6

Linear Model Selection

and Regularization

The linear model has distinct advantages in terms of inference and, on real-world problems, is often surprisingly competitive in relation to non-linear methods. Hence, before moving to the non-linear world, we discuss in this chapter some ways in which the simple linear model can be improved, by replacing plain least squares fitting with some alternative fitting procedures.

***Why might we want to use another fitting procedure instead of least squares?*** As we will see, alternative fitting procedures can yield better prediction accuracy and model interpretability .

Problem with least square:

* Provided that the true relationship between the response and the predictors is approximately linear, the least squares estimates will have low bias.
* If n >> p —that is, if n , the number of observations, is much larger than p , the number of variables—then the least squares estimates tend to also have low variance, and hence will perform well on test observations.
* However, if n is not much larger than p , then there can be a lot of variability in the least squares fit, resulting in overfitting and consequently poor predictions on future observations not used in model training.
* if p > n , then there is no longer a unique least squares coefficient estimate: the variance is infinite so the method cannot be used at all.

By constraining or shrinking the estimated coefficients, we can often substantially reduce the variance at the cost of a negligible increase in bias. This can lead to substantial improvements in the accuracy with which we can predict the response for observations not used in model training.

It is often the case that some or many of the variables used in a multiple regression model are in fact not associated with the response. Including such irrelevant variables leads to unnecessary complexity in the resulting model. By removing these variables—that is, by setting the corresponding coefficient estimates to zero—we can obtain a model that is more easily interpreted. Now ***least squares is extremely unlikely to yield any coefficient estimates that are exactly zero***. In this chapter, we see some approaches for automatically performing feature selection or variable selection —that is, for excluding irrelevant variables from a multiple regression model.

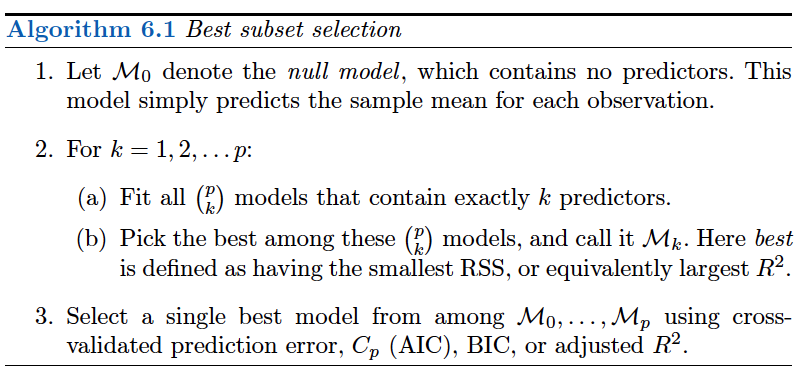
***There are many alternatives, both classical and modern, to using least squares to fit:***

* ***Subset Selection.*** This approach involves identifying a subset of the p predictors that we believe to be related to the response. We then fit a model using least squares on the reduced set of variables.
* ***Shrinkage:*** This approach involves fitting a model involving all p predictors. However, the estimated coefficients are shrunken towards zero relative to the least squares estimates. This shrinkage (also known as ***regularization***) has the effect of reducing variance. Depending on what type of shrinkage is performed, some of the coefficients may be estimated to be exactly zero. Hence, shrinkage methods can also perform variable selection.
* ***Dimension Reduction:*** This approach involves projecting the p predictors into a M-dimensional subspace, where M <p. This is achieved by computing M different linear combinations, or projections, of the variables. Then these M projections are used as predictors to fit a linear regression model by least squares.

6.1 Subset Selection

6.1.1 Best Subset Selection

we fit a separate least squares regression for each possible combination of the p predictors. That is, we fit all p models that contain exactly one predictor, all p (p− 1)/ 2 models that contain exactly two predictors, and so forth. We then look at all of the resulting models, with the goal of identifying the one that is best . The problem of selecting the best model from among the 2p possibilities considered by best subset selection is not trivial. This is usually broken up into two stages, as described in Algorithm 6.1.



Step 2 identifies the best model (on the training data) for each subset size, in order to reduce the problem from one of 2p possible models to one of p + 1 possible models.

Step 3: Now in order to select a single best model, we must simply choose among these p + 1 options. This task must be performed with care, because the RSS of these p + 1 models decreases monotonically, and the R2 increases monotonically, as the number of features included in the models increases. Therefore, if we use these statistics to select the best model, then we will always end up with a model involving all of the variables. The problem is that a low RSS or a high R2 indicates a model with a low training error, whereas we wish to choose a model that has a low test error.

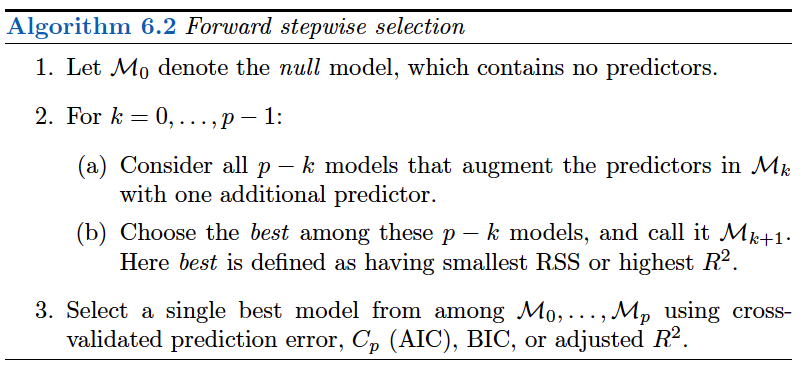
Therefore, in Step 3, we use cross-validated prediction error, Cp , BIC, or adjusted R2 in order to select among M0,M1, . . . ,Mp . These approaches are discussed in Section 6.1.3.

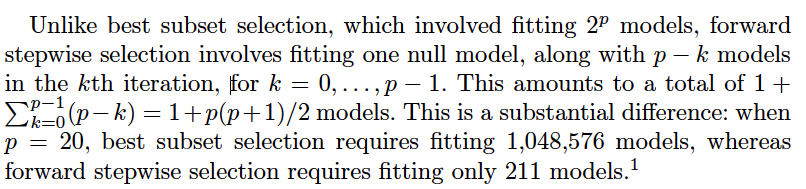
Problem with Best Subset Selection method:

best subset selection becomes computationally infeasible for values of p greater than around 40, even with extremely fast modern computers. There are computational shortcuts—so called branch-and-bound techniques—for eliminating some choices, but these have their limitations as p gets large. They also only work for least squares linear regression. We present computationally efficient alternatives to best subset selection next.

6.1.2 Stepwise Selection

Forward Stepwise Selection

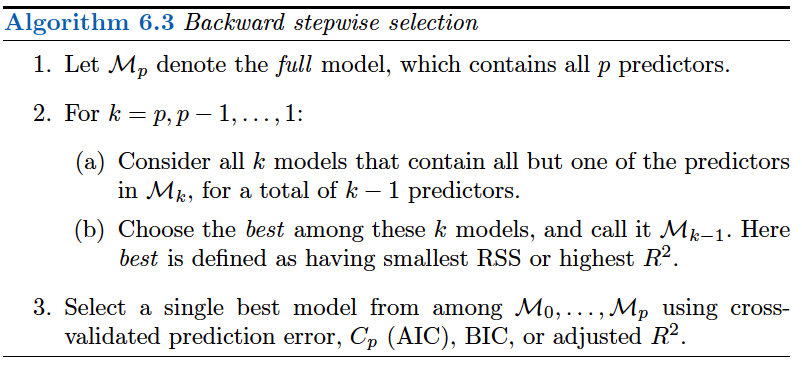




***Problem with F/w stepwise selection :*** Though forward stepwise tends to do well in practice, it is not guaranteed to find the best possible model out of all 2p models

containing subsets of the p predictors. For instance, suppose that in a given data set with p = 3 predictors, the best possible one-variable model contains X1 , and the best possible two-variable model instead contains X2 and X3 . Then forward stepwise selection will fail to select the best possible two-variable model, becauseM1 will contain X1 , soM2 must also contain X1 together with one additional variable.

