УДК 303.724.32

В. П. Козлов, канд. техн. наук, А. А. Мартыненко, О.С. Шевцова
(Украина, Днепр, Государственное высшее учебное заведение "Национальный горный университет")

НЕКОТОРЫЕ АСПЕКТЫ БАЙЕСОВСКОГО ПОДХОДА В ЛИНЕЙНОМ РЕГРЕССИОННОМ АНАЛИЗЕ

Анотація. *Показано, що байєсовські методи є засобами аналізу даних, які при малих обсягах вибірки дозволяють оцінити регресійні моделі повніше і точніше в порівнянні з класичними статистичними методами.*

Ключові слова: *байєсівський підхід, регресійна модель, апріорні параметри, метод Монте-Карло, багатовимірний нормальний розподіл.*

Аннотация*. Показано, что байесовские методы являются средствами анализа данных, которые при малых объемах выборки позволяют оценить регрессионные модели полнее и точнее по сравнению с классическими статистическими методами.*

Ключевые слова: *байесовский подход, регрессионная модель, априорные параметры, метод Монте-Карло, многомерное нормальное распределение.*

Abstract. *It is shown that Bayesian methods are data analysis tools, which for small sample sizes allow us to evaluate regression models more fully and more accurately in comparison with classical statistical methods.*

Keywords: *Bayesian approach, regression model, a priori parameters, Monte Carlo method, multidimensional normal distribution.*

Введение. Байесовские методы являются средствами анализа данных, которые вытекают из принципов байесовского статистического вывода [1, 2]. Классические статистические методы направлены на получение эффективных алгоритмов оценивания на основе данных относительно большого объема. В случае небольших выборок использование результатов асимптотической теории является необоснованным [1, 2]. В настоящее время актуальной задачей оценивания статистических моделей (в том числе регрессионных) является применение байесовской методологии, которая позволяет полнее оценивать модели и получать достаточно точные результаты в тех случаях, когда применение классического статистического подхода ограничено. Байесовская методология исследовалась во многих роботах и используется в разных областях науки и техники. В частности, А. Зельнер исследовал использование таких методов в эконометрике [2, 3].

Постановка задачи. Показать преимущества байесовского подхода с точки зрения полноты и точности полученных результатов на примере линейной регрессионной модели прогнозирования полезного отпуска тепловой энергии.

Основное содержание работы. Наиболее распространенный способ краткосрочного прогнозирования в экономике заключается в использовании линейных регрессионных моделей. Например, запасы топлива на предстоящий отопительный период определяются на основе предполагаемого отпуска тепловой энергии, который зависит от температуры атмосферы. Для расчета коэффициентов линейных регрессионных моделей целесообразно использовать данные только за прошедший отопительный период, так как использование данных за более ранний период приведет к снижению точности прогноза. Пусть полезный отпуск теплоэнергии за прошедший отопительный период приведен в табл. 1. Эти данные можно также представить в виде зависимости от среднемесячной атмосферной температуры (рис. 1).

В общем виде регрессионная модель выглядит следующим образом:

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \epsilon_i, \quad (1)$$

где $x_{i,1} = 1$ для каждого объекта наблюдения i (отпуска тепла Y_i, соответствующего среднемесячной температуре атмосферы t_i);

$x_{i,2} = 0$, если объект наблюдения i относится к группе выработки тепла с помощью ТЭЦ ($group_i = 0$), и равно 1, если – к группе выработки тепла с помощью ТЦ ($group_i = 1$);
$x_{i,3} = t_i$ (среднемесячная температура атмосферы, соответствующая объекту наблюдения i);
$x_{i,4} = x_{i,2} \times x_{i,3}$.

Для этой модели условные матожидания Y для двух разных технологий выработки тепла следующие:

$$E[Y|\mathbf{x}] = \beta_1 + \beta_3 \times t_i, \text{ если } x_2 = 0, \text{ и} \quad (2)$$

$$E[Y|\mathbf{x}] = (\beta_1 + \beta_2) + (\beta_3 + \beta_4) \times t_i, \text{если } x_2 = 1. \quad (3)$$

Другими словами, модель предполагает, что выработка тепла линейна по t_i для обеих групп, с разницей в точках пересечения с осью ординат, равной β_2, и с разницей в наклонах, равной β_4.

