Московский государственный технический университет
имени Н.Э. Баумана

Факультет «Фундаментальные науки»

Кафедра «Вычислительная математика и математическая физика»

Ю.В. Коновалов

СТАТИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ С ИСПОЛЬЗОВАНИЕМ РЕГРЕССИОННОГО АНАЛИЗА

Электронное учебное издание

Методические указания к выполнению курсовой работы по дисциплине «Компьютерное и статистическое моделирование»

Москва
(С) 2013 МГТУ им. Н.Э. БАУМАНА

УДК 519.22

Рецензент: профессор, д.т.н. Н.И. Сидняев

Коновалов Ю.В.

Статистическое моделирование с использованием регрессионного анализа: Методические указания к выполнению курсовой работы по дисциплине «Компьютерное и статистическое моделирование» / Ю.В. Коновалов. – М: Изд-во МГТУ им. Н.Э. Баумана, 2013. - 73 с.

Разработаны методические указания по получению качественных хорошо интерпретируемых регрессионных моделей. Указанные методические приемы используются на практике при применении регрессионного анализа. Главные рассматриваемые вопросы: выбор модели; интерпретация регрессионных моделей; анализ качества модели; управление выбором модели.

Для студентов направления подготовки "Математика и компьютерные науки", специальности "Прикладная математика", а также студентов технических специальностей.

Рекомендовано Учебно-методической комиссией НУК «Фундаментальные науки» МГТУ им. Н.Э. Баумана

Электронное учебное издание

Коновалов Юрий Викторович

СТАТИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ С ИСПОЛЬЗОВАНИЕМ РЕГРЕССИОННОГО АНАЛИЗА

© 2013 МГТУ имени Н.Э. Баумана

ОГЛАВЛЕНИЕ

Введение ……. 5

1. Цели и задачи методических указаний ……... 6

2. Основные понятия регрессионного анализа …….. 7

2.1. Последовательность действий при построении регрессионной модели …… 7

2.2. Набор эффектов. Вид модели …… 8

3. Предварительный анализ экспериментальной информации …… 10

3.1. Формализация априорной информации ……... 10

3.2. Графический анализ экспериментальных данных …….. 11

3.3. Выбор вида модели ……. 12

4. Анализ качества модели ……... 15

4.1. Проверка гипотезы о значимости параметров модели …… 16

4.2. Проверка согласия модели с экспериментальными данными (адекватности модели) …….. 16

4.2.1. Применение критерия Фишера для проверки гипотезы об адекватности модели …… 17

4.2.2. Дисперсия воспроизводимости …….. 18

4.2.3. Оценка дисперсии воспроизводимости ……. 21

4.2.4. Модель неадекватна ……. 24

4.3. Анализ остатков …….. 25

4.3.1. Проверка нарушения основных предположений …….. 25

4.3.2. Выбросы …… 27

4.4. Влиятельные наблюдения …….. 30

4.5. Мультиколлинеарность …….. 32

4.6. Прогноз значения откликов в области эксперимента ……. 33

4.7. Другие способы проверки качества модели ……. 36

4.7.1. Метод “складного ножа” ……. 37
4.7.2. Метод "перепроверки" …… 38
4.7.3. Метод бутстреп …… 39
4.8. Выводы ……. 39
5. Интерпретация регрессионной модели ……... 40
5.1. Типовая интерпретация полиномиальной модели второго порядка ……... 40
5.1.1. Выделение факторов, оказывающих наибольшее влияние на отклик ……. 40
5.1.2. Оценка характера, знака и степени влияния факторов на отклик ……. 41
5.1.3. Прогноз значения отклика для любых значений факторов . 42
5.1.4. Поиск и анализ оптимальных значений отклика. Решение компромиссных задач …… 45
5.1.5. Графическое представление модели …….. 48
5.2. Концепция хорошо интерпретируемой модели ……... 49
6. Управление выбором модели …….. 52
6.1. Использование центрирования факторов ……. 53
6.2. Использование данных “пассивного” эксперимента …….. 56
7. Расширения классического регрессионного анализа …… 57
7.1. Основные предположения в регрессионном анализе …….. 57
7.2. Мультиколлинеарность …….. 58
8. Задания для самостоятельной работы ……. 63
9. Вопросы для самоконтроля …….. 64
Заключение …… 65
Литература ……. 66
Приложение 1. Порядок построения и интерпретации регрессионных моделей ……... 67
Приложение 2. Бланки для формализации априорной информации ... 70

ВВЕДЕНИЕ

Регрессионный анализ - основной метод современной математической статистики. Его популярность объясняется следующими причинами:

- относительная простота регрессионных моделей и соответствующего математического аппарата;
- богатство интерпретации регрессионных моделей;
- применимость регрессионного анализа практически к любым экспериментальным данным (типа прямоугольных таблиц, содержащих зарегистрированные значения независимых и зависимых переменных);
- большая потребность в статистической обработке массивов данных (как с целью свертки, так и для извлечения из них дополнительной информации).

Литература по регрессионному анализу очень обширна. Лучшей из книг по практическому применению регрессионного анализа на русском языке является, конечно, [1]. Много ценных конкретных рекомендаций содержится в [2]. На более подготовленного читателя рассчитаны [3-6].

Многие вычислительные процедуры в регрессионном анализе автоматизированы, но основные практические аспекты применения регрессионного анализа еще не формализованы. Дело в том, что многие ключевые моменты построения и применения регрессионных моделей (в первую очередь, собственно, выбор вида модели; интерпретация модели; проблемы выбросов и влиятельных наблюдений и т.д.) требуют не только чисто статистического анализа экспериментальных данных, но и учета трудно формализуемой содержательной информации, относящейся к изучаемой задаче.

В данной работе предлагаются методические рекомендации для практического применения регрессионного анализа.

1. ЦЕЛИ И ЗАДАЧИ МЕТОДИЧЕСКИХ УКАЗАНИЙ

Целью данных методических указаний является ознакомление студентов с практическими приемами использования регрессионного анализа – основного метода современной математической статистики.

Задачи настоящих методических указаний – формирование у студентов

умения:

- анализа качества регрессионного уравнения;
- выбора «наилучшего» уравнения регрессии;

владения:

- навыками проверки значимости (надежности) параметров уравнения регрессии;
- навыками проверки согласия модели с экспериментальными данными (адекватности модели);
- другими навыками проверки качества регрессионной модели;
- навыками интерпретации регрессионной модели.

Методические указания разработаны для выполнения курсовых работ по дисциплине "Компьютерное и статистическое моделирование" Для студентов направления подготовки "Математика и компьютерные науки", специальности "Прикладная математика", а также студентов технических специальностей.

2. ОСНОВНЫЕ ПОНЯТИЯ РЕГРЕССИОННОГО АНАЛИЗА

2.1. Последовательность действий при построении регрессионной модели

Постановка задачи регрессионного анализа выглядит следующим образом. Имеются экспериментальные данные вида:

матрица факторов, X $\{x_{ij}\}$,

столбец отклика, Y $\{y_i\}$,

где x_{ij} - значение j-го фактора в i-ом опыте (j = 1, . . . , n; i = 1, . . . , N);

y_i - значение отклика в i-ом опыте;

n - количество факторов;

N - количество опытов.

Предварительный анализ экспериментальной информации описан в разделе 2. Проведение предварительного анализа является важным этапом, где закладываются основы для успешного построения модели. На этом этапе устанавливаются цели исследования и подходы к этим целям.

Для достижения целей требуется построить по экспериментальным данным модель, отражающую зависимость отклика от факторов. Необходимо оценить качество полученной модели (раздел 3).

Использование (интерпретация) качественной регрессионной модели обычно идет в двух направлениях (раздел 4):

- сглаживание имеющихся экспериментальных данных, решение задач типа аппроксимации и интерполяции; прогноз значений отклика в области эксперимента (“модель для предсказания” /1/);

- выделение факторов, оказывающих наибольшее влияние на отклик; оценка характера, знака и степени этого влияния; прогноз значений отклика для любых значений факторов; поиск и анализ оптимальных значений отклика и т.д. (“модель для управления” /1/).

2.2. Набор эффектов. Вид модели

Регрессионная модель, описывающая зависимость отклика от n факторов x_{ij} имеет вид:

$$y = A_0 + \sum_q A_q E_q ,$$

где A_0, A_q - параметры модели (q = 1, . . . , k);

$E_q = F\left(x_1,...,x_n\right)$ - некоторая функция от факторов $x_1,...,x_n$ (в регрессионном анализе E_q принято называть эффектом; в литературе встречаются также следующие термины для E_q - регрессор, предиктор, носитель, переменная и т.д.);

k - общее количество эффектов.

В линейном регрессионном анализе рассматриваются модели, линейные по параметрам.

Если модель необходима для описания поведения системы (а не для объяснения механизма явлений) и у исследователя нет других гипотез, то он может удовлетвориться гипотезой, выбранной из принципа простоты. Для решения большинства реальных прикладных задач целесообразно применять полиномиальные модели порядка m

$$y = A_0 + \sum_{j=1}^{n} A_j x_j + \sum_{j}^{n} \sum_{i=j}^{n} A_{ji} x_j x_i + ...$$

Коэффициенты полинома эквивалентны соответствующим коэффициентам ряда Тейлора. Для ситуации простой линейной модели (полинома первой степени):

$$m = 1; k = n; E_j = x_j, j = 1,...,n.$$

Часто стремятся построить модель в виде полинома второго порядка (m = 2). В этом случае в качестве эффектов наряду с линейными членами

используют эффекты второго порядка $E_q = x_j x_i$, которые подразделяются на квадратичные $E_q = x_j^2$ и чистые смешанные произведения $E_q = x_j x_i, j \neq i$. При этом $k = \frac{n(n+1)}{2}$. Аналогичным образом можно построить модель в виде полинома третьего порядка, используя наряду с линейными эффектами и эффектами второго порядка, эффекты третьего порядка и т.д.

В набор эффектов можно добавлять в какие-нибудь другие функции от факторов, например, логарифмы, экспоненты, тригонометрические функции и т.д. В некоторых случаях применяют функциональное преобразование отклика.

3. ПРЕДВАРИТЕЛЬНЫЙ АНАЛИЗ ЭКСПЕРИМЕНТАЛЬНОЙ ИНФОРМАЦИИ

3.1. Формализация априорной информации

Сбор, систематизация, анализ, представление в формализованном виде информации, накопленной к началу работ (называемой априорной информацией) являются важным этапом построения регрессионной модели. Для представления априорной информации в формализованном виде удобно использовать специальные бланки [2], приведенные в приложении. Заполненные бланки позволяют ответить на следующие вопросы:

- какие факторы исследовались;
- область существования факторов;
- число уровней и интервал варьирования каждого фактора в эксперименте;
- какие выходы исследовались;
- лучший достигнутый уровень для выходов;
- какие уровни качества для выходов можно считать "отличными", "хорошими", "удовлетворительными" и т.д.;
- формы взаимосвязи между различными факторами x_j и x_i (взаимодействия), различными откликами Y_{l_1} и Y_{l_2}, фактором x_j и откликом Y_l, известные из теоретических предпосылок и эмпирических исследований.

Бланки целесообразно заполнять до планирования эксперимента или, во всяком случае, до построения регрессионной модели.

Подразделяют следующие уровни априорного представления о модели:

1) известны лишь знаки влияния факторов (повышает, понижает);

2) известно ранжирование факторов по силе влияния (выделены сильно влияющие факторы и т.д.);

3) известен качественно вид модели (то есть набор эффектов, составляющих модель) - остается оценить параметры модели;

4) известны оценки для взаимоотношений параметров модели (например, оценки для оптимальных значений факторов);

5) известны оценки параметров модели (например, по результатам построения модели по аналогичным экспериментальным данным) .

3.2. Графический анализ экспериментальных данных

Графический анализ экспериментальных данных до построения регрессионной модели часто бывает очень полезен, особенно, если число факторов невелико и имеется четкая априорная информация о виде однофакторных зависимостей для отклика (раздел 2.1). В частности, рекомендуется строить графики зависимости отклика от фактора x_j, при фиксированных (или хотя бы близких) значениях других факторов. Такой простейший графический анализ позволяет:

- оценить качественно вид однофакторной зависимости отклика от фактора x_j (в том числе и нелинейной);
- проверить ее согласие с априорными соображениями (см. п. 2.1);
- выявить взаимодействия, оказывающие влияние на отклик;
- выделить выбросы и влиятельные наблюдения (см. ниже разделы 3.3. и 3.4);
- отметить опыты, отклоняющиеся от априорных представлений и общей тенденции расположения данных;
- выявить неожиданные свойства данных.

