Application and Challenges of Statistical Methods in Biological Genetics

Jingyi Sun *

Department of Mathematics, University of North Carolina-Chapel Hill, Chapel Hill, North Carolina, United States

* Corresponding Author Email: jingyis@ad.unc.edu

Abstract. Humans are curious about genes, from plants to animals, from breeding to diseases. For centuries, it has been considered a genetic disease. With the development of medicine, people have also realized that many diseases are heritable. With the birth of modern statistics, humans have created many models. This article focuses on the application of statistical methods in biological genetics. This paper introduces the principles and their applications of Least Absolute Shrinkage and Selection Operator Regression, the Chen-Stein Method, and Logical Regression model in different branches, such as gene set selection. These models can effectively tackle the problem of reproducibility in genetics to a certain extent when used correctly. In addition, they offer an effective means of data analysis in genetics field. Although the three models have their weaknesses, such as the use and selection of a priori, it is reasonable to believe that with the continuous improvement of the models by mathematicians, they can have better prospects.

Keywords: Genome testing, Lasso Regression, Chen-Stein Method, Logistic Regression.

1. Introduction

Genetics - is the scientific study of the laws governing biological inheritance and variation. The scientific study of genes' structure, function, transmission, and variation. The study of heredity and variation in organisms, as well as the structure, function, transmission, and expression of genes. In genetics, the parent-child does not just refer to parents and children or a family. However, it can also refer to the population, which encompasses multiple families and is the focus of population genetics research. In genetics, cells can also contribute to the formation of the notion of parenthood. Some genetic characteristics of individuals, such as the presence or absence of enzymes, can be maintained in cells grown in vitro. Somatic cytogenetics involves the genetic study of in vitro-cultivated cells. In genetics, the idea of parent-child may be extended to deoxyribonucleic acid (DNA) replication and even messenger RNA (mRNA) transcription, both of which are molecular genetics study subjects. The core of genetics is the study of genes, which reside in the nucleus, chromosomes, and DNA of living organisms. DNA and genes occur in pairs, just as chromosomes exist in pairs. The average human body consists of 23 chromosomal pairs and 46 DNA molecules. Genetics research encompasses the essence of genetic material, its transfer, and the expression of genetic information. The transmission of genetic material involves the replication of genetic material, the activity of chromosomes, genetic laws, and variations in the number of genes within a population. Modern genetics tries to identify the underlying mechanism of the entire genetic process.

In the field of statistics, lasso is a technique for conducting regression analysis that combines the selection of variables with the application of regularization to improve the precision of the statistical model's predictions. The Chen-Stein technique is a helpful tool that may be used for obtaining an error bound when calculating probabilities. In many cases, it is sufficient to just use the first and second seconds to indicate this constraint. Examples of typical applications include random graphs. The Logistic Model is a statistical model that is used to express the likelihood of an occurrence by linearly mixing one or more independent variables with the log-odds for the occurrence. This model is used to indicate the probability of an event. In the field of regression analysis, the technique known as logistic regression is used to compute an estimate of the logistic model's parameters (the coefficients in the linear combination).

This paper aims to introduce the application of Lasso, Chen-Stein Method, and Logistic regression in genetics filed. The paper will begin with the overview of the three methods: their background, mathematical traits, and principles. Then it will explain the methods' application in genetic field and other inter-discipline area.

2. Overview of Methods

2.1. Lasso

Robert Tibshirani proposed The Least Absolute Shrinkage and Selection Operator (Lasso) for the first time in 1996 [1]. Completeness of a title achieved by the selection of minimal absolute contraction and operator. This technique is a streamlined approach to the process of calculating. By producing a penalty function, it is possible to generate a model that is more refined. Some regression coefficients are compressed using this function; more specifically, a penalty function is generated using the regression coefficients that are compressed using this function. While this is going on, some regression coefficients are being turned to zero. As a result, it continues to maintain the benefit of subset shrinking while being an inadequate estimate of complicated collinear data. The issue of stretch is analogous to that of least squares [2].

The tradeoff between approximation error and sparsity is specified by the regularization parameter $\lambda = [\lambda_{min}, \lambda_{max}]$, all of which are real integers. λ is often calculated using a computerized k-fold cross-validation procedure. For this technique, the dataset is partitioned into k evenly sized subsamples at random. The first k1 subsamples are used to create a prediction model, while the remaining subsamples are use verifying the accuracy of the model. This process is repeated k times, with each iteration serving as a separate subsample for either model development or validation. Combining the outcomes of k separate validations across a spectrum of λ values and picking the preferred lambda produces an overall result, which is then used to decide the final model. With this method, overfitting can be mitigated without having to dedicate a portion of the dataset solely to internal validation [3].

