Testing for Duration Dependence with Discrete Data

by

Scott Grimshaw, James McDonald,
Grant McQueen, and Steven Thorley*

August 2001

Abstract: We use a Monte Carlo study to examine the consequences of using continuous distribution models to test for duration dependence when the data are reported in a discrete format. Our simulations show that this specification error leads to inconsistent parameter estimators and an upward bias in the likelihood ratio test. This bias gives mistaken evidence of duration dependence for sample sizes and parameter values likely to arise in empirical studies. After reporting the sample sizes and parameter values that result in a significant bias, we present and evaluate two approaches that circumvent the problem.

* Scott Grimshaw is an Associate Professor of Statistics at Brigham Young University. James McDonald is Professor of Economics at Brigham Young University. Grant McQueen is the William Edwards Professor of Finance and Steven Thorley is Driggs Fellow and an Associate Professor Finance in the Marriott School at BYU. We are grateful for comments from Dan Sichel. Research assistance from Alan Larsen and Yexiao Xu is greatly appreciated. Send correspondence to: James McDonald, 153 FOB, Brigham Young University, Provo, Utah 84602. Phone: (801) 378-3463 Email: james_mcdonald@byu.edu

I. Introduction

In theory, continuous distribution models are appropriate for economic studies because economic events do not occur at discrete times. For example, strikes, unemployment spells, and expansions do not typically end at midnight on the last day of the month. In practice, however, surveys rarely continuously monitor economic events; rather, they report data in discrete intervals. For example, strike lengths may be reported in full days, unemployment spells in weeks, and economic expansions in months.

One field of research that commonly mixes continuous models with discrete data is the analysis of duration dependence with hazard functions. For example, Kiefer (1988) tests for duration dependence in strikes using continuous distributions even though the strike data are discrete. Solon (1985) finds negative duration dependence in unemployment spells using a continuous model on unemployment data which are rounded up to the nearest week. Sichel (1991) finds evidence of duration dependence in pre-war expansions and post-war contractions using NBER business cycles which are measured in discrete monthly intervals.

Lawless (1982), Hudec (1984), Tuma and Hannan (1984), and Blossfeld, Hamerle, and Mayer (1989) discuss continuous and discrete time hazard models in general terms. We extend their work and specifically examine the consequences of using continuous models to test hypotheses when the data are in fact discrete. First, we document a bias that can arise when discrete data are used with continuous hazard functions in tests for duration dependence. Second, we specify a range of sample sizes and average durations where the bias is significant. Third, we discuss two correction for the bias.[1]

---

[1]This paper examines the parameter values which result in a bias when discrete data are treated as if they were continuous. The paper does not address potentially confounding problems

After this introductory section, Section II reviews some traditional duration models and tests for duration dependence based on continuous distributions. Section III examines the consequences of using continuous distributions with discrete data, a common misspecification. Section IV develops and tests two approaches for testing the degree of duration dependence with discrete data. Section V illustrates the problem and solution using real discrete data from stock returns. Section VI summarizes and concludes the paper.

II. Hazard Functions

The hazard function for a *continuous* variable Y, $\lambda(y)$, represents the rate that spells or runs end at duration y, given that they last until y. The hazard function for a *discrete* variable I, $\lambda(i)$, represents the probability that a spell ends at i, given that it lasts until i.[2] Thus, a hazard function

specification describes data in terms of conditional probabilities in contrast to the probability distribution specification which focuses on unconditional probabilities.

---

of heterogeneity, time-varying explanatory variables, and censoring.

[2]The hazard function for a *continuous* random variable Y is defined as:

$$\lambda(y) = \lim_{\Delta y \to 0} \frac{P(y \leq Y < y+\Delta y \mid Y \geq y)}{\Delta y}, \quad 0 < y < \infty.$$

The hazard function for a *discrete* random variable I is defined as:

$$\lambda(i) = P(I=i \mid I \geq i), \quad I = 1, 2, \ldots, \infty.$$

In order to perform tests of duration dependence, a functional form may be chosen for the density function. A natural and widely used model is the one-parameter (b>0) exponential distribution since it results in a constant hazard function with no duration dependence (i.e. memoryless). In order to allow for either positive or negative monotonic duration dependence, a two-parameter (a>0 and b>0) generalization of the exponential, the Weibull distribution, is often employed. The density and corresponding hazard functions for the two distributions are:

Exponential density function: $$f(y;b) = \frac{e^{-\frac{y}{b}}}{b}, \tag{1}$$

Exponential hazard function: $$\lambda(y;b) = \frac{1}{b}, \tag{2}$$

Weibull density function: $$f(y;a,b) = \frac{a y^{a-1} e^{-\left(\frac{y}{b}\right)^{a}}}{b^{a}}, \tag{3}$$

Weibull hazard function: $$\lambda(y;a,b) = \frac{a y^{a-1}}{b^{a}}. \tag{4}$$

Values of the parameter "a" greater than 1 in equation (4) indicate an increasing hazard function (positive duration dependence) and for values of a less than 1, a decreasing hazard function (negative duration dependence). If a equals one, the Weibull model reduces to the exponential, yielding a constant hazard rate and no duration dependence.

