Big Data

Dr. Patrick Toche

Textbook:

James H. Stock and Mark W. Watson, *Introduction to Econometrics*, 4th Edition, Pearson.

Other references:

Jeffrey M. Wooldridge, *Introductory Econometrics: A Modern Approach*, 7th Edition, Cengage Learning.
Joshua D. Angrist and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, 1st Edition, Princeton University Press.

The textbook comes with online resources and study guides. Other references will be given from time to time.

Contents

In this lesson you will learn …

- What is “Big Data.”
- The many-predictor problem.
- Ridge regression.
- Lasso.
- Principal components.

In this lesson you will learn …

- What is “Big Data.”
- The many-predictor problem.
- Ridge regression.
- Lasso.
- Principal components.

In this lesson you will learn …

- What is “Big Data.”
- The many-predictor problem.
- Ridge regression.
- Lasso.
- Principal components.

In this lesson you will learn …

- What is “Big Data.”
- The many-predictor problem.
- Ridge regression.
- Lasso.
- Principal components.

In this lesson you will learn …

- What is “Big Data.”
- The many-predictor problem.
- Ridge regression.
- Lasso.
- Principal components.

What Is Big Data?

Prediction With Many Predictors

Shrinkage estimators

- **Big data:** Many observations and/or many predictors. Includes datasets with many categories.
- **Challenges:** Storing and accessing large data sets efficiently and developing fast algorithms for estimating models.
- **OLS overfits in large studies:** Use data on 3, 932 elementary schools in California to predict school-level test scores using data on school and community characteristics. Half of these observations are used to estimate prediction models, while the other half are reserved to test their performance. Large datasets have many predictors. The analysis starts with 817 predictors and is later expanded to 2065 predictors. With so many predictors, OLS overfits the data and makes poor out-of-sample predictions.
- **Shrinkage estimators:** Better out-of-sample predictors than OLS. Shrinkage estimators are biased and may not have a causal interpretation, but the variance of the shrinkage estimator is smaller, which improves out-of-sample prediction.

Prediction With Many Predictors

Shrinkage estimators

- **Big data:** Many observations and/or many predictors. Includes datasets with many categories.
- **Challenges:** Storing and accessing large data sets efficiently and developing fast algorithms for estimating models.
- **OLS overfits in large studies:** Use data on 3, 932 elementary schools in California to predict school-level test scores using data on school and community characteristics. Half of these observations are used to estimate prediction models, while the other half are reserved to test their performance. Large datasets have many predictors. The analysis starts with 817 predictors and is later expanded to 2065 predictors. With so many predictors, OLS overfits the data and makes poor out-of-sample predictions.
- **Shrinkage estimators:** Better out-of-sample predictors than OLS. Shrinkage estimators are biased and may not have a causal interpretation, but the variance of the shrinkage estimator is smaller, which improves out-of-sample prediction.

Prediction With Many Predictors

Shrinkage estimators

- **Big data:** Many observations and/or many predictors. Includes datasets with many categories.
- **Challenges:** Storing and accessing large data sets efficiently and developing fast algorithms for estimating models.
- **OLS overfits in large studies:** Use data on $3,932$ elementary schools in California to predict school-level test scores using data on school and community characteristics. Half of these observations are used to estimate prediction models, while the other half are reserved to test their performance. Large datasets have many predictors. The analysis starts with 817 predictors and is later expanded to 2065 predictors. With so many predictors, OLS overfits the data and makes poor out-of-sample predictions.
- **Shrinkage estimators:** Better out-of-sample predictors than OLS. Shrinkage estimators are biased and may not have a causal interpretation, but the variance of the shrinkage estimator is smaller, which improves out-of-sample prediction.

Prediction With Many Predictors

Shrinkage estimators

- **Big data:** Many observations and/or many predictors. Includes datasets with many categories.
- **Challenges:** Storing and accessing large data sets efficiently and developing fast algorithms for estimating models.
- **OLS overfits in large studies:** Use data on $3,932$ elementary schools in California to predict school-level test scores using data on school and community characteristics. Half of these observations are used to estimate prediction models, while the other half are reserved to test their performance. Large datasets have many predictors. The analysis starts with 817 predictors and is later expanded to 2065 predictors. With so many predictors, OLS overfits the data and makes poor out-of-sample predictions.
- **Shrinkage estimators:** Better out-of-sample predictors than OLS. Shrinkage estimators are biased and may not have a causal interpretation, but the variance of the shrinkage estimator is smaller, which improves out-of-sample prediction.

Applications of Big Data

1. **Many 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.
2. **Categorization:** Regression with a binary dependent variable and regression with many categories.
3. **Testing multiple hypotheses:** The F-statistic of a joint hypothesis on a group of coefficients is not well suited to find out which of the treatments is effective. Special methods have been developed to test large numbers of individual hypotheses to determine which treatment effect is nonzero.
4. **Nonstandard data:** Including text and images.
5. **Deep learning:** Non-linear models are estimated ("trained") using very many observations. Useful for pattern recognition, such as facial recognition; speech recognition; multi-language translation; detecting anomalies in medical scans; interpreting network data on social media; high-frequency trading in financial markets.

Applications of Big Data

1. **Many 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.
2. **Categorization:** Regression with a binary dependent variable and regression with many categories.
3. **Testing multiple hypotheses:** The F-statistic of a joint hypothesis on a group of coefficients is not well suited to find out which of the treatments is effective. Special methods have been developed to test large numbers of individual hypotheses to determine which treatment effect is nonzero.
4. **Nonstandard data:** Including text and images.
5. **Deep learning:** Non-linear models are estimated ("trained") using very many observations. Useful for pattern recognition, such as facial recognition; speech recognition; multi-language translation; detecting anomalies in medical scans; interpreting network data on social media; high-frequency trading in financial markets.

Applications of Big Data

1. **Many 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.
2. **Categorization:** Regression with a binary dependent variable and regression with many categories.
3. **Testing multiple hypotheses:** The F-statistic of a joint hypothesis on a group of coefficients is not well suited to find out which of the treatments is effective. Special methods have been developed to test large numbers of individual hypotheses to determine which treatment effect is nonzero.
4. **Nonstandard data:** Including text and images.
5. **Deep learning:** Non-linear models are estimated ("trained") using very many observations. Useful for pattern recognition, such as facial recognition; speech recognition; multi-language translation; detecting anomalies in medical scans; interpreting network data on social media; high-frequency trading in financial markets.

Applications of Big Data

1. **Many 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.
2. **Categorization:** Regression with a binary dependent variable and regression with many categories.
3. **Testing multiple hypotheses:** The F-statistic of a joint hypothesis on a group of coefficients is not well suited to find out which of the treatments is effective. Special methods have been developed to test large numbers of individual hypotheses to determine which treatment effect is nonzero.
4. **Nonstandard data:** Including text and images.
5. **Deep learning:** Non-linear models are estimated ("trained") using very many observations. Useful for pattern recognition, such as facial recognition; speech recognition; multi-language translation; detecting anomalies in medical scans; interpreting network data on social media; high-frequency trading in financial markets.

Applications of Big Data

1. **Many 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.
2. **Categorization:** Regression with a binary dependent variable and regression with many categories.
3. **Testing multiple hypotheses:** The F-statistic of a joint hypothesis on a group of coefficients is not well suited to find out which of the treatments is effective. Special methods have been developed to test large numbers of individual hypotheses to determine which treatment effect is nonzero.
4. **Nonstandard data:** Including text and images.
5. **Deep learning:** Non-linear models are estimated ("trained") using very many observations. Useful for pattern recognition, such as facial recognition; speech recognition; multi-language translation; detecting anomalies in medical scans; interpreting network data on social media; high-frequency trading in financial markets.

Many Predictors

Many-Predictor Problem

- The analysis of the district test score data reveals nonlinearities and interactions in the test score regressions. For example, there is a nonlinear relationship between test scores and the student-teacher ratio and this relationship differs depending on whether there are a large number of English learners in the district. These nonlinearities are handled by including third-degree polynomials of the student–teacher ratio and interaction terms.
- If only the main variables are used, there are 38 regressors. Including interactions, squares, and cubes increases the number of predictors to 817.
- We can also use a larger data set with 2, 065 predictors more than the 1, 966 observations in the estimation sample. While it does not violate the Gauss–Markov theorem, OLS can produce poor predictions when the number of predictors is large relative to the sample size.

Many-Predictor Problem

- The analysis of the district test score data reveals nonlinearities and interactions in the test score regressions. For example, there is a nonlinear relationship between test scores and the student-teacher ratio and this relationship differs depending on whether there are a large number of English learners in the district. These nonlinearities are handled by including third-degree polynomials of the student–teacher ratio and interaction terms.
- If only the main variables are used, there are 38 regressors. Including interactions, squares, and cubes increases the number of predictors to 817.
- We can also use a larger data set with $2,065$ predictors more than the $1,966$ observations in the estimation sample. While it does not violate the Gauss–Markov theorem, OLS can produce poor predictions when the number of predictors is large relative to the sample size.

Many-Predictor Problem

- The analysis of the district test score data reveals nonlinearities and interactions in the test score regressions. For example, there is a nonlinear relationship between test scores and the student-teacher ratio and this relationship differs depending on whether there are a large number of English learners in the district. These nonlinearities are handled by including third-degree polynomials of the student–teacher ratio and interaction terms.
- If only the main variables are used, there are 38 regressors. Including interactions, squares, and cubes increases the number of predictors to 817.
- We can also use a larger data set with $2,065$ predictors more than the $1,966$ observations in the estimation sample. While it does not violate the Gauss–Markov theorem, OLS can produce poor predictions when the number of predictors is large relative to the sample size.

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 $= k = 38 + 38 + 38 + 703 = 817$

Mean Squared Prediction Error

- **Mean-Squared Prediction Error (MSPE):** 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.

$$MSPE = \mathsf{E}[Y^{\text{OOS}} - \hat{Y}(X^{\text{OOS}})]^2$$

where X^{OOS} and Y^{OOS} are out-of-sample observations on X and Y; where $\hat{Y}(x)$ is the predicted value of Y for some given value x.

- **Oracle prediction:** $\mathsf{E}[Y^{\text{OOS}} | X^{\text{OOS}}]$ The conditional mean minimizes the MSPE. It is not directly observable, so estimating it with the model coefficients of the prediction model introduces additional sources of error.
- The Oracle prediction is the benchmark against which to judge all feasible predictions.

Mean Squared Prediction Error

- **Mean-Squared Prediction Error (MSPE):** 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.

$$MSPE = \mathrm{E}[Y^{\mathrm{OOS}} - \hat{Y}(X^{\mathrm{OOS}})]^2$$

  where X^{OOS} and Y^{OOS} are out-of-sample observations on X and Y; where $\hat{Y}(x)$ is the predicted value of Y for some given value x.
- **Oracle prediction:** $\mathrm{E}[Y^{\mathrm{OOS}}|X^{\mathrm{OOS}}]$ The conditional mean minimizes the MSPE. It is not directly observable, so estimating it with the model coefficients of the prediction model introduces additional sources of error.
- The Oracle prediction is the benchmark against which to judge all feasible predictions.

