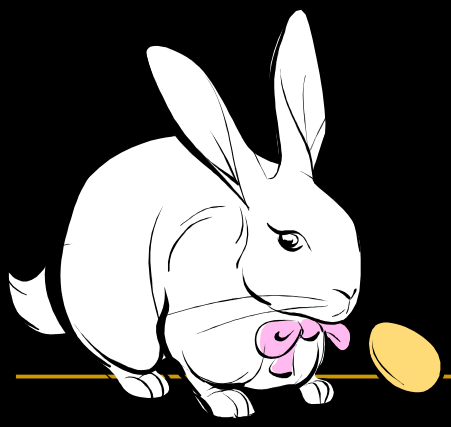
Корреляции.
Регрессионный анализ



Занятие 5

КОРРЕЛЯЦИИ

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

Мы изучали, отличается ли распределение этой переменной в одних условиях от распределения той же переменной в других условиях.

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

Это могут быть измерения одной особи или связанных пар.

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

Переменные – 1. масса; 2. длина хвоста.

КОРРЕЛЯЦИЯ (correlation)

Вопрос: в какой степени две переменные СОВМЕСТНО ИЗМЕНЯЮТСЯ? (т.е., повлечёт ли за собой увеличение одной переменной увеличение или уменьшение другой, или не повлечёт)

Коэффициент корреляции характеризует силу связи между переменными.

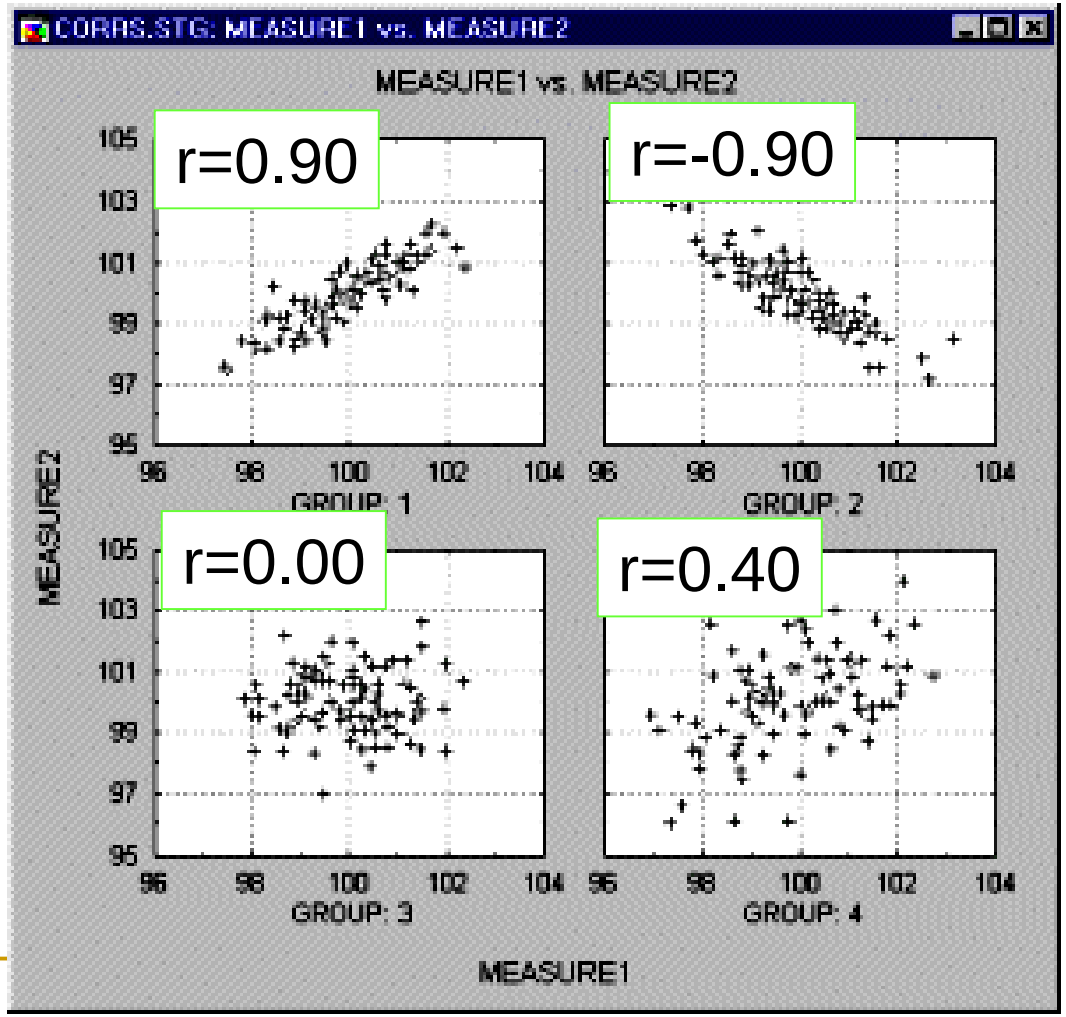
ЭТО ПРОСТО ПАРАМЕТР ОПИСАТЕЛЬНОЙ СТАТИСТИКИ

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 *тесноту* связи



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***.

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

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*

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

число строк (сусликов)

$$z_X = \frac{X - \bar{X}}{s_X}$$

$$z_Y = \frac{Y - \bar{Y}}{s_Y}$$

стандартное отклонение для веса

стандартное отклонение для хвоста

для каждого X и Y (для каждого суслика)

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

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

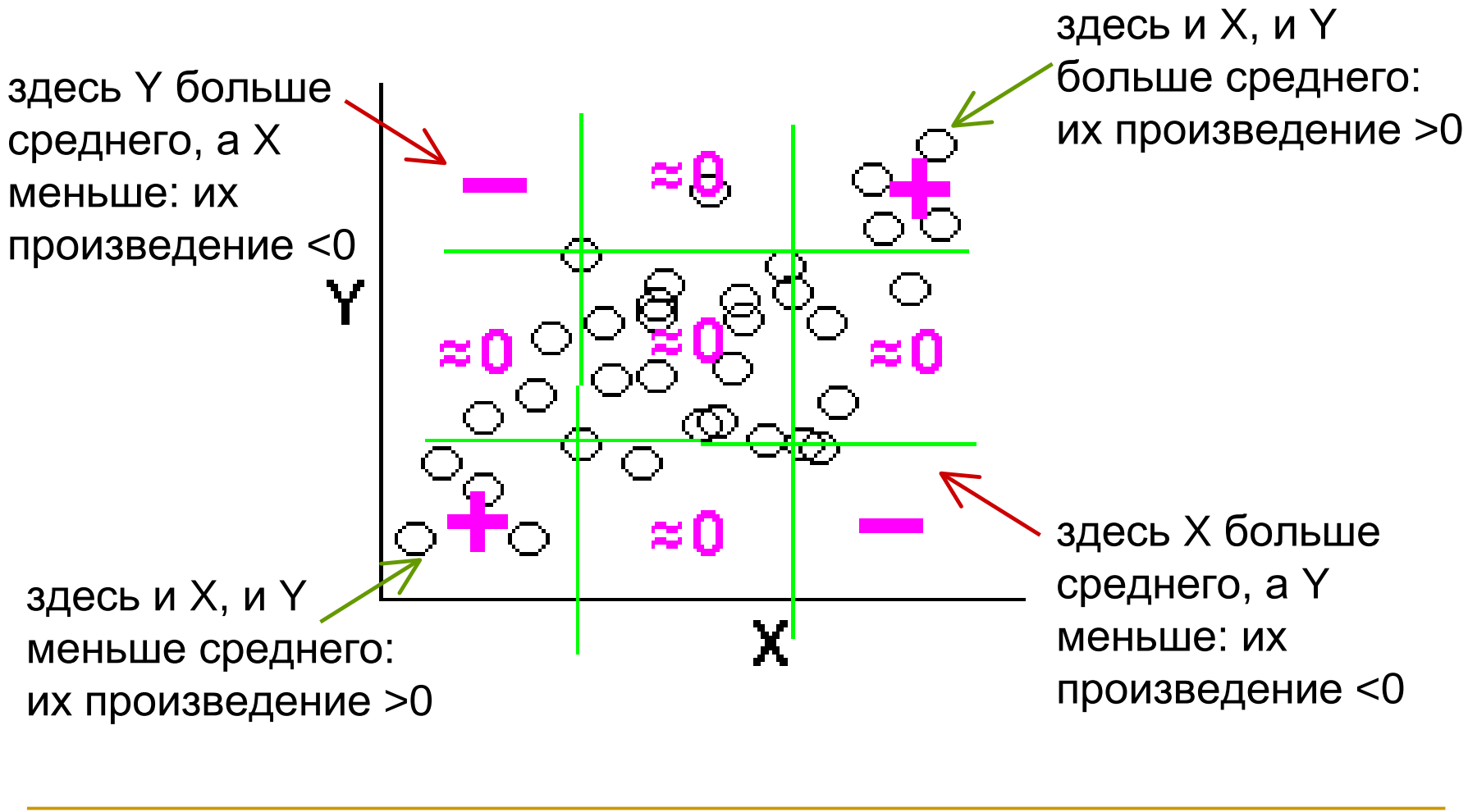
параметр ВЫБОРКИ

параметр ПОПУЛЯЦИИ

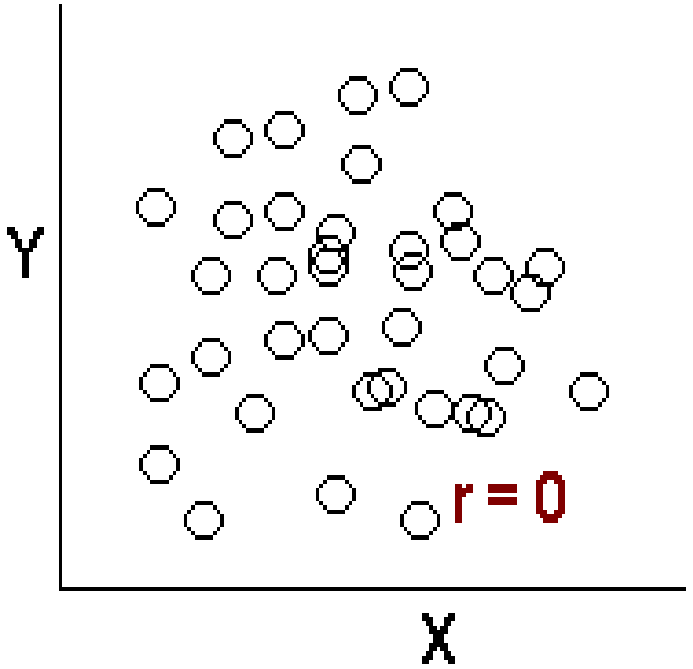
Всё как для других параметров описательной статистики: среднего, дисперсии, и т.д.!

Как определяется знак коэффициента корреляции?

Знаком $\sum z_X z_Y$:



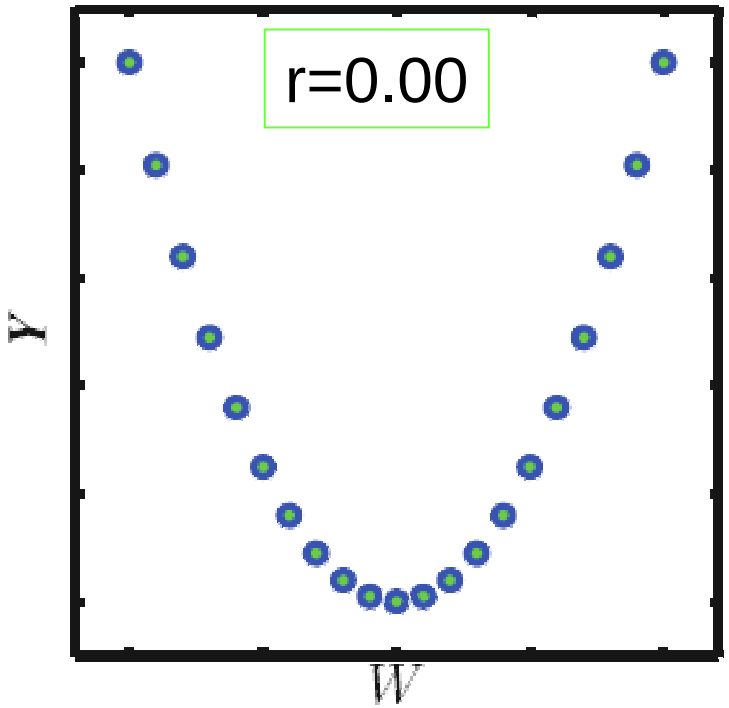
Создаётся впечатление, что близкий к нулю коэффициент корреляции говорит о том, что связи между переменными нет или почти нет.



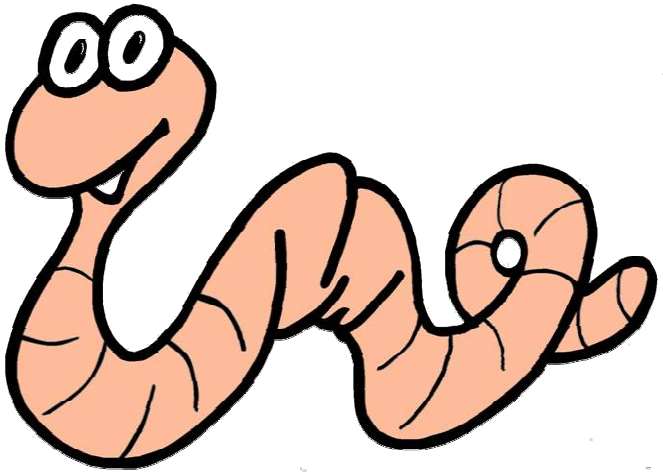
Здесь и впрямь её нет

НО это не всегда так, есть исключения.

