CHAPTER 14

Prediction with Many Regressors and Big Data

Chapter 4 began with two different questions about student performance at elementary schools. A superintendent wanted to know whether test scores would improve if she reduced the student–teacher ratio in her schools—and if they would, by how much. A father, trying to decide where to live, wanted to predict which schools had the highest-performing students. Answering the superintendent's question requires you to estimate the causal effect on test scores of the student–teacher ratio, and estimating causal effects is the focus of Chapters 4–13. In contrast, answering the father's question requires you to predict school test scores given one or more relevant variables—in Chapter 4, the student-teacher ratio, extended in Chapter 6 to include additional information on school and community characteristics.

Statistical prediction entails using data to estimate a prediction model and then applying that model to new, out-of-sample observations. The goal is accurate out-of-sample prediction. In a prediction problem, there are neither specific regressors of interest nor control variables; there are only predictors and the variable to be predicted.

If there are only a handful of predictors, ordinary least squares (OLS) works well if the least squares assumptions for prediction in Appendix 6.4 hold. But modern data sets often have many predictors. For example, the empirical application in this chapter is the prediction of school-level test scores using data on school and community characteristics. We use data on 3932 elementary schools in California; half of these observations are used to estimate prediction models, while the other half are reserved to test their performance.[1] For most of the chapter, we consider a data set with 817 predictors, which is expanded in Section 14.6 to 2065 predictors. This problem of predicting school test scores is typical of many prediction applications using cross-sectional data, such as forecasting sales for a business, predicting patient-level outcomes of medical procedures, or predicting demand for services by state and local government. In such applications, the number of predictors can be nearly as large as, or even larger than, the number of observations.

With so many predictors, OLS overfits the data and makes poor out-of-sample predictions. Fortunately, it is possible to improve upon OLS by using estimators that are broadly referred to as shrinkage estimators. These estimators are biased (they "shrink" the estimator), and the coefficients, in general, do not have a causal interpretation. Remarkably, however, when there are many predictors, introducing bias can reduce the variance of the estimator sufficiently that the overall out-of-sample prediction accuracy is improved.

[1]In California, a school district typically contains multiple individual schools. The test score data set used in Chapters 4–9 contains district-level data, while the data used here are for individual schools.

This chapter considers prediction using cross-sectional data sampled from a larger population (shoppers, patients, schools) to predict outcomes for members of the population not in the estimation sample. A related problem is prediction of future events, such as the number of jobs the economy will add next month. Predictions about the future are typically referred to as forecasts, and we adopt that terminology. Forecasting uses time series data, which introduce additional notation and technicalities. Forecasting is taken up in Part IV.

The availability of many predictors is one of the opportunities provided by very large data sets. The field of analyzing big data sets goes by multiple names, including machine learning, data science, and the term we shall use, *big data*.

14.1 What Is "Big Data"?

Data sets can be big in the sense of having many observations, or having many predictors relative to the number of observations, or both. Big data sets can be nonstandard—for example, containing text or images.

Big data sets make available new families of applications. One such family, which is the focus of this chapter, is prediction when the number of predictors k is large compared to the number of observations n. The prediction methods considered in this chapter start with linear regression, so having many predictors corresponds to having many regressors. This situation can arise if one has many distinct primitive predictors, or it can arise if one is considering predictions that are nonlinear functions of the primitive predictors. Even if one starts with only a few dozen primitive predictors, including squares, cubes, and interactions very quickly expands the number of regressors into the hundreds or thousands.

A second family of applications that arises with big data is categorization. We have encountered this problem before, in the context of regression with a binary dependent variable. The logit and probit models of Chapter 11 predict the probability that the dependent variable is 1—in the empirical application, the probability that a loan application is denied. An alternative framing of this problem is to divide the data set into two groups, or categories: those applications that are likely to be denied and those that are likely to be accepted. From a prediction perspective, the aim is to develop a model of loan applications that mimics the decision-making process of a loan officer. Said differently, by fitting that model, a machine (computer) would have learned (estimated) the decision process made by a loan officer. Using that machine learning model, the computer then can make the accept/deny decision itself for future applications. Indeed, the online home loan application industry relies heavily on machine learning, applied to very large data sets on loan applications, to assess the eligibility of an applicant for a mortgage.

A third family of applications concerns testing multiple hypotheses. In the regression context, for example, there might be a potentially large set of coefficients representing different treatments, and the econometrician might be interested in

ascertaining which, if any, of these treatments is effective. Because the F-statistic tests a joint hypothesis on a group of coefficients, it is not well suited for the problem of testing many treatments to find out *which* of the treatments is effective. Testing many individual hypotheses with the aim of determining which treatment effect is nonzero requires specialized methods that have been developed for big data applications.

A fourth family of applications concerns handling nonstandard data, such as text and images. The key step is turning these nonstandard data into numerical data, which can then be handled using techniques for high-dimensional data sets. Section 14.7 discusses methods for handling text data.

A fifth, related family of applications is pattern recognition, such as facial recognition or translating text from one language to another. This area has seen great progress using procedures such as "deep learning," which are in essence highly nonlinear models estimated ("trained") using very many observations.

A common feature of all of these problems is that handling large data sets creates computational challenges. Those challenges include storing and accessing large data sets efficiently and developing fast algorithms for estimating models. These computational issues are important; however, we do not address them in this chapter and instead leave them to computer science curricula.

The results of machine learning applied to large data sets are increasingly part of our everyday world. Examples range from software that helps doctors make diagnoses to techniques that target online advertisements to facial recognition algorithms that are used by law enforcement officials. In economics, applications include estimating local incomes based on satellite data, predicting sales for a firm using detailed customer data, interpreting network data on social media sites, searching for patterns in high-frequency asset price databases to use in computerized trading algorithms, and forecasting macroeconomic growth using up-to-the-minute data. Increasingly, computerized analysis of nonstandard data, especially text data, is playing a role in econometric applications.

This chapter cannot cover all these uses of big data, so it focuses on one of the most important for economic applications: the many-predictor problem. Although the nomenclature of this growing field—machine learning, data science, and so forth—makes it seem difficult and new, the methods discussed in this chapter are, at their core, extensions of linear regression analysis that are tailored to the opportunities and challenges of large data sets.

14.2 The Many-Predictor Problem and OLS

This chapter considers the problem of predicting test scores for a school using variables describing the school, its students, and its community. The full data set consists of data gathered on 3932 elementary schools in the state of California in 2013. The task is to use these data to develop a prediction model that will provide good out-of-sample predictions—that is, predictions for schools not in the data set. To simulate

the out-of-sample prediction problem, for most of the chapter we use half the observations ($n = 1966$) for estimating prediction models. The remaining half of the observations are reserved as a test data set to assess how the models perform and are not used until Section 14.6.

The variable to be predicted is the average fifth-grade test score at the school. The primary data set contains 817 distinct variables relating to school and community characteristics; these variables are summarized in Table 14.1. For comparison, smaller and larger data sets are used in Section 14.6. The data are described in more detail in Appendix 14.1.

If only the main variables in Table 14.1 were used, there would be 38 regressors. The analysis of the district test score data in Section 8.4, however, revealed several interesting nonlinearities and interactions in the test score regressions. For example, the regressions in Table 8.3 indicate that there is a nonlinear relationship between test scores and the student–teacher ratio and, in addition, that this relationship differs depending on whether there are a large number of English learners in the district. In Section 8.4, these nonlinearities were handled by including third-degree polynomials of the student–teacher ratio and interaction terms. As laid out in Table 14.1, including interactions, squares, and cubes increases the number of predictors to 817. In Section 14.6, we consider an even larger data set with 2065 predictors, which exceeds the 1966 observations in the estimation sample! Regression with 817 regressors, not to mention 2065 regressors, goes well beyond anything attempted so far in this text.

A natural starting point is OLS. Unfortunately, OLS can produce quite poor predictions when the number of predictors is large relative to the sample size. Fortunately, there are estimators other than OLS that can produce more reliable predictions

TABLE 14.1 Variables in the 817-Predictor School Test Score Data Set

Main variables (38)	
Fraction of students eligible for free or reduced-price lunch	Ethnicity variables (8): fraction of students who are American Indian, Asian, Black, Filipino, Hispanic, Hawaiian, two or more, none reported
Fraction of students eligible for free lunch	Number of teachers
Fraction of English learners	Fraction of first-year teachers
Teachers' average years of experience	Fraction of second-year teachers
Instructional expenditures per student	Part-time ratio (number of teachers divided by teacher full-time equivalents)
Median income of the local population	Per-student expenditure by category, district level (7)
Student–teacher ratio	Per-student expenditure by type, district level (5)
Number of enrolled students	Per-student revenues by revenue source, district level (4)
Fraction of English-language proficient students	
Ethnic diversity index	
+ Squares of main variables (38)	
+ Cubes of main variables (38)	
+ All interactions of main variables $(38 \times 37/2 = 703)$	
Total number of predictors $= k = 38 + 38 + 38 + 703 = 817$	

when the number of predictors relative to the sample size is large. This fact might seem surprising in light of the Gauss–Markov theorem, which says that the OLS estimator has the lowest variance of all unbiased estimators as long as the Gauss–Markov conditions hold (Appendix 5.2). The reason for this surprising result, and the reason it does not violate the Gauss–Markov theorem, is that these alternative estimators are biased. Although the estimators are biased, their variance is sufficiently smaller than the variance of the OLS estimator for them to produce better predictions.

The Mean Squared Prediction Error

To compare prediction models, we need a quantitative measure of predictive accuracy. As we have throughout this text, we will use the square of the error—in this case, the error from out-of-sample predictions. Using the squared prediction error means that small errors receive little weight but large errors receive great weight. This makes sense in many prediction problems, where small errors have negligible impact but very large errors can undercut the usefulness and credibility of the prediction.

The **mean squared prediction error (MSPE)** is the expected value of the square of the prediction error that arises when the model is used to make a prediction for an observation not in the data set.

Stated mathematically, the MSPE is

$$MSPE = E[Y^{oos} - \hat{Y}(X^{oos})]^2, \tag{14.1}$$

where X^{oos} and Y^{oos} are out-of-sample ("oos") observations on X and Y and $\hat{Y}(x)$ is the predicted value of Y for a value x of the predictors. As usual, X is shorthand for the k separate predictors. The notation of Equation (14.1) is taken from Appendix 6.4 (the least squares assumptions for prediction). The notation distinguishes between the n observations $(X_i, Y_i), i = 1, \ldots, n$, used to estimate the prediction model that produces $\hat{Y}(x)$ and the out-of-sample observation for which the prediction is made. The out-of-sample observation is not used to estimate the prediction model.

From the perspective of minimizing the MSPE, the best possible prediction is the conditional mean—that is, $E(Y^{oos}|X^{oos})$ (Appendix 2.2 and Exercise 14.8). This best-possible prediction, $E(Y^{oos}|X^{oos})$, is sometimes called the **oracle prediction**. Because the conditional mean is unknown, the oracle prediction cannot be used in practice (it is infeasible); however, it is the benchmark against which to judge all feasible predictions. In the regression model, the oracle prediction corresponds to the prediction that would be made using the true (unknown) population regression coefficients.

The MSPE embodies two sources of prediction errors. First, even if the conditional mean were known, the prediction would be imperfect: The oracle prediction makes the prediction error, $Y^{oos} - E(Y^{oos}|X^{oos})$. Second, $E(Y^{oos}|X^{oos})$ is unknown, and estimating its parameters—that is, estimating the coefficients of the prediction model $\hat{Y}(x)$—introduces an additional source of error.

The First Least Squares Assumption for Prediction

The school test score application uses data on some (but not all) California schools to estimate the prediction model. We can have confidence that this prediction model will generalize to other California schools; however, we have much less confidence that it will apply to schools in Europe and even less confidence that it will apply to schools in India.

The first least squares assumption for prediction makes this intuition precise. This assumption, which was introduced in Appendix 6.4, states that the out-of-sample observation is drawn from the same distribution as the in-sample observations used to estimate the model:

> **First least squares assumption for prediction**: (X^{oos}, Y^{oos}) are randomly drawn from the same population distribution as the estimation sample $(X_i, Y_i), i = 1, \ldots, n$.

Because the in- and out-of-sample observations are drawn from the same distribution, the conditional mean, $E(Y|X)$, is the oracle prediction for both in- and out-of-sample observations.

The first least squares assumption for prediction is a statement about external validity: The in-sample model can be generalized to the out-of-sample observation of interest.

Although we refer to this assumption as the first least squares assumption for prediction, the requirement applies for estimation methods other than least squares. This condition is assumed to hold for the remainder of this chapter.

The Predictive Regression Model with Standardized Regressors

This chapter uses a modified version of the linear regression model in which the regressors are all standardized; that is, they are transformed to have mean 0 and variance 1. In addition, the dependent variable is transformed to have mean 0. By using standardized regressors, all the regression coefficients have the same units, a property used in the methods of Sections 14.3–14.5.

