ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Е.А. Ботвинкин, Е.Р. Горяинова

СРАВНИТЕЛЬНЫЙ АНАЛИЗ РАЗЛИЧНЫХ МЕТОДОВ ОЦЕНИВАНИЯ В ЛИНЕЙНОЙ РЕГРЕССИИ

Препринт WP7/2014/07
Серия WP7
«Математические методы
анализа решений в экономике,
бизнесе и политике

Москва
2014

УДК 519.2
ББК 22.172
Б86

Редакторы серии WP7
«Математические методы анализа решений
в экономике, бизнесе и политике»
Ф.Т. Алескеров, В.В. Подиновский, Б.Г. Миркин

Ботвинкин, Е. А., Горяинова, Е. Р.

Б86 Сравнительный анализ различных методов оценивания в линейной регрессии : препринт WP7/2014/07 [Текст] / Е. А. Ботвинкин, Е. Р. Горяинова ; Нац. исслед. ун-т «Высшая школа экономики». – М. : Изд. дом Высшей школы экономики, 2014. – (Серия WP7 «Математические методы анализа решений в экономике, бизнесе и политике»). – 24 с. – 20 экз.

В работе рассмотрены три метода оценивания параметров в линейной регрессионной модели с неизвестным распределением шумов. Для различных распределений шумов аналитически вычислены асимптотические относительные эффективности (АОЭ) ранговых оценок по отношению к МНК-оценкам и ранговых оценок по отношению к МНМ-оценкам. С помощью метода Монте-Карло смоделированы уравнения регрессии с заданными параметрами и заданными распределениями шумов. Для выборок умеренного объёма получены усреднённые по 1000 повторений квадраты отклонений вычисленных МНК, МНМ и ранговых оценок от истинных параметров модели. Даны рекомендации по применению МНК, МНМ и ранговых оценок для различных распределений шумов.

УДК 519.2
ББК 22.172

Ботвинкин Е.А. – студент магистратуры факультета компьютерных наук НИУ ВШЭ, Москва, Россия.

Горяинова Е.Р. – Департамент математики факультета экономики НИУ ВШЭ, Москва, Россия.

**Препринты Национального исследовательского университета
«Высшая школа экономики» размещаются по адресу: http://www.hse.ru/org/hse/wp**

1. Введение

В настоящее время трудно найти область прикладных исследований, в которой не использовался бы регрессионный анализ. Задача оценивания параметров в линейной регрессии известна с XIX в. Наиболее используемым методом оценивания является разработанный Лежандром и Гауссом метод наименьших квадратов (МНК). Привлекательность этого метода обусловлена двумя причинами. Во-первых, при справедливости предположения о гауссовском распределении погрешностей модели МНК-оценка обладает важными статистическими свойствами: является эффективной и имеет нормальное распределение. Во-вторых, МНК-оценка имеет явное аналитическое выражение и легко вычисляется. Однако на практике предположение о гауссовости выполняется далеко не всегда. Обширные исследования по анализу реальных данных, представленные в [Дэниел, 1979], свидетельствуют о том, что наличие в данных отдельных резко выделяющихся наблюдений («выбросов») или небольшого процента ошибок является скорее правилом, чем исключением. Такие, казалось бы, несущественные отклонения от предполагаемой гауссовской модели способны совершенно обесценить анализ, основанный на МНК. Таким образом, возникла необходимость в разработке непараметрических методов, ориентированных на широкий класс распределений погрешностей и устойчивых к наличию «выбросов». Одним из первых альтернативных МНК-методов был метод наименьших модулей (МНМ), предложенный Лапласом. Интуитивно понятно, что большая устойчивость МНМ-оценки по сравнению с МНК-оценкой связана с тем, что МНМ основан на минимизации линейной функции остатков, а МНК – квадратичной. Известно, что МНМ является оптимальным в случае, когда погрешности модели имеют двойное экспоненциальное распределение. Однако, как будет показано в этой работе, МНМ-оценка оказывается весьма неточной в случаях, когда шумы имеют полимодальные или двумодальные распределения с «провалами» плотности в нуле. Это обстоятельство, в частности, свидетельствует о том, что в условиях неизвестного распределения шумов МНМ не всегда будет надёжной альтернативой МНК. В 1970-е годы были построены ранговые оценки парамет-

ров в линейной регрессии. Методы рангового оценивания и асимптотические свойства ранговых оценок представлены в [Jaeckel, 1972; Jureckova, 1971; Хеттманспергер, 1987]. Достоинством ранговых оценок (R-оценок) является их устойчивость к засорениям и достаточно высокая точность оценивания в случаях, когда погрешности имеют распределения с «тяжёлыми хвостами». Кроме того, в данной работе будет показано, что среди 15 наиболее распространённых непрерывных распределений шумов ранговая оценка не уступает в точности и в асимптотической эффективности оценкам МНК и МНМ одновременно. Другими словами, ни на одном из рассмотренных распределений ранговая оценка не будет наихудшей, а для таких распределений с «тяжёлыми хвостами», как распределение Тьюки и распределение Стьюдента с числом степеней свободы от 2 до 18, ранговая оценка является наилучшей в смысле точности оценивания и наибольшей асимптотической эффективности. Ещё один аспект, обусловивший интенсивное развитие рангового оценивания, связан с тем, что во многих практических задачах психологии, медицины, экспертного анализа и социологии характеристики изучаемых объектов измеряются не в количественной, а в порядковой (ординальной) шкале. Например, в [Алескеров, Юзбашев, Якуба, 2007] рассматриваются объекты, которые могут быть измерены лишь в порядковой трёхградационной шкале с категориями «плохо» – «средне» – «хорошо». В [Алескеров, Субочев, 2009] рассмотрена задача группового выбора, в которой каждый участник согласно индивидуальным предпочтениям проводит ранжирование (упорядочивание) имеющихся альтернатив. В [Zhang, Liu, Wang, 2010] исследуется зависимость между фенотипом и психическими заболеваниями пациентов, поведенческие расстройства которых измеряются в ординальной шкале.

