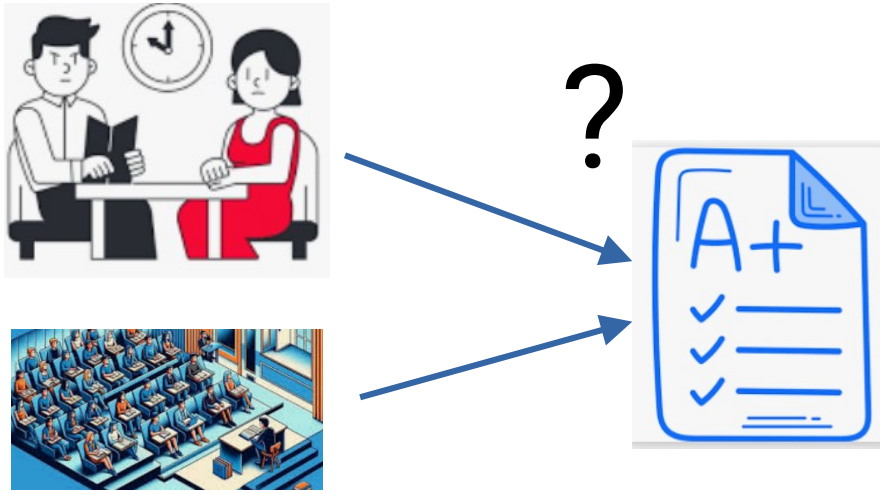
Регрессионный анализ

Регрессионный анализ – это статистический метод для количественного определения (предсказания) значения одной переменной на основании другой (других). **Как** одна (иди несколько) переменная(ых) влияет на другую переменную?



независимые переменные (предиктор, контролируемые факторы)

зависимая переменная (отклик)

Исследователь сам задает причинно следственную связь (что на что влияет).

Цели:

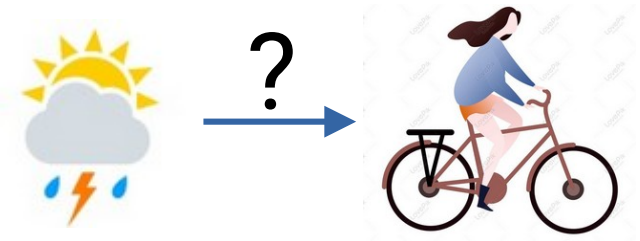
1. Предсказание значения зависимой переменной с помощью независимых переменных.

Какая будет оценка в зависимости от того, сколько часов потрачено на подготовку?

2. Измерение влияния отдельных независимых переменных на зависимую переменную (расширение корреляционного анализа на нелинейные зависимости).

Виды регрессионного анализа

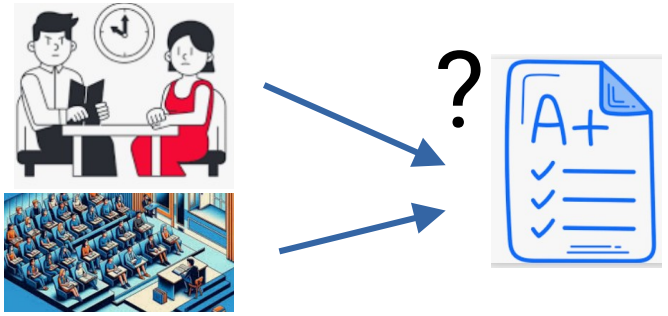
Однофакторная линейная (нелинейная) регрессия



<u>не</u>зависимая переменная ***(категориальные, порядковые, количественные)***

зависимая переменная (***количественная***)

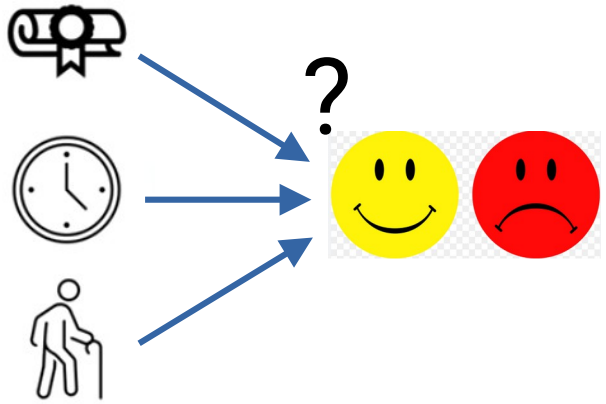
Многофакторная линейная (нелинейная) регрессия



<u>не</u>зависимая переменная ***(категориальные, порядковые, количественные)***

зависимая переменная (***количественная***)

