GR5291 Advanced Data Analysis Homework 4

YIFEI LANG UNI:yl3365

Question 1

i)

We can load the data and do some exploratory data analysis first to get some basic information about the data.

```
library(MASS)
data("birthwt")
head(birthwt, obs = 5)
```

```
##    low age lwt race smoke ptl ht ui ftv  bwt
## 85   0  19 182    2     0   0  0  1   0 2523
## 86   0  33 155    3     0   0  0  0   3 2551
## 87   0  20 105    1     1   0  0  0   1 2557
## 88   0  21 108    1     1   0  0  1   2 2594
## 89   0  18 107    1     1   0  0  1   0 2600
## 91   0  21 124    3     0   0  0  0   0 2622
```

```
dim(birthwt)
```

```
## [1] 189  10
```

```
summary(birthwt)
```

```
##       low              age             lwt             race      
##  Min.   :0.0000   Min.   :14.00   Min.   : 80.0   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:19.00   1st Qu.:110.0   1st Qu.:1.000  
##  Median :0.0000   Median :23.00   Median :121.0   Median :1.000  
##  Mean   :0.3122   Mean   :23.24   Mean   :129.8   Mean   :1.847  
##  3rd Qu.:1.0000   3rd Qu.:26.00   3rd Qu.:140.0   3rd Qu.:3.000  
##  Max.   :1.0000   Max.   :45.00   Max.   :250.0   Max.   :3.000  
##      smoke             ptl               ht                ui        
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000  
##  Median :0.0000   Median :0.0000   Median :0.00000   Median :0.0000  
##  Mean   :0.3915   Mean   :0.1958   Mean   :0.06349   Mean   :0.1481  
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000  
##  Max.   :1.0000   Max.   :3.0000   Max.   :1.00000   Max.   :1.0000  
##       ftv              bwt      
##  Min.   :0.0000   Min.   : 709  
##  1st Qu.:0.0000   1st Qu.:2414  
##  Median :0.0000   Median :2977  
##  Mean   :0.7937   Mean   :2945  
##  3rd Qu.:1.0000   3rd Qu.:3487  
##  Max.   :6.0000   Max.   :4990  
```

Then we can fit a multiple linear regression model by using the following code:

```
reg1 = lm(bwt~age+lwt+race+smoke+ptl+ht+ui+ftv,data=birthwt)
summary(reg1)
```

```
##
## 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
```

```
confint(reg1, level=0.95) # CIs for model parameters
```

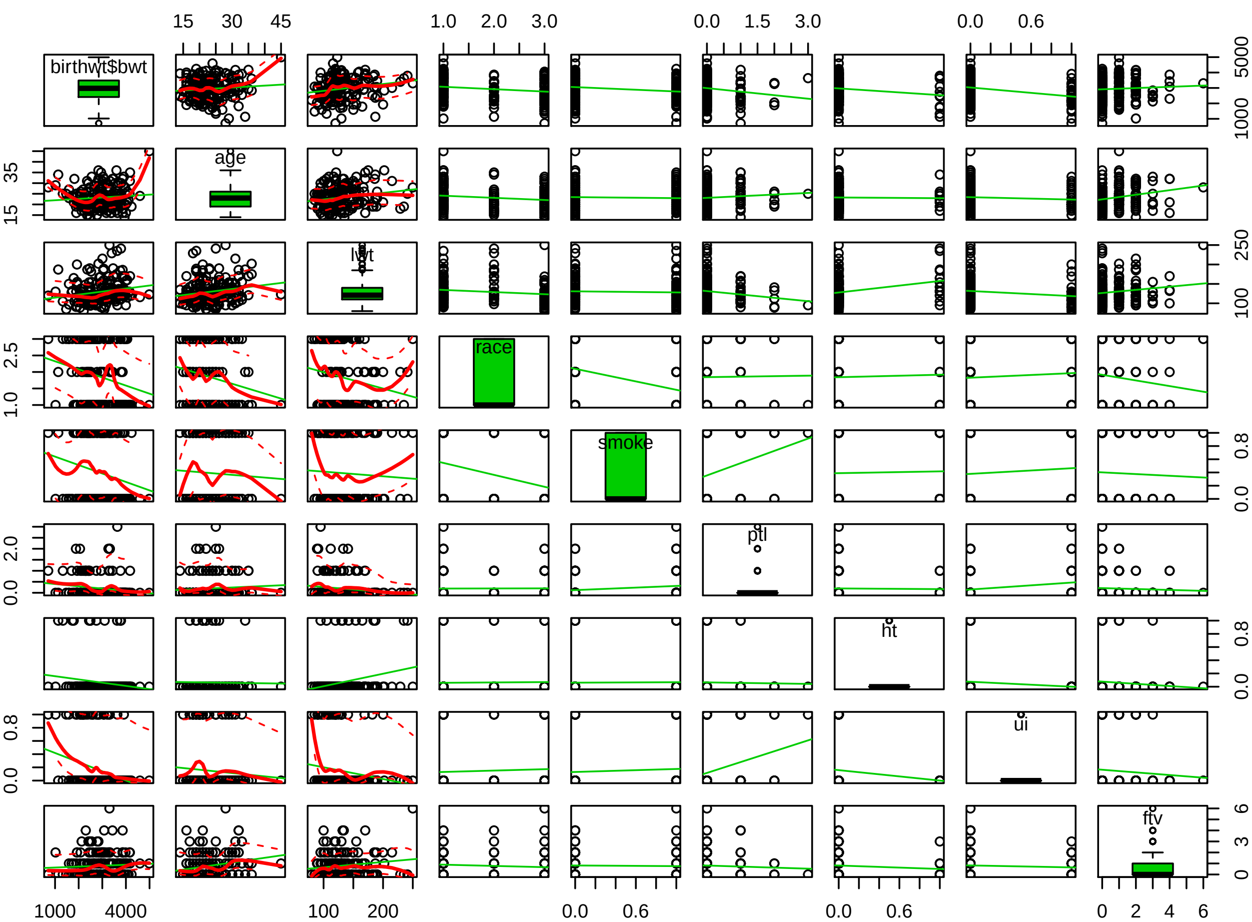
```
##                     2.5 %      97.5 %
## (Intercept)  2450.1897678 3808.729007
## age           -19.1984464   18.666826
## lwt             0.0808384    6.789424
## race         -302.4118096  -74.567219
## smoke        -570.6114961 -146.298879
## ptl          -254.3958763  152.090759
## ht          -1003.8672029 -197.425849
## ui           -788.0544699 -234.448037
## ftv          -108.1501301   77.078533
```

