Today

- Data science: classes of algorithms
- Training algorithm
- Scientific coding in Python with numpy, matplotlib, pandas

Assignments

- Yay to using functions already!
- Pay attention to indenting
- If you use LLMs, make sure you know what they're doing
  - ask it to explain the code line by line
  - if you get syntax or libraries we haven't covered in class, it's probably not what you want. Ask for a minimal example.

Algorithms in data science

- Model algorithm
- Training algorithm
- Inference (reliability) algorithm

- Workflow algorithms

Algorithms in data science

- Model algorithm
  - Intrinsic stochasticity (e.g. movement)
  - Deterministic equations + noise
  - Has parameters
- Training algorithm
- Inference algorithm

Algorithms in data science

- Model algorithm
- Training algorithm
  - An algorithm to train a model algorithm on data
  - syn. model fitting, calibration
  - e.g. least squares, maximum likelihood
  - try typing lm into the R console
- Inference algorithm

Algorithms in data science

- Model algorithm
- Training algorithm
- Inference algorithm
  - looking back: considering all the ways data could have happened
  - looking forward: predicting new data and testing against them

Workflow algorithms in DS

- Pipelines, standard conventions ...

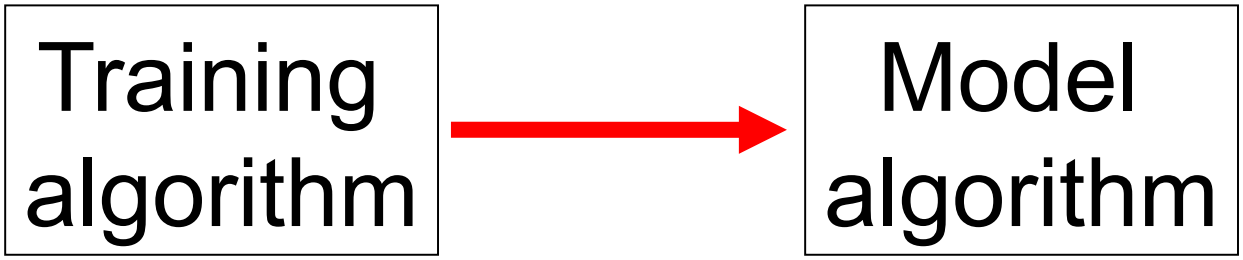
Overall Data Science Algorithm

```
Generate alternative models
Simulate models, exploration
Design data acquisition
Confront models with data
```

