

Линейная регрессия. Двухвыборочный t-тест. A/B-тестирование

Теория вероятностей
и математическая статистика / Урок 7



Линейная регрессия

В общем виде модель регрессии — это любая модель зависимости (объясняемой) количественной переменной y от другой или нескольких других переменных (факторов) x_i . Такую модель можно записать в виде:

$$y = f_b(x_1, \dots, x_m) + \varepsilon,$$

где $f_b(x)$ — некоторая функция, имеющая набор параметров b , а ε — случайная ошибка. При этом на ошибку накладывается условие, что её математическое ожидание равно 0:

$$M(\varepsilon) = 0$$

В общем виде модель регрессии — это любая модель зависимости (объясняемой) количественной переменной y от другой или нескольких других переменных (факторов) x_i . Такую модель можно записать в виде:

$$y = f_b(x_1, \dots, x_m) + \varepsilon,$$

где $f_b(x)$ — некоторая функция, имеющая набор параметров b , а ε — случайная ошибка. При этом на ошибку накладывается условие, что её математическое ожидание равно 0:

$$M(\varepsilon) = 0$$

В чём здесь суть. Участвующие в этой модели переменные удобно воспринимать как некоторые случайные величины, т.е. случайное поведение переменной y моделируется в виде некоторой функции от случайных факторов x_i , а остаток от всего этого (т.е. та часть этой зависимости, которая осталась неучтённой) — это и есть случайная ошибка ε .

Модель регрессии называется **линейной**, если функция $f_b(x)$ является линейной, т.е. модель имеет вид:

$$y = b_0 + b_1x_1 + \dots + b_mx_m + \varepsilon$$

Важным частным случаем линейной регрессии является **парная регрессия**. При парной регрессии используется только один фактор, т.е. модель имеет вид:

$$y = b_0 + b_1x + \varepsilon$$

На практике такая модель имеет вид:

$$Y = b_0 + b_1X + E,$$

где X — выборка из значений фактора x , Y — выборка из значений переменной y , а E — значения ошибок модели на каждом объекте (т.е. реализации случайной величины ε). В этом случае условие $M(\varepsilon) = 0$ трансформируется в условие $\overline{E} = 0$, где \overline{E} — выборочное среднее ошибок.

Коэффициенты парной регрессии можно найти по формуле:

$$b_1 = \frac{\sigma_{XY}}{\sigma_X^2}, \quad b_0 = \bar{Y} - b_1 \cdot \bar{X},$$

где σ_X^2 — выборочная дисперсия, σ_{XY} — выборочная ковариация.

В общем случае, когда факторов больше одного, коэффициенты можно также посчитать аналитически с помощью метода наименьших квадратов. Модель имеет следующий вид:

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_mX_m + E,$$

где X_i — выборка из значений i -го фактора.

Как правило, для удобства нахождения оптимальных коэффициентов к каждому объекту добавляется «нулевой фактор» x_0 , который равен 1 для каждого объекта. Это нужно просто чтобы модель записывалась в более симметричном виде:

$$Y = b_0X_0 + b_1X_1 + b_2X_2 + \dots + b_mX_m + E,$$

где выборка X_0 полностью состоит из единиц.

Такая форма записи модели эквивалентна матричной форме:

$$Y = X \cdot b + E,$$

где X — матрица объект признак (т.е. элемент x_{ij} из этой матрицы является j -м признаком i -го объекта), $b = (b_0, b_1, \dots, b_m)$ — вектор коэффициентов модели, « \cdot » — операция матричного умножения.

Такая форма записи модели эквивалентна матричной форме:

$$Y = X \cdot b + E,$$

где X — матрица объект признак (т.е. элемент x_{ij} из этой матрицы является j -м признаком i -го объекта), $b = (b_0, b_1, \dots, b_m)$ — вектор коэффициентов модели, « \cdot » — операция матричного умножения.

Итак, метод наименьших квадратов заключается в минимизации расстояния между векторами Y и $X \cdot b$:

$$\|Y - X \cdot b\| \rightarrow \min_b$$

Такая задача имеет аналитическое решение:

$$b = (X^T X)^{-1} X^T Y$$

Рассмотрим случайную ошибку

$$\varepsilon = y - x \cdot b$$

Коэффициенты модели линейной регрессии подбираются так, чтобы математическое ожидание ошибки было равно нулю:

$$M(\varepsilon) = 0$$

Теперь качество модели определяет дисперсия ошибки $D(\varepsilon)$. Если и математическое ожидание, и дисперсия ошибки близки к нулю, это свидетельствует о высоком качестве модели, т.е. в этом случае модель хорошо соответствует имеющимся данным. Эта интуиция приводит нас к коэффициенту детерминации:

$$R^2 = 1 - \frac{D(\varepsilon)}{D(y)}$$

Коэффициент детерминации принимает значения из интервала $[0, 1]$. Близкие к 1 значения коэффициента детерминации свидетельствуют о высоком качестве модели.

