BODY FAT PERCENTAGE

ANALYSIS

SEPTEMBER 7, 2021

KO TSZ NGA (VALERIE)
valerie.ktn@gmail.com

Introduction

This project aims at comparing the performance of different regression models in terms of predicting body fat percentage. The models include linear regression, ridge regression, LASSO, principal component regression and partial least squares.

Background of the dataset

The data set "fat" is gathered from the "faraway" library of R. Body circumference measurements (eg. Age, weight, height etc) are recorded for 252 men. Each man's percentage of body fat was accurately estimated by an underwater weighing technique. The dataset consists of 252 rows and 18 columns.

```
> dim(fat)
[1] 252  18
```

The first column (ie. **brozek**) is **the response variables** while the **remaining 17 variables** are the potential **predictors**.

Data Preparation

The dataset is split into training set and testing sets. This time, **10%** of the randomly selected data will be used as **test set**, the remaining **90%** will be our **training set (ie. 'fat1train')**.

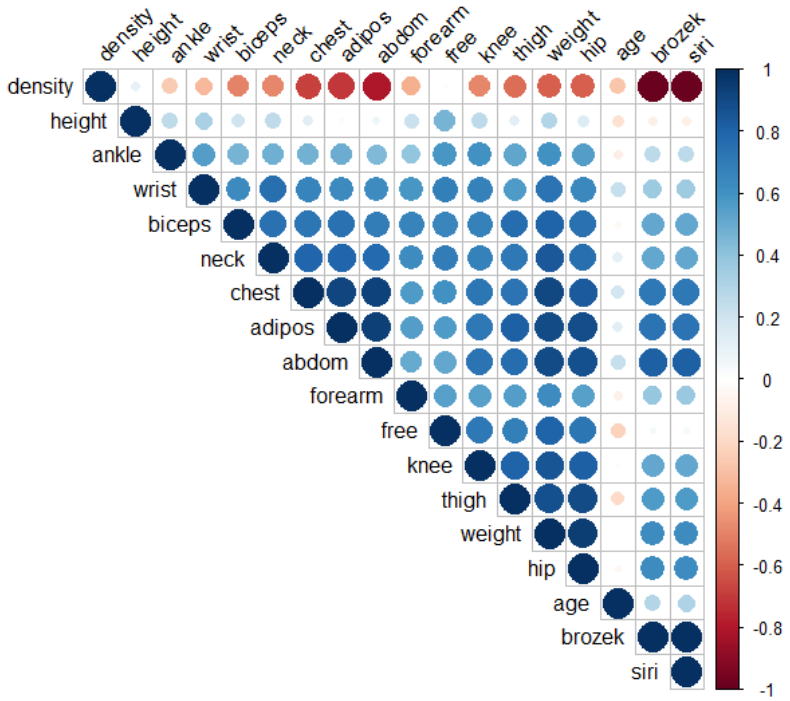
```
> dim(fat1test)
[1] 25 18
> dim(fat1train)
[1] 227  18
```

There are **25 rows in testing set** and **227 rows in training set.**

Exploratory data analysis

Exploratory data analysis is carried out for the **'fat1train'** data set.

Correlation between the variables is calculated and shown by the below correlogram.



As we can see, **'density' shows negative correlation with the other variables**. Specifically, it is strongly uncorrelated with 'brozek' and 'siri'. Besides, **'age' and 'height' shows relatively small positive correlations with the other variables**. The variables that are **strongly positively correlated with our response variable 'brozek' are: 'abdom', 'adipos', 'chest'**, and **'density' is strongly negatively correlated with our response variable 'brozek'**. This information is crucial for us to determine which predicting variables are potentially helpful in prediction in our model.

