Research Article

Taeho Kim, Benjamin Lieberman, George Luta, and Edsel A. Peña*

Prediction Regions for Poisson and Over-Dispersed Poisson Regression Models with Applications in Forecasting the Number of Deaths during the COVID-19 Pandemic

<https://doi.org/10.1515/stat-2020-0106>

Received Sep 25, 2020; accepted Nov 02, 2021

Abstract: Motivated by the Coronavirus Disease (COVID-19) pandemic, which is due to the SARS-CoV-2 virus, and the important problem of forecasting the number of daily deaths and the number of cumulative deaths, this paper examines the construction of prediction regions or intervals under the no-covariate or intercept-only Poisson model, the Poisson regression model, and a new over-dispersed Poisson regression model. These models are useful for settings with events of interest that are rare. For the no-covariate Poisson and the Poisson regression model, several prediction regions are developed and their performances are compared through simulation studies. The methods are applied to the problem of forecasting the number of daily deaths and the number of cumulative deaths in the United States (US) due to COVID-19. To examine their predictive accuracy in light of what actually happened, daily deaths data until May 15, 2020 were used to forecast cumulative deaths by June 1, 2020. It was observed that there is over-dispersion in the observed data relative to the Poisson regression model. A novel over-dispersed Poisson regression model is therefore proposed. This new model, which is distinct from the negative binomial regression (NBR) model, builds on frailty ideas in Survival Analysis and over-dispersion is quantified through an additional parameter. It has the flavor of a discrete measurement error model and with a viable physical interpretation in contrast to the NBR model. The Poisson regression model is a hidden model in this over-dispersed Poisson regression model, obtained as a limiting case when the over-dispersion parameter increases to infinity. A prediction region for the cumulative number of US deaths due to COVID-19 by October 1, 2020, given the data until September 1, 2020, is presented. Realized daily and cumulative deaths values from September 1st until September 25th are compared to the prediction region limits. Finally, the paper discusses limitations of the proposed procedures and mentions open research problems. It also pinpoints dangers and pitfalls when forecasting on a long horizon, especially during a pandemic where events, both foreseen and unforeseen, could impact point predictions and prediction regions.

Keywords: Non-homogeneous Poisson process; Non-linear regression models; Normal approximations to Poisson distribution; Over-dispersed Poisson models; Poisson models; Prediction regions; Predictions during pandemics; SARS-CoV-2 virus

2020 Mathematics Subject Classification: Primary: 62J02, 62P99; Secondary: 62F99, 62M10

***Corresponding Author: Edsel A. Peña:** Department of Statistics, University of South Carolina, Columbia, SC, 29208, USA; Email: pena@stat.sc.edu

Taeho Kim: Department of Statistics, University of Haifa, Haifa, 31905, Israel

Benjamin Lieberman: Department of Statistics, University of South Carolina, Columbia, SC, 29208, USA

George Luta: Department of Biostatistics, Bioinformatics & Biomathematics, Georgetown University, Washington, District of Columbia, 20057 USA; Department of Clinical Epidemiology, Aarhus University, Aarhus, DK-8200, Denmark; The Parker Institute, Copenhagen University Hospital, Frederiksberg, DK-2000, Denmark

1 Introduction

The main practical impetus of this work was the early phase – from March 2020 to June 2020 – of the Coronavirus Disease (COVID-19) pandemic (e.g. ECDC, 2020a), which was caused by the SARS-CoV-2 virus. The pandemic provided statisticians, data scientists, machine learners, and other modelers a real-time laboratory to test and demonstrate their forecasting skills and abilities, with the quality of their forecasts assessable in a matter of days, weeks, or months. See, for instance, <https://covid19-projections.com> from the Massachusetts Institute of Technology (MIT) and the Institute of Health Metrics (IHME)'s <https://covid19.healthdata.org/united-states-of-america> based at the University of Washington in Seattle, as well as Kramer (2020) discussing the complexities of modeling pandemics. Of particular interests were the forecasting of the numbers of daily cases¹, deaths, and hospitalizations, or the cumulative cases, deaths, and hospitalizations attributable to COVID-19 at a future date in a specified country or a locality (e.g., a county, state, or province) on the basis of currently observed cases, deaths, and hospitalizations data. Such forecasts are critical in the decision-making process by government officials, business leaders, and educational and university administrators regarding termination of lockdowns, lessening of social distancing and other mitigation regulations, opening of businesses, or continuing with online class formats in K-12 schools, colleges, and universities.

A specific motivator of this paper was the data set depicted in the left panel of Figure 1 which provides the daily number of reported deaths due to COVID-19 for the United States (US) with respect to the number of days since December 31, 2019 (coded as Day 1) and until May 15, 2020 (coded as Day 137). The data was reported by the European Center for Disease Control (ECDC) (see, ECDC, 2020b).² Because of a reporting delay, for a given date/day, including weekends, the data set reports the numbers from the preceding day. The right panel of Figure 1 is a scatterplot of the cumulative number of deaths in the US due to COVID-19. In real time during that period, given the daily and cumulative deaths until May 15, 2020, of interest was to forecast the number of cumulative deaths in the US by, say, May 25, 2020 (corresponding to Day 147), which is Memorial Day, and to ask whether by that day the cumulative number of deaths in the US due to COVID-19 will have surpassed the ominously depressing and grim milestone of 100,000 cumulative deaths. Later, for our illustration, we considered the problem of forecasting the cumulative number of deaths in the US due to COVID-19 at the end of May 2020, and to compare our forecast with what eventually transpired. Finally, we attempt to forecast the cumulative number of deaths by October 1, 2020 based on the data until September 1, 2020.

Such type of forecasting problems are clearly non-trivial because of the possibility that whatever model was fitted for the observed time period, it may not apply to the time period under forecast, the ever-present danger and risk of extrapolation. Aside from the fitted model being most likely not the true data generating model – recalling the aphorism attributed to George E. P. Box (Box, 1976) that *all models are wrong, but some are useful* – there are other factors, some beyond our control, that could impact the number of reported deaths at a future time, such as premature easing of social distancing and re-opening of business establishments, virus mutations, better diagnostic tools, changing hotspots, overburdened health care facilities, introduction of effective treatments, beneficial or detrimental actions by local, state, and/or federal entities, changing definition of deaths due to COVID-19, under- or over-reporting of deaths, timely development and introduction of vaccines, protests and riots arising from social unrests, and others. But high-level decision-makers such as government officials, business leaders, educational administrators, and society itself, demand some beacon, however dim it may be, to guide their decision-making. Statisticians, data scientists, machine learners, and other modelers are always ready and willing to provide such beacons.

This paper is in this spirit. We comprehensively examine existing procedures and develop new methods for constructing prediction regions for random variables that pertain to the number of occurrences of an event of interest. A prediction region contains more information compared to just a point prediction since it provides

¹ COVID-19 cases include confirmed and probable cases (infected people) and deaths according to a statement by the Council of State and Territorial Epidemiologists (CSTE) updated on September 14, 2021. See the Interim Report CSTE (2021).

² Copyright Statement from the European CDC Website can be found in <https://www.ecdc.europa.eu/en/copyright>.

Figure 1: Scatterplot of the number of reported daily deaths and the cumulative deaths due to COVID-19 in the United States with respect to the number of days (DayNum) starting from December 31, 2019 as reported by the European Center for Disease Control. Shown in the plots are from DayNum = 62, coinciding with March 1, 2020, until DayNum = 137, which corresponds to May 15, 2020.

information about the uncertainty inherent in the prediction, hence it is a form of uncertainty quantification (UQ). Proper UQ is critical in making crucial decisions and in order for stakeholders (e.g., scientists, ordinary citizens, decision-makers) not to lose trust in statistical, data science, or machine learning methods. Note that with a prediction region we are interested in the would-be realized value of a random variable, not the value of a parameter, hence instead of referring to it as a *confidence* region, it is called a *prediction* region. The events of particular interest are those that are ‘rare’ in the sense that, informally, the probability of an event occurring in an infinitesimal interval is also infinitesimal. Consequently, our starting point will be the Poisson distribution, which is a model for the number of occurrences of a rare event. We then transition to the more general Poisson regression model, and eventually to an over-dispersed Poisson regression model. The last model was borne out upon our realization that the Poisson regression model is a deficient model for the COVID-19 data, which exhibited over-dispersion, and this newly-introduced model turned out to be a better model than the Poisson regression model for the COVID-19 application. Under the Poisson model, the mean and the variance are identical, so over-dispersed models are those where the variance exceeds the mean. The over-dispersed regression model proposed is motivated by frailty ideas in Survival Analysis, the area dealing with the study of time-to-event variables. The proposed model is different from other over-dispersed model that have been considered previously, such as the negative binomial (NB) regression model. Moment-wise, there are some similarities to the so-called NB12 model, but distribution-wise they are different. In addition, the new model has a discrete measurement error model flavor and a viable physical interpretation. Since rare events occur in many real-life and practical fields, the methodological results in this paper will be applicable in a variety of settings.

The specific real-life and practical problem for which our methods will be applied is the construction of prediction regions for the daily and cumulative numbers of deaths due to COVID-19 in the US for a future date, given *only* the daily deaths data until a current date. Note that such predictions or forecasts could probably be improved by utilizing other information, such as the capacities of health care facilities; movements of people in a region; information about sensitivity and specificity of diagnostic tests; transmission rates (R_0) of the virus; status and implementation of mitigation guidelines, and others. Another avenue for improving forecasts is through stratification by states, cities, or counties, and then combining the results from these strata to obtain a prediction region for the whole US. However, we approach the construction of the prediction regions for the daily and cumulative deaths at a future date by *just* utilizing the observed reported daily deaths data for the whole US, which in a sense is the most reliable *available* data regarding the COVID-19 pandemic. It might be possible to utilize information about the number of cases or infected individuals, which is also reported daily, but we have reasons to believe that this is not a reliable information since it is highly dependent

on the number of tests that are performed and on the sensitivity and specificity of the diagnostic tests used. In addition, if such information is to be used in the prediction model, then we may not have their realized values at the future date at which the prediction region is desired. We point out that even though we are employing probabilistic models in the form of the Poisson or an over-dispersed Poisson model, which are derivable from intuitive conditions when dealing with rare events (cf. Resnick, 1992; Liu and Peña, 2016), our prediction methods are still purely data-driven, being only reliant on the observed data and not on other extraneous nor subjective information.

2 Poisson Model

The occurrence of a death due to COVID-19 could still be considered as a rare event when viewed in the context of the whole population, though even if it is rare, deaths are still significant and dire events. This is because to die of COVID-19, one first needs to get infected, which at this point is still a rare event, and then having been infected, to die from it. The rate of dying when infected with COVID-19, if not age-adjusted, is still rather low, less than 2% (see, for instance, the website Coronavirus (COVID-19) Mortality Rate). Because of its rarity, a plausible probability model for the number of deaths due to COVID-19 is therefore the Poisson model whose probability mass function (PMF) is given by

$$p(k|\lambda) = \frac{\exp(-\lambda)\lambda^k}{k!} I\{k \in \mathbb{Z}_{0,+}\} \quad (1)$$

with $\mathbb{Z}_{0,+} = \{0, 1, 2, \dots\}$; $\lambda > 0$ is the rate parameter, which is also the mean and variance of the distribution; and $I\{\cdot\}$ is the indicator function. For a variable Y with this Poisson distribution, we write $Y \sim POI(\lambda)$. The cumulative distribution function of a $POI(\lambda)$ is

$$P(w|\lambda) = \Pr\{Y \leq w|\lambda\} = \sum_{\{k \in \mathbb{Z}_{0,+}: k \leq w\}} p(k|\lambda) I\{w \geq 0\}. \quad (2)$$

We start our investigation with this no-covariate Poisson model, or equivalently, an intercept-only model. The results for the Poisson regression model will build on this no-covariate model.

2.1 Prediction Regions when Rate λ is Known

Suppose that $Y_0 \sim POI(\lambda)$, where for the moment we assume that we know the rate parameter λ . A $100(1 - \alpha)\%$ prediction region for Y_0 is a subset $\Gamma(\lambda, \alpha) \subset \mathbb{Z}_{0,+}$ such that

$$\Pr\{Y_0 \in \Gamma(\lambda, \alpha)|\lambda\} = \sum_{k \in \Gamma(\lambda, \alpha)} p(k|\lambda) \geq 1 - \alpha. \quad (3)$$

Note that this region will not be an interval since it is a subset of $\mathbb{Z}_{0,+}$, though if this region can be formed as the intersection between $\mathbb{Z}_{0,+}$ and an interval in $\mathbb{R}$, then we may call it imprecisely as an interval. However, the region need not always have such a structure, e.g., $\Gamma = \{1, 2, 3\} \cup \{5, 6\}$. Subject to condition (3), a desirable property of such a region is that its cardinality (its number of elements) is as small as possible. The cardinality of a region Γ is a measure of the size of Γ . When the region can be formed as the intersection of an interval in $\mathbb{R}$ and $\mathbb{Z}_{0,+}$, then a surrogate for its cardinality is the *length* of the shortest interval $S \subset \mathbb{R}$ satisfying $\Gamma = S \cap \mathbb{Z}_{0,+}$. As an example, for the region $\Gamma = \{1, 2, 3, 4, 5, 6\}$, its cardinality is 6, while its length is 5, being the length of $S = [1, 6]$. Later, in evaluating the performances of prediction regions, we will use length as defined above to measure the size of the regions. If we allow for $\Gamma(\lambda, \alpha)$ to depend on a randomizer U , a standard uniform random variable independent of Y_0 , the smallest cardinality $100(1 - \alpha)\%$ prediction region is, using a Neyman-Pearson Lemma type argument, given by

$$\Gamma_0(U; \lambda, \alpha) = A(\lambda; c(\alpha)) \bigcup \left\{ \{U \leq \gamma(\alpha)\} \cap a(\lambda; c(\alpha)) \right\}, \quad (4)$$

where, for $d \in \mathbb{R}$, we define the subsets of $\mathbb{Z}_{0,+}$ given by

$$A(\lambda; d) = \{k \in \mathbb{Z}_{0,+} : p(k|\lambda) > d\} \text{ and } a(\lambda; d) = \{k \in \mathbb{Z}_{0,+} : p(k|\lambda) = d\},$$

and $c(\alpha)$ and $\gamma(\alpha)$ determined via

$$c(\alpha) = \inf \{d \geq 0 : \Pr\{Y_0 \in A(\lambda; d)|\lambda\} \leq 1 - \alpha\} \text{ and } \gamma(\alpha) = \frac{(1-\alpha) - \Pr\{Y_0 \in A(\lambda; c(\alpha))|\lambda\}}{\Pr\{Y_0 \in a(\lambda; c(\alpha))|\lambda\}},$$

where we adopt, for notational brevity, the mathematically inaccurate convention that $0/0 = 0$. Observe that by allowing randomized prediction regions, we have $\Pr\{Y_0 \in \Gamma_0(U; \lambda, \alpha)|\lambda\} = 1 - \alpha$. If we do not admit randomized prediction regions, such as in Bejleri and Nandram (2018), tantamount to always taking $U = 0$ in $\Gamma_0(U; \lambda, \alpha)$, then unless $1 - \alpha$ is a ‘natural’ prediction coefficient, we will not achieve equality in the preceding probability statement.³ For the application of Neyman-Pearson-type arguments in constructing optimal confidence regions, see Peña and Kim (2019). The method is easily adapted in the construction of prediction regions.

