Министерство образования и науки Российской Федерации
НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

СТАТИСТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА ДАННЫХ

Методические указания к лабораторным работам и РГЗ
для студентов IV курса ФПМИ всех направлений
и специальностей

НОВОСИБИРСК
2017

УДК 519.23(076.5)
С 781

Составитель: *А.А. Попов*, д-р техн. наук, проф.

Рецензент: С.Н. Постовалов, д-р техн. наук, доц.

Работа подготовлена на кафедре теоретической
и прикладной информатики

ЛАБОРАТОРНАЯ РАБОТА № 1

ГЕНЕРАЦИЯ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ ПО СХЕМЕ ИМИТАЦИОННОГО МОДЕЛИРОВАНИЯ НА ЭВМ

1. Методические указания

Для решения задачи идентификации объекта с целью определения его математической модели, связывающей значение выходной переменной y со значениями входных переменных $x_1, x_2,\ldots,x_k$, необходим экспериментальный материал, включающий в себя достаточное количество наблюдений за входом и выходом объекта. В реальных ситуациях экспериментальный материал получают в ходе наблюдения за исследуемым объектом либо в режиме его нормального функционирования, либо в условиях активного эксперимента. В данной лабораторной работе получение необходимого экспериментального материала достигается проведением имитационного моделирования на ЭВМ. В качестве имитационной модели выбирается некоторая линейная по параметрам регрессионная модель, считающаяся истинной моделью исследуемого объекта. Линейная по параметрам θ модель представима в виде $\eta(\underline{x},\theta)=\theta^T f(\underline{x})$, где элементы вектора $f(\underline{x})$ есть действительные известные функции от факторов $\underline{x}=(x_1,...,x_k)^T$, например, $f(\underline{x})=(1,\ x_1,\ x_2,\ x_1x_2)^T$, что определяет линейную модель от двух факторов с их взаимодействием.

Процесс моделирования включает в себя следующие этапы:

1) выбор имитационной модели $\eta(\underline{x},\theta)$, линейной относительно параметров θ;

2) выбор области действия факторов $\underline{x}=(x_1,...,x_k)^T$, доступной для экспериментирования. Обычно эта область задается в виде $x_i\in[x_{\min},x_{\max}],\ i=\overline{1,k}$;

3) выбор необходимого числа n и координат точек проведения наблюдений $\underline{x}_j = (x_{j1},...,x_{jk})^T,\ j = \overline{1,n}$;

4) вычисление значений отклика y_j в выбранных точках $\underline{x}_j$:

$$y_j = u_j + e_j = \eta(\underline{x}_j, \theta) + e_j,\ j = 1,...,n,$$

где e_j – реализация случайной величины (ошибки наблюдений).

Рассмотрим каждый из этих этапов. При выборе модели объекта необходимо учесть предполагаемую характеристику зависимости выхода объекта от входных переменных. В качестве функций $f(\underline{x})$ в модели обычно выбирают: свободный член ($f(\underline{x}) \equiv 1$); линейные от входных факторов ($f(\underline{x}) = x_i,\quad i = 1,...,k$); взаимодействия факторов ($f(\underline{x}) = x_i x_j,\ i, j = 1,...,k$); квадратичные или кубические функции от входных переменных; обратные функции и т.д. В приведенных ниже вариантах заданий сложность имитационной модели будет не выше m = 6 - 10 . Числовые значения параметров $\theta_i,\ i = 1,...,m$ выбираются с учетом характеристик объекта моделирования. Выбранные значения параметров будем называть истинными значениями. В дальнейшем их необходимо будет сравнить с оценочными значениями, полученными в результате обработки экспериментальных данных (см. лаб. работу № 2).

Входные переменные (факторы) $\underline{x} = (x_1,...,x_k)^T$ в задачах моделирования, как правило, относятся к числу управляемых. Диапазон возможного изменения какой-либо входной переменной $x_i,\ i = 1,...,k$ определяется в практических ситуациях соответствующими технологическими ограничениями. В нашем случае, если нет особых требований, будем считать, что $x_i \in [-1,\ +1],\ i = 1,...,k$.

Характерной особенностью имитационного моделирования является возможность достаточно произвольно выбирать число испытаний на исследуемом объекте и условия их проведения. Мы будем придерживаться существующей эмпирической оценки: необходимое число экспериментов должно как минимум в 3–4 раза превышать число оцениваемых параметров в модели (число параметров мо-

дели – это размерность вектора θ). Выбор плана эксперимента (совокупности строк $\underline{x}_i^T$, $i=1,...,n$) является важной задачей, требующей отдельного рассмотрения. Здесь можно следовать общей рекомендации: стараться расположить точки плана $\underline{x}_i^T$, $i=1,...,n$ равномерно в допустимой области.

Выбор характеристик случайных величин *e.* В классических процедурах регрессионного анализа учитывается ряд предположений о свойствах объекта. В их числе требование о нормальном распределении ошибки наблюдения *e.* Для моделирования помехи e_j, $j=1,...,n$ необходимо использовать какой-либо доступный датчик псевдослучайных нормально распределенных величин. Моделируемая помеха должна иметь нулевое математическое ожидание и дисперсию σ^2. Дисперсию σ^2 помехи *e* целесообразно выбирать в виде некоторой доли ρ от мощности ω^2 сигнала $u=\eta(\underline{x},\theta)$. Мощность сигнала определим $\omega^2=(u-\overline{u})^T(u-\overline{u})/(n-1)$, где u – вектор истинных значений отклика, $\overline{u}$ – вектор, все элементы которого есть среднее значение сигнала по выборке. Долю ρ можно брать в пределах 5…15 %. Полученные в результате моделирования значения $(\underline{x}_i, y_i)$, $i=\overline{1,n}$ будут использоваться в следующей лабораторной работе.

2. Перечень вопросов к разработке

1. В соответствии с вариантом задания выбрать имитационную модель объекта, диапазон изменения факторов, план эксперимента.

2. Написать программу по генерации экспериментальных данных. Полученные по программе данные оформить в виде одного или двух файлов унифицированной структуры, доступных для дальнейшей обработки. Построить графики зависимости незашумленного отклика от входных факторов.

3. Оформить отчет, включающий в себя постановку задачи, обоснование принятых решений по выбору модели, порождающей данные, графики зависимости

незашумленного отклика от входных факторов, сгенерированную выборку наблюдений в виде таблицы, характеристики помехи, текст программы.

3. Варианты заданий

1. Произвести моделирование объекта, о котором известно: число факторов – два; по первому фактору зависимость выхода близка к линейной (возрастающей), по второму фактору зависимость близка к параболической. Максимальное значение отклика приходится на внутреннюю точку области действия второго фактора. Отметим, что термин " близка к линейной" не означает в точности линейную зависимость. Моделировать такое свойство объекта можно включением в состав регрессоров линейной функции от фактора и нелинейной. В качестве нелинейной функции может выступать, например, фактор во второй степени, но с малым по величине параметром при нем.

2. Условия задания такие же, как в варианте 1, но считается, что максимум выходной величины приходится на граничные точки области действия факторов.

3. Провести моделирование объекта, о котором известно: число действующих факторов – три; по всем факторам зависимость выхода близка к линейной, взаимодействия первого фактора со вторым и третьим существенны, т.е. соответствующие параметры θ при регрессорах x_1x_2, x_1x_3 значительно отличаются от нулевого значения.

4. О моделируемом объекте известно, что число действующих факторов равно двум, в точке $x_1 = 0, x_2 = 0$ значение выхода u равно 0. По первой переменной зависимость выхода близка к квадратичной, а по второй – возрастающая, близкая к линейной.

5. Условия задания такие же, как в варианте 1, но при этом известно, что первый фактор в эксперименте может варьироваться на четырех уровнях (принимать только четыре разрешенных значений), а второй на пяти уровнях.

6. Условия задания такие же, как в варианте 3, но известно, что первый фактор в эксперименте может варьироваться на трех уровнях, второй фактор варьируется на четырех уровнях, третий фактор на двух уровнях.

