HW 8

Take the *fat* data, and use the *percentage of body fat* as the response and the *other* variables as potential predictors. Split the data into train/test. Run the following models:

1. **OLS**, there is a need for regularization to improve the fit.

```
##
## Call:
## lm(formula = siri ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3285  -2.9442  -0.1046   2.9091   9.6650
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -19.82090   17.98296  -1.102  0.27162
## age           0.06717    0.03409   1.970  0.05013 .
## weight       -0.09557    0.05561  -1.718  0.08718 .
## height       -0.04456    0.11226  -0.397  0.69183
## adipos       -0.04914    0.31640  -0.155  0.87673
## neck         -0.43798    0.24846  -1.763  0.07937 .
## chest        -0.08242    0.10944  -0.753  0.45219
## abdom         1.03016    0.09780  10.533  < 2e-16 ***
## hip          -0.20410    0.15574  -1.311  0.19144
## thigh         0.25359    0.15187   1.670  0.09644 .
## knee          0.02971    0.26088   0.114  0.90944
## ankle         0.15723    0.22680   0.693  0.48891
## biceps        0.18965    0.18024   1.052  0.29391
## forearm       0.46766    0.20384   2.294  0.02275 *
## wrist        -1.74316    0.56008  -3.112  0.00211 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.324 on 212 degrees of freedom
## Multiple R-squared:  0.7591, Adjusted R-squared:  0.7432
## F-statistic: 47.71 on 14 and 212 DF,  p-value: < 2.2e-16
```

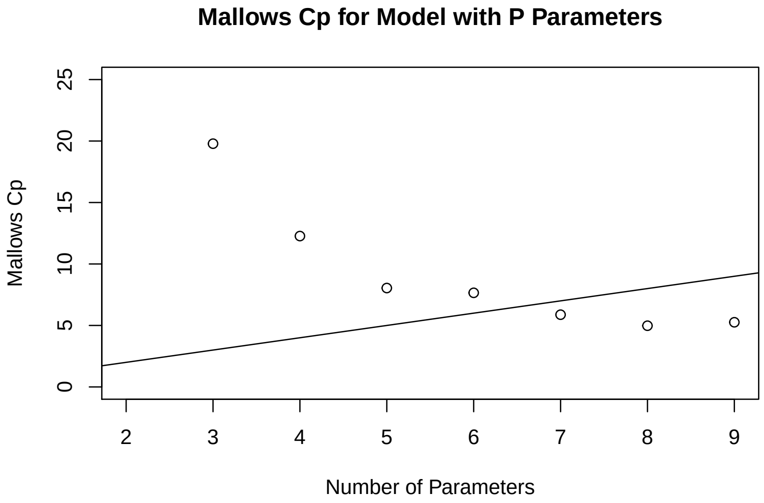
```r
#Prediction
ols_pred_train=predict(ols_fit, newdata = train)
ols_pred_test=predict(ols_fit, newdata = test)

#Root Mean Squared Error
rmse_train= sqrt((sum((train$siri-ols_pred_train)**2)/length(ols_pred_train)))
rmse_test= sqrt((sum((test$siri-ols_pred_test)**2)/length(ols_pred_test)))
rmse_train
```

```
## [1] 4.178651
```

```r
rmse_test
```

```
## [1] 4.395559
```

2.**Mallow's Cp** - The model with the minimum Mallow's Cp is the model with a total of 8 parameters. This means that the final model will have 7 predictors, and such ones being- age, weight, neck, abdom, thigh, forearm, wrist.



