Quantitative Social Science in the Age of Big Data and AI

Lecture 10: Variable Selection and Regularization

Zhaopeng Qu
Hopkins-Nanjing Center
May 22 2025

Review the last lecture

Review the last lecture

- Introduction to basic ideas of machine learning
  - Prediction vs Causal Inference
  - Supervised vs unsupervised learning
  - Regression and classification problems
- The loss function and the performance of the model
- The dangers of overfitting: the bias-variance trade off
- Spliting the data into training and testing sets
  - Using the training set to fit the model
  - Using the testing set to evaluate the performance of the model
- Cross-validation and model selection
- We use temperature data to predict the demand of the electricity for COAST area in Texas.

Linear regression using many predictors

Regression using many predictors

- Now we have many predictors Xs to predict one outcome Y, the basic model is still the same as before:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_p x_{p,i} + \varepsilon_i$$

- Then as the tricky problem we met in causal inference
  - How to choose the best predictors? As many as possible?
- Answer: NO
  - We need to balance the bias and the variance of the model to avoid overfitting.
- Two strategies:
  - Subset selection: choose a subset of predictors from the full set of predictors.
  - Shrinkage methods: shrink the coefficients of the predictors towards zero to reduce the variance of the model.

Subset selection

Subset selection

- In subset selection, we
  - Choose a subset of the p potential predictors (using some algorithm or professional judgement)
  - Estimate the chosen linear model using OLS
  - Use the model to make predictions
- Question: How to choose the subset of predictors?
- Several options:
  - Best subset selection: fit a model for every possible subset of predictors.
  - Forward stepwise selection: start with an intercept and add predictors one by one.
  - Backward stepwise selection: start with all predictors and remove predictors one by one.
  - Hybrid approaches: a combination of forward and backward selection.

Best subset selection

Best subset selection is based upon a simple idea: Estimate a model for every possible subset of variables; then compare their performances.

1. Define $\mathcal{M}_0$ as the model with no predictors.
2. For k in 1 to p:
    - Fit every possible model with k predictors.
    - Define $\mathcal{M}_k$ as the "best" model with k predictors.
3. Select the "best" model from $\mathcal{M}_0, \ldots, \mathcal{M}_p$.
4. Estimate cross-validated error for each $\mathcal{M}_k$.
5. Choose the $\mathcal{M}_k$ that minimizes the CV error.
6. Train the chosen model on the full dataset.

Best subset selection

- Question: So what's the problem? (Why do we need other selection methods?)
- Answer: "a model for **every possible subset**" can mean **a huge number** (2^p) of models.
- Example:
  - 10 predictors $\rightarrow$ 1,024 models to fit
  - 25 predictors $\rightarrow$ 33.5 million models to fit
  - 100 predictors $\rightarrow$ 1.5 trillion models to fit
- To avoid computational intensity, we need some other selection methods.

Credit data

```
str(ISLR::Credit)
```

```
#> 'data.frame':    400 obs. of  12 variables:
#>  $ ID       : int  1 2 3 4 5 6 7 8 9 10 ...
#>  $ Income   : num  14.9 106 104.6 148.9 55.9 ...
#>  $ Limit    : int  3606 6645 7075 9504 4897 8047 3388 7114 3300 6819 ...
#>  $ Rating   : int  283 483 514 681 357 569 259 512 266 491 ...
#>  $ Cards    : int  2 3 4 3 2 4 2 2 5 3 ...
#>  $ Age      : int  34 82 71 36 68 77 37 87 66 41 ...
#>  $ Education: int  11 15 11 11 16 10 12 9 13 19 ...
#>  $ Gender   : Factor w/ 2 levels " Male","Female": 1 2 1 2 1 1 2 1 2 2 ...
#>  $ Student  : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 1 1 2 ...
#>  $ Married  : Factor w/ 2 levels "No","Yes": 2 2 1 1 2 1 1 1 1 2 ...
#>  $ Ethnicity: Factor w/ 3 levels "African American",..: 3 2 2 2 3 3 1 2 3 1 ...
#>  $ Balance  : int  333 903 580 964 331 1151 203 872 279 1350 ...
```

- Outcomes: `balance`
- Predictors: `income`, `limit`, `rating`, `cards`, `age`, `education`, `gender`, `student`, `married`, `Ethnicity`

Credit data

ID	Income	Limit	Rating	Cards	Age	Education	Gender	Student	Married	Ethnicity
1	14.891	3606	283	2	34	11	Male	No	Yes	Caucasian
2	106.025	6645	483	3	82	15	Female	Yes	Yes	Asian
3	104.593	7075	514	4	71	11	Male	No	No	Asian
4	148.924	9504	681	3	36	11	Female	No	No	Asian
5	55.882	4897	357	2	68	16	Male	No	Yes	Caucasian
6	80.18	8047	569	4	77	10	Male	No	No	Caucasian
7	20.996	3388	259	2	37	12	Female	No	No	African American
8	71.408	7114	512	2	87	9	Male	No	No	Asian
9	15.125	3300	266	5	66	13	Female	No	No	Caucasian

Stepwise selection

- Stepwise selection provides a less computational intensive alternative to best subset selection.
- The basic idea behind stepwise selection
  - Start with an arbitrary model.
  - Try to find a "better" model by adding/removing variables.
  - Repeat.
  - Stop when you have the best model. (Or choose the best model.)
- The two most-common varieties of stepwise selection:
  - **Forward** starts with only intercept ($\mathcal{M}_0$) and adds variables
  - **Backward** starts with all variables ($\mathcal{M}_p$) and removes variables

Forward stepwise selection

The process...

1. Start with a model with only an intercept (no predictors), $\mathcal{M}_0$.
2. For $k = 0, \ldots, p$:
    - Estimate a model for each of the remaining $p - k$ predictors, separately adding the predictors to model $\mathcal{M}_k$.
    - Define $\mathcal{M}_{k+1}$ as the "best" model of the $p - k$ models.
3. Select the "best" model from $\mathcal{M}_0, \ldots, \mathcal{M}_p$.

- What do we mean by "best"?
- use cross-validation to choose minimized RMSE.

