Correlation and Regression

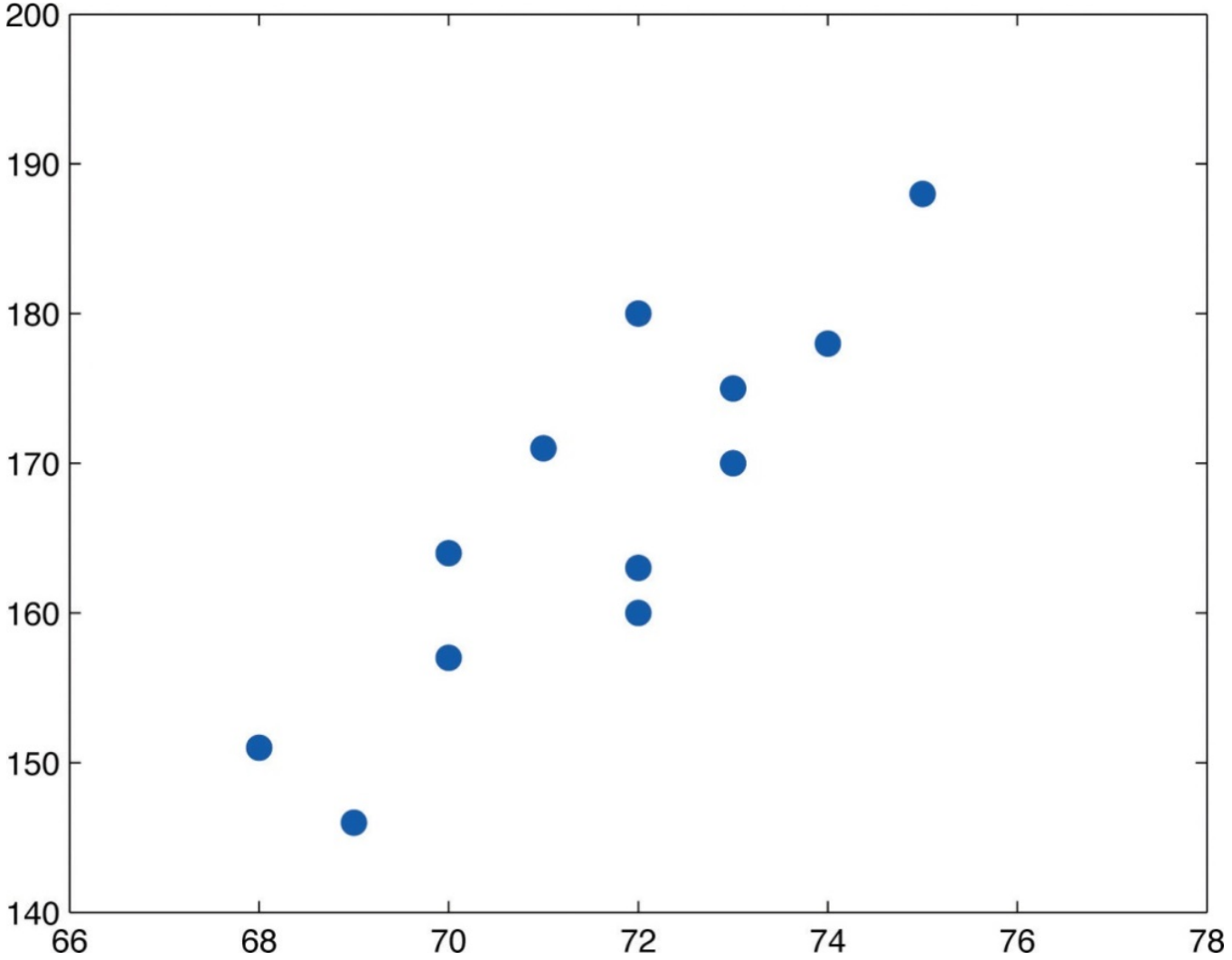
We are curious about the linear relationship between a two variables

