

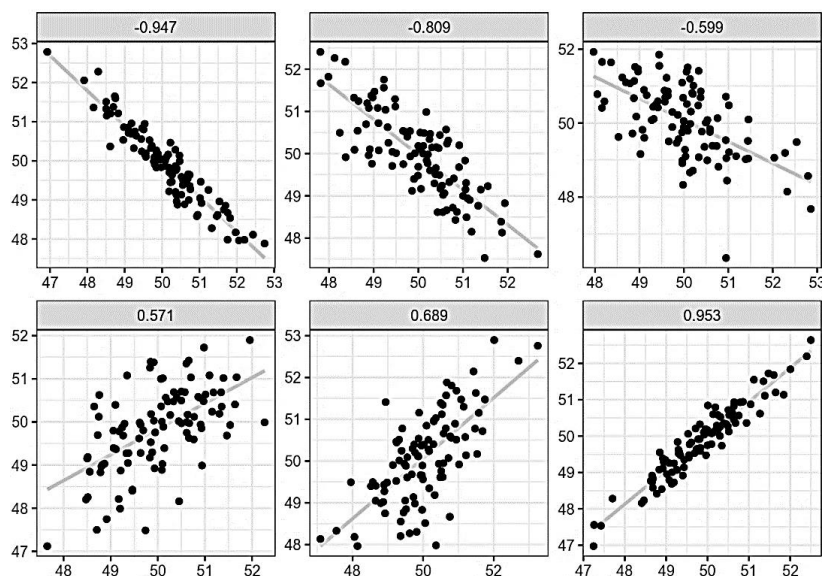
5. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ (ПАРНАЯ КОРРЕЛЯЦИЯ). РЕГРЕССИОННЫЙ АНАЛИЗ (ЛИНЕЙНАЯ РЕГРЕССИЯ)

1 Коэффициент корреляции

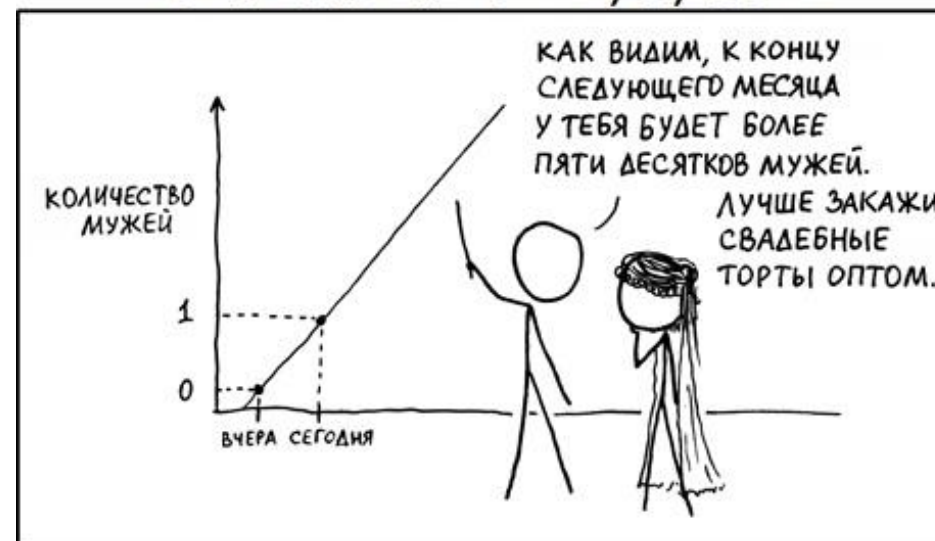
2 Ошибка коэффициента корреляции

3 Понятие о регрессии

4 Ошибки элементов уравнения прямойлинейной регрессии



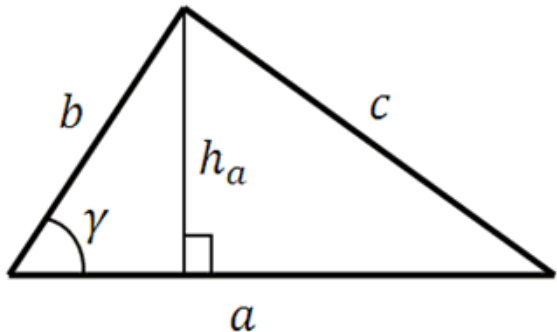
Аналитик и линейная регрессия



1. Коэффициент корреляции

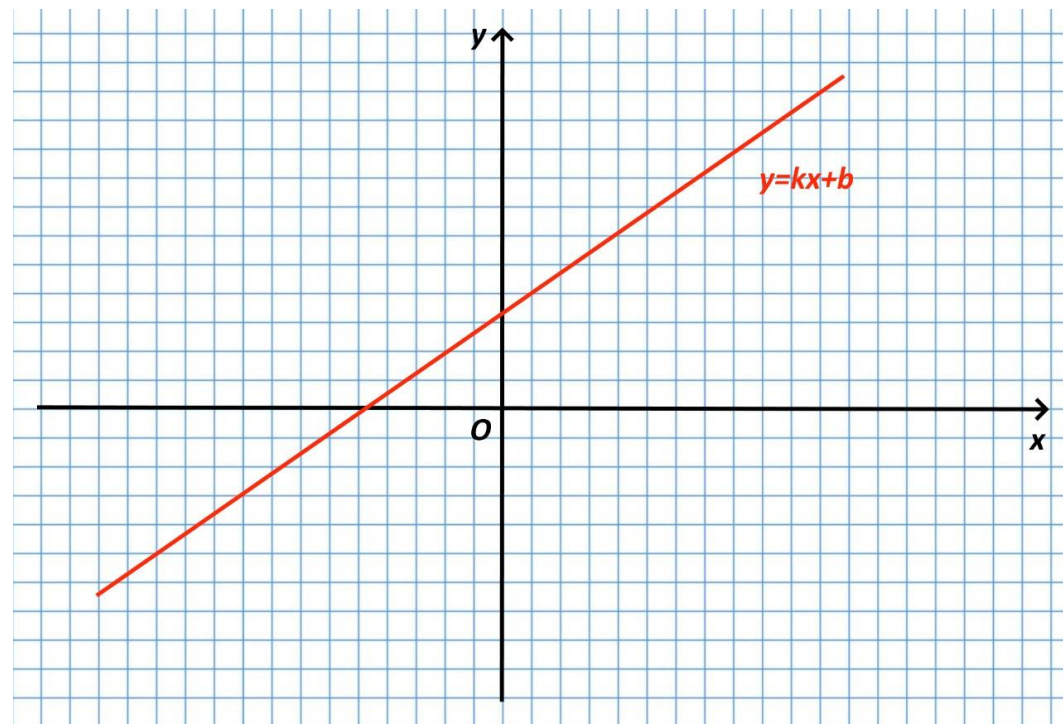
Связи между признаками

1) **Функциональные** (каждому значению первого признака всегда соответствует совершенно определенное, единственное значение второго признака).

