I J R B A T, Vol. V, Special Issue (3), Nov-2017 ISSN 2347 – 517X

INTERNATIONAL JOURNAL OF RESEARCHES IN BIOSCIENCES, AGRICULTURE AND TECHNOLOGY

© VISHWASHANTI MULTIPURPOSE SOCIETY (Global Peace Multipurpose Society) R. No. MH-659/13(N)

www.vmsindia.org

A COMPARATIVE STUDY OF REGRESSION METHODS APPLIED TO BIOLOGICAL AND AGRICULTURE DATA

S. S. Desai

Department of Statistics, Gopal Krishna Gokhale College, Kolhapur-416012 (MS) India.

ssd.stats@gmail.com

Abstract:

A popular data analysis technique used in almost all subjects including Biosciences and Agriculture is regression. In regression analysis a researcher has to face with a data on two or more variables and the interest lies in modeling the relationship between them. Mostly, this could be done by using least squares (LS) method or maximum likelihood estimator (MLE) method. Regression model is fitted under certain assumptions like, independence of predictors; error variable follows normal distribution with constant variance etc. A real life data may not satisfy some assumptions and these methods give misleading results. In this article, we use support vector machine for prediction of future values of response variable and the performance of different estimation methods is evaluated through real data.

Keywords: Least squares method; Multiple linear regression; M-estimator; Support vector regression; Prediction risk.

Introduction

The regression is the most extensively used data analysis technique in almost all fields including Biosciences, Agriculture and Technology. Mitchell-Olds and Shaw (1987), Li et al. (2014) have applied regression in Biological sciences. Overmars and Verburg (2006), Jirapure and Deshkar (2016) have used regression in agriculture. In these sciences, usually a researcher has to deal with a data on two or more variables and the interest lies in modeling the relationship between them. A researcher prefers to use multiple linear regression first. A multiple linear regression model is defined as

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e} \qquad (1.1)$$

where $\boldsymbol{Y}$ is known as response variable and is a vector of n observations, $\mathbf{X}$ is a matrix of order $(n \times k)$ of observations on $k-1$ predictors (regressors) $X_1, X_2, \ldots, X_{k-1}$ with 1's in the first column, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{k-1})'$ is a vector of unknown regression parameters, and $\boldsymbol{e}$ is a vector of errors. The assumptions on the model in (1.1) are $E(\boldsymbol{e}) = \mathbf{0}$, $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2\mathrm{I}$ and $\boldsymbol{e} \sim N_n(\mathbf{0},\ \sigma^2\mathrm{I})$, where, I is the identity matrix of order $n \times n$ and σ^2 is error variance.

Important steps in regression are to obtain the estimates of regression parameters, to choose an appropriate model for the given data and to predict the future values of response variable as accurate as possible. Least squares (LS) and maximum likelihood estimator (MLE) methods are generally used for estimation of parameters in linear regression. The performance of these methods is excellent if above assumptions on errors are true. The performance of least squares method is not satisfactory when the data contains an influential observation and/or the distribution of error variable is not normal. To tackle these situations Huber's M-estimator and rank-based estimator are used.

An alternative to these methods is to use a data dependent method such as Support Vector Machine (SVM). Boser et al. (1992) have introduced SVM in COLT. Vapnik et al. (1997) have extended SVM to regression and called it as Support Vector Regression (SVR). In this article, we use SVM for prediction of future values of response variable and the performance of different methods of parameter estimation is evaluated through real data.

The article is organized as: Section 2 gives a brief introduction to materials and some methods used for parameter estimation in linear regression. In section 3, Results of regression analysis and comparison of performance of different methods through real data is reported. Finally, Section 4 gives discussion.

2 Materials and Methods

In real life data, some of the assumptions mentioned in Section 1 are not satisfied, and we come across some of the problems like outlier observations in the data, non-normal distribution of error variable, dependency in the predictor variables, non-linearity etc. A method to be used for analysis of data depends on the problems present in the data. Below we discuss in brief two methods for parameter estimation and one for function estimation.

