Лекция 5

Корреляции
Регрессионный анализ

КОРРЕЛЯЦИИ (correlation)

До сих пор нас в выборках интересовала только одна зависимая переменная.

Мы изучали, отличается ли распределение этой переменной в одних условиях от распределения той же переменной в других условиях (*скажем, сравнивали разные группы в ANOVA*).

Обратимся к ситуации, когда зависимых переменных будет ДВЕ или более.

Нас интересует вопрос, в какой степени эти переменные связаны между собой.

Это могут быть измерения одной особи или связанных пар.

Корреляции

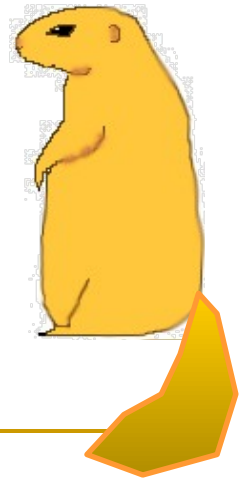
Мы исследуем сусликов. И хотим узнать, связаны ли между собой у них масса и длина хвоста?

Переменные – 1. масса; 2. длина хвоста.

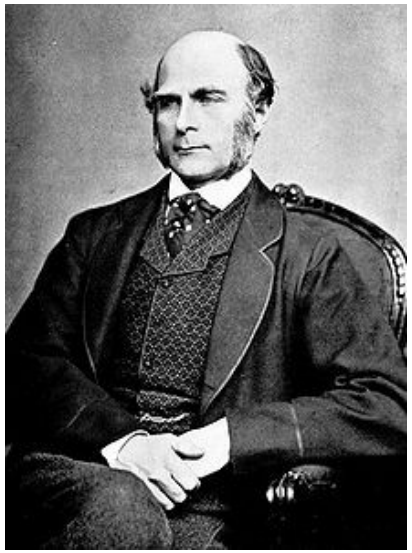
Вопрос: в какой степени две переменные СОВМЕСТНО ИЗМЕНЯЮТСЯ? (т.е., можно ли предполагать, что если у особи одна переменная принимает большое значение, то и значение второй переменной будет большим, или, наоборот, маленьким)

КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ характеризует силу связи между переменными.
ЭТО ПРОСТО ПАРАМЕТР ***ОПИСАТЕЛЬНОЙ*** *СТАТИСТИКИ*

Большой коэффициент корреляции между массой тела и длиной хвоста позволяет нам предсказывать, что у большого суслика, скорее всего, и хвост будет длинным



Коэффициент корреляции Пирсона (Pearson product-moment correlation coefficient *r*)



Francis Galton (1822-1911)

У истоков биометрии стоял Фрэнсис Гальтон (1822-1911) - двоюродный брат Чарлза Дарвина (1809-1882). Ф. Гальтон впервые (1889) ввел в употребление термин «biometrica»; им было введено понятие «регрессии» и разработаны основы корреляционного анализа. Он заложил основы новой науки и дал ей имя, но в стройную научную дисциплину ее превратил математик Карл Пирсон (1857-1936).

Karl Pearson (1857 –1936)

Коэффициент корреляции Пирсона характеризует существование *линейной зависимости между двумя величинами*.

Коэффициент корреляции

1. Может принимать значения от -1 до +1
2. Знак коэффициента показывает *направление связи* (прямая или обратная)
3. Абсолютная величина показывает *силу* связи
4. всегда основан на парах чисел (измерений 2-х переменных от одной особи или 2-х переменных от разных, но связанных особей)

r – в случае, если мы характеризуем ВЫБОРКУ

ρ - если мы характеризуем ПОПУЛЯЦИЮ

Рост братьев: коэффициент корреляции r -?

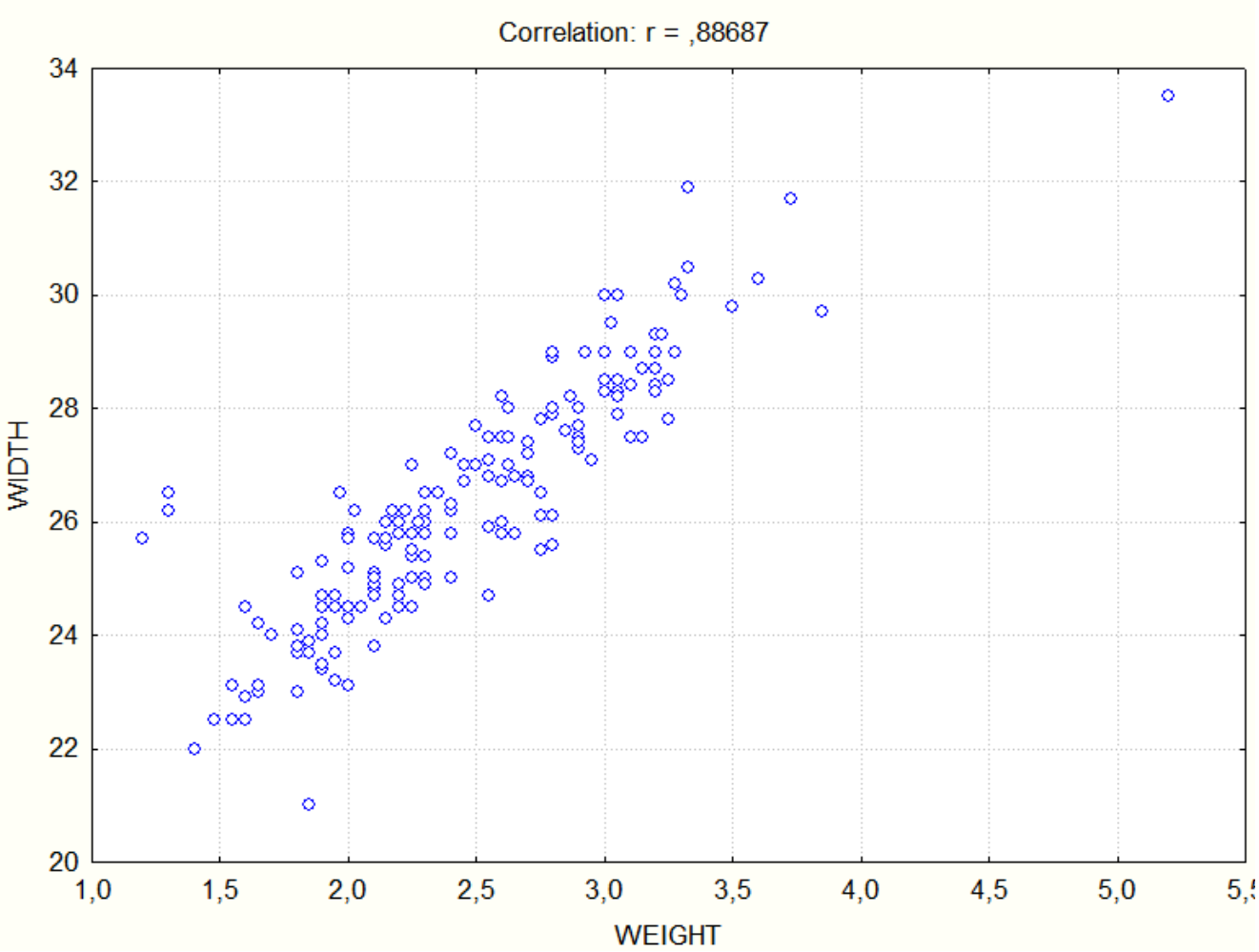
Вася

1. r=1.0: если Вася высокого роста, значит, Юра тоже высокий, это не предположение, а ***факт***.
2. r=0.7: если Вася высокий, то, ***скорее всего***, Юра тоже высокий.
3. r=0.0: если Вася высокий, то мы… не можем сказать о росте Юры ***НИЧЕГО***.

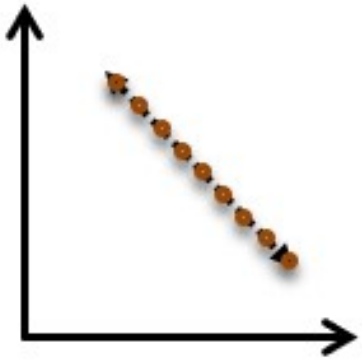
Юра

Скаттерплот

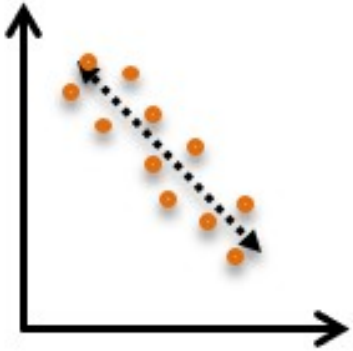
(= диаграмма рассеяния; scatterplot, scatter diagram)



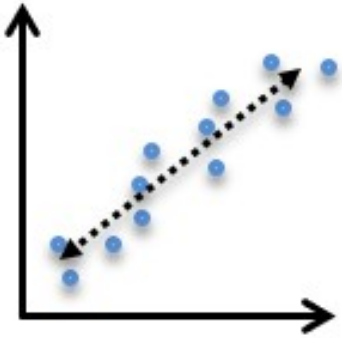
Две характеристики: – наклон (направление связи) и ширина (сила связи) воображаемого эллипса



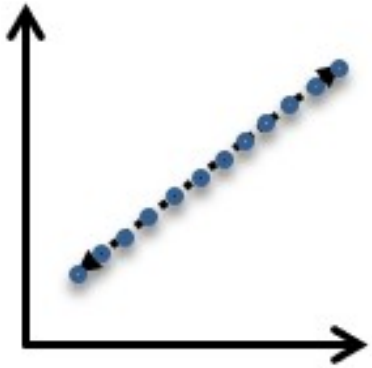
r=-1.0



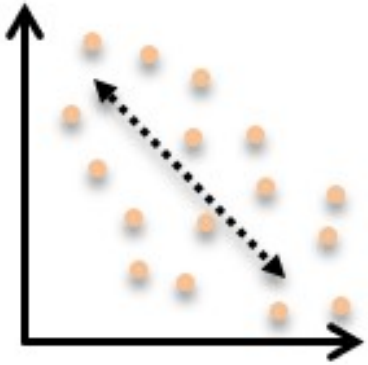
r=-0.8



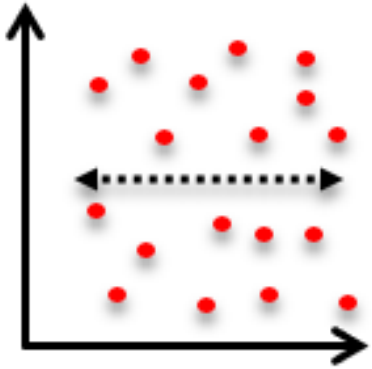
r=0.8



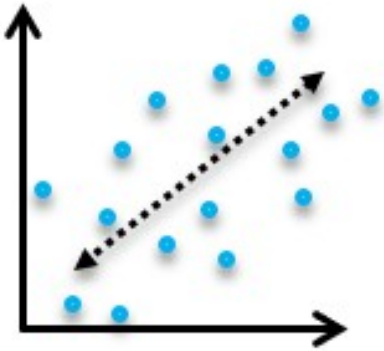
r=1.0



r=-0.3



r=0.0



r=0.3

Коэффициент корреляции Пирсона

суслик	*вес*	*хвост*
Дима	72	160
Гриша	66	144
Миша	68	154
Коля	74	210
Федя	68	182
Рома	64	159
	68,7	*168,2*

$$r = \frac{\sum z_{X_i} z_{Y_i}}{n-1}$$

z – оценки

число строк (сусликов)

$$z_{X_i} = \frac{X_i - \bar{X}}{s_X}$$

стандартное отклонение для веса

$$z_{Y_i} = \frac{Y_i - \bar{Y}}{s_Y}$$

стандартное отклонение для хвоста

для каждого X и Y (для каждого суслика)

Это одна из нескольких эквивалентных формул для коэффициента корреляции Пирсона

Для понимания величины коэффициента корреляции Пирсона необходимо вспомнить, что из себя представляет z-оценка

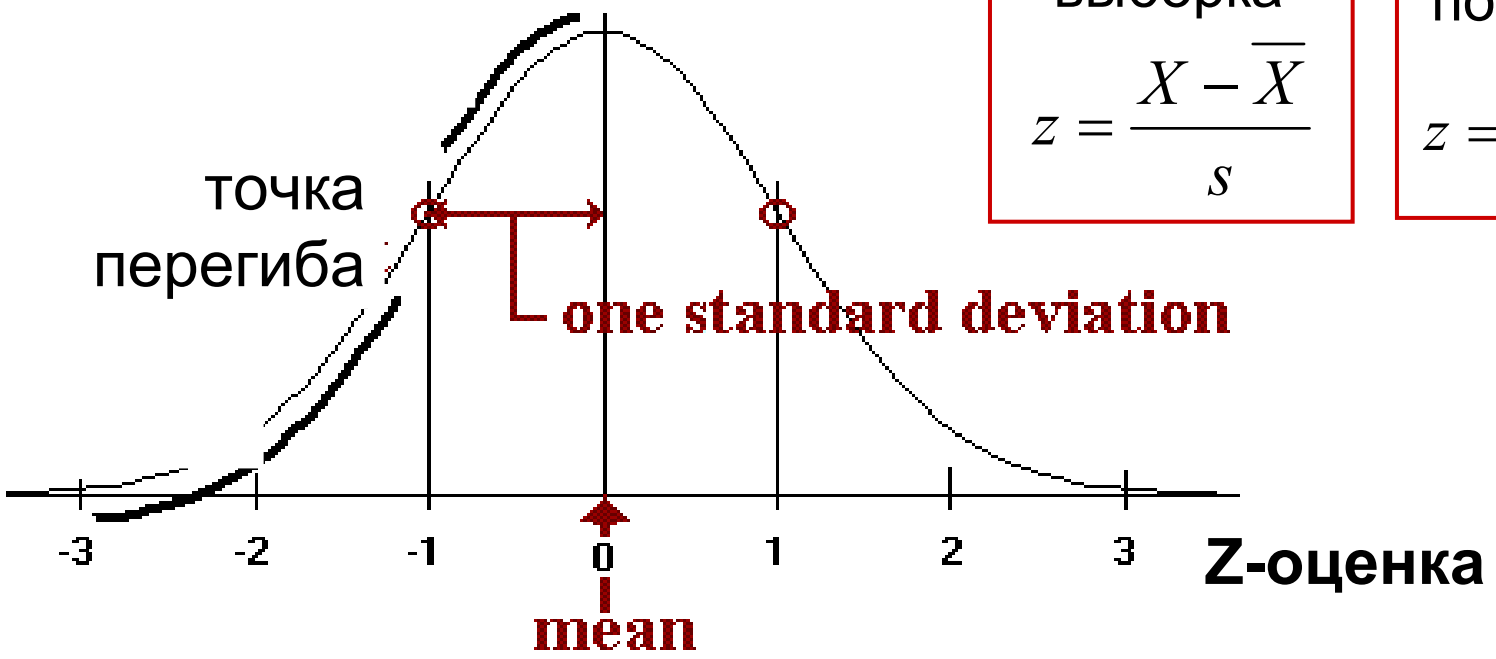
Z-оценка (z-scores) – переменная, соответствующая количеству стандартных отклонений относительно среднего значения

