Chapter 9

Model Selection

Suppose we observe a realization of a random variable Y, with distribution defined by a parameter $\boldsymbol{\beta}$

$$\prod_{\mathbf{x}_i \in N_0} f(y_i; \mathbf{x}_i, \boldsymbol{\beta}) \equiv f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}) \tag{9.1}$$

where $\mathbf{y}$ is the observed response associated with the covariates $\mathbf{X}$ and $\boldsymbol{\beta} \in \mathbb{R}^P$ is a $P \times 1$ parameter vector.

We are interested in estimating $\boldsymbol{\beta}$. Suppose that before doing so, we need to choose from amongst P competing models, generated by simply restricting the general parameter space R^P in which $\boldsymbol{\beta}$ lies.

In terms of the parameters, we represent *the full model* with P parameters as:

$$\text{Model(P): } f_{\mathbf{Y}}(\mathbf{y}; \mathbf{x}, \boldsymbol{\beta}_P), \boldsymbol{\beta}_P = (\beta_1, \ldots, \beta_p, \beta_{p+1}, \ldots, \beta_P)'.$$

We denote the "true value" of the parameter vector $\boldsymbol{\beta}$ with $\boldsymbol{\beta}^*$.

Akaike (1977) formulates the problem of statistical model identification as one of selecting a model $f_{\mathbf{Y}}(\mathbf{y}; \mathbf{x}, \boldsymbol{\beta}_p)$ based on the observations from that distribution, where the particular restricted model is defined by the constraint $\beta_{p+1} = \beta_{p+2} =$

$\ldots = \beta_P = 0$, so that

$$\text{Model(p): } f_{\mathbf{Y}}(\mathbf{y}; \mathbf{x}, \boldsymbol{\beta}_p), \boldsymbol{\beta}_p = (\beta_1, \ldots, \beta_p, 0, \ldots, 0)' \tag{9.2}$$

We will refer to p as the *number of parameters* and to Ω_p as the sub-space of $\mathbb{R}^P$ defined by restriction (9.2). For each $p = 1, \ldots, P$, we may assume model(p) to estimate the non-zero components of the vector $\boldsymbol{\beta}^*$. We are interested in a criterion that helps us chose amongst these P competing estimates.

In this Chapter we consider 3 methods for model selection.

9.1 Mallow's C_p

Mallow's C_p is a technique for model selection in regression (Mallows 1973). The C_p statistic is defined as a criteria to assess fits when models with different numbers of parameters are being compared. It is given by

$$C_p = \frac{\text{RSS}(p)}{\sigma^2} - N + 2p \tag{9.3}$$

If model(p) is correct then C_p will tend to be close to or smaller than p. Therefore a simple plot of C_p versus p can be used to decide amongst models.

In the case of ordinary linear regression, Mallow's method is based on estimating the mean squared error (MSE) of the estimator $\hat{\boldsymbol{\beta}}_p = (\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p'\mathbf{Y}$,

$$E[\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}]^2$$

via a quantity based on the residual sum of squares (RSS)

$$\begin{aligned} \text{RSS}(p) &= \sum_{n=1}^{N} (y_n - \mathbf{x}_n \hat{\boldsymbol{\beta}}_p)^2 \\ &= (\mathbf{Y} - \mathbf{X}_p \hat{\boldsymbol{\beta}}_p)'(\mathbf{Y} - \mathbf{X}_p \hat{\boldsymbol{\beta}}_p) \\ &= \mathbf{Y}'(\mathbf{I}_N - \mathbf{X}_p(\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p')\mathbf{Y} \end{aligned}$$

Here $\mathbf{I}_N$ is an $N \times N$ identity matrix. By using a result for quadratic forms, presented for example as Theorem 1.17 in Seber's book, page 13, namely

$$\mathrm{E}[\mathbf{Y}'\mathbf{A}\mathbf{Y}] = \mathrm{E}[\mathbf{Y}']\mathbf{A}\mathrm{E}[\mathbf{Y}] + \mathrm{tr}[\mathbf{\Sigma}\mathbf{A}]$$