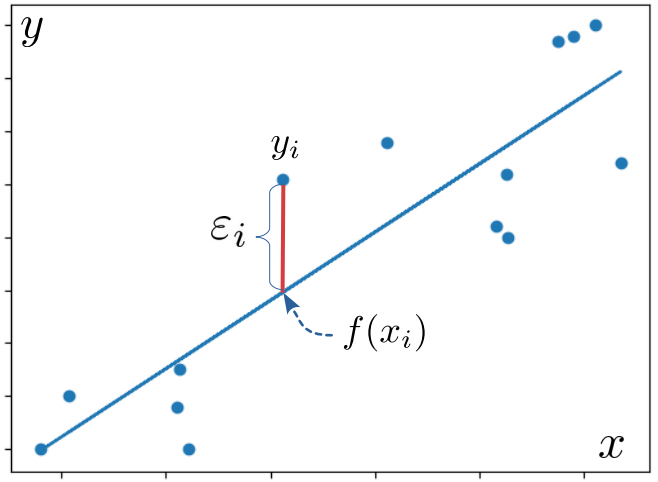
Логистическая регрессия



<u>не</u>зависимая переменная ***(категориальные, порядковые, количественные)***

зависимая переменная (***категориальная с двумя категориями***)

Регрессионный анализ



Будем предполагать, что наблюдаемое в опыте значение отклика y можно мысленно разделить на две части:

$$y = f(x)+\varepsilon$$

$f(x)$ — функциональная закономерность между y и x,

ε — некоторая случайная величина по отношению к x.

Тогда для каждого x_i: $y_i = f(x_i)+\varepsilon_i$, $i = 1...n$

Разделение на закономерную и случайную составляющие можно сделать только мысленно. Реально нам *не известны* ни $f(x_i)$, ни ε_i. Из опыта мы знаем только их сумму y_i. В связи с этим вводятся дополнительные предположения относительно величин ε_i.

В классической модели регрессионного анализа *предполагается*, что ε_i, $i = 1...n$:

а) независимые случайные величины

б) имеют одинаковый закон распределения $\varepsilon_i \in N(0, \sqrt{\sigma^2})$

Т. к. ошибка (погрешность) складывается из большого числа случайных факторов, которые мы не учитываем, то их сумма согласно ЦПТ имеет нормальный закон.

Этапы регрессионного анализа

1. Подготовка данных для анализа (убрать пустые значения, привести значения к нужному формату).
2. Визуализация данные (построение диаграмм рассеяния).
3. Выбор вида уравнения регрессии $f(x_i, \theta)$ (т. е. класса аппроксимирующих функций за счет параметров).
4. Нахождение оценок для параметров θ выбранного уравнения регрессии.
5. Нахождение доверительных интервалов и оценка значимости для параметров уравнения регрессии.
6. Оценка качества выбранного регрессионного уравнения (R^2 и $adjR^2$).
7. Оценки статистической значимости выборочной регрессии целиком (критерий Фишера).

   (Если обнаружена незначимость коэффициентов регрессионного уравнения или модели целиком, а также, если коэффициенты качества модели неудовлетворительны, то делается корректировка регрессионного уравнения и анализ проводится снова)

8. Если модель принимается, то делается прогноз значений зависимой переменной.

Выбор вида уравнения регрессии

3 Предположение о регрессионной функции делается либо на основе графического вида экспериментальных данных, либо исходя из имеющихся теоретических соображений или предыдущих исследований аналогичных данных.

Наиболее простой вид — ***линейная регрессионная модель*** $f(x) = a+bx$

Многие нелинейные модели могут быть сведены к линейной с помощью соответствующей замены переменных.

Регрессионные модели с линейной структурой

Линеаризующие функциональные преобразования $(y^* = a^* + b^* x^*)$

Исходная зависимость $y = f(x)$	Преобразование переменных		Преобразование коэффициентов	
	y^*	x^*	a^*	b^*
$y = a + \frac{b}{x}$	y	$\frac{1}{x}$	a	b
$y = \frac{a}{b + x}$	$\frac{1}{y}$	x	$\frac{a}{b}$	$\frac{1}{a}$
$y = \frac{ax}{b + x}$	$\frac{1}{y}$	$\frac{1}{x}$	$\frac{b}{a}$	$\frac{1}{a}$
$y = \frac{x}{a + bx}$	$\frac{x}{y}$	x	a	b
$y = ab^x$	$\lg y$	x	$\lg a$	$\lg b$
$y = ax^b$	$\lg y$	$\lg x$	$\lg a$	b
$y = ae^{bx}$	$\ln y$	x	$\ln a$	b
$y = ae^{\frac{b}{x}}$	$\ln y$	1	$\ln a$	b
$y = a + bx^n$	y	x^n	a	b

