Regularization Methods

Sean Hellingman ©

Regression for Applied Data Science (ADSC2020)

shellingman@tru.ca

Winter 2025

THOMPSON RIVERS UNIVERSITY

Topics

Introduction

- **Regularization methods** are used to prevent overfitting in a model.
- May be used for *feature selection*.
- By constraining or *shrinking* the estimated coefficients, we can often substantially reduce the variance at the cost of a negligible increase in bias.

Shrinkage Methods I

- Shrinkage methods differ from other model selection techniques we have covered so far.
- All potential explanatory variables are included in the model.
- Instead of removing and adding variables, the model is estimated using a method that *constrains* or *regularizes* the coefficient estimates.

Shrinkage Methods II

- **Shrinkage methods** involve the following steps:
  - Fit a regression model with all explanatory variables.
  - The estimated coefficients are *shrunken* towards zero relative to their least squares estimates.
  - This (*regularization*) approach can significantly reduce variance.

- Depending on the approach, some coefficients may be estimated to be zero.
  - Therefore, shrinkage may be used for variable selection.

- Two best-known shrinkage methods:
  1. Ridge Regression
  2. LASSO

Review: Squared Residuals

- Because some of the residuals are positive and others are negative we square them (mathematical simplicity).

- We want to minimize the sum of the squared **residuals** (observed errors):

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2. \tag{1}$$

- The best-fitting line finds the intercept and slope that minimizes this sum (*Ordinary Least Squares (OLS) Regression*).

- **When $n > k$ (parameters) guaranteed a unique solution.**

OLS Estimation

- To estimate the OLS coefficients, we minimize the following quantity:

$$\text{RSS} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2. \tag{2}$$

- The p is the number of parameters (excluding intercept).

Ridge Regression

- Ridge regression is very similar, except a slightly different quantity is minimized.
- A *penalization* term for the coefficient size is included in the estimation process.
- The penalization term shrinks the coefficient estimates towards zero.

Ridge Regression Estimation

- Ridge regression coefficient estimates b_λ^R are the values that minimize:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \text{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2. \tag{3}$$

  - $\lambda \geq 0$ is a *tuning parameter* (determined separately)
  - $\lambda \sum_{j=1}^{p} \beta_j^2$ is the *shrinkage penalty*

- When the coefficient estimates are close to zero, the penalization term is small (shrinking effect).

Some Properties of Ridge Regression

- Ridge regression shrinks the coefficient estimates towards zero but never to zero.
- May be used to perform variable selection.
- Choice of λ is important and is often done through cross-validation.
- As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias.

LASSO

- Least Absolute Shrinkage and Selection Operator (LASSO)
- Regularization method for model selection
- The LASSO solution can yield a reduction in variance at the expense of a small increase in bias

Formulation

- The LASSO coefficients, b_λ^L minimize the quantity

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \text{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|. \quad (4)$$

  - $\lambda \geq 0$ is a *tuning parameter* (determined separately)
  - $\lambda\sum_{j=1}^{p}|\beta_j|$ is the *shrinkage penalty*

- When the coefficient estimates are close to zero, the penalization term is small (shrinking effect).

Some Properties

- LASSO shrinks the coefficient estimates towards zero.
- **With a sufficiently large λ some of the coefficient estimates shrink to be exactly zero.**
- LASSO performs variable selection.
- Choice of λ is important and is often done through cross-validation

Alternative Formulations

- Ridge regression:

$$\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\} \text{ subject to } \sum_{j=1}^{p}\beta_j^2 \le s. \tag{5}$$

- LASSO:

$$\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\} \text{ subject to } \sum_{j=1}^{p}|\beta_j| \le s. \tag{6}$$

Variable Selection Property

LASSO: Ridge:

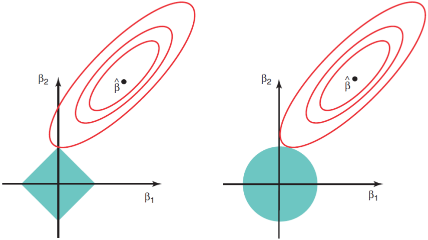

Figure: Source (1)

