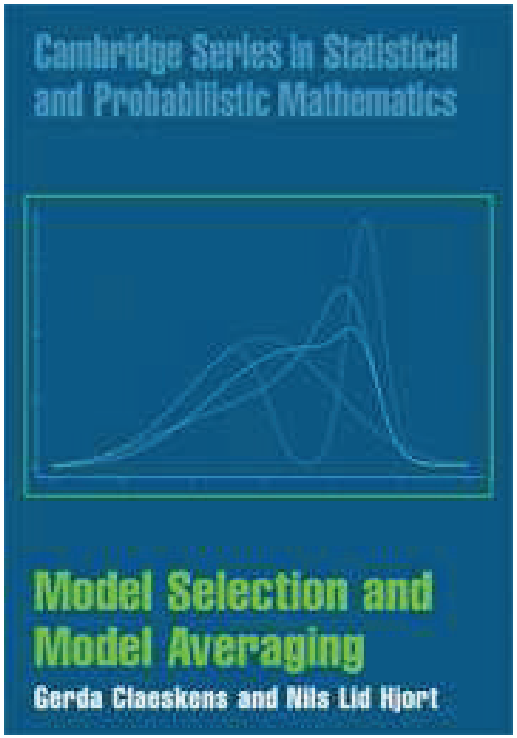
Groningen – Short course
14 March 2011

Model selection and model averaging

Gerda Claeskens **K.U.Leuven – Belgium**

Based on



Contents

- Variable selection in multiple regression (Adj. R^2, Mallows' C_p)
- AIC, TIC, AICc
- BIC, DIC, MDL, Hannan-Quinn
- Consistency, efficiency, overfitting
- Focussed model selection (linear, logistic, AFIC)
- Model averaging (frequentist, Bayesian)
- Do not ignore model selection in inference!

Introduction to variable selection

Data can often be modelled in many different ways.
When many covariates are measured: attempt to use them all, or only a subset of them.

A formal criterion for choosing one of a list of models is welcome. Many such methods exist. No exhaustive overview here, but restrict to some often used criteria.

First, some variable selection methods that can be used only in multiple regression models,
later criteria for more general models such as for example logistic regression models, Poisson models,. . .

Multiple regression – Mesquite trees data

We wish to construct a model for the total production of photosynthetic biomass of mesquite trees by using easily measured aspects of the plant as opposed to actual harvesting of the mesquite.
Data on 20 mesquite trees from the same region are collected. The variables are:
y = LEAFWT = total weight (in grams) of photosynthetic material derived from the actual harvesting of mesquite.
x_1 = DIAM1 = canopy diameter (in meters) measured along the longest axis of the tree parallel to the ground.
x_2 = DIAM2 = canopy diameter (in meters) measured along the shortest axis of the tree parallel to the ground.
x_3 = TOTHT = total height (in meters) of the tree.
x_4 = CANHT = canopy height (in meters) of the tree.

It is desired to estimate leaf weight. A multiplicative model is more natural here than a linear one since leaf weight should be nearly proportional to canopy volume, and canopy volume should be nearly proportional to the product of canopy dimensions:

$$Y = \beta_0 x_1^{\beta_1} x_2^{\beta_2} x_3^{\beta_3} x_4^{\beta_4} \varepsilon.$$

A log transformation gives us a linear model:

$$Y' = \beta_0' + \beta_1 x_1' + \beta_2 x_2' + \beta_3 x_3' + \beta_4 x_4' + \varepsilon'$$

where $\beta_0' = \log(\beta_0)$, $x_j' = \log(x_j)$ and $\varepsilon' = \log(\varepsilon)$.

Multiple regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

with $\varepsilon \sim N(0, \sigma^2)$. Residual $e_i = y_i - \hat{y}_i$ where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_k x_{ki}$.

Sums of squares:

Error sum of squares: $SSE = \sum_{i=1}^n e_i^2$

Total sum of squares: $SST = \sum_{i=1}^n (y_i - \bar{y})^2$

Regression sum of squares: $SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$

ANOVA table for regression:

Source of variation	df	SS	MS	F
Regression	k	SSR	MSR	$F = \frac{\text{MSR}}{\text{MSE}}$
Error	$n-(k+1)$	SSE	MSE $(= \hat{\sigma}^2)$	
Total	$n-1$	SST		

Adjusted R^2

A large set of predictors: select that subset which gives a good fit and results in a model which is easy to understand.

Measures of model fit: (from the anova table)

(i) the value of $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$, which is the proportion of variation of Y explained by the linear regression relationship with $\boldsymbol{x}$ in the model $Y = \boldsymbol{\beta}^t \boldsymbol{x} + \varepsilon$.

(ii) the value of

$$\text{Adj } R^2 = 1 - \frac{n-1}{n-(k+1)} \cdot \frac{SSE}{SST} = \frac{(n-1)R^2 - k}{n-1-k}.$$

We try to find a submodel with a small number of parameters p for which R^2, resp. Adj R^2, is nearly as large as R^2, resp. Adj R^2, in the complete/full model.

Selected R code and output

```
> fit1=lm(log(y)~log(x1)+log(x2)+log(x3)+log(x4))
> summary(fit1)
Call:
lm(formula = log(y)~log(x1) + log(x2) + log(x3) + log(x4))
Residuals:
     Min       1Q   Median       3Q      Max
-0.72677 -0.09396  0.03281  0.14891  0.63882
Coefficients:
            Estimate Std.Error t value Pr(>|t|)
(Intercept)   4.3043    0.3216  13.383  9.6e-10***
log(x1)       0.9579    0.6429   1.490    0.1569
log(x2)       1.0194    0.4405   2.314    0.0353*
log(x3)       1.1650    0.7259   1.605    0.1294
log(x4)      -0.6040    0.7012  -0.861    0.4026
---
Signif. codes:0'***' 0.001'**' 0.01'*' 0.05'.'
Residual standard error: 0.3685 on 15 degrees of freedom
Multiple R-Squared: 0.8841,     Adjusted R-squared: 0.8532
F-statistic:  28.6 on 4 and 15 DF,  p-value: 7.307e-07
```

Example: mesquite data

```
> library(leaps)
> x=log(mesquite[,1:4]);   y=log(mesquite[,5])
> namesvec=names(mesquite)[1:4]
> out.adjr2 = leaps(x,y,method="adjr2",names=namesvec)
> out.adjr2
$which
     x1    x2    x3    x4
1  TRUE FALSE FALSE FALSE
1 FALSE  TRUE FALSE FALSE
1 FALSE FALSE  TRUE FALSE
1 FALSE FALSE FALSE  TRUE
2 FALSE  TRUE  TRUE FALSE
2  TRUE  TRUE FALSE FALSE
2 FALSE  TRUE FALSE  TRUE
2  TRUE FALSE  TRUE FALSE
2  TRUE FALSE FALSE  TRUE
2 FALSE FALSE  TRUE  TRUE
     x1    x2    x3    x4
3  TRUE  TRUE  TRUE FALSE
3 FALSE  TRUE  TRUE  TRUE
3  TRUE  TRUE FALSE  TRUE
3  TRUE FALSE  TRUE  TRUE
4  TRUE  TRUE  TRUE  TRUE
```



```
$label
[1] "(Intercept)" "x1"          "x2"          "x3"
[5] "x4"
$size
 [1] 2 2 2 2 3 3 3 3 3 3 4 4 4 4 5
$adjr2
 [1] 0.8082405 0.7920549 0.6354619 0.5315810 0.8486957 0.8443882
 [7] 0.8202375 0.8128891 0.7981607 0.6142157 0.8555364 0.8419680
[13] 0.8387093 0.8132058 0.8531686

> which.max(out.adjr2$adjr2) # answer: [1] 11
> out.adjr2$which[which.max(out.adjr2$adjr2),]
   x1    x2    x3    x4
 TRUE  TRUE  TRUE FALSE
```


Mallows's C_p

Let SSE_p be the residual sum of squares $\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$ in the model with p regression coefficients and $\hat{\sigma}^2$ the estimated variance (MSE) in the largest model.

$$C_p = \frac{\text{SSE}_p}{\hat{\sigma}^2} + 2p - n.$$

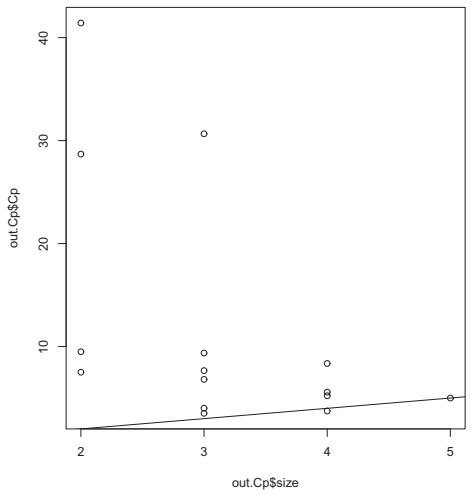
C_p estimates the scaled squared prediction error

$$\Gamma_p = \frac{E\left(\sum_{i=1}^{n}\left\{\hat{Y}_i - E(Y_i)\right\}^2\right)}{\sigma^2} = \frac{E(SSE_p)}{\sigma^2} + 2p - n$$