Таблица 1

Полезный отпуск теплоэнергии за прошедший отопительный период

Месяц отопительного периода	**Октябрь**	**Ноябрь**	**Декабрь**	**Январь**	**Февраль**	**Март**
Среднемесячная температура атмосферы, °С	8	2	-2	-4	-3	2
Отпуск теплоэнергии теплоэлектроцентралями (ТЭЦ), тыс. Гкал	1972	4758	5533	6975	7486	6493
Отпуск теплоэнергии теп-лоцентралями (ТЦ), тыс. Гкал	965	9586	11146	14051	15082	5557

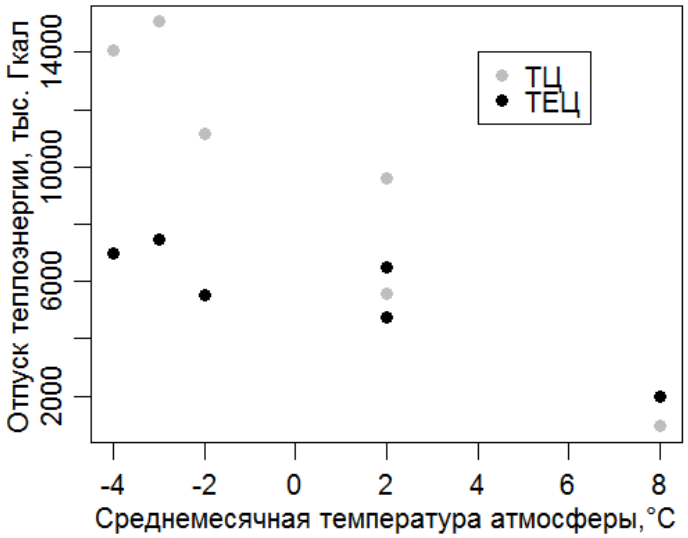

Рис. 1. Изменение полезного отпуска теплоэнергии за прошедший отопительный период

Если предположить, что $\beta_2 = \beta_4 = 0$, то линии будут идентичны для обеих групп. Если принять $\beta_4 = 0$, то получим две параллельные линии для каждой группы. При отличии всех коэффициентов от нуля получаем две несвязанные линии. Указанные регрессионные модели изображены с помощью графических средств среды R [4] на рис. 2. Приведенные на данном рисунке результаты получены на основе традиционного подхода с использованием метода наименьших квадратов (МНК) [5]. Последующие результаты, приведенные в данной работе, получены с помощью байесовского подхода.

Байесовский анализ регрессионной модели требует определения априорных параметров. Определение величин этих параметров, которые представляют фактическую априорную информацию, может оказаться сложным. Эффективный принцип построения априорного распределения для $\boldsymbol{\beta}$ основан на том, что оценка параметра должна быть инвариантной к изменениям масштаба регрессоров [1]. Данная идея выбора априорных параметров привела к варианту использования широко изученного g-prior распределения для регрессионных параметров [3]. Указанное априорное распределение для $\boldsymbol{\beta}$ при $(\mathbf{y}, \mathbf{X}, \sigma^2)$ является многомерным нормальным со средним значением и дисперсией

$$\mathrm{E}[\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2] = [\mathbf{X}^{\mathrm{T}}\mathbf{X}/(\mathrm{g}\sigma^2) + \mathbf{X}^{\mathrm{T}}\mathbf{X}/\sigma^2]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}/\sigma^2 = \frac{\mathrm{g}}{\mathrm{g}+1}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}; \quad (4)$$

$$\mathrm{Var}[\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2] = [\mathbf{X}^{\mathrm{T}}\mathbf{X}/(\mathrm{g}\sigma^2) + \mathbf{X}^{\mathrm{T}}\mathbf{X}/\sigma^2]^{-1} = \frac{\mathrm{g}}{\mathrm{g}+1}\sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}. \quad (5)$$