Methodology

Model 1: Linear regression with all predictors

```
> model1<-lm(brozek~.,data=fat1train)
> summary(model1)

Call:
lm(formula = brozek ~ ., data = fat1train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.11741 -0.04740  0.00287  0.04463  1.44638

Coefficients:
```

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	11.7936888	4.3975153	2.682	(Intercept)	0.00791 **
siri	0.8873731	0.0118486	74.893	siri	< 2e-16 ***
density	-9.5525607	3.9465795	-2.420	density	0.01636 *
age	-0.0009373	0.0014353	-0.653	age	0.51448
weight	0.0091588	0.0039170	2.338	weight	0.02032 *
height	-0.0005485	0.0046918	-0.117	height	0.90705
adipos	-0.0170554	0.0134653	-1.267	adipos	0.20670
free	-0.0109570	0.0048681	-2.251	free	0.02544 *
neck	-0.0013519	0.0105857	-0.128	neck	0.89850
chest	0.0014251	0.0047338	0.301	chest	0.76367
abdom	0.0032863	0.0049634	0.662	abdom	0.50863
hip	-0.0049879	0.0064089	-0.778	hip	0.43728
thigh	0.0162068	0.0063888	2.537	thigh	0.01192 *
knee	-0.0260943	0.0106138	-2.459	knee	0.01476 *
ankle	0.0047631	0.0096266	0.495	ankle	0.62127
biceps	-0.0143395	0.0075657	-1.895	biceps	0.05943 .
forearm	0.0154830	0.0086333	1.793	forearm	0.07436 .
wrist	0.0406682	0.0242018	1.680	wrist	0.09438 .

```
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1789 on 209 degrees of freedom
Multiple R-squared:  0.9995,    Adjusted R-squared:  0.9995
F-statistic: 2.539e+04 on 17 and 209 DF,  p-value: < 2.2e-16
```

Variables like **‘age’, ‘height’, ‘adipos’, ‘neck’, ‘chest’, ‘abdom’, ‘hip’, ‘ankle’ have p-value larger than 0.1**, which potentially make them **not statistically significant** enough.

Model 2: Linear regression with the best subset of k = 5 predictors variables

```
> summary(model2)

Call:
lm(formula = as.formula(mod5form), data = fat1train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.08059 -0.04541  0.00135  0.04106  1.59420

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.228712   4.223503   2.659  0.00842 **
siri         0.904565   0.008863 102.058  < 2e-16 ***
density     -9.240597   3.863288  -2.392  0.01760 *
thigh        0.009954   0.003928   2.534  0.01196 *
knee        -0.024832   0.009053  -2.743  0.00659 **
wrist        0.027927   0.017103   1.633  0.10392
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1795 on 221 degrees of freedom
Multiple R-squared:  0.9995,    Adjusted R-squared:  0.9995
F-statistic: 8.567e+04 on 5 and 221 DF,  p-value: < 2.2e-16
```

This time, only variable **‘wrist’** has **p-value larger than 0.1**.

Model 3: Linear regression with variables (stepwise) selected using AIC

```
> summary(model3)

Call:
lm(formula = brozek ~ siri + density + weight + adipos + free +
    thigh + knee + biceps + forearm + wrist, data = fat1train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.11212 -0.05101  0.00345  0.04885  1.47633

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.351965   4.233987   2.917  0.00390 **
siri          0.888432   0.011253  78.953  < 2e-16 ***
density     -10.119630   3.834329  -2.639  0.00892 **
weight        0.008636   0.003617   2.387  0.01783 *
adipos       -0.013668   0.008841  -1.546  0.12356
free         -0.009940   0.004564  -2.178  0.03052 *
thigh         0.014718   0.005075   2.900  0.00412 **
knee         -0.027197   0.009885  -2.751  0.00644 **
biceps       -0.015021   0.007253  -2.071  0.03953 *
forearm       0.016403   0.008086   2.029  0.04372 *
wrist         0.037131   0.020493   1.812  0.07139 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1766 on 216 degrees of freedom
Multiple R-squared:  0.9995,    Adjusted R-squared:  0.9995
F-statistic: 4.427e+04 on 10 and 216 DF,  p-value: < 2.2e-16
```

As we can see, **10** predicting variables are chosen by the algorithm. And only **‘adipos’ has p-value larger than 0.1.**