$\mathbf{\Sigma}$ being the variance matrix of $\mathbf{Y}$, we find that

$$\begin{aligned}
\mathrm{E}[\mathrm{RSS}(p)] &= \mathrm{E}[\mathbf{Y}'(\mathbf{I}_N - \mathbf{X}_p(\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p')\mathbf{Y}] \\
&= \mathrm{E}[\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}]^2 + \mathrm{tr}\left[\mathbf{I}_N - \mathbf{X}_p(\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p'\right]\sigma^2 \\
&= \mathrm{E}[\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}]^2 + \sigma^2\left(N - \mathrm{tr}\left[(\mathbf{X}_p'\mathbf{X}_p)(\mathbf{X}_p'\mathbf{X}_p)^{-1}\right]\right) \\
&= \mathrm{E}[\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}]^2 + \sigma^2(N-p)
\end{aligned}$$

where N is the number of observations and p is the number of parameters. Notice that when the true model has p parameters $\mathrm{E}[\mathrm{C}_p] = p$. This shows why, if model(p) is correct, C_p will tend to be close to p.

One problem with the C_p criterion is that we have to find an appropriate estimate of σ^2 to use for all values of p.

9.1.1 C_p for smoothers

A more direct way of constructing an estimate of PSE is to correct the ASR. It is easy to show that

$$\mathrm{E}\{\mathrm{ASR}(\lambda)\} = \left\{1 - n^{-1}\mathrm{tr}(2\mathbf{S}_\lambda - \mathbf{S}_\lambda\mathbf{S}_\lambda')\right\}\sigma^2 + n^{-1}\mathbf{v}_\lambda'\mathbf{v}_\lambda$$

notice that

$$\mathrm{PSE}(\lambda) - \mathrm{E}\{\mathrm{ASR}(\lambda)\} = n^{-1}2\mathrm{tr}(\mathbf{S}_\lambda)\sigma^2$$

This means that if we knew σ^2 we could find a "corrected" ASR

$$\mathrm{ASR}(\lambda) + 2\mathrm{tr}(\mathbf{S}_\lambda)\sigma^2$$

with the right expected value.

For linear regression $\text{tr}(\mathbf{S}_\lambda)$ is the number of parameters so we could think of $2\text{tr}(\mathbf{S}_\lambda)\sigma^2$ as a penalty for large number of parameters or for un-smooth estimates.

How do we obtain an estimate for σ^2? If we had a λ^* for which the bias is 0, then the usual unbiased estimate is

$$\frac{\sum_{i=1}^n \{y_i - f_{\lambda^*}(x_i)\}^2}{n - \text{tr}(2\mathbf{S}_{\lambda^*} - \mathbf{S}_{\lambda^*}\mathbf{S}'_{\lambda^*})}$$

The usual trick is to chose one a λ^* that does little smoothing and consider the above estimate. Another estimate that has been proposed it the first order difference estimate

$$\frac{1}{2(n-1)}\sum_{i=1}^{n-1}(y_{i+1} - y_i)^2$$

Once we have an estimate $\hat{\sigma}^2$ then we can define

$$C_p = \text{ASR}(\lambda) + n^{-1}2\text{tr}(S_\lambda)\hat{\sigma}^2$$

Notice that the p usually means number of parameters so it should be C_λ.

Notice this motivates a definition for degrees of freedoms.

9.2 Information Criteria

In this section we review the concepts behind Akaike's Information Criterion (AIC).

Akaike's original work is for IID data, however it is extended to a regression type setting in a straight forward way. Suppose that the conditional distribution of Y given $\mathbf{x}$ is know except for a P-dimensional parameter $\boldsymbol{\beta}$. In this case, the probability density function of $\mathbf{Y} = (Y_1, \ldots, Y_n)$ can be written as

$$f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}) \equiv \prod_{i=1}^n f(y_i; \mathbf{x}_i, \boldsymbol{\beta}) \tag{9.4}$$

with $\mathbf{X}$ the design matrix with rows $\mathbf{x}_i$.

