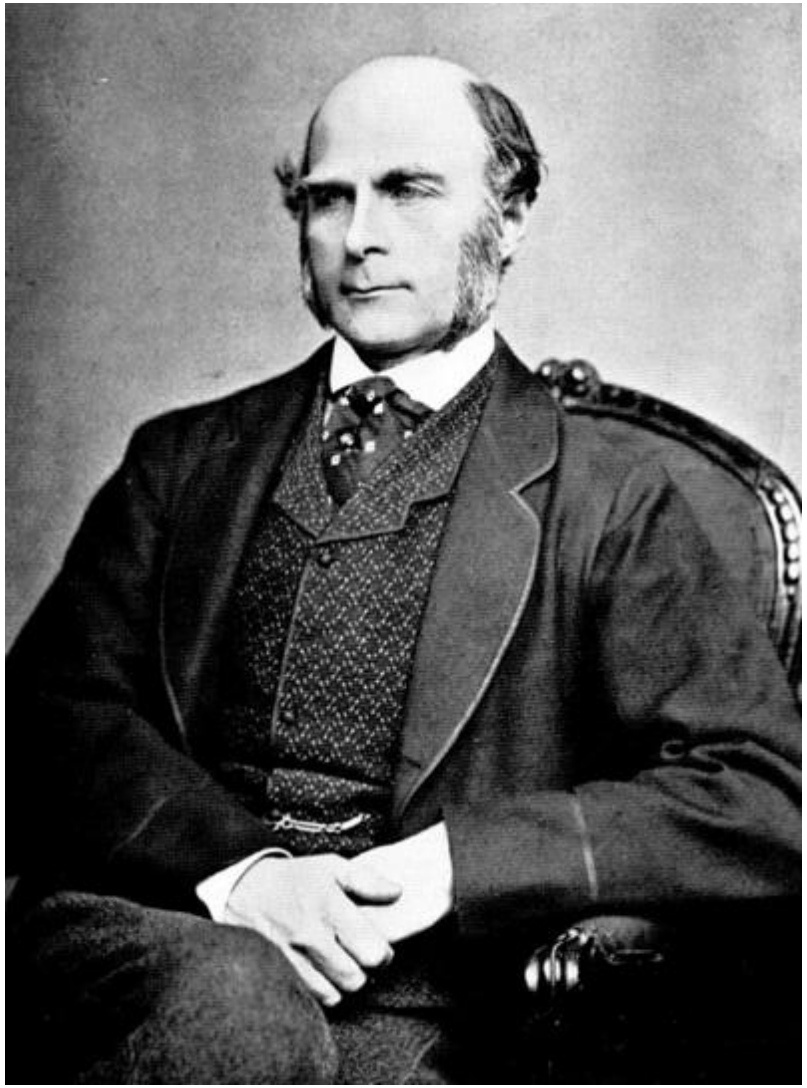
Введение в линейную регрессию

- Регрессия – от лат. regressio:
  обратное движение, возвращение
- 1886 г., Фрэнсис Гальтон



- 205 семей, 962 детей
- У более высоких родителей рождаются более высокие дети, а у более низких – низкие
- Дети высоких родителей оказываются менее высокими, а дети более низких – более высокими
- Регрессия к среднему (regression to the mean)
- Все значения лежат практически на одной линии (линии регрессии)

