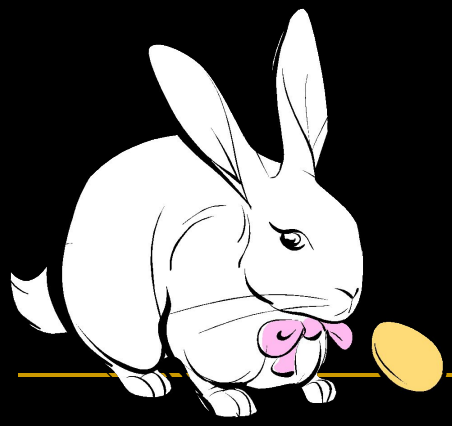
Занятие 6

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).

Настало время обратиться к ситуации, когда зависимых переменных будет ДВЕ и более.

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

Это могут быть измерения одной особи или связанных пар.

* кроме MANOVA

Корреляции

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

Переменные – 1. масса; 2. длина хвоста.

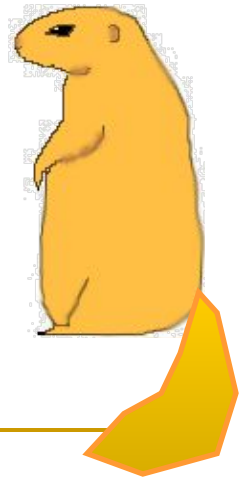
Корреляции

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

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

Корреляции

Рост братьев: коэффициент корреляции r -?

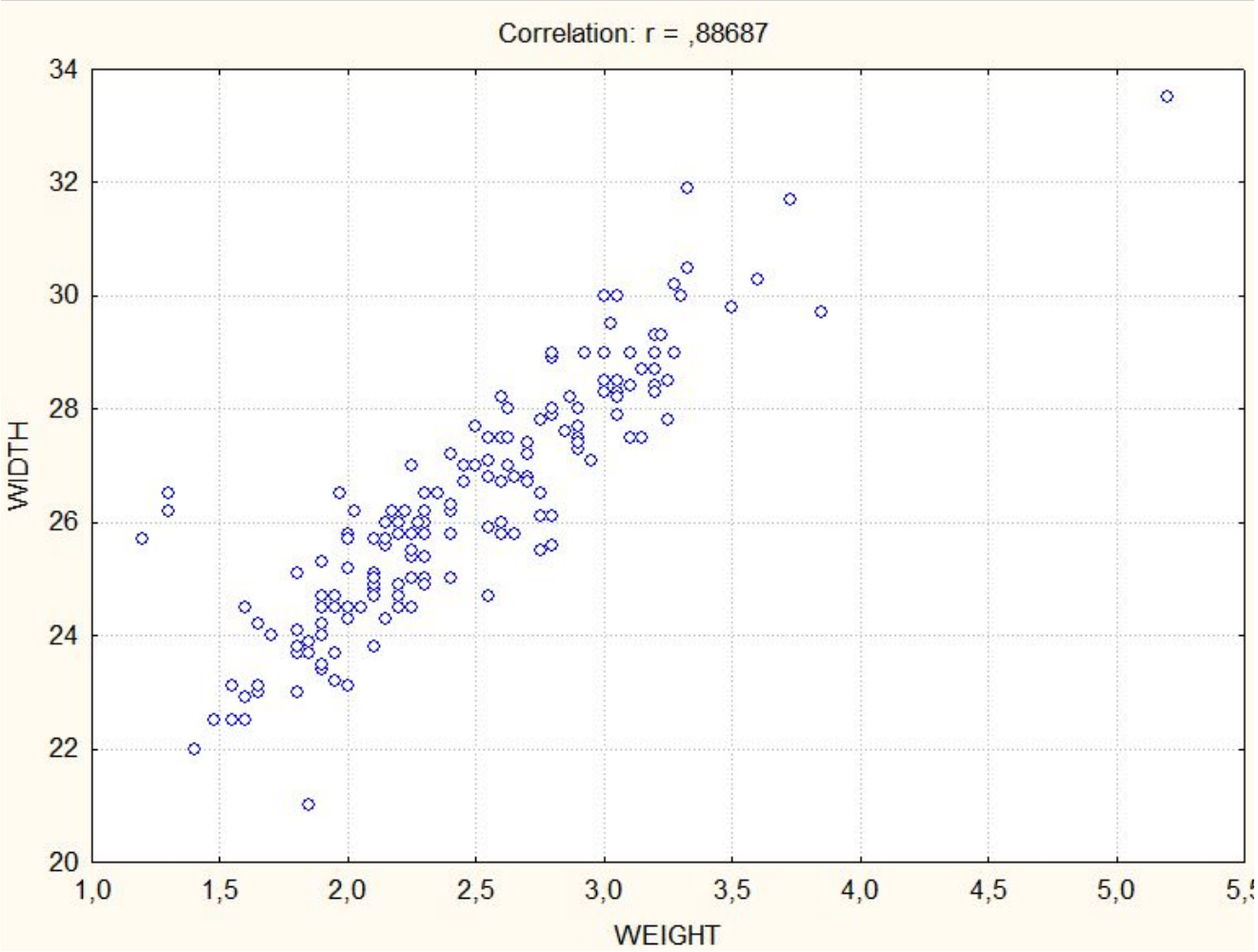
Петя

Гриша

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

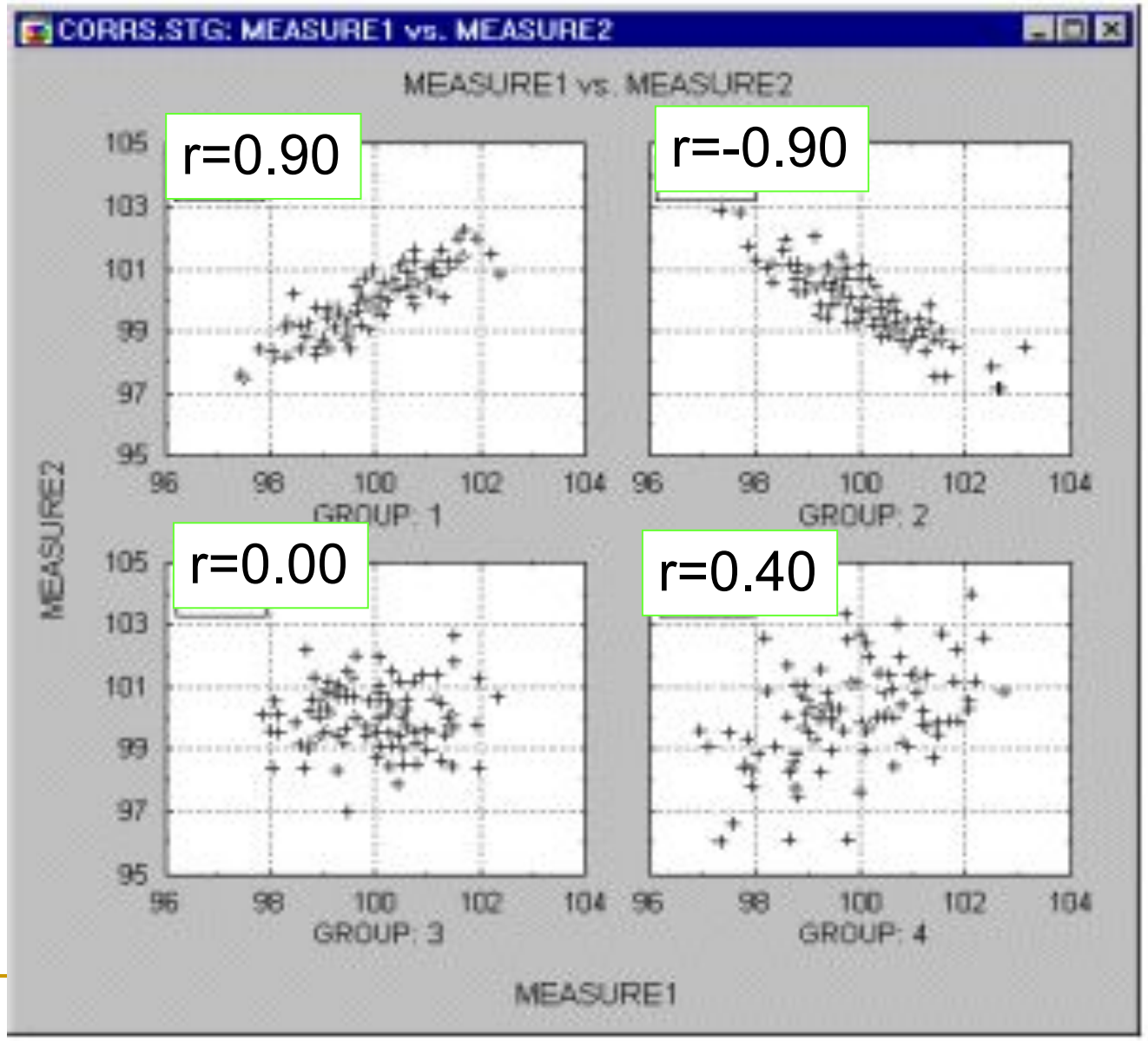
Скаттерплот

(= диаграмма рассеяния; scatterplot, scatter diagram)



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

Корреляции


CORRS.STG: MEASURE1 vs. MEASURE2
MEASURE1 vs. MEASURE2
r=0.90
r=-0.90
r=0.00
r=0.40
MEASURE2
MEASURE1
GROUP: 1
GROUP: 2
GROUP: 3
GROUP: 4
105
103
101
99
97
95
96
98
100
102
104

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

Karl Pearson (1857 –1936)

Корреляции

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 (см. занятие 1)

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

Корреляции

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

параметр ВЫБОРКИ

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

параметр ПОПУЛЯЦИИ

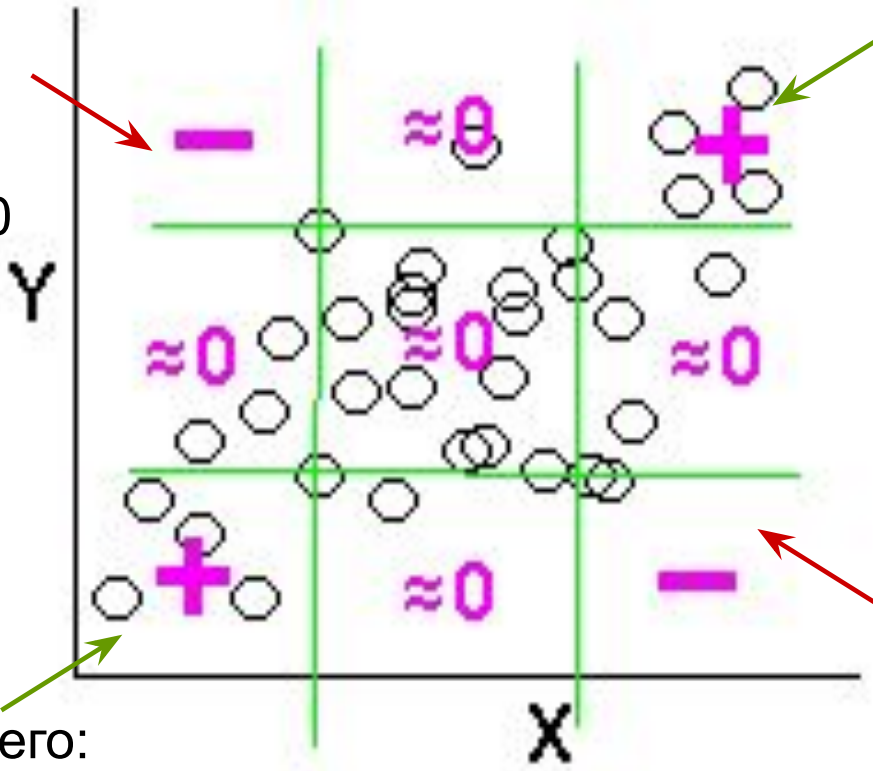
Всё как для других параметров описательной статистики: среднего, дисперсии, и т.д.!

Что определяет $\sum z_X z_Y$?

Корреляции

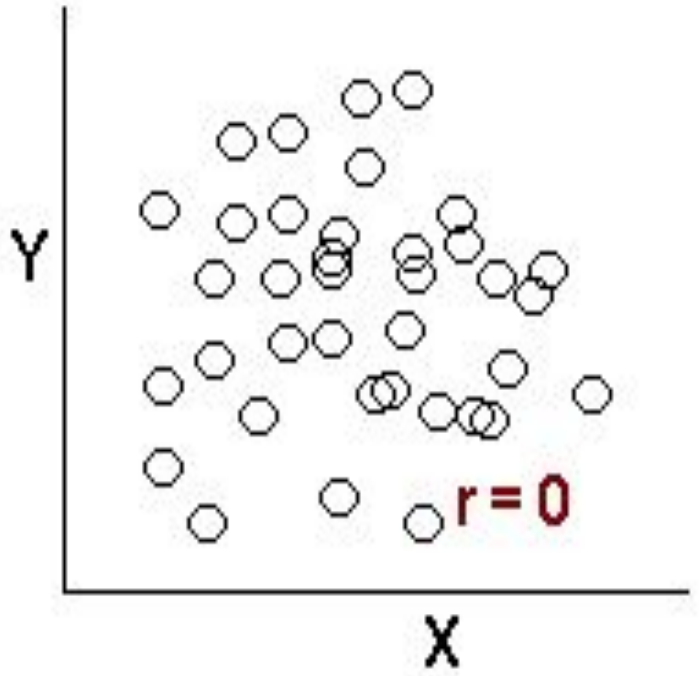
Чем определяются знак и величина коэффициента корреляции?

