Ridge Regression

Jeffrey Woo

MSDS, University of Virginia

Least Squares in Linear Regression

The least squares criterion in linear regression results in estimators for the regression coefficient that have **minimum variance among all linear unbiased estimators** (Gauss Markov Theorem).

Variance in Least Sqaures

In some circumstances, least squares estimators have high variance:

- multicollinearity is present (at least one predictor is a linear combination of other predictors), or
- when we have many predictors and a smaller number of observations ($\hat{\sigma}^2 = \frac{SSres}{n-p}$).

Bias-Variance Tradeoff

There may exist estimators (other than least squares) that are biased but may have a substantially smaller variance than least squares, and hence smaller mean-squared errors (MSE). The MSE of the estimated coefficients can be decomposed into the squared bias plus the variance of the estimated coefficients:

$$MSE(\hat{\beta}) = bias(\hat{\beta})^2 + var(\hat{\beta}) \tag{1}$$

Shrinkage Methods

Shrinkage methods regularize or shrink the estimated coefficients $\hat{\beta}$ toward 0. Shrinking coefficients toward 0 introduces bias but may reduce the variance enough to **reduce the test MSE**. We will learn about two shrinkage methods:

1. ridge regression and
2. lasso regression.

Least Squares

In least squares, we seek to find the the estimates of $\beta_0, \cdots, \beta_k$ that minimize the residual sum of squares:

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

Ridge regression is similar except that the coefficients are estimated by minimizing a slightly different quantity than (2).

Ridge Regression

The ridge regression coefficient estimates, denoted by $\hat{\beta}^R_\lambda$, are found by minimizing

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{k}\beta_j^2 = SSres + \lambda\sum_{j=1}^{k}\beta_j^2, \quad (3)$$

where λ is a tuning parameter to be determined separately.

Ridge Regression

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

- Ridge regression seeks coefficients that fit the (training) data well by making the first term, SSres, small.
- The second term, $\lambda\sum_{j=1}^{k}\beta_j^2$, called the **shrinkage penalty**, is small when the estimated coefficients are close to 0.
- The tuning parameter λ controls the relative impact of the two terms. As $\lambda \to \infty$, the impact of the shrinkage penalty grows.

Question: What happens when $\lambda = 0$?

Ridge Regression

- Unlike least squares which generates only one set of estimated coefficients, ridge regression will generate a different set of estimated coefficients, $\hat{\beta}^R_\lambda$, for each value of λ.
- Selecting a good λ is important, and is typically done by cross-validation.

Shrinkage Penalty

- The β_j^2 penalty in (3) is called an ℓ_2 penalty. The ℓ_2 norm of a vector β is given by $\|\beta\|_2 = \sum \beta_j^2$.
- Notice in (3), the shrinkage penalty is applied to the coefficients $\beta_1, \cdots, \beta_p$ but not the intercept. We want to shrink the estimated association of each variable with the response.
- The intercept is just the mean of the response when all the predictors are 0.
- As λ increases, **variance decreases but bias increases**.

Standardizing the Predictors

- The shrinkage penalty $\lambda \sum_{j=1}^{k} \beta_j^2$ depends on both the tuning parameter as well as the coefficients.
- This means the predictors should be on a similar scale, so that each predictor has **similar impact** on the shrinkage penalty.
- The predictors should be standardized.
- The `glmnet()` function that we will use for shrinkage methods performs the standardization by default.

Advantages of Ridge Regression Over Least Squares

As mentioned earlier, the variance of a least squares regression model can be high due the following reasons:

- multicollinearity is present, or
- the number of predictors is almost equal to the number of observations.

A consequence of a model with high variance is that a small change in the training data can result in a **large change** in the coefficient estimates and predicted response. Ridge regression is used to reduce the variance of a model by introducing some bias.

Computational Advantages of Ridge Regression Over Least Squares

- If $p > n$, the least squares estimators do not have a **unique solution**, whereas ridge regression can still be used.
- Fitting procedure in ridge regression is efficient: the computations required to solve (3) simultaneously for all values of λ are almost identical to those for fitting a least squares model.

1. Ridge Regression
2. Worked Example

Worked Example: Gas Mileage

We will use the mtcars data set that is available in R. The data set contains information about fuel consumption and 10 aspects of automobile design and performance for 32 classic vehicles.

