Count Models Based on Weibull Interarrival Times

Eric T. Bradlow, Peter S. Fader, Moshe (Topy) Adrian, and Blake McShane [1]

[1]Eric T. Bradlow is The K.P. Chao Professor, Professor of Marketing and Statistics and Academic Director of the Wharton Small Business Development Center, Peter S. Fader is the Frances and Pei-Yuan Chia Professor of Marketing, and Blake McShane is a doctoral student in Statistics at the Wharton School of the University of Pennsylvania. Moshe Adrian is a doctoral student in Mathematics at the University of Maryland. All correspondence on this manuscript should be sent to Eric T. Bradlow, The Wharton School of the University of Pennsylvania, 3730 Walnut Street, Philadelphia, PA 19104, ebradlow@wharton.upenn.edu, (215) 898-8255. The authors thank Rainer Winkelmann for useful suggestions and comments and for generously providing us the data used in this study.

Abstract

The widespread popularity and use of both the Poisson and negative binomial models for count data arises, in part, from their derivation as the number of arrivals in a given time period assuming exponenitally distributed interarrival times (without and with heterogeneity in the underlying base rates respectively). However, with that clean theory comes some limitations including limited flexibility in the assumed underlying arrival rate distribution and the inability to model underdispersed counts (variance less than the mean). While extant research has addressed some of these issues, there still remain numerous valuable extensions.

In this research, we present a model that, due to computational tractability, was previously thought to be infeasible. In particular, we introduce here a generalized model for count data based upon an assumed Weibull interarrival process that nests the Poisson and negative binomial models as special cases. The computational intractability is overcome by deriving the Weibull count model using a polynomial expansion which then allows for closed-form inference (integration term-by-term) when incorporating heterogeneity due to the conjugacy of the expansion and a commonly employed gamma distribution.

In addition, we demonstrate that this new Weibull count model can: (a) sometimes alleviate the need for heterogeneity suggesting that what many think is overdispersion may just be model misfit due to a different and more flexible timing model (Weibull versus exponential), (b) model both over and under dispersed count data, (c) allow covariates to be introduced straightforwardly through the hazard function, and (d) be computed in standard software. In fact, we demonstrate the efficacy of our approach using a data analysis run, including bootstrap standard errors computed via a weighted-likelihood, run in Microsoft Excel.

1 Introduction

The widespread popularity of the Poisson model for count data arises, in part, from its derivation as the number of arrivals in a given time period assuming exponentially distributed interarrival times. But of the thousands of other count models that have been developed over the years (see Wimmer and Altmann (1999) for an excellent synthesis), very few share this straightforward connection between a count model and its timing model equivalent. The connection between a count model and a timing process is more than just a theoretical nicety: in many different contexts, it is useful – if not essential – for a researcher to be able to estimate a model using one form (timing or counting) but apply it using the other. As but one example, marketing managers frequently collect interarrival time data (often in the form of a recency question) but want to make predictions of the number of arrivals (purchases) that a particular customer is likely to make over the next year.

Furthermore, the Poisson count model is truly valid only in the case where the data of interest support the restrictive assumption of *equidispersion*, i.e., where the variance of the data equals the mean. Statisticians have recognized this limitation for many years, and now routinely use models that allow for *overdispersion* (i.e., datasets marked by a fatter, longer right tail than the Poisson will accommodate). A heterogeneous gamma-Poisson model (i.e., the negative binomial or NBD) is generally the first count model invoked for this common situation. But what about datasets with the opposite problem, namely *underdispersion*? Statisticians have acknowledged and addressed this issue in different ways (King, 1989; Cameron and Trivedi 1998), but with the possible exception of a count model featuring gamma-distributed interarrival times proposed by Winkelmann (1995), none of these underdispersed count models (to the best of our knowledge) offers the conceptual elegance and usefulness of the Poisson-exponential connection.

Winkelmann (1995) readily admits the limitations of his gamma-based model. Among other reasons, he comments on the inability to obtain a closed-form hazard function for the gamma, which makes the incorporation of explanatory variables an ad hoc process when compared to the standard Poisson or NBD "regression" models. He points out that "the Weibull distribution is

preferred in duration analysis for its closed-form hazard function ..." but does not pursue such a model. The development and exploration of such a model is the main objective of the present paper.

Before we develop our Weibull count model, we first set the stage by laying out the main properties that the Weibull count model developed here embodies.

(1) The model generalizes (nests) the most commonly used extant models such as the Poisson and the NBD as special cases; thus, when a simple structure is sufficient, the researcher will see it through the estimated model parameters. Furthermore, standard inferential procedures (e.g., the likelihood ratio test) can be used to compare different specifications.

(2) The model handles both overdispersed and underdispersed data, both of which are likely to be seen in practice.

(3) Researchers who believe that the interarrival times of their dataset are Weibull distributed now have a corresponding counting model to use.

(4) The model is computationally feasible to work with. The model is estimable without requiring a formal programming language; it lends itself to implementation within a popular computing environment, such as a spreadsheet.

(5) The model allows for the incorporation of person-level heterogeneity reflecting the fact that individuals' interarrival rates may vary quite substantially across the population.

(6) The mechanism required to incorporate covariate effects is clear and simple. This process is consistent with standard "proportional hazards" methods, which represent the dominant paradigm for ordinary single-event timing models.

