Electronic Journal of Statistics > Vol. 2 (2008) open journal systems

Comparing two samples by penalized logistic regression

Konstantinos Fokianos, *University of Cyprus*

Abstract

Inference based on the penalized density ratio model is proposed and studied. The model under consideration is specified by assuming that the log--likelihood function of two unknown densities is of some parametric form. The model has been extended to cover multiple samples problems while its theoretical properties have been investigated using large sample theory. A main application of the density ratio model is testing whether two, or more, distributions are equal. We extend these results by arguing that the penalized maximum empirical likelihood estimator has less mean square error than that of the ordinary maximum likelihood estimators, especially for small samples. In fact, penalization resolves any existence problems of estimators and a modified Wald type test statistic can be employed for testing equality of the two distributions. A limited simulation study supports further the theory.

AMS 2000 subject classifications: Primary62G05; secondary 62G20.

Keywords: Empirical likelihood, biased sampling, penalty, semiparametric, shrinkage, mean square error, power.

Full Text: PDF

Fokianos, Konstantinos, Comparing two samples by penalized logistic regression, Electronic Journal of Statistics, 2, (2008), 564-580 (electronic). DOI: 10.1214/07-EJS078.

References

Albert, A. and J. A. Anderson (1984). On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71, 1–10. MR0738319

Anderson, J. A. (1972). Separate sample logistic discrimination. Biometrika 59, 19–35. MR0345332

Anderson, J. A. (1979). Multivariate logistic compounds. Biometrika 66, 17–26. MR0529143

Antoniadis, A. and J. Fan (2001). Regularization of wavelet approximations. J. Amer. Statist. Assoc. 96, 939–967. with discussion. MR1946364

Breslow, N. E. and N. E. Day (1980). The Analysis of Case–Control Data, Volume 1 of Statistical Methods in Cancer Research. World Helath Organization.

Cox, D. R. and E. J. Snell (1989). The Analysis of Binary Data (2nd ed.). London: Chapman & Hall. MR1014891

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96, 1348–1360. MR1946581

Farewell, V. T. (1979). Some results on the estimation of logistic models based on retrospective data. Biometrika 66, 27–32. MR0529144

Fokianos, K. and E. Kaimi (2006). On the effect of misspecifying the density ratio model. Annals of the Institute for Statistical Mathematics 58, 475–497. MR2327888

Fokianos, K., B. Kedem, J. Qin, and D. A. Short (2001). A semiparametric approach to the one–way layout. Technometrics 43, 56–64. MR1819908

Frank, I. E. and J. H. Friedman (1993). A statistical view of some chemometric regression tools. Technometrics 35, 109–148.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. Journal of Computational and Graphical Statistics 7, 397–416. MR1646710

Gilbert, P. B. (2000). Large sample theory of maximum likelihood estimates in semiparametric biased sampling models. The Annals of Statistics 28, 151–194. MR1762907

Gilbert, P. B., S. R. Lele, and Y. Vardi (1999). Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. Biometrika 86, 27–43. MR1688069

Gill, R. D., Y. Vardi, and J. A. Wellner (1988). Large sample theory of empirical distributions in biased sampling models. The Annals of Statistics 16, 1069–1112. MR0959189

Hastie, T. and R. Tibshirani (2004). Efficient quadratic regulariazation for expression array. Biostatistics 5, 329–340.

Hoerl, A. E. and R. W. Kennard (1970a). Ridge regression:Applications to non–orthogonal problems. Technometrics 12, 69–82.

Hoerl, A. E. and R. W. Kennard (1970b). Ridge regression:Biased estimation for non–orthogonal problems. Technometrics 12, 55–67.

Knight, K. and W. Fu (2000). Asymptotics for lasso-type estimators. Ann. Statist. 28, 1356–1378. MR1805787

Le Cessie, S. and J. C. Van Houwelingen (1992). Ridge estimators in logistic regression. Applied Statistics 41, 191–201.

McCullagh, P. and J. A. Nelder (1989). Generalized Linear Models (2nd ed.). London: Chapman and Hall. MR0727836

Murphy, S. A. and A. W. van der Vaart (2000). On profile likelihood. Journal of the American Statistical Association 95, 449–485. with discussion. MR1803168

Owen, A. B. (2001). Empirical Likleihood. Boca Raton, Florida: Chapman and Hall/CRC.

Prentice, R. L. and R. Pyke (1979). Logistic disease incidence models and case–control studies. Biometrika 66, 403–411. MR0556730

Qin, J. (1998). Inferences for case–control data and semiparametric two–sample density ratio models. Biometrika 85, 619–630. MR1665814

Qin, J. and B. Zhang (1997). A goodness of fit test for the logistic regression model based on case–control data. Biometrika 84, 609–618. MR1603924

Santner, T. J. and E. D. Duffy (1989). Statistical Analysis of Discrete Data. New York: Springer. MR1019836

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58, 267–288. MR1379242

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. The Annals of Statistics 10, 616–620. MR0653536

Vardi, Y. (1985). Empirical distribution in selection bias models. The Annals of Statistics 13, 178–203. MR0773161

Zhang, B. (2000). M-estimation under a two-sample semiparametric model. Scand. J. Statist. 27, 263–280. MR1777503

Electronic Journal of Statistics. ISSN: 1935-7524