บทคัดย่อ

ชื่อโครงการ ตัวแบบ SIMPLE AND MULTIPLE LOGISTIC MODELS สำหรับตัวแปรตอบสนองแบบ DICHOTOMOUS และตัวแปรอธิบายแบบต่อเนื่องและแบบเชิงกลุ่ม

ชื่อผู้วิจัย รองศาสตราจารย์ วีรานันท์ พงศาภักดี

หน่วยงานที่สังกัด ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร

แหล่งทุนอุดหนุนการวิจัย สถาบันวิจัยและพัฒนา มหาวิทยาลัยศิลปากร

ปีที่ทำเสร็จ 2548

ตัวแบบโลจิสติกได้ขยายขอบเขตของตัวแบบให้มีตัวแปรอธิบายทั้งแบบไม่ต่อเนื่องและแบบต่อเนื่อง ซึ่ง ในกรณีของตัวแปรอธิบายหนึ่งตัวแปรและหลายตัวแปรเรียกว่าตัวแบบโลจิสติกแบบง่ายและแบบพหุคูณ (simple and multiple logistic models) ตามลำดับ และในกรณีตัวแปรตอบสนอง Y มี 2 กลุ่ม ซึ่งเป็นกรณีที่มีการใช้และพบบ่อยในหลายสาขาทาง วิทยาศาสตร์สุขภาพ วิทยาศาสตร์การแพทย์ วิศวกรรมศาสตร์ และใช้แพร่หลายมากขึ้นทางสังคมศาสตร์ตลอดจนทางด้านการควบคุมคุณภาพ แต่ในปัจจุบันยังพบปัญหาคือ ปัญหาการประเมินความเหมาะสมของตัวแบบ มีตัวสถิติต่าง ๆ ที่อาจใช้ในการวิเคราะห์เป็นจำนวนมาก และตัวสถิติตัวใดเป็นตัวสถิติที่เหมาะสม งานวิจัยนี้ จึงสนใจการจำลองแบบตัวแบบโลจิสติกภายใต้การแจกแจงของตัวแปรอธิบายแบบ Exponential, Bernoulli และMultinomial เพื่อเปรียบเทียบและศึกษาความเหมาะสมของตัวแบบและตัวสถิติที่วัดจากเทอมต่อไปนี้ ร้อยละของการจำแนกกลุ่มถูกต้อง (%Correct) ร้อยละของกำลังการทดสอบ (%Accept) ตัวสถิติอัตราส่วนภาวะน่าจะเป็น (likelihood ratio statistic,G_M) ดัชนีประสิทธิภาพในการทำนาย (indexes of predictive efficiency, λ_P, τ_P and ϕ_P) สหสัมพันธ์แห่งการกำหนดต่าง ๆ (coefficients of determination, R^2 - analogs) และเมทริกซ์ขนาด(ค่าสัมบูรณ์)สหสัมพันธ์เพื่อตรวจสอบความเป็นอิสระต่อกันของตัวสถิติต่าง ๆ กับอัตราพื้นฐาน (Base rate) โดยการจำลองแบบจากโปรแกรมแมโครคอมพิวเตอร์ที่ประมวลผลร่วมกับโปรแกรม MINITAB ในแต่ละเงื่อนไขย่อยของพารามิเตอร์และการแจกแจงความน่าจะเป็นของ Y=1 คือ 0.05, 0.20, 0.35, 0.50 (ตาราง 1) แต่ละชุดแต่ละตัวแปร ๆ ละ 200 หน่วย ทำซ้ำ 1,000 ชุด

