Proceedings of the 2nd International Conference on Computing Innovation and Applied Physics (CONF-CIAP 2023)
DOI: 10.54254/2753-8818/5/20230427

Overview of high dimensional linear regression models

Jin Li[1]

[1]Boston University, Boston MA 02215, USA

andyli@bu.edu

Abstract. Linear regression analysis is one of the most basic and vital methods in statistical applications, especially when examining the relationship of one or more variables. This method initially used to analyze the relationship between father and child height has also been developed for nearly two hundred years. The development of computer technology has brought more efficient and accurate data processing, which allows us to use linear regression models to deal with high-dimensional and more complex scenarios. This paper mainly elaborates on the development of high-dimensional linear regression and the most classic models and tries to understand how some models can be improved. Finally, how to accurately apply model theory to practice will be a significant research direction in the future

Keywords: linear regression, high-dimension, statistical applications.

1. Introduction

Linear regression analysis and models have been a vital research branch of mathematical statistics for many years because this model has good practical value. Specifically, fields such as industry, agriculture, economics, and medical care can be analyzed using linear regression methods. With the advancement of computer science and technology, data acquisition and processing technology have been dramatically improved. For more complex data and application scenarios, traditional and classic linear regression models may no longer be applicable. We must propose more methods to improve linear regression models, including high-dimensional ones. A high-dimensional linear regression model is derived based on the traditional linear regression model suitable for high-dimensional space and data.

For the judgment of the pros and cons of the linear regression model, it is mainly to examine the fitting degree of the model to the actual situation. The main factors examined are the accuracy of model prediction, the authenticity of model prediction variable selection, the convergence calculation speed of the model, and the stability of the model. In this article, I expounded on the origin and development of high-dimensional linear regression models. I introduced some classic models, including subset selection regression, stepwise regression, ridge regression, lasso regression, least angle regression, linear models for high-dimensional data, and other regression methods. Based on these models, we have seen some optimized and improved models. The article concludes with a summary based on my learning and understanding of high-dimensional linear regression models.

2. Research background and significance of high-dimensional linear regression model

Linear regression analysis and nonlinear regression analysis represent the relationship between different variables. Linear regression analysis is a very widely used method in statistics, and it is a model used to

DOI: 10.54254/2753-8818/5/20230427

analyze and determine the relationship between multiple variables. Univariate analysis models and multi-dimensional regression models allow regression analysis to cover situations with different numbers of variables.

With the development of the economy and society, more and more people find that the data of some factors in the objects we need to show a relatively obvious linear relationship (or can become a linear relationship after linear transformation). To find out what kind of relationship there is between numbers, using regression analysis is necessary. For example, investigate the relationship between the sales of a particular product and the local population, the relationship between heart disease and age, height, weight, blood pressure, the relationship between altitude and the atmosphere, et cetera.

However, the data that can be included in the research scope of the investigator needs to be subjectively determined by the investigator, so before doing any regression analysis, the researcher needs to have a relatively correct prior knowledge of the matter. For example, to investigate the sales volume of a particular product in a specific place, in addition to the significant influence of the population and GDP of the place, it is likely to have a close relationship with the consumer price index (CPI) of the place.

There are two categories in terms of the purpose for which people apply regression analysis. One is that we need to know which factors (independent variables) lead to the outcome (response variable). For example, it is impossible for us to conduct decades-long human experiments on the effects of various heavy metal ions on brain dementia. We can only rely on regression analysis to examine the significant degree of dementia induced by various heavy metal ions [1]. The other is how, if we control for the independent variable, how does it affect the dependent variable. For example, the relationship between the indoor temperature in winter and the amount of coal burned in the furnace, we can control the indoor temperature according to the amount of coal supplied. Therefore, the research prospect of regression analysis is comprehensive. It can penetrate almost all fields, including agriculture, military, meteorology, hydrology, economy, society, et cetera. It has critical applications in helping people understand the relationship between various elements and values.

3. The history of the development of linear regression models

The regression research method was first applied to the study of the genetics of height. Specifically, the statistics found that tall fathers tended to have shorter sons than themselves, while short fathers tended to have taller sons. The overall height level shows a trend of returning to the middle, so this method is called regression analysis. The least squares method is the most basic method and concept in regression analysis, which was first proposed by the French mathematician Legendre [2]. However, distortion testing is also an essential step in determining whether the model fits correctly. Denial and Wood proposed an approximation to fit the distortion test. The idea is to use a clustering algorithm to find nearly duplicate cases and use the changes in the response variable to fit them. Combined distortion test [3]. Currently, there is no theoretical support for this type of test, and it is generally handled empirically. However, these deep-seated questions remain to be studied, and they are likely to be the focus of regression analysis research, which is widely used in various sociological studies [4].

