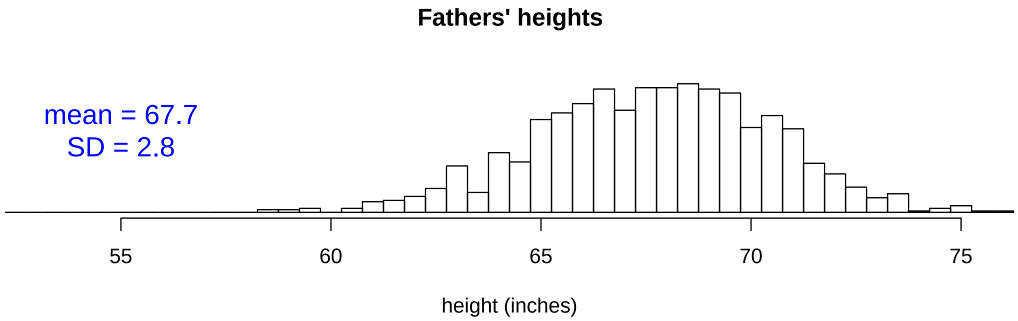
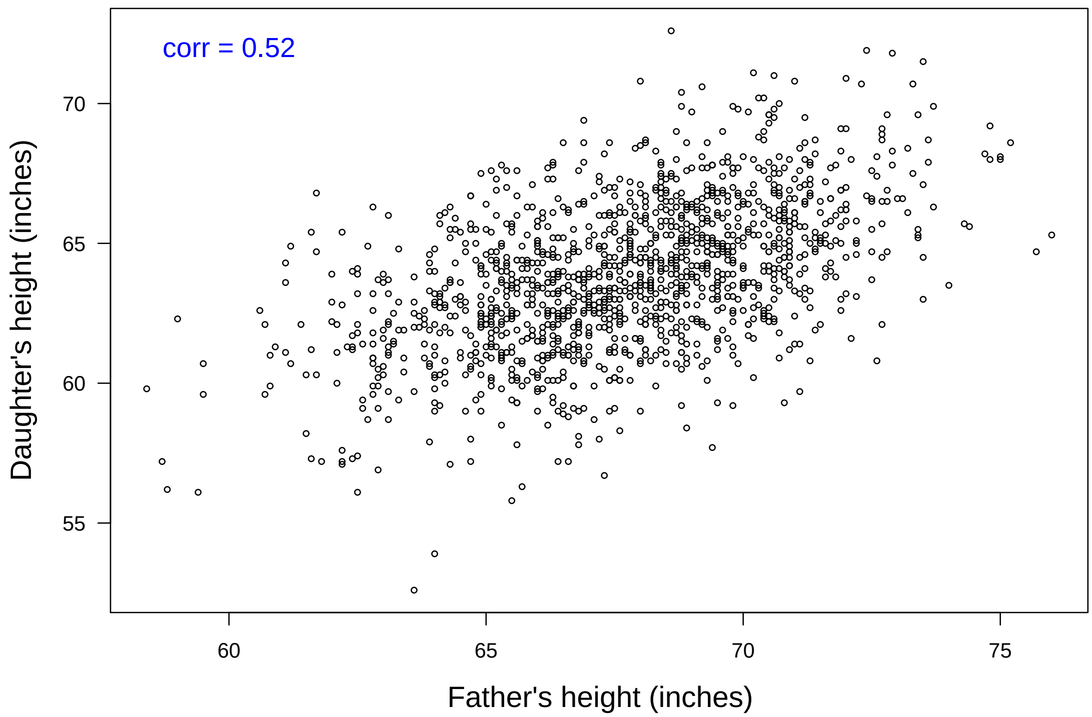
Fathers’ and daughters’ heights



Pearson and Lee (1906) Biometrika 2:357-462

1376 pairs

Fathers’ and daughters’ heights



Pearson and Lee (1906) Biometrika 2:357-462

1376 pairs

Covariance and correlation

Let X and Y be random variables with

$\mu_X = E(X)$, $\mu_Y = E(Y)$, $\sigma_X = SD(X)$, $\sigma_Y = SD(Y)$

For example, sample a father/daughter pair and let

X = the father's height and Y = the daughter's height.