7. О моделируемом объекте известно, что число факторов равно двум. По первому фактору зависимость выхода близка к обратной функции, а по второму фактору близка к квадратичной зависимости. Взаимодействие факторов незначимо.

8. Условия задачи такие же, как в варианте 7, но взаимодействие факторов оказывают значимое влияние на выход объекта.

9. Произвести моделирование объекта, о котором известно: число действующих факторов – два; по первому фактору зависимость выхода близка к квадратичной, по второму фактору зависимость близка к кубической.

10. О моделируемом объекте известно, что число действующих факторов равно двум, в точке $x_1 = 0, x_2 = 0$ значение выхода u равно 0. По первой переменной зависимость выхода близка к квадратичной, а по второй близка к кубической. Взаимодействия факторов значимо.

4. Контрольные вопросы

1. Что понимается под терминами "Machine Learning", "Data Mining", "Web Mining". Приведите примеры постановок прикладных задач анализа данных.
2. Назовите разделы многомерного статистического анализа, которые используются в задачах анализа зависимостей.
3. Прикладные цели статистического исследования зависимостей.
4. Этапы решения задачи статистического исследования зависимостей.
5. Перечислить возможные причины того факта, что результаты наблюдения за объектом есть суть случайные величины.
6. Общие требования, предъявляемые к оценкам, вычисляемым по результатам наблюдения над случайными величинами.
7. Что обозначают термины "пассивный" и "активный" эксперимент.

8. Почему этап построения зависимостей под названием "определение класса допустимых решений" называют также этапом параметризации модели.

Примечание. Для защиты лабораторных работ № 1, № 2 можно готовить один совместный отчет.

ЛАБОРАТОРНАЯ РАБОТА № 2

ОЦЕНИВАНИЕ ПАРАМЕТРОВ РЕГРЕССИОННОЙ МОДЕЛИ ПО МЕТОДУ НАИМЕНЬШИХ КВАДРАТОВ

1. Методические указания

Пусть математическая модель объекта принадлежит параметрическому семейству функций

$$E(y / x) = \eta(\underline{x}, \theta) \in F = \left\{ \theta^T f(\underline{x}) \right\}_{\theta \in R^m},$$

где $\underline{x} = (x_1, ..., x_k)^T$ – вектор независимых переменных; y –результирующая величина (отклик); E – оператор усреднения; $\theta = (\theta_1, ..., \theta_m)^T$ – параметры модели; $f(\underline{x})$ – вектор действительных функций. Модель наблюдения за объектом представляет собой уравнение $y_j = \theta^T f(\underline{x}_j) + e_j$, $j = 1, ..., n$, где e – неизвестная случайная ошибка.

Целью анализа экспериментальных данных будем считать определение вида модели $\eta(\underline{x}, \theta)$, достаточно хорошо "аппроксимирующей" наблюдаемый отклик y. При заданных функциях $f(x)$ эта задача сводится к определению числовых значений (оценок) параметров $\theta = (\theta_1, ..., \theta_m)^T$. Практическую ценность будут иметь оценки, обладающие рядом свойств, среди которых: несмещенность, состоятельность и эффективность. В классе линейных оценок, вычисляемых как линейное преобразование вектора y, такие оценки получили название наилучших

линейных оценок (НЛО). Они совпадают с оценками по методу наименьших квадратов (МНК-оценками):

$$\hat{\theta} = \operatorname{Arg}\min_{\theta} SS(\theta) = \operatorname{Arg}\min_{\theta}(y - \theta^T f(x))^T (y - \theta^T f(x)).$$

В классическом регрессионном анализе делают следующие основные предположения:

1) величины e_j, $j = \overline{1,n}$ – случайные величины, имеющие нулевое математическое ожидание. Они некоррелированны и имеют одинаковые дисперсии, т.е. $cov(e_i, e_j) = 0$, $\sigma^2(e_i) = \sigma^2$, $j = \overline{1,n}$;

2) матрица $X = \begin{bmatrix} f_1(\underline{x}_1) & \cdots & f_m(\underline{x}_1) \\ \vdots & & \vdots \\ f_1(\underline{x}_n) & \cdots & f_m(\underline{x}_n) \end{bmatrix}$ неслучайна и имеет полный столбцовый ранг, равный числу параметров в модели;

3) на значения вычисляемых параметров $\theta = (\theta_1, ..., \theta_m)^T$ не накладывается никаких ограничений.

При выполнении предположений 1–4 МНК-оценки $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_m)^T$ параметров $\theta = (\theta_1, ..., \theta_m)^T$ будут вычисляться как

$$\hat{\theta} = (X^T X)^{-1} X^T y.$$

Несмещенная оценка $\hat{\sigma}^2$ неизвестной дисперсии наблюдения равна $\hat{\sigma}^2 = \hat{e}^T \hat{e} / (n - m)$, где вектор остатков

$$\hat{e} = y - \hat{y} = y - X\hat{\theta} = (I_n - X(X^T X)^{-1} X^T) y = (I_n - P) y.$$

Вычислениями $\hat{\theta}$ и $\hat{\sigma}^2$ обычно не заканчивается решение задачи по определению модели объекта. Как минимум еще нужно получить подтверждение, что выбранная априори структура модели (вид функции $f(x)$) действительно дает хорошую аппроксимацию отклика y. Для этой цели служит так называемая процедура проверки гипотезы об адекватности модели. Гипотеза об адекватности моде-

ли имеет вид: $H_0 : E\{\hat{\sigma}^2\} = E\{\hat{\sigma}_E^2\}$, где $\hat{\sigma}^2$ – оценка дисперсии, полученная на основе остаточной суммы квадратов, $\hat{\sigma}_E^2$ – оценка дисперсии, полученная без использования модели, например, из параллельных опытов. Гипотеза H_0 не отвергается, если $F = \hat{\sigma}^2 / \hat{\sigma}_E^2 \le F_T = F_{\alpha,n-m,f_E}$, где $n–m$ – число степеней свободы для $\hat{\sigma}^2$; f_E – число степеней свободы для $\hat{\sigma}_E^2$, F_T – табличное значение квантили F-распределения. Уровень значимости α обычно задается в границе от 0,01 до 0,1, но чаще всего используется $\alpha = 0,05$. Если $F > F_T$, то полученная модель признается неадекватной. В этом случае необходимо изменить структуру модели и, возможно, заново собрать данные.

2. Перечень вопросов к разработке

1. Спроектировать и сформировать программные модули по вычислению МНК-оценок параметров для заданной параметрической модели объекта. Предусмотреть достаточно простой способ настройки программы на необходимый вид (структуру) модели.

2. Пользуясь экспериментальными данными, полученными в лабораторной работе № 1, произвести оценку параметров модели объекта.

3. Произвести проверку адекватности полученной модели. В качестве $\hat{\sigma}_E^2$ можно взять величину дисперсии σ^2, которая использовалась при зашумлении отклика в лабораторной работе № 1. Число степеней свободы $f_E = \infty$.

4. Включить в отчет постановочную часть в табличной форме выборку данных (x, y) и в дополнения к ним значения $u, \hat{y}, y - \hat{y}$, а также значения $\theta, \hat{\theta}, \sigma_E^2, \hat{\sigma}^2, F, F_T$, текст программы, принятые решения по проверке адекватности модели.

3. Контрольные вопросы

1. Понятие наилучших линейных оценок.
2. Теорема Гаусса-Маркова. Доказательство.
2. Оценки параметров модели по методу наименьших квадратов.
3. Несмещенное оценивание неизвестной дисперсии σ^2.
4. История происхождения термина "Регрессионный анализ".
5. Геометрическая интерпретация МНК.
6. Покомпонентное оценивание параметров модели по МНК.
7. Оценивание параметров модели по методу максимального правдоподобия (ММП).

