Size dependence and allocation aspects of validation

Péter Király
1st year PhD report
Supervisor: Gergely Tóth

May 31, 2019

Outline

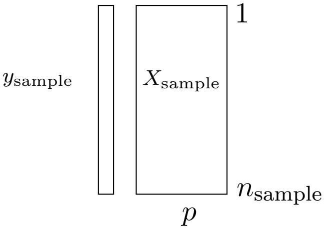
- Sample size dependence of validation parameters (VPs)
  - compared on models with weak and good linear fit
  - remarks on R^2 and CCC
  - correction of degrees of freedom: Q^2_{F3}
  - visualization of Roy-Ojha parameters
  - correlation of VPs
  - 66 VPs calculated
- Predictor allocation aspects of validation parameters
- Summary

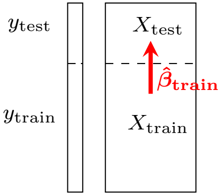
Validation schemes

Internal validation

External validation

$$y = X\beta + e_i$$





goodness of fit, robustness

$$\hat{r}_i = (y_i - \hat{y}_i)$$

predictability

$$\hat{r}_i = (y_{i,\text{test}} - \hat{y}_{i,\text{test}}(\hat{\beta}_{\text{train}}))$$

Validation parameters (VPs) - Rácz, *SAR QSAR Environ. Res.* **26**, 683 (2015)

internal VPs

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

$$Q^2_{\mathrm{LOO}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_{i/i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\mathrm{PRESS}}{\mathrm{TSS}}$$

$$\mathrm{CCC} = \frac{2\sum_{i=1}^{n}(y_i - \bar{y})\left(\hat{y}_i - \bar{\hat{y}}\right)}{\sum_{i=1}^{n}(y_i - \bar{y})^2 + \sum_{i=1}^{n}\left(\hat{y}_i - \bar{\hat{y}}\right)^2 + n\left(\bar{y} - \bar{\hat{y}}\right)^2}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad \ldots$$

external VPs

$$Q^2_{F1} = 1 - \frac{\sum_{i=1}^{n,\mathrm{test}}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n,\mathrm{test}}(y_i - \bar{y}_{\mathrm{train}})^2}$$

$$Q^2_{F2} = 1 - \frac{\sum_{i=1}^{n,\mathrm{test}}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n,\mathrm{test}}(y_i - \bar{y}_{\mathrm{test}})^2}$$

$$Q^2_{F3} = 1 - \frac{\sum_{i=1}^{n,\mathrm{test}}(y_i - \hat{y}_i)^2 / n_{\mathrm{test}}}{\sum_{i=1}^{n,\mathrm{train}}(y_i - \bar{y}_{\mathrm{train}})^2 / n_{\mathrm{train}}}$$

$$\vdots$$

Size dependence of VPs - Valinear(R)

Datasets

Concrete Compressive Strength set

Combined Cycle Power Plant set

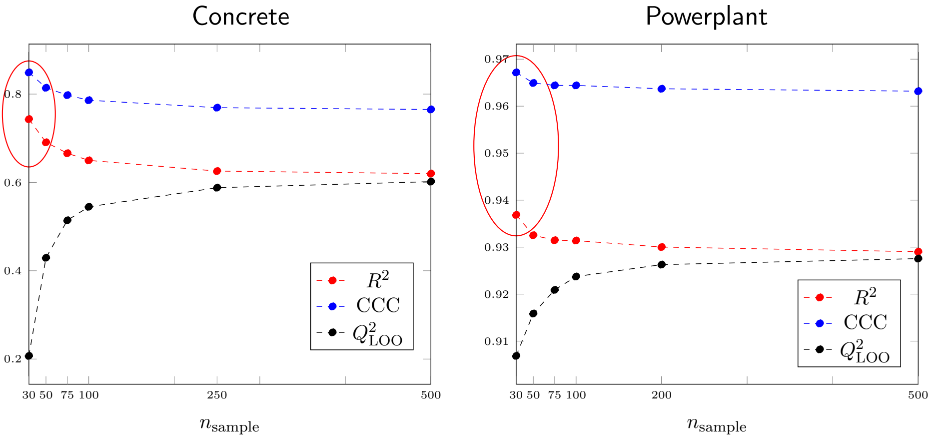
- sets are modeled with *multivariate linear regression* (MLR)

dataset	Concrete	Powerplant
# response variables	1	1
# explanatory variables	8	4
# objects	1030	9658
adequacy for MLR?	weak	good