Assume that there exists a true parameter vector $\boldsymbol{\beta}^*$ defining a true probability density denoted by $f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}^*)$. Given these assumptions, we wish to select $\boldsymbol{\beta}$, from one of the models defined as in (9.2), "nearest" to the true parameter $\boldsymbol{\beta}^*$ based on the observed data $\mathbf{y}$. The principle behind Akaike's criterion is to define "nearest" as the model that minimizes the Kullback-Leibler Information Quantity

$$\Delta(\boldsymbol{\beta}^*; \mathbf{X}, \boldsymbol{\beta}) = \int \{\log f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}^*) - \log f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta})\} f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}^*)\, d\mathbf{y}. \tag{9.5}$$

The analytical properties of the Kullback-Leibler Information Quantity are discussed in detail by Kullback (1959) . Two important properties for Akaike's criterion are

1. $\Delta(\boldsymbol{\beta}^*; \mathbf{X}, \boldsymbol{\beta}) > 0$ if $f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}^*) \neq f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta})$
2. $\Delta(\boldsymbol{\beta}^*; \mathbf{X}, \boldsymbol{\beta}) = 0$ if and only if $f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}^*) = f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta})$

almost everywhere on the range of $\mathbf{Y}$. The properties mentioned suggest that finding the model that minimizes the Kullback-Leibler Information Quantity is an appropriate way to choose the "nearest" model.

Since the first term on the right hand side of (9.5) is constant over all models we consider, we may instead maximize

$$\begin{aligned} H(\boldsymbol{\beta}) &= \int \log f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}) f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}^*)\, d\mathbf{y} \\ &= \sum_{i=1}^{n} \int \log f(y_i; \mathbf{X}, \boldsymbol{\beta})\, f(y_i; \mathbf{x}_i, \boldsymbol{\beta}^*)\, dy_i. \end{aligned} \tag{9.6}$$

Let $\hat{\boldsymbol{\beta}}_p$ be the maximum likelihood estimate under Model(p). Akaike's procedure for model selection is based on choosing the model which produces the estimate that maximizes $\mathrm{E}_{\boldsymbol{\beta}^*}\left[H(\hat{\boldsymbol{\beta}}_p)\right]$ amongst all competing models. Akaike then derives

a criterion by constructing an asymptotically unbiased estimate of $\mathrm{E}_{\boldsymbol{\beta}^*}\left[H(\hat{\boldsymbol{\beta}}_p)\right]$ based on the observed data.

Notice that $H(\hat{\boldsymbol{\beta}}_p)$ is a function, defined by (9.6), of the maximum likelihood estimate $\hat{\boldsymbol{\beta}}_p$, which is a random variable obtained from the observed data. A natural estimator of its expected value (under the true distribution of the data) is obtained by substituting the empirical distribution of the data into (9.6) resulting in the log likelihood equation evaluated at the maximum likelihood estimate under model(p)

$$l(\hat{\boldsymbol{\beta}}_p) = \sum_{i=1}^{n} \log f(y_i; \mathbf{x}_i, \hat{\boldsymbol{\beta}}_p).$$

Akaike noticed that in general $l(\hat{\boldsymbol{\beta}}_p)$ will overestimate $\mathrm{E}_{\boldsymbol{\beta}^*}\left[H(\hat{\boldsymbol{\beta}})\right]$. In particular Akaike found that under some regularity conditions

$$E_{\boldsymbol{\beta}^*}\left[l(\hat{\boldsymbol{\beta}}_p) - H(\hat{\boldsymbol{\beta}}_p)\right] \approx p.$$

This suggests that larger values of p will result in smaller values of $l(\hat{\boldsymbol{\beta}}_p)$, which may be incorrectly interpreted as a "better" fit, regardless of the true model. We need to "penalize" for larger values of p in order to obtain an unbiased estimate of the "closeness" of the model. This fact leads to the Akaike Information Criteria which is a bias-corrected estimate given by

$$\mathrm{AIC}(p) = -2l(\hat{\boldsymbol{\beta}}_p) + 2p. \tag{9.7}$$

See, for example, Akaike (1973) and Bozdogan (1987) for the details.

9.3 Posterior Probability Criteria

Objections have been raised that minimizing Akaike's criterion does not produce asymptotically consistent estimates of the correct model Notice that if we consider Model(p^*) as the correct model then we have for any $p > p*$

$$\Pr\left[AIC(p) < AIC(p^*)\right] = \Pr\left[2\{l(\hat{\boldsymbol{\beta}}_p) - l(\hat{\boldsymbol{\beta}}_{p^*})\} > 2(p - p^*)\right]. \tag{9.8}$$

