Stat 849: Theory and Application of Cross-validation in Linear Regression Models

Sündüz Keleş

Department of Statistics
Department of Biostatistics and Medical Informatics
University of Wisconsin, Madison

Outline

- Measuring performance of a candidate estimator (predictor) by its risk (expected loss). Notions of loss and risk functions.
- Various estimators of risk: resubstitution estimator, test set estimator, cross-validation estimator.
- Choosing the best estimator based on cross-validated risk.

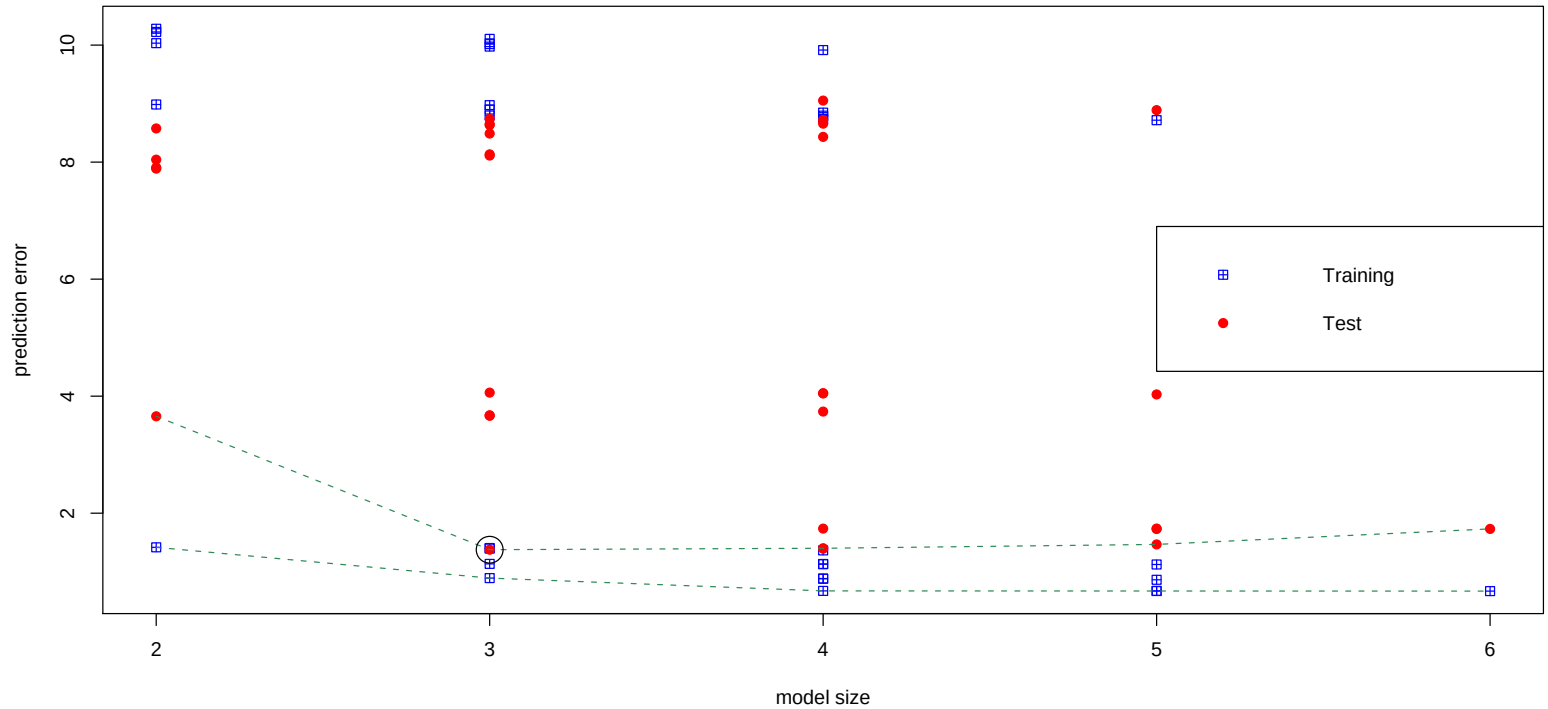


Figure: Mean residual sum of squares based on training data (data used to estimate the regression coefficients) and independent test data (new data) [code in `Training_Test_Error.R`].

