Lecture 3:

Maximum Likelihood Estimation (MLE), Linear Regression, Linear Models

Markus Hohle

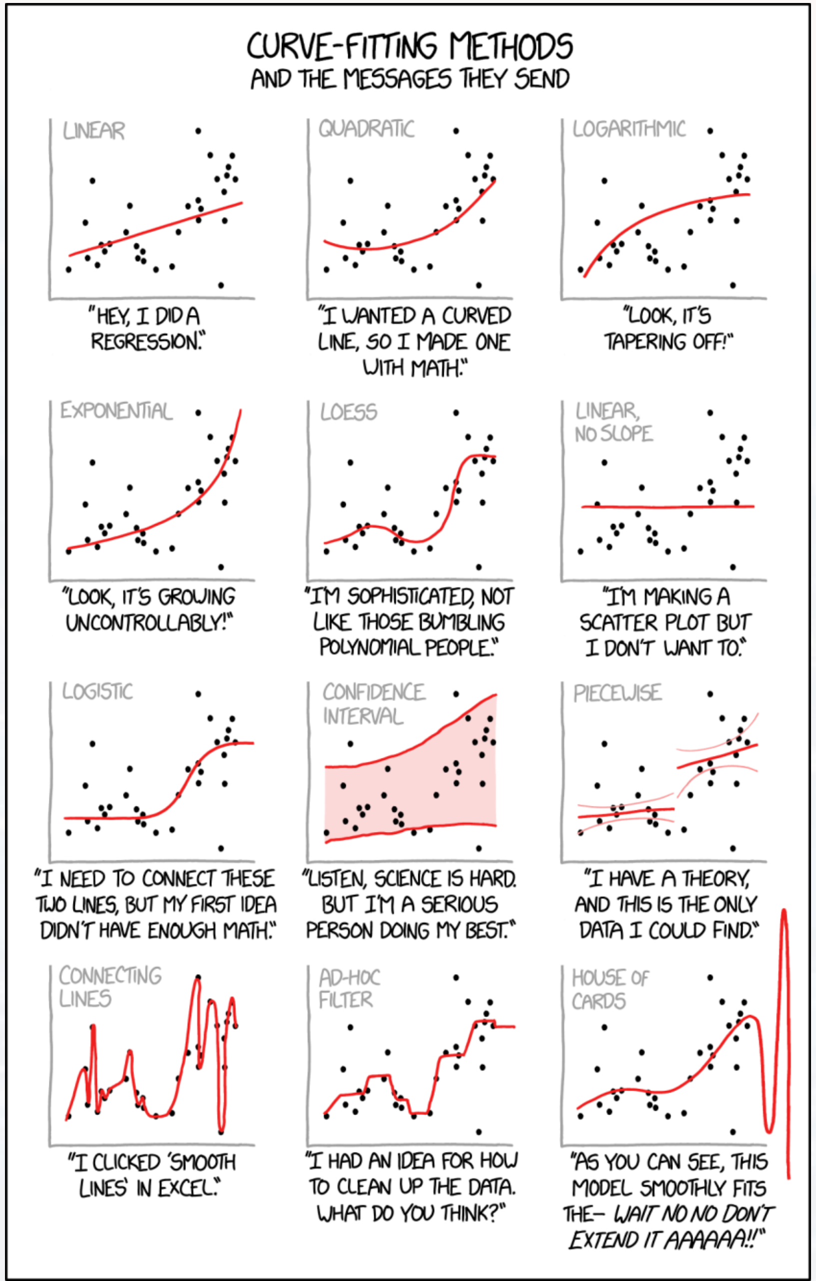
University California, Berkeley

Bayesian Data Analysis and Machine Learning for Physical Sciences

Course Map

Module 1	Maximum Entropy and Information, Bayes Theorem
Module 2	Naive Bayes, Bayesian Parameter Estimation, MAP
Module 3	**MLE, Lin Regression, Model selection: Comparing Distributions**
Module 4	Model Selection: Bayesian Signal Detection
Module 5	Variational Bayes, Expectation Maximization
Module 6	Stochastic Processes
Module 7	Monte Carlo Methods
Module 8	Markov Models, Graphs
Module 9	Machine Learning Overview, Supervised Methods
Module 10	Unsupervised Methods
Module 11	ANN: Perceptron, Backpropagation
Module 12	ANN: Basic Architecture, Regression vs Classification, Backpropagation again
Module 13	Convolution and Image Classification and Segmentation
Module 14	TBD (GNNs)
Module 15	TBD (RNNs and LSTMs)
Module 16	TBD (Transformer and LLMs)

Maximum Likelihood Estimation, Linear Regression



<u>Outline</u>

Maximum Likelihood Estimation (MLE)

