О применении метода ступенчатого оценивания при построении описательных моделей

Т.Е. Родионова[a], Г.Р. Кадырова[a]

[a] *Ульяновский государственный технический университет, 432027, ул. Северный Венец, 32, Ульяновск, Россия*

Аннотация

Рассматриваются математические регрессионные модели, используемые для описания технического объекта или процесса. Обсуждаются нарушения предположений регрессионного анализа, возникающие при обработке различных практических данных. Описывается метод ступенчатого оценивания, позволяющий преодолеть негативное влияние эффекта мультиколлинеарности. Приводятся модели, полученные при анализе лазерных и радиоинтерферометрических наблюдений, данных по физико-химическим показателям водоисточника. Сравниваются полученные оценки с результатами метода наименьших квадратов и методом пошагового оценивания. Выбор оптимальной модели производится по критериям минимума смещения. Доказывается возможность применения метода ступенчатого оценивания для построения описательных моделей.

Ключевые слова: модели описания; мультиколлинеарность; метод наименьших квадратов; регрессионный анализ; методы структурной идентификации; ступенчатое оценивание.

1. Введение

Рассмотрим описательную (параметрическую) регрессионную модель, которая претендует на описание причинно-следственных связей явления. Такая модель должна содержать как можно больше оцениваемых параметров. Пусть математическая модель имеет вид:

$$\mathbf{Y} = \eta(X, \beta) + \vec{\varepsilon} \qquad (1)$$

где Y — зависимая переменная; $\mathbf{X} = \left(x_0 x_1 \ldots x_{p-1}\right)^{\mathrm{T}}$ — вектор независимых переменных; $\vec{\beta} = \left(\beta_0 \beta_1 \ldots \beta_{p-1}\right)^{T}$ — вектор неизвестных параметров, определяемых по результатам экспериментов; $\vec{\varepsilon}$ — вектор случайных ошибок. Входящие в модель переменные X и Y являются результатами пассивного эксперимента, т.е. измеренными или вычисленными значениями. Вектор β в модели (1) предполагается не меняющимся со временем, т.е. математическая модель считается стационарной по параметрам [1,13].

На практике для оценки параметров такой математической модели используются методы регрессионного анализа и, в частности метод наименьших квадратов (МНК). При этом приходится учитывать возможные нарушения условий применения данного метода. Применение подхода регрессионного моделирования в рассматриваемой задаче подразумевает исследование и выбор оптимальных методов получения наилучших линейных оценок параметров и проверку эффективности получаемой модели по соответствующим критериям [3,12].

Можно выделить следующие нарушения применения стандартного МНК при решении практических задач:

- модели содержат незначимые (шумовые) слагаемые;
- параметры модели коррелируют друг с другом (эффект мультиколлинеарности);
- остатки также могут быть дополнительно искажены автокорреляцией и другими систематическими ошибками.

Вообще выбор способа адаптации к нарушениям условий РА-МНК зависит от типа исследуемой модели [9,14]. В данном случае, объектом внимания являются непосредственно параметры модели, а не результаты прогнозирования с ее использованием. Конечной целью адаптации являются наилучшие линейные оценки (НЛО), т.е. оценки, не обремененные заметными систематическими и случайными ошибками. Такими они могут быть, по меньшей мере, при статистической значимости и, главное, независимости друг от друга параметров модели. Очевидно, что адаптация к первому упомянутому нарушению с помощью простого устранения незначимых слагаемых затруднительна по очень простой причине: часть из них может быть взаимосвязана со значимыми.

Для преодоления эффекта мультиколлинеарности и снижения количества незначимых слагаемых в описательных моделях предлагается использовать метод ступенчатого оценивания (МСО).

2. Описание метода ступенчатого оценивания

В рассматриваемом методе поэтапное разбиение осуществляется не по отдельным переменным (как в пошаговой регрессии), а по их группам, последовательно формируемым в виде подмножеств переменных с незначимыми парными коэффициентами корреляции r_{ij}. Это означает, что группы формируются не по степени коррелированности с последовательно образуемыми откликами, а в виде отдельных структур в почти ортогональном базисе [2,4]. Краткое описание алгоритма метода ступенчатого оценивания:

1. Вычисляется оценка параметров исходной модели с использованием одной из вычислительных схем МНК

$$\hat{\Delta} = (X^T X)^{-1} X^T, \quad (2)$$

ее ковариационная матрица

$$D(\hat{\Delta}) = (X^T X)^{-1} \sigma^2, \quad (3)$$

и различные статистики, позволяющие оценить статистическую ценность каждого слагаемого и модели в целом, включая значения t- статистик и коэффициентов парных корреляций r_{ij}.