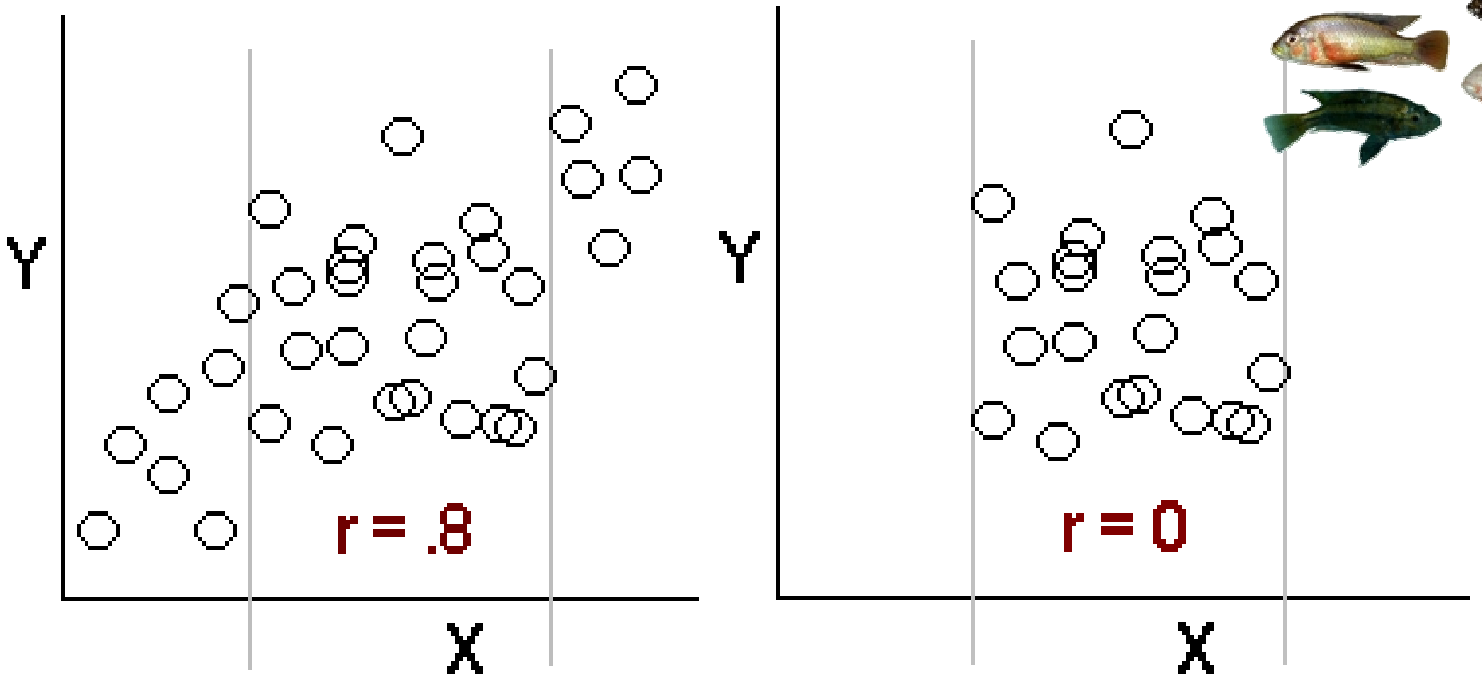
1. Коэффициент корреляции Пирсона оценивает только линейную связь переменных!
И он не покажет нам нелинейную связь



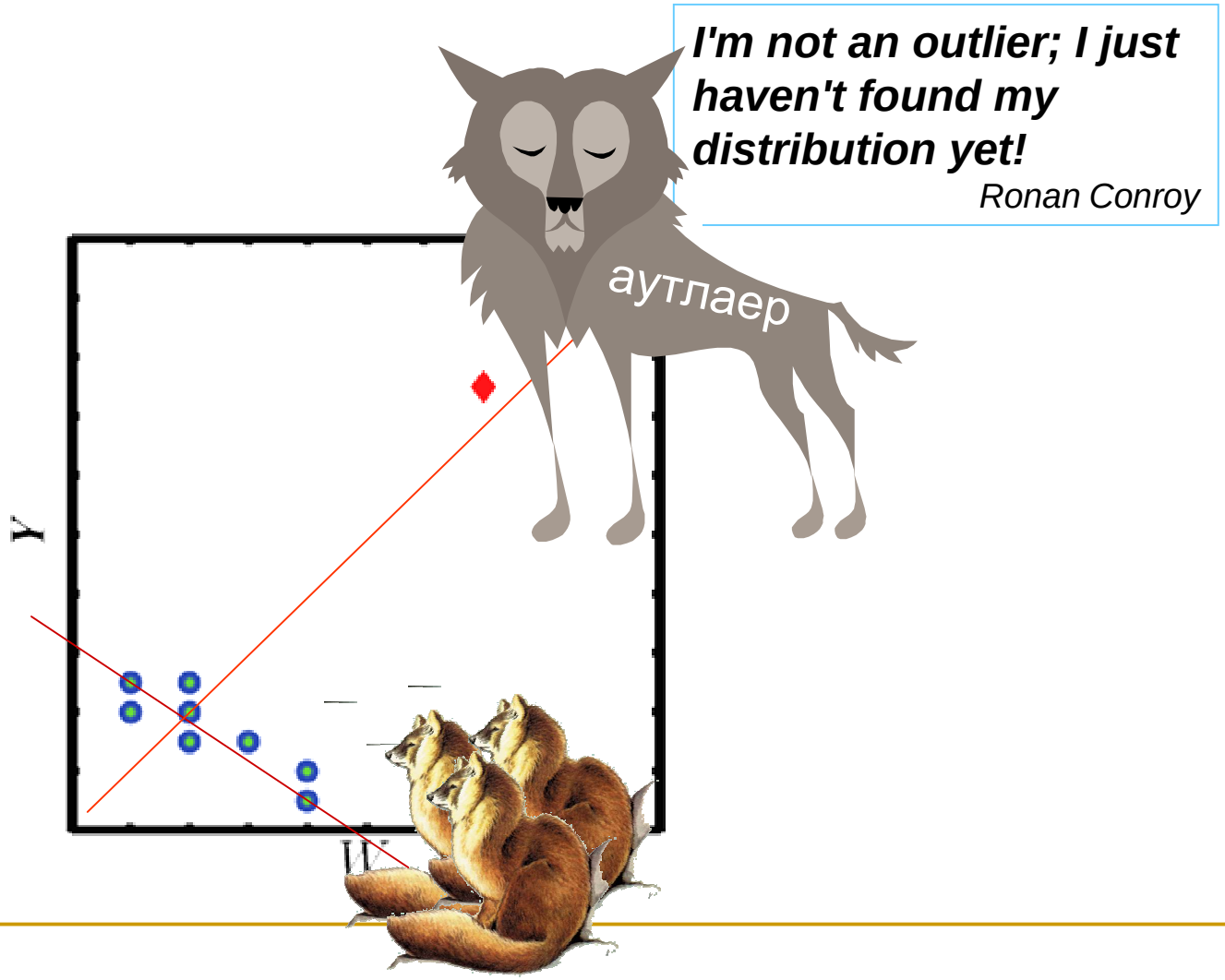
Здесь связь переменных есть, и она очень сильная, но r=0.00



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



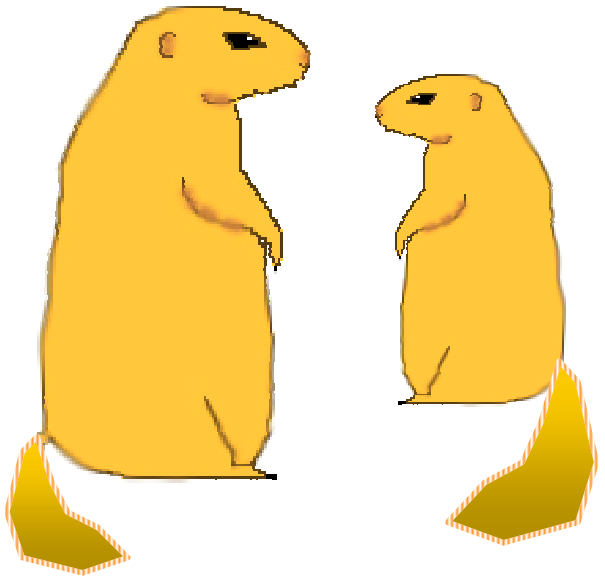
3. Коэффициент корреляции Пирсона очень чувствителен к аутлаерам.



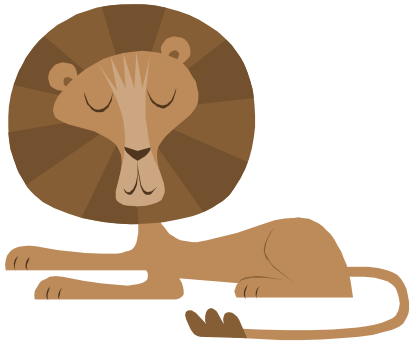
Важное замечание:

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

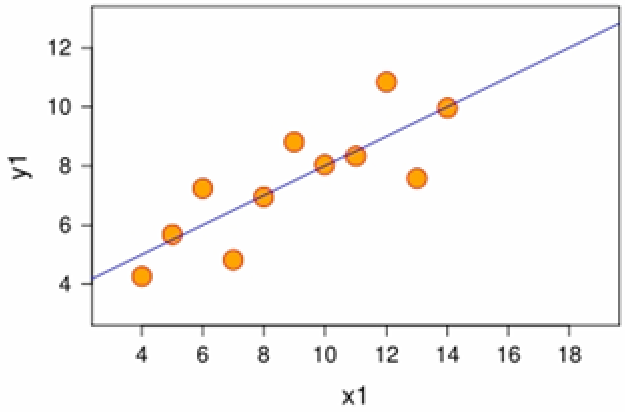
Она **ВООБЩЕ НИЧЕГО** о ней **НЕ ГОВОРИТ** (даже очень большой r)

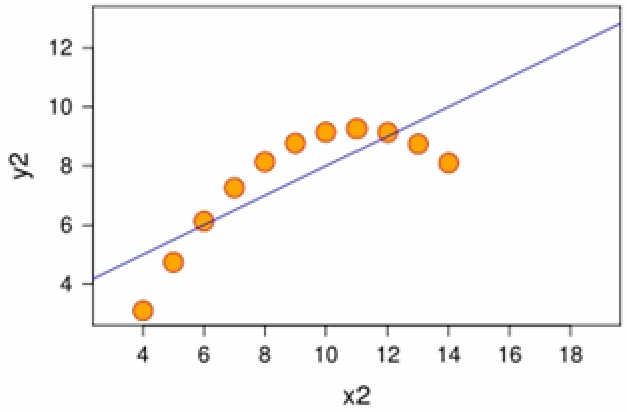


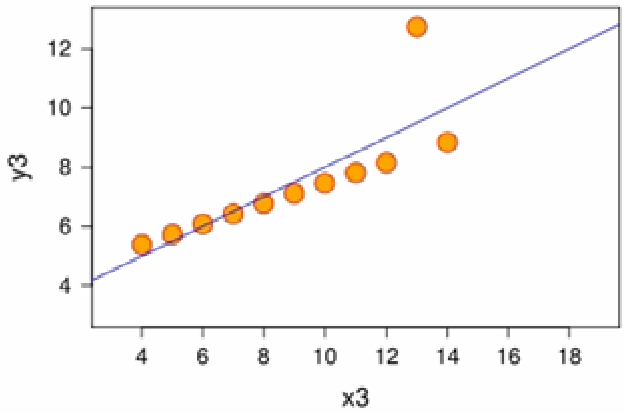
Коэффициент корреляции Пирсона – параметр выборки.
Можем ли мы на основе него судить о популяции?
Просто глядя на коэффициент – НЕТ.

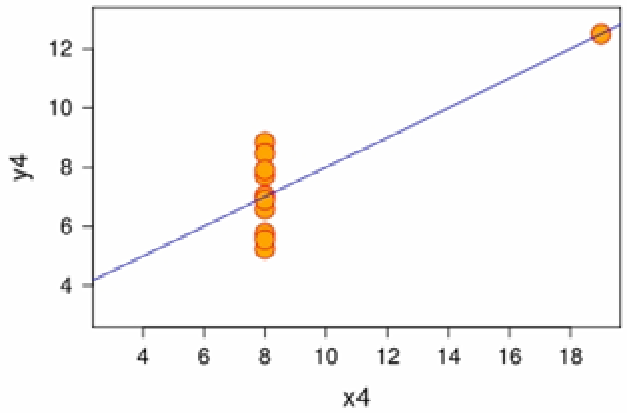


Correlation
between each x and
y = **0.816**









Мы хотим оценить коэффициент корреляции в популяции.