Forward stepwise selection with `caret` in R for the `Credit` dataset

N vars	RMSE	R2	MAE
1	231.86	0.750	175.9
2	163.30	0.878	122.3
3	104.59	0.951	84.7
4	102.18	0.953	82.1
5	100.20	0.955	79.5
6	100.44	0.955	80.1
7	100.87	0.955	80.2
8	101.16	0.954	80.6
9	101.25	0.954	80.6
10	101.66	0.954	81.1

Model selection

Backward stepwise selection

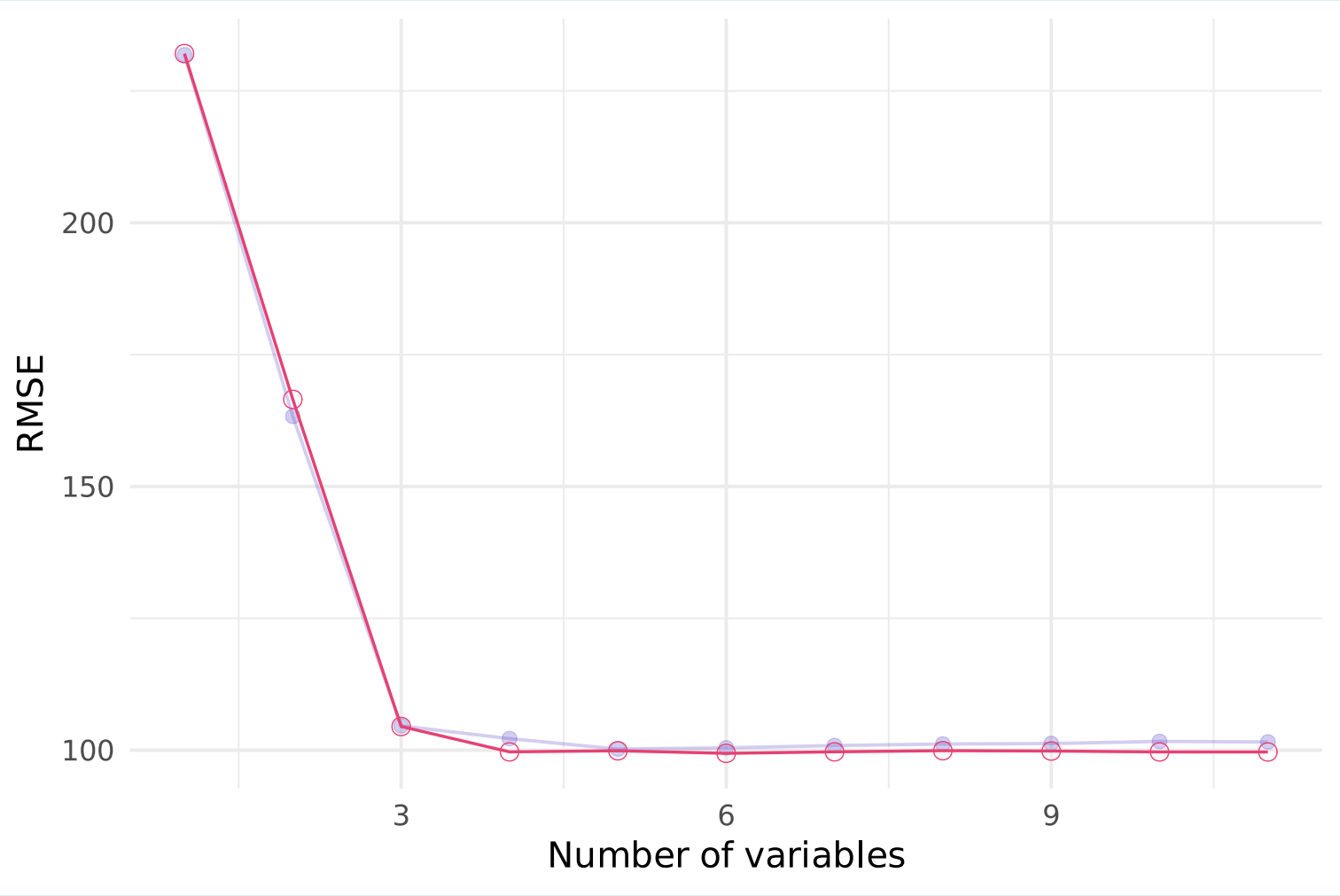
The process for backward stepwise selection is quite similar...

1. Start with a model that includes all p predictors: $\mathcal{M}_p$.
2. For $k = p, p-1, \ldots, 1$:
    - Estimate k models, where each model removes exactly one of the k predictors from $\mathcal{M}_k$.
    - Define $\mathcal{M}_{k-1}$ as the "best" of the k models.
3. Select the "best" model from $\mathcal{M}_0, \ldots, \mathcal{M}_p$.

Backward stepwise selection with `caret` in R for the `Credit` dataset

N vars	RMSE	R2	MAE
1	232.08	0.749	178.1
2	166.51	0.869	125.7
3	104.48	0.948	83.8
4	99.67	0.953	79.2
5	99.88	0.953	79.7
6	99.39	0.953	79.4
7	99.68	0.953	79.6
8	99.90	0.953	80.0
9	99.83	0.953	79.9
10	99.65	0.953	79.8

Note: **forward** and **backward** step. selection can choose different models.



Model selection

Stepwise selection

- **Less computationally intensive** (relative to best subset selection)
- With $p = 20$, BSS fits 1,048,576 models.
- With $p = 20$, foward/backward selection fits 211 models.
- However, there is **no guarantee** that stepwise selection finds the best model.
- ***Best*** is defined by your fit criterion (as always).
- Again, **cross validation is key** to avoiding overfitting.

Regularization

Regularization: some intuition

- The key to modern statistical learning is regularization: departing from optimality to stabilize a system.
- two common penalties
  - Ridge regression
  - Lasso regression

Prediction using many predictors for test scores

- Predicting test scores for a school using variable describing the school, its students, and its community.
- the full data set consists of data gathered on 3932 elementary schools in CA in 2013
- The task is to use these data to develop a prediction model that will provide good out-of-sample predictions.
- The variable to be predicted is the average fifth-grade test score at the school.

