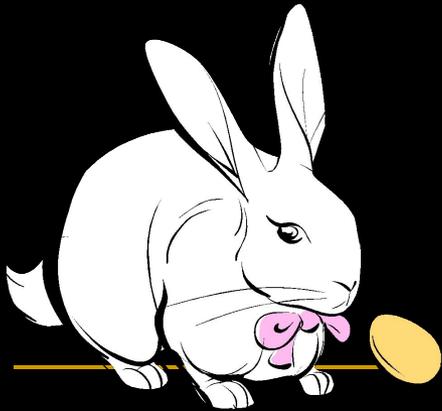


Регрессионный анализ



Вавилин К.С.

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.

Регрессионный анализ **предсказывает** значение одной переменной на основании другой.

Для этого в линейной регрессии строится прямая – **линия регрессии**.

Линейная регрессия:

Даёт нам правила, определяющие линию регрессии, которая лучше других предсказывает одну переменную на основании другой.

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}

То есть,

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

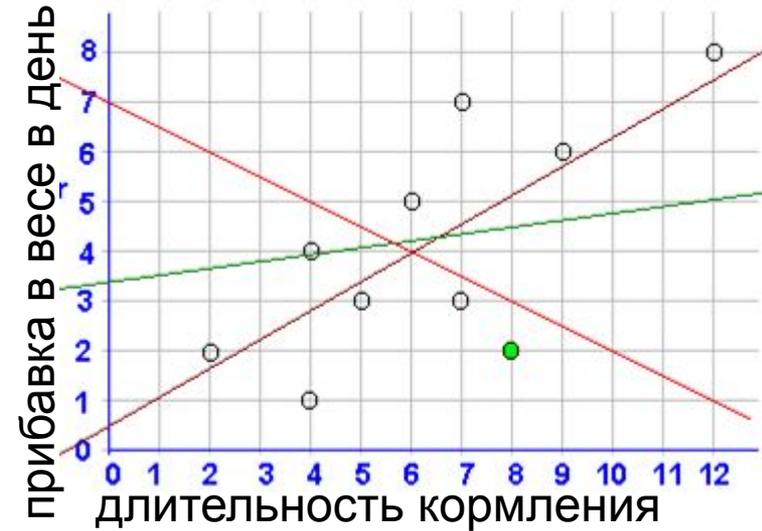
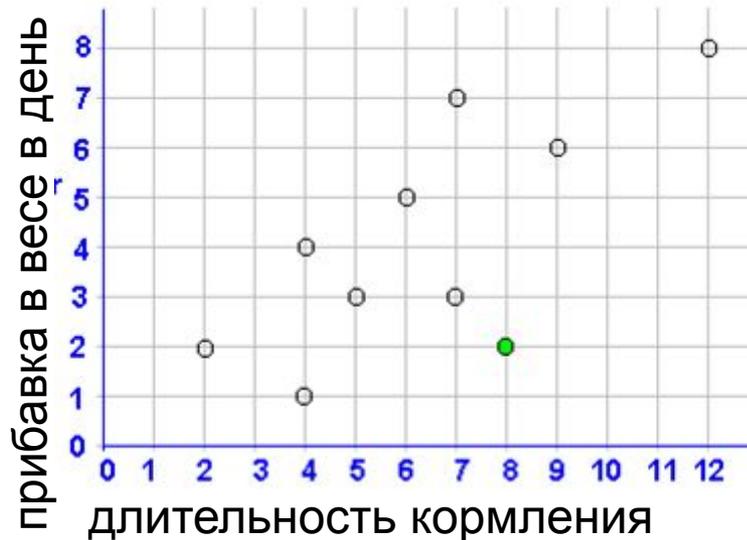
ЭТО НЕ ОДНО И ТО ЖЕ!

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 .



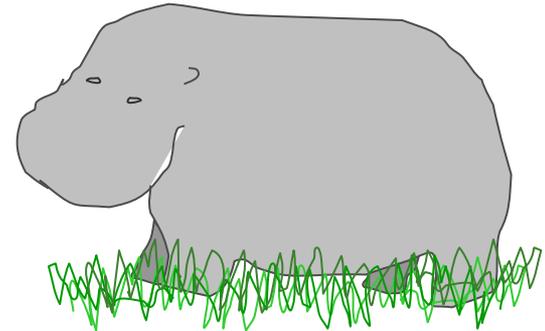
Простая линейная регрессия (*linear regression*)

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 .

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

коэффициент корреляции Пирсона!

