PREDICTING BIRTHS IN THE CURRENT EMPLOYMENT STATISTICS SURVEY

Steve Woodruff, Bureau of Labor Statistics
2 Massachusetts Ave. N.E., Suite 4985, Washington D.C. 20212

KEY WORDS: Birth and Death Process, Maximum Likelihood Estimation

1) Introduction

The Current Employment Statistics survey is a monthly survey of about 400,000 business establishments. It is used to estimate total national non-farm employment along with several other less important economic statistics. The universe for this survey (and sample) is constantly changing due to establishments going out of business and others starting up. The effect of these "births" and "deaths" remains an important source of non-sampling error.

This problem of births and deaths was studied by, Lent and Grezsiak, (1985). The operational conclusions of their work was that births and deaths tend to cancel and that the effect of these on estimation was minimal. The sample automatically picks up deaths, they report zero or just stop reporting although, in this case, nonrespondents must still be separated from deaths. Births can go unnoticed for extended periods of time (up to a year) between sample replenishments. With implementation of random sampling and possibly rotation sampling, the effect of births and deaths can be measured and these measurements used to improve the accuracy of the estimates.

The strata in this survey are industry/size-class groups. Employment in these strata may be shrinking or growing at different rates and in different directions. By observing births and deaths in the immediate past for each stratum, it is possible to estimate the "infinitesimal" birth and death rates for a Markov process that describes this birth and death behavior. This Markov model is used to predict the number of births for the current estimation period.

Modeling the number of establishments in the sampling frame, known in the Bureau as the Universe Data Base (UDB), as the realization of a birth and death process is a useful means of predicting population size when actual numbers of births is unknown.

ARIMA models were also fitted to the data on these Birth/Death processes. Autoregressive models within ARIMA seemed to provide the best fit of the various Time Series techniques. These models were used to forecast the number of reporting units in an estimation cell and these forecasts compared to the birth/death forecasts.

2) A Markov Birth and Death Process

For a sampling stratum, let X(t) be the total number of establishments in that stratum for month t. X(t) can be thought of as a stochastic process as each month, business establishments migrate to other strata or go out of business and others start up or migrate into the stratum. Karlin (1973, Ch 7), outlines several "Birth and Death Processes". One possible candidate and a method of finding it's moments is outlined next.

Let X(t) be a stationary Markov Process: $P(X(t+h)\varepsilon A | X(t)\varepsilon B) = P(X(h)\varepsilon A | X(0)\varepsilon B) \ \forall$ t, and let $P_{i,j}(h) = P(X(h) = j | P(0) = i)$. Let the transition probabilities satisfy the following conditions where λ_i *and* μ_i are positive numbers:

$$P_{i,i+1}(h) = \lambda_i h + o(h) = P(X(h) = i+1 | X(0) = i)$$

$$as \quad h \downarrow 0, i \geq 0$$

$$P_{i,i-1}(h) = \mu_i h + o(h) \quad as \quad h \downarrow 0, i \geq 1$$

$$P_{i,i}(h) = 1 - (\lambda_i + \mu_i)h + o(h) \quad as \quad h \downarrow 0, i \geq 0$$

$$P_{i,j}(0) = \delta_{ij}$$

where δ is the kronecker-δ (*)

Let M(t) = E[X(t)], be the expected value of X(t) under the stochastic structure just described and V(t) be the variance of X(t), (both assuming X(0)=M(0) is known). For a stochastic process called Linear Growth with Immigration, the infinitesimal birth rate when the population

Any views expressed here are those of the author and not the Bureau of Labor Statistics

contains n members is: $\lambda_n = n\lambda + a$ and the infinitesimal death rate is: $\mu_n = n\mu$ where a>0, λ>0, and μ>0.

M(t) and V(t) are derived by solving the "forward Kolmogorov differential equations" which are themselves derived from (*), the prime denotes derivative with respect to t; these equations are:

$$P'_{i0}(t) = -\lambda_0 P_{i0}(t) + \mu_1 P_{i1}(t)$$

$$P'_{ij}(t) = \lambda_{j-1} P_{i,j-1}(t) - (\lambda_j + \mu_j) P_{ij}(t) + \mu_{j+1} P_{i,j+1}(t)$$

$$\text{for } j \geq 1$$

When X(0) = i then be definition of expected value, M(t) is:

$$M(t) = \sum_{j=1}^{\infty} jP_{ij}(t).$$

Differentiate both sides of this equation and substitute the forward Kolmogorov differential equations to get the first order linear differential equation in M(t):

M'(t)=a +(λ-μ)M(t), M(0)=X(0)=i .

