

Модели регрессии с переключениями

Бродский Б.Е.
(ЦЭМИ РАН, Москва)

План:

1. Введение: мотивация, обзор
2. Постановки задачи
3. Метод
4. Основные результаты
5. Эксперименты: расщепление смесей
6. Аномальные наблюдения
7. Регрессии с переключениями режимов
8. Выводы

1. Введение: мотивация, обзор

Модели регрессии с переменным режимом были впервые рассмотрены **Quandt (1958)** в эконометрической литературе, хотя в статистике они имеют длинную предисторию (см. Lindgren, 1978). Простейшая модель с двумя режимами имеет вид:

$$\begin{aligned} Y_t &= X_t\beta_1 + u_{1t} && \text{в 1-м режиме} \\ Y_t &= X_t\beta_2 + u_{2t} && \text{во 2-м режиме} . \end{aligned}$$

Переменная Y_t генерируется в 1-м или 2-м режиме, но не в двух одновременно. Предположим, что определен индикатор I_t :

$$\begin{aligned} I_t &= 1, && \text{если } Y_t \text{ формируется в режиме 1} \\ I_t &= 0, && \text{если } Y_t \text{ формируется в режиме 2} . \end{aligned}$$

Если индикатор I_t наблюдаем, то Y_t можно оценить с использованием известных методов. Если же индикатор ненаблюдаем, то ситуация существенно сложнее. Обычно вводят латентную переменную Z_t , связанную с предиктором X_{1t} зависимостью

$$Z_t = X_{1t}\beta_3 + u_{3t}$$

и определяют

$$I_t = \begin{cases} 1, & \text{если } Z_t > 0 \\ 0, & \text{если } Z_t \leq 0 . \end{cases}$$

Если случайные шумы u_{3t} некоррелированы с u_{1t}, u_{2t} , то говорят о регрессионной модели с экзогенными переключениями, в противном случае - о модели с эндогенными переключениями.

Для модели с эндогенными переключениями обычные методы оценивания в регрессионных моделях неприменимы. Модели с экзогенными переключениями исследовались **Goldfeld, Quandt (1973)**, которые предложили *модель регрессии с марковскими переключениями*.

В модели регрессии с марковскими переключениями вероятности смены режимов в последовательные периоды предполагаются постоянными. Обычно они описываются посредством переходной матрицы.

Другая модификация модели регрессии с марковскими переключениями была предложена **Lee, Porter (1984)**, которые рассматривали наблюдаемый дихотомический вероятностный индикатор W_t , обеспечивающий разделение выборки.

Иначе, постулируется следующая матрица вероятностей переходов:

$$\Lambda = [p_{ij}]_{i,j=0,1}, \quad p_{ij} = P\{I_t = j | I_{t-1} = i\}.$$

Если $p_{11} = p_{01}$, то индикатор W_t не дает никакой информации о разбиении выборки на режимы. Если p_{11} и $p_{01} = 0$, то индикатор W_t обеспечивает полное разбиение выборки. Если же $p_{11} < 1$ и $p_{01} \neq p_{11}$, то мы имеем дело с вероятностным разбиением выборки на режимы. Lee, Porter (1984) рассматривают пример с ценами на железнодорожный транспорт в США за период 1880-1886, на которые оказал влияние ценовой сговор в картеле, с использованием следующей регрессионной модели:

$$\log P_t = \beta_0 + \beta_1 X_t + \beta_2 I_t + u_t,$$

где $I_t = 0$ или $I_t = 1$ в зависимости от того, объявлялись ли ценовые войны в конкретный период.