Model 4: Ridge regression

```
> c(intercept, ridge.coeffs)
                siri        density            age         weight         height         adipos
12.0033057424 0.8865902946 -9.7491306232 -0.0009322550 0.0092865470 -0.0005351683 -0.0172509330
         free          neck         chest         abdom           hip         thigh          knee
-0.0111509976 -0.0013223847 0.0014895826 0.0033093516 -0.0049629981 0.0162510851 -0.0260475063
        ankle        biceps       forearm         wrist
 0.0047756795 -0.0143179503 0.0155553954 0.0407752930
```

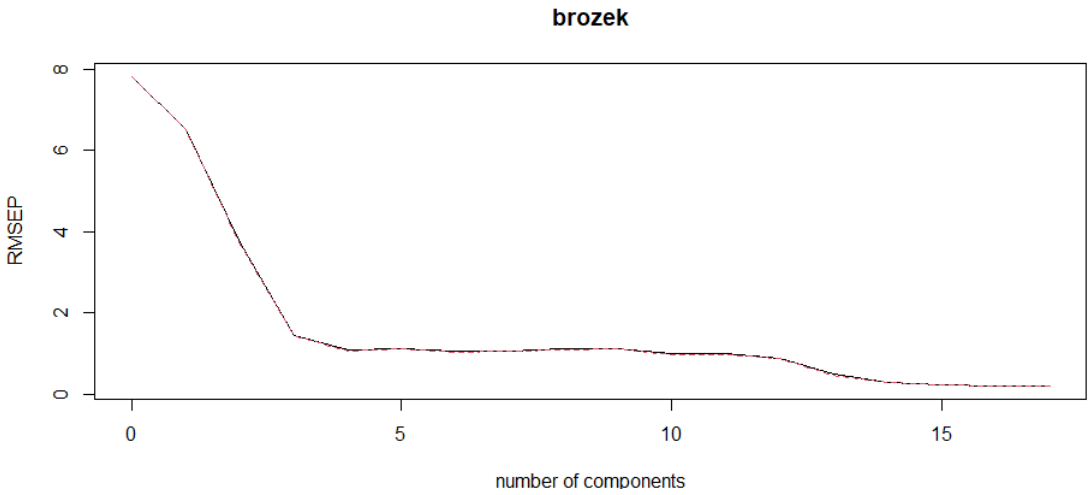
The above figure displays **the value of intercept and coefficients** of the ridge regression models.

Model 5: LASSO

```
> c(LASSOintercept, coef.lars1$coef)
                siri        density            age         weight         height         adipos
11.2964552540 0.9042772533 -9.3779257865 -0.0004611835 0.0000000000 0.0000000000 0.0000000000
         free          neck         chest         abdom           hip         thigh          knee
 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0072941819 -0.0137123158
        ankle        biceps       forearm         wrist
 0.0000000000 -0.0077138784 0.0083285937 0.0192986407
```

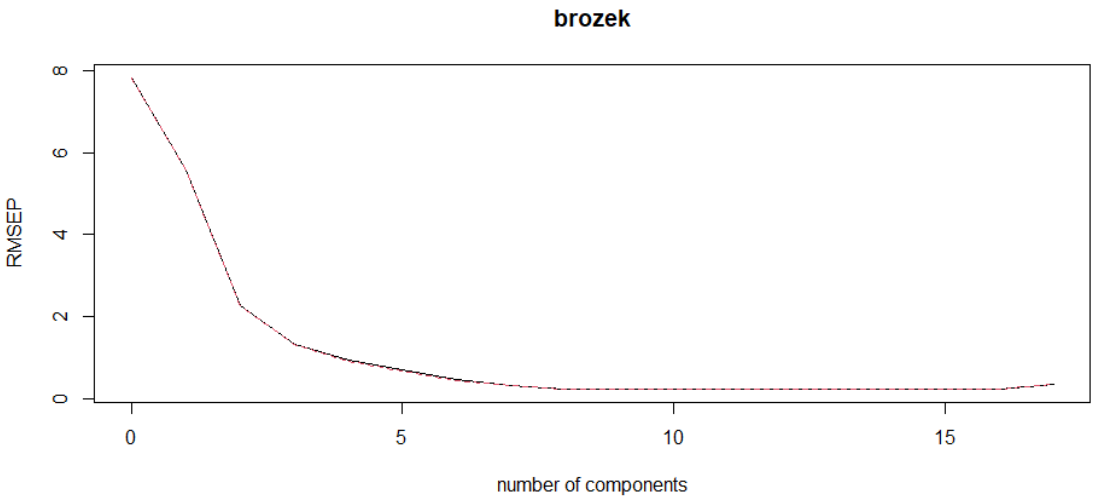
The above figure displays **the value of intercept and coefficients** of the LASSO regression models.