ЛАБОРАТОРНАЯ РАБОТА № 3

ИНТЕРВАЛЬНОЕ ОЦЕНИВАНИЕ, ПРОВЕРКА ГИПОТЕЗ И ПРОГНОЗИРОВАНИЕ

1. Методические указания

Пусть модель наблюдения за объектом представляет собой уравнение

$$y_j = \theta^T f(\underline{x}_j) + e_j,\ \ j = 1,...,n,$$

где e – неизвестная случайная ошибка. Будем предполагать, что $e \sim \mathrm{N}(0, \sigma^2 I)$. По полученным экспериментальным данным точечную оценку параметров будем находить по методу наименьших квадратов. Помимо точечных оценок параметров значительный интерес представляют их интервальные оценки. Обозначим через $\xi = A\theta$ вектор произвольных параметрических функций размерности q, а через

$\hat{\xi} = A\hat{\theta}$ их наилучшую линейную оценку. Если принять, что $\hat{\xi} = A\hat{\theta}$ мы знаем, а вектор ξ нам неизвестен, то поверхность и внутренность q-мерного эллипсоида с центром в точке $\hat{\xi} = A\hat{\theta}$, накрывающего с вероятностью $p_0 = 1 - \alpha$ вектор истинных параметрических функций $\xi = A\theta$, определяется неравенством

$$(\xi - \hat{\xi})^T [A(X^T X)^{-1} A^T]^{-1} (\xi - \hat{\xi}) \le \hat{\sigma}^2 q F_{\alpha, f_H, f_R},$$

где $f_H = q$, $f_R = n - m$, $m = \dim(\theta)$, $q = rg(A)$, матрица A имеет полный строчный ранг.

Совместное доверительное оценивание вектора параметров. Рассмотрим частный случай $A = I_m$, тогда предыдущее неравенство преобразуется к виду

$$(\theta - \hat{\theta})^T X^T X (\theta - \hat{\theta}) \le \hat{\sigma}^2 m F_{\alpha, m, n-m}$$

и определяет собой поверхность и внутренность доверительного эллипсоида для всех параметров с центром в точке $\hat{\theta}$, который с вероятностью $p_0 = 1 - \alpha$ накрывает вектор истинных значений всех параметров.

Доверительное оценивание для отдельного параметра. Доверительный интервал для отдельного параметра также можно получить исходя из общего выражения $\xi = A\theta$, выбирая в качестве A строку из нулей и одной единицей на j-й позиции:

$$(\theta_j - \hat{\theta}_j)^T \frac{1}{d_{jj}} (\theta_j - \hat{\theta}_j) \le \hat{\sigma}^2 F_{\alpha, 1, n-m},\ d_{jj} = \left((X^T X)^{-1}\right)_{jj}.$$

Учитывая, что $\sqrt{F_{\alpha, 1, f_R}} = t_{\alpha/2, f_R}$, где $t_{\alpha/2, f_R}$ – квантиль распределения Стьюдента, $f_R = n - m$, то доверительный интервал можно записать в виде $\left|\theta_j - \hat{\theta}_j\right| / \sigma(\hat{\theta}_j) \le t_{\alpha/2, f_R}$, или в виде двухстороннего неравенства:

$$\hat{\theta}_j - t_{\alpha/2, f_R} \sigma(\hat{\theta}_j) \le \theta_j \le \hat{\theta}_j + t_{\alpha/2, f_R} \sigma(\hat{\theta}_j),$$

где $\sigma(\hat{\theta}_j) = \sqrt{\hat{\sigma}^2 ((X^T X)^{-1})_{jj}}$.

Доверительное оценивание для математического ожидания. Аналогично можно записать интервальную оценку для $\eta(x,\theta)$ истинного значения математического ожидания функции отклика в точке x. Имеем:

$$\xi = A\theta = f^T(x)\theta = \eta(x,\theta),$$

$$\eta(x,\hat{\theta}) - t_{\alpha/2,f_R}\sigma(\eta(x,\hat{\theta})) \le \eta(x,\theta) \le \eta(x,\hat{\theta}) + t_{\alpha/2,f_R}\sigma(\eta(x,\hat{\theta})),$$

где $\sigma(\eta(x,\hat{\theta})) = \hat{\sigma}\sqrt{f^T(x)(X^TX)^{-1}f(x)}$, $\eta(x,\hat{\theta}) = f^T(x)\hat{\theta}$. Если мы будем предсказывать значение самого отклика, то доверительный интервал для него будет шире, поскольку дисперсия оценки отклика определяется как

$$\sigma^2(\hat{y}(x,\hat{\theta})) = \hat{\sigma}^2(1 + f^T(x)(X^TX)^{-1}f(x)).$$

В качестве $\hat{\sigma}^2$ используется несмещенная оценка $\hat{\sigma}^2 = \hat{e}^T\hat{e}/(n-m)$.

Общая линейная гипотеза. Пусть мы хотим проверить гипотезу H: $A\theta = C$, где A – известная $q \times m$ матрица полного строчного ранга, C – известный $q \times 1$ вектор. Обозначим:

$$RSS = (Y - X\hat{\theta})^T(Y - X\hat{\theta}) = (n-m)\hat{\sigma}^2,\ RSS_H = (Y - X\hat{\theta}_H)^T(Y - X\hat{\theta}_H),$$

где $\hat{\theta}_H$ – МНК оценки параметров с учетом ограничений $A\theta = C$.

Если гипотеза H верна, то статистика

$$F = \frac{(RSS_H - RSS)/q}{RSS/(n-m)} = \frac{(A\theta - C)^T\left[A(X^TX)^{-1}A^T\right]^{-1}(A\theta - C)}{q\hat{\sigma}^2}$$

имеет распределение $F_{q,n-m}$. Гипотеза H принимается, если $F < F_{\alpha,q,n-m}$, где $F_{\alpha,q,n-m}$ – критическая точка. В частном случае гипотеза о незначимости отдельного параметра имеет вид H: $\theta_j = 0$ или H: $a^T\theta = 0$, где a^T – вектор-строка, в которой на j-м месте стоит 1, на остальных местах – нули. Обозначим $d_{jj} = \left((X^TX)^{-1}\right)_{jj}$ – j-й диагональный элемент матрицы $(X^TX)^{-1}$. Тогда стати-

стика $F = \dfrac{\left(a^T\hat{\theta}\right)^T\left(a^T\hat{\theta}\right)}{\hat{\sigma}^2 d_{jj}} = \dfrac{\left(\hat{\theta}_j\right)^2}{\hat{\sigma}^2 d_{jj}}$ имеет при справедливой гипотезе H распределение $F_{1,n-m}$.

Проверка значимости уравнения регрессии. Пусть задана линейная модель:

$$y_i = \theta_0 + \theta_1 x_{1,i} + \ldots + \theta_{m-1} x_{m-1,i} + \varepsilon_i,\ i = \overline{1,n}$$

и требуется установить – является ли регрессия с заданными регрессорами значимой, т.е. дает лучшее описание, чем модель среднего. В этом случае гипотеза имеет вид H: $\theta_1 = \theta_2 = \ = \ldots = \theta_{m-1} = 0$. Гипотеза Н имеет вид: $A\theta = 0$, где $A = [0 \mid I_{m-1}]$ – $(m-1)\times$ m матрица ранга $m-1$. Применима общая теория с $q = m$-1,

$RSS = (y - X\hat{\theta})^T (y - X\hat{\theta})$ и $RSS_H = \sum_{i=1}^{n} (y_i - \bar{y})^2$.

2. Перечень вопросов к разработке

1. Изменить модель регрессии, добавив в неё дополнительный регрессор, ранее не вошедший в состав модели, порождающей данные. Не генерируя новых данных, найти точечные оценки всех параметров расширенной модели. В дальнейшем при рассмотрении этой расширенной модели анализе должно быть показано, что параметр при дополнительном регрессоре незначим.
2. Построить доверительные интервалы для каждого параметра модели регрессии.
3. Проверить гипотезу о незначимости каждого параметра модели.
4. Проверить гипотезу о незначимости самой регрессии.
5. Рассчитать прогнозные значения для математического ожидания функции отклика $\eta(x,\hat{\theta}) = f^T(x)\hat{\theta}$ для всего интервала действия одного из факторов, зафиксировав значения других факторов на границе или в центре области их определения.

