Robust Estimators of Logistic Regression with Problems Multicollinearity or Outliers Values.

Fadhil Abbul Abbas AL- Aabdi Rafid Malik Atiyah AL – Shaibani

AL Furat AL awast

Technical university

Abstract

Whenever there is a relationship between the explanatory variables (X_S). This relationship causes multicollinearity which in turn leads to inaccurate and bias estimations of the model parameters.

Therefore, this results in high discrepancy that influences the next phase of the statistical inference where (**OLS**), method loses its features having the lowest variance.

Consequently, this paper concerns itself with figuring out methods that can be applied by researchers and those who are interested in this field to overcome this problem using (**Ridge)** method. Moreover**,** the paper seeks to solve other problems such as the loss of normal distribution property or abnormalility by means of methodical means including (**Ridge)** and (**Robust Ridge).**

However this study is applied through **simulation** experiments aim at producing the data of the model. Based on these experiments and tests, the research has come up with the result that (**Robust Ridge)** is the best method that might be employed to solve the problem of has both normal and abnormal data for the estimation of the parameters of the **Logistic Regression Model.**

Keywords: Logistic Regression Model, Multicollinearity

1. Introduction

One of the main problems in regression estimation methods is multicollinearity. Multicollinearity is the term used to describe cases in which the regressors are correlated among themselves **.**The ridge regression model has been advocated in the literature as an alternative to least square estimation(LS)**,** for the multicollinearity problem in this methods, which was proposed by Hoerl and Kennard [8] (1970)**,** ridge estimators are used instead of Least Square Estimator**.**

Another common problem in regression

estimation methods is at of non - normal errors. The term simply means that the error distributions Have fatter tails than the normal distribution. These fat – tailed distributions are more prone than the normal distribution to produce outliers, or extreme observations in the data. When outliers exist in the data **,** the use of robust estimators reduces their effects **.** This work aims to investigate the way of dealing with the problem of multicollinearity .i.e. the loss

of normal distribution accompanied by outlier values in the **Logistic Regression model.**

2. Methodology

2.1 Outliers [2]

Outliers are observations that are very different from the rest of the data **.** Sample mean, sample variance and other classical estimates will be severely affeccted by these outliers and doesn't usually give good fit to the data . Barnett and Lewis (1994) defined outliers

as observations that are inconsistent with the rest of the data and most of the time will be hidden to the user since least squares residual plots fail to show these outlying point .

2.2 Logistic Regression Model [10][13][6]

Logistic regression model consists of response variable $\mathbf{Y_i}$ influenced by a set of explanatory variables $\mathbf{X_S}$ according to the relationship contains aset of parameters **βs .** If variable response $\mathbf{Y_i}$ in the life experiences of binary response of Bernoulli distribution , where response that has two levels are (1, 0).

Where: $\mathbf{Y_i \sim b(1,P_i)}$ ... (2.1)

P_i : is the probability of response when (Y =1).

(1- P_i) : is the probability of non – response when (Y = 0).

The logistic regression model is:

$$\mathbf{P_i} = \frac{\exp(\sum_{k=0}^{p} \beta_k x_{ik})}{1+\exp(\sum_{k=0}^{p} \beta_k x_{ik})} \quad \text{... (2.2)}$$

$$\mathbf{1 - P_i} = \frac{1}{1+\exp(\sum_{k=0}^{p} \beta_k X_{ik})} \quad \text{... (2.3)}$$

And that the formula (2.2) called the response function.

For the purpose of conducting a linear transformation the logistic enables the researcher (Berkson)(1944) **,** to transform the relationship between the variables (X_i) and the response variable (P_i) to the linear relationship and so draw (log P) instead of probit (P) versus (X) as follows **:**

$$\mathbf{Logit(P_i) = Log\ (\frac{P_i}{1-P_i})} \quad \text{... (2.4)}$$

Note: [16] In recent years , the evolution as a result of software that are used in the statistical analysis , including the application of (MATLAB), where they can rely on the function (**glm** fit) who writes the following formula **:**

$\widehat{\boldsymbol{\beta}}$ = **glm** fit (Y , N , X , **"distribution" ,"link","logit"**) … (2.5)

Where:

Y: response vector containing distribution counts.