Нахождение оценок для параметров θ выбранного уравнения регрессии

4

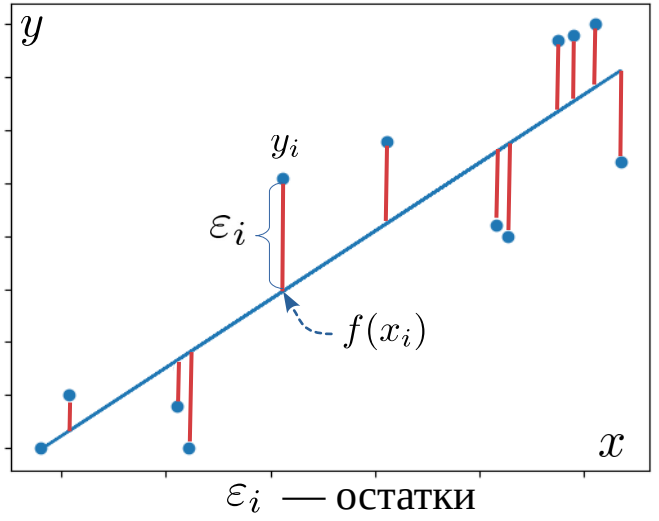
Точные значения для параметров θ регрессионной модели найти невозможно, но возможно определить оценки этих параметров $\hat{\theta}$ на основе имеющейся выборки.

Для определения оценок параметров регрессионной модели можно руководствоваться различными подходами. Наиболее естественный и распространенный — минимизировать совокупное отклонение $y_i - f(x_i,\hat{\theta})$ для $i = 1..n$. Меру близости можно выбирать по-разному (например, максимум модулей, сумма модулей и т. д.).

Наиболее простой считается

$$\Phi = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (f(x_i,\hat{\theta}) - y_i)^2 \underset{\hat{\theta}}{\rightarrow} \min$$

Такой метод называется:
методом наименьших квадратов (МНК)



ε_i — остатки

Для ***линейной*** функции $f(x) = a + xb$

a - intercept

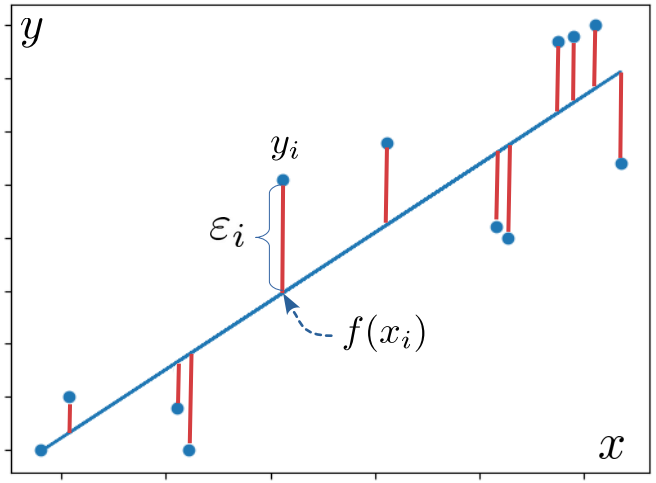
$$\begin{cases} \frac{\partial \Phi}{\partial \hat{a}} = 0 \\ \frac{\partial \Phi}{\partial \hat{b}} = 0 \end{cases} \Rightarrow \begin{cases} \hat{a}\frac{1}{n}\sum_{i=1}^{n} x_i + \hat{b} = \frac{1}{n}\sum_{i=1}^{n} y_i \\ \hat{a}\frac{1}{n}\sum_{i=1}^{n} x_i^2 + \hat{b}\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n} x_i y_i \end{cases} \Rightarrow \begin{array}{l} \hat{a} = \bar{y} - \hat{b}\bar{x} \\ \hat{b} = \dfrac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \dfrac{s_y}{s_x} r_{xy} \end{array}$$

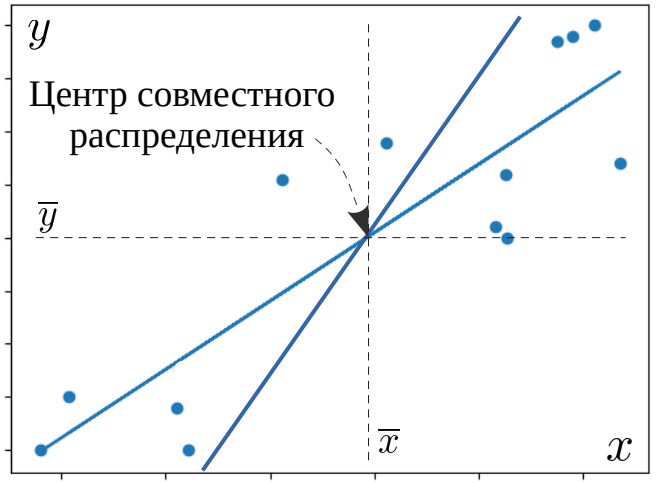
$$\overline{(*)} = \frac{\sum_{i=1}^{n} (*)_i}{n}$$

Полученные с помощью МНК оценки параметров регрессионной модели являются несмещенными и состоятельными