Let $(X^*_{1i}, \ldots, X^*_{ki}, Y^*_i), i = 1, \ldots, n$, denote the data as originally collected, where X^*_{ji} is the i^{th} observation on the j^{th} original regressor. The standardized regressors are $X_{ji} = (X^*_{ji} - \mu_{X^*_j})/\sigma_{X^*_j}$, where $\mu_{X^*_j}$ and $\sigma_{X^*_j}$ are, respectively, the population mean and standard deviation of $X^*_{j1}, \ldots, X^*_{jn}$. The transformed (demeaned) dependent variable is $Y_i = Y^*_i - \mu_{Y^*}$, where μ_{Y^*} is the population mean of $Y^*_1, \ldots, Y^*_n$.

With this notation, the **standardized predictive regression model** is the regression of Y, which has mean 0, on the k standardized X's:

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i. \tag{14.2}$$

The intercept is excluded from Equation (14.2) because all the variables have mean 0.

Because the regressors are standardized, the regression coefficients have the same units: β_j is the difference in the predicted value of Y associated with a one standard deviation difference in X_j^*, holding constant the other X's.

Because the focus of this chapter is prediction, we adopt throughout the prediction interpretation of the regression model in Appendix 6.4; that is, $E(Y|X) = \sum_{j=1}^{k} \beta_j X_j$ and $E(u|X) = 0$.

As usual, the linear structure in Equation (14.2) means that the predictions are linear in the coefficients; however, the regression function can be nonlinear in the predictors because X can include nonlinear terms such as squares or interactions.

The MSPE in the standardized predictive regression model. In the standardized regression model in Equation (14.2), the prediction for the out-of-sample value of the predictors is $\hat{Y}(X^{oos}) = \hat{\beta}_1 X_1^{oos} + \ldots + \hat{\beta}_k X_k^{oos}$. The prediction error is $Y^{oos} - (\hat{\beta}_1 X_1^{oos} + \ldots + \hat{\beta}_k X_k^{oos}) = u^{oos} - [(\hat{\beta}_1 - \beta_1)X_1^{oos} + \ldots + (\hat{\beta}_k - \beta_k)X_k^{oos}]$, where the final expression obtains using Equation (14.2), and u^{oos} is the value of the error u for the out-of-sample observation. Because u^{oos} is independent of the data used to estimate the coefficients and is uncorrelated with X^{oos}, the MSPE in Equation (14.1) for the standardized predictive regression model can be written as the sum of two components:

$$\text{MSPE} = \sigma_u^2 + E[(\hat{\beta}_1 - \beta_1)X_1^{oos} + \ldots + (\hat{\beta}_k - \beta_k)X_k^{oos}]^2. \qquad (14.3)$$

The first term in Equation (14.3), σ_u^2, is the variance of the oracle prediction error—that is, of the prediction error made using the true (unknown) conditional mean, $E(Y|X)$.

The second term in Equation (14.3) is the contribution to the prediction error arising from the estimated regression coefficients. This second term represents the cost, measured in terms of increased mean squared prediction error, of needing to estimate the coefficients instead of using the oracle prediction.

Because the mean square is the sum of the variance and the square of the bias (Equation (2.33)), the second term in Equation (14.3) is the sum of the variance of the prediction arising from estimating β and the squared bias of the prediction. When it comes to determining which estimator to use, the goal is to make this second term in Equation (14.3) as small as possible. As we shall see, when there are many predictors, this entails trading off the bias of the estimated coefficients against their variance.

Standardization using the sample means and variances. In practice, the population means and standard deviations of the original variables are not known. Accordingly, the in-sample means and variances are used to standardize the regressors, and the in-sample mean is subtracted from the dependent variable.

Because the regressors are standardized and the dependent variable is demeaned, an additional step is needed to produce the prediction for an out-of-sample observation. Specifically, the out-of-sample observation on the predictors must be standardized using the in-sample mean and standard deviation, and the in-sample mean of

the dependent variable must be added back into the prediction. Formulas are given in Appendix 14.5.

The MSPE of OLS and the Principle of Shrinkage

In the special case that the regression error u in Equation (14.2) is homoskedastic, the MSPE of OLS is given by

$$\text{MSPE}_{\text{OLS}} \cong \left(1 + \frac{k}{n}\right)\sigma_u^2. \tag{14.4}$$

The approximation in Equation (14.4) holds exactly in some special cases (Exercise 14.12), and it holds more generally as an approximation when n is large and k/n is small. In the case of a single regressor, Equation (14.4) is derived in Appendix 14.2. The derivation of Equation (14.4) for general k uses matrix algebra and is given in Appendix 19.7.

This expression has a simple interpretation. As discussed following Equation (14.3), the MSPE of the oracle prediction—that is, the prediction using the true value of β—is σ_u^2. When the k regression coefficients are estimated by OLS, the MSPE increases by the factor $(1 + k/n)$ relative to the best-possible MSPE. Thus the cost, as measured by the MSPE, of using OLS depends on the ratio of the number of regressors to the sample size.

For example, in the school test score application, suppose the 38 main regressors in Table 14.1 are used to predict test scores. Although 38 regressors sounds like a lot, $k/n = 38/1966 \approx 0.02$, so using OLS entails only a 2% loss in MSPE relative to the oracle prediction. In many applications, a loss of 2% might not be important. In the data set with 817 regressors, however, $k/n = 817/1966 \cong 0.40$, and a 40% deterioration is large enough that it is worth investigating estimators that have a lower MSPE than OLS.

Because OLS is unbiased under the prediction interpretation of Equation (14.2), the inflation factor $(1 + k/n)$ arises solely from the variance of the OLS estimator. Under the Gauss–Markov conditions, the OLS estimator has the smallest variance of all linear unbiased estimators. As a result, one might naturally be discouraged about making much headway when k/n is large. But a major conceptual breakthrough in the many-predictor problem, dating to the early 1960s, was the discovery that if one allows for biased estimators, the estimator variance can be reduced by so much that the MSPE can be less than that of OLS.

The principle of shrinkage. A **shrinkage estimator** introduces bias by "shrinking" the OLS estimator toward a specific number and thereby reducing the variance of the estimator. Because the mean squared error is the sum of the variance and the squared bias (Equation (2.33)), if the estimator variance is reduced by enough, then the decrease in the variance can more than compensate for the increase in the squared bias. The result is an estimator with a lower mean squared error than OLS.

James and Stein (1961) developed the first estimator that achieved this goal of reducing the estimator mean squared error by introducing bias. When the regressors are uncorrelated, the James–Stein estimator can be written as $\widetilde{\beta}^{JS} = c\hat{\beta}$, where $\hat{\beta}$ is the OLS estimator and c is a factor that is less than 1 and depends on the data. Because c is less than 1, the James–Stein estimator shrinks the OLS estimator toward 0 and thus is biased toward 0. It is not surprising that the James–Stein estimator has a lower mean squared error than the OLS estimator when the true β's are small. What James and Stein showed, however, is that if the errors are normally distributed, their estimator has a lower mean squared error than the OLS estimator, *regardless* of the true value of β, as long as $k \geq 3$.

James and Stein's remarkable result is the foundation of many-predictor methods used with big data. Their result leads to the family of shrinkage estimators, which includes ridge regression and the Lasso estimator, the topics of Sections 14.3 and 14.4, respectively.

Estimation of the MSPE

The MSPE is a population expectation and thus is unknown. However, it can be estimated from a sample of data. Here, we discuss two ways to estimate the MSPE. The first, split-sample estimation, draws directly on the definition of the MSPE and entails dividing the sample into two subsamples, one for estimation and one for prediction. The second, called m-fold cross validation, extends this idea but uses the data symmetrically and more efficiently by dividing the sample into m subsamples.

Estimating the MSPE using a split sample. Recall that the MSPE is the variance of the prediction error for a randomly drawn X, where the observation is not used to estimate β. This definition suggests estimating the MSPE by dividing the data set into two parts: an estimation subsample and a "test" subsample used to simulate out-of-sample prediction. The estimation subsample is used to estimate β, yielding the estimate $\widetilde{\beta}$, which could be obtained by OLS or some other estimator. This estimate is then used to make a prediction $\hat{Y}$ for each of the n_{test} observations in the test subsample. The MSPE is then estimated using the resulting n_{test} prediction errors:

$$\widehat{\text{MSPE}}_{split\text{-}sample} = \frac{1}{n_{test}} \sum_{\substack{\text{observations in} \\ \text{test subsample}}} (Y_i - \hat{Y}_i)^2. \tag{14.5}$$

Estimating the MSPE by m-fold cross validation. The split-sample procedure treats the data asymmetrically by arbitrarily splitting the observations into two subsamples that are then used for different purposes. This estimator can be improved by treating the data symmetrically. Specifically, the two subsamples can be used to produce two different estimators of the MSPE by swapping which subsample is used to estimate β and which is used to estimate the MSPE.

This idea extends to m different, randomly chosen subsamples. The resulting procedure is called m-fold cross validation. In ***m*-fold cross validation**, there are m separate estimates of the MSPE, each produced by sequentially leaving out one of

m-fold Cross Validation

KEY CONCEPT 14.1

The *m*-fold cross-validation estimator of the MSPE is determined according to the following six steps.

1. Divide the test sample into m randomly chosen subsets of approximately equal size.
2. Use the combined subsamples $2, 3, \ldots, m$ to compute $\widetilde{\beta}$, an estimate of β.
3. Use $\widetilde{\beta}$ and Equation (14.12) to compute predicted values $\hat{Y}$ and prediction errors $Y - \hat{Y}$ for the observations in subsample 1.
4. Using subsample 1 as the test sample, estimate the MSPE with the predicted values in subsample 1 and Equation (14.5); call this estimate $\widehat{\text{MSPE}}_1$.
5. Repeat steps 2–4 using subsample 2 as the left-out test sample, then subsample 3, and so forth, yielding a total of m estimates $\widehat{\text{MSPE}}_i, i = 1, \ldots, m$.
6. The *m*-fold cross-validation estimator of the MSPE is then estimated by averaging these m subsample estimates of the MSPE:

$$\widehat{\text{MSPE}}_{\text{m-fold cross validation}} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{n_i}{n/m}\right)\widehat{\text{MSPE}}_i, \tag{14.6}$$

where n_i is the number of observations in subsample i and the factor in parentheses allows for different numbers of observations in the different subsamples.

the m subsamples when estimating β and using that reserved subsample to estimate the MSPE. The *m*-fold cross-validation estimator of the MSPE is the average of the m subset estimators of the MSPE. The *m*-fold cross-validation estimator of the MSPE is summarized in Key Concept 14.1.

A loose end in *m*-fold cross validation is how to choose m. This involves a trade-off. A larger value of m produces more efficient estimators of β because more observations are used each time β is estimated. From this perspective, ideally one would use the so-called leave-one-out cross-validation estimator, for which $m = n - 1$. But a larger value of m means that β must be estimated m times. In applications in which k is large (in the hundreds or more), this can take considerable computer time, and leave-one-out cross validation takes too long computationally. As a result, the choice of m must be made taking into account practical constraints on your and your computer's time. In the school test score application in this chapter, we settle on $m = 10$ as a practical compromise given the computer we used, so that each subsample estimator of β uses 90% of the sample.

The *m*-fold cross-validation estimator can be used to estimate the MSPE in very general settings, regardless of how β is estimated. It even works for models that can be expressed only as algorithms, not in terms of parameters. This general applicability makes it widely used in empirical work with big data.

14.3 Ridge Regression

Sections 14.3 and 14.4 describe two shrinkage estimators that are designed for use with many predictors. The method discussed in this section, ridge regression, shrinks the estimated parameter to 0 by adding to the sum of squared residuals a penalty that increases with the square of the estimated parameter. By minimizing the sum of these two terms, which is called the penalized sum of squared residuals, ridge regression introduces bias into the estimator but reduces its variance. In some applications, ridge regression can result in large improvements in MSPE compared to OLS.

Shrinkage via Penalization and Ridge Regression

One way to shrink the estimated coefficients toward 0 is to penalize large values of the estimate. The ridge regression estimator is based on this idea. Specifically, the **ridge regression** estimator minimizes the penalized sum of squares, which is the sum of squared residuals plus a penalty factor that increases with the sum of the squared coefficients:

$$S^{Ridge}(b; \lambda_{Ridge}) = \sum_{i=1}^{n} (Y_i - b_1 X_{1i} - \ldots - b_k X_{ki})^2 + \lambda_{Ridge} \sum_{j=1}^{k} b_j^2, \quad (14.7)$$

where $\lambda_{Ridge} \geq 0$. The parameter λ_{Ridge} is called the ridge shrinkage parameter. The ridge regression estimator is the value of b that minimizes $S^{Ridge}(b; \lambda_{Ridge})$.

