Statistical Learning STA4042

INTERPRETBILITY VS. FLEXIBILITY

SCHOOL OF DATA SCIENCE

A general regression example

Consider a general regression problem:

- p random variables $X^{(1)}, \ldots, X^{(p)}$
- a target random variable $Y = f(X^{(1)}, \ldots, X^{(p)})$.

Statistical Learning STA4042

INTERPRETBILITY VS. FLEXIBILITY

SCHOOL OF DATA SCIENCE

A general regression example

Consider a general regression problem:

- p random variables $X^{(1)}, \ldots, X^{(p)}$
- a target random variable $Y = f(X^{(1)}, \ldots, X^{(p)})$.

ex: independent drawing = individual, $X^{(1)}, \ldots, X^{(p)}$ are some charasteristics of each individual (years of education, seniority...) Y is the level of income

Statistical Learning STA4042

INTERPRETBILITY VS. FLEXIBILITY

SCHOOL OF DATA SCIENCE

A general regression example

Consider a general regression problem:

- p random variables $X^{(1)}, \ldots, X^{(p)}$
- a target random variable $Y = f(X^{(1)}, \ldots, X^{(p)})$.

ex: independent drawing = individual, $X^{(1)}, \ldots, X^{(p)}$ are some charasteristics of each individual (years of education, seniority...) Y is the level of income

Goal: Given a training data set ($= n$ independent drawings of Y and $X = (X^{(1)}, \ldots, X^{(p)})$):

$$D \equiv \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

$\rightarrow$ Approximate f with a mapping $\hat{f}_D$ s.t. $Y \approx \hat{f}_D(X)$.

Simple solution:

Look for $\hat{f} : (x^{(1)}, \ldots, x^{(p)}) \mapsto \beta_1 x^{(1)} + \cdots + \beta_p x^{(p)}$

Statistical Learning STA4042

INTERPRETBILITY VS. FLEXIBILITY

SCHOOL OF DATA SCIENCE

A general regression example

Consider a general regression problem:

- p random variables $X^{(1)}, \ldots, X^{(p)}$
- a target random variable $Y = f(X^{(1)}, \ldots, X^{(p)})$.

ex: independent drawing = individual, $X^{(1)}, \ldots, X^{(p)}$ are some charasteristics of each individual (years of education, seniority...) Y is the level of income

Goal: Given a training data set ($= n$ independent drawings of Y and $X = (X^{(1)}, \ldots, X^{(p)})$):

$$D \equiv \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

$\rightarrow$ Approximate f with a mapping $\hat{f}_D$ s.t. $Y \approx \hat{f}_D(X)$.

Simple solution:

Look for $\hat{f} : (x^{(1)}, \ldots, x^{(p)}) \mapsto \beta_1 x^{(1)} + \cdots + \beta_p x^{(p)}$

Parameters = Interpretation

(No parameter = no interpretation)

Statistical Learning STA4042

INTERPRETBILITY VS. FLEXIBILITY

SCHOOL OF DATA SCIENCE

A general regression example

Consider a general regression problem:

- p random variables $X^{(1)}, \dots, X^{(p)}$
- a target random variable $Y = f(X^{(1)}, \dots, X^{(p)})$.

ex: independent drawing = individual, $X^{(1)}, \dots, X^{(p)}$ are some charasteristics of each individual (years of education, seniority...) Y is the level of income

Goal: Given a training data set ($= n$ independent drawings of Y and $X = (X^{(1)}, \dots, X^{(p)})$):

$$D \equiv \{(x_1, y_1), \dots, (x_n, y_n)\}$$

$\rightarrow$ Approximate f with a mapping $\hat{f}_D$ s.t. $Y \approx \hat{f}_D(X)$.

Simple solution:

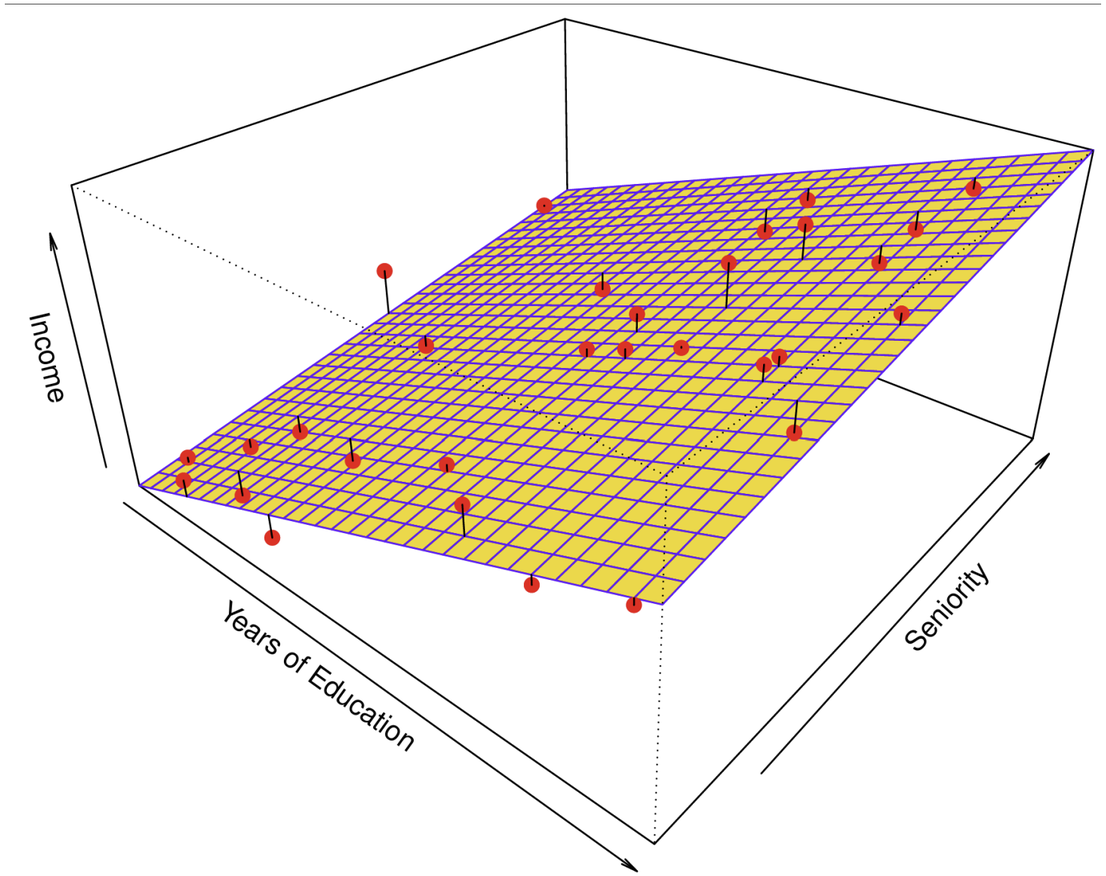
Look for $\hat{f} : (x^{(1)}, \dots, x^{(p)}) \mapsto \beta_1 x^{(1)} + \cdots + \beta_p x^{(p)}$

Parameters = Interpretation

(No parameter = no interpretation)

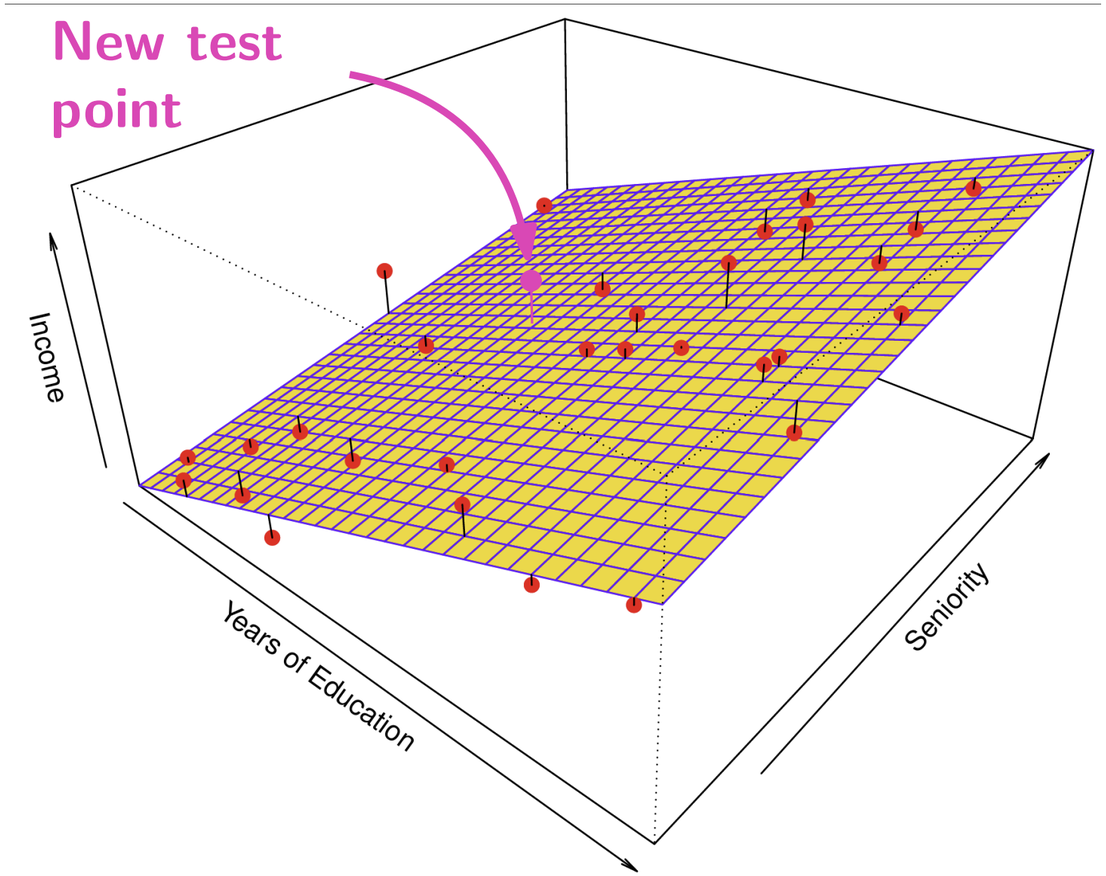
$\beta_1, \dots, \beta_p$ provide dependence on each predictor $x^{(1)}, \dots, x^{(p)}$

