Statistical learning for biological data

David Causeur

Department of Statistics and Computer Science

Agrocampus Ouest

IRMAR CNRS UMR 6625

http://www.agrocampus-ouest.fr/math/causeur/

Outline

Search for the best model

In the linear model framework

If p predictors, then $2^p - 1$ possible models

$$\binom{k}{p} \text{ models } \mathcal{M}_k \text{ of size } k$$

Still possible in R provided p is no larger than 30 (for a real-valued Y)

▶ Exhaustive search of the best model in `R`

Prediction accuracy

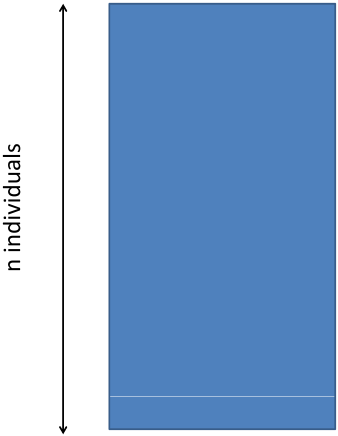
Let x_0 be the $x-$profile of an individual with response value Y_0

Prediction error: $Y_0 - \hat{Y}_0$

- $\mathbb{E}_{x_0}(Y_0 - \hat{Y}_0) = 0$
- Precision: $\sigma_p^2 = \text{Var}(Y_0 - \hat{Y}_0) = \mathbb{E}\left[(Y_0 - \hat{Y}_0)^2\right]$

