Вики

Скорректированный коэффициент детерминации (Adjusted coefficient of determination)

Показатель, выражающий долю дисперсии зависимой переменной, объясняемую регрессионной моделью с заданным набором независимых переменных, скорректированный с помощью штрафа, накладываемого на модель при увеличении числа переменных.

$$R^2_{adj} = 1 - \frac{s^2}{s^2_y} = 1 - \frac{RSS/(n-k)}{TSS/(n-1)} = 1 - (1 - R^2)\frac{(n-1)}{(n-k)} \leq R^2,$$

где

$RSS = \sum\limits_{n}^{t=1} e_t^2 = \sum\limits_{n}^{t=1} (y_t - \hat{y}_t)^2$ — сумма квадратов остатков регрессии,

$TSS = \sum\limits_{n}^{t=1} (y_t - \bar{y})^2 = n\hat{\sigma}_y^2$ — общая дисперсия,

n — количество наблюдений в наборе данных,

k — количество параметров модели.

Как и исходный коэффициент детерминации, скорректированный позволяет оценивать соответствие регрессионной модели исходным данным, а также сравнивать модели с разным числом независимых переменных.

Поскольку каждый раз при добавлении в модель новой независимой переменной доля объясненной дисперсии зависимой переменной возрастает, логично было бы включить в модель как можно больше переменных. Но на практике данный подход не дает хороших результатов, поскольку не гарантирует, что будут выбраны именно те переменные, которые вносят значимый вклад в долю объясненной дисперсии. Кроме того, падает отношение числа наблюдений к числу параметров модели, что повышает вероятность переобучения.

Скорректированный коэффициент детерминации позволяет решить данную проблему, поскольку вводит штраф модели, который увеличивается при добавлении каждой переменной.

Если исходный коэффициент детерминации непрерывно возрастает при добавлении каждой независимой переменной, то скорректированный сначала также возрастает, а потом начинает уменьшаться из-за того, что возрастание штрафа начинает «перевешивать» рост объясненной доли дисперсии. Тогда лучшей будет та модель, для которой значение скорректированного коэффициента детерминации максимально.

Значение скорректированного коэффициента детерминации изменяется в диапазоне от 0 до 1, но при этом всегда несколько меньше значения исходного. В некоторых случаях оно может принимать небольшие отрицательные значения для «бесполезных» моделей, предсказания которых хуже, чем оценки на основе простого среднего.

При хорошем согласии модели и данных исходный и скорректированный коэффициенты детерминации должны быть близки к 1 и примерно равны.