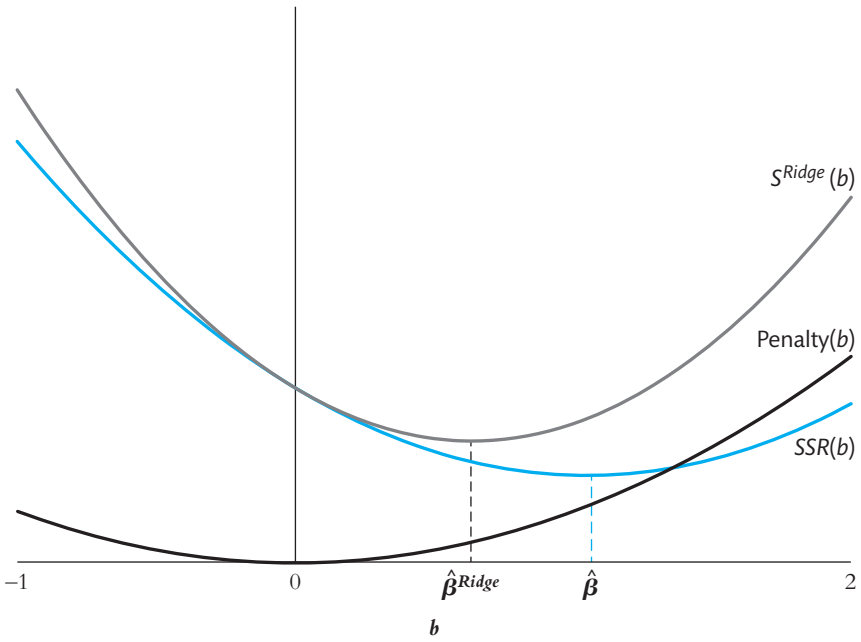
The first term on the right-hand side of Equation (14.7) is the usual sum of squared residuals for a trial coefficient value b. If this were the only term, then the ridge and OLS estimators would be the same. The second term, however, is new. This second term increases with the sum of the squared coefficients. This second term in Equation (14.7) is called a **penalty term** because it penalizes the estimator for choosing a large estimate of the coefficient. When the penalty term is scaled by the shrinkage parameter and added to the sum of squared residuals, as it is in Equation (14.7), the result is called the **penalized sum of squared residuals**.

The penalty term shrinks the ridge regression estimator toward 0. Figure 14.1 shows how ridge penalization works when there is only one regressor. Without the penalty, one would minimize the sum of squared residuals, which yields the OLS estimator. Adding in the penalty shifts the minimum of the penalized function toward 0. Thus the estimated ridge coefficient will be closer to 0 than the OLS estimator is; that is, the ridge regression estimator is shrunk toward 0.

The magnitude of the shrinkage depends on the shrinkage parameter λ_{Ridge}. If $\lambda_{Ridge} = 0$, there is no shrinkage, and the ridge regression estimator equals the OLS estimator. The larger λ_{Ridge}, the greater the penalty for a given value of b, and the greater the shrinkage of the estimator toward 0. Because we are using the standardized predictive regression model, all the coefficients have the same units, so a single shrinkage parameter λ_{Ridge} can be used for all the coefficients.

FIGURE 14.1 Components of the Ridge Regression Penalty Function

The ridge regression estimator minimizes $S^{Ridge}(b)$, which is the sum of squared residuals, $SSR(b)$, plus a penalty that increases with the square of the estimated parameter. The SSR is minimized at the OLS estimator, $\hat{\beta}$. Including the penalty shrinks the ridge estimator, $\hat{\beta}^{Ridge}$, toward 0.



The penalized sum of squared residuals in Equation (14.7) can be minimized using calculus to give a simple expression for the ridge regression estimator. This formula is derived in Appendix 14.3 for the case of a single regressor. When $k > 2$, the formula is best expressed using matrix notation, and it is given in Appendix 19.7.

In the special case that the regressors are uncorrelated, the ridge regression estimator is

$$\hat{\beta}_j^{Ridge} = \left(\frac{1}{1 + \lambda_{Ridge}/\sum_{i=1}^{n} X_{ji}^2}\right)\hat{\beta}_j, \tag{14.8}$$

where $\hat{\beta}_j$ is the OLS estimator of β_j. In this case, the ridge regression estimator shrinks the OLS estimator toward 0, like the James–Stein estimator. When the regressors are correlated, the ridge regression estimate can sometimes be greater than the OLS estimate although overall the ridge regression estimates are shrunk towards zero.

When there is perfect multicollinearity, such as when $k > n$, the OLS estimator can no longer be computed, but the ridge estimator can.

Estimation of the Ridge Shrinkage Parameter by Cross Validation

The ridge regression estimator depends on the shrinkage parameter λ_{Ridge}. While the value of λ_{Ridge} could be chosen arbitrarily, a better strategy is to pick λ_{Ridge} so that the ridge regression estimator works well for the data at hand.

One might initially think that the shrinkage parameter λ_{Ridge} could be estimated by minimizing $S^{Ridge}(b; \lambda_{Ridge})$ in Equation (14.7). However, for any trial value of b, minimizing $S^{Ridge}(b; \lambda_{Ridge})$ with respect to λ_{Ridge} simply leads to setting λ_{Ridge} to 0; but when $\lambda_{Ridge} = 0$, the ridge regression estimator is just the OLS estimator! The reason that this approach yields the OLS estimator is that it provides the best *in-sample* fit—which is given by OLS. In contrast, the goal of prediction is to have a good *out-of-sample* fit—that is, a low MSPE.

That insight suggests choosing λ_{Ridge} to minimize the estimated MSPE. This strategy can be implemented using the m-fold cross-validation estimator of the MSPE (Key Concept 14.1). Specifically, suppose you have two candidate values of λ_{Ridge}—for example, 0.1 and 0.2—and choose some value of m. Let $\widetilde{\beta}$ in Key Concept 14.1 denote the ridge regression estimator using $\lambda_{Ridge} = 0.1$. Given $\widetilde{\beta}$, compute the predictions in the test sample, and use those predictions to compute $\widehat{MSPE}$ for that estimator. Now repeat, but use $\lambda_{Ridge} = 0.2$. You now have two estimates of the MSPE, one for $\lambda_{Ridge} = 0.1$ and one for $\lambda_{Ridge} = 0.2$, so choose the value of λ_{Ridge} that provides the lowest estimated MSPE. Repeating these steps for multiple values of λ_{Ridge} yields an estimator of λ_{Ridge} that minimizes the m-fold cross-validation MSPE. Although this estimator could potentially be 0—so that the best ridge estimator is the OLS estimator—typically the best shrinkage parameter will not be 0 and the ridge estimator will differ from the OLS estimator.

Application to School Test Scores

We illustrate the use of ridge regression by fitting a predictive model for school test scores using the 817 predictors in Table 14.1 with 1966 observations.

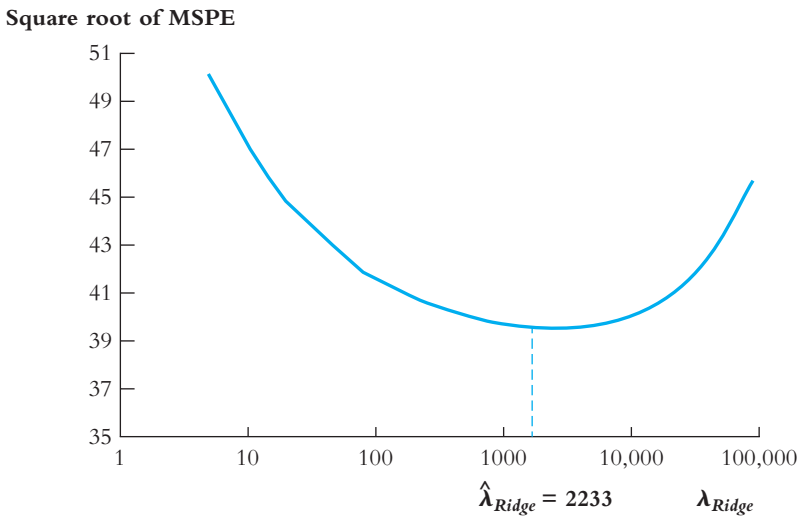
Figure 14.2 plots the square root of the 10-fold cross-validation estimator of the MSPE as a function of the ridge shrinkage parameter λ_{Ridge}. The square root of the MSPE is plotted so that it provides an estimate of the magnitude of a typical out-of-sample prediction error. For a given value of λ_{Ridge}, the MSPE was computed as described in Key Concept 14.1. The choice of $m = 10$ represents a practical balance between the desire to use as many observations as possible to estimate the parameters and the computational burden of repeating that estimation m times for each value of λ_{Ridge}.

As Figure 14.2 shows, the MSPE has a U shape. It is minimized at $\lambda_{Ridge} = 2233$, so the 10-fold cross-validation estimate of the ridge shrinkage parameter is $\hat{\lambda}_{Ridge} = 2233$.

The square root of the MSPE, evaluated at $\hat{\lambda}_{Ridge}$, is 39.5. In contrast, the root MSPE for OLS, estimated using the same 817 predictors and 1966 observations, is much larger, 78.2. Because the OLS estimator is the ridge estimator with $\lambda_{Ridge} = 0$, in principle the root MSPE of the OLS estimator could also be shown in Figure 14.2 as the point ($\lambda_{Ridge} = 0$, root MSPE $= 78.2$); however, the root MSPE for OLS is so large that it is off the scale of the figure.

FIGURE 14.2 Square Root of the MSPE for the Ridge Regression Prediction as a Function of the Shrinkage Parameter (Log Scale for λ_{Ridge})

The MSPE is estimated using 10-fold cross validation for the school test score data set with $k = 817$ predictors and $n = 1966$ observations.



The fact that the OLS MSPE is much larger than the ridge MSPE provides an empirical demonstration of the main theoretical point discussed in Section 14.2: When there are many predictors, introducing bias into the parameter estimates via shrinkage can reduce the variance of the prediction by more than enough to compensate for the bias and therefore produce much more accurate predictions.

Because $\hat{\lambda}_{Ridge}$ is chosen to minimize the cross-validated MSPE, the cross-validated MSPE evaluated at $\hat{\lambda}_{Ridge}$ is no longer an unbiased estimator of the MSPE. In Section 14.6, we use the remaining 1966 observations (not used so far) to obtain an unbiased estimator of the MSPE for ridge regression using $\hat{\lambda}_{Ridge}$.

It is also of interest to compare the ridge regression coefficients to the OLS coefficients. That comparison is made in Section 14.6, where these coefficients are also compared to the methods discussed in Sections 14.4 and 14.5, the Lasso and principal components, respectively.

14.4 The Lasso

In OLS and ridge regression, none of the estimated coefficients is exactly 0 so all the regressors are used to make the prediction. In some applications, however, only a few predictors might be useful, with the rest irrelevant. For example, among the predictors in Table 14.1, all but 38 are constructed as squares, cubes, or interactions of the 38 main variables; if the true conditional expectation is, in fact, linear in the 38 main variables, then $817 - 38 = 779$ of the variables would have a coefficient of 0.

A regression model in which the coefficients are nonzero for only a small fraction of the predictors is called a **sparse model**. If the model is sparse, predictions can be improved by estimating many of the coefficients to be *exactly* 0.

The estimator examined in this section, the Lasso (least absolute shrinkage and selection operator), is designed for sparse models. Like ridge regression, the Lasso shrinks estimated coefficients to 0. Unlike ridge regression, it sets many of the estimated coefficients exactly to 0, thereby dropping those regressors from the model. Moreover, the regressors it keeps are subject to less shrinkage than with ridge regression. Thus, the Lasso provides a way to select a subset of the regressors and then estimate their coefficients with a modest amount of shrinkage.

Like ridge regression, the Lasso can be used when $k > n$. Also like ridge regression, the Lasso has a shrinkage parameter that can be estimated by minimizing the cross-validated MSPE.

Shrinkage Using the Lasso

The **Lasso** estimator minimizes a penalized sum of squares, where the penalty increases with the sum of the absolute values of the coefficients:

$$S^{Lasso}(b; \lambda_{Lasso}) = \sum_{i=1}^{n} (Y_i - b_1 X_{1i} - \ldots - b_k X_{ki})^2 + \lambda_{Lasso} \sum_{j=1}^{k} |b_j|, \quad (14.9)$$

where λ_{Lasso} is called the Lasso shrinkage parameter. The Lasso estimator is the value of b that minimizes $S^{Lasso}(b; \lambda_{Lasso})$. As with ridge regression, if the shrinkage parameter $\lambda_{Lasso} = 0$, the Lasso estimator minimizes the sum of squared residuals in which case the Lasso is just OLS. The second term in Equation (14.9) penalizes large values of b and thus shrinks the Lasso estimate toward 0.[2]

The first part of the Lasso name—least absolute shrinkage—reflects the nature of the penalty term in Equation (14.9). Whereas the ridge regression penalty increases with the square of b, the Lasso penalty increases with its absolute value.