Данная работа имеет следующую структуру. В разделе 2 даны определения МНК, МНМ и R-оценок в линейной регрессионной модели, описаны алгоритмы вычисления этих оценок, указаны их асимптотические распределения. В разделе 3 дано определение асимптотической относительной эффективности (АОЭ) одной оценки относительно другой, приведены АОЭ ранговой оценки относительно МНК и ранговой оценки относительно МНМ для различных распределений

погрешностей регрессионной модели. В разделе 4 представлены результаты компьютерного моделирования по исследованию точности оценивания параметров тремя рассмотренными методами для выборок умеренного объёма при различном распределении шумов. В разделе 5 на примере с реальными данными исследуется влияние единичного выброса на ранговые, МНК и МНМ-оценки. В Заключении приведены выводы и рекомендации по применению МНК, МНМ и R-методов в модели линейной регрессии.

2. Алгоритм построения МНК, МНМ и R-оценок в линейной регрессионной модели

Рассмотрим модель линейной регрессии вида

$$Y = [\mathbf{1}X]\theta + \varepsilon, \qquad (1)$$

где $Y = (y_1, \dots, y_n)^T$ – вектор наблюдаемых значений зависимой переменной, n – количество наблюдений, $\mathbf{1}$ – вектор-столбец размера $n \times 1$ из единиц;

$$X = \begin{pmatrix} x_{11} & \dots & x_{1m} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nm} \end{pmatrix}$$

– матрица плана размера $n \times m$, $m < n$, в которой x_{ij} – значение j-го регрессора в i-м наблюдении, $\theta = (\theta_0, \dots, \theta_m)^T$ – вектор неизвестных параметров, а $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$ – вектор одинаково распределённых ненаблюдаемых ошибок.

Далее везде предполагается, что матрица $[\mathbf{1}X]$ имеет полный столбцовый ранг, математические ожидания компонент вектора ε нулевые $E\varepsilon_1 = 0$, ковариационная матрица K_ε вектора ε имеет вид $K_\varepsilon = \sigma^2 I$, где I – единичная матрица, а дисперсия σ^2 неизвестна.

Обозначим через $F(x)$ неизвестную функцию распределения компонент вектора ε. Введём также дополнительный вектор

$\beta = (\theta_1, \dots, \theta_m)^T$. Таким образом, вектор θ можно представить как $\theta^T=(\theta_0, \beta^T)$.

Ранговые оценки неизвестных параметров в модели (1) были предложены в [Jaeckel, 1972]. Они определяются как решение экстремальной задачи минимизации некоторой меры рассеяния остатков $D(\cdot)$, удовлетворяющей условиям $D(Z + \mathbf{1}a) = D(Z)$ и $D(-Z) = D(Z)$ для любого n-мерного вектора Z и скаляра a. Функцию потерь $D(\cdot)$, для которой выполняются указанные свойства, называют чётной, свободной от сдвига мерой рассеяния. Если рассматривать функцию $D(\cdot)$ как функцию $D(Y - X\beta)$ параметров $\theta_1, \dots, \theta_m$, то точка, в которой функция $D(Y - X\beta)$ достигает минимума, будет являться оценкой параметра β регрессионной модели. Отметим, что мера изменчивости

$$D(Y - X\beta) = D(Y - X\beta - 1\theta_0) = D(Y - X\theta)$$

не зависит от сдвига, поэтому ранговое оценивание свободного члена θ_0 для модели (1) проводится отдельно от оценивания остальных параметров.

Определение. Ранговой оценкой (R-оценкой) вектора параметров $\beta^T = (\theta_1, \dots, \theta_m)$ называется такой вектор $\hat{\beta}_R$, который минимизирует функцию

$$D(Y - X\beta) = D(\beta) = D(\theta_1, \dots, \theta_m) = \\ = \sum_{k=1}^{n} (y_k - x_{k1}\theta_1 - \cdots - x_{km}\theta_m) \left(\frac{R(y_k - x_{k1}\theta_1 - \cdots - x_{km}\theta_m)}{n+1} - \frac{1}{2}\right)\sqrt{12}, \qquad (2)$$

где $R(y_k - x_{k1}\theta_1 - \cdots - x_{km}\theta_m)$ – ранг $y_k - x_{k1}\theta_1 - \cdots - x_{km}\theta_m$ среди всех величин $y_1 - x_{11}\theta_1 - \cdots - x_{1m}\theta_m \dots y_n - x_{n1}\theta_1 - \cdots - x_{nm}\theta_m$.

Следует сказать, что определённая таким образом ранговая оценка называется R-оценкой соответствующей вилкоксоновской функции меток. В данной работе мы будем рассматривать только такой вид R-оценок.

