Prise en compte de l'incertitude en régression logistique

Fahmy, Thierry, fahmy@xlstat.com, Addinsoft, Bordeaux (Orateur)

Orateur : Thierry Fahmy est le fondateur et le dirigeant de la société Addinsoft. Thierry Fahmy est Ingénieur AgroParisTech et Dr. en Statistique.

Présentation de la société : Addinsoft est un éditeur de logiciels de mathématiques appliquées. Créée en 2000, Addinsoft est l'un des principaux acteurs du marché des logiciels d'analyse de données. L'entreprise a notamment développé le logiciel XLSTAT qui compte plus de 150 000 utilisateurs répartis dans plus de 120 pays, et s'est récemment diversifiée à travers 3 nouveaux logiciels XLOPTIM, XLRISK et ZENPLOT.

Thématique : 4 / Analyse des données

Résumé : *La régression logistique, qu'elle s'applique à des variables dépendantes binomiales, multinomiales ou ordinales, est très utilisée pour prédire l'appartenance à une classe. Néanmoins, hormis dans le cas binomial pour certaines applications, il est rare que soient prises en compte les incertitudes. Nous présentons ici une façon de prendre en compte les incertitudes, quel que soit le type de variable dépendante.*

Mots clés : *régression logistique, incertitude, matrice de confusion, indices*

1. Introduction

La régression logistique appartient à la famille des méthodes GLM (generalized linear model). C'est sans doute la méthode de prédiction la plus utilisée après la régression linéaire ordinaire. Néanmoins, lorsque la variable dépendante (à prédire) est qualitative, les utilisateurs se contentent souvent d'analyser la sensibilité et la spécificité après la conversion des probabilités prédites en la modalité associée à la probabilité maximale.

2. Méthodologie

Revenir aux intervalles de confiance autour des probabilités calculées nous semble primordial. Ceux-ci sont bien connus pour le cas de régression logistique sur variables binomiales. Nous présenterons comment procéder dans le cas de variables multinomiales et ordinales, non directement à partir d'intervalle de confiance mais à partir de ratios de probabilités. La prise en compte des incertitudes nous amène à créer une nouvelle classe d'affectation « incertaine », ajoutant ainsi une colonne à la matrice de confusion. Nous avons également mis au point une nouvelle visualisation de cette matrice. Enfin, nous proposons un nouvel indice de la qualité d'ajustement du modèle, le Goodness of Classification Index (GCI).

3. Originalité

Nous apportons trois nouveaux outils d'évaluation et d'interprétation aux utilisateurs de la régression logistique.

Références

Hosmer D.W. and Lemeshow S. (2000). Applied Logistic Regression, Second Edition. John Wiley and Sons, New York.

Lang J. B. (2014). The Pearson Score Statistic for Multinomial-Poisson Models. Communications in Statistics - Theory and Methods, 43(21), 4471-4491.

Sambamoorthi N., Ervin V.J. and Thomas G. (1994). Simultaneous prediction intervals for multinomial logistic regression models. *Communications in Statistics - Theory and Methods*, 23(3), 815-829.