Chpater 5 : Infomation Criteria

5.1.Information Criteria & 5.2.Efficient Estimation and the Fisher Information Matrix

Gayoung Moon

2025-03-16

Descendants of Lagrange
School of Mathematics, Statistics and Data Science
Sungshin Women's University

Outline

Information Criterion

- **Information criterion** is an index for evaluating the validity of a statistical model from observation data.

- Information criterion refers to the evaluation of:
  - Fitness: how much the statistical model explains the data.
  - Simplicity: how simple the statistical model is.

One of the Important Problems in Linear Regression

- One of the important problems in **linear regression**:
  To select some p covariates based on N observations
  $(x_1, y_1), \cdots, (x_N, y_N) \in \mathbb{R}^p \times \mathbb{R}$.

- If there are too many covariates, then they overfit the data and try to explain the noise fluctuation by other covariates.

Subset S used as $\mathbf{RSS}(S)$

- It isn't easy to choose $S \subseteq \{1, \cdots, p\}$ from the 2^p subsets when p is large.
  - 2^p subsets: $\{\}, \{1\}, \cdots, \{p\}, \{1, 2\}, \cdots, \{1, \cdots, p\}$.

- We express the fitness and simplicity by the residual sum of square(RSS) value $\mathrm{RSS}(S)$.
  - **RSS**(S) is based on the subset S and the cardinality $k(S) := |S|$ of S.

Property of the Subset S

$$S \subseteq S' \Longrightarrow \begin{cases} \mathrm{RSS}(S) \geq \mathrm{RSS}(S') \\ k(S) \leq k(S') \end{cases} .$$

- It means that the larger the $k = k(S)$, the smaller $\hat{\sigma}_k^2 = \frac{\mathrm{RSS}_k}{N}$ is, where $\mathrm{RSS}_k := \min_{k(S)=k} \mathrm{RSS}(S)$.

Outline

Defination of the AIC and BIC

- Akaike's Information Criterion(AIC) and the Bayesian Information Criterion(BIC) are well known.
- The AIC and BIC are defined by:
  - $\mathrm{AIC} := N \log \hat{\sigma}_k^2 + 2k$.
  - $\mathrm{BIC} := N \log \hat{\sigma}_k^2 + k \log N$.
- The coefficient of determination is:
  - $1 - \frac{\mathrm{RSS}_k}{\mathrm{TSS}}$.
  - It increases monotonically with k and reaches its maximum value at $k = p$.

Properties of the AIC and BIC

- The AIC and BIC values:
  - They decrease before reaching the minimum at some $0 \leq k \leq p$.
  - In the case of $k > p$, they increase.
- The adjusted coefficient of determination maximizes $1 - \frac{\mathrm{RSS}_k/(N-k-1)}{\mathrm{TSS}/(N-1)}$ at some $0 \leq k \leq p$.
  - It is often much larger than those of the AIC and BIC.

Outline

[Example 45] Finding the Set of Covariates that minimizes the AIC and BIC

- In ‘RSS.min(X, y, T)’ function:
  - Input values:
    X: Independent variables matrix $(n \times p)$,
    y: Dependent variable vector $(n \times 1)$,
    T: A combination(matrix) of X to select from.
  - Output values:
    value: Minimum of the RSS,
    set: Combination of X with minimum RSS.
- ‘RSS.min(X, y, T)’ function is using the following formula:

$$RSS = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2.$$

[Example 45] Finding the Set of Covariates that minimizes the AIC and BIC (Contd.)

- In 'AIC(BIC).min' function:
  - 'combn(1:p, k)' generates all combinations of k variables out of a total of p.
  - Output values:
    AIC(BIC).min: Minimum of the AIC/BIC,
    set.min: The combination of variables with the lowest AIC/BIC.
- 'AIC(BIC).min' function is using the following formula:

$$\mathrm{AIC} := N \log \hat{\sigma}_k^2 + 2k,$$

$$\mathrm{BIC} := N \log \hat{\sigma}_k^2 + k \log N.$$

[Example 45] Finding the Set of Covariates that minimizes the AIC and BIC (Contd.)

