DOI: 10.5281/zenodo.10957502

Силаев А.М., Силаева М.В.
Нижний Новгород, НИУ ВШЭ – Нижний Новгород

ОЦЕНИВАНИЕ ПАРАМЕТРОВ МОДЕЛЕЙ БИНАРНОГО ВЫБОРА С УЧЕТОМ ИЗМЕНЕНИЙ В СЛУЧАЙНЫЙ МОМЕНТ ВРЕМЕНИ

Задача оценки параметров последовательности независимых бинарных наблюдаемых случайных величин с учетом изменений в случайный момент времени решалась во многих работах. Например, в [1-3] рассматривались модели, в которых наблюдается последовательность независимых случайных величин $y_1^T \equiv \{y_1,\ y_2,\ \dots\ ,\ y_T\}$, в которой y_i распределены по Бернулли, то есть принимают значения 0 и 1 с вероятностями

$$P(y_k = 1) = \begin{cases} \theta_0, & k < \tau \\ \theta_1, & k \geq \tau \end{cases}, \quad (k = 1,\ 2,\ \dots\ ,\ T), \tag{1}$$

где θ_0, θ_1 и τ – оцениваемые параметры, $P(y_t = 0) = 1 - P(y_t = 1)$. В [1] для получения оценок параметров использовался метод максимального правдоподобия, в [2] оценки вычислялись с помощью байесовского анализа, в [3] была предложена статистика типа кумулятивных сумм, которая использовалась для оценивания вероятности гипотезы отсутствия изменений и оценивалось значение момента τ.

В настоящей работе рассматриваются модели бинарных регрессий с учетом изменений параметров в случайный момент времени, которые обобщают модель (1), поскольку предполагают, что вероятности состояний бинарного наблюдаемого процесса $y_1^T \equiv \{y_1,\ y_2,\ \dots\ ,\ y_T\}$ могут зависеть от регрессоров. Пусть наблюдаемая переменная y_k принимает бинарные значения 0 или 1 в зависимости от того, больше или меньше нуля ненаблюдаемая (латентная) переменная y_k^*:

$$y_k = \begin{cases} 1, & y_k^* \geq 0 \\ 0, & y_k^* < 0 \end{cases}, \quad (k = 1,\ 2,\ \dots\ ,\ T). \tag{2}$$

Скрытая переменная y_k^* описывается моделью линейной регрессии со скачкообразным изменением параметров в случайный момент времени τ:

$$y_k^* = \begin{cases} \beta_0 x_k + u_k, & k < \tau; \\ \beta_1 x_k + u_k, & k \geq \tau; \end{cases} \quad (k = 1,\ 2,\ \dots\ ,\ T). \tag{3}$$

Здесь x_k – вектор регрессоров, β_0 и β_1 – параметры модели, u_k – случайная ошибка. Предполагается, что u_k – независимые случайные величины с нулевым средним значением и единичной дисперсией с распределением $F(u)$. Из (2), (3) следует, что

$$P(y_k = 1) = P(y_k^* \geq 0) = \begin{cases} 1 - F(-\beta_0 x_k), & k < \tau \\ 1 - F(-\beta_1 x_k), & k \geq \tau \end{cases}, \quad (k = 1,\ 2,\ \dots\ ,\ T). \tag{4}$$

Если случайные ошибки u_k имеют симметричное относительно нуля распределение, то $F(u) = 1 - F(-u)$, поэтому в этом случае из (4) получим

$$P(y_k = 1) = \begin{cases} F(\beta_0 x_k), & k < \tau \\ F(\beta_1 x_k), & k \geq \tau \end{cases}, \quad (k = 1,\ 2,\ \ldots\ ,\ T). \qquad (5)$$

В частном случае, если $x_k = 1$ и параметры модели β_0 и β_1 также скалярные величины, то выражение (5) принимает вид (1), считая, что $\theta_0 = F(\beta_0)$ и $\theta_1 = F(\beta_1)$.

К настоящему времени по проблеме обнаружения скачкообразных изменений свойств случайных процессов и оценке их параметров опубликовано большое число работ (см., например, монографии [4-6], обзоры и библиографии [7-9]), но вопросы оценки параметров и апостериорных вероятностей момента появления скачка в моделях бинарного выбора исследованы недостаточно. При этом модели бинарного выбора широко используются в экономике, медицине, биологии, физике и других науках.

