Список ключевых слов

Модель линейной регрессии, ранговая оценка параметров линейной регрессии, оценка параметров линейной регрессии по методу наименьших квадратов (МНК-оценка), оценка параметров линейной регрессии по методу наименьших модулей (МНМ-оценка), асимптотическая относительная эффективность (АОЭ)

# Введение

Объектом исследования в этой ВКР является ранговый метод оценивания параметров регрессионной модели. Этот метод применяется при построении регрессионных моделей наряду с методом наименьших квадратов и методом наименьших модулей. Предметом исследования является предпочтительность применения рангового метода.

Основные методы исследований, используемые в данной работе, это методы теории вероятностей и математической статистики, методы компьютерного моделирования и методы оптимизации.

Актуальность темы данной ВКР заключается в том, что оценки параметров модели, полученные посредством рангового метода, в некоторых случаях являются более точными, чем МНК- и МНМ-оценки. Так, например, выбросы в данных в меньшей степени влияют на оценки параметров модели, построенные с помощью рангового метода, чем на МНК-оценки. Причиной этому является то, что функция потерь для МНК-оценки включает в себя квадраты отклонений наблюдаемых значений зависимой величины от ее оценок в рамках модели, в то время как в функцию потерь ранговой модели эти отклонения входят линейно. Также стоит отметить случаи, когда шумы в моделях имеют "двугорбое распределение" - распределение с функцией плотности, имеющей две точки максимума. Тогда эффективность МНМ-оценки параметров регрессии будет ниже по сравнению с оценками, полученными МНК и ранговым методом.

В рамках этой работы планируется выполнить следующие задачи:

· разработать и численно реализовать алгоритм построения ранговой оценки неизвестных параметров регрессии;

· смоделировать регрессионные зависимости с погрешностями, имеющими распределения с "тяжёлыми" хвостами;

· провести численный сравнительный анализ ранговых оценок с МНК- и МНМ-оценками;

· вычислить аналитически асимптотическую относительную эффективность рангового метода по отношению к МНК и МНМ при различных распределениях погрешностей;

· провести эксперимент по сравнению устойчивости ранговой, МНК- и МНМ-оценок параметров модели к выбросам в реальных данных.

Цель написания этой ВКР: сделать выводы о применимости рангового метода в задаче оценивания параметров и сформулировать конкретные рекомендации по применению одного из рассматриваемых трех методов при различных распределениях шумов.

Итак, структура данной работы следующая:

· В первой главе будет рассмотрен метод построения ранговой оценки параметров модели.

· Во второй главе будут рассмотрены некоторые распределения случайных величин, включая распределения с "тяжелыми хвостами" и "двугорбые" распределения, а так же способы их моделирования. Затем будет проведен численный сравнительный анализ ранговых оценок с МНК и МНМ-оценками для моделей, шумы которых имеют вышеупомянутые распределения.

· В третьей главе будут вычислены значения асимптотической относительной эффективности ранговой оценки по отношению к МНК и МНМ.

· В четвертой главе будет построена линейная регрессионная модель на основе реальных данных и проведен эксперимент по сравнению устойчивости ранговой, МНК- и МНМ-оценок к выбросам.

· В заключении будут обобщены полученные в рамках исследования результаты и сделаны выводы относительно эффективности рангового метода оценки.

# 1. Ранговый метод

Метод наименьших квадратов широко применяется для оценки параметров линейной регрессии, поскольку достаточно прост в вычислении и при предположении о нормальном распределении шумов в модели дает оценку параметров, совпадающую с оценкой максимального правдоподобия. К недостаткам этого метода можно отнести высокую чувствительность к выбросам в данных: даже одно наблюдение с нетипичными значениями может сильно повлиять на оценки параметров и изменить общую картину. В сравнении с методом наименьших квадратов, метод наименьших модулей в меньшей степени подвержен влиянию выбросов в данных, в случае распределения ошибок согласно закону Лапласа (двойному экспоненциальному) он дает оценку, обеспечивающую максимум функции правдоподобия.

Ранговый метод оценки параметров линейной регрессии является альтернативой двум вышеупомянутым методам. Предполагается, что он является более устойчивым к выбросам в данных и обеспечивает более точную оценку параметров регрессионной модели в случае, когда распределение шумов в модели имеет "тяжелые хвосты" (например, распределение Коши), чем МНК, и лучше, чем МНМ, оценивает значения параметров в моделях с шумами, имеющими "двугорбое" распределение. В этой части будет рассмотрен метод построения ранговой оценки параметров линейной регрессионной модели.