In 1972, Nelder and Wedderburn first introduced the term generalized linear model in an article. However, for individual exceptional cases of the generalized linear model, Fisher had applied it as early as 1919, Berkson, Dyke in the 1940s and 1950s. Moreover, Patterson et al. have also applied the most famous logistic regression model [5]. In 1983, Mccullagh and Neluer published a monograph on a systematic discussion of generalized linear models, which was republished in 1989 [6].

4. General logic of high-dimensional regression models

Generally speaking, a complete regression analysis process needs to be composed of many specific steps. These steps vary from person to person and situation, mainly including the operator's technical level, the meticulousness and accuracy of data sampling or investigation, and the accuracy of the regression analysis. Requirements of accuracy, the rigor of the equation, subjective judgment, and understanding of the matter in the survey sampling are related to many aspects.

DOI: 10.54254/2753-8818/5/20230427

The first is to collect data. To study the quantitative relationship of certain things, the data that workers care about in the thing should be investigated first, which depends on the worker's subjective understanding of the nature of the thing and the objective conditions of the worker's sampling operation. Subjective cognition refers to which quantities represent the essential relationship of the characteristics of the thing and must be investigated; which data may have an impact on the results, but it is not clear whether it has a real effect and needs to be collected and adjusted according to the situation during the research process. The data does not affect the essence of the thing and can be directly eliminated. At the same time, the data is divided into two parts, one part of the training set is used to fit the model, and the other part of the test set is used to test the model.

After the data is fully processed, it is necessary to fit the regression equation to the training set. In general, the most widely used method is the least squares estimation. Biased estimates significantly reduce forecast variance with a small fraction of bias. Next, check whether the regression equation made is meaningful and accurate. For example, a t-test or p-value is used to determine whether the regression relationship is significant, and the coefficient can be determined to determine whether the part that the dependent variable can explain is large enough. At the same time, use the test data to test the regression results to see if they are consistent with the results expected from the training data.

5. Classical high-dimensional linear regression models

5.1. Stepwise regression and subset selection

Since stepwise regression and subset selection regression focus on the selection of model variables rather than on the estimation of variable coefficients, we start with stepwise regression and subset selection regression as the description of all regression methods [7].

Forward stepwise regression begins by selecting a single predictor that best estimates the model, e.g., minimizes the sum of squared residuals, et cetera. Next, another predictor variable is entered into the model so that the first two variables reach the best estimate of the model. The third variable is entered into the model in turn so that this variable and the first two variables are combined to form the optimal model. This process continues until a certain criterion is reached, such as the number of variables reaching a specific limit or the reduction of the residual sum of squares is less than a certain threshold. For the above diabetes data, the single best predictor variable was BMI, and the variables that were subsequently selected to enter the model were S5, BP, S1, Sex, S2, S4, S5, and S6 [8]. However, this process is extremely unstable; minimal changes in the data can cause variables entering the model to be replaced by other variables, so the order of variables entering the model later is entirely different. Another form of forwarding stepwise regression is backward stepwise regression, starting with all variables entering the model and continuously removing the variables that con-tribute the least to the model fit. Meanwhile, Efroymson combined forward stepwise regression with backward stepwise regression in 1960 [9].

The above two algorithms are greedy algorithms, which do not consider the subsequent fitting effect so that the change of each step can be optimized, which is in sharp contrast to the subset selection regression. Subset selection regression compares all possible subsets of predictors under the premise of limiting the maximum number of predictors for the optimal subset [10]. The advantage of subset selection over stepwise regression is that the optimal subset of two predictors does not necessarily contain a single optimal subset of predictors. The disadvantage is that statistical inference is more biased because the regression method considers more possible models.

