Zagazig Veterinary Journal, ©Faculty of Veterinary Medicine, Zagazig University, 44511, Egypt.
Volume 46, Number 2, p. 146-153, June 2018
DOI: 10.21608/zvjz.2018.14387

RESEARCH ARTICLE

Modeling of Biological Data Based on Regression Methods

Khairy M. El-Bayomi[1], Elhoussainy A. Rady[2], Mahmoud S. El-Tarabany[1] and Hala I. Mahmoud[1*]

[1]Animal Wealth Development Department, Faculty of Veterinary Medicine, Zagazig University, 44511, Egypt

[2]Applied Statistics and Econometrics Department, Institute of Statistical Studies and Research, Cairo University

Article History: Received: 25/4/2018 Received in revised form: 15/5/2018 Accepted: 20/5/2018

Abstract

The ordinary least square estimates of multiple regression parameters is characterized by low bias and large variance leading to poor performance in both prediction and interpretation of the regression model under study. Penalized regression techniques represented in ridge, lasso and elastic net were used to improve the ordinary least square estimates performance. Categorical regression algorithm provides efficient procedure for computing the regression coefficients of ridge, lasso, and elastic Net models. The statistical analysis was done on ten single nucleotide polymorphisms simulated data with strong linkage disequilibrium as predictors of a continuous phenotypic trait. The coefficients were 39%, 34%, 29% and 28% for ridge, elastic net, lasso and stepwise multiple regression methods, respectively. The current study finished that ridge regression followed by elastic net regression performed better than the other regression methods.

Keywords: Ridge, Lasso, Penalization, Regression, Elastic net.

Introduction

Predicting response variable using the predictors is the most important objective of the model building. It is good to have a model that is reasonably easy to construct and interpret and predicts in a good way. So, the assessment of performance and goodness of fit of a predictive model is of a practical importance. The multiple linear regression models were used extensively in many scientific and biological fields. Ordinary least square (OLS) estimates of the regression parameter have low bias with arge variance leading to poor performance in both prediction and interpretation. Penalization and shrinking techniques were introduced to improve OLS estimates [1]. Single nucleotide polymorphisms (SNPs) could be highly correlated as linkage disequilibrium and the standard multiple regression analysis did not fit with these data due to their high dimensionality and correlation structure. Recently, penalized regression methods were used in the analysis of high dimensional data [2]. The penalization methods also called shrinkage methods that introduce a penalty on the size of the regression coefficients. Ridge, lasso and elastic net regression techniques are of the continuous penalization methods, but there are many other types of regression with the discrete penalization methods [3]. Ridge regression provides a means of addressing the problem of collinearity without removing variables from the original set of independent variables and it was used in a large scale data analysis scenarios, including marker selection, expression data analysis and genetic association studies when SNPs were in high linkage disequilibrium [2,4]. Ridge regression minimizes the residual sum of squares subject to a penalty of the ℓ2-norm on the regression coefficients [1]. The lasso is not robust to high correlations among predictors and will choose one and ignore the others and break down when all predictors are identical. The lasso regression estimator uses the ℓ1 penalized least squares parameter on regression coefficient. Lasso penalty expects many coefficients close to zero, and only a small subset to be larger and none zero [5]. Whether

***Corresponding author e-mail: (drhalastat@gmail.com), Animal Wealth Development Department, Faculty of Veterinary Medicine, Zagazig University, 44511, Egypt**.

there are a group of variables with high multicollinearity, lasso will select only one variable without caring which one is selected [6]. Elastic net simultaneously selects the variables automatically with continuous shrinkage and has the property to select groups of correlated variables. It shrinks the regression coefficients by combining L1-norm penalty and L2-norm penalty together. The elastic net identify a higher number of correctly influential variables than the lasso technique, and has lower false positive rate than ridge regression [6,7]. The instability of the lasso regression technique when independent variables are highly correlated as in SNPs in high linkage disequilibrium is overcome using the elastic net that was proposed for analyzing high dimensional data [5]. An important progress in the analysis of multidimensional data was the optimal assignment of quantitative values to qualitative scales. This type of optimal quantification (scaling or scoring) is a general approach to treat multivariate categorical data [8]. Categorical regression algorithm provides efficient procedure for computing the regression coefficients in the models of ridge regression, lasso and elastic Net [9]. This study aimed to assess the performance of penalized categorical regression techniques comparing to ordinary least square multiple regression performance based on coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE) and Theil's U statistic.