Знаком и величиной $\sum z_X z_Y$:



Корреляции

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



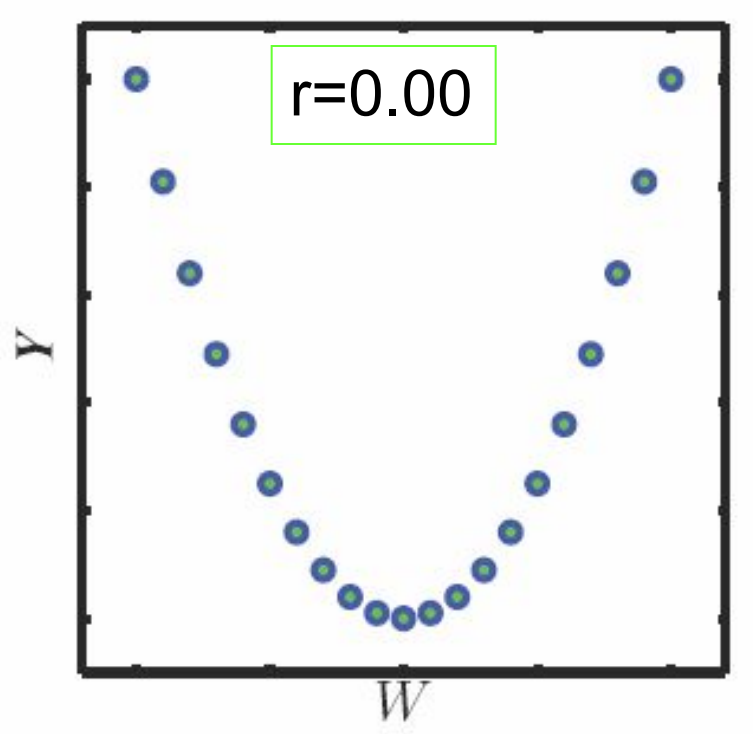
Здесь и впрямь её нет

НО это не всегда так, есть исключения.

Корреляции

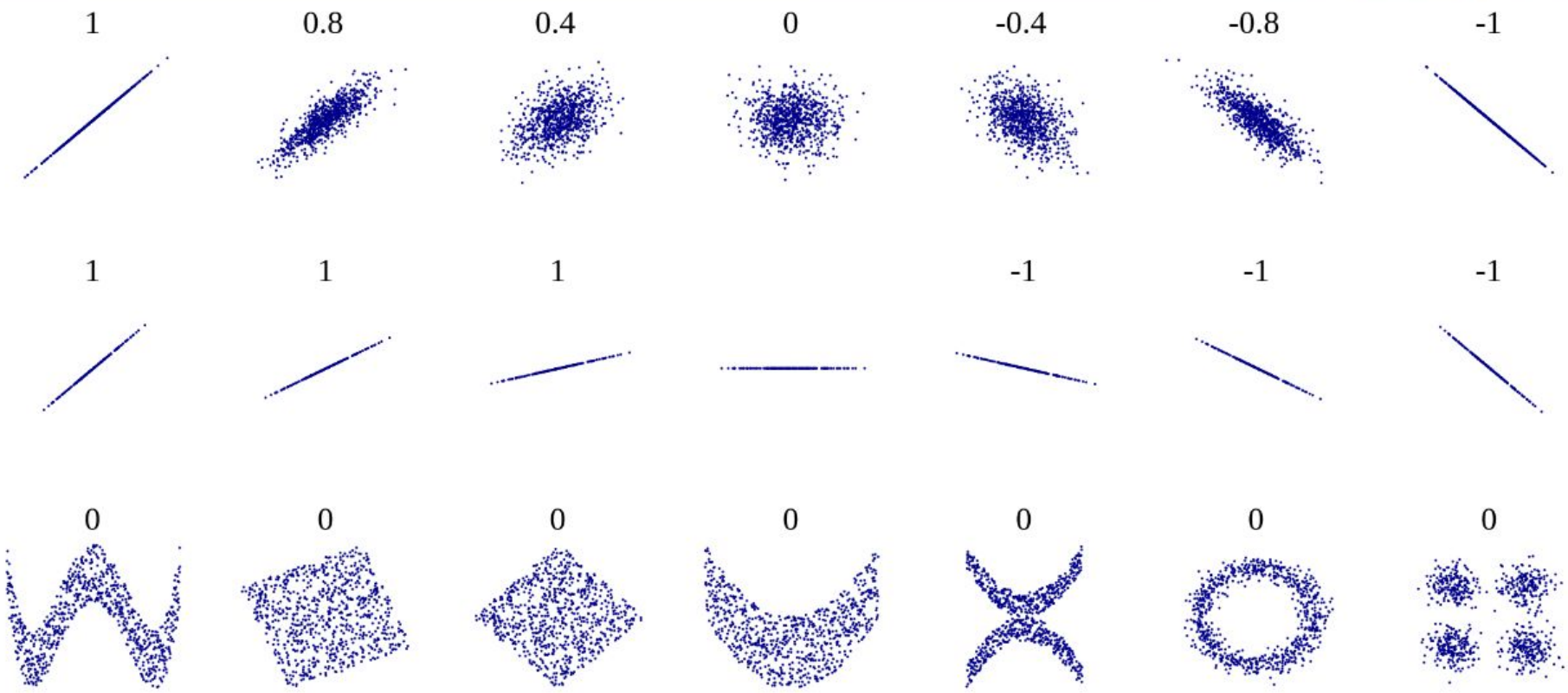
Факторы, влияющие на коэффициент корреляции

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



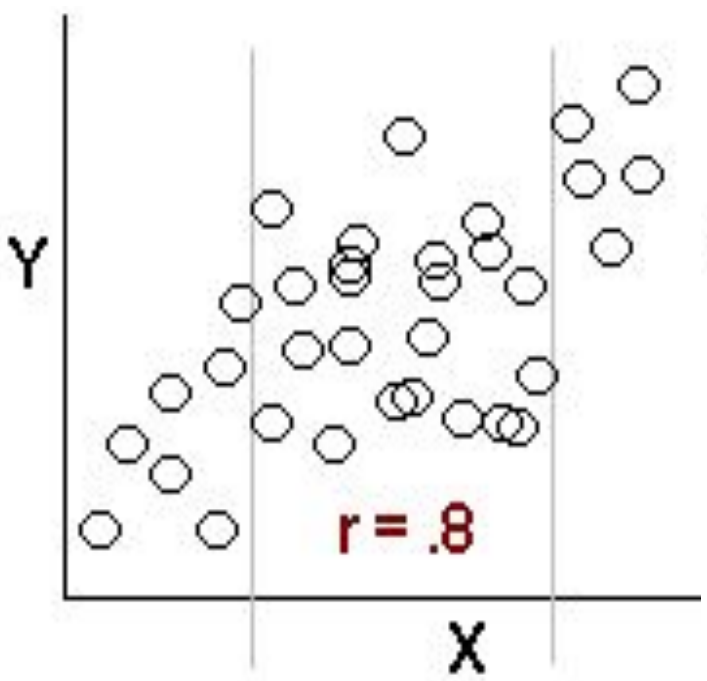
Здесь связь переменных есть,
и она очень сильная, но r=0.00

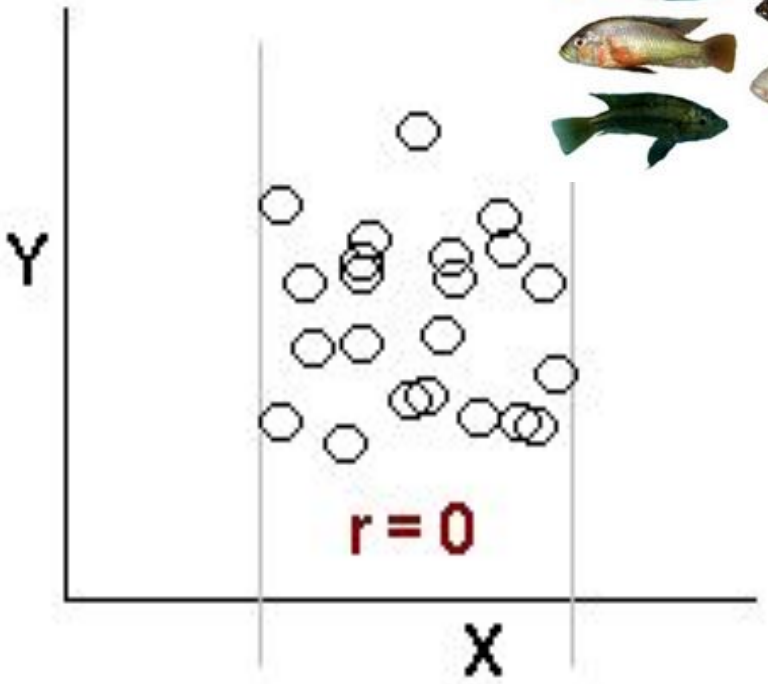
Корреляции



Корреляции

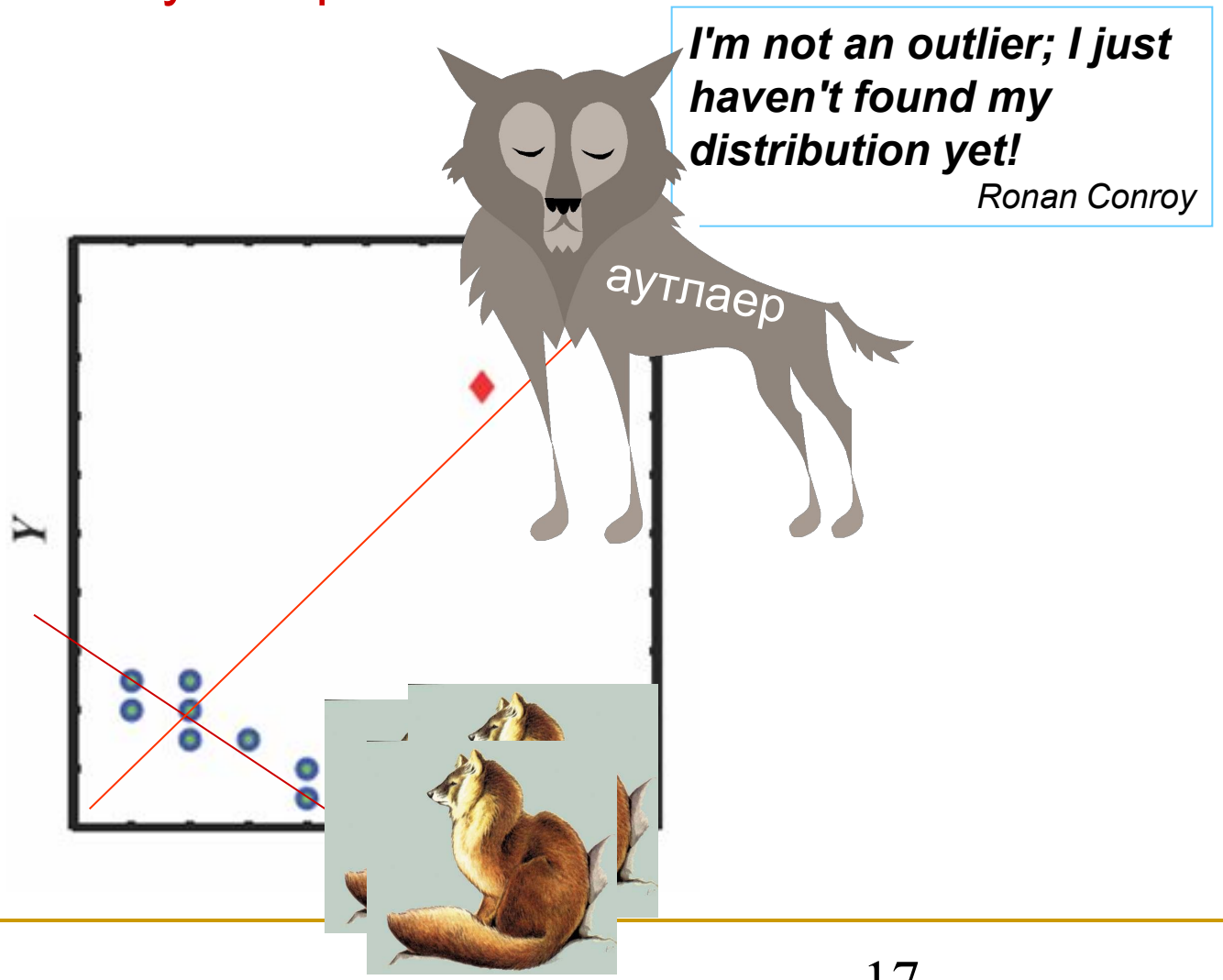
2. Необходимо, чтобы у переменных была значительная изменчивость! Если сформировать выборку изначально однотипных особей, нечего надеяться выявить там корреляции.