From the results above, the model is $bwt = 3129.46 - 0.27 * age + 3.44 * lwt - 188.49 * race - 358.46 * smoke - 51.15 * ptl - 600.65 * ht - 511.25 * ui - 15.54 * ftv$. However, we can find that the *'age'*, *'ptl'* and *'ftv'* are not significant in the model. To test whether there is **Multicollinearity** within the predictors. There are many methods:

```
library(car)
birth = cbind(birthwt$bwt, birthwt[,c(-1,-10)])
cor(birth)
```

```
##              birthwt$bwt         age         lwt         race       smoke
## birthwt$bwt   1.00000000  0.09031781  0.18573328 -0.194713487 -0.19044806
## age           0.09031781  1.00000000  0.18007315 -0.172817953 -0.04434618
## lwt           0.18573328  0.18007315  1.00000000 -0.165048544 -0.04417908
## race         -0.19471349 -0.17281795 -0.16504854  1.000000000 -0.33903074
## smoke        -0.19044806 -0.04434618 -0.04417908 -0.339030745  1.00000000
## ptl          -0.15465339  0.07160639 -0.14002900  0.007951293  0.18755706
## ht           -0.14598189 -0.01583700  0.23636040  0.019929917  0.01340704
## ui           -0.28392741 -0.07515558 -0.15276317  0.053602088  0.06215900
## ftv           0.05831777  0.21539394  0.14052746 -0.098336254 -0.02801314
##                       ptl          ht          ui         ftv
## birthwt$bwt -0.154653390 -0.14598189 -0.28392741  0.05831777
## age          0.071606386 -0.01583700 -0.07515558  0.21539394
## lwt         -0.140029003  0.23636040 -0.15276317  0.14052746
## race         0.007951293  0.01992992  0.05360209 -0.09833625
## smoke        0.187557063  0.01340704  0.06215900 -0.02801314
## ptl          1.000000000 -0.01539958  0.22758534 -0.04442966
## ht          -0.015399579  1.00000000 -0.10858506 -0.07237255
## ui           0.227585340 -0.10858506  1.00000000 -0.05952341
## ftv         -0.044429660 -0.07237255 -0.05952341  1.00000000
```

```
scatterplotMatrix(birth,var.labels = colnames(birth),diagonal = "boxplot")
```



```r
sqrt(vif(reg1))
```

```
##      age      lwt     race    smoke      ptl       ht       ui      ftv
## 1.061106 1.084950 1.106607 1.098224 1.060582 1.042774 1.042877 1.037699
```

```r
eigen(cor(birth))$value
```