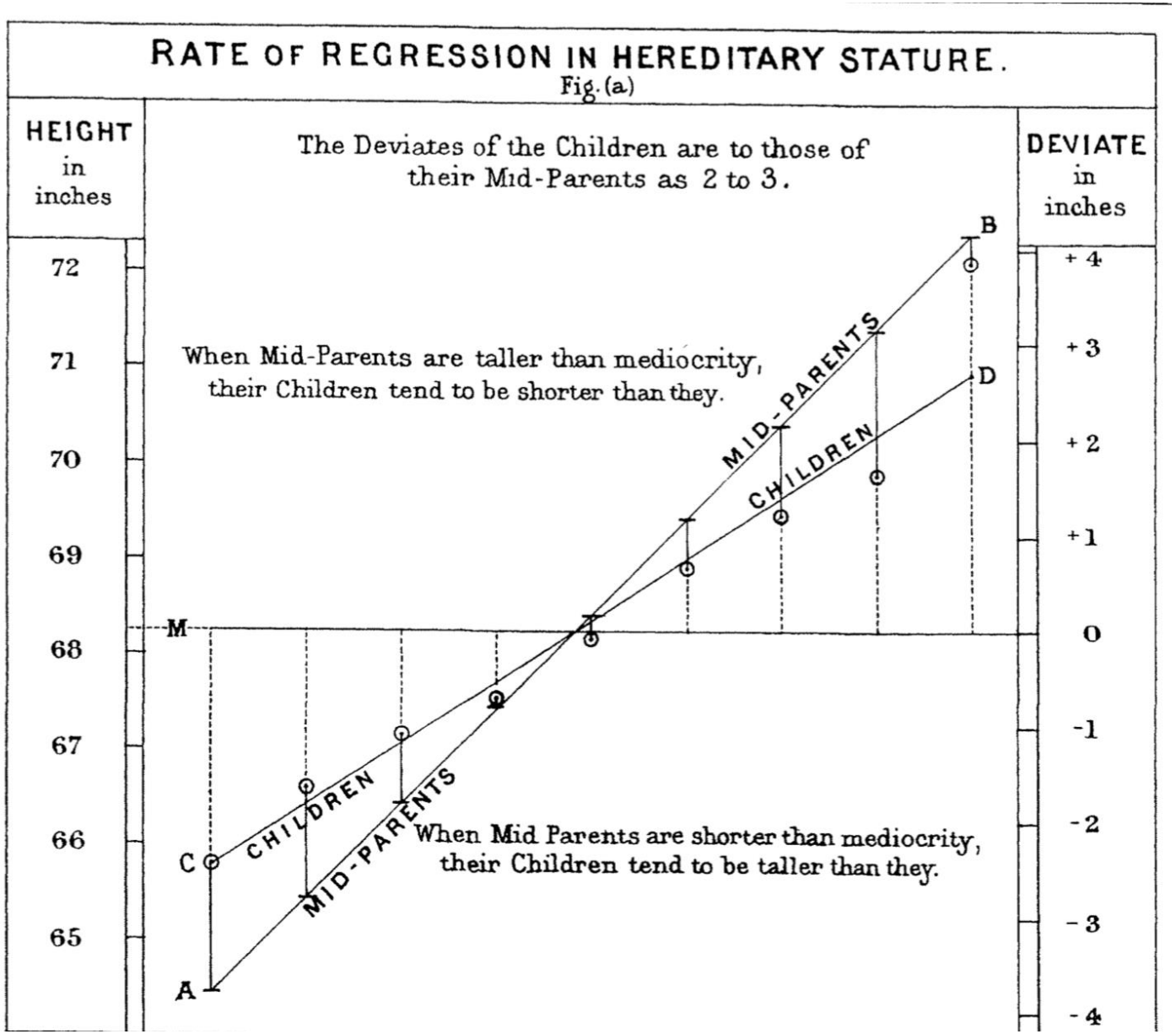
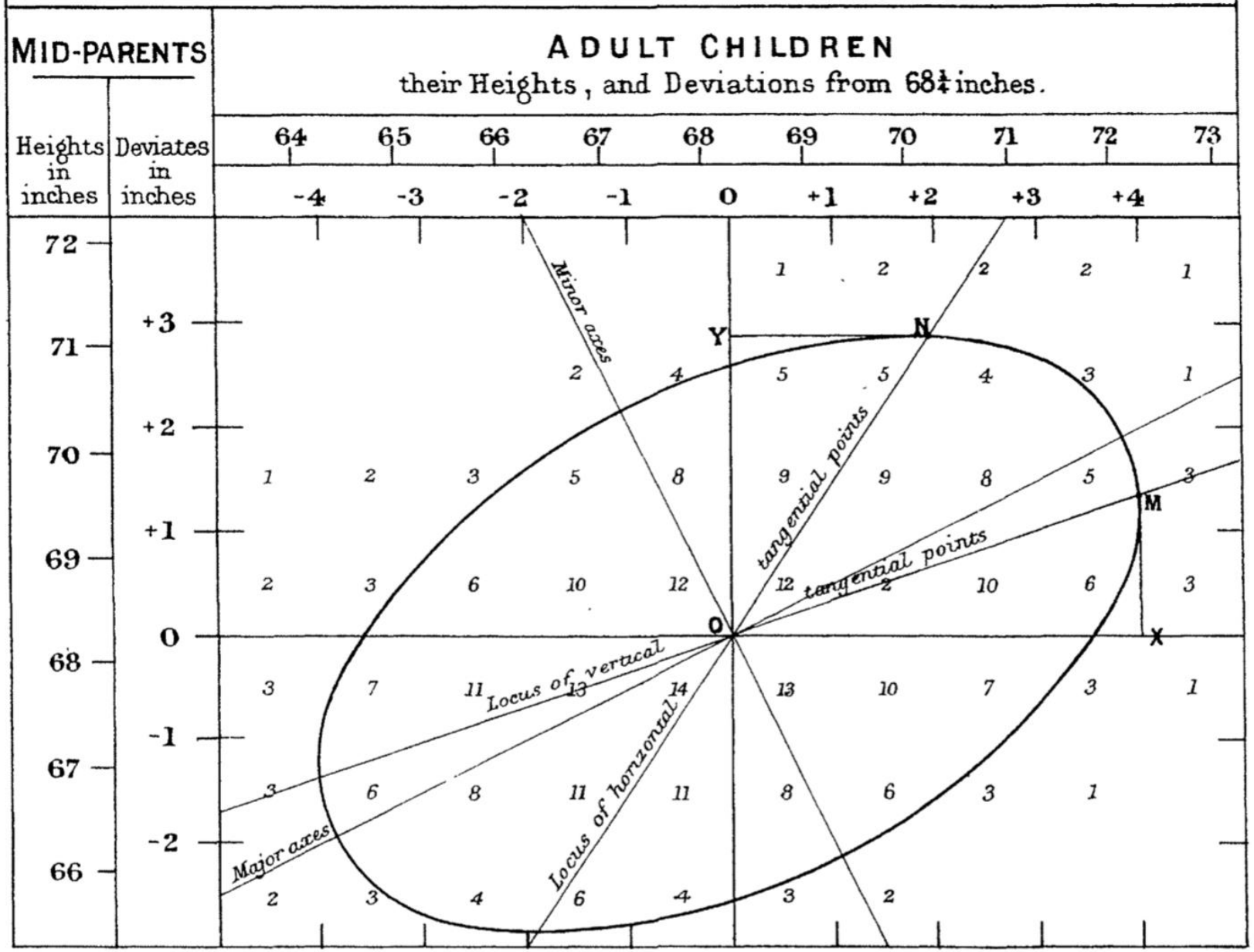
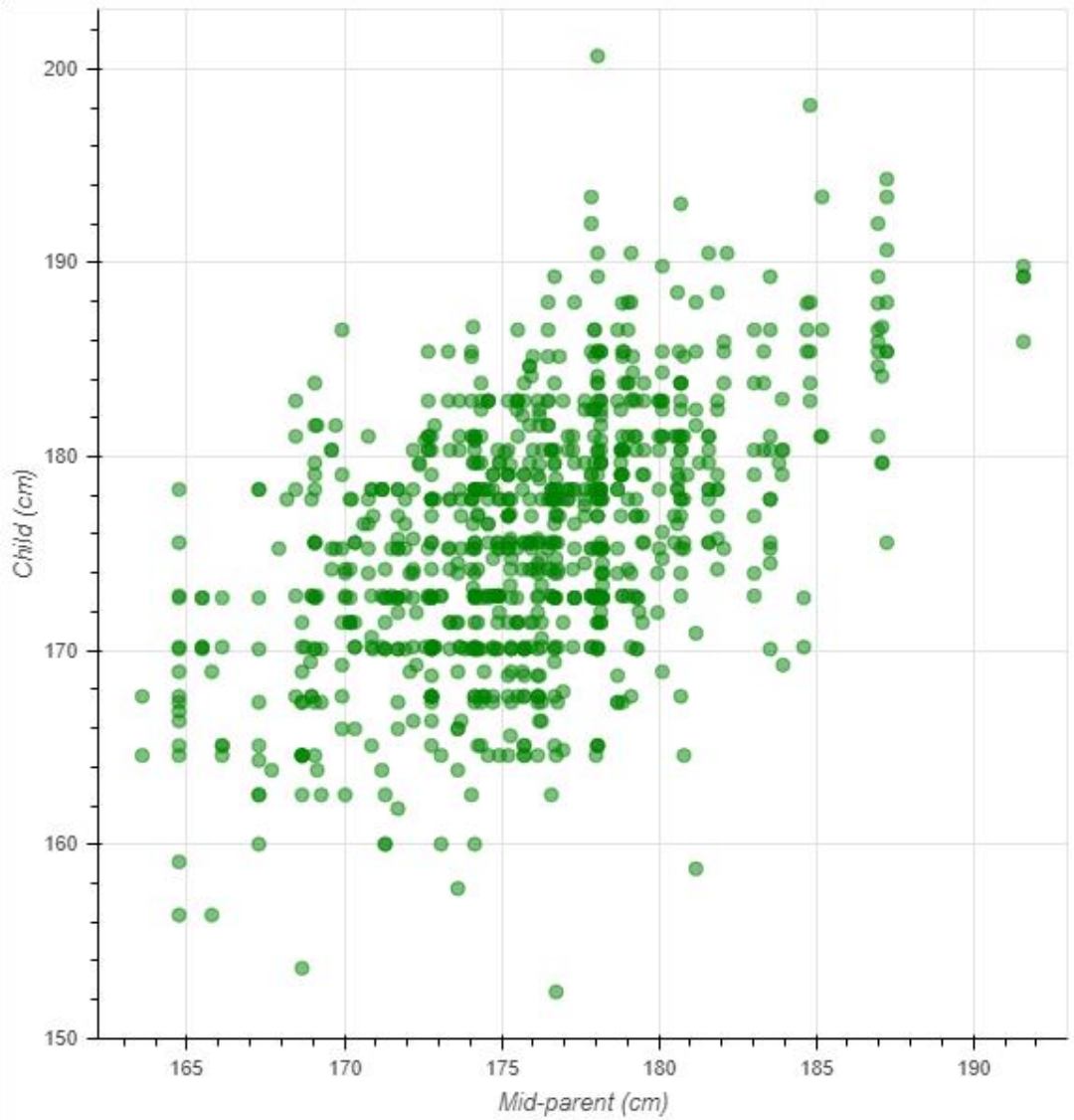
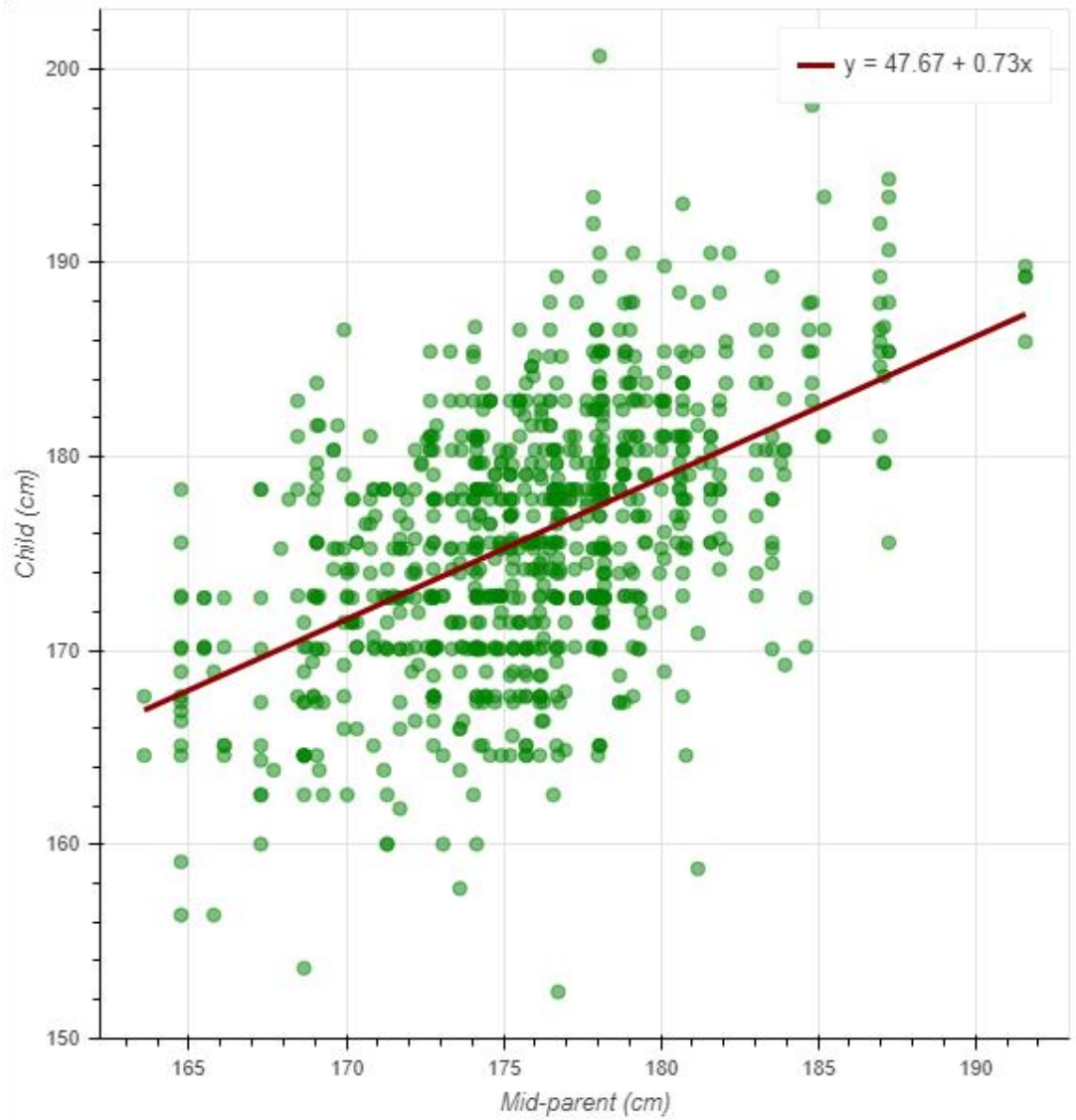


