# "Linear" Regression

Feb 26, 2018

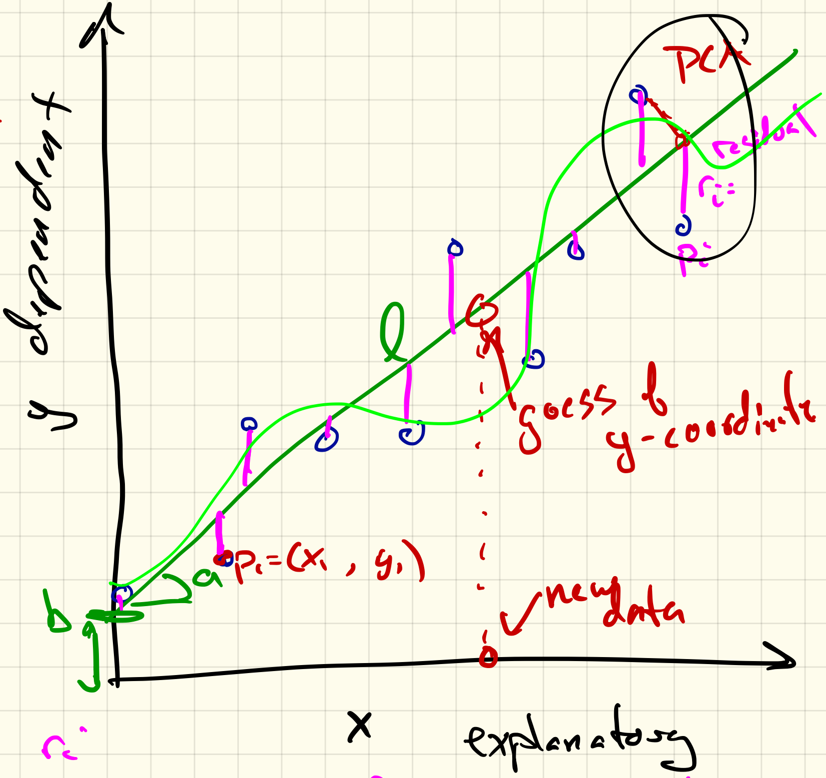
- Clustering Hw Due Wednesday
- Data Science Club: Stack Overflow (PCA)
  Tomorrow, Tue 5-6 WEB 1230
- Start of Regression / Dim-Reduction

See more in Introduction to Data Analysis (M4DA) book!

Input: $P = (p_1, p_2, \dots p_n)$

$p_i \in \mathbb{R}^{d+1}$ today $p_i \in \mathbb{R}^2$

$$p_i = (\underbrace{x_i}_{\mathbb{R}^d}, \underbrace{y_i}_{\mathbb{R}})$$




Goal: fit line

$$\ell : \ell(x) = ax + b$$

$$\operatorname*{argmin}_{\ell} \sum_{p_i \in P} (\underbrace{y_i - \ell(x_i)}_{\text{residual } r_i})^2$$

$$= \operatorname*{argmin}_{a,b} \sum_{p_i \in P} (y_i - a x_i - b)^2$$

← ordinary least squares regression

# 1. Closed Form Solution

$|P| = n$

$$\bar{P}_x = \text{average}\{x_i\} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{P}_y = \text{average}\{y_i\}$$

$$\text{Cov}[P_x, P_y] = \frac{1}{n}\sum_{p_i \in P} (P_{x_i} - \bar{P}_x)(P_{y_i} - \bar{P}_y)$$

$$\text{Var}[P_x] = \text{Cov}[P_x, P_x]$$

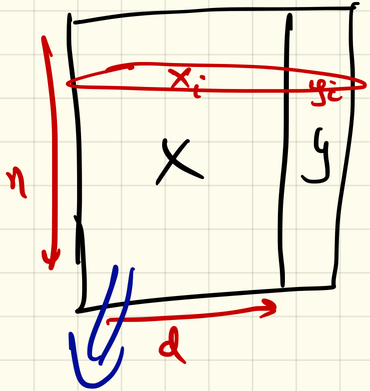
minimize least squares

$$a = \frac{\text{Cov}[P_x, P_y]}{\text{Var}[P_x]} = \left(\frac{\langle P_x, P_y\rangle}{\|P_x\|^2} \text{ if centered}\right)$$

$$b = \bar{P}_y - a\bar{P}_x$$

## 2. Extends to $x \in \mathbb{R}^d$

Input $P = (X, y)$ $\quad X \subset \mathbb{R}^{n \times d}$ $\quad y \in \mathbb{R}^n$