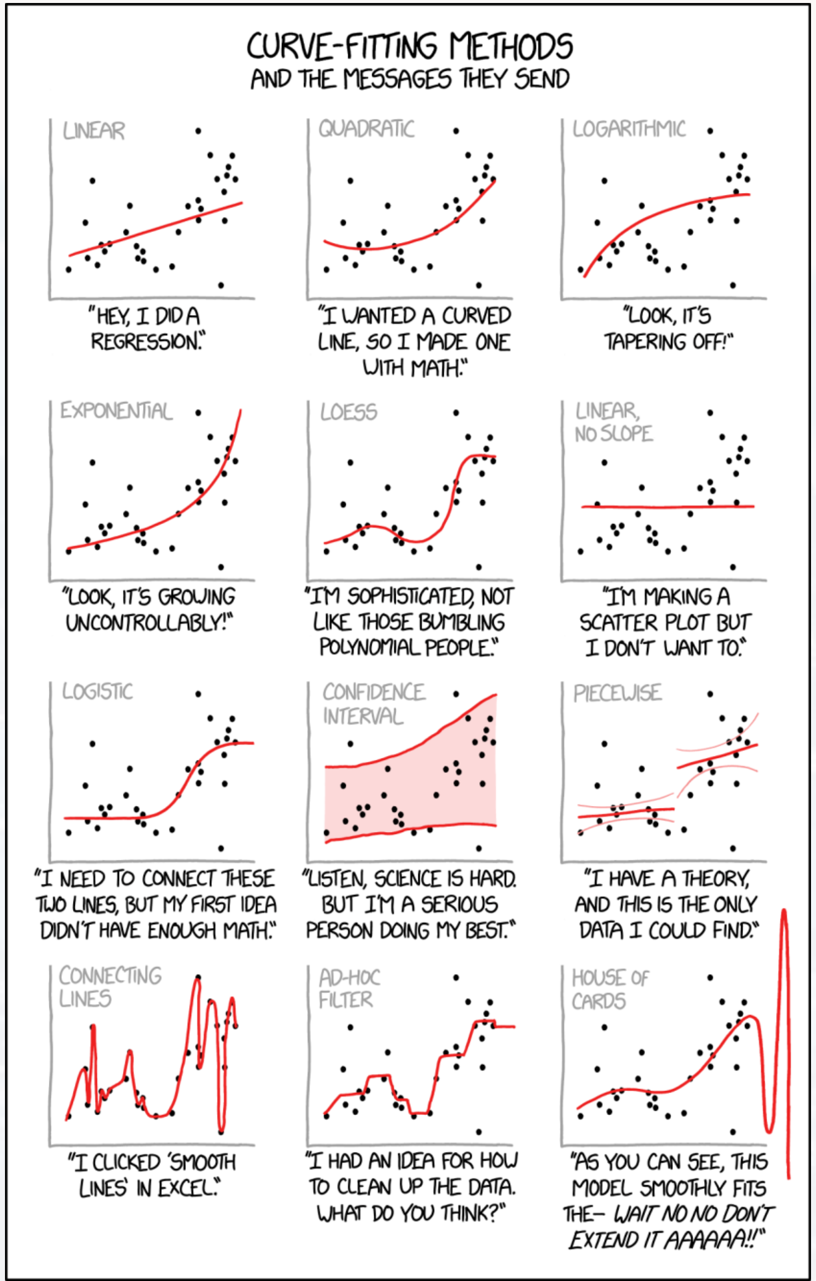
- idea
- examples

Linear Models

- classical model
- regularization
- extensions

Regression

Maximum Likelihood Estimation, Linear Regression



Outline

Maximum Likelihood Estimation (MLE)

- idea
- examples

Linear Models

- classical model
- regularization
- extensions

Regression

Maximum Likelihood Estimation, Linear Regression:

Bayesian:

$$P(\theta|D) = \frac{\overbrace{\boxed{P(D|\theta)}}^{\text{likelihood function}}\overbrace{\boxed{P(\theta)}}^{\text{prior}}}{\underbrace{\boxed{P(D)}}_{\text{evidence (const wrt q)}}}$$

$$\int P(\theta|D)\ d\theta = 1$$

parameter θ

"Classical" approach: **M**aximum **L**ikelihood **E**stimation

$x = \{x_1, x_2, \dots x_i, \dots, x_N\}$ data set of N independent observations

$\theta = \{\theta_1, \theta_2, \dots \theta_i, \dots, \theta_M\}$ set of parameters θ

$\theta = q$ (binomial)
$\theta = \{\mu, \sigma^2\}$ (normal dist)

$f_i(\theta, x_i)$ distribution f_i from which x_i has been drawn

Maximum Likelihood Estimation, Linear Regression:

<u>"Classical" approach:</u> **M**aximum **L**ikelihood **E**stimation

$x = \{x_1, x_2, \dots x_i, \dots, x_N\}$ data set of N independent observations

$\theta = \{\theta_1, \theta_2, \dots \theta_i, \dots, \theta_M\}$ set of parameters θ

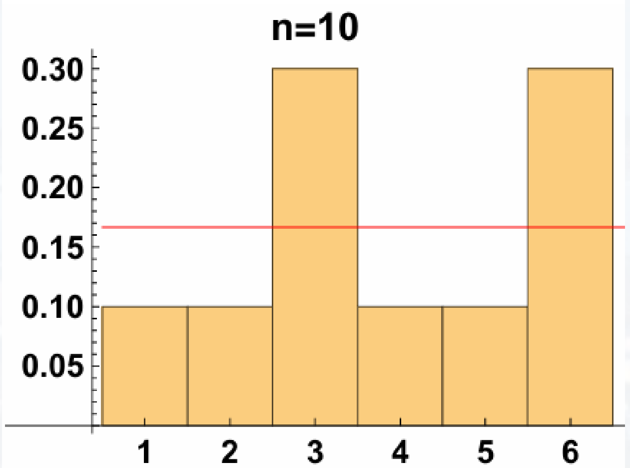
$f_i(\theta, x_i)$ distribution f_i from which x_i has been drawn

$L(\theta, x) = \prod_{i=1}^{N} f_i(\theta, x_i)$ joint density

goal: finding those $\hat{\theta}$ that maximize $L(\theta, x)$

$$\boxed{\hat{\theta} = \underset{\theta}{argmax}\{L(\theta, x)\}}$$

We aim to find the most likely one!



$\sigma = 2$, $\mu = 3.5$

$\sigma = 2$, $\mu = 5.0$

$\sigma = 1.5$, $\mu = 3.5$

$\sigma = 7.0$, $\mu = 1.0$

....and so on

Maximum Likelihood Estimation, Linear Regression:

<u>"Classical" approach:</u> **M**aximum **L**ikelihood **E**stimation

$$L(\theta, x) = \prod_{i=1}^{N} f_i(\theta, x_i)$$ joint density

goal: finding those $\hat{\theta}$ that maximize $L(\theta, x)$

$$\boxed{\hat{\theta} = \underset{\theta}{argmax}\{L(\theta, x)\}}$$

x: data set
θ: parameter
$f_i(\theta, x_i)$: density function

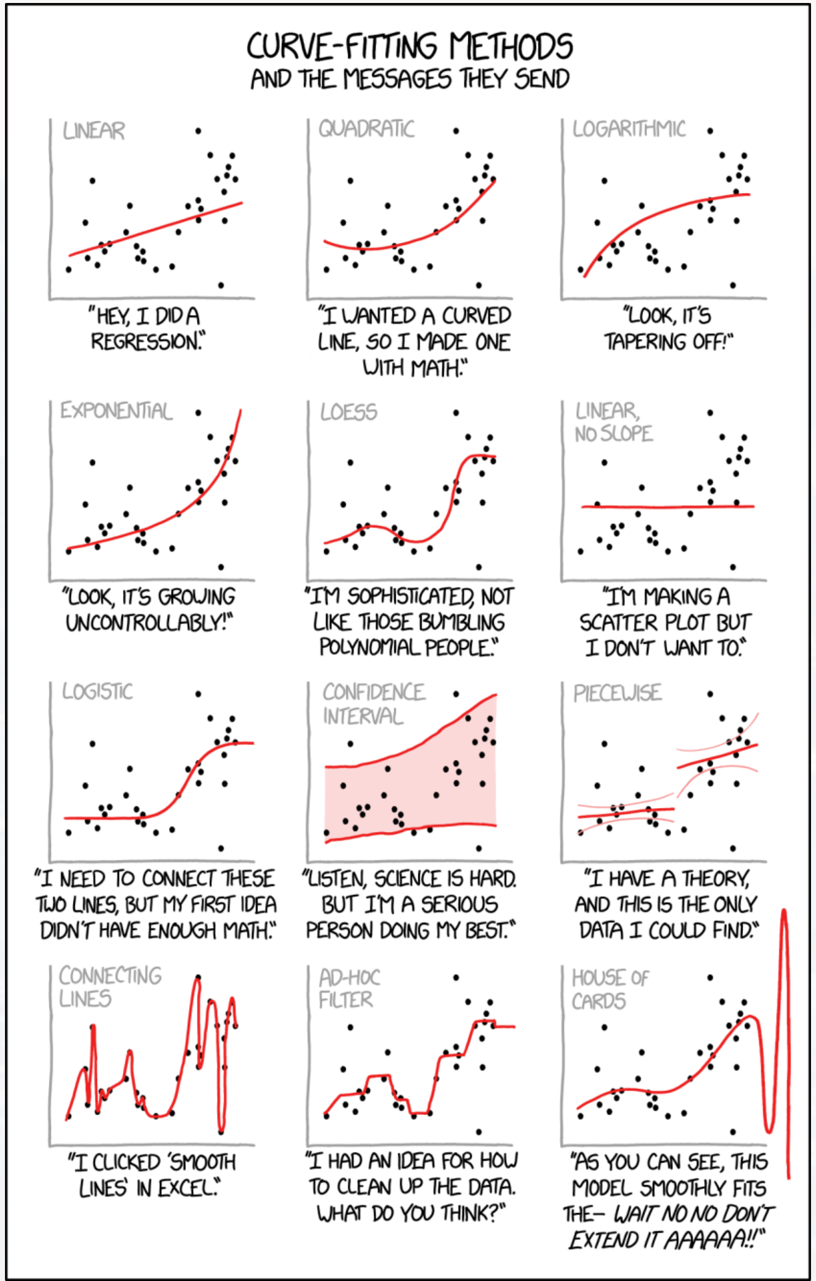
note:

- often we use $\mathrm{l}(\theta, x) = \ln[L(\theta, x)]$ for convenience
- we find $\hat{\theta}$ via $\frac{\partial L(\theta, x)}{\partial \theta_i} = 0$ for all θ_i
- equivalent to **MAP** assuming **an uniform prior**
- evaluate, if extreme is indeed a maximum:

  Hessian matrix $H = \left\{\frac{\partial^2\, \mathrm{l}(\theta, x)}{\partial \theta_i\, \partial \theta_j}\right\}$ has to be **locally** (around $\hat{\theta}$) **concave**

- we can add constrains using **Lagrangian Multipliers**

Maximum Likelihood Estimation, Linear Regression



<u>Outline</u>

Maximum Likelihood Estimation (MLE)

- idea
- examples

Linear Models

- classical model
- regularization
- extensions

Regression

Maximum Likelihood Estimation, Linear Regression:

example I: binomial distribution

$x = \{HHTHTTT \ldots.\} \qquad \theta = q$

$f(q|x) = \binom{n}{k} \boxed{q^k(1-q)^{n-k}}$ equivalent to **MAP** assuming **an uniform prior** $\dfrac{P(D|\theta)P(\theta)}{P(D)}$

$$\frac{df}{dq} = \binom{n}{k} [kq^{k-1}(1-q)^{n-k} - q^k(n-k)(1-q)^{n-k-1}] = 0$$

ignoring $q = 1$ and $q = 0$: $\qquad k(1-q) = q(n-k) \qquad q_{best} = \dfrac{k}{n}$

$$\left.\begin{aligned} q_0 &= \frac{k}{n} \\ \sigma &= \sqrt{\frac{q_0(1-q_0)}{n}} \end{aligned}\right\} \quad q = \frac{k}{n} \pm \sqrt{\frac{q_0(1-q_0)}{n}}$$

last time:
Lagrange approximation from BPE

example II: normal distribution

$$x = \{x_1, x_2, \dots x_i, \dots, x_N\} \qquad \theta = \{\mu, \sigma^2\}$$

$$f_i(\theta, x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}}$$

$$L(\theta, x) = \prod_{i=1}^{N} f_i(\theta, x_i) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2}$$

$$l(\theta, x) = \ln[L(\theta, x)] = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2$$

$$\frac{\partial l(\theta,x)}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{N}(x_i - \mu)^2 = 0$$

$$\boxed{\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

example II: normal distribution

$$x = \{x_1, x_2, \ldots x_i, \ldots, x_N\} \qquad \theta = \{\mu, \sigma^2\}$$

$$f_i(\theta, x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}}$$

$$L(\theta, x) = \prod_{i=1}^{N} f_i(\theta, x_i) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2}$$

$$l(\theta, x) = \ln[L(\theta, x)] = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2$$

$$\frac{\partial l(\theta, x)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i - \mu) = 0$$

$$\boxed{\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i}$$

$$\boxed{\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Maximum Likelihood Estimation, Linear Regression:

example II: normal distribution

$$x = \{x_1, x_2, \dots x_i, \dots, x_N\} \qquad \theta = \{\mu, \sigma^2\}$$

$$f_i(\theta, x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}}$$

$$L(\theta, x) = \prod_{i=1}^{N} f_i(\theta, x_i) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2}$$