Mean Squared Prediction Error

- **Mean-Squared Prediction Error (MSPE):** 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.

$$MSPE = \mathrm{E}[Y^{\mathrm{OOS}} - \hat{Y}(X^{\mathrm{OOS}})]^2$$

  where X^{OOS} and Y^{OOS} are out-of-sample observations on X and Y; where $\hat{Y}(x)$ is the predicted value of Y for some given value x.
- **Oracle prediction:** $\mathrm{E}[Y^{\mathrm{OOS}}|X^{\mathrm{OOS}}]$ The conditional mean minimizes the MSPE. It is not directly observable, so estimating it with the model coefficients of the prediction model introduces additional sources of error.
- The Oracle prediction is the benchmark against which to judge all feasible predictions.

Predictive Regression Model with Standardized Regressors

- **Standardized regressors:** Regressors are transformed to have mean 0 and variance 1.

$$X_{ji} = \frac{X_{ji}^{\text{original}} - \mu_{X_j^{\text{original}}}}{\sigma_{X_j^{\text{original}}}}$$

  where $\mu_{X_j^{\text{original}}}$ is the population mean of the original regressor.

- **Standardized regressand:** The regressand is transformed to have mean 0.
- **Standardized predictive regression model:** The intercept is excluded because all the variables have mean 0.

$$Y_i = \beta_1 X_{1i} + \beta_k X_{ki} + u_i$$

- The regression coefficients have the same units.
- **Interpretation:** β_j is the difference in the predicted value of Y associated with a one standard deviation difference in X_j, holding

Predictive Regression Model with Standardized Regressors

- **Standardized regressors:** Regressors are transformed to have mean 0 and variance 1.

$$X_{ji} = \frac{X_{ji}^{\text{original}} - \mu_{X_j^{\text{original}}}}{\sigma_{X_j^{\text{original}}}}$$

  where $\mu_{X_j^{\text{original}}}$ is the population mean of the original regressor.

- **Standardized regressand:** The regressand is transformed to have mean 0.
- **Standardized predictive regression model:** The intercept is excluded because all the variables have mean 0.

$$Y_i = \beta_1 X_{1i} + \beta_k X_{ki} + u_i$$

- The regression coefficients have the same units.
- **Interpretation:** β_j is the difference in the predicted value of Y associated with a one standard deviation difference in X_j, holding

Predictive Regression Model with Standardized Regressors

- **Standardized regressors:** Regressors are transformed to have mean 0 and variance 1.

$$X_{ji} = \frac{X_{ji}^{\text{original}} - \mu_{X_j^{\text{original}}}}{\sigma_{X_j^{\text{original}}}}$$

  where $\mu_{X_j^{\text{original}}}$ is the population mean of the original regressor.

- **Standardized regressand:** The regressand is transformed to have mean 0.
- **Standardized predictive regression model:** The intercept is excluded because all the variables have mean 0.

$$Y_i = \beta_1 X_{1i} + \beta_k X_{ki} + u_i$$

- The regression coefficients have the same units.
- **Interpretation:** β_j is the difference in the predicted value of Y associated with a one standard deviation difference in X_j, holding

Predictive Regression Model with Standardized Regressors

- **Standardized regressors:** Regressors are transformed to have mean 0 and variance 1.

$$X_{ji} = \frac{X_{ji}^{\text{original}} - \mu_{X_j^{\text{original}}}}{\sigma_{X_j^{\text{original}}}}$$

  where $\mu_{X_j^{\text{original}}}$ is the population mean of the original regressor.

- **Standardized regressand:** The regressand is transformed to have mean 0.
- **Standardized predictive regression model:** The intercept is excluded because all the variables have mean 0.

$$Y_i = \beta_1 X_{1i} + \beta_k X_{ki} + u_i$$

- The regression coefficients have the same units.
- **Interpretation:** β_j is the difference in the predicted value of Y associated with a one standard deviation difference in X_j, holding

Predictive Regression Model with Standardized Regressors

- **Standardized regressors:** Regressors are transformed to have mean 0 and variance 1.

$$X_{ji} = \frac{X_{ji}^{\text{original}} - \mu_{X_j^{\text{original}}}}{\sigma_{X_j^{\text{original}}}}$$

  where $\mu_{X_j^{\text{original}}}$ is the population mean of the original regressor.

- **Standardized regressand:** The regressand is transformed to have mean 0.
- **Standardized predictive regression model:** The intercept is excluded because all the variables have mean 0.

$$Y_i = \beta_1 X_{1i} + \beta_k X_{ki} + u_i$$

- The regression coefficients have the same units.
- **Interpretation:** β_j is the difference in the predicted value of Y associated with a one standard deviation difference in X_j, holding

MSPE in the Standardized Predictive Regression

Minimize the variance term, taking the bias as given

- The standardized predictive regression model can be written as the sum of two components:

$$MSPE = \sigma_u^2 + \mathsf{E}\left[(\hat{\beta}_1 - \beta_1)X_1^{\text{OOS}} + \ldots + (\hat{\beta}_k - \beta_k)X_k^{\text{OOS}}\right]^2$$

- The mean-squared error is the sum of the bias and of the variance.
- The first term σ_u^2 is the variance of the oracle prediction error: The prediction error made using the true (unknown) conditional mean.
- The second term is the contribution to the prediction error arising from the estimated regression coefficients. This cost arises from estimating the coefficients instead of using the true oracle prediction.
- **Objective:** Minimize the variance term, taking the bias as given.
- **Prediction for an out-of-sample observation:** Because the regressors are standardized and the dependent variable is demeaned, 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.

MSPE in the Standardized Predictive Regression

Minimize the variance term, taking the bias as given

- The standardized predictive regression model can be written as the sum of two components:

$$MSPE = \sigma_u^2 + \mathsf{E}\left[(\hat{\beta}_1 - \beta_1)X_1^{\text{OOS}} + \ldots + (\hat{\beta}_k - \beta_k)X_k^{\text{OOS}}\right]^2$$

- **The mean-squared error is the sum of the bias and of the variance.**
- The first term σ_u^2 is the variance of the oracle prediction error: The prediction error made using the true (unknown) conditional mean.
- The second term is the contribution to the prediction error arising from the estimated regression coefficients. This cost arises from estimating the coefficients instead of using the true oracle prediction.
- **Objective:** Minimize the variance term, taking the bias as given.
- **Prediction for an out-of-sample observation:** Because the regressors are standardized and the dependent variable is demeaned, 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.

MSPE in the Standardized Predictive Regression

Minimize the variance term, taking the bias as given

- The standardized predictive regression model can be written as the sum of two components:

$$MSPE = \sigma_u^2 + \mathsf{E}\left[(\hat{\beta}_1 - \beta_1)X_1^{\text{OOS}} + \ldots + (\hat{\beta}_k - \beta_k)X_k^{\text{OOS}}\right]^2$$

- **The mean-squared error is the sum of the bias and of the variance.**
- The first term σ_u^2 is the variance of the oracle prediction error: The prediction error made using the true (unknown) conditional mean.
- The second term is the contribution to the prediction error arising from the estimated regression coefficients. This cost arises from estimating the coefficients instead of using the true oracle prediction.
- **Objective:** Minimize the variance term, taking the bias as given.
- **Prediction for an out-of-sample observation:** Because the regressors are standardized and the dependent variable is demeaned, 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.

MSPE in the Standardized Predictive Regression

Minimize the variance term, taking the bias as given

- The standardized predictive regression model can be written as the sum of two components:

$$MSPE = \sigma_u^2 + \mathsf{E}\left[(\hat{\beta}_1 - \beta_1)X_1^{\text{OOS}} + \ldots + (\hat{\beta}_k - \beta_k)X_k^{\text{OOS}}\right]^2$$

- **The mean-squared error is the sum of the bias and of the variance.**
- The first term σ_u^2 is the variance of the oracle prediction error: The prediction error made using the true (unknown) conditional mean.
- The second term is the contribution to the prediction error arising from the estimated regression coefficients. This cost arises from estimating the coefficients instead of using the true oracle prediction.
- **Objective:** Minimize the variance term, taking the bias as given.
- **Prediction for an out-of-sample observation:** Because the regressors are standardized and the dependent variable is demeaned, 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.

MSPE in the Standardized Predictive Regression

Minimize the variance term, taking the bias as given

- The standardized predictive regression model can be written as the sum of two components:

$$MSPE = \sigma_u^2 + \mathsf{E}\left[(\hat{\beta}_1 - \beta_1)X_1^{\text{OOS}} + \ldots + (\hat{\beta}_k - \beta_k)X_k^{\text{OOS}}\right]^2$$

- **The mean-squared error is the sum of the bias and of the variance.**
- The first term σ_u^2 is the variance of the oracle prediction error: The prediction error made using the true (unknown) conditional mean.
- The second term is the contribution to the prediction error arising from the estimated regression coefficients. This cost arises from estimating the coefficients instead of using the true oracle prediction.
- **Objective:** Minimize the variance term, taking the bias as given.
- **Prediction for an out-of-sample observation:** Because the regressors are standardized and the dependent variable is demeaned, 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.

MSPE in the Standardized Predictive Regression

Minimize the variance term, taking the bias as given

- The standardized predictive regression model can be written as the sum of two components:

$$MSPE = \sigma_u^2 + \mathsf{E}\left[(\hat{\beta}_1 - \beta_1)X_1^{\text{OOS}} + \ldots + (\hat{\beta}_k - \beta_k)X_k^{\text{OOS}}\right]^2$$

- **The mean-squared error is the sum of the bias and of the variance.**
- The first term σ_u^2 is the variance of the oracle prediction error: The prediction error made using the true (unknown) conditional mean.
- The second term is the contribution to the prediction error arising from the estimated regression coefficients. This cost arises from estimating the coefficients instead of using the true oracle prediction.
- **Objective:** Minimize the variance term, taking the bias as given.
- **Prediction for an out-of-sample observation:** Because the regressors are standardized and the dependent variable is demeaned, 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.

OLS is Best Linear Unbiased

- In the special case of homoskedastic regression errors, the MSPE of OLS is given by

$$MSPE \approx \left(1 + \frac{k}{n}\right) \sigma_u^2$$

  The approximation more accurate for large n and small k/n.

- The cost of using OLS, as measured by the MSPE, depends on the ratio of the number of regressors to the sample size.
- In the school test score application with 38 regressors, using OLS has a loss of only 2% relative to the Oracle prediction. But with 817 regressors, the loss increases to 40%.

$$\frac{k}{n} = \frac{38}{1,966} \approx 0.02$$

$$\frac{k}{n} = \frac{817}{1,966} \approx 0.40$$

- Because OLS is unbiased, the loss is entirely due to the variance term. Under Gauss-Markov, this is the smallest loss in the class of linear, unbiased estimators.
- The loss can be reduced using ...biased estimators!

OLS is Best Linear Unbiased

- In the special case of homoskedastic regression errors, the MSPE of OLS is given by

$$MSPE \approx \left(1 + \frac{k}{n}\right) \sigma_u^2$$

  The approximation more accurate for large n and small k/n.

- The cost of using OLS, as measured by the MSPE, depends on the ratio of the number of regressors to the sample size.

- In the school test score application with 38 regressors, using OLS has a loss of only 2% relative to the Oracle prediction. But with 817 regressors, the loss increases to 40%.

$$\frac{k}{n} = \frac{38}{1,966} \approx 0.02$$

$$\frac{k}{n} = \frac{817}{1,966} \approx 0.40$$

- Because OLS is unbiased, the loss is entirely due to the variance term. Under Gauss-Markov, this is the smallest loss in the class of linear, unbiased estimators.

