1 Model Selection

1.1 Variable Selection

To quote Weisberg (2005) in chapter 10, “we live in an era of cheap data, but expensive information”. In many instances, we are faced with a large array of variables, but it is unclear how one can summarize this data into useful information. The task of selecting an optimal subset of covariates amongst a large set of candidates can be subdivided into two distinct goals:

1. Firstly, we need to explore the power set of the set of covariates in a computationally efficient manner. This can be done using different stepwise procedures.

2. Secondly, we need to choose a criterion for weighting the comparative values of each subset of covariates. Such a criterion will generally be motivated by some information-theoretical argument, or related to an approximation of the expected prediction error.

Variable selection is a special case of model selection, which seeks a parsimonious representation of the data at hand. Model selection criteria seeks to counterbalance a measure of goodness-of-fit, with a penalty for model complexity. Recall that measures of goodness-of-fit, such as the coefficient of determination, $R^2$, are monotonic increasing with respect to the number of variables in the model, and therefore useless from a model selection perspective. In this lecture, we will be using $d$ instead of $p^*$ to qualify the number of parameters in the models.

1.2 Variance of the Fitted Values

Recall that the variance of the predicted values can be obtained as follows,

$$\text{Var}[\hat{y}|X] = \text{Var}[Hy|X] = H\text{Var}[y|X]H^T = \sigma^2 H.$$

The variance of the $i$th fitted value is therefore given by

$$\text{Var}[\hat{y}_i|X] = \sigma^2 x_i^T (X^T X)^{-1} x_i = x_i^T \text{Var}[\hat{\beta}|X] x_i,$$

since $\text{Var}[\hat{\beta}|X] = \sigma^2 (X^T X)^{-1}$, and where $x_i$ is the $i$th row of $X$, treated as a column vector. Therefore, this calculation shows that the variance of the fitted values is monotonically related with the variance of the estimators, $\hat{\beta}_j$'s.
2 Variable Selection

2.1 Optimal Subset Selection

We are interested in identifying an optimal subset of predictors from a possibly very large set of \( d \) predictors, where \( d < n \). Thus, the mean function is of the type,

\[
E[Y_i|X] = \sum_{j=1}^{d} x_{ij} \beta_j = x_i^T \beta, \quad \forall \ i = 1, \ldots, n;
\]

where \( x_i \) is the \( i \)th row of \( X \). The number of possible subsets of variable is the size of the power set of candidate covariates,

\[
S := \{1, \ldots, d\}.
\]

The power set of \( S \) is generally denoted \( \mathcal{P}(S) \) and its cardinality is \( 2^{|S|} \) or \( 2^d \), in our case, since \( d = |S| \). Thus, the problem of testing every possible subset of covariate is \textbf{NP-hard}, since this exploration grows exponentially fast with the size of the set of candidate covariates, \( S \).

There are two types of \textbf{stepwise variable selection}, which allow one to conduct a \textbf{quadratic} exploration of the space of all subsets. Both of these methods require \( d(d-1)/2 \) steps, and therefore are of order \( O(d^2) \). We will consider in turn (i) forward selection and (ii) backward elimination. Both of these methods follow an approximate trajectory through the set of subsets of \( S \). Note that this only provides us with an exploration of \( \mathcal{P}(S) \). A selection of the optimal subset amongst the subset of \( \mathcal{P}(S) \) that we have explored is then performed using an information criterion.

2.2 Forward Selection Procedure

This popular method begins with an \textit{empty set} of candidate covariates, and progressively adds new predictors, using a linear algebraic argument. At every stage of this procedure, we maximize the \textit{partial correlation} between the fitted values, \( \hat{y}_i \)'s, at that point and the projected candidates in the space spanned by the variables already included in the model. Formally, this can be described in the following manner. Assume that at step \( t \), we have already included \( t \) variables in the model, such that the fitted values at that point are given by

\[
\hat{y}_i(t) = \sum_{j \in S(t)} x_{ij} \hat{\beta}_j, \quad \forall \ i = 1, \ldots, n.
\]

where the index set \( S(t) \) is the set of candidate covariates at time \( t \). The set of the remaining covariates at time \( t \), will be denoted by \( S^C(t) = S \setminus S(t) \).

