Subset Selection & Shrinkage Methods

By:
Alvaro Flores
Ketan B
Anto Sibi Rayan A

IÉSEG SCHOOL OF MANAGEMENT

Why do we need these methods?

- To improve the standard linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

- Typically fitted with Plain Least Squares Method

Plain Least Squares Method

- Chooses β0 and β1 to minimize the RSS: the residual sum of squares

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

- Residuals: the difference between residual the observed response value and the response value that is predicted

$$e_i = y_i - \hat{y}_i$$

Why do we need these methods?

The methods may provide:

n: observations
p: predictors (variables)

- Better prediction accuracy

n>>p	n>=p	p>n
least squares estimates with Low variance	**least squares estimates with High variance -> Overfitting-> Poor prediction**	**least squares estimates with Infinite variance -> Not possible to be used**

By shrinking or constraining the estimated coefficients through alternative methods, variance can be substantially reduced at a negligible increase to bias.

- Better Model Interpretability

Methods automatically perform feature selection → Excluding irrelevant variables → Model that is easier to interpret

Subset Selection (6.1)

Methods to identify subset of predictors (p) that we believe to be related to the response

6.1.1 Best Subset Selection

- Fitting a separate least squares regression model for each possible combination of the p predictors

$$Balance = \beta_0 + \beta_1(Income)$$

$$Balance = \beta_0 + \beta_1(Limit)$$

$$Balance = \beta_0 + \beta_1(Rating)$$

M1

$$Balance = \beta_0 + \beta_1(Income) + \beta_2(Limit)$$

$$Balance = \beta_0 + \beta_1(Income) + \beta_2(Rating)$$

$$Balance = \beta_0 + \beta_1(Income) + \beta_2(Cards)$$

M2

3,4,5 . . . predictors

- Select the best model of M1, M2, M3, etc with smallest RSS
- Select a single best model using a criteria such as : Cross validated prediction error, Cp, AIC, BIC, or adjusted R2

Disadvantages: Computational limitations to process huge quantity of models p=10 1000 models p=20 > 1M models!!!

Subset Selection (6,1)

6.1.2 Stepwise Selection
Methods that explore a far more restricted set of models.

a) Forward Stepwise Selection
Starts model with no predictors. Predictors are added to the model one at a time, based on the predictor that gives the greatest additional improvement to the fit of the model.

$$Balance = \beta_0 + \beta_1(Cards)$$ M1

$$Balance = \beta_0 + \beta_1(Cards) + \beta_2(Income)$$ M2

$$Balance = \beta_0 + \beta_1(Cards) + \beta_2(Income) + \beta_3(Student)$$ M3... 4,5 etc predictors

Select the best model of M1, M2, M3, etc with smallest RSS

Select a single best model using a criteria such as : Cp, (AIC), BIC, or adjusted R2

Disadvantage: method does not guarantee to find the best possible model, since the final model is dependent on the first predictor that is added to the null model.

Subset Selection (6,1)

b) Backward Stepwise Selection
Begins with a model that contains all of the predictors. Predictors are removed from the model one at a time, based on the predictor that is least useful towards the fit of the model.

