5291 hw4

Yijin Wang

1. Perform a multiple linear regression model of 'bwt' birth weight in grams on the explanatory variables i)Investigate whether there is any multicollinearity

ii)Run a ridge regression analysis and compare the results with the OLS results

```
#(i)
library(MASS)
data(birthwt)
mlr<-lm(bwt ~ age + lwt + race + smoke + ptl + ht + ui + ftv, data=birthwt)
mlr
```

```
##
## Call:
## lm(formula = bwt ~ age + lwt + race + smoke + ptl + ht + ui +
##     ftv, data = birthwt)
##
## Coefficients:
## (Intercept)          age          lwt         race        smoke          ptl
##   3129.4594      -0.2658       3.4351    -188.4895    -358.4552     -51.1526
##          ht           ui          ftv
##   -600.6465    -511.2513     -15.5358
```

```
summary(mlr)
```

```
##
## Call:
## lm(formula = bwt ~ age + lwt + race + smoke + ptl + ht + ui +
##     ftv, data = birthwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1816.51  -426.79    16.29   492.06  1654.01
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3129.4594   344.2424   9.091  < 2e-16 ***
## age           -0.2658     9.5947  -0.028  0.97793
## lwt            3.4351     1.6999   2.021  0.04478 *
## race        -188.4895    57.7339  -3.265  0.00131 **
## smoke       -358.4552   107.5172  -3.334  0.00104 **
## ptl          -51.1526   103.0003  -0.497  0.62006
## ht          -600.6465   204.3454  -2.939  0.00372 **
## ui          -511.2513   140.2792  -3.645  0.00035 ***
## ftv          -15.5358    46.9354  -0.331  0.74103
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 656.9 on 180 degrees of freedom
## Multiple R-squared:  0.223,	Adjusted R-squared:  0.1884
## F-statistic: 6.456 on 8 and 180 DF,  p-value: 2.232e-07
```

```
library(car)
```

```
## Loading required package: carData
```

```
vif(mlr)
```

```
##      age      lwt     race    smoke      ptl       ht       ui      ftv
## 1.125945 1.177116 1.224579 1.206096 1.124835 1.087378 1.087593 1.076820
```

```
#Since the vif for all variables are not high, there is no multicollinearity.

#(ii)
library(glmnet)
```

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

```
## Loaded glmnet 3.0-2
```