6. По полученным в п. 5 прогнозным значениям построить графики прогнозных значений и доверительной полосы для математического ожидания функции отклика и для самого отклика.

7. Заново смоделировать исходные данные (см. лаб. работу № 1), увеличив мощность случайной помехи до 50…70 % от мощности полезного сигнала и провести оценку параметров. Повторить пункты 3, 4 с новыми данными.

3. Контрольные вопросы

1. Совместное доверительное оценивание параметров.
2. Интервальные оценки для отдельных параметров модели.
3. Оценка параметров при наличии линейных ограничений.
4. Проверка общей линейной гипотезы.
5. Проверка значимости отдельных параметров регрессии.
6. Проверка значимости регрессии.
7. Доверительные интервалы для математического ожидания функции отклика.
8. Проверка структурных изменений.

ЛАБОРАТОРНАЯ РАБОТА № 4

ОЦЕНИВАНИЕ ЛИНЕЙНЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ В УСЛОВИЯХ ГЕТЕРОСКЕДАСТИЧНОСТИ ВОЗМУЩЕНИЙ

1. Методические указания

Относительно ошибки наблюдения примем общее предположение

$$E(ee^T) = \sigma^2\Omega = V,$$

где $tr\Omega = n$, Ω – симметричная положительно определенная матрица. Эффективный оцениватель для θ в этом случае состоит в использовании обобщенного метода наименьших квадратов:

$$\hat{\theta}_{\text{ОМНК}} = (X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}y = (X^TV^{-1}X)^{-1}X^TV^{-1}y.$$

Специальный случай несферических возмущений, который называется гетероскедастичностью, состоит в предположении, что $Var(e_i) \neq Var(e_j),\ i \neq j$ при сохранении предположения о независимости возмущений:

$$Var(\underline{e}) = \sigma^2\Omega = diag(\sigma_1^2,...,\sigma_n^2) \equiv V.$$

Очень часто, например дисперсия возмущения является возрастающей функцией от какого-либо фактора, группы факторов или самого отклика.

Для получения эффективной оценки $\hat{\theta}$ можно воспользоваться оцениванием по обобщенному методу наименьших квадратов (ОМНК) с V^{-1}, которая играет роль весовой матрицы, взвешивающей каждое наблюдение с весом σ_j^{-2}. Поэтому наблюдение с большей дисперсией учитывается в меньшей степени.

Для проверки данных на гетероскедастичность можно использовать следующие два теста.

1. Тест Бреуша-Пагана:

$$y_t = f^T(x_t)\theta + \varepsilon_t, \quad t = 1,2,...,n,$$

$$E\varepsilon_t = 0, \quad \forall t, \quad \sigma_t^2 = E\varepsilon_t^2 = h(z_t^T\alpha),$$

где h – определенная функция, принимающая только положительные значения, $z_t^T = (1,\ z_{1t},\ ...,\ z_{pt})$ – вектор известных переменных, $\alpha^T = (\alpha_0,\ \alpha_1,\ ...,\ \alpha_p)$ – вектор неизвестных параметров. Состав регрессоров $(z_1,\ ...,\ z_p)$ зависит от предположений относительно того как изменяется дисперсия наблюдений. Гипотеза об отсутствии гетероскедастичности (т.е. о гомоскедастичности) имеет вид $H_0 : \alpha_1 = \alpha_2 = ... = \alpha_p = 0$.

Последовательность:

оценивание исходного уравнения по МНК, с получением остатков $e_t = y_t - f(x_t)^T \hat{\theta}$ и оценивание дисперсии $\tilde{\sigma}^2 = \sum \frac{e_t^2}{n}$;

построение регрессии с откликом $c_t = \frac{e_t^2}{\tilde{\sigma}^2}$ по регрессорам $(1, z_1, ..., z_p)$ и вычисление ESS для нее; статистика $ESS/2 \sim \chi^2_{(p)}$, где ESS – объясненная сумма квадратов, связанная с регрессорами $\mathbf{z}_t^T = (1, z_{1t}, ..., z_{pt})$, но без свободного члена. Оценивая параметры α, найдем предсказанные значения нормированных квадратов остатков $\hat{c}_t = \hat{\alpha}^T \mathbf{z}_t$. Объясненная сумма квадратов по аналогии с проверкой гипотезы о незначимости регрессии рассчитывается как $ESS = \sum (\hat{c}_t - \bar{c})^2$, где $\bar{c}$ - среднее значение отклика c.

2. Тест Голдфельда-Квандтона. Предположение: источник нарушения гомоскедастичности взят в форме $E(\varepsilon_i^2) = \sigma^2 x_{ji}^2$, т.е. дисперсия возмущения возрастает пропорционально квадрату одной из объясняющей переменной.

Последовательность:

упорядочить последовательность наблюдений в соответствии с величиной x_j^2. При возможных других источниках нарушения гомоскедастичности упорядочивание наблюдений производится в соответствии с принятыми предположениями;

опустить n_c наблюдений, оказавшихся в середине упорядоченной выборки;

оценить отдельно две регрессии: по первым $(n - n_c)/2$ наблюдениям и последним $(n - n_c)/2$ наблюдениям;

при выполнении гипотезы о гомоскедастичности статистика $\frac{RSS_2}{RSS_1} \sim F_{\alpha;(n-n_c-2k)/2;(n-n_c-2k)/2}$, где k – число параметров в регрессии, RSS_1, RSS_2 – остаточные суммы квадратов для регрессий, построенных соответственно на первой и последних частях выборки. Для выбора числа исключаемых наблюдений можно воспользоваться эмпирическим правилом $n_c = n/3$.

Доступный обобщенный МНК. В случае обнаружения эффекта гетероскедастичности можно провести коррекцию оценок параметров модели с помощью взвешенного МНК с весами $\hat{\sigma}_t^{-2} = h^{-1}(z_t^T \hat{\alpha})$, где $\hat{\sigma}_t^2$ выступает в качестве оценки неизвестной дисперсии σ_t^2. Может оказаться, что для некоторых t $h^{-1}(z_t^T \hat{\alpha}) < 0$. Если число таких наблюдений невелико, то их можно просто выбросить, в противном случае использовать представление $\sigma_t^2 = exp(h(z_t^T \alpha))$.

2. Перечень вопросов к разработке

1. Провести моделирование регрессионного процесса с гетероскедастичным возмущением.
2. Полученные данные проверить по тестам на наличие гетероскедастичности.
3. Провести оценивание параметров регрессионной модели по доступному обобщенному МНК и по обыкновенному МНК.
4. Сравнить эффективность оценок в этих двух случаях по квадрату их расстояния до известных истинных значений параметров.

3. Варианты заданий

1. Дисперсия возмущений – возрастающая функция от квадрата одного из факторов.
2. Дисперсия возмущений – возрастающая функция от двух факторов (взвешенная сумма квадратов факторов).
3. Дисперсия возмущений – возрастающая функция от абсолютной величины отклика.
4. Дисперсия возмущений – возрастающая функция от расстояния точки до центра эксперимента.

5. Дисперсия возмущений – убывающая функция от модуля взаимодействия первого и второго факторов.

6. Дисперсия возмущений –экспонента от взвешенной суммы квадратов факторов.

7. Дисперсия возмущений – экспонента от абсолютной величины отклика.

8. Дисперсия возмущений – экспонента от расстояния точки до центра эксперимента.

Замечание. Рекомендуется значительно увеличить число наблюдений и моделировать возмущения с ярко выраженной гетероскедастичностью.

