СТАТЬИ

УДК 338.27

ТОЧЕЧНОЕ И ИНТЕРВАЛЬНОЕ ОЦЕНИВАНИЕ РЕГРЕССИОННЫХ МОДЕЛЕЙ БАЙЕСОВСКИМ МЕТОДОМ В ПРОГРАММНОЙ СРЕДЕ R

Бабешко Л.О.

ФГБОУ ВО «Финансовый университет при Правительстве РФ», Москва, e-mail: LBabeshko@fa.ru

Статья посвящена оцениванию регрессионных моделей при помощи байесовского подхода. Байесовский метод нашёл широкое применение при оценивании эконометрических моделей по выборочным данным малого объема, а также в случаях, когда классические методы (метод наименьших квадратов, метод максимального правдоподобия) неприменимы. Байесовский метод, как способ формализации степени уверенности в истинности некоторого утверждения и ее корректировки по мере поступления новой информации относительно исследуемого явления, позволяет значительно сузить интервальные оценки параметров регрессионных моделей по сравнению с классическим подходом. Предварительная информация, выражаемая в виде априорной функции вероятности предполагаемого результата, трансформируется в апостериорное распределение плотности вероятности, с учетом обработки выборочных данных. В работе приводятся результаты сравнительного анализа точечного и интервального оценивания модели множественной линейной регрессии в рамках байесовской регрессии и метода максимального правдоподобия. В качестве инструментальных средств выбрана программная среда R, в которой байесовская парадигма представлена в функциях многих статистических пакетов. Результаты, приведенные в статье, получены при помощи пакета MCMC (Monte Carlo Markov chain), основу которого составляет построение марковского процесса, стационарное распределение которого определяется апостериорной функцией распределения.

Ключевые слова: точечные оценки, интервальные оценки, байесовская регрессия, априорное распределение, апостериорное распределение

POINT AND INTERVAL ESTIMATION OF REGRESSION MODELS BY BAYESIAN METHOD IN THE R SOFTWARE ENVIRONMENT

Babeshko L.O.

Financial University under the Government of the Russian Federation, Moscow, e-mail: LBabeshko@fa.ru

The article is devoted to estimation of regression models by means of Bayesian approach. The Bayesian method has found wide application in the estimation of econometric models from sample data of small volume, as well as in cases where classical methods (least squares method, maximum likelihood method) are not applicable. The Bayesian method, as a way of formalizing the degree of confidence in the truth of some statement, and its correction as new information about the phenomenon under study becomes available, allows us to significantly narrow the interval estimates of the parameters of regression models in comparison with the classical approach. Preliminary information expressed in the form of a priori probability functions of the intended outcome, converted in the posterior distribution of the probability density, taking into account the processing of the sample data. The paper presents the results of a comparative analysis of point and interval estimation of the model of multiple linear regression in the framework of Bayesian regression and the maximum likelihood method. In the quality of the tools, the software environment R is chosen, in which the Bayesian paradigm is represented in the functions of many statistical packages. The results presented in the article are obtained using the package MCMC (Monte Carlo Markov chain), the basis of which is the structure of the Markov process, the stationary distribution of which is determined by the a posteriori distribution function.

Keywords: point estimates, interval estimates, Bayesian regression, a priori distribution, a posteriori distribution

Байесовский подход нашёл широкое применение при оценивании эконометрических моделей по выборочным данным малого объема, а также в случаях, когда классические методы неприменимы. В рамках классического подхода, для оценки некоторого вектора параметров модели θ по выборочным данным Y, например, методом максимального правдоподобия (ММП), выбирается целевая функция – функция правдоподобия, и находится такая оценка

$$\hat{\theta}_{ML} = f(Y_1, Y_2, ..., Y_n), \qquad (1)$$

которая её максимизирует

$$P(\theta, Y) = \max_{\theta} P(Y_1, Y_2, ..., Y_n | \theta). \qquad (2)$$

Вектор параметров θ – неслучаен, а ММП-оценка (1), вычисляемая по выборочным данным – случайна. Байесовский метод – это способ формализации *степени разумной уверенности* в некотором утверждении, и ее корректировки по мере поступления информации относительно исследуемого явления. Поэтому в байесовском подходе оцениваемый вектор параметров θ трактуется как случайный с заданным в явном виде априорным распределением $P(\theta)$. Выбор априорного распределения отражает степень незнания исследователя о неизвестных параметрах до проведения и обработки наблюдений, и задача байесовского оценивания заключается в поиске апостериорно-

