MEMORANDUM

No 23/2000

A Note on the Weibull Distribution and Time Aggregation Bias

By
Knut Røed and Tao Zhang

ISSN: 0801-1117

Department of Economics
University of Oslo

This series is published by the
University of Oslo
Department of Economics

P. O.Box 1095 Blindern
N-0317 OSLO Norway
Telephone: + 47 22855127
Fax: + 47 22855035
Internet: http://www.oekonomi.uio.no/
e-mail: econdep@econ.uio.no

In co-operation with
The Frisch Centre for Economic Research

Gaustadalleén 21
N-0371 OSLO Norway
Telephone: +47 22 95 88 20
Fax: +47 22 95 88 25
Internet: http://www.frisch.uio.no/
e-mail: frisch@frisch.uio.no

List of the last 10 Memoranda:

No 22	By Atle Seierstad: Nonsmooth maximum principle for control problems in Banach state space. 24 p.
No 21	By Diderik Lund: Imperfect loss offset and the after-tax expected rate of return to equity, with an application to rent taxation. 20 p.
No 20	By Christian Brinch: Identification of Structural Duration Dependence and Unobserved Heterogeneity with Time-varying Covariates. 19 p.
No 19	By Knut Røed and Morten Nordberg: Have the Relative Employment Prospects for the Low-Skilled Deteriorated After All? 21 p.
No 18	By Jon Vislie: Environmental Regulation under Asymmetric Information with Type-dependent outside Option. 20 p.
No 17	By Tore Nilssen and Lars Sørgard: Strategic Informative Advertising in a TV-Advertising Duopoly. 21 p.
No 16	By Michael Hoel and Perry Shapiro: Transboundary Environmental Problems with a Mobile Population: Is There a Need for Central Policy? 19 p.
No 15	By Knut Røed and Tao Zhang: Labour Market Transitions and Economic Incentives. 21 p.
No 14	By Dagfinn Rime: Private or Public Information in Foreign Exchange Markets? An Empirical Analysis. 50 p.
No 13	By Erik Hernæs and Steinar Strøm: Family Labour Supply when the Husband is Eligible for Early Retirement. 42 p.

A complete list of this memo-series is available in a PDF® format at:
http://www.oekonomi.uio.no/memo/index.shtml

1 August, 2000

A Note on the Weibull Distribution and Time Aggregation Bias

By Knut Røed and Tao Zhang*

Abstract

The application of continuous time Weibull models on discrete unemployment duration data may produce bias in the estimated shape of the hazard rate. The bias can be substantial even for weekly duration data, and it is seriously aggravated if the Weibull model is erroneously mixed with a Gamma distribution for unobserved heterogeneity.

Keywords: Unemployment duration, Weibull model, Time aggregation bias.

JEL Classification: C41

* The Frisch Centre for Economic Research, Oslo. We wish to thank the Research Council of Norway for financial support and Rolf Aaberge and Steinar Strøm for helpful comments. Correspondence to Knut Røed, The Frisch Centre for Economic Research, Gaustadalléen 21, 0349 Oslo, Norway. E-mail: knut.roed@frisch.uio.no.

Introduction

The Weibull model is a popular tool for econometric analysis of transition data. In particular, it has been extensively used in the analysis of unemployment durations (for recent examples, see e.g. Korpi, 1995; Hernæs and Strøm, 1996; Aaberge, 1996; Røed et al, 1999). While unemployment durations are often measured in weeks or months, the Weibull-distribution is continuous. This discrepancy between statistical model and data generation is often disregarded, however, and the likelihood function is specified *as if* the data were continuous. A common view seems to be that this is a minor offence, without much practical importance. Bergström and Edin (1992) demonstrated, however, that the resulting time aggregation bias could be serious, particularly with respect to duration shape parameters. On the other hand, they also found that the problems were less serious in restrictive parametric duration models. They concluded e.g. that "time aggregation does not seem to have drastic effects on the estimates in a simple parametric model like the Weibull" (Bergström and Edin, 1992, p. 22). In the present note, we show that this conclusion depends heavily on the exact way in which the time aggregation is carried out. This is, however, rarely reported in published work. One reason for this could be that the underlying exact time aggregation mechanism is unknown to the researcher. Another could be that the potential importance of this issue is not fully recognised. In this note we show that disregarded time aggregation can be responsible for substantial bias in the shape parameter of the Weibull distribution and that the bias is seriously aggravated if the Weibull distribution is mixed with a distribution of unobserved heterogeneity, such as a Gamma distribution.