N: vector of trials for each count Y

X : matrix of covariates **.**

2.3 Ridge Regression [8][5]

The regular (ridge regression) provides us with another alternative way that can be used to benefit when the explanatory variables is perpendicular to a high degree (i**.e.,** connected), and this method is a solution to the problems of multicollinearity and can be viewed as a method of interpretation of the discovery and estimation of the parameters of regression (k) when doubts the existence of the problem of multicollinearity .The estimators resulting be biased , but lead to get the mean squares error is minimize that extracted from the capabilities of the ordinary least-squares ((**OLS,** and estimates Ridge be stable in the sense that they are not affected deviations few in the data because of property smaller mean square error and transactions discretionary manner Ridge of expected to be closer to the true values of the regression coefficients in the estimates of the least-squares **OLS**)) **,** which is calculated form the following **:**

Consider the linear model

$$\mathbf{y = X\beta + \varepsilon} \quad \text{... (2.6)}$$

Where:

y : is a vector of n response values.

X: is an (n * p) matrix of rank p.

β: is a vector such that

$$E(\varepsilon) = 0, \text{and } Var(\varepsilon) = \sigma^2 I_n$$

If the columns of X are multiciollinear **,** then the least- squares estimator of β, namely

$$\widehat{\boldsymbol{\beta}}_{\mathbf{ols}} = (\mathbf{X^T X})^{-1}\mathbf{X^T y} \quad \text{... (2.7)}$$

Is an unreliable estimator due to the large variances associated with its elements. The most popular of the methods That can be used to cope with multiciollinearity is ridge regression .This method , developed by Hoerl and Kennard [8]

(1970), is based on adding a positive constant k to the diagonal element of ($\mathbf{X^T X}$). This leads to a biased estimator $\boldsymbol{\beta_{Rid}}$ **of** $\boldsymbol{\beta}$ called the ridge estimator and given by:

$$\widehat{\boldsymbol{\beta}}_{\mathbf{Rid}} = (\mathbf{X^T X} + \mathbf{K I_n})^{-1}\, \mathbf{X^T y} \quad \text{... (2.8)}$$

Where:

K: is the coefficient regression regular $0 < k < 1$

$\mathbf{I_n}$: is the unit matrix of rang p*p

P: is the number of parameters model.

The value of k can be found in several ways , including :$\mathbf{k} = \frac{\mathbf{p S^2_e}}{\widehat{\boldsymbol{\beta}}^{\mathbf{T}}_{\mathbf{ols}}.\widehat{\boldsymbol{\beta}}_{\mathbf{ols}}}$

Where: $\hat{\beta}_{ols}$, S^2_e are the estimators regression equation using the method (OLS). And

$$\mathbf{S^2_e} = \frac{\sum_{i=1}^{n} \mathbf{e_i}^2}{\mathbf{n-p}}$$

As a result of the development of computer software and Matlab applications, we can rely on the function code (**ridge**) to find the estimators of character **(R.R**) regression equation and according of the following general formula **:**

B_ridge = ridge(y,x,k) ... (2.9)

2.4 Robust Ridge Regression [1][3] [11]

When both outliers and multiciollinearity occur in a data set, it would seem beneficial to combine methods designed to deal with these problems individually**.**

Thus, robust ridge estimators will be resistant to multiciollinearity problems and will be less affected outliers .

The following formula is used to compute robust ridge estimates**:**

$$\widehat{\boldsymbol{\beta}}_{\mathbf{Robust.R}} = (\mathbf{X^T X} + \mathbf{k^* I})^{-1}.\mathbf{X^T X y} \quad \text{… (2.10)}$$

Where:

$$\mathbf{k}^* = \frac{\mathbf{p s^2_{Rob}}}{\widehat{\boldsymbol{\beta}}'_{\mathbf{Rob}}.\widehat{\boldsymbol{\beta}}_{\mathbf{Rob}}} \quad , \quad \mathbf{s^2_{Rob}} = \frac{\sum_{i=1}^{n} \mathbf{e_i}^2_{\mathbf{Rob}}}{\mathbf{n-p}}$$

2.5 Comparison of statistical measurements

For the purpose of access to the best estimate for comparison between the results as follows:

2.5.1 Mean Square Error [9] [12]

The mean squared error (MSE) of an estimator is one of many ways to quantify the difference between values implied by an estimator and the true values of the quantity being estimated .