- Two parameters ($p = 2$)
- $\hat{\beta}$: OLS solution
- Blue rectangle: $|\beta_1| + |\beta_2| \leq s$
- Blue circle: $\beta_1^2 + \beta_2^2 \leq s$
- Red ellipses: regions of constant RSS

Comments on Shrinkage

- When $\lambda = 0$, OLS estimates.
- Reduction in variance at the expense of a small increase in bias.
- Can be a useful tool for model selection.
- Models fit using *penalized maximum likelihood.*

Scaling

- Because the penalization is directly related to the size of b, the explanatory variables should be in the same scale.

- One method for scaling is the **Min-Max scaling** method.

$$\frac{x_i - min(x)}{max(x) - min(x)}$$

- All observations are from in the range [0,1].

Example 1 Preliminaries

- Suppose that we are interested in estimating a linear regression model on the response variable `hp` (horse power) from the *mtcars* dataset.
- We are worried about overfitting and would like to use regularization methods to help with the model estimation process.
- Want to model horsepower (`hp`) dependent on Miles/gallon (`mpg`), weight (`wt`), rear axle ratio (`drat`), and 1/4 mile time (`qsec`).

Example 1

- Import the *mtcars* dataset into R.
- Examine the variables under consideration. What do you notice about their scales?
- Use the provided code to perform Min-Max scaling on the variables of interest.
    - How did this change things?

Ridge Regression in R

- To estimate a model using ridge regression in R:
  - `library(glmnet)`
  - `Ridge.Model <- glmnet(x, y, alpha=0, lambda = `λ`)`
  - `coef(Ridge.Model)`
- The argument `alpha=0` is the ridge penalty.
- λ is the tuning parameter.

LASSO in R

- To estimate a model using ridge regression in R:
  - `library(glmnet)`
  - `Ridge.Model <- glmnet(x, y, alpha=1, lambda =` λ`)`
  - `coef(Ridge.Model)`
- The argument `alpha=1` is the LASSO penalty.
- λ is the tuning parameter.

Example 2

- Estimate three models each using a ridge and a LASSO penalization term with the following λ values:
  1. $\lambda = 0$ (OLS Estimate)
  2. $\lambda = 0.001$
  3. $\lambda = 5$
- *Five total models.*
- What do you notice about the estimated coefficients?

Tuning Parameter Selection

- **Find optimal lambda value that minimizes test mean squared error (MSE).**

- Perform 10-fold cross-validation to find optimal lambda value.

- Functionality in the *glmnet* R package:
    - `cv1 <- cv.glmnet(x, y, nfolds = 10, alpha = )`

    - `best_lambda <- cv1$lambda.min`
    - `best_lambda`

Example 3

- Use cross-validation to determine the best tuning parameter for your ridge regression and LASSO models.

Process Visualization

- You can also visualize the cross-validation process in R:
  - `plot(cv1)`
- Can also visualize the *shrinkage* process of the coefficients with increasing lambda:
  - `fit <- glmnet(x, y, alpha = )`
  - `plot(fit)`
- `alpha=0`: Ridge regression penalty.
- `alpha=1`: LASSO penalty.

Example 4

- Visualize the cross-validation process used to determine the best lambda.
- Visualize the *shrinkage* of the coefficients in your models.
  - Do you notice any differences?
- Examine and comment on the coefficient estimates of your final models.

Conclusions

- Penalizes β values by *shrinking* them to (or close to) zero.
- Useful for variable selection and can be applied to GLMs.
- Choice of λ is important and is often done through cross-validation.
- **Be careful with categorical variables, you can include columns of dummy variables but the order does matter.**
- Related Topics:
  - Elastic net regularization
  - Methods for dimension reduction

Exercise 1

- Take some time to try out some of these regularization methods on your own data. They are especially useful if you have *wide* data.

References & Resources

1. James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

2. Fox, J. (2015). *Applied regression analysis and generalized linear models (Third Edition).* Sage Publications.

3. Additional Resources

- glmnet
- `Shrinkage Methods`
- `glmnet()`
- `cv.glmnet()`