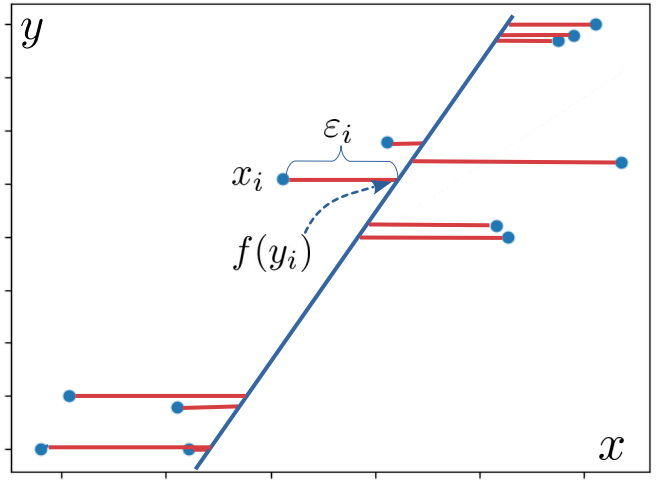
4 При регрессионном анализе исследователь сам определяет, какая их переменных будет независимой, а какая зависимой. Иногда этот выбор однозначен, иногда требуется проверка обратной зависимости.

регрессия *Y* по *X*





регрессия *X* по *Y*



В случае линейного уравнения регрессии по определенны:

$$y = \hat{a} + \hat{b}x \quad \Rightarrow \quad \hat{a} = \bar{y} - b\bar{x}, \quad \hat{b} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} \qquad \text{и} \qquad x = \hat{a} + \hat{b}y \quad \Rightarrow \quad \hat{a} = \bar{x} - b\bar{y}, \quad \hat{b} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{y^2} - \bar{y}^2}$$

Параметры *a* и *b* можно выразить через выборочный коэффициент корреляции:

$$\hat{r}_{xy} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2}\sqrt{\overline{y^2} - \bar{y}^2}}$$

$\hat{r}_{xy} = 0 : \Rightarrow$ прямые регрессии перпендикулярны

$\hat{r}_{xy} = \pm 1 : \Rightarrow$ прямые регрессии совпадают

Тогда, $\hat{b} = \hat{r}_{xy}\frac{s_y}{s_x}$ $\quad f(x) = \bar{y} - \hat{r}_{xy}\frac{s_y}{s_x}\bar{x} + \hat{r}_{xy}\frac{s_y}{s_x}x$ $\quad$ и $\quad \hat{b} = \hat{r}_{xy}\frac{s_x}{s_y}$ $\quad g(y) = \bar{x} - \hat{r}_{xy}\frac{s_x}{s_y}\bar{y} + \hat{r}_{xy}\frac{s_x}{s_y}y$

Здесь $s_x = \sqrt{\overline{x^2} - \bar{x}^2}, \quad s_y = \sqrt{\overline{y^2} - \bar{y}^2}$

Пример определения оценок для параметров θ нелинейных уравнений регрессии, которые сводятся к линейным

Для нелинейной функции $y = (a + bx)^2 \;\Rightarrow\; \sqrt{y} = a + bx$

$$\begin{cases} \frac{\partial \Phi}{\partial a} = 0 \\ \frac{\partial \Phi}{\partial b} = 0 \end{cases} \Rightarrow \begin{cases} a\frac{1}{n}\sum\limits_{i=1}^{n} x_i + b = \frac{1}{n}\sum\limits_{i=1}^{n}\sqrt{y_i} \\ a\frac{1}{n}\sum\limits_{i=1}^{n} x_i^2 + b\frac{1}{n}\sum\limits_{i=1}^{n} x_i = \frac{1}{n}\sum\limits_{i=1}^{n} x_i\sqrt{y_i} \end{cases}$$

Для показательной функции $y = \exp(a + bx) \quad \Rightarrow \quad \ln y = a + bx$

$$\begin{cases} \frac{\partial \Phi}{\partial a} = 0 \\ \frac{\partial \Phi}{\partial b} = 0 \end{cases} \Rightarrow \begin{cases} a\frac{1}{n}\sum\limits_{i=1}^{n} x_i + b = \frac{1}{n}\sum\limits_{i=1}^{n}\ln y_i \\ a\frac{1}{n}\sum\limits_{i=1}^{n} x_i^2 + b\frac{1}{n}\sum\limits_{i=1}^{n} x_i = \frac{1}{n}\sum\limits_{i=1}^{n} x_i\ln y_i \end{cases}$$