Cosslett, Lee (1985) обобщили модель Lee, Porter на случай, когда шумы u_t сериально коррелированы. Кроме того, они предполагали, что индикатор I_t управляется скрытой марковской цепью с переходной матрицей:

$$\Lambda = [\lambda_{ij}]_{i,j=0,1}, \quad \lambda_{ij} = P\{I_t = j | I_{t-1} = i\}.$$

2. Постановки задач

В этом разделе мы опишем характерные задачи, приводящие к моделям регрессии с переключениями. Одной из наиболее известных задач этого рода является

Задача расщепления смесей вероятностных распределений

В наиболее простой ситуации предположим, что функция распределения наблюдений имеет следующий вид:

$$F(x) = (1 - \epsilon)F_0(x) + \epsilon F_1(x),$$

где $F_0(x)$ - функция распределения обычных наблюдений, $F_1(x)$ - функция распределения аномальных наблюдений, $0 < \epsilon < 1$ - вероятность появления аномального наблюдения.

Необходимо по выборке наблюдений $X^N = \{x_1, x_2, \dots, x_N\}$ протестировать гипотезу о статистической однородности (отсутствии аномальных наблюдений), а при отклонении этой гипотезы построить оценку доли аномальных наблюдений в выборке (ϵ) и расклассифицировать выборку X^N на подвыборки обычных и аномальных наблюдений.

Существуют три основных подхода к решению этой задачи:

метод максимального правдоподобия,

метод моментов

метод адаптивного вероятностного обучения.

В работе Айвазян, Бухштабер, Енюков, Мешалкин (1989) анализируются преимущества и недостатки каждого из этих подходов. В частности, отмечается невысокое качество статистических оценок, полученных методами моментов и максимального правдоподобия, а также сложность вычислительных процедур, реализующих эти методы. Теоретическое исследование метода адаптивного вероятностного обучения чрезвычайно сложно. К этому следует добавить, что ни один из этих методов не позволяет обоснованно отвергать нулевую гипотезу об отсутствии аномалий (т.е. $F_0 = F_1$ или $\epsilon = 0$).

Регрессионные модели с аномальными наблюдениями

Естественным обобщением модели смеси вероятностных распределений является модель регрессии с аномальными наблюдениями:

$$Y = X\beta + \epsilon,$$

где Y - вектор зависимых наблюдений размерности $n \times 1$; X - матрица предикторов размерности $n \times k$; β - вектор регрессионных коэффициентов размерности $k \times 1$; ϵ - вектор случайных шумов, имеющих плотность распределения, описывающуюся моделью

$$f_{\epsilon}(x) = (1 - \delta)f_0(x) + \delta f_1(x),$$

где $0 \leq \delta < 1$ - вероятность появления аномального наблюдения; $f_0(x)$ - плотность распределения обычных наблюдений; $f_1(x)$ - плотность распределения аномальных наблюдений. В частности, в модели регрессии с эффектом "засорения" по Хьюберу, $f_0(\cdot) = N(0, \sigma^2)$, $f_1(\cdot) = N(0, \Lambda^2)$, где $\Lambda^2 \gg \sigma^2$.

Отметим, что в теории робастных статистик обычно используются те или иные грубые оценки доли аномальных наблюдений в выборке δ . При этом ни один из предложенных методов оценки регрессий с аномальными наблюдениями не позволяет проверить нулевую гипотезу об отсутствии аномалий.

Модели регрессии с переменными коэффициентами

Другим естественным обобщением модели смеси вероятностных распределений является модель регрессии с переменными коэффициентами. Допустим, что в режиме нормального функционирования некоторая система описывается регрессионной моделью вида

$$Y = X\beta + \epsilon,$$

где Y - вектор зависимых наблюдений размерности $n \times 1$; X - матрица предикторов размерности $n \times k$; β - вектор регрессионных коэффициентов размерности $k \times 1$; ϵ - вектор случайных шумов размерности $n \times 1$.

Однако в аномальном режиме модель наблюдений меняется на

$$Y = X\gamma + \epsilon,$$

где $\beta \neq \gamma$. Иными словами, смена режима приводит к изменению набора регрессионных коэффициентов (α), причем механизм смены режима является чисто случайным:

$$\alpha = \begin{cases} \beta & \text{с вероятностью } 1 - \epsilon \\ \gamma & \text{с вероятностью } \epsilon \end{cases}$$

Необходимо по всей выборке наблюдений проверить нулевую гипотезу об отсутствии аномалий и в случае отклонения этой гипотезы дать оценку доли аномалий

в выборке ϵ . Отметим, что вектор аномальных регрессионных коэффициентов γ зачастую неизвестен. Вместе с тем вектор обычных коэффициентов β чаще всего известен или оценивается по "обучающей" выборке.

1. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерности. М., 1989.
2. Quandt (1958) Estimation of parameters of a linear regression system obeying two separate regimes. JASA, 53, 873-880.
3. Lindgren (1978) Markov regime models for mixed distributions and switching regressions. Scandinavian Journal of Statistics, 5, 81-91.
4. Goldfeld, Quandt (1973) A Markov model for switching regressions. J. of Econometrics, 1, 3-16.
5. Lee, Porter (1984) Switching regression models with imperfect sample separation information. Econometrica, 52, 391-418.
6. Cosslett, Lee (1985) Serial correlation in discrete variable models. J. of Econometrics, 27, 79-97.

3. Метод

Начнем с задачи о расщеплении смеси двух вероятностных распределений, описывающейся следующей моделью:

$$f(x) = (1 - \epsilon)f_0(x) + \epsilon f_0(x - h),$$

где параметры ϵ, h неизвестны.

Грубый метод оценивания этих параметров состоит в следующем. Обычные и аномальные наблюдения некоторым эвристическим приемом группируются в две подвыборки, и оценка $\hat{\epsilon}$ вычисляется как доля объема подвыборки аномальных наблюдений в общем объеме выборки. Ясно, что этот метод будет работать лишь при больших значениях параметра h . Однако идея группировки наблюдений в две подвыборки может быть использована при построении более точных методов оценивания.

Итак, задача состоит в том, чтобы построить оценки параметров ϵ, h по выборке данных $X^N = \{x_i\}_{i=1}^N$. Для решения этой задачи поступим следующим образом.

1) По исходной выборке X^N построим оценку среднего:

$$\theta_N = \frac{1}{N} \sum_{i=1}^n x_i.$$

2) Назначим параметр $b > 0$ и будем классифицировать наблюдения следующим образом: если наблюдение попадает в интервал $(\theta_N - b, \theta_N + b)$, то относим его к подвыборке обычных наблюдений, в противном случае - к подвыборке аномальных наблюдений.

3) Тогда получим для каждого $b > 0$ разбиение выборки X^N на две подвыборки

$$\begin{aligned} X_1 &= \{\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_{N_1}\}, & |\tilde{x}_i - \theta_N| < b, \\ X_2 &= \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_{N_2}\}, & |\hat{x}_i - \theta_N| \geq b \end{aligned}$$

Параметр b выбирается так, чтобы подвыборки X_1 и X_2 различались наилучшим образом. Для этого рассмотрим следующую статистику:

$$\Psi_N(b) = \frac{1}{N^2} (N_2 \sum_{i=1}^{N_1} \tilde{x}_i - N_1 \sum_{i=1}^{N_2} \hat{x}_i).$$

где $N = N_1 + N_2$, $N_1 = N_1(b)$, $N_2 = N_2(b)$ - объемы подвыборок обычных и аномальных наблюдений соответственно.

4) Назначим далее пороговую границу $C > 0$ и сравним ней величину максимума $J = \max |\Psi_N(b)|$ по множеству $b > 0$. Если $J \leq C$, то принимается гипотеза H_0 об отсутствии аномальных наблюдений; если же $J > C$, то гипотеза H_0 отвергается и строятся оценки параметров ϵ и h .