выборка

$$z = \frac{X - \overline{X}}{s}$$

популяция

$$z = \frac{X - \mu}{\sigma}$$



$$r = \frac{\sum z_X z_Y}{n-1}$$

параметр ВЫБОРКИ

$$\rho = \frac{\sum z_X z_Y}{N}$$

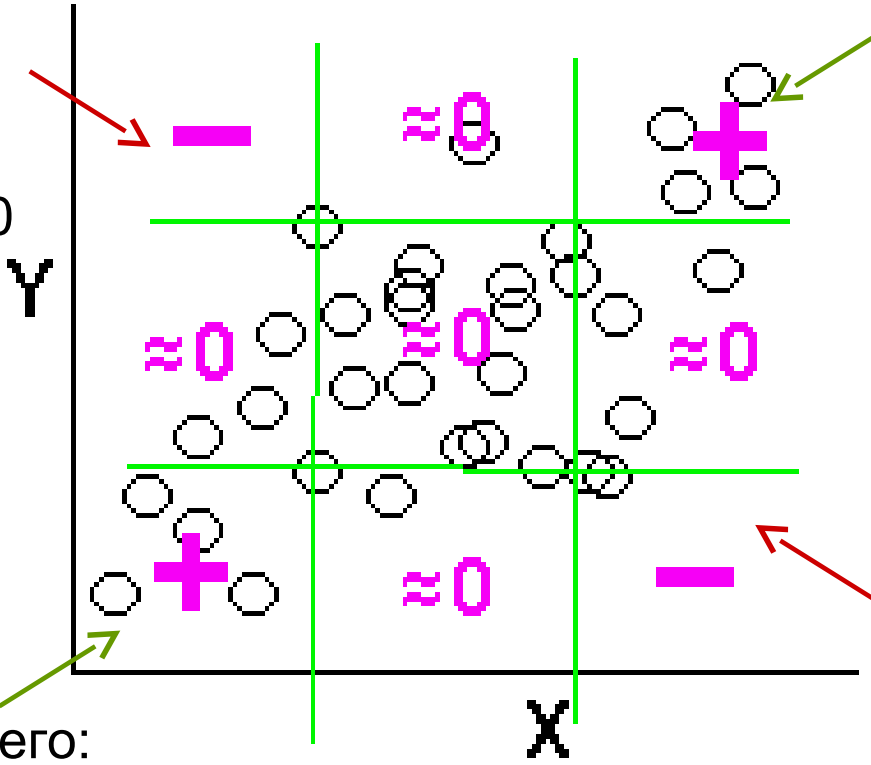
параметр ПОПУЛЯЦИИ

Всё как для других параметров описательной статистики: среднего, дисперсии, и т.д.!

Что определяет $\sum z_X z_Y$?

Чем определяются знак и величина коэффициента корреляции?

Знаком и величиной $\sum z_X z_Y$:



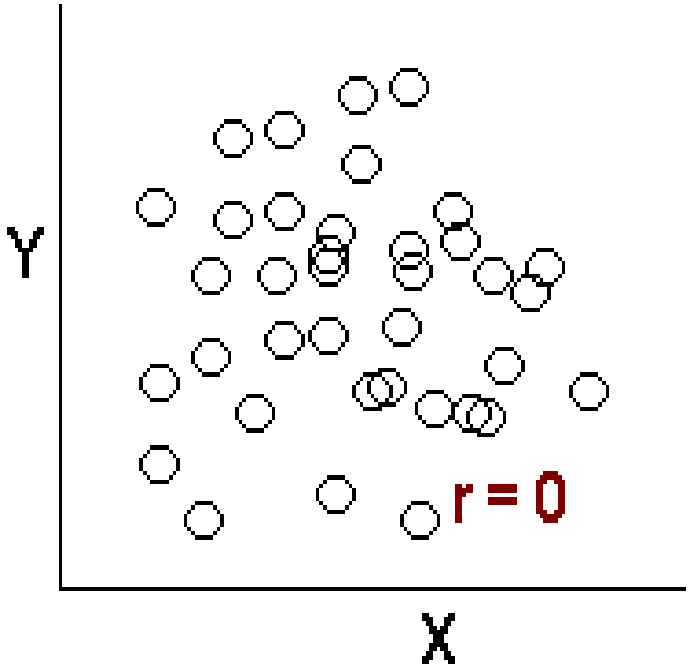
здесь Y больше среднего, а X меньше: их произведение <0

здесь и X, и Y больше среднего: их произведение >0

здесь и X, и Y меньше среднего: их произведение >0

здесь X больше среднего, а Y меньше: их произведение <0

Создаётся впечатление, что близкий к нулю коэффициент корреляции говорит о том, что связи между переменными нет или почти нет.

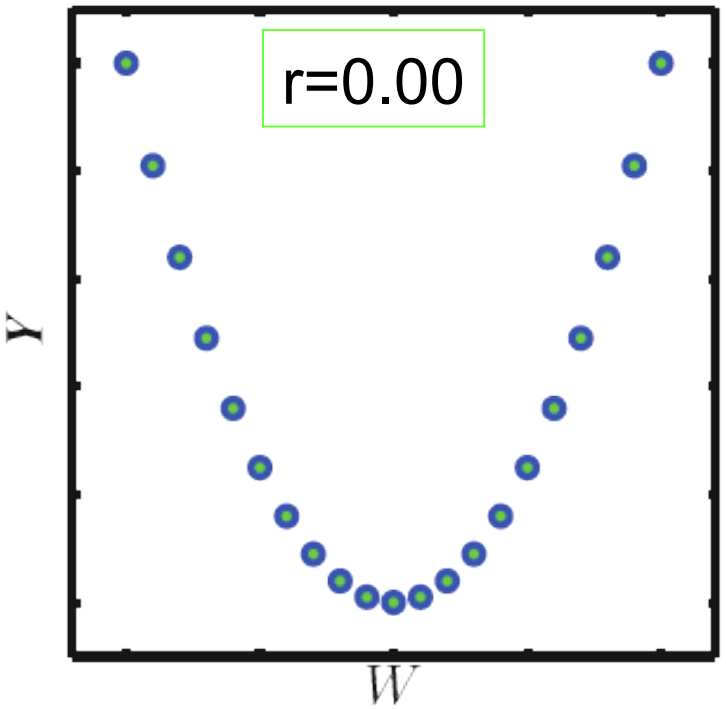


Здесь и впрямь её нет

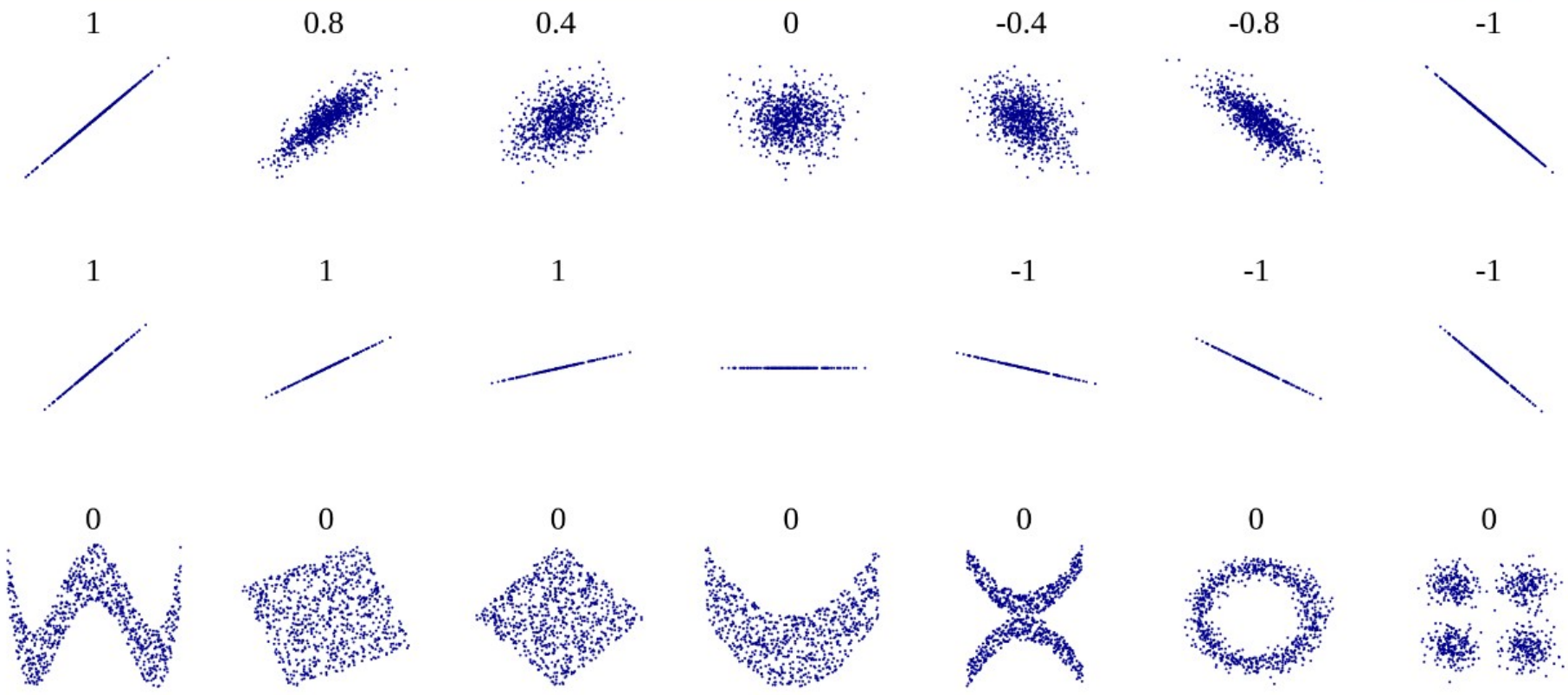
НО это не всегда так, есть исключения.

Факторы, влияющие на коэффициент корреляции

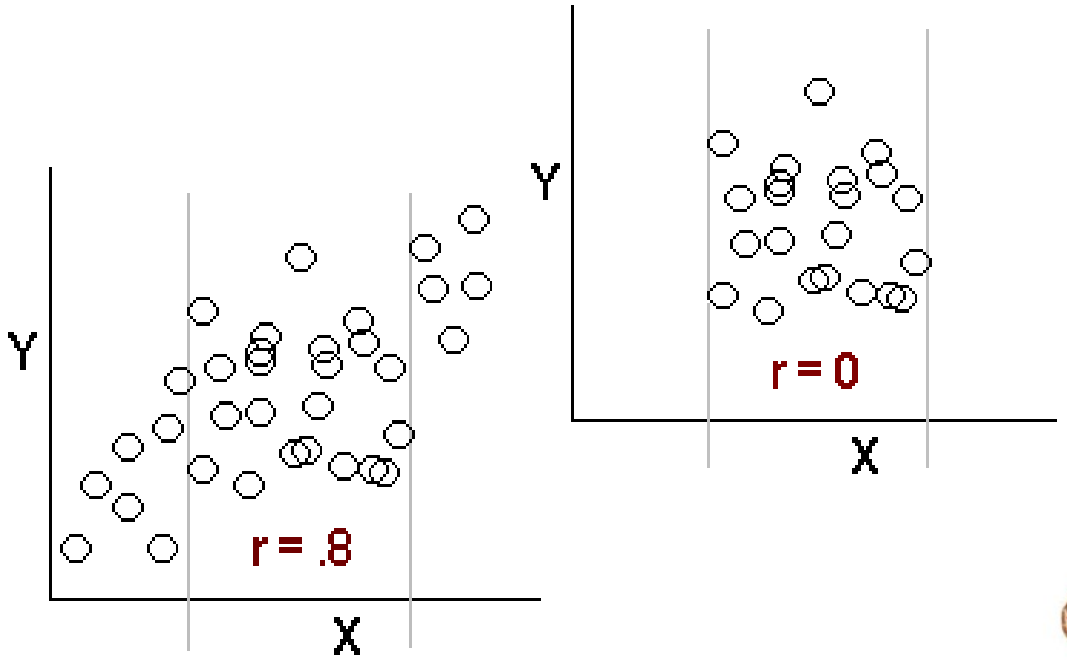
1. Коэффициент корреляции Пирсона оценивает только линейную связь переменных!
И он не покажет нам наличие нелинейной связи



Здесь связь переменных есть,
и она очень сильная, но r=0.00


1
0.8
0.4
0
-0.4
-0.8
-1
1
1
1
-1
-1
-1
0
0
0
0
0
0
0

2. Необходимо, чтобы у переменных была значительная изменчивость! Если сформировать выборку изначально однотипных особей, нечего надеяться выявить там корреляции.



3. Коэффициент корреляции Пирсона очень чувствителен к аутлаерам.

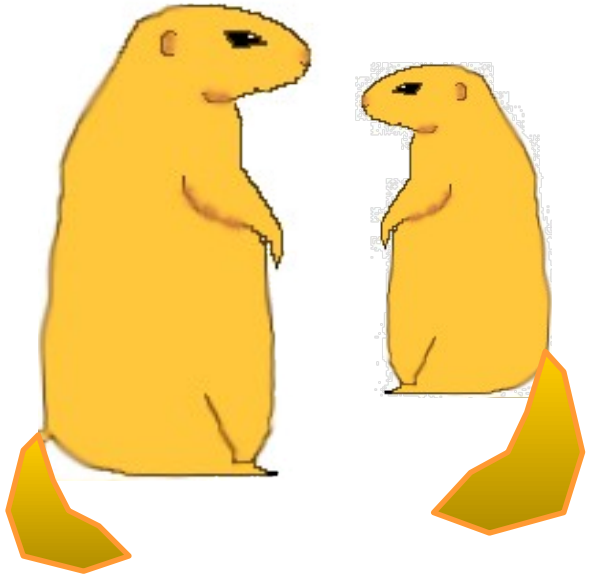
Важное замечание:

Корреляция совершенно **не подразумевает** наличие **причинно-следственной связи**!