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

(альтернативная гипотеза может быть односторонней)

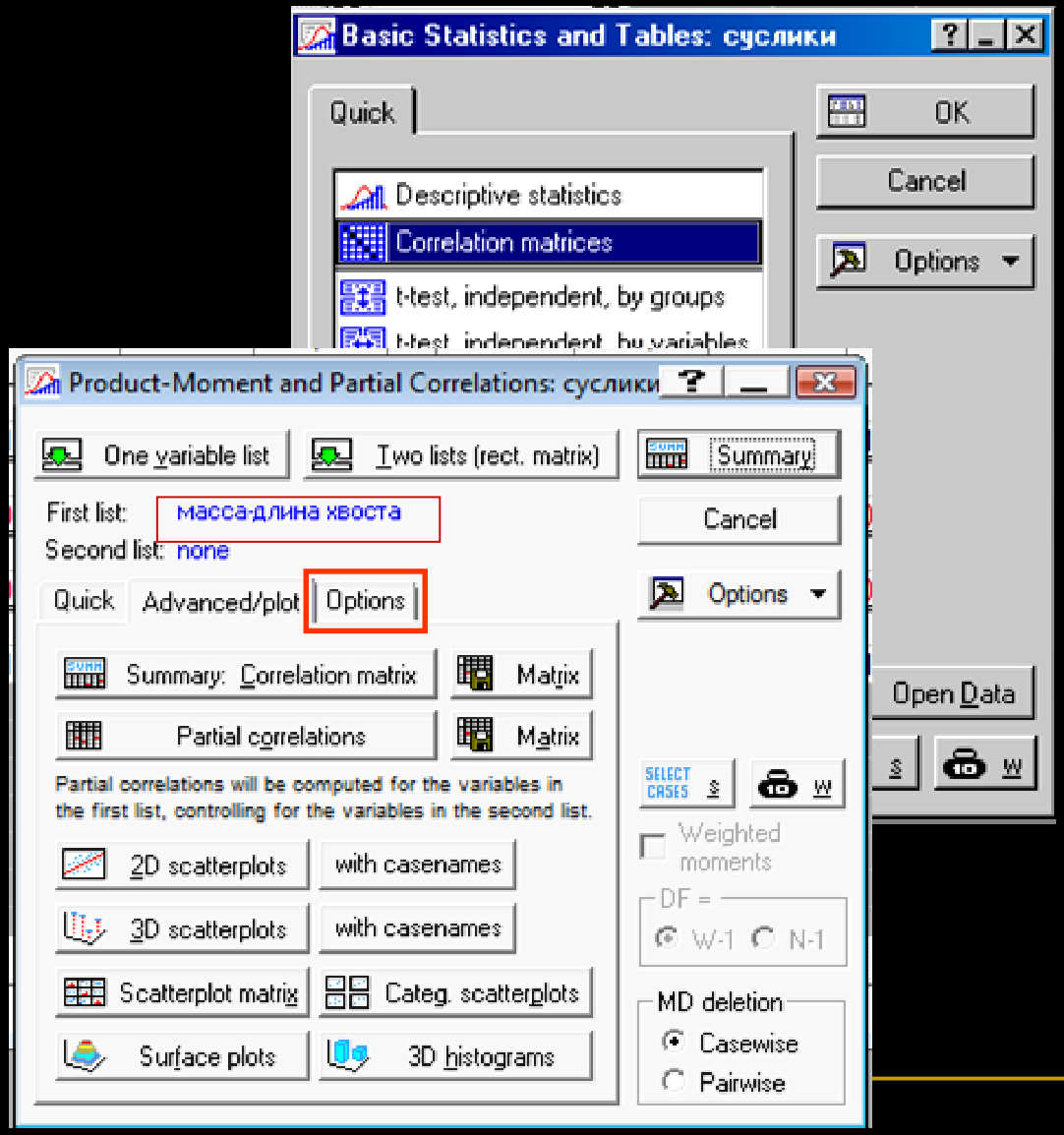
Связаны ли у сусликов масса тела и длина хвоста?

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

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

стандартная ошибка
коэффициента корреляции

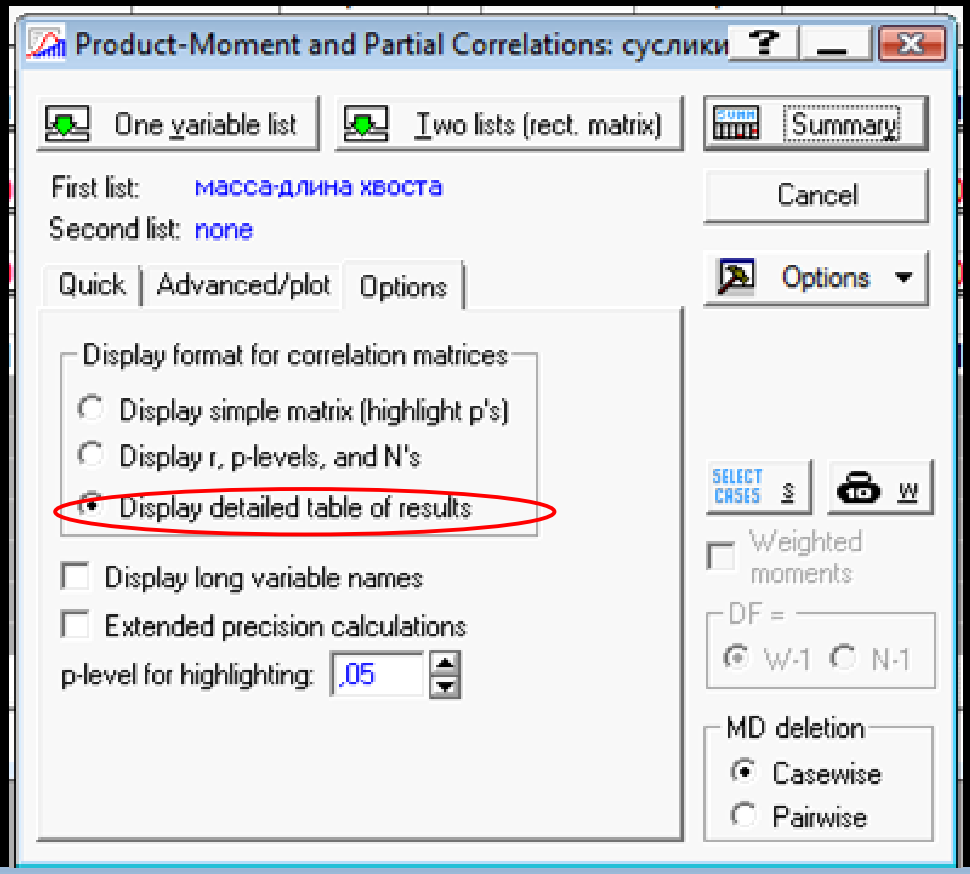
Pearson product-moment correlation coefficient *r*



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)

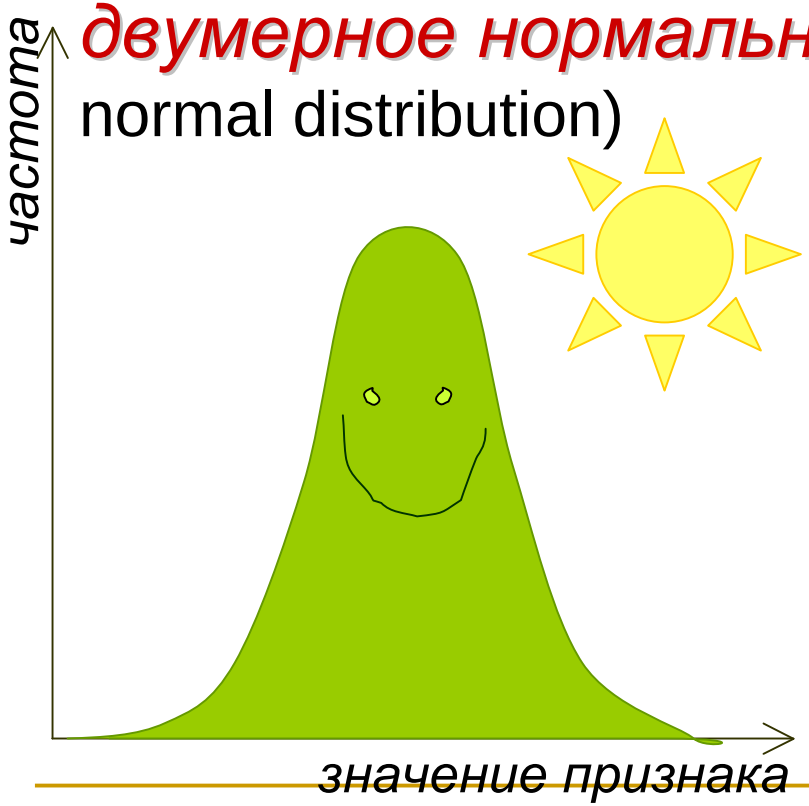
Correlations (суслики)
Marked correlations are significant at p < ,05000
(Casewise deletion of missing data)

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

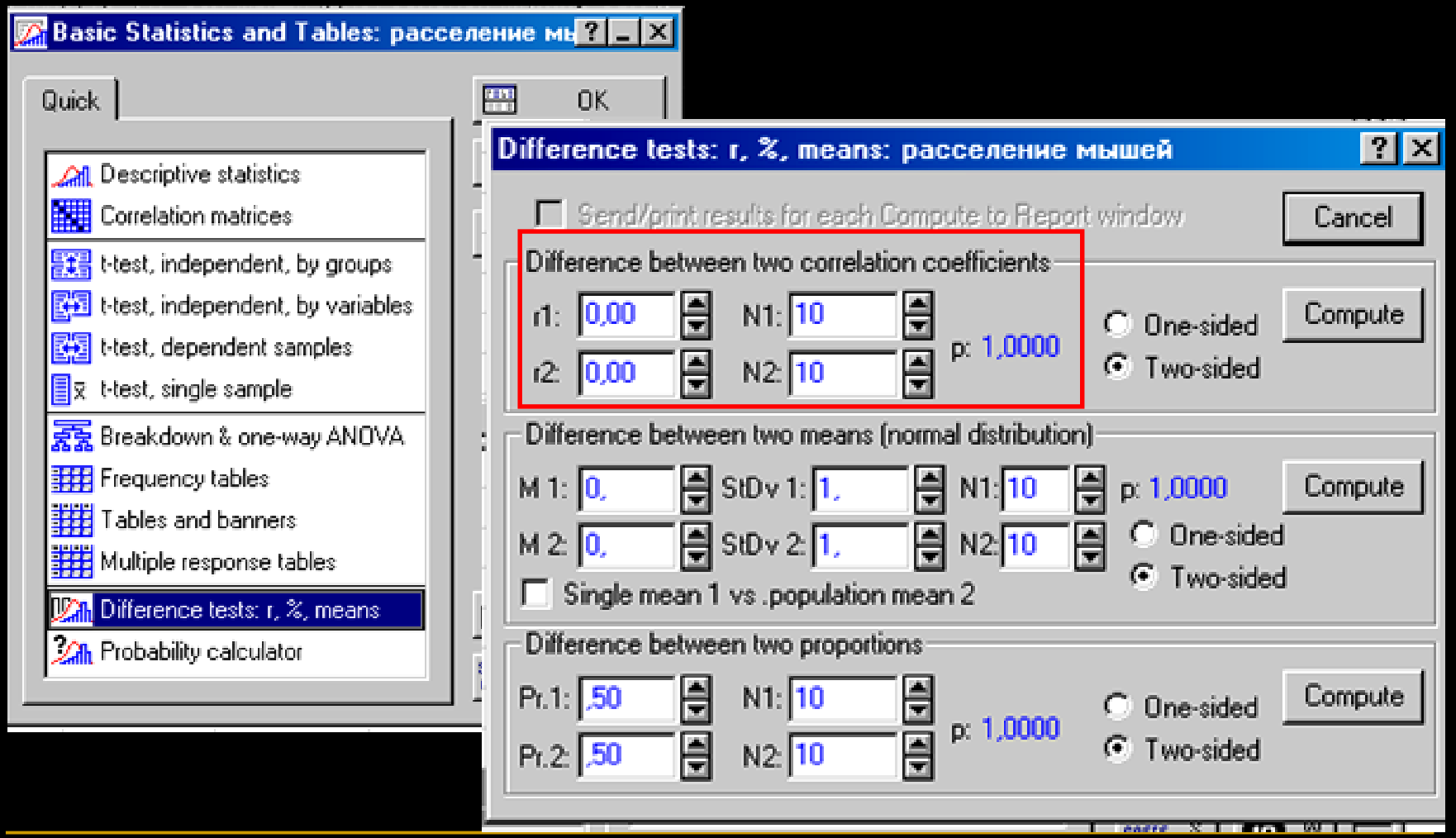
Требование к выборке для тестирования гипотезы о коэффициенте корреляции Пирсона:

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)



Можно сравнить два коэффициента корреляции от двух выборок



Для двумерного нормального распределения

Непараметрические тесты для ассоциаций (РАНГОВЫЕ)

1. Коэффициент корреляции Спирмана (*Spearman rank order correlation*)

Связана ли дистанция расселения с индексом упитанности у мыши?

Переменные – 1. дистанция расселения;
2. индекс упитанности (ранговый)

Для нашей задачи не годится коэффициент корреляции Пирсона: одна из переменных ранговая!