В качестве R-оценки свободного члена θ_0 в модели (1) предлагается использовать выборочную медиану остатков

$y_1 - (\hat{\theta}_1 x_{11} + \cdots + \hat{\theta}_m x_{1m}), \ldots, y_n - (\hat{\theta}_1 x_{n1} + \cdots + \hat{\theta}_m x_{nm})$,

где $\hat{\theta}_1, \ldots, \hat{\theta}_m$ – R-оценки параметров $\theta_1, \ldots, \theta_m$, т.е.

$$\hat{\theta}_0 = med(y_1 - \sum_{j=1}^{m} \hat{\theta}_j x_{1j}, \ldots, y_n - \sum_{j=1}^{m} \hat{\theta}_j x_{nj}). \quad (3)$$

В [Хеттманспергер, 1987] доказано, что функция $D(Y - X\beta)$ является неотрицательной, непрерывной и выпуклой функцией β. Кроме того, в [Jaeckel, 1972] показано, что если матрица плана является матрицей полного столбцового ранга, то функция (2) достигает минимума, и множество β, на котором достигается минимум, ограничено. Таким образом, в качестве R-оценки можно выбрать любое значение, минимизирующее функцию (2).

В силу этого утверждения можно искать минимум функции $D(Y - X\beta)$ при помощи численных методов по отысканию локального минимума. В данной работе при проведении экспериментов для нахождения минимума функции D используется встроенный в Matlab метод симплексного поиска. Этот метод подходит для задачи минимизации функции потерь, поскольку для непрерывных функций позволяет найти локальный минимум.

Отметим здесь, что в [Хеттманспергер, 1987] предлагается иной алгоритм нахождения R-оценки. А именно было найдено линейное приближение градиента $\nabla D(Y - X\beta)$ функции $D(Y - X\beta)$, а затем с помощью итерационных процедур находилось приближённое решение уравнения $\nabla D(Y - X\beta) = 0$. Этот алгоритм очень трудоёмкий, но его главный недостаток состоит в том, что для нахождения линейной аппроксимации функции $\nabla D(Y - X\beta)$ требуется оценивать интеграл от квадрата неизвестной плотности распределения шумов. А при небольшом объёме данных эта оценка может оказаться неточной.

Оценкой вектора $\theta = (\theta_0, \ldots, \theta_m)^T$ по методу наименьших модулей (МНМ-оценкой) называют такой вектор $\hat{\theta}_{\text{МНМ}}$, который минимизирует функцию потерь вида

$$S(\theta_0, \ldots, \theta_m) = \sum_{k=1}^{n} |y_k - \theta_0 - x_{k1}\theta_1 - \cdots - x_{km}\theta_m|. \quad (4)$$

С технической точки зрения существенный недостаток МНМ-оценки состоит в том, что отыскание минимума указанной функции потерь можно осуществить только численными методами. В данной работе для построения приближенной МНМ-оценки также использовался метод симплексного поиска.

Наиболее распространённой оценкой вектора θ является оценка, получаемая с помощью метода наименьших квадратов (МНК-оценка). МНК-оценкой вектора $\theta = (\theta_0, \dots, \theta_m)^T$ называется такой вектор $\hat{\theta}_{\text{МНК}}$, который минимизирует функцию потерь вида

$$S(\theta_0, \dots, \theta_m) = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{m} x_{ij}\theta_j \right)^2.$$

Если матрица $[\mathbf{1}X]$ имеет полный столбцовый ранг, то указанная экстремальная задача имеет единственное решение, которое явно выражается следующей формулой

$$\hat{\theta}_{\text{МНК}} = ([\mathbf{1}X]^T[\mathbf{1}X])^{-1}[\mathbf{1}X]^T Y. \quad (5)$$

Для того, чтобы провести аналитический сравнительный анализ точности оценивания параметров модели (1) рассмотренными методами, необходимо указать асимптотические распределения соответствующих оценок.

Обозначим $\hat{\beta}_{\text{МНК}} = \left(\hat{\theta}_1, \dots, \hat{\theta}_m\right)^T$, где $\hat{\theta}_1, \dots, \hat{\theta}_m$ – МНК-оценки параметров $\theta_1, \dots, \theta_m$, определённые в формуле (5), а $\hat{\beta}_{\text{МНМ}} = \left(\tilde{\theta}_1, \dots, \tilde{\theta}_m\right)^T$, где $\tilde{\theta}_1, \dots, \tilde{\theta}_m$ – МНМ-оценки параметров $\theta_1, \dots, \theta_m$, полученные минимизацией функции (4).

Будем предполагать, что при $n \to \infty$ матрица $\frac{1}{n}X^T X$ сходится к положительно определённой матрице Σ, т.е.

$$\frac{1}{n}X^T X \to \Sigma. \quad (6)$$

При выполнении условия (6) (см., например, [Себер, 1980]) МНК-оценка вектора β в модели (1) является асимптотически нормальной и

$$\sqrt{n}\left(\hat{\beta}_{\text{МНК}} - \beta\right) \xrightarrow{d} U, \quad U \sim N\left(0, \sigma^2 \Sigma^{-1}\right).$$

Асимптотические свойства МНМ-оценки исследованы в [Basset, Koenker, 1978] и [Pollard, 1991]. Теорема об асимптотической нормальности сформулирована также в [Болдин, Симонова, Тюрин, 1997].

Пусть выполнено условие (6), функция распределения $F(x)$ имеет нулевую медиану, а соответствующая ей функция плотности $f(x)$ непрерывна и положительна в нуле. Тогда МНМ-оценка в модели (1) является асимптотически нормальной и

$$\sqrt{n}\left(\hat{\beta}_{\text{МНМ}} - \beta\right) \xrightarrow{d} U, \qquad U \sim N\left(0, \frac{1}{\left(2f(0)\right)^2}\ \Sigma^{-1}\right).$$

Нижеследующая теорема об асимптотическом распределении ранговой оценки $\hat{\beta}_R$, соответствующей функции потерь (2) с вилкоксоновской функцией меток, приведена в [Хеттмансперегер, 1987].

