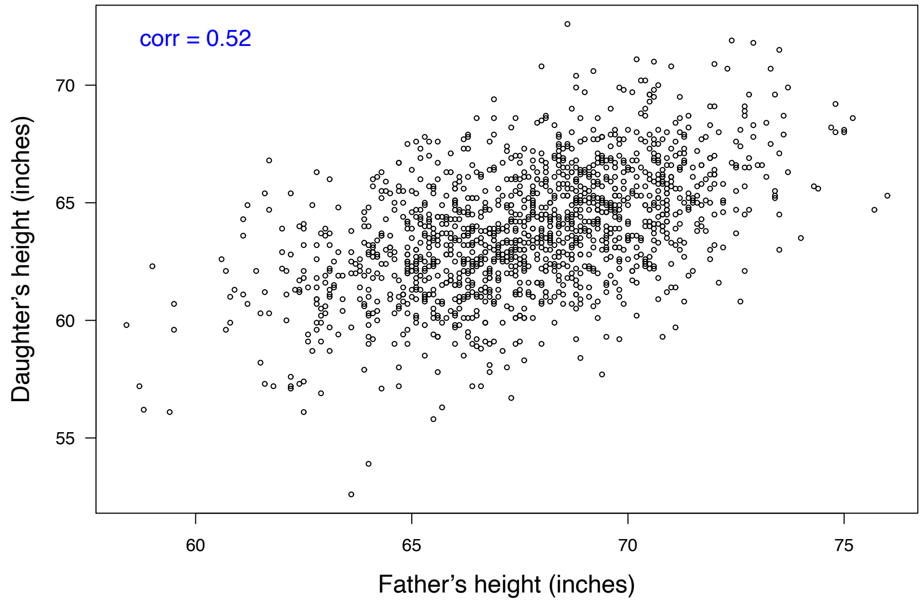
Regression and Correlation

Correlation vs Regression

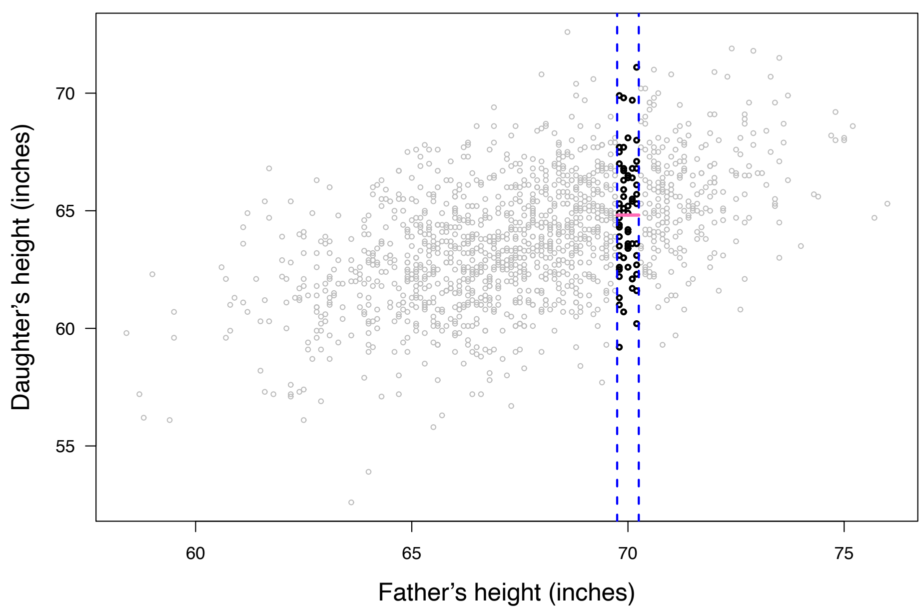
- In a correlation setting we try to determine whether two random variables vary together (covary).
- There is no ordering between those variables, and we do not try to explain one of the variables as a function of the other.