MSE measures the average of the squares of the "errors" . The error is the amount by which the value implied by the estimator differs from the quantity to be estimated . The difference occurs because of randomness or because the estimator doesn't account for information that could produce a more accurate estimate .

$$\boldsymbol{MSE} = \sum_{i=1}^{n} \frac{(Y_i - \hat{Y}_i)^2}{n-p} = \frac{\sum_{i=1}^{n} e_i^2}{n-p} \quad \text{… (2.11)}$$

The MSE of an estimator $\hat{\beta}$ with respect to estimated parameter β is define

$$\boldsymbol{MSE}\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{E}[(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^2]$$

The MSE is equal to the sum of the variance and the squared bias of the estimator **:**

$$\mathbf{MSE}(\widehat{\boldsymbol{\beta}}) = \mathbf{var}(\widehat{\boldsymbol{\beta}}) + (\mathbf{Bias}(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}))^2 \quad \text{... .(2.12)}$$

Where:

$\mathbf{var}_{\boldsymbol{\epsilon}}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X^T X})^{-1}$ **,Bias**$(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}})^2 = (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^2$

Or by the following formula:

$$\mathbf{MSE}_{\boldsymbol{\beta}} = \mathbf{det}\frac{\mathbf{1}}{\mathbf{rep.}}(\sum_{\mathbf{i=1}}^{\mathbf{rep.}}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^{\mathbf{T}}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}))$$

Where:

rep **:** number of the repetition of the experiment.

3. Application part

For the purpose of study the properties of estimators that have been mentioned in the theoretical side was relying on the implementation of simulation experiments**.**

3.1 Simulation Study [4]

Simulation is the imitation of the operation of a real-world process or system over time. The act of simulating something first requires a model to be developed; this model represents the key characteristics or behaviors of the selected physical or abstract system or process. The model represents the system itself, whereas the simulation represents the operation of the system over time.

Simulation is used in many contexts, such as simulation of technology for performance optimization, safety engineering**,** testing, training, education. Training simulators include flight simulators for training aircraft Simulation is pilots to provide them with a lifelike experience**.** Also used with scientific modeling of regular systems or human systems to gain insight into their functioning. Simulation can be used to show the eventual real effects of alternative conditions and courses of action. Simulation is also used when the real system cannot be engaged, because it may not be accessible, or it may be dangerous or unacceptable to engage, or it is being designed but not yet built, or it may simply not exist.

3.2 Experiment of Simulation Steps [7][14][15]

Simulation experiments are summarized in the

following steps:

1. Choosing a program that we can deal with as well as we have sufficient information to start the simulation, in this study we used the MATLAB program because of its properties and its flexibility compared to other programming langue. The specimen that has been relied upon in the search form will be as follows:

$$\textbf{Log } \left(\frac{P_i}{1-P_i}\right) = \sum_{k=0}^{p} \boldsymbol{\beta_i x_{ik}}$$

2. Selecting default values of the parameters, which we will use to generate random data. We have ($\boldsymbol{\beta}_{0=}$- 0.4, $\boldsymbol{\beta_{1=}}$1.1, $\boldsymbol{\beta_{2=}}$0.3) as the default values.

3. Generate random variables ($\boldsymbol{y_i}$), which are distributed according to the binary binomial

distribution. Known variable (**y**) as the number of successful cases (n) of independent Bernoulli attempts, where the likelihood of success for each attempt rate of (p) and the probability of the failure rate of (1-p). And it can be defined as (y) as a random variable for the distribution of binary and binomial probability density function (**c.d.f**) of this distribution are:

$$\boldsymbol{f(y)} = \binom{n}{y} \boldsymbol{p^y (1\text{-}p)^{n-y}} \quad , \ y = 0, 1, \ldots\ldots, n$$

Where:

y: The number of cases of success.

n: number of attempts.

P: probability of success.

4. Generate independent variables when the ratio of contamination equal to zero in any case the fact that the data not contain any abnormal values. Where: $\boldsymbol{X_1 \sim N(0,1)}$, $\boldsymbol{X_2 = 0.3X_1}$

5. Setting the sizes of samples of the most important stages that depend on them later stages. In this study, we used sample sizes as follows:

n = (10, 30, 60, 100)

6. Choosing a certain number that represents the repetition of the experiment and to control the differences between random samples, we chose **(r = 5000**) .

7. Generating contaminated data by ratio 30% with assuming the values for the parameters

$\boldsymbol{x_{11}}$ **= normrnd (0,1,1,**$\boldsymbol{n_1}$**) ,** $\boldsymbol{n_1}$**=0.7*n**

$\boldsymbol{x_{12}}$ **= lognrnd (0,1,1,**$\boldsymbol{n_2}$**) ,** $\boldsymbol{n_{2=}}$ **0.3*n**

Where:

n_1=sample size for regular data, n_2=sample size for outliers.

