ADMINISTRATIVE TAX REFORM AND THE DURATION OF TAX RETURNS IN THE OECS

by

Roland Craigwell
Research Department
Central Bank of Barbados

and

Nathaniel Samuel
Research and Information Department
Eastern Caribbean Central Bank

Administrative Tax Reform and the Duration of Tax Returns in the OECS

Section I - Introduction

This paper represents the first fruits of research into the impact of administrative tax reform on the duration of tax returns in Dominica and St Vincent & the Grenadines[1]. Specifically, it attempts to answer the following question: "Have there been any improvements in the time taken for individuals to submit their tax returns over the five-year period 1991 - 1995?" This is important because Dominica and St. Vincent have spent large sums of money in reducing the lags in processing tax information and would like to ascertain whether these changes have been successful. A detail discussion of the changes in the administrative tax procedures can be found in ECCB and ECEMP (1997).

In this respect, models of duration data are chosen as the tool of analysis. The genesis of such models originated from industrial engineering and medical research where data is often found in the form of durations or the time taken for an event to occur. For example, in testing the effectiveness of a new drug, a medical researcher may be interested in the impact of that drug on the time taken for a sample of patients to get well.

Recently, such models have been applied in economic research. By far, the most popular use has been in the analysis of unemployment data. Accurate information on the duration of unemployment are critical to job search models that answer such questions as: "Does the duration of unemployment vary across individuals and with the length of the spell?" Models of

[1]Both islands are members of the Organisation of Eastern Caribbean States (OECS).

duration data have also been used in other areas such as investigating the timing of births, and the duration of contract strikes in the manufacturing sector[2].

The first objective of this paper therefore, is to introduce the specific methodology of duration analysis to economic research in the region. As such, the paper utilises models and techniques which are sufficiently simple for illustration purposes while remaining sufficiently powerful to answer questions on tax duration. Our second objective will be to investigate whether or not these models reveal a pattern (improvement or deterioration) in the time taken to submit tax returns in Dominica and St Vincent.

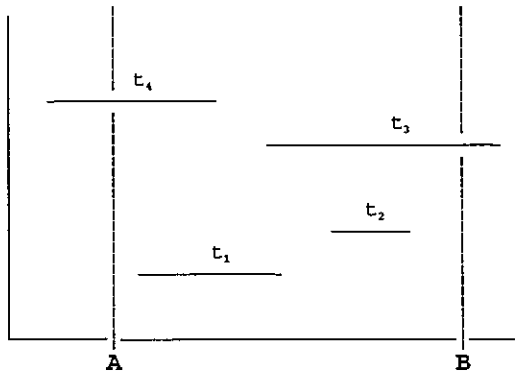
The paper proceeds as follows: Section II describes the specific methodology of duration analysis giving a brief overview of the various models utilised and defining some of the terminology specific to the topic. In Section III, we discuss the data which was taken from a study on the impact of administrative tax reform in Dominica and St Vincent and the Grenadines. Section IV presents the results of our own analysis, while Section V concludes with an appraisal of the methodology and suggestions for further research.

Section II - Methodology

Duration analysis concerns the use of data which measures the time taken for a particular state or action to be completed. For example, a researcher may be interested in finding out what is the effect of a person's age on the duration of his unemployment spell. A similar question for which duration analysis is applicable would be "Does the probability of a strike ending depend positively or negatively on the elapsed duration?"

[2] See Mc Culloch et al., (1984) and Kennan (1985).

Such data is commonly provided by population surveys. In such surveys, information on the length of unemployment, for example, of a sample of individuals is collected. Figure 1 below illustrates a special concern whenever such data is collected called censoring.



SPELL t_4 IS RIGHT CENSORED

SPELL t_4 IS LEFT CENSORED

Fig 1

The survey covers the period of time given by distance AB on the horizontal axis. Spell t_1 starts and is completed within the survey period, as is spell t_2. However, the date of completion of spell t_3 is outside the survey period and so is unknown. Spell t_3 is said to be right-censored, and its recorded length will be from its time of commencement to time period B. Conversely, spell t_4 which began before the survey period and ended during the survey, is said to be left-censored.

