Best Subset Selection

If $y_1, \ldots, y_N \in \mathbb{R}$ and $x_1, \ldots, x_N \in \mathbb{R}^p$ and

$$\text{RSS}(q) = \min_{i_1,\ldots,i_q} \min_{\beta_1,\ldots,\beta_q} \sum_{j=1}^{q} (y_j - (\beta_1 x_{j,i_1} + \ldots + \beta_q x_{j,i_q}))^2$$

is the least residual sum of squares for using any q dimensional submodel, then *what q should I choose*?

RSS is a relevant measure for comparison within all q-dimensional models, but decreases monotely with q.

Model Selection

The most important question that we have not dealt with yet is:

How do we select an appropriate model among several models of different/incommensurable complexity?

This is *model selection*. What is this besides an estimation problem?

Model Selection

We can always put all considered models into a single parameter set Θ and consider predictors f_θ for $\theta \in \Theta$. With e.g.

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f_\theta(x_i))$$

the empirical loss minimizer $f_{\hat{\theta}}$ could be a bad predictor. We divide the parameter space $\Theta = \Theta_0 \cup \Theta_1$ into disjoint sets.

- *Asymmetric test theoretic approach*: $\hat{\theta}_1 \in \Theta_1$ is preferred over $\hat{\theta}_0 \in \Theta_0$ only if it is improbable as measured by a p-value that $\hat{\theta}_0$ is as good as $\hat{\theta}_1$.
- *Symmetric prediction based approach*: Consider

  $$\text{EPE}(f_{\hat{\theta}_i}) = E(L(Y, f_{\hat{\theta}_i}(X))|\mathbf{X}, \mathbf{Y})$$

  and choose the one with the least expected prediction error.

The test theoretic conclusions are of interest for several reasons. In the following discussion of test theoretic interpretations you should have a linear regression analysis in mind and the hypothesis we test is whether a single regression parameter is equal to 0. The primary interest in a test is in the qualitative conclusions that we can draw from a test. If we reject the test we conclude that one variable does have a statistically significant effect on a response variable. Thus we can not disregard the variable in our model, and if we did the model would be biased. If data are from a carefully designed experiment we might even conclude that the variable has a specific causal effect on the response, and the estimated parameter can even provide a quantitative measure of the effect. We often talk about *confirmatory data analysis*

where we confirm an effect that is hypothesized prior to the data collection by rejecting the hypothesis that the effect is not present. If data are from an observational study we can only conclude that the variable has a non-vanishing correlation – conditionally on all other regressors – with the response. Formally this is still a confirmatory data analysis if we have a well specified hypothesis about correlation that we set up prior to the data collection, but in this case we do not get a quantitative estimate of a causal effect by intervention, say, only a quantitative measure of the (conditional) correlation for new observations sampled in the same manner.

If we don't reject the test the conclusion is that a variable does not have a statistically significant effect on a response variable. This is a vaguer conclusion, and we can not rule out that there is in fact a small affect, but if the effect is not significant it can be explained completely as a random error. Since more free parameters generally result in an increased variance on the parameter estimates we would typically disregard a non-significant variable and report only parameter estimates for significant parameters. Note, however, that theory almost never support practice. It is often the case that $\hat{\theta}_0$ is almost unbiased *conditionally* on the test statistic under the hypothesis that $\theta \in \Theta_0$. However, $\hat{\theta}_1$ is rarely unbiased *conditionally* on the test statistic. This means that if the hypothesis is true, $\hat{\theta}_0$ is a sensible estimator – even if we only consider it conditionally on having accepted the test. However, $\hat{\theta}_1$ is biased conditionally on rejecting the test. Thus if we carry out multiple, sequential tests and stop when can not accept further model reductions the resulting estimator is biased. A second problem is that p-values used on the way in such a sequential procedure are generally wrong.

In conclusion, the theory for statistical tests is poorly developed for handling any serious model selection problem and does not in its current form support any commonly used method. This is the central point when discussing *explorative data analysis* versus *confirmatory data analysis*. The test methodology is most appropriate for testing an a priory specified hypothesis and not for exploring which hypotheses the data support. The problem of model selection belongs more naturally to the world of explorative data analysis where we do not know a priory which models the data will support.