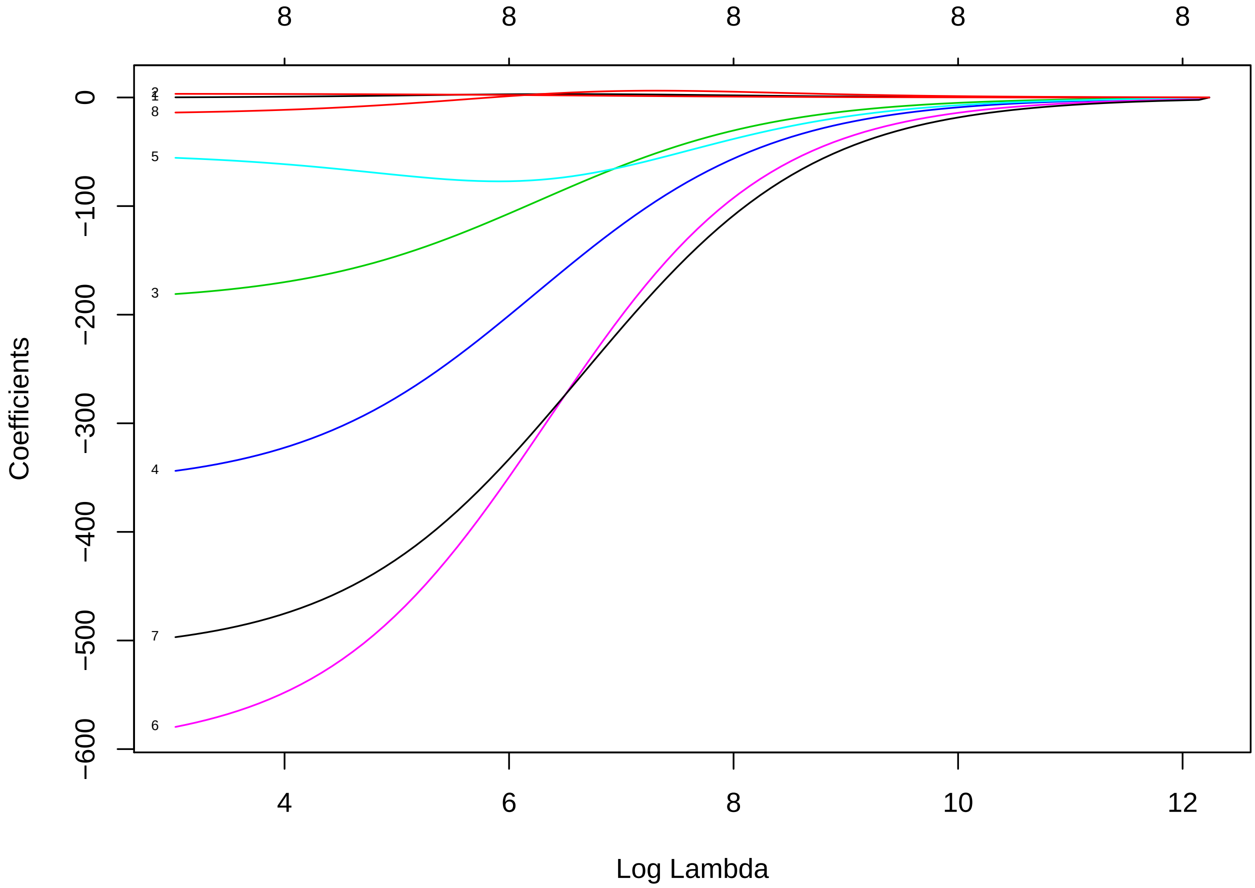
```
## [1] 1.7993879 1.4496246 1.2176971 1.1102150 0.8791301 0.7995185 0.7114200
## [8] 0.5863056 0.4467011
```

From the correlation matrix, we can find most of the correlations between the predictors are low. There is no correlation larger than 0.5, so that it is hard to say that multicollinearity exists. Then when we take a look at the scatterplot matrix, we can find that most of the predictors are categotical variables. So it is reasonable to find no multicollinearity. Finally, to have a numerical test about the collineartiy, we can check the sqare root of VIF, which also prove that there is no multicollinearity between predictors because all of the sqare root of VIF are less than 2. The eigenvalue of the correlaiton matrix also prove our conclusion since there is no eigenvalue near 0.

ii)

