Занятие 6

Корреляции. Регрессионный анализ

КОРРЕЛЯЦИЯ И РЕГРЕССИЯ

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

Мы изучали, как на значения этой переменной действуют категориальные (=группирующие) переменные.

Настало время обратиться к ситуации, когда количественных переменных в модели ДВЕ И БОЛЕЕ.

Для начала рассмотрим взаимосвязь между ДВУМЯ количественными переменными.

Корреляционный анализ

Мы хотим проанализировать взаимосвязь между двумя переменными – X и Y.

Мы исследуем сусликов. И хотим узнать, как связаны у них масса тела и длина хвоста?

Переменные – 1. масса тела; 2. длина хвоста.

Корреляции

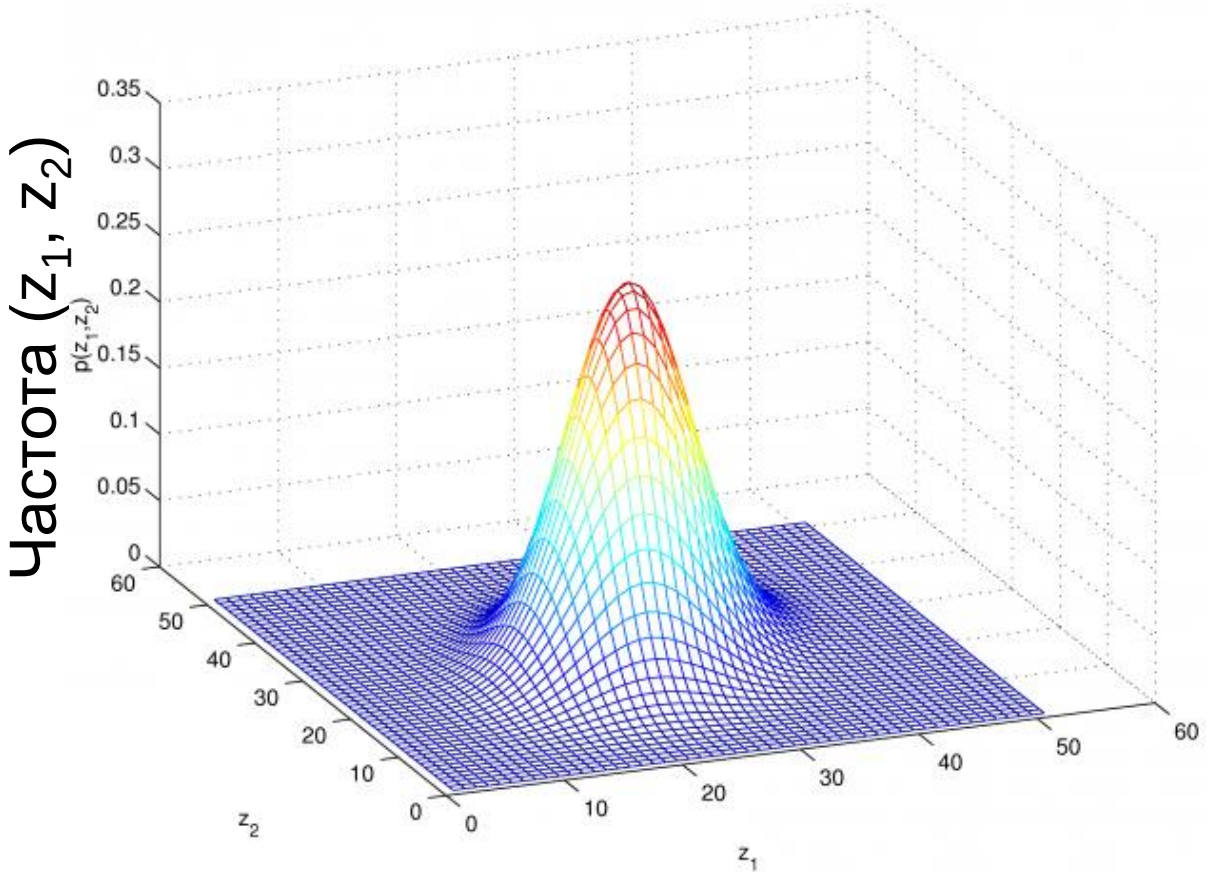
В чём смысл корреляции: мы хотим понять, в какой степени две переменные СОВМЕСТНО ИЗМЕНЯЮТСЯ: если суслик очень тяжёлый, значит ли это, что и хвост у него длинный?
А может, наоборот, короткий?

Если значения одной переменной растут, другой – тоже растут? Уменьшаются? Не изменяются?

Внутри каждой переменной есть изменчивость – большие и маленькие отклонения от среднего. И надо бы, чтоб коэффициент не зависел от размерности переменных.

Двумерное нормальное распределение

Для оценки корреляции переменные должны образовывать двумерное нормальное распределение:



Про случаи, когда это условие нарушено, поговорим позже.

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

Karl Pearson (1857 –1936)

Показатель (описательной статистики!), оценивающий силу линейной связи.

Коэффициент корреляции Пирсона

суслик	*хвост,мм*	*масса, г*
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)

Отклонения каждого значения от среднего

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

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

стандартное отклонение для веса

стандартное отклонение для хвоста

для каждого X и Y (для каждого суслика)

Это одна из нескольких эквивалентных формул для коэффициента корреляции Пирсона

Коэффициент корреляции

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

параметр ВЫБОРКИ

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

параметр ПОПУЛЯЦИИ

Рассчитывая коэффициент корреляции в выборке, мы на самом деле хотим оценить истинный коэффициент в популяции.

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

Коэффициент корреляции

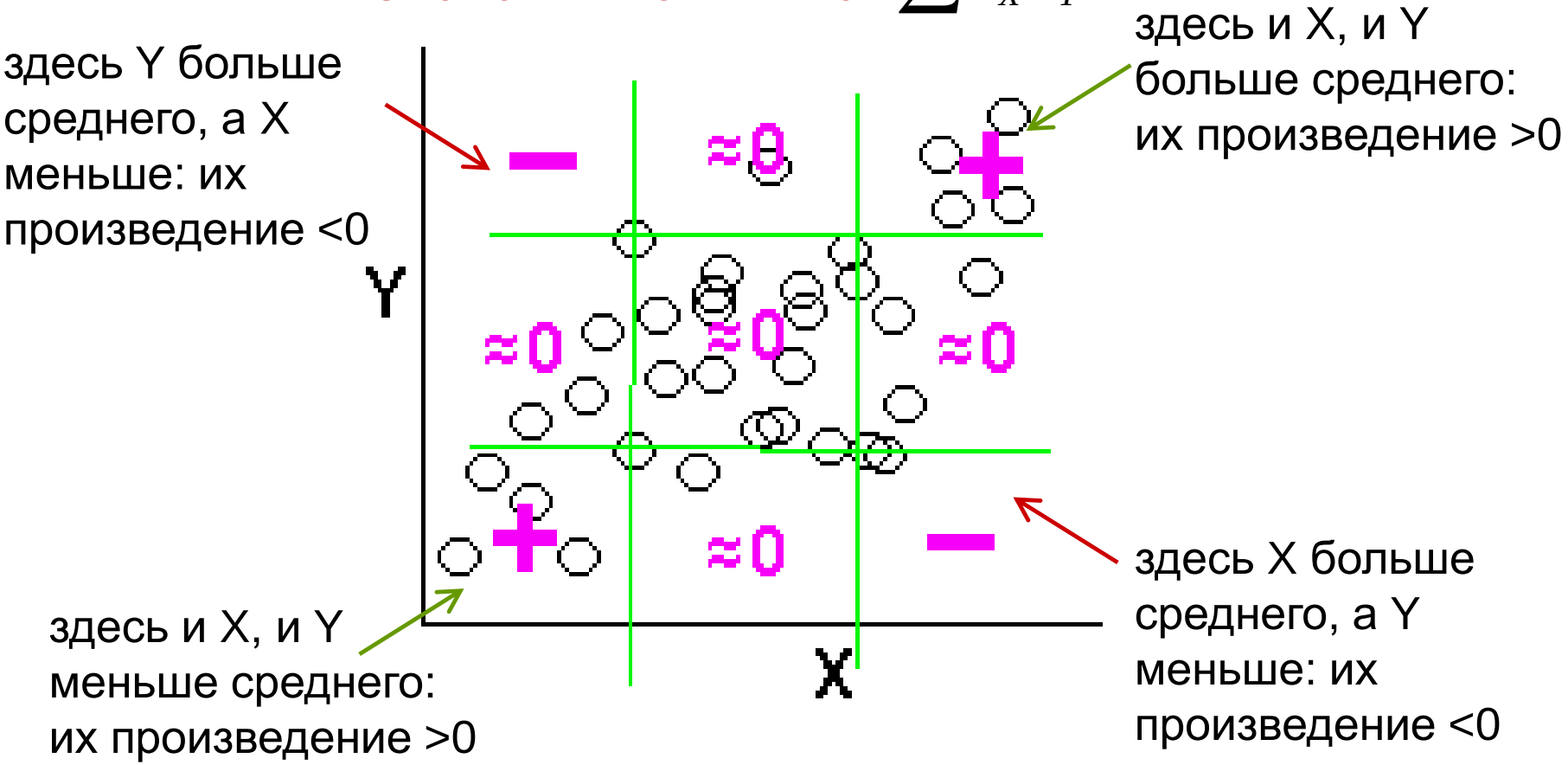
1. Может принимать значения от -1 до +1
2. Знак коэффициента показывает *направление связи* (прямая или обратная: чем больше, тем больше, или чем больше, тем меньше)
3. Абсолютная величина показывает *силу* связи
4. всегда основан на парах чисел (измерений 2-х переменных от одной особи, ручья, камня, и т.п.)

r – в случае, если мы характеризуем ВЫБОРКУ

ρ - если мы характеризуем ПОПУЛЯЦИЮ

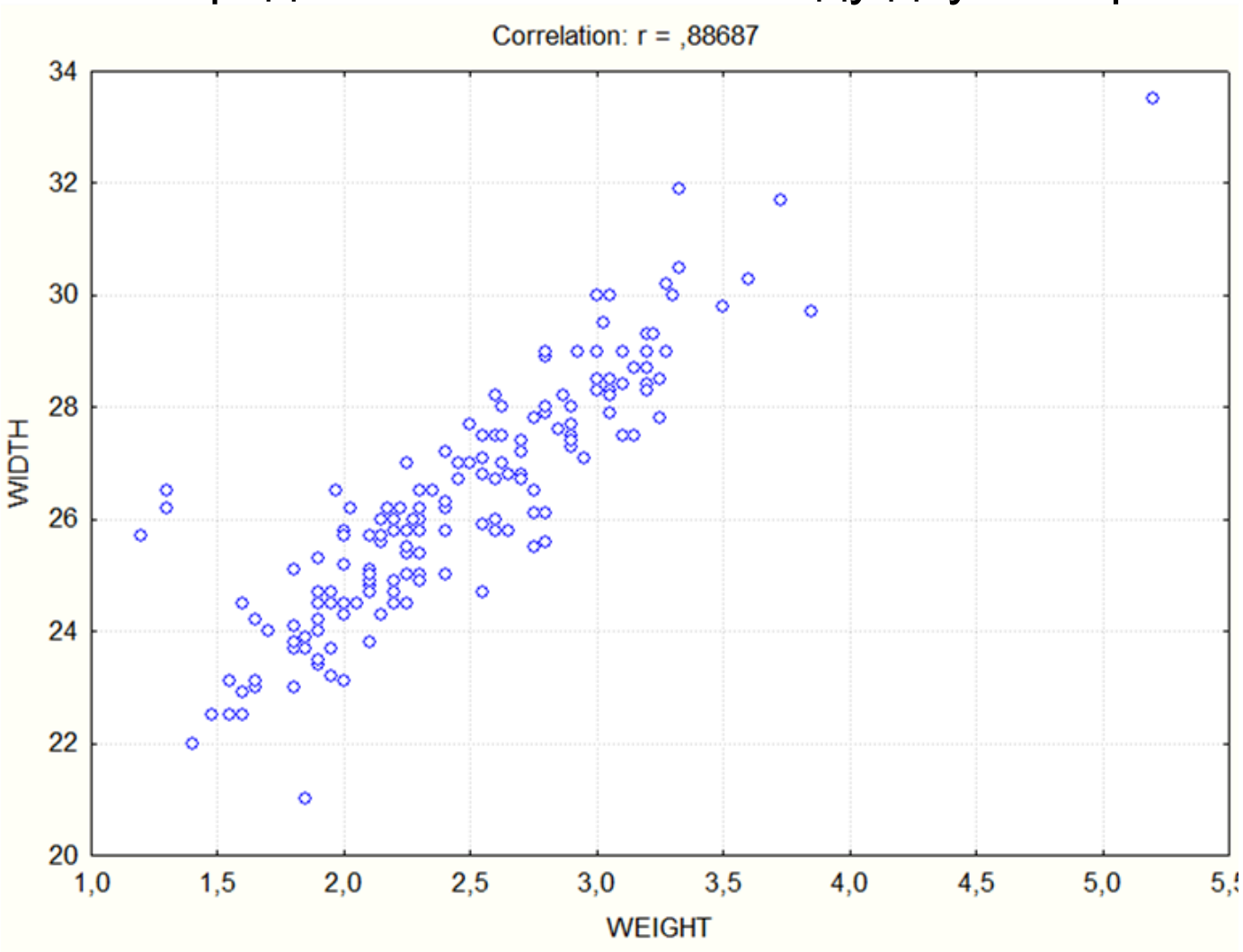
Чем определяются знак и величина коэффициента корреляции?

Знаком и величиной $\sum z_X z_Y$:



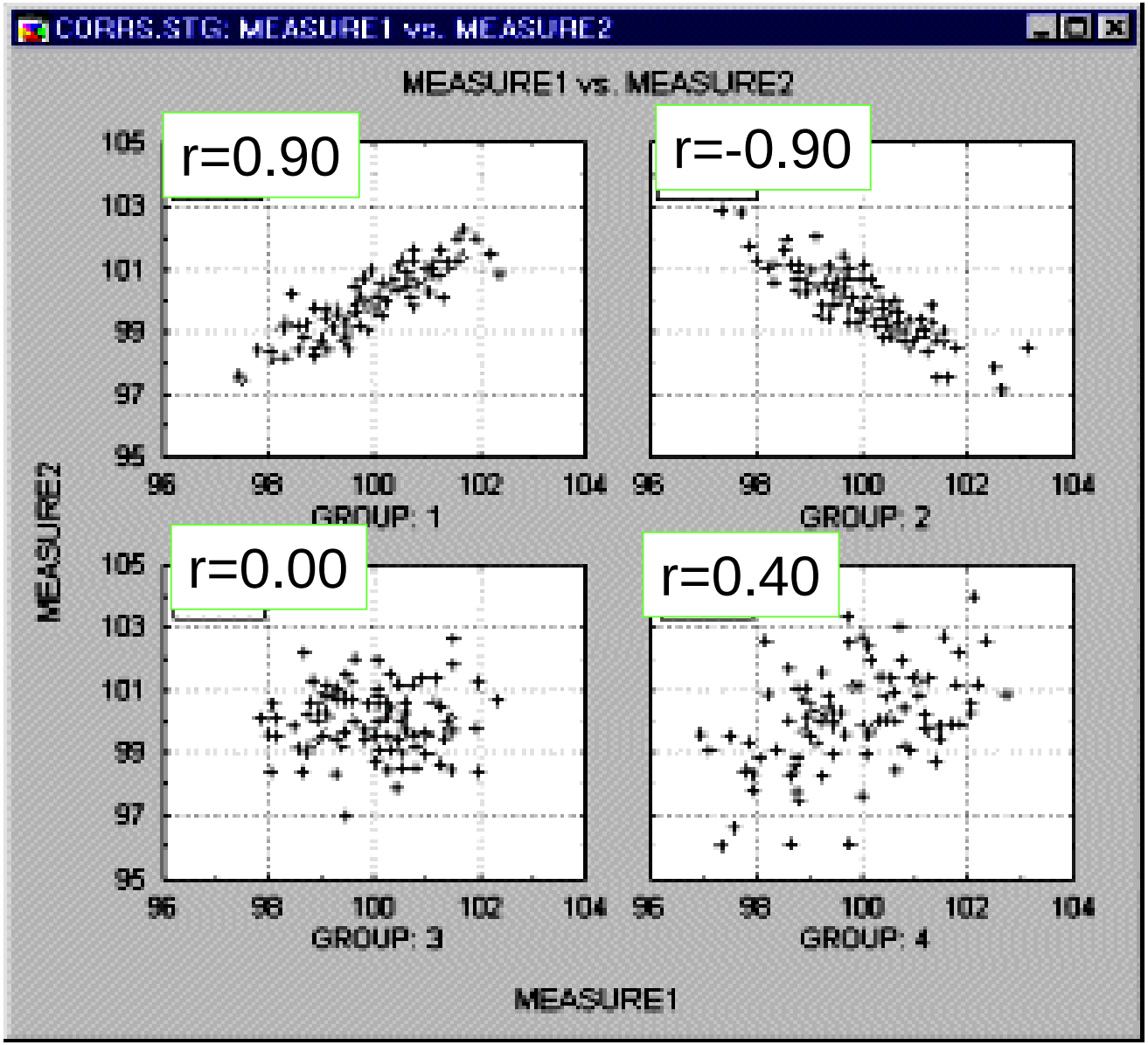
Наибольший вклад в коэффициент вносят точки, расположенные «по углам»

