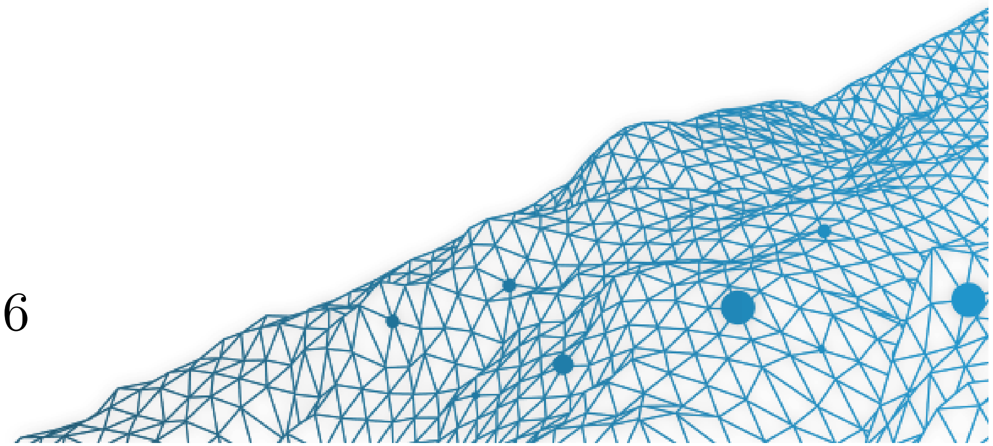
Econometrics: Learning from 'Statistical Learning' Techniques

Arthur Charpentier (Université de Rennes 1 & UQàM)

Centre for Central Banking Studies

Bank of England, London, UK, May 2016

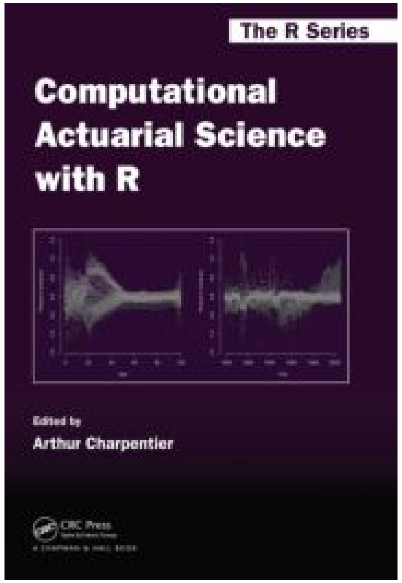
http://freakonometrics.hypotheses.org



Econometrics: Learning from 'Statistical Learning' Techniques

Arthur Charpentier (Université de Rennes 1 & UQàM)

Professor, Economics Department, Univ. Rennes 1
In charge of Data Science for Actuaries program, IA
Research Chair *actinfo* (Institut Louis Bachelier)
(previously Actuarial Sciences at UQàM & ENSAE Paristech
actuary in Hong Kong, IT & Stats FFSA)
PhD in Statistics (KU Leuven), Fellow Institute of Actuaries
MSc in Financial Mathematics (Paris Dauphine) & ENSAE
Editor of the freakonometrics.hypotheses.org's blog
Editor of Computational Actuarial Science, CRC

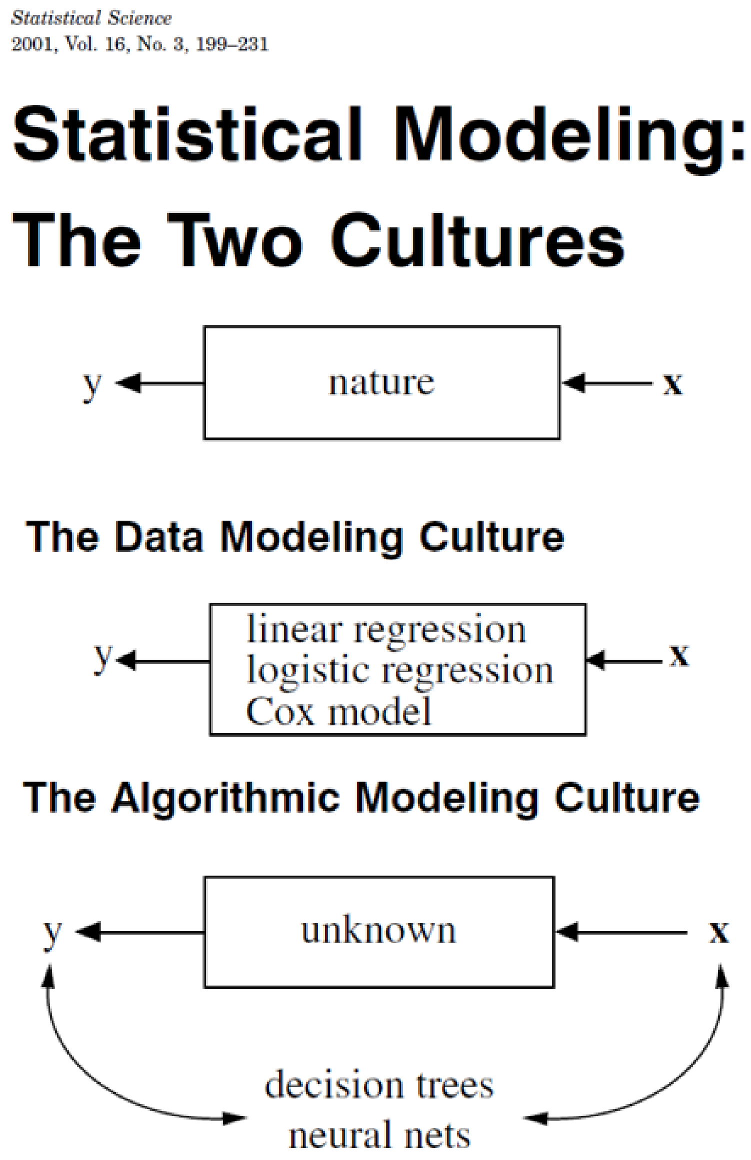


Agenda

"*the numbers have no way of speaking for themselves. We speak for them.* [· · ·] *Before we demand more of our data, we need to demand more of ourselves* " from Silver (2012).

- (big) data
- econometrics & probabilistic modeling
- algorithmics & statistical learning
- different perspectives on classification
- boostrapping, PCA & variable section

see Berk (2008), Hastie, Tibshirani & Friedman (2009), but also Breiman (2001)



Data and Models

From $\{(y_i, \boldsymbol{x}_i)\}$, there are different stories behind, see Freedman (2005)

- the **causal story** : $x_{j,i}$ is usually considered as independent of the other covariates $x_{k,i}$. For all possible $\boldsymbol{x}$, that value is mapped to $m(\boldsymbol{x})$ and a noise is attached, ε. The goal is to recover $m(\cdot)$, and the residuals are just the difference between the response value and $m(\boldsymbol{x})$.
- the **conditional distribution story** : for a linear model, we usually say that Y given $\boldsymbol{X} = \boldsymbol{x}$ is a $\mathcal{N}(m(\boldsymbol{x}), \sigma^2)$ distribution. $m(\boldsymbol{x})$ is then the conditional mean. Here $m(\cdot)$ is assumed to really exist, but no causal assumption is made, only a conditional one.
- the **explanatory data story** : there is no model, just data. We simply want to summarize information contained in $\boldsymbol{x}$'s to get an accurate summary, close to the response (i.e. $\min\{\ell(\boldsymbol{y}, m(\boldsymbol{x}))\}$) for some loss function ℓ.

See also Varian (2014)

Data, Models & Causal Inference

We cannot differentiate data and model that easily..

After an operation, should I stay at hospital, or go back home ?

as in Angrist & Pischke (2008),

$$(\text{health} \mid \text{hospital}) - (\text{health} \mid \text{stayed home}) \qquad [\text{observed}]$$

should be written

$$(\text{health} \mid \text{hospital}) - (\text{health} \mid \textit{had} \text{ stayed home}) \qquad [\text{treatment effect}]$$

$$+\ (\text{health} \mid \textit{had} \text{ stayed home}) - (\text{health} \mid \text{stayed home}) \qquad [\text{selection bias}]$$

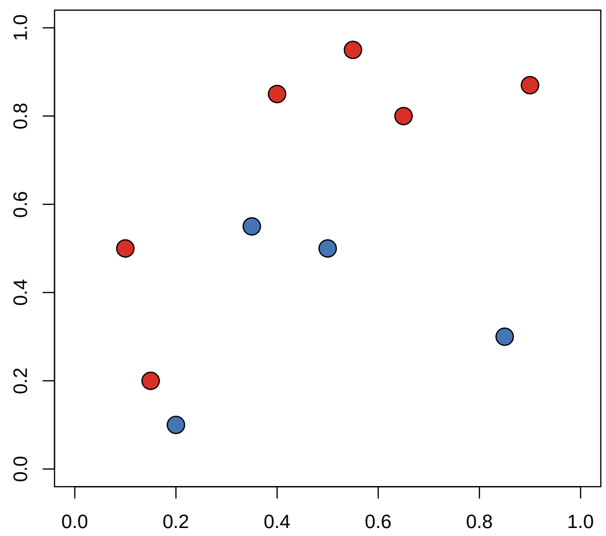
Need randomization to solve selection bias.

Econometric Modeling

Data $\{(y_i, \boldsymbol{x}_i)\}$, for $i = 1, \cdots, n$, with $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^p$ and $y_i \in \mathcal{Y}$.

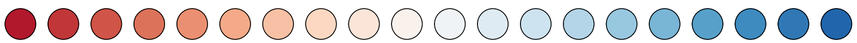
A model is a $m : \mathcal{X} \mapsto \mathcal{Y}$ **mapping**

- regression, $\mathcal{Y} = \mathbb{R}$ (but also $\mathcal{Y} = \mathbb{N}$)
- classification, $\mathcal{Y} = \{0, 1\}$, $\{-1, +1\}$, $\{\bullet, \bullet\}$ (binary, or more)



Classification models are based on two steps,

- **score** function, $s(\boldsymbol{x}) = \mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) \in [0, 1]$



- **classifier** $s(\boldsymbol{x}) \to \widehat{y} \in \{0, 1\}$.

High Dimensional Data (not to say 'Big Data')

See Bühlmann & van de Geer (2011) or Koch (2013), $\boldsymbol{X}$ is a $n \times p$ matrix

Portnoy (1988) proved that maximum likelihood estimators are asymptotically normal when $p^2/n \to 0$ as $n, p \to \infty$. Hence, **massive data**, when $p > \sqrt{n}$.

More intersting is the **sparcity** concept, based not on p, but on the effective size. Hence one can have $p > n$ and convergent estimators.

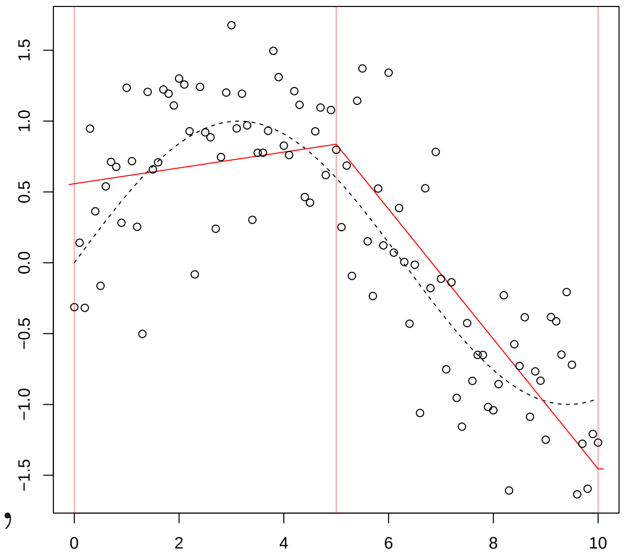
High dimension might be scary because of **curse of dimensionality**, see Bellman (1957). The volume of the unit sphere in $\mathbb{R}^p$ tends to 0 as $p \to \infty$, i.e.space is sparse.

Computational & Nonparametric Econometrics

Linear Econometrics: estimate $g : \boldsymbol{x} \mapsto \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$ by a linear function.

