Теория вероятностей и математическая статистика

Вебинар 7

Линейная регрессия. Однофакторный дисперсионный анализ. A/B-тестирование

Линейная регрессия

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:

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

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:

$$M(\varepsilon) = 0$$

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:

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

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:

$$M(\varepsilon) = 0$$

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

$$y = b_0 + b_1 x_1 + \cdots + b_m x_m + \varepsilon$$

В чём здесь суть. Участвующие в этой модели переменные удобно воспринимать как некоторые случайные величины. В этом случае можно всегда подобрать параметры b так, чтобы такое равенство выполнялось прямо на уровне случайных величин.

Парная регрессия

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

$$y = b_0 + b_1 x + \varepsilon$$

На практике такая модель имеет вид:

$$Y = b_0 + b_1 X + E,$$

где X — значения фактора x, Y — значения переменной y, 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.

Парная регрессия

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

$$y = b_0 + b_1 x + \varepsilon$$

На практике такая модель имеет вид:

$$Y = b_0 + b_1 X + E,$$

где X — значения фактора x, Y — значения переменной y, 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:

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

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

Метод наименьших квадратов

В общем случае, когда факторов больше одного, коэффициенты можно подобрать с помощью *метода наименьших квадратов*. Здесь X — это уже не просто выборка, а матрица объект-признак, т.е. элемент x_{ij} из этой матрицы является j-м признаком i-го объекта.

Для удобства записи метода наименьших квадратов в первую очередь введём дополнительный «фактор» $x_0 = 1$. Это делается для того, чтобы модель можно было записать в матричном виде:

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

где X — такая расширенная матрица объект-признак (первый столбец которой полностью состоит из единиц), $b = (b_0, b_1, \ldots, b_m)$ — вектор коэффициентов модели, операция «$\cdot$» — матричное умножение.

Метод наименьших квадратов

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

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

При этом вводится дополнительное условие на среднюю ошибку:

$$\overline{E} = 0,$$

где $E = Y - X \cdot b$. Решение такой оптимизационной задачи даёт *коэффициенты линейной регрессии*:

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

Метод наименьших квадратов

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

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

При этом вводится дополнительное условие на среднюю ошибку:

$$\overline{E} = 0,$$

где $E = Y - X \cdot b$. Решение такой оптимизационной задачи даёт *коэффициенты линейной регрессии*:

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

Замечание. У метода наименьших квадратов есть один изъян: в случае, когда в матрице X представлены линейно зависимые (или близкие к этому) признаки, вычисление обратной матрицы $(X^\top X)^{-1}$ становится проблематичным. В таких ситуациях стоит сперва избавиться от линейно зависимых признаков (это задача *факторного анализа*, который мы рассмотрим на занятии 8).

Коэффициент детерминации

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.

Коэффициент детерминации на практике

Чтобы посчитать коэффициент детерминации, построим «предсказанные» моделью значения

$$Z = X \cdot b$$

Пусть σ^2_Y — выборочная дисперсия по массиву реальных значений Y, а σ^2_{res} — *остаточная дисперсия*, т.е. выборочная дисперсия по массиву ошибок $Z - Y$. Тогда *коэффициент детерминации*:

$$R^2 = 1 - \frac{\sigma^2_{res}}{\sigma^2_y}$$

Корреляция vs детерминация

Несмотря на то, что теоретически коэффициент детерминации принимает значения от 0 до 1, значение коэффициента детерминации ниже 1 не означает, что модель построена плохо (и могла бы быть лучше).

Рассмотрим модель, построенную с помощью метода наименьших квадратов. (Напомним, что такая модель является наилучшей моделью линейной регрессии, которую можно построить на имеющихся данных.) Пусть r_{YZ} — коэффициент корреляции Пирсона между массивами Y и Z. Оказывается, в таком случае справедливо равенство:

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

Таким образом, коэффициент детерминации прямо зависит от уровня корреляции в данных и не может достигнуть 1, если в данных нет линейной зависимости.

Статистический анализ уравнения регрессии

F-критерий Фишера

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 (т.е. для модели, построенной для случайных величин) равен 0, т.е. что в имеющихся данных вообще нет никакой зависимости.

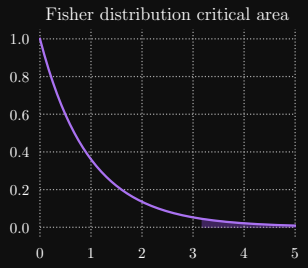
F-статистика Фишера

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 $t_{\beta,\, k_1,k_2}$ — квантиль порядка β для распределения Фишера с параметрами k_1, k_2.