The log-likelihood function can be written in terms of the probability density function, pdf, or the hazard function. Assuming individual spells or runs are independent, the density-based log-likelihood function is:

$$L(\theta) = \sum_{k=1}^{K} \ln f(Y_k;\theta) \tag{5}$$

where Y_k is the k^{th} of K observations of the continuous random variable Y, and θ is a vector of unknown parameters, e.g. $\theta = (b)$ for the exponential and $\theta = (a,b)$ for the Weibull.

Hypothesis testing is typically performed using the likelihood ratio (LR) test, although the Wald and Lagrange multiplier (LM) tests are asymptotically equivalent. This paper will focus on the null hypothesis that the hazard rate is constant (no duration dependence) or, equivalently, that the rate that runs or spells end is independent of the duration of the spell. The LR is given by:

$$LRT = 2[L(a,b) - L(b)] \sim \chi^2_1 \tag{6}$$

where $L(a,b)$ and $L(b)$ denote the maximized log-likelihood functions of the Weibull and exponential distributions. The *LRT* is asymptotically distributed χ^2 with 1 degree of freedom under the null hypothesis that a=1.

III. Problem

The problem of treating discrete data as if it were continuous is illustrated using two Monte Carlo experiments. The first experiment establishes the validity of likelihood ratio tests of duration dependence obtained by fitting continuous distributions to continuous data. The second experiment reveals a bias toward finding duration dependence when *continuous* distributions are fit to *discrete* data. The first Monte Carlo experiment follows four steps.

1. A sample of *S* observations of exponential data is randomly generated with a mean spell or run length of *b*. Sample sizes between 10 and 1,000 and mean spell or run lengths between 2 and 100 are considered.

2. Using Equation (5), MLEs of the exponential and Weibull distributions are obtained for the randomly generated exponential data.

3. The LR test in Equation (6) is evaluated and used to test the null hypothesis of no duration dependence.

4. Steps 1 through 3 are repeated 2,000 times, counting the number of times a type 1 error is made at a rejection level of 0.05.[3] Since the underlying data are generated with an exponential distribution, a type 1 error occurs when the null hypothesis that the data are exponential is falsely rejected in favor of the Weibull distribution.

Table 1 reports the percent of type 1 errors generated by the above Monte Carlo experiment with various mean run lengths and sample sizes. The probability of falsely rejecting the null hypothesis should be equal to the rejection level of the test. For example, with a rejection level of 0.05 (95 percent confidence), the probability of rejecting the null hypothesis when it is true should be 0.05.

For comparison purposes, a base case of b=10 and S=100 is emphasized. In Table 1 (third row and third column), the null hypothesis for this base case is falsely rejected 4.7 percent of the time, close to the theoretic 5 percent.[4] Under the null hypothesis of no duration dependence, the true value of the binomial rejection parameter, p, is 0.05 with a standard error of

[3]Results using rejection levels of 0.01 and 0.10 (available from the authors) are similar to those using a rejection level of 0.05.

[4]Using either the Wald (i.e., an asymptotic t-test since only a=1 is being tested) or the LM test, the null hypothesis is rejected 4.5 percent of the time.

0.005.[5] In other words, the point estimate of 4.7 percent is not significantly different from the theoretic rejection level of 5.0 percent. The validity of the test result is robust with respect to mean run length. For example, with a sample size of 100 (third row in Table 1), the highest and lowest percent of type 1 errors are 5.4 and 4.7. Even in these extreme cases, the percent of type 1 errors are less than 1 standard error from the theoretic value of 5 percent.

Although the null hypothesis is generally rejected the appropriate number of times for sample sizes of 25, 100, and 1,000, the continuous models reject too often when the sample size is only 10. For example, with a sample size of 10 and an average run length of 10, the null hypothesis is rejected 7.3 percent of the time, over four standard deviations from the rejection level of 5 percent. The shaded area in Table 1 highlights the combinations of sample size and mean run length that result in type 1 errors that are three or more standard errors from the rejection level of 5 percent. Since the LR test is *asymptotically* distributed χ^2, the tendency to reject too often with a sample size of *only* 10 is not surprising. This finding suggests that duration dependence studies based on small samples are biased toward finding dependence. For example, Sichel's (1991) finding of positive duration dependence in post World War II contractions may be overstated since the economy has experienced only 9 full post-war contractions.[6]

Having established the validity of using continuous distributions to test continuous data with sample sizes of 25 or more, the validity of using continuous distributions to test discrete

---

[5]The standard error of the binomial parameter p (reject or not reject) under the null hypothesis of no duration dependence is given by $\sigma_p = (.05(1-.05)/2000)^{1/2}$.

[6]Since the focus of this paper is on the consequences of using discrete data, not small sample sizes, the remainder of this paper deals with sample sizes of 25 or more.

data is examined with a second Monte Carlo experiment. The second experiment proceeds along the same lines as the first; the one difference is that in step 1 the randomly generated exponential data are rounded up to an integer creating a new discrete series.[7] Formally, the data are discretized using the transformation $I = T(Y)$, where I is the smallest integer greater than or equal to Y. It can be shown that the transformed exponential data follow a geometric distribution, which is the discrete analogue to the continuous exponential distribution in the sense that it exhibits no duration dependence. In other words, the rounding does not induce duration dependence although it does change the average run length.[8] In steps 2 to 4 of this second experiment, the geometric data is tested for duration dependence as if the data were continuous, paralleling the misspecification common in the literature.