Пусть выполнено условие (6), существует первая производная $f'(x)$ у плотности $f(x)$ распределения $F(x)$ и $f(x)$ имеет конечную информацию Фишера. Тогда любая точка $\hat{\beta}_R$, минимизирующая функцию потерь (2), является асимптотически нормальной и

$$\sqrt{n}\left(\hat{\beta}_R - \beta\right) \xrightarrow{d} U, \qquad U \sim N\left(0, \frac{1}{12\left(\int_{-\infty}^{\infty} f^2(x)dx\right)^2}\ \Sigma^{-1}\right).$$

Отметим, что асимптотическое распределение ранговых оценок в модели (1), соответствующих ранговым меткам общего вида, приведено, например, в [Болдин, Симонова, Тюрин, 1997].

3. Асимптотическая относительная эффективность МНК, МНМ и R-оценок

Перейдём теперь к сравнительному анализу точности МНК, МНМ и R-оценок для выборок, объём которых стремится к бесконечности. В предыдущем разделе было указано на то, что все три рассматриваемые оценки имеют одинаковые асимптотические векторы средних, которые равны вектору истинных параметров β. Значит, более точной будет та из оценок, распределение которой окажется более сконцентрированным возле истинного параметра θ. Рассеяние оценки вокруг своего среднего значения характеризуется дисперсией. Уилкс (см. [Уилкс, 1967, с. 546]) определил обобщённую дисперсию многомерной оценки как определитель ковариационной матрицы. Асимптотической относительной эффективностью (АОЭ) одной m-мерной асимптотически нормальной оценки относительно другой называют обратное отношение обобщённых асимптотических дисперсий этих оценок, возведённое в степень $1/m$.

Поскольку ковариационные матрицы асимптотических распределений МНК, МНМ и R-оценок пропорциональны матрице $(X^TX)^{-1}$, то АОЭ будет определяться отношением скалярных множителей при соответствующих ковариационных матрицах. Таким образом, если значение АОЭ первой оценки относительно второй оказывается больше единицы, то это означает, что первая оценка является более эффективной (более точной), чем вторая.

Итак, согласно введённому определению, вычислим АОЭ $e\left(\hat{\beta}_R, \hat{\beta}_{\text{МНК}}\right)$ ранговой оценки $\hat{\beta}_R$ по отношению к МНК-оценке

$$e\left(\hat{\beta}_R, \hat{\beta}_{\text{МНК}}\right) = \left(\frac{\det(\sigma^2 (X^TX)^{-1})}{\det\left(\frac{1}{12\left(\int_{-\infty}^{\infty} f^2(x)dx\right)^2} (X^TX)^{-1} \right)} \right)^{1/m} =$$

$$= 12\sigma^2 \left(\int_{-\infty}^{\infty} f^2(x)dx \right)^2,$$

где σ^2 – дисперсия шума, $f(x)$ – плотность его распределения, m – число параметров модели. АОЭ ранговой оценки по отношению к МНМ-оценке $e\left(\hat{\beta}_R, \hat{\beta}_{\text{МНМ}}\right)$ выражается как

$$e\left(\hat{\beta}_R, \hat{\beta}_{\text{МНМ}}\right) = \left(\frac{\det\left(\frac{1}{\left(2f(0)\right)^2}(X^TX)^{-1}\right)}{\det\left(\frac{1}{12\left(\int_{-\infty}^{\infty} f^2(x)dx\right)^2}(X^TX)^{-1}\right)} \right)^{1/m} = \\ = \frac{3}{f^2(0)}\left(\int_{-\infty}^{\infty} f^2(x)dx\right)^2.$$

Полученные выражения для $e\left(\hat{\beta}_R, \hat{\beta}_{\text{МНК}}\right)$ и $e\left(\hat{\beta}_R, \hat{\beta}_{\text{МНМ}}\right)$ указывают на то, что АОЭ существенно зависит от распределения погрешностей регрессионной модели. Вычислим значения АОЭ для следующих распределений: нормального, распределения Лапласа с плотностью $f(x) = \frac{\alpha}{2}e^{-\alpha|x-\gamma|}$ с параметрами $\gamma = 0$ и $\alpha = 1$, распределения Коши с плотностью $f(x) = \frac{\gamma}{\pi(x^2+\gamma^2)}$ с параметром $\gamma = 1$, распределений Стьюдента с 2, 3, 5, 13, 18 и 19 степенями свободы, распределения Симпсона (треугольного распределения), равномерного на интервале (–1; 1) распределения, логистического распределения с плотностью $f(x) = \frac{e^{-x}}{(1+e^{-x})^2}$ и «двугорбого» распределения на основе двух гауссовских с плотностью

$$f(x) = \frac{1}{2}\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-2)^2}{2}} + \frac{1}{2}\frac{1}{\sqrt{2\pi}}e^{-\frac{(x+2)^2}{2}}.$$

В табл. 1 приведены результаты числовых значений АОЭ для указанных распределений шумов регрессионной модели. Интегралы для распределений Стьюдента с различными степенями свободы, распределения Коши и логистического распределения были вычислены численно в среде Matlab, остальные взяты аналитически.

