STAT 8020 R Lab 12: Adavnced Topics II

Whitney

September 28, 2020

Contents

Regression Tree

Major League Baseball Hitters Data from the 1986–1987 season

```
library(rpart)
library(rpart.plot)
library(ISLR)
Hitters = na.omit(Hitters)
head(Hitters)
```

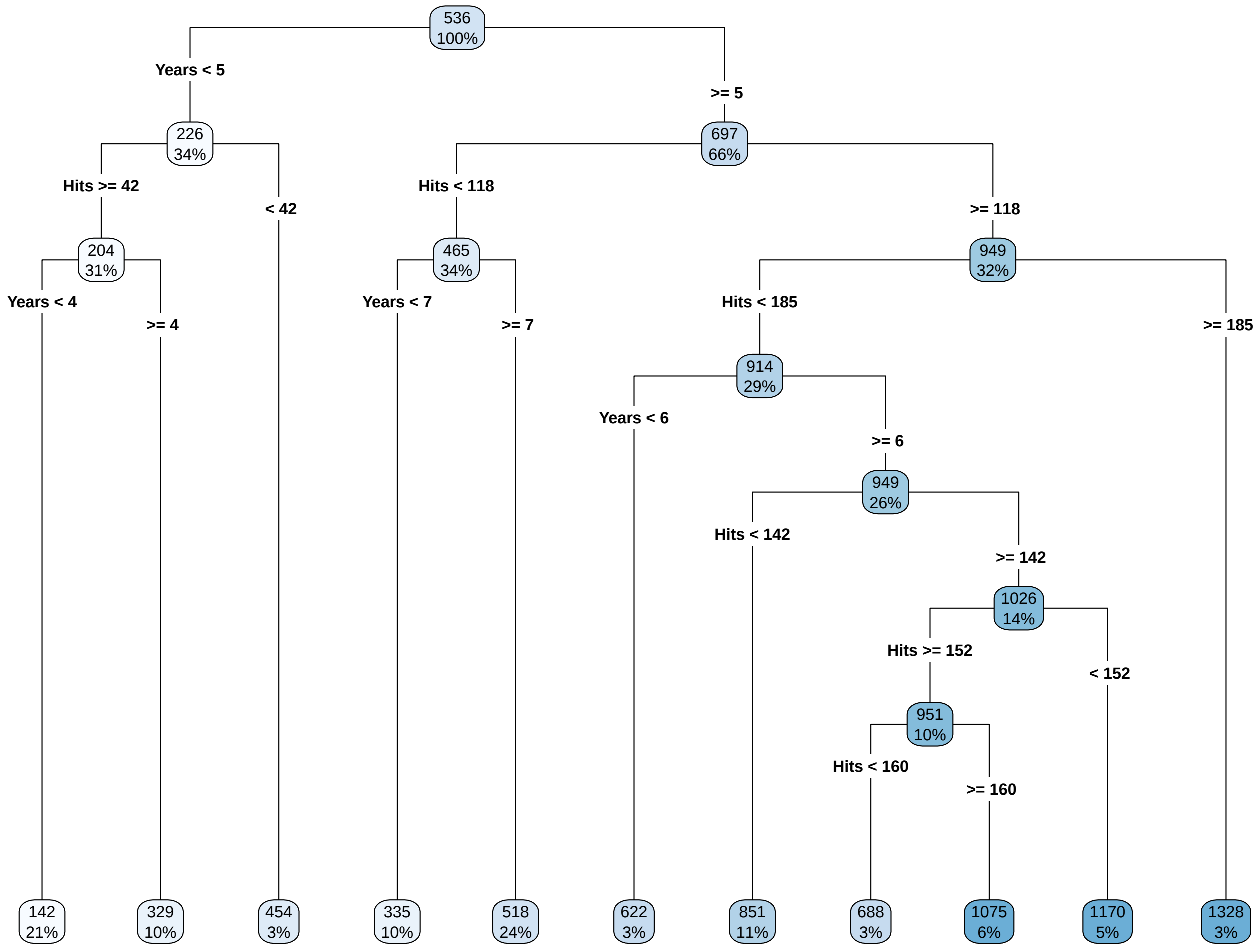
```
##                   AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun
## -Alan Ashby         315   81     7   24  38    39    14   3449   835     69
## -Alvin Davis        479  130    18   66  72    76     3   1624   457     63
## -Andre Dawson       496  141    20   65  78    37    11   5628  1575    225
## -Andres Galarraga   321   87    10   39  42    30     2    396   101     12
## -Alfredo Griffin    594  169     4   74  51    35    11   4408  1133     19
## -Al Newman          185   37     1   23   8    21     2    214    42      1
##                   CRuns CRBI CWalks League Division PutOuts Assists Errors
## -Alan Ashby         321  414    375      N        W     632      43     10
## -Alvin Davis        224  266    263      A        W     880      82     14
## -Andre Dawson       828  838    354      N        E     200      11      3
## -Andres Galarraga    48   46     33      N        E     805      40      4
## -Alfredo Griffin    501  336    194      A        W     282     421     25
## -Al Newman           30    9     24      N        E      76     127      7
##                   Salary NewLeague
## -Alan Ashby        475.0         N
## -Alvin Davis       480.0         A
## -Andre Dawson      500.0         N
## -Andres Galarraga   91.5         N
## -Alfredo Griffin   750.0         A
## -Al Newman          70.0         A
```

```
summary(Hitters)
```

