УДК 518.852

С.И. Носков

Иркутский государственный университет путей сообщения, г. Иркутск, Российская Федерация

О СООТВЕТСТВИИ НАБЛЮДЕНИЙ ВЫБОРКИ СПЕЦИФИКАЦИИ РЕГРЕССИОННОЙ МОДЕЛИ

Аннотация. *В работе предложен алгоритмический способ выявления соответствия наблюдений выборки данных спецификации регрессионной модели, а именно, ее форме и составу независимых переменных. Уровень этого соответствия предлагается оценивать, в частности, местом каждого наблюдения в их упорядочении по возрастанию суммарных по модулю ошибок аппроксимации при использовании нескольких альтернативных методов идентификации неизвестных параметров.*

Ключевые слова: *регрессионная модель, идентификация параметров, методы наименьших квадратов, модулей, антиробастного и смешанного оценивания, ошибки аппроксимации.*

S.I. Noskov

Irkutsk state University of railway engineering, Russian Federation.

ON THE COMPLIANCE OF SAMPLE OBSERVATIONS WITH THE SPECIFICATIONS OF THE REGRESSION MODEL

Annotation. *The paper proposes an algorithmic method for identifying the compliance of observations in a sample of data with the specification of a regression model, namely, its form and composition of independent variables. The level of this correspondence is proposed to be assessed, in particular, by the place of each observation in their ordering in increasing order of the total modulus of approximation errors when using several alternative methods for identifying unknown parameters.*

Key words: *regression model, methods for identifying parameters, methods of least squares, moduli, antirobust and mixed estimation, approximation errors.*

Построение математических, в частности, регрессионных моделей различных объектов часто сопровождается анализом свойств исходной информации. Так, в работе [1] изучаются конечные выборочные свойства общих регуляризованных статистических критериев при наличии псевдонаблюдений. В условиях ограниченной сильной выпуклости нештрафной функции потерь и условий регулярности штрафа получены неасимптотические границы погрешности регуляризованной М-оценки. В [2] предлагается полностью инкрементальный метод проецируемой разделительной кластеризации для потоков данных большой размерности, основанный на кластеризации высокой плотности. Метод способен идентифицировать кластеры в произвольных подпространствах, оценивать количество кластеров и обнаруживать изменения в распределении данных, которые вызывают необходимость пересмотра модели. Эмпирическая оценка предлагаемого метода на многочисленных реальных и смоделированных наборах данных показывает, что он масштабируем по размеру и количеству кластеров, устойчив к нерелевантным функциям и способен справляться с различными типами нестационарности. В статье [3] рассмотрены пять понятий глубины данных. Они в основном предназначены для функциональных данных, но их также можно адаптировать и к стандартному многомерному случаю. Эффективность этих понятий глубины, когда они используются в качестве

вспомогательных инструментов при оценке и классификации, проверяется с помощью метода Монте-Карло. Работа [4] посвящена проблеме загрязненных данных (например, данных с выбросами или конфликтами) с конкретным применением к проблеме оценки достоверного интервала для среднего значения генеральной совокупности. Выяснилось, что нормальная модель «по умолчанию», используемая в большинстве случаев анализа байесовских данных, не является надежной, и что подходы, основанные на байесовском бутстрапе, надежны только в ограниченных обстоятельствах. Простая параметрическая модель, основанная на «модели загрязненного нормального состояния» Тьюки, и модель, основанная на t-распределении, оказались заметно более надежными. В [5] исследуются числовые характеристики трех модифицированных версий метода эмпирического правдоподобия: скорректированного эмпирического правдоподобия, преобразованного эмпирического правдоподобия и преобразованного скорректированного эмпирического правдоподобия для данных с различными размерами выборки и различными пропорциями нулевых значений в ней. В работе [6] анализируются данные, связанные с оценкой уровня пожарной опасности объектов и территорий.