Nonlinear Econometrics: consider the approximation for some **functional basis**

$$g(\boldsymbol{x}) = \sum_{j=0}^{\infty} \omega_j g_j(\boldsymbol{x}) \text{ and } \widehat{g}(\boldsymbol{x}) = \sum_{j=0}^{h} \omega_j g_j(\boldsymbol{x})$$



or consider a **local model**, on the neighborhood of $\boldsymbol{x}$,

$$\widehat{g}(x) = \frac{1}{n_{\boldsymbol{x}}} \sum_{i \in \mathcal{I}_{\boldsymbol{x}}} y_i, \text{ with } \mathcal{I}_{\boldsymbol{x}} = \{\boldsymbol{x} \in \mathbb{R}^p : \|\boldsymbol{x}_i - \boldsymbol{x}\| \leq h\},$$

see Nadaraya (1964) and Watson (1964).

Here h is some **tunning parameter**: not estimated, but chosen (optimaly).

Econometrics & Probabilistic Model

> The primary goal in a regression analysis is to understand, as far as possible with the available data, how the conditional distribution of the response y varies across subpopulations determined by the possible values of the predictor or predictors. Since this is the central idea, it will be helpful to have a conve-

from Cook & Weisberg (1999), see also Haavelmo (1965).

$$(Y|\boldsymbol{X}=\boldsymbol{x}) \sim \mathcal{N}(\mu(\boldsymbol{x}),\sigma^2) \text{ with } \mu(\boldsymbol{x}) = \beta_0 + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}, \text{ and } \boldsymbol{\beta}\in\mathbb{R}^p.$$

Linear Model: $\mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}] = \beta_0 + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}$

Homoscedasticity: $\text{Var}[Y|\boldsymbol{X}=\boldsymbol{x}] = \sigma^2$.

Conditional Distribution and Likelihood

$$(Y|\boldsymbol{X}=\boldsymbol{x}) \sim \mathcal{N}(\mu(\boldsymbol{x}),\sigma^2) \text{ with } \mu(\boldsymbol{x}) = \beta_0 + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta},\ \text{ et } \boldsymbol{\beta} \in \mathbb{R}^p$$

The log-likelihood is

$$\log \mathcal{L}(\beta_0,\boldsymbol{\beta},\sigma^2|\boldsymbol{y},\boldsymbol{x}) = -\frac{n}{2}\log[2\pi\sigma^2] - \frac{1}{2\sigma^2}\underbrace{\sum_{i=1}^{n}(y_i - \beta_0 - \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta})^2}.$$

Set

$$(\widehat{\beta}_0,\widehat{\boldsymbol{\beta}},\widehat{\sigma}^2) = \text{argmax}\left\{\log \mathcal{L}(\beta_0,\boldsymbol{\beta},\sigma^2|\boldsymbol{y},\boldsymbol{x})\right\}.$$

First order condition $\boldsymbol{X}^\mathsf{T}[\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}]=\boldsymbol{0}$. If matrix $\boldsymbol{X}$ is a full rank matrix

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y} = \boldsymbol{\beta} + (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{\varepsilon}.$$

Asymptotic properties of $\widehat{\boldsymbol{\beta}}$,

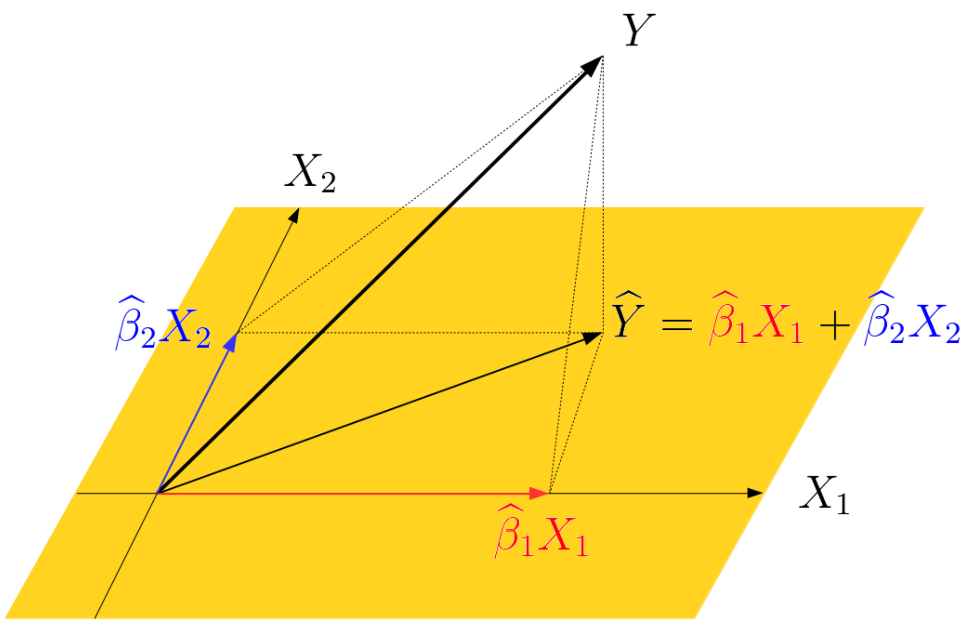
$$\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma}) \text{ as } n\to\infty$$

Geometric Perspective

Define the orthogonal projection on $\mathcal{X}$,

$$\Pi_{\boldsymbol{X}} = \boldsymbol{X}[\boldsymbol{X}^\top\boldsymbol{X}]^{-1}\boldsymbol{X}^\top$$

$$\widehat{\boldsymbol{y}} = \underbrace{\boldsymbol{X}[\boldsymbol{X}^\top\boldsymbol{X}]^{-1}\boldsymbol{X}^\top}_{\Pi_{\boldsymbol{X}}}\boldsymbol{y} = \Pi_{\boldsymbol{X}}\boldsymbol{y}.$$



Pythagoras' theorem can be writen

$$\|\boldsymbol{y}\|^2 = \|\Pi_{\boldsymbol{X}}\boldsymbol{y}\|^2 + \|\Pi_{\boldsymbol{X}^\perp}\boldsymbol{y}\|^2 = \|\Pi_{\boldsymbol{X}}\boldsymbol{y}\|^2 + \|\boldsymbol{y} - \Pi_{\boldsymbol{X}}\boldsymbol{y}\|^2$$

which can be expressed as

$$\underbrace{\sum_{i=1}^n y_i^2}_{n\times\text{total variance}} = \underbrace{\sum_{i=1}^n \widehat{y}_i^2}_{n\times\text{explained variance}} + \underbrace{\sum_{i=1}^n (y_i - \widehat{y}_i)^2}_{n\times\text{residual variance}}$$

Geometric Perspective

Define the angle θ between $\boldsymbol{y}$ and $\Pi_{\mathcal{X}}\boldsymbol{y}$,

$$R^2 = \frac{\|\Pi_{\mathcal{X}}\boldsymbol{y}\|^2}{\|\boldsymbol{y}\|^2} = 1 - \frac{\|\Pi_{\mathcal{X}^\perp}\boldsymbol{y}\|^2}{\|\boldsymbol{y}\|^2} = \cos^2(\theta)$$

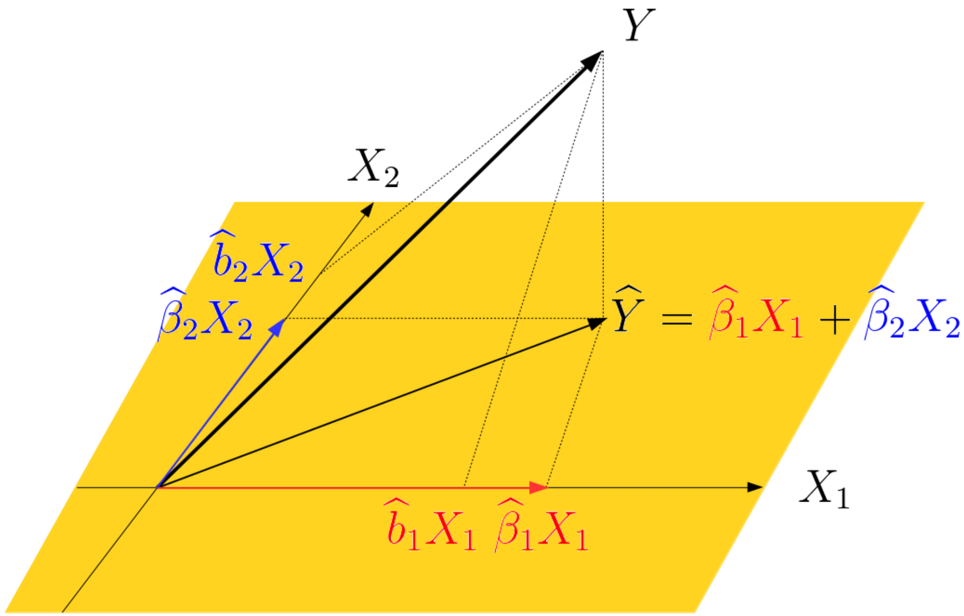
see Davidson & MacKinnon (2003)

$$\boldsymbol{y} = \beta_0 + \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \varepsilon$$

If $\boldsymbol{y}_2^\star = \Pi_{\mathcal{X}_1^\perp}\boldsymbol{y}$ and $\boldsymbol{X}_2^\star = \Pi_{\mathcal{X}_1^\perp}\boldsymbol{X}_2$, then

$$\widehat{\boldsymbol{\beta}}_2 = [\boldsymbol{X}_2^{\star\mathsf{T}}\boldsymbol{X}_2^\star]^{-1}\boldsymbol{X}_2^{\star\mathsf{T}}\boldsymbol{y}_2^\star$$

$\boldsymbol{X}_2^\star = \boldsymbol{X}_2$ if $\boldsymbol{X}_1 \perp \boldsymbol{X}_2$,
Frisch-Waugh theorem.



From Linear to Non-Linear

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\beta} = \underbrace{\boldsymbol{X}[\boldsymbol{X}^\top\boldsymbol{X}]^{-1}\boldsymbol{X}^\top}_{\boldsymbol{H}}\boldsymbol{y} \text{ i.e. } \widehat{y}_i = \boldsymbol{h}_{\boldsymbol{x}_i}^\top\boldsymbol{y},$$

with - for the linear regression - $\boldsymbol{h}_{\boldsymbol{x}} = \boldsymbol{X}[\boldsymbol{X}^\top\boldsymbol{X}]^{-1}\boldsymbol{x}$.

One can consider some smoothed regression, see Nadaraya (1964) and Watson (1964), with some smoothing matrix $\boldsymbol{S}$

$$\widehat{m}_h(x) = \boldsymbol{s}_x^\top\boldsymbol{y} = \sum_{i=1}^n s_{x,i}y_i \text{ withs } s_{x,i} = \frac{K_h(x-x_i)}{K_h(x-x_1)+\cdots+K_h(x-x_n)}$$

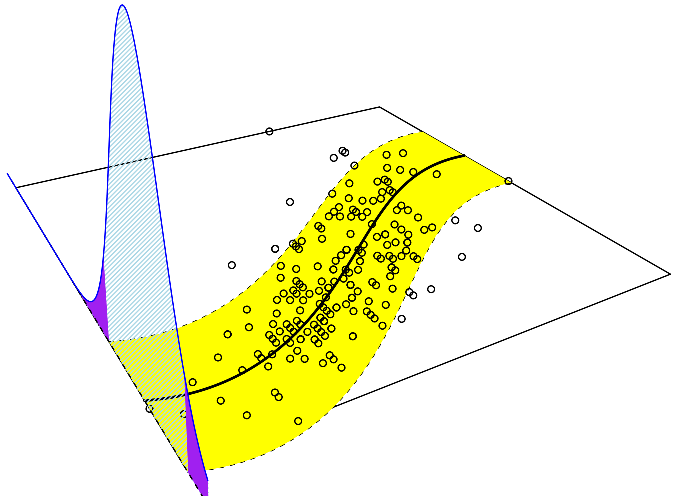
for some kernel $K(\cdot)$ and some bandwidth $h>0$.