стандартные отклонения для X и Y

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

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

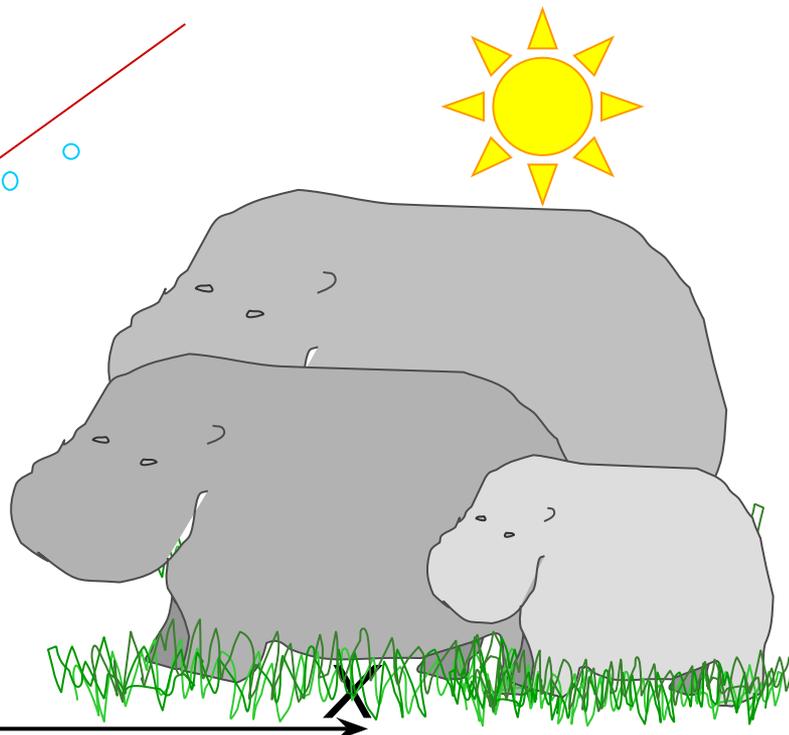
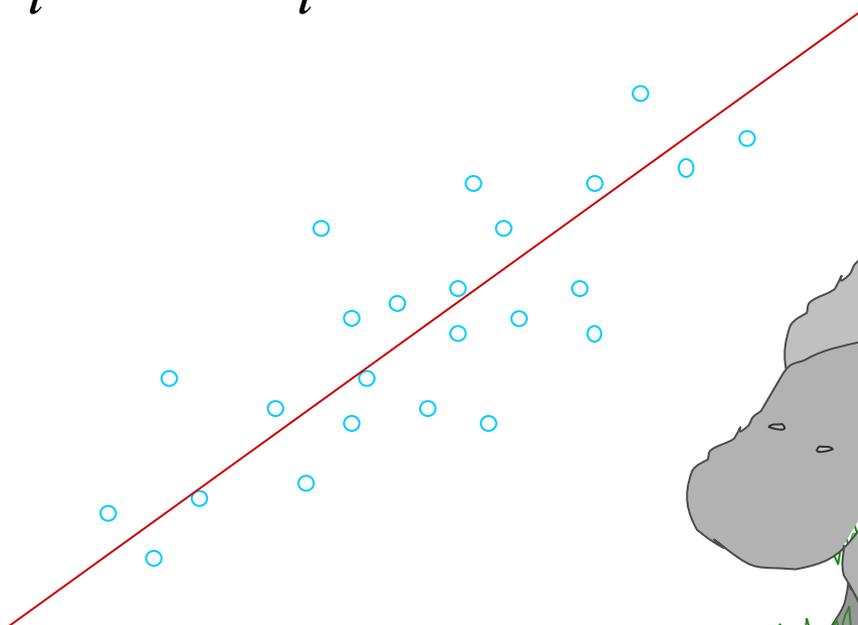
b – определяет, насколько изменится Y на единицу X ; имеет тот же знак, что и r .