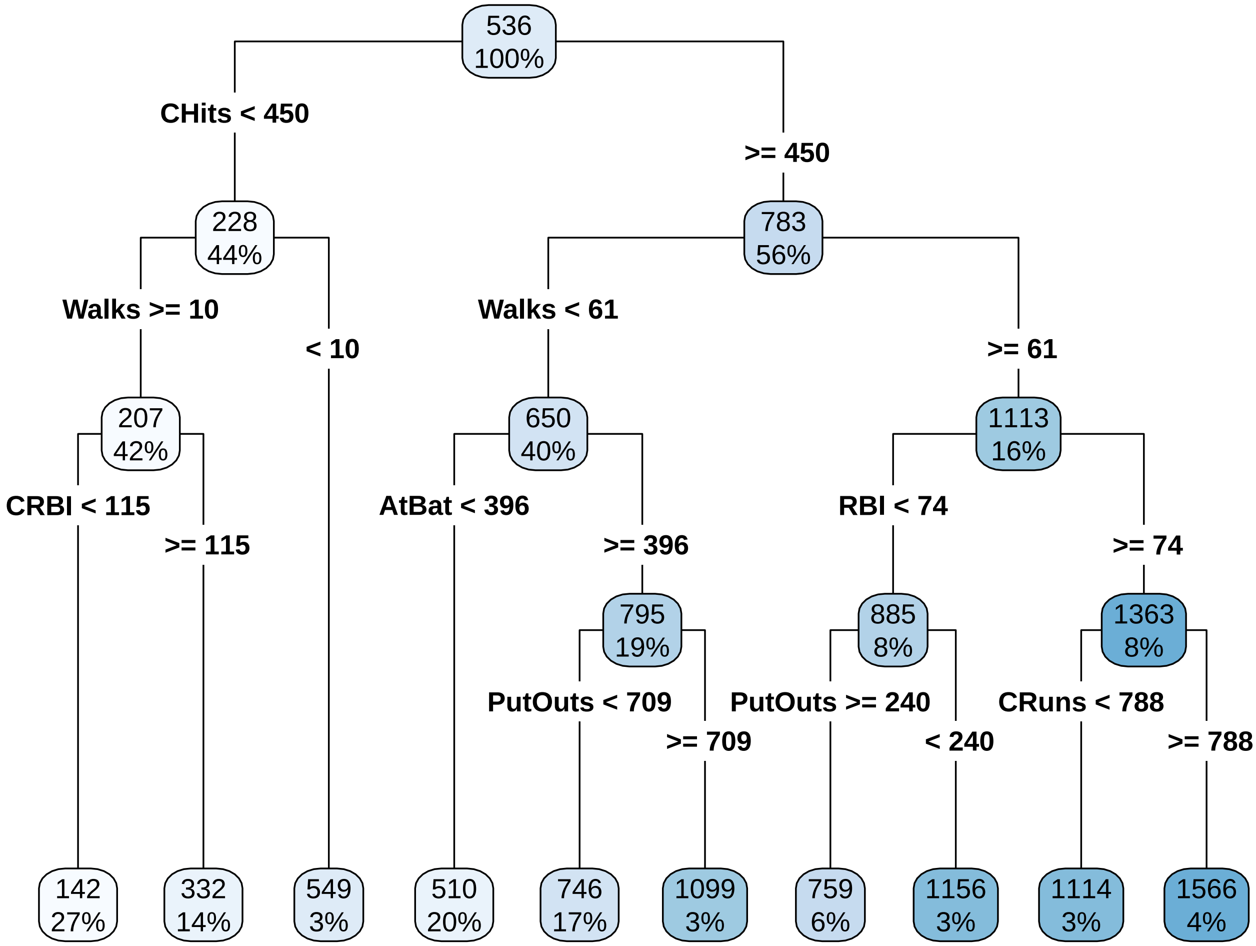
```
##      AtBat            Hits           HmRun            Runs
##  Min.   : 19.0   Min.   :  1.0   Min.   : 0.00   Min.   :  0.00
##  1st Qu.:282.5   1st Qu.: 71.5   1st Qu.: 5.00   1st Qu.: 33.50
##  Median :413.0   Median :103.0   Median : 9.00   Median : 52.00
```

```
##  Mean   :403.6   Mean   :107.8   Mean   :11.62   Mean   : 54.75  
##  3rd Qu.:526.0   3rd Qu.:141.5   3rd Qu.:18.00   3rd Qu.: 73.00  
##  Max.   :687.0   Max.   :238.0   Max.   :40.00   Max.   :130.00  
##       RBI             Walks            Years            CAtBat       
##  Min.   :  0.00   Min.   :  0.00   Min.   : 1.000   Min.   :   19.0  
##  1st Qu.: 30.00   1st Qu.: 23.00   1st Qu.: 4.000   1st Qu.:  842.5  
##  Median : 47.00   Median : 37.00   Median : 6.000   Median : 1931.0  
##  Mean   : 51.49   Mean   : 41.11   Mean   : 7.312   Mean   : 2657.5  
##  3rd Qu.: 71.00   3rd Qu.: 57.00   3rd Qu.:10.000   3rd Qu.: 3890.5  
##  Max.   :121.00   Max.   :105.00   Max.   :24.000   Max.   :14053.0  
##      CHits            CHmRun           CRuns             CRBI       
##  Min.   :   4.0   Min.   :  0.00   Min.   :   2.0   Min.   :   3.0  
##  1st Qu.: 212.0   1st Qu.: 15.00   1st Qu.: 105.5   1st Qu.:  95.0  
##  Median : 516.0   Median : 40.00   Median : 250.0   Median : 230.0  
##  Mean   : 722.2   Mean   : 69.24   Mean   : 361.2   Mean   : 330.4  
##  3rd Qu.:1054.0   3rd Qu.: 92.50   3rd Qu.: 497.5   3rd Qu.: 424.5  
##  Max.   :4256.0   Max.   :548.00   Max.   :2165.0   Max.   :1659.0  
##      CWalks        League  Division    PutOuts          Assists     
##  Min.   :   1.0   A:139   E:129    Min.   :   0.0   Min.   :  0.0  
##  1st Qu.:  71.0   N:124   W:134    1st Qu.: 113.5   1st Qu.:  8.0  
##  Median : 174.0                    Median : 224.0   Median : 45.0  
##  Mean   : 260.3                    Mean   : 290.7   Mean   :118.8  
##  3rd Qu.: 328.5                    3rd Qu.: 322.5   3rd Qu.:192.0  
##  Max.   :1566.0                    Max.   :1377.0   Max.   :492.0  
##      Errors           Salary       NewLeague
##  Min.   : 0.000   Min.   :  67.5   A:141    
##  1st Qu.: 3.000   1st Qu.: 190.0   N:122    
##  Median : 7.000   Median : 425.0            
##  Mean   : 8.593   Mean   : 535.9            
##  3rd Qu.:13.000   3rd Qu.: 750.0            
##  Max.   :32.000   Max.   :2460.0            
```

```r
#Tree 1
reg.tree <- rpart(Salary ~ Years + Hits, data = Hitters)
rpart.plot(reg.tree, type = 4)
```




```r
#Tree 2
reg.tree <- rpart(Salary ~ ., data = Hitters)
rpart.plot(reg.tree, type = 4)
```



