Investigating the Underlying Relationships in Body Dimensions

Dan Huang

Virginia Commonwealth University

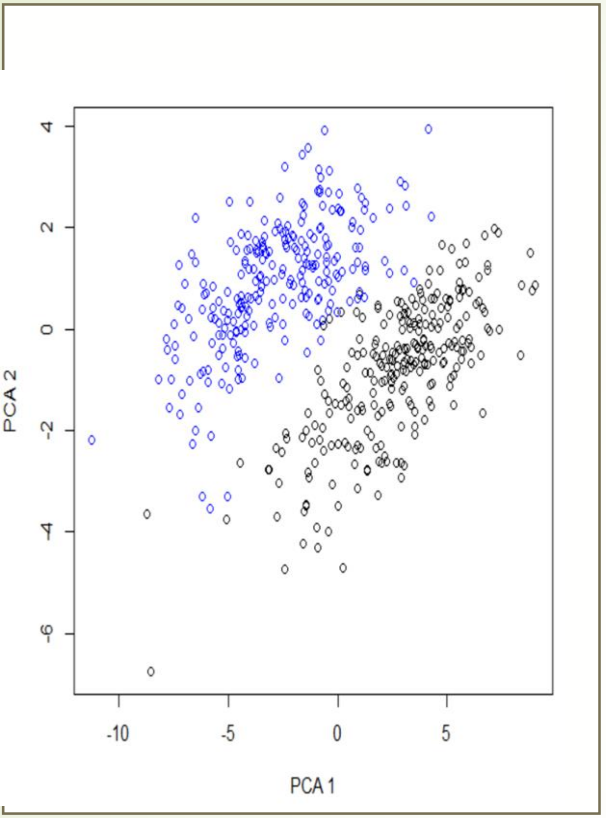
04/25/2018

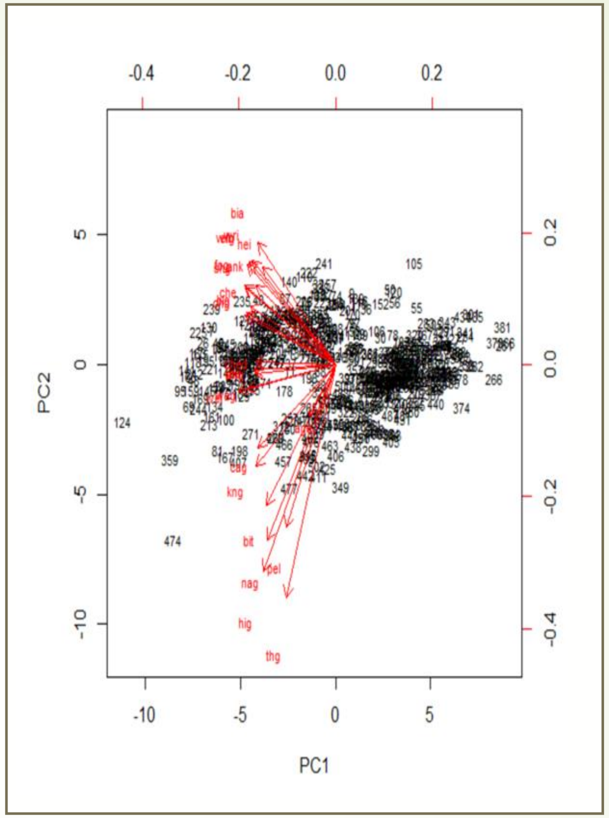
Data Introduction

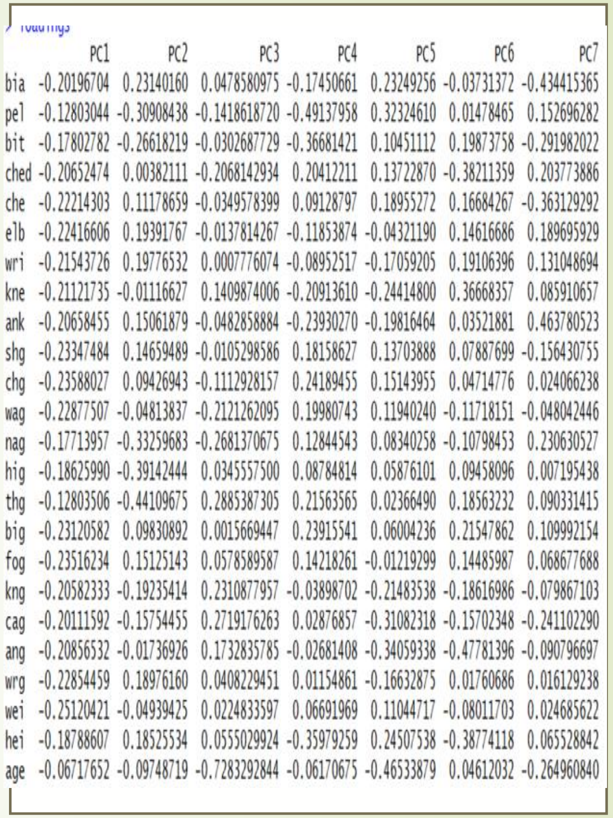
- Size: 507 observations, 25 variables
- Variables: 9 skeletal variables, 12 girth variables
- Other measurements: age, weight, height, gender
- No missing values
- The goal of this report is to investigating the relationship in the body build dimensions for commercial business or art of designs.

