Estimation of conditional mean squared error of prediction

Filip Lindskog

based on work with M. Lindholm and F. Wahl forthcoming in Annals of Actuarial Science

October 16, 2019

Conditional MSEP

- X is a random variable whose value will be observed in the future.
- Current and past observations generate the σ-algebra $\mathcal{F}_0$
- $\widehat{X}$ is an $\mathcal{F}_0$-measurable predictor of X
- How accurate is the approximation $X \approx \widehat{X}$ given $\mathcal{F}_0$?
- $\mathrm{MSEP}(X, \widehat{X}) := \mathbb{E}[(X - \widehat{X})^2]$ is not so relevant
- $\mathrm{MSEP}_{\mathcal{F}_0}(X, \widehat{X}) := \mathbb{E}[(X - \widehat{X})^2 \mid \mathcal{F}_0]$ is relevant
- How should $\mathrm{MSEP}_{\mathcal{F}_0}(X, \widehat{X})$ be estimated with limited data?

Conditional MSEP with parametric model

- Assume a stochastic model with parameter $\boldsymbol{\theta}$
- Write $h(\boldsymbol{\theta}; \mathcal{F}_0) := \mathbb{E}[X \mid \mathcal{F}_0]$ and assume $h(\boldsymbol{\theta}; \mathcal{F}_0)$ is computable
- Consider predictor $\widehat{X} := h(\widehat{\boldsymbol{\theta}}; \mathcal{F}_0)$ given $\mathcal{F}_0$-measurable estimator $\widehat{\boldsymbol{\theta}}$
- Calculate

$$
\begin{aligned}
\mathrm{MSEP}_{\mathcal{F}_0}(X, \widehat{X}) &= \mathbb{E}[(X - \widehat{X})^2 \mid \mathcal{F}_0] \\
&= \mathbb{E}[(X - \mathbb{E}[X \mid \mathcal{F}_0] + \mathbb{E}[X \mid \mathcal{F}_0] - \widehat{X})^2 \mid \mathcal{F}_0] \\
&= \mathrm{Var}(X \mid \mathcal{F}_0) + \mathbb{E}[(\widehat{X} - \mathbb{E}[X \mid \mathcal{F}_0])^2 \mid \mathcal{F}_0] \\
&\quad \cancel{+ 2\mathbb{E}[(X - \mathbb{E}[X \mid \mathcal{F}_0])(\mathbb{E}[X \mid \mathcal{F}_0] - \widehat{X}) \mid \mathcal{F}_0]} \\
&= \mathrm{Var}(X \mid \mathcal{F}_0) + (h(\widehat{\boldsymbol{\theta}}; \mathcal{F}_0) - h(\boldsymbol{\theta}; \mathcal{F}_0))^2
\end{aligned}
$$

Naive estimation of conditional MSEP

- From last page

$$\mathrm{MSEP}_{\mathcal{F}_0}(X, \widehat{X}) = \mathrm{Var}(X \mid \mathcal{F}_0) + (h(\widehat{\boldsymbol{\theta}}; \mathcal{F}_0) - h(\boldsymbol{\theta}; \mathcal{F}_0))^2$$

- The naive plug-in estimator obtained by replacing $\boldsymbol{\theta}$ by $\widehat{\boldsymbol{\theta}}$ is

$$\widehat{\mathrm{MSEP}_{\mathcal{F}_0}(X, \widehat{X})} = \widehat{\mathrm{Var}(X \mid \mathcal{F}_0)} + 0$$

  which is a bad estimator of $\mathrm{MSEP}_{\mathcal{F}_0}(X, \widehat{X})$
- The naive estimator corresponds to ignoring prediction error due to parameter estimation error

Akaike's FPE

- $X_1, \ldots, X_n$ observations from AR(p_0),

$$X_{t+1} = \phi_1 X_t + \cdots + \phi_{p_0} X_{t-p_0+1} + Z_{t+1}, \quad (Z_t) \sim \text{iid}(0, \sigma^2),$$

  with unknown p_0. Determine p_0!
- Let $\widehat{\boldsymbol{\phi}}$ be an estimator of $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^{\mathrm{T}}$ based on AR(p) data $X_1, \ldots, X_n$. Let $(X_t^{\perp})_{t=1}^n$ be an independent copy of $(X_t)_{t=1}^n$.