In this paper, we derive a new model for count data that satisfies these six criteria in the following ways. First, our count model is based upon an assumed Weibull interarrival process, which nests the exponential as a well-known special case. Second, we demonstrate that the Weibull

count model, via the shape parameter being less than, equal to, or greater than one can capture overdispersed, equidispersed, and underdispersed data respectively. Third, the Weibull interarrival time story is far richer than the exponential story, since it allows for non-constant hazard rates (duration dependence). Fourth, and a significant contribution of this research if it is to impact statistical practice, is the fact that we implement the model entirely in Microsoft Excel. This is accomplised by deriving our model using a polynomial expansion (which can be expressed in closed-form). See Bradlow, Hardie, and Fader (2002), Everson and Bradlow (2002), and Miller, Bradlow and Dayaratna (2006) for similar polynomial expansion solutions for the negative binomial, Beta-Binomial and binary logit models respectively. Fifth, and related to the previous point, once the model is expressed as a closed-form sum of polynomial terms, we can easily introduce a conjugate mixing distribution (the gamma distribution) to capture the underlying disperson in incidence rates across individuals. This ensures that our model (unlike the gamma-based form proposed by Winkelmann (1995)) nests the NBD in addition to the Poisson. Finally, we will demonstrate that we can use the proportional hazards approach to introduce covariates in a very natural manner.

The remainder of this paper is laid out as follows. In the next section, we provide a more detailed description of the major ways in which other researchers have extended basic count models (but rarely with an eye towards maintaining a known interarrival timing process). Section 3 contains the derivation of our Weibull count model, focusing specifically on the polynomial approximation that leads to the closed-form benefits. In Section 4 we re-analyze the same data used by Winkelmann (1995) and provide a set of results comparing a sequence of nested models, the most complicated of which has an underlying Weibull arrival process, heterogeneous baseline rates, and covariates. Through the sequence of models we fit, we are able to ascertain which aspects of the model are most critical. We demonstrate that when the Weibull interarrival process is utilized, the need for underlying heterogeneity is greatly reduced. (Of course our claims are limited to the dataset we analyze but we suspect it will be true more generally). Finally, we provide some concluding remarks and areas for future research in Section 5.

2 Prior Related Research

The primary way in which this research contributes to the literature on count data is by generalizing the underlying interarrival timing model to allow for greater flexibility in its hazard function, which (as described below) is how flexible forms of dispersion are accounted for. For example, Winkelmann (1995) offered a careful analysis of, a valuable framework, and his gamma counting model accounts for the relationship between the nature (i.e., slope) of the timing model hazard function and the type of dispersion seen in the equivalent count data. In particular, if we denote the mean of the interarrival distribution by μ, the variance by σ^2, and the hazard function by

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

where f(t) and F(t) are the density and cumulative probability functions respectively, we say that the distribution has negative duration dependence if $\frac{dh(t)}{dt} < 0$ and positive duration dependence if $\frac{dh(t)}{dt} > 0$. If the hazard function is monotonic, then

$$\frac{dh(t)}{dt} > 0 \Rightarrow \sigma/\mu < 1$$

$$\frac{dh(t)}{dt} = 0 \Rightarrow \sigma/\mu = 1$$

$$\frac{dh(t)}{dt} < 0 \Rightarrow \sigma/\mu > 1.$$

(see Barlow and Proschan 1965, p. 33). These three cases correspond to count data characterized by underdispersion, equidispersion, and overdispersion, respectively.

Focusing on non-constant hazard rates (as above) is but one way in which researchers have extended count models; we discuss some other methods briefly. Another way to capture the same kinds of patterns seen in duration dependent models is to assume that the probability of an event occurring depends on the *number of events* that have occurred previously, as opposed to the *arrival time* of the most recent event (duration dependence). These models are said to display contagion. For instance, they have been studied in the literature on accident proneness (Arbous and Kerrich 1951, Feller 1943). For more information, one can reference Gurland (1995) for a contagious

discrete-time model that leads to the negative binomial in which an occurrence increases and a non-occurrence decreases the probability of a future occurrence. Other models for occurrence dependence have been developed by Mullahy (1986), and Gourieroux and Visser (1997). One can also make the assumption that successive events are independent but the process intensity varies as a function of time. This class of models is known as nonhomogeneous Poisson processes and is described in Lawless (1987). We believe that a promising area for future research would be a comparison of both forms of dependence (duration and occurrence), although here we focus only on the former.

Beyond an explicit focus on any kind of time dependence, there are other many distributions that have been formulated to be able to accommodate underdispersed as well as overdispersed data. Researchers such as Benning and Korolev (2002), Cameron and Trivedi (1998), King (1989), and Shmueli et al (2005) have proposed and discussed a wide variety of generalized count models that can handle overdispersion and underdispersion. But few (if any) offer the benefits or elegance of something like the Poisson-exponential connection. In the next secion we lay out our model that fully respects this connection and also offers a great deal of flexibility in being able to capture a range of count data dispersion patterns.

2.1 A Modeling Framework

Much extant research on count data has been focused on extending the basic Poisson model (denoted here as model [0]) to allow for hyperdispersion via a non-constant hazard rate. The basic ways in which hyperdispersion have been accounted for include: (model [1]) adding covariates to the model, (model [2]) incorporating individual-level heterogeneity for the baseline rates, and (model [3]) both [1] and [2]. In particular, if we let

$$[X_{it}|\lambda_i] \quad \sim \quad \text{Poisson}(\lambda_i \exp(Z'_{it}\beta)), \tag{1}$$

a proportional-hazards framework (Cox, 1972), where X_{it} is a non-negative integer (count) for unit

$i = 1, \ldots, I$ on its $t = 1, \ldots, T_i$-th observation, λ_i is the baseline rate for unit i, $Z_{it} = (Z_{it1}, \ldots, Z_{itP})$ is a vector of covariates that describe each individual, and $\beta' = (\beta_1, \ldots, \beta_P)$ is a vector of covariate slopes: model [0] is obtained by setting $\lambda_i = \lambda$ for all i and $Z'_{it}\beta = 0$ (an intercept only); model [1] is obtained by setting $\lambda_i = \lambda$ for all i (the Poisson Regression Model); model [2] is obtained by setting $P = 1$, $Z_{it}^{\beta} = 0$ and letting $\lambda_i \sim g(\lambda_i|\theta)$ (when g is the gamma distribution then model [2] integrated over the distribution of λ_i is the Negative Binomial Distribution); and model [3] is as given in equation (1) where again $\lambda_i \sim g(\lambda_i|\theta)$. Model [3] is also sometimes referred to as the Neg-Bin II model or a random-intercepts Poisson regression model. Later in Section 4, we compare the results of models [0]-[3] to those derived in this research.

