IR Systems Evaluation via Generalized Linear Models*

Discussion Paper

Guglielmo Faggioli[1], Nicola Ferro[1] and Norbert Fuhr[2]

[1]University of Padova, Padova, Italy
[2]University of Duisburg-Essen, Duisburg, Germany

Abstract

Being able to compare Information Retrieval (IR) systems correctly is pivotal to improving their quality. Among the most popular tools for statistical significance testing, we list t-test and ANOVA that belong to the linear models family. Therefore, given the relevance of linear models for IR evaluation, a great effort has been devoted to studying how to improve them to better compare IR systems. Linear models rely on assumptions that IR experimental observations rarely meet, e.g. about the normality of the data or the linearity itself. Even though linear models are, in general, resilient to violations of their assumptions, departing from them might reduce the effectiveness of the tests. Hence, we investigate the use of the Generalized Linear Models (GLMs) framework, a generalization of the traditional linear modelling that relaxes assumptions about the distribution and the shape of the models. We discuss how GLMs can be applied in the context of IR evaluation. In particular, we focus on the *link function* used to build GLMs, which allows for the model to have non-linear shapes.

1. Introduction

Evaluation in *Information Retrieval (IR)* allows researchers and practitioners to study and compare their systems in order to understand how to improve them. To this end, sound statistical inference methods are needed to obtain robust and generalizable insights and to predict what happens when systems run in a real-world scenario. Therefore, statistical analyses, such as bootstrap, randomization tests [2, 3], t-tests, and *ANalysis Of VAriance (ANOVA)* [4, 5, 6] have been widely studied and successfully employed in IR evaluation.

In particular, t-test and ANOVA belong to the family of statistical methods called *General Linear Models (GLiMs)*, a generalization of the multiple linear regression, which is based on the following assumptions: *i)* independence of the observations, *ii)* constant variance of the data, i.e. *homoscedasticity iii)* normal distribution of the data, i.e., *normality*; last but not least and too often overlooked: *iv)* linear correlation between experimental conditions and the expectation of the response i.e., *linearity*. These assumptions allow for an analytical solution of the model and its practical computation. Furthermore, the more such assumptions are satisfied, the more accurate is the estimation of the model and the inferences drawn from it. Previous literature showed great interest in studying the empirical consequences of using data violating such assumptions, both from a theoretical standpoint [7, 8], and also considering empirical IR data [5, 9, 10, 11]. Such works show that, in general, linear models are resilient to the violation

IIR2023: 13th Italian Information Retrieval Workshop, 8th - 9th June 2023, Pisa, Italy

*This is an extended abstract of [1].

of their assumptions. At the same time, several works have explored how to make IR data closer to the GLiM assumptions, e.g. by transforming the data [5, 10]. In all the cases, the ultimate goal is to obtain models which are capable to better and more reliably distinguish among IR systems. The *Generalized Linear Model (GLM)* framework is a generalization of the GLiMs that relaxes some of the underlying assumptions to increase models' applicability. In particular, the data is no longer required to follow a normal distribution or to have constant variance. Moreover, GLMs also relax the fourth assumption, allowing the link between the response and the experimental conditions to have different forms besides the linear one. In this work, we investigate the application of Generalized Linear Models to IR evaluation and show how they can help us in better comparing and distinguishing among systems.