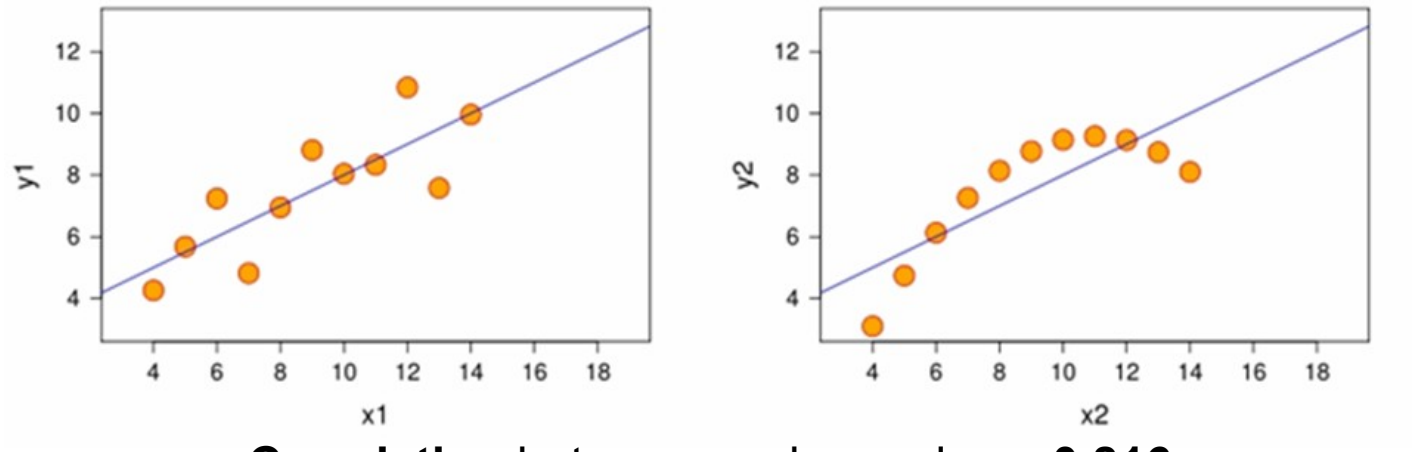
Она **ВООБЩЕ НИЧЕГО** о ней **НЕ ГОВОРИТ** (даже очень большой r)

Для пары связанных показателей *X* и *Y*, возможно:

1) *X* является причиной *Y*;
2) *Y* является причиной *X*;
3) *X* и *Y* являются следствиями одной причины *N*;
4) связь обусловлена более сложными механизмами с вовлечением большого числа показателей.



Коэффициент корреляции Пирсона – параметр выборки.
Можем ли мы на основе него судить о популяции?
Просто глядя на коэффициент – НЕТ.



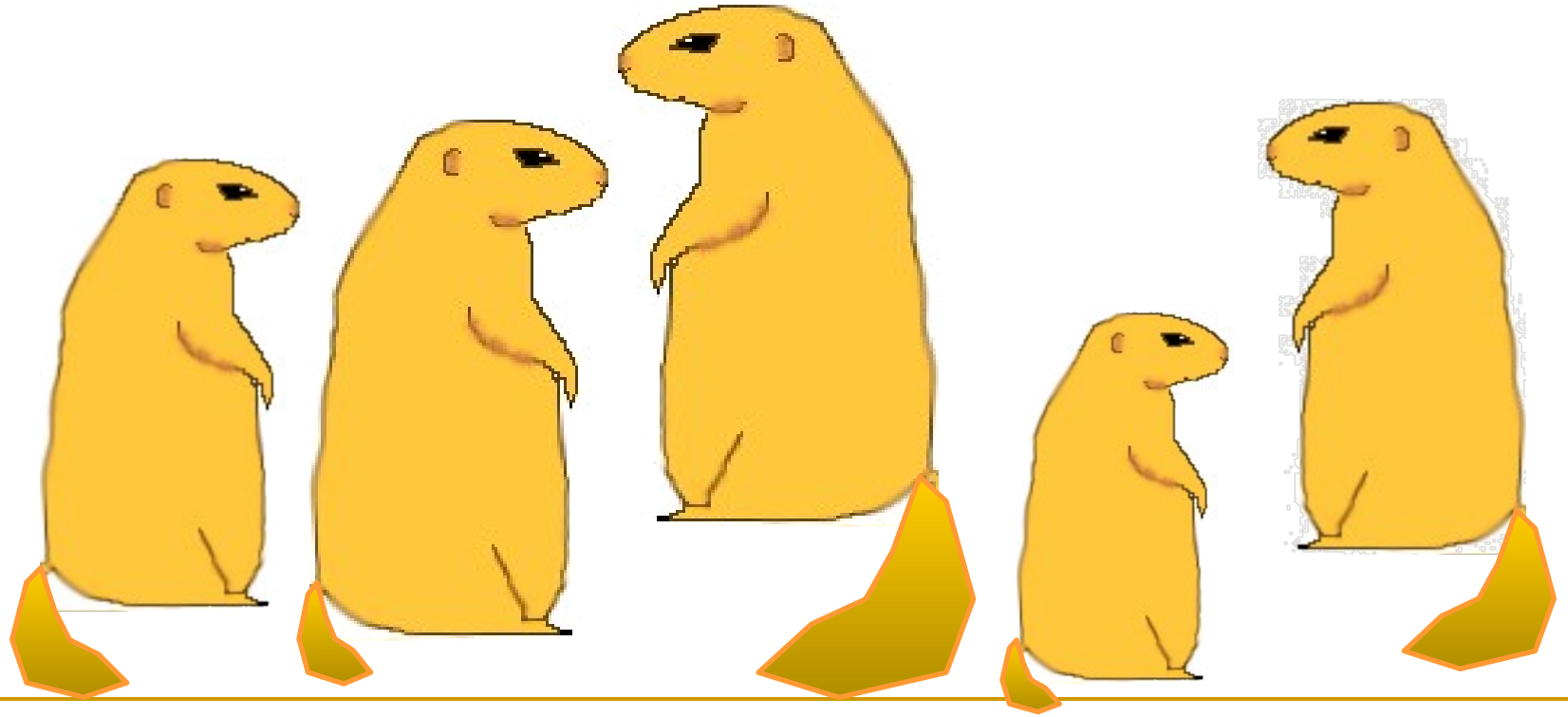
Мы хотим оценить коэффициент корреляции в популяции.

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

(альтернативная гипотеза может быть односторонней)

Связаны ли у сусликов масса тела и длина хвоста?



$$\text{Статистика} = \frac{\text{параметр выборки} - \text{параметр популяции}}{\text{стандартная ошибка параметра выборки}}$$

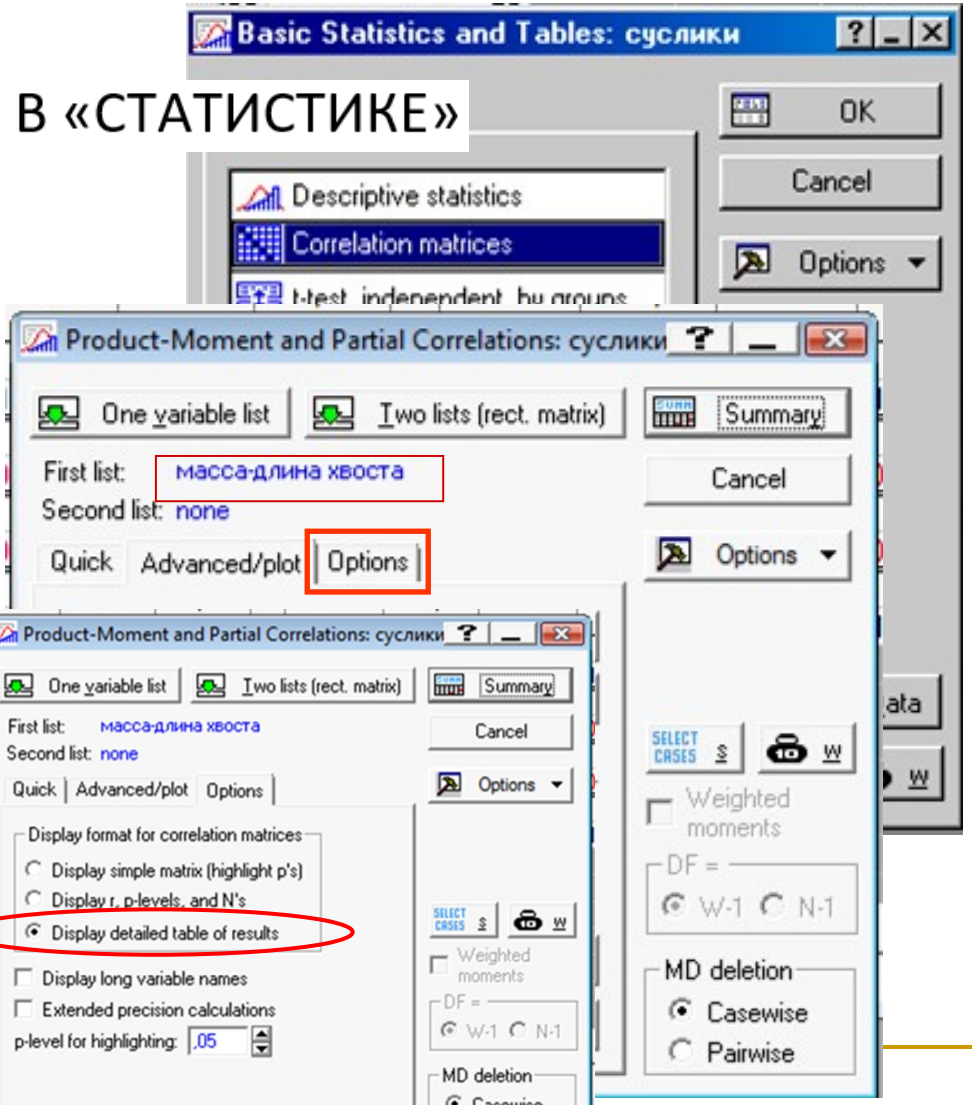
$$t = \frac{r - \rho}{s_r} \longrightarrow \boxed{t = \frac{r}{s_r}}$$

стандартная ошибка коэффициента корреляции

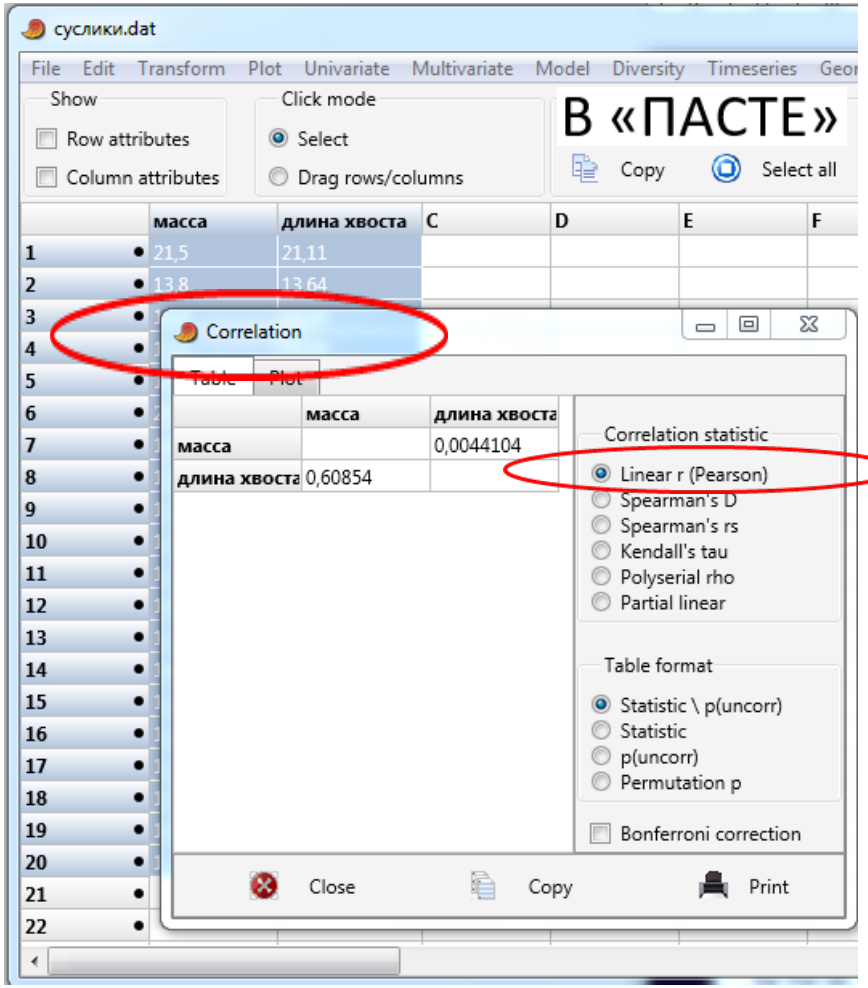
$$s_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

Pearson product-moment correlation coefficient *r*

В «СТАТИСТИКЕ»



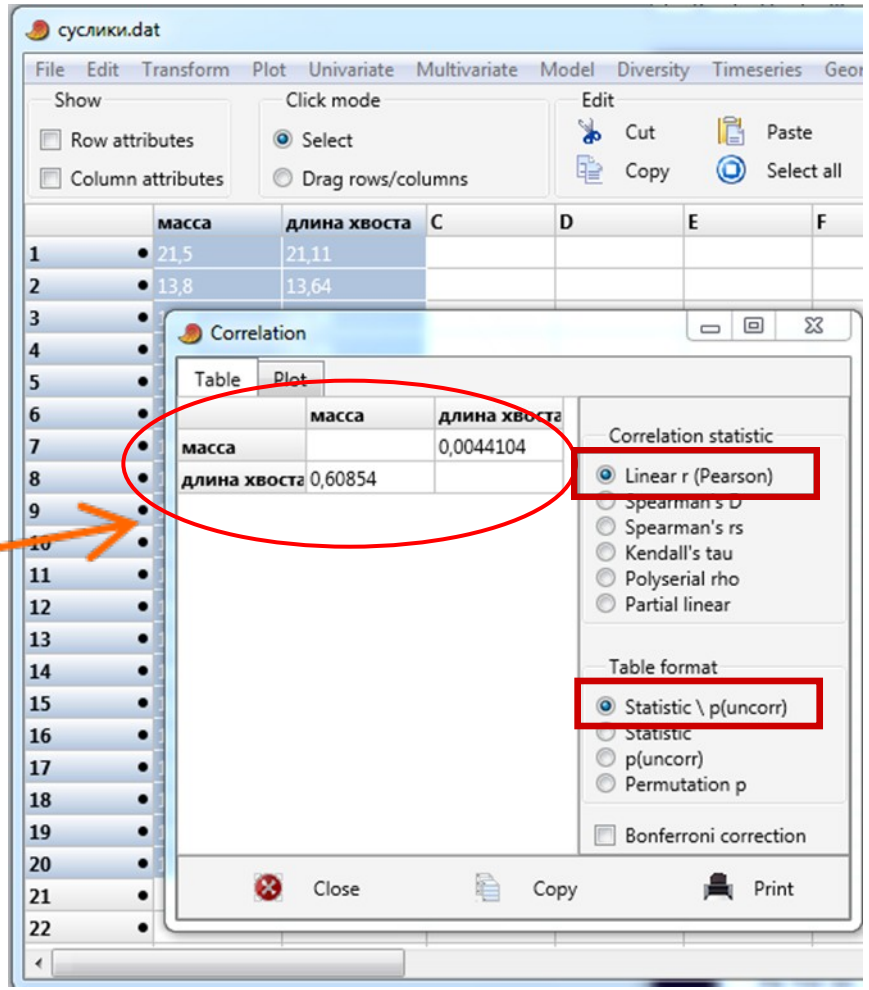
В «ПАСТЕ»



