Biometrics 65, 1254–1261
December 2009
DOI: 10.1111/j.1541-0420.2009.01191.x

Predictive Model Assessment for Count Data

Claudia Czado,[1] Tilmann Gneiting,[2,*] and Leonhard Held[3]

[1]Zentrum Mathematik, Technische Universität München, Boltzmannstr. 3, D-85748 Garching, Germany
[2]Department of Statistics, University of Washington, Box 354322, Seattle, Washington 98195, U.S.A.
[3]Institut für Sozial- und Präventivmedizin, Abteilung Biostatistik, Universität Zürich,
Hirschengraben 84, CH-8001 Zürich, Switzerland
[*]*email:* tilmann@stat.washington.edu

Summary. We discuss tools for the evaluation of probabilistic forecasts and the critique of statistical models for count data. Our proposals include a nonrandomized version of the probability integral transform, marginal calibration diagrams, and proper scoring rules, such as the predictive deviance. In case studies, we critique count regression models for patent data, and assess the predictive performance of Bayesian age-period-cohort models for larynx cancer counts in Germany. The toolbox applies in Bayesian or classical and parametric or nonparametric settings and to any type of ordered discrete outcomes.

Key words: Calibration; Forecast verification; Model diagnostics; Predictive deviance; Probability integral transform; Proper scoring rule.

1. Introduction

One of the major purposes of statistical analysis is to make predictions, and to provide suitable measures of the uncertainty associated with them. Hence, forecasts ought to be probabilistic in nature, taking the form of probability distributions over future quantities and events (Dawid, 1984).

Here, we consider the evaluation of probabilistic forecasts, or predictive distributions, for count data, as they occur in a wide range of epidemiological, ecological, environmental, climatological, demographic, and economic applications (Christensen and Waagepetersen, 2002; Gotway and Wolfinger, 2003; McCabe and Martin, 2005; Elsner and Jagger, 2006; Frühwirth-Schnatter and Wagner, 2006; Nelson and Leroux, 2006; Frühwirth-Schnatter et al., 2009). Our focus is on the low count situation in which continuum approximations fail; however, our results apply to high counts and rates as well, as they occur routinely in epidemiological projections (Knorr-Held and Rainer, 2001; Clements, Armstrong, and Moolgavkar, 2005). To this date, statistical methods for the assessment of predictive performance have been studied primarily from biomedical, meteorological, and economic perspectives (Jolliffe and Stephenson, 2003; Pepe, 2003; Clements, 2005), focusing on predictions of dichotomous events or real-valued continuous variables. Here, we consider the hybrid case of count data, in which methods developed for either type of situation continue to be relevant but require technical adaption.

Gneiting, Balabdaoui, and Raftery (2007) contend that the goal of probabilistic forecasting is *to maximize the sharpness of the predictive distributions subject to calibration*. Calibration refers to the statistical consistency between the probabilistic forecasts and the observations, and is a joint property of the predictive distributions and the observations. Sharpness refers to the concentration of the predictive distributions, and is a property of the forecasts only.

In Section 2, we introduce tools for calibration and sharpness checks, among them a nonrandomized version of the probability integral transform (PIT) that is tailored to count data, and the marginal calibration diagram. Section 3 discusses the use of scoring rules as omnibus performance measures. We stress the importance of propriety (Gneiting and Raftery, 2007), relate to classical measures of predictive performance, and identify the predictive deviance as a variant of the proper logarithmic score. Section 4 turns to a crossvalidation study, in which we apply these tools to critique count regression models for pharmaceutical and biomedical patents. The epidemiological case study in Section 5 evaluates the predictive performance of Bayesian age-period-cohort models for larynx cancer counts in Germany. The article closes with a discussion in Section 6.

For count data, a probabilistic forecast is a predictive probability distribution, P, on the set of the nonnegative integers. We denote its probability mass function by $(p_k)_{k=0}^{\infty}$ and the respective cumulative distribution function (CDF) by $(P_k)_{k=0}^{\infty}$. Generalizations of our proposed methodology to probabilistic forecasts for any type of ordered discrete data, as opposed to count data, are straightforward and given in an appendix. The tools are simple yet powerful, and they apply generally to problems of forecast evaluation, model criticism, and model diagnosis.

2. Calibration and Sharpness

As noted, probabilistic forecasts strive to maximize the sharpness of the predictive distributions subject to calibration. Calibration refers to the statistical consistency between the probabilistic forecasts and the observations, and its assessment

requires frequentist thinking (Rubin, 1984). Gneiting et al. (2007) distinguish various modes of calibration and propose tools for the assessment of calibration and sharpness for probabilistic forecasts of continuous variables. Here, we adapt their proposals to the case of count data.

2.1 *Probability Integral Transform*

Dawid (1984) proposed the use of the PIT for calibration checks. This is simply the value that the predictive CDF attains at the observation. If the observation is drawn from the predictive distribution—an ideal and desirable situation—and the predictive distribution is continuous, the PIT has a standard uniform distribution.