The solution of this is:

$$M(t) = \frac{a}{\lambda - \mu}\{e^{(\lambda-\mu)t} - 1\} + M(0)e^{(\lambda-\mu)t}$$

for λ≠μ and

M(t) = at + M(0) for λ=μ

Using the same technique, $E(X(t)^2)$ can be derived and this used to find V(t). The result is:

$$V(t) = \frac{1}{(\lambda-\mu)^2}\{(M(0)(\lambda^2-\mu^2) + a\lambda)e^{2(\lambda-\mu)t} - [(\lambda+\mu)a + M(0)(\lambda^2-\mu^2)]e^{(\lambda-\mu)t} + a\mu\}$$

for λ≠μ and

$$V(t) = a\mu t^2 + 2M(0)\mu t + at \quad \text{for} \quad \lambda = \mu$$

M(t) gives the expected value of X(t) given M(0), λ, μ, and a. The mean, M(t), will be used to predict X(t) given X(t-m) is known, m>0. In this case the prediction for X(t) is M(m) given M(0)=X(t-m).

X(t) = X(t-m) + (b(t)-d(t)) where b(t) is the number of births between months t-m and t and d(t) is the number of deaths in that same interval. Then b(t) = X(t) - X(t-m) + d(t). Since X(t-m) and d(t) are known, predicting X(t) is equivalent to predicting b(t).

If in addition to the above assumptions about {X(t)}, the {X(t)} are normally distributed (neither independent or identical), the likelihood function for this Markov process, {X(t)}, is:

$$L = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi v_i}} \text{Exp}\left(-\frac{(X(i) - m_i)^2}{2v_i}\right) \quad \text{where}$$

m_i and v_i are M(1) and V(1) evaluated with M(0) = X(i-1) for i=1,2,.....,k. That is, $m_i = \frac{a}{\lambda-\mu}\{e^{(\lambda-\mu)} - 1\} + X(i-1)e^{(\lambda-\mu)}$ and

$$v_i = \frac{1}{(\lambda-\mu)^2}\{(X(i-1)(\lambda^2-\mu^2) + a\lambda)e^{2(\lambda-\mu)} - [(\lambda+\mu)a + X(i-1)(\lambda^2-\mu^2)]e^{(\lambda-\mu)} + a\mu\}$$

for i=1,2,3,.....,k and λ ≠ μ.

For λ=μ,

$$m_i = a + X(i-1) \quad \textit{and}$$

$$v_i = a\mu + 2X(i-1)\mu + a$$

The terms in L are the conditional densities of X(i) given the outcome of X(i-1) for i=1,2,.....,k. L(X(1),X(2),X(3),.....,X(k)) is the joint density of (X(1),X(2),X(3),.....,X(k)) for this Markov process. Finding the values of λ, μ, and a which maximize this likelihood function is done by computer rather than algebraic manipulation. Using historical data for 31 months (Apr 1990 to Apr 1993) in each of 10 industries in Iowa, the values of λ, μ, and a that maximize this likelihood were found.

This initial testing indicated that the relatively simple expressions for first and second moments when μ = λ may be most appropriate, with the common value of λ and μ of about: .02.

When λ = μ, the problem reduces to estimating "*a*".

3) ARIMA Modeling and Other Alternatives

In addition to the MLE for "*a*" described above, three other estimators are tested: a regression estimator, $\hat{a}_r$, a simple slope estimator, $\hat{a}_n$, and an ARIMA based estimator, $\hat{a}_{AR}$. Let $\hat{a}_m$ denote the maximum likelihood estimator of "a" (maximizes L).

$\hat{a}_r = \dfrac{\sum_{i=1}^{n} iX(i) - M(0)\sum_{i=1}^{n} i}{\sum_{i=1}^{n} i^2}$ is the value of a that minimizes: $\sum_{i=1}^{n}(X(i) - ai - M(0))^2$.

$\hat{a}_n = \dfrac{X(n) - X(1)}{n-1}$ is the average month-to-month change from month 1 to month n.

In spite of good estimates for the mean month-to-month change, the variation around this mean of the realized change is large enough to make predictions of the change for any particular month very problematic.

The correlations between the three estimates of "a" are all about .9 This implies that little is to be gained by using a composite estimator (with optimal weights only a variance reduction of about 10% and actual weights would necessarily be only approximately optimal). .

The Bureau may be sampling for births and estimating birth employment periodically. Let y_b denote this proposed the sample estimate of birth employment. Let $\hat{n}_b = \hat{n}_{b-d} + n_d$ and let $x_b = \hat{n}_b \overline{AE}_{t-k}$ be the estimate of birth employment projected from the historical data. These two estimators are stochastically independent and there respective variances are available (estimable). $\hat{V}(y_b)$ is estimated from the sampling distribution.

$\hat{V}(x_b) = \overline{AE}^2_{t-k}\hat{V}(n_{b-d})$, and $\hat{V}(\hat{n}_{b-d}) = \hat{V}(k\hat{a}) = k^2[\hat{a}\hat{\mu}k^2 + 2\overline{AE}_{t-k}\hat{\mu}k + \hat{a}k]$

k=3 in case births are estimated for a quarter at a time and updated from the last quarter UDB employment figures.