Можно проводить графический анализ экспериментальных данных и после построения регрессионной модели, например, для объяснения

неожиданных результатов, не согласующихся с априорными представлениями. Хорошей иллюстрацией для построенной регрессионной модели является график однофакторной зависимости (по экспериментальным данным) отклика от наиболее сильно влияющего фактора с использованием в качестве признака фактора, следующего по силе влияния на отклик.

3.3. Выбор вида модели

Предварительный выбор вида модели осуществляется на основе априорной информации.

Линейную модель строят, как правило, в следующих случаях:

- малое количество опытов (например, если количество опытов примерно равно количеству факторов, то это - отсеивающий эксперимент, когда, требуется по ограниченному количеству опытов выбрать факторы, оказывающие наибольшее влияние на отклик);
- узкий интервал варьирования факторов;
- априорное предположение о линейном характере взаимосвязи отклика, и факторов (раздел 2.1);
- однофакторные зависимости отклика, полученные при графическом анализе экспериментальных данных (раздел 2.2) имеют вид, близкий к прямой линии.

Часто встречается ситуация, когда фактор x_j принимает в эксперименте только два различных значения (варьируется на двух уровнях) . Тогда не имеет смысла включать в набор эффектов x_j^2. Но смешанные произведения второго порядка $x_j x_i$ и более высоких порядков могут включаться в набор эффектов. В этом случае можно использовать модель в виде неполного полинома второй или более высокой степени

(только смешанные произведения). Такая ситуация часто возникает при построении модели по результатам планирования эксперимента для полных факторных планов 2^n и реплик от них. В табл. 2.1 приведен полный набор эффектов для некоторых простых планов эксперимента.

Таблица 3.1. Полный набор эффектов для некоторых планов эксперимента

План	Количество опытов	Полный набор эффектов
2^2	4	x_1, x_2, x_1x_2
2^3	8	$x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_1x_2x_3$
2^{3-1}	4	x_1, x_2, x_3

Главное достоинство линейной модели - простота в использовании и интерпретации.

Если же количество уровней для каждого фактора больше двух и количество опытов достаточно велико, то рекомендуется обязательно построить модель в виде полного полинома второй степени. Даже, если априорно предполагается монотонная зависимость отклика от фактора x_j, которая обычно достаточно хорошо аппроксимируется прямой, регрессионный анализ может показать наличие криволинейности у этой монотонной зависимости. Но, как правило, наличие в модели квадратичного члена означает параболическую (с минимумом или максимумом - в зависимости от знака коэффициента перед эффектом) зависимость отклика от фактора x_j. Главная ценность моделей в виде полинома второго порядка - возможность поиска области значений факторов, обеспечивающих экстремальные значения отклика. Интерпретация таких моделей существенно богаче, чем линейных моделей, но и, соответственно, сложнее.

Модели в виде полинома третьего порядка, особенно полный полином, применяются сравнительно редко. Также достаточно редко (только, если есть соответствующая априорная информация) используют функциональные преобразования факторов и отклика.

Существуют некоторые типовые ситуации, в которых применяют логарифмирование отклика:

1. Сведение нелинейной модели к модели, линейной по параметрам (например, замена модели $y = \exp(A + Bx)$ моделью $\ln y = A + Bx$;

2. Значения отклика сильно различаются между собой (на несколько порядков). В этих случаях измерения с малыми значениями отклика не оказывают никакого влияния на параметры модели и плохо описываются. Логарифмирование отклика позволяет сгладить указанные резкие различия.

3. Из априорных соображений известно, что отклик может принимать только положительные значения.

4. Ошибка измерения отклика - относительная, то есть малые значения измеряются с маленькой ошибкой, большие - с большой (в этом случае можно применять также взвешенную регрессию).

4. АНАЛИЗ КАЧЕСТВА МОДЕЛИ

В этом разделе изложена методика анализа качества модели только в статистическом смысле, не касаясь интерпретации модели.

После того как определен набор эффектов $\{E_q, q = 1,...k\}$ - кандидатов в члены модели, приступают собственно к построению модели

$$\hat{y} = A_0 + \sum_{s=1}^{p} A_s E_{q_s} \ . \tag{4.1}$$

Существенным является то обстоятельство, что в модель включаются не все эффекты, а только некоторые, оказывающие наиболее сильное влияние на отклик. Параметры модели, содержащей данное подмножество из p эффектов $\{E_{q_s}\}$, определяются методом наименьших квадратов:

$$\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \to \min. \tag{4.2}$$

Регрессионный анализ позволяет дать ответ на два существенных вопроса:

- надежно ли определены параметры модели (проверка гипотезы о значимости параметров модели);
- хорошо ли описывает модель имеющиеся экспериментальные данные (проверка гипотезы об адекватности модели).

Нарушение надежности модели может проявляться и в более завуалированной форме, а именно через:

- наличие выбросов;
- наличие влиятельных наблюдений;
- нарушение основных предположений регрессионного анализа;
- наличие мультиколлинеарности.

Анализ экстремумов модели в области эксперимента дает возможность провести быструю оценку пригодности модели для предсказания.

4.1. Проверка гипотезы о значимости параметров модели

Проверка гипотезы о значимости параметра A_s модели (4.1) производится следующим образом. Строят модель, содержащую (p - 1) эффект (без эффекта E_{q_s}, то есть $A_s = 0$) и отмечают, насколько возросла остаточная сумма квадратов по сравнению с моделью, содержащей p эффектов. Мерой возрастания остаточной суммы квадратов является величина t-статистики параметра, подчиняющаяся распределению Стьюдента [1]. Чем больше эта величина, тем более надежно определен параметр. Если величина t-статистики параметра превышает пороговое значение, зависящее от уровня надежности α и количества степеней свободы, равного $f = N - p - 1$, то параметр A_s значимо с надежностью выше α отличен от нуля. В противном случае A_s можно считать равным нулю и эффект E_{q_s} можно исключить из модели. Пригодны для использования только модели, в которых все параметры статистически значимы.

Для уровня надежности α = 95 % и достаточно большого количества степеней свободы (больше 30) соответствующее пороговое значение равно 1.96.

4.2. Проверка согласия модели с экспериментальными данными (адекватности модели)

Проверка гипотезы об адекватности модели состоит в сопоставлении ошибки экспериментального определения значений отклика (называемой в

регрессионном анализе ошибкой воспроизводимости или чистой ошибкой [1]) и ошибки, связанной с отклонением экспериментальных значений отклика от расчетных (по модели). Если первая ошибка существенно меньше второй, то говорят, что модель не адекватна. В этом случае модель неудовлетворительно описывает имеющиеся экспериментальные данные. Нежелательной является и ситуация, когда вторая ошибка значительно меньше первой. Это означает, что модель содержит избыточное количество членов, соответствующее случайным колебаниям, а не реальным изменениям значений отклика. Оптимальной является ситуация, когда эти ошибки сопоставимы (первая меньше второй). В этом случае модель удовлетворительно описывает экспериментальные данные и, говорят, что модель адекватна.

4.2.1. Применение критерия Фишера для проверки гипотезы об адекватности модели

Величина остаточной суммы квадратов (которую в регрессионном анализе принято обозначать RSS - сокращение от Residual Sum of Squares) является одной из важнейших оценок точности описания экспериментальных данных. Для этих же целей часто используются и остаточная дисперсия или средняя квадратичная ошибка, обозначаемая MSE (Mean Square Error). Величина MSE вычисляется по формуле

$$MSE = RSS / f \qquad (4.3)$$

где f - количество степеней свободы (независимых элементов информации для определения модели).

Как уже отмечалось для модели со свободным членом f = N - p - 1,

где p - количество членов модели.

Более удобной для практических целей величиной является $\sqrt{MSE}$ - средний остаток модели. $\sqrt{MSE}$ характеризует среднее отклонение

расчетных и экспериментальных значений отклика. Эта величина измеряется в тех же единицах, что и отклик.

Удобной безразмерной оценкой точности описания моделью экспериментальных данных является коэффициент множественной детерминации R^2:

$$R^2 = 1 - \frac{RSS}{\sum_{i=1}^{N}(y_i - \overline{y})^2},$$

где $\overline{y} = \frac{\sum_{i=1}^{N} y_i}{N}$.

R^2 показывает насколько предсказание по модели лучше, чем предсказание по среднему значению отклика $\overline{y}$. R^2 характеризует долю разброса отклика, описываемую регрессией ($1 \geq R^2 \geq 0$). Чем ближе R^2 к единице, тем лучше описывает модель экспериментальные данные.

Обозначение коэффициента множественной детерминации связано с тем, что

$$R^2 = R_1^2,$$

где R_1 - коэффициент множественной корреляции. R_1 можно рассматривать как обычный коэффициент парной корреляции между y и $\hat{y}$.

Если есть априорная оценка дисперсии ошибки отклика σ_e^2, то можно сравнить величины MSE и σ_e^2.Как правило, MSE > σ_e^2. С помощью F-критерия Фишера проверяют, значимо ли остаточная дисперсия MSE превышает оценку дисперсии ошибки отклика σ_e^2. С учетом того, что отношение дисперсий подчиняется F-распределению Фишера, вычисляют расчетное значение F-критерия

$F_{расч} = MSE / \sigma_e^2$

и сравнивают его с табличным значением $F_{табл}(f, f_e, \alpha)$ для f и f_e степеней свободы и уровня принятия решения α, где f и f_e - количество степеней свободы для вычисления MSE и σ_e^2, соответственно (если σ_e^2 является табличной величиной, то обычно принимают $f_e = \infty$); α обычно равно 95 %. Если $F_{расч} > F_{табл}$, то говорят, что модель неадекватна и пользоваться ею нельзя, так как она неудовлетворительно описывает имеющиеся экспериментальные данные. Если критерий адекватности не значим, то есть $F_{расч} \le F_{табл}$, то модель адекватна.

Нежелательной является редкая ситуация, когда $MSE << \sigma_e^2$. Это означает, что модель переопределена, то есть количество членов в модели необоснованно завышено. Модель описывает экспериментальные данные с большей точностью, чем они определены. Модель переопределена, если

$\sigma_e^2 / MSE > F_{табл}(f_e, f, \alpha)$.

4.2.2. Дисперсия воспроизводимости

На практике априорная оценка дисперсии ошибки отклика редко бывает известна заранее. Однако во многих случаях σ_e^2, можно оценить по экспериментальным данным.

Остановимся подробнее на смысле величины σ_e^2 (ее называют в регрессионном анализе дисперсией воспроизводимости или дисперсией опыта). Величина σ_e^2, связана с разбросом значений отклика, возникающим при повторении опытов в точках плана эксперимента. Таким образом, величина σ_e^2 включает в себя влияние неучтенных факторов.

Следует подчеркнуть отличие дисперсии воспроизводимости от дисперсии измерения (для последней часто существует априорная оценка).

Дисперсия измерения связана с неточностью измерения значения отклика. Дисперсия измерения всегда меньше дисперсии воспроизводимости (как правило, существенно меньше - в несколько раз) и является одной из ее составляющих.

На практике встречаются величины, занимающие промежуточное положение между дисперсией измерения и дисперсией воспроизводимости. Так, при построении модели зависимости механических свойств изделий, часто возникает следующая ситуация. Для повышения точности определения величины механического свойства образец разрезают на несколько частей и измеряют значение механического свойства для каждой части. Возникающий при этом разброс значений механического свойства связан, в основном, с неоднородностью изделия, и, во вторую очередь, с погрешностью измерения механического свойства. Соответствующая дисперсия занимает промежуточное положение между дисперсией измерения и дисперсией воспроизводимости. Истинная дисперсия воспроизводимости связана с разбросом значений механического свойства, измеренных на разных образцах, изготовленных по одной технологии, а не на разных частях одного образца.

Но если результаты моделирования предполагается применять, скажем, для различных видов исходного сырья, то желательно определить дисперсию воспроизводимости, связанную с разбросом значений механического свойства, измеренных на образцах, изготовленных по единой технологии из различных видов исходного сырья.