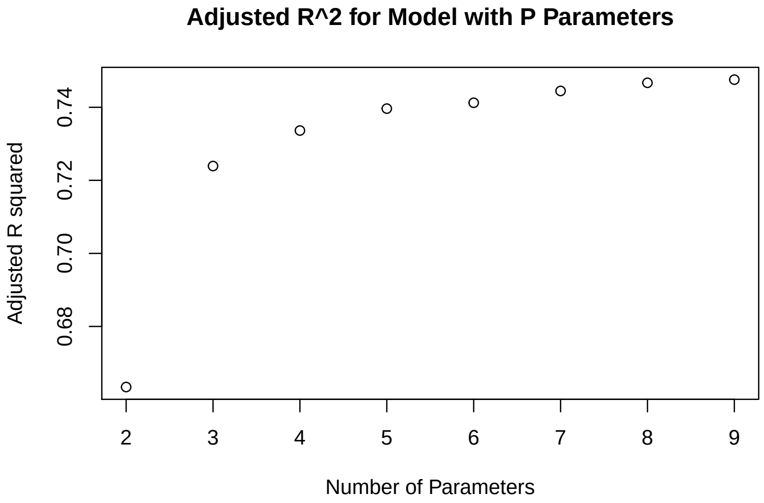
```
## Subset selection object
## Call: regsubsets.formula(siri ~ ., data = train)
## 14 Variables  (and intercept)
##         Forced in Forced out
## age         FALSE      FALSE
## weight      FALSE      FALSE
## height      FALSE      FALSE
## adipos      FALSE      FALSE
## neck        FALSE      FALSE
## chest       FALSE      FALSE
## abdom       FALSE      FALSE
## hip         FALSE      FALSE
## thigh       FALSE      FALSE
## knee        FALSE      FALSE
## ankle       FALSE      FALSE
## biceps      FALSE      FALSE
## forearm     FALSE      FALSE
## wrist       FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          age weight height adipos neck chest abdom hip thigh knee ankle biceps forearm wrist
## 1  ( 1 ) " " " "    " "    " "    " "  " "   "*"   " " " "   " "  " "   " "    " "     " "
## 2  ( 1 ) " " "*"    " "    " "    " "  " "   "*"   " " " "   " "  " "   " "    " "     " "
## 3  ( 1 ) " " "*"    " "    " "    " "  " "   "*"   " " " "   " "  " "   " "    " "     "*"
## 4  ( 1 ) " " "*"    " "    " "    " "  " "   "*"   " " " "   " "  " "   " "    "*"     "*"
## 5  ( 1 ) " " "*"    " "    " "    "*"  " "   "*"   " " " "   " "  " "   " "    "*"     "*"
## 6  ( 1 ) "*" "*"    " "    " "    " "  " "   "*"   " " "*"   " "  " "   " "    "*"     "*"
## 7  ( 1 ) "*" "*"    " "    " "    "*"  " "   "*"   " " "*"   " "  " "   " "    "*"     "*"
## 8  ( 1 ) "*" "*"    " "    " "    "*"  " "   "*"   "*" "*"   " "  " "   " "    "*"     "*"

##
```

```
## Call:
## lm(formula = siri ~ age + weight + neck + abdom + thigh + forearm +
##     wrist, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.172  -3.125  -0.264   3.089   9.315
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -33.79207    9.43053  -3.583 0.000418 ***
## age           0.07180    0.03200   2.243 0.025871 *
## weight       -0.12792    0.03548  -3.606 0.000385 ***
## neck         -0.39624    0.23121  -1.714 0.087978 .
## abdom         0.94869    0.07430  12.768  < 2e-16 ***
## thigh         0.24222    0.11828   2.048 0.041776 *
## forearm       0.53976    0.18906   2.855 0.004718 **
## wrist        -1.63732    0.53368  -3.068 0.002427 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.294 on 219 degrees of freedom
## Multiple R-squared:  0.7546, Adjusted R-squared:  0.7467
## F-statistic: 96.18 on 7 and 219 DF,  p-value: < 2.2e-16
```

```
## [1] 4.217687
```

```
## [1] 4.342456
```

3. **AdjustedR2** - The model with the highest Adjusted R^2 is the model with a total of 9 parameters. This means that the final model will have 8 predictors, and such ones being- age, weight, neck, abdom, hip, thigh, forearm, wrist.



```
## [1] 8
```

```
##
## Call:
## lm(formula = siri ~ age + weight + neck + abdom + hip + thigh +
##     forearm + wrist, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2181  -2.8832  -0.1985   2.8211   9.8197
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -23.71280   12.11193  -1.958  0.05153 .
## age           0.07011    0.03197   2.193  0.02938 *
## weight       -0.09992    0.04126  -2.422  0.01625 *
## neck         -0.46280    0.23623  -1.959  0.05138 .
## abdom         0.97661    0.07712  12.664  < 2e-16 ***
## hip          -0.19051    0.14403  -1.323  0.18732
## thigh         0.32262    0.13281   2.429  0.01594 *
## forearm       0.50778    0.19028   2.669  0.00819 **
## wrist        -1.63149    0.53279  -3.062  0.00247 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.287 on 218 degrees of freedom
## Multiple R-squared:  0.7565, Adjusted R-squared:  0.7476
## F-statistic: 84.66 on 8 and 218 DF,  p-value: < 2.2e-16
```

```
## [1] 4.200863
```

```
## [1] 4.327248
```