Пример с кол-вом удобрения на каждый кг помидоров

Прибавка в весе в день

Y

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



Длительность кормления

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



Линия регрессии в стандартной форме

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

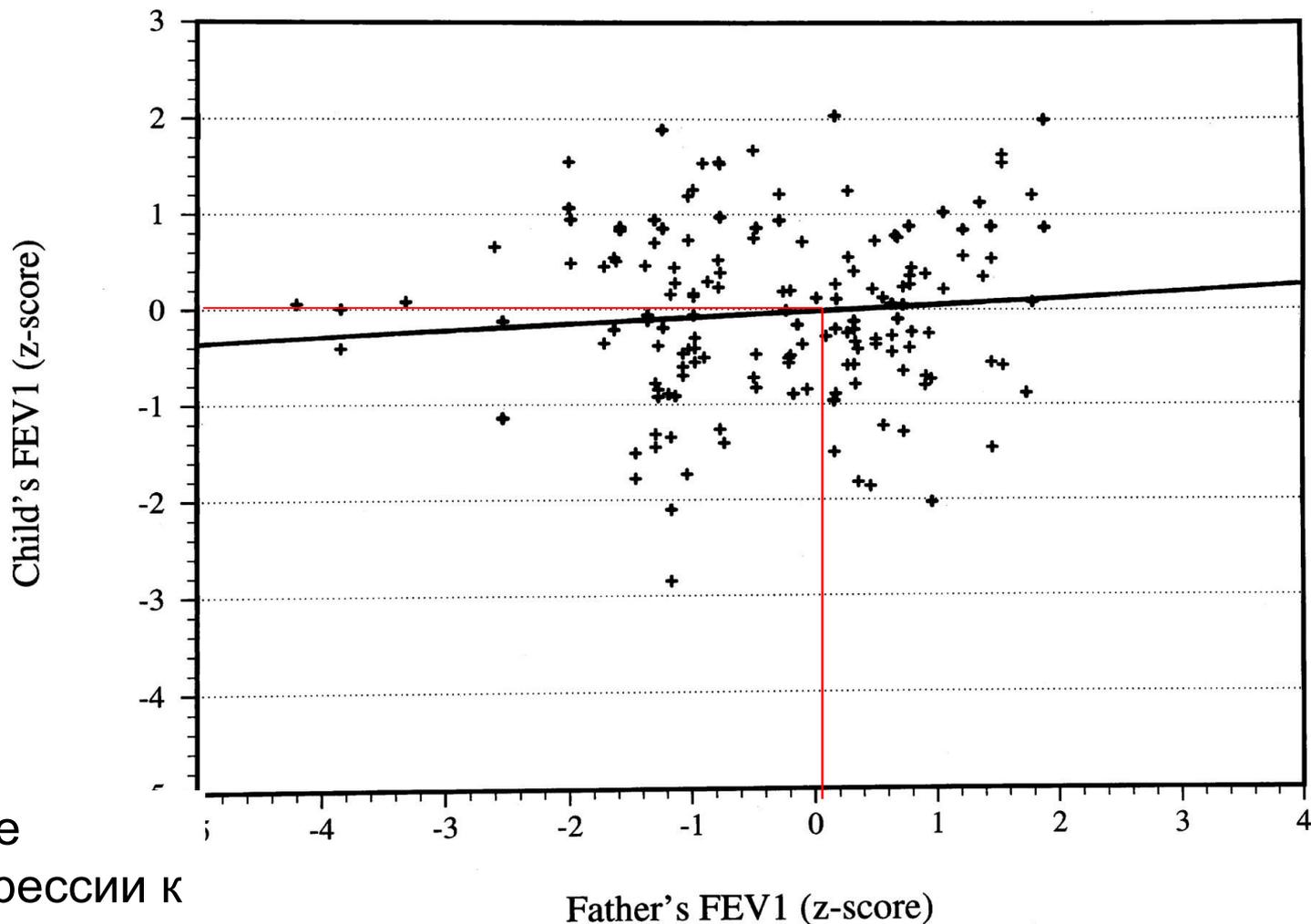
$$b = r \frac{s_X}{s_Y}$$



$$a = 0, b = r$$

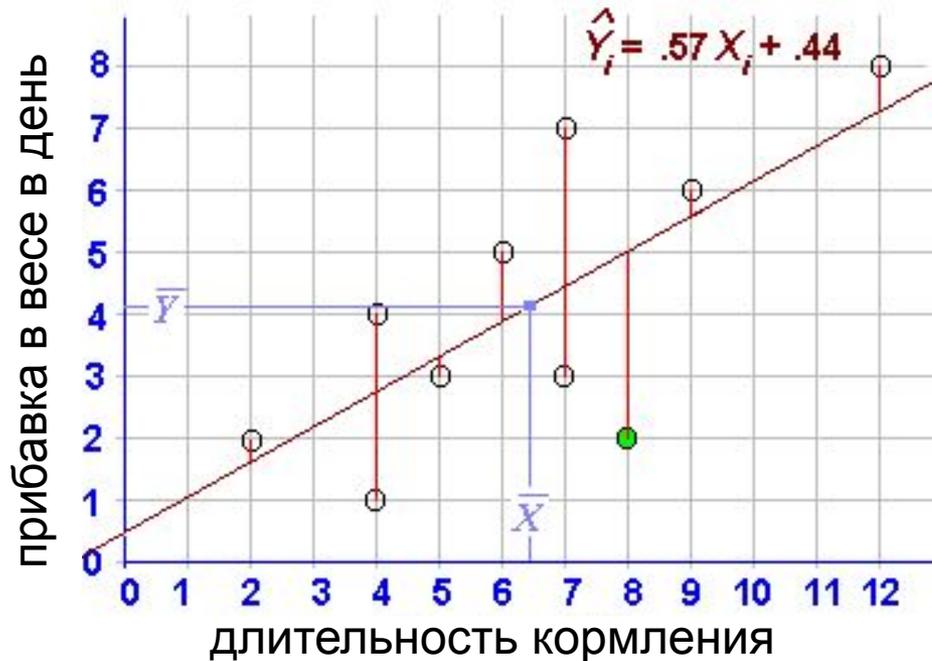