Types of Discrete Unemployment Duration Data

Discrete unemployment duration data are usually gathered from administrative registers or sample surveys. Data based on unemployment registers are created from regular recordings of labour market status. If a person was recorded as unemployed at duration $t-1$, but not at duration t, it is known that the person exited between $t-1$ and t. A common procedure (particularly when the time unit is relatively short) is to record the duration as either $t-1$ or t. Duration data from sample surveys are usually based on the respondents own recollection. A reasonable assumption is that answers are rounded

up or down to the *closest integer* of the time unit used for questioning. A side effect of this method is that the first interval is slightly longer than the other intervals.

Let *T* be the random continuous time duration variable subject to analysis, and let *T** be its discrete observed counterpart. We consider four alternative time aggregation schemes:

A: The discrete observations are obtained by rounding durations *up* to the closest integer, i.e. $T^* \in \{1,2,...,\}$, such that $t \in (t^* - 1, t^*] \rightarrow T^* = t^*$

B: The discrete observations are obtained by rounding durations *down* to the closest integer, i.e. $T^* \in \{0,1,...,\}$, such that $t \in [t^* - 1, t^*) \rightarrow T^* = t^* - 1$

C: The discrete observations are obtained by rounding durations *up or down* to the closest positive integer, i.e. $T^* \in \{1,2,...,\}$, such that $t \in (t^* - \frac{1}{2}, t^* + \frac{1}{2}] \rightarrow T^* = t^*$ for $t > \frac{1}{2}$, otherwise $T^* = 1$.

D: The discrete observations are obtained by adjusting durations to the mid-point of the duration interval in which the durations lie, i.e. $T^* \in \{\frac{1}{2}, \frac{3}{2}, ...,\}$, such that $t \in (t^* - 1, t^*] \rightarrow T^* = t^* - \frac{1}{2}$

While the first three types correspond to data typically obtained from administrative registers (A, B) and sample surveys (C), type D is a simple rule-of-thumb attempt to replace interval borders with interval midpoints, recommended by e.g. Petersen (1995, p. 498). If the durations are uniformly distributed within each interval, these mid-points would correspond to the average true duration within each interval.

Statistical Model and Data Generating Processes

Let *v* be a variable that the researcher uses to represent unobserved population hetero-geneity. In line with the mainstream literature, *v* is assumed to be Gamma distributed with expectation *E(v)=1* and variance $s^2 \geq 0$. Let *T*, conditional on *v*, be distributed according to a Weibull distribution with scale parameter λ and shape parameter α. The conditional hazard rate is then parameterised as

$$q(t \mid v) = \lim_{\Delta t \to 0} \frac{P(t \le T \le t + \Delta t \mid T \ge t)}{\Delta t} = \lambda^{\alpha} \alpha t^{\alpha - 1} v .$$

Throughout this note, we consider a situation in which *the true data* are generated by a pure continuous time Weibull model $(s^2 \equiv 0)$, but where the researcher has only access to the discrete *T**, generated according to the methods described above. We assume that the researcher disregards the discrete data collection pattern, and fits a pure continuous time model, as if the data were really continuous. We investigate how the estimates of α is affected when i) the researcher correctly adopt a pure Weibull model and ii) when the researcher erroneously fit a Gamma mixture model.