5) Определим число b_N^* из условия

$$b_N^* \in \arg \max_{b>0} |\Psi_N(b)|.$$

Тогда

$$\epsilon_N^* = N_2(b_N^*)/N, \quad h_N^* = \theta_N/\epsilon_N^*$$

являются непараметрическими оценками для ϵ и h соответственно. Как будет показано ниже, эти оценки обладают достаточной точностью для широкого класса распределений с экспоненциально убывающими "хвостами". Однако они являются смещенными. Для построения несмещенных и состоятельных оценок параметров ϵ и h воспользуемся следующими зависимостями:

$$\begin{aligned} \hat{\epsilon}_N \hat{h}_N &= \theta_N \\ \frac{1 - \hat{\epsilon}_N}{\hat{\epsilon}_N} &= \frac{f_0(\theta_N - b_N^* - \hat{h}_N) - f_0(\theta_N + b_N^* - \hat{h}_N)}{f_0(\theta_N + b_N^*) - f_0(\theta_N - b_N^*)}. \end{aligned}$$

Оценки $\hat{\epsilon}_N$ и $\hat{h}_N$ почти наверно сходятся к истинным значениям параметров ϵ и h при $N \rightarrow \infty$. Подвыборка аномальных наблюдений такова: $X_2(b_N^*)$.

Этот метод позволяет решить задачу о расщеплении смеси двух вероятностных распределений. Далее мы рассмотрим обобщение этого метода на задачи анализа регрессий с переменной структурой. Следующий раздел посвящен формальному анализу предложенного метода.

4. Основные результаты

Сформулируем предположения, при которых будет рассматриваться задача. Пусть на вероятностном пространстве $(\Omega, \mathcal{F}, P)$ задано семейство последовательностей $X^N = \{x_i\}_{i=1}^N$ независимых случайных величин x_i с плотностью распределения

$$f(x) = (1 - \epsilon)f_0(x) + \epsilon f_0(x - h),$$

где $0 \leq \epsilon < 1/2$, $h \geq 0$ - неизвестные параметры; $f_0(\cdot)$ симметрична относительно нуля.

Будем предполагать, что последовательность X^N удовлетворяет условию Крамера: $\exists H > 0$ такое, что $E \exp(tx_i) < \infty$ при $|t| < H$.

В следующей теореме получена экспоненциальная оценка для вероятности ошибки 1-го рода для предложенного метода.

Теорема 1.

Пусть $\epsilon = 0$. Тогда в предположении симметричности плотности $f_0(\cdot)$ и условия Крамера справедлива следующая оценка: для любого $C > 0$

$$P\{\max_{b>0} |\Psi_N(b)| > C\} \leq L_1 \exp(-L_2(C)N),$$

где константы $L_1, L_2 > 0$ не зависят от N .

Рассмотрим теперь характеристики описанного метода в общем случае $\epsilon h \neq 0$. Введем вероятности ошибок оценивания

$$\alpha(\delta) = P\{|\hat{\epsilon}_N - \epsilon| > \delta\}, \quad \beta(\gamma) = P\{|\hat{h}_N - h| > \gamma\}, \quad \delta, \gamma > 0.$$

Поведение этих величин определяется свойствами математического ожидания статистики $\Psi_N(b)$, допускающего следующую аппроксимацию:

$$E\Psi_N(b) = \Psi(b) + O\left(\frac{1}{N}\right) = r(b) - \epsilon h d(b) + O\left(\frac{1}{N}\right),$$

где

$$r(b) = \int_{\epsilon h - b}^{\epsilon h + b} f(x) x dx, \quad d(b) = \int_{\epsilon h - b}^{\epsilon h + b} f(x) dx.$$

Для $\Psi'(b)$ имеем

$$\begin{aligned} \Psi'(b) &= (\epsilon h + b)f(\epsilon h + b) + (\epsilon h - b)f(\epsilon h - b) - \epsilon h[f(\epsilon h + b) + f(\epsilon h - b)] \\ &= b(f(\epsilon h + b) - f(\epsilon h - b)). \end{aligned}$$

Отсюда видно, что если $f_0(\cdot) = f_1(\cdot)$ и $f_0(\cdot)$ - симметричная плотность, то $\Psi'(b) \equiv 0$ и $\Psi(b) \equiv 0$. В общем же случае $\Psi'(b)$ обращается в нуль в точках, являющихся корнями уравнения

$$f(\epsilon h + b) = f(\epsilon h - b). \quad (*)$$

Допустим, что это уравнение (*) имеет единственный корень $b^* \neq 0$ (это предположение выполнено для унимодальной плотности $f_0(\cdot)$ и $0 < \epsilon < 1/2$). Тогда справедлива следующая

Теорема 2.