Model 6: Principal component regression



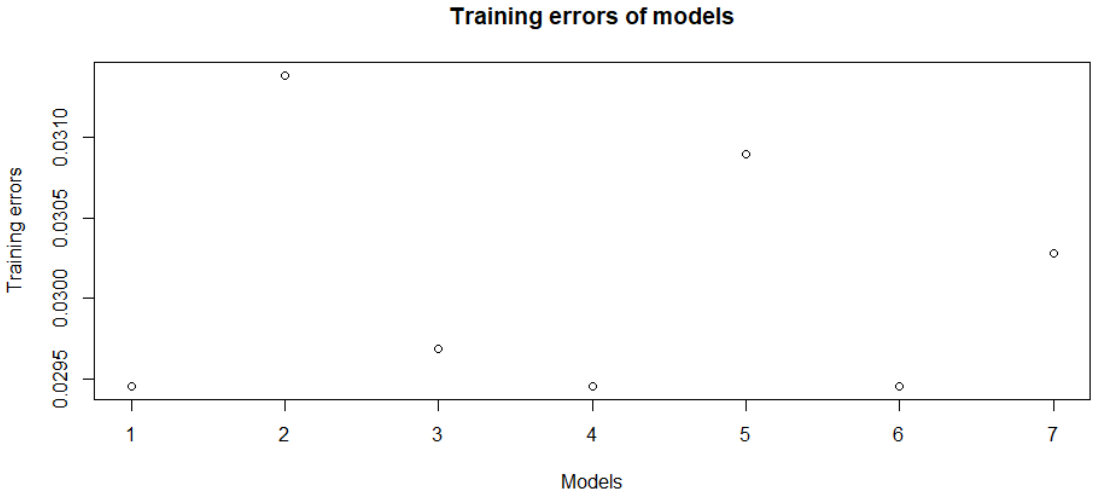
This model shows that the **optimal number of components is 17**.