$$Balance = \beta_0 + \beta_1(Cards) + \beta_2(Income) + \beta_3(Student)$$

M3... 4,5 etc predictors

$$Balance = \beta_0 + \beta_1(Cards) + \beta_2(Income)$$

M2

$$Balance = \beta_0 + \beta_1(Cards)$$

M1

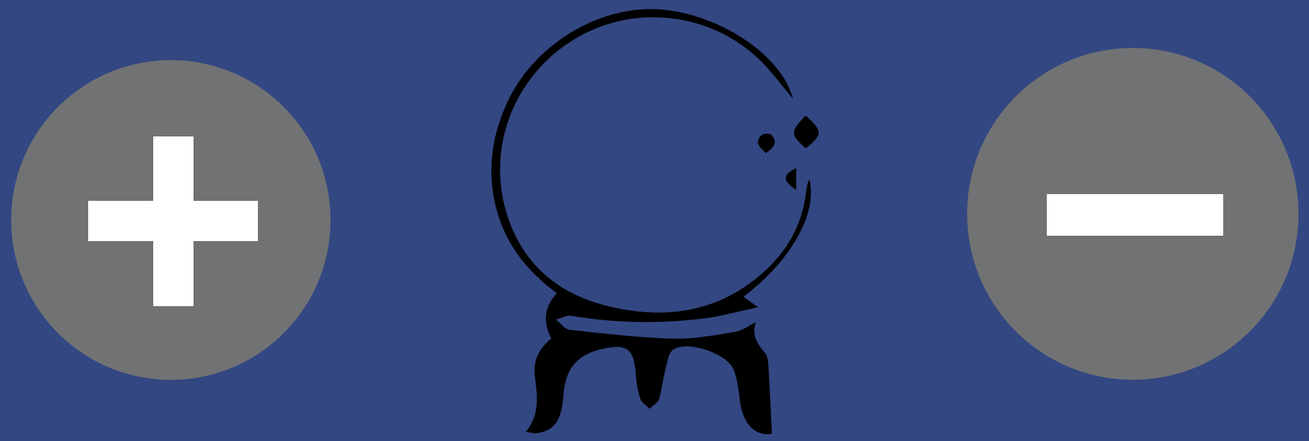
Select the best model of M1, M2, M3, etc with smallest RSS
Select a single best model using a criteria such as: Cp (AIC), BIC, or adjusted R2

Disadvantage: method does not guarantee to find the best possible model

Subset Selection (6,1)

c) Hybrid Stepwise Selection

Variables are added to the model sequentially, in analogy to forward selection. However, after adding each new variable, the method may also remove any variables that no longer provide an improvement in the model fit.



Subset Selection (6.1)

6.1.3 Choosing the optimal model

To select the best model with respect to the test error, this needs to be estimated through one of two methods:

a. making an adjustment to the training error to account for the bias due to overfitting.
b. using either a validation set approach or a cross-validation approach.

a. **Indirect Estimation of Test Error**

- Cp: The model with the lowest Cp is chosen as the best model.
- AIC: The model with the lowest AIC is chosen as the best model.
- BIC: The model with the lowest BIC is chosen as the best model.
- Adjusted-R2: The model with the highest adjusted-R2 is chosen as the best model.

Subset Selection (6.1)

b. Validation and Cross-Validation

Directly estimate the test error using these methods and then select the model with the smallest test error

