4.3 Information Criteria

1. Introduction

When we compare different models, we would like to select **the best**, i.e. the one leading to the lowest test error.

RSS and R^2 are not suitable for this task, because they are based on the training sample and *the training error can be a poor estimate of the test error*.

In order to select the best model we can

- Estimate the test error *directly through cross-validation*
- Estimate the test error *by adjustment of the training error*

2. Mallow's C_p

We now present some techniques to **adjust the training error** for the model size.

For linear models with p predictors estimated through least squares, **Mallow's C_p** is

$$C_p = \overline{\mathrm{err}} + \frac{2p\hat{\sigma}^2}{N},$$

- $\overline{\mathrm{err}} = \mathrm{RSS}/N$
- $\hat{\sigma}^2$ is an estimate of the variance of the additive noise ε in the model $Y = f(\mathrm{X}) + \varepsilon$.

C_p gives an estimate of $\mathrm{Err}_{\mathrm{in}}$ (In-sample prediction error), and is hence a criterion for model selection: *Choose the model complexity with **minimal** C_p.*

It adds a penalisation $\frac{2p\hat{\sigma}^2}{N}$ to the training error, because it tends to underestimate the test error: the penalisation increases as the number of predictors increases.

3. AIC and BIC

It is useful to consider a range of information criteria, since sometimes you will obtain conflicting information based on some of these criteria. Therefore, the more information criteria you use in your analysis, the more insight you will possess for model selection.

Akaike Information Criterion

Recall that for a linear model RSS equals (up to a constant) the negative log-likelihood. For likelihood based models, Mallow's Cp then generalises to the **Akaike Information Criterion** (AIC):

$$\mathrm{AIC} = -2\sum_{i=1}^{N} \ell(\hat{\boldsymbol{\beta}}, y_i) + 2d, \tag{4.3.1}$$

which also aims to minimise the Generalisation Error, asymptotically for large N. Here d is the number of estimated parameters in the fitted model. Example: for Gaussian regression we have $d = p + 2$, since $\beta_0, \beta_1, \ldots, \beta_p, \sigma^2$ are estimated in the fitted model.

Hurvich and Tsai (1989) developed AICc, a version of the AIC which is asymptotically equivalent to AIC as $N \to \infty$ but corrects for a bias at small sample sizes:

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2d(d+1)}{N-d-1}$$

Burnham and Anderson (2004) recommend that AICc be used instead of AIC unless $N/d > 40$.

Bayesian Information Criterion

The **Bayesian information criterion** (BIC) is

$$\mathrm{BIC} = -2\sum_{i=1}^{N} \ell(\hat{\beta}, y_i) + d\log N. \tag{4.3.2}$$

It penalises complexity more strongly than AIC if $N \geq 8$, since then $\log N > 2$. The model with the lowest BIC corresponds to the model with the highest log-posterior probability, assuming an indifferent prior.

4. R^2-adjusted criterion

R^2 is unsuitable for the choice of model parameters, since it typically increases with every added parameter. The adjusted R^2 is