Model 7: Partial least squares



This model shows that the **optimal number of components is 16.**

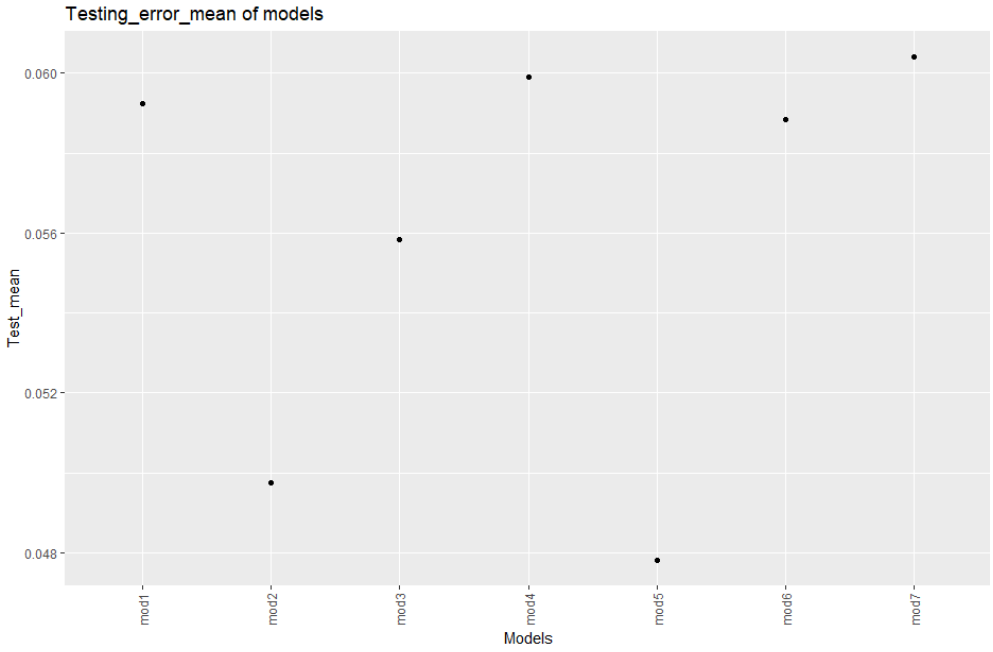
Results and Conclusions

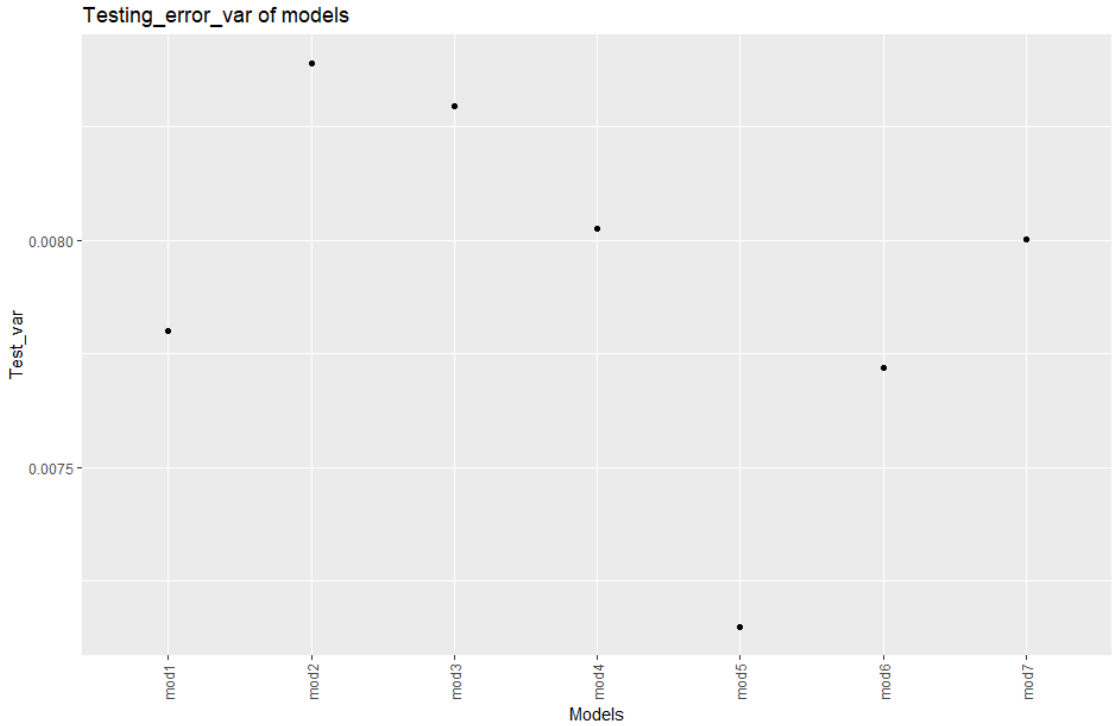


The above table shows the training error of the 7 models. As we can see, **linear regression with all predictors (Model1), Ridge regression (Model 4) and principal component regression (Model6) have the lowest training errors.**

The above table shows the testing error of the 7 models. As we can see **LASSO (Model 5) has the lowest training errors**.

Monte Carlo Cross-Validation





From the Monte Carlo Cross Validation result, **LASSO (Model 5)** stills give the **lowest average testing error** as of our result in part d. Also, we can see that it gives the **lowest average testing error variance**. **That means Model 5 is probably the best models among all.**