ECE 295: Lecture 04 Regression

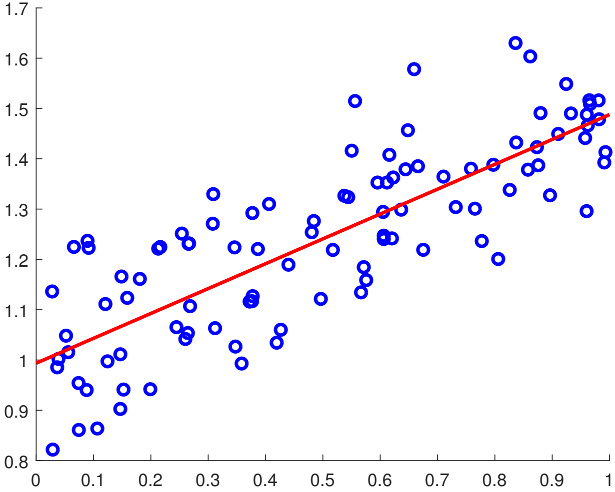
Spring 2018

Prof Stanley Chan

School of Electrical and Computer Engineering
Purdue University

Data Fitting

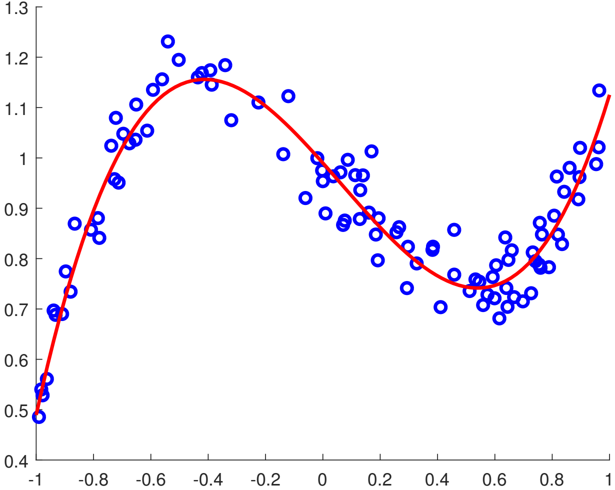
- You give me data, I find the trend.



Data Fitting

Once I find the trend, I can

- Predict values where I previously did not measure
- Extrapolate outside the range

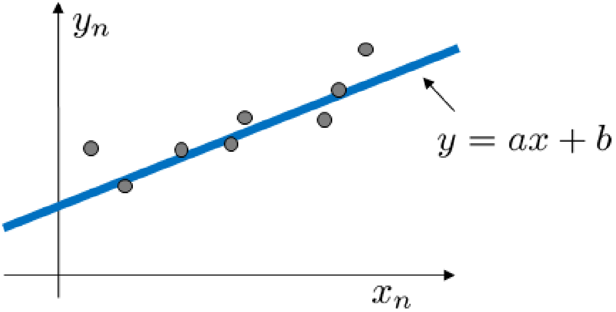


Problem Formulation

First, we need a **model**!
Let's start with this:

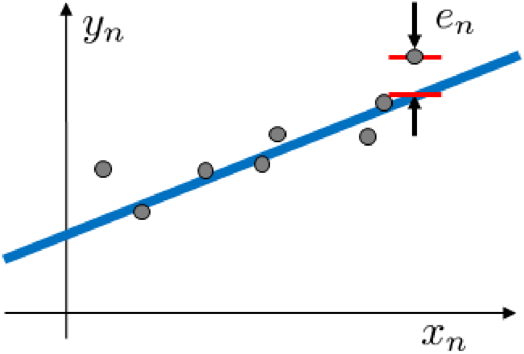
$$y_n = ax_n + b + e_n, \qquad n = 1, \ldots, N$$

This is a linear equation.



What is the error?

- y_n = true measured value
- $ax_n + b$ = estimated value
- e_n measures the difference $y_n - (ax_n + b)$



What is "best"?

We need solve this **optimization** problem:

$$\left(\widehat{a}, \widehat{b}\right) = \underset{(a,b)}{\arg\min} \sum_{n=1}^{N} (y_n - (ax_n + b))^2.$$

- argmin = find the values of the variables that can minimize the function.
- $\sum_{n=1}^{N} (y_n - (ax_n + b))^2$: sum of all the errors
- You don't have to choose $(\cdot)^2$. You can use $|\cdot|$, or $\max(\cdot)$ or whatever.
- $(\cdot)^2$ is just easier.
- How to solve this optimization?
- Take derivative, set it to zero.

