Estimating the Conflict-Crash Relationship

Extended Summary

Andrew P. Tarko

Purdue University, Lyles School of Civil Engineering, Center for Road Safety,
West Lafayette, IN 47907, USA, tarko@purdue.edu

Abstract

This paper revisits methods of estimating the Lomax distribution for predicting crashes from observed traffic conflicts. After a short introduction of the current method of traffic conflicts caused by a perception error of a driver or machine, the paper presents the ML-based estimation of the Lomax distribution shape parameter hen the scale parameter is assumed. The derived formula accounts of right-sides censored data that represent observed crashes. Then, the paper compares the OLS method proposed in another publication with the ML estimation method based on numerical simulation of cases with various scale parameter assumed. The results indicate that both the methods are comparable and the assumption of distinct scale parameters affected the results to the limited extent that deems to be acceptable until a better method of estimating both the two parameters is available. The awaited estimation method of two parameters has been a subject of a number of published papers but no practical solution seems to exist yet. Limited simulation experiments presented in the paper to evaluate a recently published paper based on reparameterization of the Lomax distribution brought mixed results. Further research on the subject is needed.

Keywords: traffic conflicts, Lomax estimation, predicting crashes

1. Introduction

Quick estimation of safety and safety effects is becoming necessity with the growing rate at which road safety changes during the current period of emerging technological advancements in both automated driving and safety improvement solutions. For this reason, the fundamental matters of how traffic conflicts are connected to crashes and how to estimate this connection from traffic conflict data is an active subject of research and refinements. There are still open questions though about traffic events that can be analytically extrapolated to related crashes, and how to efficiently estimate the probability of crash associated with such events to enable conversion of observed events to the corresponding expected number of crashes. There are two important uses of any working estimation method: (1) rapid assessment of safety at specific roads locations and evaluation of countermeasures by safety engineers, (2) modeling of safety effects by analysts based on relatively short observations at multiple locations or at limited number of locations but during extended periods. This paper focuses on first use of traffic conflicts the rapid assessment of safety by proposing a ML method of estimating the expected number of crashes and by comparing it to the proposed recently OLS-based estimation [1]. There is a two-fold motivation of the analysis: (1) The potential concern with using the OLS method rather than the ML method, and (2) Reported in the past difficulties with estimating the Lomax distribution involved in the process.

This paper starts with a brief introduction of the current state of the art in the matter and then proposes the ML estimator. Next, the single-parameter estimation of the Lomax distribution parameters is scrutinize and, specifically, the effect of assuming an incorrect scale parameter on the results. Finally, the results of the analysis and summarized with a conclusion of the usability of the ML-based single-parameter method and promising further direction of research.

Pairs of interacting vehicles are tracked to estimate the shortest time to collision τ reached during an interaction. Observed τ value shorter than sufficiently short threshold τ_c associated with an uncharacteristically strong deceleration and jerk indicates a failure-cased traffic conflict. An observed reduction of time to collision below the threshold $x = \tau_c - \tau$ is called response delay. According to the counterfactual analysis, these delays follow a certain distribution with a long or even infinite tail estimable with observed delay values. The probability of a delay longer than the time left to a hypothetical collision is the probability of crash conditioned on the observed conflict. More details can be found in [2].

Under two rather weak assumptions, Lomax distribution [3] seems to be adequate to describe the variability of response delays *X* to a hazard (road edge, obstruction, another moving, etc.) in heterogeneous conditions:

$$f(x) = k\theta(1+\theta x)^{-k-1} \tag{1}$$

$$F(x) = 1-(1+\theta x)^{-k} \tag{2}$$

where: $f(x)$ and $F(x)$ are probability density and its cumulative value for response time X, while θ and k are the distribution parameters.

The probability of crash implied with a counterfactual analysis of conflicts is the probability of response delay *X* that exceeds the threshold time to collision $x_c = \tau_c$,

$$\bar{F}(x_c) = 1-F(x_c) = (1+\theta x)^{-k} \tag{3}$$

and the expected number of crashes is:

$$Q_c = n\cdot\bar{F}(x_c) = n\cdot(1+\theta x_c)^{-k} \tag{4}$$

where θ and *k* are the scale and shape parameters of Lomax distribution to be estimated with data. The existing literature reports difficulties in obtaining ML estimates of the scale parameter θ and the shape parameters *k*. Some authors proposed non-ML methods based on moments, quantiles, or Bayesian estimators [4-8]. These studies yielded estimates burdened with large variance. The source of this problem is more fundamental than the lack of a ML-based closed analytical solution often mentioned in the literature. The matter will be further discussed in this paper.