Предположим, что выполнены условия теоремы 1 и уравнение (*) имеет единственный корень $b^* \neq 0$. Тогда

- 1) $b_N^* \rightarrow b^*$ по вероятности при $N \rightarrow \infty$;
- 2) оценки $\hat{\epsilon}_N, \hat{h}_N$ сходятся по вероятности к истинным значениям параметров ϵ, h , соответственно, при $N \rightarrow \infty$.

Аномальные наблюдения

Модель ϵ -засорения:

$$f_\epsilon(\cdot) = (1 - \epsilon)\mathcal{N}(\mu, \sigma^2) + \epsilon\mathcal{N}(\mu, \Lambda^2), \quad \Lambda^2 \gg \sigma^2, \quad 0 \leq \epsilon < 1/2.$$

В этой модели используется следующее обобщение изложенного выше метода:

1. По выборке наблюдений $X^N = \{x_1, \dots, x_N\}$ рассчитывается оценка среднего: $\hat{\mu} = \sum_{i=1}^N x_i$
2. Вычисляется диагностическая последовательность $y_i = (x_i - \hat{\mu})^2$, $i = 1, \dots, N$ и ее эмпирическое среднее: $\theta_N = \sum_{i=1}^N y_i$
3. Далее для $b \in [0, B_{max}]$ выборка $Y^N = \{y_1, \dots, y_N\}$ разбивается на два класса: при $\theta_N(1 - \phi(b)) \leq y_i \leq \theta_N(1 + b)$ полагаем $\tilde{y}_i = y_i$ (объем подвыборки $N_1 = N_1(b)$), в противном случае $\hat{y}_i = y_i$ (объем подвыборки $N_2 = N_2(b)$). При этом

$$\phi(b) = 1 - \frac{b}{e^b - 1}.$$

4. Для $b \in [0, B_{max}]$ рассчитывается статистика

$$\Psi_N(b) = \frac{1}{N^2} (N_2 \sum_{i=1}^{N_1} \tilde{y}_i - N_1 \sum_{i=1}^{N_2} \hat{y}_i).$$

где $N = N_1 + N_2$, $N_1 = N_1(b)$, $N_2 = N_2(b)$ - объемы подвыборок обычных и аномальных наблюдений, соответственно.

5. Далее, как и в основном методе, назначается пороговая граница $C > 0$ и с ней сравнивается величина максимума $J = \max |\Psi_N(b)|$ по множеству $b > 0$. Если $J \leq C$, то принимается гипотеза H_0 об отсутствии аномальных наблюдений; если же $J > C$, то гипотеза H_0 отвергается и строятся оценки параметров ϵ и h .

- 5) Определим число b_N^* из условия

$$b_N^* \in \arg \max_{b>0} |\Psi_N(b)|.$$

Тогда

$$\epsilon_N^* = N_2(b_N^*)/N.$$

Регрессионные модели с переключениями режимов

Модель наблюдений имела следующий вид:

$$y = X\beta + u = X((1 - \epsilon)\beta_0 + \epsilon\beta_1) + u,$$

где

y - $n \times 1$ - вектор зависимых наблюдений;

X - $n \times k$ матрица предикторов;

u - $n \times 1$ вектор случайных шумов;

$\beta = (1 - \epsilon)\beta_0 + \epsilon\beta_1$ - $k \times 1$ вектор коэффициентов модели. Предполагается, что коэффициент при i -м предикторе может измениться с некоторой неизвестной вероятностью ϵ_i , т.е. $\epsilon = (\epsilon_1, \dots, \epsilon_k)'$. При этом $\beta_0 \neq \beta_1$.

Необходимо проверить гипотезу об отсутствии переключений по каждому коэффициенту, т.е. $\epsilon_i = 0$, а при отклонении этой гипотезы построить оценку параметра ϵ_i для $i = 1, \dots, k$.

