Research Article

Application of logistic regression equation analysis using derivatives for optimal cutoff discriminative criterion estimation

Andrey Bokov[1]* and Svitlana Antonenko[2]

[1]Chief of Oncology and Neurosurgery Department, Federal Budgetary Institution Privolzhskiy Medical Research University, Russia

[2]Faculty of Applied Mathematics, Chief of Department of Mathematical Support of Computers, Oles Honchar Dnipro National University, Ukraine

Abstract

Background: Sigmoid curve function is frequently applied for modeling in clinical studies. The main task of scientific research relevant to medicine is to find rational cutoff criterion for decision making rather than finding just equation for probability calculation.

The objective of this study is to analyze the specific features of logistic regression curves in order to evaluate critical points and to assess their implication for continuous predictor variable dichotomization in order to provide optimal cutoff criterion for decision making.

Methods: Second order and third order derivatives were used to analyze estimated logistic regression function, critical values of independent continuous variable that correspond to zero points of second and third derivative were calculated for each logistic regression equation. Using those values continuous predictors of each logistic regression equations were converted into dichotomized scales using 1 value that correspond to second order derivative and 2 values that correspond to zero points of third derivative then receiver operating characteristics of estimated equations with dichotomized predictor were assessed.

Results: Sigmoid curve of logistic regression has the same structure with inflection point corresponding probability 0.5 (zero value of second derivative) and maximal torsion (zero values of third derivative) corresponding 0.2113 and 0.7886 probability values. Thresholds accounting for predictor values that correspond to zero values of second and third derivative provide estimation of logistic regression applying dichotomized predictor with optimal ratio of sensitivity, specificity and overall accuracy with maximal area under curve.

Conclusion: Analysis of logistic regression equation with continuous predictor applying derivatives help to choose optimal thresholds that provide maximally effective discriminative functions with priority sensitivity or specificity. Using this dichotomization discriminative function can be adjusted to the needs of particular task or study depending which characteristic is in priority – sensitivity or specificity.

Background

Sigmoid curve function is frequently applied for modeling population growth in limited resources; also sigmoid function is applied for logistic regression analysis that is also frequently used in clinical studies [1-3]. Logistic regression analysis is often used to estimate the relationships between dichotomized dependent variable and either continuous or dichotomized independent variable, result of regression analysis is conversion odds into probability [4]. The examples of logistic regression could be modeling of low energy fracture probability or implants failure in orthopedic surgery as response to characteristics of bone, for example, radiodensity or calcium density. In the majority of cases the main task of scientific

research relevant to medicine is to find rational cutoff criterion for decision making rather than finding just equation for probability calculation.

It has been estimated that sigmoid curve that is used for population growth modeling has several phases: lag phase with minimal growth responding to changes in predictor values, phase of initial acceleration, phase of exponential growth with maximal changes of dependent variable per unit change of predictor value, phase of negative acceleration and stationary phase with minimal changes of dependent variable per unit change of predictor [5]. Taking into account that sigmoid curve is used for logistic regression modeling it is assumed that sigmoid curve estimated for logistic regression modeling have the same segments that has different patterns in dependent variable growth. Knowing the borders between those segments may help to convert continuous predictor into dichotomized achieving optimal receiver operative characteristics providing optimal discrimination of cases.

An analysis of function using derivatives is used to find critical points of function and changes in patterns in dependent variable growth [6,7]. Zero values of first order derivative correspond to maximal and minimal values of function while second derivative zero value corresponds to inflection point on the graph [6,8]. The meaning of third derivative in geometry is aberrance – the torsion of curve, in mechanics it relates to jerk definition in other words it represents a rate at which acceleration changes [6,9]. Critical points that can be estimated using derivative analysis can define an inflection point of graph and points with maximal torsion that may delineate segments of graph providing optimal values for continuous data dichotomization.

The objective of this study is to analyze the specific features of logistic regression curves in order to evaluate critical points and to assess their implication for continuous predictor variable dichotomization in order to provide optimal cutoff criterion for decision making.

Methods

Three logistic regression models were used for this study

1. Logistic regression model for low energy fractures of lumbar vertebrae prediction, the continuous predictor is radiodensity, dependent dichotomized variable is probability of low energy fracture. The number of cases in study 150.

2. Logistic regression model for multilevel low energy vertebra fractures prediction, the continuous predictor is radiodensity, dependent variable – probability of multilevel fracture, the number of cases in study – 150.

3. Logistic regression equation for probability of spinal stenosis clinical presentation, the continuous predictor is square of vertebral channel.

Second order and third order derivatives were used to analyze estimated logistic regression function, critical values of independent continuous variable that correspond to zero points of second and third derivative were calculated for each logistic regression equation. Using those values continuous predictors of each logistic regression equations were converted into dichotomized scales using 1 value that correspond to second order derivative and 2 values that correspond to zero points of third derivative. Finally, 3 new logistic regression equations were estimated for every study using values of predictor that corresponds to zero value of second derivative and two zero values of third derivative. Than every equation was assessed using ROC curves analysis, the results were compared to the results of arbitrary dichotomization.

