

Лекция

Методы прогнозирования

д.б.н. Койчубеков Б.К.

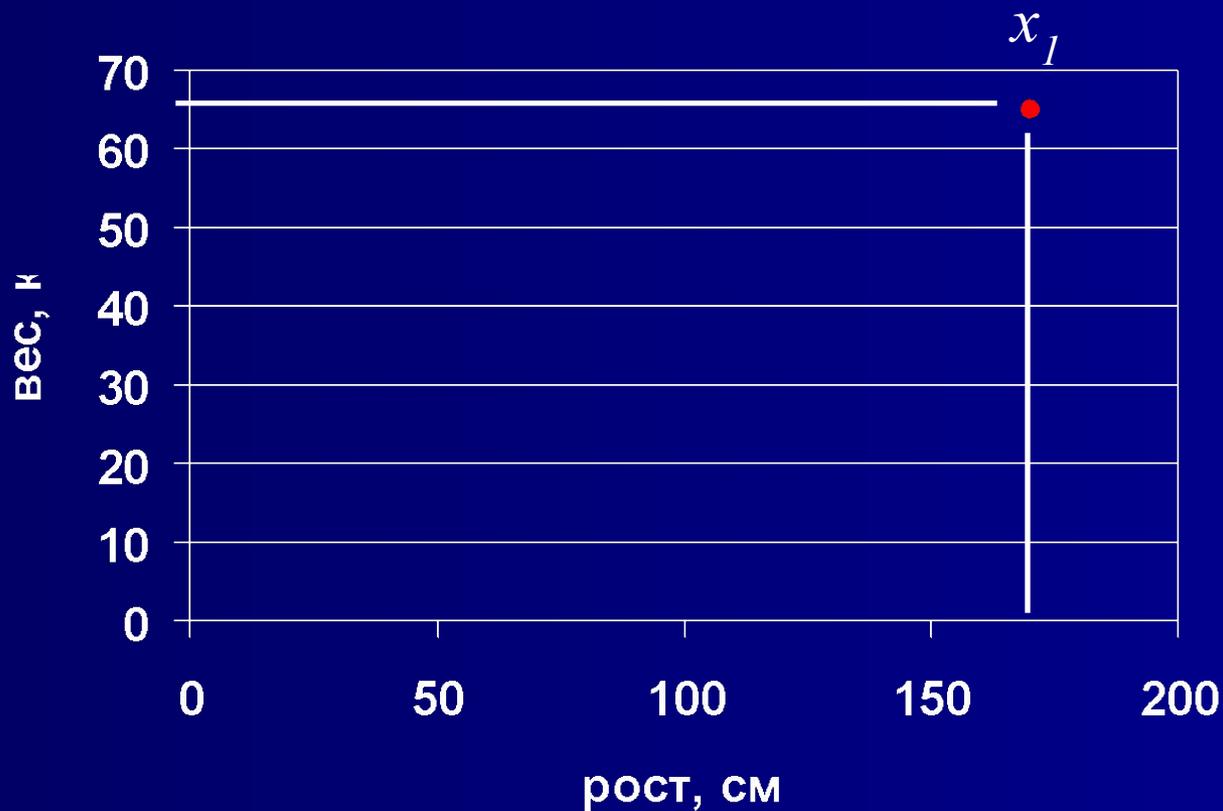
План лекции

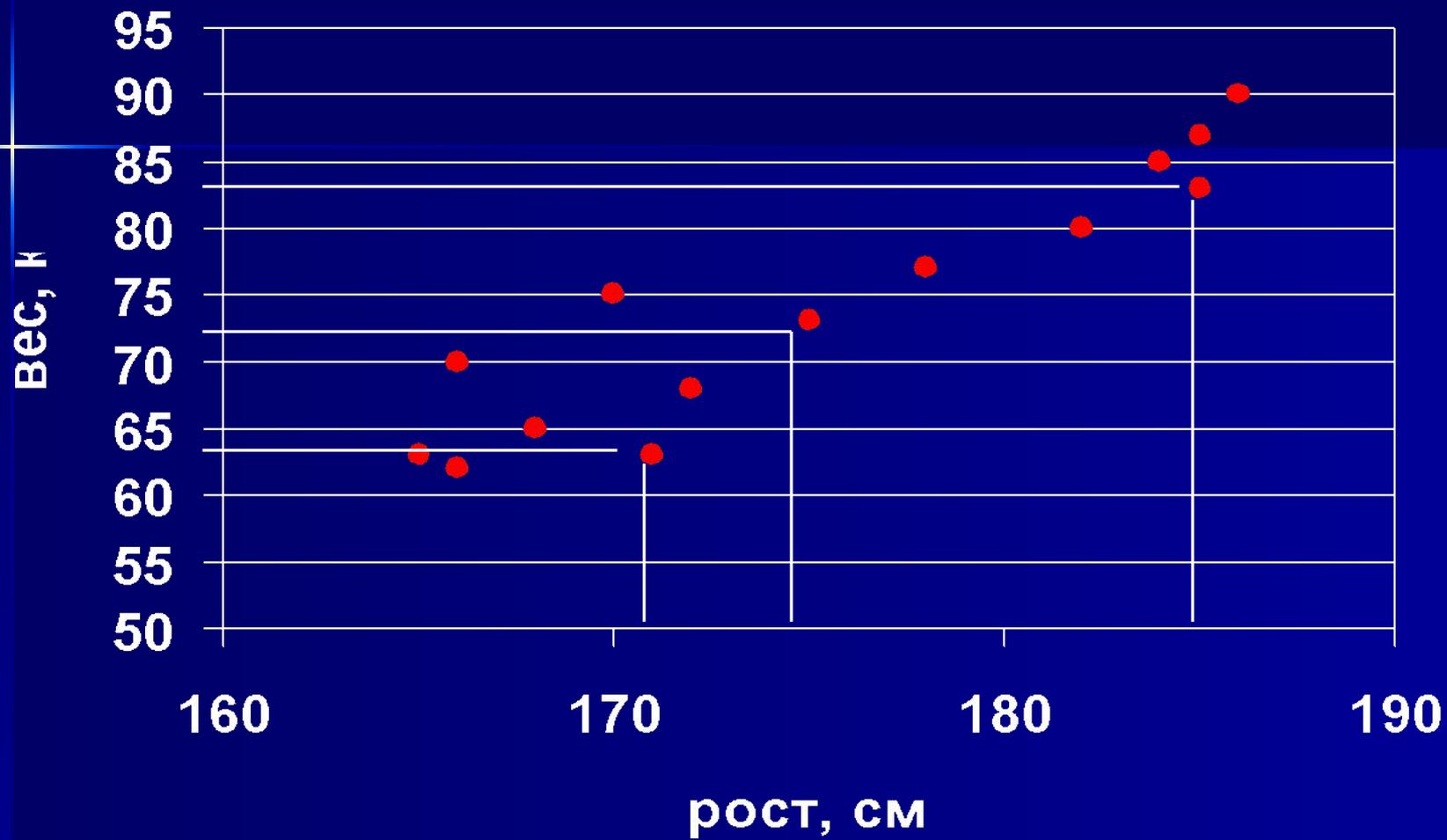
- Понятие корреляции
- Прогнозирование на основании уравнения регрессии
- Прогнозирование на основе кривой выживаемости