From Linear to Non-Linear

$$T = \frac{\|\boldsymbol{S}\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y}\|}{\text{trace}([\boldsymbol{S} - \boldsymbol{H}]^{\mathsf{T}}[\boldsymbol{S} - \boldsymbol{H}])}$$

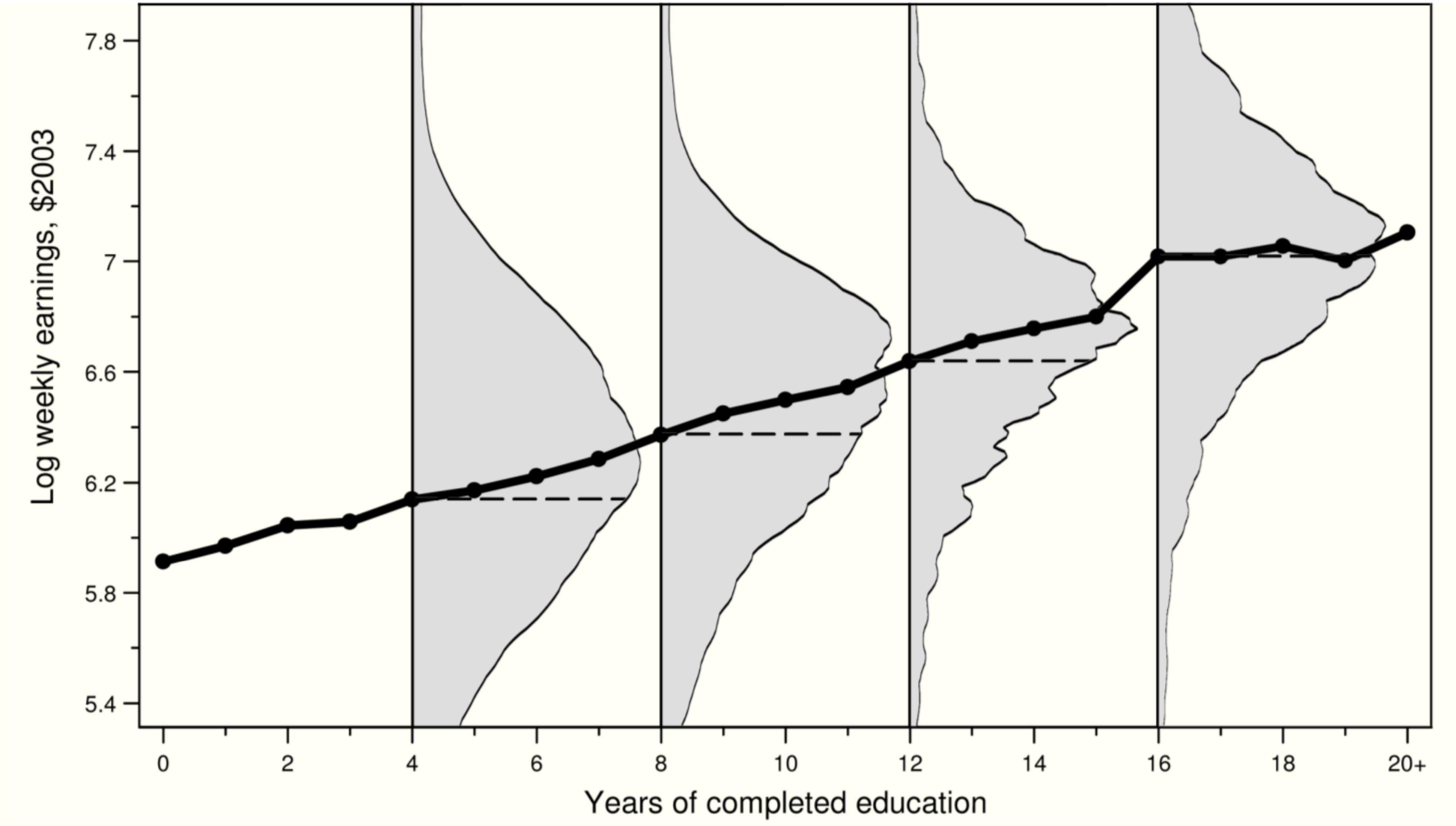
can be used to test for linearity, Simonoff (1996). trace($\boldsymbol{S}$) is the equivalent number of parameters, and $n-\text{trace}(\boldsymbol{S})$ the degrees of freedom, Ruppert *et al.* (2003).

Nonlinear Model, but Homoscedastic - Gaussian



- $(Y|\boldsymbol{X}=\boldsymbol{x}) \sim \mathcal{N}(\mu(\boldsymbol{x}), \sigma^2)$
- $\mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}] = \mu(\boldsymbol{x})$

Conditional Expectation



from Angrist & Pischke (2008), $\boldsymbol{x} \mapsto \mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}]$.

Exponential Distributions and Linear Models

$$f(y_i|\theta_i,\phi)=\exp\left(\frac{y_i\theta_i-b(\theta_i)}{a(\phi)}+c(y_i,\phi)\right)\text{ with }\theta_i=h(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta})$$

Log likelihood is expressed as

$$\log\mathcal{L}(\boldsymbol{\theta},\phi|\boldsymbol{y})=\sum_{i=1}^n\log f(y_i|\theta_i,\phi)=\frac{\sum_{i=1}^n y_i\theta_i-\sum_{i=1}^n b(\theta_i)}{a(\phi)}+\sum_{i=1}^n c(y_i,\phi)$$

and first order conditions

$$\frac{\partial\log\mathcal{L}(\boldsymbol{\theta},\phi|\boldsymbol{y})}{\partial\boldsymbol{\beta}}=\boldsymbol{X}^\mathsf{T}\boldsymbol{W}^{-1}[\boldsymbol{y}-\boldsymbol{\mu}]=\boldsymbol{0}$$

as in Müller (2001), where $\boldsymbol{W}$ is a weight matrix, function of $\boldsymbol{\beta}$.

We usually specify the **link** function $g(\cdot)$ defined as

$$\widehat{y}=m(\boldsymbol{x})=\mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}]=g^{-1}(\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}).$$

Exponential Distributions and Linear Models

Note that $\boldsymbol{W} = \text{diag}(\nabla g(\widehat{\boldsymbol{y}}) \cdot \text{Var}[\boldsymbol{y}])$, and set

$$\boldsymbol{z} = g(\widehat{\boldsymbol{y}}) + (\boldsymbol{y} - \widehat{\boldsymbol{y}}) \cdot \nabla g(\widehat{\boldsymbol{y}})$$

the the maximum likelihood estimator is obtained iteratively

$$\widehat{\boldsymbol{\beta}}_{k+1} = [\boldsymbol{X}^\mathsf{T} \boldsymbol{W}_k^{-1} \boldsymbol{X}]^{-1} \boldsymbol{X}^\mathsf{T} \boldsymbol{W}_k^{-1} \boldsymbol{z}_k$$

Set $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_\infty$, so that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, I(\boldsymbol{\beta})^{-1})$$

with $I(\boldsymbol{\beta}) = \phi \cdot [\boldsymbol{X}^\mathsf{T} \boldsymbol{W}_\infty^{-1} \boldsymbol{X}]$.

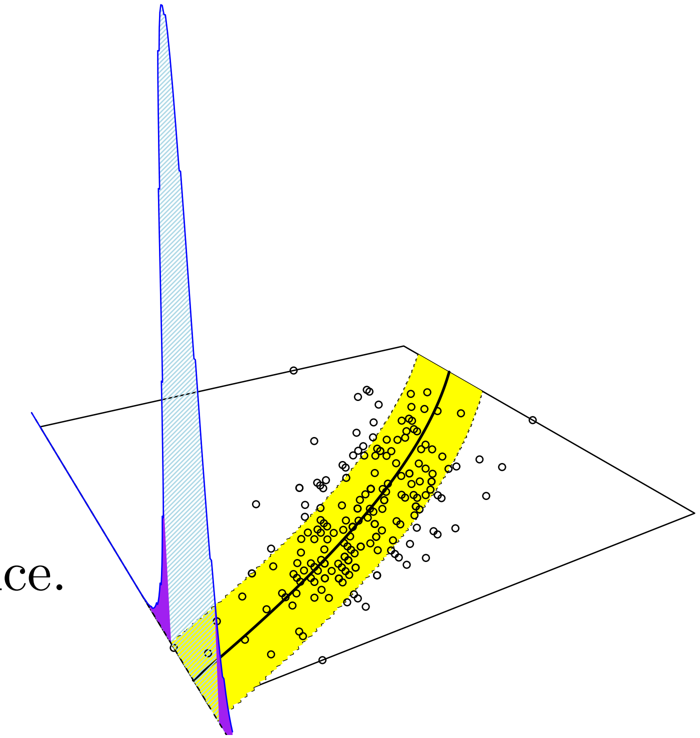
Note that $[\boldsymbol{X}^\mathsf{T} \boldsymbol{W}_k^{-1} \boldsymbol{X}]$ is a $p \times p$ matrix.

Exponential Distributions and Linear Models

Generalized Linear Model:

- $(Y|\boldsymbol{X}=\boldsymbol{x}) \sim \mathcal{L}(\theta_{\boldsymbol{x}}, \varphi)$
- $\mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}] = h^{-1}(\theta_{\boldsymbol{x}}) = g^{-1}(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta})$

e.g. $(Y|\boldsymbol{X}=\boldsymbol{x}) \sim \mathcal{P}(\exp[\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta}])$.



Use of maximum likelihood techniques for inference.

Actually, more a moment condition than a distribution assumption.

Goodness of Fit & Model Choice

From the variance decomposition

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{total variance}}=\underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}_{\text{residual variance}}+\underbrace{\frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i-\bar{y})^2}_{\text{explained variance}}$$

and define

$$R^2=\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2-\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

More generally

$$\text{Deviance}(\boldsymbol{\beta})=-2\log[\mathcal{L}]=2\sum_{i=1}(y_i-\widehat{y}_i)^2=\text{Deviance}(\widehat{\boldsymbol{y}})$$

The null deviance is obtained using $\widehat{y}_i=\overline{y}$, so that

$$R^2=\frac{\text{Deviance}(\overline{y})-\text{Deviance}(\widehat{\boldsymbol{y}})}{\text{Deviance}(\overline{y})}=1-\frac{\text{Deviance}(\widehat{\boldsymbol{y}})}{\text{Deviance}(\overline{y})}=1-\frac{\text{D}}{\text{D}_0}$$

Goodness of Fit & Model Choice

One usually prefers a penalized version

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p} = R^2 - \underbrace{(1 - R^2)\frac{p-1}{n-p}}_{\text{penalty}}$$

See also **Akaike** criteria $AIC = \text{Deviance} + 2 \cdot p$

or **Schwarz**, $BIC = \text{Deviance} + \log(n) \cdot p$

In high dimension, consider a corrected version

$$AICc = \text{Deviance} + 2 \cdot p \cdot \frac{n}{n-p-1}$$

Stepwise Procedures

Forward algorithm

1. set $j_1^\star = \underset{j\in\{\emptyset,1,\cdots,n\}}{\text{argmin}} \{AIC(\{j\})\}$
2. set $j_2^\star = \underset{j\in\{\emptyset,1,\cdots,n\}\backslash\{j_1^\star\}}{\text{argmin}} \{AIC(\{j_1^\star,j\})\}$
3. ... until $j^\star = \emptyset$

Backward algorithm

1. set $j_1^\star = \underset{j\in\{\emptyset,1,\cdots,n\}}{\text{argmin}} \{AIC(\{1,\cdots,n\}\backslash\{j\})\}$
2. set $j_2^\star = \underset{j\in\{\emptyset,1,\cdots,n\}\backslash\{j_1^\star\}}{\text{argmin}} \{AIC(\{1,\cdots,n\}\backslash\{j_1^\star,j\})\}$
3. ... until $j^\star = \emptyset$

Econometrics & Statistical Testing

Standard test for $H_0 : \beta_k = 0$ against $H_1 : \beta_k \neq 0$ is **Student-t** est $t_k = \widehat{\beta}_k/\text{se}_{\widehat{\beta}_k}$,

Use the **p-value** $\mathbb{P}[|T| > |t_k|]$ with $T \sim t_\nu$ (and $\nu = \text{trace}(\boldsymbol{H})$).

