Current Trends in Biomedical Engineering & Biosciences
ISSN: 2572-1151

Juniper Publishers
key to the Researchers

Opinion
Volume 1 Issue 1 - December 2016
DOI: 10.19080/CTBEB.2016.01.555551

Curr Trends Biomedical Eng & Biosci

Can Randomness be an Element in Multiple Regression

Richard Boire*

Boire Filler Group, Canada

Submission: November 25, 2016; **Published:** December 07, 2016

***Corresponding author:** Richard Boire, Boire Filler Group, 1101 Kingston Rd, Pickering, ON L1V 1B5, Canada, Email: richard.boire@environicsanalytics.ca

Opinion

In some of the more recent literature and discussions, discussion has ensued about the use of pure random or noise variables that end up as key regression variables. In our big data environment with millions of records and thousands of variables, intuitively one might think that random or spurious variables might be a normal outcome in many models. As a data miner, I am always intrigued by fellow colleagues who arrive at certain findings based on their research and work. When considering the validity of these comments, I harken back to my experience in building hundreds of models over the years in a variety of different industry sectors. Practitioners typically adopt techniques that ensure the unlikelihood of these variables in any final regression equation. These techniques will be discussed later on in this article.

But in initially considering how I would either refute or support this hypothesis, I initially started my exploration from a "small" data standpoint. I created a series of datasets where each dataset had a different number of records. Each dataset had four variables where the values of all 4 variables were randomly created using a random number generator. Within each dataset, I designated a fifth variable as the target variable. Listed below are the p-values of each independent variable for both the development and validation datasets where each scenario represents a different volume of observations.

In the above table, you will observe that indeed there are random variables which appear to be significant as denoted by the highlighted cells in yellow. Under four of the five # of records scenarios, a random variable appears significant at a 95% confidence interval. In the 1000 record scenario and the 5000 record scenario, the significant variable appears in development while in the 10000 record scenario and the 50000 record scenario, the variable appears in validation. Within the 500000 record scenarios, we observe that no variable is significant in either development or validation. The real finding from all these results is that no variable is significant in both development and validation under any scenario. No real trend or finding emerges whether we look at either smaller datasets or larger datasets.

These results do not necessarily refute the findings in the literature. Instead these results point to a more disciplined approach in determining whether variables are random or spurious. The need for validation, validation, and more validation is just simply reinforced as one task in trying to eliminate random variables from any predictive analytics solution (Table 1).

Table 1: Predictive analytics solution.

Variable	1000	Records	5000	Records	10000	Records	100000	Records	500000	Records
	dev	val	dev	val	dev	val	dev	val	dev	val
random variable 1	0.8429	0.7888	0.0432	0.9883	0.7675	0.22	0.2093	0.5614	0.195	0.6028
random variable 2	0.4135	0.3405	0.2105	0.5806	0.6663	0.665	0.5215	0.0106	0.4103	0.9587
random variable 3	0.0128	0.9279	0.3278	0.5457	0.8177	0.0249	0.505	0.1046	0.1896	0.3965
random variable 4	0.3884	0.1341	0.5786	0.3117	0.6967	0.2474	0.6597	0.0949	0.1563	0.6355

Yet, as the volume of records continues to increase with Big Data, the rules of statistics dictate that many more variables are more likely to be significant. Most statistical formulae which try to determine significance always have standard deviation as the denominator. A lower standard deviation will increase the probability of significance with the reverse being true with a higher standard deviation. But the calculation of standard deviation has sample size in its denominator, thereby implying that the higher the sample size, the lower the standard deviation which will result in an increased likelihood of some event or variable being significant. In a Big Data environment, the above logic not surprisingly implies that many more variables are more likely to be significant. But using measures of significance as a threshold in filtering out variables is only step. If there are hundreds of variables that are now significant, we may now want to select the top 200 variables that are ranked by the correlation coefficient versus the target variable. Then, as discussed in previous articles and in my book (Data Mining for Managers-How to Use Data(Big and Small) to Solve Business Challenges, we utilize the approach of running a series of stepwise routines as a key component in the selection of variables for a final solution.

This approach is one method that certainly mitigates the impact of random variables as being part of the final solution. But validation of the model and looking at the variables themselves in both development and validation datasets is our final check to examine the robustness of our solution. From my practical experience, random variables are more likely to occur with limited information such as acquisition of new customers.

How to cite this article: Richard B. Can Randomness be an Element in Multiple Regression. Curr Trends Biomedical Eng & Biosci. 2016; 1(1): 555551. DOI: 10.19080/CTBEB.2016.01.555551