There might a relation between

X	and	Y
Predictor	↔	Response
Temperature in C	↔	Temperature in F
Area of a house	↔	Value of the house
Age	↔	Weight

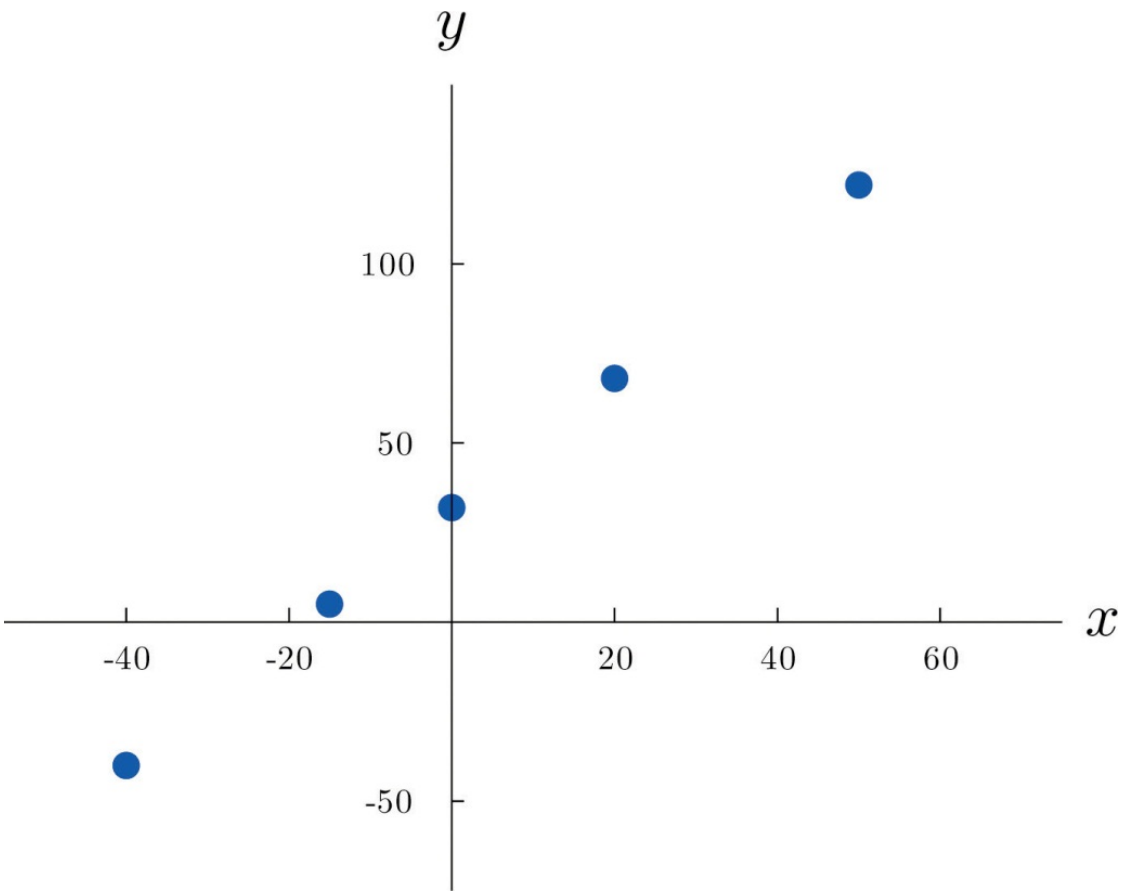
Can x be used to predict y

- How about this relation ?