Main Result

Theorem

The solution of the problem

$$\left(\widehat{a}, \widehat{b}\right) = \underset{(a,b)}{\arg\min} \; \sum_{n=1}^{N} (y_n - (ax_n + b))^2$$

is the solution to the following system of linear equations

$$\begin{bmatrix} \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} x_n & n \end{bmatrix} \begin{bmatrix} \widehat{a} \\ \widehat{b} \end{bmatrix} = \begin{bmatrix} \sum_{n=1}^{N} x_n y_n \\ \sum_{n=1}^{N} y_n \end{bmatrix} \tag{1}$$

Solution

First, let us define

$$\varphi(a, b) = \sum_{n=1}^{N} (y_n - (ax_n + b))^2.$$

Taking derivatives on both sides with respect to a and b yields

$$\frac{\partial}{\partial a}\varphi(a, b) = 2\left(\sum_{n=1}^{N} x_n y_n - a\sum_{n=1}^{N} x_n^2 - b\sum_{n=1}^{N} x_n\right) = 0$$

$$\frac{\partial}{\partial b}\varphi(a, b) = 2\left(\sum_{n=1}^{N} y_n - a\sum_{n=1}^{N} x_n - nb\right) = 0$$

Rearranging the terms, this is equivalent to

$$\begin{bmatrix} \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} x_n & n \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{n=1}^{N} x_n y_n \\ \sum_{n=1}^{N} y_n \end{bmatrix}$$

Matrix-Vector Representation

This is a 2×2 system of linear equations

$$\begin{bmatrix} \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} x_n & n \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{n=1}^{N} x_n y_n \\ \sum_{n=1}^{N} y_n \end{bmatrix}$$

This is equivalent to

$$\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta} = \boldsymbol{X}^T \boldsymbol{y}, \tag{2}$$

where

$$\boldsymbol{X} = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} a \\ b \end{bmatrix}, \tag{3}$$

Solution in Matrix-Vector Representation

- The equation
$$\boldsymbol{X}^T\boldsymbol{X}\beta = \boldsymbol{X}^T\boldsymbol{y} \tag{4}$$
is called the **normal equation** of a linear system $\boldsymbol{X}\boldsymbol{x} = \boldsymbol{\beta}$.
- To determine the vector $\boldsymbol{\beta}$, we take inverse (assuming $\boldsymbol{X}^T\boldsymbol{X}$ is invertible):
$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{5}$$
- The matrix $\boldsymbol{X}^T\boldsymbol{X}$ is invertible when there is no dependent columns of $\boldsymbol{X}^T\boldsymbol{X}$, which in turn holds when there is no dependent columns of $\boldsymbol{X}$.
- If the matrix $\boldsymbol{X}^T\boldsymbol{X}$ is close to non-invertible (i.e., having a very large condition number), then we can perturb the solution as
$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{6}$$
where $\lambda > 0$ is a constant.

General Least Squares Minimization

The normal equation can also be derived from an optimization:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \quad \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2 \tag{7}$$

Here, $\|\boldsymbol{u}\|^2$ denotes the ℓ_2-norm square of a vector $\boldsymbol{u}$:

$$\|\boldsymbol{u}\|^2 = \sum_{i=1}^{n} u_i^2.$$

Derivation of the optimal solution: (Need some matrix-calculus)

$$\begin{aligned} \frac{d}{d\boldsymbol{\beta}}\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2 = 0 \quad &\Rightarrow \quad \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}) = 0 \\ &\Rightarrow \quad \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{\beta}, \end{aligned}$$

so we obtain the same normal equation.

Example 1: Quadratic Fitting

Problem: Find the linear least squares solution for

$$y_n = ax_n^2 + bx_n + c$$

Extension: This idea can be extended high order polynomials.

Solution:

$$\boldsymbol{X} = \begin{bmatrix} x_1^2 & x_1 & 1 \\ \vdots & \vdots & \vdots \\ x_N^2 & x_N & 1 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} a \\ b \\ c \end{bmatrix},$$

The solution is

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

Example 2: Auto-Regressive Model

Problem: Find the linear least squares solution for

$$y_n = a y_{n-1} + b y_{n-2}$$

Application: Stock-prediction: We have sample y_{n-1} and y_{n-2}, we want to predict y_n.