Materials and Methods

Data source

A simulated data was obtained from a previously published study [10]. The researcher used Windows QTL Cartographer programme version 2.5 for simulation. The parameters setting used in data simulation showed in Table (1). Dependent variable defined as $y_i \sim N$ (120, 1.23), represented quantitative trait of interest and the independent variable defined as SNP genotypes, where 2, 1 and 0 were used to denote AA, Aa and aa genotypes, respectively.

Table 1: Parameters setting for data simulation

Trait and Population	
Replications	1
Sample size	100
Population	F2
Trait mean	120
Markers	
Total chromosome number	1
Marker Numbers for Ch1	10
Average marker distance	10 cM
Variations of the marker Positions (%)	0
Marker Genotypes	
AA	2
Aa	1
aa	0
QTL	
QTL No.	1
QTL	3 cM

cM: centimorgan is a genetic measure showed as the distance between chromosome positions QTL: Quantitative trait locus. A quantitative trait locus is a region of DNA which is associated with a particular phenotypic trait.

Statistical analysis

The simulated data was used to investigate the performance of penalized categorical regression models in prediction phenotypic trait from SNPs marker predictors and compare between their predictive efficiency and that for different multiple linear regression algorithms.

Multiple regression analysis

Data were described then tested for independency, linearity and homoscedasticity of the error term using histogram, scatter plot and residual analysis plot. Linkage disequilibrium (LD) estimate (D') among SNPs was calculated using Cubex online calculator [11]. Data were examined using enter, forward, backward and stepwise multiple regression methods. In general, multiple regression models is represented by the following equation

$$Y_{pop} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k$$

[12]. Each model of multiple regression methods was evaluated after rebuilding and removal of non-significant SNPs.

Penalized categorical regression analysis

Categorical ridge, lasso and elastic net regression analysis were applied then the best shrinkage parameter that resulting in an optimal model that overcomes the OLS problems was chosen. The predicting model was rebuild after removing non-significant SNPs, then the models performances were evaluated. A categorical regression version for ridge, lasso, and elastic net regression is represented as:

$$L^{ridge}(\beta_j) = \left\| y - \sum_{l \neq j} \beta_l x_l - \beta_j x_j \right\|^2 + \lambda_2 \beta_j^2 + \lambda_2 \sum_{l \neq j} \beta_l^2$$

$$L^{lasso}(\beta_j) = \left\| y - \sum_{l \neq j} \beta_l x_l - \beta_j x_j \right\|^2 + \lambda_1 w_j \beta_j + \lambda_1 \sum_{l \neq j} w_l \beta_l$$

$$L^{enet}(\beta_j) = \left\| y - \sum_{l \neq j} \beta_l x_l - \beta_j x_j \right\|^2 + \lambda_2 \beta_j^2 + \lambda_1 w_j \beta_j + \lambda_2 \sum_{l \neq j} \beta_l^2 + \lambda_1 \sum_{l \neq j} w_l \beta_l$$

Where y_i is the i_{th} observations on the dependent variable, x_{ij} is the i_{th} observation for j_{th} independent variables, β_0 is the intercept coefficients, β_i is the ith regression coefficients, $\lambda \geq 0$ is the tuning (regularization) parameter which regulates the strength of the penalty (linear shrinkage), w_l and w_j are either +1 or −1 depending on the sign of the corresponding β_l and β_j [1,5,9]. The following performance measures estimate model performance and accuracy were used to evaluate and compare between different regression models performance.