Отвергаем H_0:

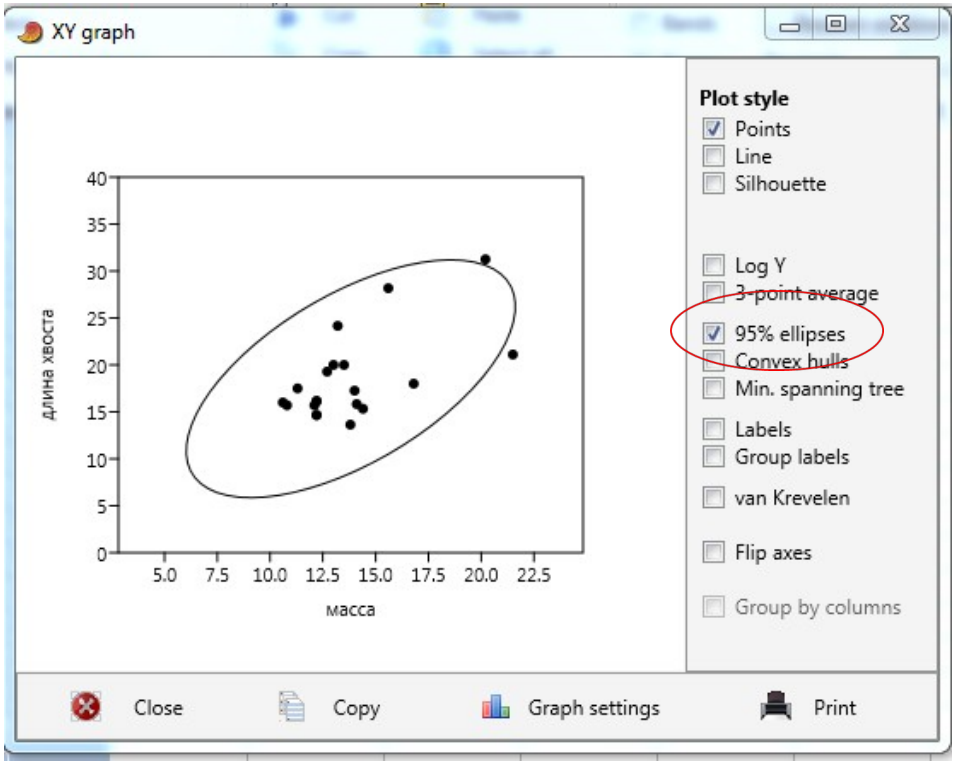
Масса тела у сусликов положительно связана с длиной хвоста

Связь положительная (знак +),
средней силы (0,6 находится в промежутке 0,3-0,7),
статистически значимая (p=0,004 – меньше 0,05)



статистическая значимость — самостоятельная характеристика, зависящая как от силы связи, так и от объёма выборки. На очень маленькой выборке достаточно сильная связь может оказаться незначимой (P> 0,05), а на очень большой выборке можно обнаружить значимой (P < 0,05) даже очень слабую связь.

Построение графика — диаграммы рассеяния (scatter plot).



Путь: Plot — XY graph.
В случае использования линейной корреляции Пирсона мы имеем право обвести облако точек 95%-ным доверительным эллипсом.
Чем у́же эллипс — тем сильнее связь.

Корреляции

В статьях обычно приводят сам коэффициент корреляции Пирсона (значение t не столь обязательно).

Он сам и является показателем практической значимости (effect size) корреляции.

В биологии и медицине обычно считается, что, если абсолютное значение коэффициента находится в интервале

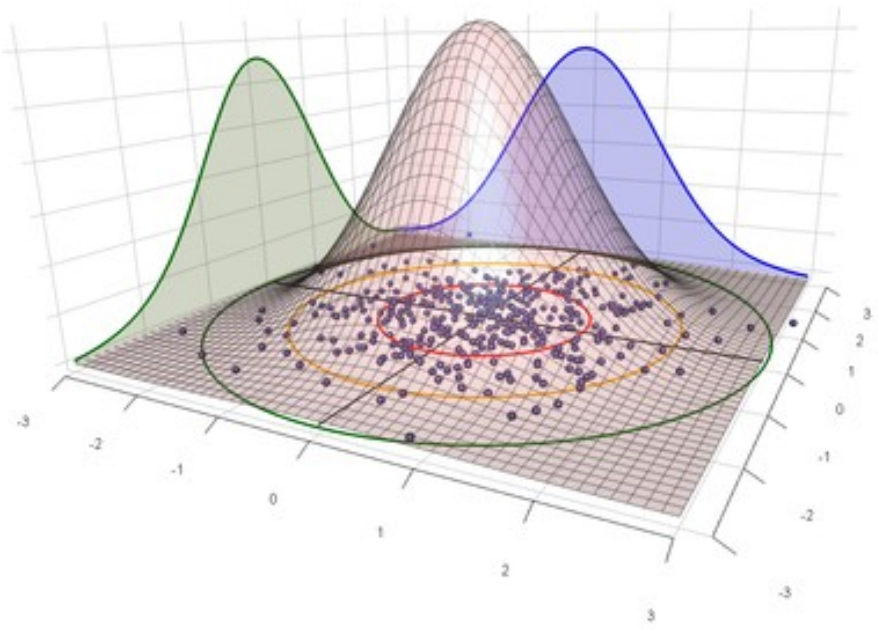
(0 - 0,3] — связь слабая,

(0,3 - 0,7] — связь средней силы,

(0,7 - 1] — связь сильная

Требование к выборке для тестирования гипотезы о коэффициенте корреляции Пирсона:

1. Для каждого X значения Y должны быть распределены нормально, и для каждого Y все X должны иметь нормальное распределение -



двумерное нормальное распределение (bivariate normal distribution)

2. Должно соблюдаться требование гомогенности дисперсии X для каждого Y и наоборот.

РЕГРЕССИОННЫЙ АНАЛИЗ

Вася

Рост братьев.

r=0.7: если Вася высокий, то, ***скорее всего***, Юра тоже высокий. Но можем ли мы предсказать, ***насколько высокий***?

Сам коэффициент корреляции этого нам не скажет.

Ответ нам даст РЕГРЕССИОННЫЙ АНАЛИЗ.

Юра

Регрессии

Регрессионный анализ – инструмент для количественного предсказания значения одной переменной на основании другой.

Для этого в линейной регрессии строится прямая – линия регрессии.

Простая линейная регрессия:

Даёт нам правила, определяющие линию регрессии, которая ЛУЧШЕ ДРУГИХ предсказывает одну переменную на основании другой (переменных всего две).

По оси Y располагают переменную, которую мы хотим предсказать (зависимую, dependent), а по оси X – переменную, на основе которой будем предсказывать (независимую, independent).

РЕГРЕССИЯ (*regression*) – предсказание одной переменной на основании другой. Одна переменная – независимая (independent), а другая – зависимая (dependent).

Пример: чем больше еды съедает каждый день детёныш бегемота, тем больше у него будет прибавка в весе за месяц

КОРРЕЛЯЦИЯ (*correlation*) – показывает, в какой степени две переменные СОВМЕСТНО ИЗМЕНЯЮТСЯ. Нет зависимой и независимой переменных, они эквивалентны.

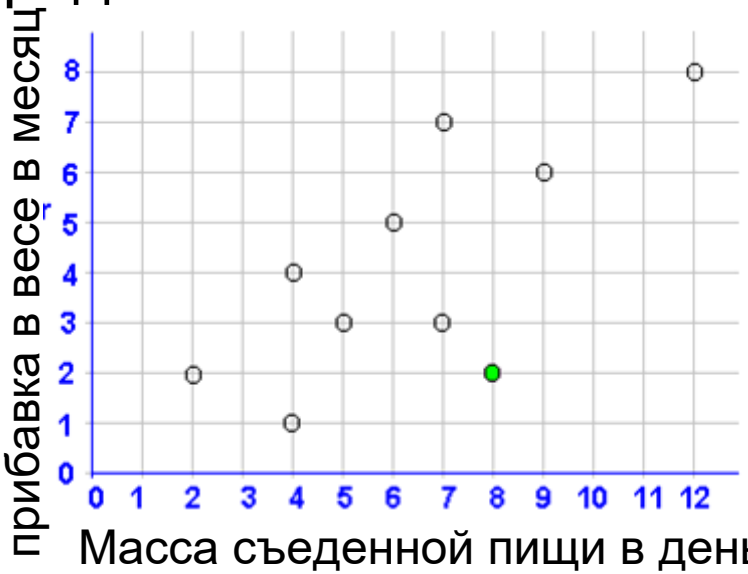
Пример: длина хвоста у суслика коррелирует положительно с его массой тела

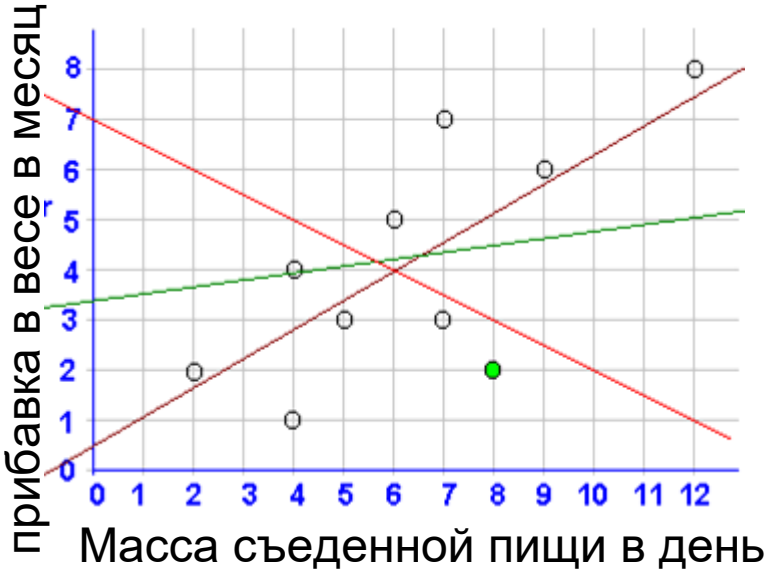
ЭТО НЕ ОДНО И ТО ЖЕ!

Мы изучаем поведение молодых бегемотов в Африке. Мы хотим узнать, как зависит прибавка в весе за месяц от количества пищи, съедаемой в день, у этих зверей?

У нас **две** переменные – 1. кол-во съедаемой в день пищи, кг (independent); 2. прибавка в весе за месяц, кг (dependent)

Мы ищем прямую, которая наилучшим образом будет предсказывать значения Y на основании значений X.





Простая линейная регрессия (*linear regression*)

Y – зависимая переменная
X – независимая переменная
a и b - коэффициенты регрессии

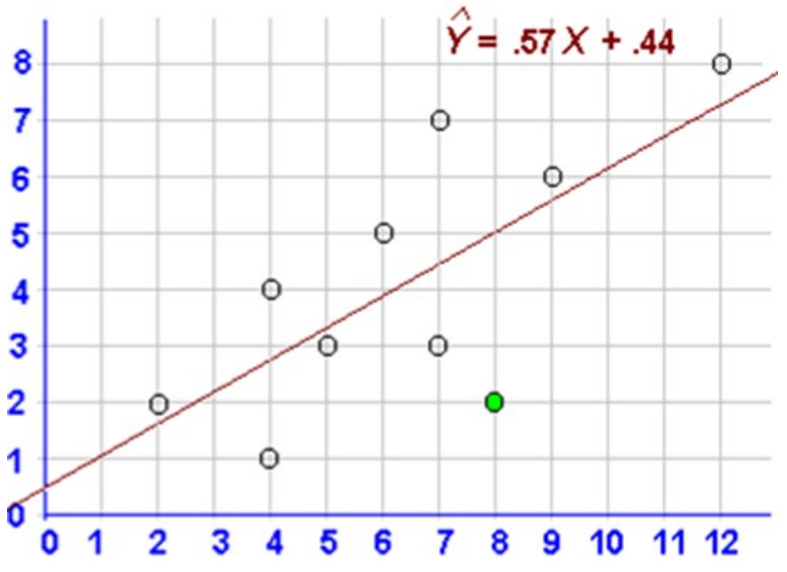
$$\hat{Y}_i = a + bX_i$$

b – характеризует **НАКЛОН** прямой (slope); это самый важный коэффициент;
a – определяет точку пересечения прямой с осью OY; не столь существенный (intercept).