$$\mathrm{R}^2_{adj} = 1 - (1 - \mathrm{R}^2)(n-1)/(n-p-1)$$

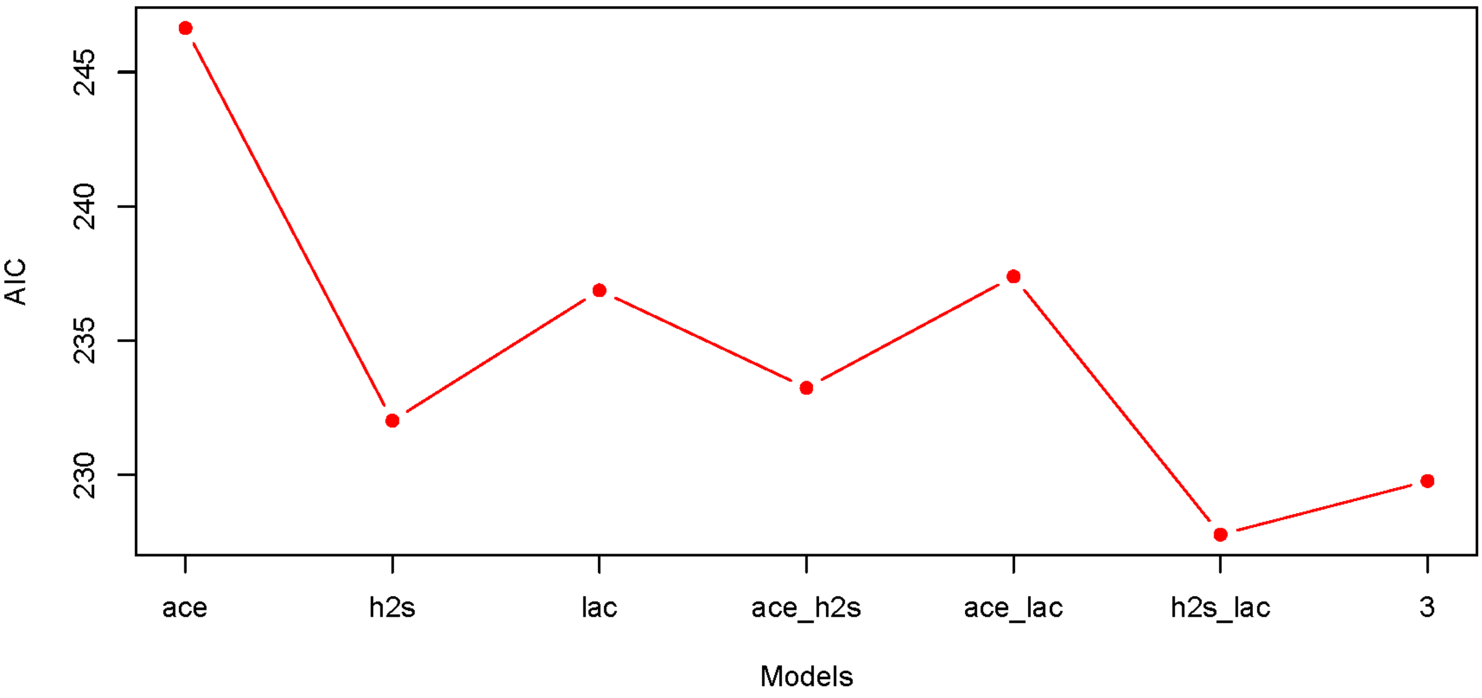
and using the fact that $\mathrm{R}^2 = 1 - \mathrm{RSS}/\mathrm{TSS}$

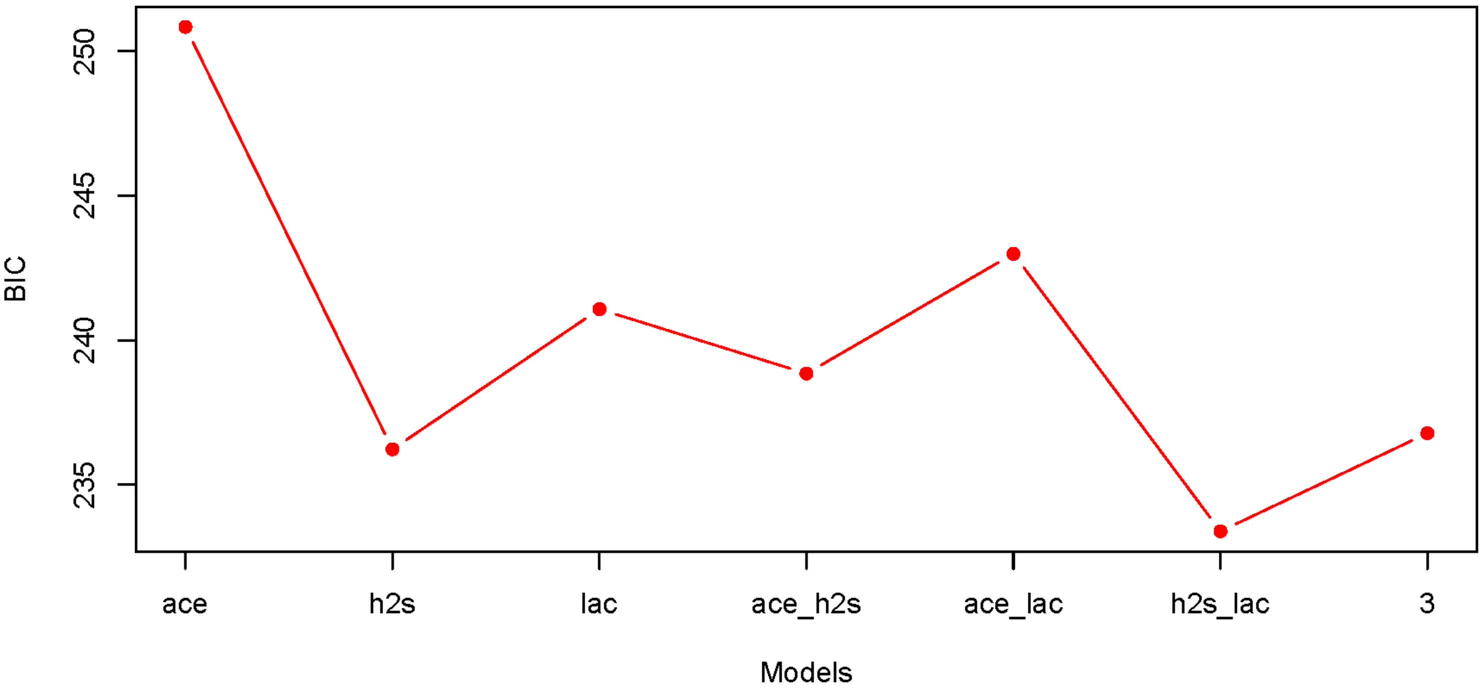
$$\mathrm{R}^2_{adj} = 1 - \frac{\mathrm{RSS}/(n-p-1)}{\mathrm{TSS}/(n-1)}$$

