NBER WORKING PAPER SERIES

PREDICTING CRIMINAL RECIDIVISM USING "SPLIT POPULATION" SURVIVAL TIME MODELS

Peter Schmidt

Ann Dryden Witte

Working Paper No. 2445

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 1987

This research was supported by the National Institute of Justice, U.S. Department of Justice. The Institute's support does not indicate their concurrence with our methods or conclusions. The research reported here is part of the NBER's research program in Labor Studies. Any opinions expressed are those of the authors and not those of the National Bureau of Economic Research.

NBER Working Paper #2445
November 1987

Predicting Criminal Recidivism Using "Split Population" Survival Time Models

ABSTRACT

In this paper we develop a survival time model in which the probability of eventual failure is less than one, and in which both the probability of eventual failure and the timing of failure depend (separately) on individual characteristics. We apply this model to data on the timing of return to prison for a sample of prison releasees, and we use it to make predictions of whether or not individuals return to prison. Our predictions are more accurate than previous predictions of criminal recidivism. The model we develop has potential applications in economics; for example, it could be used to model the probability of default and the timing of default on loans.

Peter Schmidt
Department of Economics
Michigan State University
East Lansing, MI 48824-1038

Ann Dryden Witte
Department of Economics
Wellesley College
Wellesley, MA 02181

1. Introduction

During the 1970's and early 1980's, evidence accumulated indicating that a relatively small group of offenders committed most serious offenses. These findings, coupled with increasing pressures on the budgets of criminal justice agencies, led to calls for more effective use of the public expenditures for crime control by identifying and incarcerating the most serious and persistent offenders. The extremely influential work of Greenwood (1982) is a good example of the research promoting a policy of "selective incapacitation." However, the success of such a policy clearly depends on the ability to predict accurately *ex ante* (at the time a sentencing or parole decision is to be made) which individuals would return to crime if released. Thus there has been a resurgence of interest in the question of how well one can predict criminality at the individual level. A good survey of recent work on prediction in criminology is given by Farrington (1987), who concludes that predictive ability to date is rather disappointing, with most predictive models yielding false positive and false negative rates *both* in excess of 50%.

In this paper, we generate predictions of whether or not an individual will return to prison using survival time (or "failure time") models. Surprisingly, survival models have not been used much in criminology, and explanatory variables have almost never been included in those survival models that have been used. Our predictions are therefore based on a more sophisticated statistical model than previous researchers have considered. Encouragingly, we predict return to prison more accurately than has been done in the past. However, the accuracy of our predictions is (in our opinion) still not sufficient to justify a policy of selective incapacitation.

Such a predictive exercise is of substantial interest to criminologists, but probably not to most economists. However, the model which we develop to make our predictions is novel and has potential applications in a number of areas of economics. Specifically, we consider a "split population model" in which it is assumed that some fraction of the sample would never return to prison, so that the distribution of time until return is relevant only for the remaining fraction of the sample who would eventually return. Split models tend to imply very rapidly decreasing hazard rates (because the surviving population is implicitly made up increasingly of individuals who will never fail), and they are very useful in our application because our hazard rate does indeed fall very rapidly. Furthermore, we parameterize both the probability of eventual return and the timing of return, so that we can make separate statements about the effects of explanatory variables on these two conceptually different aspects of recidivism. For example, we find that race and sex affect the probability of eventual recidivism but not its timing, while two indicators of the nature of the previous offense affect the timing of recidivism (for the eventual recidivists) but not the probability of eventual recidivism.

It is not hard to think of potential economic applications of our model. For example, in the credit-scoring problem considered by Boyes, Hoffman and Low (1988), we might wish to estimate separately the effects of individual characteristics or of features of the loan itself on the probability of eventual default and on the timing of default for those individuals who will eventually default. A split model may be very reasonable for this application because many individuals would in fact never default, no matter how long they were observed. Furthermore, while most credit-scoring analyses focus only on the probability of eventual default, the likely timing of default is also relevant to the expected profitability of a potential loan, and therefore should also be of use in deciding whether to grant credit. A traditional credit-scoring analysis that focuses only on the probability of default may fail to give proper weight to individual characteristics that affect the timing of default (conditional on eventual default), but that do not affect the probability of eventual default.

More generally, our model may be useful in the analysis of the timing of any event which does not occur for a substantial fraction of the sample. For example, if we are interested in the duration of spells of employment, it should be recognized that a substantial proportion of individuals will never be unemployed. Models which fail to recognize this point will almost surely misspecify the distribution of survival times. They will underpredict the proportion of always-employed individuals, and (because they are misspecified) they may give misleading estimates of the effects of explanatory variables.

2. Data

The data used in this paper consist of information on a cohort of releasees from the North Carolina prison system. This cohort consists of all individuals released from North Carolina prisons from July 1, 1977 through June 30, 1978. There were 9457 such individuals. This data set is far larger and more timely than is usual in criminal justice research, and it is clear what population it represents.

We also obtained and analyzed data on a second similar cohort of releasees, but to save space we will not report these results here. Further details on the results for our second cohort (and, indeed, on all aspects of our research project) can be found in Schmidt and Witte (1987, 1988).

There were 130 observations in our data that were obviously defective and had to be discarded. In almost all cases, the defect in the data leading to elimination of the observation is that the individual was in fact not released from prison during the time period which defined the data set. It is important to note that the number of defective cases was only slightly more than one percent of the original number of cases. This is a very low discard rate for release cohort data and attests to the high quality of the North Carolina record keeping system.