Корреляции

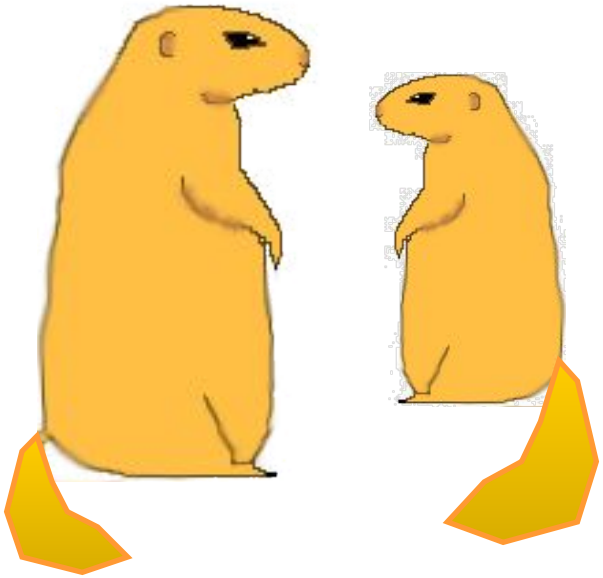
3. Коэффициент корреляции Пирсона очень чувствителен к аутлаерам.



Важное замечание:

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

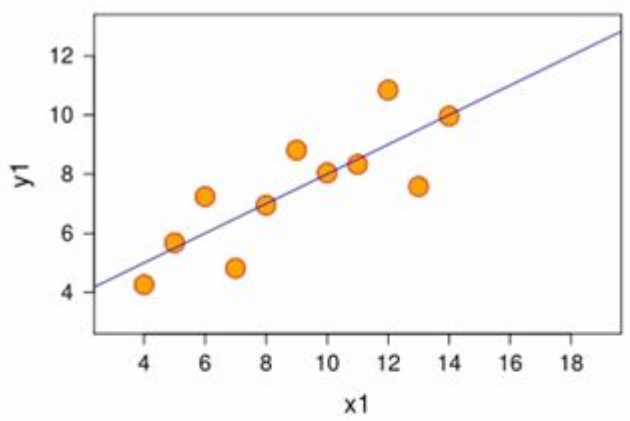
Она **ВООБЩЕ НИЧЕГО** о ней **НЕ ГОВОРИТ** (даже очень большой r)

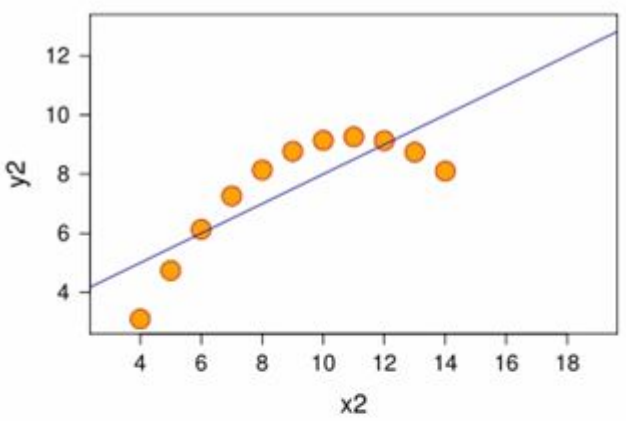


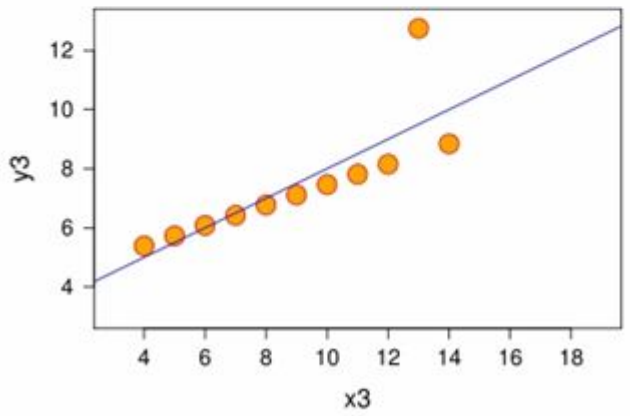
Корреляции

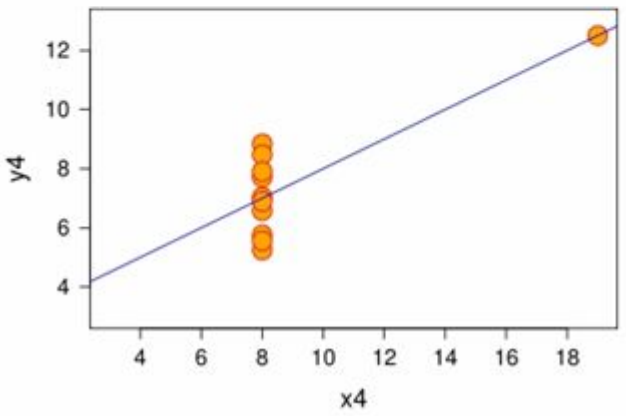
Коэффициент корреляции Пирсона – параметр выборки.
Можем ли мы на основе него судить о популяции?
Просто глядя на коэффициент – НЕТ.

Correlation
between each x and
y = **0.816**









Корреляции

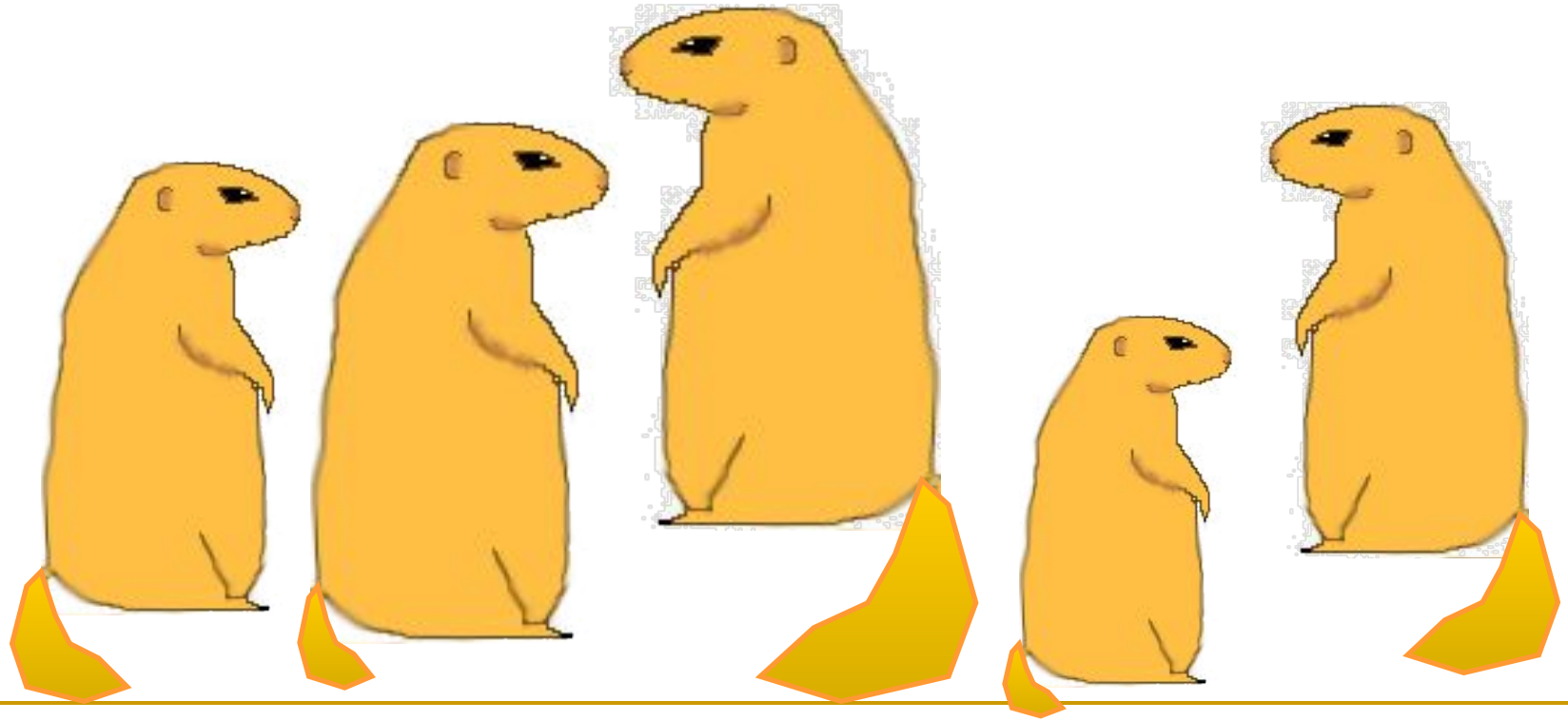
Мы хотим оценить коэффициент корреляции в популяции.

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

(альтернативная гипотеза может быть односторонней)

Связаны ли у сусликов масса тела и длина хвоста?



Корреляции

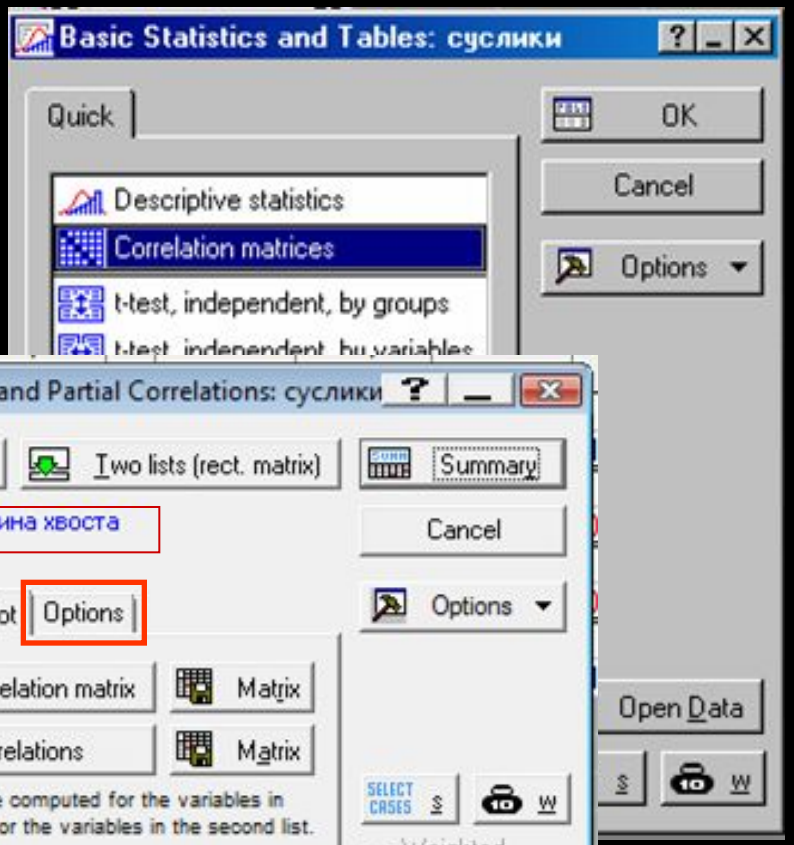
$$\text{Статистика} = \frac{\text{параметр выборки} - \text{параметр популяции}}{\text{стандартная ошибка параметра выборки}}$$

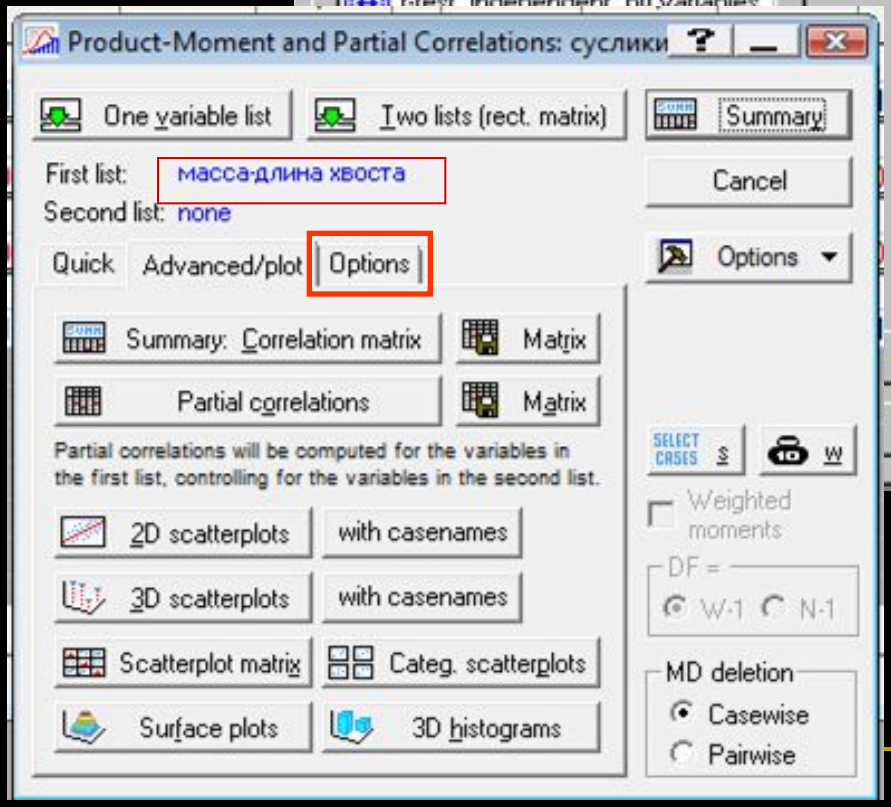
$$t = \frac{r - \rho}{s_r} \longrightarrow t = \frac{r}{s_r}$$

