Electronic Journal of Statistics > Vol. 3 (2009) open journal systems

On lasso for censored data

Brent A. Johnson, *Emory University*

Abstract

In this paper, we propose a new lasso-type estimator for censored data after one-step imputatation. While several penalized likelihood estimators have been proposed for censored data variable selection through hazards regression, many such estimators require parametric or proportional hazards assumptions. The proposed estimator, on the other hand, is based on the linear model and least-squares principles. Iterative penalized Buckley-James estimators are also popular in this setting but have been shown to be unstable and unreliable. Unlike path-based learning based on least-squares approximation, our method requires no covariance assumption and the method is valid for even modest sample sizes. Our calibration estimator is the minimizer of a convex loss function using synthetic data and yields reproducible coefficient estimates and coefficient paths. The numerical algorithms are fast, reliable, and readily available because they build on existing software for complete, uncensored data. We examine the large and small sample properties of our estimator and illustrate the method through simulation studies and application to two real data sets.

Keywords: Accelerated failure time model, Buckley-James estimator, least angle regression, survival analysis, synthetic data.

Full Text: PDF

Johnson, Brent A., On lasso for censored data, Electronic Journal of Statistics, 3, (2009), 485-506 (electronic). DOI: 10.1214/08-EJS322.

References

[1] Buckley, J. and James, I. (1979). Linear Regression with Censored Data, Biometrika, 66, 429–436.

[2] Cai, T., Huang, J. and Tian, L. (2009). Regularized Estimation for the Accelerated Failure Time Model. Biometrics, (In press).

[3] Cox, D.R. (1972). Regression Models and Life-Tables (with discussion), Journal of the Royal Statistical Society, Ser. B, 34, 187–202. MR0341758

[4] Cox, D.R. and Oakes, D. (1984). Analysis of Survival Data, London: Chapman and Hall. MR0751780

[5] Datta, S., Le-Rademacher, J. and Datta, S. (2007). Predicting Survival from Microarray Data by Accelerated Failure Time Modeling using Partial Least Squares and Lasso. Biometrics, 63, 259–271. MR2345596

[6] Dickson, E.R., Grambsch, P.M., Fleming, T.R., Fisher, L.D. and Langworth, A. (1989). Prognosis in Primary Biliary Cirrhosis: Model for Decision Making, Hepatology, 10, 1–7.

[7] Efron, B. (2005). Local False Discovery Rates, Technical Report, Department of Statistics, Stanford University.

[8] Efron, B., Hastie, T., Johnstone, I.M. and Tibshirani, R. (2004). Least Angle

Regression (with discussion). The Annals of Statistics, 32, 407–499. MR2060166

[9] Fan, J. and Gijbels, I. (1996). Local polynomial modelling and its applications. London: CRC Press. MR1383587

[10] Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. Journal of the American Statistical Association, 96, 1348–1360. MR1946581

[11] Fleming, T.A. and Harrington, D.P. (1991). Counting Processes and Survival Analyses. New York: Wiley. MR1100924

[12] Fu, W.J. (2003). Penalized Estimating Equations. Biometrics, 35, 109–148. MR1978479

[13] Gehan, E.A. (1965). A Generalized Wilcoxon Test for Comparing Arbitrarily Single-Censored Samples. Biometrika, 90, 341–353.

[14] Geyer, C.J. (1994). On the Asymptotics of Constrained M-Estimation. The Annals of Statistics, 22, 1993–2010. MR1329179

[15] Huang, J., Ma, S. and Xie, H. (2006). Regularized Estimation in the Accelerated Failure Time Model with High-Dimensional Covariates. Biometrics, 62, 813–820. MR2247210

[16] Jin, Z., Lin, D.Y. and Ying, Z. (2006). On Least-Squares Regression with Censored Data. Biometrika, 93, 147–162. MR2277747

[17] Johnson, B.A. (2008a). Variable Selection in Semiparametric Linear Regression with Censored Data, Journal of the Royal Statistical Society, Ser. B, 70, 351–370.

[18] Johnson, B.A. (2008b). Estimation in the ℓ_1-Regularized Accelerated Failure Time Model. Technical Report, Emory University. Available at http://userwww.service.emory.edu/~bajohn3/pubs.html.

[19] Johnson, B.A. (2009). Rank-based estimation in the ℓ_1-Regularized Partly Linear Model for Censored Outcomes with Application to Integrated Analyses of Clinical Predictors and Gene Expression Data. Technical Report, Emory University. Available at http://userwww.service.emory. edu/~bajohn3/pubs.html.