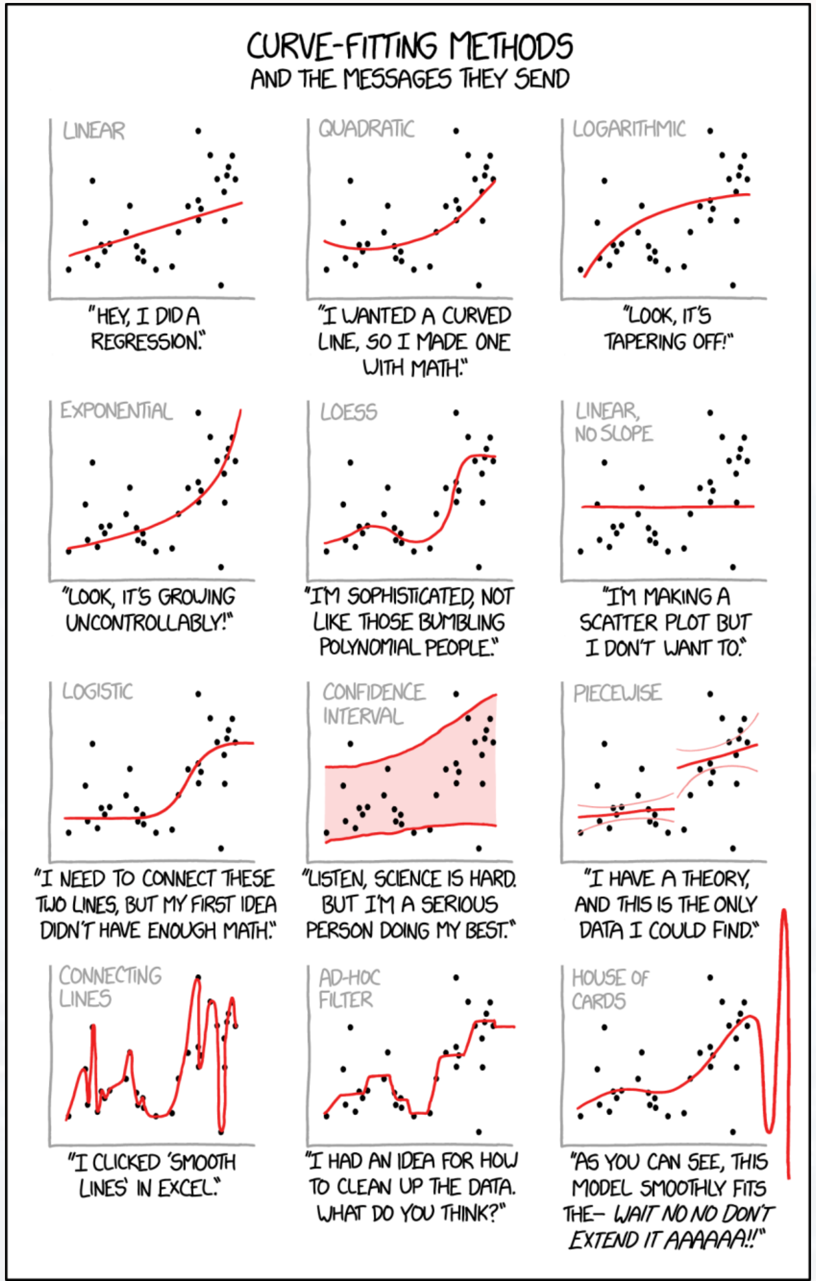
$$l(\theta, x) = \ln[L(\theta, x)] = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2$$

$$\boxed{\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2}$$

$$\boxed{\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{i=1}^{N} x_i\right)^2 = \frac{1}{N}\sum_{i=1}^{N} x_i{}^2 - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i=j}^{N} x_i\, x_j$$

Maximum Likelihood Estimation, Linear Regression



Outline

Maximum Likelihood Estimation (MLE)

- idea
- examples

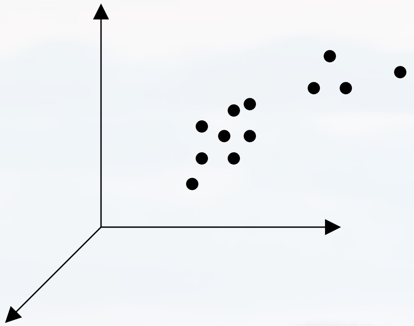
Linear Models

- classical model
- regularization
- extensions

Regression

Maximum Likelihood Estimation, Linear Regression:

<u>idea:</u> data point y_k in N dimensional feature space



$y_k = f(x_1, \dots x_n, \dots x_N) + \epsilon_k$ for each data point k

$$y_k = \beta_0 + \sum_{n=1}^{N} \beta_n x_n + \epsilon_k$$

y:	**response**
x:	**regressors** (assumed to be independent)
β:	**factors** (how a regressor contributes to the response)
β_0:	**intercept**
ε:	**error** (stochasticity of the data, assumed to be normally dist.)