A general regression example



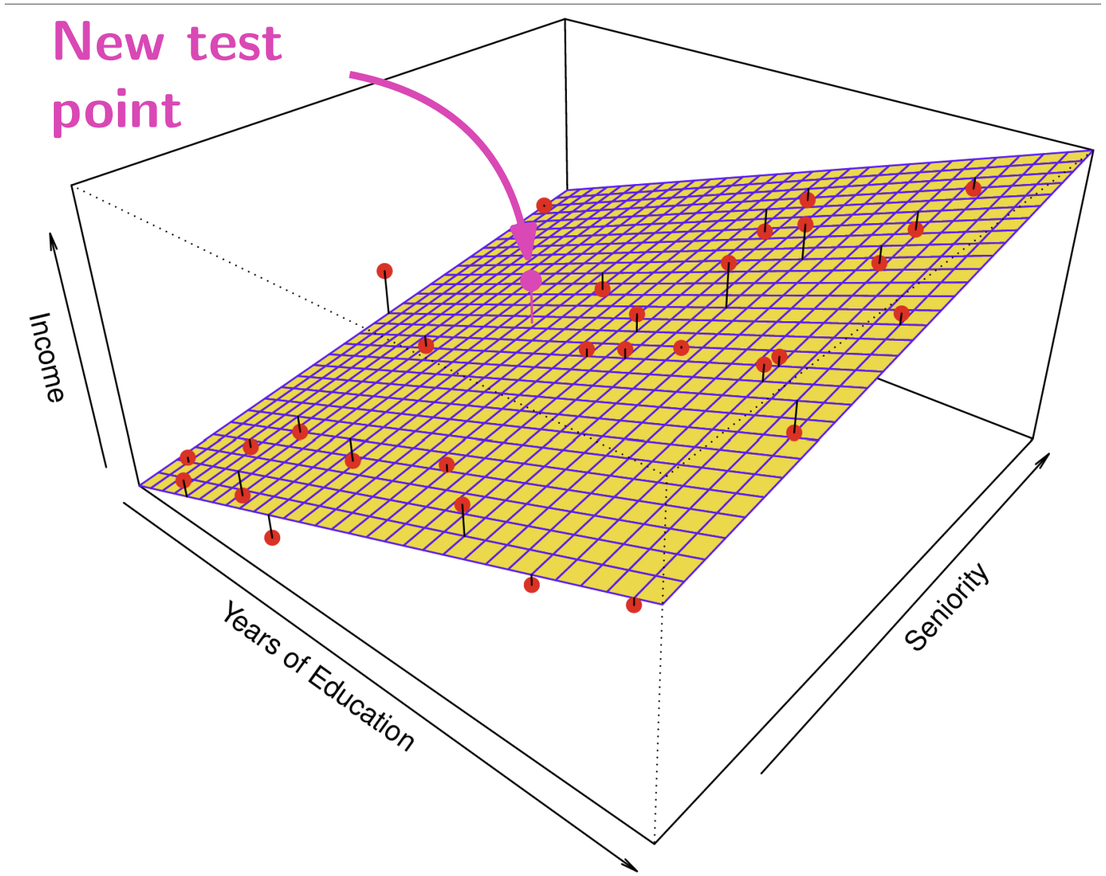
"Parametric model": look for $\beta_0, \beta_1, \beta_2$ s.t.:
"Income" $= \beta_0 + \beta_1$ "y. o. education" $+ \beta_2$ "seniority"

A general regression example



"Parametric model": look for $\beta_0, \beta_1, \beta_2$ s.t.:
"Income" $= \beta_0 + \beta_1$ "y. o. education" $+ \beta_2$ "seniority"

A general regression example

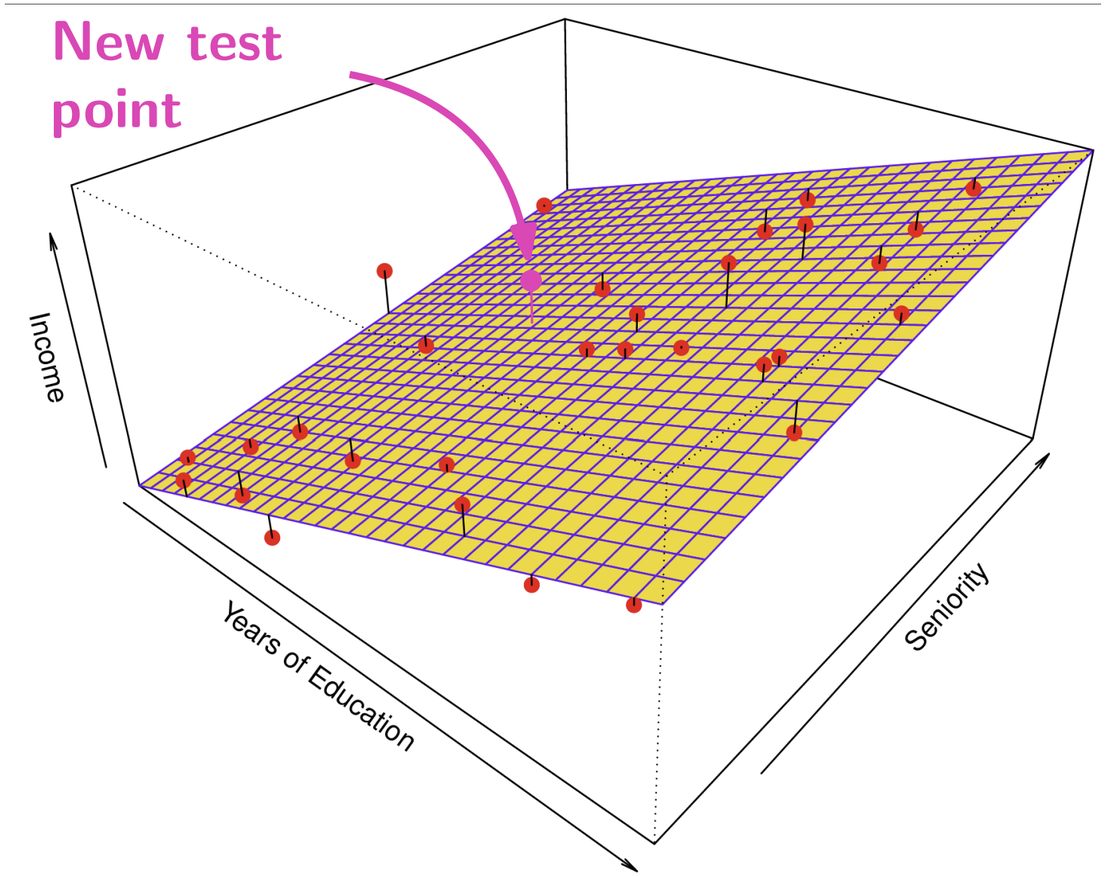


$X = (x_1, \ldots, x_n) \in \mathbb{R}^{p \times n}$: Input

"Parametric model": look for $\beta_0, \beta_1, \beta_2$ s.t.:

"Income" $= \beta_0 + \beta_1$ "y. o. education" $+ \beta_2$ "seniority"

A general regression example



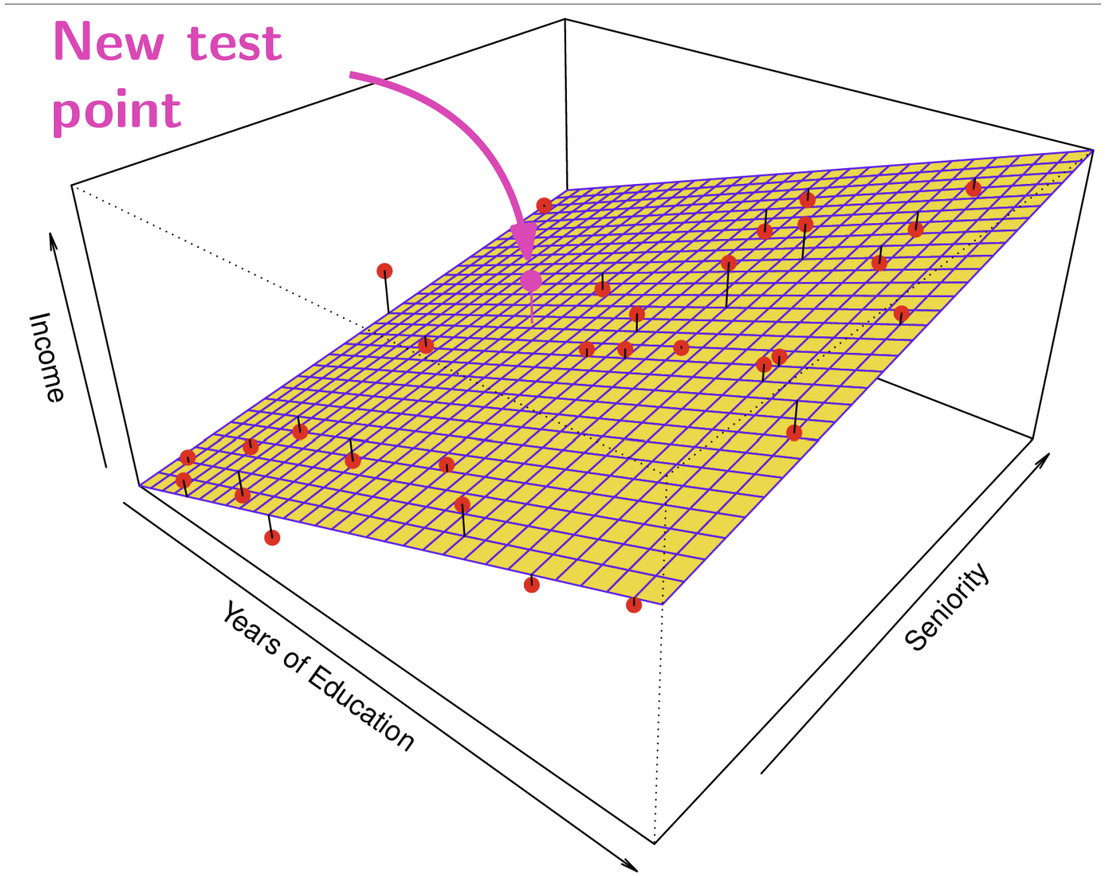
$X = (x_1, \ldots, x_n) \in \mathbb{R}^{p \times n}$: Input

$Y = (y_1, \ldots, y_n) \in \mathbb{R}^n$: Output

"Parametric model": look for $\beta_0, \beta_1, \beta_2$ s.t.:

"Income" $= \beta_0 + \beta_1$ "y. o. education" $+ \beta_2$ "seniority"