There are two other ways of constructing prediction intervals for Y_0 when λ is large using normal approximations. To obtain the prediction regions, these intervals are then intersected with $\mathbb{Z}_{0,+}$. Letting $N(\mu, \sigma^2)$ denote a normal distribution with mean μ and variance σ^2 , we recall that when λ is large and owing to the Central Limit Theorem and the Delta Method (cf. Casella and Berger, 1990), we have the normal approximations

$$Y_0 \sim N(\lambda, \lambda) \text{ and } \sqrt{Y_0} \sim N(\sqrt{\lambda}, 1/4).$$

Let $\phi(\cdot)$ and $\Phi(\cdot)$ be the probability density and cumulative distribution functions of a standard normal random variable, so that $\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}z^2\}$ and $\Phi(z) = \int_{-\infty}^z \phi(w)dw$. Let $z_\alpha = \Phi^{-1}(1 - \alpha)$ be its $(1 - \alpha)$ th quantile. Two approximate prediction regions for Y_0 when λ is large, which are based on the above normal approximations, are given by

$$\Gamma_1(\lambda, \alpha) = [\max\{0, \lambda - z_{\alpha/2}\sqrt{\lambda}\}, \lambda + z_{\alpha/2}\sqrt{\lambda}] \cap \mathbb{Z}_{0,+}; \quad (5)$$

$$\Gamma_2(\lambda, \alpha) = \left[\left(\max\{0, \sqrt{\lambda} - (z_{\alpha/2})/2 \} \right)^2, \left(\sqrt{\lambda} + (z_{\alpha/2})/2 \right)^2 \right] \cap \mathbb{Z}_{0,+}. \quad (6)$$

When λ is large, as noted in the construction of Γ_1 , we may approximate the Poisson probabilities by normal probabilities, via $p(k|\lambda) \approx \frac{1}{\sqrt{\lambda}} \phi\left(\frac{k-\lambda}{\sqrt{\lambda}}\right)$. As such, we obtain the approximation

$$A(\lambda; c(\alpha)) \approx [\lambda - z_{\alpha/2}\sqrt{\lambda}, \lambda + z_{\alpha/2}\sqrt{\lambda}] \cap \mathbb{Z}_{0,+}.$$

Consequently, when λ is large, the regions $\Gamma_0(U; \lambda, \alpha)$ and $\Gamma_1(\lambda, \alpha)$ should be close to each other.

For these prediction regions Γ_{0R} (randomized), Γ_{0N} (nonrandomized), Γ_1 , and Γ_2 , the exact coverage probabilities and their exact lengths (mean length for Γ_{0R}) could be computed under the $POI(\lambda)$ distribution, since λ is known. As was mentioned above, the length of a region, which is the difference between the upper and lower integer limits of the prediction region, is a surrogate of the cardinality of the region. Figure 2 depicts the exact coverage probabilities (CP), expressed in percentages, and their lengths (expected length for Γ_{0R}), for different values of λ . Except when λ takes small values, where the coverage probabilities of Γ_1 and Γ_2 are degraded, especially for the latter, the performance of these prediction regions are quite similar. The coverage probability of Γ_{0R} is exactly equal to $1 - \alpha$, whereas that of Γ_{0N} is always at least equal to $1 - \alpha$. Both Γ_1 and Γ_2 could have coverage probabilities below the nominal coverage level, though as λ increases, the difference between the nominal coverage level and the observed coverage probability becomes negligible. By construction, Γ_{0R} has a shorter length than Γ_{0N} ; for some values of λ , the length of Γ_{0R} exceeds that of Γ_1 and Γ_2 , but this is because the coverage probabilities of Γ_1 and Γ_2 are lower than the nominal coverage level. Based on these simulation results, Γ_1 appears preferable to Γ_2 for the reason that it performs better when λ is not large.

³ The use of the adjective ‘natural’ is analogous to its use in constructing nonparametric confidence intervals, (cf. Randles and Wolfe, 1979).

Figure 2: Exact properties of the four prediction regions $\Gamma_{OR}, \Gamma_{ON}, \Gamma_1$, and Γ_2 for $Y_0 \sim POI(\lambda)$ under the situation when the Poisson rate λ is known.

2.2 Prediction Regions when Rate λ is Unknown

In the preceding developments, we assumed that the rate parameter λ is known, which is an unrealistic assumption. How do we deal with the situation when λ is unknown? Suppose that we had observed a realization $\mathbf{y} = (y_1, y_2, \dots, y_n)$ of a random sample $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ from $POI(\lambda)$, so the components of $\mathbf{Y}$ are independent and identically distributed (IID) from $POI(\lambda)$. Our goal is to utilize $\mathbf{y}$ to construct a $100(1 - \alpha)\%$ prediction region for an unobserved Y_0 , which is independent of $\mathbf{Y}$ and whose distribution is also $POI(\lambda)$. Note that through the Sufficiency Principle, we may simplify the problem by assuming that we had observed $t = \sum_{i=1}^n y_i$, the realization of the sufficient statistic for λ given by $T = \sum_{i=1}^n Y_i$, which has a $POI(n\lambda)$ distribution. The reduced problem therefore is that we have (T, Y_0) which are independent random variables with $T \sim POI(n\lambda)$ and $Y_0 \sim POI(\lambda)$ and our goal is to construct a $100(1 - \alpha)\%$ prediction region $\tilde{T}(T, U; \alpha)$ for Y_0 , which utilizes T , and possibly a randomizer U independent of (T, Y_0) . Note that this situation is analogous to the setting in Bejleri and Nandram (2018) where for the variable T , denoted by X in their paper, the rate parameter is $s\lambda$; while for the variable Y_0 , denoted by Y in their paper, the rate parameter is $t\lambda$ (see Kim *et al.*, 2021).

Given $T = t$, the maximum likelihood estimate (MLE) of λ is $\hat{\lambda}(t) = t/n$. By virtue of the consistency of $\hat{\lambda}(T)$ for λ as $n \rightarrow \infty$, a seemingly straight-forward approach to constructing a prediction region for Y_0 is to replace λ in $\Gamma_0(U; \lambda, \alpha)$, $\Gamma_1(\lambda, \alpha)$, and $\Gamma_2(\lambda, \alpha)$ in (4), (5), and (6), respectively, by $\hat{\lambda}(t)$ to obtain

$$\tilde{\Gamma}_0(T, U; \alpha) = \Gamma_0(U; \hat{\lambda}(T), \alpha); \quad (7)$$

$$\tilde{\Gamma}_1(T; \alpha) = \Gamma_1(\hat{\lambda}(T), \alpha); \quad (8)$$

$$\tilde{\Gamma}_2(T; \alpha) = \Gamma_2(\hat{\lambda}(T), \alpha). \quad (9)$$

How do these prediction regions compare with each other in terms of their performances, as measured by their coverage probabilities and also their cardinalities or their lengths? In particular, by substituting $\hat{\lambda}(T) = T/n$ for λ , how does this impact the coverage probabilities of these prediction regions and are they still valid, even in an asymptotic sense?

It is not obvious how the substitution of λ by $\hat{\lambda}(T) = T/n$ will impact the *exact* performance of the first prediction region $\tilde{\Gamma}_0$. However, for the second and third prediction regions $\tilde{\Gamma}_1$ and $\tilde{\Gamma}_2$, we could alter them to

take into account the substitutions, provided that λ is large. As noted earlier, when λ is large, $\tilde{r}_0 \approx \tilde{r}_1$, so the alteration of $\tilde{r}_1$ should also apply, approximately, to $\tilde{r}_0$. The change in distributions of the pivotal quantities arising from these substitutions are reflected below, a consequence of the Delta-Method.

$$\frac{Y_0 - \hat{\lambda}(T)}{\sqrt{\hat{\lambda}(T)}} \dot{\sim} N\left(0, 1 + \frac{1}{n}\right) \quad \text{and} \quad \frac{\sqrt{Y_0} - \sqrt{\hat{\lambda}(T)}}{\sqrt{1/4}} \dot{\sim} N\left(0, 1 + \frac{1}{n}\right).$$

From these normal approximations, we could improve the prediction regions $\tilde{r}_1$ and $\tilde{r}_2$ into the following prediction regions, which take into account the impact of these substitutions where, for notational economy, we write $\hat{\lambda}$ for $\hat{\lambda}(T)$ and ' $\vee$ ' for max; $\lceil \cdot \rceil$ for the ceiling function; and $\lfloor \cdot \rfloor$ for the floor function:

$$\tilde{r}_1(n, T; \alpha) = \left[\left[0 \vee \left\{ \hat{\lambda} - z_{\alpha/2} \sqrt{\hat{\lambda}(1 + 1/n)} \right\}, \left\{ \hat{\lambda} + z_{\alpha/2} \sqrt{\hat{\lambda}(1 + 1/n)} \right\} \right] \cap \mathbb{Z}_{0,+}; \right. \quad (10)$$

$$\left. \tilde{r}_2(n, T; \alpha) = \left[\left[\left(0 \vee \left\{ \sqrt{\hat{\lambda}} - (z_{\alpha/2}) \sqrt{\frac{1}{4} \left(1 + \frac{1}{n} \right)} \right\} \right)^2, \left\{ \left(\sqrt{\hat{\lambda}} + (z_{\alpha/2}) \sqrt{\frac{1}{4} \left(1 + \frac{1}{n} \right)} \right)^2 \right\} \right] \cap \mathbb{Z}_{0,+}. \right. \quad (11)$$

Note that by intersecting the intervals with $\mathbb{Z}_{0,+}$, the floor and ceiling functions are actually not needed, as well as taking the maximum values between 0 and the initial lower limits, but we retain them in the formula since when we consider the 'length' of the prediction region, this will pertain to the length of the interval. Observe that if the lower limits of these intervals are not zeros, which will usually be the case for large λ , then it is straightforward to show that these two prediction intervals have the same lengths, but they are not identical prediction regions.

In trying to adapt the prediction region $\Gamma_0(U; \lambda, \alpha)$ in (4) to the situation where λ is unknown, the main idea is to replace λ by an estimate obtained from the observed data. Doing so leads to estimates of $p(k|\lambda)$, $k \in \mathbb{Z}_{0,+}$ which are then used in determining $c(\alpha)$ and $\gamma(\alpha)$ in (4). Thus, $\tilde{r}_0$ in (7) is obtained by using the ML estimates of $\{p(k|\lambda), k = 0, 1, \dots\}$ given by $\{p(k|\hat{\lambda}(n, t)), k = 0, 1, \dots\}$ with $\hat{\lambda}(n, t) = t/n$. This begs the question on whether other possible estimates of $\{p(k|\lambda), k = 0, 1, \dots\}$ could be utilized which may have better performances than the use of the ML estimates. An approach based on a second-order Taylor expansion adjusts $p(k|\hat{\lambda}(n, t))$ and leads to the approximation

$$p(k|\lambda) \approx \hat{p}_3(k; (n, t)) \equiv \frac{p(k|\hat{\lambda}(n, t))}{1 + \frac{1}{2} \left[\left(1 - \frac{k}{\hat{\lambda}} \right)^2 - \frac{k}{\hat{\lambda}^2} \right] \frac{\hat{\lambda}}{n}} I\{k \in \mathbb{Z}_{0,+}\}. \quad (12)$$

Substituting $\hat{p}_3(k; (n, t))$ for $p(k|\lambda)$ in (4) results in the prediction region denoted by $\tilde{r}_3(n, t; \alpha)$. Another intriguing possibility is to utilize the uniformly minimum variance unbiased estimator (UMVUE) (see Casella and Berger, 1990) of $p(k|\lambda)$, given the data (n, T) with $T \sim POI(n\lambda)$. The UMVUE of $p(k|\lambda)$, usually obtained through the Rao-Blackwell Theorem and Lehmann-Scheffe Theorem (see Casella and Berger, 1990), is

$$\hat{p}_4(k; (n, t)) = \binom{t}{k} \left(\frac{1}{n} \right)^k \left(1 - \frac{1}{n} \right)^{t-k} I\{k \in \{0, 1, \dots, t\}\}, \quad (13)$$

the binomial probability at k with parameters $(t, 1/n)$. Observe, however, that this approximation will lead to zero probabilities for k outside of the set $\{0, 1, \dots, t\}$. Using $\hat{p}_4(k; (n, t))$ in lieu of $p(k|\lambda)$ in (4) leads to the prediction region denoted by $\tilde{r}_4(n, t; \alpha)$.