Для решения этой задачи рассмотрим МНК-оценку вектора β :

$$\hat{\beta} = (X'X)^{-1}X'y = (1 - \epsilon)\beta_0 + \epsilon\beta_1 + (X'X)^{-1}X'u.$$

Т.к. последовательность шумов u центрирована, задача сводится к рассмотренной выше задаче обнаружения переключений по среднему случайной последовательности. Матрица предикторов X будет влиять лишь на случайную компоненту, что отразится на критическом пороге.

5. Эксперименты

Предложенный метод тестировался в имитационных экспериментах с выборками различной природы.

Расщепление смесей

1) В первой серии экспериментов исследовалась модель наблюдений следующего вида:

$$f_{\epsilon}(x) = (1 - \epsilon)f_0(x) + \epsilon f_0(x - h), \quad f_0(\cdot) = \mathcal{N}(0, 1), \quad 0 \leq \epsilon < 1/2.$$

Вначале были рассчитаны критические пороги статистики $\max |\Psi_N(b)|$ по $b > 0$. С этой целью для однородных выборок (при $\epsilon = 0$) рассчитывались α -квантили статистики $\max |\Psi_N(b)|$ по $b > 0$ ($\alpha = 0.95, 0.99$). Результаты, полученные усреднением в $k = 5000$ независимых повторениях каждого эксперимента, приведены в табл.1.

Таблица 1.

N	50	100	300	500	800	1000	1200	1500	2000	3000
$\alpha = 0.95$	0.1681	0.1213	0.0710	0.0534	0.044	0.0380	0.037	0.034	0.029	0.022
$\alpha = 0.99$	0.1833	0.1410	0.0869	0.0666	0.050	0.0471	0.0390	0.038	0.035	0.029

Значение квантили при $\alpha = 0.95$ было выбрано в качестве критического порога C при исследовании неоднородных выборок (при $\epsilon \neq 0$). При различных значениях объема выборки в $k = 5000$ независимых повторениях каждого эксперимента рассчитывалась оценка вероятности ошибки 2-го рода w_2 (т.е. частота события $\max_b |\Psi_N(b)| < C$ при $\epsilon > 0$), а также оценка $\hat{\epsilon}$ параметра ϵ . Результаты экспериментов приведены в табл. 2.

Таблица 2.

$\epsilon = 0.1$	h=2.0				h=1.5			
N	300	500	800	1000	800	1200	2000	3000
C	0.0710	0.0534	0.044	0.038	0.044	0.037	0.029	0.022
w_2	0.26	0.15	0.05	0.02	0.62	0.42	0.16	0.03
$\hat{\epsilon}$	0.104	0.101	0.097	0.099	0.106	0.103	0.102	0.0985

2) Во второй серии экспериментов исследовалась классическая модель ϵ -засорения:

$$f_{\epsilon}(\cdot) = (1 - \epsilon)\mathcal{N}(\mu, \sigma^2) + \epsilon\mathcal{N}(\mu, \Lambda^2), \quad \Lambda^2 \gg \sigma^2, \quad 0 \leq \epsilon < 1/2.$$

Вначале были рассчитаны критические пороги статистики $\max |\Psi_N(b)|$ по $b > 0$. С этой целью для однородных выборок (при $\epsilon = 0$) рассчитывались α -квантили статистики $\max |\Psi_N(b)|$ по $b > 0$ ($\alpha = 0.95, 0.99$). Результаты, полученные усреднением в $k = 5000$ независимых повторениях каждого эксперимента, приведены в табл.3.

Таблица 3.