Ridge Regression

The rest of this lab is largely based on the R lab: Ridge Regression and the Lasso of the book "Introduction to Statistical Learning with Applications in R" by *Gareth James*, *Daniela Witten*, *Trevor Hastie* and *Robert Tibshirani*. We will use the `glmnet` package to perform ridge regression and the lasso to predict `Salary` on the `Hitters` data.

Data Setup

```
library(glmnet)
```

```
## Loading required package: Matrix
## Loading required package: foreach
## Loaded glmnet 2.0-18
```

```
X <- model.matrix(Salary ~ ., data = Hitters)[, -1]
y <- Hitters$Salary
```

The `glmnet()` function has an alpha argument that determines what type of model is fit. If `alpha = 0` then a ridge regression model is fit, and if `alpha = 1` then a lasso model is fit. We first fit a ridge regression model, which minimizes

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

where $\lambda \geq 0$ is a *tuning parameter* to be determined.

Fit Ridge Regression over a grid of λ values

```r
grid <- 10^seq(10, -2, length = 100)
ridge.mod <- glmnet(X, y, alpha = 0, lambda = grid)
```

Ridge Regression Coefficents

```r
dim(coef(ridge.mod))
```

```
## [1]  20 100
```

We expect the coefficient estimates to be much smaller, in terms of ℓ_2 norm, when a large value of λ is used.

```r
ridge.mod$lambda[50] #Display 50th lambda value
```

```
## [1] 11497.57
```

```r
coef(ridge.mod)[, 50] # Display coefficients associated with 50th lambda value
```

```
##   (Intercept)         AtBat          Hits         HmRun          Runs
## 407.356050200   0.036957182   0.138180344   0.524629976   0.230701523
##           RBI         Walks         Years        CAtBat         CHits
##   0.239841459   0.289618741   1.107702929   0.003131815   0.011653637
##        CHmRun         CRuns          CRBI        CWalks       LeagueN
##   0.087545670   0.023379882   0.024138320   0.025015421   0.085028114
##     DivisionW       PutOuts       Assists        Errors    NewLeagueN
##  -6.215440973   0.016482577   0.002612988  -0.020502690   0.301433531
```

```r
sqrt(sum(coef(ridge.mod)[-1, 50]^2)) # Calculate l2 norm
```

```
## [1] 6.360612
```

In contrast, here are the coefficients when $\lambda = 705$, along with their ℓ_2 norm. Note the much larger l_2 norm of the coefficients associated with this smaller value of λ.

```r
ridge.mod$lambda[60] #Display 60th lambda value
```

```
## [1] 705.4802
```

```r
coef(ridge.mod)[, 60] # Display coefficients associated with 60th lambda value
```

```
##  (Intercept)        AtBat         Hits        HmRun         Runs          RBI
##  54.32519950   0.11211115   0.65622409   1.17980910   0.93769713   0.84718546
##        Walks        Years       CAtBat        CHits       CHmRun        CRuns
##   1.31987948   2.59640425   0.01083413   0.04674557   0.33777318   0.09355528
##         CRBI       CWalks      LeagueN    DivisionW      PutOuts      Assists
##   0.09780402   0.07189612  13.68370191 -54.65877750   0.11852289   0.01606037
##       Errors   NewLeagueN
##  -0.70358655   8.61181213
```

```r
sqrt(sum(coef(ridge.mod)[-1, 60]^2)) # Calculate l2 norm
```

```
## [1] 57.11001
```

We can use the `predict()` function for a number of purposes. For instance, we can obtain the ridge regression coefficients for a new value of λ, say 50:

```r
predict(ridge.mod, s = 50, type = "coefficients")[1:20, ]
```

```
##   (Intercept)         AtBat          Hits         HmRun          Runs
##  4.876610e+01 -3.580999e-01  1.969359e+00 -1.278248e+00  1.145892e+00
##           RBI         Walks         Years        CAtBat         CHits
##  8.038292e-01  2.716186e+00 -6.218319e+00  5.447837e-03  1.064895e-01
##        CHmRun         CRuns          CRBI        CWalks       LeagueN
##  6.244860e-01  2.214985e-01  2.186914e-01 -1.500245e-01  4.592589e+01
##     DivisionW       PutOuts       Assists        Errors    NewLeagueN
## -1.182011e+02  2.502322e-01  1.215665e-01 -3.278600e+00 -9.496680e+00
```