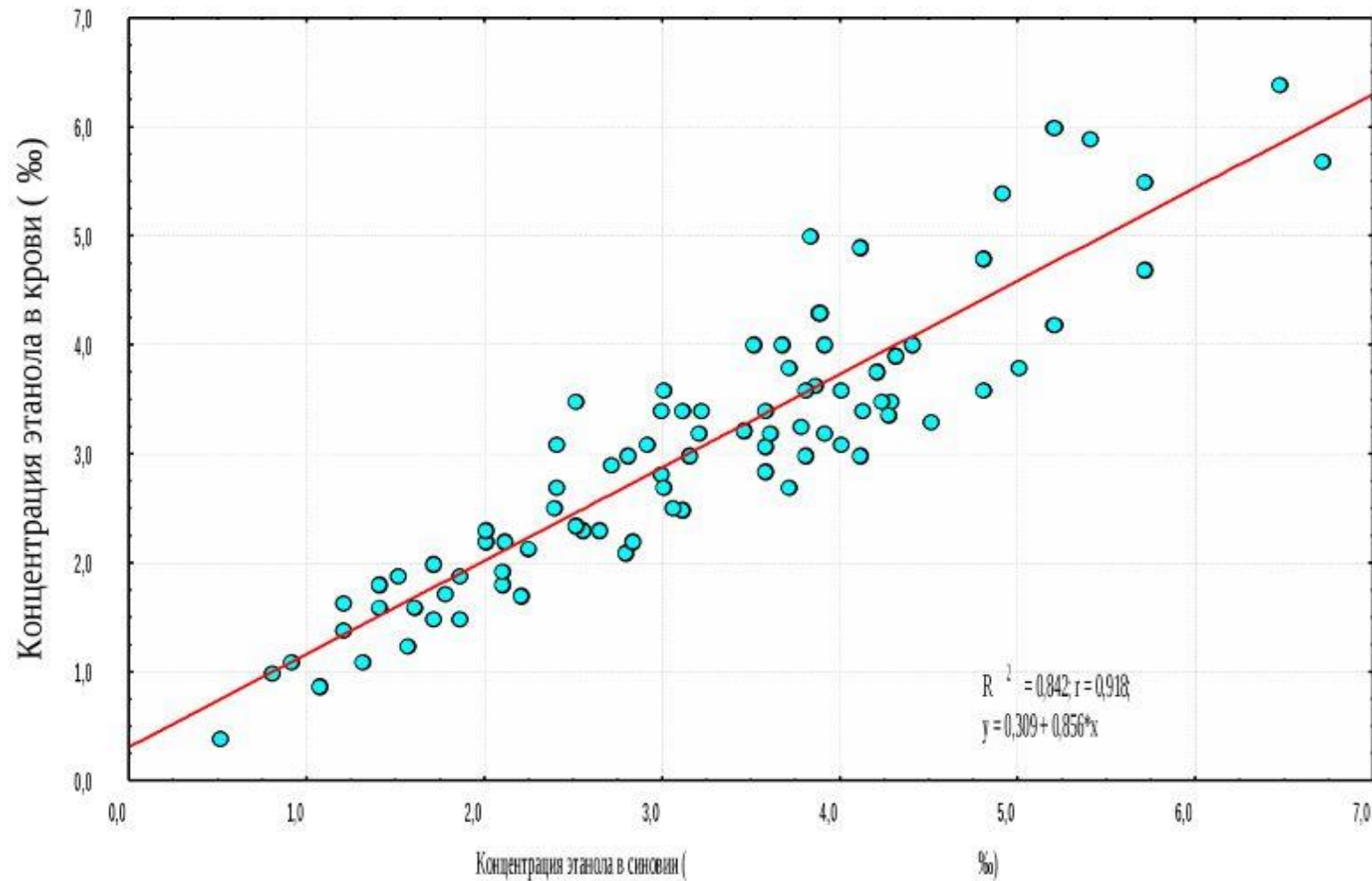


$$S = \frac{1}{2} \cdot a \cdot b \cdot \sin \gamma$$

Площадь треугольника равна половине произведения двух его соседних сторон на синус угла между ними



2) **Корреляционные, или корреляция** (каждому определенному значению первого признака соответствует не одно значение второго признака, а целое распределение этих значений – средней величины и степени разнообразия)



Корреляционные связи
определяются:

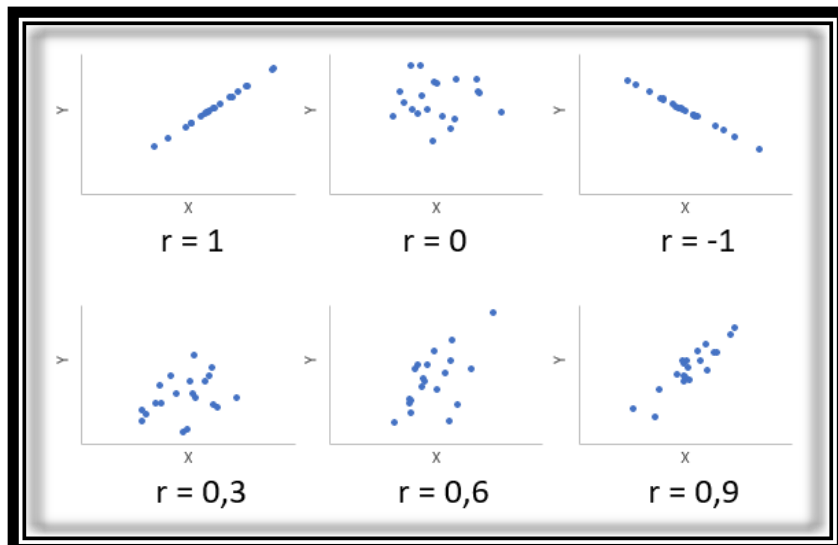
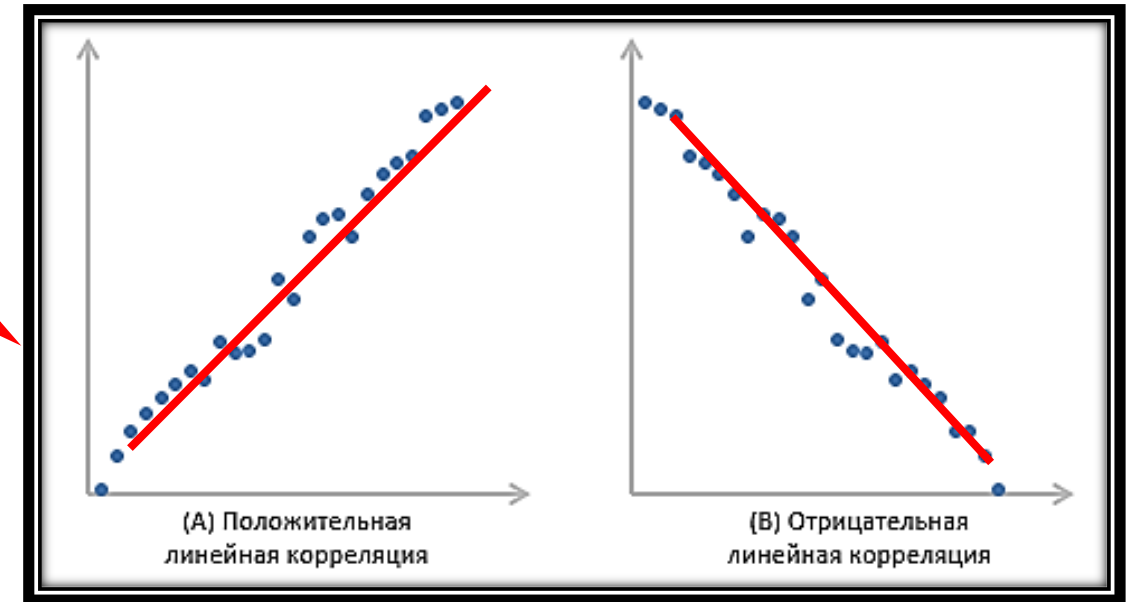
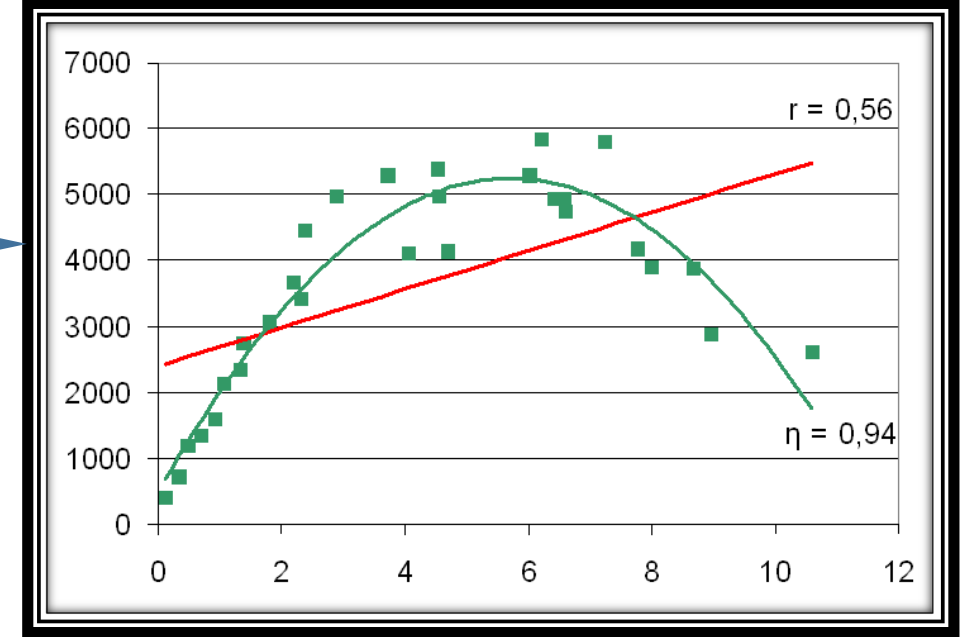
а) формой:

- прямолинейная;
- криволинейная;

