# Homework_6

Daria Barbour-Brown | Bailey Ho | Warren Kennedy

2023-05-10

## Conceptuual Problem

### Question 1

*When would you want to use ridge regression instead of a standard linear regression?*

We should explore alternative methods to enhance both the accuracy and interpretability of our model predictions. One such method is ridge regression, which is a refined version of linear regression specifically designed for scenarios where the predictor variables exhibit high correlation. In situations where multicollinearity exists, traditional linear regression often yields unstable and unreliable coefficient estimates. By incorporating regularization, ridge regression effectively addresses this challenge by shrinking the coefficient estimates, reducing their variance, and stabilizing the model.

### Question 2

*When would you not want to use ridge regression?*

Ridge regression uses regularization to shrink the coefficient estimates towards zero, although they never actually reach zero. This regularization can compromise interpretability, particularly if the aim is to interpret the significance of individual coefficients in your analysis. In such instances, alternative approaches like lasso regression may be more suitable. Lasso regression selectively identifies the most influential variables, simplifying the model and facilitating interpretation by emphasizing the most important predictors.

## Application Question

### Question 3

#### Part A

```
library(ISLR2)
data <- Hitters
data <- subset(data, select = -c(League,Division,NewLeague))
data <- data[complete.cases(data), ]
dim(data)
```

```
## [1] 263  17
```

### Part B

When a variable's coefficient is reduced to zero by LASSO regression, it signifies that the variable does not possess any substantial influence on the model's prediction or that its impact is negligible when compared to other variables.

```
set.seed(1)
train <- sample(263, 263*.8)
```

### Part C

```
lm.fit <- lm(Salary ~. , data = data, subset = train)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = Salary ~ ., data = data, subset = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -788.74 -177.36  -34.21  120.01 1912.33
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 177.00090   92.90453   1.905   0.0582 .
## AtBat        -1.74153    0.71980  -2.419   0.0165 *
## Hits          5.10323    2.85500   1.787   0.0754 .
## HmRun        -1.21466    6.91295  -0.176   0.8607
## Runs         -2.13662    3.42167  -0.624   0.5331
## RBI           3.08012    2.94988   1.044   0.2977
## Walks         4.02286    2.13287   1.886   0.0608 .
## Years       -15.47918   13.66545  -1.133   0.2587
## CAtBat       -0.24281    0.16309  -1.489   0.1382
## CHits         0.89063    0.80962   1.100   0.2727
## CHmRun       -0.39810    1.86142  -0.214   0.8309
## CRuns         0.73578    0.88780   0.829   0.4083
## CRBI          0.50915    0.84437   0.603   0.5472
## CWalks       -0.34568    0.41629  -0.830   0.4073
## PutOuts       0.44746    0.09015   4.963 1.52e-06 ***
## Assists       0.39990    0.23408   1.708   0.0892 .
## Errors       -4.67595    4.69122  -0.997   0.3201
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 315 on 193 degrees of freedom
## Multiple R-squared:  0.5259, Adjusted R-squared:  0.4866
## F-statistic: 13.38 on 16 and 193 DF,  p-value: < 2.2e-16
```

### Part D

```r
RMSE = sqrt(mean((data$Salary[-train] - predict(lm.fit, data[-train,]))^2))
RMSE
```

```
## [1] 395.7729
```

### Part E

We anticipate that the Residual Standard Error (RSE) will be smaller, mainly because the Root Mean Squared Error (RMSE) is calculated by dividing the RSS by a larger number. Additionally, it is important to note that the RSS in Part C is derived from our training data, while the RSS in Part D pertains to the testing data. In general, we typically observe that the training RSS tends to be lower than the testing RSS. This occurrence is due to the fact that our model is specifically trained on the training data, which it analyzes during the training process. By doing so, it attempts to find the underlying patterns and relationships within that data. However, when the model encounters testing data with different patterns or relationships, it may struggle to capture the relationships that are unfamiliar to our model. Consequently, this disparity in data distribution can result in a comparatively higher testing RSS when contrasted with the training RSS.

