# Прикладная статистика
## Слайды к лекции 8

С. А. Спирин

7 марта 2023

# 1. Интервальное оценивание

Доверительные интервалы

# Доверительные интервалы (confidence intervals)

Предположим, что выборка $x_1, \ldots, x_n$ представляет распределение, зависящее от неизвестного параметра $\theta$.

**Задача:** по выборке $x_1, \ldots, x_n$ найти **два** числа $\theta'$ и $\theta''$ (границы доверительного интервала) такие, что при любом значении параметра $\theta$ вероятность того, что $\theta$ попадёт внутрь интервала $(\theta', \theta'')$ больше, чем $1 - \alpha$ (где $\alpha$ — заранее заданная маленькая вероятность, например $\alpha$ = 0,05).

Вероятность нужно понимать правильно: в рамках стандартной мат. статистики $\theta$ — константа (то есть **неслучайное** число), а вот $\theta'$ и $\theta''$ — случайные величины.

# Симметричные и односторонние доверительные интервалы

Доверительный интервал не определён однозначно.

Обычно строят либо **симметричные**, либо **левосторонние**, либо **правосторонние** доверительные интервалы.

Симметричный доверительный интервал ($\theta'$, $\theta''$) означает, что вероятность того, что $\theta < \theta'$, равна вероятности того, что $\theta'' < \theta$ (и равна $\alpha/2$)

У левостороннего интервала левая граница — минус бесконечность (или левая граница допустимых значений $\theta$), то есть определяется правая граница такая, что P($\theta'' < \theta$) = $\alpha$.

Правосторонний интервал — наоборот, имеет только левую границу $\theta'$ такую, что P($\theta < \theta'$) = $\alpha$.

# Симметричный дов. интервал из точечной оценки

Начинаем с точечной оценки $\theta^* = \theta^*(x_1, \ldots, x_n)$. Это случайная величина, чьё распределение зависит от $\theta$. Обозначим её функцию распределения через $F_\theta$ .

Теперь обозначим:
$a(\theta) = F^{-1}{}_\theta(\alpha/2)$
$b(\theta) = F^{-1}{}_\theta(1 - \alpha/2)$
Таким образом, с вероятностью $1 - \alpha$ имеем $a(\theta) < \theta^* < b(\theta)$.

Теперь $\theta'$ и $\theta''$ определяются из уравнений $\theta^* = b(\theta')$ и $\theta^* = a(\theta'')$.

# Симметричный дов. интервал

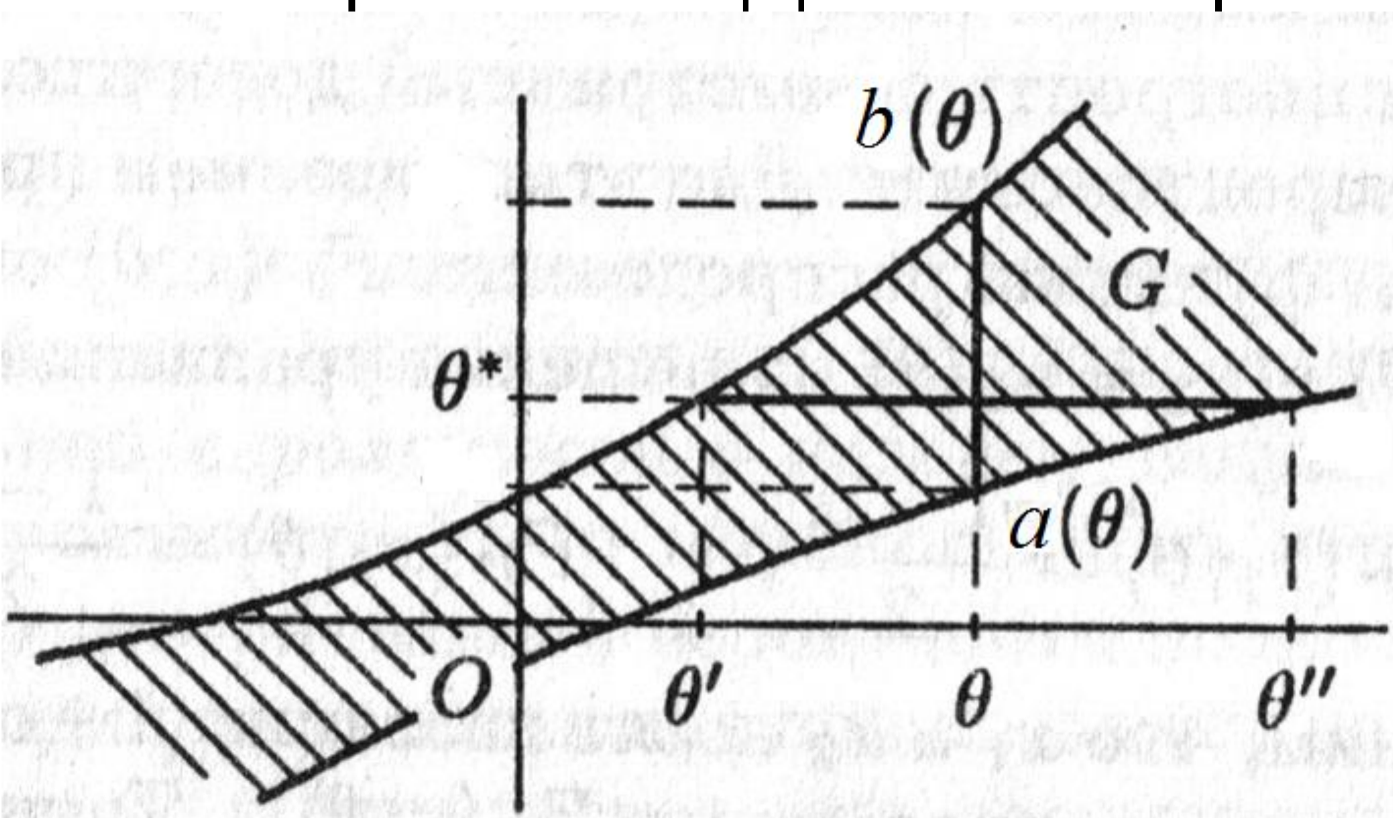



Рис. адаптирован из книги:
П.П.Бочаров, А.В.Печинкин «Теория вероятностей. Математическая статистика» Москва, 1998.