2. Methodology

Parametric statistical tests, such as t-tests or ANOVA, rely on the assumption that data can be modelled using a linear model. Given a system s and a topic t, we can compute a measure, e.g. *Average Precision (AP)*, that quantifies how well s performs on t. We refer to such a score as y_{ts} and call it *response*. y_{ts} is a realization of a random variable Y. The experimental conditions – i.e., topics and systems – are correlated with the response, thus called *covariates*. Using traditional linear models, the expectation of the response $E[Y]$, is modeled as a linear combination η of the covariates (linearity) as follows: $E[Y] = \eta = \mu + \tau_1 t_1 + ... + \tau_n t_n + \alpha_1 s_1 + ... + \alpha_m s_m$, where t_i and s_j are respectively the dummy coding variables for the topic and systems considered, τ_i is the effect due to the i-th topic, α_j is the effect due to the j-th system. The intercept μ represents the grand mean of our data. To compute the linear model and grant its inferences, we assume $Y \sim \mathcal{N}(\eta, \sigma^2)$ – Y distributes normally (normality) and has the same variance σ^2 everywhere (homoscedasticity). Without losing generality, we can say that we model $g(E[Y]) = \eta$, where g is the identity function $g(x) = x$. In this sense, g is the function that *links* $E[Y]$ to η. Summing up, fitting a linear model requires defining the following elements: 1. a linear combination η of the different explanatory variables; 2. a *link* function g to connect $E[Y]$ to η; 3. a distribution for Y. Compared to a traditional linear model, a GLM relaxes the assumptions for items 2 and 3. First, it models $g(E[Y])$ where g, the *link*, can be any monotonic continuous function. Secondly, the response Y can follow a distribution $f(\theta)$ that is not necessarily Gaussian. The *homoscedasticity* assumption is relaxed as well since the variance can change with the expected mean. Thus, a GLM can be expressed in the following form:

$$g(E[Y]) = \eta, \text{ with } Y \sim f(\theta)$$

The chosen probability distribution $f(\theta)$ must be a member of the exponential distributions family. A location parameter θ characterizes distributions belonging to the exponential family - e.g., the normal distribution's mean. If we observe that $g(E[Y]) = \theta$ for a given distribution of Y, then we say that g is the *canonical link* of such a distribution.

We are now interested in understanding the effect on the IR evaluation, switching from the traditional evaluation based on linear models to considering GLMs. As pointed out in the previous section, to fit GLMs, we need to select the link function and the response distribution. In this work, we focus only on the effect that the link function has on the IR evaluation. Any

Table 1
Deviance. The color indicates optimal (green), average (white) or low (red) results.

	Robust 04					Core 18					Core 18-wo				
link	AP	P@10	Recall	nDCG	RBP	AP	P@10	Recall	nDCG	RBP	AP	P@10	Recall	nDCG	RBP
identity	356.79	901.51	622.67	467.98	381.62	52.48	156.32	99.05	69.21	75.90	44.93	135.61	74.46	59.89	63.23
log	334.06	886.42	646.34	471.87	368.30	40.90	**132.00**	91.57	61.73	**59.00**	40.76	**128.12**	75.36	60.18	**57.76**
exp	387.52	955.56	719.93	505.65	400.98	61.31	219.01	159.33	90.92	98.80	49.19	160.79	79.90	64.29	72.55
tanh	348.91	894.19	640.07	467.54	376.76	50.30	147.13	95.89	66.00	71.15	43.81	132.13	75.84	59.80	61.00
logit	**329.46**	**882.14**	593.82	**458.04**	**366.89**	**40.50**	132.77	**85.47**	**60.51**	59.44	**40.36**	128.52	72.21	**58.93**	58.12
probit	330.36	882.66	**590.87**	458.05	367.65	40.71	133.26	85.52	60.63	59.81	40.57	128.84	**72.13**	58.99	58.42
cauchit	332.49	884.67	627.34	463.81	367.40	40.87	132.24	86.80	60.63	59.03	40.73	128.42	74.05	59.05	57.81

possible monotonic continuous function can be a suitable link. The choice of which link to use depends on the shape of the data. Therefore, we try to empirically determine the best link for IR data. In particular, we include the log, exponential and hyperbolic tangent (tanh) functions in our experiments. We also experiment with a series of sigmoidal functions: logit, probit and cauchit. Previous works on transforming AP [10, 12, 13, 14] observed that the logit transformation renders the score distribution more normal but it has the drawback of making observations for which AP is zero or one unusable. GLMs based on the logit link avoid such corner cases. However, when even the expectation of a system's performance is close to zero, using log-based links – e.g., log and logit – determines a high variance of the coefficients associated with such a system. As a consequence a larger variance increases the standard error. It is therefore advisable to remove outliers with close-to-zero expected performance.

3. Experimental Analysis

In our experimental analysis, we consider two collections for *ad-hoc* retrieval: TREC 13 Robust 04 [15] and TREC 27 Core 18 [16]. Systems mean performance is very close to 0 can challenge the log-based links. Since this happens in the case of Core 18, we also consider a second version of it, where we remove eight outlier runs, performing extremely low in terms of MAP.

