Applied Mathematical Sciences, Vol. 9, 2015, no. 110, 5487 - 5494
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ams.2015.56431

Modeling of Robust Regression in Breast Tissue Data

Siti Hasliza Ahmad Rusmili, Norizan Mohamed, Nor Azlida Aleng and Nur Farahana Zainudin

School of Informatics and Applied Mathematics
Universiti Malaysia Terengganu
21030 Kuala Terengganu, Malaysia

Abstract

In form of analyzing data that are contaminated with outliers, important method is robust regression (RR) and RR provide resistant (stable) results in the presence of outliers. The purpose of this research is to find the relationship between variable and to develop the best model in robust regression by using breast tissue data that were taken from the UCI machine learning repository. Least trimmed square, S-estimation and MM-estimation is used to develop best fit model using Statistical Analysis System (SAS) is applied to model the robust regression. Then, by using S-estimation, LTS and MM-estimation gave the R^2 values were 0.9991, 0.9990 and 0.8185 respectively. Therefore it can be concluded that all three estimations which were Least trimmed square, S-estimation and MM-estimation were comparable.

Keywords: multiple linear regressions, robust regression, S estimation, LTS, MM-estimation

Introduction

Breasts are made up of fat and breast tissue, along with nerves, veins, arteries and connective tissue that helps hold everything in place. The main chest muscle (the pectoralis muscle) is found between the breast and the ribs in the chest wall. Breast tissue is a complex network of lobules (small round sacs that produce milk) and ducts (canals that carry milk from the lobules to the nipple openings during breastfeeding) in a pattern that looks like bunches of grapes that called lobes [1].

The cancer cell is described in medical literature since about 2500g. c. Cancer cells are living cells whose growth and reproduction are out of control [2]. A tissue biopsy is typically used to diagnose in breast cancer. This process, which is a sample of suspect tissues is removed from the patient and sent away for histological and chemical analysis. The process basically takes 1 to 2 days, but some samples give inconclusive results and require the patient to have second biopsy. If a faster method of tissue analysis or to determine if a tissue sample is a good candidate for further screening, the cost of screening and stress endured by the patient in awaiting results could be reduced [3].

In this research, breast tissue data is used to be tested and analysed using robust regression approach. Somehow, data will be tasted to determine it in normal distribution by using Q-Q plot. Robust regression estimators are to detect outliers and high leverage from data and upgrade the result so best model are produced. The significant variable from previous research [4] used to analyse by using Statistical Analysis System (SAS). The performance of models were considered using statistical measurement such as determination of coefficient or R square value hence evaluated under the circumstances.

Objective

This research is to examine the relationship between the independent variable and dependents variable selected by robust regression with difference estimation thus, produce the good-fit model.

Methodology

In order to analysis and robust regression, there are a few steps to be made:

1. Identify the significant variable from previous analyses of multiple linear regressions.
2. Investigate the correlation among variables.
3. Determine outliers and diagnostics outliers by using Statistical Analysis System (SAS).
4. Estimate regression model using Least Trimmed Squared, S estimation and MM estimation.
5. Test whether independent have significant effect on the dependant variable in determination of coefficient.

The main purpose of robust regression is to detect outliers and provide resistant (stable) results in the presence of outliers. In order to achieve this stability, robust regression limits the influence of outliers. Historically, three classes of problems have been addressed with robust regression techniques:

- Problems with outliers in the y-directions(response direction)
- Problems with multivariate outliers in the x-space(that is, outliers in the covariate space, which are also referred to as leverage points)
- Problems with outliers in both the y-direction and the x-space

Many methods have been developing in response to these problems. However, in statistical application of outliers detection and robust regression, the methods most commonly used today are Huber M estimation, high breakdown value estimation and MM estimation [5].