Tasks for size dependence study

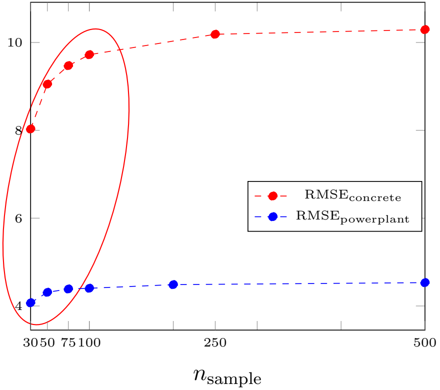
- random sampling from population
- *training/test* split of sample for external validation (80/20 for training)
- linear modeling and computation of *validation parameters* (VPs)
- tasks above for sample sizes, $n_{\text{sample}} = (30, 50, 75, 100, 250, 500)$, then repeated $1000\times$, VPs averaged

Internal VPs



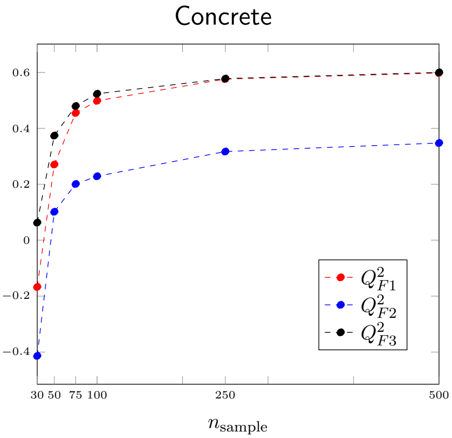
Is the smaller model the better one?

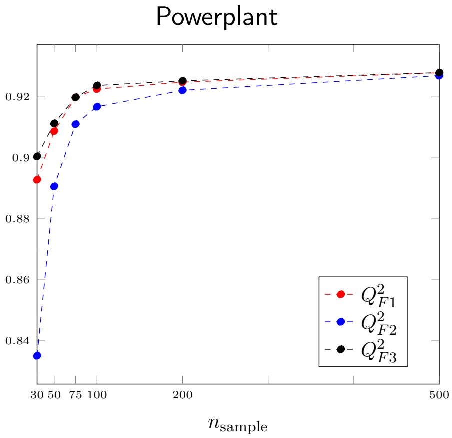
Internal VPs



Is the smaller model the better one?

External VPs



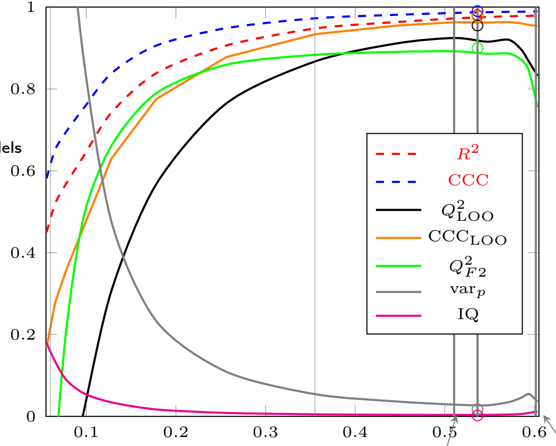


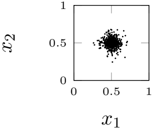
Q^2_{F2} is the most sensitive

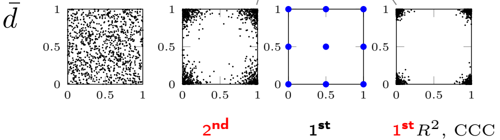
$\underline{x} \in \mathbb{R}^2 \rightarrow y$

$n_{\text{sample}} = 9$

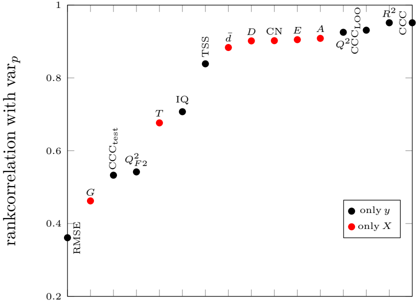
16k computer gen. models with tendency in $\underline{x}$

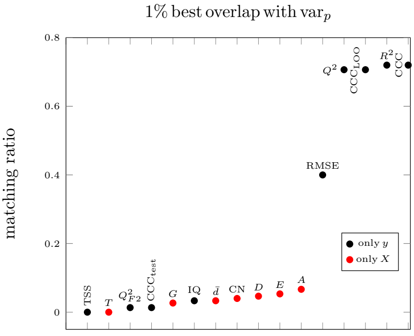






Rank correlation with var_p





significant matching only for cases where we know y values

Summary

- model validation: goodness of fit, robustness & predictivity
- sample size dependence of R^2, CCC, RMSE show anomaly
- Q^2_{F1-3} for external validation show different sensitivity for different sample size

- nonlinear modeling, PLS

Co-workers: Dániel Kovács, Gergely Tóth