- In the AIC case:

```
RSS.min=function(X,y,T){
  m=ncol(T); S.min=Inf
  for(j in 1:m){
    q=T[,j]; S=sum((lm(y~X[,q])$fitted.values - y)^2)/n
  if(S<S.min){S.min=S; set.q=q}
  }
return(list(value=S.min,set=set.q))}
library(MASS)   # We use the Boston data set in the R MASS package.
df=Boston; X=as.matrix(df[,c(1,3,5,6,7,8,10,11,12,13)]); y=df[[14]];
# We assume the 'MEDV' variable is responses
# and the remaining variables are covariates.
p=ncol(X); n=length(y)
AIC.min=Inf
for(k in 1:p){
  T=combn(1:p,k); res=RSS.min(X,y,T)
  AIC= n*log(res$value/n)+2*k ##
  if(AIC<AIC.min){AIC.min=AIC; set.min= res$set}}
```

[Example 45] Finding the Set of Covariates that minimizes the AIC and BIC (Contd.)

- AIC:

```
## Warning:    'MASS'  R   4.4.3
```

```
## [1] -1530.84
```

```
## [1]  1  3  4  6  8  9 10
```

- In the BIC case:
  - If we change the line '$n * \log(S.min) + 2 * k$' marked by ## in the AIC code with '$n * \log(S.min) + k * \log(N)$', then the quantity becomes the BIC.

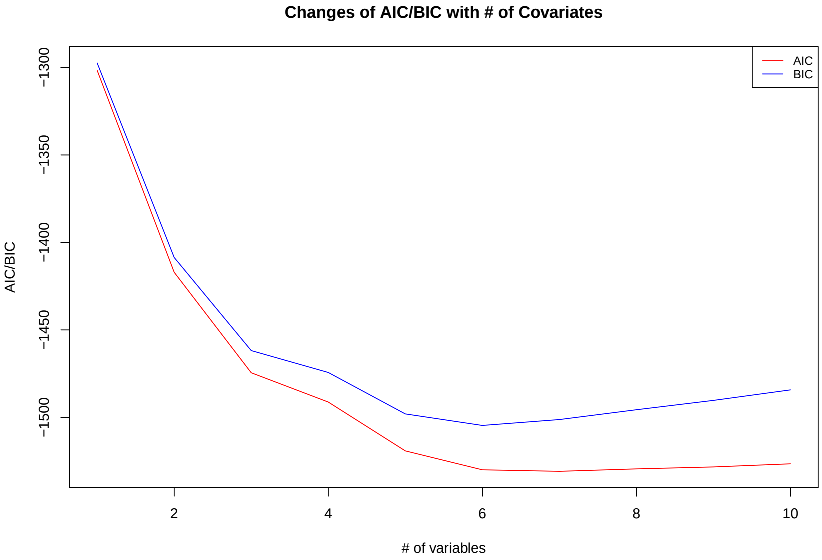
- BIC:

```
## [1] -1504.61
```

```
## [1]  3  4  6  8  9 10
```

- Both AIC and BIC values have negative values.
  → It means that AIC and BIC may not be suitable as criteria for model

[Example 46] The Plot of Changes of AIC/BIC with # of Covariates



- The BIC is larger than the AIC, but the BIC chooses a simpler model with fewer variables than the AIC.

Outline

Probability Density Function of the Observations

- Suppose that:
  - $x_1, \cdots, x_N \in \mathbb{R}^{p+1}$
  - $y_1, \cdots, y_N \in \mathbb{R}$
  - Random variables: $e_1, \cdots, e_n = \varepsilon \sim N(0, \sigma^2)$
  - Unknown constants: $\beta_0 \in \mathbb{R}$, $\beta \in \mathbb{R}^p \rightarrow \beta \in \mathbb{R}^{p+1}$
  - In other words,

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \in \mathbb{R}^{N \times (p+1)},\ y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \in \mathbb{R}^N,\ \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \in \mathbb{R}^{p+1}.$$

- The observations have been generated by the realizations $y_i = x_i \beta + \beta_0 + \varepsilon$, $i = 1, \cdots, N$ with random variables and unknown constants.

Probability Density Function of the Observations (Contd.)