If the model with p variables contains no bias, $C_p \approx p$. If there is a large bias, $C_p > p$. Values close to the corresponding p (but preferably smaller than p) indicate a good model. The best model has a small C_p value.


```
> out.Cp = leaps(x,y,method="Cp",names=namesvec)
$which
     x1    x2    x3    x4
1  TRUE FALSE FALSE FALSE
1 FALSE  TRUE FALSE FALSE
1 FALSE FALSE  TRUE FALSE
1 FALSE FALSE FALSE  TRUE
2 FALSE  TRUE  TRUE FALSE
2  TRUE  TRUE FALSE FALSE
2 FALSE  TRUE FALSE  TRUE
2  TRUE FALSE  TRUE FALSE
     x1    x2    x3    x4
2  TRUE FALSE FALSE  TRUE
2 FALSE FALSE  TRUE  TRUE
3  TRUE  TRUE  TRUE FALSE
3 FALSE  TRUE  TRUE  TRUE
3  TRUE  TRUE FALSE  TRUE
3  TRUE FALSE  TRUE  TRUE
4  TRUE  TRUE  TRUE  TRUE
$label  [1] "(Intercept)" "x1"    "x2"    "x3"    "x4"
$size   [1] 2 2 2 2 3 3 3 3 3 3 4 4 4 4 5
$Cp
 [1]7.507715 9.491905 28.688575 41.423289 3.517870 4.016583 6.812733
 [8] 7.663520 9.368762 30.665748 3.741983 5.220510 5.575614 8.354690
 [11] 5.000000
> which.min(out.Cp$Cp)     [1] 5
> out.Cp$which[which.min(out.Cp$Cp),]
   x1    x2    x3    x4
FALSE  TRUE  TRUE FALSE
```

```
> plot(out.Cp$size,out.Cp$Cp)
> abline(a=0,b=1)
```



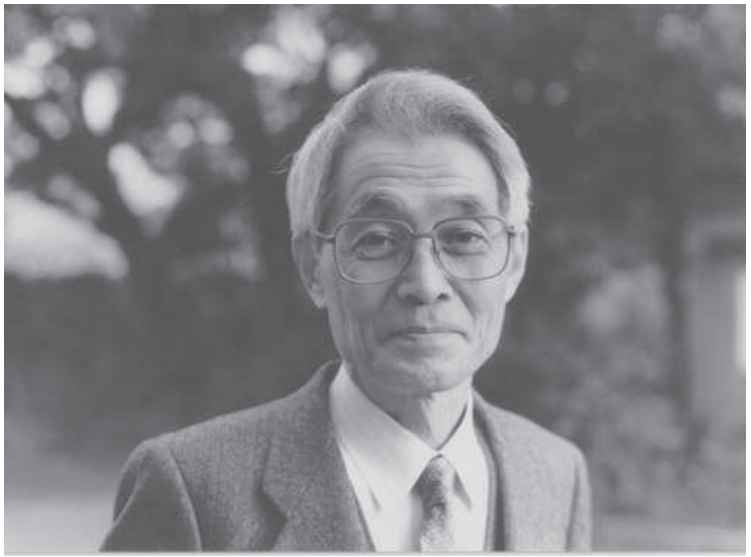
Akaike's information criterion

One of the most popular criteria is the AIC. AIC is minus twice a penalized log likelihood value, maximised using the MLE of the parameter vector θ. The likelihood function is denoted by $\mathcal{L}$:

$$\mathrm{AIC}\{f(\cdot;\theta)\} = -2\,\log \mathcal{L}(\widehat{\theta}) + 2\,\dim(\theta) = -2\,\ell_{\max} + 2\,\dim(\theta),$$

with $\dim(\theta)$ the length of the parameter vector θ.
A good model has a small value of AIC.
AIC selects the best approximating model to the unknown true data generating process, amongst the set of models under consideration.
AIC is applicable to likelihood models (thus including generalized linear models).

AIC defined by Hirotugu Akaike on 16/3/1971



Example 1: AIC for normal data

Normal multiple regression model

$$Y_i = x_{i,1}\beta_1 + \cdots + x_{i,p}\beta_p + \varepsilon_i = x_i^t\beta + \varepsilon_i \quad \text{for } i = 1,\ldots,n,$$

with $\varepsilon_1,\ldots,\varepsilon_n \sim \mathrm{N}(0,\sigma^2)$, i.i.d. and $\beta = (\beta_1,\ldots,\beta_p)^t$.
The log-likelihood function

$$\ell_n(\beta,\sigma) = \sum_{i=1}^{n}\{-\log\sigma - \tfrac{1}{2}(y_i - x_i^t\beta)^2/\sigma^2 - \tfrac{1}{2}\log(2\pi)\}.$$

Maximization with respect to β and σ yields

$$\widehat{\beta} = (X^tX)^{-1}X^tY \qquad \widehat{\sigma}^2 = n^{-1}\mathsf{SSE}(\widehat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - x_i^t\widehat{\beta})^2.$$

Plugging in $\widehat{\sigma}$ in $\ell_n(\widehat{\beta}, \sigma)$ gives
$\ell_{n,\max} = -n \log \widehat{\sigma} - \frac{1}{2}n - \frac{n}{2}\log(2\pi)$ and

$$\text{AIC} = 2n \log \widehat{\sigma} + 2(p+1) + n + n\log(2\pi).$$

Equivalently, minimize $n \log \widehat{\sigma}^2 + 2p$, across all models.

Mesquite data

```
> AIC(fit1)
[1] 23.07084
> const= 2+n+n*log(2*pi); const
[1]  58.75754
> AIC(fit1)-const
[1]  -35.68670
```


Searching through model lists

When there is a large number of variables, an all subsets search might no longer be feasible. Stepwise procedures can then be used. The function `stepAIC` can be used from within `library(MASS)`. For the **mesquite data**:

```
> stepboth=stepAIC(fit1,k=2,direction="both",
                   scope=list(upper=~.,lower=~1))
Start:  AIC=-35.69
log(y) ~ log(x1) + log(x2) + log(x3) + log(x4)
          Df Sum of Sq    RSS     AIC
- log(x4)  1   0.10075 2.1376 -36.721
<none>                 2.0368 -35.687
- log(x1)  1   0.30152 2.3384 -34.926
- log(x3)  1   0.34974 2.3866 -34.517
- log(x2)  1   0.72711 2.7639 -31.581
Step:  AIC=-36.72
```


▷

```
log(y) ~ log(x1) + log(x2) + log(x3)
          Df Sum of Sq    RSS     AIC
<none>                 2.1376 -36.721
- log(x1)  1   0.24115 2.3787 -36.583
- log(x3)  1   0.30887 2.4465 -36.022
+ log(x4)  1   0.10075 2.0368 -35.687
- log(x2)  1   0.80408 2.9417 -32.335

> stepbackw=stepAIC(fit1,k=2,direction="backward",
                    scope=list(upper=~.,lower=~1))
Start:  AIC=-35.69
log(y) ~ log(x1) + log(x2) + log(x3) + log(x4)
          Df Sum of Sq    RSS     AIC
- log(x4)  1   0.10075 2.1376 -36.721
<none>                 2.0368 -35.687
- log(x1)  1   0.30152 2.3384 -34.926
- log(x3)  1   0.34974 2.3866 -34.517
- log(x2)  1   0.72711 2.7639 -31.581
Step:  AIC=-36.72
```



```
log(y) ~ log(x1) + log(x2) + log(x3)
          Df Sum of Sq    RSS     AIC
<none>                 2.1376 -36.721
- log(x1)  1   0.24115 2.3787 -36.583
- log(x3)  1   0.30887 2.4465 -36.022
- log(x2)  1   0.80408 2.9417 -32.335

> fit0 = lm(log(y)~1,data=mesquite)
> stepforward = stepAIC(fit0,k=2,direction="forward",
            scope=list(lower=~1,upper=fit1))
Start:  AIC=-0.59
log(y) ~ 1
          Df Sum of Sq     RSS     AIC
+ log(x1)  1   14.3790  3.1921 -32.701
+ log(x2)  1   14.1096  3.4615 -31.080
+ log(x3)  1   11.5029  6.0682 -19.853
+ log(x4)  1    9.7737  7.7975 -14.839
<none>                17.5711  -0.590
Step:  AIC=-32.7
```

```
log(y) ~ log(x1)
          Df Sum of Sq    RSS     AIC
+ log(x2)  1   0.74564 2.4465 -36.022
<none>                 3.1921 -32.701
+ log(x3)  1   0.25042 2.9417 -32.335
+ log(x4)  1   0.01887 3.1732 -30.820

Step:  AIC=-36.02
log(y) ~ log(x1) + log(x2)
          Df Sum of Sq    RSS     AIC
+ log(x3)  1  0.308866 2.1376 -36.721
<none>                 2.4465 -36.022
+ log(x4)  1  0.059879 2.3866 -34.517

Step:  AIC=-36.72
log(y) ~ log(x1) + log(x2) + log(x3)
          Df Sum of Sq    RSS     AIC
<none>                 2.1376 -36.721
+ log(x4)  1   0.10075 2.0368 -35.687
```



```
Step:  AIC=-36.72
log(y) ~ log(x1) + log(x2) + log(x3)
                  Df Sum of Sq    RSS     AIC
<none>                        2.1376 -36.721
- log(x1)          1   0.24115 2.3787 -36.583
- log(x3)          1   0.30887 2.4465 -36.022
+ log(x4)          1   0.10075 2.0368 -35.687
+ log(x1):log(x2)  1   0.08511 2.0525 -35.534
+ log(x1):log(x3)  1   0.08028 2.0573 -35.487
+ log(x2):log(x3)  1   0.02523 2.1124 -34.959
- log(x2)          1   0.80408 2.9417 -32.335
```