We first generated 35 different databases, each with 10,000 observations, randomly drawn from pure Weibull models with different values of the two distribution parameters (α,λ). The parameters were selected in order to make the observations similar to typical unemployment duration data. The various parameter combinations, and the resulting expected durations, are presented in table 1.

Table 1
Expected Durations

	True scale parameter (λ)				
True shape parameter (α)	0.005	0.02	0.05	0.1	0.2
0.5	400	100	40	20	10
0.7	253	63	25	13	6
0.9	210	53	21	11	5
1.0	200	50	20	10	5
1.1	193	48	19	10	5
1.3	185	46	18	9	5
1.5	181	45	18	9	4

Low values of the scale parameter (e.g. λ=0.005), generate observations that are similar to data collected at a very high frequency, for example daily unemployment duration data. High values of the scale parameter (e.g. λ=0.2) generate observations that look more like monthly data, while the intermediate cases may be more typical for weekly unemployment duration data.

Results from Maximum Likelihood Estimation with Discrete Data

Table 2 gives the resulting Maximum Likelihood estimates for the shape parameter α (for ease of exposition, we skip the standard errors, with 10,000 observations they are typically around 0.007).

Table 2
The estimated shape parameter (α) with discrete data, but correct model.
Maximum Likelihood estimates based on 10,000 observations

		True scale parameter (λ)				
True shape parameter (α)	Type of aggregation	0.005	0.02	0.05	0.1	0.2
0.50	A	0.53	0.56	0.59	0.64	0.70
	B	0.43	0.37	0.32	0.28	0.25
	C	0.52	0.55	0.58	0.62	0.67
	D	0.52	0.53	0.55	0.58	0.61
0.70	A	0.72	0.74	0.78	0.83	0.91
	B	0.64	0.56	0.48	0.38	0.31
	C	0.71	0.72	0.76	0.80	0.87
	D	0.71	0.71	0.73	0.74	0.79
0.90	A	0.91	0.95	0.98	1.03	1.11
	B	0.85	0.78	0.68	0.53	0.40
	C	0.90	0.93	0.94	0.98	1.04
	D	0.90	0.92	0.93	0.95	0.96
1.00	A	1.02	1.03	1.07	1.12	1.21
	B	0.98	0.89	0.77	0.61	0.46
	C	1.01	1.01	1.03	1.06	1.14
	D	1.01	1.00	1.01	1.02	1.05
1.10	A	1.12	1.14	1.18	1.23	1.30
	B	1.08	1.02	0.87	0.73	0.50
	C	1.12	1.12	1.14	1.16	1.21
	D	1.11	1.11	1.12	1.13	1.13
1.30	A	1.33	1.32	1.36	1.43	1.52
	B	1.31	1.23	1.09	0.93	0.65
	C	1.32	1.30	1.31	1.35	1.40
	D	1.32	1.29	1.30	1.32	1.32
1.50	A	1.51	1.54	1.58	1.60	1.72
	B	1.50	1.46	1.34	1.10	0.78
	C	1.51	1.52	1.52	1.52	1.52
	D	1.51	1.52	1.51	1.48	1.50

The estimates depend heavily on the exact way in which time is aggregated, even for the series mimicking daily data. All the discretisations produce bias, although time aggregation of type *D* does much better than the others. The bias is of course larger the lower is the frequency of the data (the higher is the scale parameter λ). For typically weekly or monthly data, the bias is substantial.

Now, assume that the researcher does what has become standard practice, and fits a Gamma mixture model to the data, in order to take account of potential unobserved heterogeneity. If that could be done on the original continuous data series, the Maximum Likelihood estimates would correctly reveal that the Gamma variance parameter is approximately zero, and hence that the true model is really a pure Weibull. The estimated parameters of the Weibull model would hardly be affected at all. But, when the available data are discrete, it turns out that a substantial part of duration pattern embedded in the Weibull model is "thrown" over in the Gamma term, hence the bias is substantially aggravated. This is illustrated in Table 3.

