Modeling Inflated Count Data

Giles, D.E.[1]

[1]Department of Economics, University of Victoria, Victoria BC, Canada
Email: dgiles@uvic.ca

Keywords: *Count data, over-dispersion, count inflation, Hermite distribution*

EXTENDED ABSTRACT

This paper extends standard regression models for count data in two ways. The first involves allowing for excess numbers of counts (count inflation) simultaneously at several values other than the zero value. The second involves broadening the class of discrete distributions by considering the Hermite distribution (Kemp and Kemp, 1965). This distribution has not been used previously in the econometrics literature, and when it has been used elsewhere, no consideration has been given to introducing covariates into the model. Two particularly appealing features of the Hermite distribution are its abilities to model multi-modal count data without any modification, and to allow for over-dispersion in the sample. We pursue these extensions of the standard count data models both separately and jointly, and provide several empirical applications that illustrate some of their merits. Our results to date suggest that the Hermite distribution, parameterized to incorporate covariates, offers considerable potential in the modeling of discrete economic data.

1. INTRODUCTION

Count data take only non-negative integer values, and they arise in many fields. In economics, examples include the number of applicants for a job, or the number of labour strikes during a year. Such data cannot be modeled adequately by means of standard regression analysis. In addition, in practice at least two characteristics of count data require special attention: there may be an abnormally large number of observations at one or more integer values, so the distribution of the data may be multi-modal; and the data may be "over-dispersed", in the sense that the variance exceeds the mean. These complications can be handled in various ways, but in this paper we revisit these important issues and discuss some new generalizations of the basic Poisson regression model for count data. In particular, we explore the use of a flexible discrete distribution, the Hermite distribution, which has not been used previously for modeling with covariates. Our empirical work with this distribution, some of which is summarized in this paper, suggests that it can provide a powerful way of modeling over-dispersed and multi-model count data.

2. ZERO-INFLATED POISSON REGRESSION

The usual starting point for modeling count data is the Poisson distribution, whose p.m.f. is given as $\Pr.[Y = y] = \exp(-\lambda)\lambda^y / y!$; y = 0, 1, 2, …….. where λ (> 0) is both the mean and variance, so the distribution is described as "equi-dispersed". Many data are "over-dispersed" - their variance exceeds their mean - thus reducing the usefulness of the Poisson distribution. Modeling the variance by a gamma distribution, leads to the familiar Negative Binomial (NegBin II) distribution, which can capture over-dispersion. Covariates are introduced into the model by assigning $\lambda = \exp(x'\beta)$, ensuring that λ > 0. Maximum likelihood estimation is then straightforward, as the likelihood function is strictly concave (as it is for the NegBin II model). The Poisson model, and standard variants that allow for over-dispersion, cannot describe multi-modal data. More correctly, if λ is *integer*, then the Poisson distribution has modes at λ and (λ – 1), but never at non-adjacent values. If λ is non-integer, the single mode occurs at [λ]. The *zero-inflated* Poisson (ZIP) regression model is a modification of this familiar Poisson regression model that allows for an over-abundance of zero counts in the data, which is widely encountered in practice. (Mullahy, 1986; Lambert, 1992.)

The data are assumed to come from two regimes. In R_I the outcome is always a zero count, while in R_{II} the counts follow a Poisson process. Suppose that $\Pr.[y_i \in R_I] = \omega_i$; $\Pr.[y_i \in R_{II}] = (1-\omega_i)$; i = 1, 2, …, n. Then,

$\Pr.[y_i = 0] = \omega_i + (1-\omega_i)\exp(-\lambda_i)$;

$\Pr.[y_i = r] = (1-\omega_i)\exp(-\lambda_i)\lambda_i^r / r!$; r = 1, 2, 3, …..,