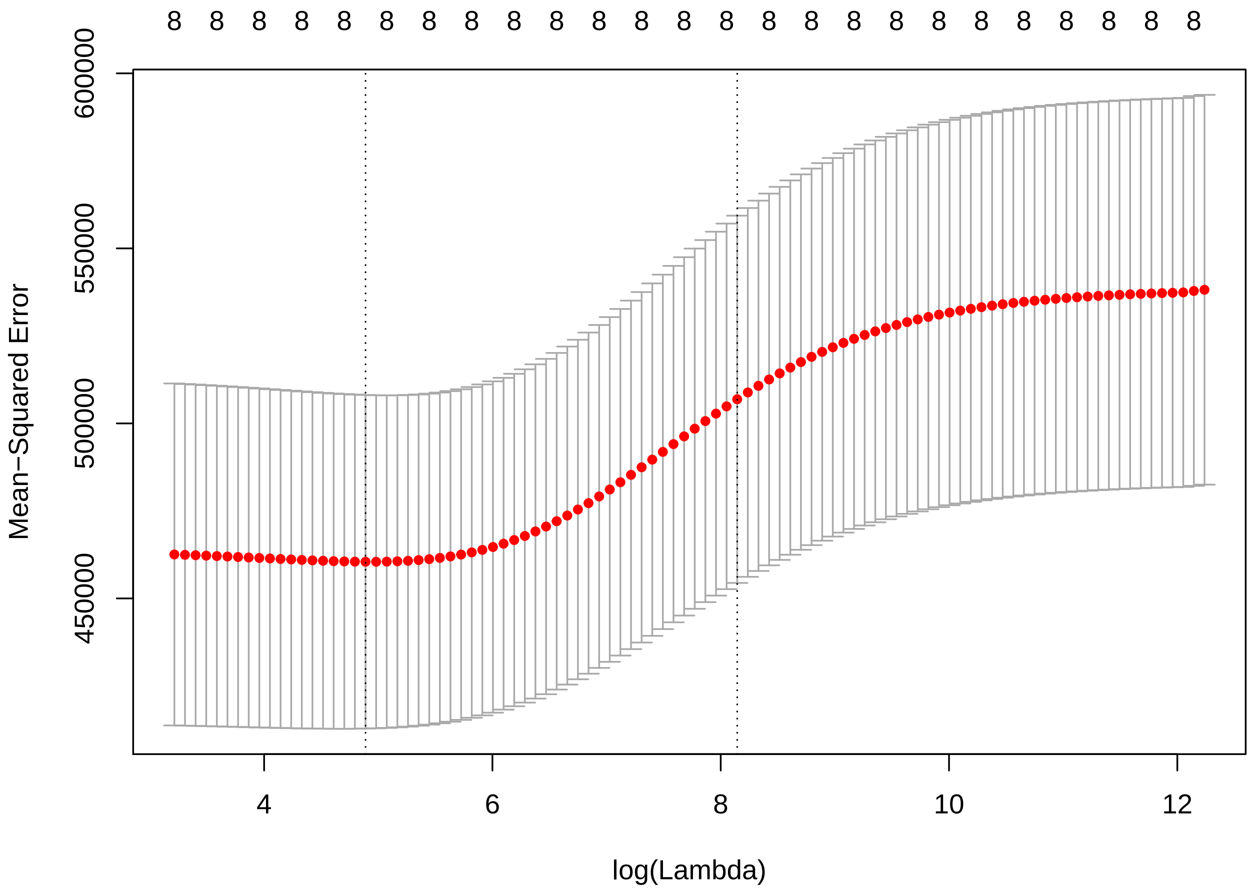
To conduce ridge regression, we can use the `fit.ridge` and `cv.ridge` to select the best shrinkage parameter and the `lm.ridge` in the library `MASS` to find the estimates of parameters.

```r
attach(birthwt)
fit.ridge = glmnet(cbind(age,lwt,race,smoke,ptl,ht,ui,ftv),bwt,alpha=0)
plot(fit.ridge,xvar="lambda",label=TRUE)
```



From the fist plot of coefficients against Log Lambda, we can find that when the shrinkage parameter is small, which means the model is close to OLS model, some of the coefficients are pretty large. And as the lambda increasing, the coefficients are concentrated. To find out when can we get the best model, we can use `cv.ridge` to conduct cross-validation to the data.

```
cv.ridge = cv.glmnet(cbind(age,lwt,race,smoke,ptl,ht,ui,ftv),bwt,alpha=0)
plot(cv.ridge)
```



From the results above, we can find that in the beginning, the mean squared error is very high, and the coefficients are restricted to be too small, and then at some point, it kind of levels off. This seems to indicate that the full model is doing a good job.

There's two vertical lines.

- The left one is at the minimum Mean-Squared Error
- The other vertical line is within one standard error of the minimum. The second line is a slightly more restricted model that does almost as well as the minimum.