В настоящей работе для оценивания параметров моделей бинарного выбора со скачкообразными изменениями в случайный момент времени предлагается алгоритм, основанный на использовании моделей марковских случайных последовательностей и априорной вероятности момента появления скачка на интервале наблюдения. В отличие от известных методов исследуемый алгоритм позволяет находить не только оценку момента скачкообразного изменения параметров, а целиком апостериорное распределение вероятности момента появления скачка, которое содержит более полную информацию и может быть полезным при анализе качества оценивания.

Будем считать, что β_0, β_1 и τ взаимонезависимы и заданы априорные вероятности $P_\tau(\tau)$ дискретных целочисленных значений случайного момента скачка $\tau \geq 1$. Задача состоит в том, чтобы по реализациям наблюдений y_1^T и регрессоров x_1^T найти оценки параметров β_0, β_1 и момента скачка τ. Для решения поставленной задачи применим вариант EM алгоритма [10]. Если параметр τ заранее известен, то весь интервал наблюдения можно разбить на участки, соответствующие значениям $k < \tau$ (отсутствия скачка к моменту k) и $k \geq \tau$ (появления скачка к моменту k).

Для интервала времени $1 \leq k < \tau$, применяя формулу Байеса, можно записать выражения для рекуррентного вычисления функции правдоподобия $l_0(\beta_0; \tau - 1) \equiv P(y_1^{\tau-1}, x_1^{\tau-1}|\beta_0)$ вектора параметров β_0 при всех значениях параметра τ из интервала $1 \leq \tau \leq T$. Аналогично для интервала времени $\tau \leq k \leq T$, применяя формулу Байеса, можно записать выражения для рекуррентного вычисления апостериорной плотности вероятности $l_1(\beta_1; \tau, T) \equiv P(y_\tau^T, x_\tau^T|\beta_1)$ вектора параметров β_1 при $1 \leq \tau \leq T$. С другой стороны, если векторы параметров β_0 и β_1 заранее известны, то можно записать выражения для рекуррентного вычисления апостериорной вероятности $W_\tau(\tau; \beta_0, \beta_1, T) \equiv P(\tau|y_1^T, x_1^T, \beta_0, \beta_1)$ момента скачка τ, используя свойство марковости процесса x_k и формулу Байеса.

В соответствии с EM алгоритмом вместо точных значений параметров модели θ_0, θ_1 и τ, которые в реальности не известны, в формулы для $l_0(\beta_0;\tau-1)$ и $l_1(\beta_1;\tau,T)$ подставляем оценки $\hat{\tau}$, а в уравнения для $W_\tau(\tau;\beta_0,\beta_1,T)$ оценки $\hat{\beta}_0$, $\hat{\beta}_1$. Можно организовать чередование вычислений оценок параметров в рассматриваемой задаче следующим образом. На первом шаге при некотором начальном значении оценки момента появления $\hat{\tau}^{(0)}$ из $l_0\left(\beta_0;\hat{\tau}^{(0)}-1\right)$ и $l_1\left(\beta_1;\hat{\tau}^{(0)},T\right)W_1\left(\theta_1;\ \hat{\tau}^{(0)},T\right)$ находим оценки для параметров $\hat{\beta}_0^{(0)}$ и $\hat{\beta}_1^{(0)}$ в соответствии с критерием максимального правдоподобия для модели бинарного выбора (логит или пробит модели в зависимости от функции распределения шумов u_k в уравнении (3)). Далее при фиксированных значениях параметров $\hat{\beta}_0^{(0)}$ и $\hat{\beta}_1^{(0)}$ вычисляется оценка момента скачка $\hat{\tau}^{(1)}$, оптимальная, например, по критерию максимума апостериорной вероятности $W_\tau\left(\tau;\hat{\beta}_0^{(0)},\hat{\beta}_1^{(0)},T\right)$. Используя $\hat{\tau}^{(1)}$, с помощью $l_0\left(\beta_0;\hat{\tau}^{(1)}-1\right)$ и $l_1\left(\beta_1;\hat{\tau}^{(1)},T\right)$ находим оценки для параметров $\hat{\beta}_0^{(1)}$ и $\hat{\beta}_1^{(1)}$, которые затем используются для оценивания на втором шаге $\hat{\tau}^{(2)}$, и т. д. В итоге вычисления производятся в количестве M итераций в соответствии со схемой:

$$\hat{\tau}^{(i-1)} \Rightarrow \hat{\beta}_0^{(i-1)},\ \hat{\beta}_1^{(i-1)}\ \Rightarrow\ \hat{\tau}^{(i)}\ \Rightarrow\ \hat{\beta}_0^{(i)},\ \hat{\beta}_1^{(i)},\ \ (i=1,2,\dots,M).$$

Можно ожидать, что в результате достаточно большого количества итераций M оценки параметров будут сходится к значениям близким к истинным.

Проверка работоспособности полученного алгоритма проводилась с помощью компьютерного моделирования ряда тестовых примеров. Генерировались реализации скрытой переменной y_k^* в виде регрессии со скачкообразными изменениями параметров:

$$y_k^* = \begin{cases}\alpha_0+\gamma_0 z_k+u_k, & k<\tau;\\ \alpha_1+\gamma_1 z_k+u_k, & k\geq\tau;\end{cases}\quad (k=1,\ 2,\ \dots\ ,\ T). \qquad (6)$$

и реализации бинарной переменной $y_1^T \equiv \{y_1,\ y_2,\ \dots\ ,\ y_T\}$ с помощью уравнения (2). Здесь $\alpha_0,\gamma_0,\alpha_1,\gamma_1$ и τ – оцениваемые параметры модели; u_k – независимые гауссовские случайные величины с нулевым средним значением и единичной дисперсией: $u_k \sim iidN(0,1)$. Процесс z_k задавался в виде белого гауссовского шума с нулевым средним значением и единичной дисперсией: $z_k \sim iidN(0,1)$. Отметим, что модель (6) принимает вид регрессии (3), если ввести обозначения для вектор-строк параметров $\beta_0=(\alpha_0,\gamma_0)$, $\beta_1=(\alpha_1,\gamma_1)$ и вектор-столбца регрессоров $x_k=(1,z_k)'$, где $'$– знак транспонирования. По реализации наблюдений с помощью полученного алгоритма вырабатывались оценки параметров $\hat{\beta}_0=(\hat{\alpha}_0,\hat{\gamma}_0)$, $\hat{\beta}_1=(\hat{\alpha}_1,\hat{\gamma}_1)$ и $\hat{\tau}$. Априорная

вероятность момента появления скачка τ задавалась равномерной на интервале времени $[1, T]$:

$$P_\tau(\tau) = \begin{cases} 0, & \tau < 0, \quad \tau > T \ ; \\ \dfrac{1}{T}, & \tau = 1,\ 2,\ \ldots,\ T. \end{cases}$$

Начальное значение оценки момента появления скачка $\hat{\tau}^{(0)}$ подбиралось для каждой наблюдаемой реализации, исходя из критерия минимизации доли неправильной классификации для модели бинарного выбора. Проводилось $M = 10$ итераций в схеме EM алгоритма для сходимости оценок параметров.

Моделирование позволило сделать выводы, что точность оценивания параметров модели с помощью исследуемого алгоритма зависит от того, как сильно различаются значения параметров в разных режимах работы, от уровня шумов в рассматриваемой модели, от числа наблюдений и других факторов. Приведем результаты моделирования для конкретных значений параметров в модели (6): $\alpha_0 = 0, \quad \gamma_0 = d, \alpha_1 = d, \gamma_1 = 2d; T = 1000$. Для вычисления оценок момента скачка $\hat{\tau}$ использовался критерий максимума апостериорной вероятности:

$$\hat{\tau} = \underset{\tau}{argmax} W_\tau\left(\tau; \hat{\beta}_0, \hat{\beta}_1, T\right).$$