As before, covariates enter the model through the conditional mean, λ_i, of the Poisson distribution: $\lambda_i = \exp(x_i'\beta)$, where x_i' is a $(1\times k)$ vector of

the i^{th} observation on the covariates, and β is $(k \times 1)$. Clearly, $E[y_i \mid x_i] = (1-\omega_i)\lambda_i$ and $Var[y_i \mid x_i] = (1-\omega_i)(\lambda_i + \omega_i \lambda_i^2)$, so this model also allows for over-dispersion of the data (if $\omega_i > 0$). This over-dispersion does not arise from heterogeneity, as when the Poisson model is generalized to the Negative Binomial model. It arises from the splitting of the data into the two regimes. In practice, over-dispersion may come from one or both of these sources (Mullahy, 1986; Greene, 2003, p.750). Following Lambert (1992), we can model ω_i using a Logit specification, so $\omega_i = [\exp(z_i'\gamma)]/[1+\exp(z_i'\gamma)]$, where z_i is a $(1 \times p)$ vector of the i^{th} observation on some covariates, and γ is a $(p \times 1)$ vector of additional parameters. The elements of z_i may include elements of x_i, and a Probit specification may be substituted for the Logit specification.

If we have n independent observations in our sample, it is readily seen that the log-likelihood function may be written as

$$\log L(\beta,\gamma) = \sum_{y_i=0} \log[\exp(z_i'\gamma) + \exp(-\exp(x_i'\beta))]$$
$$+ \sum_{y_i \neq 0} [y_i x_i'\beta - \exp(x_i'\beta) - \log(y_i!)]$$
$$- \sum_{i=1}^{n} \log[1+\exp(z_i'\gamma)]$$

The Negative Binomial regression model may be extended to allow for zero-inflation of the data in a corresponding way.

3. MULTINOMIALLY-INFLATED POISSON REGRESSION

Now, suppose that the data exhibit an excess of counts at *several* integer values (which may perhaps include the zero value). Another way of characterizing this situation is that the data have a multi-modal empirical distribution. This problem appears to have received relatively little attention in count data modeling. Santos Silva and Covas (2000) and Hellström (2006) have used modified double-hurdle models to deal with data exhibiting this characteristic. Melkersson and Rooth (1999) have also considered a simple extension of the ZIP model for the case where allowance must be made for count inflation at just the values zero and two.

We propose a full generalization of the ZIP model that allows for count-inflation at a multiplicity of values. Not only does this allow us to deal with multi-modal data, but it also has implications for modeling over-dispersion, without moving (at least initially) from the Poisson basis for the model. Other generalizations then follow. Now the data may be partitioned into $(J + 1)$ regimes, R_j (j = 0, 1, 2,…., J). The first J of these involve *different* degrees of count-inflation, and the last regime corresponds to the non-inflated counts. That is, R_J corresponds to R_{II} in the section 2. Specifically, we replace the (binomial) Logit specification that is used to model the regime probabilities in the ZIP model by a *multinomial* Logit specification. We will call the associated model the Multinomially-Inflated Poisson (MIP) model. It contains the ZIP model and the model of Melkersson and Rooth as special cases. The details are as follows.

The usual multinomial Logit specification (*e.g.*, Greene, 2003, p.721) is:

$$\Pr.[y_i = j] = \exp(z_i'\gamma_j)/[1+\sum_{l=1}^{J}\exp(z_i'\gamma_l)] \quad ; \quad j = 0, 1, 2, \ldots, J$$

where a normalization such as $\gamma_J = 0$ is imposed to take account of the fact that the $(J + 1)$ probabilities must sum to unity. So, let

$$\Pr.[y_i \in R_j] = \omega_{ij} = \exp(z_i'\gamma_j)/[1+\sum_{l=1}^{J}\exp(z_i'\gamma_l)]$$

; $j = 0, 1, 2, \ldots, J$; $\gamma_J = 0$.

Then, as in section 2, the log-likelihood function based on a sample of n independent observations is:

$$\log L(\beta,\gamma_1,\gamma_2,.....,\gamma_J) = \sum_{y_i \in R_0} \log[\omega_{i0} + (1-\sum_{l=0}^{J-1}\omega_{il})P_i]$$
$$+ \sum_{y_i \in R_1} \log[\omega_{i1} + (1-\sum_{l=0}^{J-1}\omega_{il})P_i] + ...$$
$$+ \sum_{y_i \in R_{J-1}} \log[\omega_{iJ-1} + (1-\sum_{l=0}^{J-1}\omega_{il})P_i]$$
$$+ \sum_{y_i \in R_J} \log[(1-\sum_{l=0}^{J-1}\omega_{il})P_i]$$

