few to be mentioned.

There is relatively little work on joint modeling of longitudinal and survival data which include a cure fraction. For instance, a joint longitudinal and survival mixture cure model was proposed by Law, Taylor and Sandler (2002). The authors obtained the maximum likelihood estimators of the parameters using Monte Carlo Expectation Maximization (EM) algorithm. Brown and Ibrahim (2003) also proposed a model in the Bayesian context using the BCH model for jointly modeling the longitudinal and the time-to-event data with a cure fraction. Gibbs sampler (Gelfand and Smith (1990)) was used for the estimation of the parameters.

An interesting point to be noted is that it is futile to discuss about the estimation of cured proportion when there is no such proportion. In such cases, the cure models will face the problem of estimation of parameters and therefore the non-parametric methods suggested by Maller and Zhou (1992, 1995) should be applied first, to test for the presence of the cured fraction.
2.5 Cure Model for Recurrent Event Count Data

In the previous sections, the main focus of the discussion was on the time to occurrence of an event. In diseases like epilepsy, asthma and urinary tract infection, the variable of interest could be the number of times the event occurs during the follow up period. Thus the response variable, or the variable of interest is the number of recurrences of the event under consideration. Clayton (1994) argues that the inference based on the time to occurrence of an event is also applicable to the number of recurrences of an event. He proposes a Poisson process for the number of recurrences when there is no censoring and establishes the equivalence of the inference based on the Cox's (1972) proportional hazards (PH) model and the model based on the Poisson process.

When the duration of follow up is the same for all the patients, the Poisson process reduces to a Poisson distribution. For populations consisting of cured patients, the number of zero recurrences would increase when compared to the model based on the Poisson process (distribution). Here, the random zeroes correspond to those patients for whom the recurrences of the event do not take place during the entire follow up period but are susceptible to the disease at any point of time, even though they are cured of it on treatment at the time of the study. The number of deterministic zeroes correspond to those patients who are cured and are not susceptible to the disease. This type of a situation, which consists of both deterministic as well as random zeroes, can be handled by considering a mixture of two random variables, one being degenerate, and the other, following an appropriate process (distribution).

Following Clayton (1994), we propose a new cure model which is a Poisson process (distribution) inflated at zero (Details regarding zero inflated models are deferred to chapter 3 of the thesis.) To simplify the situation, we assume that the follow up period is the same for all the patients and that there are no censored observations, i.e., there are no patients who are lost to follow up. Let \( Y \) be the random variable denoting the number of times a disease or an adverse event of interest occurs. The probability mass function (p.m.f.) of \( Y \) is given by

\[
P[Y = y] = \begin{cases} 
  p + (1 - p) \exp(-\lambda), & \text{when } y = 0 \\
  (1 - p) \exp(-\lambda) \frac{\lambda^y}{y!}, & \text{when } y = 1, 2, \ldots 
\end{cases}
\]  

(2.5.1)
Here, $p$ denotes the proportion of patients cured in the population. As in the case of the standard cure model and the BCH model, the covariates here, can be linked to the cured proportion $p$ or to the mean parameter $\lambda$ respectively, through the logit link or the log link functions given by

$$\logit p_i = \log \frac{p_i}{1 - p_i} = \alpha_0 + \alpha_i z_{i1} + \ldots + \alpha_k z_{ik}$$

and

$$\log \lambda_i = \beta_0 + \beta_i x_{i1} + \ldots + \beta_q x_{iq}$$ (2.5.2)

where $z_i$ and $x_i$ are the covariate vectors, and $p_i$ and $\lambda_i$ are the parameters associated with the $i^{th}$ individual.

The model given by (2.5.1) is a mixture distribution of two random variables. It has a direct analogy with the standard cure model (2.2.1). The new cure model (2.5.1) can also be linked to the BCH model or the promotion time cure model (2.2.8).

Suppose there are $N$ carcinogenic cells in a patient. As in the case of promotion time cure model (2.2.8), let $N$ follow the Poisson distribution with parameter $\theta$. Let the number of recurrences of a disease be $Y$. Conditional on $N$, the number of recurrences $Y$ follows the Poisson distribution with parameter $\lambda(N)$, i.e., $Y \sim P(\lambda(N))$. Here, $\lambda(N)$ denotes the average number of recurrences with $N$ carcinogenic cells left active in the cancer patient who is said to be cured at the time of observation, but is susceptible to the disease. The marginal p.m.f. of $Y$ is given by

$$P(Y = y | N) = \sum_{N=0}^{\infty} \frac{(-\lambda(N))^y}{y!} e^{-\lambda(N)} N!$$, $y = 0, 1, 2$,

(2.5.3)

If the functional form of $\lambda(N)$ is not simple, then this p.m.f. will not have a closed form solution. A simplifying assumption is that the distribution of the number of recurrences is independent of the distribution of $N$. Under this assumption, we have

$$P[Y = 0] = P[\text{There are no carcinoma cells}] + P[N \geq 1] = P[\text{The disease does not recur}]$$

$$= \exp(-\theta) + (1 - \exp(-\theta)) \exp(-\lambda)$$

and $P[Y = y] = P[N \geq 1] \frac{e^{-\lambda} \lambda^y}{y!}$

$$= \frac{(1 - \exp(-\theta)) \exp(-\lambda) \lambda^y}{y!}, \quad y = 1, 2, 3, \ldots$$ (2.5.4)
Since the Poisson distribution is used to determine the number of cells capable of metastasizing, the natural choice for the distribution of $Y$ is again Poisson with parameter $\lambda$ and the model reduces to a zero inflated Poisson model. This is the biological explanation for the use of zero inflated Poisson distribution as a cure model for recurrent event count data.

