**Chapter 6: Linear Model Selection and Regularization**

* *Feature selection*  or *variable selection* – that is, exlcluding irrelevant variables from a multiple regression model

*Subset selection*

* To perform a *best subset selection*, we fit a separate least squares regression for every possible combination of the *p* predictors. We then look at the resulting models, with the goal of identifying the one that is best. (Note: the problem of selecting the best model is not trivial…)
* One common way:
  + Let Mo denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation;
  + For *k* = 1,2…*p,* fit all models that contain exactly *k* predictors;
  + Pick the best amongst these models, and call it Mk. Here, *best* is defined as having the smallest RSS, or equivilantly largest r2
  + Select a single best model from amongst Mo to Mp, using cross validated prediction error, CP, BIC, or R2
* In the case of logistic regression, instead of R2 we use *deviance,* a measure that plays the role of RSS for a broader class of models
* Best subset selection becomes computationally unfeasable (due to the large numbers of predictors), even with extremely fast modern computers
* Best subset selection may also suffer from statistical problems when p is large. The larger the search space, the higher the chance of finding models that look good on the training data, even though they may not have any predictive power on future data.

*Forward stepwise selection*

* Forward stepwise selection begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all the predictors are in the model.
* Forward stepwise selection can be applied even in the high-dimensional setting where *n* < *p*

*Backward stepwise selection*

* Unlike forward stepwise selection, it begins with the full least squares model containing all *p* predictors, and then iteratively removes the least useful predictor, one-at-a-time.

*Hybrid stepwise selection*

* Hybrid versions of stepwise selection are available, in which variables are added to the model separately. However, after adding each new variable, the method may also remove any variables that no longer provide an improvement in the model fit.

*Choosing the optimal model*

* The training error can be a poor estimate of the test error. Therefore, RSS and R2 are not suitable for selecting the best model among a collection of models with different numbers of predictors
* In order to select the best model we need to estimate this test error. There are two common approaches:
  + We can indirectly estimate test error by making an *adjustment* to the training error to account for the bias due to overfitting;
  + We can *directly* estimate the test error, using either a validation set approach or a cross-validation approach

*Approach 1: Adjusting the training error*

* Here we consider four major approaches: CP, Akaine information criterion (AIC), Baysean information criterion (BIC), and adjusted R2.
* Essentially, the CP statistic adds a penalty of 2dpo to the training RSS in order to adjust for the fact that the training error tends to underestimate the test error.
* The CP statistic tends to take ona small value for models with a low test error, so when determining which set of models is best, we choose the model with the lowest CP value.
* The AIC criterion is defined for a large class of models ft by maximum likelihood.
* BIC is derived from a Bayesian point of view, but ends up looking similar to Cp (and AIC) as well.
* As R2 always increases as more variables are added in, the adjusted R2 is used for models with many predictors
* Unlike the R2, however, the adjusted R2 *pays a price* for the inclusion of unnecessary variables in the model
* Despite its popularity, and even though it is quite intuitive, the adjusted R2 is not was well motivated in statistical theory as AIC, BIC, and Cp.

*Validation and cross-validation*

* As an alternative to the approaches just discussed, we can directly estimate the test error using the validation set and cross-validation methods discussed in Chapter 5.
* This procedure has an advantage releative to AIC, BIC, CP and adjusted R2, in that it provides a true estimate of the test error, and makes fewer assumptions about the true underlying model
* One approach is to calculate the standard error of the estimated test MSE for each model size, and then select the smallest model for which the estimated test error is within one stnadrd error of the lowest point on the curve.

*Part Two: Shrinkage methods*

* As an alternative to some of the subset methods, we can fit a model containing all *p* predictors using a technique that constrains or regularises the coefficeitn estiamtes, or equivilantly, that *shrinks* the coefficient estimates towards zero. The two best-known techniques are ridge regression and the lasso.

*Ridge regression*

* As with least squares, ridge regression seeks coefficient estimates that fit the data well, by making the RSS small when. However, the second term in the equation, called a *shrinkage penalty*, is small when B1…BP are close to zero, and so it hasthe effect of *shrinking* the estimates of Bj towards zero. The tuning parameter serves to control the relative impact of these two terms on the regression coefficient estimates.
* As the tuning parameter increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias.
* It is best to apply ridge regression after standardising the predictors
* As the tuning parameter increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias.
* Ridge regression’s advantage over least squares is rooted in the bias-variance trade-off. A the tuning parameter increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias.
* Ridge regression works best in situations where the least squared estimates have high variance.