Если статистика попадает в критическую область, то гипотеза о равенстве нулю коэффициента детерминации отвергается. Это означает, что построенная нами модель значимо соответствует данным.

Доверительные интервалы для коэффициентов парной регрессии

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 можно построить прямо по случайным величинам. Это значит, что, построив модель по имеющимся данным, можно построить доверительные интервалы и посмотреть, насколько далеко могут быть реальные значения коэффициентов регрессии от построенных нами.

Доверительные интервалы для коэффициентов парной регрессии

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*:

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

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

Доверительные интервалы для коэффициентов парной регрессии

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

Доверительные интервалы для коэффициентов парной регрессии

Аналогично можно построить доверительный интервал для коэффициента сдвига b_0. *Стандартная ошибка коэффициента сдвига* вычисляется по формуле:

$$S_{intercept} = S_{slope} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

Статистика

$$t = \frac{\hat{b}_0 - b_0}{S_{intercept}}$$

также имеет распределение Стьюдента с параметром $df = n - 2$. Итак, *доверительный интервал для коэффициента наклона*:

$$P\left(\hat{b}_0 + t_{\alpha/2,\, n-2} \cdot S_{intercept} \leq b_0 \leq \hat{b}_0 + t_{1-\alpha/2,\, n-2} \cdot S_{intercept}\right) = p$$

Дисперсионный анализ

Дисперсионный анализ

Дисперсионный анализ — метод в математической статистике, направленный на поиск зависимостей в данных, в которых целевая переменная является *количественной*, а факторы являются *категориальными*.

В *однофакторном дисперсионном анализе* исследуется влияние одного категориального фактора x на переменную y. Допустим, у фактора x имеется k разных значений или *уровней*. На практике это означает, что у нас имеется k выборок:

$$Y_1, \ldots, Y_k,$$

и выборка Y_i соответствует значениям переменной y на i-м уровне фактора x.

Итак, нулевая гипотеза H_0 утверждает, что средние по всем этим выборкам равны:

$$H_0: \ \overline{Y_1} = \cdots = \overline{Y_k}$$

Другими словами, нулевая гипотеза заключается в том, что фактор x никак не влияет на значения переменной y.

Однофакторный дисперсионный анализ

Для проверки гипотез в дисперсионном анализе также используется *F-критерий Фишера*. Используемая статистика представляет из себя отношение дисперсии между уровнями к дисперсии внутри уровней.

Пусть в каждой выборке Y_i содержится n_i элементов. Обозначим через Y объединение всех выборок, т.е. выборку размера $n = n_1 + \cdots + n_k$.

Рассмотрим две суммы квадратов:

$$S_b^2 = \sum_{i=1}^{k} \left(\overline{Y_i} - \overline{Y}\right)^2 n_i, \;\; S_w^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \overline{Y_i}\right)^2,$$

где y_{ij} — j-й элемент i-й выборки.

Первая сумма — отклонения между группами («b» от слова Between — между), вторая — отклонения внутри групп («w» от слова Within — внутри).

Однофакторный дисперсионный анализ

По этим значениям вычисляются соответствующие несмещённые оценки дисперсий:

$$\sigma_b^2 = \frac{S_b^2}{k-1},\ \sigma_w^2 = \frac{S_w^2}{n-k}$$

Итак, статистика для проверки гипотезы H_0:

$$F = \frac{\sigma_b^2}{\sigma_w^2}$$

В предположении верности гипотезы H_0 статистика F имеет распределение Фишера с параметрами $k_1 = k-1$, $k_2 = n-k$. Как и ранее, критическая область здесь правосторонняя:

$$\Omega_\alpha = \left(t_{1-\alpha,\, k_1, k_2}.\infty\right),$$

где $t_{\beta,\, k_1, k_2}$ — квантиль порядка β для распределения Фишера с параметрами k_1, k_2.

A/B-тестирование

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 50% аудитории сайта и затем сравнивают, какая страница показывает наибольший процент конверсии.

За нулевую гипотезу берётся предположение, что конверсия на странице B не отличается от конверсии на странице A. Соответственно, обратное утверждение берётся за альтернативную гипотезу.

На следующем занятии

Двухфакторный дисперсионный анализ.
Факторный анализ. Логистическая регрессия