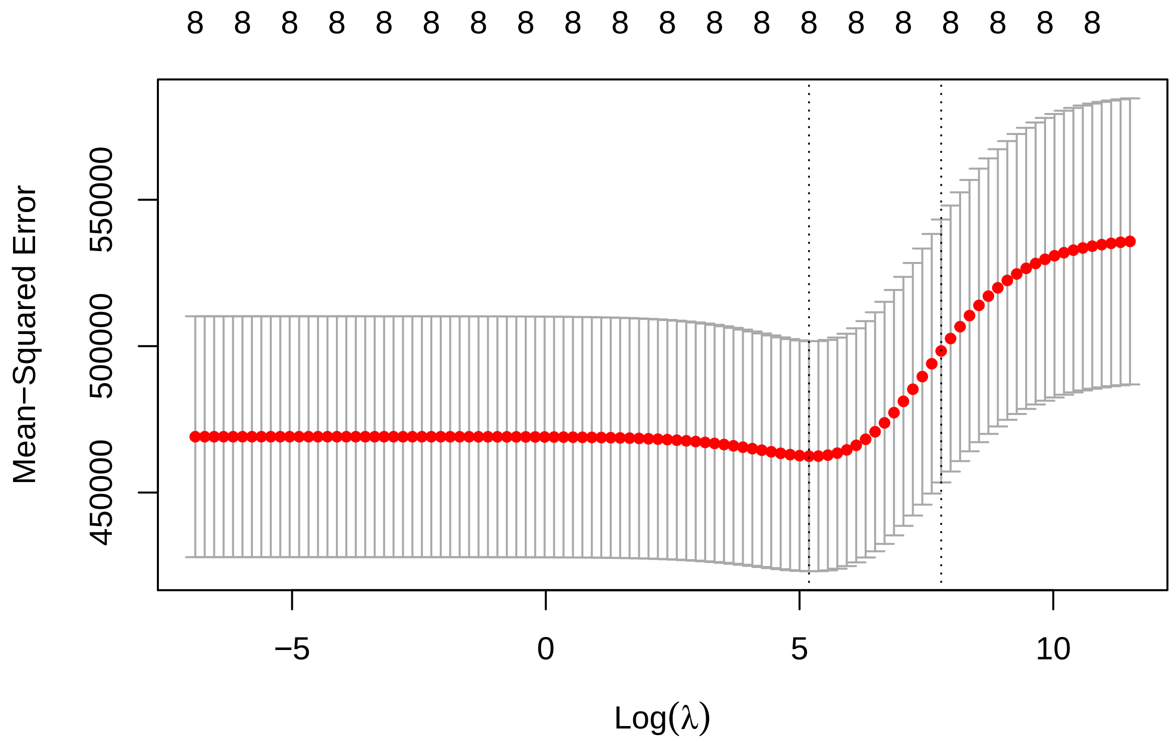
```
library(tidyverse)
```

```
## -- Attaching packages -------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.4
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ----------------------------------------- tidyverse_conflicts() --
## x tidyr::expand() masks Matrix::expand()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x tidyr::pack()   masks Matrix::pack()
## x dplyr::recode() masks car::recode()
## x dplyr::select() masks MASS::select()
## x purrr::some()   masks car::some()
## x tidyr::unpack() masks Matrix::unpack()
```

```
rr<-lm.ridge(bwt ~ age + lwt + race + smoke + ptl + ht + ui + ftv, data=birthwt)
x<-birthwt%>%select(age, lwt, race, smoke, ptl, ht, ui, ftv)%>%data.matrix()
y<-birthwt$bwt
lambdas<-10^seq(-3, 5, length.out = 100)
ridge_cv = cv.glmnet(x, y, alpha = 0, lambda = lambdas, standardize = TRUE, nfolds = 10)
plot(ridge_cv)
```



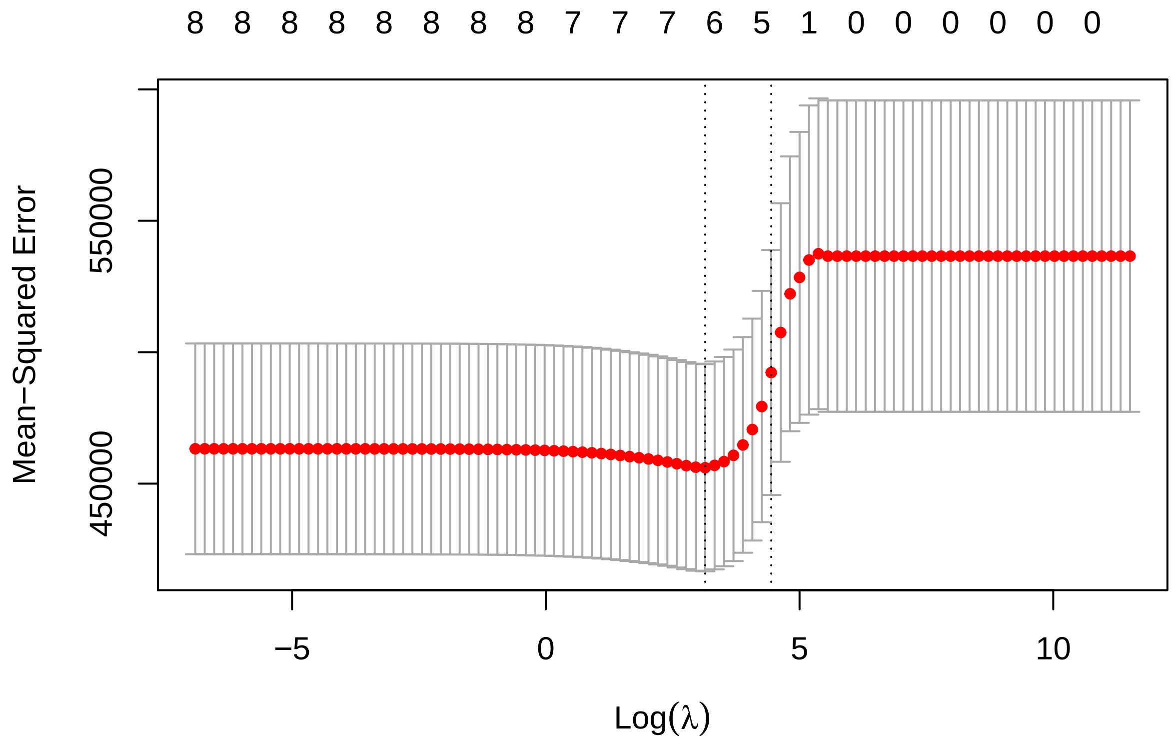

```
#best lambda
lambda_min<-ridge_cv$lambda.min
#final model
ridge_model_cv<-glmnet(x, y, alpha = 0, lambda = lambda_min, standardize = TRUE)
ridge.predict<-predict(ridge_model_cv, x)
cor(y, ridge.predict)^2
```

```
##             s0
## [1,] 0.2213511
```

