© 2019 ILEX PUBLISHING HOUSE, Bucharest, Romania
http://rjdnmd.org
Rom J Diabetes Nutr Metab Dis. 26(4):371-379
doi: 10.2478/rjdnmd-2019-0040

THE USAGE OF LASSO, RIDGE, AND LINEAR REGRESSION TO EXPLORE THE MOST INFLUENTIAL METABOLIC VARIABLES THAT AFFECT FASTING BLOOD SUGAR IN TYPE 2 DIABETES PATIENTS

Arash Farbahari [1], ***Tania Dehesh*** [1,✉], ***Mohammad Hossien Gozashti*** [2]

[1] Department of Biostatistics and Epidemiology, Kerman University of Medical Sciences, Kerman, Iran
[2] Department of Endocrinology, Kerman University of Medical Sciences, Kerman, Iran

received: November 09, 2019 *accepted*: December 14, 2019
available online: February 01, 2020

Abstract

Background and aims: *To explore the most influential variables of fasting blood sugar (FBS) with three regression methods, to identify the existence chance of type 2 diabetes based on influential variables with logistic regression (LR), and to compare the three regression methods according to Mean Squared Error (MSE) value.* ***Material and Methods:*** *In this cross-sectional study, 270 patients suffering from type 2 diabetes for at least 6 months and 380 healthy people were participated. The Linear regression, Ridge regression, and Least Absolute Shrinkage and Selection Operator (Lasso) regression were used to find influential variables for FBS.****Results:*** *Among 15 variables (8 metabolic, 7 characteristic), Lasso regression selected HbA1c, Urea, age, BMI, heredity, and gender, Ridge regression selected HbA1c, heredity, gender, smoking status, and drug use, and Linear regression selected HbA1c as the most effective predictors for FBS.* ***Conclusion:*** *HbA1c is the most influential predictor of FBS among 15 variables according to the result of three regression methods. Controlling the variation of HbA1c leads to a more stable FBS. Beside FBS that should be checked before breakfast, maybe HbA1c could be helpful in diagnosis of Type 2 diabetes.*

key words: *Type 2 Diabetes, Linear regression, Lasso regression, Ridge regression*

Background and aims

Diabetes is one of the most prevalent chronic diseases which cause discomforts and incur huge costs to the patients worldwide [1]. Diabetes is a metabolic disorder caused by impaired insulin secretion and action the most common causes of which are inheritance and environmental factors [2]. Globally, more than 520 million people are suffering from this disease [1]. According to the annual report of the World Health Organization (WHO), the number of patients with diabetes will increase to approximately 7 million people until 2030 [3]. Today overnutrition, low-fiber diets, also sedentary lifestyle, sleep deprivation, and depression are the major reasons of type 2 diabetes development [4,5]. Diabetes increases the risk of many other diseases including stroke, cerebral small vessel disease, reduced vision, as well as renal and heart disorders [2,6]. The most prevalent type of diabetes is type 2 diabetes [7].

✉ Haft Bagh Alavi, Kerman University of Medical Sciences Telephone/Fax: +98-34-31325069/ 7617647633; *corresponding author e-mail*: tania_dehesh@yahoo.com

In most, especially obese people, inability to utilize/clear glucose that has been ingested leads to type 2 diabetes [8]. One of the methods for diagnosis of type 2 diabetes is performing FBS level test. The patients with symptoms of type 2 diabetes are referred to clinical laboratories for FBS checking, where if the level of FBS is above 126 mg/dl , the FBS test must be repeated for certainty [9].

Type 2 diabetes should be managed through regular blood sugar tests. The blood metabolic variables including glycosylated hemoglobin (HbA1c), cholesterol (CHOL), triglycerides (TG), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), thyroid-stimulating hormone (TSH), creatinine (Cr), and carbamide (Urea) may affect the value of FBS [10]. So it needs to be tackled to prevent the progression. There are several studies on assessing the positive effect of physical activities on diabetes treatment [11,12], but investigating the effect of blood metabolic variables on FBS which are routinely reported in blood tests was rarely done.