Итак, рассмотрим линейную модель: ,

где y - наблюдаемое значение зависимой переменной,

x1,…, xm - значения независимых объясняющих переменных (или регрессоров),

θ0,…, θm - неизвестные параметры модели,

а ε - случайная ошибка наблюдения, имеющая нулевое математическое ожидание.

При наличии n наблюдений значения зависимой переменной и набора ее регрессоров, модель запишется в матричном виде:

,

где  - вектор наблюдаемых значений зависимой переменной,

 - матрица плана полного столбцового ранга, число столбцов которой не превышает числа строк, ее элемент  - значение j-го регрессора в i-м наблюдении,  - вектор параметров, а  - вектор независимых, одинаково распределенных ошибок с нулевым математическим ожиданием и ковариационной матрицей вида , где I -единичная матрица размера nxn.

Для построения оценки ранговым методом требуется ввести функцию D(Y-Xθ) - меру изменчивости. Мерой изменчивости (или функцией потерь) называется такая функция D(.), что  и  для любого n-мерного вектора Z и скаляра a. Если рассматривать эту функцию как функцию от m переменных - параметров θ1, … , θm, то точка, в которой функция D достигает минимума, будет являться оценкой параметров регрессионной модели. Функция изменчивости не зависит от сдвига, поэтому оценка свободного члена θ0 для модели проводится отдельно от оценки остальных параметров.

В статье Л. Джекла ранговой оценкой вектора параметров  (без свободного члена θ0) называется такой вектор , который минимизирует функцию , где yi - i-й элемент вектора Y, xi - i-я строка матрицы X (без столбца единиц), R(yi - xiθ) - ранг величины yi - xiθ среди всех величин yk - xkθ при k от 1 до n, а в качестве функции φ берется.

В книге Т. Хеттманспергера для построения оценок параметров θ1, … , θm регрессионной модели предлагается найти частные производные функции D(Y-Xθ) по переменным θi: , и решить систему уравнений:

.

Недостатком такого способа оценивания является сложность решения подобной системы уравнений при числе параметров m большем единицы. Поэтому, в данной работе предлагается построение приближения ранговой оценки вектора параметров при помощи численной минимизации функции D.

В статье Л. Джекла приведена следующая теорема: при фиксированном Y функция D(Y-Xθ) - неотрицательная, непрерывная и выпуклая функция θ.

В силу этого утверждения, можно искать минимум функции D(Y-Xθ) 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.

Метод симплексного поиска точки минимума функции k переменных f(z1,…, zk) заключается в выполнении следующих этапов:

· Сначала выбирается начальное приближение точки минимума z0 = (z10,…, zk0), в дополнение к этой точке генерируются еще k путем прибавления поочередно к каждой компоненте z0 5% ее значения. В точках z0,…,zk вычисляется значение функции f. Эти точки ранжируются по возрастанию значения функции в них, получается набор точек a0,…,ak таких, что f(ai) < f(aj) при i < j. Точки a0,…,ak образуют симплекс.

· Генерируется новая точка r, значение функции в ней сравнивается со значениями функции в вершинах симплекса. Если в какой-то вершине значение функции превосходит значение функции в данной точке r, то эта точка становится новой вершиной симплекса, а точка ak с наибольшим значением функции в ней убирается из рассмотрения. Имеющиеся вершины снова сортируются по возрастанию значения функции в них.

· Предыдущий шаг повторяется до тех пор, пока диаметр симплекса не будет меньше заданной величины. По окончании алгоритма в качестве решения задачи минимизации функции выбирается точка a0 из отсортированного набора вершин симплекса.

Данный метод подходит для задачи минимизации функции потерь, поскольку для непрерывных функций позволяет найти локальный минимум с небольшой ошибкой. В силу выпуклости функции любой конечный локальный минимум, даже не единственный, будет являться глобальным минимумом и подходить для оценки параметров.

В книге Т. Хеттманспергера также указано, что диаметр множества точек θ, на которых функция достигает минимума, сходится к нулю по вероятности. А это значит, что множество точек, минимизирующих функцию D, будет мало даже при умеренном числе наблюдений.

Согласно Т. Хеттманспергеру, свободный член θ0 в регрессионной модели можно оценить на основании уже полученных оценок  параметров θ1, … , θm. Для этого по выборке остатков модели , … , находится выборочная медиана. Тогда оценка свободного члена  равна элементу в середине упорядоченной выборки (если число элементов n нечетное), либо среднему арифметическому двух элементов, находящихся в середине упорядоченной выборки (при четном числе n).

Итак, в этой главе был описан ранговый метод оценивания параметров регрессионной модели и рассмотрены особенности его реализации в данной работе. В приложении № 1 представлен алгоритм для среды Matlab, который строит ранговую оценку параметров регрессионной модели, используя встроенную в Matlab функцию метода симплексного поиска.