Scatterplot (диаграмма рассеяния) – графическое представление связи между двумя переменными



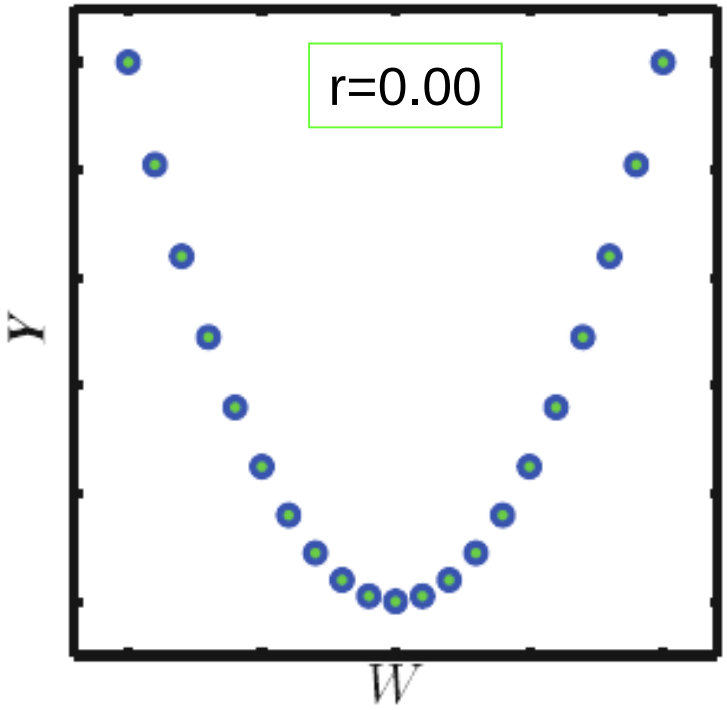
Две характеристики: – наклон (направление связи) и ширина (сила связи) воображаемого эллипса

Корреляции



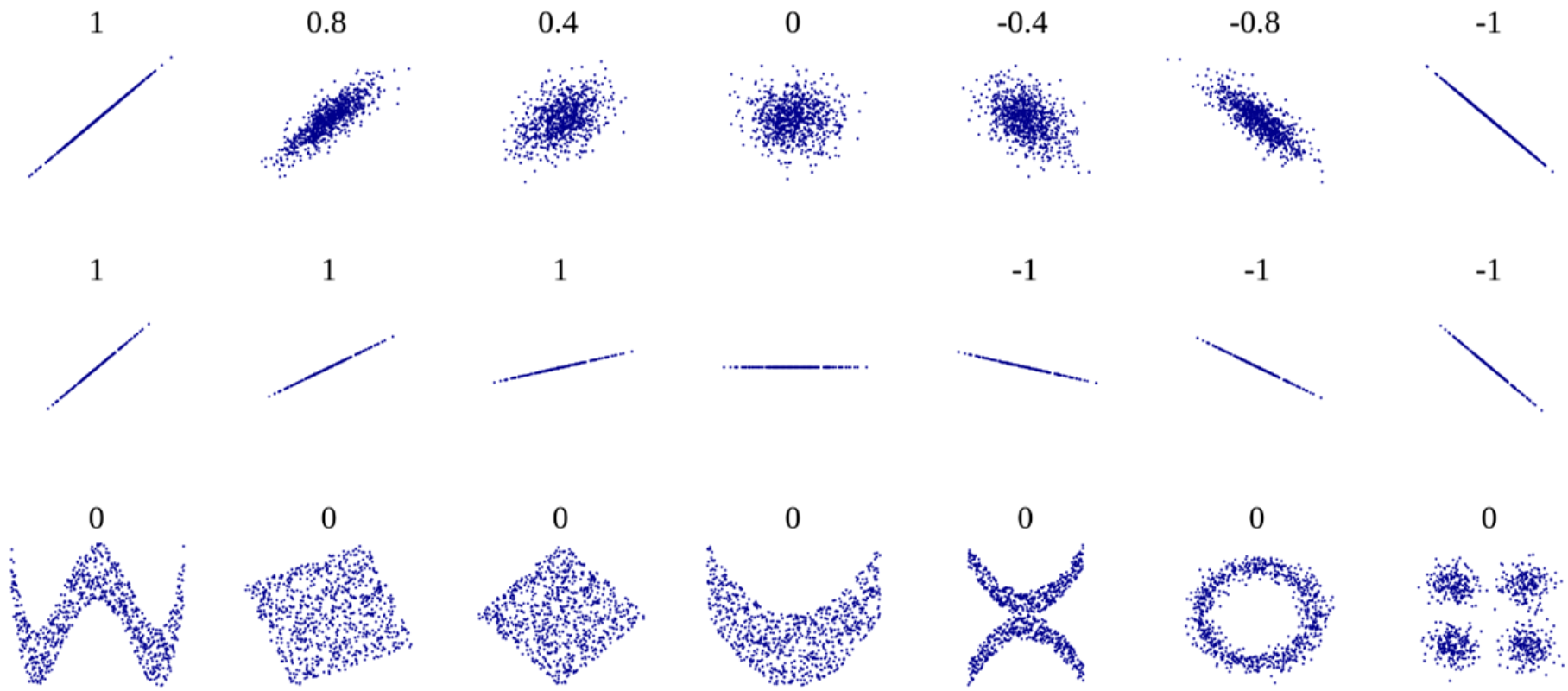
Слабее связь –шире эллипс. r=0.9: если суслик тяжёлый, то и хвост его наверняка длинный; r=1.0: если суслик тяжёлый, то и хвост его точно длинный; r=0: если суслик тяжёлый, хвост его неизвестно какой.

Коэффициент корреляции Пирсона оценивает силу только линейной связи! Он может быть близок к нулю даже при явной, но нелинейной связи.



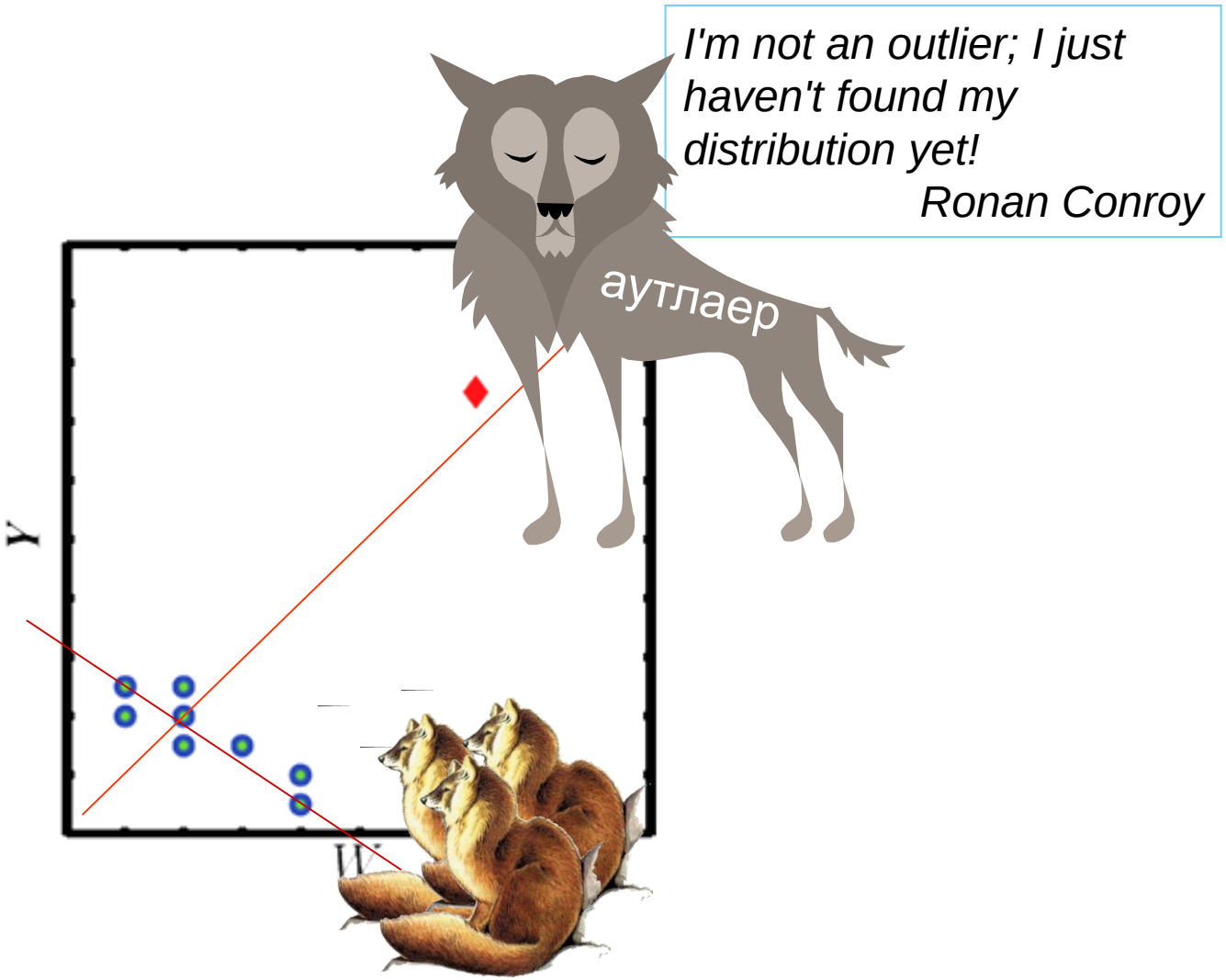
Нелинейные взаимосвязи могут становиться линейными при трансформации данных.

Корреляции



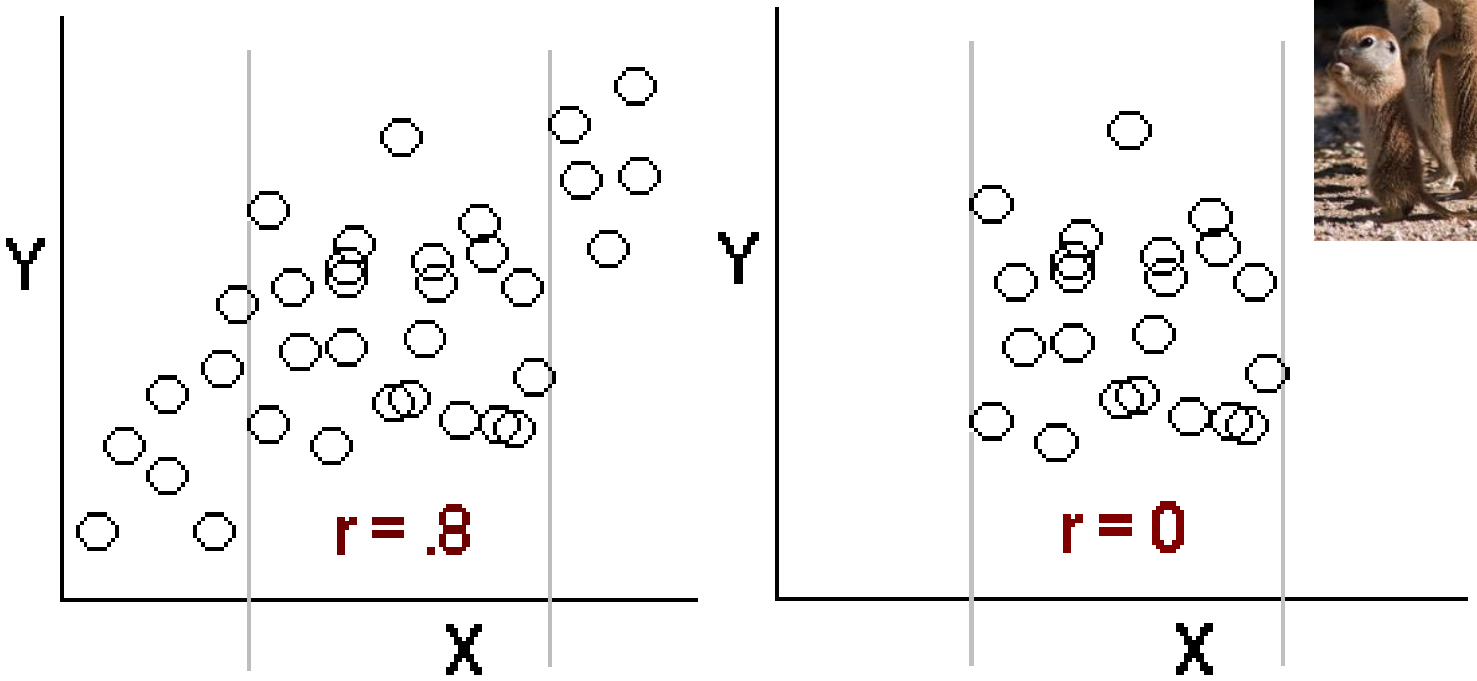
Корреляции

Коэффициент корреляции Пирсона очень чувствителен к аутлаерам. Настолько чувствителен, что один аутлаер может изменить знак *r* .



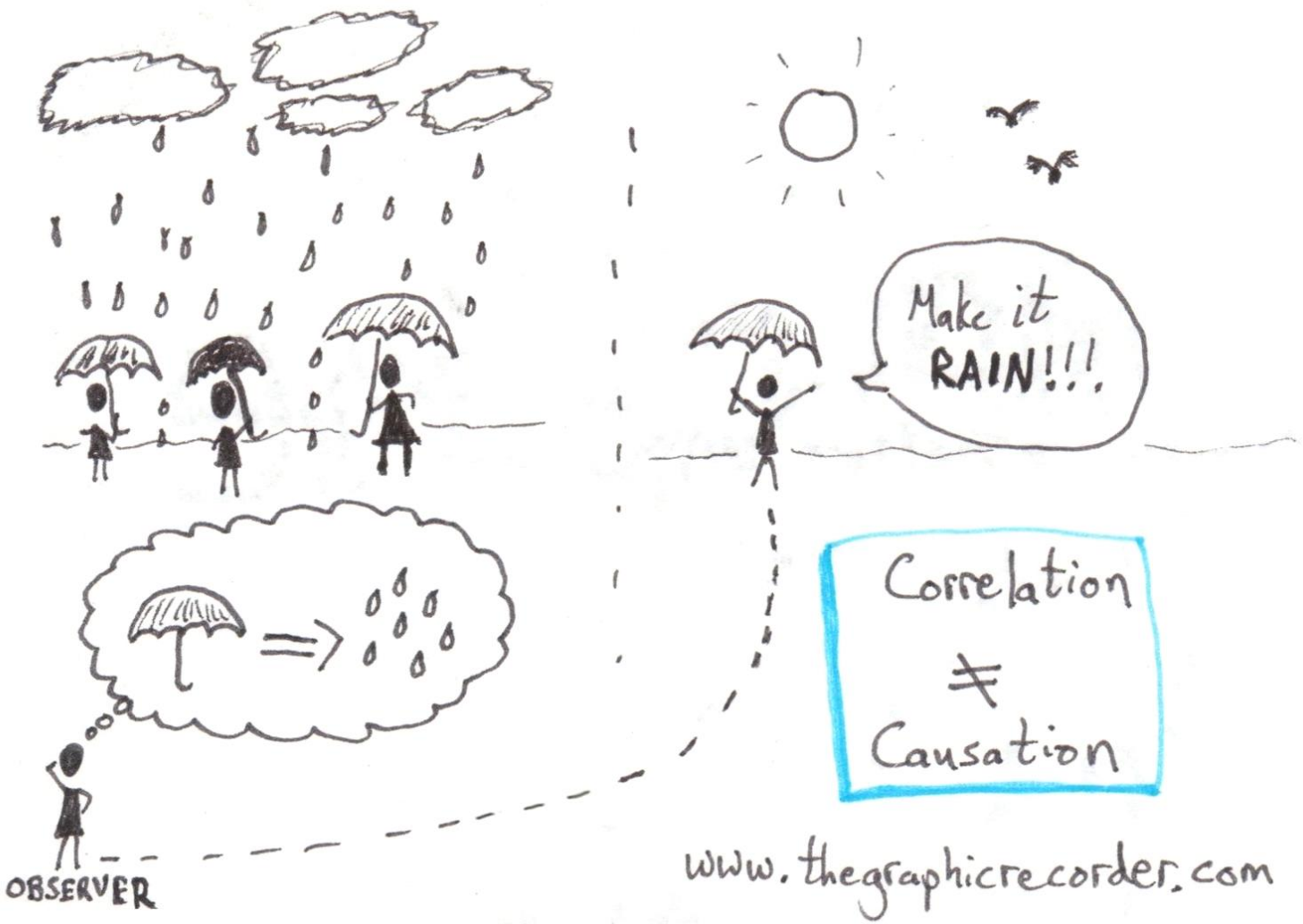
Корреляции

Секрет, важный при постановке исследований, предполагающих анализ корреляций: нужно, чтобы в исследуемых переменных была достаточно большая изменчивость. Если отобрать в выборку близких по массе зверьков, бесполезно потом пытаться выявить связь массы тела и длины хвоста.