In high dimension, consider the FDR (False Discovery Ratio).

With $\alpha = 5\%$, 5% variables are wrongly significant.

If $p = 100$ with only 5 significant variables, one should expect also 5 false positive, i.e. 50% FDR, see Benjamini & Hochberg (1995) and Andrew Gelman's talk.

Under & Over-Identification

Under-identification is obtained when the true model is $y = \beta_0 + \boldsymbol{x}_1^\top\boldsymbol{\beta}_1 + \boldsymbol{x}_2^\top\boldsymbol{\beta}_2 + \varepsilon$, but we estimate $y = \beta_0 + \boldsymbol{x}_1^\top\boldsymbol{b}_1 + \eta$.

Maximum likelihood estimator for $\boldsymbol{b}_1$ is

$$\begin{aligned}
\widehat{\boldsymbol{b}}_1 &= (\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top\boldsymbol{y} \\
&= (\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top[\boldsymbol{X}_{1,i}\boldsymbol{\beta}_1 + \boldsymbol{X}_{2,i}\boldsymbol{\beta}_2 + \varepsilon] \\
&= \boldsymbol{\beta}_1 + \underbrace{(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top\boldsymbol{X}_2\boldsymbol{\beta}_2}_{\boldsymbol{\beta}_{12}} + \underbrace{(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top\varepsilon}_{\nu_i}
\end{aligned}$$

so that $\mathbb{E}[\widehat{\boldsymbol{b}}_1] = \boldsymbol{\beta}_1 + \boldsymbol{\beta}_{12}$, and the bias is null when $\boldsymbol{X}_1^\top\boldsymbol{X}_2 = \boldsymbol{0}$ i.e. $\boldsymbol{X}_1 \perp \boldsymbol{X}_2$, see Frisch-Waugh).

Over-identification is obtained when the true model is $y = \beta_0 + \boldsymbol{x}_1^\top\boldsymbol{\beta}_1\varepsilon$, but we fit $y = \beta_0 + \boldsymbol{x}_1^\top\boldsymbol{b}_1 + \boldsymbol{x}_2^\top\boldsymbol{b}_2 + \eta$.

Inference is unbiased since $\mathbb{E}(\boldsymbol{b}_1) = \boldsymbol{\beta}_1$ but the estimator is not efficient.

Statistical Learning & Loss Function

Here, no probabilistic model, but a **loss function**, ℓ. For some set of functions $\mathcal{M}$, $\mathcal{X} \to \mathcal{Y}$, define

$$m^\star = \underset{m\in\mathcal{M}}{\text{argmin}}\left\{\sum_{i=1}^n \ell(y_i, m(\boldsymbol{x}_i))\right\}$$

Quadratic loss functions are interesting since

$$\overline{y} = \underset{m\in\mathbb{R}}{\text{argmin}}\left\{\sum_{i=1}^n \frac{1}{n}[y_i - m]^2\right\}$$

which can be writen, with some underlying probabilistic model

$$\mathbb{E}(Y) = \underset{m\in\mathbb{R}}{\text{argmin}}\left\{\|Y-m\|_{\ell_2}^2\right\} = \underset{m\in\mathbb{R}}{\text{argmin}}\left\{\mathbb{E}\left([Y-m]^2\right)\right\}$$

For $\tau \in (0,1)$, we obtain the **quantile regression** (see Koenker (2005))

$$m^\star = \underset{m\in\mathcal{M}_0}{\text{argmin}}\left\{\sum_{i=1}^n \ell_\tau(y_i, m(\boldsymbol{x}_i))\right\} \text{ avec } \ell_\tau(x,y) = |(x-y)(\tau - \mathbf{1}_{x\leq y})|$$

Boosting & Weak Learning

$$m^\star = \underset{m\in\mathcal{M}}{\text{argmin}}\left\{\sum_{i=1}^n \ell(y_i, m(\boldsymbol{x}_i))\right\}$$

is hard to solve for some very large and general space $\mathcal{M}$ of $\mathcal{X}\to\mathcal{Y}$ functions.

Consider some iterative procedure, where we learn from the errors,

$$m^{(k)}(\cdot) = \underbrace{m_1(\cdot)}_{\sim\boldsymbol{y}} + \underbrace{m_2(\cdot)}_{\sim\boldsymbol{\varepsilon}_1} + \underbrace{m_3(\cdot)}_{\sim\boldsymbol{\varepsilon}_2} + \cdots + \underbrace{m_k(\cdot)}_{\sim\boldsymbol{\varepsilon}_{k-1}} = m^{(k-1)}(\cdot) + m_k(\cdot).$$

Formely $\boldsymbol{\varepsilon}$ can be seen as $\nabla\ell$, the gradient of the loss.

Boosting & Weak Learning

It is possible to see this algorithm as a gradient descent. Not

$$\underbrace{f(\boldsymbol{x}_k)}_{\langle f,\boldsymbol{x}_k\rangle} \sim \underbrace{f(\boldsymbol{x}_{k-1})}_{\langle f,\boldsymbol{x}_{k-1}\rangle} + \underbrace{(\boldsymbol{x}_k - \boldsymbol{x}_{k-1})}_{\alpha_k} \underbrace{\nabla f(\boldsymbol{x}_{k-1})}_{\langle \nabla f,\boldsymbol{x}_{k-1}\rangle}$$

but some kind of dual version

$$\underbrace{f_k(\boldsymbol{x})}_{\langle f_k,\boldsymbol{x}\rangle} \sim \underbrace{f_{k-1}(\boldsymbol{x})}_{\langle f_{k-1},\boldsymbol{x}\rangle} + \underbrace{(f_k - f_{k-1})}_{a_k} \underbrace{\star}_{\langle f_{k-1},\nabla \boldsymbol{x}\rangle}$$

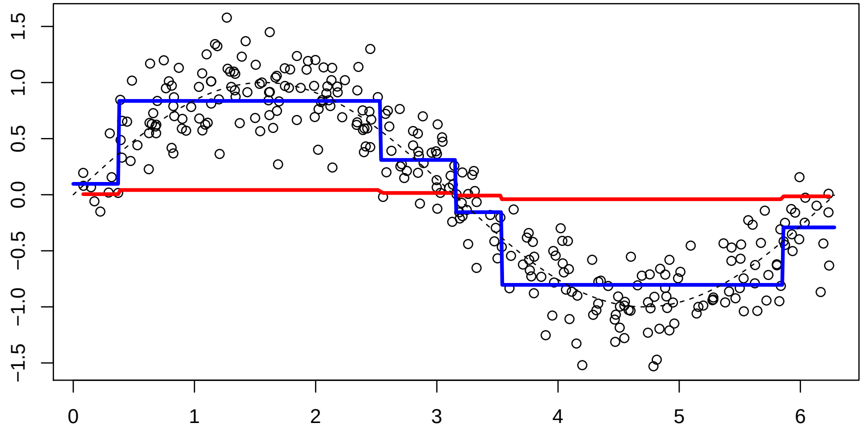
where $\star$ is a gradient is some functional space.

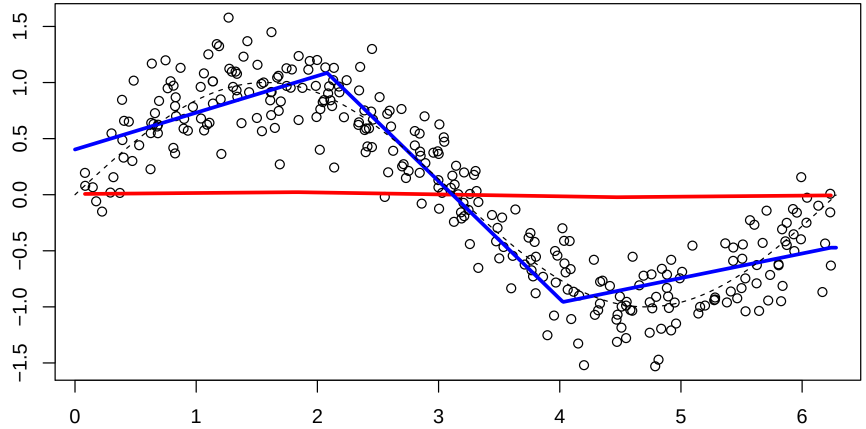
$$m^{(k)}(\boldsymbol{x}) = m^{(k-1)}(\boldsymbol{x}) + \underset{f\in\mathcal{F}}{\text{argmin}} \left\{ \sum_{i=1}^{n} \ell(y_i, m^{(k-1)}(\boldsymbol{x}) + f(\boldsymbol{x})) \right\}$$

for some simple space $\mathcal{F}$ so that we define some **weak learner**, e.g. step functions (so called stumps)

Boosting & Weak Learning

Standard set $\mathcal{F}$ are stumps functions but one can also consider splines (with non-fixed knots).





One might add a **shrinkage** parameter to learn even more weakly, i.e. set $\varepsilon_1 = y - \alpha \cdot m_1(\boldsymbol{x})$ with $\alpha \in (0, 1)$, etc.

Big Data & Linear Model

Consider some linear model $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$ for all $i = 1, \cdots, n$.

Assume that ε_i are i.i.d. with $\mathbb{E}(\varepsilon) = 0$ (and finite variance). Write

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{\boldsymbol{y}, n\times 1} = \underbrace{\begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{pmatrix}}_{\boldsymbol{X}, n\times(p+1)} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}}_{\boldsymbol{\beta}, (p+1)\times 1} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}, n\times 1}.$$

Assuming $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\mathbb{I})$, the maximum likelihood estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = \text{argmin}\{\|\boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{\beta}\|_{\ell_2}\} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

... under the assumtption that $\boldsymbol{X}^\top \boldsymbol{X}$ is a full-rank matrix.

What if $\boldsymbol{X}^\top \boldsymbol{X}$ cannot be inverted? Then $\widehat{\boldsymbol{\beta}} = [\boldsymbol{X}^\top \boldsymbol{X}]^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ does not exist, but $\widehat{\boldsymbol{\beta}}_\lambda = [\boldsymbol{X}^\top \boldsymbol{X} + \lambda \mathbb{I}]^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ always exist if $\lambda > 0$.

Ridge Regression & Regularization

The estimator $\widehat{\boldsymbol{\beta}}=[\boldsymbol{X}^\mathsf{T}\boldsymbol{X}+\lambda\mathbb{I}]^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}$ is the **Ridge** estimate obtained as solution of

$$\widehat{\boldsymbol{\beta}}=\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left\{\sum_{i=1}^n[y_i-\beta_0-\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}]^2+\lambda\underbrace{\|\boldsymbol{\beta}\|_{\ell_2}^2}_{\mathbf{1}^\mathsf{T}\boldsymbol{\beta}^2}\right\}$$

for some tuning parameter λ. One can also write

$$\widehat{\boldsymbol{\beta}}=\underset{\boldsymbol{\beta};\|\boldsymbol{\beta}\|_{\ell_2}\le s}{\operatorname{argmin}}\{\|\boldsymbol{Y}-\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}\|_{\ell_2}\}$$

There is a Bayesian interpretation of that regularization, when $\boldsymbol{\beta}$ has some prior $\mathcal{N}(\boldsymbol{\beta}_0,\tau\mathbb{I})$.

Over-Fitting & Penalization

Solve here, for some norm $\|\cdot\|$,

$$\min\left\{\sum_{i=1}^{n}\ell(y_i,\beta_0+\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta})+\lambda\|\boldsymbol{\beta}\|\right\}=\min\Big\{\text{objective}(\boldsymbol{\beta})+\text{penality}(\boldsymbol{\beta})\Big\}.$$

Estimators are **no longer unbiased**, but might have a smaller mse.

Consider some i.id. sample $\{y_1,\cdots,y_n\}$ from $\mathcal{N}(\theta,\sigma^2)$, and consider some estimator proportional to $\overline{y}$, i.e. $\widehat{\theta}=\alpha\overline{y}$. $\alpha=1$ is the maximum likelihood estimator.

