Lee, Jason D.; Sun, Dennis L.; Sun, Yuekai; Taylor, Jonathan E.

Exact post-selection inference, with application to the Lasso. (English) Zbl 1341.62061

Ann. Stat. 44, No. 3, 907-927 (2016).

Summary: We develop a general approach to valid inference after model selection. At the core of our framework is a result that characterizes the distribution of a post-selection estimator conditioned on the selection event. We specialize the approach to model selection by the lasso to form valid confidence intervals for the selected coefficients and test whether all relevant variables have been included in the model.

MSC:

62F03 Parametric hypothesis testing
62J07 Ridge regression; shrinkage estimators (Lasso)
62E15 Exact distribution theory in statistics

Keywords:

Lasso; confidence interval; hypothesis test; model selection

Software:

covTest

Full Text: DOI Euclid arXiv

References:

[1] Benjamini, Y., Heller, R. and Yekutieli, D. (2009). Selective inference in complex research. Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 367 4255-4271. · Zbl 1185.62125 · doi:10.1098/rsta.2009.0127

[2] Benjamini, Y. and Yekutieli, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. J. Amer. Statist. Assoc. 100 71-93. · Zbl 1117.62302 · doi:10.1198/016214504000001907 · masetto.asa.catchword.org

[3] Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013). Valid post-selection inference. Ann. Statist. 41 802-837. · Zbl 1267.62080 · doi:10.1214/12-AOS1077 · euclid:aos/1369836961 · arxiv:1306.1059

[4] Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels. Biometrika 62 441-444. · Zbl 0309.62014 · doi:10.1093/biomet/62.2.441

[5] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. Ann. Statist. 32 407-499. · Zbl 1091.62054 · doi:10.1214/009053604000000067 · arxiv:math/0406456 · arxiv:math/0406463 · arxiv:math/0406467 · arxiv:math/0406468 · arxiv:math/0406469 · arxiv:math/0406470 · arxiv:math/0406471 · arxiv:math/0406472 · arxiv:math/0406473 · arxiv:math/0406474

[6] Fisher, R. (1956). On a test of significance in Pearson's Biometrika Tables (No. 11). J. Roy. Statist. Soc. Ser. B. 18 56-60. · Zbl 0070.37304

[7] Fithian, W., Sun, D. and Taylor, J. (2014). Optimal inference after model selection. Preprint. Available at . arXiv:1410.2597 · arxiv.org ·

[8] Javanmard, A. and Montanari, A. (2013). Confidence intervals and hypothesis testing for high-dimensional regression. Preprint. Available at . arXiv:1306.3171 · Zbl 1319.62145 · jmlr.csail.mit.edu · arxiv:1306.3171

[9] Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. Ann. Statist. 28 1356-1378. · Zbl 1105.62357 · doi:10.1214/aos/1015957397 ·

[10] Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. Econometric Theory 21 21-59. · Zbl 1085.62004 · doi:10.1017/S0266466605050036

[11] Leeb, H. and Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? Ann. Statist. 34 2554-2591. · Zbl 1106.62029 · doi:10.1214/009053606000000821 · arxiv:math/0702703

[12] Lehmann, E. L. and Romano, J. P. (2005). Testing Statistical Hypotheses , 3rd ed. Springer, New York. · Zbl 1076.62018 · doi:10.1007/0-387-27605-X

[13] Lockhart, R., Taylor, J., Tibshirani, R. and Tibshirani, R. (2014). A significance test for the lasso (with discussion). Ann. Statist. 42 413-468. · Zbl 1305.62254 · doi:10.1214/13-AOS1175 · euclid:aos/1400592161 · arxiv:1301.7161

[14] Miller, A. (2002). Subset Selection in Regression , 2nd ed. Chapman & Hall/CRC, Boca Raton, FL. · Zbl 1051.62060

[15] Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012). A unified framework for high-dimensional analysis of \(M\)-estimators with decomposable regularizers. Statist. Sci. 27 538-557. · Zbl 1331.62350 · doi:10.1214/12-STS400 · euclid:ss/1356098555

[16] Pötscher, B. M. (1991). Effects of model selection on inference. Econometric Theory 7 163-185. · Zbl 04504752 · doi:10.1017/S0266466600004382

[17] Pötscher, B. M. and Schneider, U. (2010). Confidence sets based on penalized maximum likelihood estimators in Gaussian regression. Electron. J. Stat. 4 334-360. · Zbl 1329.62156 · doi:10.1214/09-EJS523 · euclid:ejs/1268655653 · arxiv:0806.1652

[18] Robinson, G. K. (1979). Conditional properties of statistical procedures. Ann. Statist. 7 742-755. · Zbl 0423.62005 · doi:10.1214/aos/1176344725 ·

[19] Sampson, A. R. and Sill, M. W. (2005). Drop-the-losers design: Normal case. Biom. J. 47 257-268. · doi:10.1002/bimj.200410119

[20] Sill, M. W. and Sampson, A. R. (2009). Drop-the-losers design: Binomial case. Comput. Statist. Data Anal. 53 586-595. · Zbl 1301.62115 · doi:10.1016/j.csda.2008.07.031

[21] Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the \(q\)-value. Ann. Statist. 31 2013-2035. · Zbl 1042.62026 · doi:10.1214/aos/1074290335 ·

[22] Taylor, J., Lockhart, R., Tibshirani, R. J. and Tibshirani, R. (2014). Post-selection adaptive inference for least angle regression and the lasso. Preprint. Available at . arXiv:1401.3889 · Zbl 1305.62255 · arxiv.org ·

[23] Tian, X. and Taylor, J. (2015). Asymptotics of selective inference. Preprint. Available at . arXiv:1501.0358 · Zbl 1422.62252 · arxiv.org ·

[24] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc. Ser. B 58 267-288. · Zbl 0850.62538

[25] Tibshirani, R. J. (2013). The lasso problem and uniqueness. Electron. J. Stat. 7 1456-1490. · Zbl 1337.62173 · doi:10.1214/13-EJS815 · euclid:ejs/1369148600 · arxiv:1206.0313

[26] Tibshirani, R. J., Rinaldo, A., Tibshirani, R. and Wasserman, L. (2015). Uniform asymptotic inference and the bootstrap after model selection. Preprint. Available at . arXiv:1506.0626 · Zbl 1392.62210 · arxiv.org ·

[27] van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2013). On asymptotically optimal confidence regions and tests for high-dimensional models. Preprint. Available at . arXiv:1303.0518 · Zbl 1305.62259 · doi:10.1214/14-AOS1221 · euclid:aos/1403276911 · arxiv:1303.0518

[28] Weinstein, A., Fithian, W. and Benjamini, Y. (2013). Selection adjusted confidence intervals with more power to determine the sign. J. Amer. Statist. Assoc. 108 165-176. · Zbl 06158333 · doi:10.1080/01621459.2012.737740

[29] Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. J. R. Stat. Soc. Ser. B. Stat. Methodol. 76 217-242. · doi:10.1111/rssb.12026

[30] Zhong, H. and Prentice, R. L. (2008). Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. Biostatistics 9 621-634.

[31] Zollner, S. and Pritchard, J. K. (2007). Overcoming the winner's curse: Estimating penetrance parameters from case-control data. Am. J. Hum. Genet. 80 605-615.

[32] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B. Stat. Methodol. 67 301-320. · Zbl 1069.62054 · doi:10.1111/j.1467-9868.2005.00503.x