УДК 519.2(075)

ПОСТРОЕНИЕ ЛИНЕЙНОЙ РЕГРЕССИОННОЙ МОДЕЛИ НА ОСНОВЕ БАЙЕСОВСКОГО ПОДХОДА

В.П. Козлов, А.Л. Ширин, В.С. Касьяненко
(Украина, ДВНЗ «Национальний горный университет», Днепр)

Постановка проблемы. Показать преимущества байесовского подхода с точки зрения точности полученных результатов на примере линейной регрессионной модели.

Байесовские методы являются средствами анализа данных, которые вытекают из принципов байесовского статистического вывода [1, 2]. В настоящее время имеет место значительный рост в развитии и применении байесовского подхода в статистике. Наблюдается увеличение байесовской активности по количеству опубликованных научных статей, книг, а также широкое применение этого подхода в различных областях науки и техники. Одной из причин резкого роста байесовского моделирования является появившаяся возможность вычисления интегралов, которые необходимы для байесовского апостериорного анализа. Вычислительные возможности современных компьютеров, позволяют использовать байесовскую парадигму для исследования очень сложных моделей, которые не могут быть проанализированы альтернативными частотными методами.

Для реализации байесовского подхода необходима статистическая вычислительная среда, которая должна позволять:

- писать скрипты для определения байесовской модели;
- использовать или писать функции, позволяющие получать апостериорные распределения;
- использовать функции для моделирования из апостериорного распределения;
- строить графики для иллюстрации апостериорного вывода.

Средой для статистических расчетов, которая отвечает этим требованиям, является система R [3, 4]. R предоставляет широкий спектр функций для манипулирования данными, расчетов и графических демонстраций. Кроме того, она включает в себя хорошо развитой, язык программирования, который пользователи могут расширить путем добавления новых функций. Многие из таких расширений языка в виде пакетов можно легко загружать с сайта CRAN.

Линейное регрессионное моделирование - чрезвычайно мощный инструмент анализа данных, полезный для множества задач логического вывода, таких как предсказание и оценка параметров.

Часто в регрессионном анализе мы сталкиваемся с большим количеством возможных регрессионных переменных, даже если мы предполагаем, что большинство регрессоров не имеют никакого отношения к зависимой переменной. В этом случае, включая все из возможных переменных в регрессионную модель, можно прийти к плохому статистическому результату.

Стандартный статистический совет состоит в том, что мы должны включать в регрессионную модель только те переменные, для которых есть существенные доказательства их ассоциации с зависимой переменной. Это приводит не только более простому анализу данных, но также обычно предоставляет моделям лучшие статистические свойства с точки зрения предсказания и оценки.

В материалах доклада представлена регрессионная модель, которая включает 8 объясняющих переменных и 28 переменных взаимодействия. Следовательно, в общей сложности число регрессоров велико и равняется 36. Важный аспект регрессионного моделирования – принятие решения о том, какие объясняющие переменные включать в модель. Эта проблема выбора переменных модели концептуально имеет следующее байесовское решение. Если мы полагаем, что многие регрессионные коэффициенты потенциально равняются нулю, тогда мы просто используем априорное распределение, которое отражает эту возможность. Это может быть достигнуто соглашением что у каждого регрессионного коэффициента есть некоторая отличная от нуля вероятность того, чтобы быть точно нулевым. Так как число регрессоров велико, пространство моделей исследовалось с помощью алгоритма семплирования по Гиббсу [1, 2].

Результаты тестирования показали, что среднеквадратическая ошибка прогнозирования при использовании байесовского подхода примерно в полтора раза меньше, чем для традиционной прогнозирующей модели, построенной на основе метода наименьших квадратов. Такой результат является характерным при сравнении байесовских методов с альтернативными небайесовскими.

Выводы. Байесовские методы являются средствами анализа данных, которые обеспечивают оценку параметров с хорошими статистическими свойствами.

ПЕРЕЧЕНЬ ССЫЛОК

1. Marin J., Robert C. P. Bayesian Core: A Practical Approach to Computational Bayesian Statistics. – USA: Springer Science+Business Media, 2007. –255p.
2. Rossi P. E., Allenby G. M., McCulloch R. Bayesian Statistics and Marketing. – England: John Wiley & Sons Ltd, 2005. – 364 p.
3. Adler J. R in a Nutshell. – USA: O'Reilly Media, 2012. – 697 p.
4. Pfaff B. Analysis of Integrated Series with R and Cointegrated. . – USA: Springer Science+Business Media, 2008. – 189p.