The censored observations pose special problems for regression analysis. In surveys which collect information only on individuals unemployed at the time of the survey, there may be a bias towards longer spells as they may be more frequently selected over shorter spells. This is called length-biased sampling.

Secondly, as Kiefer (1988) argues, if censored observations are treated as complete spells in a regression analysis, then the estimates derived from the regression will have an upward asymptotic bias.

Even in the absence of censoring, duration data poses special problems for regression analysis. If a model is used which expresses the length of spells as a function of a number of explanatory variables, the problem arises as to how to measure those variables whose values change over the duration of each spell.

These peculiar problems of duration data have stimulated the development in the literature of specific methods to handle duration data. Unlike classical regression analysis the central concept of these methods is the conditional probability of an event taken place. For example, they concern the probability that a person will be employed in the fifth month given that he had previously been unemployed for four months. Classical regression analysis on the other hand deals with the unconditional probabilities of an event occurring.

Out of these conditional probabilities the Hazard Function, denoted $\theta(t)$, may be specified. This may be defined as the probability that a spell ends at duration t, given that it lasts until t. Another way to look at the hazard is the rate at which spells will be completed at duration t, given that they last until t. We may write this as:-

$$\theta(t) = \lim_{h \to o} \frac{Pr(t \leq T < t+h \mid T \geq t)}{h}$$

where T is a random variable denoting duration.

The hazard function may also be written in terms of the density and distribution functions:

$$\theta(t) = \frac{f(t)}{1-F(t)}$$

Associated with the hazard function is the Survivor Function, denoted S(t). The survivor function gives the probability that the duration of an event (random variable T) will equal or exceed the value t. In terms of the distribution function we write:

$$S(t) = 1-F(t)$$

Hence the hazard function can be expressed as a function of the density and survivor functions as:

$$\theta(t) = \frac{f(t)}{S(t)}$$

The hazard and survivor functions therefore are important in revealing the pattern in the conditional probabilities of an event or sequence of events occurring. The hazard functions enable us to infer about duration dependence. Positive duration dependence exists at time t* if the hazard has a positive slope at that point, ie, if:

$$\frac{d\theta(t)}{dt} > 0 \text{ at } t = t^*$$

Positive duration dependence means that the probability of a spell ending shortly increases as the spell increases in length. This is a desirable quality for unemployment spells.

Negative duration dependence occurs where:

$$\frac{d\theta(t^*)}{dt} < 0$$

ie, where the hazard is downward sloping. This means that the probability that a spell ends shortly decreases as its length increases.

Estimation

Estimation of the above hazard and survivor functions may be classified into three types: non-parametric, parametric, and semi-parametric.

The two main non-parametric estimators used are the Kaplan-Meier estimator and Life Tables. For the Kaplan-Meier estimator, if we let:

h_i = the number of completed spells of duration I
m_i = the number of observations censored between time t_i and t_{i+1}

then the number of spells neither completed or censored before duration t; is given by

$$n_j = \sum_{i \geq j}^{k} (m_i + h_i)$$

and the Kaplan-Meier hazard and survivor functions are given by

$$\theta(t_j) = h_j/n_j \qquad \text{and} \qquad S(t_j) = \prod_{I=1}^{j} (1 - \theta i)$$

respectively.

For the Life Tables hazard, the range of t is divided into k equal intervals where h denotes the internal width. We further define

N = total number of observations

C_j = number of observations censored at time j

d_j = number of observations which 'exit' at time j

Hence we define the size of the risk set r_j as $r_j = N_j - C_j/2$ and let $q_j = d_j/r_j$ be the proportion of observations in the risk set which exited, then the hazard rate is given by

$$\theta(t) = \frac{2q_j}{h(2-q_j)}$$

The Kaplan-Meier estimor and Life Tables are used mainly for preliminary analysis of the data, ie, to see how the hazards look and to get an idea of their functional form. Once this is done we may attempt to fit known distributions to the hazard function through parametric methods. Some common examples of distributions are the Exponential, Weibull and Log-logistic distributions.

Exponential Distribution

The hazard and survivor for the exponential distribution are given by

$$\theta(t) = \gamma \qquad \text{and} \qquad S(t) = \exp(-\gamma t)$$

The hazard function is constant and so reflects no duration dependence. It has one parameter - γ. The expected or average duration in this model is given by E(T) - $1/\gamma$. One drawback of using the exponential is that the family of distributions that may be realised from the one parameter γ is very limited. The model would fit data which exhibits the characteristic of a fairly constant hazard rate.