К настоящему времени метрологический этап построения регрессионной модели (оценка точности измерений и воспроизводимости опытов) нельзя считать полностью формализованным. Следует отметить, что решение метрологических проблем эксперимента с позиции

статистического моделирования пока нельзя признать полным и законченным [2].

4.2.3. Оценка дисперсии воспроизводимости

Надежную оценку величины дисперсии воспроизводимости можно получить с помощью экспериментальных данных, содержащих повторные (другое название - параллельные) опыты. Повторными являются независимо проведенные опыты при фиксированных значениях факторов.

В предыдущем разделе показана разница между повторными опытами и повторными измерениями отклика. Имеет смысл закладывать повторные опыты на этапе планирования эксперимента.

Оценка дисперсии σ_e^2 по повторным опытам производится по формуле:

$$\sigma_e^2 = \frac{RSS_e}{f_e}, \tag{4.4}$$

где $RSS_e = \sum_{i=1}^{N_e} \sum_{j=1}^{N_i} \left(y_{ij} - \frac{\sum_{j=1}^{N_i} y_{ij}}{N_i} \right)^2$ - сумма квадратов, связанная с "чистой" ошибкой;

y_{ij} - j-ый повтор в i-ом опыте;

N_e - количество различных опытов;

N_i - количество повторов в i-ом опыте;

$f_e = \sum_{i=1}^{N_e} (N_i - 1)$ - количество степеней свободы для расчета чистой ошибки.

При расчете величины σ_e^2 полезно вычислить также дисперсию отклика в каждом опыте, равную

$$\frac{\sum_{u=1}^{N_i}\left(y_{iu}-\frac{\sum_{u=1}^{N_i}y_{iu}}{N_i}\right)^2}{N_i-1}. \qquad (4.5)$$

Далее, можно проверить массив дисперсий на неоднородность (для этого сравниваются с помощью F-критерия Фишера максимальная и минимальная дисперсии). Если массив дисперсий оказался неоднородным, то следует понять, почему. Это может быть связано с очень низким значением дисперсии в каком-либо опыте (то есть почти одинаковые значения отклика в этом опыте). Тогда можно пересчитать значение F-критерия Фишера, выбрав в качестве "минимальной" следующую по величине после минимальной дисперсии. Неоднородность массива дисперсий может быть связана с наличием выбросов (см. раздел 4.3.2).

Неоднородностью массива дисперсий можно пренебречь, если нет априорной информации о наличии зависимости ошибки эксперимента от изучаемых факторов. В противном случае рекомендуется применить взвешенную регрессию, используя в качества веса для i-го измерения, величину пропорциональную корню квадратному из дисперсии отклика, вычисляемой по формуле (4.5).

В отличие от стандартной регрессионной модели, отвечающей формуле (4.2), в случае взвешенной регрессии минимизируется взвешенная остаточная сумма квадратов

$$\sum_{i=1}^{N} w_i (y_i - \hat{y}_i)^2 \to \min,$$

где w_i - вес для i-го измерения (w_i >0).

Для удобства сравнения со стандартной регрессионной моделью вес измерений выбирают таким образом, чтобы выполнялось условие

$$\sum_{i=1}^{N} w_i = N .$$

При наличии повторных опытов проверка адекватности модели производится следующим образом. Остаточная сумма квадратов распадается на две части:

$$RSS = RSS_e + RSS_L ,$$

где RSS_e - сумма квадратов, связанная с "чистой" ошибкой (см. формулу (4.4));

$RSS_L = \sum_{i=1}^{N_e} N_i \left(\hat{y}_i - \frac{\sum_{u=1}^{N_i} y_{iu}}{N_i} \right)^2$ - сумма квадратов, связанная с неадекватностью модели.

Величина RSS_L имеет $N - f_e - p - 1$ степеней свободы.

Вычисляют отношение

$$F = \frac{RSS_L}{(N - f_e - p - 1)\sigma_e^2}, \qquad (4.6)$$

Отношение (4.6) подчиняется F-распределению Фишера. Если полученное по формуле (4.6) значение превышает табличное значение F-критерия Фишера для $N - f_e - p - 1$ и f_e степеней свободы, то модель неадекватна, если не превышает – гипотеза об адекватности модели не отвергается.

4.2.4. Модель неадекватна

Что же делать, если модель оказалась неадекватна?

В первую очередь надо проверить корректность вычисления дисперсии воспроизводимости σ_e^2. Возможно, что она была занижена (в

том смысле, как об этом сказано в разделе 4.2.2), например, в качестве дисперсии воспроизводимости использовали дисперсию измерения. В случае заведомо заниженной дисперсии воспроизводимости на неадекватность модели можно не обращать внимания.

Неадекватность модели может быть связана с присутствием выбросов (см. раздел 4.3.2) - опытов, значения отклика в которых, плохо описываются моделью. Удаление выбросов может существенно понизить остаточную сумму квадратов и сделать модель адекватной.

Если же дисперсия воспроизводимости определена корректно и неадекватность не связана с выбросами, то полученная модель неудовлетворительно описывает имеющиеся данные. Обычно это связано с неудачным выбором вида модели. Следовательно, надо изменить (как правило, усложнить) вид модели (см. разделы 3.3, 3.4). Простейшее изменение вида модели - увеличение степени полинома (полином второй степени вместо линейной модели; полином третьей степени вместо полинома второй степени). Более сложные ситуации возникают при использовании преобразований. Здесь трудно дать какие-то общие рекомендации. Использование нелинейных по параметрам преобразований может в некоторых случаях значительно уменьшить остаточную сумму квадратов, но требует применения методов нелинейной регрессии, что существенно усложняет построение модели и последующую интерпретацию. Если же неадекватность связана с тем, что априорно не учтены важные факторы, оказывающие существенное влияние на отклик, то устранить ее не удается. Модель с неустраненной неадекватностью не рекомендуется применять для практических целей.

4.3. Анализ остатков

Остатком называется разность экспериментального y_i и расчетного $\hat{y}_i$ значений отклика:

$$e_i = y_i - \hat{y}_i .$$

4.3.1. Проверка нарушения основных предположений

Остаток - это разность между тем, что фактически наблюдалось, и тем, что предсказывается с помощью модели. То есть, это величины, которые нельзя объяснить с помощью модели. Таким образом, можно считать, что остатки - это наблюдаемые проявления ошибок [1]. В разделе 7.1 описаны допущения относительно ошибок, принимаемые в регрессионном анализе, а именно, что ошибки независимы, имеют нулевое среднее, постоянную дисперсию и подчиняются нормальному распределению. Анализ остатков может показать нарушение некоторых из этих предположений. В противном случае можно утверждать, что предположения, по-видимому, не нарушены.

Предположение о том, что ошибки имеют нулевое среднее, выполняется для остатков автоматически, так как для регрессионной модели со свободным членом

$$\sum_{i=1}^{N} e_i = 0 .$$

Нарушение предположений о независимости и равноточности (постоянство дисперсий) ошибок не очень критично. В этих ситуациях все равно можно применять регрессионный анализ, только взвешенный. Видимо, поэтому проверке этих предположений уделяется, как правило, мало внимания (проверка независимости ошибок обычно вообще не производится). Для проверки равноточности рекомендуют [1] строить графики зависимости остатков от факторов.

Более важным является предположение о нормальности ошибок, так как его нарушение приводит к невозможности применения аппарата проверки статистических гипотез. Мы уже говорили о том, что остатки имеют нулевое среднее. В качестве оценки дисперсии остатков обычно выбирают величину MSE. Таким образом, надо проверить гипотезу о том, что остатки подчиняются распределению $N(0, MSE)$ - нормальному распределению с нулевым средним и дисперсией, равной MSE.

Для этого можно было бы провести стандартные проверки на нормальность по статистическим критериям (например, χ^2 Пирсона), построить эмпирическую гистограмму распределения остатков, график накопленных частот для нормальных вероятностных координат и т.д. [1]. Но все это делать не обязательно. Во-первых, остатки обычно довольно-таки хорошо подчиняются нормальному закону, особенно в центральной части: обычно в интервале $(-\sqrt{MSE}, \sqrt{MSE})$ лежит 65-70 % остатков, а в интервале $(-2\sqrt{MSE}, 2\sqrt{MSE})$ - около 95 % остатков. А, во-вторых, наибольший урон качеству регрессионной модели может принести специфическое нарушение нормальности, проявляющееся в появлении, больших остатков, значительно превышающих $\sqrt{MSE}$ по абсолютной величине (в три и более раз). Такие опыты обычно называют выбросами. Поэтому для оценки нормальности остатков достаточно проверить данные на наличие выбросов. Методика выделения выбросов и техника работы с ними описана в следующем разделе.

4.3.2. Выбросы

Выбросом называется опыт, в котором значение отклика плохо, описывается моделью. Существуют формальные критерии выделения выбросов, основанные на анализе остатков.

Так как мы считаем, что остатки должны подчиняться нормальному закону $N(0, MSE)$, то нормализованные остатки

$$e_i^н = \frac{e_i}{\sqrt{MSE}} \tag{4.7}$$

должны подчиняться стандартному нормальному закону $N(0,1)$. Тогда подозрительным на выброс являются значения отклика, для которых величины $\left|e_i^н\right| > 2$ и, очевидными выбросами при $\left|e_i^н\right| > 3$.

Более обоснованным является подход, связанный с анализом так называемых студентизованных (подчиняющихся t-распределению Стьюдента) остатков [1]. Обозначим через $\hat{\alpha}_i$, оценку значения y_i, основанную на модели, построенной по данным, содержащим (N - 1) опыт (без i-го опыта). Студентизованные остатки вычисляются по формуле

$$e_i^{nn} = \frac{y_i - \hat{\alpha}_i}{\sqrt{MSE}}. \tag{3.8}$$

Различие студентизованных и нормализованных остатков показано на рис. 4.1 и 4.2. На этих рисунках точками представлены условные экспериментальные данные измерения значения отклика Y в зависимости от фактора X. На рис. 4.1 представлен классический выброс - точка A. Существенно, что соответствующее значение фактора расположено достаточно близко к середине факторного пространства - отрезка (X_C,X_D). Для точки A при расчетах по формулам (4.7) и (4.8) будет получено большое значение как для нормализованного, так и для студентизованного остатков. Модели, построенные по данным, содержащим N опытов и (N - 1) опыт (без точки A), близки и лишь немного отличаются свободным членом.

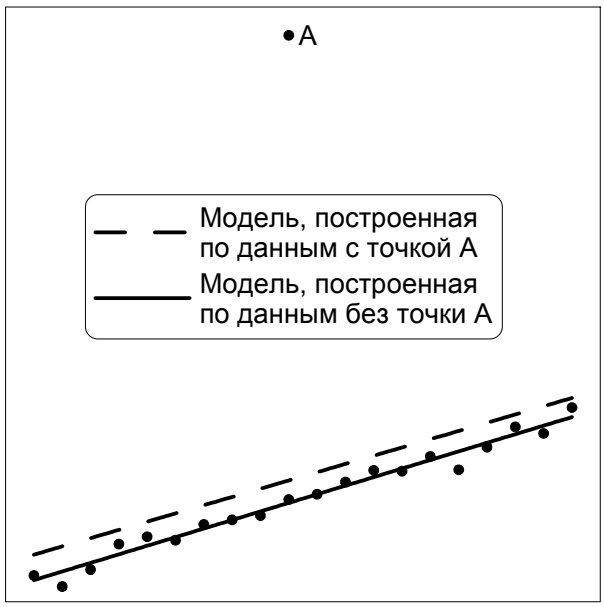


Рисунок 4.1. Выброс в середине факторного пространства

На рис. 4.2 изображена совсем другая ситуация. Ясно, что точка B плохо описывается линейной моделью. Но из традиционного анализа остатков этого можно и не увидеть, так как остаток для точки B не так уж и велик (соответствующий нормализованный остаток составляет лишь 2.2; например, остатки и нормализованные остатки для точек B и C практически одинаковы). Однако, модель, построенная по (N - 1) опыту без точки B, будет резко отличаться от модели, построенной по всем точкам. Соответствующий студентизованный остаток будет ясно показывать, что точка B является выбросом. Значение студентизованного остатка в точке B равно 3.3, что существенно выше порогового.