At the top of the plot, we actually see how many non-zero variables coefficients are in the model. There's all 8 variables in the model and no coefficient is zero. There's a coefficient function extractor that works on a cross validation object and pick the coefficient vector corresponding to the best model:

```
coef(cv.ridge)
```

```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##                         1
## (Intercept) 2902.8202681
## age            1.7685774
## lwt            0.7119818
## race         -26.9861497
## smoke        -49.9217880
## ptl          -34.5704831
## ht           -81.4725183
```

```
## ui          -96.9745124
## ftv           4.9489724
```

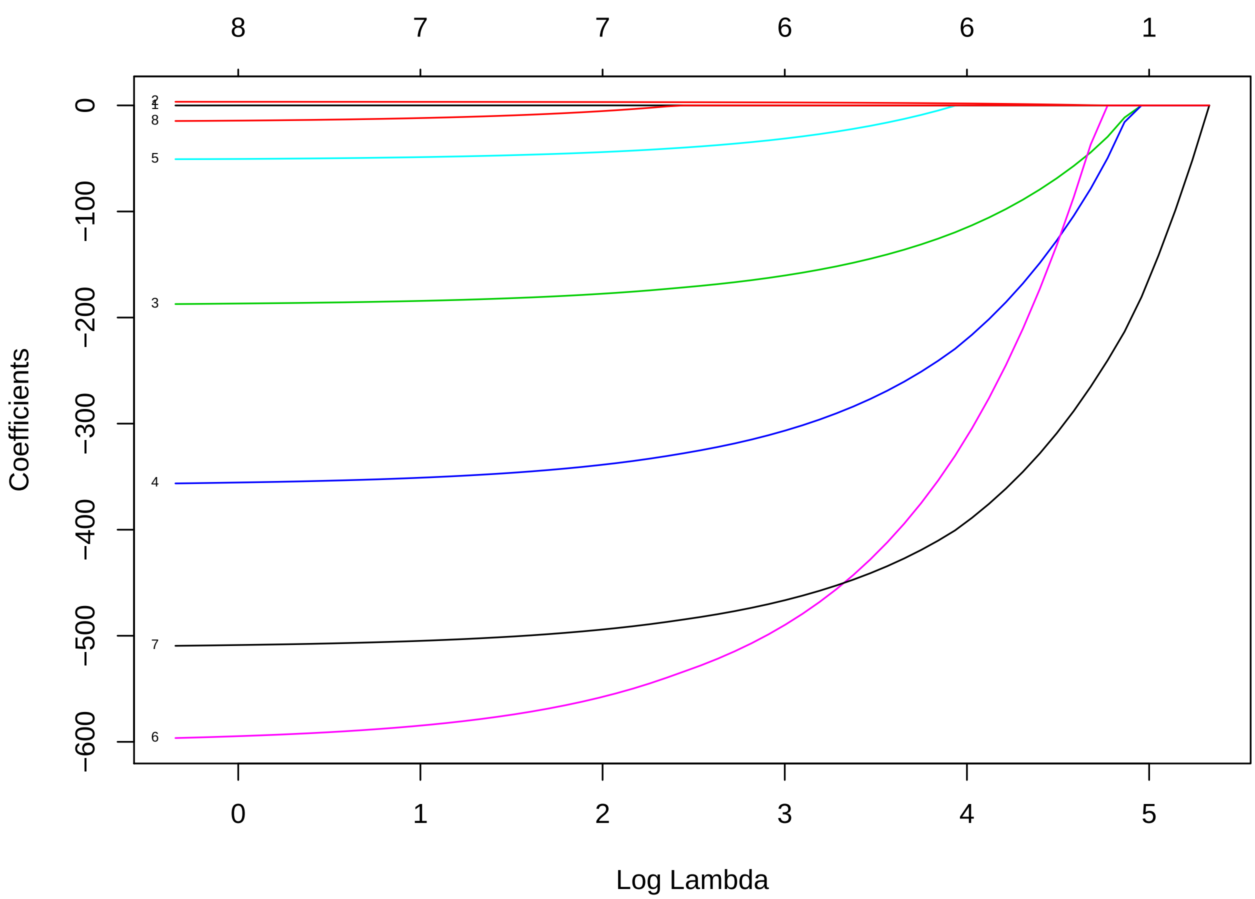
From the results above, the model is $bwt = 2914.77 + 1.18 * age + 0.45 * lwt - 16.65 * race - 30.72 * smoke - 22.60 * ptl - 49.11 * ht - 60.90 * ui + 3.53 * ftv$. Compare this model to the linear regression results, we can find that all the coefficient were shrunk, and all 8 predictors are included which is not same as the result of linear regression. The sign of the coeffecients of *'age'* and *'ftv'* are changed.