$p_i$ $\quad x_i \in \mathbb{R}^d$ $\quad x_i = (x_{i1}, x_{i2}, \dots x_{id})$

$$f(\overset{\overset{\mathbb{R}^d}{d}}{x}) = (b = a_0) + \sum_{j=1}^{d} x_{ij}\, a_j$$

<u>Goal</u>: $\underset{\substack{a=(a_0 \dots a_d) \\ f_a}}{\arg\min} \sum_{p_i \in P} \left(y_i - f_a(x_i)\right)^2$

$$\tilde{X} = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1d} \\ 1 & x_{i1} & x_{i2} & \dots & x_{id} \end{bmatrix}}_{d+1} \Big\downarrow n$$

$$\Longrightarrow \quad a = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T y$$

↑ minimize least form-lation

# 3. Polynomial Fits

<u>Input</u> $P \in \mathbb{R}^2$ $\quad p_i = (x_i, y_i) \in \mathbb{R}^2$

<u>goal</u>: $g_a$: $y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \dots a_t x^t$

$$= \sum_{j=0}^{t} a_j x^j$$

$$\underset{a \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \sum_{p \in P} \left( y_i - g_a(x_i) \right)^2$$

$$\tilde{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & \dots & x_1^t \\ & & & & & \\ 1 & x_i & x_i^2 & x_i^3 & \dots & x_i^t \\ & & & & & \end{bmatrix}$$

## 4. Gauss – Markov Thm

↳ "minimum variance solution"

Assuming

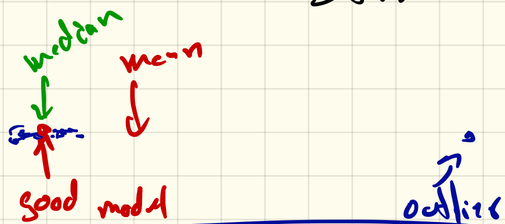
- unbias solution, expected error 0
- all errors $r_i$ are not known to correlated

---

1. minimize projection instead of vertical?
   PCA
2. $(X^T X)^{-1}$ expensive (streaming, SGD)
3. Minimize something other than $\sum r_i^2$
   → Robust Estimator (Theil-Sen Est)
4. biased solutions : regularization
   lasso

# Theil - Sen Estimator

**Breakdown point** of estimator is largest fraction of data points which can be arbitrary outliers and the estimator is still "OK".




median has a breakdown point of $1/2$.