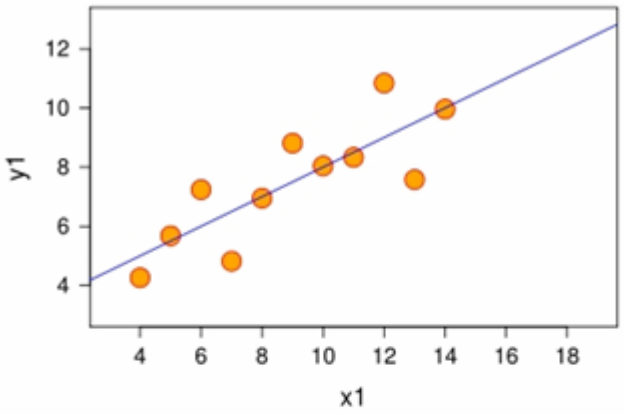
Корреляции

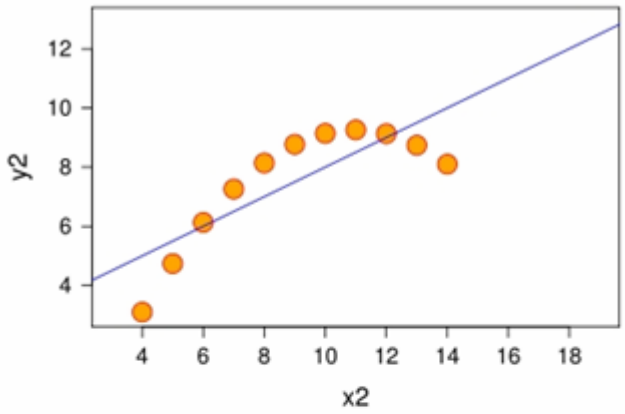
Очевидно, что корреляция между переменными ничего не говорит о причинно-следственной связи между ними!



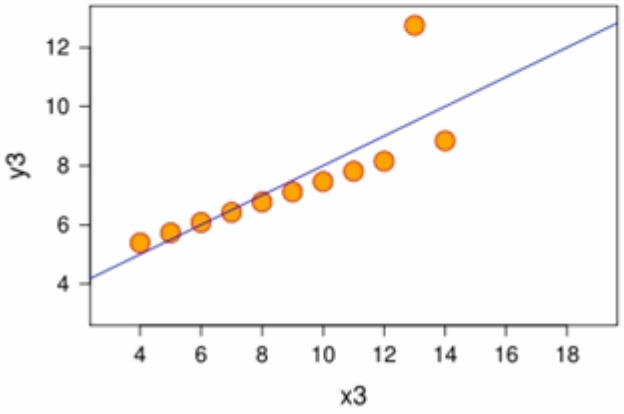
Тестирование гипотезы о коэффициенте корреляции

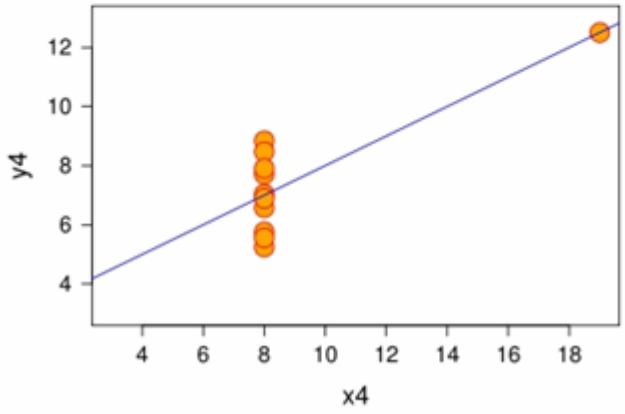
Как судить о линейной взаимосвязи между переменными в популяции, если в наших руках лишь выборка?
Сам по себе выборочный коэффициент корреляции недостаточен.





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





Мы можем проверить гипотезу о равенстве популяционного ρ нулю: так мы выясним, есть ли вообще между переменными линейная связь, или нет.

Оказывается, что если много раз из популяции, где ρ=0, формировать выборки размера N, r из этих выборок образуют нормальное распределение. Так что тестировать гипотезу о ρ очень просто!

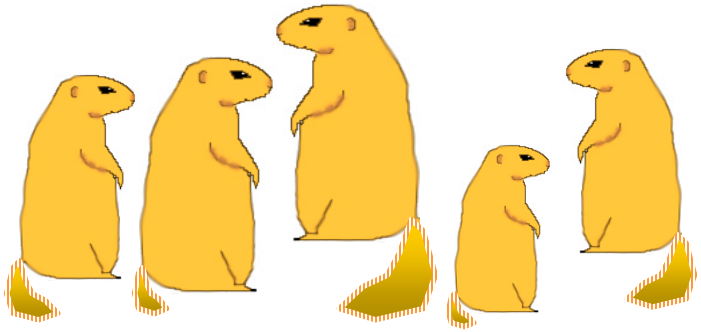
$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

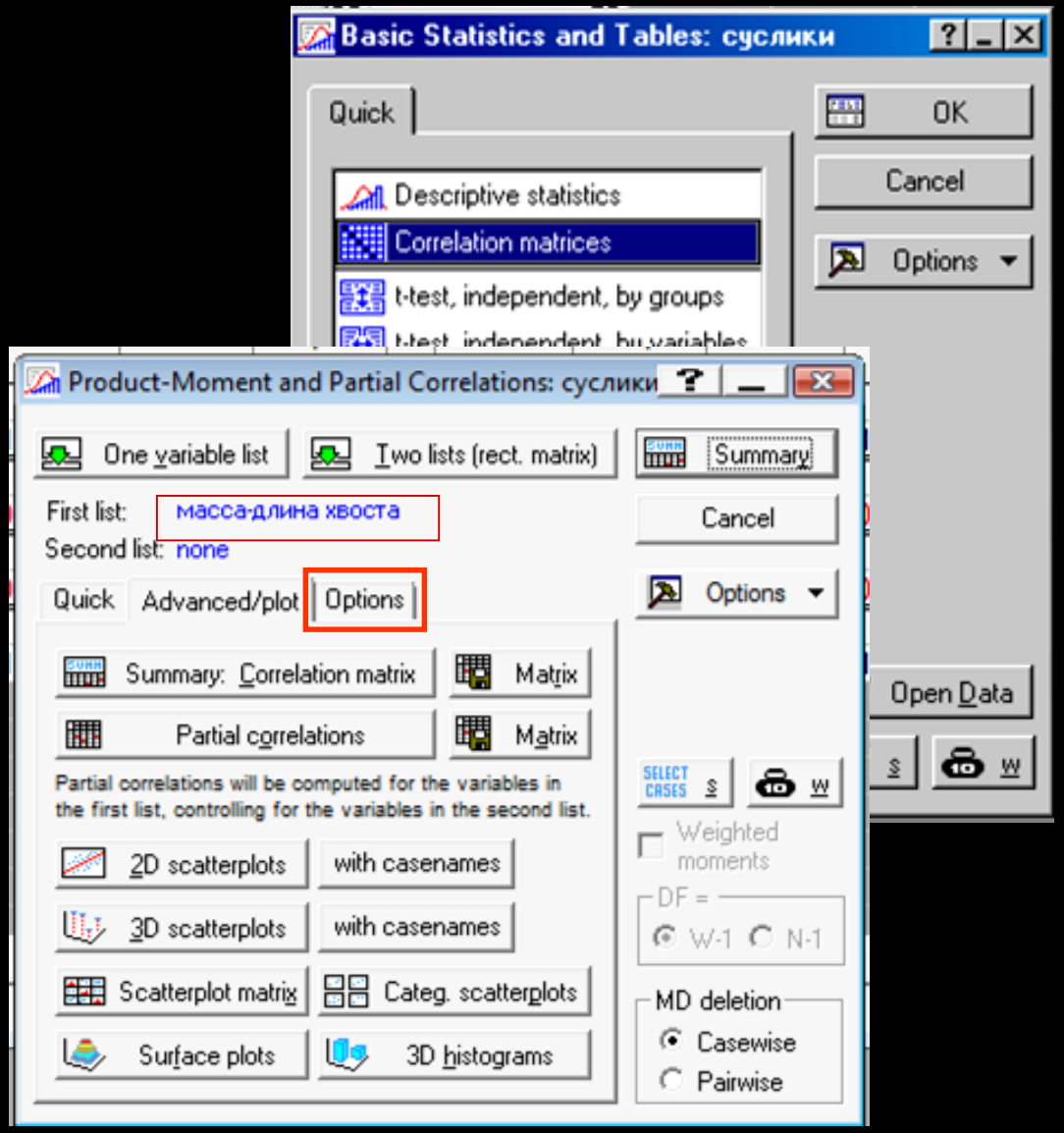
Связаны ли у сусликов масса тела и длина хвоста (линейной связью)?

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

стандартная ошибка коэффициента корреляции



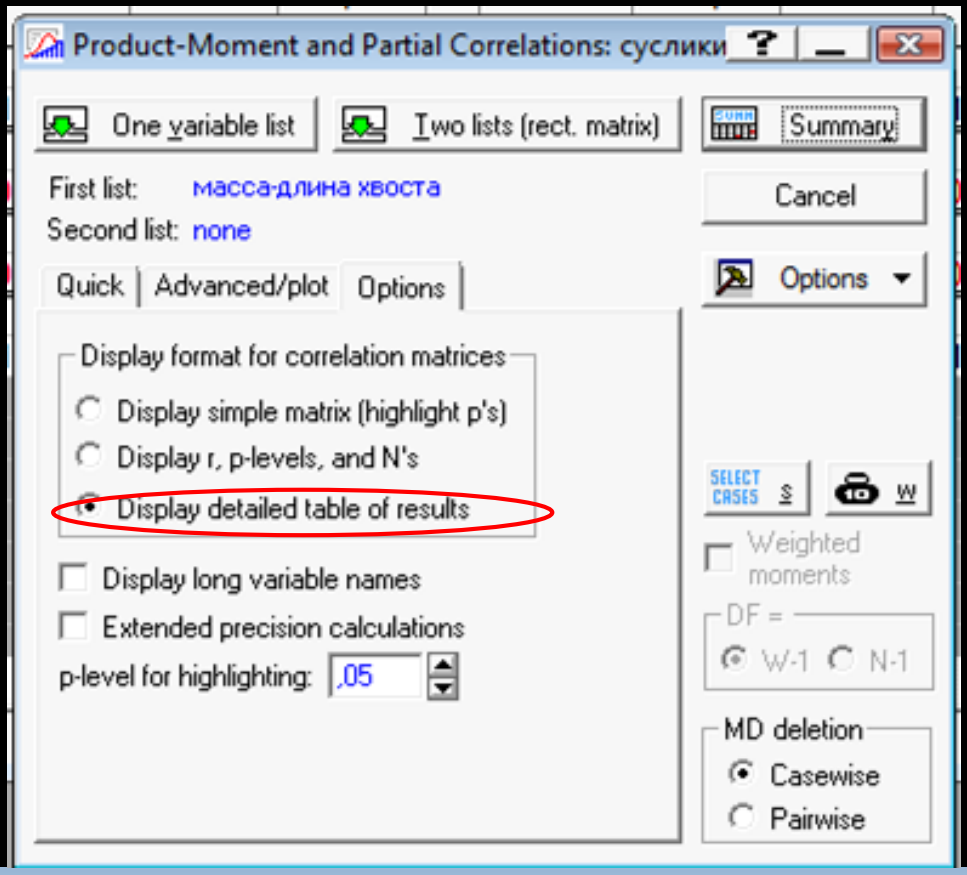
Pearson product-moment correlation coefficient *r*



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: масса тела у сусликов положительно связана с длиной хвоста.



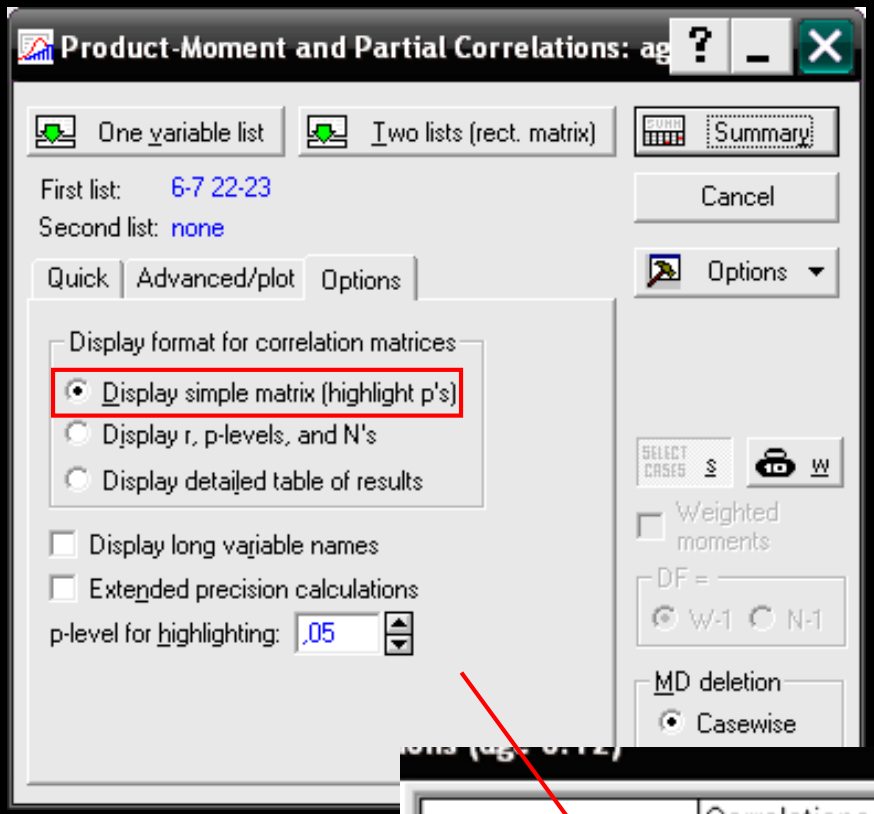
elations (суслики)

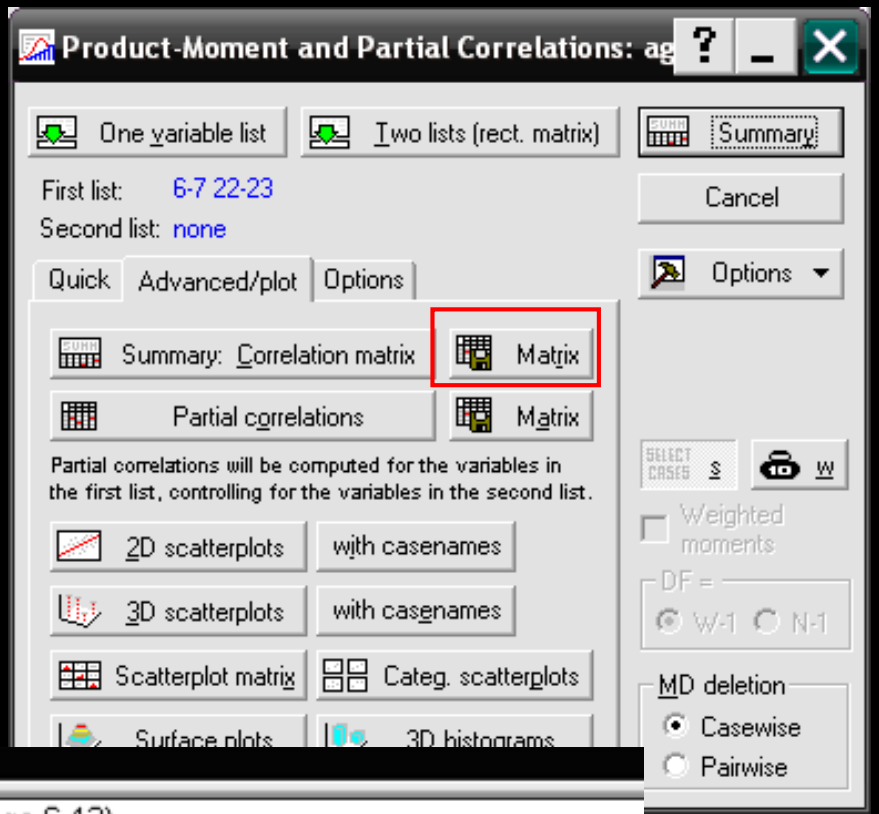
Correlations (суслики)
Marked correlations are significant at p < ,05000
(Casewise deletion of missing data)

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

Получение МАТРИЦЫ 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

Assumptions

1. Выборка должная быть случайной, а измерения в ней – независимыми: нельзя много раз измерить переменные у одного и того же зверька.
2. Двумерное нормальное распределение (проверяется тестом Шапиро-Уилкса и скаттерплотом). Отклонения от «нормальности» НЕ компенсируются увеличением размера выборки
3. Линейность взаимосвязи между переменными (проверка - скаттерплот).