Весьма интересным аспектом при построении регрессионной модели сложного объекта является выявление соответствия каждого наблюдения исходной выборки данных спецификации этой модели, а именно, ее форме и составу независимых переменных. При этом надо иметь в виду, что это соответствие сильно зависит от метода оценивания модельных параметров и от его реакции на выбросы – наблюдения, слабо согласованные со всей выборкой в целом. Какие-то методы обладают робастностью и не чувствительны к выбросам, какие-то, напротив к ним тяготеют (см., например, [7 – 10]).

Рассмотрим линейную регрессионную модель некоторого объекта:

$$y_k = \sum_{i=1}^{m} \alpha_i^s x_{ki} + \varepsilon_k^s,\ k = \overline{1,n},\ s = \overline{1,S}, \qquad (1)$$

где y − зависимая, а x_i − i-ая независимая переменные, α_i − i-ый подлежащий оцениванию параметр, ε_k − ошибки аппроксимации, k − номер наблюдения, n − длина выборки данных, s - номер метода оценивания параметров модели, S – количество используемых методов. Будем считать все переменные и ошибки модели (1) детерминированными.

Рассчитаем суммарную по модулю ошибку θ_k, $k = \overline{1,n}$ каждого наблюдения выборки при использовании всех методов оценивания:

$$\theta_k = \sum_{s=1}^{S} |\varepsilon_k^s|.$$

Введем в рассмотрение вектор $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ номеров наблюдений в их упорядочении по возрастанию суммарных по модулю ошибок θ_k, $k = \overline{1,n}$. Таким образом, если, например, σ_k=4, то это означает, что k – ое наблюдение выборки находится на четвертом месте в указанном упорядочении.

В качестве примера рассмотрим задачу построения регрессионной модели эффективности интеллектуальной деятельности четырьмя методами оценивания параметров – наименьших квадратов (МНК), наименьших модулей (МНМ), антиробастного оценивания (МАО) и смешанного оценивания (МСО) [7 – 10]. В качестве информационной базы модели используем данные из работы [11] за 2010 – 2020 г. г., всего 11 наблюдений.

Введем следующие обозначения:

y - число отечественных патентных заявок на промышленные образцы, шт.;

x_1 - количество персональных компьютеров в организациях, тыс. шт.;

x_2 - используемые передовые производственные технологии, ед.;

x_3 - внутренние текущие затраты на фундаментальные исследования, млн. руб.

С помощью указанных выше методов оценим параметры линейной регрессионной модели:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \varepsilon. \qquad (2)$$

В табл. 1 приведены соответствующие оценки.

Таблица 1

Оценки параметров модели (2)

Метод	α_0	α_1	α_2	α_3
МНК	-1544.460	0.017	0.010	0.012
МНМ	-244.618	-0.135	0.006	0.022
МАО	-3554.871	0.021	0.022	0.007
МСО	-15.110	-0.143	0.007	0.020

В табл. 2 приведены ошибки аппроксимации для каждого варианта модели (2).

Таблица 2

Ошибки аппроксимации для каждого наблюдения выборки

Метод	1	2	3	4	5	6	7	8	9	10	11
МНК	174.9	78.28	67.08	-74.92	-80.6	-421.9	-202.87	480.83	-59.14	-148.59	186.95
МНМ	147.6	0.000	102.20	0.000	-72.3	-336.5	0.000	680.32	0.000	-93.69	229.08
МАО	246.1	343.78	338.60	200.22	150.65	-343.5	-294.23	343.78	-225.54	-343.78	343.78
МСО	64.2	-47.44	64.28	-20.49	-64.2	-331.8	0.000	696.43	68.66	0.000	400.89

Наконец, в табл. 3 приведено место σ_k каждого наблюдения в их упорядочении по возрастанию суммарных по модулю ошибок.