4. Контрольные вопросы

1. Обобщенный МНК. Доступный обобщенный МНК.
2. Проверка данных на гетероскедастичность. Тесты.
3. Оценивание в условиях автокорреляции возмущений. Модель помехи - авторегрессия первого порядка. Параметр автокорреляции известен.
4. Тест Дарбина-Уотсона.
5. Оценивание параметра автокорреляции.

ЛАБОРАТОРНАЯ РАБОТА № 5

ОЦЕНИВАНИЕ ПАРАМЕТРОВ МОДЕЛИ РЕГРЕССИИ В УСЛОВИЯХ МУЛЬТИКОЛЛИНЕАРНОСТИ

1. Методические указания

Явление мультиколлинеарности возникает, если между объясняющими переменными в интервале их изменения в эксперименте существуют почти точные линейные зависимости. Мультиколлинеарность в основном появляется в задачах

пассивного эксперимента, когда исследователь, собирая данные, не может влиять на значения объясняющих переменных. Мультиколлинеарность опасна тем, что оценки параметров становятся малоэффективными, значительно возрастает их евклидова норма – фактически для них не прослеживаются свойства несмещенности и состоятельности.

В качестве мер измерения эффекта мультиколлинеарности можно рассматривать следующие.

1. Определитель информационной матрицы. $\left|X^T X\right| = \prod_{i=1}^{m} \lambda_i$.

2. Минимальное собственное число матрицы $X^T X$. Очевидно, что чем меньше $\lambda_1 = \lambda\min(X^T X)$, тем сильнее мультиколлинеарность. Предположим, что между $x_1,...,x_m$ имеется приближенная линейная зависимость, т.е.

$$v_1\underline{x}_1 + v_2\underline{x}_2 + ... + v_m\underline{x}_m \approx 0.$$

Вектор $\underline{V}^T = (v_1,...,v_m)$, соответствующий минимальному собственному числу λ_1, даёт тот набор коэффициентов, который дает максимальное приближение к нулю линейной комбинации векторов $\underline{x}_1,...,\underline{x}_m$. Помимо того, что λ_1 отражает линейную зависимость $\underline{x}_1,...,\underline{x}_m$, данная мера несет также на себе эффект выбора масштаба измерения этих переменных.

3. Мера обусловленности матрицы по Нейману-Голдстейну. Дж. Нейман и Х. Голдстейн, исследуя методы обращения матриц, заметили, что удобной характеристикой близости матрицы к вырожденной является отношение $\lambda\max(X^T X) / \lambda\min(X^T X)$. Если линейной зависимости нет, отношение $\lambda_{\max} / \lambda_{\min}$ отражает разницу в масштабах.

4. Максимальная парная сопряженность. Большую пользу при анализе регрессии приносит рассмотрение матрицы сопряженности $R = (r_{ij})$, где $r_{ij} = \cos(\underline{x}_i, \underline{x}_j)$. В качестве показателя мультиколлинеарности может выступать величина

$\max\limits_{i,j}\left|r_{ij}\right|, \quad i \neq j$. Однако она выражает только парную коллинеарность. Три вектора могут быть коллинеарны, а попарно нет. Максимальный коэффициент сопряженности не несет на себе эффекта масштаба.

5. Максимальная сопряженность. Зафиксируем независимую переменную x_i и найдем косинус угла R_i, который составляет эта переменная, т.е. вектор $\underline{x}_i \in R^n$ с подпространством S_i, натянутым на остальное множество независимых переменных $x_1,...,\underline{x}_{i-1},\underline{x}_{i+1},...,\underline{x}_m$. В качестве меры мультиколлинеарности можно взять $\max\limits_i\left|R_i\right|$. Вычислить R_i можно следующим способом $R_t^2 = 1 - \dfrac{1}{R_{ii}^{-1}}, \quad i = \overline{1,m},$ где R_{ii}^{-1} – (i, i) элемент матрицы, обратной к сопряженной R.

Один из способов оценивания параметров в условиях мультиколлинеарности состоит в управлении масштабом получаемых оценок. При этом мы вынуждены работать в классе смещенных оценок. Однако это смещение значительно меньше, чем у обычных МНК оценок.

С целью управления масштабом оценок введем в рассмотрение функцию стоимости

$$C = \sum_{i=1}^{n}(y_i - f(X_i)^T\theta)^2 + \sum_{j=1}^{m}\lambda_j\theta_j^2,$$

где второе слагаемое рассматривается как штраф при условии, что $\lambda_j \geq 0$. Минимизируя C, получим $\hat{\theta} = (X^TX + \Lambda)^{-1}X^Ty$, которые известны как ридж-оценки. Часто матрицу Λ задают диагональной в виде $\Lambda_{ii} = \lambda \cdot (X^TX)_{ii}$, $\lambda \geq 0$.

На практике возникает задача выбора оптимального значения параметра регуляризации λ. Здесь возможен некоторый компромисс между неизбежным увеличением остаточной суммы квадратов и желаемым уменьшением евклидовой нормы оценок параметров. Примерно оценить ожидаемую норму оценок параметров

можно по значениям отклика в нескольких характерных точках типа центра эксперимента или граничных точках.

2. Перечень вопросов к разработке

1. В соответствии с вариантом задания сгенерировать экспериментальные данные, в которых в явном виде присутствует эффект мультиколлинеарности.
2. Рассчитать ряд показателей, характеризующих эффект мультиколлинеарности. Определить факторы, ответственные за возникновение эффекта мультиколлинеарности.
3. Построить ридж-оценки параметров при различных значениях параметра регуляризации. Выбрать оптимальное значение параметра регуляризации. Построить графики изменения квадрата евклидовой нормы оценок параметров и остаточной суммы квадратов от параметра регуляризации.
4. Провести оценивание модели регрессии по методу главных компонентов. Перейти к описанию в исходном пространстве факторов. Сравнить решение с ридж-оцениванием по смещению оценок и точности предсказания отклика.

3. Варианты заданий

1. Регрессия на 6 факторах. Эффект мультиколлинеарности создают 3 фактора. Имеется разброс в масштабах факторов.
2. Регрессия на 7 факторах. Эффект мультиколлинеарности создают 4 фактора. Разброса в масштабах факторов нет.
3. Регрессия на 8 факторах. Эффект мультиколлинеарности от 4 факторов. Имеется разброс в масштабах факторов.
4. Регрессия на 8 факторах. Эффект мультиколлинеарности создают две пары факторов. Разброса в масштабах факторов нет.
5. Регрессия на 8 факторах. Эффект мультиколлинеарности создают две пары факторов. Имеется разброс в масштабах факторов.

6. Регрессия на 6 факторах. Эффект мультиколлинеарности создают 2 фактора. Имеется разброс в масштабах факторов.

7. Регрессия на 7 факторах. Эффект мультиколлинеарности создают две пары факторов. Разброса в масштабах факторов нет;

8. Регрессия на 9 факторах. Эффект мультиколлинеарности создают две тройки факторов. Разброса в масштабах факторов нет.

Замечание. В моделях порождающих данные целесообразно включить свободный член.

4. Контрольные вопросы

1. Меры измерения мультиколлинеарности.
2. Как освободить меры измерения мультиколлинеарности от эффекта масштаба.
3. Как определить какие именно факторы ответственны за эффект коллинеарности.
4. Ридж-оценки. Вывести формулу.
5. Свойства ридж-оценок. Какое свойство ридж-оценок можно считать главным.
6. Редуцированные оценки.
7. Метод главных компонент.