Для нашего примера

$$\mathbf{X}^{\mathrm{T}} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 8 & 2 & -2 & -4 & -3 & 2 & 8 & 2 & -2 & -4 & -3 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 8 & 2 & -2 & -4 & -3 & 2 \end{bmatrix},$$

$$\mathbf{y} = (1972, 4758, 5533, 6975, 7486, 6493,\ 965, 9586, 11146, 14051, 15082,\ 5557).$$

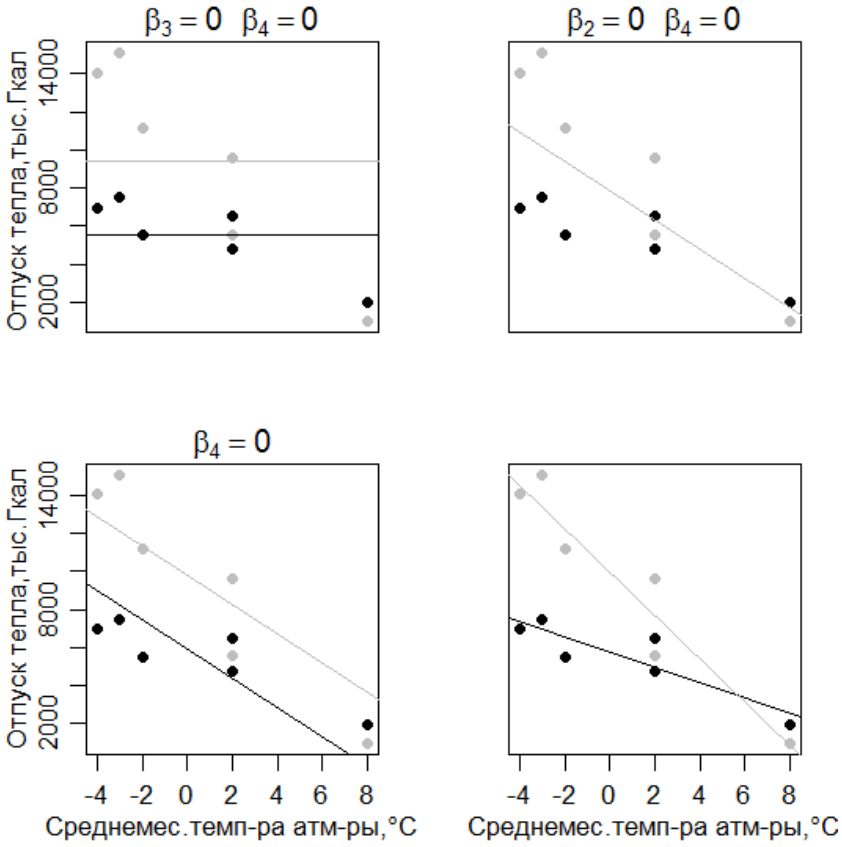

Рис. 2. Линии регрессии наименьших квадратов, полученные на основании данных о выработке тепла для четырех различных моделей

При этом априорном распределении, наряду с уравнениями (4) и (5), плотности $\mathrm{p}(\sigma^2|\mathbf{y},\mathbf{X})$ и $\mathrm{p}(\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2)$ являются обратным гамма и многомерным нормальным распределениями соответственно [1, 3]. Так как мы можем семплировать из обоих этих распределений, то значения семплов $(\sigma^2,\boldsymbol{\beta})$ из совместного апостериорного распределения $\mathrm{p}(\sigma^2,\boldsymbol{\beta}|\mathbf{y},\mathbf{X})$ могут быть получены с помощью метода Монте-Карло [6] для марковских цепей (Markov chain Monte Carlo, MCMC) следующим образом:

1. Семплировать $1/\sigma^2$ по условному гамма-распределению $\Gamma([\mathrm{v}_0+\mathrm{n}]/2, [\mathrm{v}_0\sigma_0^2 + \mathrm{SSR}_\mathrm{g}]/2)$;
2. Семплировать $\boldsymbol{\beta}$ по условному многомерному нормальному распределению $\mathrm{N}(\frac{\mathrm{g}}{\mathrm{g}+1}\widehat{\boldsymbol{\beta}}_{\text{мнк}}, \frac{\mathrm{g}}{\mathrm{g}+1}\sigma^2[\mathbf{X}^{\mathrm{T}}\mathbf{X}]^{-1})$.