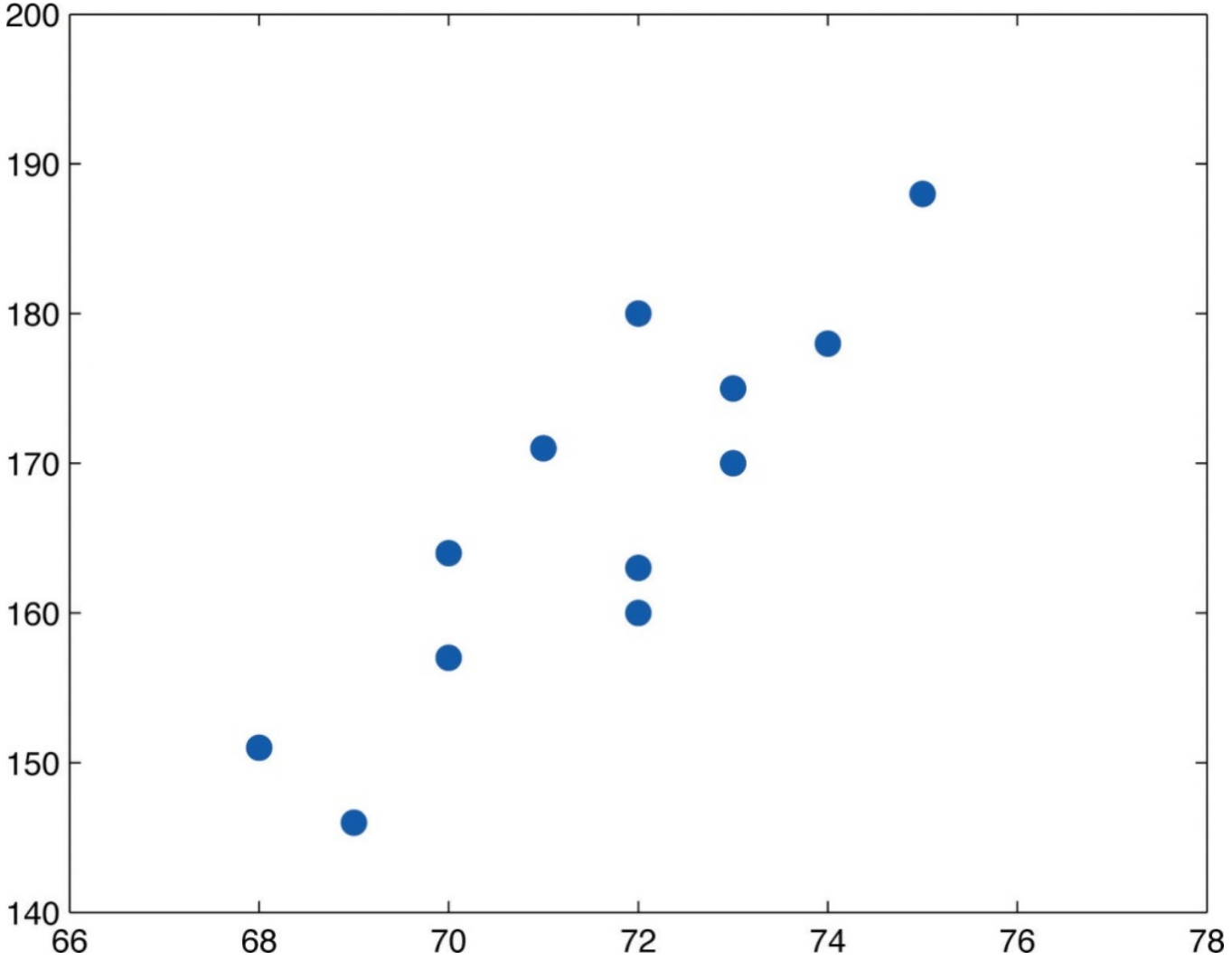
Relation between Celsius and Fahrenheit

- This would be a deterministic one



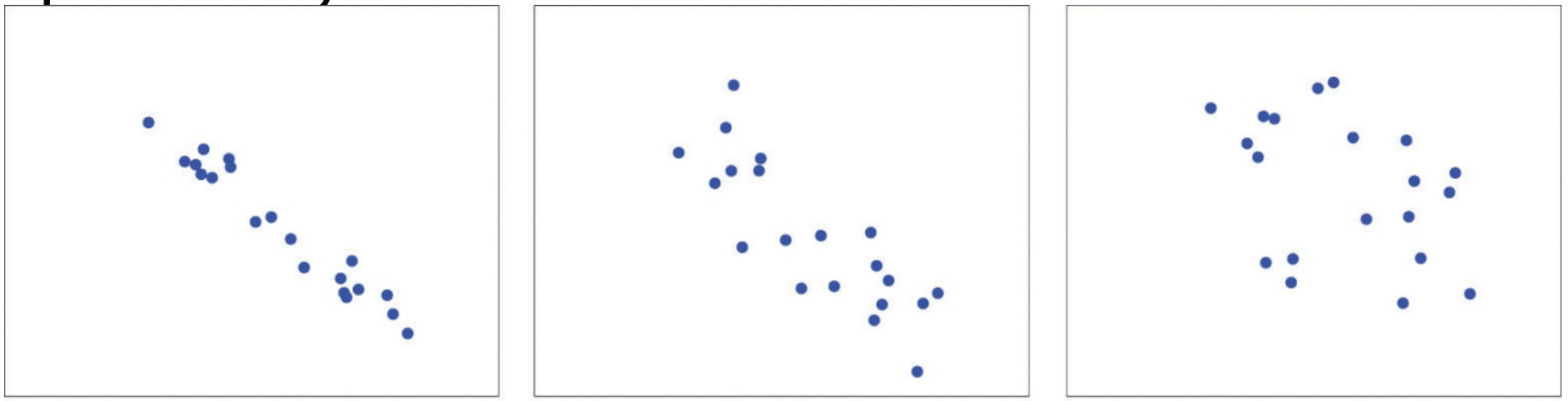
Can x be used to predict y

- How about this relation ?



Examples of correlation