This function is quite useful if you also want to investigate whether interactions between variables could make a better model. It is used in the following way.

```
> stepboth=stepAIC(fit1,k=2,direction="both",
              scope=list(upper=~.^2,lower=~1))
Start:  AIC=-35.69
log(y) ~ log(x1) + log(x2) + log(x3) + log(x4)
                  Df Sum of Sq    RSS     AIC
- log(x4)          1   0.10075 2.1376 -36.721
+ log(x1):log(x4)  1   0.27083 1.7660 -36.540
<none>                        2.0368 -35.687
+ log(x2):log(x4)  1   0.16305 1.8738 -35.355
- log(x1)          1   0.30152 2.3384 -34.926
+ log(x3):log(x4)  1   0.12211 1.9147 -34.923
+ log(x1):log(x2)  1   0.10696 1.9299 -34.766
+ log(x1):log(x3)  1   0.10658 1.9303 -34.762
- log(x3)          1   0.34974 2.3866 -34.517
+ log(x2):log(x3)  1   0.03786 1.9990 -34.062
- log(x2)          1   0.72711 2.7639 -31.581
```


Example 2: Exponential vs. Weibull

Do running computer processes have the memory-less property or not?
$Y_1, \ldots, Y_n$ are independent life-time data. Two models: (1) exponential with density $f(y, \theta) = \theta \exp(-\theta y)$.
(2) If failure rates decrease with time (or for wear-out processes increase with time): Weibull with density $f(y, \theta, \gamma) = \exp\{-(\theta y)^\gamma\}\theta^\gamma \gamma y^{\gamma-1}$.
To select the best model, we compute

$$\mathrm{AIC}_{\mathrm{exp}} = -2\sum_{i=1}^{n}(\log\widetilde{\theta} - \widetilde{\theta} y_i) + 2,$$

$$\mathrm{AIC}_{\mathrm{wei}} = -2\sum_{i=1}^{n}\{-(\widehat{\theta} y_i)^{\widehat{\gamma}} + \widehat{\gamma}\log\widehat{\theta} + \log\widehat{\gamma} + (\widehat{\gamma}-1)\log y_i\} + 4.$$

Example 3: Low birthweight data

Dataset on low birthweight (Hosmer & Lemeshow, 1999).
$n = 189$ women with newborn babies.
Low birthweight when the weight at birth $\leq$ 2500 gram.
Explanatory variables:

- $x_1 = 1$ a constant intercept,
- x_2: weight of mother just prior to pregnancy,
- x_3: age of mother,
- x_4: indicator for race 'black',
- x_5: indicator for race 'other',
- and $x_4 = x_5 = 0$: race 'white'. $\bowtie$

Models to select from

We decide to include x_1 and x_2 in all of the possible models.
Subsets of $u = (x_3, x_4, x_5)$ are possibly included.

$$P(\text{low birthweight}|x, u) = \frac{\exp(x^t\beta + u^t\gamma)}{1 + \exp(x^t\beta + u^t\gamma)}.$$

In the programming language R we fit the full model as follows.

```
fit = glm(y ~ x2 + x3 + x4 + x5,
          family=binomial)
```

The build-in output `fit$aic` gives AIC. Also: `AIC(fit)`.

Other models can be fit via the `update` command:

```
fitmin4 = update(fit,.~.-x4)
```


We choose the model with the smallest value of AIC.

Extra Cov.	AIC value	Order
$\emptyset$	232.691	
x_3	233.123	
x_4	231.075	(1)
x_5	234.101	
x_3, x_4	232.175	(3)
x_3, x_5	234.677	
x_4, x_5	231.259	(2)
x_3, x_4, x_5	232.661	

$$\text{logit}\{\widehat{\text{P}}(\text{low birthweight}|x, u)\} = 1.198 - 0.0166x_2 + 0.891x_4.$$

AIC in the i.i.d. case

Suppose $Y_1, \ldots, Y_n$ are i.i.d. from an unknown density g.
Consider a parametric model with density $f_\theta(y) = f(y, \theta)$
where $\theta = (\theta_1, \ldots, \theta_p)^{\text{t}}$ belongs to some open subset of $\mathbb{R}^p$.
MLE $\widehat{\theta}$ aims at the least false parameter value θ_0 that minimizes the Kullback-Leibler distance (KL)

$$\int g(y) \log\{g(y)/f_\theta(y)\}\, \text{d}y.$$

KL-distance between fitted and true model

$$KL = \int g(y) \log g(y)\, \text{d}y - \int g(y) \log f(y, \widehat{\theta})\, \text{d}y.$$

Since the first term is a constant across models f_θ, consider

$$R_n = \int g(y) \log f(y, \widehat{\theta})\, \text{d}y.$$

This is a random variable, dependent upon the data via $\widehat{\theta}$.

Hence, consider the expected value or average quality, i.e.

$$Q_n = \mathrm{E}_g[R_n] = \mathrm{E}_g\Big[\int g(y)\log f(y,\widehat{\theta})\,\mathrm{d}y\Big].$$

Estimate Q_n from data via

$$\widehat{Q}_n = \frac{1}{n}\sum_{i=1}^{n}\log f(Y_i,\widehat{\theta}) = \frac{1}{n}\ell_{n,\max}.$$

We shall see that the penalized form of the AIC follows naturally from properties of $\widehat{Q}_n$ as an estimator of Q_n.
The estimator tends to overshoot its target Q_n:

$$E(\widehat{Q}_n - Q_n) \approx p^*/n, \quad \text{where } p^* = \mathsf{Trace}(J^{-1}K).$$

Here we defined

$$J = -E_g\left[\frac{\partial^2 \log f(Y,\theta_0)}{\partial\theta\partial\theta^t}\right], \qquad K = \mathrm{Var}_g\left[\frac{\partial \log f(Y,\theta_0)}{\partial\theta}\right].$$

These $p\times p$ matrices are identical when $g(y)$ is actually equal to $f(y,\theta_0)$ for all y.
A bias-corrected estimator of the target Q_n:

$$\widehat{Q}_n - p^*/n = (1/n)(\ell_{n,\max} - p^*).$$

Tradition dictates transforming this to $\boxed{-2\ell_{n,\max} + 2\,p^*}$.

For a correct model

When the model actually holds, so that

$$g(y) = f(y,\theta_0),$$

then $K = J$ is the Fisher information matrix of the model, and

$$p^* = \mathrm{tr}(J^{-1}K) = p = \dim(\theta).$$

If we take $p^* = p$, the number of parameters in the model, this is what leads to the AIC criterion

$$\boxed{\mathrm{AIC} = -2\,\ell_{n,\max} + 2\,p = -2\,\log\mathcal{L}(\widehat{\theta}) + 2\dim(\theta)}.$$

Takeuchi's information criterion: TIC

When p^* is not trusted to be close to p, we estimate $p^* = \mathrm{tr}(J^{-1}(\theta_0)K(\theta_0))$, leading to Takeuchi's information criterion (1976):

$$\boxed{\mathrm{TIC} = -2\log\mathcal{L}(\hat{\theta}) + 2\,\mathrm{tr}\{J^{-1}(\hat{\theta})K(\hat{\theta})\}.}$$

Note that in case $f(\cdot;\theta) = g(\cdot)$, TIC = AIC.

When the expected values can not be calculated exactly, empirical matrices J_n and K_n might be used instead.

$$J_n = -\sum_{i=1}^{n}\frac{\partial^2\log f(y_i,\hat{\theta})}{\partial\theta\partial\theta^t}, \quad K_n = \sum_{i=1}^{n}\frac{\partial\log f(y_i,\hat{\theta})}{\partial\theta}\left(\frac{\partial\log f(y_i,\hat{\theta})}{\partial\theta}\right)^t.$$

Because of the possible estimation of J and K, this criterion is not often used in practice.

Low birthweight data, full model: $\mathrm{tr}(J_n^{-1}K_n) = 3.84$, $p = 4$.

A corrected version of AIC: AIC_C

Numerical results have shown that AIC has a tendency to overfit, it tends to pick models with more parameters than strictly necessary.

These type of studies, mostly by simulation, have been performed under the strict assumption that the correct parameterisation is known.

In linear regression models Hurvich & Tsai (1989), following Sugiura (1978), define a corrected version of AIC which has better finite sample behaviour with regard to overfitting.

