РЕАЛИЗАЦИЯ ВЗВЕШЕННОГО МЕТОДА НАИМЕНЬШИХ КВАДРАТОВ С ИСПОЛЬЗОВАНИЕМ МЕР СХОДСТВА

Носков С.И.[1], Вергасов А.С.[2]

[1]Носков Сергей Иванович – доктор технических наук, профессор;
[2]Вергасов Александр Сергеевич – соискатель,
кафедра информационных систем и защиты информации,
Иркутский государственный университет путей сообщения,
г. Иркутск

Аннотация: *в статье рассматривается подход к формировании матрицы весов наблюдений при использовании взвешенного метода наименьших квадратов в регрессионном анализе, основанный на применении элементов разработанной профессором Ю.А.Ворониным теории сходства. Рассматривается десять возможных мер сходства, задаваемых простыми арифметическими выражениями. Предлагаемый подход призван существенно повысить прогностические возможности регрессионных моделей по сравнению с традиционными методами идентификации неизвестных модельных параметров. Он также может быть применен при использовании других методов оценивания.*

Ключевые слова: *регрессионный анализ, веса наблюдений, меры сходства.*

IMPLEMENTATION OF THE SUSTAINABLE METHOD OF THE LEAST SQUARES USING ACCOUNTABILITY MEASURES

Noskov S.I.[1], Vergasov A.S.[2]

[1]Noskov Sergey Ivanovich - Doctor of Engineering Science, Professor;
[2]Vergasov Alexander Sergeevich – Job Seeker,
DEPARTMENT OF INFORMATION SYSTEMS AND INFORMATION PROTECTION,
IRKUTSK STATE TRANSPORT UNIVERSITY,
IRKUTSK

Abstract*: the article discusses the approach to the formation of the matrix of weights of observations using the weighted least squares method in regression analysis, based on the application of elements of the theory of similarity developed by Professor Yu. We consider ten possible measures of similarity given by simple arithmetic expressions. The proposed approach is designed to significantly increase the predictive capabilities of regression models compared to traditional methods for identifying unknown model parameters. Also it can be applied when using other estimation methods.*

Keywords: *regression analysis, weights of observations, measures of similarity.*

УДК 519.852

Рассмотрим линейное регрессионное уравнение

$$y_k = \sum_{i=1}^{n} \beta_i x_{ki} + \varepsilon_k, k = \overline{1,d}, \qquad (1)$$

где, y_k и x_{ki} – k-ые значения соответственно выходной и i-ой входной переменных, $\beta = (\beta_1, \dots, \beta_d)^T$ – вектор подлежащих оцениванию параметров, ε_k – ошибки аппроксимации, d – количество наблюдений выборки.

Представим уравнение (1) в векторной форме:

$$y = X\beta + \varepsilon, \qquad (2)$$

где $y = (y_1, \dots, y_d)^T$, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_d)$, $\mathrm{X} = \|x_{ki}\|, k = \overline{1,d}, i = \overline{1,n}$.

Методам оценивания неизвестных параметров уравнения (1) и критериям его верификации посвящена обширная литература (см.. например, [1-18]).

Основное направление практического использования регрессионных моделей, составным элементом которых является уравнение (1), является прогнозирование значений выходных переменных при известных значениях входных. При этом необходимо иметь в виду следующее обстоятельство. Часто исследуемый на модельном уровне объект имеет динамический характер, что предопределяет различие в информационной значимости наблюдений выборки. Для таких ситуаций вместо традиционных методов оценивания параметров – наименьших квадратов, модулей, робастных и антиробастных процедур, - более целесообразно использовать их «взвешенные» модификации, например, взвешенный метод наименьших квадратов (ВМНК), расчетная формула которого имеет вид:

$$\beta = (X^T W X)^{-1} X^T W y, \qquad (3)$$

где $\mathrm{W} = \mathrm{diag}(\omega_k)$, $k = \overline{1,d}$, $\omega_k > 0$ – веса наблюдений выборки.

