On random-parameter count models for out-of-sample crash prediction
Accounting for the variances of random-parameter distributions
Xu, Pengpeng; Zhou, Hanchu; Wong, S.C.

Published in:
Accident Analysis and Prevention

Published: 01/09/2021

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

License:
CC BY

Publication record in CityU Scholars:
Go to record

Published version (DOI):
10.1016/j.aap.2021.106237

Publication details:

Citing this paper
Please note that where the full-text provided on CityU Scholars is the Post-print version (also known as Accepted Author Manuscript, Peer-reviewed or Author Final version), it may differ from the Final Published version. When citing, ensure that you check and use the publisher's definitive version for pagination and other details.

General rights
Copyright for the publications made accessible via the CityU Scholars portal is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights. Users may not further distribute the material or use it for any profit-making activity or commercial gain.

Publisher permission
Permission for previously published items are in accordance with publisher's copyright policies sourced from the SHERPA RoMEO database. Links to full text versions (either Published or Post-print) are only available if corresponding publishers allow open access.

Take down policy
Contact lbscholars@cityu.edu.hk if you believe that this document breaches copyright and provide us with details. We will remove access to the work immediately and investigate your claim.
Short Communication

On random-parameter count models for out-of-sample crash prediction: Accounting for the variances of random-parameter distributions

Pengpeng Xu a,b,*, Hanchu Zhou c,d, S.C. Wong a,e

Abstract

One challenge faced by the random-parameter count models for crash prediction is the unavailability of unique coefficients for out-of-sample observations. The means of the random-parameter distributions are typically used without explicit consideration of the variances. In this study, by virtue of the Taylor series expansion, we proposed a straightforward yet analytic solution to include both the means and variances of random parameters for unbiased prediction. We then theoretically quantified the systematic bias arising from the omission of the variances of random parameters. Our numerical experiment further demonstrated that simply using the means of random parameters to predict the number of crashes for out-of-sample observations is fundamentally incorrect, which necessarily results in the underprediction of crash counts. Given the widespread use and ongoing prevalence of the random-parameter approach in crash analysis, special caution should be taken to avoid this silent pitfall when applying it for predictive purposes.

1. Introduction

Since its introduction by Milton et al. (2008) and Anastasopoulos and Mannering (2009), the random-parameter approach has attracted considerable research interest and has become the benchmark for crash analysis (Mannering and Bhat, 2014). Numerous studies have demonstrated its superiority in interpretability, with a substantial improvement in goodness-of-fit and the ability to account for unobserved heterogeneity (Mannering et al., 2016). Recent studies reported that the approach empirically underperformed in out-of-sample crash prediction compared with its fixed-parameter counterpart (Tang et al., 2019; Hou et al., 2020). However, this underperformance can be readily attributed to using predictions that are based only on the means of the random parameters while ignoring their variances. Our study therefore attempts to demonstrate such a bias by providing a straightforward yet analytic solution to incorporate both the means and variances of random parameters for unbiased prediction.

2. Methods

Let \( Y_i \) denote the number of crashes for observation \( i (i = 1, 2, \ldots, n) \) during a certain period. \( Z_i \) and \( X_{ij} \) represent the exposure and the \( j \)th explanatory variable, respectively. When developing the random-parameter approach to model crash frequencies, without loss of generality, we formulate the following functional form given the random, non-negative, and integer nature of crash counts:

\[
Y_i \sim \text{Poisson}(\lambda_i) \\
\lambda_i = \beta_0 Z_i^\alpha \exp \left( \sum_{j=1}^{k} \beta_j X_{ij} + u_i \right)
\]

(1)

where \( \lambda_i \) is the parameter of Poisson distribution (i.e., the expected number of crashes), \( \beta_0 \) is the intercept, \( \beta_j (j = 1, 2, \ldots, k) \) refers to the \( j \)th regression coefficients, and \( u_i \) denotes the unstructured effect. Compared with the negative binomial model (also known as the Poisson-gamma model in which \( \exp(u_i) \) follows a Gamma distribution with mean 1 and variance \( \alpha \); Lord and Miranda-Moreno, 2008), a more flexible
alternative is to assume \( u_i \) as a normal distribution as suggested by Lord and Mannering (2010):

\[
u_i \sim \text{Normal}(0, \sigma_i^2)
\]  

