Ridge Regression

Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity.

When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value.

Ridge regression is carried out on the linear regression model

$$Y = X\beta + \epsilon$$

where

Y is the $n \times 1$ vector of observations of the dependent variable

X is the $N \times K$ matrix of regressors

β is the $k \times 1$ vector of regression coefficients

ϵ is the $n \times 1$ vector of errors

Ridge Estimator

The objective function is given by

$$f(\beta) = (y - X\beta)'(y - X\beta) + \lambda\beta'\beta$$

We differentiate the function with respect to β and set the result equal to zero and have:

$$\frac{\partial f(\beta)}{\partial \beta} = -2X^T(y - X\beta) + 2\lambda\beta = 0$$

Solving for β

$$X^T X\beta + \lambda I\beta = X^T y$$

Then

$$\hat{\beta_{Ridge}} = (X^T X + \lambda I)X^T y$$

Where λ is a positive constant

Bias and Variance of Ridge Estimator

We derive the bias and the variance of the ridge estimator under the commonly made assumption that conditional on X, the errors have a zero mean and a constant variance σ^2 and are uncorrelated.

$$E[\epsilon|X] = 0$$

$$Var[\epsilon|X] = \sigma^2 I$$

where σ^2 is a positive constant and I is the $n \times n$ identity matrix.

Bias

The conditional expected value of the ridge estimator $\hat{\beta_\lambda}$ is

$$E[\hat{\beta_\lambda}|X] = (X^TX + \lambda I)^{-1}X^TX\beta$$

which is different from β unless the $\lambda = 0$

The bias of the estimator is

$$E[\hat{\beta_\lambda}|X] - \beta = \Big[(X^TX + \lambda I)^{-1} - (X^TX)^{-1}\Big]X^TX\beta$$

Proof

We can write the ridge estimator as

$$\begin{aligned}\hat{\beta_\lambda} &= (X^TX + \lambda I)^{-1}X^Ty \\ &= (X^TX + \lambda I)^{-1}X^T(X\beta) + \epsilon \\ &= (X^TX + \lambda I)^{-1}X^TX\beta + (X^TX + \lambda I)^{-1}X^T\epsilon\end{aligned} \tag{7}$$

Therefore

$$\begin{aligned}E[\hat{\beta_\lambda}] &= (X^TX + \lambda I)^{-1}X^TX\beta + (X^TX + \lambda I)^{-1}X^TE[\epsilon|X] \\ &= (X^TX + \lambda I)^{-1}X^TX\beta + (X^TX + \lambda I)^{-1}X^T \times 0 \\ &= (X^TX + \lambda I)^{-1}X^TX\beta\end{aligned} \tag{8}$$

The ridge estimator is unbiased if and only if

$$(X^TX + \lambda I)^{-1}X^TX = I$$

Variance

The covariance of the ridge estimator is given by:

$$Var[\hat{\beta_\lambda}|X] = \sigma^2(X^TX + \lambda I)^{-1}X^TX(X^TX + \lambda I)^{-1}$$

Proof

Remember that the OLS estimator $\hat{\beta}$ has conditional variance

$$Var[\hat{\beta}] = \sigma^2(X^TX)^{-1}$$

We can write the ridge estimator as a function of the OLS estimator

$$\begin{aligned}\hat{\beta_\lambda} &= (X^TX + \lambda I)^{-1}X^Ty \\ &= (X^TX + \lambda I)^{-1}X^TX(X^TX)^{-1}X^Ty \\ &= (X^TX + \lambda I)^{-1}X^TX\hat{\beta}\end{aligned} \tag{9}$$

Therefore:

$$\begin{aligned} Var[\hat{\beta_\lambda}] &= (X^TX+\lambda I)^{-1}X^TXVar[\hat{\beta_\lambda}][(X^TX+\lambda I)^{-1}X^TX]^T \\ &= (X^TX+\lambda I)^{-1}X^TXVar[\hat{\beta_\lambda}X^TX(X^TX+\lambda I)^{-1}] \\ &= (X^TX+\lambda I)^{-1}X^TX\sigma^2(X^TX)^{-1}X^TX(X^TX+\lambda I)^{-1} \\ &= \sigma^2(X^TX+\lambda I)^{-1}X^TX(X^TX+\lambda I)^{-1} \end{aligned} \tag{10}$$

How to choose λ

The most common way to find the best λ is by using leave-one-out cross-validation.