A more serious problem is that many observations lacked information on one or more variables which we used in our analyses. Only 4618 observations contained information on all variables of interest, while the other 4709 observations lacked some information. The most commonly missing piece of data was information on alcohol or drug abuse, which turns out to be a very significant predictor of post-release criminality; 4287 of the 4709 incomplete observations lacked this information. We discarded the incomplete observations entirely, and analyzed only those which were complete. Clearly, this raises the possibility of selectivity bias in our results, but we preferred this to the omission of a very important explanatory variable.[1]

Having discarded the incomplete observations, we split the sample of complete observations randomly into an "estimation sample" (or "analysis sample") of 1540 observations and a "validation sample" of 3078 observations.[2] We fit our statistical models to the estimation sample, and then used the validation sample to check the predictive accuracy of these models. This procedure reflects the generally accepted view that the predictive accuracy of a model can be checked validly only on data not used to estimate the model.

We will now define the variables used in our study. The dependent variable which we seek to explain is the length of time from an individual's release from prison in North Carolina until his or her return to prison there. Therefore the outcome variables which we define are an indicator of whether the individual returned to prison within the followup period and the length of time until return for those individuals who did return, while the explanatory variables are demographic characteristics and measures of the past criminal and correctional histories of the individuals.[3]

To be more specific, recall that our data set was defined by date of release from prison. The sentence from which the individuals were released will be called the <u>sample sentence</u>, and the conviction which resulted in the sample sentence will correspondingly be called the <u>sample conviction</u>. All explanatory variables are defined either as of the time of entry or as of the time of release from the sample sentence. The outcome variables were defined as the result of a search of North Carolina Department of Correction records in April, 1984. Thus the followup period ranged from 70 to 81 months. This followup period is quite long for a study of recidivism; most studies follow releasees for three years or less.

We define the following outcome variables, for each individual:

FOLLOW, the length of the followup period, in months.

RECID, a dummy variable equal to one if the individual returned to a North Carolina prison during the followup period, and equal to zero otherwise.

TIME, the length of time from release from prison until return to prison, rounded to the nearest month, for individuals for whom **RECID** = 1. **TIME** is undefined for individuals for whom **RECID** = 0.

We now define the following explanatory variables, for each individual:

TSERVD, the time served (in months) for the sample sentence.

AGE, age (in months) at time of release.

PRIORS, the number of previous incarcerations, not including the sample sentence, at the time of entry into the prison system for the sample sentence.

RULE, the number of prison rule violations reported during the sample sentence.

SCHOOL, the number of years of formal schooling completed at the time of entry into the prison system for the sample sentence.

WHITE, a dummy variable equal to zero if the individual is black, and equal to one otherwise.

MALE, a dummy variable equal to one if the individual is male, and equal to zero if female.

ALCHY, a dummy variable equal to one if the individual's record indicates a serious problem with alcohol (before entry into the prison system) and equal to zero otherwise.

JUNKY, a dummy variable equal to one if the individual's record indicates use of hard drugs (before entry into the prison system) and equal to zero otherwise.

MARRIED, a dummy individual equal to one if the individual was married at the time of entry into prison for the sample sentence, and equal to zero otherwise.

SUPER, a dummy variable equal to one if the individual's release from the sample sentence was supervised (e.g., parole), and equal to zero otherwise.

WORKREL, a dummy variable equal to one if the individual participated in the North Carolina prisoner work release program during the sample sentence, and equal to zero otherwise.

FELON, a dummy variable equal to one if the sample sentence was for a felony, and equal to zero if it was for a misdemeanor.

PERSON, a dummy variable equal to one if the sample sentence was for a crime against a person, and equal to zero otherwise.

PROPTY, a dummy variable equal to one if the sample sentence was for a crime against property, and equal to zero otherwise.[4]

3. Models Without Explanatory Variables

We begin our analysis by fitting various parametric models to our estimation sample, and checking how well the models fit the estimation sample and how well they predict the actual outcomes in our validation sample. This is a standard exercise in the criminological literature (see, for example, Maltz (1984) and the references therein), but there has not been sufficient attention given to the question of how well commonly-used distributions fit the data.

A useful first step is to have a look at the nature of the empirical distribution of time until recidivism. The solid line in Figure 1 gives the empirical (actual) density for the validation sample, and the density for the estimation sample would look more or less the same. Similarly, a graph of the hazard rate would again reveal more or less the same pattern. Specifically, it is important to note two important features of the hazard rate in our data. First, it is non-monotonic; the hazard first rises and then falls.[5] Second, once it begins to fall, the hazard rate falls very quickly. These features of the data are common in applications involving recidivism, and perhaps in economics as well, but they are not common in the biological or reliability applications typically discussed in the statistical literature on survival times. As a result, distributions typically found in standard texts like Kalbfleisch and Prentice (1980) or Lawless (1982) do not fit our data well.

We fit five different distributions to the data, by maximum likelihood: exponential, Weibull, lognormal, loglogistic and LaGuerre. The exponential and Weibull distributions are commonly found in the survival time literature, but they should not be expected to do well here because they can not generate a non-monotonic hazard rate. The lognormal and loglogistic distributions, on the other hand, imply a hazard that first rises and then falls, just as in our data. The LaGuerre model (Cox and Oates (1984, p. 20), Kiefer (1985)) has a density which is the product of an exponential function and a polynomial; this is known as a LaGuerre polynomial in the mathematical literature. It is intended to give a flexible approximation to an arbitrary density.[6] We use a second-degree LaGuerre distribution (that is, it contains a second-degree polynomial).