$$\mathrm{FPE}_p(X_{n+1}, \widehat{X}_{n+1}) := \mathbb{E}\Big[\Big(X_{n+1}^{\perp} - \sum_{k=1}^{p} \widehat{\phi}_k X_{n+1-k}^{\perp}\Big)^2\Big]$$

  Set $\widehat{p}_0 := \arg\min_p \widehat{\mathrm{FPE}_p(X_{n+1}, \widehat{X}_{n+1})}$
- Convergence $\sqrt{n}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \xrightarrow{d}$ multivariate normal distribution gives

$$\mathrm{FPE}_p(X_{n+1}, \widehat{X}_{n+1}) \approx \widehat{\sigma}^2(n,p)\frac{n+p}{n-p} \quad \text{for large } n$$

Akaike's FPE in a more general conditional setting

- Let $\mathcal{T} := \{\underline{t}, \underline{t}+1, \ldots, \bar{t}\} \subset \mathbb{Z}$ with $\underline{t} < 0 < \bar{t}$
- Let $(\mathbf{S}_t)_{t\in\mathcal{T}}$, $(\mathbf{S}_t^\perp)_{t\in\mathcal{T}}$ be iid copies
- Let $(\mathcal{F}_t)_{t\in\mathcal{T}}$, $(\mathcal{F}_t^\perp)_{t\in\mathcal{T}}$ be filtrations generated by $(\mathbf{S}_t)_{t\in\mathcal{T}}$, $(\mathbf{S}_t^\perp)_{t\in\mathcal{T}}$
- Let $X = f((\mathbf{S}_t)_{t\in\mathcal{T}})$, $X^\perp = f((\mathbf{S}_t^\perp)_{t\in\mathcal{T}})$ for some f
- Let $\widehat{\boldsymbol{\theta}} = g((\mathbf{S}_t)_{t\in\mathcal{T},t\leq 0})$, $\widehat{\boldsymbol{\theta}}^\perp = g((\mathbf{S}_t^\perp)_{t\in\mathcal{T},t\leq 0})$ for some g

$$\begin{aligned}
\mathrm{FPE}_{\mathcal{F}_0}(X, \widehat{X}) &:= \mathbb{E}[(X^\perp - h(\widehat{\boldsymbol{\theta}}; \mathcal{F}_0^\perp))^2 \mid \mathcal{F}_0^\perp] \\
&= \mathbb{E}[(X - h(\widehat{\boldsymbol{\theta}}^\perp; \mathcal{F}_0))^2 \mid \mathcal{F}_0] \\
&= \mathrm{Var}(X \mid \mathcal{F}_0) + \mathbb{E}[(h(\widehat{\boldsymbol{\theta}}^\perp; \mathcal{F}_0) - h(\boldsymbol{\theta}; \mathcal{F}_0))^2 \mid \mathcal{F}_0]
\end{aligned}$$

whose plug-in estimator does not make the second term vanish

Estimation of conditional FPE

- First idea: write $H(\boldsymbol{\theta}) := \mathrm{FPE}_{\mathcal{F}_0}(X, \widehat{X})$ and estimate by $H(\widehat{\boldsymbol{\theta}})$
- Problem: We do not know the distribution of $\widehat{\boldsymbol{\theta}}$, therefore $\mathbb{E}[(h(\widehat{\boldsymbol{\theta}}^{\perp}; \mathcal{F}_0) - h(\boldsymbol{\theta}; \mathcal{F}_0))^2 \mid \mathcal{F}_0]$ is not computable
- Problem: Unlikely that independent copies $\widehat{\boldsymbol{\theta}}_1^{\perp}, \ldots, \widehat{\boldsymbol{\theta}}_n^{\perp}$ are available enabling the approximation of

$$\mathbb{E}[(h(\widehat{\boldsymbol{\theta}}^{\perp}; \mathcal{F}_0) - h(\boldsymbol{\theta}; \mathcal{F}_0))^2 \mid \mathcal{F}_0]$$

by

$$\frac{1}{n} \sum_{k=1}^{n} \left(h(\widehat{\boldsymbol{\theta}}_k^{\perp}; \mathcal{F}_0) - h(\boldsymbol{\theta}; \mathcal{F}_0)\right)^2$$