$$\hat{Y}_z = r\hat{X}_z$$

(математическое объяснение регрессии к среднему)



«Лучшая» линия регрессии

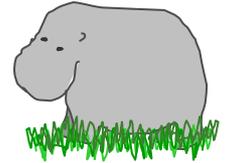
Ошибка предсказания (residual) = «остатки»



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

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

Как определить «лучшую» линию регрессии?

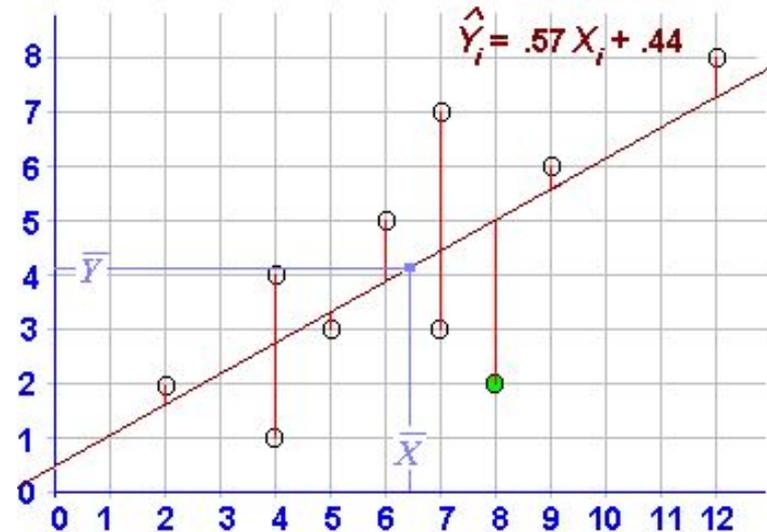
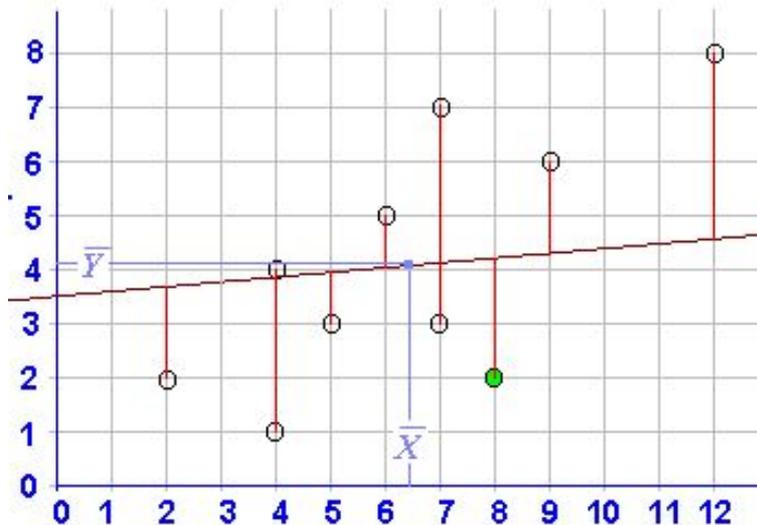


Метод наименьших квадратов:

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

$$\sum e_i = 0$$

$$\sum e_i^2 - \text{минимальна}$$



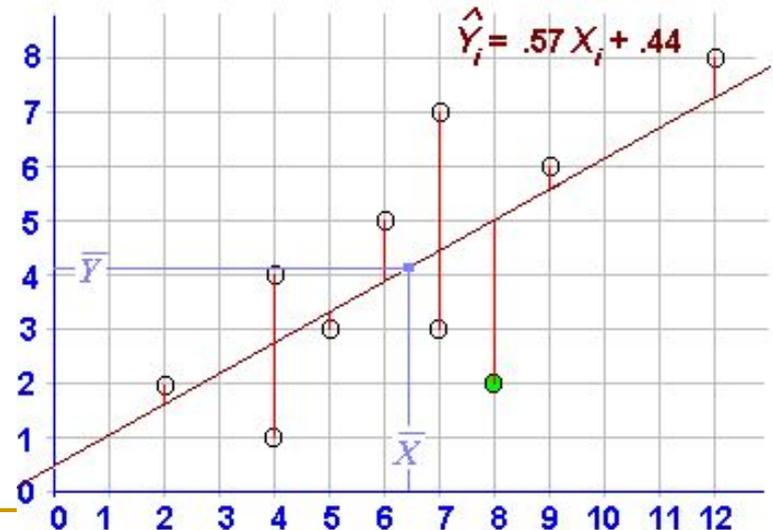
Насколько хорошо «лучшая» линия регрессии предсказывает Y?

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

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

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

зависит от квадрата коэффициента корреляции



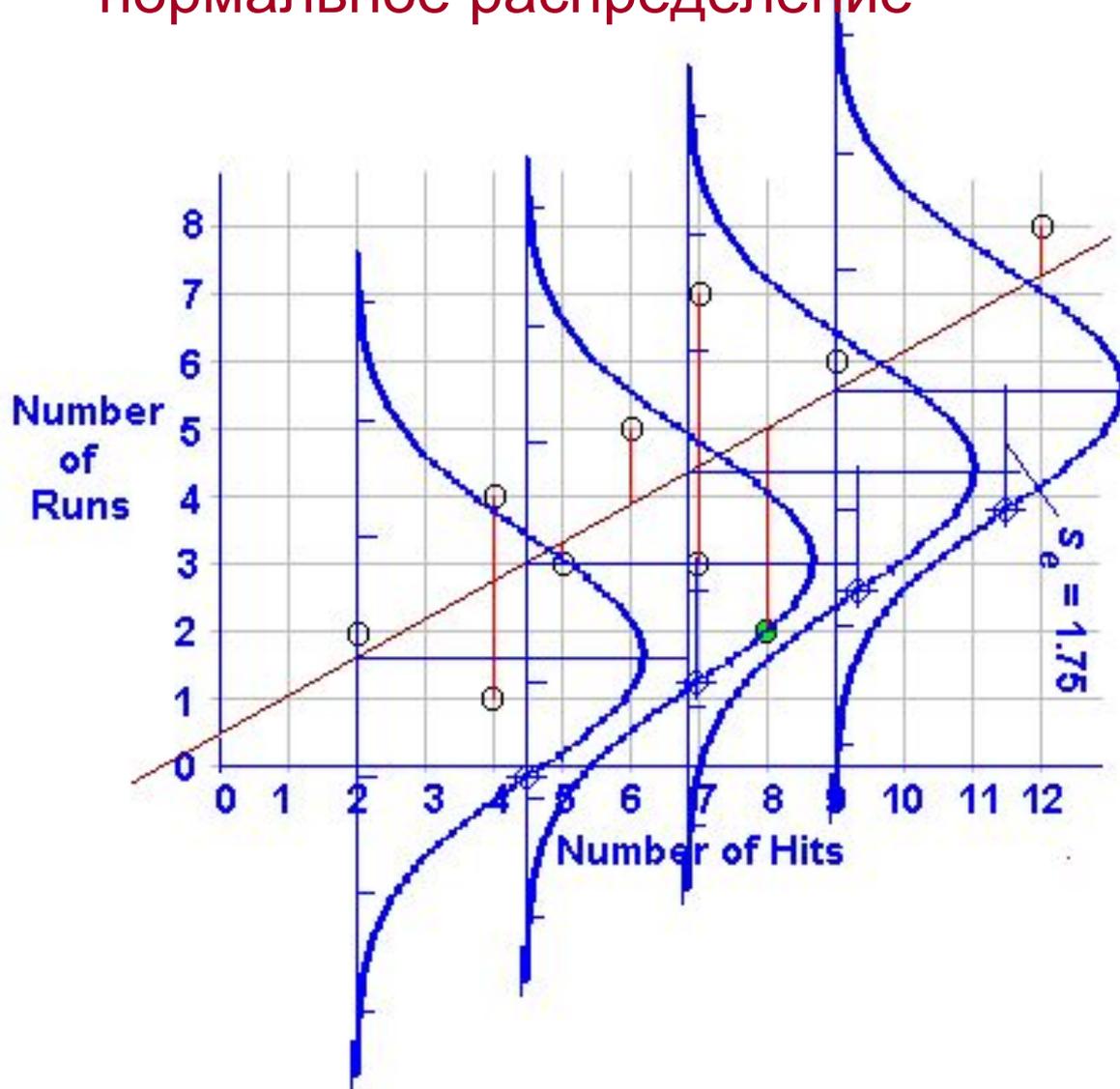
Чем *больше* коэффициент корреляции, тем *меньше* стандартное отклонение ошибки, и наоборот.