ЛАБОРАТОРНАЯ РАБОТА № 6

ВЫБОР НАИЛУЧШЕГО УРАВНЕНИЯ РЕГРЕССИИ. АЛГОРИТМЫ ВКЛЮЧЕНИЯ И ИСКЛЮЧЕНИЯ

1. Методические указания

Пусть значение исследуемого отклика описывается моделью

$$y(\underline{x}) = f_{\text{ист}}(\underline{x}) + e(\underline{x}).$$

Поскольку мы обычно не располагаем информацией о виде $f_{\text{ист}}(\underline{x})$, то собственно регрессионный анализ начинается с выбора класса допустимых решений F – класс функций, в рамках которого предполагается вести поиск наиболее подходящей аппроксимации $\hat{f}(\underline{x})$ для $f_{\text{ист}}(\underline{x})$ в соответствии с тем или иным критерием качества модели. Наиболее распространенными в статистической практике являются параметрические регрессионные схемы, когда в качестве допустимых решений выбирается некоторое параметрическое семейство функций $F = \{f(\underline{x},\theta)\}_{\theta\in\Gamma}$. В этом случае дальнейший поиск аппроксимации $\hat{f}(\underline{x})$ сводится к подбору значений параметров $\theta = (\theta_1,...,\theta_m)^T$.

Помощь в выборе F может оказать решение следующих задач:

– анализ априорной информации о содержательной сущности анализируемой зависимости (монотонность, экстремумы, асимптоты, гладкость и т.д.);

– анализ корреляционных полей (x_i, y), (x_i, x_j).

Если предположить, что $f_{\text{ист}}(\underline{x}) \in F$, то поиск наилучшей модели сведется к полному или направленному перебору всех моделей в F с сравнением их по какому-либо критерию качества. Среди них можно выделить следующие.

Статистика Малоуса: $C_p = \dfrac{RSS_p}{\hat{\sigma}^2} + 2p - n,$

где $RSS_p = (y - \hat{y}_p)^T (y - \hat{y}_p)$ – остаточная сумма квадратов, рассчитанная по модели с p регрессорами. Часто за неимением лучшего в качестве оценки неизвестной дисперсии берут $\hat{\sigma}^2 = RSS/(n-m)$, где m – полное число регрессоров (модель из F в этом случае имеет в своей структуре наибольшее число регрессоров).

Коэффициент детерминации: $R_p^2 = \dfrac{\sum(\hat{y}_i - \bar{\hat{y}})^2}{\sum(y_i - \bar{y})^2}$ *(данное выражение справедливо для моделей, содержащих свободный член).*

Среднеквадратичная ошибка предсказания по $\hat{y}_p$ истинного значения отклика y (MSEP-критерий), где в качестве статистики для ее оценки используют величину:

$$E_p = \frac{RSS_p}{n(n-p)}(1+n+\frac{p(n+1)}{n-p-2}).$$

Средняя дисперсия прогноза по всем точкам эксперимента (AEV-критерий):

$$\text{AEV} = \frac{p \cdot RSS_p}{n(n-p)}.$$

Большинство предлагаемых критериев являются функциями остаточной суммы квадратов *RSS*. При малом числе регрессоров *RSS* велика, а с добавлением новых – уменьшается. При числе регрессоров $p = n$ (неважно каких) регрессионное сглаживание превращается в интерполяционную схему с нулевой остаточной суммой квадратов.

Нахождение оптимальной модели $f^*(x)$ сводится к решению задачи структурной минимизации

$$f^*(x) = \text{Arg} \underset{f \in F}{\text{extr}} \Delta_n(f),$$

где $\Delta_n(f)$ – какой-либо критерий качества.

Генерация моделей. Полным решением задачи нахождения $f^*(x)$, очевидно, будет осуществление перебора всех моделей в *F* с запоминанием лучших по различиям критериям. Однако это довольно трудоемкая процедура и при $m > 20$ трудно реализуемая даже на современных компьютерах. На практике часто используют методы направленного перебора: метод включения и метод исключения.

В методе включения начинают с модели, состоящей из одного регрессора, например, свободного члена. Затем один за другим добавляют остальные регрессоры, а порядок их включения определяют по частным коэффициентам корреляции регрессоров с откликом. Регрессор, включенный на данном этапе, должен иметь максимальный частный коэффициент корреляции или частный

F-критерий. Частный F-критерий для вводимого регрессора $f_i(x)$ вычисляется как $F_i = (\nu_2 / \nu_1)(RSS - RSS_i) / RSS_i$, где RSS, RSS_i – остаточные суммы квадратов до и после включения в модель очередного регрессора; $\nu_1 = 1$, $\nu_2 = n - m$.

Метод исключения корректирует структуру модели в обратном порядке. Заранее оценивается самая полная модель. После этого подсчитывают значения частных F -критериев для каждого регрессора при условии, что именно он исключается из модели. В результате из модели исключается регрессор, имеющий наименьшее значение $F_i = (\nu_2 / \nu_1)(RSS_i - RSS) / RSS_i$, где RSS_i– остаточная сумма квадратов для модели без i-го регрессора; RSS –остаточная сумма квадратов для модели со структурой, включающей i-й регрессор; $\nu_2 = 1$, $\nu_1 = n - m$.

В приведенных алгоритмах мы не рекомендуем использовать какие-либо правила останова, а предлагаем доводить выполнение процедур до включения в модель всех регрессоров (в методе включения) либо до модели, содержащей один регрессор – в методе исключения.

2. Перечень вопросов к разработке

1. Модифицировать программу МНК-оценивания (см. лаб. работу № 2) под реализацию алгоритма включения (исключения).

2. Выбрать оптимальную модель для аппроксимации заданных экспериментальных данных.

3. Отчет должен содержать постановочную часть, текст программы, компьютерный листинг (структура текущей генерируемой модели, значения Cp , R_p^2, Ep, AEV для данной модели и графики остатков от предсказанных значений функции отклика) и принятое решение по выбору оптимальной модели. Для выбранной оптимальной модели привести значения y, $\hat{y}$, $y - \hat{y}$, $\hat{\theta}$.

3. Варианты заданий

1. Данные из упражнения 1 [1, с.69]. Алгоритм включения.
2. Данные из упражнения 1 [1, с.69]. Алгоритм исключения.
3. Данные из упражнения 4 [1, с.78]. Алгоритм включения.
4. Данные из упражнения 4 [1, с.78]. Алгоритм исключения.
5. Данные из приложения Б [1, с. 283]. Алгоритм включения.
6. Данные из приложения Б [1, с. 283]. Алгоритм исключения.
7. Данные из упражнения 5 [1, с.90]. Алгоритм включения.
8. Данные из упражнения 5 [1, с.90]. Алгоритм исключения.
9. Данные из п. 7.1 [1, с.104]. Алгоритм включения.
10. Данные из п. 7.1 [1, с.104]. Алгоритм исключения.

4. Контрольные вопросы

1. Алгоритмы включения и исключения поиска оптимальной структуры регрессионной модели.
2. Корреляционные поля и их анализ.
3. Критерии качества регрессионных моделей.
4. Свойства НЛО неизвестных параметров модели в случаях перебора и недобора регрессоров.
5. Анализ остатков.

ЛАБОРАТОРНАЯ РАБОТА № 7

ВЫБОР "НАИЛУЧШЕГО" УРАВНЕНИЯ РЕГРЕССИИ С ИСПОЛЬЗОВАНИЕМ ВНЕШНИХ КРИТЕРИЕВ

1. Методические указания

Для того чтобы решить задачу выбора "наилучшей" модели объекта, необходимо зафиксировать прежде всего критерий оптимальности модели. Критерий назовем внутренним, если его определение основано на использовании той же информации, тех же данных, что и для получения самой модели. К числу таких критериев относятся различные статистики, основанные на остаточной сумме квадратов. Критерий называется внешним, если его определение основано на использовании новой информации, "свежих" точках, не задействованных при синтезе тестируемой модели. Будем считать моделью оптимальной сложности модель, соответствующую минимуму внешнего критерия. При постепенном усложнении модели внешний критерий проходит через минимум, что дает возможность найти единственную для данного критерия модель.