Prediction error of best model at various model sizes.

```
> cbind(msize[PEindex], PE[PEindex])
      [,1]     [,2]
 [1,]    2 3.656369
 [2,]    3 1.376367  *BEST MODEL W.R.T. TEST DATA*
 [3,]    4 1.400373
 [4,]    5 1.466514
 [5,]    6 1.731655
> cbind(msize[RSSindex], RSS[RSSindex])
      [,1]      [,2]
 [1,]    2 1.4166478
 [2,]    3 0.8908512
 [3,]    4 0.6719834
 [4,]    5 0.6700969
 [5,]    6 0.6683968 *BEST MODEL W.R.T TRAINING (LEARNING) DATA*
```

Notions of loss and risk functions

Data: We have n i.i.d. observations $X_i = (Y_i, W_{i1}, \cdots, W_{ip})$, $i = 1, \cdots, n$ from a data generating distribution P_0, i.e., $X_i \sim P_0$, $i = 1, \cdots, n$.
We usually refer $X_1, \cdots, X_n$ as the learning set since these data are used to *estimate* or *learn* population parameters.
Model: $Y = W\beta + \epsilon$, $E[\epsilon \mid W] = 0$.
Parameter of interest. We will denote the parameter of interest by $\mu_0 = \mu(W) = E_{P_0}[Y \mid W]$.

Notions of loss and risk functions

Loss function.

- Loss functions are typically used to quantify error in prediction.
- A loss function $L : (X, \mu) \longrightarrow L(X, \mu) \in R$ is a real valued function of a candidate parameter value μ and an observation $X \sim P_0$. $L(y, \hat{y})$ elaborates the loss incurred when predicting y by $\hat{y}$.
- In the regression context, we work with the *squared error loss function* defined as $L(y, \hat{y}) = (y - \hat{y})^2$.

Notions of loss and risk functions

Risk function. For a given loss function $L(X, \mu)$, with $\mu \in \Psi$ (Ψ represents the parameter space, i.e., $\mathbb{R}^p$ for β in a linear regression model with p explanatory variables) and $X \sim P_0$, the risk is the expected value of the loss function with respect to P_0,

$$R(\mu, P_0) = E_{P_0}[L(X, \mu)] = \int L(x, \mu) dP_0(x) = \int L(x, \mu) f(x) dx.$$

Notions of loss and risk functions

Risk function. For a given loss function $L(X, \mu)$, with $\mu \in \Psi$ (Ψ represents the parameter space, i.e., $\mathbb{R}^p$ for β in a linear regression model with p explanatory variables) and $X \sim P_0$, the risk is the expected value of the loss function with respect to P_0,

$$R(\mu, P_0) = E_{P_0}[L(X, \mu)] = \int L(x, \mu) dP_0(x) = \int L(x, \mu) f(x) dx.$$

When (unrealistically) P_0 is known, it is possible to define an optimal predictor μ_{opt}, which minimizes the risk function:
$\mu_{opt} = \text{argmin}_{\mu \in \Psi} R(\mu, P_0)$.

Exercise: Show that for the squared error loss function, the optimal predictor is $\mu_{opt}(W) = E_{P_0}(Y \mid X)$. Our goal is to use the sample $X_1, \cdots, X_n$ to estimate the parameter of interest μ_0 of the unknown data generating distribution P_0. We will potentially have many candidate estimators and we would like to choose among these based on their risk.

Notions of loss and risk functions

Let P_n be the empirical distribution of the data $(X_1, \cdots, X_n)$, i.e., each data point $X_i = (Y_i, W_{i1}, \cdots, W_{ip})$ gets mass $1/n$.
Definition. An estimator $\hat{\mu}$ is a mapping from the empirical distributions to the parameter space Ψ. A realization of this mapping corresponding to a particular empirical distribution P_n is denoted by $\mu_n = \hat{\mu}(P_n)$. **E.g.**

$$\mu_n = W\hat{\beta}_{LS}, \quad \text{from the full model fit.}$$
$$\mu_n = W_1\hat{\beta}_{1,LS}, \quad \text{from a submodel fit.}$$

True risk vs resubstitution and test set estimators of risk

The *true, unknown* risk of this estimator μ_n is

$$E_{P_0}[L(X, \mu_n)] = \int L(x, \mu_n) dP_0(x), \tag{1}$$

where $L(x, \mu_n) = (y - \mu_n(w))^2$.