- The loss can be reduced using ...biased estimators!

OLS is Best Linear Unbiased

- In the special case of homoskedastic regression errors, the MSPE of OLS is given by

$$MSPE \approx \left(1 + \frac{k}{n}\right) \sigma_u^2$$

  The approximation more accurate for large n and small k/n.
- The cost of using OLS, as measured by the MSPE, depends on the ratio of the number of regressors to the sample size.
- In the school test score application with 38 regressors, using OLS has a loss of only 2% relative to the Oracle prediction. But with 817 regressors, the loss increases to 40%.

$$\frac{k}{n} = \frac{38}{1,966} \approx 0.02$$
$$\frac{k}{n} = \frac{817}{1,966} \approx 0.40$$

- Because OLS is unbiased, the loss is entirely due to the variance term. Under Gauss-Markov, this is the smallest loss in the class of linear, unbiased estimators.
- The loss can be reduced using ...biased estimators!

OLS is Best Linear Unbiased

- In the special case of homoskedastic regression errors, the MSPE of OLS is given by

$$MSPE \approx \left(1 + \frac{k}{n}\right) \sigma_u^2$$

  The approximation more accurate for large n and small k/n.
- The cost of using OLS, as measured by the MSPE, depends on the ratio of the number of regressors to the sample size.
- In the school test score application with 38 regressors, using OLS has a loss of only 2% relative to the Oracle prediction. But with 817 regressors, the loss increases to 40%.

$$\frac{k}{n} = \frac{38}{1,966} \approx 0.02$$
$$\frac{k}{n} = \frac{817}{1,966} \approx 0.40$$

- Because OLS is unbiased, the loss is entirely due to the variance term. Under Gauss-Markov, this is the smallest loss in the class of linear, unbiased estimators.
- The loss can be reduced using ...biased estimators!

OLS is Best Linear Unbiased

- In the special case of homoskedastic regression errors, the MSPE of OLS is given by

$$MSPE \approx \left(1 + \frac{k}{n}\right) \sigma_u^2$$

  The approximation more accurate for large n and small k/n.
- The cost of using OLS, as measured by the MSPE, depends on the ratio of the number of regressors to the sample size.
- In the school test score application with 38 regressors, using OLS has a loss of only 2% relative to the Oracle prediction. But with 817 regressors, the loss increases to 40%.

$$\frac{k}{n} = \frac{38}{1,966} \approx 0.02$$
$$\frac{k}{n} = \frac{817}{1,966} \approx 0.40$$

- Because OLS is unbiased, the loss is entirely due to the variance term. Under Gauss-Markov, this is the smallest loss in the class of linear, unbiased estimators.
- The loss can be reduced using ...biased estimators!

The Principle of Shrinkage

- **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, 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.
- **James-Stein estimator:** When the regressors are uncorrelated, the James-Stein estimator can be written $\tilde{\beta}^{JS} = c\hat{\beta}$, where $\hat{\beta}$ is the OLS estimator and c is a factor that is less than 1 and depends on the data. Since $c < 1$, the JS estimator shrinks the OLS estimator toward 0 and thus is biased toward 0.
- **James and Stein (1961):** If the errors are normally distributed and $k \geq 3$, their estimator has a lower mean squared error than the OLS estimator, regardless of the true value of β.
- James-Stein leads to the family of shrinkage estimators, which includes ridge regression and the Lasso estimator.

The Principle of Shrinkage

- **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, 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.
- **James-Stein estimator:** When the regressors are uncorrelated, the James-Stein estimator can be written $\tilde{\beta}^{JS} = c\hat{\beta}$, where $\hat{\beta}$ is the OLS estimator and c is a factor that is less than 1 and depends on the data. Since $c < 1$, the JS estimator shrinks the OLS estimator toward 0 and thus is biased toward 0.
- **James and Stein (1961):** If the errors are normally distributed and $k \geq 3$, their estimator has a lower mean squared error than the OLS estimator, regardless of the true value of β.
- James-Stein leads to the family of shrinkage estimators, which includes ridge regression and the Lasso estimator.

The Principle of Shrinkage

- **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, 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.
- **James-Stein estimator:** When the regressors are uncorrelated, the James-Stein estimator can be written $\tilde{\beta}^{JS} = c\hat{\beta}$, where $\hat{\beta}$ is the OLS estimator and c is a factor that is less than 1 and depends on the data. Since $c < 1$, the JS estimator shrinks the OLS estimator toward 0 and thus is biased toward 0.
- **James and Stein (1961):** If the errors are normally distributed and $k \geq 3$, their estimator has a lower mean squared error than the OLS estimator, regardless of the true value of β.
- James-Stein leads to the family of shrinkage estimators, which includes ridge regression and the Lasso estimator.

The Principle of Shrinkage

- **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, 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.
- **James-Stein estimator:** When the regressors are uncorrelated, the James-Stein estimator can be written $\tilde{\beta}^{JS} = c\hat{\beta}$, where $\hat{\beta}$ is the OLS estimator and c is a factor that is less than 1 and depends on the data. Since $c < 1$, the JS estimator shrinks the OLS estimator toward 0 and thus is biased toward 0.
- **James and Stein (1961):** If the errors are normally distributed and $k \geq 3$, their estimator has a lower mean squared error than the OLS estimator, regardless of the true value of β.
- James-Stein leads to the family of shrinkage estimators, which includes ridge regression and the Lasso estimator.

The Principle of Shrinkage

- **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, 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.
- **James-Stein estimator:** When the regressors are uncorrelated, the James-Stein estimator can be written $\tilde{\beta}^{JS} = c\hat{\beta}$, where $\hat{\beta}$ is the OLS estimator and c is a factor that is less than 1 and depends on the data. Since $c < 1$, the JS estimator shrinks the OLS estimator toward 0 and thus is biased toward 0.
- **James and Stein (1961):** If the errors are normally distributed and $k \geq 3$, their estimator has a lower mean squared error than the OLS estimator, regardless of the true value of β.
- James-Stein leads to the family of shrinkage estimators, which includes ridge regression and the Lasso estimator.

Estimation by Split Sample

Split sample:

- Estimate the MSPE by dividing the data set into two parts
  - an "estimation" subsample.
  - a "test" subsample used to simulate out-of- sample prediction.

$$MSPE_{\text{split-sample}} = \frac{1}{n_{\text{test subsample}}} \sum_{\text{test subsample}} (Y_i - \hat{Y}_i)^2$$

Estimation by Split Sample

Split sample:

- Estimate the MSPE by dividing the data set into two parts
  - an "estimation" subsample.
  - a "test" subsample used to simulate out-of- sample prediction.

$$MSPE_{\text{split-sample}} = \frac{1}{n_{\text{test subsample}}} \sum_{\text{test subsample}} (Y_i - \hat{Y}_i)^2$$

Estimation by Split Sample

Split sample:

- Estimate the MSPE by dividing the data set into two parts
  - an “estimation” subsample.
  - a “test” subsample used to simulate out-of- sample prediction.

$$MSPE_{\text{split-sample}} = \frac{1}{n_{\text{test subsample}}} \sum_{\text{test subsample}} (Y_i - \hat{Y}_i)^2$$

Estimation by m-fold Cross Validation

m-fold cross validation:

1. Estimate the MSPE by dividing the data set into two parts: an "estimation" subsample and a "test" subsample used to simulate out-of-sample prediction.
2. Use the combined sub-samples $2, 3, \ldots, m$ to compute $\tilde{\beta}$, an estimate of β.
3. Use $\tilde{\beta}$ to compute predicted values $\hat{Y}$ and prediction errors $Y - \hat{Y}$ for sub-sample 1.
4. Using sub-sample 1 as the test sample, estimate the MSPE with the predicted values in sub-sample 1.
5. Repeat steps 2-4 leaving out sub-sample 2, then 3, ..., then m.
6. The m-fold cross-validation estimator of the MSPE is estimated by averaging these m sub-sample estimates of the MSPE.

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

where n_i is the number of observations in sub-sample i.

Estimation by m-fold Cross Validation

m-fold cross validation:

1. Estimate the MSPE by dividing the data set into two parts: an "estimation" subsample and a "test" subsample used to simulate out-of-sample prediction.
2. Use the combined sub-samples $2, 3, \ldots, m$ to compute $\tilde{\beta}$, an estimate of β.
3. Use $\tilde{\beta}$ to compute predicted values $\tilde{Y}$ and prediction errors $Y - \tilde{Y}$ for sub-sample 1.
4. Using sub-sample 1 as the test sample, estimate the MSPE with the predicted values in sub-sample 1.
5. Repeat steps 2-4 leaving out sub-sample 2, then 3, ..., then m.
6. The m-fold cross-validation estimator of the MSPE is estimated by averaging these m sub-sample estimates of the MSPE.

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

where n_i is the number of observations in sub-sample i.

Estimation by m-fold Cross Validation

m-fold cross validation:

1. Estimate the MSPE by dividing the data set into two parts: an "estimation" subsample and a "test" subsample used to simulate out-of-sample prediction.
2. Use the combined sub-samples $2, 3, \ldots, m$ to compute $\tilde{\beta}$, an estimate of β.
3. Use $\tilde{\beta}$ to compute predicted values $\hat{Y}$ and prediction errors $Y - \hat{Y}$ for sub-sample 1.
4. Using sub-sample 1 as the test sample, estimate the MSPE with the predicted values in sub-sample 1.
5. Repeat steps 2-4 leaving out sub-sample 2, then 3, ..., then m.
6. The m-fold cross-validation estimator of the MSPE is estimated by averaging these m sub-sample estimates of the MSPE.

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

where n_i is the number of observations in sub-sample i.

Estimation by m-fold Cross Validation

m-fold cross validation:

1. Estimate the MSPE by dividing the data set into two parts: an "estimation" subsample and a "test" subsample used to simulate out-of-sample prediction.
2. Use the combined sub-samples $2, 3, \ldots, m$ to compute $\tilde{\beta}$, an estimate of β.
3. Use $\tilde{\beta}$ to compute predicted values $\hat{Y}$ and prediction errors $Y - \hat{Y}$ for sub-sample 1.
4. Using sub-sample 1 as the test sample, estimate the MSPE with the predicted values in sub-sample 1.
5. Repeat steps 2-4 leaving out sub-sample 2, then 3, ..., then m.
6. The m-fold cross-validation estimator of the MSPE is estimated by averaging these m sub-sample estimates of the MSPE.

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

where n_i is the number of observations in sub-sample i.

Estimation by m-fold Cross Validation

m-fold cross validation:

1. Estimate the MSPE by dividing the data set into two parts: an "estimation" subsample and a "test" subsample used to simulate out-of-sample prediction.
2. Use the combined sub-samples $2, 3, \ldots, m$ to compute $\tilde{\beta}$, an estimate of β.
3. Use $\tilde{\beta}$ to compute predicted values $\hat{Y}$ and prediction errors $Y - \hat{Y}$ for sub-sample 1.
4. Using sub-sample 1 as the test sample, estimate the MSPE with the predicted values in sub-sample 1.
5. Repeat steps 2-4 leaving out sub-sample 2, then 3, ..., then m.
6. The m-fold cross-validation estimator of the MSPE is estimated by averaging these m sub-sample estimates of the MSPE.

