Statistical Diagnostics of Reproductive Dispersion Model Based on Pena Distance

Lin Dai, Hanchi Lu and Liucang Wu*
Faculty of Science, Kunming University of Science and Technology, Kunming 650093, China
*Corresponding author

Keywords: Pena distance, Reproductive dispersion model, Data deletion model, Statistical diagnostics.

Abstract. The statistical diagnostics of the reproductive dispersion model based on the Pena distance is discussed. The expression of the Pena distance under the reproductive dispersion model is obtained, and its properties are discussed, and the discrimination of high-leverage outlier is obtained. In addition, comparing the Pena distance with the Cook distance, the conclusion that the Pena distance is better than the Cook distance under certain conditions is obtained. The model and method are scientific and reasonable through the case example analysis.

Introduction

The exponential family distribution is an important class of statistical distributions in statistics[1]. However, there is still a lot of data that cannot be fitted by exponential family distribution models. Jorgensen[2] defined a class of distributions with a wider distribution than the exponential family in his monograph "The Theory of Dispersion Models" in order to meet the needs of people for complex data analysis. He called it the reproductive dispersion model (RDM).

Due to the existence of outlier, we need to diagnose the outlier and correct them, so the diagnosis of the outlier is very important. The commonly used statistics for traditional outlier judgments are Cook distance[3], etc. The Pena distance, first proposed by American statistics professor Daniel Pena[4] in 2005, is different from the previous method. The Pena distance is that a point in the study sample that is affected by the remaining points. It is also a statistic that measures the regression value of a particular point after the point in the sample is deleted and the effect of the predicted value.Hu Jiang[5] studied the influence analysis of generalized linear and nonlinear regression models based on Pena distance. Attention has been paid to the reproductive dispersion model and statistical diagnosis in recent years, which has attracted many scholars' research. So far, Tang[6,7] systematically studied the local impact analysis and statistical diagnosis of the reproductive dispersion model. Wu Liucang[8] studied the parameter estimation of mixed reproductive dispersion. However, the data analysis of the reproductive dispersion model based on Pena distance has not been studied, and statistical diagnosis is an indispensable part of data analysis.

Based on the above analysis, this paper proposes a statistical diagnosis of the reproductive dispersion model based on Pena distance, and gives the statistical diagnosis method of the model. Through the corresponding case analysis, compare the statistical diagnostic amount to distinguish the difference between the outlier or the strong influence point. The research results show that the theory and method proposed in this paper are effective.

Generalized Linear Reproductive Dispersion Model

Reproductive Dispersion Model

Reproductive dispersion model, also known as reproductive dispersion distribution family. If the probability density function of a random variable is expressed as:

Copyright © 2019, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

$$p(y;\mu,\sigma^2)=a(y;\sigma^2)\exp\{-\frac{1}{2\sigma^2}d(y;\mu)\},y\in C\ , \tag{1}$$

Where $a(y;\sigma^2)\geq 0$ is a suitable known function; $d(y;\mu)$ is the unit deviance function defined on $C\times\Theta$; $\Theta\subseteq C\subseteq R$ is an open interval; the convex support set is the minimum interval where C contains S, and S is the probability density function Support set; $d(y;\mu)$ satisfies positive definite conditions:

$$d(y;y)=0,\quad \forall y\in\Theta;\quad d(y;\mu)>0,\quad \forall y\neq\mu,$$

Where $\mu\in\Theta$ is the positional parameter and $\sigma^2>0$ is the dispersion parameter, then Y is subject to the reproductive dispersion family of parameters μ and σ^2, abbreviated as $Y\sim RDF(\mu,\sigma^2)$. From the above discussion, we can see that the reproductive dispersion distribution family is a kind of distribution family with a wider distribution than the exponential family, which is the promotion and development of the exponential family distribution. The reproductive dispersion distribution family contains many common distributions such as normal distribution, exponential family distribution, extreme value distribution, simplex distribution and double exponential distribution.

Generalized Linear Reproductive Dispersion Model

As the introduction indicates, the reproductive dispersion distribution family is a direct extension and development of the exponential family distribution. The generalized linear model is characterized by the exponential family distribution as its "linear model" of random errors. This paper will discuss the generalized linear model with RDF as random error—the generalized linear reproductive dispersion model.