Main Methods

- Principle Component Analysis
- Factor Analysis
- Multiple Linear Regression
- Logistic Regression
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- K-Nearest Neighbors
- Classification Trees
- Bagging
- Boosting
- Random Forest





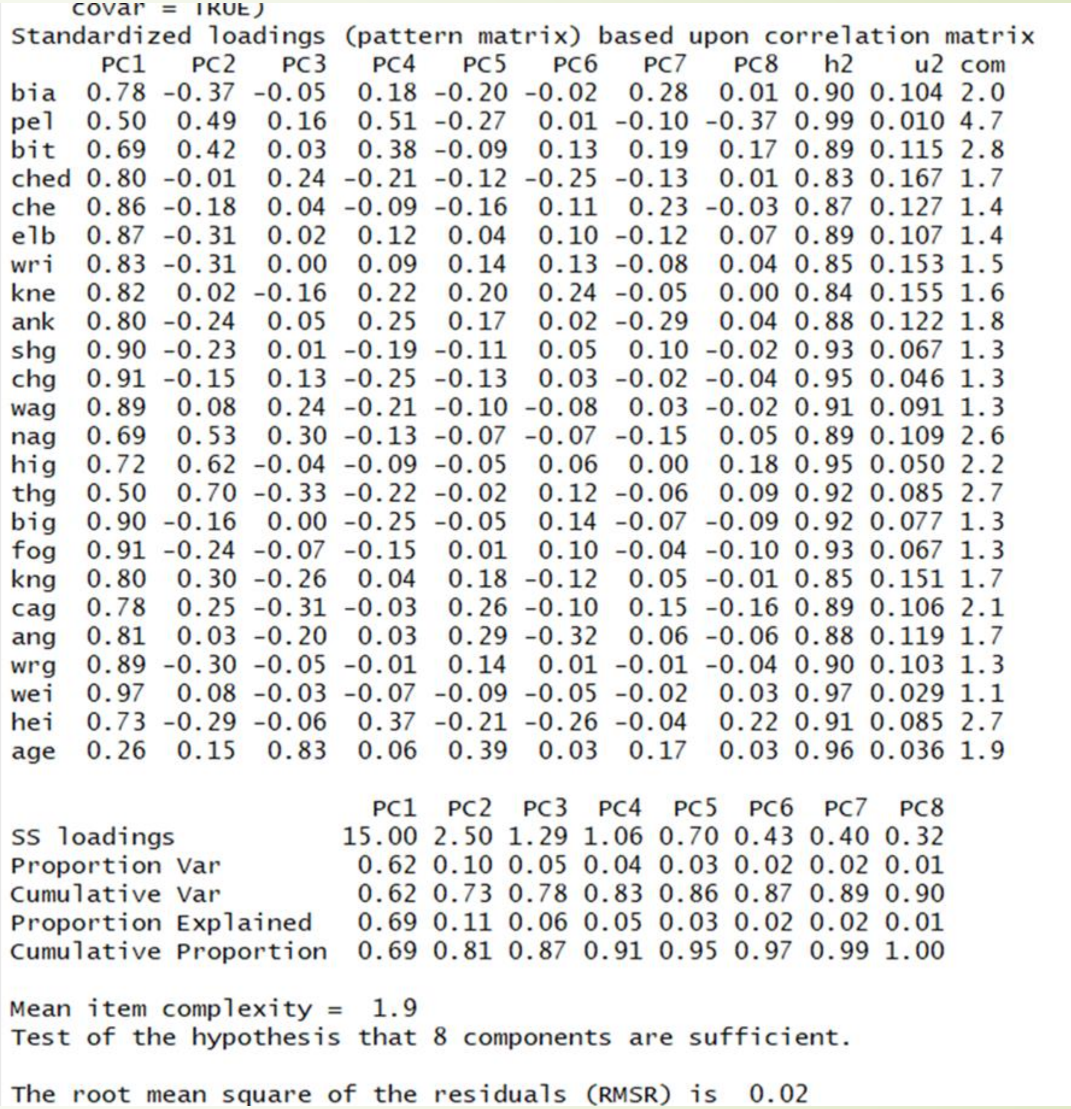

```
> loadings
            PC1         PC2          PC3         PC4         PC5         PC6          PC7
bia  -0.20196704  0.23140160  0.0478580975 -0.17450661  0.23249256 -0.03731372 -0.434415365
pel  -0.12803044 -0.30908438 -0.1418618720 -0.49137958  0.32324610  0.01478465  0.152696282
bit  -0.17802782 -0.26618219 -0.0302687729 -0.36681421  0.10451112  0.19873758 -0.291982022
ched -0.20652474  0.00382111 -0.2068142934  0.20412211  0.13722870 -0.38211359  0.203773886
che  -0.22214303  0.11178659 -0.0349578399  0.09128797  0.18955272  0.16684267 -0.363129292
elb  -0.22416606  0.19391767 -0.0137814267 -0.11853874 -0.04321190  0.14616686  0.189695929
wri  -0.21543726  0.19776532  0.0007776074 -0.08952517 -0.17059205  0.19106396  0.131048694
kne  -0.21121735 -0.01116627  0.1409874006 -0.20913610 -0.24414800  0.36668357  0.085910657
ank  -0.20658455  0.15061879 -0.0482858884 -0.23930270 -0.19816464  0.03521881  0.463780523
shg  -0.23347484  0.14659489 -0.0105298586  0.18158627  0.13703888  0.07887699 -0.156430755
chg  -0.23588027  0.09426943 -0.1112928157  0.24189455  0.15143955  0.04714776  0.024066238
wag  -0.22877507 -0.04813837 -0.2121262095  0.19980743  0.11940240 -0.11718151 -0.048042446
nag  -0.17713957 -0.33259683 -0.2681370675  0.12844543  0.08340258 -0.10798453  0.230630527
hig  -0.18625990 -0.39142444  0.0345557500  0.08784814  0.05876101  0.09458096  0.007195438
thg  -0.12803506 -0.44109675  0.2885387305  0.21563565  0.02366490  0.18563232  0.090331415
big  -0.23120582  0.09830892  0.0015669447  0.23915541  0.06004236  0.21547862  0.109992154
fog  -0.23516234  0.15125143  0.0578589587  0.14218261 -0.01219299  0.14485987  0.068677688
kng  -0.20582333 -0.19235414  0.2310877957 -0.03898702 -0.21483538 -0.18616986 -0.079867103
cag  -0.20111592 -0.15754455  0.2719176263  0.02876857 -0.31082318 -0.15702348 -0.241102290
ang  -0.20856532 -0.01736926  0.1732835785 -0.02681408 -0.34059338 -0.47781396 -0.090796697
wrg  -0.22854459  0.18976160  0.0408229451  0.01154861 -0.16632875  0.01760686  0.016129238
wei  -0.25120421 -0.04939425  0.0224833597  0.06691969  0.11044717 -0.08011703  0.024685622
hei  -0.18788607  0.18525534  0.0555029924 -0.35979259  0.24507538 -0.38774118  0.065528842
age  -0.06717652 -0.09748719 -0.7283292844 -0.06170675 -0.46533879  0.04612032 -0.264960840
```

