Multiple regression - information criteria for large data bases

Małgorzata Bogdan

University of Wroclaw

10 March 2023

Multiple regression model when $n > p$

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}, \quad \epsilon \sim N(0, \sigma^2 I_{n\times n})$$

,

Multiple regression model when $n > p$

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}, \quad \epsilon \sim N(0, \sigma^2 I_{n\times n})$$

,

$$\hat{\beta}_{LS} = argmin_{\beta \in R^p} ||Y - X\beta||^2 = (X'X)^{-1}X'Y$$

Multiple regression model when $n > p$

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}, \quad \epsilon \sim N(0, \sigma^2 I_{n\times n})$$

,

$$\hat{\beta}_{LS} = argmin_{\beta\in R^p}||Y - X\beta||^2 = (X'X)^{-1}X'Y$$

$$\hat{\beta}_{LS} \sim N(\beta, \sigma^2(X'X)^{-1})$$

Multiple regression model when $n > p$

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}, \quad \epsilon \sim N(0, \sigma^2 I_{n\times n})$$

,

$$\hat{\beta}_{LS} = argmin_{\beta\in R^p}||Y - X\beta||^2 = (X'X)^{-1}X'Y$$

$$\hat{\beta}_{LS} \sim N(\beta, \sigma^2(X'X)^{-1})$$

$$\hat{\sigma}^2 = s^2 = \frac{||Y - X\hat{\beta}_{LS}||^2}{n-p} = \frac{RSS}{n-p}$$

Selection of important variables

T-tests,

$$T_i = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)} \quad ,$$

where $s(\hat{\beta}_i) = s^2(X'X)^{-1}[i,i]$

Selection of important variables

T-tests,

$$T_i = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)} \quad ,$$

where $s(\hat{\beta}_i) = s^2(X'X)^{-1}[i, i]$

Problem - typically elements on the diagonal of $(X'X)^{-1}$ become large as p increases.

Selection of important variables

T-tests,

$$T_i = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)} \quad ,$$

where $s(\hat{\beta}_i) = s^2(X'X)^{-1}[i,i]$

Problem - typically elements on the diagonal of $(X'X)^{-1}$ become large as p increases.

If elements of X are iid from $N(0, 1/\sqrt{n})$ then $X'X$ has a Wishart distribution and the elements on its diagonal have the expected value equal to 1.

T-tests,

$$T_i = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)} \quad ,$$

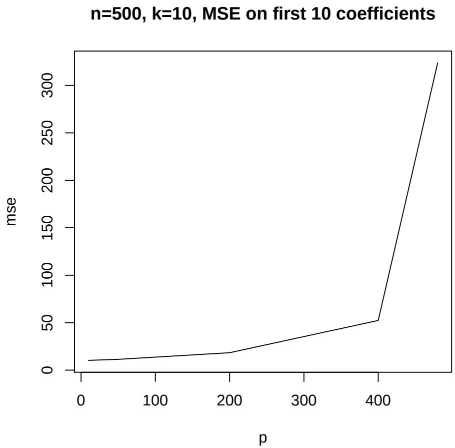
where $s(\hat{\beta}_i) = s^2(X'X)^{-1}[i, i]$

Problem - typically elements on the diagonal of $(X'X)^{-1}$ become large as p increases.

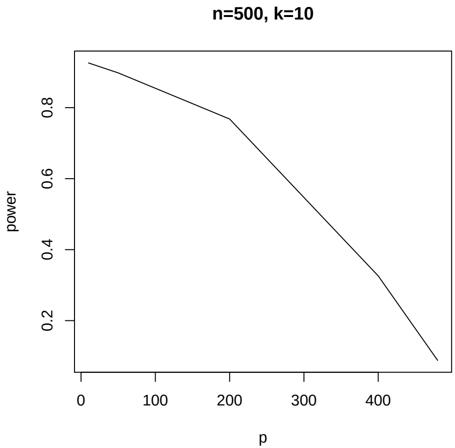
If elements of X are iid from $N(0, 1/\sqrt{n})$ then $X'X$ has a Wishart distribution and the elements on its diagonal have the expected value equal to 1.

But $(X'X)^{-1}$ has the inverse Wishart distribution and the expected values of the elements on the diagonal are equal to $\frac{n}{n-p-1}$ and become very large as p approaches n.

