Perspectives of Predictive Modeling

Arthur Charpentier

charpentier.arthur@uqam.ca

http ://freakonometrics.hypotheses.org/

(SOA Webcast, November 2013)

Agenda

- **Introduction to Predictive Modeling**
  - Prediction, best estimate, expected value and confidence interval
  - Parametric versus nonparametric models
- **Linear Models and (Ordinary) Least Squares**
  - From least squares to the Gaussian model
  - Smoothing continuous covariates
- **From Linear Models to G.L.M.**
- **Modeling a TRUE-FALSE variable**
  - The logistic regression
  - R.O.C. curve
  - Classification tree (and random forests)
- **From individual to functional data**

Prediction ? Best estimate ?

E.g. predicting someone's weight (Y)

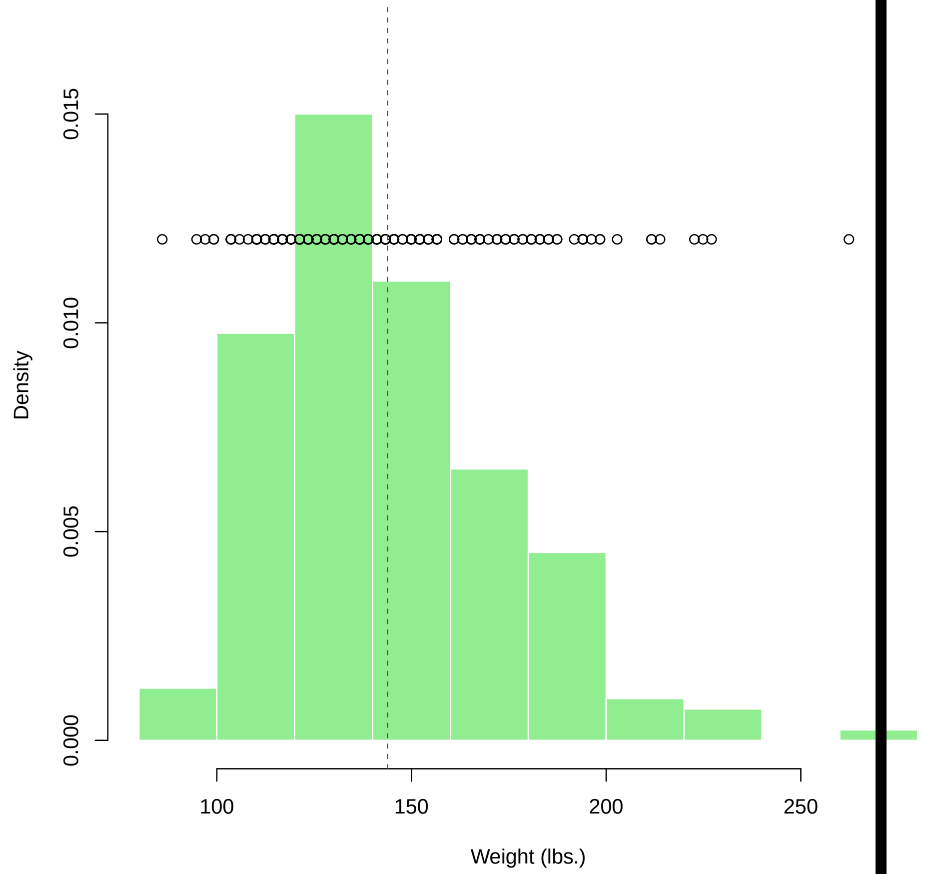
Consider a sample $\{y_1, \cdots, y_n\}$ $\longrightarrow$

Model : $Y_i = \beta_0 + \varepsilon_i$

with $\mathbb{E}(\varepsilon) = 0$ and $\text{Var}(\varepsilon) = \sigma^2$

ε is some unpredictable noise

$\widehat{Y} = \overline{y}$ is our '*best guess*'...



Predicting means estimating $\mathbb{E}(Y)$.

Recall that $\mathbb{E}(Y) = \underset{y\in\mathbb{R}}{\text{argmin}}\{\|Y - y\|_{L_2}\} = \underbrace{\underset{y\in\mathbb{R}}{\text{argmin}}\{\mathbb{E}\big([\overbrace{Y - y}^{\varepsilon}]^2\big)\}}_{\text{least squares}}$

Best estimate with some confidence

E.g. predicting someone's weight (Y)

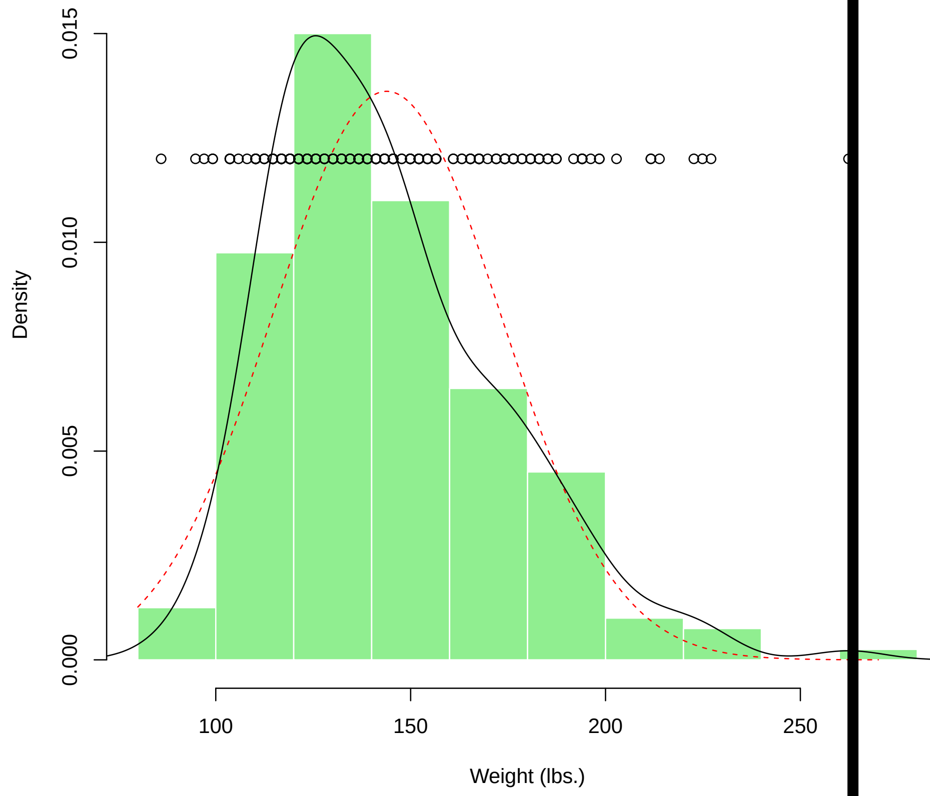
Give an interval $[y_-, y_+]$ such that

$$\mathbb{P}(Y \in [y_-, y_+]) = 1 - \alpha$$

Confidence intervals can be derived if

we can estimate the distribution of Y

$$F(y) = \mathbb{P}(Y \leq y) \text{ or } f(y) = \left.\frac{dF(x)}{dx}\right|_{x=y}$$



(related to the idea of "*quantifying uncertainty*" in our prediction...)

Parametric inference

E.g. predicting someone's weight (Y)

Assume that $F \in \mathcal{F} = \{F_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$

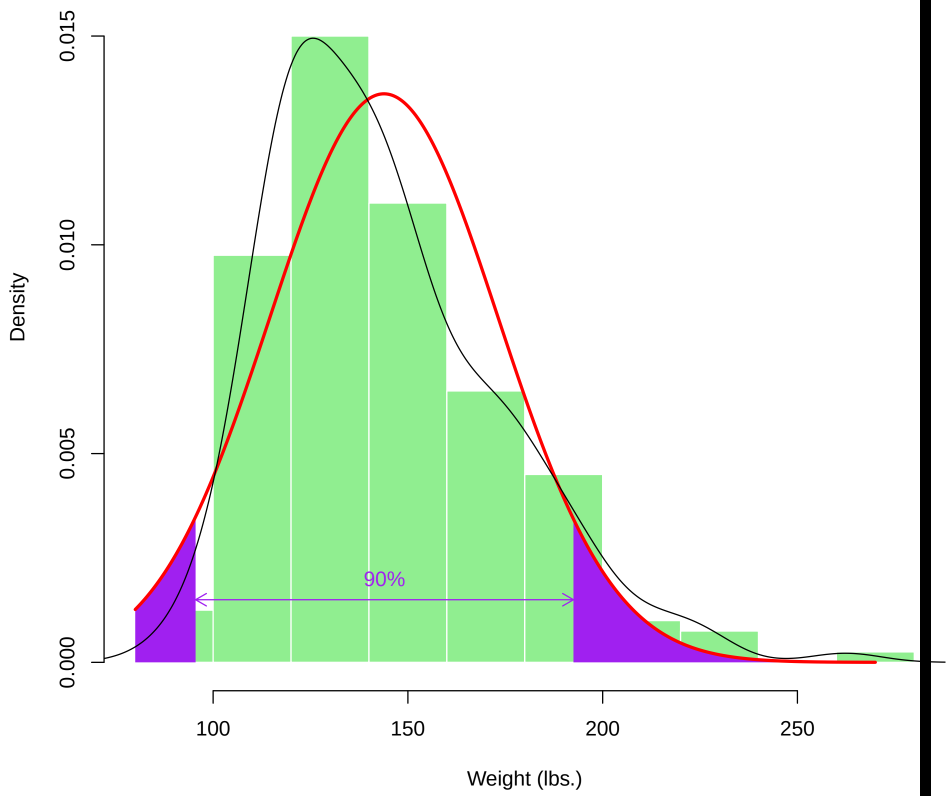
1. Provide an estimate $\widehat{\boldsymbol{\theta}}$

2. Compute bound estimates

$$\widehat{y}_- = F^{-1}_{\widehat{\boldsymbol{\theta}}}(\alpha/2)$$
$$\widehat{y}_+ = F^{-1}_{\widehat{\boldsymbol{\theta}}}(1-\alpha/2)$$

Standard estimation technique :

$\longrightarrow$ maximum likelihood techniques



$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{argmax}}\{\underbrace{\sum_{i=1}^n \log f_{\boldsymbol{\theta}}(y_i)}_{\text{log likelihood}}\}\quad \begin{cases} \text{explicit (analytical) expression for } \widehat{\boldsymbol{\theta}} \\ \text{numerical optimization (Newton Raphson)} \end{cases}$$

Non-parametric inference

E.g. predicting someone's weight (Y)

1. Empirical distribution function

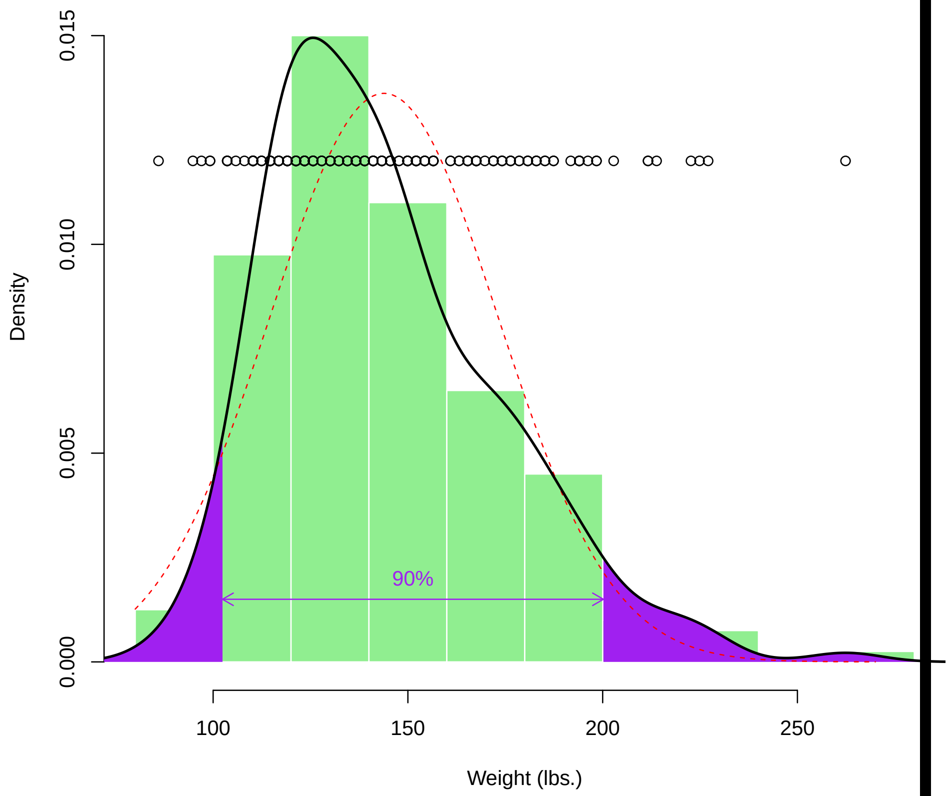
$$\widehat{F}(y) = \frac{1}{n} \underbrace{\sum_{i=1}^{n} \mathbf{1}(y_i \leq y)}_{\#\{i \text{ such that } y_i \leq y\}}$$

