Appraisal of Models for the Study of Disease Progression in Psoriatic Arthritis

Thesis submitted to the University of London for the degree of Doctor of Philosophy in the Faculty of Science

by

Rebeca Aguirre-Hernández

Department of Statistical Science
University College London

December 5, 2000

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Abstract

The subject of the thesis is the use of models for disease progression in arthritis, with special emphasis on Markov regression models. The first objective of the thesis is to propose a Pearson type goodness of fit test for stationary Markov models with covariates. The grouping technique proposed by Hosmer and Lemeshow for logistic regression models is extended to models with response variables recorded repeatedly over time. This generalization is particularly appropriate for panel data in which different numbers of observations, unequally spaced, are obtained for each sampling unit. Due to the complexity of the theoretical distribution of the test statistic, bootstrap methodology is used to calculate the distribution of the statistic under the null hypothesis. The power of the goodness of fit test is investigated for a particular model using a nested bootstrap algorithm. The proposed test is applied to a data set obtained at the University of Toronto with the objective of identifying prognostic factors for disease progression in psoriatic arthritis (PsA), measured via the number of damaged joints.

As the Markov regression model does not fit the PsA data, the second objective of the thesis is to consider potentially better models. A larger data set is analysed for this purpose . Additionally, neither the outcome variable nor the covariates are categorized. Two mixture regression models for longitudinal data are examined to determine if there is statistical evidence supporting the hypothesis that a proportion of individuals never develop damaged joints. The results indicate that a negative binomial regression model without added zeros might provide a reasonable approach. The goodness of fit of this model is examined using bootstrap methodology, comparable to that used for the Markov regression model.

Aknowledgements

I express my appreciation to Professor V. T. Farewell for the time he devoted to go through and discuss my work. His sound supervision enabled me to successfully complete the various stages of the PhD programme while making it an enjoyable activity. I am grateful to Mr. Henry Tillotson and Ms. Ina Dau for their help to correct my FORTRAN programs and to Dr. R. Chandler for updating the file "randgen.f" that produces pseudo-random numbers from several distributions. I am indebted to the people at the National University of Mexico who taught me statistics and encouraged me to pursue a PhD. Thanks are also due to Dr. D. Wright for his interest in my progress while he worked at the Statistical Science Department at UCL. I am grateful to the PhD students for their companionship, the friends that I met in London, and the old friends who have kept in touch. Particular thanks to my parents, Jesús, Ana, Rosalía, and Luis for their support. As a student at University College London I have benefited from a grant from CONACYT.

Contents

List of Tables

List of Figures

Chapter 1

Markov regression models

1.1 Introduction

Based on previous studies, Gladman, Farewell and Nadeau [1] state that psoriatic arthritis (PsA) is a disease with a variable course. Some patients develop joint deformity and destruction, as well as disability, while others follow a more benign course. The authors hypothesized that there may be identifiable prognostic indicators for disease severity. If such prognostic indicators exist, their identification could help to design a treatment for the disease. Thus, Gladman, Farewell and Nadeau conducted a study at the University of Toronto Psoriatic Arthritis Clinic in order to identify markers for severe disease in PsA.

The authors designed a prospective study lasting 14 years. It was planned to assess the patients at 6-month intervals according to a standard protocol. At each visit, clinical and laboratory assessments of both active inflammation and clinic damage were performed. A joint was considered to be damaged if the clinic assessment showed a decreased range of motion of more than 20% of the normal range that could not be attributed to active inflammation, the presence of contractures, subluxation, loosening or ankylosis, or previous surgery.

For the purpose of the analysis, the authors categorized the number of damaged joints recorded at each visit into 4 classes that reflected different stages of the disease. Based on this classification, a patient is said to be in the first stage if he/she has been diagnosed with psoriatic arthritis but has not developed damaged joints. Patients in the second stage of the disease have 1 to 4 damaged joints. Individuals in the third stage have between 5 and 9 damaged joints inclusive and subjects in the forth stage have 10 or more damaged joints. Thus, progression in damage is defined as a transition to a more severe stage of the disease.

Consecutive observations from the same patient are usually correlated while observations belonging to different individuals are considered to be independent. Gladman, Farewell, and Nadeau [1] assumed that, for every patient, the stage of the disease at the next clinic visit is independent of the previous stages given the current disease stage. This simple correlation structure between consecutive observations is known as the Markov property.

As mentioned before, clinic visits were planned every 6 months but in observational studies - like the PsA study - variations between the inter-visit periods are inevitable. In the PsA data, the extent of this variability is such that the authors assumed that the observations were made on a continuous-time scale rather than a discrete time scale. This is an important distinction as, in the two cases, the data are analysed in different ways. In continuous-time Markov models, the change from one stage to another is described by the transition rates.

The transition rates are not constant from one patient to another because, as mentioned before, the course of the disease is variable. In an initial analysis, Gladman, Farewell and Nadeau considered that the prognostic factors that could affect the transition rates are: sex, functional class, number of actively inflamed joints, number of effused joints, Lansbury index, rheumatoid factor, erythrocyte sedimentation rate, and medication level. Only the values recorded at the first clinic visit were used in the analysis. Markov models with transition rates that depend on covariates or explanatory variables are known as Markov regression models.

The authors also assumed that the transition rates are stationary so they are not affected by the time at which the clinic visits take place.

Psoriatic arthritis is a chronic disease. As it evolves, the joints become irreversibly damaged and patients progress to a more severe stage. Hence, it is sensible to assume that the only transition rates greater than zero are the ones describing a change to the next more advanced stage. Models having transition rates with this kind of structure are called progressive Markov models.

Multi-state Markov models are increasingly being used in medical applications. For example, Kalbfleisch and Lawless [2] used such a model to study smoking prevention programmes in schoolchildren and Gentleman *et. al.* [3] and Longini *et. al.* [4] considered models for HIV disease. If the response variable is qualitative or discrete, then the values it assumes are called states. Therefore, in the rest of the chapter, a disease stage is also referred to as a state.

Finally, a useful concept is that of an absorbing state. Individuals entering an absorbing state remain in it forever. Stage 4 of the PsA disease is an absorbing state. In other applications, the absorbing state is the death caused by the disease under investigation. Notice that in the first case, sample units can still be observed after they have entered the absorbing state while in the second case this is impossible. Data gathered after an individual has entered an absorbing state is disregarded as it provides no information for the estimation of the parameters of a Markov model.

The next sections deal with the theoretical aspects of time-continuous Markov regression models with stationary transition rates. I establish the notation, define the terminology and the characteristics of the model and explain the method used to estimate the parameters. These sections can be omitted by non-technical readers. In section 1.6 I describe the data collection process and define the variables measured in the Psoriatic Arthritis Clinic at the University of Toronto. As the PsA database is constantly updated, a new data set became available while I worked on my research project. Thus, in this thesis I examine three related data sets. The differences between them and some descriptive statistics are presented in section 1.7. Also, in section 1.8 I present and interpret the estimates of the model fitted by Gladman, Farewell and Nadeau.

1.2 Probability theory

Consider a random sample of size n in which each subject is observed several times in one of K different states. The measurement times may be different for each subject and unequally spaced. The total number of observations may also vary between subjects.

One way to model the correlation between the observations of each subject is to condition the future observation on the previous ones. This kind of model is known as the Markov process with discrete states in continuous time. The Markov process is different for each subject when the number of observations and the measurement times are not the same for every individual.

The m_i observations for subject i will be denoted as:

$$Y_{i,1}, \ldots, Y_{i,j}, \ldots, Y_{i,m_{i-1}}, Y_{i,m_i}$$

and the times at which they are obtained as:

$$t_{i,1} < \ldots < t_{i,j} < \ldots < t_{i,m_{i-1}} < t_{i,m_i}$$

where $Y_{i,j} \in \{1, 2, \ldots, K\}$ for all $j = 1, 2, \ldots, m_i$ and $i = 1, 2, \ldots, n$. The total number of states is K and throughout the thesis is considered to be finite *i.e.* $K < \infty$. It is said that a transition from state a to state b occurred in the time interval $(t_{i,j}, t_{i,j+1})$ if $Y_{i,j} = a$ and $Y_{i,j+1} = b$. Notice that it is not precisely known when the transition took place.

The Markov process is of order one if every observation depends only on the preceding one. This model is defined by the set of transition probabilities:

$$p_{i,j(a,b)} \equiv p_{(a,b)}(t_{i,j}, t_{i,j+1}) = P(Y_{i,j+1} = b \mid Y_{i,j} = a)$$

If these probabilities remain constant through time the Markov process is said to be stationary or time-homogeneous. Henceforth only stationary Markov processes will be considered unless otherwise specified.

The transition probabilities have the following properties:

$$
\begin{aligned}
0 \leq p_{i,j(a,b)} \quad \forall \quad a,b = 1,2,\ldots,K; \quad i = 1,2,\ldots,n \\
\sum_{b=1}^{K} p_{i,j(a,b)} = 1 \quad \forall \quad a,b = 1,2,\ldots,K; \quad i = 1,2,\ldots,n
\end{aligned}
\tag{1.1}
$$

Condition (1.1) implies that at the next observation time subject i will be observed at one of the K states including the present one. The matrix of transition probabilities for subject i at time $t_{i,j}$ is:

$$
\mathbf{P_{i,j}} = \{p_{i,j(a,b)}\} = \begin{pmatrix} p_{i,j(1,1)} & p_{i,j(1,2)} & \cdots & p_{i,j(1,K)} \\ p_{i,j(2,1)} & p_{i,j(2,2)} & \cdots & p_{i,j(2,K)} \\ \vdots & \vdots & \vdots & \vdots \\ p_{i,j(K,1)} & p_{i,j(K,2)} & \cdots & p_{i,j(K,K)} \end{pmatrix}
$$

In practice, the transition probabilities are unknown. It is possible, however, to estimate the transition rates defined as [8]:

$$
q_{i(a,b)} = \lim_{\Delta \to 0} \frac{P[Y_{i,j+\Delta} = b \mid Y_{i,j} = a]}{\Delta} \quad \text{for} \quad a \neq b \tag{1.2}
$$

where $t_{i,j+1} = t_{i,j+\Delta}$. From equation (1.1) it follows that:

$$
q_{i(a,a)} = -\sum_{b \neq a} q_{i(a,b)} \quad \text{for} \quad a = b \tag{1.3}
$$

The $K \times K$ matrix of transition rates for subject i is:

$$
\mathbf{Q_i} = \{q_{i(a,b)}\} = \begin{pmatrix} q_{i(1,1)} & q_{i(1,2)} & \cdots & q_{i(1,K)} \\ q_{i(2,1)} & q_{i(2,2)} & \cdots & q_{i(2,K)} \\ \vdots & \vdots & \vdots & \vdots \\ q_{i(K,1)} & q_{i(K,2)} & \cdots & q_{i(K,K)} \end{pmatrix}
$$

The Chapman-Kolmogorov equations:

$$p_{i,j(a,b)} = p_{a,b}(t_{i,j}, t_{i,j+1}) = \sum_{k=1}^{K} p_{a,k}(t_{i,j}, t_{i,j+\delta}) p_{k,b}(t_{i,j+\delta}, t_{i,j+1}) \quad \text{where} \quad 0 < \delta \leq 1$$

together with expressions (1.2) and (1.3) allow us to express $\mathbf{P_{i,j}}$ in terms of $\mathbf{Q_i}$ as follows:

$$\frac{\partial \mathbf{P_{i,j}}}{\partial \Delta_j} = \mathbf{P_{i,j}} \mathbf{Q_i} = \mathbf{Q_i} \mathbf{P_{i,j}} \tag{1.4}$$

The methods of ordinary differential equations with initial condition:

$$p_{i,j(a,b)} = \begin{cases} 0 & \text{if} \quad a \neq b \\ 1 & \text{if} \quad a = b \end{cases}$$

yield the following solution to (1.4):

$$\mathbf{P_{i,j}} = \exp(\mathbf{Q_i} \Delta_j) = \mathbf{I} + \sum_{r=1}^{\infty} \mathbf{Q_i^r} \frac{\Delta_j^r}{r!} \tag{1.5}$$

This series is always convergent when the number of states is finite. If $\mathbf{Q_i}$ has distinct eigenvalues: $\lambda_{i,1}, \lambda_{i,2}, \ldots, \lambda_{i,K}$ then the spectral decomposition of $\mathbf{Q_i}$ can be used to compute expression (1.5) [1]. The spectral decomposition of $\mathbf{Q_i}$ is $\mathbf{Q_i} = \mathbf{B_i} \operatorname{diag}(\lambda_{i,1}, \lambda_{i,2}, \ldots, \lambda_{i,K}) \mathbf{C_i'}$, where $\mathbf{B_i} \mathbf{C_i'} = \mathbf{I}$. The columns of $\mathbf{B_i}$ and $\mathbf{C_i'}$ are respectively the right and left eigenvectors of $\mathbf{Q_i}$. This implies that equation (1.5) can be computed as:

$$\mathbf{P_{i,j}} = \mathbf{B_i} \mathbf{D_{i,j}} \mathbf{C_i'} \tag{1.6}$$

where $\mathbf{D_{i,j}} = \operatorname{diag}(e^{\Delta_j \lambda_{i,1}}, e^{\Delta_j \lambda_{i,2}}, \ldots, e^{\Delta_j \lambda_{i,K}})$.

1.3 Transition rates that depend on covariates

Sometimes the goal of the study is to describe the pattern of transitions or to predict the future state of a subject using several explanatory variables which

[1] The Jordan canonical decomposition is used when $\mathbf{Q_i}$ has repeated eigenvalues.

are thought to influence the transition rates. The explanatory variables can be measured once - at the beginning of the study - or every time the state of the subject is recorded. Obviously the second type of data is more difficult to analyse than the former one. In this thesis I only consider the situation in which for each subject a set of $p-1$ explanatory variables or covariates are measured at the beginning of the study. The vector of covariates for subject i will be denoted as $z_i' = (1, z_{i,1}, z_{i,2}, \ldots, z_{i,p-1})$ for all $i = 1, 2, \ldots, n$.

The transition rates are numbers greater than or equal to zero, hence it is inappropriate to model them as a linear combination of z_i'. A more suitable parametrization is:

$$\ln(q_{i(a,b)}) = \beta_{0(a,b)} + \sum_{u=1}^{p-1} \beta_{u(a,b)} z_{i,u} \quad \text{for} \quad a \neq b \qquad (1.7)$$

The above expression implies that the effect of the covariates can change from one transition rate to another.

1.4 Inference for Markov regression models

An estimate of $\beta_{(a,b)}' = (\beta_{0(a,b)}, \beta_{1(a,b)}, \ldots, \beta_{p-1(a,b)})$ for $a, b \in \{1, 2, \ldots, K\}$ provides an estimate for $\mathbf{Q_i}$ and $\mathbf{P_{ij}}$.

Conditional on the first state at which subject i is observed, $y_{i,1}$, the contribution of this individual to the likelihood function is:

$$L_i(\beta) = p_{y_{i,1},y_{i,2}} p_{y_{i,2},y_{i,3}} \cdots, p_{y_{i,m_{i-1}},y_{i,m_i}} = \prod_{j=1}^{m_{i-1}} p_{y_{i,j},y_{i,j+1}}$$

where $\beta' = (\beta_{(1,1)}', \beta_{(1,2)}', \ldots, \beta_{(1,K)}', \ldots, \beta_{(K,1)}', \beta_{(K,2)}', \ldots, \beta_{(K,K)}')$ and, to simplify the notation, $p_{i,j(y_{i,j},y_{i,j+1})} \equiv P(Y_{i,j+1} = y_{i,j+1} \mid Y_{i,j} = y_{i,j})$ is denoted as $p_{y_{i,j},y_{i,j+1}}$. The likelihood function for β is defined as:

$$L(\beta) = \prod_{i=1}^{n} L_i(\beta) = \prod_{i=1}^{n} \prod_{j=1}^{m_{i-1}} p_{y_{i,j},y_{i,j+1}} \qquad (1.8)$$

To maxime (1.8) with respect to β the quasi-Newton (or scoring) procedure is used. This procedure is based on the score function and the pseudo-information matrix defined by cross products of the observed score functions. This pseudo-information matrix is used for ease of computation. The efficient score for $\beta_{u(a,b)}$ is:

$$
\begin{aligned}
S_{u(a,b)} &= \frac{\partial \log L(\beta)}{\partial \beta_{u(a,b)}} = \sum_{i=1}^{n} \sum_{j=1}^{m_i - 1} \frac{\partial \log p_{y_{i,j}, y_{i,j+1}}}{\partial \beta_{u(a,b)}} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m_i - 1} \frac{1}{p_{y_{i,j}, y_{i,j+1}}} \cdot \frac{\partial p_{y_{i,j}, y_{i,j+1}}}{\partial \beta_{u(a,b)}}
\end{aligned}
$$

Notice that $p_{Y_{i,j}, Y_{i,j+1}}$ is an entry of $\mathbf{P_{ij}}$ so the first derivative of $\mathbf{P_{ij}}$ with respect to $\beta_{u(a,b)}$ needs to be calculated for all $a, b \in \{1, 2, \ldots, K\}$. A method that enables us to calculate these derivatives without having an explicit expression for $\mathbf{P_{ij}}$ in terms of $\beta_{u(a,b)}$ is given by Kalbfleisch and Lawless [2]:

$$
\frac{\partial \mathbf{P_{ij}}}{\beta_{u(a,b)}} = B_i V_{u(a,b)} B_i^{-1} \quad \text{for all} \quad u = 0, 1, \ldots, p-1; \;\; a, b \in \{1, 2, \ldots, K\}
$$

where $V_{u(a,b)}$ is a $K \times K$ matrix with $(h, l)th$ entry:

$$
\frac{g_{h,l}^{u(a,b)} [\exp(\Delta_j \lambda_h) - \exp(\Delta_j \lambda_l)]}{\lambda_h - \lambda_l} \quad \text{if} \quad h \neq l
$$

$$
g_{h,h}^{u(a,b)} \Delta_j \exp(\Delta_j \lambda_h) \quad \text{if} \quad h = l
$$

where $g_{h,l}^{u(a,b)}$ is the (h, l)th entry of $G^{u(a,b)} = B_i^{-1} (\partial \mathbf{Q_i} / \partial \beta_{u(a,b)}) B_i$. The entries of the information matrix $I(\beta)$ are the expected values:

$$
\mathrm{E}\Big[- \frac{\partial^2 \log L(\beta)}{\partial \beta_{u'(a,b)} \partial \beta_{u(a,b)}} \Big] \qquad \text{for} \qquad u, u' = 0, 1, \ldots, p-1 \quad \text{where}
$$

$$
\begin{aligned}
\frac{\partial^2 \log L(\beta)}{\partial \beta_{u'(a,b)} \partial \beta_{u(a,b)}} &= \frac{1}{p_{y_{i,j}, y_{i,j+1}}} \cdot \frac{\partial^2 p_{y_{i,j}, y_{i,j+1}}}{\partial \beta_{u'(a,b)} \partial \beta_{u(a,b)}} \\
& \quad - \Big(\frac{1}{p_{y_{i,j}, y_{i,j+1}}} \Big)^2 \cdot \frac{\partial p_{y_{i,j}, y_{i,j+1}}}{\partial \beta_{u(a,b)}} \cdot \frac{\partial p_{y_{i,j}, y_{i,j+1}}}{\partial \beta_{u'(a,b)}}
\end{aligned}
$$

$$\text{But}\quad \frac{\partial^2 p_{y_{i,j},y_{i,j+1}}}{\partial\beta_{u'(a,b)}\partial\beta_{u(a,b)}} = 0 \quad \text{so}$$

$$\mathrm{E}\Big[-\frac{\partial^2 \log L(\beta)}{\partial\beta_{u'(a,b)}\beta_{u(a,b)}}\Big] = \Big(\frac{1}{p_{y_{i,j},y_{i,j+1}}}\Big)^2 \cdot \frac{\partial p_{y_{i,j},y_{i,j+1}}}{\partial\beta_{u(a,b)}} \cdot \frac{\partial p_{y_{i,j},y_{i,j+1}}}{\partial\beta_{u'(a,b)}}$$

If sup $L(\beta)$ is attained in the parameter space, the maximum likelihood estimator of β is the solution to:

$$S_{u(a,b)}(\hat{\beta}) = 0 \qquad \text{for} \quad u = 0, 1, \ldots, p-1; \quad a, b \in \{1, 2, \ldots, K\}.$$

These equations are solved iterativelly starting from an initial value $\beta^{(0)}$. The updated estimate of β is obtained as:

$$\beta^{(1)} = \beta^{(0)} + \{I(\beta^{(0)})\}^{-1} S(\beta^{(0)})$$

When $n \to \infty$, $\sqrt{n}(\hat{\beta} - \beta)$ has an approximate multivariate normal distribution with mean zero and covariance matriz $I(\hat{\beta})$.

1.5 Progressive Markov models.

The matrix of transition rates must be carefully defined so that it reflects the process being studied. In the medical context, Markov models have been used to describe the progression of individuals through the stages of a disease such as cancer, AIDS, and arthritis. As these disease develop, the patients are observed in the same stage or at a more advanced one. Therefore the entries below the main diagonal of $\mathbf{Q}_\mathrm{i}$ must be equal to zero. Processes with this kind of matrix of transition rates are known as birth processes.

Progressive Markov models are a special kind of birth process. They arise when only the transitions $a \to a$ and $a \to a + 1$ are allowed; in other words, $q_{i(a,b)} = 0$ for all $b \notin \{a, a+1\}$. Longini *et. al.* [4] used this kind of structure to analyse a cohort of HIV infected individuals. Also the FORTRAN program that implements the methodology described by Kalbfleisch and Lawless [2] was written to fit progressive Markov models.

1.6 The PsA study

The Psoriatic Arthritis Clinic at the University of Toronto treats a wide range of patients with psoriatic arthritis, from mild to severe disease. This is because the clinic is both a primary, secondary, and tertiary referral center, with patients being referred by family physicians, dermatologists, internists, rheumatologists, and the Psoriasis Education and Research Centre. Patients are admitted to the clinic and monitored only after a definite diagnosis of PsA is established. The diagnosis is determined on the presence of an inflammatory arthritis, usually seronegative for rheumatoid factor, in association with psoriasis. The presence of rheumatoid factor is not an exclusion criterion because approximately 15% of the psoriatic arthritis patients are seropositive.

Since 1978, the clinic runs a database that allows research to be conducted on several aspects of PsA such as the clinical course of the disease, its progression, pathogenesis, and treatment. Patients who agree to participate in the studies are periodically examined by a rheumatologist. A detailed inquiry about the onset of the disease is made at the initial visit. Physical and laboratory examinations are conducted every six months. Radiographs are taken every two years.

Each appointment is scheduled before the patient leaves the clinic and a follow-up call is made one week prior to this set date to remind patients of their appointment. In addition, patients deaths and causes of death are tracked and documented.

The information obtained at each clinic visit is recorded on a standard retrieval protocol and then entered into a SAS data base on a personal computer. The data are checked to ensure correct data entry, and modifications are made when necessary.

At the initial visit, information is obtained regarding the age of onset for both skin and joint disease, pattern of joint disease at onset, and family history of both skin and joint disease. Extra articular features including eye disease, cardiac disease, hypertension, inflammatory bowel disease, and other co-morbid illnesses are documented. The presence and duration of morning stiffness, inflammatory back pain, constitutional symptoms (such as fatigue, abdominal pain, neck pain, neck stiffness, etc.) and functional status of patients are assessed as well. Functional status refers to the degree of functional impairment and is evaluated using the American College of Rheumatology (ACR) classification scheme. Grade I is assigned to patients who are able to perform *all activities without pain or handicap.* Grade II includes those who feel *adequate for most activities of daily living (ADL) but experience some discomfort or limitation.* Grade III includes those whose *ADL are limited to self-care and/or a few daily activities.* Finally, Grade IV refers to patients who are *unable to perform (little or no) self-care activities and/or are confined to a bed or wheel-chair.* A detailed history of past and current medication use along with any side-effects associated to them are also documented. Past medication refers to the medications used by the patient before attending the PsA clinic. The medications can be: none, nonsteroidal antiinflammatory drugs (NSAID), gold or chloroquine, methotrexate or azathioprine, retinoids or psoralen ultraviolet A, and oral corticosteroids.

The physical examination consists of a general medical exam with particular attention to the skin, nails, and the peripheral and axial joints. Clinical measures of function, disease activity and severity, namely grip strength, total number of actively inflamed joints, total number of joint effusions, and total number of damaged joints (including instability and restricted range of motion due to mechanical factors) are assessed. Cervical spine limitation, sacroiliac stress pain, back movements (*i.e.* full extension, full flexion) and the presence of spinal disease are also documented.

There are 66 joints examined for activity. Actively inflamed joints refers to the number of joints with stress pain, tenderness, or effusions and is deemed present if any of these three signs occur. Effusions can be evaluated in 64 of the 66 joints examined for activity. The number of effused joints is determined by the number of joints with excess fluid. The Lansbury index is a quantity that reflects the size of the joint that is actively inflamed. Higher scores are assigned to large joints and lower scores are assigned to small joints.

For example, an inflamed jaw would be assigned 2 points while inflammation of the ankles and hips would be assigned 8 and 24 points respectively.

Laboratory measures includes complete blood counts and differential counts, erythrocyte sedimentation rate (a measure of inflammatory activity that may also reflect disease severity. The ESR is measured in mm/hour by the Westergren method), serologic HLA typing for HLA class I (HLA-A, B, C loci) and class II (HLA-DR,DQ) antigens using the microcytotoxicity assay, biochemical tests of kidney and liver functions, lipid levels (*i.e.* cholesterol and triglycerides), serum uric acid, serum protein electrophoresis and where applicable, immunoglobulin quantification.

The radiological evaluation consists of taking plain radiographs of the peripheral (*i.e* hands, wrists, feet), sacroiliac (both right and left), and spinal (*i.e* cervical, thoracic and lumbar) joints of patients.

In a first study, Gladman, Farewell and Nadeau [1] identified certain clinical predictors for disease progression in PsA. These clinical features are variables that change over time. In an attempt to identify markers for disease progression that are stable, two studies on HLA antigens were conducted [5], [6]. Gladman and Farewell [5] concluded that the B27 and B39 HLA antigens are risk factors for disease progression in PsA, as is the HLA class II antigen DQw3. The effect of these markers and clinical variables is not reassessed for the mixture models proposed in chapter 3.

1.7 The PsA data sets

In this thesis I examine three related data sets. The first is the one used by Gladman, Farewell and Nadeau [1] to fit the Markov regression model. The data set contains 271 patients with at least 2 clinic visits and no missing values on any of the covariates included in the model. All the covariates are binary and the response variable is recorded on an ordinal scale.

While I worked on the research project, a second data base became available. The database refers to a longer follow-up period and thus contains 365 patients. Some of them, however, have only one clinic visit and/or missing values on some variables. The database contains additional variables apart from the ones examined by Gladman, Farewell, and Nadeau. All the variables were recorded in their original measurement units, none was categorized. However, the information on 45 patients was incomplete. These patients have two clinic visits recorded in the data set examined by Gladman *et. al.* [1] but only the information corresponding to the first visit appears in the second database.

In order to use the second database, patients with missing values on any of the prognostic factors identified by Gladman *et. al.* [1] were eliminated. The same was done with those individuals with one clinic visit up to the date of the database creation. The 45 patients with incomplete records were also eliminated. Their missing information could not be retrieved from the data set examined by Gladman *et. al.* [1] because of the discrepancies between the number of variables and the measurement scales.

