CHAPTER 14

Births at Edendale Hospital

14.1 Introduction

Haines, Munoz and van Gelderen (1989) have described the fitting of Gaussian ARIMA models to various discrete-valued time series related to births occurring during a 16-year period at Edendale Hospital in Natal*, South Africa. The data include monthly totals of mothers delivered and deliveries by various methods at the Obstetrics Unit of that hospital in the period from February 1970 to January 1986 inclusive. Although 16 years of data were available, Haines *et al.* considered only the final eight years' observations when modelling any series other than the total deliveries. This was because, for some of the series they modelled, the model structure was found not to be stable over the full 16-year period. In this analysis we have in general modelled only the last eight years' observations.

14.2 Models for the proportion Caesarean

One of the series considered by Haines *et al.*, to which they fitted two models, was the number of deliveries by Caesarean section. From their models they drew the conclusions (in respect of this particular series) that there is a clear dependence of present observations on past, and that there is a clear linear upward trend. In this section we describe the fitting of (discrete-valued) HM and Markov regression models to this series, Markov regression models, that is, in the sense in which that term is used by Zeger and Qaqish (1988). These models are of course rather different from those fitted by Haines *et al.* in that theirs, being based on the normal distribution, are continuous-valued. Furthermore, the discrete-valued models make it possible to model the proportion (as opposed to the number) of Caesareans performed in each month. Of the models proposed here, one type is 'observation-driven' and the other 'parameter-driven'; see Cox (1981) for these terms. The most important conclusion drawn from the discrete-valued models, and one which the Gaussian ARIMA models did not provide, is that there is a strong upward time trend in the proportion of the deliveries that are by Caesarean section.

* now KwaZulu–Natal

The two models that Haines *et al.* fitted to the time series of Caesareans performed, and that they found to fit very well, may be described as follows. Let Z_t denote the number of Caesareans in month t (February 1978 being month 1 and t running up to 96), and let the process $\{a_t\}$ be Gaussian white noise, i.e. uncorrelated random shocks distributed normally with zero mean and common variance σ_a^2. The first model fitted is the ARIMA(0,1,2) model with constant term:

$$\nabla Z_t = \mu + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}. \tag{14.1}$$

The maximum likelihood estimates of the parameters, with associated standard errors, are $\hat{\mu} = 1.02 \pm 0.39$, $\hat{\theta}_1 = 0.443 \pm 0.097$, $\hat{\theta}_2 = 0.393 \pm 0.097$ and $\hat{\sigma}_a^2 = 449.25$. The second model is an AR(1) with linear trend:

$$Z_t = \beta_0 + \beta_1 t + \phi Z_{t-1} + a_t, \tag{14.2}$$

with parameter estimates as follows: $\hat{\beta}_0 = 120.2 \pm 8.2$, $\hat{\beta}_1 = 1.14 \pm 0.15$, $\hat{\phi} = 0.493 \pm 0.092$ and $\hat{\sigma}_a^2 = 426.52$.

Both of these models, (14.1) and (14.2), provide support for the conclusion of Haines *et al.* that there is a dependence of present observations on past, and a linear upward trend. Furthermore, the models are nonseasonal; the Box–Jenkins methodology used found no seasonality in the Caesareans series. The X-11-ARIMA seasonal adjustment method employed in an earlier study (Munoz, Haines and van Gelderen, 1987) did, however, find some evidence, albeit weak, of a seasonal pattern in the Caesareans series similar to a pattern that was observed in the 'total deliveries' series. (This latter series shows marked seasonality, with a peak in September, and in Haines *et al.* (1989) it is modelled by the seasonal ARIMA model $(0, 1, 1) \times (0, 1, 1)_{12}$.)

It is of some interest to model the proportion, rather than the number, of Caesareans in each month. (See Figure 14.1 for a plot of this proportion for the years 1978–1986.)

It could be the case, for instance, that any trend, dependence or seasonality apparently present in the number of Caesareans is largely inherited from the total deliveries, and a constant proportion Caesarean is an adequate model. On the other hand, it could be the case that there is an upward trend in the proportion of the deliveries that are by Caesarean and this accounts at least partially for the upward trend in the number of Caesareans. The two classes of model that we discuss in this section condition on the total number of deliveries in each month and seek to describe the principal features of the proportion Caesarean.