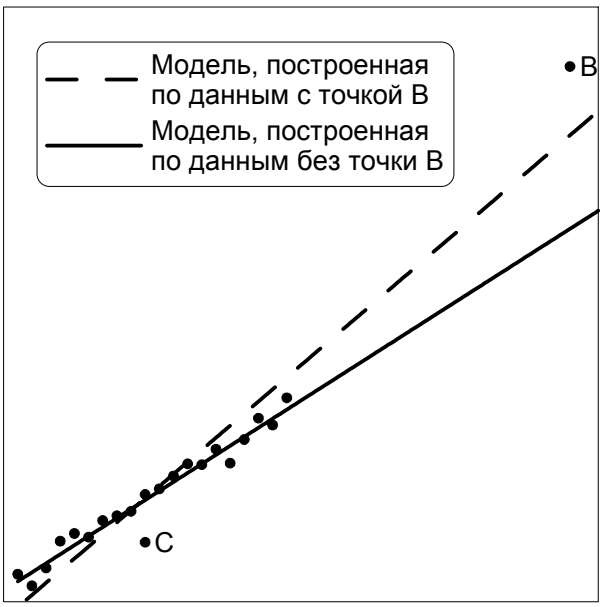


Рисунок 4.2. Выброс на краю факторного пространства

Преимущества студентизованных остатков над нормализованными не только в том, что анализ студентизованных остатков может выявить выброс, пропущенный при анализе нормализованных остатков. Ситуация на рис. 4.2 является более важной, чем ситуация на рис. 4.1, так как, если считать точку B выбросом и исключать ее из экспериментальных данных, то параметры модели значительно изменятся (тогда как в первом случае параметры моделей для N и (N-1) опыта близки). Таким образом, преимущество студентизованных остатков в том, что они позволяют находить более "важные" выбросы, оказывающие сильное влияние на параметры модели. Подробнее эти вопросы рассмотрены в разделе 4.4.

Что же делать, если обнаружен выброс? Автоматическое исключение выбросов - это далеко не всегда наиболее целесообразная процедура. Выброс - экспериментальная точка, которая не типична по отношению к остальным данным. Каждое такое наблюдение должно подвергаться тщательному исследованию с целью выяснения причин его возникновения. Полезно изучить положение точки - кандидата на выброс - на графиках однофакторных зависимостей отклика от факторов, оказывающих

наибольшее влияние на отклик (раздел 3.2). При наличии повторных опытов можно сравнить разброс (дисперсию) значений отклика, в опыте, содержащем выброс, с разбросом в других опытах, со средним разбросом (раздел 4.2.3). Выбросы должны исключаться (по согласованию с экспериментатором), если они вызваны такими причинами, как ошибки, проведение эксперимента в других условиях и т.д. Во многих случаях выбросы могут побуждать к изменению вида модели. Основные причины появления таких наблюдений:

- случайные колебания, обусловленные природой генеральной совокупности;
- нарушение условий проведения эксперимента;
- нарушение условий сбора данных;
- ошибки при регистрации данных и подготовке их для статистической обработки.

4.4. Влиятельные наблюдения

Регрессионный анализ проводится для изучения наличия, силы и формы взаимодействий между различными величинами. Однако, известно, что на параметры регрессионной модели сильное влияние может. оказывать небольшая группа из общего количества опытов таким образом, что построенная модель (параметры модели и ее статистические характеристики) может отражать скорее необычные свойства этих опытов, чем действительные соотношения между переменными. Конечно, полезно уметь выделять такие опыты.

Заметное влияние на параметры модели могут оказывать выбросы и удаленные точки, располагающиеся в факторном пространстве вне основной группы точек. Для идентификации выбросов используется значение величины студентизованного остатка, описанного в предыдущем

разделе. Характеристикой удаленности опыта от основной массы точек служит величина V_i - значение соответствующего элемента матрицы подгонки. Матрицей подгонки [7] в регрессионном анализе принято называть матрицу размером N×N:

$$H = E(E^T E)^{-1} E^T,$$

где E - матрица размером N×p эффектов, вошедших в модель (E_{iq} - значение q-го эффекта в i-ом опыте).

Диагональные элементы этой матрицы удовлетворяют следующим соотношениям:

$$\sum_{i=1}^{N} V_i = p + 1;$$

$$0 \le V_i \le 1.$$

Если V_i близко к нулю, то это означает, что i-ая точка располагается в середине факторного пространства; если V_i близко к единице, то i-ая точка является удаленной. Точки, которым соответствуют большие V_i называют точками разбалансировки [7]. Большие значения V_i служат сигналом, что i-ое наблюдение может играть большую, едва ли поддающуюся контролю, роль. В [7] предложена следующая шкала:

$V_i \le 0.2$ надежные;

$0.2 < V_i < 0.5$ рискованные;

$V_i \ge 0.5$ таких V_i лучше избегать.

Последняя рекомендация не всегда выполнима, особенно, когда количество членов модели почти равно количеству опытов.

В [1] предлагается тщательно проверять точки, в которых

$V_i > 2 (p + 1) / N$.

В любом случае нежелательными являются значения $V_i > 0.9$.

В качестве меры изменения параметров модели при удалении i-го наблюдения рекомендуется использовать величину

$$D_i = \frac{V_i \left(e_i^{nn} \right)^2}{\left(1 - V_i\right)\left(p + 1\right)}, \qquad (4.9)$$

где (p + 1) - количество параметров модели;

e_i^{nn} - значение студентизованного остатка;

V_i - значение диагонального элемента матрицы подгонки.

Из формулы (4.9) видно, что D_i будет велико, если e_i^{nn} велико (выброс) или V_i близко к единице (удаленная точка).

Наблюдения с $D_i > 1$ оказывают сильное влияние на параметры модели. При удалении таких опытов значение некоторых параметров может измениться на несколько стандартных отклонений.

Полезно изучить положение точек, соответствующих большим значениям D_i на графиках квазиоднофакторных зависимостей отклика от факторов, оказывающих заметное влияние на отклик (раздел 3.2).

4.5. Мультиколлинеарность

Одним из основных препятствий для эффективного применения аппарата регрессионного анализа является мультиколлинеарность. Подробнее о том, что такое мультиколлинеарность, негативных последствиях присутствия мультиколлинеарности в данных и мерах по ее устранению сказано в разделе 7.2.

Коротко, мультиколлинеарность - это сильная взаимосвязь между эффектами, вошедшими в модель, не позволяющая точно оценить их раздельное влияние на отклик. Для быстрого анализа качества модели нужно иметь в виду, что нежелательной является ситуация, когда коэффициенты множественной корреляции эффектов в модели по абсолютной единице близки к единице (превышают 0.95). В случае, если максимальный из этих коэффициентов превышает коэффициент

множественной корреляции отклика с эффектами в модели (то есть связь отклика с эффектами слабее, чем взаимосвязь эффектов между собой), то следует учесть, что использовать полученную регрессионную модель для интерпретации рискованно.

4.6. Прогноз значений откликов в области эксперимента

Важной характеристикой качества модели является ее предсказывающая способность. Полезно проверить расчетные значения отклика $\hat{y}_i$ вместе с 95 %-ными доверительными интервалами.

Величина доверительного интервала пропорциональна $\sqrt{V_i}$, где V_i - значение диагонального элемента матрицы подгонки. Опыты, в которых значения V_i близки к единице, имеют более широкие доверительные интервалы.

Полезно хотя бы бегло просмотреть доверительные интервалы для $\hat{y}_i$. Наличие абсурдных значений (например, отрицательных значений нижней границы, когда заведомо известно, что $\hat{y}_i$ должно быть положительно) свидетельствует о низком качестве модели (как правило, это связано с тем, что модель переопределена, то есть содержит избыточные члены). Если при построении модели используется известная дисперсия воспроизводимости, то при увеличении количества членов модели доверительные интервалы для $\hat{y}_i$ увеличиваются [5].

Следующим шагом при оценке предсказывающей способности модели в виде полинома второго порядка является анализ экстремальных значений отклика в области эксперимента. Целесообразно произвести поиск значений факторов, обеспечивающих максимальное и минимальное значение отклика в области эксперимента. В [2] для этой цели предложен диссоциативно-шаговый метод. Полезной сопроводительной информацией являются доверительные интервалы найденных экстремальных значений

отклика и “показатель экстраполяции” (подробно описанный в разделе 5.1.3).

Величина показателя экстраполяции имеет тот же смысл, что и диагональный элемент матрицы подгонки *V*, описанный в разделе 4.4.1.

Для точек плана эксперимента V ≤ 1. Значения факторов, обеспечивающие экстремальные значения отклика, формально лежат в области эксперимента. Тем не менее, для них может быть V > 1, а для ситуации с сильной мультиколлинеарностью (см. раздел 4.5) даже V >> 1. Значения V > 1 свидетельствуют о том, что такая комбинация факторов лежит в стороне от точек плана (см. раздел 5.1.3).

Области экстремумов отклика требуют внимательного изучения. Достигнутое экстремальное значение отклика должно иметь смысл (например, для выхода годного не имеют смысла значения $\hat{y}_{\min} < 0$ и $\hat{y}_{\max} > 100\%$.

Бессмысленные значения экстремального значения отклика могут возникать в следующих случаях:

1. Некачественная (обычно переопределенная, то есть содержащая избыточное количество членов) модель. О том, что модель переопределена, свидетельствуют также высокие коэффициенты множественной корреляции эффектов (превышающие 0.95). В некоторых случаях даже перекрываются доверительные интервалы экстремальных значений отклика (то есть нижняя граница максимума ниже верхней границы минимума). Это говорит о чрезвычайно низком качестве модели.

В подобной ситуации надо уменьшить количество членов переопределенной модели (см. раздел 6). Рекомендуется при построении модели по данным пассивного эксперимента включать в нее не более, чем N/10 членов, где N - количество опытов [1].

2. Экстремальное значение отклика вычислено в точке с невозможной комбинацией факторов (например, сумма концентраций компонентов больше 100 %; технологически недостижимое сочетание факторов и т.д.).

В этом случае поиск оптимальных значений отклика надо осуществлять с учетом ограничений на факторы. Найденные оптимальные значения отклика нужно дополнительно проверить.

3. В эксперименте присутствуют значения отклика, очень близкие к соответствующему ограничению (например, встречаются выходы годного, близкие к нулю или к 100 %). В подобных ситуациях зачастую экстремальные значения $\hat{y}$ могут выходить за ограничения. Если нарушается только то ограничение, которое не особенно интересует (например, получены отрицательные значения выхода годного в ситуации, когда стремятся выход годного максимизировать), то этим можно и пренебречь (и интерпретировать отрицательные значения выхода годного, как чрезвычайно низкий выход $\hat{y} \approx 0$).

Чтобы жестко выдержать ограничения, следует применять преобразования отклика. Для выполнения ограничения $\hat{y} > a$ можно использовать, например, следующие преобразования отклика:

$$\tilde{y} = \ln(y - a);$$

$$\tilde{y} = \frac{1}{y - a}$$

с последующим построением модели для преобразованного отклика $\tilde{y}$.

Можно строить и для исходного отклика в виде:

$$y = a + \exp(f(x));$$

$$y = a + \frac{1}{f(x)};$$

$$y = \begin{cases} f(x) & f(x) \geq a \\ a & f(x) < a \end{cases}$$

но для построения подобных моделей требуется использование нелинейной регрессии.

4.7. Другие способы проверки качества модели

В последнее время популярны методы, основанные на идее разбиения массива экспериментальных данных на две части. Одна часть используется для построения модели, вторая - для проверки этой модели. К этому направлению относятся методы “складного ножа” (Jack-knife), “перепроверки” (Cross-validation), бутстреп.

4.7.1. Метод “складного ножа”

Исключим из экспериментальных данных i-ый опыт и построим модель по (N-1) опыту. Пусть $\hat{\alpha}_i$ - прогноз значения отклика в i-ом опыте, рассчитанный по модели, построенной для данных без i-го опыта. В качестве реалистичной оценки остаточной дисперсии MSE можно выбрать

$$MSE_1 = \frac{\sum_{i=1}^{N} (y_i - \hat{\alpha}_i)^2}{N - p - 1} \qquad (4.10)$$

Величина MSE_1, полученная с использованием формулы (4.10), как правило, значительно (на порядок) больше, чем обычная MSE, полученная по формуле (4.3). Это связано с тем, что модель предсказывает те данные, по которым она построена, существенно точнее, чем любые другие.

Величину MSE_1 можно использовать для проверки адекватности модели и построения реалистичных доверительных интервалов.