$$\underbrace{\begin{pmatrix} y_1 \\ \dots \\ y_k \\ \dots \\ \dots \\ y_K \end{pmatrix}}_{\text{Y}} = \underbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1n} & \dots & x_{1N} \\ \dots & \dots & & & \dots & & \\ 1 & x_{k1} & & & x_{kn} & & \\ 1 & \dots & & & \dots & & \\ 1 & x_{K1} & x_{K2} & \dots & x_{Kn} & \dots & x_{KN} \end{pmatrix}}_{\text{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_n \\ \dots \\ \beta_N \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_k \\ \dots \\ \varepsilon_K \end{pmatrix}}_{\varepsilon}$$

$$\boxed{Y = X\beta + \varepsilon}$$

Maximum Likelihood Estimation, Linear Regression:

$$Y = X\beta + \varepsilon$$

***y*:**	**response**
***x*:**	**regressors** **(assumed to be independent)**
β:	**factors** **(how a regressor contributes to the response)**
$\boldsymbol{\beta_0}$**:**	**intercept**
ε:	**error** **(stochasticity of the data, assumed to be normally dist.)**

classical model
regularizations
extensions

<u>fitting:</u> finding the best β in terms of minimizing the errors

$$(Y - X\beta)^T(Y - X\beta) = \sum_k {\varepsilon_k}^2$$

$$\frac{\partial}{\partial\beta}\sum_k {\varepsilon_k}^2 = 0$$

$$\beta_{best} = \hat{\beta} = (X^T X)^{-1} X^T Y$$

the model

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

X and *Y* are all observables

hat matrix $\mathrm{H} \coloneqq \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$

$H = H^T$ (symmetry)

$HH = H \rightarrow H^n = H$ (idempotency)

$$Y = X\beta + \varepsilon$$

y:	**response**
x:	**regressors (assumed to be independent)**
β:	**factors (how a regressor contributes to the response)**
β_0:	**intercept**
ε:	**error (stochasticity of the data, assumed to be normally dist.)**

classical model
regularizations
extensions

evaluating the result:

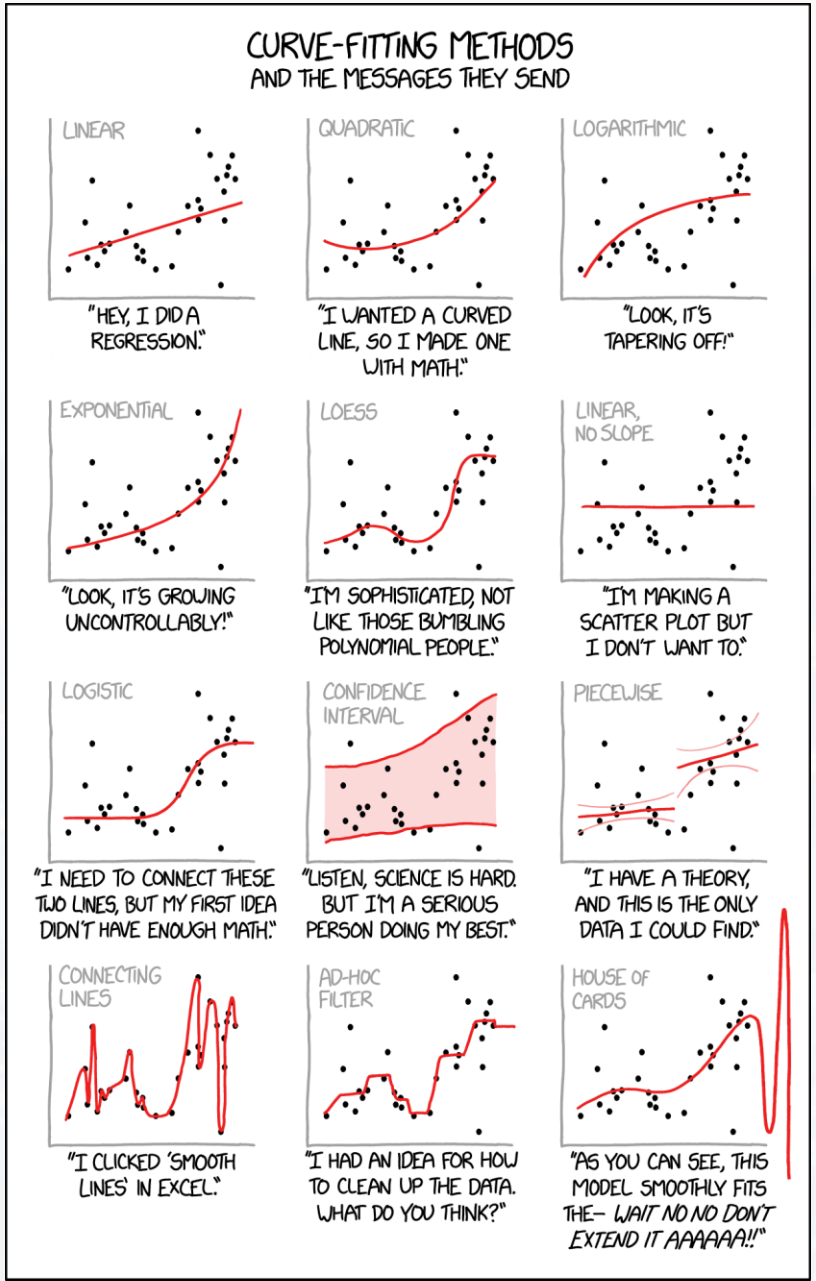
$$\hat{\varepsilon} = Y - X\hat{\beta} = Y - \hat{Y} = (I - H)Y$$

hat matrix $\mathrm{H} \coloneqq \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$

$$\hat{\varepsilon}^T\hat{\varepsilon} = [(I - H)Y]^T(I - H)Y = Y^T(I - H)^T(I - H)Y = Y^T(I - H)Y$$

sum of squared errors (SSE)

Maximum Likelihood Estimation, Linear Regression



Outline

Maximum Likelihood Estimation (MLE)

- idea
- examples

Linear Models

- classical model
- regularization
- extensions

Regression

Maximum Likelihood Estimation, Linear Regression:

classical model
regularizations
extensions

$$\hat{\beta} = \underset{\beta}{argmin}\left\{\frac{1}{N}\|Y - X\beta\|^2\right\}$$