Backward Stepwise Selection



Like forward stepwise selection, backward stepwise selection is not guaranteed to yield the best model containing a subset of the p predictors. Backward selection requires that the number of samples n is larger than the number of variables p (so that the full model can be fit). In contrast, forward stepwise can be used even when n < p , and so is the only viable subset method when p is very large.

Hybrid Approaches:

As another alternative, hybrid versions of forward and backward stepwise selection are available, in which variables are added to the model sequentially, in analogy to forward selection. However, after adding each new variable, the method may also remove any variables that no longer provide an improvement in the model fit. Such an approach attempts to more closely mimic best subset selection while retaining the computational advantages of forward and backward stepwise selection.

6.1.3 Choosing the Optimal Model:

As forward selection and backward selection methods using 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 with respect to test error, we need to estimate this test error. There are two common approaches:

1. We can indirectly estimate test error by making an adjustment to the training error to account for the bias due to overfitting.

2. We can directly estimate the test error, using either a validation set approach or a cross-validation approach, as discussed in Chapter 5.

We consider both of these approaches below.

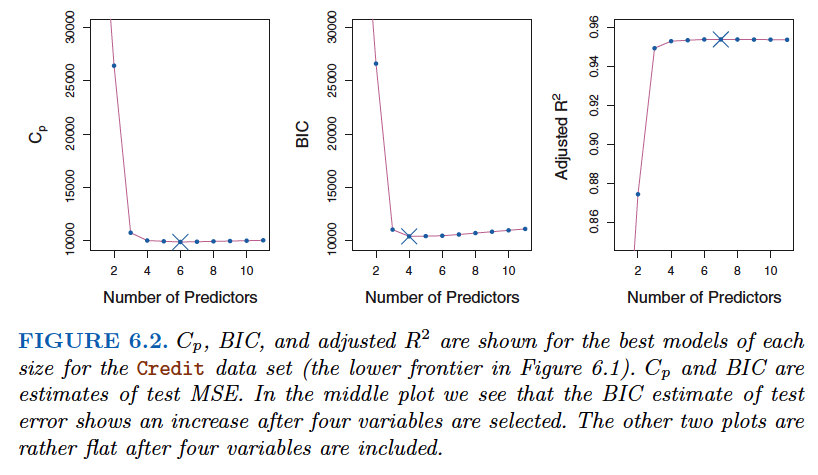
Cp, AIC, BIC, and Adjusted R2:

We show in Chapter 2 that the training set MSE is generally an underestimate of the test MSE. (Recall that MSE = RSS/n .) This is because when we fit a model to the training data using least squares, we specifically estimate the regression coefficients such that the training RSS (but not the test RSS) is as small as possible. In particular, the training error

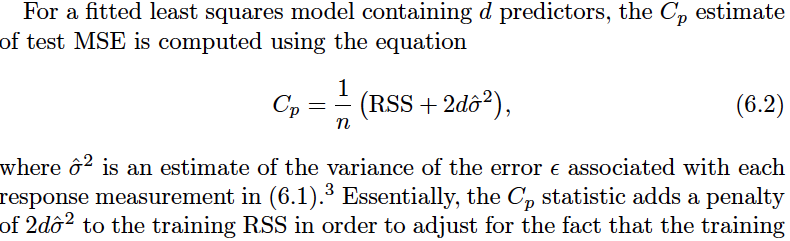
will decrease as more variables are included in the model, but the test error may not.

However, a number of techniques for adjusting the training error for the model size are available. We now consider four such approaches:

* Cp ,
* Akaike information criterion (AIC),
* Bayesian information criterion (BIC),
* adjusted R2 .