As yet another idea is to develop a procedure by borrowing from the Bayesian playbook (see Casella and Berger, 1990). We suppose that our prior knowledge of the value of the Poisson rate λ is represented by a distribution function G . Having observed $\mathbf{Y} = \mathbf{y} = (y_1, y_2, \dots, y_n)$, the posterior distribution of λ is given by

$$G(\lambda|\mathbf{y}) = \frac{\int_0^\lambda \exp\{-nw\} w^t G(dw)}{\int_0^\infty \exp\{-nw\} w^t G(dw)} \quad (14)$$

with $t = t(\mathbf{y}) = \sum_{i=1}^n y_i$. The conditional probability mass function of Y_0 , given $\mathbf{Y} = \mathbf{y}$, also called the posterior predictive PMF, is

$$p(y_0|\mathbf{y}; G) = \frac{1}{y_0!} \frac{\int_0^\infty \exp\{-(n+1)w\} w^{t+y_0} G(dw)}{\int_0^\infty \exp\{-nw\} w^t G(dw)}. \quad (15)$$

Bayesians then use the predictive PMF $p(\cdot|\mathbf{y}; G)$ in lieu of the Poisson PMF in (4) to form a prediction region for Y_0 . This approach was implemented in Bejleri and Nandram (2018) where they used a conjugate gamma distribution prior for G , as well as two improper priors obtained as limits of this gamma prior (uniform and Jeffreys). The other approach is to estimate G by a $\hat{G}(\cdot; \mathbf{y})$ based on $\mathbf{y}$. This is the essence of the Empirical Bayes (EB) approach, pioneered by Herbert Robbins; see Robbins (1956, 1980, 1983). An extreme case is to ‘estimate’ G by a *degenerate* distribution at the ML estimate $\hat{\lambda} = t/n$, which leads to just substituting $\hat{\lambda}$ in the Poisson PMF, resulting in the prediction region $\tilde{T}_0$ in (7). Another option is to estimate G non-parametrically. We restrict our implementation of this Bayesian and EB approaches by just using a family of conjugate priors. Thus, we assume that G is a gamma distribution, as in Bejleri and Nandram (2018), with mean κ/β and variance κ/β^2 , denoted by $\mathfrak{G}_{\kappa, \beta}$, with density function $g(\lambda|\kappa, \beta) = \frac{\beta^\kappa}{\Gamma(\kappa)} \lambda^{\kappa-1} \exp\{-\beta\lambda\} I\{\lambda > 0\}$ where $\kappa > 0$ is the shape parameter and $\beta > 0$ is the rate parameter. Under $G = \mathfrak{G}_{\kappa, \beta}$, simplifying (15) we obtain, for $y_0 \in \mathbb{Z}_{0,+}$,

$$p(y_0|\mathbf{y}; \mathfrak{G}_{\kappa, \beta}) = \frac{\Gamma(\kappa + t + y_0)}{\Gamma(y_0 + 1)\Gamma(\kappa + t)} \left(\frac{\beta + n}{\beta + n + 1} \right)^{t+\kappa} \left(\frac{1}{\beta + n + 1} \right)^{y_0}. \quad (16)$$

When κ is a positive integer, the PMF in (16) corresponds to a negative binomial distribution with parameters $\kappa + t$ and $(\beta + n)/(\beta + n + 1)$. The PMF in (16) could be used in place of the Poisson PMF in (4) to form a Bayesian prediction region for Y_0 , given (κ, β) , denoted by $\tilde{T}_5(U, \mathbf{Y}; (\kappa, \beta))$. An approach to specifying (κ, β) is to specify a prior mean and prior standard deviation for λ , say M and S , respectively, which yield $\kappa = M^2/S^2$ and $\beta = M/S^2$. The EB approach estimates κ and β from the data $\mathbf{y} = (y_1, y_2, \dots, y_n)$. Such estimation could be done via maximum likelihood using the likelihood function obtained from the joint marginal distribution of $(Y_1, Y_2, \dots, Y_n)$ based on the model $Y_i|\lambda_i \sim POI(\lambda_i)$ and $\lambda_i \sim \mathfrak{G}_{\kappa, \beta}$. This likelihood function is given by

$$L(\kappa, \beta|\mathbf{y}) = \frac{\prod_{i=1}^n \Gamma(\kappa + y_i)}{[\Gamma(\kappa)]^n \prod_{i=1}^n \Gamma(y_i + 1)} \left(\frac{\beta}{\beta + 1} \right)^{n\kappa} \left(\frac{1}{\beta + 1} \right)^{\sum_{i=1}^n y_i}. \quad (17)$$

A method-of-moments approach to estimating (κ, β) based on $\mathbf{y}$ fails, however, because negative estimates of κ and β are obtained when the sample variance of $\mathbf{y}$ is smaller than its sample mean.

At this point we mention previous works dealing with prediction regions under the Poisson model. It is a well-studied area both from a frequentist and Bayesian perspectives, as well as via approximate or exact approaches. For instance, Bain and Patel (1993), Weiss (1955), Nelson (2003), and Krishnamoorthy and Peng (2011) studied exact and approximate frequentist prediction regions, while Krishnamoorthy and Lv (2018) investigated exact Bayesian prediction regions. The papers Bejleri (2005) and Bejleri and Nandram (2018) also provide frequentist and Bayesian solutions to this prediction problem in the no-covariate setting. Their prediction region procedures did not allow for randomization, hence tended to be conservative. A recent review paper by Kim et al. (2021), motivated by an early version of the current paper, discusses works on this topic. Some prediction interval procedures for the Poisson model have been incorporated in the R package *envStats* (Millard, 2013). An object function in this package is `predIntPois` dealing with construction of prediction regions under the Poisson model. It provides four options for the type of prediction interval to construct. The methods are based on procedures presented in Cox and Hinkley (1974), Gibbons et al. (2009), and Nelson (2003). The option `normal.approx` in `predIntPois` coincides with the prediction region $\tilde{T}_1$ based on the normal approximation. In these earlier procedures, randomization was not utilized, hence generally conservative prediction intervals are obtained.

To compare the performances of the prediction regions $\tilde{T}_0$ (randomized version), $\tilde{T}_1$, $\tilde{T}_2$, $\tilde{T}_3$, $\tilde{T}_4$, and $\tilde{T}_5$ with $M = 50$, $S = 100$, and for $n \in \{5, 10, 15, 20, 30, 50, 70, 100\}$ and $\lambda \in \{1, 5, 15, 30, 50, 100, 200\}$, we ran simulation studies, using programs that we developed in the R programming environment (R Core Team, 2013), to determine the coverage probabilities and the lengths of the regions (recall that length is an equivalent surrogate for the cardinality of the regions since we took the ceiling and the floor of the lower and upper limits, respectively, for the intervals that leads to $\tilde{T}_1$ and $\tilde{T}_2$). For each combination of n and λ , 10000 simulation replications of the basic simulation experiment were performed. The basic simulation experiment is, for a fixed n and λ , to generate $T \sim POI(n\lambda)$ and $Y_0 \sim POI(\lambda)$. The T variable could be viewed as $T = \sum_{i=1}^n Y_i$ where $Y_1, Y_2, \dots, Y_n$ are IID from a $POI(\lambda)$. The prediction regions are then constructed based

on the observed (n, T) , with prediction coefficient of 95%. Note that since $\tilde{I}_5$ is the Bayes prediction region instead of the EB, we only needed the value of $T = \sum_{i=1}^n Y_i$, but if we also used the EB approach, then we would have needed the values of $(y_1, y_2, \dots, y_n)$ to estimate (κ, β) . After constructing the prediction regions, it is then determined if Y_0 is contained in these regions. Coverage percentage is the percentage out of the 10000 prediction regions that contain the Y_0 ; mean (standard deviation) length is the average (standard deviation) of the lengths of the 10000 prediction intervals. To avoid duplication of information, we omit the presentation of the simulation results in tabular form, except for Table 1 which shows the behavior of the standard deviation of the lengths of the prediction regions when $\lambda = 15$. Instead, we simply provide in Figure 3 the plots of the coverage probabilities (CP) and the mean lengths (ML) with respect to n for $\lambda \in \{1, 5, 30, 100\}$ of the prediction regions.

Table 1: Simulated coverage probabilities (CP), mean lengths (ML), and standard deviation of lengths (SL) of the prediction regions $\tilde{I}_0$ and $\tilde{I}_j, j = 1, 2, 3, 4, 5$ for the case $\lambda = 15$ and for different values of n . The number of replications for each combination of (n, λ) was 10000.

	n	Γ_0	Γ_1	Γ_2	Γ_3	Γ_4	Γ_5
CP	5	92.37	94.83	94.66	88.18	88.94	94.52
	10	93.65	94.81	94.59	91.88	92.18	94.68
	15	93.96	94.88	94.38	92.97	93.09	94.68
	20	94.25	94.96	94.56	93.63	93.75	94.98
	30	94.81	95.29	94.67	94.28	94.38	95.08
	50	94.83	95.22	94.80	94.58	94.57	95.11
	70	94.82	94.98	94.47	94.71	94.72	94.97
	100	94.88	95.18	94.54	94.82	94.68	95.03
ML	5	14.07	15.68	15.58	12.17	12.52	15.42
	10	14.07	14.86	14.91	13.21	13.32	14.80
	15	14.09	14.66	14.69	13.54	13.60	14.57
	20	14.10	14.56	14.54	13.69	13.73	14.46
	30	14.10	14.44	14.43	13.84	13.85	14.35
	50	14.10	14.32	14.31	13.95	13.95	14.25
	70	14.09	14.26	14.26	13.99	13.98	14.20
	100	14.10	14.22	14.21	14.01	14.02	14.17
SL	5	0.97	1.04	1.02	0.87	0.89	1.16
	10	0.75	0.75	0.78	0.71	0.72	0.78
	15	0.66	0.68	0.68	0.65	0.64	0.68
	20	0.60	0.61	0.61	0.59	0.59	0.61
	30	0.55	0.54	0.54	0.54	0.54	0.55
	50	0.47	0.48	0.48	0.48	0.48	0.49
	70	0.44	0.44	0.44	0.44	0.45	0.45
	100	0.42	0.41	0.41	0.41	0.42	0.43

Examining Figure 3 we observe that when $\lambda = 1$, the CP of $\tilde{I}_2$ is very poor and deteriorates *even* as n increases. The reason for this behavior is that the realized Y_0 tends to equal 0, but the square-root transformation has a tendency to shift to the right the prediction region, hence the region tends to miss Y_0 . This result for $\tilde{I}_2$ is similar to the result when the rate λ is known. Note further that when $\lambda = 1$, a small value, the normality approximation on which $\tilde{I}_1$ and $\tilde{I}_2$ are based will not be satisfactory even when n is large. When $n = 5$, $\tilde{I}_3$ and $\tilde{I}_4$ have unacceptably lower CPs compared to the nominal level, while $\tilde{I}_0$ also has CPs which are below the nominal level, as well as $\tilde{I}_1$ and $\tilde{I}_5$, though the last two regions have CPs closer to the desired level. The length of $\tilde{I}_0$ tends to be shorter than $\tilde{I}_1$ and $\tilde{I}_5$. As n increases, the CPs of $\tilde{I}_3$ and $\tilde{I}_4$ get closer to the desired level, and their lengths tend to be a bit shorter than $\tilde{I}_0$ and $\tilde{I}_1$. When $\lambda = 5$, the CPs of $\tilde{I}_0, \tilde{I}_2, \tilde{I}_3$, and $\tilde{I}_4$ are

all below the nominal level, whereas for $\tilde{I}_1$ and $\tilde{I}_5$, their CPs exceed or are quite close to the nominal level, except when $n = 5$. As a consequence, they ended up having longer mean lengths. These behaviors continue to hold as λ was increased, but with the CPs getting closer to the nominal level, especially as n increases. When n is small, the CPs of $\tilde{I}_3$ and $\tilde{I}_4$ are still appreciably lower than the nominal level. When λ is large, $\tilde{I}_1$, $\tilde{I}_2$, and $\tilde{I}_5$ almost have the same performance.

One may wonder why the mean lengths of the prediction regions are not generally decreasing with respect to n . For instance, the mean length for $\tilde{I}_0$ remained flat as n increased. The reason for this is that the increase

Figure 3: Simulated coverage probabilities (CP) and mean lengths (ML) of the prediction regions $\tilde{I}_0$ and $\tilde{I}_j$, $j = 1, 2, 3, 4, 5$, plotted with respect to n for values of $\lambda \in \{1, 5, 30, 100\}$. The number of replications for each combination of (n, λ) was 10000.

Figure 3: ...continued

in n only improves the estimation of the parameter λ in terms of decreasing the standard error of the estimator. However, the prediction region being constructed is for Y_0 , which has a distribution unaffected by n . This distribution is the main determinant of the length of the prediction region. Instead, the impact of the increase in n is in the variability of the prediction region length, since when n increases then the estimator of λ becomes more precise. However, as n increases, the mean length will converge to a non-zero value determined by the value of λ . The simulation results attest to this phenomenon that the standard deviation of the length is decreasing with increasing n . To see this concretely, Table 1 presents the simulated coverage probabilities and the means and standard deviations of the lengths of the prediction regions for $\lambda = 15$ and for different values

of n . Observe that the standard deviation of the lengths decreases with increasing n , while the mean length is almost constant in n and approaches a non-zero value.

Thus, summing up our observations from these simulation studies for this no-covariate or intercept-only Poisson model, in terms of adapting to the estimation of the unknown rate λ , $\check{I}_1$ and $\check{I}_5$ possess the best performance among these six prediction regions in terms of achieving the nominal level, but they also tend to be longer than the others. Again, as in the case where the λ is known, $\check{I}_1$ is preferable than $\check{I}_2$.

3 Prediction Regions for Poisson Regression

3.1 With Known Parameter Vector

Next, we consider the situation where there is a $1 \times (p + 1)$ covariate row vector $\mathbf{x} = (x_0, x_1, \dots, x_p)$ that could affect the rate parameter, where in our implementation we take $x_0 = 1$, so the model has an intercept component. Thus, we suppose that the rate parameter is $\lambda(\mathbf{x}; \theta)$, where $\theta = (\theta_0, \theta_1, \dots, \theta_p)^T$ is a $(p + 1) \times 1$ column vector of parameters. We further assume that there exists a non-negative twice continuously-differentiable function $\rho(\cdot)$, called the inverse link function, such that

$$\lambda(\mathbf{x}; \theta) = \rho(\mathbf{x}\theta) = \rho\left(\sum_{j=0}^p x_j \theta_j\right). \quad (18)$$

This is the so-called Poisson regression model and belongs to the class of generalized linear models or the class of non-linear models (see Rencher and Schaalje, 2008; Casella and Berger, 1990). If Y_0 , given $\mathbf{x} = \mathbf{x}_0$, has a Poisson distribution with rate $\lambda(\mathbf{x}_0; \theta)$, and θ is *known*, then we could construct prediction regions for Y_0 according to the methods described in the first part of Section 2 when λ was assumed known.