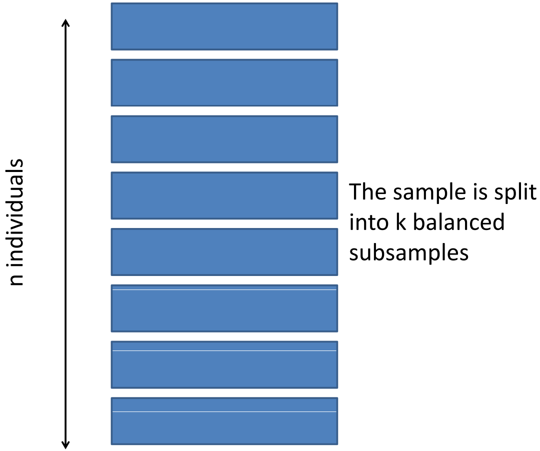
Prediction accuracy

Estimation of σ_p^2 using cross-validation



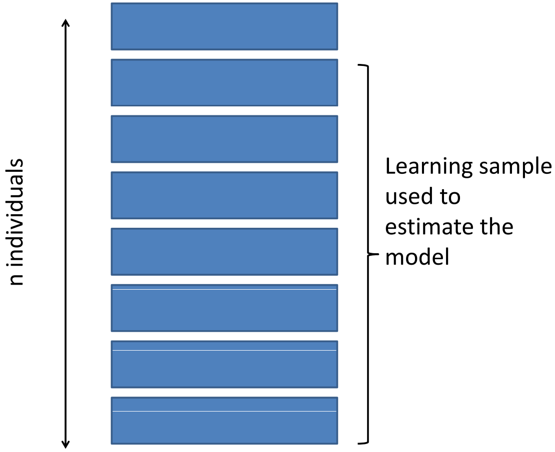
Prediction accuracy

Estimation of σ_p^2 using cross-validation



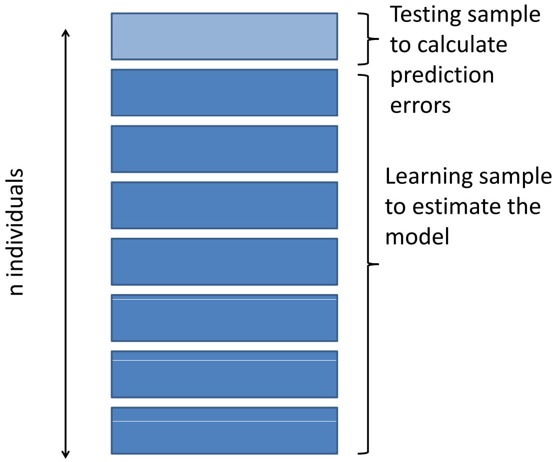
Prediction accuracy

Estimation of σ_p^2 using cross-validation



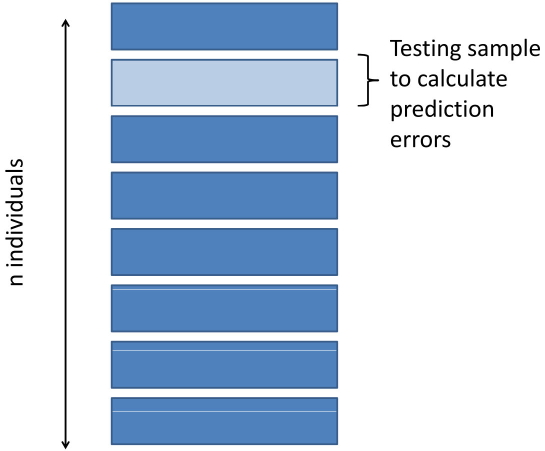
Prediction accuracy

Estimation of σ_p^2 using cross-validation



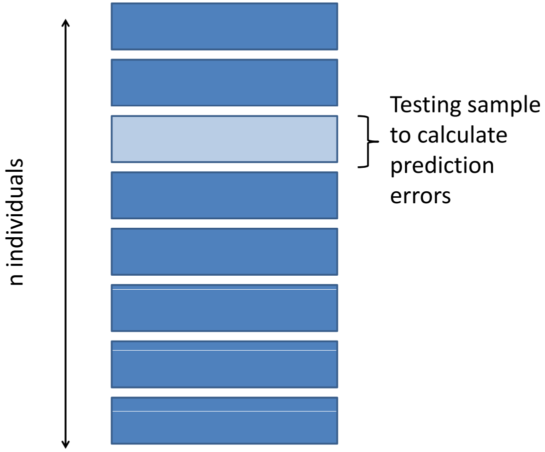
Prediction accuracy

Estimation of σ_p^2 using cross-validation



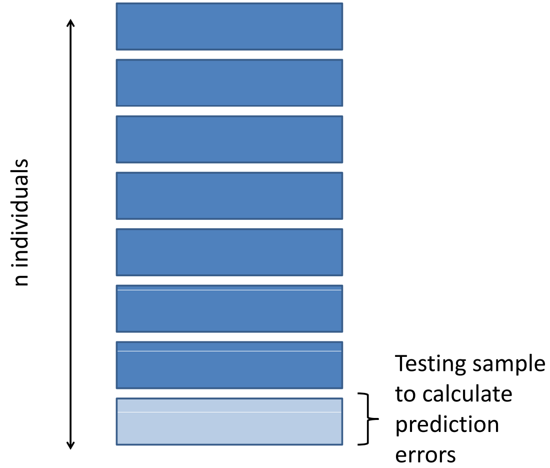
Prediction accuracy

Estimation of σ_p^2 using cross-validation



Prediction accuracy

Estimation of σ_p^2 using cross-validation



Prediction accuracy

Let x_0 be the $x-$profile of an individual with response value Y_0

Prediction error: $Y_0 - \hat{Y}_0$

- $\mathbb{E}_{x_0}(Y_0 - \hat{Y}_0) = 0$
- Precision: $\sigma_p^2 = \text{Var}(Y_0 - \hat{Y}_0) = \mathbb{E}\left[(Y_0 - \hat{Y}_0)^2\right]$

Estimation of σ_p^2 using cross-validation

$$\text{PRESS} = \sum_{i=1}^{n}\left[Y_i - \hat{Y}_{-i}\right]^2$$ [= Prediction Error Sum of Squares],

$$\hat{\sigma}_p^2 = \frac{\text{PRESS}}{n}$$ [= Mean Squared Error of Prediction]

▶ Evaluation of prediction accuracy in R

Penalized fitting performance criteria

Model comparison has to account for the complexity of the model.

The Akaike Information Criterion is given by:

$$\begin{aligned} \mathrm{AIC}(\mathcal{M}_k) &= \mathcal{D}_k + 2(k+1), \\ &\propto n \log\left[\frac{\mathrm{RSS}(\mathcal{M}_k)}{n}\right] + 2(k+1), \end{aligned}$$

where $\mathcal{D}_k$ is the residual deviance of model $\mathcal{M}_k$.

$\mathrm{AIC}(\mathcal{M}_k)$ estimates the information loss when using the estimated model rather than the unknown model that is supposed to generate the data.

Penalized fitting performance criteria

Alternatively, the Bayesian Information Criterion is defined as follows:

$$\begin{aligned} \mathrm{BIC}(\mathcal{M}_k) &= \mathcal{D}_k + \ln(n)(k+1), \\ &\propto n \log\left[\frac{\mathrm{RSS}(\mathcal{M}_k)}{n}\right] + \ln(n)(k+1). \end{aligned}$$

to measure the information loss when using the estimated model rather than the true model in the scope of parametric models considered.

Penalized fitting performance criteria

AIC or BIC?

- If the goal is to make a prediction rule, AIC is recommended.
- If the goal is just to fit the model, BIC shall be favored.

The goodness-of-fit of a model is **more penalized** by its complexity when it is evaluated by BIC than AIC.

▶ Search for the model with minimal AIC/BIC in `R`

Search for the best model

In the Logistic Linear Regression model framework

The Information Criteria of model $\mathcal{M}_k$ with p_k parameters are given by:

$$\text{IC}(\mathcal{M}_k) = \mathcal{D}_k + \lambda p_k,$$

with $\lambda = 2$ for AIC and $\lambda = \ln n$ for BIC.

Search for the submodel with minimal IC in R

- 2-class response: exhaustive search possible provided $p \leq 15$ (package `bestglm`)
- $K > 2$-class response: only stepwise search possible

▶ Search for the model with minimal AIC/BIC in `R`

Prediction performance

Let x_0 be the $x-$profile of an individual with response value Y_0

Bayes rule: $\hat{Y}_0 = y_k$ where $\hat{\mathbb{P}}(Y_0 = y_k)$ is the largest estimated class probability.

Criteria for prediction accuracy

- Accuracy: estimated probability that $\hat{Y}_0 = Y_0$.
- Weighted mean of estimated probabilities that $\hat{Y}_0 = k$ given $Y_0 = k$

▶ Cross-validated accuracy in R