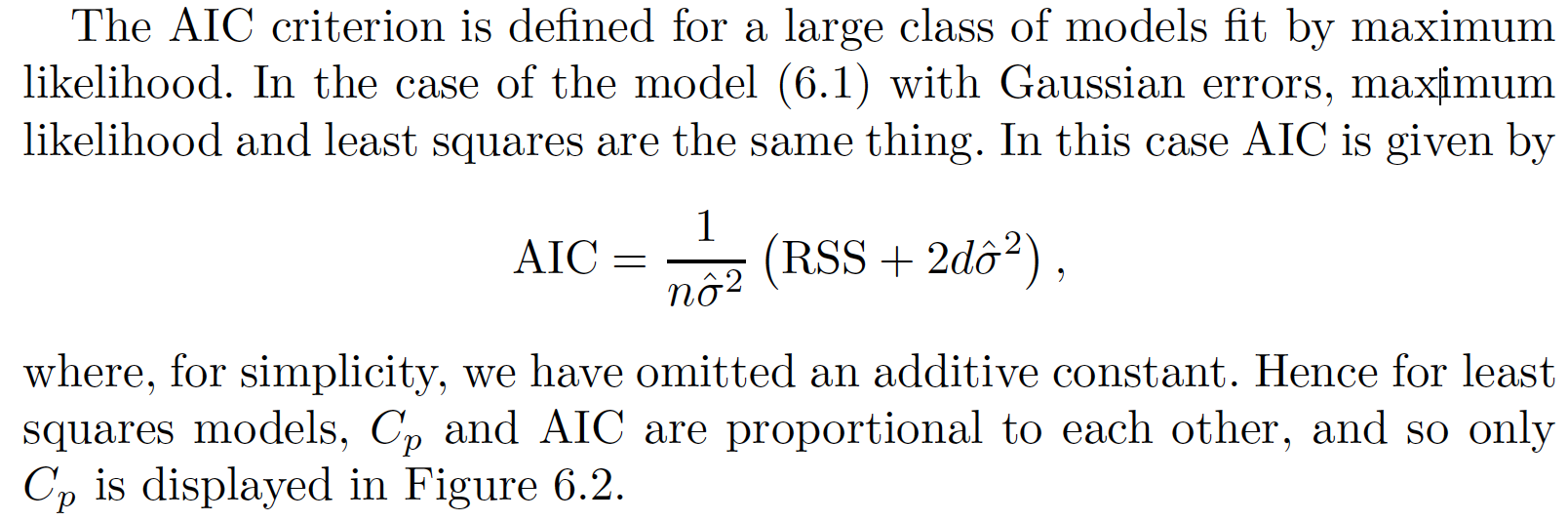
***Cp*:**



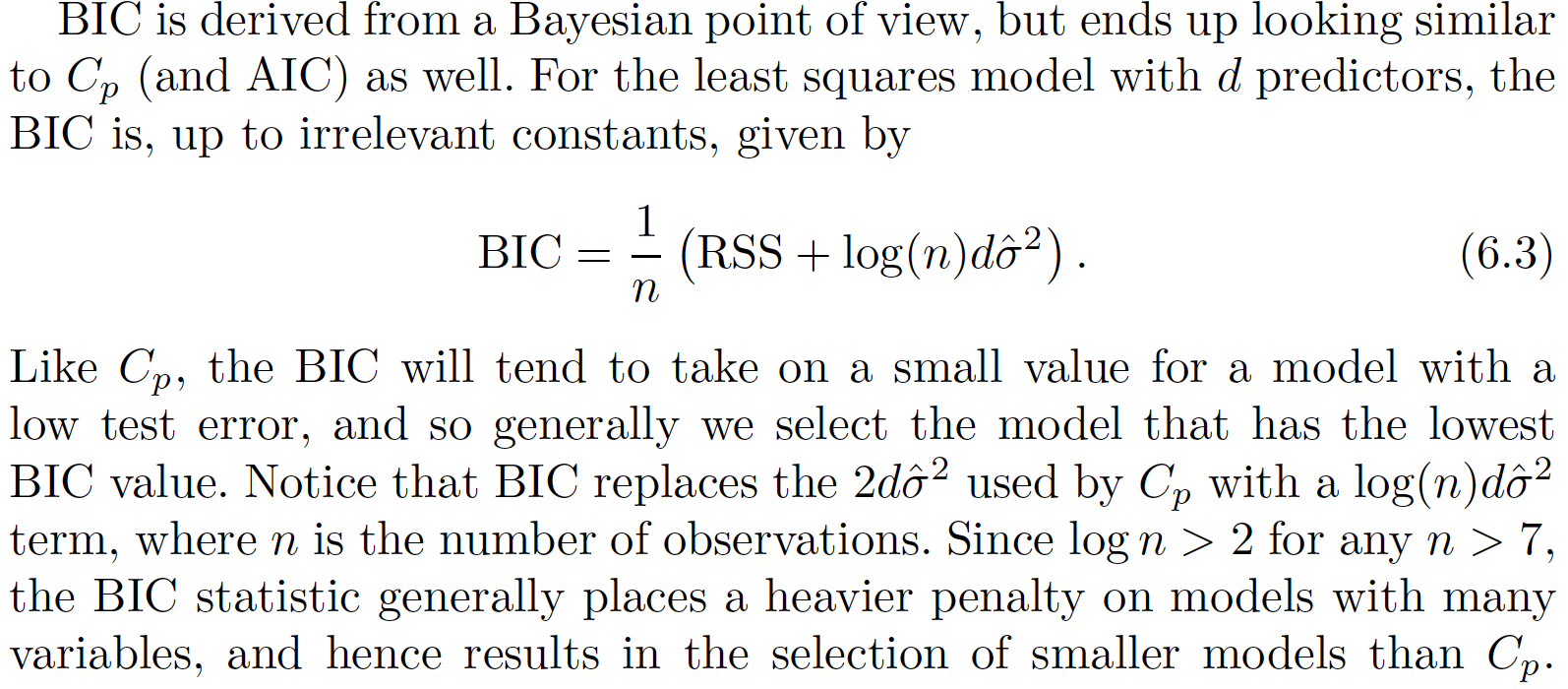
error tends to underestimate the test error. Clearly, the penalty increases as the number of predictors in the model increases; this is intended to adjust for the corresponding decrease in training RSS.

***so when determining which of a set of models is best, we choose the model with the lowest Cp value.***

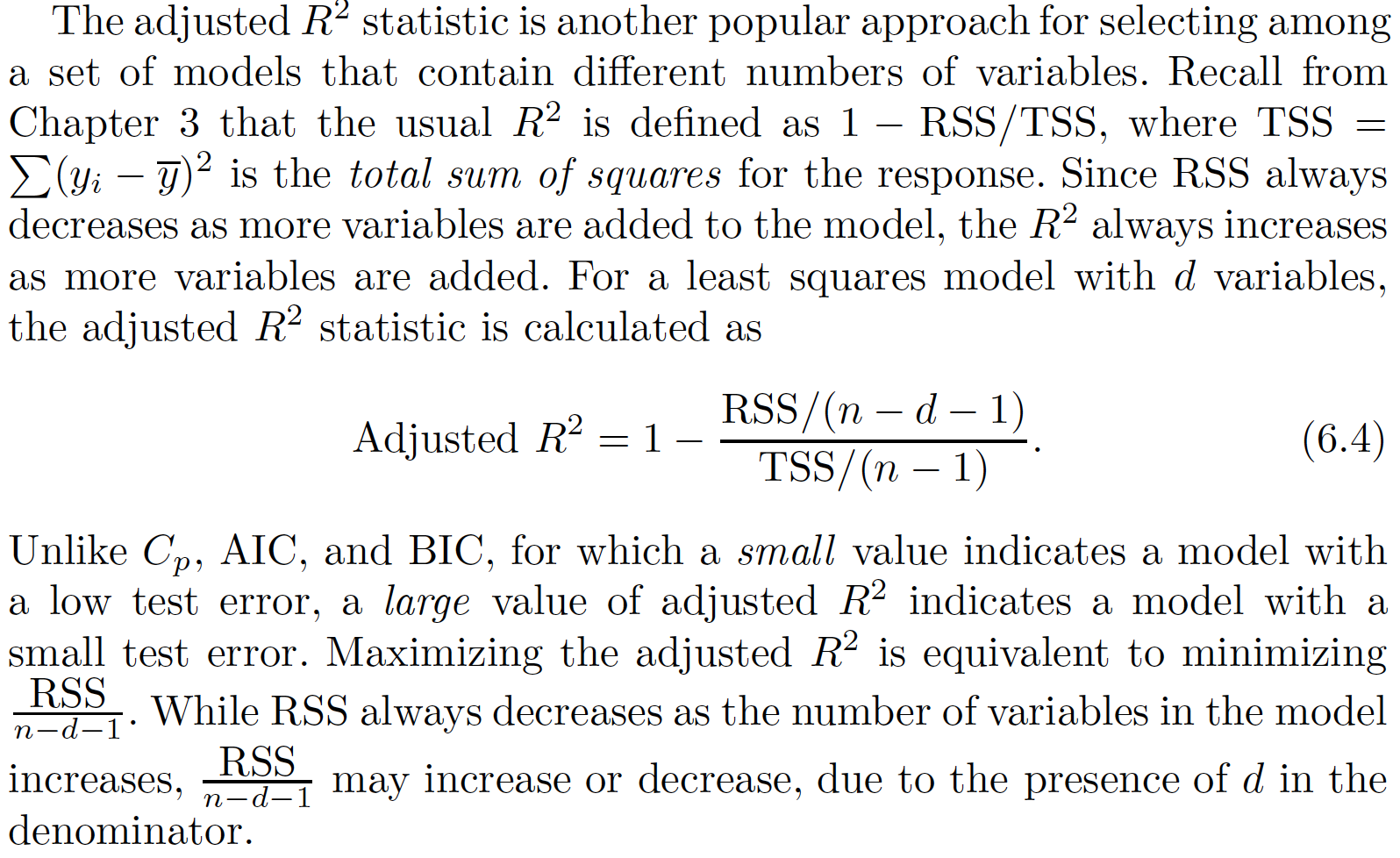
AIC: Akaike information criterion



BIC: Bayesian information criterion



**Adjusted R2:**



Therefore, in theory, the model with the largest adjusted R2 will have only correct variables and no noise variables. Unlike the R2 statistic, the adjusted R2 statistic pays a price for the inclusion of unnecessary variables in the model.

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 relative to AIC, BIC, Cp , and adjusted R2 , in that it provides a direct estimate of the test error, and makes fewer assumptions about the true underlying model.