3.2 With Unknown Parameter Vector

When θ is not known, then we need to estimate it. Let us therefore assume that we are able to observe the sample $\{(Y_i, \mathbf{x}_i), i = 1, 2, \dots, n\}$ with $Y_i|\mathbf{x}_i \sim \text{POI}(\lambda(\mathbf{x}_i; \theta))$ and with the Y_i s independent and the $\mathbf{x}_i$ s fixed. We seek to construct a prediction region for Y_0 given the covariate vector $\mathbf{x}_0$. First, we introduce the following functions:

$$\begin{aligned} \dot{\rho}(w) &= \frac{d}{dw}\rho(w) \quad \text{and} \quad \ddot{\rho}(w) = \frac{d^2}{dw^2}\rho(w); \\ \psi(w) &= \frac{d}{dw} \log \rho(w) = \frac{\dot{\rho}(w)}{\rho(w)} \quad \text{and} \quad \Psi(w) = \frac{d^2}{dw^2} \log \rho(w) = \frac{\ddot{\rho}(w)}{\rho(w)} - \psi(w)^2. \end{aligned}$$

The log-likelihood function for θ , given $\{(y_i, \mathbf{x}_i), i = 1, 2, \dots, n\}$, is $l(\theta) = -\sum_{i=1}^n \rho(\mathbf{x}_i\theta) + \sum_{i=1}^n y_i \log \rho(\mathbf{x}_i\theta)$. The associated score vector function is

$$\mathbf{U}(\theta) = \nabla_{\theta} l(\theta) = \sum_{i=1}^n [y_i - \rho(\mathbf{x}_i\theta)] \mathbf{x}_i^T \psi(\mathbf{x}_i\theta); \quad (19)$$

whereas, the *observed* Fisher information matrix is, with $\mathbf{x}^{\otimes 2} = \mathbf{x}^T \mathbf{x}$,

$$\mathbf{I}(\theta) = -\nabla_{\theta^T} \nabla_{\theta} l(\theta) = \sum_{i=1}^n \mathbf{x}_i^{\otimes 2} \ddot{\rho}(\mathbf{x}_i\theta) - \sum_{i=1}^n y_i \mathbf{x}_i^{\otimes 2} \Psi(\mathbf{x}_i\theta). \quad (20)$$

Thus, the *expected* Fisher information matrix is

$$\mathcal{I}(\theta) = \sum_{i=1}^n \mathbf{x}_i^{\otimes 2} [\psi(\mathbf{x}_i\theta)]^2 \rho(\mathbf{x}_i\theta). \quad (21)$$

The MLE of θ based on $\{(Y_i, \mathbf{x}_i), i = 1, 2, \dots, n\}$, denoted by $\hat{\theta}$, solves the equation $\mathbf{U}(\theta) = \mathbf{0}$. This will usually be obtained through iterative procedures, such as the Newton-Raphson method, with the iteration given by

$$\theta \leftarrow \theta + [\mathbf{I}(\theta)]^{-1} \mathbf{U}(\theta). \quad (22)$$

By the large-sample theory of ML estimation (cf. Casella and Berger, 1990; van der Vaart, 1998), as $n \rightarrow \infty$ and under regularity conditions on the sequence of covariate vectors $\mathbf{x}_i, i = 1, 2, \dots, n$, we have that $\hat{\theta} \sim AN[\theta, \mathcal{I}(\theta)^{-1}]$. A regularity condition is that $\|\frac{1}{n}\mathcal{I}(\theta) - \frac{1}{n}\mathcal{I}(\hat{\theta})\| \rightarrow 0$ as $n \rightarrow \infty$, so $\hat{\theta} \sim AN[\theta, \mathbf{I}(\theta)^{-1}]$. By the Delta-Method, it then follows that the ML estimator of $\lambda(\mathbf{x}_0; \theta)$ satisfies, as $n \rightarrow \infty$,

$$\widehat{\lambda(\mathbf{x}_0)} = \lambda(\mathbf{x}_0; \hat{\theta}) = \rho(\mathbf{x}_0 \hat{\theta}) \sim AN\left(\rho(\mathbf{x}_0 \theta), [\psi(\mathbf{x}_0 \theta)]^2 [\rho(\mathbf{x}_0 \theta)]^2 \text{tr}[\mathcal{I}(\theta)^{-1} \mathbf{x}_0^{\otimes 2}]\right),$$

where ‘tr’ means trace of a matrix. Using this result and when $\lambda_0 = \rho(\mathbf{x}_0 \theta)$ is large, we obtain the approximate distributions of relevant pivotal quantities for constructing prediction regions. We write $\hat{\lambda}_0$ for $\lambda(\mathbf{x}_0; \hat{\theta})$ and $\hat{\psi}_0$ for $\psi(\mathbf{x}_0 \hat{\theta})$. These pivotal quantities are:

$$\begin{aligned} \frac{Y_0 - \hat{\lambda}_0}{\sqrt{\hat{\lambda}_0}} &\sim N\left[0, \hat{V} \equiv 1 + \hat{\psi}_0^2 \hat{\lambda}_0 \text{tr}([\mathbf{I}(\hat{\theta})]^{-1} \mathbf{x}_0^{\otimes 2})\right]; \\ \frac{\sqrt{Y_0} - \sqrt{\hat{\lambda}_0}}{\sqrt{1/4}} &\sim N\left[0, \hat{V} \equiv 1 + \hat{\psi}_0^2 \hat{\lambda}_0 \text{tr}([\mathbf{I}(\hat{\theta})]^{-1} \mathbf{x}_0^{\otimes 2})\right]. \end{aligned}$$

Writing $(\mathbf{Y}, \mathbf{X}) = \{(Y_i, \mathbf{x}_i), i = 1, 2, \dots, n\}$, from these pivotal quantities, we are then able to obtain approximate prediction regions for Y_0 given by:

$$\tilde{I}_1(\mathbf{x}_0, (\mathbf{Y}, \mathbf{X}); \alpha) = \left[0 \vee \left(\hat{\lambda}_0 - z_{\alpha/2} \sqrt{\hat{\lambda}_0} \sqrt{\hat{V}}\right), \hat{\lambda}_0 + z_{\alpha/2} \sqrt{\hat{\lambda}_0} \sqrt{\hat{V}}\right] \cap \mathbb{Z}_{0,+}; \quad (23)$$

$$\tilde{I}_2(\mathbf{x}_0, (\mathbf{Y}, \mathbf{X}); \alpha) = \left[\left\{0 \vee \left(\sqrt{\hat{\lambda}_0} - z_{\alpha/2} \left(\frac{1}{2}\right) \sqrt{\hat{V}}\right)\right\}^2, \left\{\sqrt{\hat{\lambda}_0} + z_{\alpha/2} \left(\frac{1}{2}\right) \sqrt{\hat{V}}\right\}^2\right] \cap \mathbb{Z}_{0,+}. \quad (24)$$

Under the given setting, $\tilde{I}_1$ can be considered as a special case of prediction regions developed under the GLM class; see, for instance, Myers and Montgomery (1997). We could also have a prediction region based on Γ_0 from Section 2 given by

$$\tilde{I}_0(U, \mathbf{x}_0, (\mathbf{Y}, \mathbf{X}); \alpha) = \Gamma_0(U; \hat{\lambda}_0, \alpha), \quad (25)$$

where we note that the dependence on $\mathbf{x}_0$ and $(\mathbf{Y}, \mathbf{X})$ is through $\hat{\lambda}_0$. Observe, however, that we are simply plugging in the estimate of $\lambda(\mathbf{x}_0; \theta)$, but *without* taking into consideration the variability inherent in the estimator $\lambda(\mathbf{x}_0; \hat{\theta})$.

A specific inverse link function $\rho(\cdot)$, which we will consider in the application to forecasting deaths in the US due to COVID-19, is the exponential function $\rho(w) = \exp(w)$, so that

$$\rho(w) = \dot{\rho}(w) = \ddot{\rho}(w) = \exp(w); \quad \psi(w) = 1; \quad \text{and} \quad \Psi(w) = 0.$$

For this special inverse link function, we obtain the simplifications of the score vector and information matrix functions given by

$$\mathbf{U}(\theta) = \sum_{i=1}^n \mathbf{x}_i^T [y_i - \rho(\mathbf{x}_i \theta)] \quad \text{and} \quad \mathbf{I}(\theta) = \mathcal{I}(\theta) = \sum_{i=1}^n \mathbf{x}_i^{\otimes 2} \rho(\mathbf{x}_i \theta). \quad (26)$$

We also mention the extension of the Bayesian/EB approaches to constructing prediction regions in the regression setting. We suppose that the parameter θ in $\lambda(\mathbf{x}; \theta) = \rho(\mathbf{x} \theta)$ takes values in a parameter space Θ . The approach then proceeds by starting with a prior distribution $\Pi(\cdot)$ on Θ which quantifies our

prior knowledge about θ . The posterior predictive distribution of Y_0 , the response at $\mathbf{x}_0$, given the data $(\mathbf{Y}, \mathbf{X}) = \{(y_i, \mathbf{x}_i), i = 1, 2, \dots, n\}$, is given by

$$p(y_0|\mathbf{x}_0, (\mathbf{Y}, \mathbf{X})) = \frac{1}{y_0!} \frac{H[(\mathbf{Y}_0, \mathbf{X}_0)]}{H[(\mathbf{Y}, \mathbf{X})]} \quad (27)$$

where $(\mathbf{Y}_0, \mathbf{X}_0) = \{(Y_i, \mathbf{x}_i), i = 0, 1, 2, \dots, n\}$ and

$$H[(\mathbf{Y}, \mathbf{X})] = \int_{\Theta} \exp \left\{ - \sum \rho(\mathbf{x}_i \theta) \right\} \left[\prod \rho(\mathbf{x}_i \theta)^{y_i} \right] \Pi(d\theta),$$

where the sum and the product are taken over the index set associated with $(\mathbf{Y}, \mathbf{X})$, so it will be over $\{1, 2, \dots, n\}$ for $(\mathbf{Y}, \mathbf{X})$, and $\{0, 1, 2, \dots, n\}$ for $(\mathbf{Y}_0, \mathbf{X}_0)$. Generally, there will be no family of conjugate prior distributions on θ with respect to the Poisson regression model, so the function H will not be in a closed analytical form, hence it has to be computed numerically, for instance, using Markov Chain Monte Carlo (MCMC) algorithms. Nevertheless, upon obtaining the posterior predictive distribution of Y_0 given in (27), a prediction region is then obtained by using this PMF $p(y_0|\mathbf{x}_0, (\mathbf{Y}, \mathbf{X}))$ in lieu of the Poisson PMF in (4), analogously to the development of the prediction region $\tilde{I}_5$ for the intercept-only model. The prior distribution Π will involve hyper-parameters, for example, if $\theta = \mathbb{R}_{p+1}$, Π could be specified to be a multivariate normal distribution with mean vector μ and covariance matrix Σ , so (μ, Σ) will be the hyper-parameters. A Bayesian will assign values to these hyper-parameters, unless an improper prior distribution (e.g., Lebesgue measure), which need not involve unknown hyper-parameters, is adopted; whereas, an empirical Bayesian will estimate these hyper-parameters using the data $(\mathbf{Y}, \mathbf{X})$. Because of the need to approximate the posterior predictive PMF through numerical methods, these Bayesian and EB approaches to constructing a prediction region for Y_0 are clearly computationally-intensive, especially if used in a simulation study to investigate their properties, such as their coverage probabilities and their lengths. Because of the need to specify a non-conjugate prior and its hyper-parameters and the need for intensive computations, these Bayesian and EB procedures are not included in our illustrations, simulations, and applications. We mention in passing that there are also previous works on the construction of prediction regions under Poisson regression such as in Wood (2005), which studied a one-sided prediction region based on Chebyshev's inequality; in Olive et al. (2021), which introduced a procedure based on a nonparametric approach combined with bootstrapping; and in Davison and Hinkley (1997), which contains a chapter discussing a bootstrapping approach.

3.3 Illustrations of Prediction Regions

We demonstrate these prediction regions, depicted as intervals in the plots, via the following experiment. We specify a sample size n and an order p . We then generate IID realizations $w_i, i = 1, 2, \dots, n, n+1$, from either a $N(\mu, \sigma^2)$ distribution or a standard uniform distribution, and form the covariate vectors $\mathbf{x}_i = (1, w_i, w_i^2, \dots, w_i^p), i = 1, 2, \dots, n, n+1$. For a specified $\theta = (\theta_0, \theta_1, \dots, \theta_p)^T$, the Poisson rates

$$\lambda_i = \lambda(\mathbf{x}_i; \theta) = \rho(\mathbf{x}_i \theta) = \exp\{\mathbf{x}_i \theta\}, i = 1, 2, \dots, n, n+1.$$

are computed. The i th response y_i is a realization of a random draw from a $POI(\lambda_i)$. The response vector is $\mathbf{y} = (y_1, y_2, \dots, y_n, y_{n+1})^T$. The goal is to construct a prediction region for $Y_0 \equiv Y_{n+1}$, given the data $\{(y_i, \mathbf{x}_i), i = 1, 2, \dots, n\}$ and $\mathbf{x}_0 \equiv \mathbf{x}_{n+1}$. We construct the 95% prediction regions $\tilde{I}_0(\mathbf{x}_0, (\mathbf{Y}, \mathbf{X}); .05)$, $\tilde{I}_1(\mathbf{x}_0, (\mathbf{Y}, \mathbf{X}); .05)$, and $\tilde{I}_2(\mathbf{x}_0, (\mathbf{Y}, \mathbf{X}); .05)$ given in (25), (23), and (24), respectively. The procedures were coded into R functions, with these programs available upon request from the authors. We present the results pictorially via a scatterplot of $\{(w_i, y_i), i = 1, 2, \dots, n, n+1\}$. The realized value $y_0 \equiv y_{n+1}$ of Y_0 is highlighted and the three prediction intervals for Y_0 are also plotted. Included in the plot is the theoretical curve for the $\lambda(\mathbf{x}; \theta)$ as a function of w and we also super-impose the fitted curve. Prediction regions for Y_i at $w_i, i = 1, 2, \dots, n$, are also depicted in the plot. Figure 4 provides four plots under the following model settings.

- (i) Case 1 Model: $n = 30, p = 1$, with $W_i \sim U[0, 1], i = 1, 2, \dots, 31$, and with $\theta = (2, 3)^T$.

- (ii) Case 2 Model: $n = 100$, $p = 2$, with $W_i \sim N(\mu = 2, \sigma = 2)$, $i = 1, 2, \dots, 101$, and with $\theta = (.3, -.2, .05)$.
- (iii) Case 3 Model: $n = 200$, $p = 3$, with $W_i \sim N(\mu = 1, \sigma = 2)$, and with $\theta = (3, .2, -.1, -.05)$.
- (iv) Case 4 Model: $n = 100$, $p = 5$, with $W_i \sim U[0, 1]$, $i = 1, 2, \dots, 101$, and with $\theta = (3, -1, 3, -2, 1, -.5)$.

The top left plot shows a realization from the Case 1 model. In this case, $\lambda_i = \exp\{2 + 3w_i\}$, $i = 1, 2, \dots, 31$, which is a sharply and exponentially increasing function of w_i . The Y_{31} being predicted was at $w_{31} = .8139$ and the realized value was $y_{31} = 1222$, which ended up being contained in all three realized prediction regions, which in interval form are $\tilde{T}_0 = [1091, 1223]$, $\tilde{T}_1 = [1087, 1226]$, and $\tilde{T}_2 = [1088, 1227]$. The top right plot is from the Case 2 model. The value being predicted is at $w_{101} = 5.0622$ and the realized value of Y_{101} was $y_{101} = 1$. The three prediction regions were $\tilde{T}_0 = [0, 3]$, $\tilde{T}_1 = [0, 3]$, and $\tilde{T}_2 = [1, 4]$, which all contained y_{101} . Two things to observe from this plot are (1) the prediction region $\tilde{T}_2$ was shifted to the right relative to $\tilde{T}_1$, and (2) the prediction regions $\tilde{T}_0$ with respect to the w -values are scissor-like or jagged. The latter is a consequence of the randomization approach in the construction of the prediction regions and this

Figure 4: Plot panels are realizations from the four model specifications: *top left* is from Case 1; *top right* is from Case 2; *bottom left* is from Case 3; and *bottom right* is from Case 4.

non-smooth behavior becomes more apparent since the realized values of the Y_i s are small. The bottom left plot is from the Case 3 model. The rate curve $\lambda(w)$ as a function of w goes to zero as w increases, but goes to ∞ as w decreases, with a local minimum and maximum close to $w = -2$ and $w = 1$, respectively. The target of the prediction regions was Y_{201} which took value $y_{201} = 23$ at $w_{201} = -3.2509$. The realized prediction regions were $\tilde{I}_0 = [10, 25]$, $\tilde{I}_1 = [9, 26]$, and $\tilde{I}_2 = [10, 27]$, so all regions contained $y_{201} = 23$. Once again observe that $\tilde{I}_2$ tends to be shifted to the right relative to $\tilde{I}_1$. Finally, the bottom right plot is from the Case 4 model. The rate curve is not highly varying compared to that in the preceding model since w is restricted on $[0, 1]$. The target is the value of Y_{101} , which turned out to be $y_{101} = 29$ at $w_{101} = .8944$. This realized value of Y_{101} was contained in the realized prediction regions $\tilde{I}_0 = [21, 42]$, $\tilde{I}_1 = [20, 42]$, and $\tilde{I}_2 = [22, 43]$.