4.7.2. Метод "перепроверки"

В этом методе производится проверка качества модели на независимой выборке. При этом разбиение всех экспериментальных данных на две части - для построения модели и ее проверки - производится до начала работы. Существуют рекомендации по разбиению экспериментальных данных. Основной смысл их состоит в следующем: две части должны содержать примерно одинаковое количество опытов и соответствующие точки должны быть распределены примерно равномерно и примерно одинаковым образом по всей области эксперимента.

По независимой выборке можно затем получить статистически более обоснованную, чем традиционная, оценку остаточной дисперсии MSE_2

$$MSE_2 = \frac{\sum_{i=1}^{N_1}(y_i - \hat{y}_i)^2}{N_1} \qquad (4.11)$$

где N_1 - размер независимой выборки.

Величину MSE_2 можно использовать для проверки надежности (адекватности) модели. Отметим, что резервирование части данных только для проверки качества модели целесообразно проводить еще на этапе планирования эксперимента.

Полезна и традиционная практика, состоящая в проведении дополнительных опытов для проверки рекомендуемых оптимальных режимов.

При этом происходит проверка надежности модели в области, наиболее важной для прикладных целей. Проверочные точки рекомендуется строить с использованием методов планирования эксперимента.

4.7.3. Метод бутстреп

Сущность метода состоит в следующем: по выборке экспериментальных данных, содержащей N опытов, формируются так называемые бутстреп-выборки. Каждая бутстреп-выборка состоит тоже из N опытов, которые выбираются случайным образом (с возвращением) из исходной выборки. Таким образом, каждый опыт бутстреп-выборки входит в исходную выборку, но некоторые опыты исходной выборки могут входить в бутстреп-выборку несколько раз, а некоторые - могут отсутствовать в бутстреп-выборке. Построив большое количество (рекомендуется 500-1000) бутстреп-выборок, можно по каждой выборке рассчитать регрессионную модель. Полученные наборы значений параметров регрессионной модели можно использовать для построения эмпирической функции распределения этих параметров или любых других величин, относящихся к модели. Привлекательность методики бутстреп связана с возможностью определения с любой желаемой точностью эмпирической функции распределения изучаемой величины, что сулит многочисленные преимущества (существенно расширяется класс соответствующих статистических выводов, появляется возможность освободиться от гипотезы о нормальном распределении ошибок и т.д.). Затраты на использование этого метода связаны только со значительным увеличением сложности расчетов. Некоторые математики с известным скептицизмом относятся к методу бутстреп. Основные возражения:

- слабая теоретическая обоснованность метода;
- определенная зависимость от датчика псевдослучайных чисел (бутстреп требует использования большого количества случайных чисел);
- недостаточно выявлены типы данных, для которых бутстреп заметно лучше традиционных методов.

4.8. Выводы

В таблице 4.1 перечислены основные возможные недостатки регрессионных моделей и способы их выявления.

Таблица 4.1. Основные возможные недостатки регрессионных моделей и способы их выявления

Возможные недостатки	Способы их выявления
Неадекватность модели (то есть недостаточная по сравнению с экспериментом точность описания данных)	Если есть повторные опыты, то можно проверить адекватность модели (то есть сравнить ошибку предсказания с экспериментальной)
Данные содержат выбросы - опыты, плохо описываемые моделью	Анализ студентизованных остатков (которые должны быть не больше трех по абсолютной величине)
Имеются влиятельные наблюдения (то есть наблюдения, удаление которых приведет к резкому изменению параметров модели)	Анализ меры влияния (должны быть не больше единицы) и диагональных элементов матрицы подгонки (должны быть не больше 0.9)
В модели присутствует мультиколлинеарность (сильная взаимосвязь эффектов)	Анализ коэффициентов множественной корреляции эффектов в модели (должны быть не больше 0.95)
Модель непригодна для предсказания в области эксперимента (дает невозможные предсказанные значения отклика), что может быть связано с низким качеством модели	Анализ экстремумов модели

5. ИНТЕРПРЕТАЦИЯ РЕГРЕССИОННОЙ МОДЕЛИ

5.1. Типовая интерпретация полиномиальной модели второго порядка

Регрессионные модели в виде полинома носят, как правило, формальный характер и не отражают физической сущности изучаемых объектов. Тем не менее, полиномиальные модели весьма удобны для решения практических прикладных задач. Ниже перечислены основные возможности интерпретации таких моделей.

5.1.1. Выделение факторов, оказывающих наибольшее влияние на отклик

Традиционно принято считать, что факторы, вошедшие в модель, оказывают заметное влияние на отклик. И, напротив, факторы, не вошедшие в модель, оказывают слабое влияние на отклик (при изменении факторов в зоне эксперимента; так что слабое влияние фактора может в частности объясняться узостью его интервала варьирования). Рекомендуется обязательно строить линейную модель зависимости отклика от факторов. При построении линейной модели следует обратить особое внимание на коэффициенты парной корреляции отклика с факторами и корреляционную матрицу факторов. Вообще говоря, все факторы, имеющие достаточно большие по абсолютной величине коэффициенты корреляции с откликом (больше, чем пороговое значение для уровня 95 %) должны бы войти в модель. Тем не менее, часто встречается ситуация, когда фактор имеет большой по абсолютной величине коэффициент корреляции с откликом, а в модель он не вошел. Как правило, причиной этого является сильная взаимосвязь этого фактора x_j с прочими факторами, вошедшими в модель “вместо” фактора x_j. Эту

взаимосвязь можно проследить в корреляционной матрице факторов, где она проявляется в больших по абсолютной величине коэффициентах корреляции фактора x_j с факторами, вошедшими в модель. Значения коэффициентов корреляции между факторами, близкие по абсолютной величине к единице, показывают, что эксперимент поставлен таким образом, что различить влияние двух (или большего количества) соответствующих коррелированных факторов невозможно. Только использование дополнительной априорной информации может решить вопрос о том, какой фактор "на самом деле" должен войти в модель. Применение методов планирования эксперимента может помочь избежать подобных нежелательных ситуаций.

Подробнее о негативных последствиях при наличии взаимосвязи между факторами сказано в разделе 7.2.

5.1.2. Оценка характера, знака и степени влияния факторов на отклик

Характер влияния фактора x_j на отклик определяется наличием в модели квадратичного эффекта x_j^2. Если эффект x_j^2 отсутствует, то фактор x_j влияет на отклик линейно, в противном случае - криволинейно (параболически). Знак коэффициента A_{jj} при эффекте x_j^2 определяет направление ветвей параболы:

- $A_{jj}>0$ - парабола с минимумом (в форме буквы U);

- $A_{jj}<0$ - парабола с максимумом.

Значение x_j, при котором достигается экстремальное значение отклика, может находиться вне области эксперимента. В этом случае зависимость отклика от x_j будет монотонной, хотя и криволинейной.

В линейной модели знак влияния фактора x_j на отклик определяется значением коэффициента A_j при x_j. Если A_j>0, то при увеличении фактора x_j значение отклика растет, A_j<0 - значение отклика снижается при увеличении x_j. В линейной модели степень влияния фактора на отклик определяется величиной

$$L_j = A_j I_j \qquad (5.1)$$

где I_j - длина интервала варьирования фактора x_j (максимальное значение x_j в зоне эксперимента минус минимальное значение x_j). Удобно, что величина L_j измеряется в тех же единицах, что и отклик.

Задача об оценке роли фактора x_j в многофакторной модели (оценка знака и степени влияния) существенно усложняется, если в модели присутствуют эффекты типа $x_j x_i$, характеризующие взаимодействия факторов. В этом случае степень влияния x_j на отклик (а зачастую и знак этого влияния) зависят от значений других факторов, то есть в одной зоне факторного пространства фактор может влиять сильнее, в другой - слабее; в одной - повышать отклик, в другой - понижать. Поэтому можно говорить лишь о локальном влиянии фактора x_j на отклик в определенной зоне факторного пространства. Как правило, важно анализировать влияние факторов на отклик в зоне достижения экстремальных значений отклика (см. раздел 5.1.4).

5.1.3. Прогноз значения отклика для любых значений факторов

С использованием построенной регрессионной модели можно определить расчетное значение отклика и 95%-ный доверительный интервал для любых значений факторов. Соответствующий расчет принято называть интерполяцией, если точка факторного пространства лежит

близко к экспериментальным данным (точкам плана) и экстраполяцией в противном случае.

Мерой близости проверяемой точки к точкам плана является показатель экстраполяции V, определенный в разделе 4.4. В разделах 4.4 и 4.6 описан смысл этой величины. Для точек плана V принимает значения $0 \le V \le 1$. Соответственно, принято считать, что значения показателя экстраполяции для проверяемой точки $0 \le V \le 1$ говорят о том, что расчет значения отклика в этой точке является интерполяцией; $V > 1$ - экстраполяцией (причем, чем больше значение V, тем более далекой экстраполяцией). Отметим, что значения $V>1$ могут быть получены и для точек, формально лежащих внутри области эксперимента, но находящихся в стороне от точек плана. На рис. 5.1 показана точка A, лежащая внутри области эксперимента, но в стороне от точек плана. Величина V для точки A составляет 17.

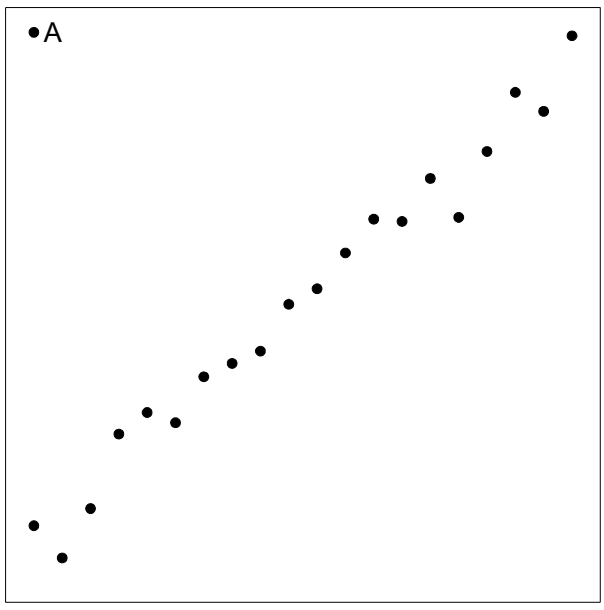


Рисунок 5.1. Пример точки, лежащей внутри области эксперимента, но в стороне от точек плана

Величина доверительного интервала пропорциональна $\sqrt{V}$. Необходимо отметить, что, если законность проведения интерполяции не вызывает никаких сомнений, то экстраполяция, особенно, далекая экстраполяция, является весьма рискованной операцией. Экстраполяция опирается на предположение, что вне области эксперимента функция отклика ведет себя так же, как и внутри области эксперимента, хотя для выдвижения такого предположения нет никаких оснований. Даже экстраполяция по одному фактору на расстояние одного интервала варьирования (при поддержании значений других факторов в зоне эксперимента) может дать ошибочные результаты. Поэтому полученное расчетное значение отклика должно быть обязательно проверено с точки зрения здравого смысла. Часто при экстраполяции получаются бессмысленные результаты (отрицательные значения для заведомо положительных величин и т.д.) или чрезмерно широкие доверительные интервалы, особенно при построении модели в виде полиномов высоких порядков или, если модель содержит избыточные факторы. Следует применять экстраполяцию с очень большой осторожностью. Чрезмерно широкие доверительные интервалы сигнализируют о ненадежном расчете значения отклика.

Для повышения надежности экстраполяции можно применять следующую стратегию. Допустим, если нужно добиться увеличения значения отклика, то следует стремиться к экстраполяции, при которой наблюдается рост нижней границы доверительного интервала, а не рост предсказанного значения отклика.

5.1.4. Поиск и анализ оптимальных значений отклика. Решение компромиссных задач

Как уже отмечалось ранее после построения модели в виде полинома второго порядка можно осуществить поиск значений факторов, обеспечивающих в области эксперимента глобальные экстремальные (минимальное и максимальное) значения отклика (диссоциативно-шаговым методом [2]). В процессе поиска может быть получена также информация о всех локальных экстремумах функции отклика. Нужно определить значение экстремума - расчетное значение функции отклика, соответствующий 95%-ный доверительный интервал и значения факторов, при которых достигается экстремум.

