Linear regression models an outcome variable (Y) in terms of one or more predictor variables (X). The model asserts that Y is a linear combination of columns of X plus some noise. The noise is assumed to be Gaussian with some variance σ^2. The residual variance is assume to be the same for all data points).

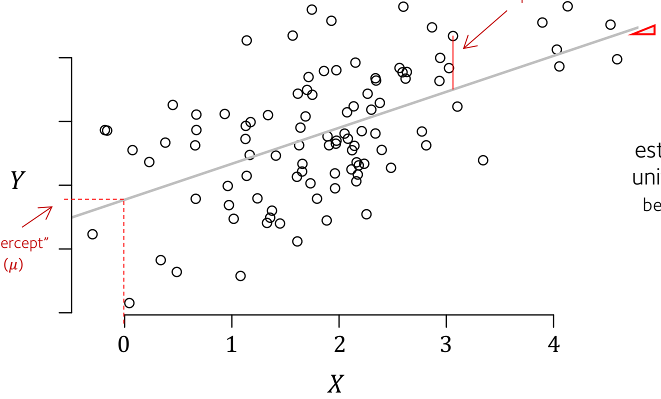
$$Y = \mu + X_1\beta_1 + X_2\beta_2 + \cdots X_d\beta_d + \epsilon \qquad \epsilon \sim N(0, \sigma^2)$$

Or using matrix notation:

$$Y = \mu + X\beta + \epsilon \qquad \epsilon \sim N(0, \sigma^2)$$

Matrix multiplication of the *d*-dimensional *row* vector of predictors X and the d-dimensional *column vector* of parameters β

"residual" error for this data point



Regression "slope" ($\hat{\beta}$)

"intercept" (μ)

Interpretation: $\hat{\beta}$ is the estimated increase in Y for a unit increase in X. (The hat has been added to indicate this is an estimate.)

The likelihood function. The regression likelihood composes the above into a single formula – the likelihood of Y given X and the parameters. (It is simplest to write this if we instead imagine μ to be the first entry of β . This works out if we add a single 1 as the first entry of X:

For a single sample: $$P(Y|X,\beta) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\frac{(Y-X\beta)^2}{\sigma^2}}$$

Squared error (distance) from regression line

The outcome values are assumed independent of each other (probabilities multiply). So for multiple samples the likelihood is:

For multiple samples: $(n = 1, \ldots, N)$ $$P(Y|X,\beta) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\frac{\sum_n (Y_n - X_n\beta)^2}{\sigma^2}}$$

"sum of squared errors"

The exponent is negative. Maximising the likelihood is therefore the same as minimizing the sum of squared errors – it finds the 'best-fitting line'.