Now let n_t denote the total number of deliveries in month t. A very general possible model for $\{Z_t\}$ which allows for trend, dependence on previous observations and seasonality, in the proportion Caesarean, is as follows. Suppose that, conditional on the history $\mathbf{Z}^{(t-1)} = \{Z_s : s \leq t–1\}$,

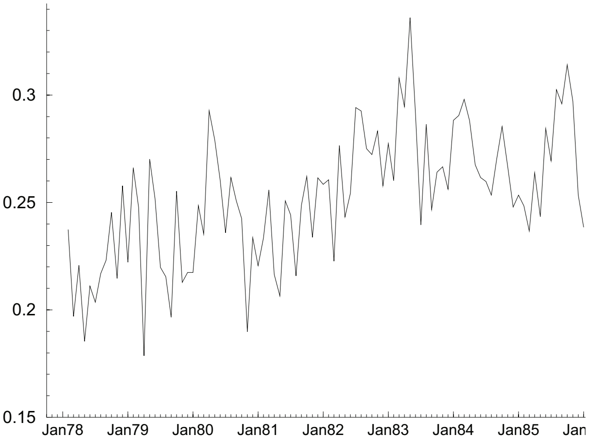

Figure 14.1 *Edendale births: monthly numbers of deliveries by Caesarean section, as a proportion of all deliveries, February 1978 – January 1986.*

Z_t is distributed binomially with parameters n_t and ${}_tp$, where, for some positive integer q,

$$\begin{aligned} \text{logit}\,{}_tp \quad = \quad & \alpha_1 + \alpha_2 t + \beta_1(Z_{t-1}/n_{t-1}) + \beta_2(Z_{t-2}/n_{t-2}) + \cdots \\ & + \beta_q(Z_{t-q}/n_{t-q}) + \gamma_1 \sin(2\pi t/12) + \gamma_2 \cos(2\pi t/12). \end{aligned} \tag{14.3}$$

This is, in the terminology of Zeger and Qaqish (1988), a Markov regression model, and generalizes the model described by Cox (1981) as an 'observation-driven linear logistic autoregression', in that it incorporates trend and seasonality and is based on a binomial rather than a Bernoulli distribution. It is observation-driven in the sense that the distribution of the observation at a given time is specified in terms of the observations at earlier times. Clearly it is possible to add further terms to the above expression for logit ${}_tp$ to allow for the effect of any further covariates thought relevant, e.g. the number or proportion of deliveries by various instrumental techniques.

It does not seem possible to formulate an unconditional maximum likelihood procedure to estimate the parameters α_1, α_2, β_1, …, β_q, γ_1 and γ_2 of the model (14.3). It is, however, straightforward to compute estimates of these parameters by maximizing a conditional likelihood. If for instance no observations earlier than Z_{t-1} appear in the model, the product

$$\prod_{t=1}^{96} \binom{n_t}{z_t} {}_tp^{z_t}(1 - {}_tp)^{n_t - z_t}$$

is the likelihood of $\{Z_t : t = 1, \ldots, 96\}$, conditional on Z_0. Maximization thereof with respect to $\alpha_1, \alpha_2, \beta_1, \gamma_1$ and γ_2 yields estimates of these parameters, and can be accomplished simply by performing a logistic regression of Z_t on t, Z_{t-1}/n_{t-1}, $\sin(2\pi t/12)$ and $\cos(2\pi t/12)$.

In the search for a suitable model the following explanatory variables were considered: t (i.e. the time in months, with February 1978 as month 1), t^2, the proportion Caesarean lagged one, two, three or twelve months, sinusoidal terms at the annual frequency (as in (14.3)), the calendar month, the proportion and number of deliveries by forceps or vacuum extraction, and the proportion and number of breech births. Model selection was performed by means of AIC and BIC. The **R** function used, `glm`, does not provide l, the log-likelihood, but it does provide the deviance, from which the log-likelihood can be computed (McCullagh and Nelder, 1989, p. 33). Here the maximum log-likelihood of a full model is -321.0402, from which it follows that $-l = 321.04 + \frac{1}{2} \times \text{deviance}$.

The best models found with between one and four explanatory variables (other than the constant term) are listed in Table 14.1, as well as several other models that may be of interest. These other models are: the model with constant term only; the model with Z_{t-1}/n_{t-1}, the previous proportion Caesarean, as the only explanatory variable; and two models which replace the number of forceps deliveries as covariate by the proportion of instrumental deliveries. The last two models were included because the proportion of instrumental deliveries may seem a more sensible explanatory variable than the one it replaces; it will be observed, however, that the models involving the number of forceps deliveries are preferred by AIC and BIC. (By 'instrumental deliveries' we mean those which are either by forceps or by vacuum extraction.) Note, however, that both AIC and BIC indicate that one could still gain by inclusion of a fifth explanatory variable in the model.