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

where n_i is the number of observations in sub-sample i.

Estimation by m-fold Cross Validation

m-fold cross validation:

1. Estimate the MSPE by dividing the data set into two parts: an "estimation" subsample and a "test" subsample used to simulate out-of-sample prediction.
2. Use the combined sub-samples $2, 3, \ldots, m$ to compute $\tilde{\beta}$, an estimate of β.
3. Use $\tilde{\beta}$ to compute predicted values $\hat{Y}$ and prediction errors $Y - \hat{Y}$ for sub-sample 1.
4. Using sub-sample 1 as the test sample, estimate the MSPE with the predicted values in sub-sample 1.
5. Repeat steps 2-4 leaving out sub-sample 2, then 3, ..., then m.
6. The m-fold cross-validation estimator of the MSPE is estimated by averaging these m sub-sample estimates of the MSPE.

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

where n_i is the number of observations in sub-sample i.

Estimation by m-fold Cross Validation

The tradeoff

- Choosing the value of m involves a tradeoff between efficiency of the estimators and computational requirements.
- **More observations:** A larger value of m produces more efficient estimators of β, because more observations are used to estimate β. **leave-one-out cross-validation estimator:** Set $m = n - 1$. This maximizes the number of observations used.
- **More computations:** A larger value of m implies that β must be estimated m times. The leave-one-out cross validation may demand too much computational power.
- **School test score application:** A compromise value $m = 10$ is selected, meaning that each sub-sample estimator of β uses 90% of the sample.

Estimation by m-fold Cross Validation

The tradeoff

- Choosing the value of m involves a tradeoff between efficiency of the estimators and computational requirements.
- **More observations:** A larger value of m produces more efficient estimators of β, because more observations are used to estimate β. **leave-one-out cross-validation estimator:** Set $m = n - 1$. This maximizes the number of observations used.
- **More computations:** A larger value of m implies that β must be estimated m times. The leave-one-out cross validation may demand too much computational power.
- **School test score application:** A compromise value $m = 10$ is selected, meaning that each sub-sample estimator of β uses 90% of the sample.

Estimation by m-fold Cross Validation

The tradeoff

- Choosing the value of m involves a tradeoff between efficiency of the estimators and computational requirements.
- **More observations:** A larger value of m produces more efficient estimators of β, because more observations are used to estimate β. **leave-one-out cross-validation estimator:** Set $m = n - 1$. This maximizes the number of observations used.
- **More computations:** A larger value of m implies that β must be estimated m times. The leave-one-out cross validation may demand too much computational power.
- **School test score application:** A compromise value $m = 10$ is selected, meaning that each sub-sample estimator of β uses 90% of the sample.

Estimation by m-fold Cross Validation

The tradeoff

- Choosing the value of m involves a tradeoff between efficiency of the estimators and computational requirements.
- **More observations:** A larger value of m produces more efficient estimators of β, because more observations are used to estimate β. **leave-one-out cross-validation estimator:** Set $m = n - 1$. This maximizes the number of observations used.
- **More computations:** A larger value of m implies that β must be estimated m times. The leave-one-out cross validation may demand too much computational power.
- **School test score application:** A compromise value $m = 10$ is selected, meaning that each sub-sample estimator of β uses 90% of the sample.

Ridge Regression

Ridge Regression

Penalized Sum of Squared Residuals

- **Penalty:** To shrink the estimated coefficients toward 0, penalize large values of the estimate.
- **Ridge regression estimator:** Minimizes the penalized sum of squares —the sum of squared residuals plus a penalty factor that increases with the sum of the squared coefficients:

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

- **Ridge shrinkage parameter:** $\lambda_{\text{Ridge}} \geq 0$.
- First term: Sum of squared residuals for candidate estimator b.
- Second term: Penalizes the estimator for choosing a large estimate of the coefficient.
- In the special case that the regressors are uncorrelated, the ridge regression estimator is:

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

where $\hat{\beta}_j$ is the OLS estimator of β_j.

Ridge Regression

Penalized Sum of Squared Residuals

- **Penalty:** To shrink the estimated coefficients toward 0, penalize large values of the estimate.
- **Ridge regression estimator:** Minimizes the penalized sum of squares —the sum of squared residuals plus a penalty factor that increases with the sum of the squared coefficients:

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

- **Ridge shrinkage parameter:** $\lambda_{\text{Ridge}} \geq 0$.
- First term: Sum of squared residuals for candidate estimator b.
- Second term: Penalizes the estimator for choosing a large estimate of the coefficient.
- In the special case that the regressors are uncorrelated, the ridge regression estimator is:

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

where $\hat{\beta}_j$ is the OLS estimator of β_j.

Ridge Regression

Penalized Sum of Squared Residuals

- **Penalty:** To shrink the estimated coefficients toward 0, penalize large values of the estimate.
- **Ridge regression estimator:** Minimizes the penalized sum of squares —the sum of squared residuals plus a penalty factor that increases with the sum of the squared coefficients:

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

- **Ridge shrinkage parameter:** $\lambda_{\text{Ridge}} \geq 0$.
- First term: Sum of squared residuals for candidate estimator b.
- Second term: Penalizes the estimator for choosing a large estimate of the coefficient.
- In the special case that the regressors are uncorrelated, the ridge regression estimator is:

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

where $\hat{\beta}_j$ is the OLS estimator of β_j.

Ridge Regression

Penalized Sum of Squared Residuals

- **Penalty:** To shrink the estimated coefficients toward 0, penalize large values of the estimate.
- **Ridge regression estimator:** Minimizes the penalized sum of squares —the sum of squared residuals plus a penalty factor that increases with the sum of the squared coefficients:

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

- **Ridge shrinkage parameter:** $\lambda_{\text{Ridge}} \geq 0$.
- First term: Sum of squared residuals for candidate estimator b.
- Second term: Penalizes the estimator for choosing a large estimate of the coefficient.
- In the special case that the regressors are uncorrelated, the ridge regression estimator is:

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

where $\hat{\beta}_j$ is the OLS estimator of β_j.

Ridge Regression

Penalized Sum of Squared Residuals

- **Penalty:** To shrink the estimated coefficients toward 0, penalize large values of the estimate.
- **Ridge regression estimator:** Minimizes the penalized sum of squares —the sum of squared residuals plus a penalty factor that increases with the sum of the squared coefficients:

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

- **Ridge shrinkage parameter:** $\lambda_{\text{Ridge}} \geq 0$.
- First term: Sum of squared residuals for candidate estimator b.
- Second term: Penalizes the estimator for choosing a large estimate of the coefficient.
- In the special case that the regressors are uncorrelated, the ridge regression estimator is:

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

where $\hat{\beta}_j$ is the OLS estimator of β_j.

Ridge Regression

Penalized Sum of Squared Residuals

- **Penalty:** To shrink the estimated coefficients toward 0, penalize large values of the estimate.
- **Ridge regression estimator:** Minimizes the penalized sum of squares —the sum of squared residuals plus a penalty factor that increases with the sum of the squared coefficients:

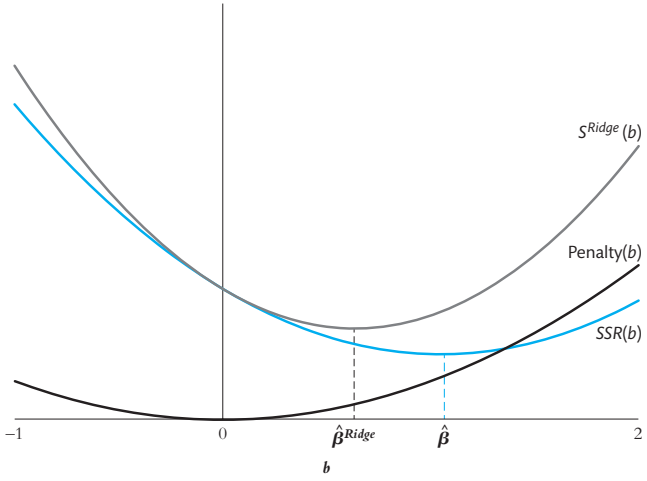
$$S^{\text{Ridge}}(b; \lambda_{\text{Ridge}}) = \sum_{i=1}^{n}(Y_i - b_1 X_{1i} - \ldots - b_k X_{ki})^2 + \lambda_{\text{Ridge}} \sum_{j=1}^{k} b_j^2 \rightarrow \min$$

- **Ridge shrinkage parameter:** $\lambda_{\text{Ridge}} \geq 0$.
- First term: Sum of squared residuals for candidate estimator b.
- Second term: Penalizes the estimator for choosing a large estimate of the coefficient.
- In the special case that the regressors are uncorrelated, the ridge regression estimator is:

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

where $\hat{\beta}_j$ is the OLS estimator of β_j.

Ridge Regression: Penalty Function



The ridge regression estimator minimizes $S^{\text{Ridge}}(b)$, 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, $\hat{\beta}^{\text{Ridge}}$, toward 0.

Ridge Regression: Shrinkage

Choosing the Shrinkage Parameter

- Choose λ_{Ridge} to minimize the estimated MSPE, using the m-fold cross-validation estimator of the MSPE.
- Suppose you have two candidate values 0.1 and 0.2. Let $\tilde{\beta}$ be the Ridge estimator for a given value of λ_{Ridge}. Compute the predictions in the test sample and corresponding $\widehat{MSPE}$. Compare the values of MSPE obtained for $\lambda_{\text{Ridge}} = 0.1$ and $\lambda_{\text{Ridge}} = 0.2$. Select the smaller of the two values.
- Typically the Ridge estimator that minimizes the m-fold cross-validation MSPE differs from the OLS estimator.

Ridge Regression: Shrinkage

Choosing the Shrinkage Parameter

- Choose λ_{Ridge} to minimize the estimated MSPE, using the m-fold cross-validation estimator of the MSPE.
- Suppose you have two candidate values 0.1 and 0.2. Let $\tilde{\beta}$ be the Ridge estimator for a given value of λ_{Ridge}. Compute the predictions in the test sample and corresponding $\widehat{MSPE}$. Compare the values of MSPE obtained for $\lambda_{\text{Ridge}} = 0.1$ and $\lambda_{\text{Ridge}} = 0.2$. Select the smaller of the two values.
- Typically the Ridge estimator that minimizes the m-fold cross-validation MSPE differs from the OLS estimator.

Ridge Regression: Shrinkage

Choosing the Shrinkage Parameter

- Choose λ_{Ridge} to minimize the estimated MSPE, using the m-fold cross-validation estimator of the MSPE.
- Suppose you have two candidate values 0.1 and 0.2. Let $\tilde{\beta}$ be the Ridge estimator for a given value of λ_{Ridge}. Compute the predictions in the test sample and corresponding $\widehat{MSPE}$. Compare the values of MSPE obtained for $\lambda_{\text{Ridge}} = 0.1$ and $\lambda_{\text{Ridge}} = 0.2$. Select the smaller of the two values.
- Typically the Ridge estimator that minimizes the m-fold cross-validation MSPE differs from the OLS estimator.