A general regression example



$X = (x_1, \ldots, x_n) \in \mathbb{R}^{p \times n}$: Input

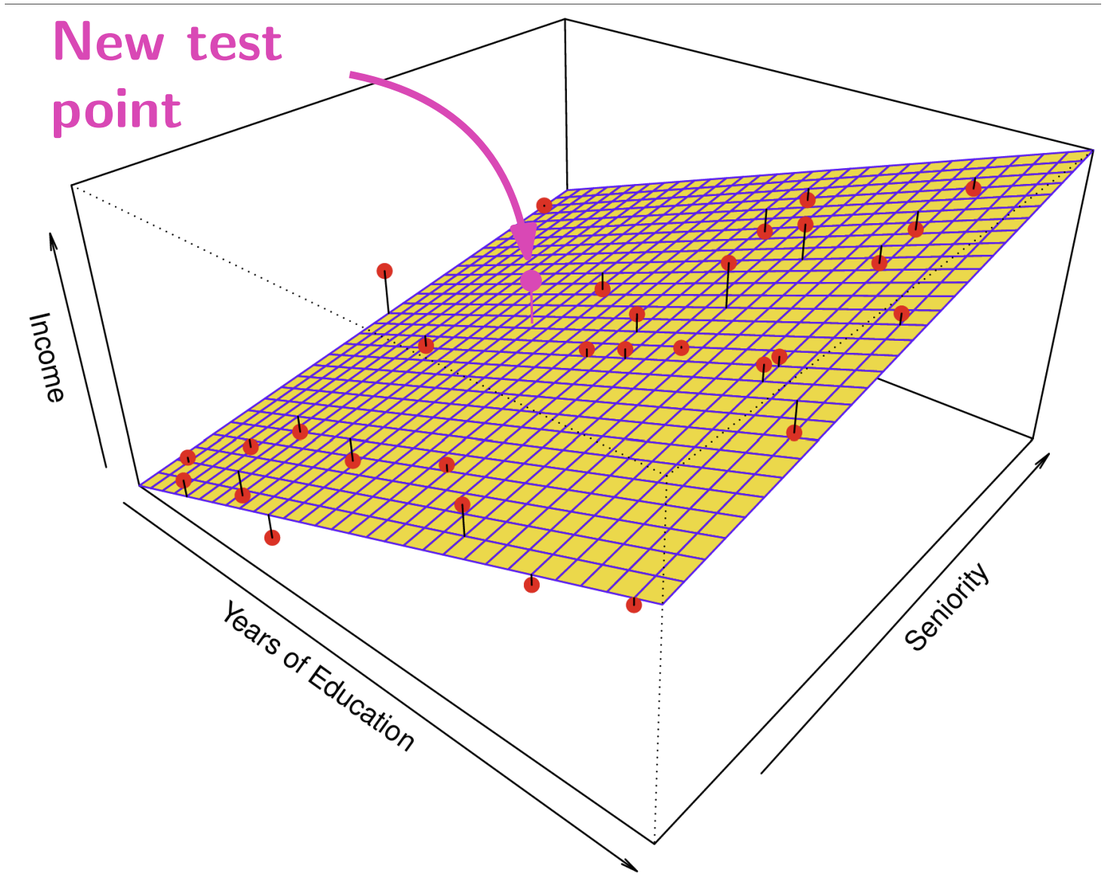
$Y = (y_1, \ldots, y_n) \in \mathbb{R}^n$: Output

β minimizes $\|\beta X - Y\|^2$

"Parametric model": look for $\beta_0, \beta_1, \beta_2$ s.t.:

"Income" $= \beta_0 + \beta_1$ "y. o. education" $+ \beta_2$ "seniority"

A general regression example



"Parametric model": look for $\beta_0, \beta_1, \beta_2$ s.t.:
"Income" $= \beta_0 + \beta_1$ "y. o. education" $+ \beta_2$ "seniority"

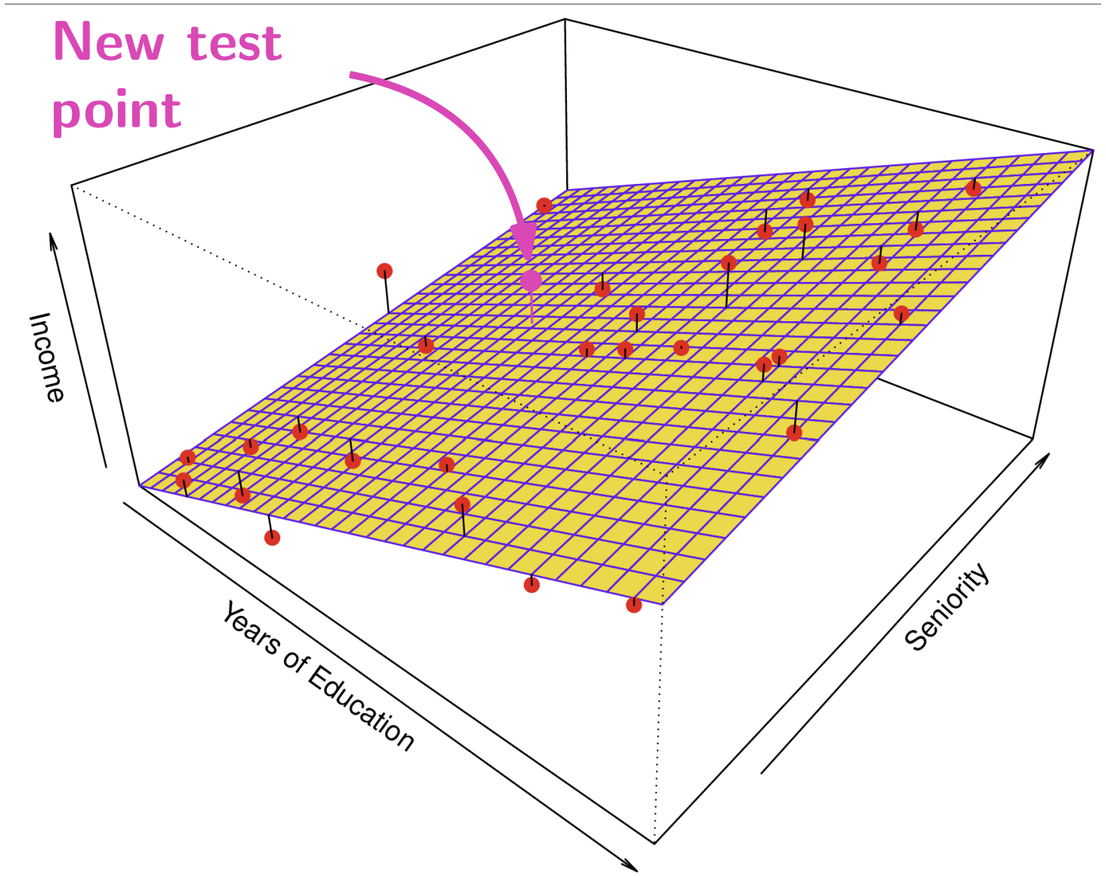
$X = (x_1, \ldots, x_n) \in \mathbb{R}^{p \times n}$: Input

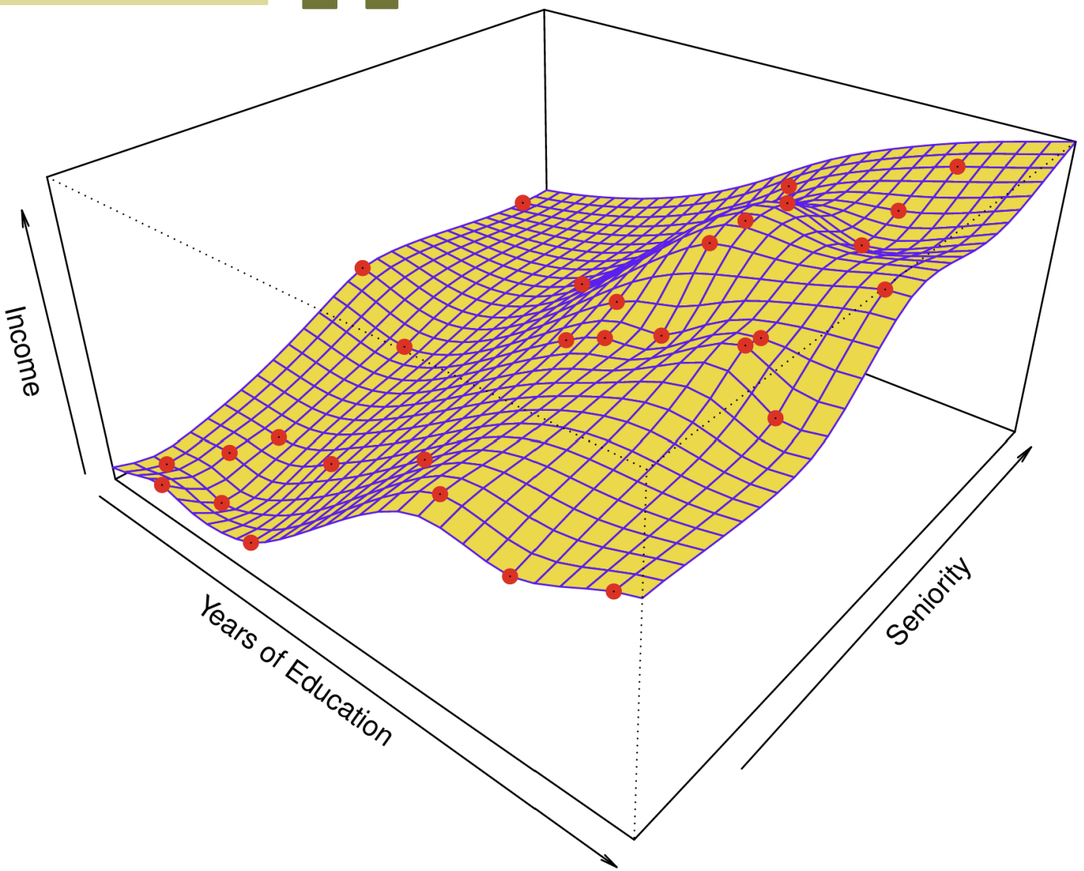
$Y = (y_1, \ldots, y_n) \in \mathbb{R}^n$: Output