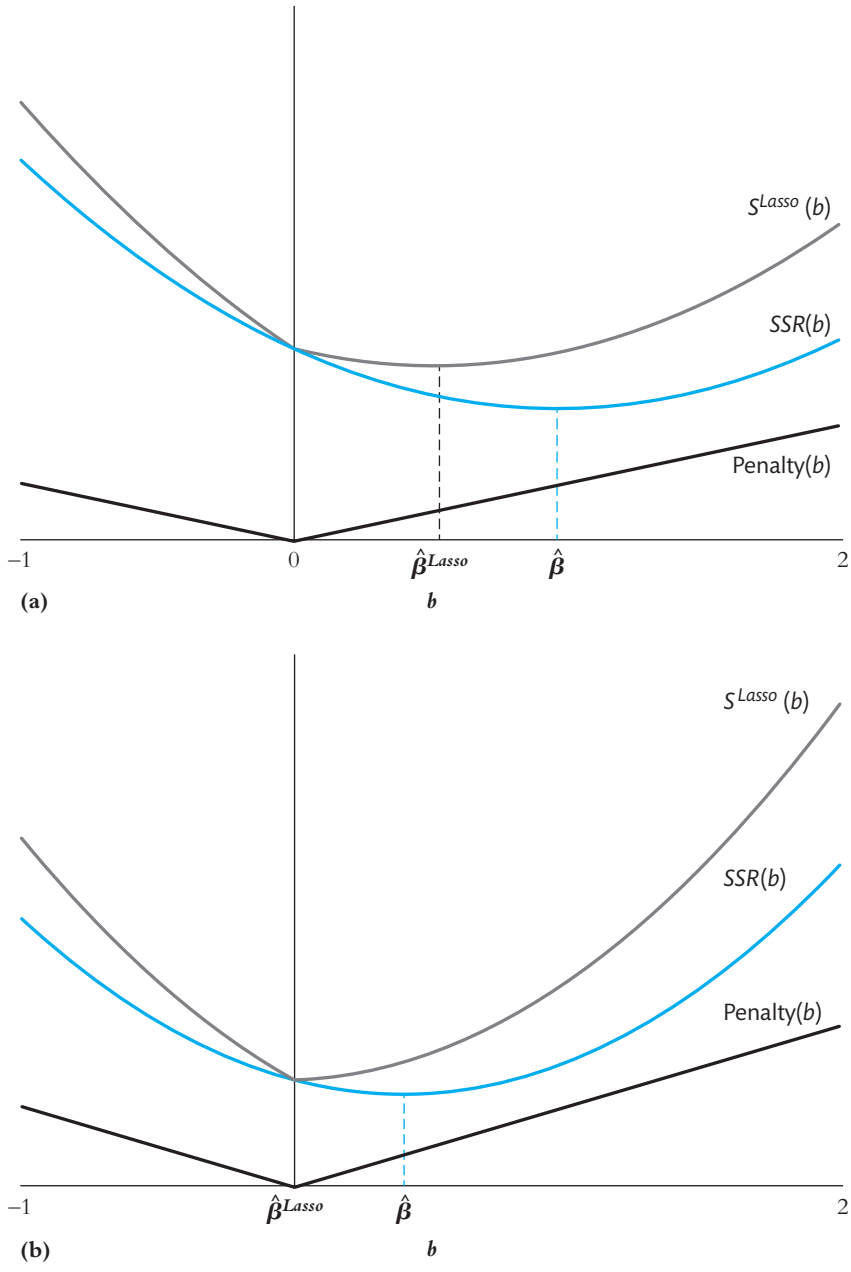
The second part of the Lasso name—selection operator—arises because the Lasso estimates many coefficients to be exactly 0, thereby dropping some of the predictors. Thus the Lasso, in effect, selects a subset of the predictors to be used in the model.

The reason that the Lasso estimates some coefficients to be exactly 0 is illustrated in Figure 14.3 for $k = 1$. This figure shows the sum of squared residuals, the Lasso penalty, and the combined Lasso minimization function in Equation (14.9). Parts a and b of Figure 14.3 differ only in the value of the OLS estimate, which minimizes the first term in Equation (14.9). In Figure 14.3a, the OLS estimate is far from

[2]The ridge and Lasso penalty terms can both be written as $\lambda \sum_{j=1}^{k} |b_j|^p$, where $p = 2$ for ridge and $p = 1$ for Lasso. The expression $(\sum_{j=1}^{k} |b_j|^p)^{1/p}$ is called the L_p length of b, where $p = 2$ corresponds to the usual Euclidean distance. As a result, the ridge is sometimes called L_2 penalization, and the Lasso is sometimes called L_1 penalization.

FIGURE 14.3 The Lasso Estimator Minimizes the Sum of Squared Residuals Plus a Penalty That Is Linear in the Absolute Value of *b*

For a single regressor, (a) when the OLS estimator is far from zero, the Lasso estimator shrinks it toward 0; (b) when the OLS estimator is close to 0, the Lasso estimator becomes exactly 0.



0 ($\hat{\beta} = 1.0$), and the Lasso shrinks it to a smaller value ($\hat{\beta}^{Lasso} = 0.5$). In Figure 14.3b, the curve representing the sum of squared residuals is shifted to the left, so the OLS estimate is smaller ($\hat{\beta} = 0.4$), and the Lasso estimate is exactly 0 ($\hat{\beta}^{Lasso} = 0$). This estimate of exactly 0 arises because the sum of squared residuals function in Figure 14.3b is so flat near 0 that the penalty term takes over from the sum of squared residuals and drives the estimate to 0.

Appendix 14.4 provides a formula for the Lasso estimator when $k = 1$. The formula shows mathematically that for sufficiently small values of the OLS estimator, the Lasso estimator is exactly 0.

The ridge and Lasso estimators also behave differently when the OLS estimate is large. For large values of b, the ridge penalty exceeds the Lasso penalty. Thus, when the OLS estimate is large, the Lasso shrinks it less than ridge, but when the OLS estimate is small, the Lasso shrinks it more than ridge—in some cases, all the way to 0.

Figure 14.3 considers the case of a single regressor, for which the Lasso always shrinks the OLS estimator toward 0. If there are multiple predictors, then the Lasso generally shrinks the OLS estimates toward 0; however, it is possible that the Lasso estimate of some of the coefficients could be larger than the OLS estimate.

Computation of the Lasso estimator. Unlike OLS and ridge regression, there is no simple expression for the Lasso estimator when $k > 1$, so the Lasso minimization problem must be done numerically using a computer. One of the many computational advances in machine learning is the development of specialized algorithms to compute the Lasso estimator. Some econometric software packages incorporate these algorithms and make it straightforward to use the Lasso estimator.

Estimation of the shrinkage parameter by cross validation. As in ridge regression, the Lasso tuning parameter can be estimated by minimizing an estimate of the MSPE. The algorithm for estimating λ_{Lasso} is the same as that laid out in Section 14.3 for estimating λ_{Ridge}.

A word of warning about the ridge and Lasso estimators. The ridge and Lasso estimators differ from all the other estimators used in this text in an important way. In OLS, the fit of the regression model is the same whether one uses the k original regressors or k linear combinations of the regressors as long as one avoids perfect multicollinearity. For example, one can use an intercept and a dummy variable for *male*, or an intercept and a dummy variable for *female*, or both a *male* dummy and a *female* dummy and no intercept; all yield identical fits of the OLS regression and identical predictions. Moreover, which of these three specifications is used makes no difference for the other estimated coefficients in the model.

In contrast, with ridge and Lasso the regression fit, the estimated coefficients, and the predictions in general depend on the specific choice of the linear combination of regressors used. This is easiest to see for the Lasso because the population values of the coefficients change as you change linear combinations. For example, the

coefficient on male in the (intercept, *male*) specification differs from that in the (*female*, *male*) specification. Thus the Lasso might drop *male* from the (intercept, *male*) specification but not from the (*female*, *male*) specification. If so, the (intercept, *male*) and (*female*, *male*) specifications would have different selected predictors and thus would make different predictions.

The reason that the choice of linear combinations matters for ridge is more subtle and stems from the fact that different linear combinations will have different correlations with each other. An explanation of this result for ridge regression is given in Appendix 19.7.

The dependence of the ridge and Lasso estimators on the choice of linear combination of regressors implies that one needs to put thought into choosing the regressors when using these estimators—a decision that does not matter for OLS or for the principal components method of Section 14.5 (or, for that matter, for logit, probit, or IV regression).

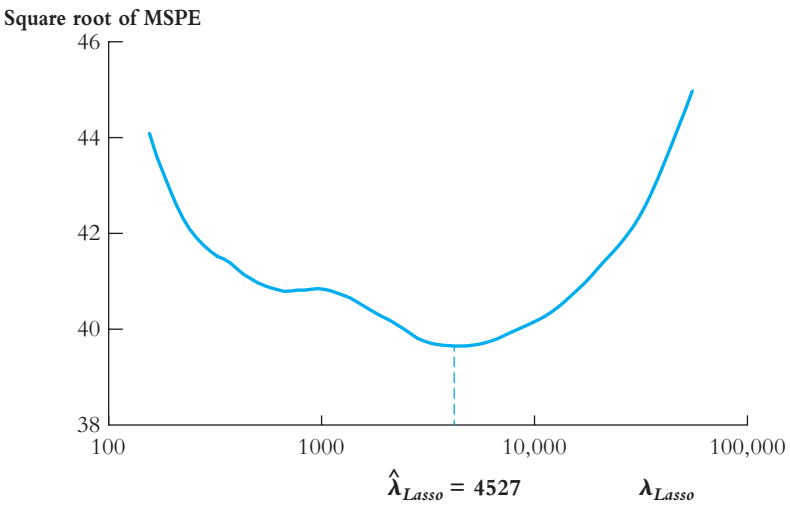
Application to School Test Scores

We now turn to estimation of a Lasso prediction model using the same 817 regressors and 1966 observations as in Section 14.3.

Figure 14.4 plots the square root of the 10-fold cross-validation estimate of the MSPE as a function of the Lasso shrinkage parameter λ_{Lasso}. The MSPE is minimized when the shrinkage parameter is 4527, so $\hat{\lambda}_{Lasso} = 4527$. At this estimated value of λ_{Lasso}, the MSPE is 39.7. This MSPE is much less than the MSPE of OLS, 78.2, which is equivalent to the Lasso estimator for $\lambda_{Lasso} = 0$. The Lasso MSPE is close to, but slightly greater than, the minimized ridge MSPE of 39.5 (from Section 14.3).

FIGURE 14.4 Square Root of the MSPE for the Lasso Prediction as a Function of the Lasso Shrinkage Parameter (Log Scale for λ_{Lasso})

The MSPE is estimated by 10-fold cross validation using the school test score data set with $k = 817$ predictors and $n = 1966$ observations.



The Lasso estimates nonzero coefficients on only 56 of the 817 predictors; thus the Lasso estimator excludes 761, or 93%, of the candidate predictors in Table 14.1. Of the retained predictors, all but 4 are interactions among the 38 main predictors in Table 14.1.

14.5 Principal Components

When the regressors are perfectly collinear, at least one of them can be dropped from the data set without any loss of information because the dropped regressor can be perfectly reconstructed from the retained regressors. This observation suggests that there might be little loss of information from dropping a variable that is highly, but imperfectly, correlated with the other regressors. This insight forms the basis for an alternative strategy for handling many predictors: Exploit the correlations among the regressors to reduce the number of regressors while retaining as much of the information in the original regressors as possible. Principal components analysis implements this strategy and can reduce sharply the number of regressors so that estimation and prediction can proceed using OLS.

This section begins by showing how principal components analysis works when there are two regressors. We then turn to the more relevant case when the number of regressors is large.

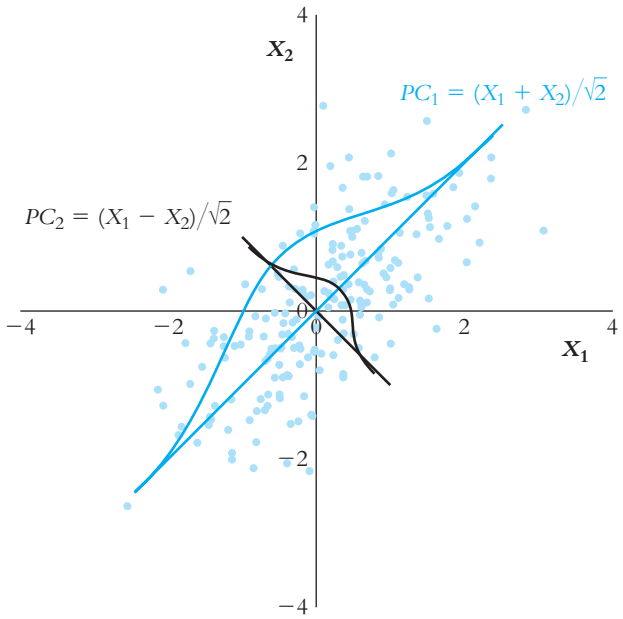
Principal Components with Two Variables

The **principal components** of a set of standardized variables X are linear combinations of those variables, where the linear combinations are chosen so that the principal components are mutually uncorrelated and sequentially contain as much of the information in the original variables as possible. Specifically, the linear combination weights for the first principal component are chosen to maximize its variance, in this sense capturing as much of the variation of the X's as possible. The linear combination weights for the second principal component are chosen so that it is uncorrelated with the first principal component and captures as much of the variance of the X's as possible, controlling for the first principal component. The third principal component is uncorrelated with the first two and captures as much of the variance of the X's as possible, controlling for the first two principal components, and so forth. If $k \leq n$ and there is no perfect multicollinearity, then the total number of principal components is k. If $k > n$, then the total number of principal components is n.