λ *Lagrangian Multiplier*

$$\hat{\beta} = \underset{\beta}{argmin}\left\{\frac{1}{N}\|Y - X\beta\|^2 + \boldsymbol{\lambda\|\beta\|^1}\right\}$$

the Loss Function
$\boldsymbol{L(X, Y, \lambda)}$

L1 or **L**east **a**bsolute **s**hrinkage and **s**election **o**perator
- encourages **sparsity** of β
- reduces **overfitting**

$$\hat{\beta} = \underset{\beta}{argmin}\left\{\frac{1}{N}\|Y - X\beta\|^2 + \boldsymbol{\lambda\|\beta\|^2}\right\}$$

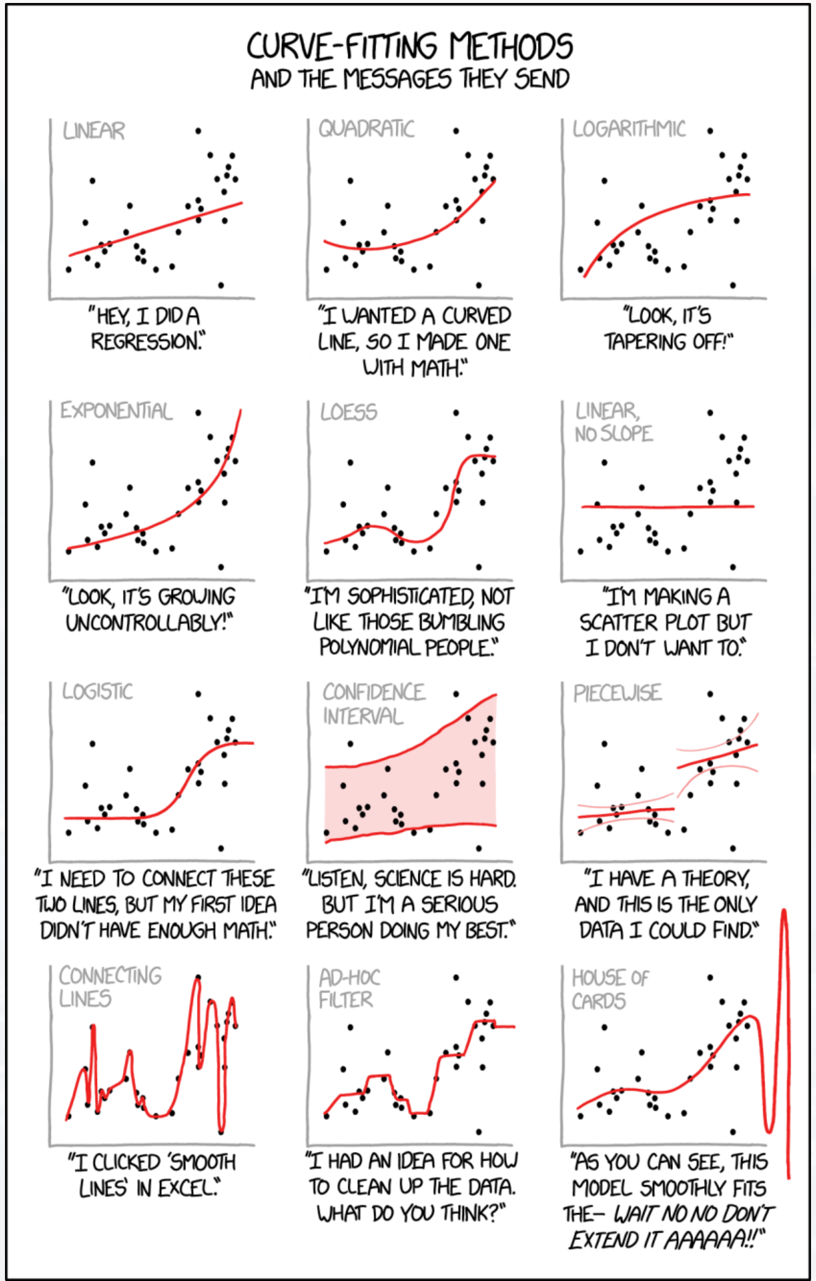
L2 or **Ridge**
- **penalizes large $\boldsymbol{\beta}$**

$$\hat{\beta} = \underset{\beta}{argmin}\left\{\frac{1}{N}\|Y - X\beta\|^2 + \boldsymbol{\lambda\, max(0, -\beta)}\right\}$$

- **penalizes negative $\boldsymbol{\beta}$**

...and so on

Maximum Likelihood Estimation, Linear Regression



Outline

Maximum Likelihood Estimation (MLE)

- idea
- examples

Linear Models

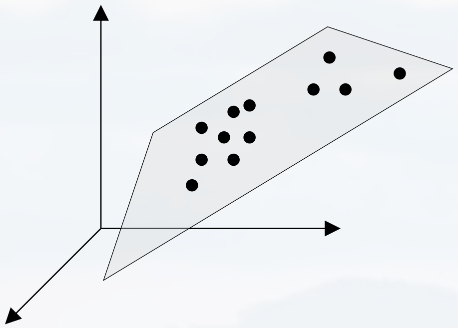
- classical model
- regularization
- extensions

Regression

Maximum Likelihood Estimation, Linear Regression:

$$y_k = \beta_0 + \sum_{n=1}^{N} \beta_n x_n + \epsilon_k$$

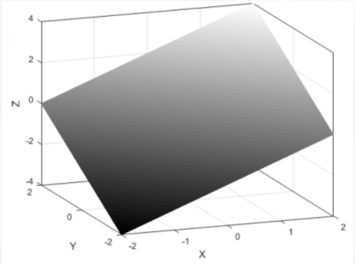
the classical model is very limiting



general: linear refers to the **factors**

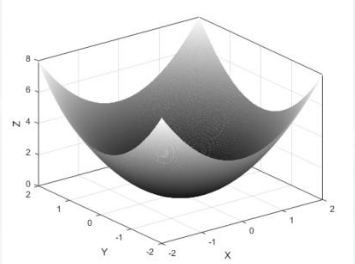
$$y_k = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