Training/Testing

We now split the samples into a training set and a test set in order to estimate the test error of ridge regression and later on the lasso.

```r
set.seed(1)
train <- sample(1:nrow(X), nrow(X) / 2)
test <- (-train)
y.test <- y[test]

# Fit Ridge regression to the training data
ridge.mod <- glmnet(X[train,], y[train], alpha = 0, lambda = grid, thresh = 1e-12)
# Predcit the salary to the testing data with lambda = 4
ridge.pred <- predict(ridge.mod, s = 4, newx = X[test,])
# Calculate the Root Mean Square Error (RMSE)
sqrt(mean((ridge.pred - y.test)^2))
```

```
## [1] 377.093
```

```r
# Compute the RMSE for the intercept-only model
sqrt(mean((mean(y[train]) - y.test)^2))
```

```
## [1] 473.9936
```

```r
# Change to a much larger lambda
ridge.pred <- predict(ridge.mod, s = 1e10, newx = X[test,])
sqrt(mean((ridge.pred - y.test)^2))
```

```
## [1] 473.9935
```

```r
# Change lambda to 0
ridge.pred <- predict(ridge.mod, s = 0, newx = X[test,])
sqrt(mean((ridge.pred - y.test)^2))
```

```
## [1] 409.6215
```

```r
lm(y ~ X, subset = train)
```

```
##
## Call:
## lm(formula = y ~ X, subset = train)
##
## Coefficients:
## (Intercept)       XAtBat        XHits       XHmRun        XRuns         XRBI
##    274.0145      -0.3521      -1.6377       5.8145       1.5424       1.1243
##      XWalks       XYears      XCAtBat       XCHits      XCHmRun       XCRuns
##      3.7287     -16.3773      -0.6412       3.1632       3.4008      -0.9739
```

```