None of these distributions fits our data adequately. In each case they overpredicted recidivism for the first few months after release and they underpredicted it during the intermediate period of roughly one to three years after release. Furthermore, all of the distributions except the LaGuerre overpredicted recidivism in the tail of the distribution. The lognormal and loglogistic distributions fit noticably better than the exponential or Weibull, as expected, but they still did not generate a sufficiently rapid increase in the hazard at first or a sufficiently rapid decrease in the hazard in the tail of the distribution. The LaGuerre model fit better than any of the others, especially in the tail, but it still was quite inadequate for the first two years after release.

The superiority of the LaGuerre model is evident from the likelihood values achieved, which were -3431, -3405, -3370, -3390 and -3356, for the exponential, Weibull, lognormal, loglogistic and LaGuerre distributions, respectively. It is also clear from any reasonable measure of the quality of the predictions generated for the validation sample. For example, if we measure the quality of these predictions by the maximum difference between the predicted and the actual cdf, the value for the LaGuerre distribution is .020, compared to .079, .049, .037, and .040 for the other four distributions.

The dashed line in Figure 1 displays the density of time until recidivism predicted by the LaGuerre distribution. The inadequacy of the fit is evident. While we could continue to experiment with other distributions, one lesson that should be clear is that "off-the-shelf" models from the biostatistical or operations research literatures are not necessarily adequate for applications in other fields.

4. Split Models

The parametric models considered in the last section all assumed some form of the cumulative distribution function for the time until recidivism. Any such cumulative distribution function approaches one as time at risk becomes sufficiently large. In the present context, this implies that every individual would eventually return to prison, and this implicit assumption can be argued to be unreasonable. In this section, we will consider "split population models" (or simply "split models") in which the probability of eventual recidivism is an additional parameter to be estimated, and may be less than one. A distribution of failure times is also specified, as before, but this is understood to apply only to those individuals who will eventually fail. Split models were introduced to the criminological literature by Maltz and McCleary (1977), with previous treatments in the statistical literature dating back to Anscombe (1961), and they have been further developed in a line of research well summarized by Maltz (1984) or Schmidt and Witte (1988, chapter 5). They do not appear to have been used in economics, but in our opinion they are likely to be useful there as well.

Using the notation of Schmidt and Witte (1984, section 6.4), we can express a split model as follows. First, let F be an unobservable variable indicating whether an individual would or would not eventually fail. Specifically, let F equal one for individuals who would eventually fail, and zero for individuals who would never fail. Then we assume

(1) $P(F = 1) = \delta \quad , \quad P(F = 0) = 1 - \delta \quad .$

The parameter δ is of course the eventual recidivism rate. Second, we assume some cumulative distribution function $G(t|F=1)$ for the individuals who would ultimately fail, and we let $g(t|F=1)$ be the corresponding density. We note explicitly that such a distribution is defined <u>conditional</u> on F=1, and is irrelevant for individuals for whom F=0.

Now let **T** be the length of the followup period, and let **C** be the observable dummy variable indicating whether or not the individual has returned to prison by the end of the followup period. For the recidivists in the sample, we observe **C** = 1 and the failure time **t**, and of course we know that **F** = 1. The appropriate density is therefore

(2) $P(F=1)\ g(t|F=1) = \delta\ g(t|F=1).$

On the other hand, for the non-recidivists in the sample we observe only C=0, and the probability of this event is

(3) $$P(C=0) = P(F=0) + P(F=1)P(t>T|F=1)$$
$$= 1 - \delta + \delta\ [1 - G(T|F=1)].$$

The likelihood is then made up of terms like (2) for recidivists and (3) for non-recidivists (individuals who have not returned to prison by the end of the followup period).

We fit split models to our data using the same five distributions as were considered in the previous section (exponential, Weibull, lognormal, loglogistic and LaGuerre). The likelihood values achieved were -3358, -3346, -3342, -3341, and -3349, respectively, while the maximum differences between the actual and the predicted cdf in the validation sample were .024, .017, .005, .011 and .021. The loglogistic model fits the estimation sample slightly better than the lognormal model, while the lognormal model generates slightly better predictions for the validation sample, but there is in fact little basis on which to choose between the two distributions, and either of them dominates the other three distributions considered. However, what is more striking is the extent to which the split models dominate the simple (not split) models of the previous section. In terms of likelihood value or quality of predictions, the worst of our split models (split exponential) is comparable to the best of our simple models (LaGuerre), even though it contains fewer parameters. The best of our split models, say the split lognormal, is <u>much</u> better than any of the models of the last section. Introduction of the splitting parameter into the lognormal model increases the likelihood value by approximately 50, and reduces the maximum difference between the predicted and the actual cdf from .037 to .005; these are certainly impressive improvements in the model to be achieved by the introduction of a single parameter.[7]

The dotted line in Figure 1 gives the predicted density for the split lognormal model. The fit of the model to the data appears to be quite adequate, and this conclusion is confirmed by more formal tests reported in Schmidt and Witte (1988, section 5.3).

The value of the "splitting parameter" δ generated by the split lognormal model is 0.45. This is the long-run (eventual) recidivism rate. By way of contrast, the long-run recidivism rate is by definition equal to one in non-split models, and this is reflected in very large implied recidivism rates for long but finite followups. For example, the 25-year recidivism rates implied by our models of the last section were 0.94, 0.89, 0.78, 0.78, and 0.52 for the exponential, Weibull, lognormal, loglogistic and LaGuerre distributions, respectively. Based on a limited number of studies with very long followup periods, such as McCord (1978) and Kitchener, Schmidt and Glaser (1977), a long-run failure rate of approximately 0.5 appears to be reasonable for our data and definition of recidivism.

5. Models With Explanatory Variables

We now consider models with explanatory variables. This is obviously necessary if we are to make predictions for individuals, or even if we are to make potentially accurate predictions for groups which differ systematically from our original sample (for example, to evaluate the effectiveness of a correctional program which is applied to a non-random sample of the population of releasees). Furthermore, in many applications in economics or criminology the coefficients of the explanatory variables may be of obvious interest.

