Regression and curve fitting

Benjamin Naecker

NENS 230

20 Nov 2014

Outline

- What is it good for?
- Concepts
  - Least-squares
  - Linear vs nonlinear regression
  - Parametric vs non-parametric methods
  - Overfitting and regularization
- Examples!

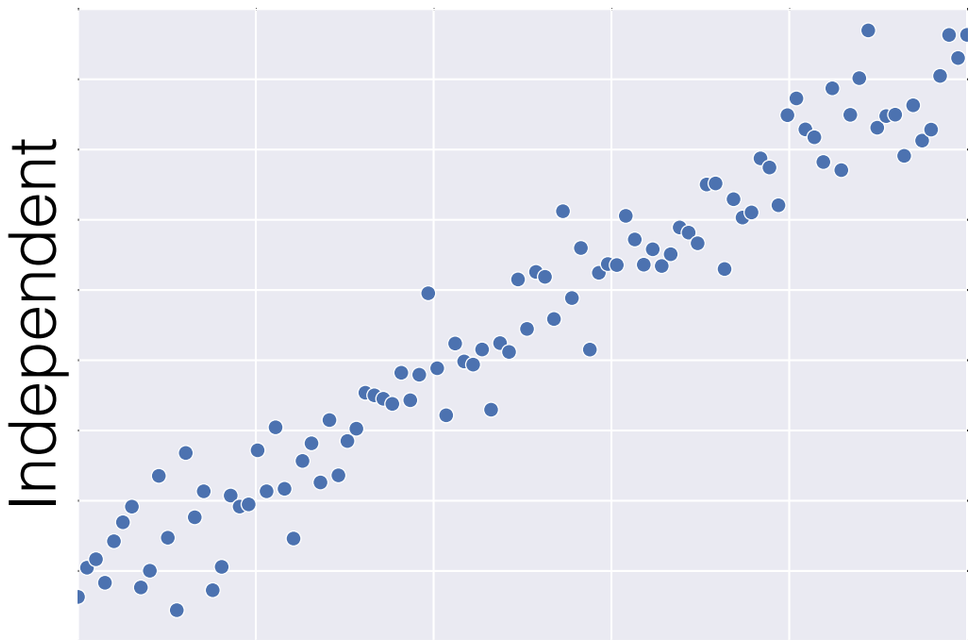
Why fit curves?

Expectation

Dependent

$$y = f(x)$$

Reality



Dependent

$$y_{\mathrm{meas}} = f(x) + \mathrm{noise}$$

Given noisy measurements,
how can we uncover “true” relationships?

Regression analysis FTW!