What is of interest to note is that all of these extensions use the Poisson model (with associated exponential interarrival times) as their kernel. That is, these extensions to the model have not been done at the core unit of analysis, i.e., the underlying arrival time distribution, but instead work strictly with the count model from an assumed simple arrival time distribution. What we do in this research is to enhance the flexibility of the arrival time model to account for richer patterns. In particular, instead, we assume that the underlying arrival time distribution for Y_{ik}, the $k-th$ arrival for unit i follows a Weibull with density given by:

$$f(Y_{ik} = y|\lambda_i, \beta, c) \quad = \quad \lambda_i c y^{c-1} \exp(-\lambda_i y^c) \tag{2}$$

Later, when we introduce covariates into the model, we do it through the hazard function:

$$h(t) \quad = \quad \lambda c t^{c-1} \tag{3}$$

which is monotonically increasing for $c > 1$, monotonically decreasing for $c < 1$, and constant (and equal to λ) when $c = 1$.

Using the standard proportional hazards framework, we then boost this "baseline" hazard (given in (3)) by a weighted vector of the covariates $h(t) = h_0(t)\exp(\beta' Z)$, and then rely on the well-known relationship between the hazard function and the CDF

$$F(t) \quad = \quad 1 - \exp(-\int (h(u)du))$$

to arrive at the Weibull regression model

$$f(Y_{ik} = y|\lambda_i, \beta, c) \quad = \quad \lambda_i \exp(Z'_{it}\beta) c y^{c-1} \exp(-\lambda_i \exp(Z'_{it}\beta) y^c) \tag{4}$$

We note that when $c = 1$, equation (4) simplifies to a heterogeneous exponential arrival time model with covariates that leads to count models [0]-[3] above.

Thus, directly analogous to models [0]-[3] which are based on an exponential interarrival time, our interest lies in looking at various reduced-form specifications of model (4). Specifically, we denote as model [4], the Weibull model without heterogeneity and without covariates (model [0] analog) such that $\lambda_i = \lambda$ and $Z'_{it}\beta = 0$. We label model [5] as the Weibull regression model (without heterogeneity) such that $\lambda_i = \lambda$. Model [6] is the model, to be discussed in section 3.2, in which we allow for heterogeneity in baseline rates λ_i but do not include covariates ($Z'_{it}\beta = 0$). Finally, model [7] is the fully parameterized model that includes heterogeneity and covariates. All eight of these models will be fit and results compared in Section 4.

3 Basic Theory and Definitions

Before discussing the Weibull count model itself, we describe the general framework utilized to derive the model that is based upon the relationship between interarrival times and their count model equivalent. Let Y_n be the time from the measurement origin at which the n-th event occurs. Let $X(t)$ denote the number of events that have occurred up until time t. The relationship between interarrival times and the number of events is

$$Y_n \leq t \Leftrightarrow X(t) \geq n.$$

We can restate this relationship by saying that the amount of time at which the n-th event occurred from the time origin is less than or equal to t if and only if the number of events that have occurred by time t is greater than or equal to n.

We therefore have the following relationships that allow us to derive our Weibull count model $C_n(t)$:

$$\begin{aligned} C_n(t) = P(X(t) = n) &= P(X(t) \geq n) - P(X(t) \geq n+1) \\ &= P(Y_n \leq t) - P(Y_{n+1} \leq t). \end{aligned} \tag{5}$$

If we let the cumulative density function (cdf) of Y_n be $F_n(t)$, then $C_n(t) = P(X(t) = n) = F_n(t) - F_{n+1}(t)$. In the case where the measurement time origin (and thus the counting) process coincides with the occurrence of an event, then $F_n(t)$ is simply the n-fold convolution of the common interarrival time distribution which may or may not have a closed-form solution. Based upon (5), we derive our Weibull count model next based upon a polynomial expansion of F(t).

3.1 Weibull Count Model

We derive the basic Weibull count model, model [4] from above, by assuming that the interarrival times are independent and identically distributed Weibull with probability density function (pdf) $f(t) = \lambda c t^{c-1} e^{-\lambda t^c}, (c, \lambda \in R^+)$, and corresponding cdf $F(t) = 1 - e^{-\lambda t^c}$, which simplifies to the exponential model when $c = 1$.

The challenge in deriving the Weibull count model arises in the need to be able to evaluate convolutions of the form $\int_0^t F(t-s)f(s)ds$. While this integral is easily solved for the exponential density, as well as the gamma with an integer-value shape parameter (a.k.a. the Erlang distribution), it does not have a proper solution for the Weibull. Thus, our approach is to handle this intergral (and derive the Weibull count model as a whole) using a Taylor series approximation to the Weibull density.