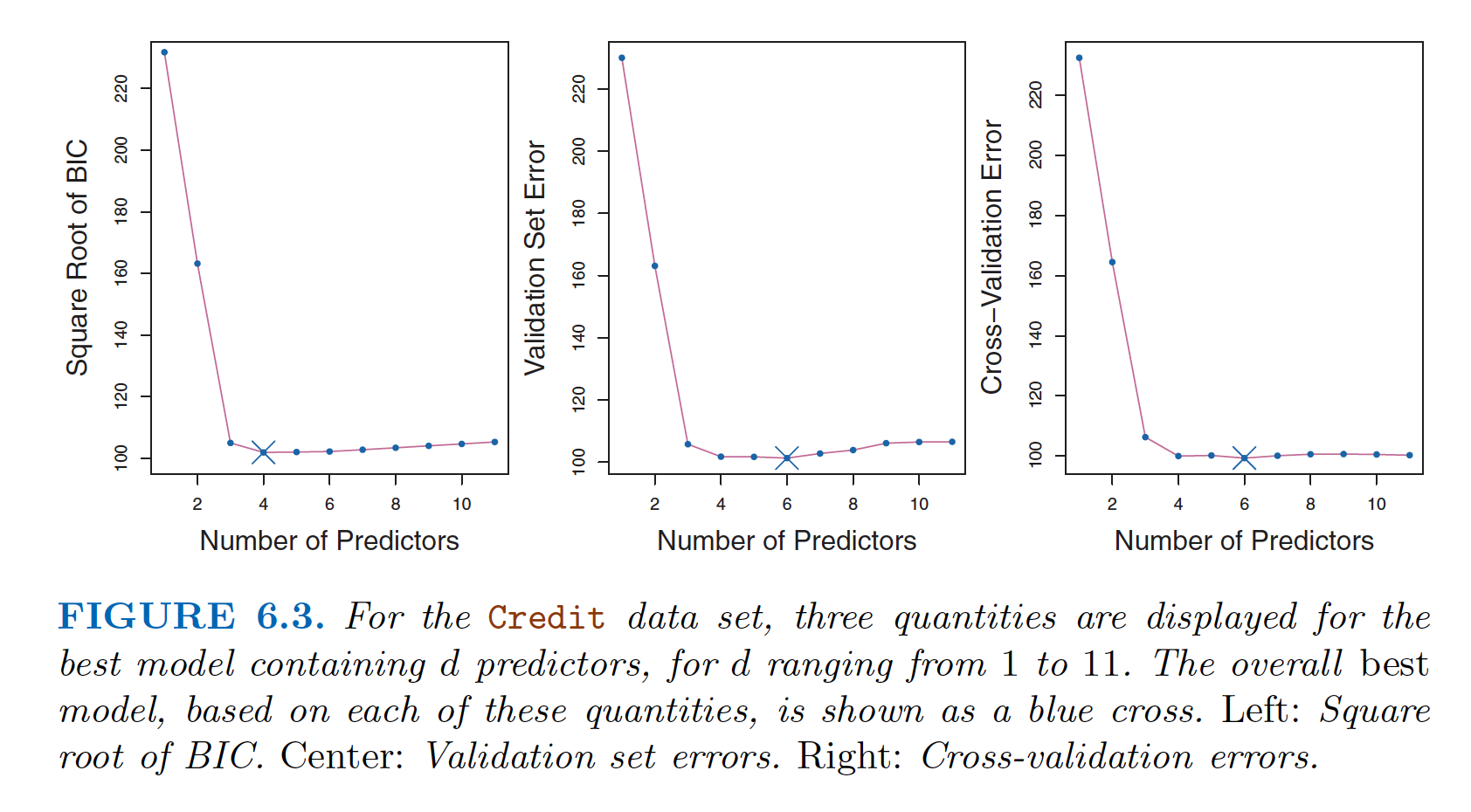


Figure 6.3 displays, as a function of d , the BIC, validation set errors, and cross-validation errors on the Credit data, for the best d -variable model. However, all three approaches suggest that the four-, five-, and six-variable models are roughly equivalent in terms of their test errors.

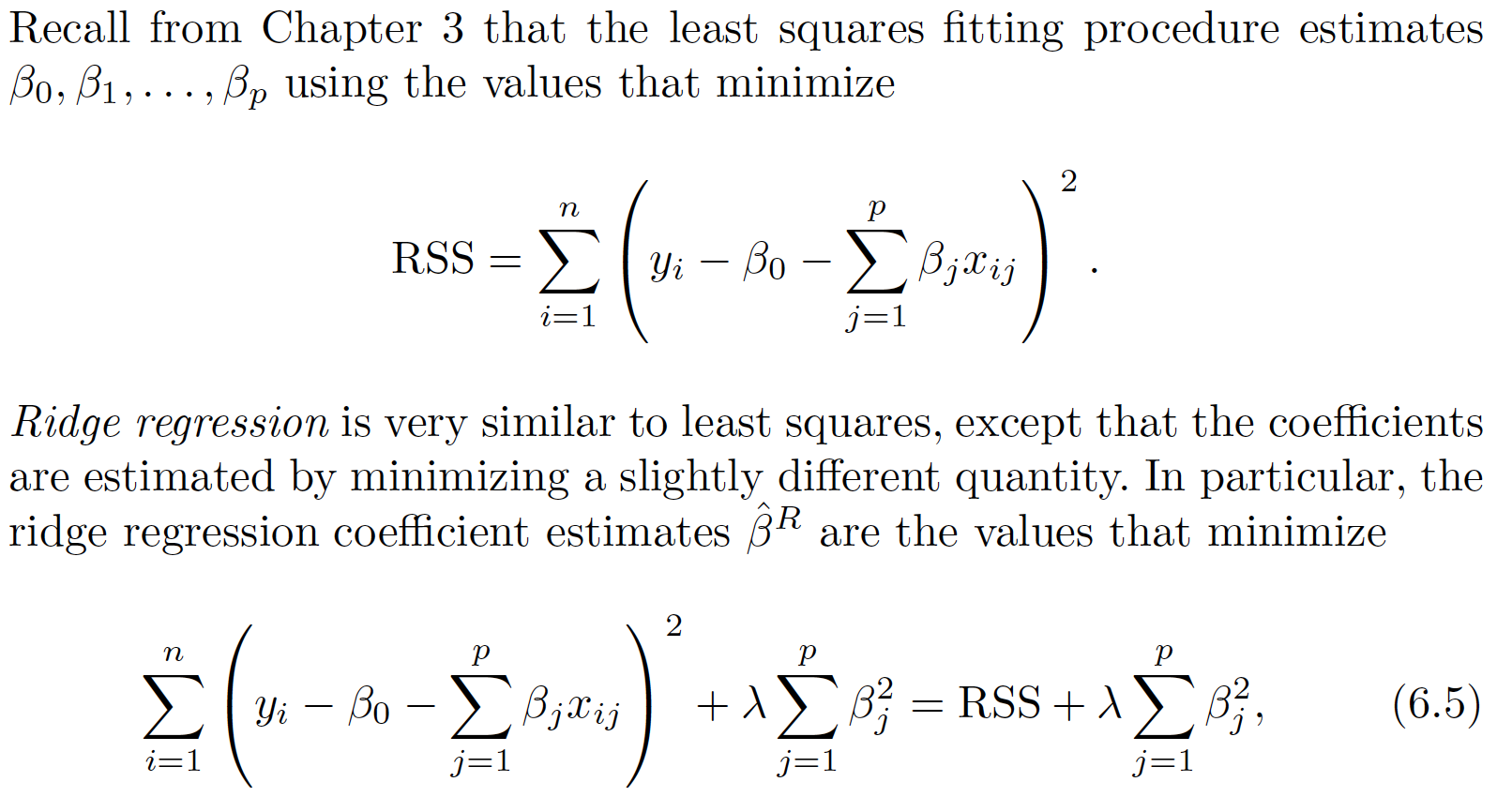
In fact, the estimated test error curves displayed in the center and righthand panels of Figure 6.3 are quite flat.While a three-variable model clearly has lower estimated test error than a two-variable model, the estimated test errors of the 3- to 11-variable models are quite similar. Furthermore, if we repeated the validation set approach using a different split of the data into a training set and a validation set, or if we repeated cross-validation using a different set of cross-validation folds, then the precise model with the lowest estimated test error would surely change. In this setting, we can select a model using the ***one-standard-error rule*** . We first 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 standard error of the lowest point on the curve. The rationale here is that if a set of models appear to be more or less equally good, then we might as well choose the simplest model—that is, the model with the smallest number of predictors. In this case, applying the one-standard-error rule to the validation set or cross-validation approach leads to selection of the three-variable model.