- When $f(y|x,\beta)$ follows a multivariate Gaussian distribution, the probability density function(PDF) can be written as follows:

$$f(y|x,\beta) := \frac{1}{\sqrt{(2\pi\sigma^2)^N}} \exp\Big\{ -\frac{1}{2\sigma^2} \|y - x\beta\|^2 \Big\}.$$

The Value of $\widehat{\beta}^{LSE}$ and the Likelihood

- $y = X\beta + \varepsilon$
- In the **least squares method**, we estimated β by:

$$\widehat{\beta}^{LSE} = (X^\top X)^{-1} X^\top y.$$

- $\widehat{\beta}^{LSE}$ coincides with the $\beta \in \mathbb{R}^{p+1}$ that maximizes the likelihood

$$\begin{aligned} L :&= \prod_{i=1}^{N} f(y_i | x_i, \beta) \\ &= \prod_{i=1}^{N} \left(\frac{1}{\sqrt{(2\pi\sigma^2)}} \exp\left\{ -\frac{1}{2\sigma^2} (y_i - x_i\beta)^2 \right\} \right). \end{aligned}$$

Caculating log-likelihood

- The log-likelihood is written by:

$$
\begin{aligned}
\ell :&= \log L \\
&= \log \prod_{i=1}^{N} \left(\frac{1}{\sqrt{(2\pi\sigma^2)}} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - x_i\beta)^2 \right\} \right) \\
&= \log(2\pi\sigma^2)^{-\frac{N}{2}} + \log \prod_{i=1}^{N} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - x_i\beta)^2 \right\} \\
&= -\frac{N}{2}\log(2\pi\sigma^2) - \prod_{i=1}^{N} \frac{1}{2\sigma^2}(y_i - x_i\beta)^2 \\
&= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|y - X\beta\|^2.
\end{aligned}
$$

- If $\sigma^2 > 0$ is fixed,
  maximizing l is equivalent to minimizing $\|y - X\beta\|^2$.

Differentiate log-likelihood with σ^2

- If we partially differentiate ℓ w.r.t. σ^2:

$$
\begin{aligned}
\frac{\partial \ell}{\partial \sigma^2} &= \frac{\partial}{\partial \sigma^2}\Big(-\frac{N}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\|y - X\beta\|^2\Big) \\
&= \frac{\partial}{\partial \sigma^2}\Big(-\frac{N}{2}\log\left(2\pi\sigma^2\right)\Big) - \frac{\partial}{\partial \sigma^2}\Big(\frac{1}{2\sigma^2}\|y - X\beta\|^2\Big) \\
&= -\frac{N}{2}\frac{1}{2\pi\sigma^2}2\pi - \|y - X\beta\|^2\big(-\frac{1}{2}\big)\frac{1}{(\sigma^2)^2} \\
&= -\frac{N}{2\sigma^2} + \frac{\|y - X\beta\|^2}{2(\sigma^2)^2} = 0.
\end{aligned}
$$

Differentiate log-likelihood with σ^2 (Contd.)

- Using $\hat{\beta} = (X^\top X)^{-1} X^\top y$, we find:

$$\hat{\sigma}^2 := \frac{1}{N} \|y - X\hat{\beta}\|^2 = \frac{1}{N} \|y - \hat{y}\|^2 = \frac{RSS}{N}.$$

- It is the maximum likelihood estimate of $\hat{\sigma}^2$.

Efficient Estimator and the Fisher Information Matrix

- Efficient Estimator:
  - When there are multiple unbiased estimators, the estimator with the smallest variance is called an efficient estimator.

- Let $\nabla\ell$ be the vector consisting of $\frac{\partial\ell}{\partial\beta_j}$, $j = 0, 1, \cdots, p$,
  $\rightarrow$ The fisher information matrix:
  The covariance matrix J of $\nabla\ell$ divided by N.

Differentiation of the Score Function

- For $f^N(y|x,\beta) := \prod_{i=1}^N f(y_i|x_i,\beta)$,

$$
\begin{aligned}
\nabla \ell &= \frac{\nabla f^N(y|x,\beta)}{f^N(y|x,\beta)} \\
&= \frac{\nabla \prod_{i=1}^N f(y_i|x_i,\beta)}{f^N(y|x,\beta)} \\
&= \frac{\nabla \prod_{i=1}^N \left(\frac{1}{\sqrt{(2\pi\sigma^2)}} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - x_i\beta)^2 \right\} \right)}{f^N(y|x,\beta)} \\
&= \frac{f^N(y|x,\beta) \cdot \nabla \left(-\frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - x_i\beta)^2 \right)}{f^N(y|x,\beta)} \\
&= \frac{f^N(y|x,\beta) \cdot \left(-\frac{1}{\sigma^2} \sum_{i=1}^N (y_i - x_i\beta) x_i \right)}{f^N(y|x,\beta)} \\
&= -\frac{1}{\sigma^2} \sum_{i=1}^N (y_i - x_i\beta) x_i.
\end{aligned}
$$