Вероятности того, что реальное значение параметра $\theta$ окажется слева и справа от симметричного доверительного интервала, равны между собой.

# Односторонние доверительные интервалы

Бывает, что нужно найти **одно** число $\theta'$ такое, что при любом значении параметра $\theta$ вероятность того, что $\theta > \theta'$, больше, чем $1 - \alpha$ (нижняя доверительная граница)
или число $\theta''$ такое, что при любом значении $\theta$ вероятность того, что $\theta < \theta''$, больше, чем $1 - \alpha$ (верхняя доверительная граница).

Опять начинаем с точечной оценки $\theta^* = \theta^*(x_1, \ldots, x_n)$, и её функции распределения через $F_\theta$ . Обозначим:
$c(\theta) = F^{-1}{}_\theta(\alpha)$
$d(\theta) = F^{-1}{}_\theta(1 - \alpha)$
Таким образом, с вероятностью $1 - \alpha$ имеем $c(\theta) < \theta^*$ и с такой же вероятностью $\theta^* < d(\theta)$.

Нижняя доверительная граница определяется из уравнения $\theta^* = d(\theta')$, а верхняя — из уравнения $\theta^* = c(\theta'')$.

Из двух в реальности нужна какая-нибудь одна

# Доверительный интервал для среднего

Пусть нам откуда-нибудь известно истинное значение дисперсии $\sigma^2$, но не известно значение математического ожидания $\theta$.

Оценка среднего: $\theta^* = \bar{x} = (x_1 + \ldots + x_n)/n$.

Если исходное распределение нормальное, то $\bar{x}$ распределено нормально со средним $\theta$ и дисперсией $\sigma^2/n$.
(А если исходное распределение другое, но $n$ достаточно велико, то распределение $\bar{x}$ всё равно очень близко к нормальному с такими же параметрами)

Иными словами, $F_\theta(x) = F(\sigma x/n^{1/2} + \theta)$, где $F$ – функция распределения стандартного нормального распределения N(0;1).

Поэтому (обозначим $F^{-1}$ через $\varphi$):
$a(\theta) = F^{-1}{}_\theta(\alpha/2) = \sigma\,\varphi(\alpha/2)/n^{1/2} + \theta$
$b(\theta) = F^{-1}{}_\theta(1 - \alpha/2) = \sigma\,\varphi(1 - \alpha/2)/n^{1/2} + \theta$

Например, для $\alpha = 0{,}05$ имеем $a(\theta) = \theta - 1{,}96\,\sigma/n^{1/2}$ и $b(\theta) = \theta + 1{,}96\,\sigma/n^{1/2}$.

# Доверительный интервал для среднего (продолжение)

Границы доверительного интервала $\theta'$ и $\theta''$ находим из уравнений:

$\theta^* = \theta' + \sigma\, \varphi(1 - \alpha/2)/n^{1/2}$

$\theta^* = \theta'' - \sigma\, \varphi(1 - \alpha/2)/n^{1/2}$

Откуда $\theta' = \bar{x} - \sigma\, \varphi(1 - \alpha/2)/n^{1/2}$ и $\theta'' = \bar{x} + \sigma\, \varphi(1 - \alpha/2)/n^{1/2}$

Иными словами, симметричный доверительный интервал $(\theta', \theta'')$ для среднего содержит точечную оценку среднего $\theta^* = \bar{x}$ в своей середине, а половина его длины равна соответствующему квантилю стандартного нормального распределения, умноженному на **стандартную ошибку** $\sigma/n^{1/2}$.

Если $\sigma$ неизвестно, то вместо него используют $s$ — квадратный корень из $s^2 = \sum(x_i - \bar{x})^2/(n-1)$ , а вместо квантилей стандартного нормального распределения берут квантили распределения Стьюдента с $n - 1$ степенью свободы.

# Согласованность гипотез и доверительных интервалов

Если рассмотреть равенство $\theta = 0$ в качестве $H_0$ против двусторонней альтернативы $\theta \neq 0$ , то критерием для непринятия $H_0$ может служить непопадание нуля в доверительный интервал.

То, что получается, совпадает с критерием Стьюдента.

# Доверительный интервал для разности средних

Есть две выборки из нормальных распределений с неизвестными средними и равными дисперсиями $x_1, \ldots, x_m$ и $y_1, \ldots, y_n$.

Нас интересует разность их средних.

Оцениваем средние $\overline{x}$ и $\overline{y}$ и общую дисперсию:
$s^2 = (\sum(x_i - \overline{x})^2 + \sum(y_i - \overline{y})^2) / (m + n - 2)$

Границы доверительного интервала

$\theta' = \overline{x} - \overline{y} - s \cdot (1/m + 1/n)^{½} T^{-1}{}_{m+n-2}(1 - \alpha/2)$

$\theta'' = \overline{x} - \overline{y} + s \cdot (1/m + 1/n)^{½} T^{-1}{}_{m+n-2}(1 - \alpha/2)$

где $T_{m+n-2}(x)$ — функция распределения Стьюдента с $m + n - 2$ степенями свободы.

# Доверительный интервал для дисперсии

Факт (теорема Фишера): если $x_1, \ldots, x_n$— выборка из нормального распределения с дисперсией $\sigma^2$, то величина

$(n-1)\, s^2/\, \sigma^2$ (где $s^2 = \sum(x_i - \overline{x})^2/(n-1)$) имеет распределение хи-квадрат с $n-1$ степенью свободы.

Отсюда путём несложных выкладок (см. слайд 5) выводятся следующие границы симметричного доверительного интервала для дисперсии:

$\theta' = \sum(x_i - \overline{x})^2/\, C^{-1}{}_{n-1}(1-\alpha/2)$ , $\theta'' = \sum(x_i - \overline{x})^2/\, C^{-1}{}_{n-1}(\alpha/2)$

где $C_{n-1}(x)$ — функция распределения хи-квадрат с $n-1$ степенью свободы.

Как видно, в данном случае точечная оценка $s^2$ уже не является серединой симметричного доверительного интервала!