Что делать, если нарушены условия 2 и 3?
Трансформировать данные! Часто нелинейность связи и возникает по причине скошенных распределений; трансформация может разом исправить и ненормальность, и нелинейность!
Не помогает? Тогда см. лекцию 7.

Корреляции

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

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

Cohen, 1988:

ρ = 0.1 - слабая корреляция;

ρ = 0.3 – корреляция средней силы;

ρ = 0.5 - сильная корреляция.

РЕГРЕССИОННЫЙ АНАЛИЗ

В анализе корреляций у нас переменные были равнозначны; всё, что мы хотели – просто сказать, есть ли между ними линейная связь, и прикинуть, сильна ли она.

Теперь мы хотим проанализировать влияние количественной переменной (независимой = PREDICTOR) на другую, зависимую, переменную (RESPONSE).
Эта задача и вся эта процедура очень близки ANOVA.

Сейчас мы будем создавать модель!

Регрессии

Зачем всё это нужно?

1. Регрессионный анализ оценивает, НА СКОЛЬКО изменится значение одной переменной на единицу изменения другой (на сколько вырастет урожай помидоров, если мы внесём 100 кг удобрения?);
2. Мы можем на основании значения одной переменной ПРЕДСКАЗАТЬ значение другой переменной, не измеряя её (какой длины будет хвост суслика массой 500 г?);
3. В регрессионном анализе можно изучить влияние сразу НЕСКОЛЬКИХ ПРЕДИКТОРОВ на зависимую переменную!

Регрессии

То есть,

РЕГРЕССИЯ (*regression*) – предсказание одной переменной на основании другой. Одна переменная – независимая, другая – зависимая, и мы предполагаем, что предиктор дает биологическое объяснение значениям зависимой переменной.

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

ЭТО НЕ ОДНО И ТО ЖЕ

Регрессии

Выполнить задачи регрессионного анализа нам помогла бы формула вроде

Response = «что-то» x «predictor» + ошибка

(без ошибки не получится, ведь мы имеем дело с живыми объектами! Это та часть изменчивость зависимой переменной, которая не объясняется моделью)

Например,

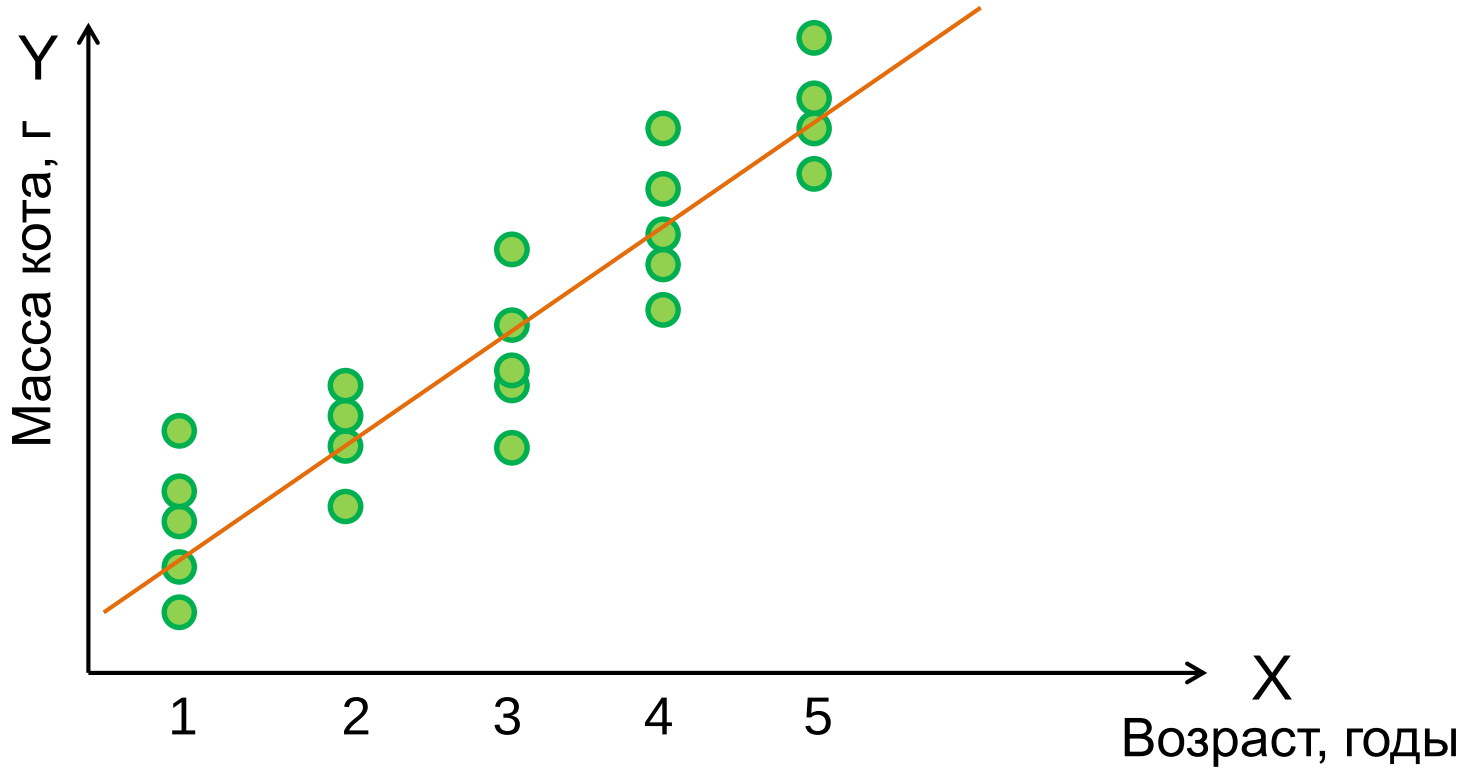
Масса кота = (500 грамм) x возраст кота в годах + ошибка

Такая формула нам позволит запросто предсказать массу кота, зная, что ему 5 лет.

Регрессии

Простая линейная регрессия (X – predictor, Y - response):

1. Описывает линейную взаимосвязь между X и Y;
2. Позволяет предсказать новые значения Y для новых значений X;
3. Определяет, какая часть изменчивости Y объясняется X, а какая остается не объяснённой.



Регрессии

Представим группу из n котов (i= от 1 до n), для которых известны масса и возраст.

$$Y_i = a + bX_i + e_i$$

Для выборки

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

Для популяции

Это – уравнение линии регрессии. В нём есть коэффициенты:

β (b) – характеризует **НАКЛОН** прямой (slope); это самый важный коэффициент;

α (a) – определяет точку пересечения прямой с осью OY; не столь существенный (**intercept**).

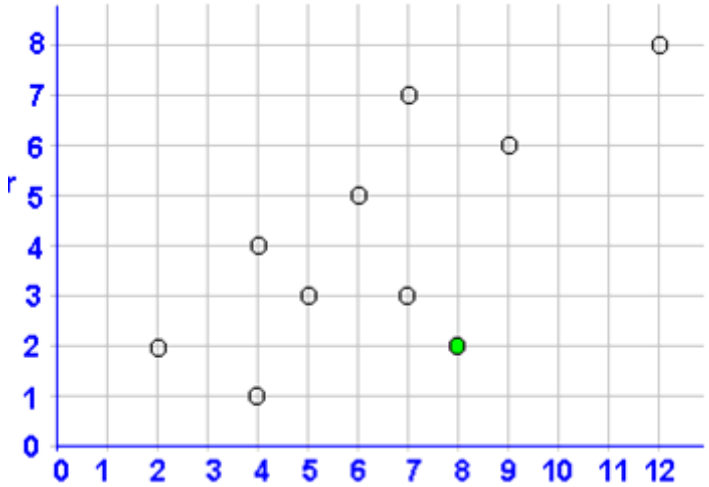
ε – ошибка, показывает, насколько реальное значение Y_i отличается от предсказанного моделью (от точки на прямой).

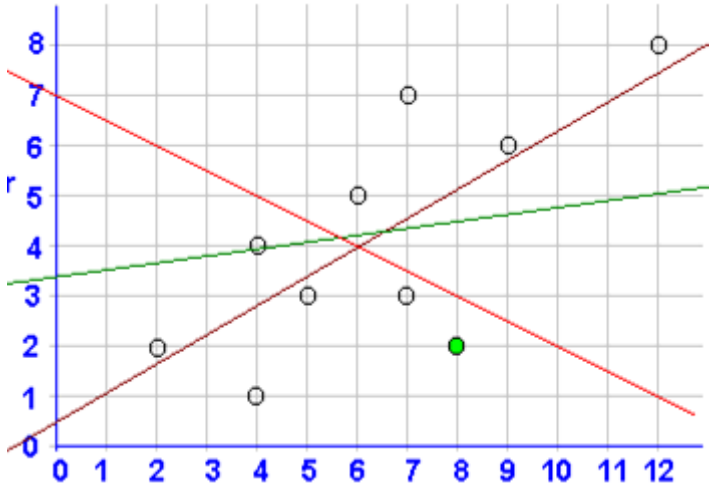
Мы хотим оценить эти коэффициенты.
Сперва научимся считать их для выборки!

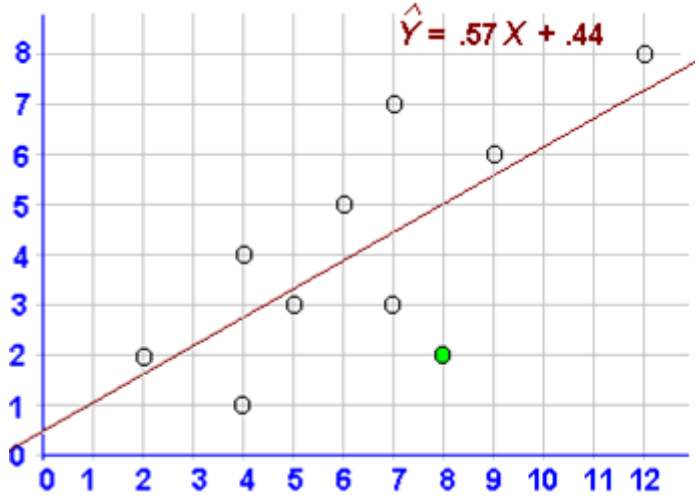
Регрессии

Строим ЛИНИЮ регрессии, чтобы на основе своей выборки оценить ПОПУЛЯЦИОННЫЕ коэффициенты линейной модели.

Поиск «лучшей» прямой:



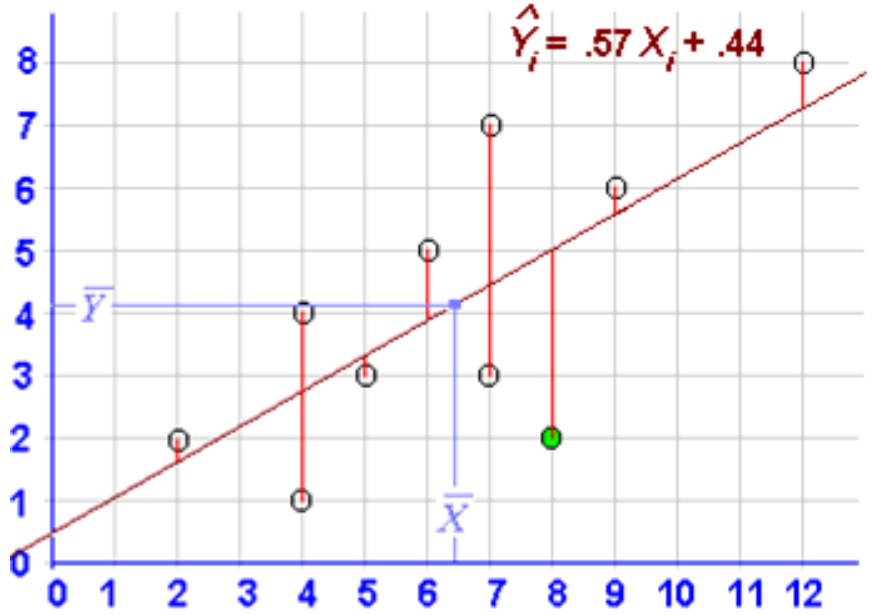




Регрессии

Очевидно, что провести прямую через все точки разом невозможно. Точки будут отстоять от прямой на некоторое расстояние – ошибку предсказания (RESIDUAL) = «остатки»

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



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

Регрессии

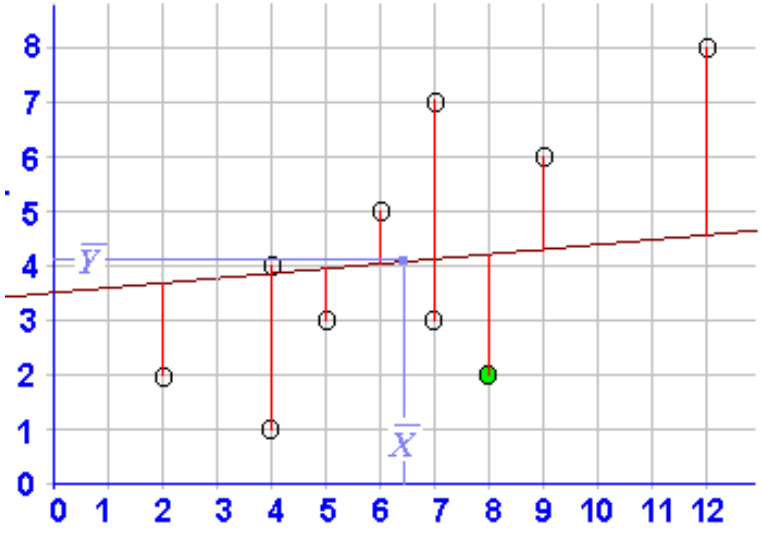
Как определить «лучшую» линию регрессии?

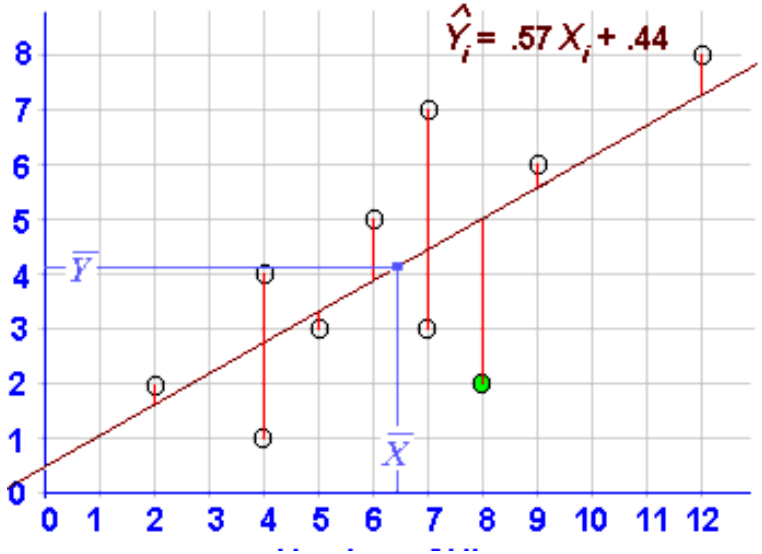
Метод наименьших квадратов:

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

Рассчитываем коэффициенты a и b (по ним будем оценивать популяционные коэффициенты):

b:

$$b_1 = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

covariance

SS для X

Связь с коэффициентом корреляции:

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

стандартные отклонения для X и Y

коэффициент корреляции Пирсона

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

Регрессии

a:

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

Поэтому считаем a, просто подставив в уравнение средние значения.