Какова прогнозируемая смертность в 2015 г. и на чем основан прогноз



- Отложим на графике рост и вес каждого обследованного из 200 жителей





- Из графика видно, что между ростом и весом есть определенная взаимосвязь – чем выше рост, тем больше вес. Эта связь линейная.

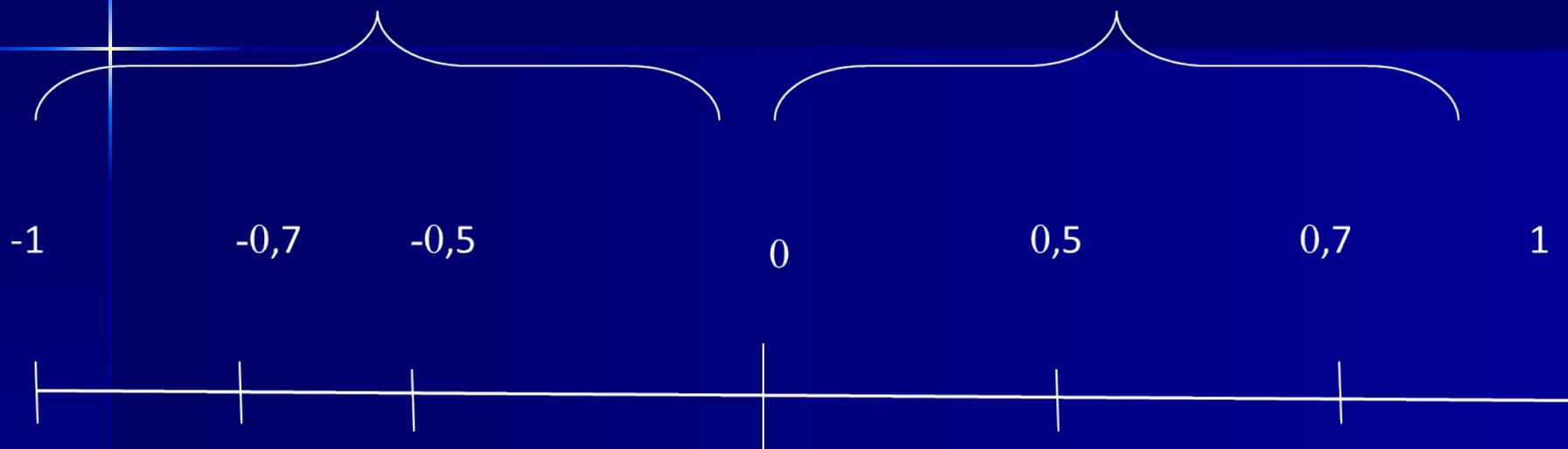


- Степень выраженности связи между вариационными рядами отражает понятие *корреляциию*
- Связь может быть **слабой, средней, сильной**. Связь может и отсутствовать.
- Количественно взаимосвязь между случайными величинами определяет *коэффициент корреляции - r*

- Коэффициент корреляции лежит в пределах
$$-1 \leq r \leq 1.$$
- Если $r < 0$, то это означает, что с увеличением величины X_1 соответствующие им значения X_2 второго вариационного ряда в среднем также уменьшаются.
- Если $r > 0$, то с увеличением значений одной величины другая также в среднем возрастает.
- Если $r = 0$, то это означает, что случайные величины X_1 и X_2 абсолютно независимы.
- При $r = 1$ между параметрами существует прямо пропорциональная функциональная зависимость (в медико-биологических исследованиях крайне редкий случай).