The most common goodness-of-fit statistics under the GLM framework is the *deviance* [17], which is analogous to *sum of squares of residuals (RSS)* under the GLiM framework. Table 1 illustrates the deviance measured for different GLMs using several link functions, IR measures, and experimental collections. The traditional GLiM approach based on the *identity link*, (corresponding to the current evaluation methodology) presents a low goodness-of-fit, given its high deviance compared to other links. This evidence supports the idea of investigating and using GLMs instead. The *exponential link* is the worst, systematically underperforming on all experimental conditions and its high deviance indicates poor goodness-of-fit compared to all the other links. Its poor capability in fitting IR data leads to overall instability, especially concerning shallow performing systems, and to convergence problems when fitting the model – highlighted by the increased iterations to reach convergence, not reported here due to space constraints. The *log link* shows improved goodness-of-fit compared to identity one, especially for the Core 18 collection (both with and without outliers). The *tanh link* exhibits an intermediate behaviour in all scenarios: it appears slightly better than the identity without providing substantial improvements. Finally, The *logit link* has the best goodness-of-fit in most cases, achieving the lowest deviance. Logit, *probit* and *cauchit links* tend to perform quite similarly.

Table 2 contains the number of *statistically significantly different (ssd)* pairs detected by the GLMs based on different links, using Tukey's HSD [18] test. On Robust 04 collection, all the links outperform the traditional modelling strategy – i.e., identity link – using AP, P@10, and RBP as performance measures. Logit is the best-performing link: it detects 7.9%, 8.9%, and 7.0% more ssd pairs compared to identity, when regarding AP, P@10, and RBP, respectively. On the other hand, considering Recall and *Normalized Discounted Cumulated Gain (nDCG)*, log and tanh fail to identify more pairs than identity, while logit, probit and cauchit increase the number of ssd pairs found. Concerning the Core 18 collection, Table 2 show that when log-based links are used in presence of low-performing outliers, they tend to underperform compared to the identity link. In particular, we observe that log, logit and cauchit almost always fail to outperform the identity baseline. Indeed, in our specific case, almost all the ssd pairs lost against the identity link correspond to the eight outlier runs – 8 runs appear in 540 pairwise comparisons. Notice that these runs have extremely low mean performance, being their MAP between 0.003 and 0.007. Probit outperforms the identity on all the measures except AP. This might be due to the shape of the link functions. The cauchit function is the steepest and thus the most vulnerable to outliers. The logit function has intermediate steepness, exhibiting medium vulnerability to outliers. Finally, probit is the least steep and the more resilient to outliers. If we consider Core 18-wo collection all the new links obtain a consistent improvement over the identity baseline for what concerns AP, P@10, and RBP. Logit and probit links are the best, gaining 17.4% new pairs on the AP. Logit, probit and cauchit perform well also with Recall and nDCG. Similarly to Robust 04, both tanh and log links lose several ssd pairs with respect to identity when using Recall and nDCG.

4. Conclusions and Future Work

We studied GLMs, an extension of the traditional linear models typically used in IR evaluation to compare systems. GLMs overcome the main reasons of departure of IR data from assumptions underlying linear models: non-normality and heteroscedasticity of the data and non-linearity of the empirical mean. In this work, we focused on the latter and studied how to address it using different *link functions*. We observed that log, logit, tanh, prob, and probit provide general improvements concerning the identity link used today. We then dug into the log and logit links, which were the most promising ones, and we found out that they can detect a sizeably greater number of consistent ssd pairs than the identity link. In future work, we plan to consider different distributions, to deal with the non-normality and heteroscedasticity of the data.

