Linear Methods for Regression

Zichao Yang

Zhongnan University of Economics & Law

Date: December 4, 2023

Roadmap

- Ordinary Least Squares (OLS)
- Subset Selection Methods
- Shrinkage Methods
- A Comparison of the Selection and Shrinkage Methods

Linear Regression Models

A linear regression model assumes that the regression function $E(Y|X)$ is linear in the input $X_1, ..., X_p$.

These models were largely developed in the pre-computer age of statistics. They are simple, but often provide an adequate and **interpretable** description of how the inputs affect the output.

For prediction purposes, linear regression models can sometimes outperform fancier nonlinear models, especially in situations with small number of training cases, low signal-to-noise ratio or sparse data.

Linear Regression Models

Suppose we have an input vector $X^\intercal = \{X_1, X_2, ..., X_p\}$, and want to predict a real-valued output Y. The linear regression model has the form:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

Note: A linear model either assumes that the regression function $E(Y|X)$ is linear, or that the linear model is a reasonable approximation.

Linear Regression Models

The variables X_j can come from different sources:

- quantitative inputs
- transformations of quantitative inputs, such as log, square-root or square
- basis expansions, such as $X_2 = X_1^2, X_3 = X_1^3$, leading to a polynomial representations
- numeric or "dummy" coding of the levels of qualitative inputs
- interactions between variables, like $X_3 = X_1 \cdot X_2$

Remember: no matter the source of the X_j, the model is linear in the parameters.

Ordinary Least Squares

The most popular estimation method is **least squares**, in which we pick the coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^\intercal$ to minimize the residual sum of square. If we apply this method on linear regressions, we usually call it **ordinary least squares (OLS)**.

$$\begin{aligned} \text{RSS}(\boldsymbol{\beta}) &= \sum_{i=1}^{N} (y_i - f(x_i))^2 \\ &= \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \end{aligned}$$

Q: What assumptions should we impose on our data to make the OLS a valid estimation method?

Ordinary Least Squares

Q: What assumptions should we impose on our data to make the OLS a valid estimation method?

(1) From a statistical point of view, the least squares method is valid if the training observations $(\boldsymbol{x}_i, y_i)$ represent independent random draws from their population.

(2) Even if $\boldsymbol{x}_i$ is not drawn randomly, this method is still valid if the y_i is conditionally independent given the input $\boldsymbol{x}_i$.

An Illustration of the OLS method

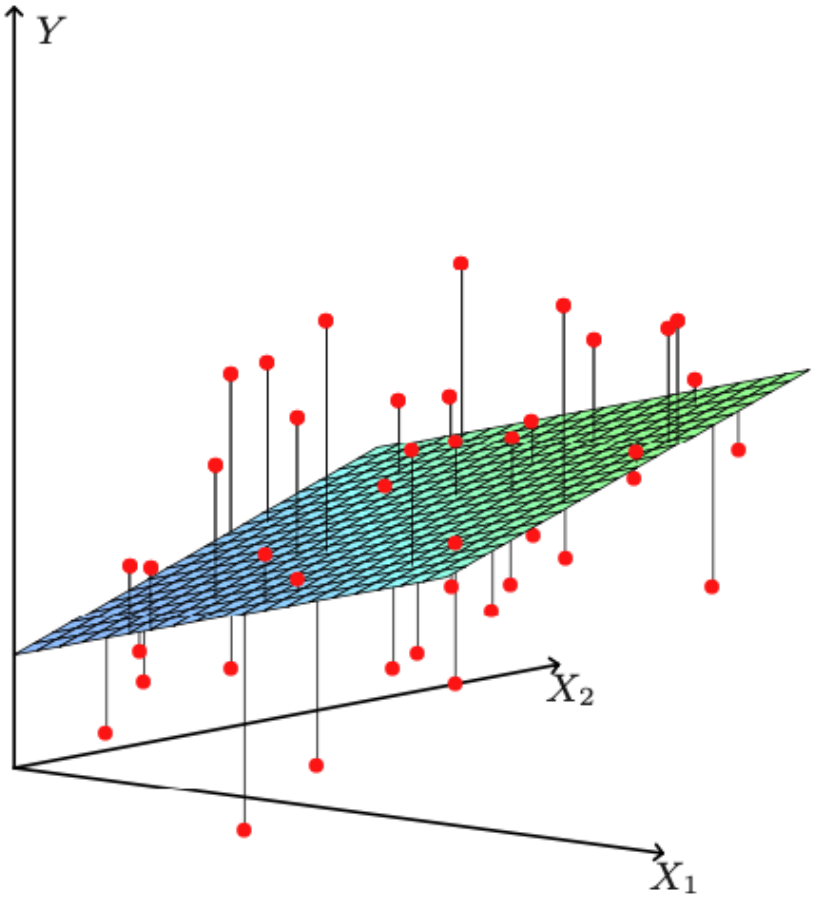


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

Source: Hastie, Trevor, et al. *The elements of statistical learning: data mining, inference, and prediction.* New York: springer, 2009.