2*D* plane in 3*D* space



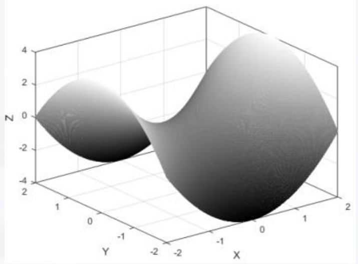
$$y_k = \beta_0 + \beta_1 {x_1}^2 + \beta_2 {x_2}^2$$

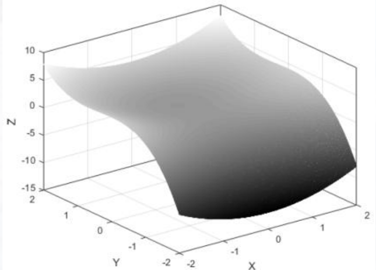
2*D* parabolic

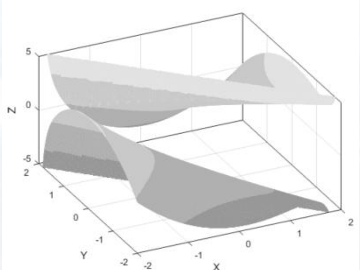


$$y_k = \beta_0 + \beta_1 {x_1}^2 - \beta_2 {x_2}^2$$

2*D* hyperbolic





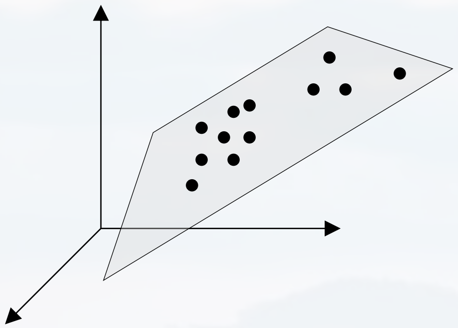


...and many more...

Maximum Likelihood Estimation, Linear Regression:

$$y_k = \beta_0 + \sum_{n=1}^{N} \beta_n x_n + \epsilon_k$$

the classical model is very limiting

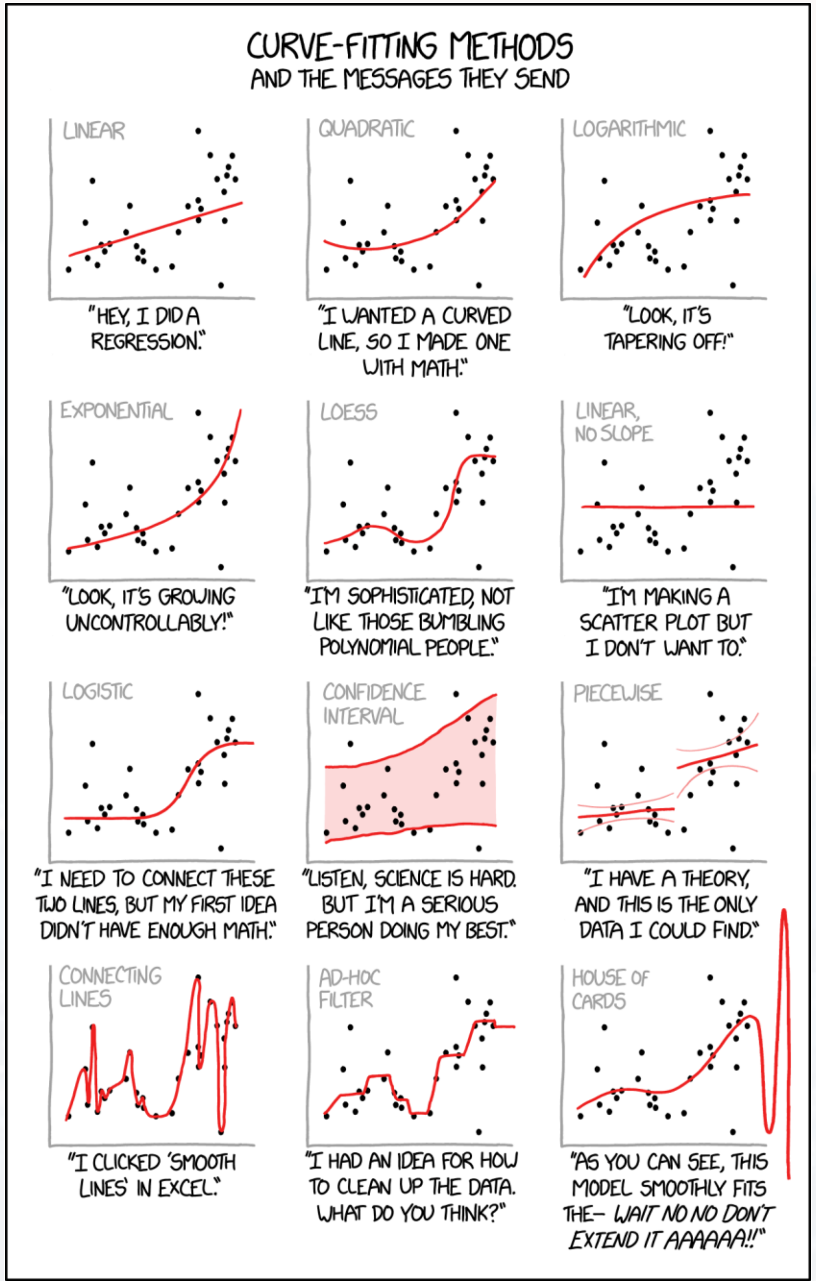


general: linear refers to the **factors**

- x_n can be subject to any **basis function** $\varphi_n(x_n)$
- the matrix X is then called **design matrix** $\phi_{i,j} = \varphi_j(x_i)$
- $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T Y \ \rightarrow (\phi^T \phi)^{-1} \phi^T Y$ is called **normal equations**
- most textbooks: rows = number of observations/samples (here K)
  columns = number of features (here N)
  → one φ_j per column j
- 1D curve fit → **one feature** x: $y_k = \beta_0 + \sum_{n=1}^{N} \beta_n \, \varphi_n(x_n) + \epsilon_k$

  polynomial fit $y_k = \beta_0 + \sum_{n=1}^{N} \beta_n x_{k,n}^n + \epsilon_k$

Maximum Likelihood Estimation, Linear Regression



<u>Outline</u>

Maximum Likelihood Estimation (MLE)