Кроме того, с помощью $W_\tau\left(\tau; \hat{\beta}_0, \hat{\beta}_1, T\right)$ вычислялась также апостериорная дисперсия момента появления скачка $D_\tau\left(T; \hat{\beta}_0, \hat{\beta}_1\right)$. На рис. 1, 2 представлены графики соответственно средних значений смещений вырабатываемой оценки $\Delta = \overline{\hat{\tau} - \tau_0}$ и апостериорных среднеквадратичных отклонений $\delta = \overline{\sqrt{D_\tau\left(T; \hat{\beta}_0, \hat{\beta}_1\right)}}$ в зависимости от истинного момента появления скачка τ_0 при различных величинах параметра d, изменяющихся от 0,1 до 100. Для иллюстрации на рис. 3 представлен график апостериорных вероятностей $W(\tau) = \overline{W_\tau\left(\tau; \hat{\beta}_0, \hat{\beta}_1, T\right)}$, вырабатываемых алгоритмом для исследуемой модели при $\alpha_0 = 0, d = 1, \quad \gamma_0 = 1, \alpha_1 = 1, \gamma_1 = 2, \tau_0 = 400$. Верхняя черта при вычислении Δ, δ и апостериорных вероятностей $W(\tau)$ означает усреднение, которое проводилось с помощью генерации и обработки 100 независимых реализаций в модели (2), (6).

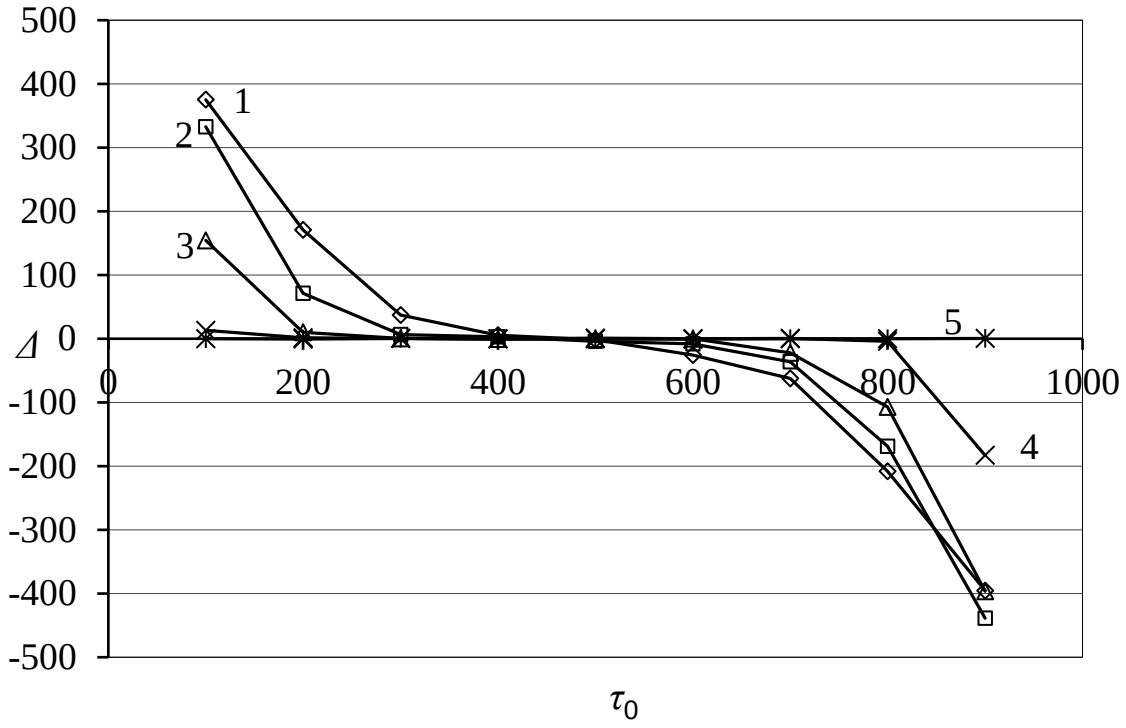

Рис.1. Зависимость смещения оценки Δ от истинного момента появления скачка τ_0. 1: $d = 0{,}1$; 2: $d = 0{,}2$; 3: $d = 1$; 4: $d = 10$; 5: $d = 100$