Note that

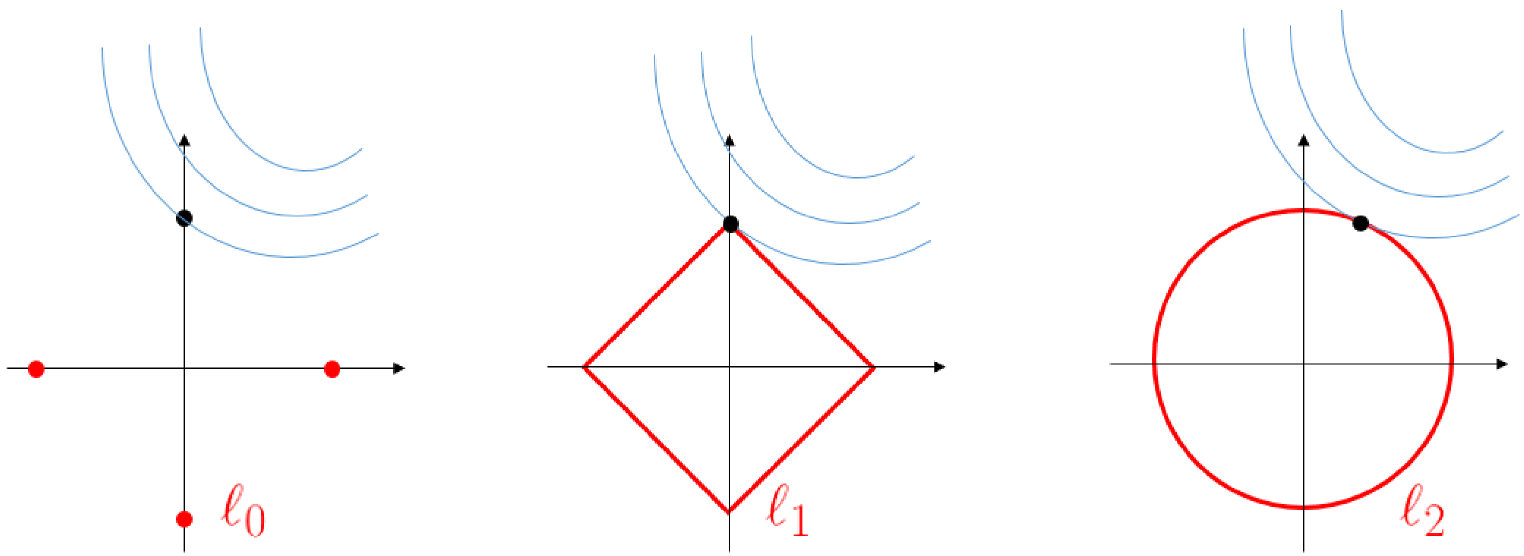
$$\text{mse}[\widehat{\theta}]=\underbrace{(\alpha-1)^2\mu^2}_{\text{bias}[\widehat{\theta}]^2}+\underbrace{\frac{\alpha^2\sigma^2}{n}}_{\text{Var}[\widehat{\theta}]}$$

and $\alpha^\star=\mu^2\cdot\left(\mu^2+\dfrac{\sigma^2}{n}\right)^{-1}<1.$

$$(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) = \text{argmin}\left\{\sum_{i=1}^n \ell(y_i, \beta_0 + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|\right\},$$

can be seen as a **Lagrangian** minimization problem

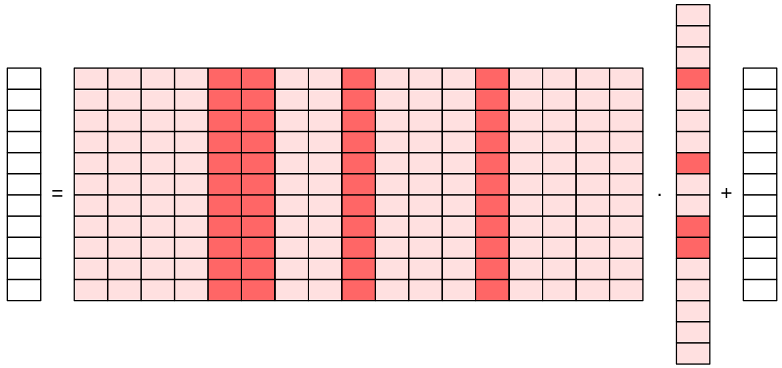
$$(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) = \underset{\boldsymbol{\beta};\|\boldsymbol{\beta}\|\leq s}{\text{argmin}}\left\{\sum_{i=1}^n \ell(y_i, \beta_0 + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta})\right\}$$



LASSO & Sparcity

In severall applications, p can be (very) large, but a lot of features are just noise: $\beta_j = 0$ for many j's. Let s denote the number of **relevent features**, with $s << p$, cf Hastie, Tibshirani & Wainwright (2015),

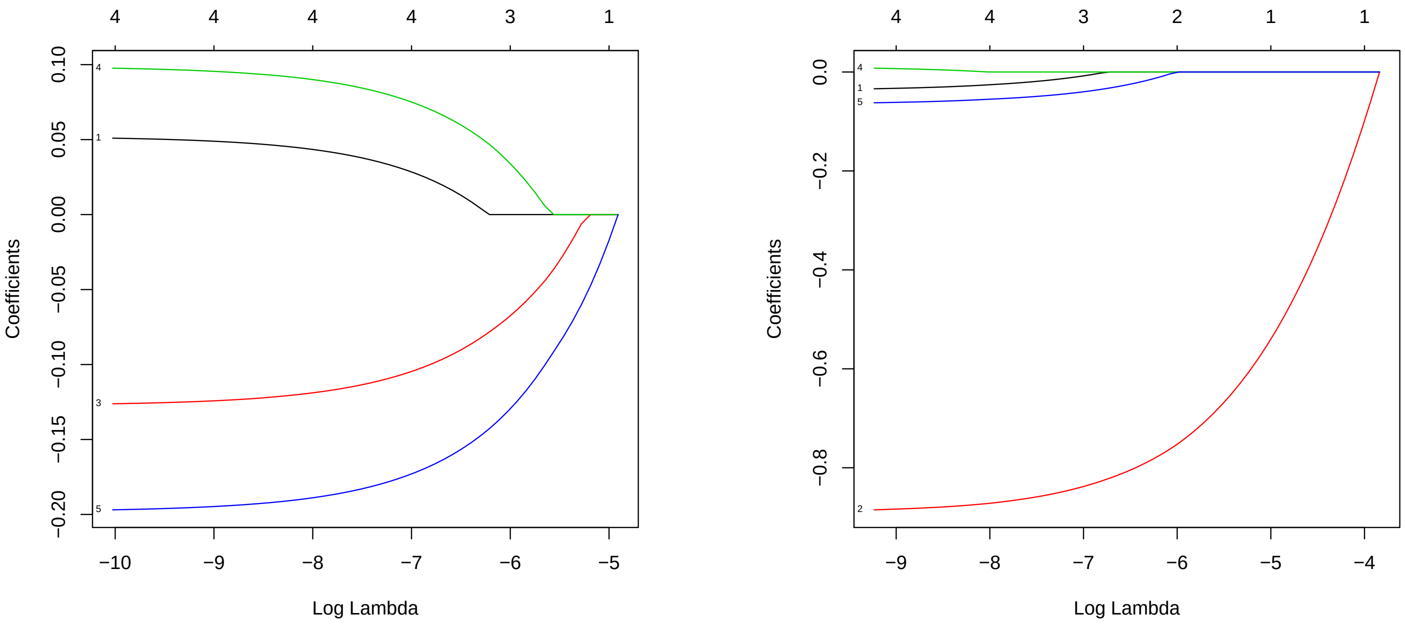
$$s = \text{card}\{\mathcal{S}\} \text{ where } \mathcal{S} = \{j; \beta_j \neq 0\}$$



The true model is now $y = \boldsymbol{X}_{\mathcal{S}}^{\top}\boldsymbol{\beta}_{\mathcal{S}} + \varepsilon$, where $\boldsymbol{X}_{\mathcal{S}}^{\top}\boldsymbol{X}_{\mathcal{S}}$ is a full rank matrix.

LASSO & Sparcity

Evolution of $\widehat{\boldsymbol{\beta}}_\lambda$ as a function of $\log\lambda$ in various applications



In-Sample & Out-Sample

Write $\widehat{\boldsymbol{\beta}}=\widehat{\boldsymbol{\beta}}((\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_n,y_n))$. Then (for the linear model)

$$\text{Deviance}_{\mathsf{IS}}(\widehat{\boldsymbol{\beta}})=\sum_{i=1}^{n}[y_i-\boldsymbol{x}_i^{\mathsf{T}}\widehat{\boldsymbol{\beta}}((\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_n,y_n))]^2$$

Withe this "in-sample" deviance, we cannot use the central limit theorem

$$\frac{\text{Deviance}_{\mathsf{IS}}(\widehat{\boldsymbol{\beta}})}{n}\not\to\mathbb{E}\left([Y-\boldsymbol{X}^{\mathsf{T}}\beta]\right)$$

Hence, we can compute some "out-of-sample" deviance

$$\text{Deviance}_{\mathsf{OS}}(\widehat{\boldsymbol{\beta}})=\sum_{i=n+1}^{m+n}[y_i-\boldsymbol{x}_i^{\mathsf{T}}\widehat{\boldsymbol{\beta}}((\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_n,y_n)]^2$$

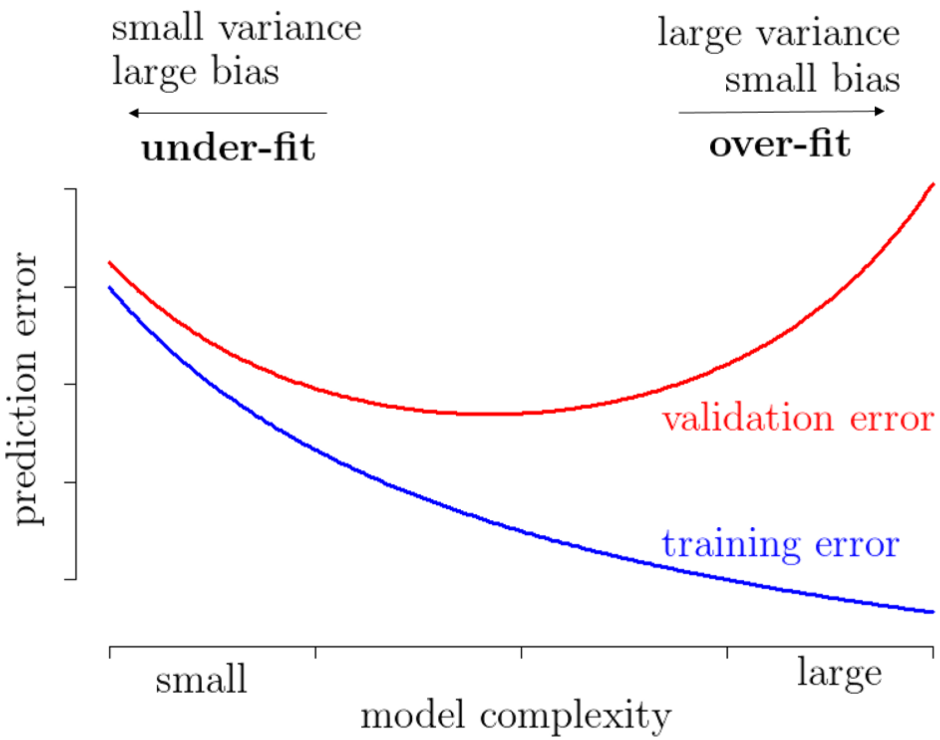
In-Sample & Out-Sample

Observe that there are connexions with Akaike penaly function

$$\text{Deviance}_{\,\mathsf{IS}}(\widehat{\beta}) - \text{Deviance}_{\,\mathsf{OS}}(\widehat{\beta}) \approx 2 \cdot \text{degrees of freedom}$$