Inflation of MSE



Loss of Power



Model selection in multiple regression - identification of important variables

Model selection

Model selection in multiple regression - identification of important variables

Error in the training sample $RSS = ||Y - \hat{Y}||^2$ never increases when we add new variables into the model. Thus, minimization of RSS is not a good criterion for model selection.

Model selection

Model selection in multiple regression - identification of important variables

Error in the training sample $RSS = ||Y - \hat{Y}||^2$ never increases when we add new variables into the model. Thus, minimization of RSS is not a good criterion for model selection.

Also, RSS is not a good measure of the prediction error.

Training and prediction error

Let's consider a new sample

$$Y^* = X\beta + \epsilon^* \quad ,$$

where ϵ^* is independent on the noise term ϵ in the training sample

Training and prediction error

Let's consider a new sample

$$Y^* = X\beta + \epsilon^* \quad ,$$

where ϵ^* is independent on the noise term ϵ in the training sample

We use our training sample to build a good predictive model, i.e. the model which minimizes

$$PE = E||Y^* - \hat{Y}||^2$$

Training and prediction error

Let's consider a new sample

$$Y^* = X\beta + \epsilon^* \quad ,$$

where ϵ^* is independent on the noise term ϵ in the training sample

We use our training sample to build a good predictive model, i.e. the model which minimizes

$$PE = E||Y^* - \hat{Y}||^2$$

If $\mu = E(Y) = X\beta$, then $PE = E||\mu - \hat{\mu}||^2 + n\sigma^2 = E||\mu - \hat{Y}||^2 + n\sigma^2$

Training and prediction error

Let's consider a new sample

$$Y^* = X\beta + \epsilon^* \quad ,$$

where ϵ^* is independent on the noise term ϵ in the training sample

We use our training sample to build a good predictive model, i.e. the model which minimizes

$$PE = E||Y^* - \hat{Y}||^2$$

If $\mu = E(Y) = X\beta$, then $PE = E||\mu - \hat{\mu}||^2 + n\sigma^2 = E||\mu - \hat{Y}||^2 + n\sigma^2$

RSS measures the fit within the training sample, i.e. it adjusts to the specific realization of the noise term ϵ - this is overfitting. PE measures the fit with respect to the true expected value of Y, which indeed is an indication of predictive properties (i.e. how well we can predict new observations with different noise terms).

Prediction error of linear operators

If $\hat{Y} = M_{n\times n} Y$ then
$PE = E(RSS) + 2\sigma^2 Tr(M)$

Prediction error of linear operators

If $\hat{Y} = M_{n \times n} Y$ then
$PE = E(RSS) + 2\sigma^2 Tr(M)$
Proof by Stein's identity :$\hat{\mu} = \hat{Y} + Y - Y$

Prediction error of linear operators

If $\hat{Y} = M_{n\times n}Y$ then
$PE = E(RSS) + 2\sigma^2 Tr(M)$
Proof by Stein's identity :$\hat{\mu} = \hat{Y} + Y - Y$
$g(Y) = \hat{Y} - Y = MY - Y$

Prediction error of linear operators

If $\hat{Y} = M_{n \times n} Y$ then
$PE = E(RSS) + 2\sigma^2 Tr(M)$
Proof by Stein's identity :$\hat{\mu} = \hat{Y} + Y - Y$
$g(Y) = \hat{Y} - Y = MY - Y$
$E||\hat{\mu} - \mu||^2 = n\sigma^2 + E(||g(Y)||^2 + 2\sigma^2 div\ g(Y))$

Prediction error of linear operators

If $\hat{Y} = M_{n \times n} Y$ then
$PE = E(RSS) + 2\sigma^2 Tr(M)$
Proof by Stein's identity :$\hat{\mu} = \hat{Y} + Y - Y$
$g(Y) = \hat{Y} - Y = MY - Y$
$E||\hat{\mu} - \mu||^2 = n\sigma^2 + E(||g(Y)||^2 + 2\sigma^2 div\ g(Y))$
$||g(Y)||^2 = RSS$