Table 2 reports the results of the second Monte Carlo experiment using the geometric (rounded exponential) data. For the base case of b=10 and S=100, the null hypothesis of no duration dependence, a=1, is rejected 664 out of 2,000 times. Thus, a type 1 error is made 33.2 percent of the time using a rejection level of 5 percent.[9] In other words, when discrete data is treated as if it were continuous, a very strong bias toward finding duration dependence is introduced.

---

[7]Monte Carlo experiments on discrete data are also performed after rounding the exponential data *down* or to the *nearest integer*. These additional tests yield results similar to those reported in the paper which are based on rounding *up*. When the data are rounded down (or to the nearest integer), the observations that fall between 0 and 1 (or 0 and 0.5) are truncated since these runs would not be observable.

[8]For example, rounding up increases the average spell or run length by approximately 0.5.

[9]For the sample size of 100, the Wald and the LM tests make type 1 errors 29.5 and 29.4 percent of the time, respectively.

The shaded area of Table 2 highlights the combinations of sample size and mean run length that result in type 1 errors that are three or more standard errors from the theoretic rejection level of 5.0 percent. The Table 2 results indicate that the magnitude of the bias is sensitive to both the sample size and to the mean run length. As the mean run length increases, the bias toward rejecting the true exponential distribution decreases. For example, with a sample size of 100 (second row), the percent of type 1 errors decreases from 33.2 to 5.8 as the mean run length increases from the base case of 10 to 100. Table 2 also indicates that large sample sizes exacerbate the problem, raising concerns about the consistency of estimators. For a mean run length of 10 (third column), the percent of type 1 errors increases from 33.2 to 99.9 as the sample size increases from 100 to 1,000. The issues of mean run length and sample size are further examined in Tables 3 and 4.

Table 3 gives more detail on three of the cells in Table 2 to examine the relationship between mean run length and the percent of type 1 errors. Specifically, for a sample size of 100, Table 3 reports summary statistics of the Monte Carlo experiments with mean run lengths of 2, 10, and 50. In each of the 2,000 replications of the experiment, maximum likelihood estimates are found for the two Weibull parameters, a and b. $\bar{a}$ and $\bar{b}$ denote the means, and σ_a and σ_b the standard deviations, of the 2,000 MLEs of a and b, respectively. The true value of a is 1 for the underlying exponential data. However, when the data are geometric (rounded exponential) the misspecification of using discrete data with continuous hazard functions results in $\bar{a}$ values of 1.477, 1.136, and 1.047 for mean run lengths of 2, 10, and 50, respectively.[10] The Kolmogorov-

[10] $\bar{a}$ is significantly different from 1 (at the 1 percent rejection level) for all three levels of b in Table 3. With 2,000 observations, the standard deviations of $\bar{a}$, $\sigma_{\bar{a}}$, are 0.0025, 0.0018, and 0.0017 for b=2, 10, and 50, respectively.

Smirnov (KS) statistic confirms the inappropriateness of using discrete data with continuous hazard functions. The 2,000 likelihood ratio test values (for the null hypothesis of $a = 1$) should be asymptotically distributed χ_1^2. KS levels above 0.0364 (rejection level of 0.01) indicate that the actual LR values do not fit the theoretic distribution very well. The null hypothesis, which states that the actual distribution of the LR values is χ_1^2, can be rejected at the 99 percent confidence level for b=2 and b=10, but not for b=50.[11]

In practice, average run lengths that are less than 50 times the unit of measure are common. For example, the strike data in Kennan (1985) and Kiefer (1988) have an average duration of about 35 days. The weekly retrospective data from Waves 14 and 15 of the University of Michigan Panel Study of Income Dynamics used by Katz and Meyer (1990) and Dynarski and Sheffrin (1990) have average unemployment spells of 16.7 and 12.5 weeks, respectively. Solon (1985) uses the Georgia Continuous Wage and Benefit History Program data which has average unemployment spells of 8.4 to 10.8 weeks. The Census Bureau's Current Population Survey used by Butler and McDonald (1986) and others exhibits unemployment spells with an average duration of approximately 25 weeks.

Since in practice many economic series are reported as discrete data and have average durations less than 50, the use of continuous models will lead to biased results. One obvious solution to the problem is to increase the frequency at which events are monitored. For example, if certain unemployment data with an average duration of 25 weeks were gathered to the nearest week, then regathering the same data to the nearest day would result in an average duration of 175 days, thus eliminating the bias caused by using discrete data. The average spell of

[11]See Siegel (1956, p. 251) for KS critical values.

unemployment does not change; however, the run length relative to the unit of measure increases by a factor of 7, shifting the data out of the range of bias highlighted in Table 2. By increasing the frequency at which observations are measured, the data become "less discrete," limiting the degree of misspecification. In practice, however, measuring the data in finer units may not be possible. Less costly and more practical methods for dealing with discrete data are discussed in Section IV.