Thus, the data set examined in chapter 3 contains 285 patients with at least two clinic visits and no missing values on any covariate. The third data set is a subset of the second one and was obtained after eliminating 31 patients with 10 or more damaged joints at presentation. The third data set is examined in chapter 4 and contains 254 patients with information on 2 or more clinic visits.

For each data set, Tables 1.1, 1.2, and 1.3 show some characteristics of the patients at presentation to the clinic. Between 61.8% (Table 1.2) and 69.3% (Table 1.3) of the patients had zero damaged joints when first examined. In fact, a striking feature of the data sets is the percentage of individuals who do not develop damaged joints throughout the study period. These percentages are: 42.8%, 36.8%, and 41.3% for the 1st., 2nd., and 3rd. data sets respectively.

Histograms representing the distributions of the number of clinic visits per patient are shown in Figures 1.1, 1.3, and 1.5. As mentioned before, the 1st. data set corresponds to a shorter follow-up period. Therefore, the median of the number of clinic visits in the data set examined by Gladman *et. al.* [1] is 4 while that for the 2nd. and 3rd. data sets is 6 (see Table 1.4).

Despite the efforts done to assess the patients at 6 month intervals, the time elapsed between clinic visits varied from 0.04 years to 9.78 and 15.35 years (Table 1.5). The average length of the time intervals between visits is 1 year and the median is 0.6 years (Table 1.5). For each data set, a histogram for the time gap between measurements is presented in Figures 1.2, 1.4, 1.6.

In long term studies, the number of patients who are lost to follow-up increases over time. This patient attrition is considered a threat to the study; it may lead to systematic bias, and missing data may reflect pathological factors. Brubacher *et. al.* [7] found that 33% of the patients in the PsA clinic have been lost to followup over a period of 12 years - a percentage which is comparable to other longitudinal prospective studies. Furthermore, the authors concluded that patients who attend the clinic on a regular basis are similar in clinical characteristics to patients who attend the clinic occasionally.

1.8 Results for the PsA study

Over a 14 year period, 305 patients were followed prospectively at the Psoriatic Arthritis Clinic at the University of Toronto. The response (or outcome) variable analysed by Gladman, Farewell and Nadeau [1] is the number of damaged joints divided into 4 categories: 0, 1 to 4, 5 to 9, and 10 or more. These were denoted as states 1 to 4. Clinic assessments were planned at the initial visit and at 6 month intervals although, in practice, considerable variation occurred. After examining the univariate and multivariate effect of several covariates, the authors decided to model the transition rates in terms of: the number of effused joints recorded in the first clinic visit, < 5 or ≥ 5;

Table 1.1: Clinical characteristics of patients at presentation, 1st. data set.

		Number	Percentage
Number of patients entered to computer		271	100
Gender	Female	127	46.9
	Male	144	53.1
Functional Status	Poor (III, IV)	22	8.1
	Medium (II)	161	59.4
	Good (I)	88	32.5
Active joints	High (> 4)	186	68.6
	Medium (1-5)	70	25.8
	Low (0)	15	5.5
Effusions	High (> 4)	44	16.2
	Medium (1-4)	139	51.3
	Low (0)	88	32.5
Lansbury index	High (> 30)	103	38.0
	Low (< 31)	168	62.0
Erythrocyte Sed. Rate	High (> 30)	83	30.6
	Medium (15-30)	98	36.2
	Low (< 15)	90	33.2
Previous medication	Corticosteroids	93	34.3
	High	27	10.0
	None/NSAID	151	55.7
Damaged joints	$5-9$	15	5.5
	$1-4$	71	26.2
	0	185	68.3

	Mean	Range
Age at presentation (years)	42.2	15.5 - 79.2
Duration of arthritis (years)	6.9	0.1 - 47.3

Table 1.2: Clinical characteristics of patients at presentation, 2nd. data set.

		Number	Percentage
Number of patients entered to computer		285	100
Gender	Female	155	54.4
	Male	130	45.6
Functional Status	Poor (III, IV)	31	10.9
	Medium (II)	167	58.6
	Good (I)	87	30.5
Active joints	High (> 4)	199	69.8
	Medium (1-5)	69	24.2
	Low (0)	17	6.0
Effusions	High (> 4)	55	19.3
	Medium (1-4)	143	50.2
	Low (0)	87	30.5
Lansbury index	High (> 30)	119	41.8
	Low (< 31)	166	58.2
Erythrocyte Sed. Rate	High (> 30)	96	33.7
	Medium (15-30)	96	33.7
	Low (< 15)	93	32.6
Previous medication	Corticosteroids	114	40.0
	High	29	10.2
	None/NSAID	142	49.8
Damaged joints	10 or more	31	10.9
	$5 - 9$	16	5.6
	$1 - 4$	62	21.8
	0	176	61.8

	Mean	Range
Age at presentation (years)	43.2	16.3 - 79.2
Duration of arthritis (years)	7.5	0.1 - 47.7

Table 1.3: Clinical characteristics of patients at presentation, 3rd. data set.

		Number	Percentage
Number of patients entered to computer		254	100
Gender	Female	138	54.3
	Male	116	45.7
Functional Status	Poor (III, IV)	22	8.7
	Medium (II)	148	58.3
	Good (I)	84	33.1
Active joints	High (> 4)	175	68.9
	Medium (1-5)	64	25.2
	Low (0)	15	5.9
Effusions	High (> 4)	44	17.3
	Medium (1-4)	133	52.4
	Low (0)	77	30.3
Lansbury index	High (> 30)	98	38.6
	Low (< 31)	156	61.4
Erythrocyte Sed. Rate	High (> 30)	79	31.1
	Medium (15-30)	92	36.2
	Low (< 15)	83	32.7
Previous medication	Corticosteroids	91	35.8
	High	25	9.9
	None/NSAID	138	54.3
Damaged joints	$5-9$	16	6.3
	$1-4$	62	24.3
	0	176	69.3

	Mean	Range
Age at presentation (years)	42.2	16.3 - 79.2
Duration of arthritis (years)	6.6	0.1 - 47.3

Figure 1.1: Histogram representing the distribution of the number of clinic visits per patient in the 1st. data set.

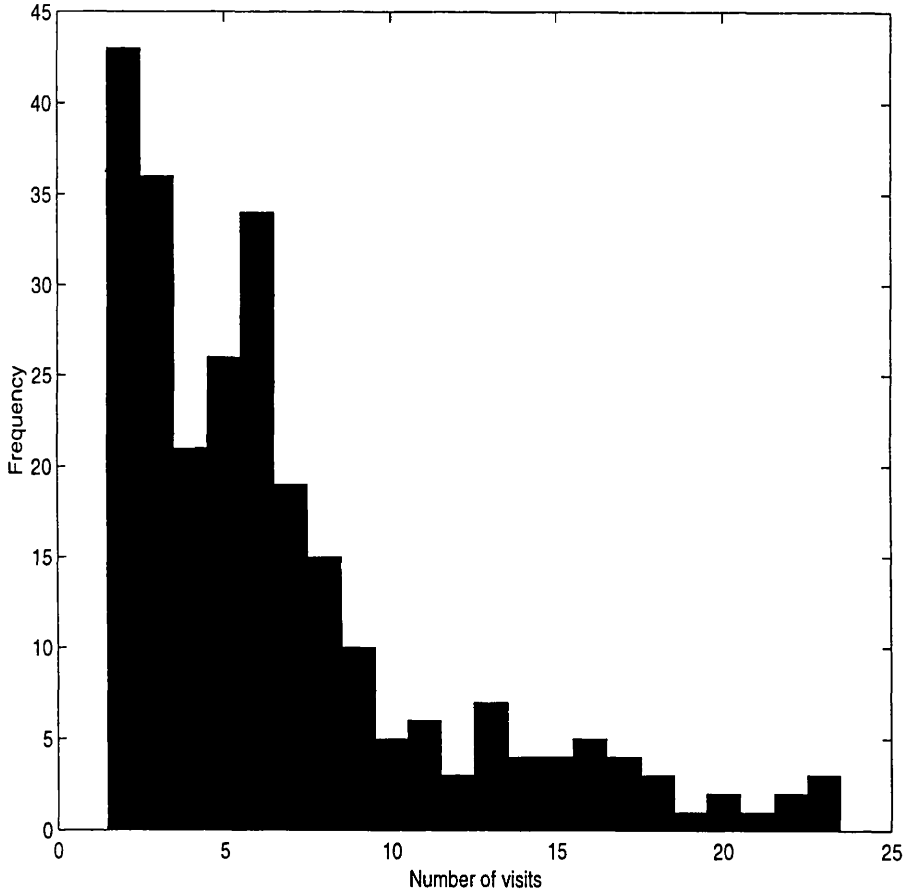


Figure 1.2: Histogram showing the distribution of the time elapsed between consecutive clinic visits in the 1st. data set.

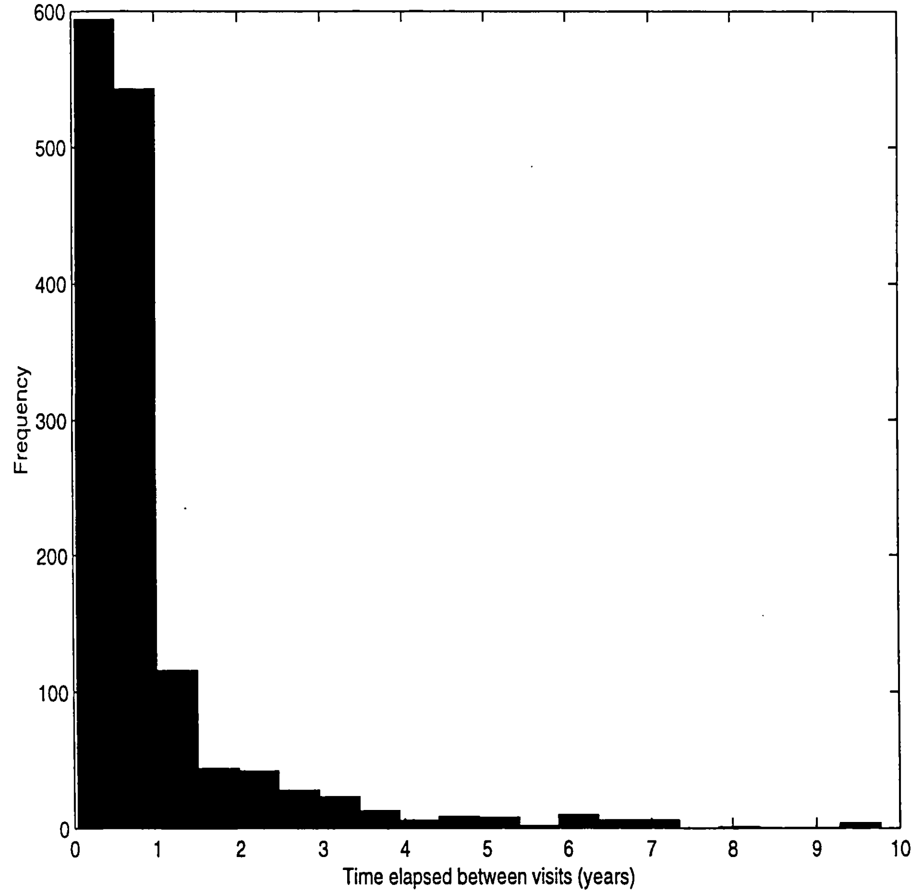


Figure 1.3: Histogram representing the distribution of the number of clinic visits per patient in the 2nd. data set.

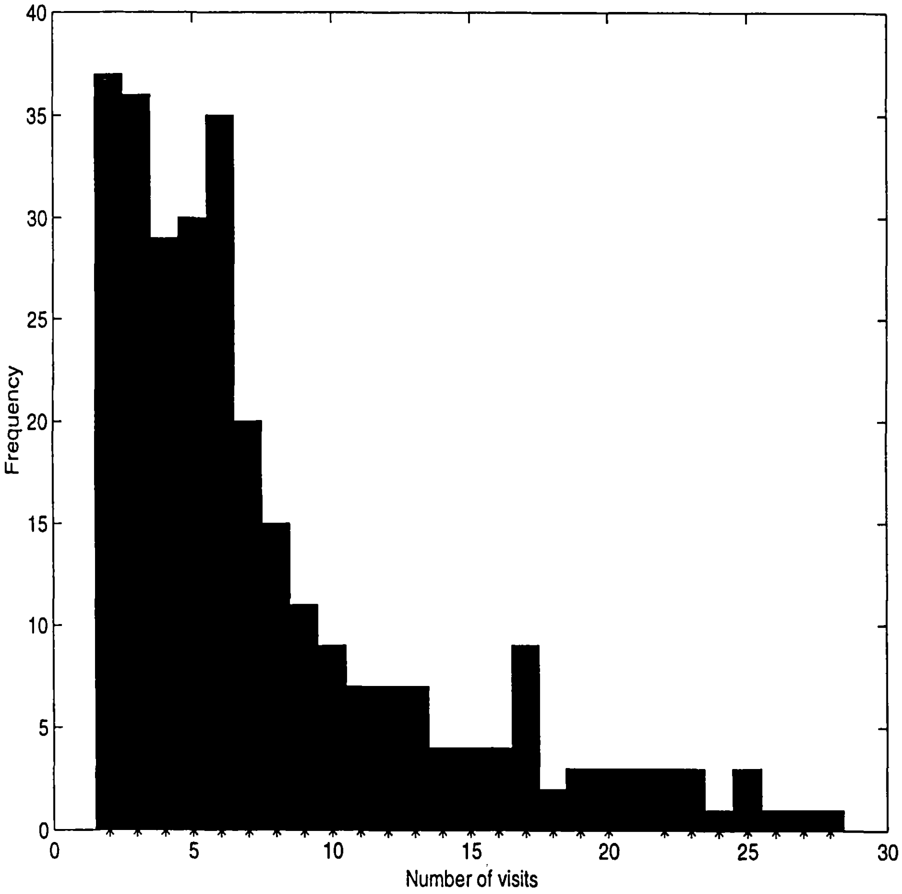


Figure 1.4: Histogram showing the distribution of the time elapsed between consecutive clinic visits in the 2nd. data set.

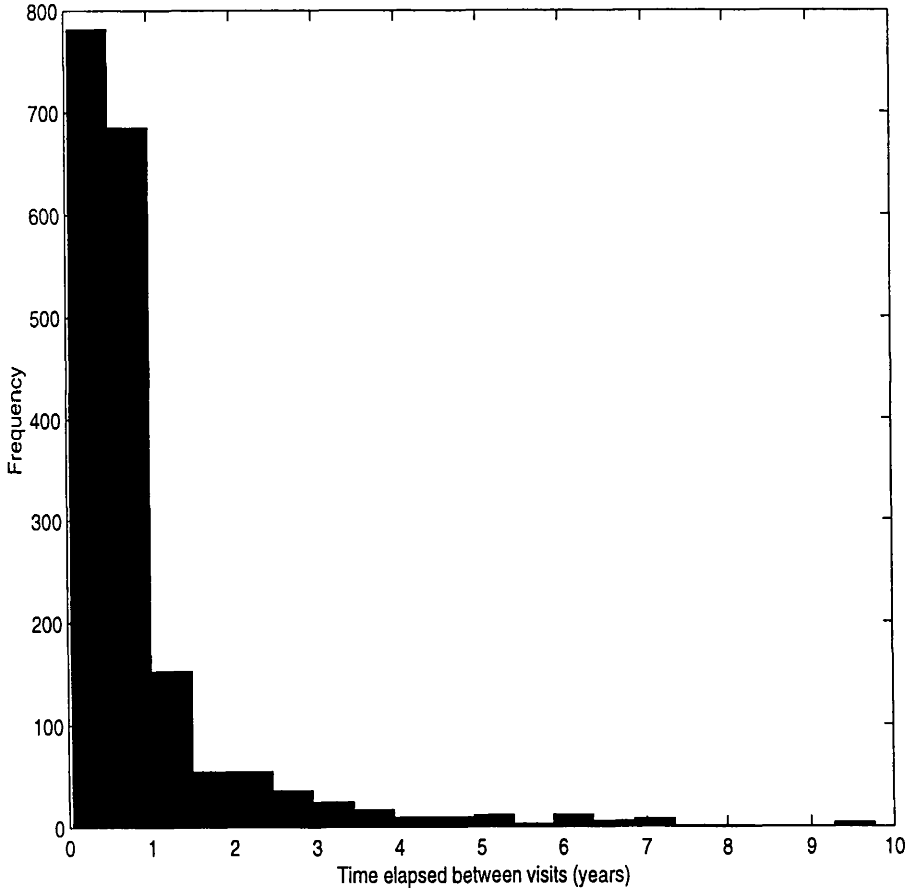


Figure 1.5: Histogram representing the distribution of the number of clinic visits per patient in the 3rd. data set.

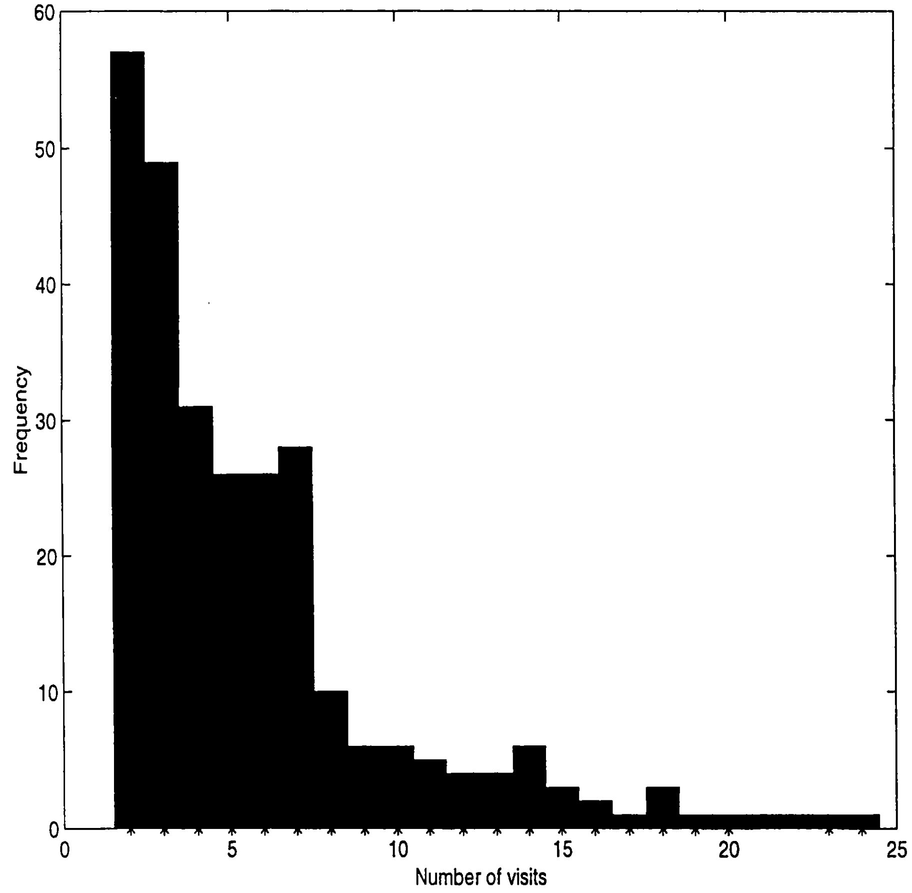


Figure 1.6: Histogram showing the distribution of the time elapsed between consecutive clinic visits in the 3rd. data set.

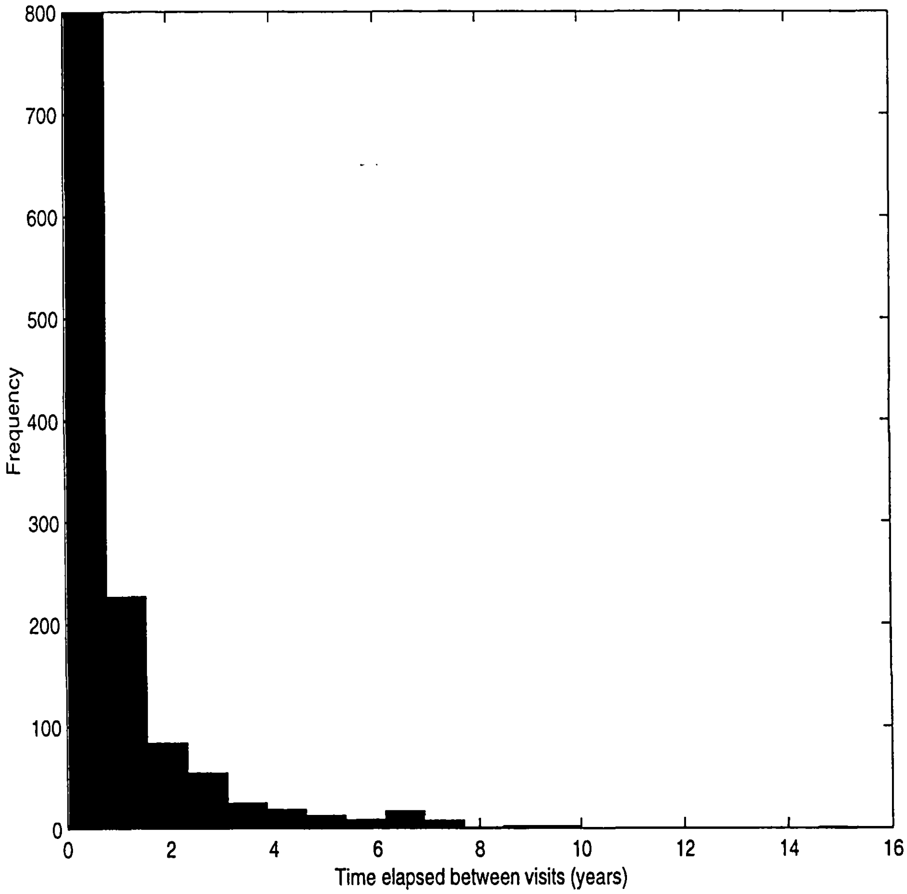


Table 1.4: Statistics for the number of clinic visits.

Data set	Min.	Median	Mean	Max.
1st.	2	4	5.7	24
2nd.	2	6	7.6	28
3rd.	2	6	6.7	23

Table 1.5: Statistics for the time elapsed between clinic visits.

Data set	Min.	Median	Mean	Max.
1st.	0.04	0.61	1.18	15.35
2nd.	0.04	0.56	0.98	9.78
3rd.	0.04	0.57	1.00	9.78

erythrocyte sedimentation rate (ESR) at the initial visit, < 15 mm/h or ≥ 15 mm/h; and the type of medication taken before participating in the study, none or nonsteroidal antiinflammatory medications, disease modifying drugs (DMD) or oral corticosteroids. Patients were also stratified by their initial state in order to adjust for any differences in the referral pattern, although these were not expected to be marked.

Table 1.6 shows the parameter estimates for the stationary Markov regression model fitted by Gladman, Farewell, and Nadeau after eliminating 34 subjects with missing values on at least one covariate. The first column indicates the condition which was coded as one. The numbers in brackets are the standard deviations.

Notice that a different intercept, $\beta_{0(a,a+1)}$, was used to model each transition rate. When no covariates or stratification variables are included in the model, $1/\exp(\hat{\beta}_{0(a,a+1)})$ provides an estimate for the average (mean) time spent in state a, where $a = 1, 2, 3$. Such a model applied to the PsA data suggests that patients with psoriatic arthritis do not develop damaged joints for an average of 11 years. Also, patients with PsA remain in states 2 and 3 for an average of 6.23 and 4.03 years respectively.

As expected, the stratification variables are not significantly different from zero indicating that the state at entry to the PsA Clinic has no effect on the progression of the disease. The estimated relative risks shown in Table 1.7 are the ratio of the transition rate of a patient with a given covariate coded as 1 and the transition rate of a patient with that same covariate coded as 0. It is assumed that both individuals have the same values for the stratifying variables and for the other covariates.

Thus, a patient taking disease modifying drugs (DMD) before entering the study has a risk 1.84 times higher of moving to the next state as compared with a patient taking none or nonsteroidal antiinflammatory medications. Also, subjects using oral corticosteroids prior to the study have a risk that is 1.57 times higher of moving to the next state as compared with subjects not taking oral corticosteroids. The transition rate of a patient with 5 or more effused joints is 1.63 times bigger than the transition rate of a patient with the same characteristics but having less than 5 effused joints. Patients with an erythrocyte sedimentation rate less than 15 mm/h have a smaller risk of moving to the next stage of the disease as compared to patients with an ESR of 15 mm/h or more.

Therefore, based on the information recorded on the first clinic visit, the model suggests that 5 or more effused joints, disease modifying drugs and oral corticosteroids predict progression in damage while an erythrocyte sedimentation rate less than 15 mm/h prevents from such progression.

Table 1.6: Estimated parameters and standard deviations for the Markov regression model fitted by Gladman *et. al.*

	Transition Rates		
Parameter	$1 \to 2$	$2 \to 3$	$3 \to 4$
Constant term	-2.52 (0.14)	-1.67 (0.18)	-1.70 (0.24)
Effused joints ≥ 5	0.49 (0.18)	0.49 (0.18)	0.49 (0.18)
ESR < 15 mm/h	-0.54 (0.19)	-0.54 (0.19)	
Corticosteroids, Yes	0.45 (0.15)	0.45 (0.15)	0.45 (0.15)
DMD Yes	0.61 (0.20)	0.61 (0.20)	0.61 (0.20)
Initial state is 2		-0.51 (0.24)	-0.00 (0.30)
Initial state is 3			-0.64 (0.40)

Table 1.7: Estimated relative risks for each prognostic factor in the Markov regression model fitted by Gladman *et. al.*

		Transition rates		
Covariate	Condition	$1 \to 2$	$2 \to 3$	$3 \to 4$
Number of effused joints	< 5	1	1	1
	≥ 5	1.63	1.63	1.63
Erythrocyte sedimentation rate	≥ 15 mm/h	1	1	
	< 15 mm/h	0.58	0.58	
Use of corticosteroids	No	1	1	1
	Yes	1.57	1.57	1.57
Disease modifying drugs	No	1	1	1
	Yes	1.84	1.84	1.84

Chapter 2

Goodness of fit for Markov regression models

2.1 Introduction

Goodness-of-fit statistics measure the conformity of a sample of data with a hypothesized distribution specified by H_0. The alternative hypothesis is usually very vague - it gives little or no information on the distribution of the data, and simply states that H_0 is false. The majority of the goodness-of-fit techniques proposed in the literature are for univariate data. Methods for multivariate data are much less well developed.

Pearson-type, or chi-square tests, are particularly attractive for categorical data. They are also well known and easy to interpret. If the null hypothesis is rejected, the examination of the contingency table of observed and expected counts can give some information about the nature of the model misspecification.

It is generally difficult to calculate the exact discrete distribution of Pearson-type statistics when H_0 is assumed to be correct. Thus, a continuous distribution is frequently used to approximate the exact null distribution of the test statistic. The accuracy of this approximation depends on the total sample size, the dimension of the contingency table, and the magnitude of the expected counts. Alternatively, simulation techniques can be used to estimate the exact distribution of the statistic under the null hypothesis.