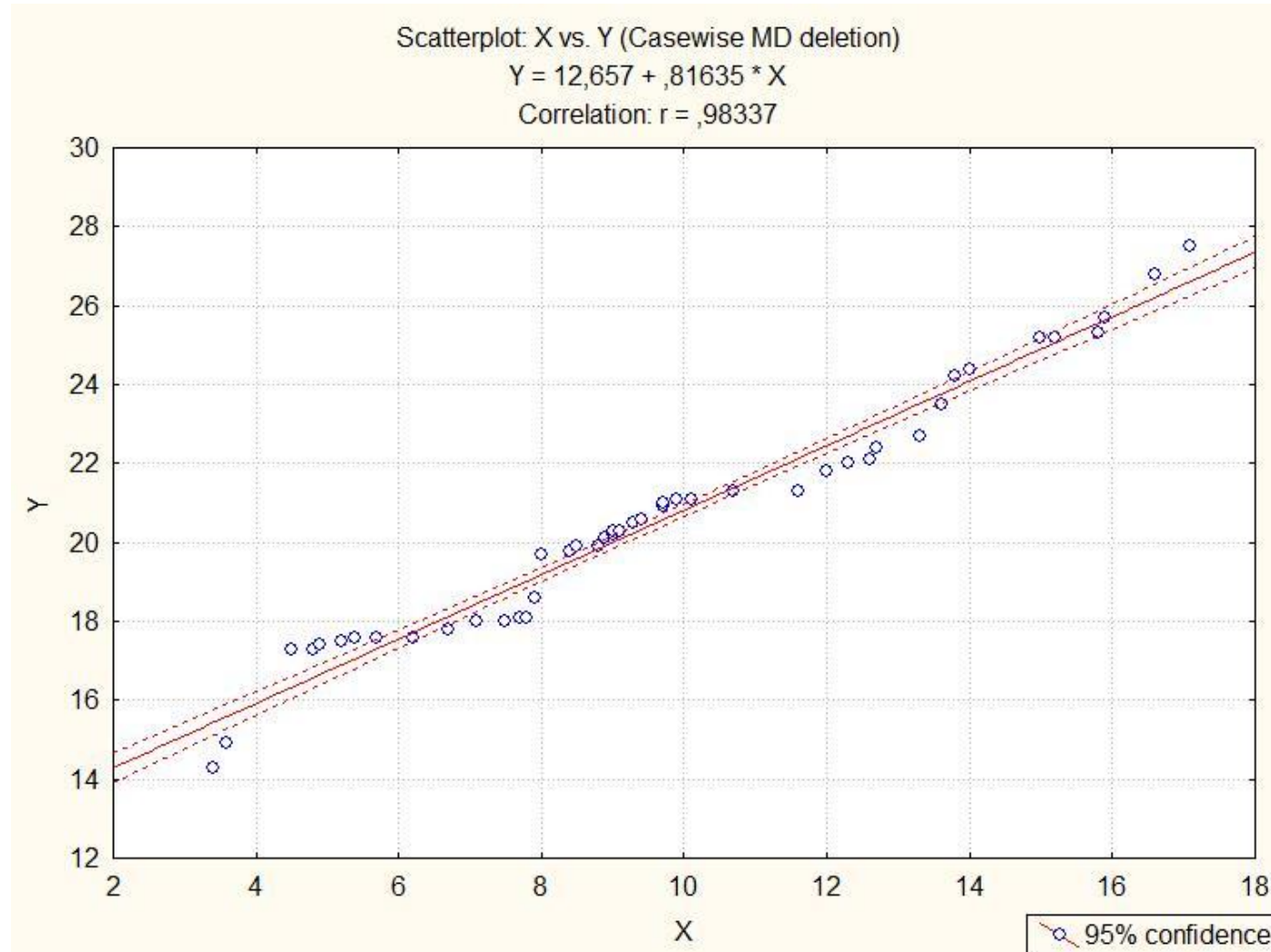
б) направлением:

- прямая;
- обратная.

в) степенью



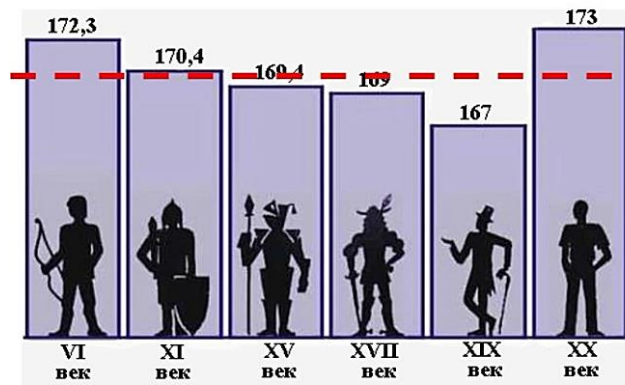
Прямолинейная связь между признаками – это такая связь, при которой равномерным изменениям первого признака соответствуют равномерные (**в среднем!**) изменения второго признака при незначительных и беспорядочных отклонениях от этой равномерности.



При измерении степени связи между разными признаками приходится сравнивать величины, выраженные в разных единицах измерения. Проводить такие сравнения возможно через использование **нормированного отклонения**:

$$\bar{X}_i = \frac{X_i - \mu}{\sigma}$$

где X_i – значение i-варианты, μ – среднее значение вариантов, σ – стандартное отклонение



$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

$$\bar{x} = 170.18 \text{ см}$$

$$\sigma = \sqrt{\frac{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}}$$

$$\sigma = \sqrt{\frac{(172.3 - 170.18)^2 + \dots + (173 - 170.18)^2}{6}} = 2.03$$

Нормированное отклонение служит универсальной и неименованной мерой развития признаков. Эти свойства нормированного отклонения и позволили сконструировать основной показатель корреляционной связи – *коэффициент корреляции*. Основная формула этого показателя имеет простую структуру:

$$r = \frac{\sum \bar{x}_1 \cdot \bar{x}_2}{v}$$

где r – коэффициент корреляции; $\bar{x}_1, \bar{x}_2$ - нормированные отклонения данных по первому и второму признаку; v – число степеней свободы (число сравниваемых пар без одной)

Сумма произведений нормированных отклонений, входящая в формулу для коэффициента корреляции, обладает следующими 3 особыми свойствами:

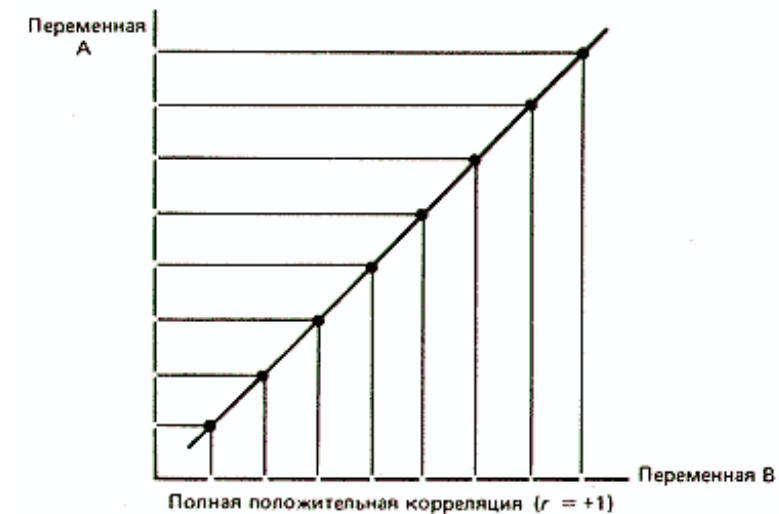
- 1) Если оба признака изменяются параллельно, то сумма произведений их нормированных отклонений даёт положительную величину.
- 2) Если при увеличении одного признака другой уменьшается, то приходится умножать положительные числа на отрицательные и вся сумма произведений нормированных отклонений даёт отрицательную величину.

