

# **COUNT DATA MODELLING AND TOURISM DEMAND**

**by**

**Jörgen Hellström**

**Author and title:**

Jörgen Hellström (2002)

Count Data Modelling and Tourism Demand. Umeå Economic Studies No. 584, 100 pages.

**Abstract**

This thesis consists of four papers concerning modelling of count data and tourism demand. For three of the papers the focus is on the integer-valued autoregressive moving average model class (INARMA), and especially on the INAR(1) model. The fourth paper studies the interaction between households' choice of number of leisure trips and number of overnight stays within a bivariate count data modelling framework.

Paper[I] extends the basic INAR(1) model to enable more flexible and realistic empirical economic applications. The model is generalized by relaxing some of the model's basic independence assumptions. Results are given in terms of first and second conditional and unconditional order moments. Extensions to general INAR( $p$ ), time-varying, multivariate and threshold models are also considered. Estimation by conditional least squares and generalized method of moments techniques is feasible. Monte Carlo simulations for two of the extended models indicate reasonable estimation and testing properties. An illustration based on the number of Swedish mechanical paper and pulp mills is considered.

Paper[II] considers the robustness of a conventional Dickey-Fuller (DF) test for the testing of a unit root in the INAR(1) model. Finite sample distributions for a model with Poisson distributed disturbance terms are obtained by Monte Carlo simulation. These distributions are wider than those of AR(1) models with normal distributed error terms. As the drift and sample size, respectively, increase the distributions appear to tend to  $t(T-2)$  and standard normal distributions. The main results are summarized by an approximating equation that also enables calculation of critical values for any sample and drift size.

Paper[III] utilizes the INAR(1) model to model the day-to-day movements in the number of guest nights in hotels. By cross-sectional and temporal aggregation an INARMA(1,1) model for monthly data is obtained. The approach enables easy interpretation and econometric modelling of the parameters, in terms of daily mean check-in and check-out probability. Empirically approaches accounting for seasonality by dummies and using differenced series, as well as forecasting, are studied for a series of Norwegian guest nights in Swedish hotels. In a forecast evaluation the improvements by introducing economic variables is minute.

Paper[IV] empirically studies household's joint choice of the number of leisure trips and the total night to stay on these trips. The paper introduces a bivariate count hurdle model to account for the relative high frequencies of zeros. A truncated bivariate mixed Poisson lognormal distribution, allowing for both positive as well as negative correlation between the count variables, is utilized. Inflation techniques are used to account for clustering of leisure time to weekends. Simulated maximum likelihood is used as estimation method. A small policy study indicates that households substitute trips for nights as the travel costs increase.

**Key words:** Time series, Count data, INARMA, Unit root, Aggregation, Forecasting, Tourism, Truncation, Inflation, Simulated maximum likelihood, Bivariate hurdle model.

**Distribution and author's address:** Department of Economics, University of Umeå, SE-901 87 Umeå, Sweden.

ISSN: 0348-1018

ISBN: 91-7305-215-9

## Acknowledgements

During the process of writing this thesis I have received excellent supervision from Kurt Brännäs. Kurt has been generous with his time and expertise and shown a never ending patience in reading, commenting, and improving earlier drafts of the papers. I'm also very grateful for all help concerning practical matters, as regarding Scientific Workplace and Fortran. Thank You Kurt!

Another person who has contributed greatly to this thesis is my second co-author (apart from Kurt) Jonas Nordström. Particularly for the last paper of the thesis Jonas' help has been invaluable. Thank You Jonas!

A number of other colleagues at the Department of Economics have contributed by commenting earlier drafts of the papers and for this I'm grateful. Thank You Linda Andersson, Thomas Aronsson, Erik Bergkvist, Runar Brännlund, Xavier de Luna, Niklas Nordman, Magnus Wikström!

I would also like to acknowledge Colin Cameron (opponent on my licenciate thesis) whose comments have improved the papers. Niklas Rudholm is thanked for providing me with data and the absolute truth during coffee breaks. Niklas Hanes has provided invaluable practical help (particularly since I always seem to forget files on my computer in Umeå) and been a good friend over the years. Marie Hammarstedt is thanked for providing excellent practical support. To all other colleagues whom I don't mention by names, Thanks for making these years an enjoyable time.

The financial support from J C Kempes Minnes Akademiska forskningsfond and from KFB/VINNOVA is also greatly acknowledged.

Finally, I would like to thank my mother, brother, sister and my life companions Ulrica and Raymond, for love and moral support.

Also: To the kidz in CrossRoads. Thanks for the Rock'n Roll.

Örnsköldsvik, February 25, 2002

Jörgen Hellström

The thesis consists of this summary and the following four papers:

- [I] Brännäs, K. and Hellström, J. (2001). Generalized Integer-Valued Autoregression. *Econometric Reviews* **20**, 425-443.
- [II] Hellström, J. (2001). Unit Root Testing in Integer-Valued AR(1) Models. *Economics Letters* **70**, 9-14.
- [III] Brännäs, K., Hellström, J. and Nordström, J. (2002). A New Approach to Modelling and Forecasting Monthly Guest Nights in Hotels. *International Journal of Forecasting* **18**, 19-30.
- [IV] Hellström, J. (2002). An Endogenously Stratified Bivariate Count Data Model for Household Tourism Demand. *Umeå Economic Studies* **583**. Umeå University.

Papers I-III are reproduced by permission from the journals.

## 1. Introduction

Count data regression modelling techniques have become important tools in empirical studies of economic behavior and their applicability continues to grow in various areas of economics. In the specialized literature, e.g., Cameron and Trivedi (1998) and Winkelmann (2000), numerous examples of economic studies utilizing count data methodology is mentioned, covering topics in, e.g., health economics, labour economics, financial economics, industrial organization, and transportation and travel. Some examples of studies are the number of doctor visits, the number of bank failures, the number of shopping trips as well as studies of takeover activity, entry and exit in industries and labor mobility.

Although count data regression modelling techniques have a rather recent origin, the statistical analysis of counts has a long and rich history. Most of the early statistical count analyses concerned univariate independent and identically distributed random variables within the framework of discrete parametric distributions (Johnson, Kotz and Kemp, 1992). The benchmark for the development of count data models is the Poisson distribution. The Poisson distribution was derived as a limiting distribution of the binomial by Siméon-Denis Poisson, who gave the limiting distribution in his 1837 book. The negative binomial distribution, a standard generalization of the Poisson, was derived by Greenwood and Yule (1920) and by Eggenberger and Polya (1923). In economics an early application of the Poisson regression model is Jorgenson (1961), who modelled consumer demand. Gouriéroux, Monfort and Trognon (1984ab) and Hausman, Hall and Griliches (1984) are early contributors who have influenced the development of applied count data regression analysis.

Economic studies usually begin with a specification of a theoretical model trying to explain agents' (households, firms or other institutions) behavior or choice as depending on other variables. From the theoretical model an empirically feasible regression model is formulated. The regression model is an empirical counterpart to the theoretical model where the choice or outcome on the variable of interest, is explained by explanatory variables specified by theory. In count data regression, the main focus is the effect of covariates on the frequency of an event, measured by non-negative and integer-valued counts.