Чтобы посчитать коэффициент детерминации, построим массив из ошибок модели:

$$E = Y - X \cdot b$$

Пусть $SS_Y = \sum_{i=1}^n (y_i - \bar{Y})^2$ — сумма квадратов отклонений значений массива Y от среднего, а SS_{res} — остаточная сумма квадратов, т.е. сумма квадратов отклонений элементов массива E от их среднего.

Коэффициент детерминации:

$$R^2 = 1 - \frac{SS_{res}}{SS_y}$$

Для модели, построенной методом наименьших квадратов, коэффициент детерминации можно посчитать иначе. Он равен квадрату коэффициента корреляции Пирсона между массивом Y и массивом значений модели $Z = X \cdot b$:

$$R^2 = r_{YZ}^2$$

В случае парной регрессии коэффициент детерминации вовсе можно посчитать напрямую по входным данным. Он равен квадрату коэффициента корреляции Пирсона между массивами X и Y :

$$R^2 = r_{XY}^2$$

Это наблюдение приводит нас к важному **замечанию** касательно коэффициента детерминации: несмотря на то, что абсолютный максимум коэффициента детерминации равен 1, это не означает, что если он меньше 1, то модель в каком-то смысле *плохая*, что её ещё можно было бы улучшить.

Как мы видим, качество модели линейной регрессии напрямую зависит от уровня линейной зависимости в данных. Другими словами, если в данных такой зависимости нет, то и коэффициент детерминации не достигнет 1.

Метод наименьших квадратов можно применить для мощного обобщения линейной регрессии, а именно **полиномиальной регрессии**. Как это делается — читайте в дополнительных материалах к уроку.

Статистический анализ уравнения регрессии

Итак, ранее мы установили, что верхняя граница коэффициента детерминации для модели линейной регрессии, построенной по имеющимся данным, не всегда равна 1. Так как же тогда определить, какой коэффициент детерминации означает значимый уровень соответствия модели данным, а какой — нет?

Для таких целей существует т.н. F-тест Фишера. Формально при таком тесте проверяется нулевая гипотеза о том, что теоретический коэффициент детерминации равен 0. Другими словами, если нулевая гипотеза верна, то между факторами и целевой переменной вообще нет никакой значимой зависимости, а отличие коэффициента детерминации от нуля обусловлено лишь случайностью процесса.

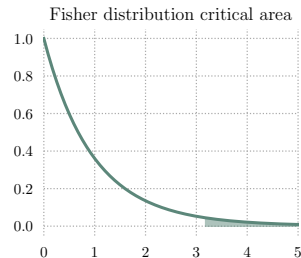
При F-тесте используется статистика:

$$F = \frac{R^2/m}{(1 - R^2)/(n - m - 1)},$$

где R^2 — коэффициент детерминации, n — число наблюдений, m — число факторов. Такая статистика в предположении верности нулевой гипотезы имеет F-распределение Фишера с параметрами $k_1 = m$, $k_2 = n - m - 1$.

Распределение Фишера имеет один хвост, поэтому рассматривается правосторонняя критическая область $\Omega_\alpha = (t_{1-\alpha, k_1, k_2}, \infty)$, где $t_{1-\alpha, k_1, k_2}$ — квантиль порядка $1 - \alpha$ для распределения Фишера с параметрами k_1, k_2 .

Если статистика попадает в критическую область, то гипотеза о равенстве нулю коэффициента детерминации отвергается. Уравнение признаётся значимым.



В случае парной регрессии можно построить доверительные интервалы для коэффициентов регрессии.

Смысл доверительных интервалов тут в том, что, как мы уже отмечали ранее, модель линейной регрессии формально определена прямо для случайных величин, т.е. существуют какие-то *теоретические* коэффициенты модели. В свою очередь, построенные нами (например, с помощью метода наименьших квадратов) коэффициенты представляют собой *оценки* этих коэффициентов.

Таким образом, построив доверительные интервалы для этих коэффициентов, мы можем выяснить, насколько далеко могут быть реальные значения коэффициентов регрессии от построенных нами.

