Коэффициент детерминации (Coefficient of determination)

> Синонимы: Коэффициент смешанной корреляции, Коэффициент R-квадрат
> Разделы: Метрики
> Loginom: Статистика (визуализатор)

Статистический показатель, отражающий объясняющую способность регрессии $f : X \to Y$ и определяемый как доля дисперсии зависимой переменной, объясненная регрессионной моделью с данным набором независимых переменных. Обычно определяется как единица минус доля необъясненной дисперсии, т.е:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y})^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{SSE}{SST}, (1)$$

где:

- $SSE = \sum_i (y_i - \hat{y}_i)^2$ — сумма квадратов остатков (ошибок) регрессии (sum square of errors),
- $SST = \sum_i (y_i - \bar{y}_i)^2$ — полная сумма квадратов (sum square total), т.е. сумма квадратов отклонений точек данных от среднего значения,
- $X^n = (x_i, y_i)_{i=1}^n$ — набор данных из n наблюдений,
- $y_i \in Y, \bar{y}_i = \frac{\sum_i y_i}{n}$,
- $\widehat{y_i} = f(x_i)$.

Коэффициент детерминации является статистической мерой согласия, с помощью которой можно определить, насколько модель линейной регрессии соответствует данным, на которых она построена.

Коэффициент детерминации изменяется в диапазоне от $-\infty$ до 1. Если он равен 1, это соответствует идеальной модели, когда все точки наблюдений лежат точно на линии регрессии, т.е. сумма квадратов их отклонений равна 0. Если коэффициент детерминации равен 0, это означает, что связь между переменными регрессионной модели отсутствует, и вместо нее для оценки значения выходной переменной можно использовать простое среднее ее наблюдаемых значений.

На практике, если коэффициент детерминации близок к 1, это указывает на то, что модель работает очень хорошо (имеет высокую значимость), а если к 0, то это означает низкую значимость модели, когда входная переменная плохо «объясняет» поведение выходной, т.е. линейная зависимость между ними отсутствует. Очевидно, что такая модель будет иметь низкую эффективность.

Кроме того, бывают случаи, когда коэффициент детерминации принимает отрицательные значения (обычно небольшие). Это случается, когда ошибка модели простого среднего становится меньше ошибки регрессионной модели. Таким образом, добавление в модель с константой некоторой переменной только ухудшает ее.

Иногда коэффициент детерминации вводят как отношение:

$$R^2 = \frac{\sum_i \left(\hat{y}_i - \bar{y}_i\right)^2}{\sum_i \left(y_i - \bar{y}_i\right)^2},$$

где $SSR = \sum_i \left(\hat{y}_i - \bar{y}_i\right)^2$ — сумма квадратов регрессии (sum square of regression). Хотя данное определение является более простым, оно может использоваться только для регрессии с константой (если свободный член в уравнении регрессии не равен нулю), когда знаменатель не обращается в 0. В противном случае необходимо использовать общее соотношение (1).

Недостатком коэффициента детерминации при его применении в качестве меры значимости регрессионных моделей заключается в том, что его значение возрастает (по крайней мере не уменьшается) при добавлении в модель новых зависимых переменных, даже если они никак не связаны с независимой. Это делает сравнение регрессионных моделей с разными наборами предикторов с использованием коэффициента детерминации некорректным.

Поэтому для сравнения моделей используют скорректированный коэффициент детерминации, при вычислении которого вводится штраф за дополнительно вводимые в модель переменные. Скорректированный (adjusted) коэффициент детерминации вычисляется по формуле:

$$R^2_{adj} = 1 - \frac{\sum_i (y_i - \hat{y})^2 (n-k)}{\sum_i \left(y_i - \bar{y}\right)^2 (n-1)} = 1 - \frac{SSE(n-k)}{SST(n-1)},$$

где k — число независимых переменных модели, n — количество наблюдений в наборе данных.

Очевидно, что $R^2_{adj} \leq R^2$. При этом скорректированный коэффициент детерминации может принимать значения в том же диапазоне, что и обычный от $-\infty$ до 1.