стандартная ошибка
коэффициента корреляции

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

Pearson product-moment correlation coefficient *r*

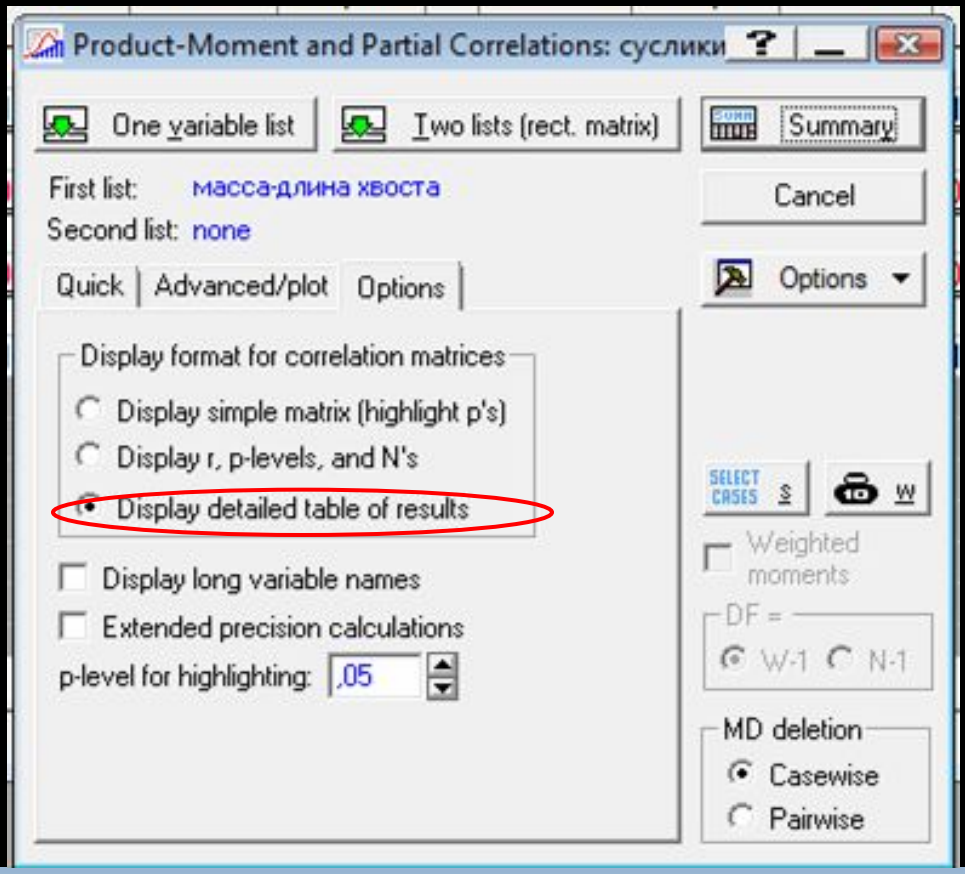




Data: суслики* (11v by 20c)

	1 зверёк	2 масса	3 длина хвоста
1	1	21,5	21,11
2	2	13,8	13,64
3	3	16,8	18,00
4	4	13,5	20,00
5	5	14,0	17,27
6	6	20,2	31,25
7	7	14,1	15,83
8	8	13,0	20,00
9	9	11,3	17,50
10	10	12,2	16,15
11	11	12,2	16,15
12	12	10,8	15,71
13	13	12,1	15,71
14	14	14,4	15,33
15	15	12,2	14,67
16	16	12,2	14,67
17	17	13,2	24,17
18	18	15,6	28,18
19	19	10,6	16,00
20	20	12,7	19,29

Отвергаем H_0:
Оказалось, что масса тела у сусликов положительно связана с длиной хвоста.



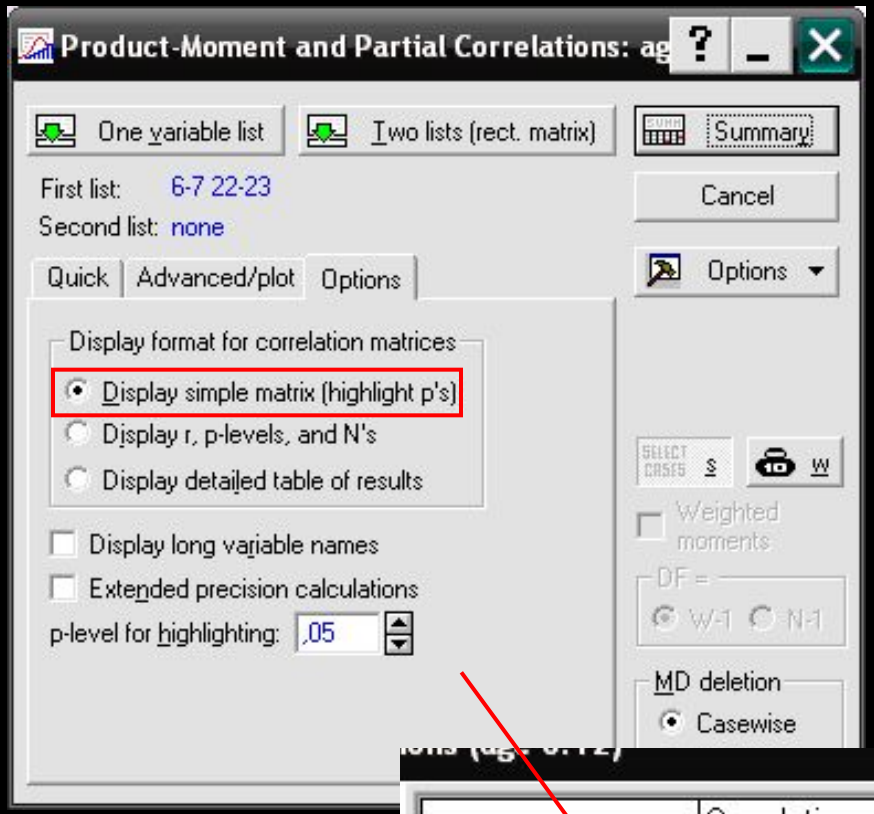
elations (суслики)

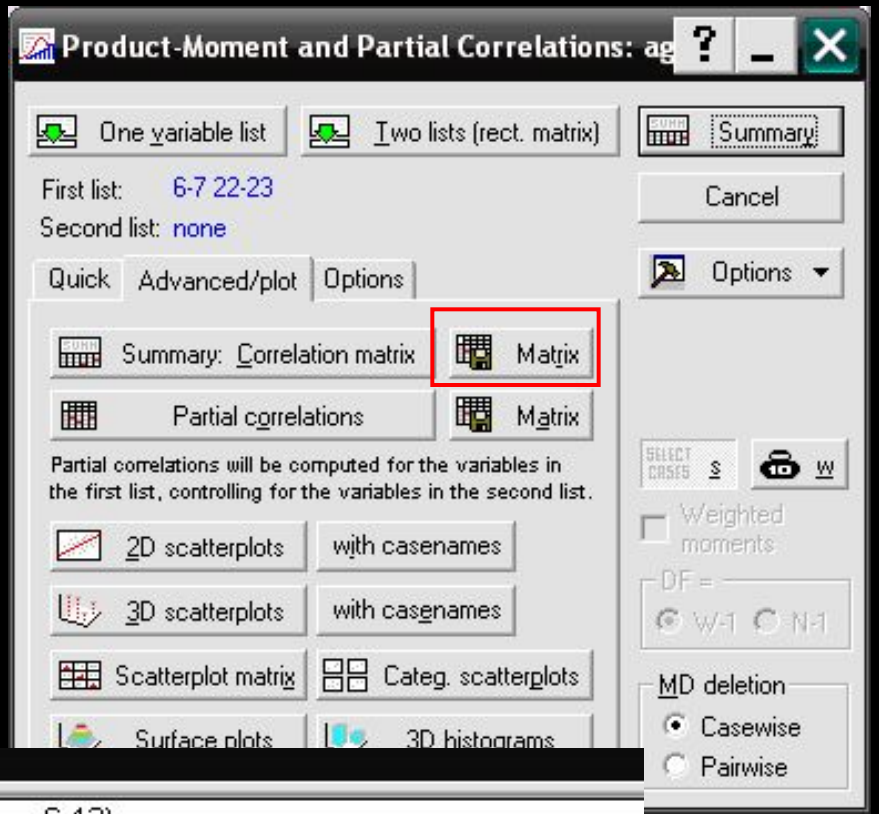
Correlations (суслики)
Marked correlations are significant at p < ,05000
(Casewise deletion of missing data)

Var. X & Var. Y	Mean	Std.Dv.	r(X,Y)	r?	t	p	N	Constant dep: Y	Slope dep: Y	Constant dep: X	Slope dep: X
масса	13,82845	2,838194									
длина хвоста	18,53203	4,622850	0,611508	0,373942	3,278920	0,004171	20	4,758573	0,996024	6,870880	0,375435

Коэффициенты a и b

Бывают задачи, когда нам необходимо получить МАТРИЦУ КОРРЕЛЯЦИЙ (для многомерных методов анализа)

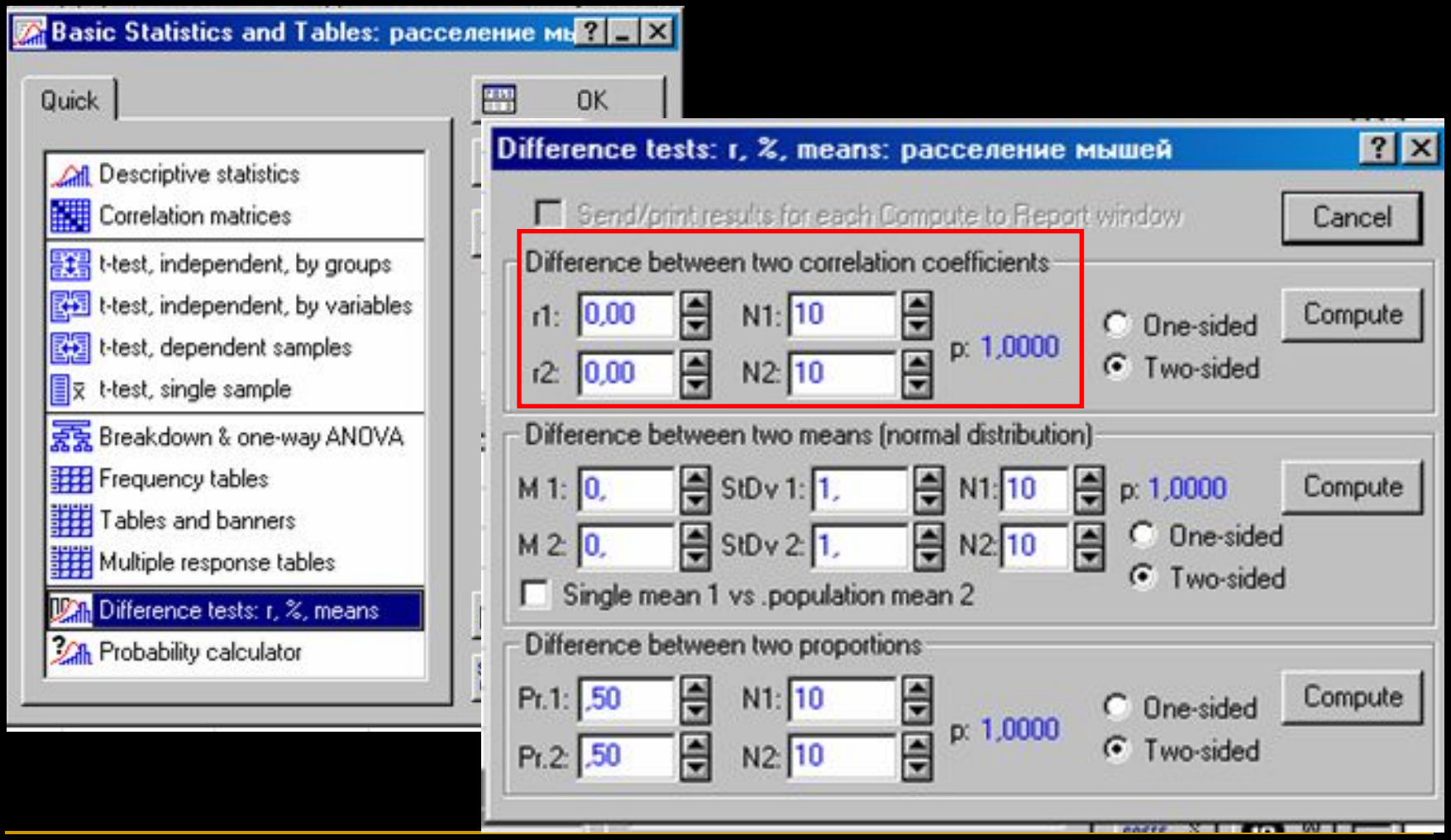




Correlations (age 6.12)
Marked correlations are significant at p < ,05000
N=14 (Casewise deletion of missing data)

Variable	масса	упитанность	масса детёныша	масса выводка
масса	1,00	0,98	-0,03	-0,35
упитанность	0,98	1,00	-0,02	-0,27
масса детёныша	-0,03	-0,02	1,00	0,40
масса выводка	-0,35	-0,27	0,40	1,00