Обратная корреляция

Прямая корреляция



Сильная
обратная

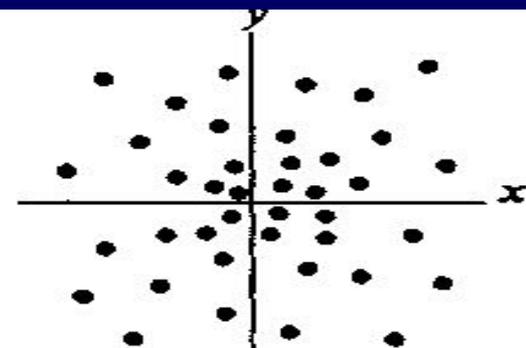
Средняя
обратная

Слабая
обратная

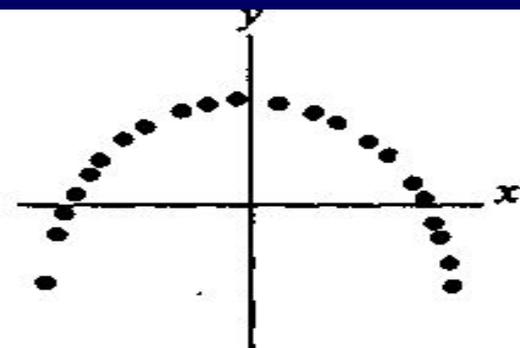
Слабая
прямая

Средняя
прямая

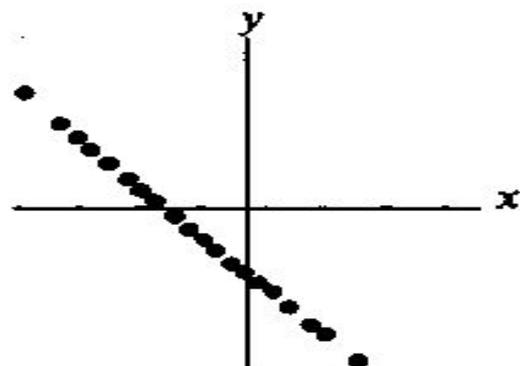
Сильная
прямая



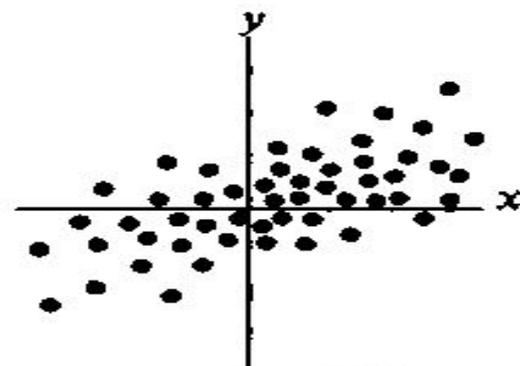
a $r_{xy}=0$



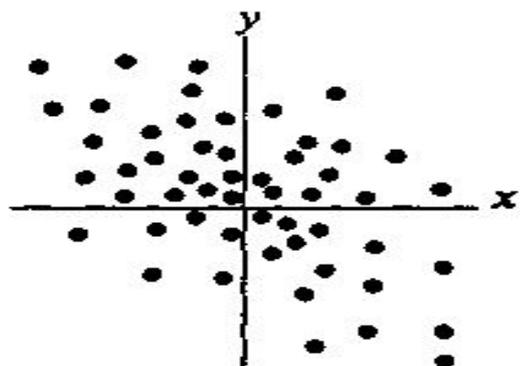
б $r_{xy}=0$



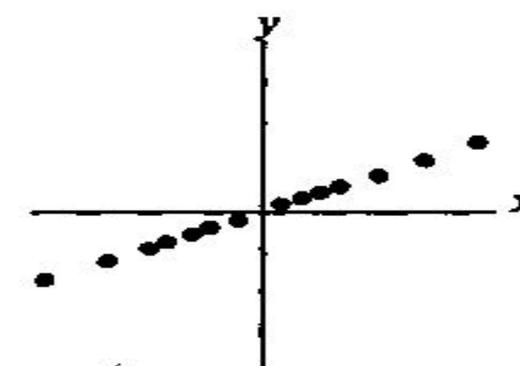
в $r_{xy}=-1$



г $r_{xy}=+0,5$



д $r_{xy}=-0,30$



е $r_{xy}=+1$

Коэффициент корреляции Пирсона

- Для двух случайных величин X_1 и X_2 (n - объем каждой выборки), если они **нормально** распределены, их **линейную** взаимосвязь можно вычислить

$$r = \frac{\sum_{i=1}^n (x_{1i} - \bar{X}_1) \times (x_{2i} - \bar{X}_2)}{\sqrt{\sum_{i=1}^n (x_{1i} - \bar{X}_1)^2 \times \sum_{i=1}^n (x_{2i} - \bar{X}_2)^2}}$$

Проверка на статистическую значимость

- $H(0): r=0$

- Проверяется гипотеза по критерию Стьюдента:

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

- $t_{\text{крит}}$ по таблице для заданного уровня значимости α и числа степеней свободы $f=n-2$
- Если $|t_{\text{выч}}| \geq t_{\text{крит}}$ то принимается $H(1)$ и делается вывод, что между величинами существует значимая корреляция.
- Если $|t_{\text{выч}}| < t_{\text{крит}}$ то принимается $H(0)$ и делается вывод о независимости исследуемых величин (коэффициент корреляции незначим).