True risk vs resubstitution and test set estimators of risk

The *true, unknown* risk of this estimator μ_n is

$$E_{P_0}[L(X, \mu_n)] = \int L(x, \mu_n) dP_0(x), \tag{1}$$

where $L(x, \mu_n) = (y - \mu_n(w))^2$.
Note that this risk is a random variable as it depends on the data $X_1, \cdots, X_n$ via the empirical distribution P_n.
If we knew the true data generating distribution P_0, we could compare various candidate estimators μ_n based on their risk.

True risk vs resubstitution and test set estimators of risk

The *true, unknown* risk of this estimator μ_n is

$$E_{P_0}[L(X, \mu_n)] = \int L(x, \mu_n) dP_0(x), \tag{1}$$

where $L(x, \mu_n) = (y - \mu_n(w))^2$.
Note that this risk is a random variable as it depends on the data $X_1, \cdots, X_n$ via the empirical distribution P_n.
If we knew the true data generating distribution P_0, we could compare various candidate estimators μ_n based on their risk.
Resubstitution risk estimator. The empirical or resubstitution risk estimator for $\mu_n = \hat{\mu}(P_n)$ replaces the unknown data generating distribution P_0 by the known empirical distribution P_n.
Any guesses what this corresponds to in the linear regression model, i.e., plug in P_n for P_0 in the equation 1?

Resubstitution risk estimator

$$
\begin{aligned}
E_{P_n}[L(X,\mu_n)] &= \int L(X,\mu_n)dP_n(X) \\
&= \frac{1}{n}\sum_{i=1}^{n} L(X_i,\mu_n) = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu_n(W_i))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(Y_i-W_i\hat{\beta})^2.
\end{aligned}
$$

Resubstitution risk estimator

$$
\begin{aligned}
E_{P_n}[L(X, \mu_n)] &= \int L(X, \mu_n) dP_n(X) \\
&= \frac{1}{n} \sum_{i=1}^{n} L(X_i, \mu_n) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu_n(W_i))^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} (Y_i - W_i \hat{\beta})^2.
\end{aligned}
$$

Resubstitution estimator of risk equals mean residual sum of squares.

This estimator is severely biased downward due to overfitting. The learning data used to estimate the parameter of interest is also used to estimate its risk!

True risk, resubstitution and test set estimators of risk for candidate linear regression estimators of $E_{P_0}[Y \mid W]$

E.g. Define $X = (Y, W)$ where $W \sim \mathcal{N}(4, 1)$ and $Y = 3 - 5W + \epsilon$ with $\epsilon \sim \mathcal{N}(0, 2)$, thus $Y \mid W \sim \mathcal{N}(\mu_0(W) = 3 - 5W, \sigma_0^2 = 2)$. Parameter of interest $\mu_0 = E_{P_0}[Y \mid W] = 3 - 5W$. Let $\beta_0 = 3$, $\beta_1 = -5$.

```
set.seed(1)
w = rnorm(50, 4, 1)
y = 3-5 * w + rnorm(50, 0, 2)

lm1 = lm(y ~ w)
> betan = lm(y ~ w)$coef
> betan
(Intercept)           w
   3.608191   -5.091097
```

We have $\mu_n(W) = \beta_{n,0} + \beta_{n,1} W$ where $\beta_{n,0} = 3.608191$ and $\beta_{n,1} = -5.091097$.

First let's look at the true risk of this estimator $\mu_n(W) = \beta_{n,0} + \beta_{n,1} W$.

First let's look at the true risk of this estimator $\mu_n(W) = \beta_{n,0} + \beta_{n,1} W$.

$$
\begin{aligned}
E_{P_0}[(Y - \mu_n(W))^2] &= E_{P_0}[(Y - \mu_0(W) + \mu_0(W) - \mu_n(W))^2] \\
&= E_{P_0}[(Y - \mu_0(W))^2] + E_{P_0}[(\mu_0(W) - \mu_n(W))^2] \\
&\quad + E_{P_0}[(Y - \mu_0(W))(\mu_0(W) - \mu_n(W))] \\
&\quad \text{Cross-term vanishes (\textit{hint:} first condition on } W\text{):} \\
&= \sigma_0^2 + E_{P_0}[(\beta_0 + \beta_1 W - \beta_{0,n} - \beta_{1,n} W)^2] \\
&= \sigma_0^2 + (\beta_0 - \beta_{0,n})^2 + 2(\beta_0 - \beta_{n,0})(\beta_1 - \beta_{n,1}) E_{P_0}(W) \\
&\quad + (\beta_1 - \beta_{n,1})^2 E_{P_0}(W^2).
\end{aligned}
$$

This is a quantity that we can calculate when we know P_0 (or equivalently the marginal distribution of W and the conditional distribution of $Y \mid W$).