Since the R^2 for both regression is very similar, we think both regression work.

2. Compare models selected using LASSO and a stepwise procedure to predict 'bwt' birth weight in grams using the above set of predictors.

```
#Lasso
lasso_cv<-cv.glmnet(x, y, alpha = 1, lambda = lambdas, standardize = TRUE, nfolds = 10)
plot(lasso_cv)
```




```
#best lambda
lambda_lasso<-lasso_cv$lambda.min
#final model
lasso_model_cv<-glmnet(x, y, alpha = 1, lambda = lambda_lasso, standardize = TRUE)
lasso.predict<-predict(lasso_model_cv, x)
cor(y, lasso.predict)^2
```

```
##              s0
## [1,] 0.2219398
```

```
#install.packages("lars")
library(lars)
```

```
## Loaded lars 1.2
```

```
lar1<-lars(x, y, type = "lasso")
lar1
```

```
##
## Call:
## lars(x = x, y = y, type = "lasso")
## R-squared: 0.223
## Sequence of LASSO moves:
##      ui race smoke lwt ht ptl ftv age
## Var   7    3     4   2  6   5   8   1
## Step  1    2     3   4  5   6   7   8
```

```
lar1$Cp[which.min(lar1$Cp)]
```

```
##        6
## 5.433202
```

```
#Lasso Predict
lar1.predict<-predict(lar1,x,s=7)$fit
```

```r
cor(y, lar1.predict)^2
```

```
## [1] 0.2223486
```

```r
#Stepwise
lar2<-lars(x, y, type = "stepwise")
lar2
```

```
##
## Call:
## lars(x = x, y = y, type = "stepwise")
## R-squared: 0.223
## Sequence of Forward Stepwise moves:
##      ui race smoke ht lwt ptl ftv age
## Var   7    3     4  6   2   5   8   1
## Step  1    2     3  4   5   6   7   8
```

```r
lar2$Cp[which.min(lar2$Cp)]
```

```
##        5
## 3.369843
```

```r
#Stepwise Predict
stepwise.predict<-predict(lar2,x,s=6)$fit
cor(y, stepwise.predict)^2
```

```
## [1] 0.2213582
```

Both models perform similarly in predicting y.

	OLS	Ridge	Lasso	Elastic Net
performance when p>>n	3	2	2	1
performance under multicollinearity	3	2	2	1
Unbiased estimators	1	2	2	2
Model selection capability	3	3	1	2
Simplicity: Computation Inference, Interpretation	1	2	3	3