Calibration then is checked empirically, by plotting the empirical CDF of a set of PIT values and comparing to the identity function, or by plotting the histogram of the PIT values and checking for uniformity (Diebold, Gunther, and Tay, 1998; Gneiting et al., 2007). The PIT histogram is typically used informally as a diagnostic tool; formal tests can also be employed though they require care in interpretation (Hamill, 2001; Jolliffe, 2007). Deviations from uniformity hint at reasons for forecast failures and model deficiencies. U-shaped histograms indicate underdispersed predictive distributions, hump or inverse-U shaped histograms point at overdispersion, and skewed histograms occur when central tendencies are biased.

In the case of count data, the predictive distribution is discrete. Here, the PIT is no longer uniform under the hypothesis of an ideal forecast, for which the observed count is a random draw from the predictive distribution. To remedy this, several authors have suggested a *randomized* PIT. Specifically, if P is the predictive distribution, $x \sim P$ is the observed count, and v is standard uniform and independent of x, then

$$u = P_{x-1} + v(P_x - P_{x-1}), \tag{1}$$

where we define $P_{-1} = 0$, is standard uniform (Smith, 1985; Frühwirth-Schnatter, 1996; Liesenfeld, Nolte, and Pohlmeier, 2006; Brockwell, 2007). For time series data one typically considers one-step (or k-step) ahead predictions, based on a time series model fitted on past and current data, and checks for the independence of the randomized PIT, in addition to checks for uniformity.

Here we propose a *nonrandomized* yet uniform version of the PIT histogram. To this end, we replace the randomized PIT value in (1) by its conditional CDF given the observed count x, that is, by

$$F(u \,|\, x) = \begin{cases} 0, & u \le P_{x-1}, \\ (u - P_{x-1})/(P_x - P_{x-1}), & P_{x-1} \le u \le P_x, \\ 1, & u \ge P_x, \end{cases} \tag{2}$$

similarly to the *discrete grade transformation* in relative distribution methodologies for the social sciences (Handcock and Morris, 1999, p. 180). Calibration can be assessed by aggregating over a relevant set of n predictions and comparing the mean PIT,

$$\bar{F}(u) = \frac{1}{n} \sum_{i=1}^{n} F^{(i)}(u \,|\, x^{(i)}), \qquad 0 \le u \le 1, \tag{3}$$

where $F^{(i)}$ is based on the predictive distribution $P^{(i)}$ and the observed count $x^{(i)}$, to the CDF of the standard uniform law, that is, the identity function.

We prefer to perform this comparison by plotting a nonrandomized PIT histogram, which can be interpreted diagnostically in the ways described above. Specifically, we pick the number of bins, J, compute

$$f_j = \bar{F}\Big(\frac{j}{J}\Big) - \bar{F}\Big(\frac{j-1}{J}\Big)$$

for equally spaced bins $j = 1, \ldots, J$, plot a histogram with height f_j for bin j, and check for uniformity. Under the hypothesis of calibration, that is, if $x^{(i)} \sim P^{(i)}$ for all forecast cases $i = 1, \ldots, n$, it is straightforward to verify that $\bar{F}(u)$ has expectation u, so that we expect uniformity. Principled guidelines for the selection of the number of bins remain to be developed; however, $J = 10$ or $J = 20$ are typical choices that lead to visually informative displays.

2.2 *Marginal Calibration Diagram*

We now consider what Gneiting et al. (2007) refer to as *marginal calibration*. The idea is straightforward: If each observed count is a random draw from the respective probabilistic forecast, and if we aggregate over the individual predictive distributions, $P^{(i)}$, we expect the resulting composite distribution and the histogram of the observed counts to be statistically compatible. A *marginal calibration diagram* plots the predicted frequency,

$$\hat{p}_x = \sum_{i=1}^{n} \big(P_x^{(i)} - P_{x-1}^{(i)}\big) \quad \text{or} \quad \hat{p}_{(x_a, x_b]} = \sum_{i=1}^{n} \big(P_{x_b}^{(i)} - P_{x_a}^{(i)}\big),$$

for specific x values or intervals $(x_a, x_b]$, and compares to the empirical counterpart,

$$f_x = \sum_{i=1}^{n} \mathbf{1}\big(x^{(i)} = x\big) \quad \text{or} \quad f_{(x_a, x_b]} = \sum_{i=1}^{n} \mathbf{1}\big(x_a < x^{(i)} \le x_b\big).$$

Major discrepancies hint at reasons for forecast failures and model deficiencies. An example of this type of diagnostic tool is shown in the bottom panel of Figure 2 below. In the case in which all $P^{(i)}$ are equal, this type of diagnostic check is routine in the literature and mostly presented in tabular form.

2.3 *Sharpness*

Sharpness refers to the concentration of the predictive distributions. In the context of prediction intervals, this can be rephrased simply: The shorter the intervals, the sharper, and the sharper the better, subject to calibration. Prediction intervals for continuous predictive distributions are uniquely defined, and Gneiting et al. (2007) suggest to tabulate their average width, or to plot *sharpness diagrams*, which can be used as a diagnostic tool. Sharpness continues to be critical for count data; however, we have found these tools to be less useful for discrete predictive distributions, for the ambiguities in specifying prediction intervals. Our preferred way of addressing sharpness is indirectly, via proper scoring rules; see below.