Estimation of conditional FPE

- Second idea: approximating

$$h(\widehat{\boldsymbol{\theta}}^{\perp}; \mathcal{F}_0) \approx h(\boldsymbol{\theta}; \mathcal{F}_0) + \nabla h(\boldsymbol{\theta}; \mathcal{F}_0)^{\mathrm{T}} (\widehat{\boldsymbol{\theta}}^{\perp} - \boldsymbol{\theta})$$

approximates $H(\boldsymbol{\theta}) := \mathrm{FPE}_{\mathcal{F}_0}(X, \widehat{X})$ by $H^{\nabla}(\boldsymbol{\theta})$ given by

$$\mathrm{Var}(X \mid \mathcal{F}_0) + \nabla h(\boldsymbol{\theta}; \mathcal{F}_0)^{\mathrm{T}} \mathbb{E}[(\widehat{\boldsymbol{\theta}}^{\perp} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}^{\perp} - \boldsymbol{\theta})^{\mathrm{T}}] \nabla h(\boldsymbol{\theta}; \mathcal{F}_0)$$

- Estimate $\mathrm{FPE}_{\mathcal{F}_0}(X, \widehat{X})$ and $\mathrm{MSEP}_{\mathcal{F}_0}(X, \widehat{X})$ by $H^{\nabla}(\widehat{\boldsymbol{\theta}})$
- If $\mathbb{E}[\widehat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$, then

$$H^{\nabla}(\boldsymbol{\theta}) = \mathrm{Var}(X \mid \mathcal{F}_0) + \nabla h(\boldsymbol{\theta}; \mathcal{F}_0)^{\mathrm{T}} \, \mathrm{Cov}(\widehat{\boldsymbol{\theta}}) \nabla h(\boldsymbol{\theta}; \mathcal{F}_0)$$

- The only remaining challenge is to estimate $\mathrm{Cov}(\widehat{\boldsymbol{\theta}})$

Data

$C_{i,j}$ denotes the total amount paid to policyholders during development periods $1, \ldots, j$ due to accidents in accident period i.

	1	2	3	4	5
i_0	$C_{i_0,1}$	$C_{i_0,2}$	$C_{i_0,3}$	$C_{i_0,4}$	$C_{i_0,5}$
$\vdots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
-5	$C_{-5,1}$	$C_{-5,2}$	$C_{-5,3}$	$C_{-5,4}$	$C_{-5,5}$
-4	$C_{-4,1}$	$C_{-4,2}$	$C_{-4,3}$	$C_{-4,4}$	$C_{-4,5}$
-3	$C_{-3,1}$	$C_{-3,2}$	$C_{-3,3}$	$C_{-3,4}$	$C_{-3,5}$
-2	$C_{-2,1}$	$C_{-2,2}$	$C_{-2,3}$	$C_{-2,4}$	$C_{-2,5}$
-1	$C_{-1,1}$	$C_{-1,2}$	$C_{-1,3}$	$C_{-1,4}$	$C_{-1,5}$

$S_0 = (C_{-1,1}, \ldots, C_{-5,5})^{\mathrm{T}}$, $S_{-1} = (C_{-2,1}, \ldots, C_{-6,5})^{\mathrm{T}}$, $\ldots$, $S_{i_0} = C_{i_0,1}$,
$X = C_{-1,5} + \cdots + C_{-4,5}$

Development-period (column) dynamics

- For vector-valued functions $\mathbf{F}_{i,j}$ and scalar-valued functions $G_{i,j}$,

$$C_{i,j+1} = \mathbf{F}_{i,j}\Big((C_{i,k})_{k\leq j}\Big)^{\mathrm{T}}\boldsymbol{\beta}_j + \sigma_j G_{i,j}\Big((C_{i,k})_{k\leq j}\Big)e_{i,j+1},$$

$\{C_{i,1}, e_{i,2}, \ldots, e_{i,J}\}$ and $\{C_{k,1}, e_{k,2}, \ldots, e_{k,J}\}$ independent if $i \neq k$,

$$\mathbb{E}[e_{i,j+1} \mid C_1, \ldots, C_j] = 0, \quad \mathbb{E}[e_{i,j+1}^2 \mid C_1, \ldots, C_j] = 1$$

Note: $e_{i,1}, \ldots, e_{i,J}$ not necessarily independent

- Development-period dynamics as vector-valued time series:

$$\mathbf{C}_{j+1} = \mathbf{A}_j\boldsymbol{\beta}_j + \sigma_j\mathbf{D}_j\mathbf{e}_{j+1}$$

- Mack's Chain Ladder: $C_{i,j+1} = C_{i,j}\beta_j + \sigma_j\sqrt{C_{i,j}}e_{i,j+1}$ when $\mathbf{F}_{i,j}\big((C_{i,k})_{k\leq j}\big) = \mathbf{F}\big((C_{i,k})_{k\leq j}\big) = C_{i,j}$ and $G_{i,j}\big((C_{i,k})_{k\leq j}\big) = G\big((C_{i,k})_{k\leq j}\big) = \sqrt{C_{i,j}}$