where

$$P_i = \Pr.[Y_i = y_i] = \exp(-\lambda_i)\lambda_i^{y_i} / y_i!$$

is the Poisson probability, and covariates are introduced by setting $\lambda_i = \exp(x_i'\beta)$. As $\omega_{iJ} = (1-\sum_{l=0}^{J-1}\omega_{il})$, the log-likelihood function can be expressed more compactly as

$$\log L(\beta,\gamma_1,\gamma_2,.....,\gamma_J) = \{\sum_{l=0}^{J-1}\sum_{y_i \in R_l} \log[\omega_{il} + \omega_{iJ}P_i]\}$$
$$+ \sum_{y_i \in R_J} \log(\omega_{iJ}P_i)$$

with P_i and the ω_{il} 's are defined as above.

Now consider the (conditional) mean and variance of the data in this more general MIP model. Let the value of y_i when $y_i \in R_j$ be r_j. Then,

$$\Pr.[Y_i = r_j \mid x_i, z_i] = \omega_{ij} + \omega_{iJ} \exp(-\lambda_i)\lambda_i^{r_j} / r_j! \quad ;$$

$j = 0, 1, 2, \ldots, J\text{-}1$

and

$$\Pr.[Y_i = r_J \mid x_i, z_i] = \omega_{iJ} \exp(-\lambda_i)\lambda_i^{r_J} / r_J!$$

The first and second raw moments for the Poisson distribution are λ_i and $(\lambda_i + \lambda_i^2)$, so

$$E[Y_i \mid x_i, z_i] = \omega_{iJ}\lambda_i + \sum_{j=0}^{J-1} r_j \omega_{ij}$$

$$Var[Y_i \mid x_i, z_i] = \omega_{iJ}(\lambda_i + \lambda_i^2) + \sum_{j=0}^{J-1} r_j^2 \omega_{ij} - [\omega_{iJ}\lambda_i + \sum_{j=0}^{J-1} r_j \omega_{ij}]^2 \quad ,$$

and the ω_{ij}'s are functions of the z_i's. These two expressions collapse to their counterparts in the section 2 in the binary Logit case, and to equations (9) and (10) in Melkersson and Rooth (1999) when the only two inflated values are zero and two. The mean and variance expressions indicate that the model may account for either over-dispersion or under-dispersion in the data. As Melkersson and Rooth (1999, p. 195) note, even in the case where $J = 2$, the degree of departure from equi-dispersion is a complicated function of the parameters.

4. MORE GENERAL MODELS

4.1 The Hermite Distribution

Let us now consider a distributional basis for modeling count data that allows for both multi-modality and departures from equi-dispersion, and that has, to our knowledge, not been exploited previously in the econometrics literature. This is the Hermite distribution (Kemp and Kemp, 1965), which is a form of generalized Poisson distribution and is so-named because the expressions for its probabilities and factorial moments can be expressed in terms of the coefficients of (modified) Hermite polynomials.

Kemp and Kemp show that this distribution arises naturally in several ways. For instance, the bivariate Poisson and the Poisson-Binomial distributions are both special cases of the Hermite distribution. The sum of an ordinary Poisson variate and an independent Poisson 'doublet' variate is also Hermite-distributed; and the sum of two *correlated* Poisson variates has a distribution (first established by McKendrick, 1926) that is also of the Hermite form. This also applies to the distribution of a sum of a finite number of correlated Poisson variates, derived by Maritz (1952). In the current context, this latter interpretation of the Hermite distribution is appealing, as it allows for a situation where several separate but correlated Poisson processes (with different means) may be generating the data, and this may be reflected in count-inflation and multi-modality at different values. As we will see below, the Hermite distribution exhibits over-dispersion, adding to its appeal here.

The p.m.f. for the Hermite distribution can be expressed in several ways, but the following representation is convenient:

$$\Pr.[Y_i = r] = \exp\{-(a_1 + a_2)\} \sum_{l=0}^{[r/2]} \frac{a_1^{r-2l} a_2^l}{(r-2l)!\,l!} \quad ; \quad r = 0, 1, 2, \ldots\ldots$$

where [.] denotes the integer part of the argument, and a_1 and a_2 are the (non-negative) parameters of the distribution. The mean and variance of this distribution are (a_1+2a_2) and (a_1+4a_2) respectively. Unless $a_2 = 0$ (the Poisson case), we have over-dispersion.