The main motivation for using specialized methods to handle count data is that the standard Gaussian linear regression model ignores the restricted support (non-negativity and the integer-valued character) of the dependent count data variable. This may lead to significant deficiencies unless the mean of the counts is high, in which case the normal approximation and related regression methods may be satisfactory. The results in Paper [II] gives an example of the latter since it indicates that as the mean in the count data model increase, the better is the approximation of the model based on normality.

From an econometric point of view this thesis focuses on the modelling of time series of count data and on the analysis of repeated and bivariate cross-sectional count data.

## 2. The Poisson regression model

The starting point for most cross-sectional count data analyses is the Poisson regression model. The model is based on the assumption that the count variable  $y_i$ , given the vector of explanatory variables  $\mathbf{x}_i$ , is independently Poisson distributed with density

$$f(y_i|\mathbf{x}_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \dots$$

with exponential mean function

$$\mu_i = \exp(\mathbf{x}_i\boldsymbol{\beta}), \quad \mu_i \geq 0,$$

where  $\boldsymbol{\beta}$  is a  $k \times 1$  parameter vector. The choice of the exponential mean function is made to insure a non-negative mean. The model implies that the conditional mean and the conditional variance are equal, i.e.  $E(y_i|\mathbf{x}_i) = V(y_i|\mathbf{x}_i) = \mu_i$ . Note that the Poisson regression model is intrinsically heteroskedastic.

A number of estimators (e.g., Cameron and Trivedi, 1998, ch. 3) can be used to obtain estimates of  $\boldsymbol{\beta}$ , e.g., maximum likelihood (ML) and pseudo maximum likelihood (PML).

For a sample of  $K$  observations the log-likelihood function for the Poisson regression model is given by

$$l = \sum_{i=1}^K \ln f(y_i|\mathbf{x}_i) = \sum_{i=1}^K y_i \mathbf{x}_i \boldsymbol{\beta} - \exp(\mathbf{x}_i \boldsymbol{\beta}) - \ln y_i!.$$

The ML estimator is consistent and asymptotically normally distributed if the data generating process is Poisson. The assumption of correct underlying distribution may, however, be relaxed. This leads to a class of pseudo- and quasi-ML estimators (Gouriéroux et al., 1984b, McCullagh and Nelder, 1983). The robustness of these estimators depends on a correctly specified conditional mean function.

## 3. Generalizations of the basic Poisson model

In many empirical applications the assumptions underlying the basic Poisson regression model are too restrictive with regard to features of the observed data (Cameron and Trivedi, 1998, ch. 4). Some commonly encountered deviations, from the basic assumptions, and those relevant for this thesis are briefly reviewed in this summary.

A common deviation from the basic Poisson model is the failure of the equal conditional mean and conditional variance restriction. If the conditional variance of the data exceeds the conditional mean overdispersion is present. The most commonly used explanation for overdispersion is that unobserved heterogeneity is present, i.e. there are omitted variables in the mean function. Other explanations are measurement errors in explanatory variables and that structural parameters are random (Brännäs and Rosenqvist, 1994).

A second frequently encountered feature of empirical data is a higher relative frequency of zero observations than is consistent with any standard count regression model (Mullahy, 1986, Lambert, 1992). From a statistical point of view two common approaches dealing with the "excess zero" problem, are to use a hurdle model or a zero-inflated model.

A third common deviation from the basic Poisson model is truncation and censoring. The observed counts are left truncated if small counts are excluded from the sample (zero truncation being the most frequently encountered) and right censored if counts exceeding some value are aggregated.

### 3.1. Unobserved heterogeneity and overdispersion

To deal with unobserved heterogeneity and overdispersion one may allow for random variation in the conditional mean by introducing a multiplicative error term. This generates a family of models usually referred to as mixed Poisson models. The mean function of the Poisson distribution conditional on a multiplicative error term is specified as

$$\lambda_i = \mu_i v_i.$$

The random variables,  $v_i$ , representing random unobserved heterogeneity, are usually assumed independent and identically distributed, possibly with a known parametric distribution, and independent of the explanatory variables. The marginal density of the mixed Poisson family of models is for a continuous mixing distribution of the form

$$h(y_i|\mathbf{x}_i) = \int f(y_i|\mathbf{x}_i, v_i)g(v_i) dv_i.$$

If  $f(\cdot)$  and  $g(\cdot)$  are conjugate families the resulting model is expressible in a closed form. The specific form of the mixed Poisson distribution depends on the choice of  $g(v_i)$ . Examples of mixture models derived, by different assumptions regarding the mixing density  $g(\cdot)$ , are Poisson inverse Gaussian mixture (Dean, Lawless and Willmot, 1989), discrete lognormal (Shaban, 1988), generalized Poisson (Consul, 1989, Consul and Jain, 1973) and Gauss-Poisson (Johnson, Kotz and Kemp, 1992). Also, one of the oldest and frequently applied alternatives to the basic Poisson regression model, the negative binomial model, may be derived as a Poisson-gamma mixture (Greenwood and Yule, 1920).

Mixing based on multiplicative heterogeneity implies two important features of mixed models. One is that the conditional variance exceeds the conditional variance of the parent Poisson distribution, i.e. overdispersion is explained by the neglect of unobserved heterogeneity. The other feature is that mixing causes the proportion of zero counts to increase compared with the corresponding proportion of zeroes in the parent Poisson distribution. This is a fact irrespectively of the form of  $g(\cdot)$  for nondegenerate distributions (Feller, 1943, Mullahy, 1997).

The estimation of parameters in mixed Poisson models is feasible with ML estimation techniques for mixtures that generate closed form marginal distributions. For mixtures

without closed form marginal distribution, computer intensive methods such as simulated maximum likelihood (SML) or simulated method of moments (SMM) are feasible estimation methods. For further details about simulation based estimation, see, e.g., McFadden (1989), Gouriéroux and Monfort (1991), Hajivassiliou and Ruud (1994). Paper [IV] of the thesis deals with a bivariate mixture model without explicit closed form solution and SML is used as the estimation method.

### 3.2. Truncated counts

A count data distribution is truncated if the distribution is not observable over the whole range of non-negative integers.<sup>1</sup> A common example are distributions "truncated at zero". Such a distribution arises in "on-site" sampling (Shaw, 1988). Since the survey is conducted only among those participating in the event, only positive integer values can be observed. The zero truncation case is usually referred to as left truncation or truncation from below. Right truncation or truncation from above, may also arise.

