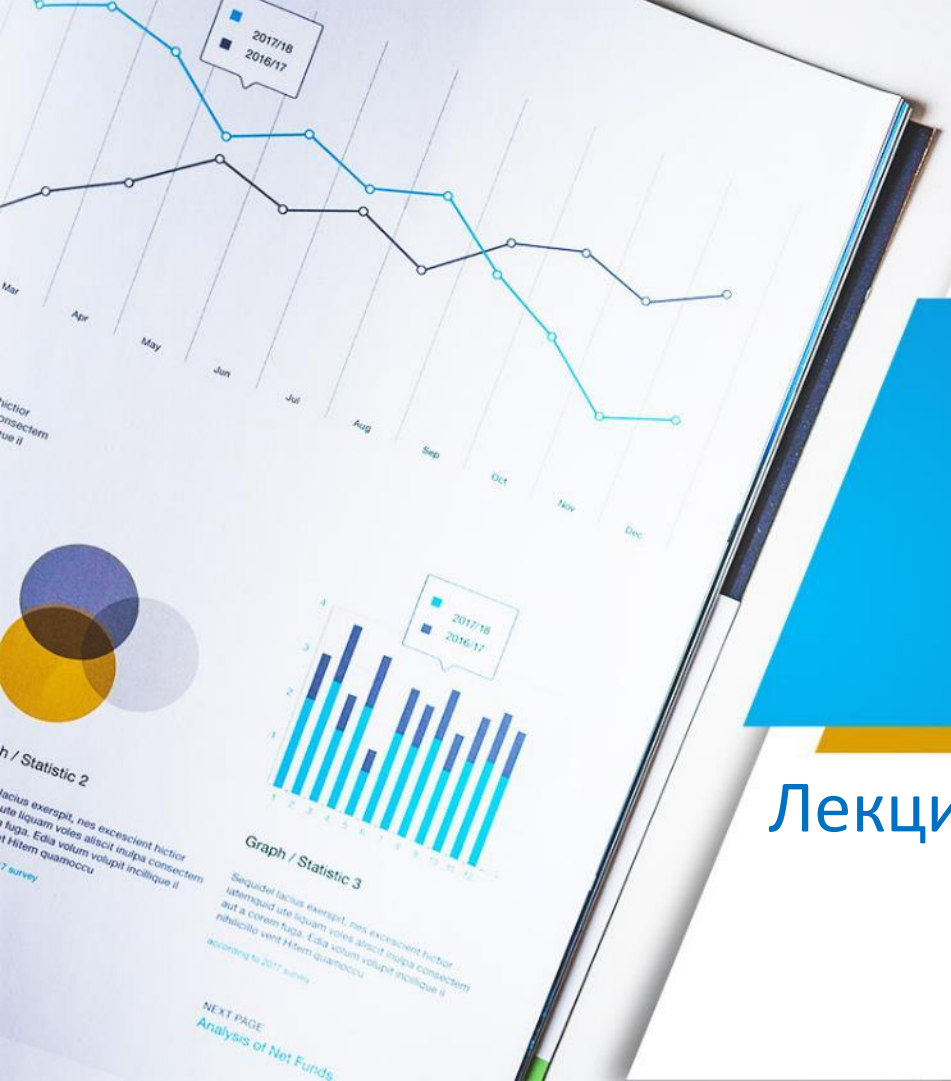


СТАТИСТИКА БОЛЬШИХ ДАННЫХ

Лекция 17. Регрессионный анализ. Часть 1

Введение

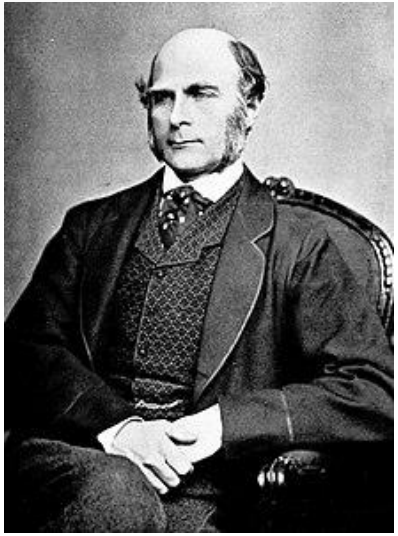
В этой лекции мы рассматриваем постановку задачи и основные понятия регрессионного анализа, важного раздела математической статистики.