Least Trimmed Squared (LTS) Estimator is a high breakdown value method by Rousseeuw (1984) [6]. The breakdown value is a measure of the proportion of contamination that a procedure can withstand and still maintains its robustness [7]. Least trimmed to square defined

$$\hat{\beta}^{(LTS)} = \arg min = \sum_{i=1}^{h} r_{(i)}^2(\beta).$$

Where $r_{(1)}^2 \leq \cdots \leq r_{(n)}^2$ are the ordered squared residuals. There always exists a solution for the LTS-estimator. The LTS estimator is regression equivariant, scale equivariant and affine equivariant. If $p > 1, h = \left[\frac{n}{2}\right] + \left[\frac{p+1}{2}\right]$ then the breakdown point of the LTS-estimator $\varepsilon^* := \frac{\frac{[n-p]}{2}+1}{n}$. The LTS can be very sensitive to a very small change of data or to a deletion of even one point from data set (i.e. small change of data can really cause a large change of the estimate). In Linear model $y_i = \alpha + \beta_1 x_{i1} + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i = \boldsymbol{x_i'}\beta + \varepsilon_i$, ith of n observations. The an estimator b for β, the fitted model is $\hat{y}_i = a + y_i = \alpha + b_1 x_{i1} + b_1 x_{i1} + \cdots + b_k x_{ik} + e_i = \boldsymbol{x_i'}b$ and the residuals are given by $e_i = y_i - y_i$.

Based on [6], the estimates b are determined by minimizing a particular objective function over all b,

$$\sum_{i=1}^{n} \rho(e_i) = \sum_{i=1}^{n} \rho(y_i - \boldsymbol{x_i'}b)$$

Where the function ρ gives the contribution of each residual to the objective function. A reasonable ρ should have the following properties such as always nonnegative, $\rho(e) \geq 0$, equal to zero when its argument is zero, $\rho(0) = 0$, symmetric, $\rho(e) = \rho(-e)$ and monotone in $|e_i|, \rho(e_i) \geq \rho e_{i'}$ for $|e_i| > |e_i|$.

MM estimation, introduced by Yohai (1987), combines high breakdown values estimation and M estimation. It has both the high breakdown property and a higher statistical efficiency than S estimation [5]. The MM estimation procedure is to estimate the regression parameter using S estimation which minimize the scale of the residual from M estimation and then proceed with M estimation. MM estimation aims to obtain estimates that have a high breakdown value and more efficient. Breakdown value is a common measure of the proportion of outliers that can be addressed before these observations affect the model [7]. MM estimator is the solution of

$$\sum_{i=1}^{n} \rho_1'(ui) X_{ij} = 0 \; or \; \sum_{i=1}^{n} \rho_1' \left(\frac{Y_i - \sum_{j=0}^{k} \widehat{\beta_J}}{s_{MM}} \right) X_{ij} = 0$$

where SMM is the standard deviation obtained from the residual of S estimation and ρ is a Tukey's biweight function:

$$\rho(u_i) = \begin{cases} \frac{u_i^2}{2} - \frac{u_i^4}{2c^2} + \frac{u_i^6}{26}, & -c \le u_i \le c; \\ \frac{c}{6}, & u_i < -c \; or \; u_i > c. \end{cases}$$

Set Data

The EIS data analysed here was taken from a study by of 106 freshly excised breast tissue samples. EIS was performed in the range of 488 Hz to MHz and histology was performed to classify each tissue sample as either carcinoma, fibro-adenoma, mastopathy, glandular, connective or adipose tissue. This data is available publicly at the UCI Machine Learning Repository [2]. This dataset include 106 instances and each instance belongs to one class. Six classes of freshly excised tissue were studied using electrical impedance measurements. Table I present the description about six classes of tissue while Table 2 describes the description significant variable of the characteristics of the breast tissue dataset from the previous study [4].

Table I. Description about the Six Classes of Tissue

Six Classes of Tissue		
Class		Total
Car	Carcinoma	21
Fa	Fibro-adenoma	15
Mas	Mastopathy	18
Gla	Glandular	16
Con	Connective	1
Adi	Adipose	22

Table II. Description Significant Variable about the Characteristics (Attribute) of the Breast Tissue Dataset

Attributes		
Variable	Attribute	Description
y	I0	Impedivity (ohm) zero frequency
x_1	HFS	High-frequency slope of phase angle
x_2	AREA	Area under spectrum
x_3	A/DA	Area normalized by DA
x_4	DR	Distance between I0 and real part of the maximum frequency point
x_5	P	Length of the spectral curve

Result and Discussion

Wisconsin Breast Tissue Data from UCI machine learning repository [8] are analyses by using robust regression with different estimation. Statistical Analysis System (SAS) give chi-square value using 0.05 significant level same as t test and F test. Data significant variable of breast tissue are proceed in SAS, thus to be analyses by using robust regression so can produce better result. Table III discussed about the diagnostics summary by using least trimmed squared.