Можно сравнить два коэффициента корреляции от двух выборок



Для двумерного нормального распределения

Корреляции

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

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

Cohen, 1988:

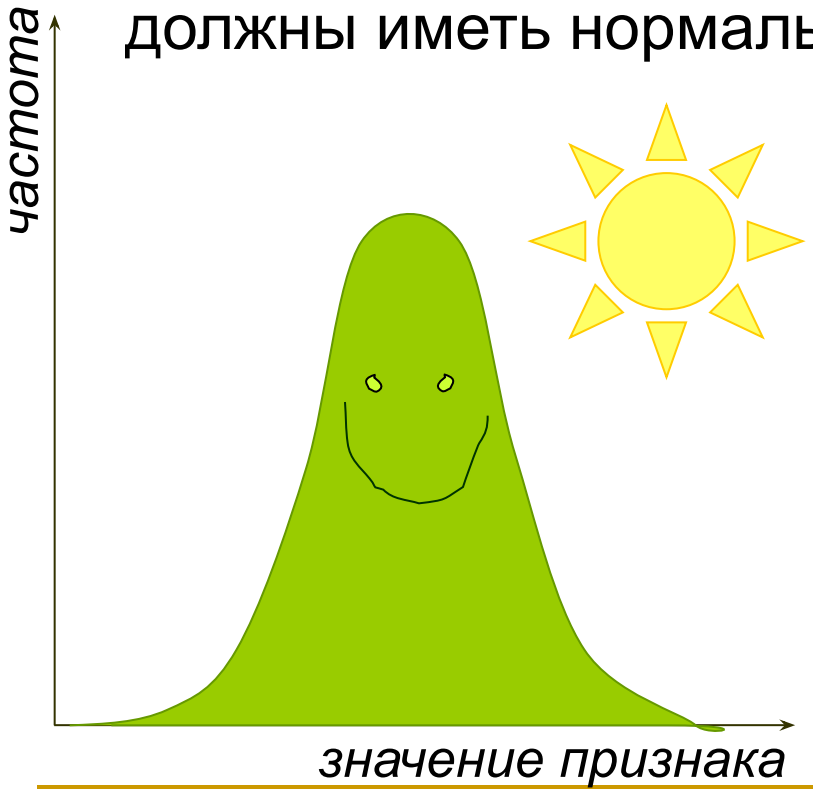
$\rho = 0.1$ - слабая корреляция;

$\rho = 0.3$ – корреляция средней силы;

$\rho = 0.5$ - сильная корреляция.

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

Рост братьев.

Петя

Гриша

r=0.7: если Петя высокий, то, ***скорее всего***, Гриша тоже высокий. Но можем ли мы предсказать, ***насколько высокий***? Сам коэффициент корреляции этого нам не скажет.

Ответ нам даст РЕГРЕССИОННЫЙ АНАЛИЗ.

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).
Предсказанное значение Y обычно обозначают как $\hat{Y}$

Регрессии

То есть,

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!

Регрессии

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?

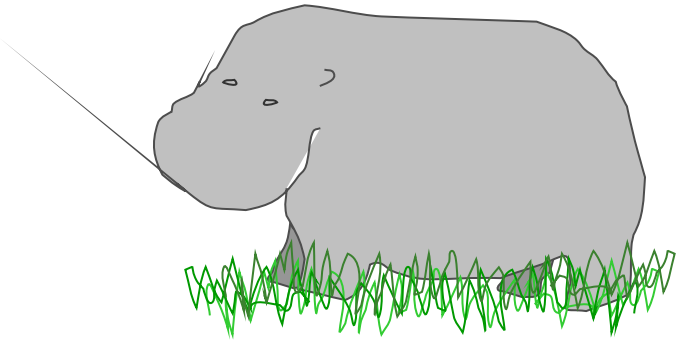
У нас **две** переменные – 1. кол-во съедаемой в день пищи, кг (independent); 2. прибавка в весе за месяц, кг (dependent)

Регрессии

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

Регрессии

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

Пояснить размерность b и a

Регрессии

Задача сводится к поиску коэффициентов a и b.

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

стандартные отклонения для X и Y

коэффициент корреляции Пирсона

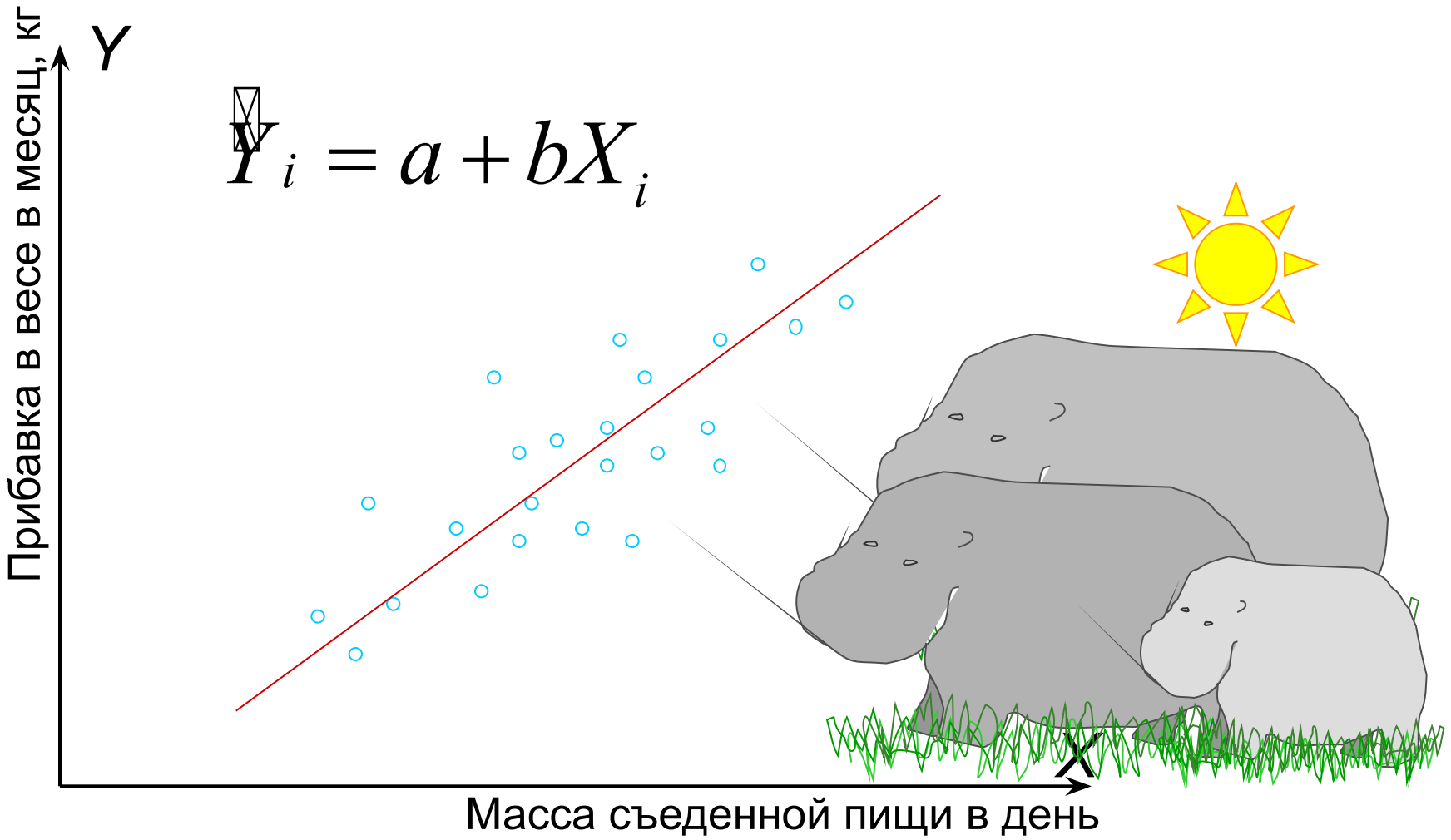
$$\bar{Y} = a + b\bar{X} \longrightarrow a = \bar{Y} - b\bar{X}$$

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.

Пример с кол-вом удобрения на каждый кг помидоров

Регрессии


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

Регрессии

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**.

Пример про Dr. Nostat, который отобрал 100 самых глупых учеников, подверг их специальной программе и потом протестировал повторно, и их IQ оказался в среднем выше.

Пример про очень умную 5-летнюю девочку

Регрессии

Линия регрессии в стандартной форме

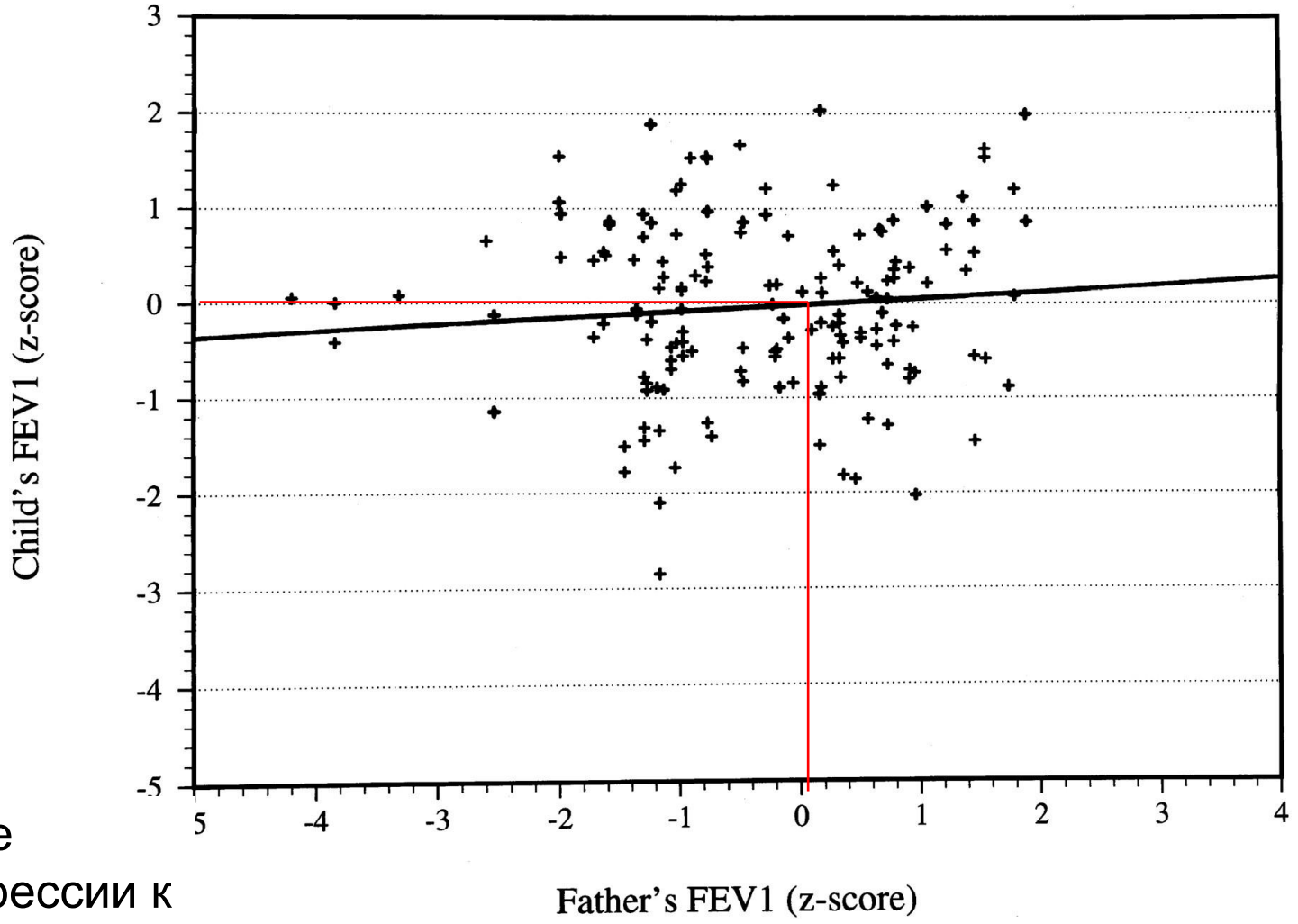
$$a = \bar{Y} - b\bar{X}$$

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

a = 0, b = r

$$Y_z = rX_z$$

(математическое объяснение регрессии к среднему)



Регрессии

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.

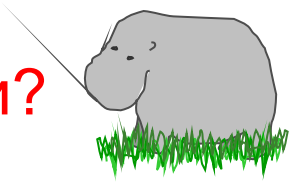
$$Y_i = \alpha + \beta X_i + \varepsilon_i$$ Для популяции

$$Y_i = a + bX_i + e_i$$ Для выборки

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

Регрессии

Как определить «лучшую» линию регрессии?