5.2. Ridge Regression

Ridge regression includes all predictors, but the estimated coefficients of the predictors are smaller than the estimated coefficients of the least squares regression in the usual case. Ridge estimation coefficients minimize the penalty residual sum of squares:

$$\hat{\beta}^{ridge} = arg\max_{\beta}\left\{(||Y - X\beta||_2^2) + \theta \sum_{j=1}^{p} \beta_j^2\right\} \quad (1)$$

DOI: 10.54254/2753-8818/5/20230427

Here $\boldsymbol{\theta}$ is a positive scalar, and $\boldsymbol{\theta = 0}$ corresponds to the most common least squares regression. In the actual operation process, the variables must be standardized first, so that the change of the original data will not affect the penalty coefficient $\boldsymbol{\theta}$.
We can express the above formula in matrix form:

$$\hat{\beta}^{ridge} = arg\max_{\beta}\left\{(Y - X\beta)^T(Y - X\beta) + \theta\sum_{j=1}^{p}\beta_j^2\right\} \quad (2)$$

The solution of Ridge regression can be obtained:

$$\hat{\beta}^{ridge} = (X^TX + \lambda I)^{-1}X^TY \quad (3)$$

where $\boldsymbol{I}$ is the identity matrix. Note that the quadratic penalty $\boldsymbol{\beta^T\beta}$ is selected, and the solution of the ridge regression is a linear function of Y again. This solution adds a positive constant to the diagonal of the matrix $\boldsymbol{X^TX}$ before the inversion of the matrix $\boldsymbol{X^TX}$, which makes the problem non-singular even if the matrix $\boldsymbol{X^TX}$ is not of full rank. This is the main motivation for statisticians to elicit ridge regression in the first place.

5.3. Lasso regression

The set of regression is one of the most classic models in high-dimensional linear regression [11]. Tibshirani proposed to minimize the residual sum of squares under the constraint of the sum of the absolute values of the regression coefficients $\boldsymbol{\sum_{j=1}^{p}|p_j| \le t}$, as follows:

$$\hat{\beta}^{Lasso} = arg\min_{\beta}\{(||Y - X\beta||_2^2)\} \quad (4)$$

$$s.t.\sum_{j=1}^{p}|p_j| \le t \quad (5)$$

This is equivalent to adding an $\boldsymbol{L_1}$ penalty to the regression coefficients based on minimizing the residual sum of squares, which we can express as follows:

$$\hat{\beta}^{Lasso} = arg\min_{\beta}\left\{(||Y - X\beta||_2^2) + \theta\sum_{j=1}^{p}|\beta_j|\right\} \quad (6)$$

Similar to the ridge regression problem, the L2 penalty is replaced by the locking set $\boldsymbol{\sum_{j=1}^{p}\beta_j^2}$, the latter constraint makes the solution nonlinear in the response variable, and the set regression is solved with a quadratic programming algorithm. Due to the nature of this constraint, making the constraint t sufficiently small results in some regression coefficients being exactly 0, which makes the lasso regression method do some kind of continuous subset selection. If the selected t is greater than $\boldsymbol{\sum_{j=1}^{p}|\widehat{\beta_j}|^{LS}}$, then the result of the set of regression estimation is the least square estimation.

5.4. Forward Stagewise Regression

The forward piecewise regression method seems to be very different from the lasso regression method, but the results of the two regression methods have very similar properties. This regression method is based on stepwise regression to eliminate the negative effects of greedy behaviour at each step of stepwise regression. In stepwise regression, the most valuable predictors are gradually selected into the model at each step. The regression coefficients are gradually changed from zero to the regression coefficients of the least squares regression.

The forward stepwise regression is the same as the forward stepwise regression, selecting the same predictor variable in the first step. However, the change in the correlation coefficient at each step is minimal. The model continuously selects the variable with the highest correlation with the current residual (it may also be the variable that has appeared in the previous steps), and then advances a small step in the direction of that variable, and then goes on, which is the forward piecewise regression method. If a variable is more dominant than other variables, the model will advance several steps in the direction of this variable. Once several variables are entered into the model, the model will alternately advance in

DOI: 10.54254/2753-8818/5/20230427

the direction of these variables, forwarding the regression of piecewise regression. The coefficient results are much more stable than the stepwise regression coefficients.

Although forward piecewise regression is very different in form from lasso regression, an idealized piecewise regression (where the forward step size tends to zero) has very similar properties to lasso regression. In the diabetes data, before the eighth variable entered the model, the forward piecewise regression and the lasso regression had consistent results. Furthermore, overall, there is little difference in the results of the two regression methods for estimating regression coefficients.