The next covariate can then be selected by maximizing the absolute value of the \textit{partial correlation} between the residuals at time \( t \),

\[
\hat{e}_i(t) = y_i - \hat{y}_i(t);
\]

and the residuals of any other covariate in the set \( S^C(t) \), regressed on the covariates already included in the model, \( S(t) \). That is, for any \( k \in S^C(t) \),

\[
\hat{e}_{ik}(t) = x_{ik} - \hat{x}_{ik}(t),
\]

with

\[
\hat{x}_{ik}(t) = \sum_{j \in S(t)} x_{ij} \hat{\beta}_{jk}, \quad \forall \ i = 1, \ldots, n;
\]

and where we have emphasized the dependence of \( \hat{\beta}_{jk} \) on \( k \), since these coefficients differ from the ones computed in equation (1). We then choose the \( j' \) that maximizes the absolute value of the partial correlation.
between the \( \hat{e}_i(t) \)'s and the \( \hat{e}_{ij}(t) \)'s. These stepwise procedures formalize a law of diminishing returns, when considering variable selection.

### 2.3 Backward Elimination Procedure

In this case, we start with a full model, and progressively remove covariates. At each stage, we are comparing two models that only differ by a single covariate from the set of remaining candidate covariates. Therefore, we can use the fact that an \( F \)-test on two models that solely differ by a single covariate is equivalent to the square of a \( t \)-test.

### 2.4 Hybrid Selection Method

In R, the `step` function also implements a hybrid procedure for selecting the best subset of variables at every stage of the algorithm. In this case, the forward and backward methods are compared using an information criterion called the AIC, which is described below. Finally, note that when one has finally obtained a subset of variables, one should not use the standard errors of the resulting \( \hat{\beta}_j \)'s, as these do no reflect the variable selection procedure.

## 3 Information Criteria

All of these methods are trying to strike a trade-off between goodness-of-fit and complexity. There is a vast literature on the subject. We are here only scratching the surface. These information criteria are defined with respect to a space model. Each model is a probability distribution function associated with a parameter space, \( \Theta \), such that the \( j \)th candidate model, for instance, may be denoted as follows,

\[
\mathcal{M}_j := \{ p(y, \theta) : \theta \in \Theta^d \},
\]

and the space of all candidate models will simply be denoted by \( \mathcal{M} \). In the case of multiple linear regression, each model of interest will be a normal distribution, indexed by some vector of slope parameters, and an estimate of the variance, such that

\[
\{ N(y; X\beta, \sigma^2) : (\beta, \sigma^2) \in \Theta^{d+1} \},
\]

for a set of \( n \) data points. In general, we will assume that there exists a true model, \( p_0(y) \), from which a vector of realizations, \( y_i \)'s, has been drawn. One may think of this true model as the population distribution function.

### 3.1 Akaike Information Criterion (AIC)

This criterion for model selection is justified using the Kullback-Leibler (KL) divergence. This quantity is a distance function (although this is not strictly speaking a metric, since it is not symmetric). Given the true model \( p_0 \) and any candidate model \( p \in \mathcal{M} \), the KL divergence between these two models is given by

\[
\text{KL}(p_0, p) = \int_{\mathbb{R}^n} p_0(y) \log \left( \frac{p_0(y)}{p(y; \hat{\theta})} \right) dy,
\]

where \( \hat{\theta} \) represents the MLE of \( p(y; \theta) \) for some vector of realizations \( y \). Hence, the model has already been fitted, producing the MLE, \( \hat{\theta}(y) \). Thereafter, we are computing the AIC for model selection purposes. Observe that the KL divergence is not a metric, since it is not symmetric, and it does not satisfy the
triangle inequality. However, the KL divergence is a premetric, since it is non-negative for every pair of distributions, and it also satisfies the coincidence axiom. That is, if \( p_1 = p_2 \), then \( \text{KL}(p_1, p_2) = 0 \).

The above expression can be straightforwardly simplified, since one can observe that

\[
\text{KL}(p_0, p) = \int_{\mathbb{R}^n} p_0(y) \log p_0(y) dy - \int_{\mathbb{R}} p_0(y) \log p(y; \hat{\theta}) dy.
\]

Moreover, the first term in the latter equation does not depend on \( p(y; \hat{\theta}) \). Therefore, the task of finding the model, which is the closest to the true model \( p_0(y) \) in terms of KL divergence, reduces to a maximization of the following quantity

\[
\bar{K}(p) := \int_{\mathbb{R}} p_0(y) \log p(y; \hat{\theta}) dy.
\]