# 2. Численный сравнительный анализ

Итак, в рамках данной работы рассматриваются такие распределения случайных величин, как распределения Гаусса и Лапласа, треугольное распределение (распределение Симпсона) и "двугорбые" распределения, моделируемые на основе гауссовских и треугольных распределений. Помимо этого рассматриваются распределение Коши, распределение Стьюдента с небольшим числом степеней свободы, распределение Тьюки и логистическое распределение. Последние распределения относятся к распределениям с "тяжелыми хвостами".

Распределение Гаусса с дисперсией σ2 > 0 и математическим ожиданием m имеет функцию плотности . В проводимых экспериментах гауссовская случайная величина с математическим ожиданием m=0 и дисперсией σ2=1 моделируется при помощи встроенной в Matlab процедуры. На рисунке 2.1 изображена функция плотности этой величины.



Рис. 2.1. График плотности распределения Гаусса

Распределение Лапласа со сдвигом β и коэффициентом масштаба α > 0 имеет плотность . В экспериментах рассматривается величина с β=0 и α=1. Величина с таким распределением моделируется как разность двух величин с экспоненциальным распределением. Каждая из этих величин, в свою очередь, вычисляется как логарифм равномерно распределенной на отрезке от 0 до 1 случайной величины, поделенный на - α: . График функции плотности распределения Лапласа представлен на рисунке 2.2.



Рис. 2.2. График плотности распределения Лапласа

Распределение Коши с коэффициентом масштаба γ > 0 и сдвигом x0 имеет плотность . Распределение Стьюдента с n степенями свободы имеет плотность , где Г - гамма-функция Эйлера, . Случайные величины с распределениями Коши и Стьюдента получаются при помощи функций, обратных их функциям распределения, в которые в качестве аргумента была подставлена величина с равномерным на отрезке от 0 до 1 распределением. Рассматривалось распределение Стьюдента с 2, 3, 5 и 13 степенями свободы, распределение Коши имеет параметры γ = 1 и x0 = 0. На рисунке 2.3 синими точками отмечена плотность распределения Коши, синей линией - плотность распределения Стьюдента с 2 степенями свободы, красной - с 3 степеням, зеленой - с 5 степенями и черной - с 13 степенями.



Рис. 2.3. Графики плотностей распределений Стьюдента с 2, 3, 5 и 13 степенями свободы и Коши

Плотность распределения Тьюки равна

,

где 0 < γ < 1 -доля зашумления выборки, σ12 > σ22. Величина с распределением Тьюки в экспериментах моделируется как смесь двух гауссовских случайных величин: с вероятностью (1-γ) величина имеет нормальное распределение с дисперсией σ22 и нулевым математическим ожиданием, а с вероятностью γ она имеет дисперсию σ12. Для реализации такой комбинации дополнительно используется равномерно распределенная на отрезке от 0 до 1 случайная величина. Если значение этой величины оказывается меньше (1-γ), то генерируется величина с меньшей дисперсией, иначе же генерируется величина с большей дисперсией. На рисунке 2.4 изображена плотность распределения Тьюки.



Рис. 2.4. График плотности распределения Тьюки

Плотность "двугорбого" распределения на основе двух гауссовских величин описывается формулой

,

где m - одна из двух симметричных мод распределения. В экспериментах случайная величина с таким распределением моделируется следующим образом: с вероятностью 0,5 величина имеет гауссовское распределение с математическим ожиданием m = 2 и дисперсией σ2 = 1, и с вероятностью 0,5 величина имеет математическое ожидание -m = -2 и дисперсию σ2 = 1. Для реализации комбинации, аналогично распределению Тьюки, используется вспомогательная случайная величина с равномерным распределением. На рисунке 2.5 изображен график плотности такого распределения.



Рис. 2.5. График плотности "двугорбого" распределения на основе комбинации двух гауссовских

Плотность треугольного распределения на отрезке от а до b равна

.

В экспериментах случайная величина с треугольным распределением на отрезке от -1 до 1 вычисляется как сумма двух случайных величин, каждая из них распределена равномерно на отрезке от -0,5 до 0,5. На рисунке 2.6 изображена плотность треугольного распределения.

ранговый регрессия асимптотический погрешность



Рис. 2.6. График плотности треугольного распределения

"Двугорбое" распределение на основе двух треугольных (на отрезках .[-b,-a] и [a,b]) имеет плотность

.

