Chapter 4

Model Selection

The aim of model selection is to find from a set of predictor variables those which are **relevant** for the response variable. Relevance in this context means that variability of the predictor is associated with variability of the response variable. Furthermore this co-existence of variability of predictors and response has to be quantifiable by a linear function, such as the one given in the model (4.1).

In a practical data analysis setting, the dataset used as input to the analysis may have many predictor variables. But it is not guaranteed that all of them have an influence on the response variable. Because we want to model the responses with a linear function of the predictor variables, every additional predictor variable introduces an additional coefficient that must be estimated. Every estimated coefficient leads to more variability in the predicted response values of a given model. Hence if a model should be used to predict new responses based on observed predictor values, the increased variability decreases the predictive power.

We assume the following linear model

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \quad (i = 1, \ldots, n) \tag{4.1}$$

where $\epsilon_1, \ldots, \epsilon_n$ are identically independently distributed (i.i.d) with $E(\epsilon_i) = 0$ and $var(\epsilon_i) = \sigma^2$. The model selection problem can be stated by the following question.

> "Which of the predictor variables should be used in the linear model?"

As already mentioned, it may be that not all of the p predictor variables included in the full model shown in (4.1) are relevant. Predictors that are not

relevant should not be included in a model because every coefficient of a predictor must be estimated and leads to increased variability of the fitted model. In case where this variability is caused by non-relevant predictor variables, the predictive power of the estimated model is lowered. As a consequence, we are often looking for an **optimal** or the **best** model given the available input dataset.

4.1 Bias-Variance Trade-Off

What was explained above can be formalized a bit more. Suppose, we are looking for optimizing the prediction

$$\sum_{r=1}^{q} \hat{\beta}_{j_r} x_{ij_r} \tag{4.2}$$

which includes q relevant predictor variables with indices taken from the vector j with $j_1, \ldots, j_q \in \{1, \ldots, p\}$. The average mean squared error of the prediction in (4.2) can be computed as

$$\begin{aligned} MSE &= n^{-1} \sum_{i=1}^{n} E\left[(m(x_i) - \sum_{r=1}^{q} \hat{\beta}_{j_r} x_{ij_r})^2 \right] \\ &= n^{-1} \sum_{i=1}^{n} \left(E\left[\sum_{r=1}^{q} \hat{\beta}_{j_r} x_{ij_r} \right] - m(x_i) \right)^2 + n^{-1} \sum_{i=1}^{n} var(\sum_{r=1}^{q} \hat{\beta}_{j_r} x_{ij_r}) \end{aligned} \tag{4.3}$$

where $m(.)$ denotes the linear function in the true model with p predictor variables. The systematic error $n^{-1} \sum_{i=1}^{n} \left(E\left[\sum_{r=1}^{q} \hat{\beta}_{j_r} x_{ij_r} \right] - m(x_i) \right)^2$ is called squared bias and this quantity is expected to decrease as the number of predictors q increases. But the variance term increases with the number of predictors q. This fact is called the **bias-variance trade-off** which is present in many applications in statistics. Now finding the best model corresponds to finding the model that optimizes the bias-variance trade-off. This process is also referred to as **regularization**.

4.2 Mallows C_p Statistic

The mean square error in (4.3) is unknown because we do not know the magnitude of the bias. But MSE can be estimated.

Let us denote by $SSE(\mathcal{M})$ the residual sum of squares in the model $\mathcal{M}$. Unfortunately $SSE(\mathcal{M})$ cannot be used to estimate MSE because $SSE(\mathcal{M})$ becomes

smaller the more predictors are included in the model $\mathcal{M}$. The number of predictors in the model $\mathcal{M}$ is also often referred to as the size of the model and is written as $|\mathcal{M}|$.