Table 3
The estimated shape parameter (α) with discrete data and incorrect Gamma mixture model. Maximum Likelihood estimates based on 10,000 observations

		True scale parameter (λ)				
True shape parameter (α)	Type of aggregation	0.005	0.02	0.05	0.1	0.2
0.50	A	0.57	0.70	0.96	-	-
	B	0.43	0.37	0.32	0.28	0.25
	C	0.56	0.67	0.94	2.19	-
	D	0.54	0.60	0.69	0.86	2.15
0.70	A	0.75	0.82	0.94	1.67	-
	B	0.64	0.56	0.48	0.38	0.31
	C	0.74	0.79	0.89	1.14	1.93
	D	0.72	0.75	0.79	0.85	1.05
0.90	A	0.92	0.99	1.09	1.25	1.67
	B	0.85	0.78	0.68	0.53	0.40
	C	0.91	0.95	1.02	1.15	1.88
	D	0.90	0.93	0.96	0.99	1.12
1.00	A	1.04	1.06	1.17	1.30	1.63
	B	0.98	0.89	0.77	0.61	0.46
	C	1.01	1.02	1.09	1.19	1.59
	D	1.02	1.00	1.04	1.06	1.17

Table 3
The estimated shape parameter (α) with discrete data and incorrect Gamma mixture model. Maximum Likelihood estimates based on 10,000 observations

		True scale parameter (λ)				
True shape parameter (α)	Type of aggregation	0.005	0.02	0.05	0.1	0.2
1.10	A	1.13	1.17	1.24	1.38	1.60
	B	1.08	1.02	0.88	0.73	0.50
	C	1.12	1.12	1.16	1.26	1.52
	D	1.11	1.11	1.12	1.15	1.18
1.30	A	1.34	1.34	1.41	1.57	1.78
	B	1.31	1.23	1.10	0.93	0.65
	C	1.33	1.30	1.32	1.43	1.61
	D	1.32	1.30	1.30	1.35	1.36
1.50	A	1.52	1.57	1.65	1.72	1.92
	B	1.51	1.46	1.34	1.10	0.78
	C	1.51	1.53	1.54	1.55	1.70
	D	1.51	1.52	1.53	1.49	1.50

Note: The symbol – indicates lack of convergence.

The punch line is that any bias produced by time aggregation is seriously aggravated by the imposition of a gamma mixture model. In case of negative duration dependence, even the type *D* adjustment imposes serious distortions.

To illustrate the potential severity of the problem, consider the type of data that may arise from a typical sample survey, questioning persons about how many weeks they have been unemployed. Assume that the respondents round their answers to the closest integer, according to type *C* data. Assume furthermore that the true shape parameter is 0.9 (negative duration dependence), and that $\lambda = 0.05$ (the expected duration is 21 weeks). Figure 1 displays the shape of the true hazard rate, as well as the two estimated hazard rates obtained with continuous time methods based on a pure Weibull model and a Gamma mixture model. The degree of negative duration dependence is seriously under-reported, and in the case of the Gamma-mixture, the negative duration dependence is even interpreted as positive duration dependence.