- In regression settings we describe the dependence of one variable on the other variable.
- There is an ordering of the variables, often called the dependent variable and the independent variable.

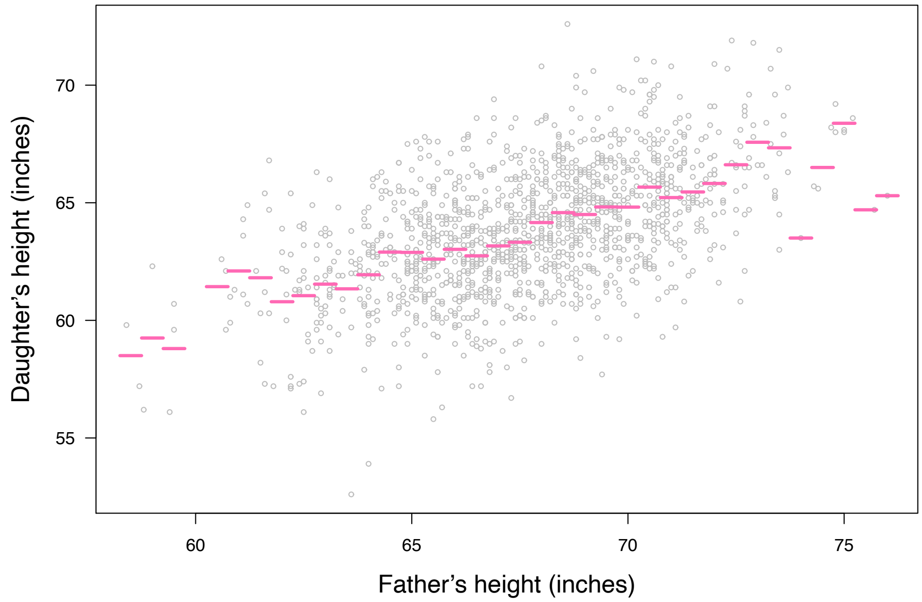