We begin by fitting the proportional hazards model to our data. The point of this exercise is to see which explantory variables are worth including, without making a specific distributional assumption. The estimates are based on the usual "partial MLE" method; the ties in the data are handled using the approximation of Peto (1972), as reported also by Kalbfleisch and Prentice (1980, equation (4.8)). Our estimates are given in Table 1, using the 15 explanatory variables defined in section 2. (Note that a few variables have been rescaled, to make the coefficients of a more conveniently magnitude.) The "t ratios" reported are the asymptotic standard normal statistics used to test the hypothesis that the coefficient is zero.

Looking under the heading "ORIGINAL SPECIFICATION," we see that six coefficients are individually insignificant at the 5% level. They are also jointly insignificant, as judged by the likelihood ratio test, and we dropped the corresponding six variables (**RULE**, **MARRIED**, **SCHOOL**, **WORKREL**, **PERSON**, and **SUPER**) from the model. Interestingly, an exponential model with the log of the mean depending linearly on the explanatory variables led to exactly the same decision about which variables to drop, and indeed to almost exactly the same t ratios and likelihood ratio statistic. Our results indicate that the type of individual most likely to have a small value of time until recidivism is a young, black male with a large number of previous incarcerations, who is a drug addict and/or alcoholic, and whose previous incarceration was lengthy and was for a crime against property. These findings, with the possible exception of the findings on race, are consistent with the conclusions of one of the two most comprehensive surveys available (Wilson and Herrnstein (1985)), with most of the conclusions of the second such survey (Blumstein et al. (1986)), and with our own previous work (Schmidt and Witte (1984)).

We now turn to a parametric model based on the lognormal distribution. As noted above, we also considered the exponential distribution, but it did not fit the data as well as the lognormal, so we will not discuss it here. The model in its most general form is a split model in which the probability of eventual recidivism follows a logit model, while the distribution of time until recidivism (conditional on eventual recidivism) is lognormal, with its mean depending on explanatory variables.

To be more explicit, we follow the notation of section 4. For individual **i**, there is an unobservable variable $\mathbf{F_i}$ which indicates whether or not individual **i** will eventually return to prison. The probability of eventual failure for individual **i** will be denoted $\boldsymbol{\delta_i}$, so that $P(F_i = 1) = \delta_i$. Let $\mathbf{X_i}$ be a (row) vector of individual characteristics (explanatory variables), and let $\boldsymbol{\alpha}$ be the corresponding vector of parameters. Then we assume a logit model for eventual recidivism:

$$(4) \qquad \delta_i = 1 \; / \; [1 + e^{X_i \alpha}] \quad .$$

Next, we assume that the distribution of time until recidivism (given eventual recidivism) is lognormal, so that the distribution of $\ln t_i$ is normal, with mean μ_i and variance σ^2. The mean of this distribution is assumed to depend on individual characteristics X_i, so that $\mu_i = X_i\beta$.

It should be noted that the way in which we have parameterized the model implies that a positive coefficient (in either α or β) indicates that the corresponding variable has a positive influence on time until recidivism (i.e., it makes recidivism either less likely or longer in coming, or both). This is the opposite of the case for the proportional hazards model, in which a positive coefficient indicates a positive effect on the hazard rate, and therefore a negative effect on the survival time. We should therefore expect (or hope) that most coefficients will be opposite in sign in this model as opposed to the proportional hazards model.

The likelihood function for this model is:

$$(5) \qquad \ln L = \sum_{i=1}^{N} \{ C_i [\ln \delta_i - 0.5 \ln (2\pi) - 0.5 \ln \sigma^2 - (\ln t_i - X_i\beta)^2/2\sigma^2] + (1 - C_i) \ln P_i \} ,$$

where

$$(6) \qquad P_i = 1 - \delta_i + \delta_i \Phi[(X_i\beta - \ln t_i)/\sigma] ,$$

where Φ is the standard normal cdf, and where C_i is the dummy indicator of return to prison during the followup period.

We can now define special cases of this general model. First, the model in which $\delta_i = 1$, but in which the mean time until recidivism depends on individual characteristics, will be called the <u>lognormal model</u> (with explanatory variables). This model has been considered by Kalbfleisch and Prentice (1980, Section 3.6) and Lawless (1982, Section 6.5), among others. Witte and Schmidt (1977) have used a very similar model to analyze recidivism. It is not a split model. Second, the model in which δ_i is replaced by a single parameter δ will be referred to as the split lognormal model (with explanatory variables). In this model the probability of eventual recidivism is a constant, though not necessarily equal to one, while the mean of the distribution of time until recidivism varies over individuals. Third, the model in which μ_i is replaced by a single parameter μ will be called the logit lognormal model. In this model the probability of eventual recidivism varies over individuals, while the distribution of time until recidivism (for the eventual recidivists) does not depend on individual characteristics. Finally, the general model as presented above will be called the logit/individual lognormal model. In this model both the probability of eventual recidivism and the distribution of time until recidivism vary over individuals.

In the lognormal, split lognormal and logit lognormal models, only one aspect of recidivism (probability or timing) depends on explanatory variables. Interestingly, these three models generate very similar results. Table 2 gives the results for the split lognormal model and the logit lognormal model. The results from these two models are very similar, as is evident at a glance, and are in turn very similar to those from the proportional hazards model (Table 1). In fact, this robustness of results goes beyond what is displayed in this paper. Essentially the same results are obtained from the lognormal model, and also from models like these three models but based on the exponential distribution.