Начнём с коэффициента наклона b_1 . Допустим, мы получили коэффициент наклона \hat{b}_1 , и пусть b_1 — реальное значение этого коэффициента. Рассмотрим статистику

$$t = \frac{\hat{b}_1 - b_1}{S_{slope}},$$

где S_{slope} — стандартная ошибка коэффициента наклона:

$$S_{slope} = \sqrt{\frac{\frac{1}{n-2} \sum_{i=1}^n e_i^2}{\sum_{i=1}^n (x_i - \bar{X})^2}}$$

Здесь e_i — значение ошибки на i -м объекте, т.е. $e_i = y_i - z_i$.

Статистика t имеет распределение Стьюдента с параметром $df = n - 2$. Отсюда можно, имея доверительную вероятность p , построить доверительный интервал для коэффициента наклона по формуле:

$$P\left(\hat{b}_1 + t_{\alpha/2, n-2} \cdot S_{slope} \leq b_1 \leq \hat{b}_1 + t_{1-\alpha/2, n-2} \cdot S_{slope}\right) = p,$$

где $\alpha = 1 - p$, $t_{x, n-2}$ — квантиль порядка x для распределения Стьюдента.

Этот интервал можно интерпретировать так: если у нас появятся новые данные из того же распределения, то с вероятностью p модель линейной регрессии на этих новых данных будет иметь коэффициент наклона в пределах этого интервала.

Аналогичным образом можно построить доверительный интервал для коэффициента сдвига b_0 . Как это делать — читайте в дополнительных материалах к уроку.

Кроме того, доверительные интервалы можно также строить и для коэффициентов многомерной регрессии (т.е. когда имеется несколько факторов). По этому поводу также имеется дополнительный материал.

Итак, подытожим весь описанный выше регрессионный анализ и возникающие в процессе величины:

- Непосредственно факт наличия линейной взаимосвязи проверяется с помощью корреляционного анализа.
- Если линейная зависимость наблюдается, можно построить модель линейной регрессии. Она укажет на характер этой зависимости (т.е. на то, каким именно образом изменяется переменная под влиянием факторов).
- С помощью F-критерия Фишера можно проверить, является ли уровень зависимости в данных статистически значимым.
- С помощью доверительных интервалов можно оценить реальный вклад каждого фактора в изменение переменной.

Двухвыборочный t-тест

Ранее с помощью распределения Стьюдента мы научились проверять гипотезы о математическом ожидании и о коэффициенте корреляции.

Распределение Стьюдента также можно применять для следующей задачи.

Имеются две независимые выборки X_1, X_2 , взятые, соответственно, из распределений D_1, D_2 . Требуется проверить гипотезу о том, что математические ожидания двух соответствующих распределений равны, т.е.

$$H_0 : M(D_1) = M(D_2)$$

Допустим, размеры выборок равны, соответственно, n_1 и n_2 . Обозначим через

$$\sigma_{\Delta} = \sqrt{\frac{\sigma_{X_1}^2}{n_1} + \frac{\sigma_{X_2}^2}{n_2}}$$

среднее квадратическое отклонение разности между выборочными средними выборок X_1 и X_2 . Здесь $\sigma_{X_i}^2$ — несмещённая оценка дисперсии по выборке X_i .

Для проверки данной гипотезы используется следующая статистика:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sigma_{\Delta}}$$

В предположении нулевой гипотезы данная статистика имеет распределение Стьюдента. Число степеней свободы распределения определяется из следующего равенства:

$$df = \frac{\left(\frac{\sigma_{X_1}^2}{n_1} + \frac{\sigma_{X_2}^2}{n_2} \right)^2}{\frac{(\sigma_{X_1}^2/n_1)^2}{n_1 - 1} + \frac{(\sigma_{X_2}^2/n_2)^2}{n_2 - 1}}$$

Дальнейшая процедура проверки гипотезы стандартна: выбирается уровень значимости α , по нему строится двухсторонняя критическая область (с использованием квантилей $t_{x, df}$ распределения Стьюдента) и наконец проводится статистический тест.


A/B-тестирование

A/B-тестирование (или сплит-тестирование) — маркетинговый метод, который используется для оценки эффективности веб-страниц и управления ими.

При A/B-тестировании сравнивают страницы A и B, имеющие разные элементы дизайна (например, цвета кнопки заказа товара). На каждую страницу случайным образом запускают около 50% аудитории сайта и затем сравнивают, какая страница показывает наибольший процент конверсии.

За нулевую гипотезу берётся предположение, что конверсия на страницах A и B не отличается. Соответственно, обратное утверждение берётся за альтернативную гипотезу.

Как правило, для проверки такой гипотезы как раз используется двухвыборочный t-тест.



Спасибо за внимание