Если (вдруг) среднее заранее известно, то можно использовать его вместо $\overline{x}$ и хи-квадрат с $n$ вместо $n-1$ степенями свободы — интервал будет меньше при той же надёжности.

# Доверительный интервал для частоты успеха

Имеется $n$ испытаний с неизвестной вероятностью успеха $\theta$.

Считаем, что $n$ достаточно велико, чтобы оценка $\theta^* = (x_1 + \ldots + x_n)/n$ была распределена нормально со средним $\theta$ и дисперсией $\theta(1-\theta)/n$.

Тогда

$$a(\theta) = \theta - \varphi(\alpha/2)\,(\theta(1-\theta)/n)^{1/2}$$

$$b(\theta) = \theta + \varphi(\alpha/2)\,(\theta(1-\theta)/n)^{1/2}$$

Тем самым нужно решить (относительно $\theta'$ и $\theta''$) уравнения

$$\theta^* = \theta'' - \varphi(\alpha/2)\,(\theta''(1-\theta'')/n)^{1/2}$$

$$\theta^* = \theta' + \varphi(\alpha/2)\,(\theta'(1-\theta')/n)^{1/2}$$

Что равносильно

$$(\theta'' - \theta^*)^2 = \varphi(\alpha/2)^2\,\theta''(1-\theta'')/n$$

$$(\theta^* - \theta')^2 = \varphi(\alpha/2)^2\,\theta'(1-\theta')/n$$

Это уже квадратные уравнения.
Точные выражения для $\theta'$ и $\theta''$ выглядят достаточно сложно.

# Пример 1

У 10 мышей измеряли уровень потребления кислорода до и после воздействия неким препаратом. После воздействия получили следующие показатели (в процентах от исходного уровня): 102%, 109%, 107%, 103%, 104%, 111%, 104%, 104%, 109%, 107%.

1. Предполагая, что относительное изменение этой величины распределено нормально, найти границы симметричного доверительного интервала для среднего значения такого изменения на уровне доверия $\alpha = 0{,}05$.
2. Найти левую границу правостороннего доверительного интервала для среднего значения (среднее изменение не меньше такого-то) на том же уровне доверия
3. Как найти границы симметричного доверительного интервала для дисперсии такого изменения?

# Пример 1, вопрос 1

## симметричный доверительный интервал для среднего

Значения: 102%, 109%, 107%, 103%, 104%, 111%, 104%, 104%, 109%, 107%,  всего 10.

Среднее десяти значений: $m$ = 106%

Складываем квадраты разностей значений и среднего и делим на 9, получаем оценку дисперсии: $s^2 = 0{,}00091$ .
Извлекая корень, получаем стандартное отклонение: s = 3,02%

Делим на квадратный корень из 10, получаем стандартную ошибку: SE = 0,955%

Уровень значимости $\alpha = 0{,}05$,
функция распределения Стьюдента с 9 степенями свободы принимает значение $\alpha/2 = 0{,}025$ при аргументе $-t = -2{,}26$ (соответственно значение $1 - \alpha/2$ при $t = 2{,}26$).

Для получения левой границы симметричного доверительного интервала нужно умножить $t$ на SE и результат вычесть из $m$, а для получения правой — прибавить к $m$.

Доверительный интервал: ($m - t{\cdot}$SE, $m + t{\cdot}$SE) = (103,8%; 108,2%)

# Пример 1, вопрос 2

## правосторонний доверительный интервал для среднего

Все действия те же, но теперь находим такое $t$, при котором функция распределения Стьюдента с 9 степенями свободы принимает значение $1 - \alpha$, это $t = 1{,}83$.
Для получения левой границы одностороннего доверительного интервала нужно умножить такое $t$ на SE и результат вычесть из $m$.

Доверительный интервал: $(m - t \cdot \mathrm{SE}, \infty) = (104{,}25\%; \infty)$

# Пример 1, вопрос 3

## симметричный доверительный интервал для дисперсии

Если реальное значение дисперсии равно $\theta$, то точечная оценка дисперсии (то есть $s^2$) по 10 наблюдениям распределена как $\chi^2 \cdot \theta / 9$, где $\chi^2$ распределена по хи-квадрат с 9 степенями свободы.

Пусть $F$ — функция распределения $\chi^2$. Найдём числа A и B такие, что $F(A) = \alpha/2$ и $F(B) = 1 - \alpha/2$. Это A = 2,7 и B = 19,02

Значит, значение $\chi^2$ с вероятностью $1 - \alpha$ попадает в интервал (2,7; 19,02), с вероятностью $\alpha/2$ оказывается левее его и с той же вероятностью $\alpha/2$ правее.

Отсюда следует, что оценка $s^2$ с вероятностью $\alpha$ попадает в интервал $(2{,}7\ \theta / 9;\ 19{,}02\ \theta / 9)$, с вероятностью $\alpha/2$ оказывается левее его и с той же вероятностью $\alpha/2$ правее.

С вероятностью $\alpha/2$ величина $19{,}02\ \theta / 9$ окажется меньше $s^2$, и с той же вероятностью величина $2{,}7\ \theta / 9$ окажется больше $s^2$.
Равносильное утверждение: $P(\theta < 9s^2/ 19{,}02) = P(\theta > 9s^2/ 2{,}7) = \alpha/2$

Поэтому симметричный доверительный интервал: $(9s^2/ 19{,}02;\ 9s^2/ 2{,}7)$.

Подставляя $s^2 = 0{,}00091$, получаем симметричный доверительный интервал для дисперсии (0,0004; 0,003)

Для $\sigma$ это (2,08%; 5,51%). Заметим, что точечная оценка $s = 3{,}02\%$ не совпадает с серединой интервала.

# Пример 2

Исследовали заражённость моллюсков личиночной стадией печёночного сосальщика. Для этого проанализировали 100 экземпляров и у 33 нашли паразита.

1. Как найти симметричный доверительный интервал для процента заражённости?
2. Тот же вопрос для одностороннего доверительного интервала (наибольшее значение процента заражённости).

В обоих случаях полагаем уровень доверия $\alpha = 0{,}05$.