The general approach to dealing with left truncated<sup>2</sup> count data is to specify the count data density  $h(y_i, \boldsymbol{\theta})$ , with corresponding distribution  $H(y_i, \boldsymbol{\theta}) = \Pr(Y_i \leq y_i)$ , where  $\boldsymbol{\theta}$  is a parameter vector. If observed counts less than a positive integer  $r$  are omitted, the left truncated count distribution is given by

$$f(y_i, \boldsymbol{\theta} | y_i \geq r) = \frac{h(y_i, \boldsymbol{\theta})}{1 - H(r-1, \boldsymbol{\theta})}, \quad y_i = r, r+1, \dots$$

For the Poisson distribution  $h(y_i, \boldsymbol{\theta}) = e^{-\mu_i} \mu_i^{y_i} / y_i!$  the left truncated Poisson distribution is given by

$$f(y_i, \mu_i | y_i \geq r) = \frac{\mu_i^{y_i}}{\left[ e^{\mu_i} - \sum_{j=0}^{r-1} \mu_i^j / j! \right] y_i!}, \quad y_i = r, r+1, \dots$$

The truncated mean and variance of the left truncated Poisson model are given in, e.g., Cameron and Trivedi (1998, ch. 4). The mean of the left truncated random variable exceeds the corresponding mean of the untruncated variable, whereas the variance of the truncated variable is smaller.

The log-likelihood for the left truncated Poisson model is given by

$$l = \sum_{i=1}^K \left[ y_i \ln(\mu_i) - \mu_i - \ln \left( 1 - e^{-\mu_i} \sum_{j=0}^{r-1} \mu_i^j / j! \right) - \ln(y_i!) \right].$$

The ML estimator is consistent and asymptotically normally distributed when the true data generating process is Poisson. The first conditional truncated moment depends on

---

<sup>1</sup>For censored count data observations are also only available for a restricted range of the count variable, but in this case explanatory variables are observed. This is in contrast to truncation where all observations, both the counts and the explanatory variables, are lost for some range of the count variable.

<sup>2</sup>Analogous results may be derived for right truncated distributions.

the correct probability for the non-observable counts, i.e. the probability of observing  $y_i < r$  for the left truncated case. A misspecification of the underlying distribution will therefore lead to a misspecification in the first conditional truncated moment and will then result in an inconsistent parameter estimator.

Paper [IV] of this thesis deals with both left and right truncated data, i.e. the number of trips to a particular destination is only observed for one and, at most, two trips as a consequence of the survey construction.

### 3.3. The count data hurdle regression model

The hurdle count data regression model was developed by Mullahy (1986) and builds on ideas from Cragg (1971). Recent applications include Creel and Loomis (1990), Wilson (1992), Pohlmeier and Ulrich (1995), and Arulampalam and Booth (1997). Paper [IV] of this thesis introduces a bivariate version of the hurdle count data regression model. The hurdle model, which is flexible and allows for both under- and overdispersion, has intuitive appeal thanks to its interpretation as mirroring a two-part decision process. This is a plausible feature for many situations in economics. The two-part model specifies the first part as a binary outcome model and the second part as a zero truncated model. The first part governs participation, i.e. choosing a positive number or not, while the second part governs the positive choice. The count data hurdle model is also one possible approach to dealing with the "excess zero" problem.

The hurdle version of the Poisson regression model is obtained by letting  $\mu_{1i} = \exp(\mathbf{x}_i\boldsymbol{\beta}_1)$  be the mean function for the binary choice model and  $\mu_{2i} = \exp(\mathbf{x}_i\boldsymbol{\beta}_2)$  be the mean function governing the positive outcomes of the positive set  $J = \{1, 2, \dots\}$ . The probabilities of observing a zero count, respectively, a one count (1, for  $y_i > 0$ ), in the binary part of the model, is given by

$$\begin{aligned}\Pr(y_i = 0|\mathbf{x}_i) &= \exp(-\mu_{1i}) \\ \Pr(y_i = 1|\mathbf{x}_i) &= \sum_{y_i \in J} f(y_i|\mathbf{x}_i) = 1 - \exp(-\mu_{1i}).\end{aligned}$$

For the second part of the hurdle model the truncated-at-zero Poisson distribution is given by

$$\Pr(y_i|\mathbf{x}_i, y_i > 0) = \frac{e^{-\mu_i}\mu_i^{y_i}}{(1 - e^{-\mu_i})y_i!}, \quad y_i = 1, 2, \dots$$

The log-likelihood function for the Poisson hurdle model is given by

$$\begin{aligned}l &= l_1 + l_2 \\ l_1 &= \sum_{i=1}^K (1 - I_i) \ln \Pr(y_i = 0|\mathbf{x}_i) + I_i \ln[1 - \Pr(y_i = 0|\mathbf{x}_i)] \\ l_2 &= \sum_{i=1}^K I_i \ln \Pr(y_i|\mathbf{x}_i, y_i > 0),\end{aligned}$$

where  $I_i = 1$  if  $y_i > 0$  and  $I_i = 0$  if  $y_i = 0$ . The joint likelihood is maximized by separately maximizing  $l_1$  and  $l_2$ , assuming functional independence between  $l_1$  and  $l_2$ .

### 3.4. Inflated models

The zero-inflated Poisson (ZIP) model addresses, as the Poisson hurdle model, the "excess zero" problem. The high frequency of zeros is accounted for by allowing extra probability mass at zero and reducing the probability mass for other frequencies. The zero-inflated Poisson model was introduced by Lambert (1992). The model is specified as

$$\begin{aligned}\Pr(y_i = 0) &= \pi_i + (1 - \pi_i)e^{-\mu_i} \\ \Pr(y_i = r) &= (1 - \pi_i)\frac{e^{-\mu_i}\mu_i^r}{r!}, \quad r = 1, 2, \dots\end{aligned}$$

From this specification zeros arise from two independent sources. It is either generated from the Poisson distribution with probability  $(1 - \pi_i)$  or independently by the extra probability  $\pi_i$ . The higher frequencies are reduced by the factor  $(1 - \pi_i)$ . Lambert (1992) parameterized  $\pi_i$  as a logistic function of the observable vector of covariates  $\mathbf{z}_i$ , ensuring non-negativity of  $\pi_i$ . The log-likelihood of the constant  $\pi$  ZIP model is given by

$$l = \sum_{i=1}^K (1 - I_i) \ln(\pi + (1 - \pi)e^{\mu_i}) + I_i(\ln(1 - \pi) - \mu_i + r_i \ln \mu_i - \ln r!),$$

where  $I_i$  is an indicator variable that takes values 1 if  $y_i > 0$  and 0 if  $y_i = 0$ . Lambert used the parameterized version of this model to study the occurrence of defects in manufacturing. Other applications of the model are given by Green (1994), Crépon and Duguet (1997) and Grootendorst (1995). Recent applications include Santos Silva and Covas (2000) and Melkersson and Rooth (2000), who both modelled completed fertility. Santos Silva and Covas combine a hurdle-at-zero model with inflation/deflation at one. Melkersson and Rooth introduce a model inflating both the "zero" and the "two" outcome. In Paper [IV] a model allowing for inflation in a bivariate count hurdle setting is utilized. The model allows for inflation at counts (1, 1), (1, 2), (2, 2), (2, 3) and (2, 4).

### 3.5. Bivariate models for count data

The discussion so far has concerned univariate models. Paper [IV] in this thesis is, however, concerned with a bivariate analysis of count data. This section therefore briefly reviews bivariate count data models.