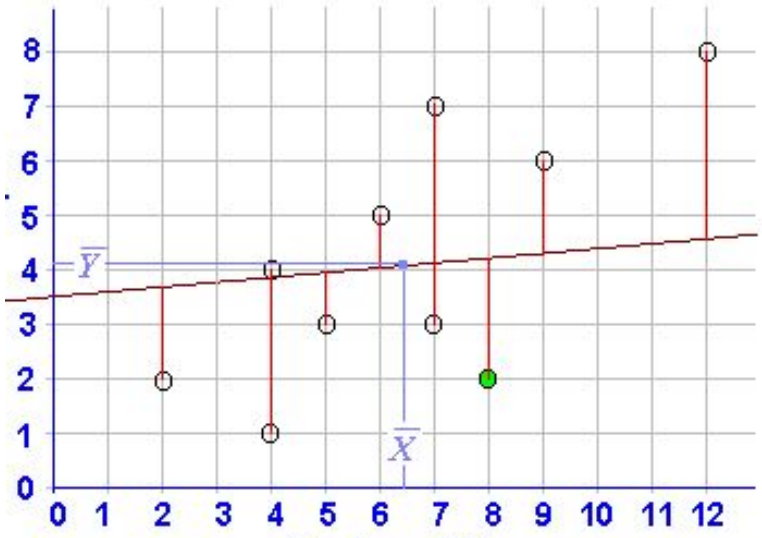


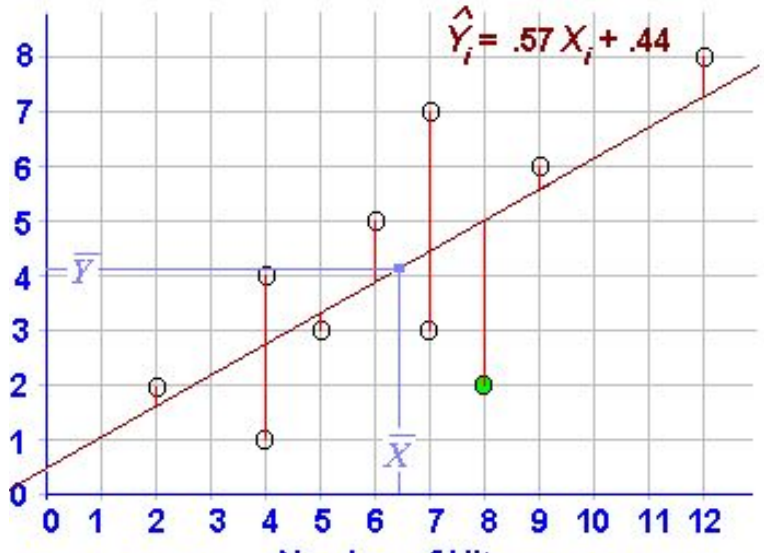
Метод наименьших квадратов:

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

Регрессии

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)

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

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

Можно тестировать гипотезу и о том, что intercept ( ) = 0

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

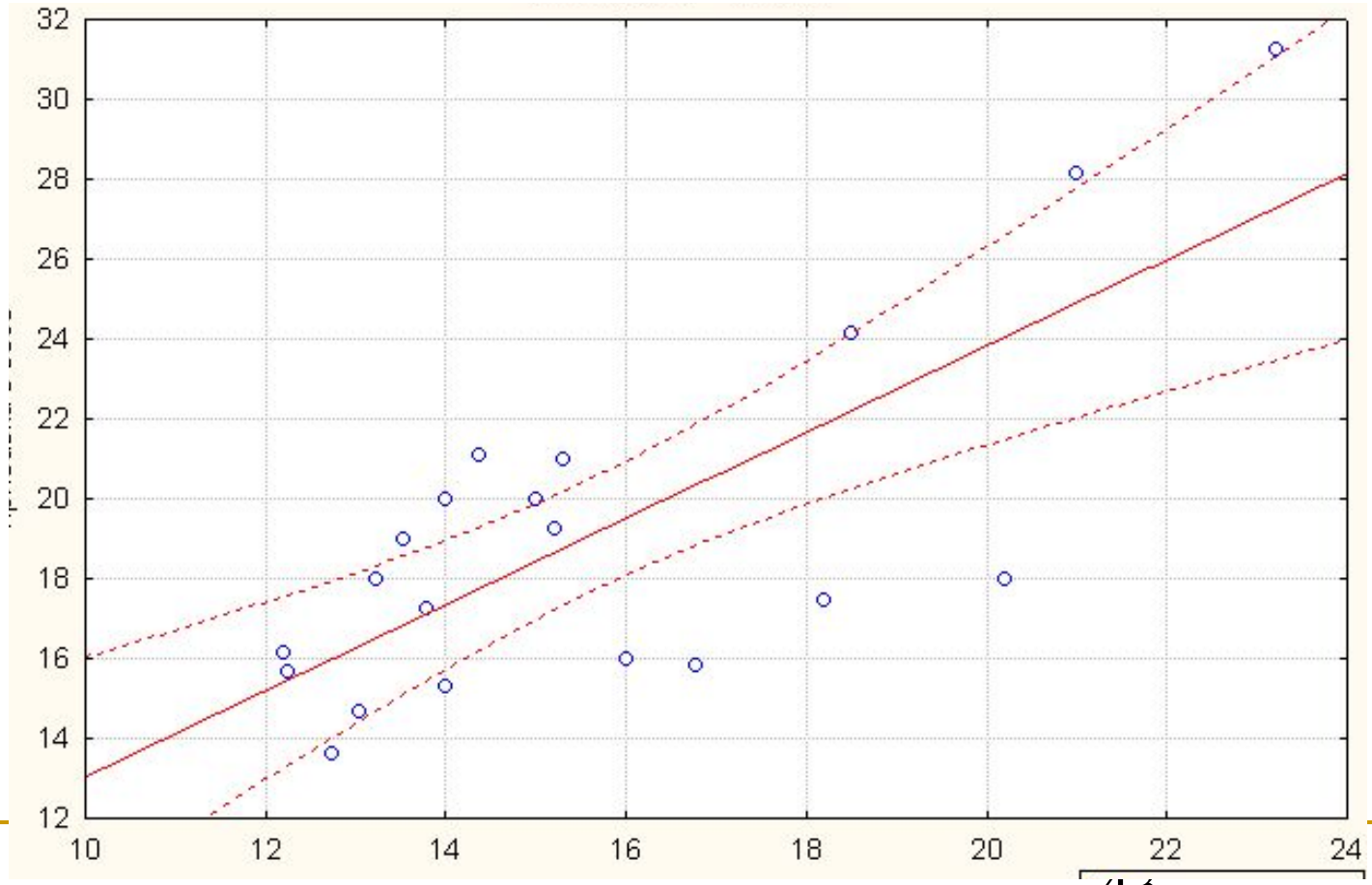
$$R^2 = \frac{SS_{regression}}{SS_{total}} = \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2}$$

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

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

Регрессии

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



Регрессии

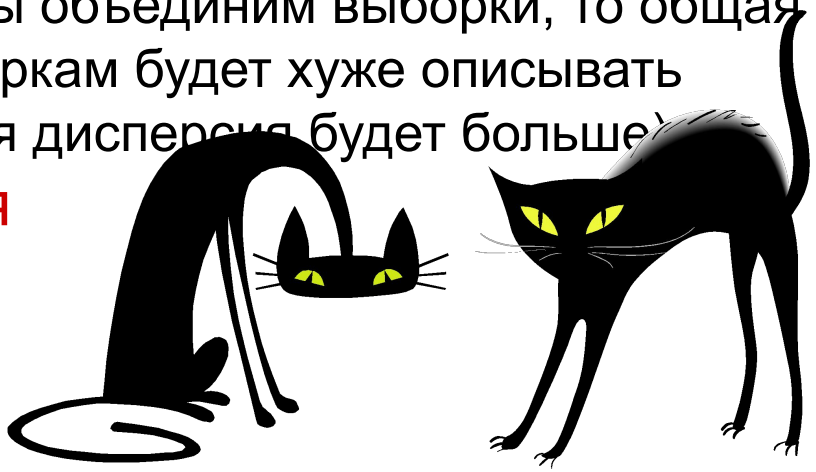
Сравнение двух (и более) уравнений линейной регрессии

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



линии регрессии

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 взаимосвязь нескольких переменных, среди которых невозможно выделить зависимую.

Регрессии

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).

Регрессии

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.

Регрессии

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.

Регрессии

Коэффициент детерминации (coefficient of determination)

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

Считается потому 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:
лучше, чем просто R^2, так как не увеличивается с ростом кол-ва переменных в модели

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

Добавление переменных в модель:

- $SS_{regression}$ увеличивается, поэтому R^2 растёт.
- При этом F может уменьшаться.

Для каждой переменной по отдельности можно протестировать гипотезу $\beta = 0$ -

Partial regression coefficients.

Регрессии

Multicollinearity = ill-conditioning

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;
- ✔ общее F для всей модели достоверно, а отдельные t-тесты для каждой переменной – нет;
- ✔ при пошаговом анализе выбирая разные способы анализа мы получаем разные результаты.

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

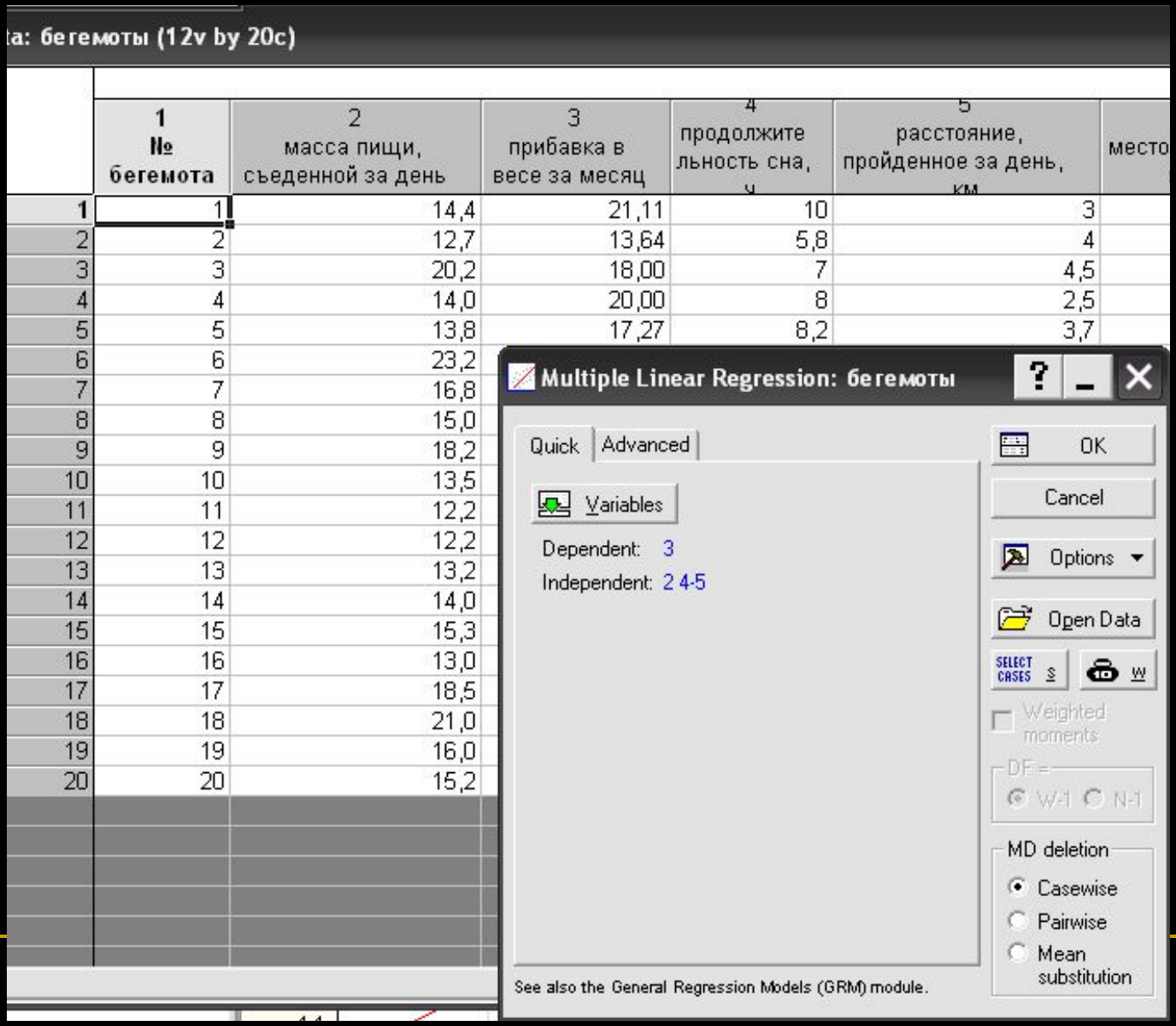
Выбор «лучших» независимых переменных

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

Используют пошаговые модели:

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

Simple linear regression



a: бегемоты (12v by 20c)

	1 № бегемота	2 масса пищи, съеденной за день	3 прибавка в весе за месяц	4 продолжительность сна,	5 расстояние, пройденное за день, км	место
1	1	14,4	21,11	10	3	
2	2	12,7	13,64	5,8	4	
3	3	20,2	18,00	7	4,5	
4	4	14,0	20,00	8	2,5	
5	5	13,8	17,27	8,2	3,7	
6	6	23,2				
7	7	16,8				
8	8	15,0				
9	9	18,2				
10	10	13,5				
11	11	12,2				
12	12	12,2				
13	13	13,2				
14	14	14,0				
15	15	15,3				
16	16	13,0				
17	17	18,5				
18	18	21,0				
19	19	16,0				
20	20	15,2				