One of the most widely used statistical methods in prediction and exploration is the regression methods [13]. Linear regression model examines the effect of a number of explanatory (predictor) variables on a dependent (response) variable [14]. In clinical researches usually predictor variables are strongly correlated, and this correlation must be considered. New regression methods have recently been introduced, which could better manage the multicollinearity (correlations between predictor variables) and lead to better predictors selection [15].

As mentioned earlier, type 2 diabetes is primarily diagnosed based on FBS value, where there are many metabolic variables which may influence the FBS value. The aim of the present study is first to find the most influential blood metabolic variables (and characteristic variables) that affect FBS value in patients with type 2 diabetes based on three different regression methods (Linear regression, Lasso regression, Ridge regression), second to calculate the chance of catching type 2 diabetes based on these variables with LR regression, and third to compare the three regression methods according to MSE value.

Material and Methods

Data Collection and Preparation

The data of this cross-sectional study were collected from October 2017 to April 2018. The study population consisted of 270 type 2 diabetic patients over 18 years of age who had been referred to Besat laboratory (the main laboratory of diabetics patients), in the center of Kerman province, southeastern Iran. The metabolic variables affecting FBS including HbA1c, TG, Urea, Cr, CHOL, HDL-C, LDL-C, and TSH were collected from these individuals' blood tests under the supervision of an endocrinologist. The characteristic variables (inclusive age, gender, smoking status, drug use, and heredity) were also collected by self-report. In addition, BP and BMI were collected as two other effective predictors of FBS. BP was measured by Cuff in sitting manner twice (first, on blood test day and the second on the next day, when they came to get the test result). The participants under the age of 18 and pregnant women as well as patients with other chronic diseases (type 1 diabetes, kidney disorder), were excluded from the study (8 patients). There were also 380 healthy people who did not have any chronic diseases and came only to check their annual blood test, and they completely match with patients group based on characteristic variables and were participated for LR regression. Ethical approval was granted by the Ethics Committee of Kerman University (reference number:

IR.KMU.REC.1397.174). This study was conducted in compliance with the Helsinki Declaration. All the participants gave their informed consent.

Linear Regression

The most conventional regression method widely used across all kinds of research is linear regression. Basically, this method involves evaluating the relationship and the effect of some predictor variables on a quantitative response variable. This method was introduced to statistics by Sir Francis Galton in the late 19th century [16]. All regression methods have similar equation as follows:

$$Y_i = B_0 + B_1X_1 + \cdots + B_pX_p$$

Where, Y_i represents the response variable, B_i is the coefficient of, X_i which shows the rate of X effects on Y variable.

Lasso Regression

Lasso is a regression analysis method which performs penalized estimation in order to enhance the prediction accuracy and interpretability via the best predictor selection. This method was first proposed by Tibshirani in 1996 [17]. Lasso regression has great performance and efficiency when the number of predictors are greater than observations. This method yields some non-significant coefficients toward zero and provides a desirable subset of predictor variables [18]. In fact, this method completely eliminates unimportant predictors.

Ridge Regression

Another penalized regression method proposed by Hoerl and Kennard in 1970 is Ridge regression. The goal of this method is to select predictor variables when there is multicollinearity (severe correlation between predictors) between them [19]. This method does not eliminate any of the predictor variables from the model. The result of Ridge estimation method is bias parameters, but with minimal variance [20].

The difference between Lasso and Ridge is that in Lasso regression, some insignificant coefficients can become zero and be eliminated from the final model, but in Ridge method all coefficients remain in the model even if they are not influential. The researchers of the present study compared the above three methods according to their MSE values. The MSE measures the average squared difference between the estimated values and what is estimated. Low MSE values in regression models could be used as goodness of fit index. The variance inflation factor (VIF) is an index for assessing the multicollinearity. If the VIF value exceeds 10, it can be assumed that the regression coefficients have been poorly estimated due to multicollinearity [16].

Logistic Regression

Logistic regression is a type of prediction regression model in which the dependent (response) variable is not continuous. In fact the main usage of logistic regression is to predict the dichotomous (the variable that have two values) variable by the other independent variables [16].

The analysis of this study was performed using software R version 4.3.3 with MASS and glmnet packages.

Results

Table 1 compares blood metabolic and characteristics variables between diabetic and healthy people. This table reveals that there were statistically significant differences between diabetic and healthy people in terms of FBS, HbA1c, and age. FBS and HbA1c are significantly higher in diabetic patients than healthy people ($P < 0.001$).