1. $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_1)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
2. RMSE = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_1)^2}$
3. MAE = $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_1|$
4. MAPE = $\frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_1|}{y_i} \times 100$
5. Theil'sU-statistic = $\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_1)^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_1)^2}}$ [1].

The data were statistically analyzed using SPSS (Statistical Package for Social Science) version 21 [13].

Results

It is too helpful to perform regression diagnostic analysis as exploratory steps to determine the characteristics of the traits under the study. The phenotypic trait is normally distributed, linearly related with the predictor SNPs and there is no heteroskadisticity in variance of the residuals. Examining the linkage disequilibrium among SNPs showed that D' estimate reached to 0.8 and more (Table 2).

Table 2: Matrix of D' of all pairwise linkage disequilibrium between SNPs

	SNP2	**SNP3**	**SNP4**	**SNP5**	**SNP6**	**SNP7**	**SNP8**	**SNP9**	**SNP10**
SNP1	0.89	0.69	0.53	0.39	0.37	0.30	0.22	0.12	0.05
SNP2		0.89	0.74	0.60	0.53	0.37	0.35	0.18	0.02
SNP3		-	0.85	0.71	0.63	0.52	0.45	0.30	0.11
SNP4			-	0.85	0.71	0.61	0.55	0.41	0.18
SNP5				-	0.83	0.69	0.64	0.53	0.33
SNP6					-	0.83	0.80	0.62	0.40
SNP 7						-	0.89	0.75	0.49
SNP8							-	0.86	0.60
SNP9								-	0.80

SNP: Single neocleotide polymoephism

Each SNP consider a different marker for a phenotypic trait

Multiple regression analysis results

Regression coefficients of all SNPs resulting from enter method was non-significant (P > 0.05), while stepwise and forward methods selected SNP1 and SNP3 that had significant coefficients with P value 0.002 and 0.048, respectively. Backward method selected SNP1, SNP3 that were significant predictors and SNP9 which was non-significant with P value 0.077 was deleted. Backward** (** means backward regression after removal of non-significant SNP9) gave a model including a significant SNP1 and SNP3 with P value 0.002 and 0.048, respectively.

The final model size of other multiple regression variable selecting methods (stepwise, forward and backward**) after removal of any non-significant SNPs included only 2 SNPs (SNP1 and SNP3) that is about 20% of SNPs under the study. Regression coefficients of SNP1 and SNP3 corresponding to stepwise, forward and backward** were 0.618 and 0.403, respectively (Table 3).

Table 3: Significance of SNPs Regression Coefficient in multiple regression methods

Multiple regression methods	SNPs included in the Model	β regression coefficient	Std. error	Significance
Enter	SNP1	0.550	0.287	0.059
	SNP2	0.144	0.390	0.770
	SNP3	0.241	0.373	0.520
	SNP4	0.163	0.355	0.647
	SNP5	0.029	0.347	0.934
	SNP6	0.002	0.360	0.995
	SNP7	-0.008	0.347	0.983
	SNP8	0.253	0.352	0.475
	SNP9	-0.533	0.309	0.088
	SNP10	0.109	0.229	0.636
Stepwise	SNP1	0.618	0.196	0.002*
	SNP3	0.403	0.201	0.048*
Forward	SNP1	0.618	0.196	0.002*
	SNP3	0.403	0.201	0.048*
Backward	SNP1	0.593	0.195	0.003*
	SNP3	0.489	0.205	0.019*
	SNP9	-0.245	0.137	0.077
Backward **	SNP1	0.618	0.196	0.002*
	SNP3	0.403	0.201	0.048*

*Significant (p-value ≤ 0.05)
Std. error: standard error
** After removal of non-significant SNPs

Penalized categorical regression results

Categorical ridge regression was performed, the optimal model was chosen at penalty parameter 0.9 at which the mean square error (MSE) that represents the variance of the regression coefficient decreased to the minimum (0.739) against other penalty parameters. The non-significant SNPs were deleted remaining SNP1, SNP3 and SNP9 with P value less than 0.05 at 0.1 penalty.