It is easiest to see how this procedure works when there are two X's. Figure 14.5 illustrates this case when X_1 and X_2 are standard normal random variables with a correlation of 0.7. The first principal component is the weighted average, $PC_1 = w_1X_1 + w_2X_2$, with the maximum variance, where w_1 and w_2 are the principal component weights. Choosing the weights corresponds to choosing a direction in which to add the variables or, equivalently, choosing a direction in which the spread

FIGURE 14.5 Scatterplot of 200 Observations on Two Standard Normal Random Variables, X_1 and X_2, with Population Correlation 0.7

The first principal component (PC_1) maximizes the variance of the linear combination of these variables, which is done by adding X_1 and X_2. The second principal component (PC_2) is uncorrelated with the first and is obtained by subtracting the two variables. The principal component weights are normalized so that the sum of squared weights adds to 1.



of the variables is greatest. As Figure 14.5 illustrates, the spread of the variables is greatest in the direction of the 45° line. Along this direction, the variables are added together with equal weights.

Without further restrictions, the variance of the linear combination can always be increased simply by increasing both w_1 and w_2. Thus, for the principal components problem to have a solution, it is necessary to restrict the weights. This is done by requiring the sum of squared weights to equal 1; that is, $w_1^2 + w_2^2 = 1$. Along the 45° line, the weights are equal, so $w_1 = w_2 = 1/\sqrt{2}$ and $PC_1 = (X_1 + X_2)/\sqrt{2}$, a result derived mathematically in Exercise 14.11.

The second principal component is chosen to be uncorrelated with the first principal component, and the sum of its squared weights also equals 1. When there are two variables, these two requirements imply that $PC_2 = (X_1 - X_2)/\sqrt{2}$. This corresponds to adding the variables along the downward-sloping 45° line in Figure 14.5. As illustrated in the figure, the spread of the variables is minimized in this direction. Thus, when there are only two variables, the first principal component maximizes the variance of the linear combination, while the second principal component minimizes the variance of the linear combination.

The variances of the two principal components are $\text{var}(PC_1) = 1 + |\rho|$ and $\text{var}(PC_2) = 1 - |\rho|$, where $\rho = \text{corr}(X_1, X_2)$ (Exercise 14.11). These expressions confirm that if the variables are correlated, PC_1 has a greater variance than PC_2.

These expressions for the variances of PC_1 and PC_2 have another, more subtle feature: $\text{var}(PC_1) + \text{var}(PC_2) = \text{var}(X_1) + \text{var}(X_2)$.[3] This provides an R^2 interpretation of principal components: The fraction of the total variance explained by the first principal component is $\text{var}(PC_1)/[\text{var}(X_1) + \text{var}(X_2)]$, and the fraction explained by the second is $\text{var}(PC_2)/[\text{var}(X_1) + \text{var}(X_2)]$. Together, the two principal components explain all the variance of X. For the two variables in Figure 14.5, the correlation is 0.7, so the first principal component explains $(1 + \rho)/2 = 85\%$ of the variance of X, while the second principal component explains the remaining 15% of the variance of X.

If there are only two variables, there is little reason to reduce their number using principal components. The utility of principal components arises when there are many correlated variables, in which case much or most of the variation in those variables can be captured by a smaller number of principal components.

Principal Components with k Variables

The principal components of the k variables $X_1, \ldots, X_k$ are the linear combinations of those variables that are mutually uncorrelated, have squared weights that sum to 1, and maximize the variance of the linear combination controlling for the previous principal components. Assuming there is no perfect multicollinearity among the variables, the number of principal components of X is the minimum of n and k.

Expressions for the principal component weights for $k > 2$ are more complicated than when $k = 2$. Fortunately, there is a fast method for computing the principal components and their weights. Because this method entails matrix calculations, it is deferred to Appendix 19.7. This procedure for computing principal components is widely available in standard statistical software.

Principal components with k variables is summarized in Key Concept 14.2.

The scree plot. The equality in Equation (14.10) leads to a useful graph, known as a scree plot, for visualizing the amount of variation in X that is captured by the j^{th} principal component.

A **scree plot** is the plot of the sample variance of the j^{th} principal component relative to the total sample variance in the X's (that is, the sample value of $\text{var}(PC_j)/\sum_{j=1}^{k}\text{var}(X_j)$) against the number of the principal component, j. Because this ratio has the interpretation of the R^2 of the j^{th} principal component, the scree plot makes it possible to read off the fraction of the sample variance of the Xs explained by any particular principal component. Because the principal components are mutually uncorrelated, the cumulative sum of these ratios through the p^{th} principal component is the fraction of the total sample variance of X explained by the first p principal components.

Figure 14.6 is the scree plot for the first 50 principal components of the 817-variable data set in Table 14.1. The first principal component explains 18% of the

[3]For $k = 2$, this can be verified by adding the two expressions for the variances of the principal components: $\text{var}(PC_1) + \text{var}(PC_2) = (1 + |\rho|) + (1 - |\rho|) = 2 = \text{var}(X_1) + \text{var}(X_2)$, where the final equality follows because X_1 and X_2 are standardized and thus have unit variance.

The Principal Components of X

KEY CONCEPT
14.2

The principal components of the k variables $X_1, \ldots, X_k$ are the linear combinations of X that have the following properties:

(i) The squared weights of the linear combinations sum to 1;

(ii) The first principal component maximizes the variance of its linear combination;

(iii) The second principal component maximizes the variance of its linear combination, subject to its being uncorrelated with the first principal component; and

(iv) More generally, the j^{th} principal component maximizes the variance of its linear combination, subject to its being uncorrelated with the first $j - 1$ principal components.

- Assuming there is no perfect multicollinearity in X, the number of principal components is the minimum of n and k.
- The sum of the sample variances of the principal components equals the sum of the sample variances of the X's:

$$\sum_{j=1}^{\min(n,k)} \text{var}(PC_j) = \sum_{j=1}^{k} \text{var}(X_j). \tag{14.10}$$

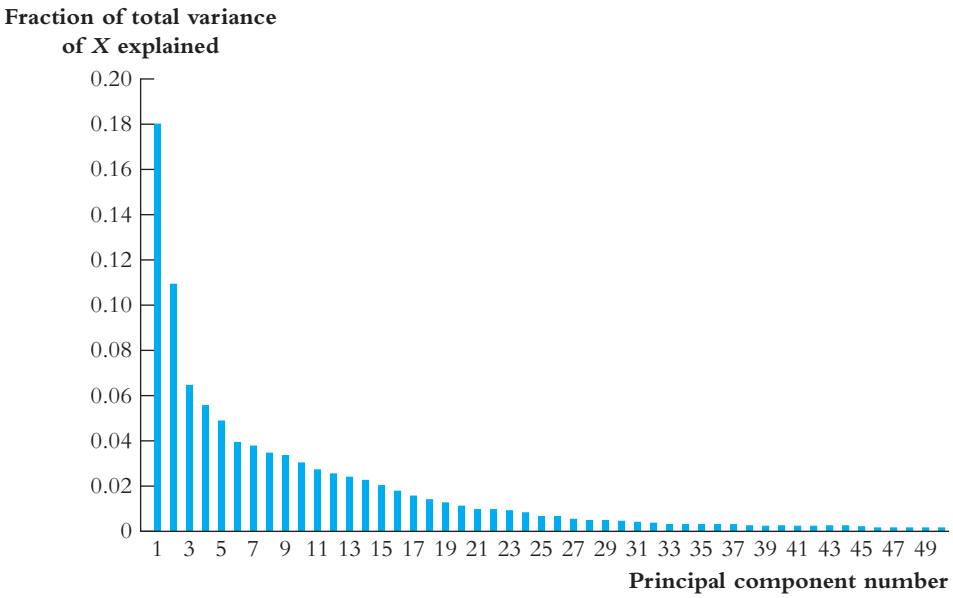
- The ratio $\text{var}(PC_j) / \sum_{j=1}^{k} \text{var}(X_j)$ is the fraction of the total variance of the X's explained by the j^{th} principal component. This measure is like an R^2 for the total variance of the X's.

total sample variance of the 817 X's, and the second principal component explains 11% of the total variance. Thus 29%, or more than one-fourth, of the total variance of the 817 variables is explained by just these two principal components. The first 10 principal components explain 63% of the total variance of the 817 X's, and the first 40 principal components explain 92% of the total variance.

The flattening in Figure 14.6 after the first few principal components is typical of many data sets in which the variables are highly correlated, as they are in the 817-variable school test score data set. This feature gives the scree plot its name: It looks like a cliff, with boulders, or scree, cascading into a valley.

Prediction using principal components. The fact that so much of the variation in the 817 predictors is captured by the first 10, or 50, principal components suggests that one could replace the 817 predictors with far fewer principal components and use

FIGURE 14.6 Scree Plot for the 817-Variable School Data Set (First 50 Principal Components)



Plotted values are the fraction of the total variance of the 817 regressors explained by the indicated principal component. The first principal component explains 18% of the total variance of the 817 X's, and the first 10 principal components together explain 63% of the total variance.

those principal components as regressors. With many fewer regressors, the coefficients can be estimated using OLS.

A key question is how many principal components p to include in the regression. Like the ridge and Lasso shrinkage parameters, the number of principal components p can be estimated by minimizing the MSPE, where the MSPE is estimated by m-fold cross validation.

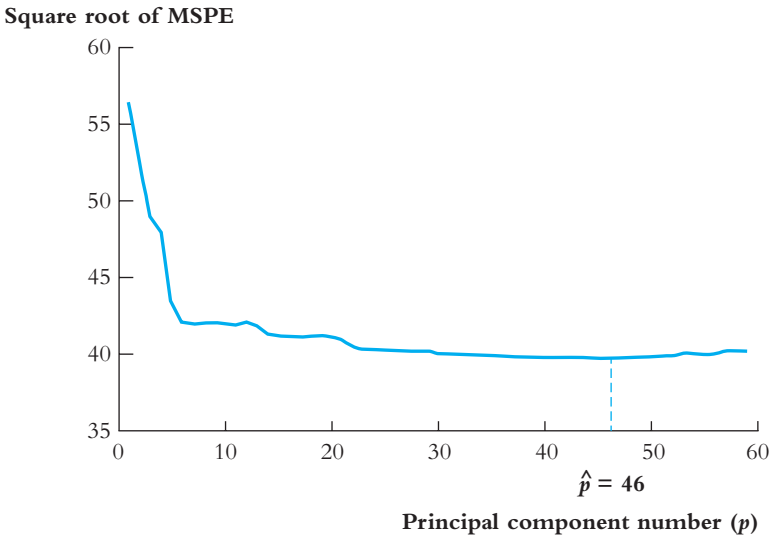
As discussed following Equation (14.3), computing the predicted value for an out-of-sample observation requires standardizing the observation using the in-sample mean and variance of each predictor. In the case of principal components regression, the out-of-sample values of the principal components must, in addition, be computed by applying the weights (the w's) estimated using in-sample data values to the standardized X's. The details are discussed in Appendix 14.5.

Application to School Test Scores

Figure 14.7 plots the square root of the 10-fold cross-validation estimate of the MSPE of the principal components predictor of school test scores as a function of the number p of principle components used as regressors; the principle components were computed using the same 817 predictors and 1966 observations as in Sections 14.3

FIGURE 14.7 Square Root of the MSPE for the Principal Components Prediction as a Function of the Number of Principal Components p Used as Predictors

The MSPE is estimated using 10-fold cross validation for the school test score data set with $k = 817$ predictors and $n = 1966$ observations.



and 14.4. Initially, increasing the number of principal components used as predictors results in a sharp decline in the MSPE. After $p = 5$ principal components, the improvement slows down, and after $p = 23$ principal components, the MSPE is essentially flat in the number of predictors. The MSPE is minimized at 46 predictors, so this is the cross-validation estimate of p; that is, $\hat{p} = 46$. Using 46 principal components, the MSPE is 39.7, the same as for Lasso and just slightly more than for ridge.

14.6 Predicting School Test Scores with Many Predictors

Do the many-predictor methods improve upon test score predictions made using OLS with a small data set and, if so, how do the many-predictor methods compare? To find out, we predict school test scores using small ($k = 4$), large ($k = 817$), and very large ($k = 2065$) data sets. For the small data set, the predictions are made using OLS. For the other data sets, they are made using OLS, ridge regression, the Lasso, and principal components.

As was stressed in Section 14.2, the predictive performance that matters is performance out of sample. Because the m-fold MSPE is used to estimate the ridge and Lasso shrinkage parameters and the number of included principal components p, the MSPE no longer provides a true out-of-sample comparison among the prediction methods. We therefore have reserved half the observations for assessing the performance of the estimated models; we call these remaining observations the reserved test sample.

Specifically, we use the following procedure, explained here for ridge regression, to assess predictive performance. Using the 1966 observations in the estimation sample, we estimate the shrinkage parameter λ_{Ridge} by 10-fold cross validation; for the 817-predictor data set, this yields the estimate $\hat{\lambda}_{Ridge}$ reported in Section 14.3. Using this estimated shrinkage parameter, the ridge regression coefficients are reestimated using all 1966 observations in the estimation sample. Those estimated coefficients are then used to predict the out-of-sample values Y^{*oos} for all the observations in the reserved test sample. Analogous procedures are used for the Lasso and principal components.