Таблица 1. АОЭ ранговой оценки по отношению к МНК и МНМ-оценкам

Распределение	АОЭ ранговой оценки к МНК	АОЭ ранговой оценки к МНМ
Нормальное распределение	$3/\pi \approx 0{,}9549$	1,5
Распределение Лапласа	1,5	0,75
Распределение Коши	∞	0,75
Распределение Стьюдента с 2 степенями свободы	∞	1,0416
Распределение Стьюдента с 3 степенями свободы	1,8998	1,1725
Распределение Стьюдента с 5 степенями свободы	1,2412	1,3553
Распределение Стьюдента с 13 степенями свободы	1,0252	1,4162
Распределение Стьюдента с 18 степенями свободы	1,0023	1,438
Распределение Стьюдента с 19 степенями свободы	0,9993	1,4417
Треугольное распределение	$8/9 \approx 0{,}8889$	$4/3 \approx 1{,}3333$
Логистическое распределение	$\pi^2/9 \approx 1{,}0966$	$4/3 \approx 1{,}3333$
Равномерное распределение	1	3
«Двугорбое» распределение на основе комбинации гауссовских	1,236	21,22

Особое внимание уделим распределению Тьюки, представляющему загрязнённое нормальное распределение. С помощью распределения Тьюки можно описать реальную ситуацию наличия в стандартной гауссовской выборке небольшой (1–15%) доли выбросов. Плотность распределения Тьюки с долей загрязнения γ ($0 \le \gamma < 1$) и дисперсией загрязняющего распределения σ_1^2 (где $\sigma_1^2 > \sigma_2^2$) имеет вид

$$f(x) = \gamma \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma_1^2}} + (1-\gamma)\frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma_2^2}}.$$

Отметим здесь, что загрязняющее распределение необязательно должно быть гауссовским.

Плотность распределения Тьюки с долей загрязнения $\gamma = 0{,}1$ и дисперсиями смешиваемых гауссовских величин $\sigma_1^2 = 100$ и $\sigma_2^2 = 1$

показана на рис. 1 пунктирной линией, плотность стандартного гауссовского распределения – сплошной линией. Рис. 1 наглядно показывает, что плотность распределения Тьюки на бесконечности убывает медленнее, чем плотность гауссовского распределения (имеет более «тяжёлые хвосты»).

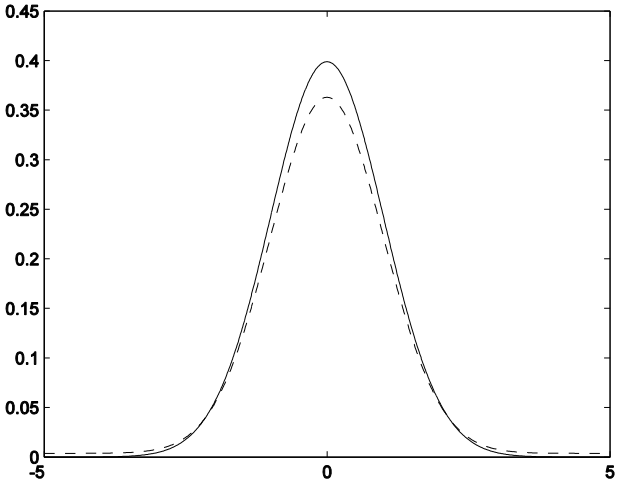

Рис. 1. Плотность стандартного гауссовского распределения (сплошная линия) и распределения Тьюки с параметрами $\gamma = 0{,}1$, $\sigma_1^2 = 100$, $\sigma_2^2 = 1$ (пунктирная линия)

В табл. 2 представлены асимптотические относительные эффективности $e\left(\hat{\beta}_R, \hat{\beta}_{\text{МНК}}\right)$ и $e\left(\hat{\beta}_R, \hat{\beta}_{\text{МНМ}}\right)$ для случая, когда погрешности имеют распределение Тьюки с различными параметрами γ, σ_1^2 и $\sigma_2^2 = 1$.

Таблица 2. АОЭ ранговой оценки по отношению к МНК и МНМ-оценкам для распределения Тьюки

γ	σ_1^2	$e\left(\hat{\beta}_R, \hat{\beta}_{\text{МНК}}\right)$	$e\left(\hat{\beta}_R, \hat{\beta}_{\text{МНМ}}\right)$
0,05	9	1,196	1,436
0,05	25	1,815	1,405
0,05	100	4,767	1,381
0,1	9	1,373	1,375
0,1	25	2,410	1,316

γ	σ_1^2	$e(\hat{\beta}_R, \hat{\beta}_{\text{МНК}})$	$e(\hat{\beta}_R, \hat{\beta}_{\text{МНМ}})$
0,1	100	7,208	1,268
0,15	9	1,496	1,319
0,15	25	2,794	1,231
0,15	100	8,755	1,157

Проведённые вычисления позволяют сделать следующие выводы:

• Ранговая оценка имеет более высокую асимптотическую эффективность, чем МНК-оценка, в случаях, когда погрешности в модели (1) имеют логистическое распределение, распределение Лапласа, распределение Коши, распределение Тьюки, распределение Стьюдента с менее чем 19 степенями свободы.

• Ранговая оценка имеет более высокую асимптотическую эффективность, чем МНМ-оценка, в случаях, когда погрешности в модели (1) имеют логистическое распределение, треугольное распределение, нормальное распределение, равномерное распределение, распределение Стьюдента с двумя и более степенями свободы, двугорбое на основе двух гауссовских и распределение Тьюки с небольшими и умеренными параметрами загрязнения.