Table III. Analyse Using Least Trimmed Square

Diagnostics Summary		
Observations	Proportion	Cutoff
Outliers	0.2358	3.0000
Leverage	0.3679	3.5822
R Square = 0.9990		

The outliers are detect by using LTS, about 25 observation due to standardize robust residuals exceed the cutoff value in absolute value, while 39 observations from 106 observations are leverage. The performance of R square shows high value R square (R2=0.9990) close to 1, it can verify that high correlation between dependent and independent variable. The value R square close to 1, indicating that a greater proportion of variance is accounted for by the model, thus LTS has strong good fit models. Table IV shows the parameter estimates of tissue each variable and use to create equation of the best fit model of least trimmed square.

Table IV. Parameter Estimates of Tissue

Parameter Estimates		
Parameter		Estimates
Y	Intercept	-2.7452
	x_1	-65.3921
	x_2	-0.0045
	x_3	-3.2250
	x_4	0.4039
	x_5	1.0129

The equation of the best fit model of least trimmed square,

$$\hat{y} = -2.7452 - 65.3921x_1 - 0.0045x_2 - 3.2250x_3 + 0.4039x_4 + 1.0129x_5$$

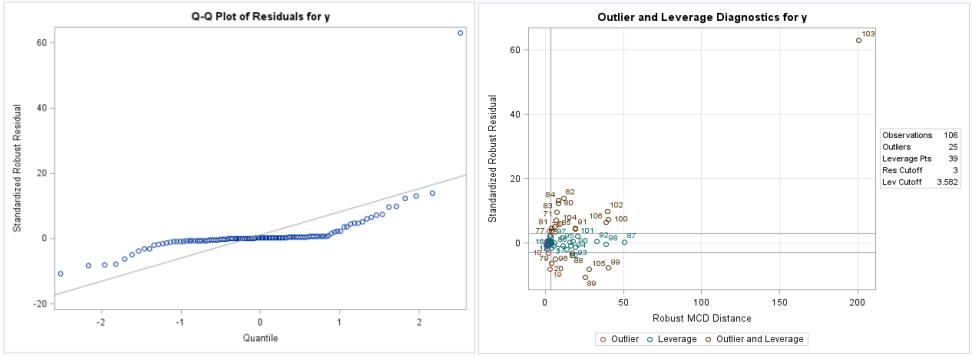

Figure 1. Q-Q Plot of Residuals for y in Breast Tissue Data and Graph Outliers and leverage diagnostics for y. A data in Q-Q Plot of Residual for y appear to be normal distribution, so the points will cluster tightly around the reference line. Graph outliers and leverage diagnostics for y displays revealing outliers and the leverage points.

Table V. Parameter Estimates Description about Six Classes of Tissue Using MM-Estimation

Parameter Estimates				
Parameter		Estimates	Standard Error	Pr>Chi Sq
Y	Intercept	-0.7601	4.4880	0.8655
	x_1	-99.3397	23.0940	<0.0001
	x_2	-0.0055	0.0007	<0.0001
	x_3	-3.8735	0.1951	<0.0001
	x_4	0.7328	0.0202	<0.0001
	x_5	0.9828	0.0047	<0.0001

Significant p value= <0.05

Parameter of x_1, x_2, x_3, x_4 and x_5 achieve significant because lower than significant p value. By using MM estimation, observation 19, 20, 79, 87, 91, 102 and 103 are detected as outliers due to standardize robust residuals exceed the cutoff (cutoff=3.000) value in absolute value.. Then, 39 observations are at leverage point. The fitted linear model of MM estimation:

$$\hat{y} = -0.7601 - 99.3397x_1 - 0.0055x_2 - 3.8735x_3 + 0.7328x_4 + 0.9828x_5$$