Обычно представляет интерес какой-нибудь один из экстремумов (соответствующий только минимальному или только максимальному значению отклика). Тогда можно провести анализ влияния факторов в зоне соответствующего экстремума (пусть для определенности нас интересует максимум функции). Если значение фактора x_j лежит на нижней границе интервала варьирования;, то можно считать, что в зоне максимума фактор x_j снижает отклик; если значение x_j лежит на верхней границе, то в зоне максимума фактор x_j повышает отклик. Для сравнения силы влияния различных факторов в зоне экстремума применяется следующий прием [2]: фиксируются значения всех факторов, соответствующие экстремальному значению отклика, кроме фактора x_j, а фактор x_j принимает наихудшее с точки зрения рассматриваемого экстремума значение. Полученная таким образом величина L_j отклонения значения отклика от экстремального (в связи с изменением фактора x_j) является мерой влияния фактора x_j в зоне экстремума. Величина L_j измеряется в тех же единицах, что и отклик и является обобщением величины L_j, определяемой по формуле (5.1). Для

удобства сравнения степени влияния факторов на отклик между собой величины L_j можно вычислить значения величин L_j также и в условных процентах (сумма всех L_j принимается за 100 %).

Если количество откликов - два или больше, то могут быть решены компромиссные задачи, являющиеся важнейшими и наиболее сложными технико-экономическими задачами [2]. Для качественного решения подобных задач важна четкая формализация: назначение приоритета различным откликам, учет уже достигнутых уровней оптимальности, формулировка обобщенного критерия оптимальности. Для формальной одновременной оптимизации обобщенной многооткликовой модели в можно использовать следующий алгоритм. Последовательно строятся полиномиальные модели для каждого s-го отклика $\hat{y}_s$. Вычисляется обобщенная модель, равная взвешенной сумме частных моделей $\hat{y}_s$:

$$\hat{y}_{nnn} = \sum_{s=1}^{S} \left\{ \frac{(-1)^{G_s}\ \hat{y}_s}{\sqrt{MSE_s}} \right\}, \qquad (5.2)$$

где S - общее количество откликов;

$\sqrt{MSE_s}$ - средний стандарт ошибки предсказания для s-ой модели.

Согласно формуле (5.2) частная модель $\hat{y}_s$ с хорошей предсказывающей способностью входит в обобщенную модель с большим весом). Слагаемое в формуле (5.2) берется со знаком “+”, если желательным является максимум $\hat{y}_s$, и со знаком “-” в противном случае.

После получения обобщенной модели осуществляется поиск значений факторов, обеспечивающих максимум обобщенной модели (5.2). Для этих значений факторов вычисляются также расчетные значения частных откликов $\hat{y}_s$ (в натуральных единицах и в условных процентах по шкале экстремумов частной модели $\hat{y}_s$ в зоне эксперимента).

Рассмотрим некоторые другие подходы к решению компромиссных задач. Формальная одновременная оптимизация многооткликовой модели на основе так называемой “функции желательности” предложена Харрингтоном [1] в следующем виде:

$$D = \sqrt[S]{\prod_{s=1}^{S} d_s} ,$$

где D - обобщенная функция желательности;

d_s - функция желательности s-го отклика.

Каждое d_s принимает значение от нуля до единицы в соответствии со шкалой желательности для s-го отклика; причем d_s=0, когда $\hat{y}_s$ принимает нежелательные значения; d_s=1, когда $\hat{y}_s$ принимает наиболее желательные значения. Ясно, что D, являющееся средним геометрическим d_s, также принимает значения от 0 до 1, то есть является функцией желательности.

Часто задают компромиссную задачу в виде “оптимизация одного отклика, при ограничениях на другие отклики”. Задачи такого типа решают с применением сложной техники многомерной нелинейной оптимизации (например, методом скользящего допуска или методом штрафных функций [8]).

Рассмотрим применение метода штрафных функций для задачи

$$\begin{cases} \hat{y}_1 \to \min \\ \hat{y}_2 \ge A_2 \\ \dots \\ \hat{y}_S \ge A_S \end{cases} \qquad (5.3)$$

Для этого требуется найти значения факторов, минимизирующих функционал

$$\Omega = \hat{y}_1 + \sum_{s=2}^{S} Z_s ,$$

где $Z_s = \begin{cases} 0 & \hat{y}_s \geq A_s \\ W_s(A_s - \hat{y}_s) & \hat{y}_s < A_s \end{cases}$;

W_s - штраф за нарушение s-го неравенства.

Зачастую сложно корректно поставить задачу в ситуации, когда требуется одновременно оптимизировать два или несколько откликов. Облегчить формализацию подобной задачи для случая откликов может позволить следующий методический прием. Рассмотрим, например, задачу (5.3) при S=2 и построим в координатах $(\hat{y}_2, \hat{y}_1)$ график следующую функцию F($\hat{y}_2$):

F(A) = min($\hat{y}_1$) при условии, $\hat{y}_2 \geq$A.

По графику можно выбрать наилучшее соотношение $(\hat{y}_2, \hat{y}_1)$, а затем и соответствующий режим, обеспечивающий эти значения отклика.

5.1.5. Графическое представление модели

Под графическим представлением модели далее понимается изображение в графическом виде функции отклика (сюда не относятся графики остатков от модели, предсказанных значений отклика и т.д., используемые для анализа качества модели и описанные в разделе 4.3.1). Графическое представление очень наглядно (что особенно важно для пользователя с трудом представляющего себе уравнение регрессии) и может оказать существенную помощь при решении задач управления (в том числе компромиссных задач) и задач типа экономии ресурса.

Проще всего построить графические однофакторные зависимости отклика (откликов) при фиксированных значениях других факторов. По такому графику можно выбрать значение фактора, обеспечивающее заданные значения по нескольким откликам. По оси абсцисс можно откладывать не только изменения значения одного фактора, но и движение вдоль любой прямой в факторном пространстве.

Для решения задач типа экономии ресурсов удобно строить линии уровня поверхности отклика $\hat{y}$. То есть в координатах (x_j, x_i), где x_j, x_i - два выбранных фактора (значения остальных факторов фиксированы) изображаются кривые, для каждой точки которых, выполняется условие $\hat{y} = C_h$, где C_h - значение h-го уровня. Анализ подобных кривых позволяет выбрать значения факторов (x_j, x_i) таким образом, что $\hat{y} \geq C_h$, а расход ресурсов x_j, x_i минимален. Такая же с математической точки зрения задача может формулироваться, например, следующим образом: выбрать наиболее удобный с технологической точки зрения режим (x_j, x_i) (если x_j, x_i характеризуют режим изготовления или обработки изделия), обеспечивающий $\hat{y} \geq C_h$.

5.2. Концепция хорошо интерпретируемой модели

Основываясь на механизме проверки гипотез, Ефроимсон предложил концепцию “наилучшей” модели. Разработаны различные статистические критерии и методы достижения “наилучшей” модели [1].

Идея “наилучшей” модели не является единственно возможной. В.В.Федоров предложил идею полимодельности, когда различные модели могут отражать различные свойства изучаемого объекта. Н.Дрейпер, Г.Смит [1] подразделяют функциональную модель, модель для управления, модель для предсказания. В.А.Вознесенский [2] предлагает использовать несколько моделей для различных уровней значимости (90%, 95%, 99%) и делать на основе этих моделей, соответственно, рискованные, надежные и очень надежные выводы.

В настоящей работе рассматривается концепция хорошо интерпретируемой модели, основанная на идеях [9, 10].

Необходимо затратить известные усилия, чтобы получить регрессионную модель, позволяющую сделать содержательные и надежные прикладные выводы. Под “хорошо интерпретируемой моделью” понимается регрессионная модель вместе с соответствующими прикладными рекомендациями [9, 10]. Качество традиционной регрессионной модели определяется чисто ее статистическими характеристиками, (скажем, чем выше коэффициент множественной детерминации R^2, тем модель считается лучше). Качество хорошо интерпретируемой модели определяется ценностью и надежностью соответствующих прикладных рекомендаций. Ценность рекомендаций - трудно формализуемая величина, выражаемая в терминах конкретной прикладной задачи. Надежность рекомендаций определяется согласием полученных выводов с имеющейся априорной информацией и статистическими характеристиками модели. При этом не надо стремиться получить статистические характеристики как можно выше - достаточно построить адекватную, статистически значимую, качественную (см. раздел 3) модель. Разумеется, лучше иметь содержательную модель с чуть худшими характеристиками, чем бесполезную с прикладной точки зрения модель с прекрасными статистическими характеристиками. Не бывает верных моделей, но бывают полезные [1]. Конечно, все вышесказанное не означает, что надо стремиться подогнать модель под априорные представления.

Процесс построения хорошо интерпретируемой модели включает следующие стадии.

Основа для последующего построения хорошо интерпретируемой модели закладывается на этапе предварительного анализа экспериментальной информации (раздел 3). Формализация априорной информации позволяет четко установить цели исследования и показатели

качества (с содержательной точки зрения) для построения модели. Графическое представление показывает структуру и качество экспериментальных данных. Выбор вида модели во многом определяет возможности последующей интерпретации.

На этапе собственно построения модели следует стремиться построить несколько различных регрессионных моделей, близких по статистическим характеристикам и выбрать лучшую из них с точки зрения содержательности. Прежде, чем оценивать модель с точки зрения содержательности, необходимо провести анализ качества регрессионной модели, в соответствии с рекомендациями раздела 3.

Заключительным этапом получения хорошо интерпретируемой модели является собственно тщательная интерпретация уравнения регрессии (раздел 5.1), проверка согласия полученных рекомендаций с имеющейся априорной информацией. Особое внимание следует обратить на проведение анализа корреляционной матрицы факторов и коэффициентов корреляции эффектов с откликом (п. 5.1.1). Несоблюдение этого положения может привести к ошибочным выводам.

6. УПРАВЛЕНИЕ ВЫБОРОМ МОДЕЛИ

Одним из лучших автоматических методов получения “наилучшей” регрессионной модели является шаговый метод включения-исключения [1]. Этот эмпирический метод прекрасно зарекомендовал себя при обработке самых разных групп данных. Но при наличии возможности интерактивного построения регрессионной модели не следует ограничиваться автоматическими методами. Полученная регрессионная модель должна быть не только адекватной и значимой. Необходимо проверять качество модели (раздел 4), соответствие модели априорной информации (раздел 4). Надо стремиться получить высококачественную содержательную модель. Поэтому метод не должен быть “черным ящиком” - пользователь должен отчетливо представлять себе, как он работает, какие есть средства для управления выбором модели.

Шаговый метод ориентирован на получение статистически значимой модели (раздел 4.1). При наличии информации о дисперсии воспроизводимости производится проверка гипотезы об адекватности модели (раздел 4.2).

При последовательном увеличении количества членов модели показатели значимости параметров модели и адекватности обычно ведут себя следующим образом. Модель, содержащая малое количество наиболее сильно влияющих на отклик эффектов, как правило, будет значимой (все параметры значимы) и неадекватной (среднее отклонение расчетных значений отклика от экспериментальных существенно превышает ошибку воспроизводимости). С увеличением количества членов модели значимость параметров модели будет иметь тенденцию к снижению (и, начиная с некоторого количества членов, модель станет незначимой), а отклонение экспериментальных и расчетных значений отклика будет уменьшаться (и, начиная с некоторого количества членов, модель станет

адекватной}. Наилучшей является ситуация, когда можно получить значимую и адекватную модель, причем такую, что при исключении одного из членов, она становится неадекватной, а при включении новых членов - незначимой. Но такую модель можно построить не всегда.

6.1. Использование центрирования факторов

При построении полиномиальных регрессионных моделей часто возникают большие коэффициенты корреляции между эффектами вида x_j и x_j^2, x_j и $x_j x_i$, и т.д. Это приводит к тому, что затрудняется включение в модель коррелированных эффектов. Если коррелированные эффекты все же включаются в модель, то возникает мультиколлинеарность со всеми ее негативными последствиями (см. разделы 4.5 и 7.2). Причем сильная корреляция между некоторыми эффектами в полиномиальных моделях обусловлена не спецификой сбора экспериментальных данных или фактически существующими взаимосвязями между факторами (как для линейных моделей), а неудачным выбором вида модели.

Использование центрированных значений факторов для образования эффектов при построении полиномиальной модели второго порядка позволяет, как правило, существенно снизить соответствующие коэффициенты корреляции.