They estimate the expected Kullback-Leibler distance directly for a multiple linear regression model

$$Y_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i,$$

with independent normally distributed errors $N(0, \sigma^2)$. The true model is of the same form:

$$Y_i = \beta_{01} x_{i1} + \ldots + \beta_{0q} x_{iq} + \varepsilon_{0i},$$

with independent normally distributed errors $N(0, \sigma_0^2)$.
The KL-distance between $f(y_1, \ldots, y_n; \beta_1, \ldots, \beta_p)$ and the true $g(y_1, \ldots, y_n)$ is obtained as:

$$\text{KL}(g, f) = \frac{n}{2}\{\log(\sigma^2/\sigma_0^2) - 1 - \sigma_0^2/\sigma^2\} + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i\beta_0 - x_i\beta)^2.$$

This leads for linear regression models with normal errors to the corrected AIC criterion $\boxed{\text{AIC}_C = -n\log(\hat{\sigma}_p^2) - \frac{n(n+p)}{n-p-2}}$.

Although AIC_C has only been derived for linear regression models and autoregressive models (Hurvich & Tsai, 1989), the criterion can be defined more generally for any likelihood function as:

$$\begin{aligned}\text{AIC}_C &= 2\log\mathcal{L}(\hat{\theta}) - \frac{2n\dim(\theta)}{n - \dim(\theta) - 1} \\ &= \text{AIC} - \frac{2\dim(\theta)(\dim(\theta)+1)}{n - \dim(\theta) - 1}.\end{aligned}$$

Use with care outside normal regression or autoregressive models.

The Bayesian information criterion

About equally popular as the AIC is the Bayesian information criterion BIC. The construction goes back to both Akaike (1978) and Gideon Schwarz (1978).

$$\boxed{\text{BIC} = -2\,\log\mathcal{L}(\widehat{\theta}) + \log(n)\dim(\theta) = -2\,\ell_{\max} + \log(n)\dim(\theta)}$$

with $\dim(\theta)$ the length of the parameter vector θ.

A good model has a small value of BIC.
The values of BIC can easily be obtained in R via the function

```
AIC(fitted.object, k=log(sample.size))
```

The default argument for the penalty `k` is the value 2, corresponding to the AIC.

Example 1: BIC for normal data

Normal multiple regression model

$$Y_i = x_{i,1}\beta_1 + \cdots + x_{i,p}\beta_p + \varepsilon_i = x_i^t\beta + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

with $\varepsilon_1, \ldots, \varepsilon_n \sim \mathrm{N}(0, \sigma^2)$, i.i.d. and $\beta = (\beta_1, \ldots, \beta_p)^t$.
Maximized log-likelihood function
$\ell_{n,\max} = -n \log \widehat{\sigma} - \frac{1}{2}n - \frac{n}{2}\log(2\pi)$

$$\mathrm{BIC} = 2n \log \widehat{\sigma} + n + n\log(2\pi) + \log(n) \cdot (p+1).$$

Equivalently, minimize $n \log \widehat{\sigma}^2 + \log(n) \cdot p$, across all models.
Mesquite data

```
> BIC=AIC(fit1,k=log(nrow(mesquite)))    [1]  29.04523
> logLik(fit1)                           'log Lik.' -5.535419 (df=6)
> -2*logLik(fit1)+log(nrow(mesquite))*6 [1] 29.04523
```



```
> stepboth=stepAIC(fit1,k=log(nrow(mesquite)),
direction="both",scope=list(upper=~.,lower=~1))
Start:  AIC=-30.71
log(y) ~ log(x1) + log(x2) + log(x3) + log(x4)
          Df Sum of Sq    RSS     AIC
- log(x4)  1   0.10075 2.1376 -32.738
- log(x1)  1   0.30152 2.3384 -30.943
<none>                 2.0368 -30.708
- log(x3)  1   0.34974 2.3866 -30.535
- log(x2)  1   0.72711 2.7639 -27.599

Step:  AIC=-32.74
log(y) ~ log(x1) + log(x2) + log(x3)
          Df Sum of Sq    RSS     AIC
- log(x1)  1   0.24115 2.3787 -33.596
- log(x3)  1   0.30887 2.4465 -33.035
<none>                 2.1376 -32.738
+ log(x4)  1   0.10075 2.0368 -30.708
- log(x2)  1   0.80408 2.9417 -29.348
```



```
Step:  AIC=-33.6
log(y) ~ log(x2) + log(x3)
          Df Sum of Sq    RSS     AIC
<none>                 2.3787 -33.596
+ log(x1)  1    0.2411 2.1376 -32.738
+ log(x4)  1    0.0404 2.3384 -30.943
- log(x3)  1    1.0828 3.4615 -29.089
- log(x2)  1    3.6895 6.0682 -17.862

> n=nrow(mesquite)
> const= n+n*log(2*pi)+log(n)
> BIC-const
[1] -30.70804
```


Example 2: Exponential vs. Weibull

Indep. life-time data $Y_1, \ldots, Y_n$ are modelled either via an exponential, or via the Weibull model.

$$\mathrm{BIC}_{\mathrm{exp}} = -2\sum_{i=1}^{n}(\log \widetilde{\theta} - \widetilde{\theta} y_i) + \log(n),$$

$$\mathrm{BIC}_{\mathrm{wei}} = -2\sum_{i=1}^{n}\{-(\widehat{\theta} y_i)^{\widehat{\gamma}} + \widehat{\gamma}\log\widehat{\theta} + \log\widehat{\gamma} + (\widehat{\gamma}-1)\log y_i\} + 2\log(n).$$

The best model has the smallest BIC value.

Example 3: Low birthweight data

Corresponding to the logistic regression model:

$$\text{BIC} = -2\sum_{i=1}^{n}\{y_i \log \widehat{p}_i + (1-y_i)\log(1-\widehat{p}_i)\} + \dim(\theta)\log(n),$$

where $\widehat{p}_i$ is the estimated probability for $Y_i = 1$ under the model and $\dim(\theta)$ is the number of estimated parameters.

The sample size $n = 189$, with log(189) $\approx$ 5.2417.

Smallest model: intercept only.
Full model: adds x_2, x_3, x_4, x_5.

Extra cov.	BIC value	order	Extra cov.	BIC value
—	239.914	(2)	x_3, x_4	246.471
x_2	239.174	(1)	x_3, x_5	246.296
x_3	242.395	(4)	x_4, x_5	245.387
x_4	243.502		x_2, x_3, x_5	247.644
x_5	243.382		x_2, x_4, x_5	244.226
x_2, x_3	242.849	(5)	x_3, x_4, x_5	249.094
x_2, x_4	240.800	(3)	x_2, x_3, x_4	245.142
x_2, x_5	243.826		full	248.869

BIC selected model:

$$\widehat{P}\{\text{low birthweight} \mid x_2\} = \frac{\exp(0.998 - 0.014\, x_2)}{1 + \exp(0.998 - 0.014\, x_2)}.$$

AIC leads to:

$\text{logit}\{\widehat{\text{P}}(\text{low birthweight}|x,u)\} = 1.198 - 0.0166x_2 + 0.891x_4.$

For the full model:

Variable	Estimate	Std. Error	z-value	$2P(Z > \lvert z\rvert)$
x_1	1.306741	1.069786	1.221	0.2219
x_2	-0.014353	0.006523	-2.200	0.0278 *
x_3	-0.025524	0.033252	-0.768	0.4427
x_4	1.003822	0.498014	2.016	0.0438 *
x_5	0.443461	0.360257	1.231	0.2183

Highest posterior probabilities

The "B" in BIC stands for "Bayesian".

A Bayesian procedure selects that model which is a posteriori most likely.

We calculate the posterior probability of each model and select the model with the biggest posterior probability. Denote the models by $M_1, M_2, \ldots$, and use Y as a vector notation for the vector of iid $Y_1, \ldots, Y_n$.
Via Bayes theorem:

$$P(A_j|B) = \frac{P(A_j)P(B|A_j)}{P(B)}.$$

$$P(M_j|Y) = \frac{P(M_j)}{f(Y)}\int_\Theta f(Y|M_j,\theta)\pi(\theta|M_j)\mathrm{d}\theta.$$

BIC

The 'B' is for 'Bayesian'. Select that model that is a posteriori most likely.

$$P(M_j|Y) = \frac{P(M_j)}{f(Y)} \overbrace{\int_\Theta \underbrace{f(Y|M_j,\theta_j)}_{\text{likelihood}} \underbrace{\pi(\theta_j|M_j)}_{\text{prior}} d\theta_j}^{P(Y|M_j)}$$

Laplace approximation to integral gives that

$$-2\log\{\int_\Theta f(Y|M_j,\theta_j)\pi(\theta_j|M_j)d\theta_j\} \approx \text{BIC} = -2\log \text{likelihood} + p\log n$$

Thus

$$P(M_j\,|\,Y) \approx \frac{P(M_j)\exp(-\frac{1}{2}\text{BIC}_{n,j})}{\sum_{j'=1}^k P(M_{j'})\exp(-\frac{1}{2}\text{BIC}_{n,j'})}.$$

For BIC no prior information is needed, no complicated calculations of posteriors.