2.4 *Simulation Study*

We consider the negative binomial distribution NB(λ, a) with mean $\lambda \ge 0$ and dispersion parameter $a \ge 0$, hence variance $\lambda(1 + a\lambda)$. If $a = 0$, this is simply the Poisson distribution P(λ). We sample 200 counts from an NB(5, $\frac{1}{2}$) distribution,

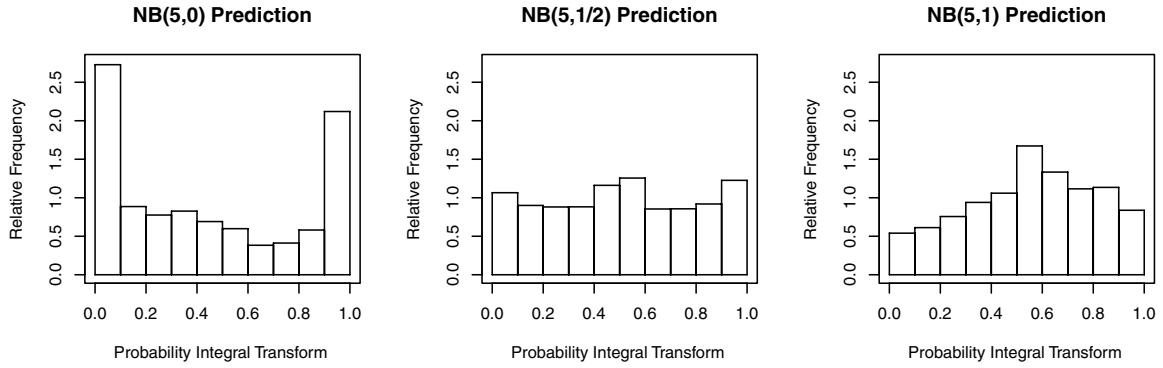

Figure 1. Nonrandomized PIT histograms for probabilistic forecasts for a sample of 200 counts from the negative binomial distribution NB(λ, a) with mean $\lambda = 5$, dispersion parameter $a = \frac{1}{2}$ and variance $\lambda(1 + a\lambda)$. The predictive distribution is negative binomial with mean $\lambda = 5$ and dispersion parameter $a = 0$, $a = \frac{1}{2}$, and $a = 1$ (from left to right). The PIT histograms are U-shaped, uniform, and inversely U-shaped, indicating underdispersed, well calibrated, and overdispersed predictive distributions, respectively.

and consider probabilistic forecasters whose predictive distribution is NB(5,0) = P(5), NB(5, $\frac{1}{2}$) and NB(5,1). Figure 1 shows nonrandomized PIT histograms with $J = 10$ equally spaced bins for these three cases. The PIT histograms are U-shaped, uniform, and inversely U-shaped, indicating underdispersed, well calibrated, and overdispersed predictive distributions, respectively.

In a web-based supplement, we provide code in the software environment `R` (Ihaka and Gentleman, 1996) that replicates this experiment.

3. Scoring Rules

Scoring rules provide summary measures in the evaluation of probabilistic forecasts, by assigning a numerical score based on the predictive distribution and the observation. We take scoring rules to be negatively oriented penalties that a forecaster wishes to minimize. Specifically, if the forecaster quotes the predictive distribution P and the count x realizes, the penalty is $s(P, x)$. We write $s(P, Q)$ for the expected value of $s(P, \cdot)$ under Q. In practice, scores are reported as averages over suitable sets of probabilistic forecasts, and we use upper case to denote a mean score; say

$$S = \frac{1}{n}\sum_{i=1}^{n} s(P^{(i)}, x^{(i)}),$$

where $P^{(i)}$ and $x^{(i)}$ refer to the ith predictive distribution and the ith observed count. In particular, the tables in our article show mean scores.

3.1 *Propriety*

Suppose, then, that the forecaster's best judgment is the predictive distribution Q. The forecaster has no incentive to predict any $P \neq Q$, and is encouraged to quote her true belief, $P = Q$, if

$$s(Q, Q) \leq s(P, Q), \tag{4}$$

with equality if and only if $P = Q$. A scoring rule with this property is said to be *strictly proper*. If $s(Q, Q) \leq s(P, Q)$ for all P and Q, the scoring rule is said to be *proper*. Propriety is an essential property of a scoring rule that encourages honest and coherent predictions (Bröcker and Smith, 2007; Gneiting and Raftery, 2007). Strict propriety ensures that both calibration and sharpness are being addressed (Winkler, 1996).

3.2 *Examples of Proper Scoring Rules*

The *logarithmic score* is defined as

$$\mathrm{logs}(P, x) = -\log p_x. \tag{5}$$

This is the only proper scoring rule that depends on the predictive distribution P only through the probability mass p_x at the observed count (Good, 1952).

There is a close relationship between the logarithmic score and the *predictive deviance*, defined as

$$\mathrm{dev}(P, x) = -2\log p_x + 2\log f_x,$$

where f_x is "some fully specified standardizing term that is a function of the data alone" (Spiegelhalter et al., 2002, p. 587). If the predictive distribution is a member of a one-parameter exponential family, such as the binomial or Poisson, the standardizing term is routinely taken to be the saturated deviance (McCullagh and Nelder, 1989; Knorr-Held and Rainer, 2001; Spiegelhalter et al., 2002; Clements et al., 2005). However, if the predictive distributions come from possibly distinct parametric or nonparametric families, it is vital that the standardizing terms in the deviance are common (Spiegelhalter et al., 2002). We contend that the choice is rather arbitrary and propose, for simplicity, that the standardizing term is taken to be zero as in Gschlößl and Czado (2007, 2008), which corresponds to the use of the logarithmic score.