Covariance

$$\text{cov}(X,Y) = E\{(X - \mu_X)\,(Y - \mu_Y)\}$$

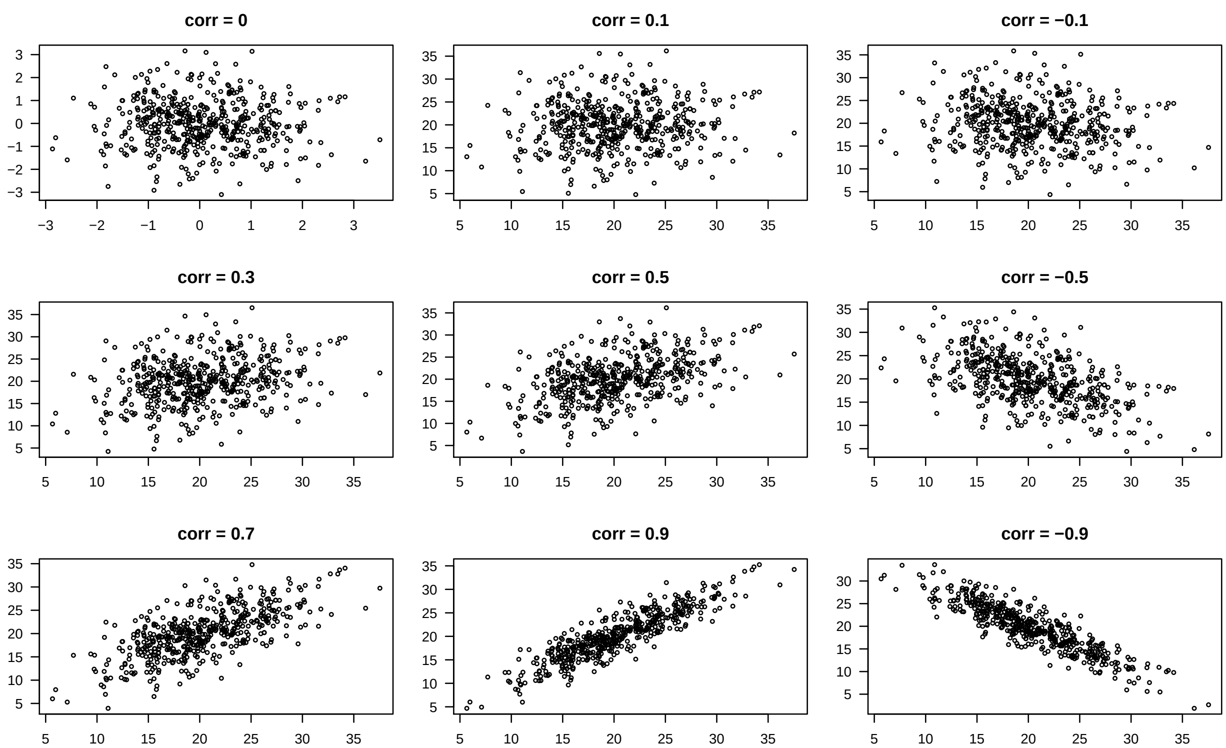
cov(X,Y) can be any real number.

Correlation

$$\text{cor}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

$$-1 \leq \text{cor}(X, Y) \leq 1$$

Examples



Estimated correlation

Consider n pairs of data: $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$

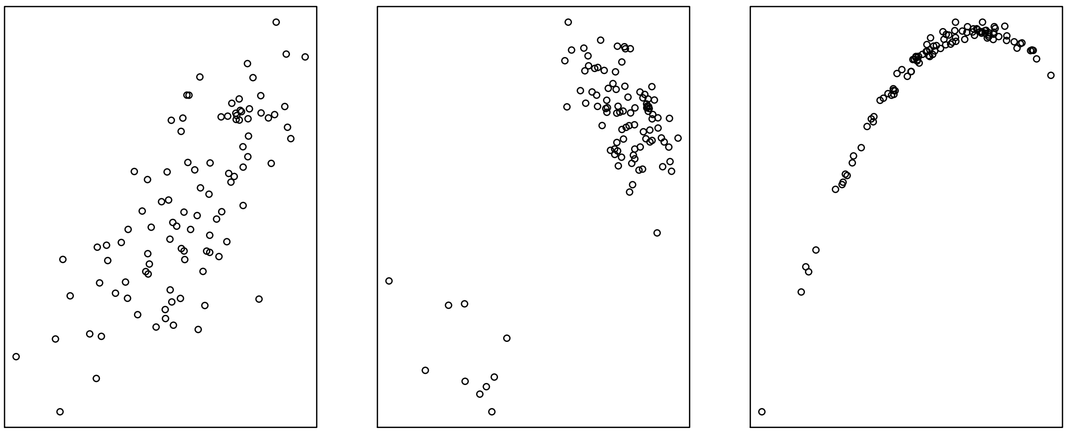
We consider these as independent draws from some bivariate distribution.