Нам подходит коэффициент корреляции Спирмана

1. Ранжируем данные для каждой переменной от меньшего к большему;
2. Если встретились одинаковые значения (***tied ranks***), присваиваем им средние ранги (как для Манн-Уитни теста);
3. Считаем разности рангов в каждой строчке (паре);
4. Считаем коэффициент r_s

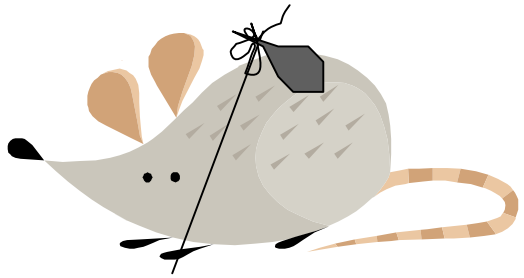
$$r_s = 1 - \frac{6\sum D_i^2}{n(n^2-1)}$$

разности рангов

число строк (размер выборки)

$H_0 : \rho_s = 0$

$H_1 : \rho_s \neq 0$

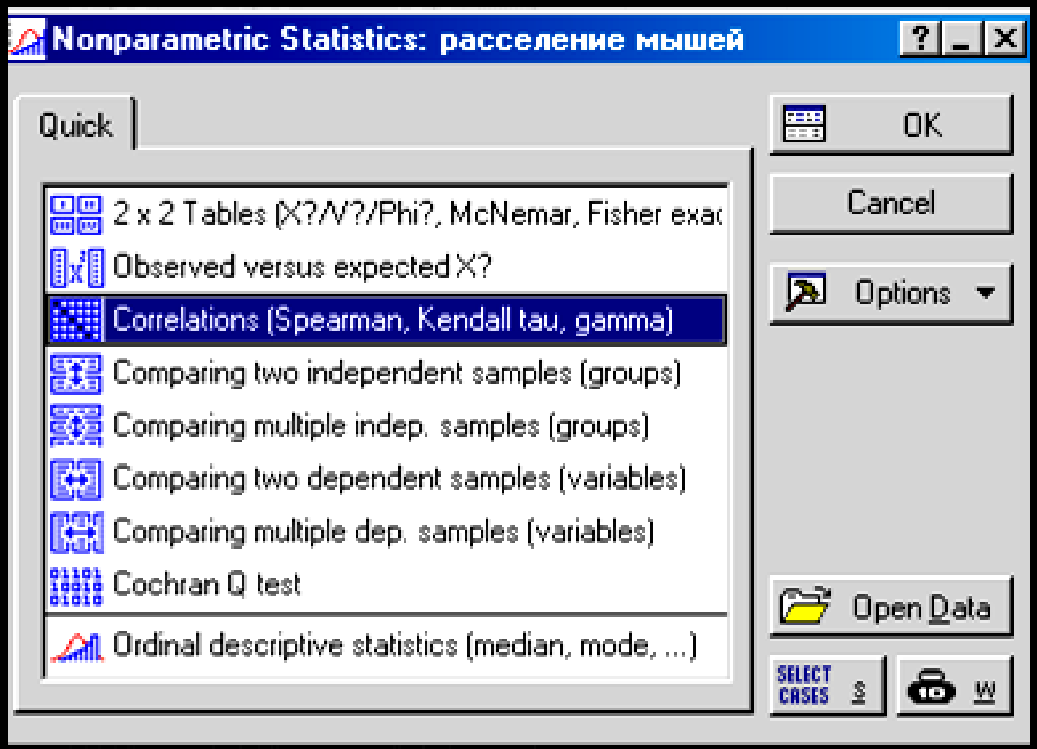


Статистика критерия – сам коэффициент корреляции Спирмана

Коэффициент Спирмана – аналог коэффициента корреляции Пирсона, стремится к нему в больших выборках.

Подходит для *2-х и более переменных*, лучший для дробных количественных признаков. Размер выборки ≥ 10.

Spearman Rank Order Correlations



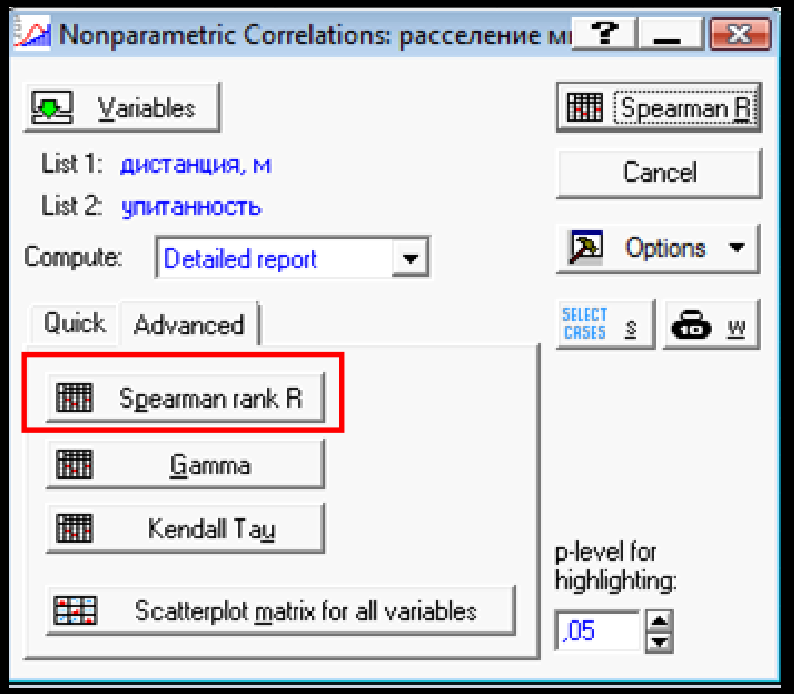
Data: расселение мышей.sta (10v by 20c)

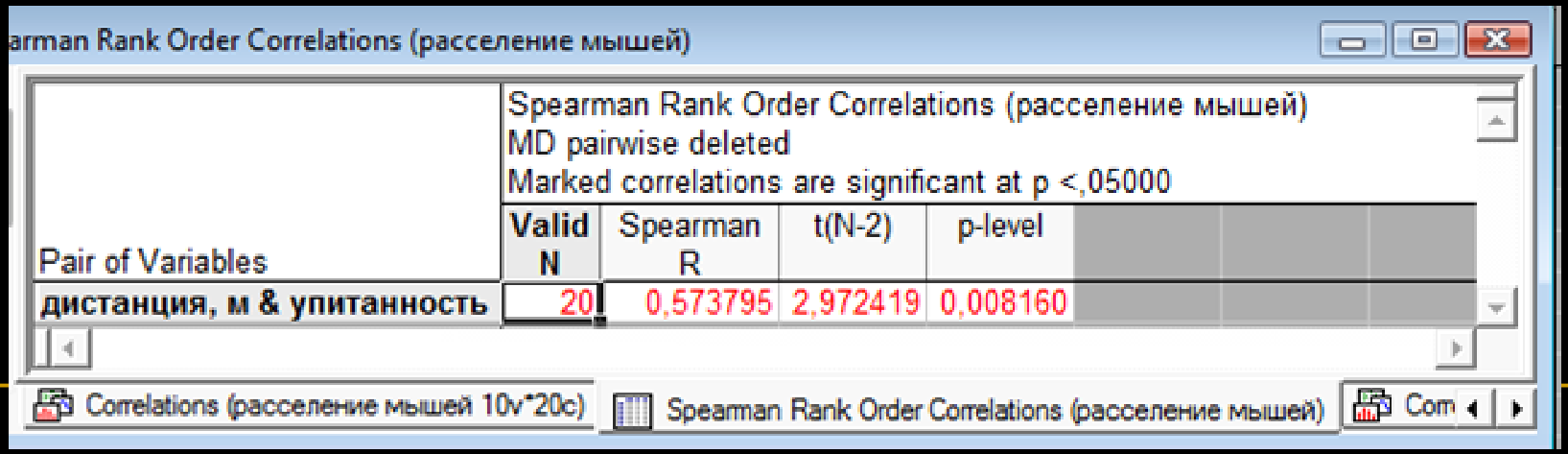
	1 № мыши	2 дистанция, м	3 упитанность
1	1	96,00	15
2	2	43,00	8
3	3	67,05	11
4	4	40,00	12
5	5	55,93	10
6	6	102,00	19
7	7	42,00	5
8	8	38,00	12
9	9	35,00	11
10	10	41,00	10
11	11	40,00	10
12	12	15,00	9
13	13	14,00	8
14	14	78,00	6
15	15	58,00	7
16	16	49,00	11
17	17	52,88	15
18	18	120,00	17

Spearman Rank Order Correlations

Отвергаем H_0:

Оказалось, что дистанция расселения положительно связана с упитанностью у мыши.





2. Коэффициент корреляции Кендалла (*Kendall's coefficient of rank correlation, Kendall-τ*)

Он оценивает разность между вероятностью того, что порядок данных в обеих переменных одинаков, и вероятностью того, что порядки разные.

Считается совсем не так, как коэффициент Спирмана.

Связана ли дистанция расселения с упитанностью у экзотических зелёных мышей?

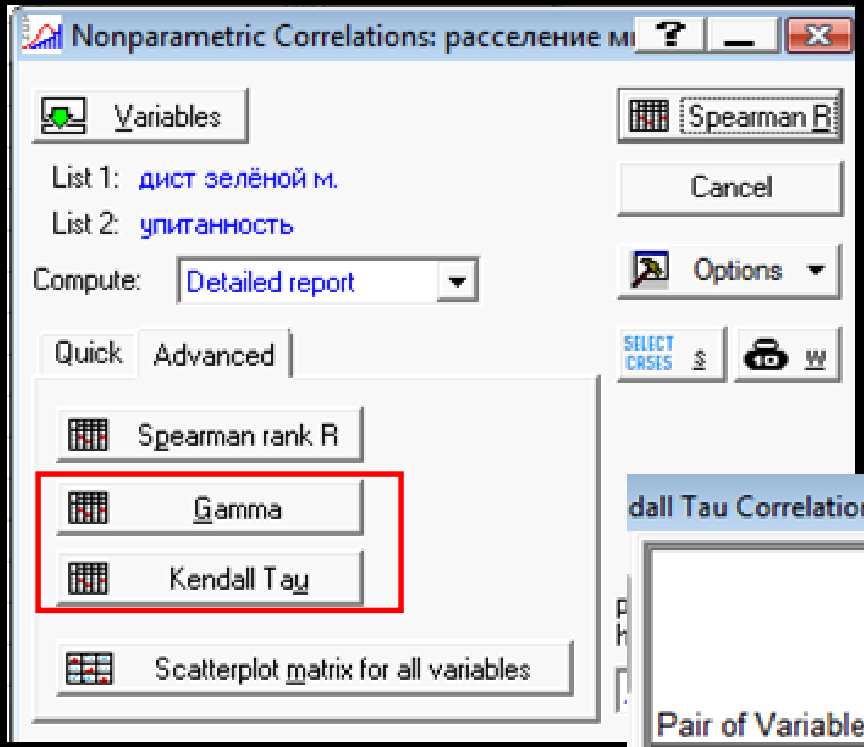
Только для 2-х переменных; подходит для маленьких выборок (?).

3. Гамма-статистика (*Gamma*)

Почти как коэффициент корреляции Кендалла, её лучше использовать, если в выборке много совпадающих значений (она их учтёт).

Kendall's coefficient of rank correlation, Kendall-τ

Отвергаем H_0: дистанция расселения у зелёных мышей отрицательно связана с упитанностью.



Data: расселение мышей.sta (10v by 20c)

	5 № зел мыши	6 дист зелёной м.	7 упитанность
1	1	122	23
2	2	245	9
3	3	101	17
4	4	287	7
5	5	98	31
6	6	50	36
7	7	264	10
8	8	175	13
9			

dall Tau Correlations (расселение мышей)

Kendall Tau Correlations (расселение мышей)
MD pairwise deleted
Marked correlations are significant at p <,05000

Pair of Variables	Valid N	Kendall Tau	Z	p-level	p-exact 1-tailed
дист зелёной м. & упитанность	8	-0,857143	-2,96923	0,002985	,001

3. Коэффициент конкордантности Кендалла (*Kendall's coefficient of concordance*)

Лучше всего подходит для сравнения ранговых признаков например, при сравнении мнений разных экспертов (6 детей и 3 типа пирожных). Переменных может быть 3 и более

Считается он на основе коэффициентов корреляции Спирмана.

