Big idea in data science

Using one algorithm (training algorithm) to train
another (model algorithm) on data.
Legendre 1805 - orbits of comets, SSQ.
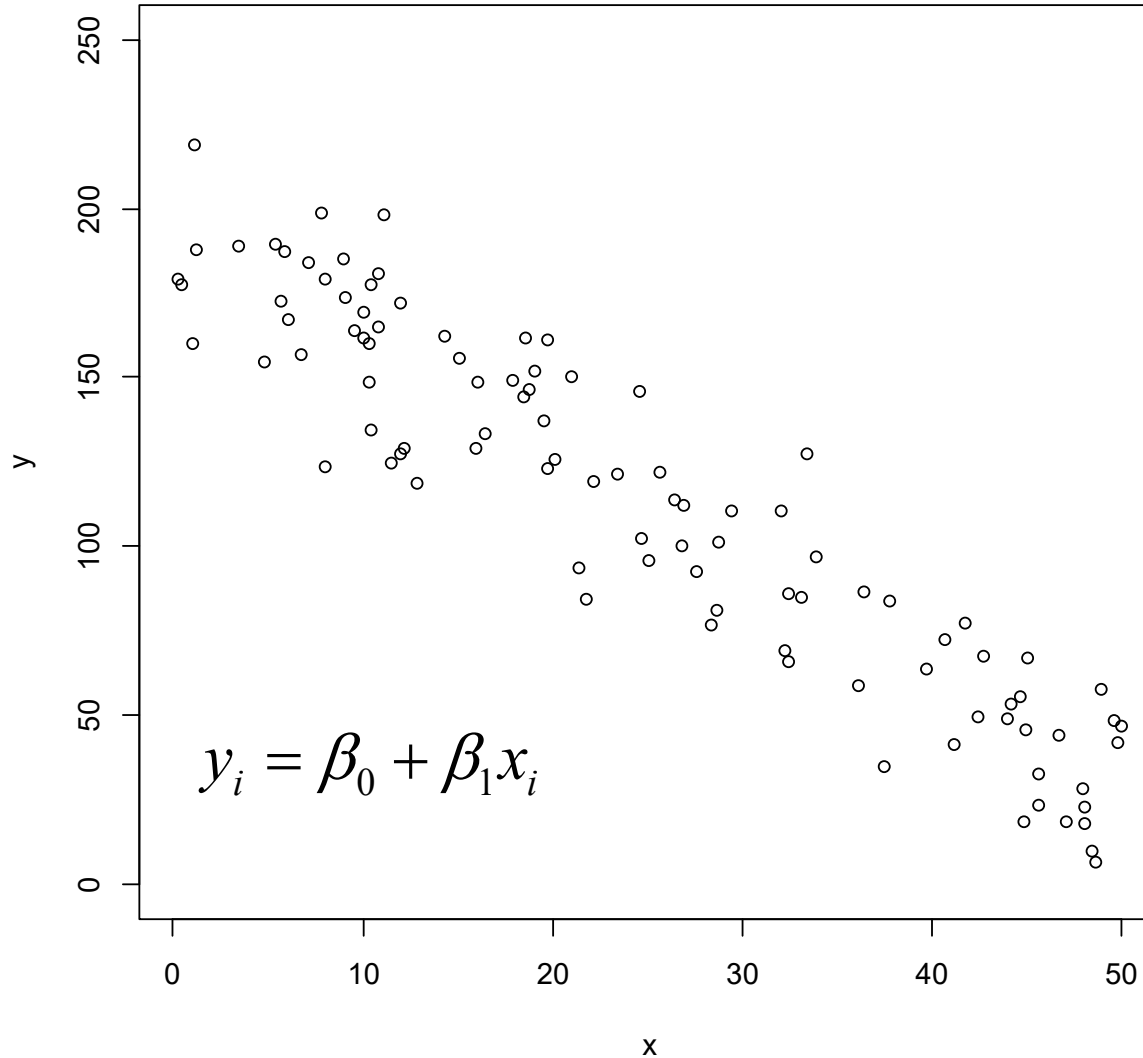
Model algorithm