[R code in handout_110308.R]

```
#Resubstitution (empirical) risk estimate for \mu_n
> sum(lm1$resid^2)/50
[1] 3.673798

#True risk
sigma0 = 2
beta0 = 3
beta1 = -5
betan0 = betan[1]
betan1 = betan[2]

muw = 4 #E[W]

muwsq = 17 #E[W^2] obtained using var(W) = E[W^2]-E[W]^2

#Plug in the formula for the true risk

sigma0^2 + (beta0 - betan0)^2 + 2*(beta0 - betan0)
* (beta1 - betan1) * muw + (betan1 - beta1)^2 * muwsq
   4.067739 #True risk
```

Test set risk estimator

Test set risk estimator. Another risk estimate is based on an independent sample.

Test set risk estimator

Test set risk estimator. Another risk estimate is based on an independent sample. Let $(X_1^{TS}, \cdots, X_{n_{TS}}^{TS})$be an independent sample of size n_{TS}. Then, the test set risk estimate is given by

$$\frac{1}{n_{TS}} \sum_{i=1}^{n_{TS}} (Y_i^{TS} - \mu_n(W_i^{TS}))^2.$$

Test set risk estimator

Test set risk estimator. Another risk estimate is based on an independent sample. Let $(X_1^{TS}, \cdots, X_{n_{TS}}^{TS})$be an independent sample of size n_{TS}. Then, the test set risk estimate is given by

$$\frac{1}{n_{TS}} \sum_{i=1}^{n_{TS}} (Y_i^{TS} - \mu_n(W_i^{TS}))^2.$$

Compute a test set risk estimator for the above example:

```
set.seed(1)
wts = rnorm(10000, 4, 1)
yts = 3 - 5 * wts + rnorm(10000, 0, 2)
mean((yts - betan0 - betan1 * wts)^2)
[1] 4.000341
```

So, if we have an independent data set (a data set that we have not touched while *learning* the parameter of interest, i.e., while constructing the candidate estimator), we could get a better estimate of the risk. What happens when $n_{TS} \longrightarrow \infty$?

However, very rarely, we have the luxury of setting aside a portion of the dataset "untouched". Cross-validation tries to bypass this, by reusing the learning data set in a clever way.

However, very rarely, we have the luxury of setting aside a portion of the dataset "untouched". Cross-validation tries to bypass this, by reusing the learning data set in a clever way.

Again remember that we would like to have a good estimator of the risk for a given candidate predictor, because we will be using the estimated risk to choose among predictors.

Cross-validation

Cross-validation is a general approach for the following two tasks:

- **Risk estimation.** Given a candidate estimator $\mu_n = \hat{\mu}(P_n)$ of a parameter $\mu_0 = \mu(P_0)$, we wish to estimate the risk of μ_n with respect to the unknown true data generating distribution P_0, that is
$$\int L(X, \mu_n) dP_0.$$
- **Estimator selection.** Select an optimal (in terms of risk) estimator among K possible candidate estimators
$$\{\mu_{n,k} = \hat{\mu}_k(P_n) : k = 1, \cdots, K\},$$
for a parameter $\mu_0 = \mu(P_0)$.
E.g.
$$\begin{aligned} \mu_{n,1} &= \hat{\beta}_0 + \hat{\beta}_1 W_1 \\ \mu_{n,2} &= \hat{\beta}_0 + \hat{\beta}_1 W_1 + \hat{\beta}_2 W_2 \\ \mu_{n,3} &= \hat{\beta}_0 + \hat{\beta}_1 W_1 + \hat{\beta}_2 W_2 + \hat{\beta}_3 W_3 \\ &\vdots \end{aligned}$$

Cross-validation

- The main idea in CV is to divide the available learning data into two sets: a training set and a validation set.
- Observations in the training set are used to compute or *train* the estimators and the validation set is used to assess the risk (or *validate*) these estimators.
- Define a binary random n-vector or *split* vector, $B_n \in \{0,1\}^n$, independent of the empirical distribution P_n.