In particular, the Taylor series approximations obtained by expanding the exponential pieces

$(e^{\lambda t^c})$ respectively, for both the cdf and pdf of the Weibull are:

$$F(t) = \sum_{j=1}^{\infty} \frac{(-1)^{j+1}(\lambda t^c)^j}{\Gamma(j+1)} \tag{6}$$

and

$$f(t) = \sum_{j=1}^{\infty} \frac{(-1)^{j+1} cj \lambda^j t^{cj-1}}{\Gamma(j+1)}. \tag{7}$$

Utilizing, as in (5), that $C_n(t) = F_n(t) - F_{n+1}(t)$, we obtain the following recursive relationship that we utilize in deriving the Weibull count model:

$$\begin{aligned} C_n(t) &= \int_0^t F_{n-1}(t-s) f(s) ds - \int_0^t F_n(t-s) f(s) ds \\ &= \int_0^t C_{n-1}(t-s) f(s) ds. \end{aligned} \tag{8}$$

Before proceeding to develop the general solution to the problem, we note that $F_0(t)$ is 1 for all t and $F_1(t) = F(t)$. Therefore, we have $C_0(t) = F_0(t) - F_1(t) = e^{-\lambda t^c} = \sum_{j=0}^{\infty} \frac{(-1)^j (\lambda t^c)^j}{\Gamma(j+1)}$. Using the recursive formula in (8), we can therefore compute $C_1(t)$:

$$\begin{aligned} C_1(t) &= \int_0^t C_0(t-s) f(s) ds \\ &= \int_0^t (\sum_{j=0}^{\infty} \frac{(-1)^j (\lambda (t-s)^c)^j}{\Gamma(j+1)}) (\sum_{k=1}^{\infty} \frac{(-1)^{k+1} ck \lambda^k s^{ck-1}}{\Gamma(k+1)} ds^{F1} \\ &= \sum_{j=0}^{\infty} \sum_{k=1}^{\infty} \frac{(-1)^j (-1)^{k+1} (\lambda)^j (\lambda)^k}{\Gamma(j+1)\Gamma(k+1)} \int_0^t ck (t-s)^{cj} s^{ck-1} ds \\ &= \sum_{j=0}^{\infty} \sum_{k=1}^{\infty} \frac{(-1)^j (-1)^{k+1} (\lambda)^j (\lambda)^k}{\Gamma(j+1)\Gamma(k+1)} \frac{(t)^{cj} (t)^{ck} \Gamma(cj+1)\Gamma(ck+1)}{\Gamma(cj+ck+1)} \end{aligned} \tag{9}$$

Then, by using a change of variables $m = j$ and $l = m + k$, we obtain:

$$= \sum_{l=1}^{\infty} (\sum_{m=0}^{l-1} \frac{(-1)^m (-1)^{l-m+1} (\lambda)^m (\lambda)^{l-m}}{\Gamma(m+1)\Gamma(l-m+1)} \frac{(t)^{cm} (t)^{cl-cm} \Gamma(cm+1)\Gamma(cl-cm+1)}{\Gamma(cm+cl-cm+1)})$$

$$
\begin{aligned}
&= \sum_{l=1}^{\infty} \frac{(-1)^{l+1}(\lambda t^c)^l}{\Gamma(cl+1)} \left(\sum_{m=0}^{l-1} \frac{\Gamma(cm+1)\Gamma(cl-cm+1)}{\Gamma(m+1)\Gamma(l-m+1)}\right.\\
&= \sum_{l=1}^{\infty} \frac{(-1)^{l+1}(\lambda t^c)^l \alpha_m^l}{\Gamma(cl+1)}
\end{aligned}
$$

where $\alpha_m^l = \sum_{m=0}^{l-1} \frac{\Gamma(cm+1)\Gamma(cl-cm+1)}{\Gamma(m+1)\Gamma(l-m+1)}$.

This suggests a general form for $C_n(t)$, namely: $\sum_{l=n}^{\infty} \frac{(-1)^{l+n}(\lambda t^c)^l \alpha_l^n}{\Gamma(cl+1)}$ which is confirmed by

$$
\begin{aligned}
C_{n+1}(t) &= \int_0^t C_n(t-s) f(s) ds \qquad (10)\\
&= \int_0^t \left(\sum_{j=n}^{\infty} \frac{(-1)^{j+n}(\lambda(t-s)^c)^j \alpha_j^n}{\Gamma(cj+1)}\right)\left(\sum_{k=1}^{\infty} \frac{(-1)^{k+1} ck \lambda^k s^{ck-1}}{\Gamma(k+1)} ds\right.\\
&= \sum_{j=n}^{\infty}\sum_{k=1}^{\infty} \frac{(-1)^{j+n}(-1)^{k+1}(\lambda)^j(\lambda)^k \alpha_j^i}{\Gamma(cj+1)\Gamma(k+1)} \int_0^t ck(t-s)^{cj} s^{ck-1} ds\\
&= \sum_{j=n}^{\infty}\sum_{k=1}^{\infty} \frac{(-1)^{j+n}(-1)^{k+1}(\lambda)^j(\lambda)^k \alpha_j^n}{\Gamma(cj+1)\Gamma(k+1)} \frac{(t)^{cj}(t)^{ck}\Gamma(cj+1)\Gamma(ck+1)}{\Gamma(cj+ck+1)}\\
&= \sum_{l=n+1}^{\infty} \frac{(-1)^{l+n+1}(\lambda t^c)^l}{\Gamma(cl+1)} \left(\sum_{m=n}^{l-1} \alpha_m^n \frac{\Gamma(cl-cm+1)}{\Gamma(l-m+1)}\right)\\
&= \sum_{l=n+1}^{\infty} \frac{(-1)^{l+1}(\lambda t^c)^l \alpha_l^{n+1}}{\Gamma(cl+1)}
\end{aligned}
$$

where $\alpha_l^{n+1} = \sum_{m=n}^{l-1} \alpha_m^n \frac{\Gamma(cl-cm+1)}{\Gamma(l-m+1)}$.