Пример: Найти параметры зависимости между x и y для выборки

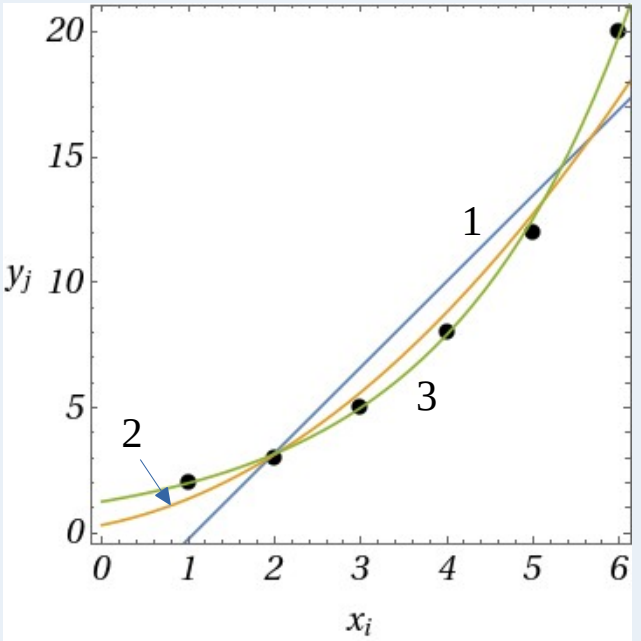
							$\sum_{i=1}^{6}$	$\frac{1}{n}\sum_{i=1}^{6}$
x_i	1	2	3	4	5	6	21	3,5
y_i	2	3	5	8	12	20	50	8,33
x_i^2	1	4	9	16	25	36	91	15,17
y_i^2	4	9	25	64	144	400	646	107,7
$x_i y_i$	2	6	15	32	60	120	235	39,17
$\sqrt{y_i}$	1,41	1,73	2,24	2,83	3,46	4,47	16,15	2,69
$x_i\sqrt{y_i}$	1,41	3,46	6,72	11,32	17,3	26,82	67,03	11,18
$\ln y_i$	0,69	1,10	1,61	2,08	2,48	3,00	10,96	1,83
$x_i \ln y_i$	0,69	2,20	4,83	8,32	12,4	18	46,44	7,74

Рассмотрим:

1) $y = ax + b$
2) $y = (ax + b)^2$
3) $y = e^{ax + b}$

В результате:

1) Для линейной зависимости $a = 3.43$ и $b = -3.67$;
2) Для квадратичной зависимости $a = 0.6$ и $b = 0.57$;
3) Для показательной зависимости $a = 0.46$ и $b = 0.23$;



y_i	Номер на рис.	2	3	5	8	12	20	$\sum(y_i - y_i^*)^2$
$(y_i)_{лин}$	1	–0,24	3,19	6,62	10,05	13,48	16,91	23,62
$(y_i)_{кв}$	2	1,37	3,13	5,62	8,82	12,75	17,40	8,79
$(y_i)_{показ}$	3	1,99	3,16	5,00	7,93	12,55	19,87	0,35

5 Доверительные интервалы для параметров *a* и *b* уравнения линейной регрессии

Полученные с помощью МНК оценки параметров регрессионной модели являются несмещенными, эффективными и состоятельными, которые принадлежат нормальному закону распределения:

$$\hat{b} \in N\left(M[\hat{b}], \sqrt{D[\hat{b}]}\right), \quad M[\hat{b}] = b, \quad D[\hat{b}] = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{a} \in N\left(M[\hat{a}], \sqrt{D[\hat{a}]}\right), \quad M[\hat{a}] = a, \; D[\hat{a}] = \frac{\sigma^2}{n} + \bar{x}^2 D[\hat{b}]$$

Определение *доверительных интервалов* для параметров *a* и *b* соответствует определению доверительных интервалов для математического ожидания случайных величин $\hat{a}$ и $\hat{b}$ при неизвестном среднеквадратичном отклонении σ^2.

Как рассчитывается оценка s_e для неизвестного среднеквадратичного отклонении σ^2 смотри через один слайд.

$\hat{a} - \varepsilon_a \leqslant a \leqslant \hat{a} + \varepsilon_a$, где $\varepsilon_a = \frac{s_e}{\sqrt{n}} F^{-1}_{T(n-2)}\left(\frac{\gamma}{2}\right)$ — Стандартная ошибка параметра *a*

$\hat{b} - \varepsilon_b \leqslant b \leqslant \hat{b} + \varepsilon_b$, где $\varepsilon_b = \frac{s_e}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} F^{-1}_{T(n-2)}\left(\frac{\gamma}{2}\right)$ — Стандартная ошибка параметра *b*

γ -- доверительная вероятность

5 Оценка значимости для параметра *b* уравнения линейной регрессии == Насколько статистически значима взаимосвязь переменных?

Может оказаться, что фактор x не влияет на результат y, что эквивалентно $b = 0$, однако при этом выборочный коэффициент $\hat{b}$, вообще говоря отличен от нуля.

Другими словами, это делается для того, чтобы понять, **можем ли** мы нашу регрессионную модель, **рассчитанную на основе выборки**, применять ко **всей генеральной совокупности**?
Т.к. на основании только ограниченных данных выборки мы по умолчанию не можем утверждать, сработает ли это для всей генеральной совокупности

Поэтому необходимо проверить *гипотезу о значимости коэффициентов уравнения регрессии*:

Для линейного уравнения: $H_0 : b = 0,\ H_1 : b \neq 0$

Используется критерий Стьюдента: $t = \dfrac{\hat{b} - b}{s_b} = \dfrac{\hat{b}}{s_b} \in T_{n-l}$

Определяется p-value, которое затем сравнивается с уровнем значимости α

Проверка данной гипотезы эквивалентна задаче определения доверительных интервалов для параметров модели.