Пусть модель объекта подчиняется следующему уравнению наблюдения

$$Y = \hat{Y} + \varepsilon = \hat{X}\hat{\theta} + \varepsilon, \qquad (1)$$

где $\hat{Y} - (n \times 1) -$ вектор ненаблюдаемого незашумленного выхода объекта, $\hat{X} - (n \times m) -$ расширенная матрица плана, соответствующая истинному набору регрессоров $\hat{x}_1, ..., \hat{x}_m$, $\varepsilon - (n \times 1) -$ вектор ненаблюдаемых случайных ошибок измерения, относительно которых выполнены предположения

$$E(\varepsilon) = 0_n, \;\; E(\varepsilon\varepsilon^T) = \sigma^2 I_n,$$

где 0_n – вектор, состоящий из нулей, σ^2 – неизвестная дисперсия наблюдения. Набор регрессоров $\hat{x}_1, ..., \hat{x}_m$ образует множество $\hat{X}$, о котором известно, что $\hat{X} \subset \Re$, где $\Re -$ некоторое расширенное множество регрессоров. Пусть в

результате наблюдения объекта получена $Z-(n\times p)-$ расширенная матрица плана из *n* наблюдений над *p* регрессорами из $\Re$ и требуется определить множество $\widehat{X}$ и получить оценку параметров $\widehat{\theta}$. Для поиска наилучшей аппроксимации для (1) воспользуемся каким-либо переборным алгоритмом. Пусть $X-\left(n\times s\right)-$ расширенная матрица для текущей модели из *s* регрессоров, образующих множество $L\subset\Re$. Регрессия отклика y по L будет определяться по уравнению наблюдения

$$y = X\beta + e. \tag{2}$$

J – оптимальная модель определяется решением задачи

$$f^* = \operatorname{Arg}\min_{f\in\Omega_f} J(f), \tag{3}$$

где Ω_f – множество всех возможных моделей, формируемых на основе наблюдаемой *Z*. Теоретический (идеальный) $J(f)$ определяет собой среднеквадратичную ошибку предсказания истинного отклика либо на всей выборке, либо на прогнозной части *B*:

$$J(f) = \frac{1}{n} E\left\|\hat{y} - X\hat{\theta}\right\|^2,$$

$$J_B(f) = \frac{1}{n_B} E\left\|\hat{y}_B - X_B\hat{\theta}_A\right\|^2.$$

При решении задачи (3) минимуму $J(f)$ соответствует оптимальная сглаживающая модель, а минимуму $J_B(f)-$ оптимальная прогнозирующая. В условиях шума с нулевой дисперсией минимумы этих критериев приходятся на модель сложности $s^0 = m$. При дисперсии наблюдения $\sigma^2 > 0$ функции $J(f)$ и $J_B(f)$ имеют единственный минимум в точке $s^* \le s^0$. С ростом σ^2 сложность s^* уменьшается – *J*-оптимальной становится все более простая модель. В качестве оценок для идеальных критериев $J(f)$ и $J_B(f)$ могут выступать внешние критерии $\Delta^2(B)$, d^2, S^2, $n^2_{\text{см}}$, $\Delta^2_{\text{СК}}$. Исследования показывают, что внешние критерии

$\Delta^2(B)$, d^2, S^2, Δ^2_{CK} несмещенно оценивают *J*-оптимальную модель. В то же время критерий в виде остаточной дисперсии RSS как и его скорректированная величина $RSS/(n-s)$ не обладают необходимыми свойствами помехоустойчивости, поскольку минимум этих критериев при $\sigma^2 > 0$ не находится в области $s \le s^0$.

Приведем упомянутые здесь основные внешние критерии.

Критерий регулярности

$$\Delta^2(B) = \Delta^2(B/A) = \left\| y_B - X_B \hat{\theta}_A \right\|^2,$$

где запись $\Delta^2(B/A)$ означает “ошибка” на B модели, коэффициенты, которой получены на A.

Критерий симметричной регулярности

$$d^2 = \Delta^2(B/A) + \Delta^2(A/B) = \left\| y_B - X_B \hat{\theta}_A \right\|^2 + \left\| y_A - X_A \hat{\theta}_B \right\|^2;$$

критерий стабильности:

$$S^2 = \Delta^2(A/A) + \Delta^2(B/A) = \left\| y_W - X_W \hat{\theta}_A \right\|^2 + \left\| y_W - X_W \hat{\theta}_B \right\|^2;$$

критерий вариативности:

$$V^2 = (X_W \hat{\theta}_A - X_W \hat{\theta}_W)^T (X_W \hat{\theta}_W - X_W \hat{\theta}_B).$$

Очевидный подход к получению “наилучшей” модели состоит в вычислении всех возможных уравнений регрессии, которые можно получить по 0, 1, 2,...,*m* регрессоров из совокупности $x_1, x_2, ..., x_m$. Поскольку для каждого регрессора мы имеем только две возможности: он либо включается в уравнение, либо не включается в него, то всего имеем 2^m возможных уравнений регрессии. Если учесть, что уравнение без регрессоров нас не интересует, то всего может быть $2^m - 1$ уравнений. Если *m* велико, мы сталкиваемся с необходимостью сравнения очень большого количества уравнений. Поэтому нам, прежде всего, необходим эффективный алгоритм для получения всех возможных регрессий. Переход от одной модели к

другой осуществляется путем так называемого выметания информационной матрицы по диагональному элементу, соответствующему включаемому или исключаемому регрессору. В силу обратимости и коммутативности оператора выметания выполнение его по некоторому ведущему элементу матрицы приводит либо к включению в модель соответствующего регрессора, либо к исключению этого регрессора из модели, если он в ней присутствовал. Используя оператор выметания, можно построить эффективную процедуру построения всех возможных моделей. Главное в ней – это последовательность проводимых выметаний. Для $m = 2$ она имеет вид "121", а для $m = 3$ "1213121". Дальнейшие обобщения очевидны.

Рассмотренный алгоритм построения моделей получил название комбинаторного. Применение его в чистом виде возможно для m порядка 15-20 из-за большого времени вычислений. Модификация алгоритма может заключаться в отказе от просмотра моделей любой сложности. Просмотр моделей может осуществляться в диапазоне сложности $1 \le p_{\min} \le p \le p_{\max}$, где p – число регрессоров, включаемых в модель.

Для случая относительно большого числа m регрессоров-претендентов может использоваться многорядный селекционный алгоритм. Работа алгоритма состоит в следующем. На первом шаге выбирается некоторое число лучших моделей, состоящих из одного регрессора. На втором шаге к этим моделям добавляются различные регрессоры по одному и отбираются лучшие модели с двумя регрессорами и т.д. до построения полной модели.

Данная лабораторная работа выполняется с использованием программной системы ОДА 1.0 (Объективный анализ данных).

Назначение. Программная система ОДА предназначена для решения задачи поиска модели оптимальной сложности в классе линейных по параметрам моделей. В качестве входных действующих факторов могут выбираться факторы количественные, качественные или смешанной природы.

Структура. Фрейм-задача "Структурная оптимизация линейных моделей" состоит из 4 подфреймов: **Модель, Данные, Вычисления, Просмотр**.

Фрейм **Модель**. Назначение и способ означивания отдельных слотов.

– *Базовая модель.* В этом слоте пользователь может предварительно выбрать в качестве исходной одну из 4 видов моделей различной сложности. По умолчанию всегда предлагается *"Линейная"* модель.

– *Число факторов.* Слот задает число факторов в базовой модели. Число факторов может быть выбрано в диапазоне от 1 до 40.

– *Регрессоры*. Слот определяет список регрессоров выбранной базовой модели в зависимости от числа факторов. Список отображается в окне просмотра.

– *Число параметров в модели.* Слот является информационным. Значение его определяется заданной моделью. Пересчитывается при изменениях в слотах *Базовая модель*, *Число факторов*, *Использовать модель*, *Модель пользователя*.

– *Модель пользователя.* Назначение слота – формирование списка регрессоров нетиповой модели. Первоначально список регрессоров *Модель пользователя* пуст. Для его первоначального заполнения можно воспользоваться списком *Регрессоры* соответствующей *Базовой модели*. Для переноса отдельных или всех элементов списка необходимо воспользоваться кнопками ">" или ">>" соответственно. И наоборот – для того чтобы удалить регрессоры из списка *Модель пользователя*, необходимо воспользоваться управляющими кнопками "<", "<<".

