Variable and model selection method in linear regression for analysis

Borui Zhang

University of California, Davis, CA, 95618

erozhang@ucdavis.edu

Abstract. When making decisions, probabilistic reasoning is used to utilize the known information to predict or determine those unobserved factors and variables that are crucial for the outcome. To find the relationship between outcome and one or more predictors, linear regression is commonly used in statistics and machine learning algorithms. In this article, the basic concept of linear regression and hypothesis testing are reviewed. The common modern methods for variable and model selection including Stepwise selection, Akaike's information criterion, Bayesian information criterion, and Mallow's C_p are discussed and reviewed. Each method and criterion have its own uniqueness and limitation depending on the dataset and the purpose of analysis. Through discussing this, this paper aims to inform and explain each different method for variable and model selection in linear regression and provide information to help choose the most suitable methods to predict or find the relationship between response variables and the independent variables for analysis.

Keywords: Linear regression, model selection, variable selection, machine learning.

1. Introduction

Probabilistic reasoning is used to utilize known information to predict or determine those unobserved factors and variables that are crucial for decision-making. It includes five crucial elements: general knowledge and relevant factors, supply knowledge in each situation or probabilistic model, the questions or query, the inference algorithm to predict the outcome, and the answer as probabilities. The model utilizes all the relevant general knowledge about a situation and uses the evidence and queries to answer questions and get an outcome. A probabilistic reasoning system is commonly used to predict future events, infer the cause of events, and learn from past events to better predict future events based on knowledge and logic. Most things happen in real life are dominated by random variables. These random variables are things that humans hard to control and predict. Probabilistic programming has been gradually discovered and used in order to better help people to control the events that did not happen . Logic programing can help people use what they know to make more accurate predictions for future. For example, to predict the future sales data of an upcoming product, we can use the data of another similar product in the previous year to analyze practical probroms filled with uncertainty. The Probabilistic programming could provide a way to use the data and logic given to answer the specified questions. The idea of probabilistic programming is to introduce theories and inference algorithms from statistics, and make the models and applications used by machine learning supported by effective inference evaluators, such as compilers and formal semantics in programming languages. An analysis and prediction method for quantitative or continuous data is called linear regression. It is a simple approach to supervised learning for modeling the relationship between one response value Y and one or more predictors X and can show the relationship between or among each variable in a dataset including intensity, quantitative function, accuracy, linearity, and synergy. Linear regression was developed by Sir Francis Galton in the late 19th century and can support the probabilistic reasoning system by distinguishing the relationship between response variables and explanatory factors to assist in assessing the impact of confusion for observation purposes or making a crucial decision [1-23]. Linear regression is used to establish linear relationships between response and explanatory variables from analysis and learning to knowledge and results. However, the original/full model is not suitable for all situations in real-life data analysis and prediction, thus transformation and model selection are needed to find the best fit model for factors analysis and prediction. In this article, a summary of linear regression and methodology is performed. The goal of this paper is to discuss and review the different methods used to find the most proper model for linear regression analysis and further study.

2. Notation and Basic concept

2.1. Regression analysis

Regression analysis is a statistical methodology used to predict a response variable from the other via utilizing the relation among quantitative variables. Regression models can be applied to experimental data from a fully randomized design as well as observational data. Regression models must meet certain requirements to apply to the data at hand, whether they are experimental or observational.

2.2. Simple linear regression

Simple linear regression is a type of linear regression when there is just one independent variable or predictor. $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. It is used to determine how independent variable X affects and relates to response variable Y. There is only one independent variable, X, making it simple. Since the parameters are not an exponent or are not multiplied by or divided by another parameter, it is linear in the parameters. Because this variable only appears in the first power, it is also linear in the predictor. If the independent variable and model are linear, the variable can also be referred to as a first-order model.