natural estimator for $\mathbb{P}(Y \leq y)$

2. Compute bound estimates

$$\widehat{y}_- = \widehat{F}^{-1}(\alpha/2)$$

$$\widehat{y}_+ = \widehat{F}^{-1}(1-\alpha/2)$$



Prediction using some covariates

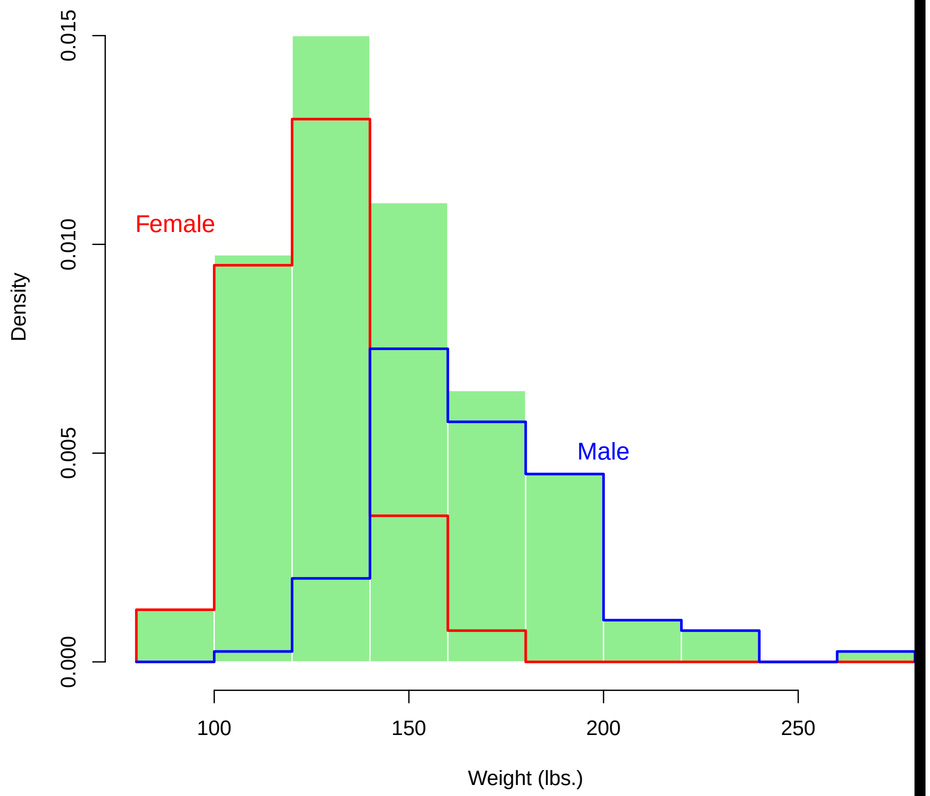
E.g. predicting someone's weight (Y)

based on his/her sex (X_1)

$$\text{Model} : Y_i = \begin{cases} \beta_F + \varepsilon_i \text{ if } X_{1,i} = \text{F} \\ \beta_H + \varepsilon_i \text{ if } X_{1,i} = \text{M} \end{cases}$$

$$\text{or } Y_i = \underbrace{\beta_0}_{\beta_M} + \underbrace{\beta_1}_{\beta_F - \beta_M} \mathbf{1}(X_{1,i} = \text{F}) + \varepsilon_i$$

with $\mathbb{E}(\varepsilon) = 0$ and $\text{Var}(\varepsilon) = \sigma^2$



Prediction using some (categorical) covariates

E.g. predicting someone's weight (Y)

based on his/her sex (X_1)

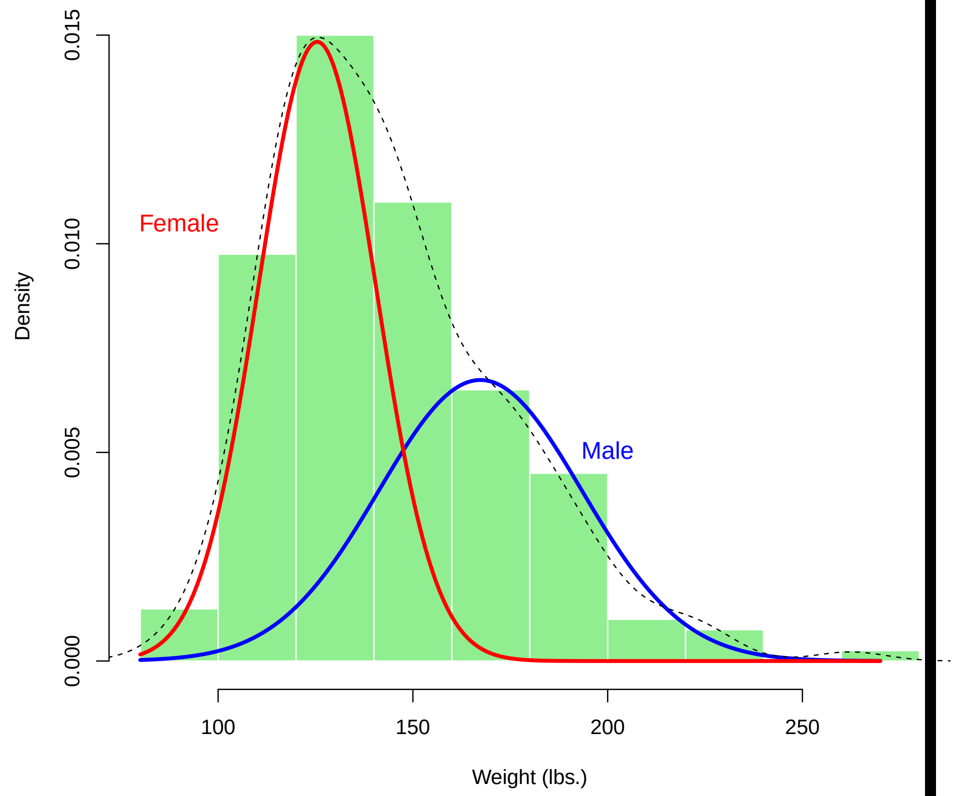
Conditional parametric model

assume that $Y|X_1 = x_1 \sim F_{\boldsymbol{\theta}(x_1)}$

i.e. $Y_i \sim \begin{cases} F_{\boldsymbol{\theta}_F} \text{ if } X_{1,i} = \text{F} \\ F_{\boldsymbol{\theta}_M} \text{ if } X_{1,i} = \text{M} \end{cases}$

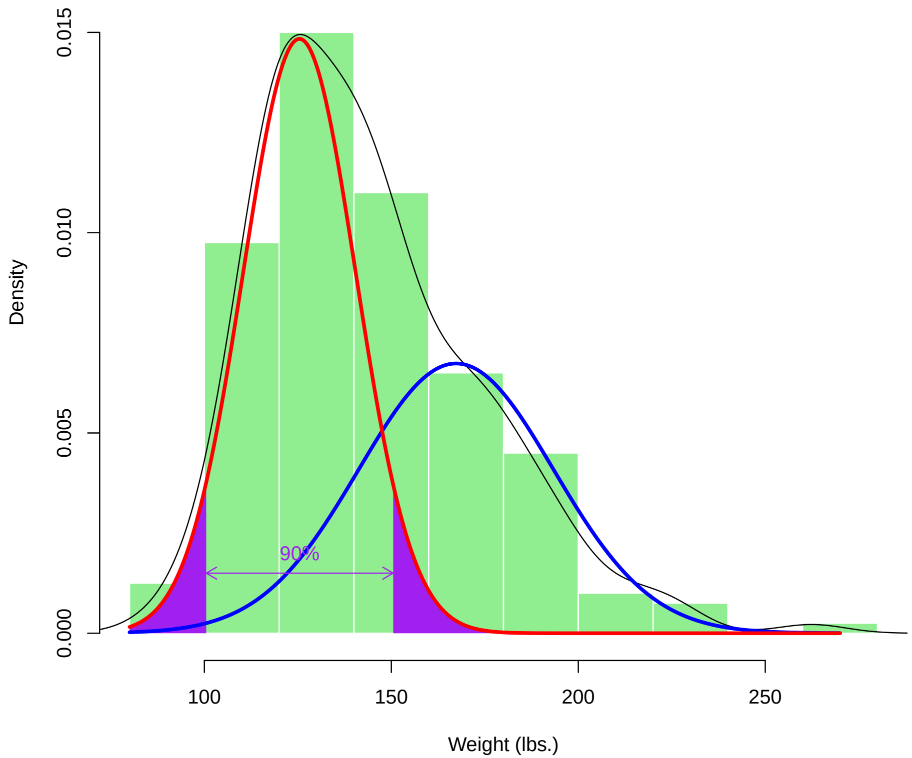
$\longrightarrow$ our prediction will be

conditional on the covariate

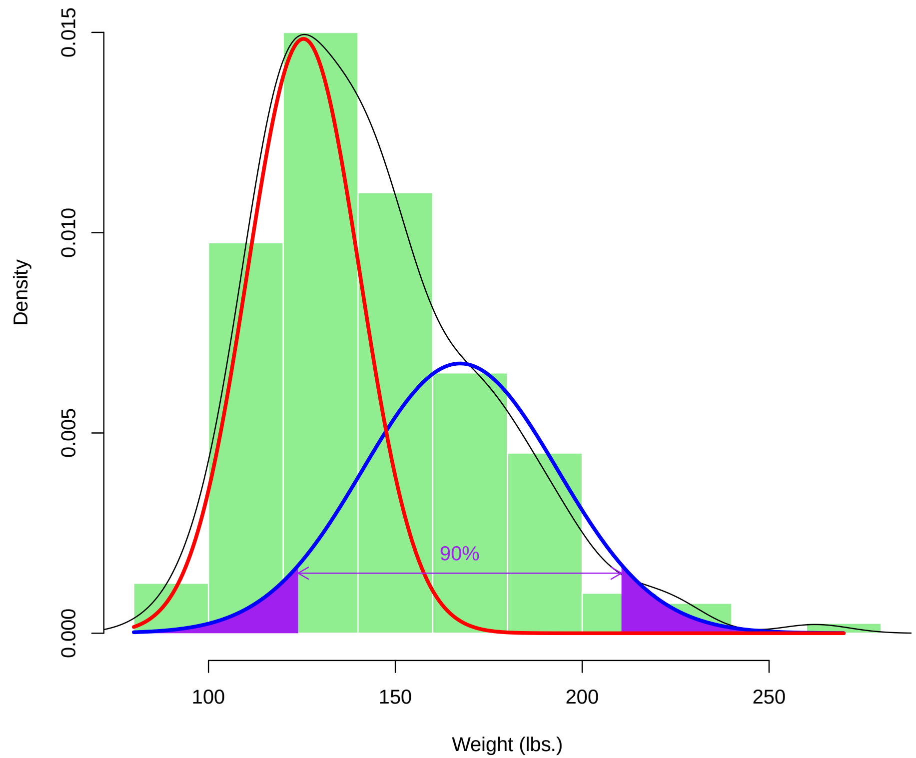


Prediction using some (categorical) covariates

Prediction of Y when $X_1 = \mathrm{F}$



Prediction of Y when $X_1 = \mathrm{M}$



Linear Models, and Ordinary Least Squares

E.g. predicting someone's weight (Y)

based on his/her height (X_2)

Linear Model : $Y_i = \beta_0 + \beta_2\ X_{2,i} + \varepsilon_i$

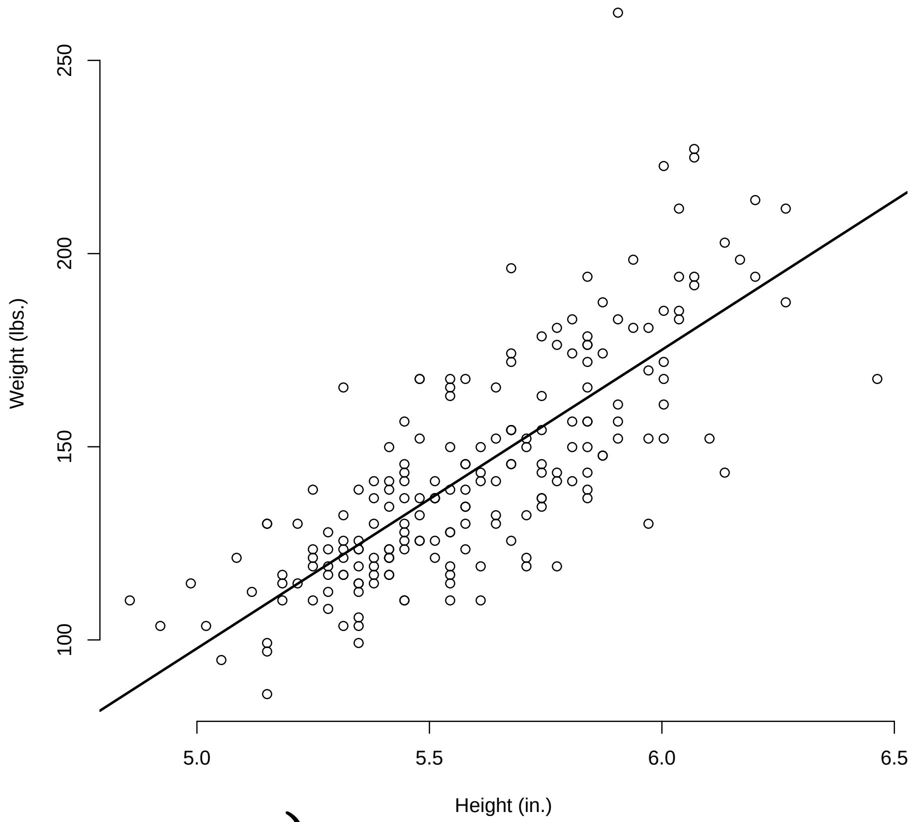
with $\mathbb{E}(\varepsilon) = 0$ and $\text{Var}(\varepsilon) = \sigma^2$