Table 14.2 lists the three sets of predictors. The 4 predictors in the small set are similar to some regressors in Chapters 5–9 for the district-level test score regressions. The 817 predictors are those in Table 14.1. The very large set augments the 38 main variables in Table 14.1 with demographic data on residents in the neighborhood of the school (age distribution, sex, marital status, education, and immigrant status), as well as some binary descriptors of the school and district, for a total of 65 main variables. For the very large data set, these 65 main variables are further augmented by all interactions, squares, and cubes, for a total of 2065 predictors—more than the number of observations (1966) in the estimation sample!

TABLE 14.2 The Three Sets of Predictors, School Test Score Data Set

Small ($k = 4$)
School-level data on Student–teacher ratio
Median income of the local population
Teachers' average years of experience
Instructional expenditures per student
Large ($k = 817$) The full data set in Table 14.1
Very Large ($k = 2065$)
The main variables are those in Table 14.1, augmented with the 27 variables below, for a total of 65 main variables, 5 of which are binary: Population Age distribution variables in local population (8) Fraction of local population that is male Local population marital status variables (3) Local population educational level variables (4) Fraction of local housing that is owner occupied Immigration status variables (4) Charter school (binary) School has full-year calendar (binary) School is in a unified school district (large city) (binary) School is in Los Angeles (binary) School is in San Diego (binary)
+ Squares and cubes of the 60 nonbinary variables $(60 + 60)$
+ All interactions of the nonbinary variables $(60 \times 59/2 = 1770)$
+ All interactions between the binary variables and the nonbinary demographic variables $(5 \times 22 = 110)$
Total number of variables $= 65 + 60 + 60 + 1770 + 110 = 2065$

TABLE 14.3 Out-of-Sample Performance of Predictive Models for School Test Scores

Predictor Set	OLS	Ridge Regression	Lasso	Principal Components
Small ($k = 4$)				
Estimated λ or p	—	—	—	—
In-sample root MSPE	53.6	—	—	—
Out-of-sample root MSPE	52.9	—	—	—
Large ($k = 817$)				
Estimated λ or p	—	2233	4527	46
In-sample root MSPE	78.2	39.5	39.7	39.7
Out-of-sample root MSPE	64.4	38.9	39.1	39.5
Very large ($k = 2065$)				
Estimated λ or p	—	3362	4221	69
In-sample root MSPE	—	39.2	39.2	39.6
Out-of-sample root MSPE	—	39.0	39.1	39.6

Notes: The in-sample MSPE is the 10-fold cross-validation estimate computed using the 1966 observations in the estimation sample. For the many-predictor methods, the shrinkage parameter or p was estimated by minimizing this in-sample MSPE. The out-of-sample MSPE is a split-sample estimate, computed with the 1966 observations in the reserved test sample and using the model estimated from the full estimation sample.

The results of this comparison are summarized in Table 14.3. Four features stand out. First, the MSPE of OLS is much less using the small data set than using the large data set (OLS cannot be computed in the very large data set because $k > n$). When there are many regressors, OLS is unable to use the additional information to improve out-of-sample prediction.

Second, for the many-predictor methods, there are substantial gains from increasing the number of predictors from 4 to 817, with the square root of the MSPE falling by roughly one-fourth. There are no further gains, however, from going to the very large set of regressors.

Third, the in-sample estimates of MSPE (the 10-fold cross-validation estimates) are similar to the out-of-sample estimates. In fact, the out-of-sample MSPEs are slightly less than the in-sample MSPEs. There are two reasons for this surprising result. First, the 10-fold MSPE uses only 90% of the data for estimating the coefficients at any one time (that is, $0.9 \times 1966 = 1769$ observations), whereas the coefficients used for the out-of-sample estimate of the MSPE are estimated using all

TABLE 14.4 Coefficients on Selected Standardized Regressors, 4- and 817-Variable Data Sets

Predictor	$k = 4$	$k = 817$			
	OLS	OLS	Ridge Regression	Lasso	Principal Components
Student–teacher ratio	4.51	118.03	0.31	0	0.25
Median income of the local population	34.46	−21.73	0.38	0	0.30
Teachers' average years of experience	1.00	−79.59	−0.11	0	−0.17
Instructional expenditures per student	0.54	−1020.77	0.11	0	0.19
Student–teacher ratio × Instruction expenditures per student		−89.79	0.72	2.31	0.84
Student–teacher ratio × Fraction of English learners		−81.66	−0.87	−5.09	−0.55
Free or reduced-price lunch × Index of part-time teachers		29.42	−0.92	−8.17	−0.95

Notes: The index of part-time teachers measures the fraction of teachers who work part-time. For OLS, ridge, and Lasso, the coefficients in Table 14.4 are produced directly by the estimation algorithms. For principal components, the coefficients in Table 14.4 are computed from the principal component regression coefficients (the γ's in Equation ((14.13)), combined with the principal component weights. The formula for the β coefficients for principal components is presented using matrix algebra in Appendix 19.7.

1966 observations in the estimation sample. As a result, those latter coefficient estimates are more precise. Second, there is random sampling variation in both estimates. The more general point is that the in-sample 10-fold MSPEs provide a good guide to the out-of-sample MSPE.

Fourth, the MSPE in the reserved test sample is generally similar for all the many-predictor methods. This is not always the case; it just happens to be so in this application. For these data, ridge regression has a slight edge, and the lowest out-of-sample MSPE is obtained using ridge in the large data set.

Table 14.4 lists the coefficients on 7 of the variables in the 817-predictor data set; 4 of the 7 are those in the small data set. Although none of these coefficients has a causal interpretation, comparing them across the different methods and data sets gives insights into how the various methods work. Because the regressors are standardized, all the coefficients have the same units, points on the test per standard deviation of the original predictor.[4]

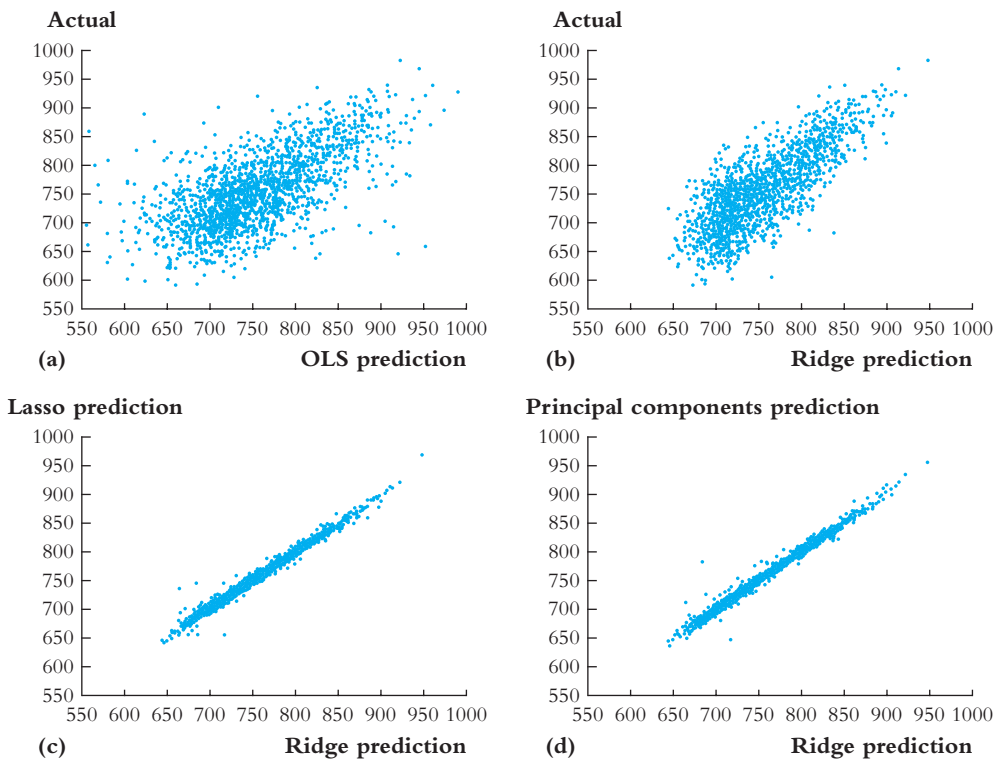
Table 14.4 has several striking features. For the small model, the magnitudes of the coefficients accord with the findings of Chapter 9 using the district-level data; for example, a one-standard-deviation greater value of median income predicts a

[4]The coefficients on the X's in the principal components column are obtained by combining the two steps of prediction using principal components. Specifically, the principal components are linear combinations of the X's, and the principal components regression model is a linear combination of the principal components. Thus the prediction can be written as a linear combination of the X's, where the weights involve both the principal components weights and the regression weights. The relevant formulas are given in Appendix 19.7.

34-point higher score on the test (the standard deviation of the test scores across schools is 64 points). In the large data set, however, many of the OLS coefficients are extremely large, and the pattern is erratic. With many regressors, OLS can fit individual observations by estimating large coefficients on specific variables, and this seems to be what is happening. This overfitting is why the predictive performance of OLS deteriorates moving from the small to the large data set. In contrast, the estimated coefficients for the many-predictor methods are substantially smaller and do not exhibit wild values. For the seven predictors in the table, the ridge and principal components coefficients are numerically similar. The Lasso coefficients, however, differ substantially from the ridge and principal components coefficients. Most notably, many of the Lasso coefficients (92% in all) are 0, including the coefficients on the four variables in the small data set. For the three coefficients in the table that are nonzero, they have the same sign as the ridge and principal components coefficients but are much larger, an empirical illustration of the tendency of Lasso to shrink more than ridge for small coefficients but to shrink less than ridge for large ones.

Another way to compare predictive models is to look at their predictions. Figure 14.8 shows scatterplots of the four sets of predictions for the 817-variable model, where the predictions are for the 1966 observations in the reserved test set.

FIGURE 14.8 Scatterplots for Out-of-Sample Predictions Using the 817-Predictor Data Set



(a) Actual versus OLS, (b) actual versus ridge, (c) Lasso versus ridge, and (d) principal components versus ridge.

Specifically, Figure 14.8a shows a scatterplot of the actual test scores versus the OLS predictions, and Figure 14.8b is the scatterplot of the actual test scores versus the ridge predictions. Figure 14.8c and Figure 14.8d are scatterplots of the Lasso versus the ridge predictions and the principal components versus the ridge predictions, respectively.

In Figure 14.8a and Figure 14.8b, the tighter the spread of the scatter along the 45° line, the better the prediction. Ridge has a tighter scatter than OLS, and it makes better out-of-sample predictions. (These scatterplots understate the improvement of ridge over OLS because some of the OLS predictions are outside the vertical scale of the plot.)

The clustering of the points along the 45° line in Figure 14.8c and Figure 14.8d indicate that the ridge, Lasso, and principal components predictions are generally quite similar. Still, one can see quite a few schools for which the predictions differ by at least 15 points, a substantial amount. Thus, while the three models have quite similar performance as measured by the MSPE (Table 14.3), for any given school the predictions can differ meaningfully.

The most important conclusion from this application is that for the large data set the many-predictor methods succeed where OLS fails. The reason for this success is that the many-predictor methods allow the coefficient estimates to be biased in a way that reduces their variance by enough to compensate for the increased bias. Another important conclusion is that the m-fold MSPE is close to the MSPE computed using the reserved test sample. One finding that does not generalize, however, is that the three methods happen to perform equally well in these data.

14.7 Conclusion

The coefficients in the predictive regression model do not have a causal interpretation. This does not matter, however, when the goal is prediction; the aim simply is to make out-of-sample predictions that are as accurate as possible, where accuracy is measured by the MSPE.

This chapter presented three methods for making predictions with many predictors. These methods provide different ways to overcome the poor performance of OLS predictions when the number of regressors is large relative to the sample size. The methods covered in this chapter—ridge regression, the Lasso, and principal components regression—all introduce bias into the estimator of the β's. However, this bias is introduced in a way that reduces the variance of the prediction by enough to yield a smaller MSPE.

Although ridge regression, the Lasso, and principal components regression all reduce variance by introducing bias, they do so in quite different ways. The Lasso sets many of the coefficients exactly to 0, in effect discarding those predictors. This approach works well when the oracle prediction model is sparse or approximately so. Principal components regression is most appropriate when the predictors, or groups of predictors, are highly correlated, in which case most of the variation in the regressors can be captured by a relatively small number of linear combinations of the

Text as Data

Text contains a lot of information! That is why you read the newspaper or posts on social media. That information keeps you abreast of political developments and helps you decide what to do tonight. By reading these sources, you use textual information—textual data—to make predictions about outcomes that are relevant to you.

A major accomplishment of statistics and machine learning is figuring out how to use computers to read text and to make predictions using textual data. At a conceptual level, it is a big leap to go from analyzing numbers to analyzing texts. The key step in doing so is turning text data into numerical data.

One way to turn text data into numerical data is to develop a list of words or phrases and then count the number of times that these words or phrases occur in a given text excerpt (for example, a newspaper article or blog post). These counts of words or phrases are numerical data that summarize the text. The unit of observation is the text excerpt, and the number of observations is the number of excerpts analyzed. This method of distilling a set of texts into occurrence counts of words or phrases was developed by Frederick Mosteller and David Wallace (1963) and is the basis of the field of stylometrics (see the box titled "Who Invented Instrumental Variables Regression?").

