Информационный критерий Акаике (Akaike's information criterion, AIC)

К. А Мурашев

5 октября 2021 г.

Современный оценщик в своей практике часто сталкивается с необходимостью выбора конкретной регрессионной модели с точки зрения включения либо невключения в неё отдельных предикторов (ценообразующих факторов). В данном фрагменте рассматривается Информационный критерий Акаике, предназначенный для осуществления выбора такой регрессионной модели, которая позволяет в достаточной мере описать данные, используя при этом минимальное число предикторов, что отчасти позволяет устранить проблему переобучения модели.

1. Общие сведения

Критерий Акаике представляет собой информационный критерий [5] выбора наилучшей модели из нескольких параметризованных регрессионных моделей, имеющих разное число предикторов. Критерий основан на понятии расстояния Кульбака"– Лейбнера [7], являющегося мерой удалённости двух вероятностных распределений относительно друг друга и способного помочь определить расстояние между двумя моделями. Применение данного критерия основывается на Принципе Оккама [1], согласно которому, применительно к регрессионному анализу, можно сказать, что лучшей моделью является та, которая в достаточной мере полно описывает данные, используя при этом наименьшее число предикторов. Данный критерий был разработан в начале 1970-х годов японским исследователем Хироцугу Акаике [4].

2. Описание критерия

Расстояние Кульбака–Лейблера между двумя непрерывными функциями представляет собой интеграл.

$$I(f,g) = \int f(x) ln \frac{f(x)}{g(x|\theta)} \times d(x) \tag{1}$$

Оценка расстояния между двумя моделями при этом может быть осуществлена на основе величины:

$$E_{\hat{\theta}}[I(f, \hat{g})], \tag{2}$$

где $\hat{\theta}$ — оценка вектора параметров, в состав которого входят параметры модели и случайные величины,

$\hat{g} = g(\cdot|\hat{\theta})$.

При этом максимум логарифмической функции правдоподобия и оценка матожидания связаны следующим выражением:

$$\log(L(\hat{\theta}|y)) - K = Const - \hat{E}_{\hat{\theta}}[I[f, \hat{g}]], \tag{3}$$

где K — число параметров модели,

L — максимум логарифмической функции правдоподобия [8, 3].

Таким образом вместо вычисления расстояния между моделями можно ввести оценивающий критерий.

$$AIC = 2K - 2log(L(\hat{\theta}|y)) \tag{4}$$

В интересующем нас случае применения критерия с целью выбора наилучшей регрессионной модели можно использовать следующую формулу данного критерия, основанную на сумме квадратов остатков (SSE):

$$AIC = 2K + n[ln(\delta^2)], \tag{5}$$

где

$$SSE = |f(x_i) - y_i| = \sum_{(} i = 1)^(n)(y_i - f(\omega, x_i))^2 \tag{6}$$

$$\delta^2 = \frac{SSE}{N-2} \tag{7}$$

В случае использования моделей с различным количеством наблюдений (объектов-аналогов) выражение принимает вид.

$$AIC = 2K + n[\ln(\frac{2\pi RSS}{n}) + 1] \tag{8}$$

Наилучшей является та модель, значение AIC которой минимально. При этом само значение AIC не является содержательным и служит только для сравнения моделей.

3. Особенности применения критерия

- Критерий не только вознаграждает за качество приближения, но и штрафует за использование излишнего количества предикторов.
- Штраф за число предикторов ограничивает значительный рост сложности модели.
- Порядок выбора моделей неважен.

4. Модификации критерия

$\mathbf{AIC_c}$ используется в случае работы с относительно небольшим числом наблюдений, когда $\frac{n}{K} \leq 40$. В то же время при наличии значительного числа наблюдений $\frac{n}{K} \geq 40$ возможно применение обоих вариантов, хотя чаще рекомендуется использование базового варианта AIC. Особенность критерия AIC_c заключается в том, что функция штрафа умножается на поправочный коэффициент:

$$AIC_c = AIC + \frac{2K(K+1)}{n-K-1}, \tag{9}$$

При этом данное выражение эквивалентно:

$$AIC_c = \ln\frac{SSE}{n} + \frac{n+K}{n-K-2} \tag{10}$$

QAIC следует использовать для моделей, в которых часть переменных предикторов являются случайными величинами с простыми дискретными распределениями (биномиальное, пуассоновское и т. д.). В таких случаях используется более общая модель, которая получается из рассматриваемой добавлением параметра обобщённого распределения. Оценка параметра определяется как распределение χ^2. В таком случае значение параметра как правило находится на отрезке $c \in [1:4]$. В случае, когда $\hat{c} \leq 1$ следует выполнить замену на $\hat{c} = 1$. При $\hat{c} = 1$ QAIC сводится к AIC.

$$QAIC = 2K - \frac{ln(L)}{\hat{c}} \tag{11}$$

$$QAIC_c = QAIC + \frac{2K(K+1)}{n-K-1} \tag{12}$$

5. Практическая реализация

В языках R и Python существуют функции, позволяющие осуществлять автоматический отбор моделей на основе AIC и его модификаций. В частности, в языке R существует библиотека «MASS», содержащая функцию **stepAIC**, позволяющую автоматически отобрать регрессионную модель, являющуюся наилучшей с точки зрения AIC. В языке Python нет отдельной функции, позволяющей оптимизировать модель по критерию AIC, однако данный критерий можно указать в качестве аргумента при построении лассо-регрессии. Также возможно написание собственного кода, выполняющего пошаговую оптимизацию модели. Для оценщика, понимающего суть метода и формулы, приведённые выше, также не составит труда написать алгоритм расчёта AIC в табличном процессоре. Однако, последнее решение не соответствует лучшим практикам и потому не будет рассматриваться в данной работе.