(2)

Unlike the conventional Poisson-lognormal model with fixed parameters, by specifying \( \beta_j \) as a predefined distribution with mean \( \overline{\beta}_j \) and variance \( \sigma_j^2 \), each observation now has its individual coefficients. In theory, \( \beta_j \) can be assumed to follow any distributions, but the normal distribution is mostly used, given its better statistical fit (Anastasopoulos and Manning, 2009; El-Basyouny and Sayed, 2009; Ukkusuri et al., 2011; Venkataraman et al., 2013; Chen and Tarko, 2014; Xu and Huang, 2015; Amoh-Gyimah et al., 2016; Meng et al., 2017; Cai et al., 2018; Hou et al., 2018, 2021; Tang et al., 2019).

A challenge arises when applying the estimated random-parameter model to predict the number of crashes for out-of-sample observations, as in reality, it appears to be impossible to know the exact value of \( \beta_{ji} \) for sample \( i \neq j \). If \( \overline{\beta}_j \) is used as the default to estimate the number of crashes (i.e., \( \lambda_{ji} \); Mannering et al., 2016; Tang et al., 2019; Hou et al., 2020), we then have:

\[
\lambda_{ji} \approx \lambda_{ji} = \beta_{ji} \lambda_{ji} \exp \left( \sum_{j=1}^{n} Z_{ji} \beta_{ji} X_{ij} \right)
\]

(3)

However, the expectation of \( \lambda_{ji} \) is not only dependent on \( \overline{\beta}_j \) but also on the variances \( \sigma_j^2 \) and \( \sigma_i^2 \). Indeed, if we take a Taylor series expansion near the point (\( \beta_{ji} = \overline{\beta}_j \), \( u_{ji} = 0 \)), \( \lambda_{ji} \) can be closely approximated using a quadratic polynomial as follows (Wong and Wong, 2019):

\[
\lambda_{ji} \approx \beta_{ji} \lambda_{ji} \exp \left( \sum_{j=1}^{n} Z_{ji} \beta_{ji} X_{ij} \right) + \sum_{j=1}^{n} \beta_{ji} \frac{\partial \lambda_{ji}}{\partial \beta_{ji}} \left( \beta_{ji} - \overline{\beta}_j \right) + \frac{1}{2} \sum_{j=1}^{n} \beta_{ji} \frac{\partial^2 \lambda_{ji}}{\partial \beta_{ji}^2} \left( \beta_{ji} - \overline{\beta}_j \right)^2 + \sum_{j=1}^{n} \frac{1}{2} \beta_{ji} \frac{\partial^2 \lambda_{ji}}{\partial u_{ji}^2} \left( u_{ji} - 0 \right)^2
\]

(4)

where \( \frac{\partial \lambda_{ji}}{\partial \beta_{ji}} \) and \( \frac{\partial \lambda_{ji}}{\partial u_{ji}} \) refer to the first-order partial derivative of \( \lambda_{ji} \) to \( \beta_{ji} \) and \( u_{ji} \), respectively. \( \frac{\partial^2 \lambda_{ji}}{\partial \beta_{ji} \partial \beta_{ji}} \) and \( \frac{\partial^2 \lambda_{ji}}{\partial u_{ji} \partial u_{ji}} \) are the corresponding second-order partial derivatives.

Taking the expectation on both sides of Eq. (4), we have:

\[
E(\lambda_{ji}) = \beta_{ji} \lambda_{ji} \exp \left( \sum_{j=1}^{n} Z_{ji} \beta_{ji} X_{ij} \right) + \sum_{j=1}^{n} \beta_{ji} \frac{\partial \lambda_{ji}}{\partial \beta_{ji}} \left( \lambda_{ji} - \overline{\lambda}_j \right) + \frac{1}{2} \sum_{j=1}^{n} \beta_{ji} \frac{\partial^2 \lambda_{ji}}{\partial \beta_{ji}^2} \left( \lambda_{ji} - \overline{\lambda}_j \right)^2 + \sum_{j=1}^{n} \frac{1}{2} \beta_{ji} \frac{\partial^2 \lambda_{ji}}{\partial u_{ji}^2} \left( u_{ji} - 0 \right)^2
\]