Table (3.1): Simulation Results for **Ridge regression** method when β_0= - 0.4 , β_1=1.1, β_2=0.3 in **normal** data.

Sample size	**Bias of the parameters**			***MSE***	MSE_{model}
	$\widehat{\boldsymbol{\beta}}_0$	$\widehat{\boldsymbol{\beta}}_1$	$\widehat{\boldsymbol{\beta}}_2$		
10	0.4959	0.9196	0.0684	0.0198	1.5958e-06
30	0.5971	0.8357	0.0622	0.0056	2.2063e-07
60	0.6313	0.8122	0.0604	0.0031	6.0475e-08
100	0.6488	0.8024	0.0597	0.0023	2.3880e-08

Table (3.2): Simulation Results for **Robust Ridge regression** method when β_0=- 0.4, β_1=1.1, β_2=0.3 in **normal** data.

Sample Size	**Bias of the parameters**			***MSE***	MSE_{model}
	$\widehat{\boldsymbol{\beta}}_0$	$\widehat{\boldsymbol{\beta}}_1$	$\widehat{\boldsymbol{\beta}}_2$		
10	0.6679	0.7769	0.0581	7.0027e-04	7.5719e-08
30	0.6727	0.7833	0.0583	0.0011	5.2035e-09
60	0.6737	0.7851	0.0585	0.0013	1.2200e-09
100	0.6741	0.7855	0.0584	0.0013	4.3216e-10

Table (3.3)**:** Simulation results for **Ridge regression** method when β_0= -0.4, β_1=1.1, β_2=0.3 in **multicollinearity** and without **Outliers.**

Sample Size	**Bias of the parameters**			MSE	MSE_{model}
	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$		
10	0.4955	1.0158	0.0435	0.0253	4.7929e-07
30	0.5988	0.9884	0.0378	0.0118	9.3431e-08
60	0.6350	0.9809	0.0363	0.0095	2.6948e-08
100	0.6508	0.9778	0.0357	0.0087	1.0076e-08

Table (٣.4): Simulation results for **Robust Ridge regression** method when β_0=-0.4, β_1=1.1, β_2=0.3 in **multicollinearity** without **Outliers.**

Sample size	**Bias of the parameters**			MSE	MSE_{model}
	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$		
10	0.6696	0.7848	0.0556	9.8159e-04	6.2070e-11
30	0.6743	0.7910	0.0561	0.0014	6.8701e-12
60	0.6755	0.7929	0.0562	0.0015	1.8384e-12
100	0.6760	0.7939	0.0563	0.0015	6.5424e-13

Table (3.5): Simulation Results for **Ridge regression** method whenβ_0=-0.4 β_1=1.1, β_2 =0.3 With **Outliers.**

Sample size	**Bias of the parameters**			**MSE**	$\mathbf{MSE_{model}}$
	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$		
10	0.5883	1.0011	0.0405	0.0254	4.9371e-07
30	0.7222	0.9814	0.0365	0.0162	1.4423e-07
60	0.7693	0.9800	0.0362	0.0178	6.9743e-08
100	0.7911	0.9812	0.0364	0.0192	4.2426e-08

Table (3.6): Simulation Results for **Robust Ridge regression** method when β_0=- 0.4, β_1=1.1, β_2=0.3 With **Outliers**

Sample size	Bias of the parameters			MSE	MSE_{model}
	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$		
10	0.6569	0.8376	0.0598	0.0067	2.6747e-09
30	0.6799	0.8795	0.0632	0.0091	5.3833e-10
60	0.6905	0.9064	0.0654	0.00122	2.4119e-10
100	0.6962	0.9198	0.0664	0.0139	1.2118e-10

5. The Conclusions:

- (**Robust Ridge**) Has proved method in the case of normal data their efficiency when there is a problem multicollinearity where the (MSE=0.0087), and the best (MSE_{model}=1.0076x10^{-8}).
- Proven method (Robust Ridge) in the case of contaminated data and there is a problem where multi-linear efficiency given less bias in the sample size (10) while given best value of (MSE=0.067), and the best MSE_{model} is equals (1.2118x10^{-10}) in the sample size (100).
- There is need to rely on test data the problem of multicollinearity in order to address this problem.
- Researchers should be familiar with the test problem and its exact distribution.
- Further studies are to be done to acldress the effect the problem of multicollinearity hypothesis testing.