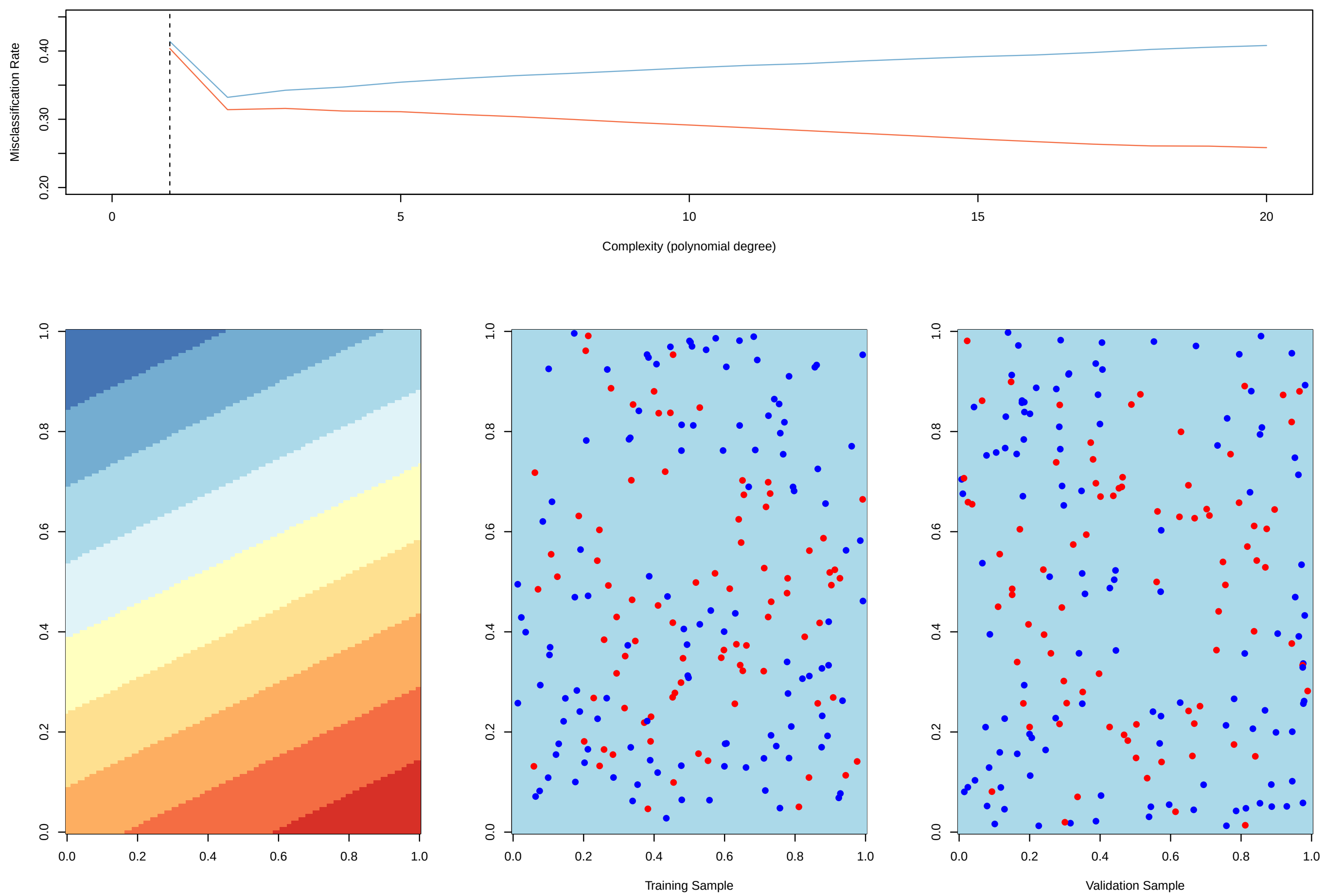
From Stone (1977), minimizing AIC is closed to cross validation,

From Shao (1997) minimizing BIC is closed to k-fold cross validation with $k = n/\log n$.



Overfit, Generalization & Model Complexity

Complexity of the model is the degree of the polynomial function



Cross-Validation

See Jacknife technique Quenouille (1956) or Tukey (1958) to reduce the bias.

If $\{y_1, \cdots, y_n\}$ is an i.id. sample from F_θ, with estimator $T_n(\boldsymbol{y}) = T_n(y_1, \cdots, y_n)$, such that $\mathbb{E}[T_n(\boldsymbol{Y})] = \theta + O\left(n^{-1}\right)$, consider

$$\widetilde{T}_n(\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^n T_{n-1}(\boldsymbol{y}_{(i)}) \text{ avec } \boldsymbol{y}_{(i)} = (y_1, \cdots, y_{i-1}, y_{i+1}, \cdots, y_n).$$

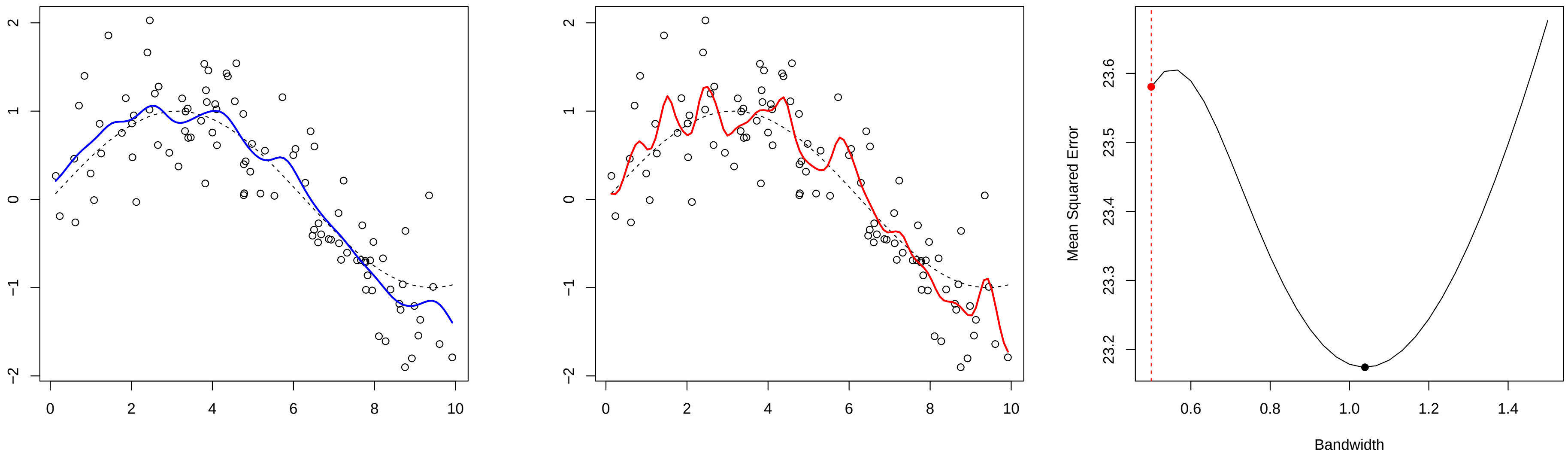
Then $\mathbb{E}[\widetilde{T}_n(\boldsymbol{Y})] = \theta + O\left(n^{-2}\right)$.

Similar idea in **leave-one-out cross validation**

$$\text{Risk} = \frac{1}{n}\sum_{i=1}^n \ell(y_i, \widehat{m}_{(i)}(\boldsymbol{x}_i))$$

Rule of Thumb vs. Cross Validation

$$\widehat{m}^{[h^\star]}(x) = \widehat{\beta}_0^{[x]} + \widehat{\beta}_1^{[x]}x \text{ with } (\widehat{\beta}_0^{[x]}, \widehat{\beta}_1^{[x]}) = \underset{(\beta_0,\beta_1)}{\text{argmin}} \left\{ \sum_{i=1}^{n} \omega_{h^\star}^{[x]} [y_i - (\beta_0 + \beta_1 x_i)]^2 \right\}$$



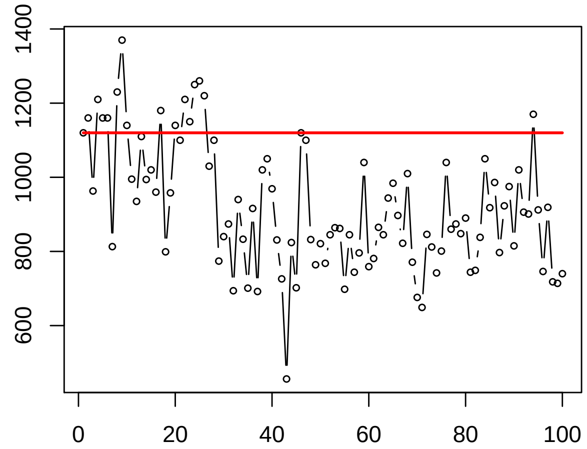
$$\text{set } h^\star = \text{argmin}\{\text{mse}(h)\} \text{ with } \text{mse}(h) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \widehat{m}_{(i)}^{[h]}(x_i)\right]^2$$

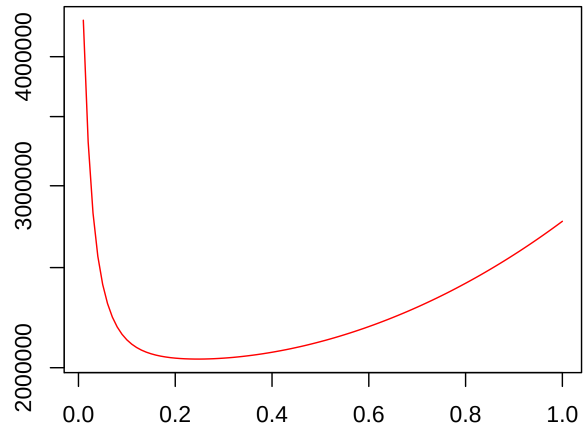
Exponential Smoothing for Time Series

Consider some exponential smoothing filter, on a time series (x_t), $\widehat{y}_{t+1} = \alpha\widehat{y}_t + (1-\alpha)y_t$, then consider

$$\alpha^\star = \text{argmin}\left\{\sum_{t=2}^{T} \ell(\widehat{y}_t, y_t)\right\},$$

see Hyndman *et al.* (2003).

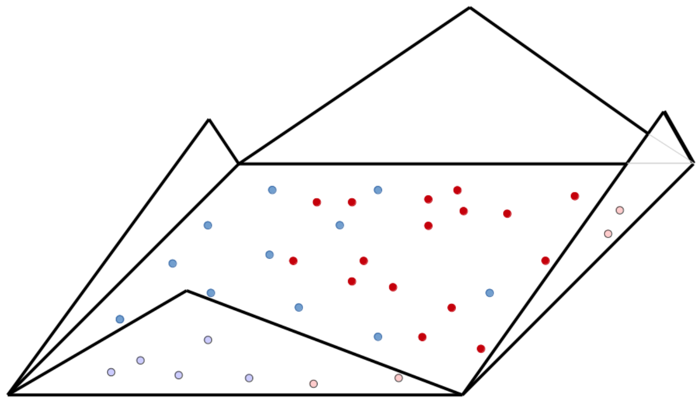




Cross-Validation

Consider a partition of $\{1,\cdots,n\}$ in k groups with the same size, $\mathcal{I}_1,\cdots,\mathcal{I}_k$, and set $\mathcal{I}_{\bar{j}}=\{1,\cdots,n\}\backslash\mathcal{I}_j$. Fit $\widehat{m}_{(j)}$ on $\mathcal{I}_{\bar{j}}$, and

$$\text{Risk}=\frac{1}{k}\sum_{j=1}^{k}\text{Risk}_j \text{ where } \text{Risk}_j=\frac{k}{n}\sum_{i\in\mathcal{I}_j}\ell(y_i,\widehat{m}_{(j)}(\boldsymbol{x}_i))$$



Randomization is too important to be left to chance!

Consider some **bootstraped** sample, $\mathcal{I}_b = \{i_{1,b}, \cdots, i_{n,b}\}$, with $i_{k,b} \in \{1, \cdots, n\}$

Set $n_i = \mathbf{1}_{i \notin \mathcal{I}_1} + \cdots + \mathbf{1}_{i \notin v_B}$, and fit $\widehat{m}_b$ on $\mathcal{I}_b$

$$\text{Risk} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{b: i \notin I_b} \ell(y_i, \widehat{m}_b(\boldsymbol{x}_i))$$

Probability that ith obs. is not selection $(1 - n^{-1})^n \to e^{-1} \sim 36.8\%$,

see training / validation samples (2/3-1/3).