Formal statistical properties of bivariate discrete models are found in e.g., Kocherlakota and Kocherlakota (1992) and Johnson, Kotz and Balakrishnan (1997). One class of bivariate models may be derived by trivariate reduction techniques, e.g., Kocherlakota and Kocherlakota (1992). Other bivariate models are proposed by, e.g., Gouriéroux, Monfort and Trognon (1984ab), Marshall and Olkin (1990), Delgado (1992), Gurmu and Elder

(2000ab) and Cameron and Johansson (1997). A shortcoming with most of these suggested bivariate models is that the correlation between count variables is restricted to be positive.

A flexible model that allows for both positive as well as negative correlation is the bivariate Poisson lognormal model. The multivariate Poisson lognormal model was suggested by Aitchison and Ho (1989) as an extension to the univariate Poisson lognormal model. Assume that the two count variables  $y_1$  and  $y_2$  have independent Poisson distributions conditional on random unobserved heterogeneity components  $\varepsilon_1$  and  $\varepsilon_2$  and explanatory variables  $\mathbf{x}_1$  and  $\mathbf{x}_2$ :

$$\begin{aligned} y_1 | \mathbf{x}_1, \varepsilon_1 &\sim P(\mu_1), \mu_1 \geq 0 \\ y_2 | \mathbf{x}_2, \varepsilon_2 &\sim P(\mu_2), \mu_2 \geq 0, \end{aligned}$$

where the mean parameters are specified as  $\mu_1 = \exp(\mathbf{x}_1 \boldsymbol{\beta}_1 + \varepsilon_1)$  and  $\mu_2 = \exp(\mathbf{x}_2 \boldsymbol{\beta}_2 + \varepsilon_2)$ . The unobservable variables  $\varepsilon_1$  and  $\varepsilon_2$  are assumed jointly normally distributed, i.e.

$$(\varepsilon_1, \varepsilon_2) \sim N\{(0, 0), (\sigma_1^2, \rho\sigma_1\sigma_2, \sigma_2^2)\}, \quad |\rho| \in [0, 1].$$

The joint density function of  $y_1$  and  $y_2$ , conditional on explanatory variables, is given by

$$\begin{aligned} \Pr(y_1, y_2 | \mathbf{x}_1, \mathbf{x}_2) &= \int \int \Pr(y_1, y_2 | \mathbf{x}_1, \mathbf{x}_2, \varepsilon_1, \varepsilon_2) dF(\varepsilon_1, \varepsilon_2) \\ &= \int \int \Pr(y_1 | \mathbf{x}_1, \varepsilon_1) \Pr(y_2 | \mathbf{x}_2, \varepsilon_2) f(\varepsilon_1, \varepsilon_2) d\varepsilon_1 d\varepsilon_2 \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{\exp(-\mu_1) \mu_1^{y_1}}{y_1!} \frac{\exp(-\mu_2) \mu_2^{y_2}}{y_2!} \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\ &\quad \times \exp\left[-\frac{1}{2(1-\rho^2)} \left\{ \left(\frac{\varepsilon_1}{\sigma_1}\right)^2 - 2\rho\frac{\varepsilon_1\varepsilon_2}{\sigma_1\sigma_2} + \left(\frac{\varepsilon_2}{\sigma_2}\right)^2 \right\}\right] d\varepsilon_1 d\varepsilon_2. \end{aligned}$$

Although there is no apparent simplification of this multiple integral its moments can easily be obtained through conditional expectation results and standard properties of the Poisson and lognormal distribution (Aitchison and Ho, 1989). The sign of the correlation between  $y_1$  and  $y_2$  is determined by the sign of  $\rho$ .

The model lacks a closed form solution for the multiple integral and estimation of the parameters of the model has to be accomplished either by numerical integration or by simulation techniques. Munkin and Trivedi (1999) suggest and study estimation of this model by simulated maximum likelihood. Chib and Winkelmann (2001) use Markov Chain Monte Carlo methods for the same model. Empirical applications in economics of this model are still rare, mainly due to prior difficulties concerning estimation. Paper [IV] in this thesis provides an empirical application of a truncated and inflated version of this model.

#### 4. Time-series models for count data

The literature on count data has mainly focused on cross-sectional analysis. In recent years, however, the increased availability of economic data, especially at the micro level, has stimulated the development of models and methods for analyzing longitudinal (panel) and time series of count data.

A time series of count data is an integer-valued non-negative sequence of count observations observed over time. While there exists a large literature concerning conventional time series modelling the specialization to count data analysis is still not as well developed. This, despite the fact that several models for time series analysis of count data have been proposed. Examples are DARMA models suggested by Jacobs and Lewis (1978ab, 1983), the Zeger and Qaqish (1988) model, serially correlated error models suggested by Zeger (1988) and integer-valued autoregressive moving average (INARMA) models suggested by McKenzie (1985, 1986) and Al-Osh and Alzaid (1987). For a review of these and other models, see, e.g., Cameron and Trivedi (1998, ch. 7). In this thesis the INARMA model class and especially the integer-valued autoregression of order one [INAR(1)] is the focal point and an introduction to the model is given below.

Count data time series approaches have empirically been used in, for example, studies of motorway casualties (Johansson, 1996), the number of bank failures (Davutyan, 1989), the number of homicides in California (Grogger, 1990), and generic competition in the Swedish pharmaceutical market (Rudholm, 2001). As the development of models and methods for analyzing time series of count data proceeds, the number of empirical applications is expected to increase.

##### 4.1. INARMA models

The INARMA model class has evolved as an analogue to the continuous ARMA class of models and builds on the earlier work on non-Gaussian time-series by mainly Jacobs and Lewis (1977). The models were independently proposed by McKenzie (1985, 1986) and Al-Osh and Alzaid (1987). The simplest form of INARMA model is the INAR(1) process defined as

$$y_t = \alpha \circ y_{t-1} + \varepsilon_t,$$

where  $\{\varepsilon_t\}$  is a sequence of non-negative and integer-valued random variables with  $E(\varepsilon_t) = \lambda$  and  $V(\varepsilon_t) = \delta$ . The model resembles the conventional AR(1) model in that it explicitly models the serial correlation in terms of lags of the endogenous variable. The major difference is that scalar multiplication is replaced by a binomial thinning operation. The operator, that was introduced in this framework by Steutel and Van Harn (1979), is defined as  $\alpha \circ y = \sum_{i=1}^y u_i$ , where  $u_i$  is a sequence of binary random variables and where each individual component  $i$ , either survives ( $u_i = 1$ ) with probability  $\alpha$  or dies ( $u_i = 0$ ) with probability  $(1 - \alpha)$ ,  $\alpha \in [0, 1]$ . The basic assumptions that are made in the model are:

$$\text{A.1: } E(u_i u_j) = E(u_i)E(u_j)$$

$$\text{A.2: } E(u_i \varepsilon_t) = E(u_i)E(\varepsilon_t)$$

$$\text{A.3: } E(u_i y_{t-1}) = E(u_i)E(y_{t-1})$$

$$\text{A.4: } \text{Cov}(\varepsilon_t, \varepsilon_s) = 0, \text{ for all } t \neq s$$

$$\text{A.5: } E(\varepsilon_t y_{t-1}) = E(\varepsilon_t)E(y_{t-1}).$$

Properties like conditional and unconditional first and second order moments for the basic INAR(1) model are presented in Paper [I]. Among the attractive features of the model are its relative simplicity and interpretational appeal. Since the  $\{\varepsilon_t\}$  sequence may be seen as the arrival of individual components to the series, with  $\lambda$  as the mean entry, and the binomial thinning operator may be seen to represent the number of survivors from period  $t - 1$  to  $t$ , this makes the model attractive for studying birth-death related processes. Different marginal distributions for  $y_t$  are obtained by different distributional choices for  $\varepsilon_t$ . Al-Osh and Alzaid (1987) consider Poisson distributed  $\varepsilon_t$  [Poisson INAR(1)] resulting in a Poisson marginal distribution for  $y_t$ . Al-Osh and Alzaid (1988) also noted that Poisson is the only distribution for  $\varepsilon_t$  that produces the same distribution for  $y_t$ .