Optimization versus Simplicity

One can say that we have two opposing philosophical directions of how to draw inference from empirical data:

- Trust all aspects of the information in the data on the relation between y and x as implemented in *parameter estimators* based on pure *optimization*.
- Keep it simple. Don't choose a complicated model over a simpler model if the simpler model suffice (Occam's razor).

Model selection/test theory work by the second principle to compensate for the fact that the optimization procedure may overfit to the given dataset.

Model or Predictor Assessment

Another problem of some importance is

Having fitted a final predictor $\hat{f}$, how will it actually perform?

We have the *training error*

$$\bar{\text{err}} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

which generally *underestimates* $\text{EPE}(\hat{f})$. The *generalization* or *test* error is

$$\text{Err} = E(\text{EPE}(\hat{f}))$$

is the expected EPE.

$\text{EPE}(\hat{f})$ can only really be estimated if we have an independent *test dataset*.

In the book at this point they introduce the notion of conditional test error – conditioning on the training data. This is nothing but the expected prediction error for the estimated predictor $\hat{f}$.

Figure 7.1 – The Bias-Variance Tradeoff

Realizations of training error $\bar{\text{err}}$ and expected prediction error $\text{EPE}(\hat{f})$ estimated on an independent test dataset as functions of model complexity. Also the estimates of the expectation of $\bar{\text{err}}$ and $\text{EPE}(\hat{f})$ are shown.

The Train-Validate-Test Idea

In a data rich situation we split the data *before doing anything else* into three subsets.

- On the *training* data we estimate all parameters besides tuning parameters (model complexity parameters).
- On the *validation* data we estimate prediction error for the estimated predictors and optimize over tuning parameters and models.
- On the *test* data we estimate the expected prediction error for the chosen predictor – no model selection here, please.

Problem: We are almost never in a data rich situation.

Can we justify to throw away data that can be used for estimation and thus reduction of variance to estimate parameters of secondary importance?

Figure 7.2 – Space of Models

Digesting Figure 7.2 provides a core understanding of the bias-variance tradeoff between complex and simple models. This understanding should be obtained in close connection with reading about the bias-variance decomposition in Section 7.3. One should note that the nice additive decomposition into a squared bias term and a variance term of the expectation of the prediction error in x_0 is a consequence of the choice of the loss function being the squared error loss. For the 0-1 loss often used in classification things work out differently, see Exercise 7.2.

Figure 7.3 – Quadratic Loss vs. 0-1 Loss

Setup

In the following discussion $(X_1, Y_1), \ldots, (X_N, Y_N)$ denote N i.i.d. random variables, with X_i a p-dimensional vector.

A concrete realization is denoted $(x_1, y_1), \ldots, (x_N, y_N)$ and we use boldface, e.g. $\mathbf{Y} = (Y_1, \ldots, Y_N)^T$ and $\mathbf{y} = (y_1, \ldots, y_N)^T$ to denote vectors.

We can not distinguish in notation between $\mathbf{X}$ – the matrix of random variables $X_1, \ldots, X_N$ – and $\mathbf{X}$ – the matrix of a concrete realization $x_1, \ldots, x_N$.

Mallows' C_p

With $\hat{\mathbf{f}} = P\mathbf{y}$ where P is a projection onto a d-dimensional subspace define

$$\bar{\text{err}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\mathbf{f}}_i)^2 = \frac{1}{N}||\mathbf{y} - \hat{\mathbf{f}}||^2.$$

By a standard decomposition

$$\frac{1}{N}E(||\mathbf{Y}^{\text{new}} - \hat{\mathbf{f}}||^2|\mathbf{X}) = \frac{1}{N}E(||\mathbf{Y} - \hat{\mathbf{f}}||^2|\mathbf{X}) + \frac{2d}{N}\sigma^2$$

The *in-sample error*

$$\text{Err}_{\text{in}} = \frac{1}{N}E(||\mathbf{Y}^{\text{new}} - \hat{\mathbf{f}}||^2|\mathbf{X})$$

can thus be estimated by

$$C_P = \hat{\text{Err}}_{\text{in}} = \bar{\text{err}} + \frac{2d}{N}\hat{\sigma}^2.$$