Если 0 входит в доверительный интервал для параметра *b*, то гипотеза принимается и коэффициент наклона считается незначимым. В противоположном случае, гипотеза отвергается, коэффициент значим.

5 Проверка значимости (пригодности) уравнения линейной регрессии

В соотношения для доверительных интервалов *a* и *b* параметров регрессионного уравнения входит переменная s_e — оценка для стандартного отклонения случайных отклонений (ошибок наблюдения) от регрессионной модели. **Как ее определять?**

Для линейного уравнения регрессии есть точное математическое выражение, которое получается из следующих рассуждений:

Как упоминалось ранее в классической модели регрессионного анализа *предполагается*, что ε_i , i =1...n :
а) независимые *б*) имеют одинаковый закон распределения $N(0, \sigma^2)$.

Тогда

$$\sum_{i=1}^{n} \frac{\varepsilon_i^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(y_i - f(x_i, \theta))^2}{\sigma^2} = \chi^2 \in \mathrm{H}^2(n-l),$$

здесь l — количество параметров в регрессионной модели (для линейной модели l = 2)

Поскольку известно, что математическое ожидание χ^2 равно числу степеней свободы, то

$$M[\chi^2] = (n-l), \Rightarrow M\left[\sum_{i=1}^{n} \frac{\varepsilon_i^2}{(n-l)}\right] = M\left[\sum_{i=1}^{n} \frac{(y_i - f(x_i, \theta))^2}{(n-l)}\right] = \sigma^2$$

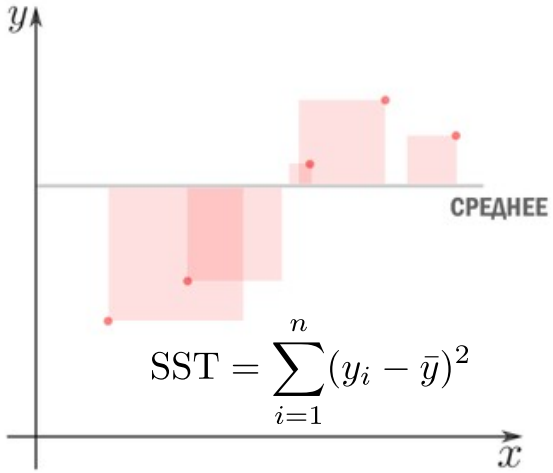
Условие для несмещенности оценки

Таким образом, хотя мы и не знаем истинное значение σ^2 , но можем для него получить точечную несмещенную оценку.

$$\widehat{\sigma^2} \equiv s_e^2 = \frac{\sum_{i=1}^{n} (y_i - f(x_i, \hat{\theta}))^2}{(n-l)}$$

Различные виды вариаций (сумм квадратов отклонений) и дисперсий регрессионных моделей

SST (Sum of Square **Total**)
общая вариация исходных данных
s^2 -- общая дисперсия

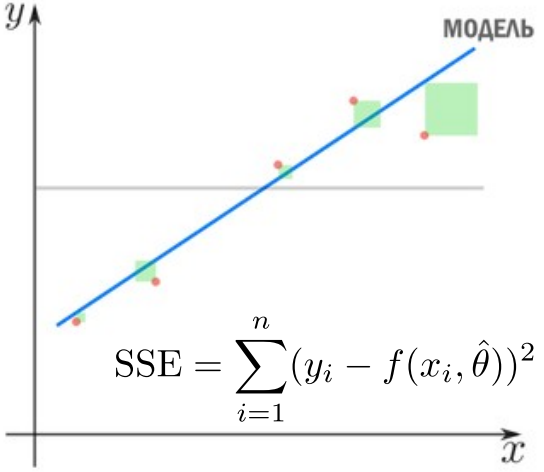


$$\mathrm{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{(n-1)}$$

SST - характеризует случайную ошибку для всей выборки, т. е. оценивает несоответствие между конкретными (текущими) значениями результата эксперимента и средним арифметическим.

SSE (Sum of Square **Error**)
вариация ошибок (остатков),
s_e^2 — дисперсия ошибок (остатков)

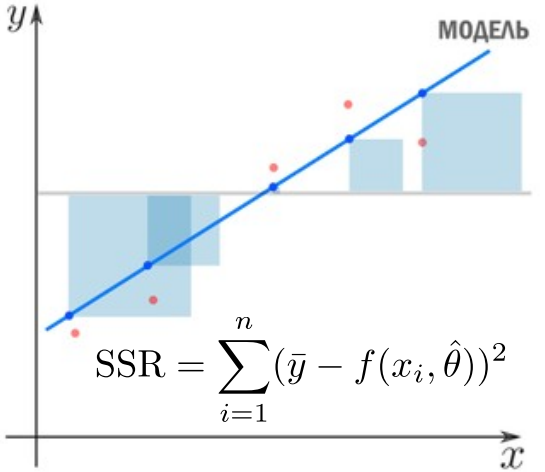


$$\mathrm{SSE} = \sum_{i=1}^{n}(y_i - f(x_i, \hat{\theta}))^2$$

$$s_e^2 = \frac{\sum_{i=1}^{n}(y_i - f(x_i, \hat{\theta}))^2}{(n-l)}$$

SSE - характеризует величину среднего разброса экспериментальных точек относительно линии(Sum of Square Error) регрессии. Позволяет оценить ошибку, с которой уравнение регрессии предсказывает фактический результат.