However, while the choice of model does not have much effect on the estimated coefficients, there is considerable variation in the quality of the fit and the predictions which are generated. In both respects the lognormal models dominate the exponential models, and the logit lognormal model dominates the lognormal and split lognormal models. For example, the likelihood value of -3265 for the logit lognormal model is noticably higher than the values for the lognormal and split lognormal models (-3273 and -3256). The maximum difference between the predicted and the actual cdf is also smaller (.006 versus .030 and .034).

We now turn to the logit/individual lognormal model, in which both the probability of eventual recidivism and the distribution of time until recidivism vary according to individual characteristics. These parameter estimates are given in Table 3. They are somewhat more complicated to discuss than the results from our other models, in part because there are simply more parameters, and some of them turn out to be statistically insignificant. However, every coefficient that is statistically significantly different from zero has the expected sign (the same sign as in our previous models), and in that sense the results are still essentially the same as before.

In Table 3, we can see that four variables have statistically significant effects on both the probability of eventual recidivism and on the mean time until recidivism: **TSERVD**, **AGE**, **PRIORS**, and **ALCHY**. Three variables have statistically significant effects on the probability but not the timing of recidivism: **WHITE**, **JUNKY**, and **MALE**. The remaining two variables, **FELON** and **PROPTY**, have statistically significant effects on the timing of recidivism but not on the probability of eventual recidivism. Thus it appears that we are indeed able to separate out the effects of individual characteristics on the probability of eventual failure from their effects on the timing of failure (for those who will ultimately fail), an optimistic result.

Furthermore, these results are reasonably similar to the results we obtained using a logit/individual exponential model; see Table 4. The difference is that **AGE**, **PRIORS**, and **ALCHY** did not have significant effects on the mean time until recidivism, in the exponential case, while they did in the lognormal case. Thus our results are reasonably robust to our distributional assumptions. Perhaps not surprisingly, however, they are less robust in this model than they were in the simpler models previously considered.

The extent to which our results are sensitive to distributional assumptions is an important issue, because there is seldom much reason to believe strongly in one's distributional assumptions. However, a counterargument is that this simply indicates that one should take care in investigating the adequacy of such assumptions, which we have done.

6. Predictions for Individuals

We now return to the problem of prediction at the individual level, using the models estimated in section 5. This is a fairly standard use of such models; for example, the studies included in Farrington and Tarling (1985) include predictions of failure on parole, of recidivism, and of absconding from institutions for young offenders. The desire to make predictions for individuals undoubtedly derives from a desire to use such predictions as the basis for differential treatments for individuals. Because our data are on length of time until recidivism, it is natural for us to regard recidivism as the event to be predicted, and to ask how well our models predict it.

Recidivism is a discrete event, while our models yield a probability of this event for each individual. This immediately raises the question of how to summarize the accuracy of the models' predictions. One possibility is to use statistical measures (akin to correlations) of the degree of association between the models' probabilities and the observed binary outcome. While Kendall's tau has often been suggested for this purpose in the criminological literature, a more standard statistical measure would be the biserial correlation coefficient, which is a form of the correlation coefficient used when one variable is continous and the other is binary. However, we will not pursue such measures here, since they are not of obvious practical use.

A more readily interpreted summary of predictive success, in the present context, is simply to predict that individuals with probabilities above some chosen level will return to prison (and that individuals with probabilities below that level will not), and to calculate the error rate of these predictions. This seems to us to be reasonable because it evaluates the accuracy of exactly the procedure which would be followed in a practical application of these models.

Following the usual practice in the criminological literature (e.g., see Wilbanks (1985)), we begin by predicting the "correct" proportion of failures, and evaluate the extent to which we have correctly predicted which individuals will fail. The failure rate in the estimation sample is 0.366, and so we predict recidivism for the 36.6% of the validation sample who have the highest probabilities of recidivism (regardless of the absolute magnitudes of these probabilites). Basing our predictions on the proportional hazards model, we predict recidivism for 1127 individuals (36.6% of the 3078 individuals in the 1978 validation sample), of whom 595 actually returnded to prison and 532 did not; and we predict 1951 individuals not to return to prison, of whom 540 returned to prison and 1411 did not. We therefore have a false positive rate of 532 / 1127 = 0.472, and a false negative rate of 540 / 1951 = 0.277. Although we do not report the results here, the logit lognormal model and the logit/individual lognormal model generate more or less the same error rates.[8]

The predictive accuracy of our models compares quite favorably with the accuracy of the models recently surveyed by Farrington (1987). He reports that Greenwood (1982) had a false positive rate of 56% and a false negative rate of 46%

for his estimation sample. (Greenwood had no validation sample.) Janus' (1985) predictions resulted in an even poorer record, with a 62% false positive rate and a 64% false negative rate. Blumstein, Farrington and Moitra's (1985) false negative rate of 35% is lower than that of either Greenwood or Janus, although more than ten percentage points higher than our own, and their false positive rate is higher than either we or Greenwood obtain. However, while we are able to predict more accurately than the studies surveyed by Farrington, our false positive rate is in our opinion still much too high to justify using our models to implement a policy of selective incapacitation.

On the other hand, not all potential users of prediction need to predict recidivism for the "correct" proportion of the sample. For example, a policy of selective incapacitation might be considered viable if we could predict recidivism with considerable assurance even for a very limited proportion of the sample. Similarly, models like ours might be useful in deciding on candidates for early release if they could predict success (non-recidivism) with assurance, for some proportion of the sample.

