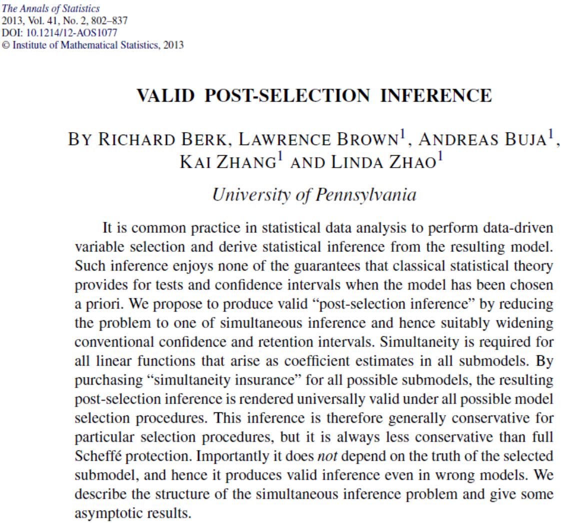
Berk, Brown, Buja, Zhang and Zhao (2013, Annals)

Yasuyuki Matsumura (Kyoto University)

Last Updated: May 3, 2025

https://yasu0704xx.github.io



The Annals of Statistics
2013, Vol. 41, No. 2, 802–837
DOI: 10.1214/12-AOS1077
© Institute of Mathematical Statistics, 2013

VALID POST-SELECTION INFERENCE

BY RICHARD BERK, LAWRENCE BROWN[1], ANDREAS BUJA[1],
KAI ZHANG[1] AND LINDA ZHAO[1]

University of Pennsylvania

It is common practice in statistical data analysis to perform data-driven variable selection and derive statistical inference from the resulting model. Such inference enjoys none of the guarantees that classical statistical theory provides for tests and confidence intervals when the model has been chosen a priori. We propose to produce valid "post-selection inference" by reducing the problem to one of simultaneous inference and hence suitably widening conventional confidence and retention intervals. Simultaneity is required for all linear functions that arise as coefficient estimates in all submodels. By purchasing "simultaneity insurance" for all possible submodels, the resulting post-selection inference is rendered universally valid under all possible model selection procedures. This inference is therefore generally conservative for particular selection procedures, but it is always less conservative than full Scheffé protection. Importantly it does *not* depend on the truth of the selected submodel, and hence it produces valid inference even in wrong models. We describe the structure of the simultaneous inference problem and give some asymptotic results.

This slide is available on
https://github.com/yasu0704xx/ArticleReview.

Contents

Introduction: The Problem with Statistical Inference after Model Selection

Targets of Inference and Assumptions

Estimation and its Targets in Submodels

Universally Valid Post-Selection Confidence Intervals

The Structure of the PoSI Problem

Illustrative Examples and Asymptotic Results

Summary and Discussion

References

References i

- Angrist, J. D. and Pischke, J. S. (2009). *Mostly Harmless Econometrics*. Princeton Univ. Press, Princeton.
- Bahadur, R. R. (1966). A note on quantiles in large samples. *Ann. Math. Statist.* **37**, 577-580. MR0189095
- Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013). Supplement to " Valid post-selection inference. " DOI:10.1214/12-AOS1077SUPP.
- Brown, L. (1967). The conditional level of Student ' s t test. *Ann. Math. Statist.* **38**, 1068-1071. MR0214210

References ii

- Buehler, R. J. and Feddersen, A. P. (1963). Note on a conditional property of Student ’ s t. *Ann. Math. Statist.* **34**, 1098-1100. MR0150864

- Claeskens, G. and Hjort, N. L. (2003). The focused information criterion (with discussion). *J. Amer. Statist. Assoc.* **98**, 900-945. MR2041482

- Dijkstra, T. K. and Veldkamp, J. H. (1988). Data-driven selection of regressors and the bootstrap. In *On Model Uncertainty and Its Statistical Implications* (T. K. Dijkstra, ed.) 17-38. Springer, Berlin.

References iii

- Hall, P. and Carroll, R. J. (1989). Variance function estimation in regression: The effect of estimating the mean. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **51**, 3-14. MR0984989

- Moore, D. S. and McCabe, G. P. (2003). *Introduction to the Practice of Statistics*, 4th ed. Freeman, New York.

- Olshen, R. A. (1973). The conditional level of the F-test. *J. Amer. Statist. Assoc.* **68**, 692-698. MR0359198

- Potscher, B. M. (1991). Effects of model selection on inference. *Econometric Theory* **7**, 163-185. MR1128410

References iv

- Potscher, B. M. (2006). The distribution of model averaging estimators and an impossibility result regarding its estimation. In *Time Series and Related Topics. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **52**, 113 -29. IMS, Beachwood, OH. MR2427842

- Potscher, B. M. and Leeb, H. (2009). On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *J. Multivariate Anal.* **100**, 2065-2082. MR2543087

- Potscher, B. M. and Schneider, U. (2009). On the distribution of the adaptive LASSO estimator. *J. Statist. Plann. Inference* **139**, 2775-2790. MR2523666

References v

Potscher, B. M. and Schneider, U. (2010). Confidence sets based on penalized maximum likelihood estimators in Gaussian regression. *Electron. J. Stat.* **4**, 334-360. MR2645488

Potscher, B. M. and Schneider, U. (2011). Distributional results for thresholding estimators in high-dimensional Gaussian regression models. *Electron. J. Stat.* **5**, 1876-1934. MR2970179

Scheffe, H. (1959). *The Analysis of Variance*. Wiley, New York. MR0116429

Sen, P. K. (1979). Asymptotic properties of maximum likelihood estimators based on conditional specification. *Ann. Statist.* **7**, 1019-1033. MR0536504

References vi

- Sen, P. K. and Saleh, A. K. M. E. (1987). On preliminary test and shrinkage M-estimation in linear models. *Ann. Statist.* **15**, 1580-1592. MR0913575
- Wyner, A. D. (1967). Random packings and coverings of the unit n-sphere. *Bell System Tech. J.* **46**, 2111-2118. MR0223979