2. Путем сравнения значений r_{ij} формируется первое подмножество поправок Δ_1, обладающих незначимыми значениями r_{ij}.

3. Оцениваются параметры ортогональной структуры

$$Y = X_1 \Delta_1, \quad (4)$$

вычисляется первый вектор остатков

$$e_1 = Y_1 - \hat{Y}_1, \quad (5)$$

который рассматривается как очередной вектор отклика для формирования следующего подмножества поправок из множества оставшихся.

4. Этапы 2, 3 повторяются до завершения процесса формирования подмножеств $\Delta_1,\Delta_2,\ldots\Delta_k$.

Для улучшения качества получаемых оценок в вычислительную схему включено ортогональное преобразование Хаусхолдера. При этом численная устойчивость, характерная для ортогональных преобразований, сочетается с гибкостью, позволяющей легко приспосабливаться к последовательному накапливанию данных, что очень важно при решении задач большой размерности. Дополнительно сокращается требование к памяти компьютера, увеличивается скорость выполнения и точность; следует отметить защищенность от «машинных нулей» и переполнения. С помощью первой стратегии этого алгоритма сделана попытка раздельно оценить взаимосвязанные регрессоры за счет оценивания их на разных стадиях метода (МСО1). В качестве недостатка такого алгоритма можно отметить, что в число оцениваемых параметров попадают и незначимые по статистике Стьюдента.

Вторая стратегия данного метода включает в итоговую модель только те регрессоры, которые оказались значимыми по t-критерию на каждой стадии работы (МСО2). Она наиболее близка к методу пошаговой регрессии, но за счет того, что расчет идет по отдельным подмножествам, позволяет оценить во много раз больше параметров исходной модели, так как регрессор незначимый на одной стадии может оказаться значимым на последующих. Это очень важно для задачи параметрического оценивания, где необходимо получить как можно более полную модель. Недостаток этой стратегии состоит в отсутствии анализа взаимозависимости включаемых параметров.

Третья стратегия представляет собой совокупность первой и второй. Отбор во множество оцениваемых параметров идет сразу по двум признакам: значимости и ортогональности (МСО3).

3. Описание исходных данных и выявленных нарушений предположений регрессионного анализа

Для апробации рассматриваемого метода оценивания параметров математической модели использовались следующие данные: данные по лазерной локации Луны; РСДБ-наблюдения внегалактических источников; результаты физико-химического контроля питьевой воды.

Обрабатываемые двухгодичные светолокационные данные получены при использовании уголкового отражателя КК “Аполлон-15” в обсерватории Мак Дональд (Техас, США) с августа 1971 по ноябрь 1973 года (всего 549 наблюдений). Исходные данные в виде коэффициентов условных уравнений были подготовлены сотрудниками Института теоретической астрономии (ИТА АН СССР) .

Рассматриваемые РСДБ-наблюдения представляют собой 1262 условных уравнения для определения 203 поправок к постоянным теории орбитального движения и вращения Земли. К этим данным добавлены 4 уравнения связи, определяющие равенство нулю параллельного переноса земной системы координат и поворота земной и небесной систем координат, а также ограничения, накладываемые на вектора баз. Данные для расчетов подготовлены профессором В.Е.Жаровым (ГАИШ МГУ) [6,7].

В качестве третьего примера рассматривались результаты физико-химического контроля питьевой воды (отклики $y_1 - y_7$) и воды водоисточника (оцениваемые параметры $x_1 - x_8$), используемых для очистки воды [8,10]. Исходный файл представляет собой результаты контроля указанных параметров за год.