Let $\|p\|^2 = \sum_{k=0}^{\infty} p_k^2$, which can frequently be computed analytically. For example, this is true for the Poisson and the negative binomial distribution. The *quadratic score* or *Brier score* and the *spherical score* are then defined as

$$\mathrm{qs}(P, x) = -2p_x + \|p\|^2, \tag{6}$$

and

$$\mathrm{sphs}(P, x) = -\frac{p_x}{\|p\|}, \tag{7}$$

respectively. Wecker (1989) proposed the use of the quadratic score in the assessment of time series predictions of counts.

The *ranked probability score* (Epstein, 1969) was originally introduced for ranked categorical data. It is easily adapted to count data, by defining

$$\mathrm{rps}(P, x) = \sum_{k=0}^{\infty} \{P_k - \mathbf{1}(x \le k)\}^2 . \tag{8}$$

Equation (14) in Gneiting and Raftery (2007) implies an alternative representation expressed in terms of expectations, which we now assume to be finite, namely

$$\mathrm{rps}(P, x) = E_P \,|\, X - x| - \frac{1}{2} E_P \,|\, X - X'|,$$

where X and X' are independent copies of a random variable with distribution P. The ranked probability score generalizes the absolute error, to which it reduces if P is a point forecast. Hence, it provides a direct way of comparing point forecasts and predictive distributions. The scores introduced in this section are strictly proper, except that the ranked probability score requires Q to have finite first moment for strict inequality in equation (4) to hold.

There is no automatic choice of a proper scoring rule to be used in any given situation, unless there is a unique and clearly defined underlying decision problem. However, in many types of situations probabilistic forecasts have multiple simultaneous uses, and it may be appropriate to use a variety of diagnostic tools and scores, to take advantage of their differing emphases and strengths. For instance, there is a distinct difference between the ranked probability score and the other scores discussed in this section, in that the former blows up score differentials between competing forecasters in cases in which predicted and/or observed counts are unusually high (Candille and Talagrand, 2005). Hence, a single high count case can dominate and obscure differences in the mean score. This type of behavior might be desirable, if the high count cases are the crucial ones, or might be undesirable, depending on the application at hand.

3.3 *Classical Measures of Predictive Performance*

We now discuss further traditional summary measures of predictive performance. For simplicity, we assume hereinafter that all moments considered are finite. Suppose first that $\mu \in \mathbb{R}$ is point forecast and the count x realizes. Typically, one uses the absolute error, $\mathrm{ae}(\mu,\, x) = |x - \mu|$, or the squared error, $\mathrm{se}(\mu,\, x) = (x - \mu)^2$, as a measure of predictive performance, averaging, again, over suitable sets of forecasts, to obtain the mean absolute error and mean squared error, respectively. Of course, these measures apply to probabilistic forecasts as well. For example, we can define the *squared error score*,

$$\mathrm{ses}(P, x) = (x - \mu_P)^2 , \tag{9}$$

where μ_P is the mean of the predictive distribution P. Viewed as a scoring rule for probabilistic forecasts, this score is proper, but not strictly proper (Gneiting and Raftery, 2007).

We now turn to studentized errors. It has frequently been argued that the *squared Pearson residual* or *normalized squared error score*,

$$\mathrm{nses}(P, x) = \left(\frac{x - \mu_P}{\sigma_P}\right)^2 , \tag{10}$$

where μ_P and σ_P^2 denote the mean and the variance of P, ought be approximately one when averaged over the predictions (Carroll and Cressie, 1997; Liesenfeld et al., 2006). Gotway and Wolfinger (2003) call the mean normalized squared error score the average *empirical-to-model variability ratio*, arguing also that it should be close to one. One way of justifying this is by noting that the function

$$f\left(\mu_P\,, \sigma_P^2\right) = \left(\mathrm{nses}(P, Q) - 1\right)^2 ,$$

has a minimum at $\mu_P = \mu_Q$ and $\sigma_P^2 = \sigma_Q^2$. Still, we follow Frühwirth-Schnatter (1996) in arguing that the PIT histogram is a more informative and more robust tool for unmasking dispersion errors. Of course, the normalized squared error score is improper, in that the expected score decays to zero as the predictive standard deviation, σ_P, tends to infinity.

The scores in this section depend on the predictive distribution P only through the first two moments. Dawid and Sebastiani (1999) provide a comprehensive study of proper scoring rules for which this property holds. A particularly appealing example is the scoring rule

$$\mathrm{dss}(P, x) = \left(\frac{x - \mu_P}{\sigma_P}\right)^2 + 2 \log \sigma_P , \tag{11}$$

to which we refer as the *Dawid–Sebastiani score*. It was proposed by Gneiting and Raftery (2007) as a proper alternative to the *predictive model choice criterion* of Gelfand and Ghosh (1998).