As shown in Table 1, from all 650 people who referred to the laboratory, the mean age of

270 diabetic patients (52.7 years) were significantly higher than 380 healthy people (43.6 years) (P=0.014). The mean BMI value of all the people (27.7) was higher than normal range relative to their age, but the mean BMI value of diabetic patients was not significantly higher than healthy people.

Table 1. The comparison of metabolic and characteristic variables between healthy and patient people

Variable	Total (n=650)	Healthy (n=380)	Patient (n=270)	P- Value
Smoke status (n; %)				
Yes	40 (6.2%)	30 (7.9%)	10 (3.7%)	0.488
No	610 (93.8%)	350 (92.1%)	260 (96.3%)	
Drug use (n; %)				
Yes	70 (10.8%)	40 (10.5%)	30 (11.1%)	0.940
No	580 (89.2%)	340 (89.5%)	240 (88.9%)	
Heredity (n; %)				
Yes	240 (36.9%)	110 (28.9%)	130 (48.1%)	0.114
No	410 (63.1%)	270 (71.1%)	140 (51.9%)	
Gender (n; %)				
men	270 (41.5%)	150 (39.5%)	120 (44.4%)	0.689
women	380 (58.5%)	230(60.5%)	150(55.6%)	
Age (years) (mean; SD)	47.42 (13.9)	43.66 (13.9)	52.70 (10.5)	0.014
BMI (mean; SD)	27.72 (5.1)	27.42 (6.3)	28.14 (4.2)	0.440
BP (mean; SD)	141.86 (21.7)	139.42 (21.9)	145.30 (19.9)	0.274
FBS (mean; SD)	138.32 (58.0)	100.50 (11.5)	191.56 (55.5)	< 0.001
TSH (mean; SD)	2.78 (1.8)	2.62 (1.6)	2.99 (2.1)	0.719
Urea (mean; SD)	29.89 (7.2)	29.26 (6.3)	30.78 (8.5)	0.414
Cr (mean; SD)	0.85 (0.2)	0.85 (0.27)	0.87 (0.2)	0.565
CHOL mean; SD)	170.34 (41.1)	173.13 (38.6)	166.41 (44.9)	0.521
TG (mean; SD)	179.22 (119.6)	167.26 (127.7)	196.0 (107.4)	0.058
HDL-C (mean; SD)	42.57 (11.8)	44.74 (11.1)	39.52 (12.4)	0.081
LDL-C (mean; SD)	107.05 (35.0)	108.60 (28.5)	104.88 (43.0)	0.676
HbA1c (mean; SD)	6.71 (2.3)	5.44 (0.6)	8.50 (2.7)	< 0.001

For discrete variables P –Value calculated with chi-square test. For continuous variables after checking normality test, used independent T-test for normal variables and Mann-Whitney for abnormal variables.

Table 2 presents the VIF values for all the 15 (8 metabolic and 7 characteristics) predictors. The VIF values were above 10 for CHOL and LDL-C, confirming the existence of high multicollinearity.

The results of three regression coefficients are provided in Table 3, with MSE employed for comparison. As it can be observed, Ridge regression keeps all variables even if they are not

very important, but Lasso deletes insignificant predictors. Lasso regression keeps HbA1c, urea, age, BMI, heredity, and gender as the most influential variables for FBS. According to the results of Ridge regression, HbA1c, heredity, gender, smoking status and drug use are the most effective predictors for FBS, respectively. HbA1c has been selected in linear regression. Having a family history of diabetes in all the models has a positive effect on FBS. Lasso regression had the lowest MSE value among three models.