OLS Formulas

We can also write the residual sum-of-squares as:

$$\mathrm{RSS}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\intercal(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

This is a quadratic function in the $p + 1$ parameters. Differentiating with respect to $\boldsymbol{\beta}$ we obtain:

$$\frac{\partial \mathrm{RSS}}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^\intercal(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

$$\frac{\partial^2 \mathrm{RSS}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\intercal} = 2\boldsymbol{X}^\intercal\boldsymbol{X}$$

Assuming that $\boldsymbol{X}$ has **full column rank**, and hence $\boldsymbol{X}^\intercal\boldsymbol{X}$ is positive definite, we set the first derivative to zero:

$$\boldsymbol{X}^\intercal(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = 0$$

so we can obtain the unique solution:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{y}$$

OLS Formulas

The fitted values at the training inputs are:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{y}$$

The matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal$ is called the "hat" matrix because it puts the hat on $\boldsymbol{y}$. Or it is called the projection matrix. See the following picture:

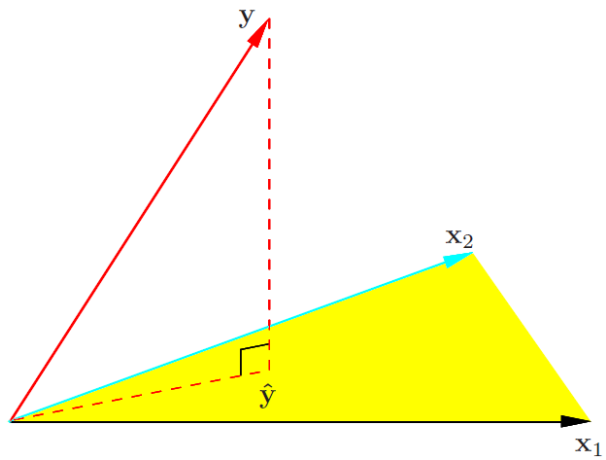


FIGURE 3.2. *The N-dimensional geometry of least squares regression with two predictors. The outcome vector* $\mathbf{y}$ *is orthogonally projected onto the hyperplane spanned by the input vectors* $\mathbf{x}_1$ *and* $\mathbf{x}_2$*. The projection* $\hat{\mathbf{y}}$ *represents the vector of the least squares predictions*

Source: Hastie, Trevor, et al. *The elements of statistical learning: data mining, inference, and prediction.* New York: springer, 2009.

What If $\boldsymbol{X}$ is NOT of Full Rank?

If two of the inputs are perfectly correlated, then $\boldsymbol{X}^\intercal \boldsymbol{X}$ is singular and the coefficient $\hat{\boldsymbol{\beta}}$ are not uniquely defined.

However, the fitted values $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ are still the projection of $\boldsymbol{y}$ onto the column space of $\boldsymbol{X}$; there is just more than one way to express the projection in terms of the column vectors of $\boldsymbol{X}$.

Usually modern regression software packages can automatically drop redundant perfectly correlated columns.

The Gauss-Markov Theorem

The Gauss-Markov Theorem: under certain conditions, the ordinary least squares (OLS) estimator of the coefficients of a linear regression model is the best linear unbiased estimator (**BLUE**), that is, the estimator that has the **smallest variance** among those that are **unbiased** and linear in the observed output variables.

Those conditions: the errors in the linear regression model are uncorrelated, have equal variances and expectation value of zero. The errors do not need to be normal, nor do they need to be independent and identically distributed.

Proof 1: $\hat{\boldsymbol{\beta}}$ is Unbiased

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \boldsymbol{y} \\
&= (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\
&= (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \boldsymbol{\varepsilon} \\
&= \boldsymbol{\beta} + (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \boldsymbol{\varepsilon}
\end{aligned}$$

When we condition on $\boldsymbol{X}$, we can treat $\boldsymbol{X}$ as a constant matrix. Then:

$$\begin{aligned}
E[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] &= \boldsymbol{\beta} + (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal E[\boldsymbol{\varepsilon}|\boldsymbol{X}] \\
&= \boldsymbol{\beta} + (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \mathbf{0} \\
&= \boldsymbol{\beta}
\end{aligned}$$

The Law of Iterated Expectations implies that:

$$E[\hat{\boldsymbol{\beta}}] = E[E[\hat{\boldsymbol{\beta}}|\boldsymbol{X}]] = E[\boldsymbol{\beta}] = \boldsymbol{\beta}$$

Proof 2: Smallest Variance Among Unbiased Estimators

Now we assume another estimator $\tilde{\boldsymbol{\beta}} = \boldsymbol{C}\boldsymbol{y}$, and we define: $\boldsymbol{D} = \boldsymbol{C} - (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal$. Then we can write:

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{C}\boldsymbol{y} = \boldsymbol{D}\boldsymbol{y} + (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \boldsymbol{y} = \boldsymbol{D}\boldsymbol{y} + \hat{\boldsymbol{\beta}}$$

Then we can get:

$$\begin{aligned} E[\tilde{\boldsymbol{\beta}}|\boldsymbol{X}] &= E[\boldsymbol{D}\boldsymbol{y} + \hat{\boldsymbol{\beta}}|\boldsymbol{X}] \\ &= E[\boldsymbol{D}\boldsymbol{y}|\boldsymbol{X}] + E[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] \\ &= E[\boldsymbol{D}\boldsymbol{y}|\boldsymbol{X}] + \boldsymbol{\beta} \\ &= E[\boldsymbol{D}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})|\boldsymbol{X}] + \boldsymbol{\beta} \\ &= \boldsymbol{D}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta} \end{aligned}$$