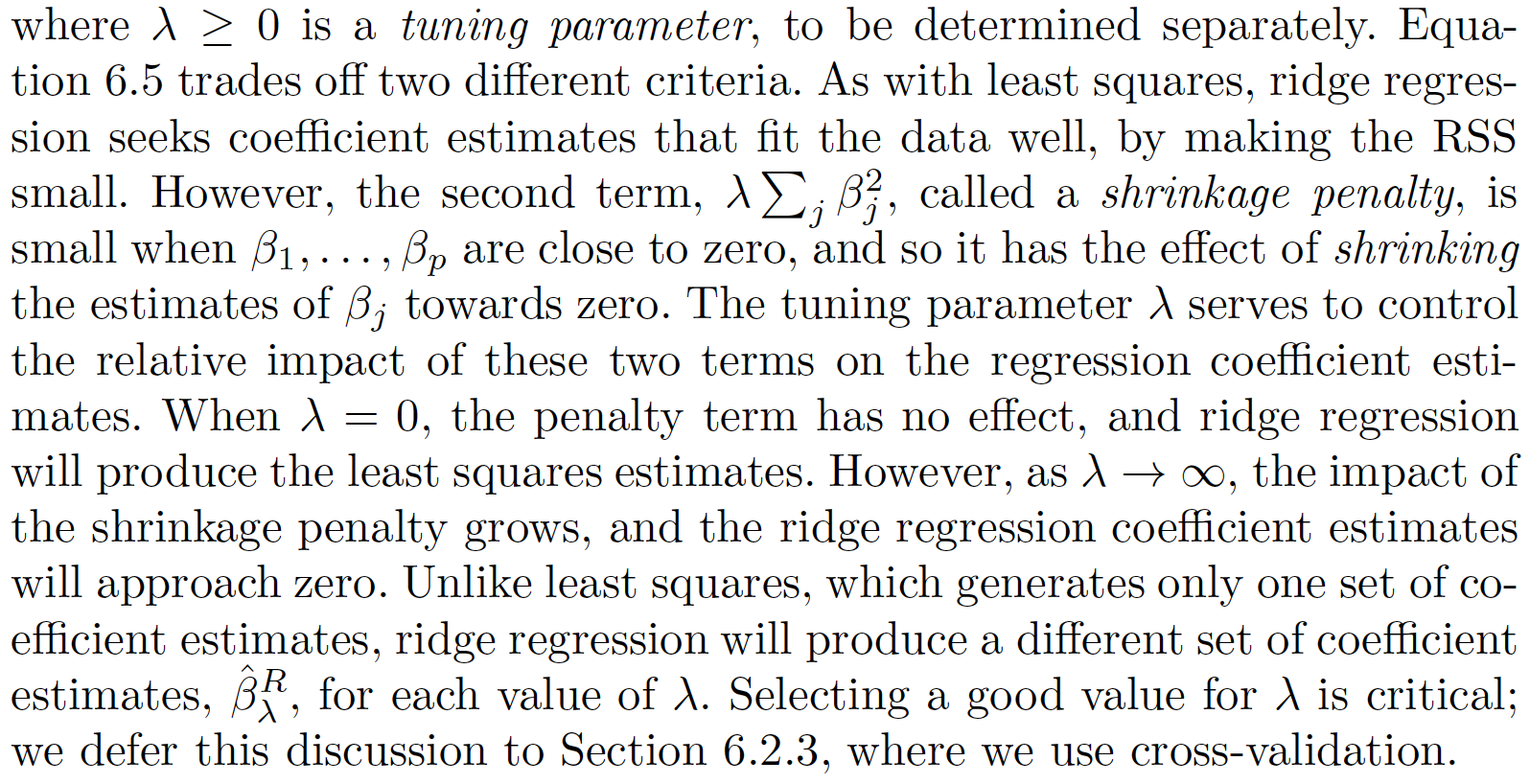
6.2 Shrinkage Methods:

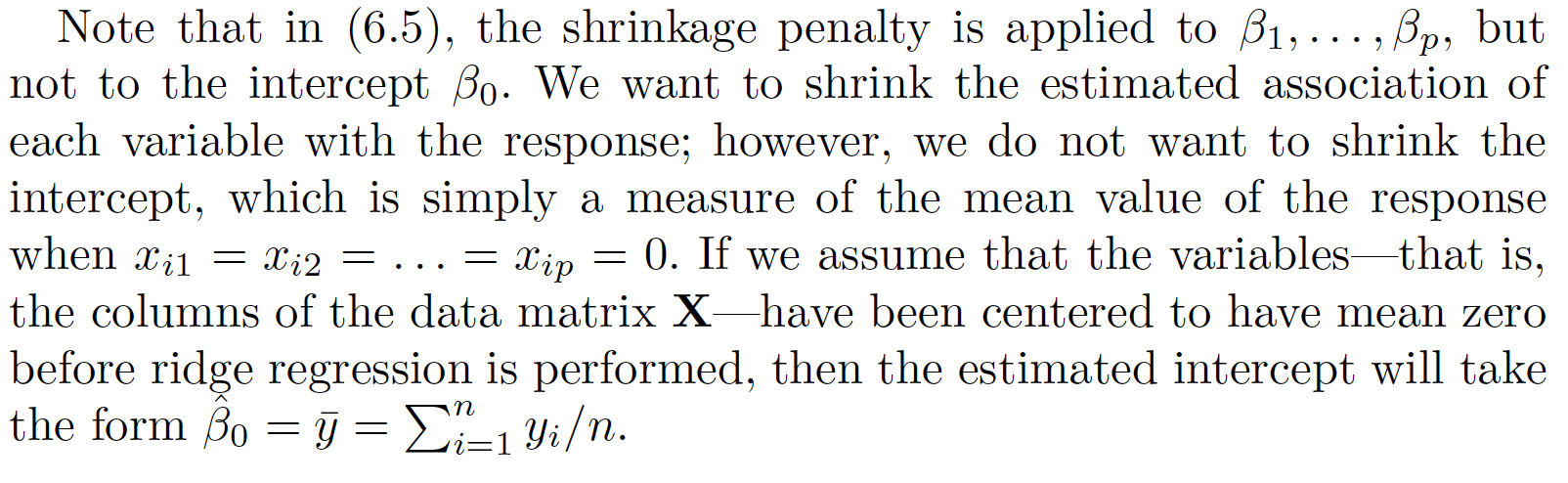
The subset selection methods described in Section 6.1 involve using least squares to fit a linear model that contains a subset of the predictors. As an alternative, we can fit a model containing all p predictors using a technique that constrains or regularizes the coefficient estimates, or equivalently, that shrinks the coefficient estimates towards zero. It may not be immediately obvious why such a constraint should improve the fit, but it turns out that shrinking the coefficient estimates can significantly reduce their variance.

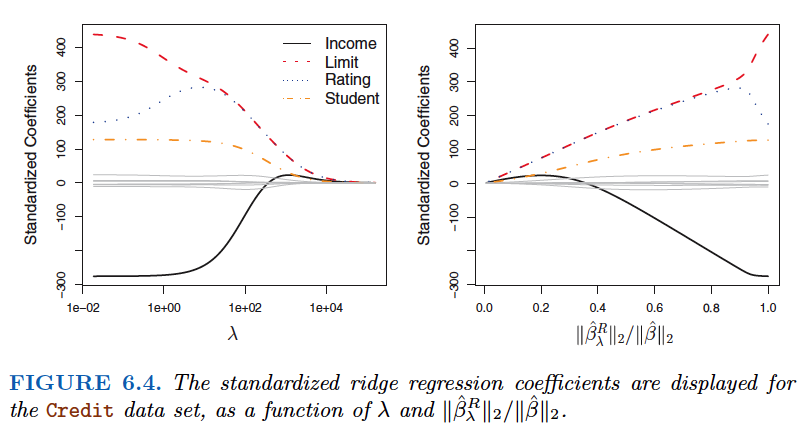
The two best-known techniques for shrinking the regression coefficients towards zero are ridge regression and the lasso .