Prediction error of linear operators

If $\hat{Y} = M_{n\times n} Y$ then

$PE = E(RSS) + 2\sigma^2 Tr(M)$

Proof by Stein's identity :$\hat{\mu} = \hat{Y} + Y - Y$

$g(Y) = \hat{Y} - Y = MY - Y$

$E||\hat{\mu} - \mu||^2 = n\sigma^2 + E(||g(Y)||^2 + 2\sigma^2 div\ g(Y))$

$||g(Y)||^2 = RSS$

$div\ g(Y) = TrM - n$

Prediction error of linear operators

If $\hat{Y} = M_{n\times n} Y$ then
$PE = E(RSS) + 2\sigma^2 Tr(M)$
Proof by Stein's identity :$\hat{\mu} = \hat{Y} + Y - Y$
$g(Y) = \hat{Y} - Y = MY - Y$
$E||\hat{\mu} - \mu||^2 = n\sigma^2 + E(||g(Y)||^2 + 2\sigma^2 div\ g(Y))$
$||g(Y)||^2 = RSS$
$div\ g(Y) = TrM - n$
$PE = n\sigma^2 + E(SURE(\hat{\mu})) = n\sigma^2 + E(RSS) + 2\sigma^2 TrM - n\sigma^2$

Prediction error in least squares regression

In least squares estimation

$$M = X(X'X)^{-1}X'$$

is the matrix of the orthogonal projection on the space spanned by columns of X and $Tr(M) = rank(X)$.

Prediction error in least squares regression

In least squares estimation

$$M = X(X'X)^{-1}X'$$

is the matrix of the orthogonal projection on the space spanned by columns of X and $Tr(M) = rank(X)$.

If $rank(X) = p$ then the unbiased estimator of the prediction error is equal to

$$\hat{P}E = RSS + 2\sigma^2 p \quad .$$

Prediction error in least squares regression

In least squares estimation

$$M = X(X'X)^{-1}X'$$

is the matrix of the orthogonal projection on the space spanned by columns of X and $Tr(M) = rank(X)$.

If $rank(X) = p$ then the unbiased estimator of the prediction error is equal to

$$\hat{P}E = RSS + 2\sigma^2 p \ \ .$$

Minimizing $\hat{P}E$ coincides with AIC criterion which suggests selecting the model for which $RSS + 2\sigma^2 p$ is minimal.

Prediction error in least squares regression

In least squares estimation

$$M = X(X'X)^{-1}X'$$

is the matrix of the orthogonal projection on the space spanned by columns of X and $Tr(M) = rank(X)$.

If $rank(X) = p$ then the unbiased estimator of the prediction error is equal to

$$\hat{P}E = RSS + 2\sigma^2 p \ .$$

Minimizing $\hat{P}E$ coincides with AIC criterion which suggests selecting the model for which $RSS + 2\sigma^2 p$ is minimal.

Leave-one-out cross-validation:

$$CV = \sum_{i=1}^{n} (Y_i - \hat{Y}[i])^2 = \sum_{i=1}^{n} \left(\frac{Y_i - \hat{Y}_i}{1 - M_{ii}} \right)^2$$

Akaike Information Criterion

$X = (X_1, \ldots, X_n)$ - vector of iid random variables from the model M_k: $f(x,\theta)$, $\theta \in R^k$

Akaike Information Criterion

$X = (X_1, \ldots, X_n)$ - vector of iid random variables from the model M_k: $f(x, \theta)$, $\theta \in R^k$

$$L(X, \theta) = \prod_{i=1}^{n} f(X_i, \theta)$$

Akaike Information Criterion

$X = (X_1, \ldots, X_n)$ - vector of iid random variables from the model M_k: $f(x, \theta)$, $\theta \in R^k$

$$L(X, \theta) = \prod_{i=1}^{n} f(X_i, \theta)$$

$$AIC(M_k) = \ln L(X, \hat{\theta}_{MLE}) - k$$

Akaike Information Criterion in Linear Regression, σ known

$\epsilon_1 = Y_1 - X_1\beta, \ldots, \epsilon_n = Y_n - X_n\beta$ - iid from $N(0, \sigma^2)$, $\beta \in R^k$

Akaike Information Criterion in Linear Regression, σ known

$\epsilon_1 = Y_1 - X_1\beta, \ldots, \epsilon_n = Y_n - X_n\beta$ - iid from $N(0, \sigma^2)$, $\beta \in R^k$

$$L(Y|X, \beta, \sigma) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{\frac{-||Y - X\beta||^2}{2\sigma^2}}$$