DIAGRAM BASED ON TABLE I.

(all female heights are multiplied by 1·08)




200
190
180
170
160
150
Child (cm)
165
170
175
180
185
190
Mid-parent (cm)


y = 47.67 + 0.73x
200
190
180
170
160
150
Child (cm)
165
170
175
180
185
190
Mid-parent (cm)

- $Y_i = a_0 + a_1 X_i + \varepsilon_i$ – реальные значения
- $\hat{Y}_i = a_0 + a_1 X_i$ – предполагаемая линия регрессии

- X_i – предиктор/фактор, Y_i – отклик
- a_0 – свободный член (intercept)
- a_1 – угловой коэффициент/коэффициент регрессии
- ε_i – случайная ошибка

Метод наименьших квадратов (1795)

- $\sum \varepsilon_i^2 \to \min$
- $\begin{cases} \hat{a}_1 = \frac{\overline{XY} - \bar{X} \cdot \bar{Y}}{\overline{X^2} - \bar{X}^2} \\ \hat{a}_0 = \bar{Y} - \hat{a}_1 \bar{X} \end{cases}$
- $\widehat{Y}_i = \hat{a}_0 + \hat{a}_1 X_i$, тогда

$$\widehat{Y}_i - \bar{Y} = \hat{a}_1 (X_i - \bar{X})$$

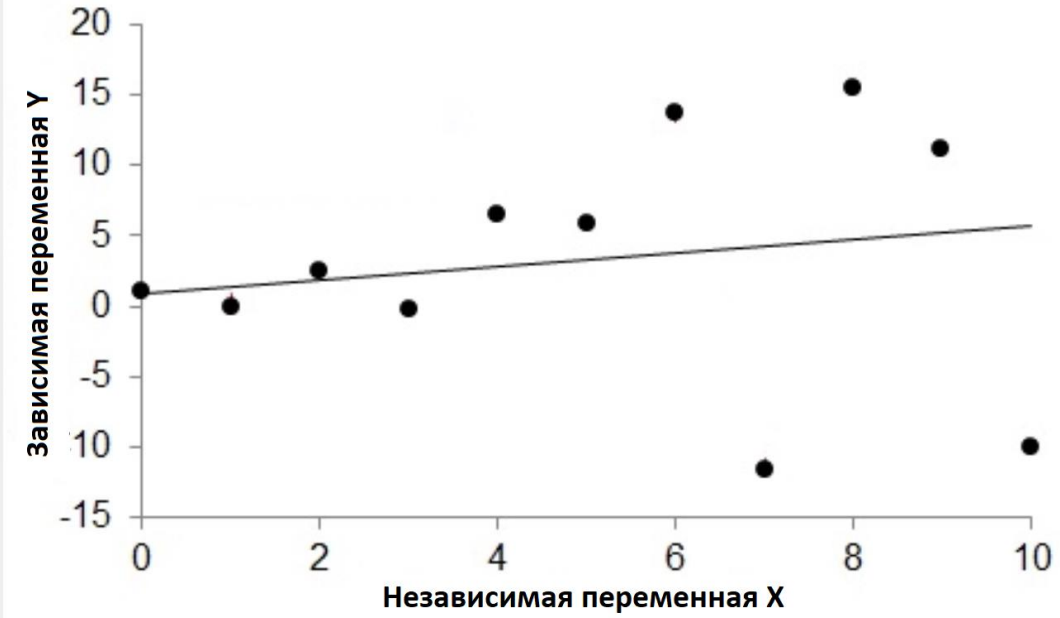
Условия Гаусса-Маркова

- $\boldsymbol{E}\varepsilon_i = 0, \boldsymbol{D}\varepsilon_i = \sigma^2$
- $corr(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$
- $corr(X_i, \varepsilon_j) = 0, \forall i, j$