Consider the generalized linear reproductive dispersion model below:

$$\begin{cases} y_i\sim RDF(\mu_i,\sigma^2), \\ g(\mu_i)=x_i^T\beta, \\ i=1,2,...,n. \end{cases} \tag{2}$$

Where y_i is an independent response variable and $x_i=(x_{i1},x_{i2},\cdots,x_{ip})^T$ is an explanatory variable whose dimension is p. $\beta=(\beta_1,\beta_2,\cdots,\beta_p)^T$ is an unknown parameter of the azimuth model whose dimension is p. In fact, the dispersion parameters can also be modeled, which is what needs to be studied in the future.

Pena Distance

The Cook distance study is the effect of deleting a (group) point on an estimate or predictor. The Pena distance study is that a certain point in the sample is affected by the remaining points. The estimated values $\hat{\beta}$ and $\hat{\beta}_i$ of the parameter maximum likelihood estimation are obtained by the Gauss-Newton iteration method. The Pena distance is defined as follows:

$$S_i=\frac{(\hat{\beta}_i-\hat{\beta}_{i(j)})^T H^T H(\hat{\beta}_i-\hat{\beta}_{i(j)})}{p\hat{\sigma}^2}$$

Where $H=X(X^TX)^{-1}X^T$ is the hat matrix, p is the dimension corresponding to the explanatory variable, and $\hat{\sigma}^2$ is the estimated value of the model variance after deleting the i point. $\hat{\theta}_{i(j)}$ indicates the parameter estimate of the i data point after deleting the j data point. For the specific analysis, the S_i is also calculated at a certain point after deleting each point, and a scatter plot is drawn. Larger S_i may be an outlier or a strong influence point.

For the size of the statistical diagnosis, since there is no uniform standard to judge the strong influence point and the outlier, the determination of the critical value is very difficult. Here we use $\bar{M}+2SM$ [9,10]as the critical value for judging strong influence points and outlier, where $\bar{M}$ and SM are the mean and standard deviation of the statistical diagnostics.

The detailed process is given below. We define the Pena distance[4]as follows: $S_i=\frac{s_i^T s_i}{pVar(\hat{y}_i)}$, Therefore, the Pena distance corresponding to the model (2) is as follows:

$$\hat{S}_i=\frac{s_i^T s_i}{pVar(\hat{y}_i)}=\frac{1}{p\hat{h}_{ii}\hat{\sigma}^2}\sum_{j=1}^{n}\frac{\hat{h}_{ij}^2}{(1-\hat{h}_{jj})^2}\hat{e}_j=\frac{1}{p\hat{h}_{ii}}\sum_{j=1}^{n}\frac{\hat{h}_{ij}^2}{(1-\hat{h}_{jj})}\hat{r}_j^{\,2}$$

According to Wei Bocheng's statistical diagnosis: $E(\hat{r}_j^{\,2})=1$, so $E(\hat{S}_i)=\frac{1}{p\hat{h}_{ii}}\sum_{j=1}^{n}\frac{\hat{h}_{ij}^2}{(1-\hat{h}_{jj})}$.

Let $h^*=\max_{1\le j\le n}h_{jj}$, we have $E(\hat{S}_i)=\frac{1}{p\hat{h}_{ii}}\sum_{j=1}^{n}\frac{\hat{h}_{ij}^2}{(1-\hat{h}_{jj})}\le\frac{1}{p(1-h^*)}=\to\frac{1}{p}$ （$h^*\to 0$）.

On the contrary, when $h_{jj}\ge\frac{1}{n}$, we have $E(\hat{S}_i)=\frac{1}{p\hat{h}_{ii}}\sum_{j=1}^{n}\frac{\hat{h}_{ij}^2}{(1-\hat{h}_{jj})}\ge\frac{1}{p(1-\frac{1}{n})}\to\frac{1}{p}(n\to\infty)$.

In summary, when the sample size n is large $(n\to\infty)$ and h^* is small $\left(h^*\to 0\right)$, the Pena distance of all sample points is expected to approach $\frac{1}{p}\left(E(\hat{S}_i)\to\frac{1}{p}\right)$.When the Pena distance of a sample point differs greatly from $\frac{1}{p}$, it can be judged that the sample is an outlier.