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

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

Вот уравнение линии регрессии, где Ŷ – предсказанное (predicted) значение Y.

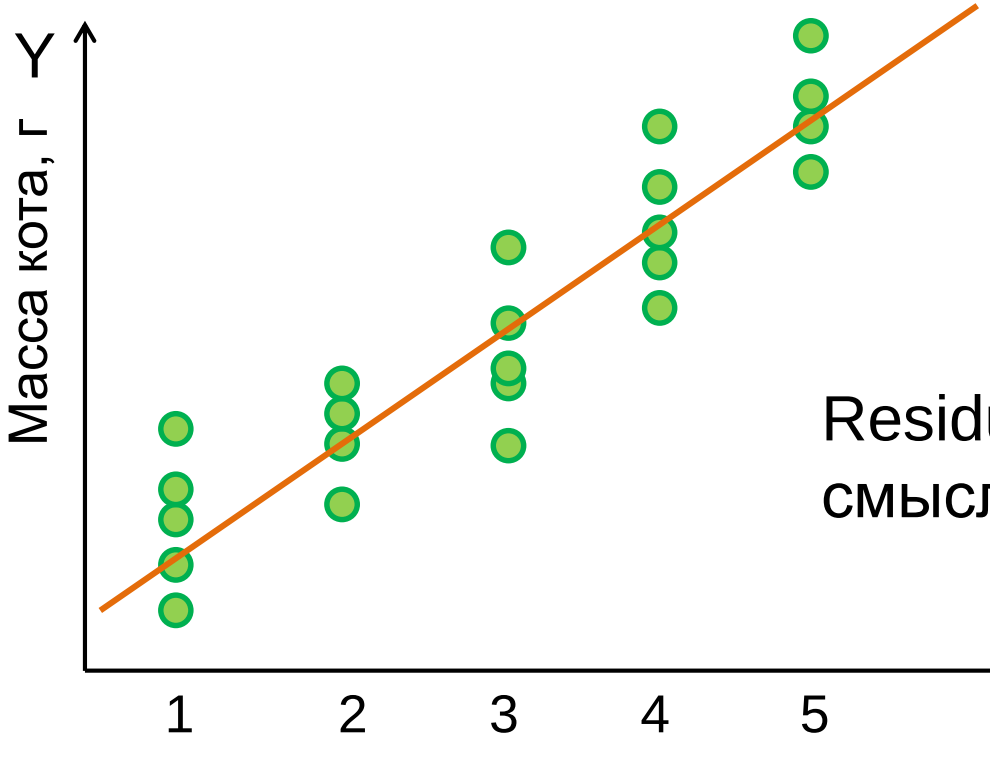
важно: нельзя пытаться предсказывать Y на основе значений X, лежащих за пределами размаха X в выборке.

Регрессии

Масса кота = 200 г + (500 г) x возраст кота в годах + ошибка

a

b



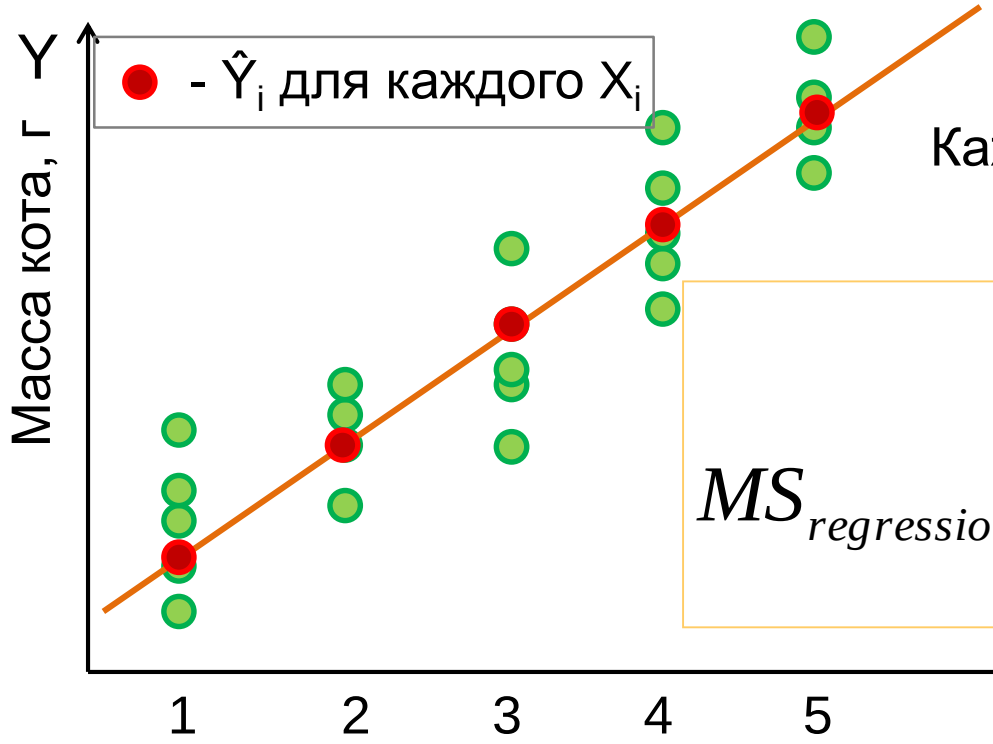
Residuals имеют тот же смысл, что и в ANOVA!

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**.

Регрессии ANOVA в регрессионном анализе.

Мы создали модель (наше уравнение регрессии), и теперь хотим оценить её качество: какую часть изменчивости зависимой переменной она объясняет.
Подход – разделение изменчивости (partitioning of variation).

«Объяснённая» изменчивость – изменчивость предсказанных значений (для каждого X_i это точка на прямой).

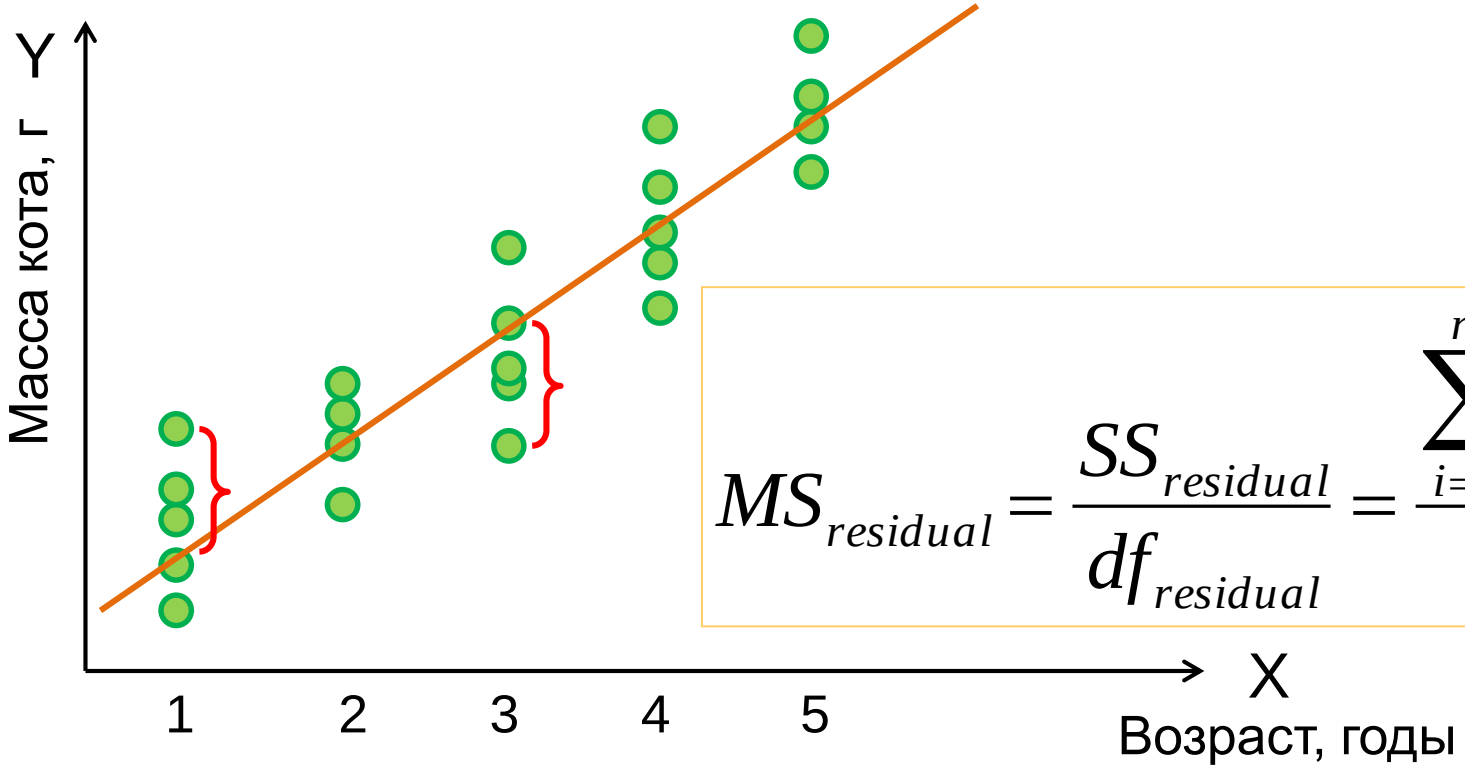


Каждое X – как маленькая группа.

$$MS_{regression} = \frac{SS_{regression}}{df_{regression}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{1}$$

Регрессии

«Остаточная» изменчивость – доля изменчивости Y, которая не объясняется связью с X, измеряется отклонениями Y_i от $\hat{Y}_i$ - residuals.

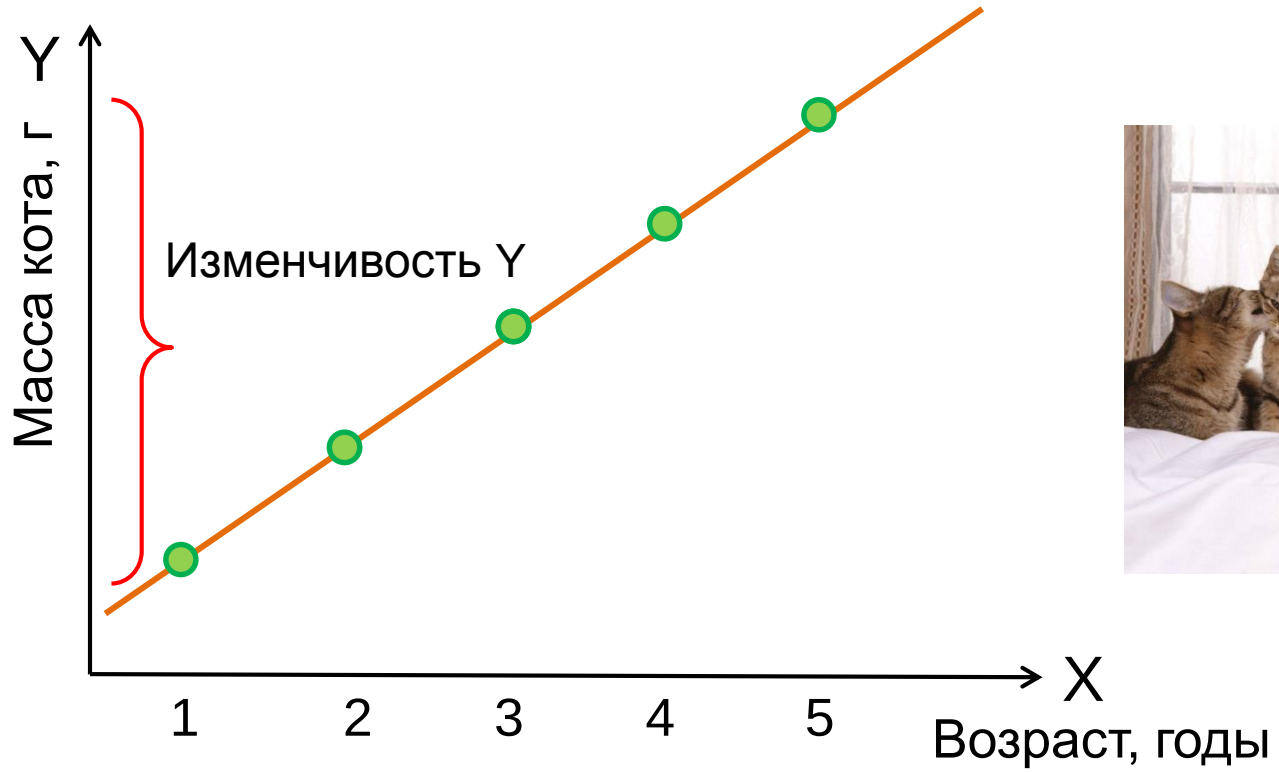


$$MS_{residual} = \frac{SS_{residual}}{df_{residual}} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

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

Регрессии

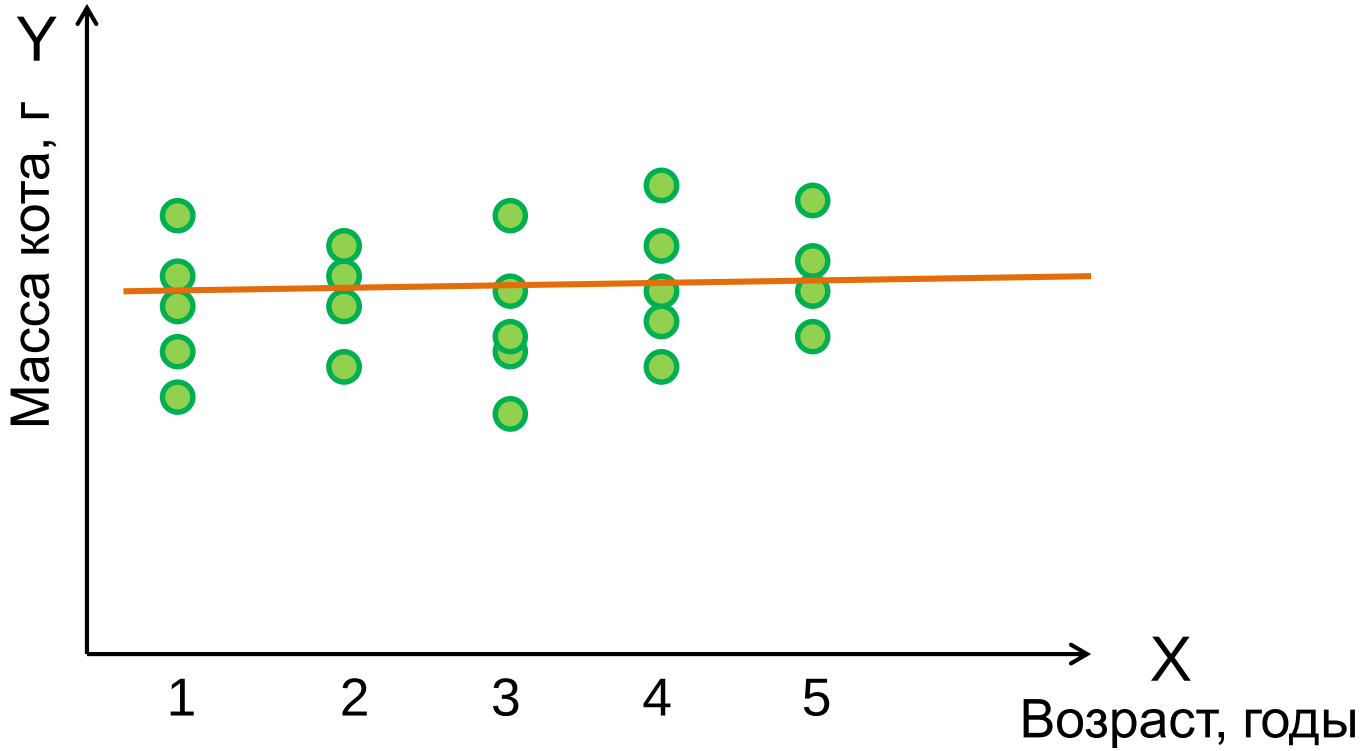
Экстремальный вариант, когда в модели нет остаточной изменчивости:



Все коты каждый год прибавляют в весе 500 грамм и ни граммом меньше, и возрастные группы отличаются ровно на 500 грамм

Регрессии

Теперь в модели только остаточная изменчивость:



Масса котов с возрастом не меняется вообще, хотя коты – ровесники различаются по массе.