SSR (Sum of Square **Regression**)
вариация регрессии (фактора),



$$\mathrm{SSR} = \sum_{i=1}^{n}(\bar{y} - f(x_i, \hat{\theta}))^2$$

$$s_r^2 = \frac{\sum_{i=1}^{n}(\bar{y} - f(x_i, \hat{\theta}))^2}{(l-1)}$$

Общая сумма квадратов = «необъясненная» сумма квадратов + «объясненная» сумма квадратов

$$\boxed{\mathrm{SST} = \mathrm{SSE} + \mathrm{SSR}}$$

* В литературе можно встретить другие обозначения для аналогичных понятий (с противоположными буквами):
ESS (**Explained** Sum of Square) = **SSR** (Sum of Square **Regression**)
RSS (**Residual** Sum of Square) = **SSE** (Sum of Square **Error**)

здесь l — количество параметров в регрессионной модели (для линейной модели $l = 2$)

6 Качество выбранной регрессионной модели

Помимо проверки каждого коэффициента модели важно знать, насколько она хороша в целом.

В качестве меры того, насколько хорошо регрессия описывает изучаемую систему, служит **коэффициент детерминации =** доля «объясненной» дисперсии -- это отношение объясненной части вариации ко всей вариации в целом.

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = \frac{\mathrm{SSR}}{\mathrm{SSR} + \mathrm{SSE}} \qquad R^2 \in [0; 1]$$

Чем меньше SSE, тем «ближе» этот коэффициент к 1, и тем лучше описание.

Возможные проблемы: Проблемы с использованием R^2 заключаются в том, что его значение не уменьшается при добавлении в уравнение факторов, сколь плохи бы они ни были.

Он гарантированно будет равен 1, если мы добавим в модель столько факторов, сколько у нас наблюдений. Поэтому сравнивать модели с разным количеством факторов, используя R^2, не имеет смысла.

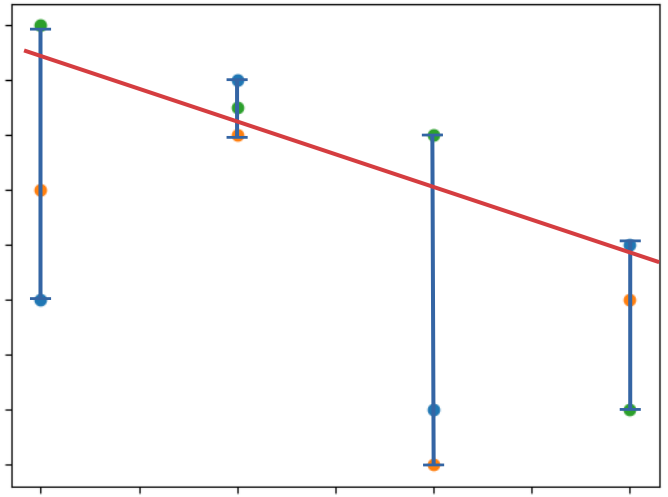
Для более адекватной оценки модели используется скорректированный коэффициент детерминации (adj R^2). Как видно из названия, этот показатель представляет собой скорректированную версию R^2, накладывая «штраф» за каждый добавленный фактор:

$$\mathrm{adj}.R^2 = 1 - (1 - R^2)\frac{N - 1}{N - l + 1}$$

здесь l — количество параметров в регрессионной модели (для линейной модели l = 2)

7 Проверка значимости влияния фактора на отклик

Имеющиеся у нас данные y_i — это только реализация выборки. И если бы мы могли провести опыт еще несколько раз, то получили бы всегда разные результаты. Т.е. на графике это бы выглядело примерно так:



Можно построить регрессионную модель и получить какие-то ненулевые коэффициенты, но поскольку это только выборочные данные, то надо проверить гипотезу о том, что на самом деле коэффициенты формы регрессионной модели равны нулю (для линейной b=0).

Если это так, то $f(x)=a$ и, следовательно фактор не влияет на отклик (т. е. x не влияет на y), а ненулевые параметры регрессионной модели связаны только с погрешностями (дисперсией) измерения отклика, а не с влиянием фактора.