In Table 4, three of the cells from Table 2 are examined in more detail in order to understand the relationship between the sample size and the percent of type 1 errors. Specifically, for a mean run length of 10, Table 4 reports summary statistics of the Monte Carlo experiments with sample sizes of 25, 100, and 1,000. As the sample size increases, the percent of type 1 errors increases from 12.1 to 33.2 to 99.9 percent. An examination of $\bar{a}$ and σ_a helps explain why. For exponential data, the true value of a is 1. The *absolute* difference between $\bar{a}$ and the reference value for discrete data does decrease with the larger sample size. However, because the standard deviation, σ_a, is decreasing at a faster rate than $\bar{a}$, the *relative* distance between $\bar{a}$ and the reference value of 1 is increasing. In other words, the specification error leads to an inconsistent estimator. This inconsistency is illustrated with a histogram in Figure 1. The 2,000 MLEs for each of the three Monte Carlo experiments (S=25, S=100, and S=1,000) are used to create a histogram by grouping the observations into 70 intervals of width 0.025. The number of observations in each interval is plotted against the interval midpoint and a line is drawn through the midpoints.

IV. Solutions

The previous results clearly illustrate the impact of specification errors which arise when continuous duration models are fit to discrete data. In this section, two solutions to this problem are reviewed. First, the multinomial-based log-likelihood function is introduced as a natural way to obtain MLE while salvaging conventional data collection procedures. Second, duration models based on discrete density functions (e.g., geometric) are shown to be appropriate when the data are inherently discrete, not just a discretized transformation of continuous data.

A. Multinomial Model

The multinomial approach explicitly recognizes that a discretized run or spell *could* have ended anywhere in a given interval with the corresponding probability equal to the integral of the pdf over the interval (i.e., difference between the cumulative distribution function values, cdf's, evaluated at the interval end points).[12] A Monte Carlo experiment verifies that the LR test can be used to test hypotheses about the corresponding hazard functions.

Consider a sample of *K* observations of a random variable *Y* drawn from the population defined by the cdf, $F(y;\theta)$. Let the interval $[0,\infty)$ be partitioned into G intervals of the form $I_i = [y_{i-1}, y_i)$. Let n_i denote the number of observations contained in the i^{th} interval. The probability of an observation being contained in I_i is given by

$$p(i;\theta) = F(y_i;\theta) - F(y_{i-1};\theta) \quad . \qquad (7)$$

[12]The likelihood function for interval data is introduced in Lancaster (1979) and is applied in Butler and McDonald (1986) and Butler, Anderson, and Burkhauser (1989). Similar formulations are used in the analysis of grouped data (see for example, Prentice and Gloeckler (1978) and McDonald (1984)) and with censored data arising in duration models.

The multinomial-based log-likelihood function is

$$L(\theta) = \ln n! - \sum_{i=1}^{G} \ln n_i! + \sum_{i=1}^{G} n_i \ln p(i;\theta) . \tag{8}$$

The first two terms with factorials do not affect parameter estimation and can be ignored in both MLE and LR tests. Since the $p(i;\theta)$ are expressed in terms of the cdf, there is an advantage to selecting cdf's such as the exponential and Weibull, which have simple closed forms. Cox and Hinkley (1974) show that under appropriate regularity conditions the MLE of θ is consistent and has an asymptotic normal distribution, with a variance-covariance matrix given by $\Sigma = [-E(d^2L(\theta)/d\theta^2)]^{-1}$. In addition, they show that the corresponding LR test has an asymptotic χ^2 distribution.

Table 5 reports the results of a Monte Carlo study using the multinomial log-likelihood function for geometric (rounded exponential) data. Table 5 generally exhibits very close agreement with the theoretic 5 percent level. These results confirm that the multinomial-based maximum likelihood estimation provides an approach to circumvent the problem of fitting continuous distributions (and related hazard functions) to discrete data.

B. Discrete Model

An alternative solution to the specification error of using continuous models on discrete data is to develop duration models based on discrete density functions as discussed in Lawless (1982), Tuma and Hannan (1984), and Blossfeld, Hamerle, and Mayer (1989). A disadvantage of the discrete model is the lack of well-known multi-parameter discrete density functions. For example, a commonly used discrete density analogue to the two-parameter Weibull function does not exist. The lack of multi-parameter discrete density functions motivates the development of discrete *hazard* function-based estimation similar to Cox's (1972) logistic model and Kennan's (1985) beta-logit model.

Let $\lambda(i;\theta)$ represent a general discrete hazard function with unknown parameters θ. $\lambda(i;\theta)$ is the conditional probability that an observation or run, I_k, has length i, given that its length is i or greater.[13] Let n_i denote the number of runs that last exactly i periods. Let m_i be the sum of all n_j, $j > i$ (i.e., m_i is number of runs that last longer than i periods). Ignoring factorial terms that do not affect parameter estimation, the hazard-based log-likelihood function is

$$L(\theta) = \sum_i [n_i \ln \lambda(i;\theta) + m_i \ln(1 - \lambda(i;\theta))]. \tag{9}$$

The choice of functional form for $\lambda(i;\theta)$ is guided by the requirement that it conform to the (*0*,1) probability space as well as the desire to test simple hazard function hypotheses (i.e. constant, decreasing, or increasing hazard function). One simple functional form that meets these requirements is the logistic function

$$\lambda(i;\alpha,\beta) = \frac{1}{1 + e^{-(\alpha + \beta i)}}, \tag{10}$$

which maps the unbounded values of $\alpha + \beta\, i$ into the (0,1) probability space.[14] The constant