Conditional parametric model

assume that $Y|X_2 = x_2 \sim F_{\boldsymbol{\theta}(x_2)}$

E.g. Gaussian Linear Model

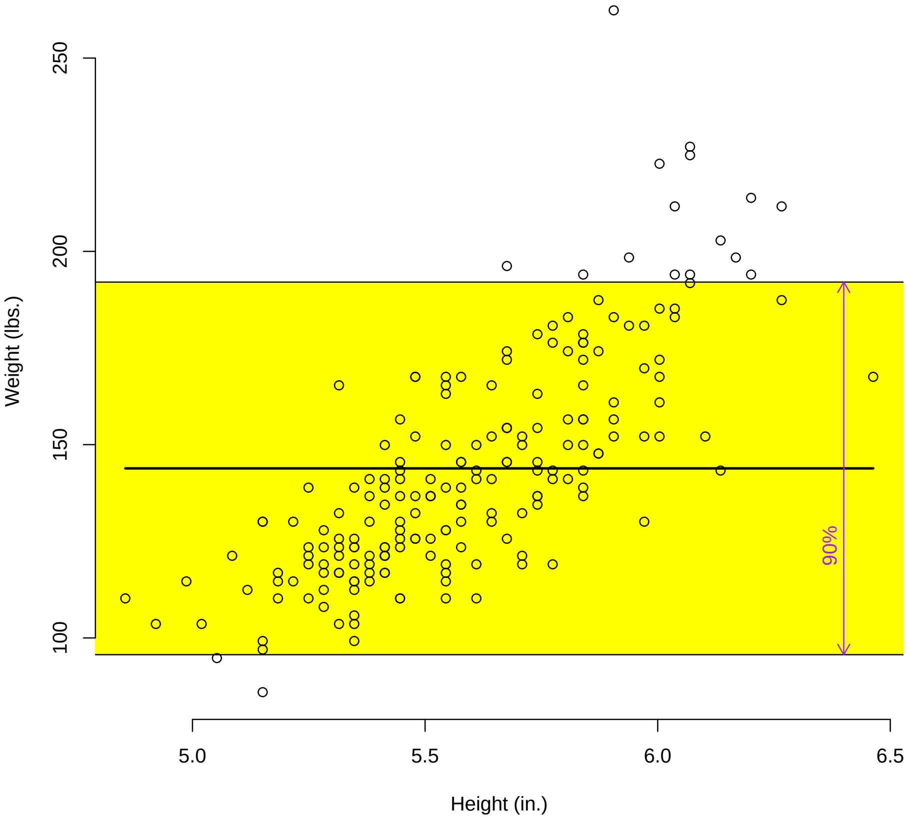
$$Y|X_2 = x_2 \sim \mathcal{N}(\underbrace{\mu(x_2)}_{\beta_0+\beta_2 x_2}, \underbrace{\sigma^2(x_2)}_{\sigma^2})$$

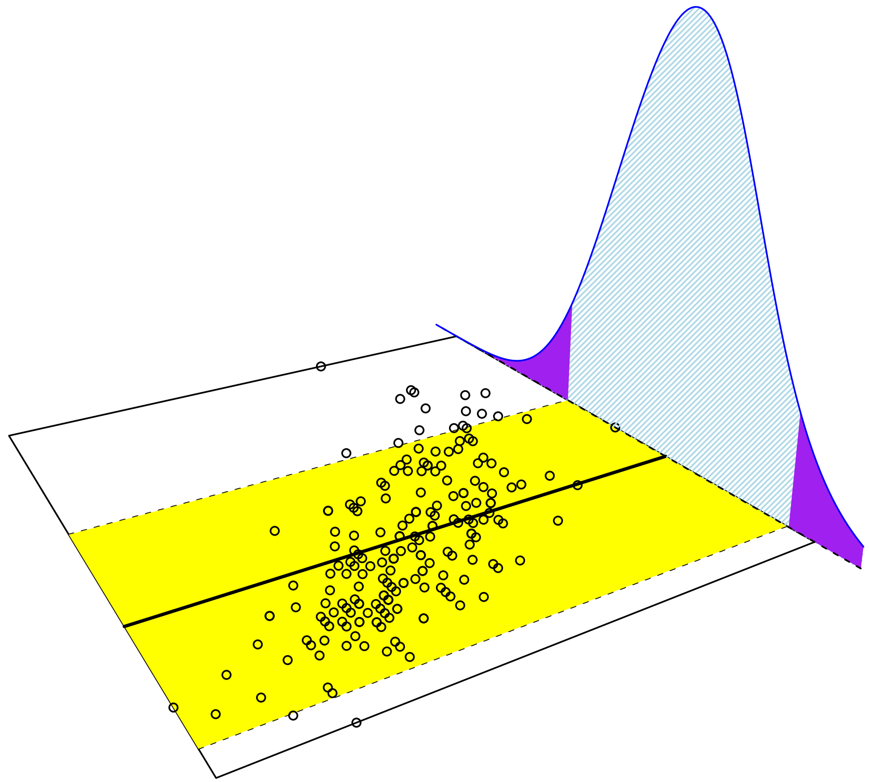


$$\longrightarrow \text{ordinary least squares, } \widehat{\boldsymbol{\beta}} = \text{argmin}\left\{\sum_{i=1}^{n}[Y_i - \boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\beta}]^2\right\}$$

$\widehat{\boldsymbol{\beta}}$ is also the M.L. estimator of $\boldsymbol{\beta}$

Prediction using no covariates





Prediction using a categorical covariates

E.g. predicting someone's weight (Y)

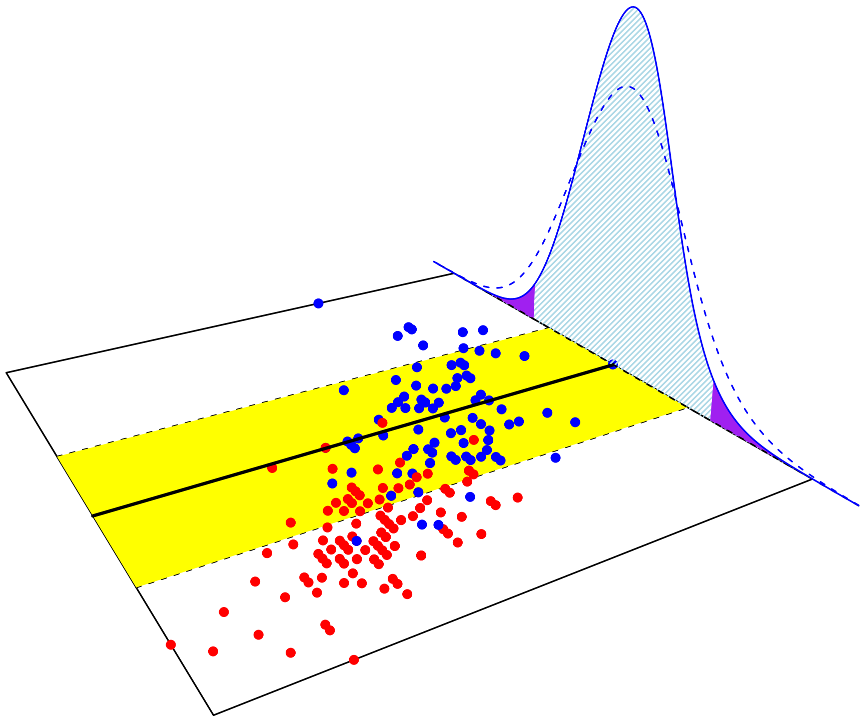
based on his/her sex (X_1)

E.g. Gaussian linear model

$$Y|X_1 = \text{M} \sim \mathcal{N}(\mu_M, \sigma^2)$$

$$\widehat{\mathbb{E}}(Y|X_1 = \text{M}) = \frac{1}{n_M} \sum_{i:X_{1,i}=\text{M}} Y_i = \widehat{Y}(\text{M})$$

$$Y \in \left[\widehat{Y}(\text{M}) \pm \underbrace{u_{1-\alpha/2}}_{1.96} \cdot \widehat{\sigma}\right]$$



Remark In the linear model, $\text{Var}(\varepsilon) = \sigma^2$ does not depend on X_1.

Prediction using a categorical covariates

E.g. predicting someone's weight (Y)

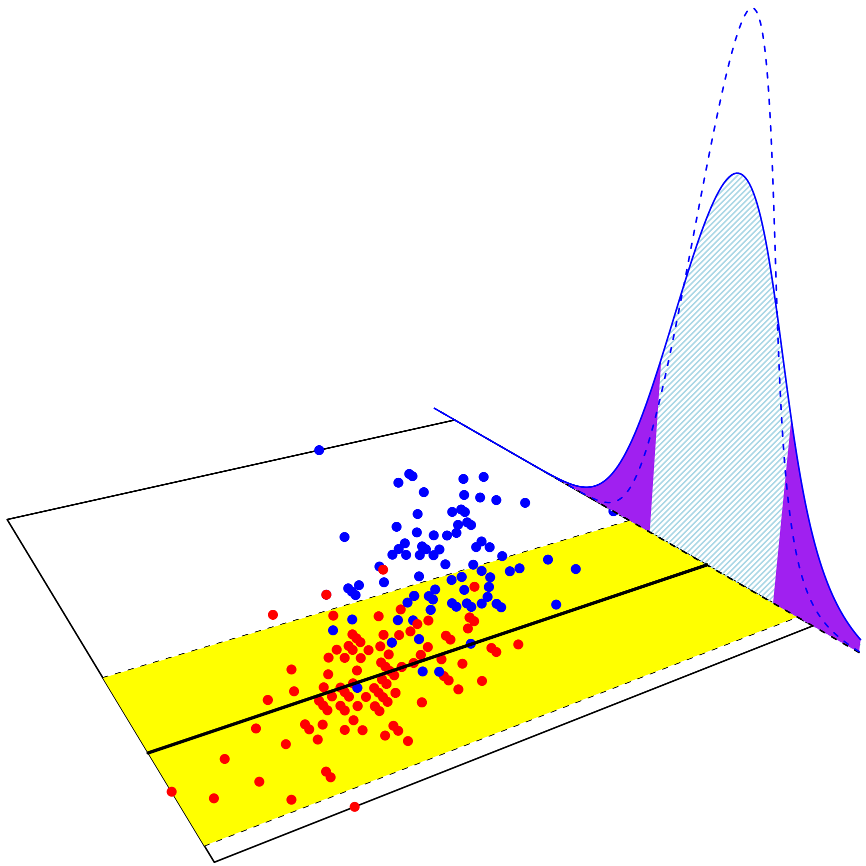
based on his/her sex (X_1)

E.g. Gaussian linear model

$$Y|X_1=\text{F}\sim\mathcal{N}(\mu_F,\sigma^2)$$

$$\widehat{\mathbb{E}}(Y|X_1=\text{F})=\frac{1}{n_F}\sum_{i:X_{1,i}=\text{F}}Y_i=\widehat{Y}(\text{F})$$

$$Y\in\left[\widehat{Y}(\text{F})\pm\underbrace{u_{1-\alpha/2}}_{1.96}\cdot\widehat{\sigma}\right]$$



Remark In the linear model, $\text{Var}(\varepsilon)=\sigma^2$ does not depend on X_1.

Prediction using a continuous covariates

E.g. predicting someone's weight (Y)

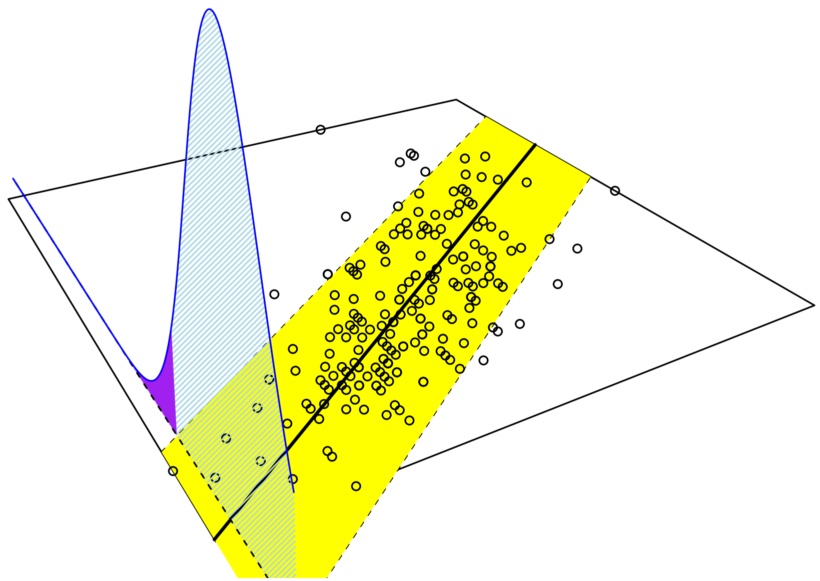
based on his/her height (X_2)

E.g. Gaussian linear model

$$Y|X_2 = x_2 \sim \mathcal{N}(\beta_0 + \beta_1 x_2, \sigma^2)$$

$$\widehat{\mathbb{E}}(Y|X_2 = x_2) = \widehat{\beta}_0 + \widehat{\beta}_1 x_2 = \widehat{Y}(x_2)$$

$$Y \in \left[\widehat{Y}(x_2) \pm \underbrace{u_{1-\alpha/2}}_{1.96} \cdot \widehat{\sigma}\right]$$



Improving our prediction ?

(Empirical) residuals, $\widehat{\varepsilon}_i = Y_i - \underbrace{\boldsymbol{X}_i^\mathsf{T}\widehat{\beta}}_{\widehat{Y}_i}$

R^2 or log-likelihood

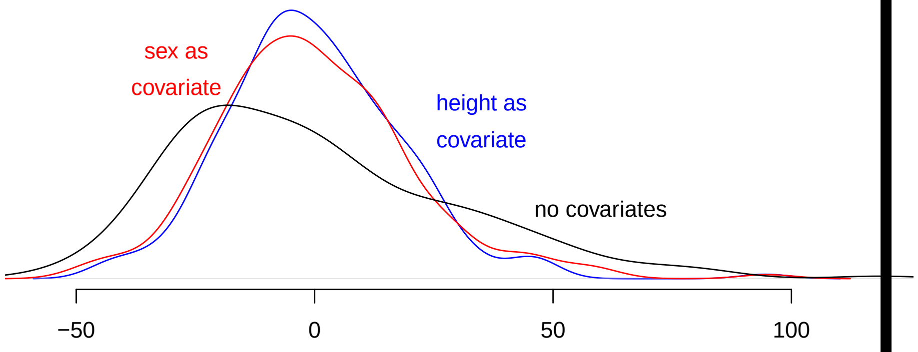
parsimony principle ?