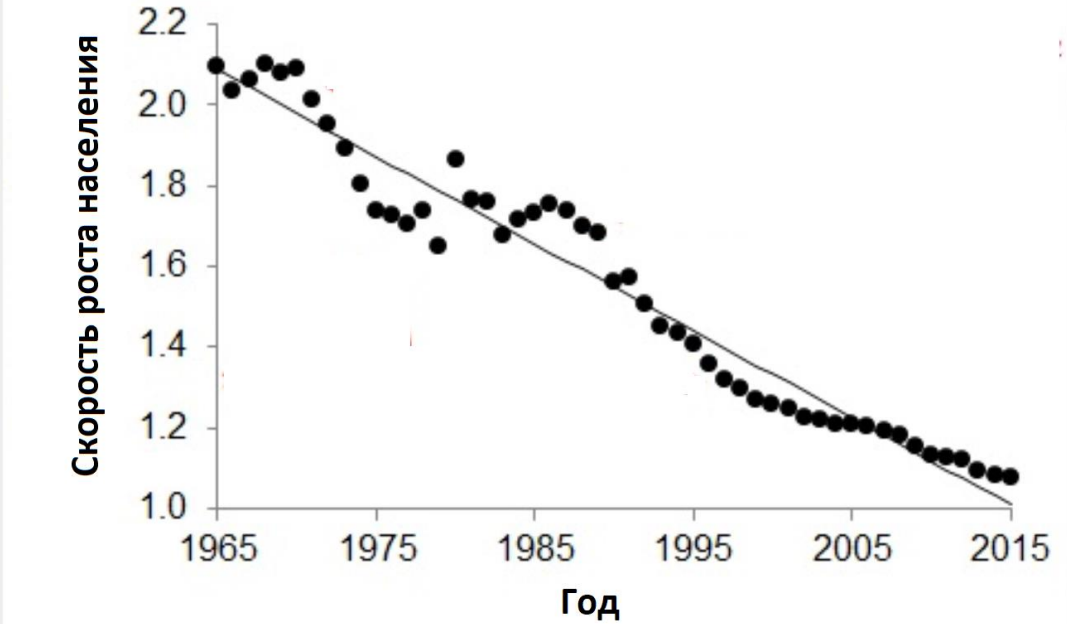
Если $\varepsilon_i \sim N_{0,\sigma^2}$, то оценка МНК совпадает с оценкой ММП

Пример нарушения условий Гаусса-Маркова

Разная дисперсия наблюдений



Автокорреляция наблюдений



Для нашей модели регрессии получили линию

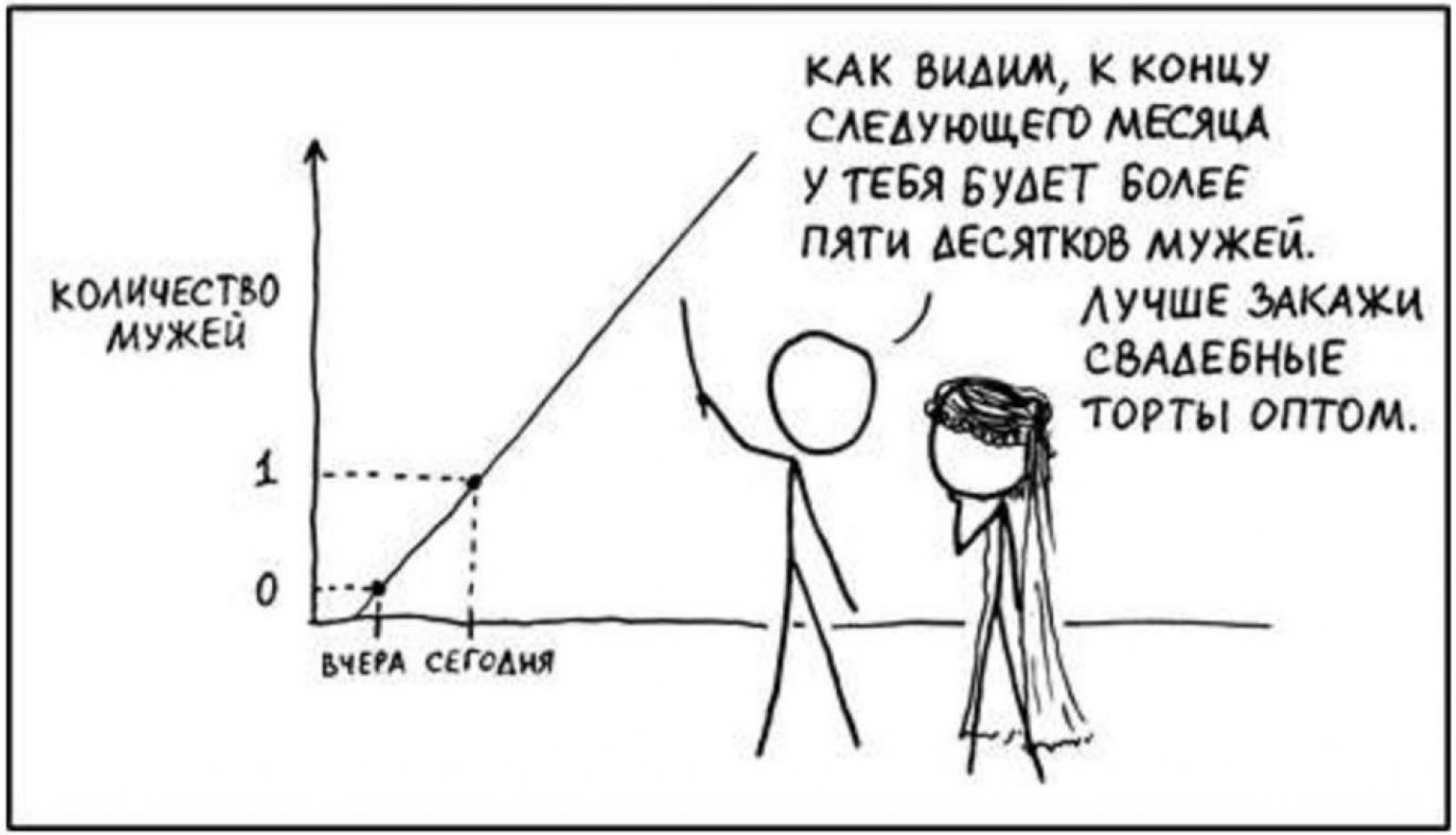
$$\hat{Y}_i = 47.67 + 0.73X_i$$

Предсказания:

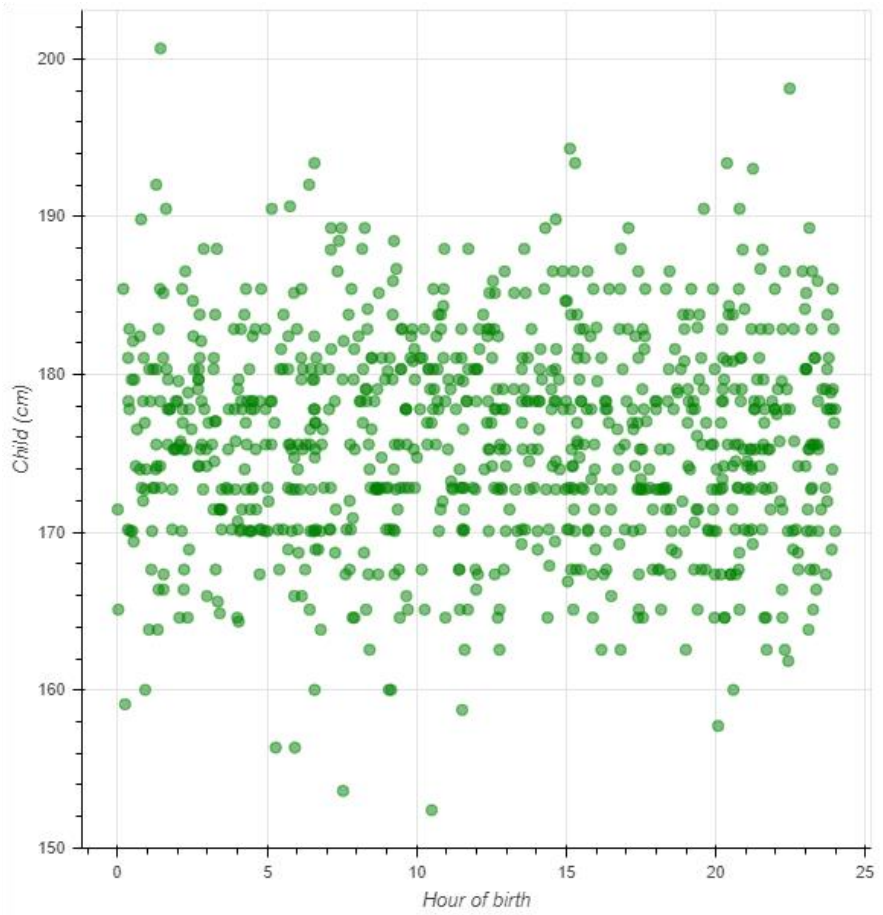
$$X_i = 150\text{см}, \hat{Y}_i = 157.17\text{см}$$

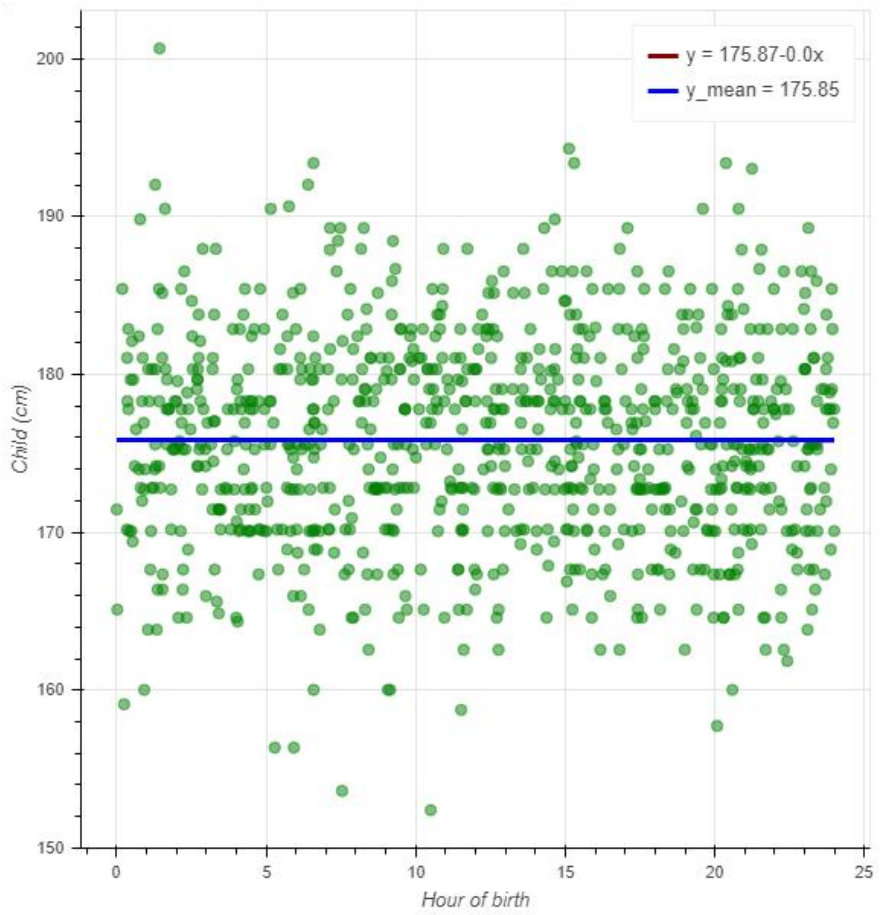
$$X_i = 190\text{см}, \hat{Y}_i = 186.37\text{см}$$