Finally, as is the case for all generalized Poisson distributions, the probabilities follow a simple recursion scheme. For the Hermite distribution if $p_r = \Pr.[Y_i = r]$, then:

$$p_{r+1} = (a_1 p_r + 2a_2 p_{r-1})/(r+1); \quad r = 2, 3, \ldots\ldots$$

$$p_1 = (a_1 p_0)$$

$$p_0 = \exp\{-(a_1 + a_2)\}$$

We are not aware of any previous discussion of introducing covariates into models based on the Hermite distribution. The mean involves *both* of the parameters a_1 and a_2, so matters are less straightforward than in the Poisson case. One possibility is to introduce the covariates by assigning $a_{1i} = \exp(x_{1i}'\beta_1)$, and retaining a_2 (> 0) as a parameter to be estimated. In this case $E[y_i \mid x_{1i}] = \exp(x_{1i}'\beta_1) + 2a_2$;

$Var[y_i \mid x_{1i}] = \exp(x_{1i}'\beta_1) + 4a_2$.

Alternatively, we can assign $a_{1i} = \exp(x_{1i}'\beta_1)$ and $a_{2i} = \exp(x_{2i}'\beta_2)$, so the conditional mean is $E[y_i \mid x_{1i}] = \exp(x_{1i}'\beta_1) + 2\exp(x_{2i}'\beta_2) \quad > 0$.

Expressing both the mean and variance as positive functions of the covariates would not guarantee the positivity of a_{1i} and a_{2i} at all sample points, in general. However, in the case of just two inflated counts, the fact that a Hermite variate can be interpreted as the sum of two correlated Poisson variates can be exploited. Then from Kemp and Kemp (1965, p.390, eq. 58), the covariates could be introduced directly

through the means of the underlying variates by setting $a_{1i} = \exp(x_{1i}'\beta_1) + \exp(x_{2i}'\beta_2) - 2\sigma_{12}$ and $a_{2i} = \sigma_{12}$, where σ_{12} denotes the covariance between the two implicitly underlying variates, and is a parameter to be estimated. Using any of these approaches, the Hermite distribution can accommodate covariates in a very flexible way, and can simultaneously model over-dispersed and multi-modal (multinomially-inflated) data.

4.2 An Alternative Parameterization

Estimation of the Hermite distribution by MLE can be challenging, especially in the presence of covariates. Typically, the ikelihood function has several local maxima. Sometimes it is computationally convenient to re-parameterize the Hermite distribution in terms of $\theta_1 = (2a_2)^{1/2}$ and $\theta_2 = a_1 (2a_2)^{-1/2}$. Then,

$p_{r+1} = (\theta_1\theta_2 p_r + \theta_1^2 p_{r-1})/(r+1)$; $r = 2,3, \ldots\ldots$

$p_1 = (\theta_1\theta_2 p_0)$

$p_0 = \exp\{-(\theta_1\theta_2 + \theta_1^2/2)\}$

Then the mean and variance are $\theta_1(\theta_1 + \theta_2)$ and $\theta_1(2\theta_1 + \theta_2)$. Again, covariates can be introduced in various ways, the simplest by assigning $\theta_{1i} = \exp(x_{1i}'\beta_1)$, and retaining θ_2 (> 0) as a parameter to be estimated. Then,

$E[y_i \mid x_{1i}] = \exp(x_{1i}'\beta_1)[\exp(x_{1i}'\beta_1) + \theta_2]$ and

$Var[y_i \mid x_{1i}] = \exp(x_{1i}'\beta_1)[2\exp(x_{1i}'\beta_1) + \theta_2]$.

This alternative parameterization of the model is used in sections 5.2 and 5.3 below.