го распределения, скорректированного по результатам наблюдений:

$$P(\theta\Big|Y) = \frac{P(\theta, Y)}{P(Y)} = \frac{P(\theta)\cdot P(Y|\theta)}{P(Y)}, \quad (3)$$

где $P(\theta)$ – плотность априорного распределения, $P(Y|\theta)$ – функция правдоподобия. Поскольку $P(Y)$ выполняет роль нормирующего множителя, и не зависит от вектора параметров, то (3) можно записать в виде

$$P(\theta|Y) \sim P(\theta)\cdot P(Y|\theta), \quad (4)$$

где символ ~ означает пропорциональность левой и правой частей выражения (4) с точностью до нормирующей константы. Имея выборочные данные и вычислив функцию правдоподобия, можно найти условный закон распределения при данной выборке, по которому рассчитать точечные и интервальные оценки эконометрической модели.

Цель исследования: сравнительный анализ оценок параметров эконометрической модели в рамках байесовской регрессии и метода максимального правдоподобия. В качестве инструментальных средств выбрана программная среда R, в которой байесовская парадигма представлена в функциях многих статистических пакетов. Результаты, приведенные в статье, получены при помощи пакета MCMC (Monte Carlo Markov chain), основу которого составляет построение марковского процесса, стационарное распределение которого определяется апостериорной функцией распределения [1, 2].

Результаты исследования и их обсуждение

Алгоритм байесовского оценивания имеет следующую последовательность [3]:

1) выбор априорного распределения $P(\theta)$ параметра θ;

2) сбор статистических данных: Y_1, Y_2,..., Y_n (случайная выборка из анализируемой генеральной совокупности);

3) вычисление функции правдоподобия, в предположении статистической независимости наблюдений:

$$P(Y_1, Y_2, ..., Y_n|\theta) = P(Y_1|\theta)\cdot P(Y_2|\theta)\cdot \cdots \cdot P(Y_n|\theta)\ ; \quad (5)$$

4) вычисление апостериорного распределения параметра θ: $P(\theta|Y_1, Y_2, ..., Y_n)$ по формуле (4);

5) заключение о значении параметра θ: точечная или интервальная оценка.

Под байесовской точечной оценкой параметра понимается математическое ожидание или мода случайной величины, имеющей апостериорное распределение (4), например, для непрерывного случая:

$$\hat{\theta}_{ср} = E\left(\theta|Y_1, Y_2, ..., Y_n\right) = \int \theta\cdot p\left(\theta|Y_1, Y_2, ..., Y_n\right)d\theta, \quad (6)$$

$$\hat{\theta}_{мод} = \arg\max_{\theta} p\left(\theta|Y_1, Y_2, ..., Y_n\right). \quad (7)$$

Интервальные оценки параметров так же определяются через функцию апостериорного закона распределения вектора параметров (*Highest Posterior Density*, *HPD* – интервал высокой апостериорной плотности) [4].

При практической реализации байесовского подхода, в частности выбора априорного распределения, существенную роль играют распределения, сопряжённые с функцией правдоподобия. В этом случае общий вид априорного закона распределения известен, нужно только «уточнить» его параметры при переходе к апостериорному. Сопряженное семейство априорных распределений существует, если функцию правдоподобия можно представить в виде произведения достаточных статистик:

$$P(Y|\theta) = v\left(T(Y);\theta\right)\cdot u(Y), \quad (8)$$

где $v\left(T(Y);\theta\right)$ – неотрицательная функция, зависящая от Y только через $T(Y)$, $u(Y)$ – положительная функция от выборочных данных, независящая от параметров [5].