2.2. Chen-Stein Method

Charles Stein invented a fresh approach for his statistics lecture near the end of the 1960s [4]. In 1970, he gave a crucial paper at the sixth Berkeley Symposium. Later, Louis Chen Hsiao Yun, his Ph.D. student, updated the approach to produce Poisson distribution approximation findings [5]. As a result, the Stein technique applied to the issue of Poisson approximation, known as the Stein-Chen approach [6]. Instead of using Fourier methods, Stein's technique depended on the elementary differential equation:

$$f'(x) - xf(x) = h(x) - Nh \tag{1}$$

In the preceding equation, h is used to test convergence in distributions, and Nh = E[h(Z)], where Z is the standard normal distribution. Chen's study has resulted in advancements in the theory of Poison approximation and contributed to the development and improvement of several fascinating applications and instances [7].

2.3. Logistic Regression

Logistic regression analysis, often known as LRA, is an extension of the concept of multiple regression analysis that is used to investigate variables that may be classed as finite. The findings have been made public and are very apparent. When analyzing an educational strategy, for instance, it is possible to forecast two outcomes, such as success or failure, or improvement or improvement. There is a possibility that the existence or absence of the sickness will have analogous effects on the workings of the medical facility. Although LRA technology may be utilized to provide three or more

different sorts of outcomes (such as more accurate, identical, or not too terrible), the primary emphasis of this work is on variable results that produce binary results. The logic function that was developed in the 19th century serves as an indication of both the self-development and advancement of chemical processes, as well as the expansion of the population [8]. Understanding the traditional binary regression model is necessary to comprehend ordinal logistic regression. Let's start off simply by thinking about the scenario where the impact of one explanatory variable (co-variate) X on the response variable Y is examined. The simplest link between Fand X is a straight line provided by the simple linear regression model if the measurement levels of X and Y are continuous, for instance if X = height and Y = forced expiratory volume in 1 second (FEVI).

$$Y = \alpha + \beta \tag{2}$$

This approach presupposes that Y and X are, at the very least, roughly, linearly connected. If this presumption is incorrect, additional, more complex models, such as nonlinear ones, should be considered since the simple linear regression is not relevant.

The simple linear regression model is flawed if Y is not continuous but rather binary (i.e., only 1/0 type responses like "success/fail" or " yes/no" are permitted), as it assumes that Y can take any numerical value between minus infinity and plus infinity. Furthermore, the common assumption of homogenous variance is broken if Y is a binary variable. The secret to accurately characterizing the link between Y and X is to represent the likelihood of an occurrence rather than Y itself, i.e., p=P(Y=1) instead Y itself. Unlike p, which may accept any number between 0 and 1, F only has two possible values: 1 and 0. The odds p/(1-p) can be any positive number, while the odds ln [p/ (1-p)], also known as the logit, have a logarithmic range from minus infinity to plus infinity. Consequently, it is reasonable to infer that the logit and X have a linear relationship [9]:

$$logit = ln(\frac{p}{1-p}) = \alpha + \beta X \tag{3}$$

It is mathematically equivalent to the expression.

$$p = \frac{exp(\alpha+\beta X)}{1+exp(\alpha+\beta X)} \tag{4}$$

3. Application Field

3.1. Lasso Penalized Regression (SLPR) in Gene Set Selection

Genome testing is a crucial method in the biological field that allows researchers to go beyond single genome variables to study the biological importance of genome association. When compared to study that focuses on single genome variables, this strategy may considerably enhance statistical ability, biological interpretation, and replication by concentrating on a limited number of functional genomes [9, 10]. The analysis of "big data" in biology absolutely requires the use of this technology. These data are gathered via the use of innovative and efficient techniques of analysis on a big scale. Under a variety of experimental settings, these technologies can measure the distribution, change, and alteration of hundreds of thousands or hundreds of thousands of biological molecular elements or decigrams. One example of "big data" is the transcription of data sets that measure hundreds of distinct mRK molecules [11]. data sets including genome-wide associations that evaluate the genotypic variance of over one million genetic markers [12].