We estimate the correlation in the underlying distribution by:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

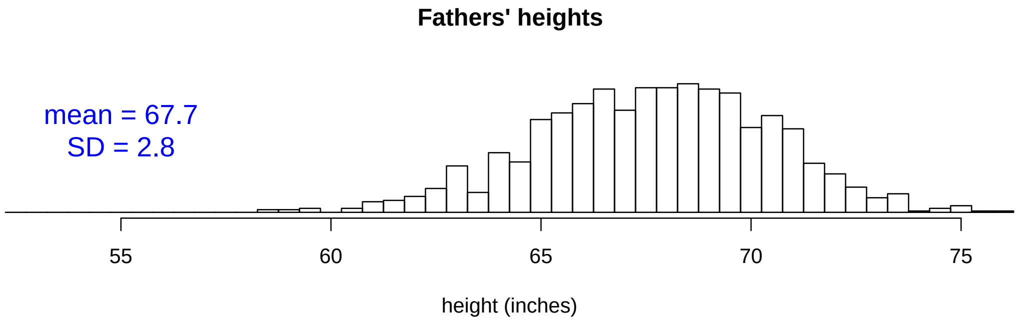
This is sometimes called the correlation coefficient.

Correlation measures linear association



All three plots have correlation ≈ 0.7!