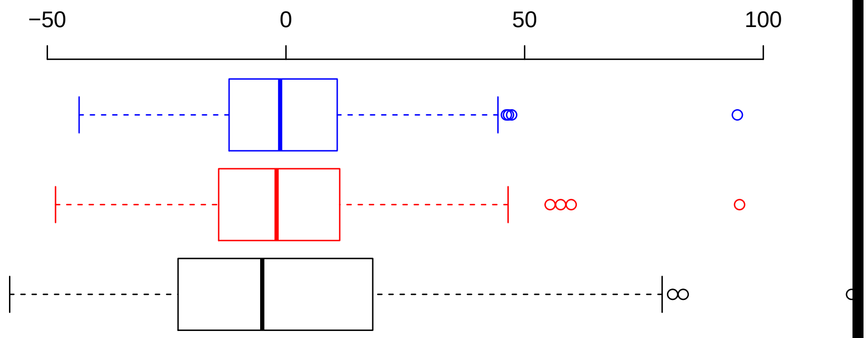
$\longrightarrow$ penalizing the likelihood with

the number of covariates

Akaike (AIC) or

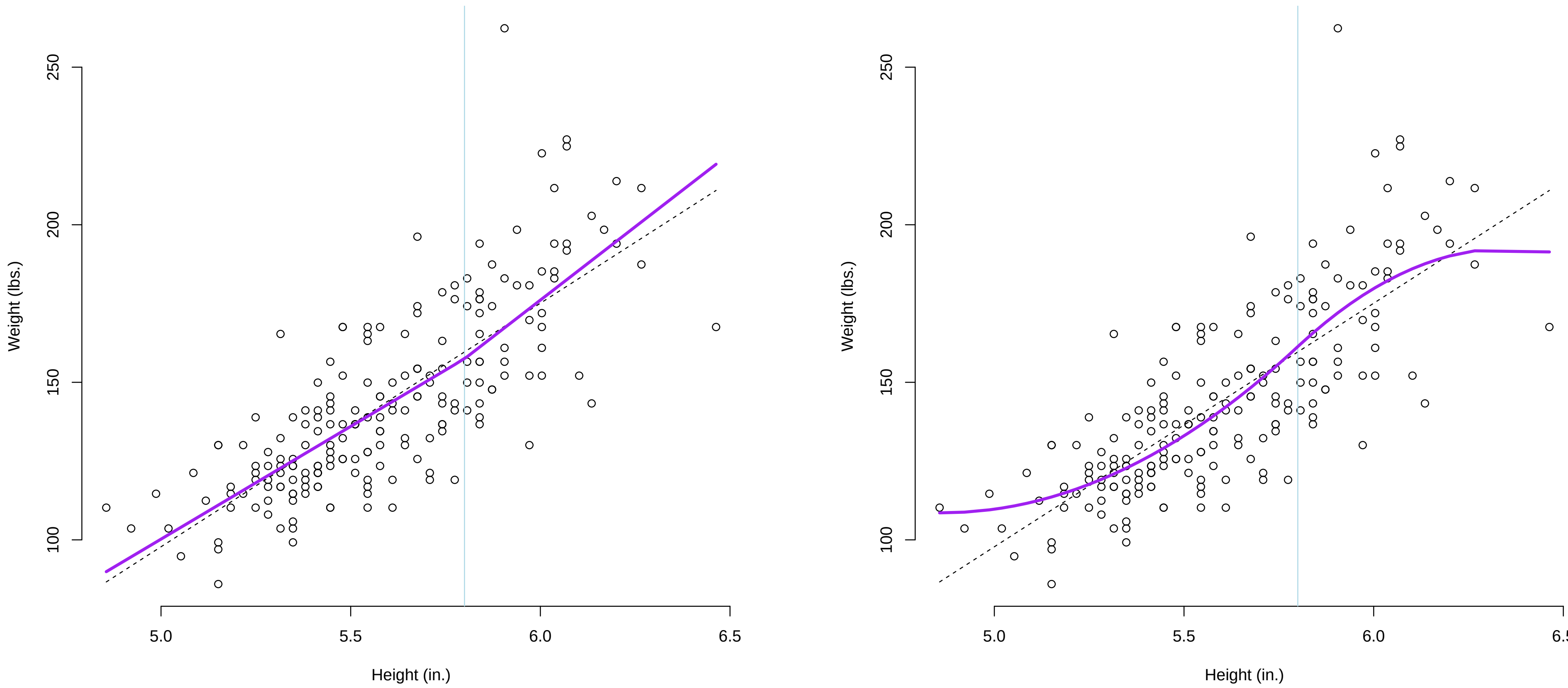
Schwarz (BIC) criteria





Relaxing the linear assumption in predictions

Use of b-spline function basis to estimate $\mu(\cdot)$ where $\mu(\boldsymbol{x}) = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})$



Relaxing the linear assumption in predictions

E.g. predicting someone's weight (Y)

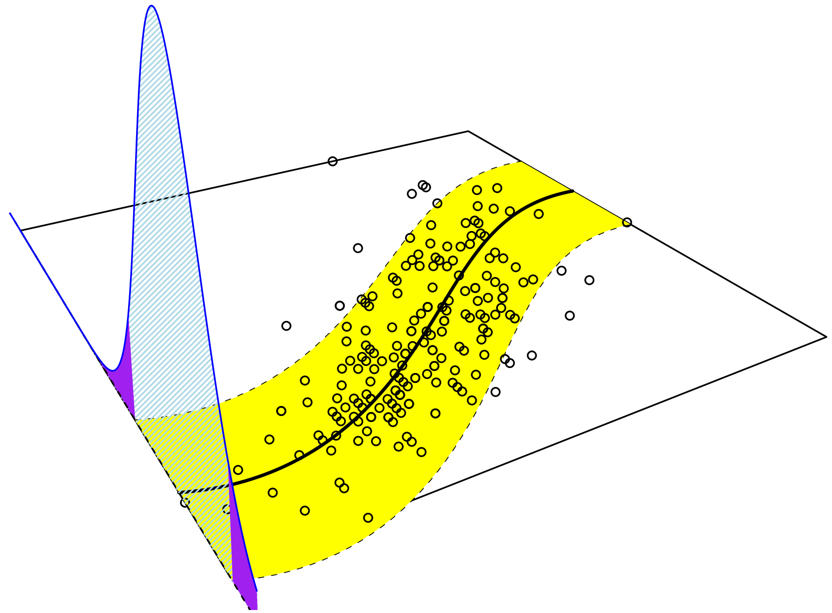
based on his/her height (X_2)

E.g. Gaussian linear model

$$Y|X_2 = x_2 \sim \mathcal{N}(\mu(x_2), \sigma^2)$$

$$\widehat{\mathbb{E}}(Y|X_2 = x_2) = \widehat{\mu}(x_2) = \widehat{Y}(x_2)$$

$$Y \in \left[\widehat{Y}(x_2) \pm \underbrace{u_{1-\alpha/2}}_{1.96} \cdot \widehat{\sigma}\right]$$



Gaussian model : $\mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x}) = \mu(\boldsymbol{x})$ (e.g. $\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta}$) and $\text{Var}(Y|\boldsymbol{X} = \boldsymbol{x}) = \sigma^2$.

Nonlinearities and missing covariates

E.g. predicting someone's weight (Y)

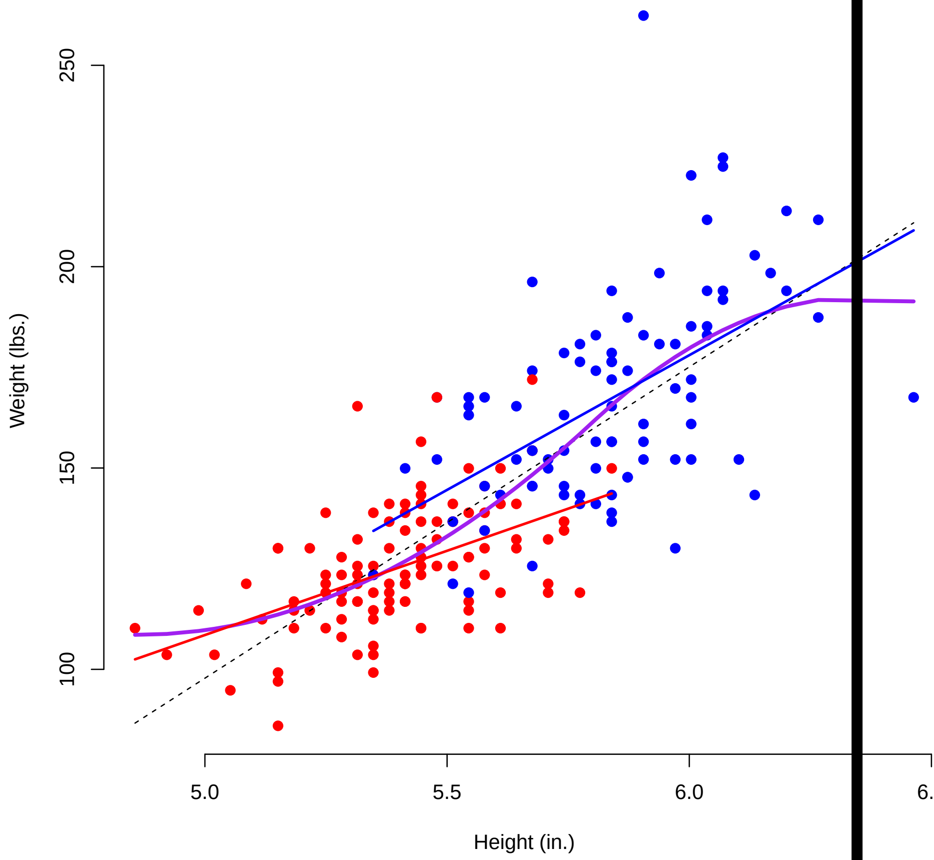
based on his/her height **and** sex

$\longrightarrow$ nonlinearities can be related to

model mispecification

E.g. Gaussian linear model

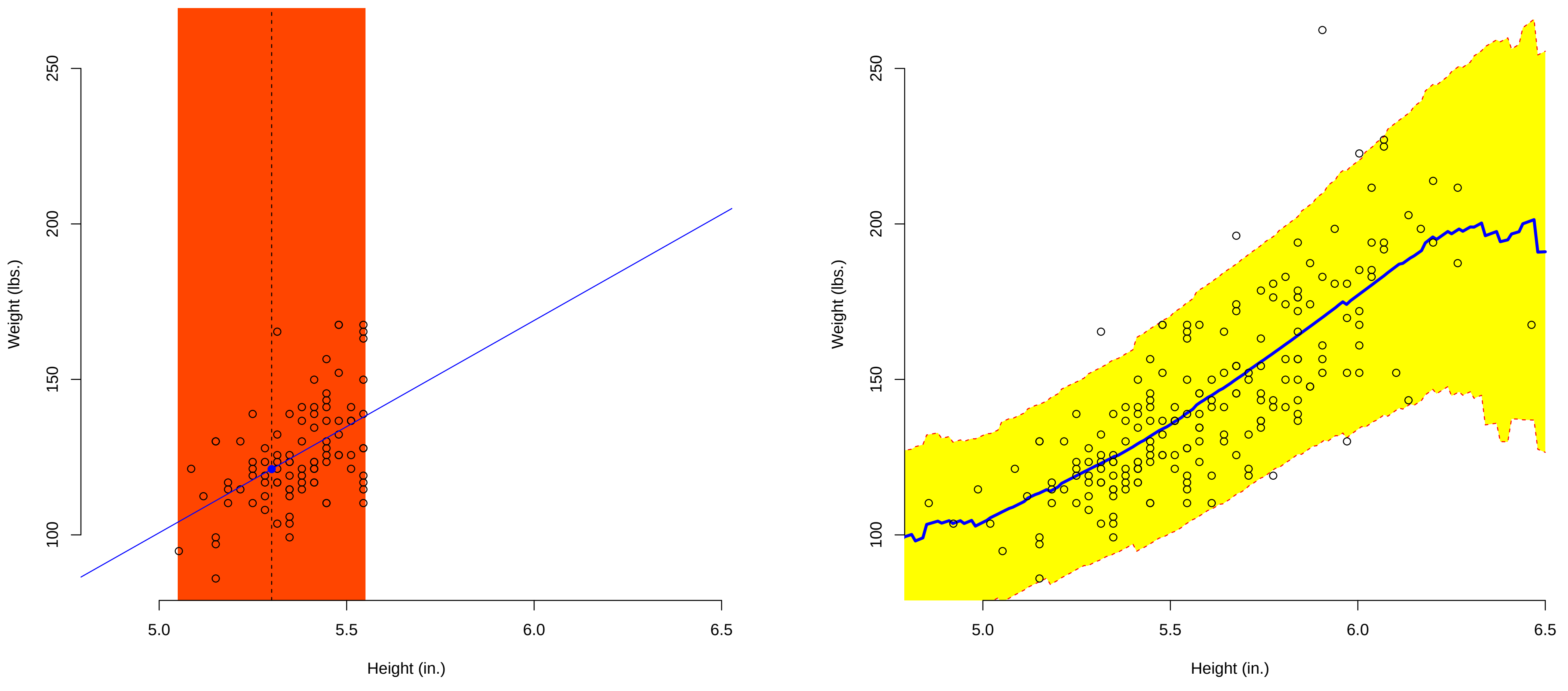
$$Y_i = \begin{cases} \color{red}{\beta_{0,F} + \beta_{2,F} X_{2,i}} + \varepsilon_i \text{ if } X_{1,i} = \text{F} \\ \color{blue}{\beta_{0,M} + \beta_{2,M} X_{2,i}} + \varepsilon_i \text{ if } X_{1,i} = \text{M} \end{cases}$$