Now it is computationally possible with modern computers

Shrinkage Methods

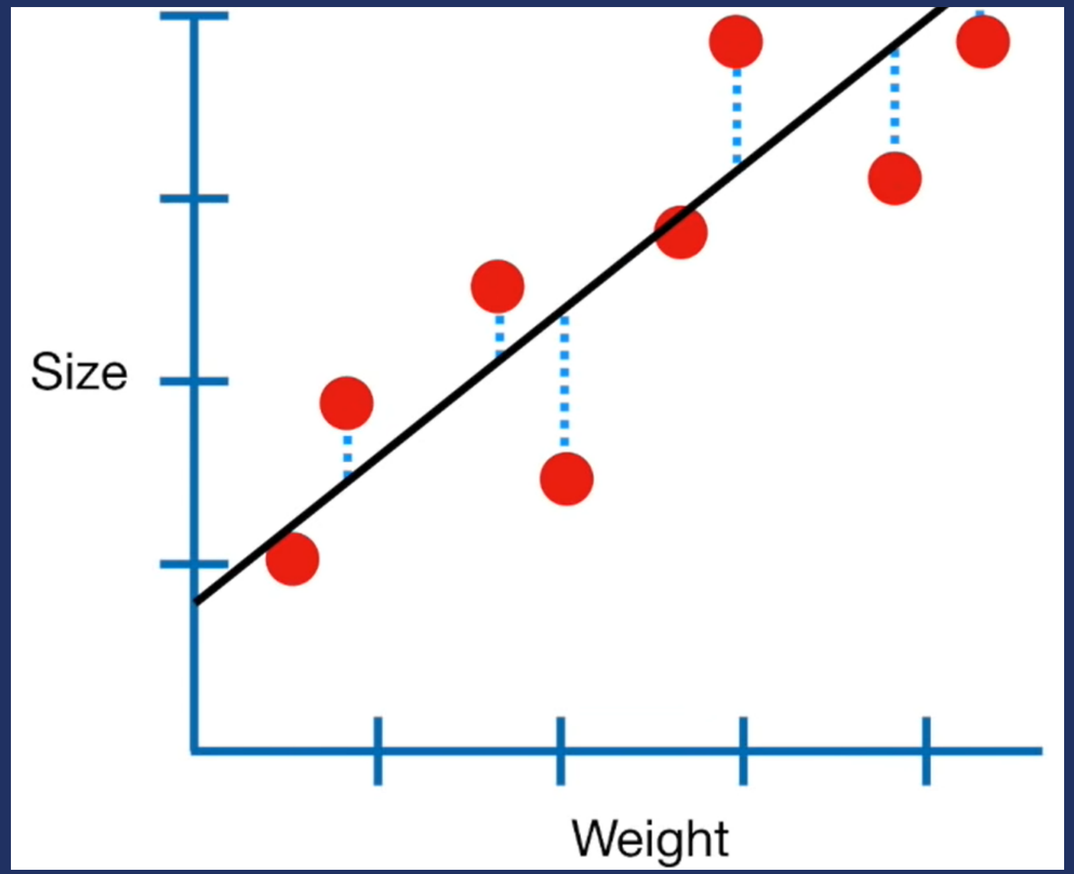
Ridge

reduce errors caused by overfitting on training data.