β minimizes $\|\beta X - Y\|^2$

Solution: $\beta = (XX^T)^{-1}XY \in \mathbb{R}^p$

A general regression example

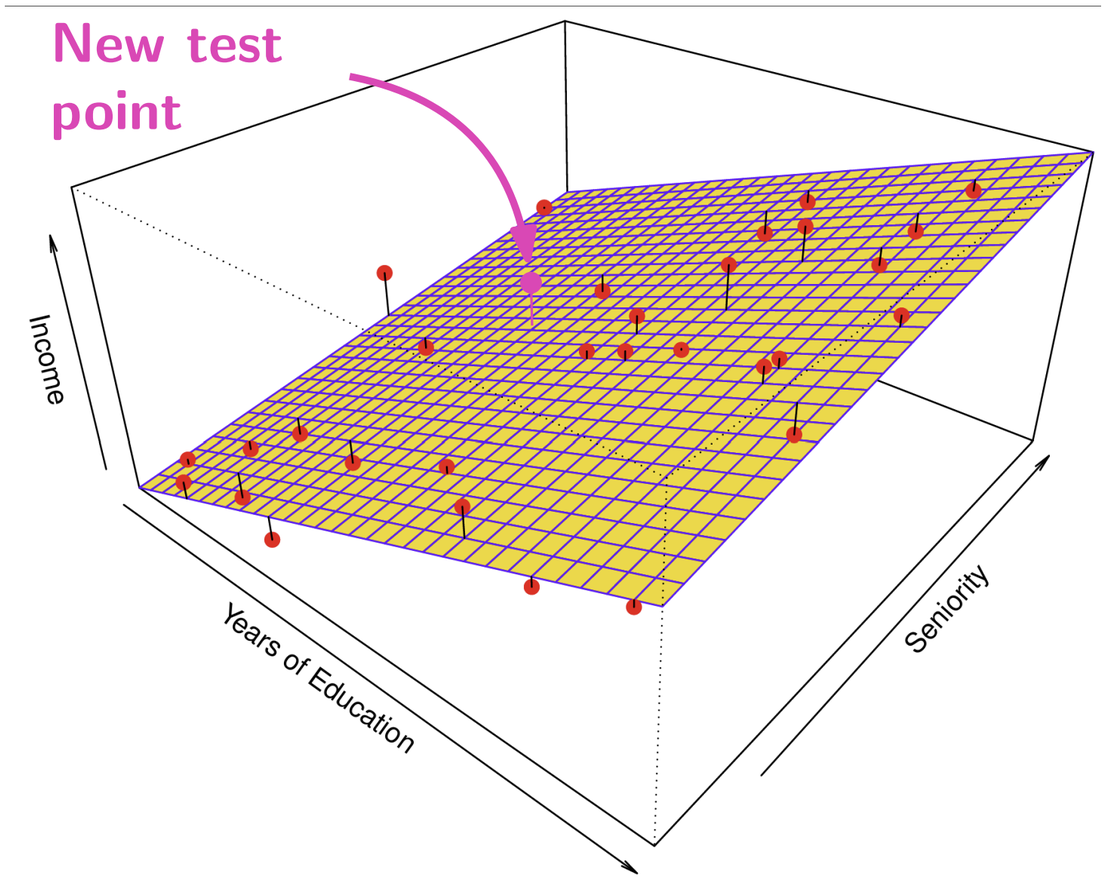


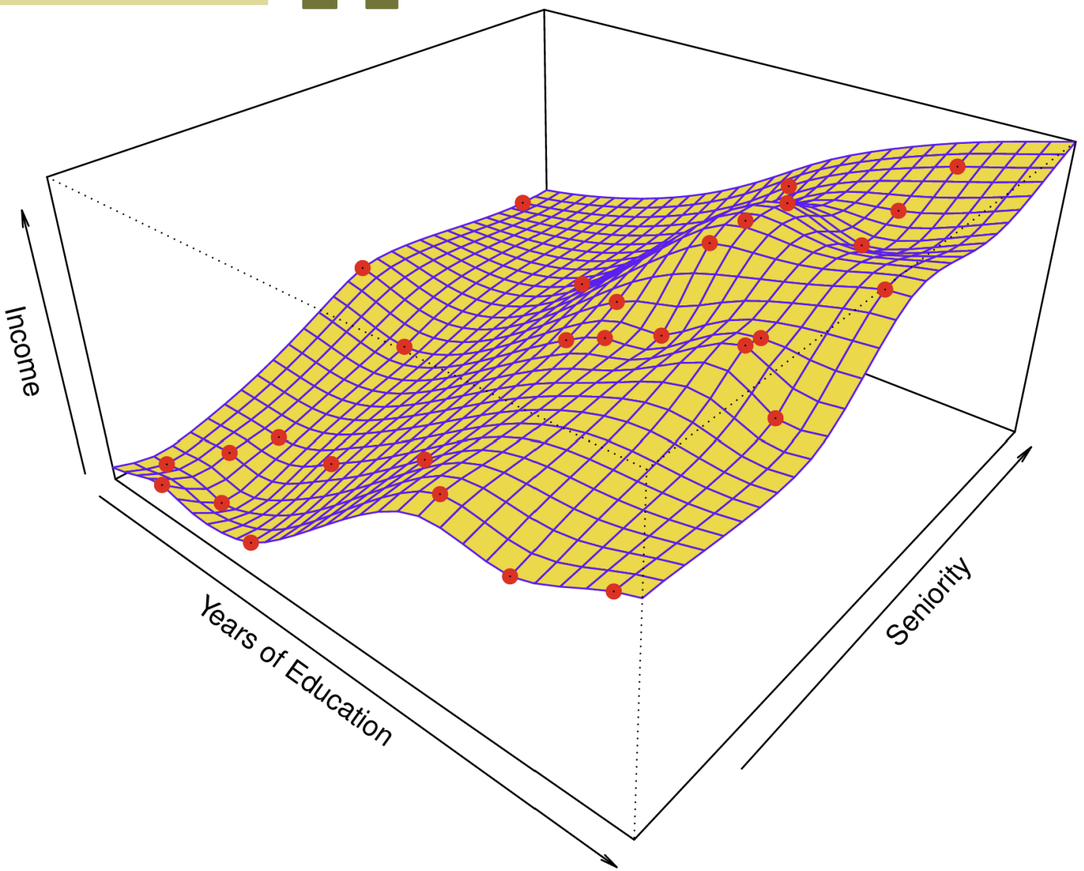


"Parametric model": look for $\beta_0, \beta_1, \beta_2$ s.t.:
"Income" $= \beta_0 + \beta_1$ "y. o. education" $+ \beta_2$ "seniority"

Best polynomial fit to the data $Y = P(X)$

A general regression example





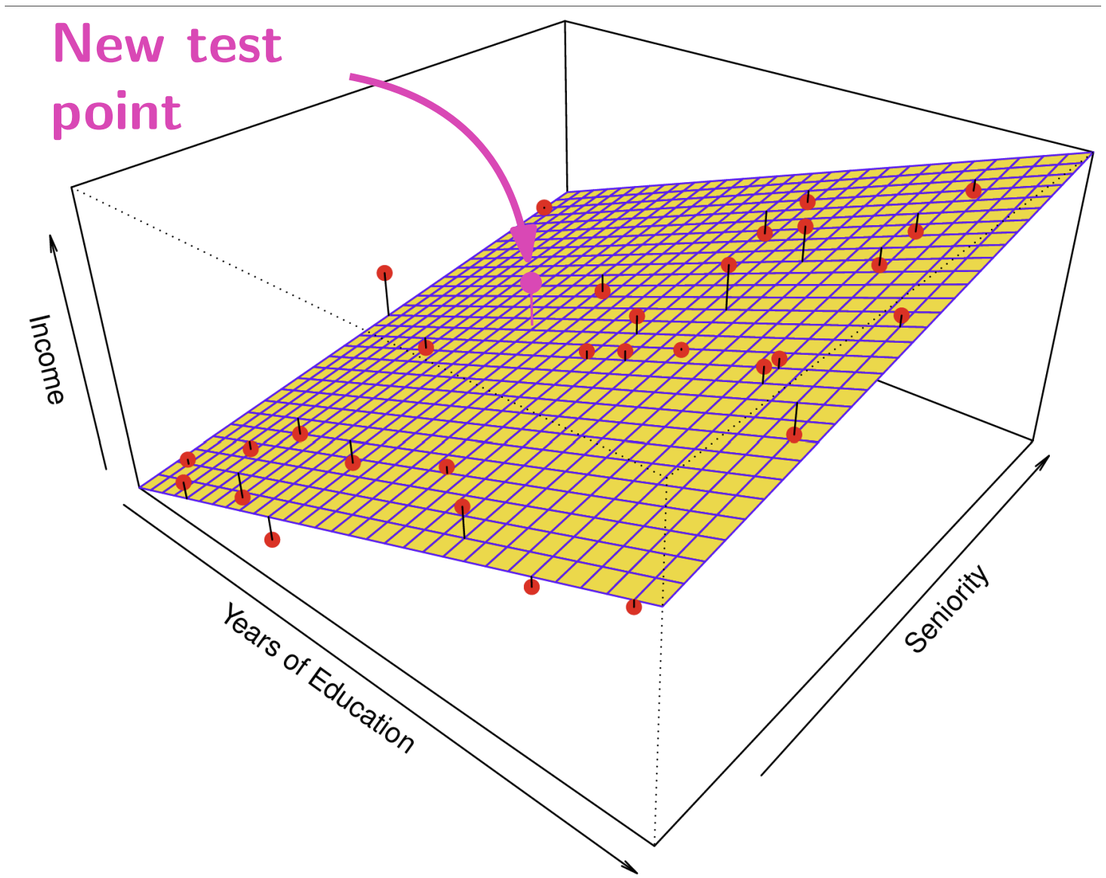
"Parametric model": look for $\beta_0, \beta_1, \beta_2$ s.t.:
"Income" $= \beta_0 + \beta_1$ "y. o. education" $+ \beta_2$ "seniority"

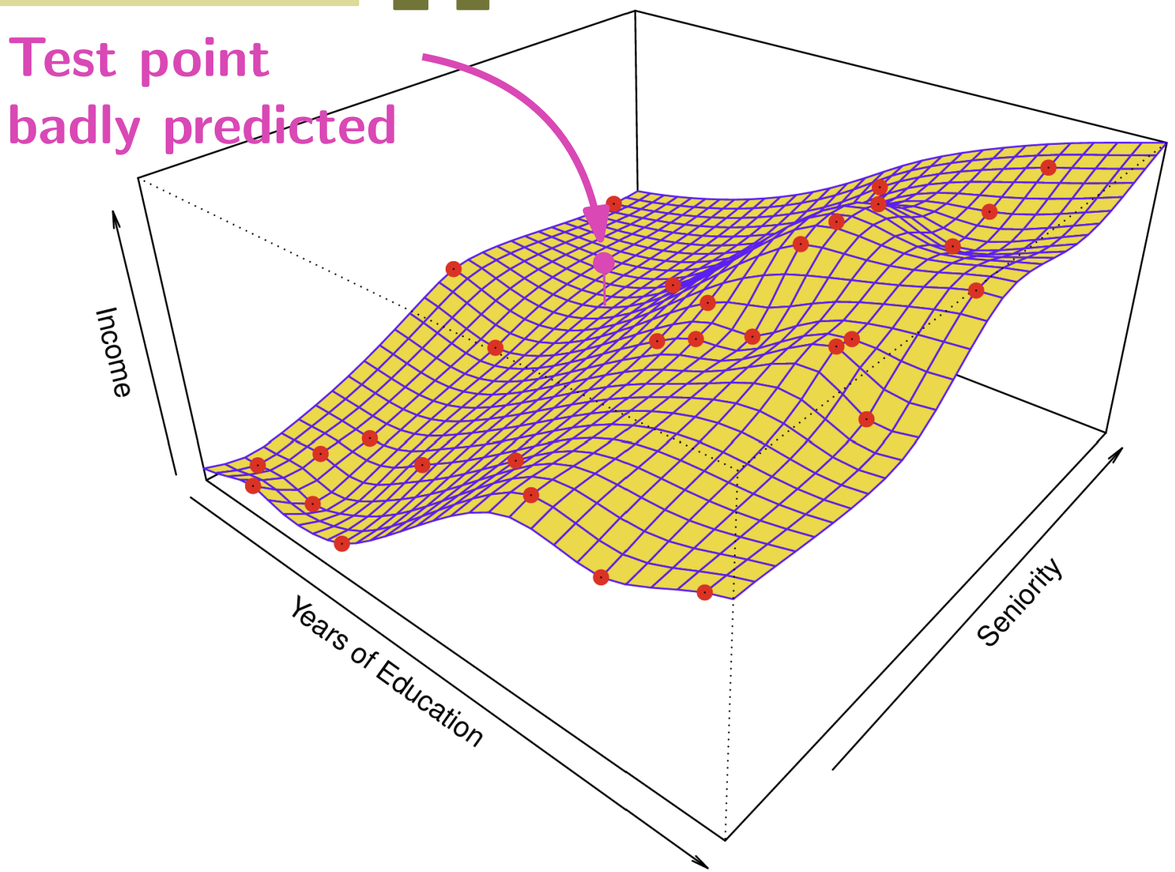
INTERPRETABLE

Best polynomial fit to the data $Y = P(X)$

FLEXIBLE

A general regression example





"Parametric model": look for $\beta_0, \beta_1, \beta_2$ s.t.:
"Income" $= \beta_0 + \beta_1$ "y. o. education" $+ \beta_2$ "seniority"

INTERPRETABLE

Best polynomial fit to the data $Y = P(X)$

FLEXIBLE

A non parametric method instead of regression: KNN

k-**nearest neighbors** method:

Supervised method:
$D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ given

A non parametric method instead of regression: KNN

k-**nearest neighbors** method:

Supervised method:
$D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ given

New data test $x \in \mathbb{R}^p$:

- Look for $i_1, \dots, i_k$ s.t.: $x_{i_1}, \dots, x_{i_k}$ are the k closest neighbors of x:

$$\forall i \in [n] : \quad \|x - x_i\| \geq \sup_{j \in [k]} \|x - x_{i_k}\|.$$

- $\hat{f}_D(x) \equiv \frac{1}{k} \sum_{j=1}^{k} y_{i_j}.$

A non parametric method instead of regression: KNN

k-**nearest neighbors** method:

Supervised method:
$D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ given

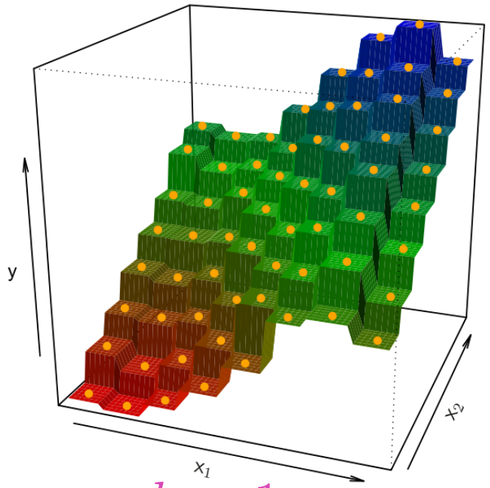
New data test $x \in \mathbb{R}^p$:

- Look for $i_1, \dots, i_k$ s.t.: $x_{i_1}, \dots, x_{i_k}$ are the k closest neighbors of x:

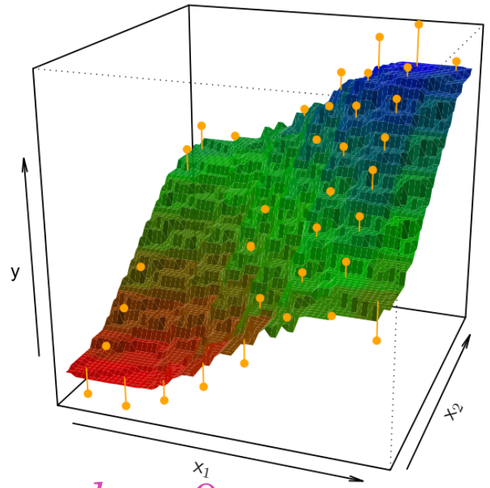
$$\forall i \in [n] : \quad \|x - x_i\| \geq \sup_{j \in [k]} \|x - x_{i_k}\|.$$

- $\hat{f}_D(x) \equiv \frac{1}{k} \sum_{j=1}^{k} y_{i_j}.$

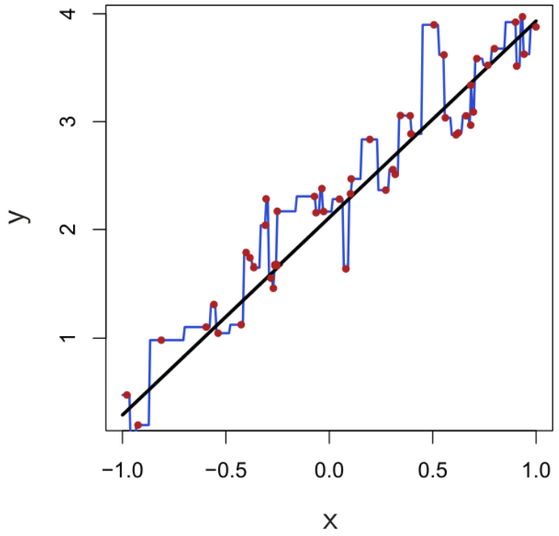
Method highly relies on the number of neighbors chosen

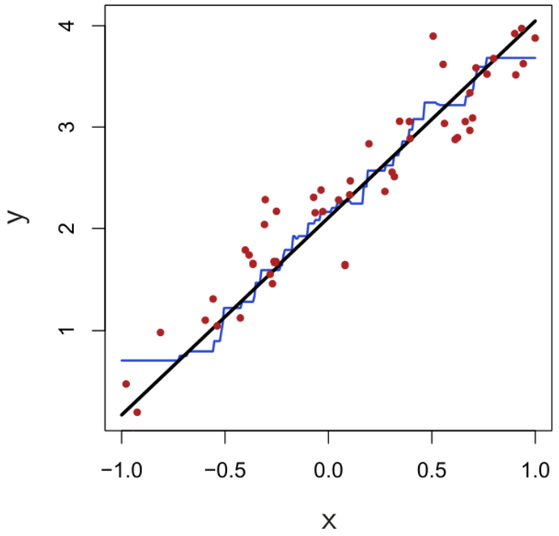


$k = 1$



$k = 9$





A non parametric method instead of regression: KNN

Training dataset: $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ given

Least squared regresion **(LSR)**:

$$\hat{f}_D(x) \equiv \hat{\beta}_0 x^{(1)} + \cdots + \hat{\beta}_p x^{(p)}$$

where $\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \|\beta^T x - y\|^2.$

A non parametric method instead of regression: KNN

Training dataset: $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ given

Least squared regresion **(LSR)**:

$$\hat{f}_D(x) \equiv \hat{\beta}_0 x^{(1)} + \cdots + \hat{\beta}_p x^{(p)}$$

where $\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \|\beta^T x - y\|^2.$

k-nearest neighbors method **(KNN)**:

$x_{i_1}, \dots, x_{i_1}$ k nearest neighbors of x and:

$$\hat{f}_D(x) \equiv \frac{1}{k} \sum_{j=1}^{k} y_{i_j}.$$

A non parametric method instead of regression: KNN

Training dataset: $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ given

Least squared regresion **(LSR)**:

$$\hat{f}_D(x) \equiv \hat{\beta}_0 x^{(1)} + \cdots + \hat{\beta}_p x^{(p)}$$

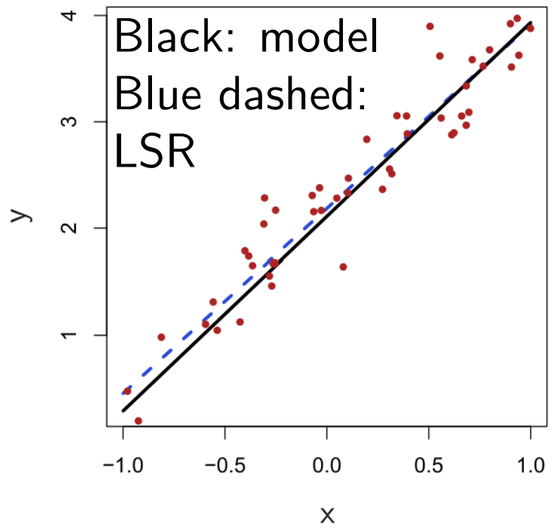
where $\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \|\beta^T x - y\|^2.$

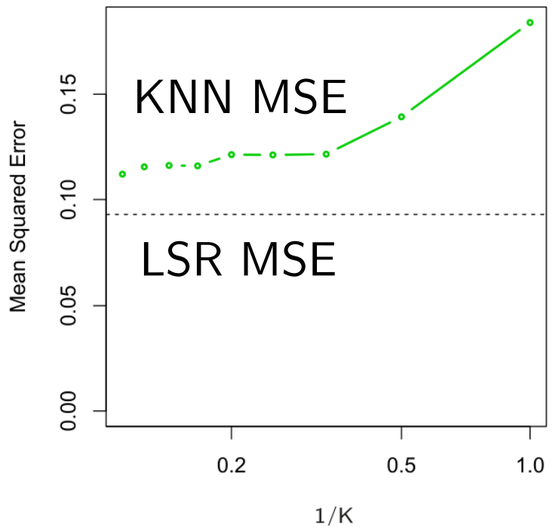
k-nearest neighbors method **(KNN)**:

$x_{i_1}, \ldots, x_{i_1}$ k nearest neighbors of x and:

$$\hat{f}_D(x) \equiv \frac{1}{k} \sum_{j=1}^{k} y_{i_j}.$$

- With linear model, LSR better





A non parametric method instead of regression: KNN

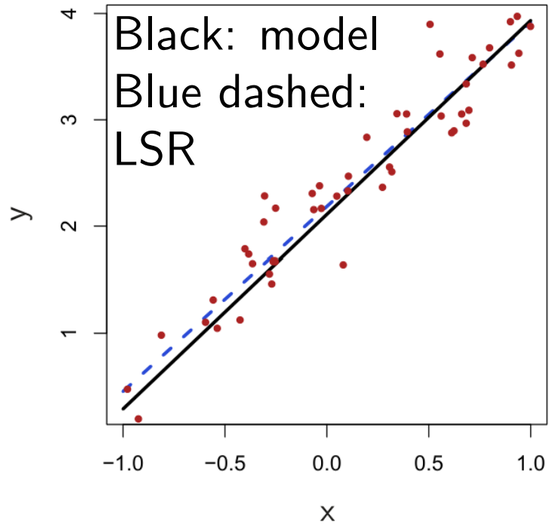
Training dataset: $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ given

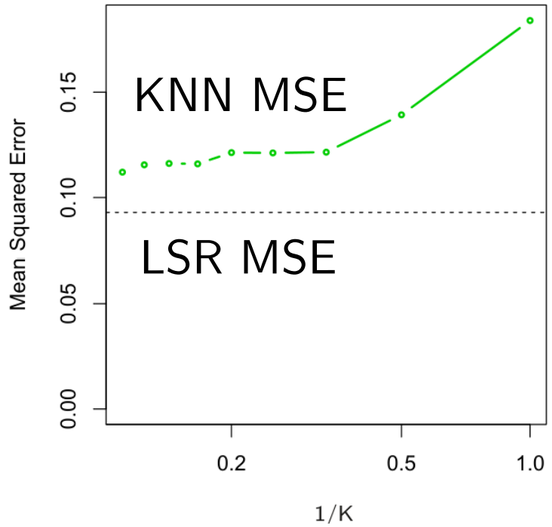
Least squared regresion **(LSR)**:

$$\hat{f}_D(x) \equiv \hat{\beta}_0 x^{(1)} + \cdots + \hat{\beta}_p x^{(p)}$$

where $\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \|\beta^T x - y\|^2.$