Training algorithms

Training algorithm

aka model fitting, model calibration

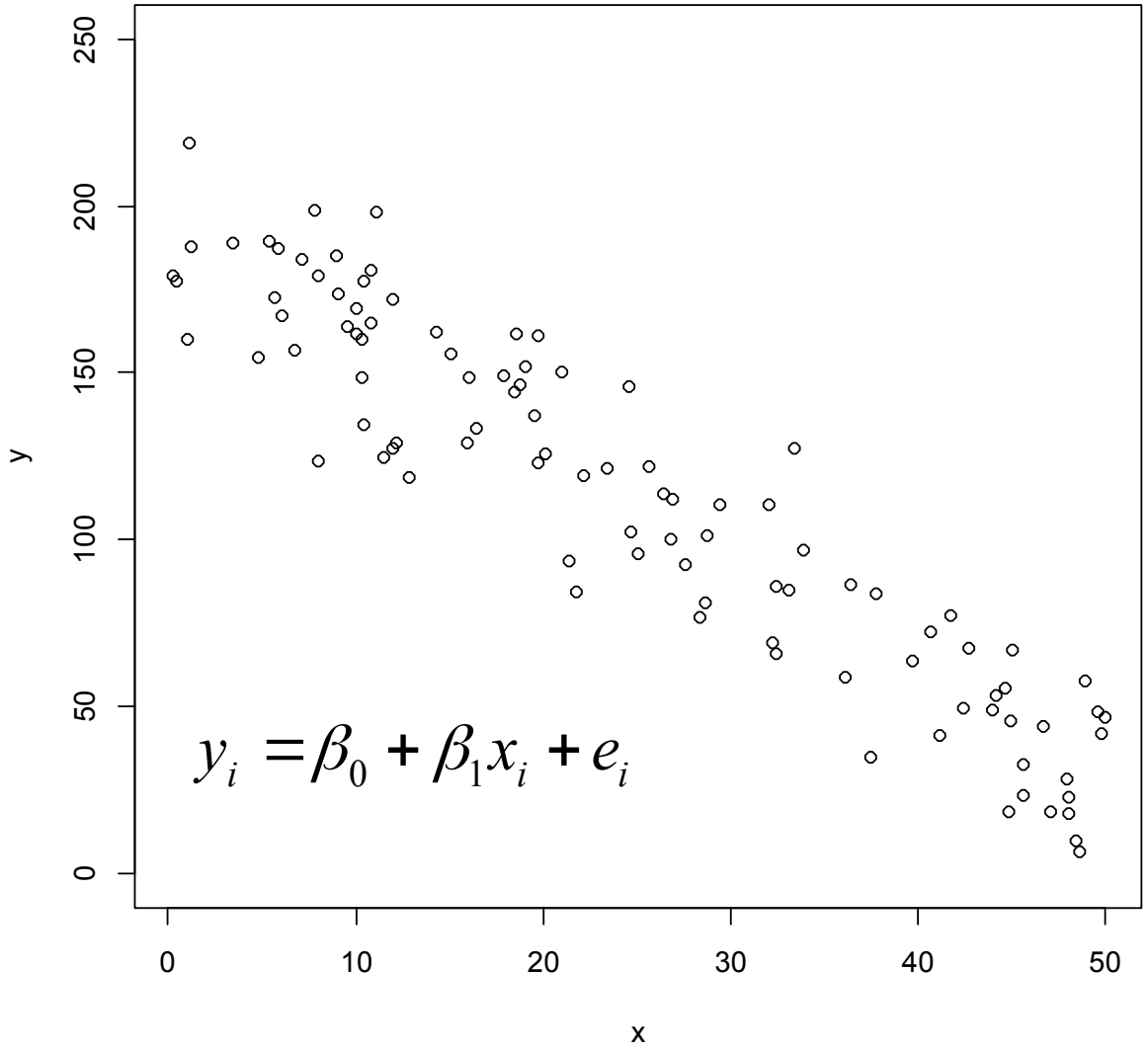


Big idea in data science

Legendre 1805: comet orbits, SSQ

Use one algorithm (training algorithm) to train another (model algorithm) on data