Hence, if $\tilde{\boldsymbol{\beta}}$ is unbiased, then we should have $\boldsymbol{D}\boldsymbol{X} = 0$

Proof 2: Smallest Variance Among Unbiased Estimators

Now let us calculate the variance of $\tilde{\boldsymbol{\beta}}$:

$$\begin{aligned}
\mathrm{Var}[\tilde{\boldsymbol{\beta}}|X] &= \mathrm{Var}[\boldsymbol{D}\boldsymbol{y} + \hat{\boldsymbol{\beta}}|\boldsymbol{X}] \\
&= \mathrm{Var}[\boldsymbol{D}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}\boldsymbol{\varepsilon} + (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})|\boldsymbol{X}] \\
&\underline{\underline{\boldsymbol{D}\boldsymbol{X}=0}}\ \mathrm{Var}[\boldsymbol{\varepsilon}|\boldsymbol{X}]\boldsymbol{D}\boldsymbol{D}^\intercal + \mathrm{Var}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] \\
&= \sigma^2\boldsymbol{D}\boldsymbol{D}^\intercal + \mathrm{Var}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] \\
&\geq \mathrm{Var}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}]
\end{aligned}$$

So, the $\hat{\boldsymbol{\beta}}$ from OLS is the best linear unbiased estimator (**BLUE**).

Best-Subset Selection

Let p be the total number of predictors. The goal is to find the subset of predictors S with size k, where $0 \leq k \leq p$, that minimizes the chosen criterion:

$$\min_{S,|S|=k} \text{Criterion}(S)$$

where Criterion(S) is the chosen measure of model fit for the subset S. This involves evaluating the criterion for all possible combinations of predictors of size k and selecting the one with the best (minimum) value.

In mathematical terms:

$$\text{Criterion}(S) = \text{Measure of Model Fit}(y, X_S)$$

Here, y is the response variable, and X_S is the matrix of predictors corresponding to the subset S.

Forward-stepwise Selection

The goal of Forward-Stepwise Selection is to iteratively build a model by adding one predictor at each step. Starting with an empty set of predictors, the algorithm selects the predictor that provides the greatest improvement in the chosen criterion, such as the residual sum of squares (RSS) or coefficient of determination (R^2).

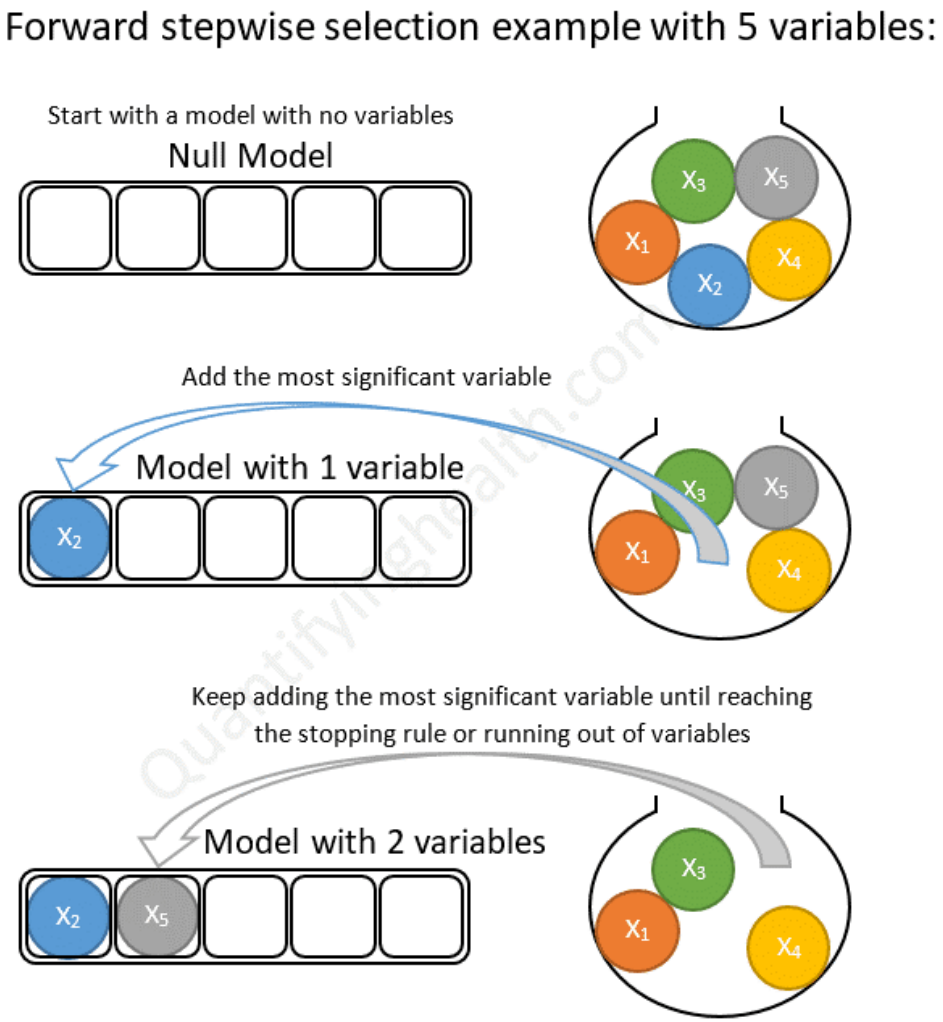
The forward stepwise selection process can be represented mathematically as:

(1) Choose $j_k = \underset{j \notin \text{selected predictors}}{\arg\min}$ Criterion(selected predictors + $\{j\}$)

(2) Add X_{j_k} to the selected predictors

where Criterion(selected predictors) is the chosen measure of model fit, and X_{j_k} is the predictor with index j_k.