В качестве первой проблемы при обработке данных можно назвать проблему достаточности объема наблюдений. В рассматриваемой работе мы имеем дело со следующей ситуацией: лазерные данные по Луне для определения 24 неизвестных поправок содержат 549 условных уравнений, т.е. превосходят количество оцениваемых поправок в 22 раза; по радиоинтерферометрическим данным по Земле мы имеем соотношение – 203 неизвестные поправки и 1289 условных уравнения (включая 27 уравнений связи), таким образом, количество наблюдений всего в 6 раз превосходит количество параметров; по данным водоисточника для определения 8 параметров воды имеется 365 наблюдений (количество

наблюдений в 45 раз превосходит количество параметров). В регрессионном анализе между числом определяемых параметров p и количеством наблюдений n в эксперименте должно выполняться соотношение n =5p÷15p.

Исследование данных начиналось с анализа модели полученной методом множественной регрессии. Исследовалось количество незначимых параметров модели и матрица парных коэффициентов корреляции. Для этого использовался пакет СПОР, позволяющий получать регрессионные модели и определять их меры качества [5,11,15].

Наличие аномальных наблюдений в выборке можно считать второй проблемой стоящей перед исследователем. В рассматриваемых исходных данных из файла с лазерными наблюдениями Луны были удалены 4 аномальных наблюдений, в файле с РСДБ-наблюдениями было обнаружено 18 выбросов, а в данных по водоисточнику для разных откликов количество выбросов колеблется от 1 до 4.

Следующая проблема напрямую связана с матрицей исходных данных: среди аргументов (переменных) не должно быть линейно зависимых. Однако, на практике, это предположение соблюдается не всегда. При нарушении этого условия, между анализируемыми переменными существует линейная функциональная или статистическая связь. Это явление называется мультиколлинеарностью и имеет весьма отрицательные последствия для оценивания коэффициентов регрессии. В вычислительной математике этим понятиям соответствует вырожденность и плохая обусловленность матрицы X^TX, т.е. для последней не существует $(X^TX)^{-1}$ и определитель её близок к нулю. Последствия этого нарушения особенно серьезны для моделей, оцениваемые параметры которых подлежат физической интерпретации. Один из способов решения проблемы мультиколлинеарности может заключаться в том, что в уравнении должны находиться только некоррелированные друг с другом члены.

При анализе радиоинтерферометрических данных были выявлено, что матрица коэффициентов корреляции содержала 76 коэффициентов, превышающих по модулю 0.5. Из них 30 значений коэффициентов больше 0.95, что свидетельствует о практически линейной взаимосвязи между оцениваемыми параметрами. При исследовании данных по водоисточнику для каждого из рассматриваемых откликов $y_1 - y_7$, было выявлено от 1 до 3 коррелирующих параметров модели . Стоит также отметить, что в рассматриваемых математических моделях данные о факторах и об отклике имеют разный физический смысл и разные физические размерности. Это вызывает вычислительные неудобства, поскольку приходится работать как с очень большими, так и с очень маленькими числами, что может привести к вычислительным ошибкам.

Таким образом, наличие в полученных моделях незначимых слагаемых, а также наличие взаимной корреляции между оцениваемыми параметрами аномальных наблюдений позволяет сделать вывод о нарушениях предположений регрессионного анализа.

4. Применение метода ступенчатого оценивания для адаптации к выявленным нарушениям

Для устранения эффекта мультиколлинеарности и наличия в моделях незначимых параметров был применен описанный выше метод ступенчатого оценивания. Далее приведены результаты использования МСО для обработки различных наборов данных. Основная задача при создании описательных моделей – определение максимального числа параметров с наивысшей точностью. Для лазерных данных применение метода ступенчатой ортогонализации (МСО1) позволило оценить все параметры модели. При выборе только значимых параметров метода ступенчатого оценивания (МСО2) были получены оценки 10 поправок, а алгоритм МСО3 дала оценки для 9 поправок. Метод пошаговой регрессии, который был использован для сравнения, позволил оценить только 9 параметров из 24 возможных.

Для радиоинтерферометрических данных: методом пошаговой регрессии была получена модель, содержащая оценки 6 значимых параметров из 203 возможных; по стратегии МСО1 метода ступенчатого оценивания были определены 188 поправок, стратегия МСО2 дала оценки 136 поправок, а стратегия МСО3 - 51 поправку.