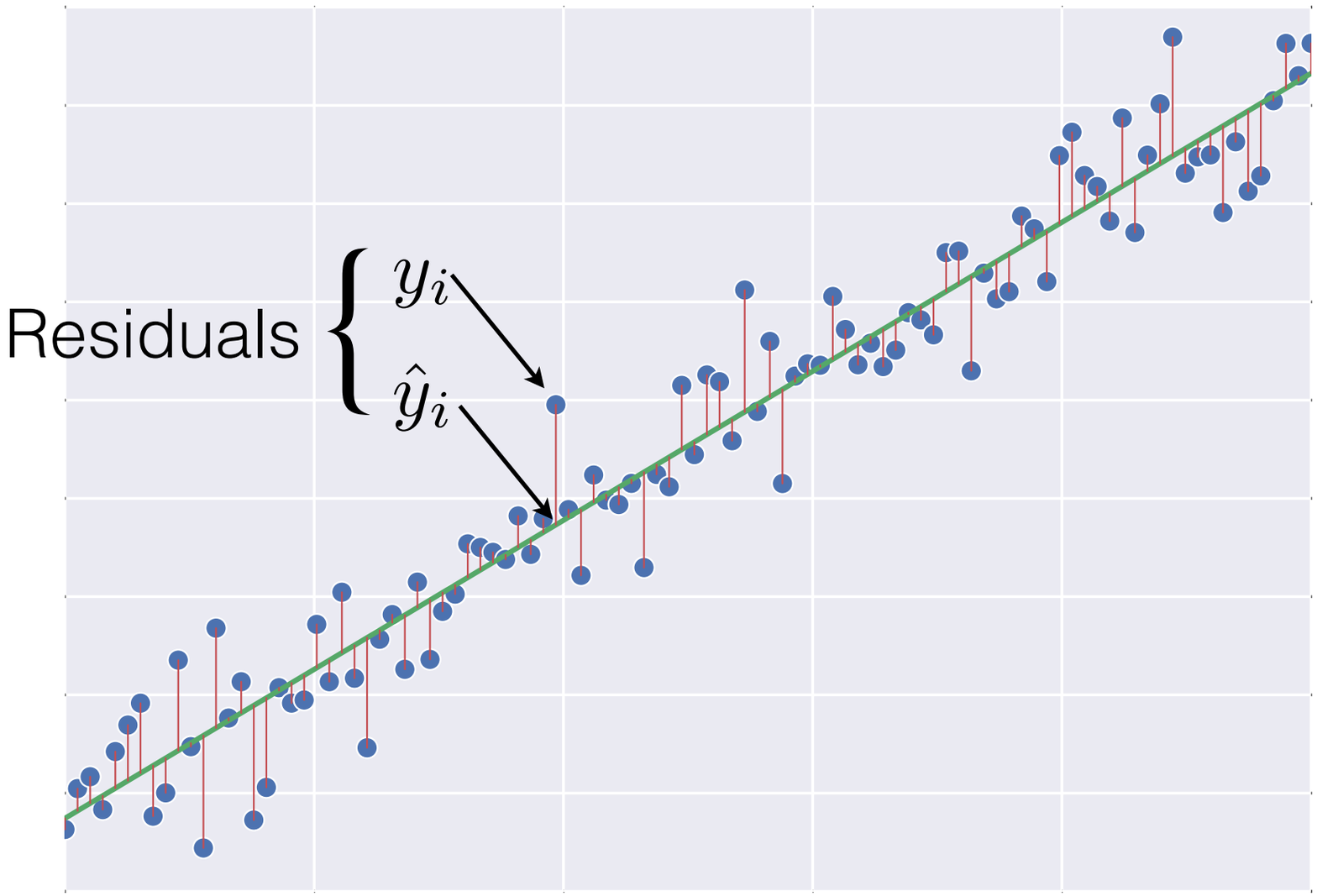
Regression analysis

Goal of regression and curve fitting is to find a *line of best-fit*

Recover the “true” relationships in the data

$$y = f(x)$$

Least-squares



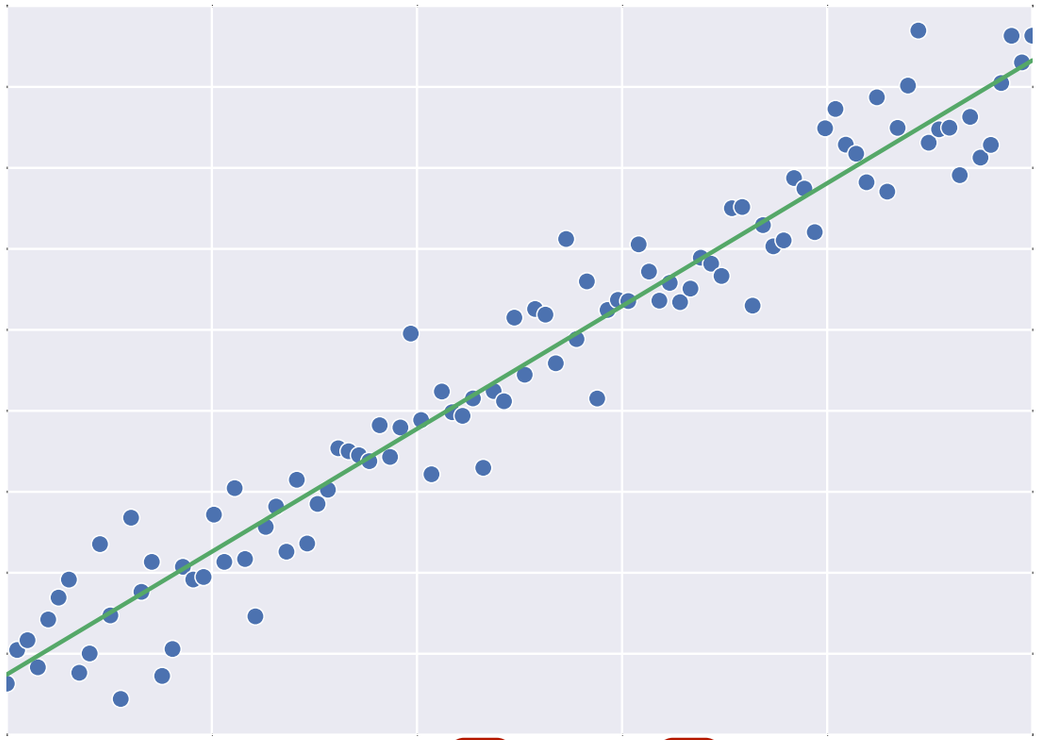
$$\text{minimize} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