Здесь $\mathrm{SSR}_\mathrm{g} = \mathbf{y}^{\mathrm{T}}(\mathbf{I} - \frac{\mathrm{g}}{\mathrm{g}+1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{y}$; $\widehat{\boldsymbol{\beta}}_{\text{мнк}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = (5736.1, 4234.2, -399.4, -745.6)$.

Для нашего примера $\mathrm{g} = \mathrm{n} = 12, \mathrm{v}_0 = 1$ и $\sigma_0 = 1443.8$ Апостериорное среднее $\boldsymbol{\beta}$ может быть получено непосредственно из уравнения (4). Так как $\mathrm{E}[\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2]$ не зависит от σ^2, то мы имеем $\mathrm{E}[\boldsymbol{\beta}|\mathbf{y},\mathbf{X}] = \mathrm{E}[\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2] = \frac{\mathrm{g}}{\mathrm{g}+1}\widehat{\boldsymbol{\beta}}_{\text{мнк}}$. Поэтому апостериорные средние четырех регрессионных параметров равны

(5185.1, 4016.9, -362.5, -689.2). Апостериорные стандартные отклонения этих параметров $-(1142.,1593.,278.,406.)$. Совместное апостериорное распределение для β_2 и β_4 изображено на рис. 3. Здесь вероятности попадания в области, ограниченные замкнутыми кривыми 1, … , 5, равны 0.975, 0.75, 0.5, 0.25, 0.025 соответственно. Расчеты проведены в среде R на основе 50000 семплов Монте-Карло.

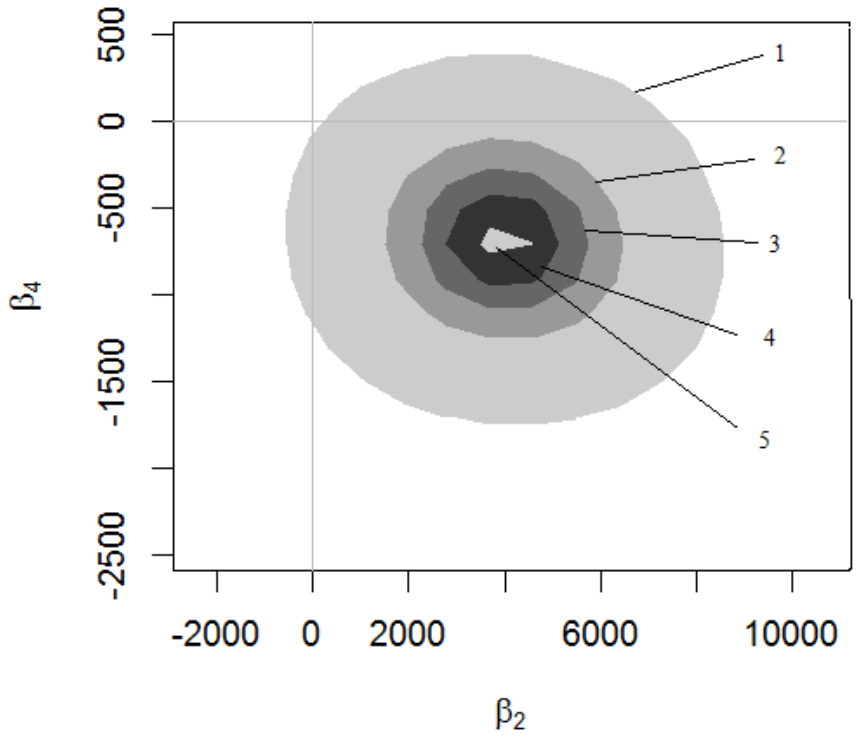

Рис. 3. Совместное апостериорное распределение для β_2 и β_4

Важным направлением регрессионного моделирования является решение того, какие объясняющие переменные включать в модель. Очевидно, что в регрессионную модель необходимо включать только те переменные, для которых имеются существенные доказательства их ассоциации с y. Это приводит к получению более простых моделей с лучшими статистическими свойствами.