$\epsilon_1 = Y_1 - X_1\beta, \ldots, \epsilon_n = Y_n - X_n\beta$ - iid from $N(0,\sigma^2)$, $\beta \in R^k$

$$L(Y|X,\beta,\sigma) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{\frac{-||Y-X\beta||^2}{2\sigma^2}}$$

$$\ln L(Y|X,\beta,\sigma) = C - n\log(\sigma) - \frac{||Y-X\beta||^2}{2\sigma^2}$$

Akaike Information Criterion in Linear Regression, σ known

$$\epsilon_1 = Y_1 - X_1\beta, \ldots, \epsilon_n = Y_n - X_n\beta \text{ - iid from } N(0,\sigma^2),\ \beta \in R^k$$

$$L(Y|X,\beta,\sigma) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{\frac{-||Y-X\beta||^2}{2\sigma^2}}$$

$$\ln L(Y|X,\beta,\sigma) = C - n\log(\sigma) - \frac{||Y-X\beta||^2}{2\sigma^2}$$

$$AIC(M_k) = C(n,\sigma) - \frac{RSS}{2\sigma^2} - k$$

Akaike Information Criterion in Linear Regression, σ known

$\epsilon_1 = Y_1 - X_1\beta, \ldots, \epsilon_n = Y_n - X_n\beta$ - iid from $N(0, \sigma^2)$, $\beta \in R^k$

$$L(Y|X, \beta, \sigma) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{\frac{-||Y - X\beta||^2}{2\sigma^2}}$$

$$\ln L(Y|X, \beta, \sigma) = C - n\log(\sigma) - \frac{||Y - X\beta||^2}{2\sigma^2}$$

$$AIC(M_k) = C(n, \sigma) - \frac{RSS}{2\sigma^2} - k$$

Maximizing AIC corresponds to minimizing $RSS + 2\sigma^2 k$

Akaike Information Criterion in Linear Regression, σ unknown

$$\hat{\sigma}^2_{MLE} = \frac{RSS}{n}$$

Akaike Information Criterion in Linear Regression, σ unknown

$$\hat{\sigma}^2_{MLE} = \frac{RSS}{n}$$

$$\ln L(Y|X, \hat{\beta}, \hat{\sigma}) = C - n/2 \log(RRS/n) - \frac{RSS}{2}\frac{n}{RSS}$$

Akaike Information Criterion in Linear Regression, σ unknown

$$\hat{\sigma}^2_{MLE} = \frac{RSS}{n}$$

$$\ln L(Y|X, \hat{\beta}, \hat{\sigma}) = C - n/2 \log(RRS/n) - \frac{RSS}{2}\frac{n}{RSS}$$

$$AIC(M_k) = C(n) - n/2 \log(RSS) - k$$

Akaike Information Criterion in Linear Regression, σ unknown

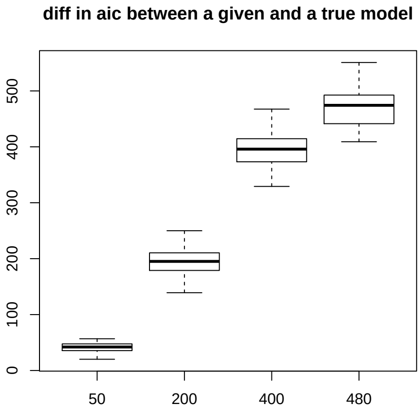
$$\hat{\sigma}^2_{MLE} = \frac{RSS}{n}$$

$$\ln L(Y|X, \hat{\beta}, \hat{\sigma}) = C - n/2 \log(RRS/n) - \frac{RSS}{2} \frac{n}{RSS}$$

$$AIC(M_k) = C(n) - n/2 \log(RSS) - k$$

Maximizing AIC corresponds to minimizing $n \log(RSS) + 2k$

In our example AIC identifies the true model among 5 models with different dimensions, $p = 500$, $k = 10$.



Can we use AIC to select variables in large data bases ?

Problem 1: Discrete optimization over 2^p of possible models - not doable in polynomial time.

Can we use AIC to select variables in large data bases ?

Problem 1: Discrete optimization over 2^p of possible models - not doable in polynomial time.

In practice we often resort to heuristics which with large probability return models closed to being optimal.