Если

- выборка мала
- Или распределение выборки не соответствует нормальному
- Или имеем дело с качественными ординальными величинами, то используется **коэффициент корреляции рангов К. Спирмена**

$$r_s = 1 - \frac{6 \times \sum_{i=1}^n d_i^2}{n \times (n^2 - 1)}$$

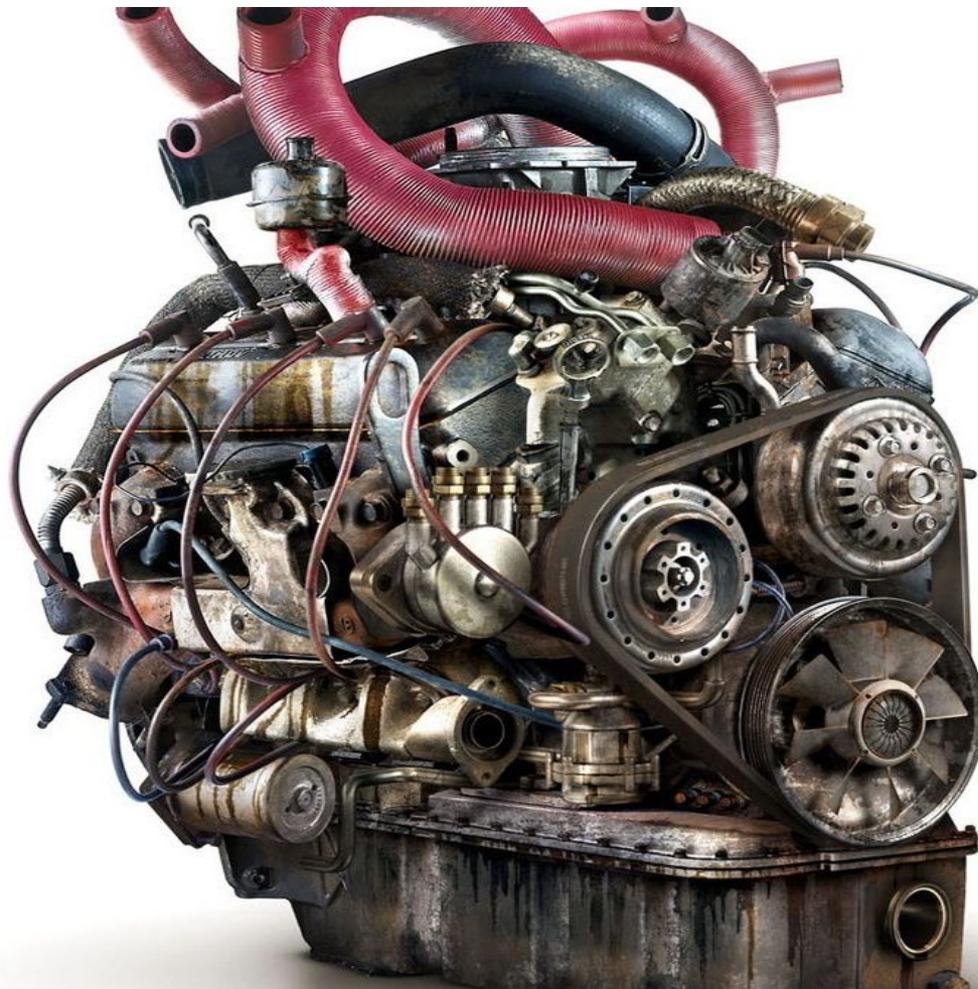
Проверка на статистическую значимость

- Для проверки гипотезы о значимости коэффициента корреляции Спирмена можно воспользоваться таблицей критических значений .
- Если вычисленный коэффициент корреляции превышает табличное значение, то связь между величинами признается достоверной.