The forward piecewise regression method is closely related to the popular boosting algorithm in machine learning. However, the difference between the two methods is not reflected in the model's fitting; in fact, the predictor variables of forwarding piecewise regression are determined in advance. In contrast, the following variable of the boosting algorithm is dynamically determined.

6. Comparison of important high-dimensional linear regression models

In the usual high-dimensional case, the minimum Angle regression, lasso regression, and piecewise regression show similar results, but their regression results are not the same. Minimum Angle regression has the advantage of computational speed because it does not move variables out of the model once they enter it. Therefore, after the least-angle regression step, all variables will be added to the model, and the complete least-squares solution will finally be obtained. However, for lasso and piecewise regression, variables can be removed from the model or re-entered, so both methods may require more than p steps to achieve the entire model [12]. The above model is for a sample size more significant than the number of variables.

The Lasso regression method is straightforward and based on the least squares regression method with an additional L1 penalty. This approach is also common in other areas of machine learning. Boosting algorithms or slow learning methods in machine learning include piecewise regression algorithms. The minimum angle regression method is implemented based on Newton's method, and the minimum angle regression method can be extended to nonlinear regression models such as generalized linear models [11].

7. Conclusions

This paper introduces the research significance and research background of the subject. It introduces some classical methods of generalized high-dimensional linear models, such as the stepwise regression method, subset selection method, ridge regression, lasso regression, least angle regression, and other methods.

Based on the model, in the practical process of high-dimensional linear regression analysis, two purposes are generally achieved: the accuracy of model prediction and the authenticity of variable selection. If the focus is on finding an interpretable model or trying to find a true model as much as possible, then the accuracy of the model's predictions is secondary. As an example of network modelling in biology, if the model's prediction accuracy is the first consideration, then as long as the coefficient of each predictor is small enough, the prediction variance is significantly reduced, and the model includes some other predictors. is also acceptable.

The least-angle regression method has great potential in model computation speed, interpretability, estimation stability, and graphical representation of regression coefficient paths. However, we still have much work to tap this huge potential in practical applications. In practical problems, many factors have to be considered, such as the order of predictors. And we can learn that there are many improvement methods for classical regression methods, such as SCAD, adaptive lasso regression, group lasso regression, et cetera. The basic ideas of these methods can be extended to high-dimensional linear regression models and Investigate the data structures that fit each model. Therefore, there is still a vast research space for related methods of high-dimensional linear regression. At the same time, these methods should not only stay in the academic discussion stage, but more importantly, they should be applied in production practice to discover the advantages and disadvantages of different methods, et

DOI: 10.54254/2753-8818/5/20230427

cetera, so that the models can be better based on different needs. With different degrees of improvement, these issues are worthy of our further study.

References

[1] Dette, H., & Munk, A. (1998). Validation of linear regression models. Annals of Statistics, 778-800.

[2] Bohmanova, J., Miglior, F., Jamrozik, J., Misztal, I., & Sullivan, P. G. (2008). Comparison of random regression models with Legendre polynomials and linear splines for production traits and somatic cell score of Canadian Holstein cows. Journal of Dairy Science, 91(9), 3627-3638.

[3] Atkinson, A. C. (1982). Regression diagnostics, transformations, and constructed variables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *44*(1), 1-22.

[4] Atkinson, A. C. (1983). Diagnostic regression analysis and shifted power transformations. *Technometrics*, *25*(1), 23-33.

[5] Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. Journal of the Royal Statistical Society: Series A (General), 135(3), 370-384.

[6] Pregibon, D. (1984). Generalized linear models.

[7] Miller, A. J. (1984). Selection of subsets of regression variables. Journal of the Royal Statistical Society: Series A (General), 147(3), 389-410.

[8] Hastie, T., Rosset, S., Tibshirani, R., & Zhu, J. (2004). The entire regularization path for the support vector machine. Journal of Machine Learning Research, 5(Oct), 1391-1415.

[9] Freedman, D. A., & Freedman, D. A. (1983). A note on screening regression equations. the American statistician, 37(2), 152-155.

[10] Furnival, G. M., & Wilson, R. W. (2000). Regressions by leaps and bounds. Technometrics, 42(1), 69-79.

[11] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

[12] Yuan, Z., & Yang, Y. (2005). Combining linear regression models: When and how?. Journal of the American Statistical Association, 100(472), 1202-1214.