Multiple Linear Regression: бегемоты

Quick | Advanced

Variables

Dependent: 3

Independent: 2 4-5

OK

Cancel

Options

Open Data

SELECT CASES s

w

Weighted moments

DF =

W-1 N-1

MD deletion

Casewise

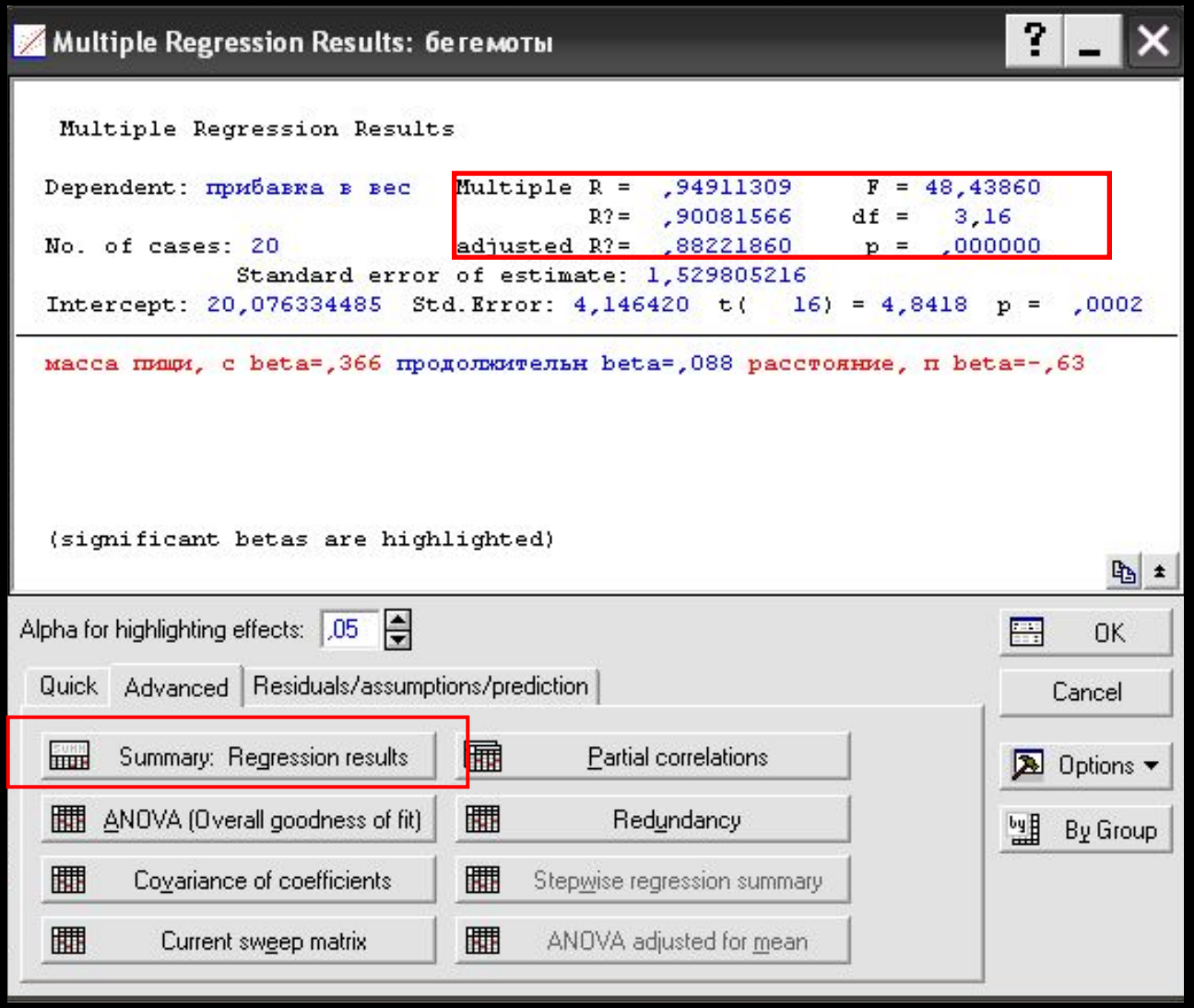
Pairwise

Mean substitution

See also the General Regression Models (GRM) module.

linear regression

У бегемотов прибавка в весе зависела от этих переменных



n Summary for Dependent Variable: прибавка в весе за месяц (бегемоты)

Regression Summary for Dependent Variable: прибавка в весе за месяц (бегемоты)
R= ,94911309 R?= ,90081566 Adjusted R?= ,88221860
F(3,16)=48,439 p<,00000 Std.Error of estimate: 1,5298

N=20	Beta	Std.Err. of Beta	B	Std.Err. of B	t(16)	p-level
Intercept			20,07633	4,146420	4,84185	0,000180
масса пищи, съеденной за день	0,366070	0,099078	0,52591	0,142339	3,69478	0,001964
продолжительность сна, ч	0,087805	0,119511	0,19430	0,264456	0,73470	0,473152
расстояние, пройденное за день, км	-0,631234	0,114110	-3,16357	0,571888	-5,53179	0,000046

Коэффициенты наклона в стандартной форме

Коэффициенты a и b

Часто «остатки» используют как самостоятельную переменную

Multiple Regression Results: бегемоты

Multiple Regression Results

Dependent: прибавка в вес Multiple R = ,94911309 F = 48,43860
R?= ,90081566 df = 3,16
No. of cases: 20 adjusted R?= ,88221860 p = ,000000
Standard error of estimate: 1,529805216
Intercept: 20,076334485 Std.Error: 4,146420 t(16) = 4,8418 p = ,0002

масса пищи, с beta=,366 продолжительн beta=,088 расстояние, п beta=-,63

(significant betas are highlighted)

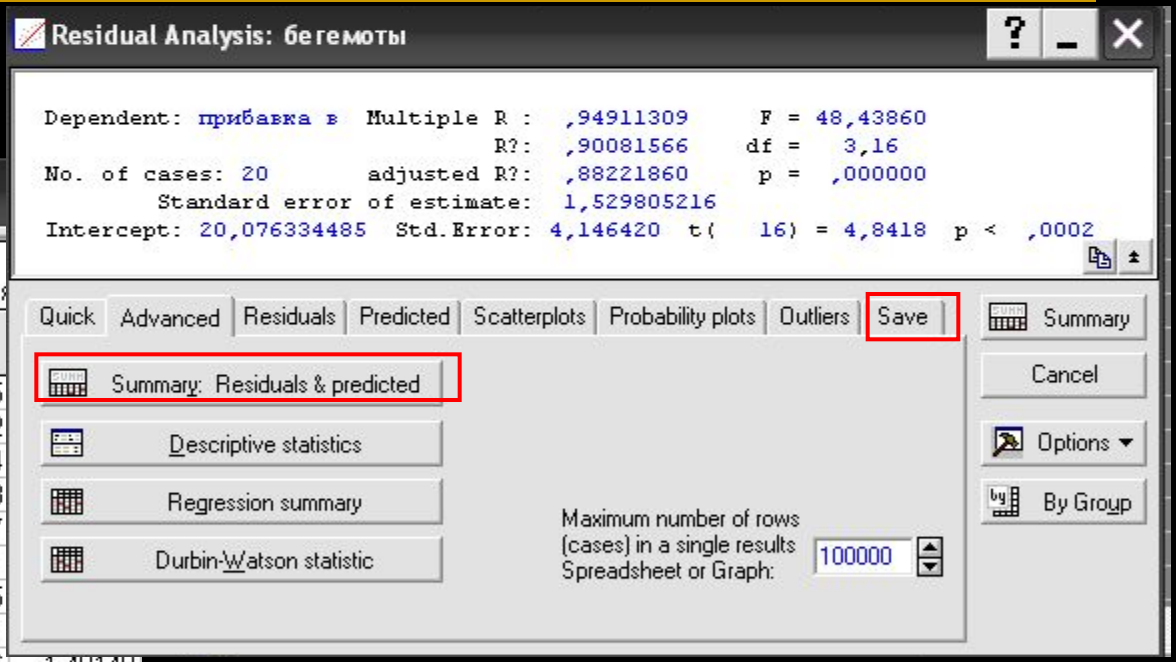
Alpha for highlighting effects: ,05

Quick | Advanced | Residuals/assumptions/prediction

Perform residual analysis
Descriptive statistics
Code generator

Predict values
Predict dependent variable
Compute confidence limits Alpha: ,05
Compute prediction limits

OK
Cancel
Options
By Group



& Residual Values (бегемоты)

Predicted & Residual Values (бегемоты)
Dependent variable: прибавка в весе за мес

Case No.	Observed Value	Predicted Value	Residual	Standard Pred. v.	
1	21,11111	20,08197	1,02914	0,23085	
2	13,63636	15,24849	-1,61213	-0,91162	
3	18,00000	17,82153	0,17847	-0,30344	
4	20,00000	21,08455	-1,08455	0,46783	
5	17,27273	17,20708	0,06565	-0,44867	
6	31,25000	29,86369	1,38631	2,54291	
7	15,83333	19,55820	-3,72486	0,10705	
8	20,00000	19,90662	0,09338	0,18941	
9	17,50000	19,76637	-2,26637	0,15626	-1,48148
10	19,00000	18,15647	0,84353	-0,22427	0,55140
11	16,15385	15,72465	0,42919	-0,79907	0,28055
12	15,71429	14,96591	0,74838	-0,97841	0,48920
13	18,00000	18,28773	-0,28773	-0,19324	-0,18808
14	15,33333	14,57290	0,76043	-1,07130	0,49708
15	21,00000	21,82652	-0,82652	0,64321	-0,54028
16	14,66667	12,15947	2,50720	-1,64176	1,63890
17	24,16667	23,20703	0,95963	0,96951	0,62729
18	28,18182	27,40240	0,77942	1,96115	0,50949
19	16,00000	16,70276	-0,70276	-0,56788	-0,45938
20	19,28572	18,56154	0,72418	-0,12852	0,47338
Minimum	13,63636	12,15947	-3,72486	-1,64176	-2,43486
Maximum	31,25000	29,86369	2,50720	2,54291	1,63890
Mean	19,10529	19,10529	0,00000	-0,00000	0,00000
Median	18,00000	18,42463	0,30383	-0,16088	0,19861

Predicted & Residual Values (бегемоты)

Residual Analysis: бегемоты

Dependent: прибавка в Multiple R : ,94911309 F = 48,43860
R?: ,90081566 df = 3,16
No. of cases: 20 adjusted R?: ,88221860 p = ,000000
Standard error of estimate: 1,529805216
Intercept: 20,076334485 Std.Error: 4,146420 t(16) = 4,8418 p < ,0002

Quick | Advanced | Residuals | Predicted | Scatterplots | Probability plots | Outliers | Save

Summary
Cancel
Options
By Group

Predicted vs. residuals
Observed vs. squared residuals
Predicted vs. sguared residuals
Residuals vs. deleted residuals
Predicted vs. observed
Bivariate correlation
Observed vs. residuals
Partial residual plot

й за день vs. прибавка в весе за месяц
ь = 5,6335 + ,52281 * прибавка в весе за месяц
rrelation: r = ,75110

масса пищи, съеденной за день

прибавка в весе за месяц

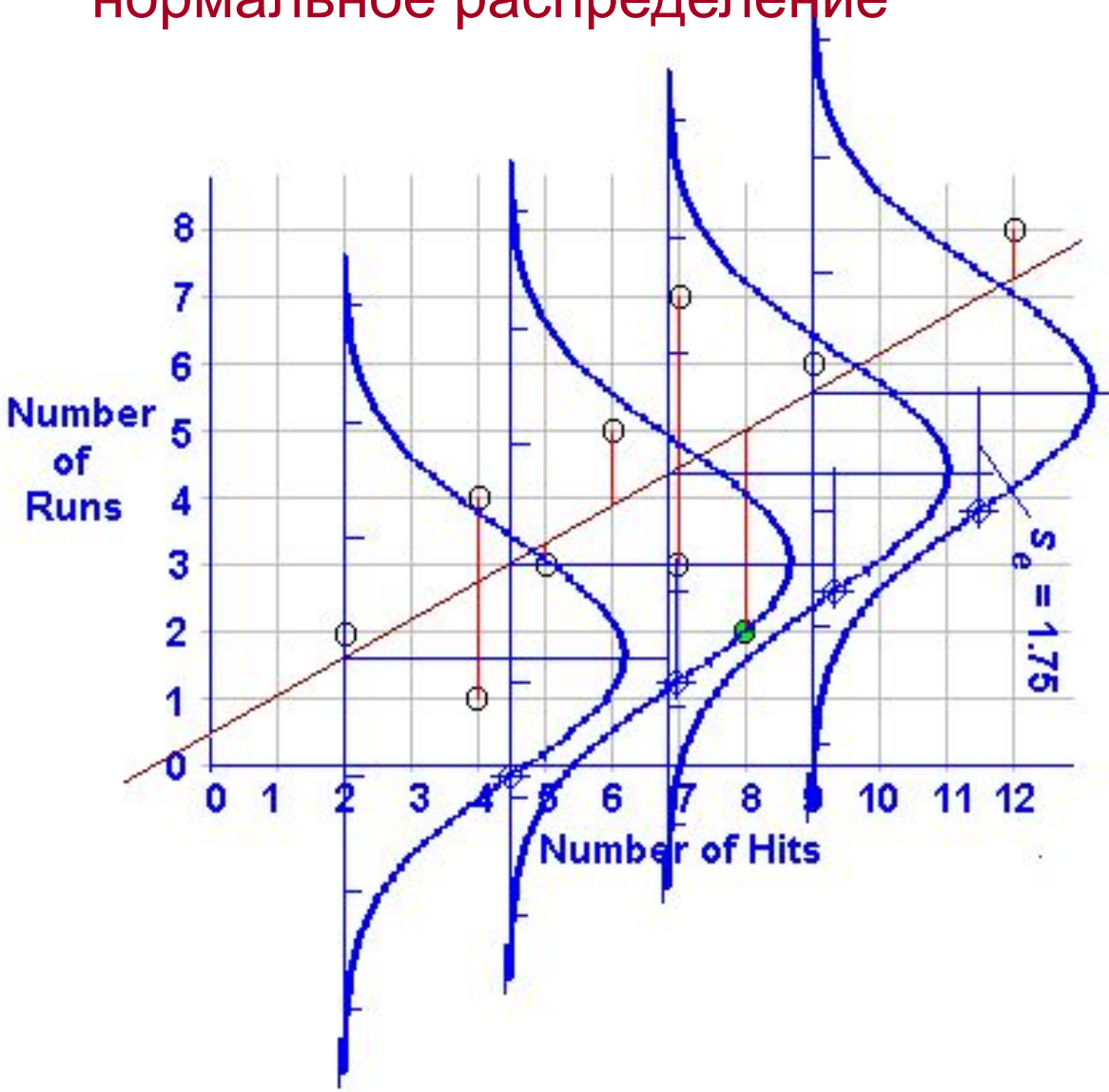
95% confidence

Регрессии
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 быть не меньше, чем в 10 раз превосходить число переменных в анализе (лучше – в 20 раз).
6. Следует исключить аутлаеры

Регрессии

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.