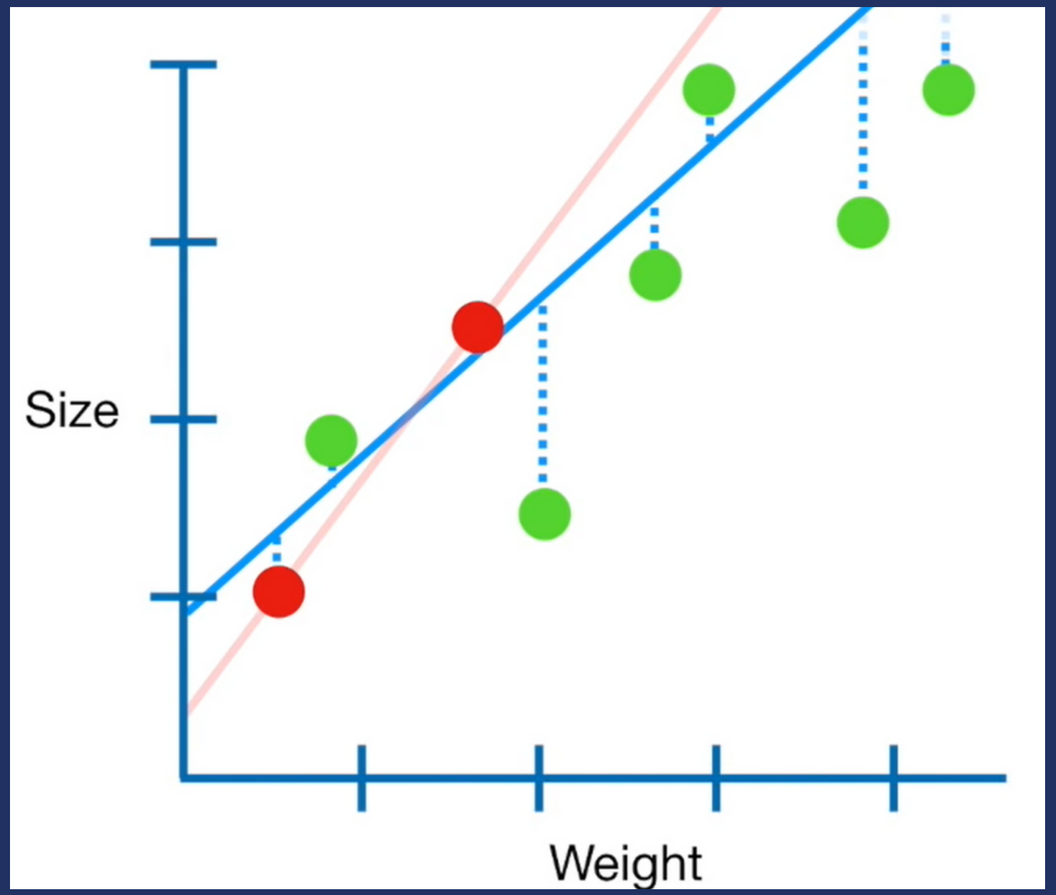
Lasso

simple, sparse models (i.e. models with fewer parameters).

Ridge Regression



Test Data



Test & Train Data

Variance would be high

For train data, sum of squared residuals would be 0

Ridge Regression

Created by Arthur Hoerl and Robert Kennard in 1970

Constrains the size of the regression coefficients

Ridge Regression is advantageous when,

- Number of variables to predict are high
- High multicollinearity between predictors
- Least squares estimates have high variance

Ridge Regression vs Least Squares

Least squares has overfit and high variance

Ridge Regression has a complexity penalty to shrink the coefficients

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2,$$

Slope squared adds a penalty to least squares method

Lambda determines how severe that penalty is

Ridge Regression vs Least Squares

Interpretation of Output

- A coefficient estimate for every predictor variable
- Coefficient estimate shrunk towards zero
- Degree of shrinkage depends on the tuning parameter lambda

Comparatively, Ridge Regression has

- Lower variance but slightly higher bias
- Yields lower test mean squared error

Ridge Regression

To standardize the predictors, we use the formula,

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}}$$

For Logistic Regression,

Optimizes the sum of likelihoods instead of squared residuals as Logistic Regression is solved using Maximum Likelihood

Ridge Regression

Advantages

Reduces variance of coefficient estimates

Can handle multicollinearity & overfitting - no of variables can be high

Disadvantage

Includes all predictors in the model - Challenging for model interpretation

Challenge

Selecting the optimum value for tuning parameter
[Using Ten fold cross validation method]

Lasso Régression

term coined by Robert Tibshirani.

Why

Automatically select most important features.

Helps

preventing overfitting

When

when only some predictors matter

How

minimizing a cost function

Lasso Régression

Ordinary least squares (OLS) linear regression by adding a penalty term to the cost function.

Coefficients change

least → **close to 0**

Previous formula

$$\text{Cost OLS} = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2$$

The addition

$+\lambda \sum_{j=1}^{p} |\beta_j|$, where λ is the regularization parameter.

$$\text{Cost OLS} = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2$$

sum of squared error (SSE)

$(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2$

L1 Regularization / sum of absolute values of coefficients

Lasso

Similar to Ridge Regression

$$\text{Lasso} = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Similarities	**Explanation**
use shrinkage	central point, like the mean.
regularization	adds a penalty
tuning parameter λ	λ increases, coefficients become smaller and closer to zero
when λ =0	least squares fit

In R, you can use the **glmnet** package for Lasso & Ridge Regression.

(λ = Lambda): is represented in $|\beta_j|$ for Lasso

Lasso

Differences to Ridge Regression

$$\text{Lasso} = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Differences	**Explanation**
Lasso = L1 regularization	penalty equal to the absolute value of the magnitude of coefficients,
sparse models	automatic feature selection
Ridge only shrinks	Lasso when not all features are equally important.
Feature selection	Ridge retains all features in the model,

Summarizing Results

Subset Selection =	interpretability and a clear subset of predictors are critical
Ridge Regression =	multicollinearity and you want to retain all predictors prediction > variable selection.
Lasso Regressio =	"sparse model" = where only a small number of features (predictors or variables)

Cross-validation

Thank You!