Integral of $\nabla f^N(y|x, \beta)$

- If we partially differentiate both sides of $\int f^N(y|x,\beta)dy = 1$ w.r.t. β, we have that $\int \nabla f^N(y|x,\beta)dy = 0$.

$$
\begin{aligned}
\int \nabla f^N(y|x,\beta)\, dy &= \int \nabla \prod_{i=1}^{N} \left(\frac{1}{\sqrt{(2\pi\sigma^2)}} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - x_i\beta)^2 \right\} \right) dy \\
&= \int f^N(y|x,\beta) \cdot \nabla \left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - x_i\beta)^2 \right) dy \\
&= \int f^N(y|x,\beta) \cdot \left(-\frac{1}{\sigma^2} \sum_{i=1}^{N} (y_i - x_i\beta)x_i \right) dy \\
&= -\frac{1}{\sigma^2} \sum_{i=1}^{N} x_i \int (y_i - x_i\beta)\, f^N(y|x,\beta)\, dy \\
&= -\frac{1}{\sigma^2} \sum_{i=1}^{N} x_i\, E[y_i - x_i\beta] \\
&= -\frac{1}{\sigma^2} \sum_{i=1}^{N} x_i\, (x_i\beta - x_i\beta) = 0.
\end{aligned}
$$

Expectation of the Score Function

- We have that:

$$
\begin{aligned}
E\nabla\ell &= \int \nabla l \cdot f^N(y|x,\beta)\,dy \\
&= \int \frac{\nabla f^N(y|x,\beta)}{f^N(y|x,\beta)} f^N(y|x,\beta)\,dy \\
&= \int \left(-\frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i - x_i\beta)x_i \right) f^N(y|x,\beta)\,dy \\
&= -\frac{1}{\sigma^2}\sum_{i=1}^{N} x_i \int (y_i - x_i\beta)\, f^N(y|x,\beta)\,dy \\
&= -\frac{1}{\sigma^2}\sum_{i=1}^{N} x_i\, E[y_i - x_i\beta] \; = -\frac{1}{\sigma^2}\sum_{i=1}^{N} x_i\left(E[y_i] - x_i\beta\right) \\
&= -\frac{1}{\sigma^2}\sum_{i=1}^{N} x_i\left(x_i\beta - x_i\beta\right) \; = \int \nabla f^N(y|x,\beta)\,dy \; = 0.
\end{aligned}
$$

- And

$$
\begin{aligned}
0 &= \nabla \otimes [E\nabla\ell] \\
&= \nabla \otimes \int (\nabla l) f^N(y|x,\beta) dy \\
&= \int (\nabla^2 l) f^N(y|x,\beta) dy + \int (\nabla l)\{\nabla f^N(y|x,\beta)\} dy \\
&= \int \nabla \Big\{ -\frac{1}{\sigma^2} \sum_{i=1}^N (y_i - x_i\beta) x_i \Big\} f^N(y|x,\beta)\, dy + \int \Big\{ -\frac{1}{\sigma^2} \sum_{i=1}^N (y_i - x_i\beta) x_i \Big\} \{\nabla f^N(y|x,\beta)\} dy \\
&= \int \frac{1}{\sigma^2} \sum_{i=1}^N x_i x_i^\top\, f^N(y|x,\beta)\, dy\, + \int \Big\{ -\frac{1}{\sigma^2} \sum_{i=1}^N (y_i - x_i\beta) x_i \Big\} \Big\{ -\frac{1}{\sigma^2} \sum_{i=1}^N (y_i - x_i\beta) x_i \Big\}\, dy \\
&= \frac{1}{\sigma^2} \sum_{i=1}^N x_i x_i^\top \int f^N(y|x,\beta)\, dy\, + \int \Big\{ -\frac{1}{\sigma^2} \sum_{i=1}^N (y_i - x_i\beta) x_i \Big\}^2 dy \\
&= \frac{1}{\sigma^2} \sum_{i=1}^N x_i x_i^\top\, + \int \Big\{ -\frac{1}{\sigma^2} \sum_{i=1}^N (y_i - x_i\beta) x_i \Big\}^2 dy \\
&= E[\nabla^2\ell] + E[(\nabla\ell)^2].
\end{aligned}
$$