3.4 Simulation Studies

In each of these illustrative realizations, the three prediction regions did not vary much from each other in terms of their lengths, except in the first case where $\tilde{I}_0$ was shorter but barely covered the value being predicted. A question that now arises is how their coverage probabilities and their mean lengths compare with each other? To gain some insights into these comparisons, we performed simulation studies under the four different models described above, with each simulation run having 10000 replications. The sample sizes considered were $n \in \{30, 50, 100, 200\}$. The results are summarized in Table 2 which presents the coverage probabilities (CP), mean lengths (ML), and standard deviation of lengths (SL).

Examining Table 2, it appears that $\tilde{I}_0$ has coverage probabilities that are below the nominal level (between 3% and 4% below for Case 4 when $n = 30$), with the discrepancy being more pronounced when the sample size is small. As the sample size is increased, these observed coverage probabilities get closer to the nominal level. This deficiency is due to the estimation of the θ parameter and, as previously noted, the $\tilde{I}_0$ does not take into consideration the variability in the resulting estimator of $\lambda(\mathbf{x}_0; \theta)$. On the other hand, $\tilde{I}_1$ and $\tilde{I}_2$ both achieve coverage probabilities that are quite close to the nominal level, especially $\tilde{I}_1$, when n is large. $\tilde{I}_0$, on the other hand, tends to have a lower mean length compared to the mean lengths of $\tilde{I}_1$ and $\tilde{I}_2$, with the

Figure 5: An extreme realization from the third model with $p = 3$, $n = 30$, $\theta = (3, .2, -.1, -.05)$, and with $W \sim N(\mu = 1, \sigma = 2)$. The resulting prediction regions were $\tilde{I}_0 = [257, 323]$, $\tilde{I}_1 = [0, 1100]$, $\tilde{I}_2 = [0, 1665]$. The realized value of Y_{31} was $y_{31} = 47$ associated with $w_{31} = -4.0493$.

Table 2: Simulation results for the coverage probability (CP), mean length (ML), and standard deviation of length (SL) for the four model cases in the illustrations for three prediction regions and different sample sizes n . The number of replications was 10,000.

Case 1 Model					Case 2 Model				
	n	Γ_0	Γ_1	Γ_2		n	Γ_0	Γ_1	Γ_2
CP	30	94.36	95.26	94.98	CP	30	93.10	95.27	95.01
	50	94.36	94.99	94.78		50	94.13	94.99	94.79
	100	94.95	95.24	95.15		100	94.10	94.68	94.41
	200	94.82	94.92	94.84		200	94.52	94.76	94.59
ML	30	77.92	82.29	82.28	ML	30	16.78	19.18	19.21
	50	77.77	80.29	80.29		50	16.69	17.86	17.87
	100	76.94	78.15	78.15		100	16.73	17.30	17.30
	200	78.74	79.37	79.37		200	16.72	17.01	17.01
SL	30	54.04	60.53	60.52	SL	30	3.66	21.32	23.37
	50	54.40	58.01	58.01		50	3.30	11.27	11.62
	100	53.84	55.50	55.50		100	3.32	6.08	6.08
	200	54.39	55.22	55.22		200	3.12	4.05	4.05

Case 3 Model					Case 4 Model				
	n	Γ_0	Γ_1	Γ_2		n	Γ_0	Γ_1	Γ_2
CP	30	92.33	94.74	94.44	CP	30	91.32	94.96	94.81
	50	92.67	94.30	93.48		50	92.96	94.99	94.75
	100	94.13	95.08	94.22		100	94.36	95.13	94.99
	200	94.91	95.51	94.36		200	94.71	95.08	94.84
ML	30	13.62	2011.21	5679.31	ML	30	17.68	21.43	22.93
	50	13.28	150.32	237.33		50	17.65	19.11	19.12
	100	13.44	617.14	1010.62		100	17.65	18.35	18.35
	200	13.07	47.16	47.30		200	17.63	17.99	17.99
SL	30	57.83	197118.78	561658.43	SL	30	2.50	35.46	124.53
	50	22.62	7945.90	12852.71		50	2.12	3.70	3.77
	100	46.33	58624.75	96369.77		100	2.02	2.39	2.38
	200	17.05	3327.74	3327.77		200	1.97	2.08	2.08

differences in mean lengths becoming alarmingly large for the Case 3 model. Recall that for this model, the rate curve increases to ∞ as w decreases to $-\infty$, and since the w_i 's are generated from a normal distribution, on some occasions, w_{n+1} falls outside the range of $\{w_i, i = 1, 2, \dots, n\}$. Depending on how different w_{n+1} is from the mean of $w_1, w_2, \dots, w_n$, this could lead to a large estimate of the standard error of $\hat{\lambda}_{n+1}$, thus leading to very wide prediction regions for $\tilde{T}_1$ and $\tilde{T}_2$. Since $\tilde{T}_0$ simply utilized the estimate of $\hat{\lambda}_{n+1}$, but was totally oblivious to its variability, it was not much affected in such a situation. However, because of its rigidity with respect to this added variability, it could dramatically suffer. We demonstrate this situation by plotting an extreme realization in Figure 5 from the third model with $n = 30$ and with the w_{31} outside and far to the left of the values of $w_1, \dots, w_{30}$.

Observe here that the three prediction regions $\tilde{T}_0 = [257, 323]$, $\tilde{T}_1 = [0, 1100]$, and $\tilde{T}_2 = [0, 1665]$ are very different in terms of their lengths (66, 1100, and 1665, respectively). But in this case $\tilde{T}_0$ did *not* cover, by a wide margin, the realized value of Y_{31} , which was $y_{31} = 49$ associated with $w_{31} = -4.0493$. This particular demonstration warns us of the danger and pitfalls of making a prediction for a response variable that is associated with a covariate vector outside the convex hull of the covariate vectors used in the construction of the prediction regions and when the Poisson rate hyper-surface generated by the map $\mathbf{x} \mapsto \rho(\mathbf{x}\theta)$ is of a complicated form. As a word of caution, when performing extrapolation to do predictions, be forewarned of sinkholes littering the forecasting road — and, if it could be avoided, make no forecasts on long, especially

very long, horizons. But, unfortunately, this is the type of forecasting problem that is actually *realistic* and of most interest, such as that of predicting the number of cases or deaths due to COVID-19 at a future date, given the observed data up to a certain date. Based on the results of these simulation studies, the prediction region $\tilde{I}_1$ appears to be the most preferable among the three prediction regions. For our illustration using the COVID-19 data set in Section 4, we will therefore only present the prediction region provided by $\tilde{I}_1$.

4 COVID-19 Pandemic Application

We now present in this section an illustration of the potential application of the procedures discussed in the preceding sections. One of the interesting questions during this COVID-19 pandemic is the forecasting of the number of cumulative deaths in the US at a given date, for example, at the end of May 31, 2020, given information up to a certain date, say May 15, 2020. Such forecasts are of critical importance since they could partly be the basis of highly consequential and possibly controversial decisions by federal, state, and local government officials, school and university administrators, executives of big corporations and small businesses, religious leaders, and many others. Such decisions could pertain to instituting stay-in-place directives, when to issue social distancing or social easing guidelines, when to open business establishments, when to open public places such as shopping malls and ocean beaches, when to allow religious gatherings, when to ramp-up vaccinations if vaccines are available, etc. Data for daily deaths and cumulative deaths, as well as cases, for different countries and states due to the COVID-19 are available from either the US Center for Disease Control and Prevention <https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html>, Johns Hopkins University <https://coronavirus.jhu.edu/data>, or the European CDC <https://www.ecdc.europa.eu/en/publications-data/> (ECDC, 2020b).

Clearly, the sequence of cumulative deaths does not satisfy the independence assumption, so a non-homogeneous Poisson process model (see Resnick, 1992) is not an appropriate model for cumulative deaths when viewed as a continuous-time stochastic process. However, a non-homogeneous Poisson process could plausibly model the number of occurrences of deaths in continuous-time, from which it follows that the sequence of daily deaths will be independently Poisson distributed with possibly different rates depending on the number of days from the time origin and the specific day of the week, as well as other features such as, for example, the quality of the health care facilities, which is hard to quantify and not available in the European CDC data set. Our novel idea therefore is to utilize Poisson regression to predict the number of daily deaths according to the methods developed earlier, and then to aggregate these daily forecasts to obtain forecasts of the cumulative deaths.

We used the data set for the US provided by the European CDC (ECDC, 2020b) plotted in Figure 1 which are the observed numbers of daily deaths attributed to COVID-19 starting on March 1, 2020, the day after the first reported death due to COVID-19, until May 15, 2020. Note that, technically, this will be the deaths data at the end of May 14, 2020. Using this data set on May 15th, and given the cumulative number of deaths until then, the goal is to forecast the cumulative number of deaths in the US due to COVID-19 by the end of May 31, 2020, that is, June 1, 2020. We limit our illustration to simply utilizing the variable *DayNum*, which is the number of days starting from December 31, 2019; *Day* which is the day of the week; and *Deaths*, which is the variable representing the daily number of deaths. We surmise that the deaths data set is the most reliable among the data sets that were compiled, for instance, compared to the data set pertaining to the number of cases or infected people. However, the deaths data set need not also be totally reliable and could be subject to misclassification error and competing causes of deaths. For example, a patient who contracted COVID-19 who dies primarily because of pneumonia may be classified as having died of COVID-19, but could also be classified as having died, not of COVID-19, but of pneumonia. See also, for instance, the April 16, 2020 *Wall Street Journal* article (WSJ, 2020), the April 16, 2020 *The Economist* article (ECO, 2020), and the June 18, 2020 *BBC News* article (BBC, 2020a), the last two discussing the notion of “excess deaths,” which are deaths that may have been due to the pandemic, but which are not included in the reported COVID-19 deaths data set. Certainly, we could have used other information such as the number of reported cases; by performing separate

forecasts in each of the 50 states and the District of Columbia, then aggregating; or even by utilizing counties or metropolitan cities as strata, and then combining forecasts from these strata to obtain an overall forecast for the whole US. However, for illustrative purposes, we decided to keep things simple. The important point to note is that constructing prediction regions via these more elaborate set-ups will typically build on the approach developed for the simpler setting discussed above.

From March 1 (DayNum = 62) to May 15 (DayNum = 137), consisting of 76 days, we have the daily number of deaths, hence also the cumulative number of deaths. The other variable used in the modeling is Day (e.g., Sunday, Monday, etc.) associated with each value of DayNum, which is a categorical or factor variable. We fitted a Poisson regression model using the `glm` function in R with a log-link, with response being $Y = \text{Deaths}$ and covariate vector $\mathbf{X} = (1, W, W^2, W^3, W^4, W^5, \text{Day})$, where $W = \text{DayNum}$, and Day is considered as a factor variable hence is converted into six, instead of seven since we already have an intercept term, dummy variables in the design matrix. We chose this 5th-order model with respect to DayNum since the Akaike Information Criterion (AIC) values, computed under the Poisson regression model, appear to stabilize starting at this model and adhering to the Law of Parsimony (Occam's Razor). The AIC values associated with the 8th- and 9th-order models were actually smaller than for the 5th-order model; however, these models possess highly unstable predicted values, possibly partly due to insufficient sample size. To automate the choice of the model order, an *ad hoc* procedure for choosing the order is to specify an $\epsilon > 0$ and a maximum model order $p_{\max}$, and then to choose the smallest order less than or equal to $p_{\max}$ satisfying

$$\frac{\text{AIC}(k) - \min_{j \leq p_{\max}} \text{AIC}(j)}{\min_{j \leq p_{\max}} \text{AIC}(j)} \leq \epsilon.$$

The choice of $p = 5$ in the application above arose, for instance, by setting $\epsilon = .05$ and $p_{\max} = 9$. Table 3 summarizes the AIC values for the different models, both without and with Day in the model. It also contains the estimates of ξ , the over-dispersion parameter in a model that will be introduced shortly, and as we will then see, larger values of ξ are indicative of the Poisson regression model becoming a more adequate model.

Table 3: AIC values for the fitted models under different order (p) for DayNum. The second column does not include Day in the model, whereas the third column includes Day in the model as a categorical variable. The fourth column is the estimate of ξ without Day, while the fifth column is the estimate of ξ with Day. ξ is the additional parameter in the over-dispersed model.

p	AIC_{ND}	AIC_D	$\hat{\xi}_{ND}$	$\hat{\xi}_D$
1	43755.86	40750.61	2.66	2.99
2	12289.82	8504.21	7.26	10.96
3	8551.98	5117.39	9.33	16.25
4	8486.66	5098.87	9.40	16.27
5	8372.16	4889.52	9.47	16.89
6	8374.09	4875.94	9.46	17.14
7	8374.27	4867.00	9.47	17.20
8	8216.72	4797.71	9.73	17.80
9	7917.52	4600.89	10.08	18.79

The histogram and time plot of the residuals with respect to DayNum are provided in Figure 6. Recall that the i th residual in Poisson regression is defined as $R_i = (Y_i - \hat{\lambda}_i) / \sqrt{\hat{\lambda}_i}$, $i = 1, 2, \dots, n$, where $\hat{\lambda}_i$ is the fitted value associated with $\mathbf{x}_i$. As such, if the Poisson regression model is adequate, one should see a histogram similar to that associated with a centered (at zero) Poisson distribution with unit rate, but this will just be an approximation since the rates are estimated, hence that affects the distribution of the residuals. Similarly, the time plot of the residuals should be randomly distributed on the zero horizontal line. The histogram and time plot in Figure 6, with the time plot also showing a *lowess* fit, do not appear to contradict expected behaviors under the Poisson regression model assumptions of Poissonness, independence, and an adequate rate function. In the time plot, observe the outliers at DayNum values of 64 and 81, with mild outliers at DayNum

of 70, 71, 82, 84, and 108. The outlying values at DayNum 64 and 81 could be a consequence of the uncertainty on who to count as COVID-19 deaths since they occurred in the early days of recording COVID-19 deaths, while on Daynum 81, which was March 20th, there were zero reported COVID-19 deaths, highly suspect since two days before and two days after this day there were 23, 42, 110, and 80 COVID-19 reported deaths. The outlier on DayNum 108, corresponding to April 16th, was due to a one-time adjustment of 3778 that was made because of a change in criteria on what are considered as COVID-19 deaths (see the April 15, 2020 *BBC News* article (BBC, 2020b) and WSJ (2020)). The two days prior to April 16th, there were 1541 and 2408 deaths reported, while the three days after April 16th, there were 2299, 3770, and 1856 deaths reported. Thus, the reported number on April 16th of 4928, the highest number of daily deaths reported, is an outlier which could be explained by the adjustment that was made. However, we still included these perceived outliers in the fitting of the fifth-order, with respect to DayNum, Poisson regression model and the over-dispersed Poisson regression model.