2.1 Least squares method

The method of least squares was discovered by Legendre A. M. in France around 1805 (see Birkes and Dodge, 1993). The LS estimator of $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}. \qquad (2.1)$$

The vector of predicted values $\widehat{\boldsymbol{Y}}$ of the response variable $\boldsymbol{Y}$ based on $\widehat{\boldsymbol{\beta}}_{LS}$ is

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{LS} \qquad (2.2)$$

The values of the residual measure the deviation between the observed data and the predicted values. The i^{th} residual denoted by r_i is defined as

$$r_i = y_i - \hat{y}_i,\ i\ =\ 1, 2, .., n. \qquad (2.3)$$

where y_i denote the i^{th} observation on the response variable and $\hat{y}_i$ is the corresponding predicted value.

For more details about LS estimator and residuals one may refer Montgomery et al. (2006). The performance of LS estimator and inference based on it is excellent for clean data which satisfies the assumptions mentioned in Section 1.

2.2 M- estimator method

Huber in 1964 has proposed a robust loss function defined as;

$$L_H = \begin{cases} \frac{(r_i)^2}{2}\ , & \text{for } |r_i|\ \le \delta \\ \delta|r_i| - \frac{\delta^2}{2}\ , & \text{otherwise} \end{cases} \qquad (2.4)$$

An estimator $\widetilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ which minimizes a function ρ of residuals that is

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(r_i) \qquad (2.5)$$

for some function $\rho(\cdot)$ is called as M–estimator.

An iterative reweighted least squares method is used to obtain the M–estimator of unknown regression parameters and it is given by

$$\widetilde{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{W}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{W}\mathbf{Y} \qquad (2.6)$$

where $\boldsymbol{W} = \text{diag}(w_1, w_2, \dots, w_n)$, is the weight matrix. Huber's M-estimator is designed to perform well when error distribution is non normal but it is close to normal (Birkes and Dodge, 1993, pp. 111). The vector of predicted values $\widehat{\boldsymbol{Y}}$ of the response variable $\boldsymbol{Y}$ based on the M–estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{Y}} = \mathbf{X}\widetilde{\boldsymbol{\beta}} \qquad (2.7)$$

We have taken the value of bending point constant δ =1.345 for calculating M- estimator,

2.3 Support vector regression

In SVR, based on a data set $(\boldsymbol{x}_i, y_i),\ i\ =\ 1, 2, \dots, n$ of input vectors $\boldsymbol{x}_i \in R^{k-1}$ (i^{th} row of design matrix $\mathbf{X}$) and associated targets $y_i \in R$, an unknown regression function $f(\boldsymbol{x}_i)$ can be obtained in the form,

$$y_i = f(\boldsymbol{x}_i) +\ e_i,\ \ i = 1, 2, \dots, n. \qquad (2.8)$$

where, e_i is an error term. Using the ε -insensitive loss function (Vapnik, 1995)

$$L_\varepsilon(\ y_i, f(\boldsymbol{x}_i)) =\ Max.\{\ |f(\boldsymbol{x}_i) - y_i| - \varepsilon,\ 0\ \} \qquad (2.9)$$

the regression problem can be written as convex optimization problem as:

$$\text{Minimize} \quad \tfrac{1}{2}\|\boldsymbol{w}\|^2 \qquad (2.10)$$

$$\text{Subject to:} \quad y_i - (\boldsymbol{x}_i\boldsymbol{w} + b\ )\ \le\ \varepsilon\ ,\ i\ =\ 1, 2, \dots, n. \qquad (2.11)$$

$$(\boldsymbol{x}_i\boldsymbol{w} + b\ ) - y_i\ \le\ \varepsilon\ ,\ i\ =\ 1, 2, \dots, n. \qquad (2.12)$$

where, ε > 0 is a pre-defined constant. Using Langaranges method of multipliers and dual theory the above problem can be solved and the weight vector is given by

$$\boldsymbol{w}'\ = \sum_{i=1}^{n_{nsv}} (\alpha_i - \alpha_i^*)\, \boldsymbol{x}_i \qquad (2.13)$$