Observed data

	1	2	3	4	5
i_0	$C_{i_0,1}$	$C_{i_0,2}$	$C_{i_0,3}$	$C_{i_0,4}$	$C_{i_0,5}$
$\vdots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
-5	$C_{-5,1}$	$C_{-5,2}$	$C_{-5,3}$	$C_{-5,4}$	$C_{-5,5}$
-4	$C_{-4,1}$	$C_{-4,2}$	$C_{-4,3}$	$C_{-4,4}$	$C_{-4,5}$
-3	$C_{-3,1}$	$C_{-3,2}$	$C_{-3,3}$	$C_{-3,4}$	$C_{-3,5}$
-2	$C_{-2,1}$	$C_{-2,2}$	$C_{-2,3}$	$C_{-2,4}$	$C_{-2,5}$
-1	$C_{-1,1}$	$C_{-1,2}$	$C_{-1,3}$	$C_{-1,4}$	$C_{-1,5}$

Example: $\widetilde{\mathbf{C}}_4 = \widetilde{\mathbf{A}}_3\boldsymbol{\beta}_3 + \sigma_3\widetilde{\mathbf{D}}_3\widetilde{\mathbf{e}}_4$

Weighted least-squares estimation

The estimators

$$\widehat{\boldsymbol{\beta}}_j := \left(\widetilde{\mathbf{A}}_j^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_j^{-1}\widetilde{\mathbf{A}}_j\right)^{-1}\widetilde{\mathbf{A}}_j^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_j^{-1}\widetilde{\mathbf{C}}_{j+1}, \quad \widetilde{\boldsymbol{\Sigma}}_j := \widetilde{\mathbf{D}}_j^2, \quad \text{(WLS)}$$

$$\widehat{\boldsymbol{\sigma}}_j^2 := \frac{1}{n_j - p_j}\left(\widetilde{\mathbf{C}}_{j+1} - \widetilde{\mathbf{A}}_j\widehat{\boldsymbol{\beta}}_j\right)^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_j^{-1}\left(\widetilde{\mathbf{C}}_{j+1} - \widetilde{\mathbf{A}}_j\widehat{\boldsymbol{\beta}}_j\right)$$

lead to

- $\mathbb{E}[\widehat{\boldsymbol{\beta}}_j \mid \widetilde{\mathbf{A}}_j, \widetilde{\boldsymbol{\Sigma}}_j] = \mathbb{E}[\widehat{\boldsymbol{\beta}}_j] = \boldsymbol{\beta}_j$
- $\mathbb{E}[\widehat{\sigma}_j^2 \mid \widetilde{\mathbf{A}}_j, \widetilde{\boldsymbol{\Sigma}}_j] = \mathbb{E}[\widehat{\sigma}_j^2] = \sigma_j^2$
- $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_j, \widehat{\boldsymbol{\beta}}_k) = \mathbf{0}$ if $j \neq k$
- $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_j \mid \widetilde{\mathbf{A}}_j, \widetilde{\boldsymbol{\Sigma}}_j) = \sigma_j^2\left(\widetilde{\mathbf{A}}_j^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_j^{-1}\widetilde{\mathbf{A}}_j\right)^{-1}$

Estimating covariance matrix of regression coefficients

The estimator $\boldsymbol{\Lambda}(\widehat{\boldsymbol{\sigma}}^2; \mathcal{F}_0)$ of $\mathrm{Cov}(\widehat{\boldsymbol{\beta}})$ given by

$$\begin{bmatrix} \widehat{\sigma}_1^2 \big(\widetilde{\mathbf{A}}_1^{\mathrm{T}} \widetilde{\boldsymbol{\Sigma}}_1^{-1} \widetilde{\mathbf{A}}_1\big)^{-1} & 0 & \dots & 0 \\ 0 & \ddots & & \\ \vdots & & & \\ 0 & & & \widehat{\sigma}_{J-1}^2 \big(\widetilde{\mathbf{A}}_{J-1}^{\mathrm{T}} \widetilde{\boldsymbol{\Sigma}}_{J-1}^{-1} \widetilde{\mathbf{A}}_{J-1}\big)^{-1} \end{bmatrix}$$