To avoid the mentioned estimation issue, a single-parameter estimation (SPE) method was proposed in [2] by setting the scale parameter θ at $1/x_c$ and calculating the shape parameter *k* as follows:

$$k = \frac{-\sum_{i=1}^{n}\log(1-(i-0.5)/n)\log(1+x_i/x_c)}{\sum_{i=1}^{n}[\log(1+x_i/x_c)]^2} \tag{5}$$

In spite of the indications of the SPE method' usefulness, the selection of scale parameter $\theta = 1/\tau_c$ is arbitrary and the use of OLS instead of ML method may also raise concerns. Some may argue that used θ parameters may be quite distant from the unknown true value. First, an ML estimator of *k* parameter with known scale parameter θ will be presented. The estimator accounts for the right-side truncation of response delays *X* at $x_c = \tau_c$ to properly include crashes that occur when observing conflicts. The OLS and ML estimators are re-applied to the SHRP2 data and the results are compared. Then, the effect of θ selection on the most important result – the expected number of crashes Q_c – is demonstrated by applying the k OLS and ML estimators to assumed values of θ. The effect of θ selection is discussed and alternative remedies discussed. Next, an alternative parametrization of Lomax distribution is proposed and the distribution's usefulness for estimating both the distribution parameters briefly discussed. Conclusions and further research needs finalize the paper.

2. ML Estimator of k Parameter for Right-Censored Data

Response delay is Lomax-distributed; its distribution has an infinite tail. Obviously, counterfactual response delays longer than x_c are not observable and observed *x* values must be treated as right-censored at x_c. Let assume that *n* traffic events: *c* crashes and $(n-c)$ traffic conflicts, are observed in the observation period under the x_c threshold $(x_c = \tau_c)$. With the distribution density of *X* according to equation (1) and the probability of observing crash according to equation (3), the log-likelihood *LL* of a sample with observed (*n*-*c*) response delays x and observed c crashes $(X > x_c)$ is:

$$LL = k\theta\sum_{i=1}^{n-c}\ln(1+\theta x_i)^{-k-1} + c\ln(1+\ \theta x_c)^{-k} \tag{6}$$

Setting $\partial LL/\partial k$ at zero and solving for *k* yields:

$$k = \frac{n-c}{\sum_{i=1}^{n-c}\ln(1+\theta x_i\ ) + c\ln(1+\theta x_c)} \tag{7}$$

Setting parameter θ at $1/x_c$ produces:

$$k = \frac{n-c}{\sum_{i=1}^{n}\ln(1+x_i/x_c\ ) + c\ln(2)} \tag{8}$$

3. Comparison of the ML and OLS Methods Applied to SHRP2 Data

The SHRP2 rear-end conflicts data [9] were re-used to estimate the rear-end crash rates with the proposed ML formula. The results are compared to the OLS-based estimates. The first conclusion is that the x_c thresholds obtained with the two estimation methods are the same in each of the considered driver cases. On the other hand, the ML method tends to produce slightly lower estimates of the parameter *k* which leads to lower bot the crash probabilities and the expected number of crashes when compared to the OLS results. Additional studies are needed if one wants to confirm if these observations indicate a systematic trend or it is just a coincident. Nevertheless, the differences between the results produced with the two different estimation methods are rather limited.

Both the methods require assuming scale parameter θ. The effect of this assumption on the results needs to be evaluated. After all, it is possible that a poorly selected θ parameter may considerably skew the results. Thus, the assumption of an acceptable effect of the θ selection on the results must be checked by evaluating the potential bias that may be present in the estimates. It must be emphasized that the primary interest of the presented here analysis is on exceedances that are equivalent to collision occurrence. This prompts that the focus of a Lomax distribution estimation should be less on the accuracy of the parameter estimates and more on the exceedance probability $\bar{F}(x_c)$ and the resulted expected number of crashes. The next section investigates these two effects when *k* parameter is estimated with the OLS and ML methods while the θ parameter is assumed.

Obviously, fair evaluation requires also consideration of methods that estimate both the parameters. Thus, the availability of such comprehensive methods will be discussed together with the sources of difficulties faced.

4. Simulation Experiments with ML and OLS Estimates

The case of young male drivers studied with SHRP2 data [9] is approximated numerically by simulating the Lomax distribution with the parameters similar to the ones reported in the mentioned publication. Specifically, 100 conflict observations were numerically generated by drawing them randomly from the Lomax distribution. Samples were drawn 200 times to allow statistical analysis of the results for each case. Parameter *k* was then estimated with the OLS and ML methods repeatedly for assumed eleven θ parameters.