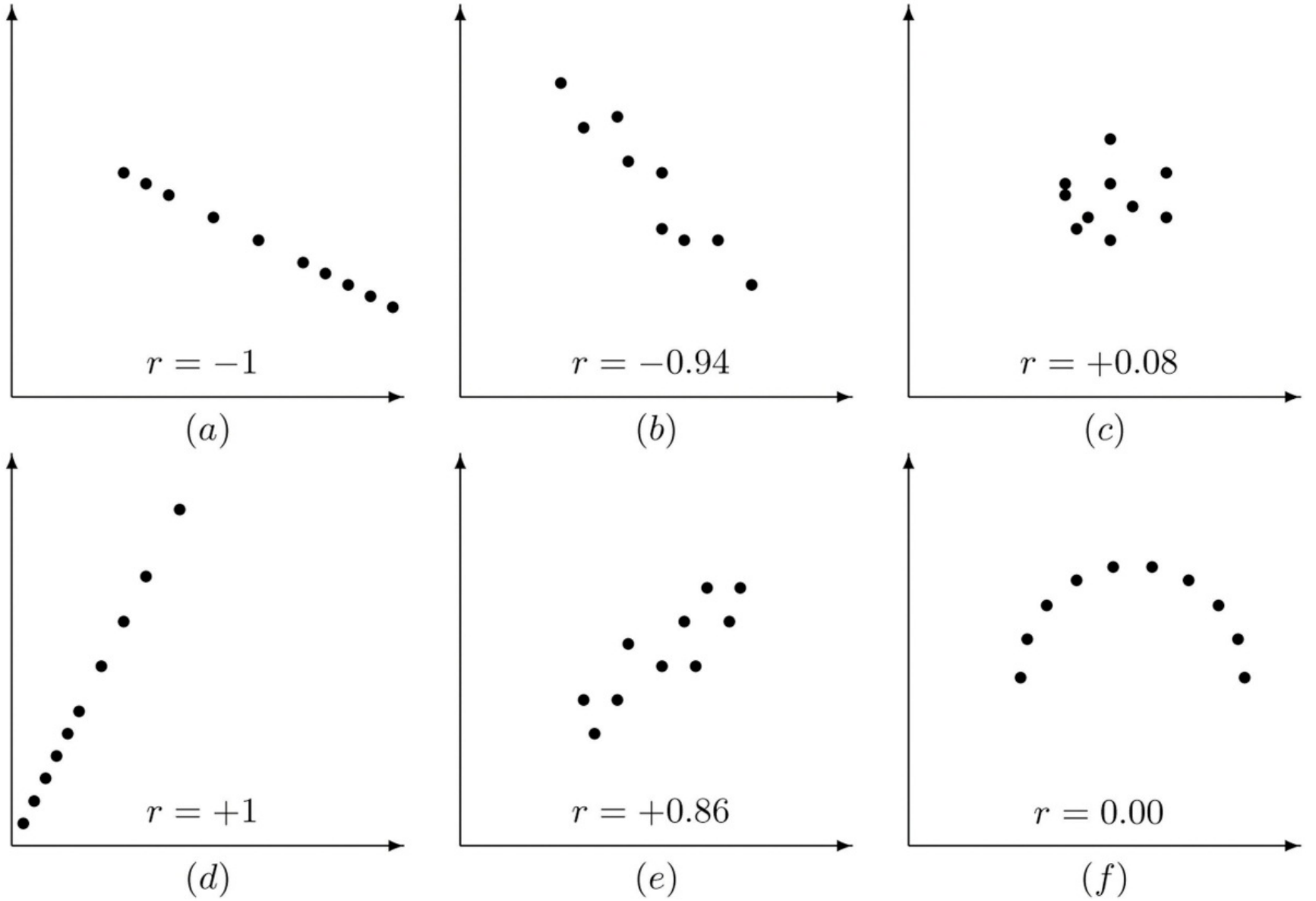
- (a), x could serve as a useful predictor of y , it would be less useful in the situation illustrated in panel (b), and in the situation of panel (c) the linear relationship is so weak as to be practically nonexistent