– *Преобразование*. Слот определяет список функций, которые можно использовать для осуществления преобразований того или иного регрессора в списке *Модель пользователя*.

– *Протекция регрессорам.* Этот слот определяет список регрессоров, которые должны обязательно присутствовать в результирующей модели оптимальной сложности. По умолчанию данный список пуст.

– *База моделей.* Слот является процедурным. По кнопке *"Сохранить"* сформированная пользователем спецификация модели сохраняется в специальном файле.

Файл выбирается пользователем в стандартном диалоговом окне. По кнопке *"Выбрать"* можно выбрать одну из ранее сформированных спецификаций модели.

– *Использовать модель.* Важный слот. Выбором одной из двух опций слота пользователь окончательно фиксирует ситуацию – какую именно модель он будет в дальнейшем использовать: *Базовую* или *Пользователя*.

Фрейм **Данные**. Фрейм *Данные* можно использовать автономно для подготовки файлов данных эксперимента. Для этого используется электронная таблица, расположенная в правой части экранной формы. Рассмотрим назначение отдельных слотов.

– *Число факторов.* Определяет число факторов в таблице данных.

– *Число наблюдений.* Определяет число наблюдений в таблице данных. При изменении значений слотов *Число факторов*, *Число наблюдений* соответствующим образом меняется размерность таблицы данных.

База файлов данных. Структурный слот, определяющий работу с файлами данных. По кнопке "Выбрать" осуществляется выбор и загрузка файла данных из внешнего файла. Поиск файлов осуществляется по маске "EXS*.txt" или "EXS*.prn". в последнем случае считается, что файлы подготовлены в EXCEL 7.0.

– *Разбиение выборки.* Слот управляет процедурой разбиения выборки данных на части A и B. По определению часть A считается обучающей выборкой, а часть B – проверочной. Внутренний слот *Часть A* определяет размерность обучающей выборки. Означиванием этого слота управляет пользователь. Слот *Часть B* – информационный – для отображения оставшейся части выборки. Внутренний слот *Оптимизация разбиения выборки* предоставляет дополнительные возможности по разбиению выборки на части A и B. Если эта опция отключена, то по умолчанию в часть A попадают точки с первыми номерами $1,\ldots,n_A$. Если эта опция включена, то производится (на этапе выполнения) D – оптимальное разбиение выборки: часть A составят точки D – оптимального плана, построенного из точек таблицы данных.

Фрейм **Вычисления**. Основное назначение этого фрейма – осуществить этап Вычислений решения задачи структурной оптимизации. Процедурно это осуще-

ствляется нажатием кнопки "Вычисления". Кнопка "Вычисления" доступна, если выполнены минимально необходимые требования по спецификации задачи. Рассмотрим другие слоты, которые управляют процессом вычислений.

– *Критерий основной.* Этот слот предлагает альтернативный выбор критерия селекции моделей, по которому будет осуществляться выбор лучших моделей.

– *Критерий дополнительный.* Слот предлагает альтернативный выбор критерия селекции моделей, который будет использоваться на заключительном этапе выбора модели.

– *Алгоритмы*. Слот предлагает альтернативный выбор алгоритма, который будет использоваться при поиске модели оптимальной сложности.

– *Коридор сложности моделей.* Этот слот требует означивания двух его составных частей: *Нижний уровень* и *Верхний уровень*. Значения этих слотов определяют коридор сложности перебираемых моделей.

– *Буфер моделей.* Данный слот состоит из двух частей: *Общий буфер* и *Буфер одного ряда*. *Общий буфер* определяет число лучших моделей, отбираемых по основному критерию. **Необходимо выбирать его значение большим 1**. *Буфер одного ряда* определяет число лучших моделей, предаваемых из ряда в ряд в многорядном алгоритме селекции. По умолчанию оба слота имеют значение 1. В многорядном алгоритме для более глубокой структурной оптимизации рекомендуется выбирать значение слота *Буфер одного ряда* больше, чем 1.

– *Шкалирование факторов.* Слот может определять список факторов, которые для повышения вычислительной устойчивости алгоритмов приводятся к интервалу [–1, 1].

– *Вычисления*. По данному слоту кнопка "Вычисления" может быть в одном из двух состояний: доступна и недоступна. Когда кнопка недоступна, она высвечивает свое название значительно более тусклым цветом. Если кнопка "Вычисления" недоступна, то возможные причины этого можно узнать по *Справке*, которая вызывается нажатием соответствующей кнопки.

– *Файл результатов*. Слот требует обязательного своего означивания. При нажатии кнопки "Назначить" открывается стандартное диалоговое окно выбора

файла. Имя выбранного файла высвечивается в информационной строке. Файлы выбираются по маске Res*.txt.

Фрейм **Просмотр**. Представляет возможность просматривать содержимое файлов результатов в скроллируемом окне.

2. Перечень вопросов к разработке

1. Изучить функциональные возможности программного комплекса ОДА, пользуясь данными методическими рекомендациями и встроенной контекстной помощью.

2. Подготовить исходные данные из лабораторных работ № 2, № 6.

3. Решить задачу выбора "наилучшей" модели регрессии для задач из лабораторных работ № 2 и № 6.

4. Оформить отчеты по лабораторной работе, аналогичные предыдущим.

3. Контрольные вопросы

1. Внешние критерии селекции моделей.

2. Внешние критерии селекции моделей и проверка гипотез.

3. Помехоустойчивость внешних критериев.

4. Эффективные комбинаторные и многорядные алгоритмы перебора моделей регрессии.

Расчетно-графическое задание по дисциплине "Статистические методы анализа данных"

ЗАДАЧА: Провести полный цикл исследований, связанных с построением регрессионной зависимости по имеющимся экспериментальным данным.

В перечень исследований как обязательные части должны входить:

1. проверка данных на мультиколлинеарность;
2. проверка данных на гетероскедастичность (предположительно, что чем дальше от центра эксперимента проведено наблюдение, то возможно дисперсия его больше);
3. проверка данных на автокорреляцию (упорядоченность наблюдений по своим номерам считать упорядоченностью по времени);
4. выбор предварительного состава регрессоров с использованием корреляционных полей. В качестве регрессоров-кандидатов предположительно могут выступать: свободный член, сами факторы, их взаимодействия (двух-трех факторов), квадраты факторов;
5. выбор модели оптимальной сложности проводить с использованием критериев Мэллоуса, скорректированного коэффициента детерминации, внешних критериев;
6. дополнительно провести проверку адекватности выбранной модели с использованием повторных наблюдений (последние 6 наблюдений выборки), по которым необходимо будет вычислить оценку дисперсии наблюдений;
7. построение графиков остатков в различных координатах (по номеру наблюдений, по факторам, по отклику);
8. определение, опираясь на построенную модель, точки в факторном пространстве, имеющей максимальное значение математического ожидания отклика. Вычисление для этой точки доверительного интервала. Координаты

такой точки не обязательно должны совпадать с какой-либо точкой из имеющихся в таблице наблюдений.

Дополнительные комментарии. При обнаружении эффектов мультиколлинеарности, гетероскедастичности, автокорреляции применить для получения оценок параметров соответствующие методы. Модель оптимальной сложности искать среди регрессоров, в числе которых должны быть те, что были выявлены при анализе корреляционных полей.

Экспериментальные данные представляются преподавателем в виде таблицы наблюдений типа "вход-выход" в формате xlsx. Номер варианта задания (v.1, v.2,….) соответствует порядковому номеру студента в списке группы.

ЛИТЕРАТУРА

1. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ: В 2 кн. Кн. 2. – М.: Финансы и статистика. 1987.

2. *Вучков И., Бояджиева Л., Салаков Е*. Прикладной линейный регрессионный анализ. – М.: Финансы и статистика, 1987.