Extensions of the INAR(1) model to INAR( $p$ ) have been considered by Alzaid and Al-Osh (1990) and Jin-Guan and Yuan (1991). The integer valued moving average model of  $q$ th order, INMA( $q$ ) was introduced by Al-Osh and Alzaid (1988) and in a slightly different form by McKenzie (1988). INARMA models can be found in McKenzie (1985) and Al-Osh and Alzaid (1991). Extensions of the models to include explanatory variables are considered by Brännäs (1995) and to panel data situations by Berglund and Brännäs (2001) and Blundell, Griffith and Windmeijer (1999). A finite mixture version of Poisson INAR regression model is proposed by Böckenholt (1999).

Maximum likelihood (conditional and exact) estimation (CMLE and MLE) of the Poisson INAR(1) model is studied in Al-Osh and Alzaid (1987), Brännäs (1994) and Ronning and Jung (1992). Al-Osh and Alzaid (1987) also considered Yule-Walker and conditional least squares (CLS) estimation of the Poisson INAR(1). Brännäs (1994) studied estimation by generalized method of moments (GMM) for the Poisson and generalized Poisson INAR(1) model. Brännäs and Hall (2001) studied estimation of the INMA model.

In the field of economics INARMA models are relatively new. However, some empirical economic applications are Berglund and Brännäs (2001), who studied entry and exit, Blundell et al. (1999), who studied the number of patents in firms, Böckenholt (1999), who studied consumer purchase behavior, and Rudholm (2001), who studied competition in the generic pharmaceuticals market.

As economics, in general, is concerned with describing the behavior of economic agents (households, firms and institutions) or the interaction (dependence) between them, the assumptions of the basic INAR(1) model may seem too restrictive for economic applications.

Paper [I] of this thesis considers relaxing some of the stated independence assumptions (A.1-A.5) in the model. The primary aim is to make the model more readily available for empirical economic applications. The second paper (Paper [II]) deals with testing a special case of the model, namely, whether the INAR(1) has a unit root ( $\alpha = 1$ ). While unit roots have received a lot of attention in the conventional time series literature it has not previously been considered for time series of count data. The third paper (Paper [III]) builds on the basic INAR(1) model and utilizes cross-sectional and temporal aggregation to arrive at an INARMA(1,1) model. The paper considers modelling and forecasting a monthly time series of guest nights in hotels. The paper shows the usefulness of the INAR(1) model, particularly emphasizing its interpretational appeal, as well as offering a new and intuitively appealing approach to the modelling and forecasting of aggregated time series of count data.

## **5. Tourism demand**

An important and continuously growing sector in most economies is the tourism and travel sector. In this thesis two of the papers (Paper [III] and Paper [IV]) are focused on the modelling of tourism related issues.

Paper [III] is concerned with international tourism and focuses on the modelling and forecasting of guest nights in hotels. Although international tourism has grown rapidly over the recent 20 years, the rate of growth has varied considerably from year to year (Nordström, 1999b). For this reason accurate modelling and forecasting has become important for many actors in the tourism industry.

The paper introduces a new approach to the modelling and forecasting of guest nights in hotels that is a mix between economic demand models and pure time series analytical models. Other mixed forms are presented by, e.g., Martin and Witt (1989) and Witt and Witt (1995). Economic demand modelling approaches are given by, e.g., Adamowicz (1994), Melenberg and van Soest (1996) and Nordström (1999a). Pure time series models for forecasting of guest nights in hotels are given by, e.g., Garcia-Ferrer and Queralt (1997) and Gustavsson and Nordström (2001).

Traditionally, the study of consumer demand for leisure- and tourist products/services has received major attention. Various theories and empirical models (e.g., Burt and Brewer 1971, Williams 1979, Stynes and Peters 1984) have been developed over the years to study consumer preferences and choice behavior. These studies have identified the relative importance of the attributes that influence consumer choice behavior. An individual's travel decision over a given period of time includes choices about, e.g., number of trips, time to stay, destination, transportation mode, accommodation and activities.

Many studies concerning tourism and travel focus on one of the travel decisions (e.g., number of trips, time to stay, destination, transportation mode, accommodation and ac-

tivities) independently of the others. Some focus on the demand for trips (Creel and Loomis, 1990, Grogger and Carson, 1991, Terza and Wilson, 1990) others on the choice of destination (Woodside and Lysonski, 1989) and some on the transportation mode choice (Ben-Akiva and Lerman, 1985).

A popular way of modelling trip demand is the travel cost method. It uses the number of trips as a quantity measure and travel costs as the price of the trips.<sup>3</sup> Count data models are a rather recent innovation in this literature (Hellerstein, 1991). Up to date most count data approaches consider single destinations and do not include substitute prices in the demand function (Feather and Shaw, 1999). Paper [IV] of this thesis provides an empirical study including substitute destinations into the demand analyses.

If the decisions about the number of trips and the number of overnight stays or the choice of destination and transportation mode are systematically related, studies considering only one of the decisions may be misleading and fail to account for the interaction between decisions. This fact has received some recent attention in the literature: Hausman, Leonard and McFadden (1995) study the demand for trips (number of trips) and destination choice in a two-stage budgeting approach. Dellaert, Borgers and Timmermans (1997) study the choice of destination along the choice of transportation mode. McConnell (1992), Larson (1993) and Berman and Kim (1999) study the simultaneous choice of number of trips and time to stay for different types of recreational trips. Paper [IV] of this thesis follows this line of research and considers the household's joint choice of the number of trips and the number of nights to stay. In the paper, in contrast to previous studies, count data modelling techniques are utilized.

## 6. Summary of the papers

### Paper [I] Generalized Integer-Valued Autoregression

To enhance the usefulness of INAR models in empirical economic applications we relax some of the independence assumptions (A.1-A.5) of the basic INAR model. The discussion is in terms of entry and exit of firms in a stock of firms model, which serves as a working example.