В теории байесовского подхода доказывается, что если априорное распределение генеральной совокупности имеет функции сопряжённые с функцией распределения, то уже первый переход от априорного к апостериорному распределению по формуле (4) приводит к семейству распределений, сопряжённому с наблюдаемой генеральной совокупностью, даже если априорное распределение не несёт никакой информации об оцениваемых параметрах (САЗ – скудность априорных знаний [3], априорные распределения). Это позволяет упростить процедуру выбора априорного распределения для оцениваемого параметра:

$$p_{САЗ}\left(\theta\right) = \text{const} \quad (9)$$

– для параметра, принимающего значения на конечном $\left[\theta_{\min}, \theta_{\max}\right]$ или бесконечном $\left(-\infty; +\infty\right)$ интервалах;

$$p_{САЗ}(\theta) \sim 1/\theta \quad (10)$$

– для параметра, принимающего любые положительные значения, и в качестве априорных распределений неизвестных параметров рекомендуется использовать равномерные распределения.

Выбор семейства априорных распределений, сопряженных с наблюдаемой генеральной совокупностью, осуществляется в результате следующих шагов:

1) выполняется проверка условия существования семейства априорных распределений (8), сопряженных с функцией правдоподобия для наблюдаемой генеральной совокупности;

2) выполняется вывод САЗ-апостериорного распределения, которое и определяет общий вид семейства априорных распределений, сопряженных с наблюдаемой генеральной совокупностью:

$$P_{\text{САЗ}}\left(\theta\middle|Y\right) \sim P_{\text{САЗ}}\left(\theta\right)\cdot P\left(Y\middle|\theta\right). \qquad (11)$$

Оценим параметры линейной регрессионной модели зависимости среднедушевых сбережений Y от доходов X у одинаковых по численному составу домохозяйств (по данным таблицы) в рамках ММП и байесовского подхода.

Выборочные данные переменных модели (в условных единицах) [6]

№	Y	X	№	Y	X
1	0,6	15,6	9	9,3	116
2	0,2	20	10	15	123,2
3	2	28,8	11	18,6	156
4	1,6	40	12	15	174
5	4,4	53,2	13	15,9	200,8
6	5	72	14	26,4	219,6
7	4	77,6	15	27,6	244
8	7,6	89,2	16	27,6	244

Спецификация оцениваемой модели $Y_t = \beta_1 + \beta_2 \cdot X_t + \varepsilon_t$, $t = 1,..., n$, или в матричном виде

$$Y = X\beta + \varepsilon, \qquad (12)$$

где $Y = (Y_1,...,Y_t,...,Y_n)'$ – вектор-столбец значений эндогенной переменной, X – детерминированная $(n\times k)$-матрица регрессоров, $\varepsilon = (\varepsilon_1,...,\varepsilon_t,...,\varepsilon_n)'$ – вектор-столбец возмущений, $\beta = (\beta_1,\beta_2,...,\beta_k)'$ – вектор-столбец параметров модели, n – объем выборки, k – число параметров, t – номер наблюдения, $\varepsilon \sim N\left(0,\sigma^2 I_n\right)$, σ^2 – дисперсия случайного возмущения, $h = \left(Var\left(\varepsilon_t\right)\right)^{-1}$ – параметр точности (*precision metrics*), $Y \sim N\left(X\beta,\sigma^2 I_n\right)$ – вектор эндогенной переменной, плотность (5) которого представляет собой априорную функцию распределения:

$$f(Y) = \left(2\pi\right)^{-n/2} \times$$
$$\times h^{n/2} \exp\left(-\frac{h}{2}\left(Y - X\beta\right)^T \left(Y - X\beta\right)\right). \qquad (13)$$