The MIP model of section 3 can be generalized further by replacing the Poisson distribution with the Negative Binomial or Hermite distributions. Count-inflation could then be modeled in an even more flexible manner, but recall that the Hermite distribution is already capable of capturing count-inflation in its own right. Using the primary approach to incorporating covariates suggested in section 4.1, if we generalize the MIP model by using Hermite probabilities, we have what we will call the MIH model:

$$\log L(\beta, \gamma_1, \gamma_2, \ldots\ldots, \gamma_J) = \{\sum_{l=0}^{J-1} \sum_{y_i \in R_l} \log[\omega_{il} + \omega_{iJ} P_i]\}$$

$$+ \sum_{y_i \in R_J} \log(\omega_{iJ} P_i)$$

where:

$$\Pr.[y_i \in R_j] = \omega_{ij} = \exp(z_i'\gamma_j)/[1 + \sum_{l=1}^{J} \exp(z_i'\gamma_l)]$$

; $j = 0, 1, 2, \ldots, J$

$$P_i = \Pr.[Y_i = y_i] = \exp\{-(a_{1i} + a_{2i})\} \sum_{l=0}^{[y_i/2]} \frac{a_{1i}^{y_i - 2l} a_{2i}^{l}}{(y_i - 2l)!\,l!}$$

$a_{1i} = \exp(x_{1i}'\beta_i)$.

Let the value of y_i when $y_i \in R_j$ be r_j. Then,

$$\Pr.[Y_i = r_j \mid x_{1i}, z_i] = \omega_{ij} + \omega_{iJ} \exp\{-(a_{1i} + a_2)\}$$

$$\sum_{l=0}^{[r_j/2]} \frac{a_{1i}^{r_j - 2l} a_2^{l}}{(r_j - 2l)!\,l!}; \quad \text{j} = 0, 1, 2, \ldots., \text{J - 1}$$

$$\Pr.[Y_i = r_J \mid x_{1i}, z_i] = \omega_{iJ} \exp\{-(a_{1i} + a_2)\}$$

$$\sum_{l=0}^{[r_J/2]} \frac{a_{1i}^{r_J - 2l} a_2^{l}}{(r_J - 2l)!\,l!}$$

So, as the mean and variance of the Hermite distribution are (a_{1i}+2a_2) and (a_{1i}+4a_2), we have:

$$E[Y_i \mid x_{1i}, z_i] = \omega_{iJ}(a_{1i} + 2a_2) + \sum_{j=0}^{J-1} r_j \omega_{ij}$$

$$Var[Y_i \mid x_{1i}, z_i] = \omega_{iJ}[(a_{1i} + 4a_2) + (a_{1i} + 2a_2)^2]$$

$$+ \sum_{j=0}^{J-1} r_j^2 \omega_{ij} - [\omega_{iJ}(a_{1i} + 2a_2) + \sum_{j=0}^{J-1} r_j \omega_{ij}]^2$$

Again, the model may account for either over-dispersion or under-dispersion in the data, once count-inflation is explicitly built into the model.

5. EMPIRICAL APPLICATIONS

5.1 Leucocyte Data

The (over-dispersed) count data for bacteria in leucoctyes studied by McKendrick (1926) and Kemp and Kemp (1965) are summarized in Figure 1. Table 1 provides the results of estimating a Poisson model, an Hermite model, and a "two-inflated" Poisson (2IP) model to these data by MLE. By construction, the 2IP model must *exactly* account for the number of "2" values in the data, given the absence of covariates. Whether this constraint is imposed or not, the basic Hermite model out-performs the Poisson models in terms of the fit of the counts.

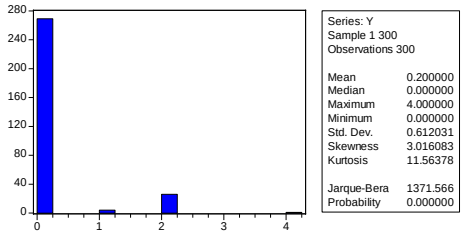

Figure 1. Number of Bacteria in Leucocytes. Source: McKendrick (1926)

Table 1. MLE Results (Bacteria in Leucocytes)

	Poisson	**Hermite**	**2IP**

$log(\lambda)$	-1.6094		
	(0.129)		
a_1		0.0135	
		(0.007)	
a_2		0.0932	
		(0.018)	
γ_0			-2.3601
			(0.206)
β_0			0.0301
			(0.007)
$logL$	-177.77	-116.34	-127.76

Asymptotic standard errors are in parentheses.