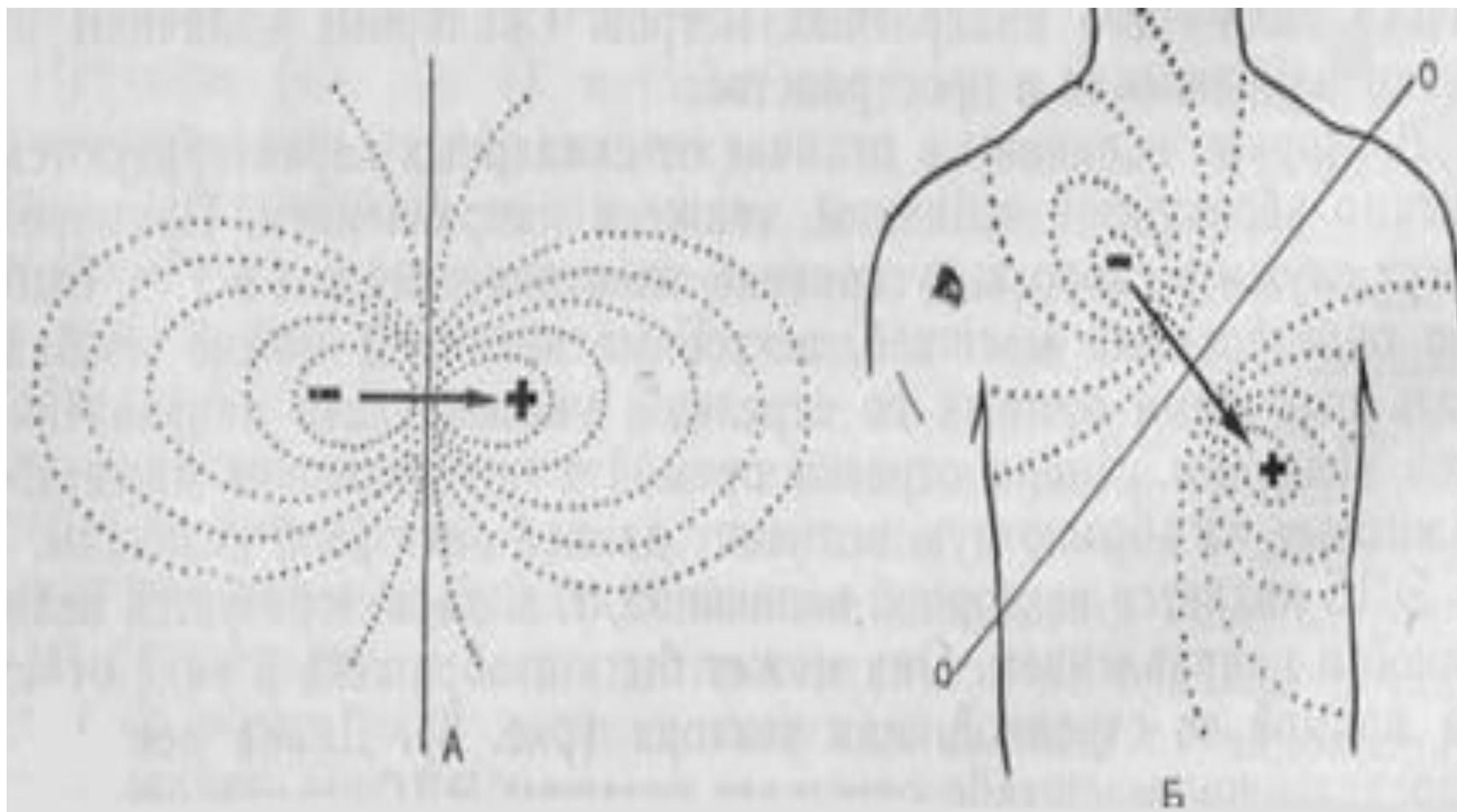
- где d_i — разность между рангами сопряженных признаков, n — число парных членов ряда.
- При полной связи ранги признаков совпадут и разность между ними будет равна 0, соответственно коэффициент корреляции будет равен 1. Если же признаки варьируются независимо, коэффициент корреляции получится равным 0

