

ПОСТРОЕНИЕ БАЙЕСОВСКОЙ ГАММА-РЕГРЕССИИ
ДЛЯ ИССЛЕДОВАНИЯ РИСК-ФАКТОРОВ ПРИ ПРОГНОЗИРОВАНИИ ДИАБЕТА

В.О. Алексеев

Национальный исследовательский Томский политехнический университет,

Россия, г. Томск, пр. Ленина, 30, 634050

E-mail: valeks@tpu.ru

Известно, что значение уровня гликозилированного гемоглобина выше семи принимается за подтверждение диагноза диабета. Мы выбрали данную характеристику в качестве целевой для формирования регрессионной модели. Цель работы – проанализировать данные и обнаружить признаки возникновения диабета. Для создания обобщенной линейной модели использовано двухпараметрическое гамма-распределение

$$Y_i \sim \Gamma\left(v, \frac{v}{\exp(z_i^T w)}\right)_{i=1, 2, \dots, n} \quad (1)$$

где v – параметр формы с экспоненциальным распределением и $\lambda = 1/100$, z_i – это набор ковариаций для i -го элемента, w – вектор рассматриваемых регрессоров с нормальным распределением. Модель (1) представляет собой простую байесовскую линейную модель. Введем μ для обозначения ожидаемого значения распределения. В работе [1] показано, что при $\mu > 0$ каноническая функция связи $g(\mu) = -\mu^{-1}$ не всегда адекватна, поэтому мы использовали в качестве функции связи выражение

$$g(\mu) = \log \mu \quad (2)$$

Зададим результирующий признак η_i в качестве линейной комбинации ковариаций виде $\eta_i = z_i^T w$, тогда из функции связи (2) получаем $\mu_i = \exp(\eta_i)$. Для проведения моделирования уровня гликозилированного гемоглобина были использована выборка $n=403$ наблюдений по следующим показателям: уровень холестерина, стабилизированный уровень глюкозы, уровень липопротеина высокой плотности, местоположение, возраст, пол, рост, вес, фрейм (малый, средний, большой), первое систолическое и диастолическое артериальное давление, обхват талии, обхват бедер [2].

Для расчета была использована библиотека РунМС3, для генерации выборки мы использовали два метода Монте-Карло и Метрополиса. Для сравнения методов были использован тест Гельмана-Рубина (Gelman-Rubin test) и визуальная оценка цепей Маркова для одного и того же параметра. Метод Метрополиса показал преимущества по сравнению с методом Монте-Карло.

Итоговая оценка, основанная на двух методах, выявила уровень холестерина, стабилизированный уровень глюкозы, индекс массы тела и отношение waist/hip в качестве основных риск-факторов при прогнозировании диабета. С точки зрения физиологии человека эти факторы являются существенными признаками ожирения, которое рассмотрено, в качестве главного фактора риска возникновения диабета в работе [2].

СПИСОК ЛИТЕРАТУРЫ

1. Paulino C.D., Amaral Turkman M.A., Murteira B., Silva G.L. (2018). Bayesian Statistics. Calouste Gulbenkian Foundation, Lisbon.
2. Willems J.P., Saunders J.T., Hunt D.E., Schorling J.B. (1997) Prevalence of coronary heart disease risk factors among rural blacks: A community-based study. Southern Medical Journal. V. 90, p. 814-820.