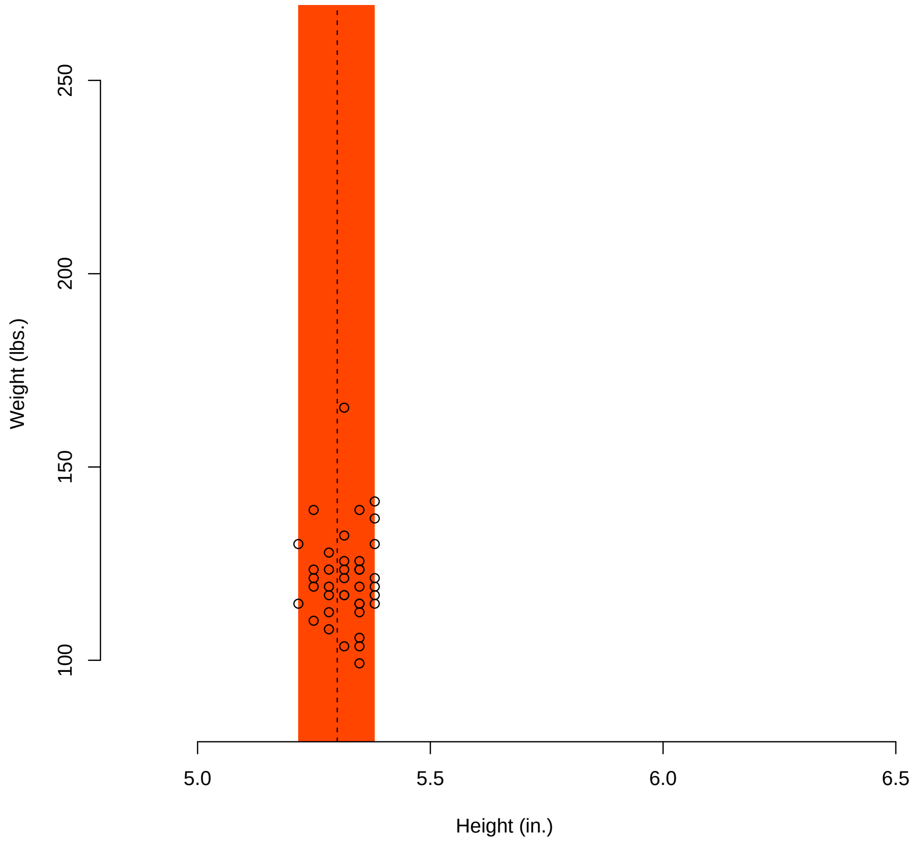
$\longrightarrow$ local linear regression, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{x}} = \text{argmin}\left\{\sum_{i=1}^{n} \omega_i(\boldsymbol{x}) \cdot [Y_i - \boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\beta}]^2\right\}$

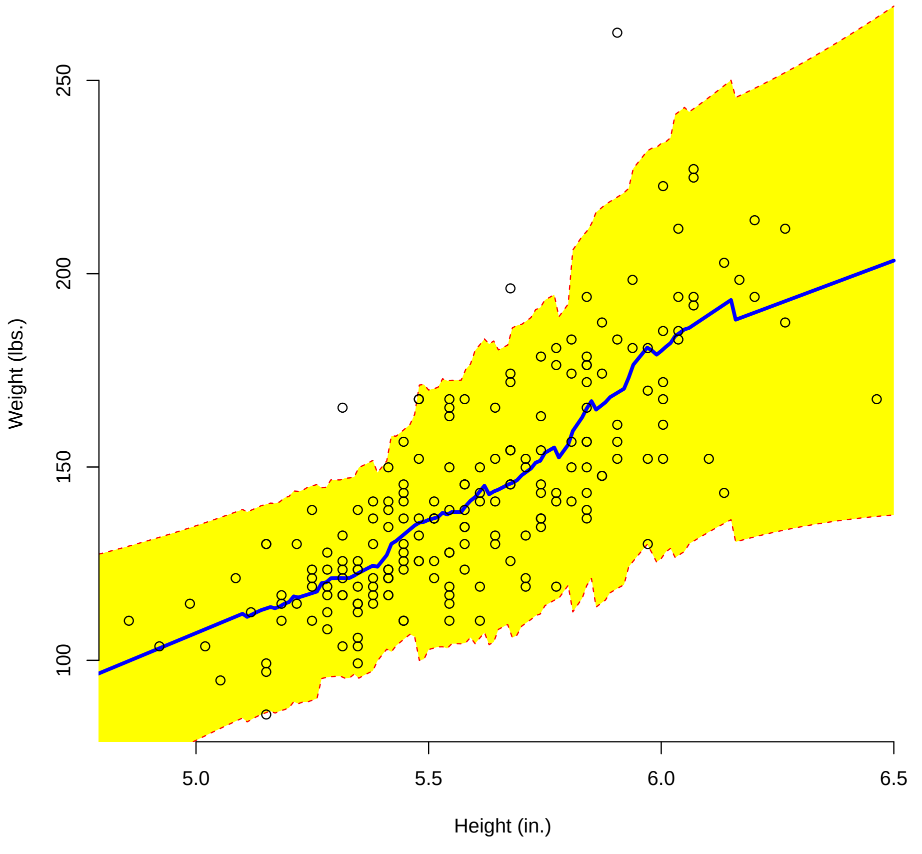
and set $\widehat{Y}(\boldsymbol{x}) = \boldsymbol{x}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{x}}$

Local regression and smoothing techniques

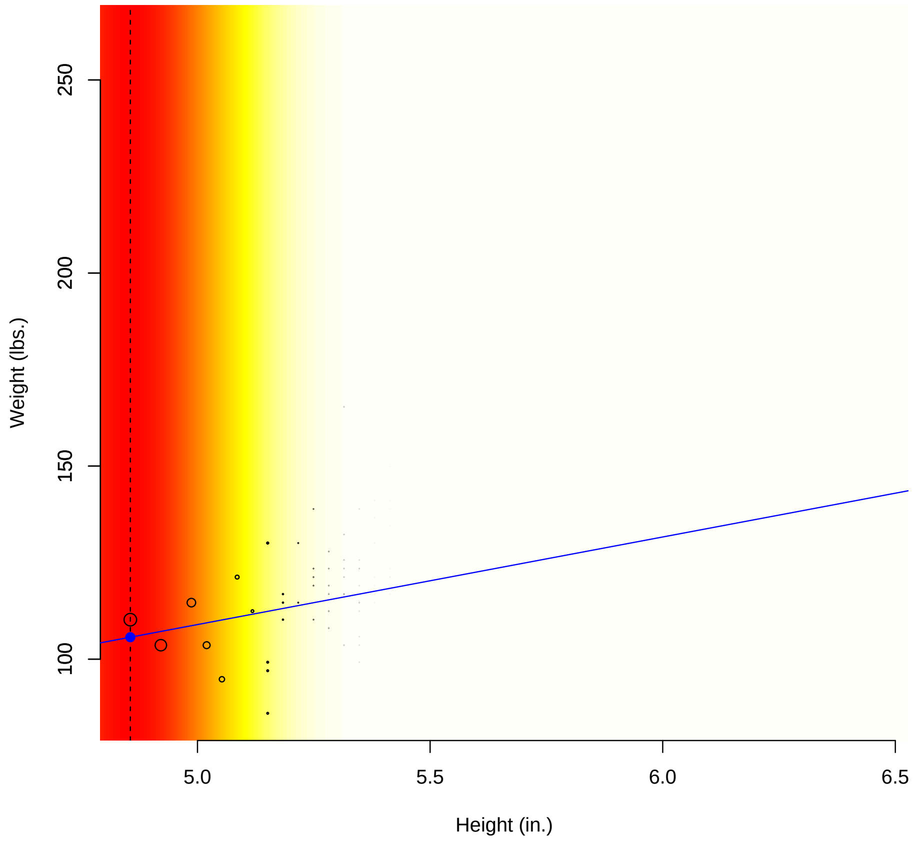


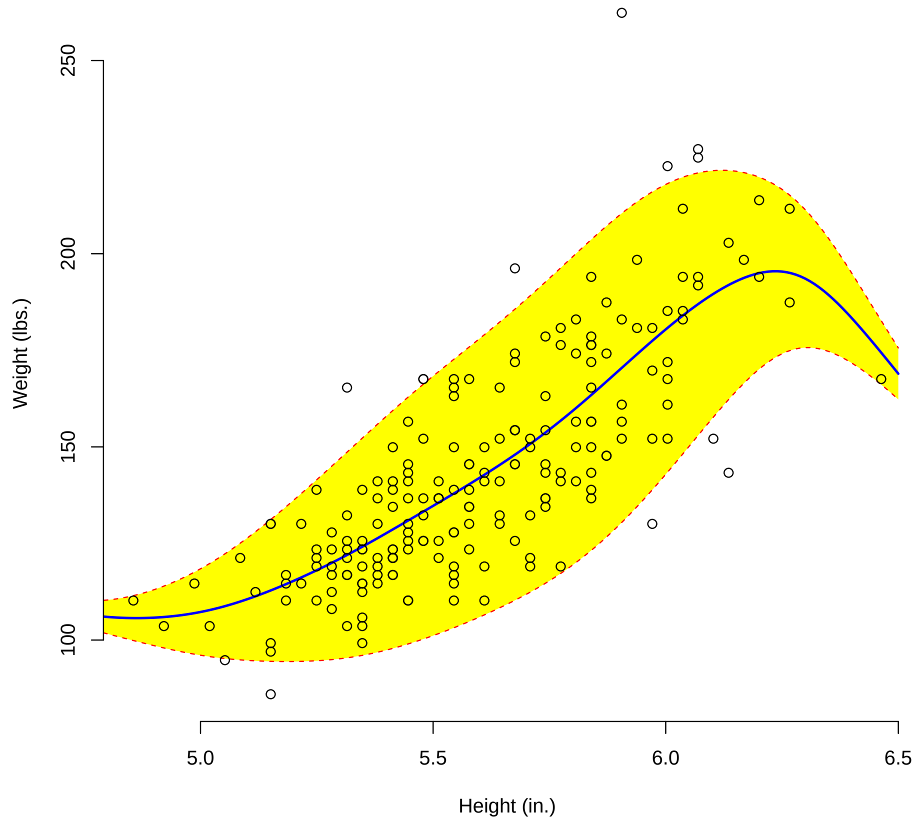
k-nearest neighbours and smoothing techniques





Kernel regression and smoothing techniques





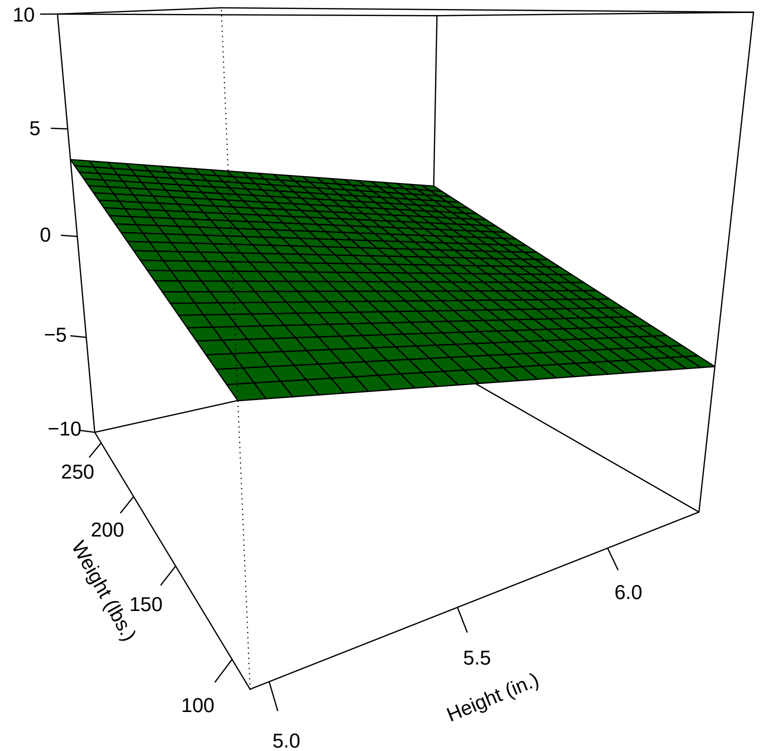
Multiple linear regression

E.g. predicting someone's misperception of his/her weight (Y) based on his/her $\underbrace{\text{height}}_{X_2}$ **and** $\underbrace{\text{weight}}_{X_3}$

$\longrightarrow$ linear model

$$\mathbb{E}(Y|X_2, X_3) = \beta_0 + \beta_2 X_2 + \beta_3 X_3$$

$$\text{Var}(Y|X_2, X_3) = \sigma^2$$



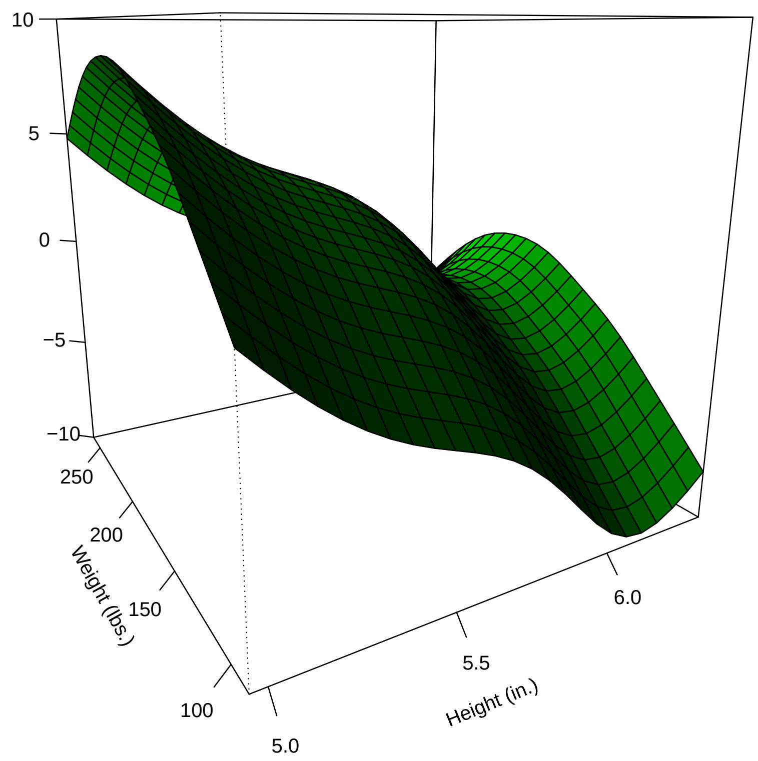
Multiple non-linear regression

E.g. predicting someone's misperception of his/her weight (Y) based on his/her $\underbrace{\text{height}}_{X_2}$ **and** $\underbrace{\text{weight}}_{X_3}$