4. **Ridge regression**, after standardizing the predictors, the first step was to check out the diag(X'X) for possible values of lambda. The plausible range values were from 0 to 1.5. The best lambda is 1.09 (yields smallest test error). The coefficients that go with this lambda are shown in the output below.

```
## [1] "Lambda values"
```

```
##         siri          age       weight       height       adipos         neck        chest          abd
## 1.130649e-05 9.882871e-03 1.429308e-01 9.044444e-03 7.126626e-02 1.952698e-02 4.551466e-02 6.082933e-
```

```
##         hip       thigh        knee       ankle      biceps     forearm       wrist
## 0.066576323 0.034050276 0.021171259 0.007901782 0.015863259 0.009082273 0.014626280
```

```
## [1] "Coefficents"
```

```
##                     age      weight      height      adipos        neck       chest       abdom
## 19.31622176  0.96031824 -2.36140066 -0.18161031  0.02750306 -1.07830316 -0.55241476 10.37269974
```

```
##         hip       thigh        knee       ankle      biceps     forearm       wrist
## -1.33678364  1.29763598  0.03757994  0.21631607  0.50169488  0.91796568 -1.67469614
```

```
## [1] 4.183839
```

```
## [1] 4.282531
```

Models Performance

OLS, this model has a total of 4 significant predictors (age, abdom, forearm, wrist), and a residual standard error of 4.324. The R^2 is 0.7591, which is quite high indeed. This model is not too bad in terms of performance, but does have too many insignificant predicotrs & as a result the analysis can be improved (through regularization). Train and test errors are presented in the table.

Mallow's Cp, this model has 7 predictors and a total of 6 significant predictors (age, weight, abdom, thigh, forearm, wrist), and a residual standard error of 4.294 (smaller than OLS). The R^2 is 0.7546, which is quite high indeed. This model is quite good in terms of performance- it has less predictors than OLS & the oveall performance is similar. Train and test errors are presented in the table.

Adjusted R^2, this model has 8 predictors and a total of 6 significant predictors (age, weight, abdom, thigh, forearm, wrist), and a residual standard error of 4.28 (smaller than OLS & MCp). The R^2 is 0.7566, which is quite high indeed. This model is quite good in terms of performance, but it has insignifcnat predictors compared to Mallow's Cp where the oveall performance is similar. Train and test errors are presented in the table.

Ridge,this model best lambda is 1.09. The coefficients that results from the best lambda are

age	weight	height	adipos	neck	chest	abdom
0.96031824	-2.36140066	-0.18161031	0.02750306	-1.07830316	-0.55241476	10.37269974

hip	thigh	knee	ankle	biceps	forearm	wrist
-1.33678364	1.29763598	0.03757994	0.21631607	0.50169488	0.91796568	-1.67469614

In all models the training error is always smaller than the testing error (in-sample vs out-of-sample error). This is typical, and reflective of the bias/variance trade of. Also, training error tends to be smaller given that our models are models is trained on that data; it will be biased in a certain way to give nice results. Test data error, gives a better idea of the performance of the model given that the model has not seen that data in its modeling phase. The predictor *abdom* seems to be the most significant in most models. Finally, the **best model*** is **Ridge Regression** with the smallest test root mean squared error **4.28**. Ridge tends to be biased towards smaller coefficents.

Model	**Train RMSE**	**Test RMSE**
OLS	4.178651	4.395559
Mallow's Cp	4.217687	4.342456
Adjusted R^2	4.200863	4.327248
Ridge	4.183839	4.282531