Это уравнение регрессии для ВЫБОРКИ

$$Y_i = \alpha + \beta X_i$$

уравнение для популяции



Задача сводится к поиску коэффициентов a и b.

$$b = r\frac{s_X}{s_Y}$$

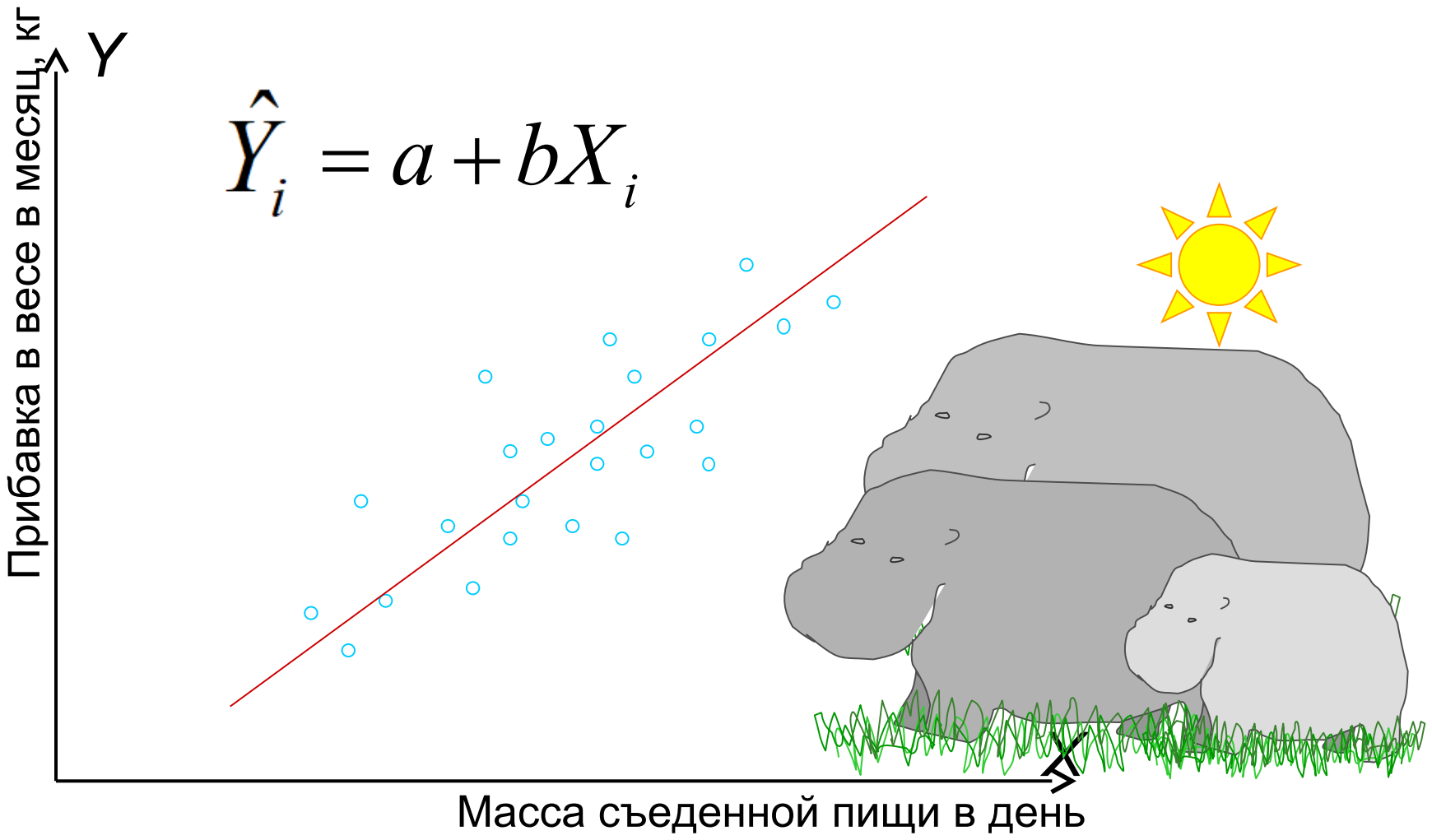
стандартные отклонения для X и Y

коэффициент корреляции Пирсона

$$\bar{Y} = a + b\bar{X} \longrightarrow \boxed{a = \bar{Y} - b\bar{X}}$$

Линия регрессии всегда проходит через точку $(\bar{X}, \bar{Y})$, то есть через середину графика.

b – определяет, насколько изменится Y на единицу X; имеет тот же знак, что и r.

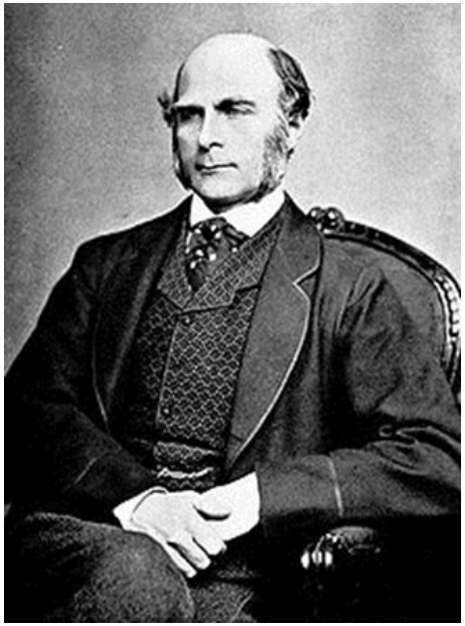

Y
Прибавка в весе в месяц, кг
$\hat{Y}_i = a + bX_i$
X
Масса съеденной пищи в день

Если r=0.0, линия регрессии всегда горизонтальна. Чем ближе r к нулю, тем труднее на глаз провести линию регрессии. А **чем больше r**, тем **лучше предсказание**.

Важная особенность нашего предсказания: предсказанное значение Y всегда ближе к среднему значению, чем то значение X, на основе которого оно было предсказано – **регрессия к среднему**.

Закон регрессии к среднему Фрэнсиса Гальтона

Фрэнсис Гальтон
в 1886-1889 годах провёл серию измерений, в том числе изучил **205** пар родителей и **930** их взрослых детей и опубликовал ряд статей, в которых им был сформулирован **«закон регрессии к среднему»** или, как иногда его переводят: **«закон регрессии к посредственности»**.



Francis Galton
(1822-1911)

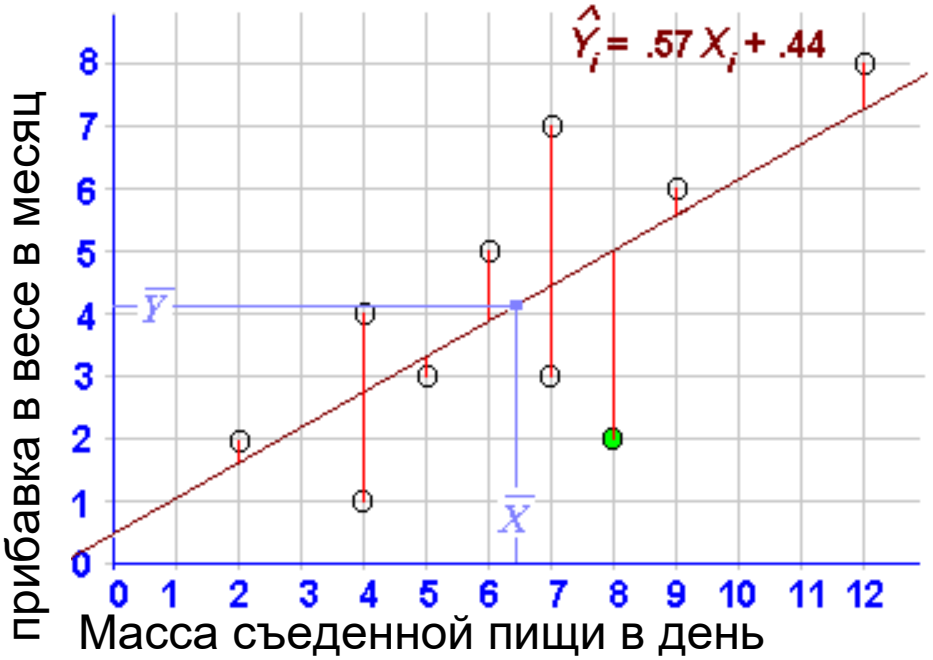
Закон регрессии к среднему Фрэнсиса Гальтона

Было установлено, что для многих непрерывных признаков, таких как рост и интеллект, взрослое потомство данного родителя отклоняется в меньшей степени от среднего значения для данной популяции, чем родитель, то есть, потомки «регрессируют» к среднему для популяции.

Фрэнсис Гальтон этому наблюдению дал название «закон дочерней регрессии к посредственности». Он считал его фундаментальным законом наследственности.

Ошибка предсказания и поиск «лучшей» линии

Очевидно, что точки не лежат на самой линии регрессии.



$$e_i = Y_i - \hat{Y}_i$$

Ошибка предсказания (residual) = «остатки»

e положительно для точек **над** прямой и отрицательно для точек **под** прямой.

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$ Для популяции

$$Y_i = a + bX_i + e_i$$ Для выборки

важно: нельзя пытаться предсказывать Y на основе значений X, лежащих за пределами размаха X в выборке.

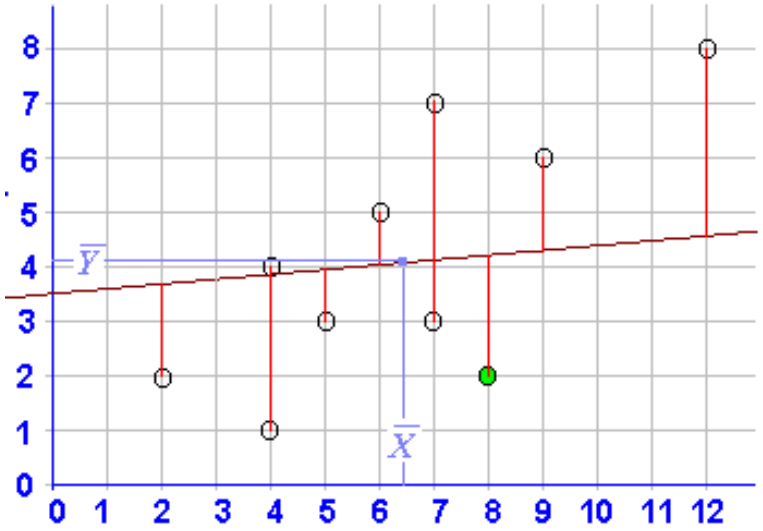
Как определить «лучшую» линию регрессии?

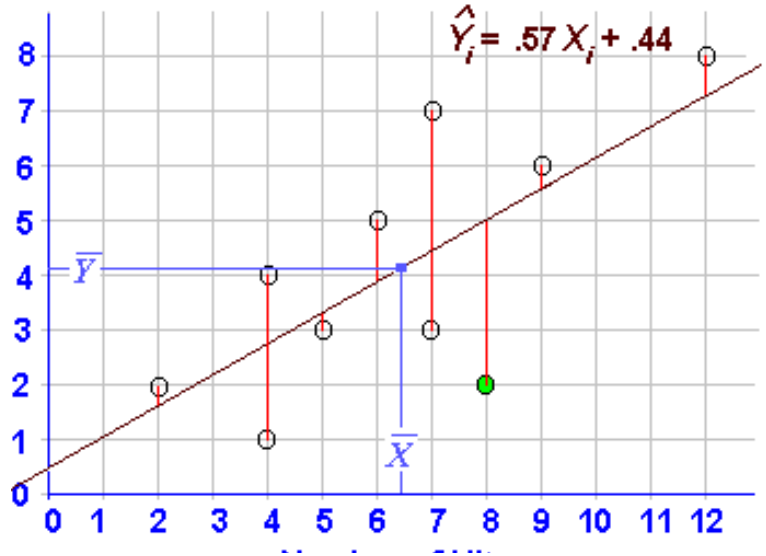
Метод наименьших квадратов:

линию регрессии подбирают такую, чтобы общая сумма квадратов ошибок (residuals) была наименьшей

$\sum e_i = 0$

$\sum e_i^2$ - минимальна





$\sum e_i^2$ - residual sum of squares = residual SS

Насколько хорошо «лучшая» линия регрессии предсказывает Y?

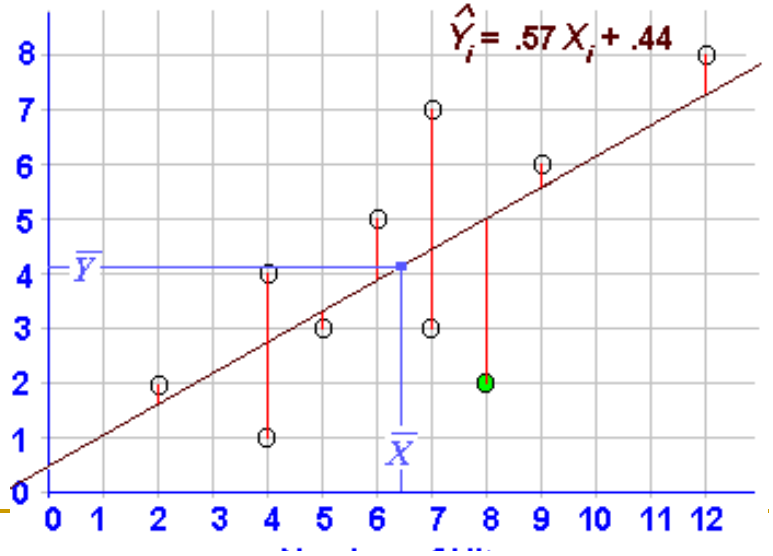
Чем меньше **стандартное отклонение ошибок e_i** (standard error of estimate), тем точнее предсказание (потому, что оно напрямую зависит от размера самих ошибок).

$$s_e = \sqrt{\frac{\sum (e_i - \bar{e})^2}{n-2}} = \sqrt{\frac{\sum e_i^2}{n-2}}$$

$$s_e = s_Y \sqrt{1 - r^2} \sqrt{\frac{n-1}{n-2}}$$

≈ 1

зависит от квадрата коэффициента корреляции



В регрессионном анализе, как и в ANOVA, используют разные **суммы квадратов отклонений** (SS) для разных источников изменчивости, и на их основе тестируют гипотезы.

$$SS_{total} = \sum_i (Y_i - \bar{Y})^2$$

$$SS_{regression} = \sum_i (\hat{Y}_i - \bar{Y})^2$$

$$SS_{residual} = \sum_i (Y_i - \hat{Y}_i)^2$$

$$SS_{total} = SS_{regression} + SS_{residual}$$

Для каждого SS считают соответствующий MS = SS/DF
(df=1 и df=n-2)

$H_0: \beta = 0$
$H_1: \beta \neq 0$

$$F = \frac{MS_{regression}}{MS_{residual}}$$

Эту же гипотезу можно протестировать с помощью t-статистики:

$$t = \frac{b - \beta_0}{s_b} = \frac{b}{s_b}$$

Причём $t^2 = F$

На самом деле,
если r достоверно отличается от нуля, то и $\beta \neq 0$!
То есть, если мы отвергаем H_0 о том, что r=0, то нулевая гипотеза о коэффициенте β отвергается автоматически.