2.3. Multiple linear regression

A linear regression model containing two or more independent variables or predictors is called a multiple linear regression model. If there are p – 1 independent variable, X1, … Xp-1. The regression model should be:$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{I,p-1} + \varepsilon$. It is the most used model in real-life analysis to predict the relation and influence of several explanatory variables X on responsible Y. This can be also present in general linear regression model matrix terms.

$$Y_{n\times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \dots \\ Y_n \end{bmatrix} \tag{1}$$

$$X_{n\times p} = \begin{bmatrix} 1 & X_{11} & X_{i1} & \cdots & X_{i,p-1} \\ 1 & X_{21} & X_{i1} & \cdots & X_{i,p-1} \\ \vdots & \vdots & & \vdots & \\ 1 & X_{n1} & X_{n1} & \dots & X_{n,p-1} \end{bmatrix} \tag{2}$$

$$\beta_{n\times 1} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \dots \\ \beta_n \end{bmatrix} \tag{3}$$

$$\varepsilon_{n\times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \dots \\ \varepsilon_n \end{bmatrix} \tag{4}$$

$$Y_{n\times 1} = X_{n\times p}\, \beta_{n\times 1} + \varepsilon_{n\times 1} \tag{5}$$

3. Method and Selection

3.1. Method of least-squares

The method of least-squares is used to find regression parameters estimators for estimating $\boldsymbol{\beta}.$ To estimate $\boldsymbol{\beta}$, Consider the objective function $G(b) = (Y - Xb)^T(Y - Xb)b; b \in R^p.$

A vector b = $\hat{\beta}$ that minimizes G(b) is called a least squares estimator of β. The minimum of G(b) is attained at $\hat{\beta}$ if and only if $X^T X\hat{\beta} = X^T Y$. If X has full column rank p, $X^T X$ is nonsingular, then the least squares estimator of β is unique and is given by $b = \hat{\beta} = (X^T X)^{-1} X^T Y.$

3.2. F-test

In linear regression, F-test test for significance of regression by using the analysis of variance. It can evaluate whether the independent variable and the response variable are linearly related.

$$F = \frac{\sum_{i=1}^{n} \frac{(\hat{y}_i - \bar{y}_i)^2}{p-1}}{\sum_{i=1}^{n} \frac{(y_i - \bar{y}_i)^2}{n-p}} \tag{6}$$

The p is the number of parameters. The function is also known as $F = \frac{MSR}{MSE}$.

3.3. t-test

t-test in linear regression can check the significance of individual regression coefficients.

$$t_j = \frac{\hat{\beta}_j}{\sqrt{C_{jj}}\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}{(n-p)}}} \tag{7}$$

C_{jj} is j<u>th</u> component on matrix $(\acute{x}x)^{-1}$

3.4. Variable selection

3.4.1. Stepwise Feature Selection

Stepwise methods begin with the null model or with certain variables to enhance the model's performance, by selecting or eliminating a single variable at each step. The two most commonly used method is the forward selection and backward selection/elimination.

3.4.2. Forward Selection

Forward Selection begins with a null model, which contains only an intercept without any independent variable or predictor. The variable with the lowest residual sum of squares, or RSS, is added to the null model., and p simple linear regressions in the model are fitted in the model. Then, among all two-variable models, the variable with the lowest residual sum of squares is added to the model. The process will stop until some conditions are met, like when the p-value for every variable in the model under forward selection is higher than the threshold

3.4.3. Backward Selection

Backward selection starts with a full model with every variable, the process is used to eliminate the variable with the largest p-value, and this indicates that the least statistically significant variable is eliminated in each step during backward selection. The variable with the largest p-value is eliminated, and a new (p-1) variable model will be fit. The process will stop until some conditions are met. For example, the process can be stopped when every last variable meets some criteria for significance and has a significant p-value.

3.5. Information criteria

Information criteria are based on the fit of a model that includes a penalty for complexity when used to measure the quality of a statistical model. The most commonly used information criteria for linear regression analysis are AIC and BIS/SIC.

3.5.1. AIC

Akaike Information Criterion (also known as AIC) is formulated by Hirotugu Akaike and is a method to measure the quality of each of the available statistical models for the dataset.

$$AIC = -2 * ln(L) + 2 * k \tag{8}$$

k is the number of estimated parameters, and L is the value of the likelihood. The smaller the AIC score, the better the model is.

3.5.2. BIC

Bayesian information criterion (also known as BIC) is one variation of AIC and is formulated by Schwarz Bayesian. It is more conservative than AIC and more likely to select a simpler model due to the greater penalty.

$$BIC = -2 * ln(L) + 2 * ln(N) * k \tag{9}$$

k is the number of estimated parameters, N is the number of recorded measurements or sample size, and L is the value of the likelihood. The smaller the BIC score, the better the model is.

3.5.3. Mallow's cp

Mallow's Cp is formulated by Colin Lingwood Mallows, and is a method to assess the fit of the regression model by using ordinary least squares (also known as OLS)

$$C_p = \frac{RSS}{s^2} - n + 2(p + 1) \tag{10}$$

RSS is the residual sum of squares (also known as SSE or SSR, the sum of the squared estimate of errors, the sum of squared residuals), n is the sample size, and p is the number of parameters.