Case Analysis

Below we use the daily minimum temperature data to illustrate the practical application of the proposed model and method. We collected daily weather data from the Kunming platform in 2017 from the China Meteorological Data Network. The data contains the Y —day minimum temperature (Celsius) and four explanatory variables: X_1 —Daily average water vapor pressure (hPa), X_2 — Daily precipitation (millimeter) at 20-20 o'clock, X_3 —Daily precipitation (millimeter) at 08-08 o'clock, X_4 —Daily average wind speed (m/s).We will establish the relationship between the Y — day minimum temperature and the four explanatory variables.

Since the collection of meteorological data for the whole year, we consider dividing the data into the first half of the year (January-June) and the second half (July-December).The daily minimum temperature is the lowest daily temperature in 2017.Therefore, the daily minimum temperature, the minimum daily temperature in the first half of the year, and the daily minimum temperature in the second half of the year are subject to extreme value distribution. This data can be used for in-depth analysis and research using the reproductive dispersion model proposed in this paper.

According to model (2), the estimated results of daily minimum temperature data for January-June (181 days) and the daily minimum temperature data for July-December (184 days) are shown in Table 1:

Table 1. Estimated results of daily minimum temperature data parameters

Model	Data	Constant	X_1	X_2	X_3	X_4
$\hat{\beta}$	January-June	-14.8425	1.3903	-0.1701	-0.0203	3.9429
	July-December	-4.0891	1.2448	-0.0737	-0.0270	-0.7412

It is known from the table that the maximum likelihood estimates of the parameters $\hat{\beta}$ of the data for the first half of the year(January-June) and the second half of the year(July-December) are:

$$\hat{\beta}_1 = (\text{-}14.8425\ ,\ 1.3903\ ,\ -0.1701\ ,\ 0.0203\ ,\ 3.0429)^T,$$

$$\hat{\beta}_1 = (\text{-}4.0891\ ,\ 1.2448\ ,\ -0.0737\ ,\ 0.0270\ ,\ -0.7412)^T.$$

Thus, the corresponding statistic can be obtained according to the formula of statistical diagnosis.

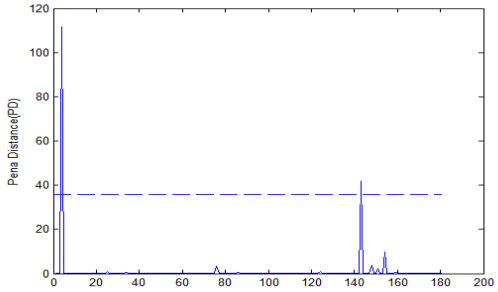

Figure 1. Normal distribution Pena distance

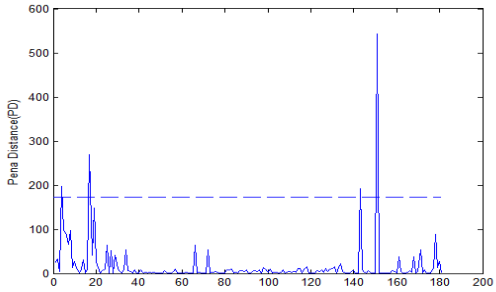

Figure 2. Extreme value distribution Pena distance

As we mentioned earlier, the meteorological data in the first half of the year is subject to extreme distribution which is asymmetrical distribution. Here we use the Pena distance under the normal distribution in the symmetric distribution and the Pena distance under the extreme value distribution for lateral comparison. The results are shown in Figures 1 and 2.

We see that 4, 143 points are outliers in Figure 3 and 4, 17, 143, and 151 points are outliers in Figure 4.This is in line with the daily meteorological situation of the Kunming platform in 2017 in the China Meteorological Data Network. The Pena distance under the reproductive dispersion distribution diagnosed the four outliers in the first half of the year, while the Pena distance under the normal distribution only diagnosed two outliers. The Pena distance effect under the reproductive dispersion distribution is better than the Pena distance effect under the normal distribution in comparison.

