Empirical Exercise: Chapter 3

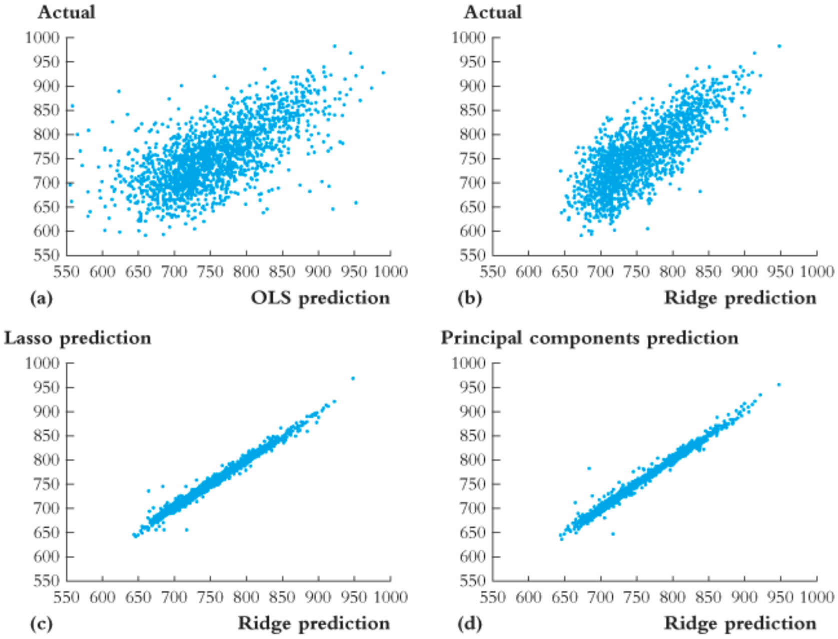
In the folder, you will find a data set **CASchools_EE14_InSample** that contains a subset of schools from the data set used in this chapter. Included are data on test scores and 20 of the primitive predictor variables; see **CASchools_EE141_Description** for a description of the variables. In this exercise, 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.

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 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. In the folder, 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[1]**. What do you learn from the plots?
i. Estimate $\lambda_{Ridge}, \lambda_{Lasso}$ and the number of principal components using 10-fold cross validation from the in-sample data set.
j. Use the estimated values of $\lambda_{Ridge}, \lambda_{Lasso}$ 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)?

Empirical Exercise: Chapter 3

1.

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.