• В случае распределения Тьюки при увеличении доли загрязнения или при увеличении дисперсии загрязняющего распределения АОЭ ранговой оценки по отношению к МНМ-оценке имеет тенденцию к снижению.

• Ранговая оценка уступает в эффективности МНК-оценке в моделях с шумами, имеющими нормальное распределение, распределение Стьюдента с не менее чем 19 степенями свободы и треугольное распределение.

• Ранговая оценка уступает МНМ в моделях с шумами, имеющими распределения Лапласа и Коши.

4. Численный сравнительный анализ точности МНК, МНМ и R-оценок

Результаты предыдущего раздела относились к случаю большого объёма наблюдений ($n \to \infty$). Теперь, для того чтобы выяснить насколько точными являются рассматриваемые оценки при умеренных

объёмах выборок, построим компьютерную модель линейной регрессии вида (1) с $m+1=3$ параметрами, включая свободный член, и числом наблюдений $n=50$. В качестве распределений погрешностей модели будут выбраны те распределения, которые указаны в предыдущем разделе. Помимо этих распределений рассмотрим ещё одно двумодальное распределение на основе двух треугольных с плотностью

$$f(x)=\begin{cases}1-|1-2x|, & x\in[0,1],\\ 1-|1+2x|, & x\in[-1,0],\\ 0, & \text{иначе.}\end{cases}$$

Отметим, что значение плотности этого распределения в нуле обращается в ноль (т.е. $f(0)=0$). Поэтому на таком распределении нельзя определить обобщённую асимптотическую дисперсию МНМ-оценки.

Для каждой из указанных плотностей и одного и того же значения параметров была проведена 1000 генераций данных. Используя смоделированные данные, при каждом из 1000 повторов проводилось оценивание параметров ранговым методом, методом наименьших квадратов и методом наименьших модулей.

Назовём ошибкой оценивания истинного вектора параметров $\theta=(\theta_0,\dots,\theta_m)^T$ сумму квадратов разностей

$$d^2(\hat{\theta},\theta)=\sum_{i=0}^{m}\left(\hat{\theta}_i-\theta_i\right)^2, \qquad (7)$$

в которой $\hat{\theta}=(\hat{\theta}_0,\dots,\hat{\theta}_m)^T$ – вектор оценок. Критерием качества оценки будет выступать усреднённая по 1000 повторов ошибка оценивания. Таким образом, наилучшим среди трёх рассмотренных будет считаться тот метод, у которого будет наименьшая усреднённая ошибка оценивания. Напомним, что МНК-оценка вектора θ определяется формулой (5), МНМ-оценка – точка минимума функции (4), а $\hat{\theta}_R=(\hat{\theta}_0,\hat{\beta}_R^T)^T$, где $\hat{\theta}_0$ определяется формулой (3), а $\hat{\beta}_R$ – любая точка, минимизирующая функцию (2). Результаты проведенных экспериментов сведены в табл. 3.

Таблица 3. Усреднённая ошибка оценивания параметров

$d^2(\hat{\theta}, \theta)$ **для:**	**Ранговой оценки**	**МНК**	**МНМ**
Стандартное нормальное распределение	0,1959	0,1768	0,2677
Распределение Лапласа	0,2416	0,3332	0,2464
Распределение Коши	0,6909	16958,23	0,5641
Распределение Стьюдента с 2 степенями свободы	0,3399	1,6814	0,3652
Распределение Стьюдента с 3 степенями свободы	0,2766	0,4770	0,3337
Распределение Стьюдента с 5 степенями свободы	0,2488	0,3054	0,3102
Распределение Стьюдента с 13 степенями свободы	0,2006	0,1947	0,2740
«Двугорбое» распределение на основе комбинации гауссовских	1,3196	0,8360	3,0581
Треугольное распределение	0,0328	0,0271	0,0477
«Двугорбое» распределение на основе комбинации треугольных	0,0796	0,0482	0,2046
Логистическое распределение	0,5243	0,5530	0,6847
Равномерное на [–1; 1] распределение	0,0764	0,0573	0,1507

Усреднённая ошибка оценивания $d^2(\hat{\theta}, \theta)$ для распределения Тьюки с различными параметрами загрязнения представлена в табл. 4.

Таблица 4. Усреднённая ошибка оценивания параметров при распределении Тьюки с параметрами γ, σ_1^2 и σ_2^2

Доля зашумления γ	**Дисперсии** σ_1^2 и σ_2^2	**Ранговая оценка**	**МНК-оценка**	**МНМ-оценка**
0,05	$\sigma_1^2 = 9, \sigma_2^2 = 1$	0,2038	0,2156	0,2808
0,05	$\sigma_1^2 = 25, \sigma_2^2 = 1$	0,2102	0,3608	0,2943
0,05	$\sigma_1^2 = 100, \sigma_2^2 = 1$	0,2288	1,0443	0,2948
0,1	$\sigma_1^2 = 9, \sigma_2^2 = 1$	0,2362	0,3065	0,3111
0,1	$\sigma_1^2 = 25, \sigma_2^2 = 1$	0,2671	0,6023	0,3145
0,1	$\sigma_1^2 = 100, \sigma_2^2 = 1$	0,2963	1,7633	0,3530
0,15	$\sigma_1^2 = 9, \sigma_2^2 = 1$	0,2539	0,3463	0,3081
0,15	$\sigma_1^2 = 25, \sigma_2^2 = 1$	0,3259	0,8290	0,3697
0,15	$\sigma_1^2 = 100, \sigma_2^2 = 1$	0,3414	2,6568	0,3713

Результаты численного эксперимента позволяют сделать следующие выводы:

• МНК наиболее точен для оценивания параметров регрессионной модели с шумами, имеющими распределение Гаусса, Стьюдента с 13 и более степенями свободы, «двугорбое» распределение на основе гауссовских величин, треугольное распределение и «двугорбое» распределение на основе треугольных. Этот метод дает наихудшую оценку при распределении Лапласа, Коши, Тьюки и Стьюдента с менее чем 5 степенями свободы. Более того, для распределения Коши и распределения Тьюки с высоким уровнем загрязнения МНК-оценка существенно уступает в точности МНМ и R-оценкам.