Figure 6: Histogram and Time Plot of the Residuals from the Fitted Fifth-Order Poisson Regression Model of Deaths with respect to DayNum.

Based on the fitted model's residuals, we further assessed the independence assumption of the daily deaths. We do this by creating a contingency table for DayNum with six intervals and with Residuals being either negative or positive and then performing a test for independence. The observed contingency table is presented in Table 4. A chi-square test for independence based on this table yielded $\chi^2_c = 4.4685$ on 5 degrees-of-freedom, with associated p -value of 0.4841, hence the null hypothesis of independence cannot be rejected. Observe, however, that the fit of the model in the early days is not satisfactory, and between DayNum 99 to 112, there was a preponderance of negative residuals, possibly owing to the influence of the adjusted reported daily deaths on DayNum 108.

It may appear surprising and counter intuitive to include a Day effect in the model, since one could argue that when a patient is dying, he/she does not really have any control or choice of which day he/she should die, so that deaths due to COVID-19 should be uniformly distributed over the days of a week. However, the data pertains to *reported* daily deaths, and so the number of daily deaths could be affected by reporting delays due possibly to limited health or hospital personnel during certain days of the week. The fitted model based on data up to May 15th did demonstrate that there is a Day effect. Based on the estimates of the regression

Table 4: Contingency Table of DayNum and Residuals from the Fifth-Order Poisson Regression Model.

DayNum Interval	# Residuals ≤ 0	# Residuals > 0
$61 < \text{DayNum} \leq 74$	7	6
$74 < \text{DayNum} \leq 87$	6	7
$87 < \text{DayNum} \leq 99$	5	7
$99 < \text{DayNum} \leq 112$	10	3
$112 < \text{DayNum} \leq 125$	7	6
$125 < \text{DayNum} \leq 137$	5	7

coefficients, Sunday has the lowest Day effect, followed by Monday, then Saturday, then Friday, then almost equally by Tuesday and Thursday, and finally Wednesday, which has the largest effect.

Recall that our main objective is to obtain a prediction region for the cumulative number of deaths at a specified date, in our case June 1, 2020 (DayNum = 154). This means we were predicting the cumulative deaths at the end of May 31, 2020. From DayNum = 138 to DayNum = 154, there were a total of 17 days. If we denote by $Y_j, j = 62, 63, \dots, 154$, the random variable denoting the daily number of deaths for DayNum = j , the random variable denoting the cumulative number of deaths until DayNum = k is $S_k = \sum_{j=62}^k Y_j$. Thus, we are seeking a prediction region for S_{154} , given that $S_{137} = 85906$. Under the fitted Poisson regression model, Y_j , given $\mathbf{x}_j = (1, w_j, w_j^2, w_j^3, w_j^4, w_j^5, d_{1j}, d_{2j}, d_{3j}, d_{4j}, d_{5j}, d_{6j})$ with $w_j = \text{DayNum}_j$ and $d_{kj}, k = 1, 2, \dots, 6$, the dummy variables representing whether Day is a Tuesday, a Wednesday, a Thursday, a Friday, a Saturday, or a Sunday, respectively, has Poisson distribution with rate $\lambda(\mathbf{x}_j; \theta) = \exp\{\mathbf{x}_j \theta\}$. We mention that in our R code for fitting this model, for computational stability, we first centered and standardized the non-constant columns of the $\mathbf{x}_j$'s. Let $\alpha^* \in (0, 1)$, and the Y_j s being independent. From the preceding section we know how to construct a $100(1 - \alpha^*)\%$ prediction interval $[a_j, b_j]$ for Y_j , where, as mentioned earlier, we will simply utilize the prediction region $\tilde{I}_1$. By the independence, we will have

$$(1 - \alpha^*)^{17} \leq \Pr \left\{ \bigcap_{j=138}^{154} [Y_j \in [a_j, b_j]] \mid S_{137} \right\} \leq \Pr \left\{ S_{154} - S_{137} = \sum_{j=138}^{154} Y_j \in [a_\bullet \equiv \sum_{j=138}^{154} a_j, b_\bullet \equiv \sum_{j=138}^{154} b_j] \mid S_{137} \right\}.$$

Thus, if we wanted $S_{137} + [a_\bullet, b_\bullet]$ to be a $100(1 - \alpha)\%$ prediction interval for S_{154} , given S_{137} , we could choose $\alpha^* = 1 - (1 - \alpha)^{1/17}$. This procedure will guarantee a conservative $100(1 - \alpha)\%$ prediction interval for S_{154} , given S_{137} . This is the approach we followed in constructing a conservative 95% prediction region for S_{154} , the cumulative number of deaths due to COVID-19 in the US by the end of May 31, 2020.

However, another approach that is expected to produce a prediction region which is not conservative and potentially with a shorter length than the conservative prediction region is to note from the above derivation that it suffices to obtain a prediction region for $T = \sum_{j=138}^{154} Y_j$, given $\{Y_{62}, Y_{63}, \dots, Y_{137}\}$, since with such a prediction region for T , we simply need to add S_{137} to this prediction region to obtain a prediction region for S_{154} . Since the Y_j s are independent Poisson random variables, then conditionally and unconditionally on S_{137} ,

$$T_0 = \sum_{j=138}^{154} Y_j \sim \text{POI} \left(\Lambda_0(\theta) \equiv \sum_{j=138}^{154} \lambda(\mathbf{x}_j; \theta) = \sum_{j=138}^{154} \exp\{\mathbf{x}_j \theta\} \right).$$

Again, we could approximate the distribution of T_0 by a normal distribution with mean and variance $\Lambda_0(\theta)$. Using the Delta-Method, with $\hat{\theta}$ the estimator of θ based on $Y_j, j = 62, 63, \dots, 137$, we obtain the result

$$[T_0 - \Lambda_0(\hat{\theta})] \sim N[0, \hat{V}_0^2 \equiv \Lambda_0(\hat{\theta}) + \dot{\Lambda}_0(\hat{\theta})^T \mathbf{I}(\hat{\theta})^{-1} \dot{\Lambda}_0(\hat{\theta})].$$

Note that $\dot{\Lambda}_0(\hat{\theta}) = \sum_{j=138}^{154} \mathbf{x}_j \lambda(\mathbf{x}_j; \hat{\theta})$. Using this result, a prediction region for T_0 is

$$[\Lambda_0(\hat{\theta}) - z_{\alpha/2} \hat{V}_0, \Lambda_0(\hat{\theta}) + z_{\alpha/2} \hat{V}_0] \cap \mathbb{Z}_{0,+}.$$

Therefore, an approximate prediction region for S_{154} , given S_{137} , is

$$[(S_{137} + \Lambda_0(\hat{\theta})) - z_{\alpha/2} \hat{V}_0, (S_{137} + \Lambda_0(\hat{\theta})) + z_{\alpha/2} \hat{V}_0] \cap \mathbb{Z}_{0,+}. \quad (28)$$

The preceding discussion is for the specific situation where the data used is until $\text{DayNum} = 137$, and we are forecasting for $\text{DayNum} = 154$. In general, we could let n (which equals 137 in the illustration) to be the value of DayNum for the end date of the data used in developing the prediction model, while m (154 in this illustration) to be the value of DayNum for which the forecast is desired, so that $T_0 = \sum_{j=n+1}^m Y_j$ and $\Lambda_0(\theta) = \sum_{j=n+1}^m \lambda(\mathbf{x}_j; \theta)$.

4.1 Over-Dispersed Poisson Regression Model

We implemented the above procedure and also constructed the prediction regions at each of the observed DayNum which are depicted in the left-panel of Figure 7. The right-panel is the result associated with the over-dispersed Poisson model which will be described shortly. Examining this plot, note that there are more observed Deaths outside the prediction curves than what is expected nominally. This indicates that either there is more variability inherent in the stochastic mechanism generating the observed number of daily deaths relative to a purely Poisson regression model, or the fifth-order Poisson rate model is still inadequate, or both. We therefore propose an approach that introduces over-dispersion with the Poisson regression model serving as a hidden model. We mention that our Occam's Razor-type solution is motivated by frailty modeling in Survival Analysis (see, for instance, Andersen *et al.*, 1993). Our model assumes the existence of an unobserved positive latent variable Z_j of mean 1 at $w_j = \text{DayNum}_j$, and the reported number of deaths Y_j is the integer part of $Z_j Y_j^*$, with $\{Y_j^*\}$ arising from a Poisson regression model and with Z_j and Y_j^* independent. Recall that frailty models in Survival Analysis are used specifically to model correlations among observations; whereas, in our model it serves as an unobserved random contamination component in the observed number of daily deaths. In our implementation, we shall take Z_j to have a gamma distribution with mean one and variance $1/\xi$ (see Casella and Berger, 1990). This ξ is then an additional parameter in the regression model aside from the parameter vector θ . Such a model leads to an over-dispersed Poisson regression model, with the purely Poisson regression model embedded in this model and obtainable as a limiting case when $\xi \rightarrow \infty$. The idea behind this model is that Y_j^* represents the true number of occurrences of the event of interest, but the observed value Y_j is a 'blurring' or a 'noisy' version of Y_j^* . In fact, ignoring the flooring operation in $Y_j = \lfloor Y_j^* Z_j \rfloor$ and taking logarithm, we have $\log(Y_j) = \log(Y_j^*) + \log(Z_j)$, so that $\log(Z_j)$ serves as a noise contamination. One could actually obtain the exact distribution of Y_j under $Y_j^* \sim \text{POI}(\lambda_j)$. Its probability mass function, in infinite-sum

Figure 7: Scatterplot of daily deaths data until May 15, 2020, together with the 95% prediction regions under (i) the Poisson regression model in the left-panel; (ii) the 5th-order over-dispersed Poisson model in the right-panel.

form, is

$$\Pr\{Y_j = k|\lambda_j, \xi\} = \exp(-\lambda_j) \left\{ I\{k=0\} + \sum_{m=1}^{\infty} \frac{\lambda_j^m}{m!} w_m(k|\xi) \right\}, \quad k = 0, 1, 2, \dots,$$

where $w_m(k|\xi) = \mathfrak{G}_{\xi, \xi}((k+1)/m) - \mathfrak{G}_{\xi, \xi}(k/m)$, with $\mathfrak{G}_{\alpha, \beta}(\cdot)$ the CDF of a gamma variable with mean α/β and variance α/β^2 .

Inference for such a model requires further study, with possible use of an EM-type algorithm, though this could be difficult to implement since the distributions of the Y_i s are not in closed forms. However, we may implement a Z-estimation approach (see, for instance, van der Vaart, 1998). We first note that

$$\begin{aligned} E(Y_i) &\approx E[E(Y_i|Z_i)] = E[Z_i\lambda_i] = \lambda_i; \\ \text{Var}(Y_i) &\approx E[\text{Var}(Y_i|Z_i)] + \text{Var}[E(Y_i|Z_i)] \\ &= E[Z_i^2\lambda_i] + \text{Var}[Z_i\lambda_i] = \lambda_i(1 + 1/\xi) + \lambda_i^2(1/\xi) = \lambda_i \left[1 + \left(\frac{1 + \lambda_i}{\xi} \right) \right]. \end{aligned}$$

The approximate higher-order moments of the Y_i s for the over-dispersed Poisson are also obtainable.

We briefly pause at this point to differentiate this model from the negative binomial (NB) model discussed in Hilbe (2011), a model that is over-dispersed relative to the Poisson distribution. Without covariates, in the NB model the ratio of the variance and the mean is equal to $1/\omega$, where ω is the success probability in this model, hence the over-dispersion property: variance exceeds mean. There are several versions of the negative binomial regression model, e.g., see Table 8.1 of Hilbe (2011) for 22 different NB models. Denoting by λ_i the mean for the i th unit, the so-called NB1 model postulates that $\omega_i^{-1} = (1 + \kappa)$, while in the so-called NB2 model it is postulated that $\omega_i^{-1} = (1 + \kappa\lambda_i)$, where $\kappa > 0$ is the over-dispersion parameter. Specifying λ_i and ω_i is equivalent to specifying the parameter r_i , the number of ‘successes’ desired and the ‘success’ probability ω_i , though there may be a need to perform a floor or ceiling operation since r_i is an integer parameter. For our proposed model, the ratio between the variance and mean is approximately equal to $1 + (1 + \lambda_i)/\xi$, which always exceeds 1. Thus, in a sense, the newly-introduced model is some sort of a hybrid version of the NB1 and NB2 models, with $\kappa = 1/\xi$, at least from the moments perspective.⁴ Though the first two moments of the NB models and our proposed model are quite similar, their distributions could be quite different, though we also

Figure 8: Plots of PMFs for the Poisson (red), negative binomial (blue), and the over-dispersed Poisson (black). The parameters are $\lambda = 10$, $r = 10$ is the size for the negative binomial, and $\xi \in \{2, 10, 100\}$ for panels 1, 2, and 3, respectively. The success probabilities for the negative binomial were chosen so the respective means equal λ .

⁴ There is another model called NB12 which specifies the variance and mean ratio to be $\omega_i^{-1} = (\kappa_1 + \kappa_2\lambda_i)$. With respect to the first two moments, our model can therefore be considered as a special case of this NB12 model with $\kappa_1 = 1 + 1/\xi$ and $\kappa_2 = 1/\xi$. This NB12 model has only been considered in a limited number of publications, e.g., Lindén and Mäntyniemi (2011).

mention that the Poisson model is a limit in distribution of each of these models: (i) by taking $\xi \uparrow \infty$ in the new model, and (ii) by taking $\kappa \downarrow 0$ in the NB models. We pictorially depict the probability mass functions for the Poisson, NB, and the over-dispersed Poisson proposed under three sets of parameter values in Figure 8. Note that as ξ increases, the over-dispersed Poisson becomes closer to the Poisson. If the size parameter r of the NB was also increased with the mean remaining equal to λ , it will also become closer to the Poisson, since the Poisson is a limit of NBs. Looking at the first plot panel, we observe that these three PMFs could be quite different. An appealing property of our new over-dispersed regression model over the NB regression model is its physical interpretability in the sense of having the flavor of a discrete measurement error model. Such a physical or measurement error interpretation is not evident with the NB regression model. We also mention in passing that alternative approaches for handling over-dispersion are the quasi-Poisson approach (McCullagh and Nelder, 1989) or the mixed-Poisson approach (Cameron and Trivedi, 2013).