- With linear model, LSR better



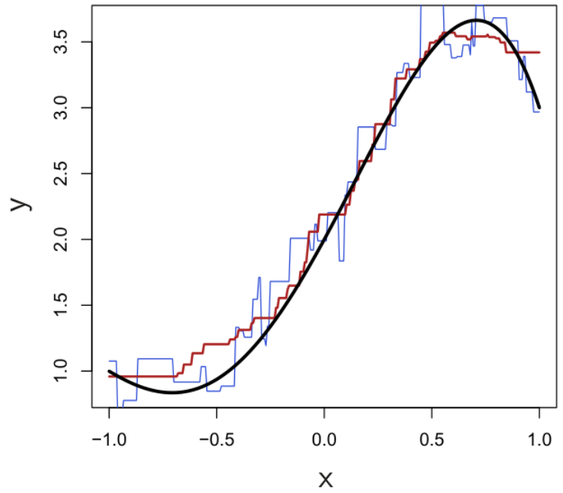


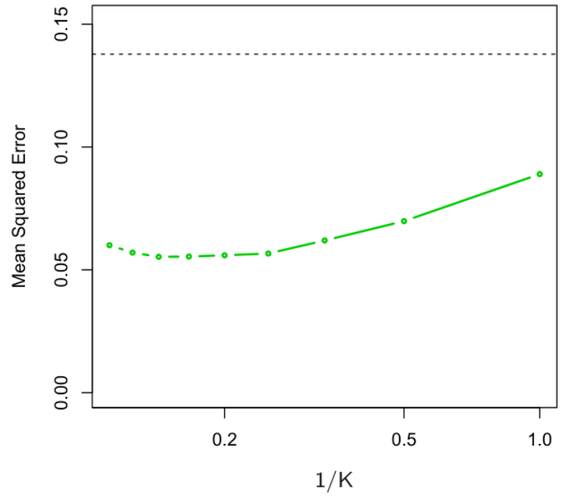
k-nearest neighbors method **(KNN)**:

$x_{i_1}, \ldots, x_{i_1}$ k nearest neighbors of x and:

$$\hat{f}_D(x) \equiv \frac{1}{k} \sum_{j=1}^{k} y_{i_j}.$$

- More complex model of data: KNN better

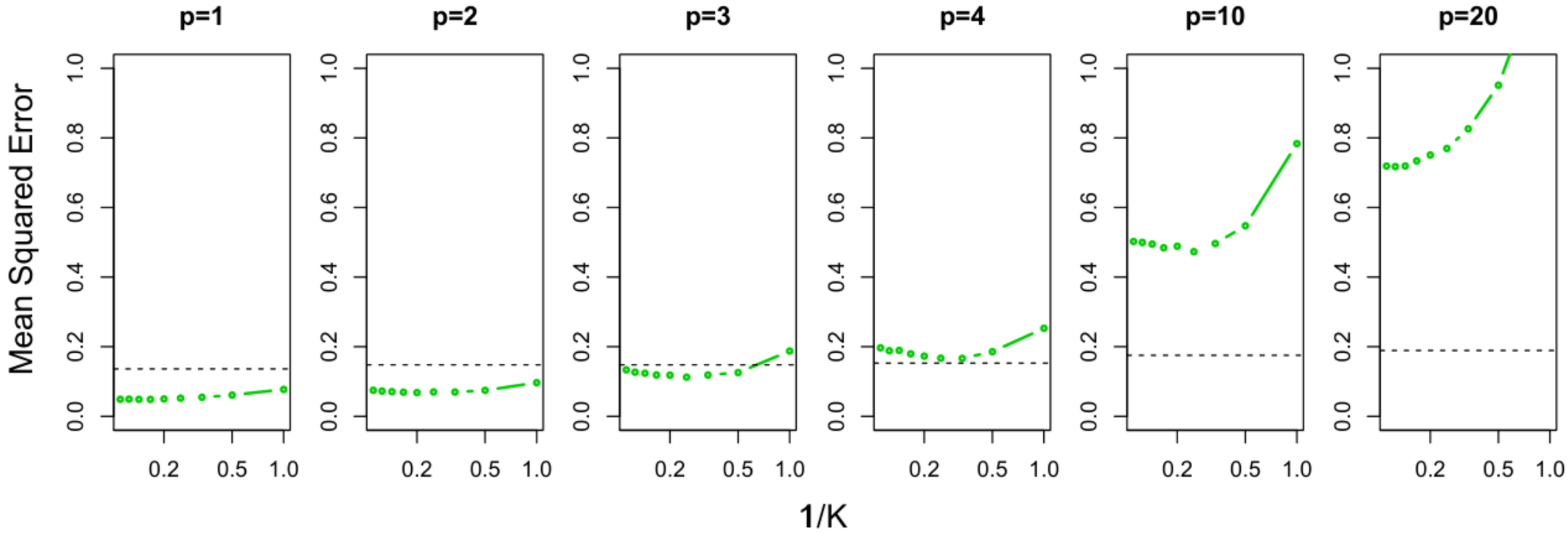




A non parametric method instead of regression: KNN

When dimension grows, with fixed amount of data in training data set (n constant)
KNN highly dependent on the variance of the data: lack of interpretability.

(Here model of data non-linear because otherwise least squares regression would be always performing better)



Maximum likelihood (interpretable parametric method)

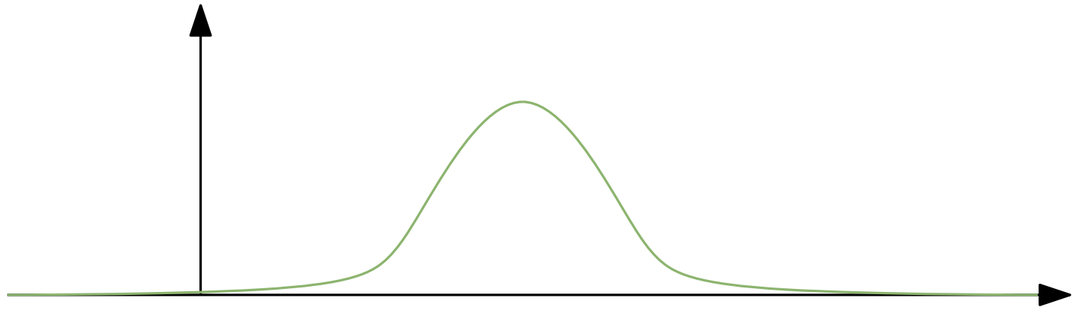
We know the model of X, we want to "fit the model"

Maximum likelihood (interpretable parametric method)

We know the model of X, we want to "fit the model"

→ Find the parameters defining the model

Ex: X follows a Gaussian distribution

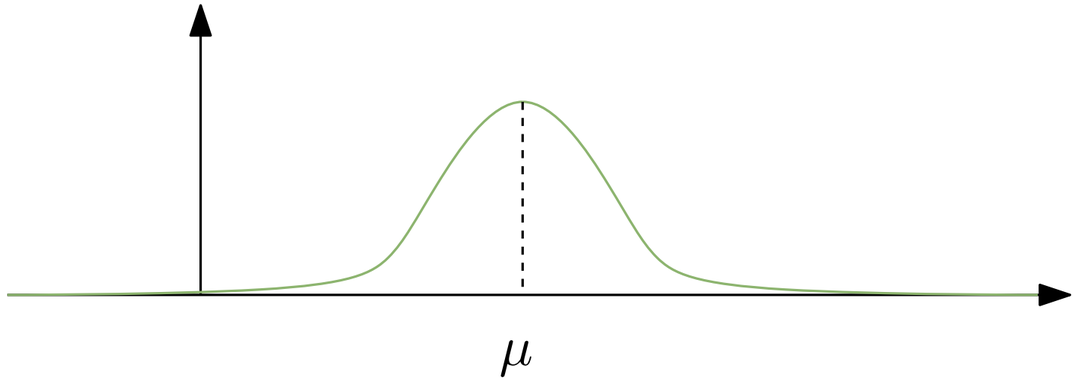


Maximum likelihood (interpretable parametric method)

We know the model of X, we want to "fit the model"

→ Find the parameters defining the model

Ex: X follows a Gaussian distribution

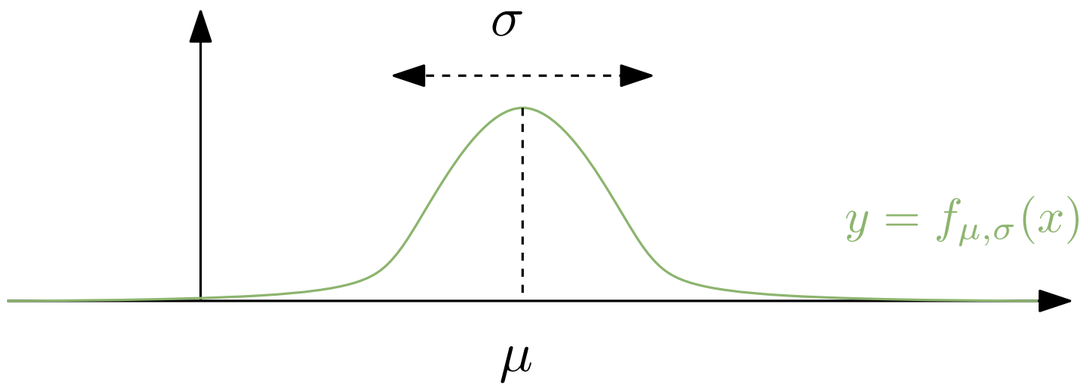


Maximum likelihood (interpretable parametric method)

We know the model of X, we want to "fit the model"

→ Find the parameters defining the model

Ex: X follows a Gaussian distribution

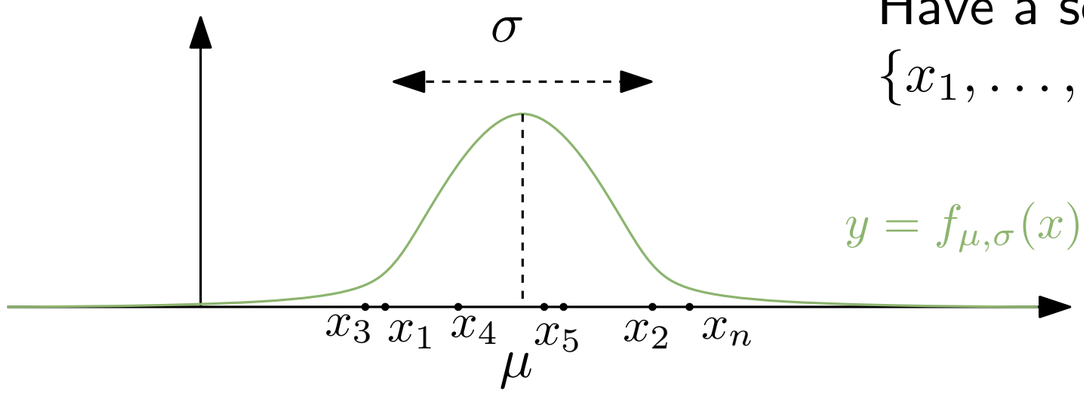


Maximum likelihood (interpretable parametric method)

We know the model of X, we want to "fit the model"

→ Find the parameters defining the model

Ex: X follows a Gaussian distribution



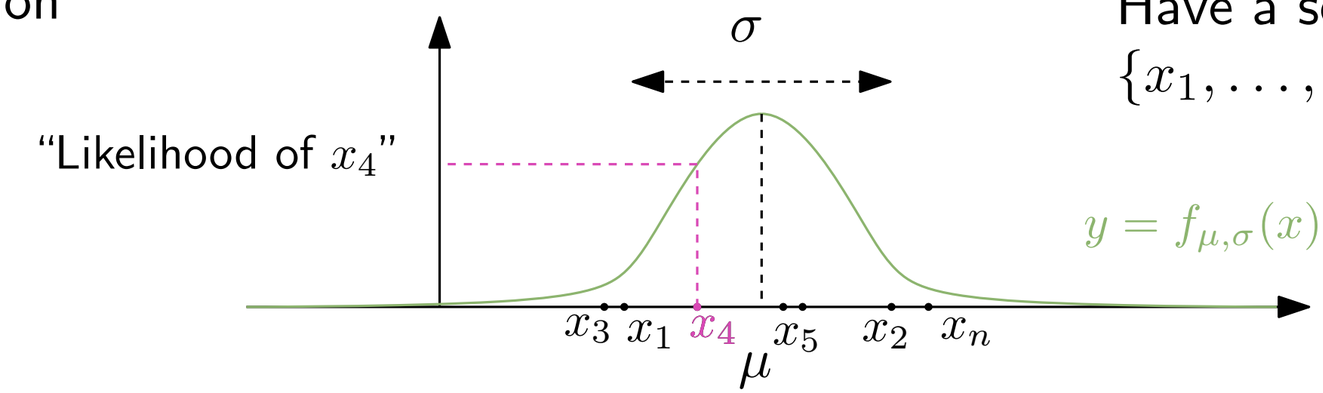
Have a set of drawings $\{x_1, \dots, x_n\}$