Application: 817 predictors

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
- Fraction of students eligible for free lunch
- Fraction of English learners
- Teachers' average years of experience
- Instructional expenditures per student
- Median income of the local population
- Student–teacher ratio
- Number of enrolled students
- Fraction of English-language proficient students
- Ethnic diversity index
- Ethnicity variables (8): fraction of students who are American Indian, Asian, Black, Filipino, Hispanic, Hawaiian, two or more, none reported
- Number of teachers
- Fraction of first-year teachers
- Fraction of second-year teachers
- Part-time ratio (number of teachers divided by teacher full-time equivalents)
- Per-student expenditure by category, district level (7)
- Per-student expenditure by type, district level (5)
- Per-student revenues by revenue source, district level (4)

+ Squares of main variables (38)

+ Cubes of main variables (38)

+ All interactions of main variables ($38 \times 37/2 = 703$)

Total number of predictors $= k = 38 + 38 + 38 + 703 = 817$

Ridge Regression

Ridge regression and OLS regression

- Recall in OLS regression, our model is

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + u_i$$

- Least-squares regression finds $\hat{\beta}_j$ by minimizing SSR

$$\begin{aligned} \min_{\hat{\beta}} \text{SSR} &= \min_{\hat{\beta}} \sum_{i=1}^{n} \hat{u}_i^2 \\ &= \min_{\hat{\beta}} \sum_{i=1}^{n} \left(Y_i - \underbrace{\left[\hat{\beta}_1 X_{i,1} + \cdots + \hat{\beta}_p X_{i,p} \right]}_{=\hat{Y}_i} \right)^2 \end{aligned}$$

- Ridge regression makes a small change
  - adds a .hi-purple[shrinkage penalty]: the sum of squared coefficents $\left(\lambda \sum_j \beta_j^2 \right)$
  - still minimizes the (weighted) sum of SSR and the shrinkage penalty.

Ridge regression and OLS regression

- The ridge regression estimator minimizes the penalized sum of squared residuals $SSR^{ridge}(b)$

$$\min_{\hat{\beta}} R^{ridge} = \min_{\hat{\beta}} \sum_{i=1}^{n} \left(Y_i - \underbrace{\left[\hat{\beta}_1 X_{i,1} + \cdots + \hat{\beta}_p X_{i,p} \right]}_{=\hat{y}_i} \right)^2 + \lambda_R \sum_{j=1}^{p} \beta_j^2$$

  - λ (≥ 0) is a tuning parameter for the harshness of the penalty.
  - $\lambda = 0$ implies no penalty: back to OLS.
- Each value of λ produces a new set of coefficients.
- Ridge's approach to the bias-variance tradeoff: Balance
  - reducing SSR, *i.e.*, $\sum_i \left(Y_i - \hat{Y}_i \right)^2$
  - reducing coefficients'magnitudes.
  - λ determines how much ridge "cares about" these two quantities.

Ridge regression estimator

- When $k = 1$, based on F.O.C, then

$$\frac{\partial SSR^{ridge}}{\partial b} = -2\sum_{i=1}^{n} X_i \left(Y_i - bX_i\right) + 2\lambda b = 0$$

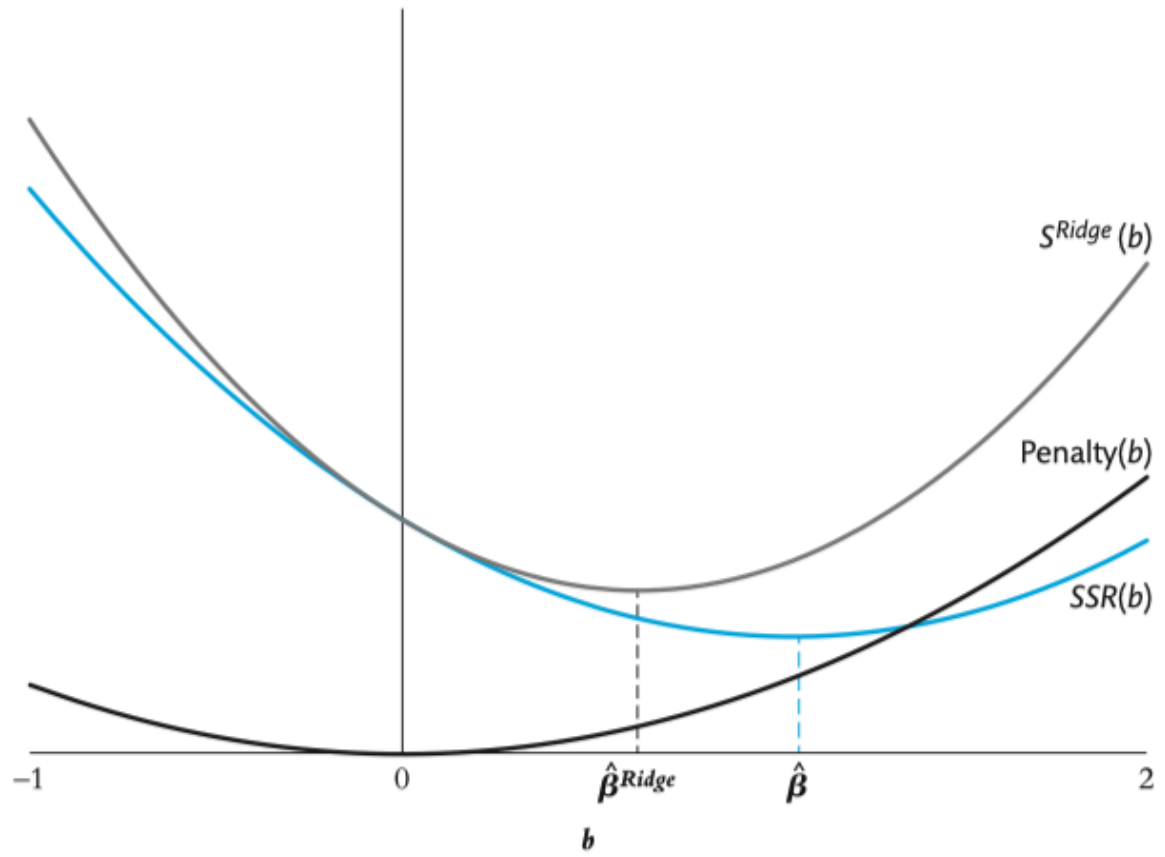
- Then we have Ridge regression estimator

$$\begin{aligned}
\hat{\beta}^{\text{Ridge}} &= \frac{\sum_{i=1}^{n} X_i Y_i}{\left(\sum_{i=1}^{n} X_i^2 + \lambda\right)} \\
&= \left(\frac{1}{1 + \lambda / \sum_{i=1}^{n} X_i^2}\right) \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \\
&\equiv \left(\frac{1}{1 + \lambda / \sum_{i=1}^{n} X_i^2}\right) \hat{\beta} < \hat{\beta}
\end{aligned}$$

Ridge regression estimator

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.



How to obtain the Parameter λ

- Question: Can we use the F.O.C to obtain it as we did for β as we did in OLS?
- Answer: No.
  - Then the optimal value of λ will be ZERO.
- Instead, it can be chosen by minimizing the m-fold cross-validated estimate of the RMSE.
  - Choose some value of λ , and estimate the MSPE by m-fold cross-validation
  - Repeat for many values of λ, and choose the one that yields the lowest RMSE.