Can we use AIC to select variables in large data bases ?

Problem 1: Discrete optimization over 2^p of possible models - not doable in polynomial time.

In practice we often resort to heuristics which with large probability return models closed to being optimal.

Forward selection - we start from the empty model and add variables one by one. At each step we select the one which leads to the largest improvement of the criterion. We stop when the criterion is no longer improved.

Can we use AIC to select variables in large data bases ?

Problem 1: Discrete optimization over 2^p of possible models - not doable in polynomial time.

In practice we often resort to heuristics which with large probability return models closed to being optimal.

Forward selection - we start from the empty model and add variables one by one. At each step we select the one which leads to the largest improvement of the criterion. We stop when the criterion is no longer improved.

Backward elimination - we start from the full model and remove variables one by one until criterion is no longer improved.

Can we use AIC to select variables in large data bases ?

Problem 1: Discrete optimization over 2^p of possible models - not doable in polynomial time.

In practice we often resort to heuristics which with large probability return models closed to being optimal.

Forward selection - we start from the empty model and add variables one by one. At each step we select the one which leads to the largest improvement of the criterion. We stop when the criterion is no longer improved.

Backward elimination - we start from the full model and remove variables one by one until criterion is no longer improved.

Step-wise selection: alternating between forward selection and backward elimination

Can we use AIC to select variables in large data bases ?

Problem 1: Discrete optimization over 2^p of possible models - not doable in polynomial time.

In practice we often resort to heuristics which with large probability return models closed to being optimal.

Forward selection - we start from the empty model and add variables one by one. At each step we select the one which leads to the largest improvement of the criterion. We stop when the criterion is no longer improved.

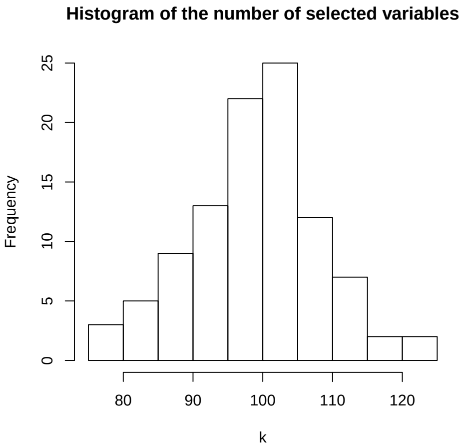
Backward elimination - we start from the full model and remove variables one by one until criterion is no longer improved.

Step-wise selection: alternating between forward selection and backward elimination

More complicated heuristics: genetic algorithms, simulated annealing etc.

Can we use AIC to select important variables in large data bases ?

bigstep - *R* library with many different search strategies, optimizing a variety of model selection criteria; $p = 500$, $k = 10$.



Multiple testing explanation (1)

Assume that $X'X = I$

Multiple testing explanation (1)

Assume that $X'X = I$

$$\hat{\beta} = (X'X)^{-1}X'Y = X'Y, \ \ \hat{\beta}' = Y'X$$

Multiple testing explanation (1)

Assume that $X'X = I$

$$\hat{\beta} = (X'X)^{-1}X'Y = X'Y, \;\; \hat{\beta}' = Y'X$$

$$RSS = (Y - X\hat{\beta})'(Y - X\hat{\beta}) = Y'Y + \hat{\beta}'X'X\hat{\beta} - 2Y'X\hat{\beta}$$

Multiple testing explanation (1)

Assume that $X'X = I$

$$\hat{\beta} = (X'X)^{-1}X'Y = X'Y, \quad \hat{\beta}' = Y'X$$

$$RSS = (Y - X\hat{\beta})'(Y - X\hat{\beta}) = Y'Y + \hat{\beta}'X'X\hat{\beta} - 2Y'X\hat{\beta}$$

$$RSS = Y'Y - \hat{\beta}'\hat{\beta} = Y'Y - \sum_{i=1}^{k} \hat{\beta}_i^2$$

Multiple testing explanation (2)

Thus AIC selects variables which satisfy

$$|\hat{\beta}_i| \geq \sqrt{2}\sigma \quad .$$

Multiple testing explanation (2)

Thus AIC selects variables which satisfy

$$|\hat{\beta}_i| \geq \sqrt{2}\sigma \quad .$$

When $\beta_i = 0$ then $\hat{\beta}_i \sim N(0, \sigma^2)$.