Note that $\mathbf{Y}^{\text{new}} - P\mathbf{Y}$ and $\mathbf{Y} - P\mathbf{Y}^{\text{new}}$ have the same conditional distributions given $\mathbf{X}$ and note that $\mathbf{Y} - P\mathbf{Y}^{\text{new}} = (I-P)\mathbf{Y} + P(\mathbf{Y} - \mathbf{Y}^{\text{new}})$ where the two terms are orthogonal. This implies that

$$\begin{aligned} E(||\mathbf{Y}^{\text{new}} - P\mathbf{Y}||^2|\mathbf{X}) &= E(||\mathbf{Y} - P\mathbf{Y}^{\text{new}}||^2|\mathbf{X}) \\ &= E(||(I-P)\mathbf{Y}||^2|\mathbf{X}) + E(||P(\mathbf{Y} - \mathbf{Y}^{\text{new}})||^2|\mathbf{X}) \end{aligned}$$

Since the vector $\mathbf{Y} - \mathbf{Y}^{\text{new}}$ has (conditional) mean 0 and (conditional) covariance matrix $2\sigma^2 I$ the second expectation above equals $2\sigma^2 d$.

Mallows' C_p

$$C_P = \hat{\mathrm{Err}}_{\mathrm{in}} = \bar{\mathrm{err}} + \frac{2d}{N}\hat{\sigma}^2.$$

is an equivalent of *Mallows' C_p statistics* – with $\hat{\sigma}^2$ estimated from a "low-bias" model with p degrees of freedom;

$$\hat{\sigma}^2 = \frac{1}{N-p}||\mathbf{y} - Q\mathbf{y}||^2$$

where Q is a projection on a p-dimensional space.

If $\mathbf{S}_\lambda$ is a smoother and $\hat{\mathbf{f}} = \mathbf{S}_\lambda \mathbf{y}$ one can generalize C_p as

$$\hat{\mathrm{Err}}_{\mathrm{in}} = \bar{\mathrm{err}} + \frac{2\mathrm{trace}(\mathbf{S}_\lambda)}{N}\hat{\sigma}^2$$

with $\hat{\sigma}^2$ estimated from a "low-bias", or small λ, model, e.g.

$$\hat{\sigma}^2 = \frac{1}{N - \mathrm{trace}(2\mathbf{S}_\lambda - \mathbf{S}_\lambda^2)}||\mathbf{y} - \mathbf{S}_\lambda \mathbf{y}||^2.$$

The complete justification of the above generalization of Mallows' C_p to general smoothers is sketched in the book and treated in Q 1.1-1.3 in Theo.6.

Using C_p

The classical use of C_p is when $\mathbf{X}$ is $N \times p$ of rank p and $Q = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

For any choice of d columns we compute C_p and select the model with the smallest value of C_P.

This is equivalent to best subset selection for each d followed by choosing d that minimizes

$$NC_p = \mathrm{RSS}(d) + \frac{2d}{N-p}\mathrm{RSS}(p)$$

As a function of d the normal defintion of C_p,

$$\tilde{C}_p = \frac{NC_p(N-p)}{\mathrm{RSS}(p)} - N = \frac{(N-p)\mathrm{RSS}(d)}{\mathrm{RSS}(p)} + 2d - N,$$

is a monotonely increasing function of C_p.

Minimizing C_p or $\tilde{C}_p$ is equivalent.

The correct, historical definition of Mallows' C_p is as $\tilde{C}_p$ above in the framework of multiple linear regression, see e.g. Wikipedia. When used as a model selection tool in this framework we can just as well consider C_p as we have defined. They select the same models. Our C_p is, however, easier to generalize and compare to other methods.

Generalization Error

Instead of the in-sample error we can consider the *generalization* or *test error*

$$\text{Err} = E(L(Y, \hat{f}(X))) = E(E(L(Y, \hat{f}(X))|\mathbf{X}, \mathbf{Y})) = E(\text{EPE}(\hat{f}))$$

Here (X, Y) is independent of $(X_1, Y_1), \ldots, (X_N, Y_N)$ that enter through $\hat{f}$.

Err is the *expectation* over the dataset of the *expected prediction error* for the estimated predictor $\hat{f}$.

A small value of Err tells us that *the estimation methodology is good* and will on average result in estimators with a small EPE. It *does not* guarantee that a concrete realization $\hat{f}$ has a small EPE!