The steps are as follows:

- We choose a grid of p possible values of $\lambda_1, \lambda_2, ..., \lambda_p$ for the penalty parameter
- for $i = 1, ..., N$ we exclude the $i-th$ observation (y_i, x_i) from the sample and use the remaining $n-1$ observations to compute p ridge estimates of β denoted by $\hat{\beta_{\lambda p,i}}$ and compute p out-of-sample predictions of the excluded observation
- We compute the MSE of the predictions

$$MSE_\lambda = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_{\lambda p})^2$$

- We choose as the optimal penalty parameter λ the one that minimizes the MSE of the predictions

Example

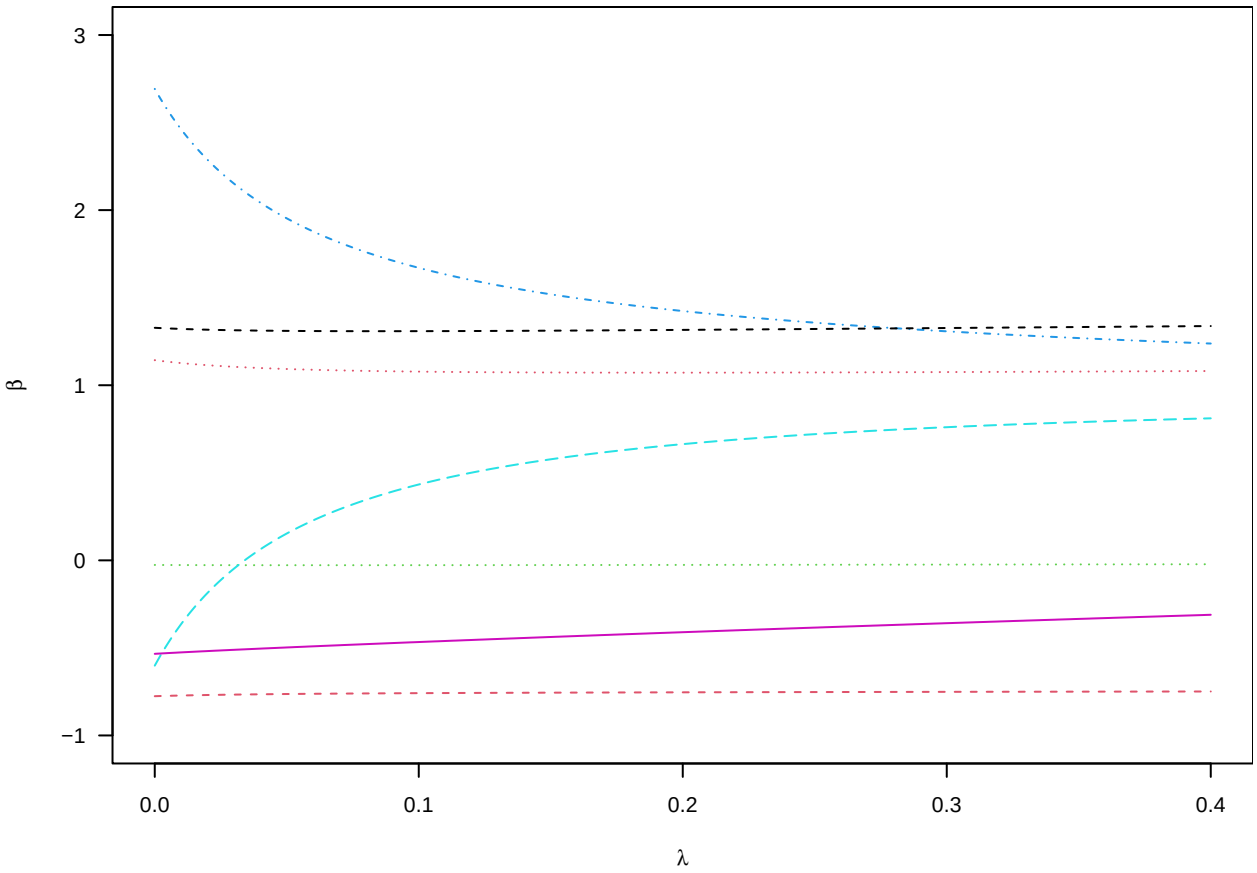
As the beginning of ridge regression, it is recommended to standardize the predictors. You can still carry out ridge regression without doing so, but standardization would improve the effect of ridge regression, as it makes the shrinking fair to each coefficients. Luckily, the function that we are going to use here automatically standardizes the data, so we don't need to do the standardization by ourselves.