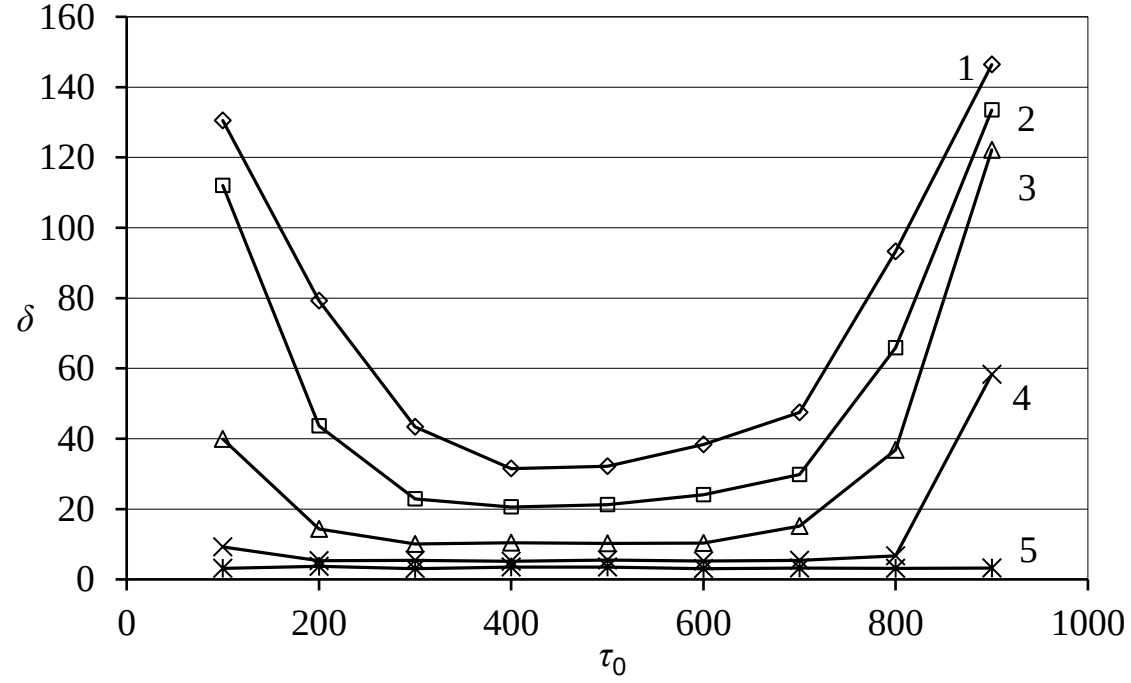

Рис.2. Зависимость среднеквадратичного отклонения δ от момента появления скачка τ_0. 1: $d = 0{,}1$; 2: $d = 0{,}2$; 3: $d = 1$; 4: $d = 10$; 5: $d = 100$