- Введение и постановка задачи регрессионного анализа
- Простая линейная регрессия
- Проверка адекватности модели простой линейной регрессии
- Заключение

Регрессионные зависимости — один из самых распространенных видов моделей зависимости одной случайной величины от ряда других случайных или неслучайных величин. Регрессионная зависимость или регрессия записывается в виде функции, линейной или нелинейной, представляющей эту зависимость "в среднем", т.е., приближенно, как записываются практически все физические законы с параметрами и инженерные зависимости.

Введение

Термин "регрессия" ввел в науку выдающийся британский ученый Ф. Гальтон (Galton, 1885), которым он обозначил зависимость размеров потомков от размеров родителей: в ней наблюдался регресс, т.е., рост детей в среднем оказался меньше роста родителей, если рост родителей был выше среднего, и наоборот. Это явление можно объяснить высокой устойчивостью средних значений в наблюдаемой зависимости.



Фрэ́нсис Га́льтон
(1822 — 1911)

Постановка задачи регрессионного анализа

В практике обработки данных часто приходится изучать зависимость наблюдаемой случайной величины Y от одной или нескольких других случайных (или неслучайных) величин $X_1, \dots, X_p$.
Величину Y называют **откликом,** а $X_1, \dots, X_p$ – **факторами** (*регрессорами*, *предикторами*), влияющими на отклик.

Для суждения о значениях отклика y (прогнозе для Y) в зависимости от значений факторов $x_1, \dots, x_p$ желательно иметь функциональную связь $y = f(x_1, \dots, x_p)$. Строгая функциональная зависимость между случайными величинами практически никогда не наблюдается, по причине влияния дополнительных, неучтенных факторов, помех и ошибок. Предполагают *статистическую зависимость* отклика от факторов, отражающую связь между откликом и факторами лишь приближенно, "в среднем".

Рассмотрим регрессионную модель, дающую пример статистической зависимости между переменными.

Постановка задачи регрессионного анализа

Пусть значение отклика y представляет собой сумму двух слагаемых

$$y = f(x) + e,$$

где $f(x)$ — функция наблюдаемых значений факторов $x = (x_1, \dots, x_p)$,
e — случайная величина, называемая **ошибкой**.
Случайная ошибка e отражает влияние случайных факторов, изменчивость, погрешности результатов измерений и т.п.

Постановка задачи регрессионного анализа: по выборке (x_i, y_i) наблюдаемых (экспериментальных) данных о значениях факторов и отклика

$$y_i = f(x_i) + e_i, \qquad i = 1, \dots, n$$

требуется построить оценку функциональной зависимости y от x:

$$\hat{y} = \hat{f}(x).$$

Постановка задачи регрессионного анализа

Замечания:

Рассмотрение аддитивной модели регрессионной зависимости отклика от факторов с неслучайным x и случайным e является упрощающим допущением. Возможны и другие модели связи отклика и факторов: мультипликативная $y = e\,f(x)$, смешанная $y = e_1 f(x) + e_2$ и т.д.

В регрессионной модели $y_i = f(x_i) + e_i$ принимают, что $e_1, \dots, e_n$ — независимые, одинаково распределенные случайные величины с нулевым математическим ожиданием $Ee_i = 0$ и дисперсией $De_i = \sigma^2$, $i = 1, \dots, n$.

Относительно регрессионной функции f, как правило, предполагают, что ее вид задан с точностью до неизвестного векторного параметра $\beta = (\beta_1, \dots, \beta_l)$:

$$y_i = f(x_i, \beta) + e_i, \qquad i = 1, \dots, n.$$

Построение оценки регрессии сводится к построению оценок неизвестных параметров этой функции.

Постановка задачи регрессионного анализа

Важный на практике класс регрессионных моделей возникает при рассмотрении линейных относительно неизвестных параметров β регрессионных зависимостей $f(x,\beta)$ — это задачи **линейного регрессионного анализа.** При этом зависимость отклика от факторов x может быть и нелинейной.