Maximum likelihood (interpretable parametric method)

We know the model of X, we want to "fit the model"

→ Find the parameters defining the model

Ex: X follows a Gaussian distribution



Have a set of drawings $\{x_1, \ldots, x_n\}$

Maximum likelihood (interpretable parametric method)

We know the model of X, we want to "fit the model"

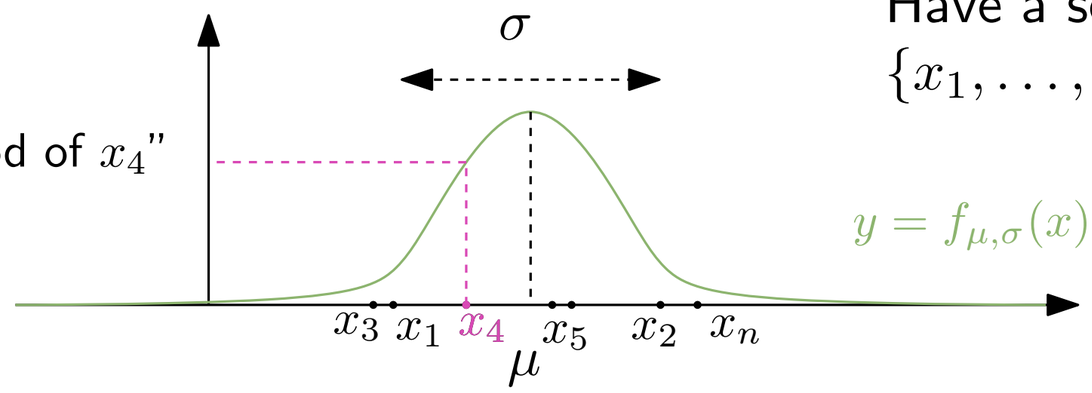
→ Find the parameters defining the model

Ex: X follows a Gaussian distribution

Have a set of drawings $\{x_1, \dots, x_n\}$

"Likelihood of x_4"

→ Depends on x_4



Maximum likelihood (interpretable parametric method)

We know the model of X, we want to "fit the model"

Find the parameters defining the model

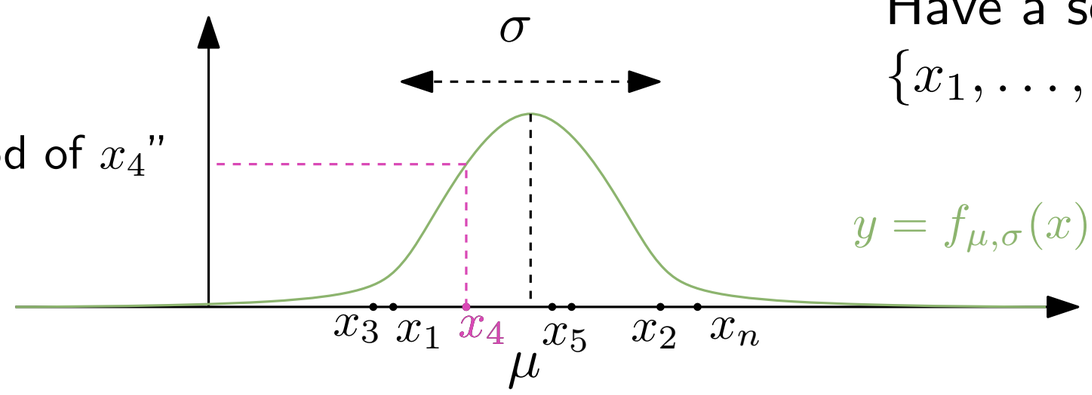
Ex: X follows a Gaussian distribution

Have a set of drawings $\{x_1, \dots, x_n\}$

"Likelihood of x_4"

Depends on x_4

But also on μ and σ!



Maximize: $\prod_{i=1}^{n} f_{\mu,\sigma}(x_i), \quad \mu, \sigma > 0$

Maximum likelihood (interpretable parametric method)

We know the model of X, we want to "fit the model"

Find the parameters defining the model

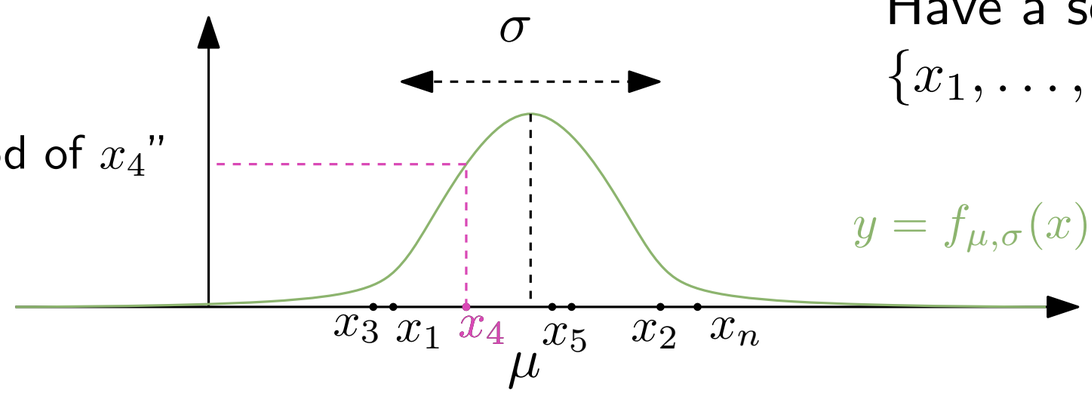
Ex: X follows a Gaussian distribution

Have a set of drawings $\{x_1, \dots, x_n\}$

"Likelihood of x_4"

Depends on x_4

But also on μ and σ!



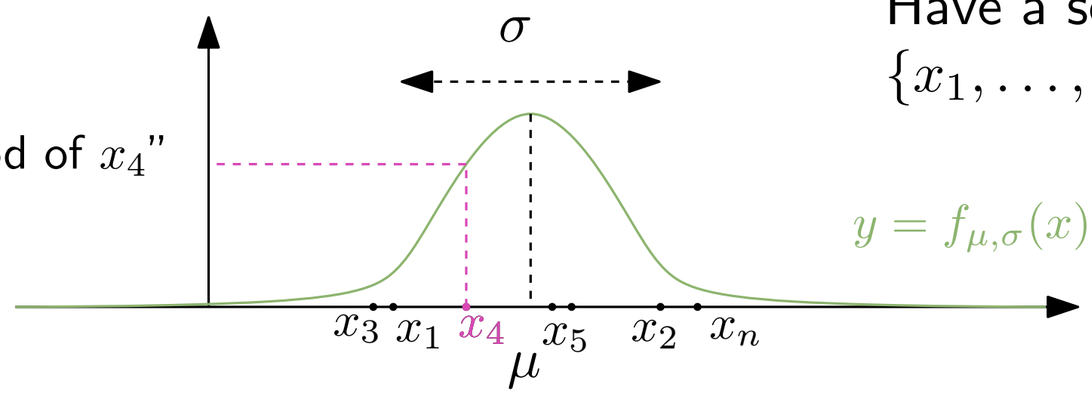
Maximize: $\prod_{i=1}^{n} f_{\mu,\sigma}(x_i), \quad \mu, \sigma > 0 \quad \Longleftrightarrow$ Maximize: $\sum_{i=1}^{n} \log(f_{\mu,\sigma}(x_i)) \quad \mu, \sigma > 0$

Maximum likelihood (interpretable parametric method)

We know the model of X, we want to "fit the model"

→ Find the parameters defining the model

Ex: X follows a Gaussian distribution



Have a set of drawings $\{x_1, \dots, x_n\}$

"Likelihood of x_4"

Depends on x_4

But also on μ and σ!

"loglikelihood minimization"

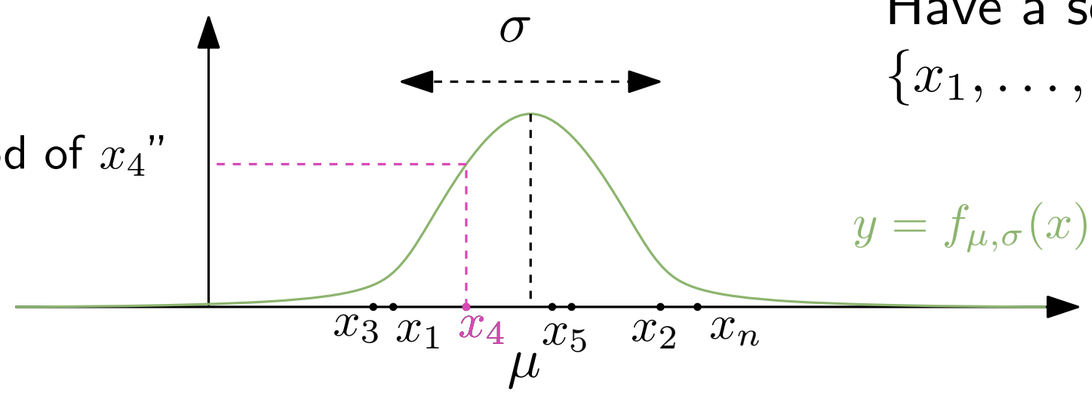
Maximize: $\prod_{i=1}^{n} f_{\mu,\sigma}(x_i), \quad \mu, \sigma > 0 \iff$ Maximize: $\sum_{i=1}^{n} \log(f_{\mu,\sigma}(x_i)) \quad \mu, \sigma > 0$

Maximum likelihood (interpretable parametric method)

We know the model of X, we want to "fit the model"

→ Find the parameters defining the model

Ex: X follows a Gaussian distribution



Have a set of drawings $\{x_1, \dots, x_n\}$

"Likelihood of x_4"

Depends on x_4

But also on μ and σ!

"loglikelihood minimization"

Maximize: $\prod_{i=1}^{n} f_{\mu,\sigma}(x_i), \quad \mu, \sigma > 0 \iff$ Maximize: $\sum_{i=1}^{n} \log(f_{\mu,\sigma}(x_i)) \quad \mu, \sigma > 0$

Gaussian likelihood - Ordinary least square regression (LSR).

Gaussian likelihood - Ordinary least square regression (LSR).

Model: $Y = \beta^T X + \varepsilon$, with:

$Y \in \mathbb{R}$: Target
$X \in \mathbb{R}^p$: data
$\varepsilon \in \mathbb{R}$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$: noise
$\beta \in \mathbb{R}^p$: regression vector

Gaussian likelihood - Ordinary least square regression (LSR).