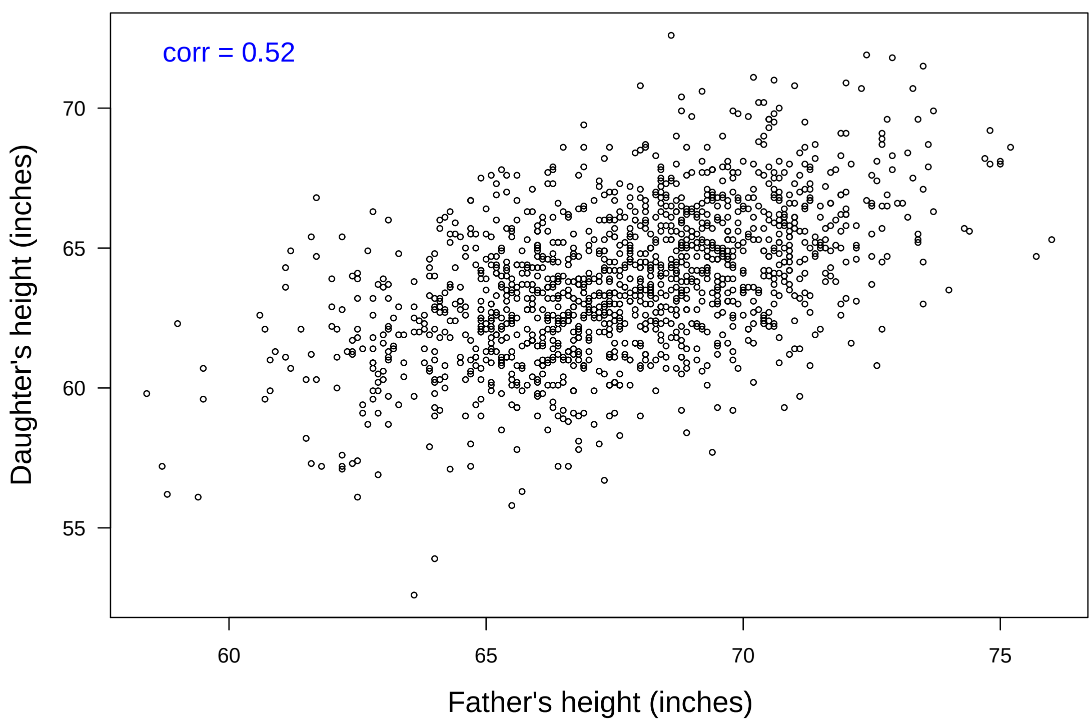
Fathers' and daughters' heights



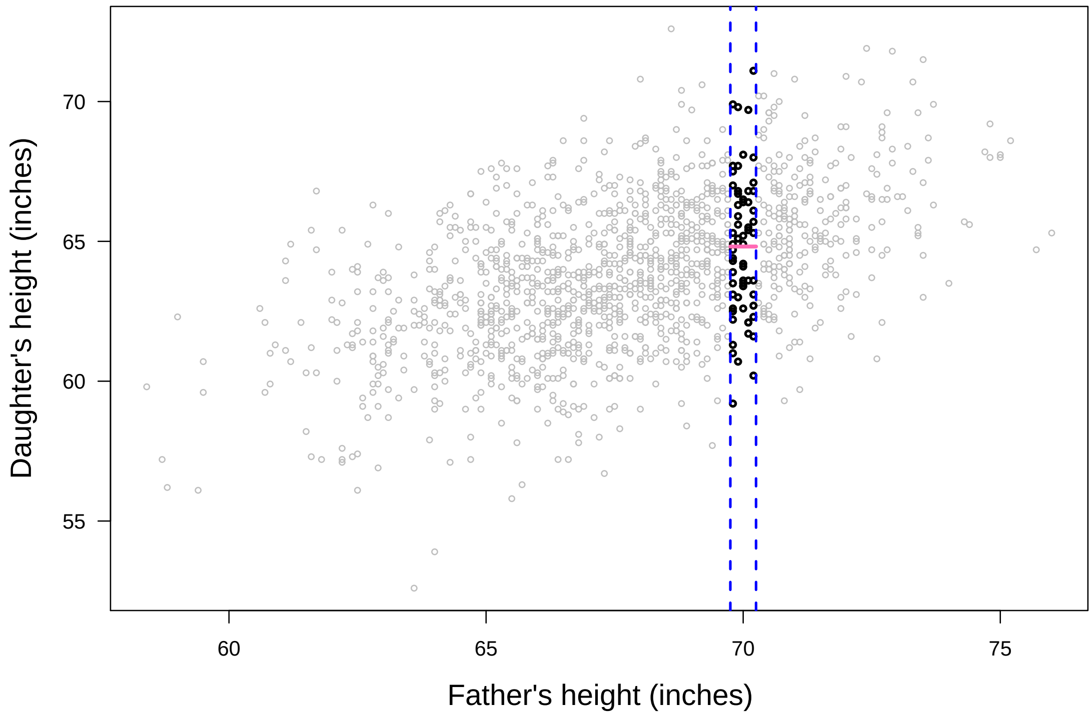
Pearson and Lee (1906) Biometrika 2:357-462