*The lasso*

* Ridge regression does have one obvious disadvantage. Unlike best subset, forward stepwise, and backward stepwise selection, which will generally select models that involve just a subset of the variables, ridge regression will include all *p* predictors in the final model. The penalty will shrink all of the coefficients towards zero, but it will not set any of them exactly to zero.
* This may not be a problem for prediction accuracy, but it can create a challenge in model interpretation in settings in which the number of varibales *p* is quite large.
* As with ridge regression, the lasso srhinks the coefficient estimates towards zero. However, in the case of the lasso, the l1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter is sufficiently large.
* In general, one might expect the lasoo to perform better in a setting where a relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or that equal zero. Ridge regression will perofrm better when the response is a function of many predictors, all with coefficients of roughly equal size.
* A technique such as cross-validation can be used to determine which approach is better on a particular data set.
* In ridge regression, each least squares coefficient estimate is shrunken by the same proportion. In constrast, the lasso shrinks each least squares coefficient towards zero by a constant amount. The type of shrinkage perofmred byt eh lasso in this simple setting is known as softthresholding.

*Selecting the tuning paramatr*

* Just as the subset selection approaches considered in Section 6.1 require a method to determine which of the models under construction is best, implementing ridge regression and the lasso requires a method for selecting a value for the tuning parameter.
* Cross validation provides a simple way to tackle this problem. We choose a grid of tuning parameter values, and compute the cross-validation error for each value of the tuning parameter, as descibred in Chapter 5. We then select the tuning parameter value for which the cross-validation error is smallest. Finally, the model is re-fit using all of the available observations and the selected value of the tuning parameter.

*Part three: Dimension reduction methods*

* All of the methods discussed so far are defined using the original predictors. We now explore a class of approaches that *transform* the predictors and then fit a least square model using the transformed variables. We will refer to these techniques as *dimension reduction* methods.
* All dimension methods work in two steps. First, the transformed predictors are obtained. Second, the model is fit using these *M* predictors.

*Principal component regression*

* PCA is a popular approach for deriving a low-dimension set of features from a large set of features from a large set of variables.
* PCA is a technique for for reducingthe dimension of a n x p data matrix X. The *first principal component* direction of the data is that along which the observations *vary the most*.
* The principal components regression (PCR) approach involves constructing the first *M* principal components, and then using these components as the least squares predictors in a linear regression model that is fit using least squares.
* In other words, we assume that the directions in which x1….xp show the most variation are the directions that are associated with Y.
* As more principal components are used in the regression model, the bias decreases, but the variance increases.
* PCR will tend to do well in cases where the first few principal components are sufficient to capture most of the variation in the predictors as well as the relationship with the response.
* In PCR, the number of principal components, *M*, is typically chosen by cross-validaiton.
* When performing PCR, we generally recommend *standardising* each predictor. In the absece of standardisation, the high variance variables will tend to play a larger role in the principal components obtained, and the scale on which the variabes are measured will ultimately have an effect on the final PCR model.

*Partial least squares*

* The PCR approach involves identifying linear combinations, or *directions*, that best represent the predicors X1 to Xp. These directions are identified in an *unsupervised* way, since the response *Y* is not used to help determine the principal component directions. That is, the response does not *supervise* the identification of the principal components. Consequently, PCR suffers from a drawback: there is no guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response.
* PLS is a dimension reduction method, which first identifies a new set of features Z1….Zm, that are linear combinations of the original features, and then fits a linear model via least squares using these *M* new features.
* Unlikes PCR, PLS identifies these new features in a supervised way – that is, it makes use of the response *Y* in order to identify new features that no only approximate the old features well, but also that *are related to the response*.
* After standarising the *p* predictors, PLS computes the first direction Z1 by setting each o1 equal to the coefficient from the simple linear regression of Y onto Xj. Hence, PLS places the highest weight on variables that are most strongly related to the response.
* To identify the second PLS direction, we first *adjust* each of the variables for Z1, by regressing each variable on Z1 and taking residuals. These residuals can be interpreted as the remaining information that has not been explained by the first PLS direction. We then compute Z2 using this *orthoganlized* data in exactly the same fashion as Z1 was computed based on the original data.

*Part Four: Considerations in High Dimensions*

* Datasets containing more features than observations are often referred to as *high-dimensional*. Classical approaches such as linear regression are not appropriate in this setting.
* Unfortunately, the Cp, AIC and BIC approaches are not appropriate in the high-dimensional setting, because estimating o2 is problematic
* *The curse of dimensionality*: In general, adding additional signal features that are truly associated with the response will improve the fitted model. However, adding noise features that are not truly associated with the response will lead to a deterioration in the fitted model, and consequently an increased test set error.
* In the high-dimensional setting, the multicollinearity problem is extreme: any variable in the model can be written as a linear combination of all of the other variables in the model.
* One should *never* use sum of squared errors, p-values, R2 statistics, or other traditional measures of model fit on the training data as evidence of a good model fit in the high-dimensional setting.