##       XCRBI      XCWalks     XLeagueN   XDivisionW     XPutOuts     XAssists
##     -0.6005       0.3379     119.1486    -144.0831       0.1976       0.6804
##     XErrors  XNewLeagueN
##     -4.7128     -71.0951
```

```r
predict(ridge.mod, s = 0, type = "coefficients")[1:20,]
```

```
##  (Intercept)        AtBat         Hits        HmRun         Runs          RBI
##  274.2089049   -0.3699455   -1.5370022    5.9129307    1.4811980    1.0772844
##        Walks        Years       CAtBat        CHits       CHmRun        CRuns
##    3.7577989  -16.5600387   -0.6313336    3.1115575    3.3297885   -0.9496641
##         CRBI       CWalks      LeagueN    DivisionW      PutOuts      Assists
##   -0.5694414    0.3300136  118.4000592 -144.2867510    0.1971770    0.6775088
##       Errors   NewLeagueN
##   -4.6833775  -70.1616132
```

Instead of arbitrarily choosing $\lambda = 4$, it would be better to use cross-validation (CV) to choose the tuning parameter λ. We can do this using the built-in cross-validation function, `cv.glmnet()`. By default, the function performs 10-fold cross-validation, though this can be changed using the argument `folds`.

Cross-Validation (CV)

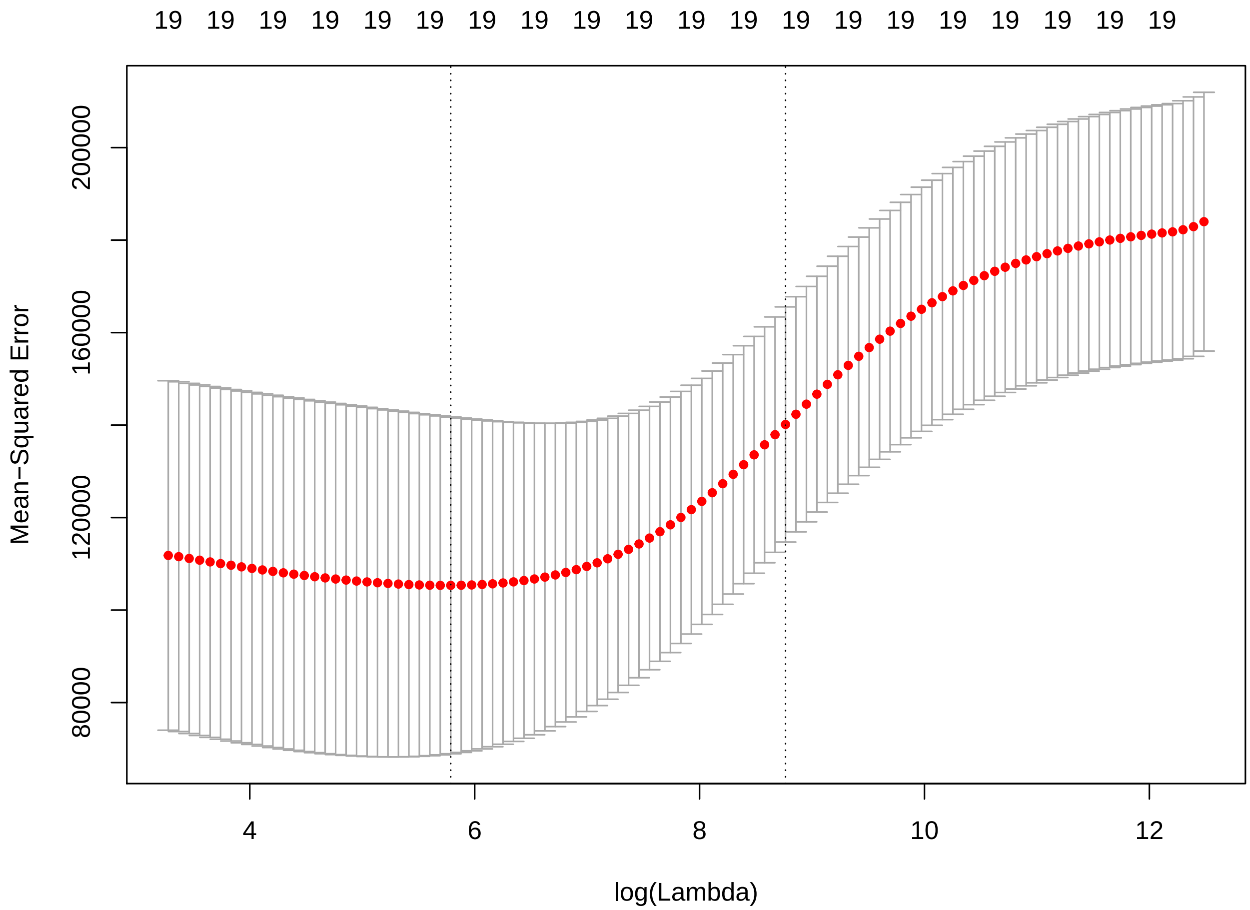
```r
set.seed(1)
# Fit ridge regression model on training data
cv.out <- cv.glmnet(X[train,], y[train], alpha = 0)
# Select lamda that minimizes training MSE
bestLambda = cv.out$lambda.min
bestLambda
```

```
## [1] 326.0828
```

```r
ridge.pred <- predict(ridge.mod, s = bestLambda, newx = X[test,])
sqrt(mean((ridge.pred - y.test)^2))
```

```
## [1] 373.9741
```

```r
plot(cv.out) # Draw plot of training MSE as a function of lambda
```