This expression can be estimated using the empirical distribution function (EDF), since we know that the EDF is a good approximation of the true population CDF. Thus, we define our estimated model as

\[
\bar{K}(p) \doteq \frac{1}{n} \sum_{i=1}^{n} \log p(y_i; \hat{\theta}(y)) = \frac{1}{n} \mathcal{L}(\hat{\theta}; y),
\]

where \( \mathcal{L}(\hat{\theta}; y) \) denotes the loglikelihood estimated at its maximum. Crucially, however, observe that we have used the data twice. Firstly, we have estimated the MLE, \( \hat{\theta}(y) \). Secondly, we have used \( y \) again for obtaining an empirical estimate of the true distribution \( p_0 \). The resulting estimator \( \hat{d} \) of the true model is therefore biased. Akaike showed that, under some regulatory conditions, this bias can be approximated as \( p/n \), where \( p := \text{dim}(\Theta) \), such that

\[
\bar{K}(p) := \frac{1}{n} \mathcal{L}(\hat{\theta}; y) - \frac{d}{n}.
\]

Therefore, by convention, we normalize the above expression by \(-2n\), which gives the following for any \( f \in \mathcal{M} \),

\[
\text{AIC}(p) := -2n\bar{K}(p) = -2\mathcal{L}(\hat{\theta}; y) + 2d.
\]

As a result, we obtain a criterion that strikes a balance between the likelihood of the data and the complexity of the model, as measured by the number of effective parameters in your model. Thus, we wish to find the model \( f \in \mathcal{M} \) that minimizes \( \text{AIC}(p) \).

For the case of multiple regression with normal errors, we obtain the following simplification of the AIC,

\[
\text{AIC} = n \log \left( \frac{\text{RSS}(\hat{\beta})}{n} \right) + 2d.
\]

This last derivation is left as an exercise.

### 3.2 Bayes Information Criterion (BIC)

This criterion is motivated by a Bayesian perspective to statistical inference. In this framework, we specify a prior probability distribution on the space of possible parameter values, \( p(\theta) \). The likelihood is denoted by the conditional distribution, \( p(y|\theta) \). Altogether, Bayes’ theorem states that

\[
p(\theta|y) = \frac{p(y|\theta) \times p(\theta)}{p(y)},
\]
where each component is given the following name,

\[
\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}.
\]

The main quantity of interest to quantify the usefulness of the model is the **evidence**, which is computed as follows using standard probability theory,

\[
p(y) = \int_{\Theta} p(y|\theta)p(\theta)d\theta.
\]

The evidence is often referred to as the evidence of the model, given the data. Using a decision-theoretic framework, we are indeed implicitly assuming that \(p(y)\) is conditional on a particular choice of model, say \(M_j\), such that \(p(y|M_j)\). The evidence can be interpreted as the amount of support given to that model by the data, since integration over the parameter space, \(\Theta\), automatically applies an **Occam's razor**, leading to the selection of the most parsimonious model.

The evidence, \(p(y)\), can be approximated using **Laplace's method**. This relies on a Taylor expansion, which is conducted within the integral, thereby producing a Gaussian integral, which can be easily integrated. Altogether, this approximation reduces to the following information criterion,

\[
\text{BIC} = -2 \log p(y|\hat{\theta}) + d \log(n).
\]

where \(y\) is a \(n\)-dimensional vector of realizations. For a multiple linear regression, this criterion reduces to

\[
\text{BIC} = n \log \left( \frac{\text{RSS}(\hat{\beta})}{n} \right) + d \log(n).
\]

As for the AIC, we then choose the model that attains the smallest value amongst the set of candidate models.

### 3.3 Cross-Validation

A method for model selection that has become extremely popular in the machine learning literature is **cross-validation**. This is a general framework for estimating and comparing models. For instance, various high-level or nuisance parameters can be estimated using a **leave-one out** cross-validation. In this case, we would be fitting the model, estimating the parameters of interest based on the reduced data set, and then calculating the classification rate of the model, or some other criteria.

The PRESS method used in R is somewhat similar to leave-one out classification, but here, we attempt to estimate the \(i\)th data point using \(\hat{\beta}_{(i)}\) —that is, using the vector of OLS parameters, which were based on the reduced data set. This gives

\[
\text{PRESS} = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \hat{\beta}_{(i)})^2 = \sum_{i=1}^{n} \left( \frac{\hat{e}_i}{1 - h_{ii}} \right)^2.
\]

The AIC and cross-validation tend to display very similar behavior on a range of different data sets.

### 3.4 Conclusion

Altogether, the three criteria used in R, when selecting a subset of variables are AIC, BIC and Mallows’ \(C_d\),

\[
\text{AIC} = n \log \left( \frac{\text{RSS}(\hat{\beta})}{n} \right) + 2d,
\]

\[
\text{BIC} = n \log \left( \frac{\text{RSS}(\hat{\beta})}{n} \right) + d \log(n),
\]

\[
C_d = \frac{\text{RSS}(\hat{\beta})}{\hat{\sigma}^2} + 2d - n.
\]