# Chapter 4

# Least Absolute Shrinkage And Selection Operator (LASSO)

The linear model

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \tag{4.1}$$

for an observation $i$ (with $i = 1, \dots, n$) is used to model the relationship between the predictor variables $x_{i1}, \dots, x_{ip}$ and a response variable $y_i$. In a fixed linear effect model the unknown parameters $\beta_j$ (with $j = 1, \dots, p$) are estimated with least squares.

The values for $\beta_j$ and for the error terms $\epsilon_i$ are unknown. We expect the predictor variables $x_{i1}, \dots, x_{ip}$ to be known without any error. For a given dataset with $n$ observations the model can be written in matrix-vector notation as

$$y = X\beta + \epsilon \tag{4.2}$$

## 4.1 Stochastic Error Component

The error terms $\epsilon$ are random effects in the model (4.2). The expected value $E(\epsilon) = 0$ and the variance $var(\epsilon) = I * \sigma^2$. This means the error terms of the different observations are assumed to be uncorrelated. The single error terms $\epsilon_i$

are not so interesting, but the variance component $\sigma^2$ is besides the vector $\beta$ of coefficients an additional unknown parameter.

## 4.2 Parameter Estimation

For the fixed linear effects model we have seen that the parameters $\beta$ can be estimated using **Least Squares**. The condition implied by least squares corresponds to

$$\hat{\beta}_{LS} = argmin_{\beta}||y - X\beta||^2 \tag{4.3}$$

where $||.||$ corresponds to the euclidean norm. The solution of (4.3) for $\hat{\beta}_{LS}$ leads to the least squares normal equations given by

$$(X^T X)\hat{\beta}_{LS} = X^T y \tag{4.4}$$

If the matrix $X$ has full column rank, the inverse of $(X^T X)$ exists and we can write the least squares estimator $\hat{\beta}_{LS}$ as

$$\hat{\beta}_{LS} = (X^T X)^{-1} X^T y \tag{4.5}$$

In all cases where $X$ does not have full column rank, $(X^T X)$ is singular and one solution to (4.4) can be written in terms of a generalized inverse $(X^T X)^-$ of $(X^T X)$. This solution is called $b_0$ and can be written as

$$b_0 = (X^T X)^- X^T y \tag{4.6}$$

The solution in (4.6) are called $b_0$ and not $\hat{\beta}$ because $b_0$ is not an estimate of $\beta$. Furthermore, $b_0$ is not unique, because $(X^T X)^-$ is not unique. The solution $b_0$ can be used to generate estimates of estimable functions of $\beta$. An estimable function is a linear function of the parameters for which an estimator can be found from $b_0$ that is invariant to whatever solution is used for the normal equations. Any linear function of the parameters $\beta$ is defined as estimable, if it is identically equal to a linear function of the expected value of the observations $(y)$. This means that the linear function $q^T \cdot \beta$ is estimable if $q^T \cdot \beta = t^T \cdot E(y)$ for some vector $t$. For more details on estimable functions of parameters, we refer to section 4 of chapter 5 in [Searle, 1971].

Although, least squares provides us a tool to get estimates of either the unknown parameters or of estimable functions of the parameters, we are still not able to determine the important predictor variables out of a large set of available variables that characterize our responses in a given dataset. In statistics terms this problem is referred to as model selection or variable selection. In what follows, a method is described that solves the problem of variable selection.

## 4.3 Alternatives To Least Squares

The fixed linear effect model ((4.1)) is a very useful tool. Least Squares provides a very well established and an efficient method to estimate the unknown parameters. In the recent past, with the advent of a phenomenon called `Big Data` which stands for recent tendencies of systematically collecting large amounts of data, we have datasets available where each response has a very large number of potentially meaningful predictor variables. Finding the relevant predictors for a given response has become an important problem. Possible solutions can be divided into the following three classes of methods.

1. **Subset Selection**: Out of a set of $p$ predictor variables, a subset of "relevant" variables are selected. All other variables are ignored. The relevant variables are often selected based on the significance of the hypothesis test against the Null-Hypothesis ($H_0$) of a given model coefficient $\beta_j$ being 0 which means $H_0 : \beta_j = 0$.
2. **Regularisation (Shrinkage)**: All $p$ parameters are used in the model. The estimated coefficients are "forced" towards the origin. This process is called **shrinkage**. This causes a reduction of the variability of the estimates which is called *regularisation*.
3. **Dimension Reduction**: The $p$ predictors are reduced to $m$ linear combinations of the predictors. This reduction is achieved with techniques such as principal components analysis (PCA) or factor analysis (FA).