- We can verify that it is as follows:

$$E[\nabla^2\ell] + E[(\nabla\ell)^2] = 0.$$

$$\therefore E[(\nabla\ell)^2] = -E[\nabla^2\ell].$$

- Then, the above equation implies that:

$$J = \frac{1}{N}E[(\nabla\ell)^2] = -\frac{1}{N}E[\nabla^2\ell].$$

Outline

$Cram\acute{e}r - Rao$ **Inequality**

- $Cram\acute{e}r - Rao$ inequality is:
    - Inequality that gives a lower bound on the variance of the discomfort estimate.
    - It is expressed as the inverse matrix of the fisher information matrix.

$$Var(\tilde{\beta}) \geq (NJ)^{-1}.$$

- $\tilde{\beta} \in \mathbb{R}^{(p+1)}$: unbiased estimator.
- $(NJ)^{-1}$: the fisher information matrix $\in \mathbb{R}^{(p+1)\times(p+1)}$.

Definition of log-likelihood Function and Score Function

- We defined the log-likelihood function above as follows:

$$\ell := \log L = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|y - X\beta\|^2.$$

- And we can also define the score function as the slope of the log-likelihood:

$$\nabla\ell := -\frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i - x_i\beta)x_i.$$

Definition of the Fisher Information Matrix

- The fisher information matrix J is expressed as the square of the expectation of the score function.

$$J = \frac{1}{N} E[(\nabla \ell)^2] = \frac{1}{N\sigma^4} E[\sum_{i=1}^{N} (y_i - x_i \beta)^2 x_i]$$

- In Gaussian distribution, $E[(y_i - x_i \beta)^2] = \sigma^2$:

$$J = \frac{1}{N\sigma^2} \sum_{i=1}^{N} x_i.$$

Propositoin of $Cramér - Rao$ Inequality

- The least squares estimate satisfies the equality part of the inequality.

- We know $\int f^N(y|x,\beta) = 1$ and this end, if we partially differentiate both sides of

$$\int \tilde{\beta}_i f^N(y|x,\beta)dy = \beta_i$$

w.r.t β_j, we have the following equation:

$$\int \tilde{\beta}_i \frac{\partial}{\partial \beta_j} f^N(y|x,\beta)dy = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases} \quad .$$

Propositoin of $Cramér - Rao$ Inequality (Contd.)

- If we write this equation in terms of its covariance matrix, we have that $E[\tilde{\beta}(\nabla\ell)^\top] = I$, where I is a unit matrix of size $(p+1)$.

- And we know $E[\nabla\ell] = 0$, we rewrite the above equation as follows:

$$E[(\tilde{\beta} - \beta)(\nabla\ell)^\top] = I.$$

Propositoin of $Cramér - Rao$ Inequality (Contd.)

- Then, the covariance matrix of the vector of size $2(p+1)$ is:

$$\begin{bmatrix} V(\tilde{\beta}) & I \\ I & NJ \end{bmatrix}.$$

- Because both $V(\tilde{\beta})$ and J are covariance matrices, they are non-negative definite.
    - Non-negative definite:
      Let A be an $n \times n$ real symmetric matrix, A is **non-negative definite** if:

$$xAx \geq 0, \text{ for any } x \in \mathbb{R}^n.$$

Propositoin of $Cramér - Rao$ Inequality (Contd.)

- Then, we claim that both sides of matrixes are non-negative definite:

$$\begin{bmatrix} V(\tilde{\beta}) - (NJ)^{-1} & 0 \\ 0 & NJ \end{bmatrix} = \begin{bmatrix} I & -(NJ)^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} V(\tilde{\beta}) & I \\ I & NJ \end{bmatrix} \begin{bmatrix} I & 0 \\ -(NJ)^{-1} & I \end{bmatrix}.$$

- For an arbitrary $x \in \mathbb{R}^n$, if $xAx \geq 0$, for an arbitrary $B \in \mathbb{R}^{n \times m}$ and $y \in \mathbb{R}^m$, we have that $yBABy \geq 0$, which means that $V(\tilde{\beta}) - (NJ)^{-1}$ is non-negative definite.

- So, we have the conclusion:

$$\therefore V(\tilde{\beta}) \geq (NJ)^{-1}.$$

Q & A

Thank you :)