Ordinal Regression with Fenton-Wilkinson Order Statistics: A Case Study of an Orienteering Race

Joonas Pääkkönen

Abstract—In sports, individuals and teams are typically interested in final rankings. Final results, such as times or distances, dictate these rankings, also known as *places*. Places can be further associated with ordered random variables, commonly referred to as *order statistics*. In this work, we introduce a simple, yet accurate order statistical ordinal regression function that predicts relay race places with changeover-times. We call this function the *Fenton-Wilkinson Order Statistics* model. This model is built on the following educated assumption: individual leg-times follow log-normal distributions. Moreover, our key idea is to utilize Fenton-Wilkinson approximations of changeover-times alongside an estimator for the total number of teams as in the notorious German tank problem. This original place regression function is sigmoidal and thus correctly predicts the existence of a small number of elite teams that significantly outperform the rest of the teams. Our model also describes how place increases linearly with changeover-time at the inflection point of the log-normal distribution function. With real-world data from Jukola 2019, a massive orienteering relay race, the model is shown to be highly accurate even when the size of the training set is only 5% of the whole data set. Numerical results also show that our model exhibits smaller place prediction root-mean-square-errors than linear regression, mord regression and Gaussian process regression.

Keywords—Fenton-Wilkinson approximation, German tank problem, log-normal distribution, order statistics, ordinal regression, orienteering, sports analytics, sports modeling.

I. Introduction

CLASSIFICATION refers to machine learning methods where the target variable is a discrete class, while regression is typically associated with continuous variables. However, when the number of classes is large, yet discrete, it becomes challenging to make a distinction between classification and regression. *Ordinal regression*, also known as *ordinal classification*, refers to regression with a target that is discrete and ordered. It can thus be regarded as a hybrid mixture of both classification and regression.

Typical applications of ordinal classification include age estimation with an integer-valued target, advertising systems, recommender systems, and movie ratings. For additional insight into recent developments of related machine learning methods, the reader is kindly directed, *e.g.*, to [1] for a survey on ordinal regression, and to [2] for a survey on deep learning.

Ordinal regression lends itself especially well to ordered sets. In sports, all result lists are ordered sets with respect to results such as times, distances or points. Thus, for a given result, ordinal regression could predict the final rankings, *i.e.*, the *places* of teams or individual athletes. Here we conduct a case study of ordinal regression on the ranks of sorted sums of random variables of the duration of a relay. To be more specific, we study a large number of realizations of the changeover-times of an orienteering relay race.

We compare three widely-used regression schemes to an original ordinal classification method, the derivation of which is attributed to algebraic manipulations and well-known results concerning ordered random variables. Such random variables are known as *order statistics*, and they represent a branch of mathematical statistics closely related to *extreme value theory* (EVT). While sports analytics has seen several EVT applications for record values [3], [4], in this work we do not focus on extreme values but rather order statistics in general.

As an underpinning educated assumption, we say that individual leg-times are log-normal. We furthermore assume the log-normality of changeover-times, which is due to the Fenton-Wilkinson approximation [5]–[7]. We also note that while there exist explicit expressions for the expectations of log-normal order statistics [8], for our purposes these expressions are unnecessary as scaling the log-normal distribution function directly produces a place predictor.

According to the principle of maximum entropy [9], one could argue that the amount of uncertainty in the relay system increases with time, and that changeover-times thus tend to follow a maximum entropy distribution, such as the log-normal distribution. In practice, though, the log-normality assumption follows from the observation that marathon finish-times exhibit log-normality [10]. We show that a log-normal shape also fits orienteering data, which is to be expected given that both marathon running and orienteering are endurance sports.

Unlike prediction models for individual marathon race finish-times [11], [12], here we consider orienteering relay team place prediction. As a distinctive element of our work, rather than predicting times, we are interested in predicting places. It is often the place that is the hard, quantitative result that many teams wish to minimize. Thus, place prediction is of particular interest.