---

[13] The unconditional probability that an observation I_k has length i is:

$$p(i;\theta) = \lambda(i;\theta) \prod_{j=1}^{i-1} (1 - \lambda(j;\theta))$$

[14] Equation (10) is one of many functional forms that incorporate two parameters, just as the Weibull is one of many two-parameter continuous distributions. Like the Weibull distribution, it is chosen based on its computational simplicity, not because it conforms to any economic theory or model.

hazard function (i.e., geometric density) is a special case of $\lambda(i;\alpha,\beta)$ where $\beta=0$. This is analogous to the description of the exponential distribution as a special case of the Weibull where $a = 1$ for continuous probabilities. Positive (negative) values of β yield monotonically increasing (decreasing) hazard functions. Substituting equation (10) into (9) and maximizing with respect to θ yields maximum likelihood estimates needed in a LR test statistic for the null hypothesis of no duration dependence, $\beta=0$. The MLE for the 1-parameter logistic hazard function has a closed form solution for α, while the 2-parameter function must be estimated numerically.

Table 6 reports the results of a Monte-Carlo study using the discrete model described above on the same data set as Tables 2 and 5. Table 6 exhibits very close agreement between the observed percent of type 1 errors and the theoretic 5 percent rejection level. However, the insignificant deviations from 5 percent in each cell of Table 5 (multinominal estimation) are similar to those in Table 1 (continuous data), while the deviations in Table 6 do not necessarily parallel the deviations in Table 1. Tables 1 and 5 both use the Weibull density function to model non-constant hazard functions. On the other hand, Table 6 is based on the 2-parameter logistic function, which is not simply a discrete version of the Weibull. However, the Monte-Carlo results suggest that the use of discrete probability models or the multinomial model avoids the bias associated with the use of continuous models on discrete data.

V. Illustration

Thus far, this paper has used computer-generated data in Monte Carlo experiments to examine the bias that results from treating discrete data as if it were continuous. This section illustrates the problem and solutions using real data from McQueen and Thorley's (1994) study of stock return runs.

A stylized version of the efficient markets hypothesis holds that a sequence of holding period returns on a risky asset should be serially random. Thus, stock returns should not exhibit duration dependence. Furthermore, the efficient markets hypothesis predicts that stock runs (series of either positive or negative returns) should have very short durations which fall into the range of mean run lengths that have been shown to result in a bias.

A competing hypothesis posits that asset prices contain "bubbles" which grow each period until they "burst," causing the stock market to crash. McQueen and Thorley (1994) argue that bubbles cause runs of positive, but not negative, stock returns to exhibit *negative* duration dependence[15]. The monthly stock data in McQueen and Thorley (1994) is inherently discrete as the runs of positive or negative monthly returns last, say, 2 or 3 months, but not, by construction, 2.5 months. Thus, the Monte Carlo evidence would suggest that if the discrete stock run data were treated as if it was continuous, then *positive* duration dependence could be found and both the efficient market hypothesis and the bubble alternative could be falsely rejected.

Plugging the McQueen and Thorley (1994) data into the Weibull density model, Equation (3), and maximizing the log-likelihood function, Equation (5), yields estimates of a which are significantly greater than 1 (a = 1.380 for runs of positive returns and a = 1.623 for runs of

---

[15]Speculative bubble returns are skewed, characterized by a long run-up in price followed by a crash. For such bubbles to be rational, the bubble must always be positive and explosive; that is, the expected value of the bubble must be increasing over time to compensate the investor for the possibility of a crash. The skewness and explosiveness of bubbles, combined with serially random innovations in fundamental value, result in observed abnormal returns that exhibit duration dependence in runs of positive returns only.

negative returns).[16] Thus, the misspecified *continuous* model rejects the efficient market hypothesis since stock returns appear to have memory. The continuous model also rejects the bubbles hypothesis since the apparent memory leads to positive duration dependence in runs of both positive and negative returns. However, since the average run length of stock returns is only about 2 months, the Monte Carlo experiments cast doubt on these rejections.

When the same data is plugged into the discrete model, Equation (10), and the discrete log-likelihood function, Equation (9), is maximized, the estimates of β are significantly negative ($\beta = -0.109$) for the runs of positive returns but not significantly different from zero ($\beta = -0.022$) for runs of negative returns, consistent with the bubble hypothesis. A similar result is obtained using the multinomial-based log-likelihood function. For example, if the discrete run lengths of 2 months are treated as if they ended somewhere between 1.5 and 2.5 months, then the difference in the cdf's, Equation (7), can be substituted into the log-likelihood function, Equation (8), to find the MLEs. For the stock run data, this process yields estimates of a that are significantly less than 1 ($a = 0.801$) for runs of negative returns but not significantly different from 1 ($a = 0.941$) for runs of positive returns.

Thus, when the model is misspecified and the data are treated as if they were continuous, the bubble hypothesis is falsely rejected. However, when the discrete stock data is appropriately treated as discrete or when runs are treated as if they ended sometime within an interval, the efficient market hypothesis is rejected in favor of the bubbles hypothesis.

---

[16]The duration dependence tests are performed on continuously compounded real monthly returns for an equally-weighted portfolio of all New York Stock Exchange stocks from 1926 to 1991. See McQueen and Thorley (1994) for a full description of the data and tests.