# Пример 2, вопрос 1
## симметричный доверительный интервал для доли

Начинаем с точечной оценки $\theta^* = (x_1 + \ldots + x_{100})/100 = 0{,}33$
Считаем, что она пришла из нормального распределения со средним $\theta$ и дисперсией $\theta(1-\theta)/100$
С вероятностью α/2 = 0,025 она могла оказаться левее, чем
$a(\theta) = \theta - 1{,}96\,(\theta(1-\theta)/100)^{1/2}$ (1,96 — это 2,5% перцентиль стандартного нормального распределения)
и с той же вероятностью правее, чем
$b(\theta) = \theta + 1{,}96\,(\theta(1-\theta)/100)^{1/2}$
Решаем уравнения:

$$\theta^* = \theta'' - 1{,}96\,(\theta''(1-\theta'')/100)^{1/2}$$

$$\theta^* = \theta' + 1{,}96\,(\theta'(1-\theta')/100)^{1/2}$$

которые равносильны уравнению:
$(x - \theta^*)^2 = (1{,}96)^2\, x(1-x)/100$ (меньший из корней — $\theta'$, больший — $\theta''$)
Подставляем $\theta^* = 0{,}33$ и приводим к стандартному виду:
$(1 + (1{,}96)^2/100)\, x^2 - (0{,}66 + (1{,}96)^2/100)\, x + (0{,}33)^2 = 0$
$1{,}038\, x^2 - 0{,}698\, x + 0{,}109 = 0$
$\theta', \theta'' = (0{,}698 \pm 0{,}188)/(2\cdot 1{,}038) = (0{,}265;\ 0{,}46)$
(опять-таки точечная оценка, то есть $\theta^*$, не является серединой симметричного дов. интервала)
Вывод: с надёжностью 95% заражённость находится в интервале от 26,5% до 46%.

# Пример 2, вопрос 2

## левосторонний доверительный интервал для доли

Правая граница левосторонного интервала получается из уравнения:

$0{,}33 = \theta'' - 1{,}645\,(\theta''\,(1-\theta'')/100)^{1/2}$ (1,645 — это 5% перцентиль стандартного нормального распределения)

Его решение — больший из двух корней уравнения:

$(x - 0{,}33)^2 = (1{,}645)^2\, x(1-x)/100$

Решая, получаем: $\theta'' = 0{,}433$

Вывод: с надёжностью 95% заражённость не выше 43,3% (как и должно быть, правая граница левостороннего интервала меньше правой границы симметричного двустороннего интервала)

*Такой подход (использование нормального распределения) для интервальной оценки доли можно использовать только если как число успехов, так и число неуспехов достаточно велико (более 10). При малом числе успехов нужен перебор возможных значений доли в генеральной совокупности и вычисление вероятности наблюдений, больших, равных и меньших данного, при каждом значении этой доли.*

# Пример 3

Из десяти испытаний в трёх наблюдался успех.
Каков 95% симметричный доверительный интервал для вероятности успеха?

**Решение**

Обозначим вероятность успеха через $\theta$.
Посчитаем для разных значений $\theta$ вероятность такого или меньшего, а также такого или большего числа успехов. Найдём значение $\theta'$ такое, что вторая вероятность близка к 2,5%, и значение $\theta''$ такое, что первая вероятность близка к 2,5%. Это и будут границы симметричного доверительного интервала.

| $\theta$ | $P(k \le 3 \mid \theta)$ | $P(k \ge 3 \mid \theta)$ |
|---|---|---|
| 0,065 | 0,9973 | 0,0233 |
| 0,066 | 0,9971 | 0,0243 |
| **0,067** | 0,9970 | **0,0253** |
| 0,068 | 0,9968 | 0,0263 |
| 0,069 | 0,9966 | 0,0273 |

| $\theta$ | $P(k \le 3 \mid \theta)$ | $P(k \ge 3 \mid \theta)$ |
|---|---|---|
| 0,65 | 0,0260 | 0,9952 |
| 0,651 | 0,0256 | 0,9953 |
| **0,652** | **0,0252** | 0,9954 |
| 0,653 | 0,0248 | 0,9955 |
| 0,654 | 0,0244 | 0,9956 |

**Ответ:** (0,067; 0,65)

# Пример 4

В трёх испытаниях не случилось ни одного успеха.
Каков 95% левосторонний доверительный интервал для вероятности успеха?

**Решение**
Обозначим вероятность успеха через $\theta$.
Посчитаем для разных значений $\theta$ вероятность нулевого числа успехов. Найдём значение $\theta''$ такое, что эта вероятность близка к 5%.

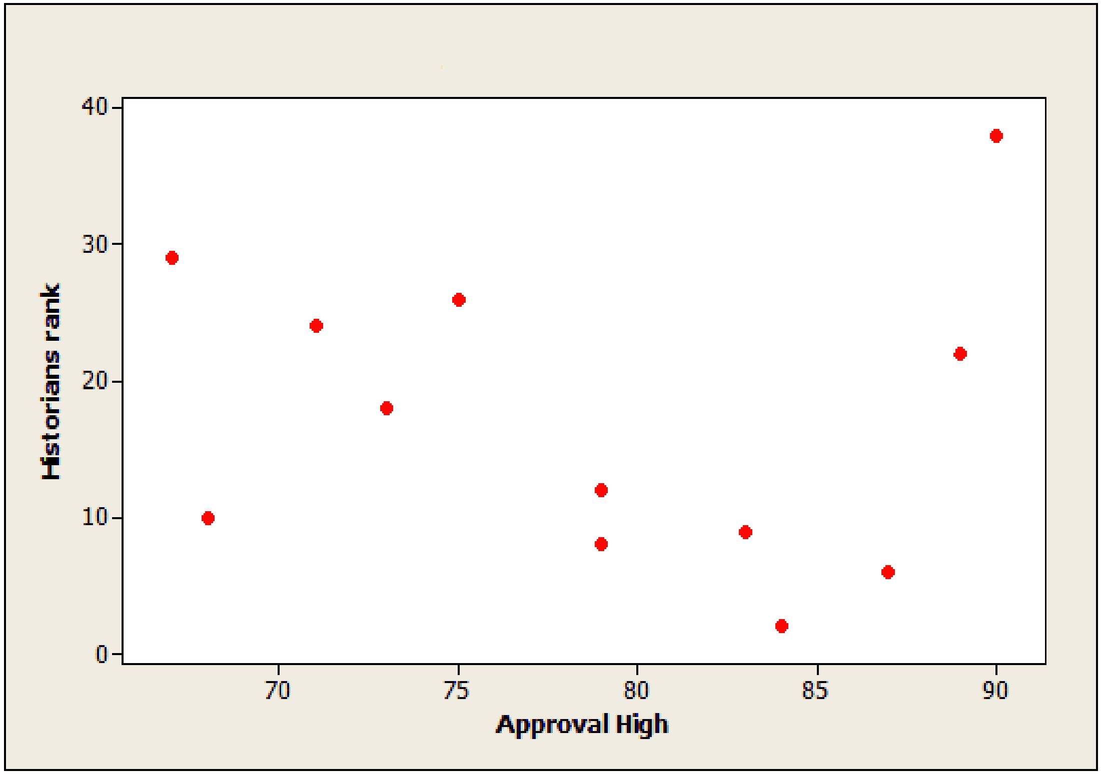
| $\theta$ | P($k$=3 \| $\theta$) |
|---|---|
| 0.61 | 0.059 |
| 0.62 | 0.055 |
| **0.63** | **0.051** |
| 0.64 | 0.047 |
| 0.65 | 0.043 |