For any (sub-) model $\mathcal{M}$ which involves some (or all) of the predictor variables, the mean square error (MSE) can be estimated by

$$\widehat{MSE} = n^{-1}SSE(\mathcal{M}) - \hat{\sigma}^2 + 2\hat{\sigma}^2|\mathcal{M}|/n \tag{4.4}$$

where $\hat{\sigma}^2$ is the error variance estimate in the full model and $SSE(\mathcal{M})$ is the residual sum of squares in the sub-model $\mathcal{M}$. Hence to find the best model, we could search for the sub-model $\mathcal{M}$ that minimizes $\widehat{MSE}$. Because $\hat{\sigma}^2$ and n are constants with respect to sub-models $\mathcal{M}$, we can also consider the well-known C_p statistic

$$C_p(\mathcal{M}) = \frac{SSE(\mathcal{M})}{\hat{\sigma}^2} - n + 2|\mathcal{M}| \tag{4.5}$$

and search for the sub-model $\mathcal{M}$ minimizing the C_p statistic.

4.3 Searching For The Best Model With Respect To C_p

If the full model has p predictor variables, there are $2^p - 1$ sub-models (every predictor can be considered in a sub-model or not. The empty sub-model without any predictors is excluded here).

Therefore, an exhaustive search for the sub-model $\mathcal{M}$ minimizing C_p is only feasible if p is less than 16 which results in $2^{16} - 1 = 6.5535 \times 10^4$ sub-models to be tested. If p is much larger, we can proceed with one of the two following stepwise algorithms.

4.3.1 Forward Selection

1. Start with the smallest model $\mathcal{M}_0$ with only a general mean as the current model
2. Include the predictor variable to the current model which reduces the residual sum of squares the most.
3. Continue with step 2 until all predictor variables have been chosen or until a large number of predictor variables have been selected. This produces a sequence of sub-models $\mathcal{M}_0 \subseteq \mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \ldots$
4. Choose the model in the sequence $\mathcal{M}_0 \subseteq \mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \ldots$ with the smallest C_p value.

4.3.2 Backward Selection

1. Start with the full model $\mathcal{M}_0$ as the current model. The full model is the model including all p predictor variables
2. Exclude the predictor variable from the current model which increases the residual sum of squares the least.
3. Continue with step 2 until all predictor values have been deleted (or a large number of variables have been deleted). This produces a sequence of sub-models $\mathcal{M}_0 \supseteq \mathcal{M}_1 \supseteq \mathcal{M}_2 \supseteq \ldots$.
4. Choose the model in the sequence $\mathcal{M}_0 \supseteq \mathcal{M}_1 \supseteq \mathcal{M}_2 \supseteq \ldots$ which has the smallest C_p value.

4.3.3 Considerations

Backward selection (4.3.2) typically leads to better results than forward selection, but it is computationally more expensive. But in the case where $p \geq n$, the full model cannot be fitted and backward selection is not possible. Forward selection might then be a possibility, but alternative estimation procedures such as LASSO might be a better solution.

4.4 Alternative Model Selection Criteria

Other popular criteria to estimate the predictive potential of an estimated model are Akaike'w information criterion (AIC) and the Bayesian information criterion (BIC). Both of them are based on the likelihood and require therefore assumptions about the distribution of the data.

The goodness of the fit of the linear model for explaining the data is quantified by the coefficient of determination which is typically abbreviated by R^2 where

$$R^2 = \frac{||\hat{y} - \bar{y}||^2}{||y - \bar{y}||^2} \tag{4.6}$$

where $||\hat{y} - \bar{y}||^2$ are the sum of squares explained by the model and $||y - \bar{y}||^2$ stands for the total sum of squares around the global mean $\bar{y}$. The coefficient of determination R^2 is always increasing the more predictor variables are included in the model. This behavior can be corrected as proposed in [Yin and Fan, 2001]. This correction includes the number of predictor variables an hence reduces the favoring of the full model. The result of the correction is the adjusted R^2 which is computed as

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{4.7}$$

where R^2 is the unadjusted coefficient of determination given by (4.6), n stands for the number of observations and p is the number of predictor variables. The formula in (4.7) holds for sub-models that include an intercept term. For sub-models without intercept, the -1 in both numerator and the denominator of (4.7) can be dropped.

The adjusted coefficient of determination (R^2_{adj}) allows to assess the goodness of fit of a model. That assessment considers the number of predictor variables included in the model.