VI. Conclusions

Many economic studies fit continuous probability density functions to continuous data that are reported in a grouped or discrete format. Examples are especially common in the analysis of duration dependence in economic events such as the duration of unemployment spells, strike length, or economic expansions. This paper examines the consequences of using continuous models to test hypotheses about duration dependence when the data are reported in a discrete format.

Monte Carlo simulations are used to examine the validity of the likelihood ratio test for investigating duration dependence. The simulations show that mixing continuous models with either discrete or grouped data may lead to inconsistent parameter estimators and a large upward bias in the size of the likelihood ratio test. This bias can give mistaken evidence of duration dependence for sample sizes and parameter values that often arise in empirical studies.

After reporting the sample sizes and parameter values which lead to a significant bias, this paper presents and evaluates two approaches to circumventing the problem. The first approach is based on maximum likelihood estimation of the continuous model after acknowledging that the discretized runs or spells in the data could have ended anywhere in a given interval. Thus the probability of the event is given by the integral of the pdf over the interval (difference of the cdf's) and the likelihood function is in a multinomial form which is maximized over the underlying distributional parameters. The corresponding estimators will have an asymptotic normal distribution and provide the basis for LR, LM, and Wald tests. The second approach is to parallel the use of continuous distributions and tests with a corresponding development of discrete distributions and tests. Both the multinomial and the discrete distribution approaches avoid the bias associated with the use of continuous models on discrete data.

References

Blossfeld, Hans-Peter, Alfred Hamerle, and Karl Mayer (1989), Event History Analysis, Lawrence Erlbaum Associates, Inc., Hillsdale.

Butler, J.S., Kathryn Anderson, and Richard Burkhauser (1989), "Work and Health After Retirement: A Competing Risks Model with Semiparametric Unobserved Heterogeneity," The Review of Economics and Statistics, 71, 46-52.

Butler, Richard J. and James B. McDonald (1986), "Trends in Unemployment Duration Data," The Review of Economics and Statistics, 68, 545-557.

Cox, D. R. (1972), "Regression Models and Life-Tables," Journal of the Royal Statistical Society, Series B, 34, 187-220.

Cox, D. R. and D. V. Hinkley (1974), Theoretical Statistics, Chapman and Hall, London.

Dynarski, Mark and Steven M. Sheffrin (1990), "The Behavior of Unemployment Durations over the Cycle," The Review of Economics and Statistics, 72, 350-356.

Hudec, Marcus (1984), "Survival Analysis in Heterogeneous Populations--Statistical Models and Concepts," in A. Diekmann and P. Mitler (eds.), Stochastic Modelling of Social Processes, Academic Press, Inc., Orlando.

Katz, Lawrence F. and Bruce D. Meyer (1990), "The Impact of the Potential Duration of Unemployment Benefits on the Duration of Unemployment," Journal of Public Economics, 41, 45-72.

Kennan, John (1985), "The Duration of Contract Strikes in U.S. Manufacturing," Journal of Econometrics, 28, 5-28.

Kiefer, Nicholas M. (1988), "Economic Duration Data and Hazard Functions," Journal of Economic Literature, 26, 646-679.

Lancaster, Tony (1979), "Econometric Methods for the Duration of Unemployment," Econometrica, 47, 939-956.

Lawless, Jerald (1982), Statistical Models and Methods for Lifetime Data, John Wiley & Sons, New York.

McDonald, James B. (1984), "Some Generalized Functions for the Size Distribution of Income," Econometrica, 52, 647-663.

McQueen, Grant R. and Steven Thorley (1994), "Bubbles, Stock Returns, and Duration Dependence," Journal of Financial and Quantitative Analysis, 29, 379-401.

Prentice R. L. and L. A. Gloeckler (1978), "Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data," Biometrics, 34, 57-67.

Sichel, Daniel E. (1991), "Business Cycle Duration Dependence: A Parametric Approach," The Review of Economics and Statistics, 73, 254-260.

Siegel, Sidney (1956), Nonparametric Statistics for the Behavioral Sciences, McGraw-Hill, New York.

Solon, Gary (1985), "Work Incentive Effects of Taxing Unemployment Benefits," Econometrica, 53, 295-306.

Tuma, Nancy and Michael Hannan (1984), Social Dynamics: Models and Methods, Academic Press, Inc., Orlando.

TABLE 1
Percent of Type 1 Errors for Continuous Models at a Rejection Level of 0.05
Using Randomly Generated Exponential (Continuous) Data with Various Means and Sample Sizes*

Weibull hazard function: $\lambda(y) = \frac{a\,y^{a-1}}{b^a}$.

Exponential hazard function: $\lambda(y) = \frac{1}{b}$.

Null hypothesis: $a=1$

Mean spell or run length (b)

Sample Size (# of spells or runs per replication)	2	5	10	25	50	75	100
10	7.3	6.4	7.3	6.7	7.6	6.9	8.3
25	6.4	5.5	5.8	4.9	5.4	4.9	6.4
100	5.4	5.2	4.7	4.9	5.1	5.4	5.3
1,000	5.1	4.3	5.6	5.0	5.3	4.6	5.3

* A type 1 error is rejecting the null hypothesis of the exponential distribution (no duration dependence) in favor of the Weibull distribution (monotonic duration dependence) when the data, in fact, are exponential. Each cell in the table equals 100 times the number of false rejections at the 95% confidence level, divided by the total number of replications, 2,000. Under the null hypothesis of no duration dependence, the asymptotic standard error for each cell in the table is 0.5%. The shaded area highlights the combinations of sample size and mean run length that result in type 1 errors that are three or more standard errors from the rejection level of 5%.