В экспериментах один треугольник располагается на отрезке [-1,0], а второй - на отрезке [0,1]. Такое распределение моделируется следующим образом: с вероятностью 0,5 генерируется сумма двух величин с равномерным распределением на отрезке от 0 до 0,5, иначе генерируется сумма двух величин с равномерным на отрезке от -0,5 до 0 распределением. График плотности такого распределения представлен на рисунке 2.7.



Рис. 2.7. График плотности "двугорбого" распределения на основе комбинации двух треугольных

Логистическое распределение с параметрами сдвига μ и масштаба s > 0 имеет функцию распределения вида . В экспериментах такая величина с параметрами μ = 0 и s = 1 моделируется при помощи функции, обратной функции распределения, в которую в качестве аргумента подставляется случайная величина с равномерным на отрезке от 0 до 1 распределением. На рисунке 2.8 изображен график плотности этого распределения.



Рис. 2.8. График плотности логистического распределения

Для проведения сравнительного анализа требуется построить регрессионные зависимости. В экспериментах рассматриваются модели с n = 50 наблюдениями и (m+1) = 3 параметрами, включая свободный член. Данные генерируются следующим образом:

· Сначала случайным образом генерируется матрица X с данными, равномерно распределенными на некотором отрезке. Матрица имеет n строк и m столбцов.

· К матрице X приписывается столбец из единиц для того, чтобы строить модели со свободным членом.

· Согласно рассматриваемому распределению генерируется n-мерный вектор-столбец погрешностей ε.

· Задаётся m-мерный вектор-столбец θ с реальными значениями параметров линейной регрессионной модели.

· Строится вектор наблюдений Y = Xθ + ε

Для каждой построенной регрессионной зависимости в экспериментах вычисляются ранговая, МНК- и МНМ-оценки вектора ее параметров.

Согласно Дж. Себеру, МНК-оценка вектора параметров линейной регрессионной модели имеет вид

.

Для построения приближенной МНМ-оценки используется метод симплексного поиска с целью минимизации функции потерь

.

Для построения приближенной ранговой оценки тем же методом проводится минимизация рассмотренной в предыдущей главе функции потерь

.

Поскольку данные генерируются случайно, то следует избегать влияния какого-то конкретного разброса данных на общую картину. Поэтому для одного и того же значения заданных параметров данные генерируются 1000 раз в проводимых экспериментах, каждый раз на основе этих данных строятся оценки параметров, вычисляются значения критерия качества оценок, и затем считается выборочное среднее качества оценок параметров регрессионных моделей для каждого метода. Критерием качества оценки в этом случае будет выступать сумма квадратов разностей истинного значения параметра и его оценки , где  - вектор с истинными значениями параметров,  - вектор с оценками параметров. Наилучшим будет тот метод, для которого выборочное среднее ошибки оценивания будет меньше. Результаты проведенных экспериментов для удобства сведены в таблицу 2.1. Из этих данных следуют 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, а так же "двугорбое" распределение на основе треугольного. Этот метод дает наихудшую оценку при распределении Лапласа, Коши, Тьюки и Стьюдента с менее чем 5 степенями свободы.

· 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 распределении Гаусса, распределении Стьюдента с 5 и более степенями свободы, "двугорбом" распределении на основе гауссовских величин, треугольном распределении, а так же "двугорбом" распределении на основе треугольного.

· Ранговый метод наиболее точен для оценивания параметров регрессионной модели с шумами, имеющими распределение Стьюдента со степенями свободы меньше 13 (но не с 1 степенью, т.к. это распределение Коши), логистическое распределение, а так же при распределении Тьюки.

Таблица 2.1

|  |  |  |  |
| --- | --- | --- | --- |
| для | рангового метода | МНК | МНМ |
| Стандартное нормальное распределение | 0,1959 | 0,1768 | 0,2677 |
| Распределение Лапласа | 0,2416 | 0,3332 | 0,2464 |
| Распределение Коши | 0,6909 | 16958,2274 | 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 |
| Распределение Тьюки (γ=0,1, σ12=100, σ22=1) | 0,2991 | 1,8910 | 0,3551 |
| Треугольное распределение | 0,0328 | 0,0271 | 0,0477 |
| Двугорбое распределение на основе комбинации треугольных | 0,0796 | 0,0482 | 0,2046 |
| Логистическое распределение | 0,5243 | 0,553 | 0,6847 |

Так же был проведен ряд дополнительных экспериментов с распределением Тьюки с различными значениями параметров: уровнем зашумления γ и значениями дисперсий σ12 и σ22. Результаты представлены в таблице 2.2. Из этих результатов следует, что ранговый метод дает наиболее точную оценку при умеренном значении дисперсий комбинируемых величин и небольшом значении уровня зашумления. При увеличении каждого из этих параметров в определенный момент МНМ-оценка становится более точной, МНК-оценка достаточно быстро теряет свою точность.