The composite estimator for birth employment is:

$C_b = \dfrac{y_b\hat{V}(x_b) + x_b\hat{V}(y_b)}{\hat{V}(x_b) + \hat{V}(y_b)}$. and the variance of this composite estimator is:

$$V(C_b) = \frac{[\hat{V}(x_b)]^2 V(y_b) + [\hat{V}(y_b)]^2 V(x_b)}{[\hat{V}(x_b) + \hat{V}(y_b)]^2}$$

and this is estimated with

$$\hat{V}(C_b) = \frac{\hat{V}(x_b)\hat{V}(y_b)}{\hat{V}(x_b) + \hat{V}(y_b)} \leq \begin{cases} \hat{V}(x_b) \\ \hat{V}(y_b) \end{cases}.$$ The variance of the composite estimator is minimized when $\hat{V}(x_b) = \hat{V}(y_b)$ and is half their common value.

The Time Series programs that were used to fit ARIMA models, test them, and predict the industry sizes (number of reporting units) from a fitted model for this study was done using S-PLUS software, Venables and Ripley (1994), on a PC. As was noted above, when birth and death processes were used to model this problem, the noise in these stochastic processes overwhelmed the signal. The same was true when ARIMA models were fitted to the same data.

AR(1) to AR(3) processes seemed to best describe the behavior of the data as measured by the Akaike Information Criteria. The Autoregression coefficients were estimated via maximum likelihood assuming a Gaussian distribution. The ARIMA estimates of "*a*" denoted $\hat{a}_{AR}$ (or multiples of "*a*" as tabulated in the table above) appear uncorrelated with the other estimates. This fact suggests that a composite of the ARIMA estimates and one of the other three may give an improved estimate of "*a*".

4) An Empirical Study

Historical data for Iowa from SICs 16, 23, 24, 25, 28, 30, 53, 63, 73, and 75 for 37 months (March 91 to March 94) was used to test the three estimators 6*a* (half year ahead projections of SIC growth in number of reporting units). The MLE was generally best (possibly because it tends to be more conservative, smaller) and is based on a more specific stochastic description of birth and death processes than the other predictions. For these six month ahead projections, data for the first 31 months was used to estimate change between month 31 and month 37. These projections are given below in Table 1.

Table 1. Six month ahead changes for four predictors and the observed change.

SIC→ Est↓	16	23	24	25	28	30	53	63	73	75
$\hat{a}_n$	6.1	1.1	3.1	2.2	-2.4	3.1	-3.6	0	3.6	2.7
$\hat{a}_r$	6.6	2.1	4.1	3.5	-1.3	3.2	-.7	1.1	4.3	3.5
$\hat{a}_m$	10.3	.4	4.5	1.7	.1	2.9	0	0	4.1	4.5
$\hat{a}_{AR}$	-13	0	-1	-1	4	-3	3	-1.5	-6.5	-11
change	-24	-6	7	1	3	-3	0	2	3	7

Table 1a. Root MSEs (across SICs) of Six month ahead changes (smallest in bold). These come from Table 1, root of average of ([change row] - [predictor row])^2.

MSE → Est ↓	All SICs	W/O SIC 16
$\hat{a}_{AR}$	**8.13**	7.75
$\hat{a}_m$	11.31	**3.37**
$\hat{a}_r$	10.43	4.12
$\hat{a}_n$	10.37	4.34

SIC 16 is highly seasonal and the ARIMA predictor did a much better job of picking up this seasonality than the other methods as shown in Table 1a. When SIC 16 was removed (second column of Table 1a, the MLE was the best method.

These estimated changes in number of reporting units must be translated into change in employment. Since births and deaths generally predominate in the smaller size classes, it would give biased estimates of employment change to simply multiply average reporting unit employment in an SIC by the estimated change in the number of reporting units. The historical data can also be used to determine month-to-month change in employment as well as month-to-month change in the number of reporting units. The ratio of the one sequence to the other then estimates the change in employment per unit change in reporting units. Some weighted average of these ratios can be multiplied by the projected change in number of reporting units to estimate projected change in employment.

5) Conclusions

It may be futile to try to predict changes in number of reporting units in an estimation cell based on past numbers of births and deaths. A random sampling and estimation plan is a necessary adjunct to the indirect techniques studied in this paper. One current plan is to sample for births on a "just in time" basis. Look for births between frame refinements and include them in the estimation process in a more timely fashion than is now occurring. At present, the effect of births is only accounted for during annual benchmark revisions. We propose getting birth employment by sampling new Unemployment Insurance accounts and screening these to remove pre existing establishments while asking new establishments for their employment.

The indirect techniques studied here may be useful in a composite estimator where the other component is a direct sample based estimator for birth employment. The MLE of *"a"* that is derived from the Birth-Death process seems to be better (except in highly seasonal industries) than the other alternatives considered: the regression estimates, naive estimates, and the Time Series estimates. The MLE can be improved if seasonality is first removed from the historical series before "a" is estimated, then this estimate of "a" is used to predict the 6-month ahead nonseasonal component of the time series, and then the seasonal component added back to get the final prediction.

References

Grzesiak T. J.and Lent J. (1988) *Estimating Business Birth Employment in the Current Employment Statistics Program*, Proceeding of the American Statistical Association, Section on Survey Research Methods.

Karlin Samuel (1973). *A First Course in Stochastic Processes*, Academic Press.

Venables W. N. and Ripley B. D. (1994). *Modern Applied Statistics with S-Plus*, Springer-Verlag.