$\longrightarrow$ **non**-linear model

$$\mathbb{E}(Y|X_2, X_3) = h(X_2, X_3)$$

$$\text{Var}(Y|X_2, X_3) = \sigma^2$$



Away from the Gaussian model

Y is not necessarily Gaussian

Y can be a counting variable,

E.g. Poisson $Y \sim \mathcal{P}(\lambda(\boldsymbol{x}))$

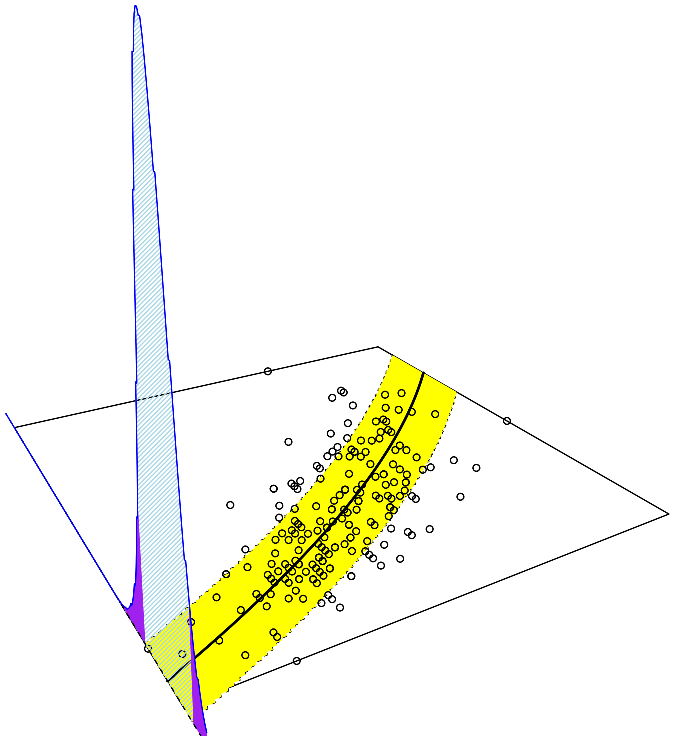
Y can be a FALSE-TRUE variable,

E.g. Binomial $Y \sim \mathcal{B}(p(\boldsymbol{x}))$

(see next section)

$\longrightarrow$ Generalized Linear Model

E.g. $Y|X_2 = x_2 \sim \mathcal{P}\left(e^{\beta_0+\beta_2 x_2}\right)$



Remark With a Poisson model, $\mathbb{E}(Y|X_2 = x_2) = \text{Var}(Y|X_2 = x_2)$.

Logistic regression

E.g. predicting someone's misperception of his/her weight (Y)

$$Y_i = \begin{cases} 1 \text{ if prediction } > \text{ observed weight} \\ 0 \text{ if prediction } \leq \text{ observed weight} \end{cases}$$

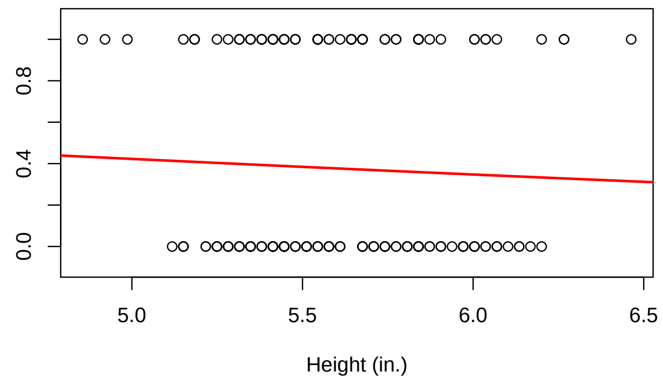
Bernoulli variable,

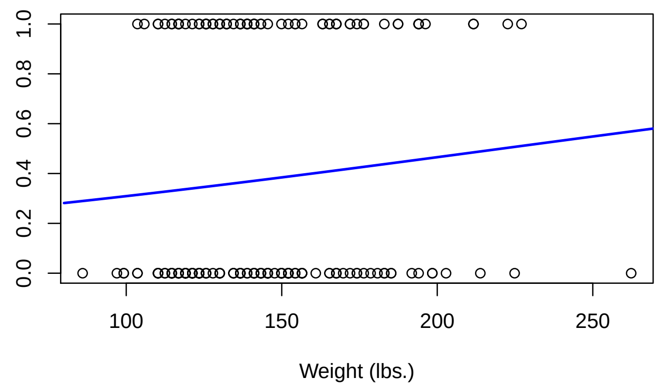
$\mathbb{P}(Y = y) = p^y(1-p)^{1-y}$, where $y \in \{0, 1\}$

$\longrightarrow$ logistic regression

$\mathbb{P}(Y = y|\boldsymbol{X} = \boldsymbol{x}) = p(\boldsymbol{x})^y(1 - p(\boldsymbol{x}))^{1-y}$,

where $y \in \{0, 1\}$





Logistic regression

$\longrightarrow$ logistic regression

$\mathbb{P}(Y=y|\boldsymbol{X}=\boldsymbol{x})=p(\boldsymbol{x})^y(1-p(\boldsymbol{x}))^{1-y}$,

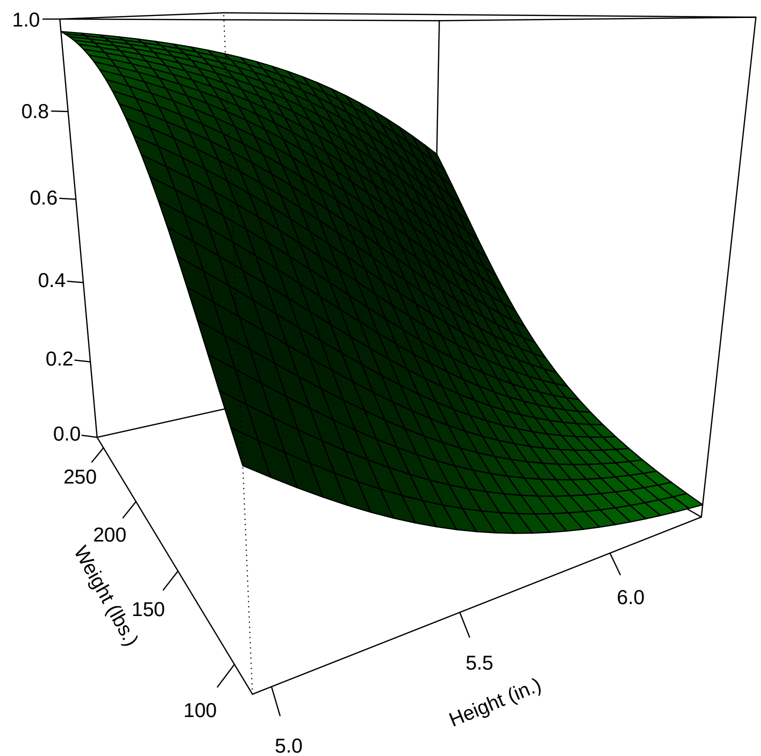
where $y\in\{0,1\}$

Odds ratio $\dfrac{\mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x})}{\mathbb{P}(Y=0|\boldsymbol{X}=\boldsymbol{x})}=\exp\left(\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}\right)$

$$\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x})=\frac{\exp\left(\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}\right)}$$

Estimation of $\boldsymbol{\beta}$?

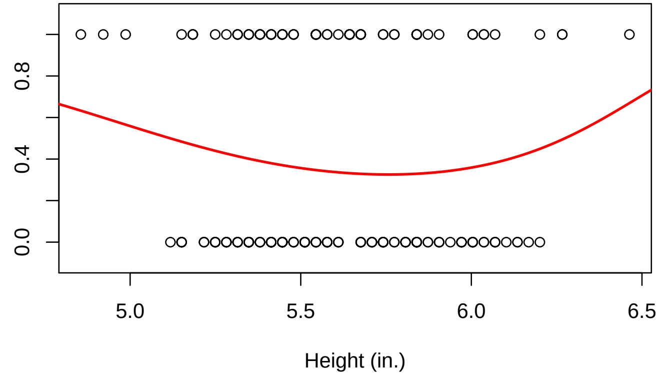
$\longrightarrow$ maximum likelihood $\widehat{\boldsymbol{\beta}}$ (Newton - Raphson)



Smoothed logistic regression

GLMs are linear since

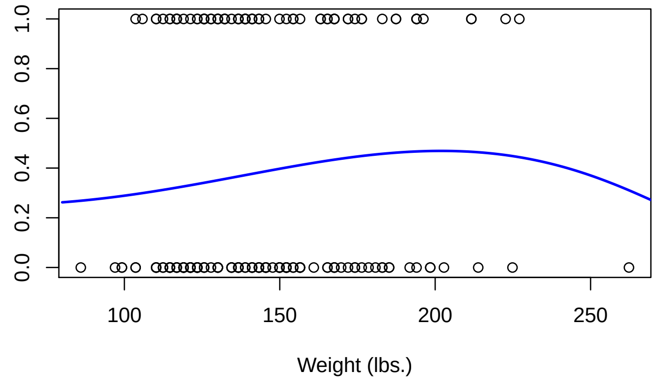
$$\frac{\mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x})}{\mathbb{P}(Y=0|\boldsymbol{X}=\boldsymbol{x})}=\exp\left(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta}\right)$$



$\longrightarrow$ smooth nonlinear function instead

$$\frac{\mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x})}{\mathbb{P}(Y=0|\boldsymbol{X}=\boldsymbol{x})}=\exp\left(h(\boldsymbol{x})\right)$$

$$\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x})=\frac{\exp\left(h(\boldsymbol{x})\right)}{1+\exp\left(h(\boldsymbol{x})\right)}$$



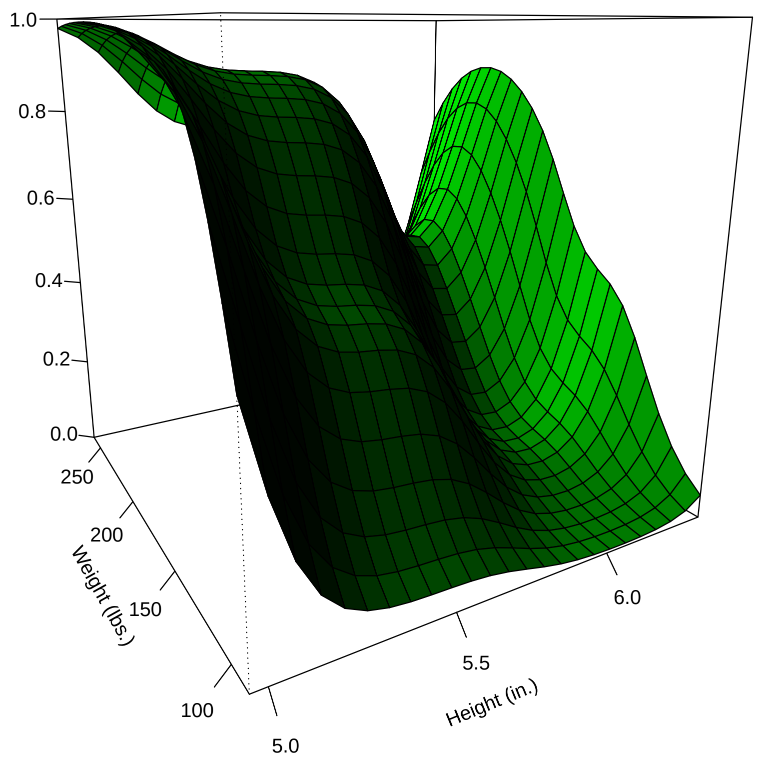
Smoothed logistic regression

$\longrightarrow$ non linear logistic regression

$$\frac{\mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x})}{\mathbb{P}(Y=0|\boldsymbol{X}=\boldsymbol{x})}=\exp\left(h(\boldsymbol{x})\right)$$

$$\mathrm{E}(\mathrm{Y}|\boldsymbol{X}=\boldsymbol{x})=\frac{\exp\left(h(\boldsymbol{x})\right)}{1+\exp\left(h(\boldsymbol{x})\right)}$$

Remark we do not predict Y here, but $\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x})$.



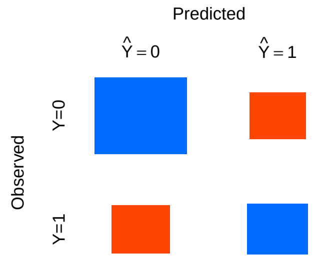
Predictive modeling for a $\{0,1\}$ variable

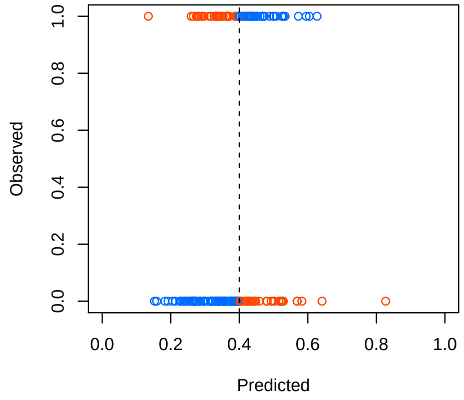
What is a good $\{0,1\}$-model ?