Центрированным значением j-го фактора в i-ом опыте называется

$$x_{ij}^n = x_{ij} - \overline{x}_j,$$

где $\overline{x}_j$ - среднее значение j-го фактора в эксперименте

$$\overline{x}_j = \frac{\sum_{i=1}^{N} x_{ij}}{N}$$

При построении регрессионной модели можно использовать два различных режима центрирования факторов.

Для первого из этих режимов факторы центрируются только при образовании квадратичных эффектов, то есть используются эффекты следующего вида :

$$\begin{aligned} &E = x_j; \\ &E = x_j x_i; \\ &E = (x_j - \overline{x}_j)^2. \end{aligned} \qquad (6.1)$$

Для второго режима факторы центрируются при образовании любых эффектов

$$\begin{aligned} &E = x_j - \overline{x}_j; \\ &E = (x_j - \overline{x}_j)(x_i - \overline{x}_i); \\ &E = (x_j - \overline{x}_j)^2. \end{aligned} \qquad (6.2)$$

При использовании режимов центрирования факторов происходит расчет модели с участием эффектов соответствующего вида. Для окончательной модели можно произвести пересчет к натуральным значениям факторов.

С точки зрения последующей интерпретации модели предпочтительнее использовать режим (6.1). Эффект вида $(x_j - \overline{x}_j)(x_i - \overline{x}_i)$ затруднительно интерпретировать. Пусть, например, соответствующий параметр в модели положителен. Тогда, это должно означать следующее: отклик повышается при одновременном увеличении или уменьшении x_j и x_i; отклик снижается, если x_j увеличивают, а x_i уменьшают или x_j уменьшают, а x_i увеличивают. Достаточно редко встречаются процессы, которые в действительности устроены подобным образом. Присутствие в модели также и линейных эффектов $(x_j - \overline{x}_j)$ и $(x_i - \overline{x}_i)$ делает интерпретацию эффекта $(x_j - \overline{x}_j)(x_i - \overline{x}_i)$ более осмысленной.

Но есть две часто встречающиеся ситуации, в которых лучше использовать режим (6.2):

1. Построение модели по данным спланированного эксперимента. Известно, что для использования преимуществ таких данных (к которым относятся, в частности, небольшие значения коэффициентов корреляции между эффектами) рекомендуется строить модель для кодированных значений факторов. Кодированным значением j-го фактора в i-ом опыте называется

$$x_{ij}^{K} = \frac{x_{ij} - ОУ_j}{ИВ_j},$$

$ОУ_j$ - основной уровень;

$ИВ_j$ - полуширина интервала варьирования j-го фактора.

Аналогичный результат (использование преимуществ, связанных с применением планирования эксперимента) дает построение модели с использованием режима центрирования факторов (6.2), так как для спланированных данных обычно $ОУ_j \approx \overline{x}_j$.

2. Имеются факторы, для которых выполняется условие $\overline{x}_j >> \sqrt{D_j}$, где $\sqrt{D_j}$ - характеристика разброса j-го фактора. Это означает, что длина интервала варьирования фактора существенно меньше, чем его среднее значение (например, в эксперименте температура меняется в диапазоне 990-1010 °C; длина интервала варьирования, равная 20 °C, существенно меньше среднего значения, равного 1000 °C). В этой ситуации возникают очень большие коэффициенты корреляции не только между эффектами x_j и x_j^2, но и между x_j и $x_j x_i$. Для уменьшения этих коэффициентов корреляции нужно использовать режим центрирования факторов (6.2).

6.2. Использование данных “пассивного” эксперимента

Принято различать данные, полученные методом активного эксперимента (то есть с применением методов планирования эксперимента) и методом пассивного эксперимента (то есть просто зарегистрированные и собранные данные). Данные, полученные методом пассивного эксперимента, могут иметь следующие типовые дефекты:

- наличие ошибок как в зависимых, так и в независимых переменных;
- сильная корреляция между факторами;
- важные факторы могут иметь слишком узкие интервалы варьирования;
- выборка может быть непредставительной по отношению к генеральной совокупности.

Таким образом, данные, полученные методом пассивного эксперимента, обычно хуже (с точки зрения построения регрессионной модели), чем спланированные данные. Большие объемы выборок могут несколько компенсировать недостатки, присущие собранным данным. С учетом этого соображения в [1] предлагается, чтобы модель, построенная по данным пассивного эксперимента, содержала не более, чем N/10 членов (то есть было бы порядка 10 наблюдений на каждый член модели).

7. РАСШИРЕНИЯ КЛАССИЧЕСКОГО РЕГРЕССИОННОГО АНАЛИЗА

7.1. Основные предположения в регрессионном анализе

Основные допущения, которые делают при проведении регрессионного анализа, следующие:

1. Функция отклика действительно описывается линейной по параметрам зависимостью от эффектов вида

$$Y = A_0 + \sum_s A_s E_s .$$

2. При регистрации значений отклика возникают ошибки ε. Фактически наблюдаемое значение отклика записывается в виде:

$$Y_i^{набл} = Y_i^{эксп} + \varepsilon .$$

При этом предполагается, что значения факторов измеряются точно.

3. Ошибка ε является случайной величиной с нулевым средним, постоянной дисперсией σ_e^2; значения ошибки в различных опытах не коррелируют между собой (независимы).

4. Ошибки ε подчиняются нормальному распределению N(0, σ_e^2).

Предположение 1 является фундаментальным. Нарушение этого предположения называется неадекватностью. Техника проверки гипотезы о неадекватности модели описана в разделе 4.2.1.

Предположение 2 означает, что значения факторов измеряются точно, а отклик - с ошибкой. На практике требование отсутствия ошибок в измерении фактора эквивалентно выполнению условия [1]

$$\Delta_y >> A_j \Delta_j$$

для модели

$$Y = A_0 + \sum_j A_j x_j ,$$

где Δ_y и Δ_j - экспериментальная ошибка определения отклика и j-го фактора, соответственно.

Если же ошибками в независимых переменных пренебречь нельзя, то можно воспользоваться специальными методами типа “квазиортогональной регрессии” [4], [6] (так как использование обычного метода наименьших квадратов дает модели со смещенными значениями параметров) .

Анализ остатков (раздел 4.3) позволяет проверить нарушения предположений 3 и 4. Существует так называемый робастный подход к построению регрессионных моделей [7], для которого не требуется выполнения предположений 3 и 4.

7.2. Мультиколлинеарность

Мультиколлинеарностью называется почти точная линейная зависимость в корреляционной матрице эффектов, вошедших в модель (сильная взаимосвязь эффектов модели) .

Мерой мультиколлинеарности является максимальный из коэффициентов множественной корреляции эффекта в модели с остальными эффектами, вошедшими в модель. Если он превышает 0.95, то говорят о сильной мультиколлинеарности. Если же он превышает коэффициент множественной корреляции отклика с эффектами (равный $\sqrt{R^2}$, где R^2 - коэффициент множественной детерминации), то взаимосвязь отклика с эффектами слабее, чем взаимосвязь эффектов между собой.

Мерой участия эффекта в мультиколлинеарности является коэффициент множественной корреляции этого эффекта с остальными эффектами в модели. Аналогично, если он превышает 0.95, то эффект сильно связан с другими эффектами.

Если только два коэффициента превышают 0.95, то мультиколлинеарность обусловлена сильной взаимосвязью этих эффектов между собой.

Может встретиться следующая ситуация, обусловленная мультиколлинеарностью: коэффициент парной корреляции эффекта и отклика и коэффициент для эффекта в модели имеют противоположные знаки. Поясним эту ситуацию на следующем примере. Пусть фактор x_1 имеет положительный и даже довольно значительный коэффициент корреляции с откликом $K(x_1, y) = 0.5$. Однако, фактор x_1 тесно связан с фактором x_2 ($K(x_1, x_2) = 0.8$), который в свою очередь тесно связан с откликом ($K(x_2, y) = 0.7$). Тогда знак параметра A_1 в модели

$$Y = A_0 + A_1 x_1 + A_2 x_2$$

будет определяться знаком частного коэффициента корреляции x_1 с откликом (при устраненном влиянии x_2). Частный коэффициент корреляции вычисляется по формуле [10]:

$$\tilde{K}(x_1, y) = \frac{K(x_1, y) - K(x_2, y)K(x_1, x_2)}{\sqrt{(1 - K(x_2, y)^2)(1 - K(x_1, x_2)^2)}}.$$

В рассматриваемом примере $\tilde{K}(x_1, y) = -0.14$ и, соответственно, $A_1 < 0$.

Отрицательные последствия мультиколлинеарности проявляются в следующем [3]:

- неустойчивость оценок, проявляющаяся в том, что добавление или исключение небольшого количества опытов может привести к резкому изменению параметров модели;
- численная неустойчивость расчета оценок, вызванная djpvj;ysvb ошибками машинного округления при работе с почти вырожденными матрицами;

- могут быть получены абсурдные значения параметров модели (ввиду их сильной корреляции между собой), что делает невозможной их разумную интерпретацию;
- резко увеличиваются фактические дисперсии параметров модели (причем увеличение дисперсий оценок может не проявляться в значениях t-статистик параметров, так как общая доверительная область в таких ситуациях обычно является узким, сильно вытянутым эллипсоидом).

Отметим, что аппроксимирующие свойства регрессии в условиях мультиколлинеарности остаются на высоком уровне. Таким образом, если использовать уравнение регрессии только для предсказания, то на мультиколлинеарность можно не обращать внимания. Но прогноз с помощью этой модели может производиться только для данных, расположенных в том же узком диапазоне, что и экспериментальные данные, использовавшиеся для построения модели. Выход за пределы экспериментальной области легко диагностировать по большим значениям (больше единицы) показателя экстраполяции (раздел 5.1.3) и резкому увеличению доверительного интервала предсказанного значения отклика.

Основные причины появления мультиколлинеарности:

1. Наличие реальной функциональной взаимосвязи между факторами. Например, в смесях всегда концентраций компонентов равна 100 %. Часто присутствует ложная мультиколлинеарность между несколькими переменными, связанная с тем, что все они выражаются через некоторую другую истинную независимую переменную. Причем эта мультиколлинеарность будет присутствовать для любого конкретного набора экспериментальных данных. В этом случае надо удалить из модели эффекты, вызывающие мультиколлинеарность, связанную с функциональной взаимосвязью.

2. Ограниченность данных. Имеется в виду, что экспериментальные данные лежат в узком подпространстве факторного пространства. Эта мультиколлинеарность связана с определенным набором данных и, возможно, не будет присутствовать в других данных. Мультиколлинеарность этого вида встречается очень часто.

Самое лучшее в таких ситуациях использовать метод планирования эксперимента (в том числе и для пополнения уже имеющихся экспериментальных данных [3]). Если этого сделать нельзя, то необходимо использовать априорную информацию о параметрах модели или использовать RIDGE-регрессию.

В некоторых случаях встречается другой вид мультиколлинеарности, связанный с данными, - мультиколлинеарность, порожденная наличием небольшого количества точек, далеко лежащих в пространстве факторов по отношению к остальным. Наличие таких точек легко обнаружить по близким к единице значениям диагональных элементов матрицы подгонки (см. раздел 4.4.1). Удаление таких точек позволяет устранить мультиколлинеарность этого вида.

3. Неудачный выбор вида модели. Возникает в двух типовых ситуациях. Во-первых, при построении регрессионных моделей в виде полинома второго порядка часто возникают большие коэффициенты корреляции между эффектами x_j и x_j^2, x_j и $x_j x_i$. Эта мультиколлинеарность носит искусственный характер. Использование центрированных значений факторов для образования эффектов (раздел 6.1) позволяет, как правило, существенно снизить эту мультиколлинеарность.

Вторая типовая ситуация - переопределенная модель, содержащая чрезмерно большое количество членов (см. раздел 4.6). О том, что модель переопределена, свидетельствуют абсурдные значения экстремальных значений отклика, слабый прирост коэффициента множественной

детерминации R^2 на заключительных этапах построения модели (порядка 1 %), малое количество степеней свободы и, соответственно, завышенное значение R^2 (близкое к единице). Рекомендуется при построении модели по данным пассивного эксперимента включать в нее не более, чем N/10 членов, где N - количество опытов. Для борьбы с мультиколлинеарностью этого вида достаточно удалить лишние члены из модели. Техника регулирования количества членов в модели описана в разделе 5.