## Question 4

### Part A

```r
library(glmnet)
```

```
## Loading required package: Matrix
```

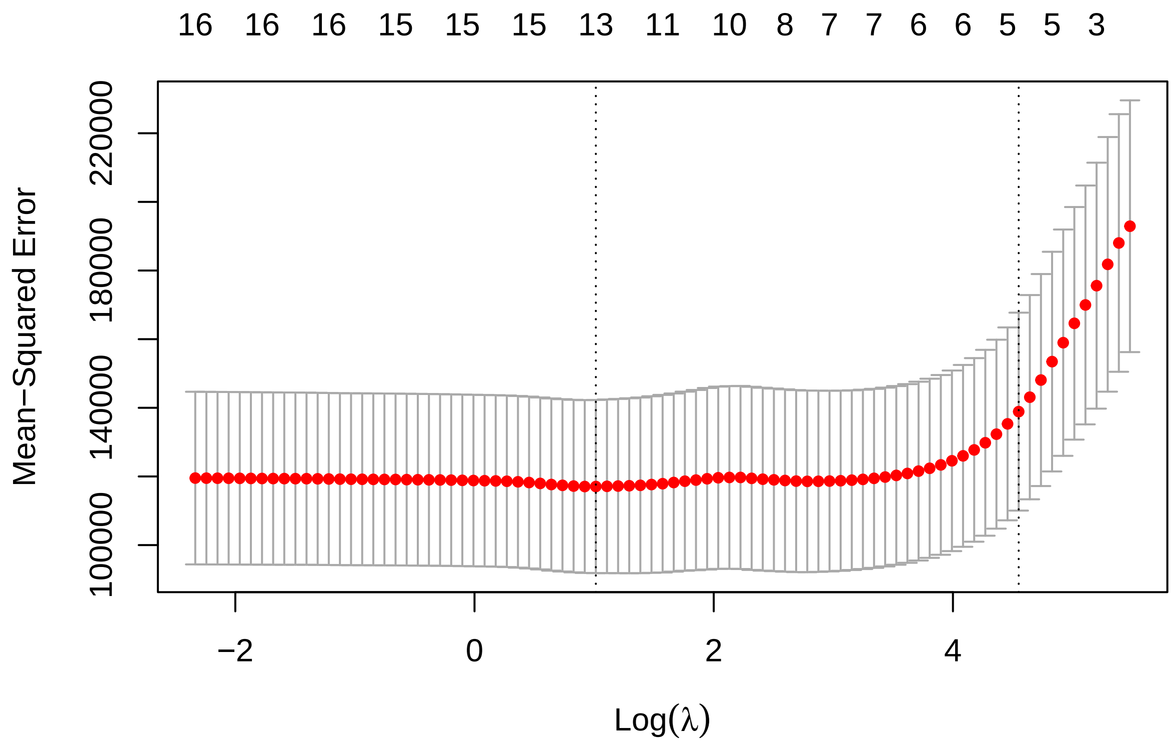
```
## Loaded glmnet 4.1-7
```

```r
x <- model.matrix(Salary ~ ., data)[, -1]
y <- data$Salary

set.seed(1)
lasso.fit <- cv.glmnet(x[train, ], y[train], alpha = 1)
lasso.fit
```

```
##
## Call:  cv.glmnet(x = x[train, ], y = y[train], alpha = 1)
##
## Measure: Mean-Squared Error
##
##     Lambda Index Measure    SE Nonzero
## min   2.76    49  117059 25231      13
## 1se  94.61    11  138884 28841       5
```

```r
plot(lasso.fit)
```




```
bestlam.lss <- lasso.fit$lambda.min
bestlam.lss
```

```
## [1] 2.757799
```

**Part B**

```
coef(lasso.fit)
```

```
## 17 x 1 sparse Matrix of class "dgCMatrix"
##                       s1
## (Intercept) 223.5746052
## AtBat         .
## Hits          0.5396439
## HmRun         .
## Runs          .
## RBI           .
## Walks         1.2242333
## Years         .
## CAtBat        .
## CHits         0.1174027
## CHmRun        .
## CRuns         .
```

```
## CRBI          0.2219827
## CWalks        .
## PutOuts       0.1721853
## Assists       .
## Errors        .
```