Penalization and standardization

- Note: Scale of X can drastically affect ridge regression results.
- Because the scale of X will affect $\hat{\beta}$ and ridge is very sensitive to β.
  - Ridge regression pays a much larger penalty for different β
- Therefore, you have to standardize variables firstly before you use ridge regression, thus you have to standardize the predictors.

$$X_{i,j}^{*} = \frac{X_{i,j} - \bar{X}_j}{s_j}$$

- where $\bar{x}_j$ is the mean of x_j and s_j is the standard deviation of x_j.

Predicting School-level test scores

- λ estimated by minimizing the 10-fold cross-validated RMSE.

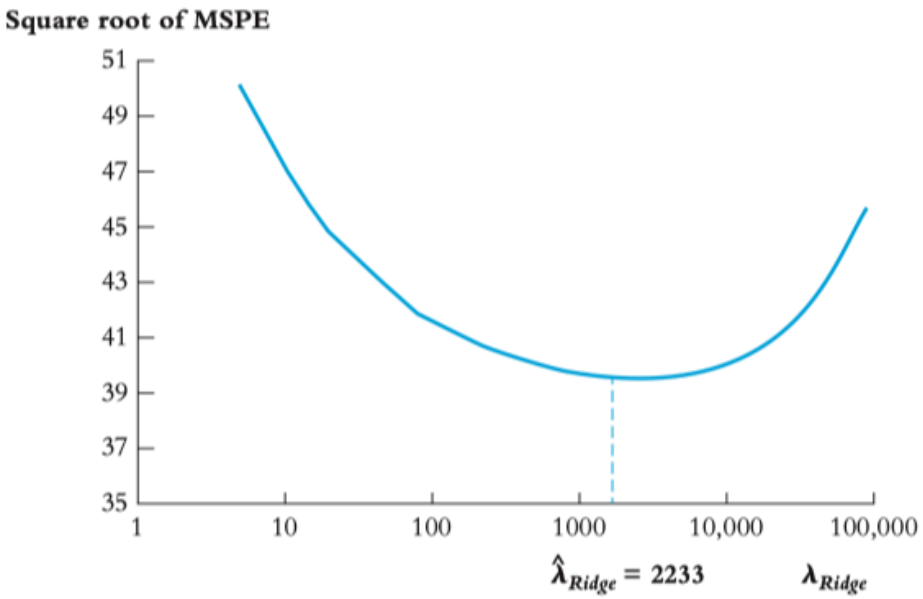
$$\hat{\lambda} = 2233$$

- Ridge have a smaller RMSE than OLS

$$RMSE_{ols} = 78.2; RMSE_{ridge} = 39.5$$

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.



Lasso Regression

Introduction

- Least Absolute Shrinkage and Selection Operator(LASSO): it simply replaces ridge's squared coefficients with absolute values.

$$\min_{\hat{\beta}} R^{lasso} = \min_{\hat{\beta}} \sum_{i=1}^{n} \left(Y_i - \underbrace{\left[\hat{\beta}_1 X_{i,1} + \cdots + \hat{\beta}_p X_{i,p} \right]}_{=\hat{y}_i} \right)^2 + \lambda_L \sum_{j=1}^{p} |\beta_j|$$

  - where $\lambda_L \sum_{j=1}^{p} |\beta_j|$ is the penalty.

- Unlike ridge, lasso's penalty does not increase with the size of β.
- The only way to avoid lasso's penalty is to set βs to zero.
- This feature has two benefits

  1. Some coefficients will be set to zero.
  2. Lasso can be used for subset/feature selection.

Estimate the Lasso estimator

- For simplicity, $p = 1$

$$R^{lasso} = \sum_{i=1}^{n} \left(Y_i - \hat{\beta} X_{i,1}\right)^2 + \lambda_L |\beta|$$

- Suppose $\hat{\beta} > 0$, then

$$R^{lasso} = \sum_{i=1}^{n} \left(Y_i - \hat{\beta} X_{i,1}\right)^2 + \lambda_L \beta$$

- Suppose $\hat{\beta} < 0$, then

$$R^{lasso} = \sum_{i=1}^{n} \left(Y_i - \hat{\beta} X_{i,1}\right)^2 - \lambda_L \beta$$

Estimate the Lasso estimator

- Then F.O.C

$$
\begin{aligned}
(-2)\sum_{i=1}^{n} X_i \left(Y_i - \hat{\beta}_{\text{Lasso}} X_i\right) + \lambda_{\text{Lasso}} &= 0 \\
\Rightarrow \sum_{i=1}^{n} X_i^2 \hat{\beta}_{\text{Lasso}} - \sum_{i=1}^{n} X_i Y_i &= -\frac{1}{2}\lambda_{\text{Lasso}} \\
\Rightarrow \hat{\beta}_{\text{Lasso}} &= \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} - \frac{1}{2}\frac{\lambda_{\text{Lasso}}}{\sum_{i=1}^{n} X_i^2} = \hat{\beta}_{ols} - \frac{1}{2}\frac{\lambda^{\text{Lasso}}}{\sum_{i=1}^{n} X_i^2}
\end{aligned}
$$

- Because we suppose $\hat{\beta} > 0$, then

$$
\hat{\beta}_{\text{Lasso}} = \max\left(\hat{\beta} - \frac{1}{2}\frac{\lambda_{\text{Lasso}}}{\sum_{i=1}^{n} X_i^2}, 0\right) \text{ when } \hat{\beta} > 0
$$