The strongest conclusion we may draw from these models is that there is indeed a marked upward time trend in the proportion Caesarean. Secondly, there is positive dependence on the proportion Caesarean in the previous month. The negative association with the number (or proportion) of forceps deliveries is not surprising in view of the fact that delivery by Caesarean and by forceps are in some circumstances alternative techniques. As regards seasonality, the only possible seasonal pattern found in the proportion Caesarean is the positive 'October effect'. Among the calendar months only October stood out as having some explanatory power. As can be seen from Table 14.1, the indicator variable specify-

Table 14.1 *Edendale births: models fitted to the logit of the proportion Caesarean.*

explanatory variables	coefficients	deviance	AIC	BIC
constant	−1.253	208.92	855.00	860.13
t (time in months)	0.003439			
constant	−1.594	191.70	839.78	847.47
t	0.002372			
previous proportion Caesarean	1.554			
constant	−1.445	183.63	833.71	843.97
t	0.001536			
previous proportion Caesarean	1.409			
no. forceps deliveries in month t	−0.002208			
constant	−1.446	175.60	**827.68**	**840.50**
t	0.001422			
previous proportion Caesarean	1.431			
no. forceps deliveries in month t	−0.002393			
October indicator	0.08962			
constant	−1.073	324.05	968.13	970.69
constant	−1.813	224.89	870.97	876.10
previous proportion Caesarean	2.899			
constant	−1.505	188.32	838.40	848.66
t	0.002060			
previous proportion Caesarean	1.561			
proportion instrumental deliveries in month t	−0.7507			
constant	−1.528	182.36	834.44	847.26
t	0.002056			
previous proportion Caesarean	1.590			
proportion instrumental deliveries in month t	−0.6654			
October indicator	0.07721			

ing whether the month was October was included in the 'best' set of four explanatory variables. A possible reason for an October effect is as follows. An overdue mother is more likely than others to give birth by Caesarean, and the proportion of overdue mothers may well be highest in October because the peak in total deliveries occurs in September.

Since the main conclusion emerging from the above logit-linear models

Table 14.2 *Edendale births: the three two-state binomial–HMMs fitted to the proportion Caesarean. (The time in months is denoted by t, and February 1978 is month 1.)*

$\text{logit}\,{}_t p_i$	$-l$	AIC	BIC	γ_{12}	γ_{21}	α_1	α_2	β_1	β_2
α_i	420.322	848.6	858.9	0.059	0.086	−1.184	−0.960	–	–
$\alpha_i + \beta t$	402.350	**814.7**	**827.5**	0.162	0.262	−1.317	−1.140	0.003298	
$\alpha_i + \beta_i t$	402.314	816.6	832.0	0.161	0.257	−1.315	−1.150	0.003253	0.003473

is that there is a marked upward time trend in the proportion Caesarean, it is of interest also to fit HMMs with and without time trend. The HMMs we use in this application have two states and are defined as follows. Suppose $\{C_t\}$ is a stationary homogeneous Markov chain on state space $\{1, 2\}$, with transition probability matrix

$$\mathbf{\Gamma} = \begin{pmatrix} 1 - \gamma_{12} & \gamma_{12} \\ \gamma_{21} & 1 - \gamma_{21} \end{pmatrix}.$$

Suppose also that, conditional on the Markov chain, Z_t has a binomial distribution with parameters n_t and p_i, where $C_t = i$. A model without time trend assumes that p_1 and p_2 are constants and has four parameters. One possible model which allows p_i to depend on t has $\text{logit}\,{}_t p_i = \alpha_i + \beta t$ and has five parameters. A more general model yet, with six parameters, has $\text{logit}\,{}_t p_i = \alpha_i + \beta_i t$.

Maximization of the likelihood of the last eight years' observations gives the three models appearing in Table 14.2 along with their associated log-likelihood, AIC and BIC values. It may be seen that, of the three models, that with a single time-trend parameter and a total of five parameters achieves the smallest AIC and BIC values.

In detail, that model is as follows, t being the time in months and February 1978 being month 1:

$$\mathbf{\Gamma} = \begin{pmatrix} 0.838 & 0.162 \\ 0.262 & 0.738 \end{pmatrix},$$

$$\text{logit}\,{}_t p_1 = -1.317 + 0.003298t,$$

$$\text{logit}\,{}_t p_2 = -1.140 + 0.003298t.$$

The model can be described as consisting of a Markov chain with two fairly persistent states, along with their associated time-dependent probabilities of delivery being by Caesarean, the (upward) time trend being the same, on a logit scale, for the two states. State 1 is rather more likely than state 2, because the stationary distribution is (0.618, 0.382), and has associated with it a lower probability of delivery being by Caesarean. For state 1 that probability increases from 0.212 in month 1 to 0.269 in month 96, and for state 2 from 0.243 to 0.305. The corresponding unconditional probability increases from 0.224 to 0.283. It may or may not be possible to interpret the states as (for instance) nonbusy and busy periods in the Obstetrics Unit of the hospital, but without further information, e.g. on staffing levels, such an interpretation would be speculative.