Weibull Distribution

The Weibull distribution is characterised by two parameters: γ and α. The hazard and survivor functions are

$$\theta(t) = \gamma\alpha t^{\alpha-1} \qquad \text{and} \qquad S(t) = \exp(-\gamma t^{\alpha})$$

The shape of the hazard depends on the size of the parameter α. If $\alpha > 1$ the hazard function increases monotonically in duration and if $\alpha < 1$ it decreases. If $\alpha = 1$ then the hazard is constant and we get the exponential distribution. Hence the Weibull distribution is a simple generalisation of the exponential distribution.

Log-logistic Distribution

The log-logistic distribution allows for a hazard which is non-monotic. The hazard and survivors are given by

$$\theta(t) = \frac{\gamma\alpha t^{\alpha-1}}{(1 + t^{\alpha}\gamma)} \qquad \text{and} \qquad S(t) = \frac{1}{1 + t^{\alpha}\gamma}$$

Again it is a two parameter distribution, but with $\alpha, \gamma > 0$. For $\alpha > 1$ the hazard first increases with duration then decreases. With $0 < \alpha \leq 1$ the hazard decreases with duration.

Other distributions which may be found in the literature include the nomal, gompertz and gamma. However we limit the analysis to the those described above. Estimation of the parameters of these distributions is done by maximum likelihood.

Semi-Parametric estimators include the Cox Partial Likelihood method and the Piecewise Constant models. However we do not use these estimators in our analysis as they commonly require the use of exogenous explanatory variables.

Section III - Data

The data utilised is taken from a survey which was part of a study on the impact of tax administrative reform in St Vincent and the Grenadines and Dominica. The data involves the time taken for individual taxpayers to submit their tax return. Five years data were collected for each individual covering the period 1991 - 1995. The sample sizes of the panel were as follows:

Table 1: Sample sizes of Tax Data

DOMINICA			ST VINCENT & THE GRENADINES	
YEAR	NO. OF OBS	NO. OF OBS CENSORED	NO. OF OBS	NO. OF OBS CENSORED
1991	589	19	1476	97
1992	1766	91	1539	108
1993	1866	53	1435	80
1994	1868	57	1198	74
1995	1795	2	1183	19

* Durations in excess of 364 days were censored at 365 days or 1 year.

All estimation was done using LIMDEP.

Section IV - Results

The estimated hazard functions and tables of parametric estimates are provided in the appendix. From the results of the non-parametric estimation it is clear that the hazards for the duration data are non-standard. The hazards appear to increase up to a peak around the 3 month period and then fall sharply afterwards. This suggests that on average most taxpayers take about 3 months to submit their tax returns and then afterwards the rate of submission falls drastically. This is true for both Dominica and for St Vincent and the Grenadines. The frequent fluctuations on the hazard functions however, as well as the fact that they appear to be non-standard, makes it very difficult to assess duration dependence and to determine whether the hazards have been affected by changes in the countries administrative tax procedures.

To assess whether there have been any improvements in the rate of collection over the five year period we therefore looked at the median survival times from the non parametric estimation. Apart from the year 1994 the median survival time in St Vincent and the Grenadines appears to be declining providing some evidence that individuals are submitting their returns at an earlier date on average. Although the results for Dominica seem less conclusive at first, if median value for 1991 is discounted because of the smaller sample size in that year, then a declining pattern is also evident from 1992/3 - 1995.

As expected the parametric models did not perform too well. The plots of the estimated parametric hazards do not resemble the non-parametric hazards at all[3]. The exponential hazard is a horizontal line with an average parameter estimate of 0.01282 over the five year period. This translates to an average duration for returns of 78 days. Even if this hazard was not representative of its non-parametric counterpart, it is interesting to note that the

[3] Apart from these visual tests, specification tests using the estimated integrated hazard function were performed, however the conclusions remained the same.

parameter estimates increase over the last four years in Dominica translating into shorter durations over that period.