Table 2. Actual and Predicted Counts (Bacteria in Leucocytes)

r	Actual	Poisson	Hermite	2IP
0	269	245.7	269.6	266.0
1	4	49.1	3.7	8.0
2	26	4.9	25.2	26.0
3	0	0.3	0.3	0.0
4	1	0.0	1.2	0.0

5.2 **Botswana Fertility Data**

The next example uses data on fertility in Botswana to model the number of living children for 4,361 women in a 1988 survey reported by Drovandi (2006). The data in Figure 2 are over-dispersed, but there is no evidence of count-inflation. To simplify the maximization of the likelihood function we use the alternative parameterization of the Hermite distribution with $\theta_1 = \sqrt{2a_2}$ and $\theta_2 = a_1 / \sqrt{2a_2}$. Then, because $E[y_i \mid x_{1i}] = \exp(x_{1i}'\beta_1)[\exp(x_{1i}'\beta_1) + \theta_2]$, the marginal effect for the k^{th} covariate at observation i is $\partial E[y_i \mid x_{1i}] / \partial x_{1ik} = [\exp(x_{1i}'\beta_1)\beta_{1k}] \times [\exp(x_{1ik}'\beta_1) + \theta_2]$. We need $\theta_2 > 0$ to ensure the positivity of the conditional mean and variance, so the marginal effects have the same signs as the associated coefficients.

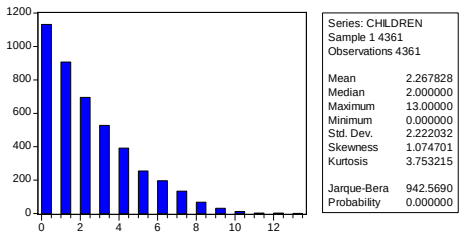

Figure 2. Number of Children. Source: Wooldridge (2002).

Table 3. MLE Results (Fertility Data)

	Poisson	NegBinII	Hermite
$log(\lambda)$	0.8188	0.8188	
	(0.010)	(0.016)	
$log(\eta^2)$		-0.4763	
		(0.042)	
θ_1			1.2147
			(0.015)
θ_2			0.6524
			(0.032)
$logL$	-9598.56	-8704.56	-8806.86

Asymptotic standard errors are in parentheses.

Table 3 reports the basic MLE results with no covariates. The Poisson model is rejected in favour of the NegBinII models on the basis of a likelihood ratio test. However, in Table 4 we see that despite the over-dispersion of the data, the counts for the children predicted by the NegBinII model are inferior to those from the Poisson model, which in turn is clearly dominated by the Hermite model. We focus on the latter model when introducing covariates.

Table 4. Actual and Predicted Counts (Fertility Data)

r	Actual	Poisson	NegBinII	Hermite
0	1132	451.5	2249.5	944.2
1	907	1024.0	1221.0	748.2
2	696	1161.1	537.2	993.0
3	528	877.7	217.9	630.2
4	392	497.6	84.7	491.1
5	255	225.7	32.0	263.8
6	197	85.3	11.9	155.6
7	134	27.6	4.4	73.2
8	68	7.8	1.6	35.9
9	32	2.0	0.6	15.2
10	13	0.4	0.2	6.5
11	3	0.1	0.1	2.5
12	3	0.0	0.0	1.1
13	1	0.0	0.0	0.3

As there is no allowance for count inflation in these specifications, the Poisson models are nested within their Hermite counterparts. Thus, the significance of the estimates of θ_2 lends further support for the Hermite specifications over the Poisson models in each case.

Estimated Poisson and Hermite models with various covariates are reported in Table 5. The covariates are: CATH = 1 if respondent is Catholic, = 0 otherwise; EVERMARR = 1 if respondent has ever been married, = 0 otherwise; EDUC = years of formal education; and URBAN = 1 if respondent lives in an urban centre, = 0 otherwise; and USEBC = 1 if the respondent ever uses birth control, = 0 otherwise. The signs of the

coefficients (and hence the marginal effects) are the same whether a Poisson or Hermite specification is used. When the obvious covariate, "age of the woman" was considered, we were unable to maximize the likelihood function for the Hermite model. In the case of the Poison model this covariate had a positive marginal effect, as expected, but in that model the CATH dummy variable was insignificant.