Forward-stepwise Selection Illustration



Source: Understand Forward and Backward Stepwise Regression

Backward-stepwise Selection

The goal of Backward-Stepwise Selection is to iteratively build a model by removing one predictor at each step. Starting with the full set of predictors, the algorithm removes the predictor that provides the least contribution in terms of the chosen criterion, such as the residual sum of squares (RSS) or coefficient of determination (R^2).

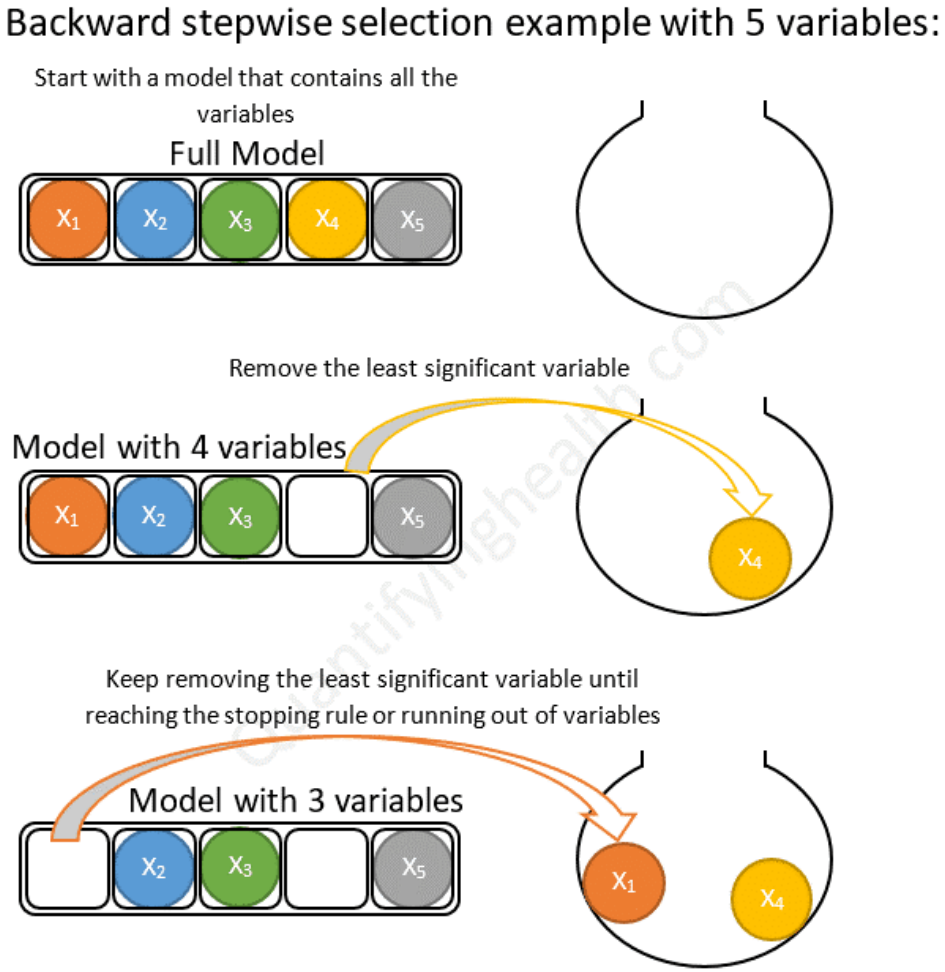
The backward stepwise selection process can be represented mathematically as:

(1) Choose $j_k = \underset{j \in \text{selected predictors}}{\arg\min}$ Criterion(selected predictors $- \{j\}$)

(2) Remove X_{j_k} from the selected predictors

where Criterion(selected predictors) is the chosen measure of model fit, and X_{j_k} is the predictor with index j_k.

Backward-stepwise Selection Illustration



Source: Understand Forward and Backward Stepwise Regression

Shrinkage Methods

Earlier on, we show that the OLS estimator is the best linear unbiased estimator (BLUE), which means it has the smallest variance among all unbiased estimators. What if we want to achieve an even smaller variance?

There is a trade off between unbiasedness and variance, if we are willing to sacrificing some accuracy (unbiasedness) for stability (smaller variance), the following shrinkage methods can be useful:

- Ridge Regression
- Lasso Regression
- Elastic-net Penalty Regression
- Least Angle Regression

Ridge Regression

Ridge regression shrinks the regression coefficients by imposing a penalty on their size. The ridge coefficients minimize a penalized residual sum of squares:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

Here $\lambda \geq 0$ and controls the amount of shrinkage: the larger the λ is, the greater the amount of shrinkage. An equivalent way to write the ridge problem:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2,$$

$$\text{s.t.} \sum_{j=1}^{p} \beta_j^2 \leq t$$

There is a one-to-one correspondence between λ and t.