Fathers’ and daughters’ heights



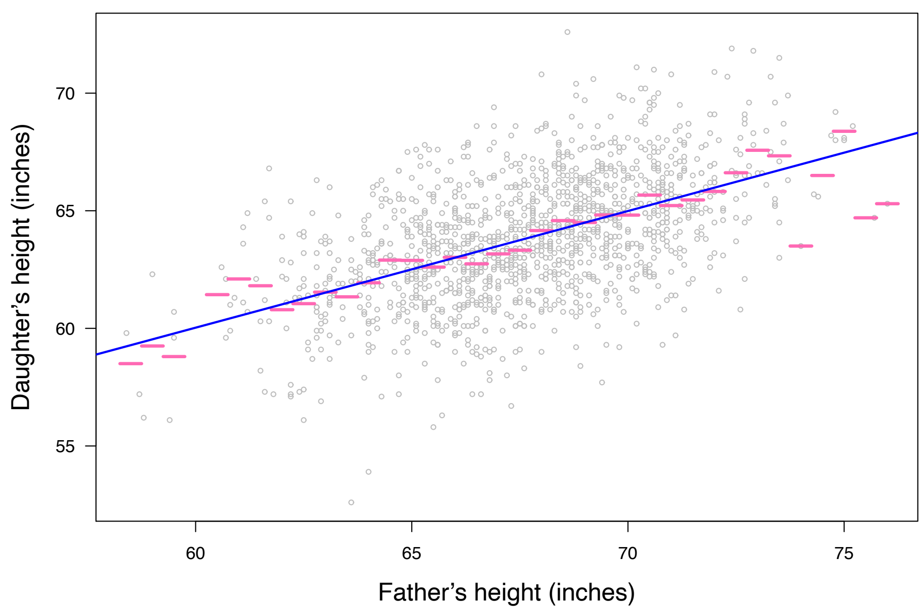
Linear regression



Linear regression

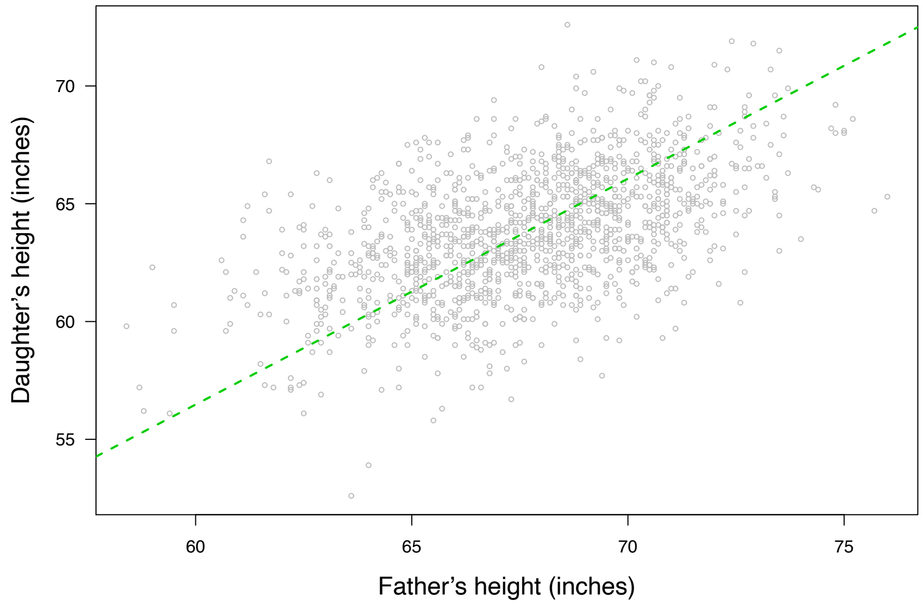


Regression line



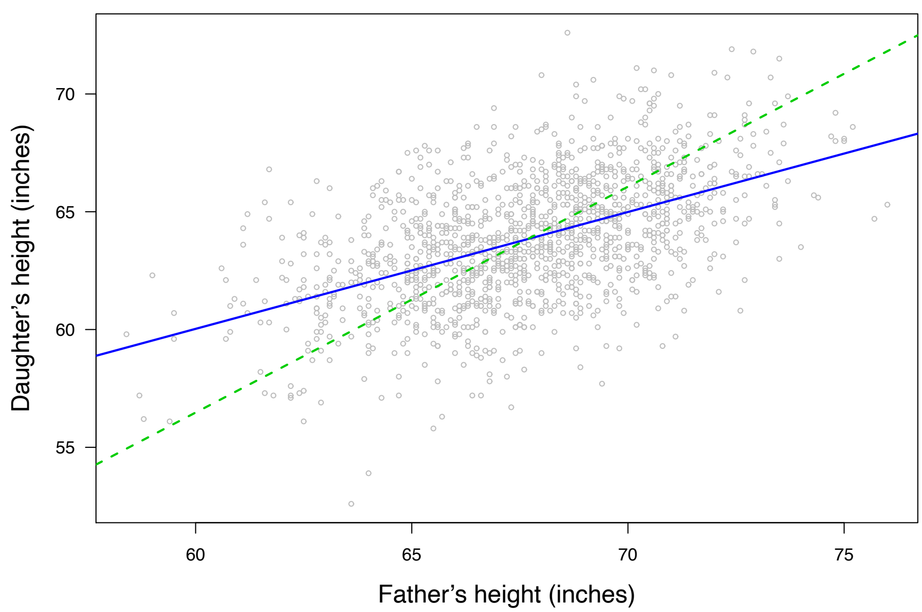
⟶ Slope = r × SD(Y) / SD(X)