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 . Определяет, какую долю изменчивости зависимой переменной объясняет независимая переменная (т.е., насколько точно предсказание).

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

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



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**.
3. Для любого значения X_i выборки для Y должны иметь **одинаковую дисперсию** (homoscedasticity).
4. Для любого значения X_i выборки для Y должны быть **независимы** друг от друга.

Тестирование гипотезы в регрессионном анализе: отличен ли от нуля наклон линии регрессии?

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

$b_{\text{population}}$ часто обозначается как β , в т.ч. в Statistica

Если r достоверно отличается от нуля, то и $b \neq 0$!
То есть, если мы отвергаем H_0 о том, что $r=0$, то эта гипотеза отвергается автоматически.

linear regression

Data: бегемоты.sta (10v by 20c)

	1	2	3
	№ бегемота	длительность кормления	прибавка в весе
1	1	14,4	21,11
2	2	12,7	13,64
	3	20,2	18,00
	4	14,0	20,00
	5	13,8	17,27
	6	12,2	31,25
	7	16,8	15,83
	8	15,0	20,00
	9	18,2	17,50
	10	13,5	19,00
	11	12,2	16,15
	12	12,2	15,71
	13	13,2	18,00
	14	14,0	15,33
	15	15,3	21,00
	16	13,0	14,67
	17	15,6	24,17
	18	14,1	28,18
	19	16,0	16,00

Multiple Linear Regression: бегемоты

Quick | **Advanced**

Variables

Dependent: прибавка в весе
Independent: длительность кормления

Input file: Raw Data

Advanced options (stepwise or ridge regression)
 Review descriptive statistics, correlation matrix
 Extended precision computations
 Batch processing/reporting
 Print/report residual analysis

Specify all variables for the analysis; additional models (indep./dep. vars) can be specified later. For stepwise regression etc. check the advanced options check box.

See also the General Regression Models (GRM) module.

OK
Cancel
Options
Open Data
SELECT CASES
Weighted moments
DF =
W1 N1
MD deletion
 Casewise
 Pairwise
 Mean substitution

linear regression

У бегемотов прибавка в весе положительно зависела от длительности кормления

Multiple Regression Results

Dependent: прибавка в вес

No. of cases: 20

Standard error of estimate: 3,023493998

Intercept: 2,248207760 Std. Error: 3,557217 t(18) = ,63201 p = ,5353

Multiple R =	,75109618	F =	23,29818
R ² =	,56414547	df =	1,18
adjusted R ² =	,53993133	p =	,000135

длительность beta = ,751

(significant betas are highlighted)

Alpha for highlighting effects: .05

Quick | Advanced | Residuals/assumptions/prediction

- Summary: Regression results
- ANOVA (Overall goodness of fit)
- Covariance of coefficients
- Current sweep matrix
- Partial correlations
- Redundancy
- Stepwise regression summary
- ANOVA adjusted for mean

Buttons: OK, Cancel, Options

Summary for Dependent Variable: прибавка в весе (бегемоты)

Regression Summary for Dependent Variable: прибавка в весе (бегемоты)						
R= ,75109618 R ² = ,56414547 Adjusted R ² = ,53993133						
F(1,18)=23,298 p<,00014 Std. Error of estimate: 3,0235						
	Beta	Std. Err. of Beta	B	Std. Err. of B	t(18)	p-level
N=20						
Intercept			2,248208	3,557217	0,632013	0,535324
длительность кормления	0,751096	0,155609	1,079058	0,223555	4,826819	0,000135

Коэффициент наклона в стандартной форме

Коэффициенты а и b

Часто «остатки» используют как самостоятельную переменную

Multiple Regression Results

Dependent: прибавка в вес Multiple R = ,75109618 F = 23,29818
R²= ,56414547 df = 1,18
No. of cases: 20 adjusted R²= ,53993133 p = ,000135
Standard error of estimate: 3,023493998
Intercept: 2,248207760 Std. Error: 3,557217 t(18) = ,63201 p = ,5353