Sofrware used for this study: wolfram mathematica, statistica 12, SPSS 20.

Results

It has been estimated, that logistic regression curve has the same structure, zero point of second derivative – an inflection point corresponds to 0.5 value of probability while zero points of third derivative correspond to probability values 0.2113 and 0.7886. Those three dichotomizations using cutoff values that correspond to zero points of second and third order derivatives provided three logistic regression models for each study. The estimated three logistic regression models for each study have different sensitivity, specificity and accuracy. To assess discriminative value of estimated critical points of graph receiver operative characteristics analysis, calculation of sensitivity specificity and overall accuracy were performed. Characteristics of equations estimated using dichotomization with derivative analysis were compared to equations with arbitrary dichotomization. It has been found that predictor value that corresponds to zero point of second order derivative provide equation that has the same receiver operative characteristics as equation that uses continuous predictor. Points of maximal graph torsion that corresponds to zero points of third order derivative zero values provide two equations: first minor value of predictor provides equation with maximally achievable accuracy and maximal value of specificity, the second major value provides equation with maximally achievable accuracy with maximal value of sensitivity. Using ROC curve analysis it has been defined that estimated regression equations using those three values that were calculated using derivatives provide three maximal values of area under curve Table 1. Illustrates application of continuous data dichotomization using derivative analysis for the study on low energy vertebra fracture prediction based on radiodensity in Hounsfield Units (HU) (Table 2).

Initial equation: $Y = e^{-1.6632 + 0.0445 \cdot x}; 1 + e^{-1.6632 + 0.0445 \cdot x}$

Receiver operating characteristics of initial regression model were: overall accuracy 82%, sensitivity 60%, specificity 91%.

Third derivative equation:

$$Y = 0.0445^3 * \frac{e^{-1.6632+0.0445*x}}{1+e^{-1.6632+0.0445*x}} - 7*0.0445^3 * \frac{\left(e^{-1.6632+0.0445*x}\right)^2}{\left(1+e^{-1.6632+0.0445*x}\right)^2} + 12*0.0445^3$$

$$* \frac{\left(e^{-1.6632+0.0445*x}\right)^3}{\left(1+e^{-1.6632+0.0445*x}\right)^3} - 6*0.0445^3 * \frac{\left(e^{-1.6632+0.0445*x}\right)^4}{\left(1+e^{-1.6632+0.0445*x}\right)^4}$$

Dichotomization of continuous predictor variables applying analysis of initial logistic regression equation with derivatives provides an opportunity of a threshold selection with optimal discriminative function for cases.

Table 1: Receiver operating characteristics of three models estimated using predictor thresholds corresponding to zero values of second and third order derivatives.

	Sensitivity	Specificity	Overall accuracy	ROC curves analysis area under curve
Value 97.2 HU - major zero value of third derivative	82	77	79	0.801
Value 67.6 HU – zero of second order derivative	60	91	83	0.755
Value 38.0 HU – minor zero value of third order derivative	32	99	82	0.658

Table 2: Receiver operating characteristics of three models estimated using arbitrary cutoff criteria.

Value of radiodensity HU	Sensitivity	Specificity	Overall accuracy	ROC curves analysis area under curve
40 HU	37	99	82	0.683
50 HU	37	93	79	0.656
60 HU	50	95	83	0.728
70 HU	62	91	84	0.768
80 HU	67	83	79	0.757
90 HU	77	79	79	0.785
100 HU	85	75	78	0.800
110 HU	0	100	NA	0.776
120 HU	0	100	NA	0.752

Discussion

Estimation of discriminative function for classification of cases is one of the most frequent goals of medical research [4]. Logistic regression is frequently used for those purposes if dependent variable is registered in dichotomized scale. The result of the analysis is an equation for probability calculation that can be used to forecast likelihood of complication or probability of unsatisfactory results of particular treatment modality application [4,10]. In certain cases practitioners need also optimal thresholds for decision making, however equations with continuous predictors do not provide this information consequently optimal dichotomization of continuous data is required [11].

The effectiveness of discrimination system worked out using logistic regression analysis is assessed using ROC-curves analysis that reflects the ability to classify cases correctly.

The better ability of discriminative function has the greater area under curve [12,13]. Finally logistic regression function can be characterized by accuracy of classification, sensitivity and specificity [14]. The problem of sensitivity and specificity balance is akin to the problem of type I error and type 2 errors in statistics: the decrease in one error probability results in an inevitable increase in rate of another error. A good example could be selection of patients with pain syndromes caused by degenerative diseases of spine for interventional pain management. With getting criteria for selection stricter at first the number of failures is getting decreased but in applying further restriction a certain number of patients who may benefit from minimal invasive interventions will be missed [15]. It has been clearly defined that interventions applied for treatment of those patients are associated with minimal risks of complication as a consequence selection criteria should be based on optimal accuracy with maximally achievable sensitivity. Conversely in patient selection for risky traumatic surgery maximal specificity in terms of clinically significant result with optimal accuracy should be used.