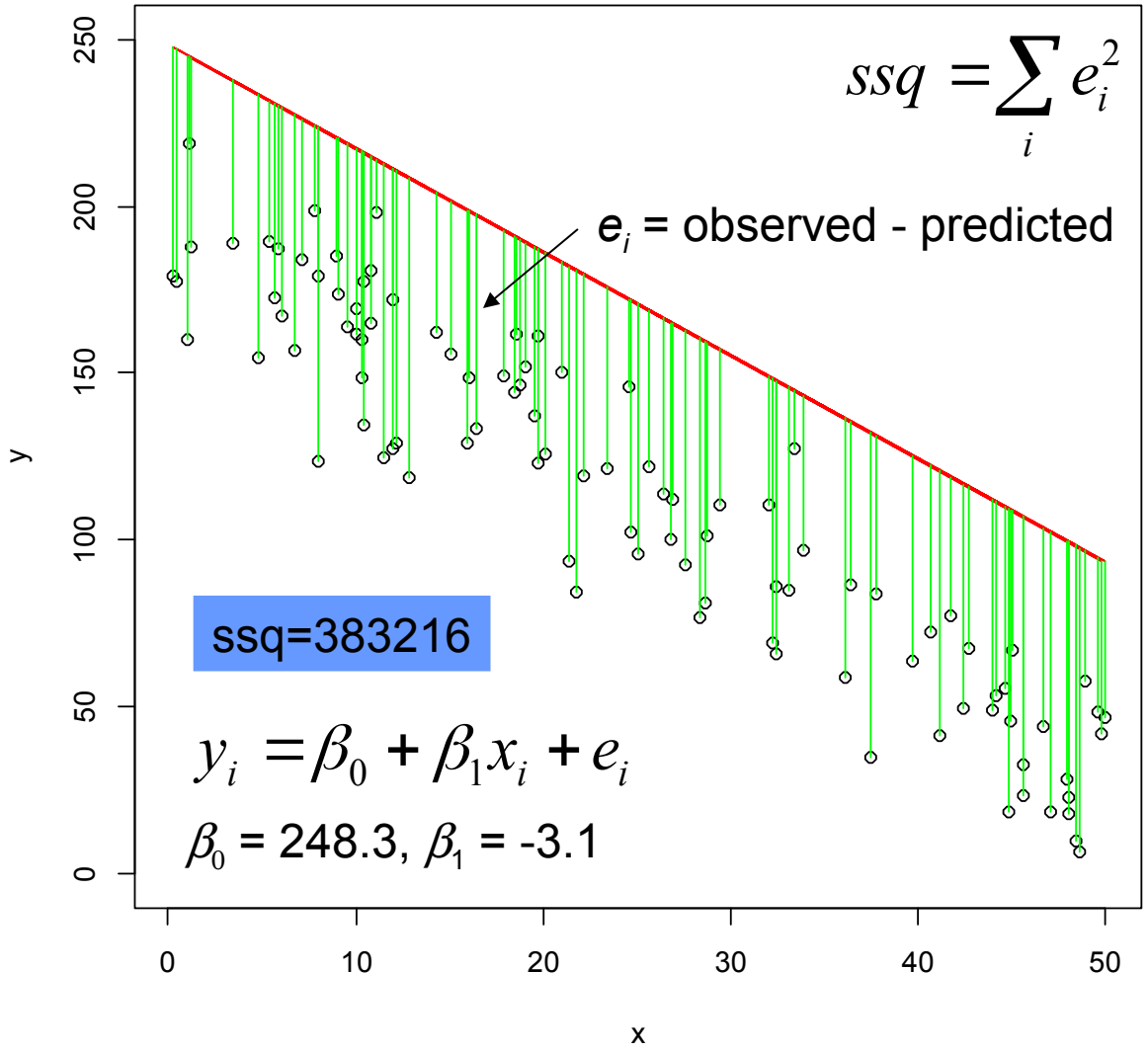
Least squares algorithm



General algorithmic idea:

Vary model parameters until we find the parameter values that minimize the distance of the model's deterministic skeleton from the data

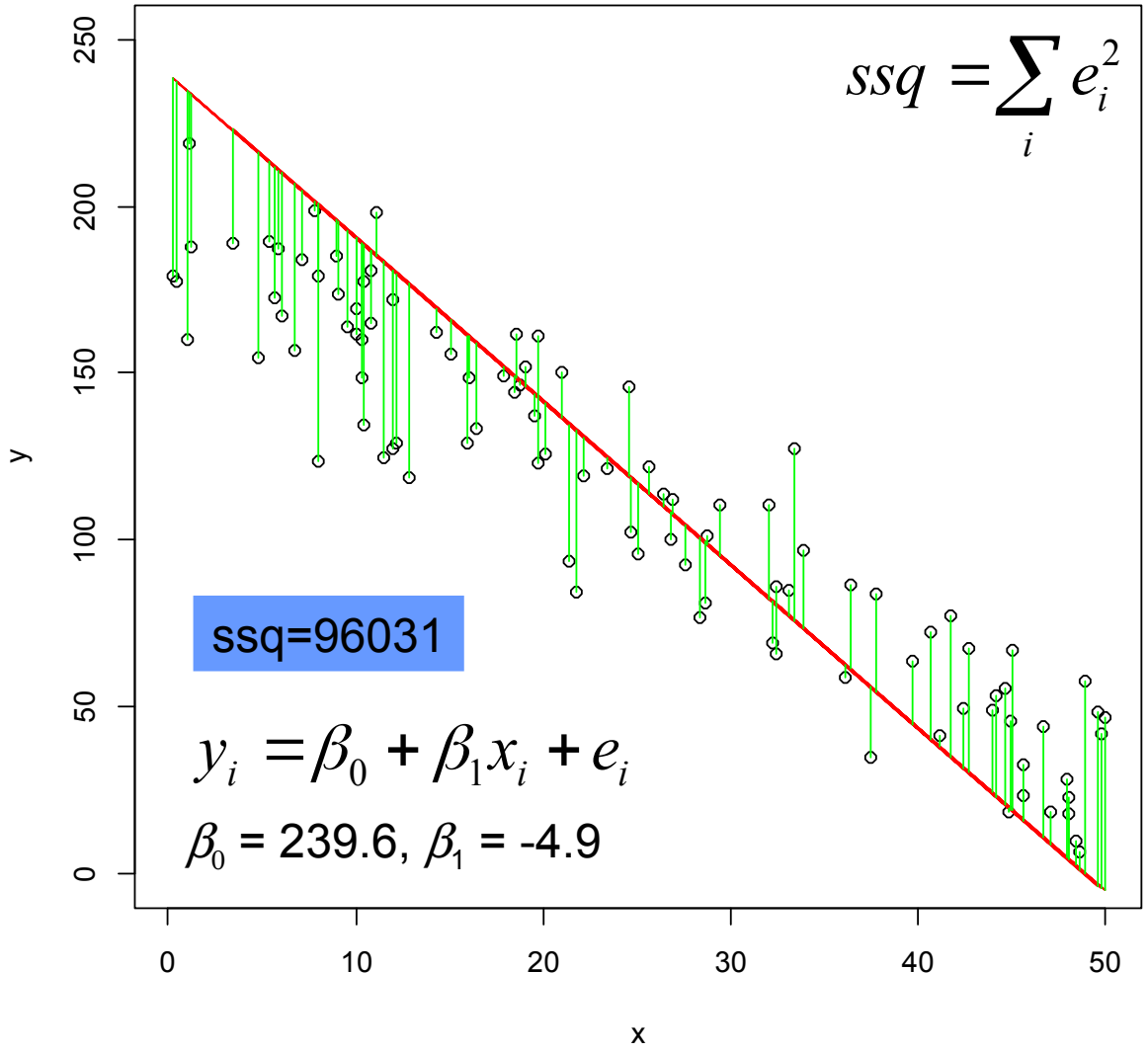
Least squares algorithm



General algorithmic idea:

Vary model parameters until we find the parameter values that minimize the distance of the model's deterministic skeleton from the data

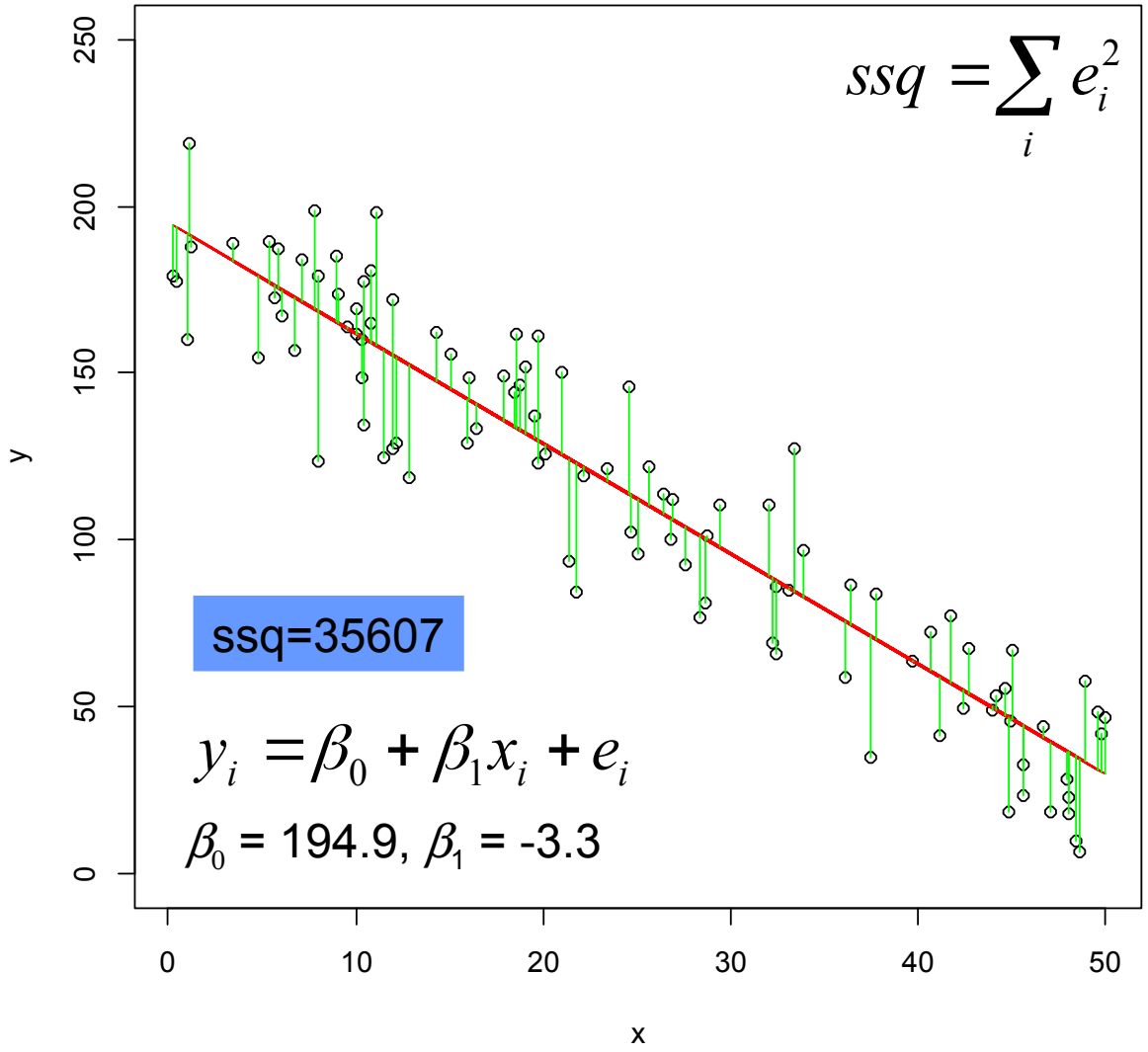
Least squares algorithm



General algorithmic idea:

Vary model parameters until we find the parameter values that minimize the distance of the model's deterministic skeleton from the data

Least squares algorithm



General algorithmic idea:

Vary model parameters until we find the parameter values that minimize the distance of the model's deterministic skeleton 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

Develop a training 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

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

Translate this to Python and use it to train the model with your data