Multiple testing explanation (2)

Thus AIC selects variables which satisfy

$$|\hat{\beta}_i| \geq \sqrt{2}\sigma \quad .$$

When $\beta_i = 0$ then $\hat{\beta}_i \sim N(0, \sigma^2)$.

Thus probability of the type I error

$$P(X_i \text{ is selected} | \beta_i = 0) = 2(1 - \Phi(\sqrt{2})) = 0.16$$

Multiple testing explanation (2)

Thus AIC selects variables which satisfy

$$|\hat{\beta}_i| \geq \sqrt{2}\sigma \quad .$$

When $\beta_i = 0$ then $\hat{\beta}_i \sim N(0, \sigma^2)$.

Thus probability of the type I error

$$P(X_i \text{ is selected}|\beta_i = 0) = 2(1 - \Phi(\sqrt{2})) = 0.16$$

When $p = 500$ and $k = 10$ we expect to see on average $490 \times 0.16 = 78$ false discoveries and the typical size of the selected model should be around k=88

Multiple testing explanation (2)

Thus AIC selects variables which satisfy

$$|\hat{\beta}_i| \geq \sqrt{2}\sigma \quad .$$

When $\beta_i = 0$ then $\hat{\beta}_i \sim N(0, \sigma^2)$.

Thus probability of the type I error

$$P(X_i \text{ is selected} | \beta_i = 0) = 2(1 - \Phi(\sqrt{2})) = 0.16$$

When $p = 500$ and $k = 10$ we expect to see on average $490 \times 0.16 = 78$ false discoveries and the typical size of the selected model should be around k=88

In our simulations $\hat{k} \approx 100$ due to additional disturbance by the sample correlations between columns of the design matrix and using the form of AIC with unknown σ

Would BIC help ?

BIC selects the model which minimizes

$$RSS + \sigma^2 k \log n$$

Would BIC help ?

BIC selects the model which minimizes

$$RSS + \sigma^2 k \log n$$

Thus BIC selects variables which satisfy

$$|\hat{\beta}_i| \geq \sqrt{\log n}\sigma \quad .$$

Would BIC help ?

BIC selects the model which minimizes

$$RSS + \sigma^2 k \log n$$

Thus BIC selects variables which satisfy

$$|\hat{\beta}_i| \geq \sqrt{\log n}\sigma \quad .$$

The probability of the type I error

$$P(X_i \text{ is selected}|\beta_i = 0) = 2(1 - \Phi(\sqrt{\log n}),$$

Would BIC help ?

BIC selects the model which minimizes

$$RSS + \sigma^2 k \log n$$

Thus BIC selects variables which satisfy

$$|\hat{\beta}_i| \geq \sqrt{\log n}\sigma \quad .$$

The probability of the type I error

$$P(X_i \text{ is selected}|\beta_i = 0) = 2(1 - \Phi(\sqrt{\log n}),$$

which for $n = 500$ is equal to 0.013

Would BIC help ?

BIC selects the model which minimizes

$$RSS + \sigma^2 k \log n$$

Thus BIC selects variables which satisfy

$$|\hat{\beta}_i| \geq \sqrt{\log n}\sigma \quad .$$

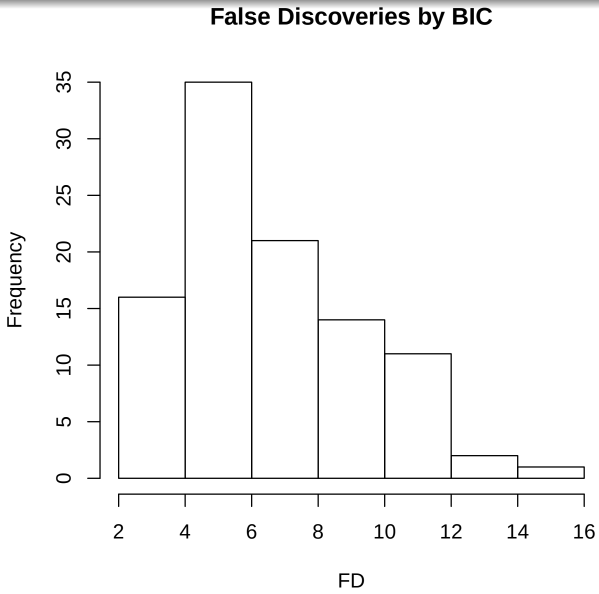
The probability of the type I error

$$P(X_i \text{ is selected} | \beta_i = 0) = 2(1 - \Phi(\sqrt{\log n}),$$

which for $n = 500$ is equal to 0.013

Thus we expect to see on average $p_0 * 0.013 = 490 * 0.013 \approx 6.5$ false discoveries