длительность beta=,751

(significant betas are highlighted)

Alpha for highlighting effects: .05

Quick | Advanced | **Residuals/assumptions/prediction**

Perform residual analysis
Descriptive statistics

Predict values
? Predict dependent variable
 Compute confidence limits Alpha: .05
 Compute prediction limits

OK Cancel Options

Residual Analysis: бегемоты

Dependent: прибавка в вес (multiple R : ,75109618 F = 23,29818
 R?: ,56414547 df = 1,18
 No. of cases: 20 adjusted R?: ,53993133 p = ,000135
 Standard error of estimate: 3,023493998
 Intercept: 2,248207760 Std. Error: 3,557217 t(18) = ,63201 p < ,5353

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

Summary: Residuals & predicted

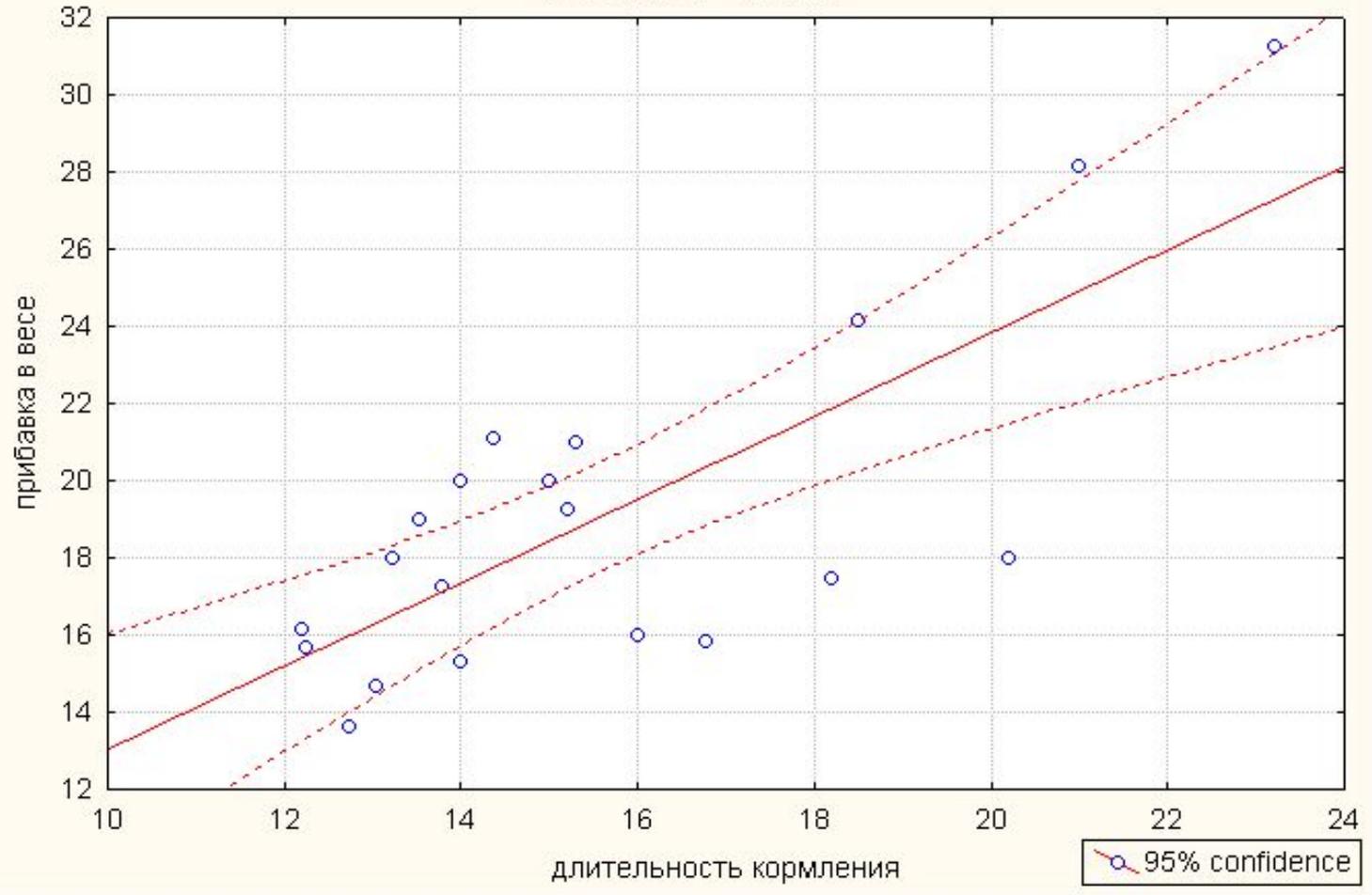
Descriptive statistics
 Regression summary
 Durbin-Watson statistic