Петя

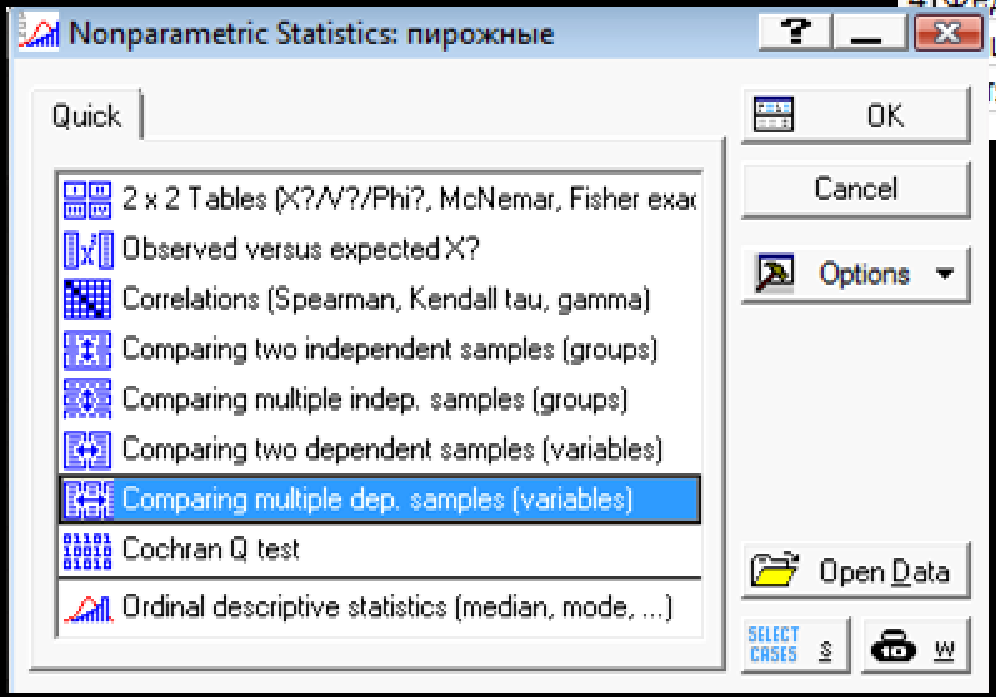
Гриша

Гурвинек

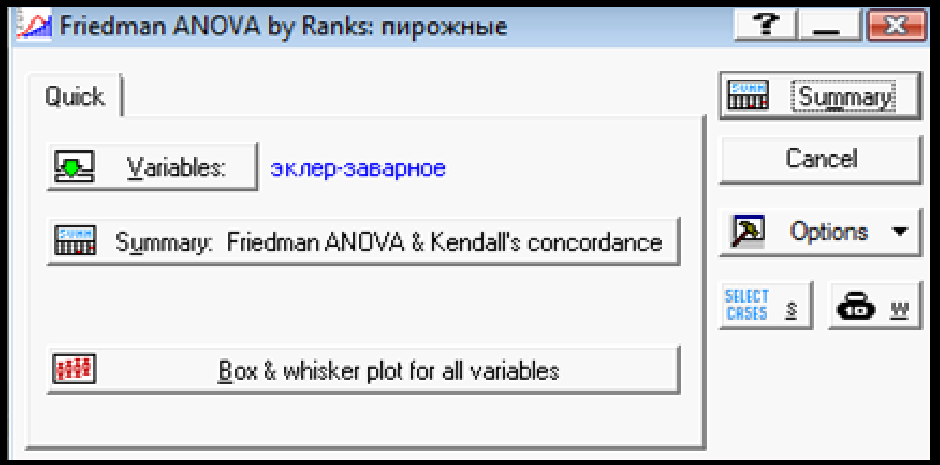
Коэффициент конкордантности Кендалла

Data: пирожные.sta (11v by 10c)

	1 тип пирожного	2 эклер	3 картошка	4 заварное
1	Петя	2	6	9
2	Гриша	1	7	6
3	Гурвинек	3	5	8
4	Федя	2	7	8
	ша	2	4	7
	тя	3	7	9



Чем ближе коэффициент к 1, тем выше согласие экспертов по сравнению со случайным.
Чем ближе к нулю, тем меньше согласие экспертов.



edman ANOVA and Kendall Coeff. of Concordance (пирожные)

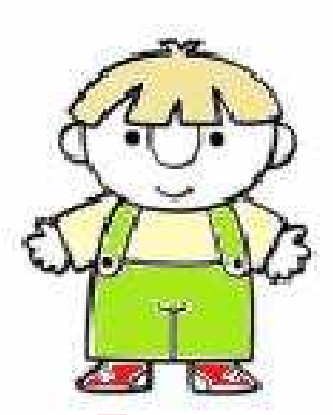
Friedman ANOVA and Kendall Coeff. of Concordance (пирожные
ANOVA Chi Sqr. (N = 6, df = 2) = 10,33333 p < ,00570
Coeff. of Concordance = ,86111 Aver. rank r = ,83333

Variable	Average Rank	Sum of Ranks	Mean	Std.Dev.
эклер	1,000000	6,00000	2,166667	0,752773
картошка	2,166667	13,00000	6,000000	1,264911
заварное	2,833333	17,00000	7,833333	1,169045

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.

Итак,

Задача регрессионного анализа – предсказать значение одной переменной на основании другой.
Для этого в линейной регрессии строится прямая – линия регрессии.

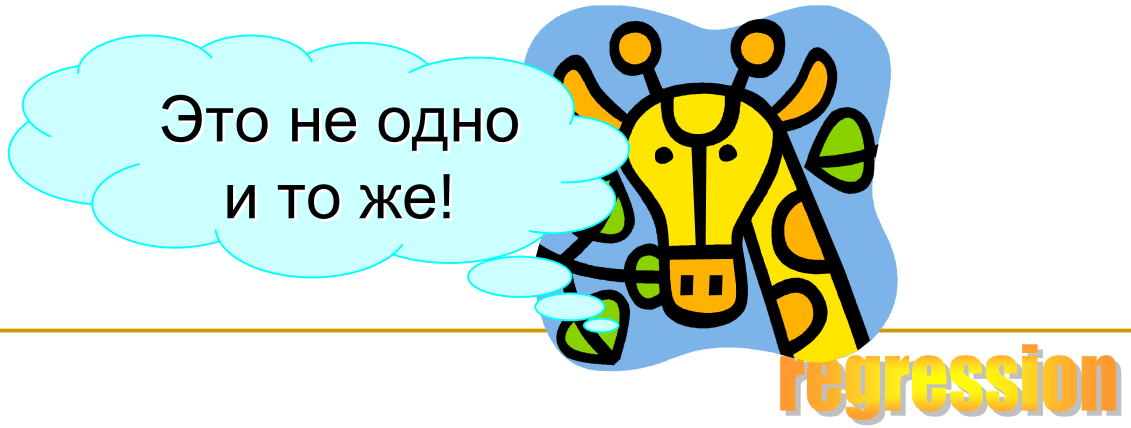
По оси Y располагают переменную, которую мы хотим предсказать, а по оси X – переменную, на основе которой будем предсказывать.
Предсказанное значение Y обычно обозначают как $\hat{Y}$

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

Пример: скорость набора веса у бегемота растёт с увеличением продолжительности кормления; долго кормившийся бегемот быстрее набирает вес

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

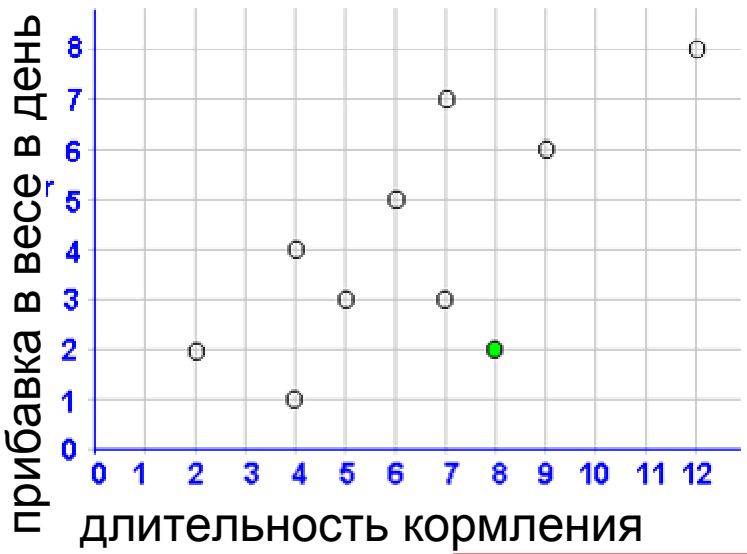
Пример: длина хвоста у суслика коррелирует положительно с его массой тела

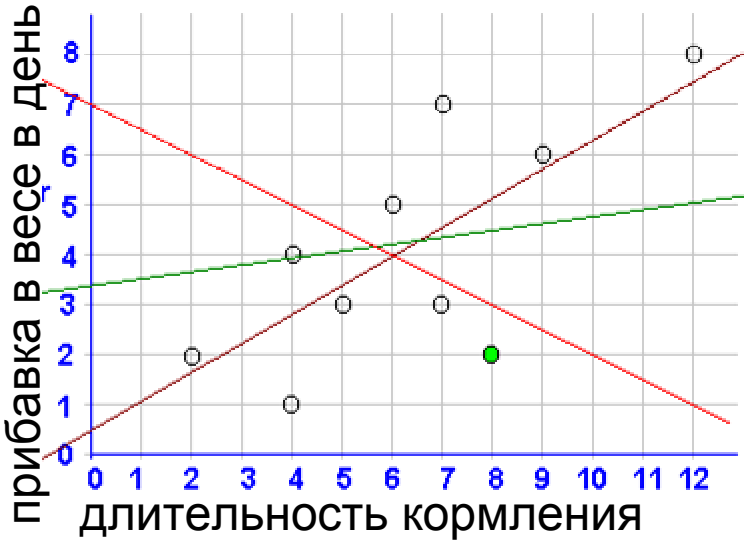


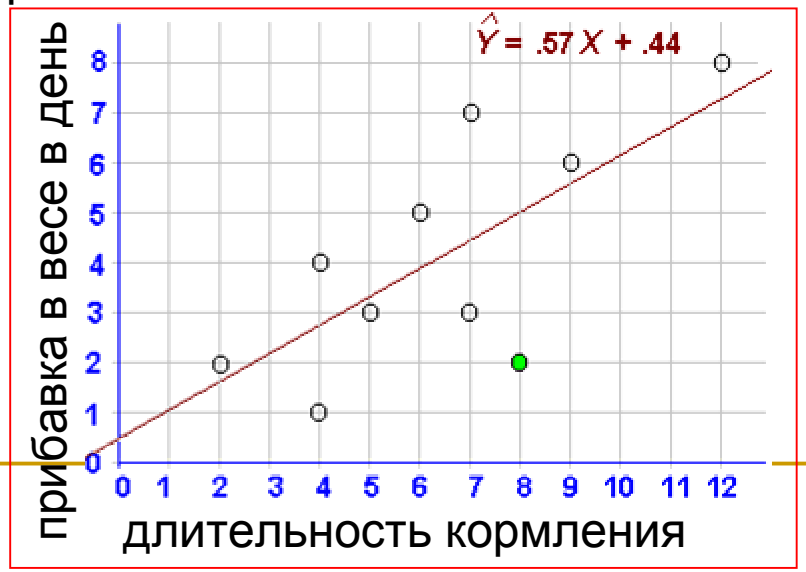
Мы изучаем поведение бегемотов в Африке. Мы хотим узнать, как связана длительность кормления со скоростью набора веса у этих зверей?

У нас **две** переменные – 1. длительность кормления в день (independent); 2. скорость набора веса в день (dependent)

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



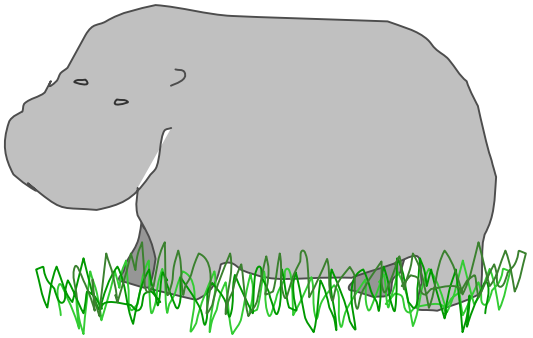




Простая линейная регрессия (*linear regression*)

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

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



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

Задача сводится к поиску коэффициентов a и b.

