Bertsimas, Dimitris; **Li, Michael Lingzhi**
Scalable holistic linear regression. (English) Zbl 07204099
Oper. Res. Lett. 48, No. 3, 203-208 (2020).

Summary: We propose a new scalable algorithm for holistic linear regression building on *D. Bertsimas* and *A. King* [Oper. Res. 64, No. 1, 2–16 (2016; Zbl 1338.90272)]. Specifically, we develop new theory to model significance and multicollinearity as lazy constraints rather than checking the conditions iteratively. The resulting algorithm scales with the number of samples n in the 10,000s, compared to the low 100s in the previous framework. Computational results on real and synthetic datasets show it greatly improves from previous algorithms in accuracy, false detection rate, computational time and scalability.

MSC:

68 Computer science
62 Statistics

Keywords:

holistic linear regression; multicollinearity and significance in linear regression; mixed-integer optimization

Software:

UCI-ml

Full Text: DOI

References:

[1] Bertsimas, D.; Copenhaver, M. S., Characterization of the equivalence of robustification and regularization in linear and matrix regression, European J. Oper. Res., 270, 931-942 (2018) · Zbl 1403.62040

[2] Bertsimas, D.; King, A., An algorithmic approach to linear regression, Oper. Res., 64, 1, 2-16 (2016) · Zbl 1338.90272

[3] Carrizosaa, E.; Olivares-Nadal, A. V.; Ramırez-Cobob, P., Enhancing interpretability by tightening linear regression models (2017)

[4] Chung, S.; Park, Y. W.; Cheong, T., A mathematical programming approach for integrated multiple linear regression subset selection and validation (2017), arXiv preprint arXiv:1712.04543

[5] D. Dua, C. Graff, UCI machine learning repository, 2017.

[6] Eicker, F., Asymptotic normality and consistency of the least squares estimators for families of linear regressions, Ann. Math. Stat., 34, 2, 447-456 (1963) · Zbl 0111.34003

[7] Hocking, R. R., A biometrics invited paper. the analysis and selection of variables in linear regression, Biometrics, 32, 1, 1-49 (1976) · Zbl 0328.62042

[8] Lazaridis, A., A note regarding the condition number: the case of spurious and latent multicollinearity, Qual. Quant., 41, 1, 123-135 (2007)

[9] Mansfield, E. R.; Helms, B. P., Detecting multicollinearity, Amer. Statist., 36, 3a, 158-160 (1982)

[10] O'brien, R. M., A caution regarding rules of thumb for variance inflation factors, Qual. Quant., 41, 5, 673-690 (2007)

[11] R. Tamura, K. Kobayashi, Y. Takano, R. Miyashiro, K. Nakata, T. Matsui, Mixed integer quadratic optimization formulations for eliminating multicollinearity based on variance inflation factor. Optimization Online, Optimization Online, 2016. · Zbl 1421.90093

[12] Tamura, R.; Kobayashi, K.; Takano, Y.; Miyashiro, R.; Nakata, K.; Matsui, T., Best subset selection for eliminating multicollinearity, J. Oper. Res. Soc. Japan, 60, 3, 321-336 (2017) · Zbl 1382.90068

[13] Tibshirani, R., Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B Stat. Methodol., 58, 1, 267-288 (1996) · Zbl 0850.62538