Ridge Regression for School Test Scores

- Fit a predictive model for school test scores using 817 predictors with $1,966$ observations.
- The square root of the MSPE estimates the magnitude of a typical out-of-sample prediction error.
- 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}.
- The MSPE is minimized for $\hat{\lambda}_{\text{Ridge}} = 2,233$.
- The square-root of MSPE evaluated at $\lambda_{\text{Ridge}} = \hat{\lambda}_{\text{Ridge}}$ is about 39.5.
- The square-root of MSPE evaluated at $\lambda_{\text{Ridge}} = 0$ — the OLS estimator — is about 78.2.
- Because $\hat{\lambda}_{\text{Ridge}}$ minimizes the cross-validated MSPE, the cross-validated MSPE evaluated at $\hat{\lambda}_{\text{Ridge}}$ is not an unbiased estimator of the MSPE. We therefore use the remaining $1,966$ observations to obtain an unbiased estimator of the MSPE for ridge regression using $\hat{\lambda}_{\text{Ridge}}$.

Ridge Regression for School Test Scores

- Fit a predictive model for school test scores using 817 predictors with $1,966$ observations.
- The square root of the MSPE estimates the magnitude of a typical out-of-sample prediction error.
- 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}.
- The MSPE is minimized for $\hat{\lambda}_{\text{Ridge}} = 2,233$.
- The square-root of MSPE evaluated at $\lambda_{\text{Ridge}} = \hat{\lambda}_{\text{Ridge}}$ is about 39.5.
- The square-root of MSPE evaluated at $\lambda_{\text{Ridge}} = 0$ — the OLS estimator — is about 78.2.
- Because $\hat{\lambda}_{\text{Ridge}}$ minimizes the cross-validated MSPE, the cross-validated MSPE evaluated at $\hat{\lambda}_{\text{Ridge}}$ is not an unbiased estimator of the MSPE. We therefore use the remaining $1,966$ observations to obtain an unbiased estimator of the MSPE for ridge regression using $\hat{\lambda}_{\text{Ridge}}$.

Ridge Regression for School Test Scores

- Fit a predictive model for school test scores using 817 predictors with $1,966$ observations.
- The square root of the MSPE estimates the magnitude of a typical out-of-sample prediction error.
- 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}.
- The MSPE is minimized for $\hat{\lambda}_{\text{Ridge}} = 2,233$.
- The square-root of MSPE evaluated at $\lambda_{\text{Ridge}} = \hat{\lambda}_{\text{Ridge}}$ is about 39.5.
- The square-root of MSPE evaluated at $\lambda_{\text{Ridge}} = 0$ — the OLS estimator — is about 78.2.
- Because $\hat{\lambda}_{\text{Ridge}}$ minimizes the cross-validated MSPE, the cross-validated MSPE evaluated at $\hat{\lambda}_{\text{Ridge}}$ is not an unbiased estimator of the MSPE. We therefore use the remaining $1,966$ observations to obtain an unbiased estimator of the MSPE for ridge regression using $\hat{\lambda}_{\text{Ridge}}$.

Ridge Regression for School Test Scores

- Fit a predictive model for school test scores using 817 predictors with $1,966$ observations.
- The square root of the MSPE estimates the magnitude of a typical out-of-sample prediction error.
- 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}.
- The MSPE is minimized for $\hat{\lambda}_{\text{Ridge}} = 2,233$.
- The square-root of MSPE evaluated at $\lambda_{\text{Ridge}} = \hat{\lambda}_{\text{Ridge}}$ is about 39.5.
- The square-root of MSPE evaluated at $\lambda_{\text{Ridge}} = 0$ — the OLS estimator — is about 78.2.
- Because $\hat{\lambda}_{\text{Ridge}}$ minimizes the cross-validated MSPE, the cross-validated MSPE evaluated at $\hat{\lambda}_{\text{Ridge}}$ is not an unbiased estimator of the MSPE. We therefore use the remaining $1,966$ observations to obtain an unbiased estimator of the MSPE for ridge regression using $\hat{\lambda}_{\text{Ridge}}$.

Ridge Regression for School Test Scores

- Fit a predictive model for school test scores using 817 predictors with $1,966$ observations.
- The square root of the MSPE estimates the magnitude of a typical out-of-sample prediction error.
- 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}.
- The MSPE is minimized for $\hat{\lambda}_{\text{Ridge}} = 2,233$.
- The square-root of MSPE evaluated at $\lambda_{\text{Ridge}} = \hat{\lambda}_{\text{Ridge}}$ is about 39.5.
- The square-root of MSPE evaluated at $\lambda_{\text{Ridge}} = 0$ — the OLS estimator — is about 78.2.
- Because $\hat{\lambda}_{\text{Ridge}}$ minimizes the cross-validated MSPE, the cross-validated MSPE evaluated at $\hat{\lambda}_{\text{Ridge}}$ is not an unbiased estimator of the MSPE. We therefore use the remaining $1,966$ observations to obtain an unbiased estimator of the MSPE for ridge regression using $\hat{\lambda}_{\text{Ridge}}$.

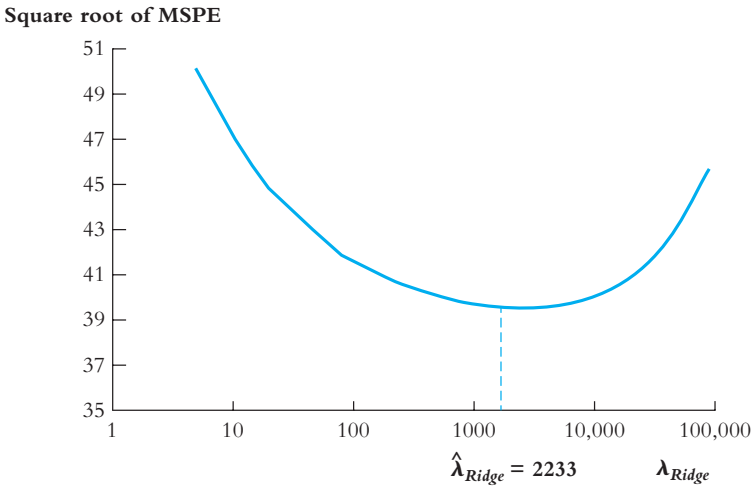
Ridge Regression for School Test Scores

- Fit a predictive model for school test scores using 817 predictors with $1,966$ observations.
- The square root of the MSPE estimates the magnitude of a typical out-of-sample prediction error.
- 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}.
- The MSPE is minimized for $\hat{\lambda}_{\text{Ridge}} = 2,233$.
- The square-root of MSPE evaluated at $\lambda_{\text{Ridge}} = \hat{\lambda}_{\text{Ridge}}$ is about 39.5.
- The square-root of MSPE evaluated at $\lambda_{\text{Ridge}} = 0$ — the OLS estimator — is about 78.2.
- Because $\hat{\lambda}_{\text{Ridge}}$ minimizes the cross-validated MSPE, the cross-validated MSPE evaluated at $\hat{\lambda}_{\text{Ridge}}$ is not an unbiased estimator of the MSPE. We therefore use the remaining $1,966$ observations to obtain an unbiased estimator of the MSPE for ridge regression using $\hat{\lambda}_{\text{Ridge}}$.

Ridge Regression for School Test Scores

- Fit a predictive model for school test scores using 817 predictors with $1,966$ observations.
- The square root of the MSPE estimates the magnitude of a typical out-of-sample prediction error.
- 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}.
- The MSPE is minimized for $\hat{\lambda}_{\text{Ridge}} = 2,233$.
- The square-root of MSPE evaluated at $\lambda_{\text{Ridge}} = \hat{\lambda}_{\text{Ridge}}$ is about 39.5.
- The square-root of MSPE evaluated at $\lambda_{\text{Ridge}} = 0$ — the OLS estimator — is about 78.2.
- Because $\hat{\lambda}_{\text{Ridge}}$ minimizes the cross-validated MSPE, the cross-validated MSPE evaluated at $\hat{\lambda}_{\text{Ridge}}$ is not an unbiased estimator of the MSPE. We therefore use the remaining $1,966$ observations to obtain an unbiased estimator of the MSPE for ridge regression using $\hat{\lambda}_{\text{Ridge}}$.

Square Root of the MSPE for the Ridge Regression Prediction



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

Lasso Regression

Penalized Sum of Squared Residuals

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

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

  where λ_{Lasso} is the Lasso shrinkage parameter.

- **Sparse model:** A regression model in which the coefficients are nonzero for only a small fraction of the predictors. In sparse models, predictions can be improved by estimating many of the coefficients to be exactly 0. The Lasso regression sets many of the estimated coefficients exactly to 0. The regressors it keeps are subject to less shrinkage than with Ridge regression.

Lasso Regression

Penalized Sum of Squared Residuals

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

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

  where λ_{Lasso} is the Lasso shrinkage parameter.

- **Sparse model:** A regression model in which the coefficients are nonzero for only a small fraction of the predictors. In sparse models, predictions can be improved by estimating many of the coefficients to be exactly 0. The Lasso regression sets many of the estimated coefficients exactly to 0. The regressors it keeps are subject to less shrinkage than with Ridge regression.
  - Ridge regression penalty increases with the square b^2.
  - Lasso regression penalty increases with the absolute value $|b|$.
  - For large values of b, the ridge penalty exceeds the Lasso penalty. If the OLS estimate is large, the Lasso shrinks it less than Ridge; if the OLS estimate is small, the Lasso shrinks it more.

Lasso Regression

Penalized Sum of Squared Residuals

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

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

  where λ_{Lasso} is the Lasso shrinkage parameter.

- **Sparse model:** A regression model in which the coefficients are nonzero for only a small fraction of the predictors. In sparse models, predictions can be improved by estimating many of the coefficients to be exactly 0. The Lasso regression sets many of the estimated coefficients exactly to 0. The regressors it keeps are subject to less shrinkage than with Ridge regression.
  - Ridge regression penalty increases with the square b^2.
  - Lasso regression penalty increases with the absolute value $|b|$.
  - For large values of b, the ridge penalty exceeds the Lasso penalty. If the OLS estimate is large, the Lasso shrinks it less than Ridge; if the OLS estimate is small, the Lasso shrinks it more.

Lasso Regression

Penalized Sum of Squared Residuals

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

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

  where λ_{Lasso} is the Lasso shrinkage parameter.

- **Sparse model:** A regression model in which the coefficients are nonzero for only a small fraction of the predictors. In sparse models, predictions can be improved by estimating many of the coefficients to be exactly 0. The Lasso regression sets many of the estimated coefficients exactly to 0. The regressors it keeps are subject to less shrinkage than with Ridge regression.
  - Ridge regression penalty increases with the square b^2.
  - Lasso regression penalty increases with the absolute value $|b|$.
  - For large values of b, the ridge penalty exceeds the Lasso penalty. If the OLS estimate is large, the Lasso shrinks it less than Ridge; if the OLS estimate is small, the Lasso shrinks it more.