Due to the vagueness of the alternative hypothesis, goodness-of-fit techniques have, in general, low power to detect specific deviations from the hypothesized distribution. Therefore, the modelling process should not stop when a goodness-of-fit test indicates there is no lack of fit. Instead, special techniques should be used to check more thoroughly the adequacy of the model. Some of these techniques are designed to identify systematic departures between the model and the data while others detect isolated deviations. Also, some model-checking techniques are graphical, like residual or probability plots, while others are powerful statistical tests that focus on a specific class of alternatives. Thus, a model that appropriately describes the phenomenon under investigation can be found by successively modifying and checking the model originally proposed.

Several authors have proposed Pearson-type goodness-of-fit tests to examine the adequacy of stationary Markov models for qualitative response variables. All these tests are designed for models with transition rates that do not depend on covariates or explanatory variables.

In this chapter I propose a Pearson-type goodness-of-fit statistic for stationary Markov regression models with a qualitative response variable. The statistic can also be used to examine the fit of non-stationary Markov regression models if the transition probabilities are estimated in an appropriate way. A partition of the covariate space for the Markov regression model is defined in the same way as Hosmer and Lemeshow did for logistic regression models. This means that when the covariates are qualitative, the estimated transition probabilities are grouped into different categories defined by the covariate patterns. Similarly, if some or all the covariates are continuous, the estimated transition probabilities are grouped into equiprobable categories defined by the quantiles of the estimated transition rates. A method has been proposed in the literature to deal with panel data. Here I propose a different method that generalizes the above technique proposed by Hosmer and Lemeshow to response variables measured repeatedly over time.

The exact null distribution of the goodness-of-fit statistic proposed here for Markov regression models is intractable. A parametric bootstrap algorithm is proposed to estimate such distribution. The estimated distribution is then compared to the asymptotic "naive" distribution. This is considered to be a chi-square with degrees of freedom equal to the number of independent cells in the contingency table minus the number of estimated parameters. This comparative study is done for several stationary Markov models - with and without covariates - fitted to different observation patterns. The bootstrap methodology is also used to estimate the power of the proposed statistic to identify Markov models with non-stationary transition rates. Also, the goodness-of-fit procedure proposed in this chapter is applied to examine the adequacy of the Markov model fitted by Gladman, Farewell, and Nadeau [1] to the PsA data.

Sections 2.3, 2.5, 2.6, and 2.9 may be skipped by non-technical readers.

2.2 Goodness of fit statistics for Markov models

Several goodness of fit statistics have been proposed in the literature to determine the adequacy of Markov models. These statistics assume that the number of observations and/or the frequency at which they are obtained is controlled by the researcher and that covariates do not affect the transition rates.

The simplest case is that in which no covariates are measured and every individual has the same number of observations ($m_i = m$) equally spaced in time ($t_{i,j+1} - t_{i,j} = t_{j+1} - t_j$ for all $i = 1, 2, \ldots, n$). This type of longitudinal data can be displayed in a contingency table as explained by Bishop, Fienberg, and Holland [9], chapter 7. The authors show that the analysis of these tables is formally equivalent to certain contingency table analyses based on log-linear models. Without referring to log-linear models, Kalbfleisch and

Lawless [3] also proposed the construction of contingency tables for the observed and expected transitions at each time interval (t_j, t_{j+1}). The authors defined the observed transition counts, $n_{j(a,b)}$, as the total number of subjects in state a at time t_j and in state b at time t_{j+1}. The corresponding expected counts, $e_{j(a,b)}$, are obtained by multiplying the total number of subjects in state a at time t_j, $n_{j(a,.)} = \sum_{k=1}^{K} n_{j(a,k)}$, by the probability of observing a transition from a to b in the time interval (t_j, t_{j+1}), *i.e.* $e_{j(a,b)} = n_{j(a,.)}\hat{p}_{j(a,b)}$. If none of the $\hat{p}_{j(a,b)}$ is restricted to be zero, the likelihood ratio statistic is:

$$\Lambda = 2\sum_{j=1}^{m-1}\sum_{a=1}^{K}\sum_{b=1}^{K} n_{j(a,b)} \log\left(\frac{n_{j(a,b)}}{e_{j(a,b)}}\right)$$

and the Pearson statistic is:

$$X^2 = \sum_{j=1}^{m-1}\sum_{a=1}^{K}\sum_{b=1}^{K} \frac{(n_{j(a,b)} - e_{j(a,b)})^2}{e_{j(a,b)}}$$

The authors state that Λ has an asymptotic (m fixed and $n \to \infty$) chi-square distribution with $(m-1)K(K-1) - \eta$ degrees of freedom; where η is the dimension of the vector that parameterizes the matrix of transition rates.

Stavola [10] also proposed the Pearson statistic mentioned above.

Gentleman *et. al.* [3] generalized the above Pearson statistic for the case in which no covariates are measured and the number and periodicity of the observations varies between individuals. A partition of the time scale was suggested by the authors. Approximate transition or prevalence counts are then calculated at each cutpoint by assuming that individuals not observed at the cutpoints have remained in their last observed state. The observed and expected transition counts are defined as proposed by Kalbfleisch and Lawless [2]. The observed prevalence counts, $n_{j(a,.)}$, are the number of subjects in state a at time j, with $a = 1, 2, \ldots, K$. The expected prevalence counts are defined as the number of subjects being studied at time t_j, $n_j = \sum_{a=1}^{K} n_{j(a,.)}$, multiplied by the estimated probability of observing a transition from state 1 to state a in the time interval $(0, t_j)$ *i.e.* $\hat{p}_{0(1,a)}$, where $a = 1, 2, \ldots, K$. This requires that the time of the disease onset must be known. The authors suggested the use of the Pearson statistic to compare the observed and expected counts. No discussion was given of the distribution of the statistic.

2.3 The estimated transition probabilities

In this section I explain how the distribution of the sojourn times is linked to the transition probabilities for a stationary Markov model with covariates. The sojourn times are defined as the time that an individual spends in the various states of the process. The procedure is illustrated for a progressive model although it is applicable to any kind of stationary Markov model.

In stationary Markov models, the time that an individual spends in each state is governed by an exponential distribution. Let $T_{i(a)}$ denote the time that individual i spends in state a before moving to state $a+1$. The random variable $T_{i(a)}$ has an exponential distribution with parameter $\lambda_{i(a)} = q^{-1}_{i(a,a+1)}$. *i.e.*

$$f_{T_{i(a)}} = \frac{1}{\lambda_{i(a)}} \exp\Big(-\frac{t_{i(a)}}{\lambda_{i(a)}}\Big).$$

The probability that patient i is observed in state $a+1$ at time $t_{i,j+1}$ given that $Y_{i,j} = a$ can be expressed in terms of the sojourn times in states a and $a+1$. The argument is as follows. Given that $Y_{i,j} = a$, individual i can only be observed in state $a+1$ at time $t_{i,j+1}$ if the transition to state $a+1$ occurrs in a time span not greater than $t_{i,j+1} - t_{i,j}$. Before the transition takes place, individual i can remain in state a for a period not greater than $t_{i,j+1} - t_{i,j}$. Once the transition has taken place, individual i remains in state $a+1$ for a period greater than $t_{i,j+1} - t_{i,j} - t_{i(a)}$. Then

$$\begin{aligned} p_{i,j(a,a+1)} &= P[Y_{i,j+1} = a+1 \mid Y_{i,j} = a] \\ &= P[T_{i(a)} < t_{i,j+1} - t_{i,j} \quad \text{and} \quad T_{i(a+1)} > t_{i,j+1} - t_{i,j} - t_{i(a)} \mid Y_{i,j} = a] \end{aligned}$$

The second conditional probability is equal to the unconditional probability:

$$P[T_{i(a)} < t_{i,j+1} - t_{i,j} \quad \text{and} \quad T_{i(a+1)} > t_{i,j+1} - t_{i,j} - t_{i(a)}]$$

because, by the Markov property, the remaining sojourn time in state a is independent of the amount of time that individual i has already spend in state a. The Markov property also implies that $T_{i(a)}$ and $T_{i(a+1)}$ are independent. In other words, the sojourn time in state a is independent of the sojourn time in state $a+1$. Furthermore, $T_{i(a)}$ is a random variable that assumes values between 0 and $t_{i,j+1} - t_{i,j}$ because the transition to state $a+1$ can occur just after the j-th observation is made or just before observation $j+1$ is obtained. Therefore

$$\begin{aligned} p_{i,j(a,a+1)} &= \int_0^{t_{i,j+1}-t_{i,j}} f_{T_{i(a)}} \times \left\{1 - P[T_{i(a+1)} \leq t_{i,j+1} - t_{i,j} - t_{i(a)}]\right\} dt_{i(a)} \\ &= \int_0^{t_{i,j+1}-t_{i,j}} f_{T_{i(a)}} \times \left(1 - \int_0^{t_{i,j+1}-t_{i,j}-t_{i(a)}} f_{T_{i(a+1)}} dt_{i(a+1)}\right) dt_{i(a)} \end{aligned}$$

The explicit solution is:

$$p_{i,j(a,a+1)} = \frac{\lambda_{i(a+1)}}{\lambda_{i(a+1)} - \lambda_{i(a)}} \left[\exp\left(\frac{t_{i,j} - t_{i,j+1}}{\lambda_{i(a+1)}}\right) - \exp\left(\frac{t_{i,j} - t_{i,j+1}}{\lambda_{i(a)}}\right)\right] \quad (2.1)$$

Explicit expressions for $p_{i,j(a,a+2)}, \ldots, p_{i,j(a,K)}$ for all $a = 1, 2, \ldots, K-1$ are calculated in a similar way. As $\sum_{k=a}^{K} p_{i,j(a,k)} = 1$ then $p_{i,j(a,a)} = 1 - \sum_{k=a+1}^{K} p_{i,j(a,k)}$. An estimate of the transition probabilities is obtained by replacing the parameter of the exponential distribution by its maximum likelihood estimate. The formulas to calculate the nine transition probabilities of a progressive Markov regression model with 4 states are included in Appendix A.

A patient who participated in the PsA study was observed in states 2, 3, 3, 3, and 4 at 3.18, 6.14, 6.91, 7.54, and 8.62 years respectively after the disease was diagnosed. The estimated transition rates for this individual, derived from the model fitted by Gladman *et. al.* [1], are: $\hat{q}_{i(1,2)} = 0.21$, $\hat{q}_{i(2,3)} = 0.29$, and $\hat{q}_{i(3,4)} = 0.47$. At the first clinic visit the patient was observed in state 2 so at the next observation time, the individual can remain in state 2 or move to state 3 or 4. The estimated transition probabilities, calculated as explained above, are: $\hat{p}_{i,1(2,2)} = 0.43$, $\hat{p}_{i,1(2,3)} = 0.28$ and $\hat{p}_{i,1(2,4)} = 0.29$.

At the second clinic visit the subject had entered state 3. As the fitted model is stationary, the above transition rates are used again to calculate $\hat{p}_{i,2(3,3)}$ and $\hat{p}_{i,2(3,4)}$ with $t_{i,2} = 6.14$ and $t_{i,3} = 6.91$. In a similar way, the estimated transition probabilities for the third and fourth observed states are obtained. Notice that if m_i observations are made for patient i then m_{i-1} matrices of transition probabilities need to be estimated.

2.4 The goodness of fit statistic

In determining goodness of fit statistics for logistic regression models, the number of distinct values for expectations is given by the total number of covariate patterns in the data. If some covariates are continuous then the number of different covariate patterns becomes approximately equal to the sample size. In this situation, Hosmer and Lemeshow [11] proposed to group the expected values according to the quantiles (*e.g.* deciles) of the estimated probabilities. Each category thus defined is equiprobable and the number of observations in each category increases as the sample size increases. This produces an asymptotic null chi-square distribution for the Pearson type goodness of fit statistics.

In Markov regression models, the estimated transition probabilities depend on the covariate pattern, the sequence of observed states, and the time elapsed between consecutive observations. For non-stationary Markov models, the time at which the observations are made also affects the value of the estimated transition probabilities. When the model is stationary and it is wished to test if the stationarity assumption is valid, the behaviour of the estimated transition probabilities across time should also be examined. All

Table 2.1: Example of a contingency table for the observed and expected counts of a Markov regression model.

Observation	Time		Covariate values		States at times $t_{i,j}$ and $t_{i,j+1}$			
period	t_j	t_{j+1}	X_1 (Temp.)	X_2 (Pressure)	$a \to a$	$a \to b$	$b \to a$	$b \to b$
First ($j = 1$)	0	2	Low	Low				
			Low	High				
			High	Low				
			High	High				
	1	3	Low	Low				
			Low	High				
			High	Low				
			High	High				
Second ($j = 2$)	2	4	Low	Low				
			Low	High				
			High	Low				
			High	High				
	3	5	Low	Low				
			Low	High				
			High	Low				
			High	High				
Third ($j = 3$)	4	6	Low	Low				
			Low	High				
			High	Low				
			High	High				
	5	7	Low	Low				
			Low	High				
			High	Low				
			High	High				

these factors need to be considered in order to group the estimated transition probabilities to calculate a Pearson-type goodness of fit statistic.

In order to illustrate how a contingency table of observed and expected counts might be constructed, consider Table 2.1. It refers to a hypothetical prospective study in which the evolution of a phenomenon was examined through a response variable that assumes two values a and b. Previous studies suggested that the covariates, X_1 and X_2 (*e.g.* temperature and pressure), affect the value of the response variable. At the beginning of the study, the values of the covariates were randomly fixed at a high or low level for each sample unit. Four observations ($m = 4$) were obtained for each sample unit every 2 days (*i.e.* $t_{i,j+1} - t_{i,j} = 2$ days for $j = 1, 2, \ldots, m - 1$). However, it was not possible to observe all the sample units on the same day. Therefore, some of them were assessed on days 0, 2, 4, and 6 while others were observed on days 1, 3, 5, and 7. The data gathered in this way could be summarized in Table 2.1.

Consider now the case in which X_1 (temperature) and X_2 (pressure) are recorded in a continuous scale. In this situation, the quantiles of a linear predictor, such as $\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$, can be used, for each transition rate, to group the observations according to the values of the covariates.

In some studies, the sample units are observed at random times $t_{i,1}$, $t_{i,2}$, $\ldots$, t_{i,m_i}. In this case I propose to classify the first pair of observations, $y_{i,1}$ and $y_{i,2}$, based on the quantiles of the distances $t_{i,2} - t_{i,1}$. Similarly, the second pair of observations, $y_{i,2}$ and $y_{i,3}$, are grouped according to the quantiles of the time spans $t_{i,3} - t_{i,2}$. The same procedure is used to classify the remaining observations.

The procedure just described to construct the contingency table of observed and expected counts is more formally presented below.

In experimental studies, the number of observations and the spacing between them is constant between individuals and fixed in advance. Also, frequently, the covariates are qualitative. For this type of data, a method for the calculation of aggregate summary statistics is the following. Group the estimated transition probabilities according to: the covariate pattern, the time at which the measurements are made, the time elapsed between observations, and the sequence of observed states. This method is appropriate for both stationary and non-stationary Markov models. As recommended by D'Agostino and Stephens [12], chapter 3, the categories defined should be equiprobable and sparse cells should be avoided if the asymptotic distribution of the goodness-of-fit statistic is to be used.

In panel data, the number of observations and the frequency at which they are made vary between individuals. Also, the covariates can be qualitative or quantitative. In order to classify the estimated transition probabilities according to the time and the periodicity at which the observations are made I generalize the grouping technique proposed by Hosmer and Lemeshow [11] as follows. Classify the estimated transition probabilities $\{\hat{p}_{i,j(a,b)}\}$ of the j-th observation period, $(t_{i,j}, t_{i,j+1})$, into L levels defined by the quantiles (*e.g.* deciles or quintiles) of $t_{i,j+1} - t_{i,j}$, where $j = 1, 2, \ldots, \max\{m_{i-1}\}$. This procedure yields different categories for the estimated transition probabilities of each observation period. For example, the PsA data set discussed in chapter 4 yields the following quartiles for the first inter-visit period $\xi^{(1)}_{0.25} = 0.49$ years, $\xi^{(1)}_{0.50} = 0.69$ years, $\xi^{(1)}_{0.75} = 2.15$ years, and $\xi^{(1)}_{1.0} = 9.78$ years. Therefore, if for individual i, the time elapsed between the first and the second clinic visits is 1 year then the first two observed states, $y_{i,1}$ and $y_{i,2}$, and $\hat{p}_{i,1(a,b)}$ are classified in the category defined by $\xi^{(1)}_{0.50} = 0.69$ and $\xi^{(1)}_{0.75} = 2.15$. Analogously, the quartiles obtained for the time elapsed between the second and the third measurements are $\xi^{(2)}_{0.25} = 0.50$ years, $\xi^{(2)}_{0.50} = 0.62$ years, $\xi^{(2)}_{0.75} = 1.16$ years, and $\xi^{(2)}_{1.0} = 9.31$ years. If for individual i, the length of the interval between the 2nd. and the 3th. observations is 0.33 years then $(y_{i,2}, y_{i,3})$ and the estimated transition probabilities $\hat{p}_{i,2(a,b)}$ are assigned to the category with upper bound given by $\xi^{(2)}_{0.25} = 0.50$.

If the set of all possible values for the response variable is $\{1, 2, \ldots, K\}$ then a total of K^2 different transitions between states can be observed. Therefore, the estimated transition probabilities $\{\hat{p}_{i,j(a,b)}\}$ for the j-th time interval can also be classified into one of K^2 different groups according to the values of a and b.

When all the covariates are qualitative the $\{\hat{p}_{i,j(a,b)}\}$ should be grouped according to the covariate pattern associated with individual i. If some covariates are quantitative I propose to use the quantiles of the estimated transition rates, $\{\hat{q}_{i(a,b)}\}$, to define a partition of the covariate space. The transition rates are used here because they do not depend on time.

Thus, I propose to use four classification criteria to group the observed states and the estimated transition probabilities. This can produce sparse cells when the number of observations between individuals is not fairly constant, or when few individuals reach a given state or when some covariate patterns are uncommon in the population. In these situations, some categories might need to be collapsed. For example, if few individuals have more than m' observations, the estimated transition probabilities of time intervals $(t_{i,j}, t_{i,j+1})$, where $j \geq m'$, can be grouped together into L levels defined by the quantiles of $t_{i,j+1} - t_{i,j}$ for $j \geq m'$ and $i = 1, 2, \ldots, n$.

Let H denote the dimension of the contingency table associated with the number of observation periods. Then, $H = m - 1$ if all the individuals have the same number of observations equally spaced on time. If the number of observations varies between individuals then $H = \max\{m_{i-1}\}$, if the $\{\hat{p}_{i,j(a,b)}\}$ are not collapsed across observation periods, and $H = m'$, if the estimated transition probabilities of some observation periods are classified together. If the time at which the observations are made is unimportant, the observation period can be ignored. In this case $H = 1$ and the estimated transition probabilities are classified into L categories defined by the quantiles of the time span between all consecutive observations.

The letter h will denote the h-th category defined on the basis of the observation periods, so $h = 1, 2, \ldots, H$. The estimated transition probabilities of category h are also classified into one of L different levels defined by the quantiles (*e.g.* deciles or quintiles) of the time elapsed between observations. A particular level will be denoted by l with $l = 1, 2, \ldots, L$. Analogously,

R will denote the number of groups in which the transitions between states are classified. Notice that $R = K^2$ if $a, b \in \{1, 2, \ldots, K\}$ and none of the $a \to b$ transitions are grouped together. Letter r will refer to the r-th group with $r = 1, 2, \ldots, R$. Similarly, C will represent the total number of classes in which the covariate patterns are grouped or the total number of levels defined by the quantiles of the estimated transition rates; c will denote the c-th category with $c = 1, 2, \ldots, C$. Therefore, I propose to construct a 4 dimensional contingency table with $H \times L \times R \times C$ cells.

The quantity $\hat{p}_{h,l,r,c}$ will be called the expected number of transitions in cell (h, l, r, c). It is the sum of the estimated transition probabilities between any two states classified in group r of individuals studied in an observation period classified in category h, with an elapsed time between measurements in level l, and a covariate vector grouped in category c. Similarly, $n_{h,l,r,c}$ will denote the total number of observed transitions from a to b classified in group r that belong to individuals observed in a time interval in category h, with a time span between observations classified in level l and covariate vector in category c. The Pearson-type goodness of fit statistic that I propose to examine the adequacy of stationary Markov regression models is:

$$T = \sum_{h=1}^{H} \sum_{l=1}^{L} \sum_{r=1}^{R} \sum_{c=1}^{C} \frac{(n_{h,l,r,c} - \hat{p}_{h,l,r,c})^2}{\hat{p}_{h,l,r,c}} \qquad (2.2)$$

This statistic can be generalized to test the goodness-of-fit of a non-stationary Markov model if the transition probabilities are appropriately estimated.

2.5 Distribution of the statistic

Given that individual i is in state a at time $t_{i,j}$, the probabilities of observing transitions to the states $a, a+1, a+2, \ldots, K$ in the time interval $t_{i,j+1} - t_{i,j}$ follow a multinomial distribution with parameters 1 and $\theta_{ij} = (\hat{p}_{i,j(a,a)}, \hat{p}_{i,j(a,a+1)}, \ldots, \hat{p}_{i,j(a,K)})$. (A similar distribution is obtained for non-progressive Markov models). Individual i will, in general, have a different multinomial distribution at time $t_{i,j+1}$ because the vector of parameters

$\theta_{i,j+1}$ depends on the time elapsed between $y_{i,j+1}$ and $y_{i,j+2}$. As θ_{ij} also depends on the explanatory variables in the model, different subjects also have different multinomial distributions. Therefore, if the total number of observation periods is taken as fixed, $P[N_{1111} = n_{1111}, \ldots, N_{HLRC} = n_{HLRC}]$ is the sum of several independent and non-identical multinomial distributions. As the total number of observation periods is not fixed since entry into an absorbing state terminates observation of one subject, the exact distribution of the proposed test statistic is particularly intractable.

A method of estimating the distribution of (2.2) is by generating B independent bootstrap samples from the model specified by the null hypothesis, and calculating the goodness of fit statistic for each sample. As B goes to infinity, the bootstrap distribution of (2.2) will approach the true null distribution of the test statistic [13].

Here the bootstrap distribution is compared to the "naive" asymptotic distribution of T. This "naive" distribution is assumed to be a chi-square with degrees of freedom equal to the number of independent cells in the contingency table minus the number of estimated parameters.

2.6 The hypothesis testing procedure

A parametric bootstrap procedure can be carried out to calculate the significance level by generating data from an estimated Markov model specified by the null hypothesis. If this hypothesis states that a stationary Markov regression model of order one fits the data then the time elapsed between transitions follow exponential distributions with parameters $\lambda_{i(a)} = q^{-1}_{i(a,a+1)}$ for $a = 1, \ldots, K - 1$.

The bootstrap states for individual i will be denoted as $Y^*_{i1}, Y^*_{i2}, \ldots$ where $Y^*_{i1} = Y_{i1}$ for $i = 1, 2, \ldots, n$. Thus the first bootstrap state is taken to be the first observed state. Assume that $Y^*_{i1} = a$. The rest of the Y^*_{ij} are obtained by simulating the times at which individual i enters states $a + 1, \ldots, K$ (if the model is progressive). Let $t^*_{i(k)}$ be an observation simulated from the exponential distribution with parameter $\hat{\lambda}_{i(k)}$ for $k = a, \ldots, K-1$; $t^*_{i(k)}$ represents the time that individual i remains in state k before moving to state $k+1$. Therefore $t^*_{i(a)}, \sum_{r=a}^{a+1} t^*_{i(r)}, \ldots, \sum_{r=a}^{K-1} t^*_{i(r)}$ are the simulated times at which individual i enters state $a+1, a+2, \ldots, K$ respectively. The bootstrap state Y^*_{ij} is then obtained as follows:

$$\text{If} \quad t_{i,j} < t^*_{i(a)} \quad \text{then} \quad Y^*_{i,j} = a, \quad \text{otherwise}$$

$$\text{if} \quad \sum_{r=a}^{k} t^*_{i(r)} \leq t_{i,j} < \sum_{r=a}^{k+1} t^*_{i(r)} \quad \text{then} \quad Y^*_{i,j} = k+1 \quad \text{for} \quad k = a, \ldots, K-2,$$

$$\text{otherwise if} \quad t_{i,j} \geq \sum_{r=a}^{K-1} t^*_{i(r)} \quad \text{then} \quad Y^*_{i,j} = K \tag{2.3}$$

The above inequalities state that $Y^*_{i,j}$ is equal to a if $t_{i,j}$ is less than the simulated time at which the transition to state $a+1$ occurrs. Similarly, individual i remains in state $a+1$ until $t_{i,j}$ is greater than or equal to the simulated time at which the transition to state $a+2$ occurrs, etc. If the transition to the absorbing state takes place at a simulated time which is less than t_{i,m_i} then several bootstrap states are equal to K. Whenever $Y^*_{i,s_i} = Y^*_{i,s_{i+1}} = \ldots = Y^*_{i,m_i} = K$ the last $m_i - s_i$ states are ignored. Thus the total number of bootstrap states for individual i will be denoted as s_i for $i = 1, 2, \ldots, n \quad (s_i \leq m_i)$.

Once a sequence of states is generated for each individual, the Markov model is estimated based on the bootstrap data and the test statistic is calculated. This process is repeated several times. Finally, the value of the statistic from the original data is compared with the bootstrap distribution of values to compute the significance level.