Therefore, we have the main result of this paper, the Weibull count model:

$$
P(N(t)=n) = C_n(t) = \sum_{j=n}^{\infty} \frac{(-1)^{j+n}(\lambda t^c)^j \alpha_j^n}{\Gamma(cj+1)} \quad n = 0,1,2,... \qquad (11)
$$

where $\alpha_j^0 = \frac{\Gamma(cj+1)}{\Gamma(j+1)}$ $j = 0,1,2,...$ and $\alpha_j^{n+1} = \sum_{m=n}^{j-1} \alpha_m^n \frac{\Gamma(cj-cm+1)}{\Gamma(j-m+1)}$, for $n = 0,1,2,...$ for $j = n+1, n+2, n+3, ...$

We note in addition that the expectation of this count model is

$$
E(N) = \sum_{n=1}^{\infty}\sum_{j=n}^{\infty} \frac{n(-1)^{j+n}(\lambda t^c)^j \alpha_j^n}{\Gamma(cj+1)}
$$

with variance given by

$$\begin{aligned} Var(N) &= E(N^2) - (E(N))^2 \\ &= \sum_{n=2}^{\infty}\sum_{j=n}^{\infty} \frac{n^2(-1)^{j+n}(\lambda t^c)^j \alpha_j^n}{\Gamma(cj+1)} - (\sum_{n=1}^{\infty}\sum_{j=n}^{\infty} \frac{n(-1)^{j+i}(\lambda t^c)^j \alpha_j^n}{\Gamma(cj+1)})^2. \end{aligned}$$

3.2 The Benefits of the Weibull Count Model

We now revisit the properties listed in Section 1, point-by-point (and provided in italics below), both to describe those aspects that the basic Weibull count model (without covariates and without heterogeneity) given in (11) provides, and those that require extensions.

(1) *The model generalizes (nests) the most commonly used extant models such as the Poisson and the NBD as special cases; thus, when a simple structure is sufficient, the researcher will clearly see it through the estimated model parameters. Furthermore, standard inferential procedures (e.g., the likelihood ratio test) can be used to compare different specifications.*

We note that when we set $c = 1$ in (11), we do in fact get the Poisson count model as $P(N(t) = n) = \sum_{j=n}^{\infty} \frac{(-1)^{j+n}(\lambda)^j \alpha_j^n}{\Gamma(j+1)}$, a standard result. With regards to the negative binomial model, we discuss this with respect to item [5] below, when λ is allowed to vary across the population.

(2) *The model handles both overdispersed and underdispersed data, both of which are likely to be seen in practice.*

Through extensive simulations (because the result is unavailable in closed-form), we have verified that for $0 < c < 1$, the probability mass function assocaited with the Weibull count model displays overdispersion, whereas for $c > 1$, underdispersion is displayed. That is, the underlying interarrival times have a decreasing (increasing) hazard for $0 < c < 1$ $(c > 1)$. Thus, negative duration dependence is associated with overdispersion, positive duration dependence with underdispersion

(Winkelmann 1995). A lack of duration dependence leads to the Poisson distribution with equal mean and variance.

As one demonstration of these findings, Figures 1 and 2 display probability histograms for the Weibull and Poisson count models with different parameter values. Both the Weibull and the Poisson were intentionally chosen to have identical means (set to 2); yet their dispersion is quite different. In Figure 1, we have the probability histograms for an underdispersed Weibull with parameters $c = 1.5$ and $\lambda = 2.93$, and a Poisson with $\lambda = 2$. The variance of the Weibull count model in this case is 0.880. In Figure 2, we have the probability histograms for an overdispersed Weibull with parameters $c = .5$ and $\lambda = 1.39$, and again the Poisson with $\lambda = 2$. The variance of the Weibull count model in this case is 3.40, which is greater than the mean, as expected.

Insert Figures 1 and 2 here

(3) *Researchers who believe that the interarrival times of their dataset are Weibull distributed now have a corresponding counting model to use.*

As (11) is derived from the Weibull timing model, the link between the timing model and its counting model equivalent is maintained. Hence, in those cases where an analysis of the interarrival times (if the data are available) suggests that a more flexible timing model is needed, it can now be incorporated via its count model equivalent. Furthermore, in those cases where one only has count data, but would like to make forecasts of the next arrival time, this can now be done given the timing and count model link that is now achieved.

(4) *The model is computationally feasible to work with. The model is estimable without requiring a formal programming language; it lends itself to implementation within a popular computing environment, such as a spreadsheet.*

Although the summations shown in the expressions above may seem a bit daunting at first, they are easy to manage from an operational standpoint. We will demonstrate in Section 4 that the model is tractable enough that we perform parameter estimtion, etc., in Microsoft Excel.

(5) *The model allows for the incorporation of person-level heterogeneity reflecting the fact that individuals' interarrival rates may vary quite substantially across the population.*

One nice feature of the model presented in (11) is that introducing heterogeneity across units in their rate parameters, λ_i, is straightforward. If, as is standard in many timing models, we assume that the underlying rates are drawn from a gamma distribution, $\lambda_i \sim \text{gamma}(r, \alpha)$, we can increase the model flexibility at the expense of only one additional model parameter and also, as per item 1, when $c = 1$ nest the negative binomial model. Thus, when we combine our polynomial expansion Weibull count model in (11) with a gamma mixing distribution, we get a count model that nests the Poisson and negative binomial.

In particular, the derivation of the heterogeneous Weibull count model, model [6] from Section 2.1, is given as follows:

$$\begin{aligned} P(N(t)=n) &= \int_0^\infty [\sum_{j=n}^{\infty} \frac{(-1)^{j+n}(\lambda_i t^c)^j \alpha_j^n}{\Gamma(cj+1)}] g(\lambda_i|r,\alpha) d\lambda_i \\ &= \int_0^\infty [\sum_{j=n}^{\infty} \frac{(-1)^{j+n}(\lambda_i t^c)^j \alpha_j^n}{\Gamma(cj+1)}] \frac{\alpha^r (\lambda_i)^{r-1} e^{-\alpha\lambda_i}}{\Gamma(r)} d\lambda_i \\ &= [\sum_{j=n}^{\infty} \frac{(-1)^{j+n}(t^c)^j \alpha_j^n}{\Gamma(cj+1)}] \int_0^\infty \lambda_i^j \frac{\alpha^r (\lambda_i)^{r-1} e^{-\alpha\lambda_i}}{\Gamma(r)} d\lambda_i \\ &= [\sum_{j=n}^{\infty} \frac{(-1)^{j+n}(t^c)^j \alpha_j^n}{\Gamma(cj+1)}] \frac{\Gamma(r+j)}{\Gamma(r)\alpha^j}; \end{aligned} \tag{12}$$

This expression is simply a weighted sum of the j-th moments of the gamma distribution around zero, $\frac{\Gamma(r+j)}{\Gamma(r)\alpha^j}$, as λ_i^j enters the polynomial approximated likelihood in a linear way. Hence, the conjugacy of the gamma mixing distribution, and the polynomial approximated likelihood is directly obtained.