False Discoveries by BIC



Solution - multiple testing correction

In Risk Inflation Criterion (Foster and George 1994) the penalty depends on p

$$RSS + \sigma^2 2k \log p$$

In Risk Inflation Criterion (Foster and George 1994) the penalty depends on p

$$RSS + \sigma^2 2k \log p$$

Thus RIC selects variables which satisfy

$$|\hat{\beta}_i| \geq \sigma\sqrt{2\log p} \quad .$$

Solution - multiple testing correction

In Risk Inflation Criterion (Foster and George 1994) the penalty depends on p

$$RSS + \sigma^2 2k \log p$$

Thus RIC selects variables which satisfy

$$|\hat{\beta}_i| \geq \sigma\sqrt{2\log p} \ .$$

The probability of the type I error

$$P(X_i \text{ is selected}|\beta_i = 0) = 2(1 - \Phi(\sqrt{2\log p})) \approx \frac{1}{\sqrt{\pi}}\frac{1}{p\sqrt{\log p}}.$$

In Risk Inflation Criterion (Foster and George 1994) the penalty depends on p

$$RSS + \sigma^2 2k \log p$$

Thus RIC selects variables which satisfy

$$|\hat{\beta}_i| \geq \sigma\sqrt{2\log p} \ .$$

The probability of the type I error

$$P(X_i \text{ is selected}|\beta_i = 0) = 2(1 - \Phi(\sqrt{2\log p})) \approx \frac{1}{\sqrt{\pi}}\frac{1}{p\sqrt{\log p}}.$$

Accuracy of approximation: for $p = 500$

$$2(1 - \Phi(\sqrt{2\log p}) = 0.000423, \quad \frac{1}{\sqrt{\pi}}\frac{1}{p\sqrt{\log p}} = 0.000453$$

Solution - multiple testing correction

In Risk Inflation Criterion (Foster and George 1994) the penalty depends on p

$$RSS + \sigma^2 2k \log p$$

Thus RIC selects variables which satisfy

$$|\hat{\beta}_i| \geq \sigma\sqrt{2\log p} \quad .$$

The probability of the type I error

$$P(X_i \text{ is selected}|\beta_i = 0) = 2(1 - \Phi(\sqrt{2\log p})) \approx \frac{1}{\sqrt{\pi}}\frac{1}{p\sqrt{\log p}}.$$

Accuracy of approximation: for $p = 500$

$$2(1 - \Phi(\sqrt{2\log p}) = 0.000423, \quad \frac{1}{\sqrt{\pi}}\frac{1}{p\sqrt{\log p}} = 0.000453$$

Here the expected number of false discoveries is smaller than 1 and decreases with p

Modified BIC

In mBIC (Bogdan et al. 2004) the penalty depends on p and n,

$$mBIC = RSS + k\sigma^2 \left(\log n + 2 \log \left(\frac{p}{C} \right) \right) \quad ,$$

where C is the prior expected number of nonzero regression coefficients. In the lack of the prior knowledge the value $C = 4$ is suggested. It is motivated by controlling the probability of at least one false discovery.

Modified BIC

In mBIC (Bogdan et al. 2004) the penalty depends on p and n,

$$mBIC = RSS + k\sigma^2 \left(\log n + 2 \log \left(\frac{p}{C} \right) \right) \quad ,$$

where C is the prior expected number of nonzero regression coefficients. In the lack of the prior knowledge the value $C = 4$ is suggested. It is motivated by controlling the probability of at least one false discovery.

mBIC2 (Żak-Szatkowska and Bogdan (CSDA, 2011), Frommlet et al. (2011))

$$mBIC2 := RSS + \sigma^2 (k \log(n) + 2k \log(p/4) - 2 \log(k!) \quad .$$

The last relaxing term $2 \log(k!)$ is motivated by the desire to control FDR instead of FWER.