Model: $Y = \beta^T X + \varepsilon$, with:

$Y \in \mathbb{R}$: Target
$X \in \mathbb{R}^p$: data
$\varepsilon \in \mathbb{R}$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$: noise
$\beta \in \mathbb{R}^p$: regression vector

Given data set $(x_1, y_1), \ldots, (x_n, y_n)$,
$\longrightarrow$ Estimate β

Gaussian likelihood - Ordinary least square regression (LSR).

Model: $Y = \beta^T X + \varepsilon$, with:

$Y \in \mathbb{R}$: Target
$X \in \mathbb{R}^p$: data
$\varepsilon \in \mathbb{R}$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$: noise
$\beta \in \mathbb{R}^p$: regression vector

Given data set $(x_1, y_1), \ldots, (x_n, y_n)$,
$\longrightarrow$ Estimate β

$y_i \sim \mathcal{N}(\beta^T x_i, \sigma^2)$ thus $f_\beta(y_i) = \frac{\exp(\frac{-(y_i - \beta^T x_i)^2}{2\sigma^2}}{\sqrt{2\pi}\sigma}$.

Gaussian likelihood - Ordinary least square regression (LSR).

Model: $Y = \beta^T X + \varepsilon$, with:

$Y \in \mathbb{R}$: Target
$X \in \mathbb{R}^p$: data
$\varepsilon \in \mathbb{R}$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$: noise
$\beta \in \mathbb{R}^p$: regression vector

Given data set $(x_1, y_1), \ldots, (x_n, y_n)$,
$\longrightarrow$ Estimate β

$y_i \sim \mathcal{N}(\beta^T x_i, \sigma^2)$ thus $f_\beta(y_i) = \frac{\exp(\frac{-(y_i - \beta^T x_i)^2}{2\sigma^2}}{\sqrt{2\pi}\sigma}$.

Negative Log-likelihood:
$L(\beta) = \sum_{i=1}^n \frac{(y_i - \beta^T x_i)^2}{2\sigma^2} + \frac{\log(2\pi)}{2} + \log(\sigma)$

Gaussian likelihood - Ordinary least square regression (LSR).

Model: $Y = \beta^T X + \varepsilon$, with:

$Y \in \mathbb{R}$: Target
$X \in \mathbb{R}^p$: data
$\varepsilon \in \mathbb{R}$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$: noise
$\beta \in \mathbb{R}^p$: regression vector

Given data set $(x_1, y_1), \ldots, (x_n, y_n)$,
$\longrightarrow$ Estimate β

$y_i \sim \mathcal{N}(\beta^T x_i, \sigma^2)$ thus $f_\beta(y_i) = \frac{\exp(\frac{-(y_i - \beta^T x_i)^2}{2\sigma^2}}{\sqrt{2\pi}\sigma}$.

Negative Log-likelihood:
$L(\beta) = \sum_{i=1}^n \frac{(y_i - \beta^T x_i)^2}{2\sigma^2} + \frac{\log(2\pi)}{2} + \log(\sigma)$

Minimize $L(\beta)$ $\iff$ Minimize $\|Y - \beta^T X\|^2$

With $Y = (y_1, \ldots, y_n) \in \mathbb{R}^n$,
$X = (x_1, \ldots, x_n) \in \mathbb{R}^{p \times n}$

Gaussian likelihood - Ordinary least square regression (LSR).

Model: $Y = \beta^T X + \varepsilon$, with:

$Y \in \mathbb{R}$: Target
$X \in \mathbb{R}^p$: data
$\varepsilon \in \mathbb{R}$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$: noise
$\beta \in \mathbb{R}^p$: regression vector

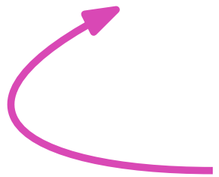
Given data set $(x_1, y_1), \dots, (x_n, y_n)$,
$\longrightarrow$ Estimate β

$y_i \sim \mathcal{N}(\beta^T x_i, \sigma^2)$ thus $f_\beta(y_i) = \frac{\exp(\frac{-(y_i - \beta^T x_i)^2}{2\sigma^2}}{\sqrt{2\pi}\sigma}$.

Negative Log-likelihood:
$L(\beta) = \sum_{i=1}^n \frac{(y_i - \beta^T x_i)^2}{2\sigma^2} + \frac{\log(2\pi)}{2} + \log(\sigma)$

Minimize $L(\beta)$ $\iff$ Minimize $\|Y - \beta^T X\|^2$

With $Y = (y_1, \dots, y_n) \in \mathbb{R}^n$,
$X = (x_1, \dots, x_n) \in \mathbb{R}^{p \times n}$



"Ordinary least square regression"

Poisson likelihood - Poisson Regression.

Poisson distribution is used to model number of apperence of a random event in a given interval.

Poisson likelihood - Poisson Regression.

Poisson distribution is used to model number of apperence of a random event in a given interval.

Ex: number of accident in one year, number of travels in one year, number of calls in one hour, number of day cases during an epidemy...

Poisson likelihood - Poisson Regression.

Poisson distribution is used to model number of apperence of a random event in a given interval.

Ex: number of accident in one year, number of travels in one year, number of calls in one hour, number of day cases during an epidemy...

Law: $f(k) = \frac{\lambda^k}{k!} e^{-\lambda}$

Poisson likelihood - Poisson Regression.

Poisson distribution is used to model number of apperence of a random event in a given interval.

Ex: number of accident in one year, number of travels in one year, number of calls in one hour, number of day cases during an epidemy...

Law: $f(k) = \frac{\lambda^k}{k!} e^{-\lambda}$

Poisson likelihood - Poisson Regression.

Poisson distribution is used to model number of apperence of a random event in a given interval.

Ex: number of accident in one year, number of travels in one year, number of calls in one hour, number of day cases during an epidemy...

Law: $f(k) = \frac{\lambda^k}{k!} e^{-\lambda}$ λ: expected number of appearence.

Poisson likelihood - Poisson Regression.

Poisson distribution is used to model number of apperence of a random event in a given interval.

Ex: number of accident in one year, number of travels in one year, number of calls in one hour, number of day cases during an epidemy...

Law: $f(k) = \frac{\lambda^k}{k!} e^{-\lambda}$ λ: expected number of appearence.

For the Poisson regression the mean is not $\lambda = \beta^T x$ but $\lambda = e^{\beta^T x}$

Poisson likelihood - Poisson Regression.

Poisson distribution is used to model number of apperence of a random event in a given interval.

Ex: number of accident in one year, number of travels in one year, number of calls in one hour, number of day cases during an epidemy...

Law: $f(k) = \frac{\lambda^k}{k!} e^{-\lambda}$ λ: expected number of appearence.

For the Poisson regression the mean is not $\lambda = \beta^T x$ but $\lambda = e^{\beta^T x}$

Given data set $(x_1, y_1), \ldots, (x_n, y_n)$,
$\longrightarrow$ Estimate β

Poisson likelihood - Poisson Regression.

Poisson distribution is used to model number of apperence of a random event in a given interval.

Ex: number of accident in one year, number of travels in one year, number of calls in one hour, number of day cases during an epidemy...

Law: $f(k) = \frac{\lambda^k}{k!} e^{-\lambda}$ λ: expected number of appearence.

For the Poisson regression the mean is not $\lambda = \beta^T x$ but $\lambda = e^{\beta^T x}$

Given data set $(x_1, y_1), \ldots, (x_n, y_n)$,
$\longrightarrow$ Estimate β

Negative log likelihood: $L(\beta) = \sum_{i=1}^{n} e^{\beta^T x_i} - y_i \beta^T x_i + \log(y_i!)$

Poisson likelihood - Poisson Regression.

Poisson distribution is used to model number of apperence of a random event in a given interval.

Ex: number of accident in one year, number of travels in one year, number of calls in one hour, number of day cases during an epidemy...

Law: $f(k) = \frac{\lambda^k}{k!} e^{-\lambda}$ λ: expected number of appearence.

For the Poisson regression the mean is not $\lambda = \beta^T x$ but $\lambda = e^{\beta^T x}$

Given data set $(x_1, y_1), \dots, (x_n, y_n)$,
$\longrightarrow$ Estimate β

Negative log likelihood: $L(\beta) = \sum_{i=1}^n e^{\beta^T x_i} - y_i \beta^T x_i + \log(y_i!)$

No closed-form solution for $\beta = \text{Argmin}_{\beta \in \mathbb{R}^p} L(\beta)$.

Poisson likelihood - Poisson Regression.

Poisson distribution is used to model number of apperence of a random event in a given interval.

Ex: number of accident in one year, number of travels in one year, number of calls in one hour, number of day cases during an epidemy...

Law: $f(k) = \frac{\lambda^k}{k!} e^{-\lambda}$ λ: expected number of appearence.

For the Poisson regression the mean is not $\lambda = \beta^T x$ but $\lambda = e^{\beta^T x}$

Given data set $(x_1, y_1), \ldots, (x_n, y_n)$,
$\longrightarrow$ Estimate β

Negative log likelihood: $L(\beta) = \sum_{i=1}^{n} e^{\beta^T x_i} - y_i \beta^T x_i + \log(y_i!)$

No closed-form solution for $\beta = \text{Argmin}_{\beta \in \mathbb{R}^p} L(\beta)$.

$\rightarrow$ Use an optimizer to compute β.

Interpretability or flexibility?

No good answer!

Most of the techniques using loglikelihood can be classified under the larger class of "Empirical risk minimization" method:

Interpretability or flexibility?

No good answer!

Most of the techniques using loglikelihood can be classified under the larger class of "Empirical risk minimization" method:

$$\text{Minimize: } R(h) = \sum_{i=1}^{n} L(h(x_i), y_i), \quad h : \mathbb{R}^p \to \mathbb{R}^q$$

Interpretability or flexibility?

No good answer!

Most of the techniques using loglikelihood can be classified under the larger class of "Empirical risk minimization" method:

$$\text{Minimize: } R(h) = \sum_{i=1}^{n} L(h(x_i), y_i), \quad h : \mathbb{R}^p \to \mathbb{R}^q$$

h: "hypothesis", in regression tasks, $h_\beta : x \mapsto \beta^T x$

Interpretability or flexibility?

No good answer!

Most of the techniques using loglikelihood can be classified under the larger class of "Empirical risk minimization" method:

$$\text{Minimize: } R(h) = \sum_{i=1}^{n} L(h(x_i), y_i), \quad h : \mathbb{R}^p \to \mathbb{R}^q$$

h: "hypothesis", in regression tasks, $h_\beta : x \mapsto \beta^T x$

Multiple choices of L, $h \to$ more flexible methods.

Interpretability or flexibility?

No good answer!

Most of the techniques using loglikelihood can be classified under the larger class of "Empirical risk minimization" method:

$$\text{Minimize: } R(h) = \sum_{i=1}^{n} L(h(x_i), y_i), \quad h : \mathbb{R}^p \to \mathbb{R}^q$$

h: "hypothesis", in regression tasks, $h_\beta : x \mapsto \beta^T x$

Multiple choices of L, $h \to$ more flexible methods.

When interpretation is available, always prefer interpretable methods.