(6) *The mechanism required to incorporate covariate effects is clear and simple. This process is consistent with standard "proportional hazards" methods, which represent the dominant paradigm for ordinary single-event timing models.*

Now that we have the closed-form solution for the heterogeneous count model with an underlying Weibull interarrival process, we extend it to allow for the inclusion of covariates, i.e., models [5] and [7] from Section 2.1. We define the Weibull regression model, without heterogeneity, as

$$\begin{aligned} P(N(t)=n) &= \sum_{j=n}^{\infty} \frac{(-1)^{j+n}(\lambda e^{x_i'\beta} t^c)^j \alpha_j^n}{\Gamma(cj+1)} \\ &= (\sum_{j=n}^{\infty} \frac{(-1)^{j+n}(\lambda t^c)^j \alpha_j^n}{\Gamma(cj+1)})(e^{x_i'\beta})^j \end{aligned} \tag{13}$$

where x_i' denotes the covariate vector for unit i and β a set of covariate slopes. In an analogous manner, we derive model [7], our most complex model which allows for Weibull interrival times, covariate heterogeneity, and parameter heterogeneity and is given by:

$$P(N(t)=n) = \sum_{j=n}^{\infty} \frac{(-1)^{j+n}(t^c)^j \alpha_j^n}{\Gamma(cj+1)} \frac{\Gamma(r+j)}{\Gamma(r)\alpha^j} (e^{x_i'\beta})^j. \tag{14}$$

after integrating over $\lambda_i \sim \text{gamma}(r, \alpha)$.

We next describe an application of these models using a data set initially described and analyzed by Winkelmann (1995) that is an underdispersed count data set with covariates.

4 Testing and Results

Besides the derivation of the Weibull count model, with and without covariates and with and without heterogeneity, an additional goal of this research was to provide an empirical demonstration of our model with two aspects in mind. First, that the polynomial expansion and conjugate prior derived here, which then allows for a closed-form solution has computational advantages that should not be trivialized. Remarkably enough, the computational approach for our class of models, including the computation of bootstrap standard errors (Efron, 1982), was conducted entirely in Microsoft Excel, an aspect we believe makes our approach widely accessible. The spreadsheets that were utilized are available upon request.

Specifically, to compute the standard errors of coefficients under the series of models, we utilized a bootstrap procedure in which 30 replicate data sets for each model were generated by sampling individual respondents from the original data set with replacement. The results reported for the standard errors are the standard deviation of the coefficients across those samples. We note that for our model, the bootstrapping procedure can be implemented by using a weighted likelihood approach where each observation's weight in the likelihood is the number of times that it appears in the replicate sample; a procedure easily implemented within Excel. This equivalence of using a weighted likelihood approach to compute bootstrap standard errors we believe is not specific to this model, can be utilized in a large number of research domains, and hence can be applied in software packages that contain just random number generation and function maximizer (e.g. Microsoft Excel solver) capabilities.

Secondly, one research question we wished to investigate was whether a more flexible (and perhaps more realistic, in many cases) timing model (e.g., Weibull versus exponential) might alleviate the need for heterogeneity – whether brought in through the underlying rates or via covariates. Thus, as we fit a sequence of models with increasing complexity (Poisson, Poisson with covariates, negative binomial, negative binomial with covariates, Weibull, Weibull with covariates, Weibull with gamma heterogeneity, and Weibull with gamma heterogeneity and covariates, as described in Section 2.1), but differing in the source of that complexity, we focus on which aspects of the model are doing the "heavy lifting". Therefore, if in fact we find that a richer underlying kernel timing model can provide an adequate fit when compared to a model that requires heterogeneity, this is important from a scientific perspective. Perhaps researchers' long-standing faith in the validity and robustness of the exponential distribution may be misplaced.

We apply our series of models to a data set initially (and more fully) described by Winkelmann (1995) which contains as a dependent variable the number of children born to a random sample of females. A number of explanatory variables are available including the female's general education (measured as the number of years of school), a series of dummy variables for post-secondary education (either vocational training or university), nationality (German or not), rural or urban dwelling,

religious denomination (Catholic, Protestant, and Muslim, with other or none as reference group), and continuous variables for year of birth and age at marriage.

This data set was chosen for a number of reasons. First, the paper by Winkelmann (1995) acted as a motivation for this research; hence utilizing the identical data set made sense. Secondly, for this data set, the variance of the number of births is less than the mean (2.3 versus 2.4), thus we have an opportunity to demonstrate the ability of the Weibull family of count models to handle underdispersion. And finally, as Winkelmann (1995) already contained the results for the Poisson regression model (model [1] here) and the gamma-based count model which he derived in that paper, we already had results that will let us confirm the accuracy of our computational approach, and will also provide a strong benchmark (the gamma-based model) to which we can compare the Weibull.

Tables 1 and 2 below list the results of the non-regression models (without covariates) and the regression models, respectively. We note that the log-likelihood values computed using our approach, for both the regular Poisson (LL = -2186.8) and Poisson regression (LL = -2101.8) are identical to those in Table 1 (p. 471) of Winkelmann (1995), thus verifying the accuracy of our polynomial expansion approach. In addition, the last column in Table 2, the results of the gamma count model, is taken directly from Table 1 (p. 471) from Winkelmann (1995). We first describe our findings with respect to the models without and then with covariates.