Таблица 2.2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Доля зашумления γ | Дисперсии σ12 и σ22 | Ошибка ранговой оценки | Ошибка МНК-оценки | Ошибка МНМ-оценки |
| 0,05 | σ12 = 10, σ22 = 1 | 0,2085 | 0,2411 | 0,285 |
| 0,05 | σ12 = 200, σ22 = 1 | 0,2154 | 1,9202 | 0,2911 |
| 0,05 | σ12 = 100, σ22 = 5 | 1,133 | 1,7767 | 1,4342 |
| 0,1 | σ12 = 10, σ22 = 1 | 0,2544 | 0,3317 | 0,3408 |
| 0,1 | σ12 = 200, σ22 = 1 | 0,2815 | 3,7429 | 0,3218 |
| 0,1 | σ12 = 100, σ22 = 5 | 1,2936 | 2,5161 | 1,5501 |
| 0,2 | σ12 = 10, σ22 = 1 | 0,3149 | 0,4624 | 0,3745 |
| 0,2 | σ12 = 200, σ22 = 1 | 0,4239 | 6,9605 | 0,4164 |
| 0,2 | σ12 = 100, σ22 = 5 | 1,6651 | 3,9654 | 1,7696 |

В приложении № 2 представлены коды для Matlab, с помощью которых генерировались случайные величины в экспериментах. В приложении № 3 представлен пример кода, вычисляющий результаты эксперимента для гауссовского распределения шумов в модели.

Итак, в этой главе были рассмотрены используемые в экспериментах распределения случайных величин, методы их моделирования и приведен алгоритм проведения экспериментов. Так же были представлены результаты экспериментов и сделаны выводы, касающиеся точности методов оценивания параметров моделей при различных распределениях шумов.

# Асимптотическая относительная эффективность

Вычисление АОЭ рангового метода по отношению к МНК и МНМ позволяет сделать выводы о том, какой метод лучше применять для оценки параметров в моделях с большим объемом выборки.

Согласно Т. Хеттманспергеру, АОЭ одного метода по отношению к другому определяется как корень степени p обратного отношения обобщенных дисперсий оценок параметров, полученных этими методами. Под обобщенной дисперсией вектора оценок параметров понимается определитель матрицы ковариаций оценок параметров, p - число параметров в модели. Если полученное число меньше единицы, то альтернативный метод эффективнее. Иначе более эффективным считается данный метод.

Ковариационная матрица МНК-оценки, согласно Дж. Себеру, имеет вид , где X - матрица плана, а σ2 - дисперсия шумов модели. В случае, когда предполагается, что вектор шумов имеет гауссовское n-мерное распределение с нулевым математическим ожиданием и ковариационной матрицей σ2 In (диагональная, на диагонали - дисперсии σ2), можно утверждать, что вектор  МНК-оценок параметров регрессии имеет гауссовское m-мерное распределение с математическим ожиданием θ - вектором реальных значений параметров, и ковариационной матрицей .

В книге Т. Хеттманспергера приведена следующая теорема:

Пусть  - любая точка, минимизирующая функцию D(Y-Xθ). Тогда, если вектор θ содержит истинные значения параметров и выполняются следующие предположения:

· в функции D(Y-Xθ) используются веса ,

· матрица [1X] (со столбцом из единиц) имеет полный столбцовый ранг,

· матрица n-1[1X]T[1X] сходится к положительно определенной матрице, и матрица n-1XTX сходится к положительно определенной матрице Σ, то вектор  по распределению сходится к величине, имеющей m-мерное гауссовское распределение с нулевым математическим ожиданием и ковариационной матрицей , где f(x) - плотность распределения шума. В статье Д. Полларда приведена теорема о распределении МНМ-оценок параметров регрессионной модели:

Пусть шумы εi независимы, одинаково распределены, с нулевой медианой и непрерывной, положительной функцией плотности f(.) в окрестности нуля. Пусть {xi} - детерминированная последовательность, для которой матрица  имеет положительно определенный квадратный корень Vn. Если  при , то вектор  имеет асимптотическое многомерное нормальное распределение с нулевым математическим ожиданием и единичной ковариационной матрицей.

Условия на {xi} выполняются, если существует положительно определенная матрица V, такая что . Тогда вектор  имеет m-мерное гауссовское распределение с нулевым математическим ожиданием и ковариационной матрицей .

Тогда АОЭ рангового метода по отношению к МНК выражается как

,

где σ2 - дисперсия шума, f(x) - плотность его распределения, p - число параметров модели.