Split vector for cross-validation

A realization of $B_n = (B_n(1), \cdots, B_n(n))$ defines a particular split of the learning set of n observations into a training and a validation set.

$$B_n(i) = \begin{cases} 0 & \text{i-th observation is in the training set} \\ 1 & \text{i-th observation is in the validation set.} \end{cases}$$

Let P^1_{n,B_n} and P^0_{n,B_n} denote the empirical distributions of the training and the validation sets. $p_n = n_1/n$ be the proportion of the observations in the validation set, where $n_1 = \sum_i I(B_n(i) = 1)$.

Cross-validation risk estimator

A general definition of the cross-validation risk estimator for $\mu_n = \hat{\mu}(P_n)$ is

$$E_{B_n} \int L(x, \hat{\mu}(\underbrace{P^0_{n,B_n}}_{\text{training}})) \underbrace{dP^1_{n,B_n}}_{\text{validation}}(x)$$

$$= E_{B_n} \frac{1}{n_1} \sum_{i:B_n(i)=1} L(X_i, \hat{\mu}(P^0_{n,B_n})).$$

The particular distribution of the split vector B_n defines the type of cross-validation procedure. This representation covers many types of CV.

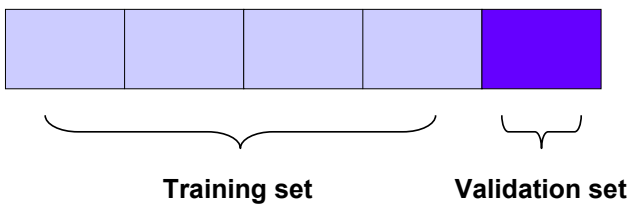
Commonly used cross-validation schemes

LOOCV (Leave-one-out CV). Each observation in the learning set is used in turn as the validation set and the remaining $n-1$ observations are used as the training set. The corresponding distribution B_n places mass $1/n$ on each of the n binary vectors, $b_n = (b_n(1), \cdots, b_n(n))$ such that $\sum_i b_n(1) = 1$ ($p_n = 1/n$).

Commonly used cross-validation schemes

LOOCV (Leave-one-out CV). Each observation in the learning set is used in turn as the validation set and the remaining $n-1$ observations are used as the training set. The corresponding distribution B_n places mass $1/n$ on each of the n binary vectors, $b_n = (b_n(1), \cdots, b_n(n))$ such that $\sum_i b_n(1) = 1$ $(p_n = 1/n)$.

V-fold CV. The learning set is randomly divided into V mutually exclusive and exhaustive sets, and each set is used in turn as the validation set. The corresponding distribution of B_n places mass $1/V$ on each of V binary vectors $b_n^v = (b_n(1)^v, \cdots, b_n^v(n))$, $v = 1, \cdots, V$ such that $\sum_i b_n(i)^v \approx n/v$ and $\sum_v b_n^v(i) = 1$ $(P_n = 1/V)$.



Selecting an optimal estimator

Estimators. $\mu_{n,1}, \cdots, \mu_{n,K}$ where $\mu_{n,i} = \hat{\mu}_i(P_n)$, e.g., estimator from the model fit using k predictors. We would like to select $\tilde{k}$ such that the true risk of the estimator with respect to P_0 is minimized

$$\tilde{k} = \text{argmin}_{k=1,\cdots,K} \int L(X, \hat{\mu}_k(P_n)) dP_0(x).$$

$\tilde{k}$ is usually referred to as *oracle or benchmark* selector.

Selecting an optimal estimator

Problem. P_0 is usually unknown!
To bypass this, we use cross-validated risk estimator to choose among the candidate estimators

$$\hat{k} = \text{argmin}_{k=1,\cdots,K} E_{B_n} \int L(X, \hat{\mu}_k(P^0_{n,B_n})) dP^1_{n,B_n}(X).$$

That is, the cross-validation estimator $\hat{\mu}_{n,\hat{k}}$ is chosen to have the best performance on the validation set.