Finally, we refit our ridge regression model on the full data set, using the value of λ chosen by cross-validation, and examine the coefficient estimates.

```
# Fit ridge regression model on full dataset
out <- glmnet(X, y, alpha = 0)
# Display coefficients using lambda chosen by CV
predict(out, type = "coefficients", s = bestLambda)[1:20,]
```

```
##  (Intercept)        AtBat         Hits        HmRun         Runs          RBI
##  15.44383135   0.07715547   0.85911581   0.60103107   1.06369007   0.87936105
##        Walks        Years       CAtBat        CHits       CHmRun        CRuns
##   1.62444616   1.35254780   0.01134999   0.05746654   0.40680157   0.11456224
##         CRBI       CWalks      LeagueN    DivisionW      PutOuts      Assists
##   0.12116504   0.05299202  22.09143189 -79.04032637   0.16619903   0.02941950
##       Errors   NewLeagueN
##  -1.36092945   9.12487767
```

The Lasso

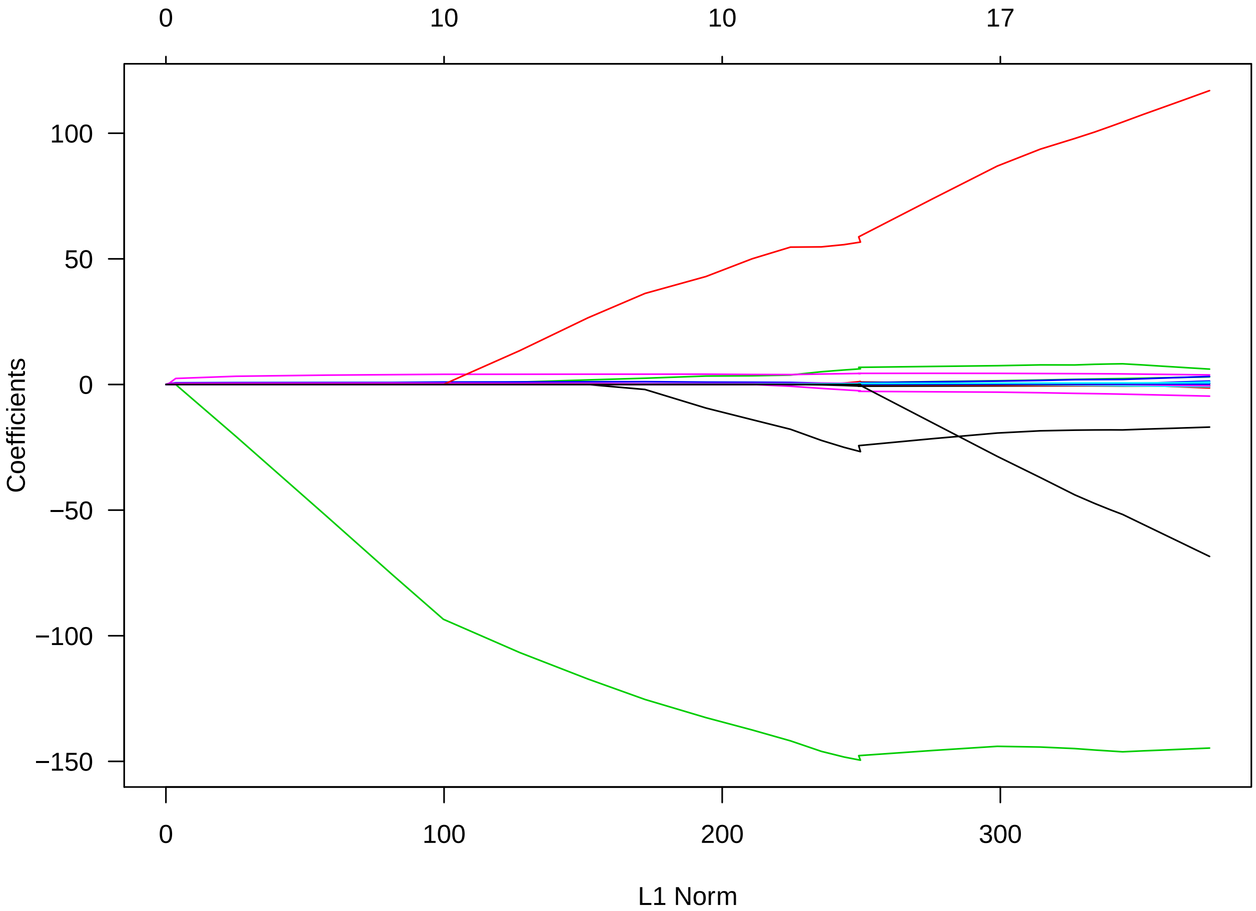
We saw that ridge regression with a wise choice of λ can outperform least squares as well as the null model on the Hitters data set. We now ask whether the lasso, which minimizes

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

can yield either a more accurate or a more interpretable model than ridge regression. In order to fit a lasso model, we once again use the `glmnet()` function; however, this time we use the argument `alpha=1`.