ММП-оценки параметров при регрессорах совпадают с МНК-оценками, поэтому оценим их в программной среде R при помощи функции lm пакета lmtest:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.592462   1.088620  -2.381    0.032 *
X            0.118612   0.007747  15.310    3.88e-10 ***
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 2.406 on 14 degrees of freedom
Multiple R-squared:  0.9436,    Adjusted R-squared:  0.9396
F-statistic: 234.4 on 1 and 14 DF,  p-value: 3.88e-10
confint(fm,level=0.90)
# интервальная оценка параметров
                5 %       95 %
(Intercept) -4.5098597 -0.6750641
X            0.1049671  0.1322574
# автоковариационная матрица вектора оценок параметров
            (Intercept)            X
(Intercept)  1.185093948 -7.029648e-03
X           -0.007029648  6.001834e-05
```

Для оценки параметров в рамках байесовского подхода необходимо проверить выполнение условия (8) существования семейства сопряженного априорного распределения $p(\beta; h)$. Достаточной статистикой функции (13) является функции $T(Y, X)$, которая определяется произведениями $Y'Y$, $X'Y$, $X'X$, что означает, что существует априорное распределение неизвестных параметров β и σ^2, сопряжённое с функцией правдоподобия. Преобразуем отклонение $Y - X\beta$ в формуле (13) следующим образом:

$$\left(Y - X\hat{\beta} + X\hat{\beta} - X\beta\right) = \left(\left(Y - X\hat{\beta}\right) + X\left(\hat{\beta} - \beta\right)\right),$$

тогда аргумент функции $\exp(\cdot)$ принимает вид

$$\left(Y - X\hat{\beta}\right)' \left(Y - X\hat{\beta}\right) + \left(\hat{\beta} - \beta\right)' X'X\left(\hat{\beta} - \beta\right). \qquad (14)$$

Выражая первое слагаемое в (14) через несмещенную оценку дисперсии возмущений, $\left(Y - X\hat{\beta}\right)' \left(Y - X\hat{\beta}\right) = (n-k)\hat{\sigma}^2$, и подставляя (14) в (13), получим

$$f(Y) = (2\pi)^{-n/2} \cdot h^{n/2} \cdot \exp\left(-\frac{h\left((n-k)\hat{\sigma}^2\right)}{2} - \frac{h}{2}\left(\hat{\beta}-\beta\right)' X'X\left(\hat{\beta}-\beta\right)\right). \quad (15)$$

Определим САЗ-апостериорное распределение для параметров модели множественной регрессии. Так как параметр точности h принимает положительные значения, то (11), с учетом правила (10) и формулы (15), принимает вид

$$P_{\text{САЗ}}\left(\beta;h\middle|X;Y\right) \sim P_{\text{САЗ}}(\beta;h) \cdot P_{\text{САЗ}}(X;Y|\beta;h) =$$

$$= \frac{1}{h} \cdot \left[h^{n/2} \cdot \exp\left(-\frac{h\left((n-k)\hat{\sigma}^2\right)}{2}\right) \cdot \exp\left(-\frac{h}{2}\left(\hat{\beta}-\beta\right)' X'X\left(\hat{\beta}-\beta\right)\right)\right] =$$

$$= \left[h^{(n-k)/2-1} \cdot \exp\left(-\frac{h\left((n-k)\hat{\sigma}^2\right)}{2}\right) \cdot h^{k/2} \cdot \exp\left(-\frac{h}{2}\left(\hat{\beta}-\beta\right)' X'X\left(\hat{\beta}-\beta\right)\right)\right] =$$

$$= \left[h^{k/2} \cdot \exp\left(-\frac{h}{2}\left(\hat{\beta}-\beta\right)' X'X\left(\hat{\beta}-\beta\right)\right) \cdot h^{(n-k)/2-1} \cdot \exp\left(-\frac{h\left((n-k)\hat{\sigma}^2\right)}{2}\right)\right]. \quad (16)$$

Перепишем (16) вводя обозначения: $\alpha = (n-k)/2$, $\theta = (n-k)\hat{\sigma}^2/2$,

$$p(\beta;h) \sim h^{k/2} \cdot \left|\Lambda_0\right|^{1/2} \cdot \exp\left(-\frac{h}{2}\left(\hat{\beta}-\beta\right)' \Lambda_0\left(\hat{\beta}-\beta\right)\right) \cdot h^{\alpha-1} \cdot \exp(-h\theta). \quad (17)$$

Распределение (17) представляет собой (с точностью нормирующего множителя, независящего от параметров) многомерное гамма-нормальное распределение с параметром сдвига $\hat{\beta}$, матрицей точности $X'X$ и параметрами α и θ.

При реализации байесовского подхода необходимо знать параметры сопряженного с наблюдаемой генеральной совокупностью априорного распределения. В большинстве случаев они определяются при помощи метода моментов по оценкам их математического ожидания и среднеквадратическим ошибкам. Для этой цели обычно используется любая априорная информация, например экспертное оценивание. Воспользуемся ММП-оцениванием. Так как частное распределение параметра точности h нормальной части распределения (17) имеет гамма-распределение с параметрами α и θ, его числовые характеристики определяются по формулам