6.2.1 Ridge Regression:

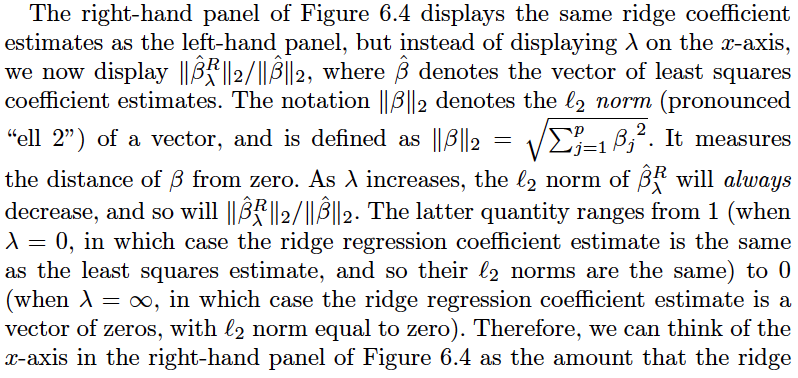






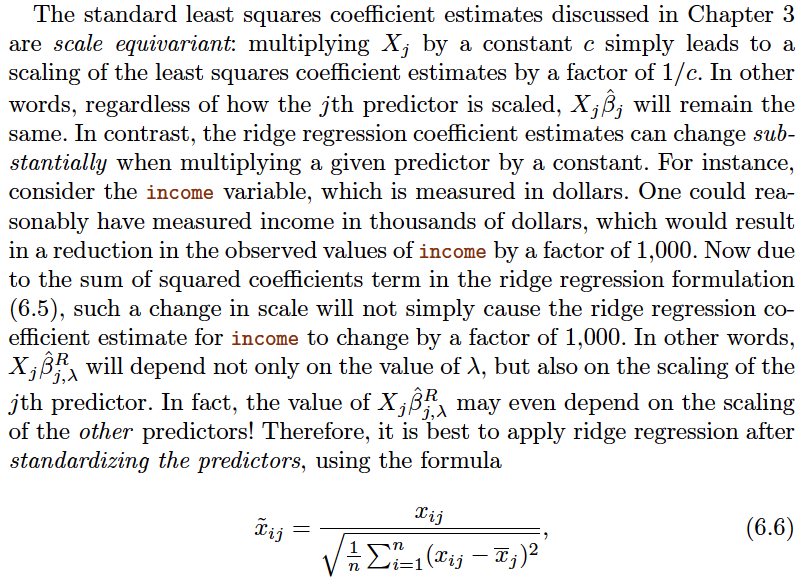


When λ is extremely large, then all of the ridge coefficient estimates are basically zero; this corresponds to the null model that contains no predictors.



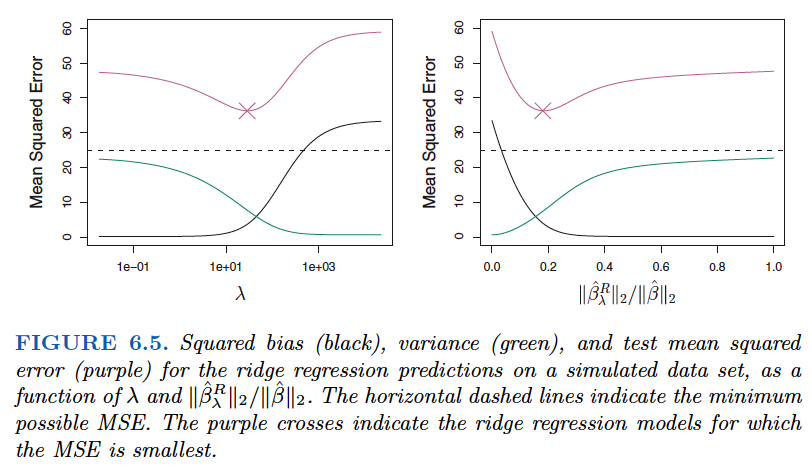
regression coefficient estimates have been shrunken towards zero; a small value indicates that they have been shrunken very close to zero.

Standardizing the predictors needed:



Why Does Ridge Regression Improve Over Least Squares?

Ridge regression’s advantage over least squares is rooted in the bias-variance trade-off . As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias.



The right-hand panel of Figure 6.5 displays the same curves as the lefthand panel, this time plotted against the l2 norm of the ridge regression coefficient estimates divided by the l2 norm of the least squares estimates. Now as we move from left to right, the fits become more flexible, and so the bias decreases and the variance increases.

In general, in situations where the relationship between the response and the predictors is close to linear, the least squares estimates will have low bias but may have high variance. This means that a small change in the training data can cause a large change in the least squares coefficient estimates. In particular, when the number of variables p is almost as large as the number of observations n , as in the example in Figure 6.5, the least squares estimates will be extremely variable. And if p > n , then the least squares estimates do not even have a unique solution, whereas ridge regression can still perform well by trading off a small increase in bias for a large decrease in variance.

***Advantages:***

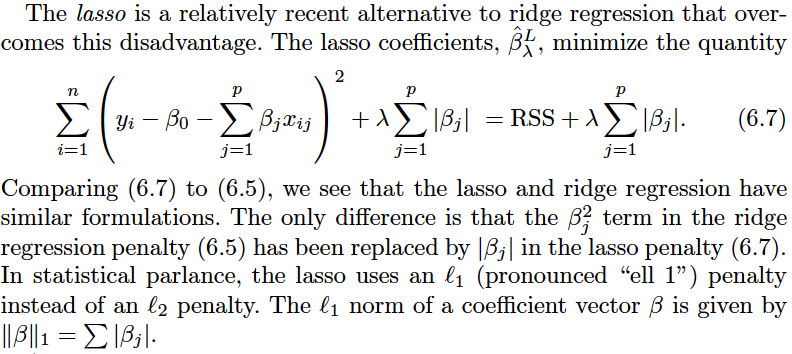
Hence, ridge regression works best in situations where the least squares estimates have high variance.

Ridge regression also has substantial computational advantages over best subset selection, which requires searching through 2^p models.

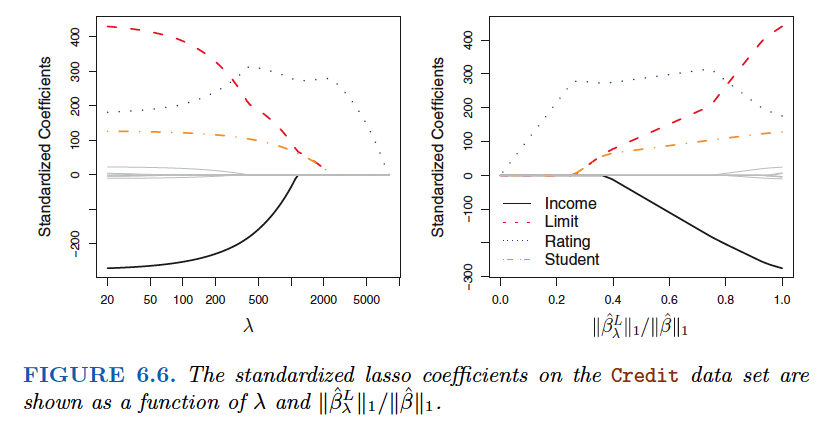
***Dis-advantages:***

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 l2 norm of coeff s in (6.5) will shrink all of the coefficients towards zero, but it will not set any of them exactly to zero (unless λ = ∞ ). 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 variables p is quite large.

The Lasso:



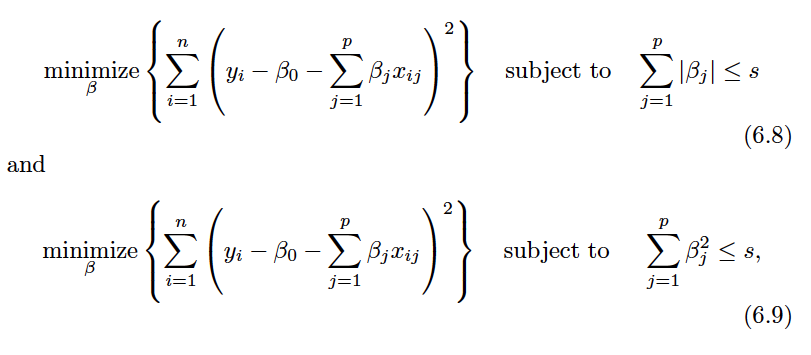
* As with ridge regression, the lasso shrinks the coefficient estimates towards zero.
* However, in the case of the lasso, the l1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large.
* Hence, much like best subset selection, the lasso performs variable selection . As a result, models generated from the lasso are generally much easier to interpret than those produced by ridge regression.
* As in ridge regression, selecting a good value of λ for the lasso is critical;



Hence, depending on the value of λ , the lasso can produce a model involving any number of variables. In contrast, ridge regression will always include all of the variables in the model, although the magnitude of the coefficient estimates will depend on λ .