TABLE 2
Percent of Type 1 Errors for Continuous Models at a Rejection Level of 0.05
Using Randomly Generated Geometric (Discrete) Data with Various Means and Sample Sizes*

Weibull hazard function: $\lambda(i) = \dfrac{a\, i^{a-1}}{b^{a}}.$

Exponential hazard function: $\lambda(i) = \dfrac{1}{b}.$

Null hypothesis: a=1

Mean spell or run length (b)

Sample Size (# of spells or runs per replication)	2	5	10	25	50	75	100
25	79.0	25.6	12.1	6.7	6.3	5.2	6.3
100	100.0	77.1	33.2	10.9	6.6	5.9	5.8
1,000	100.0	100.0	99.9	63.3	24.5	13.3	10.8

* A type 1 error is rejecting the null hypothesis of the exponential distribution (no duration dependence) in favor of the Weibull distribution (monotonic duration dependence) when the data, in fact, are exponential. Each cell in the table equals 100 times the number of false rejections at the 95% confidence level, divided by the total number of replications, 2,000. Under the null hypothesis of no duration dependence, the asymptotic standard error for each cell in the table is 0.5%. The shaded area highlights the combinations of sample size and mean run length that result in type 1 errors that are three or more standard errors from the rejection level of 5%.

TABLE 3

Summary Statistics from the Monte Carlo Simulation Fitting the Weibull Hazard Function to Geometric (Rounded Exponential) Data with Various Run Lengths*

Weibull hazard function: $$\lambda(i) = \frac{a i^{a-i}}{b^a}.$$

Statistic	Mean Spell or Run Length (b) $b = 2$	$b = 10$	$b = 50$
$\bar{a}$	1.477	1.136	1.047
σ_a	0.111	0.081	0.078
$\bar{b}$	2.829	10.993	51.231
σ_b	0.213	1.049	5.373
KS	0.989	0.444	0.004
% Type 1	100.0%	33.2%	6.6%

* The Monte Carlo simulation first randomly generates 2000 data sets, each containing 100 observations (*S*=100) of exponential (continuous) data with average run lengths, *b*, of 2, 10, or 50. Second, the continuous data is rounded up to the nearest integer creating geometric (discrete) data. Third, the maximum likelihood estimates of the two Weibull parameters, *a* and *b*, are estimated for each of the 2000 geometric data sets. $\bar{a}$ and $\bar{b}$ are the mean of the 2000 MLE's of *a* and *b*, and σ_a and σ_b are the standard deviations of *a* and *b* respectively. KS is the Kolmogorov-Smirnov statistic testing whether the 2000 likelihood ratio test values (for the null hypothesis of a = 1) are truly distributed χ^2_1. % Type 1 is the number of times the null hypothesis is falsely rejected at the control level of 0.05 times 100 and divided by 2000, the number of replications.

TABLE 4

Summary Statistics from the Monte Carlo Simulation Fitting the Weibull Hazard Function to Geometric (Rounded Exponential) Data with Various Sample Sizes*

Weibull hazard function: $$\lambda(i) = \frac{a i^{a-1}}{b^{a}}.$$

	Sample Size (S)		
Statistic	$S = 25$	$S = 100$	$S = 1{,}000$
$\bar{a}$	1.187	1.136	1.122
σ_a	0.178	0.081	0.025
$\bar{b}$	11.023	10.993	10.992
σ_b	2.096	1.049	0.334
KS	0.177	0.444	0.984
% Type 1	12.1%	33.2%	99.9%

* The Monte Carlo simulation first randomly generates 2000 data sets, each containing 25, 100, or 1000 observations, S, of exponential (continuous) data with an average run length, b, of 10. Second, the continuous data is rounded up to the nearest integer creating geometric (discrete) data. Third, the maximum likelihood estimates of the two Weibull parameters, a and b, are estimated for each of the 2000 geometric data sets. $\bar{a}$ and $\bar{b}$ are the mean of the 2000 MLE's of a and b, and σ_a and σ_b are the standard deviations of a and b, respectively. KS is the Kolmogorov-Smirnov statistic testing whether the 2000 likelihood ratio test values (for the null hypothesis of $a = 1$) are truly distributed χ^2_1. % Type 1 is the number of times the null hypothesis is falsely rejected at the control level of 0.05 times 100 and divided by 2000, the number of replications.

TABLE 5

Percent of Type 1 Errors for Multinomial Estimations at a Rejection Level of 0.05
Using Randomly Generated Geometric (Discrete) Data with Various Means and Sample Sizes*

Weibull hazard function: $\lambda(i) = \dfrac{a\, i^{a-1}}{b^{a}}.$

Exponential hazard function: $\lambda(i) = \dfrac{1}{b}.$

Null hypothesis: a=1

Mean spell or run length (b)

Sample Size (# of spells or runs per replication)	2	5	10	25	50	75	100
25	5.7	5.9	5.9	4.8	5.3	5.1	6.4
100	5.3	5.2	4.9	5.2	5.1	5.9	5.6
1,000	4.9	4.4	5.3	4.8	5.3	4.5	5.3

* A type 1 error is rejecting the null hypothesis of the exponential distribution (no duration dependence) in favor of the Weibull distribution (monotonic duration dependence) when the data, in fact, are exponential. Each cell in the table equals 100 times the number of false rejections at the 95% confidence level, divided by the total number of replications, 2,000. Under the null hypothesis of no duration dependence, the asymptotic standard error for each cell in the table is 0.5%.