Регрессионный анализ

Механическая модель сердца



Электрическая модель сердца

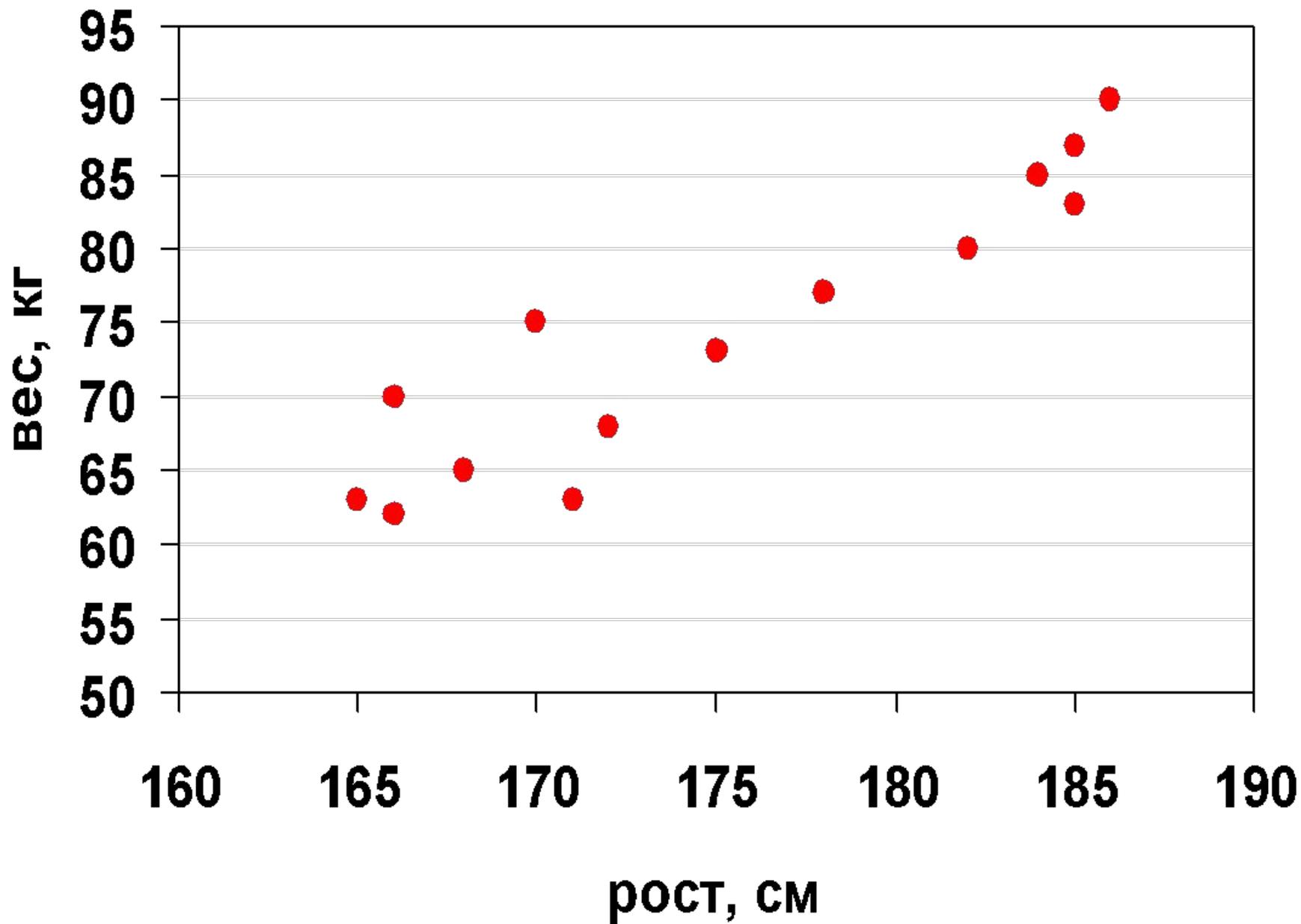


Математическая модель сердца

- **$CO = z * Q * DP * S * T * 1333 / ((T - S) * C_{\text{э}})$** , где:
 - z - фактор поправки (отношение длины артериального русла ко всему сосудистому руслу), который лежит в пределах от 0,48 до 0,607 (для человека z в большинстве случаев около 0,6 и потому обычно берут эту величину);
 - DP — истинная пульсовая амплитуда,
 - S - время изгнания;
 - T — время полной сердечной инволюции
 - $(T - S)$ — время диастолического периода в секундах;
 - $C_{\text{э}}$ — скорость распространения пульсовой волны по артериям эластического типа;
 - Q — площадь поперечного сечения аорты.

Вернемся к нашему графику зависимости веса от роста.

Как мы уже указали есть определенная взаимосвязь между этими величинами, которая оценивается коэффициентом корреляции.



- Из графика видно, что при увеличении роста вес также увеличивается, хотя и не во всех случаях.
- Попробуем вывести некоторую функцию, связывающую эти величины

$$y = f(x)$$

y - зависимая величина (вес)

x - независимая величина (рост)