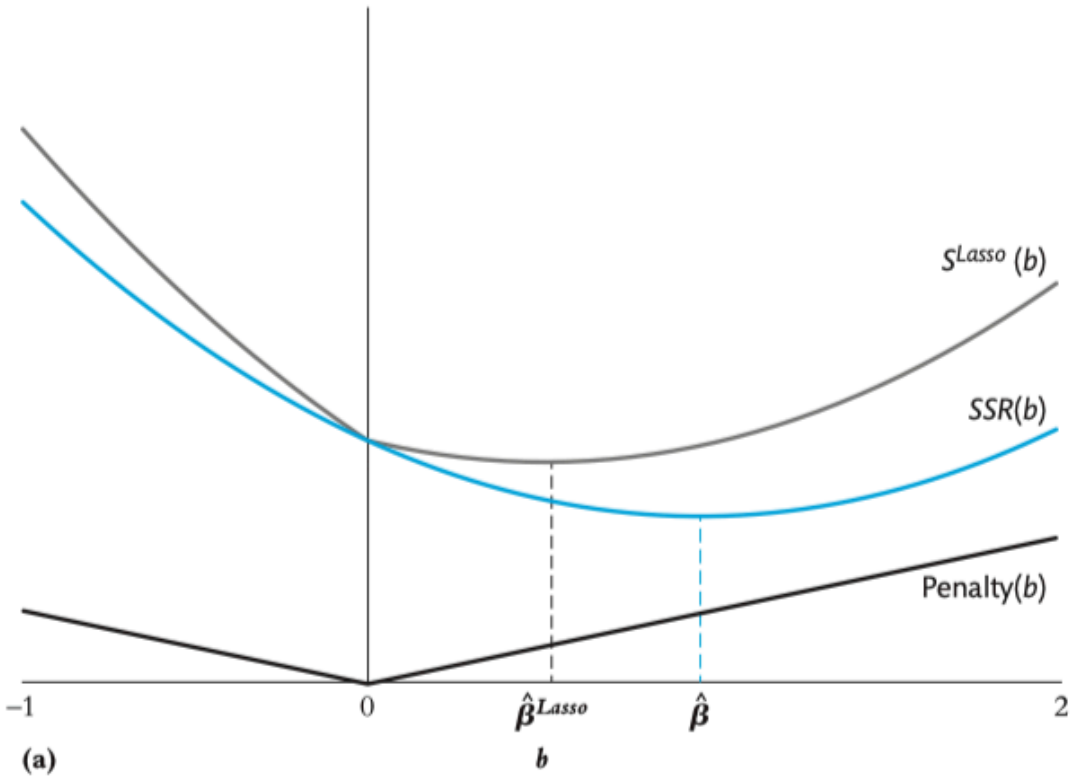
- Similar reasoning shows that when $\hat{\beta} < 0$

$$
\hat{\beta}_{\text{Lasso}} = \min\left(\hat{\beta} + \frac{1}{2}\frac{\lambda_{\text{Lasso}}}{\sum_{i=1}^{n} X_i^2}, 0\right) \text{ when } \hat{\beta} < 0
$$

The Lasso estimator

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.



- When the OLS estimator is large, the Lasso estimator shrinks it slightly towards zero— less than ridge.

The Lasso estimator

- Lasso sets many βs to ZERO, which means "select" some useful predictors for prediction and drops the others.
- It means that Lasso can work especially well when in reality many of the predictors are irrelevant.
- Models in which most of the true β are zero are called sparse models.

Predicting School-level test scores

- λ estimated by minimizing the 10-fold cross-validated RMSE.

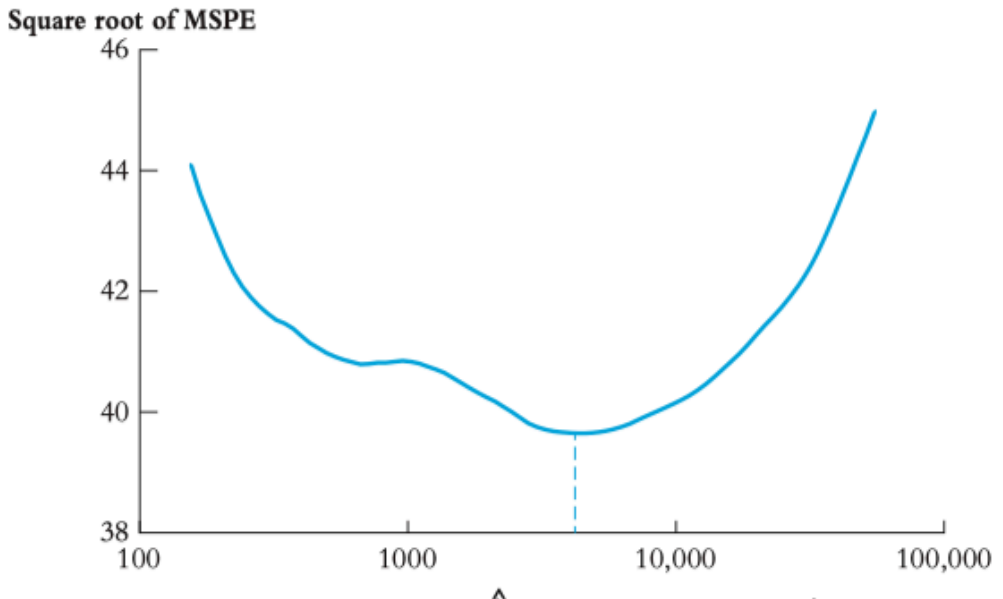
$$\hat{\lambda}_{Lasso} = 4527$$

- Only use $p = 56$ predictors and Lasso have a smaller RMSE than OLS

$$RMSE_{Lasso} = 39.7$$

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.



Principal Components

Introduction

- Ridge and Lasso reduce the RMSE by shrinking (biasing) the estimated coefficients to zero.
  - In the case of Lasso, by eliminating many of the regressors entirely.
- Instead, Principal components regression(PCR) collapses the very many predictors(k) into a much smaller number(p) of linear combinations of the predictors.
- These linear combinations – called the principal components of X – are computed so that they capture as much of the variation in the original X's as possible.
- Because the number p of principal components is small, OLS can be used, with the principal components as (new) regressors.

Principal Components(k=2)

- The easy way to combine X_1 and X_2 is a linear equation

$$aX_1 + bX_2$$

- Question : What values of a and b should be used?
- The Principal Components solution is to choose a and b to solve