Question 2

Lasso

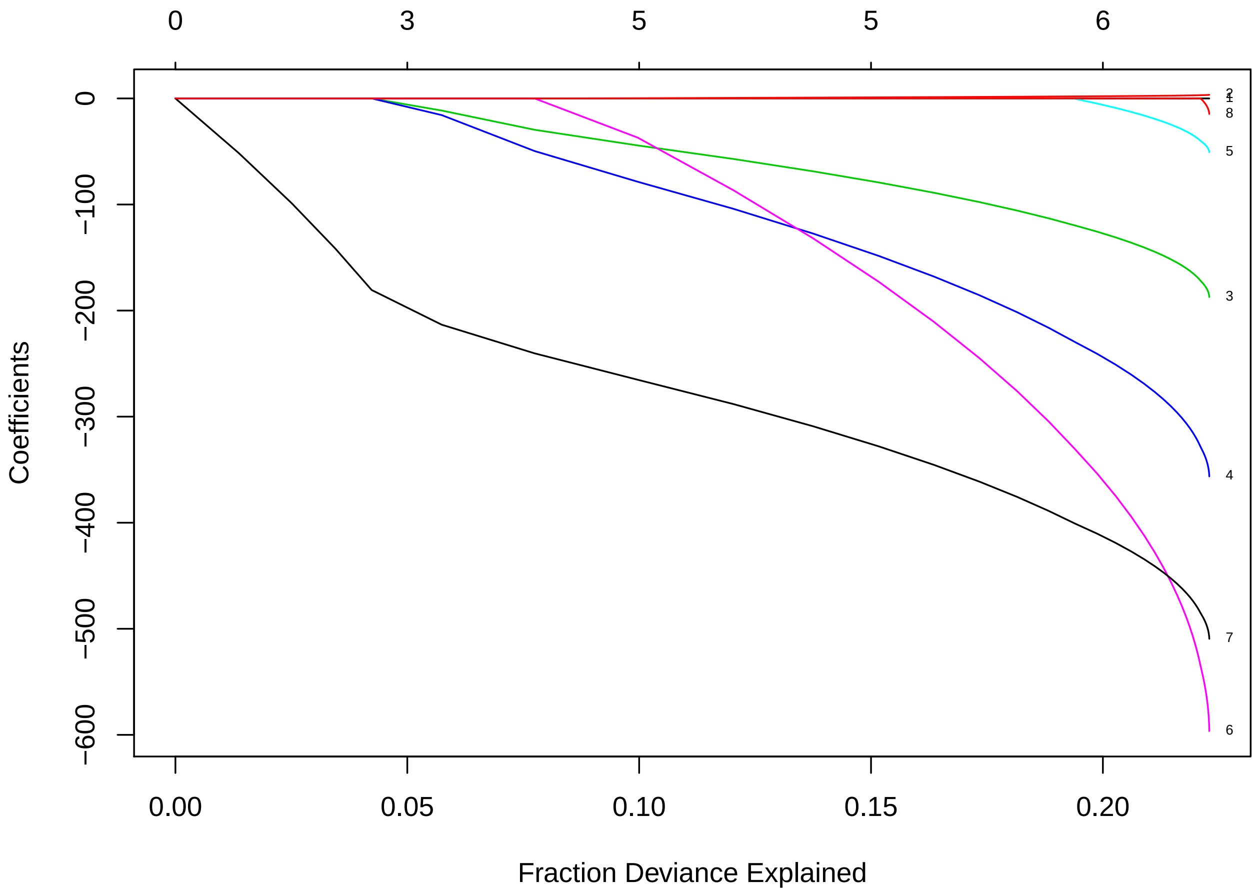
We can do lasso regression by using the similar codes as ridge regression:

```r
fit.lasso = glmnet(cbind(age,lwt,race,smoke,ptl,ht,ui,ftv),bwt,alpha=1)
plot(fit.lasso,xvar="lambda",label=TRUE)
```



From the plot of coefficients against Log Lambda, we can find that when the shrinkage parameter is small, which means the model is the same as OLS model, some of the coefficients are pretty large. And as the lambda increasing, the coefficients are concentrated. The plot has various choices. The deviance shows the percentage of deviance explained, (equivalent to r squared in case of regression)