Laplace approximation

In p dimensions

$$\begin{aligned} P(M_j|Y) &= (2\pi)^{p/2} e^{\ell_n(\hat\theta)} n^{-p/2} \\ &\quad \times \left\{\pi(\hat\theta|M_j)|\text{Det}J(\hat\theta)|^{-1/2} + O(n^{-1})\right\} \frac{P(M_j)}{f(Y)}, \end{aligned}$$

with $J(\hat\theta) = -\frac{\partial^2}{\partial\theta\partial\theta^t}\log f(Y|M_j,\widehat\theta)$.

Taking logarithms

$$\begin{aligned} \log\{P(M_j|Y)\} &= \ell_n(\hat\theta) - \frac{p}{2}\log n + (p/2)\log(2\pi) + \log P(M_j) \\ &\quad - \log f(Y) + \log\{\pi(\hat\theta|M_j)|\text{Det}J(\hat\theta)|^{-1/2} \\ &\quad + O(n^{-p/2-1})\}. \end{aligned}$$

Schwarz's approximation

- $f(Y)$ is a common proportionality constant for all models, and hence can be ignored.
- All models are equally likely, ignore prior probabilities $P(M_j)$.
- All last four terms bounded in n for given Y and M_j.

This is the motivation to define

$$\boxed{\text{BIC}(M_j) = -2\log\mathcal{L}_n(\hat\theta) + p\log n.}$$

where $p = \dim(\theta)$ and $\mathcal{L}_n(\hat\theta)$ is the maximized likelihood. We select that model M_k for which $\text{BIC}(M_k)$ is the smallest.

Deviance information criterion: DIC

DIC is an alternative to the use of Bayes factors for model comparison (Spiegelhalter, Best, Carlin & VanderLinde 2002).

The deviance for a model $S \stackrel{\text{def}}{=} 2\times$ difference in maximised log likelihood of a saturated model and of the model S.

$$D(y,\theta) = 2\{\ell_n(\theta) - \ell_n(\hat\theta_S)\}.$$

Regression: saturation by estimating $E(Y_i) = \mu_i$ by $\hat\mu_i = Y_i$.

In Bayesian modelling θ is a random variable with posterior mean $E(\theta|Y) = \bar\theta$. Penalty p_D is defined as

$$p_D = E_{\theta|Y}[2\{\ell_n(\bar\theta) - \ell_n(\theta)\}|\text{data}] = \overline{\text{D(y},\theta)} - \text{D(y},\bar\theta).$$

Deviance information criterion

$$\boxed{\text{DIC} = D(y,\bar{\theta}) + 2p_D = \overline{D(y,\theta)} + p_D.}$$

We search for the model with the lowest value of DIC.
p_D is the effective number of parameters in the model.
We can show that $p_D \xrightarrow{P} p^* = \text{tr}(J^{-1}K)$.
Does this ring a bell?

Exact formula's for DIC are difficult. Computation is easy by simulation.
$\bar{\theta}$: observed mean of a large number of simulated θ from the posterior distribution.
p_D: observed mean of a large number of simulated $d(y;\theta,\bar{\theta})$ values.

Minimum description length: MDL

[For an excellent treatment, see Grünwald (2007, MIT)]

MDL tries to measure the complexity of the modelling process and selects that model which is least complex. Complexity $\hookrightarrow$ shortest code length $\hookrightarrow$ maximizing a probability.

$$\theta_{mdl} = \arg\min_{\theta\in\Theta}\{-\log P(Y|\theta) + L_C(\theta)\}$$

Choice of L_C is *not* unique.

1. k-dim. estimator with $1/\sqrt{n}$ consistency:
   complexity $\approx -k\log(1/\sqrt{n})$
2. Minimax results (Rissanen). Stochastic complexity code.

$$L_{SC}(Y|\mathcal{M}) = -\log P(Y|\hat{\theta}_{ML}) + \frac{k}{2}\log\left(\frac{n}{2\pi}\right) + \log\int\sqrt{|J(\theta)|} + o(1)$$

The Hannan and Quinn criterion

Hannan and Quinn's 1979 criterion replaces the $\log n$ factor in BIC by the slower diverging quantity $\log\log(n)$. This gives

$$\boxed{\text{HQ} = -2\log\mathcal{L}(\hat{\theta}) + \log\log(n)\dim(\theta)}.$$

The criterion was originally derived to determine the order in an autoregressive time series model.

The double logarithmic term arises from an application of the law of the iterated logarithm.

Low birthweight data

Extra Cov.	AIC value	Order	HQ	Order	BIC	Order
$\emptyset$	232.69	5	232.00	5	239.17	1
x_3	233.12	6	232.09	6	242.85	3
x_4	231.07	1	230.04	2	240.80	2
x_5	234.10	7	233.07	7	243.83	4
x_3, x_4	232.17	3	230.80	3	245.14	6
x_3, x_5	234.68	8	233.30	8	247.64	7
x_4, x_5	231.26	2	229.89	1	244.23	5
x_3, x_4, x_5	232.66	4	230.94	4	248.87	8

Penalty constants. AIC: 2, BIC: $\log(n) = 5.2417$,
HQ: $\log\log(n) = 1.6567$.

Consistency

There exists one true model that generated the data. Under the (strong) assumption:

(A1) this true model is one of the candidate models,

we want the model selection method to identify this true model.

Definition. A model selection method is weakly consistent if with probability tending to one the selection method is able to select the true model from the candidate models. Strong consistency is obtained when the selection of the true model happens almost surely.

Consistent criteria: BIC and HQ are strongly consistent.

Efficiency

If we do NOT believe assumption (A1), then there exist other ways of defining 'good' models. For example, we might want the selected model to have the smallest expected prediction error.

Example. Select the best set of variables in the model

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_k x_{k,i} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\mathrm{Var}\,\varepsilon_i = \sigma^2$, with the specific purpose of predicting a new (independent) outcome variable $\widehat{Y}_i$ at the observed covariates $x_i = (x_{1,i}, \ldots, x_{k,i})^t$, for $i = 1, \ldots, n$.
Select that set of covariates $x_j, j \in M$ for which the expected prediction error $\mathrm{PE}(M)$ is as small as possible.

$$\mathrm{PE}_n(M) = \sum_{i=1}^{n} E\{(\widehat{Y}_{M,i} - Y_{\mathrm{true},i})^2\}.$$

For the regression situation this reads

$$\mathrm{PE}_n(M) = E\{(\widehat{\beta}_M - \beta_{\mathrm{true}})^t X^t X (\widehat{\beta}_M - \beta_{\mathrm{true}})\} + n\sigma^2.$$

Denote M^*: index set for which the minimum value of the expected prediction error is attained. Let $\widehat{M}$ be the set of indices in the selected model. The notation $E_{\widehat{M}}$ denotes that the expectation is taken with respect to all random quantities except for $\widehat{M}$.

Definition. The criterion used to select $\widehat{M}$ is efficient when

$$\frac{\sum_{i=1}^n E_{\widehat{M}}\{(\widehat{Y}_{\widehat{M},i} - Y_{\mathrm{true},i})^2\}}{\sum_{i=1}^n E\{(\widehat{Y}_{M^*,i} - Y_{\mathrm{true},i})^2\}} = \frac{\mathrm{PE}_n(\widehat{M})}{\mathrm{PE}_n(M^*)} \xrightarrow{P} 1, \text{ as } n \to \infty.$$

Efficient criteria: AIC, AIC$_c$, Mallows' C_p.

Over- and underfitting

Let M_0 be the true model and $\hat{M}$ the selected model.

Probability of underfitting: $P(M_0 \not\subseteq \hat{M})$

Probability of overfitting: $P(M_0 \subset \hat{M}, M_0 \neq \hat{M})$

Asymptotically, consistent criteria do not overfit nor underfit.

Asymptotic probabilities of overfitting for AIC?

Assume that $M_0 \subset M_1 \subset \ldots \subset M_k \subset \ldots$ and that $k = 0$ is the correct model order. dim(M_j) =dim$(M_0) + j$.

AIC and the arc sine distribution

Woodroofe (1982) calculates that for $k \to \infty$ the expected number of superfluous parameters is 0.946, while the probability of correctly identifying the true model is 0.712.

Representation of AIC via χ^2 random variables

$\hat{M} = M_{\hat{k}}$ minimizes over $k = 0, 1, \ldots,$

$$\mathrm{AIC}(M_k) = -2 \log \mathcal{L}(\widehat{\theta}_k) + 2 \dim(\theta_k).$$

$\Leftrightarrow \hat{M}$ maximizes over $k = 0, 1, \ldots,$

$$\mathrm{AIC}(M_k) - \mathrm{AIC}(M_0) = 2 \log \left\{ \frac{\mathcal{L}(\widehat{\theta}_k)}{\mathcal{L}(\widehat{\theta}_0)} \right\} - 2\,k.$$

If M_0 is the true model,

$$2 \log \left\{ \frac{\mathcal{L}(\widehat{\theta}_k)}{\mathcal{L}(\widehat{\theta}_0)} \right\} \to \chi^2_k.$$

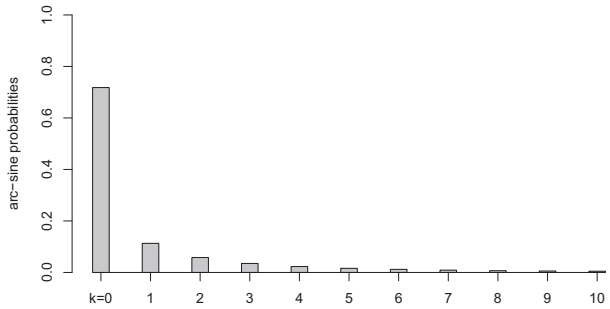
Asymptotically AIC finds the model order $\hat{k}$ such that

$$2\left(\tfrac{1}{2}\chi^2_k - k\right) \text{ is maximal.}$$

Let $Z_1, Z_2, \ldots$ i.i.d. $N(0,1)$. Then $\chi^2_k = \sum_{j=1}^k Z_j^2$.