При построении модели в ситуациях 1 и 2 в последнее время часто применяют так называемую RIDGE-регрессию [1], [4], [6]. Смысл RIDGE-регрессии состоит в следующем. Для ситуации мультиколлинеарности корреляционная матрица эффектов в модели является практически вырожденной. Предлагается использовать специальную небольшую положительную добавку к диагональным элементам этой матрицы для улучшения ее обусловленности. Разработаны методики для выбора оптимального значения величины добавки. Полученные параметры уравнения регрессии будут смещенными, но могут иметь существенно более низкую дисперсию ошибки, чем стандартно определяемые параметры.

8. ЗАДАНИЯ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ

Таблица 8.1. Данные для самостоятельной работы

X_1	X_2	Y
-1	-1	4.25
-0.9	0.1	-2.38
-1.05	1	-6.88
0.1	-1.1	1.32
-0.05	0	1.49
0	0.9	0.75
1	-0.95	-1.49
0.95	-0.05	2.65
1.1	1.05	8.66

Дисперсия воспроизводимости равна 1;
количество степеней свободы – 10.

Для данных, приведенных в таблице, построить регрессионные модели:

1. $Y = A + B \cdot X_1 + C \cdot X_2$
2. $Y = A + B \cdot X_1 + C \cdot X_1 \cdot X_2$
3. $Y = A + B \cdot X_1 + C \cdot X_2 + D \cdot X_1 \cdot X_2$
4. $Y = A + B \cdot X_1 + C \cdot X_2 + D \cdot X_1^2$

Ответьте на следующие вопросы:

1. Какая из этих моделей является адекватной?

2. Какая из этих моделей является значимой (все параметры значимы)?

3. Какая из этих моделей является наилучшей? Почему?

4. Каким образом факторы X_1, X_2 влияют на отклик Y?

5. Содержат ли данные выбросы или влиятельные наблюдения?

9. ВОПРОСЫ ДЛЯ САМОКОНТРОЛЯ

1. В чем заключается анализ качества модели?

2. Что делать, если модель незначима?

3. Что такое «адекватность модели»?

4. Может ли регрессионная модель быть значима, но не адекватна? Адекватна, но не значима?

5. Что делать, если модель не адекватна?

6. В чем заключается анализ остатков?

7. Что такое «выбросы»? Почему нежелательно наличие выбросов в данных?

8. Что такое «влиятельные наблюдения»?

9. Что такое «мультиколлинеарность»?

ЗАКЛЮЧЕНИЕ

В работе представлены методические приемы, использующиеся на практике при применении регрессионного анализа. Главные рассматриваемые вопросы:

- выбор модели;
- интерпретация регрессионных моделей;
- анализ качества модели;
- управление выбором модели.

Даны основные приемы для построения хорошо интерпретируемых моделей.

ЛИТЕРАТУРА

1. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. В 2-х книгах. М.: Финансы и статистика, т.1 - 1986, т.2 - 1987.

2. Вознесенский В.А. Статистические методы планирования эксперимента в технико-экономических исследованиях. М.: Финансы и статистика, 1981.

3. Вучков И, Бояджиева Л., Солаков Е. Прикладной линейный регрессионный анализ. М. :Финансы и статистика, 1987.

4. Демиденко Е.З. Линейная и нелинейная регрессии. М.: Финансы и статистика, 1981.

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

6. Петрович М.Л. Регрессионный анализ и его математическое обеспечение. М.: Финансы и статистика, 1982.

7. Хьюбер П. Робастность в статистике. М.: Мир, 1984.

8. Химмельблау Д. Прикладное нелинейное программирование. М.: Мир, 1975.

9. Васильев В.В., Ихильчик А.Р. Некоторые вопросы построения хорошо интерпретируемых статистических моделей физико-металлургических процессов. В книге “Математическое моделирование металлургических и сварочных процессов”. М.: Металлургия, 1986.

10. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Исследование зависимостей. М.: Финансы и статистика, 1985.

ПРИЛОЖЕНИЕ 1

ПОРЯДОК ПОСТРОЕНИЯ И ИНТЕРПРЕТАЦИИ РЕГРЕССИОННЫХ МОДЕЛЕЙ

К построению регрессионной модели следует приступать только после тщательной проверки экспериментальных данных. Рекомендуется использовать графическое представление экспериментальных данных. Некоторые грубые ошибки в данных могут быть выявлены и при построении регрессионной модели.

Рекомендуется хотя бы бегло просмотреть значения средних и дисперсий для факторов к отклика - соответствуют ли их значения истине.

Следует внимательно проанализировать корреляционную матрицу факторов. Надо учитывать, что, если для двух факторов абсолютная величина коэффициента корреляции $|K|$ достаточно велика, то возникают следующие ситуации:

1. $|K|=1$ - факторы совершенно неразличимы при построении регрессионной модели;

2. $|K|\geq 0.95$ - факторы практически неразличимы;

3. $|K|\geq 0.90$ - факторы трудноразличимы;

4. $|K|$ превышает соответствующее пороговое значение - факторы достаточно тесно связаны.

Формальный статический подход в первых трех ситуациях невозможен. Решить вопрос о том, какой именно из двух (или нескольких) коррелированных факторов должен войти в модель, можно только с использованием априорной информации.

Если обрабатываются данные, полученные с применением метода планирования эксперимента, то необходимо проверить соответствие корреляционной матрицы с теоретической.

По коэффициентам парной корреляции факторов с откликом можно грубо выделить:

- факторы, оказывающие наибольшее влияние на отклик (фактор с максимальным значением $|K|$, а также факторы имеющие близкие значения $|K|$);

- факторы, оказывающие заметное влияние на отклик (факторы, для которых $|K|$ превышает пороговое значение);

- факторы, слабо влияющие на отклик (факторы, для которых $|K| \approx 0$).

Следует отметить, что для двух факторов, у которых между собой $|K| \geq 0.9$ заведомо будут близкие значения коэффициентов корреляции с откликом.

В разделе 6.2 описана ситуация, когда фактор коррелирует с откликом с одним знаком, а в модель входит с противоположным. Такая перемена знака происходит из-за сильной взаимосвязи фактора с остальными членами модели.

Для выбранного для включения в модель набора факторов определяются значения параметров и следующие статистические характеристики модели:

- t-статистики параметров;
- F-отношение для проверки общей значимости модели;
- коэффициент множественной детерминации R^2;
- остаточная сумма квадратов;
- остаточная дисперсия MSE.

Если используется дисперсия воспроизводимости, то производится проверка адекватности модели.

Смысл значений t-статистик параметров модели описан в разделе 3.1. Чем выше значение t-статистики, тем более надежно определен параметр. Параметры модели, для которых значение t-статистики ниже порогового (обычно выбирается пороговое значение для уровня надежности 95 %, равное 1.96) незначимы и определены ненадежно.

Коэффициент множественной детерминации R^2 является безразмерной характеристикой качества соответствия расчетных и экспериментальных значений отклика. R^2 - это доля разброса отклика, описываемая регрессией. Желательными являются значения R^2, близкие к единице.

Наличие мультиколлинеарности можно обнаружить при анализе корреляционной матрицы и коэффициентов множественной корреляции факторов, вошедших в модель (см. разделы 3.5, 6.2).

При анализе остатков для каждого опыта необходимо определить:

- экспериментальное значение отклика;
- расчетное значение отклика;
- 95 %-ный доверительный интервал расчетного значения отклика;
- значения остатка;
- значение нормализованного остатка (см. раздел 3.3.2);
- значение студентизованного остатка (см. раздел 3.3.2);
- значение диагонального элемента матрицы подгонки (см. раздел 3.4.1);
- значение меры влияния опыта на параметры модели (см. раздел 3.4.2).

Важны следующие измерения:

- максимальное по абсолютной величине значение остатка (наибольшее значение отклонения экспериментального и расчетного значений отклика);
- максимальное по абсолютной величине значение диагонального элемента матрицы подгонки (измерения, лежащего дальше всех от центра в пространстве факторов);
- максимальное по абсолютной величине значение студентизованного остатка (это измерение хуже всего описывается моделью с учетом его положения в пространстве факторов};
- максимальное значение меры влияния (при удалении этого измерения параметры модели изменятся сильнее всего по сравнению с возможным удалением других измерений);
- все измерения, для которых величина нормализованного остатка превышает по модулю три.

Все эти выделенные измерения должны быть тщательно проанализированы (см. раздел 3.3, 3.4). Необходимо также помнить, что эти измерения - первые кандидаты на наличие в них грубых ошибок.

Для модели в виде полинома второго порядка нужно произвести анализ положений глобальных экстремумов отклика. Техника анализа экстремумов подробно описана в разделах 3.6 и 4.1.4.

Если использовалось центрирование или кодирование факторов для образования эффектов при построении регрессионной модели, то наряду с моделью для преобразованных значений эффектов, нужно вычислить параметры модели для натуральных значений факторов.

ПРИЛОЖЕНИЕ 2

БЛАНКИ ДЛЯ ФОРМАЛИЗАЦИИ АПРИОРНОЙ ИНФОРМАЦИИ

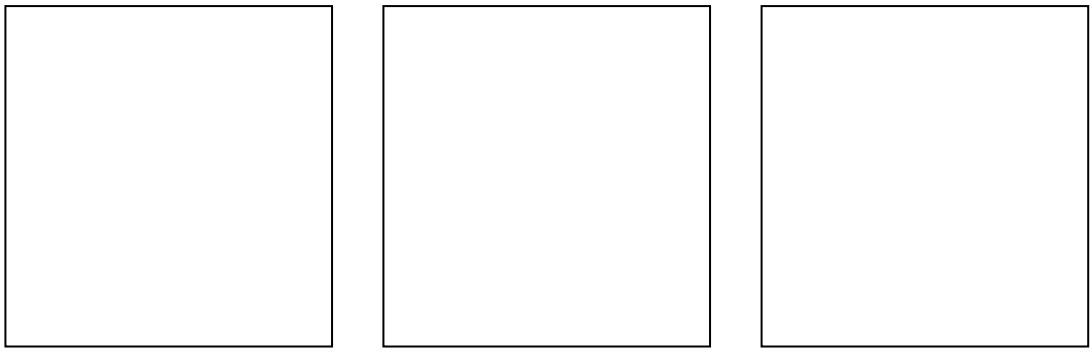
Бланк априорной информации о факторе

1. Наименование
2. Условное обозначение
3. Единица измерения
4. Область существования:

   возможная: минимум максимум

   в эксперименте: минимум максимум
5. Точность управления или стабилизации:

   абсолютная относительная
6. Количество уровней варьирования
7. Качественные однофакторные зависимости фактора в пределах эксперимента на исследуемые выходы:



8. Возможные эффекты взаимодействия с другими факторами

9. Литература

Бланк априорной информации о выходе

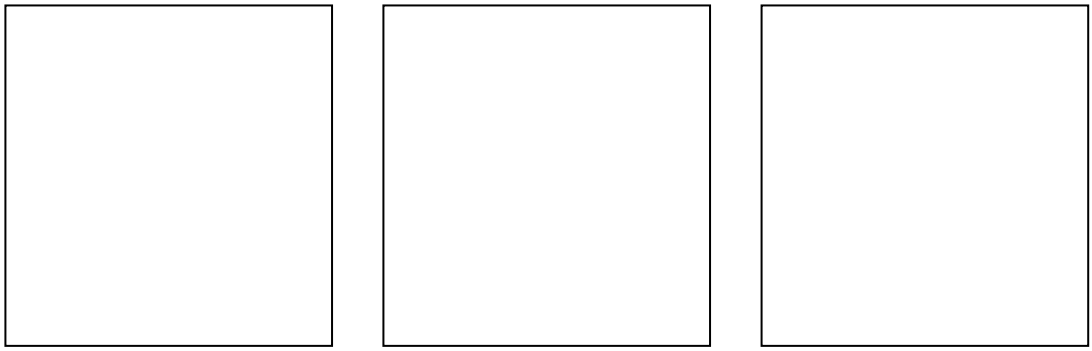
1. Наименование
2. Условное обозначение
3. Единица измерения
4. Лучший результат:

   в экспериментах автора в литературе
5. Целевой уровень
6. Приоритет
7. Шкала оценок: “Удовлетворительно”, “Хорошо”, “Отлично”:

8. Ошибка измерения

   абсолютная относительная
9. Ошибка воспроизводимости

10. Качественные однофакторные зависимости влияния на выход основных факторов в пределах эксперимента:



8. Сведения о взаимосвязи с другими факторами

9. Литература