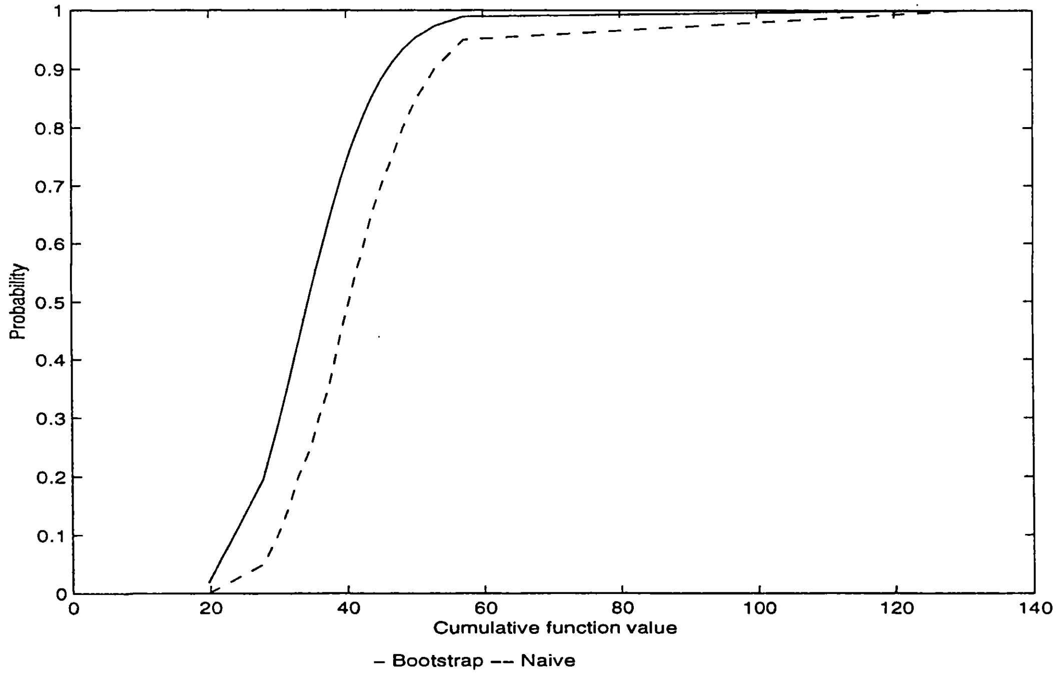
2.7 Results for the psoriatic arthritis data

The model proposed by Gladman *et. al.* [1] for the PsA data has 6 binary covariates that produced 30 different covariate patterns. (The maximum number of different covariate patterns is $2^4 \times 3 = 48$ because the two

Table 2.2: Contingency table for the observed and expected counts for the Markov regression model fitted by Gladman *et. al.*

			Transition					
Time	P. Factor		$1 \to 1$	$1 \to \bar{1}$	$2 \to 2$	$2 \to \bar{2}$	$3 \to 3$	$3 \to \bar{3}$
	Zero	Obs.	57	0	19	0	17	1
0.0384		Exp.	55.32	1.68	18.11	0.89	16.75	1.25
	One	Obs.	51	2	38	**7**	16	2
to		Exp.	51.14	1.86	42.59	**2.41**	16.45	1.55
	Two +	Obs.	20	**3**	18	1	6	1
0.4791		Exp.	22.01	**0.99**	17.69	1.31	6.21	0.79
	Zero	Obs.	50	3	18	1	6	1
0.4791		Exp.	50.87	2.13	17.55	1.45	6.37	0.63
	One	Obs.	53	5	31	**7**	21	3
to		Exp.	55.09	2.91	35.32	**2.68**	21.68	2.32
	Two +	Obs.	16	1	19	3	6	2
0.5394		Exp.	16.04	0.96	19.97	2.03	7	1
	Zero	Obs.	47	**6**	24	1	11	1
0.5394		Exp.	50.42	**2.58**	22.94	2.06	10.69	1.31
	One	Obs.	51	**8**	42	2	23	2
to		Exp.	55.95	**3.05**	39.86	4.14	21.76	3.24
	Two +	Obs.	15	1	14	2	5	2
0.7474		Exp.	14.72	1.28	14.09	1.91	5.50	1.50
	Zero	Obs.	57	3	21	2	1	**2**
0.7474		Exp.	55.39	4.61	20.32	2.68	2.50	**0.51**
	One	Obs.	51	6	27	2	25	**1**
to		Exp.	52.01	4.99	24.59	4.41	20.51	**5.49**
	Two +	Obs.	15	2	15	3	11	4
1.4012		Exp.	15.70	1.30	14.27	3.73	10.78	4.22
	Zero	Obs.	38	13	21	**3**	4	2
1.4012		Exp.	39.16	11.84	16.17	**7.83**	3.80	2.20
	One	Obs.	67	**12**	25	9	14	5
to		Exp.	57.88	**21.12**	22.49	11.51	11.43	7.57
	Two +	Obs.	13	5	10	6	2	4
15.3484		Exp.	12.21	5.79	9.40	6.60	2.33	3.67
		Total Obs.	601	70	342	49	168	33
		Total Exp.	603.89	67.11	335.36	55.64	163.76	37.24

Figure 2.1: Bootstrap and naive distribution functions for the Markov regression model fitted by Gladman *et. al.*



covariates for the initial state can not both be equal to one). Sixty patients with a total of 330 transitions had a zero covariate vector. The least common covariate vector corresponds to a patient with only two observed states. Due to the discrepancy between the frequencies of the covariate patterns, the partition of the covariate space was done using the number of prognostic factors coded as one. The prognostic factors are: number of effused joints, ESR, and type of medication taken before participating in the study. Patients were classified into $C = 3$ categories based on these prognostic factors. Level one refers to subjects with all prognostic factors coded as zero, level two corresponds to patients with only one prognostic factor coded as one, and level three contains individuals with two or more prognostic factors coded as one.

Gladman, Farewell and Nadeau [1] tested the appropriateness of the assumption that the transition rates are stationary. They allowed the transition probabilities to depend on a power of time but did not find evidence for this dependence. Therefore, I decided to group together observations obtained at different time intervals so $H = 1$.

Even though clinic visits were planned every 6 months, the mean time between assessments is 1.18 years with a standard deviation of 1.42 years and a median of 0.61 years. The time elapsed between observations was categorized into $L = 5$ levels defined by the quintiles of $t_{i,j+1} - t_{i,j}$ for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m_i$.

Nine types of transitions were observed; 47.60% are of the form $1 \rightarrow 1$ but only 0.55% are of the type $1 \rightarrow 4$. Cross-classifying the 9 types of transitions by the $H \times L \times C = 1 \times 5 \times 3 = 15$ categories previously defined produces a table with sparse cells. Therefore, transitions of the form $a \rightarrow b$ with $b > a$ were grouped together in a category denoted as $a \rightarrow \bar{a}$. Consequently. $R = 6$ types of transitions were considered: $a \rightarrow a$ and $a \rightarrow \bar{a}$ with $a = 1, \ldots. K - 1$.

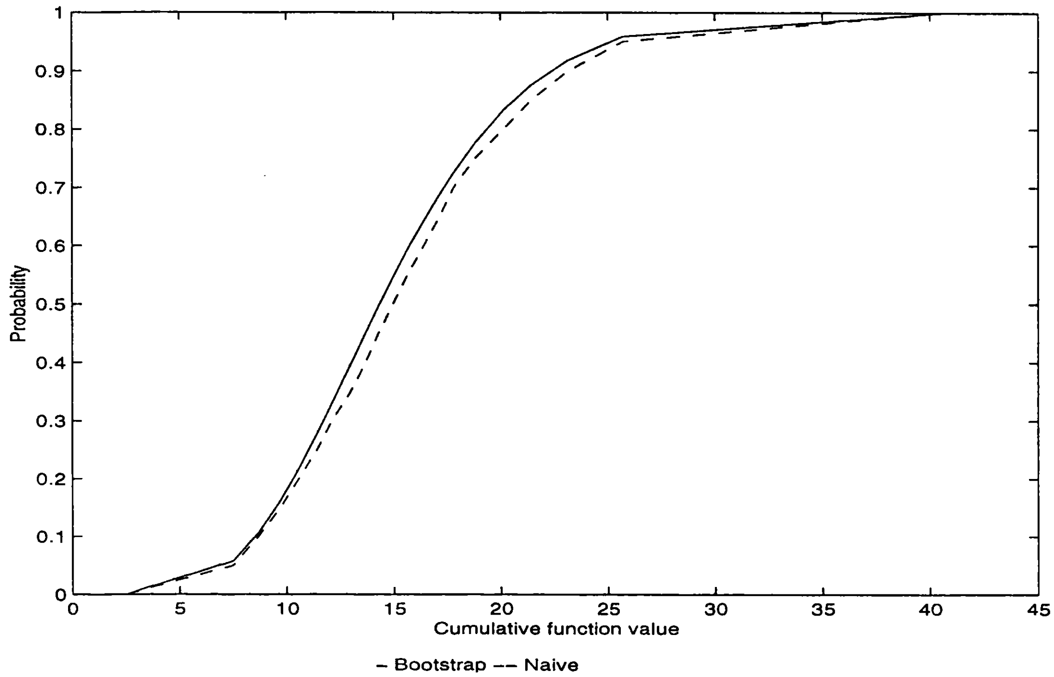
The observed and expected transitions for the PsA data are shown in Table 2.2. The first two rows contain the observed and expected transitions of those patients with all the prognostic factors coded as zero and with an elapsed time between visits less than or equal to 0.4791 years. Analogously, the observed and expected transitions in the 3th and 4th rows correspond to subjects with only one prognostic factor coded as one and with $t_{i,j+1} - t_{i,j} \leq 0.4791$, etc. As ties occurred at some quintiles, the $L = 5$ groups defined for the time elapsed between observations do not have the same number of transitions. The nine cells with bold numbers contribute 67.7% to the value of the goodness of fit statistic: $T = 69.95$.

One thousand bootstrap replications were carried out. In 44 of them the bootstrap algorithm did not generate any observations from state 3 so the associated contingency tables contain some empty cells. These cells were ignored in the calculation of the bootstrap goodness of fit statistic. The p-value thus obtained is $9/1000 = 0.009$ indicating that the fitted stationary Markov regression model does not describe adequately the PsA data. Note that all the cells with bold numbers occur in the columns labeled as $a \rightarrow \bar{a}$ with $a = 1, 2, 3$. In six of these cells the observed count is bigger than the expected count. These six cells contain patients with an elapsed time between clinic visits less than or equal to 1.4 years. Therefore, it may be that patients who experienced a rapid progression in damage were prompted to visit the clinician in a shorter time interval than reflected by the majority of patients. Thus, the non-random distribution of the bold numbers indicates that some clinic visits did not occur at random times.

Furthermore, two outliers were detected in the contingency table formed by using the 9 original transitions instead of the 6 collapsed categories. The outliers are patients who had a 1 → 4 transition in less than 2 years and thus make a large contribution to the value of the corresponding goodness of fit statistic.

Table 2.2 has 3 independent columns because the number of patients in state a at the beginning of every observation period is known so the expected number of transitions from a to $\bar{a}$ depends on the expected number of transitions within state a, for $a = 1, 2, 3$. Then, the number of independent cells is 45 ($= H \times L \times (R - 3) \times C$). Furthermore, Table 1.6 shows that $\eta = 10$

Figure 2.2: Bootstrap and naive distribution functions for model 1.



parameters were estimated to fit the Markov regression model.

Naively, the expected degrees of freedom would be 35 ($= H \times L \times (R - 3) \times C - \eta$), which would correspond to the mean of the distribution of the statistic if it was chi-square. The mean value of the bootstrap goodness of fit statistic was 41.12.

Figure 2.1 shows the cumulative probability function of the statistic calculated from the 1000 bootstrap data sets and the "naive" asymptotic chi-square distribution.

2.8 Other bootstrap analysis

To better understand the distribution of the proposed statistic, the bootstrap algorithm was implemented for different underlying models and observation patterns. The psoriatic arthritis data were used as a basis for the study.

The first scenario considered, termed model 1, was that of regularly

Figure 2.3: Bootstrap and naive distribution functions for model 2.

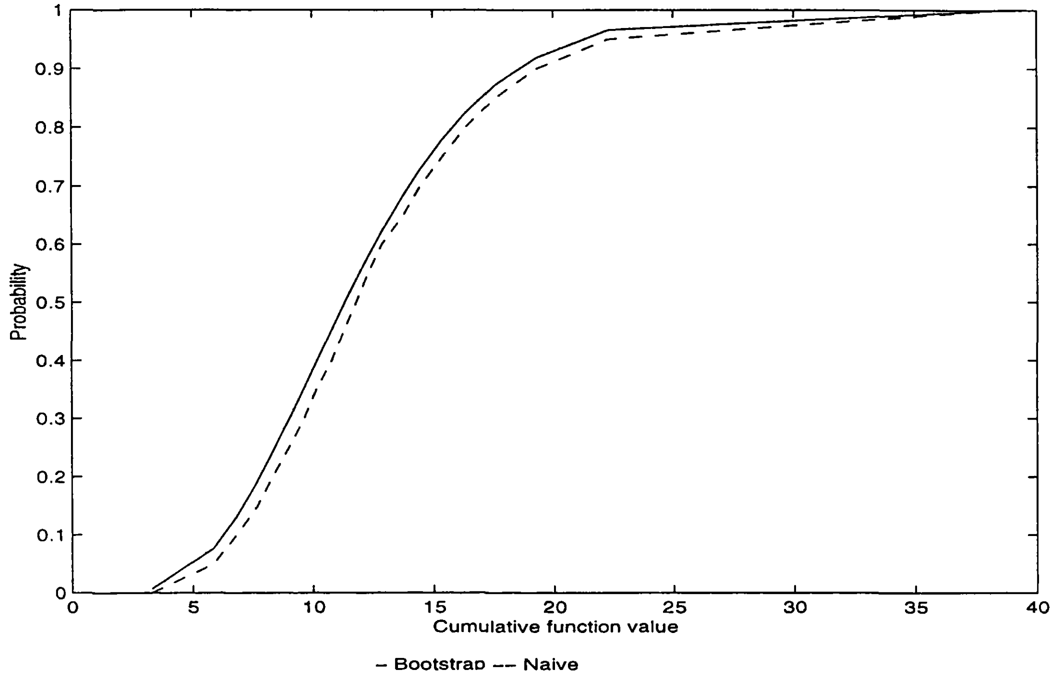


Figure 2.4: Bootstrap and naive distribution functions for model 3.

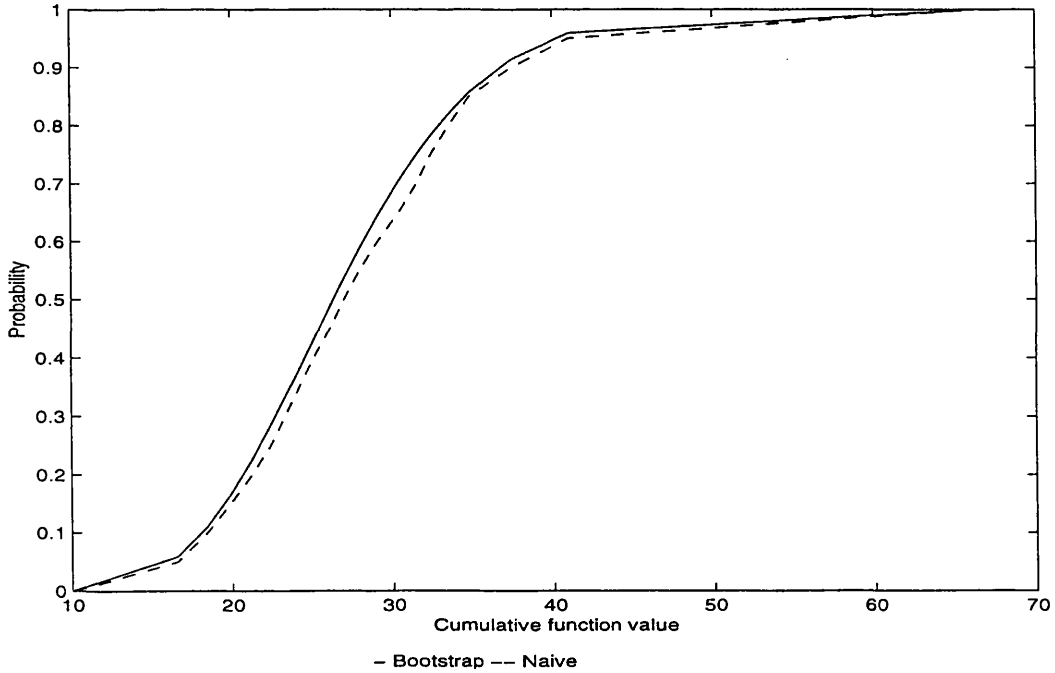


Figure 2.5: Bootstrap and naive distribution functions for model 4.

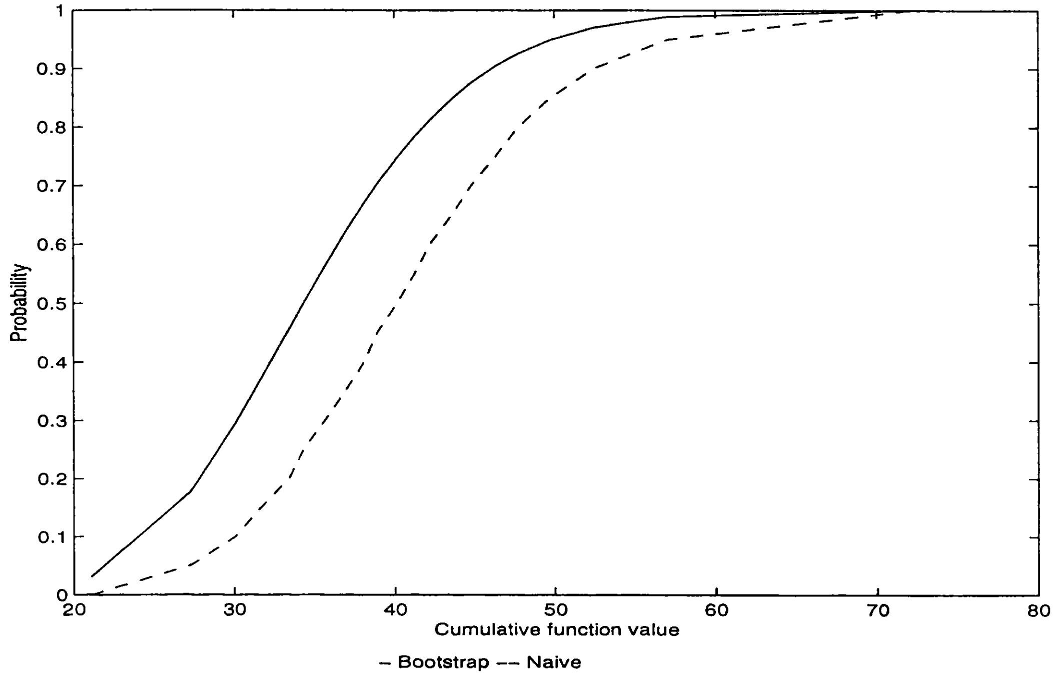
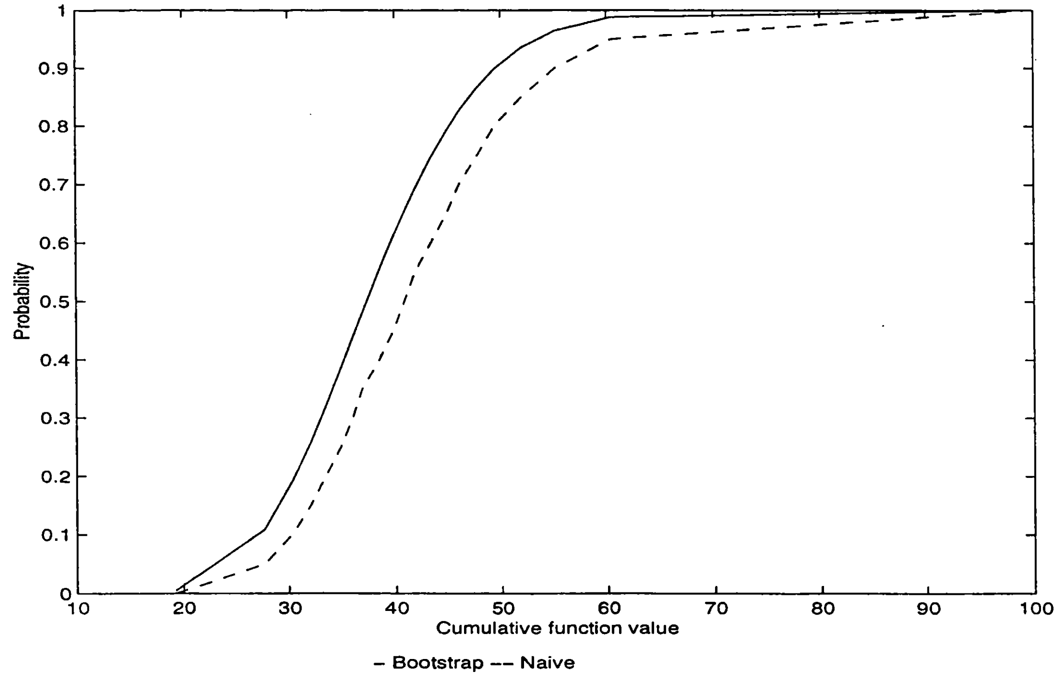


Figure 2.6: Bootstrap and naive distribution functions for model 5.



spaced observations 0.60 years apart and a Markov model, not involving explanatory variables, with transition rates: $q_{i(1,2)} = 0.184$, $q_{i(2,3)} = 0.303$, and $q_{i(3,4)} = 0.431$. Seven states were simulated for each individual unless they first reached the absorbing state. The initial states were taken to be as observed in the psoriatic arthritis study. As $t_{i,j+1} - t_{i,j} = 0.6$ for all i, j the contingency table can not be constructed using the quantiles of the time elapsed between observations. Therefore, the sequence of observed states were classified according to the observation period (1st. observation period, 2nd. observation period, etc.) and according to their type: $a \rightarrow a$ (transitions to the same state) or $a \rightarrow \bar{a}$ (transitions to a different state) with $a = 1, 2, 3$. Both classification criteria have 6 levels. The "naive" degrees of freedom for the proposed statistic are $6 \times 3 - 3 = 15$ which is well in accordance with 15.53, the mean of the 1,000 bootstrap statistics. This is the situation considered by Kalbfleisch and Lawless [2] except for the fact that the number of transitions decreases as the number of observation periods increases because some individuals reach the absorbing state before the seventh bootstrap state is obtained. According to these authors, the statistic should have an asymptotic chi-squared distribution. The cumulative probability function of the goodness-of-fit statistic calculated from the bootstrap data and the "naive" distribution function for model 1 are plotted in Figure 2.2.

In model 2, the original m_i states for patient i were used under the assumption that they were equally spaced every 0.6 years. For the PsA data, if no covariates are included in the model, the estimated transition rates are: $\hat{q}_{i(1,2)} = 0.184$, $\hat{q}_{i(2,3)} = 0.303$, and $\hat{q}_{i(3,4)} = 0.431$. The observed states were again classified according to transition type ($a \rightarrow a$ or $a \rightarrow \bar{a}$) and observation period as follows. The 1st. pair of observed states were categorized in one class, the 2nd. ones in another, the 3th and the 4th were grouped in a third category, the 5th and the 6th were collapsed in a fourth class and the remaining ones were classified together in a fifth category. One thousand bootstrap data sets were simulated from model 2. The mean of the bootstrap goodness of fit statistic is 12.64 while the degrees of freedom of the "naive" asymptotic chi-square distribution are $12 = 5 \times 3 - 3$. The simulated distribution function of the proposed test statistic and the "naive" distribution function are shown in Figure 2.3.

Model 3 is based on the original PsA data in which there are m_i unequally spaced states for patient i. No covariates are included in the model and the estimated transition rates are: $\hat{q}_{i(1,2)} = 0.091$, $\hat{q}_{i(2,3)} = 0.161$, and $\hat{q}_{i(3,4)} = 0.249$. A contingency table of two dimensions was constructed using the deciles of the time elapsed between all the measurements and the transition type ($a \rightarrow a$ or $a \rightarrow \bar{a}$). Again, one thousand bootstrap data sets were generated from the fitted model. This produced a mean of 27.73 for the boostrap goodness of fit statistic. The mean of the "naive" chi-square distribution is $27 = 10 \times 3 - 3$. The distribution function of the test statistic estimated from the bootstrap data and the "naive" distribution function are plotted in Figure 2.4.

Model 4 is similar to model 2 except for the transition rates that depend on the covariates as proposed by Gladman *et. al.* [1]. A three way contingency table was constructed using the transition type, the observation period and the number of prognostic factors coded as one. The transition type and the observation period were categorized as done for model 2. The partition of the covariate space was done in the same way as for the original PsA data. One thousand bootstrap samples were generated but 177 of the resulting contingency tables had empty cells. These cells were eliminated to calculate the bootstrap goodness of fit statistic. Thus, the average of the 1000 bootstrap statistics is 40.82. This value is bigger than $35 = 5 \times 3 \times 3 - 10$, the degrees of freedom of the "naive" chi-square distribution. Figure 2.5 shows the distribution function of the statistic as estimated by the bootstrap replications and the "naive" chi-square approximation.

In the last scenario, termed model 5, a stationary Markov regression model was fitted to the original data. The patients were not stratified by their first observed state and it was assumed that the rest of the covariates had the same effect on all the transition rates. The contingency tables were constructed in the same way as for the model proposed by Gladman *et. al.* Thirty six out of 1000 bootstrap data sets produced a contingency table with empty cells. These cells were ignored in the calculation of the bootstrap test statistic. The mean of the bootstrap statistics is 41.92 and the mean of the "naive" chi-square distribution is $38 = 5 \times 3 \times 3 - 7$. The distribution function of the statistic calculated from the bootstrap data and the "naive" chi-square distribution are shown in Figure 2.6.

2.9 An illustrative power calculation

In this section I examine the power of the test statistic (2.2) in a particular situation. Longitudinal data sets were simulated with time elapsed between transitions having a Weibull distribution. The null hypothesis was taken to be the Markov stationary model described in scenario 5. The non-stationary longitudinal data was generated using the following result. If $T_{i(a)}$ has an exponential distribution with parameter $\lambda_{i(a)}$ and $\alpha > 0$, then $W_{i(a)} = T_{i(a)}^{1/\alpha}$ has a Weibull distribution with parameters α and $\gamma_{i(a)} = \lambda_{i(a)}^{1/\alpha}$. A proof of this well known result is presented in Appendix B for completeness. Therefore, with a few modifications, the procedure described in Section 2.6 to generate the bootstrap states can also be applied to simulate non-stationary sequences of data. For example, if $\alpha = 2$ then, in equations (2.3), $t_{i,j}$ and $t_{i(a)}^*$ must be replaced by $\sqrt{t_{i,j}}$ and $w_{i(a)} = \sqrt{t_{i(a)}^*}$ respectively.

The nested bootstrap algorithm as described by Shao and Tu [14], chapter 4, was used to calculate the power of the test statistic. A total of $B_1 = 200$ independent longitudinal samples were generated with elapsed time between transitions following a Weibull distribution. The stationary Markov regression model described in scenario 5 was fitted to each sample. The contingency table was also constructed in the same way as for model 5. For each of the 200 data sets, a total of $B_2 = 500$ independent bootstrap sub-samples were generated from the model specified by the null hypothesis. Model 5 was fitted to each of these B_2 sub-samples and the test statistic was computed. The p-value of each of the $B_1 = 200$ models was obtained by comparing their goodness of fit statistic with the respective $B_2 = 500$ statistics. For 25% of the B_1 samples, the scoring algorithm did not converge for up to three bootstrap sub-samples. The significance level was calculated using the remaining B_2' data sets. Also, for each of the B_1 hypothesis tests, an average of 5.25 sub-samples produced a contingency table with empty cells. These cells were ignored in the calculation of the test statistic. If the type I error is set at 5%, 107 out of 200 stationary models were rejected, so the power at the 5% level is 54%. The power of the proposed test at the 1% critical level is 26% ($= 100\% \times 52/200$).

2.10 Conclusions

A contingency table needs to be constructed to calculate the proposed goodness of fit statistic. The classification criteria that need to be considered are: (1) the transition type, (2) the covariate pattern, (3) the time elapsed between observations, and (4) the observation period (or the time at which the measurements are made). It will only be possible to use them all when the total number of observed states is large or when each classification criterion has a small number of levels. The later situation arises in experimental studies where the researcher assigns the treatments and decides when and how many observations each sampling unit will provide. In observational studies, where panel data are common, some experience and insight are needed to define the contingency table.

Classification criterion 4 will automatically induce criterion 3 when all the subjects have the same number of equally spaced observations. The fourth criterion is of paramount importance if the fitted Markov model is nonstationary or if interest lies in testing whether the model is stationary or not. When the time elapsed between measurements is not constant, we propose to classify the observed transitions and the estimated transition probabilities using the quantiles of $t_{i,j+1} - t_{i,j}$. This is justified by the fact that two individuals with the same covariate pattern having a transition from state a to state b also have similar transition probabilities if $t_{i,j+1} - t_{i,j} \sim t_{i',j+1} - t_{i',j}$. Analogously, if the covariates are continuous we propose to use the deciles of the estimated transition rates, $\{\hat{q}_{i(a,b)}\}$, to generate a partition of the covariate space.