3.4 *Simulation Study*

We continue the simulation study of Section 2.4, which the reader can replicate using `R` code in a web-based supplement. We sample 200 counts from an $\mathrm{NB}(5, \frac{1}{2})$ distribution, and suppose that the predictive distribution is NB(5,0) = P(5), $\mathrm{NB}(5, \frac{1}{2})$ and NB(5, 1), respectively. Table 1 shows the mean score for each probabilistic forecast and each of six scoring rules (logarithmic, quadratic, spherical, ranked probability, squared error, and Dawid–Sebastiani scores). The $\mathrm{NB}(5, \frac{1}{2})$ forecast is correctly identified as superior. Of course, the predictive mean, and therefore the mean squared error score, is

Table 1

Scoring rules in a simulation experiment in which 200 *counts from an* $\mathrm{NB}(5, \frac{1}{2})$ *distribution are predicted as* NB(5,0), $\mathrm{NB}(5, \frac{1}{2})$, *and* NB(5,1). *The three forecasts are compared by the mean logarithmic, quadratic, spherical, ranked probability, Dawid–Sebastiani, and squared error scores.*

Model	LogS	QS	SphS	RPS	DSS	SES
NB(5,0) = P(5)	3.33	−0.052	−0.252	2.51	5.45	19.2
$\mathrm{NB}(5, \frac{1}{2})$	2.69	−0.083	−0.287	2.33	3.96	19.2
NB(5,1)	2.74	−0.073	−0.272	2.38	4.04	19.2

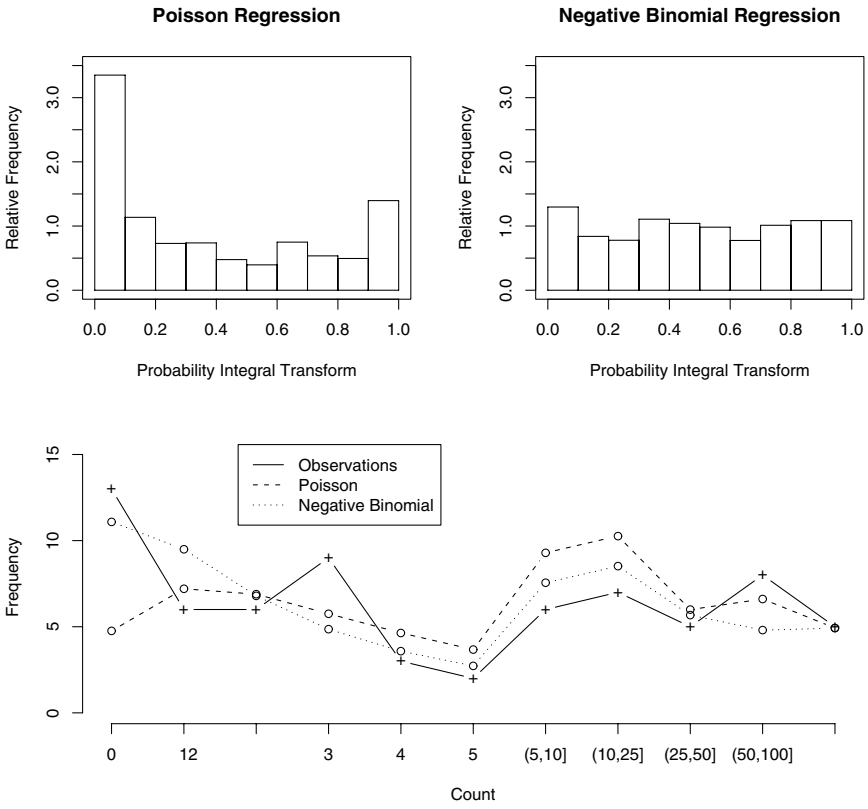

Figure 2. Nonrandomized PIT histograms and marginal calibration diagram for patent data count regressions. The marginal calibration diagram compares observed and hypothesized numbers of outcomes. For example, 13 out of the 70 observed patent counts are zero. Under the crossvalidated Poisson and negative binomial models, 4.8 and 11.1 outcomes equal to zero are expected.

the same for all three forecasts. The respective mean normalized squared error scores are 3.84, 1.10, and 0.64, thereby supporting the dispersion assessments that the PIT histograms in Figure 1 make more powerfully.

4. Case Study: Model Critique for Count Regression

Count data often show substantial extra variation or overdispersion relative to a Poisson regression model (Winkelmann, 2008). Various alternatives have been suggested to accommodate this, such as negative binomial and mixed Poisson models (Lawless, 1987). In this section, we investigate whether the nonrandomized PIT histogram, the marginal calibration diagram, and proper scoring rules are effective tools for model criticism (O'Hagan, 2003) in this context. We adopt a leave-one-out crossvalidation approach, in which the prediction for each observation is based on a count regression model fitted on the remaining data only. The Akaike information criterion (AIC) then is asymptotically equivalent to the mean logarithmic score (Stone, 1977).