Таблица 1. Номера параметров, вошедших в модель для разных схем обработки

Отклик	ПР	МСО1	МСО2	МСО3
y_1	1, 2, 3, 5	3, 4, 5, 6, 7, 8	-	-
y_2	2, 5, 6, 7, 8	3, 4, 5, 6, 7, 8	2, 4, 6, 7	2, 3, 6, 7
y_3	3, 5, 7	3, 4, 5, 6, 7, 8	1, 2, 4, 6, 7	1, 2, 3, 4, 6
y_4	2, 3, 5, 6, 7, 8	3, 4, 5, 6, 7, 8	1, 2, 3, 4, 6, 7,8	4, 6, 8
y_5	1, 2, 7, 8	3, 4, 5, 6, 7, 8	1, 2, 7	1, 2, 3, 4, 5, 6, 7, 8
y_6	2, 3, 4, 5, 6, 7	3, 4, 5, 6, 7, 8	4, 6, 7	-
y_7	2, 5, 7	3, 4, 5, 6, 7, 8	2, 5, 6, 7	5, 6, 7

В таблице 1 приведены наборы параметров, вошедшие в модели, полученные разными вычислительными схемами (ПР – пошаговая регрессия) в рамках обработки данных по водоочистке. Из таблицы видно, что метод ступенчатого оценивания (стратегия МСО1 – выбор только ортогональных параметров) позволяет оценить больше параметров модели чем пошаговая регрессия, что очень важно для описания технологического процесса. Для рассматриваемого набора данных стратегии МСО2 (выбор на каждом шаге только значимых) и МСО3 (выбор на каждом шаге значимых и одновременно ортогональных параметров) не позволили получить модели лучше пошаговой регрессии. В таблице показана структура модели, какие из восьми регрессоров значимы и входят в состав модели. Приведенные данные

позволяют сделать вывод о том, что для различных выборок в модель ПР вошли разные параметры, при этом ни в одну из моделей не вошли оба управляемых параметра x_7 и x_8. В модель, полученную стратегией МСО1 для всех показателей качества функционирования объекта $y_1 - y_7$, набор показателей совпадает и практически во всех случаях x_7 и x_8 значимы [16].

Сравнивая оценки для одинаковых параметров, полученные различными методами оценивания можно сделать вывод о том, что мы получили достаточно близкие друг к другу значения. Если взять за эталон значения полученные методом пошаговой регрессии, то очень малое количество оценок, полученных другими методами, сильно отличается от эталона. Рассматривая соотношение стандартных ошибок приведенных оценок, полученных разными методами оценивания, что точность оценивания неизвестных параметров в рассматриваемых методах ПР и МСО1 практически совпадает. Таким образом, можно сделать вывод о применимости полученных моделей для описания данного технологического процесса.

Следующий этап исследования - задача выбора наилучшей описательной модели. При её решении следует учитывать, что внутренние критерии, т.е. критерии, не использующие никакой дополнительной информации, при наличии помех не могут решить задачу выбора наилучшей описательной модели, какими являются описательные модели. При использовании внешних мер следует очень серьезно подходить к разбиению исходной выборки на две части. Необходимо учитывать физический смысл и время наблюдения, так как исходные данные – это объединение нескольких выборок. Предлагается выбирать модель по критерию минимума смещения - непротиворечивости, по которому требуется, чтобы модели, полученные по обучающей выборке, возможно меньше отличались от моделей, полученных по тестовой выборке. Анализируя полученные результаты обработки радиоинтерферометрических, лазерных наблюдений и данных по водоочистке можно сделать вывод о том, что методы ступенчатого оценивания являются эффективными.

5. Заключение

Проведенные численные эксперименты позволяют сделать следующие выводы: - метод ступенчатого оценивания позволяет оценить большее количество параметров модели; - оценки метода ступенчатого оценивания близки к оценкам пошаговой регрессии. Таким образом, метод ступенчатого оценивания можно использовать для оценок параметров математической модели, а также для описания технических объектов и технологических процессов. Анализируя полученные значения критериев минимума смещения, для указанных наблюдений (как радиоинтерферометрических и лазерных наблюдений, так и данных по водоочистке), можно сделать вывод о том, что методы ступенчатого оценивания являются эффективными и позволяют с достаточной точностью описать исследуемый объект.