- idea
- examples

Linear Models

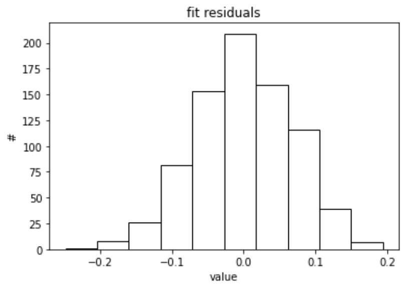
- classical model
- regularization
- extensions

Regression

Maximum Likelihood Estimation, Linear Regression:

We now want to apply MLE to our problem and see how it leads to the same structure for $\hat{\boldsymbol{\beta}}$

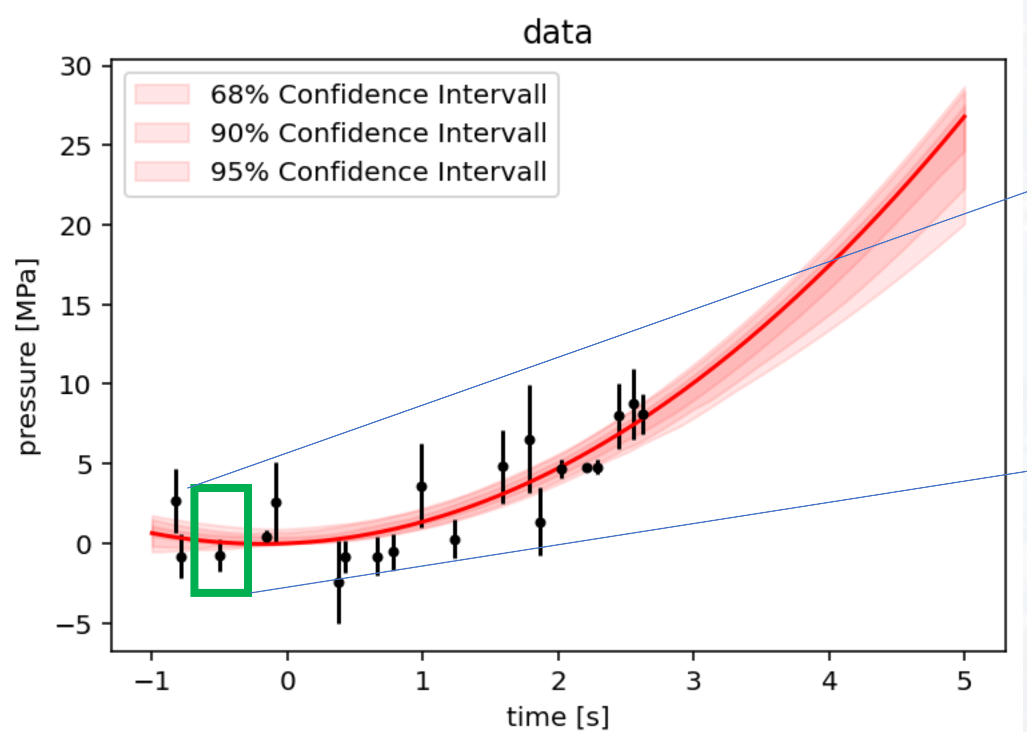
$Y = X\beta + \varepsilon$ ε: additive gaussian noise around $\mu = 0$

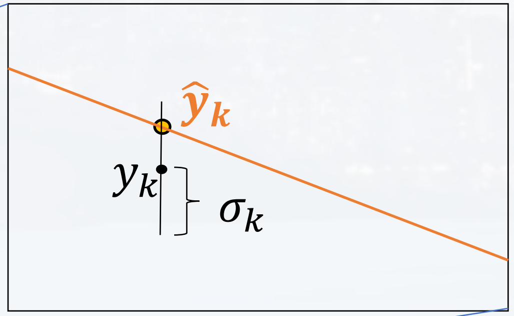


function $\boldsymbol{y(x, \beta)}$

we aim to predict a target value $\hat{\boldsymbol{y}}$

$\hat{\boldsymbol{y}} = \boldsymbol{y(x, \beta)} + \varepsilon$ for each data point *k*

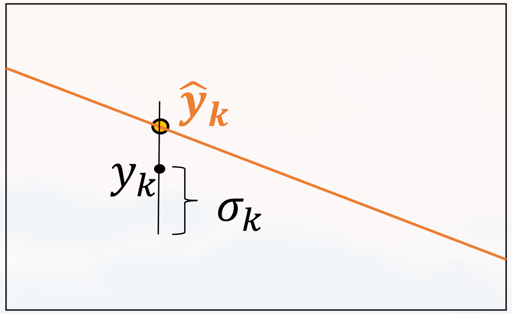




Maximum Likelihood Estimation, Linear Regression:

$$p(\hat{y}|\beta,\sigma) = \prod_{k=1}^{K} N(\hat{y}_k|\beta^T\phi(x_k),\sigma)$$

note: we assume $\sigma_k = const = \sigma$ for simplification



$$\ln[p(\hat{y}|\beta,\sigma)] = -\frac{K}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^{K}[\hat{y}_k - \beta^T\phi(x_k)]^2$$

We want to find $\hat{\boldsymbol{\beta}}$ via MLE: gradient of $\ln[p(\hat{y}|\beta,\sigma)]$ wrt β

$$grad\{\ln[p(\hat{y}|\beta,\sigma)]\}_\beta = \frac{1}{\sigma^2}\sum_{k=1}^{K}[\hat{y}_k - \beta^T\phi(x_k)]\phi(x_k)^T$$

setting the gradient to zero: $\boxed{\hat{\boldsymbol{\beta}} = (\phi^T\phi)^{-1}\phi^T\hat{y}}$ as before

for different σ_k we find $\chi^2 = \sum_{k=1}^{K}\left(\frac{\hat{y}_k - \beta^T\phi(x_k)}{\sigma_i}\right)^2$

Thank you very much for your attention!