Table 5. MLE Results With Covariates (Fertility Data)

	Poiss1	**Herm1**	**Poiss2**	**Herm2**
β_0	0.6381	-0.0785	0.4828	-0.2521
	(0.025)	(0.030)	(0.026)	(0.036)
β_1	0.0850	0.0562	0.0646	0.0443
(CATH)	(0.035)	(0.030)	(0.035)	(0.031)
β_2	0.9407	0.6351	0.8677	0.6086
(EVERMARR)	(0.023)	(0.020)	(0.023)	(0.021)
β_3	-0.0617	-0.0406	-0.0754	-0.0514
(EDUC)	(0.003)	(0.002)	(0.003)	(0.003)
β_4	-0.1611	-0.1053	-0.1970	-0.1334
(URBAN)	(0.021)	(0.017)	(0.021)	(0.018)
β_5			0.4676	0.3201
(USEBC)			(0.022)	(0.018)
θ_2		1.0652		1.2566
		(0.071)		(0.092)
logL	-8001.5	-7809.0	-7770.2	-7653.4

Asymptotic standard errors are in parentheses.

5.3 Hot 100 Hits

Our final application relates to data for the 965 "number 1" hits on the Hot 100 chart over the period January 1955 to December 2003. The start of this sample period enables us to capture the rock and roll era, and all seventeen of Elvis Presley's number one hits. The end of the sample avoids the recent impact of downloading digital music on the internet. The data were compiled and analyzed in various ways by Giles (2006, 2007a, 2007b). For a recording that reaches the number one spot, TOP measures the number of weeks that it stays at number one beyond the minimum of one week. We also allow for re-entry into the top spot after having being relegated to a lower position in the chart. The maximum number of "extra" weeks at the top is 15 (Figure 3) and the data are moderately over-dispersed.

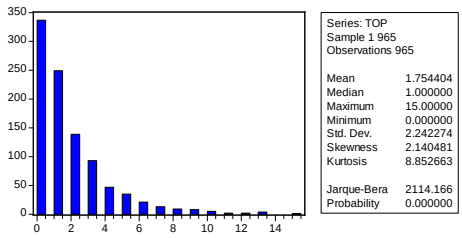

Figure 3: Number of "Extra" Weeks at Top of the Hot 100 Chart. Source: Giles (2006).

Table 6. MLE Results (Hot 100 Data)

	Poisson	**ZIP**	**NegBinII**	**Hermite**
$log(\lambda)$	0.5621		0.5621	
	(0.041)		(0.041)	
$log(\eta^2)$			-0.0222	
			(0.083)	
θ_1				1.0338
				(0.025)
θ_2				0.6633
				(0.063)
γ_0		-0.9007		
		(0.083)		
β_0		2.4672		
		(0.043)		
logL	-2063.7	-1884.5	-1741.4	-1829.1

Asymptotic standard errors are in parentheses.

Table 7. Actual and Predicted Counts (Hot 100 Data)

r	**Actual**	**Poisson**	**ZIP**	**NegBinII**	**Hermite**
0	337	167.0	337.0	616.6	284.9
1	249	292.9	143.6	223.6	195.3
2	139	256.9	177.2	80.2	219.2
3	93	150.3	145.7	28.7	119.7
4	47	65.9	89.9	10.2	79.1
5	35	23.1	44.3	3.6	36.4
6	21	6.8	18.2	1.3	18.2
7	13	1.7	6.4	0.5	7.3
8	9	0.4	2.0	0.2	3.1
9	8	0.1	0.5	0.1	1.1
10	5	0.0	0.1	0.0	0.4
11	2	0.0	0.0	0.0	0.1
12	2	0.0	0.0	0.0	0.0
13	4	0.0	0.0	0.0	0.0
14	0	0.0	0.0	0.0	0.0
15	1	0.0	0.0	0.0	0.0

Table 6 shows the results of estimating some basic models by MLE. The Poisson model is clearly rejected in favour of the NegBinII model using a LRT, and the ZIP and Hermite models out-perform the Poisson and NegBinII models in terms of the predicted counts reported in Table 7. Table 8 shows the results of introducing

covariates into the Poisson, ZIP and Hermite models. The covariates are: ELVIS = 1 if the recording was by Elvis Presley, = 0 otherwise; FEMALE = 1 if the artist was a solo female, = 0 otherwise; INST = 1 if the recording was purely instrumental, = 0 otherwise; and NONCON = 1 if the recording topped the charts in non-consecutive weeks, = 0 otherwise. Other covariates, including a dummy variable for recordings by the Beatles, were found to be insignificant. Importantly, the covariates that are significant in these count data models are exactly those that were significant in the survival models reported by Giles (2007a).