The main contributions of this work are the introduction and the validation of what we refer to as the Fenton-Wilkinson Order Statistics (FWOS) model. For a case study of an orienteering race with real-world data, numerical results show that FWOS accurately predicts places even with very few training examples. Further, FWOS plots correctly illustrate that place increases sigmoidally with changeover-time.

II. System Model

Consider an orienteering relay race. Let n denote the number of finishing teams as we ignore disqualified and retired teams. There are m runners on each team and each runner runs one *leg*. Each leg is immediately followed by another at a *changeover* until the end of the relay.

J. Pääkkönen is with the Department of Informatics, School of Technology and Business Studies, Dalarna University, Börlänge, Sweden (e-mail: jpa@du.se).

ISSN: 2692-5079

The log-normal c.d.f. is

$$F_{T^{(l)}}(t) = \Phi\left(\frac{\log t - \mu_l}{\sigma_l}\right), \tag{9}$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{\tau^2}{2}\right) \mathrm{d}\tau$$

is the standard normal c.d.f., and μ_l and σ_l are the log-normal parameters.

We plug (8) and (9) into (7) and, after rounding, arrive at

$$\left[\Phi\left(\frac{\log t - \mu_l}{\sigma_l}\right)(n+1)\right] \approx r, \tag{10}$$

where $[\cdot]$ denotes rounding to the nearest integer.

Maximum likelihood estimation (MLE) for the normal distribution yields log-normal estimators for μ_l and σ_l as

$$(\hat{\mu_l}, \hat{\sigma_l}) = \left(\frac{1}{c}\sum_{i=1}^{c} q_i^{(l)}, \sqrt{\frac{1}{c}\sum_{i=1}^{c}\left(q_i^{(l)} - \hat{\mu_l}\right)^2}\right) \tag{11}$$

by setting $q_i^{(l)} := \log t_i^{(l)}$.

What remains to be done is finding an estimate for the total number of teams n to estimate the scaling factor $(n+1)$ in (10). We assume that there are no ties, which is equivalent to stating that the elements in the training set $\boldsymbol{r}$ are unique. Thus, $\boldsymbol{r}$ is a length-c sample, without replacement, of the discrete uniform distribution $\mathcal{U}[1, n]$.

Now recall that $\boldsymbol{r}$ corresponds to an ordered B_c (an ordered proper c-subset of $\mathbb{N}_n$). Let D denote a random variable that follows $\mathcal{U}[1, n]$. Estimating the parameter n of $\mathcal{U}[1, n]$, with a sample drawn without replacement, is in the literature known as the *German tank problem* [21]. A uniformly minimum-variance unbiased estimator (UMVUE) for this parameter is given in [22] as

$$\hat{n} = \left(1 + \frac{1}{c}\right) r_{(c)} - 1, \tag{12}$$

where

$$r_{(c)} = \max_{i \in \mathbb{N}_c} r_i$$

is the realization of the c^{th} order statistic (maximum) of a length-c sample of D.

We plug the pair $(\hat{\mu_l}, \hat{\sigma_l})$ of (11) into (10). We plug (12) into the n of (10). This concludes the derivation.

References

[1] P. A. Gutierrez, M. Perez-Ortiz, J. Sanchez-Monedero, F. Fernandez-Navarro, and C. Hervas-Martinez, "Ordinal regression methods: Survey and experimental study," *IEEE Trans. Knowl. and Data Eng.*, vol. 28, no. 1, pp. 127–146, 2016.

[2] M. Raghu and E. Schmidt. (2020, March) A survey of deep learning for scientific discovery. [Online]. Available: arXiv:2003.11755

[3] M. Strand and D. Boes, "Modeling road racing times of competitive recreational runners using extreme value theory," *Am. Stat.*, vol. 52, no. 3, pp. 205–210, 1998.

[4] H. Spearing, J. A. Tawn, D. B. Irons, T. Paulden, and G. A. Bennett. (2020, June) Ranking, and other properties, of elite swimmers using extreme value theory. [Online]. Available: arXiv:1910.10070

[5] L. F. Fenton, "The sum of log-normal probability distibutions in scattered transmission systems," *IRE Trans. Commun. Syst.*, vol. 8, pp. 57–67, 1960.