The results revealed that under the assumed scale parameters θ, the 90% confidence interval of *k* estimate does not include the true *k* value (k=7.82) for seven out of 11 θ values tried. On the other hand, all the 90% confidence intervals of the expected crashes estimate Q_c include the true value (Q_c=0.441). This funding applies to both the ML and OLS estimation methods. The length of the 90% confidence interval of *k* estimate normalized with the average value is at or below 2 in all the studied cases.

Three sources of the variability in Q_c estimates were analyzed: different values of scale parameter θ, randomness, and differences between drivers. Variability among Q_c estimates caused by assumed parameter θ should be limited be able to claim that the SPE method is practical. To investigate this variability, samples of 100 Lomax-distributed response delays were drawn 200 times from the same population and the θ parameter was estimated 200 times. The average estimate has the randomness effect reduced to the level that may be ignored. This procedure was applied to several values of θ parameters ranging from 0.38 to 1.15. The differences between the maximum and minimum average Q_c estimates for the OLS and ML methods were compared to the values obtained for other sources of Q_c variability and the results are discussed in the last paragraph of this section.

The second source of variability is randomness. The spread of the estimate values is measured with the length of the 90% confidence interval. Again, samples of 100 response delays were drawn 200 times from the Lomax distribution with the assumed true parameter θ =.77 and the k parameter estimated with the OLS and ML methods. The results indicate that the spread of the estimates is much higher than in the previous case and it is slightly worse for the ML method.

Finally, the Q_c values obtained for the two investigated types of drivers: young males and mature females are compared. Tarko and Lizarazo [11/9] reported the estimated rear-end crash rates for young male drivers and mature female drivers responsively: 1086 and 137 crashes per 100 million miles of car following. The conducted here simulation experiments for young drivers used the Lomax parameters estimated in that study. Thus, the estimated Q_c for mature female drivers under the same exposure may be assessed as: 0.494·137/1086 = 0.0623 (ML method) and 0.527·137/1086=0.0665 (OLS method). The corresponding differences between the two types of drivers are: 0.432 and 0.460.

The results indicate a strong randomness of crash estimates. This result should not be a surprise since a similar uncertainty is present in crash counts during much longer period of data collection. On the other hand, assuming

a scale parameter from a relatively large range of values has a weaker effect on the results than the randomness. Nevertheless, this added variability makes obtaining conclusive results more challenging. This observation would prompt for trying to include the scale parameter among estimated ones. The published by other authors research on estimating both the Lomax parameters based on the ML-like methods indicates a large variance of the estimates in such attempts.

5. Reparametrized Lomax Distribution

Although the analytical analysis of the sample $LL(\theta, k)$ surface indicated that it was strictly concave [10], the crosscut of the sample LL surface along ML estimates of k values along the assumed θ values was almost invariant in the discussed here case. This indicates that there are infinite number of pairs of θ and k that correspond to sample LL values indistinguishable due to the observation errors and the limited accuracy of numerical optimization. As the result, multiple attempts to numerically maximize sample LL for various initial θ and k parameter values yielded solutions with the θ value unchanged from the initial one. More importantly, the exceedance probabilities at x_c (crash probability) were considerably different from estimation to estimation.

Furthermore, the OLS-based optimization (minimizing SS) produced a trivial solution at θ=0 in all the attempts. Figure 6 provides a clear explanation. In the light of the above findings, the Bayesian estimation method applied to the Lomax with θ and k parameters must be reevaluated from the point of view discussed here. It is critical to confirm that obtained solutions are independent of the starting point.

The problems with estimating the scale and shape parameters of Lomax probability function might motivated Altun [11] to reparametrize the Lomax probability function by replacing scale parameter θ with a location parameter μ (mean x) whose connection with the other parameters is:

$$\theta = \frac{1}{\mu(k-1)}. \tag{9}$$

The sample LL obtained with the reparametrized distribution turned out to be much more sensitive to the shape and location parameters. Limited numerical experiments with the reparametrized LL function, not presented here in detail due to the space limitation, indicated a potential of efficient estimating the two Lomax parameters: location and shape parameters.

Another potential avenue to explore is applying a more flexible distribution than negative exponential with unobserved Gamma heterogeneity. One option tis to replace the negative exponential distribution with the Weibull distribution [12]. The Weibull distribution includes an additional parameter that relaxes the assumption of a fixed response rate enforced by the negative binomial distribution.