**Ответ:** (0; 0,63)

# 2. Линейная регрессия

Статистические задачи (14 марта)

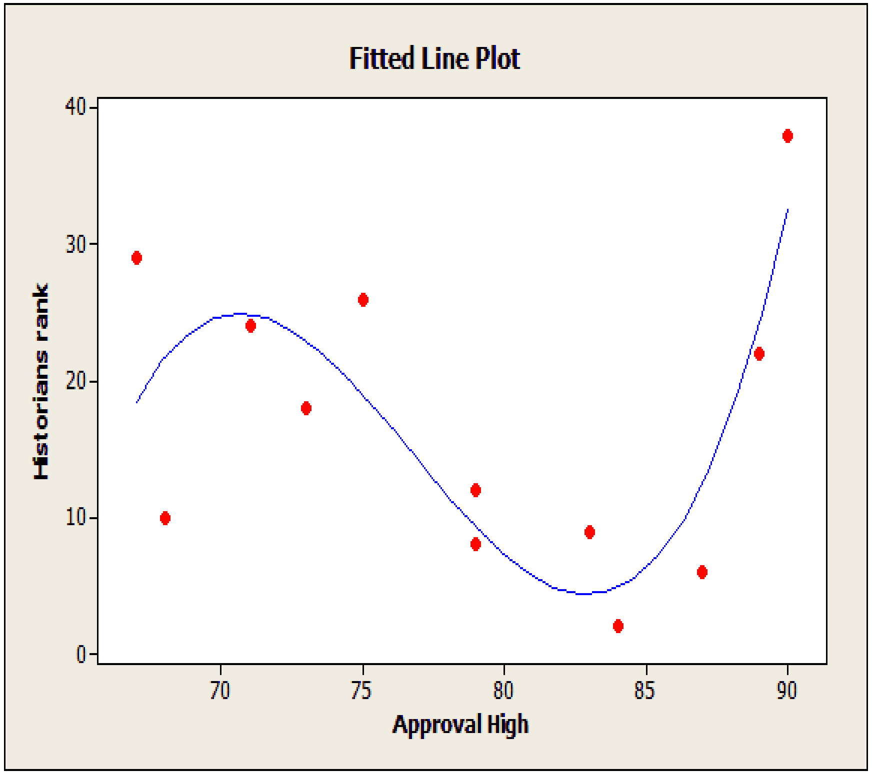
# Набор наблюдений при различных условиях




$x_1, \ldots, x_n$ — условия
$y_1, \ldots, y_n$ — наблюдения

https://blog.minitab.com/hubfs/Imported_Blog_Media/overfitlineplotnoequ-1.gif

# Регрессия

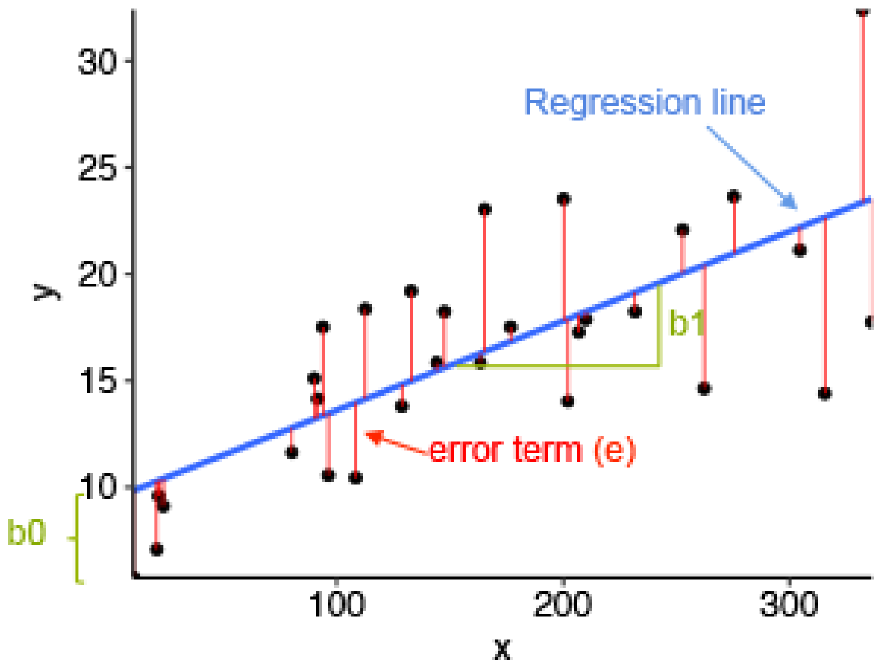



$x_1, \ldots, x_n$ — условия
$y_1, \ldots, y_n$ — наблюдения
$y = F(x) + \varepsilon$ — **модель**
здесь $F(x)$ — некоторая функция,
а $\varepsilon$ — «ошибка измерения» (случайная величина)
(например, если $F(x) = bx + a$, то это «линейная регрессия»)
$\hat{y}_i = F(x_i)$ — предсказанные значения.
$r_i = \hat{y}_i - y_i$ — невязки.