Bootstrap

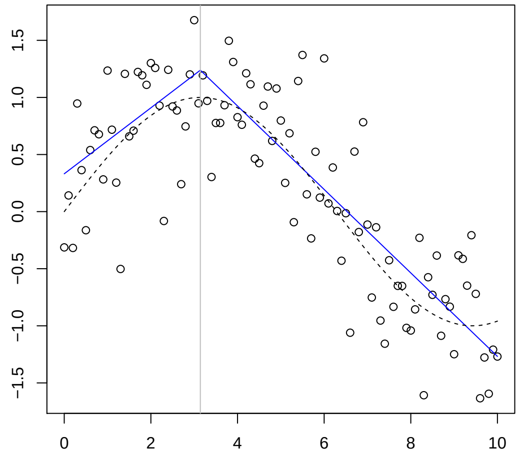
From Efron (1987), generate samples from $(\Omega, \mathcal{F}, \mathbb{P}_n)$

$$\widehat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(y_i \leq y) \text{ and } \widehat{F}_n(y_i) = \frac{\text{rank}(y_i)}{n}.$$

If $U \sim \mathcal{U}([0,1])$, $F^{-1}(U) \sim F$

If $U \sim \mathcal{U}([0,1])$, $\widehat{F}_n^{-1}(U)$ is uniform

on $\left\{\frac{1}{n}, \cdots, \frac{n-1}{n}, 1\right\}$.



Consider some **boostraped sample**,

- either $(y_{i_k}, \boldsymbol{x}_{i_k})$, $i_k \in \{1, \cdots, n\}$
- or $(\widehat{y}_k + \widehat{\varepsilon}_{i_k}, \boldsymbol{x}_k)$, $i_k \in \{1, \cdots, n\}$

Classification & Logistic Regression

Generalized Linear Model when Y has a **Bernoulli distribution**, $y_i \in \{0,1\}$,

$$m(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}] = \frac{e^{\beta_0+\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}}}{1+e^{\beta_0+\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}}} = H(\beta_0+\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta})$$

Estimate $(\beta_0,\boldsymbol{\beta})$ using maximum likelihood techniques

$$\mathcal{L} = \prod_{i=1}^n \left(\frac{e^{\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}}}\right)^{y_i}\left(\frac{1}{1+e^{\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}}}\right)^{1-y_i}$$

$$\text{Deviance} \propto \sum_{i=1}^n \left[\log(1+e^{\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}}) - y_i\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}\right]$$

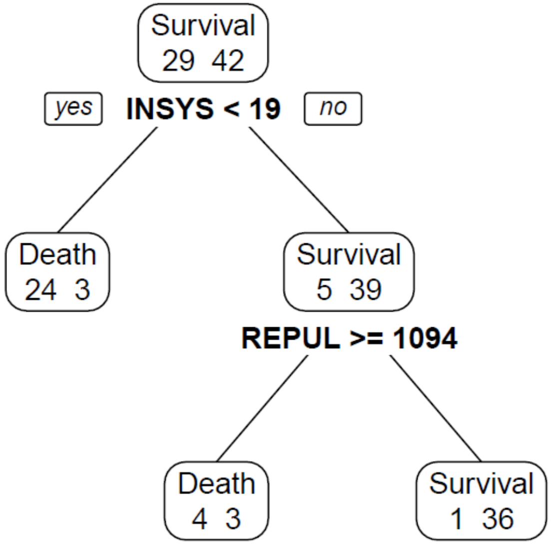
Observe that

$$\mathrm{D}_0 \propto \sum_{i=1}^n [y_i\log(\overline{y}) + (1-y_i)\log(1-\overline{y})]$$

Classification Trees

To split $\{N\}$ into two $\{N_L, N_R\}$, consider

$$\mathcal{I}(N_L, N_R) = \sum_{x\in\{L,R\}} \frac{n_x}{n}\mathcal{I}(N_x)$$



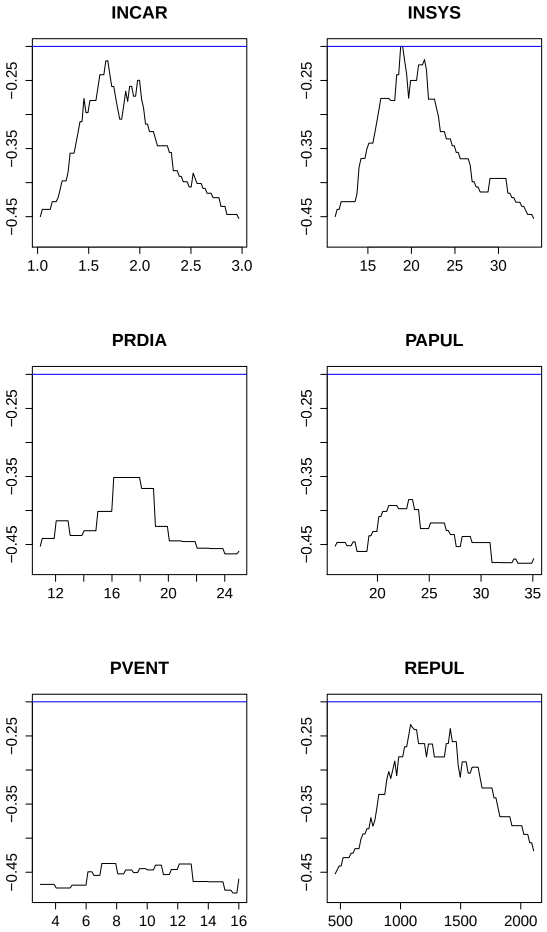
e.g. **Gini index** (used originally in CART, see Breiman *et al.* (1984))

$$\text{gini}(N_L, N_R) = -\sum_{x\in\{L,R\}} \frac{n_x}{n} \sum_{y\in\{0,1\}} \frac{n_{x,y}}{n_x}\left(1 - \frac{n_{x,y}}{n_x}\right)$$

and the **cross-entropy** (used in C4.5 and C5.0)

$$\text{entropy}(N_L, N_R) = -\sum_{x\in\{L,R\}} \frac{n_x}{n} \sum_{y\in\{0,1\}} \frac{n_{x,y}}{n_x}\log\left(\frac{n_{x,y}}{n_x}\right)$$

Classification Trees

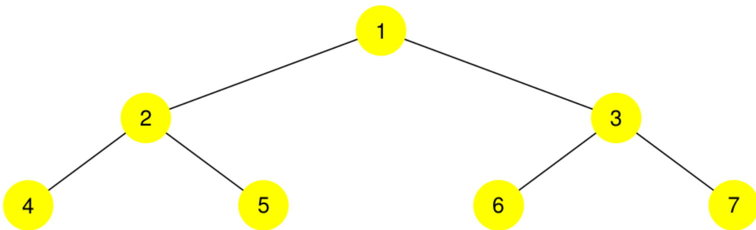


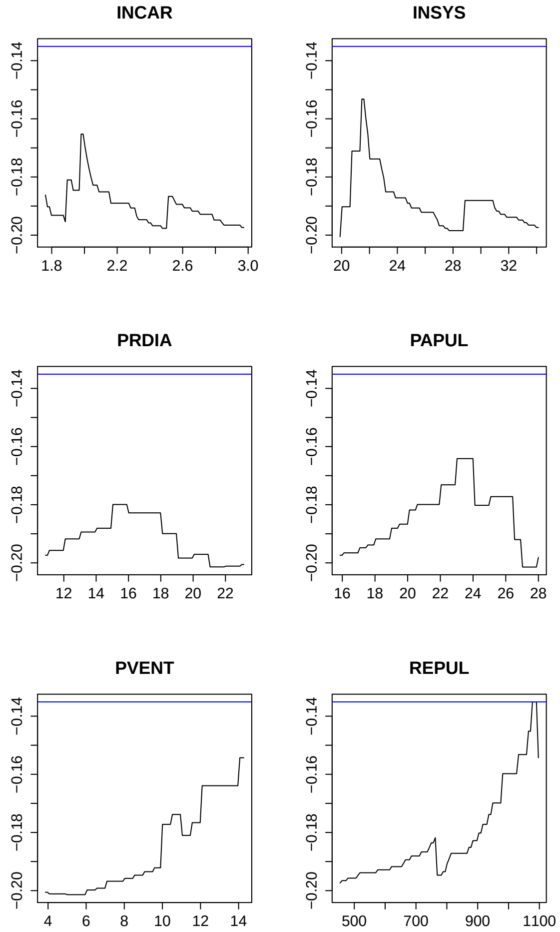
$N_L\colon \{x_{i,j} \le s\}$ $N_R\colon \{x_{i,j} > s\}$

solve $\displaystyle\max_{j\in\{1,\cdots,k\},s} \{\mathcal{I}(N_L,N_R)\}$

$\longleftarrow$ first split

second split $\longrightarrow$





Trees & Forests

Boostrap can be used to define the concept of **margin**,

$$\text{margin}_i = \frac{1}{B}\sum_{b=1}^{B}\mathbf{1}(\widehat{y}_i^{(b)} = y_i) - \frac{1}{B}\sum_{b=1}^{B}\mathbf{1}(\widehat{y}_i^{(b)} \neq y_i)$$

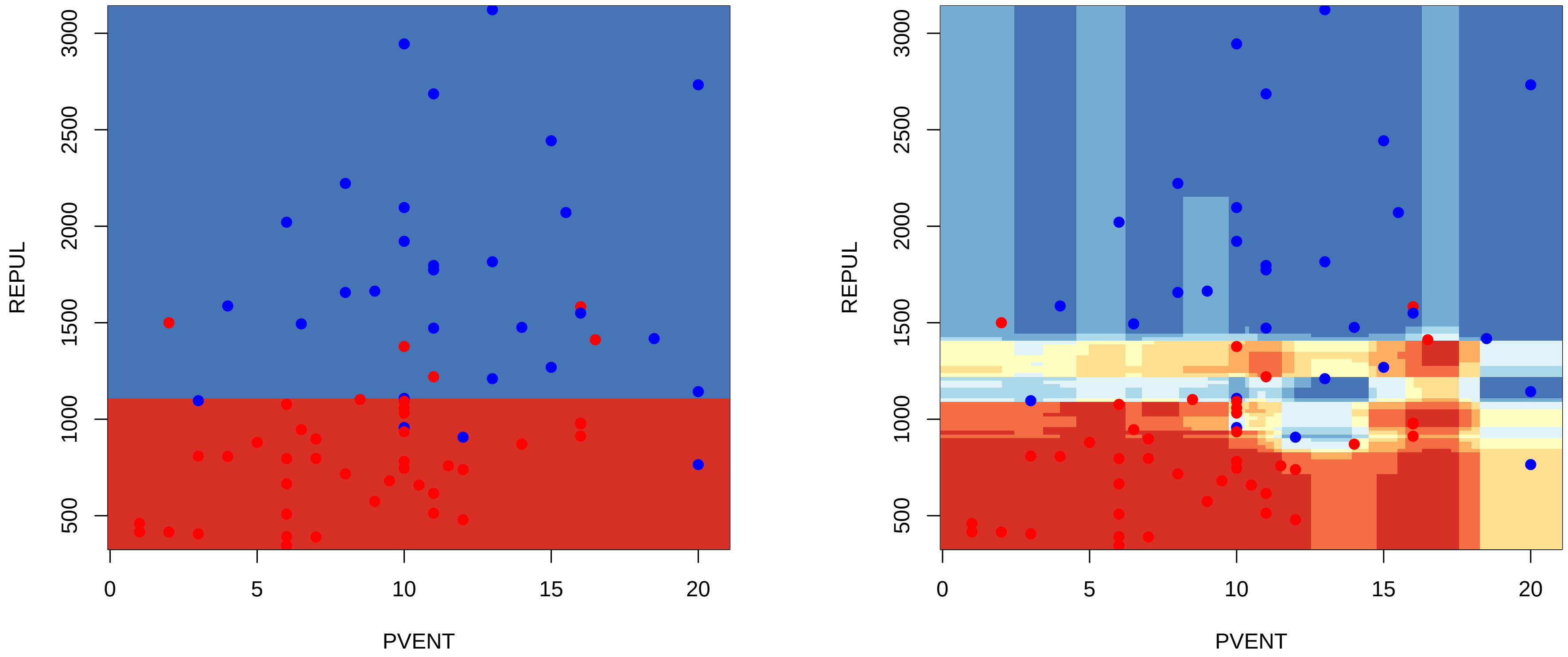
Subsampling of variable, at each knot (e.g. $\sqrt{k}$ out of k)

Concept of **variable importance**: given some random forest with M trees,

$$\text{importance of variable } k \quad I(X_k) = \frac{1}{M}\sum_{m}\sum_{t}\frac{N_t}{N}\Delta\mathcal{I}(t)$$

where the first sum is over all trees, and the second one is over all nodes where the split is done based on variable X_k.