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

Residual Values (бегемоты)

Case No.	Predicted & Residual Values (бегемоты)				
	Observed Value	Predicted Value	Residual	Standard Pred. v.	S
1	21,11111	17,74611	3,36500	-0,40596	
2	13,63636	15,99417	-2,35781	-0,92923	
3	18,00000	24,04059	-6,04059	1,47408	
4	20,00000	17,35502	2,64498	-0,52277	
5	17,27273	17,10871	0,16402	-0,59634	0,05425 0,792576 0,355622 0,1
6	31,25000	27,28236	3,96764	2,44233	1,31227 1,824011 5,964968 6,2
7	15,83333	20,33500	-4,50167	0,36729	-1,48890 0,722483 0,134902 -4,7
8	20,00000	18,43408	1,56592	-0,20048	0,51792 0,690227 0,040192 1,6
9	17,50000	21,88707	-4,38707	0,83086	-1,45099 0,888378 0,690330 -4,8
10	19,00000	16,85329	2,14671	-0,67263	0,71001 0,821434 0,452428 2,3
11	16,15385	15,40924	0,74461	-1,10394	0,24627 1,021481 1,218682 0,8
12	15,71429	15,46548	0,24881	-1,08714	0,08229 1,012775 1,181875 0,2
13	18,00000	16,51289	1,48711	-0,77430	0,49185 0,863444 0,599542 1,6
14	15,33333	17,33524	-2,00191	-0,52868	-0,66212 0,769125 0,279502 -2,1
15	21,00000	18,75780	2,24220	-0,10379	0,74159 0,679896 0,010772 2,3
16	14,66667	16,30658	-1,63991	-0,83592	-0,54239 0,890659 0,698763 -1,7
17	24,16667	22,21078	1,95588	0,92755	0,64689 0,933283 0,860347 2,1
18	28,18182	24,90843	3,27339	1,73328	1,08265 1,379321 3,004269 4,1
19	16,00000	19,51314	-3,51314	0,12182	-1,16195 0,681334 0,014839 -3,7
20	19,28572	18,64989	0,63582	-0,13602	0,21029 0,682625 0,018501 0,6
Minimum	13,63636	15,40924	-6,04059	-1,10394	-1,99788 0,679896 0,010772 -7,2
Maximum	31,25000	27,28236	3,96764	2,44233	1,31227 1,824011 5,964968 6,2
Mean	19,10529	19,10529	0,00000	0,00000	0,00000 0,915623 0,950000 0,0
Median	18,00000	18,09010	0,69022	-0,30322	0,22828 0,842439 0,525985 0,7

длительность кормления vs. прибавка в весе
прибавка в весе = $2,2482 + 1,0791 * \text{длительность кормления}$
Correlation: $r = ,75110$



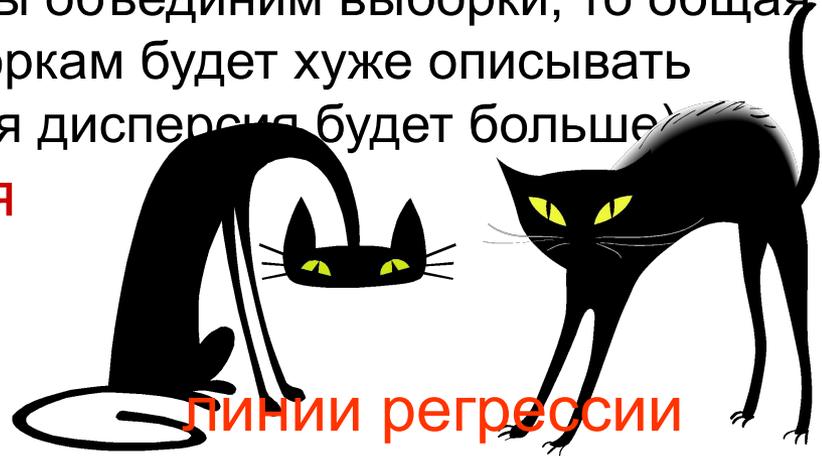
Сравнение двух линий регрессии

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



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