Based on the approximate moments for the introduced over-dispersed regression model, we could form the set of estimating equations, where we recall that $\rho(w) = \exp(w)$:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^T [Y_i - \rho(\mathbf{x}_i \theta)] &= 0; \\ \frac{1}{n} \sum_{i=1}^n \{ [Y_i - \rho(\mathbf{x}_i \theta)]^2 - \rho(\mathbf{x}_i \theta) [1 + (1 + \rho(\mathbf{x}_i \theta))/\xi] \} &= 0. \end{aligned}$$

By Z-Estimation Theory (see van der Vaart, 1998) and under regularity conditions, it will follow that, for some

$(p+1) \times (p+1)$ matrix $\Xi = \begin{bmatrix} \Xi_{11} & \Xi_{12} \\ \Xi_{21} & \Xi_{22} \end{bmatrix}$, we have

$$\begin{pmatrix} \tilde{\theta} \\ \tilde{\xi} \end{pmatrix} \sim AN \left[\begin{pmatrix} \theta \\ \xi \end{pmatrix}, \frac{1}{n} \Xi \right].$$

In fact, let us introduce the $(p+1) \times 1$ vector functions:

$$\mathbf{U}((y, \mathbf{x}); \theta, \xi) = \begin{bmatrix} \mathbf{U}_1((y, \mathbf{x}); \theta, \xi) \\ \mathbf{U}_2((y, \mathbf{x}); \theta, \xi) \end{bmatrix} = \begin{bmatrix} \mathbf{x}^T [y - \rho(\mathbf{x}\theta)] \\ [y - \rho(\mathbf{x}\theta)]^2 - \rho(\mathbf{x}\theta) [1 + (1 + \rho(\mathbf{x}\theta))/\xi] \end{bmatrix}.$$

Denote by $\mathbf{H}$ the $(p+1) \times (p+1)$ matrix function consisting of the derivatives of $\mathbf{U}$ with respect to (θ, ξ) . The components of this matrix function are:

$$\begin{aligned} \mathbf{H}_{11}((y, \mathbf{x}); \theta, \xi) &= -\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^{\otimes 2} \rho(\mathbf{x}_i \theta) \quad \text{and} \quad \mathbf{H}_{12}((y, \mathbf{x}); \theta, \xi) = \mathbf{0}; \\ \mathbf{H}_{21}((y, \mathbf{x}); \theta, \xi) &= -\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^T \rho(\mathbf{x}_i \theta) \{ 2(y_i - \rho(\mathbf{x}_i \theta)) + [1 + (1 + \rho(\mathbf{x}_i \theta))/\xi] + \rho(\mathbf{x}_i \theta)/\xi \}; \\ H_{22}((y, \mathbf{x}); \theta, \xi) &= \frac{1}{n} \frac{1}{\xi^2} \sum_{i=1}^n \rho(\mathbf{x}_i \theta) [1 + \rho(\mathbf{x}_i \theta)]. \end{aligned}$$

We could then obtain the estimates via the Newton-Raphson (NR) method with iteration step

$$\begin{bmatrix} \theta \\ \xi \end{bmatrix} \leftarrow \begin{bmatrix} \theta \\ \xi \end{bmatrix} - \left[\frac{1}{n} \sum_{i=1}^n \mathbf{H}((y_i, \mathbf{x}_i); \theta, \xi) \right]^{-1} \left[\frac{1}{n} \sum_{i=1}^n \mathbf{U}((y_i, \mathbf{x}_i); \theta, \xi) \right].$$

It turns out that a simpler way to obtain the estimates of θ and ξ is to first obtain the estimate $\tilde{\theta}$ of θ from the first estimating equation. This could be done by using the `glm` object function in R (R Core Team, 2013) with the Poisson family and logarithm inverse link. The estimate $\tilde{\xi}$ of ξ is then obtained from the second estimating equation using a one-variable NR iteration with θ replaced by $\tilde{\theta}$. Define the $(p+1) \times (p+1)$ matrices Σ and Ω according to

$$\Sigma = \text{plim} \frac{1}{n} \sum_{i=1}^n \mathbf{U}((Y_i, \mathbf{x}_i); \theta, \xi)^{\otimes 2} \quad \text{and} \quad \Omega = \text{plim} \frac{1}{n} \sum_{i=1}^n \mathbf{H}((Y_i, \mathbf{x}_i); \theta, \xi),$$

with plim denoting “in-probability limit” as $n \rightarrow \infty$. Then, the asymptotic covariance matrix of $(\tilde{\theta}^T, \tilde{\xi})^T$ is

$$\Xi = \begin{bmatrix} \Xi_{11} & \Xi_{12} \\ \Xi_{21} & \Xi_{22} \end{bmatrix} = \Omega^{-1} \Sigma [\Omega^{-1}]^T,$$

which could be consistently estimated by $\hat{\Xi} = \hat{\Omega}^{-1} \hat{\Sigma} [\hat{\Omega}^{-1}]^T$, with

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \mathbf{U}((Y_i, \mathbf{x}_i); \tilde{\theta}, \tilde{\xi})^{\otimes 2} \quad \text{and} \quad \hat{\Omega} = \frac{1}{n} \sum_{i=1}^n \mathbf{H}((Y_i, \mathbf{x}_i); \tilde{\theta}, \tilde{\xi}).$$

We note that another way of obtaining estimates of Σ and Ω is to obtain their theoretical expressions using higher-order moments of the Y_j s. These expressions depend on θ and ξ , so estimates could then be obtained by replacing θ and ξ by $\hat{\theta}$ and $\hat{\xi}$. We also point out that Ξ_{11} is generally not equal to Σ_{11}^{-1} when ξ is finite. In fact, it is imperative that Ξ_{11} should be used instead of Σ_{11}^{-1} since it takes into account the *impact* of the estimation of ξ .

Applying the Delta-Method (see van der Vaart, 1998), we then have the pivotal quantity result, as $n \rightarrow \infty$, given by

$$\frac{Y_0 - \hat{\lambda}_0}{\sqrt{\hat{\lambda}_0[1 + (1 + \hat{\lambda}_0)/\tilde{\xi}] + \hat{\lambda}_0^2 \hat{\psi}_0^T \hat{\Xi}_{11} \hat{\psi}_0 / n}} \dot{\sim} N(0, 1),$$

where quantities with “ $\dot{\sim}$ ” are estimates obtained by plugging in $\tilde{\theta}$ and $\tilde{\xi}$ for θ and ξ in their respective expressions. From this pivotal quantity, it follows that an approximate $100(1 - \alpha)\%$ prediction region for Y_0 is given by

$$\tilde{I}_6 \equiv \left[\hat{\lambda}_0 \pm z_{\alpha/2} \sqrt{\frac{\hat{\lambda}_0(1 + \hat{\lambda}_0)}{\tilde{\xi}} + \hat{\lambda}_0 + \frac{1}{n} \hat{\lambda}_0^2 \hat{\psi}_0^T \hat{\Xi}_{11} \hat{\psi}_0} \right] \cap \mathbb{Z}_{0,+}. \quad (29)$$

Because of the term $1/\tilde{\xi}$, when ξ is small, this prediction interval will be wider than the prediction regions under the purely Poisson regression model. As $\xi \rightarrow \infty$, which makes the model approach the Poisson regression model, then this prediction interval will approach $\tilde{I}_1$.

The form of the prediction region in (29) could now be used to construct a conservative prediction region for S_{n+m} , the cumulative number of deaths at DayNum = $n + m$, given S_n . One first constructs the prediction regions for Y_{n+j} , $j = 1, 2, \dots, m$, using $\alpha^* = 1 - (1 - \alpha)^{1/m}$. Then the lower bounds of these prediction regions, and also the upper bounds, are summed up to obtain the conservative prediction region. However, we found that this approach produces highly conservative prediction regions. Another way to construct an approximate prediction region for S_{n+m} is to extend (28) for this over-dispersed Poisson regression model. The resulting prediction region for S_{n+m} is given by

$$\left[\left(S_n + \sum_{j=n+1}^m \lambda(\mathbf{x}_j; \hat{\theta}) \right) \pm z_{\alpha/2} \hat{V}_0^* \right] \cap \mathbb{Z}_{0,+}, \quad (30)$$

where

$$(\hat{V}_0^*)^2 = \sum_{j=n+1}^m \lambda(\mathbf{x}_j; \hat{\theta}) + \sum_{j=n+1}^m \frac{\lambda(\mathbf{x}_j; \hat{\theta})(1 + \lambda(\mathbf{x}_j; \hat{\theta}))}{\tilde{\xi}} + \left(\sum_{j=n+1}^m \lambda(\mathbf{x}_j; \hat{\theta}) \mathbf{x}_j \right)^T \hat{\Xi}_{11} \left(\sum_{j=n+1}^m \lambda(\mathbf{x}_j; \hat{\theta}) \mathbf{x}_j \right).$$

This second approach is the method we used in the COVID-19 application. There are different prediction interval procedures based on the other approaches for dealing with over-dispersion. For instance, the procedure in Olive et al. (2021) can be constructed from an NB model, the procedure in Hakulinen and Dyba (1994) uses a quasi-Poisson model, while the method in Wood (2005) relied on the mixed-Poisson models in Ash et al. (2021). Implementing this procedure in (29) based on this 5th-order over-dispersed Poisson regression model, we first examine the *pointwise* prediction regions at each of the observed DayNum values, which are shown in the right-panel of Figure 7. We now see that these approximate 95% prediction regions cover most of the observed daily deaths. This indicates that the over-dispersed Poisson regression model is a more appropriate

Figure 9: Pointwise prediction regions for daily and cumulative deaths under the over-dispersed 5th-order Poisson regression model using data from $\text{DayNum} = 62$ (March 1, 2020) until $\text{DayNum} = 137$ (May 15, 2020). Red points in the plot were the subsequent observed number of deaths after $\text{DayNum} = 137$. The last value of DayNum is 154 coinciding with the end of May 31, 2020.

model for the observed daily deaths data than the purely Poisson regression model whose approximate 95% prediction regions are shown in the left-panel of Figure 7. The estimate of ξ turned out to be $\hat{\xi} = 16.89016$.

The left-panel of Figure 9 is the scatterplot of daily deaths, but now including the actual observed values after DayNum 137 and until DayNum 154 (these are the red dots) and the *pointwise* prediction regions for the daily deaths past DayNum 137. The right-panel of Figure 9 displays the *pointwise* prediction interval for S_{154} , given S_{137} ; in fact, this also displays the pointwise prediction regions for S_k , $k = 138, 139, \dots, 153$, given S_{137} . The red dots are the actual observed values past DayNum equal to 137. The predicted cumulative deaths on $\text{DayNum} = 154$ was $\hat{S}_{154} = 96876$ and the approximate pointwise prediction interval for S_{154} was $[85906, 109258]$. In both of these plots, notice that the 5th-order prediction model did not perform well past $\text{DayNum} = 150$, though the prediction interval for S_{154} did still cover what was *actually* observed, which was 104383. From Figure 9, an elevated number of daily deaths occurred starting at $\text{DayNum} = 150$ (May 28), which could be a consequence of some relaxation of social distancing and other mitigation guidelines. The grim milestone of 100000 cumulative deaths due to COVID-19 in the US was also surpassed on this day.

4.2 Sensitivity Analysis

To partly assess the sensitivity of the procedure to the data set used in the prediction, if we had used the data until $\text{DayNum} = 145$ (May 23rd), the predicted value for S_{154} is $\hat{S}_{154} = 101010$ and the approximate 95% prediction region, given S_{145} , is $[98430, 103590]$ (written as an interval), which does **not** cover what was eventually observed. The associated plots in this case are provided in Figure 10. On the other hand, if we had used the data until June 1, 2020 ($\text{DayNum} = 154$), the predicted value is 104383, which coincided with the *observed* cumulative number of deaths. That the predicted cumulative number of deaths and the observed cumulative number of deaths on the last day were equal is actually a consequence of the estimating equation, hence this is not a surprising result. The associated plots in this case are provided in Figure 11. Observe in the right-panel of Figure 11 that the model for the cumulative deaths based on this 5th-order model is quite excellent, lending support to our novel approach of modeling the daily deaths data, instead of the cumulative deaths data, for the purpose of making predictions for the cumulative deaths. We summarize the results of this sensitivity analysis in Figure 12, where we plot the predicted values and the prediction regions for S_{154} under scenarios where data used in the model fitting is up to the different days from DayNum equal to 137 up to 153. Based on this analysis, the 5th-order model appears to possess stability since the predictions and

the prediction regions for S_{154} remain somewhat consistent as the amount of data being used in the model fitting varies. As to be expected, note that as the forecasting horizon shortens, then the prediction interval also narrows. This is the case, *even* though we still considered the outlier on April 16th, which was the adjusted Deaths data point, as a legitimate observation.

However, any model, especially higher-order models, will have a breakdown point in the sense of yielding seemingly unreasonable predictions, perhaps due to a long forecasting horizon, insufficient amount of data, or wildly changing data points drastically altering estimates which highly impact forecasts. For this 5th-order model, it appears to break down when the data used is on or before DayNum 132. One possible cause appears to be sharp increases and decreases in the observed daily deaths. For instance, on DayNum = 120 the reported daily deaths was 1369, but on the next two days they were 2110 and 2611, and these led to a huge jump in the predicted value for S_{154} . Also, from DayNum ranging from 125 to 133, the reported daily deaths were 1317, 1297, 1252, **2144**, **2353**, **2239**, 1510, 1624, and 734, and the predictions were highly unstable and only started to stabilize after DayNum = 132. This is a clear warning on the danger of fitting higher-order models or

Figure 10: Pointwise prediction regions for daily and cumulative deaths under the over-dispersed 5th-order Poisson regression model using data from DayNum = 62 (March 1, 2020) until DayNum = 145 (May 23, 2020). Red points in the plot were the subsequent observed number of deaths and cumulative deaths after Daynum = 145. The last value of DayNum is 154 coinciding with the end of May 31, 2020.

Figure 11: Prediction curve and regions for daily deaths and observed cumulative deaths together with its associated fitted prediction curve under the over-dispersed 5th-order Poisson regression model using data from DayNum = 62 (March 1, 2020) until DayNum = 154 (June 1, 2020).