We use the `MASS` package in R

```
# loading the data

data <- read.csv("ridge.csv")

# package to use

library(MASS)

# model with a range of lambdas

fit = lm.ridge(hipcenter ~ ., data, lambda = seq(0, .4, 1e-3))
```

We can observe how the coefficients shrink as λ grows larger:

```
par(mar = c(4, 4, 0, 0), cex = 0.7, las = 1)
matplot(fit$lambda, coef(fit), type = "l", ylim = c(-1, 3),
        xlab = expression(lambda), ylab = expression(hat(beta)))
```



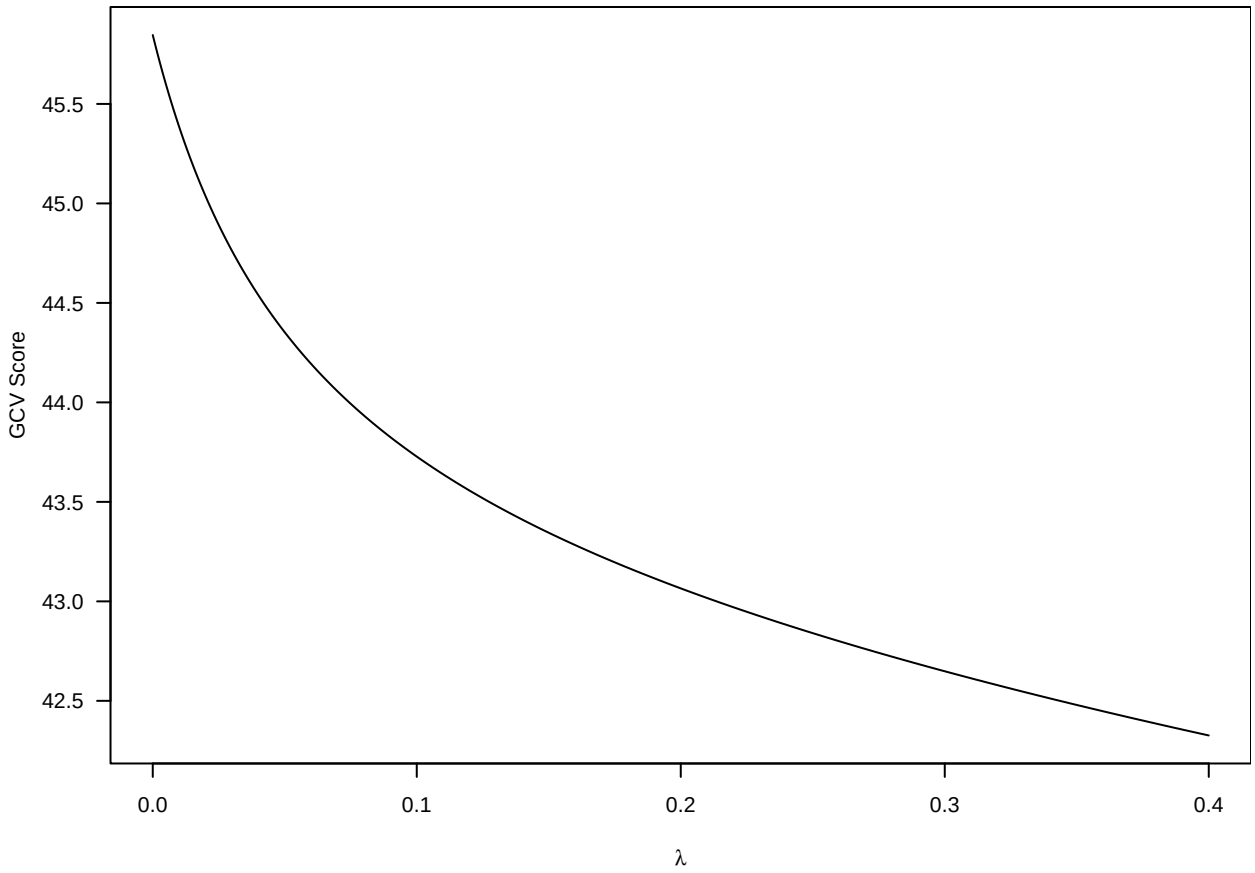
To select the optimal value of λ we use `select` function

```
select(fit)
```

```
## modified HKB estimator is 5.425415
## modified L-W estimator is 3.589434
## smallest value of GCV  at 0.4
```

So the optimal value of λ is at 0.4

```
par(mar = c(4, 4, 0, 0), cex = 0.7, las = 1)
plot(names(fit$GCV), fit$GCV, type = 'l',
     xlab = expression(lambda), ylab = "GCV Score")
```



Detecting Outliers in Regression Models

Outliers are observations that appear inconsistent with the rest of dataset.

