*УДК 519.237.5*

ВЗВЕШЕННЫЙ МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

*Д.Ч. Абдильдинова, О.Н. Канева, Е.С. Дубейко*

Федеральное учреждение Омский государственный технический университет, Россия г.Омск

*В процессе исследования был изучен взвешенный метод наименьших квадратов, разработано и реализовано программное обеспечение. Проведены численные эксперименты.*

*Ключевые слова: аппроксимация, интерполяция, уравнение регрессии, метод наименьших квадратов, взвешенный метод наименьших квадратов, гетероскедастичность*

Развитие новых технологий, применение новых наблюдательных методик и компьютеризация способствуют повышению точности измерений. Для реализации точности вычислений методы обработки данных также должны постоянно совершенствоваться.

Метод наименьших квадратов (МНК)  — математический метод, применяемый для решения различных задач, основанный на минимизации суммы квадратов отклонений аппроксимирующих функций от искомых переменных. МНК является одним из базовых методов [регрессионного анализа](https://ru.wikipedia.org/wiki/%D0%A0%D0%B5%D0%B3%D1%80%D0%B5%D1%81%D1%81%D0%B8%D0%BE%D0%BD%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7) и используется для оценки неизвестных параметров регрессионных моделей по выборочным данным.

Обычно в качестве регрессионного уравнения используется функция линейная относительно своих параметров[1]. Общий вид такой функции:

$\hat{Y}=f\left(b\_{1},b\_{2},…, b\_{k},X\right)=b\_{1}φ\_{1}\left(X\right)+b\_{2}φ\_{2}\left(X\right)+…+b\_{k}φ\_{k}\left(X\right),$ (1)

где $φ\_{i}-$ заданные функции от факторов (i=$\overbar{1,k}$), X=(t, $X\_{1 }, X\_{2 },…, X\_{m }$).

Критерием МНК является сумма квадратов ошибок, которая в данном случае выглядит следующим образом:

SSE= $\sum\_{i=1}^{n}(y\_{i}-(b\_{1}φ\_{1}\left(x\_{i}\right)+b\_{2}φ\_{2}\left(x\_{i}\right)+…+b\_{k}φ\_{k}\left(x\_{i}\right)))^{2}=e^{T}e\rightarrow min$ (2)

где $x\_{i}$=($t\_{i}$, $x\_{i1}, x\_{i2}, …, x\_{im} $), (i=$\overbar{1,n}$); e $–$ вектор ошибок.

Введем векторно-матричное обозначение:

$A=\left[φ\_{1}\left(X\right),φ\_{2}\left(X\right),…,φ\_{k}\left(X\right)\right]=\left[\begin{matrix}\begin{matrix}φ\_{1}\left(x\_{1}\right)&φ\_{2}\left(x\_{1}\right)\\φ\_{1}\left(x\_{2}\right)&φ\_{2}\left(x\_{2}\right)\end{matrix}&…&φ\_{k}\left(x\_{1}\right)\\… …&…&φ\_{k}\left(x\_{2}\right)\\φ\_{1}\left(x\_{n}\right) φ\_{2}\left(x\_{n}\right)&…&φ\_{k}\left(x\_{n}\right)\end{matrix}\right]$ (3)

$Y=\left[\begin{array}{c}y\_{1}\\y\_{2}\\…\\y\_{n}\end{array}\right] $, $B=\left[\begin{array}{c}b\_{1}\\b\_{2}\\…\\b\_{k}\end{array}\right]$,

где y - вектор-столбец наблюдений объясняемой переменной.

Тогда вектор оценок параметров регрессионного уравнения находятся по формуле:

B=$(A'A)^{-1}A'Y$, (4)

где $A^{'}-транспонированная матрица, A^{-1}- $обратная.

Для того чтобы вектор оценок B полученный с помощью МНК являлся состоятельным, несмещенным и эффективным должен выполнятся ряд условий, называемых условиями выполнимости метода наименьших квадратов. Одним из таких условий является условие гомоскедастичности данных, т.е. дисперсии ошибок модели должны быть одинаковы для всех точек данных. Невыполнение этого условия называется гетероскедастичностью данных. Наличие гетероскедастичности случайных ошибок приводит к [неэффективности оценок](https://ru.wikipedia.org/wiki/%D0%AD%D1%84%D1%84%D0%B5%D0%BA%D1%82%D0%B8%D0%B2%D0%BD%D0%B0%D1%8F_%D0%BE%D1%86%D0%B5%D0%BD%D0%BA%D0%B0), полученных с помощью обычного [метода наименьших квадратов](https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D1%82%D0%BE%D0%B4_%D0%BD%D0%B0%D0%B8%D0%BC%D0%B5%D0%BD%D1%8C%D1%88%D0%B8%D1%85_%D0%BA%D0%B2%D0%B0%D0%B4%D1%80%D0%B0%D1%82%D0%BE%D0%B2). Кроме того, в этом случае оказывается [смещённой](https://ru.wikipedia.org/wiki/%D0%9D%D0%B5%D1%81%D0%BC%D0%B5%D1%89%D0%B5%D0%BD%D0%BD%D0%B0%D1%8F_%D0%BE%D1%86%D0%B5%D0%BD%D0%BA%D0%B0) и [несостоятельной](https://ru.wikipedia.org/wiki/%D0%A1%D0%BE%D1%81%D1%82%D0%BE%D1%8F%D1%82%D0%B5%D0%BB%D1%8C%D0%BD%D0%B0%D1%8F_%D0%BE%D1%86%D0%B5%D0%BD%D0%BA%D0%B0) классическая оценка ковариационной матрицы МНК-оценок параметров.