median in $\mathbb{R}^1$

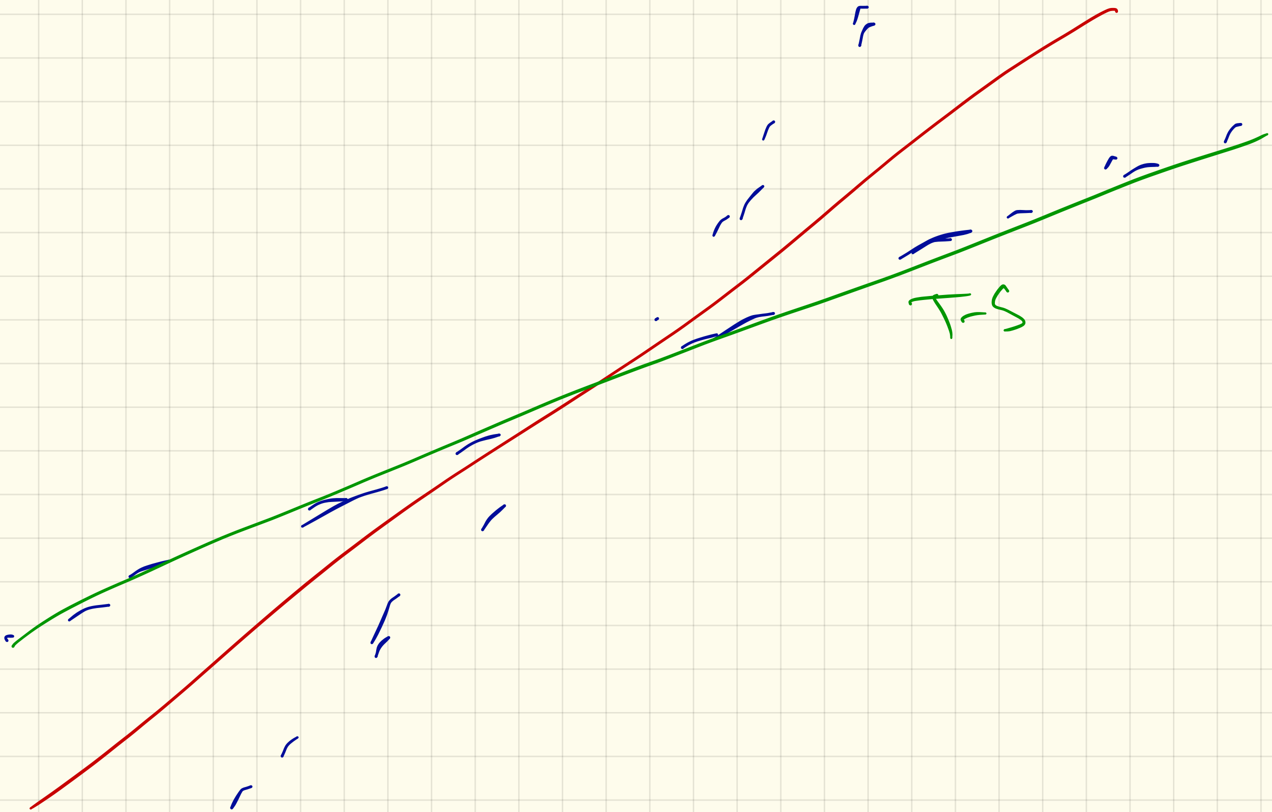
- half points left, half right
- $\underset{m \in \mathbb{R}}{\operatorname{argmin}} \sum_{p \in P} |p - x|$

$$S = \left\{ s_{ij} \;\middle|\; \frac{y_j - y_i}{x_j - x_i},\ x_i < x_j \right\}$$

$$a = \operatorname{median}\{ s_{ij} \in S \}$$

$$b = \operatorname{median}\{ y_i - a\, x_i \}$$

bkdn pt 0.293


T-S

# Tikhonov Regularization (Ridge Regression)

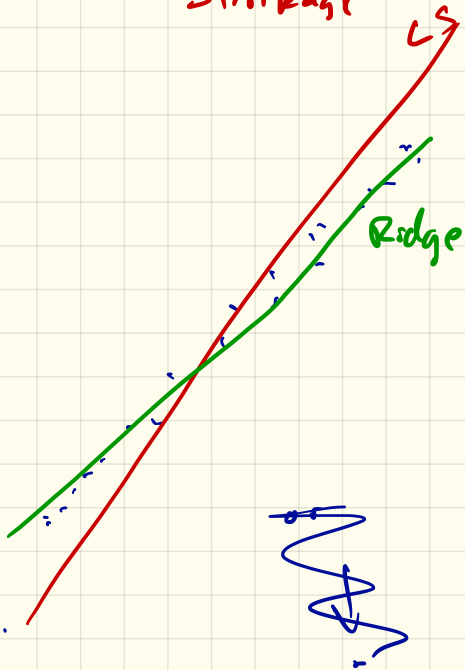
Cost $L_{2,s}(P, a) = \sum_{p_i \in P} (y_i - a x_i)^2 + s\|a\|_2^2$

shrinkage

closed form soln

$$a = (\hat{X}^T \hat{X} + s^2 I)^{-1} \hat{X}^T y$$

$$= \frac{\langle P_x, P_y \rangle}{\langle P_x, P_x \rangle + s^2}$$




- Choosing $s$
  - + cross-validation
  - + $\exists s$ variance at least as small as LS.

Lasso (basis pursuit)

Cost: $L_{1,s}(P,a) = \sum_{p_i \in P} (y_i - a x_i)^2 + s\|a\|_1$

$\rightarrow$ no "simple" closed form soln.

when $a \in \mathbb{R}^d$ $d$ large

then biases toward "sparse" $a$.

$\hookrightarrow$ many $a_i = 0$

Even works when $d >> n$