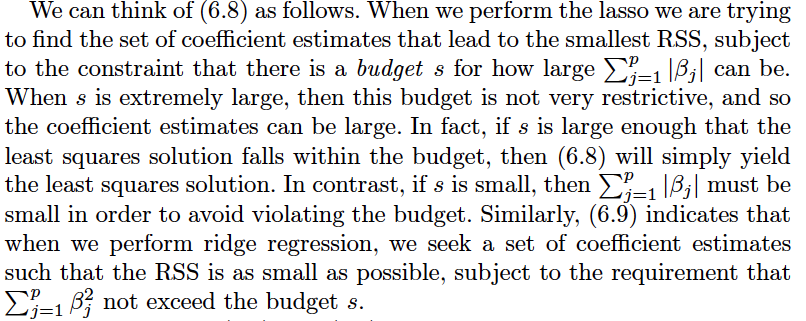
Another Formulation for Ridge Regression and the Lasso

One can show that the lasso and ridge regression coefficient estimates solve the problems



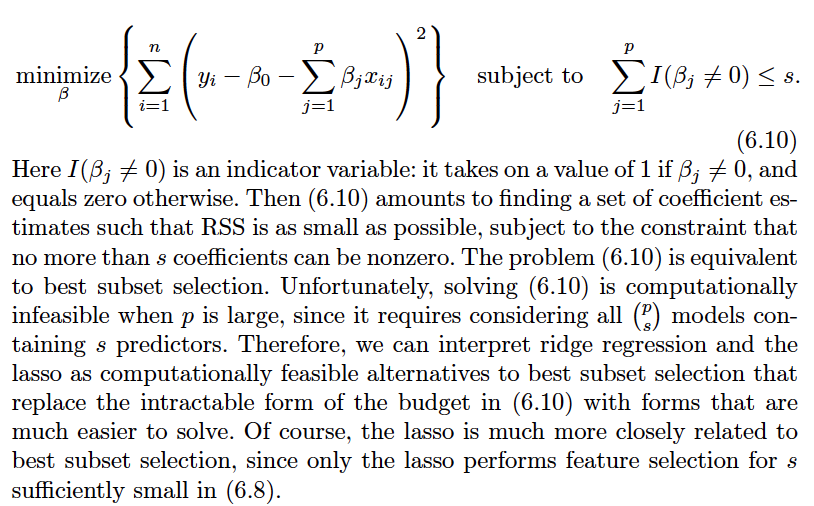
In other words, for every value of λ , there is some s such that the Equations (6.7) and (6.8) will give the same lasso coefficient estimates. Similarly, for every value of λ there is a corresponding s such that Equations (6.5) and (6.9) will give the same ridge regression coefficient estimates.

When p = 2, then (6.8) indicates that the lasso coefficient estimates have the smallest RSS out of all points that lie within the diamond defined by |β1| + |β2| ≤ s . Similarly, the ridge regression estimates have the smallest RSS out of all points that lie within the circle defined by



Formulation for Best subset selection Method:

The formulations (6.8) and (6.9) reveal a close connection between thelasso, ridge regression, and best subset selection. Consider the problem



The Variable Selection Property of the Lasso

Why is it that the lasso, unlike ridge regression, results in coefficient estimates that are exactly equal to zero?

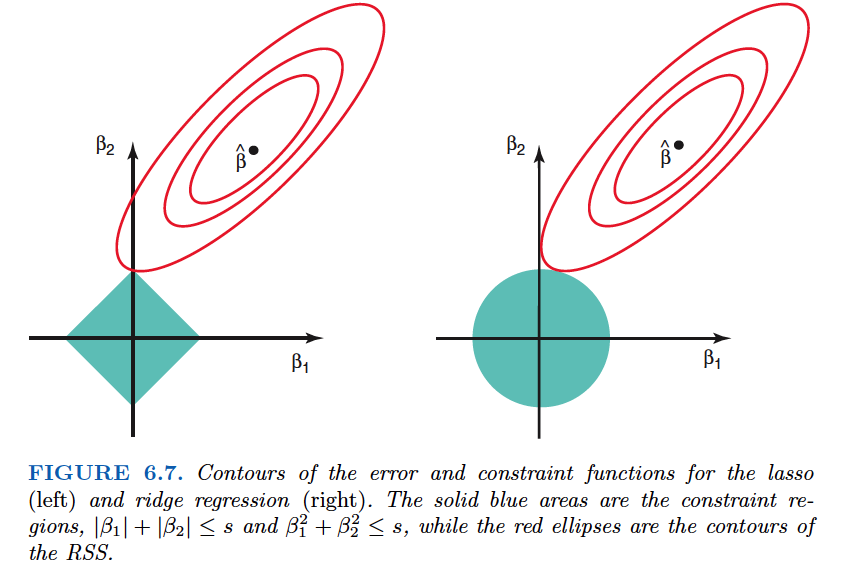


Figure 6.7 illustrates the situation. The least squares solution is marked as ˆβ , while the blue diamond and circle represent the lasso and ridge regression constraints in (6.8) and (6.9), respectively. If s is sufficiently large, then the constraint regions will contain ˆ β , and so the ridge regression and lasso estimates will be the same as the least squares estimates. (Such a large value of s corresponds to λ = 0 in (6.5) and (6.7).) However, in Figure 6.7 the least squares estimates lie outside of the diamond and the circle, and so the least squares estimates are not the same as the lasso and ridge regression estimates.

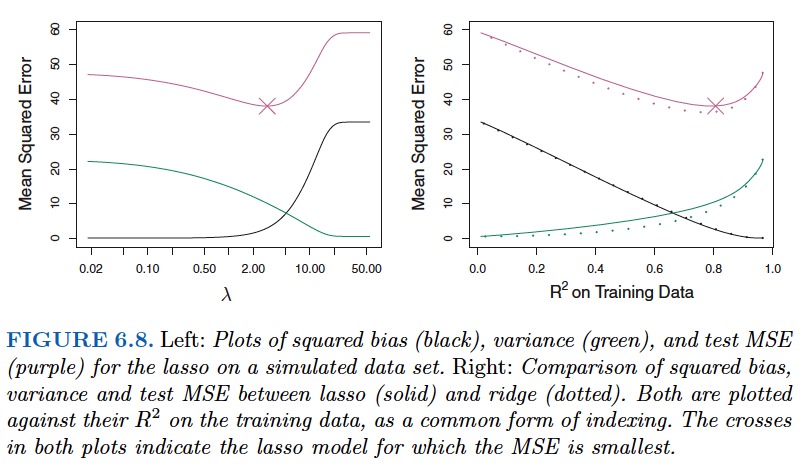
Since ridge regression has a circular constraint with no sharp points, this intersection will not generally occur on an axis, and so the ridge regression coefficient estimates will be exclusively non-zero. However, the lasso constraint has corners at each of the axes, and so the ellipse will often intersect the constraint region at an axis. When this occurs, one of the coefficients will equal zero. In higher dimensions, many of the coefficient estimates may equal zero simultaneously. In Figure 6.7, the intersection occurs at β1 = 0, and so the resulting model will only include β2 .

In Figure 6.7, we considered the simple case of p = 2. When p = 3, then the constraint region for ridge regression becomes a sphere, and the constraint region for the lasso becomes a polyhedron. When p > 3, the constraint for ridge regression becomes a hypersphere, and the constraint for the lasso becomes a polytope. However, the key ideas depicted in Figure 6.7 still hold. In particular, the lasso leads to feature selection when p > 2 due to the sharp corners of the polyhedron or polytope.

Comparing the Lasso and Ridge Regression:

It is clear that the lasso has a major advantage over ridge regression, in that it produces simpler and more interpretable models that involve only a subset of the predictors.

However, which method leads to better prediction accuracy?



However, the data in Figure 6.8 were generated in such a way that all 45 predictors were related to the response—that is, none of the true coefficients β1, . . . , β45 equaled zero. The lasso implicitly assumes that a number of the coefficients truly equal zero. Consequently, it is not surprising that ridge regression outperforms the lasso in terms of prediction error in this setting.