The non-regression models show that the Weibull model has a better log-likelihood than the Poisson (which it must as it nests it) and the NBD. The latter two models are identical for this data set, because the underdispersion will drive the NBD heterogeneity to zero (r and α are extremely large). (The presence of gamma heterogeneity around the Poisson process would overdisperse, not underdisperse, the fertility counts, so it wouldn't help in this case.) Similarly, the log-likelihood of the Weibull model with heterogeneity is effectively equal to that of the simple Weibull model, i.e., heterogeneity is still unnecessary.

Although these results are not especially dramatic, they do provide initial evidence that duration

dependence plays a distinctly different role when compared to heterogeneity. It is valuable to have a model that can distinguish between these two factors. If the underlying data set were instead overdispersed, one could use the heterogeneous Weibull count model to determine whether the "non-Poisson" dispersion effects were coming from the timing process or from cross-sectional differences. We leave this deeper comparison for future research.

Notice finally that the value of c in both Weibull models is 1.116, slightly more than two standard errors above 1. This is consistent with our earlier discussion result that when c is greater than 1, the Weibull count model's variance is less than the mean – underdispersion. It also indicates that the "arrival process" for babies is not completely random. A mother is unlikely to have a baby immediately after the birth of a previous child (which fits the laws of nature quite well), but the odds (or hazard) of delivering another child steadily increases thereafter.

Insert Table 1 Here

Turning our attention to the models with covariates, we first note that the two Weibull regression models provide the best fits, i.e., a slight improvement in log-likelihood compared to the Poisson and Winkelmann's gamma count model. Once again, adding heterogeneity to the Poisson and Weibull models add very little. The values of c for the Weibull regression and heterogeneous Weibull regression models are slightly higher than before, and still significantly greater than 1. The coefficients for the covariates show very small differences across the models. The coefficients of all variables are identical in sign as those in Winkelmann (1995), are extremely stable across the class of models, and have comparable standard errors such that the variables that are significant coincide in both sets of models[1].

Insert Table 2 Here

[1]The year of birth and age of marriage variables were centered in Winkelmann, and not here, hence the difference in size of the coefficients. However, the Poisson regression models as indicated by the log-likelihoods are the same.

5 Conclusions

In this research, we have derived and provided an empirical demonstration for an entirely new class of count models derived from a Weibull interarrival time process. The new model has many nice features such as its closed-form nature, computational simplicity, the ability to nest both the Poisson and NBD models, and the ability to bring in both heterogeneity and covariates in a natural way. The key to the derivation is the use of a Taylor series expansion to get around the fact that, unlike the exponential or gamma distributions, there is no simple way to obtain a convolution of two (or more) Weibulls.

From an empirical standpoint, we showed that the Weibull count model offers a slight improvement in log-likelihood when compared to the gamma count model of Winkelmann (1995) and a dramatic improvement over extant models commonly used. Admittedly the differences, compared to Winkelmann's gamma count model, are small, and it's impossible to generalize from one data set, but these results provide encouraging signs about the model's usefulness and validity. More importantly, the model provides a sizeable improvement over the more traditional Poisson/NBD model (with and without covariates). This may have important implications in many cases, because most researchers have always turned to heterogeneity as the first explanation/correction for data sets that do not conform well to the simple assumption of Poisson counts (and, implicitly, exponential interarrival times). Now researchers have a very plausible second explanation available (i.e., Weibull interarrival times), and unlike Winkelmann's model, they can further explore it using conventional techniques such as proportional hazards for covariates and a parametric mixing distribution for heterogeneity. This is a powerful combination of old and new methods that has substantial promise for a wide variety of application areas.

References

Arbous A.G., Kerrich J.E. (1951), "Accident Statistics and the Concept of Accident-Proneness", *Biometrics*, 341-433.

Barlow, R. E., Proschan F. (1965), "Mathematical Theory of Reliability", *John Wiley and Sons*, New York.

Benning, V. E., and Korolev, V. Y. (2002), *Generalized Poisson Models and Their Applications in Insurance and Finance*, Brill Academic Publishers, Leiden (Netherlands).

Bradlow, E.T., Hardie, B.G.S., Fader, P.S. (2002), "Bayesian Inference for the Negative Binomial Distribution Via Polynomial Expansions", *Journal of Computational and Graphical Statistics*, Vol. 11, No. 1, 189-201.

Cameron, A. C. and Trivedi, P.K. (1998), *Regression Analysis of Count Data*, Cambridge University Press, Cambridge (UK).

Cameron, A. C. and Johansson, P. (1997), "Count Data Regression Using Series Expansion: With Applications", *Journal of Applied Econometrics*, May, Vol 12, No. 3, 203-223.

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

Efron, B. (1982), "Bootstrap Methods : Another Look at the Jackknife, *Annals of Statistics*, Vol. 7, 1-26.

Everson, P.J. and Bradlow, E.T. (2002), "Bayesian Inference for the Beta-Binomial Distribution via Polynomial Expansions", *Journal of Computational and Graphical Statistics*, Vol. 11, No. 1, 202-207.

Feller, W. (1943), "On a General Class of 'Contagious' Distributions", *Annals of Mathematical*

Statistics, Vol. 14, pp. 389-400.

Gourieroux, C., Visser, M. (1997), "A Count Data Model with Unobserved Heterogeneity", *Journal of Econometrics*, Vol. 79, 247-268.

Gurland, J., Sethuraman, J. (1995), "How Pooling Failure Data May Reverse Increasing Failure Rates", *Journal of the American Statistical Association*, Vol. 90 1416-1423.