In-sample Error and Generalization Error

Note the decomposition

$$\begin{aligned}
\text{Err} &= E(L(Y, \hat{f}(X))) \\
&= E(E(L(Y, \hat{f}(X))|\mathbf{X}, X)) \\
&\simeq E\left(\frac{1}{N}\sum_{i=1}^{N} E(L(Y, \hat{f}(x_i))|\mathbf{X}, X = x_i)\right) \\
&= E\left(\frac{1}{N}\sum_{i=1}^{N} E(L(Y_i^{\text{new}}, \hat{f}(x_i))|\mathbf{X})\right) \\
&= E\left(\text{Err}_{in}\right)
\end{aligned}$$

If the p-dimensional model is unbiased $E(\hat{\sigma}^2|\mathbf{X}) = \sigma^2$ and

$$E(C_p) = E(\hat{\text{Err}}_{\text{in}}) = E(\bar{\text{err}}) + \frac{2d}{N}\sigma^2 = E(\text{Err}_{\text{in}}),$$

in which case C_p can also be seen as an estimator of Err.

The approximation above, $\simeq$, is an approximation of the true, unknown distribution of X by the empirical distribution $\varepsilon_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}$. This may in fact not be a completely innocent approximation.

Linear Smoothers

For $\hat{\mathbf{f}} = \mathbf{S}_\lambda \mathbf{Y}$ with $\mathbf{S}_\lambda = \mathbf{S}_\lambda(\mathbf{X})$ *not* depending upon $\mathbf{Y}$ we had

$$\hat{\text{Err}}_{\text{in}} = \bar{\text{err}} + \frac{2\text{trace}(S)}{N}\hat{\sigma}^2,$$

It justifies the definition of trace$(\mathbf{S}_\lambda)$ as the *effective degrees of freedom* for model selection – but trace$(\mathbf{S}_\lambda)$ is now $\mathbf{X}$-dependent.

The fact that trace$\mathbf{S}_\lambda$ is $\mathbf{X}$-dependent makes it somewhat more difficult to try to relate the estimate of the in-sample error to the generalization error. For C_p the only approximation involved is the approximation of the marginal distribution of X by the empirical distribution of the observed values $x_1, \ldots, x_N$. For the general, linear smoother we can not easily decouple trace$\mathbf{S}_\lambda$ and $\hat{\sigma}^2$ – not even if we are able to construct an unbiased estimator $\hat{\sigma}^2$ of σ^2.

Likelihood Loss

The *generalized* decision theoretic setup has sample spaces E and F, *action space* $\mathcal{A}$, *decision rule* $f : E \to \mathcal{A}$ and loss functions $L : F \times \mathcal{A}$. If h_a for $a \in \mathcal{A}$ denotes a collection of densities on F we define the *minus-log-likelihood* loss function as

$$L(y, a) = -\log h_a(y)$$

The empirical loss for $(x_1, y_1), \ldots, (x_N, y_N)$ when using decision rule f is

$$\frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) = -\log \prod_{i=1}^{N} h_{f(x_i)}(y_i)$$

With $\mathcal{F}$ a class of decision rules *empirical risk minimization* over $\mathcal{F}$ coincides with *conditional maximum likelihood estimation* of $f \in \mathcal{F}$. Expected prediction error equals the expectation of (conditional) cross entropies.

$$\mathrm{EPE}(f) = \int \underbrace{\int -\log h_{f(x)}(y) g(y|x) \mathrm{d}y}_{\text{cross entropi}} g_1(x) \mathrm{d}x$$

The standard example in this context of the need for the general setup as compared to the setup where $\mathcal{A} = F$ and f is simply the predictor is when F is discrete. For instance, if $F = \{0, 1\}$ we might want "the action space" to be the set of probability measures on F – represented as $\mathcal{A} = [0, 1]$ and $p \in [0, 1]$ is the probability of $Y = 1$. A "decision" can then be the computation of $f(x) = \Pr(Y = 1|X = x)$ – the conditional probability that $Y = 1$ given $X = x$. What is perhaps more common in this case is that $\mathcal{A} = \mathbb{R}$ and a "decision" is the computation of the logit of $\Pr(Y = 1|X = x)$, that is

$$f(x) = \mathrm{logit}(\Pr(Y = 1|X = x)) = \log \frac{\Pr(Y = 1|X = x)}{\Pr(Y = 0|X = x)}$$

the log-odds of $Y = 1$ conditionally on $X = x$.