Dependence between exit decisions ( $E(u_i u_j) \neq E(u_i)E(u_j)$  for  $i \neq j$ ), entry decisions (dependence in the  $\varepsilon_t$  process), entry and exit decisions ( $E(u_i \varepsilon_t) \neq E(u_i)E(\varepsilon_t)$ ) as well as extensions to INAR( $p$ ), multivariate INAR and threshold INAR(1) models are considered. The changes in the dependence structure of the basic INAR model will generally not change the first order moments, but will change higher order moments. Since the INARCH effect as well as the variance properties vary substantially with model type, we argue that empirical discrimination between the model types should be possible.

---

<sup>3</sup>Another popular way of modelling trip demand is the random utility model (RUM).

The cost of the extensions is paid by not being able to obtain full distributional results. For most of the models the first and second conditional and unconditional moments may, however, be obtained. Estimation of the extended models by conditional least squares, generalized method of moments (GMM) and Yule-Walker techniques are discussed as well as issues regarding specification testing and forecasting.

A Monte-Carlo simulation indicates that the models may be estimated satisfactorily by conditional least squares or generalized method of moments techniques. The simulation study also reveals that the dependence parameters are quite precisely estimated and that a Wald test statistic based on GMM have reasonable properties.

An illustration in terms of the number of Swedish mechanical paper and pulp mills is included. The illustrative study indicates a positive but insignificant correlation between exits and a significant positive correlation between the entry and the exit decision.

## **Paper [II] Unit Root Testing in Integer-Valued AR(1) Models**

In this paper we address the question whether the conventional Dickey-Fuller (DF) test of unit root is robust in the integer-valued autoregressive framework. A unit root in the INAR(1) model is considered mainly as an empirical feature. Although, a unit root process may seem unrealistic in the long run, it may be a reasonable characterization in finite empirical samples.

The testing of unit root in the INAR(1) model implies testing whether the first difference follows a random walk with drift. Previous research shows that for non-zero and sufficiently large drifts the DF distribution is inappropriate and that instead the standard normal or  $t$  distributions should be used.

By simulation we obtain the small sample distributions of the DF test for a model with Poisson distributed and hence skewed errors. We also provide a small sample approximation equation to enable calculation of small sample critical values. The results indicate that using DF critical values can be quite misleading. The finite sample distributions have heavier tails than distributions based on normal error terms and the quantiles differ most for small drifts and sample sizes. We further note that negative first differences are inconsistent with a unit root in the INAR(1) model. This provides a simple diagnostic for the presence of unit root.

The paper further provides an illustrative study where a test of unit root is considered for a quarterly time series of the number of generics for different medical substances in Sweden (1972-1986). Using the results of the paper the null hypothesis of a unit root can not be rejected at the 5 percent level. This conclusion is, however, contradicted by the presence of two negative first differences.

**Paper [III] A New Approach to Modelling and Forecasting Monthly Guest Nights in Hotels**

An approach to model and forecast monthly guest nights in hotels is proposed. The approach, which is a mix of time series and econometric modelling, starts from the micro level and models the day-to-day number of guest nights for specific hotels by an INAR(1) model. By cross-sectional and temporal aggregation an aggregate INARMA(1,1) model is obtained. In most empirical applications the AR-part of the INARMA(1,1) can be disregarded and the model then simplifies to an INMA(1) model. The model offers intuitive parameter interpretations in terms of mean-check-in and check-out probabilities.

Empirical specification relating the mean check-out and check-in probabilities to explanatory variables are obtained by a logistic distributional function for the former and an exponential function for the latter.

First, as well as, quarterly and annual differencing is discussed as a method of eliminating trends and seasonal patterns. The models based on the differenced series have the same parameters as the aggregate model based on the undifferenced series, so regardless of whether we prefer estimating at level or at differenced series levels, we may still interpret the parameter estimates in an interesting way. Details concerning estimation, by conditional least squares, and forecasting is also discussed.

The model and the different specifications of it are empirically used to model and forecasts Norwegian guest nights in Swedish Hotels. Strong seasonal patterns are found in the mean-check-in and check-out probabilities and a forecast evaluation against a conventional ARIMA(1,0,1)(0,1,1)<sub>12</sub> shows that the model performs quite well. The gain in forecasting accuracy from including explanatory variables in the model is small.

**Paper [IV] An Endogenously Stratified Count Data Model for Tourism Demand**

The household choice of the number of trips and the total number of nights is empirically studied on Swedish travel data. The study consider trips to the three largest city areas in Sweden: Stockholm, Gothenburg and Malmö.

A bivariate Poisson lognormal model, allowing negative correlation between the count variables, is utilized for the simultaneous estimation of the demand for trips and the demand for nights. Truncation arises from above for the number of trips as counts larger than two are never observed due to the sampling procedure. The total number of nights are truncated "at zero" conditional on households making one trip and "at one" conditional on households making two trips, since the data chosen for the study excludes daytrips. The bivariate Poisson lognormal distribution is adjusted to account for these truncations in the empirical application.

Since an explicit solution for the truncated bivariate Poisson lognormal distribution

can not be obtained the model is estimated by simulated maximum likelihood (SML) techniques.

A bivariate hurdle model is introduced to handle the relatively high frequency of zeros in the sample. In the part of the hurdle model accounting for the positive outcomes count inflation techniques are utilized to allow for possible clustering of counts due to the allocation of leisure time to weekends. The inflated models provide the best fits to the data giving some support for the weekend hypothesis.

For most models the number of trips and nights are negatively correlated. Own price effects in the demand for trips and the demand for nights are negative while cross-price effects show mixed results. A small policy study is provided to illustrate the effects of the model. The policy study reveals the importance of including substitute destination prices in the model as well as illustrates the substitution of trips for nights due to increased transportation costs.

## References

- Adamowicz, W.L. (1994). Habit Formation and Variety Seeking in a Discrete Choice Model of Recreation Demand. *Journal of Agricultural and Resource Economics* **19**, 19-31.
- Aitchison, J. and Ho, C.H. (1989). The Multivariate Poisson Lognormal Distribution. *Biometrika* **76**, 643-653.
- Al-Osh, M.A. and Alzaid, A.A. (1987). First-Order Integer Valued Autoregressive Process. *Journal of Time Series Analysis* **8**, 261-275.
- Al-Osh, M.A. and Alzaid, A.A. (1988). Integer-Valued Moving Average (INMA) Process. *Statistical Papers* **29**, 281-300.
- Al-Osh, M.A. and Alzaid, A.A. (1991). Binomial Autoregressive Moving Average Models. *Communications in Statistics: Stochastic Models* **7**, 261-282.
- Alzaid, A.A. and Al-Osh, M.A. (1990). An Integer-Valued  $p$ th-Order Autoregressive Structure (INAR( $p$ )) Process. *Journal of Applied Probability* **27**, 314-324.
- Arulampalam, W. and Booth, A.L. (1997). Who Gets Over the Training Hurdle? A Study of the Training Experiences of Young Men and Women in Britain. *Journal of Population Economics* **10**, 197-217.
- Ben-Akiva, M. and Lerman, S.R. (1985). *Discrete Choice Analysis, Theory and Application to Travel Demand*. MIT Press, Cambridge, MA.
- Berglund, E. and Brännäs, K. (2001). Plants' Entry and Exit in Swedish Municipalities. *The Annals of Regional Science* **35**, 431-448.
- Berman, M.D. and Kim, H.J. (1999). Endogenous On-Site Time in the Recreation Demand Model. *Land Economics* **75**, 603-19.
- Blundell, R., Griffith, R. and Windmeijer, F. (1999). Individual Effects and Dynamics in Count Data Models. Working Paper W99/3, Institute of Fiscal Studies, London.
- Brännäs, K. (1994). Estimation and Testing in Integer-Valued AR(1) models. *Umeå Economic Studies* **335**.
- Brännäs, K. (1995). Explanatory Variables in the AR(1) Count Data Model. *Umeå Economic Studies* **381**.
- Brännäs, K. and Hall, A. (2001). Estimation in Integer-Valued Moving Average Models. *Applied Stochastic Models in Business and Industry* **17**, 277-291.
- Brännäs, K. and Rosenqvist, G. (1994). Semiparametric Estimation of Heterogenous Count Data Models. *European Journal of Operational Research* **76**, 247-258.
- Burt, O. and Brewer, D. (1971). Estimation of Net Social Benefits from Outdoor Recreation. *Econometrica* **52**, 32-41.
- Böckenholt, U. (1999). Mixed INAR(1) Poisson Regression Models: Analyzing Heterogeneity and Serial Dependencies in Longitudinal Count Data. *Journal of Econometrics* **89**, 317-338.