Коэффициент детерминации

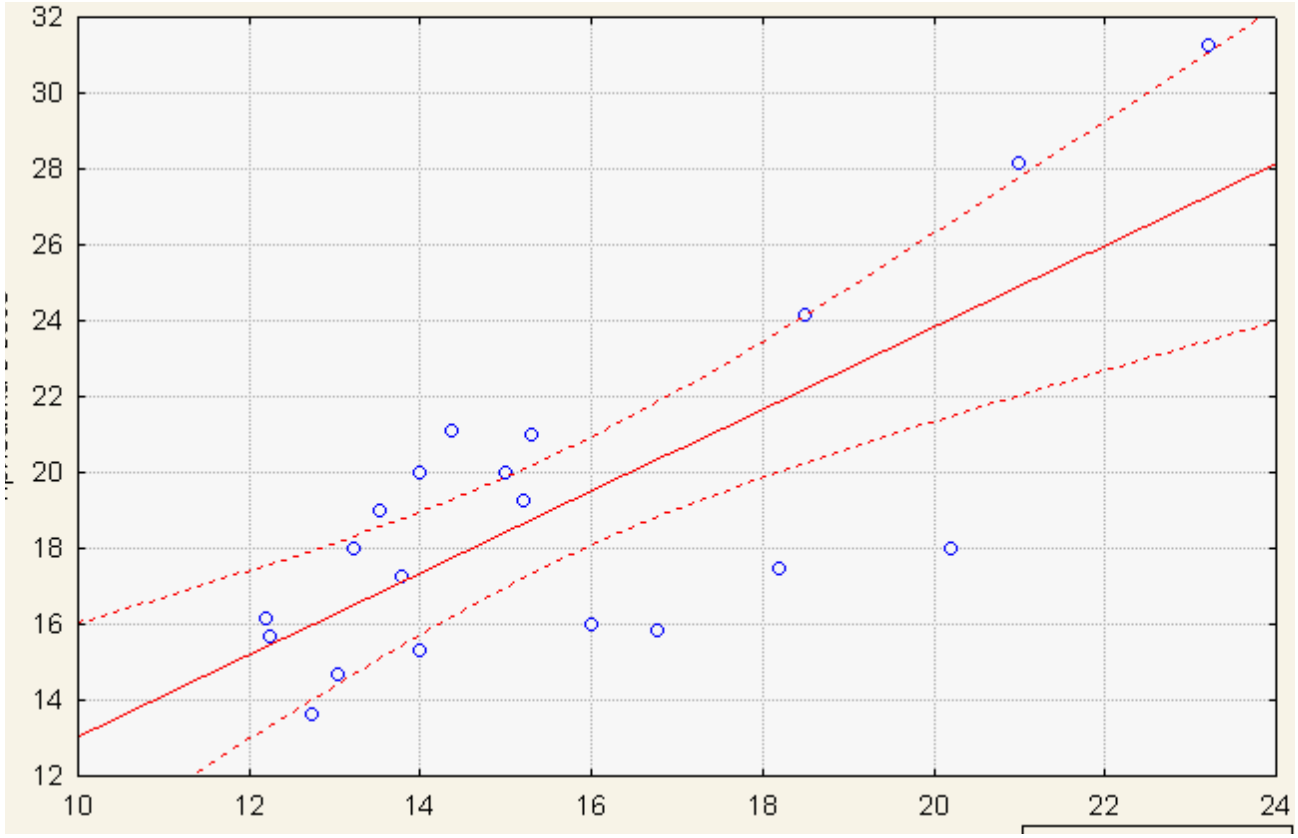
$$R^2 = \frac{SS_{regression}}{SS_{total}} = \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2}$$

Показывает, какую долю изменчивости (буквально, её даже можно выразить в процентах) зависимой переменной (Y) объясняет независимая переменная (регрессионная модель)

r – коэффициент корреляции, $r^2 = R^2$

Насколько велик или мал коэффициент корреляции 0.3?
$0.3^2 = 0.09$, независимая переменная объясняет только около 1/10 изменчивости зависимой переменной.

Доверительный интервал для значений зависимой переменной: строится для каждого значения X, причём наименьшая ошибка получается для среднего Y.



С ростом числа наблюдений доверительный интервал сужается к линии регрессии

Сравнение двух (и более) уравнений линейной регрессии

1. Сравнение коэффициентов наклона b_1 b_2
2. Сравнение коэффициентов сдвига a_1 и a_2

На основе критерия Стьюдента

3. Сравнение двух линий регрессии в целом (предполагается, что если линии для 2-х выборок у нас сильно различаются, и мы объединим выборки, то общая линия по этим двум выборкам будет хуже описывать изменчивость, остаточная дисперсия будет больше) – на основе F-критерия

Множественная линейная регрессия и корреляция (multiple regression)

Простая линейная регрессия: одна зависимая переменная и одна независимая.

Множественная регрессия: исследуется влияние НЕСКОЛЬКИХ независимых переменных на ОДНУ зависимую.

Множественная корреляция: исследуется связь нескольких переменных, среди которых невозможно выделить зависимую.

Например, мы хотим узнать, как на прибавку в весе у бегемотов (1 зависимая переменная) влияют: средняя масса пищи, съедаемой в день; продолжительность сна в сутки; подвижность бегемота (км/день) (3 независимых непрерывных переменных).

Уравнение регрессии:

для популяции

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_m X_{mi} + \varepsilon_i$$

для выборки

$$\hat{Y}_i = a + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_m X_{mi}$$

Это уже не прямая, это уже либо плоскость (для 3-х переменных), либо пространство.

Тестирование гипотез для множественной регрессии:

Если для простой регрессии можно было проверить только гипотезу относительно коэффициента корреляции, в множественной регрессии без SS, MS и F не обойтись – этот анализ тоже называется ANOVA

$$H_o : \beta_1 = \beta_2 = \ldots = \beta_m = 0$$

$$F = \frac{MS_{regression}}{MS_{residual}}$$

Source of variation	Sum of squares (SS)	DF*	Mean square (MS)
Total	$\sum(Y_j - \overline{Y})^2$	$n - 1$	
Regression	$\sum(\hat{Y}_j - \overline{Y}_j)^2$	m	$\frac{\text{regression SS}}{\text{regression DF}}$
Residual	$\sum(Y_j - \hat{Y}_j)^2$	$n - m - 1$	$\frac{\text{residual SS}}{\text{residual DF}}$

*n = total number of data points (i.e., total number of Y values); m = number of independent variables in the regression model.

Коэффициент детерминации (coefficient of determination)

$$R^2 = \frac{SS_{regression}}{SS_{total}}$$

Считается по тому же принципу, что и для простой регрессии, и тоже показывает, какую долю общей изменчивости зависимой переменной объясняет модель, т.е., совместное влияние всех независимых переменых.

Multiple correlation coefficient:
аналогичен коэффициенту корреляции Пирсона

$$R = \sqrt{R^2}$$

Adjusted coefficient of determination:
(Отрегулированный КД) лучше, чем просто R^2, так как не увеличивается с ростом кол-ва переменных в модели

$$R_a^2 = 1 - \frac{MS_{residual}}{MS_{total}}$$

Multicollinearity = ill-conditioning

Множественная взаимозависимость переменных = многократная вырожденность

У нас много переменных, поэтому расчёт коэффициентов и статистик сопряжён с операциями над матрицами.
Если какие-то независимые переменные сильно коррелируют между собой, возникает принципиальная проблема в расчётах (матрицы оказываются вырожденными) – коэффициенты регрессии не могут быть рассчитаны.

Признаки:

✓При удалении (добавлении) какой-либо переменной принципиально меняются коэффициенты при других переменных;

✓при пошаговом анализе выбирая разные способы анализа мы получаем разные результаты.

Что делать? Искать коррелирующие переменные и исключать одну из них из модели.

Выбор «лучших» независимых переменных

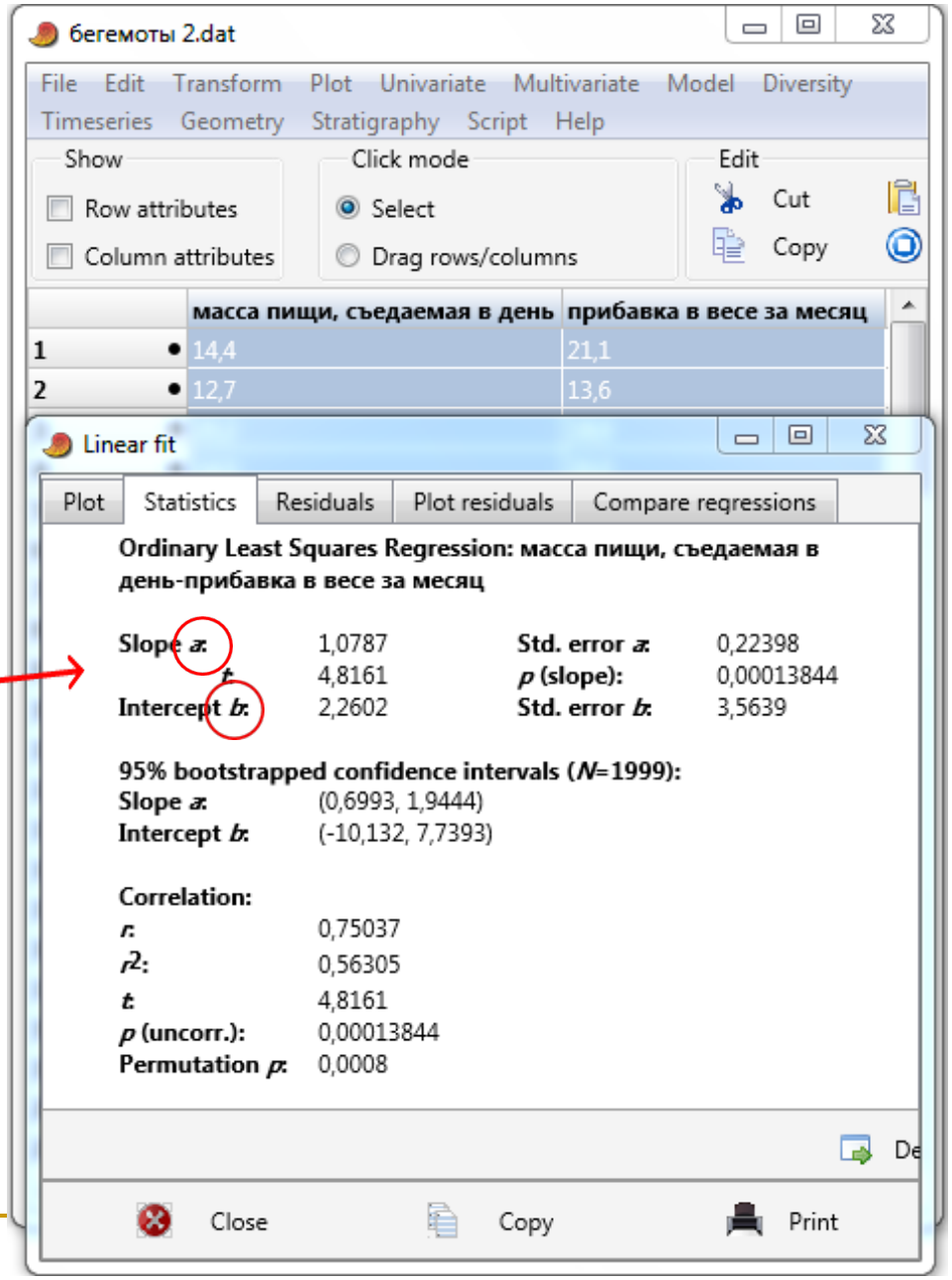
Как выбрать лучшую модель, чтобы наименьшим числом независимых переменных описать наибольшую долю изменчивости Y?

Используют пошаговые модели:

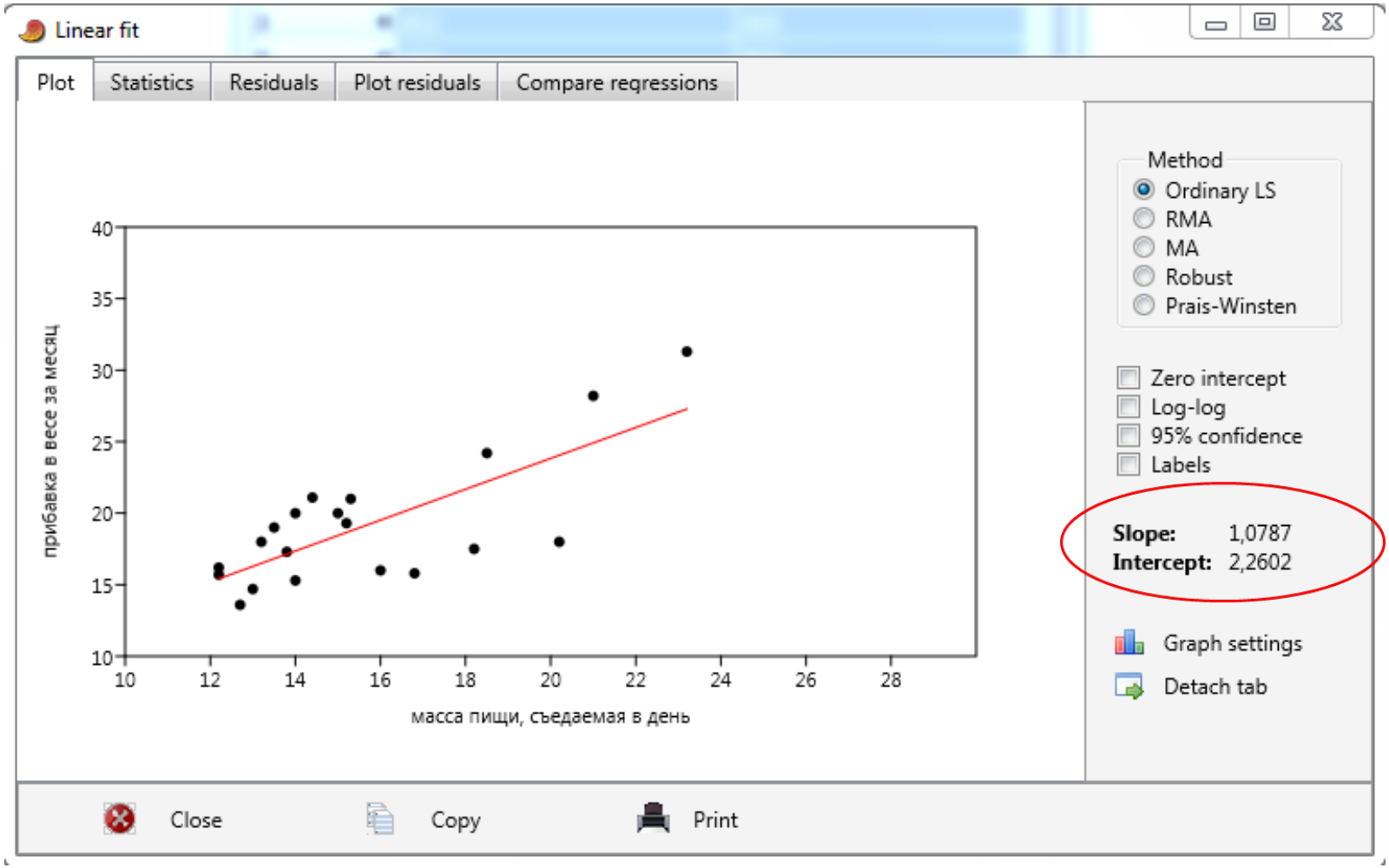
- ✓Backward elimination – **постепенное удаление переменных из модели**.
- ✓Forward selection – **постепенное добавление переменных в модель**
- ✓**Смешанный пошаговый метод анализа**.