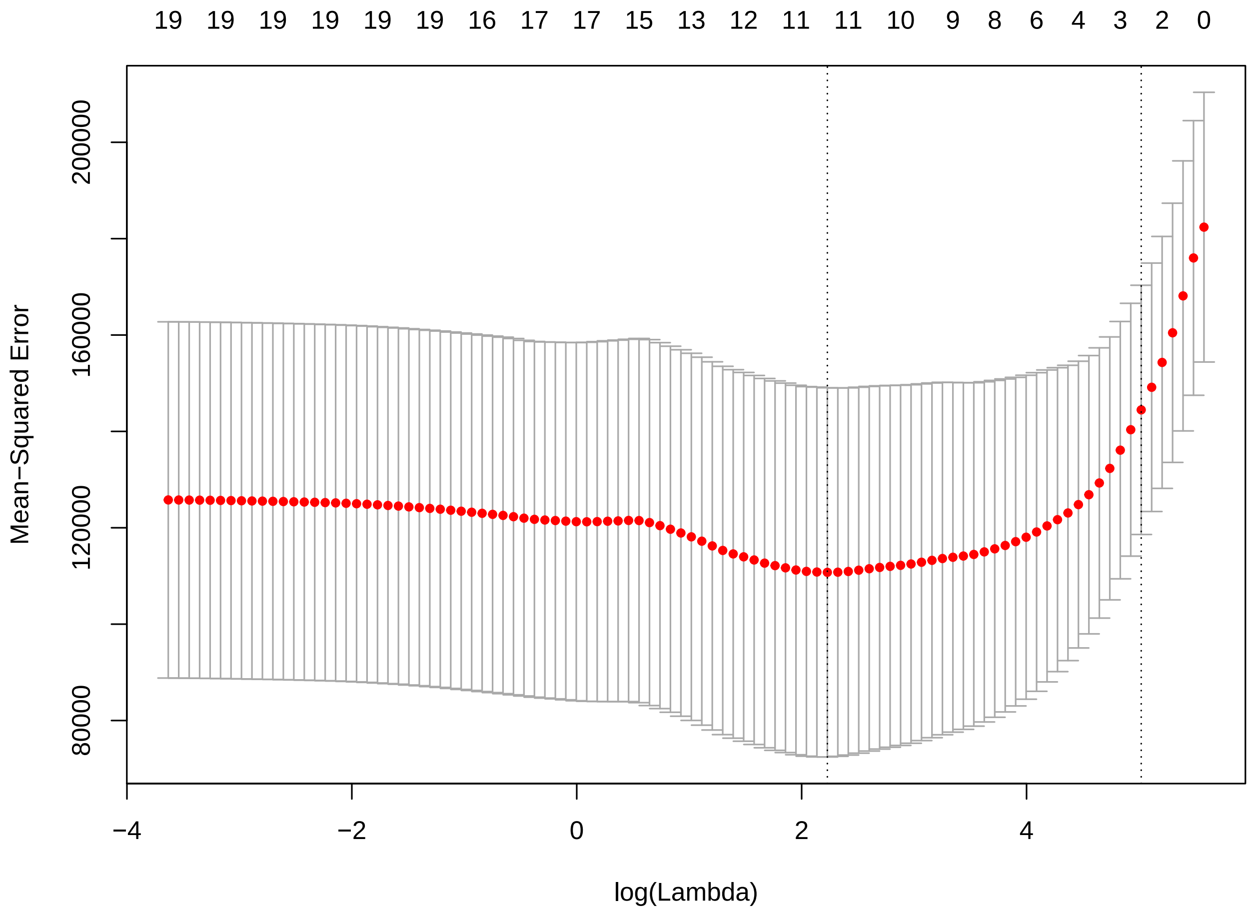
```r
# Fit lasso model on training data
lasso.mod <- glmnet(X[train,], y[train], alpha = 1, lambda = grid)
# Draw plot of coefficients
plot(lasso.mod, las = 1)
```

```
## Warning in regularize.values(x, y, ties, missing(ties)): collapsing to unique
## 'x' values
```



Notice that in the coefficient plot that depending on the choice of tuning parameter, some of the coefficients are exactly equal to zero. We now perform cross-validation and compute the associated test error:

```r
set.seed(1)
# Fit lasso model on training data
cv.out <- cv.glmnet(X[train,], y[train], alpha = 1)
# Draw plot of training MSE as a function of lambda
plot(cv.out)
```




```r
# Select lamda that minimizes training MSE
bestLambda <- cv.out$lambda.min
# Use best lambda to predict test data
lasso.pred <- predict(lasso.mod, s = bestLambda, newx = X[test,])
# Calculate test RMSE
sqrt(mean((lasso.pred - y[test])^2))
```

```
## [1] 379.043
```

This is substantially lower than the test set RMSE of the null model and of least squares, and very similar to the test RMSE of ridge regression with λ chosen by cross-validation.

However, the lasso has a substantial advantage over ridge regression in that the resulting coefficient estimates are sparse. Here we see that 8 of the 19 coefficient estimates are exactly zero:

```r
# Fit lasso model on full dataset
out <- glmnet(X, y, alpha = 1, lambda = grid)
# Display coefficients using lambda chosen by CV
lasso.coef <- predict(out, type = "coefficients", s = bestLambda)[1:20,]
lasso.coef
```

```
##   (Intercept)        AtBat         Hits        HmRun         Runs
##    1.27479059  -0.05497143   2.18034583   0.00000000   0.00000000
##           RBI        Walks        Years       CAtBat        CHits
##    0.00000000   2.29192406  -0.33806109   0.00000000   0.00000000
##        CHmRun        CRuns         CRBI       CWalks      LeagueN
##    0.02825013   0.21628385   0.41712537   0.00000000  20.28615023
```

```
##     DivisionW       PutOuts       Assists        Errors    NewLeagueN
## -116.16755870    0.23752385    0.00000000   -0.85629148    0.00000000
```

```
lasso.coef[lasso.coef != 0] # Display only non-zero coefficients
```

```
##   (Intercept)         AtBat          Hits         Walks         Years
##    1.27479059   -0.05497143    2.18034583    2.29192406   -0.33806109
##        CHmRun         CRuns          CRBI       LeagueN     DivisionW
##    0.02825013    0.21628385    0.41712537   20.28615023 -116.16755870
##       PutOuts        Errors
##    0.23752385   -0.85629148
```