- Cameron, A.C. and Johansson, P. (1997). Count Data Regressions Using Series Expansions with Applications. *Journal of Applied Econometrics* **12**, 203-224.
- Cameron, A.C. and Trivedi, P. (1998). *Regression Analysis of Count Data*. Cambridge University Press, Cambridge.
- Chib, S. and Winkelmann, R. (2001). Markov Chain Monte Carlo Analysis of Correlated Count Data. *Journal of Business and Economic Statistics* **19**, 428-435.
- Consul, P.C. (1989). *Generalized Poisson Distributions: Properties and Applications*. Marcel Dekker, New York.
- Consul, P.C. and Jain, G.C. (1973). A Generalization of the Poisson Distribution. *Technometrics* **15**, 791-799.
- Cragg, J.G. (1971). Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods. *Econometrica* **39**, 829-844.
- Creel, M.D. and Loomis, J.B. (1990). Theoretical and Empirical Advantages of Truncated Count Data Estimators for Analysis of Deer Hunting in California. *Journal of Agricultural Economics* **72**, 434-441.
- Crépon, B. and Duguet, E. (1997). Research and Development, Competition and Innovation: Pseudo Maximum Likelihood and Simulated Maximum Likelihood Methods Applied to Count Data Models with Heterogeneity. *Journal of Econometrics* **79**, 355-378.
- Davutyan, N. (1989). Bank Failures as Poisson Variates. *Economic Letters* **29**, 333-338.
- Dean, C., Lawless, J.F. and Willmot, G.E. (1989). A Mixed Poisson-Inverse Gaussian Regression Model. *Canadian Journal of Statistics* **17**, 171-182.
- Delgado, M.A. (1992). Semiparametric Generalized Least Squares in the Multivariate Nonlinear Regression Model. *Econometric Theory* **8**, 203-222.
- Dellaert, B.G.C., Borgers, A.W.J. and Timmermans, H.J.P. (1997). Conjoint Models of Tourist Portfolio Choice: Theory and Illustration. *Leisure Science* **19**, 31-58.
- Eggenberger, F. and Polya, G. (1923). Über die Statistik Verketteter Vorgänge. *Zeitschrift für Angewandte Mathematik und Mechanik* **1**, 279-289.
- Feather, P. and Shaw, W.D. (1999). Possibilities for Including the Opportunity Cost of Time in Recreation Demand Systems. *Land Economics* **75**, 592-602.
- Feller, W. (1943). On a General Class of Contagious Distributions. *Annals of Mathematical Statistics* **14**, 389-400.
- Garcia-Ferrer, A. and Queralt, R.A. (1997). A Note on Forecasting International Tourism Demand in Spain. *International Journal of Forecasting* **13**, 539-549.
- Gouriéroux, C. and Monfort, A. (1991). Simulation based Inference in Models with Heterogeneity. *Annales d'Economie et de Statistique* **20/21**, 69-107.
- Gouriéroux, C., Monfort, A. and Trognon, A. (1984a). Pseudo Maximum Likelihood Methods: Theory. *Econometrica* **52**, 681-700.
- Gouriéroux, C., Monfort, A. and Trognon, A. (1984b). Pseudo Maximum Likelihood Methods: Applications to the Poisson Models. *Econometrica* **52**, 701-720.

- Green, W.H. (1994). Accounting for Excess of Zeros and Sample Selection in Poisson and Negative Binomial Regression Models. Discussion Paper EC-94-10, Department of Economics, New York University, New York.
- Greenwood, M. and Yule, G.U. (1920). An Inquiry into the Nature of Frequency Distributions of Multiple Happenings, with Particular Reference to the Occurrence of Multiple Attacks of Disease or Repeated Accidents. *Journal of the Royal Statistical Society* **83**, 255-279.
- Grogger, J. (1990). The Deterrent Effect of Capital Punishment: An Analysis of Daily Homicide Counts. *Journal of the American Statistical Association* **85**, 295-303.
- Grogger, J.T. and Carson, R.R. (1991). Models for Truncated Counts. *Journal of Applied Econometrics* **6**, 225-238.
- Grootendorst, P.V. (1995). A Comparison of Alternative Models of Prescription Drug Utilization. *Health Economics* **4**, 183-198.
- Gurmu, S. and Elder, J. (2000a). Generalized Bivariate Count Data Regression Models. *Economic Letters* **68**, 31-36.
- Gurmu, S. and Elder, J. (2000b). Estimation of Multivariate Count Regression Models With Applications to Health Care Utilization. Mimeo. Economics Department, Georgia State University.
- Gustavsson, P. and Nordström, J. (2001). The Impact of Seasonal Unit Roots and Vector ARMA Modeling on Forecasting Monthly Tourism Flows. *Tourism Economics* **7**, 117-133.
- Hajivassiliou, V. and Ruud, P. (1994). Classical Estimation Methods for LDV Models using Simulation. In Engle, R. and McFadden, D., eds., *Handbook of Econometrics IV*. North-Holland, Amsterdam.
- Hausman, J.A., Hall, B.H. and Griliches, Z. (1984). Econometric Models for Count Data with an Application to the Patents - R and D Relationship. *Econometrica* **52**, 909-938.
- Hausman, J.A., Leonard, G.K. and McFadden, D. (1995). A Utility-Consistent, Combined Discrete Choice and Count Data Model: Assessing Recreational Use Losses due to Natural Resource Damage. *Journal of Public Economics* **56**, 1-30.
- Hellerstein, D.M. (1991). Using Count Data Models in Travel Cost Analysis with Aggregate Data. *American Journal of Agricultural Economics* **73**, 860-866.
- Jacobs, P.A. and Lewis P.A.W. (1977). A Mixed Autoregressive Moving Average Exponential Sequence and Point Process. *Advances in Applied Probability* **9**, 87-104.
- Jacobs, P.A. and Lewis P.A.W. (1978a). Discrete Time Series Generated by Mixtures I: Correlational and Runs Properties. *Journal of the Royal Statistical Society* **B40**, 94-105.
- Jacobs, P.A. and Lewis P.A.W. (1978b). Discrete Time Series Generated by Mixtures II: Asymptotic Properties. *Journal of the Royal Statistical Society* **B40**, 222-228.
- Jacobs, P.A. and Lewis P.A.W. (1983). Stationary Discrete Autoregressive Moving Average Time Series Generated by Mixtures. *Journal of Time Series Analysis* **4**, 19-36.