As in the case of the exponential, the Weibull hazards did not fit their non-parametric counterparts at all. These hazards are monotonically increasing and do not capture the peaks at the three month period or the substantial drop afterwards. The closest fit was obtained with the log-logistic estimated hazard which peaked around the 90 day period and then fell away - but only gradually. This is the most promising parametric estimate so far and suggests the need to consider other non-monotonic distributions in future parametric estimation.

Section V - Conclusions

This paper attempted to analyse data on the time taken for the submission of personal tax returns in Dominica and St Vincent & the Grenadines using models of duration data. Such analysis is important to these two countries who are interested in finding out whether tax administrative reforms implemented over the period 1991 - 1995 were successful.

We used non-parametric techniques to plot the hazard functions for the respective countries, and then we attempted to fit these distributions using parametric estimation. From the non-parametric analysis it is clear that the hazard functions for the data are non standard. The plot shows a peak around the 90 day period suggesting that on average people take about 3 months to submit their tax returns. After that period the rate of submission falls drastically.

Median survival times fell over the five year period providing evidence, though not conclusive, that the duration before submission have become shorter on average.

Given the non standard nature of the non-parametric hazards, parametric estimation did not perform well. The Log-logistic hazard was the closest fit as judged by the residual plot, peaking at the 90 day period but then tailing off only gradually. As such we were not confident in any of the estimates of the parameters of the hazard function underlying the data.

This suggests the use of semi-parametric models in the future. Models like the Piecewise-Constant method are promising as they divide the durations into ranges and attempt to fit the best exponential model in each range. Such models may better capture the substantial variation in the hazards which give them their non standard nature.

Further research will also include data on the corporate sector to see whether tax administrative reforms have had any effect on corporate returns. We also intend to add explanatory variables to the model providing that the data is available.

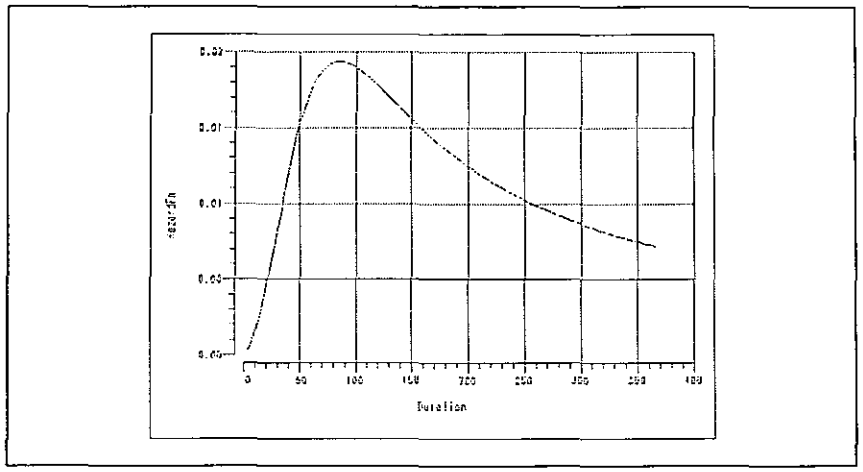
TABLE OF MEDIAN VALUES FROM THE ESTIMATED DISTRIBUTIONS

DOMINICA				ST VINCENT & THE GRENADINES			
YEAR	MEDIAN			YEAR	MEDIAN		
	EXP	WIEB	LOG		EXP	WIEB	LOG
1991	47.51	55.09	53.98	1991	73.34	89.9	77.49
1992	63.84	74.42	70.63	1992	73.36	85.22	77.08
1993	59.13	74.71	73.55	1993	68.34	81.54	74.41
1994	55.4	67.83	66.75	1994	73.45	87.59	80.32
1995	48.05	65.12	64.41	1995	62.25	79.33	76.15

TABLE SHOWING MEDIAN VALUES OF ESTIMATED SURVIVAL FUNCTION

DOMINICA		ST.VINCENT & THE GRENADINES	
YEAR	MEDIAN	YEAR	MEDIAN
1991	87.044	1991	104.515
1992	98.242	1992	103.224
1993	99.152	1993	100.719
1994	96.458	1994	105.684
1995	93.907	1995	102.112