МОЁ ХОББИ: ЭКСТРАПОЛИРОВАТЬ



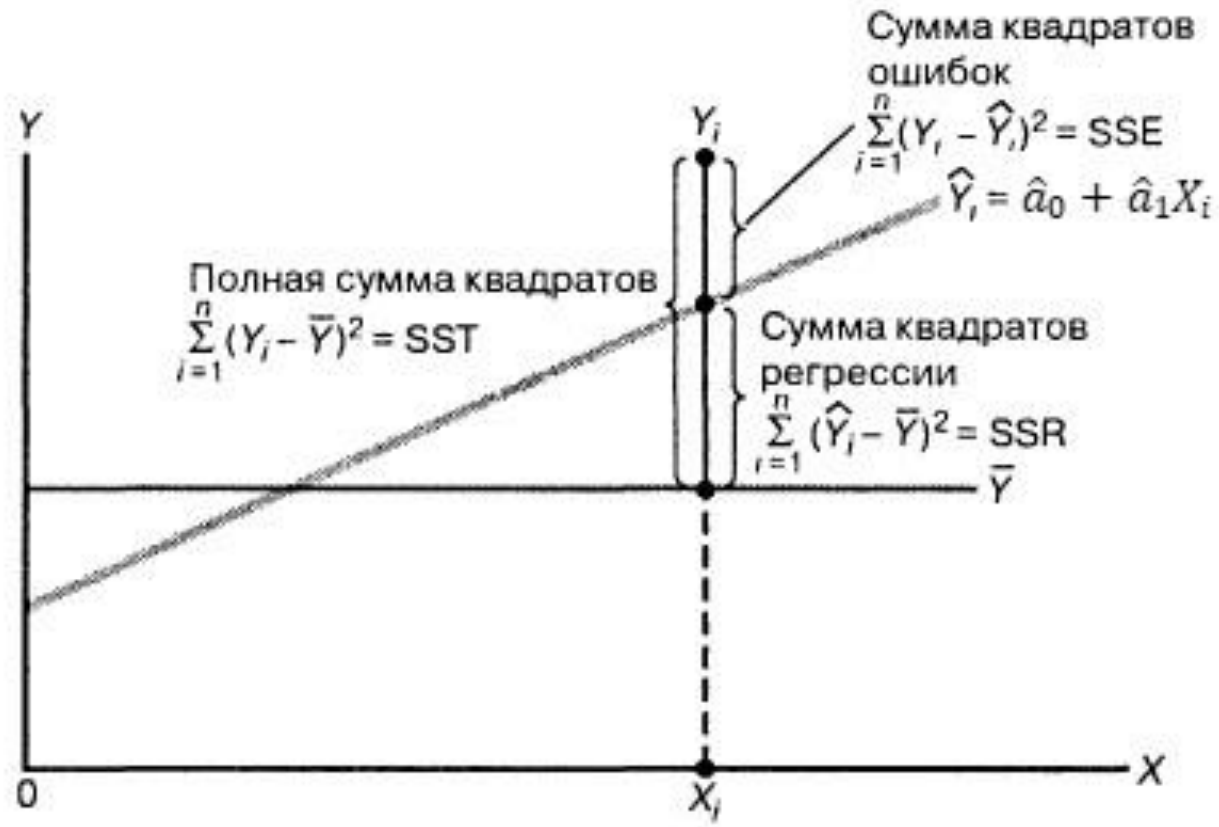
Случай, когда X и Y независимы

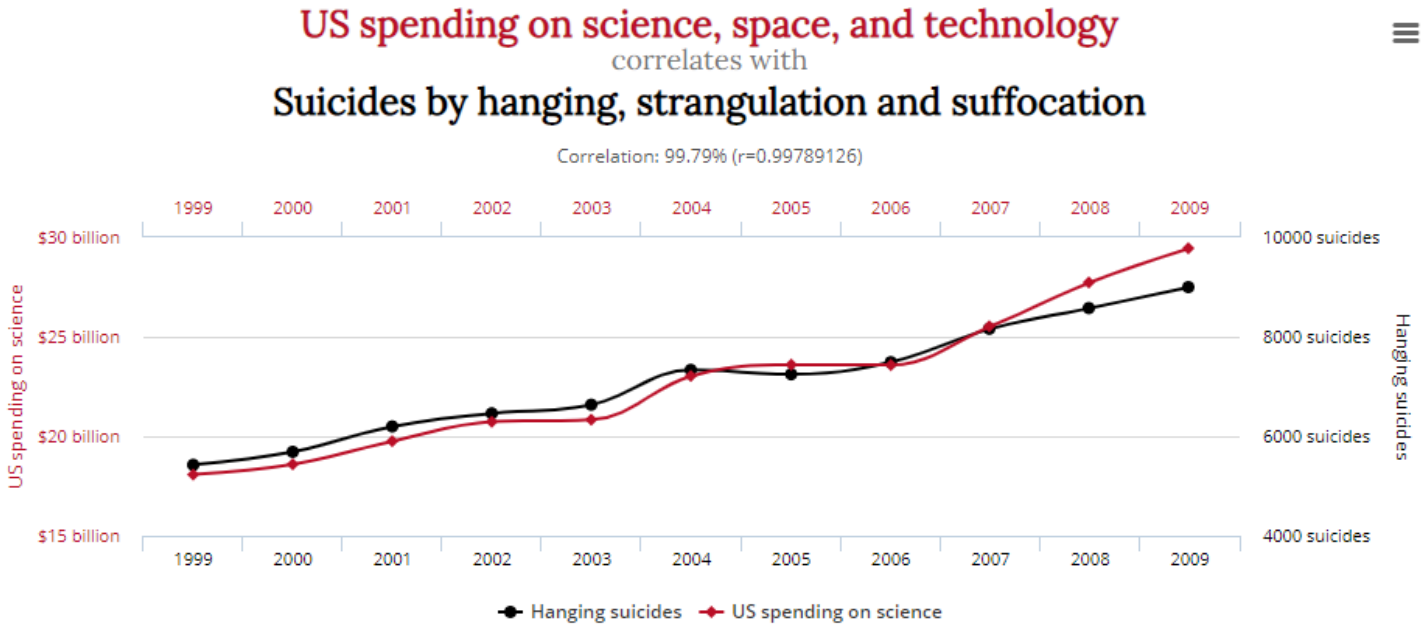




Коэффициент детерминации R^2

- $SST = \sum(Y_i - \bar{Y})^2$
- $SSE = \sum\left(Y_i - \hat{Y}_i\right)^2$
- $R^2 = 1 - \frac{SSE}{SST}$
- $R^2 \in [0; 1]$