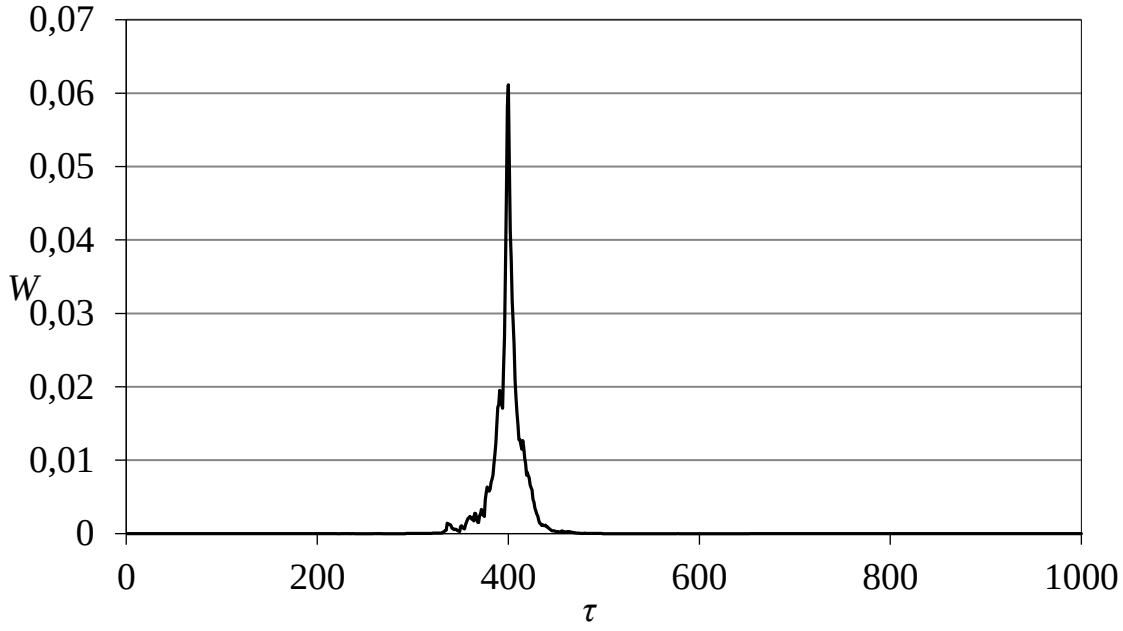

Рис.3. График зависимости апостериорной вероятности $W(\tau)$ от τ при $\alpha_0 = 0, \quad \gamma_0 = 1, \alpha_1 = 1, \gamma_1 = 2, \tau_0 = 400$

Как видно из рис. 1, 2 при увеличении параметра d, то есть при увеличении соотношения сигнал/шум в рассматриваемой модели смещение оценки Δ и среднеквадратичное отклонение δ стремятся к нулю, что свидетельствует об асимптотической несмещенности и состоятельности вырабатываемых оценок момента появления скачка τ. Но, если истинный момент появления скачка τ_0 расположен слишком близко к началу или к концу интервала наблюдения $[1, T]$, то точность оценивания снижается, так как становится недостаточно данных для оценивания параметров модели.

Проведенное компьютерное моделирование ряда тестовых примеров подтверждает работоспособность предлагаемого алгоритма оценки параметров моделей бинарных регрессий с учетом изменений параметров в случайный момент времени. Точность оценивания зависит от отношения сигнал/шум в конкретных рассматриваемых задачах. При этом исследуемый алгоритм позволяет находить не только оценки, а целиком апостериорные распределения вероятности моментов появления скачков τ, которые содержат более полную информацию о случайных величинах τ и могут быть полезными при анализе качества оценивания.

Список использованной литературы:

1. Hinkley D. V., Hinkley E. A. Inference about the change-point in a sequence of binomial variables / D. V. Hinkley, E. A. Hinkley // Biometrika. – 1970. – Vol. 57. – No. 3. – P. 477-488.

2. Smith A. F. M. A Bayesian approach to inference about a change-point in a sequence

of random variables // Biometrika. – 1975. – Vol. 62. – No. 2. – P. 407-416.

3. Pettitt, A. N. A Simple Cumulative Sum Type Statistic for the Change-Point Problem with Zero-One Observations // Biometrika. – 1980. – Vol. 67. – No. 1. – P. 79-84.

4. Обнаружение изменения свойств сигналов и динамических систем / Ред. М. Бассвиль, А. Банвениста. – М. : Мир, 1989. - 278 с.

5. Chen J., Gupta A. L. Parametric Statistical Change Point Analysis. With Applications to Genetics, Medicine, and Finance. Second Edition. 2012. Springer Science+Business Media.

6. Brodsky B. Change-Point Analysis in Nonstationary Stochastic Models. 2017. Taylor & Francis Group.

7. Khodadadi A. Change-point Problem and Regression: An Annotated Bibliography / Khodadadi A., Asgharian M. // COBRA Preprint Series. Nov. 2008. Working Paper 44.

8. Tze-San Lee. Change-Point Problems: Bibliography and Review // Journal of Statistical Theory and Practice. – 2010. – V. 4. – No. 4. – P. 643-662.

9. Aminikhanghahi S. A Survey of Methods for Time Series Change Point Detection / Aminikhanghahi S., Cook D. J. // Knowledge and Information Systems. – 2017. – V. 51. – Iss. 2. – P. 339-367.

10. Dempster A.P. Maximum Likelihood from Incomplete Data via the EM Algorithm / A.P. Dempster, N.M. Laird, D.B. Rubin // Journal of the Royal Statistical Society. – 1977. – Series B (Methodological). – Vol. 39. – No. 1. – P. 1-38.