Ridge Regression

The residual sum-of-squares of the ridge regression is:

$$\text{RSS} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\intercal(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^\intercal\boldsymbol{\beta}$$

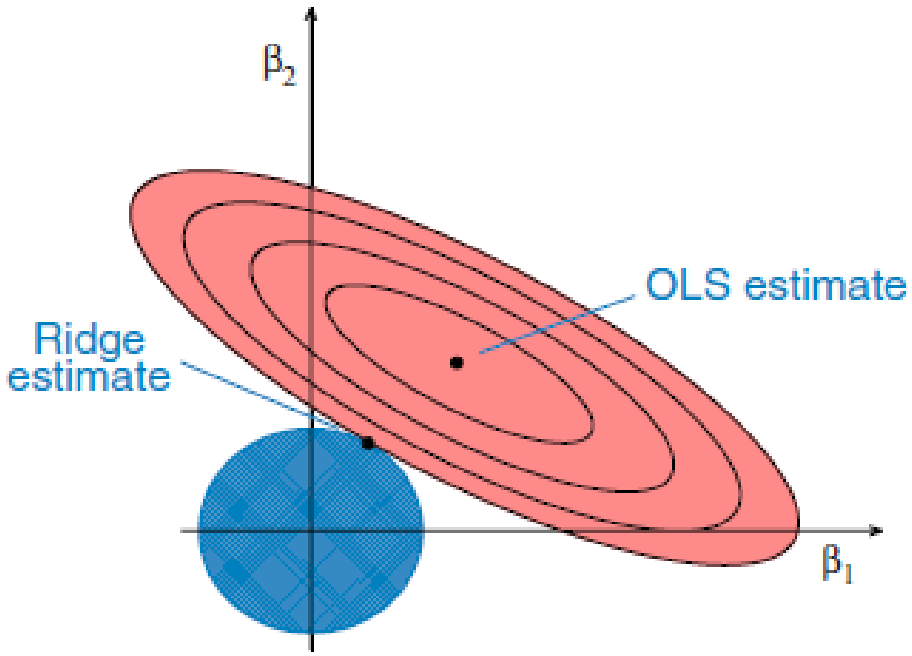
Then the ridge regression solutions are:

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = (\boldsymbol{X}^\intercal\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^\intercal\boldsymbol{y}$$

The solution adds a positive constant to the diagonal of $\boldsymbol{X}^\intercal\boldsymbol{X}$ and makes the problem non-singular, even if $\boldsymbol{X}^\intercal\boldsymbol{X}$ is not of full rank. This was the main motivation for ridge regression was first introduced by Hoerl and Kennard (1970).

In the case of orthonormal inputs, the ridge estimates are just a scaled version of the OLS estimates: $\hat{\boldsymbol{\beta}}^{\text{ridge}} = \hat{\boldsymbol{\beta}}/(1+\lambda)$.

Geometric Interpretation of Ridge Regression



Source: PennState STAT 897D

The ellipses correspond to the contours of residual sum of squares (RSS): the inner ellipse has smaller RSS, and RSS is minimized at ordinal least square (OLS) estimates.

We are trying to minimize the ellipse size and circle simultanously in the ridge regression. The ridge estimate is given by the point at which the ellipse and the circle touch.

Ridge Regression: Standardization

In Ridge Regressions, we need to standardize our inputs because of the following reasons:

- **Scaling Coefficients Equally:** Ridge regression penalizes the sum of squared coefficients. If the variables have different scales, those with larger scales will contribute more to the penalty term.
- **Multicollinearity Mitigation:** Standardizing inputs helps alleviate multicollinearity issues, as the regularization penalty is then applied uniformly to all variables, preventing any single variable from dominating the regression.
- **Algorithm Convergence:** Variables with large scales can cause numerical instability and slow convergence, and standardization helps mitigate these issues and improve model performance.

Ridge Regression: How to do Standardization?

The **standardization** process includes two parts:

- **Centering:** subtract the mean from each variable, resulting in centered data where the mean of each variable is zero.
- **Scaling:** divide variables by their standard deviations, which makes them unitless and bring them to a comparable scale.

$$\boldsymbol{X_i} = \frac{\boldsymbol{X_i} - \mu_i}{\sigma_i}$$

After doing the standardization, the β_0 in original equation actually is no longer needed. We can think that $\beta_0 = \bar{\boldsymbol{y}} = \frac{1}{N}\sum_1^N y_i$, and it has been removed in the standardization procedure.

Ridge Regression: How to do Standardization?

Q: The **standardization** process should be applied on:

(1) the whole dataset

(2) the whole training set (3) the whole test set

(4) only X in the training set (5) only X in the test set

(6) only y in the training set (7) only y in the test set

Ridge Regression: How to do Standardization?

None of them is 100% correct. (Yup, I deceived you!)

Here is the recommended procedure of standardization:

- **Fit the scaler on the training set:** Calculate the mean and standard deviation of the features (only X, not y) in the **training set**.
- **Transform the training set:** Standardize the features in the training set using the calculated mean and standard deviation.
- **Transform the test set:** Use the same scaler to standardize the features in the **test set**.