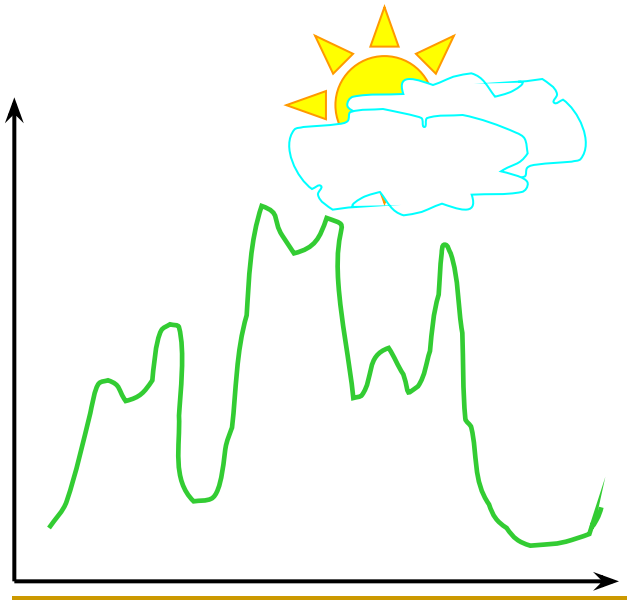
$$Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \cdots + \beta_m X_i^m + \epsilon_i$$

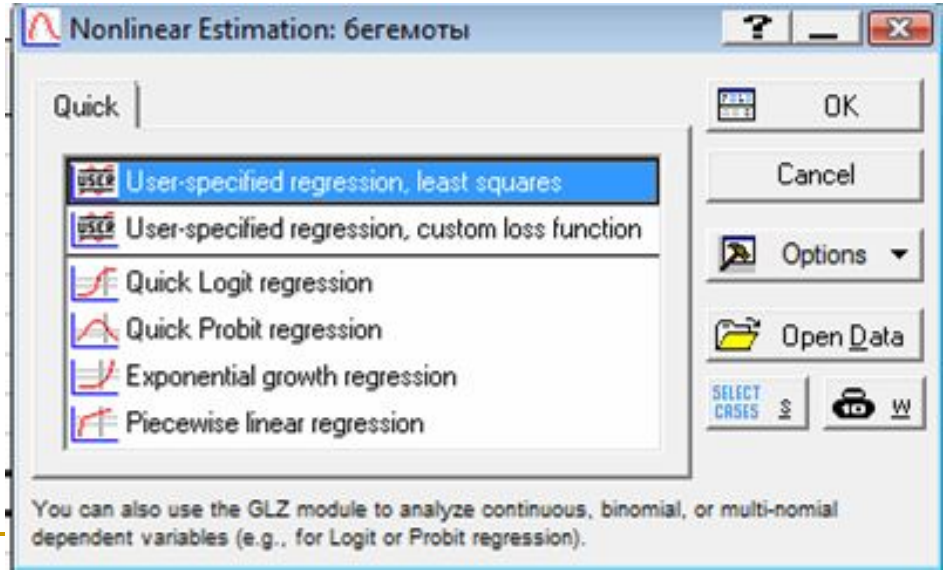
В статистке каждый X^m обозначают как новую переменную и дальше анализируют почти как линейную модель.

Регрессии

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



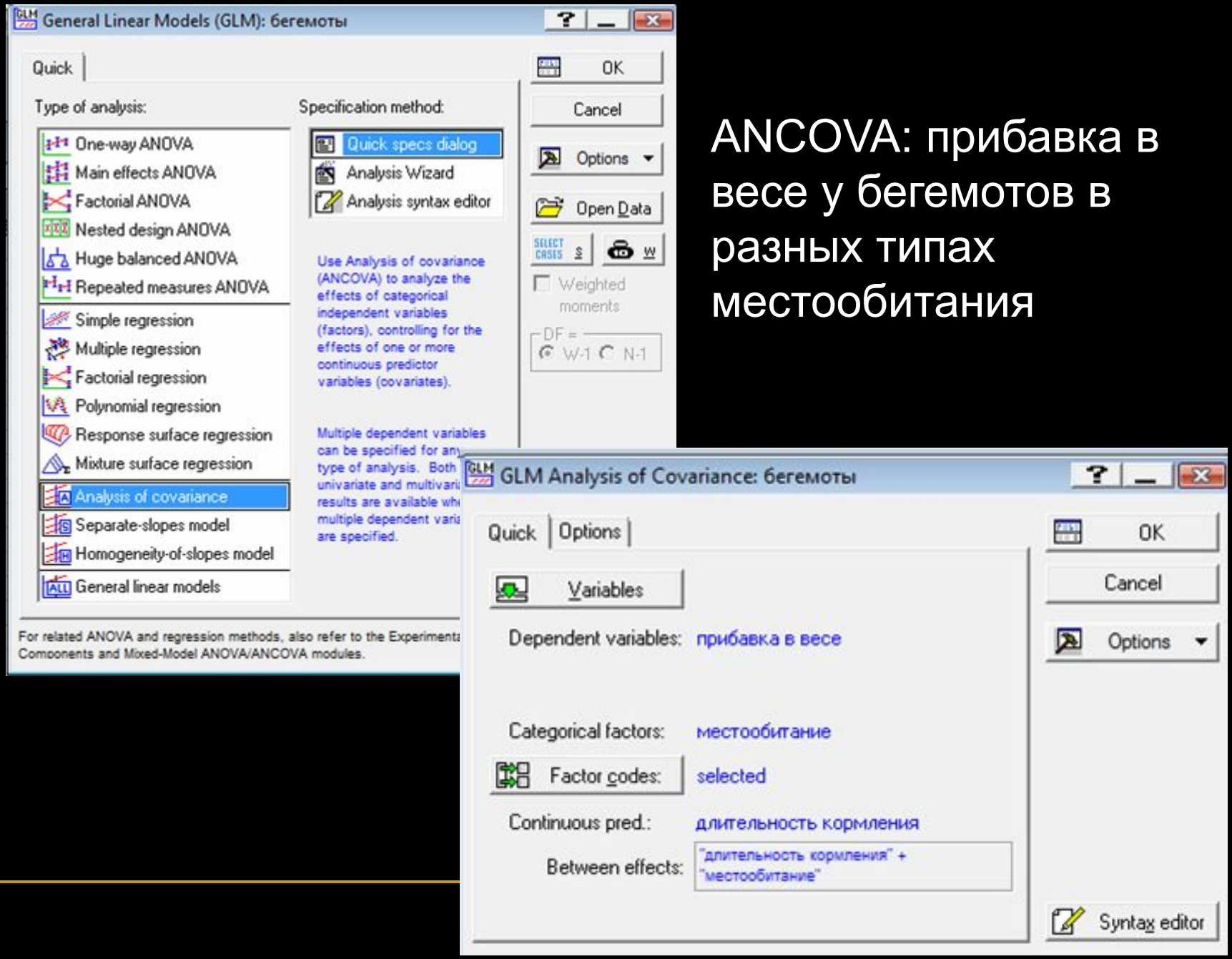


ANCOVA

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

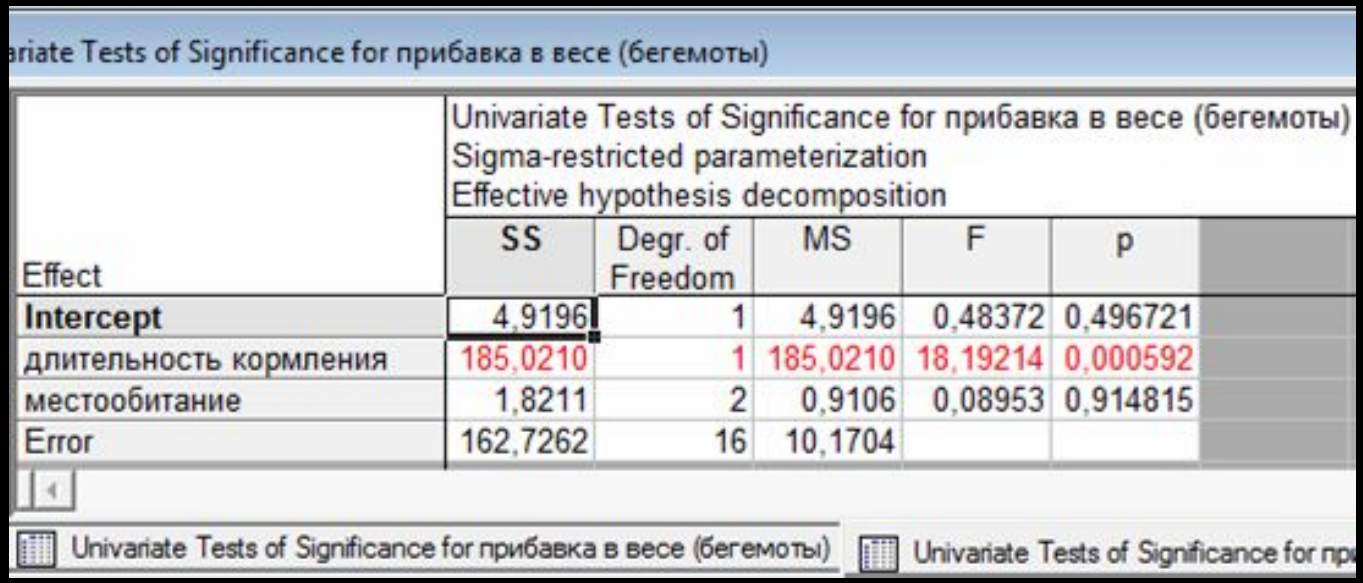
Пример: мы анализируем влияние типа пищи (группирующая независимая) и уровня кортикостероидов в крови (непрерывная независимая) на массу тигров (непрерывная зависимая).

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



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

Независимые переменные	Зависимые переменные	Модель
Одна группирующая	Одна непрерывная	One-way ANOVA
Много группирующих	Одна непрерывная	Factorial ANOVA (two-, multiway). Main effect ANOVA
Одна или много группирующих	Много непрерывных	MANOVA (multivariate ANOVA)
Одна непрерывная	Одна непрерывная	Simple regression
Много непрерывных	Одна непрерывная	Multiple regression
Одна группирующая (или много) + одна непрерывная (или много)	Одна непрерывная	ANCOVA

«Много» = 2 и больше

1. исследователь решил узнать, как зависит размер дома у семьи от дохода семьи (в год). Собрал данные от 50 семей. H_0? Статистический критерий? Как изменится результат теста, если доходы семей увеличатся каждая на 5000$ в год?

2. педиатры изучают прибавку в весе у младенцев (её оценивают как разницу в массе ребёнка в 2 мес и при рождении). При этом, в их выборке есть дети, которые вскармливаются искусственно, а есть те, которые находятся на грудном вскармливании. Кроме того, некоторые матери кормят младенцев по требованию, другие же – строго по расписанию. Влияют ли тип пищи и распорядок вскармливания на прибавку в весе? H_0? Статистический критерий?

3. владелец бассейна думает, что количество хлора, которое ежедневно затрачивается на то, чтобы содержать бассейн в чистоте, зависит от температуры воздуха и дня недели. Он стал отмечать, сколько каждый раз у него уходит хлора на очистку, и взял из газет данные о дневных температурах. Так он делал в течение полугода. Зависит ли количество хлора от температуры и дня недели? H_0? Статистический критерий?

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