$$\left.\begin{aligned} E(h) = \alpha/\theta = h_0 = 1/\hat{\sigma}^2 = 0{,}173 \\ Var(h) = \alpha/\theta^2 = \Delta_h^2 = 2h_0^2/(n-1) = 0{,}004 \end{aligned}\right\}. \quad (18)$$

Выражая из (18) параметры распределения через числовые характеристики параметра точности, получаем

$$\alpha = h_0^2/\Delta_h^2 = 7{,}5, \quad \theta = h_0/\Delta_h^2 = 43{,}41. \quad (19)$$

Частное распределение параметра β есть обобщённое $(k+1)$-мерное распределение Стьюдента с 2α числом степеней свободы, параметром сдвига β и матрицей точности $\Lambda_0 = (\alpha/\theta)\Delta$, поэтому его числовые характеристики определяются по формулам

$$E(\beta) = \beta_0 = \begin{pmatrix} -2{,}592 \\ 0{,}119 \end{pmatrix}, \quad \Delta = \begin{pmatrix} 1{,}09^2 & 0 \\ 0 & 0{,}008^2 \end{pmatrix},$$

$$\Lambda_0 = \frac{2\alpha}{2\alpha - 2}\left(\frac{\alpha}{\theta}\Delta\right)^{-1} = \frac{\theta}{\alpha - 1}\cdot\Delta = \frac{\theta}{\alpha - 1}\begin{pmatrix} s_{\beta_1}^2 & 0 \\ 0 & s_{\beta_2}^2 \end{pmatrix}^{-1} = \begin{pmatrix} 5,635 & 0 \\ 0 & 111272,6 \end{pmatrix}, \quad (20)$$

где s_j^2 – заданные значения априорных дисперсий элементов вектора параметров β.

По параметрам априорного распределения (19) и (20), выборочным данным (Y, X), вычисляются точечные оценки параметров апостериорного распределения (17):

$$\tilde{\beta}_0 = E\left(\beta \middle| X, Y\right) = \left(\tilde{\Lambda}_0\right)^{-1}\left(\Lambda\cdot\beta_0 + X'Y\right) = \left(-2,592 \quad 0,119\right)', \quad (21)$$

где

$$\tilde{\Lambda}_0 = \left(\Lambda + X'X\right) = \begin{pmatrix} 21,635 & 1874 \\ 1874 & 427201,1 \end{pmatrix} \quad (22)$$

– матрица точности; и параметры частного апостериорного гамма-распределения параметра точности h:

$$\tilde{\alpha} = \alpha + n/2 = 7,5 + 8 = 15,5, \quad (23)$$

$$\tilde{\theta} = \theta + \frac{1}{2}\left[\left(Y - X\tilde{\beta}_0\right)' Y + \left(\beta_0 - \tilde{\beta}_0\right)' C_\beta\cdot\beta_0\right] = 83,925, \quad (24)$$

параметр точности:

$$\tilde{h} = E\left(h \middle| X, Y\right) = \tilde{\alpha}/\tilde{\theta} = 0,1847. \quad (25)$$

При построении интервальных оценок, в рамках байесовского подхода, используется блочная структура матрицы точности:

$$\Lambda = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} = \tilde{h}\cdot\tilde{\Lambda}_0 = \begin{pmatrix} 3,996 & 346,106 \\ 346,106 & 78899,05 \end{pmatrix},$$

$$\Lambda(1) = \Lambda_{11} - \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21} = 3,996 - 346,106^2 / 78899,05 = 2,478,$$

$$\Lambda(2) = \Lambda_{22} - \Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12} = 78899,05 - 346,106^2 / 3,996 = 48920,2 \quad (26)$$

и интервальные оценки параметров модели определяются по формулам:

$$\left.\begin{aligned} \beta_1^- &= \beta_1 - \sqrt{\Lambda(1)}\cdot t_\alpha = -2,592 - \sqrt{2,478}\cdot 1,753 = -3,71 \\ \beta_1^+ &= \beta_1 + \sqrt{\Lambda(1)}\cdot t_\alpha = -2,592 + \sqrt{2,478}\cdot 1,753 = -1,48 \end{aligned}\right\}, \quad (27)$$

$$\left.\begin{aligned} \beta_2^- &= \beta_2 - \sqrt{\Lambda(2)}\cdot t_\alpha = 0,119 - \sqrt{48920,2}\cdot 1,753 = 0,111 \\ \beta_2^+ &= \beta_2 + \sqrt{\Lambda(2)}\cdot t_\alpha = 0,119 + \sqrt{48920,2}\cdot 1,753 = 0,127 \end{aligned}\right\}, \quad (28)$$

$$\left.\begin{aligned} h^- &= \chi^2_{0,05}(2\tilde{\alpha})/2\tilde{\theta} = 0,115 \\ h^+ &= \chi^2_{0,95}(2\tilde{\alpha})/2\tilde{\theta} = 0,268 \end{aligned}\right\}, \quad (29)$$

с учетом $t_{0,05}(2\cdot 7,5) = 1,753$, $\chi^2_{0,05}(2\tilde{\alpha}) = \chi^2_{0,05}(2\cdot 15,5) = 19,281$, $\chi^2_{0,95}(2\tilde{\alpha}) = 44,985$.

Получим точечные и интервальные оценки параметров регрессионной модели зависимости среднедушевых сбережений Y от доходов X, по данным таблицы, при помощи функции *MCMCregress* пакета *MCMCpack*:

```
Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000
1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:
             Mean      SD  Naive SE          Time-series SE
(Intercept) -2.5893 0.263544 2.635e-03      2.602e-03
   X         0.1186 0.001891 1.891e-05      1.891e-05
sigma2       5.8883 2.491154 2.491e-02      2.491e-02
2. Quantiles for each variable:
              2.5%     25%     50%     75%   97.5%
(Intercept) -3.1076 -2.7657 -2.5873 -2.4127 -2.0698
   X         0.1148  0.1173  0.1186  0.1199  0.1224
 sigma2      2.8866  4.2309  5.3662  6.9210 12.0544
```

Заключение

Как следует из сравнительного анализа результатов оценивания, интервальные оценки параметров регрессионной модели, полученные в рамках байесовского подхода, при непосредственном вычислении по формулам (18)–(29) уже, по сравнению с ММП-оценками: для параметра β_1 – в 1,72 раза, для параметра β_2 – в 1,75 раз, и при вычислении при помощи функции *MCMCregress* в программной среде R – для параметра β_1 – в 3,7 раза, для параметра β_2 – в 4 раза.

Список литературы

1. Martin A.D., Quinn K.M., Park J.H. MCMCpack: Markov Chain Monte Carlo in R. Journal of Statistical Software. 2011. Vol. 42. Issue 9. P. 1–21.

2. Fornalski K.W. Applications of the robust Bayesian regression analysis. International Journal of Society Systems Science. 2015. Vol. 7. no. 4. P. 314–333.

3. Айвазян С.А., Фантаццини Д. Эконометрика-2: Продвинутый курс с приложениями в финансах: учебник. М.: Магистр: Инфра-Ь, 2014. 944 с.

4. Шитиков В.К., Розенберг Г.С. Рандомизация и бутстреп: статистический анализ в биологии и экологии с использованием R. Тольятти: «Кассандра», 2013. 305 с.

5. Де Гроот М. Оптимальные статистические решения. М.: МИР, 1974. 491 с.

6. Бабешко Л.О., Бич М.Г., Орлова И.В. Эконометрика и эконометрическое моделирование: учебник. М.: Вузовский учебник: ИНФРА-М, 2017. 385 с.