АОЭ рангового метода по отношению к МНМ имеет вид

.

В таблице 3.1 приведены результаты вычислений АОЭ для различных распределений шумов регрессионной модели. Интегралы для распределений Стьюдента с различными степенями свободы, распределения Коши и логистического распределения были вычислены численно в среде Matlab, остальные были взяты аналитически.

Таблица 3.1

|  |  |  |
| --- | --- | --- |
|  | ОАЭ рангового метода к МНК | ОАЭ рангового метода к МНМ |
| Нормальное распределение | 3/π ≈ 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 ≈ 0,8889 | 4/3 ≈ 1,3333 |
| Логистическое распределение | π2/9 ≈ 1,0966 | 4/3 ≈ 1,3333 |
| Распределение Тьюки с γ = 0,1, σ12 = 100, σ22 = 1 | 7,2804 | 1,267 |

Из полученных данных следуют такие выводы:

· Ранговый метод уступает МНК в моделях с шумами, имеющими распределение Гаусса, Стьюдента с не менее чем 19 степенями свободы и треугольное распределение.

· Ранговый метод уступает МНМ в моделях с шумами, имеющими распределения Лапласа и Коши.

Итак, в этой главе были рассмотрены и приведены результаты вычисления АОЭ рангового метода по отношению к МНК и МНМ. Так же были сделаны выводы об эффективности рангового метода по отношению к рассматриваемым альтернативным ему методам.

# 3. Построение модели на реальных данных

Для построения линейной регрессионной модели на основе реальных данных при помощи рангового метода оценивания параметров был выбран достаточно известный набор данных "ирисы Фишера". Эти данные были собраны американским ботаником Эдгаром Андерсоном, они включают в себя измеренные в миллиметрах длину и ширину чашелистика и лепестка у 150 экземпляров цветка ириса - по 50 экземпляров каждого из трех видов: ирис щетинистый, ирис виргинский и ирис разноцветный.

В этой работе строится линейная регрессионная модель зависимости длины лепестка от длины и ширины чашелистика для вида ирис разноцветный. Такая модель имеет 2 регрессора и 50 наблюдений, для построения модели надо оценить при помощи рангового метода 3 параметра: свободный член и коэффициенты перед двумя регрессорами.

В результате запуска алгоритма, вычисляющего ранговую оценку параметров модели, получилось, что свободный член равен 0,3141, коэффициент перед длиной чашелистика равен 0,5429, а коэффициент перед шириной чашелистика равен 0,3571. На рисунках 4.1 и 4.2 представлены графики, изображающие построенную зависимость в виде плоскости и реальные наблюдения в виде кружков. Так же для данной модели были построены МНК- и МНМ-оценки параметров. Затем в данные были внесены изменения: у одного случайно выбранного наблюдения значение зависимой переменной было увеличено в 10 раз - как в случае ошибки с порядком. Ранговая оценка, МНК- и МНМ-оценки параметров были пересчитаны, их изменение d2 было измерено вычислением суммы квадратов разностей соответствующих компонент старого и нового векторов для каждой оценки параметров. Результаты для наглядности представлены в таблице 4.1.



Рис. 4.1. График линейной регрессии и рассеивание данных



Рис. 4.2. График линейной регрессии и рассеивание данных

Таблица 4.1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Оценка: | Ранговая оценка | | МНК-оценка | | МНМ-оценка | |
| Модель: | До изменений | После изменений | До изменений | После изменений | До изменений | После изменений |
| Θ0 | 0,3141 | 0,757 | -1,1089 | 21,0407 | 0,9023 | 0,903 |
| Θ1 | 0,5429 | 0,5899 | 0,578 | 1,5217 | 0,5574 | 0,5583 |
| Θ2 | 0,3571 | 0,2446 | 0,3394 | -2,3672 | 0,3108 | 0,3092 |
| d2 | 0,211 | | 498,8213 | | 3,7607\*10-6 | |

Таким образом, можно сделать вывод, что наилучшим образом на выбросы в данных реагирует МНМ-оценка, наихудшим - МНК-оценка. Притом ранговая оценка показывает вполне удовлетворительные результаты.

На рисунке 4.3 изображен график линейной регрессии с ранговыми оценками параметров модели после внесения изменений в данные, на которых так же видно точку-выброс.



Рис. 4.3. График линейной регрессии после внесения изменений в данные

На рисунке 4.4 изображено изменение МНК-оценок параметров линейной регрессии в результате изменения данных: более бледная плоскость соответствует модели без внесенных выбросов, а более яркая - с выбросом.