(5)

The difference between Eq. (5) and (3) can then be quantified as:

\[
E(\lambda_{ji}) - \lambda_{ji} = \left( \frac{\sigma_j^2}{2} \left[ \ln(E(\lambda_{ji})) \right]^2 + \sum_{j=1}^{n} \frac{\sigma_i^2}{2} \left[ \lambda_{ji} \right] + \frac{1}{2} \beta_{ji} \frac{\partial \lambda_{ji}}{\partial \beta_{ji}} \left( \lambda_{ji} - \overline{\lambda}_j \right) + \frac{1}{2} \sum_{j=1}^{n} \beta_{ji} \frac{\partial^2 \lambda_{ji}}{\partial u_{ji}^2} \left( u_{ji} - 0 \right)^2 \right)
\]

(6)

That is, using \( \overline{\lambda}_j \) as the prediction will always result in a downward bias, given the convex property of the link function in Eq. (1). This silent pitfall has been ignored by safety analysts when applying random-parameter models for predictive purposes (Tang et al., 2019; Hou et al., 2020).

To validate, we conducted the following numerical experiment, given its advantages in maneuverability and reproducibility. We set \( n = 5000 \), \( Z_i \sim \text{Uniform}(200, 20000) \), \( X_i \sim \text{Uniform}(0, 20) \), \( \beta_0 = 0.004 \), \( \beta_{ji} \sim \text{Normal}(0.70, 0.06^2) \), \( \beta_{ji} \sim \text{Normal}(0.03, 0.03^2) \), and \( u_i \sim \text{Normal}(0.60^2) \). Only one explanatory variable was considered here for simplicity. We then sequentially sampled the values of \( Z_i, X_i, \beta_0, \beta_{ji} \), and \( u_i \) from the aforementioned distributions, calculated \( \lambda_{ji} \), and generated \( Y_i \) using the freeware RStudio (R Core Team, 2013). Such a simulated dataset matched well with the empirical data collected by Tang et al. (2019) for rural two-lane roads in Pennsylvania. Detailed descriptive statistics of our data under investigation are listed in Table 1.

To evaluate the out-of-sample predictive performance of the random-parameter count model, we performed a 10-fold cross-validation (James et al., 2013; Hou et al., 2021), in which the original 5000 observations were randomly partitioned into 10 mutually exclusive subsets of equal size. Each time, one subset was left out as the validation while the remaining subsets were combined for model estimation. This process was repeated 10 times. We repeatedly predicted the crash frequencies based on the simulated models. Reliable parameters for the random-parameter count models could be estimated via 600 Halton draws based on the simulated maximum likelihood method (Tang et al., 2019; Hou et al., 2019; Hou et al., 2020) to visualize any anomalies. Models with a CURE curve oscillating around zero within the 95% confidence interval (CI) are regarded as reliable in predictive performance (Hauer, 2015). Consistent

<table>
<thead>
<tr>
<th>Y_i</th>
<th>5.129</th>
<th>3.000</th>
<th>7.442</th>
<th>0.000</th>
<th>161,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>\lambda_j</td>
<td>5.047</td>
<td>2.897</td>
<td>6.874</td>
<td>0.031</td>
<td>154,089</td>
</tr>
<tr>
<td>Z_i</td>
<td>10234.448</td>
<td>10364.573</td>
<td>5759.323</td>
<td>201,152</td>
<td>19997.846</td>
</tr>
<tr>
<td>X_i</td>
<td>9.967</td>
<td>10.080</td>
<td>5.749</td>
<td>0.001</td>
<td>19,999</td>
</tr>
</tbody>
</table>

Table 1: Descriptive statistics for the simulated data.
absolute values according to the MAE results, E(λ̂) was less sensitive and more robust to outliers, given a lower RMSE value. It is also interesting to find that when E(λ̂) was used as the predictive value, the predictive performance of the random-parameter model was comparable with that of its fixed-parameter counterpart, as both models produced similar MAE and RMSE values. However, the MBE value of E(λ̂) was the closest to zero, suggesting that E(λ̂) is least likely to be affected by systematic over- or under-prediction bias. This finding becomes even more plausible when comparing the MBE values of Ŷ and E(λ̂).