We study the relationship of the number of patent applications to research and development (R&D) spending and sales using data from 1976 for 70 pharmaceutical and biomedical companies. The same data set was studied by Wang, Cockburn, and Puterman (1998), who used a mixed Poisson regression approach to address the overdispersion that is commonly observed in patent counts (Czado et al., 2007). Here we take a simpler approach and compare Poisson regression to negative binomial regression, using the specification

$$\log \lambda = \beta_0 + \beta_1 \frac{\text{R\&D}}{\text{sales}} + \beta_2 \left(\text{R\&D}\right)^{1/5},$$

for the predictive mean, λ. The top two panels of Figure 2 show nonrandomized PIT histograms based on the leave-one-out predictive distributions, using Poisson and negative binomial count regression models fitted with the `R` functions `glm()` and `glm.nb()`, which are available in the `R` library `MASS`. The PIT histogram for the Poisson case indicates underdispersion of the Poisson regression model. The histogram for the negative binomial case does not show any lack of model fit. The bottom panel of Figure 2 shows a marginal calibration diagram, as introduced in Section 2.2, which hints at zero inflation relative to the Poisson model.

Table 2 shows mean scores for the two competing methods. The scores prefer the negative binomial over the Poisson model, with the exception of the Dawid–Sebastiani score. For comparison, the AIC is 629.1 and 415.5 for the Poisson and negative binomial model. Vuong (1989) developed a comparison test for nonnested models, which is applicable here, because the dispersion parameter of the Poisson regression is at the boundary of the parameter space for the negative binomial regression model. The Vuong statistic is −3.01, indicating

Table 2

Mean scores for patent data count regressions. The Poisson and negative binomial models are compared by the mean logarithmic, quadratic, spherical, ranked probability, Dawid–Sebastiani, and squared error scores. The best value in each column is shown in bold face.

Model	LogS	QS	SphS	RPS	DSS	SES
Poisson	10.00	−0.10	−0.27	13.7	**14.6**	3476
Negative binomial	**4.33**	**−0.12**	**−0.31**	**7.5**	94.7	**562**

that the negative binomial model outperforms the Poisson model at the 0.01 level.

In conclusion, our diagnostic tools point at the superiority of the negative binomial regression model. The PIT histogram shows that the Poisson model is strongly underdispersed, and the marginal calibration diagram hints at zero inflation relative to the Poisson regression. The results in this section can be replicated using the `R` code that is available in a web-based supplement.

5. Case Study: Predicting Cancer Incidence

Bayesian age-period-cohort models are used increasingly to project cancer incidence and mortality rates. Data from younger age groups (typically age < 30 years) for which rates are low are often excluded from the analysis. However, a recent empirical comparison (Baker and Bray, 2005) based on data from Hungary suggests that age-specific predictions based on data from all age groups are more accurate. A natural question arises here in how to quantify the quality of the predictive distributions.

Baker and Bray (2005) predict mortality rates, using what they call the *sum of squared standardized residuals* to assess the quality of the forecasts. From personal communication with the authors, the standardization is not based on the predictive variance, so the aforementioned residuals are not the squared Pearson residuals in equation (10). Instead, Baker and Bray (2005) use the traditional standard error of a rate estimate for standardization and argue, in discussing their Table 1, that smaller values of this quantity correspond to more accurate predictions. Clements, Hakulinen, and Moolgavkar (2006) question the assessment in Baker and Bray (2005) and suggest the use of the predictive deviance, originally proposed by Knorr-Held and Rainer (2001). They argue that the Bayesian age-period-cohort model suffers from very wide credible intervals, but do not relate the width of the intervals to properties of calibration, and do not specifically recommend the use of proper scoring rules.

In this section, we use scoring rules to investigate whether the conclusion drawn by Baker and Bray (2005) applies to larynx cancer in Germany, 1952–2002. Our assessment is based on counts rather than rates. We fit four different predictive models depending on whether or not data from age groups <30 years have been included in the analysis, and whether or not the model allows for overdispersion, as shown in Table 3. Because the different models are based on different data, a comparison based on the AIC or related model fit criteria is not feasible here.

Table 3

Four predictive models for larynx cancer counts in Germany, 1998–2002. Models 1 and 2 use data from younger age groups; models 1 and 3 allow for overdispersion. The predictive distributions from the four methods are compared by the logarithmic, quadratic, spherical, ranked probability, Dawid–Sebastiani, and squared error scores. The best value in each column is shown in bold face.

Model	age	disp	LogS	QS	SphS	RPS	DSS	SES
1	+	+	**4.27**	**−0.041**	**−0.153**	14.0	**6.74**	852.9
2	+	−	4.35	−0.040	−0.152	12.9	6.89	684.4
3	−	+	4.29	−0.040	−0.152	14.2	6.78	870.0
4	−	−	4.35	−0.039	−0.151	**12.2**	6.90	**564.8**

Let n_{ij} be the number of persons at risk in age group i and year j. We assume that the respective number of deaths, X_{ij}, is binomially distributed with parameters n_{ij} and π_{ij}. A Poisson model would be a nearly identical choice. Following Besag et al. (1995) and Knorr-Held and Rainer (2001), we decompose the logarithmic odds, $\eta_{ij} = \log\{\pi_{ij}/(1-\pi_{ij})\}$, additively into an overall level, μ, age effects, θ_i, period effects, φ_j, and cohort effects, ψ_k, namely,

$$\eta_{ij} = \mu + \theta_i + \varphi_j + \psi_k.$$

Note that there is a problem in defining cohorts because age groups (in 5-year steps) and periods (in 1-year steps) are not on the same grid. We follow Knorr-Held and Rainer (2001) and use the cohort index $k = 5 \cdot (I - i) + j$, where I is the number of age groups.