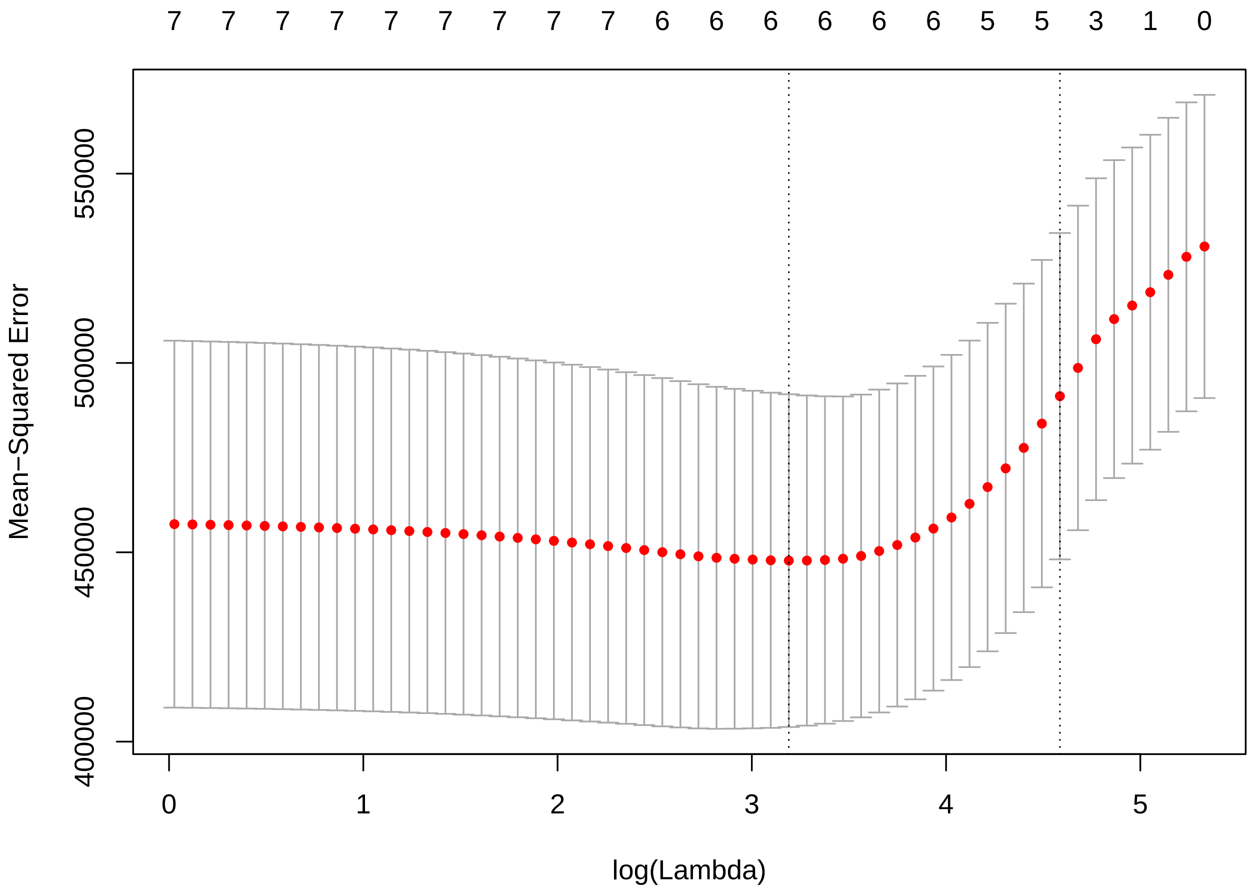
```r
plot(fit.lasso,xvar="dev",label=TRUE)
```



A lot of the r squared was explained for quite heavily shrunk coefficients. And towards the end, with a relatively small increase in r squared from between 0.1 and 0.2, coefficients grow very large. This may be an indication that the end of the path is overfitting

To find out when can we get the best model, we can use `cv.lasso` to conduct cross-validation to the data.

```r
cv.lasso = cv.glmnet(cbind(age,lwt,race,smoke,ptl,ht,ui,ftv),bwt,alpha=1)
plot(cv.lasso)
```



The fit.lasso will have the whole path of coefficients, which is 100 coefficient vectors dependent on each value, indexed by different values of lambda. We can find when we do lasso regression, it shows that two optimal selection of the predictors are the model with 6 predictors and 5 predictors.

To find out which is the best model, there's a coefficient function extractor that works on a cross validation object and pick the coefficient vector corresponding to the best model:

```
coef(cv.lasso)
```

```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##                         1
## (Intercept) 3073.1306531
## age            .
## lwt            0.5068003
## race         -57.0762981
## smoke       -104.0834359
## ptl            .
## ht           -86.5835626
## ui          -288.1511936
## ftv            .
```

The output above has 6 non-zero coefficients which shows that the function has chosen the first vertical line on the cross-validation plot. Which is a similar result as the OLS linear regression. The model is $bwt = 3070.09 + 0.22 * lwt - 44.33 * race - 78.54 * smoke - 36.93 * ht - 265.21 * ui$.