Trees & Forests



See also discriminant analysis, SVM, neural networks, etc.

Model Selection & ROC Curves

Given a scoring function $m(\cdot)$, with $m(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$, and a threshold $s \in (0,1)$, set

$$\widehat{Y}^{(s)} = \mathbf{1}[m(\boldsymbol{x}) > s] = \begin{cases} 1 \text{ if } m(\boldsymbol{x}) > s \\ 0 \text{ if } m(\boldsymbol{x}) \leq s \end{cases}$$

Define the confusion matrix as $\boldsymbol{N} = [N_{u,v}]$

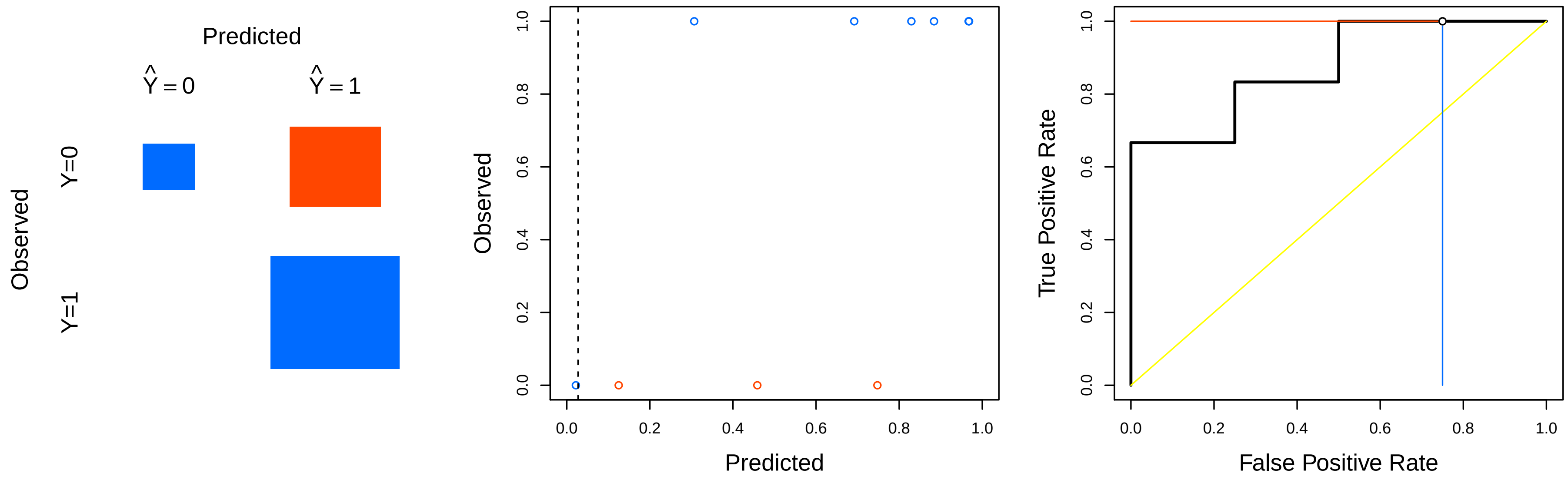
$$N^{(s)}_{u,v} = \sum_{i=1}^{n} \mathbf{1}(\widehat{y}^{(s)}_i = u, y_j = v) \text{ for } (u,v) \in \{0,1\}.$$

	$Y = 0$	$Y = 1$	
$\widehat{Y}_s = 0$	TN_s	FN_s	TN_s+FN_s
$\widehat{Y}_s = 1$	FP_s	TP_s	FP_s+TP_s
	TN_s+FP_s	FN_s+TP_s	n

Model Selection & ROC Curves

ROC curve is

$$\text{ROC}_s = \left(\frac{\text{FP}_s}{\text{FP}_s + \text{TN}_s}, \frac{\text{TP}_s}{\text{TP}_s + \text{FN}_s}\right) \text{ with } s \in (0,1)$$



Model Selection & ROC Curves

In machine learning, the most popular measure is κ, see Landis & Koch (1977). Define $\boldsymbol{N}^{\perp}$ from $\boldsymbol{N}$ as in the chi-square independence test. Set

$$\text{total accuracy} = \frac{\text{TP} + \text{TN}}{n}$$

$$\text{random accuracy} = \frac{\text{TP}^{\perp} + \text{TN}^{\perp}}{n} = \frac{[\text{TN+FP}] \cdot [\text{TP+FN}] + [\text{TP+FP}] \cdot [\text{TN+FN}]}{n^2}$$

and

$$\kappa = \frac{\text{total accuracy} - \text{random accuracy}}{1 - \text{random accuracy}}.$$

See Kaggle competitions.

Reducing Dimension with PCA

Use **principal components** to reduce dimension (on centered and scaled variables): we want d vectors $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_d$ such that

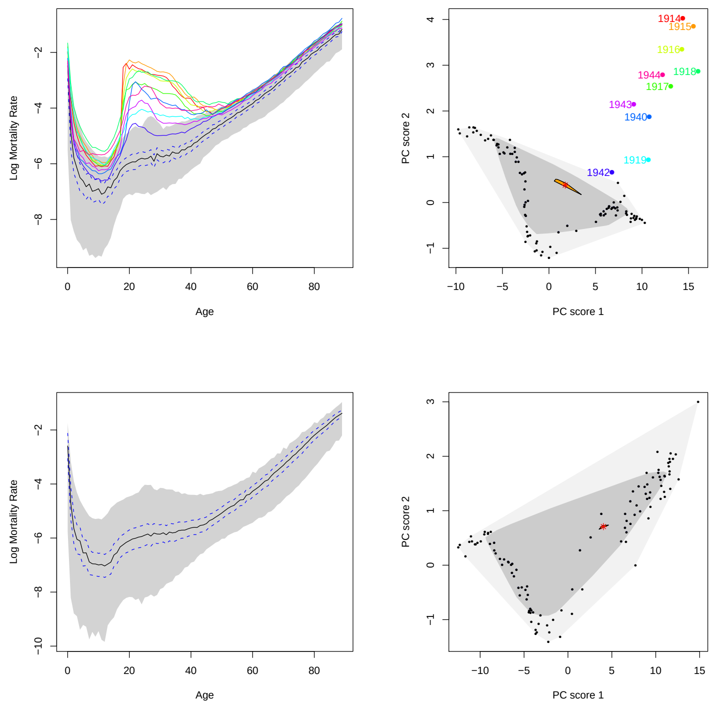
First Compoment is $\boldsymbol{z}_1 = \boldsymbol{X}\boldsymbol{\omega}_1$ where

$$\boldsymbol{\omega}_1 = \underset{\|\boldsymbol{\omega}\|=1}{\text{argmax}} \left\{ \|\boldsymbol{X} \cdot \boldsymbol{\omega}\|^2 \right\} = \underset{\|\boldsymbol{\omega}\|=1}{\text{argmax}} \left\{ \boldsymbol{\omega}^\mathsf{T} \boldsymbol{X}^\mathsf{T} \boldsymbol{X} \boldsymbol{\omega} \right\}$$

Second Compoment is $\boldsymbol{z}_2 = \boldsymbol{X}\boldsymbol{\omega}_2$ where

$$\boldsymbol{\omega}_2 = \underset{\|\boldsymbol{\omega}\|=1}{\text{argmax}} \left\{ \|\widetilde{\boldsymbol{X}}^{(1)} \cdot \boldsymbol{\omega}\|^2 \right\}$$

with $\widetilde{\boldsymbol{X}}^{(1)} = \boldsymbol{X} - \underbrace{\boldsymbol{X}\boldsymbol{\omega}_1}_{\boldsymbol{z}_1} \boldsymbol{\omega}_1^\mathsf{T}$.



Reducing Dimension with PCA

A regression on (the d) principal components, $y = \boldsymbol{z}^\top \boldsymbol{b} + \boldsymbol{\eta}$ could be an interesting idea, unfortunatley, principal components have no reason to be correlated with y. First compoment was $\boldsymbol{z}_1 = \boldsymbol{X}\boldsymbol{\omega}_1$ where

$$\boldsymbol{\omega}_1 = \underset{\|\boldsymbol{\omega}\|=1}{\text{argmax}} \left\{ \|\boldsymbol{X} \cdot \boldsymbol{\omega}\|^2 \right\} = \underset{\|\boldsymbol{\omega}\|=1}{\text{argmax}} \left\{ \boldsymbol{\omega}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{\omega} \right\}$$

It is a non-supervised technique.

Instead, use **partial least squares**, introduced in Wold (1966). First compoment is $\boldsymbol{z}_1 = \boldsymbol{X}\boldsymbol{\omega}_1$ where

$$\boldsymbol{\omega}_1 = \underset{\|\boldsymbol{\omega}\|=1}{\text{argmax}} \left\{ < \boldsymbol{y}, \boldsymbol{X} \cdot \boldsymbol{\omega} > \right\} = \underset{\|\boldsymbol{\omega}\|=1}{\text{argmax}} \left\{ \boldsymbol{\omega}^\top \boldsymbol{X}^\top \boldsymbol{y} \boldsymbol{y}^\top \boldsymbol{X} \boldsymbol{\omega} \right\}$$

(etc.)

Instrumental Variables

Consider some instrumental variable model, $y_i = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon_i$ such that

$$\mathbb{E}[Y_i|\boldsymbol{Z}] = \mathbb{E}[\boldsymbol{X}_i|\boldsymbol{Z}]^{\mathsf{T}}\beta + \mathbb{E}[\varepsilon_i|\boldsymbol{Z}]$$

The estimator of $\boldsymbol{\beta}$ is

$$\widehat{\beta}_{\mathsf{IV}} = [\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{X}]^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{y}$$

If $\dim(\boldsymbol{Z}) > \dim(\boldsymbol{X})$ use the Generalized Method of Moments,

$$\widehat{\beta}_{\mathsf{GMM}} = [\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\Pi}_Z\boldsymbol{X}]^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\Pi}_Z\boldsymbol{y} \text{ with } \boldsymbol{\Pi}_Z = \boldsymbol{Z}[\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}]^{-1}\boldsymbol{Z}^{\mathsf{T}}$$

Instrumental Variables

Consider a standard two step procedure

1) regress colums of $\boldsymbol{X}$ on $\boldsymbol{Z}$, $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{\eta}$, and derive predictions $\widehat{\boldsymbol{X}} = \boldsymbol{\Pi}_{\boldsymbol{Z}}\boldsymbol{X}$

2) regress Y on $\widehat{\boldsymbol{X}}$, $y_i = \widehat{\boldsymbol{x}}_i^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon_i$, i.e.

$$\widehat{\beta}_{\mathsf{IV}} = [\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{X}]^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{y}$$

See Angrist & Krueger (1991) with 3 up to 1530 instruments : 12 instruments seem to contain all necessary information.

Use LASSO to select necessary instruments, see Belloni, Chernozhukov & Hansen (2010)

Take Away Conclusion

Big data mythology

- $n \to \infty$: 0/1 law, everything is simplified (either true or false)
- $p \to \infty$: higher algorithmic complexity, need variable selection tools

Econometrics vs. Machine Learning

- probabilistic interpretation of econometric models
  (unfortunately sometimes misleading, e.g. p-value)
  can deal with non-i.id data (time series, panel, etc)
- machine learning is about predictive modeling and generalization
  algorithmic tools, based on bootstrap (sampling and sub-sampling),
  cross-validation, variable selection, nonlinearities, cross effects, etc

Importance of visualization techniques (forgotten in econometrics publications)