Регрессии

Тестирование гипотез о коэффициентах регрессии

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

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

Идея в том, что если H_0 верна, вся изменчивость Y = остаточная изменчивость. Считают F и сравнивают с критическим значением.

Эту же гипотезу можно протестировать с помощью t-статистики:

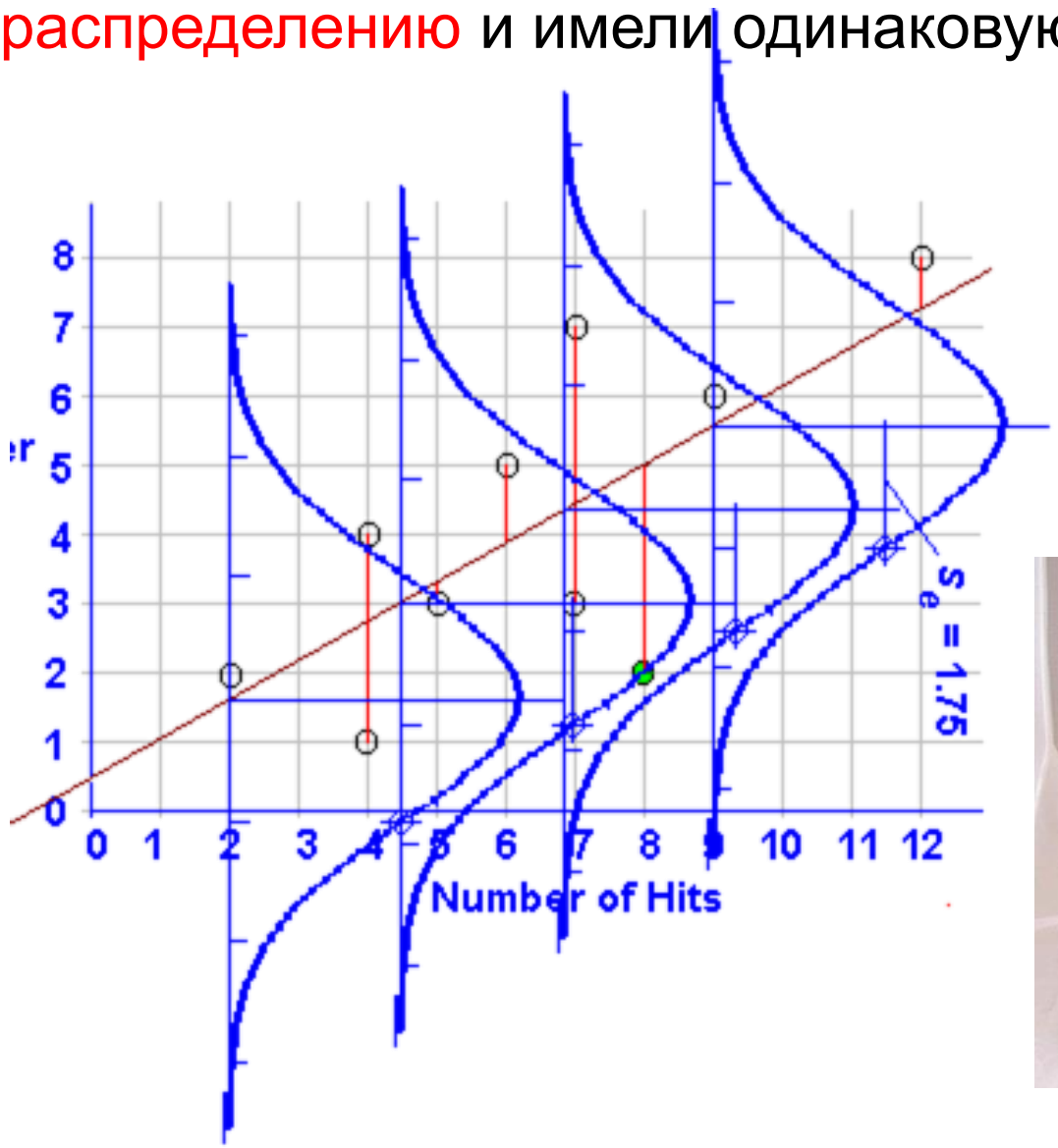
$$t = \frac{b - \beta_0}{s_b} = \frac{b}{s_b}$$

Причём $t^2 = F$

Можно проверять гипотезы и о α.

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 тоже будет отвергнута.

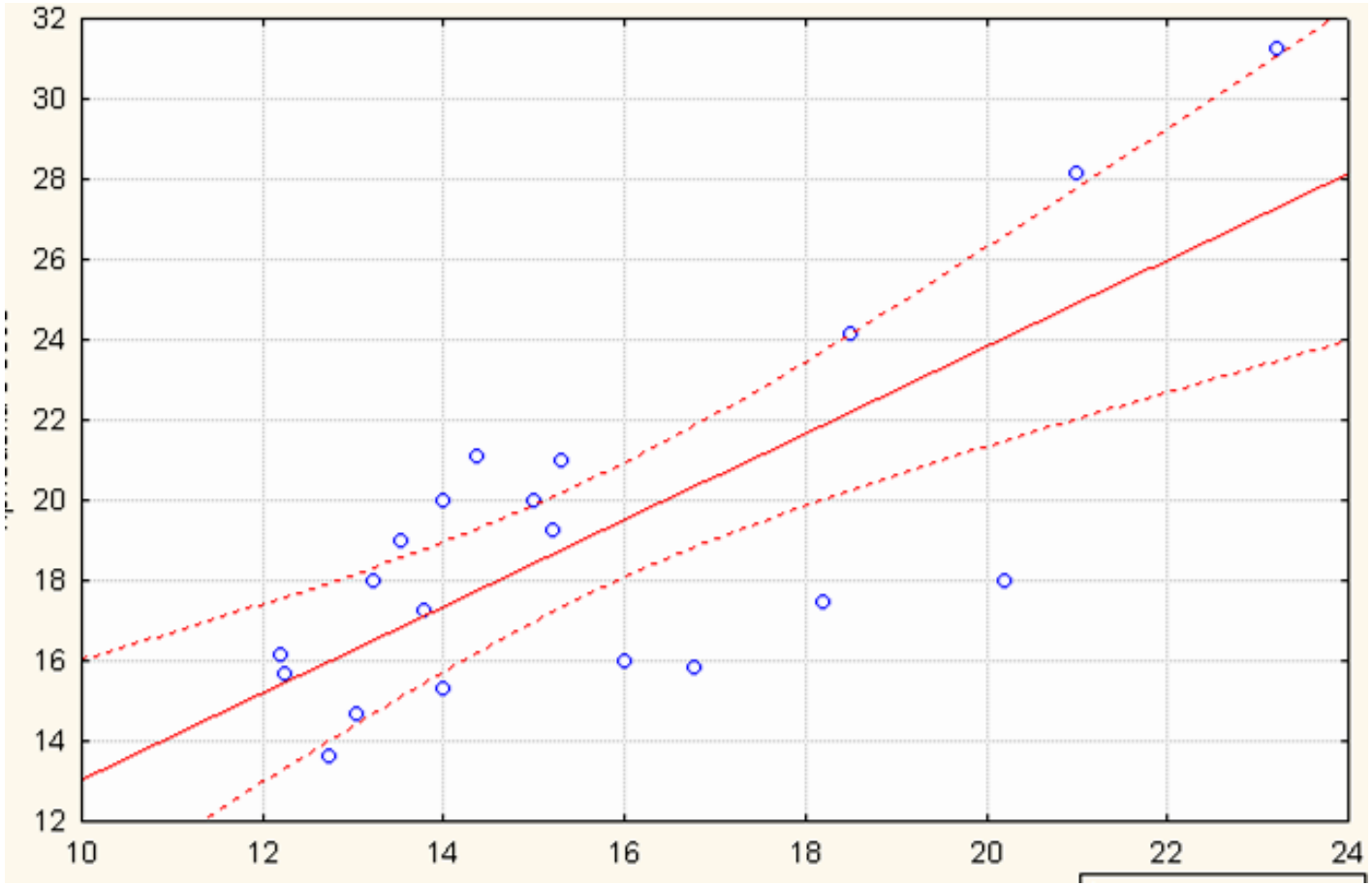
Поскольку в тестировании гипотез мы оцениваем параметры распределений, необходимо, чтобы для любого значения X_i значения Y и их residuals соответствовали нормальному распределению и имели одинаковую дисперсию



В противном случае регрессионный анализ невозможен.

Регрессии

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



Регрессии

Оценка качества модели (fit) и сравнение моделей

1. Качество модели определяется долей «объяснённой» изменчивости.
2. можно оценить качество модели методом максимального правдоподобия (maximum likelihood, ML): для нашей конфигурации данных компьютер оценивает, какие коэффициенты будут наиболее «правдоподобны», считает «правдоподобие» того, что в популяции действительно эти коэффициенты, потом считает «правдоподобие» для альтернативной модели и считает отношение «правдоподобий» (likelihood ratio) для этих двух моделей.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$y_i = \beta_0 + \varepsilon_i$$

К примеру, если мы сравним эти две модели, мы оценим, насколько важен β в модели = есть ли линейная связь.

Именно этот подход сравнения моделей используется в сложных моделях.

Регрессии

Коэффициент детерминации

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

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

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

Он показывает точность предсказания зависимой переменной и силу связи между переменными.

Беда в том, что он зависит от числа переменных в модели и от размерности данных, поэтому для сравнения разных моделей используют adjusted R^2

Assumptions

1. Для любого значения X_i Y должна иметь нормальное распределение, и residuals тоже должны быть распределены нормально.
Проверка: общая проверка распределения переменной на **нормальность**; построение **гистограммы остатков**.

Решение проблем: трансформация данных, использование ранговых корреляций.

Не компенсируется размером выборки.

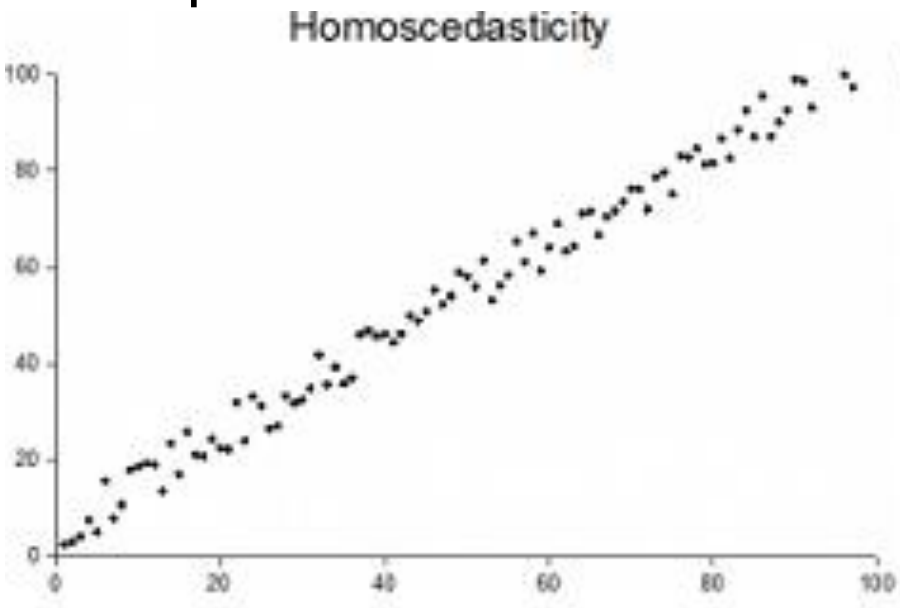
Регрессии

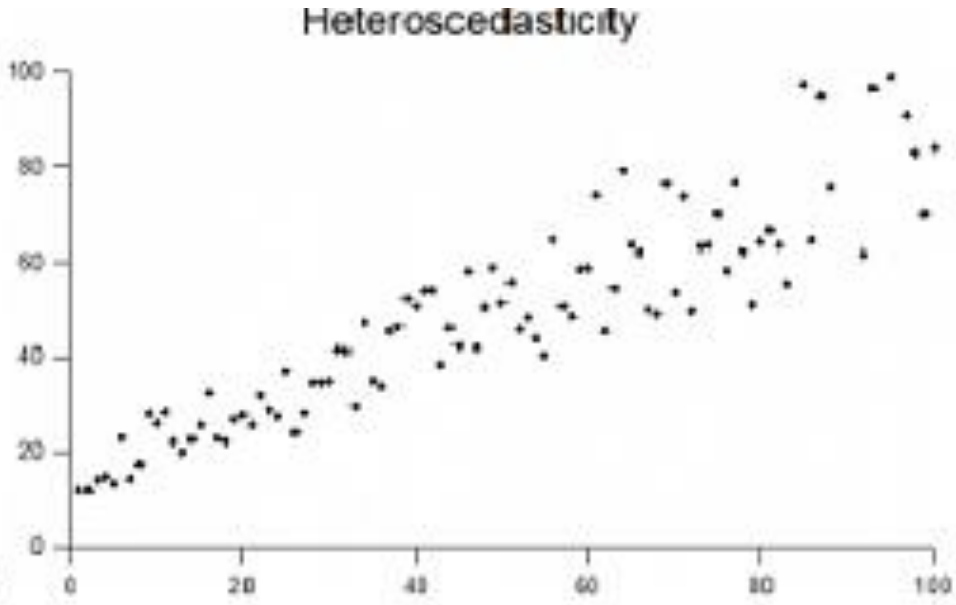
2. Гомогенность дисперсии: для любого значения X_i Y должны иметь одинаковую дисперсию.

Это очень важное условие!

Проверка: анализ скаттерплота исходных переменных («эллипс» не должен расширяться или сужаться); скаттерплот «остатки» - predicted ($\hat{Y}_i$)

Решение проблем: трансформация данных, использование ранговых корреляций. Часто негомогенность связана с «ненормальностью»





Регрессии

3. Ожидаемая зависимость переменной Y от X должна быть линейной.
Проверка: АНАЛИЗ СКАТТЕРПЛОТА исходных переменных; скаттерплот «остатки - predicted ($\hat{Y}_i$) ».
Решение проблем: трансформация данных, использование ранговых корреляций, анализ нелинейной регрессии.

4. Для любого значения X_i Y должны быть независимы друг от друга (тест Durbin-Watson – диагностика автокорреляций).

5. Размер выборки – от 15-20.

Регрессии

6. Отсутствие аутлаеров.

Проверка: анализ скаттерплота исходных переменных; скаттерплот «остатки» - predicted ($\hat{Y}_i$); Cook's distances, Mahalanobis distances.

Cook's distances: показывает, насколько данное наблюдение влияет на результат. >1 – повод для беспокойства.

Попробовать исключить аутлаер: если сильно изменились коэффициенты, результаты просто так докладывать нельзя.

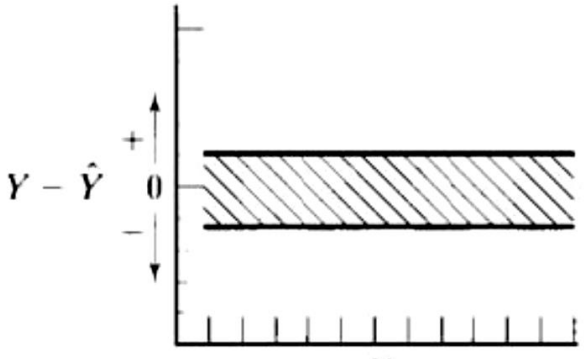
Решение проблем: трансформация данных, использование ранговых корреляций.

Трансформация данных часто решает все проблемы, но нужно быть внимательным при интерпретации коэффициентов.

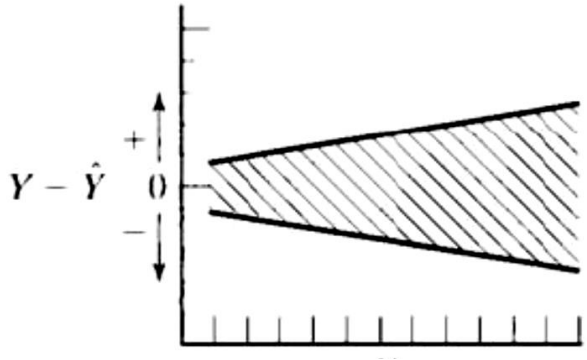
Регрессии

«Анализ остатков» (residual analysis)

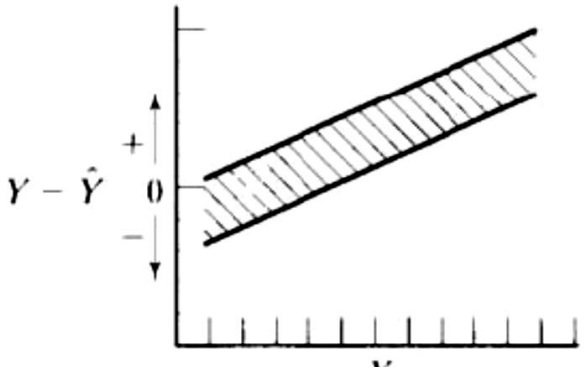
1. Позволяет оценить корректность анализа (assumptions) и линейность связи (по картинке);
2. Можно использовать «остатки» как новую переменную.