The theoretical distribution of the proposed goodness of fit statistic is intractable when covariates are used to model the transition rates and the measurement times are not fixed in advance . In this paper, the bootstrap methodology was applied to estimate the distribution of the test statistic under the null hypothesis. Several stationary Markov models were considered, some with covariates and others without covariates.

In all the scenarios considered, the mean value of the bootstrap goodness of fit statistic is bigger than the "naive" degrees of freedom given by the number of independent cells in the table minus the number of estimated parameters. For the models without covariates, the difference between this mean and the "naive" degrees of freedom is less than one unit but for the Markov regression models the discrepancy is bigger. Also, for the models without covariates, the bootstrap distribution of the proposed goodness of fit statistic is well approximated by a chi-square distribution with the "naive" degrees of freedom.

The number of observed states varied greatly between the patients who participated in the PsA study. A median of 4 states were observed for the individuals who participated in the study. Only 2 states were recorded for 21% of the patients while 24 states were recorded for a single subject. The 271 patients analysed gave a total of 1236 transitions.

The number of bootstrap states, s_i generated by the algorithm proposed here is smaller or equal than the number of observed states m_i. The longitudinal series produced by the bootstrap algorithm is smaller when the absorbing state is simulated in less than m_{i-1} observation periods. Sometimes, the observed and simulated series have the same length but the absorbing state is only reached in the observed data. This means that the rate at which the patients progress to the absorbing state is not always the same in the PsA data and in the bootstrap data. In the nested bootstrap analysis carried out for the power calculations, the longitudinal series of the sub-samples become even shorter for the same reason.

The parametric bootstrap algorithm that I propose does not guarantee that the proportion of individuals that reach the absorbing state is the same in the PsA data and in the bootstrap sample. The boostrap algorithm was modified so that additional states were generated for those individuals who, in the original data, reached the final state provided that their observation period does not exceed the mean observation period of the subjects who did not reach this absorbing state. The p-value obtained for the model fitted by Gladman *et. al.* is similar to the value mentioned in section 2.7. This may be due to the fact that only 20.3% of the patients reached state 4. With a higher rate of progression to the absorbing state a different result might be obtained.

Some contingency tables constructed from the simulated data to analyse the fit of Markov models with transition rates that depend on covariates had empty cells. This is a consequence of the difference in length between the bootstrap and the observed series and the number of classification criteria used to construct the contingency table. I computed the bootstrap test statistic by ignoring the empty cells instead of collapsing adjacent categories.

Chapter 3

Mixture and contagious models

3.1 Introduction

Table 2.2 shows that 601 (47.59%) transitions in the PsA study discussed in chapter 2 are of the form 1 → 1. In fact, 116 (42.8 %) patients with a total of 463 (36.66 %) transitions did not develop damaged joints during the course of the study. Thus arises the question of whether a subpopulation of individuals with PsA never develops damaged joints. If this hypothesis is clinically true, a mixture model for repeated observations can be used to describe the PsA data. Mixture models assume that patients who do not develop damaged joints have a different behaviour (distribution) from the rest of the population. These models can be fitted without knowledge of the group to which each individual belongs and this gives an estimate of the proportion of subjects in each group.

Another way of motivating the use of mixture models is the following. Patients remaining in state 1 for a long time and patients not susceptible to damaged joints produce a large number of transitions from state 1 to state 1. This implies that an increase of 0 damaged joints between visits is frequently recorded. The increase in the number of damaged joints in a given time interval is a discrete variable. Some discrete distributions, like the negative binomial, can be used to describe data sets with a large proportion of zeros.

This is not the case for the Poisson distribution. The excess of zeros in the data relative to the Poisson, or any other distribution, can be modelled by an additional parameter. In this way, a distribution with added zeros is obtained. These distributions are special kinds of mixture models. The additional parameter is interpreted as the proportion of zeros not accounted for by the original distribution or as the proportion of individuals that can not experience the event of interest (damaged joints) if the population is formed by two subgroups.

In this chapter I investigate the use of mixture regression models and negative binomial regression models for longitudinal (panel) data. The correlation between a series of observations is modelled in a similar way as for the Markov regression model. The models are used to describe a PsA data set which is larger than the one analysed by Gladman, Farewell, and Nadeau [1]. The new data is based on a longer observation period in which more patients joined the PsA study and the ones already participating on it continued to be assessed. Also, observations corresponding to 10 or more damaged joints were incorporated into the analysis. As a result, 285 patients with a total of 1875 transitions are analysed here.

The model proposed by Gladman, Farewell and Nadeau [1] has the characteristic that, not only the response variable, but also the covariates were categorized. This can lead to some loss of information. Therefore, in this chapter, I avoid the categorization of explanatory variables and examine the fit of models for discrete response variables measured repeatedly over time.

The response variable studied here is the increase in the number of damaged joints from one clinic visit to the next. The correlation between consecutive response variables is modelled by considering, as an extra covariate, the total number of damaged joints recorded up to the last assessment. Thus, the average increase in the number of damaged joints between visits depends on: a baseline value, the number of damaged joints recorded up to the last assessment, the type of medication taken before participating in the study and the erythrocyte sedimentation rate and the number of effused joints observed on the first clinic visit.

It was thought that the age of the patient at the time of the disease onset was the only covariate in the data that could be related to the additional parameter - representing the proportion of individuals never developing damaged joints or the proportion of zeros not explained by a given distribution. The logit of this parameter is modelled as a linear function of the patient's age when he/she began suffering from PsA.

Three models for longitudinal data are examined in this chapter: a Poisson regression model with added zeros for repeated observations, a negative binomial regression model for repeated observations and a negative binomial regression model with added zeros for repeated observations. The maximum likelihood estimates of the parameters are computed using the quasi-Newton algorithm implemented in the NAG subroutine e04jaf. An estimate of the asymptotic covariance matrix of the parameter estimates is obtained by calculating the inverse of the observed information matrix. This matrix is computed using the NAG subroutine e04xaf. When possible, for the models with added zeros, a statistical hypothesis test is carried out to determine if the additional parameter is significantly greater than zero. The total number of observed increments equal to $0, 1, 2, \ldots, 8, 9$, and 10 or more damaged joints is compared to the corresponding total predicted by each model. A measure summarizing the discrepancies between the observed and expected counts is also calculated.

Readers wishing to avoid the technical material may skip sections 3.2, 3.3.1, 3.3.2, 3.3.3, 3.4.1, 3.4.2, 3.5.1, 3.5.2.

3.2 Statistical background

Some probability distributions can be regarded as a combination or overlapping of other distributions. Usually, the resulting distribution is more complex than the original ones and is known as a contagious or mixture distribution. More formally, let $h(x \mid \boldsymbol{\theta})$ be a conditional density function that depends on the vector of parameters $\boldsymbol{\theta}$. Suppose also that $\boldsymbol{\theta} \in \Re^m$ is subject to random variation according to the probability law $q(\boldsymbol{\theta})$. Then the contagious density function $f(\boldsymbol{x})$ is defined as:

$$f(\boldsymbol{x}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\boldsymbol{x} \mid \boldsymbol{\theta}) q(\boldsymbol{\theta}) d\theta_1 \ldots d\theta_m. \tag{3.1}$$

When $\boldsymbol{\theta}$ assumes a finite number of values: $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G$, each with probability $p_g = q(\boldsymbol{\theta}_g)$, where $\sum_{g=1}^{G} p_g = 1$, then $q(\boldsymbol{\theta})$ has a discrete multivariate distribution and

$$f(\boldsymbol{x}) = \sum_{g=1}^{G} p_g h(\boldsymbol{x} \mid \boldsymbol{\theta}_g) \tag{3.2}$$

is called a finite mixture or compound distribution. Johnson and Kotz, [15] chapter 8, describe several density functions of the form (3.1) and (3.2).

Sometimes, contagious distributions are used to describe heterogeneous populations. For example, in accident proneness, Greenwood and Yule [16] derived the negative binomial distribution by assuming that the number of accidents per individual follows a Poisson distribution with parameter θ. The authors assumed that θ varies from individual to individual according to a gamma distribution. A negative binomial regression model is obtained when θ depends on several explanatory variables. For this model, Lawless [17] compared the efficiency and robustness properties of maximum likelihood estimators with those of weighted least-squares along with moment estimation of the dispersion parameter.

Finite mixture distributions are applied when a population is formed by G distinct subpopulations. Sometimes it is known to which subpopulation each individual belongs so the primary aim is to estimate the mixing proportions $p_1, \ldots, p_G$ in (3.2). In other situations it is impossible to observe the variable(s) that split the individuals into different groups. This means that there is no available information for each conditional distribution separately but only for the combined mixture distribution. Several examples are given by Everitt and Hand [18]. In this situation, the objective is to estimate both the mixing proportions and the parameters of the conditional distributions in (3.2). Usually, in this context, G is fixed by the theoretical background of the problem under investigation. In fact, some authors, like Everitt and Hand [18] and Farewell [19], state that this type of mixture distributions should only be used when there is strong scientific evidence for the existence of two or more subpopulations. The reason is that interpretation problems may arise since a mixture distribution can always be fitted to the data by choosing a sufficiently large number of groups, G.

In spite of this, mixtures of distributions are frequently used in cluster analysis where there is no *a priori* knowledge about any grouping structure in the population. The aim is to model heterogeneous data and to obtain some insight into the problem by the formation of several clusters. An example of this kind of approach is given by McLachlan and Basford [20].

When the population is divided into $G = 2$ groups, expression (3.2) becomes:

$$f(x) = p_1 h(x \mid \theta_1) + (1 - p_1) h(x \mid \theta_2)$$

Here, X can be viewed as depending on a binary variable V that is equal to one with probability p_1 and equal to zero with probability $(1 - p_1)$. In other words, $q(\theta)$ has a Bernoulli distribution with parameter p_1. If covariates are available, their effect on p_1 can be assessed by fitting a logistic model. Farewell [21] and Struthers and Farewell [22] applied this model to time to event data. In these two articles, p_1 represents the proportion of individuals that experience the event of interest (*e.g.* AIDS or relapse of a disease) and X is the time until the event occurs. It is assumed that the conditional distribution of X given $V = 1$ follows an exponential or Weibull distribution.

Consider now the situation in which X represents the number of events that occur in a specified period of time. The Poisson distribution is a natural choice for X. However, count data may have an excess of zeros as compared with a Poisson distribution. This is one example of the phenomenon known as overdispersion. It arises when a proportion, p_1, of individuals can not experience the event of interest. Their zero count is a structural zero. Other individuals have a zero count by chance; these are sampling zeros. Several methods for modelling such overdispersed data have been proposed in the literature. The use of mixture distributions is one of them. In the situation just described, X is a discrete variable and the population is again divided into $G = 2$ groups: individuals that can experience the event of interest, $V = 0$, and those who can not, $V = 1$. Therefore, expression (3.2) can be rewritten more conveniently as:

$$P(X = x) = p_1 P(X = x \mid V = 1) + (1 - p_1) P(X = x \mid V = 0)$$

or equivalently:

$$\begin{aligned} P(X = 0) &= p_1 + (1 - p_1) P(X = 0 \mid V = 0) \\ P(X = x) &= (1 - p_1) P(X = x \mid V = 0) \quad \text{if} \quad x = 1, 2, \ldots \end{aligned}$$

This particular type of mixture distribution is known as a Poisson distribution with added zeros or as a zero-inflated Poisson (ZIP) distribution because the proportion of zeros has been increased by a constant p_1. If measured, covariates can be used to model both the binomial parameter p_1 and the mean of the Poisson distribution.

Lambert [23] compares several zero-inflated Poisson regression models for experimental data obtained at AT & T Bell Laboratories. The objective of the experiment was to study the influence of five qualitative factors on the number of soldering defects on printed wiring boards. When a reliable manufacturing process is in control, the number of defects on an item should be Poisson distributed. Nevertheless, the Bell Laboratories data have many more items without defects than would be expected from a Poisson distribution. The author postulates that slight, unobserved changes in the environment cause the process to move randomly back and forth between a perfect state ($V = 1$) and an imperfect state ($V = 0$). Lambert considered three types of ZIP models. In the first one, the probability that the process is in the perfect state, p_1, does not depend on the factors. In the second model, the Poisson parameter and the Bernoulli parameter are not functionally related but both depend on the same factors. Finally, in the third model, p_1 is a simple function of the Poisson parameter that depends on the factors. For each model, the author discusses the interpretation of the parameters and the algorithm to calculate the maximum likelihood estimates. Simulations showing the appropriateness of the asymptotic results are also presented.

In an unpublished work, Ridout, Demétrio and Hinde [24] fitted several regression models to experimental data from horticulture. The aim of the experiment was to evaluate the effect of 4 hormone concentrations and 2 periods of light exposure on the number of roots produced by a plant cutting. The regression models examined by the authors are: Poisson, Poisson with added zeros, negative binomial and negative binomial with added zeros. Several variations of each model were examined as the parameter accounting for the extra zeros and the dispersion parameter were sometimes expressed as a function of the exposure to light. The authors analysed the significance of the factors and compared non-nested models using the Akaike information criterion and the BIC statistic.

Several methods have been proposed to calculate the maximum likelihood estimates of the parameters of a negative binomial model. These are the Newton-Raphson method, the conditional maximum likelihood approach for the estimation of the dispersion parameter and the maximum extended quasi-likelihood method. For mixture models describing populations formed by G subgroups, the maximum likelihood estimates are usually obtained via the EM algorithm or the Newton-Raphson method.

3.3 Poisson regression model with added zeros for repeated observations

The Poisson regression model with added zeros examined here for longitudinal data assumes that a subpopulation of individuals with PsA never develops damaged joints. Besides this, it is assumed that, in a fixed time interval, the average rate at which the joints are damaged varies from one person to another. The source of this variability is considered to be known and measured by the researchers. Thus, the average increase in damaged joints between clinic visits is expressed as a function of several covariates.

3.3.1 Description of the model

Let $J_{i,j}$ represent the total number of damaged joints for patient i up to time $t_{i,j}$. Then, $D_{i,j} = J_{i,j+1} - J_{i,j}$ is the number of joints damaged between times $t_{i,j}$ and $t_{i,j+1}$ with $j = 1, 2, \ldots, m_{i-1}$ and $i = 1, 2, \ldots, n$. The $D_{i,j}$ are discrete variables so in a first approach I assume they have a Poisson distribution with mean $\mu_{i,j}$. I also assume that $D_{i,j}$ is independent of $D_{i,j-1}, \ldots, D_{i,1}$ given the value of $J_{i,j}$. This assumption is similar to the one made for the Markov model in which the state occupied by individual i at time $t_{i,j}$ depends only on the previous state. Therefore, the expected number of damaged joints in an interval of length $t_{i,j+1} - t_{i,j}$ is expressed as a function of the vector of covariates $z_i' = (1, z_{i,1}, \ldots, z_{i,p-2})$ and $J_{i,j}$ *i.e*

$$\begin{aligned}\mu_{i,j} &= (t_{i,j+1} - t_{i,j}) \exp(\alpha_0 + \alpha_1 z_{i,1} + \ldots + \alpha_{p-2} z_{i,p-2} + \alpha_{p-1} J_{i,j}) \\ &= (t_{i,j+1} - t_{i,j}) \exp(\boldsymbol{\alpha}' \boldsymbol{z}_{i,j})\end{aligned}$$

where $\boldsymbol{z}_{i,j}' = (z_i', J_{i,j})$ and $\alpha' = (\alpha_0, \alpha_1, \ldots, \alpha_{p-2}, \alpha_{p-1})$. Notice that $J_{i,j}$ is the only variable in $\boldsymbol{z}_{i,j}'$ that varies over time. A more general model would allow the other covariates, $z_{i,u}$, to change over time. In this case, $\mu_{i,j} = E(D_{i.j})$ would be a function of the covariates measured at time $t_{i,j}$.

The probability density function of $D_{i,j}$ is given by:

$$P(D_{i,j} = d_{i,j} \mid z_{i,j}) = \frac{\exp(-\mu_{i,j})\mu_{i,j}^{d_{i,j}}}{d_{i,j}!}, \qquad d_{i,j} = 0, 1, \ldots$$

As suggested by the results obtained in chapter 2, an excess of zero increments in the number of damaged joints may occur relative to the Poisson distribution. If the excess of zeros can be explained by the existence of a subpopulation of individuals who never develop damaged joints during the course of the disease, a mixture model can be used to describe the data. Let V_i be a binary variable where $V_i = 1$ indicates that individual i will not develop damaged joints and $V_i = 0$ indicates that individial i is susceptible to damaged joints. The probability that patient i is not susceptible to damaged joints can be described by a logistic model:

$$\theta_i = P(V_i = 1; \boldsymbol{z}_i^*) = \frac{\exp(\boldsymbol{\beta}'\boldsymbol{z}_i^*)}{1 + \exp(\boldsymbol{\beta}'\boldsymbol{z}_i^*)}$$

where $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_{r-1})$ and $\boldsymbol{z}_i^{*'} = (1, z_{i,1}^*, \ldots, z_{i,r-1}^*)$. It will be assumed that the covariate vectors, $\boldsymbol{z}_i^*$ and $\boldsymbol{z}_{i,j}$, modelling the Poisson and the Bernoulli parameters are, in general, different.

The probability of not observing any damaged joints for patient i is:

$$\begin{aligned} & P(D_{i,1} = 0, D_{i,2} = 0, \ldots, D_{i,m_{i-1}} = 0 \mid \boldsymbol{z}_{i,1}, \boldsymbol{z}_{i,2}, \ldots, \boldsymbol{z}_{i,m_{i-1}}) \\ = \; & P(V_i = 1 \mid \boldsymbol{z}_i^*) P(D_{i,1} = 0, \ldots, D_{i,m_{i-1}} = 0 \mid V_i = 1, \boldsymbol{z}_{i,1}, \ldots, \boldsymbol{z}_{i,m_{i-1}}) \\ + \; & P(V_i = 0 \mid \boldsymbol{z}_i^*) P(D_{i,1} = 0, \ldots, D_{i,m_{i-1}} = 0 \mid V_i = 0, \boldsymbol{z}_{i,1}, \ldots, \boldsymbol{z}_{i,m_{i-1}}) \\ = \; & \theta_i + (1 - \theta_i) \prod_{j=1}^{m_{i-1}} P(D_{i,j} = 0 \mid V_i = 0, \boldsymbol{z}_{i,j}) \\ = \; & \theta_i + (1 - \theta_i) \prod_{j=1}^{m_{i-1}} \exp(-\mu_{i,j}) \\ = \; & \theta_i + (1 - \theta_i) \exp(- \sum_{j=1}^{m_{i-1}} \mu_{i,j}) \qquad (3.3) \end{aligned}$$

For patient i, who developed damaged joints, the probability of the observed $d_{i,j}$ values is:

$$\begin{aligned} & P(D_{i,1} = d_{i,1}, \ldots, D_{i,m_{i-1}} = d_{i,m_{i-1}} \mid \boldsymbol{z}_{i,1}, \boldsymbol{z}_{i,2}, \ldots, \boldsymbol{z}_{i,m_{i-1}}) \\ = \; & (1 - \theta_i) \prod_{j=1}^{m_{i-1}} P(D_{i,j} = d_{i,j} \mid V_i = 0, \boldsymbol{z}_{i,j}) \\ = \; & (1 - \theta_i) \prod_{j=1}^{m_{i-1}} \left[\frac{\exp(-\mu_{i,j}) \mu_{i,j}^{d_{i,j}}}{d_{i,j}!} \right], \qquad (3.4) \end{aligned}$$

where $d_{i,j} > 0$ for some $j = 1, 2, \ldots, m_{i-1}$. Note that expressions (3.3) and (3.4) retain the usual form for a distribution with added zeros.

Let $J_i' = (J_{i,1}, J_{i,2}, \ldots, J_{i,m_i})$. The likelihood function for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is:

$$\begin{aligned} L(\boldsymbol{\alpha},\boldsymbol{\beta}) &= \prod_{i|J_i=0} [\theta_i + (1-\theta_i)\exp(-\sum_{j=1}^{m_{i-1}} \mu_{i,j})] \\ &\times \prod_{i|J_i\neq 0} \left\{(1-\theta_i) \prod_{j=1}^{m_{i-1}} \left[\frac{\exp(-\mu_{i,j})\mu_{i,j}^{d_{i,j}}}{d_{i,j}!}\right]\right\} \end{aligned} \tag{3.5}$$

Therefore, the logarithm of the likelihood function is proportional to:

$$\begin{aligned} l(\boldsymbol{\alpha},\boldsymbol{\beta}) &= \sum_{i|J_i=0} \ln[\theta_i + (1-\theta_i)\exp(-\sum_{j=1}^{m_{i-1}} \mu_{i,j})] \\ &+ \sum_{i|J_i\neq 0} \ln(1-\theta_i) - \sum_{i|J_i\neq 0}\sum_{j=1}^{m_{i-1}} \mu_{i,j} + \sum_{i|J_i\neq 0}\sum_{j=1}^{m_{i-1}} d_{i,j}\ln\mu_{i,j} \end{aligned}$$

3.3.2 Goodness of fit analysis

In this section I explain how to evaluate the fit of the Poisson model with added zeros for repeated observations through the calculation and analysis of the observed and expected counts.

Based on the model defined by equations (3.3) and (3.4), I propose to calculate expectations based on the quantities:

$$e_{i,j}(k) = \hat{P}(D_{i,j} = k \mid z_{i,j}) \qquad \text{for} \quad k = 0, 1, \ldots \quad \text{and} \quad z_{i,j}' = (z_i', J_{i,j}),$$

defined for the interval $(t_{i,j}, t_{i,j+1})$. For a patient with zero damaged joints up to time $t_{i,j}$, *i.e* $J_{i,j} = 0$, the above probabilities are calculated as:

$$\begin{aligned} e_{i,j}(0) = \hat{P}(D_{i,j} = 0 \mid z_{i,j}) &= \hat{\theta}_i + (1-\hat{\theta}_i)\exp(-\hat{\mu}_{i,j}) \\ e_{i,j}(k) = \hat{P}(D_{i,j} = k \mid z_{i,j}) &= \frac{(1-\hat{\theta}_i)\exp(-\hat{\mu}_{i,j})\hat{\mu}_{i,j}^k}{k!}, \qquad k = 1, 2, \ldots \end{aligned}$$

$$\text{where} \quad \hat{\theta}_i = \frac{\exp(\hat{\boldsymbol{\beta}}' z_i^*)}{1 + \exp(\hat{\boldsymbol{\beta}}' z_i^*)} \quad \text{and} \quad \hat{\mu}_{i,j} = (t_{i,j+1} - t_{i,j})\exp(\hat{\boldsymbol{\alpha}}' z_{i,j});$$

if $J_{i,j} > 0$ then the probability of k damaged joints in an interval of length $t_{i,j+1} - t_{i,j}$ is:

$$e_{i,j}(k) = \quad \hat{P}(D_{i,j} = k \mid z_{i,j}) = \frac{\exp(-\hat{\mu}_{i,j})\hat{\mu}_{i,j}^k}{k!} \qquad \text{for} \quad k = 0, 1, 2, \ldots$$

The probabilities $e_{i,j}(k)$ depend on the values of the response variable, $D_{i,j}$, the time elapsed between observations, the covariates modelling the Poisson parameter, the number of damaged joints observed up to time $t_{i,j}$, and, if $J_{i,j} = 0$, the covariates affecting the Bernoulli parameter. All this information needs to be considered if the $e_{i,j}(k)$ are to be grouped to form a contingency table. As done for the Markov regression model, if the covariates are continuous and the observation times are variable, the $e_{i,j}(k)$ can be classified into equiprobable categories defined by the quantiles of $\exp(\hat{\alpha}' z_{i,j})$, $\hat{\theta}_i$, and $t_{i,j+1} - t_{i,j}$. Nevertheless, with several classification criteria sparse cells may occur (Bishop *et. al.* [9]). To have a preliminary indication of how accurately the model fits the PsA data, I collapse the estimated probabilities across all the variables just mentioned except the values of $D_{i,j}$ *i.e.*

$$e(k) = \sum_{i=1}^{n} \sum_{j=1}^{m_{i-1}} e_{i,j}(k) = \sum_{i=1}^{n} \sum_{j=1}^{m_{i-1}} \hat{P}(D_{i,j} = k \mid z_{i,j}) \qquad \text{where} \quad k = 0, 1, \ldots$$

Furthermore, I only consider eleven categories for the increase in the number of damaged joints, namely: $0, 1, \ldots, 9, \geq 10$. The total number of observed increments equal to k is:

$$n(k) = \sum_{i=1}^{n} \sum_{j=1}^{m_{i-1}} 1_{\{D_{i,j}=k\}}$$

where $1_{\{D_{i,j}=k\}}$ is an indicator variable equal to one if $D_{i,j} = k$. The ratios $(n(k) - e(k))^2 / e(k)$ are calculated to evaluate the discrepancies between the observed and expected counts. The sum of these ratios is used to measure the overall discrepancy between the fitted model and the data.

3.3.3 Tests to determine the significance of the additional parameter

Models should not be used for descriptive or predictive purposes without checking them carefully. The model checking phase is often overlooked in practice. It comprises several techniques to establish the validity of the assumptions, to detect any misspecification of the various components of the model and to identify outliers and influential observations.

When a mixture model is based on the assumption that the population is divided into G different groups, it is important to check that the proportion of individuals in each group is greater than zero.

The simplest situation is to determine if the population is divided into two groups defined by a Bernoulli parameter that does not depend on covariates: *i.e.*

$$H_0 : \theta = 0 \quad vs. \quad H_1 : \theta > 0$$

This is a non-standard hypothesis test because, under H_0, θ lies on the boundary of the parameter space. Note that this problem persists if the logit of θ is equal to β_0 and the hypothesis is expressed in terms of β_0:

$$H_0 : \beta_0 = -\infty \quad vs. \quad H_1 : \beta_0 > -\infty \tag{3.6}$$

Self and Liang [25] and Ghitany, Maller and Zhou [26] proved that the deviance statistic for testing this type of hypothesis has, asymptotically, not a chi-squared distribution with one degree of freedom, but the distribution of X, where

$$P(X \leq x) = 0.5 + 0.5P(\chi^2_1 \leq x) \tag{3.7}$$

A more general test was proposed by Ghitany, Maller and Zhou [26] to determine if the proportion of individuals that do not experience the event of interest differs between levels in a one-way classification. The authors assumed that the proportions do not depend on additional covariates.

When covariates affect the proportion of individuals in each subpopulation, the test to determine if the proportions are greater than zero has an additional non-standard feature. Under the null hypothesis, the intercept is equal to $-\infty$ and the effect of the covariates is therefore overriden by β_0. Essentially, the regression coefficients disappear under the null hypothesis. A heuristic approach is to determine if the effect of the covariates is significantly different from zero. If it is not then the problem is reduced to the situation described above. When the effect of the covariates is significant, the sampled units can be stratified based on the value of the covariates. The hypothesis that the proportion of individuals in each subpopulation is greater than zero is investigated in each strata.

Confidence intervals based on the profile likelihood for β_0 are an alternative procedure to examine if the proportion of subjects in each subpopulation is greater than zero. Profile likelihoods are often used to construct confidence regions when the maximum likelihood estimate of a parameter does not have an asymptotic normal distribution. In the rest of the section I explain how to calculate the profile likelihood for β_0 and how to use it to construct an asymptotic confidence interval for β_0. I assume that no covariates affect the Bernoulli parameter so $\text{logit}(\theta) = \beta_0$.