LOG-LOGISTIC DISTRIBUTION - HAZARD FUNCTION



EXPONENTIAL DISTRIBUTION - HAZARD FUNCTION

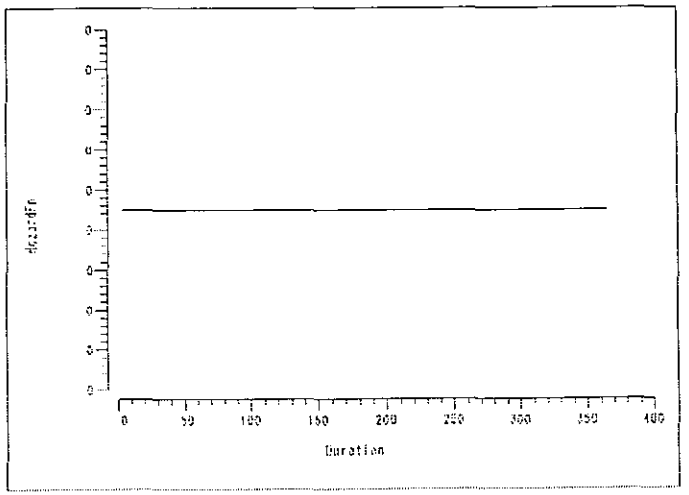


TABLE OF PARAMETERS FOR LOG-LOGISTIC DISTRIBUTION

DOMINICA			ST VINCENT & THE GRENADINES		
YEAR	PARAMETER		YEAR	PARAMETER	
	γ	α		γ	α
1991	0.01852	2.41967	1991	0.01291	3.042
1992	0.01416	2.65392	1992	0.01297	2.97454
1993	0.01359	3.78465	1993	0.01344	3.3162
1994	0.01498	3.0089	1994	0.01245	3.11384
1995	0.01553	3.33777	1995	0.01313	3.47233

TABLE OF PARAMETERS FOR EXPONENTIAL DISTRIBUTION

DOMINICA		ST VINCENT & THE GRENADINES	
YEAR	PARAMETER γ	YEAR	PARAMETER γ
1991	0.01459	1991	0.00945
1992	0.01086	1992	0.00945
1993	0.01172	1993	0.01014
1994	0.01251	1994	0.00904
1995	0.01442	1995	0.0113