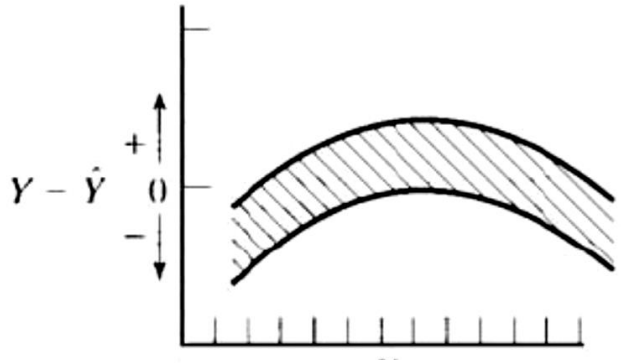
Как должно быть



Нелинейность - негомогенность



негомогенность

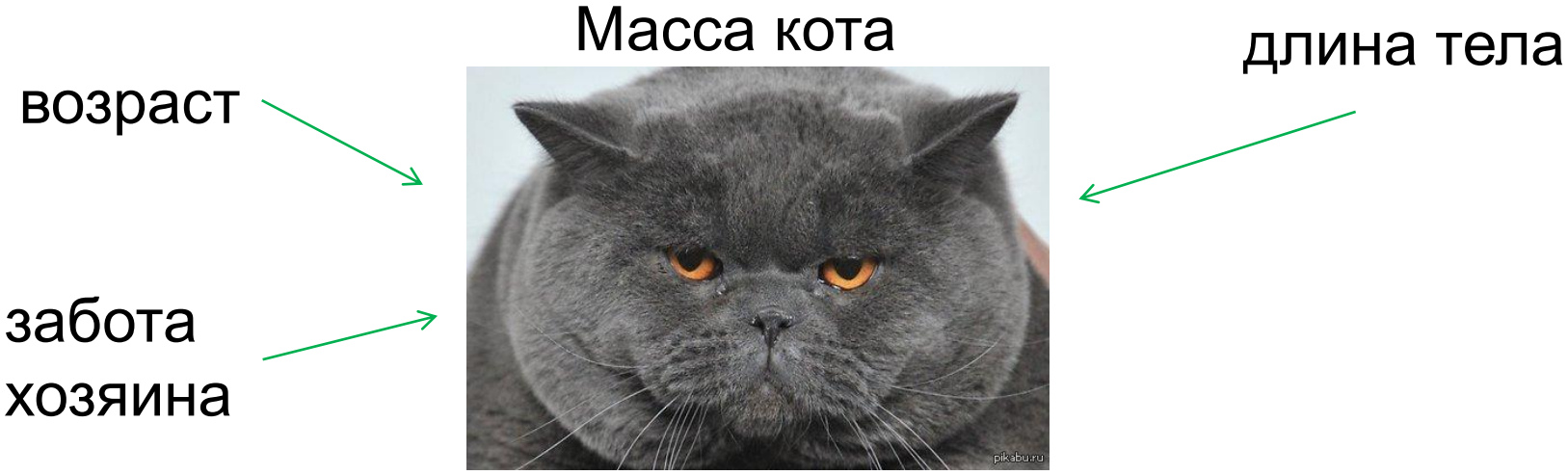


нелинейность

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.



Регрессии

возраст

длина тела

забота хозяина

Масса кота

Зависимая переменная – Y

Независимые (predictors) – X_1, X_2, X_3

Как и в простой регрессии, строим линейную модель, ищем коэффициенты.

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

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

На основе выборки строим «гиперплоскость» предсказанных значений:

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

Регрессии

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

Коэффициенты β_i: это коэффициенты ЧАСТИЧНОЙ регрессии (partial regression slopes) – показывают, на сколько изменяется Y на единицу изменения X_i, при условии, что все остальные X «зафиксированы» - остаются постоянными.

Это не то же самое, что коэффициенты простой регрессии – в них учитывается присутствие в модели остальных переменных.

Важный момент: если предикторы измерены в разных шкалах, для того, чтобы их вес в модели не зависел от единиц измерения, лучше их стандартизировать.

Регрессии

Поиск коэффициентов методом наименьших квадратов не так прост, как в простой регрессии.

Он предполагает операции над матрицами.

К примеру, вот матрица значений зависимых переменных для 4-х переменных:

$$X = \begin{pmatrix} X_{11} & X_{12} & X_{13} & X_{14} \\ X_{21} & X_{22} & X_{23} & X_{24} \\ X_{31} & X_{32} & X_{33} & X_{34} \\ X_{41} & X_{42} & X_{43} & X_{44} \end{pmatrix}$$

Эти манипуляции над матрицами не дадут корректного результата, если какие-нибудь переменные коррелируют между собой: чем выше корреляция, тем меньше точность оценки коэффициентов.

Так что корреляций между предикторами (~r>0.8) следует избегать!

Регрессии

Как и в простой регрессии: **ANOVA** используется для разделения изменчивости зависимой переменной на составляющие – $SS_{regression} + SS_{residual} = SS_{total}$

На основе MS тестируются гипотезы о равенстве коэффициентов нулю:

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

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

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

Доля объяснённой изменчивости:

Коэффициент детерминации (coefficient of determination)

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

Тот же принцип, что и для простой регрессии; показывает, какую долю изменчивости зависимой переменной объясняет модель, т.е., совместное влияние всех независимых переменных.

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

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

Регрессии

Оценка влияния КАЖДОЙ переменной:

Тестируем гипотезы о равенстве нулю КАЖДОГО коэффициента β!

Процедура такая же, как и в простой регрессии: считаем статистику t, и узнаем наконец, что же наиболее важно коту для набора веса.

Assumptions

Почти как в простой регрессии:

1. Нормальность
2. Гомогенность
3. Независимость измерений
4. Линейность взаимосвязей
5. Отсутствие аутлаеров

Диагностика – как в простой регрессии

6. Предикторы не слишком коррелируют между собой (collinearity)
7. Размер выборки должен не меньше, чем в 10 раз превосходить число переменных в анализе (лучше – в 20).

Необходимо строить **скаттерплоты**: между переменными, остатки – predicted и остатки – независимые переменные. Многие проблемы решает **трансформация**, и даже простая стандартизация.

Регрессии

Multicollinearity = ill-conditioning

Расчёт коэффициентов и статистик связан с операциями над матрицами. Если какие-то предикторы сильно коррелируют между собой, возникает принципиальная проблема в расчётах, коэффициенты регрессии не могут быть рассчитаны.

Признаки:

✓При удалении (добавлении) 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.

✓Высокие коэффициенты корреляции между предикторами;

✓ Tolerance < 0.1

Что делать:

✓Искать коррелирующие переменные и исключать одну из них из модели.

✓Использовать анализ главных компонент для уменьшения числа переменных.

Регрессии Выбор «лучших» независимых переменных

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

Есть несколько способов сделать это!

1. Partial regression – просто сравнение величин коэффициентов.
2. Используют пошаговые модели:

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

3. Сравнение моделей на основе максимального правдоподобия и информационных критериев.

В статьях приводят:

1. Общие характеристики модели: R^2; перечень предикторов;
2. Вклад отдельных переменных: коэффициенты β в стандартном виде ± их ошибки; значения t и p, соответственно;

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.

ANCOVA

Модель, когда исследуется действие группирующей переменной с поправкой на действие непрерывных предикторов на непрерывную зависимую переменную

Пример: мы анализируем влияние типа пищи (группирующая независимая) на массу тигров (непрерывная зависимая) с поправкой на размер тела (непрерывная независимая.

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

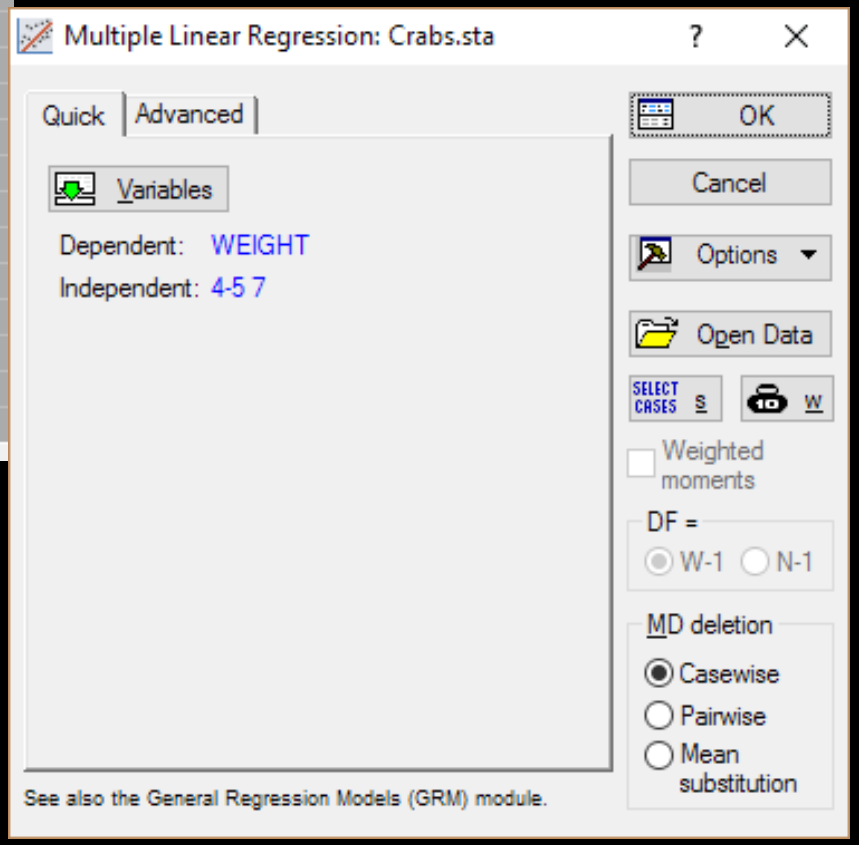
Регрессии

Multiple linear regression

rabs.sta (7v by 173c)

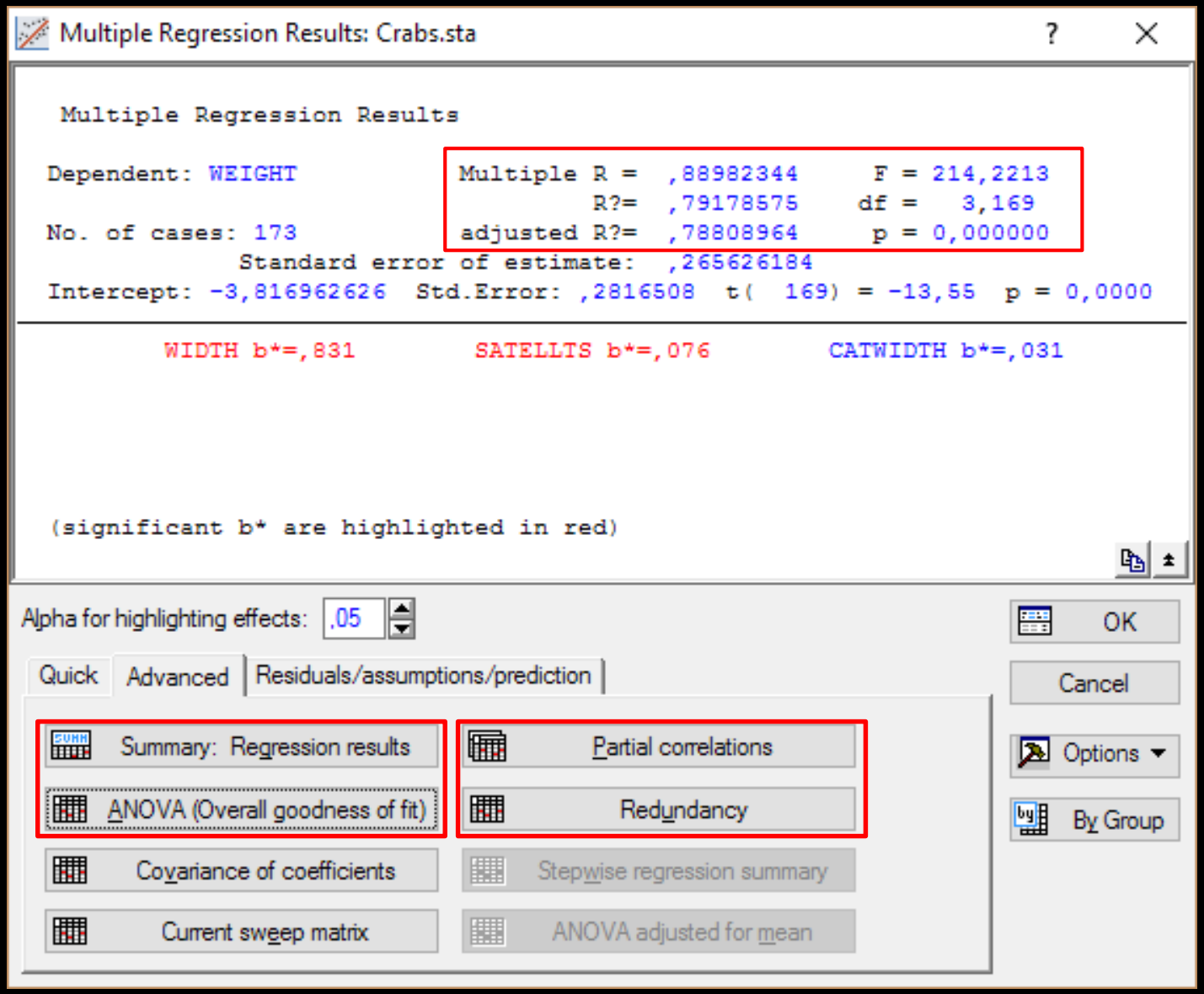
Number of crab satellites by female's color, spine condition, width, and weight

	1 Y	2 COLOR	3 SPINE	4 WIDTH	5 SATELLTS	6 WEIGHT	7 CATWIDTH
1	1	medium	bothworn	28,3	8	3,05	28,75
2	0	darkmed	bothworn	22,5	0	1,55	22,75
3	1	lightmed	bothgood	26,0	9	2,30	25,75
4	0	darkmed	bothworn	24,8	0	2,10	24,75
5	1	darkmed	bothworn	26,0	4	2,60	25,75
6	0	medium	bothworn	23,8	0	2,10	23,75
7	0	lightmed	bothgood	26,5	0	2,35	26,75
8	0	darkmed	oneworn	24,7	0	1,90	24,75
9	0	medium	bothgood	23,7	0	1,95	23,75
10	0	darkmed	bothworn	25,6	0	2,15	25,75
11	0	darkmed	bothworn	24,3	0	2,15	24,75
12	0	medium	bothworn	25,8	0	2,65	25,75
13	1	medium	bothworn	28,2	11	3,05	27,75
14	0	dark	oneworn	21,0	0	1,85	22,75
15	1	medium	bothgood	26,0	14	2,30	25,75
16	1	lightmed	bothgood	27,1	8	2,95	26,75
17	1	medium	bothworn	25,2	1	2,00	24,75
18	1	medium	bothworn	29,0	1	3,00	28,75



Файл с крабами:
зависимость массы от
ширины, ширины клешни
и сателлитов

Регрессии



Влияние двух переменных достоверно, третьей - нет

Регрессии

ression Summary for Dependent Variable: WEIGHT (Crabs.sta)

Regression Summary for Dependent Variable: WEIGHT (Crabs.sta)
R= ,88982344 R?= ,79178575 Adjusted R?= ,78808964
F(3,169)=214,22 p<0,0000 Std.Error of estimate: ,26563

N=173	b*	Std.Err. of b*	b	Std.Err. of b	t(169)	p-value
Intercept			-3,81696	0,281651	-13,5521	0,000000
WIDTH	0,831072	0,158027	0,22738	0,043235	5,2591	0,000000
SATELLTS	0,076468	0,037336	0,01402	0,006843	2,0481	0,042098
CATWIDTH	0,030575	0,157857	0,00890	0,045976	0,1937	0,846653

Коэффициенты наклона в стандартной форме (для стандартизированных переменных)

Коэффициенты a и b

lysis of Variance; DV: WEIGHT (Crabs.sta)

Analysis of Variance; DV: WEIGHT (Crabs.sta)