Categorical lasso regression was also performed, the optimal model was chosen at penalty parameter 0.2 at which MSE decreased to the minimum (0.77) against other penalty parameters. The non-significant SNPs were deleted remaining only SNP1 with P value less than 0.0001 at 0.00 penalty. Elastic net (EN) is a combined procedure of ridge and lasso regression. Categorical EN was done, the penalty parameters at which the optimal model was build were 1 for ridge procedures and 0.4 for lasso procedures. At these penalties, MSE decreased to the minimum (0.722) against other penalty parameters. The non-significant SNPs were deleted remaining only SNP1, SNP 2 and SNP3 with P value less than 0.0001at 0.8 ridge penalty and 0.3 lasso penalty. Significance of SNPs regression coefficient in penalized regression methods was illustrated (Table 4).

Table 4: Significance of SNPs Regression Coefficient in Penalized regression methods

Penalized categorical regression methods	SNPs included in the Model	β regression coeffecient	Std. error	Significance
Ridge (penalty = 0.1)	SNP1 SNP3 SNP9	0.656 0.319 -0.176	0.077 0.079 0.083	0.000* 0.000* 0.013*
Lasso (penalty = 0.0)	SNP1	0.869	0.153	0.000*
Elastic net (Ridge penalty=0.8) (Lasso penalty= 0.3)	SNP1 SNP2 SNP3	0.238 0.203 0.187	0.064 0.056 0.057	0.000* 0.000* 0.000*

*Significant (p-value ≤ 0.05)
Std. error: standard error

Table 5: Comparative performance of multiple regression methods and penalized regression models

Regression methods	R^2	RMSE	MAE	MAPE	Theil's U-statistic ($\times 10^{-4}$)
Enter	0.31	0.92	0.76	0.63	0.6396
Stepwise	0.28	0.94	0.77	0.64	0.6535
Forward	0.28	0.94	0.77	0.64	0.6535
Backward	0.28	0.94	0.77	0.64	0.6535
Ridge	0.39	0.79	0.66	0.55	0.5492
Lasso	0.29	0.96	0.77	0.64	0.6674
Elastic net	0.34	0.81	0.68	0.57	0.5657

Enter method has the highest R^2 (31%) against other methods of multiple regression. Ridge regression and elastic net models R^2 were 39% and 34% respectively, while other regression models coefficient of determination ranged from 28% to 31%. Ridge regression and elastic net models RMSE were 0.79 and 0.81 respectively, while other regression models RMSE ranged from 0.92 to 0.96. Ridge regression and elastic net models MAE were 0.66 and 0.68, respectively, while other regression models MAE ranged from 0.76 to 0.77. Ridge regression and elastic net models MAPE were 0.55 and 0.57, while other regression models MAPE ranged from 0.63 to 0.64. Ridge regression and elastic net models U- statistic were 0.5492×10^{-4} and 0.5657×10^{-4}, while other regression models MAPE ranged from 0.6396×10^{-4} to 0.6674×10^{-4} (Table 5).

Discussion

The current research addressed an important problem in the field of genetics and biological science, concerning how to deal with highly correlated explanatory variables or predictors. Identifying the SNPs biomarkers significantly affect specific phenotypic trait was our aim to build a predictive model with high performance. In a previous study, it was reported that linkage disequilibrium D' estimate of more than 0.55 represented strong linkage disequilibrium [14]. In this study, D' reached 0.80 and more, therefore we considered the SNPs in strong linkage disequilibrium. Traditional multiple regression procedures are highly problematic when strong LD exists among the predictors and the multicollinearity (LD) among the SNPs predictors can produce misleading results and interpretations. Herein, traditional multiple linear regression with enter method resulting in 100% non-significant SNPs with P values ranged from 0.059 to 0.995. This result agreed with Malo and Coauthors [14], who found that multiple regression model suffering from multicollinearity among SNPs resulting in several missing coefficient values and non-significant P value that ranged from 0.118 to 0.914.