3) При полных связях, когда изменения обоих признаков строго соответствуют друг другу и корреляционная связь превращается в функциональную

коэффициент корреляции может определять направление связи: при прямых связях он положителен, а при обратных связях отрицателен

Итс тру!

АБСОЛЮТЛИ



В биологических работах наиболее приемлема формула, предложенная для малых групп:

$$r = \frac{\sum X_1 \cdot X_2 - \frac{\sum X_1 \cdot \sum X_2}{n}}{\sigma_1 \cdot \sigma_2}$$

где: X_1 , X_2 – данные первого и второго признаков; n – число сравниваемых пар данных, или объектов, у которых измерено по 2 признака; σ_1 , σ_2 – стандартные отклонения по первому и по второму признаку

2. Ошибка коэффициента корреляции

Как и всякая выборочная величина (по данным выборки из генеральной совокупности), коэффициент корреляции имеет свою **ошибку репрезентативности**, вычисляемую для больших выборок по формуле:

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

коэффициент корреляции в генеральной совокупности, из которой взята выборка



эта формула может применяться **только** в исключительных случаях, когда заранее известна или предполагается степень корреляции в **генеральной совокупности**



Для проверки гипотезы о том, что коэффициент корреляции между детьми и родителями $r = +0,5$, была сопоставлена плодовитость 226 лисиц и их дочерей в соответствующем возрасте и в сходных условиях. Коэффициент корреляции оказался равным $+0,45$. Подтверждает или опровергает этот результат гипотезу?

$$s_r = \frac{1 - 0,5^2}{\sqrt{225}} = \frac{0,75}{15} = 0,05$$

критерий достоверности разности не превышает даже первого порога достоверности ($t_1 = 2,0$; $p_1 = 0,05$):

$$t_{(r-r)} = \frac{0,05}{0,05} = 1$$

Гипотеза не опровергнута, так как эмпирический коэффициент корреляции недостоверно отличается от гипотетического

В большинстве исследований значение коэффициента корреляции в генеральной совокупности неизвестно, поэтому **вместо точного значения** ошибки коэффициента корреляции **берут приближенное значение**:

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

где r – выборочное значение коэффициента корреляции, n – число сравниваемых пар данных или число объектов, у которых измерены два признака

2.1 Достоверность выборочного коэффициента корреляции

$$t_r = \frac{r}{s_r} \geq t_{st} \{v = n - 2\}$$

где t_r – критерий достоверности коэффициента корреляции; r – выборочный коэффициент корреляции; n – число коррелированных пар данных; t_{st} – стандартное значение критерия Стьюдента, определяемое по таблице для установленного числа степеней свободы и порога вероятности безошибочных прогнозов

При $t \geq t_{st}$ выборочный коэффициент корреляции достоверен. В этом случае с определенной вероятностью можно считать, что между коррелируемыми признаками имеется связь и в генеральной совокупности такая же по знаку, какая получилась в выборке (прямая или обратная).

При $t < t_{st}$ выборочный коэффициент корреляции недостоверен, что не дает возможности сделать какое-либо заключение о связи признаков в генеральной совокупности. Для выяснения этого вопроса требуется провести повторные исследования на более многочисленном материале.



For Example!



При проверке гипотезы о связи крупноплодности с жирномолочностью был рассчитан коэффициент корреляции между процентом жира в молоке у 50 коров и весом при рождении телят от этих же коров. Получено:
коэффициент корреляции: $r = +0,21$ $t_{st} = \{2,0 - 2,7 - 3,5\}$

его ошибка: $s_r = \sqrt{\frac{1 - 0,21^2}{50 - 2}} = 0,14$

критерий достоверности: $t_r = \frac{0,21}{0,14} = 1,5 \{v = 48\}$

Итог:
Выборочный коэффициент оказался недостоверным.

На основе проведенного исследования нельзя ожидать связи между крупноплодностью и жирномолочностью у всех коров вообще

2.2 Доверительные границы коэффициента корреляции

Доверительный интервал (confidence interval)

Доверительный интервал – вычисленный на основе выборки интервал значений признака, который с известной вероятностью содержит оцениваемый параметр генеральной совокупности.

«Мы на 95% уверены, что доля людей которым известна наша торговая марка находится где-то между 23,2% и 38,0%».

«Параметр находится где-то здесь
с 95% вероятностью»



For Example!

При разработке способов определения веса устриц определенного вида по их длине было измерено и взвешено 200 экземпляров и определен коэффициент корреляции между весом и длиной $r = +0,85$



Доверительные границы генерального значения коэффициента корреляции находятся общим способом по формуле:

$$\bar{r} = r \pm \Delta$$

$\Delta = t_{st} \cdot s_r$ – возможная погрешность при определении генерального параметра (t_{st} – критерий Стьюдента при числе степеней свободы $n = n - 2$; s_r – ошибка коэффициента корреляции)

Ошибка r : $s_r = \sqrt{\frac{1 - 0,85^2}{200 - 2}} = 0,037$

Число степеней свободы и критерий Стьюдента:
 $v = n - 2 = 198$, $t_{st} = \{2,0 - 2,6 - 3,3\}$

Возможная погрешность при прогнозе генерального параметра:

$$\Delta = t_{st} \times s_r = 2,0 \times 0,037 = 0,074$$

Доверительные границы:

$$r = +0,85 \pm 0,074 \text{ [не более } +0,85 + 0,074 = \mathbf{0,92}; \text{ не менее } 0,85 - 0,074 = \mathbf{0,78}]$$

Даже минимальная граница (гарантированный минимум) оказалась достаточно высокой. Это указывает на возможность практического использования вскрытой закономерности путем разработки формулы регрессии для определения веса устриц по их длине с практически достаточной точностью