Effect	Sums of Squares	df	Mean Squares	F	p-value
Regress.	45,34461	3	15,11487	214,2213	0,00
Residual	11,92418	169	0,07056		
Total	57,26879				

undancy of Independent Variables; DV: WEIGHT (Crabs.sta)

Redundancy of Independent Variables; DV: WEIGHT (Crabs
R-square column contains R-square of respective variable with all other independent variables

Variable	Toleran.	R-square	Partial Cor.	Semipart Cor.
WIDTH	0,049336	0,950664	0,375019	0,184595
SATELLTS	0,883817	0,116183	0,155626	0,071889
CATWIDTH	0,049442	0,950558	0,014897	0,006799

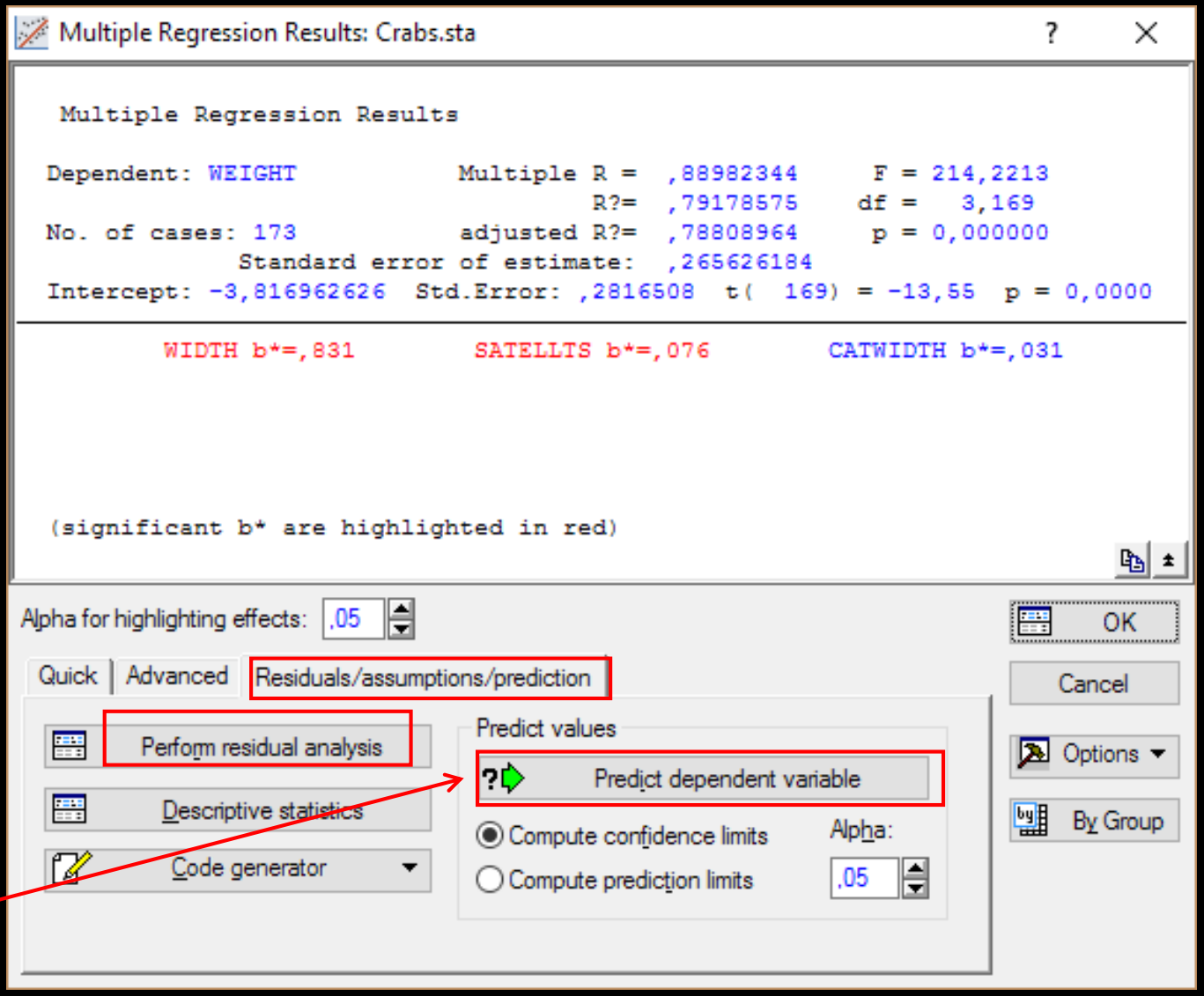
Диагностика коллинеарности

Регрессии

Анализ остатков

Часто «остатки» используют как самостоятельную переменную

Предсказание зависимой переменной



Регрессии

Диагностика модели

Residual Analysis: Crabs.sta

```
Dependent: WEIGHT      Multiple R :  ,88982344     F = 214,2213
                              R?:  ,79178575    df =   3,169
No. of cases: 173      adjusted R?:  ,78808964     p = 0,000000
        Standard error of estimate:   ,265626184
Intercept: -3,816962626  Std.Error: ,2816508  t(  169) = -13,55  p < 0,0000
```

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

- Summary: Residuals & predicted
- Descriptive statistics
- Regression summary
- Durbin-Watson statistic

Maximum number of rows (cases) in a single results Spreadsheet or Graph: 100000

Summary | Cancel | Options | By Group

cted & Residual Values

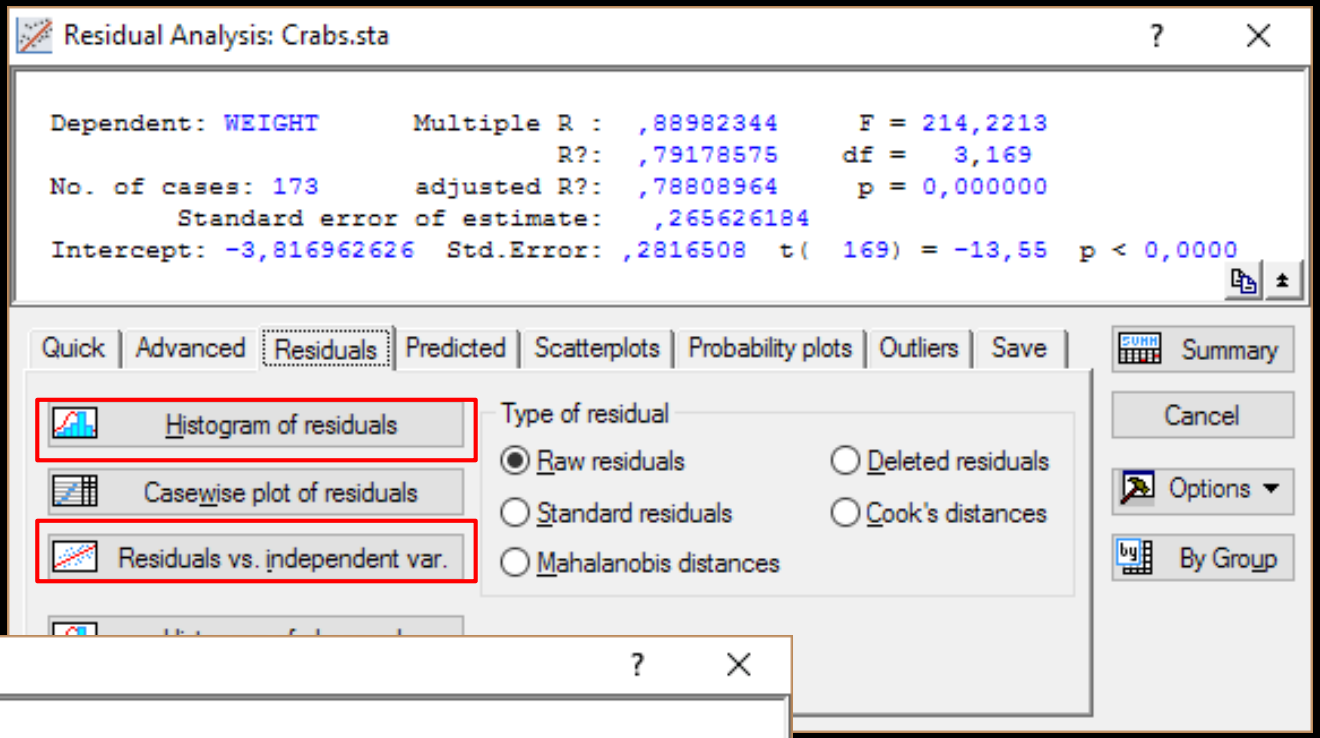
Predicted & Residual Values
WEIGHT

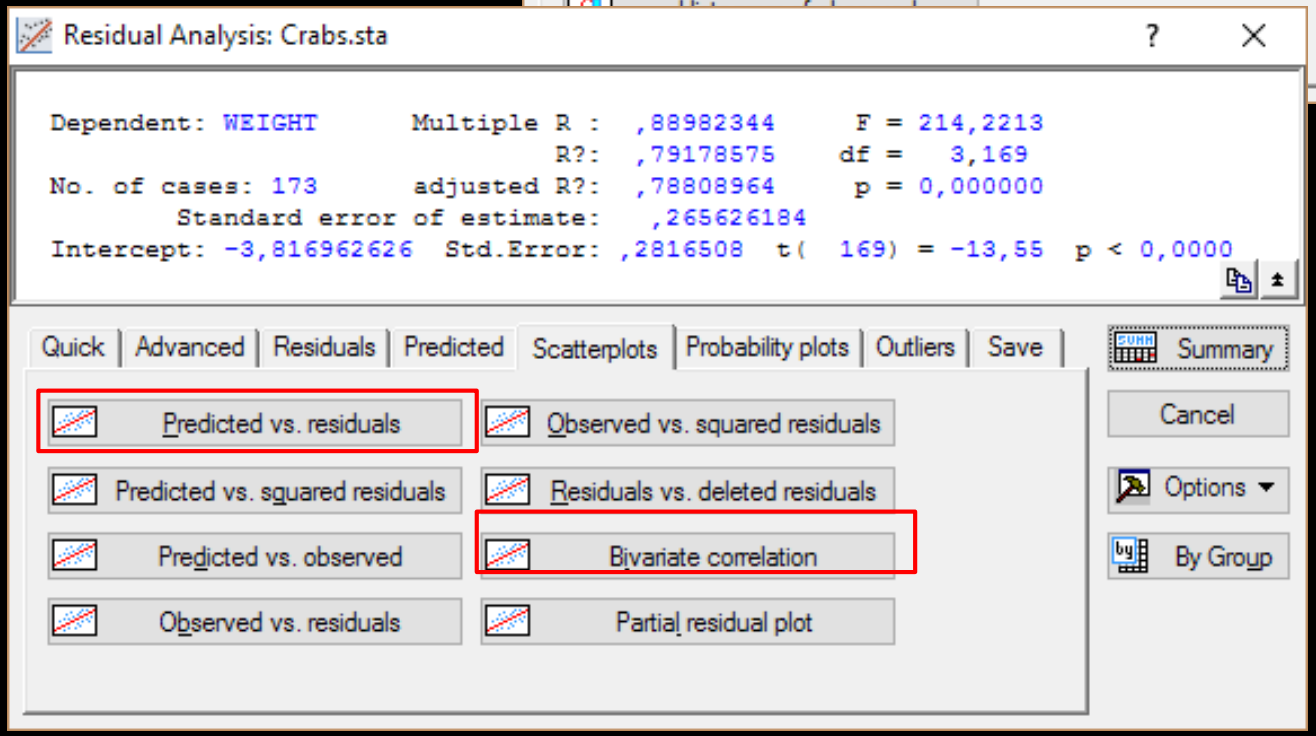
	Observed Value	Predicted Value	Residual	Standard Pred. v.	Standard Residual	Std.Err. Pred.Val	Mahalanobis Distance	Deleted Residual	Cook's Distance
1	3,050000	2,985908	0,064092	1,068686	0,241285	0,050187	5,145688	0,066464	0,000559
2	1,550000	1,501579	0,048421	-1,822204	0,182289	0,042248	3,356926	0,049677	0,000221
3	2,300000	2,450244	-0,150244	0,025423	-0,565623	0,048386	4,713122	-0,155401	0,002839
4	2,100000	2,042353	0,057647	-0,768988	0,217021	0,029174	1,080647	0,058350	0,000146
5	2,600000	2,380169	0,219831	-0,111056	0,827594	0,024005	0,410534	0,221641	0,001422
6	2,100000	1,806073	0,293927	-1,229170	1,106545	0,034334	1,879392	0,298921	0,005289
7	2,350000	2,446702	-0,096702	0,018524	-0,364053	0,033221	1,696087	-0,098239	0,000535

Predicted Residual Values

Регрессии

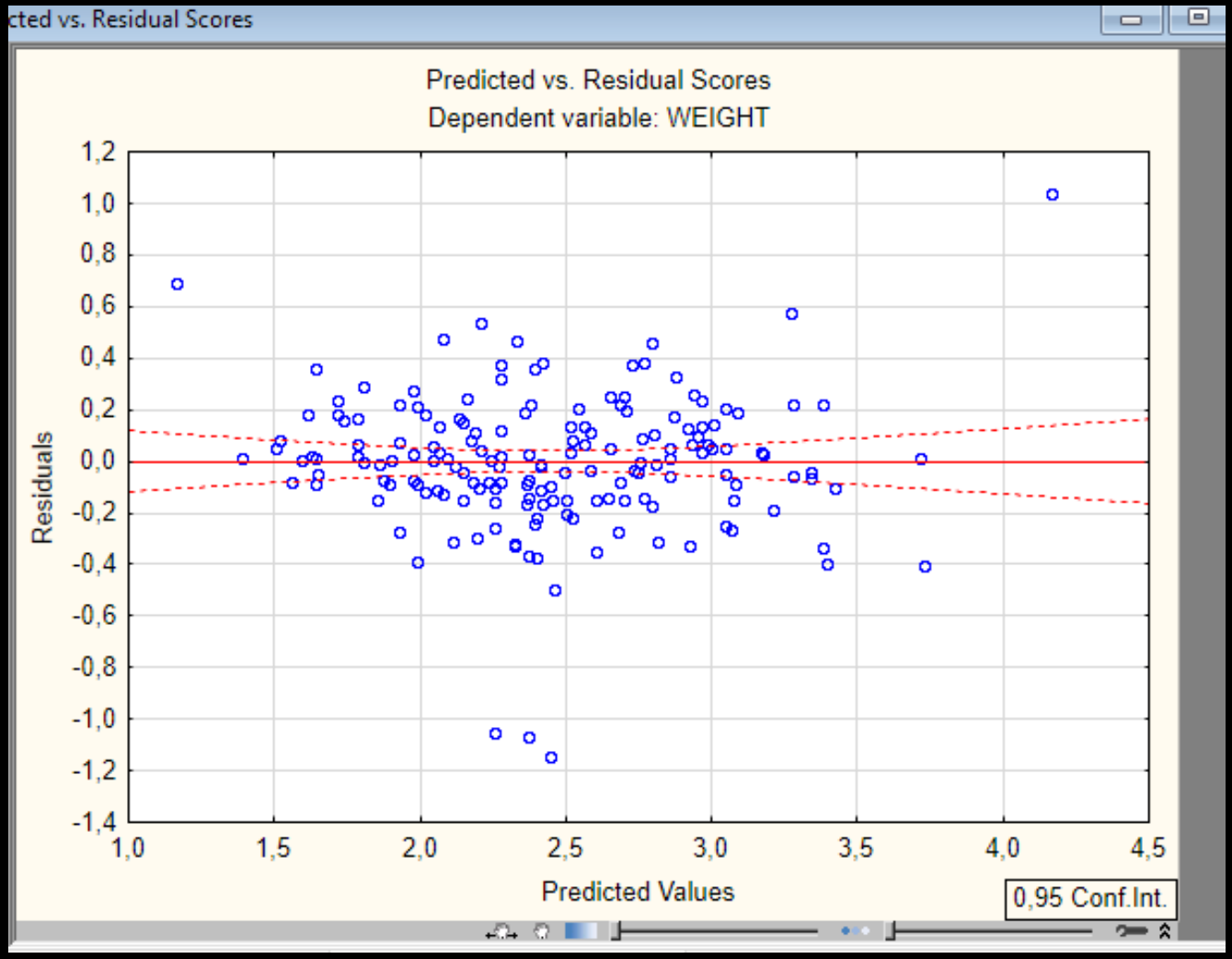
Диагностика модели





Регрессии

Диагностика модели



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

Регрессии

1. построить скаттерплот правой кнопкой и в Basic stats Crabs
2. Basic stat – корреляц матрица, коэф корреляции Crabs
3. Mult regeression Crabs Job_prof.sta
4. ANCOVA Crabs Ancova