Solution:

$$\boldsymbol{X} = \begin{bmatrix} y_2 & y_1 \\ y_3 & y_2 \\ \vdots & \vdots \\ y_{N-1} & y_{N-2} \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_3 \\ y_4 \\ \vdots \\ y_N \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} a \\ b \end{bmatrix},$$

The solution is

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}.$$

Interpreting the Results

city	funding	hs	not-hs	college	college4	crime rate
1	40	74	11	31	20	478
2	32	72	11	43	18	494
3	57	70	18	16	16	643
4	31	71	11	25	19	341
5	67	72	9	29	24	773
⋮	⋮	⋮	⋮	⋮		
50	66	67	26	18	16	940

https://web.stanford.edu/~hastie/StatLearnSparsity/data.html

$$\boldsymbol{X} = \begin{bmatrix} 1 & 40 & 74 & 11 & 31 & 20 \\ 1 & 32 & 72 & 11 & 43 & 18 \\ & & \vdots & & & \\ 1 & 66 & 67 & 26 & 18 & 16 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 478 \\ 494 \\ \vdots \\ 940 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_5 \end{bmatrix},$$

Interpreting the Results

Run regression analysis (with $\lambda = 1000$). Here is the result:

- $\beta_1 = 10.9934$: police funding
- $\beta_2 = 1.1451$: high school
- $\beta_3 = 10.1812$: no high school
- $\beta_4 = 2.7386$: college
- $\beta_5 = -0.7781$: college at least 4 years

That means:

- Crime rate is more influenced by police funding
- and number of residents without high school
- Other factors are not quite relevant

The term β_0 is known as the bias, or the DC term in circuit terminology.

Solution Trajectory

Recall that $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ is equivalent to

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\arg\min} \quad \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2.$$

We can show that $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ is equivalent to

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\arg\min} \quad \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{\beta}\|^2. \tag{8}$$

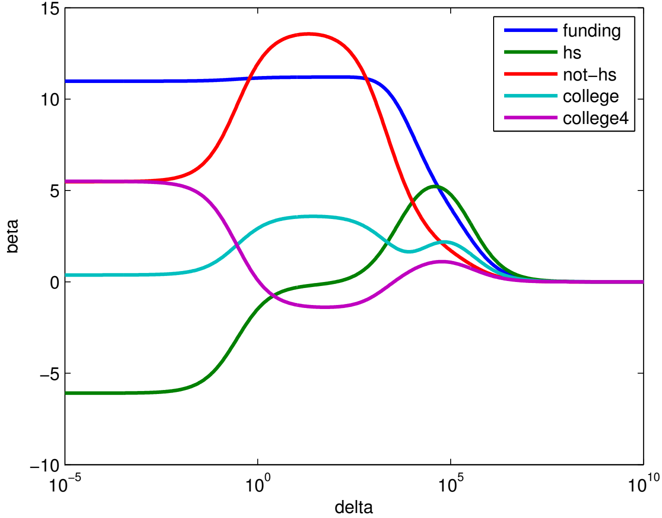
Why?

$$\begin{aligned} \frac{d}{d\boldsymbol{\beta}}(\cdot) = 0 \quad &\Rightarrow \quad \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}) + \lambda\boldsymbol{\beta} = 0 \\ &\Rightarrow \quad (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{y}. \end{aligned}$$

Now, consider $\widehat{\boldsymbol{\beta}}$ as a function of λ:

$$\widehat{\boldsymbol{\beta}}_\lambda = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Solution Trajectory



Beyond Least Squares

It is possible to use other forms of optimization, e.g.,

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\arg\min} \quad \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{\beta}\|_1, \tag{9}$$

where $\|\cdot\|_1$ is called the ℓ_1-norm:

$$\|\boldsymbol{u}\|_1 = \sum_{i=1}^{n} |u_i|.$$

This is called the Least Absolute Shrinkage and Selection Operation (LASSO).

- Solving the LASSO problem is beyond the scope of this course. (See ECE 695 Sparse Modeling and Algorithms)
- It requires convex optimization algorithms.
- LASSO makes $\widehat{\boldsymbol{\beta}}$ *sparse*.
- Essential if $\boldsymbol{X}$ is short and fat. ($\boldsymbol{X}^T\boldsymbol{X}$ is not invertible.)