Секция «Теория вероятностей и математическая статистика»

Выбор методов регрессионного анализа при наличии взаимосвязей между предикторами модели.

Научный руководитель – Яровая Елена Борисовна

Куценко Владимир Александрович
Студент (специалист)
Московский государственный университет имени М.В.Ломоносова,
Механико-математический факультет, Кафедра теории вероятностей, Москва, Россия
E-mail: vlakutsenko@ya.ru

Основная проблема применения регрессионного анализа - наличие взаимосвязей между предикторами модели. В тезисах мы остановимся на изучении регрессионных моделей вида $Y{=}X^T\beta{+}\epsilon$ при наличии выраженных линейных взаимосвязей между предикторами и проиллюстрируем их применение в медико-биологических исследованиях. В таких случаях говорится, что модель множественной регрессии применяется в условиях мультиколлинеарности. Напомним, что коллинеарность - это наличие линейной зависимости между предикторами модели, мультиколлинеарность - наличие высокой корреляции между ними. В случае коллинеарности матрица регрессоров (предикторов модели) X имеет ранг меньше, чем её размерность, что приводит к невозможности обращения матрицы XX^T и, как следствие, оценка наименьших квадратов для модели не существует. В случае мультиколлинеарности матрица XX^T обратима, но алгоритмически невычислима, или же вычислима, но чувствительна к небольшим вариациям данных.

О наличии проблемы мультиколлинеарности в выборке можно узнать, рассмотрев фактор инфляции дисперсии *(VIF),* см., напр., [1], суть которого - построение линейной регрессии между выбранным регрессором и остальными, исключая отклик. Для полученной модели рассчитывается коэффициент детерминации R^2, и, если он достаточно большой, то считается, что выбранный регрессор зависим от остальных. В таком случае исследователю приходится выбирать из большого количества методов, разработанных для решения проблемы мультиколлинеарности.

В работе рассматриваются основные существующие методы: алгоритмы последовательного добавления или удаления признаков, среди которых регрессионный метод наименьших углов *(LARS)* [2], ступенчатая регрессия [3], методы ортогонализации [4]; алгоритмы с введением ограничения на функционал правдоподобия: гребневая [5] и мостовая [6] регрессии, лассо-регрессия [7], эластичная сеть [8], алгоритмы перебора моделей: методы группового учёта, генетический алгоритм, полный перебор, алгоритмы добавления и удаления признаков (пошаговые регрессионные модели). Для каждого метода в докладе будет представлено его краткое теоретическое описание, особенности применения, и практическая реализация, а также будет произведено сравнение методов. Особое внимание уделено сравнению моделей регуляризации.

В результате сравнений были получены следующие результаты: у алгоритмов последовательного добавления есть общий недостаток - сходимость к локально наилучшему набору параметров, а не к глобально наилучшему, в отличие от алгоритмов добавления и удаления признаков, которые позволяют тестировать уже добавленные признаки. Для моделей регуляризации были получены следующие результаты: если размеры эффектов небольшие, но ненулевые, то рекомендуется использовать гребневую регрессию, если же считается, что размеры эффектов большие, то следует использовать регрессию с лассо, так как она, в отличие от гребневой регрессии, обладает эффектом обнуления параметров.

В случае, если исследователь не уверен, какое из предположений принимать, следует использовать эластичную сеть или мостовую регрессию, помня о том, что эластичная сеть, в отличие от мостовой регрессии обладает эффектом обнуления параметров.

На имеющихся данных, полученных в лаборатории молекулярной генетики ФГБУ «Государственный научно-исследовательский центр профилактической медицины» МЗ РФ были протестированы все обозреваемые методы и объяснён выбор наилучшей модели.

Источники и литература

1) Стрижов В.В., Крымова Е.А. Методы выбора регрессионных моделей. Вычислительный центр РАН, Москва, 2010.
2) Efron B., Hastie T., Johnstone I., Tibshirani R. Least angle regression // The Annals of Statistics. 2004. Vol. 32, no. 3. Pp. 407–499.
3) Draper N. R., Smith H. Appied Regression Analysis. John Wihley and Sons, 1998.
4) Chen Y. W., Billings C. A., Luo W. Orthogonal least squares methods and their application to non-linear system identification // International Journal of Control. 1989. Vol. 2, no. 50. Pp. 873–896.
5) Hoerl, A. and Kennard, R. (1988) Ridge regression. In Encyclopedia of Statistical Sciences, vol. 8, pp. 129–136.New York: Wiley.
6) Wenjiang J. Fu. Penalized Regressions: The Bridge Versus the Lasso, 1998 Journal of Computational and Graphical Statistics, Volume 7, Number 3, Pages 397–416
7) Tibshirani, R. Regression shrinkage and selection via the lasso. J. R. Statist. Soc. B (1996), 58, 267–288.
8) Zou H. and Hastie T. Regularization and variable selection via the elastic net. J. R. Statist. Soc. B (2005) 67 , Part 2 , pp. 301–320