Akaike's Information Criteria – AIC

We take $\mathcal{A} = \{f_\theta(x, \cdot)\}_{\theta \in \Theta, x \in E}$ with Θ being d-dimensional and $f_\theta : E \times F \to [0, \infty)$ such that $f_\theta(x, \cdot)$ is a probability density on F. Let $\hat{\theta}_N$ denote the MLE.

With likelihood loss we define the equivalent of the in-sample error

$$\mathrm{Err}_{\mathrm{loglik,in}} = -\frac{1}{N}\sum_{i=1}^{N} E(\log f_{\hat{\theta}_N}(x_i, Y_i^{\mathrm{new}})|\mathbf{X})$$

Then one derives the approximation

$$\mathrm{Err}_{\mathrm{loglik,in}} \simeq \frac{1}{N} E(l_N(\hat{\theta}_N)) + \frac{d}{N}$$

where the minus-log-likelihood function in $\hat{\theta}_N$

$$l_N(\hat{\theta}_N) = -\frac{1}{N}\sum_{i=1}^{N} \log f_{\hat{\theta}_N}(x_i, y_i)$$

is the equivalent of $\bar{\text{err}}$ when using likelihood loss.

We let $Y_1, \ldots, Y_N$ and $Y_1^{\text{new}}, \ldots, Y_N^{\text{new}}$ be conditionally independent with the same distribution given $X_1 = x_1, \ldots, X_N = x_N$. The minus-log-likelihood

$$l_N(\theta) = -\sum_{i=1}^{N} \log f_\theta(x_i, Y_i)$$

and the minus-log-likelihood for the new data

$$l_N^*(\theta) = -\sum_{i=1}^{N} \log f_\theta(x_i, Y_i^{\text{new}}).$$

Letting $\hat{\theta}_N$ denote the MLE for the original dataset and $\tilde{\theta}_N$ the MLE for the new dataset then a Taylor expansion of l_N^* around $\tilde{\theta}_N$ yields

$$l_N^*(\hat{\theta}_N) = l_N^*(\tilde{\theta}_N) + \frac{1}{2}(\hat{\theta}_N - \tilde{\theta}_N)^T D^2 l_N(\tilde{\theta}_N)(\hat{\theta}_N - \tilde{\theta}_N) + \text{remainder}_N.$$

Under suitable regularity assumptions there is a θ_0 such that

$$\frac{1}{\sqrt{N}} D_\theta l_N(\theta_0)^T \xrightarrow{\mathcal{D}} N(0, K(\theta_0))$$

and

$$\frac{1}{N} D_\theta^2 l_N(\theta_0) \xrightarrow{P} I(\theta_0)$$

and the two esimators are independent and asymptotically $N(\theta_0, \frac{1}{N} I(\theta_0)^{-1} K(\theta_0) I(\theta_0)^{-1})$-distributed. Consequently

$$\sqrt{N}(\hat{\theta}_N - \tilde{\theta}_N) \xrightarrow{\mathcal{D}} N(0, 2I(\theta_0)^{-1} K(\theta_0) I(\theta_0)^{-1})$$

$$\begin{aligned} E\left(\frac{1}{N} l_N^*(\hat{\theta}_N) | \mathbf{X}\right) &\simeq E\left(\frac{1}{N} l_N^*(\tilde{\theta}_N) | \mathbf{X}\right) + \frac{1}{N} \text{trace}(E\left(N(\hat{\theta}_N - \tilde{\theta}_N)(\hat{\theta}_N - \tilde{\theta}_N)^T | \mathbf{X}\right) I(\theta_0)) \\ &\simeq E\left(\frac{1}{N} l_N(\hat{\theta}_N) | \mathbf{X}\right) + \frac{1}{N} \text{trace}(I(\theta_0)^{-1} K(\theta_0)) \end{aligned}$$