1376 pairs

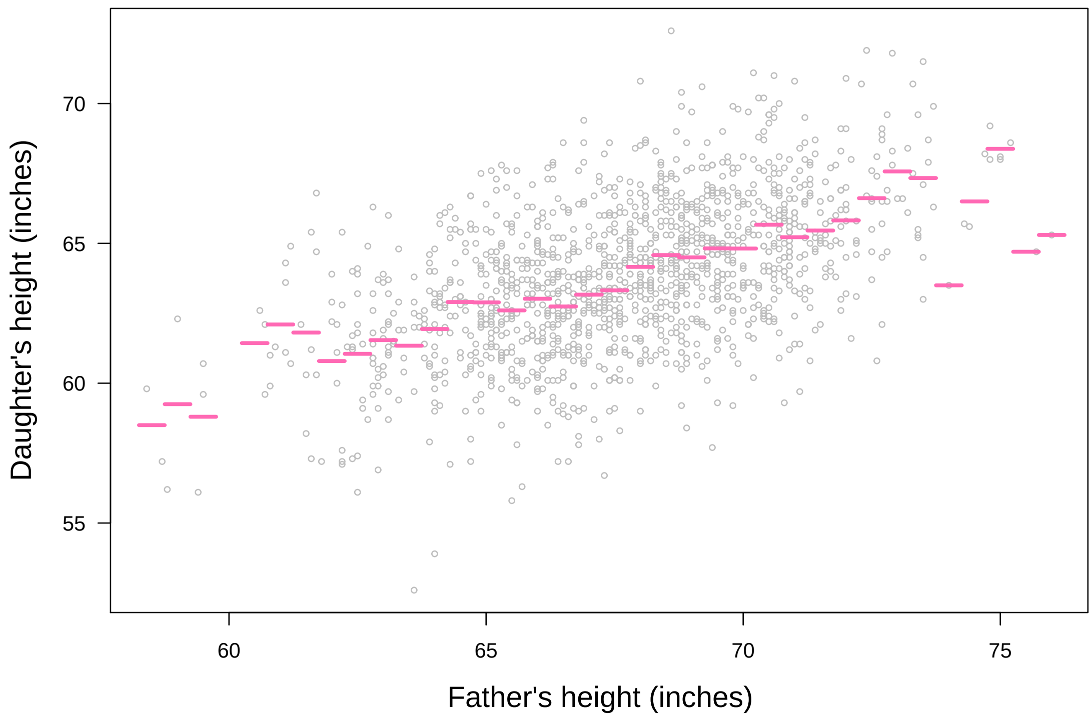
Fathers' and daughters' heights



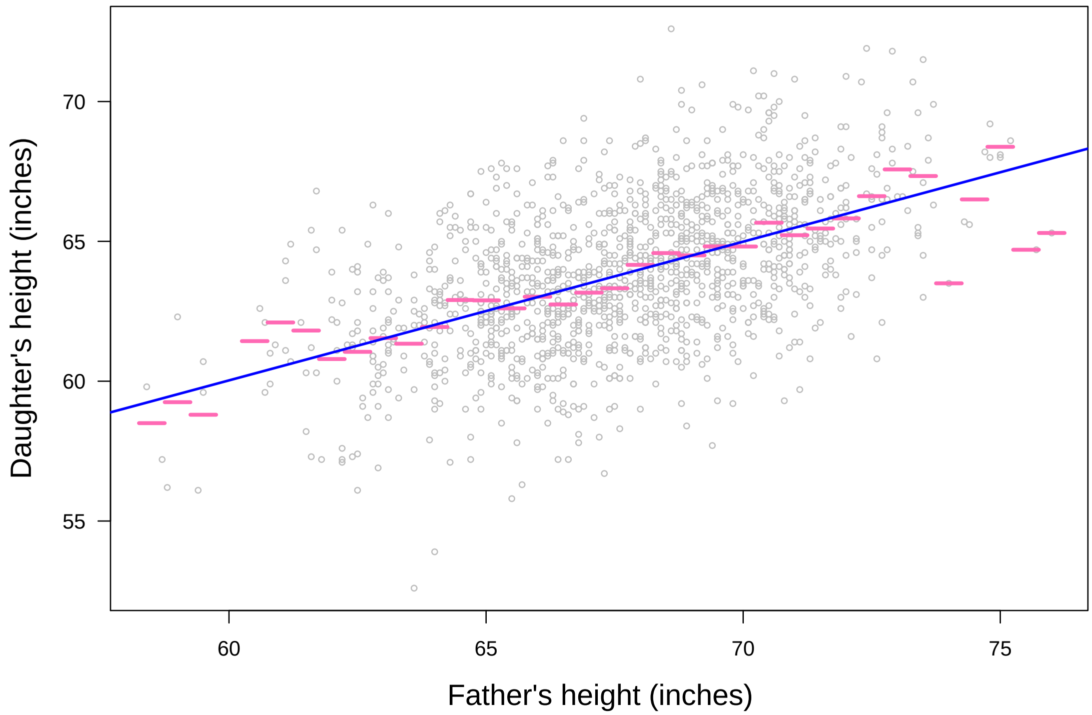
Linear regression



Linear regression