If $\hat{k}_{\mathrm{aic}} > 0$ there is overfitting.

$P(\hat{k}_{\mathrm{aic}} = k)$



Consistency and efficiency?

Both AIC and BIC have good properties:

- AIC is efficient
- BIC is consistent.

Can the consistency of the BIC be combined with the efficiency of the AIC?
The answer is NO.
Changing the penalty constant 2 in the AIC to some other value takes away the favorable situation of optimality.
Formal proofs are given in Yang (2005, Biometrika).

Model selection for best performance

The focussed information criterion: FIC

- What is “focussed" selection?
- Is it more difficult than the AIC?
- Can we have less focus without falling back to no focus?

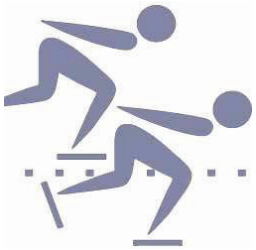
A ‘best model’ should depend on the parameter under focus, such as the mean, the variance, etc.
The FIC allows and encourages different models to be selected for different parameters of interest.

The focus is more important than the model

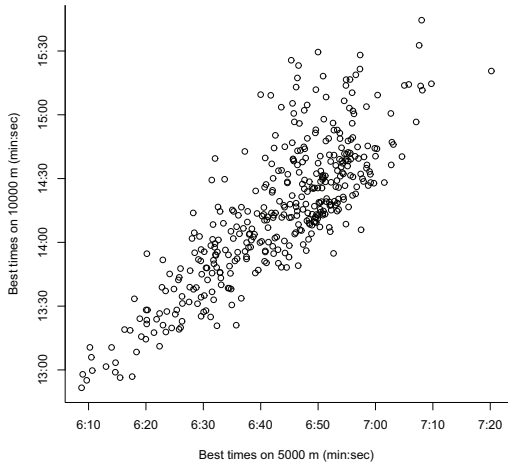
Speedskating data

Top of the Adelskalender, as of 25 March 2006: this is a list of the best speedskaters ever, sorted by the point-sum based on the skaters' personal bests over the four classic distances 500 m, 1500 m, 5000 m, 10000 m. The point-sum is $X_1 + X_2/3 + X_3/10 + X_4/20$.

- How can we (best) predict 10km time from 5km time for a top-level skater?
- How can we (best) predict 10km time from 5km time for a median-level skater?



			500 m	1500 m	5000 m	10000 m	pointsum
1	C. Hedrick	USA	35.58	1.42.78	6.09.68	12.55.11	145.563
2	S. Davis	USA	35.17	1.42.68	6.10.49	13.05.94	145.742
3	E. Fabris	ITA	35.99	1.44.02	6.10.23	13.10.60	147.216
4	J. Uytdehaage	NED	36.27	1.44.57	6.14.66	12.58.92	147.538
5	S. Kramer	NED	36.93	1.46.80	6.08.78	12.51.60	147.988
6	E. Ervik	NOR	37.03	1.45.73	6.10.65	12.59.69	148.322
7	C. Verheijen	NED	37.14	1.47.42	6.08.98	12.57.92	148.740
8	D. Parra	USA	35.88	1.43.95	6.17.98	13.33.44	149.000
9	I. Skobrev	RUS	36.00	1.45.36	6.21.40	13.17.54	149.137
10	D. Morrison	CAN	35.34	1.42.97	6.24.13	13.45.14	149.333



Speedskating data. Personal best times (in min) for the 400 first listed skaters on the 5km and 10km distances.

Four models to consider

$Y_i = a + bx_i + cx_i^2 + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma_i^2), \sigma_i = \sigma \exp(\phi x_i)$

M_0: linear & homoscedastic, $c = \phi = 0$

M_1: quadratic & homoscedastic, $\phi = 0$;

M_2: linear & heteroscedastic, $c = 0$;

M_3: quadratic & heteroscedastic.

A no-focus model search:

AIC $=$ $-2\log$ likelihood $+ 2 \times$ number of parameters

BIC $=$ $-2\log$ likelihood $+ \log(n) \times$ number of parameters.

Both choose model M_2.

Example 1: a median level skater with x_0 equal to 6:35.00
Example 2: a top level skater with x_0 equal to 6:15.00

Estimated 10% quantiles of the 10km time for the 2 skaters:

5km time:	median (6:35)	top (6:15)
M_0	13:37.25	12:49.35
M_1	13:37.89	12:48.13
M_2	13:38.05	12:57.55
M_3	13:38.12	12:57.48

Properties of a good estimator:

- Small or no bias
- Small variance

$\hookrightarrow$ small MSE $=$ bias^2 $+$ var

Focussed model selection

Select that model for which the estimated MSE of the estimator $\hat{\mu}(x_0, q)$ is the lowest.
$\hookrightarrow$ Need to estimate the MSE in each of the models.
Likelihood function:

Smallest model: $f(Y_i, (\sigma, a, b), (0, 0))$
Smallest variance, largest bias.

Biggest model: $f(Y_i, \underbrace{(\sigma, a, b)}_{=\theta}, \underbrace{(c, \phi)}_{=\gamma})$
Largest variance, smallest bias.

True model somewhere in between:

$$f(Y_i, \sigma, a, b, 0 + \frac{\delta_1}{\sqrt{n}}, 0 + \frac{\delta_2}{\sqrt{n}}).$$

We need a local model, otherwise the bias will dominate.

Mean squared error of $\widehat{\mu}_S$

MSE of the limit distribution Λ_S of $\sqrt{n}(\widehat{\mu}_S - \mu_{\rm true})$ for each model S. Add bias squared and variance
The idea of the FIC is to estimate MSE(S) for each of the models S and choose that model which gives the smallest estimated mean squared error.

$$\text{FIC}(S) = \widehat{\text{Bias}^2(S)} + \widehat{\text{Var}(S)}$$

We choose the model with the smallest value of FIC.

Ingredients for an FIC analysis

(1) Specify focus of interest as $\mu(\theta, \gamma)$

(2) Decide on the list of candidate models

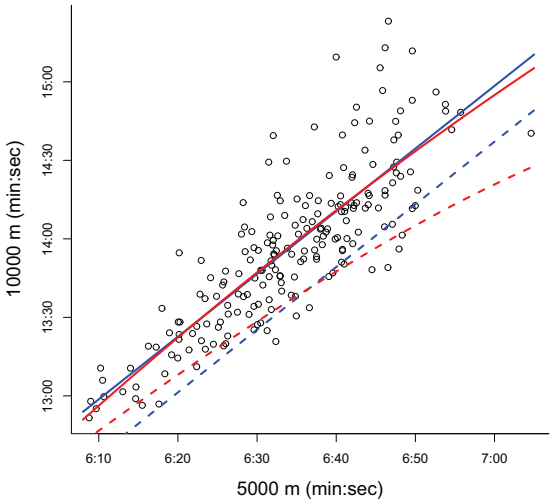
(3) Estimate $J_{\rm wide}$, and use this to obtain submatrices

(4) Estimate γ in the wide model, form $\widehat{\delta} = \sqrt{n}(\widehat{\gamma} - \gamma_0)$

(5) Estimate $\omega = J_{10} J_{00}^{-1} \frac{\partial \mu}{\partial \theta} - \frac{\partial \mu}{\partial \gamma}$

Speedskating data

Model	Var$_S$ − c	$\widehat{\mathrm{sqb}}_3(S)$	FIC	$\sqrt{\mathrm{MSE}/n}$	
10% quantile of 10km time for skater with 5km time = 6:35					
M_0	0.000	0.000	0.000	1.589	(1)
M_1	152.788	27.380	180.168	1.851	(4)
M_2	1.103	0.000	1.103	1.590	(2)
M_3	153.891	0.000	153.891	1.815	(3)
10% quantile of 10km time for better skater with 5km time = 6:15					
M_0	0.000	17749.40	17749.404	9.796	(3)
M_1	623.211	19027.45	19650.661	10.269	(4)
M_2	766.466	0.00	766.466	3.323	(1)
M_3	1389.677	0.00	1389.677	3.762	(2)



Personal best times (200 best skaters). Linear regression (M_0) and quadratic regression (M_2), with estimated 10% quantile for 10km, predicted from 5km results.