The approach of distilling text into counts of words or phrases has its own jargon. The list of words in a text is called a *bag of words*. The list of words and phrases of interest is called the *dictionary*. The dictionary may include only the words or phrases that are relevant to the prediction problem at hand, or it may contain all the words in the bag of words, excluding (for example) articles, pronouns, and conjunctions.

The word counts now can be used as predictors (X's) to predict a variable Y of interest. Thus, this bag-of-words approach has turned a seemingly intractable problem of combining text and numerical data into a regression problem.

Because the dictionary typically contains many words, the number of predictors can be large relative to the number of texts (n). If so, OLS would tend to produce poor predictions, but the methods in this chapter can be applied directly. For example, principal components analysis can be a useful tool in this setting because often words appear in groups (think of the words used in an article about a baseball game compared to an article about macroeconomic conditions). Putting all these pieces together results in predictive models that take text as the input and yield a prediction of Y.

variables—specifically, by their first few principal components. Because these principal components are relatively few in number, they can be the regressors used in a multiple regression model estimated by OLS. Ridge regression shrinks the OLS estimates toward 0 but does not rely on there being sparsity or on the regressors being highly correlated; thus it provides a useful approach when the regressors are not highly correlated and there is no a priori reason to assume sparsity. As it happens, in the school test score data, the three methods perform similarly, but this coincidence does not occur in general.

As discussed in Section 14.1, making predictions using many predictors that take on numerical values is only one of the opportunities provided by the methods of machine learning. For example, the box "Text as Data" describes how the

tools of this section can be used to analyze text data. Similarly, principal components analysis and its extensions can be used to turn images into numerical data, which then can be analyzed by the many-predictor methods described in this chapter. While many of the procedures in machine learning are new and the computational algorithms and tools are sophisticated, at their core are the key ideas of regression analysis, estimation, and testing that are at the heart of Parts I–III of this text.

The use of machine learning in economics is young, and many exciting applications await. For some examples and further reading, see Jean et. al. (2016) (predicting poverty using satellite imagery), Davis and Heller (2017) (examining treatment heterogeneity for a summer jobs program), and Kleinberg et. al. (2018) (application of machine learning to criminal sentencing).[5]

Summary

1. The goal of prediction is to make accurate predictions for out-of-sample observations—that is, for observations not used to estimate the prediction model.
2. The coefficients in prediction models do not have a causal interpretation.
3. One of the opportunities provided by big data sets is making predictions using many predictors. However, OLS works poorly for prediction when the number of regressors is large relative to the sample size.
4. The shortcomings of OLS can be overcome by using prediction methods that have lower variance at the cost of introducing estimator bias. These many-predictor methods can produce predictions with substantially better predictive performance than OLS, as measured by the MSPE.
5. Ridge regression and the Lasso are shrinkage estimators that minimize a penalized sum of squared residuals. The penalty introduces a cost to estimating large values of the regression coefficient. The weight on the penalty (the shrinkage parameter) can be estimated by minimizing the m-fold cross-validation estimator of the MSPE.
6. The principal components of a set of correlated variables capture most of the variation in those variables in a reduced number of linear combinations. Those principal components can be used in a predictive regression, and the number of principal components included can be estimated by minimizing the m-fold cross-validation MSPE.

[5]The field of machine learning is growing rapidly. A textbook introduction to this area, which is accessible to students after completing Parts I–III of this text, is Gareth James et al., *An Introduction to Statistical Learning* (2013).

Key Terms

mean squared prediction error (518)
oracle prediction (518)
first least squares assumption for prediction (519)
standardized predictive regression model (519)
shrinkage estimator (521)
m-fold cross validation (522)
ridge regression (524)
penalty term (524)
penalized sum of squared residuals (524)
sparse model (528)
Lasso (528)
principal components (532)
scree plot (534)

MyLab Economics Can Help You Get a Better Grade

MyLab Economics If your exam were tomorrow, would you be ready? For each chapter, **MyLab Economics** Practice Tests and Study Plan help you prepare for your exams. You can also find the Exercises and all Review the Concepts Questions available now in **MyLab Economics**. To see how it works, turn to the **MyLab Economics** spread on the inside front cover of this text and then go to **www.pearson.com/mylab/economics**.

For additional Empirical Exercises and Data Sets, log on to the Companion Website at **www.pearsonglobaleditions.com**.

Review the Concepts

14.1 Using data from a random sample of elementary schools, a researcher regresses average test scores on the fraction of students who qualify for reduced-price meals. The regression indicates a negative coefficient that is highly statistically significant and yields a high R^2. Is this regression useful for determining the causal effect of school meals on student test scores? Why or why not? Is this regression useful for predicting test scores? Why or why not?

14.2 Cross-validation uses in-sample observations. How does it estimate the MSPE for out-of-sample observations, even though the econometrician does not have those observations?

14.3 Regression coefficients estimated using shrinkage estimators are biased. Why might these biased estimators yield more accurate predictions than an unbiased estimator?

14.4 Ridge regression and Lasso are two regression estimators based on penalization. Explain how they are similar and how they differ.

14.5 Suppose a data set with 10 variables produces a scree plot that is flat. What does this tell you about the correlation of the variables? What does this suggest about the usefulness of using the first few principal components of these variables in a predictive regression?

Exercises

14.1 A researcher is interested in predicting average test scores for elementary schools in Arizona. She collects data on three variables from 200 randomly chosen Arizona elementary schools: average test scores (*TestScore*) on a standardized test, the fraction of students who qualify for reduced-priced meals (*RPM*), and the average years of teaching experience for the school's teachers (*TExp*). The table below shows the sample means and standard deviations from her sample.

Variable	Sample Mean	Sample Standard Deviation
TestScore	750.1	65.9
RPM	0.60	0.28
TExp	13.2	3.8

After standardizing *RPM* and *TEXP* and subtracting the sample mean from *TestScore*, she estimates the following regression:

$$\widehat{TestScore} = -48.7 \times RPM + 8.7 \times TExp, SER = 44.0$$

a. You are interested in using the estimated regression to predict average test scores for an out-of-sample school with $RPM = 0.52$ and $TEXP = 11.1$.

i. Compute the transformed (standardized) values of *RPM* and *TEXP* for this school; that is, compute the X^{oos} values from the X^{*oos} values, as discussed preceding Equation (14.2).

ii. Compute the predicted value of average test scores for this school.

b. The actual average test score for the school is 775.3. Compute the error for your prediction.

c. The regression shown above was estimated using the standardized regressors and the demeaned value of *TestScore*. Suppose the regression had been estimated using the raw data for *TestScore*, *RMP*, and *TExp*. Calculate the values of the regression intercept and slope coefficients for this regression.

d. Use the regression coefficients that you computed in (c) to predict average test scores for an out-of-sample school with $RPM = 0.52$ and $TExp = 11.1$. Verify that the prediction is identical to the prediction you computed in (a.ii).

14.2 A school principal is trying to raise funds so that all her students will receive reduced-price meals; currently, only 40% qualify for reduced-priced meals. Can she use the regression in Exercise 14.1 to estimate the effect of the new policy on test scores? Explain why or why not.

14.3 Describe the relationship, if any, between the standard error of a regression and the square root of the MSPE of the regression's out-of-sample predictions.

14.4 A large online retailer sells thousands of products. The retailer has detailed data on the products purchased by each of its customers. Explain how you would use these data to predict the next product purchased by a randomly selected customer.

14.5 Y is a random variable with mean $\mu = 2$ and variance $\sigma^2 = 25$.

a. Suppose you know the value of μ.

i. What is the best (lowest MSPE) prediction of the value of Y? That is, what is the oracle prediction of Y?

ii. What is the MSPE of this prediction?

b. Suppose you don't know the value of μ but you have access to a random sample of size $n = 10$ from the same population. Let $\overline{Y}$ denote the sample mean from this random sample. You predict the value of Y using $\overline{Y}$.

i. Show that the prediction error can be decomposed as $Y - \overline{Y} = (Y - \mu) - (\overline{Y} - \mu)$, where $(Y - \mu)$ is the prediction error of the oracle predictor and $(\mu - \overline{Y})$ is the error associated with using $\overline{Y}$ as an estimate of μ.

ii. Show that $(Y - \mu)$ has a mean of 0, that $(\overline{Y} - \mu)$ has a mean of 0, and that $Y - \overline{Y}$ has a mean of 0.

iii. Show that $(Y - \mu)$ and $(\overline{Y} - \mu)$ are uncorrelated.

iv. Show that the MSPE of $\overline{Y}$ is MSPE $= E(Y - \mu)^2 + E(\overline{Y} - \mu)^2 = \text{var}(Y) + \text{var}(\overline{Y})$.

v. Show that MSPE $= 25(1 + 1/10) = 27.5$.

14.6 In Exercise 14.5(b), suppose you predict Y using $\overline{Y}/2$ instead of $\overline{Y}$.

a. Compute the bias of the prediction.

b. Compute the mean of the prediction error.

c. Compute the variance of the prediction error.

d. Compute the MSPE of the prediction.

e. Does $\overline{Y}/2$ produce a prediction with a lower MSPE than the $\overline{Y}$ prediction?

f. Suppose $\mu = 10$ (instead of $\mu = 2$). Does $\overline{Y}/2$ produce a prediction with a lower MSPE than the $\overline{Y}$ prediction?

g. In a realistic setting, the value of μ is unknown. What advice would you give someone who is deciding between using $\overline{Y}$ and $\overline{Y}/2$?

14.7 In Exercise 14.5(b), suppose you predict Y using $\overline{Y} - 1$ instead of $\overline{Y}$.

a. Compute the bias of the prediction.

b. Compute the mean of the prediction error.

c. Compute the variance of the prediction error.

d. Compute the MSPE of the prediction.

e. Does $\overline{Y} - 1$ produce a prediction with a lower MSPE than the $\overline{Y}$ prediction?

f. Does $\overline{Y} - 1$ produce a prediction with a lower MSPE than the $\overline{Y}/2$ prediction from Exercise 14.6?

14.8 Let X and Y be two random variables. Denote the mean of Y given $X = x$ by $\mu(x)$ and the variance of Y by $\sigma^2(x)$.

a. Show that the best (minimum MSPE) prediction of Y given $X = x$ is $\mu(x)$ and the resulting MSPE is $\sigma^2(x)$. (*Hint:* Review Appendix 2.2.)

b. Suppose X is chosen at random. Use the result in (a) to show that the best prediction of Y is $\mu(X)$ and the resulting MSPE is $E[Y - \mu(X)]^2 = E[\sigma^2(X)]$.

14.9 You have a sample of size $n = 1$ with data $y_1 = 2$ and $x_1 = 1$. You are interested in the value of β in the regression $Y = X\beta + u$. (Note there is no intercept.)

a. Plot the sum of squared residuals $(y_1 - bx_1)^2$ as function of b.

b. Show that the least squares estimate of β is $\hat{\beta}^{OLS} = 2$.

c. Using $\lambda_{Ridge} = 1$, plot the ridge penalty term $\lambda_{Ridge}b^2$ as a function of b.

d. Using $\lambda_{Ridge} = 1$, plot the ridge penalized sum of squared residuals $(y_1 - bx_1)^2 + \lambda_{Ridge}b^2$.

e. Find the value of $\hat{\beta}^{Ridge}$.

f. Using $\lambda_{Ridge} = 0.5$, repeat (c) and (d). Find the value of $\hat{\beta}^{Ridge}$.

g. Using $\lambda_{Ridge} = 3$, repeat (c) and (d). Find the value of $\hat{\beta}^{Ridge}$.

h. Use the graphs that you produced in (a)–(d) for the various values of λ_{Ridge} to explain why a larger value of λ_{Ridge} results in more shrinkage of the OLS estimate.

14.10 You have a sample of size $n = 1$ with data $y_1 = 2$ and $x_1 = 1$. You are interested in the value of β in the regression $Y = X\beta + u$. (Note there is no intercept.)

a. Plot the sum of squared residuals $(y_1 - bx_1)^2$ as function of b.

b. Show that the least squares estimate of β is $\hat{\beta}^{OLS} = 2$.

c. Using $\lambda_{Lasso} = 1$, plot the Lasso penalty term $\lambda_{Lasso}|b|$ as a function of b.

d. Using $\lambda_{Lasso} = 1$, plot the Lasso penalized sum of squared residuals $(y_1 - bx_1)^2 + \lambda_{Lasso}|b|$.

e. Find the value of $\hat{\beta}^{Lasso}$.

f. Using $\lambda_{Lasso} = 0.5$, repeat (c) and (d). Find the value of $\hat{\beta}^{Lasso}$.

g. Using $\lambda_{Lasso} = 5$, repeat (c) and (d). Find the value of $\hat{\beta}^{Lasso}$.

h. Use the graphs that you produced in (a)–(d) for the various values of λ_{Lasso} to explain why a larger value of λ_{Lasso} results in more shrinkage of the OLS estimate.