Let's try it out in python!

Ridge Regression: How to Pick the λ?

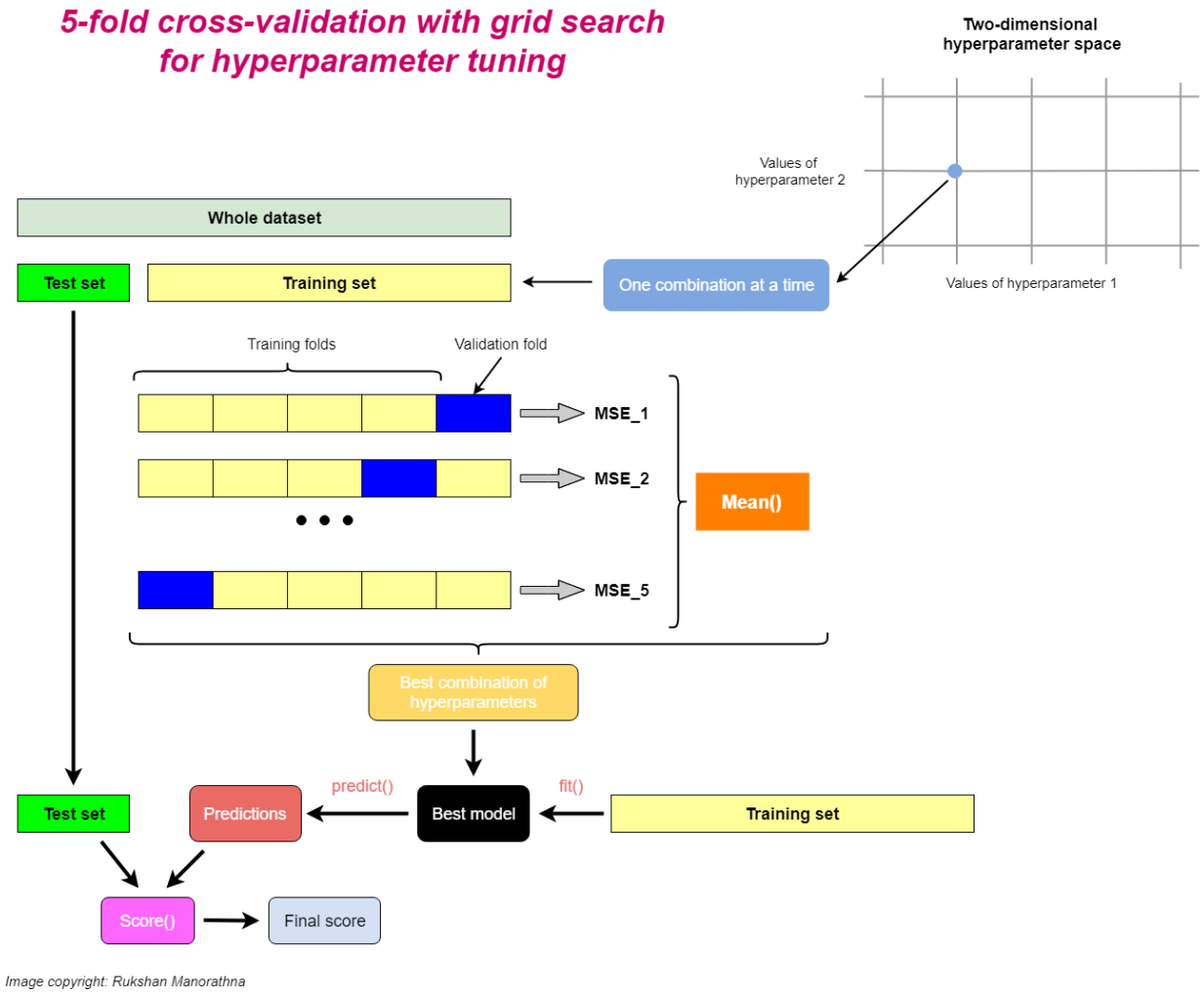
There are two ways to pick the regularization parameter, λ:

1. **AIC, BIC:** A traditional approach would be to choose λ such that some information criterion, e.g., AIC or BIC, is the smallest.

2. **Cross-validation:** A more machine learning-like approach is to perform cross-validation and select the value of λ that minimizes the cross-validated sum of squared residuals (or some other measure). To choose λ through cross-validation, you should choose a set of P different values of λ to test, split the dataset into K folds.

The first approach emphasizes the model's fit to the data, while the latter one is more focused on its predictive performance.

Ridge Regression: How to Pick the λ?



Source: k-fold cross-validation explained in plain English

Ridge Regression: How to Pick the λ?

K-fold cross-validation algorithm

for p in 1 : P:

  for k in 1 : K:

    keep fold k as hold-out data

    use the remaining folds and $\lambda = \lambda_p$ to estimate $\hat{\beta}^{ridge}$

    predict hold-out data: $\hat{y}_{train,k} = \mathbf{X}_{train,k}\hat{\beta}^{ridge}$

    compute a sum of squared residuals: $SSR_k = ||y_{train,k} - \hat{y}_{train,k}||^2$

  end for k

  average SSR over the folds: $SSR_p = \frac{1}{K}\sum_{k=1}^{K} SSR_k$

end for p

choose optimal value: $\lambda_{opt} = \arg\min_p SSR_p$

Let's try it out in python!