The SLPR system is based on a biological basis. The SLPR model expressly implies that gene-level summary statistics may be described by linear functions relating to gene sets including the gene, while set-level statistics measure the overall process or pathway activity level. A crucial aspect of this approach is the notion that the activities of different gene sets have extra impacts on the statistical data at the gene level. In this situation, the activity of two genomes with independent roles, identical

amplitude, and direction, and both containing the same gene is considered to provide gene-level statistics that are about twice those produced when only one genome is active. The SLPR approach may support additional, more complicated gene activity models in addition to covariate adjustment, gene set or gene weight, and single-sample gene set testing [13].

In the initial phase, the R glmnet package is utilized to solve the Lasso penalty version of the regression model [14] To minimize cross-validation errors, active sets are chosen from gene sets. There is no penalty regression after the Lasso penalty, and only the prediction variables are utilized in Lasso fitting. These two steps are called Gauss Lasso [15, 16] method retains Lasso's advantage in model selection and generates coefficient estimates that do not shrink. Although the unpunished model gives p values for the gene set predictor, because to the preceding Lasso-based coefficient selection, these second-stage P values cannot be utilized for inference.

3.2. Chen-Stein Method in Somatic cell hybrids

To determine which chromosomes, contain which human genes, somatic cell hybrids have been widely used. Partially human chromosome- and gene-containing mouse-human hybrid cell lines [17]. Thus, normal human cells are combined with genetically engineered rat cells to produce somatic cell hybrids. All rat chromosomes survive in the ensuing hybrid cells, but sporadic human chromosomes are lost. Clones with stable subsets of human chromosomes may be found many generations following cell fusion. In most cases, chromosomes in both humans and rodents maintain their normal functionality.

Important presumptions must be met for the gene allocation method to work. All the genes in a human being should be located on either a single chromosome or on two similar chromosomes. Second, any rodent genes that are related to the human gene within the clone must be easily recognizable. Protein electrophoresis or direct annealing of a suitable DNA probe to a gene region are common techniques for identifying genes. Next, every one of the 24 unique human chromosomes must be detectable throughout the cloning process, either by cytology or biochemistry. Cytologically, chromosomes may be distinguished from one another based on their size, centromere location, and distinctive banding patterns when the appropriate dye is used. Chromosome identification may also be accomplished using in situ hybridization or isozyme detection with the use of large fluorescent DNA probes. It is possible that the different proteins may be determined by analyzing isozymes [18].

Lange analyzed the information of somatic cell hybrids using the Chen-Stein method [19]. n represents the total number of mixed clones inside the panel. As the y chromosome contains few genes of relevance, hybridization typically begins in female human cells. This produces in 23 distinct chromosomes, including 22 chromosomes and x, which displays the n=9 mixed clone panel. in Figure. 1, Each column inside the panel indicates a unique clone. Every one of 23 columns represents a chromosome. Clone i has a j chromosome, as shown by the first line of the panel and the first line of the list j. 0 indicates the absence of chromosomes in the clone. When evaluating each clone for the presence of human genes, columns 0 and 1 are added. If there are no analytical mistakes or fundamental assumptions, just one of the panels will match the test column. In this instance, the gene is associated with the proper chromosome.

When constructing random panels of somatic cell hybrids, three valid assumptions may be made. During the first stages of the development of a stable clone, each genetic chromosome is eliminated or preserved independently. Second, P is the shared likelihood of chromosomal pairing retention. With probability p, each pair of homologous pairs keeps at least one member.

0 1 0 1 0 0 0 1 0 0 0 0 0 0 1 0 1 1 0 1 1 1 1
1 0 1 0 1 1 0 0 1 0 0 0 0 1 0 0 1 0 1 0 1 1 1
0 1 1 1 1 0 1 0 0 0 0 0 1 0 0 1 1 0 1 1 0 1 1
1 1 1 0 0 1 1 0 0 1 0 1 0 0 0 1 1 1 0 0 1 0 1
0 0 0 1 1 1 1 0 0 0 1 1 1 1 1 0 1 0 0 0 1 1 0
0 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 1 0 0 0 0 0
0 0 1 0 1 0 1 1 0 1 1 1 0 0 0 0 1 1 1 1 1 0 0
0 0 0 1 0 1 1 1 0 0 0 1 0 1 1 1 1 0 1 0 1 0 1
1 0 0 0 1 1 0 0 0 1 0 1 1 0 1 0 1 0 1 1 0 0 1

Figure 1. A somatic cell hybrid panel

When producing somatic cell hybrid panels at random, three acceptable assumptions may be made. Initially, each human chromosome is destroyed or retained individually during the creation of a stable clone. Second, P is the retention probability shared by all chromosomal pairings. Each pair of homologous chromosomes preserves at least one member with probability P. Rushton [20] expects p to fall between 0.07 to 0.75. Our theory is greatly simplified by the parameter p =0.5. Third, the retention patterns of distinct clones are independent of one another.