Отметим, что каких-либо формальных правил назначения таких весов, строго теоретически обоснованных, не существует, а используются какие-либо эвристические содержательно обоснованные приемы. Для регрессионных моделей динамических объектов, в которых индекс k представляет собой время (например, номер года), эти веса обычно задают в виде ω_k=g(k), где g – монотонно возрастающая функция. В тривиальном случае g(k)=k.

Представляется, что такой подход обладает одним весьма существенным недостатком – далеко неочевидно, что более ранние наблюдения выборки данных автоматически обладают заведомо меньшей значимостью по сравнению с более поздними. Ведь текущие тенденции функционирования объекта необязательно будут соответствовать наблюдениям с близкими к d номерами.

Гораздо более оправданным, по-видимому, является подход, основанный на постулате – чем ближе в некотором заданном смысле вектор значений входных переменных прогнозного периода к соответствующему наблюдению периода основания прогноза (то есть самой выборки), тем большим весом это наблюдение должно обладать, а, значит, тем выше должен быть его вес ω_k в (3).

Мера оценки указанной близости может быть основана на разработанной Ю.А.Ворониным теории сходства (см., например,[19]). В [19] приведены десять возможных мер сходства. Разберемся с их формальным представлением.

Пусть для s некоторых объектов задана матрица H характеризующих их n признаков:

$$\mathrm{H}= \|h_{ki}\|, k = \overline{1,s}, i = \overline{1,n}.$$

Введем обозначения:

$h_i^- = \min_k h_{ki}$, $h_i^+ = \max_k h_{ki}$.

Для каждого a-го объекта рассчитаем значения:

$$f_i^a = (\mathrm{h}_{ai} - h_i^-) / (h_i^+ - h_i^-), i = \overline{1,n}.$$

Очевидно, что для всех a и i справедливы неравенства

$$0 \le f_i^a \le 1, a = \overline{1,s}, i = \overline{1,n}.$$

Тогда аналитические выражения для мер сходства между произвольными объектами k u l имеют вид [19]:

1) $1 - \sum_{i=1}^n \alpha_i \left| f_i^k - f_i^l \right|,\ \sum_{i=1}^n \alpha_i = 1,\ \alpha_i \ge 0.$

2) $\frac{1-\sqrt{\sum_{i=1}^n \alpha_i^2 (f_i^k - f_i^l)^2}}{\sqrt{\sum_{i=1}^n \alpha_i}}.$

3) $1 - \max_i \left| f_i^k - f_i^l \right|.$

4) $\sum_{i=1}^n \alpha_i \frac{\min(f_i^k, f_i^l)}{\max(f_i^k, f_i^l)},\ \sum_{i=1}^n \alpha_i = 1,\ \alpha_i \ge 0.$

5) $\frac{1}{1+\sum_{i=1}^n |f_i^k - f_i^l|}.$

6) $1 - \frac{\sum_{i=1}^n |(f_i^k - f_i^l)| + |\sum_{i=1}^n (f_i^k - f_i^l)|}{2}.$

7) $\frac{\sum_{i=1}^n (f_i^k f_i^l)}{(\sum_{i=1}^n (f_i^k)^2)^{\frac{1}{2}} (\sum_{i=1}^n (f_i^l)^2)^{\frac{1}{2}}}.$

8) $1 - e^{-\left(\sum_{i=1}^n \left(f_i^k - f_i^l\right)^2\right)^{\frac{1}{2}}}.$

9) $\sum_{i=1}^n \alpha_i \left(1 - \left|f_i^k - f_i^l\right|\right) * \frac{\sum_{i=1}^n (f_i^k f_i^l)}{\left(\sum_{i=1}^n (f_i^k)^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^n (f_i^l)^2\right)^{\frac{1}{2}}}.$

10) $\sum_{i=1}^{n_1} \alpha_i \left(1 - \left|f_i^k - f_i^l\right|\right) * \prod_{i=n_1}^{n} \left(1 - \left|f_i^k - f_i^l\right|\right), \alpha_i > 0.$

Здесь α_i - весовые коэффициенты признаков, которые в простейшем случае могут быть приняты равными, например, 1/n.