In Table 5, we rank individuals by their predicted probabilities of recidivism (generated by the logit lognormal model) and then report the actual proportions of recidivists in groups representing various percentiles of the "distribution" of predicted probabilities of recidivism. For example, we can see in Table 5 that the 20% of the sample (616 individuals) with the highest predicted probabilities of recidivism had an actual recidivism rate of 59.9%, whereas the remaining 80% of the sample (2462 individuals) had a recidivism rate of 31.1%. Obviously our models have at least some predictive power, since individuals with higher predicted probabilities of failure do indeed fail more often than individuals with lower probabilites.

In considering a policy of selective incapacitation, it is important that there be some group of individuals whom we can predict to fail with near certainty. Using this model, any such group would have to be very small. For example, our predicted "worst" 1% of the sample has a recidivism rate of 83.9%, but it is a group of only 31 individuals. The recidivism rate falls to 70.1% if we include the 154 individuals in the upper 5% of the probability ranking. These probabilities of a false positive error seem rather high. A public official who is considering selective incapacitation for the "worst" 1% of a potential release cohort would probably not like to think that over 15% of those "worst" individuals would in fact not return to prison after a four-year followup.

We are much more successful in predicting individuals who will not fail. For example, in the group of 31 individuals who represent the predicted "best" 1% of the sample, the failure rate is only 6.5%. Even if we enlarge the group considerably to include 308 individuals (the 10% of the sample with the lowest predicted probabilities of recidivism), the failure rate is only 13.0%, and this false negative rate is much lower than the corresponding false positive rate (32.5%) for the corresponding upper 10% of the sample. The fact that it is easier for us to identify individuals who are likely to succeed than it is for us to identify individuals who are likely to fail is natural, because in our data less than half of all individuals fail.

7. Conclusions

One purpose of this paper is to interest econometricians and economists in "split-population" models which have been used in criminology. These are failure time models in which it is explicitly recognized that some individuals will never fail. Split models tend to imply a very rapidly falling hazard rate, and they will tend to fit well data with this characteristic. They tend to avoid the overprediction of the long-run failure rate, a major problem with more traditional failure time models in criminology. Furthermore, we can allow explanatory variables to affect both the probability of eventual failure and the timing of failure for those who will ultimately fail, so that we can separate out the effects of the explanatory variables on these two conceptually different features of behavior.

A second purpose of the paper is to make a contribution to the criminological literature on prediction of recidivism. There is substantial interest in the question of how well we can predict the future criminal behavior of individuals. Previous attempts to predict individual outcomes using statistical models have had somewhat mixed success: statistical methods have consistently outperformed clinical or judgemental methods of prediction, but have still suffered from high error rates. Most previous attempts at prediction in criminology have relied on rather simple models, and our analysis is an attempt to see whether we can improve the accuracy of such predictions by using a more sophisticated statistical model. We succeed in predicting recidivism more accurately than others have done. However, it is still not clear that predictions of the accuracy that we attain are useful in a practical sense. Unsurprisingly, there is still a need for better models and better data.

REFERENCES

Anscombe, F. J. (1961), "Estimating a Mixed-Exponential Response Law," Journal of the American Statistical Association, 56, 493-502.

Blumstein, A., D. Farrington and S. Moitra (1985), "Delinquency Careers: Innocents, Desisters and Persisters," Crime and Justice: An Annual Review of Research, 7, 187-220.

Blumstein, A., J. Cohen, J. A. Roth and C. A. Visher (1986), Criminal Careers and "Career Criminals", Washington, D.C.: National Academy Press.

Boyes, W. J., D. L. Hoffman and S. A. Low (1988), "An Econometric Analysis of the Bank Credit Scoring Problem," Journal of Econometrics, this issue.

Burgess, E. W. (1928), "Factors Determining Success or Failure on Parole," in A. A. Bruce, E. W. Burgess and A. J. Harno, The Workings of the Indeterminate Sentence Law and the Parole System in Illinois, Springfield: Illinois State Board of Parole.

Cox, D. R. and D. Oates (1984), Analysis of Survival Data, London: Chapman and Hall.

Farrington, D. P. (1987), "Predicting Individual Crime Rates," in Prediction and Classification: Criminal Justice Decision Making (Crime and Justice: A Review of Research), 9, 53-101.

Farrington, D. P. and R. Tarling (1985), Prediction in Criminology, Albany: State University of New York Press.

Goldfeld, S. M. and R. E. Quandt (1981), "Econometric Modelling with Non-Normal Disturbances," Journal of Econometrics, 17, 41-156.

Greenwood, P. (1982), Selective Incapacitation, Santa Monica, California: Rand Corporation.

Hoffman, P. B. and B. Stone-Meierhoefer (1979), "Post Release Arrest Experiences of Federal Prisoners: A Six-year Follow-up," Journal of Criminal Justice, 7, 193-216.

Janus, M. C. (1985), "Selective Incapacitation: Have We Tried It? Does It Work?" Journal of Criminal Justice, 13, 117-129.

Kalbfleisch, J. D. and R. L. Prentice (1980), The Statistical Analysis of Failure Time Data, New York: Wiley.

Kiefer, N. M. (1985), "Specification Diagnostics Based on LaGuerre Alternatives for Econometric Models of Duration," Journal of Econometrics, 28, 135-154.

Kitchener, H., A. K. Schmidt and D. Glaser (1977), "How Persistent is Post-prison Success?" Federal Protection, 41, 9-15.

Lawless, J. F. (1982), Statistical Models and Methods for Lifetime Data, New York: Wiley.

Lutkepohl, H. (1980), "Approximation of Arbitrary Distributed Lag Structures by a Modified Polynomial Lag: An Extension," Journal of the American Statistical Association, 75, 428-430.

Maltz, M. D. (1984), Recidivism, Orlando, Florida: Academic Press.