Correlation

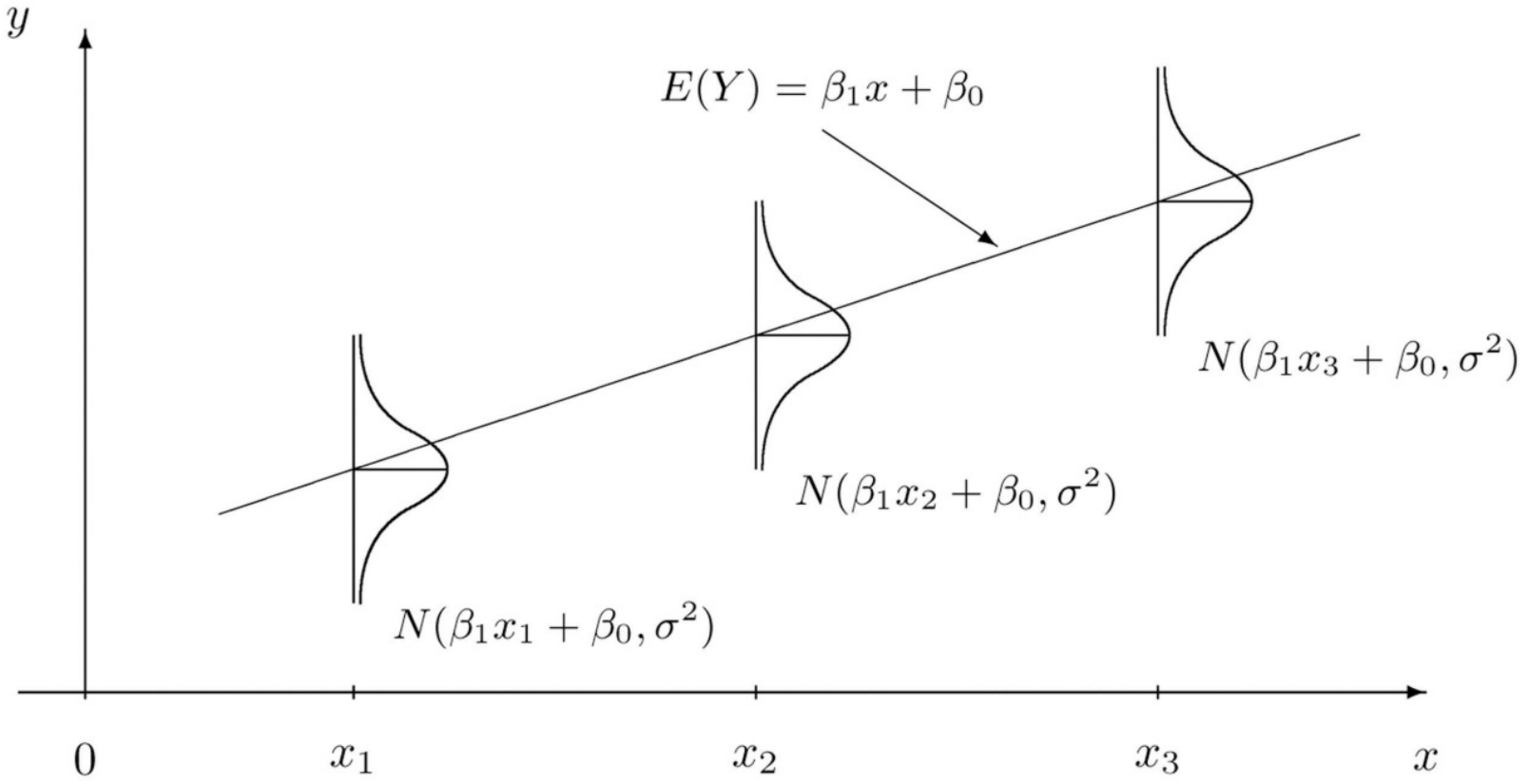
- This linear relationship can also be expressed in terms of correlation.
- The correlation is between -1 and 1
- If correlation < 0 y decreases when x increases
- If correlation > 0 y increases when x increases
- |correlation| is near to 1, the relation is strong
- |correlation| is near to 0, the relation is weak