Lasso Regression

Penalized Sum of Squared Residuals

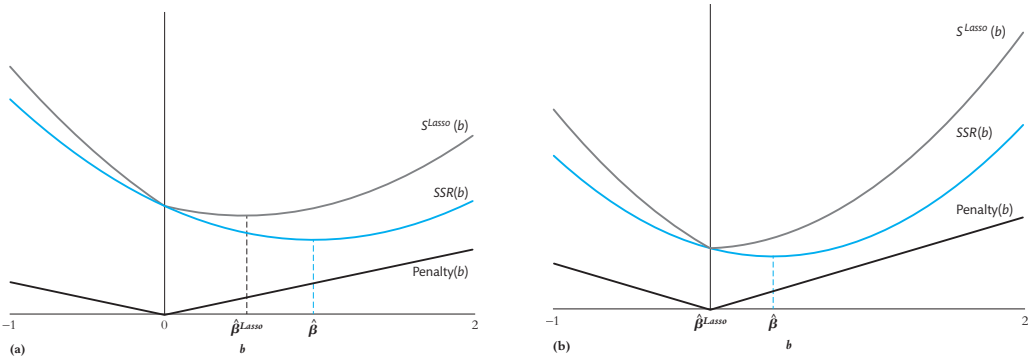
- The Lasso estimator minimizes a penalized sum of squares, where the penalty increases with the sum of the absolute values of the coefficients:

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

  where λ_{Lasso} is the Lasso shrinkage parameter.

- **Sparse model:** A regression model in which the coefficients are nonzero for only a small fraction of the predictors. In sparse models, predictions can be improved by estimating many of the coefficients to be exactly 0. The Lasso regression sets many of the estimated coefficients exactly to 0. The regressors it keeps are subject to less shrinkage than with Ridge regression.
  - Ridge regression penalty increases with the square b^2.
  - Lasso regression penalty increases with the absolute value $|b|$.
  - For large values of b, the ridge penalty exceeds the Lasso penalty. If the OLS estimate is large, the Lasso shrinks it less than Ridge; if the OLS estimate is small, the Lasso shrinks it more.

Lasso Regression: Penalty Function



The Lasso Estimator Minimizes the Sum of Squared Residuals Plus a Penalty. 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.

Lasso Regression: Shrinkage

Choosing the Shrinkage Parameter

- 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 using specialized algorithms. Recent advances in machine learning have made it easier to compute Lasso problems.
- With Ridge and Lasso regressions, the estimated coefficients depend on the specific choice of the linear combination of regressors used. Selecting regressors requires more care because predictions are more sensitive to the choice of regressors.
- Using the California Schools data on test scores yields:

$$\text{MSPE}_{\text{OLS}} = 78.2$$
$$\hat{\lambda}^{\text{Lasso}} = 4,527 \rightarrow \text{MSPE}_{\text{Lasso}} = 39.7$$
$$\hat{\lambda}^{\text{Ridge}} = 2,233 \rightarrow \text{MSPE}_{\text{Ridge}} = 39.5$$

Lasso Regression: Shrinkage

Choosing the Shrinkage Parameter

- 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 using specialized algorithms. Recent advances in machine learning have made it easier to compute Lasso problems.
- With Ridge and Lasso regressions, the estimated coefficients depend on the specific choice of the linear combination of regressors used. Selecting regressors requires more care because predictions are more sensitive to the choice of regressors.
- Using the California Schools data on test scores yields:

$$\text{MSPE}_{\text{OLS}} = 78.2$$

$$\hat{\lambda}^{\text{Lasso}} = 4,527 \rightarrow \text{MSPE}_{\text{Lasso}} = 39.7$$

$$\hat{\lambda}^{\text{Ridge}} = 2,233 \rightarrow \text{MSPE}_{\text{Ridge}} = 39.5$$

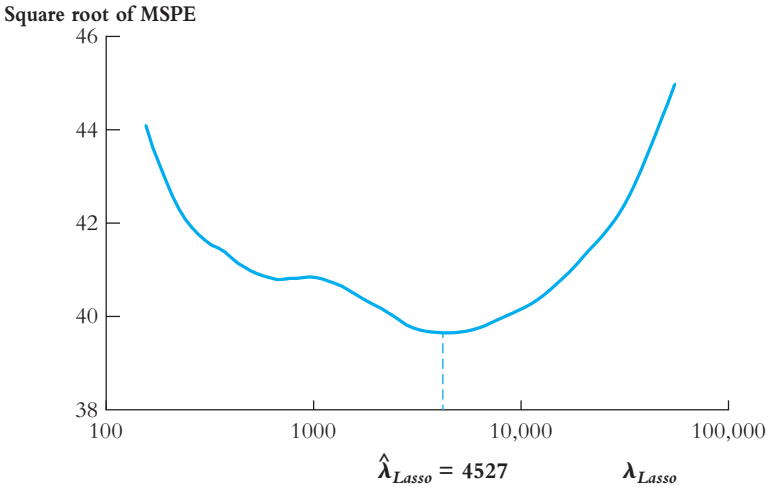
Lasso Regression: Shrinkage

Choosing the Shrinkage Parameter

- 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 using specialized algorithms. Recent advances in machine learning have made it easier to compute Lasso problems.
- With Ridge and Lasso regressions, the estimated coefficients depend on the specific choice of the linear combination of regressors used. Selecting regressors requires more care because predictions are more sensitive to the choice of regressors.
- Using the California Schools data on test scores yields:

$$\begin{aligned} \text{MSPE}_{\text{OLS}} &= 78.2 \\ \hat{\lambda}^{\text{Lasso}} = 4,527 &\rightarrow \text{MSPE}_{\text{Lasso}} = 39.7 \\ \hat{\lambda}^{\text{Ridge}} = 2,233 &\rightarrow \text{MSPE}_{\text{Ridge}} = 39.5 \end{aligned}$$

Square Root of the MSPE for the Lasso Prediction



The MSPE is estimated by 10-fold cross validation using 46 the school test score data set with $k = 817$ predictors and $n = 1966$ observations. The MSPE is minimized when the shrinkage parameter is $\hat{\lambda} = 4,527$, with $\text{MSPE}_{\text{Lasso}} = 39.7 < \text{MSPE}_{\text{OLS}} = 78.2$. The Lasso and Ridge MSPE are similar, $\text{MSPE}_{\text{Ridge}} = 39.5$.

Principal Components

Principal Components

- If two regressors are perfectly collinear, one of them must be dropped — to avoid falling into the dummy variable trap.
- This suggests dropping a variable if it is highly correlated (even imperfectly) with the other regressors
- Principal components analysis implements this strategy — to avoid falling into the "too many variables trap". Linear combinations of variables selected so that the principal components are mutually uncorrelated and keep as much information as possible.
- **Principal Components with 2 Variables:**
  - The linear combination weights for the first principal component are chosen to maximize its variance — to capture as much of the variation as possible.
  - The linear combination weights for the second principal component are chosen to be uncorrelated with the first principal component, and to capture as much of the variance as possible after controlling for the first principal component.
  - The linear combination weights for the third principal component are chosen to be uncorrelated with the first two components and, again, to capture as much of the variance as possible after controlling for the first two components.
  - And so on for the fourth, fifth, ..., all components.

Principal Components

- If two regressors are perfectly collinear, one of them must be dropped — to avoid falling into the dummy variable trap.
- This suggests dropping a variable if it is highly correlated (even imperfectly) with the other regressors
- Principal components analysis implements this strategy — to avoid falling into the "too many variables trap". Linear combinations of variables selected so that the principal components are mutually uncorrelated and keep as much information as possible.
- **Principal Components with 2 Variables:**
  - The linear combination weights for the first principal component are chosen to maximize its variance — to capture as much of the variation as possible.
  - The linear combination weights for the second principal component are chosen to be uncorrelated with the first, also to carry as much as possible of the variance as possible after controlling for the first principal component.
  - The linear combination weights for the third principal component are chosen to be uncorrelated with the first two components and, again, to capture as much of the variance as possible after controlling for the first two components.
  - And so on for the fourth, fifth, ..., nth components.

Principal Components

- If two regressors are perfectly collinear, one of them must be dropped — to avoid falling into the dummy variable trap.
- This suggests dropping a variable if it is highly correlated (even imperfectly) with the other regressors
- Principal components analysis implements this strategy — to avoid falling into the "too many variables trap". Linear combinations of variables selected so that the principal components are mutually uncorrelated and keep as much information as possible.
- **Principal Components with 2 Variables:**
  - The linear combination weights for the first principal component are chosen to maximize its variance — to capture as much of the variation as possible.
  - The linear combination weights for the second principal component are chosen to be uncorrelated with the first principal component, and to capture as much of the variance as possible after controlling for the first principal component.
  - The linear combination weights for the third principal component are chosen to be uncorrelated with the first two components and, again, to capture as much of the variance as possible after controlling for the first two components.
  - And so on for the fourth, fifth, ..., nth components.

Principal Components

- If two regressors are perfectly collinear, one of them must be dropped — to avoid falling into the dummy variable trap.
- This suggests dropping a variable if it is highly correlated (even imperfectly) with the other regressors
- Principal components analysis implements this strategy — to avoid falling into the "too many variables trap". Linear combinations of variables selected so that the principal components are mutually uncorrelated and keep as much information as possible.
- **Principal Components with 2 Variables:**
  - The linear combination weights for the first principal component are chosen to maximize its variance — to capture as much of the variation as possible.
  - The linear combination weights for the second principal component are chosen to be uncorrelated with the first principal component and to capture as much of the variance as possible after controlling for the first principal component.
  - The linear combination weights for the third principal component are chosen to be uncorrelated with the first two components and, again, to capture as much of the variance as possible, after controlling for the first two components.
  - And again for the fourth, fifth, ..., nth components.

Principal Components

- If two regressors are perfectly collinear, one of them must be dropped — to avoid falling into the dummy variable trap.
- This suggests dropping a variable if it is highly correlated (even imperfectly) with the other regressors
- Principal components analysis implements this strategy — to avoid falling into the "too many variables trap". Linear combinations of variables selected so that the principal components are mutually uncorrelated and keep as much information as possible.
- **Principal Components with 2 Variables:**
  - The linear combination weights for the first principal component are chosen to maximize its variance — to capture as much of the variation as possible.
  - The linear combination weights for the second principal component are chosen to be uncorrelated with the first principal component and to capture as much of the variance as possible after controlling for the first principal component.
  - The linear combination weights for the third principal component are chosen to be uncorrelated with the first two components and, again, to capture as much of the variance as possible, after controlling for the first two components.
  - And again for the fourth, fifth, ..., nth components.

Principal Components

- If two regressors are perfectly collinear, one of them must be dropped — to avoid falling into the dummy variable trap.
- This suggests dropping a variable if it is highly correlated (even imperfectly) with the other regressors
- Principal components analysis implements this strategy — to avoid falling into the "too many variables trap". Linear combinations of variables selected so that the principal components are mutually uncorrelated and keep as much information as possible.
- **Principal Components with 2 Variables:**
  - The linear combination weights for the first principal component are chosen to maximize its variance — to capture as much of the variation as possible.
  - The linear combination weights for the second principal component are chosen to be uncorrelated with the first principal component and to capture as much of the variance as possible after controlling for the first principal component.
  - The linear combination weights for the third principal component are chosen to be uncorrelated with the first two components and, again, to capture as much of the variance as possible, after controlling for the first two components.
  - And again for the fourth, fifth, ..., nth components.

Principal Components

- If two regressors are perfectly collinear, one of them must be dropped — to avoid falling into the dummy variable trap.
- This suggests dropping a variable if it is highly correlated (even imperfectly) with the other regressors
- Principal components analysis implements this strategy — to avoid falling into the "too many variables trap". Linear combinations of variables selected so that the principal components are mutually uncorrelated and keep as much information as possible.
- **Principal Components with 2 Variables:**
  - The linear combination weights for the first principal component are chosen to maximize its variance — to capture as much of the variation as possible.
  - The linear combination weights for the second principal component are chosen to be uncorrelated with the first principal component and to capture as much of the variance as possible after controlling for the first principal component.
  - The linear combination weights for the third principal component are chosen to be uncorrelated with the first two components and, again, to capture as much of the variance as possible, after controlling for the first two components.
  - And again for the fourth, fifth, ..., nth components.