Table 2. The correlation matrix between predictors

Variable	Smoke status	Drug use	Heredity	Gender	Age	BMI	BP	TSH	Urea	Cr	CHOL	TG	HDL-C	LDL-C	HbA1c
VIF	1.55	1.82	1.22	2.27	1.95	1.50	1.69	1.21	1.77	2.23	17.39	5.25	3.73	14.01	1.38

Table 3. The coefficient of different metabolic and characteristic variables on FBS in Distribution regression methods

	Ridge regression		Lasso regression		Linear regression	
variable	Coefficient	MSE	Coefficient	MSE	Coefficient	MSE
Intercept	34.484		–0.876		–26.578	
Smoke status	12.455		0		13.446	
Drug use	4.794		0		16.517	
Heredity	8.548		4.932		17.156	
Gender	–7.379		–3.035		–21.638	
Age	0.670		0.473		0.791	
BMI	0.604		0.262		1.333	
BP	0.059		0		0.276	
TSH	–0.163		0		–0.892	
Urea	0.825		0.835		1.476	
Cr	–1.995		0		–31.318	
CHOL	–0.035		0		–0.701	
TG	0.012	1496.689	0	1393.062	0.095	1528.710
HDL-C	–0.384		0		0.299	
LDL-C	0.035		0		0.791	
HbA1c	9.965		15.204		17.283	

Table 4 shows the results of LR regression. As it is observed, increasing one unit (mmol/mol) in HbA1c increases the chance of Type 2 diabetes more than fifty times. The odds ratio (OR) of gender and heredity are 6.264 and 4.457 respectively, but their effects are not statistically significant. The other variables (BMI, urea, gender, age) did not have significant effects.

Table 4. The effect of important metabolic and characteristic variables on type 2 diabetes

	Coefficient	SE. coef	P-value	OR	95% CI for OR	
Heredity	1.494	1.036	0.149	4.457	0.585	33.976
Gender	1.835	1.473	0.213	6.264	0.349	112.319
Age	-0.018	0.049	0.714	0.982	0.893	1.081
BMI	-0.233	0.147	0.112	0.792	0.594	1.056
Urea	-0.048	0.096	0.614	0.953	0.789	1.150
HbA1c	4.024	1.413	0.004	55.944	3.506	892.704

Discussion

According to the researchers' knowledge, investigating the effect of metabolic variables that were routinely reported in blood tests on FBS with three different regression models is rarely done. This investigation could confirm the role of predictors in controlling the variation of FBS and also diabetes disease. The result of Lasso regression shows that HbA1c had a high positive effect on FBS. So, it seems essential to control HbA1c in order to stabilize FBS variation. Other studies also confirmed that high HbA1c could be a main risk factor of type 2 diabetes [21,22]. The age variable has a positive effect on FBS. This result was expectable because an increase in age leads to an increase in the chance of catching all diseases [23]. BMI has positive effect on FBS. This result is in line with the result of a previous study which showed BMI as the only significant predictor of diabetes. In that study diabetic patients had a higher mean BMI than healthy people [24].

Ridge and Lasso regression models confirmed the effect of gender simultaneously. They showed that men had significantly higher FBS levels relative to women. This result is completely in accordance with the result of previous study [25]. In the present study; smoking status and drug use have positive effect on FBS. Therefore, Active smoking is associated with an increased risk of type 2 diabetes. There is evidence that smokers (especially heavy smokers) tend to have higher BMIs than lighter smokers and even some nonsmokers [26], and increasing the BMI leads to an increase in the chance of type 2 diabetes.

Note that as a main total result, three models jointly introduced HbA1c as the most effective predictor of FBS, which is in line with the result of previous study [27]. Using HbA1c as an influential predictor of FBS could be helpful in diagnosis of type 2 diabetes. In fact it could be used instead of FBS in order to diagnose type 2 diabetes. The WHO consultation concluded that HbA1c can be used as a diagnostic test for diabetes [28]. High value of HbA1c shows the

danger of type 2 diabetes, like the high value of FBS. This metabolic variable does not have the problem of day-to-day variability of glucose values. Checking HbA1c also does not need to be fasting and to have preceding dietary preparations [29].

This study compared three regression models based on MSE values. The lowest value of MSE represented the better model. The results of this study indicated that linear regression is not acceptable, especially in the existence of multicollinearity. As it can be observed, this model has the biggest MSE compared to other models. Lasso regression model had the lowest MSE compare to others. This result is completely in accordance with the results of previous studies that performed simulation to compare Lasso regression with three types of Ridge regression. Using the MSE criterion, they found that Lasso regression performs better than Generalized Ridge Regression (GRR) and Jackknifed Ridge Regression (JRR) [30]. In another study, LASSO method was proposed as a novel method to predict financial market behavior, and the results indicate that the proposed model outperforms the ridge linear regression model [31].