A more precise definition, they are observations that are distinct from most of the data points in the sample.

There are many methods of detecting outliers in regression models. They include:

- Graphical methods
- Analytical methods

Graphical Methods

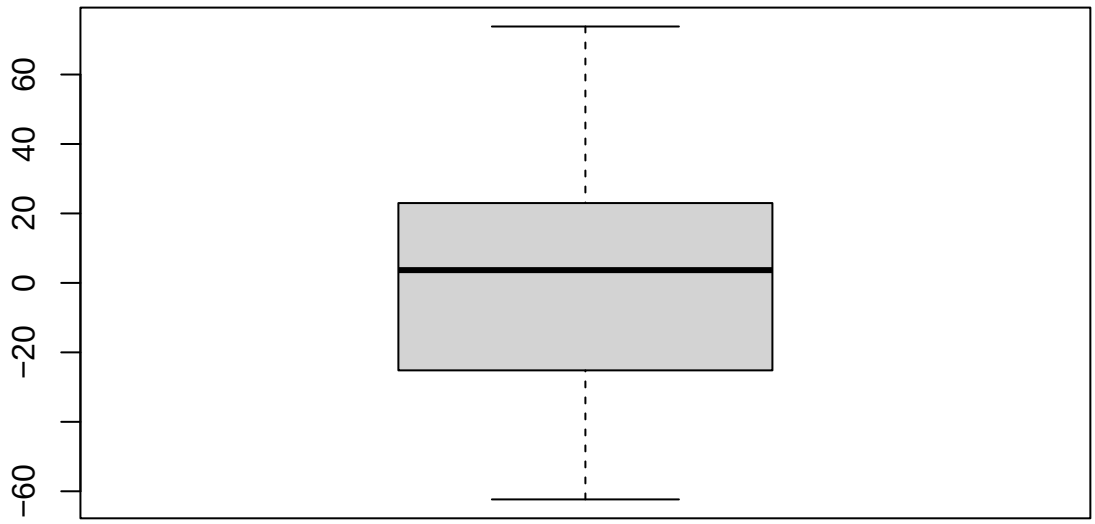
The graphical methods include scatter graph, boxplot, williams graph, Q-Q plot and graph of predicted residuals

Scatter and Box plot

Scatter plot is a line of best fit (alternatively called "trendline") drawn in order to study the relationship between the variables measured. For a set of data variables (dimensions) $X_1, X_2, ..., X_k$the scatter plot matrix shows all the pairwise scatter plots of the variables on the dependent variable.

A box plot is a method for graphically depicting groups of numerical data through their quartiles (i.e. Mean, Median Mode, quartiles). Box plots may also have lines extending vertically from the boxes (whiskers) indicating variability outside the upper and lower quartiles. It is also called box-and-whisker plot and box-and-whisker diagram. Outliers may be plotted as individual points and it can be used for outlier detection in regression model, where the primary aim here is not to fit a regression model but find out outliers using regression and to improve a regression model by removing the outliers.

```r
# loading the data

data <- read.csv("ridge.csv")

# model

fit = lm(hipcenter ~ ., data)

# extract the residuals

res <- fit$residuals

# boxplot of the residuals

boxplot(res)
```



Analytical Methods

The analytical methods include:

- predicted residuals
- Standardized residuals
- Jack-nife residuals
- Cook's distance
- Atkinsons measure

Studentized and Standardized Residuals

The Standardized residuals are given by:

$$\epsilon_{S.i} = \frac{\hat{\epsilon_i}}{\hat{\sigma}\sqrt{1-h_i}} = \frac{\hat{\epsilon_i}}{\sqrt{\hat{\sigma^2}(1-h_i)}}$$

Studentized residuals with large absolute values are considered large. If the regression model is appropriate, with no outlying observations, each Studentized residual follows a t distribution with n-k-1 degrees of freedom.

If the Studentized residual is divided by the estimates of its standard error so that the outcome is a residual with zero mean and standard deviation one, it becomes standardized residual denoted by

$$\epsilon_{ST.i} = \frac{\hat{\epsilon_i}}{sd(\sigma)}$$

The standardized residuals, $d_i > 3$ potentially indicate outlier