Principal Components

- If two regressors are perfectly collinear, one of them must be dropped — to avoid falling into the dummy variable trap.
- This suggests dropping a variable if it is highly correlated (even imperfectly) with the other regressors
- Principal components analysis implements this strategy — to avoid falling into the "too many variables trap". Linear combinations of variables selected so that the principal components are mutually uncorrelated and keep as much information as possible.
- **Principal Components with 2 Variables:**
  - The linear combination weights for the first principal component are chosen to maximize its variance — to capture as much of the variation as possible.
  - The linear combination weights for the second principal component are chosen to be uncorrelated with the first principal component and to capture as much of the variance as possible after controlling for the first principal component.
  - The linear combination weights for the third principal component are chosen to be uncorrelated with the first two components and, again, to capture as much of the variance as possible, after controlling for the first two components.
  - And again for the fourth, fifth, ..., nth components.

Principal Components: Two Variables

- Let X_1, X_2 be standard normal random variables with population correlation $\rho = 0.7$.
- The first principal component is the weighted average, $PC_1 = w_1 X_1 + w_2 X_2$, with the maximum variance, where w_1 and w_2 are the principal component weights.
- The second principal component is chosen to be uncorrelated with the first. This minimizes the spread of the variables.
- 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.
- Together the two principal components explain all of the variance of X. The fraction of the total variance explained by the principal components are:

$$\frac{\text{var}(PC_1)}{\text{var}(X_1) + \text{var}(X_2)} \quad \text{and} \quad \frac{\text{var}(PC_2)}{\text{var}(X_1) + \text{var}(X_2)}$$

- The variances are $\text{var}(PC_1) = 1 + |\rho|$ and $\text{var}(PC_2) = 1 - |\rho|$, where $\text{cov}(X_1, X_2) = \rho$.

Principal Components: Two Variables

- Let X_1, X_2 be standard normal random variables with population correlation $\rho = 0.7$.
- The first principal component is the weighted average, $PC_1 = w_1 X_1 + w_2 X_2$, with the maximum variance, where w_1 and w_2 are the principal component weights.
- The second principal component is chosen to be uncorrelated with the first. This minimizes the spread of the variables.
- 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.
- Together the two principal components explain all of the variance of X. The fraction of the total variance explained by the principal components are:

$$\frac{\text{var}(PC_1)}{\text{var}(X_1) + \text{var}(X_2)} \quad \text{and} \quad \frac{\text{var}(PC_2)}{\text{var}(X_1) + \text{var}(X_2)}$$

- The variances are $\text{var}(PC_1) = 1 + |\rho|$ and $\text{var}(PC_2) = 1 - |\rho|$, where $\text{cov}(X_1, X_2) = \rho$.

Principal Components: Two Variables

- Let X_1, X_2 be standard normal random variables with population correlation $\rho = 0.7$.
- The first principal component is the weighted average, $PC_1 = w_1 X_1 + w_2 X_2$, with the maximum variance, where w_1 and w_2 are the principal component weights.
- The second principal component is chosen to be uncorrelated with the first. This minimizes the spread of the variables.
- 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.
- Together the two principal components explain all of the variance of X. The fraction of the total variance explained by the principal components are:

$$\frac{\text{var}(PC_1)}{\text{var}(X_1) + \text{var}(X_2)} \quad \text{and} \quad \frac{\text{var}(PC_2)}{\text{var}(X_1) + \text{var}(X_2)}$$

- The variances are $\text{var}(PC_1) = 1 + |\rho|$ and $\text{var}(PC_2) = 1 - |\rho|$, where $\text{cov}(X_1, X_2) = \rho$.

Principal Components: Two Variables

- Let X_1, X_2 be standard normal random variables with population correlation $\rho = 0.7$.
- The first principal component is the weighted average, $PC_1 = w_1 X_1 + w_2 X_2$, with the maximum variance, where w_1 and w_2 are the principal component weights.
- The second principal component is chosen to be uncorrelated with the first. This minimizes the spread of the variables.
- 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.
- Together the two principal components explain all of the variance of X. The fraction of the total variance explained by the principal components are:

$$\frac{\text{var}(PC_1)}{\text{var}(X_1) + \text{var}(X_2)} \quad \text{and} \quad \frac{\text{var}(PC_2)}{\text{var}(X_1) + \text{var}(X_2)}$$

- The variances are $\text{var}(PC_1) = 1 + |\rho|$ and $\text{var}(PC_2) = 1 - |\rho|$, where $\text{cov}(X_1, X_2) = \rho$.

Principal Components: Two Variables

- Let X_1, X_2 be standard normal random variables with population correlation $\rho = 0.7$.
- The first principal component is the weighted average, $PC_1 = w_1 X_1 + w_2 X_2$, with the maximum variance, where w_1 and w_2 are the principal component weights.
- The second principal component is chosen to be uncorrelated with the first. This minimizes the spread of the variables.
- 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.
- Together the two principal components explain all of the variance of X. The fraction of the total variance explained by the principal components are:

$$\frac{\text{var}(PC_1)}{\text{var}(X_1) + \text{var}(X_2)} \quad \text{and} \quad \frac{\text{var}(PC_2)}{\text{var}(X_1) + \text{var}(X_2)}$$

- The variances are $\text{var}(PC_1) = 1 + |\rho|$ and $\text{var}(PC_2) = 1 - |\rho|$, where $\text{cov}(X_1, X_2) = \rho$.

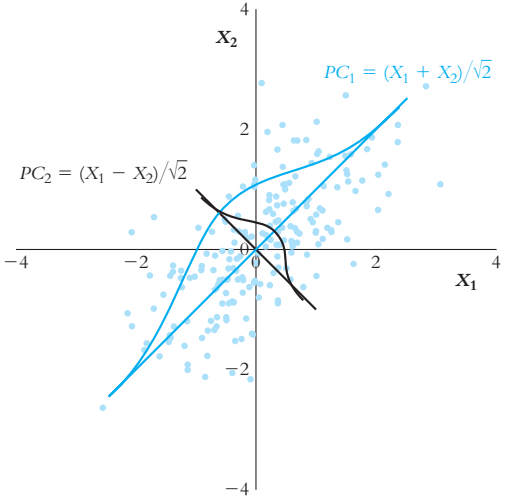
Principal Components: Two Variables

- Let X_1, X_2 be standard normal random variables with population correlation $\rho = 0.7$.
- The first principal component is the weighted average, $PC_1 = w_1 X_1 + w_2 X_2$, with the maximum variance, where w_1 and w_2 are the principal component weights.
- The second principal component is chosen to be uncorrelated with the first. This minimizes the spread of the variables.
- 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.
- Together the two principal components explain all of the variance of X. The fraction of the total variance explained by the principal components are:

$$\frac{\text{var}(PC_1)}{\text{var}(X_1) + \text{var}(X_2)} \quad \text{and} \quad \frac{\text{var}(PC_2)}{\text{var}(X_1) + \text{var}(X_2)}$$

- The variances are $\text{var}(PC_1) = 1 + |\rho|$ and $\text{var}(PC_2) = 1 - |\rho|$, where $\text{cov}(X_1, X_2) = \rho$.

First and Second Principal Components



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. The spread of the variables is greatest in the direction of the $45°$ line. Along the $45°$ line, the weights are equal, so $w_1 = w_2 = 1/\sqrt{2}$ and $PC_1 = (X_1 + X_2)/\sqrt{2}$. The first component explains $(1+\rho)/2 = 85\%$ − The second component explains 15%.

Principal Components

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

1. The squared weights of the linear combinations sum to 1.
2. The first principal component maximizes the variance of its linear combination.
3. The second principal component maximizes the variance of its linear combination, subject to its being uncorrelated with the first principal component.
4. The jth principal component maximizes the variance of its linear combination, subject to its being uncorrelated with the first $j-1$ principal components.

- 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 Xs:

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

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

Principal Components

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

1. The squared weights of the linear combinations sum to 1.
2. The first principal component maximizes the variance of its linear combination.
3. The second principal component maximizes the variance of its linear combination, subject to its being uncorrelated with the first principal component.
4. The jth principal component maximizes the variance of its linear combination, subject to its being uncorrelated with the first $j-1$ principal components.

- 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 Xs:

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

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

Principal Components

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

1. The squared weights of the linear combinations sum to 1.
2. The first principal component maximizes the variance of its linear combination.
3. The second principal component maximizes the variance of its linear combination, subject to its being uncorrelated with the first principal component.
4. The jth principal component maximizes the variance of its linear combination, subject to its being uncorrelated with the first $j-1$ principal components.

- 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 Xs:

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

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

Principal Components

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

1. The squared weights of the linear combinations sum to 1.
2. The first principal component maximizes the variance of its linear combination.
3. The second principal component maximizes the variance of its linear combination, subject to its being uncorrelated with the first principal component.
4. The jth principal component maximizes the variance of its linear combination, subject to its being uncorrelated with the first $j-1$ principal components.

- 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 Xs:

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

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

Principal Components

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

1. The squared weights of the linear combinations sum to 1.
2. The first principal component maximizes the variance of its linear combination.
3. The second principal component maximizes the variance of its linear combination, subject to its being uncorrelated with the first principal component.
4. The jth principal component maximizes the variance of its linear combination, subject to its being uncorrelated with the first $j-1$ principal components.

- 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 Xs:

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

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

Principal Components

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

1. The squared weights of the linear combinations sum to 1.
2. The first principal component maximizes the variance of its linear combination.
3. The second principal component maximizes the variance of its linear combination, subject to its being uncorrelated with the first principal component.
4. The jth principal component maximizes the variance of its linear combination, subject to its being uncorrelated with the first $j-1$ principal components.

- 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 Xs:

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

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

Principal Components

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

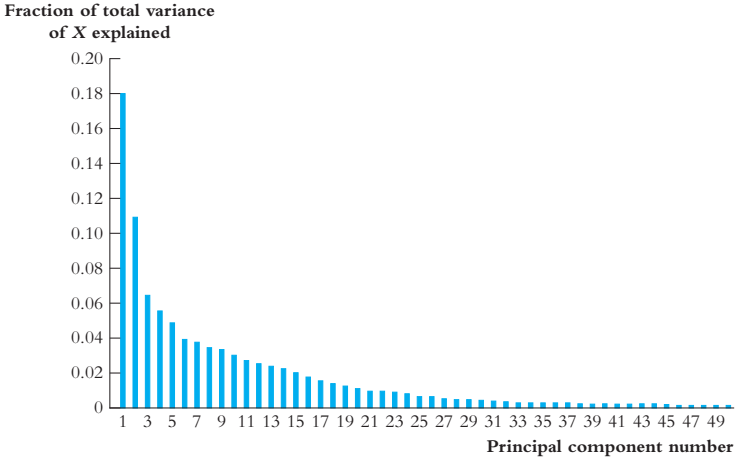
1. The squared weights of the linear combinations sum to 1.
2. The first principal component maximizes the variance of its linear combination.
3. The second principal component maximizes the variance of its linear combination, subject to its being uncorrelated with the first principal component.
4. The jth principal component maximizes the variance of its linear combination, subject to its being uncorrelated with the first $j-1$ principal components.