4. Discussion

When considering the model selection and variable selection in linear regression, the fitness, and the contribution are important in decision-making because, for further linear regression analysis, the variables that help to predict and explain the response value are the primary focus of analysis.

The p-value of the t-test and F-test is not a good estimator for variable selection or model selection. Trying to keep significant values and eliminate non-significant values depending on the p-value of hypothesis testing cannot help the decision-making for analysis, because the test statistics and p-value can be affected by the coefficient, noise, sample size, variance, and correlation. When coefficients are large, the $\hat{\beta}_j$, the test statistics will be large which means more significant. Reducing the noise around the regression line will also increase the test statistics and make all independent variables more significant. A larger sample size means larger test statistics which makes variables more significant. More variance in an independent variable while all else being equal will increase the test statistics making the independent variable more significant. Also, the correlation between each independent variable will decrease the test statistics which means less significance. For test statistic and p-value, it explains how well the articular coefficient is, but it does not provide any information related to the relevance between the response variable and independent variable.

Stepwise variable selection based on some criterion like p-value, $adjusted\ R^2$, Mallow’s Cp, AIC, or cross-validation. Forward stepwise is a good estimator for a dataset with a large number of variables, even larger than the sample size since it starts with the null model without considering the full model. Backward stepwise will be a good estimator when considering collinearity. This is because, the backward stepwise start with the null model, and it is forced to keep the variables with correlation. However, since stepwise is based on some criterion or rule, it cannot consider all possible combinations of independent variables, and it might not choose the best model with biased regression coefficients, p-value, and $adjusted\ R^2$, because of the rule of adding or eliminating variables.

AIC is a good estimator to predict future observations because AIC tries to keep more information with selected variables or models. In this case, AIC will deal with underfitting by keeping predictors with important information, while also eliminating the noise to reduce the risk of overfitting.

BIC works best to select a correct model since it is used to select the simplest true model from the set of candidate models. This is because BIC will choose this model with probability 1 when the simplest true model is included in the list of candidates and n is approximately infinity, while the probability of AIC is less than 1. If there is two or more model is true, AIC may not choose the simplest true model while BIC will. In this case, it is hard for BIC to model the dataset whose sample size is much smaller than the number of parameters, and it cannot handle high-dimension complex model space.

Mallow’s Cp is used to assess the fit of the regression model and addresses the overfitting in variable selection or model selection because this method tries to find the model with the smallest RSS. Nevertheless, Mallow’s Cp is only valid for large sample sizes, because it relies on the OLS, error sum of squares, and the sample size is a crucial element in its equation. Also, Mallow’s Cp cannot be used to select models in complex collections when doing model selection or variable selection.

Since AIC, BIC, and Mallow’s Cp are information criteria, they are limited to computing models for the known complexity. For linear regression, the complexity is known, but if there is no strong linear relationship between the independent variables and response variables, the result of these three information criteria will have errors and be not precise.

5. Conclusion

When predicting a future event in a probabilistic reasoning system, linear regression is a useful method to predict and explain the relationship between predictor or independent variables and response variables. Since the original dataset might contain some unknown problems like noise and overfitting, variable selection and model selection are performed to select the wanted model for analysis. The model and method used in linear regression depending on the purpose of the analysis. Stepwise selection is good with customizable rules or criteria for variable and model selection in linear regression. If predicting the future event or observation is the primary goal, AIC might be the best choice for model selection, since it returns the model with crucial information and eliminates the variables causing overfitting. BIC can be used if the simplest true model is preferred for further analysis. And Mallow’s Cp can be a good information criterion to deal with overfitting and find the most crucial variables for decision-making.

References

[1] Boisbunon, Aurélie, et al. “Akaike’s Information Criterion, Cp and Estimators of Loss for Elliptically Symmetric Distributions.” International Statistical, vol. 82, no. 3, 2014, pp. 422–39. JSTOR, http://www.jstor.org/stable/43299006. Accessed 6 Oct. 2022.

[2] Bozdogan, Hamparsum. “Model Selection and Akaike’s Information Criterion (AIC): The General Theory and Its Analytical Extensions.” Psychometrika, vol. 52, no. 3, Sept. 1987, pp. 345–70. Crossref, https://doi.org/10.1007/bf02294361.