## 4.4 LASSO

Some procedures to estimate parameters can be found in more than one of the above three classes. Such methods are very popular, because they combine multiple of the above described properties which are desirable. An example of such a procedure is LASSO. LASSO is an abbreviation for Least Absolute Shrinkage and Selection Operator. It combines subset selection and regularization. The regularization is achieved by adding a penalty term to the least squares condition given in (4.3).

### 4.4.1 Regularisation With LASSO

With LASSO a penalty term is added to the least squares condition. The penalty term corresponds to $\lambda \sum_{j=1}^{p} |\beta_j|$. This term is a penalty, because when the sum of the absolute values of all $\beta_j$ parameters is larger, the contribution due to that term is also larger and since the least squares criterion must be minimized, larger values are acting like a penalty. As a consequence of that parameter values with smaller absolute values are preferred and this leads to the desired

effect of regularization. The resulting LASSO criterion can be expressed by the following formula.

$$\begin{aligned} \hat{\beta}_{LASSO} &= argmin_{\beta} \left\{ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \\ &= argmin_{\beta} \left\{ RSS + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \end{aligned} \tag{4.7}$$

### 4.4.2 Subset Selection With LASSO

The penalty term $\lambda \sum_{j=1}^{p} |\beta_j|$ in (4.7) is also responsible for the effect of subset selection. Due to the absolute value operator in the penalty term, some of the coefficients $\beta_j$ in the linear model are explicitly set to zero. Why this effect occurs is shown in Figure 4.1.

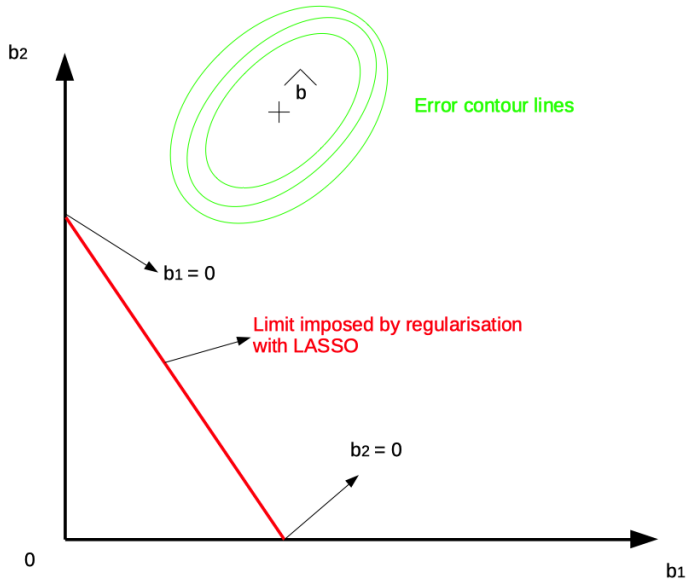


Figure 4.1: Subset Selection With LASSO

In Figure 4.1 a simple case with $p = 2$ parameters. The coefficients are called $b$ and not $\beta$. With an infinite amount of data the coefficients $b_j$ would be estimated with minimal error at the point which is labelled $\hat{b}$ in Figure 4.1. The green ellipses denote the contour lines that have a constant value for the error.

The red line symbolizes the limit which is created by the regularization effect imposed by LASSO. The regularization forces the parameter estimates to be inside of the triangle defined by the red line and the coordinate axes. Further more we want parameter estimates with minimal error. Hence the best estimate is at the intersection of a green ellipsis with the red line. This intersection is very likely to happen at one of the corners of the regularization triangle. At these corners, one of the coefficients is set to zero which is the source of the desirable property of subset selection.

## 4.5 Determine $\lambda$

The penalty term in (4.7) contains the parameter $\lambda$. This parameter is used to determine the strength of the regularization and it has to be estimated from the data. One possibility to determine $\lambda$ is via a procedure called **cross validation** (CV). In a cross validation the dataset is divided randomly into a training set and a test set. The complete data set is separated such that the test set is smaller than the training set. For a given training set with an assumed value of $\lambda$ the other parameters are estimated. With the estimated parameters, the data in the test set are tried to be predicted. This is repeated for different values of $\lambda$ and different separations into training and test sets. The value of $\lambda$ with minimal prediction error is selected as optimal estimate for $\lambda$.