$E\left(\frac{1}{N} l_N^*(\hat{\theta}_N) | \mathbf{X}\right)$ is for the likelihood loss the equivalent of the in-sample error for quadratic loss. If $I(\theta_0) = K(\theta_0)$ the trace simplifies to the trace of the $d \times d$ identity matrix and is thus equal to d. This always happens if Θ contains the true parameter. To make likelihood loss for the Gaussian model (with known variance) equivalent to squared error loss we usually multiply everything by 2 and define the esimator of twice this in-sample error as

$$\text{AIC} = \frac{2}{N} l_N(\hat{\theta}_N) + \frac{2d}{N}.$$

For a more general quantity that does not rely on the model being true we need to replace d by $\text{trace}(I(\theta_0)^{-1} K(\theta_0))$ with the latter quantity having the obvious draw-back that it depends upon the unknown matrices $I(\theta_0)$ and $K(\theta_0)$, which have to be estimated also. Simple estimators are

$$\hat{K} = \frac{1}{N} \sum_{i=1}^{N} D_\theta \log f_{\hat{\theta}_N}(x_i, y_i)^T D_\theta \log f_{\hat{\theta}_N}(x_i, y_i) \quad \text{and} \quad \hat{I} = \frac{1}{N} D_\theta^2 l_N(\hat{\theta}_N)$$

which gives

$$\text{NIC} = \frac{2}{N} l_N(\hat{\theta}_N) + \frac{2}{N} \text{trace}(\hat{I}^{-1}\hat{K}).$$

If the distribution of Y given $X = x$ is $N(f(x), \sigma^2)$ for an unknown mean value function $f(x)$, if we take $\Theta = \mathbb{R}^p$, if we assume that σ^2 is fixed, and if we let $f_\theta = X^T\theta$ then

$$2\sigma^2 l_N(\hat{\theta}_N) = ||\mathbf{y} - \mathbf{X}\hat{\theta}_N||^2$$

and we see in this case that $\sigma^2\text{AIC} = C_p$. We derived C_p and thus AIC as a valid model selection quantity even if the model, as in this general case, is wrong. It is no problem to show explicitly (and perhaps surprisingly) in this case that the identity $I(\theta_0) = K(\theta_0)$ in fact holds. Here θ_0 is the θ that minimizes $E((f(X) - X^T\theta)^2)$. If we consider logistic regression instead this result does not hold. For logistic regression let $p(x) = \Pr(Y = 1|X = x)$ denote the true conditional probability and let W denote the $N \times N$ diagonal matrix with $p(x_i)(1-p(x_i))$ in the diagonal. Then it is straight forward to show that $I(\beta_0) = \mathbf{X}^T W(\beta_0)\mathbf{X}$ but $K(\beta_0) = \mathbf{X}^T W\mathbf{X}$ – which does not depend upone β_0 – hence

$$\text{trace}(I(\theta_0)^{-1}K(\theta_0)) = \text{trace}((\mathbf{X}^T W(\beta_0)\mathbf{X})^{-1}\mathbf{X}^T W\mathbf{X}).$$

With

$$p_\beta(x) = \frac{\exp((1, x^t)\beta}{1 + \exp((1, x^t)\beta)}$$

the β_0 is the minimizer of

$$E(-p(X)\log p_\beta(X) - (1-p(X))\log(1-p_\beta(X))) = E(-p(X)(1, X^T)\beta + \log(1+\exp((1, X^T)\beta))).$$

One good starting point for a more theoretical treatment of AIC and other aspects of statistical decision theory and model selection is *Pattern Recognition and Neural Networks* by Brian D. Ripley. I have also heard very positive things about the relatively new book *Model selection and model averaging* by Gerda Claeskens and Nils Lid Hjort, but I have not yet read it myself.

AIC

$$\text{AIC} = \frac{2}{N} l_N(\hat{\theta}_N) + \frac{2d}{N}$$

We use AIC for model selection by choosing the model among several possible that *minimizes* AIC.

Assumptions and extensions:

- The models considered *must be true*. If they are *not*, d must in general be replaced by a more complicated quantity d^* leading to the model selection criteria

$$\text{NIC} = \frac{2}{N} l_N(\hat{\theta}_N) + \frac{2d^*}{N}.$$

- For linear regression with Gaussian errors and fixed variance $d^* = d$ even when the model is wrong, but this does not hold in general, e.g. logistic regression.
- The estimator $\hat{\theta}_N$ must be the MLE. Extensions to non-MLE and non-likelihood loss setups are possible with d replaced again by a more complicated d^*.