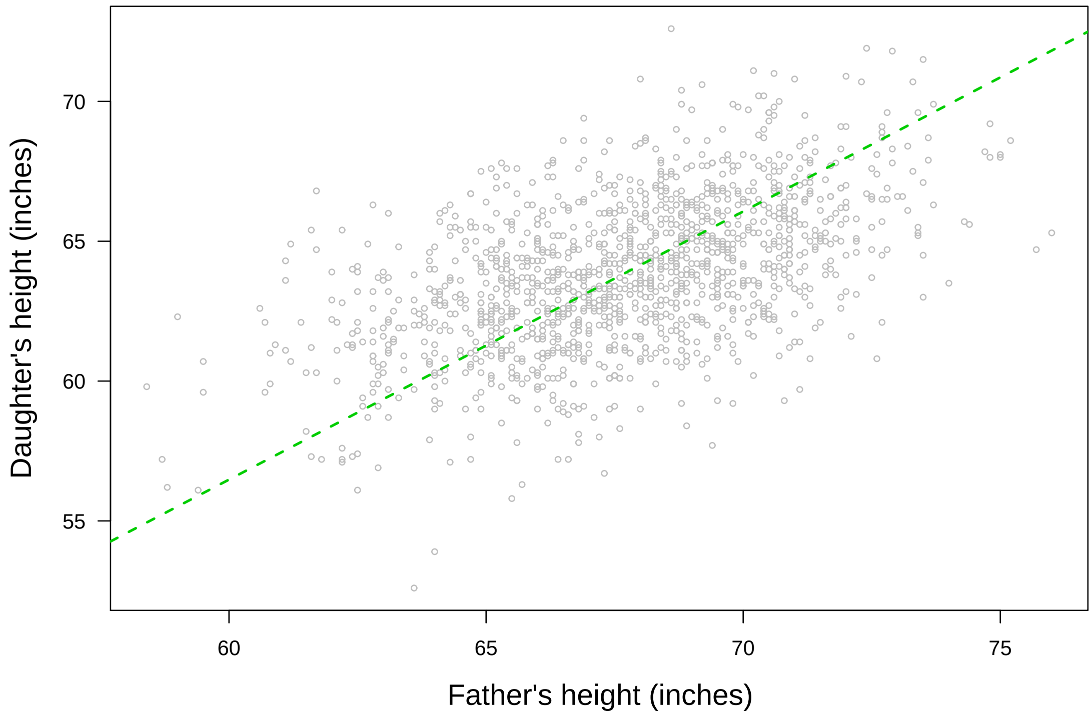


Regression line



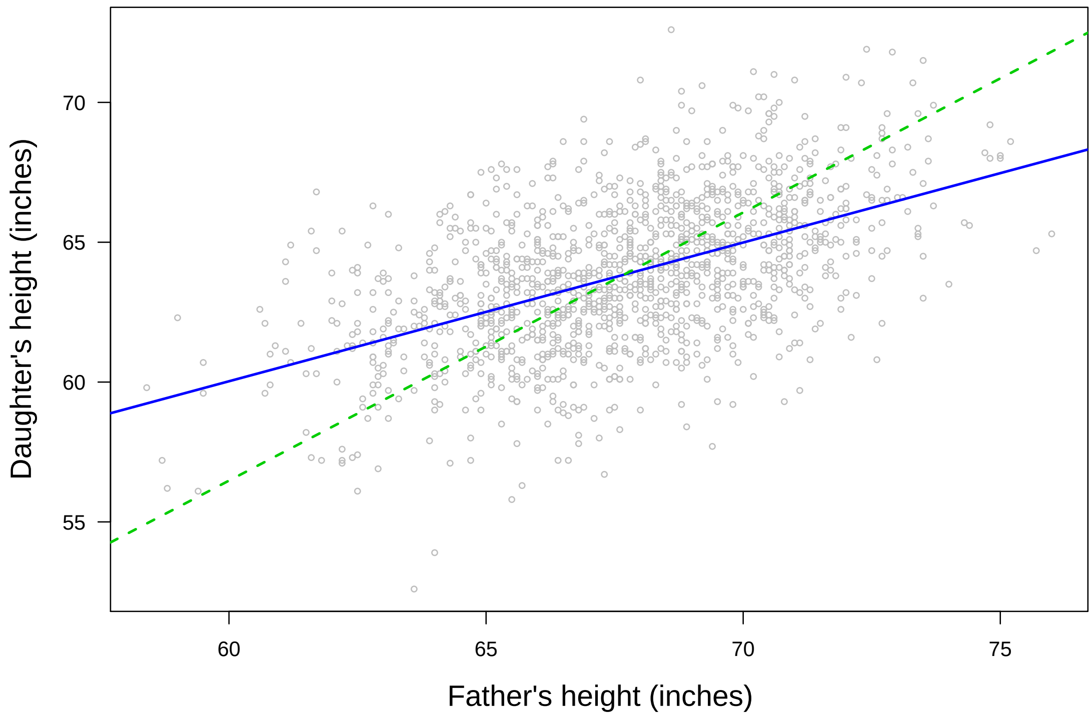
Slope = r $\times$ SD(Y) / SD(X)

SD line



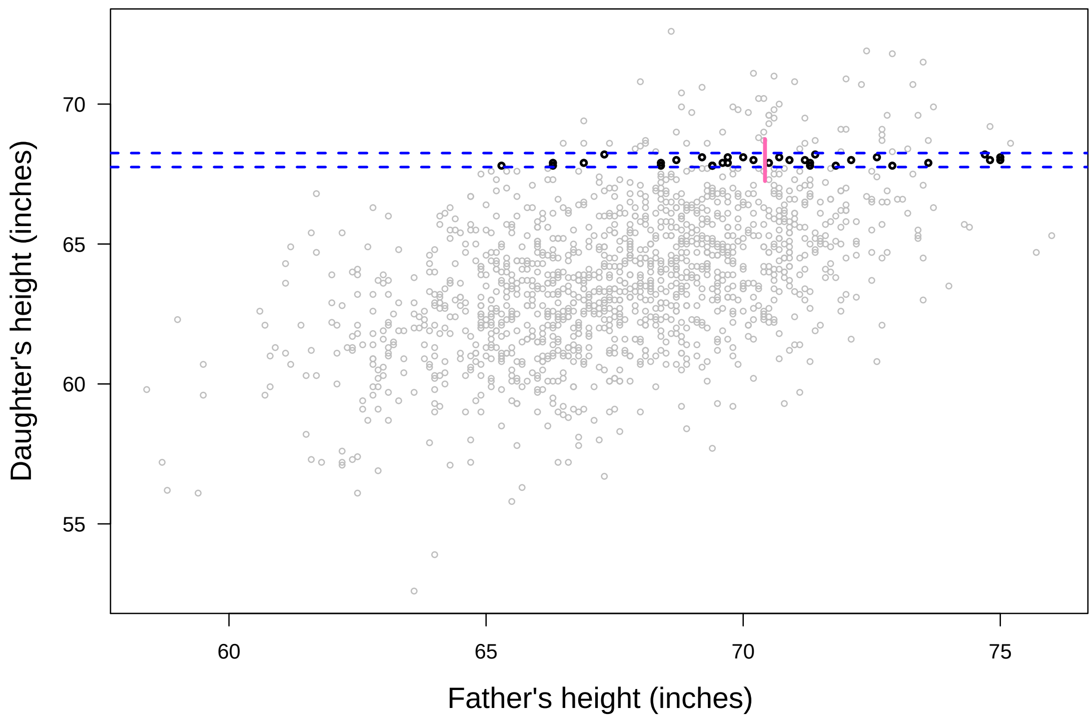
Slope = SD(Y) / SD(X)