Рис. 4.4. Изменение МНК-оценок в результате изменения данных

В приложении № 4 представлен код для среды Matlab, при помощи которого была вычислена оценка параметров регрессии и построены графики, а так же проведен эксперимент на реальных данных.

В данной главе был рассмотрен пример применения рангового метода оценивания параметров линейной регрессионной модели к реальным данным и была построена линейная модель зависимости длины лепестка цветка ириса от длины и ширины его чашелистика. Так же было рассмотрено и сравнено влияние выбросов в данных на ранговую, МНК- и МНМ-оценки параметров модели с реальными данными.

# Заключение

В данной работе был рассмотрен ранговый метод оценивания параметров регрессионной модели, был предложен способ приближенного вычисления ранговой оценки параметров линейной регрессионной модели посредством минимизации соответствующей функции потерь при помощи одного из численных методов нахождения точки минимума функции многих переменных - метода симплексного поиска.

Так же в этой работе с помощью компьютерного моделирования был проведен численный сравнительный анализ рангового метода, МНК и МНМ. Результатом этого анализа стали следующие выводы о применимости рассматриваемых методов при различных распределениях шумов в регрессионных моделях с небольшим числом наблюдений:

· МНК рекомендуется применять в случае, когда есть основания полагать, что шумы в модели имеют распределение Гаусса, распределение Стьюдента с 13 и более степенями свободы, треугольное распределение, "двугорбое" распределение на основе гауссовских случайных величин или величин с треугольным распределением.

· МНМ лучше применять, когда предполагается, что шумы в модели имеют распределение Лапласа или распределение Коши, а так же распределение Тьюки с большими значениями дисперсий комбинируемых величин и высокой долей зашумления.

· Ранговый метод точнее оценит параметры модели, если шумы имеют распределение Лапласа, распределение Стьюдента с менее чем 13 степенями свободы (кроме случая с одной степенью свободы), логистическое распределение или распределение Тьюки с умеренными значениями дисперсий комбинируемых величин и уровня зашумления.

· Поскольку ни в одном эксперименте ранговый метод не показал худших результатов, в случае отсутствия каких-либо предположений о распределении шумов в модели рекомендуется применять этот метод.

Помимо этого, сравнительный анализ был дополнен вычислением АОЭ рангового метода по отношению к МНК и МНМ. На основании полученных результатов были сделаны следующие выводы:

· Если предполагается, что шумы в модели имеют распределение Лапласа, Коши, Стьюдента с не более чем 18 степенями свободы, логистическое распределение, то вместо МНК лучше использовать ранговый метод.

· Ранговый метод более эффективен, чем МНМ, при нормальном распределении шумов, распределении Стьюдента с не менее 2 степенями свободы, треугольном распределении и логистическом распределении.

Также был приведен пример применения рангового метода для оценивания параметров линейной регрессионной модели, построенной на реальных данных, и вычислены ранговые оценки параметров этой модели. Был смоделирован выброс в реальных данных и оценен эффект, оказанный на оценки параметров модели, в результате чего сделан вывод: МНМ дает более устойчивую к выбросам в данных оценку, чем ранговый метод, и намного более устойчивую, чем МНК-оценка.

Основной вывод по результатам проведенных вычислений и экспериментов таков: ранговый метод является надежной альтернативой МНК и МНМ в задаче оценивания неизвестных параметров линейной регрессионной модели.

# Список литературы

) Хеттманспергер Т. Статистические выводы, основанные на рангах. -М.: Финансы и статистика, 1987.

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

) Ю. И. Ингстер и др., Основные алгоритмы численного анализа. -СПб.: СПбГЭТУ "ЛЭТИ", 2009. -URL: http://www.hse.ru/pubs/lib/data/access/ram/ticket/47/14003652778464aba14a844ad5813be5adab26198a/matlab\_stat\_s3.pdf (дата доступа 03.02.2014).

4) Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions/ J.C. Lagarias, J. A. Reeds, M. H. Wright, P. E. Wright // SIAM Journal of Optimization, Vol. 9, Number 1, pp.112-147, 1998. -URL: http://jasoncantarella.com/downloads/SJE000112.pdf (дата доступа 07.05.2014).

) Louis A. Jaeckel. Estimating Regression Coefficients by Minimizing the Dispersion of the Residuals// The Annals of Mathematical Statistics, Vol. 43, Number 5 (1972), pp. 1449-1458, 1972. -URL: http://projecteuclid.org/download/pdf\_1/euclid.aoms/1177692377 (дата доступа 27.04.2014).

6) Себер Дж. Линейный регрессионный анализ. -М.: Мир, 1980.