Training algorithm

An algorithm to train a model algorithm on data
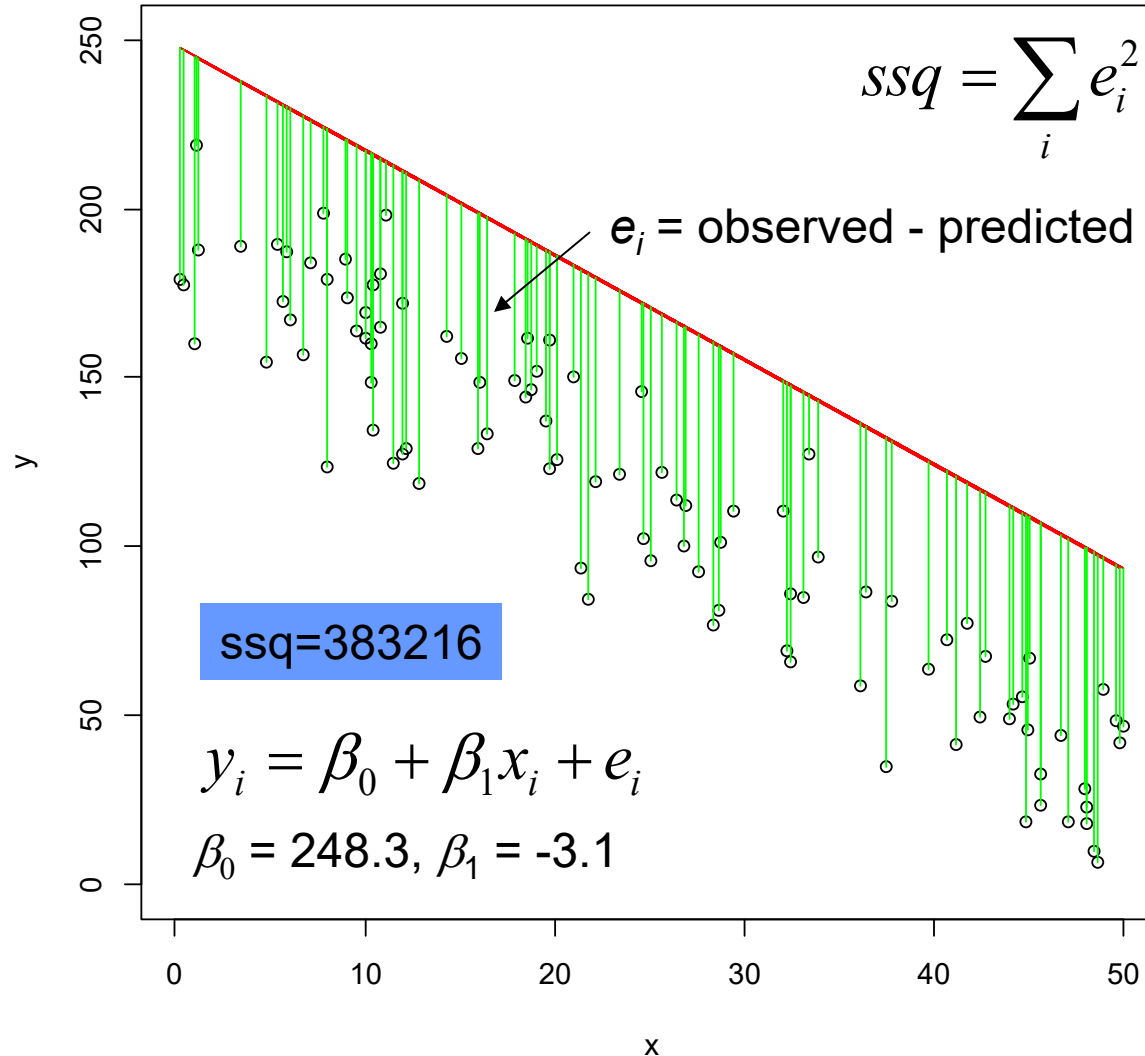syn. model fitting, calibration

Least squares algorithms

$$y_i = \beta_0 + \beta_1 x_i$$

General algorithmic idea:

Vary model parameters until we find the parameter values that minimize the distance of the model from the data

Least squares algorithms

$$ssq = \sum_i e_i^2$$

e_i = observed - predicted

ssq=383216

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

β_0 = 248.3, β_1 = -3.1

General algorithmic idea:

Vary model parameters until we find the parameter values that minimize the distance of the model from the data