Lasso Regression

Similar to the Ridge Regression, the Lasso can be written as:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Equivalent to:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2,$$

$$\text{s.t.} \sum_{j=1}^{p} |\beta_j| \leq t$$

In lasso regression, the penality $\sum_{j=1}^{P} |\beta_j|$ makes the solutions nonlinear in the y_i, and there is no closed form expression as in ridge regression.

Lasso VS. Ridge

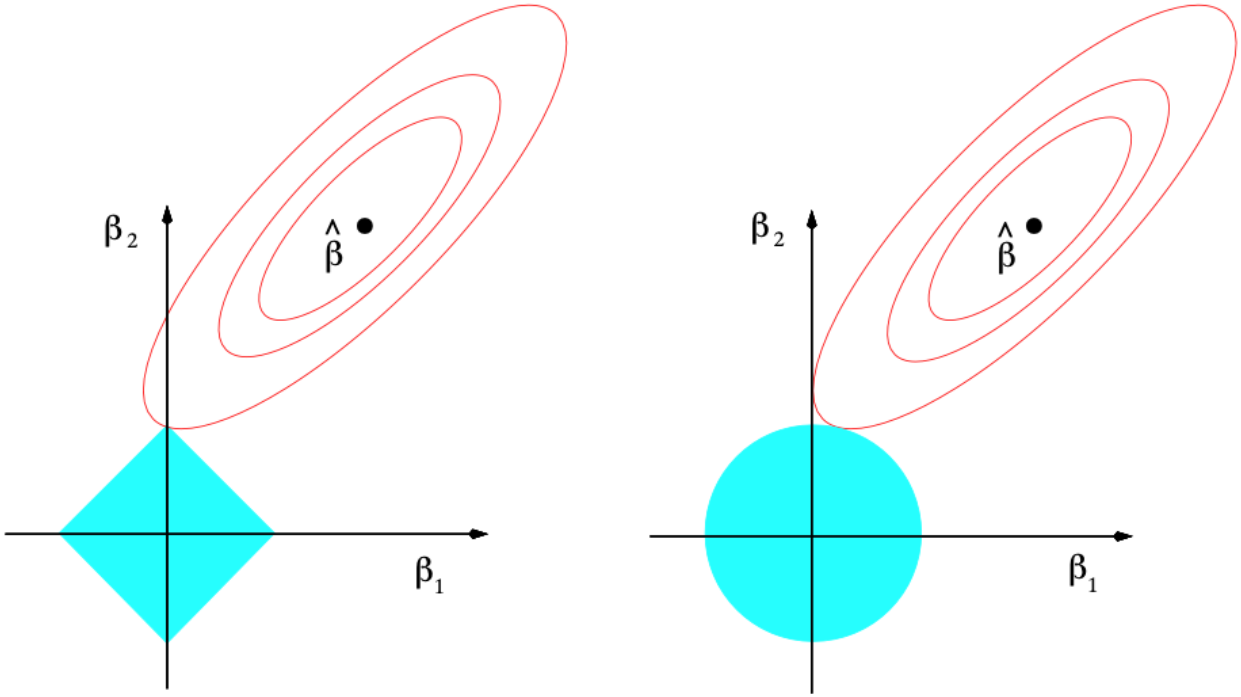


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

Source: Hastie, Trevor, et al. *The elements of statistical learning: data mining, inference, and prediction.* New York: springer, 2009.

Now let's try to apply the lasso regression to the Boston Housing Data.

Lasso VS. Ridge

- **Feature Selection:** Lasso translates each coefficient by a constant factor λ, truncating at 0. It can automatically select a subset of relevant features, effectively performing feature selection during the modeling process. Ridge does a proportional shrinkage. It reduces the impact of less important features, but it generally keeps all features in the model.
- **Handling Multicollinearity:** Lasso naturally selects one variable among highly correlated variables and sets the coefficients of the rest to zero. Ridge regression tends to handle multicollinearity by reducing the impact of correlated variables.
- **Solution Stability:** Ridge regression tends to have more stable solutions in the presence of multicollinearity.

The Group Lasso

In some problems, the predictors belong to pre-defined groups. In this situation it may be desirable to shrink and select the members of a group together. The **group lasso** is one way to achieve this.

Suppose that the p predictors are divided into L groups, with p_l the number in group l. We use $\boldsymbol{X}_l$ to represent the predictors corresponding to the lth group, with corresponding coefficient vector β_l. The group lass minimize the convex criterion:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ ||\boldsymbol{y} - \beta_0 \mathbf{1} - \sum_{l=1}^{L} \boldsymbol{X}_l \beta_l||_2^2 + \lambda \sum_{l=1}^{L} \sqrt{p_l} ||\beta_l||_2 \right\}$$

You can conduct group lasso regression using the Group Lasso Package. Let's see an example.