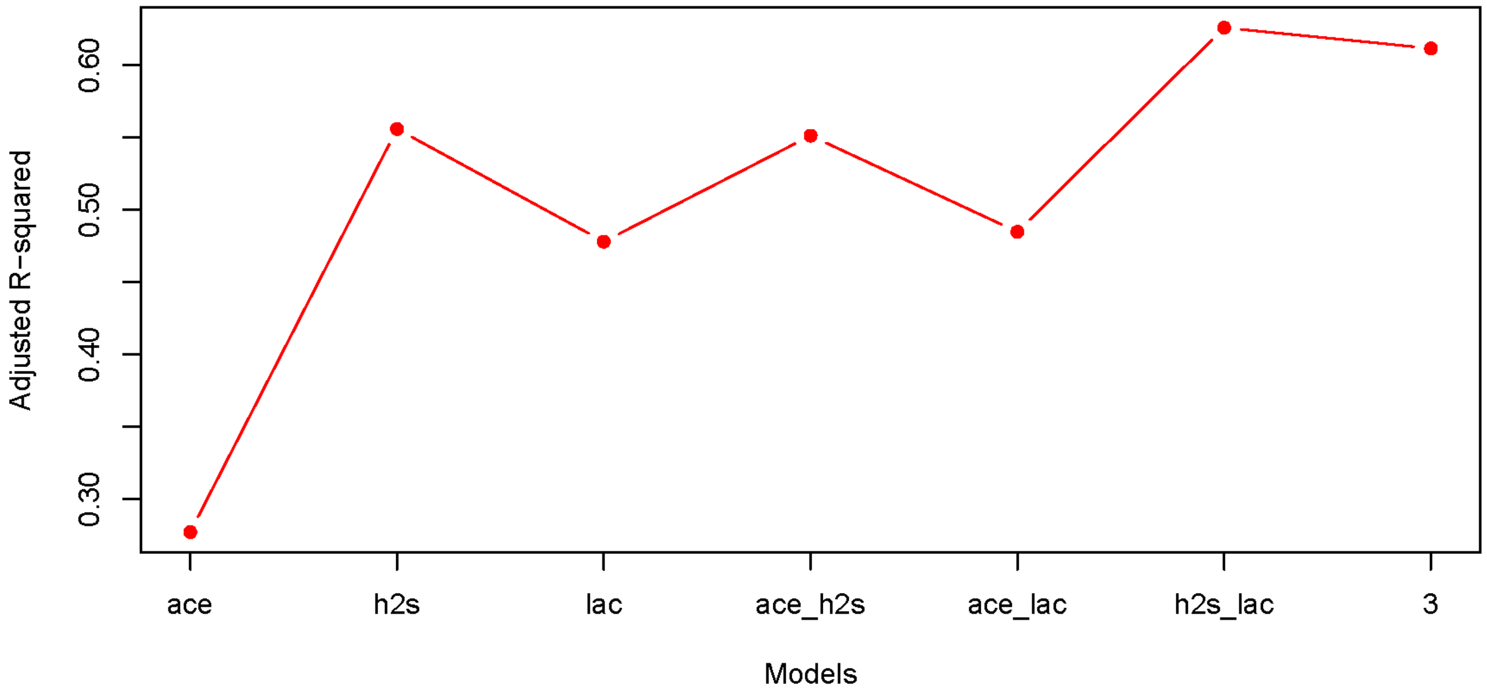
where p is the number of predictors in the current model. It can be used for model selection, but still tends to "overfit".