Функция $F$ может быть подобрана из разных соображений. Чаще всего стараются минимизировать сумму квадратов невязок.

https://blog.minitab.com/hubfs/Imported_Blog_Media/overfitlineplotnoequ-1.gif

# Линейная регрессия




$$y = bx + a + \varepsilon$$

https://autotis.ru/wp-content/uploads/2019/08/linear-regression.png

# Линейная регрессия

Пусть $x_1, \ldots, x_n; y_1, \ldots, y_n$ — две выборки чисел одинаковой длины.

Гипотеза состоит в том, что значения $y$ зависят от значений $x$ линейно с точностью до ошибки измерения, которая (ошибка) нормально распределена со средним 0 :

$y = bx + a + \varepsilon$

$\varepsilon \sim \mathrm{N}(0;\sigma^2)$

($\sigma$ неизвестно, но предполагается постоянным)

Наша задача — оценить $a$ и $b$.

Для этого минимизируем сумму квадратов невязок:

$Q(a, b) = \sum(bx_i + a - y_i)^2$

Приравняем $\partial Q/\partial a$ и $\partial Q/\partial b$ к 0, получим уравнения:

$$\begin{cases} \sum(bx_i + a - y_i) = 0 \\ \sum x_i(bx_i + a - y_i) = 0 \end{cases}$$

или:

$$\begin{cases} na = \sum y_i - b\sum x_i & \Rightarrow a = \overline{y} - b\overline{x} \\ b\sum x^2{}_i + a\sum x_i = \sum x_i y_i \end{cases}$$

# Линейная регрессия: решение

$$\hat{b} = \sum (x_i - \overline{x})(y_i - \overline{y}) / \sum (x_i - \overline{x})^2$$

$$\hat{a} = \overline{y} - \hat{b}\,\overline{x}$$

# Связь угла наклона и коэффициента корреляции

При линейной регрессии $\hat{b} = r\, s_y / s_x$

где $r$ — коэффициент корреляции между $y_1, \ldots, y_n$ и $x_1, \ldots, x_n$ ,

а $s_y$ и $s_x$ — стандартные отклонения выборок $y_1, \ldots, y_n$ и $x_1, \ldots, x_n$

# Доверительные интервалы для параметров линейной регрессии

Практически важные вопросы:

(1) значимо ли отличие угла наклона $b$ от нуля?

(2) каким может быть значение $y$ при данном значении $x$ ?
(прежде всего для таких $x$ , для которых не измерено $y$, то есть экстраполяция)

Стандартная ошибка для $b$ :

$$SE_b = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{(n-2)\sum_i (x_i - \bar{x})^2}}$$

Оценка для $b$ , центрированная на реальное значение $b$ и делённая на $SE_b$, распределена по Стьюденту с $n$ – 2 степенями свободы, поэтому симметричный доверительный интервал:

$$\left(\hat{b} - SE_b T^{-1}_{n-2}(1-\alpha/2), \hat{b} + SE_b T^{-1}_{n-2}(1-\alpha/2)\right)$$

где $T_{n-2}$ — функция распределения Стьюдента, α — уровень надёжности.

# Экстраполяция линейной регрессии

Пусть мы определили параметры линейной регрессии по наблюдениям $x_1, \ldots, x_n; y_1, \ldots, y_n$ — и пусть нам хочется предсказать значение $y$ для какого-либо нового $x$.

На самом деле тут **две разные** задачи:

(1) Оценить **среднее** значение $y$ при таком $x$

(2) Оценить границы доверительного интервала значений $y$ при таком $x$

Интуитивно понятно, что точечная оценка для среднего значения $y$ — это

$\hat{y} = \hat{a} + \hat{b}\, x$

Но вот границы доверительного интервала для самого $y$ и для его среднего разные.

Для среднего это

$$\hat{y} \pm T_{n-2}^{-1}(1-\alpha/2)\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-2}\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$$

А для самого $y$ :

$$\hat{y} \pm T_{n-2}^{-1}(1-\alpha/2)\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-2}\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}+1\right)}$$

# Задача

Эксклюзионная хроматография позволяет определять молекулярную массу вещества по скорости прохождения хроматографической колонки. При калибровке колонки тестовыми веществами с известной молекулярной массой были получены следующие значения:

| $M$ (kDal.) | $T$ (min) |
|---|---|
| 2.512 | 21.79 |
| 6.214 | 19.93 |
| 8.159 | 17.77 |
| 10.700 | 16.03 |
| 14.404 | 11.73 |
| 16.949 | 10.85 |
| 18.021 | 10.05 |

В каких пределах может находиться молекулярная масса вещества, чьё время прохождения равно 25 минутам?

# Множественная линейная регрессия

Всё то же самое, только вместо одного набора $x$ имеем несколько:
$x^{(1)}, \ldots, x^{(k)}$ (например, $k = 2$, $x^{(1)}$ — температура, $x^{(2)}$ — давление).

Ищем такие $a_0, a_1, \ldots, a_k$ , чтобы:
$y = \sum_{j=1}^{k} a_j x^{(j)} + a_0 + \varepsilon$
$\varepsilon \sim \mathrm{N}(0;\sigma)$

Решение аналогичное: составляем сумму квадратов невязок, рассматриваем её как функцию от коэффициентов $a_0, a_1, \ldots, a_k$ , находим точку минимума этой функции, что сводится к решению системы из $k$+1 линейного уравнения от $k$+1 неизвестной.

Получившаяся система хорошо (=устойчиво по отношению к небольшим вариациям входных данных) решается, если только между различными наборами $x$ нет сильной линейной зависимости.

# Регрессия, сводящаяся к линейной

Предполагаем зависимость $y$ от $x$ в виде

$$y = \sum a_k f_k(x) + \varepsilon$$

где $f_k(x)$ — заранее заданные функции, а мы ищем наиболее подходящие значения $a_k$. Задача решается так же, как задача множественной линейной регрессии: решением системы линейных уравнений.

# Модель, объясняющая зависимость наблюдений от условий (общий случай)

$x_1, \ldots, x_n$ — условия
$y_1, \ldots, y_n$ — наблюдения
$y = F(x) + \varepsilon$ — модель (например, если $F(x) = bx + a$, то это  линейная регрессия)
$\hat{y}_i = F(x_i)$ — предсказанные значения.
$r_i = \hat{y}_i - y_i$ — невязки.