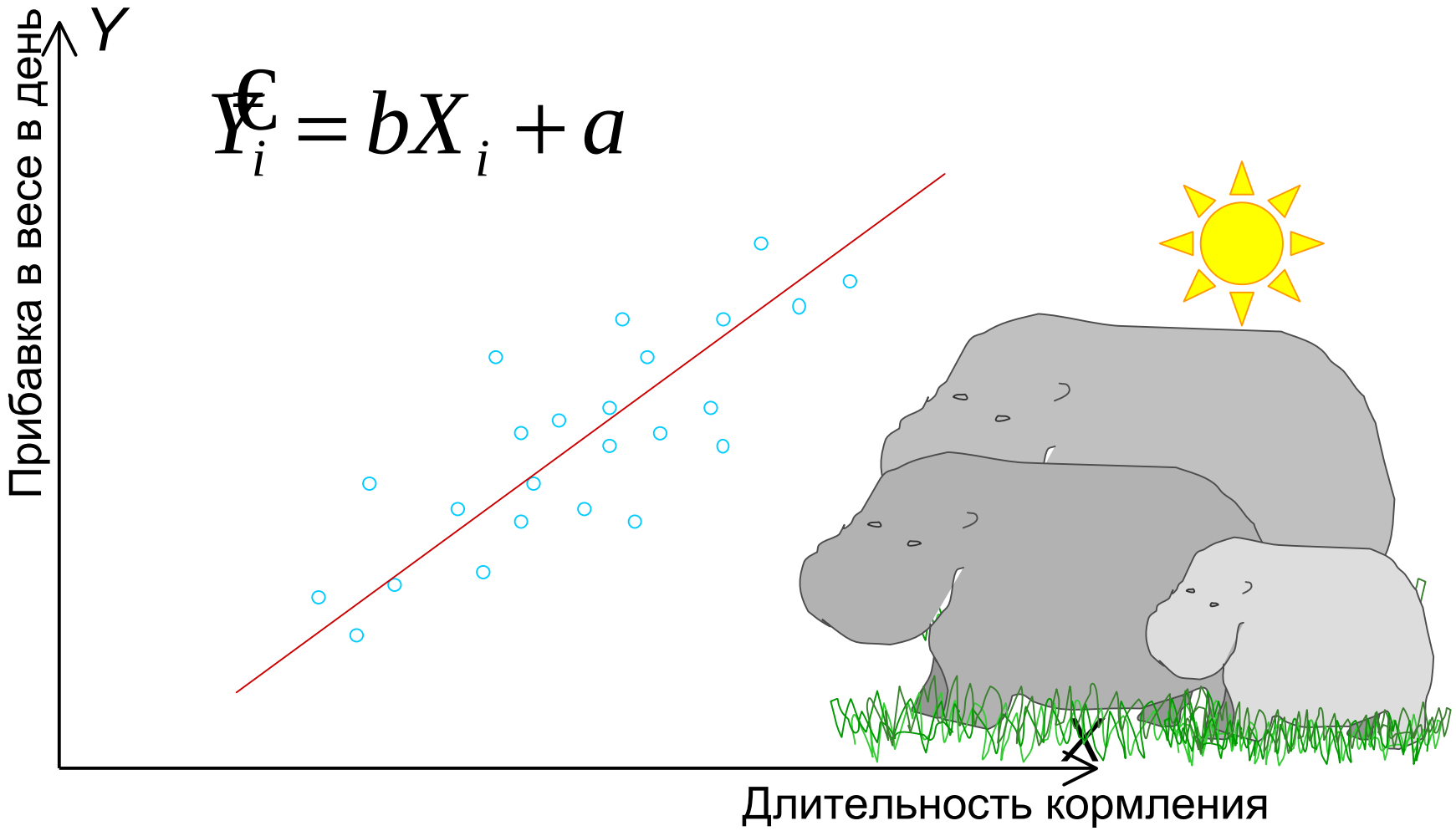
$$b = r\frac{s_X}{s_Y}$$

стандартные отклонения для X и Y

коэффициент корреляции Пирсона!

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

Следствие: линия регрессии всегда проходит через точку $(\bar{X}, \bar{Y})$, то есть через середину графика.
b – имеет тот же знак, что и r.

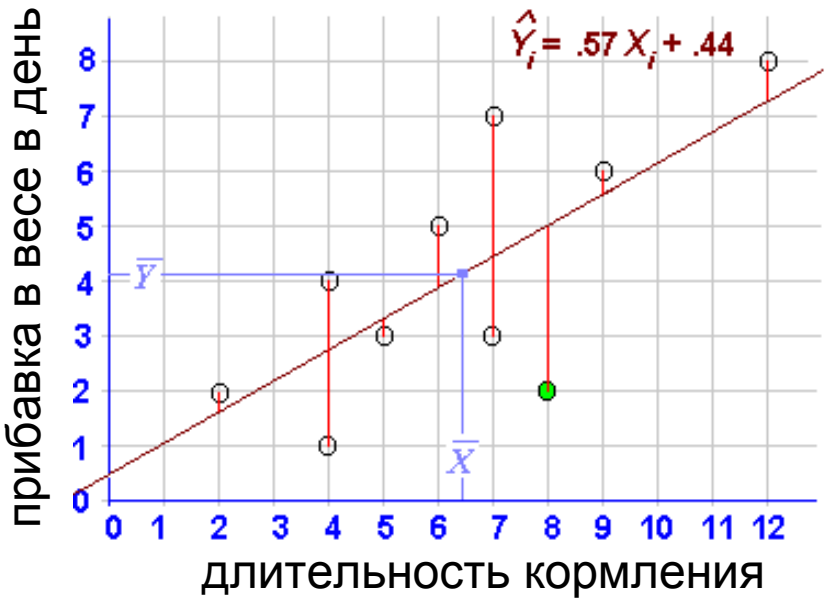

Y
Прибавка в весе в день
$\hat{Y}_i = bX_i + a$
X
Длительность кормления

Если r=0.0, линия регрессии всегда горизонтальна. Чем ближе r к нулю, тем труднее на глаз провести линию регрессии.

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*.

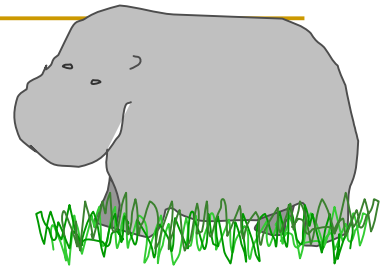
пример про Dr. Nostat, который отобрал 100 самых глупых учеников, подверг их специальной программе и потом протестировал повторно, и их IQ оказался в среднем выше.

Ошибка предсказания (residual)



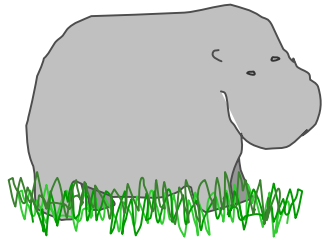
$$e_i = Y_i - \hat{Y}_i$$

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

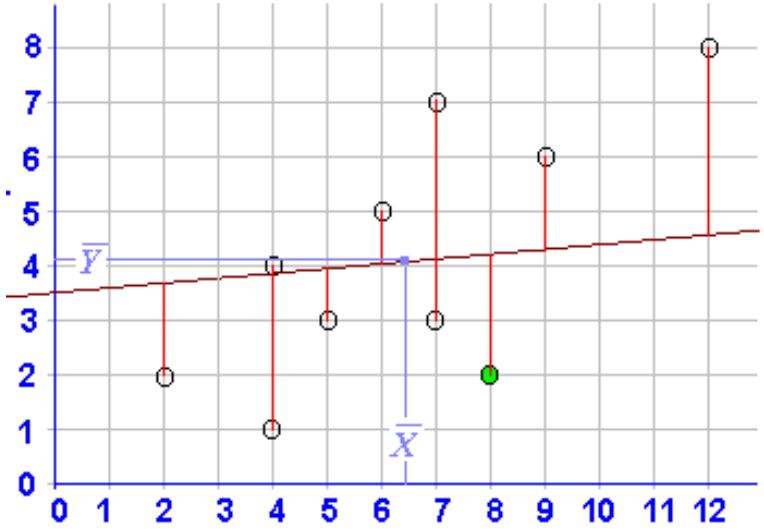


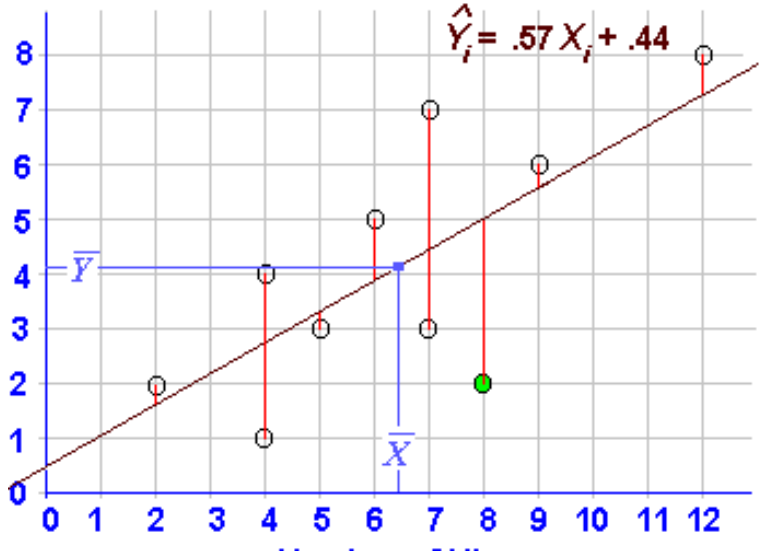
Метод наименьших квадратов:

линию регрессии подбирают такую, чтобы общая сумма квадратов отклонений (residuals) от неё была наименьшей.



$$\sum e_i^2$$ - минимальна





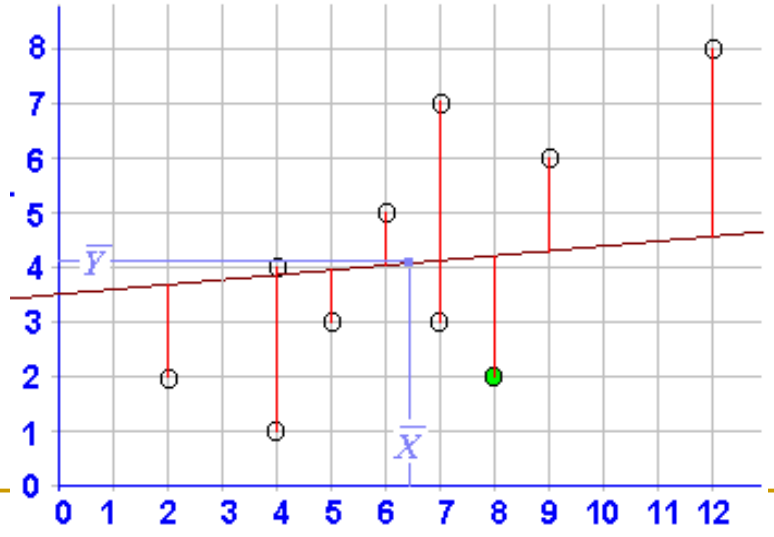
regression

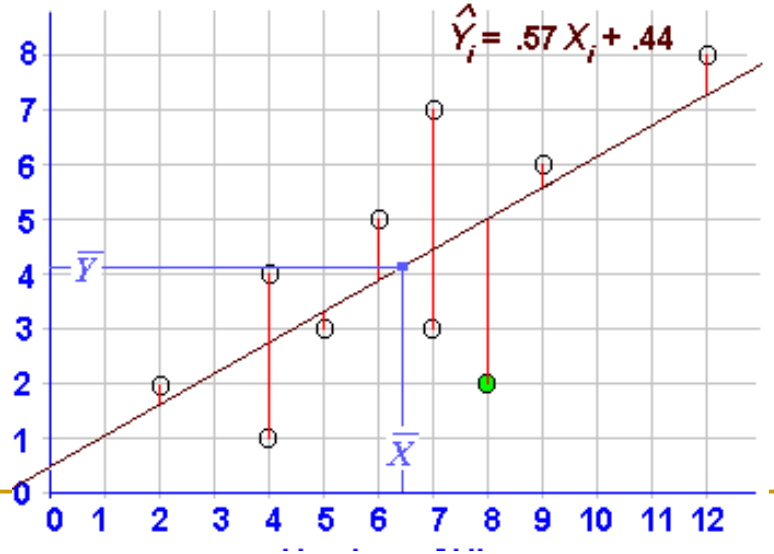
Как оценить, насколько хороша наша линия регрессии, как выбрать лучшую?

Показатель, насколько хороша линия регрессии – стандартное отклонение ошибок e_i: чем оно меньше, тем она лучше!

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

зависит от квадрата коэффициента корреляции

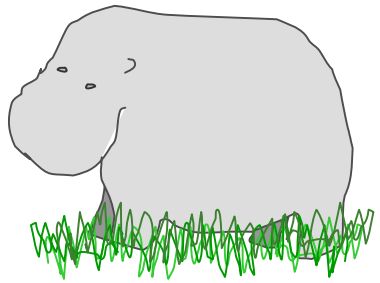




Чем *больше* коэффициент корреляции, тем *меньше* стандартное отклонение ошибки, и наоборот.

Важное требование к выборке: размер этой стандартной ошибки должен быть независимым от X!

Квадрат коэффициента корреляции Пирсона называется коэффициент детерминации (coefficient of determination) - r^2 или 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.

Тестирование гипотезы: отличен ли от нуля наклон линии регрессии?