Таблица 3

Места наблюдений в их упорядочении по возрастанию суммарных по модулю ошибок

Наблюдение	1	2	3	4	5	6	7	8	9	10	11
Место	8	4	6	1	3	10	5	11	2	7	9

Таким образом, наиболее полным соответствием модели (2) обладает четвертое наблюдение, наименее – восьмое. Вся содержащаяся в табл. 3 информация может служить

основанием для проведения дальнейших исследований, связанных с моделированием эффективности интеллектуальной деятельности.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Poignard B., Fermanian J.D. The finite sample properties of sparse M-estimators with pseudo-observations // Annals of the Institute of Statistical Mathematics. – 2022. – V. 74. – P. 1–31.
2. Hofmeyr D.P., Pavlidis N.G., Eckley I.A. Statistics and Computing. – 2016. – V. 26. - P. 1101–1120.
3. Cuevas A., Febrero M., Fraiman R. Robust estimation and classification for functional data via projection-based depth notions // Computational Statistics. – 2007. – V. 22. - P. 481–496.
4. Kennedy L.A., Navarro D.J., Perfors A., Briggs N. Not every credible interval is credible: Evaluating robustness in the presence of contamination in Bayesian data analysis // Behavior Research Methods. – 2017. – V. 49. – P. 2219–2234.
5. Stewart P., Ning W. Modified empirical likelihood-based confidence intervals for data containing many zero observations // Computational Statistics. – 2020. – V. 35. – P. 2019–2042.
6. Носков С.И., Удилов В.П. Управление системой обеспечения пожарной безопасности на региональном уровне. - Иркутск: ВСИ МВД России. - 2003. - 151с.
7. Носков С. И. Компромиссные паретовские оценки параметров линейной регрессии // Математическое моделирование. - 2020. - Т. 32. - № 11. - С.70-78.
8. Носков С.И. Метод антиробастного оценивания параметров линейной регрессии: число максимальных по модулю ошибок аппроксимации // Южно-Сибирский научный вестник. - 2020. - № 1 (29). - С. 51-54.
9. Носков С. И. Метод смешанного оценивания параметров линейной регрессии: особенности применения // Вестник ВГУ. Серия: Системный анализ и информационные технологии. - 2021. - № 1. - С. 126-132.
10. Носков С. И. Выбор метода оценивания параметров линейной регрессии на основе выявления аномальных наблюдений // Вестник Воронежского государственного технического университета. - 2021. - т. 17. - № 2. - С. 24-29.
11. Пашков Д.В, Носков С.И. Реализация конкурса регрессионных моделей эффективности интеллектуальной деятельности // Электронный сетевой политематический журнал "Научные труды КубГТУ". – 2022. - № 6. - С. 40-51.

Информация об авторе

Сергей Иванович Носков – д.т.н., профессор, профессор кафедры «Информационные системы и защита информации», Иркутский государственный университет путей сообщения, г. Иркутск, e-mail: sergey.noskov.57@mail.ru

Author

Sergey Ivanovich Noskov, Doctor of Technical Science, Professor, the Subdepartment Information systems and information security, Irkutsk State Transport University, Irkutsk, e-mail: sergey.noskov.57@mail.ru

Для цитирования

Носков С.И. О соответствии наблюдений выборки спецификации регрессионной модели // «Информационные технологии и математическое моделирование в управлении сложными системами»: электрон. науч. журн. – 2023. – №.3 – С.– Режим доступа:

http://ismm-irgups.ru/toma/ , свободный. – Загл. с экрана. – Яз. рус., англ. (дата обращения)

For citation

Noskov S.I. O sootvetstvii nablyudenij vyborki specifikacii regressionnoj modeli [On the compliance of sample observations with the specifications of the regression model] // Informacionnye tehnologii i matematicheskoe modelirovanie v upravlenii slozhnymi sistemami: ehlektronnyj nauchnyj zhurnal [Information technology and mathematical modeling in the control of complex systems: electronic scientific journal], 2023. No.3. P. . [Accessed]