N	50	100	300	500	800	1000	1200	1500	2000	3000
0.95	0.3031	0.2330	0.1570	0.1419	0.1252	0.1244	0.1146	0.1107	0.1075	0.1019
0.99	0.3699	0.2862	0.1947	0.1543	0.1436	0.1331	0.1269	0.1190	0.1157	0.1087

Значение квантили при $\alpha = 0.95$ было выбрано в качестве критического порога C при исследовании неоднородных выборок (при $\epsilon \neq 0$). При различных значениях объема выборки в $k = 5000$ независимых повторениях каждого эксперимента рассчитывалась оценка вероятности ошибки 2-го рода w_2 (т.е. частота события $\max_b |\Psi_N(b)| < C$ при $\epsilon > 0$), а также оценка $\hat{\epsilon}$ параметра ϵ . Результаты экспериментов приведены в табл. 4, 5.

Таблица 4.

$\Lambda = 3.0$	$\epsilon = 0.05$			
N	300	500	800	1000
C	0.1570	0.1419	0.1252	0.1244
w_2	0.27	0.15	0.06	0.04
$\hat{\epsilon}$	0.064	0.056	0.052	0.05

Таблица 5.

$\Lambda = 5.0$	$\epsilon = 0.01$				
N	1000	1200	1500	2000	3000
C	0.1244	0.1146	0.1107	0.1075	0.1019
w_2	0.25	0.20	0.15	0.10	0.04
$\hat{\epsilon}$	0.0135	0.013	0.012	0.011	0.010

Регрессии с аномальными наблюдениями

3) В третьей серии экспериментов исследовалась следующая регрессионная модель с аномальными наблюдениями:

$$y_i = 1 + 2 * i + \epsilon_i, \quad f_\epsilon(\cdot) = (1 - \epsilon)\mathcal{N}(0, \sigma^2) + \epsilon\mathcal{N}(0, \Lambda^2).$$

Случай модели регрессии с аномальными наблюдениями сводится к предыдущему рассмотрению последовательности регрессионных остатков и применении к ней описанной выше процедуры. Вначале были рассчитаны критические пороги статистики $\max |\Psi_N(b)|$ по $b > 0$. С этой целью для однородных выборок (при $\epsilon = 0$) рассчитывались α -квантили статистики $\max |\Psi_N(b)|$ по $b > 0$ ($\alpha = 0.95, 0.99$). Рассчитанные значения квантилей полностью совпадают в данными, приведенными в табл.3. Значение квантили при $\alpha = 0.95$ было выбрано в качестве критического порога C при исследовании неоднородных выборок (при $\epsilon \neq 0$). При различных значениях объема выборки в $k = 5000$ независимых повторениях каждого эксперимента рассчитывалась оценка вероятности ошибки 2-го рода w_2 (т.е. частота события $\max_b |\Psi_N(b)| < C$ при $\epsilon > 0$), а также оценка $\hat{\epsilon}$ параметра ϵ . Результаты экспериментов приведены в табл. 6, 7.

Таблица 6.

$\Lambda = 3.0$	$\epsilon = 0.05$				
N	300	500	800	1000	1500
C	0.16	0.14	0.124	0.120	0.113
w_2	0.27	0.15	0.06	0.04	0.01
$\hat{\epsilon}$	0.062	0.057	0.052	0.051	0.050

Таблица 7.

$\Lambda = 5.0$	$\epsilon = 0.01$				
N	1000	1200	1500	2000	3000
C	0.120	0.116	0.113	0.106	0.101
w_2	0.24	0.20	0.15	0.10	0.04
$\hat{\epsilon}$	0.014	0.012	0.011	0.01	0.010

Регрессии с переключениями режимов

4) В 4-й серии экспериментов исследовались регрессионные модели с переключениями режимов.

Модель наблюдений имела следующий вид:

$$y = X\beta + u = X((1 - \epsilon)\beta_0 + \epsilon\beta_1) + u,$$

где

y - $n \times 1$ - вектор зависимых наблюдений;

X - $n \times k$ матрица предикторов;

u - $n \times 1$ вектор случайных шумов;

$\beta = (1 - \epsilon)\beta_0 + \epsilon\beta_1$ - $k \times 1$ вектор коэффициентов модели. Предполагается, что коэффициент при i -м предикторе может измениться с некоторой неизвестной вероятностью ϵ_i , т.е. $\epsilon = (\epsilon_1, \dots, \epsilon_k)'$. При этом $\beta_0 \neq \beta_1$.