Now, identify column s of a randomly selected panel of n clones with C_s^n. Let $X_{\{s,t\}}^n$ to be the indicator of the event $\rho(C_s^n, C_t^n) < d$, d is the fixed Hamming distance between any two different columns: C_s^n and C_t^n. The variable $Y_d^n = \sum_{\{s,t\}} X_{\{s,t\}}^n$ is 0 when the lowest Hamming distance is equal to or greater than d. In the index set I, there are $\binom{23}{2}$pair $\alpha = \{s, t\}$, and each related X_α^n has the same mean

$$p_\alpha = \sum_{i=0}^{d-1} \binom{n}{i} q^i (1-q)^{n-i} \tag{5}$$

Where, $q = 2q(1-q)$ is the probability that C_s^n and C_t^n differ in any entry.

This method generally combines with Poisson approval to solve problems and has an upper limit value. Therefore, more improvements may be needed in the future to improve the efficiency and upper limit of data processing. Furthermore, all the literature records are not very new, or the literature related to genes is limited, so there may be a better way to exist.

3.3. Logistic Regression

In the field of genetic classification, several researchers have modified the commonly used method of logistic regression, which makes use of data from microarrays, to develop prediction models for binary outcomes. However, a step for feature (gene) selection must be introduced into the logistic modeling process to account for this limitation [21]. T Logistic regression's derivative is a valuable tool. Based on the basic logistic regression model, statisticians may create a variety of logistic regression and other models. These models make it simpler for genetics experts to examine data, resulting in more accurate results. Authors J.G. Liao and Khew-Von Chin propose a parametric bootstrap model based on logistic regression that is adapted to the microarray data to assess the prediction error more accurately. This approach was developed based on extensive research into the process of discovering genes with altered expression, especially how frequent the local false discovery can be. The proposed method considers two of the most important aspects of model selection: the overall number of genes that should be included into the model, as well as the correct level of downsizing for penalized logistic regression. Khew-Voon Chin demonstrates that picking more than 20 genes does not always result in a further reduction of prediction error. The results, which produced accurate prediction models, made use of both the data on cervical cancer and the data on leukemia that Golub had collected. Multiple logistic regression is widely used and important in determining the association between genes and illnesses. Researchers investigated the BPD outcomes of 108 pairs of twins at risk for the condition. The disadvantage of logistic regression is obvious at first since it was limited to just two categorical variables, which does not seem to be a viable approach

in genetics investigations. However, as more people construct more complicated models on it and its limitation is lifted, it is being used in an increasing number of genetic investigations. As previously stated, multiple logistic regression is an integral derivative based on logistic regression. In addition, new models like as sparse logistic regression, ridge regression, and others were introduced to the study of genetics.

The researchers were looking for a strategy to appropriately define breast cancer subtypes, which is critical since it directly influences therapy choices. Traditional statistical approaches may fail to uncover outliers in Segaert's research owing to their substantial effect, necessitating the use of robust statistics. Therefore, the researcher used robust sparse logistic regression, and the findings revealed 36 important genes, more than 60% of which had previously been identified as having biological significance to triple-negative breast cancer [22].

4. Discussion

In the subject of Genetics, the statistical methods of Lasso Regression, Chen-Stein Method, and Logistic Regression are commonly employed, but they have drawbacks. Several times, it has been proved that Lasso Regression outperforms conventional methods. It does not, however, obviate the need to validate a model against an external dataset and is not a solution for overfitting and optimism bias. In addition, the Lasso approach compensates for any bias in predicting individual parameters by generating a more accurate total forecast. The Chen-Stein Method, which has a maximum value, is frequently used in conjunction with Poisson acceptance to solve issues. Traditional statistical procedures may fail to detect outliers when Logistic Regression is applied for genetic data due to their significant effect, requiring the necessity for robust statistics. Furthermore, even if the current concepts and tools have been created and propagated throughout time, this technique still has a distinct barrier for certain researchers.

However, genetic research cannot exist without these statistical methods. By combining a priori and a posteriori characteristic, the Lasso Regression, Chen-Stein Method, and Logistic Regression methods can help tackle the problem of reproducibility in genetics to a certain extent when used correctly. In addition, these statistical tools offer an effective means of data analysis. In summary, Lasso Regression, Chen-Stein Method, and Logistic Regression give the genetical study a new notion and perspective, which is favorable to the development of Genetics. In addition, there is a gradual push for creating and disseminating these three methodologies and accompanying technologies. It is reasonable to assume that their application potential in the field of Genetics will be promising.