WEIBULL DISTRIBUTION - HAZARD FUNCTION

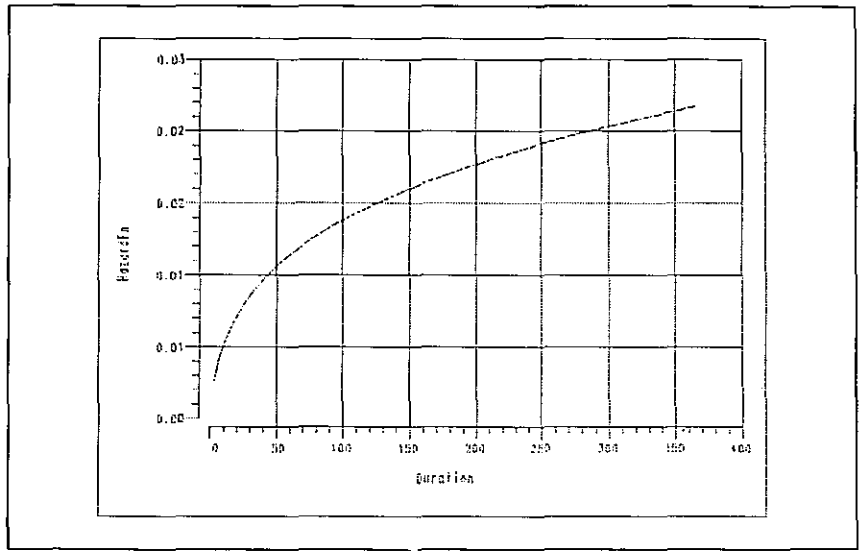
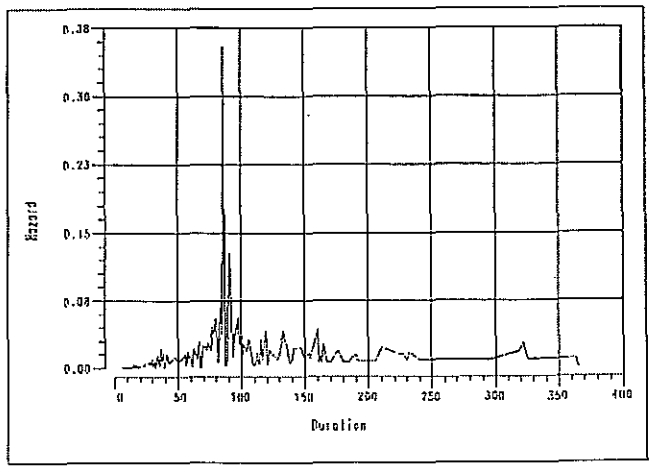
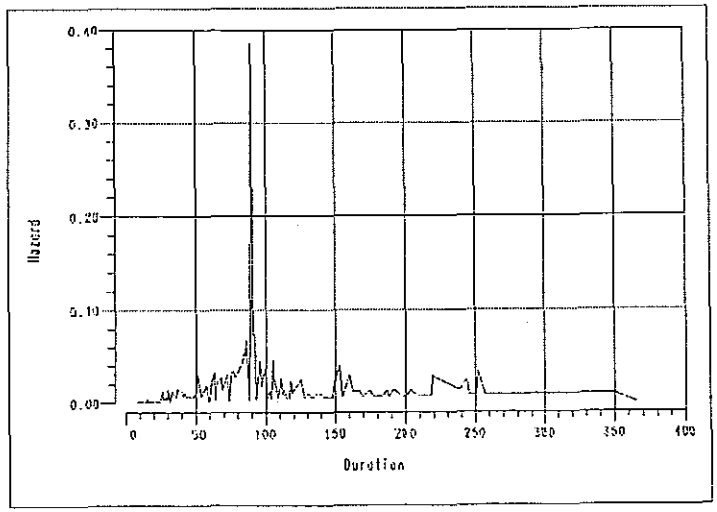


TABLE OF PARAMETERS FOR WEIBULL DISTRIBUTION

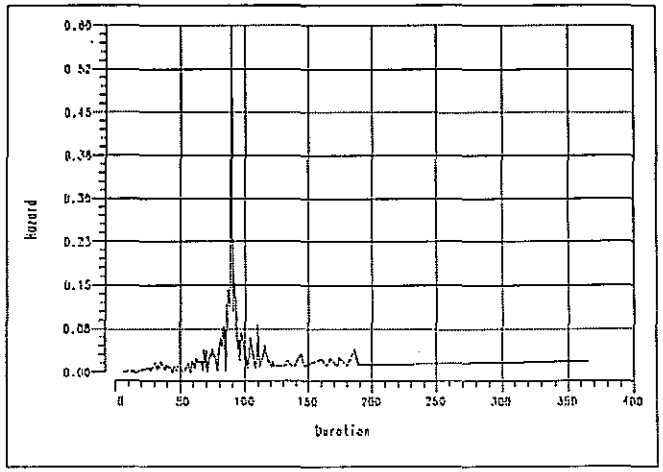
DOMINICA			ST VINCENT & THE GRENADINES		
YEAR	PARAMETER		YEAR	PARAMETER	
	γ	α		γ	α
1991	0.01361	1.27124	1991	0.00884	1.33204
1992	0.01015	1.30536	1992	0.00888	1.31244
1993	0.01063	1.58873	1993	0.00938	1.37431
1994	0.01145	1.45123	1994	0.00877	1.38698
1995	0.01282	2.02671	1995	0.01003	1.60363

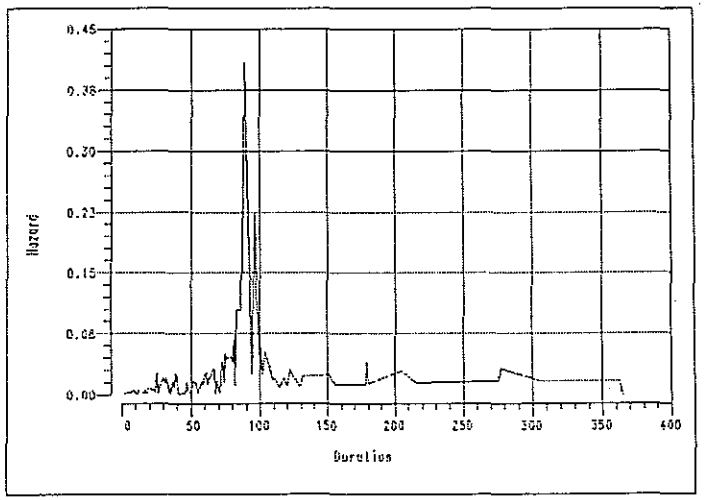
ST VINCENT &THE GRENADINES HAZARD FUNCTIONS - 1991, 1992