Beyond Ridge and Lasso

We can generalize ridge and lasso regression, and view them as Bayes estimates. Consider the criterion:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}, \quad q \geq 0$$

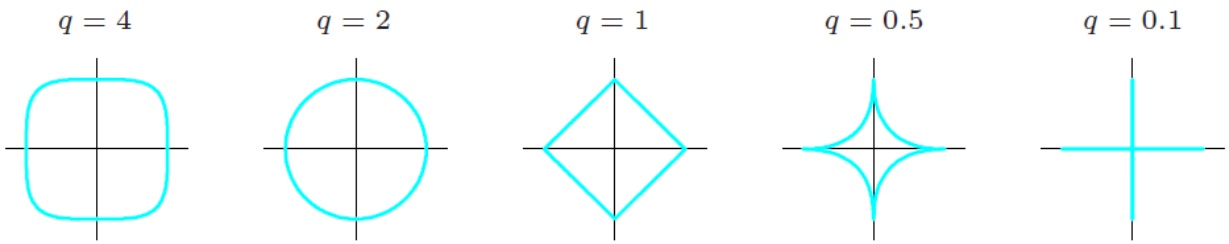


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

Source: Hastie, Trevor, et al. *The elements of statistical learning: data mining, inference, and prediction.* New York: springer, 2009.

The case $q = 1$ (lasso) is the smallest q such that the constraint region is convex; non-convex constraint regions make the optimization problem more difficult.

Elastic-net Penalty Regression

Zou and Hastie (2005) introduced the **elastic-net** penalty:

$$\lambda \sum_{j=1}^{p} (\alpha \beta_j^2 + (1-\alpha)|\beta_j|)$$

The elastic-net selects variables like the lasso, and shrinks together the coefficients of correlated predictors like ridge. It also has considerable computational advantages over the L_q penalties.

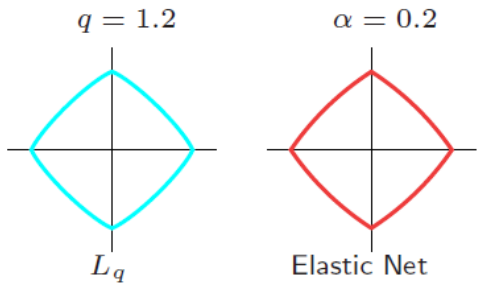


FIGURE 3.13. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for* $q = 1.2$ *(left plot), and the elastic-net penalty* $\sum_j (\alpha\beta_j^2 + (1-\alpha)|\beta_j|)$ *for* $\alpha = 0.2$ *(right plot). Although visually very similar, the elastic-net has sharp (non-differentiable) corners, while the* $q = 1.2$ *penalty does not.*

Source: Hastie, Trevor, et al. *The elements of statistical learning: data mining, inference, and prediction.* New York: springer, 2009.

Least Angle Regression

Least angle regression (LAR) is a relative newcomer (Efron et al. 2004). It uses a similar strategy as the forward stepwise regression, but LAR only enter "as much" of a predictor as it deserves.

At the first step, LAR identifies the variable most correlated with the response. Rather than fit this variable completely, LAR move the coefficient of this variable continuously toward its least-squares value.

As soon as anohter variable "catch up" in terms of correlation with the residual, the second variable join the set, and their coefficient are moved together.

This process is continued untile all the variables are in the model.

Least Angle Regression Algorithm

Algorithm 3.2 *Least Angle Regression.*

1. Standardize the predictors to have mean zero and unit norm. Start with the residual $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}$, $\beta_1, \beta_2, \ldots, \beta_p = 0$.
2. Find the predictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$.
3. Move β_j from 0 towards its least-squares coefficient $\langle \mathbf{x}_j, \mathbf{r} \rangle$, until some other competitor $\mathbf{x}_k$ has as much correlation with the current residual as does $\mathbf{x}_j$.
4. Move β_j and β_k in the direction defined by their joint least squares coefficient of the current residual on $(\mathbf{x}_j, \mathbf{x}_k)$, until some other competitor $\mathbf{x}_l$ has as much correlation with the current residual.
5. Continue in this way until all p predictors have been entered. After $\min(N-1, p)$ steps, we arrive at the full least-squares solution.

Source: Hastie, Trevor, et al. *The elements of statistical learning: data mining, inference, and prediction.* New York: springer, 2009.

A Comparison of the Selection and Shrinkage Methods

Subset Selection Methods:

Pros:

- **Interpretability:** Subset selection provides a clear interpretation of the selected variables. And the results model is often simpler and more comprehensible.
- **Improved Prediction Accuracy:** In some cases, subset selection can lead to improved predictive performance by excluding irrelevant or redundant features.

Cons:

- **Computational Complexity:** Subset selection involves evaluating multiple combinations of variables, making it computationally expensive.
- **Overfitting Risk:** If the sample size is small relative to the number of predictors, it may lead to overfitting, where the model fits the noise in the data rather than the underlying patterns.

A Comparison of the Selection and Shrinkage Methods

Shrinkagge Methods:
Pros:

- **Regularization:** Shrinkage methods introduce regularization penalties that help prevent overfitting by penalizing the magnitude of coefficients.
- **Computational Efficiency:** Shrinkage methods are often computationally more efficient than subset selection, especially for high-dimensional data.

Cons:

- **Bias:** Shrinkage methods introduce bias in parameter estimates, which can lead to a potential loss of information in certain cases.
- **Difficulty in Interpretation:** The resulting models from shrinkage methods might be less interpretable than those from subset selection.