14.11 Let X_1 and X_2 be two positively correlated random variables, both with variance 1.

a. (Requires calculus) The first principal component, PC_1, is the linear combination of X_1 and X_2 that maximizes $\text{var}(w_1X_1 + w_2X_2)$, where $w_1^2 + w_2^2 = 1$. Show that $PC_1 = (X_1 + X_2)/\sqrt{2}$. (*Hint:* First derive an expression for $\text{var}(w_1X_1 + w_2X_2)$ as a function of w_1 and w_2.)

b. The second principal component is $PC_2 = (X_1 - X_2)/\sqrt{2}$. Show that $\text{cov}(PC_1, PC_2) = 0$.

c. Show that $\text{var}(PC_1) = 1 + \rho$ and $\text{var}(PC_2) = 1 - \rho$, where $\rho = \text{cor}(x_1, x_2)$.

14.12 Consider the fixed-effects panel data model $Y_{jt} = \alpha_j + u_{jt}$ for $j = 1, \ldots, k$ and $t = 1, \ldots, T$. Assume that u_{jt} is i.i.d. across entities j and over time t with $E(u_{jt}) = 0$ and $\text{var}(u_{jt}) = \sigma_u^2$.

a. The OLS estimator of α_j is the value of a_j that makes the sum of squared residuals $\sum_{j=1}^{k}\sum_{t=1}^{T}(Y_{jt} - a_j)^2$ as small as possible. Show that the OLS estimator is $\hat{\alpha}_j = \overline{Y}_j = \frac{1}{T}\sum_{t=1}^{T} Y_{jt}$.

b. Show that

i. $\hat{\alpha}_j$ is an unbiased estimator of α_j.

ii. $\text{var}(\hat{\alpha}_j) = \sigma_u^2/T$.

iii. $\text{cov}(\hat{\alpha}_i, \hat{\alpha}_j) = 0$ for $i \neq j$.

c. You are interested in predicting an out-of-sample value for entity j—that is, for $Y_{j,T+1}$—and use $\hat{\alpha}_j$ as the predictor. Show that MSPE $= \sigma_u^2 + \sigma_u^2/T$.

d. You are interested in predicting an out-of-sample value for a randomly selected entity—that is, for $Y_{j,T+1}$, where j is selected at random. You again use $\hat{\alpha}_j$ as the predictor. Show the MSPE $= \sigma_u^2 + \sigma_u^2/T$.

e. The total number of in-sample observations is $n = kT$. Show that in both (c) and (d) MSPE $= \sigma_u^2(1 + k/n)$.

Empirical Exercises

E14.1 On the text website, **http://www.pearsonglobaleditions.com**, you will find a data set **CASchools_EE14_InSample** that contains a subset of $n = 500$ schools from the data set used in this chapter. Included are data on test scores and 20 of the primitive predictor variables; see **CASchools_EE141_Description** for a description of the variables. In this exercise, you will construct prediction models like those described in the text and use these models to predict test scores for 500 out-of-sample schools. (Please read **EE141_SoftwareNotes** on the text website before solving the exercise.)

a. From the 20 primitive predictors, construct squares of all the predictors, along with all of the interactions (that is, the cross products $X_{ji}X_{ki}$ for all j and k). Collect the 20 primitive predictors, their squares,

and all interactions into a set of k predictors. Verify that you have $20 + 20 + (20 \times 19)/2 = 230$ predictors. One of the primitive predictors is the binary variable *charter_s*. Drop the predictor $(charter\_s)^2$ from the list of 230 predictors, leaving 229 predictors for the analysis. Why should $(charter\_s)^2$ be dropped from the original list of predictors?

b. Compute the sample mean and standard deviation of each of the predictors, and use these to compute the standardized regressors. Compute the sample mean of *TestScore*, and subtract the sample mean from *TestScore* to compute its demeaned value.

c. Using OLS, regress the demeaned value of *TestScore* on the standardized regressors.

i. Did you include an intercept in the regression? Why or why not?

ii. Compute the standard error of the regression.

d. Using ridge regression with $\lambda_{Ridge} = 300$, regress the demeaned value of *TestScore* on the standardized regressors. Compare the OLS and ridge estimates of the standardized regression coefficients.

e. Using Lasso with $\lambda_{Lasso} = 1000$, regress the demeaned value of *TestScore* on the standardized regressors. How many of the estimated Lasso coefficients are different from 0? Which predictors have a nonzero coefficient.

f. Compute the scree plot for the 229 predictors. How much of the variance in the standardized regressors is captured by the first principal component? By the first two principal components? By the first 15 principal components?

g. Compute 15 principal components from the 229 predictors. Regress the demeaned value of *TestScore* on the 15 principal components.

h. On the text website, you will find a data set **CASchools_EE14_OutOfSample** that contains data from another $n = 500$ schools.

i. Predict the average test score for each of these 500 schools using the OLS, ridge, Lasso, and principal components prediction models that you estimated in (c), (d), (e), and (g). Compute the root mean square prediction error for each of the methods.

ii. Construct four scatter plots like those in Figure 14.8. What do you learn from the plots?

i. Estimate λ_{Ridge}, λ_{Lasso}, and the number of principal components using 10-fold cross validation from the in-sample data set.

j. Use the estimated values of λ_{Ridge}, λ_{Lasso}, and the number of principal components from (i) to construct predictions of test scores for the out-of-sample schools. Are these predictions more accurate than the predictions you computed in (h)? Is the difference in line with what you expected from the cross-validation calculations in (i)?

APPENDIX

14.1 The California School Test Score Data Set

The test scores used in this chapter are from the California Standards Tests (part of California's Standardized Testing and Reporting program) given to fifth-grade students in the spring of 2013. The average test score for each of California's schools is available from the California Department of Education, where you can also find much of the other school and district data used in the chapter. The remaining school and district data were obtained from ED-Data (www.ed-data.org). All school and district data are for the 2012–13 academic year. In addition to school and district data, demographic data for 2013 are constructed from the census tracts making up the zip code for each school. These data are available from the American Community Survey (see factfinder.census.gov). More detail is available in the replication files for the chapter at **http://www.pearsonglobaleditions.com.**

APPENDIX

14.2 Derivation of Equation (14.4) for $k = 1$

With a single regressor, the OLS prediction in the standardized predictive regression model (Equation (14.2)) for a given value $X = x$ is $\hat{Y}(x) = \hat{\beta}x$. The second term in Equation (14.3) is $E[(\hat{\beta} - \beta)X^{OOS}]^2 = E(\hat{\beta} - \beta)^2E(X^{OOS})^2 = E(\hat{\beta} - \beta)^2$, where the first equality uses the independence of $\hat{\beta}$ and X^{oos} ($\hat{\beta}$ is estimated using the in-sample data) and the second equality uses the fact that the regressors are standardized, so $E(X^{OOS})^2 = \text{var}(X^{OOS}) = 1$. Because the OLS estimator is unbiased in the prediction model, $E(\hat{\beta} - \beta)^2 = \text{var}(\hat{\beta}) = \sigma_u^2/(n\sigma_X^2) = \sigma_u^2/n$, where the second equality uses the large-n formula for the variance of the OLS estimator under homoskedasticity in Equation (5.27) and the final equality uses the fact that $\sigma_X^2 = 1$ because the regressors in Equation (14.2) are standardized using the population mean and variance. It follows from Equation (14.3) that, with $k = 1$ under homoskedasticity, the MSPE of OLS $\cong (1 + 1/n)\sigma_u^2$ for large n, which is Equation (14.4) with $k = 1$.

APPENDIX

14.3 The Ridge Regression Estimator When $k = 1$

When $k = 1$, the ridge estimator minimizes the penalized sum of squares, $S^{Ridge}(b; \lambda_{Ridge}) = \sum_{i=1}^n (Y_i - bX_i)^2 + \lambda_{Ridge}b^2$. Taking the derivative of $S^{Ridge}(b; \lambda_{Ridge})$ with respect to b and setting the derivative equal to 0 yields $-\sum_{i=1}^n X_i(Y_i - \hat{\beta}^{Ridge}X_i) + \lambda_{Ridge}\hat{\beta}^{Ridge} = 0$. Solving for $\hat{\beta}^{Ridge}$ yields $\hat{\beta}^{Ridge} = \sum_{i=1}^n X_iY_i/(\sum_{i=1}^n X_i^2 + \lambda_{Ridge}) = (1 + \lambda_{Ridge}/\sum_{i=1}^n X_i^2)^{-1}\hat{\beta}$, where $\hat{\beta} = \sum_{i=1}^n X_iY_i/\sum_{i=1}^n X_i^2$ is the OLS estimator.

APPENDIX

14.4 The Lasso Estimator When $k = 1$

When $k = 1$, the Lasso minimizes the penalized sum of squared residuals, $S^{Lasso}(b; \lambda_{Lasso}) = \sum_{i=1}^{n}(Y_i - bX_i)^2 + \lambda_{Lasso}|b|$. Inspection of Figure 14.3 shows that $\hat{\beta}$ and $\hat{\beta}^{Lasso}$ must have the same sign when $k = 1$. Suppose $\hat{\beta}$ is positive. Then, over the relevant range $b \geq 0$, the Lasso minimizes $\sum_{i=1}^{n}(Y_i - bX_i)^2 + \lambda_{Lasso}b$, and its derivative with respect to b is $-2\sum_{i=1}^{n}X_i\,(Y_i - bX_i) + \lambda_{Lasso}$. For $\hat{\beta}^{Lasso} > 0$, setting this derivative equal to 0 implies $-2\sum_{i=1}^{n}X_i\,(Y_i - \hat{\beta}^{Lasso}X_i) + \lambda_{Lasso} = 0$; otherwise, $\hat{\beta}^{Lasso} = 0$. Solving for $\hat{\beta}^{Lasso}$ yields

$$\hat{\beta}^{Lasso} = \max\left(\hat{\beta} - \tfrac{1}{2}\lambda_{Lasso}/\sum_{i=1}^{n}X_i^2, 0\right) \text{ when } \hat{\beta} \geq 0. \tag{14.11}$$

Similar reasoning shows that $\hat{\beta}^{Lasso} = \min\left(\hat{\beta} + \tfrac{1}{2}\lambda_{Lasso}/\sum_{i=1}^{n}X_i^2, 0\right)$ when $\hat{\beta} < 0$.

APPENDIX

14.5 Computing Out-of-Sample Predictions in the Standardized Regression Model

The estimators of this chapter are all computed using the standardized predictive regression model in Equation (14.2). Computing the prediction for an out-of-sample observation entails first standardizing the out-of-sample predictors, then computing the demeaned out-of-sample prediction, then adding back in the in-sample mean of Y. These transformations must all be done using the same means, variances, and weights for the out-of-sample data as for the in-sample data. Details are provided first for ridge regression and the Lasso, and then for principal components regression.

Out-of-Sample Predictions Using the Standardized Regression Model of Equation (14.2) (Ridge and Lasso)

Following Section 14.2, let $X_1^{*oos}, \ldots, X_k^{*oos}$ denote an out-of-sample observation on the original, untransformed values of the k predictors, and let Y^{*oos} denote the out-of-sample observation on the variable to be predicted. The transformed out-of-sample value of the j^{th} predictor is $X_j^{oos} = (X_j^{*oos} - \overline{X}_j^*)/s_{X_j^*}$, where $\overline{X}_j^*$ and $s_{X_j^*}$ are the in-sample mean and standard deviation of the j^{th} predictor. Let $\widetilde{\beta}_j$ be some estimator of β_j, e.g., the ridge regression or Lasso estimator. Then the predicted value of the original dependent variable in terms of the original predictors is

$$\hat{Y}^{*oos} = \overline{Y}^* + \sum_{j=1}^{k}\widetilde{\beta}_j\left(\frac{X_j^{*oos} - \overline{X}_j^*}{s_{X_j^*}}\right), \tag{14.12}$$

where $\overline{Y}^*, \overline{X}_j^*, s_{X_j^*}$, and $\widetilde{\beta}_j$ $(j = 1, \ldots, k)$ are all computed using the estimation sample.

Out-of-Sample Predictions Using Principal Components Regression

To compute the predicted value for an out-of-sample observation using principal components regression, it is necessary, in addition, to compute the out-of-sample values of the principal components using the in-sample weights. Let γ denote the coefficients in the regression of Y on the first p principal components:

$$Y_i = \gamma_1 PC_{1i} + \gamma_2 PC_{2i} + \ldots + \gamma_p PC_{pi} + v_i, \tag{14.13}$$

where v_i is an error term. The prediction of Y^{*oos} is computed in the following steps:

1. Compute the principal components in the estimation sample:
   a. Compute the demeaned Y and standardized X for the in-sample observations on Y^* and X^* as described preceding Equation (14.2).
   b. Compute the in-sample principal components of X; call these $PC_1, \ldots, PC_{min(n,k)}$.
2. Given p, estimate the regression coefficients in Equation (14.13); call these estimates $\hat{\gamma}_1^{PC}, \ldots, \hat{\gamma}_p^{PC}$.
3. Compute the out-of-sample values of the principal components:
   a. Standardize the out-of-sample predictors X^{*oos} using the in-sample mean and standard deviation from step 1(a). Denote this transformed observation as X^{oos}.
   b. Compute the principal components for the out-of-sample observation using the in-sample weights from step 1(b); call these $PC_1^{oos}, \ldots, PC_p^{oos}$.
4. Compute the predicted value for the out-of-sample observation as $\hat{Y}^{*oos} = \overline{Y}^* + \sum_{j=1}^{p} \hat{\gamma}_j^{PC} PC_j^{oos}$.