Simple linear regression *(простая линейная регрессия)*

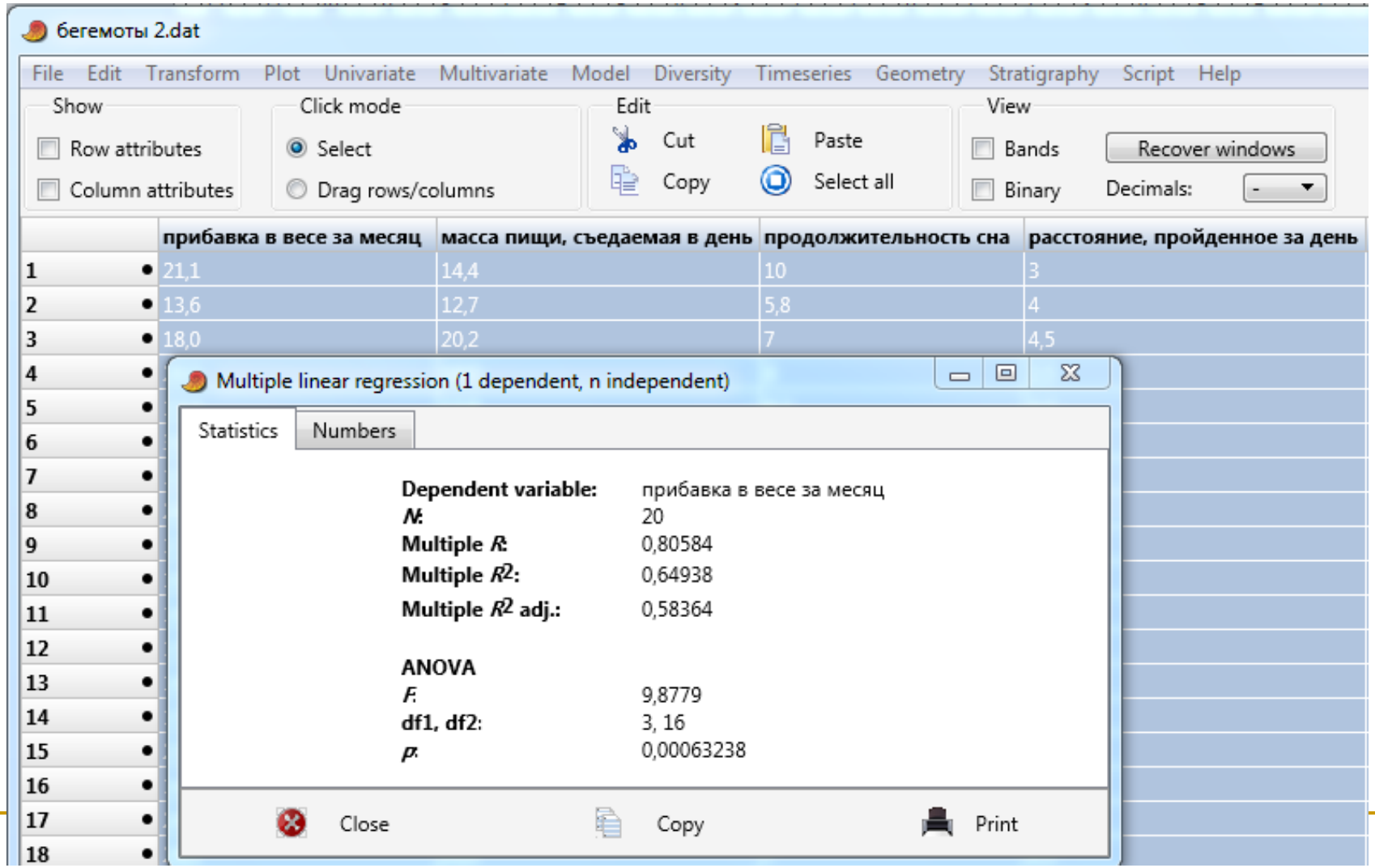
Коэффициенты
a и b



У бегемотов прибавка в весе положительно зависела от массы пищи, съедаемой в день


Linear fit
Plot
Statistics
Residuals
Plot residuals
Compare regressions
Method
Ordinary LS
RMA
MA
Robust
Prais-Winsten
Zero intercept
Log-log
95% confidence
Labels
Slope: 1,0787
Intercept: 2,2602
Graph settings
Detach tab
прибавка в весе за месяц
масса пищи, съедаемая в день
40
35
30
25
20
15
10
10
12
14
16
18
20
22
24
26
28
Close
Copy
Print

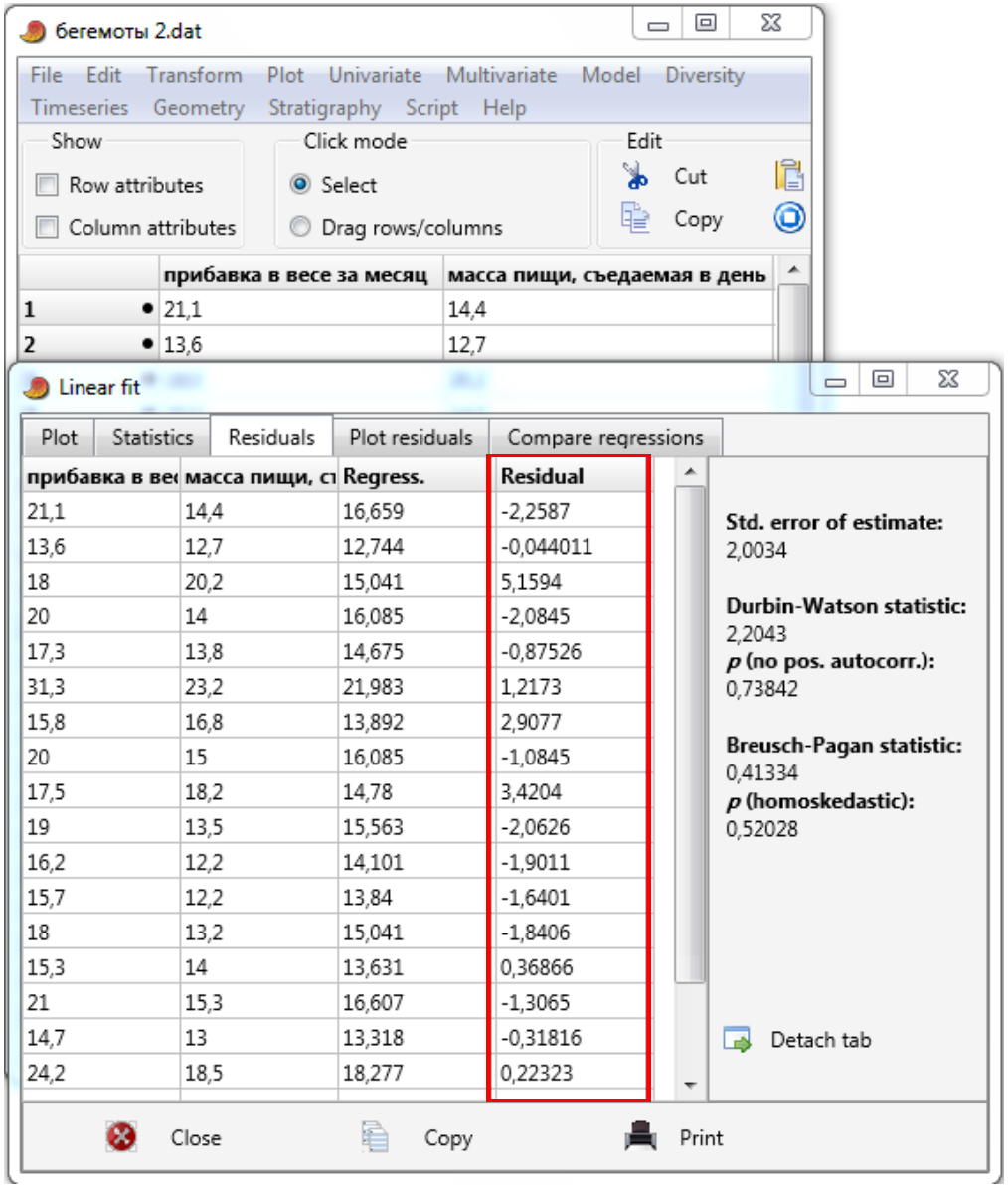
Multiple linear regression (множественная линейная регрессия)

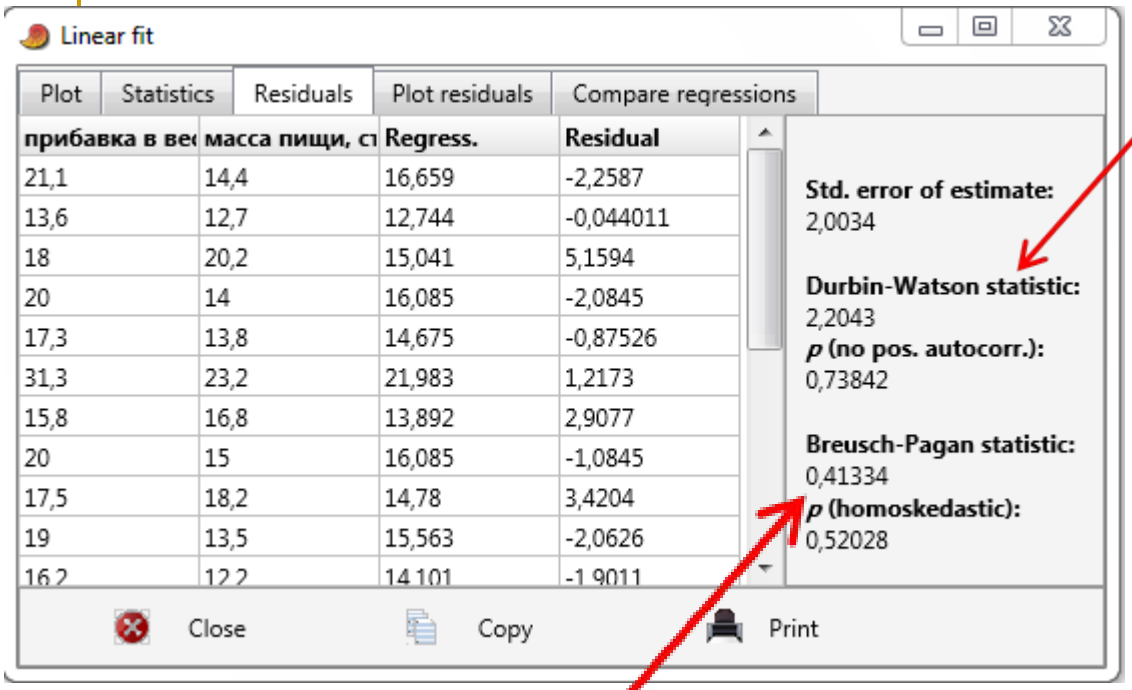


Закладки Residuals Plot residuals — анализ остатков.

Остаток — отклонение реального значения Y от предсказанного для данной точки регрессией.

Несмотря на значимость линейной регрессии, её форма может не быть оптимальной для данных.
Поэтому полезно проанализировать остатки: они должны быть случайно и нормально распределены относительно линии регрессии, со средним, равным 0.





Критерий Дарбина — Уотсона (Durbin-Watson test) проверяет случайность распределения остатков.

Если $p \leq 0{,}05$, значит существует **автокорреляция**: каждое последующее значение остатка зависит от предыдущего.

Критерий проверяет условие независимости наблюдений друг от друга – обязательное условие применения однофакторного линейного регрессионного анализа.

Критерий Бройша — Пагана (Breusch-Pagan test) проверяет однородность дисперсии остатков.

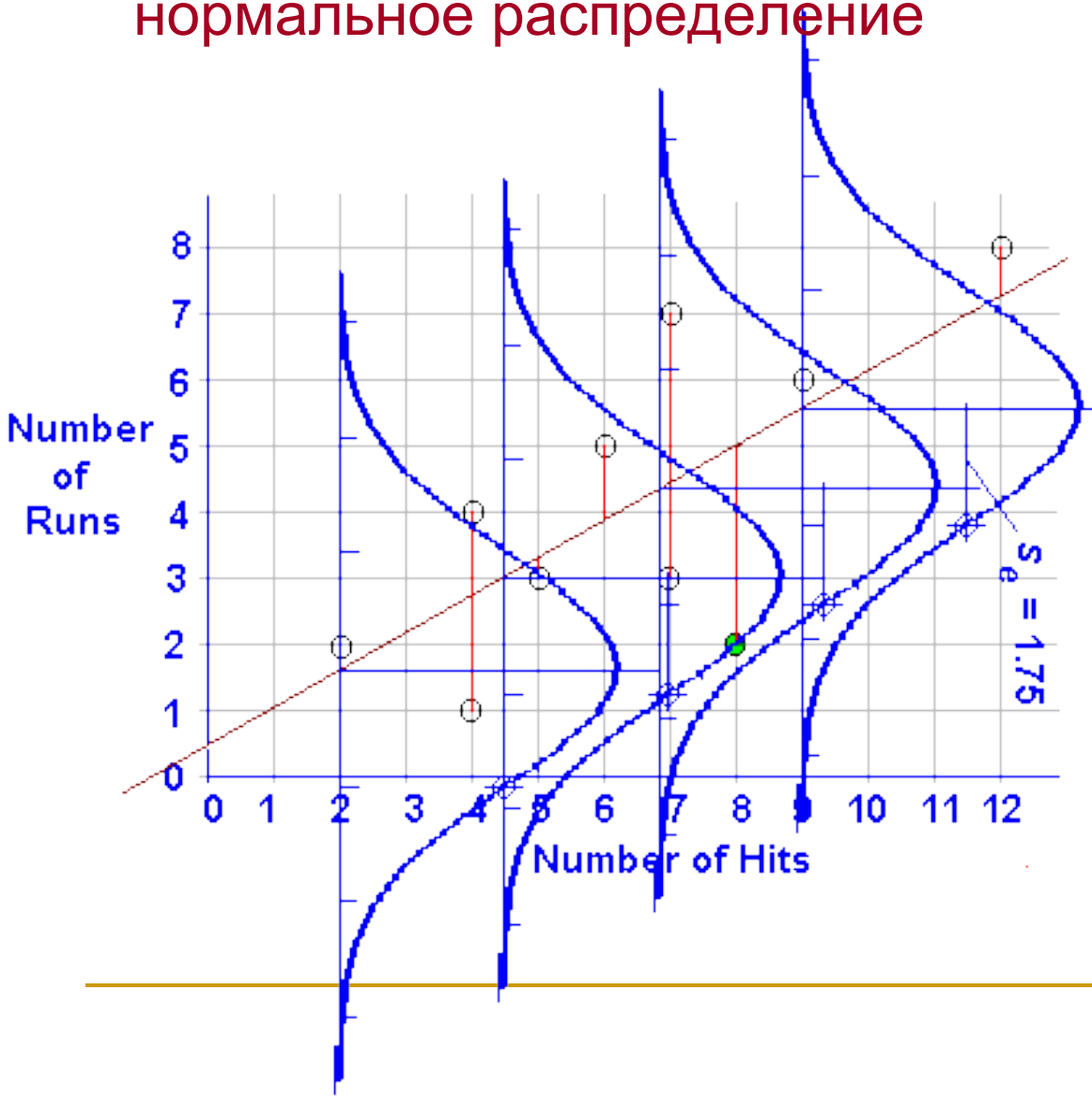
Если $p \leq 0{,}05$, значит дисперсии остатков неоднородны