More importantly, as illustrated in Fig. 1, when plotting the CURE curves over the entire range of a variable of interest, only Ŷ consistently exhibited a monotonically increasing trend, resulting in a substantial deviation from zero in the end. By comparison, although the CURE curve for E(λ̂) was not always close to zero, it oscillated broadly within the 95% CI. A similar fashion was also observed for the fixed-parameter model.

4. Discussion

A major challenge faced by the random-parameter count models in transferability and extrapolation is the unavailability of unique coefficients for out-of-sample observations. By virtue of the Taylor series expansion, we provided an analytic solution that includes both the

Table 2

<table>
<thead>
<tr>
<th>Round</th>
<th>Fixed-parameter model</th>
<th>Random-parameter model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MBE</td>
<td>MAE</td>
</tr>
<tr>
<td>1</td>
<td>0.694</td>
<td>4.079</td>
</tr>
<tr>
<td>2</td>
<td>0.677</td>
<td>3.866</td>
</tr>
<tr>
<td>3</td>
<td>−0.019</td>
<td>3.800</td>
</tr>
<tr>
<td>4</td>
<td>0.388</td>
<td>3.846</td>
</tr>
<tr>
<td>5</td>
<td>1.023</td>
<td>3.968</td>
</tr>
<tr>
<td>6</td>
<td>0.387</td>
<td>4.201</td>
</tr>
<tr>
<td>7</td>
<td>0.188</td>
<td>3.852</td>
</tr>
<tr>
<td>8</td>
<td>0.461</td>
<td>3.843</td>
</tr>
<tr>
<td>9</td>
<td>−0.479</td>
<td>4.198</td>
</tr>
<tr>
<td>10</td>
<td>−0.160</td>
<td>3.287</td>
</tr>
<tr>
<td>Average</td>
<td>0.316</td>
<td>3.894</td>
</tr>
</tbody>
</table>

MBE, MAE, and RMSE refer to the mean bias error, mean absolute error, and root-mean-square error, which are calculated as \( \frac{1}{500} \sum_i (Y_i - \hat{Y}_i) \), \( \frac{1}{500} \sum_i |Y_i - \hat{Y}_i| \), and \( \sqrt{\frac{1}{500} \sum_i (Y_i - \hat{Y}_i)^2} \), respectively. Here, \( Y_i \) and \( \hat{Y}_i \) represent the observed and predicted number of crashes for validation sample \( i \), respectively.

\[ \hat{Y}_i = (1 + 0.5 \sigma^2_u |\ln(Z_{i2})|) + 0.5 \sigma^2_u |\ln(Z_{i2})| + 0.5 \sigma^2_u |\ln(Z_{i2})| \exp(\beta_{i2}), \] where \( \beta_{i2}, \hat{\beta}_{i1}, \hat{\beta}_{i2} \) are the parameters estimated by random-parameter model.

\[ \hat{Y}_i = E(\lambda(\hat{Y})) = (1 + 0.5 \sigma^2_u |\ln(Z_{i2})|)^2 + 0.5 \sigma^2_u |\ln(Z_{i2})| + 0.5 \sigma^2_u |\ln(Z_{i2})| \exp(\beta_{i2}), \] where \( \hat{\beta}_{i1}, \hat{\beta}_{i2}, \sigma_1, \sigma_2, \) and \( \sigma_u \) are the parameters estimated by random-parameter model.

\[ \hat{Y}_i = \sqrt{\frac{1}{500} \sum_i (Y_i - \hat{Y}_i)^2}, \] whereas Res.RP.1 and Res.RP.2 were calculated as \( Y_i - \hat{Y}_i \) and \( Y_i - (1 + 0.5 \sigma^2_u |\ln(Z_{i2})|)^2 + 0.5 \sigma^2_u |\ln(Z_{i2})| + 0.5 \sigma^2_u |\ln(Z_{i2})| \exp(\beta_{i2}) \) in random-parameter models, respectively.

