Shi, Pixu; **Zhang, Anru**; **Li, Hongzhe**

Regression analysis for microbiome compositional data. (English) Zbl 1398.62346

Ann. Appl. Stat. 10, No. 2, 1019-1040 (2016).

Summary: One important problem in microbiome analysis is to identify the bacterial taxa that are associated with a response, where the microbiome data are summarized as the composition of the bacterial taxa at different taxonomic levels. This paper considers regression analysis with such compositional data as covariates. In order to satisfy the subcompositional coherence of the results, linear models with a set of linear constraints on the regression coefficients are introduced. Such models allow regression analysis for subcompositions and include the log-contrast model for compositional covariates as a special case. A penalized estimation procedure for estimating the regression coefficients and for selecting variables under the linear constraints is developed. A method is also proposed to obtain debiased estimates of the regression coefficients that are asymptotically unbiased and have a joint asymptotic multivariate normal distribution. This provides valid confidence intervals of the regression coefficients and can be used to obtain the p-values. Simulation results show the validity of the confidence intervals and smaller variances of the debiased estimates when the linear constraints are imposed. The proposed methods are applied to a gut microbiome data set and identify four bacterial genera that are associated with the body mass index after adjusting for the total fat and caloric intakes.

MSC:

62P10 Applications of statistics to biology and medical sciences; meta analysis
62J07 Ridge regression; shrinkage estimators (Lasso)
62H12 Estimation in multivariate analysis
62F25 Parametric tolerance and confidence regions

Keywords:

compositional coherence; coordinate descent method of multipliers; high dimension; log-contrast model; model selection; regularization; asymptotic multivariate normal distribution; confidence intervals

Software:

CVX; MEGAN

Full Text: DOI Euclid

References:

[1] Aitchison, J. (1982). The statistical analysis of compositional data. J. Roy. Statist. Soc. Ser. B 44 139-177. · Zbl 0491.62017

[2] Aitchison, J. (2003). The Statistical Analysis of Compositional Data . Blackburn Press, Cadwell, NJ. · Zbl 0491.62017

[3] Aitchison, J. and Bacon-Shone, J. (1984). Log contrast models for experiments with mixtures. Biometrika 71 323-330.

[4] Bertsekas, D. P. (1996). Constrained Optimization and Lagrange Multiplier Methods . Athena Scientific, Belmont. · Zbl 0572.90067

[5] Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. Bernoulli 19 1212-1242. · Zbl 1273.62173 · doi:10.3150/12-BEJSP11 · euclid:bj/1377612849 · arxiv:1202.1377

[6] Cornell, J. A. (2002). Experiments with Mixtures : Designs , Models , and the Analysis of Mixture Data , 3rd ed. Wiley, New York. · Zbl 1001.62024

[7] Efron, B. (2014). Estimation and accuracy after model selection. J. Amer. Statist. Assoc. 109 991-1007. · Zbl 1368.62071 · doi:10.1080/01621459.2013.823775

[8] Grant, M. and Boyd, S. (2013). CVX: Matlab software for disciplined convex programming, version 2.0 beta. Technical report. Available at . · cvxr.com

[9] Huson, D. H., Auch, A. F., Qi, J. and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. Genome Res. 17 377-386.

[10] James, G. M., Paulson, C. and Rusmevichientong, P. (2015). Penalized and constrained regression. Unpublished manuscript.

[11] Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. J.

Mach. Learn. Res. 15 2869-2909. · Zbl 1319.62145 · jmlr.csail.mit.edu · arxiv:1306.3171

[12] Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J. and Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. PLoS Computational Biolology 11 e1004226.

[13] Lam, Y. Y., Ha, C. W. Y., Campbell, C. R., Mitchell, A. J., Dinudom, A., Oscarsson, J., Cook, D. I., Hunt, N. H., Caterson, I. D., Holmes, A. J. and Storlien, L. H. (2012). Increased gut permeability and microbiota change associate with mesenteric fat inflammation and metabolic dysfunction in diet-induced obese mice. PLoS ONE 7 e34233.

[14] Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. Ann. Statist. 44 907-927. · Zbl 1341.62061 · doi:10.1214/15-AOS1371 · euclid:aos/1460381681

[15] Ley, R. E., Bäckhed, F., Turnbaugh, P., Lozupone, C. A., Knight, R. D. and Gordon, J. I. (2005). Obesity alters gut microbial ecology. Proc. Natl. Acad. Sci. USA 102 11070-11075.

[16] Ley, R. E., Turnbaugh, P. J., Klein, S. and Gordon, J. I. (2006). Microbial ecology: Human gut microbes associated with obesity. Nature 444 1022-1023.

[17] Lin, W., Shi, P., Feng, R. and Li, H. (2014). Variable selection in regression with compositional covariates. Biometrika 101 785-797. · Zbl 1306.62164 · doi:10.1093/biomet/asu031

[18] Manichanh, C., Borruel, N., Casellas, F. and Guarner, F. (2012). The gut microbiota in IBD. Nat. Rev. Gastroenterol. Hepatol. 9 599-608.

[19] Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T. et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464 59-65.

[20] Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D. et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490 55-60.

[21] Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O. and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. Nat. Methods 9 811-814.

[22] Shi, P., Zhang, A. and Li (2016). Supplement to "Regression analysis for microbiome compositional data." . · Zbl 1398.62346 · doi:10.1214/16-AOAS928 · · dx.doi.org

[23] Snee, R. D. (1973). Techniques for the analysis of mixture data. Technometrics 15 517-528.

[24] Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. Biometrika 99 879-898. · Zbl 1452.62515 · doi:10.1093/biomet/ass043

[25] Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R. and Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444 1027-1031.

[26] Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R. and Gordon, J. I. (2007). The human microbiome project. Nature 449 804-810.

[27] van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Statist. 42 1166-1202. · Zbl 1305.62259 · doi:10.1214/14-AOS1221 · euclid:aos/1403276911 · arxiv:1303.0518

[28] Walker, A. W., Ince, J., Duncan, S. H., Webster, L. M., Holtrop, G., Ze, X., Brown, D., Stares, M. D., Scott, P., Bergerat, A., Louis, P., McIntosh, F., Johnstone, A. M., Lobley, G. E., Parkhill, J. and Flint, H. J. (2011). Dominant and diet-responsive groups of bacteria within the human colonic microbiota. ISME J. 5 220-230.

[29] Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., Sinha, R., Gilroy, E., Gupta, K., Baldassano, R., Nessel, L., Li, H., Bushman, F. D. and Lewis, J. D. (2011). Linking long-term dietary patterns with gut microbial enterotypes. Science 334 105-108.

[30] 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