Stepwise

We can do stepwise selection towards the model by `stepAIC` in the library `MASS`.

```
fit.step = stepAIC(reg1, direction="both")
```

```
## Start:  AIC=2461.08
## bwt ~ age + lwt + race + smoke + ptl + ht + ui + ftv
##
##          Df Sum of Sq      RSS    AIC
## - age     1       331 77681278 2459.1
## - ftv     1     47283 77728230 2459.2
## - ptl     1    106439 77787385 2459.3
## <none>                77680946 2461.1
## - lwt     1   1762311 79443257 2463.3
## - ht      1   3728637 81409584 2467.9
## - race    1   4599967 82280914 2470.0
## - smoke   1   4796844 82477791 2470.4
## - ui      1   5732239 83413186 2472.5
##
## Step:  AIC=2459.09
## bwt ~ lwt + race + smoke + ptl + ht + ui + ftv
##
##          Df Sum of Sq      RSS    AIC
## - ftv     1     50343 77731620 2457.2
## - ptl     1    110047 77791325 2457.3
## <none>                77681278 2459.1
## + age     1       331 77680946 2461.1
## - lwt     1   1789656 79470934 2461.4
## - ht      1   3731126 81412404 2465.9
## - race    1   4707970 82389248 2468.2
## - smoke   1   4843734 82525012 2468.5
## - ui      1   5749594 83430871 2470.6
##
## Step:  AIC=2457.21
## bwt ~ lwt + race + smoke + ptl + ht + ui
##
##          Df Sum of Sq      RSS    AIC
## - ptl     1    108936 77840556 2455.5
## <none>                77731620 2457.2
## + ftv     1     50343 77681278 2459.1
## + age     1      3390 77728230 2459.2
## - lwt     1   1741198 79472818 2459.4
## - ht      1   3681167 81412788 2463.9
## - race    1   4660187 82391807 2466.2
## - smoke   1   4810582 82542203 2466.6
## - ui      1   5716074 83447695 2468.6
##
## Step:  AIC=2455.47
## bwt ~ lwt + race + smoke + ht + ui
##
##          Df Sum of Sq      RSS    AIC
## <none>                77840556 2455.5
## + ptl     1    108936 77731620 2457.2
```

```
## + ftv    1     49231 77791325 2457.3
## + age    1     10218 77830338 2457.4
## - lwt    1   1846738 79687294 2457.9
## - ht     1   3718531 81559088 2462.3
## - race   1   4727071 82567628 2464.6
## - smoke  1   5237430 83077987 2465.8
## - ui     1   6302771 84143327 2468.2
```

```
fit.step$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## bwt ~ age + lwt + race + smoke + ptl + ht + ui + ftv
##
## Final Model:
## bwt ~ lwt + race + smoke + ht + ui
##
##
##    Step Df    Deviance Resid. Df Resid. Dev      AIC
## 1                            180   77680946 2461.085
## 2 - age  1    331.2213       181   77681278 2459.085
## 3 - ftv  1  50342.6420       182   77731620 2457.208
## 4 - ptl  1 108935.8760       183   77840556 2455.473
```

```
fit.step$coefficients
```

```
## (Intercept)         lwt        race       smoke          ht          ui
## 3104.438081    3.433744 -187.848529 -366.134929 -595.820177 -523.419297
```

We can find the results gave a similar model as we obtained from the lasso regression, but with larger coefficients. This phenomenon proved the shrinkage property of the lasso method. The model is $bwt = 3104.44 + 3.43 * lwt - 187.85 * race - 366.13 * smoke - 595.82 * ht - 523.42 * ui$.

Question 3

Rating of performance of procedures below:

1 = Good, 2 = Fair, 3= Poor

	OLS	Ridge	LASSO	Elastic Net
1. Performance when p >> n	3	1	1	1
2. Performance under multicollinearity	3	1	3	1
3. Unbiased estimators	1	3	3	3
4. Model selection capability	1	3	1	1
5. Simplicity:	NA	NA	NA	NA
Computation	1	2	2	2
Inference	1	3	2	2
Interpretation	2	3	1	1

The properties for different methods:

OLS (Ordinary Least Square)

PROS

- OLS has the minimum MSE among unbiased linear estimator
- Explicit form
- Computation $O(np^2)$
- Confidence Interval, Significance of Coefficient can be calculated

CONS

- Multicollinearity leads to high variance of estimators
- Requires $n > p$
- Prediction error increases linearly as a function of p
- Hard to interpret when the number of predictors is large

Ridge Regression

PROS

- $p >> n$
- Explicit solution
- Multicollinearity
- Biased but smaller variance and smaller MSE

CONS

- Shrink coefficients to zero but can not produce a parsimonious model
- Not good for variable selection (model selection)
- A ridge solution can be hard to interpret because it is not sparse

Lasso Regression

PROS

- Allow $p >> n$
- Enforce sparcity in parameters
- Quadratic programming problem
- When $/lambda$ goes to 0, OLS solution
- Good for variable selection

CONS

- If a group of predictors are highly correlated among themselves, LASSO tends to pick only one of them and shrink others to zero

- Can not do grouped selection

Elastic Net

PROS

- It shares all of the advantages of Ridge and LASSO regression
- No limitation on the number of selected variable
- Enforce sparcity in parameters
- Encourge grouping effect in the presence of highly correlated predictors

CONS

- Naive elastic net suffers from double shrinkage