- Jin-Guan, D. and Yuan, L. (1991). The Integer-Valued Autoregressive (INAR( $p$ )) Model. *Journal of Time Series Analysis* **12**, 129-142.
- Johansson, P. (1996). Speed Limitation and Motorway Casualties: A Time Series Count Data Regression Approach. *Accident Analysis and Prevention* **28**, 73-87.
- Johnson, N.L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. John Wiley, New York.
- Johnson, N.L., Kotz, S., and Kemp, A.W. (1992). *Univariate Discrete Distributions*, 2. ed., New York, John Wiley.
- Jorgenson, D.W. (1961). Multiple Regression Analysis of a Poisson Process. *Journal of the American Statistical Association* **56**, 235-246.
- Kocherlakota, S. and Kocherlakota, K. (1992). *Bivariate Discrete Distributions*. Marcel Dekker, New York.
- Lambert, D. (1992). Zero-Inflated Poisson Regression with an Application to Defects in Manufacturing. *Technometrics* **34**, 1-14.
- Larson, D.M. (1993). Joint Recreation Choices and Implied Value of Time. *Land Economics* **69**, 270-86.
- Marshall, A.W. and Olkin, I. (1990). Multivariate Distributions Generated from Mixtures of Convolution and Product Families. In *Topics in Statistical Dependence*, eds. Block, H.W., Sampson, A.R. and Savits, T.H., IMS Lecture Notes Monograph Series **16**, 371-393, Hayward, CA.
- Martin, A. and Witt, S.F. (1989). Forecasting Tourism Demand: A Comparison of the Accuracy of Several Quantitative Methods. *International Journal of Forecasting* **5**, 1-13.
- McConnell, K.E. (1992). On-site time in the Demand for Recreation. *American Journal of Agricultural Economics* **74**, 918-25.
- McCullagh, P. and Nelder, J.A. (1983). *Generalized Linear Models*. 1st ed., Chapman and Hall, London.
- McFadden, D. (1989). A Method of Simulated Moments for Estimation of Discrete Choice Models without Numerical Integration. *Econometrica* **57**, 995-1026.
- McKenzie, E. (1985). Some Simple Models for Discrete Variate Time Series. *Water Resources Bulletin* **21**, 645-650.
- McKenzie, E. (1986). Autoregressive Moving-Average Processes with Negative Binomial and Geometric Marginal Distributions. *Advances in Applied Probability* **18**, 679-705.
- McKenzie, E. (1988). Some ARMA Models for Dependent Sequences of Poisson Counts. *Advances in Applied Probability* **20**, 822-835.
- Melenberg, B. and van Soest, A. (1996). Parametric and Semi-Parametric Modelling of Vacation Expenditures. *Journal of Applied Econometrics* **11**, 59-76.
- Melkersson, M. and Rooth, D. (2000). Modelling of Household Fertility Using Inflated Count Data Models. *Journal of Population Economics* **13**, 189-203.
- Mullahy, J. (1986). Specification and Testing in some Modified Count Data Models.

- Journal of Econometrics* **33**, 341-365.
- Mullahy, J. (1997). Heterogeneity, Excess Zeros and the Structure of Count Data Models. *Journal of Applied Econometrics* **12**, 337-350.
- Munkin, M.K. and Trivedi, P.K. (1999). Simulated Maximum Likelihood Estimation of Multivariate Mixed-Poisson Regression Models, with Application. *Econometrics Journal* **2**, 29-48.
- Nordström, J. (1999a). Unobserved Components in International Tourism Demand. *Umeå Economic Studies* **502**.
- Nordström, J. (1999b). Tourism and Travel: Accounts, Demand and Forecasts. *Umeå Economic Studies* **509**.
- Pohlmeier, W. and Ulrich, V. (1995). An Econometric Model of the Two-part Decision Making Process in the Demand for Health Care. *Journal of Human Resources* **30**, 339-361.
- Ronning, G. and Jung, R.C. (1992). Estimation of a First Order Autoregressive Process with Poisson Marginals for Count Data. In Fahrmeir, L., Francis, B., Gilchrist, R., Tutz, G., eds., *Advances in GLIM and Statistical Modelling*. Springer-Verlag, New York.
- Rudholm, N. (2001). Entry and the Number of Firms in the Swedish Pharmaceuticals Market. *Review of Industrial Organization* **19**, 351-364.
- Santos Silva, J.M.C. and Covas, F. (2000). A Modified Hurdle Model for Completed Fertility. *Journal of Population Economics* **13**, 173-188.
- Shaban, S.A. (1988). Poisson-Lognormal Distributions. In Crow, E.L. and Shimizu, K., eds, *Lognormal Distributions*. Marcel Dekker, New York.
- Shaw, D. (1988). On-Site Samples, Regression Problems of Non-negative Integers, Truncation and Endogenous Stratification. *Journal of Econometrics* **37**, 211-223.
- Stutel, F.W. and Van Harn, K. (1979). Discrete Analogues of Self-Decomposability and Stability. *Annals of Probability* **7**, 893-899.
- Stynes, D. J. and Peters, G. L. (1984). A review of logit models with implications for modeling recreation choices. *Journal of Leisure Research* **16**, 295-310.
- Terza, J.V. and Wilson, P.W. (1990). Analyzing Frequencies of Several Types of Events: A Mixed Multinomial-Poisson Approach. *Review of Economics and Statistics* **72**, 108-115.
- Williams, H.C.W.L. (1979). On the foundation of travel demand models and economic evaluation measures of user benefit. *Environment and Planning* **9**, 285-344.
- Wilson, P.W. (1992). Count Data Models Without Mean-Variance Restrictions. Working Paper presented at the European Meeting of the Econometric Society, Brussels.
- Winkelmann, R. (2000). *Econometric Analysis of Count Data*, 3rd ed., Springer-Verlag, Berlin.
- Witt, S.F. and Witt, C.A. (1995). Forecasting Tourism Demand: A Review of Empirical Research. *International Journal of Forecasting* **11**, 447-475.
- Woodside, A.G. and Lysonski, S. (1989). A General Model of Traveler Destination Choice.

*Journal of Travel Research* **27**, 8-14.

Zeger, S.L. (1988). A Regression Model for Time-series of Counts. *Biometrika* **75**, 621-629.

Zeger, S.L. and Qaqish, B. (1988). Markov Regression Models for Time Series: A Quasi-Likelihood Approach. *Biometrics* **44**, 1019-1031.