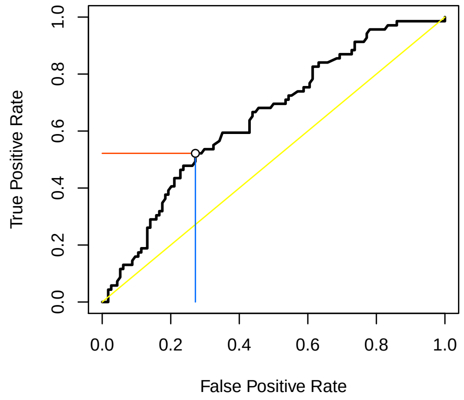
$\longrightarrow$ decision theory

$$\begin{cases} \text{if } \mathbb{P}(Y|\boldsymbol{X}=\boldsymbol{x}) \leq s, & \text{then } \widehat{Y}=0 \\ \text{if } \mathbb{P}(Y|\boldsymbol{X}=\boldsymbol{x}) > s, & \text{then } \widehat{Y}=1 \end{cases}$$

	$\widehat{Y}=0$	$\widehat{Y}=1$
$Y=0$	fine	error
$Y=1$	error	fine







R.O.C. curve

True positive rate

$$TP(s) = \mathbb{P}(\widehat{Y}(s) = 1|Y = 1)$$
$$= \frac{n_{\widehat{Y}=1,Y=1}}{n_{Y=1}}$$

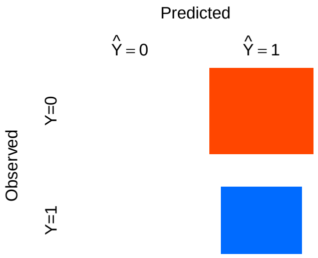
False positive rate

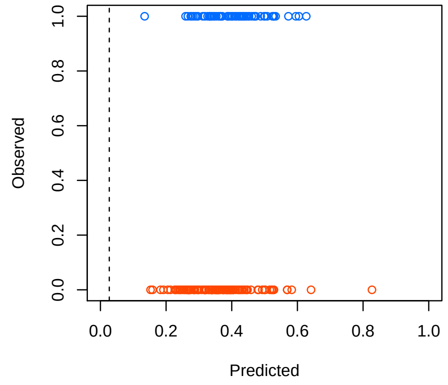
$$FP(s) = \mathbb{P}(\widehat{Y}(s) = 1|Y = 0)$$
$$= \frac{n_{\widehat{Y}=1,Y=1}}{n_{Y=1}}$$

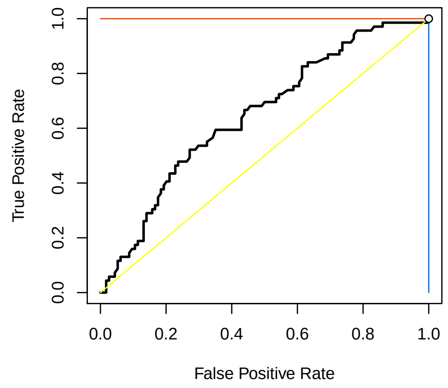
R.O.C. curve is

$$\{FP(s), TP(s)), s \in (0,1)\}$$

(see also model gain curve)







Classification tree (CART)

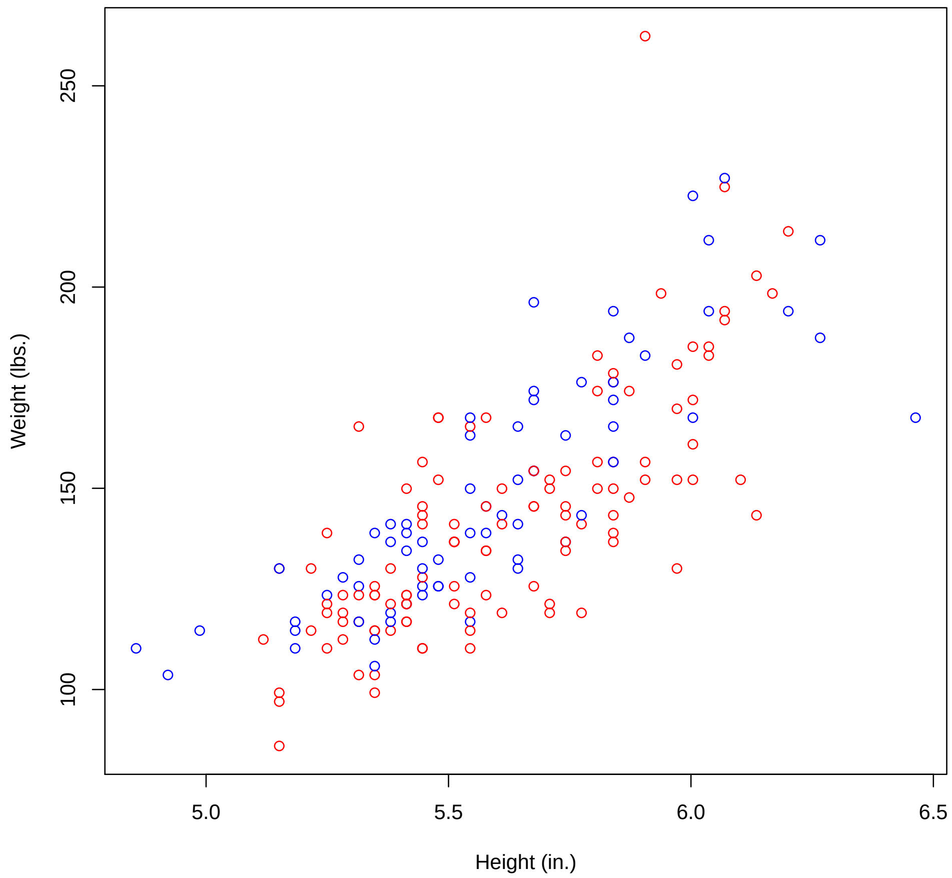
If Y is a TRUE-FALSE variable

prediction is a classification problem.

$$\mathbb{E}(Y|\boldsymbol{X}=\boldsymbol{x}) = p_j \text{ if } \boldsymbol{x} \in A_j$$

where $A_1, \cdots, A_k$ are disjoint

regions of the $\boldsymbol{X}$-space.



Classification tree (CART)

$\longrightarrow$ iterative process

Step 1. Find two subset of indices

either $A_1 = \{i, X_{1,i} < s\}$

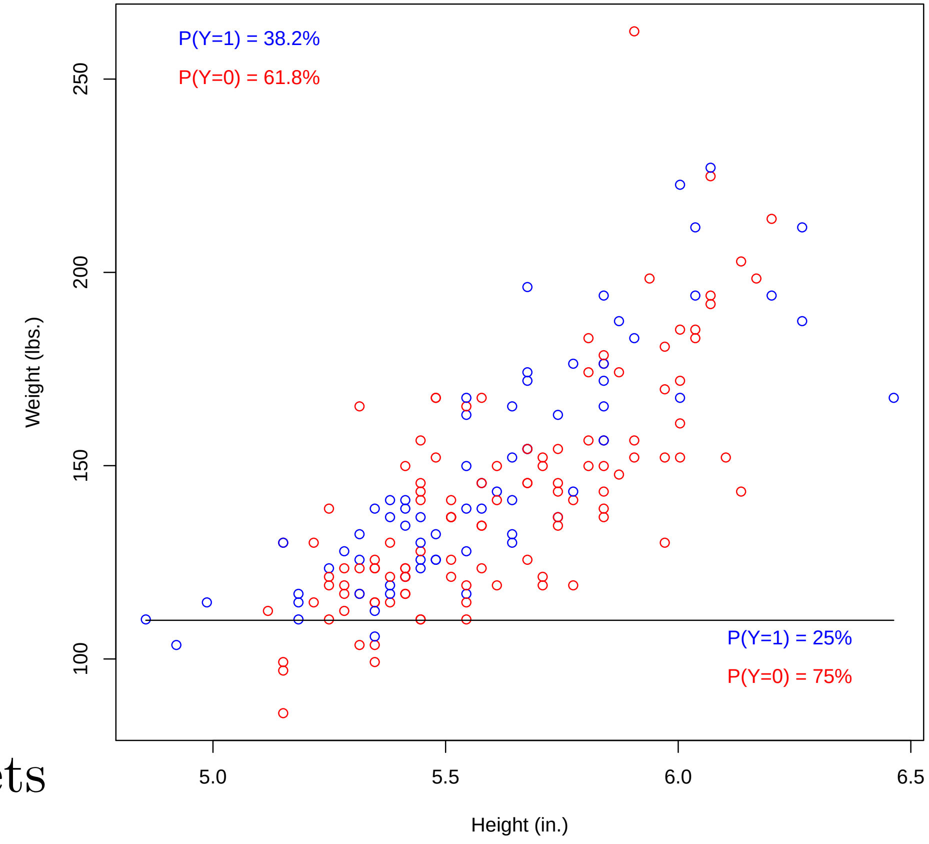
and $A_2 = \{i, X_{1,i} > s\}$

or $A_1 = \{i, X_{2,i} < s\}$

and $A_2 = \{i, X_{2,i} > s\}$

maximize homogeneity within subsets

& maximize heterogeneity between subsets

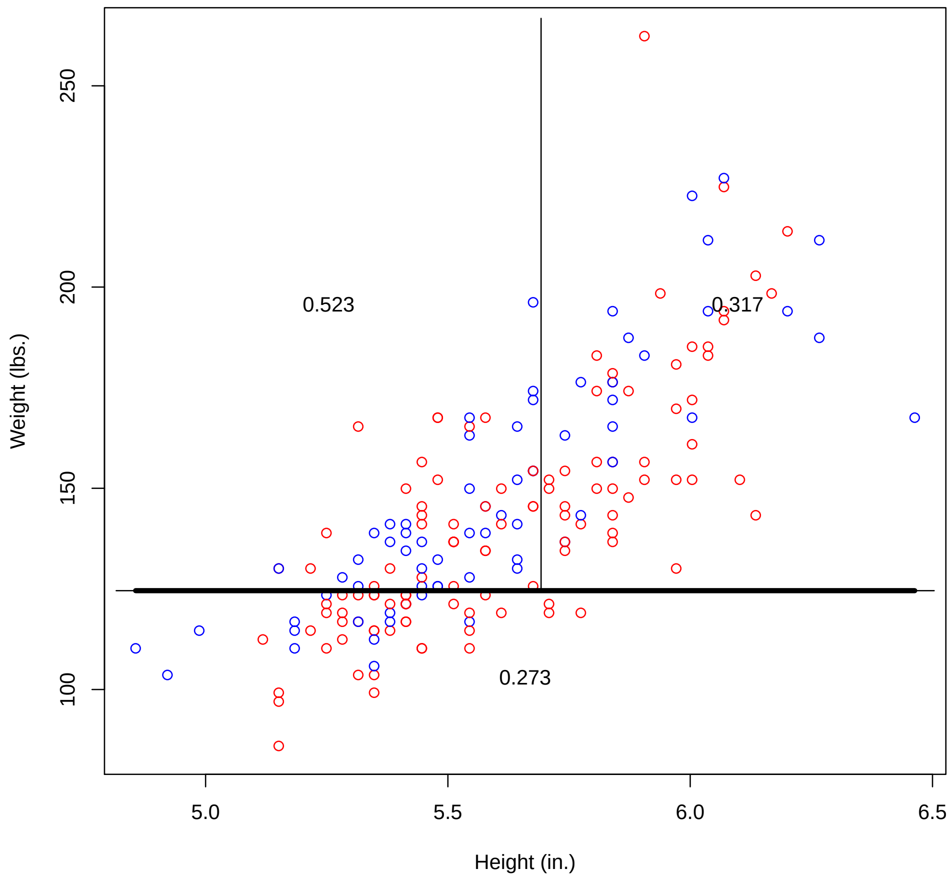


Classification tree (CART)

Need an impurity criteria

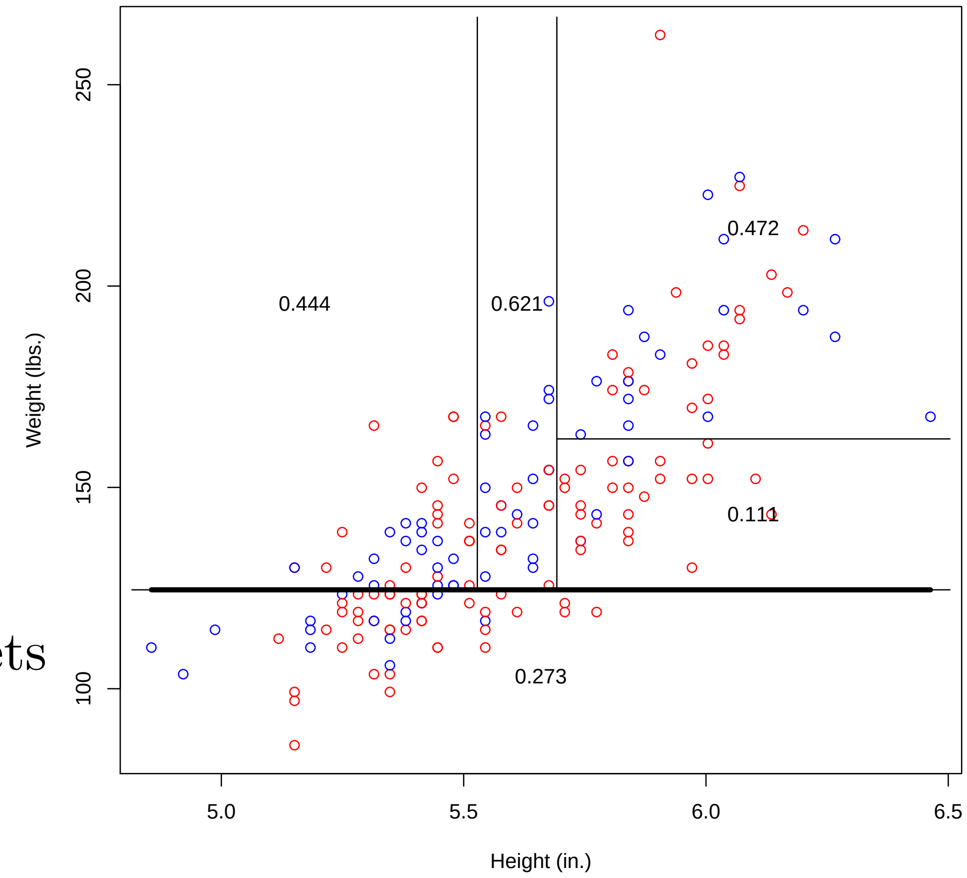
E.g. Gini index

$$-\sum_{\boldsymbol{x}\in\{A_1,A_2\}}\frac{n_x}{n}\sum_{y\in\{0,1\}}\frac{n_{x,y}}{n_x}\cdot\left(1-\frac{n_{x,y}}{n_x}\right)$$