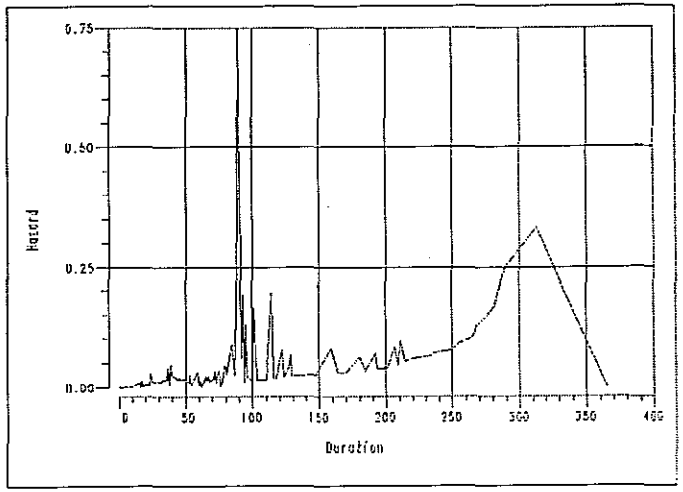


DOMINICA HAZARD FUNCTIONS - 1993,1994

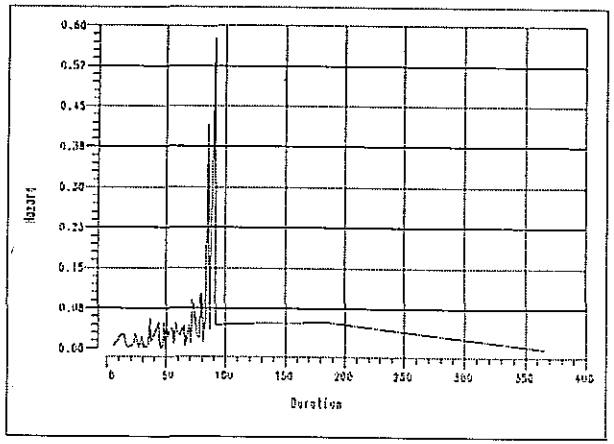


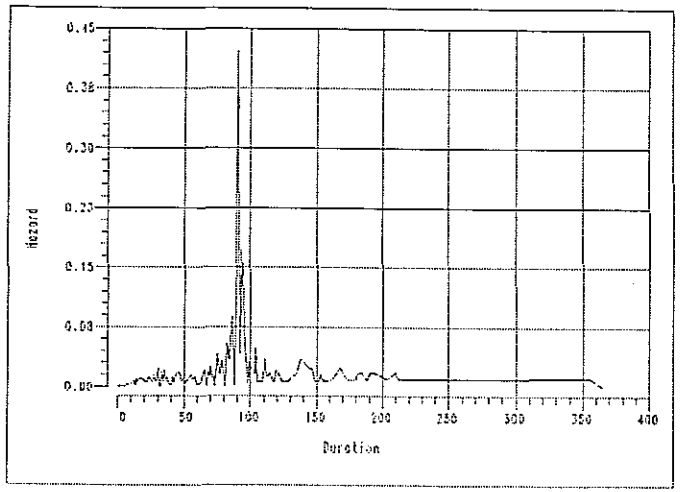


DOMINICA HAZARD FUNCTION - 1995

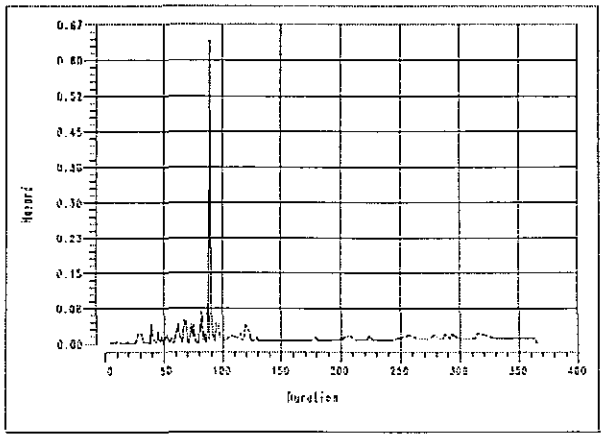


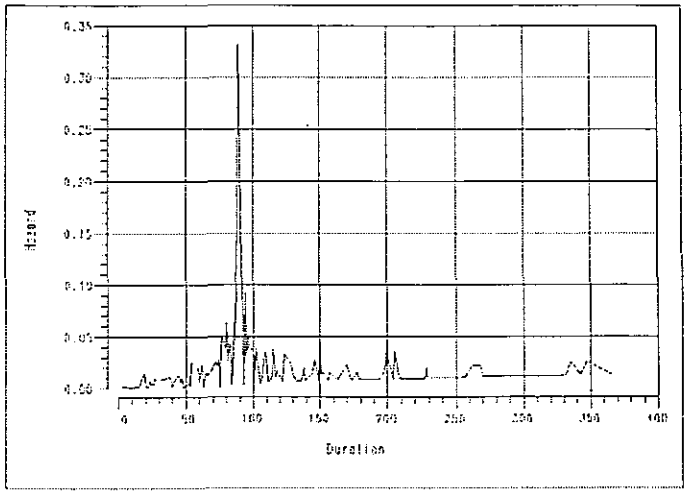
DOMINICA HAZARD FUNCTIONS - 1991, 1992



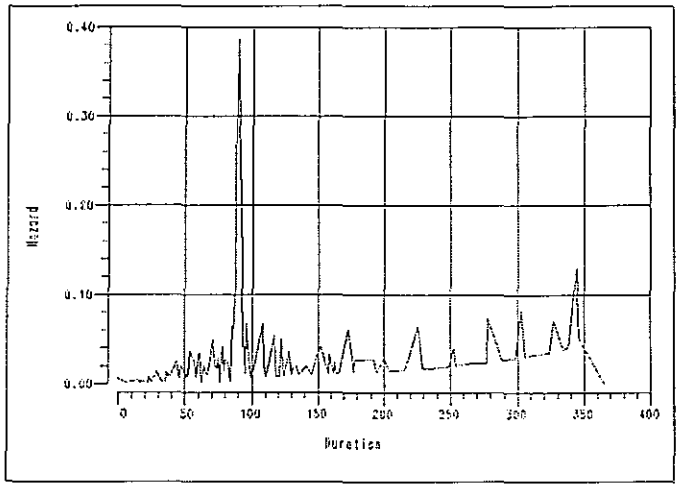


ST VINCENT & THE GRENADINES HAZARD FUNCTIONS - 1993, 1994





ST VINCENT & THE GRENADINES HAZARD FUNCTION - 1995



BIBLIOGRAPHY

ECCB/ECEMP 1997 "The Impact of Administrative Tax Reform".

Gloeckler, L.A. & R.L. Prentice. 1978. "Regression Analysis of Grouped Survival Data With Application to Breast Cancer Data". Biometrica 34.

Heckman, J.J. and G.J. Borja. 1980. "Does Unemployment Cause Future Unemploment? Definitions, Questions and Answers from a Continuous Time Model of Heterogeneity and State Dependence". Economica 47.

Kennan, J. 1985. "The Duration of Contract Strikes in U.S. Manufacturing". Journal of Econometrics 28, No.1.

Kiefer, N.M. 1988. "Economic Duration Data and Hazard Functions". Journal of Economic Literature XXVI.

Lancaster, T. 1979. "Econometric Methods For the Duration of Unemployment". Econometrica 47, #4.

----------. 1990. The Econometric Analysis of Transition Data. Cambridge England: Cambridge University Press.

McCulloch, C.E. and J.L. Newman. 1984. "A Hazard Rate Approach to The Timing of Births". Econometrica 52, No.4.