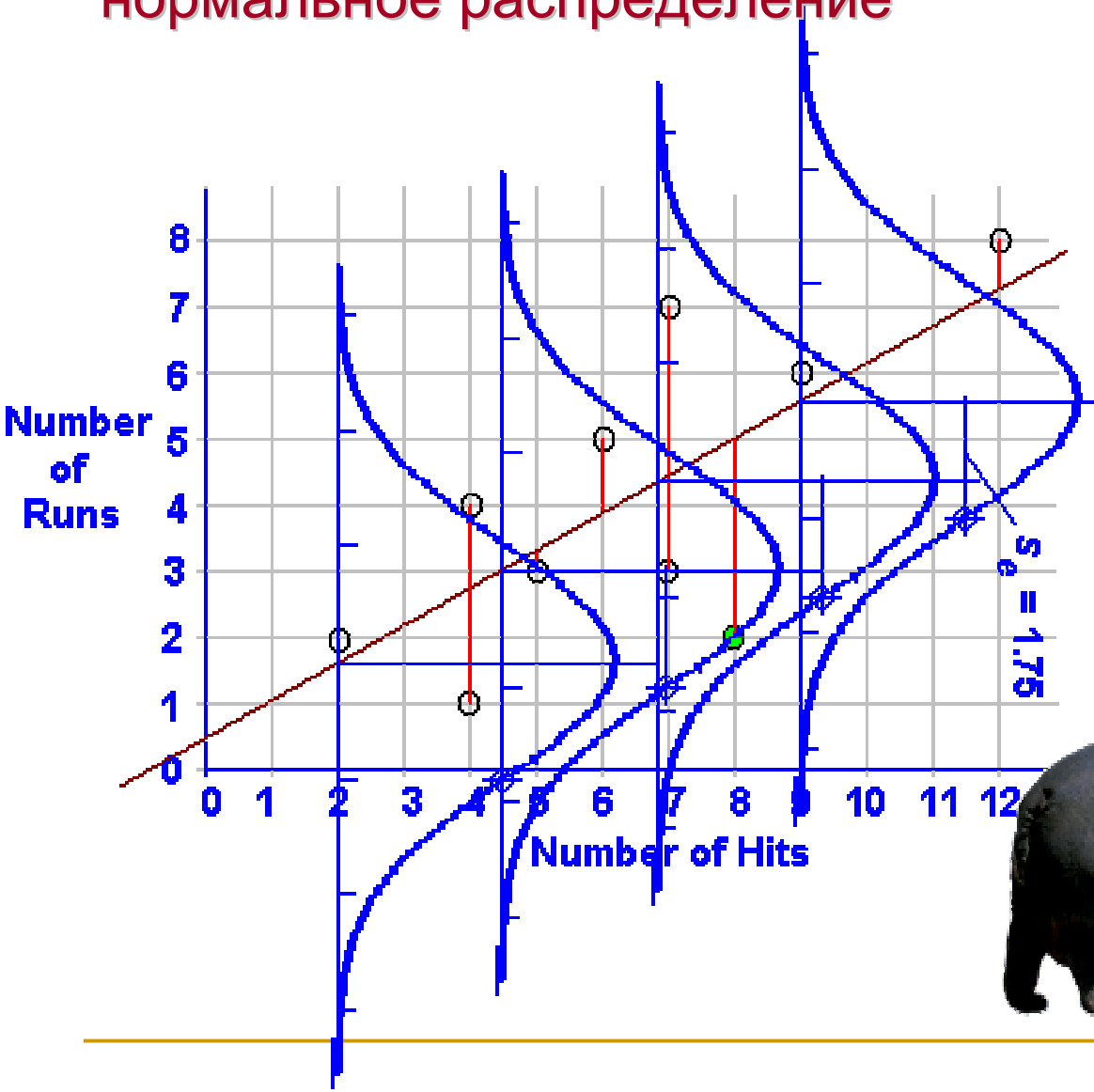
$$H_0: b_{population} = 0$$

Если r достоверно отличается от нуля, то и $b \neq 0$!
То есть, статистикой критерия будет *коэффициент корреляции Пирсона.*

Требования к выборке для построения линии регрессии

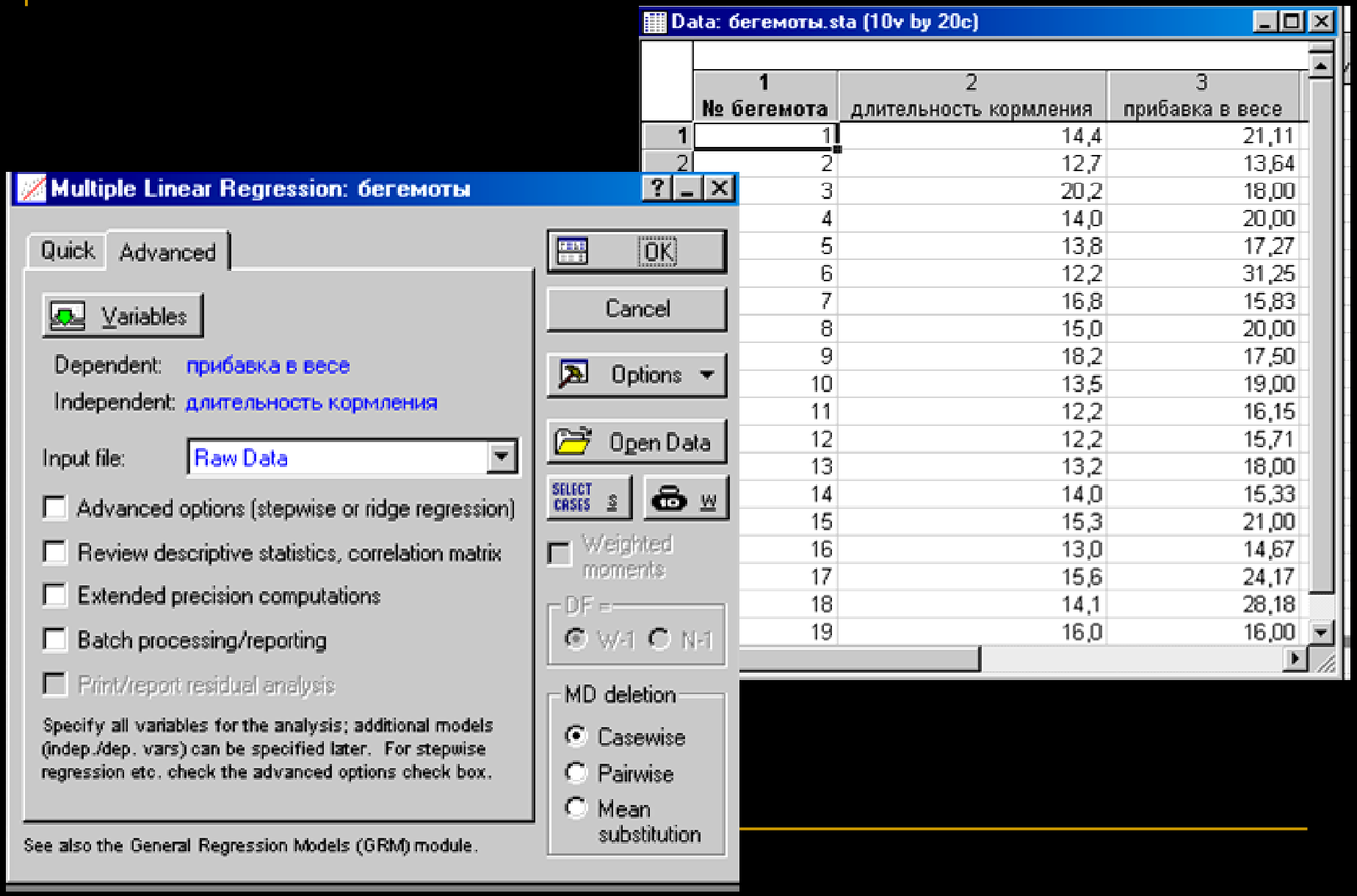
1. Ожидаемая зависимость переменной Y от X должна быть линейной
2. Для любого значения X_i Y должна иметь нормальное распределение *(и ошибка тоже)*
3. Для любого значения X_i выборки для Y должны иметь одинаковую дисперсию
4. Для любого значения X_i выборки для Y должны быть независимы друг от друга

Для любого значения X_i Y должна иметь нормальное распределение



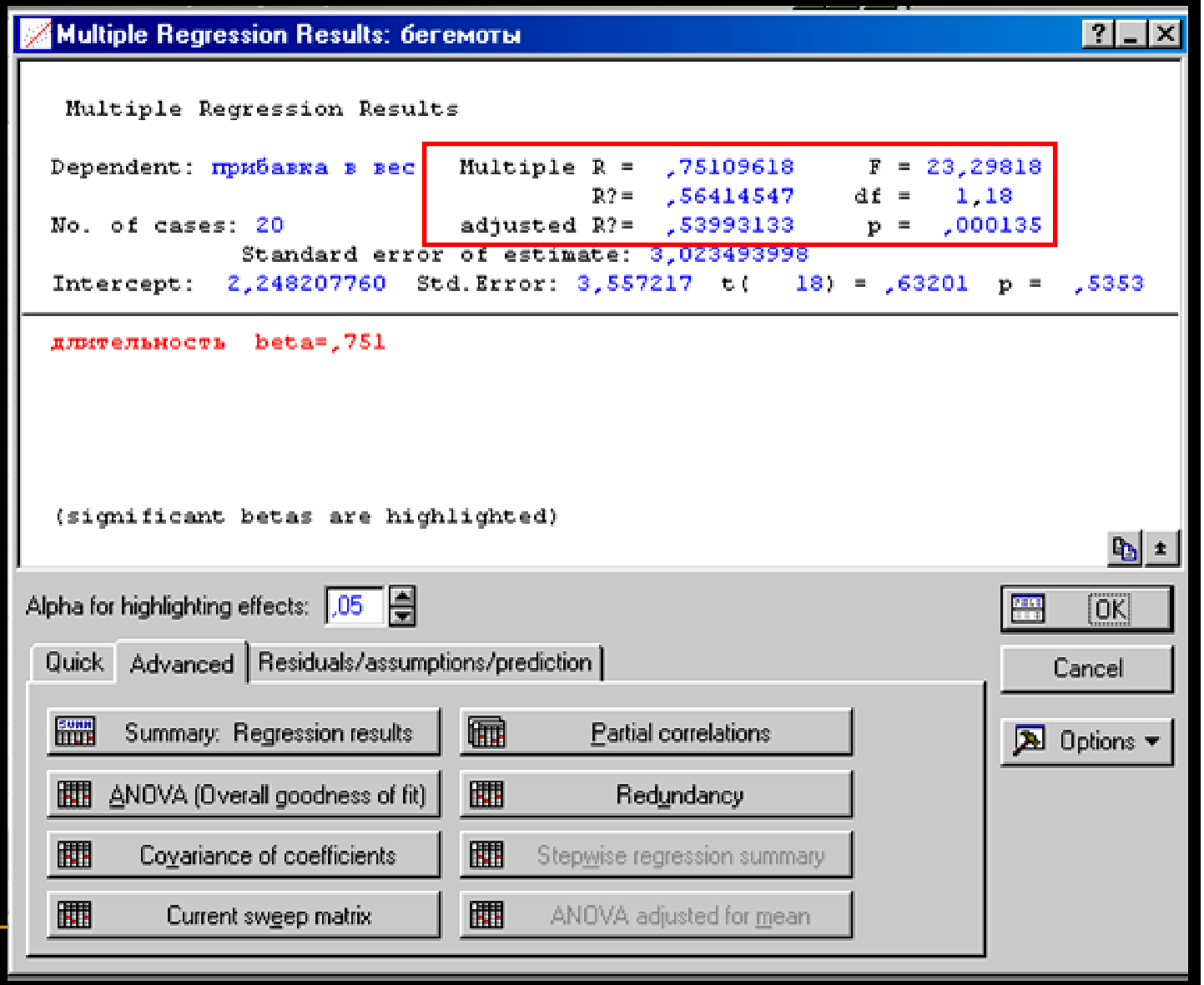
То есть прибавка в весе для всех бегемотов, кормившихся по 20 часов в день имеет нормальное распределение