satisfies

$$\mathbb{E}[\boldsymbol{\Lambda}(\widehat{\boldsymbol{\sigma}}^2; \mathcal{F}_0)] = \mathrm{Cov}(\widehat{\boldsymbol{\beta}}), \quad \widehat{\boldsymbol{\beta}} := \big(\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}, \dots, \widehat{\boldsymbol{\beta}}_{J-1}^{\mathrm{T}}\big)^{\mathrm{T}}$$

which can be shown using covariance decomposition

$$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_j) = \mathrm{Cov}(\mathbb{E}[\widehat{\boldsymbol{\beta}}_j \mid \mathcal{H}_j]) + \mathbb{E}[\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_j \mid \mathcal{H}_j)]$$

for suitably chosen $\mathcal{H}_j$ together with conditional unbiasedness of $\widehat{\boldsymbol{\beta}}_j$.

Putting the pieces together

Recall that we want to estimate

$$\mathrm{Var}(X \mid \mathcal{F}_0) + \nabla h(\boldsymbol{\theta}; \mathcal{F}_0)^{\mathrm{T}} \, \mathrm{Cov}(\widehat{\boldsymbol{\theta}}) \nabla h(\boldsymbol{\theta}; \mathcal{F}_0)$$

which, for the considered model class with $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ and X is such that $h(\boldsymbol{\beta}, \boldsymbol{\sigma}^2; \mathcal{F}_0) = \mathbb{E}[X \mid \mathcal{F}_0]$ does not depend on $\boldsymbol{\sigma}^2$, equals

$$\mathrm{Var}(X \mid \mathcal{F}_0) + \nabla_{\boldsymbol{\beta}} h(\boldsymbol{\beta}, \boldsymbol{\sigma}^2; \mathcal{F}_0)^{\mathrm{T}} \, \mathrm{Cov}(\widehat{\boldsymbol{\beta}}) \nabla_{\boldsymbol{\beta}} h(\boldsymbol{\beta}, \boldsymbol{\sigma}^2; \mathcal{F}_0)$$

which is estimated by

$$\widehat{\mathrm{Var}(X \mid \mathcal{F}_0)} + \nabla_{\boldsymbol{\beta}} h(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}^2; \mathcal{F}_0)^{\mathrm{T}} \boldsymbol{\Lambda}(\widehat{\boldsymbol{\sigma}}^2; \mathcal{F}_0) \nabla_{\boldsymbol{\beta}} h(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}^2; \mathcal{F}_0)$$

Tedious calculations show that the above expression coincides with estimators of conditional MSEP proposed for special models, including Mack's Chain Ladder.

Generalisation of conditional FPE

- We estimated $\mathrm{MSEP}_{\mathcal{F}_0}(X, \widehat{X})$ by estimating

$$\mathrm{FPE}_{\mathcal{F}_0}(X, \widehat{X}) = \mathbb{E}[(X - h(\widehat{\boldsymbol{\theta}}^{\perp}; \mathcal{F}_0))^2 \mid \mathcal{F}_0]$$

  Why $\widehat{\boldsymbol{\theta}}^{\perp}$? Any $\widehat{\boldsymbol{\theta}}^{*}$ conditionally independent of X given $\mathcal{F}_0$ makes most of the analysis stay the same when considering

$$\mathrm{MSEP}^{*}_{\mathcal{F}_0}(X, \widehat{X}) = \mathbb{E}[(X - h(\widehat{\boldsymbol{\theta}}^{*}; \mathcal{F}_0))^2 \mid \mathcal{F}_0]$$

- $\mathrm{MSEP}^{*}_{\mathcal{F}_0}(X, \widehat{X})$ together with development-period time-series dynamics $(\mathbf{C}_j)$ with iid residuals $(\mathbf{e}_j)$ with iid components

$$\mathbf{C}_{j+1} = \mathbf{A}_j \boldsymbol{\beta}_j + \sigma_j \mathbf{D}_j \mathbf{e}_{j+1}$$

  allows for estimation of $\mathrm{MSEP}^{*}_{\mathcal{F}_0}(X, \widehat{X})$ by parametric bootstrap. $\mathrm{FPE}_{\mathcal{F}_0}(X, \widehat{X})$ does not.

Natural next step

Recalling the originally intended use of Akaike's FPE, a natural next step would be to...

...use $\mathrm{MSEP}^*_{\mathcal{F}_0}(X, \widehat{X})$ together with machine learning ideas for automatic predictor selection including assessment of prediction accuracy.