Концептуально байесовское решение проблемы выбора модели следующее [1]. Если мы полагаем, что многие регрессионные коэффициенты потенциально равняются нулю, то нужно ввести в модель априорное распределение, которое отражает эту возможность. Это может быть достигнуто определением того, что у каждого регрессионного коэффициента есть некоторая отличная от нуля вероятность того, что он может быть равен нулю. Удобный способ представить это заключается в том, чтобы записать регрессионный коэффициент для переменной j как $\beta_j = z_j \times b_j$, где $z_j \in \{0,1\}$, а b_j являются некоторыми действительными числами. С такой параметризацией наше регрессионное уравнение примет вид

$$y_i = z_1 b_1 x_{i,1} + \cdots + z_p b_p x_{i,p} + \epsilon_i \, .$$

Таким образом, z_j указывают на то, какие регрессионные коэффициенты отличны от нуля. В случае нашего примера

$$E[Y|\mathbf{x}, \mathbf{b}, \mathbf{z} = (1,0,1,0)] = b_1 x_1 + b_3 x_3 = b_1 + b_3 \times t_i;$$
$$E[Y|\mathbf{x}, \mathbf{b}, \mathbf{z} = (1,1,0,0)] = b_1 x_1 + b_2 x_2 = b_1 + b_2 \times \text{group};$$
$$E[Y|\mathbf{x}, \mathbf{b}, \mathbf{z} = (1,1,1,0)] = b_1 x_1 + b_2 x_2 + b_3 x_3 = b_1 + b_2 \times \text{group} + b_3 \times t_i.$$

Каждая величина $\mathbf{z} = (z_1, \ldots, z_p)$ соответствует различной модели или, более определенно, различному набору переменных, имеющих регрессионные коэффициенты, отличные от нуля. Например, считаем, что модель с $\mathbf{z} = (1,0,1,0)$ является линейной регрессионной моделью для y как функции температуры. Модель с $\mathbf{z} = (1,1,1,0)$ является регрессионной моделью для y как функции температуры, но с определенной для группы точкой пересечения с осью ординат. С такой параметризацией выбор того, какую переменную включать в регрессионную модель, эквивалентен выбору, какая z_i равна 0, а какая равняется 1.

Выбор байесовской модели заключается в получении апостериорного распределения для $\mathbf{z}$. Конечно, для этого необходимо знать совместное априорное распределение для $\{\mathbf{z}, \boldsymbol{\beta}, \sigma^2\}$. Здесь необходимо использовать указанное выше g-prior распределение, которое позволяет оценить $\mathrm{p}(\mathbf{y}|\mathbf{X}, \mathbf{z})$ для каждой возможной модели z по формуле

$$\mathrm{p}(\mathbf{y}|\mathbf{X}, \mathbf{z}) = \iint \mathrm{p}(\mathbf{y}, \boldsymbol{\beta}, \sigma^2|\mathbf{X}, \mathbf{z})\mathrm{d}\boldsymbol{\beta}\mathrm{d}\,\sigma^2. \quad (6)$$

Учитывая априорное распределение $\mathrm{p}(\mathbf{z})$ по моделям, можно вычислять апостериорную вероятность для каждой регрессионной модели:

$$\mathrm{p}(\mathbf{z}|\mathbf{y}, \mathbf{X}) = \frac{\mathrm{p}(\mathbf{z})\mathrm{p}(\mathbf{y}|\mathbf{X}, \mathbf{z})}{\Sigma_{\tilde{\mathbf{z}}}\mathrm{p}(\tilde{\mathbf{z}})\mathrm{p}(\mathbf{y}|\mathbf{X}, \tilde{\mathbf{z}})}. \quad (7)$$