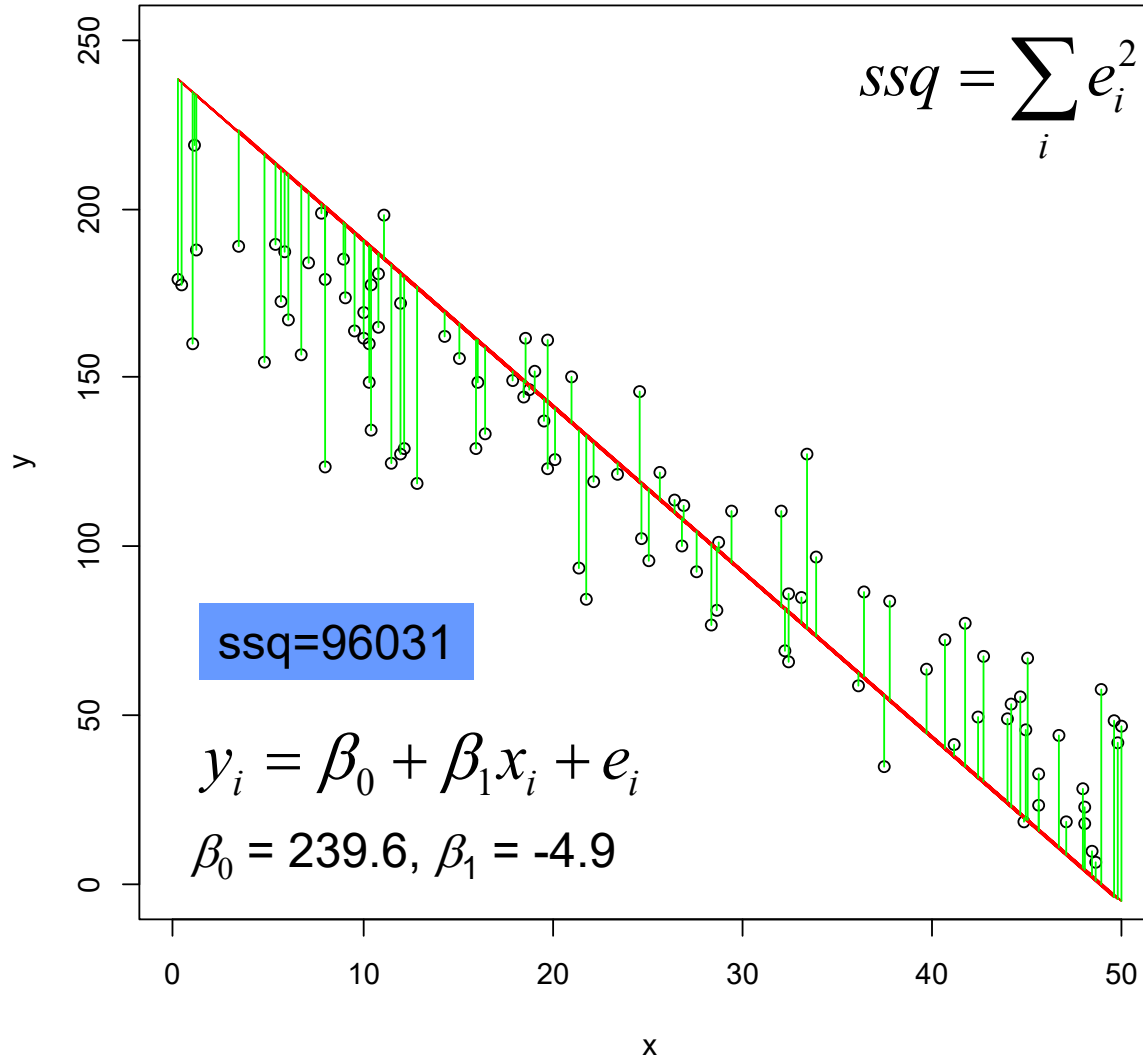# Least squares algorithms

$$ssq = \sum_i e_i^2$$

ssq=96031

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

β_0 = 239.6, β_1 = -4.9

General algorithmic idea:

Vary model parameters until we find the parameter values that minimize the distance of the model from the data