[20] Johnson, B.A. and Peng, L. (2008). Rank-based Variable Selection, Journal of Nonparametric Statistics, 20, 241–252. MR2421768

[21] Johnson, B.A., Lin, D.Y. and Zeng D. (2008). Penalized Estimating Functions and Variable Selection in Semiparametric Regression Models. Journal of the American Statistical Association, 103, 672–680. MR2435469

[22] Kalbfleisch, J.D. and Prentice, R.L. (2002). The Statistical Analysis of Failure Time Data, 2nd Ed., Hoboken, Wiley. MR1924807

[23] Knight, K. and Fu, W. (2000). Asymptotics for Lasso-Type Estimators. The Annals of Statistics, 28, 1356–1378. MR1805787

[24] Koul, H., Susarla, V. and Van Ryzin, J. (1981). Regression Analysis with Randomly Right-Censored Data. The Annal of Statistics, 9, 1276–1288. MR0630110

[25] Lai, T.L. and Ying, Z. (1991), Large Sample Theory of a Modified Buckley-James Estimator for Regression Analysis with Censored Data. The Annals of Statistics, 19, 1370–1402. MR1126329

[26] Mantel, N. (1966). Evaluation of Survival Data and Two New Rank Order Statistics Arising in its Considerations. Cancer Chemo. Rep., 50, 163–170.

[27] Miller, R.G. and Halpern, J. (1982). Regression with Censored Data. Biometrika,

69, 521–531. MR0695199

[28] Osborne, M.R., Presnell, B. and Turlach, B.A. (2000). On the Lasso and its Dual. Journal of Computational and Graphical Statistics, 9, 319–337. MR1822089

[29] Prentice, R.L. (1978), Linear Rank Tests with Right-Censored Data. Biometrika, 65, 167–179. MR0497517

[30] Reid, N. (1994). A Conversation with Sir David Cox. Statistical Science, 9, 439–455. MR1325436

[31] Ritov, Y. (1990). Estimation in a Linear Regression Model with Censored Data. The Annals of Statistics, 18, 303–328. MR1041395

[32] Rubin, D. and van der Lann, M.J. (2007). A Doubly Robust Censoring Unbiased Transformation. The International Journal of Biostatistics, 3, 2007. MR2306842

[33] Schmid, M. and Hothorn, T. (2008). Flexible Boosting of Accelerated Failure Time Models. Technical Report 018-2008, Department of Statistics, University of Munich.

[34] Tibshirani, R.J. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society, Ser. B, 58, 267–288. MR1379242

[35] Tibshirani, R.J. (1997). The Lasso Method for Variable Selection in the Cox Model. Statistics in Medicine, 16, 385–395.

[36] Tsiatis, A.A. (1990). Estimating Regression Parameters using Linear Rank Tests for Censored Data. The Annals of Statistics, 18, 354–372. MR1041397

[37] Tsiatis, A.A. (2006). Semiparametric theory and missing data. Springer: New York. MR2233926

[38] van der Laan, M.J. and Robins, J.M. (2003). Unified Methods for Censored Longitudinal Data and Causality. Springer: New York. MR1958123

[39] Wang, H. and Leng, C. (2007). Unified LASSO Estimation by Least Squares Approximation. Journal of the American Statistical Association, 102, 1039–1048. MR2411663

[40] Wang, H., Li, G. and Jiang, G. (2007). Robust Regression Shrinkage and Consistent Variable Selection through the Lad-Lasso. Journal of Business and Economic Statistics, 11, 1–6.

[41] Wang, S., Nan, B., Zhu, J. and Beer, D.G. (2008). Doubly Penalized Buckley-James Method for Survival Data with High-Dimensional Covariates. Biometrics, 64, 132–140.

[42] Wang, N. and Robins, J.M. (1999). Large-sample Theory for Parametric Multiple Imputation Procedures. Biometrika, 85, 935–948. MR1666715

[43] Zeng, D. and Lin, DY. (2007). Efficient Estimation for the Accelerated Failure Time Model. Journal of the American Statistical Association, 102, 1387–1396. MR2412556

[44] Zhang, H.H. and Lu, W. (2007). Adaptive-Lasso for Cox's Proportional Hazards Model. Biometrika, 94, 691–703. MR2410017

[45] Zou, H. (2006). The Adaptive Lasso and its Oracle Properties. Journal of the American Statistical Association, 101, 1418–1429. MR2279469

[46] Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society, Ser. B, 67, 301–320. MR2137327

Electronic Journal of Statistics. ISSN: 1935-7524