Table 8. MLE Results With Covariates (Hot 100 Data)

	Poisson	**ZIP1**	**ZIP2**	**Hermite**
γ_0		-0.9483	-0.9204	
		(0.086)	(0.086)	
γ_1			-2.0842	
(ELVIS)			(1.377)	
β_0	0.4608	0.8283	0.8315	-0.0377
	(0.050)	(0.021)	(0.021)	(0.027)
β_1	0.9098	0.6043	0.5874	0.5516
(ELVIS)	(0.186)	(0.097)	(0.097)	(0.091)
β_2	0.1809	0.0845	0.0831	0.1127
(FEMALE)	(0.099)	(0.046)	(0.046)	(0.039)
β_3	0.3852	0.3143	0.3123	0.2373
(INST)	(0.216)	(0.085)	(0.085)	(0.073)
β_4	0.6293	0.2874	0.2846	0.3838
(NONCON)	(0.120)	(0.082)	(0.082)	(0.081)
θ_2				0.6836
				(0.066)
logL	-1994.25	-1869.67	-1866.97	-1803.13

Asymptotic standard errors are in parentheses.

6. CONCLUSIONS

We have suggested various ways of extending the traditional models for count data to allow for over-dispersion and for excess counts at multiple values. One of these suggestions is to employ the Hermite distribution as a basis for the modeling. This distribution has the advantage of being able to deal with *both* of the phenomena in which we are interested. We have presented several empirical illustrations that lend credence to the use of this distribution, and for the first time we show how the Hermite distribution can be used to model with covariates. These empirical illustrations lay the groundwork for the use of Hermite regression to model count data in more complete studies. Work in progress includes using the Hermite models with covariates to analyze currency crises and alcohol consumption behaviour, and the development of various specification tests.

7. REFERENCES

Drovandi, S. (2006), Fertility behaviour and its correlates in Botswana, 1988[1], Dipartimento di Statistica, Università degli Studi di Firenze.

Giles, D.E.A. (2006), Superstardom in the U.S. popular music industry revisited, *Economics Letters*, 92, 68-74.

Giles, D.E.A. (2007a), Survival of the hippest: life at the top of the Hot 100, *Applied Economics*, in press.

Giles, D.E.A. (2007b), Increasing returns to information in the U.S. popular music industry, *Applied Economics Letters*, 14, 327-331.

Greene, W.E. (1993), *Econometric Analysis*, 5th ed., Prentice Hall, Upper Saddle River NJ.

Hellström, J. (2006), A bivariate count data model for tourism demand, *Journal of Applied Econometrics*, 21, 213- 226.

Kemp, C.D. and A.W. Kemp (1965), Some properties of the 'Hermite' distribution, *Biometrika*, 52, 381-394.

Lambert, D. (1992), Zero-inflated Poisson regression, with an application to defects in manufacturing, *Technometrics*, 34, 1-14.

McKendrick, A.G. (1926), Applications of mathematics to medical problems, *Proceedings of the Edinburgh Mathematical Society*, 44, 98-130.

Maritz, J.S. (1952), Note on a certain family of discrete distributions, *Biometrika*, 39, 196-198.

Mullahy, J. (1986), Specification and testing of some modified count data models, *Journal of Econometrics*, 33, 341- 365.

Melkersson, M. and D-O. Rooth (2000), Modeling female fertility using inflated count data models, *Journal of Population Economics*, 13, 189-203.

Santos Silva, J.M.C. and F. Covas (2000), A modified hurdle model for completed fertility, *Journal of Population Economics*, 13, 173-188.

Wooldridge, J.M. (2002), Econometric Analysis of Cross Section and Panel Data, MIT Press, Cambridge MA.