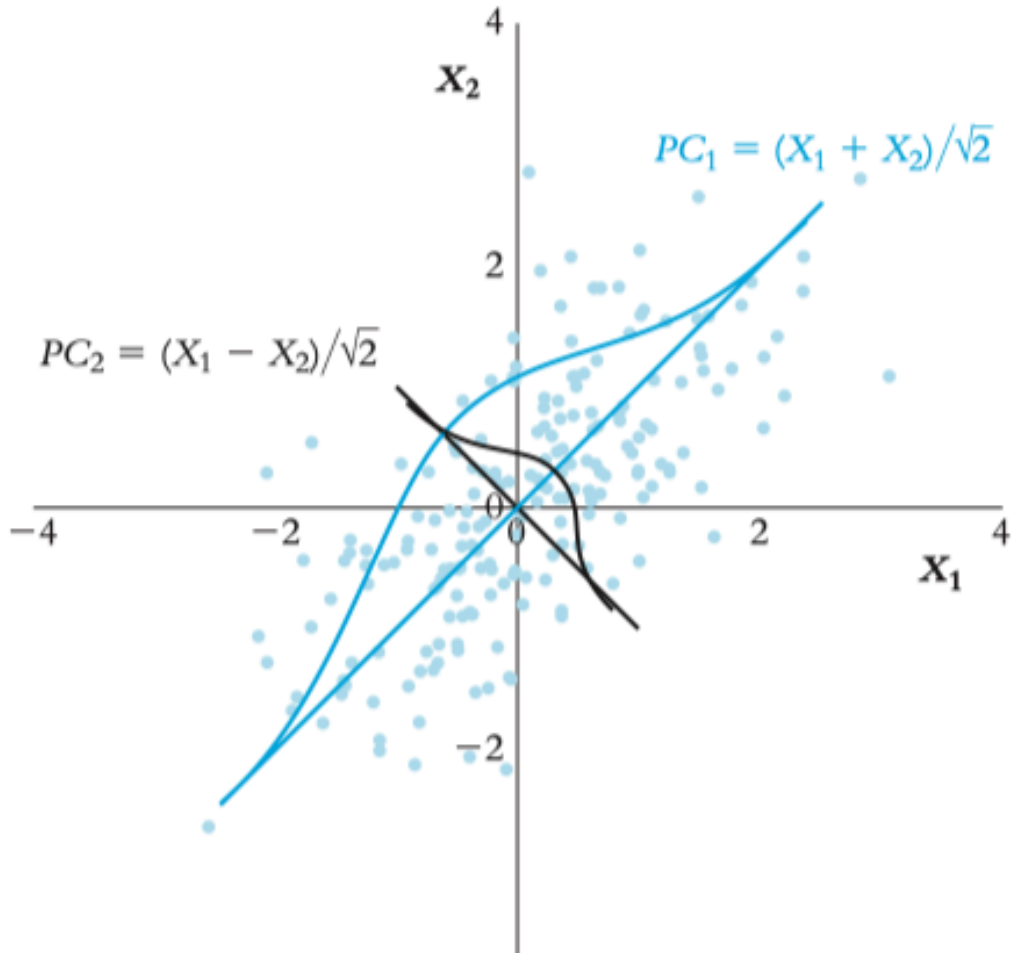
$$\max \mathrm{Var}(aX_1 + bX_2), \text{subject to } a^2 + b^2 = 1$$

- For X_1 and X_2 are positively correlated, then $a = b = \frac{1}{\sqrt{2}}$,
  - the first principal component(PC_1) is $(X_1 + X_2)/\sqrt{2}$
  - the second principal component(PC_2) is $(X_1 - X_2)/\sqrt{2}$, which is uncorrelated with the first.
- The principal component weights are normalized so that the sum of squared weights adds to 1.

Principal Components(k=2)

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.



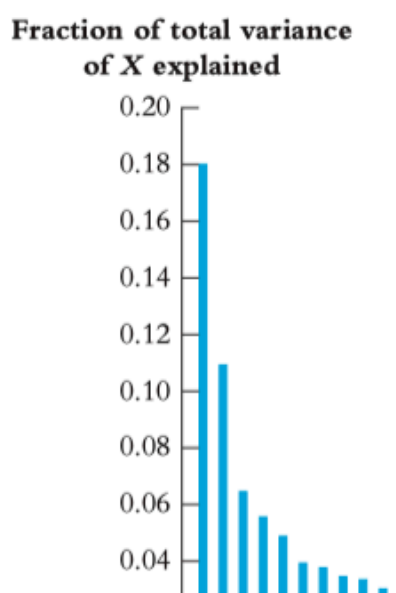
Principal Components

- Principal components can be thought of as a data compression tool, so that the compressed data have fewer regressors with as little information loss as possible.
- Data compression is used all the time to reduce very large data sets to smaller ones.
    - eg. image compression, where the goal is to retain as many of the features of the image (photograph) as possible, while reducing the file size.

Principal Components

- PC_1 explains 18% of the variation of all Xs.
- The first 10 PCs thus PC_1 ... PC_{10} explains 63%.
- The first 40 PCs explains 92%.
- The first 40 PCs explains 92%.

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

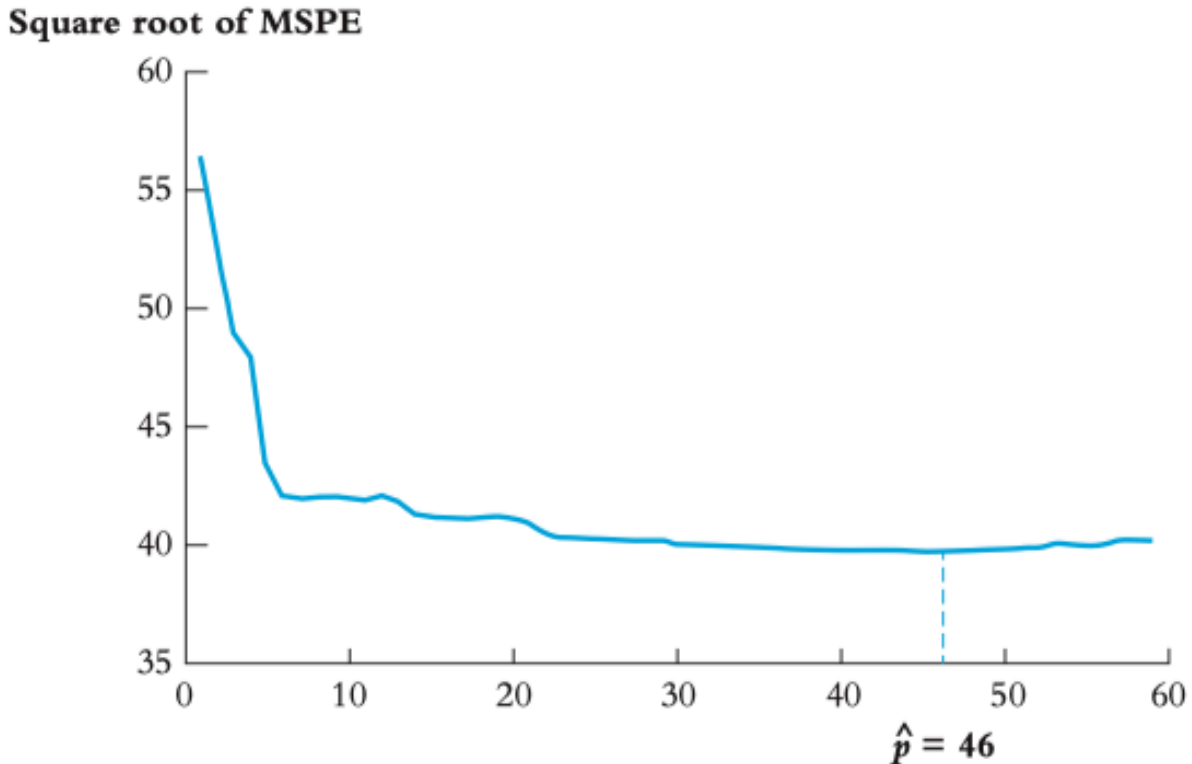


Principal Components

- It turns out that $p = 46$ by CV method.
- $$RMSE_{PC} = 39.7$$

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.