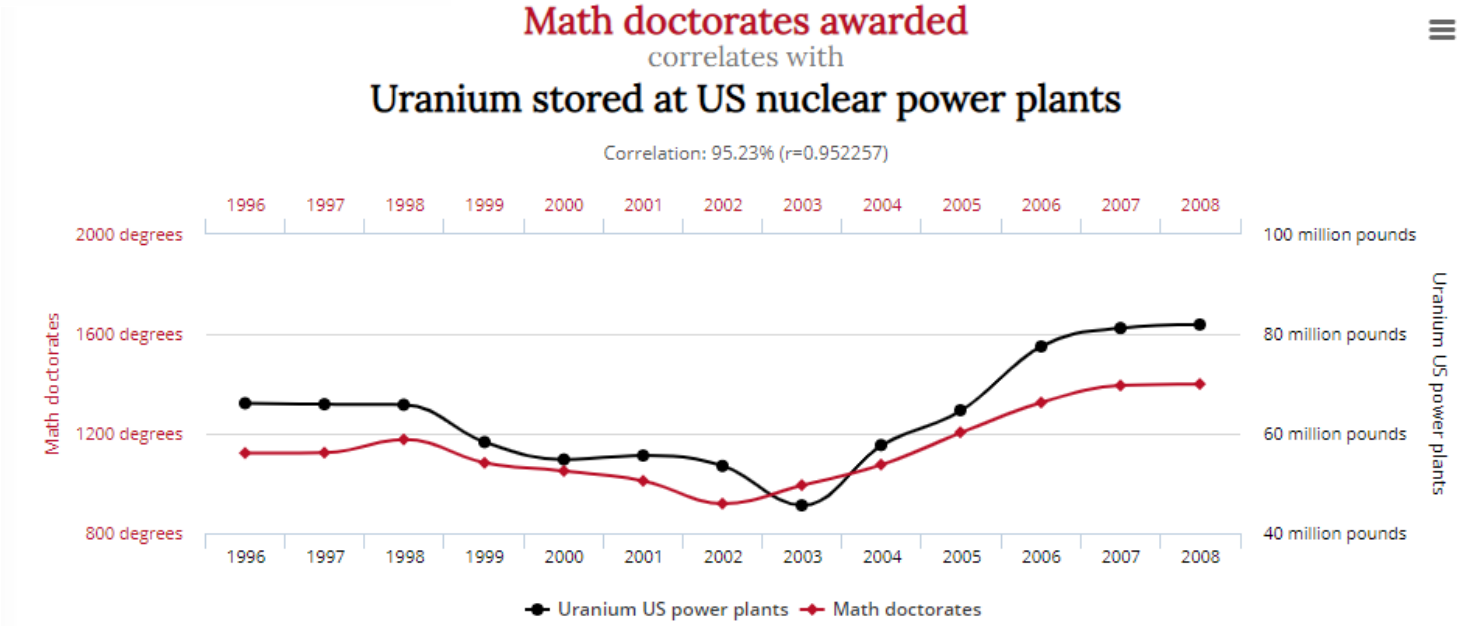




Data sources: U.S. Office of Management and Budget and Centers for Disease Control & Prevention

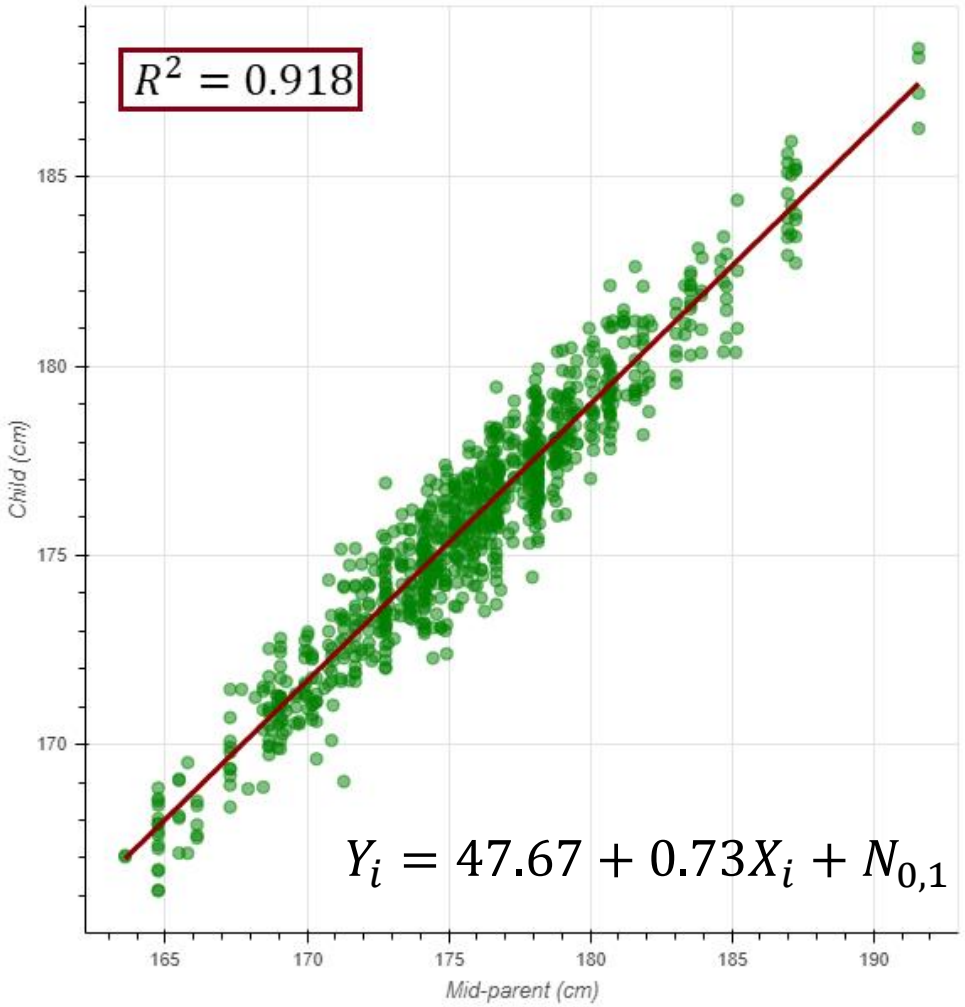
В случае парной линейной регрессии коэффициент детерминации равен квадрату коэффициента корреляции

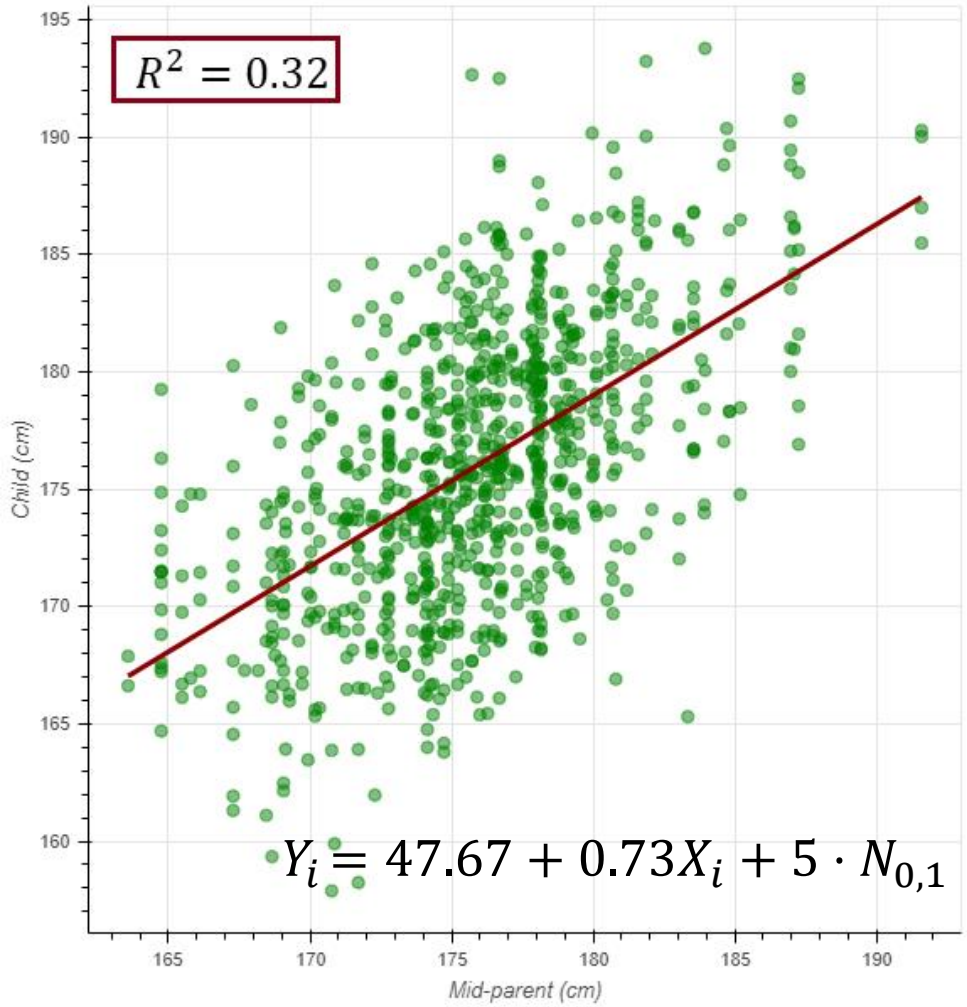
На графиках показаны не связанные друг с другом величины, имеющие при этом высокий коэффициент корреляции