Clayton (1994) does not deal with censored cases but suggests that the censored data can be handled by treating them as missing observations. In chapter 5 of this thesis, we propose a new method for handling censored data and the details are deferred to that chapter.

2.5.1 Example

The applications of zero inflated distributions based on the Poisson, the negative binomial and the generalized Poisson distributions are available in literature. Noteworthy papers based on these are that of Lambert (1992) and Bhattacharya, et al (2008). However, the inflated distributions have not been used as cure models. Since at the time of writing the thesis, we could not get a real life data set on the recurrent event count data, we considered the analysis of a data set which relates to the current status of an event for illustrating the application of our model. The data set is from a dental epidemiological study conducted to estimate the proportion of people who do not have any oral health problems based on their DMFT (Decayed, Missing, Filled Teeth) indices. In dental epidemiology, the DMFT index is a measure used in determining the oral health status of individuals. It gives the number of decayed, missing or filled permanent teeth in the mouth of an adult. It ranges from 0 to 32. An individual is said to have good oral health if his DMFT index is zero.

The present example is an example for count data. It also relates to a current status data. Therefore, the use of cure models may not be fully justified. However, Andersson (2007) has used cure models for the analysis of current status data on cancer patients. He has pointed out the usefulness of the cure models for valid inference on current status time series data. This motivated us to consider the present cross sectional data from a dental epidemiological study.
The data set consisted of 2000 subjects who were examined during a survey conducted by A. B. Shetty Memorial Institute of Dental Sciences, Karnataka, India. The survey was conducted to estimate the proportion of people who did not have any dental problems, based on their DMFT index. This proportion of people can be considered as the cured proportion. The DMFT index along with the age, gender, occupation, dietary habits, brushing habits, viz., the brushing frequency and the method, the present and the past history of medication of the patients, whether self or prescribed by a physician or a medical practitioner, were recorded.

We have used a zero inflated Poisson distribution to estimate the proportion of people who did not have any dental problems at the time of the survey. In the model, a log-link function was used to relate the parameter \( \lambda \). The inflate parameter \( p \) was treated as a constant. The analysis was carried out using the statistical software STATA 7.5. The fitting of the model was done using backward elimination procedure. The results are summarized in Table 2.1 given below.

Table 2.1 ZIP Regression model to estimate the inflate parameter

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient ( \beta )</th>
<th>Standard Error</th>
<th>( p ) - value</th>
<th>95% confidence interval for ( \beta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>0.0039</td>
<td>0.0012</td>
<td>0.001</td>
<td>(0.0016, 0.0062)</td>
</tr>
<tr>
<td>gender</td>
<td>0.0319</td>
<td>0.0346</td>
<td>0.357</td>
<td>(-0.0359, 0.0096)</td>
</tr>
<tr>
<td>diet</td>
<td>-0.0793</td>
<td>0.0355</td>
<td>0.026</td>
<td>(-0.1489, -0.0097)</td>
</tr>
<tr>
<td>Brushing method</td>
<td>-0.0307</td>
<td>0.0293</td>
<td>0.295</td>
<td>(-0.0882, 0.0268)</td>
</tr>
<tr>
<td>Brushing frequency</td>
<td>-0.3195</td>
<td>0.0474</td>
<td>&lt;0.001</td>
<td>(-0.4125, -0.0265)</td>
</tr>
<tr>
<td>Past history of prescribed medication</td>
<td>0.4353</td>
<td>0.1358</td>
<td>0.001</td>
<td>(0.1693, 0.7018)</td>
</tr>
<tr>
<td>Self medication</td>
<td>-0.4117</td>
<td>0.1366</td>
<td>0.003</td>
<td>(-0.6793, -0.1441)</td>
</tr>
</tbody>
</table>

A glance at the above table shows that the variables influencing the oral health status are the age of the individual, diet, brushing frequency, past history of prescribed medication and self medication. These variables have a \( p \)-value less than 0.05. The estimate of logit \( p \) was -0.8408 with a standard error of 0.0554 and a highly significant \( p \)-value less than 0.001. Thus, the estimate of the inflate parameter \( p \) which denotes the proportion insusceptible to oral diseases is 0.3014 (30.14%). In this example for count data, \( p \) actually denotes the proportion of people having their
DMFT index as zero. This proportion did not require any dental treatment at the time of the survey

2.6 Conclusion

In this chapter, an exposure to the cure models has been given by reviewing the existing cure models in univariate as well as in multivariate set up. The applications of cure models include the study of onset of secondary complications of a disease in two or more organs of a patient and the study of onset of several other diseases in an individual who is already suffering from one disease. The cure models are also being used in jointly modeling the overall risk of a disease and the distribution of the age-at-onset of the disease for the diseased individuals. The existing cure models are based on the time to occurrence of an event such as a disease. We have proposed a cure model for recurrent event count data which is an extension of the work by Clayton (1994). A biological explanation for this model is given and it has been observed that the proposed model is a natural extension of the cure model developed by Yakovlev, et al. (1993). An application of the model to a data set from a dental health set up, provided good information regarding the percentage of the population who did not suffer from any dental problems at a given point of time. It has been found that age, diet, brushing frequency, past history of prescribed medication and self medication influence the oral health status of individuals. The conclusions coincide with that of Bohning, et al. (1999) and Lewsey and Thomson (2004).

A paper based on this chapter was presented at the International Indian Statistical Association (IISA) Conference on Statistics, Probability and Related Areas organized by the Department of Statistics, Cochin University of Science and Technology (CUSAT), Cochin, India, during January 2-5, 2007. The paper has also been published in the electronic indexed journal INTERSTAT in the November 2008 issue.