Требования к выборке для проведения регрессионного анализа

1. Ожидаемая зависимость переменной Y от X должна быть линейной
2. Для любого значения X_i Y должна иметь нормальное распределение, и residuals тоже должны быть распределены нормально
3. Для любого значения X_i выборки для Y должны иметь одинаковую дисперсию (homogeneity)
4. Для любого значения X_i выборки для Y должны быть независимы друг от друга
5. Размер выборки должен более чем в 10 раз превосходить число переменных в анализе (лучше – в 20 раз)
6. Следует исключить аутлаеры

Для любого значения X_i Y должна иметь

нормальное распределение



Например, прибавка в весе для всех бегемотов, съедавших по 20 кг в день имеет нормальное распределение

Нелинейная регрессия

Иногда связь между зависимой и независимой переменной нелинейная. Например:

$$Y_i = \alpha\beta^{X_i} + \epsilon_i$$ экспоненциальный рост

$$Y_i = \alpha - \beta(e^{-\gamma X_i}) + \epsilon_i$$ асимптотическая регрессия

$$Y_i = \frac{\alpha}{1 + \beta\delta^{X_i}} + \epsilon_i$$ логистический рост

$$Y_i = \alpha X_i^{\beta}\epsilon_i$$

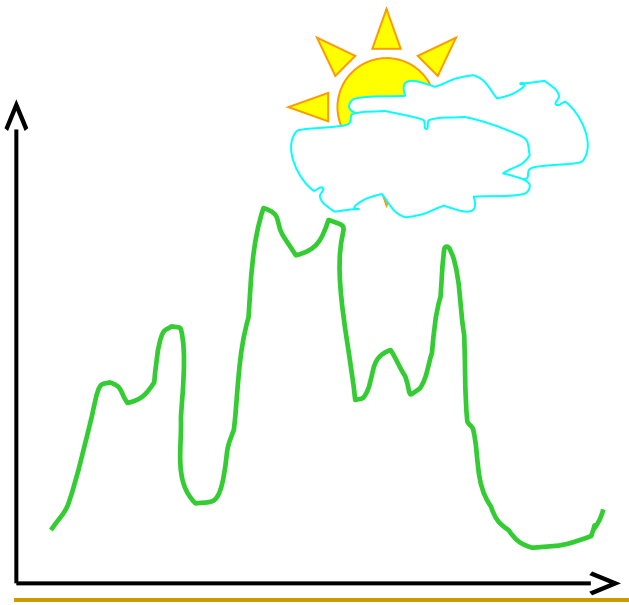
Отдельный случай – полиномиальная регрессия.

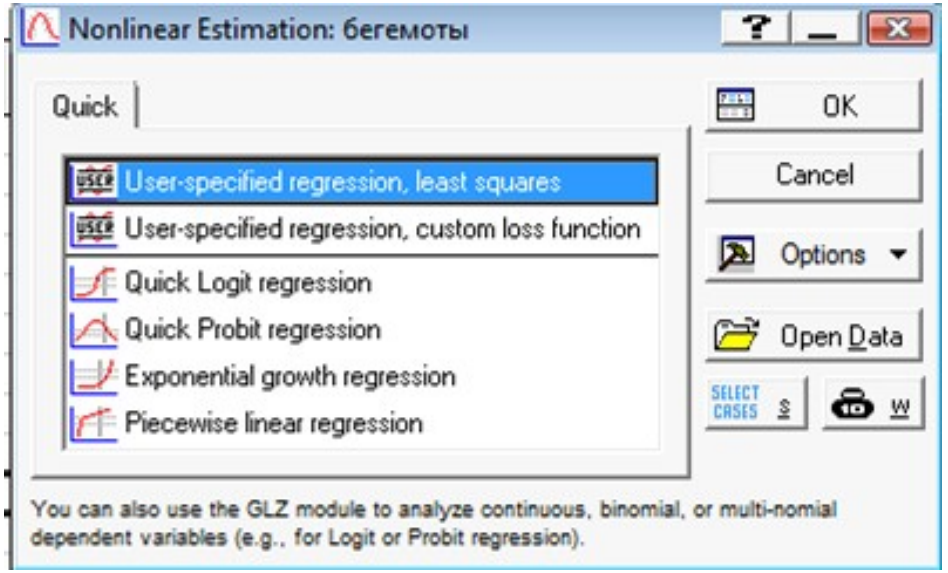
$$Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \cdots + \beta_m X_i^m + \epsilon_i$$

В статистке каждый X^m обозначают как новую переменную и дальше анализируют почти как линейную модель.

В случае, если наши переменные связаны друг с другом принципиально не линейной зависимостью:

1. можно трансформировать данные и привести зависимость к линейной (логарифмирование, извлечение квадратного корня и пр.);
2. Можно предположить (или угадать) функцию, которая их связь отражает и потом сравнить данные с ней

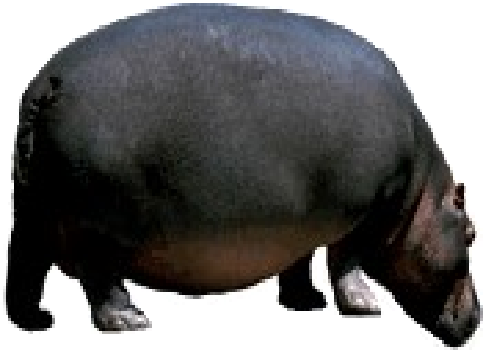




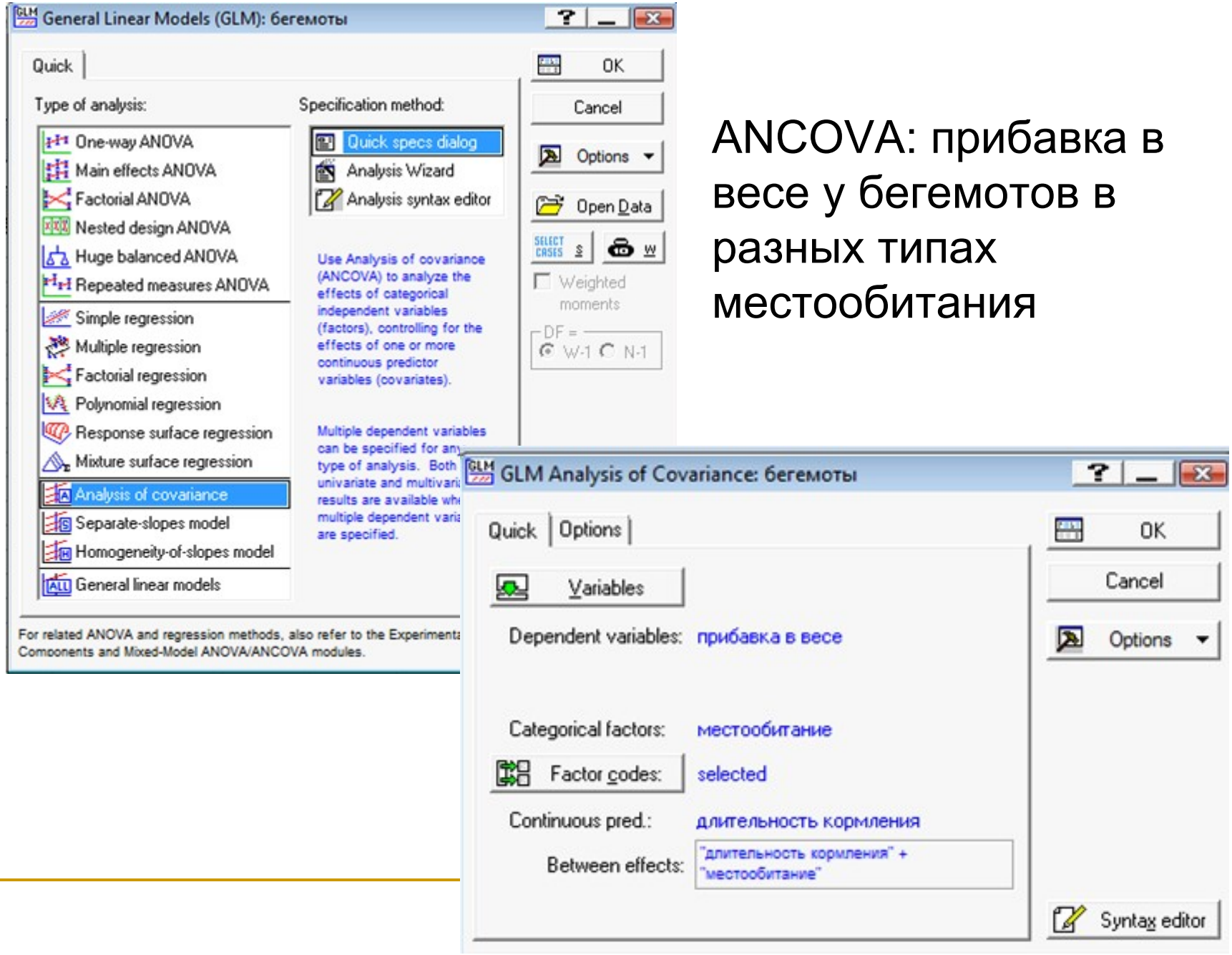
ANCOVA

Модель, когда исследуется действие **и** группирующей, и непрерывной независимых переменных на непрерывную зависимую переменную

Пример: мы анализируем влияние типа местообитания (группирующая независимая переменная) и длительности кормления (непрерывная независимая переменная) на прибавку в весе у бегемотов (непрерывная зависимая переменная).

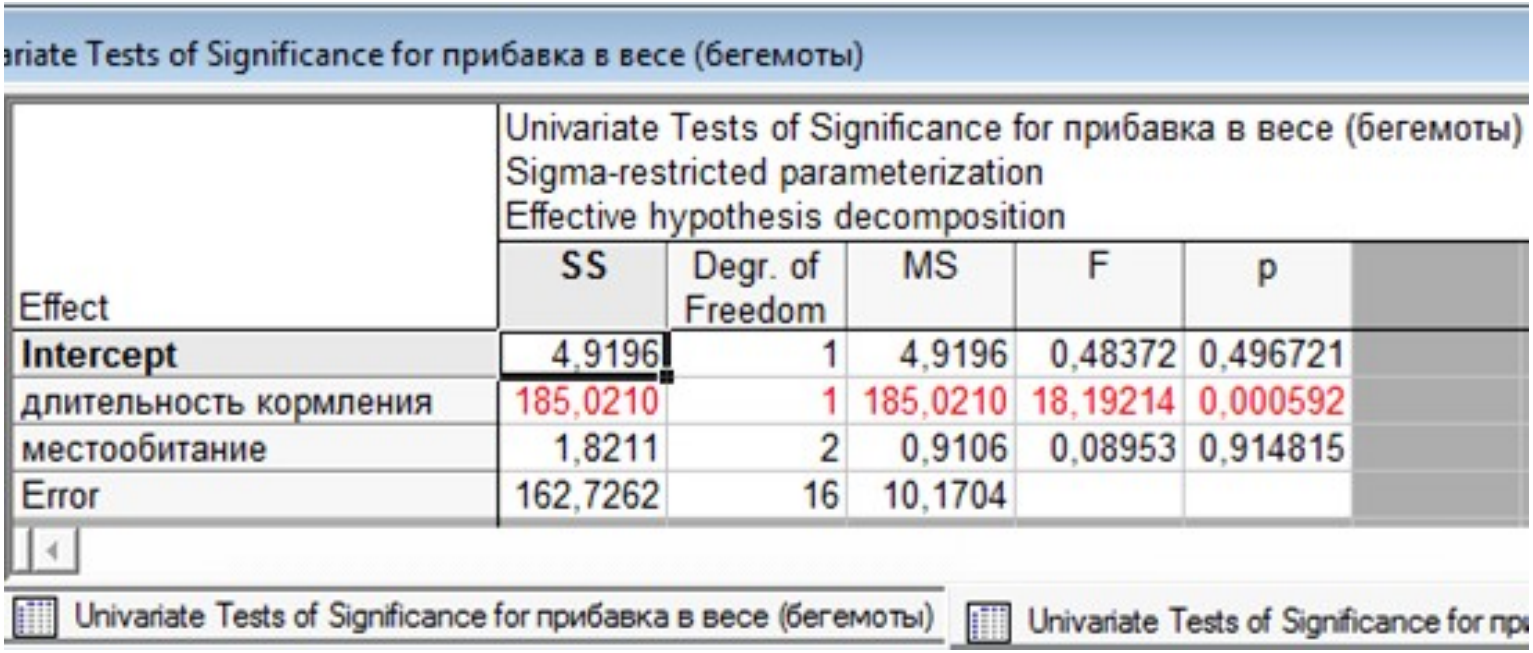


ANCOVA (analysis of covariance) – комбинированный тип анализа – ANOVA + регрессионный анализ



ANCOVA: прибавка в весе у бегемотов в разных типах местообитания

Тип местообитания не влиял на прибавку в весе, она зависела только от длительности кормления.



ariate Tests of Significance for прибавка в весе (бегемоты)

Univariate Tests of Significance for прибавка в весе (бегемоты)
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	4,9196	1	4,9196	0,48372	0,496721
длительность кормления	185,0210	1	185,0210	18,19214	0,000592
местообитание	1,8211	2	0,9106	0,08953	0,914815
Error	162,7262	16	10,1704		

Univariate Tests of Significance for прибавка в весе (бегемоты)

Univariate Tests of Significance for пр

Выбор модели в GLM (Обобщенных Линейных Моделях)

Независимые переменные	Зависимые переменные	Модель
Одна группирующая	Одна непрерывная	One-way ANOVA Однофакторный дисперсионный анализ
Много группирующих	Одна непрерывная	Factorial ANOVA (two-, multiway). Main effect ANOVA Многофакторный дисперсионный анализ
Одна или много группирующих	Много непрерывных	MANOVA (multivariate ANOVA) Многомерный дисперсионный анализ
Одна непрерывная	Одна непрерывная	Simple regression Простая линейная регрессия
Много непрерывных	Одна непрерывная	Multiple regression Множественная линейная регрессия
Одна группирующая (или много) + одна непрерывная (или много)	Одна непрерывная	ANCOVA ANOVA + регрессионный анализ

«Много» = 2 и больше

Спасибо за внимание!