If β_0 is replaced by a fixed value β_0^*, equation (3.5) becomes:

$$\begin{aligned} L(\alpha \mid \beta_0 = \beta_0^*) &= \prod_{i \mid J_i = 0} [\theta^* + (1-\theta^*)\exp(-\sum_{j=1}^{m_{i-1}} \mu_{i,j}^*)] \\ &\quad \times \prod_{i \mid J_i \neq 0} \left\{ (1-\theta^*) \prod_{j=1}^{m_{i-1}} \left[\frac{\exp(-\mu_{i,j}^*){\mu_{i,j}^*}^{d_{i,j}}}{d_{i,j}!} \right] \right\}. \end{aligned}$$

The profile likelihood for β_0, denoted as $PL(\beta_0)$, is obtained by substituting $\boldsymbol{\alpha}$ by its maximum likelihood estimate ($\hat{\boldsymbol{\alpha}}^*$) in the above expression *i.e.*

$$PL(\beta_0^*) = L(\hat{\boldsymbol{\alpha}}^* \mid \beta_0 = \beta_0^*)$$

The likelihood ratio statistic for the hypothesis:

$$H_0 : \beta_0 = \beta_0^* \quad vs. \quad H_1 : \beta_0 \neq \beta_0^*$$

can be expressed in terms of the profile likelihood for β_0 as:

$$\Lambda = -2\ln\left(\frac{PL(\beta_0^*)}{L(\hat{\alpha}, \hat{\beta}_0)}\right) = -2\ln\left(\frac{PL(\beta_0^*)}{PL(\hat{\beta}_0)}\right)$$

The test statistic Λ has an asymptotic chi-square distribution with one degree of freedom. Therefore a $(1-\alpha)\cdot 100\%$ confidence interval for β_0 is the set of values of β_0 for which:

$$\frac{PL(\beta_0)}{L(\hat{\alpha}, \hat{\beta}_0)} \geq \exp(-0.5\chi^2_{1,1-\alpha})$$

or equivalently: $S = \ln(PL(\beta_0)) - \ln(L(\hat{\alpha}, \hat{\beta}_0)) \geq -0.5\chi^2_{1,1-\alpha}$ If $\alpha = 0.05$ then $-0.5\chi^2_{1,0.95} = -1.921$.

3.3.4 Results for the PsA data

For the PsA data, the proportion of individuals not susceptible to damaged joints was modelled with a dependence on the patient's age at the time of the PsA onset. The asymptotic 95% confidence interval for the corresponding parameter contains the value zero: $0.0136 \pm 1.96 \times 0.0112 = (-0.0084, 0.0356)$. Furthermore, the deviance statistic to test the significance of this parameter is equal to 1.42 with a significance level of 0.233. Therefore, there is no evidence that the patient's age at the time of the disease onset is related to the chance of developing damaged joints. Consequently, I assume that the logit of this proportion is constant. The mean number of damaged joints between successive assessments is modelled by the sedimentation rate and the number of effused joints recorded on the first clinic visit, the type of medication taken before participating in the study and the previous number of damaged joints. Table 3.1 shows the maximum likelihood estimates for the regression parameters; the numbers in brackets are the standard errors.

Table 3.1: Estimated parameters and standard deviations for the Poisson regression model with added zeros for repeated observations.

Parameter		Estimates
Binomial intercept		-0.772 (0.139)
Poisson intercept		-0.318 (0.062)
Erythrocyte sedimentation rate		0.001 (0.001)
Num. effused joints		0.051 (0.007)
Disease modifying drugs	Yes	0.027 (0.095)
Use of corticosteroids	Yes	0.109 (0.061)
Prev. number damaged joints		0.014 (0.002)

The number of damaged joints observed until the previous visit and the number of effused joints recorded in the first assessment are the only covariates with a significant effect on the mean number of damaged joints between visits.

The logarithm of the likelihood function evaluated at the estimated parameters is equal to -1395.07. Table 3.2 lists the values obtained for the logarithm of the profile likelihood (S) when the Bernoulli parameter takes different values around its maximum likelihood estimate. The values are plotted on Figure 3.1. The curve is fairly symmetrical around -0.77. For this reason, the asymptotic 95% confidence interval based on the profile likelihood, (-1.05,-0.5), is equal to the 95% confidence interval obtained by assuming that the additional parameter has an asymptotic normal distribution, $-0.77 \pm 1.96 \times 0.14 = (-1.05, -0.5)$. Calculation of the inverse of the logit transformation gives: $(\exp(-1.05)/(1+\exp(-1.05)), \exp(-0.5)/(1+\exp(-0.5))) = (0.259, 0.378)$. This means that with 95% confidence, the percentage of patients with PsA that are not susceptible to damaged joints is estimated to lie between 25.9% and 37.8%. The point estimate is $31.6\%(=$

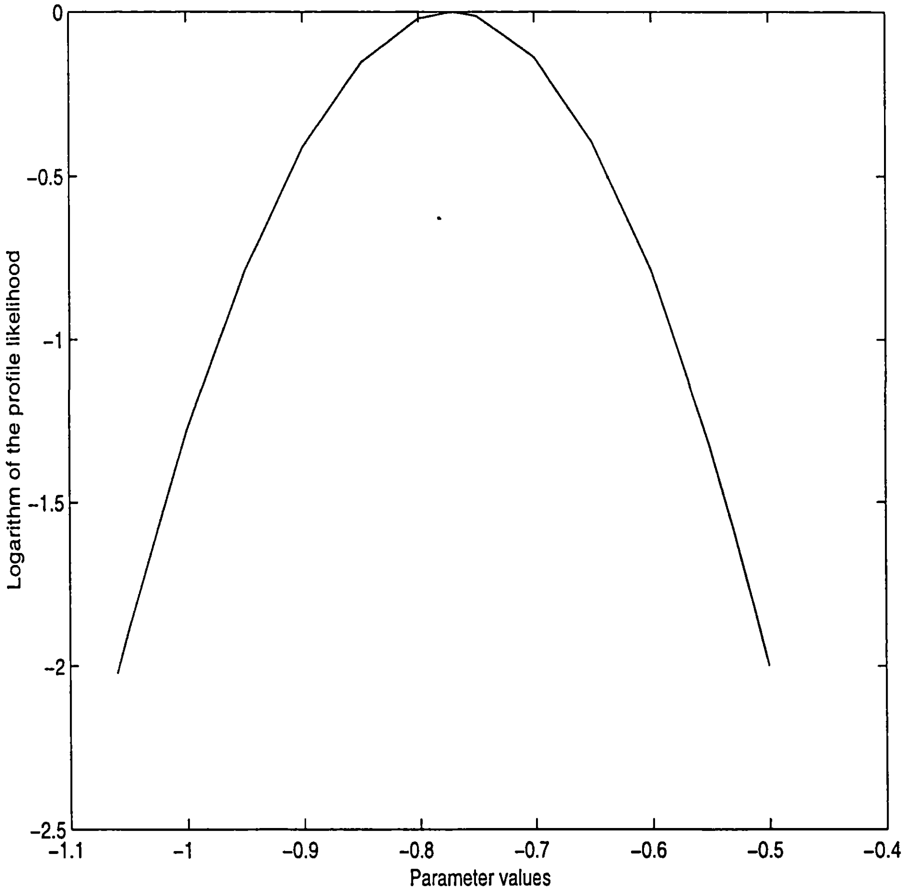
Table 3.2: Statistic based on the logarithm of the profile likelihood for the Poisson regression model with added zeros for repeated observations.

Binomial intercept	Statistic
-1.06	-2.023
-1.053	-1.928
-1.05	-1.888
-1.00	-1.282
-0.95	-0.788
-0.90	-0.411
-0.85	-0.153
-0.80	-0.020
-0.77	-0.000
-0.75	-0.013
-0.70	-0.137
-0.65	-0.394
-0.60	-0.789
-0.55	-1.323
-0.53	-1.576
-0.51	-1.853
-0.50	-2.000

Table 3.3: Total number of observed and expected counts for the Poisson regression model with added zeros for repeated observations.

Increase in the Num. of Damaged Joints	Observed Count	Expected Count	Scaled Differences
0	1497	1032.20	209.30
1	146	481.87	234.11
2	87	190.31	56.08
3	27	75.08	30.79
4	20	35.80	6.97
5	19	20.60	0.12
6	17	13.15	1.13
7	8	8.70	0.06
8	10	5.78	3.08
9	5	3.81	0.37
10 +	39	7.70	127.13
Total	1875	1875.00	669.15

Figure 3.1: Logarithm of the profile likelihood for the Bernoulli parameter of the Poisson regression model with added zeros for repeated observations.



$100\% \times \exp(-0.77)/(1+\exp(-0.77)))$. The deviance statistic for testing if this percentage is greater than zero is equal to 395.65 with a significance level of 0.0 based on (3.7). Therefore, the confidence interval and the hypothesis test indicate that the proportion of patients not susceptible to damaged joints is significantly greater than zero.

The total numbers of observed and expected counts are shown on Table 3.3. Although the proportion of individuals not developing damaged joints is estimated to be greater than zero, the model predicts fewer zero increments than observed in the data. In contrast, the expected number of increments equal to 1, 2, 3 and 4 damaged joints overestimates the corresponding observed count. The model also underestimates the total number of increments with 10 or more damaged joints. The scaled differences in the last column of Table 3.3 are the ratio of the squared difference between each observed and expected count divided by the corresponding expected count. The sum of the scaled differences is 669.15 and suggests that the proposed model does not fit the PsA data.

3.4 Negative binomial regression model for repeated observations

The negative binomial regression model is frequently used as an alternative to the Poisson regression model (without added zeros). The. difference between them is that the negative binomial model assumes that the average rate at which the joints are damaged between clinic visits depends on several covariates and on a random (unknown) component. Unlike the model examined in the previous section, the negative binomial regression model for repeated observations assumes that all patients with PsA are susceptible to damaged joints.

3.4.1 Description of the model

Recall that $D_{i,j} = J_{i,j+1} - J_{i,j}$ represents the increase in the number of damaged joints for patient i in the period of length $t_{i,j+1} - t_{i,j}$. Here I assume that the random variables $D_{i,j}$ are independent and follow a negative binomial distribution given the number of damaged joints recorded up to the last assessment, $J_{i,j}$. The mean of $D_{i,j}$ is expressed as a function of the vector of covariates: $z_i' = (1, z_{i,1}, \ldots, z_{i,p-2})$ and $J_{i,j}$ as follows:

$$\mu_{i,j} = (t_{i,j+1} - t_{i,j}) \exp(\boldsymbol{\alpha}' \boldsymbol{z}_{i,j})$$

where $\boldsymbol{z}_{i,j}' = (z_i', J_{i,j})$ and $\boldsymbol{\alpha}' = (\alpha_0, \alpha_1, \ldots, \alpha_{p-2}, \alpha_{p-1})$. Thus,

$$P(D_{i,j} = d_{i,j} \mid \boldsymbol{z}_{i,j}) = \frac{\Gamma(d_{i,j} + \gamma^{-1})}{d_{i,j}!\Gamma(\gamma^{-1})} \Big(\frac{\gamma\mu_{i,j}}{1 + \gamma\mu_{i,j}}\Big)^{d_{i,j}} \Big(\frac{1}{1 + \gamma\mu_{i,j}}\Big)^{\gamma^{-1}} \qquad (3.8)$$

where $\gamma \geq 0$ is the dispersion parameter, $\Gamma(\cdot)$ is the gamma function and $d_{i,j} = 0, 1, 2, \ldots$.

Therefore, the logarithm of the likelihood function for $\boldsymbol{\alpha}$ and γ is:

$$\ln(L(\boldsymbol{\alpha}, \gamma)) = \sum_{i=1}^{n} \sum_{j=1}^{m_{i-1}} \ln(P(D_{i,j} = d_{i,j} \mid \boldsymbol{z}_{i,j}))$$

$$\begin{array}{llll}
\text{if} & d_{i,j} = 0 & \text{then} & \ln(P(D_{i,j} = d_{i,j} \mid \boldsymbol{z}_{i,j})) = -\frac{1}{\gamma}\ln(1 + \gamma\mu_{i,j}), \\
\text{if} & d_{i,j} > 0 & \text{then} & \ln(P(D_{i,j} = d_{i,j} \mid \boldsymbol{z}_{i,j})) \\
 & & = & \ln\Big\{\frac{\Gamma(d_{i,j} + \gamma^{-1})}{d_{i,j}!\Gamma(\gamma^{-1})}\Big\} + d_{i,j}\ln\Big(\frac{\gamma\mu_{i,j}}{1 + \gamma\mu_{i,j}}\Big) - \frac{1}{\gamma}\ln(1 + \gamma\mu_{i,j}) \\
 & & = & \ln\Big\{\Big(\frac{1}{d_{i,j}!}\Big)\gamma^{-1}(\gamma^{-1} + 1)(\gamma^{-1} + 2)\ldots(\gamma^{-1} + d_{i,j} - 2)(\gamma^{-1} + d_{i,j} - 1)\Big\} \\
 & & + & d_{i,j}\ln\Big(\frac{\gamma\mu_{i,j}}{1 + \gamma\mu_{i,j}}\Big) - \frac{1}{\gamma}\ln(1 + \gamma\mu_{i,j})
\end{array}$$

3.4.2 Model appraisal

Expectations related to the number of damaged joints occurring in the interval $(t_{i,j}, t_{i,j+1})$ are again defined in terms of the probabilities of $k = 0, 1, 2, \ldots$ damaged joints in a period of length $t_{i,j+1} - t_{i,j}$, *i.e.*

$$e_{i,j}(k) = \hat{P}(D_{i,j} = k \mid z_{i,j}) = \frac{\Gamma(k + \hat{\gamma}^{-1})}{k!\Gamma(\hat{\gamma}^{-1})} \Big(\frac{\hat{\gamma}\hat{\mu}_{i,j}}{1 + \hat{\gamma}\hat{\mu}_{i,j}}\Big)^k \Big(\frac{1}{1 + \hat{\gamma}\hat{\mu}_{i,j}}\Big)^{\hat{\gamma}^{-1}}.$$

Grouping the $e_{i,j}(k)$ according to the values of k regardless of the patient's characteristics or the time elapsed between observations gives:

$$e(k) = \sum_{i=1}^{n} \sum_{j=1}^{m_{i-1}} e_{i,j}(k) = \sum_{i=1}^{n} \sum_{j=1}^{m_{i-1}} \hat{P}(D_{i,j} = k \mid z_{i,j})$$

for the expected number of observations of k new damaged joints between visits.

The $e(k)$ become smaller as k tends to infinity so I only consider eleven expected counts: $e(0), e(1), \ldots, e(9), e(10+)$, where

$$e(10+) = \sum_{i=1}^{n} \sum_{j=1}^{m_{i-1}} [1 - \sum_{k=0}^{9} e_{i,j}(k)]$$

represents the expectation of the number of observations of 10 or more damaged joints. Analogously, $n(k) = \sum_{i=1}^{n} \sum_{j=1}^{m_{i-1}} 1_{\{D_{i,j}=k\}}$ is the total number of observed increments equal to k.

3.4.3 Results for the PsA data

The logarithm of the likelihood function evaluated at the maximum likelihood estimates is -88.43. Table 3.4 contains the parameter estimates along with their asymptotic standard errors in brackets. The magnitude of the estimated dispersion parameter, compared with its standard deviation, suggests that it contributes to the explanation of the high incidence of zero increments in the data. The covariate that has the most significant effect in predicting the mean increase in the number of damaged joints between clinic visits is the number of damaged joints until the last assessment. The type of medication taken before participating in the study has a smaller but significant effect on the average increase in damaged joints.

Table 3.4: Estimated parameters and standard deviations for the negative binomial regression model for repeated observations.

Parameter		Estimates	
Dispersion parameter		7.475	(0.548)
Intercept		-1.092	(0.162)
Erythrocyte sedimentation rate		0.006	(0.004)
Num. effused joints		0.051	(0.026)
Disease modifying drugs	Yes	0.647	(0.247)
Use of corticosteroids	Yes	0.481	(0.158)
Prev. number damaged joints		0.046	(0.008)

Table 3.5: Total number of observed and expected counts for the negative binomial regression model for repeated observations.

Increase in the Num. of Damaged Joints	Observed Count	Expected Count	Scaled Differences
0	1497	1499.03	0.003
1	146	152.29	0.260
2	87	67.45	5.665
3	27	38.42	3.396
4	20	24.64	0.872
5	19	16.99	0.238
6	17	12.33	1.772
7	8	9.29	0.180
8	10	7.22	1.072
9	5	5.74	0.097
10 +	39	41.60	0.162
Total	1875	1875.00	13.716

In general, the observed and expected counts in Table 3.5 show good agreement. The negative binomial regression model for repeated observations predicts quite accurately the number of increments equal to zero damaged joints. Nevertheless, the model underestimates the number of increments equal to 2 damaged joints and overestimates the number of increments equal to 3 damaged joints. These two categories contribute 9.06 units to the sum of the scaled differences: 13.72.

3.5 Negative binomial regression model with added zeros for repeated observations

The model fitted in the previous section is based on an unrealistic assumption if the population of patients with PsA is formed by a group of individuals who never develop damaged joints and a group of persons susceptible to damaged joints. In this case a natural alternative is the negative binomial regression model with added zeros for repeated observations. Here I examine the fit of this model.

3.5.1 Description of the model

As before, the increase in the number of damaged joints between times $t_{i,j}$ and $t_{i,j+1}$ is denoted as $D_{i,j} = J_{i,j+1} - J_{i,j}$ for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m_{i-1}$. In the group of individuals susceptible to damaged joints, the distribution of $D_{i,j}$ conditional on the vector of covariates $\boldsymbol{z}_{i,j}' = (1, z_{i,1}, \ldots, z_{i,p-2}, J_{i,j})$ is described by a negative binomial regression model with mean $\mu_{i,j} = (t_{i,j+1} - t_{i,j})\exp(\boldsymbol{\alpha}' \boldsymbol{z}_{i,j})$, see expression (3.8).

Thus, the probability of not observing damaged joints for patient i during the course of the study is:

$$P(\boldsymbol{D}_i = \boldsymbol{0} \mid \boldsymbol{z}_{i,1}, \ldots, \boldsymbol{z}_{i,m_{i-1}}) = \theta_i + (1 - \theta_i) \prod_{j=1}^{m_{i-1}} \left(\frac{1}{1 + \gamma\mu_{i,j}}\right)^{\gamma^{-1}} \tag{3.9}$$

while the probability that patient i develops damaged joints in the pattern specified by $\boldsymbol{d}_i$ is:

$$
\begin{aligned}
&P(\boldsymbol{D}_i = \boldsymbol{d}_i \mid \boldsymbol{z}_{i,1}, \ldots, \boldsymbol{z}_{i,m_{i-1}}) = \\
&\qquad (1-\theta_i) \prod_{j=1}^{m_{i-1}} \frac{\Gamma(d_{i,j}+\gamma^{-1})}{d_{i,j}!\Gamma(\gamma^{-1})} \left(\frac{\gamma\mu_{i,j}}{1+\gamma\mu_{i,j}}\right)^{d_{i,j}} \left(\frac{1}{1+\gamma\mu_{i,j}}\right)^{\gamma^{-1}} \qquad (3.10)
\end{aligned}
$$

where $\theta_i = P(V_i = 1 \mid \boldsymbol{z}_i^*) = \exp(\boldsymbol{\beta}'\boldsymbol{z}_i^*)/(1+\exp(\boldsymbol{\beta}'\boldsymbol{z}_i^*))$ is the probability that subject i belongs to the subpopulation of individuals not susceptible to damaged joints.

The likelihood function for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and γ is:

$$
\begin{aligned}
L(\boldsymbol{\alpha},\boldsymbol{\beta},\gamma) &= \prod_{i|J_i=0} \left\{\theta_i + (1-\theta_i) \prod_{j=1}^{m_{i-1}} \left(\frac{1}{1+\gamma\mu_{i,j}}\right)^{\gamma^{-1}}\right\} \\
&\times \prod_{i|J_i\neq 0} \left\{(1-\theta_i) \prod_{j=1}^{m_{i-1}} \frac{\Gamma(d_{i,j}+\gamma^{-1})}{d_{i,j}!\Gamma(\gamma^{-1})} \left(\frac{\gamma\mu_{i,j}}{1+\gamma\mu_{i,j}}\right)^{d_{i,j}} \left(\frac{1}{1+\gamma\mu_{i,j}}\right)^{\gamma^{-1}}\right\}
\end{aligned}
$$

Thus the logarithm of $L(\boldsymbol{\alpha},\boldsymbol{\beta},\gamma)$ is proportional to:

$$
\begin{aligned}
l(\boldsymbol{\alpha},\boldsymbol{\beta},\gamma) &= \sum_{i|J_i=0} \ln\left\{\theta_i + (1-\theta_i) \prod_{j=1}^{m_{i-1}} \left(\frac{1}{1+\gamma\mu_{i,j}}\right)^{\gamma^{-1}}\right\} \\
&+ \sum_{i|J_i\neq 0} \ln(1-\theta_i) + \sum_{i|J_i\neq 0} \sum_{j=1}^{m_{i-1}} \ln\left(\frac{\Gamma(d_{i,j}+\gamma^{-1})}{d_{i,j}!\Gamma(\gamma^{-1})}\right) \\
&+ \sum_{i|J_i\neq 0} \sum_{j=1}^{m_{i-1}} d_{i,j} \ln(\gamma\mu_{i,j}) - \sum_{i|J_i\neq 0} \sum_{j=1}^{m_{i-1}} d_{i,j} \ln(1+\gamma\mu_{i,j}) \\
&-\gamma^{-1} \sum_{i|J_i\neq 0} \sum_{j=1}^{m_{i-1}} \ln(1+\gamma\mu_{i,j})
\end{aligned}
$$

where

$$
\frac{\Gamma(d_{i,j}+\gamma^{-1})}{d_{i,j}!\Gamma(\gamma^{-1})} = \begin{cases} 1 & \text{if} \quad d_{i,j} = 0 \\[2ex] \left(\frac{1}{d_{i,j}!}\right)\gamma^{-1}(\gamma^{-1}+1)(\gamma^{-1}+2)\cdots & \\ \qquad \ldots(\gamma^{-1}+d_{i,j}-2)(\gamma^{-1}+d_{i,j}-1) & \text{if} \quad d_{i,j} = 1,2,\ldots \end{cases}
$$

3.5.2 Model appraisal

For the i-th patient, the estimated probability that $D_{i,j} = k$ in the interval $(t_{i,j}, t_{i,j+1})$ is again used as the estimate of the expected number of increases equal to k damaged joints in the period $t_{i,j+1} - t_{i,j}$ *i.e.*

$$e_{i,j}(k) = \hat{P}(D_{i,j} = k \mid z_{i,j}) \quad \text{where} \quad k = 0, 1, 2, \ldots.$$

If $J_{i,j} = 0$, the model described by equations (3.9) and (3.10) implies that $e_{i,j}(k)$ is calculated as:

$$e_{i,j}(k) = \begin{cases} \hat{\theta}_i + (1 - \hat{\theta}_i)\left(\dfrac{1}{1 + \hat{\gamma}\hat{\mu}_{i,j}}\right)^{\hat{\gamma}^{-1}} & \text{if} \quad k = 0 \\ \\ (1 - \hat{\theta}_i)\dfrac{\Gamma(k + \hat{\gamma}^{-1})}{k!\Gamma(\hat{\gamma}^{-1})}\left(\dfrac{\hat{\gamma}\hat{\mu}_{i,j}}{1 + \hat{\gamma}\hat{\mu}_{i,j}}\right)^k \left(\dfrac{1}{1 + \hat{\gamma}\hat{\mu}_{i,j}}\right)^{\hat{\gamma}^{-1}} & \text{if} \quad k = 1, 2, \ldots \end{cases}$$

but if $J_{i,j} > 0$ then

$$e_{i,j}(k) = \frac{\Gamma(k + \hat{\gamma}^{-1})}{k!\Gamma(\hat{\gamma}^{-1})}\left(\frac{\hat{\gamma}\hat{\mu}_{i,j}}{1 + \hat{\gamma}\hat{\mu}_{i,j}}\right)^k \left(\frac{1}{1 + \hat{\gamma}\hat{\mu}_{i,j}}\right)^{\hat{\gamma}^{-1}} \quad \text{for} \quad k = 0, 1, 2, \ldots$$

where $\hat{\theta}_i$, $\hat{\mu}_{i,j}$, and $\hat{\gamma}$ are the maximum likelihood estimates of θ_i, $\mu_{i,j}$ and γ respectively.

As for the previous models, the grouped expected counts: $e(k) = \sum_{i=1}^{n} \sum_{j=1}^{m_i-1} e_{i,j}(k)$ are compared with the total number of observed increments equal to k damaged joints: $n(k) = \sum_{i=1}^{n} \sum_{j=1}^{m_i-1} 1_{\{D_{i,j}=k\}}$ where $k = 0, 1, \ldots, 9, \geq 10$.

The hypothesis test described in section 3.3.3 is also applied to determine if there is evidence that the proportion of patients not susceptible to damaged joints is greater than zero.

3.5.3 Results for the PsA data

First I fitted the model in which the logit of the additional parameter is expressed as a function of the patient's age at the time of the PsA onset.

Figure 3.2: Logarithm of the profile likelihood for the Bernoulli parameter of the negative binomial regression model with added zeros for repeated observations

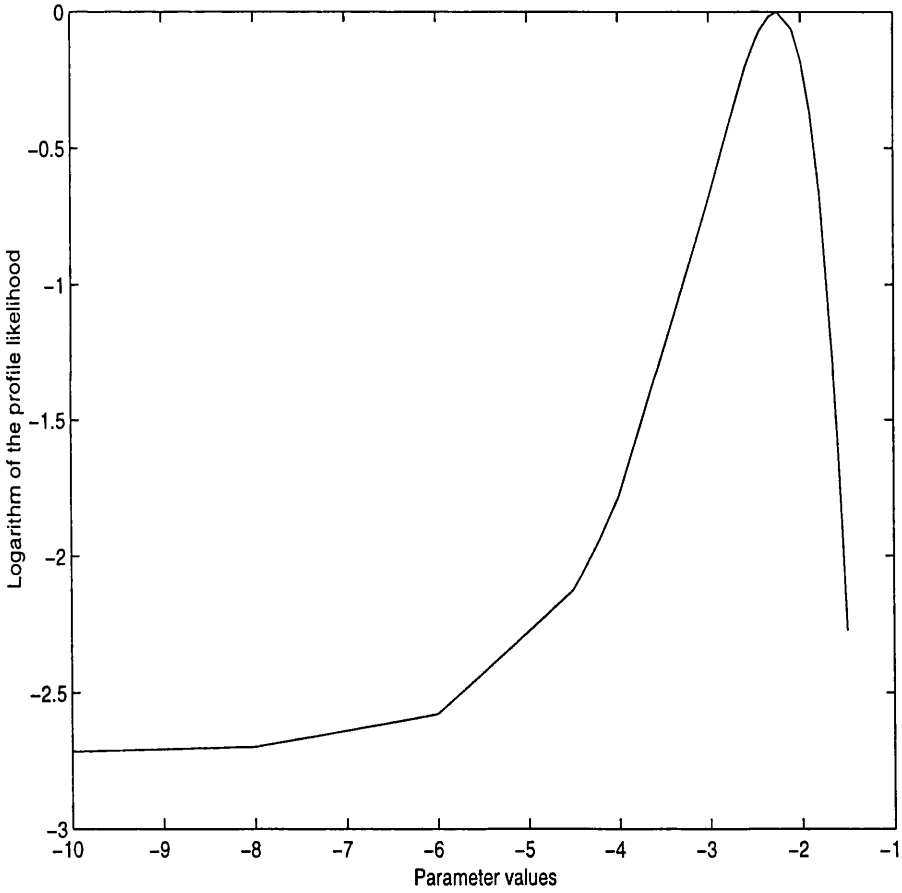


Table 3.6: Estimated parameters and standard deviations for the negative binomial regression model with added zeros for repeated observations.