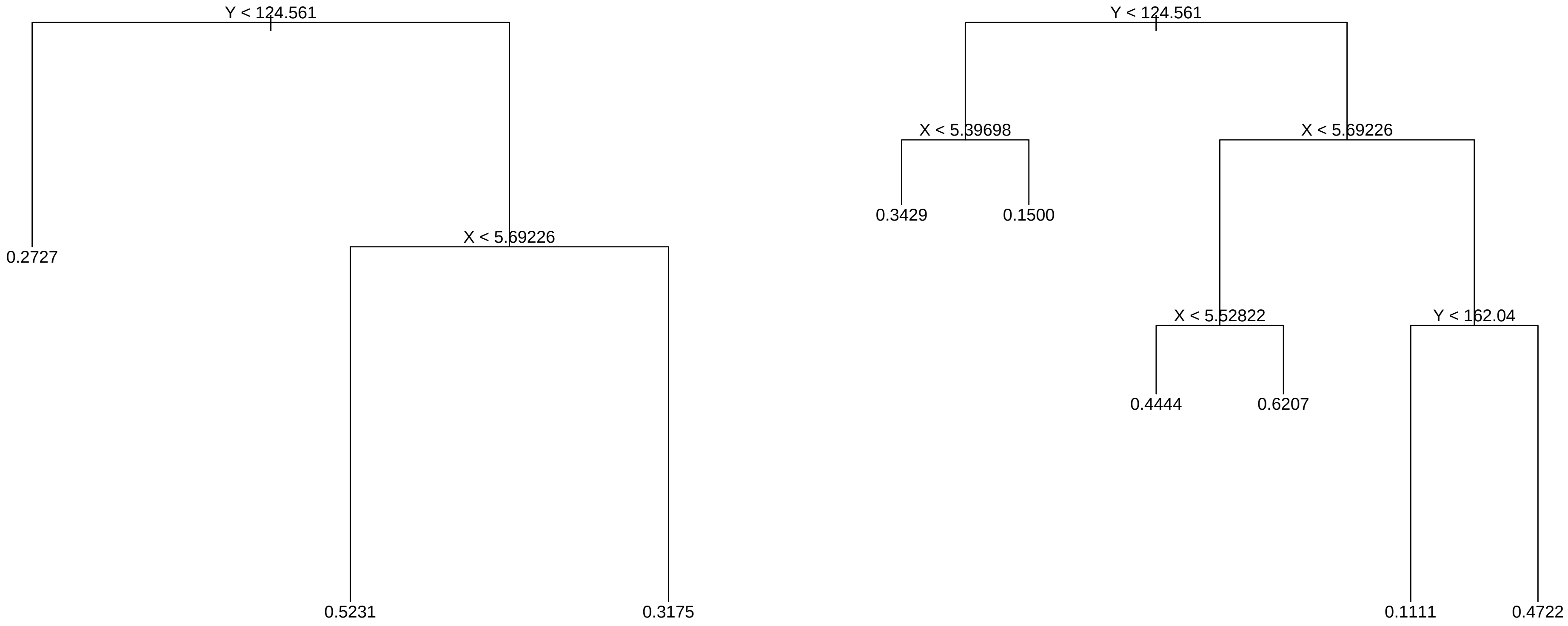


Classification tree (CART)

Step k. Given partition $A_1, \cdots, A_k$ find which subset A_j will be divided, either according to X_1 or according to X_2 maximize homogeneity within subsets & maximize heterogeneity between subsets



Visualizing classification trees (CART)



From trees to forests

Problem CART tree are not robust

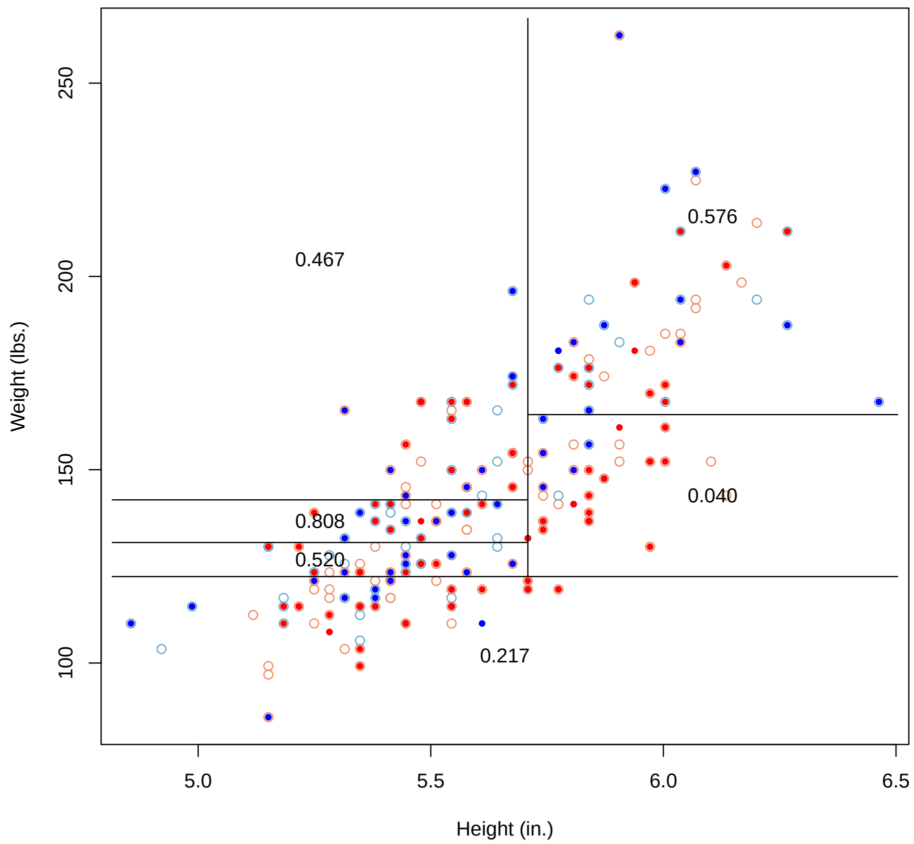
⟶ boosting and bagging

use bootstrap : resample in the data

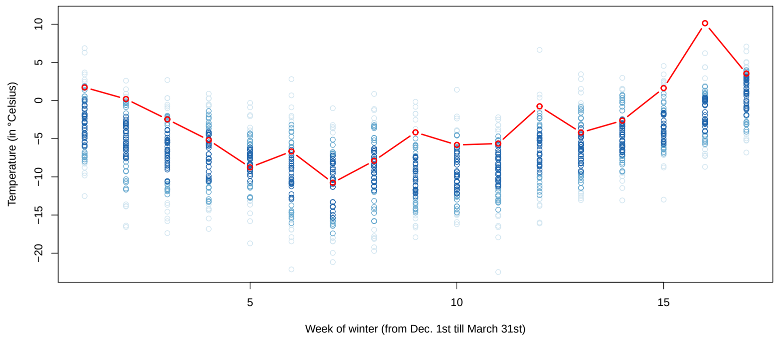
and generate a classification tree

repeat this resampling strategy

Then aggregate all the trees



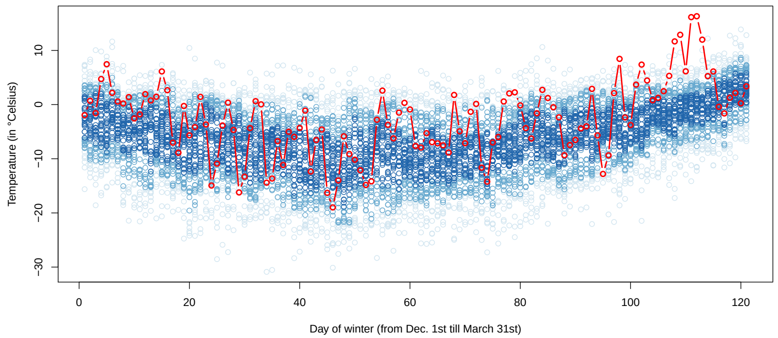
A short word on functional data



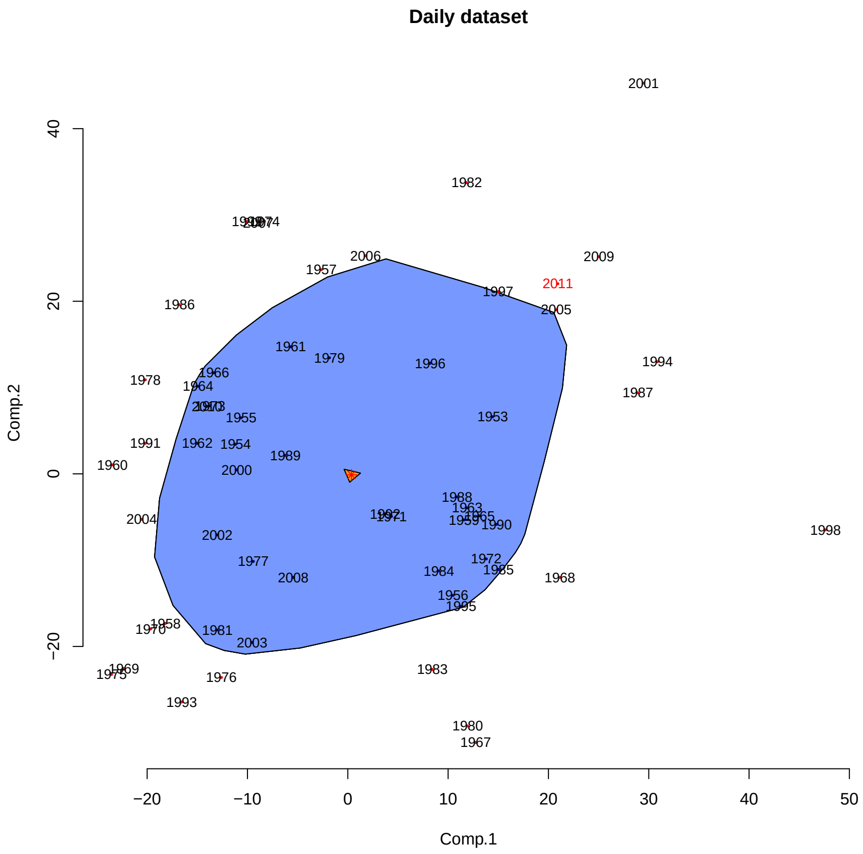
Individual data, $\{Y_i, (X_{1,i}, \cdots, X_{k,i}, \cdots)\}$

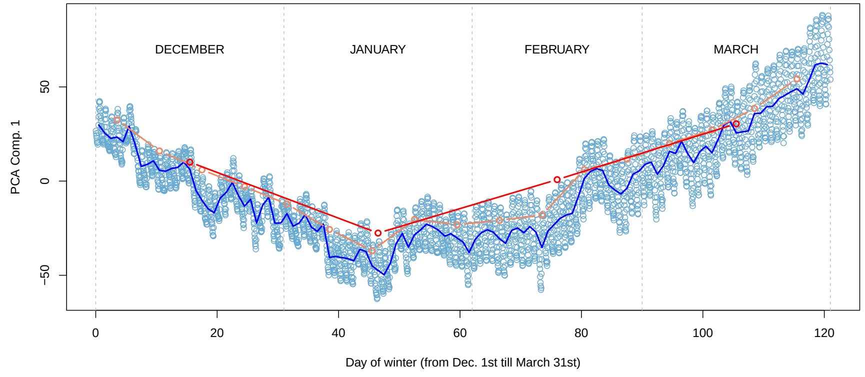
Functional data, $\{\boldsymbol{Y}_i = (Y_{i,1}, \cdots, Y_{i,t}, \cdots)\}$

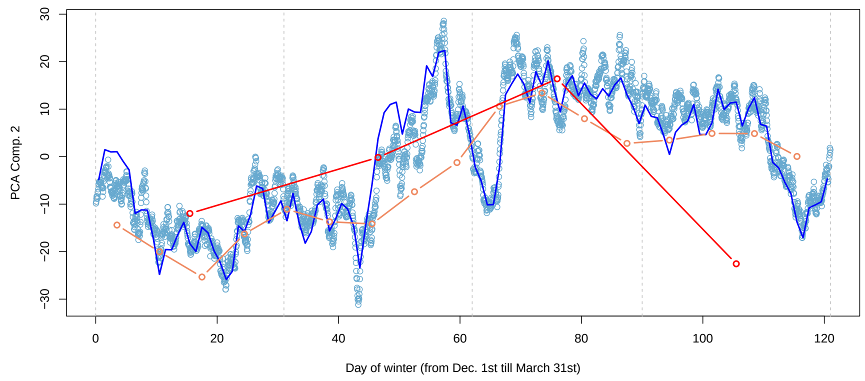
E.g. Winter temperature, in Montréal, QC



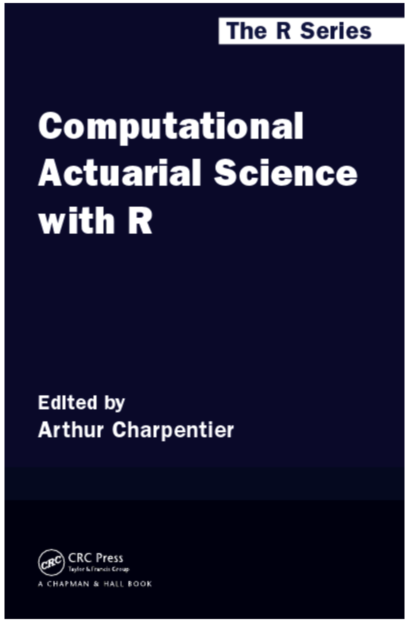
A short word on functional data







To go further...



forthcoming book entitled

Computational Actuarial Science with R