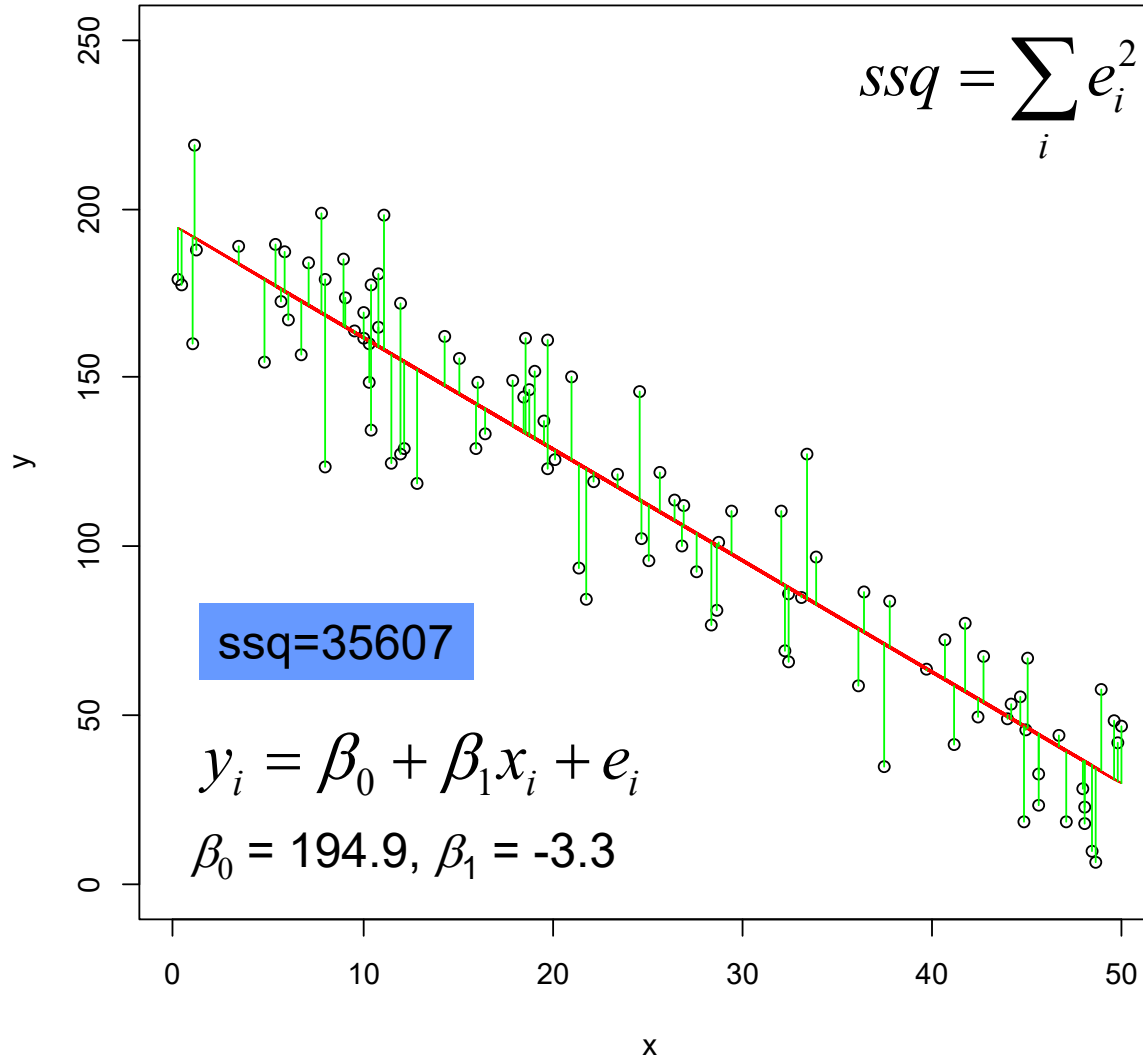# Least squares algorithms

$$ssq = \sum_i e_i^2$$

ssq=35607

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

$\beta_0 = 194.9$, $\beta_1 = -3.3$

General algorithmic idea:

Vary model parameters until we find the parameter values that minimize the distance of the model from the data

Optimization algorithms

- Systematically try all combinations for β_0 and β_1 - Grid search algorithms

- Narrowing in - keep changing parameters in the direction that leads to lower SSQ - Descent algorithms

- Try random values for β_0 and β_1 - Monte Carlo algorithms

- Solve for parameters using math - Analytical or numerical algorithms

Developing an algorithm: an example

Key points:

Pseudocode

3 Phases

Top down refinement

Grid search algorithm
Pseudocode

For each value of β_0
      For each value of β_1
            Calculate sum of squares

Grid search algorithm Pseudocode

Read in data
Set up values of β_0 and β_1 to try
Set up storage for ssq, β_0, β_1
For each value of β_0
    For each value of β_1
        Calculate sum of squares
    Store ssq, β_0, β_1
Plot sum of squares profiles (ssq vs β_0, ssq vs β_1)
Report best ssq, β_0, β_1
Plot fitted model with the data

Grid search algorithm Pseudocode

Read in data
Set up values of β_0 and β_1 to try
Set up storage for ssq, β_0, β_1
For each value of β_0
    For each value of β_1
        Calculate sum of squares
        Store ssq, β_0, β_1
Plot sum of squares profiles (ssq vs β_0, ssq vs β_1)
Report best ssq, β_0, β_1
Plot fitted model with the data

Initialization phase

Calculation phase

Termination phase

Grid search algorithm Pseudocode

Read in data
Set up values of β_0 and β_1 to try
Set up storage for ssq, β_0, β_1
For each value of β_0
     For each value of β_1
          Calculate sum of squares
          Store ssq, β_0, β_1
Plot sum of squares profiles (ssq vs β_0, ssq vs β_1)
Report best ssq, β_0, β_1
Plot fitted model with the data

Top down refinement

Grid search algorithm Pseudocode

Read in data
Set up values of β_0 and β_1 to try
Set up storage for ssq, β_0, β_1
For each value of β_0

    For each value of β_1

        Calculate model predictions
        Calculate deviations
        Sum squared deviations
        Store ssq, β_0, β_1

Plot sum of squares profiles (ssq vs β_0, ssq vs β_1)
Report best ssq, β_0, β_1
Plot fitted model with the data

Top down refinement