ผลการวิจัยพบว่า % Correct (77-99 %) และ % Accept (94-96 %) ให้ผลสอดคล้องกันในทุกการแจกแจง และภายใต้การแจกแจงแบบ Exponential ควรใช้ตัวสถิติ R^2_M, R^2_C, R^2_O โดยเฉพาะ R^2_M รวมถึงตัวสถิติดัชนีประสิทธิภาพ λ_P, τ_P และ ϕ_P ซึ่งพบว่า λ_P และ ϕ_P ใช้ได้ดีกว่า τ_P สำหรับการแจกแจงแบบ Bernoulli พบว่า R^2_O, R^2_M, R^2_C ให้ผลดี แต่เน้นที่ R^2_O ที่เด่นกว่าตัวสถิติอื่น ๆ ซึ่งรวมถึงตัวสถิติดัชนีประสิทธิภาพที่พบว่า λ_P และ τ_P เด่นกว่า ϕ_P ส่วนการแจกแจงแบบ Multinomial ให้ผลโดยทั่วไปคล้ายคลึงกับผลของการแจกแจงแบบ Bernoulli แต่มีแนวโน้มดีกว่า โดยเฉพาะผลของตัวสถิติดัชนีประสิทธิภาพของการแจกแจงแบบ Multinomial ให้ผลลัพธ์ดีกว่าของการแจกแจงแบบ Bernoulli อย่างไรก็ตามยังพบข้อควรระวังในกรณีของการแจกแจงแบบ Exponential คือ เมื่อความน่าจะเป็นของ Y=1 (P_1) เข้าใกล้ 0.5 ตัวสถิติ % Correct มีค่าต่ำกว่าของกรณีอื่น ๆ โดยเฉพาะให้ค่าพิสัยหรือการกระจายมากกว่าของกรณีอื่น ๆ ดังนั้นการวิจัยครั้งต่อไป ควรศึกษาในรายละเอียดของกรณีนี้เพิ่มเติม และอาจศึกษากรณีของขนาดตัวอย่างที่ใหญ่ขึ้นเพื่อให้ผลลัพธ์ชัดเจนยิ่งขึ้น

ABSTRACT

177707

Research Title SIMPLE AND MULTIPLE LOGISTIC MODELS ASSOCIATED WITH THE DICHOTOMOUS RESPONSE AND THE COMBINATION OF CONTINUOUS AND CATEGORICAL EXPLANATORY VARIABLES

Researcher Assoc. Prof. Veeranun Pongsapukdee
Office Department of Statistics, Faculty of Science, Silpakorn University
Research Grants Institute of Research and Development, Silpakorn University
Year 2005

The logistic model to allow for one or several explanatory variables of which the model is also called simple or multiple logistic model, respectively. In the usual case of logistic models, the basic random variable Y is dichotomous response which is commonly used procedure in many disciplines in health sciences research, medical sciences, engineering settings, and is becoming increasingly popular in the behavioral and social sciences and in quality control. In this model data taking the value 1 with the success probability P_1, and the value 0 with the failure probability ($1-P_1$). Problems arise with different proposed statistics for assessing the fit of the models and which one of them is more suitable. In this article, 1,000 computer simulation experiments in each condition of the probability of Y=1 (P_1), calculated parameters and X's distributions, were generated to evaluate the performance of several statistics, all of which were used for assessing the goodness-of-fit of the models. Ten statistics were computed for each combination of base rate levels and model conditions (Table 1): the likelihood ratio statistics (G_M), the indexes of predictive efficiency which consist of λ_P, τ_P and ϕ_P (Menard, 1995), the coefficients of determination or R^2 - analogs which consist of R^2_C (the contingency coefficient R^2), R^2_L (the log likelihood ratio R^2, R^2_M (the geometric mean squared improvement per observation R^2), R^2_N (the adjusted geometric mean squared improvement R^2), and R^2_O (the ordinary least squares R^2). Moreover, the correlation matrices for determining their magnitude (absolute values) of the measures of independence from the base rate levels, the percentages of correct classification of the model (%Correct) and the type II error rates, corresponding to the percentages of power of the tests (%Accept) were computed.

The results of the simulation studies show that, for hypothesis testing goodness-of-fit of models, both of the %Correct (77-99 %) and the %Accept (94-96 %) are all satisfied. The average of %Correct, when X is Exponential is around 77% and when X's are Bernoulli and multinomial distributed, they are approximately equal to 99%. Similarly for the average of %Accept which are also approximately equal to 94.%. For X~ Exponential, R^2_C, R^2_M, and R^2_O are preferable and for X~ Bernoulli R^2_C, R^2_M, R^2_O are still preferable but R^2_o outperforms. For (X_1, X_2)~ Multinomial, the results are similar but slightly superior to those of X~ Bernoulli. The indexes of predictive efficiency of the multinomial case when the success probability P_1 is high, the λ_P, τ_P statistics may be used as alternatives of the R^2_C, R^2_M and R^2_O.

Some recommendations are made for logistic models with dichotomous response and exponential explanatory variable distributed. That are the statistics R^2_C, R^2_M, R^2_O, λ_P and ϕ_P probably be interesting to use ;however, when P_1 is closed to 0.5 the %Correct is low and it's range is high. Therefore, further studies in more details for the exponential explanatory variable and the increased sample sizes would be recommended. The logistic models with dichotomous response and Bernoulli and multinomial exponential explanatory variables are much improved, Then, the statistics R^2_C, R^2_M, R^2_O, λ_P and τ_P are probably appropriated, especially the R^2_O statistic.