Литература

[1] Валеев, С.Г. Система поиска оптимальных регрессий: учебное пособие / С.Г. Валеев, Г.Р. Кадырова. – Казань : ФЭН, 2003. – 160 с.

[2] Валеев, С.Г. Метод ступенчатой ортогонализации базиса и его применение при решении задач МНК / С.Г. Валеев, Т.Е. Родионова // Изв. вузов. Геодезия и аэрофотосъемка, 2003. –№ 6. –С.3–14.

[3] Валеев, С.Г. Анализ методов оценки параметров при мультиколлинеарности переменных / С.Г. Валеев, Т.Е. Родионова // Известия Вузов. Серия: Геодезия и аэрофотосъемка. 1999, –№5. –С.20-28

[4] Валеев, С.Г. Программное обеспечение для решения задач структурно-параметрического оценивания при обработке данных / С.Г. Валеев, Т.Е. Родионова .// Известия Вузов. Серия: Геодезия и аэрофотосъемка. 2004, –№1. –С.25-34.

[5] Валеев, С.Г. Автоматизированная система для решения задач метода наименьших квадратов / С.Г. Валеев, Г.Р.Кадырова // Известия Вузов. Сер.: Геодезия и аэрофотосъемка. – 1999. – № 6. – С. 124–130.

[6] Валеев, С.Г. Методика статистической обработки РСДБ- наблюдений / С.Г. Валеев, Т.Е. Родионова, В.Е. Жаров // Изв. вузов. Геодезия и аэрофотосъемка, 2008. –№ 1. –С.13–18.

[7] Валеев, С.Г. Вычислительные эксперименты по обработке РСДБ- наблюдений / С.Г. Валеев, Т.Е. Родионова, В.Е. Жаров // Изв. вузов. Геодезия и аэрофотосъемка, 2008. –№ 2. –С.94–100.

[8] Родионова, Т.Е. Применение адаптивного регрессионного моделирования для описания функционирования технического объекта / Т.Е. Родионова // Известия Самарского научного центра Российской академии наук. 2014. –Т. 16. № 6-2. –С. 572-575.

[9] Кадырова, Г.Р. Оценка и прогнозирование состояния технического объекта по регрессионным моделям регрессий / Г.Р.Кадырова // Автоматизация процессов управления. – 2015. – № 4(42). – С. 90–95.

[10] Родионова, Т.Е. Статистические методы оценки показателей качества питьевой воды / Т.Е. Родионова, В.Н. Клячкин // Доклады АН ВШ РФ №2-3 (23-24) апрель-сентябрь 2014 . –С.101-110

[11] Валеев, С.Г. Программная система поиска оптимальных регрессий / С.Г. Валеев, Г.Р. Кадырова, А.А. Турченко // Вопросы современной науки и практики. Сер. Технические науки. – 2008. – № 4(14), т. 2. – С. 97–101.

[12] Кадырова, Г.Р. Модификация метода пошаговой регрессии для получения математических моделей прогноза поведения объекта / Г.Р. Кадырова // Автоматизация процессов управления. – 2016. – № 3(45). – С. 65–70.

[13] Валеев, С.Г. Последовательная ортогонализация базиса в задачах метода наименьших квадратов / С.Г. Валеев, Т.Е. Родионова // Вестник Ульяновского государственного технического университета. – 1999. –№1(6). –С.4-9.

[14] Кадырова , Г.Р. Программная система поиска оптимальных регрессионных моделей прогноза / Г.Р. Кадырова // Путь науки. – 2014.– № 7 (7). – С. 10–11.

[15] Кадырова, Г.Р. Система поиска оптимальной модели. Состояние дел и перспективы развития / Г.Р. Кадырова // Потенциал современной науки. – 2015. – № 4 (12). – С. 8–10.

[16] Родионова, Т.Е. Сравнение регрессионных моделей показателей качества питьевой воды / Т.Е. Родионова // Материалы 3-й научно-практической internet-конференции 20-21 февраля 2014 Тольятти «Междисциплинарные исследования в области математического моделирования и информатики» – С. 159-162.