Задача оценивания $f(x,\beta)$ при *нелинейной* зависимости от β называется **нелинейным регрессионным анализом.**

Исторически уравнением регрессии y по x называют условное математическое ожидание E_x отклика Y при заданных значениях факторов x. Действительно, имеем

$$E_x Y = E_x[f(x,\beta) + e] = f(x,\beta) + E_x e = f(x,\beta),$$

так как $E_x e = 0$.

Простая линейная регрессия

Рассмотрим модель **простой линейной регрессии:**

$$y_i = \beta_0 + \beta_1 x_i + e_i, \qquad i =, \dots, n,$$

где $x_1, \dots, x_n$ — заданные числа (значения фактора); $y_1, \dots, y_n$ — наблюдаемые значения отклика; $e_1, \dots, e_n$ — независимые, нормально распределенные $N(0, \sigma)$ с нулевым математическим ожиданием и одинаковой (неизвестной) дисперсией случайные величины (ненаблюдаемые); β_0, β_1 — неизвестные параметры, подлежащие оцениванию.

В этой модели отклик y зависит только от одного фактора x , и весь разброс экспериментальных точек объясняется только погрешностями наблюдений (результатов измерений) отклика y. Погрешности результатов измерений x в этой модели полагают существенно меньшими погрешностей результатов измерений y, так что ими можно пренебречь.

Простая линейная регрессия

Метод наименьших квадратов

Вводится мера (критерий) рассогласования отклика и регрессионной функции, и оценки параметров регрессии определяются так, чтобы сделать это рассогласование наименьшим. Достаточно простые расчетные формулы для оценок получают при выборе критерия в виде суммы квадратов отклонений значений отклика от значений регрессионной функции — суммы квадратов остатков (СКО) (the residual sum of squares (RSS)):

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \to \min_{\beta_0, \beta_1}.$$

Простая линейная регрессия

Подход, основанный на использовании метода наименьших квадратов, был предложен Лежандром (Legendre, 1805) и Гауссом (Gauss, 1809) дает следующие оценки для коэффициентов регрессии:

$$\hat{\beta}_1 = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_1^n (x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

где $\bar{y}$ и $\bar{x}$ — выборочные средние.

Проверка адекватности модели

Точность оценок параметров простой линейной регрессии

Стандартные ошибки $(\boldsymbol{SE})$ оценок $\hat{\beta}_0$ и $\hat{\beta}_1$ имеют вид:

$$\left(SE(\hat{\beta}_0)\right)^2 = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_1^n (x_i - \bar{x})^2}\right),$$

$$\left(SE(\hat{\beta}_1)\right)^2 = \frac{\sigma^2}{\sum_1^n (x_i - \bar{x})^2}.$$

Стандартные ошибки могут быть использованы для вычисления **доверительных интервалов.** В случае простой линейной регрессии $95\% - \text{е}$ доверительные интервалы для β_0, β_1 имеют вид

$$\hat{\beta}_0 \pm 2 \cdot SE(\hat{\beta}_0),$$

$$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1).$$

Проверка адекватности модели

Стандартные ошибки используются для проверки гипотез о коэффициентах регрессии, например

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0.$$

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)},$$

На практике для этого вычисляется t-статистика которая измеряет отклонение оценки $\hat{\beta}_1$ от 0. Если нет связи между X и Y, то t-статистика имеет t-распределение Стьюдента с $n - 2$ степенями свободы.

Удобнее использовать p-значение (p-value) - вероятность наблюдения любого значения случайной величины T, распределенной по Стьюденту, равной или большей $|t|$ в предположении, что $\beta_1 = 0$, т.е., при справедливости нулевой гипотезы H_0

$$p = P(T \geq |t|) \quad \text{при} \quad \beta_1 = 0.$$

Мы отвергаем нулевую гипотезу, если p-значение достаточно мало. Типичное значение порога для p-значения в этом случае равно 5% или 1%. При $n = 30$ это соответствует значению t-статистики около 2 и 2.75.