Fig. 1. CURE plots based on 10-fold cross-validation (for fixed-parameter models, Res.FP was computed as \( Y_i - (1 + 0.5 \sigma^2_u |\ln(Z_{i2})|) \), whereas Res.RP.1 and Res.RP.2 were calculated as \( Y_i - \hat{Y}_i \) and \( Y_i - (1 + 0.5 \sigma^2_u |\ln(Z_{i2})|)^2 + 0.5 \sigma^2_u |\ln(Z_{i2})| + 0.5 \sigma^2_u |\ln(Z_{i2})| \exp(\beta_{i2}) \) in random-parameter models, respectively).
means and variances of the random parameters for unbiased crash prediction. Compared with the simulation method¹ (Alogaili and Mannering, 2020; Islam et al., 2020; Hou et al., 2021), our approach is not only theoretically sound but also readily applicable without much extra computational cost. Based on a well-designed numerical experiment, we further demonstrated that simply using the means of random parameters to estimate the number of crashes for out-of-sample observations is essentially incorrect, which necessarily results in an underprediction of crash counts in the long run. This finding should be robust, particularly after adjusting for confounders arising from small sample sizes (Lord and Miranda-Moreno, 2008), misspecification of functional forms (Wang et al., 2020), and reporting of crashes (Imprailou and Quddus, 2019; Zeng et al., 2020), excess zeros (Lord et al., 2007; Liu et al., 2018), measurement errors in exposure (Xie et al., 2018; Kamed and Sayed, 2020), omitted variables (Mitra and Washington, 2012), spatial correlation (Xu et al., 2017; Dong et al., 2020; Ziaikopoulos and Yannis, 2020; Cui and Xie, 2021), and temporal instability (Mannering, 2018; Alnawmisi and Mannering, 2019; Behnood and Mannering, 2019; Islam et al., 2020), all of which are frequently encountered in empirical studies. Accordingly, the underperformance of random-parameter count models reported in previous studies (Tang et al., 2019; Hou et al., 2020) arises very likely from the convenient use of $\hat{\lambda}_y$ as the prediction. That is, the conclusion that random-parameter count models were less reliable and inaccurate for out-of-sample crash prediction reached by Tang et al. (2019) and Hou et al. (2020) is implausible and misleading, which should not be attributed to the modeling technique itself, but rather to the artificial and incorrect use of model estimates by analysts. Given the widespread use and ongoing prevalence of the random-parameter approach in crash analysis, special attention should be paid to this fundamental issue when applying it to predict crash frequencies for out-of-sample observations. For future research, since the data structures in empirical settings are much more complicated and intractable than anticipated (Mannering et al., 2020), a thorough comparison of the predictive performance of random-parameter count models with that of other techniques, such as the emerging deep learning methods (Bao et al., 2019; Cai et al., 2019), is highly recommended.

Funding
This study was supported by grants from the Guangdong – Hong Kong – Macau Joint Laboratory Program of the 2020 Guangdong New Innovative Strategic Research Fund, Guangdong Science and Technology Department (Project No. 2020B1212030009) and the Fundamental Research Funds for the Central Universities of Central South University (Project No. 2019zzts868). Pengpeng Xu and S.C. Wong were supported by the YS and Christabel Lung Postgraduate Scholarship and the Francis S Y Bong Professorship in Engineering, respectively. The funders had no interests or personal relationships that could have appeared to influence the work reported in this paper.

Author statement
Pengpeng Xu: Conceptualization, Methodology, Data curation, Software, Visualization, and Writing-original draft; Hanchu Zhou: Writing-review & editing; S.C. Wong: Supervision, Funding acquisition, and Writing-review & editing. We would like to thank Prof. Mannering Fred from University of South Florida and Prof. Huang Helai from Central South University for providing the insightful, thoughtful, and constructive comments that have helped us to substantially improve the quality and readability of our manuscript.

Declaration of Competing Interest
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References
Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. Geosci. Model Dev. 7 (3), 1247–1250.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning. Springer Press, NY.

¹ For more details, please refer to Hou et al. (2021) who used the Monte Carlo simulation to approximate $E(\hat{\lambda}_y)$. Similar simulated methods were also applied by Alogaili and Mannering (2020) and Islam et al. (2020) for out-of-sample probability prediction in injury-severity analysis.