Data sources: National Science Foundation and Dept. of Energy

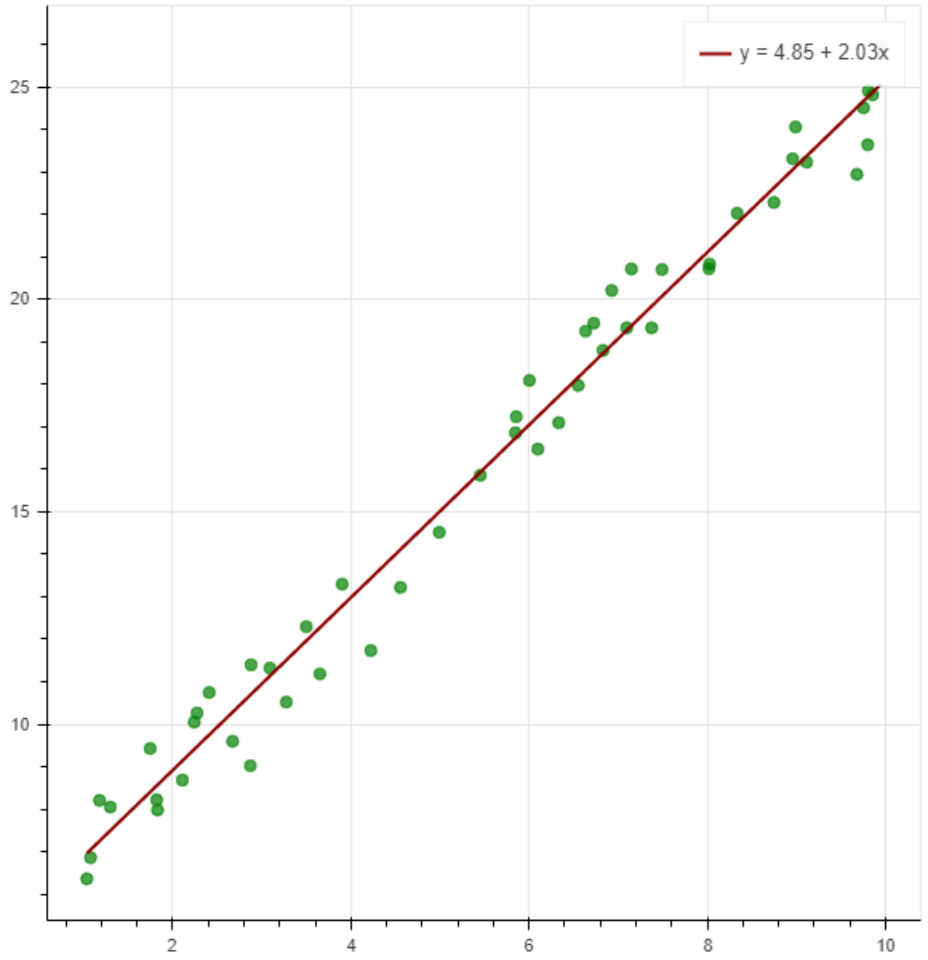
При большом разбросе данных R^2 будет низким даже для очень хорошей модели

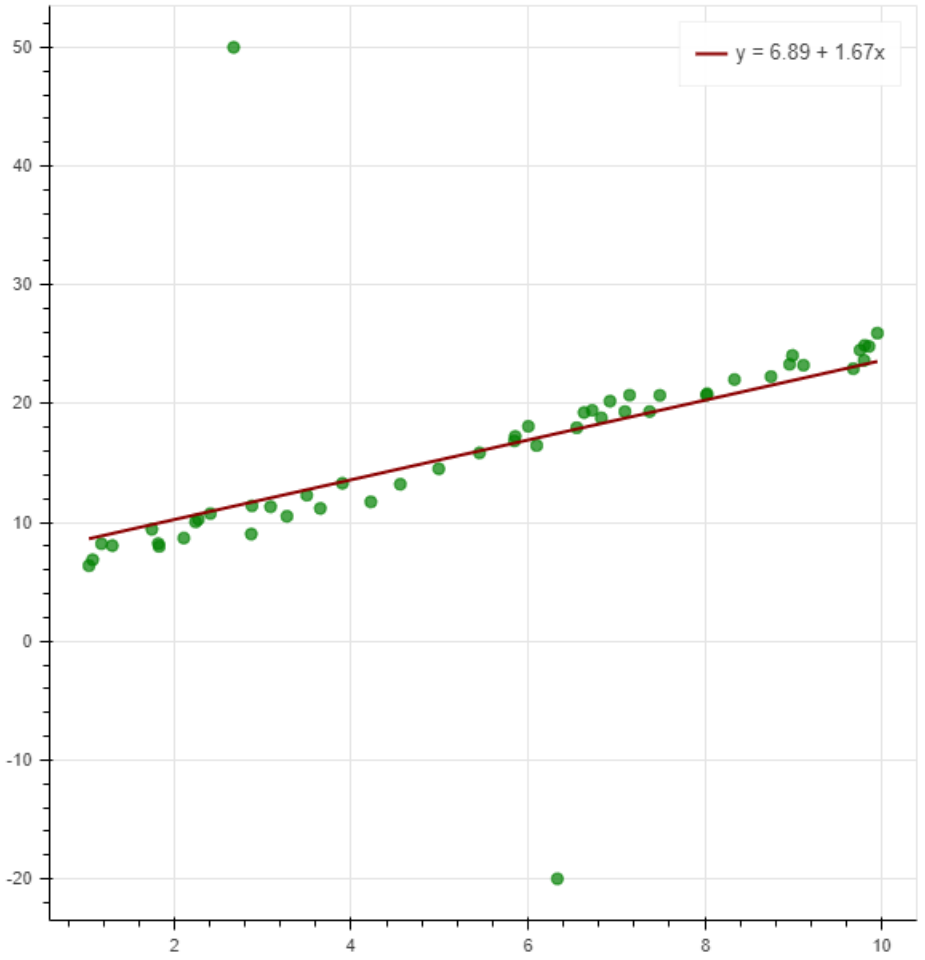




Выбросы влияют на R^2 и на оценки МНК

$$Y_i = 5 + 2X_i + N_{0,1}$$





Множественная линейная регрессия

- $Y_i = \sum a_j X_{ij} + \varepsilon_i$

Перепишем в виде

- $Y = Xa + \varepsilon,$

  где Y, a, ε – вектор-столбцы, X – матрица.

Тогда

- $SSE = (Y - Xa)^T (Y - Xa)$
- $\hat{a}_{\text{МНК}} = (X^T X)^{-1} X^T Y$