And the regression function is given by,

$$f(\boldsymbol{x})\ = \sum_{i=1}^{n_{nsv}} (\alpha_i - \alpha_i^*)\, \boldsymbol{x}_i\boldsymbol{x}' + b \qquad (2.14)$$

where n_{nsv} – number of support vectors and $\alpha_i,\ \alpha_i^*$ for $i = 1, 2, \dots, n$ are Lagrange's multipliers. For computing $f(\boldsymbol{x})$ the value of $\boldsymbol{w}$ does not need to be calculated explicitly. The value of bias b is given by (Gunn, 1998),

$$b = -\tfrac{1}{2}(\boldsymbol{x}_r\ +\ \boldsymbol{x}_s)\boldsymbol{w} \qquad (2.15)$$

where $\boldsymbol{x}_r$ and $\boldsymbol{x}_s$ are the support vectors (i.e. any input vector which has nonzero value of either α_i or α_i^* respectively). For details refer Desai and Kashid (2015).

To perform SVR, we have taken $\mathrm{C}\ =\ Max\ (|Me - 3Q.D.|, |Me\ +\ 3Q.D.|)$ suggested by Desai and Kashid (2015) and $\varepsilon\ =\ \mathrm{C} * 10^{-6}$.

Comparison of performance of methods

To compare the performance of various methods, we obtain the mean absolute percentage error (MAPE) defined as,

$$\text{MAPE} = \sum_{i=1}^{n} [(\ |y_i - \hat{y}_i| / y_i) * 100] / n \qquad (2.16)$$

Results

In this section, we do regression analysis of two real data sets, one from Biology and another from Agriculture. We compare the performance of above methods using analysis of these data sets.

3.1 Blood fat contents data (Kleinbaum and Kupper, 1978):

This data contains 25 observations on response variable blood fat content (Y), two predictor variables weight in kilograms (X_1) and age in years (X_2). We apply all three methods discussed in Section 2. Obtained predicted values (assuming the same data as test data) and MAPE values. To introduce outlier we changed the observation y_8 = 3850 instead of original value 385. The values of regression parameters and MAPE for original and outlier data are shown in the Table 1. Using SVR we get regression function, but for comparison purpose we report regression parameters.

From Table 1, we observe that there is small variation in values of regression parameters, MAPE values and the performance of three methods is same for actual data. But for outlier data there is large variation in regression parameters and MAPE values. The performance of LS is poor (MAPE=54.64), that of SVR is good (MAPE=25.934) and that of M-est. method is better (MAPE=13.886) for blood fat contents data.

3.2 Oil extraction from peanuts data (Montgomery, 2006, pp 574):

This data contains 16 observations on response variable total oil yield (Y) and five predictor variables, X_1- CO2 pressure (bar), X_2- CO2 temperature (in degrees Celsius), X_3-peanut moisture (percent by weight), X_4- CO2 flow rate (L/min) and X_5- peanut particle size (mm). All the predictors are categorical taking either of two values. We coded all five predictors in 0 and 1 separately. We apply all three methods to the coded predictors and response. Obtained predicted values (assuming the same data as test data) and MAPE values using every method. To introduce outlier we changed the observation y_7 = 350 instead of original value = 71. The values of regression parameters and MAPE for original and outlier data are shown in the Table 2.