AIC

- For model comparison there is theoretical evidence that

$$\mathrm{AIC}_1 - \mathrm{AIC}_2 = \frac{2}{N}(l_N^1(\hat{\theta}_N^1) - l_N^2(\hat{\theta}_N^2)) + \frac{2(d_1 - d_2)}{N}$$

  can be a (much) better approximation of the difference in $\mathrm{Err}_{\mathrm{loglik,in}}$ when the models are nested than if the models are non-nested.

- The *deviance* equals $2l_N(\hat{\theta}_N)$ up to an additive constant – often the value of twice the minus-log-likelihood in the "saturated model". For model comparisons we can replace $2l_N(\hat{\theta}_N)$ by the deviance, but make sure that all models considered use the same reference model/additive constant in their definition of the deviance.

Figure 7.4 – AIC Used for Model Selection

This figure provide some empirical justification of using AIC in a context where there is no theoretical justification. The 0-1 loss is not a minus-log-likelihood.

Practical BIC

With the same framework as for AIC

$$\mathrm{BIC} = 2l_N(\hat{\theta}_N) + d\log(N)$$

We choose among several models the one with the smallest BIC.

Up to the scaling by $1/N$, BIC is from a practical point of view AIC with 2 replaced by $\log(N)$. The theoretical derivation is, however, completely different.

For $N > e^2 \simeq 7.4$, BIC penalizes complex models more than simple models compared to AIC.

All the preceeding computations with AIC and BIC have been done in the framework of the *conditional* distribution of Y given X. This framework with the likelihood loss has the strongest resemblence to the pure prediction-loss statistical decision theoretic framework, though we have to allow for a more general "action space" to accomodate all situations of practical interest. We can also consider AIC and BIC in the framework of the joint distribution of (X, Y).

AIC in Unconditional Models

With $(P_\theta)_{\theta\in\Theta}$ a parameterized family of probability measures on $E \times F$, h_θ the joint density of (X, Y),

$$l_N(\theta) = -\sum_{i=1}^{N} \log h_\theta(x_i, y_i)$$

the joint minus-log-likelihood function, and $\hat{\theta}_N$ the MLE then

$$\text{AIC} = \frac{2}{N} l_N(\hat{\theta}_N) + \frac{2d^*}{N}.$$

Under suitable regularity conditions

$$E(\text{AIC}) = -E(\log h_{\hat{\theta}_N}(X,Y)) = E(\underbrace{-E(\log h_{\hat{\theta}_N}(X,Y)|\mathbf{X},\mathbf{Y})}_{\text{cross entropy}})$$

is the expected cross entropy of $P_{\hat{\theta}}$ from the true distribution of (X,Y) If the model is true then $d^* = d$ where d is the dimension of Θ.

The cross entropi of P_θ from the true distribution, lets call it Q and assume that it has density q, is

$$H(Q, P_\theta) = -\int \log h_\theta(x) q(x) \mathrm{d}x$$

is a reasonable measure of how well P_θ approximates Q. The minimal cross entropi of any measure from Q is $H(Q) = H(Q,Q)$, which is known as the entropy of Q. If $P_\theta \neq Q$ then $H(Q, P_\theta) > H(Q)$. The *Kullback-Leibler* divergence of P_θ from Q is

$$D(Q||P_\theta) = \int \log \frac{q(x)}{h_\theta(x)} q(x) \mathrm{d}x = H(Q, P_\theta) - H(Q).$$

This is a non-symmetric, non-metric "distance" measure of P_θ from Q.

For BIC there is nothing really changes either and

$$\text{BIC} = 2 l_N(\hat{\theta}_N) + d \log(N).$$

Other Ideas and Methods

An alternative to the *approximations of expectations* we can consider *upper bounds*.

Let $\eta \in [0,1]$ and $h(\mathcal{F})$ denote a number – a complexity measure – for the class $\mathcal{F}$ such that with probability at least $1 - \eta$

$$EPE(\hat{f}) \leq \bar{\text{err}} + g(h(\mathcal{F}), \bar{\text{err}})$$

The theory by Vapnik based on the Vapnik-Chervonenkis dimension (VC dimension) provides such upper bounds. General upper bounds are nice but almost always *extremely* pessimistic. The theoretical justification is extremely hard. But it does not rule out the practical use of the upper bounds for model selection ....