[6] R. I. Wilkinson, "Unpublished, cited in 1967," *Bell Telephone Labs*, 1934.

[7] B. R. Cobb, R. Rumí, and A. Salmerón, "Approximating the distribution of a sum of log-normal random variables," in *Proc. 6th Eur. Workshop Probab. Graph. Models*, 2012, pp. 67–74.

[8] S. Nadarajah, "Explicit expressions for moments of log normal order statistics," *Economic Quality Control*, vol. 23, no. 2, pp. 267–279, 2008.

[9] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, vol. 106, no. 4, pp. 620–630, 1957.

[10] E. J. Allen, P. M. Dechow, D. G. Pope, and G. Wu, "Reference-dependent preferences: Evidence from marathon runners," *Manag. Sci.*, vol. 63, no. 6, pp. 1657–2048, 2017.

[11] D. Ruiz-Mayo, E. Pulido, and G. Martíñoz, "Marathon performance prediction of amateur runners based on training session data," in *Proc. Mach. Learn. and Data Min. for Sports Anal.*, 2016.

[12] J. Esteve-Lanao, S. D. Rosso, E. Larumbe-Zabala, C. Cardona, A. Alcocer-Gamboa, and D. A. Boullosa, "Predicting recreational runners' marathon performance time during their training preparation," *J. Strength Cond. Res. doi: 10.1519/JSC.0000000000003199 [Epub ahead of print]*, 2019.

[13] K. A. Wang, G. Pleiss, J. R. Gardner, S. Tyree, K. Q. Weinberger, and A. G. Wilson, "Exact gaussian processes on a million data points," in *Proc. Adv. Neural Inf. Process. Syst. 32*, 2019, pp. 14 648–14 659.

[14] C. E. Rasmussen and C. K. I. Williams, "Gaussian processes for machine learning," *The MIT Press*, 2006.

[15] Gpytorch regression tutorial. [Online]. Available: https://gpytorch.readthedocs.io/en/latest/examples/01_Exact_GPs/Simple_GP_Regression.html

[16] Mord: Ordinal regression in python. [Online]. Available: https://pythonhosted.org/mord/

[17] F. Pedregosa-Izquierdo, "Feature extraction and supervised learning on fmri: from practice to theory," Ph.D. dissertation, Université Pierre-et-Marie-Curie, 2015.

[18] Jukola 2019. [Online]. Available: https://results.jukola.com/tulokset/en/j2019_ju/

[19] E. Limpert, W. A. Stahel, and M. Abbt, "Log-normal distributions across the sciences: Keys and clues," *Bioscience*, vol. 51, pp. 341–352, 2001.

[20] P. Chen, R. Tong, G. Lu, and Y. Wang, "Exploring travel time distribution and variability patterns using probe vehicle data: Case study in beijing," *J. Adv. Transp.*, pp. 1–13, 2018.

[21] R. Ruggles and H. Brodie, "An empirical approach to economic intelligence in world war ii," *J. Am. Stat. Assoc.*, vol. 42, no. 237, pp. 72–91, 1947.

[22] L. A. Goodman, "Serial number analysis," *J. Am. Stat. Assoc.*, vol. 47, no. 270, pp. 622–634, 1952.

Joonas Pääkkönen received the MSc and PhD degrees in communications engineering from Aalto University, Finland, in 2012 and 2018, respectively. During his graduate studies, he worked as part of the Communications Theory Research Group under Prof. Olav Tirkkonen and as part of the Algebra, Number Theory and Applications Research Group under Prof. Camilla Hollanti. Later, he worked as the WOC Team Coach on the National Team of Orienteering Canada in 2019.

His research interests include wireless communications, distributed storage, coding theory, probability, and mathematical statistics. More recently, his research interests have expanded to numerical methods in sports science, as well as computational sports analytics, athlete training program planning, performance analysis and recovery analysis.

Dr. Pääkkönen currently works as a researcher and lecturer at the Department of Informatics, School of Technology and Business Studies at Dalarna University, Borlänge, Sweden.