Selecting an optimal estimator

Problem. P_0 is usually unknown!
To bypass this, we use cross-validated risk estimator to choose among the candidate estimators

$$\hat{k} = \text{argmin}_{k=1,\cdots,K} E_{B_n} \int L(X, \hat{\mu}_k(P^0_{n,B_n})) dP^1_{n,B_n}(X).$$

That is, the cross-validation estimator $\hat{\mu}_{n,\hat{k}}$ is chosen to have the best performance on the validation set.
Can show that $\hat{k}$ is optimal in the sense that it performs (in terms of risk) asymptotically as well as the optimal benchmark or oracle selector $\tilde{k}$ based on the true unknown data generating distribution P_0.

```
library(DAAG)
cvv2 = cv.lm(data.frame(cbind(y, w)), y ~ w, 2)

cv.lm(data.frame(cbind(y, w)), y ~ w, 2)

#cross-validation risk estimate with v = 2 is 3. 73.
#cross-validation risk estimate with v = 3 is 3. 94.
#cross-validation risk estimate with v = 5 is 3. 83.
#cross-validation risk estimate with v = 10 is 3. 95.

#In practice, typically 5 or 10 fold CV is commonly used.
```

```
#A cross-validation function is
#available through the boot package.

library(boot)

glm1 = glm (y ~ x1, data = egdata, family = gaussian)
cv1 = cv.glm(egdata, glm1, K = 5)$delta[1]

glm2 = glm (y ~ x1 + x3, data = egdata, family =  gaussian)
cv2 = cv.glm(egdata, glm2, K = 5)$delta[1]

glm3 = glm (y ~ x1 + x3 + x2, data = egdata, family =  gaussian)
cv3 = cv.glm(egdata, glm3, K = 5)$delta[1]

glm4 = glm (y ~ x1 + x3 +  x2 + x4, data = egdata, family = gaussian)
cv4 = cv.glm(egdata, glm4, K = 5)$delta[1]
```

Summary

If we believe that the *correct* model is in fact

$$Y = X_k \beta_k + \epsilon,$$

where X_k is the design matrix with only k variables and the full design matrix is of dimension $p \geq k$.

- (Nishi, 1984) For fixed p and $n \longrightarrow \infty$, AIC, C_p and CV(1) are asymptotically the same, and all tend to overfit, e.g., the probability of selecting a subset properly containing the true subset converges to a positive number rather than 0. The probability of underfitting converges to 0.

Summary

If we believe that the *correct* model is in fact

$$Y = X_k \beta_k + \epsilon,$$

where X_k is the design matrix with only k variables and the full design matrix is of dimension $p \geq k$.

- (Nishi, 1984) For fixed p and $n \longrightarrow \infty$, AIC, C_p and CV(1) are asymptotically the same, and all tend to overfit, e.g., the probability of selecting a subset properly containing the true subset converges to a positive number rather than 0. The probability of underfitting converges to 0.
- Under the same asymptotics, BIC selects the true model with probability converging to 1.

Summary

If we believe that the *correct* model is in fact

$$Y = X_k \beta_k + \epsilon,$$

where X_k is the design matrix with only k variables and the full design matrix is of dimension $p \geq k$.

- (Nishi, 1984) For fixed p and $n \longrightarrow \infty$, AIC, C_p and CV(1) are asymptotically the same, and all tend to overfit, e.g., the probability of selecting a subset properly containing the true subset converges to a positive number rather than 0. The probability of underfitting converges to 0.
- Under the same asymptotics, BIC selects the true model with probability converging to 1.
- When the number of coefficients is small, the AIC-like criteria tend to overfit.

Summary

If we believe that the *correct* model is in fact

$$Y = X_k \beta_k + \epsilon,$$

where X_k is the design matrix with only k variables and the full design matrix is of dimension $p \geq k$.

- (Nishi, 1984) For fixed p and $n \longrightarrow \infty$, AIC, C_p and CV(1) are asymptotically the same, and all tend to overfit, e.g., the probability of selecting a subset properly containing the true subset converges to a positive number rather than 0. The probability of underfitting converges to 0.
- Under the same asymptotics, BIC selects the true model with probability converging to 1.
- When the number of coefficients is small, the AIC-like criteria tend to overfit.
- Cross-validation asymptotics hold even when the underlying linear model is just an approximation of the true model.

References

- Sandrine Dudoit and Mark J. van der Laan (February 5, 2003) Asymptotics of Cross-Validated Risk Estimation in Model Selection and Performance Assessment. http://www.bepress.com/ucbbiostat/paper126/. [A very comprehensive paper which establishes some of the asymptotic optimality properties of CV.]