Here we use nonparametric smoothing priors within a hierarchical Bayesian framework, for which model-based extrapolation of period and cohort effects for future periods is straightforward (Besag et al., 1995). This choice has the additional advantage that adjustments for overdispersion are easy to make. Inference and prediction based on Markov chain Monte Carlo techniques is done as described in Knorr-Held and Rainer (2001).

To assess the predictive performance of the different models, we predict mortality counts for the five years 1998–2002. For all different models, we consider predictions in the 12 age groups with age ≥ 30 years. Table 3 shows mean scores, averaged over all $12 \cdot 5 = 60$ projections. Interestingly, the scores do not agree. One set of scores (logarithmic, quadratic, spherical, and Dawid–Sebastiani scores) points to model 1 as the best, which includes data from the very young age groups and adjusts for overdispersion. The other set of scores (ranked probability and squared error scores) prefers model 4.

The disagreement can be explained as follows. The scores in the first set are roughly independent of the size of the counts in the different age groups. In contrast, the ranked probability and squared error scores are highly dependent on the size of the counts (Candille and Talagrand, 2005). Hence the results in the mid age groups, where the counts are highest and model 4 is more competitive, dominate the mean score.

These results might support Baker and Bray's (2005) contention that age-specific predictions based on full data yield sharper predictive distributions, and more accurate point forecasts for the younger age groups that benefit from the strong

cohort effect that is present here, particularly for the younger birth cohorts. However, the mean differences of all proper scores pointing to model 1 are statistically insignificant compared with model 4, based on a permutation test for paired observations. In contrast, the ranked probability and squared error scores suggest highly significant differences in mean scores, in favor of model 4. Of course, these findings are tentative, being based on 60 (dependent) predictions only, and further analysis is called for.

6. Discussion

We have demonstrated a toolbox for the assessment of the predictive performance of probabilistic forecasts for count data, which includes a nonrandomized PIT histogram, the marginal calibration diagram, and proper scoring rules. Simplicity, generality, and interpretability are attractive features of these tools; they apply in Bayesian or classical and parametric or nonparametric settings, and do not require models to be nested, nor be related in any way. Typically, they are used diagnostically, to identify model deficiencies and facilitate model comparison and model selection. Formal inference is often feasible (Clements, 2005; Jolliffe, 2007), but may not be the goal.

The same set of tools can be used for any type of ordered discrete outcomes, with Grammig and Kehrle (2008) giving one such example. Consider a probabilistic forecast P of a quantity that can attain the values $(x_j)_{j=-\infty}^{\infty}$, where $x_{j-1} < x_j < x_{j+1}$ for all j. If $(P_j)_{j=-\infty}^{\infty}$ denotes the associated CDF, we replace x by j to generalize the definition of the PIT in equations (1) and (2). Similarly, the marginal calibration diagram and proper scoring rules adapt easily.

The toolbox applies to two apparently distinct, yet closely related tasks. One is the evaluation of probabilistic forecasts that take the form of predictive distributions for future counts. Here, the PIT histogram and the marginal calibration diagram are employed diagnostically, and proper scoring rules allow us to rank competing forecasters. The other task is the critique of statistical models (O'Hagan, 2003); frequently, models can be fitted in crossvalidation mode, and can be assessed based on the quality of the ensuing probabilistic forecasts. We have demonstrated the toolbox in case studies in both types of situations. It is our belief that it can provide similar guidance in a very wide range of applied statistical problems for ordered discrete outcomes.

7. Supplementary Materials

`R` code for the analyses in Sections 2 to 4 is available under the Paper Information link at the *Biometrics* website, `http://www.biometrics.tibs.org`. The code includes functions for the nonrandomized PIT histogram, the marginal calibration diagram, and proper scoring rules, and reproduces Figures 1 and 2 and Tables 1 and 2.

Acknowledgements

The authors are grateful to the editors and referees for exceptionally helpful and constructive reports. C.C. was supported by the German Research Foundation under grant CZ 86/1–3. T.G. acknowledges support by the National Science Foundation under Awards ATM-0724721 and DMS-0706745 and by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research under Grant N00014-01-10745. L.H. acknowledges support by the Swiss National Science Foundation.

References

Baker, A. and Bray, I. (2005). Bayesian projections: What are the effects of excluding data from younger age groups? *American Journal of Epidemiology* **162,** 798–805.

Besag, J. E., Green, P. J., Higdon, D. M., and Mengersen, K. L. (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science* **10,** 3–41.

Bröcker, J. and Smith, L. A. (2007). Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting* **22,** 382–388.

Brockwell, A. E. (2007). Universal residuals: A multivariate transformation. *Statistics and Probability Letters* **77,** 1473–1478.

Candille, G. and Talagrand, O. (2005). Evaluation of probabilistic prediction systems of a scalar variable. *Quarterly Journal of the Royal Meteorological Society* **131,** 2131–2150.