Parameter		Estimates	
Binomial intercept		-2.259	(0.476)
Dispersion parameter		6.762	(0.563)
Negative binomial intercept		-0.830	(0.196)
Erythrocyte sedimentation rate		0.005	(0.004)
Num. effused joints		0.042	(0.026)
Disease modifying drugs	Yes	0.579	(0.253)
Use of corticosteroids	Yes	0.418	(0.162)
Prev. number damaged joints		0.038	(0.008)

The asymptotic 95% confidence interval for the parameter measuring the effect of this covariate contains the value zero: $-0.011 \pm 1.96 \times 0.027 = (-0.064, 0.042)$. Also, the deviance statistic for testing the significance of this parameter is equal to 0.163 with a significance level of 0.687. This means that the data do not provide evidence that the patient's age at the time of the disease onset is related to the proportion of patients not susceptible to damaged joints. Therefore, the model I discuss below assumes that the logit of the additional parameter is constant and that the mean increase in damaged joints between consecutive visits depends on the erythrocyte sedimentation rate and the number of effused joints recorded on the first assessment, the type of medication taken before joining the PsA study and the number of damaged joints observed until the last visit.

The logarithm of the likelihood function evaluated at the maximum likelihood estimates is equal to -85.71. The estimates of the parameters and of the standard deviations are presented in Table 3.6. Because of the additional parameter, the estimate of the dispersion parameter is slightly smaller than the one obtained for the negative binomial regression model for repeated observations. Nevertheless, the values in Table 3.6 are similar to the ones in

Table 3.7: Statistic based on the logarithm of the profile likelihood for the negative binomial regression model with added zeros for repeated observations.

Binomial intercept	Statistic
-10.00	-2.716
-8.00	-2.699
-6.00	-2.579
-4.50	-2.126
-4.40	-2.068
-4.20	-1.938
-4.00	-1.785
-3.00	-0.676
-2.60	-0.196
-2.50	-0.106
-2.45	-0.069
-2.35	-0.017
-2.26	0.000
-2.10	-0.063
-2.00	-0.181
-1.90	-0.377
-1.80	-0.665
-1.64	-1.364
-1.60	-1.592
-1.55	-1.912
-1.50	-2.272

Table 3.8: Total number of observed and expected counts for the negative binomial regression model with added zeros for repeated observations.

Increase in the Num. of Damaged Joints	Observed Count	Expected Count	Scaled Differences
0	1497	1475.48	0.31
1	146	160.33	1.28
2	87	72.50	2.90
3	27	41.73	5.20
4	20	26.89	1.76
5	19	18.56	0.01
6	17	13.45	0.94
7	8	10.11	0.44
8	10	7.82	0.61
9	5	6.19	0.23
10 +	39	41.94	0.21
Total	1875	1875.00	13.89

Table 3.4. Again, the number of damaged joints recorded until the last visit is the most significant covariate. It is followed by the covariates representing the type of medication taken before entering the study.

Figure 3.2 shows the logarithm of the profile likelihood for the additional parameter. The curve is not symmetrical with respect to -2.26, the maximum likelihood estimate of the additional parameter. This suggests that the estimate does not have an asymptotic normal distribution. Therefore, the 95% confidence interval based on such a distribution, $-2.26 \pm 1.96 \times 0.48 = (-3.19, -1.33)$, is of questionable validity. A 95% confidence interval based on Figure 3.2, or equivalently, on Table 3.7 is: (-4.2,-1.55). Calculating the inverse of the logit transformation gives: $(\exp(-4.2)/(1 + \exp(-4.2)), \exp(-1.55)/(1+\exp(-1.55))) = (0.015, 0.175)$. This means that with a confidence of 95%, the proportion of individuals not susceptible to damaged joints is estimated to lie between 1.5% and 17.5%. The deviance statistic for testing if this proportion is greater than zero is equal to 5.44 with a significance level of 0.0099 based on (3.7). Both, the confidence interval and the hypothesis test indicate that the estimate of the proportion of patients that never develop damaged joints is significantly greater than zero. The point estimate is $9.45\% = (100\% \times \exp(-2.26)/(1 + \exp(-2.26)))$.

Table 3.8 lists the observed and expected counts as well as their scaled differences. Although the estimate of the additional parameter is significantly greater than zero, the estimated number of increments equal to zero damaged joints is not as accurate as the estimate produced by the negative binomial regression model for repeated observations (see Table 3.5). The fitted model also underestimates and overestimates the number of increments equal to 2 and 3 damaged joints respectively. These two categories contribute with 8.1 units to the sum of the scaled differences: 13.89.

3.6 Conclusions

Three models for discrete response variables measured repeatedly over time were examined in this chapter. The Poisson regression model with added zeros for repeated observations assumes that a subpopulation of patients with PsA never develops damaged joints. The negative binomial regression model for repeated observations is not based on this assumption but its dispersion parameter can describe data with a large proportion of zeros. The negative binomial regression model with added zeros for repeated observations combines the effect of the dispersion parameter and of the additional parameter. Therefore, it is useful for modelling data with a high proportion of zeros some of which correspond to individuals not susceptible to damaged joints.

The only covariate that was available in the data set that could be related to the additional parameter is the age of the patient at the time of the disease onset. However, the results indicate that it does not have a significant effect on the proportion of patients that never develop damaged joints. These results - obtained for the two mixture models examined in this chapter - allowed me to test if such a proportion is greater than zero.

The data analysed contains 105 (36.84%) patients with no damaged joints in the entire follow-up period. In other words, about a third of the sampled individuals did not develop damaged joints during the course of the study. The two mixture models for repeated observations investigated here assume that a fraction of these patients are those who will never develop damaged joints. With 95% confidence, the estimated percentage of patients not susceptible to damaged joints lies between 25.9% and 37.8% for the Poisson mixture model for repeated observations and between 1.5% and 17.5% for the negative binomial mixture model for repeated observations. The first confidence interval contains values greater than 36.84%. This means that damaged joints were observed for nearly all the patients susceptible to present them and that the majority of the patients with zero-damaged joints will remain like that forever. In other words, the data collection period coincided with the time at which damaged joints occurred for those patients susceptible to present them. Samples like this seldom occur in practice. Therefore, the confidence interval might indicate that the additional parameter explains between 25.9% and 37.8% of the zero-increases in damaged joints, regardless of the subpopulation to which the individuals belong. The confidence interval produced by the negative binomial mixture model indicates that a proportion as small as 1.5% of the population with PsA might never develop damaged joints. Such a small proportion might be clinically unimportant.

Thus, the estimate of the additional parameter is significantly greater than zero for the two models with added zeros. However, in the Poisson regression model, care must be taken to interpret it as the proportion of individuals who are not susceptible to damaged joints. In the negative binomial regression model, the estimated proportion of patients not susceptible to damage joints can be so small that its practical usefulness is arguable. Thus, on balance, I believe the results obtained do not establish that individuals with PsA are in fact divided into two groups. This hypothesis needs to be investigated more thoroughly with additional data.

In the three models examined here, the number of damaged joints recorded up to the last assessment is the most significant covariate for predicting the mean increase in damaged joints between consecutive visits. The parameter estimating the effect of this covariate is positive so the more damaged joints observed until the last assessment, the bigger the increase in the number of damaged joints becomes. For the two negative binomial regression models, the type of medication taken before participating in the study also has a significant effect on the mean increase in damaged joints. However, for the Poisson regression model with added zeros, a discrete variable - the number of effused joints observed on the first clinic visit - also has a significant effect on the mean increase in damaged joints.

The tables of observed and expected counts show that the Poisson regression model with added zeros for repeated observations produces a poor fit compared with that of the negative binomial regression models. Furthermore, the parameter estimates and the contingency tables of the two negative binomial regression models examined here are similar. The one without the additional parameter is preferred unless further studies confirm that a subpopulation of patients with PsA do not develop damaged joints and it is clinically important to identify these cases. Also, the negative binomial regression model for repeated observations is more parsimonious so it is easier to interpret and to fit using standard statistical software.

In summary, the three models examined in this chapter suggest that the rate at which the joints are damaged between clinic visits is better explained by a Poisson model with a random patient specific effect, *i.e.* the negative binomial model, rather than by a subpopulation of individuals who never develop damaged joints. Only some of the covariates used to model the transition rates of the Markov regression model had a significant effect on the mean of the negative binomial regression model for repeated observations. It can not be expected that the same covariates have a comparable or even significant effect on models with different response variables. The fit of the negative binomial regression model might be improved by considering other covariates. However, in this chapter I focused on techniques to measure the goodness of fit of models for repeated observations and not on covariate selection procedures.

Chapter 4

Comparison between the Markov and the negative binomial regression models

4.1 Introduction

In this chapter I examine the fit of the negative binomial regression model for repeated observations. I also compare it to the Markov regression model on the basis of goodness of fit. This comparison method has the limitation that it does not indicate formally whether one model is significantly better than the other.

A reasonable comparison between the two alternative models can only be done by fitting them to the same data. The data used to fit the Markov regression model is a subset of the one used to fit the negative binomial regression model for repeated observations. The largest data set refers to a longer follow-up period in which additional observations were obtained for some individuals and new patients entered the PsA study. Also, observations corresponding to 10 or more damaged joints were retained in the largest data set to fit the negative binomial regression model.

Use of the smallest data set for comparative purposes would disregard valuable information recorded in the extended follow-up period. On the other hand, the largest data set contains observations (transitions within the absorbing state) that are not used to fit the Markov regression model. Therefore, a third data set was created by eliminating from the largest one patients with 10 or more damaged joints in their first clinic visit and by omitting observations after 10 or more damaged joints were reached. In this way, a data set with 254 patients and 1455 transitions was obtained.

The Markov and the negative binomial regression models examined in previous chapters differ in several aspects. The response variable of the two models are different. In the negative binomial model for repeated observations the response is the increase in the number of damaged joints between consecutive assessments, so it is a discrete variable. In the Markov regression model, the response variable is the damage state occupied by an individual at every clinic visit, so it is a categorical (ordinal) variable.

The linear predictor of the two models is also different. In chapters 1 and 2, a different intercept was used in the model for each transition rate and all the prognostic factors were dichotomized. However, in chapter 3, the rate of damage between clinic visits depends on a baseline value and on the number of damaged joints recorded until the most recent assessment. Also, the erythrocyte sedimentation rate and the number of effused joints were analysed as discrete variables. In this chapter I redefine the linear predictor of the negative binomial regression model for repeated observations so that it resembles that of the Markov regression model. Therefore, the baseline value and the number of damaged joints recorded up to the last clinic visit, $J_{i,j}$, are replaced by an intercept that depends on whether $J_{i,j}$ is equal to zero, between 1 and 4 or in the range 5 to 9. The erythrocyte sedimentation rate and the number of effused joints recorded on the first clinic visit were categorized as Gladman, Farewell and Nadeau [1] did for the Markov regression model.

The techniques described in chapter 2 are applied to evaluate the goodness-of-fit of the Markov regression model and of the negative binomial regression model for repeated observations. This means that a Pearson-type goodness-of-fit statistic is calculated to measure the discrepancy between the observed and the expected counts. Its significance level is computed using bootstrap methodology.

The theoretical material related to the negative binomial regression model for repeated observations examined in this chapter is presented in sections 4.2, 4.3, and 4.4. These sections can be avoided by readers without a statistical background.

4.2 Description of the model

The response variable, $D_{i,j}$, is the number of joints presenting damage between consecutive clinical assessments. Here, I denote the covariate vector in the same way as in chapter 1, *i.e.* $z_i' = (1, z_{i,1}, \ldots, z_{i,p-1})$. Based on the results obtained in the previous chapter, the distribution of $D_{i,j}$ given z_i' is assumed to be well approximated by a negative binomial regression model with dispersion parameter γ and mean:

$$\mu_{i,j} = (t_{i,j+1} - t_{i,j}) \exp(\delta_{0a} + \sum_{u=1}^{p-1} \delta_u z_{i,u}) \tag{4.1}$$

where $a = 1, 2, 3$. The baseline value of the linear predictor is δ_{0a} if $J_{i,j}$ (the number of damaged joints observed up to time $t_{i,j}$) is contained in set S_a where $S_1 = \{0\}$, $S_2 = \{1, 2, 3, 4\}$, and $S_3 = \{5, 6, 7, 8, 9\}$. The likelihood function for the estimation of γ and $\delta' = (\delta_{01}, \delta_{02}, \delta_{03}, \delta_1, \ldots, \delta_{p-1})$ is:

$$L(\gamma, \delta) = \prod_{i=1}^{n} \prod_{j=1}^{m_i} \frac{\Gamma(d_{i,j} + \gamma^{-1})}{d_{i,j}! \Gamma(\gamma^{-1})} \left(\frac{\gamma \mu_{i,j}}{1 + \gamma \mu_{i,j}} \right)^{d_{i,j}} \left(\frac{1}{1 + \gamma \mu_{i,j}} \right)^{\gamma^{-1}}$$

4.3 The goodness of fit statistic

The expected counts are estimated by first calculating $e_{i,j}(k)$, the probability of an increase of k ($k = 0, 1, 2, \ldots$) damaged joints for patient i in an interval of length $t_{i,j+1} - t_{i,j}$ where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m_i$, *i.e.*

$$e_{i,j}(k) = \hat{P}(D_{i,j} = k \mid z_i') = \frac{\Gamma(k+\hat{\gamma}^{-1})}{k!\Gamma(\hat{\gamma}^{-1})}\Big(\frac{\hat{\gamma}\hat{\mu}_{i,j}}{1+\hat{\gamma}\hat{\mu}_{i,j}}\Big)^k\Big(\frac{1}{1+\hat{\gamma}\hat{\mu}_{i,j}}\Big)^{\hat{\gamma}^{-1}}$$

Ideally, for the first observation period the probabilities $e_{i,1}(k)$ should be grouped according to the value of k, the covariate pattern, and the time elapsed between the first two clinical assessments. This process should be repeated for each observation period $j = 1, 2, \ldots, m_{i-1}$. In panel data, the heterogeneity between the number of assessments (m_i) and the variability between the time elapsed between clinical visits $(t_{i,j+1} - t_{i,j})$ can be a hinderance in applying such procedure. In a first instance, the observation period can be ignored as the proposed model assumes that $D_{i,j}$ depends solely on the distance: $t_{i,j+1} - t_{i,j}$, regardless of the time at which the assessments are made. However, in order to detect different departures from the hypothesised model it is advisable to group the $e_{i,j}(k)$ in several ways by considering different classification criteria in each case.

Here I classify the estimated probabilities $e_{i,j}(k)$ according to the value of the response variable $(D_{i,j})$, the quantiles of the time elapsed between successive observations, and the covariate pattern. Theoretically, the response variable can assume an infinite number of values. These should be grouped into R levels with approximately the same number of observations. These levels will be indexed by the letter r $(r = 1, 2, \ldots, R)$. The letter L will denote the number of categories obtained by calculating the quantiles of $t_{i,j+1} - t_{i,j}$; let l represent the l-th category. The lower and upper bounds of category l are, respectively, the $\frac{l-1}{L}100\%$ and the $\frac{l}{L}100\%$ quantiles of $t_{i,j+1} - t_{i,j}$. Analogously, C will denote the total number of groups defined by the partition of the covariate space; c will refer to the c-th group.

Let e_{lrc} be the sum over all the $e_{i,j}(k)$ such that k is contained in level r, z_i' belongs to group c, and the width of the interval $(t_{i,j}, t_{i,j+1})$ is contained in category l. Similarly, n_{lrc} will represent the total number of response variables whose value is contained in level r, associated to a covariate vector in group c and a time interval contained in category l. The ratios $(n_{lrc} - e_{lrc})^2/e_{lrc}$

measure the discrepancy between the observed and expected counts. The sum of these ratios is defined as the goodness of fit statistic for the negative binomial regression model for repeated observations.

$$T_{NB} = \sum_{l=1}^{L} \sum_{r=1}^{R} \sum_{c=1}^{C} \frac{(n_{lrc} - e_{lrc})^2}{e_{lrc}} \tag{4.2}$$

The contingency table of observed and expected counts has $L \times R \times C$ cells but only $L \times (R-1) \times C$ are independent because $\sum_{k=0}^{\infty} e_{i,j}(k) = 1$ so $\sum_{r=1}^{R} e_{lrc}$ has a fixed value.

4.4 Testing the adequacy of the model

In this section I describe the bootstrap algorithm used to calculate a significance level for the statistic (4.2). The idea is to estimate the distribution of the goodness of fit statistic under the hypothesis that the negative binomial regression model for repeated observations fits the data. This is accomplished by simulating several data sets from the hypothesized model. Each data set is then used to estimate the parameters of the model defined by the null hypothesis and the goodness of fit statistic.

The number of damaged joints for patient i produced by the bootstrap algorithm up to time $t_{i,j}$ is denoted as $J^*_{i,j}$. It is assumed that $J^*_{i,1} = J_{i,1}$, where $J_{i,1}$ is the number of damaged joints recorded for patient i in the first clinic visit. The increase in the number of damaged joints given by the bootstrap algorithm, $D^*_{i,j}$, is obtained by simulating an observation from the negative binomial regression model with parameters $\hat{\gamma}$ and $\hat{\mu}_{i,j} = (t_{i,j+1} - t_{i,j}) \exp(\hat{\delta}_{0a} + \sum_{u=1}^{p-1} \hat{\delta}_u z_{i,u})$. The value of $J^*_{i,j+1}$ is then computed as $J^*_{i,j+1} = J^*_{i,j} + D^*_{i,j}$. If $J^*_{i,j+1} \leq 9$ and $j+1 < m_i$ then $D^*_{i,j+1}$ and $J^*_{i,j+2}$ are calculated in the same way, otherwise the process is stopped. Thus, the number of response variables generated for patient i, s_i, is less than or equal to the total number of clinical assessments (m_i). Once a sequence of observations has been generated for each patient, the negative binomial regression model with mean given by (4.1) is fitted to the bootstrap data and statistic

Table 4.1: Estimated parameters and asymptotic standard deviations for the negative binomial regression model for repeated observations (third PsA data set).

Parameter		Estimates	
Dispersion parameter		8.9795	(0.8706)
Baseline constant for the number of damaged joints up to the last assessment			
	Equal to 0	-1.2135	(0.1684)
	Between 1 and 4	-0.4877	(0.3052)
	Between 5 and 9	0.4121	(0.3002)
ESR,	< 15 mm/h	-0.3410	(0.2046)
Num. effused joints,	≥ 5	0.4114	(0.2611)
Disease modifying drugs,	Yes	0.7647	(0.3314)
Use of corticosteroids,	Yes	0.4345	(0.2113)
Initial state 2		-0.0189	(0.3006)
Initial state 3		-1.4083	(0.5087)

(4.2) is calculated. This process is repeated several times. The p-value is the proportion of bootstrap statistics greater than the value of the statistic obtained for the original data.

4.5 Goodness-of-fit for the negative binomial regression model for repeated observations

The negative binomial regression model for repeated observations fitted to the third PsA data set gives the estimated parameters and asymptotic

Table 4.2: Relative rates of damage for each prognostic factor in the negative binomial regression model for repeated observations (third PsA data set).

Prognostic factor	Relative rate of damage
Erythrocyte sedimentation rate	
≥ 15 mm/hr	1
< 15 mm/hr	0.711
Number of effused joints	
< 5	1
≥ 5	1.509
Disease modifying drugs	
No	1
Yes	2.148
Use of corticosteroids	
Yes	1
No	1.544

standard deviations (in brackets) shown in Table 4.1.

The table shows that the prognostic factors for type of medication are the only ones with a significant effect on the rate at which the joints are damaged between clinic visits.

The first baseline value in Table 4.1 indicates that patients with all prognostic factors coded as zero and with zero damaged joints up to the most recent clinic visit develop 0.297 ($= e^{-1.214}$) damage joints within 1 year (or 0.149 damage joints in 6 months). Analogously, individuals with all prognostic conditions equal to zero and having between 5 and 9 damaged joints up to their last assessment develop 1.5 damage joints in one year.

Relative rates should always be viewed as a comparison of two patients with identical inter-visit periods. The relative rates of damage quantify the risk of developing damage joints by comparing two individuals with the same characteristics except that one has certain condition coded as one while the other has that condition coded as zero. For example, Table 4.2 indicates that patients who took disease modifying drugs (DMD) before participating in the study have a risk that is 2.148 times higher of developing damaged joints as compared to patients who did not take DMD. More explicitly, patients with 0 damaged joints on their latest assessment, having less than 5 effused joints and an erythrocyte sedimentation rate of 15 mm/hr or more in their first clinic visit and taking disease modifying drugs (except corticosteroids) before entering the study develop an average of 0.64 ($= e^{-1.214+0.765} = 0.297 \times 2.148$) damage joints in one year. Analogously, subjects with the same characteristics but taking none or nonsteroidal antiinflammatory medications before participating in the study develop an average of 0.297 ($= e^{-1.214}$) damage joints in 1 year. Notice that 0.64/0.297 is equal to 2.148, the relative rate of damage associated to DMD in Table 4.2.

A three dimensional contingency table was constructed to examine the fit of the negative binomial regression model for repeated observations. The estimated probabilities and the observed increases in damaged joints were

Table 4.3: Contingency table of observed and expected counts for the negative binomial regression model for repeated observations (third PsA data set).

Increase in the number of damaged joints

Prognostic	Inter-visit	Zero		One or Two		Three +		Total
Factors	Period (years)	Obs.	Exp	Obs.	Exp	Obs.	Exp	Obs.
Zero	0.038 - 0.479	91	87.38	6	7.45	**0**	**2.18**	97
	0.479 - 0.518	88	79.94	**2**	**7.67**	**0**	**2.39**	90
	0.518 - 0.652	83	83.00	6	8.66	**6**	**3.35**	95
	0.652 - 1.062	88	86.39	9	10.01	4	4.60	101
	1.062 - 9.777	73	74.08	13	10.53	8	9.39	94
One	0.038 - 0.479	116	116.54	10	10.80	5	3.66	131
	0.479 - 0.518	125	128.85	15	13.41	8	5.74	148
	0.518 - 0.652	119	123.97	**19**	**13.57**	6	6.46	144
	0.652 - 1.062	120	111.57	9	13.74	5	8.68	134
	1.062 - 9.777	107	108.42	**23**	**15.94**	12	17.64	142
Two +	0.038 - 0.479	54	55.72	8	5.83	2	2.45	64
	0.479 - 0.518	45	45.99	3	5.42	**6**	**2.60**	54
	0.518 - 0.652	46	44.04	4	5.17	2	2.80	52
	0.652 - 1.062	42	42.84	7	5.98	5	5.18	54
	1.062 - 9.777	33	40.81	9	6.23	**13**	**7.96**	55
Totals		1230	1229.54	143	140.41	82	85.08	1455

Table 4.4: Table of observed and expected counts for the negative binomial regression model for repeated observations (third PsA data set).

Increase in the Num. of Damaged Joints	Observed Count	Expected Count	Scaled Differences
0	1230	1229.545	0.0002
1	96	98.952	0.0881
2	47	41.445	0.7446
3	12	22.764	5.0895
4	11	14.221	0.7296
5	13	9.617	1.1900
6	10	6.871	1.4252
7	6	5.113	0.1539
8	7	3.927	2.4059
9	2	3.092	0.3858
10 +	21	19.454	0.1229
Total	1455	1455.001	12.3356

classified into $R = 3$ levels depending on whether the response variable was equal to an increase of 0 damaged joints (level 1), an increase of one or two damaged joints (level 2), or an increase of 3 or more damaged joints (level 3). Based on the number of prognostic conditions equal to one, the estimated probabilities and the observed increases in damaged joints were classified into $C = 3$ categories. The prognostic factors are the erythrocyte sedimentation rate and the number of effused joints recorded on the first visit as well as the type of medication taken before participating in the study. Category 1 refers to individuals with all prognostic factors coded as zero, category 2 corresponds to subjects with one prognostic factor coded as one, and category 3 refers to patients with 2 or more prognostic conditions equal to one. With respect to the time elapsed between successive clinic visits, the estimated probabilities and the observed increases in damaged joints were classified into $L = 5$ categories defined by the quintiles of the inter-visit periods. (For the third PsA data set, the median and the mean of the inter-visit periods are 0.575 years and 0.995 years respectively.)

Table 4.3 contains the observed and expected counts for the negative binomial regression model for repeated observations. The contingency table has $45 = L \times R \times C$ cells but only $30 = L \times (R - 1) \times C$ are independent. The value of the goodness of fit statistic defined by expression (4.2) is 40.06.

The column totals show that 1230 (84.5%) increments are equal to zero damaged joints, 143 (9.8%) increments are equal to 1 or 2 damaged joints, and the remaining 82 (5.6%) increments are equal to 3 or more damaged joints. This suggests that, in PsA, the process leading to damaged joints is slow. For some time, patients have none or a constant number of damaged joints.

From a clinical point of view it is reasonable to expect that more damaged joints occur the longer the time elapsed between visits. This is reflected in the columns for an increase of one or two and three or more damaged joints where the cells containing more observations correspond to patients assessed over periods longer than 1.062 years - regardless of the number of prognostic factors coded as one [1].

Nearly 50% (699 × 100%/1455) of the observed increments in damaged joints correspond to patients with one prognostic factor coded as one.

There is good agreement between the observed and expected counts in the first column of Table 4.3. The bold numbers in the next columns highlight the observed and expected counts that contribute with more than two units to the value of the goodness of fit statistic. Note that the column that refers to increments of one or two damaged joints has three cells with bold numbers while the column containing increments of three or more damaged joints has five cells with bold numbers. Thus, the negative binomial regression model produces accurate predictions for increases of zero damaged joints but not for increments of one or more damaged joints.

The highlighted counts in the first two rows of Table 4.3 indicate that the negative binomial regression model overestimates the number of response variables equal to one or more damaged joints when the time elapsed between visits is less than 6 months and all the prognostic factors are equal to zero. In all the other cells with bold numbers, the expected count is smaller than the observed count. Most of these cells contain data of patients having at least one prognostic factor coded as one and intervals between visits of 6 months or more. This non-random pattern of observed and expected counts suggests that patients with a large number of damaged joints tend to delay their visit to the PsA clinic. Perhaps these patients are reluctant to see the clinician unless they feel bad.

The p-value obtained by generating one thousand bootstrap data sets from the estimated negative binomial regression model for repeated observations is 0.076 = 76/1000. Then, although the proposed model is not entirely satisfactory it is not rejected at the 5% and 1% critical levels usually used in practice.

[1] The distribution of the inter-visit periods greater than 1.062 years is the following: 134 (9.21%) time intervals are between 1.062 and 2 years while 34 (2.34%) elapsed times are greater than 5 years.