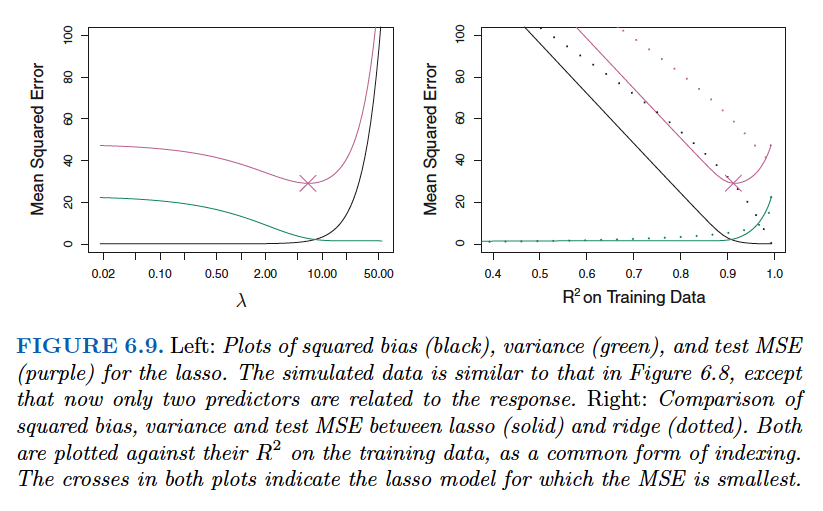


Figure 6.9 illustrates a similar situation, except that now the response is a function of only 2 out of 45 predictors. Now the lasso tends to outperform ridge regression in terms of bias, variance, and MSE.

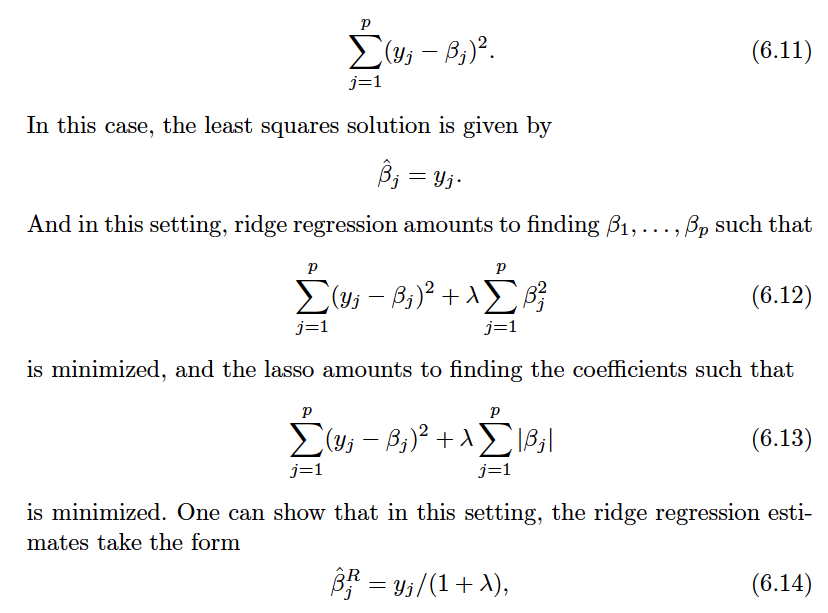
These two examples illustrate that neither ridge regression nor the lasso will universally dominate the other.

* In general, one might expect the lasso 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 perform better when the response is a function of many predictors, all with coefficients of roughly equal size.

*However, the number of predictors that is related to the response is never known a priori for real data sets. A technique such as cross-validation can be used in order to determine which approach is better on a particular data set.*

A Simple Special Case for Ridge Regression and the Lasso

In order to obtain a better intuition about the behavior of ridge regression and the lasso, consider a simple special case with n = p , and X a diagonal matrix with 1’s on the diagonal and 0’s in all off-diagonal elements. To simplify the problem further, assume also that we are performing regression without an intercept. With these assumptions, the usual least squares problem simplifies to finding β1, . . . , βp that minimize



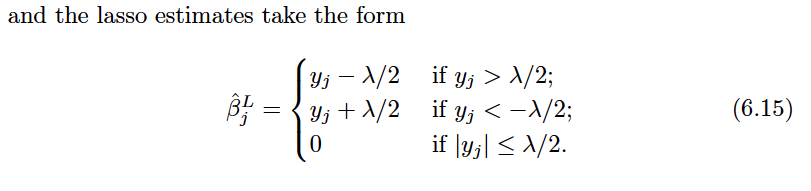
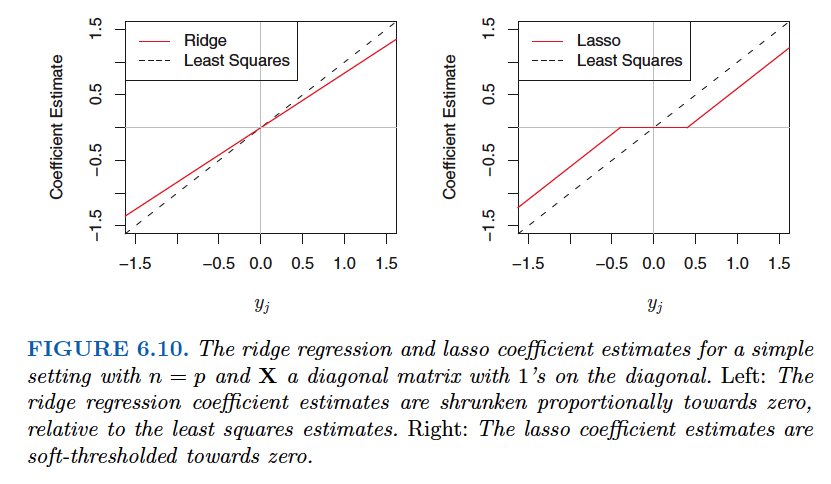


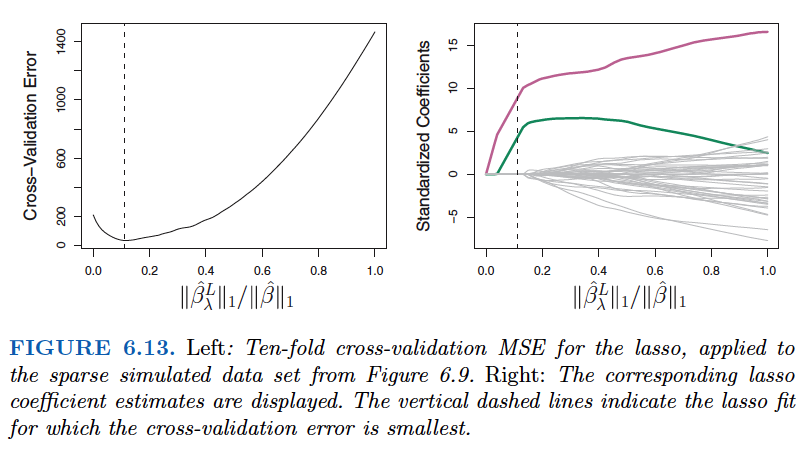
Figure 6.10 displays the situation. We can see that ridge regression and the lasso perform two very different types of shrinkage. In ridge regression, each least squares coefficient estimate is shrunken by the same proportion. In contrast, the lasso shrinks each least squares coefficient towards zero by a constant amount, λ/ 2; the least squares coefficients that are less than λ/ 2 in absolute value are shrunken entirely to zero. The type of shrinkage performed by the lasso in this simple setting (6.15) is known as softthresholding . The fact that some lasso coefficients are shrunken entirely to zero explains why the lasso performs feature selection.

In the case of a more general data matrix X , the story is a little more complicated than what is depicted in Figure 6.10, but the main ideas still hold approximately: ridge regression more or less shrinks every dimension of the data by the same proportion, whereas the lasso more or less shrinks all coefficients toward zero by a similar amount, and sufficiently small coefficients are shrunken all the way to zero.



Bayesian Interpretation for Ridge Regression and the Lasso

Study Later\*\*



6.3 Dimension Reduction Methods

The methods that we have discussed so far in this chapter have controlled variance in two different ways, either by using a subset of the original variables, or by shrinking their coefficients toward zero. All of these methods are defined using the original predictors, X1,X2, . . . , Xp . We now explore a class of approaches that transform the predictors and then fit a least squares model using the transformed variables. We will refer to these techniques as dimension reduction methods.