Examples of correlations



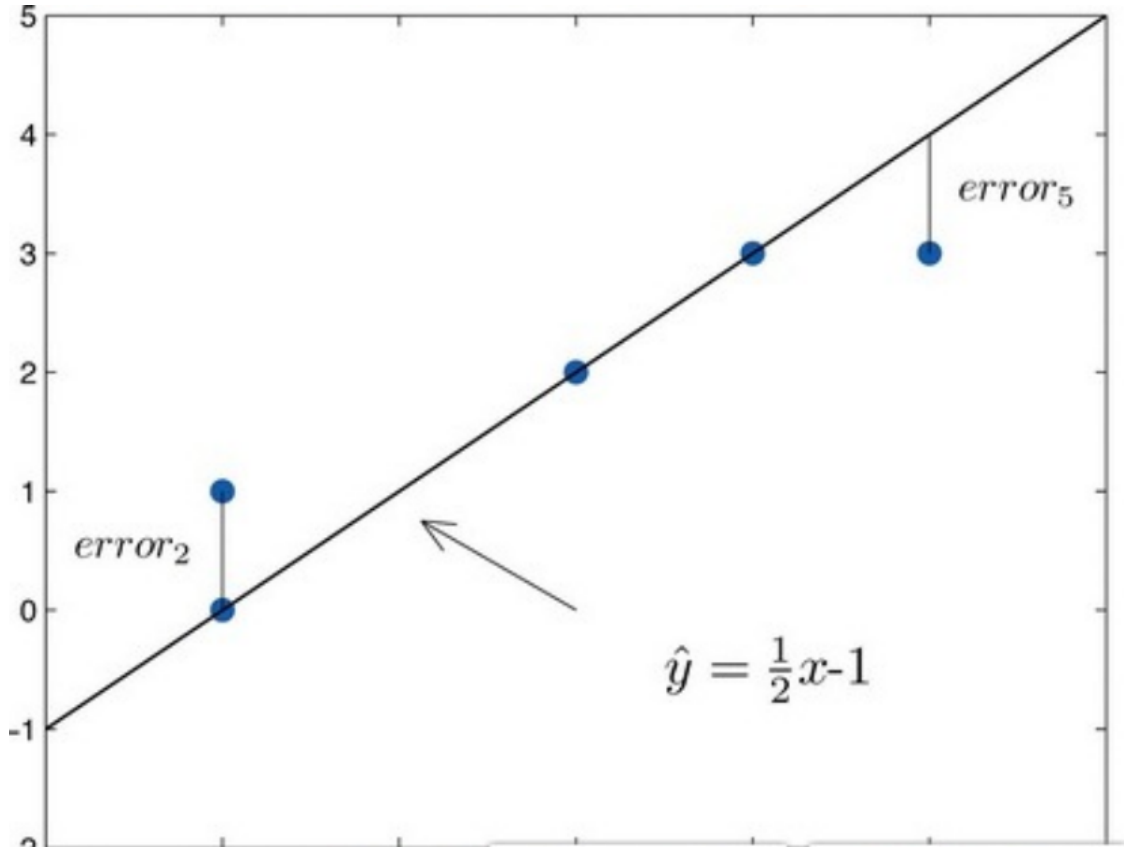
Modelling Linear Relationships with Randomness Present