Linear vs nonlinear regression

$$y = ax + b$$

$$y = ax + bx^2 + cx^3$$

Parametric curve fitting

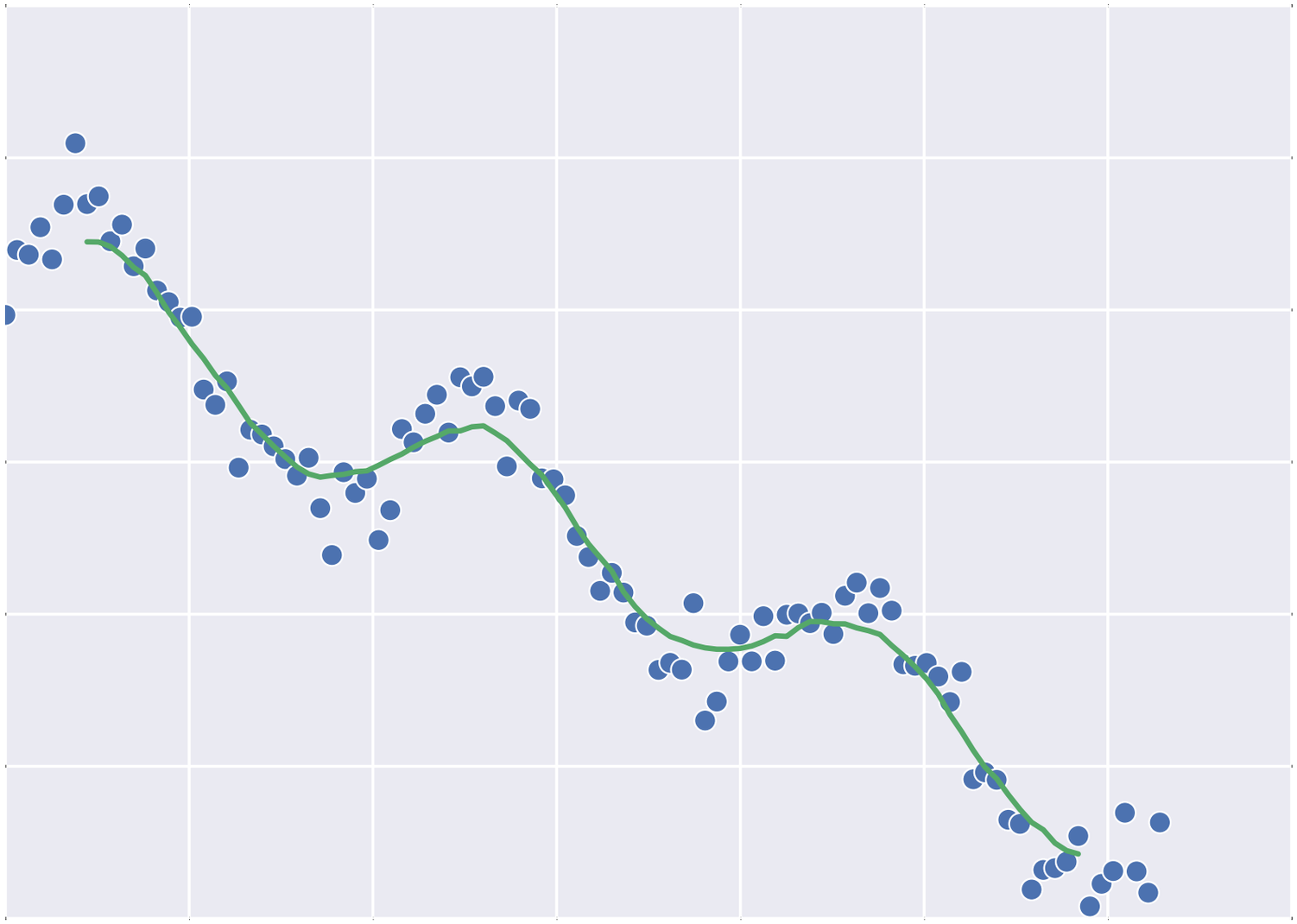


$$y = ax + b$$

Parameters

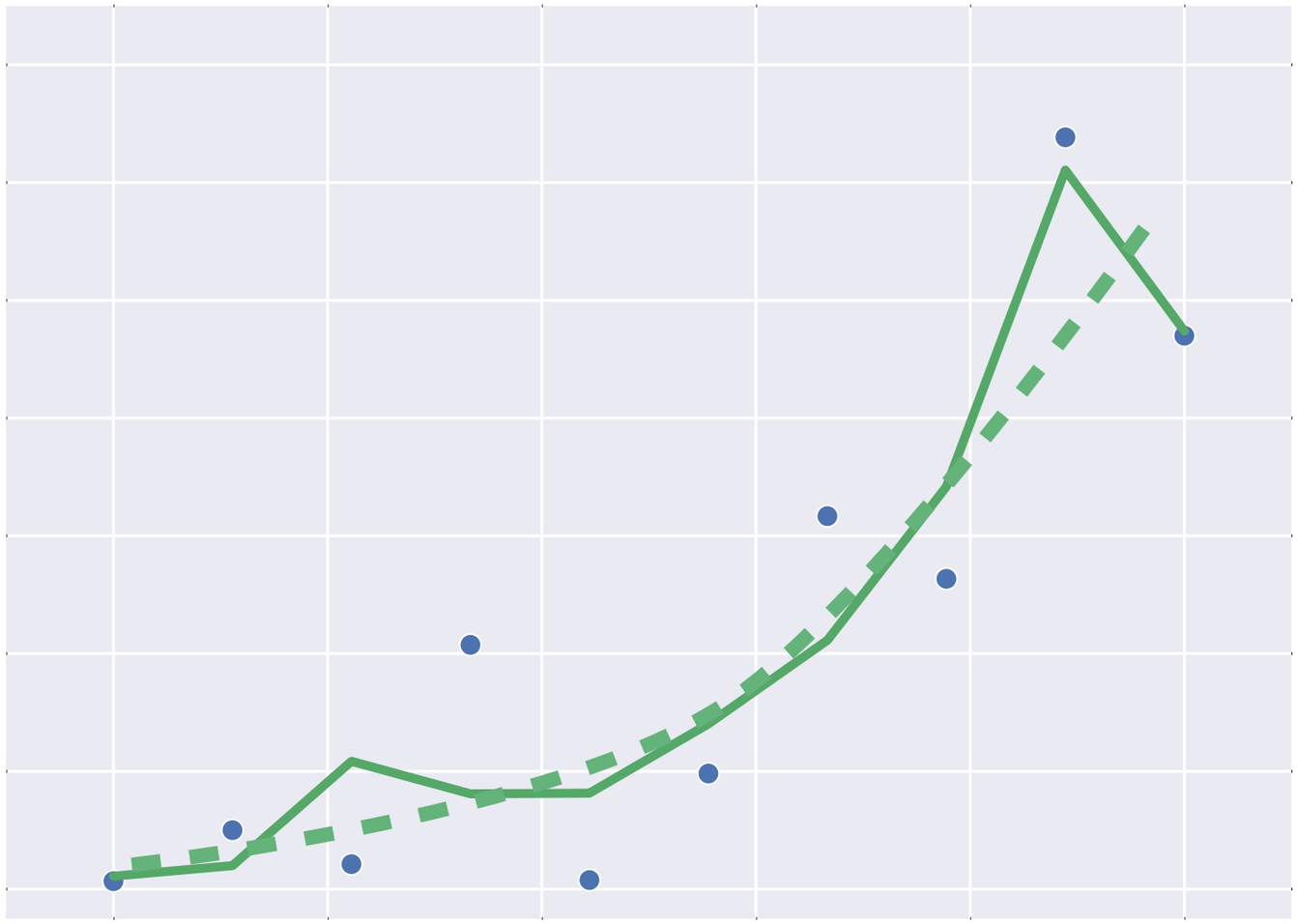
We want to know something about the actual relationship in our data

Non-parametric curve fitting



No model for relationships between variables.
Want to *de-noise* or *smoooooth* the data.

Overfitting and regularization



Not all variability in data is meaningful

Regularization *penalizes* fitting the noise

A note on terminology

- Independent variables
  - *data, predictors, design variables, design matrix*
- Dependent variables
  - *observations, response variables, outcome*
- Statistics
  - *summary statistics, goodness-of-fit, e.g, r-square, F-statistics etc*

Examples!