Table 1: Parameters and MAPE values for Blood fat data using different methods

Data	Actual data			Outlier data		
Method	LS	M-Est.	SVR	LS	M-Est.	SVR
β_0	77.983	74.974	46.629	105.650	74.974	60.687
β_1	0.417	0.425	0.785	4.293	0.425	4.791
β_2	5.217	5.228	5.373	1.248	5.228	1.460
MAPE	11.330	11.139	11.513	54.640	13.886	25.934

Table 2: Regression parameters and MAPE values for Oil extraction data using different methods

Data	Actual data			Outlier data		
Method	LS	M-Est.	SVR	LS	M-Est.	SVR
β_0	62.250	62.312	54.078	97.125	66.694	71.288
β_1	7.500	7.438	9.580	−27.375	3.056	−19.856
β_2	19.750	20.759	20.469	54.625	24.194	53.033
β_3	1.250	1.313	4.024	36.125	5.694	36.588
β_4	0.000	− 0.062	2.913	−34.875	−4.444	−26.523
β_5	−44.500	− 44.562	−36.642	−79.375	−48.944	−66.078
MAPE	12.939	13.045	15.055	77.818	13.997	63.837

For this data also we observe the same type of results as reported in Section 3.1.

Discussion

In this article, we discussed LS and M-estimator methods for estimating parameters for linear regression and SVM for regression function approximation. There are other methods also available in the literature for estimation. A researcher may find difficult to choose one of them. Naturally, if one uses a method without knowing the nature of the data, then the results may be misleading. It is important to understand the nature of the data and problems associated with it. Based on the problems in the data, an appropriate method should be chosen. For clean data the performance of LS is excellent, for data containing outliers M-estimator method performance better. When nature of data is unknown one may use SVR, but it require to choose the parameters of SVR properly.

References

[1] Birkes, D. and Dodge, Y. (1993): Alternative Methods of Regression, *John Wiley & Sons.*

[2] Boser, B. E., Guyon, I. M. and Vapnik, V. N. (1992): A Training Algorithm for Optimal Margin Classifiers, *5th Annual ACM Workshop on COLT 92, Pittsburgh, pp. 144-152.*

[3] Desai, S. S. and Kashid, D. N. (2015): Estimation of Regression Parameters Using SVM with New Methods for Meta Parameter, *International Journal of Data Mining, Modeling and Management, 7(3), 239-256.*

[4] Gunn, S. (1998): Support Vector Machines for Classification and Regression, *Technical Report. School of Electronics and Computer Science, University of Southampton.*

[5] Huber, P. (1964): Robust estimation of a location parameter, *Annals of Mathematical Statistics 35, 73-101.*

[6] Jirapure P. V., Deshkar P. (2016): Regression method and Cloud computing llTechnology in the field of Agriculture, *International Journal of Innovative Research in Computer and Communication Engineering Vol. 4(4), 6758-6765.*

[7] Kleinbaum D. G. and Kupper L. L. (1978): Applied Regression Analysis and Other Multivariable Methods, *Duxbury Press, 1978, page 149.*

[8] Li Y, Liang M, and Zhang Z (2014): Regression Analysis of Combined Gene Expression Regulation in Acute Myeloid Leukemia. *PLOS Computational Biology, Vol. 10(10).*

[9] Mitchell-Olds, T., and Shaw, R. G. (1987). Regression analysis of natural selection, *statistical inference and biological interpretation. Evolution, 41(6), 1149-1161.*

[10] Montgomery D. C., Peck E. A. and Vining G. G. (2006): Introduction to Linear Regression Analysis, *Third edition - John Wiley and Sons Inc.*

[11] Overmars K. P. and Verburg P. H. (2006): Multilevel modeling of land use from field to village level in the Philippines, *Agricultural Systems 89, 435–456.*

[12] Vapnik, V. (1995): The Nature of Statistical Learning Theory, *Springer, New York*

[13] Vapnik, V., Golowich, S. and Smola, A. (1997): Support Vector Method for Function Approximation, Regression Estimation and Signal Processing, *In Mozer M., Jordan M. and Petshe T. editors, NIPS, Vol. 9, pp. 281-287, Cambridge, MA, MIT Press.*