7) D. Pollard. Asymptotics for lest absolute deviation regression estimators// Econometric Theory, 7, pp. 186-199, 1991. -URL: http://www.math.pku.edu.cn/teachers/xirb/Courses/QR2013/Pollard91ET.pdf (дата доступа 12.05.2014).

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

# Приложение № 1

Функция, вычисляющая приближенное значение вектора оценки параметров регрессионной модели с матрицей данных X (без столбца единиц) и вектором значений зависимой переменной Y:

function rkT = rankreg(X,Y)

[n,m]=size(X);=ones(m,1);

[rkT,D]=fminsearch(@(T) Dfunc(Y,X,T), T0);=median(Y-X\*rkT);=[a;rkT];

end

Отдельным файлом прописывается функция D:

function D = Dfunc(Y, X, T)=Y-X\*T;

[R, trash]=tiedrank(A);=length(Y);=0;i=1:N=D+(A(i)\*sqrt(12)\*(2\*R(i)-(N+1))/(2\*N+2));

end

# Приложение № 2

1) Код, генерирующий 50 случайных величин с распределением Лапласа:

n=50;=-log(rand(n,1));=-log(rand(n,1));=z1-z2;

2) Код, генерирующий 50 случайных величин с распределением Коши:

=50;=tan(pi\*rand(n,1) - pi/2);

%либо вместо последней строчки

%ep=tinv(rand(n,1),1);

3) Код, генерирующий 50 случайных величин с распределением Стьюдента с N степенями свободы:

=50;=2;=tinv(rand(n,1),N);

4) Код, генерирующий 50 случайных величин с распределением Тьюки:

n=50;=rand(n,1);=randn(n,1);=sqrt(100)\*randn(n,1);=find(p>=0.9);

ep(I)=eps(I);

5) Код, генерирующий 50 случайных величин с двугорбым распределением на основе двух гауссовских:

n=50;=rand(n,1);=randn(n,1)+2;=randn(n,1)-2;=find(p>=0.5);

ep(I)=eps(I);

6) Код, генерирующий 50 случайных величин с треугольным распределением:

n=50;=rand(n,1)-0.5;=rand(n,1)-0.5;

ep=u1+u2;

7) Код, генерирующий 50 случайных величин с двугорбым распределением на основе двух треугольных:

n=50;=rand(n,1);=rand(n,1)\*0.5;=rand(n,1)\*0.5;=u1+u2;=-rand(n,1)\*0.5;=-rand(n,1)\*0.5;=v1+v2;=find(p>=0.5);(I)=eps(I);

8) Код, генерирующий 50 случайных величин с логистическим распределением:

n=50;=-log((1./rand(n,1)) -1);

# Приложение № 3

Пример кода, вычисляющий результаты эксперимента по сравнинию ранговой оценки, МНК- и МНМ-оценок параметров для модели с гауссовским распределением шумов:

=2;=50;=[2;5;-3];=0;=0;=0;k=1:1000=ones(n,1);=5\*rand(n,m);=[X1,Xr];=randn(n,1);=X\*reT+ep;=((X'\*X)^-1)\*X'\*Y;=ls+sum((lsT-reT).^2)/1000;=rankreg(Xr,Y);=rk+sum((rkT-reT).^2)/1000;

[laT,S]=fminsearch(@(T) LAVal(Y,X,T), ones(m+1,1));=la+sum((laT-reT).^2)/1000;

rk

ls

la

Отдельным файлом для вычисления приближенной МНМ-оценки параметров прописывается функция потерь для МНМ:

S = LAVal(Y, X, T)=sum(abs(Y-X\*T));

end

# Приложение № 4

Программный код, вычисляющий оценку параметров линейной регрессии на используемых данных "ирисы Фишера" и строящий график, изображающий полученную зависимость и рассеивание данных:

iris.dat=iris(51:100,3);=iris(51:100,1);=iris(51:100,2); Xc=[X1 X2];=rankreg(Xc,Y)

[x1 x2]=meshgrid(0:0.1:75,0:0.1:40);=rkT1(1) + rkT1(2)\*x1 + rkT1(3)\*x2;(x1,x2,y,'FaceAlpha',0.6,'EdgeColor','none')on(X1,X2,Y,'ko')=[ones(50,1) Xc];=((X'\*X)^-1)\*X'\*Y

[laT1,S]=fminsearch(@(T) LAVal(Y,X,T), ones(3,1));=fix(50\*rand(1));(it)=5\*Y(it);=rankreg(Xc,Y)=((X'\*X)^-1)\*X'\*Y

[laT2,S]=fminsearch(@(T) LAVal(Y,X,T), ones(3,1));=sum((rkT1-rkT2).^2)=sum((lsT1-lsT2).^2)=sum((laT1-laT2).^2)