Application: FIC in logistic regression

Low birthweight data
Weight just prior to pregnancy (x_2), age (x_3), indicator for race 'black' (x_4), and for race 'other' (x_5). Constant $x_1 = 1$.

$$p(x,u) = \mathrm{P}\{\text{low birth weight} \mid x,u\} = \frac{\exp(x^t\beta + u^t\gamma)}{1+\exp(x^t\beta + u^t\gamma)},$$

where $x = (1, x_2)^t$ is always in the model while subsets of $u = (x_3, x_4, x_5)^t$ are considered for possible inclusion.
AIC: best submodel is 'x_4'.
BIC chooses the narrow model as the best one.

For a logistic regression model estimate in full model

$$J_{n,\mathrm{full}} = n^{-1}\sum_{i=1}^n p_i(1-p_i)\begin{pmatrix} x_ix_i^t & x_iu_i^t \\ u_ix_i^t & u_iu_i^t \end{pmatrix},$$

with $p_i = \exp(x_i^t\beta + u_i^t\gamma)/\{1+\exp(x_i^t\beta + u_i^t\gamma)\}$.
$\delta/\sqrt{n}$ measures the departure between narrow and true model and is estimated by $\hat\delta/\sqrt{n} = (\hat\gamma - \gamma_0)$.
Narrow model: $\gamma_0 = (0,0,0)^t$.
$\widehat\delta = \sqrt{n}\widehat\gamma = (-0.351, 13.799, 6.096)^t$.
Another use of $\widehat\delta$ is to test for $\gamma = 0$ inside the wide model, where the approximate χ^2_3 test statistic is $\widehat\delta^t\widehat{Q}^{-1}\widehat\delta = 5.927$,

Focus parameters

A first focus parameter is $p(x,u)$ itself, for different (x,u) corresponding to different strata of mothers.
A very specific model search for precise values of the covariates.

(1) Women of race 'white': $x=(1,132.05)^t$ and $u=(24.29,0,0)^t$, average weight and age in that group.
(2) Women of race 'black': $x=(1,146.81)^t$ and $u=(21.54,1,0)^t$; the average woman here is younger but a bit heavier than in the previous group.
(3) Women of race 'other': $x=(1,120.01)^t$ and $u=(22.39,0,1)^t$, average weight and age in that group.

Second focus parameter ratio $\mu=p(x',u')/p(x,u)$: (x',u') corresponding to the average black and (x,u) to the average white mother.

x_i	p(wh)	FIC	p(bl)	FIC	p(oth)	FIC	ratio	FIC
$\emptyset$	0.298	0.860	0.256	5.099	0.334	0.158	0.861	291.806
x_3	0.288	0.654	0.272	4.171	0.337	0.140	0.945	231.353
x_4	0.269	0.375	0.412	2.813	0.310	0.694	1.533	110.376
x_5	0.279	0.695	0.242	6.481	0.369	0.797	0.868	272.466
x_3,x_4	0.264	0.315	0.413	2.813	0.314	0.625	1.564	106.519
x_4,x_5	0.231	0.383	0.414	2.813	0.368	0.795	1.794	110.938
$x_{3,4,5}$	0.230	0.385	0.414	2.813	0.367	0.796	1.801	111.016

Estimates and FIC values for $p(\text{white})$, $p(\text{black})$, $p(\text{other})$ and the ratio $p(\text{black})/p(\text{white})$.

Need less focus?

One extreme: person specific focus and model search: FIC
Other extreme: no focus (AIC, BIC, . . .)
Large area in between
One possibility: AVERAGED FIC (Sec. 6.9)

- Keep the focus
- Average over the covariate space using weights

Minimise the estimated average MSE $\hookrightarrow$ AFIC.

$$L_n(S)=n\int\{\hat\mu_S(u)-\mu_{\text{true}}(u)\}^2\,dW_n(u),$$

Examples: (1) $W_n=$ empirical distribution, (2) W_n gives equal weight to the deciles $u=0.1,0.2,\ldots,0.9$ for estimation of the quantile distribution, (3) gliding covariate window, (4) robust weights,. . . . Easy expression of AFIC for GLM.

Model averaging

- What is model averaging?
- Do we ever need this?
- Why do Bayesians like this so much?

Model averaging estimators

Model selection schemes, like the AIC, BIC and the FIC, take the form

$$\widehat{\mu} = \sum_S c_n(S|\widehat{\delta}_{\rm full})\widehat{\mu}_S,$$

where $c_n(S|\widehat{\delta})$ is indicator for the chosen set (0/1 weight).

More generally: any data-dependent $c_n(S|\widehat{\delta})$ with sum 1.

May smooth across all models, or only over some of them, like the nested sequence $\emptyset, \{1\}, \{1,2\}, \ldots, {\rm full}$.

Main result for FMA

(Theorem 7.1) For a general FMA (frequentist model average) estimator, if $c_n(S|\widehat{\delta}) \stackrel{d}{\to} c(S|D)$ for each S,

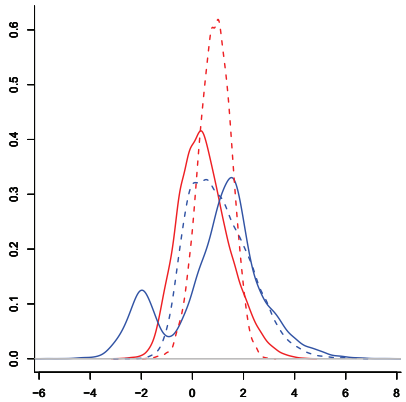
$$\sqrt{n}\Big\{\sum_S c(S|D_n)\widehat{\mu}_S - \mu_{\rm true}\Big\} \to_d \sum_S c(S|D)\Lambda_S$$

$$= (\tfrac{\partial\mu}{\partial\theta})^t J_{00}^{-1} M + \omega^t[\{I - G(D)\}^t\delta - G(D)^t W]$$

$M \sim N_p(0, J_{00}), W \sim N_q(0, Q)$ indep.; $G(d) = \sum_S c(S|d)G_S$

and $D \sim N_q(\delta, Q)$ with $Q = J^{11}$.

Example: Danish melanoma data



Densities Λ for $\sqrt{n}(\widehat{\mu}_2 - \mu_2)$, for FIC, AIC (solid), smoothed FIC, smoothed AIC (dashed).

Fig is based on 10 000 simulations from Λ at $\widehat{\delta} = \sqrt{n}\widehat{\gamma}_{\rm full}$.

Nonlinear combination of normals

Each Λ_S is normal, but random weights $c(S|D)$. Only if $c(S|D)$ are nonrandom constants $c(S)$ then limit Λ is again normal.

$$E(\Lambda) = \omega^t\left[\delta - \sum_S E\{c(S|D)G_S D\}\right]$$

$$\text{Var}(\Lambda) = \tau_0^2 + \omega^t \text{Var}\left\{\sum_S c(S|D)G_S D\right\}\omega$$

$$\text{MSE} = E(\Lambda^2) = \tau_0^2 + E\{\omega^t\widehat{\delta}(D) - \omega^t\delta\}^2.$$

Choice of weights

This is quite flexible. Often taken choices are

- 0/1 weights coming from model selection method, e.g. AIC, BIC, FIC
$$c(S) = I\{S = \widehat{S}_{\text{aic}}\}$$
- smoothed weights from model selection, e.g. AIC, BIC, FIC [*Akaike weights – Burnham & Anderson, 2002*]

$$c_{\text{aic}}(S) = \frac{\exp(-\frac{1}{2}AIC_S)}{\sum_{\text{all}S'}\exp(-\frac{1}{2}AIC_{S'})}, \quad c_{\text{bic}}(S) = \frac{\exp(-\frac{1}{2}BIC_S)}{\sum_{\text{all}S'}\exp(-\frac{1}{2}BIC_{S'})}$$

Remember from the derivation of the BIC:

$$P(M_j \mid Y) \approx \frac{P(M_j)\exp(-\frac{1}{2}\text{BIC}_{n,j})}{\sum_{j'=1}^{k} P(M_{j'})\exp(-\frac{1}{2}\text{BIC}_{n,j'})}.$$

Bayesian model averaging

- Specify the list of models used
$$\mathcal{M} = \{M_1, \ldots, M_k\}$$
(everything is conditional on this set).
- Set prior probabilities $P(M_j)$ for all models M_j.
- Set prior probabilities $\pi(\theta_j|M_j)$ for all parameters θ_j in M_j, for all $j = 1, \ldots, k$.

Then compute/simulate the posterior distribution of the parameter of interest (focus) μ.
Choice of priors

- Informative priors
- Noninformative priors Bayesian analysis independent of investigator's own beliefs.

Noninformative priors are often improper.

We need priors for the parameters inside each model:
For example Jeffreys' prior $f(\theta_j|M_j) \propto |\det(J_{\theta_j})|^{1/2}$
Prior probabilities for each of the models $M_1, \ldots, M_K$:

- $1/K$
- if $p_1 = \pi$ prior belief in M_1: $p_j = (1-\pi)/(K-1)$
- If nested models with $M_1 \subset M_2 \subset \ldots$ Smaller prior probability on models of larger dimension.
Jeffreys' improper prior $p_j = 1/(j+1)$
Rissanen's noninformative prior for integers:
$\pi(m) = p^m(1-p),\ m = 0, 1, \ldots$ with $p = p_1$.
- More than one model with the same dimension: equal probability. (Berger & Pericchi (1996, JASA).