2.3 Достоверность разности двух коэффициентов корреляции

Достоверность разности коэффициентов корреляции определяется так же, как и достоверность разности средних, по обычной формуле:

$$t_d = \frac{d}{s_d} \geq t_{st} \{v = n_1 + n_2 - 4\}$$

$d = r_1 - r_2$ - разность коэффициентов корреляции; $s_d = \sqrt{s_1^2 + s_2^2}$ - ошибка разности, равная корню квадратному из суммы квадратов ошибок обоих сравниваемых коэффициентов корреляции

$s^2 = \frac{1-r^2}{n-2}$ - квадрат ошибки коэффициента корреляции



For Example!

$$d = 0,80 - 0,60 = 0,20$$

$$s_d^2 = 0,0032 + 0,0024 = 0,0056$$

$$s_d = \sqrt{0,0056} = 0,075; \quad t_d = \frac{0,200}{0,075} = 2,7$$

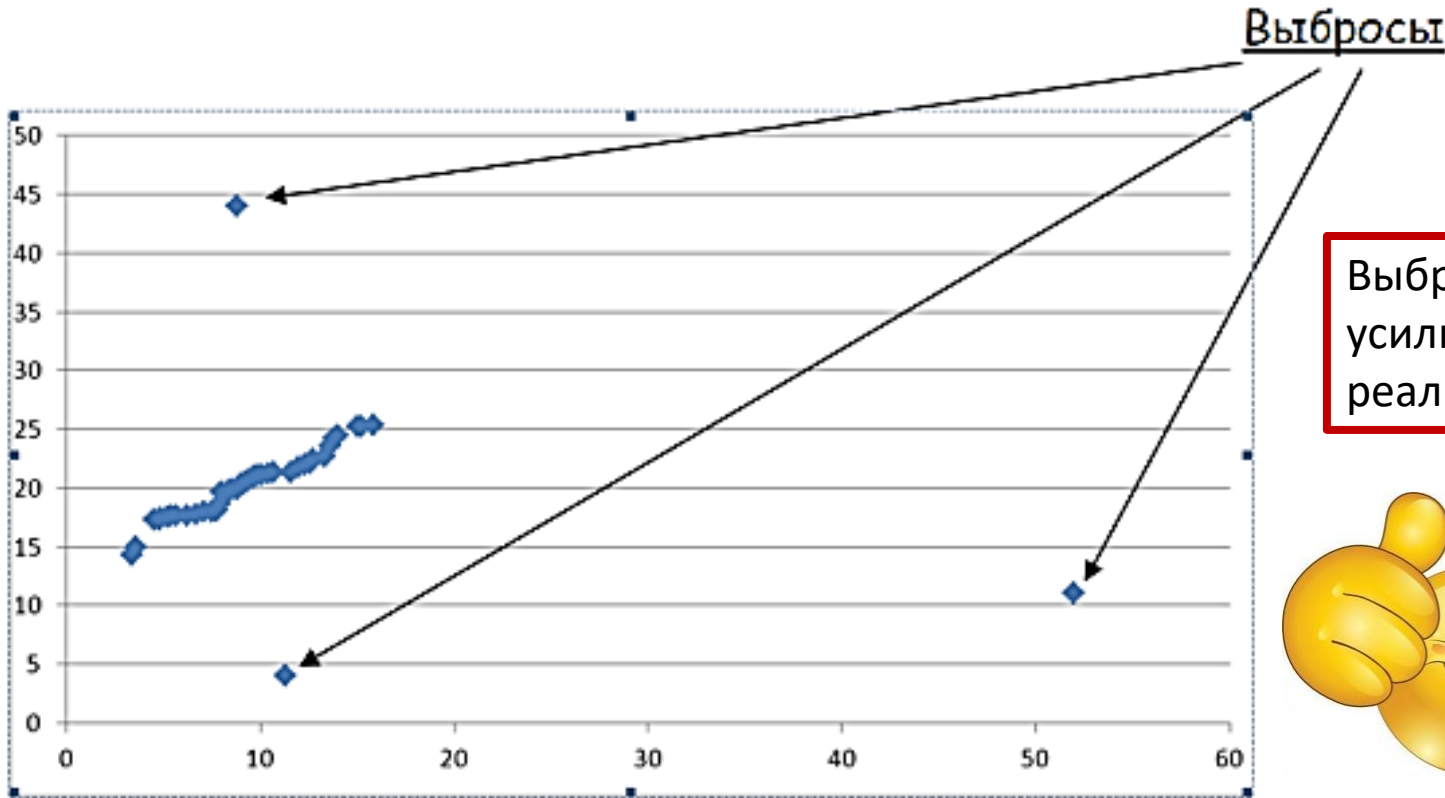
$$v = (200 + 150) - 4 = 346, \quad t_{st} = \{2,0 - 2,6 - 3,3\}$$

Для выяснения возможности применения единой формулы пересчета обхвата на высоту потребовалось выяснить: достоверно ли различие связи высоты с обхватом между лиственной и хвойной породами деревьев. Получены следующие результаты:

Сравниваемые породы достаточно достоверно (по второму порогу вероятности) различаются по степени связи между высотой и обхватом дерева. Поэтому для этих пород нельзя пользоваться единой формулой пересчета обхвата на высоту

Выбросы

Выбросы являются нетипичными, резко выделяющимися наблюдениями. Так как при построении прямой регрессии используется сумма квадратов расстояний наблюдаемых точек до прямой, то выбросы могут существенно повлиять на наклон прямой и, следовательно, на значение коэффициента корреляции.