TABLE 6
Percent of Type 1 Errors for Discrete Models at a Rejection Level of 0.05
Using Randomly Generated Geometric (Discrete) Data with Various Means and Sample Sizes*

2-parameter logistic hazard function: $\lambda(i) = \dfrac{1}{1 + e^{-(\alpha+\beta i)}}$.

1-parameter logistic hazard function: $\lambda(i) = \dfrac{1}{1+e^{-\alpha}}$.

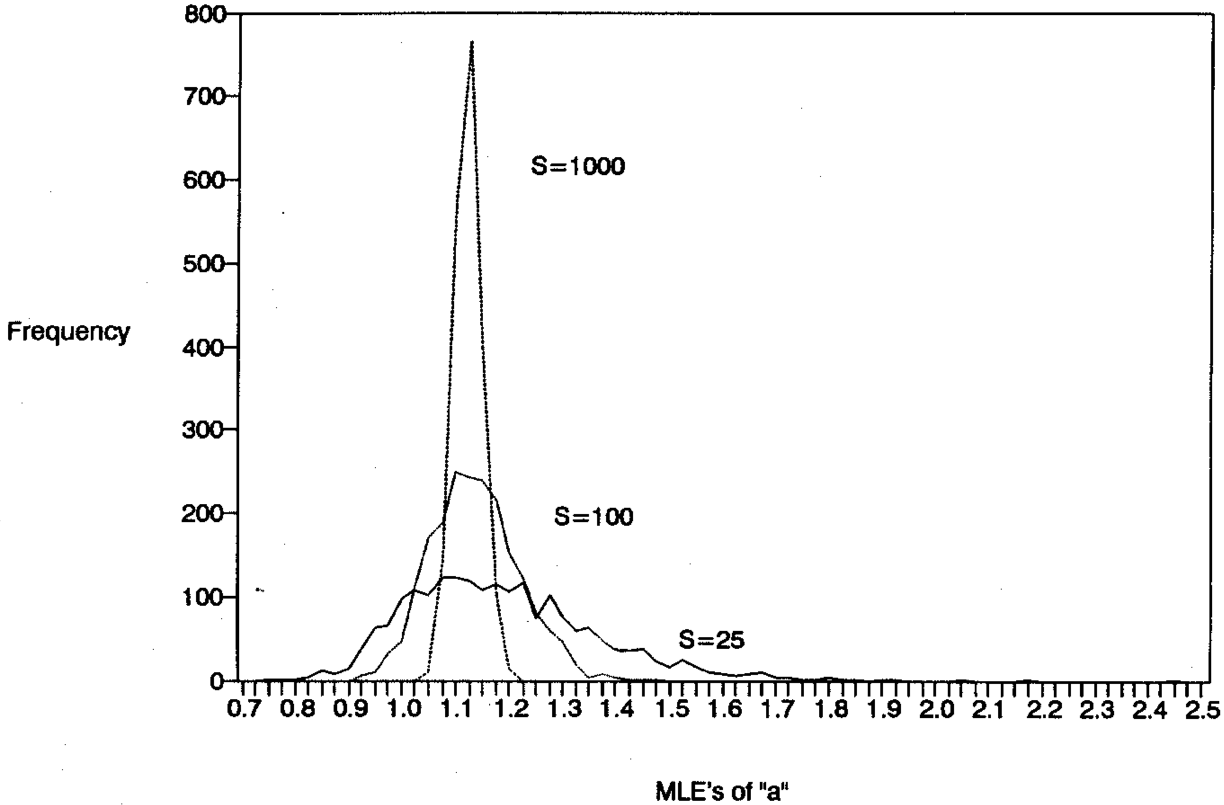
Null hypothesis: $\beta = 0$

Mean spell or run length (b)

Sample Size (# of spells or runs per replication)	2	5	10	25	50	75	100
25	5.3	6.3	5.1	5.6	6.0	5.2	5.9
100	4.7	4.6	4.8	4.9	4.4	5.3	5.6
1,000	5.3	4.9	5.3	4.4	4.4	4.9	4.8

* A type 1 error is rejecting the null hypothesis of the 1-parameter logistic distribution (no duration dependence) in favor of the 2-parameter logistic distribution (monotonic duration dependence) when the data, in fact, are exponential. Each cell in the table equals 100 times the number of false rejections at the 95% confidence level, divided by the total number of replications, 2,000. Under the null hypothesis of no duration dependence, the asymptotic standard error for each cell in the table is 0.5%.

Figure 1
Histogram of 2000 Maximum Likelihood
Estimates of "a" for Three Sample Sizes*



* "a" is a parameter in the Weibull density function. The true value of "a" in the Monte Carlo generated data is 1 indicating no duration dependence. S denotes the sample size of each of the 2000 replications in the simulation. The histogram groups the data into intervals of width 0.025. The number of observations in each interval is plotted against the interval midpoint.