Проверка адекватности модели

Если нулевая гипотеза отвергнута и решение принято в пользу альтернативы, естественно попробовать количественно оценить степень связи между откликом и фактором. Это обычно делается с помощью двух статистик: стандартной ошибки остатков (the residual standard error (*RSE*)) и корреляционное отношение η^2 (другое обозначение - коэффициент R^2):

$$RSE = \sqrt{\frac{1}{n-2} RSS}$$

$$\eta^2 = \frac{TSS - RSS}{TSS},$$

$$RSS = \sum_{1}^{n} (y_i - \hat{y}_i)^2$$

$$TSS = \sum (y_i - \bar{y})^2$$

TSS измеряет полную дисперсию (изменчивость) отклика, и его можно рассматривать как количество изменчивости внутренне присущей отклику до того как была реализована процедура регрессии. Напротив, RSS измеряет количество изменчивости, которое осталось необъясненным после процедуры регрессии

Проверка адекватности модели

TSS-RSS измеряет количество изменчивости в отклике, которое объясняется (или устраняется) процедурой регрессии, и η^2 измеряет долю изменчивости в отклике Y, которая может быть объяснена при использовании фактора X. Значения η^2 близкие к 1 указывают, что большая доля изменчивости отклика может быть объяснено регрессией. Значения близкие к 0 указывают, регрессия не объяснила большую часть изменчивости отклика; это может быть по причине неадекватности линейной модели, или потому, что внутренняя дисперсия ошибок σ^2 высока, или по обеим причинам.

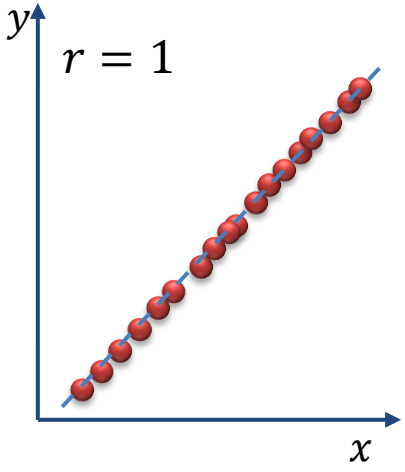
Статистика η^2 имеет преимущество над статистикой *RSE* еще потому, что она является мерой линейной зависимости между X и Y, и в модели простой линейной регрессии равна квадрату выборочного коэффициента корреляции Пирсона r между случайными величинами X и Y:

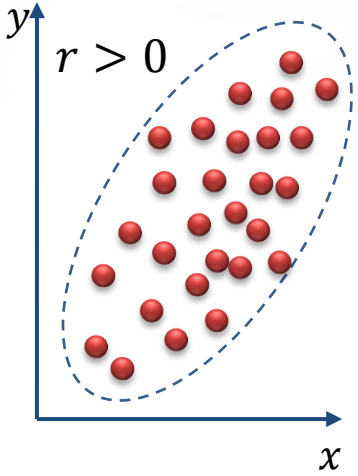
$$\eta^2 = r^2,$$

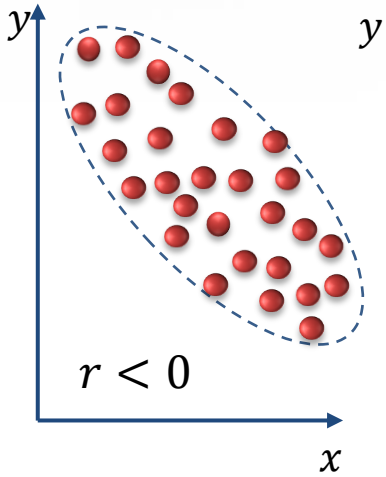
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}.$$

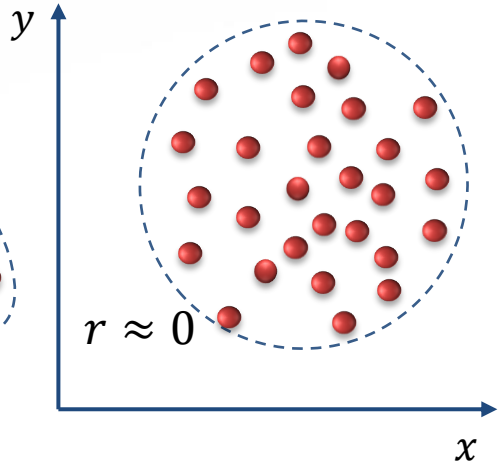
Коэффициент корреляции Пирсона

$$r = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n s_x s_y}.$$









15