[3] Burnham KP, Anderson DR. Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach. 2nd ed. New York: Springer; 2002

[4] Chakrabarti, Arijit, and Jayanta K. Ghosh. “AIC, BIC and Recent Advances in Model Selection.” Philosophy of Statistics, North-Holland, 17 June 2011.

[5] Gideon Schwarz. "Estimating the Dimension of a Model." Ann. Statist. 6 (2) 461 - 464, March, 1978. https://doi.org/10.1214/aos/1176344136

[6] Gilmour, Steven G. “The Interpretation of Mallows’ Cp Statistic.” Journal of the Royal Statistical Society. Series D (The Statistician), vol. 45, no. 1, 1996, pp. 49–56. JSTOR, https://doi.org/10.2307/2348411. Accessed 6 Oct. 2022.

[7] Giraud, C, Introduction to high-dimensional statistics, Chapman & Hall/CRC, 2015, ISBN 9781482237948

[8] H. Akaike, "A new look at the statistical model identification," in IEEE Transactions on Automatic Control, vol. 19, no. 6, pp. 716-723, December 1974, doiaking: 10.1109/TAC.1974.1100705.

[9] H.-I. Lim, "A Linear Regression Approach to Modeling Software Characteristics for Classifying Similar Software," in 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), 2019, pp. 942-943.

[10] J. Wu, C. Liu, W. Cui, and Y. Zhang, "Personalized Collaborative Filtering Recommendation Algorithm based on Linear Regression," in 2019 IEEE International Conference on Power Data Science (ICPDS), 2019, pp. 139-142.

[11] Mallows, C. L. "Some Comments on Cp". Technometrics. 15 (4): *661–675,* 1973. *doi:10.2307/1267*

[12] M. R. Sarkar, M. G. Rabbani, A. R. Khan and M. M. Hossain, "Electricity demand forecasting of Rajshahi city in Bangladesh using fuzzy linear regression model," *2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, 2015, pp. 1-3, doi: 10.1109/ICEEICT.2015.7307424.

[13] Roopa, H. and T. Asha. “A Linear Model Based on Principal Component Analysis for Disease Prediction.” *IEEE Access* 7 2019: 105314-105318.

[14] Scavia, Donald, Joseph V. DePinto, and Isabella Bertani. "A multi-model approach to evaluating target phosphorus loads for Lake Erie." Journal of Great Lakes Research 42.6 (2016): 1139-1150.

[15] Daghighi, Amin. "Harmful algae bloom prediction model for western lake erie using stepwise multiple regression and genetic programming." (2017).

[16] Zhou, Zheng-Xi, Ren-Cheng Yu, and Ming-Jiang Zhou. "Resolving the complex relationship between harmful algal blooms and environmental factors in the coastal waters adjacent to the Changjiang River estuary." Harmful Algae 62 (2017): 60-72.

[17] Lou, Inchio, et al. "Integrating support vector regression with particle swarm optimization for numerical modeling for algal blooms of freshwater." Advances in Monitoring and Modelling Algal Blooms in Freshwater Reservoirs. Springer, Dordrecht, 2017. 125-141.

[18] Yun, Hongwon. "Prediction model of algal blooms using logistic regression and confusion matrix." International Journal of Electrical and Computer Engineering 11.3 (2021): 2407.

[19] Lee, Sangmok, and Donghyun Lee. "Improved prediction of harmful algal blooms in four Major South Korea’s Rivers using deep learning models." International journal of environmental research and public health 15.7 (2018): 1322.

[20] Bouquet, Aurélien, et al. "Prediction of Alexandrium and Dinophysis algal blooms and shellfish contamination in French Mediterranean Lagoons using decision trees and linear regression: a result of 10 years of sanitary monitoring." Harmful Algae 115 (2022): 102234.

[21] Xia, Jingjing, and Jin Zeng. "Early warning of algal blooms based on the optimization support vector machine regression in a typical tributary bay of the Three Gorges Reservoir, China." Environmental Geochemistry and Health (2022): 1-15.

[22] Chen, Qiuwen, and Arthur E. Mynett. "Predicting Phaeocystis globosa bloom in Dutch coastal waters by decision trees and nonlinear piecewise regression." Ecological modelling 176.3-4 (2004): 277-290.

[23] Chang, K-W., Y. Shen, and P-C. Chen. "Predicting algal bloom in the Techi reservoir using Landsat TM data." International Journal of Remote Sensing 25.17 (2004): 3411-3422.