SD line vs regression line

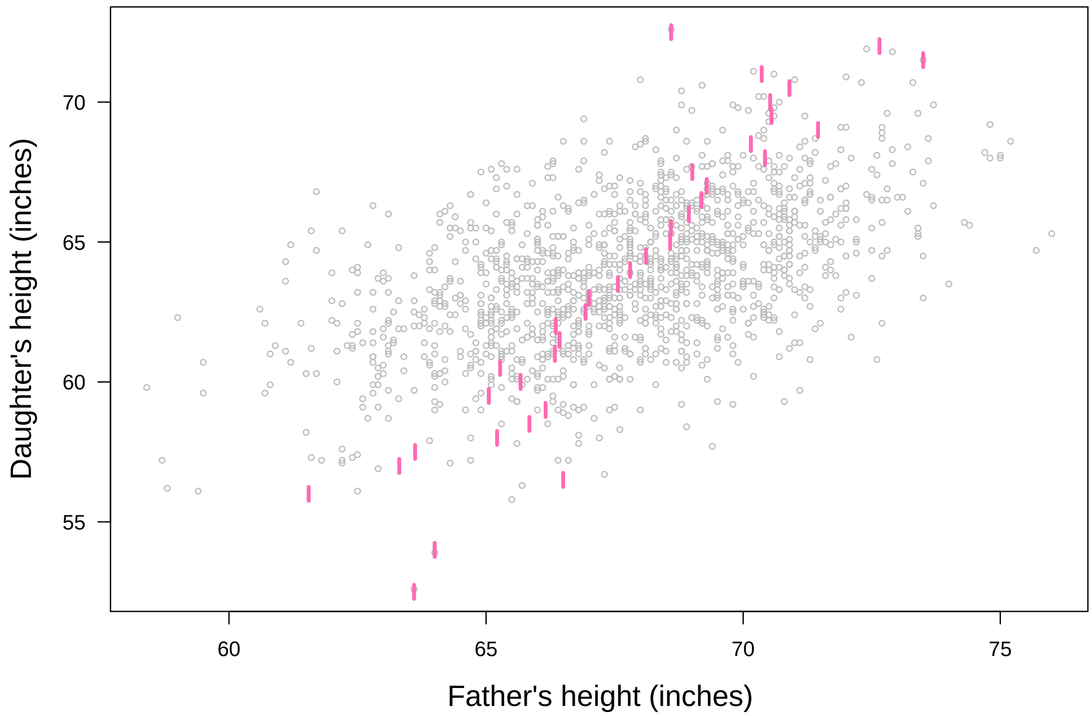


Both lines go through the point ($\bar{X}$, $\bar{Y}$).

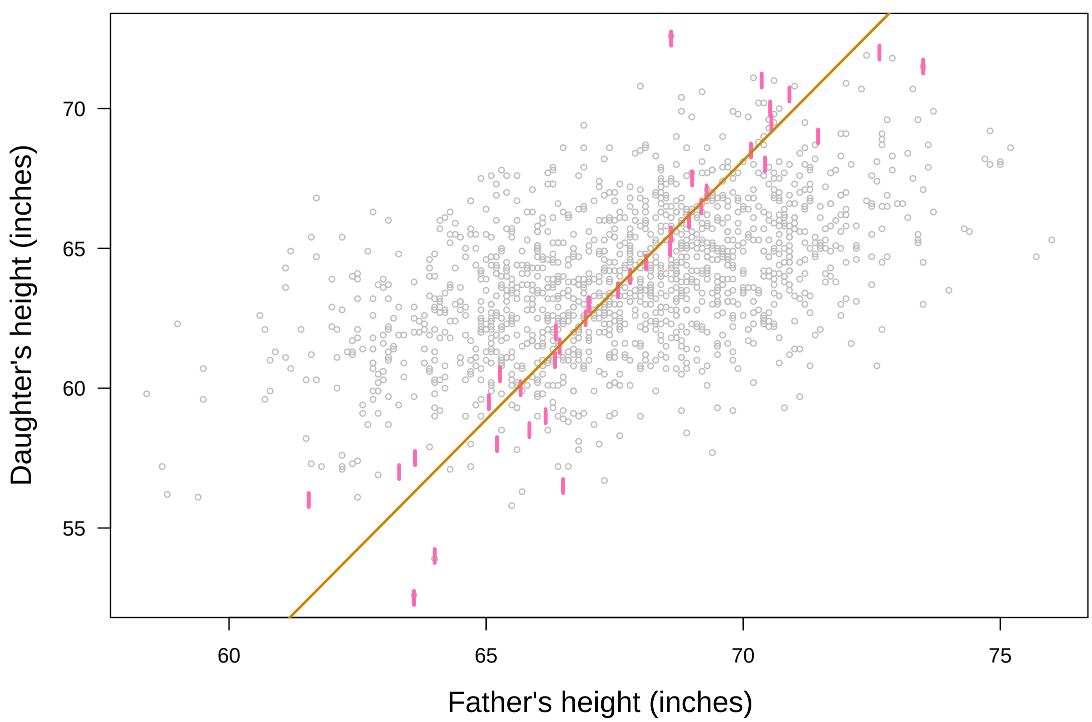
Predicting father's ht from daughter's ht



