Econometrics & Machine Learning

A. Charpentier, E. Flachaire & A. Ly

https://arxiv.org/abs/1708.06992

Probabilistic Foundations of Econometrics

THE PROBABILITY APPROACH
IN ECONOMETRICS

By
Trygve Haavelmo
RESEARCH ASSOCIATE
COWLES COMMISSION FOR
RESEARCH IN ECONOMICS

Supplement to Econometrica, Volume 12, July, 1944

THE ECONOMETRIC SOCIETY
The University of Chicago
Chicago 37, Illinois

see Haavelmo (1944)

THE FORMATION OF ECONOMETRICS

A Historical Perspective

QIN DUO
秦朵

see Duo (1997)

Probabilistic Foundations of Econometrics

There is a probabilistic space $(\Omega, \mathcal{F}, \mathbb{P})$ such that data $\{y_i, \boldsymbol{x}_i\}$ can be seen as realizations of random variables $\{Y_i, \boldsymbol{X}_i\}$.

Consider a conditional distribution for $Y|\boldsymbol{X}$, e.g.

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

for the linear model, with some extension when it is in the exponential family (GLM).

Then use maximum likelihood for inference, to derive confidence interval (quantification of uncertainty), etc.

Importance of unbiased estimators (Cramer Rao lower bound for the variance).

Loss Functions

Gneiting in a statistical context: a statistics T is said to be ellicitable if there is $\ell : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_+$ such that

$$T(Y) = \underset{x \in \mathbb{R}}{\text{argmin}} \left\{ \int_{\mathbb{R}} l(x,y) dF(y) \right\} = \underset{x \in \mathbb{R}}{\text{argmin}} \left\{ \mathbb{E}\big[l(x,Y) \big] \text{ where } Y \overset{\mathcal{L}}{\sim} F \right\}$$

(e.g. $T(\boldsymbol{x}) = \overline{x}$ and $\ell = \ell_2$, or $T(\boldsymbol{x}) = \text{median}(\overline{x})$ and $\ell = \ell_1$).

In machine-learning, we want to solve

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

using optimization techniques, in a not-too-complex space $\mathcal{M}$.

Overfit ?

Underfit: true $y_i = \beta_0 + \boldsymbol{x}_1^\top \boldsymbol{\beta}_1 + \boldsymbol{x}_2^\top \boldsymbol{\beta}_2 + \varepsilon_i$ vs. fitted model $y_i = b_0 + \boldsymbol{x}_1^\top \boldsymbol{b}_1 + \eta_i$.

$$\widehat{\boldsymbol{b}}_1 = (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{y} = \boldsymbol{\beta}_1 + \underbrace{(\boldsymbol{X}_1' \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_2 \boldsymbol{\beta}_2}_{\boldsymbol{\beta}_{12}} + \underbrace{(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \varepsilon}_{\nu_i}$$

i.e. $\mathbb{E}[\widehat{\boldsymbol{b}}_1] = \boldsymbol{\beta}_1 + \boldsymbol{\beta}_{12} \neq \boldsymbol{\beta}_1$, unless $\boldsymbol{X}_1 \perp \boldsymbol{X}_2$ (see Frish-Waugh Theorem).

Overfit: true $y_i = \beta_0 + \boldsymbol{x}_1^\top \boldsymbol{\beta}_1 + \varepsilon_i$ vs. fitted model $y_i = b_0 + \boldsymbol{x}_1^\top \boldsymbol{b}_1 + \boldsymbol{x}_2^\top \boldsymbol{b}_2 + \eta_i$.

In that case $\mathbb{E}[\widehat{\boldsymbol{b}}_1] = \boldsymbol{\beta}_1$ but no-longer efficient.

Occam's Razor and Parsimony

Importance of penalty in order to avoid a too complex model (overfit).

Consider some linear predictor, $\widehat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$ where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$, or more generally $\widehat{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}$ for some smoothing matrix $\boldsymbol{S}$.

In econometrics, consider some ex-post penalty for model selection,

$$AIC = -2\log\mathcal{L} + 2p = \text{deviance} + 2p$$

where p is the dimension (i.e. trace($\boldsymbol{S}$) in a more general setting).

In machine-learning, penalty is added in the objective function, see Ridge or LASSO regression

$$(\widehat{\beta}_{0,\lambda}, \boldsymbol{\beta}_\lambda) = \underset{(\beta_0,\widehat{\boldsymbol{\beta}})}{\text{argmin}}\left\{\sum_{i=1}^n \ell(y_i, \beta_0 + \boldsymbol{x}_i \top \boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|\right\}$$

Goodhart's law, "when a measure becomes a target, it ceases to be a good measure"

Boosting, or learning from previous errors

Construct a sequence of models

$$m^{(k)}(\boldsymbol{x}) = m^{(k-1)}(\boldsymbol{x}) + \alpha \cdot f^\star(\boldsymbol{x})$$

where

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

for some set of weak learner $\mathcal{W}$.

Problem: where to stop to avoid overfit...