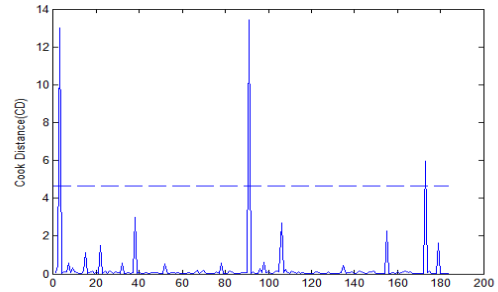

Figure 3. Cook distance in the second half of the year

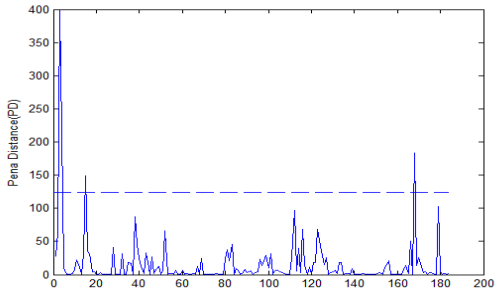

Figure 4. Pena distance in the second half of the year

The Pena distance is longitudinally compared with the Cook distance after the Pena distance is horizontally compared. The results are shown in Figures 3 and 4.

3, 91 ,168 points are outliers in Figure 3 and 3, 15 ,168 points are outliers in Figure 4.Comparing the daily meteorological conditions of the Kunming platform in 2017 in China's meteorological data network, it is found that the minimum temperature of the three days of 3, 15, and 168 is abnormal, and the minimum temperature of 91 is normal. The Cook distance of the reproductive dispersion distribution not only detects the abnormal point of 15 but also detects the abnormal point by mistake. This shows that the Pena distance diagnosis of the reproductive dispersion distribution is better than the Cook distance.

The comparison between Figure 4 and Figure 6 shows that the statistical diagnostics under the data for the first half of the year and the second half of the year are significantly different. This is

consistent with the difference in climatic temperatures between the first half of the year and the second half of the actual life.

Summary

The model proposed in this paper has the following two advantages compared with the models and methods proposed in other literature:

First, the reproductive dispersion model of this paper is more widely distributed than the exponential family and it has a wider application range.

Second, the statistical diagnosis of reproductive dispersion model is studied systematically based on Pena distance in this paper.

Simulations and case studies show that the theory and method presented in this paper are effective. The results obtained in this paper broaden the statistical diagnostic method of the reproductive dispersion model and the application range of Pena distance, enriching its theory and method.

Acknowledgement

L.C. Wu's research was supported by a grant from the National Natural Science Foundation of China (No.11861041).

Reference

[1] Weiss R E, Cho M. Bayesian marginal influence assessment[J]. Journal of Statistical Planning & Inference, 1998, 71(12): 163-177.

[2] Jorgensen B. The Theory of Dispersion Models[M]. London: Chapman and Hall, 1997: 3—29.

[3] Cook R D. Assessment of local influence (with discussion) [J]. Journal of the Royal Statistical Society, Series B, 1986, 48: 133-169.

[4] Pena D. A New Statistic for Influence in Linear Regression[J]. Technometrics, 2005, 47(1): 1-12.

[5] Hu Jiang, Lin Jinguan, Zhao Yanyong. Influence Analysis of Generalized Linear Regression Model Based on Pena Distance[J]. Applied Mathematics, 2017, 30(3): 539-546.

[6] Tang N S, Wei B C, Wang X R. Local influence diagnostics in nonlinear reproductive

dispersion models[J]. Statistics & Probability Letters, 2000, 46(1): 59-68.

[7] Tang N S, Wei B C, Zhang W Z. Influence diagnostics in nonlinear reproductive

dispersion mixed models[J]. Statistics, 2006, 40(3): 227-246.

[8] Wu Liucang, Kong Xiangchao, Dai Lin. Parameter Estimation of Mixed Regeneration Divergence Model[J]. Journal of Statistics and Information, 2017, 32(8): 3-8.

[9] Lee S Y , Tang N S . Local influence analysis of nonlinear structural equation models[J]. Psychometrika, 2004, 69(4):573-592.

[10] Zhu H T , Lee S Y . Local Influence for Incomplete-Data Models[J]. J. R. Stat. Soc. Ser. B, 2001, 63(1):111-126.