Эти-то выражения и предлагается использовать для расчета весов наблюдений при реализации ВМНК. Описанию механизма такого использования авторы намерены посвятить следующую публикацию.

Список литературы / References

1. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ. М.: Финансы и статистика, 1981.Т. 1. 366 с., Т. 2. 351с.
2. *Носков С.И.* Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных. Иркутск: Облинформпечать, 1996. 320 с.

3. *Носков С.И.* Идентификация параметров кусочно-линейной функции риска. Транспортная инфраструктура Сибирского региона, 2017. Т. 1. С. 417-421.
4. *Иванова Н.К., Носков С.И.* Организация прогнозных расчетов по регрессионным моделям // Информационные технологии и проблемы математического моделирования сложных систем, 2017. № 18. С. 78-80.
5. *Носков С.И., Баенхаева А.В.* Множественное оценивание параметров линейного регрессионого уравнения // Современные технологии. Системный анализ. Моделирование, 2016. № 3 (51). С. 133-138.
6. *Носков С.И., Быкова О.В., Некипелова О.Е., Соколова Л.Е.* Возможный способ поиска компромиссного решения в задаче линейного программирования с векторной целевой функцией // Фундаментальные исследования, 2014. № 6-3. С. 502-505.
7. *Носков С.И.* Критерий «согласованность поведения» в регрессионном анализе // Современные технологии. Системный анализ. Моделирование, 2013. № 1 (37). С. 107-110.
8. *Носков С.И.* Оценивание параметров аппроксимирующей функции с постоянными пропорциями // Современные технологии. Системный анализ. Моделирование, 2013. № 2 (38). С. 135-136.
9. *Лакеев А.В., Носков С.И.* Метод наименьших модулей для линейной регрессии: число нулевых ошибок аппроксимации // Современные технологии. Системный анализ. Моделирование, 2012. № 2 (34). С. 48-50.
10. *Носков С.И.* Проблема единственности парето-оптимального решения в задаче линейного программирования с векторной целевой функцией // Современные технологии. Системный анализ. Моделирование, 2011. № S-4 (32). С. 283-285.
11. *Базилевский М.П., Носков С.И*. Анализ систем программирования для решения вычислительной задачи проведения «конкурса» регрессионных моделей // Информационные технологии и проблемы математического моделирования сложных систем, 2011. № 9. С. 47-51.
12. *Носков С.И.* Точечная характеризация множества парето в линейной многокритериальной задаче // Современные технологии. Системный анализ. Моделирование, 2008. № 1 (17). С. 99-101.
13. *Носков С.И., Зырянов С.И.* Применение критерия смещения при построении регрессионных уравнений // Современные технологии. Системный анализ. Моделирование, 2004. № 2. С. 93.
14. *Носков С.И.* L-множество в многокритериальной задаче оценивания параметров регрессионных уравнений // Информационные технологии и проблемы математического моделирования сложных систем, 2004. № 1. С. 64.
15. *Носков С.И.* Построение эконометрических зависимостей с учетом критерия «согласованность поведения» // Кибернетика и системный анализ, 1994. № 1. С. 177.
16. *Головченко В.Б., Носков С.И.* Выбор класса линейной по параметрам регрессии на основе экспертных высказываний // Кибернетика и системный анализ, 1992. № 5. С. 109.
17. *Носков С.И., Потороченко Н.А.* Диалоговая система реализации «конкурса» регрессионных зависимостей // Управляющие системы и машины, 1992. № 2-4. С. 111.
18. *Golovchenko V.B., Noskov S.I.* Estimation of an econometric model using statistical data and expert information // Automation and Remote Control, 1991. V. 52. № 4. P.542-548.
19. *Воронин Ю.А.* Начала теории сходства. Новосибирск: ВЦ СО АН СССР, 1989. 120 с.