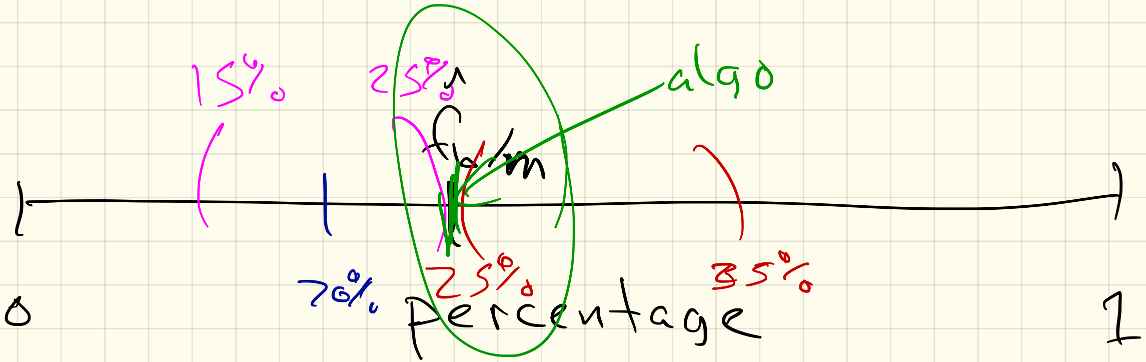
L13: Regression



might?
must?

Regression

Data (x, y)

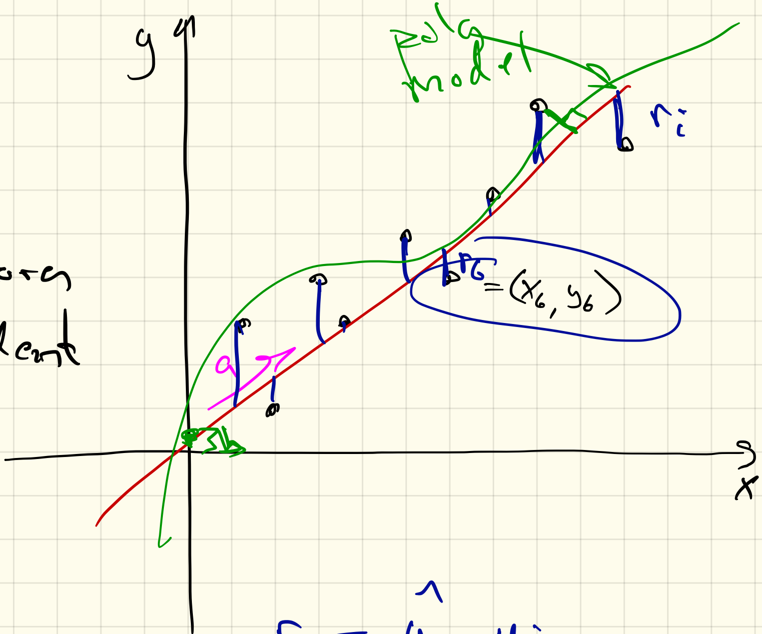
$x \in \mathbb{R}$ explanatory

$y \in \mathbb{R}$ dependent

fit (x, y) w line

$\hat{y}_i = f(x_i) = a x_i + b$

$0.8 x_i + 0.2$



$r_i = \hat{y}_i - y_i$

minimize $\sum_i r_i^2$

Goal Find a, b

$$\underset{a,b}{\text{minimize}} \sum_i (y_i - a x_i - b)^2$$

$\bar{x} = \frac{1}{n} \sum_i x_i \qquad \bar{y} = \frac{1}{n} \sum_i y_i$

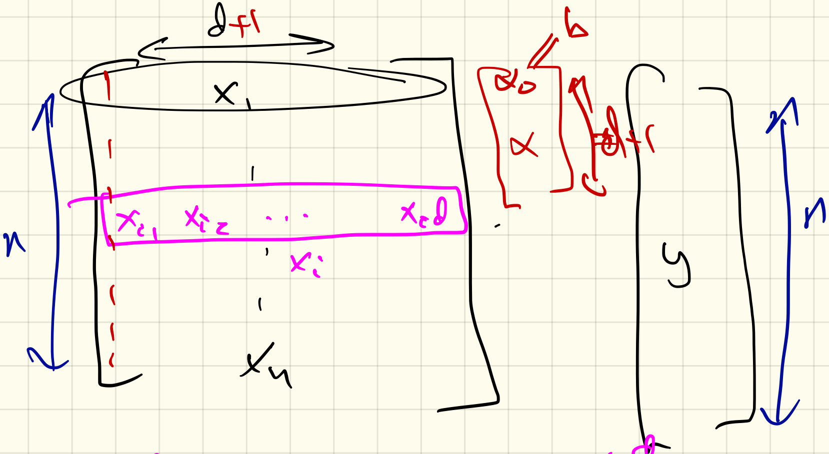
$\mathrm{Cov}(x, y) = \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y})$

$a = \frac{\mathrm{Cov}(x, y)}{\mathrm{Cov}(x, x)} \qquad b = \bar{y} - a\bar{x}$