• МНМ дает наиболее точную оценку при шумах в модели, имеющих распределение Коши и оценку, сопоставимую по точности с ранговой, при распределении Лапласа. Этот метод в меньшей степени точен, чем рассматриваемые альтернативы, при нормальном распределении, треугольном распределении и распределении Стьюдента с 5 и более степенями свободы. МНМ существенно уступает в точности при «двугорбом» распределении на основе гауссовских величин, и «двугорбом» распределении на основе треугольных.

• Ранговый метод наиболее точен для оценивания параметров регрессионной модели с шумами, имеющими логистическое распределение, распределение Тьюки с различными параметрами загрязнения и распределение Стьюдента со степенями свободы меньше 13 и больше единицы.

• Численный эксперимент, проведенный на выборках умеренного объема, в целом подтверждает аналитические асимптотические результаты.

• Важно отметить, что ранговая оценка ни на одном из рассмотренных распределений не имела наихудшей точности.

5. Изменение оценок при искусственном внесении выбросов в реальные данные

Для построения линейной регрессионной модели на основе реальных данных был выбран достаточно известный набор данных «ирисы Фишера». Эти данные были собраны американским ботаником Эдгаром Андерсоном и включают в себя измеренные в миллиметрах длину и ширину лепестка и длину и ширину чашелистика у 150 экземпляров цветка ириса – по 50 экземпляров каждого из трех видов: «ирис щетинистый», «ирис виргинский» и «ирис разноцветный». Опишем зависимость длины лепестка (Y) от длины (X_1) и ширины (X_2) чашелистика для вида «ирис разноцветный» с помощью линейной регрессионной модели вида

$$y_i = \theta_0 + x_{i1}\theta_1 + x_{i2}\theta_2 + \varepsilon_i, \qquad i = 1, \dots, 50.$$

Оценим вектор параметров $\theta = (\theta_0, \theta_1, \theta_2)^T$, используя ранговый метод, МНК и МНМ. Полученные оценки $\hat{\theta}_R$, $\hat{\theta}_{\text{МНК}}$ и $\hat{\theta}_{\text{МНМ}}$ указаны во втором, четвёртом и шестом столбцах табл. 5 соответственно.

Проведём теперь искусственное загрязнение данных, а именно – умножим на 10 одно случайно выбранное значение зависимой переменной. Такое засорение имитирует ошибочно поставленную запятую. Проведём новое оценивание вектора параметров $\theta = (\theta_0, \theta_1, \theta_2)^T$ для загрязнённых данных тремя указанными методами. Пересчитанные

оценки $\hat{\theta}_R$, $\hat{\theta}_{\text{МНК}}$ и $\hat{\theta}_{\text{МНМ}}$ указаны в третьем, пятом и седьмом столбцах табл. 5 соответственно. В ситуации с реальными данными истинное значение параметров модели неизвестно. Поэтому мы назовём ошибкой оценивания при внесении засорения сумму квадратов разностей (7) между старой оценкой, построенной по реальным данным, и новой оценкой, построенной по данным с загрязнением. Значение величины d^2 для всех трёх типов оценок приведено в последней строке табл. 5.

Таблица 5. Изменение R, МНК и МНМ-оценок при внесении загрязнения

Оценка	**Ранговая оценка**		**МНК-оценка**		**МНМ-оценка**	
Модель	**До засорения**	**После засорения**	**До засорения**	**После засорения**	**До засорения**	**После засорения**
Θ_0	0,3141	0,757	–1,1089	21,0407	0,9023	0,903
Θ_1	0,5429	0,5899	0,578	1,5217	0,5574	0,5583
Θ_2	0,3571	0,2446	0,3394	–2,3672	0,3108	0,3092
d^2	0,211		498,8213		$3{,}7607 \cdot 10^{-6}$	

Этот пример показывает, что наилучшим образом на единичный выброс в данных реагирует МНМ-оценка, наихудшим – МНК-оценка, ранговая оценка показывает вполне удовлетворительные результаты. Можно сказать, что этот результат подтверждает выводы двух предыдущих разделов, поскольку указанный тип засорения можно трактовать как внесение небольшой доли (2%) загрязнения с очень большой дисперсией.

6. Заключение

В работе проведён сравнительный анализ ранговой, МНК и МНМ-оценок в модели линейной регрессии при различных распределениях шумов. Для выборок, объём которых стремится к бесконечности, вычислены АОЭ ранговых оценок по отношению к МНК-оценкам и МНМ-оценкам. Установлено, что среди трёх рассмотренных оценок МНК-оценки имеют наибольшую эффективность для шумов с нормальным распределением, треугольным распределением и распределе-

ниями Стьюдента с числом степеней свободы не меньше 19; МНМ-оценки имеют наибольшую эффективность для распределения Коши и распределения Лапласа (двойного экспоненциального распределения); R-оценки имеют наибольшую эффективность для логистического распределения, распределения Стьюдента с числом степеней свободы от 2 до 18, распределения Тьюки с различными параметрами загрязнения, «двугорбого» распределения на основе комбинации двух гауссовских.