Выбросы могут не только искусственно усилить впечатление о взаимосвязи, но также реально уменьшить существующую корреляцию



**Выбросы
исключаются из
выборки попарно!**

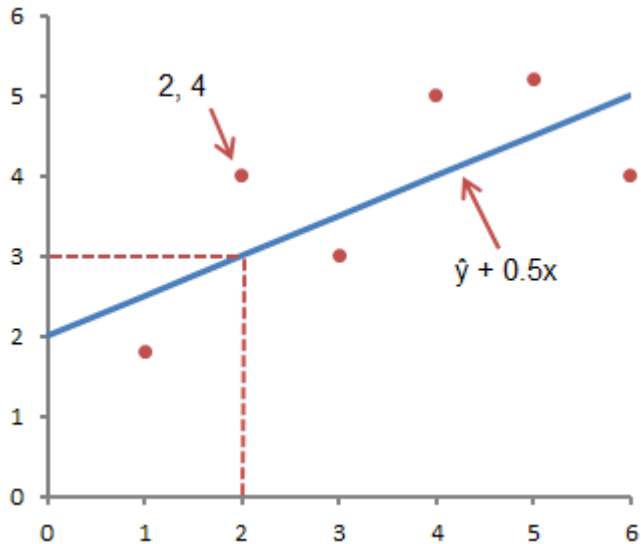
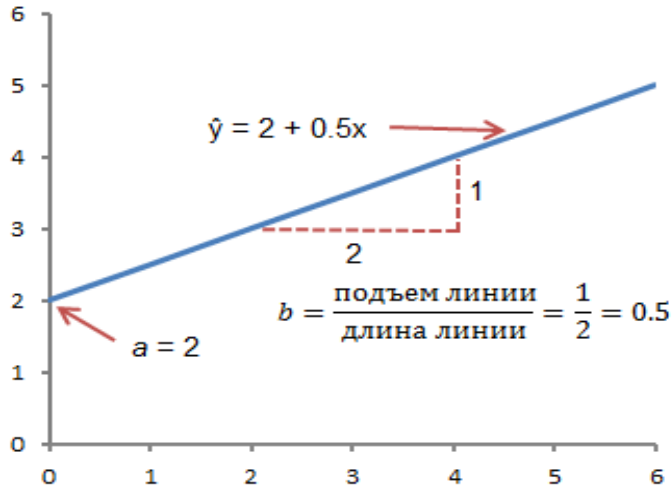
3. Понятие о регрессии

Прямолинейная корреляция характеризуется тем, что при этой форме связи каждому из одинаковых изменений первого признака соответствует вполне определенное и также одинаковое в среднем изменение другого признака, связанного с первым или зависящего от первого.

Величина, на которую в среднем изменяется второй признак, при изменении первого на единицу измерения – это **коэффициент регрессии**. Рассчитывается он по формуле:

$$R_{2/1} = \frac{\sigma_2}{\sigma_1} \cdot r_{1-2}$$

где $R_{2/1}$ – коэффициент регрессии второго признака по первому;
 σ_2 – стандартное отклонение второго признака, который изменяется в связи с изменением первого; σ_1 – стандартное отклонение первого признака, в связи с изменением которого изменяется второй признак; r_{1-2} – коэффициент корреляции между первым и вторым признаками



Ошибка коэффициента регрессии

$$S_R = \frac{\sigma_2}{\sigma_1} \cdot S_r = \frac{\sigma_2}{\sigma_1} \cdot \sqrt{\frac{1-r^2}{n-2}}$$

Критерий достоверности коэффициента регрессии

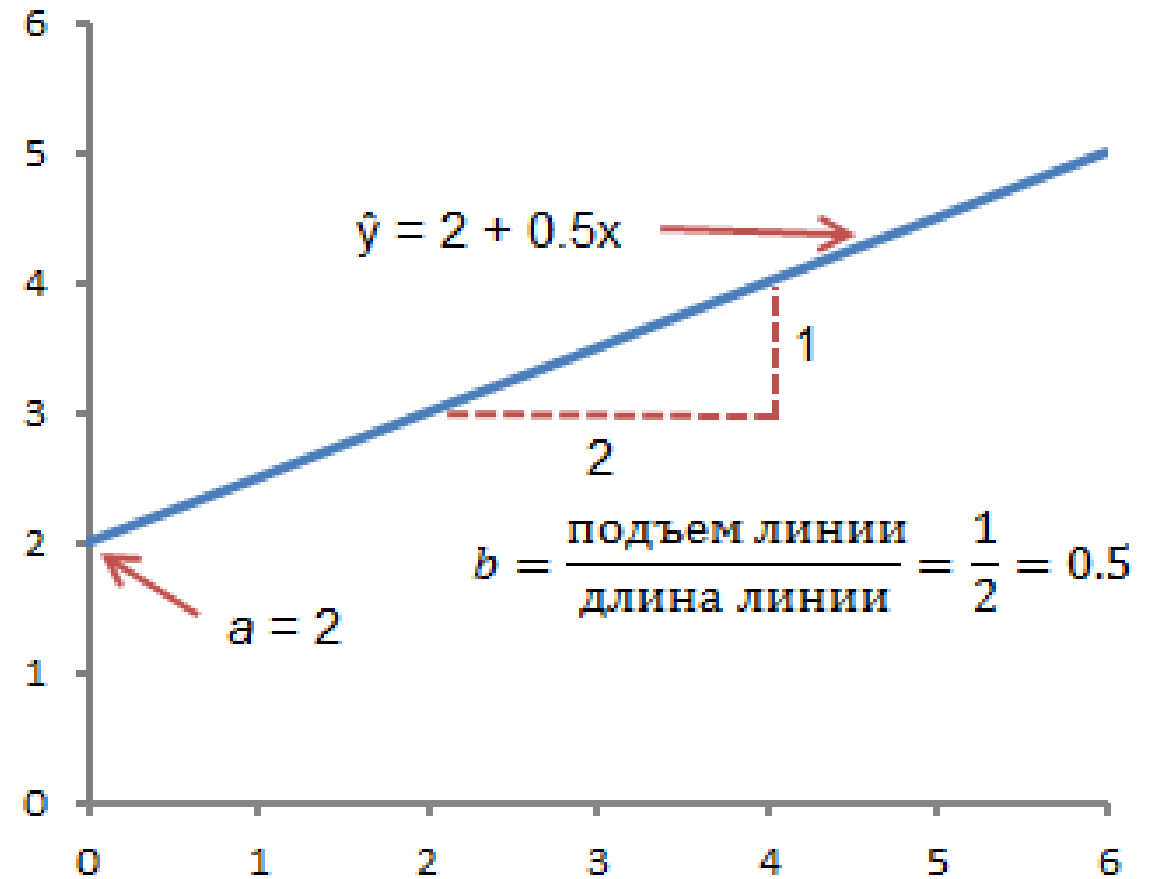
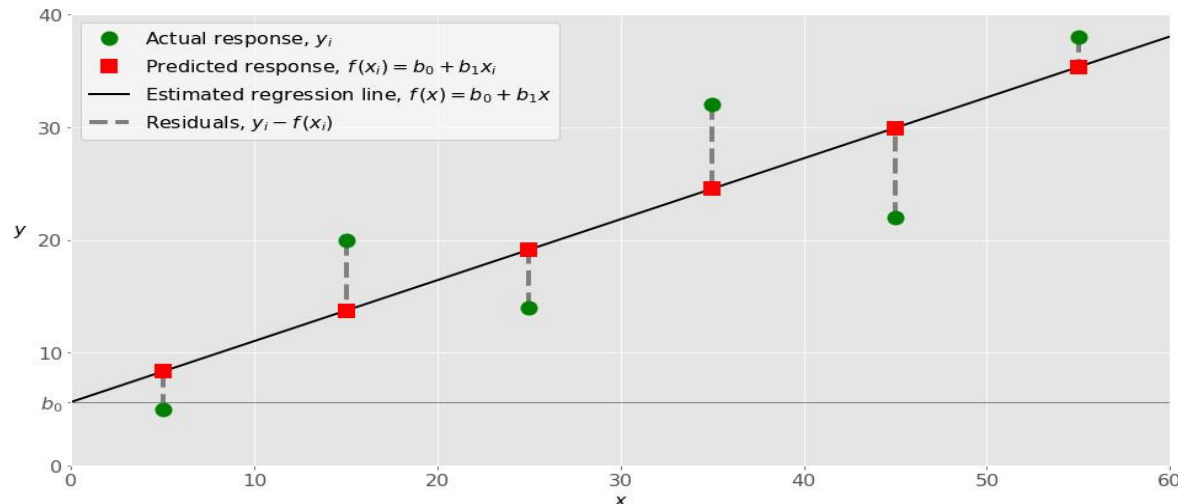
$$t_R = \frac{R}{S_R} = \frac{\frac{\sigma_2}{\sigma_1} \cdot r_{1-2}}{\frac{\sigma_2}{\sigma_1} \cdot S_r} = \frac{r}{S_r} = t_r$$