Figure 12: Predictions and prediction regions for the cumulative deaths by June 1, 2020 when the data used in the model fitting is up until the different days between May 15, 2020 and June 1, 2020. The gray horizontal dashed line represents the actual observed cumulative number of deaths on June 1, 2020, which was 104383.

extrapolating with a long forecasting horizon. As the adage goes, attributed to Niels Bohr (see Ulam, 1976), with similar versions attributed to Mark Twain, Yogi Berra, and others: *It is difficult to make predictions, especially about the future.*

4.3 Forecasting One Month Ahead

With dire thoughts of the inherent dangers of forecasting over a long horizon, and fully cognizant of the many eventualities (e.g., re-opening of economy; nationwide protests and riots due to police brutality; changing hotspots; adjustments on counts; etc.), which we are not taking into account in order to be purely data-driven, but which could drastically alter trajectories of daily and cumulative deaths, on the basis of the observed data up to September 1, 2020, in which the cumulative deaths is 183598, we seek to forecast the cumulative deaths on October 1, 2020. We should mention that on June 26th, there was an adjustment of 1854 which occurred

Figure 13: Based on the available data until September 1, 2020 (DayNum = 246), point predictions and prediction regions of the deaths and cumulative deaths by October 1, 2020 (DayNum = 276). The red points after September 1 until September 25 (DayNum = 270) are the eventual realized daily and cumulative deaths values. These values were not used in developing the prediction model.

“following a state review of death certificates and prior outbreaks” in the State of New Jersey; see the June 25, 2020 *USA Today* article (USA, 2020). Together with the 3778 adjustment made by the State of New York on April 16, 2020 (BBC, 2020b), this is the second documented non-trivial adjustment made on the daily deaths counts.

Figure 13 presents the point prediction and the prediction interval for October 1, 2020 based on the 5th-order over-dispersed Poisson regression model and using data until September 1st. The plots also include the realized values until September 25, 2020, the date the original version of this manuscript was finalized, which were not included in the model-fitting. These are the points indicated by red dots. Observe that most of the realized daily deaths values are falling within the lower and upper limits of the prediction regions, except for the two observations when adjustments were made. This somewhat confirms that the over-dispersed model is properly accounting for the increased variability, hence the prediction regions are providing proper uncertainty quantification. The prediction curve for the cumulative deaths also appear to be tracking well the observed cumulative deaths, with a light degradation after September 15th. Observe that all the realized cumulative deaths values after September 1st are inside the cone of uncertainty. The predicted cumulative deaths on October 1st, based on data until September 1st, is 203258 with a prediction interval of [183645, 222871]. When the observed data until September 25th is used, the predicted cumulative deaths on October 1st is 207774 with a prediction interval of [206289, 209258]. The observed cumulative number of deaths on September 25th is 202810. Note that the prediction for the cumulative deaths on October 1st based on data until September 25th has increased from the one when using the data until September 1st, while the prediction interval has narrowed considerably.

It remains to be seen if this increase in the predicted cumulative deaths value by October 1st, when using data from September 1st to September 25th, is the effect of the lessening of social distancing guidelines; re-opening of business establishments and beaches; re-opening of schools, colleges and universities; of people gathering because of social unrests; or changing hotspots in the country. It should be noted that in forecasting settings with new data points accruing frequently – daily in this COVID-19 pandemic – forecasts should be updated as each new data point accrues.

5 Concluding Remarks

Motivated by the COVID-19 pandemic, we examined the problem of constructing prediction regions for a Poisson distributed random variable, both under the no-covariate (that is, intercept-only) and with-covariate settings. We compared the performances of the different prediction regions through simulation studies. In the regression setting, we also introduced an over-dispersed Poisson regression model upon observing over-dispersion in the COVID-19 reported (daily) *Deaths* data relative to a purely Poisson regression model. This model was motivated by ideas from frailty models in Survival Analysis and is intrinsically different from a negative binomial model, which also induces over-dispersion. With the ultimate goal of predicting cumulative deaths due to COVID-19 at a future date, we first studied how to construct prediction regions for the number of daily deaths, and then utilized these prediction regions to construct the prediction region for the number of cumulative deaths. The final fitted prediction models involved a 5th-order model in the variable *DayNum*, and also included the factor variable *Day*. Based on data until September 1, 2020, prediction and prediction interval for the October 1, 2020 cumulative deaths were obtained. Prediction regions between these two dates were also constructed and the realized values until the date of final preparation of the original version of this manuscript, which was September 25th, were also ascertained to be almost all within the bounds of the prediction regions, lending credence that the proposed over-dispersed Poisson regression model is effective in accounting for the increased variability in this COVID-19 data set. The methodologies developed have the potential to be used in the monitoring of daily and cumulative deaths during epidemics or pandemics through the construction of prediction regions, which could then be used by decision-makers regarding implementation of social distancing/easing guidelines and deciding on the closure/opening of business, educational, government, and other establishments. However, further studies are needed to compare our methodologies to other methods that have been proposed during this pandemic.

The prediction and prediction region procedures we developed still possess some limitations, aside from the fact that they did not incorporate other potential informative covariates such as the status of mitigation guidelines, degree of face mask usage, observance of social distancing, presence of relevant vaccines, etc. First, in contrast to the Susceptible-Exposed-Infected-Recovered (SEIR) compartment model cf., Allen (2017), which is based on a continuous-time Markov Chain, our model does not posit an upper bound to the number of people that could die, which clearly is not the case. Second, especially since it involves higher-order terms in DayNum , they are highly sensitive to outliers, such as when huge adjustments are made as in the cases for the States of New York and New Jersey (see WSJ, 2020; USA, 2020). It would be desirable to develop procedures that are robust to such non-trivial adjustments, or procedures that impose constraints on the rate of increase or decrease of the prediction curve via regularization in order to curb the influential impact of such adjustments, though this will entail developing new theory for the construction of prediction regions. Due to these first two limitations, the proposed methods are not suitable for use in long-horizon forecasting, hence our decision to simply forecast one month forward. Third, the procedures are not adaptive in their choice of the prediction model. Possible improvements may be possible by choosing the prediction model in a data-dependent manner, but then model choice uncertainty needs to be accounted for in constructing prediction regions. Fourth, there could be an advantage in utilizing other bases functions to transform the variable DayNum , such as by using Laguerre polynomials, Legendre polynomials, trigonometric functions, or even splines or wavelets. These limitations of the proposed methods generate several potential research avenues for further studies and improvements on the forecasting procedures.

Finally, we mention an interesting open problem, which is the construction of a simultaneous prediction region in two-dimensional space $\mathbb{Z}_{0,+}^2$ when there are two count variables of interest. An example of this situation is when we want to *simultaneously* predict (Y_0^1, Y_0^2) , where Y_0^1 is the number of COVID-19 deaths and Y_0^2 is the number of COVID-19 cases. A simultaneous treatment of these two variables is important since Y_0^1 and Y_0^2 are going to be correlated, and this must be taken into account in the simultaneous prediction. In the COVID-19 or pandemic setting, one may then extend the problem to constructing simultaneous prediction regions for each of their cumulative totals at some future time, given their joint data up to the current time.

Acknowledgement: We thank the editor, associate editor, and referees for their useful and insightful comments and criticisms which helped in improving the manuscript. T. Kim thanks Professor Alexander Goldenshluger for providing the post-doctoral research opportunity at the University of Haifa in Israel. E. Peña is currently Program Director in the Directorate of Mathematical and Physical Sciences (MPS), Division of Mathematical Sciences (DMS) at the US National Science Foundation (NSF). As a consequence of this position, he receives support for research, which included work in this manuscript, under NSF Grant 2049691 to the University of South Carolina.

Disclaimer: Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Conflict of interest: Prof. Edsel A. Peña is a member of the Editorial Advisory Board of Open Statistics. Prof. George Luta is a member of the Editorial Board of Open Statistics. They were not involved in the review process of this article.

References

- Allen, L. J. (2017), "A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis," *Infectious Disease Modelling*, 2, 128–142.
- Andersen, P., Borgan, O., Gill, R., and Keiding, N. (1993), *Statistical Models Based on Counting Processes*, New York: Springer-Verlag.
- Ash, J. E., Zou, Y., Lord, D., and Wang, Y. (2021), "Comparison of confidence and prediction intervals for different mixed-Poisson regression models," *Journal of Transportation Safety & Security*, 13, 357–379.

- Bain, L. J., and Patel, J. K. (1993), "Prediction intervals based on partial observations for some discrete distributions," *IEEE Transactions on Reliability*, 42, 459–463.
- BBC (2020a), "Coronavirus: What is the true death toll of the pandemic?" <https://www.bbc.com/news/world-53073046>.
- BBC (2020b), "Coronavirus: Why is NYC reporting surge in virus deaths?" <https://www.bbc.com/news/world-us-canada-52303739>.
- Bejleri, V. (2005), *Prediction intervals for the Poisson model with applications to Atlantic storms data*, ProQuest LLC, Ann Arbor, MI, Thesis (Ph.D.)—The American University.
- Bejleri, V., and Nandram, B. (2018), "Bayesian and frequentist prediction limits for the Poisson distribution," *Comm. Statist. Theory Methods*, 47, 4254–4271.
- Box, G. E. P. (1976), "Science and statistics," *J. Amer. Statist. Assoc.*, 71, 791–799.
- Cameron, A. C., and Trivedi, P. K. (2013), *Regression analysis of count data*, Vol. 53, Cambridge University Press.
- Casella, G., and Berger, R. L. (1990), *Statistical inference*, The Wadsworth & Brooks/Cole Statistics/Probability Series, Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA.
- Cox, D., and Hinkley, D. (1974), *Theoretical Statistics*, New York: Chapman and Hall.
- CSTE (2021), "Update to standardized surveillance case definition and national notification for 2019 novel coronavirus disease (COVID-19)," <https://asprtracie.hhs.gov/technical-resources/resource/8322/standardized-surveillance-case-definition-and-national-notification-for-2019-novel-coronavirus-disease-covid-19>.
- Davison, A. C., and Hinkley, D. V. (1997), *Bootstrap methods and their application*, Vol. 1 of *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge University Press.
- ECDC (2020a), "Coronavirus disease (COVID-19) pandemic: increased transmission in the EU/EEA and the UK – seventh update," <https://www.ecdc.europa.eu/sites/default/files/documents/RRA-seventh-update-Outbreak-of-coronavirus-disease-COVID-19.pdf>.
- ECDC (2020b), "Historical data (to 14 December 2020) on the daily number of new reported COVID-19 cases and deaths worldwide," SOURCE: <https://opendata.ecdc.europa.eu/covid19/casedistribution/csv>.
- ECO (2020), "Tracking Covid-19 excess deaths across countries," <https://www.economist.com/graphic-detail/2020/04/16/tracking-covid-19-excess-deaths-across-countries>.
- Gibbons, R., Bhaumik, D., and Aryal, S. (2009), *Statistical Methods for Groundwater Monitoring* (Second Edition ed.), Hoboken: John Wiley & Sons.
- Hakulinen, T., and Dyba, T. (1994), "Precision of incidence predictions based on Poisson distributed observations," *Statistics in medicine*, 13, 1513–1523.
- Hilbe, J. M. (2011), *Negative binomial regression*, Cambridge University Press.
- Kim, T., Lieberman, B., Luta, G., and Peña, E. A. (2021), "Prediction intervals for Poisson-based regression models," *WIREs: Computational Statistics*, 1, e1568.
- Kramer, D. (2020), "COVID-19 pandemic modeling fraught with uncertainties," *Physics Today*, 73, 25–27.
- Krishnamoorthy, K., and Lv, S. (2018), "Highest posterior mass prediction intervals for binomial and Poisson distributions," *Metrika*, 81, 775–796.
- Krishnamoorthy, K., and Peng, J. (2011), "Improved closed-form prediction intervals for binomial and Poisson distributions," *Journal of Statistical Planning and Inference*, 141, 1709–1718.
- Lindén, A., and Mäntyniemi, S. (2011), "Using the negative binomial distribution to model overdispersion in ecological count data," *Ecology*, 92, 1414–1421.
- Liu, P., and Peña, E. A. (2016), "Sojourning with the homogeneous Poisson process," *The American Statistician*, 70, 413–423.
- McCullagh, P., and Nelder, J. A. (1989), *Generalized linear models*, Vol. 37, Chapman & Hall.
- Millard, S. P. (2013), *EnvStats: An R Package for Environmental Statistics*, New York: Springer.
- Myers, R. H., and Montgomery, D. C. (1997), "A tutorial on generalized linear models," *Journal of Quality Technology*, 29, 274–291.
- Nelson, W. (2003), *Applied life data analysis*, Wiley Series in Probability and Statistics, John Wiley & Sons.
- Olive, D. J., Rathnayake, R. C., and Haile, M. G. (2021), "Prediction intervals for GLMs, GAMs, and some survival regression models," *Communications in Statistics-Theory and Methods*.
- Peña, E. A., and Kim, T. (2019), "Median confidence regions in a nonparametric model," *Electronic Journal of Statistics*, 13, 2348–2390.
- R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
- Randles, R., and Wolfe, D. (1979), *Introduction to the theory of nonparametric statistics*, John Wiley & Sons.
- Rencher, A. C., and Schaale, G. B. (2008), *Linear models in statistics* (Second ed.), Wiley-Interscience [John Wiley & Sons], Hoboken, NJ.
- Resnick, S. (1992), *Adventures in stochastic processes*, Birkhäuser Boston, Inc., Boston, MA.
- Robbins, H. (1956), "An empirical Bayes approach to statistics," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I*, University of California Press, Berkeley and Los Angeles, pp. 157–163.
- (1980), "An empirical Bayes estimation problem," *Proc. Nat. Acad. Sci. U.S.A.*, 77, 6988–6989.
- (1983), "Some thoughts on empirical Bayes estimation," *Ann. Statist.*, 11, 713–723.
- Ulam, S. M. (1976), *Adventures of a mathematician*, Charles Scribner's Sons, New York.

- USA (2020), "NJ says 1,854 additional residents likely died of coronavirus after review of death records," <https://www.northjersey.com/story/news/2020/06/25/nj-deaths-residents-coronavirus-review-death-certificates-records/3236853001/>.
- van der Vaart, A. W. (1998), *Asymptotic statistics*, Vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge University Press, Cambridge.
- Weiss, L. (1955), "A note on confidence sets for random variables," *The Annals of Mathematical Statistics*, 26, 142–144.
- Wood, G. (2005), "Confidence and prediction intervals for generalised linear accident models," *Accident Analysis & Prevention*, 37, 267–273.
- WSJ (2020), "Why New York's Coronavirus Death Count Jumped: the Stories of Patients Who Died at Home," <https://www.wsj.com/articles/why-new-yorks-coronavirus-death-count-jumped-the-stories-of-patients-who-died-at-home-11587029400>.