- 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 Xs:

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

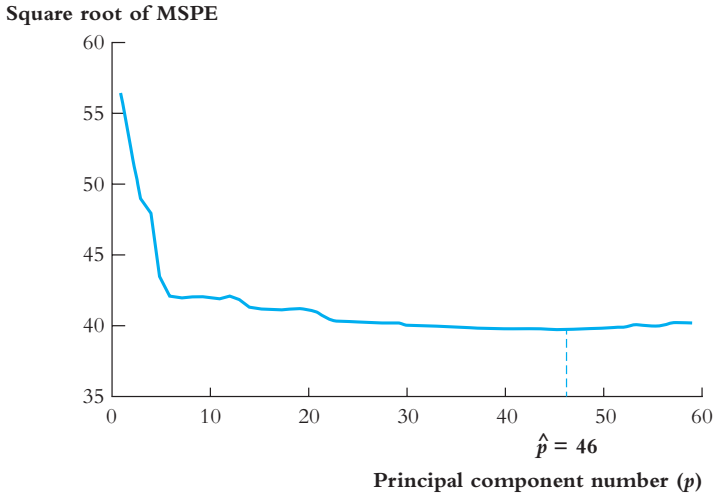
- The ratio $\text{var}(PC_j) / \sum_{j=1}^{k} \text{var}(X_j)$ is the fraction of the total variance of the Xs explained by the jth principal component. This measure is like an R^2 for the total variance of the Xs.

Scree Plot of the 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; the first 10 principal components together explain 63% of the total variance; and the first 40 principal components explain 92% of the total variance. A typical scree looks like a cliff, with boulders, or scree, cascading into a valley.

Square Root of the MSPE for the Principal Components Prediction



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

Principal Components

Application to School Test Scores:

- 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. The cross-validation estimate is $\hat{p} = 46$, with MSPE $=$ 39.7, similar to Lasso and Ridge.

Principal Components

Application to School Test Scores:

- 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. The cross-validation estimate is $\hat{p} = 46$, with MSPE $=$ 39.7, similar to Lasso and Ridge.

Application to School Test Scores:

- 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. The cross-validation estimate is $\hat{p} = 46$, with MSPE $= 39.7$, similar to Lasso and Ridge.

Test Scores With Many Predictors

Prediction Procedure

- 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?
- Predict school test scores using small ($k = 4$), large ($k = 817$), and very large ($k = 2065$) data sets, using OLS, Ridge, Lasso, and PC.
- We reserve half the observations for assessing the performance of the estimated models. Using the 1966 observations in the estimation sample, we estimate the shrinkage parameter by 10-fold cross validation. Using this estimated shrinkage parameter, the regression coefficients are re-estimated using all 1966 observations in the estimation sample. Those estimated coefficients are then used to predict the out-of-sample values for all the observations in the reserved test sample.

Prediction Procedure

- 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?
- Predict school test scores using small ($k = 4$), large ($k = 817$), and very large ($k = 2065$) data sets, using OLS, Ridge, Lasso, and PC.
- We reserve half the observations for assessing the performance of the estimated models. Using the 1966 observations in the estimation sample, we estimate the shrinkage parameter by 10-fold cross validation. Using this estimated shrinkage parameter, the regression coefficients are re-estimated using all 1966 observations in the estimation sample. Those estimated coefficients are then used to predict the out-of-sample values for all the observations in the reserved test sample.

Prediction Procedure

- 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?
- Predict school test scores using small ($k = 4$), large ($k = 817$), and very large ($k = 2065$) data sets, using OLS, Ridge, Lasso, and PC.
- We reserve half the observations for assessing the performance of the estimated models. Using the 1966 observations in the estimation sample, we estimate the shrinkage parameter by 10-fold cross validation. Using this estimated shrinkage parameter, the regression coefficients are re-estimated using all 1966 observations in the estimation sample. Those estimated coefficients are then used to predict the out-of-sample values for all the observations in the reserved test sample.

Prediction Procedure

Standout Features:

1. The MSPE of OLS is much less using the small data set than using the large data set.
2. 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, but not much beyond that.
3. The in-sample estimates of MSPE (the 10-fold cross-validation estimates) are similar to the out-of-sample estimates. This is mainly because the coefficients used for the out-of-sample estimate of the MSPE are estimated using all 1966 observations in the estimation sample.
4. The MSPE in the reserved test sample is generally similar for all the many-predictor methods. The lowest out-of- sample MSPE is obtained using Ridge in the large data set.
5. For the large data set, the many-predictor methods succeed where OLS fails: 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.

Prediction Procedure

Standout Features:

1. The MSPE of OLS is much less using the small data set than using the large data set.
2. 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, but not much beyond that.
3. The in-sample estimates of MSPE (the 10-fold cross-validation estimates) are similar to the out-of-sample estimates. This is mainly because the coefficients used for the out-of-sample estimate of the MSPE are estimated using all 1966 observations in the estimation sample.
4. The MSPE in the reserved test sample is generally similar for all the many-predictor methods. The lowest out-of- sample MSPE is obtained using Ridge in the large data set.
5. For the large data set, the many-predictor methods succeed where OLS fails: 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.

Prediction Procedure

Standout Features:

1. The MSPE of OLS is much less using the small data set than using the large data set.
2. 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, but not much beyond that.
3. The in-sample estimates of MSPE (the 10-fold cross-validation estimates) are similar to the out-of-sample estimates. This is mainly because the coefficients used for the out-of-sample estimate of the MSPE are estimated using all 1966 observations in the estimation sample.
4. The MSPE in the reserved test sample is generally similar for all the many-predictor methods. The lowest out-of- sample MSPE is obtained using Ridge in the large data set.
5. For the large data set, the many-predictor methods succeed where OLS fails: 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.

Prediction Procedure

Standout Features:

1. The MSPE of OLS is much less using the small data set than using the large data set.
2. 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, but not much beyond that.
3. The in-sample estimates of MSPE (the 10-fold cross-validation estimates) are similar to the out-of-sample estimates. This is mainly because the coefficients used for the out-of-sample estimate of the MSPE are estimated using all 1966 observations in the estimation sample.
4. The MSPE in the reserved test sample is generally similar for all the many-predictor methods. The lowest out-of- sample MSPE is obtained using Ridge in the large data set.
5. For the large data set, the many-predictor methods succeed where OLS fails: 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.

Prediction Procedure

Standout Features:

1. The MSPE of OLS is much less using the small data set than using the large data set.
2. 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, but not much beyond that.
3. The in-sample estimates of MSPE (the 10-fold cross-validation estimates) are similar to the out-of-sample estimates. This is mainly because the coefficients used for the out-of-sample estimate of the MSPE are estimated using all 1966 observations in the estimation sample.
4. The MSPE in the reserved test sample is generally similar for all the many-predictor methods. The lowest out-of- sample MSPE is obtained using Ridge in the large data set.
5. For the large data set, the many-predictor methods succeed where OLS fails: 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.

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	Immigration status variables (4)
Age distribution variables in local population (8)	Charter school (binary)
Fraction of local population that is male	School has full-year calendar (binary)
Local population marital status variables (3)	School is in a unified school district (large city) (binary)
Local population educational level variables (4)	School is in Los Angeles (binary)
Fraction of local housing that is owner occupied	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 $= 65 + 60 + 60 + 1770 + 110 = 2065$

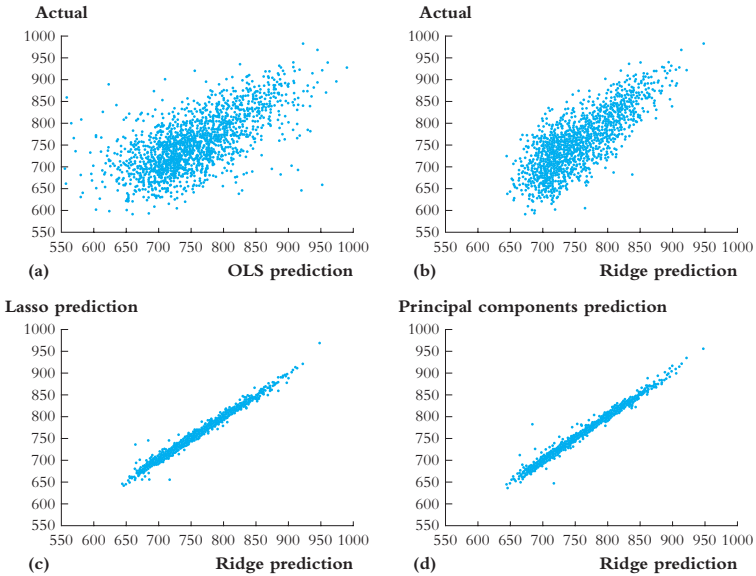
Out-of-Sample Performance of Predictive Models for School Test Scores

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

Coefficients on Selected Standardized Regressors

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

Scatterplots for Out-of-Sample Predictions



Summary

Summary

1. The goal of prediction is to make accurate predictions for out-of-sample observations. The coefficients in prediction models do not have a causal interpretation.
2. OLS works poorly for prediction when the number of regressors is large relative to the sample size.
3. 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.
4. 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.
5. 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.

Summary

1. The goal of prediction is to make accurate predictions for out-of-sample observations. The coefficients in prediction models do not have a causal interpretation.
2. OLS works poorly for prediction when the number of regressors is large relative to the sample size.
3. 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.
4. 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.
5. 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.

Summary

1. The goal of prediction is to make accurate predictions for out-of-sample observations. The coefficients in prediction models do not have a causal interpretation.
2. OLS works poorly for prediction when the number of regressors is large relative to the sample size.
3. 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.
4. 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.
5. 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.

Summary

1. The goal of prediction is to make accurate predictions for out-of-sample observations. The coefficients in prediction models do not have a causal interpretation.
2. OLS works poorly for prediction when the number of regressors is large relative to the sample size.
3. 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.
4. 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.
5. 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.

Summary

1. The goal of prediction is to make accurate predictions for out-of-sample observations. The coefficients in prediction models do not have a causal interpretation.
2. OLS works poorly for prediction when the number of regressors is large relative to the sample size.
3. 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.
4. 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.
5. 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.

Problems & Applications

Problems and Applications

Stock & Watson, Introduction (4th), Chapter 14, Exercise 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 substracting the sample mean from $TestScore$, she estimates the following regression:

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

Problems and Applications

Stock & Watson, Introduction (4th), Chapter 14, Exercise 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 substracting the sample mean from $TestScore$, she estimates the following regression:

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

Problems and Applications

Stock & Watson, Introduction (4th), Chapter 14, Exercise 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 substracting the sample mean from $TestScore$, she estimates the following regression:

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

Problems and Applications

Stock & Watson, Introduction (4th), Chapter 14, Exercise 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 substracting the sample mean from $TestScore$, she estimates the following regression:

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

Problems and Applications

Stock & Watson, Introduction (4th), Chapter 14, Exercise 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 substracting the sample mean from $TestScore$, she estimates the following regression:

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

Problems and Applications

Stock & Watson, Introduction (4th), Chapter 14, Exercise 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 substracting the sample mean from $TestScore$, she estimates the following regression:

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