### Part C

```r
set.seed(1)
test <- (-train)
y.test <- y[test]

lasso.pred <- predict(lasso.fit, s = bestlam.lss, newx = x[test, ])
mean((lasso.pred - y.test)^2)
```

```
## [1] 155073.8
```

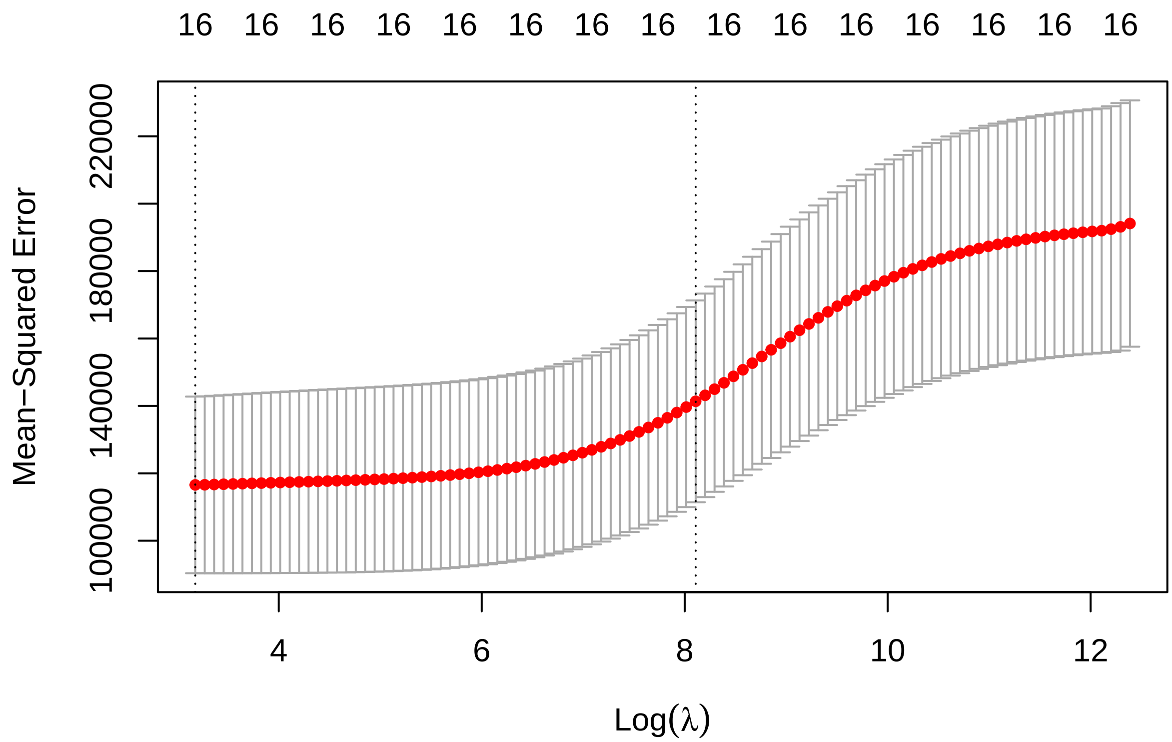
## Question 5

### Part A

```r
ridge.fit <- cv.glmnet(x[train, ], y[train], alpha = 0)
ridge.fit
```

```
##
## Call:  cv.glmnet(x = x[train, ], y = y[train], alpha = 0)
##
## Measure: Mean-Squared Error
##
##     Lambda Index Measure    SE Nonzero
## min     24   100  116566 26201      16
## 1se   3322    47  141365 29937      16
```

```r
plot(ridge.fit)
```




```
bestlam.rr <- ridge.fit$lambda.min
bestlam.rr
```

```
## [1] 23.98593
```