Проверяется гипотезу о том, что все факторы одновременно (кроме константы) являются незначимыми:

$$H_0 : b_i = 0,\ i = 1..(l-1), \quad H_1 : \text{хотя бы один коэффициент не равен нулю}$$

Для количественной проверки данной гипотезы используется **F-критерий** (критерий Фишера):

$$F = \frac{s_r^2}{s_e^2} = \frac{\text{SSR}}{\text{SSE}} \cdot \frac{(n-l)}{(l-1)}$$

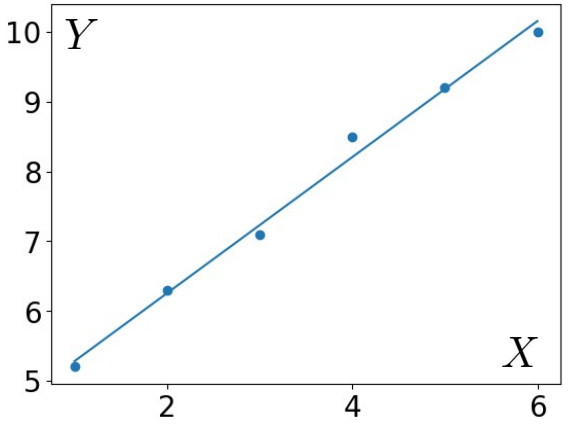
$$F_{fisher}(k_1, k_2), \quad k_1 = l-1,\ k_2 = n-l$$

l — количество параметров в регрессионной модели (для линейной модели l = 2)

Значение F-статистики также сравнивают с критическими значениями заданного уровня значимости α (или уровень значимости сравнивается с p-value). Если F>Fкр., то гипотеза принимается и считается, что фактор не влияет на отклик.

П р и м е р: предполагается линейная зависимость переменных *X* и *Y*.
Установить значимость линейного уравнения регрессии с надежностью 0.05.

X	Y	f(X) = aX +b	(f(X)-Y)**2	(Y - mean(Y))**2
1	5.2	5.280952	0.006553	6.333611
2	6.3	6.255238	0.002004	2.006944
3	7.1	7.229524	0.016776	0.380278
4	8.5	8.203810	0.087729	0.613611
5	9.2	9.178095	0.000480	2.200278
6	10.0	10.152381	0.023220	5.213611



Из МНК: $y = 0.974x + 4.307$

$n = 6,\ l = 2$

$$s_e^2 = \frac{\sum_{i=1}^{n}(f(x_i,\hat{\theta}) - y_i)^2}{n-l} \approx 0.034$$

$$s_r^2 = \frac{\sum_{j=1}^{n}(f(x,\hat{\theta}) - \bar{y})^2}{l-1} \approx 16.612$$

$$\Rightarrow\ F_{cal} = \frac{s_r^2}{s_e^2} = 485.854 > F_{kp} = F^{-1}_{2-1,6-2}(1-0.05) = 7.708$$

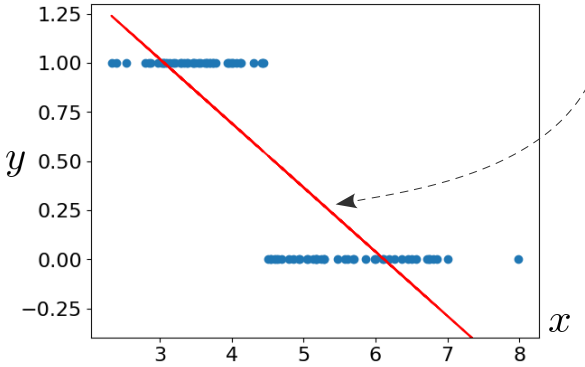
Вывод: линейное уравнение значимо, т. е. фактор *X* значимо влияет на отклик *Y*.

Логистическая регрессия (статистический метод классификации)

зависимая переменная
(категориальная с двумя категориями)

независимая
переменная
(категориальные, порядковые, количественные)

Представим экспериментальные данные на графике и построим для них обычную линейную регрессию $y = f(x) = ax + b$:



Однако такое уравнение будет давать предсказания y не только в диапазоне от [0,1]

В отличие от обычной регрессии, в методе логистической регрессии регрессионное уравнение $f(x)$ применяется не для предсказание значения отклика исходя из выборки исходных значений. Вместо этого, значением функции является вероятность $p(x)$ того, что данное исходное значение принадлежит к определенному классу, которая задается в виде **сигмоиды**. Далее, на основе полученной функции вероятностей $p(x)$ строится классификатор (0; 1).

$$y = ax + b \;\Rightarrow\; p(x) = \frac{1}{1+\exp(-y)} \;\Rightarrow\; \begin{cases} 1, & \text{if } p(x) \geqslant 0.5 \\ 0, & \text{if } p(x) < 0.5 \end{cases}$$

Для оценки качества построенной модели исходная выборка разделяется на части: обучающую (trained) и тестовых (tested). Параметры сигмоиды определяются по обущающей выборке. Далее по данным тестовой выборки строятся предсказания с помощью полученной сигмоиды, которые сравниваются с тестовыми значениями. По количеству ошибок определяется качество полученного классификатора