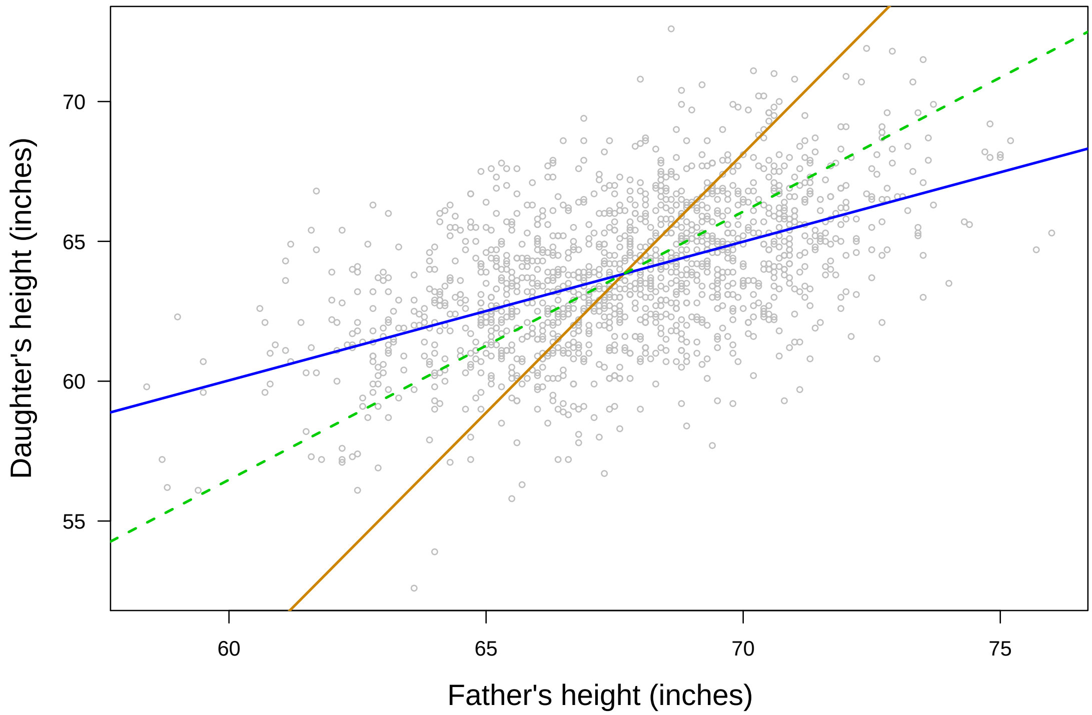
Predicting father's ht from daughter's ht



Predicting father's ht from daughter's ht



There are two regression lines!



The regression lines

Predicting y from x

$$\left(\frac{y-\bar{y}}{s_y}\right) = r \times \left(\frac{x-\bar{x}}{s_x}\right)$$

Predicting x from y

$$\left(\frac{x-\bar{x}}{s_x}\right) = r \times \left(\frac{y-\bar{y}}{s_y}\right)$$

The regression effect

- Tall fathers have, on average, daughters who are not so tall.
- Short fathers have, on average, daughters who are not so short.
- Tall daughters have, on average, fathers who are not so tall.
- Short daughters have, on average, fathers who are not so short.

The regression fallacy

The regression fallacy: Ascribing important meaning to the regression effect.

Example: the “sophomore slump”

Also think:

Exam grade = skill + luck