The results in Table 3 show that that SNP1 is more associated with the phenotypic trait more than SNP3. We decided to apply penalized regression procedures that overcome multicollinearity problems [15]. The stepwise selection procedures in ordinary regression are poorly in variable selection, regression coefficients estimation and its standard errors, especially when multicollinearity is present.

Categorical ridge regression was performed. The optimal model that was chosen at 0.9 penalty has MSE=0.739 that was less than multiple regression MSE which was 0.94. This indicated that ridge regression decreased the variation in predicted error which is comparable with others [1]. All SNPs included in the model with 30% SNPs were significant. After deleting non-significant SNPs and rebuilding the model, it was found that SNP1 has the highest regression coefficient that equals 0.356.

Categorical lasso regression was also performed and the optimal model that was chosen at 0.2 penalty has MSE=0.76 that was also less than multiple regression MSE that was 0.94. This indicated that lasso regression decreased the variation in predicted error which was previously stated by others [1]. Sixty percent (60%) of SNPs included in the model (SNP1, SNP2, SNP3, SNP4, SNP9 and SNP10) and all are non-significant except SNP1.

Other SNPs coefficients were shrunk to zero as reported in another work [5] that the lasso penalty expected many regression coefficients to be close to 0 and only a small subset may be larger (and non-zero). Categorical elastic net regression was also performed and the optimal model that was chosen at 1 and 0.4 penalties of ridge and lasso procedures has MSE=0.72 that was also less than multiple regression MSE that was 0.94 as documented previously [1]. This also indicated that elastic net regression decreased the variation in predicted error. SNP 4 was non-significant and deleted from the model, then it was found that SNP1 has the highest regression coefficient that equals 0.238.

Ridge regression outperformed ordinary multiple regression followed by elastic net regression in the prediction of phenotypic trait based on SNPs predictors. A previous study [14] stated that ridge regression is better than ordinary multiple regression and traditional single-locus-based analyses. Several authors discouraged the use of stepwise algorithms and reported that lasso procedures performed better than backward and forward stepwise algorithm which was slightly agreed with our results as R^2 of lasso, backward and forward stepwise algorithms were 24%, 26% and 26% respectively [15]. Our results were also supported by previously published articles [5,16], on which the performance of shrinkage ridge regression estimators is superior to lasso estimators when predictors are highly correlated. The performance parameter results of the current study was contrary to several studies [1-3] as the prediction error, elastic net performed better than the ordinary multiple regression, ridge regression and other regulatory regression methods, but we agreed the elastic net was better than lasso. The difference may be due to different analytical statistical programs, different sample size or different experimental units and variables.

Conclusion

The Ridge regression and elastic net technique outperformed the other methods ordinary multiple linear regression methods (forward, backward, stepwise and enter) and other penalized regression technique for predicting the quantitative phenotypic trait regressed on SNPs predictors with strong LD.

Conflict of interest

None of the authors have any conflict of interest.

References

[1]Kumar, H. and Hooda, B.K. (2015): Comparison of penalized and multiple regression for prediction of milk yield in cross bred cattle. Int J Agric Stat Sci, 11(1): 151-154.

[2]Cule, E.; Vineis, P. and De Lorio M. (2011): Significance testing in ridge regression for genetic data. BMC Bioinformatics, 12: 372.

[3]Acharjee, A.; Finkers, R; Visser, R.G. and Maliepaard, C. (2013): Comparison of regularized regression methods for omics data. Metabolomics, 3(3): 126.

[4] McDonald, G.C. (2009): Ridge Regression. WIREs Comp Stat, 1: 93-100.

[5]Ogutu, J.O.; Schulz-Streeck, T. and Piepho, H. (2012): Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC Proceedings, 6(suppl 2): S10.