Так как

SSE=$e^{T}We\rightarrow min,$ (5)

где W$-$ковариационная матрица ошибок, которая в случае гетероскедастичности данных является диагональной матрицей.

Если ковариационная матрица ошибок диагональная (имеется гетероскедастичность ошибок, но нет автокорреляции), то обобщённая сумма квадратов является фактически взвешенной суммой квадратов, где веса обратно пропорциональны дисперсиям ошибок. В этом случае говорят о взвешенном МНК.

$Cov\left(ε\_{i}\right)=δ=\left(\begin{array}{c}\begin{matrix}δ\_{1}^{2}&0 \\0&δ\_{2}^{2}\end{matrix}\begin{matrix}…&0\\…&0\end{matrix}\\\begin{matrix}…& …\end{matrix} \begin{matrix} …&…\end{matrix}\\\begin{matrix} 0& 0\end{matrix}\begin{matrix} …&δ\_{n}^{2}\end{matrix}\end{array}\right)$ (6)

Как и в общем случае, дисперсии ошибок неизвестны и их необходимо оценить из тех же данных. Поэтому делают некоторые упрощающие предположения о структуре гетероскедастичности.

В данной работе рассмотрены два подхода:

1. Дисперсия ошибки пропорциональна некоторой переменной. В основе этого подхода ложится предположение о том, что дисперсия ошибок будет изменяться при изменении значений какого-либо фактора, т.е.
2. величины ошибок и фактора X должны быть коррелированны.
3. рассчитывается коэффициент корреляции Спирмена по формуле:

$r\_{s}=1-\frac{6}{n^{3}-n}\sum\_{i=1}^{n}\left(r\_{i}-q\_{i}\right)^{2}$.

1. Проводится его тест на значимость. Если тест подтверждает значимость коэффициента корреляции, то разделить все переменные на  фактор X (включая константу, то есть появится новая переменная 1/ X).
2. К преобразованным данным применяется обычный МНК для получения оценок параметров.

2) Однородные группы наблюдений[2].

1. Определяем число групп по формуле Стерджесса:

$n=1+3.322lgN=log\_{2}N+1$ (7)

1. Величина интервала группировки находится по формуле: $i=(X\_{max}-X\_{min})/n$.
2. Модель оценивают обычным МНК и находят остатки.
3. По остаткам внутри каждой группы оценивают дисперсии $δ\_{j}^{2}$.
4. Данные каждой j-й группы наблюдений делятся на $δ\_{j}$ .
5. К преобразованным подобным образом данным применяется обычный МНК.

Для анализа эффективности взвешенного метода наименьших квадратов были проведены численные эксперименты для различных временных рядов. Пример сравнения оценок качества одной из этих моделей представлен на рисунке 1. На рисунке изображена модель до преобразования и после, так же показаны графики остатков.



Рисунок 1 –Результат численного эксперимента

На рисунке 2 приведен пример реализации метода однородных групп наблюдений в MS Excel и на языке программирования C#. На рисунке 3 приведен пример реализации метода, когда дисперсия ошибок пропорциональна фактору X.



Рисунок 2 – Результат численного эксперимента для однородных групп наблюдения



Рисунок 3 – Результат численного эксперимента при гетероскедастичности данных

Библиографический список:

1. Дрейпер, Н. Прикладной регрессионный анализ [Текст] : пер. с англ. Ю. П. Адлером, В. Г. Горским. / Н. Дрейпер, Г. Смит. – книга 2, 2-е изд. – М. : Финансы и статистика, 2012. – 304 с.
2. Понятский, В. М. Использование метода группового учета аргументов для выбора структуры модели динамического объекта [Текст] / В. М. Понятский, С. И. Велешки, А. В. Жирнова. // Известия Тульского государственного университета. Технические науки. – 2013. – №2.