King, G. (1989), "Variance Specification in Event Count Models: From Restrictive Assumptions to a Generalized Estimator", *American Journal of Political Science*, August, Vol. 33, No. 3, 762-784.

Lawless, J. F. (1987), "Regression Methods for Poisson Process Data", *The Journal of the American Statistical Association*, Vol. 82, 808-815.

Miller, S.J., Bradlow, E.T., and Dayartna, K. (2004), "Closed-Form Bayesian Inferences for the Logit Model via Polynomial Expansions", working paper.

Mullahy, J. (1986), "Specification and Testing of Some Modified Count Data Models", *Journal of Econometrics*, Vol. 33, 341-351.

Shmueli, G., Minka, T.P., Kadane, J.B., Borle, S., and Boatwright, P., (2005), "A useful distribution for fitting discrete data: revival of the ConwayMaxwellPoisson distribution", *Journal of the Royal Statistical Society Series C*, Vol 54, No. 1, 127-142.

Trivedi, P. K., Cameron A. C. (1996), "Applications of Count Data Models to Financial Data", *Chapter 12 in Handbook of Statistics Vol 14: Statistics in Finance*, North Holland, 363-391.

Trivedi, P. K., Deb. P. (1997), "The Demand for Health Care by the Elderly: A Finite Mixture Approach", *Journal of Applied Econometrics*, Vol 12, No. 3, 313-332.

Wimmer, G. and Altmann, G. (1999), *Thesaurus of Univariate Discrete Probability Distributions*,

STAMM VERLAG, Germany.

Winkelmann, R. (1995), "Duration Dependence and Dispersion in Count-Data Models", *Journal of Business and Economic Statistics*, October, Vol 13, No. 4, 467-474.

Winkelmann, R. (2003), "Econometric Analysis of Count Data", Fourth Edition, Heidelberg, New York: Springer".

Table 1: Non-regression model results for total marital fertility.

	Model							
	Poisson		NBD		Weibull		Het. Weibull	
Variable	Coef	SE	Coef	SE	Coef	SE	Coef	SE
λ	2.38	0.046	-	-	2.635	0.099	-	-
c	-	-	-	-	1.116	0.050	1.116	0.051
r	-	-	35183010	684118	-	-	17292.3	85.04
α	-	-	14753795	323713	-	-	6561.7	221.3
Log Likelihood	-2186.8	-	-2186.8	-	-2180.4	-	-2180.3	-

Table 2: Regression model results for total marital fertility.

	Model									
	Poisson		NBD		Weibull		Het. Weibull		Gamma	
Variable	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE
German	-0.200	0.077	-0.200	0.077	-0.222	0.086	-0.222	0.095	-0.190	0.060
Years of Schooling	0.033	0.039	0.033	0.039	0.038	0.045	0.039	0.043	0.032	0.027
Vocational Training	-0.153	0.036	0.153	0.036	-0.173	0.040	-0.174	0.039	-0.144	0.037
University	-0.155	0.158	-0.155	0.158	-0.174	0.181	-0.204	0.177	-0.146	0.130
Catholic	0.218	0.071	0.218	0.071	0.242	0.080	0.249	0.079	0.206	0.059
Protestant	0.113	0.079	0.113	0.079	0.123	0.089	0.128	0.087	0.107	0.063
Muslim	0.548	0.077	0.548	0.077	0.639	0.092	0.651	0.087	0.523	0.070
Rural	0.059	0.046	0.059	0.046	0.068	0.053	0.067	0.052	0.055	0.032
Year of Birth	0.242	0.176	0.242	0.176	0.231	0.200	0.240	0.199	-0.002	0.002
Age at Marriage	-3.044	0.663	-3.044	0.663	-3.403	0.771	-3.370	0.791	-0.290	0.006
λ	3.150	1.020	-	-	4.044	1.590	-	-	-	-
c	-	-	-	-	1.236	0.054	1.254	0.054	-	-
r	-	-	1766.28	147	-	-	17011	354.0	-	-
α	-	-	560.3	190.3	-	-	5023.7	1414.8	1.439	0.233
Log Likelihood	-2101.8	-	-2101.8	-	-2077.0	-	-2076.3	-	-2078.2	-

Figure 1: Poisson and Weibull models displaying underdispersion.

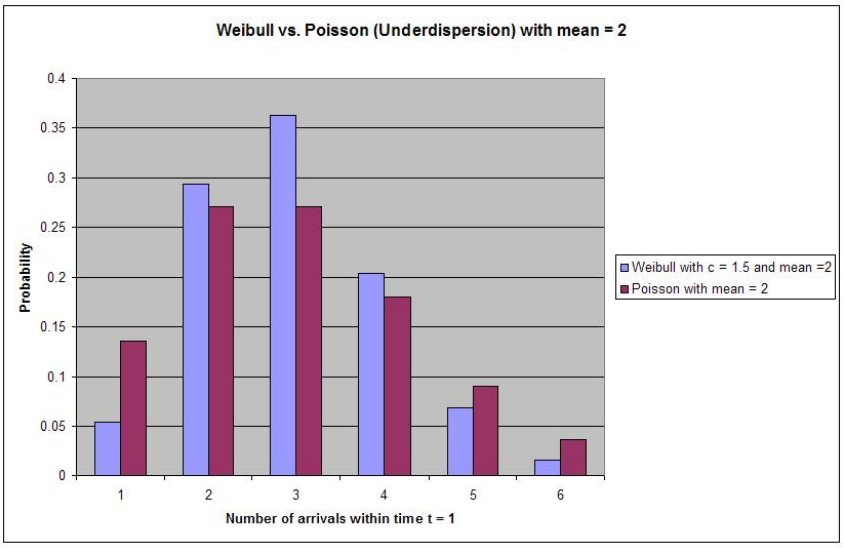

Figure 2: Poisson and Weibull models displaying overdispersion.