Table VI. Parameter Estimates Description about Six Classes of Tissues Using S Estimation

Parameter Estimates				
Parameter		Estimates	Standard Error	Pr>ChiSq
Y	Intercept	-1.7887	4.3514	0.6810
	x_1	-84.1535	23.0758	0.0003
	x_2	-0.0058	0.0007	<0.0001
	x_3	-3.8733	0.1914	<0.0001
	x_4	0.7389	0.0201	<0.0001
	x_5	0.9831	0.0048	<0.0001
			Significant p value= <0.05	

All parameters are shown significant because p value is less than significant p value. Analysis using SAS, robust regression using S estimation detect outliers and leverage. Observations 19, 20, 79, 87, 91, 100, 102 and 103 shows outliers based on standardized robust residuals exceed the cutoff value (cutoff=3. 00) in absolute value while 39 observations have been detected as leverage. The robust regression fitted model using S estimation:

$$\hat{y} = -1.7887 - 84.1535x_1 - 0.0058x_2 - 3.8733x_3 + 0.7389x_4 + 0.9831x_5$$

Table VII. Least Trimmed Square, S Estimation and Mm Estimation of R Square

Estimations		
Least Trimmed Squares	MM estimation	S estimation
R^2= 0.9990	R^2=0.8185	R^2=0.9991

Robust regression using least trimmed squares estimation, S estimation and MM estimation produce a higher value of R square. Based on this result, independent and dependent variable show strong correlation between variable. Then, both variable have strong relation between each other. Thus, can simplify that, the model fit is strong in robust regression by using difference estimators. Furthermore, outliers and leverage do not give influence because it has a minimum amount in this breast tissue data. The results show that, least trimmed squares, MM estimation and S estimation of determination of coefficient or R^2 are 0.9990, 0,8185 and 0.9991 respectively. So, the robust regression using different estimation by using breast tissue Wisconsin data is comparable.

Conclusions

This research more focus on robust regression using least trimmed squared, S estimation and MM estimation. Based on the result, the R^2 of least trimmed squares, MM estimation and S estimation are 0.9990, 0.8185 and 0.9991 respectively. The three value of R^2 gave strong correlation and relation between variables, so it shown that strong good fit model. Based on the performance of R^2 that given above it can concluded that all three estimations were comparable.

Acknowledgments. Special acknowledgement to Research Management Centre, Universiti Malaysia Terengganu and the Ministry of Science, Technology and Innovation (MOSTI), Malaysia for the financial and moral support in the form of FRGS Grant no. 59266 for the delivery of this paper.

References

[1] Susan G. Komen. (2015). http://ww5.komen.org/BreastCancer/TheBreast.html

[2] J. King, *Breast Cancer Answer: Practical Tips and Personal Advice from a Survivor*, The Career Press Inc, pp 13. 2004.

[3] M. Nonte, *Classification of Breast Tissue Using Electrical Impedance Spectroscopy*, University of Wisconsin-Madison, pp 1-8.

[4] S.H. Ahmad Rusmili et. al, *Multiple Linear Regression and Robust Regression Approach in Breast Tissue Data*, The 10th IMT-GT International Conference Mathematics, Statistics and its Application (ICMSA). 2014.

[5] SAS/STAT 12.3 User's Guide THE ROBUSTREG Procedure (Chapter), SAS Institute Inc.. 2013. Cary, NC, USA, 6810-6902.

[6] P.J. Rousseeuw and A.M. Leroy, *Robust Regression and Outlier Detection*, Wiley-Interscience, New York (Series in Applied Probability and Statistics) 1987. http://dx.doi.org/10.1002/0471725382

[7] C. Chen, Robust Regression and Outlier Detection with the Robustreg Procedure, 2002, SAS Institute Inc., Cary, NC, 265-27.

[8] UCI Machine Learning Repository, (2014). https://archive.ics.uci.edu/ml/datasets/Breast+Tissue

Received: June 28, 2015; Published: August 27, 2015