It is true, however, that other models can reasonably be considered. One possibility, suggested by inspection of Figure 14.1, is that the proportion Caesarean was constant until January 1981, then increased linearly to a new level in about January 1983. Although we do not pursue such a model here, it is possible to fit an HMM incorporating this feature. (Models with change-points are discussed and used in Chapter 15.)

If one wishes to use the chosen model to forecast the proportion Caesarean at time 97 for a given number of deliveries, what is needed is the one-step-ahead forecast distribution of Z_{97}, i.e. the distribution of Z_{97} conditional on $Z_1, \dots, Z_{96}$. This is given by the likelihood of Z_1, …, Z_{97} divided by that of Z_1, …, Z_{96}. More generally, the k-step-ahead forecast distribution, i.e. the conditional probability that $Z_{96+k} = z$, is given by a ratio of likelihoods, as described in Section 5.2.

The difference in likelihood between the HMMs with and without time trend is convincing evidence of an upward trend in the proportion Caesarean, and confirms the main conclusion drawn above from the logit-linear models. Although Haines *et al.* concluded that there is an upward trend in the number of Caesareans, it does not seem possible to draw any conclusion about the proportion Caesarean from their models, or from any other ARIMA models.

It is of interest also to compare the fit of the five-parameter HMM to the data with that of the logistic autoregressive models. Here it should be noted that the HMM produces a lower value of $-l$ (402.3) than does the logistic autoregressive model with four explanatory variables (408.8), without making use of the additional information used by the logistic autoregression. It does not use z_0, the number of Caesareans in January 1978, nor does it use information on forceps deliveries or the calendar month. It seems therefore that HMMs have considerable potential as simple yet flexible models for examining dependence on covariates (such as time) in the presence of serial dependence.

14.3 Models for the total number of deliveries

If one wishes to project the number of Caesareans, however, a model for the proportion Caesarean is not sufficient; one needs also a model for the

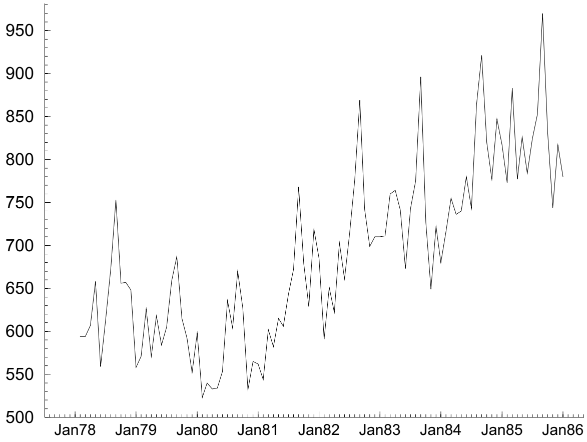

Figure 14.2 *Edendale births: monthly totals of deliveries, February 1978 – January 1986.*

total number of deliveries, which is a series of unbounded counts. The model of Haines *et al.* for the total deliveries was the seasonal ARIMA model $(0, 1, 1) \times (0, 1, 1)_{12}$ without constant term. For this series (unlike the others) they used all 16 years' data to fit the model, but we continue here to model only the final eight years' observations.

First, five two-state Poisson–HMMs. were fitted to the monthly totals of deliveries (depicted in Figure 14.2). One was a model without covariates, that is,

$$\log {}_t\lambda_i = a_i; \tag{14.4}$$

and the other four models for $\log {}_t\lambda_i$ were of the following forms:

$$\begin{aligned}
\log {}_t\lambda_i &= a_i + bt; && (14.5)\\
\log {}_t\lambda_i &= a_i + bt + c\cos(2\pi t/12) + d\sin(2\pi t/12); && (14.6)\\
\log {}_t\lambda_i &= a_i + bt + c\cos(2\pi t/12) + d\sin(2\pi t/12) + fn_{t-1}; && (14.7)\\
\log {}_t\lambda_i &= a_i + bt + c\cos(2\pi t/12) + d\sin(2\pi t/12) + fn_{t-1} + gn_{t-2}. && (14.8)
\end{aligned}$$

Models (14.7) and (14.8) are examples of the incorporation of extra dependencies at observation level as described in Section 8.6; that is, they do not assume conditional independence of the observations $\{n_t\}$ given the Markov chain. But this does not significantly complicate the likelihood evaluation, as the state-dependent probabilities can just treat n_{t-1} and n_{t-2} in the same way as any other covariate. The models fitted are summarized in Table 14.3.