Another contingency table was constructed to examine the fit of the negative binomial regression model for repeated observations. The estimated probabilities and the observed number of increments in damaged joints were grouped into eleven categories defined by the values of the response variable, regardless of the covariate pattern and the time elapsed between assessments. The observed and expected counts as well as their weighted discrepancies are shown in Table 4.4. The data set contains 59 (4.05%) increments between 3 and 8 damaged joints. Their observed counts range from 13 to 6 while the expected counts decrease from 22.76 to 3.93. The goodness of fit statistic, defined as the sum of the scaled differences, is equal to 12.34. The smallest contribution to the goodness of fit statistic comes from the category containing increments of zero damaged joints. The largest contributions to the statistic correspond to increments of 3 and 8 damaged joints. In the first case, the expected count is almost twice as big as the observed count while in the second case the expected count is nearly half the value of the observed count. The one thousand bootstrap data sets used to calculate the p-value for Table 4.3 were used to compute the significance level of the test statistic associated to Table 4.4. The p-value thus obtained is 0.199 = 199/1000. As previously concluded, the negative binomial regression model for repeated observations is not rejected.

Tables 4.4 and 3.5 suggest that the overall fit of the negative binomial models examined here and in the previous chapter are similar although they were fitted to different data sets.

Table 4.5: Estimated parameters and standard deviations for the Markov regression model fitted to the third PsA data set.

	Transition Rates		
Parameter	1 → 2	2 → 3	3 → 4
Intercept	-2.46 (0.14)	-1.65 (0.20)	-1.30 (0.22)
Effused joints ≥ 5	0.31 (0.18)	0.31 (0.18)	0.31 (0.18)
ESR < 15 mm/h	-0.37 (0.18)	-0.37 (0.18)	
Corticosteroids, Yes	0.35 (0.15)	0.35 (0.15)	0.35 (0.15)
Disease modifying drugs, Yes	0.48 (0.21)	0.48 (0.21)	0.48 (0.21)
Initial state 2		-0.60 (0.25)	-0.36 (0.30)
Initial state 3			-1.30 (0.45)

4.6 Results for the Markov regression model

The Markov regression model fitted to the third PsA data yields the estimated parameters and asymptotic standard deviations shown in Table 4.5. The estimates are comparable to the ones obtained by Gladman, Farewell, and Nadeau, see Table 1.6.

The contingency table of observed and expected counts is presented in Table 4.6. It has $90 = 3 \times 5 \times 6$ cells but only 45 are independent. The test statistic is equal to 72.161. The bold numbers highlight the observed and expected counts that contribute with more than two units to the goodness of fit statistic. The majority of the highlighted counts are found in the group with one prognostic condition equal to one. As in Table 2.2, the majority of the highlighted observed counts are bigger than their expected count. Only three pairs of bold numbers are such that the observed count is smaller than the expected count. They are located in the group with one prognostic factor coded as one and have an elapsed time between observations greater than 0.65 years.

Table 4.6: Contingency table of observed and expected counts for the Markov regression model fitted to the third PsA data set.

Prognostic Factors	Inter-visit Period (years)	Count	Transition					
			1 → 1	1 → 1	2 → 2	2 → 2	3 → 3	3 → 3
Zero	0.038 - 0.479	Obs	62	1	18	0	15	1
		Exp	61.00	2.00	17.08	0.92	14.73	1.27
	0.479 - 0.518	Obs	60	1	20	0	9	0
		Exp	58.44	2.56	18.57	1.43	7.94	1.06
	0.518 - 0.652	Obs	53	**7**	19	2	13	1
		Exp	57.10	**2.90**	19.59	1.41	12.21	1.79
	0.652 - 1.062	Obs	59	4	28	1	6	3
		Exp	58.75	4.25	26.37	2.64	7.31	1.69
	1.062 - 9.777	Obs	52	11	23	3	3	2
		Exp	51.29	11.71	19.80	6.20	2.99	2.01
One	0.038 - 0.479	Obs	65	1	38	**6**	18	3
		Exp	63.70	2.30	41.55	**2.45**	19.21	1.79
	0.479 - 0.518	Obs	69	5	30	**6**	33	5
		Exp	70.74	3.26	33.51	**2.49**	34.29	3.71
	0.518 - 0.652	Obs	69	**9**	34	4	25	3
		Exp	73.97	**4.03**	34.86	3.14	24.51	3.49
	0.652 - 1.062	Obs	54	3	46	**2**	28	**1**
		Exp	52.99	4.013	42.28	**5.72**	23.90	**5.10**
	1.062 - 9.777	Obs	70	14	30	8	18	**2**
		Exp	64.13	19.87	26.75	11.25	13.97	**6.03**
Two +	0.038 - 0.479	Obs	21	**4**	28	1	8	2
		Exp	23.95	**1.05**	26.91	2.09	9.14	0.86
	0.479 - 0.518	Obs	16	2	25	2	7	2
		Exp	17.07	0.93	24.91	2.09	8.20	0.80
	0.518 - 0.652	Obs	27	0	17	2	5	1
		Exp	25.47	1.53	16.98	2.02	4.97	1.03
	0.652 - 1.062	Obs	14	3	22	1	10	4
		Exp	15.50	1.50	19.74	3.26	10.94	3.06
	1.062 - 9.777	Obs	21	6	14	8	4	2
		Exp	20.40	6.60	15.28	6.72	3.43	2.57
Total		Obs	712	71	392	46	202	32

The significance level of the test statistic associated with Table 4.6 was obtained by the bootstrap algorithm described in section 2.6. One thousand bootstrap data sets were generated independently from the ones used to test the goodness of fit of the negative binomial regression model for repeated observations. Twenty of the resulting contingency tables had cells with zero observed and expected counts. These cells were ignored in the calculation of the bootstrap goodness of fit statistic. Only 8 bootstrap statistics were greater than 72.161, therefore the p-value is $0.008 = 8/1000$. As concluded in chapter 2, the Markov regression model does not give an adequate description of the PsA data.

4.7 Conclusions

Ten parameters were estimated to fit the Markov and the negative binomial regression models. The Markov regression model has three parameters, the transition rates, that describe the progression in damage while the negative binomial model has only one parameter for this purpose. Nevertheless, the only difference between the linear predictors of the two models is the way in which the effect of the erythrocyte sedimentation rate (ESR) was expressed. In the Markov regression model it was assumed that the ESR has no effect on the progression of damage once a patient had reached state 3 (*i.e.* once the individual has 5 or more damaged joints). However, in the negative binomial regression model it was assumed that the effect of the ESR is constant, regardless of the number of damaged joints observed up to the last clinic visit. This discrepancy between the linear predictors of the two models is unlikely to produce a substantially better fit for any model.

The contingency table constructed for the Markov regression model, Table 4.6, has 90 cells while that for the negative binomial regression model, Table 4.3, has 45 cells. The difference is due to the way in which the values of the response variable were classified. In the Markov regression model, the response variable was classified into six categories depending on whether or not a change of state was observed. Only three of these six categories are independent so the contingency table has 45 independent cells. For the negative binomial regression model, the response variable was classified into 3 levels defined by the increase in the number of damaged joints between consecutive assessments. The observed and expected counts in any one level depend on the values obtained for the other two levels. Consequently, the contingency table has 30 independent cells. The difference in size of the two contingency tables explains why only for the Markov regression model some bootstrap data sets produced cells with zero observed and expected counts.

Empty cells can be avoided by reducing the number of categories of one or more classification criteria. For example, transitions to the same damage state (*i.e.* $1 \to 1$, $2 \to 2$, and $3 \to 3$) can be grouped in one category while the rest of the transitions can be classified together in another category. This means that the estimated transition probabilities and the observed states would be classified into $R = 2$ classes instead of six. Such a coarse classification can hide patterns in the expected counts and produce a spurious significance level. Furthermore, empty cells and small expected counts should be avoided when the asymptotic distribution of the test statistic is used but it is less critical when the significance level is simulated.

The total number of observed and expected counts in the 3 groups that define the partition of the covariate space and in the 5 classes in which the time elapsed between visits were classified coincide in Tables 4.3 and 4.6. The marginal totals for the response variable are not comparable. In Table 4.6, a total of $1306 = 712 + 392 + 202$ transitions are contained in the categories referring to no change in damage state. This total is greater than 1230, the number of increments equal to zero damaged joints in Table 4.3. This discrepancy is caused by the fact that a transition from state 2 to state 2 represents an increase of 0, 1, 2, or 3 damaged joints. Analogously, a transition from state 3 to state 3 is equivalent to an increase of 0, 1, 2, 3, or 4 damaged joints.

The number of zero increments in damaged joints, representing 84.5% of the data, are well described by the negative binomial regression model for repeated observations. On the other hand, 89% of the observations referring to sequences of data in which a particular damage state is repeatedly recorded are adequately described by the Markov regression model. Thus, the Markov model fits well observations in which the number of damaged joints recorded at consecutive visits is the same or different provided that they belong to the same damage state.

Few highlighted observed counts are smaller than their expected count in Tables 4.3 and 4.6. These cases are concentrated in one category defined by the prognostic factors and correspond to patients observed either for long (in the Markov model) or short (in the negative binomial model) periods. The highlighted observed counts that are greater than their expected count correspond to increments greater than zero damaged joints. This means that the high rate of damage experienced by some patients was not appropriately described by either of the two models. A covariate or another model can be pursued to explain the behaviour of a small proportion of individuals, 15.5% or less, who experienced a rapid progression in damage. Recall that time varying covariates were not considered and, in the negative binomial model, the effect of some prognostic factors was not significantly different from zero.

Based on Table 4.3, the significance level associated to the negative binomial regression model for repeated observations is 0.076 while that of the Markov regression model is 0.008, based on Table 4.6. Thus, the negative binomial regression model is not rejected like the Markov regression model but this does not mean that the two models give a significantly different fit to the PsA data. Special techniques have been proposed in the literature to compare and, ultimately, to choose between two alternate statistical models. The main obstacle to apply these techniques to compare the Markov and the negative binomial regression models is the difference between their response variables.

Chapter 5

Conclusions

Approximately one third of the patients who attended the Psoriatic Arthritis Clinic at the University of Toronto did not develop damaged joints during the course of the study. The PsA data set is also characterized by having few patients with a large number of damaged joints developing between clinic visits.

Gladman, Farewell and Nadeau proposed the use of a stationary Markov regression model to identify prognostic indicators for disease severity in psoriatic arthritis. The response variable defined by the authors is the damaged state recorded at each clinic visit. In this thesis I propose three alternative models to describe the rate at which joints are damaged between consecutive assessments. The response variable of these models is the increase in the number of damaged joints between clinic visits. The goodness-of-fit analyses suggest that the Markov regression model and the negative binomial regression model for repeated observations are the models that best describe the PsA data sets.

A goodness of fit test of the negative binomial regression model for repeated observations produced a significance level of 0.076 not leading to rejection. The model best describes increments of zero damaged joints between clinic visits. In this model, the erythrocyte sedimentation rate and the number of effused joints recorded at the initial visit do not have a significant role. In contrast, a goodness-of-fit test indicated that the Markov regression model does not provide an adequate explanation of the data leading to a significance level of 0.008. Nevertheless, the model appropriately describes transitions to the same damage state which represent increments of zero or more damaged joints. In this model,the three prognostic factors have a significant effect on the transition rates.

In the negative binomial regression model for repeated observations, 15.5% of the observations represent increases of one or more damaged joints. Some of these observations are not adequately described by the model. They belong to individuals who visited the clinician after 0.652 years when they had experienced an increase of several damaged joints. The time elapsed between the visits of these patients might not be independent of the damage. Analogously in the Markov regression model, the transitions to a different damage state represent 11% of the data. Some of these observations are not well described by the model. They come from patients who experienced a rapid progression in damage and who attended the clinic sooner than expected, in 0.652 years or less. Thus, the visits of these patients might not be randomly spaced on time violating the assumption that the times of clinic visits are independent of the response variable. Summarizing, patients with a large number of damaged joints are not well described by the negative binomial regression model while patients with a rapid progression in damage are not well represented by the Markov regression model.

The Markov and the negative binomial regression models for repeated observations appropriately describe the patients who did not develop damaged joints during the course of the study. No further improvement was obtained by assuming that a subpopulation of individuals with PsA are not susceptible to damaged joints and by fitting a mixture model to the data. This conclusion was reached even though an estimate, significantly greater than zero, was obtained for the proportion of individuals who never develop damaged joints. The Poisson mixture model with added zeros for longitudinal data seems to overestimate such proportion. On the other hand, the estimate produced by the negative binomial mixture model for repeated observations may not have clinical relevance.

A Pearson-type goodness-of-fit statistic was proposed to examine the fit of the models. These statistics are appropriate for categorical data. They are also well known and easy to interpret. General guidelines exist to construct the contingency table. Nevertheless, decisions must always be made concerning the dimension of the table, the number of categories of each classification criterion and the way in which the limits between the categories are defined. These decisions affect the value of the test statistic, its null distribution and the significance level of the test.

Contingency tables of three or four dimensions were defined to examine the goodness of fit of the models fitted to the PsA data. The classification factors considered were: the values assumed by the response variable, the time at which the measurements are made, the time span between observations and the values of the covariates. A method was proposed here to classify panel data according to the time elapsed between observations. This method is an extension of Hosmer and Lemeshow's technique to group the estimated probabilities of a logistic regression model.

When the time elapsed between observations is variable and covariates are measured, the exact distribution of the proposed statistic is intractable. Bootstrap methodology was used to estimate the exact distribution of the test statistic under the null hypothesis. The simulations suggest that if the transition rates do not depend on covariates the estimated distribution is well approximated by a chi-square distribution with degrees of freedom equal to the number of independent cells in the table minus the number of estimated parameters. For Markov regression models, this approximation is not very accurate. To some extent, this may be caused by the existence of cells with small expected values or the omission of cells without observations. For Markov regression models, the true distribution of the test statistic might be approximated by a chi-square with more degrees of freedom than given by the number of independent cells in the contingency table minus the number of estimated parameters.

Models that provide a better fit to the PsA data over a wider range of values of the response variable need to be investigated in the future. A first approach would be to fit a negative binomial regression model for repeated observations with other prognostic factors different from the ones used to fit the stationary Markov regression model. Threshold models have successfully been used to describe univariate observations with a small proportion of non-zero values. These non-parametric models could be generalized to describe correlated data. Alternatively, more sophisticated models can be explored in which the response variable at the next observation time depends on the present outcome and the current covariate values. This means that the effect of the covariates is not assumed to be constant over time but instead varies along with the response variable.

I assumed that, in the Markov regression models, the future damage state depends only on the current state. Similarly, in the models for discrete longitudinal data, the increase in the number of damaged joints between clinic visits depends on the number of damaged joints recorded up to the last assessment. Other approaches to model the correlation between the response variables of an individual can also be considered. For example, the Markov models can be generalized so that the future observation depends not only on the present one but also on past observations. The generalized estimating equations (GEE) approach is frequently used to make inferences about marginal models for response variables. This is accomplished by making weak assumptions about the correlation structure of the data. Random effects models for repeated observations should be considered if the patients with PsA are not homogeneous (*e.g.* because they do not have the same susceptibility to damage joints). In this case, an unobserved random variable, specific to each person, is incorporated into the model. Thus, these models have two random terms, one is the subject-specific effect and the other is the error. Consequently, random effects models are more difficult to fit than the models considered in this thesis.

Tests for separate families of hypothesis might be generalized to compare and choose between two models with different response variables. Also more powerful goodness-of-fit tests, such as those based on likelihood ratio statistics, should be investigated to measure the adequacy of models for discrete response variables measured repeatedly over time. This includes mixture models for longitudinal data.

Appendix A

Formulas for the estimated transition probabilities

When $K = 4$, the transition probabilities $\hat{p}_{i,j(1,2)}$ and $\hat{p}_{i,j(2,3)}$ are obtained by substituting $a = 1$ and $a = 2$ in expression (2.1) respectively. This expression is only valid when $a + 1 < K$ (if K is an absorbing state) so it can not be used to calculate $\hat{p}_{i,j(3,4)}$.

Proceeding in an analogous way as with $\hat{p}_{i,j(a,a+1)}$ it can be shown that $\hat{p}_{i,j(1,3)}$ is calculated as:

$$\int\limits_{0}^{t_{i,j+1}-t_{i,j}} f_{T_{i(1)}} \int\limits_{0}^{t_{i,j+1}-t_{i,j}-t_{i(1)}} f_{T_{i(2)}} \left(1 - \int\limits_{0}^{t_{i,j+1}-t_{i,j}-t_{i(1)}-t_{i(2)}} f_{T_{i(3)}}\, dt_{i(3)}\right) dt_{i(2)}\, dt_{i(1)}$$
$$= \left(\frac{\lambda_{i(3)}}{\lambda_{i(2)} - \lambda_{i(3)}}\right)\left(\frac{\lambda_{i(2)}}{\lambda_{i(1)} - \lambda_{i(2)}}\right)\left[\exp\left(\frac{t_{i,j} - t_{i,j+1}}{\lambda_{i(1)}}\right) - \exp\left(\frac{t_{i,j} - t_{i,j+1}}{\lambda_{i(2)}}\right)\right]$$
$$-\left(\frac{\lambda_{i(3)}}{\lambda_{i(2)} - \lambda_{i(3)}}\right)\left(\frac{\lambda_{i(3)}}{\lambda_{i(1)} - \lambda_{i(3)}}\right)\left[\exp\left(\frac{t_{i,j} - t_{i,j+1}}{\lambda_{i(1)}}\right) - \exp\left(\frac{t_{i,j} - t_{i,j+1}}{\lambda_{i(3)}}\right)\right].$$

The estimated transition probability from state 1 to state 4, $\hat{p}_{i,j(1,4)}$, is obtained by solving:

$$\begin{aligned}
&\int\limits_0^{t_{i,j+1}-t_{i,j}} f_{T_{i(1)}} \int\limits_0^{t_{i,j+1}-t_{i,j}-t_{i(1)}} f_{T_{i(2)}} \int\limits_0^{t_{i,j+1}-t_{i,j}-t_{i(1)}-t_{i(2)}} f_{T_{i(3)}}\, dt_{i(3)}\, dt_{i(2)}\, dt_{i(1)} \\
&= 1-\left(\frac{\lambda_{i(1)}}{\lambda_{i(1)}-\lambda_{i(2)}}\right)\left(\frac{\lambda_{i(1)}}{\lambda_{i(1)}-\lambda_{i(3)}}\right)\exp\left(\frac{t_{i,j}-t_{i,j+1}}{\lambda_{i(1)}}\right) \\
&\quad +\left(\frac{\lambda_{i(2)}}{\lambda_{i(1)}-\lambda_{i(2)}}\right)\left(\frac{\lambda_{i(2)}}{\lambda_{i(2)}-\lambda_{i(3)}}\right)\exp\left(\frac{t_{i,j}-t_{i,j+1}}{\lambda_{i(2)}}\right) \\
&\quad -\left(\frac{\lambda_{i(3)}}{\lambda_{i(2)}-\lambda_{i(3)}}\right)\left(\frac{\lambda_{i(3)}}{\lambda_{i(1)}-\lambda_{i(3)}}\right)\exp\left(\frac{t_{i,j}-t_{i,j+1}}{\lambda_{i(3)}}\right).
\end{aligned}$$

The estimate, $\hat{p}_{i,j(1,1)}$, and in general $\hat{p}_{i,j(a,a)}$ $\forall\ a = 1, 2, \ldots, K-1$, can be calculated as:

$$\hat{p}_{i,j(a,a)} = 1 - \sum_{b>a}^{K} \hat{p}_{i,j(a,b)}$$

or as:

$$\hat{p}_{i,j(a,a)} = P(T_{i(a)} > t_{i,j+1} - t_{i,j}) = 1 - P(T_{i(a)} \leq t_{i,j+1} - t_{i,j}).$$

In either case:

$$\hat{p}_{i,j(a,a)} = \exp\left[\frac{t_{i,j}-t_{i,j+1}}{\lambda_{i(a)}}\right] \qquad \forall \quad a = 1, 2, \ldots, K-1. \qquad \text{(A.1)}$$

Patients in state 2 at time $t_{i,j}$ can be observed in states 2, 3 or 4 at time $t_{i,j+1}$. It has already been explained how to calculate $\hat{p}_{i,j(2,2)}$ and $\hat{p}_{i,j(2,3)}$. The estimated transition probability $\hat{p}_{i,j(2,4)}$ is obtained as follows:

$$\begin{aligned}
\hat{p}_{i,j(2,4)} &= \int\limits_0^{t_{i,j+1}-t_{i,j}} f_{T_{i(2)}} \int\limits_0^{t_{i,j+1}-t_{i,j}-t_{i(2)}} f_{T_{i(3)}}\, dt_{i(3)}\, dt_{i(2)} \\
&= 1+\left(\frac{\lambda_{i(2)}}{\lambda_{i(3)}-\lambda_{i(2)}}\right)\exp\left(\frac{t_{i,j}-t_{i,j+1}}{\lambda_{i(2)}}\right) \\
&\quad -\left(\frac{\lambda_{i(3)}}{\lambda_{i(3)}-\lambda_{i(2)}}\right)\exp\left(\frac{t_{i,j}-t_{i,j+1}}{\lambda_{i(3)}}\right).
\end{aligned}$$

Finally, for patients in state 3, $\hat{p}_{i,j(3,3)}$ is given by expression (A.1) and $\hat{p}_{i,j(3,4)}$ is calculated as:

$$\begin{aligned}\hat{p}_{i,j(3,4)} &= P(T_{i(3)} < t_{i,j+1} - t_{i,j}) \\ &= 1 - \exp\left(\frac{t_{i,j} - t_{i,j+1}}{\lambda_{i(3)}}\right)\end{aligned}$$

Appendix B

Derivation of the Weibull distribution function

Let $T_{i(a)}$ be an exponential random variate with parameter $\lambda_{i(a)} = q^{-1}_{i(a,a+1)}$. Then, $W_i = T^{1/\alpha}_{i(a)}$, where $\alpha > 0$, is a continuous random variable with domain $(0, \infty)$.

$$P(W_i < w_i) = P(T^{1/\alpha}_{i(a)} < w_i) = P(T_{i(a)} < w_i^\alpha) = 1 - \exp\Big(-\frac{w_i^\alpha}{\lambda_{i(a)}}\Big)$$

If $\lambda_{i(a)} = \varrho^\alpha_{i(a)}$ then

$$P(W_i < w_i) = F_{W_i}(w_i; \varrho_{i(a)}, \alpha) = 1 - \exp\Big\{-\Big(\frac{w_i}{\varrho_{i(a)}}\Big)^\alpha\Big\}$$

and

$$f_{w_i}(w_i; \varrho_{i(a)}, \alpha) = \frac{d}{dw_i} F_{W_i}(w_i; \varrho_{i(a)}, \alpha) = \frac{\alpha}{\varrho^\alpha_{i(a)}} w_i^{\alpha-1} \exp\Big\{-\Big(\frac{w_i}{\varrho_{i(a)}}\Big)^\alpha\Big\}$$

Therefore, $W_i = T^{1/\alpha}_{i(a)}$ has a Weibull distribution with parameters α and $\varrho_{i(a)} = \lambda^{1/\alpha}_{i(a)}$.

Bibliography

[1] Gladman, D.D., Farewell, V.T., and Nadeau, C. (1995), "Clinical Indicators of Progression in Psoriatic Arthritis: Multivariate Relative Risk Model", *Journal of Rheumatology*, **22**, 675-679.

[2] Kalbfleisch, J.D., and Lawless, J.F.(1985), "The Analysis of Panel Data Under a Markov Assumption", *Journal of the American Statistical Association*, **80**, 863-871.

[3] Gentleman,R.C., Lawless, J.F., Lindsey, J.C., and Yan,P. (1994), "Multi-State Markov Models for Analysing Incomplete Disease History Data with Illustrations for HIV Disease", *Statistics in Medicine*, **13**, 805-821.

[4] Longini Jr.,I.M., Clark, W.S., Byers,R.H., Ward,J.W., Darrow,W.W., Lemp,G.F., Hethcote, H.W. (1989), "Statistical Analysis of the Stages of HIV Infection Using a Markov Model", *Statistics in Medicine*, **8**, 831-843.

[5] Gladman, D.D., and Farewell, V.T. (1995), "The role of HLA antigens as indicators of disease progression in psoriatic arthritis", *Arthritis and Rheumatism*, **38**, 845-850.

[6] Gladman, D.D., Farewell, V.T., Kopciuk, K.A., and Cook, R.J. (1998), "HLA markers and progression in psoriatic arthritis", *The Journal of Rheumatology*, **25**, 730-733.

[7] Brubacher, B., Gladman, D.D., Buskila, D., Langevitz, P., and Farewell, V.T. (1992), "Followup in psoriatic arthritis: Relationship to disease characteristics", *The Journal of Rheumatology*, **19**, 917-920.

[8] Karlin, S., and Taylor, H.M.(1975), *A First Course in Stochastic Processes*,(2nd ed.),New York: Academic Press, Inc.

[9] Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge: MIT.

[10] Stavola, B. L. de (1988), "Testing Departures from Time Homogeneity in Multistate Markov Processes", *Applied Statistics*, **37**, 242-250.

[11] Hosmer, D.W., and Lemeshow, S. (1989), *Applied Logistic Regression*, New York: John Wiley and Sons, Inc.

[12] D'Agostino, R.B., and Stephens, M.A. (1986), *Goodness-of-Fit Techniques*, New York: Marcel Dekker, Inc.

[13] Efron, B., and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*, New York: Chapman & Hall.

[14] Shao, J., and Tu, D. (1995), *The Jackknife and Bootstrap*, New York: Springer-Verlag.

[15] Johnson, N. L. and Kotz, S. (1969), *Discrete Distributions*, Boston: Houghton Mifflin Company.

[16] Greenwood, M. and Yule, G. U. (1920), "An enquiry into the nature of frequency distributions of multiple happenings, with particular reference to the occurence of multiple attacks of disease or repeated accidents", *Journal of the Royal Statistical Society, Series A*,**83**, 255-279.

[17] Lawless, J. F. (1987), "Negative binomial and mixed Poisson regression",*The Canadian Journal of Statistics*, **15**, 209-225.

[18] Everitt, B. S. and Hand, D. J. (1981), *Finite Mixture Distributions*, New York: Chapman and Hall.

[19] Farewell, V. T. (1986), "Mixture models in survival analysis: Are they worth the risk?",*The Canadian Journal of Statistics*, **14**, 257-262.

[20] McLachlan, G. J. and Basford, K. E. (1988), *Mixture models: inference and applications to clustering*, New York: Marcel Dekker, Inc.

[21] Farewell, V. T. (1977), "A model for a binary variable with time-censored observations",*Biometrika*, **64**, 43-46.

[22] Struthers, C. A. and Farewell, V. T. (1989), "A mixture model for time to AIDS data with left truncation and an uncertain origin", *Biometrika*, **76**, 814,817.

[23] Lambert, D. (1992), "Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing", *Technometrics*, **34**, 1-14.

[24] Ridout, M., Demétrio, C. G. B., and Hinde, J. (1998), "Models for count data with many zeros", International Biometric Conference, Cape Town.

[25] Self, S. G. and Liang, K. (1987), "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions",*Journal of the American Statistical Association*, **82**, 605-610.

[26] Ghitany, M. E., Maller, R. A., and Zhou, S. (1994), "Exponential Mixture Models with Long-Term Survivors and Covariates", *Journal of Multivariate Analysis*, **49**, 218-241.