Data (X, y)

$X \in \mathbb{R}^{n \times d}$

$y \in \mathbb{R}^{n}$



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

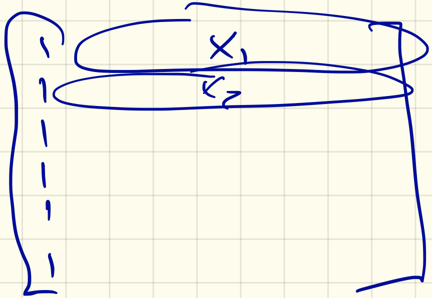
Goal $\alpha \in \mathbb{R}^{d+1}$ $\quad x_i \in \mathbb{R}^{d}$

$$\sum_i \left(y_i - \langle \alpha, \tilde{x}_i \rangle\right)^2$$

$\downarrow$ soln

$$\alpha = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T y$$

$\tilde{X} =$



$\in \mathbb{R}^{n \times d+1}$

$\tilde{x}_i = [\underset{x_{i0}}{1}, x_{i1}, x_{i2}, \ldots x_{id}]$

Polynomial Regression

Model $\hat{y}_i = M(x_i) = \alpha_0 x_i^0 + \alpha_1 x_i + \alpha_2 x_i^2 + \alpha_3 x_i^3$ ... p

$(+\alpha_4 \sin(x_i))$

Input (x, y) $x_i \in \mathbb{R}$

$x \in \mathbb{R}$

$y \in \mathbb{R}$

map

$\hookrightarrow X \in \mathbb{R}^{p+1}$

Goal

$\alpha \in \mathbb{R}^{p+1}$

minimize

$$\sum_i \left(y_i - \left(\sum_{j=0}^{p} \alpha_j \cdot x_i^j \right) \right)^2$$

$$\tilde{X}_p = \overset{n}{\updownarrow} \underbrace{\begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ 1 & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & & x_n^p \end{bmatrix}}_{p+1}$$

soln

$$\alpha = (\tilde{X}_p^T \tilde{X}_p)^{-1} \tilde{X}_p^T y$$

Gauss-Markov Thm

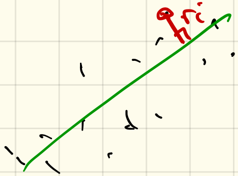
- $\alpha = (X^TX)^{-1}X^Ty$ optimal

(1) Goal $\sum_i (y_i - M(x_i))^2$ — MLE Normal Noise; good algo

(2) Assume residuals $r_i = y_i - \hat{y}$ unbiased, uncorrelated

(3) Want unbiased solution

Dim. Red



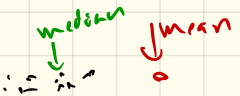
PCA

OLS: ok if units on x_i, y_i don't match

$\hat{y} = \alpha_0 + \alpha_1 x_i$

Robust Regression

mean vs. median



Theil-Sen Estimator $\left\{\frac{(y_i - y_{i'})}{(x_i - x_{i'})}\right\}$

a = median all slopes

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

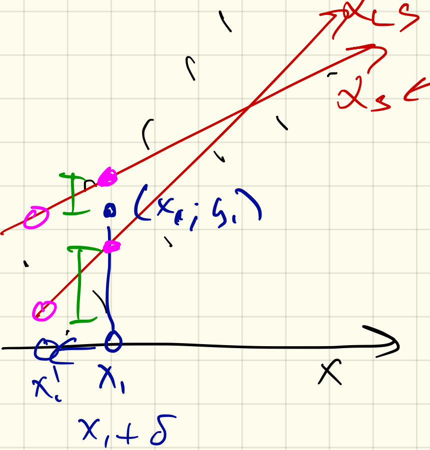
Ridge Regression (Tikhonov Regularization)

$$\text{Goal} \quad \sum_i \left(y_i - \langle \alpha, x_i \rangle\right)^2 + \underbrace{s\|\alpha\|_2^2}_{\text{penalty on complex model } \alpha}$$

$\exists\, s$ s.t. RR α_s is better than α_{LS} (unseen data)

$$\alpha_s = \left(X^T X + s^2 \underbrace{I}_{}\right)^{-1} X^T y$$

$\hookrightarrow$ $d \times d$ identity matrix $\begin{bmatrix} s^2 & & 0 \\ & s^2 & \\ 0 & & \ddots \; s^2 \end{bmatrix}$



smaller slope

$=$

$\|\alpha\|^2$ smaller

Lasso (basis pursuit)

$$L_{1,s}(x,y,\alpha) = \sum_i \left(y_i - \langle x_i, \alpha \rangle\right)^2 + \underbrace{s\|\alpha\|_1}_{L_1\text{-regularization}}$$

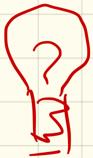
convex
not smooth

no simple linear algebra soln

→ Induce sparsity in α

↳ many coefficients $\alpha_i = 0$

Not perfect variable selection



only select variable iff $\alpha_i \neq 0$