Table 14.3 *Edendale births: summary of the five two-state Poisson–HMMs fitted to the number of deliveries. (The time in months is denoted by t, and February 1978 is month 1.)*

log $_t\lambda_i$	$-l$	AIC	γ_{12}	a_1	b	c	f	g
		BIC	γ_{21}	a_2		d		
(14.4)	642.023	1292.0	0.109	6.414	–	–	–	–
		1302.3	0.125	6.659		–		
(14.5)	553.804	1117.6	0.138	6.250	0.004873	–	–	–
		1130.4	0.448	6.421		–		
(14.6)	523.057	1060.1	0.225	6.267	0.004217	−0.0538	–	–
		1078.1	0.329	6.401		−0.0536		
(14.7)	510.840	**1037.7**	0.428	5.945	0.002538	−0.0525	0.000586	–
		1058.2	0.543	6.069		−0.0229		
(14.8)	510.791	1039.6	0.415	5.939	0.002511	−0.0547	0.000561	0.0000365
		1062.7	0.537	6.063		−0.0221		

It may be seen that, of these five models, model (14.7), with a total of eight parameters, achieves the smallest AIC and BIC values. That model is as follows:

$$\begin{aligned}\log {}_t\lambda_i &= 5.945/6.069 + 0.002538t - 0.05253\cos(2\pi t/12)\\ &\qquad - 0.02287\sin(2\pi t/12) + 0.0005862 n_{t-1}.\end{aligned}$$

In passing, we may mention that models (14.5)–(14.8) were fitted by means of the **R** function `constrOptim` (for constrained optimization). It is possible to refine these models for the monthly totals by allowing for the fact that the months are of unequal length, and by allowing for seasonality of frequency higher than annual, but these variations were not pursued.

A number of log-linear models were then fitted by `glm`. Table 14.4 compares the models fitted by `glm` to the total deliveries, and from that table it can be seen that of these models BIC selects the model incorporating time trend, sinusoidal components at the annual frequency, and the number of deliveries in the previous month (n_{t-1}). The details of this model are as follows. Conditional on the history, the number of deliveries in month t (N_t) is distributed Poisson with mean ${}_t\lambda$, where

$$\begin{aligned}\log {}_t\lambda &= 6.015 + 0.002436t - 0.03652\cos(2\pi t/12)\\ &\qquad - 0.02164\sin(2\pi t/12) + 0.0005737 n_{t-1}.\end{aligned}$$

(Here, as before, February 1978 is month 1.)

Both the HMMs and the log-linear autoregressions reveal time trend,

Table 14.4 *Edendale births: models fitted by* `glm` *to the log of the mean number of deliveries.*

Explanatory variables	Deviance	$-l$	AIC	BIC
t	545.9778	674.3990	1352.8	1357.9
t, sinusoidal terms *	428.6953	615.7577	1239.5	1249.8
t, sinusoidal terms, n_{t-1}	356.0960	579.4581	1168.9	**1181.7**
t, sinusoidal terms, n_{t-1}, n_{t-2}	353.9086	578.3644	**1168.7**	1184.1
sinusoidal terms, n_{t-1}	464.8346	633.8274	1275.7	1285.9
t, n_{t-1}	410.6908	606.7555	1219.5	1227.2
n_{t-1}	499.4742	651.1472	1306.3	1311.4

* This is a one-state model similar to the HMM (14.6), and its log-likelihood value can be used to some extent to check the code for the HMM. Similar comments apply to the other models listed above.

seasonality, and dependence on the number of deliveries in the previous month. On the basis of both AIC and BIC the Poisson–HMM (14.7) is the best model.

14.4 Conclusion

The conclusion is therefore twofold. If a model for the number of Caesareans, given the total number of deliveries, is needed, the binomial–HMM with time trend is best of all of those considered (including various logit-linear autoregressive models). If a model for the total deliveries is needed (e.g. as a building-block in projecting the number of Caesareans), then the Poisson–HMM with time trend, 12-month seasonality and dependence on the number in the previous month is the best of those considered — contrary to our conclusion in MacDonald and Zucchini (1997), which was unduly pessimistic about the Poisson–HMMs.

The models chosen here suggest that there is a clear upward time trend in both the total deliveries and the proportion Caesarean, and seasonality in the total deliveries.