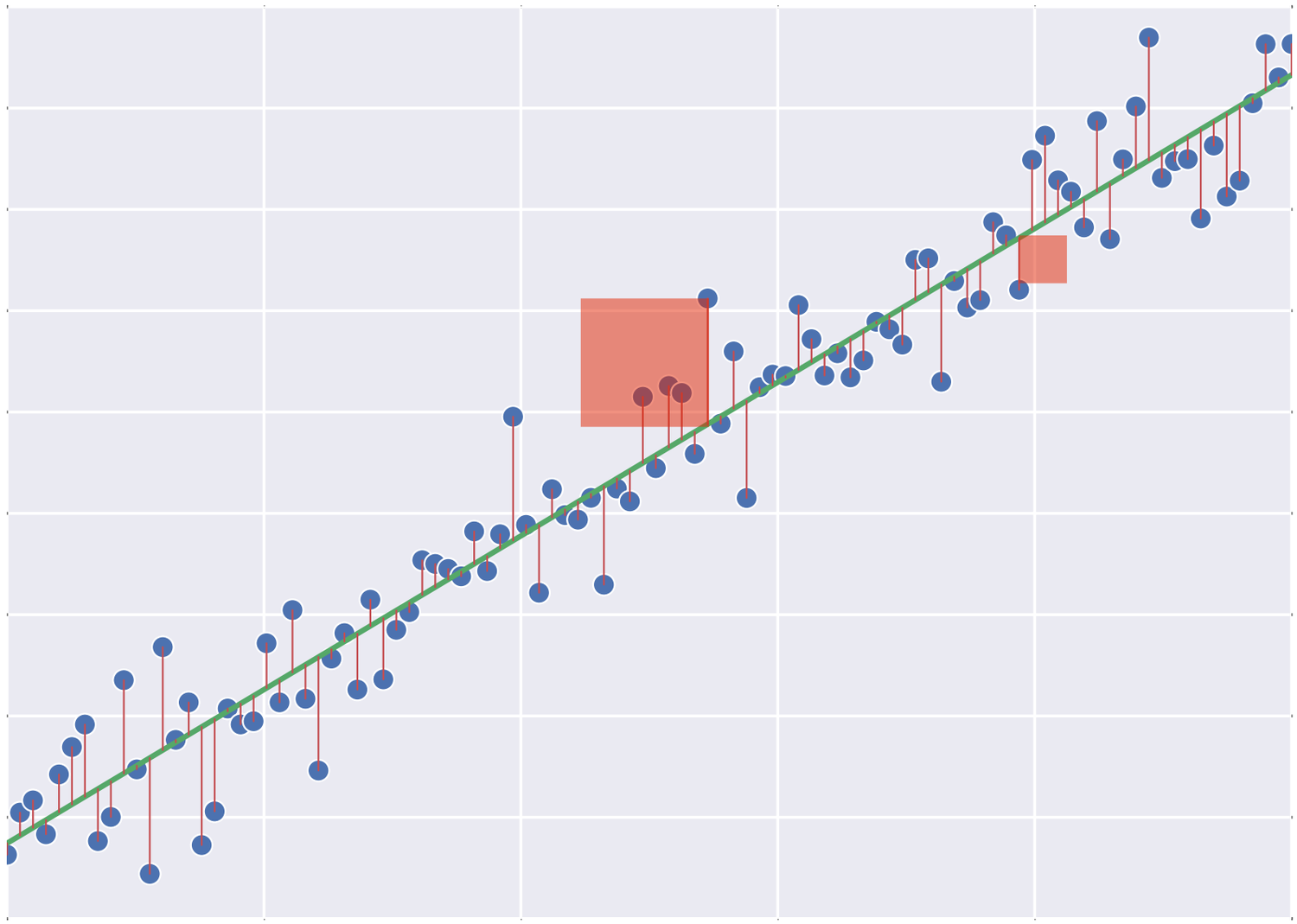
All examples on course website:

nens230.stanford.edu/lecture9.html/regression-examples.zip

Simple linear regression

- `\ (mldivide)`
- `regress`
- `robustfit`

Why robust least-squares



Ordinary least-squares heavily weights outliers.
A residual twice as big costs *four times* as much.

Multiple and multivariate linear regression

- `regress`
- `mvregress`

Nonlinear regression

- `polyfit`
- `nlinfit`
- Overfitting!
- `cftool`

Smoothing and interpolation

- `smooth`
- `interp1`

Final projects

- Goal is to use concepts learned here in real applications
- Scope of about two assignments, but very flexible
  - Data analysis: your own data, someone else's, or freely available set
  - Simulation and analysis thereof

Final projections

- Will include code, figures and summary
- Code
  - Neat, documented, and reasonably efficient
- Figures
  - Include captions, don’t use Matlab’s defaults
- Summary
  - 1 page of text
  - Include clear distinction of your own code versus borrowed or inherited code
  - Basic introduction to your dataset, where it came from and how it was collected

Final projections

Do something useful!

Have fun!