SD line



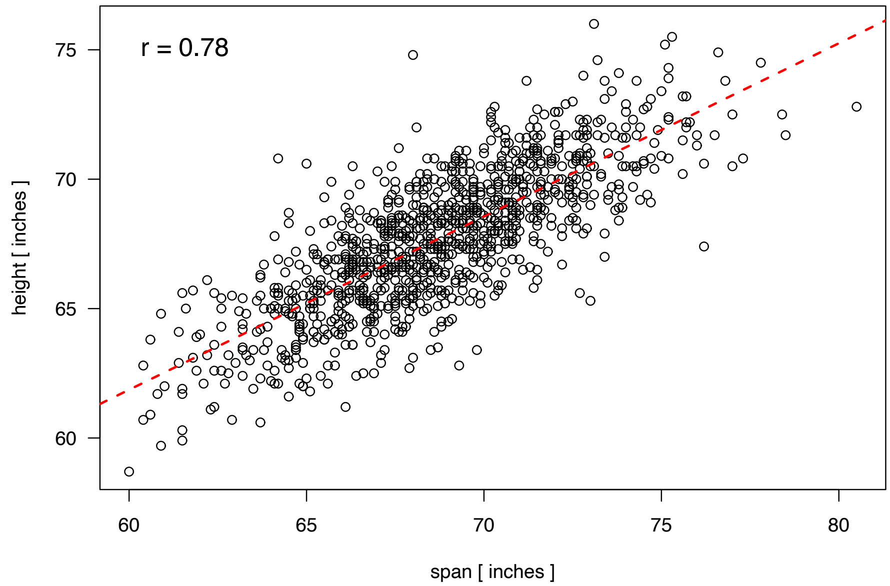
⟶ Slope = SD(Y) / SD(X)

SD line vs regression line



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

Span and height



Variability

Having no information about x,

Predict y as $\bar{y}$

Typical prediction error: SD(y)

For predicting height, SD(y) $\approx$ 2.73

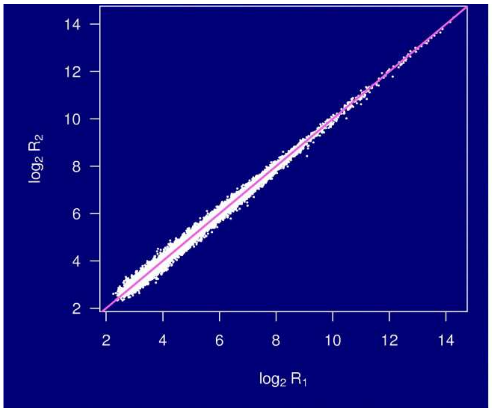
Having been told about x,

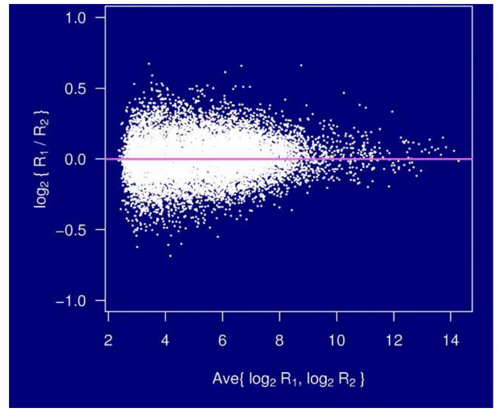
Predict y using the regression line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Typical prediction error: SD(y) $\sqrt{1 - r^2}$

For predicting height from span, SD(y) $\sqrt{1 - r^2} \approx 1.71$

Gene expression





Concordance

If you want to show that two sets of measurements are alike (such as gene expression from two technical replicates of the same sample) use the concordance correlation coefficient.

The concordance correlation between two random variables X and Y is defined as

$$\rho_{\text{CC}}(X,Y) = \frac{2 \times \text{cov}(X,Y)}{\sigma_X^2 + \sigma_Y^2 + (\mu_X - \mu_Y)^2}.$$

Unlike the Pearson correlation coefficient, the concordance correlation is not invariant to changes in location and scale, and assesses the actual agreement between X and Y, rather then their correlation alone.

Correlation..?