Notice that, in this case, the random variable $2\{l(\hat{\boldsymbol{\beta}}_p) - l(\hat{\boldsymbol{\beta}}_{p^*})\}$ is the logarithm of the likelihood ratio of two competing models which, under certain regularity conditions, is known to converge in distribution to $\chi^2_{p-p^*}$, and thus it follows that the probability in Equation (9.8) is not 0 asymptotically. Some have suggested multiplying the penalty term in the AIC by some increasing function of n, say $a(n)$, that makes the probability

$$\Pr\left[2\{l(\hat{\boldsymbol{\beta}}_p) - l(\hat{\boldsymbol{\beta}}_{p^*})\} > 2a(n)(p - p^*)\right]$$

asymptotically equal to 0. There are many choices of $a(n)$ that would work in this context. However, some of the choices made in the literature seem arbitrary.

Schwarz (1978) and Kashyap (1982) suggest using a Bayesian approach to the problem of model selection which, in the IID case, results in a criterion that is similar to AIC in that it is based on a penalized log-likelihood function evaluated at the maximum likelihood estimate for the model in question. The penalty term in the Bayesian Information Criteria (BIC) obtained by Schwarz (1978) is the AIC penalty term p multiplied by the function $a(n) = \frac{1}{2}\log(N)$.

The Bayesian approach to model selection is based on maximizing the posterior probabilities of the alternative models, given the observations. To do this we must define a strictly positive prior probability $\pi_p = \Pr[\text{Model}(p)]$ for each model and a conditional prior $d\mu_p(\boldsymbol{\beta})$ for the parameter given it is in Ω_p, the subspace defined by Model(p). Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be the response variable and define the distribution given $\boldsymbol{\beta}$ following (9.4)

$$f_{\mathbf{Y}}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) \equiv \prod_{i=1}^{n} f(y_i; \mathbf{x}_i, \boldsymbol{\beta})$$

The posterior probability that we look to maximize is

$$\Pr\left[\text{Model}(p)|\mathbf{Y} = \mathbf{y}\right] = \frac{\int_{\Omega_p} \pi_p f_{\mathbf{Y}}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) d\mu_p(\boldsymbol{\beta})}{\sum_{q=1}^{P} \int_{\Omega_q} \pi_q f_{\mathbf{Y}}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) d\mu_q(\boldsymbol{\beta})}$$

Notice that the denominator depends neither on the model nor the data, so we need only to maximize the numerator when choosing models.

Schwarz (1978) and Kashyap (1982) suggest criteria derived by taking a Taylor expansion of the log posterior probabilities of the alternative models. Schwarz

(1978) presents the following approximation for the IID case

$$\log \int_{\Omega_p} \pi_p f_{\mathbf{Y}}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) d\mu_p(\boldsymbol{\beta}) \approx l(\hat{\boldsymbol{\beta}}_p) - \frac{1}{2} p \log n$$

with $\hat{\boldsymbol{\beta}}_p$ the maximum likelihood estimate obtained under Model(p).

This fact leads to the Bayesian Information Criteria (BIC) which is

$$\text{BIC}(p) = -2l(\hat{\boldsymbol{\beta}}_p) + p \log n \tag{9.9}$$

9.3.1 Kyphosis Example

The AIC and BIC obtained for the gam are:

```
AIC(Age) =  83              BIC(Age) = 90
AIC(Age,Start) = 64         BIC(Age,Start) = 78
AIC(Age,Number) = 73        BIC(Age,Number) = 86
AIC(Age,Start,Number) = 60  BIC(Age,Start,Number) = 81
```

Bibliography

[1] Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle," in Petrov, B. and Csaki, B., editors, *Second International Symposium on Information Theory*, pp. 267–281, Budapest: Academiai Kiado.

[2] Bozdogan, H. (1987), "Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions," *Psychometrica*, 52, 345–370.

[3] Bozdogan, H. (1994), "Mixture-model cluster analysis using a new informational complexity and model selection criteria," in Bozdogan, H., editor, *Multivariate Statistical Modeling*, volume 2, pp. 69–113, The Netherlands: Dordrecht.

[4] Kullback, S. (1959), *Information Theory and Statistics*, New York: John Wiley & Sons.

[5] Schwarz, G. (1978), "Estimating the dimension of a model," *Annals of Statistics*, 6, 461–464.

[6] Shibata, R. (1989), "Statistical aspects of model selection," in Williems, J. C., editor, *From Data to Model*, pp. 215–240, New York: Springer-Verlag.