Maltz, M. D. and R. McCleary (1977), "The Mathematics of Behavioral Change: Recidivism and Construct Validity," Evaluation Quarterly, 1, 421-438.

McCord, J. (1978), "A Thirty Year Follow-up of Treatment Effects," American Psychologist, 284-289.

Peto, R. (1972), "Contribution to the Discussion of Paper by D. R. Cox," Journal of the Royal Statistical Society, Series B, 34, 472-475.

Schmidt, P. and W. R. Mann (1977), "A Note on the Approximation of Arbitrary Distributed Lag Structures by a Modified Almon Lag," Journal of the American Statistical Association, 72, 442-443.

Schmidt, P. and A. D. Witte (1984), An Economic Analysis of Crime and Justice: Theory, Methods, and Applications, Orlando, Florida: Academic Press.

Schmidt, P. and A. D. Witte (1987), How Long Will They Survive? Predicting Recidivism using Survival Models, Report to the National Institute of Justice.

Schmidt, P. and A. D. Witte (1988), Predicting Recidivism using Survival Models, New York: Springer-Verlag.

Wilbanks, W. L. (1985), "Predicting Failure on Parole," in Prediction in Criminology, D. P. Farrington and R. Tarling, editors, Albany: State University of New York Press.

Wilson, J. Q. and R. J. Herrnstein (1985), Crime and Human Nature, New York: Simon and Schuster.

Witte, A. D. and P. Schmidt (1977), "An Analysis of Recidivism, Using the Truncated Lognormal Distribution," Applied Statistics, 26, 302-311.

FOOTNOTES

1. A discussion of the extent to which selection into the sample correlates with our explanatory variables can be found in Schmidt and Witte (1987, chapter 2).

2. We placed approximately one third of the sample of complete observations into our analysis sample because this yielded a sample size large enough to yield precise results, but not so large as to exhaust our computer budget prematurely.

3. Note that our data contain information only on return to prison in North Carolina. While this is the variable of interest to the North Carolina Department of Correction, it is certainly not ideal. In particular, some of our releasees certainly must have returned to prison elsewhere than in North Carolina. A similar problem is that some of the releasees will have died during the followup period. Variables which correlate positively with geographical mobility or with mortality will have a spurious correlation with our dependent variable. However, given the nature of our data, there is nothing we can do about this. There is some prior evidence suggesting that relatively few individuals should be expected to return to

prison outside of North Carolina, but that the death rate is considerably higher than in the general population; see Witte (1975).

4. Some convictions are for crimes not classified as crimes against a person or crimes against property, so that **PROPTY** and **PERSON** do contain independent information.

5. An interesting question is whether the non-monotonicity of the density of time until reimprisonment is due solely to procedural delays between the commission of a criminal offense and return to prison. Unfortunately, our data do not allow us to answer this question.

6. For a sufficiently high degree of polynomial, the LaGuerre distribution can approximate <u>any</u> survival time distribution arbitrarily well. See Schmidt and Mann (1977) and Lutkepohl (1980).

7. The change in the likelihood value of 50 would generate a likelihood ratio test statistic of 100, for a test of the restriction (δ = 1) that reduces the split lognormal model to the simple lognormal model of section 3. While the likelihood ratio test statistic does not have its usual chi-squared distribution here (the restriction is on the boundary of the parameter space), there can be little doubt that this restriction is soundly rejected by the data.

8. The logit lognormal and logit/individual lognormal models generate rather different probabilities of recidivism than the proportional hazards model, but the <u>rankings</u> of individuals are very similar for all three models. If we fix the proportion of the sample for whom we will predict recidivism, only the rankings are relevant.