References

[1] Arslan, O., & Bill, N. (1996). **"Robust ridge regression estimation based on the GM-estimators"**. Journal of Mathematical and Computational Science, 9 (1), 1-9.

[2] Barnet, V. and Lewis, T., **"Outliers in Statistical Data"** John wiley & Sons, 3rd edition, 1944 http://en.wikipedia.org/wiki/outlier

[3] Bianco, A. And Yohai, V. J. (1996), **"Robust Estimation In The Logistic Regression Model , Robust Statistics"** , Data Analysis And Computer Intensive Methods, Proceedings Of The Work Shop In Honor Of Peter J. Huber, H Rieder (Ed.),Lecture Notes Statistics 109 ,17 -34 , New York: Springer.

[4] Banks , J. Carson , B. Nelson , D. Nicol" **Discrete Event System Simulation"** prentice Hall . p.3 ISBN 0- 13 – 088702 -1 , 2001.

Htt://en.wikipedia.org /wiki/Simulation

[5] Dorugade , A.V, Kashid, D.N.(2011). **"Parameter Estimation Method in Ridge Regression".**

[6] Esteban, F. & Jose, G. (2001). **" Robust Logistic Regression For Insurance Risk Classification "** , Universidad Carlos I I I de Madrid , call Madrid , 126 . **www.docubib.uc3m.es/WORKINGPAPERS/wb016413.pdf**

[7] Gramacy, R.B., Polson, N.G.(2010). **"Simulation – based regularized Logistic regression.**

[8] Hoerl , A. E. and R. W. Kennard .(1970) , *" Ridge Regression Applications to No orthogonal problems "* , Technimetrics , Vol.12 , No.1 , pp.69-82 .

[9] Kapur . J . N., and Saxena . H.C., **"Mathematical Statistic",** S . CHAND & Company LTD., 2009 .

[10] Patrick L. Harrington Jr.,(2011), **"Robust Logistic Regression with Bounded Data Uncertainties".**

[11] Peter, J. Hebber ,Elvezio ,M . (2009), **"Robust statistics"**, 2nd Edition, John Wiley, sonsltd Canada.

[12] Simpson, J. R., & Montgomery, D. C. (1996). **"A biased robust regression technique for combined outlier-multicollinearity problem"**

[13] Srivastava. N. (2005). **"A Logistic Regression Model For Predicting The Occurrence Of Intense Geomagnetic Storms" .** Annales Geophysicae, 23, 2969 -2974. **www.ann-geophys.net /23/2969/2005.pdf**

[14] Wisnowski, J. W., Simpson, J. R., & Montgomery, D. C. (2002), **"An improved compound estimator for robust regression".**Communications in Statistical Simulations,31(4), 653-672, with Bounded Data Uncertainties.

[15] Wendy , L. ,Angel , R. (2008) , **"Computational statistics Hand book with Matlab "** , 2nd Edition , chapman and Hall , USA .

المقدرات الحصينة لنماذج الانحدار اللوجستي مع وجود مشكلة التعدد الخطي أو

البيانات الشاذة

الخلاصة

مشكلة التعدد الخطي تظهر في حالة وجود العلاقة بين المتغيرات التفسيرية (X_s')

مما يسبب عدم دقة مقدرات معالم الأنموذج وظهور التحيز فيها وبالتالي تصبح ذات تباين عالي مما يؤثر على المرحلة اللاحقة من الاستدلال الإحصائي(الاختبارات). حيث تفقد طريقة المربعات الصغرى الاعتيادية(OLS) خصائصها في امتلاكها اقل تباين، لذلك يجب البحث عن طرائق لها القابلية على تجاوز هذه المشكلة ومنها طريقة (Ridge).

بالإضافة إلى أن بحثنا يهتم بمعالجة المشاكل الأخرى ومنها فقدان خاصية التوزيع الطبيعي أو مشكلة التلوث من خلال استخدام طرائق مزدوجة مع (Ridge) ومنها (Robust Ridge).

حيث نفذت هذه الدراسة من خلال تجارب المحاكاة في توليد بيانات النماذج حيث اظهر البحث أن طريقة (Robust Ridge) هي من أفضل الطرائق لمعالجة هذه المشكلة في البيانات الملوثة والطبيعة لتقدير معلمات أنموذج الانحدار اللوجستي.