[6]Liu, W. and Li, Q. (2017): An efficient elastic net with regression coefficients method for variable selection of spectrum data. PLOS One, 12(2): e0171122

[7]Waldmann, P.; Mészáros, G.; Gredler, B.; Fuerst, C. and Sölkner, J. (2013): Evaluation of the lasso and the elastic net in genome-wide association studies. Front Genet, 4: 270.

[8] Meulman, J.J. (1998): Optimal scaling methods for multivariate categorical data analysis. SPSS white article. SPSS Inc.

[9] Van der K.A.J. (2007): Prediction accuracy and stability of regression with optimal scaling transformations. PhD Thesis, Faculty of Social and Behavioral Sciences, Leiden University, Netherlands.

[10]Yousef, M.A. (2014): Statistical Modelling for Analysis of Some Genetic Traits. MVSc Thesis, Faculty of Veterinary Medicine, Zagazig University, Egypt.

[11]Gaunt, T.R.; Rodriguez, S. and Day, I.N.M. (2007): Cubic exact solutions for the estimation of pairwise haplotype frequencies: implications for linkage disequilibrium analyses and a web tool 'CubeX'. BMC Bioinformatics, 8: 428.

[12]Aviva, P. and Watson, P. (2013): Statistics for veterinary and animal science. 3rd edition. USA: John Wiley & Sons.

[13] SPSS V21 (2012): IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp.

[14]Malo, N.; Libiger, O. and Schork, N.J. (2008): Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. The Am J of Hum Genet, 82 (2): 375-385.

[15]Morozova, O.; Levina, O.; Uuskülam, A. and Heimer, R. (2015): Comparison of subset selection methods in linear regression in the context of health-related quality of life and substance abuse in Russia. BMC Med Res Methodol, 15: 71.

[16]Yuzbasi, B.; Ahmed, S.E. and Gungor, M. (2017): Improved penalty strategies in linear regression models. REVSTAT Stat J, 15(2): 251-276.

الملخص العربي

نمذجة البيانات البيولوجية إعتمادا علي طرق الإنحدار

خيري محمد البيومي[1]، الحسيني عبد البر راضي[2]، محمود صلاح الطرباني[1] وهالة اسماعيل محمود[1]*

[1]قسم تنمية الثروة الحيوانية – كلية الطب البيطري – جامعة الزقازيق

[2] قسم الإحصاء والإقتصاد القياسي – معهد الدراسات والبحوث الإحصائية – جامعة القاهرة

تتميز طريقة تقديرات المربعات الصغري لمعلمات الإنحدار المتعدد عموماً بوجود تحيز منخفض وتباين كبيرمما يؤدي إلى ضعف الأداء في كل من التنبؤ والتفسيرات الخاصه بنموذج الإنحدار تحت الدراسه. أستخدمت تقنيات الإنحدار الجزائي ممثلة في إنحدار الطرف، إنحدار Lasso، وإنحدار الشبكة المرنة. خوارزمية الإنحدار الفئوي يقوم بإجراءات فعالة لحساب معامل الإنحدار في نماذج إنحدار الطرف، لاسو، والشبكة المرنة، حيث تم إجراء التحليل الإحصائي على بيانات محاكية لعشرة متغيرات لتعدد أشكال النيوكليوتيدات المفرده ذات الإختلال في التوازن الإرتباطي كمتغير متنبئ لصفة مظهرية مستمره. نسبة معامل التحديد لنموذج إنحدار الطرف ونموذج إنحدار الشبكة المرنة كان 39٪ و34٪ على التوالي وكانت 29٪ و28٪ لطرق Lasso والإنحدار المتعدد علي التوالي. وعليه فإنه قد تم إستنتاج أن إنحدار الطرف يمكن الإعتماد عليه بشكل أفضل من طرق الإنحدار الأخرى يليه تقنية إنحدار الشبكة المرنة.