- Наиболее простой является линейная функция, которая имеет вид

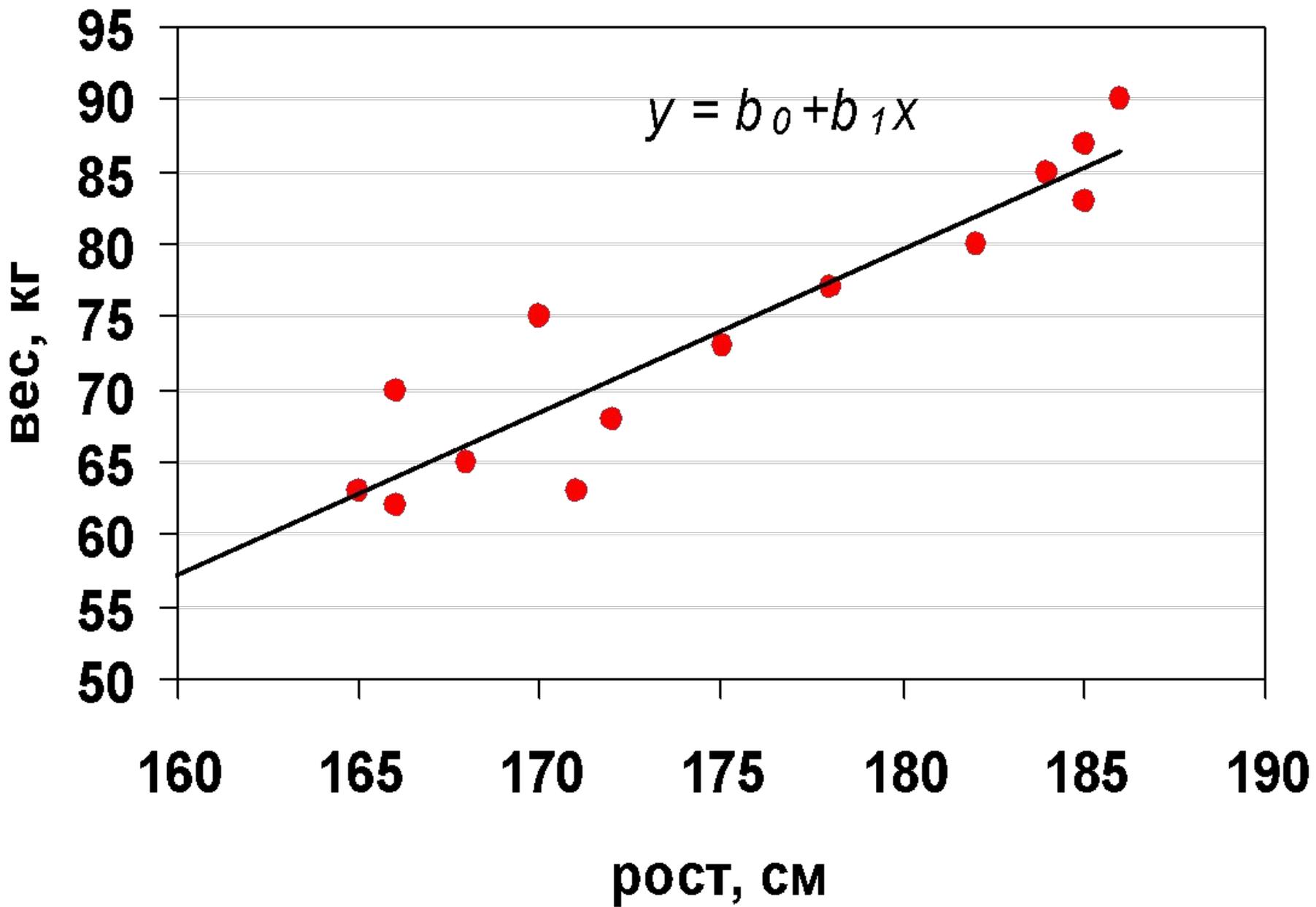
$$y = b_0 + b_1x$$

и называется уравнением регрессии

- **b_0 и b_1** - постоянные коэффициенты
 b_1 – коэффициент регрессии

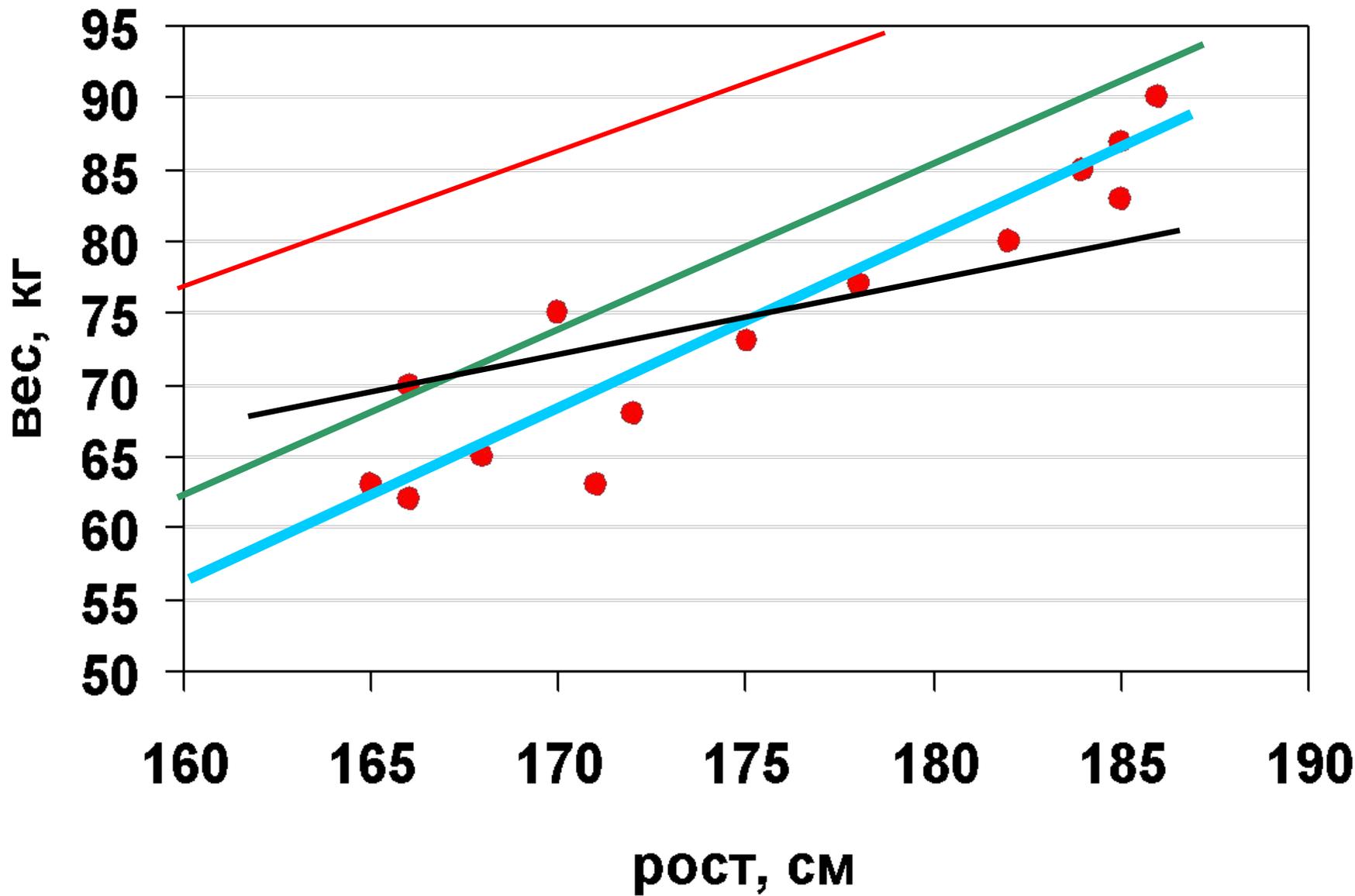
- Иногда запись имеет вид

$$y = a + bx$$

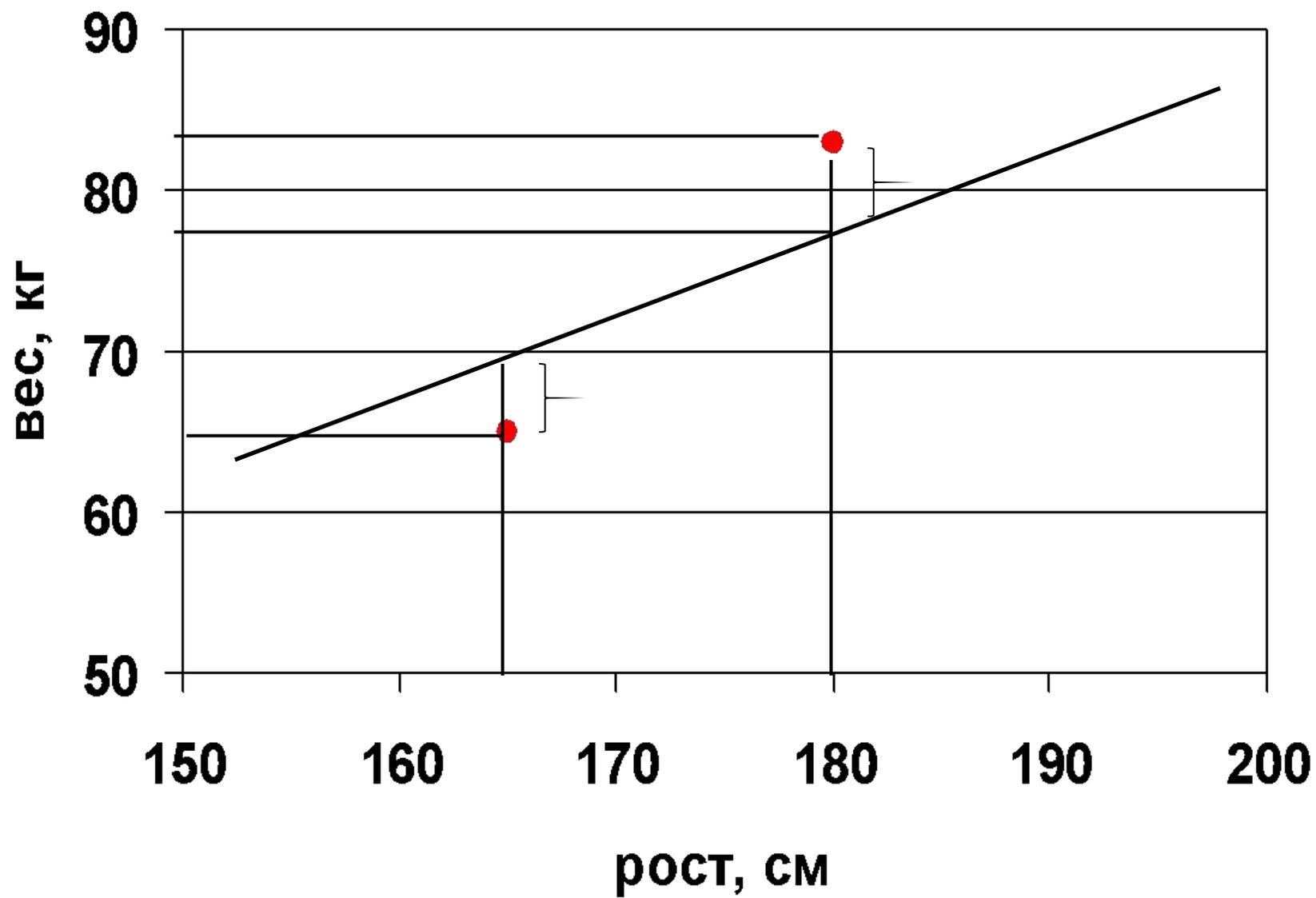


- Это функция показывает как в среднем меняется величина y при изменении величины x
- Т.е. по этой функции зная величину x можно вычислить (предсказать) величину y

- Через точки на графике можно провести сколько угодно много прямых



- Каждая прямая отличается от других значениями коэффициентов b_0 и b_1
- Для выбора наиболее оптимального служит метод наименьших квадратов, который позволяет выбрать такие коэффициенты b_0 и b_1 , что прямая регрессии наилучшим образом отражает взаимосвязь между изучаемыми величинами



- Из графика видно, что реальные данные и данные, полученные по уравнению отличаются на некоторую величину (т.е. существует отклонение)
- Необходимо выбрать такую линию, чтобы сумма всех отклонений была минимальной

- Полученная функция является математической моделью взаимосвязи двух случайных величин
- Т.к. мы рассмотрели зависимость только от **одной** независимой переменной X эта зависимость носит **линейный** характер, то такая модель носит название ***простой линейной регрессии***

- Как оценить полученную модель, т.е. насколько хорошо модель отражает взаимосвязь между исследуемыми величинами.
- Можно использовать **коэффициент детерминации R^2**

Он показывает сколько процентов исходных (выборочных) данных вписывается в полученную модель

- Если независимых переменных много

$$x_1, x_2, \dots, x_n$$

- То возможно построение уравнение множественной линейной регрессии

$$y = b_0 + b_1 x_1 + b_2 x_2 \dots + b_n x_n$$

- Например, вес зависит также от места проживания, типа питания, социального положения, наследственных факторов, хронических заболеваний и т.д.

- Возможны также нелинейные модели

$$y = b_0 + b_1 x_1^2 + b_2 x_2$$

- Математические модели можно использовать для прогнозирования