Рассмотрим пример. Предположим, что существует зависимая переменная y, а также 8 предикторов, записанных в переменные v1, v2 ... v7, v8, записанные в единый датафрейм **df**. Задача состоит в том, чтобы построить модель, включающую в себя

Листинг 1. Реализация на языке R

```
model_aic <- stepAIC(lm(y ~ (.), data = df))
summary(model)
```

Листинг 2. Реализация на языке Python

```
AICs = {}
for k in range(1,len(predictorcols)+1):
for variables in itertools.combinations(predictorcols, k):
predictors = train[list(variables)]
predictors['Intercept'] = 1
res = sm.OLS(target, predictors).fit()
AICs[variables] = 2*(k+1) - 2*res.llf
pd.Series(AICs).idxmin()
```

минимальный необходимый и достаточный набор предикторов. В качестве допущения укажем, что переменные являются независимыми. Для построения модели, оптимизированной методом AIC, достаточно написать простой код, примеры которого приводятся в листингах 1, 2.

6. Выводы

Проверка качества регрессионной модели и её оптимизация является важной частью процесса оценки. К сожалению, некоторые оценщики используют лишь один критерий — коэффициент детерминации (R^2), забывая о необходимости проверки p-значений как всей модели, так и каждого предиктора, а также необходимости её оптимизации, одним из методов которой является AIC.

Источники информации

[1] Machinelearning.ru. *Бритва Оккама.* URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%91%D1%80%D0%B8%D1%82%D0%B2%D0%B0_%D0%9E%D0%BA%D0%BA%D0%B0%D0%BC%D0%B0 (дата обр. 05.10.2021).

[2] Machinelearning.ru. *Критерий Акаике.* URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%9A%D1%80%D0%B8%D1%82%D0%B5%D1%80%D0%B8%D0%B9_%D0%90%D0%BA%D0%B0%D0%B8%D0%BA%D0%B5 (дата обр. 05.10.2021).

[3] Machinelearning.ru. *Метод наибольшего правдоподобия.* URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%B5%D1%82%D0%BE%D0%B4_%D0%BD%D0%B0%D0%B8%D0%B1%D0%BE%D0%BB%D1%8C%D1%88%D0%B5%D0%

B3%D0%BE_%D0%BF%D1%80%D0%B0%D0%B2%D0%B4%D0%BE%D0%BF%D0%BE%D0%B4%D0%BE%D0%B1%D0%B8%D1%8F (дата обр. 05.10.2021).

[4] Wikipedia. *Акаике, Хироцугу.* URL: https://ru.wikipedia.org/wiki/%D0%90%D0%BA%D0%B0%D0%B8%D0%BA%D1%8D,_%D0%A5%D0%B8%D1%80%D0%BE%D1%86%D1%83%D0%B3%D1%83 (дата обр. 05.10.2021).

[5] Wikipedia. *Информационный критерий.* URL: https://ru.wikipedia.org/wiki/%D0%98%D0%BD%D1%84%D0%BE%D1%80%D0%BC%D0%B0%D1%86%D0%B8%D0%BE%D0%BD%D0%BD%D1%8B%D0%B9_%D0%BA%D1%80%D0%B8%D1%82%D0%B5%D1%80%D0%B8%D0%B9 (дата обр. 05.10.2021).

[6] Wikipedia. *Информационный критерий Акаике.* URL: https://ru.wikipedia.org/wiki/%D0%98%D0%BD%D1%84%D0%BE%D1%80%D0%BC%D0%B0%D1%86%D0%B8%D0%BE%D0%BD%D0%BD%D1%8B%D0%B9_%D0%BA%D1%80%D0%B8%D1%82%D0%B5%D1%80%D0%B8%D0%B9_%D0%90%D0%BA%D0%B0%D0%B8%D0%BA%D0%B5 (дата обр. 05.10.2021).

[7] Wikipedia. *Расстояние Кульбака—Лейблера.* URL: https://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D1%81%D1%81%D1%82%D0%BE%D1%8F%D0%BD%D0%B8%D0%B5_%D0%9A%D1%83%D0%BB%D1%8C%D0%B1%D0%B0%D0%BA%D0%B0_%E2%80%94_%D0%9B%D0%B5%D0%B9%D0%B1%D0%BB%D0%B5%D1%80%D0%B0 (дата обр. 05.10.2021).

[8] Wikipedia. *Функция правдоподобия.* URL: https://ru.wikipedia.org/wiki/%D0%A4%D1%83%D0%BD%D0%BA%D1%86%D0%B8%D1%8F_%D0%BF%D1%80%D0%B0%D0%B2%D0%B4%D0%BE%D0%BF%D0%BE%D0%B4%D0%BE%D0%B1%D0%B8%D1%8F (дата обр. 05.10.2021).