6. Conclusions

ML estimates of k parameter are consistent if the assumed θ parameter is correct, while OLS estimates tend to be slightly overestimated. On the other hand, ML estimates exhibits larger bias in estimating k parameters than OLS if assumed k is incorrect. It should be noted that estimates of the conditional probability of crash are more important than correctly estimated distribution parameters. Also, estimation efficiency is more important than consistency. In most current applications, samples used to estimate the crash probability are limited while the estimation error should be limits to make the results useable.

The above comments are particularly adequate when the reparametrized distribution with location and shape parameters is considered. Although the location parameter is closely approximated, the shape parameter estimates exhibit strong estimation error. And yet, the estimated probability of crash (exceedance probability) is estimated with reasonable accuracy that makes the results applicable. Also, the OLS estimates seem to be more practical than ML when the scale parameter must be assumed. Although there is an estimation limited inconsistency even when the scale parameter is correct, the estimation accuracy is higher them ML estimator when the scale parameter is incorrect.

These comments are applicable to the early phase of safety analysis when traffic conflicts are expected to available at lower numbers. With time and development of the safety estimation methods and proliferation of machine-learning algorithms for traffic perception, the massive data available for safety estimation and its modeling will increase the importance of estimation consistency.

The presented here research focused on estimation of the distribution parameters without reaching for regression analysis. Incorporating regression variables and alternative methods of models' estimation is a growing and needed area of surrogate measures of safety. It must be stressed that use traffic events that are indeed safety relevant has fundamental importance for the results validity. This element of research should also continue.

It was demonstrated that neither of the two discussed estimation methods, ML and OLS, can deliver estimates of the two Lomax parameters. The sample *LL* function of the two parameters is practically flat while the minimizing the OLS function generates a trivial solution that is obviously incorrect. The past attempts of resolving the estimation issues did not delivered satisfactory results. Relatively recent work on correcting biased ML estimates of Lomax two parameters is worth of attention but it awaits scrutiny in the light of safety estimation with traffic conflicts.

References

1. Tarko, A. P. (2018). Estimating the expected number of crashes with traffic conflicts and the Lomax Distribution – A theoretical and numerical exploration. *Accident Analysis and Prevention*. Vol 113, Pages 63-73. https://doi.org/10.1016/j.aap.2018.01.008
2. Tarko, A.P. (2020). *Measuring Road Safety with Surrogate Events*. Elsevier. Amsterdam, Oxford, Cambridge.
3. Lomax, K. S. (1954). Business Failures; Another example of the analysis of failure data. *Journal of the American Statistical Association*, 49, 847–852.
4. Greenwood J.A., Landwehr, J.M., Matalas, N.C.,Wallis, J.R. (1979). Probability weighted moments: Definition and relation to parameters of several distributions expressible in inverse form. *Water Resources Research*, 15(5), 1049-1054.
5. Hosking, J.R.M., Wallis, J.R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics* 29, 339–349.
6. Castillo, E., Hadi, A.S. (1997). Fitting the generalized Pareto distribution to data. *Journal of the American Statistical Association*, 92, 1609-1620.
7. Zhang, J., Stephens, M.A. (2009). A new and efficient estimation method for the generalized Pareto distribution. *Technometrics*, 51(3), 316-325.
8. Shakeel, M., Rehmat, N., Ul Haq, M.A. (2017). Comparison of the robust parameters estimation methods for the two-parameter Lomax distribution. *Cogent Mathematics* (2017), 4: 1279397. Available at: http://dx.doi.org/10.1080/23311835.2017.1279397. Accessed on 12/9/2021.
9. Tarko, A.P, Lizarazo, C.G. (2021). Validity of failure-caused traffic conflicts as surrogates of rear-end collisions in naturalistic driving studies. *Accident Analysis and Prevention*, Vol 149. https://www.sciencedirect.com/science/article/pii/S0001457520316833
10. Giles, D.E., Feng, H., Godwin, R. T. (2013). On the Bias of the Maximum Likelihood Estimator for the Two-Parameter Lomax Distribution, Communication in Statistics - Theory and Methods, 42(11), https://doi.org/10.1080/03610926.2011.600506 (accessed on 12/9/2021).
11. Altun, E. (2021) The Lomax regression model with residual analysis: an application to insurance data, Journal of Applied Statistics, 48:13-15, 2515-2524, https://doi.org/10.1080/02664763.2020.1834515 (accessed on 12/9/2021).
12. Fréchet, Maurice (1927), "Sur la loi de probabilité de l'écart maximum", Annales de la Société Polonaise de Mathématique, Cracovie, 6: 93–116.