Worked Example: Gas Mileage

```
> head(mtcars, 8)
                   mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
```

Worked Example: Gas Mileage

```
mpg:  Miles/(US) gallon
cyl:  Number of cylinders
disp: Displacement (cu.in.)
hp:   Gross horsepower
drat: Rear axle ratio
wt:   Weight (1000 lbs)
qsec: 1/4 mile time
vs:   Engine (0 = V-shaped, 1 = straight)
am:   Transmission (0 = automatic, 1 = manual)
gear: Number of forward gears
carb: Number of carburetors
```

Worked Example: Gas Mileage

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.30337   18.71788   0.657   0.5181
xcyl        -0.11144    1.04502  -0.107   0.9161
xdisp        0.01334    0.01786   0.747   0.4635
xhp         -0.02148    0.02177  -0.987   0.3350
xdrat        0.78711    1.63537   0.481   0.6353
xwt         -3.71530    1.89441  -1.961   0.0633 .
xqsec        0.82104    0.73084   1.123   0.2739
xvs          0.31776    2.10451   0.151   0.8814
xam          2.52023    2.05665   1.225   0.2340
xgear        0.65541    1.49326   0.439   0.6652
xcarb       -0.19942    0.82875  -0.241   0.8122

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared:  0.869,     Adjusted R-squared:  0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

Worked Example: Gas Mileage

Consider using ridge regression to reduce variance of the model. Randomly split the data set into 2 equal parts, a training and a test set.

Worked Example: Gas Mileage

Check the test MSE for various values of $\lambda = 0, 4, 10^{10}$.

```
ridge.pred.0<-predict(ridge.mod,s=0,newx=x.test)
mean((ridge.pred.0-y.test)^2)
[1] 12.66556

ridge.pred.4<-predict(ridge.mod,s=4,newx=x.test)
mean((ridge.pred.4-y.test)^2)
[1] 7.723544

ridge.pred.l<-predict(ridge.mod,s=1e10,newx=x.test)
mean((ridge.pred.l-y.test)^2)
[1] 40.4016
```

Worked Example: Gas Mileage

Use cross-validation to find the λ that is optimal. By default, the function `cv.glmnet` uses 10-fold cross validation.

```
set.seed(12)
cv.out<-cv.glmnet(x.train,y.train,alpha=0)
plot(cv.out)
bestlam<-cv.out$lambda.min
bestlam
[1] 2.643695
ridge.pred<-predict(ridge.mod,s=bestlam,newx=x.test)
mean((ridge.pred-y.test)^2)
[1] 7.398034
```

Worked Example: Gas Mileage

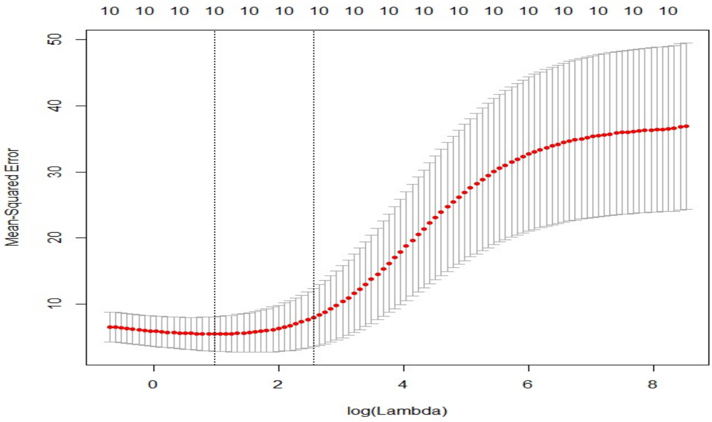

Figure: MSE with ridge regression against various values of tuning parameter (in logarithm).

Worked Example: Gas Mileage

Fit ridge regression using all observations using optimal λ, and compare coefficients with least squares.

```
cbind(coefficients(out.ridge), coefficients(out.ols))
(Intercept) 21.164674099 12.30334580
cyl         -0.371614420 -0.11143684
disp        -0.005238229  0.01333516
hp          -0.011645632 -0.02148211
drat         1.052609540  0.78711436
wt          -1.244587522 -3.71529660
qsec         0.162770783  0.82103979
vs           0.763638514  0.31776368
am           1.635361847  2.52022649
gear         0.545524257  0.65541591
carb        -0.552818552 -0.19942219
sqrt(sum(coefficients(out.ridge)[-1]^2))
[1] 2.585053
sqrt(sum(coefficients(out.ols)[-1]^2))
[1] 4.693825
```

Worked Example: Gas Mileage

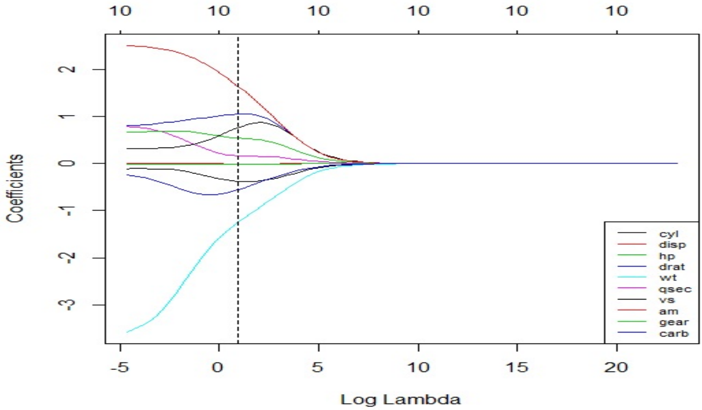

Figure: Ridge coefficients against various values of tuning parameter (in logarithm).