- Regression model – "all models are wrong"

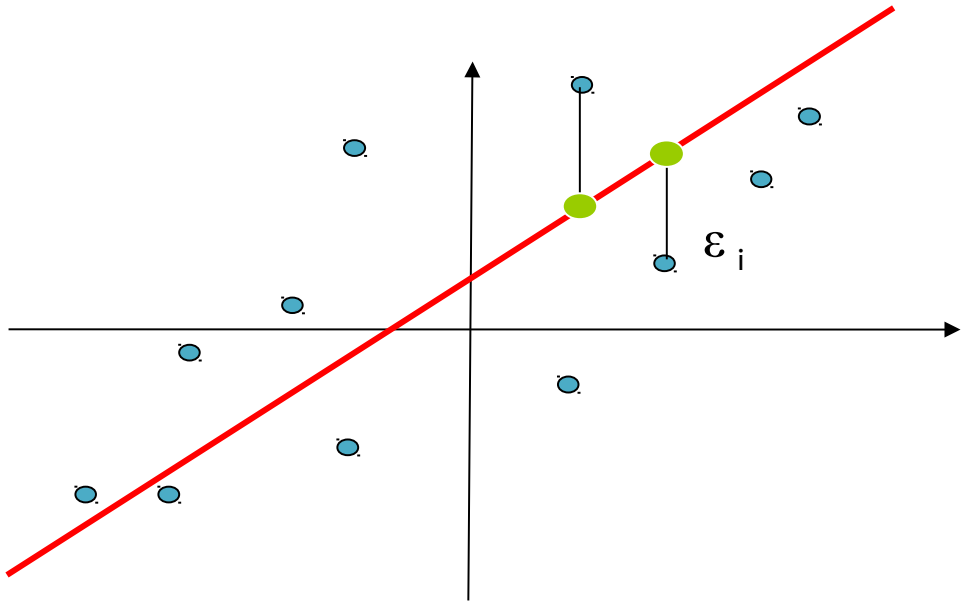


Least square regression

- What line best to fit? Sum of error might cancel out. So better fit sum of squared errors.



Fitting data with a line



= $\hat{y}_i$, predicted

= y_i, observed

ε_i = residual

Residual error (ε_i): Difference between obtained and predicted values of y (i.e. $y_i - \hat{y}_i$)

Best fit line (values of *b* and *a*) is the one that minimises the sum of squared errors (SS_{error}) $\Sigma(y_i - \hat{y}_i)^2$

Test for regression weights

- Does a variable contribute to the prediction? Given the estimated parameter and variance, we can use the T statistic. T = (beta – B0)/s df=n-p and estimate its p-value