5. Conclusion

This paper introduces some applications of Lasso Regression, Chen-Stein Method, and Logistic Regression method in the field of genetics. Firstly, this paper briefly introduces the primary connotation of Lasso Regression, Chen-Stein Method, and Logistic Regression. Then, three different uses of the methods of Lasso Regression, Chen-Stein Method, and Logistic Regression method in genetic and related fields are introduced. These advanced statistic technologies provide the genetical study a new notion and perspective, which is favorable to the development of Genetics. Although Lasso Regression, Chen-Stein Method, and Logistic Regression still have some problems in the use and selection of a priori, by further exploring to improve the accuracy and efficiency of these methods, the field of genetics will have expected development in the future.

References

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

[2] S. L. Kukreja, J. Löfberg, M. J. Brenner, "A least absolute shrinkage and selection operator (LASSO) for nonlinear system identification," *IFAC proceedings volumes*, 2006, 39 (1): 814 - 819.

[3] J. Ranstam, J. A. Cook, "LASSO regression," *British Journal of Surgery*, Volume 105, Issue 10, September 2018, Page 1348.

[4] C. Stein, "The Invariant, the Direct and the" Pretentious"," Creative Minds, Charmed Lives: Interviews at Institute for Mathematical Sciences, National University of Singapore. 2010. 282 - 287.

[5] L. H. Y. Chen, "Poisson approximation for dependent trials," *The Annals of Probability*, 1975, 3 (3): 534 - 545.

[6] K. Lange, "Mathematical and statistical methods for genetic analysis," New York: Springer, 2002.

[7] R. Arratia, L. Goldstein, L. Gordon, "Poisson approximation and the Chen-Stein method," Statistical Science, 1990: 403 - 424.

[8] Wilson, R. Jeffrey, and Kent A. Lorenz. "Short history of the logistic regression model." *Modeling Binary Correlated Responses using SAS*, SPSS and R. Springer, Cham, 2015. 17 - 23.

[9] Bender, Ralf, and Ulrich Grouven. "Ordinal logistic regression in medical research." *Journal of the Royal College of physicians of London* 31.5 (1997): 546.

[10] Allison, B. David, et al. "Microarray data analysis: from disarray to consolidation and consensus." *Nature reviews genetics* 7.1 (2006): 55 - 65.

[11] Khatri, Purvesh, Marina Sirota, and Atul J. Butte. "Ten years of pathway analysis: current approaches and outstanding challenges." *PLoS computational biology* 8.2 (2012): e1002375.

[12] Barrett, Tanya, et al. "NCBI GEO: archive for functional genomics data sets—update." Nucleic acids research 41.D1 (2012): D991 - D995.

[13] Visscher, Peter M., et al. "Five years of GWAS discovery." The American Journal of Human Genetics 90.1 (2012): 7 - 24.

[14] J. H. Friedman, T. Hastie, R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," J. Stat. Softw.2010; 33: 1 – 22.

[15] R. Tibshirani, "Regression shrinkage and selection via the Lasso," J. R. Stat. Soc. B (Stat. Methodol.). 1996; 58: 267 – 288.

[16] A. Javanmard, A. Montanari, "Model selection for high-dimensional regression under the generalized irrepresentability condition," *Proceedings of the 26th International Conference on Neural Information Processing Systems*. 2013; Curran Associates Inc. 3012 – 3020.

[17] D'Eustachio, Peter, and Frank H. Ruddle. "Somatic cell genetics and gene families." *Science* 220.4600 (1983): 919 - 924.

[18] K. Lange, "Mathematical and statistical methods for genetic analysis," New York: Springer, 2002.

[19] T. M. Goradia, K. Lange, "Applications of coding theory to the design of somatic cell hybrid panels," Mathematical biosciences, 1988, 91 (2): 201 - 219.

[20] A. R. Rushton, "Quantitative analysis of human chromosome segregation in man-mouse somatic cell hybrids." *Cytogenetic and Genome Research* 17.5 (1976): 243 - 253.

[21] J. G. Liao, Khew-Voon Chin, "Logistic regression for disease classification using microarray data: model selection in a large p and small n case," *Bioinformatics*, Volume 23, Issue 15, August 2007, Pages 1945 - 1951.

[22] Segaert, Pieter, et al. "Robust identification of target genes and outliers in triple-negative breast cancer data." *Statistical methods in medical research* 28.10 - 11 (2019): 3042 - 3056.