В общем случае не обязательно минимизировать именно сумму квадратов невязок, функция $F$ может быть подобрана и из других соображений.

# Коэффициент детерминации

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \qquad \text{total sum of squares}$$

$$SS_{res} = \sum_i (y_i - \hat{y}_i)^2 \qquad \text{residual sum of squares}$$

$$SS_{ex} = \sum_i (\hat{y}_i - \bar{y})^2 \qquad \text{explained sum of squares}$$

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

Здесь $\hat{y}_i = F(x_i)$ — предсказанные моделью значения, $\bar{y}$ — среднее значение $y_i$

Величина $R^2$ называется «коэффициент детерминации»

$R^2 = 1$ означает идеальное соответствие модели наблюдениям (все $\hat{y}_i = y_i$)

Близкое к 0 или тем более отрицательное значение $R^2$ означает, что модель ничего не объяснила (остаточная сумма квадратов практически равна полной сумме квадратов)

# Разложение суммы квадратов

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

Если модель линейная: $\hat{y}_i = bx_i + a$
и **оба коэффициента** получены методом наименьших квадратов (!) ,
то полная сумма квадратов равна сумме объяснённой суммы квадратов и остаточной суммы квадратов.

*Упражнение: докажите это*

Поэтому $R^2$ в этом случае равен доле объяснённой суммы квадратов в полной сумме квадратов (или, что то же самое, доле объяснённой дисперсии в полной дисперсии — дисперсия получается из суммы квадратов делением на $n$):

$$R^2 = SS_{ex}/SS_{tot}$$

Кроме того, $R^2$ в этом случае равен квадрату коэффициента корреляции между $x_1, \ldots, x_n$ и $y_1, \ldots, y_n$ .

Как следствие, в этом случае $0 \leq R^2 \leq 1$

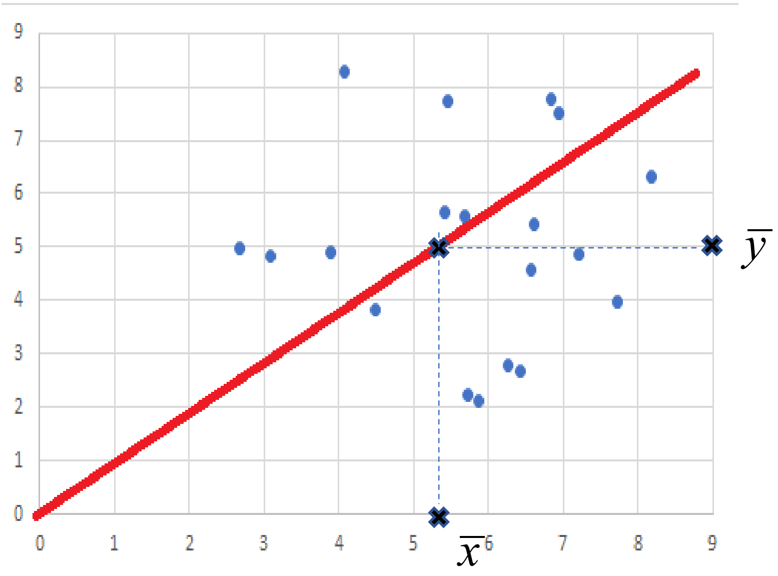
# Пример, когда разложение суммы квадратов неприменимо

Пусть мы предполагаем, что y **пропорционален** x с точностью до ошибок измерения:

$y = bx + \varepsilon\,,\;\; \varepsilon \sim \mathrm{N}(0;\sigma^2)$

и хотим из измерений оценить коэффициент пропорциональности, подбирая $b$ методом наименьших квадратов.

Тогда решение, очевидно: $\hat{b} = \overline{y}\,/\,\overline{x}$



# Пример, когда разложение суммы квадратов неприменимо

Пусть мы предполагаем, что y **пропорционален** x с точностью до ошибок измерения:

$y = bx + \varepsilon\,,\ \ \varepsilon \sim \mathrm{N}(0;\sigma^2)$

и хотим из измерений оценить коэффициент пропорциональности, подбирая $b$ методом наименьших квадратов.

Тогда решение, очевидно: $\hat{b} = \overline{y}\,/\,\overline{x}$

В этом случае коэффициент детерминации **не будет** равен отношению объяснённой и полной сумм квадратов.

Представьте себе, что $y$ реально не зависит от $x$.

Заменим все $y_i$ на $(y_i - \overline{y}\ )\cdot a + \overline{y}$ , где $0 < a < 1$, то есть приблизим $y_i$ к их среднему, не меняя само среднее.