Следует отметить, что МНК-оценка имеет очень низкую эффективность для распределения с «тяжёлыми хвостами», а именно для распределения Коши, распределения Стьюдента с двумя степенями свободы и распределения Тьюки даже с небольшими параметрами загрязнения. МНМ-оценка имеет очень низкую эффективность для равномерного распределения и двумодальных распределений с «провалами» в нуле, а именно для «двугорбого» распределения на основе двух гауссовских и на основе двух треугольных. Ранговая оценка ни на одном из рассмотренных распределений не показала наихудшего результата.

С помощью численного моделирования была сравнена точность оценивания параметров для выборок умеренного объёма равного 50 наблюдениям. Результаты компьютерного моделирования в целом подтвердили аналитические результаты.

Итак, согласно проведённому исследованию, можно утверждать, что в условиях априорной стохастической неопределённости ранговая оценка является более предпочтительной по отношению к МНК и МНМ-оценкам.

Литература

Алескеров Ф.Т., Юзбашев Д.А., Якуба В.И. Пороговое агрегирование трёхградационных ранжировок // А и Т. 2007. № 1. С. 147–152.

Алескеров Ф.Т., Субочев А.Н. Об устойчивых решениях в ординальной задаче группового выбора // ДАН. 2009. Т. 426. № 3. С. 318–320.

Болдин М.В., Симонова Г.И., Тюрин Ю.Н. Знаковый статистический анализ линейных моделей. М.: Наука; Физматлит, 1997.

Дэниел К. Применение статистики в промышленном эксперименте. М.: Мир, 1979.

Ингстер Ю.И. и др. Основные алгоритмы численного анализа. СПб.: СПбГЭТУ «ЛЭТИ», 2009.

Мудров В.И., Кушко В.Л. Метод наименьших модулей. М.: Знание, 1971.

Себер Дж. Линейный регрессионный анализ. М.: Мир, 1980.

Уилкс С. Математическая статистика / пер. с англ.; под ред. Ю.В. Линника. М.: Наука, 1967.

Робастность в статистике. Подход на основе функций влияния / Хампель Ф., Рончетти Э., Рауссеу П., Штаэль В. М.: Мир, 1989.

Хеттмансперггер Т. Статистические выводы, основанные на рангах. М.: Финансы и статистика, 1987.

Basset G., Koenker R. Asymptotic theory of least absolute error regression // JASA. 1978. Vol. 73. No. 363. P. 618–622.

Jaeckel L.A. Estimating Regression Coefficients by Minimizing the Dispersion of the Residuals // The Annals of Mathematical Statistics. 1972. Vol. 43. No. 5. P. 1449–1458.

Jureckova J. Nonparametric estimate of regression coefficients // The Annals of Mathematical Statistics. 1971. Vol. 42. P. 1328–1338.

Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions // SIAM Journal of Optimization / *J.C. Lagarias, J.A. Reeds, M.H. Wright, P.E. Wright*. 1988. Vol. 9. No. 1. P. 112–147.

Pollard D. Asymptotics for lest absolute deviation regression estimators // Econometric Theory. 1991. Vol. 7. P. 186–199.

Zhang H., *Liu* C.-T., *Wang* X. An Association Test for Multiple Traits Based on the Generalized Kendall's Tau // JASA. 2010. Vol. 105. No. 490. P. 473–481.

Botvinkin, E. A., Goryainova, E. R.
Comparative analysis of different methods of estimation in linear regression : Working paper WP7/2014/07 [Text] / E. A. Botvinkin, E. R. Goryainova ; National Research University Higher School of Economics. – Moscow : Higher School of Economics Publ. House, 2014. – 24 p. – 20 copies.

This paper focuses on three methods of parameter estimation in linear regression model with unknown distribution of noises. For different distributions of noises there were analytically calculated asymptotic relative efficiencies (ARE) of rank estimations towards LS-estimations and LAD-estimations. There were also simulated regression equations with specific parameters and distributions of noises applying the Monte Carlo method. For datasets with moderate number of entities there were calculated mean values of squared differences between estimation vectors and a real parameter vector over a thousand of simulated regression models. There were made some recommendations on the application of the LS method, the LAD method and the rank method for cases of different distributions of noises.

Botvinkin E.A. – Graduate Faculty of Computer Science NRU HSE, Moscow, Russia.

Goryainova E.R. – Department of Mathematics for Economics NRU HSE, Moscow, Russia.

Препринт WP7/2014/07
Серия WP7
Математические методы анализа решений
в экономике, бизнесе и политике

Ботвинкин Е.А., Горяинова Е.Р.

Сравнительный анализ различных методов оценивания в линейной регрессии

Зав. редакцией оперативного выпуска *А.В. Заиченко*
Технический редактор *Ю.Н. Петрина*

Отпечатано в типографии
Национального исследовательского университета
«Высшая школа экономики» с представленного оригинал-макета
Формат 60×84 1/16. Тираж 20 экз. Уч.-изд. л. 1,5.
Усл. печ. л. 1,4. Заказ № . Изд. № 1903

Национальный исследовательский университет
«Высшая школа экономики»
125319, Москва, Кочновский проезд, 3
Типография Национального исследовательского университета
«Высшая школа экономики»