Таким образом, в нашем примере можно оценить должны ли β_2 или β_4 равняться нулю, вычисляя вероятность $\mathrm{p}(\mathbf{z}|\mathbf{y}, \mathbf{X})$ для множества конкурирующих моделей. В табл. 2 приведен список из четырех указанных на рис. 2 различных регрессионных моделей, для которых мы хотим рассмотреть эти апостериорные распределения.

Таблица 2

Апостериорные распределения $\mathrm{p}(\mathbf{z}|\mathbf{y}, \mathbf{X})$ для четырех различных моделей

z	**Модель**	$\mathrm{p}(\mathbf{z}|\mathbf{y}, \mathbf{X})$
(1,1,0,0)	$\beta_1 + \beta_2 \times \mathrm{group_i}$	0.00
(1,0,1,0)	$\beta_1 + \beta_3 \times \mathrm{t_i}$	0.11
(1,1,1,0)	$\beta_1 + \beta_2 \times \mathrm{group_i} + \beta_3 \times \mathrm{t_i}$	0.36
(1,1,1,1)	$\beta_1 + \beta_2 \times \mathrm{group_i} + \beta_3 \times \mathrm{t_i} + \beta_4 \times \mathrm{group_i} \times \mathrm{t_i}$	0.53

Апостериорные вероятности определялись для каждой модели в соответствии с выражением (7). Здесь значения $\mathrm{p}(\mathbf{y}|\mathbf{X}, \mathbf{z})$ вычислены для каждой из четырех рассматриваемых величин $\mathbf{z}$ с использованием g-prior распределения для $\boldsymbol{\beta}$. При этом принято априорное распределение для моделей $\mathrm{p}(\mathbf{z}) = (0.25, 0.25, 0.25, 0.25)$. Приведенные в табл. 2 результаты вычислений показывают, что наибольшую вероятность имеет модель, соответствующая $\mathbf{z} = (1,1,1,1)$. Необходимость включения в модель в качестве независимой переменной температуры атмосферы следует из того, что апостериорные вероятности трех моделей, учитывающих температуру атмосферы, в сумме равны едини-це. Доказательства значимости независимой переменной group не такие убедительные, так как суммарная вероятность для трех моделей с переменной group равна 0.00 + 0.36 + 0.53 = 0.89. Однако эта вероятность существенно выше, чем соответствующая априорная вероятность 0.25 + 0.25 + 0.25 = 0.75 для этих трех моделей.

Выводы. Таким образом, байесовские методы являются средствами анализа данных, которые при малых объемах выборки позволяют оценить регрессионные модели полнее и точнее по сравнению с классическими статистическими методами.

Список литературы

1. Hoff P. D. A First Course in Bayesian Statistical Methods / Peter D. Hoff. – Springer New York, 2009. – 276 p.
2. Зельнер А. Байесовские методы в эконометрии / А. Зельнер. – М.: Статистика, 1980. – 434 с.
3. Zellner A. On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions / A. Zellner // Bayesian Inference and Decision Techniques : Essays in Honor of Bruno de Finetti. – New York, North Holland Publishing Co., 1986. – Vol. 6. – P. 233–243.
4. Adler J. R in a Nutshell / J. Adler. – USA: O'Reilly Media, 2012. – 697 p.
5. Дрейпер Н., Смит Г. Прикладной регрессионный анализ / Н. Дрейпер , Г. Смит . – 3-е изд. – М.: Диалектика, 2007. – 912 с.
6. Ермаков С.М. Метод Монте-Карло в вычислительной математике. Вводный курс / С.М. Ермаков. – СПб: Бином. Лаборатория знаний, 2009. –192 с.

Рекомендовано до друку д-ром техн. наук, проф. Куваєвим В.М.