**Part B**

The intercept in ridge regression exhibits a greater magnitude compared to our linear estimates. Additionally, the coefficient estimates in ridge regression are, in many cases, closer to zero when compared to the estimates in our linear model.

```
coef(ridge.fit)
```

```
## 17 x 1 sparse Matrix of class "dgCMatrix"
##                       s1
## (Intercept) 250.383896676
## AtBat         0.069678812
## Hits          0.286890025
## HmRun         0.906816756
## Runs          0.454298095
## RBI           0.476771168
## Walks         0.672986779
## Years         2.070773930
## CAtBat        0.007352379
```

```
## CHits          0.029216322
## CHmRun         0.172419128
## CRuns          0.055475339
## CRBI           0.055900249
## CWalks         0.058234872
## PutOuts        0.060078683
## Assists        0.010632032
## Errors        -0.152639646
```

```
coef(lm.fit)
```

```
## (Intercept)       AtBat        Hits       HmRun        Runs         RBI
## 177.0008998  -1.7415256   5.1032332  -1.2146591  -2.1366221   3.0801204
##       Walks       Years      CAtBat       CHits      CHmRun       CRuns
##   4.0228564 -15.4791836  -0.2428135   0.8906328  -0.3980951   0.7357843
##        CRBI      CWalks     PutOuts     Assists      Errors
##   0.5091505  -0.3456766   0.4474601   0.3999002  -4.6759469
```

### Part C

```
ridge.pred <- predict(ridge.fit, s = bestlam.rr, newx = x[test, ])
ridge.pred
```

```
##                           s1
## -Andre Dawson      949.85837
## -Andres Galarraga  549.92155
## -Al Newman         106.43309
## -Andres Thomas     154.58464
## -Alex Trevino      315.09079
## -Barry Bonds       342.20301
## -Bill Buckner     1556.54955
## -Carlton Fisk      925.32168
## -Chris Speier      583.55044
## -Doug DeCinces     802.33537
## -Darrell Evans    1351.17563
## -Dan Gladden       371.68972
## -Dave Henderson    425.56860
## -Dale Murphy       933.26812
## -Don Slaught       473.66637
## -Eddie Milner      422.00707
## -Glenn Braggs       94.47897
## -George Brett     1151.53968
## -George Hendrick   895.41437
## -Gary Redus        345.50554
## -Gary Ward         658.28299
## -Howard Johnson    198.11654
## -Jose Cruz        1088.49440
## -Jeffrey Leonard   457.46062
## -Jerry Mumphrey    619.66122
## -Jim Rice         1445.93949
## -Joel Skinner      207.40256
```

```
## -Kevin Bass        548.28906
## -Ken Griffey       859.06851
## -Ken Phelps        323.10433
## -Len Dykstra       442.60248
## -Lee Lacy          614.61700
## -Larry Sheets      238.44927
## -Mike Kingery      139.19749
## -Mike Marshall     346.30605
## -Ozzie Virgil      541.43760
## -Phil Bradley      583.72239
## -Paul Molitor      638.33103
## -Pete Rose        1740.04439
## -Pat Tabler        762.90128
## -Ron Hassey        519.72387
## -Rickey Henderson  884.50191
## -Ray Knight        623.68819
## -Rick Schu         140.94392
## -Steve Balboni     810.46817
## -Steve Garvey     1609.18080
## -Steve Jeltz       305.40352
## -Tony Gwynn        690.13904
## -Ted Simmons      1001.61748
## -Vince Coleman     359.85531
## -Wally Joyner      931.92586
## -Willie Randolph   923.46115
## -Wayne Tolleson    328.17524
```

```
mean((ridge.pred - y.test)^2)
```

```
## [1] 150137.8
```

## Question 6

*Which model would you recommend using if the General Manager of a baseball team is interested knowing which variables are most important for predicting a players salary?*

In this scenario, our preference would be to utilize the LASSO estimates when reporting back to the General Manager. Although both LASSO and ridge regression can be employed for predictive modeling, LASSO's advantage lies in its ability to exclude less significant predictors, whereas ridge regression retains all predictors, thereby reducing model interpretability. Given the General Manager's primary interest in the most important variables, we opt for LASSO as it enables subset selection, emphasizing the variables of greatest relevance.