The evidence presented in this study confirmed the usage of Lasso regressions, especially for the clinical researchers. As we know, clinical variables usually show severe correlation. Therefore, the usage of lasso regression is preferred, especially in the existence of many predictors.

The results of the present study indicated that apart from age and gender which are out of our control, people should have structured exercise training that is associated with HbA1c and FBS reduction in patients with type 2 diabetes in order to prevent it [32]. Finally, as we know, diagnosis is more important than remedy, and finding the influential predictors is important in diagnosis. HbA1c is an important predictor of FBS and also type 2 diabetes and sometimes it may work better than other predictors in diagnosis of Type 2 disease.

This study has some limitations. As the main purpose of this study was investigating the effect of metabolic variables, the effect of stress, alcohol consumption, nutrient intakes, and work-related physical activity were not assessed. These factors may have effects on FBS which can be investigated in future researches. This study was a cross-sectional study; therefore, a temporal relationship between predictor variables and diabetes cannot be inferred from these results.

Conclusion

HbA1c has a high role in the variation of FBS. This role is more than many other metabolic variables. HbA1c testing can be performed at any time of the day and without special patient preparation. This advantage makes it more valuable. Considerable caution should be warranted when using linear regression, especially in clinical researches. In clinical research, the Lasso regression is preferred because of multicollinearity.

Conflict of Interest. The authors declare that they have no competing interests.

Acknowledgements. There was no external funding received for this study.

REFERENCES

1. **Png ME, Yoong J, Tan CS, Chia KS**. Excess Hospitalization Expenses Attributable to Type 2 Diabetes Mellitus in Singapore. *Value in health regional issues*: 15:106-11, 2018.

2. **Wang M, Li J, Yeung V, Zee B, Yu R, Ho S et al**. Four pairs of gene–gene interactions associated with increased risk for type 2 diabetes (CDKN2BAS–KCNJ11), obesity (SLC2A9–IGF2BP2, FTO–APOA5), and hypertension (MC4R–IGF2BP2) in Chinese women. *Meta gene* 2:384-91, 2014.

3. **Organization WH**. Global report on diabetes. 2016.

4. **Selvakumar G, Shathirapathiy G, Jainraj R, Paul PY.** Immediate effect of bitter gourd, ash gourd, knol-khol juices on blood sugar levels of patients with Type 2 diabetes mellitus: A pilot study. *Journal of traditional and complementary medicine.* 7(4): 526-31, 2017.

5. **Kolb H, Mandrup-Poulsen T**. The global diabetes epidemic as a consequence of lifestyle-induced low-grade inflammation. *Diabetologia*. 53(1): 10-20.2010.

6. **Wang B, Aw TY, Stokes KY**. N-acetylcysteine attenuates systemic platelet activation and cerebral vessel thrombosis in diabetes. *Redox biology* 14: 218-28, 2018.

7. **Shigemizu D, Abe T, Morizono T et al**. The construction of risk prediction models using GWAS data and its application to a type 2 diabetes prospective cohort. PLoS One 9(3): e92549, 2014.

8. **Inzucchi SE, Bergenstal R, Buse J et al**. Management of hyperglycaemia in type 2 diabetes: a patient-centered approach. Position statement of the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). Diabetologia 55(6): 1577-96, 2012.

9. **Kerner W, Brückel J**. Definition, classification and diagnosis of diabetes mellitus. Experimental and Clinical Endocrinology & Diabetes 122(07): 384-6, 2014.

10. **Haffner SM, Alexander CM, Cook TJ et al**. Reduced coronary events in simvastatin-treated patients with coronary heart disease and diabetes or impaired fasting glucose levels: subgroup analyses in the Scandinavian Simvastatin Survival Study. *Archives of Internal Medicine*. 159(22): 2661-7, 1999.

11. **Jeon CY, Lokken RP, Hu FB, Van Dam RM.** Physical activity of moderate intensity and risk of type 2 diabetes: a systematic review. *Diabetes care* 30(3): 744-52, 2007.