Using Schwarz's approximation:

$$P(M_j|Y) \approx \exp\{-\frac{1}{2}\text{BIC}(M_j)\}$$

Using this as weights, and rescale such that the weights sum to one, we get that

$$w(M_j) = \frac{\exp\{-\frac{1}{2}\text{BIC}(M_j)\}}{\sum_{k=1}^{K}\exp\{-\frac{1}{2}\text{BIC}(M_k)\}}.$$

This motivates the smooth BIC weights in frequentist model averaging.
Easy to compute approximation to posterior probabilities.

Bayesian calculations

If Laplace approximation not used, need to compute the integrals directly,

$$A = \int_\Theta f(Y|M_j,\theta) f(\theta|M_j) \mathrm{d}\theta.$$

Taking θ as a random variable, and keeping Y and M_j fixed, $A = E_{\theta|Y,M_j}[f(Y|M_j,\theta)]$. This is used by the MCMC technique. If we sample a large number $\theta_{(k)}$ values from the distribution of θ given M_j,

$$A \approx \frac{1}{k_{\text{sim}}} \sum_{k=1}^{k_{\text{sim}}} f(Y|M_j,\theta_{(k)})$$

If sampling from $\theta|M_j$ is difficult, construct a Markov chain which converges to this distribution, e.g. Gibbs sampler, Metropolis-Hastings algorithms (see, e.g. Robert, 2001).

Data generating method

- Generate a model M_j from $\{M_1,\ldots,M_k\}$.
- Generate a parameter vector θ_j from $\pi(\theta_j|M_j)$.
- Generate data y from $P(Y|\theta_j,M_j)$.

Then compute the posterior probability $P(\mu|Y)$ using Bayes theorem.

$$P(\mu|Y) = \sum_{j=1}^{k} P(\mu|M_j,Y) P(M_j|Y)$$

with

$$P(M_j|Y) = \frac{P(M_j)\int L_j(y,\theta_j)\pi(\theta_j|M_j)d\theta_j}{\sum_{\ell=1}^{k}\int L_\ell(y,\theta_\ell)\pi(\theta_\ell|M_\ell)d\theta_\ell}.$$

Properties of the posterior density

- Posterior density of μ is a weighted average of conditional posterior densities
$$P(\mu|Y) = \textstyle\sum_{j=1}^{k} P(\mu|M_j,Y)P(M_j|Y).$$
- Posterior mean is a weighted average of posterior means in separate models
$$E(\mu|Y) = \textstyle\sum_{j=1}^{k} E(\mu|M_j,Y)P(M_j|Y).$$
- Posterior variance is a mixture

$$\begin{aligned}\mathrm{Var}(\mu|Y) &= \sum_{j=1}^{k} P(M_j|Y)\big[\mathrm{Var}(\mu|M_j,Y) + \\ &\quad \big\{E(\mu|M_j,Y) - E(\mu|Y)\big\}^2\big].\end{aligned}$$

No mistake of ignoring model uncertainty

What goes wrong?

. . . when we ignore model selection

- Asymptotic distribution after model selection is non-normal.
- Variance estimator in model $\widehat{S}$ underestimates true variance of $\widehat{\mu}_{\widehat{S}}$
- Estimator $\widehat{\mu}_{\widehat{S}}$ has a non-negligible bias because of the model selection.

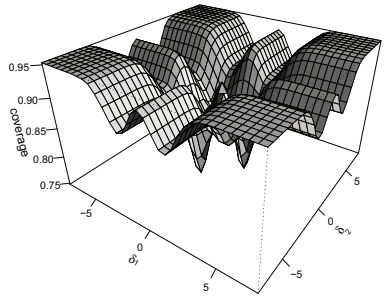
Ignoring uncertainties involved in model selection leads to too optimistic inference results.

Hence, “typical” confidence interval

$$\widehat{\mu}_{\widehat{S}} \pm 1.96\, \widehat{\sigma}_{\widehat{S}}/\sqrt{n}$$

can have much lower coverage than 95% because (1) variance estimate too small, (2) bias not taken into account, (3) critical value 1.96 from assumed normal distribution.
Similar problem with tests after model selection.
$\hookrightarrow$ Simulate from Λ distribution, or use biggest model’s variance + bias correction.

True coverage prob – naive use of AIC

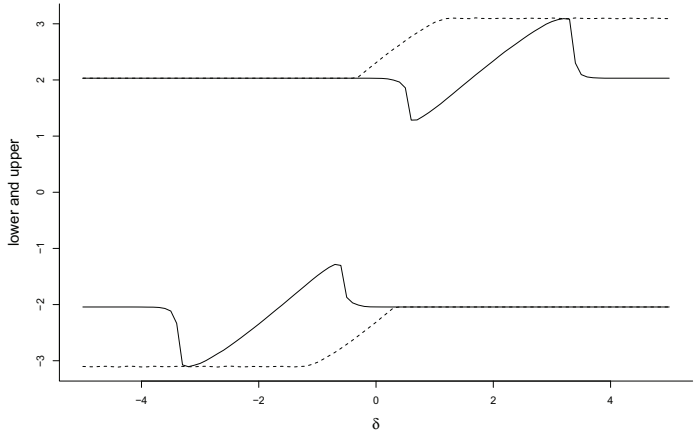


Intended coverage 0.95, AIC selects among four models, for $q = 2$. The situation corresponds to $\omega = (1,1)^t$ and $Q = \mathrm{diag}(1,1)$.

Better confidence intervals

- Try to correct for the variance (location might still be wrong)
- Correct the bias using the wide variance
  Estimate $E(\Lambda)$ by $\hat{\omega}^t\big(\hat{\delta} - \sum_S c(S|\hat{\delta})\hat{G}_S\hat{\delta}\big)$ and use $\hat{\tau}_{\mathrm{wide}}$.
  (Intervals are probably (much) too wide)
- Simulate from the Λ distribution
  $P\big(a(\delta) \le \Lambda(\delta) \le b(\delta)\big) = 0.95$

Pointwise $a(\delta)$ and $b(\delta)$



AIC selection between two models. Pointwise bounds that give for each δ 0.95 coverage for $\Lambda(\delta)$ (solid line).

Simulation methods

- One-stage simulation: plug in $\hat{\delta}$ (doesn't work well)
- Two-stage simulation: make confidence ellipsoid CE for δ, calculate $a(\delta)$ and $b(\delta)$ for each δ in CE

$$\begin{aligned}\hat{a} &= \min\{a(\delta) : \{(D_n-\delta)^t\hat{Q}^{-1}(D_n-\delta)\}^{1/2} \leq (\chi_{q,0.95})^{1/2}\}\\ \hat{b} &= \max\{b(\delta) : \{(D_n-\delta)^t\hat{Q}^{-1}(D_n-\delta)\}^{1/2} \leq (\chi_{q,0.95})^{1/2}\}\end{aligned}$$

$$\mathsf{CI}_n^* = [\hat{\mu}_{\text{avg}} - \frac{\hat{b}}{\sqrt{n}}, \hat{\mu}_{\text{avg}} - \frac{\hat{a}}{\sqrt{n}}]$$

Conservative confidence level.

Some points to remember & remaining issues

- Questions about 'which model is best' are difficult to answer. Conflicting recommendations might arise from different criteria. This stresses the importance of learning the aims and properties of the selection method.
- Not a single criterion can be best everywhere
- General weighting schemes allow to bypass the model selection step and work with averaged estimators.
- FMA distributions for post-model selection estimators take the model selection step into account. Use these!
- Correct inference methods for use in selected models is still an open problem.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Csáki, F., editors, *Second International Symposium on Information Theory*, pages 267–281. Akadémiai Kiadó, Budapest.

Akaike, H. (1978). A new look at the Bayes procedure. *Biometrika*, 65:53–59.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91:109–122.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference* (2nd Ed). Springer.

Claeskens, G. and Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association*, 98:900–916. With discussion and a rejoinder by the authors.

Claeskens, G. and Hjort, N.L. (2008). *Model Selection and Model Averaging*. Cambridge University Press.

Grünwald, P. (2007). *The Minimum Description Length Principle*. MIT Press.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B*, 41:190–195.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, 98:879–899. With discussion and a rejoinder by the authors.

Hosmer, D. W. and Lemeshow, S. (1999). *Applied Logistic Regression*. John Wiley & Sons Inc., New York.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76:297–307.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Rissanen, J. J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42:40–47.

Robert, C. P. (2001). *The Bayesian Choice*. Springer-Verlag, New York.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64:583–639. With a discussion and a rejoinder by the authors.

Spitzer, F. (1956). A combinatorial lemma and its applications to probability theory. *Transactions of the American Mathematical Society*, 82:323–339.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite sample corrections. *Communications in Statistics. Theory and Methods*, 7:13–26.

Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku (Mathematical Sciences)*, 153:12–18. In Japanese.

Woodroofe, M. (1982). On model selection and the arc sine laws. *The Annals of Statistics*, 10:1182–1194.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? *Biometrika*, 92:937–950.