linear regression

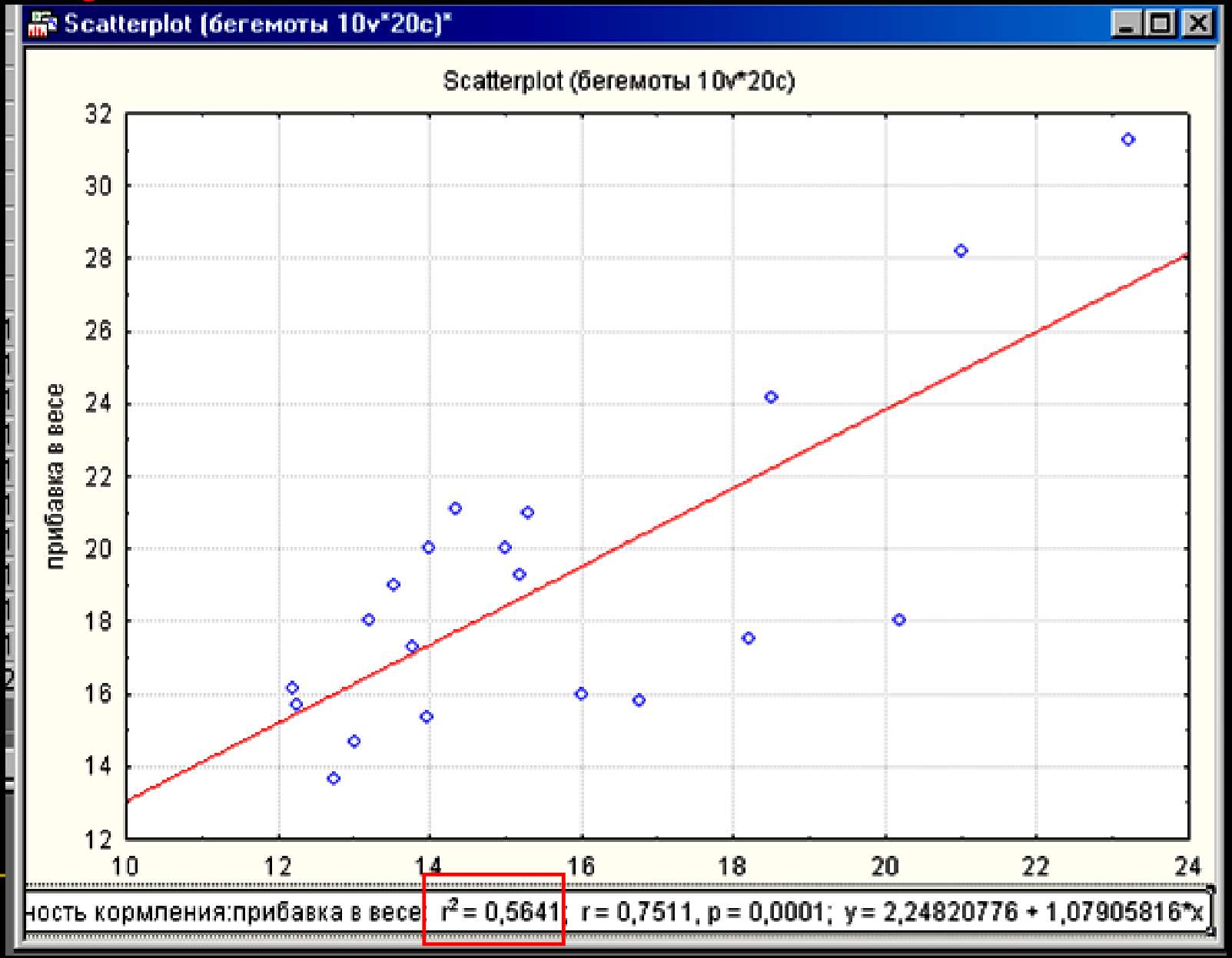


linear regression

У бегемотов прибавка в весе положительно зависела от длительности кормления



linear regression

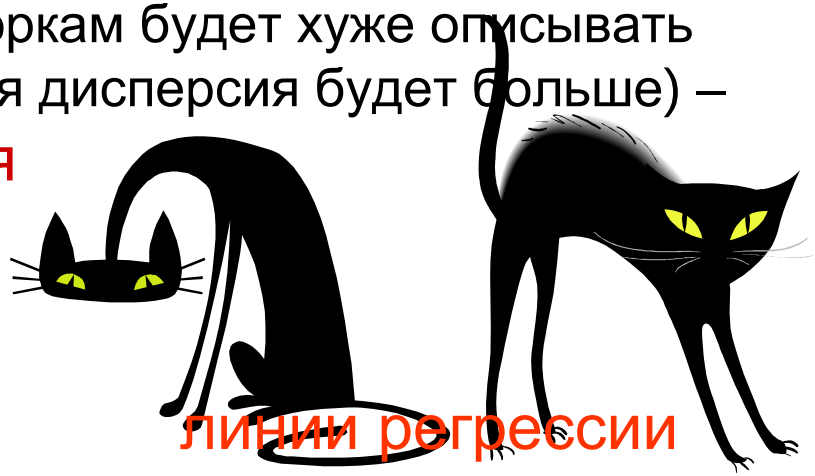


Сравнение двух линий регрессии

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

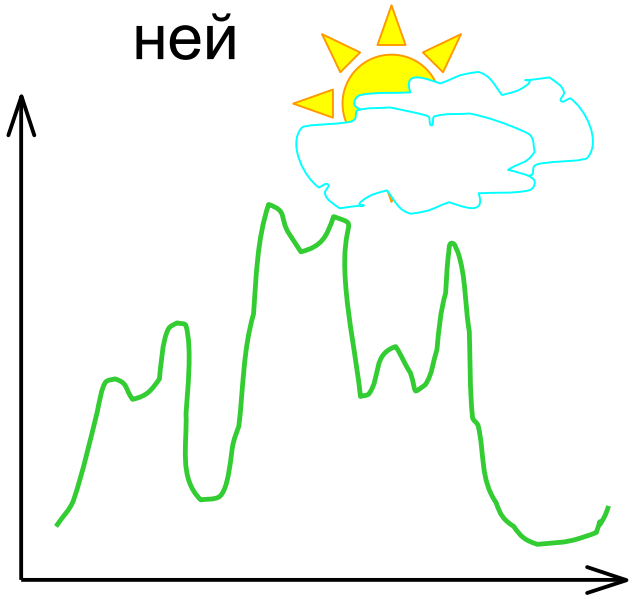


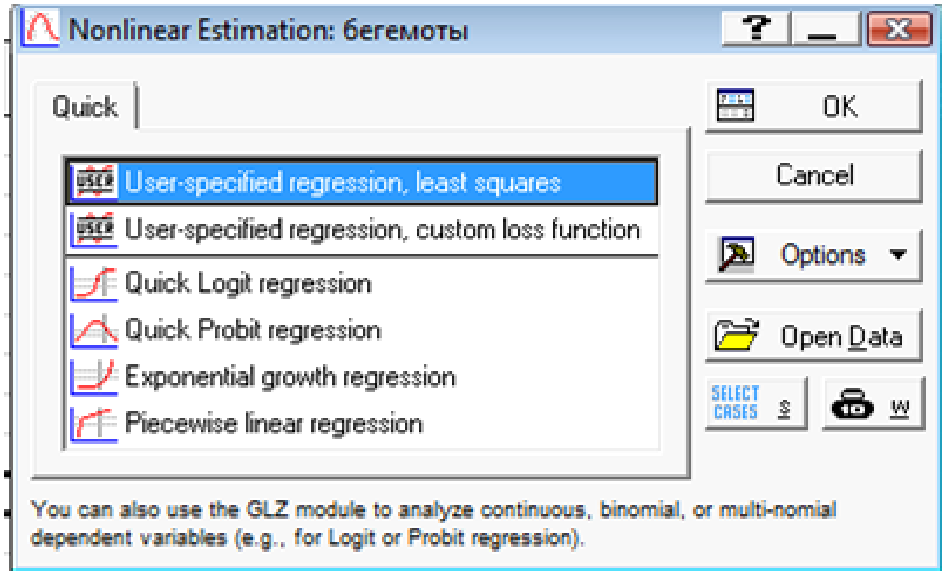
regression

Трансформация в регрессии

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

1. можно трансформировать данные и привести зависимость к линейной;
2. Можно угадать или как-то предположить функцию, которая их связь отражает и потом сравнить данные с ней





regression

Непараметрические методы

1. *Kendall's robust line-fit method*

   Выборку упорядочивают по возрастанию НЕЗАВИСИМОЙ переменной и считают все отношения $Y_j - Y_i$ к $X_j - X_i$. На основе этого считают b.

   Минимальное число измерений - 5

2. *L-test of ordered alternatives*

   В случае, если наши данные в принципе ранговые

В СТАТИСТИКЕ ОТСУТСТВУЮТ

regression

ANCOVA

Когда мы в ANOVA собирались анализировать действие какого-то фактора, стремились к тому, чтобы всякая посторонняя изменчивость была поменьше.

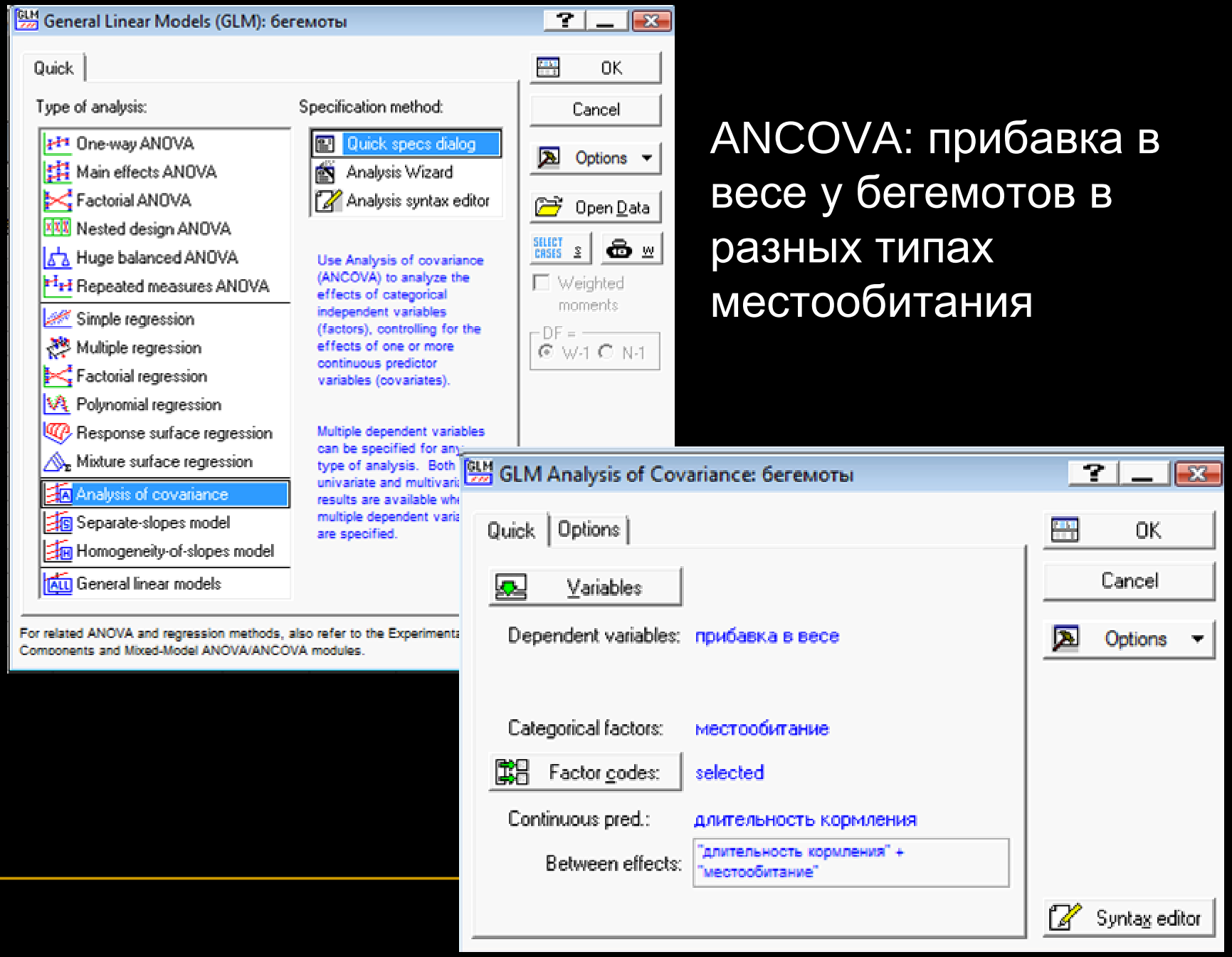
Пример: чтобы проанализировать влияние питания на вес тигров, мы постараемся взять тигров одного возраста и исходно близкой массы.

Но: как быть, если наши тигры изначально разные по весу? Или по возрасту?

(это непрерывные переменные)

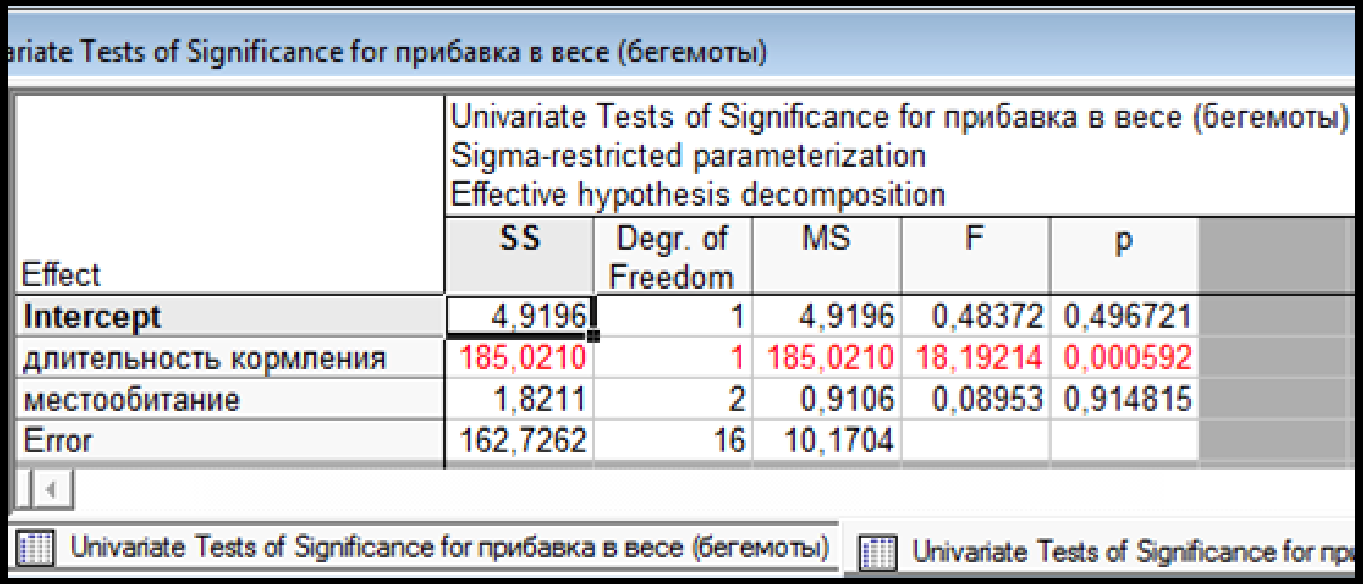
Комбинированный тип анализа –
ANOVA + регрессионный анализ = ANCOVA

regression



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