Коэффициент прямолинейной регрессии показывает, на сколько от своей средней отклоняется второй признак, если первый признак от своей средней отклонился на единицу измерения. Это можно выразить следующей формулой:

$$(X_2 - \mu_2) = R_{2/1} (X_1 - \mu_1)$$

Обозначая X_1 через x , X_2 через y , $R_{1/2}$ через b можно получить **рабочую формулу** прямолинейной регрессии:

$$y = a + bx \quad \left\{ \begin{array}{l} a = \mu_y - b\mu_x \\ b = R_{y/x} \end{array} \right.$$





Для разработки способа определения веса лошадей без взвешивания по обхвату груди было взвешено 1618 лошадей и у каждой из них измерен обхват груди (х). Получены следующие показатели:

$$n = 1618, \mu_x = 174 \text{ см}, \sigma_x = 7,9 \text{ см}$$

$$n = 1618, \mu_y = 424 \text{ кг}, \sigma_y = 56,8 \text{ кг}$$

$$\text{Коэффициент корреляции } r_{x/y} = +0,89 \pm 0,011$$

Коэффициент регрессии массы по обхвату груди:

Ошибка коэффициента регрессии:

$$R_{y/x} = \frac{\sigma_y}{\sigma_x} \cdot r_{y/x} = \frac{56,8}{7,9} (+0,89) = +6,4$$

$$s_R = \frac{\sigma_y}{\sigma_x} \times s_r = \frac{56,8}{7,9} \times 0,011 = 0,08$$

Достоверность коэффициента регрессии:

Возможная максимальная погрешность:

$$t_R = \frac{6,4}{0,08} = 80,0$$

$$\Delta = t_{st} \times s_R = 2,0 \times 0,08 = 0,16$$

Доверительные границы:

$$v = 1618 - 2 = 1616, t_{st} = \{2,0 - 2,6 - 3,3\} \quad \bar{R}_{y/x} = +6,4 \pm 0,16 = \{6,24 - 6,56\}$$



Таким образом, можно ожидать, что при увеличении (или уменьшении) обхвата груди на 1 см вес лошадей увеличится (или уменьшится) в среднем на 6,4 кг при гарантированном минимуме изменения +6,24 кг и возможном максимуме +6,56 кг, если учитывать изменения признаков в обе стороны от их средней величины

Формула

$$\begin{aligned} a &= \mu_y - R_{y/x} \cdot \mu_x = \\ &= 424 - (+6,4) \cdot 174 = -690, \\ b &= R_{y/x} = +6,4 \\ \mathbf{y} &= \mathbf{a + bx = -690 + 6,4x} \end{aligned}$$

4 Ошибки элементов уравнения прямой регрессии

1) Ошибка коэффициента регрессии:

$$s_R = \frac{\sigma_y}{\sigma_x} \cdot \sqrt{\frac{1-r^2}{n-2}} = \frac{\sigma_y}{\sigma_x} \cdot s_r$$

2) Ошибка уравнения регрессии

$$s_{y_x} = \sigma_y \cdot \sqrt{\frac{1-r^2}{n-2}}$$

$$s_{y_x} = 56,8 \cdot 0,011 = 0,62$$

Следовательно, максимальная погрешность в определении уровня точек линии регрессии при первом пороге вероятности безошибочных прогнозов ($\beta_1 = 0,95$, $t_1=2,0$) будет $\Delta = t \cdot s_{y_x} = 2 \cdot 0,62 \text{ кг} = \pm 1,24 \text{ кг}$

3) Ошибка индивидуальных определений функции

$$s_y = \sigma_y \sqrt{1-r^2}$$

$$s_y = 56,8 \sqrt{1-0,89^2} = 26,2$$

Следовательно, индивидуальная погрешность в определении веса лошадей по обхвату груди по найденной формуле регрессии, принимая первый порог вероятности безошибочных прогнозов ($\beta_1 = 0,95$, $t_1 = 2,0$), в крайних случаях не будет превышать $\Delta = 2 \times 26 \text{ кг} = \pm 52 \text{ кг}$

Спасибо за внимание!!!