Carroll, S. S. and Cressie, N. (1997). Spatial modeling of snow water equivalent using covariances estimated from spatial and geomorphic attributes. *Journal of Hydrology* **190,** 42–59.

Christensen, O. F. and Waagepetersen, R. (2002). Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics* **58,** 280–286.

Clements, M. P. (2005). *Evaluating Econometric Forecasts of Economic and Financial Variables*. Basingstoke, U.K.: Palgrave Macmillan.

Clements, M. S., Armstrong, B. K., and Moolgavkar, S. H. (2005). Lung cancer rate predictions using generalized additive models. *Biostatistics* **6,** 576–589.

Clements, M. S., Hakulinen, T., and Moolgavkar, S. H. (2006). Re: Bayesian projections: What are the effects of excluding data from younger age groups? *American Journal of Epidemiology* **164,** 292–293.

Czado, C., Erhardt, V., Min, A., and Wagner, S. (2007). Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates. *Statistical Modelling* **7,** 125–153.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society, Series A, General* **147,** 278–292.

Dawid, A. P. and Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design. *Annals of Statistics* **27,** 65–81.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review* **39,** 863–883.

Elsner, J. B. and Jagger, T. H. (2006). Prediction models for annual U.S. hurricane counts. *Journal of Climate* **19,** 2935–2952.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology* **8,** 985–987.

Frühwirth-Schnatter, S. (1996). Recursive residuals and model diagnostics for normal and non-normal state space models. *Environmental and Ecological Statistics* **3,** 291–309.

Frühwirth-Schnatter, S. and Wagner, H. (2006). Auxiliary mixture sampling for parameter-driven models of time series of counts with applications to state space modelling. *Biometrika* **93,** 827–841.

Frühwirth-Schnatter, S., Frühwirth, R., Held, L., and Rue, H. (2009). Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data. *Statistics and Computing*, in press.

Gelfand, A. E. and Ghosh, S. K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika* **85,** 1–11.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102,** 359–378.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society, Series B, Statistical Methodology* **69,** 243–268.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society, Series B, Methodological* **14,** 107–114.

Gotway, C. A. and Wolfinger, R. D. (2003). Spatial prediction of counts and rates. *Statistics in Medicine* **22,** 1415–1432.

Grammig, J. and Kehrle, K. (2008). A new marked point process model for the federal funds rate target: Methodology and forecast evaluation. *Journal of Economic Dynamics and Control* **32,** 2370–2396.

Gschlößl, S. and Czado, C. (2007). Spatial modelling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial Journal* **107,** 202–225.

Gschlößl, S. and Czado, C. (2008). Modelling count data with overdispersion and spatial effects. *Statistical Papers* **49,** 531–552.

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review* **129,** 550–560.

Handcock, M. S. and Morris, M. (1999). *Relative Distribution Methods in the Social Sciences*. New York: Springer.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5,** 299–314.

Jolliffe, I. T. (2007). Uncertainty and inference for verification measures. *Weather and Forecasting* **22,** 637–650.

Jolliffe, I. T. and Stephenson, D. B. (2003). *Forecast Verification: A Practicioner's Guide in Atmospheric Science*. Chichester, U.K.: John Wiley and Sons.

Knorr-Held, L. and Rainer, E. (2001). Projections of lung cancer mortality in West Germany: A case study in Bayesian prediction. *Biostatistics* **2,** 109–129.

Lawless, J. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics* **15,** 209–225.

Liesenfeld, R., Nolte, I., and Pohlmeier, W. (2006). Modeling financial transaction price movements: A dynamic integer count data model. *Empirical Economics* **30,** 795–825.

McCabe, B. P. M. and Martin, G. M. (2005). Bayesian predictions of low count time series. *International Journal of Forecasting* **21,** 315–330.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edition. London: Chapman and Hall.

Nelson, K. P. and Leroux, B. G. (2006). Statistical models for autocorrelated count data. *Statistics in Medicine* **25,** 1413–1430.

O'Hagan, A. (2003). HSSS model criticism. In *Highly Structured Stochastic Systems*, P. J. Green, N. L. Hjort, and S. Richardson (eds), 423–453. Oxford: Oxford University Press.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford: Oxford University Press.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* **12,** 1151–1172.

Smith, J. Q. (1985). Diagnostic checks of non-standard time series models. *Journal of Forecasting* **4,** 283–291.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B, Statistical Methodology* **64,** 583–639.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B, Methodological* **39,** 44–47.

Vuong, Q. H. (1989). Likelihood tests for model selection and nonnested hypotheses. *Econometrica* **57,** 307–333.

Wang, P., Cockburn, I. M., and Puterman, M. L. (1998). Analysis of patent data—a mixed-Poisson-regression-model approach. *Journal of Business and Economic Statistics* **16,** 27–41.

Wecker, W. B. (1989). Assessing the accuracy of time series model forecasts of count observations. *Journal of Business and Economic Statistics* **7,** 418–419.

Winkelmann, R. (2008). *Econometric Analysis of Count Data*, 5th edition. Berlin: Springer.

Winkler, R. L. (1996). Scoring rules and the evaluation of probabilities. *Test* **5,** 1–60.

Received September 2007. *Revised August* 2008.
Accepted September 2008.