Необходимо проверить гипотезу об отсутствии переключений по каждому коэффициенту, т.е. $\epsilon_i = 0$, а при отклонении этой гипотезы построить оценку параметра ϵ_i для $i = 1, \dots, k$.

В следующем эксперименте рассматривалась регрессия с одним предиктором:

$$y_i = c_1 + c_2 * i + u_i, \quad u_i \sim N(0; 1), \quad i = 1, \dots, n.$$

$$\xi \sim U[0; 1]$$

$$\beta = [c_1; c_2] = \begin{cases} \beta_1, & \epsilon_1 < \xi \leq 1 \\ \beta_2, & 0 \leq \xi \leq \epsilon_1 \end{cases}$$

Таблица 8.

$\epsilon = 0.05$	$\beta_1 = [1; 1], \beta_2 = [1; 2]$			
N	300	500	800	1000
C	0.07	0.05	0.04	0.03
w_2	0.87	0.59	0.14	0.004
$\hat{\epsilon}$	0.08	0.059	0.052	0.05

Таблица 9.

$\epsilon = 0.1$	$\beta_1 = [1; 1], \beta_2 = [1; 1.5]$			
N	300	500	800	1000
C	0.07	0.05	0.04	0.03
w_2	0.83	0.65	0.13	0.0
$\hat{\epsilon}$	0.15	0.12	0.102	0.10

В следующем эксперименте рассматривалась регрессия с двумя предикторами:

$$y_i = c_1 + c_2 * i + c_3 * \log(i) + u_i, \quad u_i \sim N(0; 1), \quad i = 1, \dots, n.$$

$$\xi \sim U[0; 1]$$

$$\beta = [c_1; c_2; c_3] = \begin{cases} \beta_1, & \epsilon_1 + \epsilon_2 < \xi \leq 1 \\ \beta_2, & 0 \leq \xi \leq \epsilon_1 \\ \beta_3, & \epsilon_1 < \xi \leq \epsilon_1 + \epsilon_2. \end{cases}$$

Таблица 10.

$\epsilon_1 = 0.05; \epsilon_2 = 0.1$	$\beta_1 = [1; 1; 1], \beta_2 = [1; 2; 1], \beta_3 = [1; 1; 1.5]$			
n	300	500	800	1000
C	0.07	0.05	0.04	0.03
$w_2^1; w_2^2$	0.87; 0.95	0.63; 0.84	0.156; 0.152	0.018; 0.006
$\hat{\epsilon}_1; \hat{\epsilon}_2$	0.08; 0.15	0.06; 0.13	0.051; 0.103	0.0503; 0.10

Выводы

1. Предложен новый метод расщепления смесей вероятностных распределений, выделения аномальных наблюдений и различных режимов регрессии.
2. Предложенный метод позволяет обоснованно отвергнуть нулевую гипотезу об отсутствии переключений (аномальных наблюдений) и построить состоятельные оценки параметров различных регрессионных режимов
3. Свойства метода исследованы как теоретически (теоремы 1 и 2 об экспоненциально сходимости к нулю вероятностей ошибок), так и экспериментально: имитационные выборки смесей распределений, регрессий с аномальными наблюдениями и переключениями коэффициентов

Литература

1. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерности. М., 1989.
2. Quandt (1958) Estimation of parameters of a linear regression system obeying two separate regimes. JASA, 53, 873-880.
3. Lindgren (1978) Markov regime models for mixed distributions and switching regressions. Scandinavian Journal of Statistics, 5, 81-91.
4. Goldfeld, Quandt (1973) A Markov model for switching regressions. J. of Econometrics, 1, 3-16.
5. Lee, Porter (1984) Switching regression models with imperfect sample separation information. Econometrica, 52, 391-418.
6. Cosslett, Lee (1985) Serial correlation in discrete variable models. J. of Econometrics, 27, 79-97.