Table 2
statistically significantly different pair found. The color indicates good (green), sota (white) or bad (red) results.

	robust 04 - (5995 systems pairs)					core 18 - (2556 systems pairs)					core 18-wo - (2016 systems pairs)				
link	AP	P@10	Recall	nDCG	RBP	AP	P@10	Recall	nDCG	RBP	AP	P@10	Recall	nDCG	RBP
identity	3427	2347	3848	3704	2837	1210	1054	1115	1270	1247	789	596	427	786	803
log	3556	2383	3622	3550	2946	925	934	672	1097	1220	878	635	384	748	941
tanh	3509	2354	3639	3641	2905	**1301**	**1130**	1086	1283	**1361**	843	633	380	766	892
logit	**3700**	**2557**	4018	**3773**	**3035**	976	1034	1267	1251	1257	**926**	**713**	594	818	929
probit	3693	2541	**4027**	3766	3034	974	1079	**1304**	**1340**	1341	**926**	710	**597**	815	928
cauchit	3682	2552	3929	3764	3016	848	739	877	872	968	796	**713**	**597**	**823**	**937**

References

[1] G. Faggioli, N. Ferro, N. Fuhr, Detecting significant differences between information retrieval systems via generalized linear models, in: M. A. Hasan, L. Xiong (Eds.), Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022, ACM, 2022, pp. 446–456. URL: https://doi.org/10.1145/3511808.3557286. doi:10.1145/3511808.3557286.

[2] T. Sakai, Evaluating Evaluation Metrics Based on the Bootstrap, in: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '06, 2006, p. 525–532.

[3] M. D. Smucker, J. Allan, B. Carterette, A Comparison of Statistical Significance Tests for Information Retrieval Evaluation, in: Proceedings of the 16th ACM Conference on Information and knowledge Management, CIKM '07, 2007, pp. 623–632.

[4] A. Rutherford, Introducing ANOVA and ANCOVA: a GLM approach, Sage, 2001.

[5] J. M. Tague-Sutcliffe, J. Blustein, A Statistical Analysis of the TREC-3 Data, in: Proceedings of The 3rd Text REtrieval Conference, TREC '94, 1994, pp. 385–398.

[6] D. Banks, P. Over, N.-F. Zhang, Blind Men and Elephants: Six Approaches to TREC Data, Information Retrieval Journal (IRJ) 1 (1999) 7–34.

[7] P. Ito, 7 robustness of anova and manova test procedures, in: Analysis of Variance, volume 1 of *Handbook of Statistics*, 1980, pp. 199–236.

[8] S. M. Scariano, J. M. Davenport, The Effects of Violations of Independence Assumptions in the One-Way ANOVA, The American Statistician 41 (1987) 123–129.

[9] D. Hull, Using statistical testing in the evaluation of retrieval experiments, in: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '93, 1993, p. 329–338.

[10] G. V. Cormack, T. R. Lynam, Statistical Precision of Information Retrieval Evaluation, in: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '06, 2006, p. 533–540.

[11] B. Carterette, Multiple Testing in Statistical Analysis of Systems-based Information Retrieval Experiments, ACM Transactions on Information Systems (TOIS) 30 (2012) 1–34.

[12] S. E. Robertson, E. Kanoulas, On Per-Topic Variance in IR Evaluation, in: Proceedings of the 33rd ACM SIGIR Conference on Research and Development on Information Retrieval, SIGIR '12, 2012, p. 891–900.

[13] S. Robertson, On Smoothing Average Precision, in: Advances in Information Retrieval, 2012, pp. 158–169.

[14] A. Berto, S. Mizzaro, S. Robertson, On Using Fewer Topics in Information Retrieval Evaluations, in: Proceedings of the 2013 Conference on the Theory of Information Retrieval, ICTIR '13, 2013, p. 30–37.

[15] E. M. Voorhees, Overview of the TREC 2004 Robust Retrieval Track, in: Proceedings of The 13th Text REtrieval Conference, TREC '13, 2004.

[16] J. Allan, D. K. Harman, E. Kanoulas, E. M. Voorhees, TREC 2018 Common Core Track Overview, in: The Twenty-Seventh Text REtrieval Conference Proceedings (TREC 2018), 2019.

[17] P. McCullagh, J. A. Nelder, Generalized Linear Models, Springer, 1989. URL: https://doi.

org/10.1007/978-1-4899-3242-6. doi:10.1007/978-1-4899-3242-6.
[18] J. W. Tukey, Comparing Individual Means in the Analysis of Variance, Biometrics (1949) 99–114.