Predicting School Test Scores

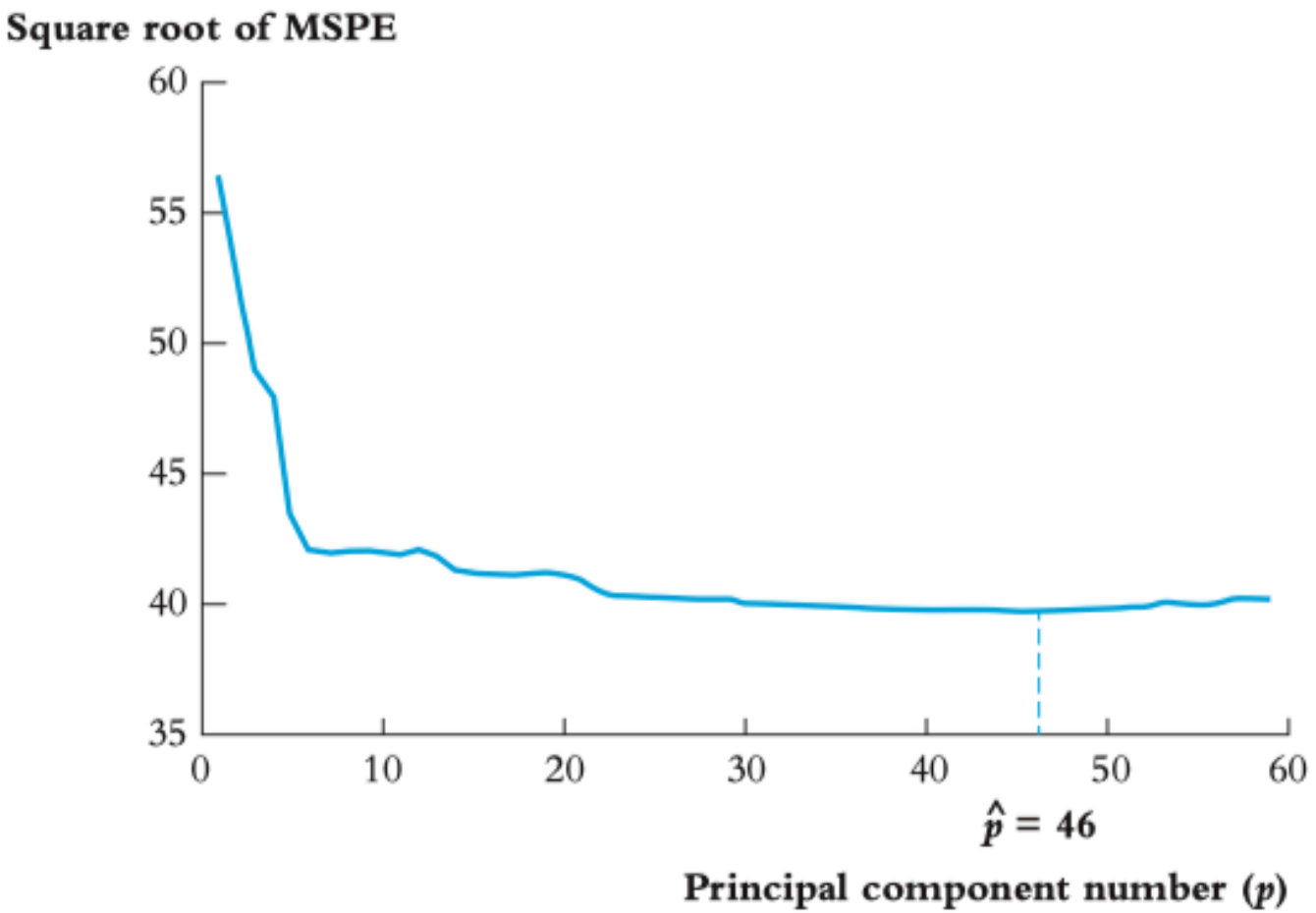
Basic procedure

- Split observations into two parts.
    1. The first half for model training.
    2. The second half for testing.
- There sets of predictors are used
    1. Small(k=4): Student-teacher ratio, median local income, teacher's average years of experience, instructional expenditures per student.
    2. Large(k=817)
    3. Vary Large(k=2065):Additional school and demographic variables, squares and cubes, and interactions. Note $k > n$.

The Three Sets of Predictors

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.