12. **Sigal RJ, Kenny GP, Wasserman DH, Castaneda-Sceppa C, White RD**. Physical activity/exercise and type 2 diabetes: a consensus statement from the American Diabetes Association. *Diabetes care* 29(6): 1433-8, 2006.

13. **Lipovetsky S**. Two-parameter ridge regression and its convergence to the eventual pairwise model. *Mathematical and Computer Modelling* 44(3-4): 304-18, 2006.

14. **Spencer B, Alfandi O, Al-Obeidat F**. A Refinement of Lasso Regression Applied to Temperature Forecasting. *Procedia computer sci*ence 130: 728-35, 2018.

15. **Pusponegoro NH, Muslim A, Notodiputro KA, Sartono B**. Group LASSO for Rainfall Data Modeling in Indramayu District, West Java, Indonesia. *Procedia computer scienc*e 116:190-7, 2017.

16. **Neter J, Kutner MH, Nachtsheim CJ, Wasserman W**. Applied linear statistical models: Irwin Chicago; 1996.

17. **Iturbide E, Cerda J, Graff M**. A comparison between LARS and LASSO for initialising the time-series forecasting auto-regressive equations. *Procedia Technology* 7: 282-8, 2013.

18. **Kamkar I, Gupta SK, Phung D, Venkatesh S**. Stable feature selection for clinical prediction: Exploiting ICD tree structure using Tree-Lasso. *Journal of biomedical informatics* 53: 277-90, 2015.

19. **Saleh AME**. A ridge regression estimation approach to the measurement error model. *Journal of Multivariate Analysis* 123: 68-84, 2014.

20. **Budka M, Gabrys B**. Ridge regression ensemble for toxicity prediction. *Procedia Computer Science* 1(1): 193-201, 2010.

21. **Bennett C, Guo M, Dharmage S**. HbA1c as a screening tool for detection of type 2 diabetes: a systematic review. Diabetic medicine 24(4): 333-43, 2007.

22. **Currie CJ, Peters JR, Tynan A et al**. Survival as a function of HbA1c in people with type 2 diabetes: a

retrospective cohort study. *The Lancet* 375(9713): 481-9, 2010.

23. Wannamethee SG, Shaper AG, Whincup PH, Lennon L, Sattar N. Impact of diabetes on cardiovascular disease risk and all-cause mortality in older men: influence of age at onset, diabetes duration, and established and novel risk factors. *Archives of internal medicine* 171(5): 404-10, 2011.

24. Regenold WT, Thapar RK, Marano C, Gavirneni S, Kondapavuluru PV. Increased prevalence of type 2 diabetes mellitus among psychiatric inpatients with bipolar I affective and schizoaffective disorders independent of psychotropic drug use. *Journal of affective disorders* 70(1):19-26, 2002.

25. Yang W, Lu J, Weng J et al. Prevalence of diabetes among men and women in China. *New England Journal of Medicine* 362(12): 1090-101, 2010.

26. Willi C, Bodenmann P, Ghali WA, Faris PD, Cornuz J. Active smoking and the risk of type 2 diabetes: a systematic review and meta-analysis. *Jama* 298(22): 2654-64, 2007.

27. Sung K, Bae S. Effects of a regular walking exercise program on behavioral and biochemical aspects in elderly people with type II diabetes. *Nursing & health sciences* 14(4): 438-45, 2012.

28. World Health Organization. Use of glycated haemoglobin (HbA1c) in diagnosis of diabetes mellitus: abbreviated report of a WHO consultation. Geneva: World Health Organization, 201.1

29. Rohlfing CL, Little RR, Wiedmeyer H-M et al. Use of GHb (HbA1c) in screening for undiagnosed diabetes in the US population. *Diabetes care* 23(2): 187-91, 2000.

30. Batah FSM, Ramanathan TV, Gore SD. The efficiency of modified jackknife and ridge type regression estimators: a comparison. *Surveys in Mathematics & its Applications*. 2008; 3.

31. Abraham A, Krömer P, Snasel V. Afro-European Conference for Industrial Advancement: Springer; 2015.

32. Umpierre D, Ribeiro PA, Kramer CK et al. Physical activity advice only or structured exercise training and association with HbA1c levels in type 2 diabetes: a systematic review and meta-analysis. *Jama* 305(17): 1790-9, 2011.