Analysis using derivatives of function is frequently used in physics and differential geometry for graph analysis [6–8]. Using derivatives of second and third order it is possible to estimate points of graph inflection and points with maximal curve torsion that delineate the borders of exponential growth in probability. Those points correspond to critical values of predictors that correspond to critical changes in patterns of probability growth. If those values are used for dichotomization of continuous predictor, three logistic regression model equations with different characteristics can be estimated. Logistic regression equation that applies value of predictor that correspond to zero of second derivative has the same receiver operating characteristics as equation with continuous predictor. Minor value corresponding to zero value of third derivative used for logistic regression equation provides model with optimal overall accuracy and maximal specificity while second – major value of predictor that correspond to the second zero point of third derivative provide classification system with optimal overall accuracy and maximal sensitivity. Those three equations have three maximal values of area under ROC curve compared to dichotomization using arbitrary values. The results of our study demonstrate that those predictor values constantly correspond to 0.5; 0.2113 and 0.7886 probability values.

It is well known fact, that in studies relevant to clinical practice it is hardy possible to work out ideal discrimination function for prediction. In certain cases scientist of practitioner should choose which characteristic of discrimination function is in priority: specificity or sensitivity. Application of two values of continuous predictor that correspond to zero values of third derivative for dichotomization provide equations with maximal sensitivity or specificity along with optimal accuracy.

Limitation

Feasibility of suggested analysis can be limited by goodness-of-fit of initial equation with continuous predictor.

Conclusion

Analysis of logistic regression equation with continuous predictor applying derivatives help to choose optimal thresholds that provide maximally effective discriminative functions with priority sensitivity or specificity. Using this dichotomization discriminative function can be adjusted to the needs of particular task or study depending which characteristic is in priority – sensitivity or specificity.

References

1. Peleg M, Corradini MG, Normand MD (2007) The logistic (Verhulst) model for sigmoid microbial growth curves revisited. Food Research International 40: 808-818. **Link:** https://bit.ly/31XB5mK

2. Harre FE, Lee KL, Pollock BG (1988) Regression models in clinical studies: determining relationships between predictors and response. J Natl Cancer Inst 80: 1198-1202. **Link:** https://bit.ly/3gfrlJT

3. Meurer WJ, Tolles J (2017) Logistic regression diagnostics: understanding how well a model predicts outcomes. JAMA 317: 1068-1069. **Link:** https://bit.ly/3aBwdro

4. Boateng EY, Abaye DA (2019) A Review of the Logistic Regression Model with Emphasis on Medical Research. Journal of Data Analysis and Information Processing 7: 190-207.**Link:** https://bit.ly/324Fg02

5. Yoshinaga T, Hagiwara A, Tsukamoto K (2001) Why do rotifer populations present a typical sigmoid growth curve?. Hydrobiologia 446: 99-105. **Link:** https://bit.ly/34bQW3Q

6. Kobayashi S (2020) Differential geometry of curves and surfaces. Springer doi: 10.1007/978-981-15-1739-6

7. Schot SH (1978) Aberrancy: Geometry of the third derivative. Mathematics Magazine 51: 259-275. **Link:** https://bit.ly/2CHvrwA

8. Christopoulos DT (2016) On the efficient identification of an inflection point. International Journal of Mathematics and Scientific Computing 6. **Link:** https://bit.ly/3iPJ389

9. Herişanu N, Marinca V (2016) Approximate Analytical Solutions to Jerk Equations. In: Awrejcewicz J. (eds) Dynamical Systems: Theoretical and Experimental Analysis. Springer Proceedings in Mathematics & Statistics, 182. **Link:** https://bit.ly/2Q02O0D

10. Peng CYJ, Lee KL, Ingersoll GM (2002) An introduction to logistic regression analysis and reporting. The Journal of Educational Research 96: 3-14. **Link:** https://bit.ly/3iRWSTA

11. Steyerberg EW, Eijkemans MJ, Harrell FE, Habbema JDF (2001) Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Medical Decision Making 21: 45-56. **Link:** https://bit.ly/2Q37sLg

12. Carter JV, Pan J, Rai SN, Galandiuk S (2016) ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. Surgery 159: 1638-1645. **Link:** https://bit.ly/3kXgbNp

13. Obuchowski NA, Bullen JA (2018) Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine. Phys Med Biol 63: 07TR01. **Link:** https://bit.ly/3kVGQdu

14. Grigoryev SG, Lobzin YV, Skripchenko NV (2016) The role and place of logistic regression and ROC analysis in solving medical diagnostic task. Journal Infectology 8: 36-45. **Link:** https://bit.ly/2YaZSCQ

15. Bokov A, Perlmutter O, Aleynik A, Rasteryaeva M, Mlyavykh S (2013) The potential impact of various diagnostic strategies in cases of chronic pain syndromes associated with lumbar spine degeneration. Journal of Pain Research 6: 289-296. **Link:** https://bit.ly/34c6ox0