Compare predictive models by predictions

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

Compare predictive models by predictions

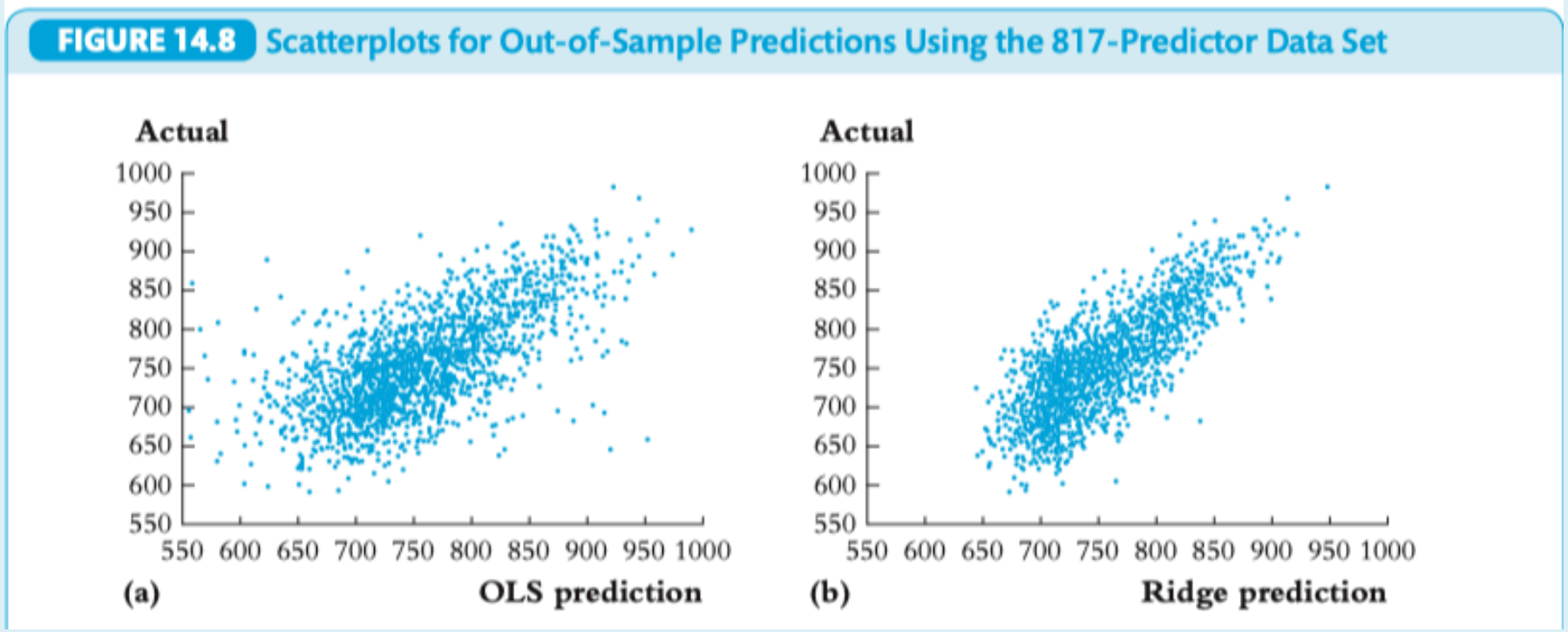
TABLE 14.4 Coefficients on Selected Standardized Regressors, 4- and 817-Variable Data Sets

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

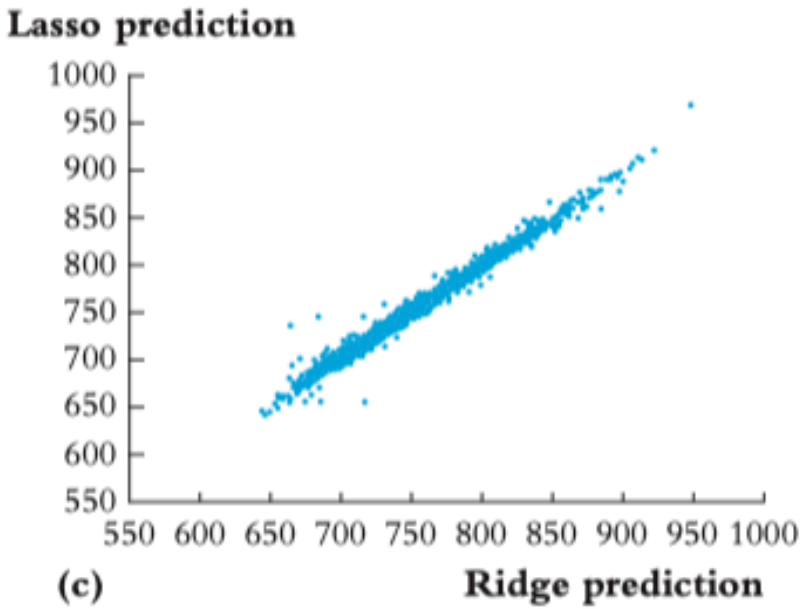
Notes: The index of part-time teachers measures the fraction of teachers who work part-time. For OLS, ridge, and Lasso, the coefficients in Table 14.4 are produced directly by the estimation algorithms. For principal components, the coefficients in Table 14.4 are computed from the principal component 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.

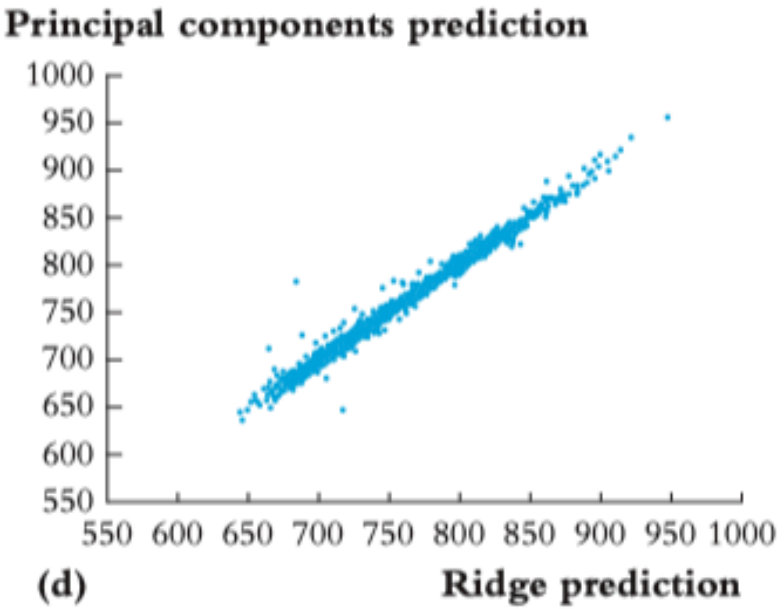
Compare predictive models by predictions

- The tighter the spread of the scatter along the 45 degree line, the better the prediction.



Compare predictive models by predictions





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

Compare predictive models by predictions

- The most important conclusion from this application is that for the large data set the many-predictor methods succeed where OLS fails.
- Because the many-predictor methods allow the coefficients to be biased in a way that reduces their variance by enough to compensate for the increased bias.
- One finding that may not generalize: three methods happen to perform equally well in these data.

Summary

- With many predictors, OLS will produce poor out-of-sample predictions.
- By introducing the right type of bias— shrinkage towards zero— the variance of the prediction can be reduced by enough to offset the bias and result in smaller RMSE.
- Ridge and Lasso reduce the RMSE by shrinking (biasing) the estimated coefficients to zero— and in the case of Lasso, by eliminating many of the regressors entirely.
- Principal components collapses X into fewer uncorrelated linear combinations that capture as much of the variation of the X's as possible. Predictions are then made using the OLS regression of Y on the principal components.