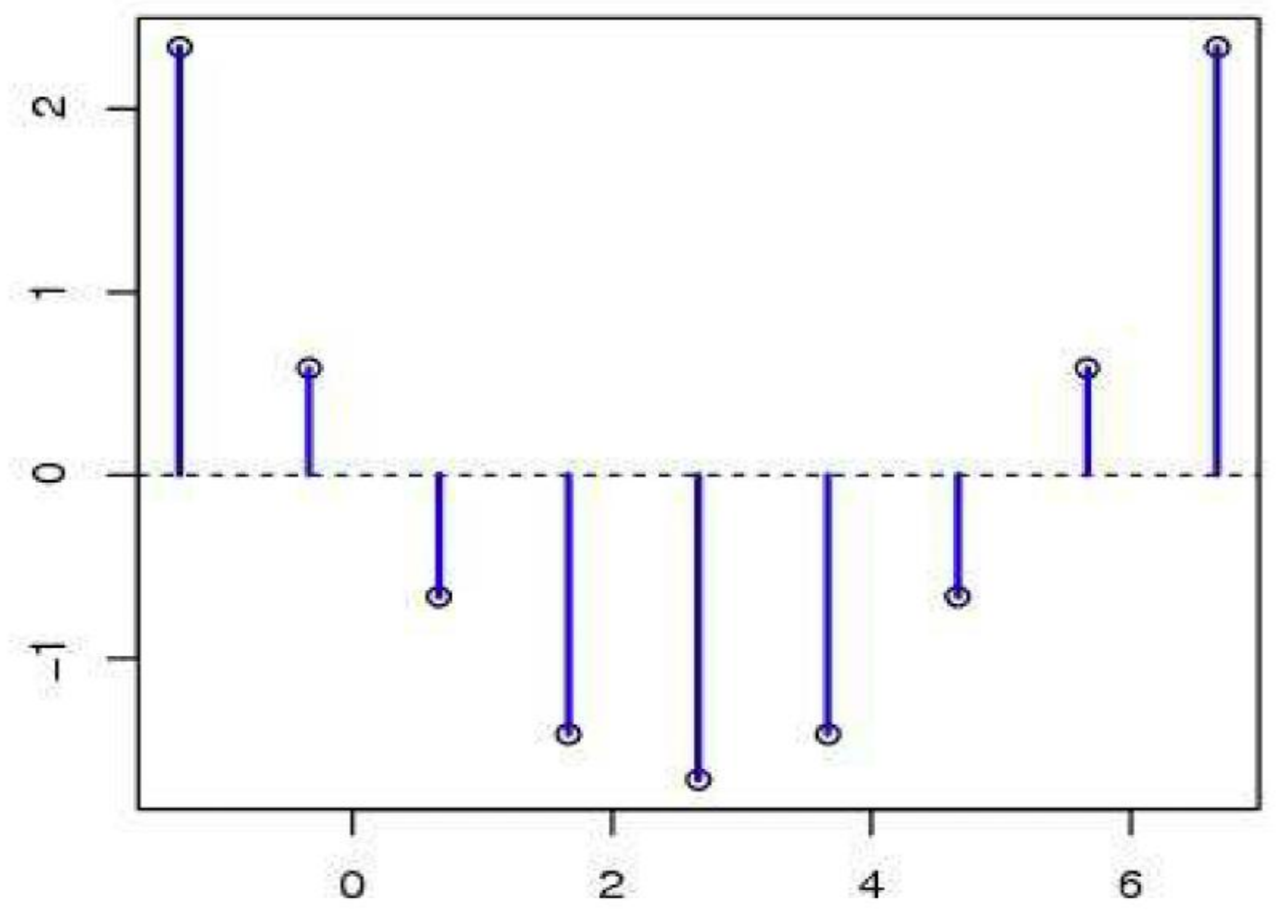
Тогда модель не изменится (потому что среднее то же самое), объяснённая сумма квадратов тоже не изменится.

Полная сумма квадратов уменьшится, поэтому отношение объяснённой и полной сумм квадратов увеличится.

А вот остаточная сумма квадратов будет примерно той же (в пределе $a \to 0$ она будет стремится к объяснённой).

Поэтому отношение остаточной и полной сумм квадратов будет увеличиваться, значит $R^2$ будет неограниченно уменьшаться.

# Пример нелинейной зависимости

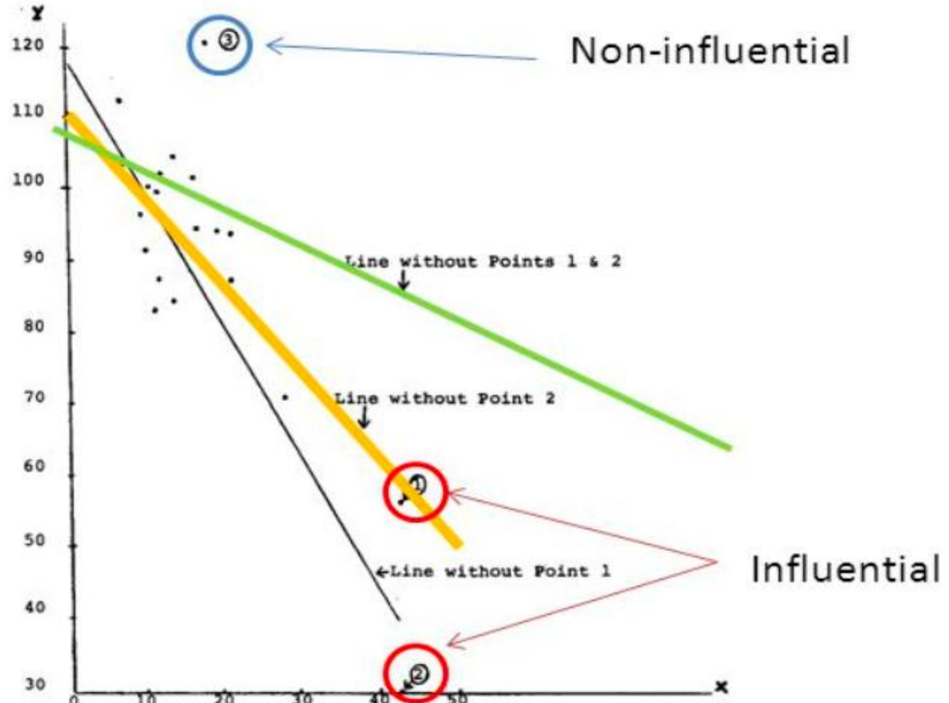


# Выбросы и влиятельные значения

В регрессии выбросом (outlier) называется пара x, y с большой (по модулю) невязкой.
Точнее: обычно выбросами называются те пары, для которых невязки выбиваются из нормального распределения, в то время как для большинства пар они хорошо соответствуют гипотезе нормальности.

Влиятельным (influential) значением называется пара x, y, сильно влияющая на параметры регрессии, то есть такая, после выкидывания которой параметры регрессии существенно меняются.

# Выбросы и влиятельные значения




| Line without | Regression coeffs. | $r_i$ | $h_{ii}$ | $CD_i$ | $D_i^2$ |
|---|---|---|---|---|---|
| Point 2 | 111.51, -1.18 | -.753 | .42 | .508 | 10.604 |
| Point 1 | 118.64, -1.76 | -2.00 | .42 | 3.583 | 10.604 |
| Points 1 & 2 | 107.92, -.89 | | | | |

*Figure 1.* Regression lines and diagnostics for Mickey, Dunn, & Clark (1967) data. (A variation (42.30) on data point 1 has been added.)

Точка 3, хотя и выброс, но не влиятельное значение

# Что здесь выбросы и что — влиятельные значения?