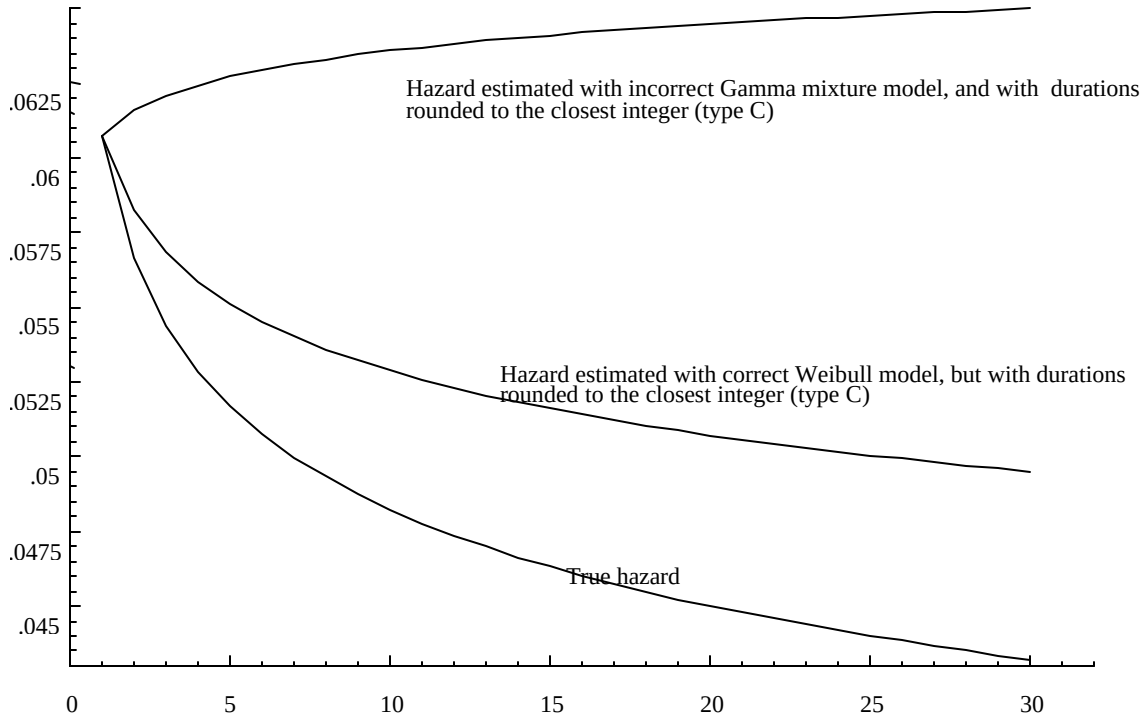

Figure 1 Estimated Weibull hazards when the true shape parameter is equal to 0.9

Conclusion

The application of continuous time Weibull models on discrete duration data may produce substantial bias in the estimated shape of the hazard rate. The bias may be present even for daily unemployment duration data. To the extent that the transformation from continuous to discrete data is known completely, there are simple adjustments that can be made to reduce (or even eliminate) the problem. However, the data generating process may not be known completely. Hence, it is probably much better to follow the route suggested by Meyer (1990) and Narendranathan and Stewart (1993), and derive the likelihood function for the discrete data at hand.

References

Bergström, R. and Edin, P. A. (1992) Time Aggregation and the Distributional shape of Unemployment Duration. *Journal of Applied Econometrics*, Vol. 7, 5-30.

Hernæs, E. and Strøm, S. (1996). 'Heterogeneity and Unemployment Duration.' *Labour*, Vol. 10, No. 2, 269-296.

Korpi (1995) Effects of Manpower Policies on Duration Dependence in Reemployment Rates: The Example of Sweden. *Economica*, Vol. 62, 353-371.

Meyer, B. (1990). 'Unemployment Insurance and Unemployment Spells.' *Econometrica*, Vol. 58, No. 4, 757-782.

Narendranathan, W. and Stewart, M. B. (1993). 'Modelling the Probability of Leaving Unemployment: Competing Risks Models with Flexible Base-line Hazards.' *Applied Statistics*, Vol. 42, No. 1, 63-83.

Petersen, T. (1995) Analysis of Event Histories. Chapter 9 in the Handbook of Statistical Modeling for the Social and Behaviour Sciences, edited by Arminger, Clogg and Sobel, Plenum Press, New York.

Røed, K. Raaum, O. and Goldstein, H. (1999). 'Does Unemployment Cause Unemployment? Micro Evidence from Norway.' *Applied Economics*, Vol. 31, No. 10 (1999).

Aaberge, R. (1996) Unemployment Duration Models with Non-Stationary Inflow and Unobserved Heterogeneity. *Ricerche Economiche*, Vol. 50, 163-172.