5. Example: Cheddar cheese

Going back to the **`cheddar`** dataset, we evaluate the AIC, BIC and adjusted R^2 for each of the proposed models.







```
library(faraway)

data("cheddar")

res_lm_ace <- lm(taste ~ Acetic, data=cheddar)
res_lm_h2s <- lm(taste ~ H2S, data=cheddar)
res_lm_lac <- lm(taste ~ Lactic, data=cheddar)
res_lm2_ace_h2s <- lm(taste ~ Acetic + H2S, data=cheddar)
res_lm2_ace_lac <- lm(taste ~ Acetic + Lactic, data=cheddar)
res_lm2_h2s_lac <- lm(taste ~ H2S + Lactic, data=cheddar)
res_lm3 <- lm(taste ~ ., data=cheddar)

AIC <- AIC(res_lm_ace, res_lm_h2s, res_lm_lac, res_lm2_ace_h2s,
           res_lm2_ace_lac, res_lm2_h2s_lac, res_lm3)
BIC <- BIC(res_lm_ace, res_lm_h2s, res_lm_lac, res_lm2_ace_h2s,
           res_lm2_ace_lac, res_lm2_h2s_lac, res_lm3)

Adj.R2 <- vector(length=7)
Adj.R2[1] <- summary(res_lm_ace)$adj.r.squared
Adj.R2[2] <- summary(res_lm_h2s)$adj.r.squared
Adj.R2[3] <- summary(res_lm_lac)$adj.r.squared

Adj.R2[4] <- summary(res_lm2_ace_h2s)$adj.r.squared
Adj.R2[5] <- summary(res_lm2_ace_lac)$adj.r.squared
Adj.R2[6] <- summary(res_lm2_h2s_lac)$adj.r.squared
Adj.R2[7] <- summary(res_lm3)$adj.r.squared

models <- c("ace", "h2s", "lac", "ace_h2s", "ace_lac","h2s_lac", "3")

plot(AIC$AIC, xaxt="n", xlab="Models", ylab="AIC", col=2, type="b", pch=16)
axis(1, at=1:7, labels=models)

plot(BIC$BIC, xaxt="n", xlab="Models", ylab="BIC", col=2, type="b", pch=16)
axis(1, at=1:7, labels=models)
```

```
plot(Adj.R2, xaxt="n", xlab="Models", ylab="Adjusted R-squared", col=2,
     type="b", pch=16)
axis(1, at=1:7, labels=models)
```

6. Activity in R: Model Selection

Consider the **`swiss`** dataset in R including standardised fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888.

```
head(swiss)
```

```
             Fertility  Agriculture  Examination  Education  Catholic
Courtelary        80.2         17.0           15         12      9.96
Delemont          83.1         45.1            6          9     84.84
Franches-Mnt      92.5         39.7            5          5     93.40
Moutier           85.8         36.5           12          7     33.77
Neuveville        76.9         43.5           17         15      5.16
Porrentruy        76.1         35.3            9          7     90.57
             Infant.Mortality
Courtelary               22.2
Delemont                 22.2
Franches-Mnt             20.2
Moutier                  20.3
Neuveville               20.6
Porrentruy               26.6
```

Calculate AIC and BIC for the multiple regression model with **`Fertility`** as a response and all other variables as explanatory variables.

Additionally, consider a model with **`Examination`** removed from the set of the explanatory variables. Compare the AIC and BIC for the simplified model and the full model.

Hint: Use functions $AIC()$ and $BIC()$ for this exercise.

Additional Activity

Question *Submitted Feb 7th 2024 at 3:58:01 pm*

Which of the following statements are true about the information criteria introduced in this section:

☐ R^2 is used for model selection from models with differing number of variables: the higher R^2 indicates better fitting model;

☑ Mellow's Cp gives an estimate of Err_{in};

☑ for likelihood based models, Mallow’s Cp generalises to the Akaike Information Criterion;

☐ AIC penalises complexity more strongly than BIC if $N \geq 8$.