Figure 1

Predicted versus actual recidivism

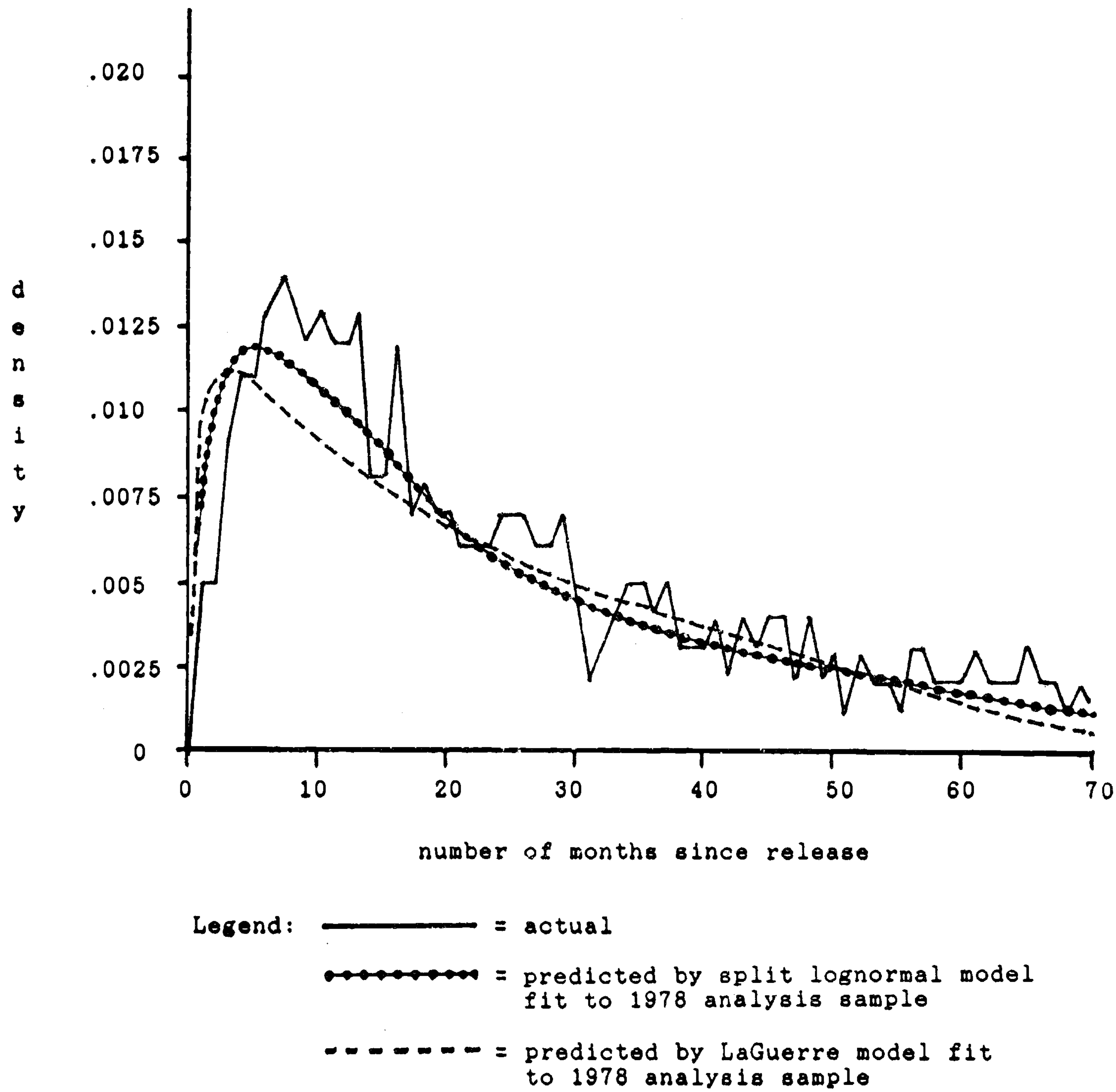

TABLE 1

PROPORTIONAL HAZARDS MODEL

VARIABLE	Final Specification COEFFICIENT	t RATIO	Original Specification COEFFICIENT	t RATIO
TSERVD/100	1.3712	8.15	1.1620	5.92
AGE/1000	-3.4969	-7.09	-3.3445	-6.43
PRIORS/10	.89883	6.75	.83602	6.09
WHITE	-.44041	-5.07	-.44475	-5.07
FELON	-.57342	-4.10	-.57866	-3.54
ALCHY	.41250	3.98	.42850	4.11
JUNKY	.31512	3.28	.28204	2.91
PROPTY	.40483	3.02	.39012	2.47
MALE	.70252	2.92	.67569	2.78
RULE/100			3.0861	1.83
MARRIED			-.15290	-1.42
SCHOOL/10			-.25082	-1.29
WORKREL			.086048	.96
PERSON			.075544	.31
SUPER			-.0087688	-.09
ln L	-3970.7		-3967.0	

TABLE 2

VARIABLE	SPLIT LOGNORMAL MODEL COEFFICIENT	t RATIO	LOGIT LOGNORMAL MODEL COEFFICIENT	t RATIO
TSERVD/100	-1.9750	-5.96	-2.8713	-5.03
AGE/1000	3.5721	7.48	4.3424	6.74
PRIORS/10	-1.4551	-6.66	-1.9857	-5.26
WHITE	.48400	4.10	.66509	4.79
FELON	.94958	4.73	1.0043	4.01
ALCHY	-.61275	-4.21	-.63419	-3.60
JUNKY	-.31317	-2.18	-.44104	-2.73
PROPTY	-.66631	-3.55	-.55841	-2.35
MALE	-.79656	-3.14	-.88252	-2.96
CNST	4.0828	13.33	.067918	.19

Split lognormal model: $\delta = .70852$, $\sigma = 1.4901$, ln L = -3265.1

Logit lognormal model: $\mu = 3.2159$, $\sigma = 1.2001$, ln L = -3256.5

TABLE 3

LOGIT/INDIVIDUAL LOGNORMAL MODEL

VARIABLE	Equation for P(Never Fail) COEFFICIENT	t RATIO	Equation for Duration, Given Eventual Failure COEFFICIENT	t RATIO
TSERVD/100	-1.5841	-4.63	-1.1532	-3.89
AGE/1000	3.7653	5.12	1.2498	2.08
PRIORS/10	-1.1543	-4.71	-.66517	-4.01
WHITE	.64818	4.34	.017090	.13
FELON	.46363	1.71	.68911	3.06
ALCHY	-.44280	-2.49	-.31282	-2.07
JUNKY	-.45455	-2.60	.004483	.03
PROPTY	-.20422	-.79	-.56749	-2.85
MALE	-.88117	-2.27	-.097835	-.17
CNST	-.001198	-.00	3.2381	5.43

σ = 1.1212
ln L = -3239.3

TABLE 4

LOGIT/INDIVIDUAL EXPONENTIAL MODEL

VARIABLE	Equation for P(Never Fail) COEFFICIENT	t RATIO	Equation for Duration, Given Eventual Failure COEFFICIENT	t RATIO
TSERVD/100	-1.3105	-3.78	-1.4441	-4.75
AGE/1000	3.7754	4.34	1.4229	1.56
PRIORS/10	-1.9868	-4.91	.039529	.20
WHITE	.74094	4.67	-.13145	-.82
FELON	.35430	1.30	.79184	3.39
ALCHY	-.46571	-2.55	-.24999	-1.48
JUNKY	-.37281	-2.23	-.11489	-.72
PROPTY	-.11487	-.46	-.67599	-3.71
MALE	-.93485	-2.67	.13731	.25
CNST	.12416	.28	3.3319	5.29

ln L = -3255.4