Principle Component Analysis

The first 8 principle components can explain 90.481% of the total variances which is good.

Factor Analysis

- First factor could be considered girth build factor; the second factor could be considered thigh girth factor; the third factor could be considered as age factor. Compared with the communities of the two factors analysis, the communities of the eight factors analysis are much better. Moreover, the p-value of chi square test is less than 0.86, which implies that null hypothesis can not be rejected and conclude that the eight-factor model is adequate.



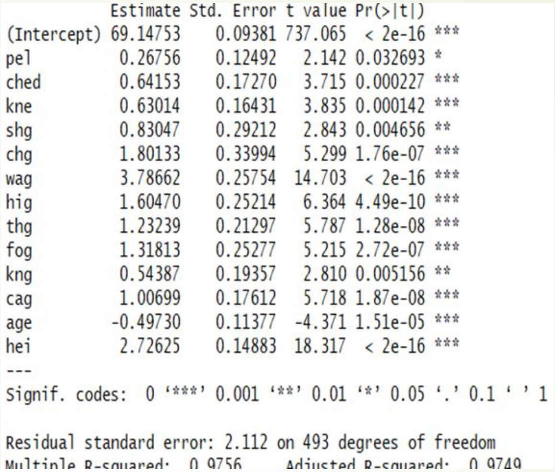
```
    covar = TRUE)
Standardized loadings (pattern matrix) based upon correlation matrix
      PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   h2    u2 com
bia  0.78 -0.37 -0.05  0.18 -0.20 -0.02  0.28  0.01 0.90 0.104 2.0
pel  0.50  0.49  0.16  0.51 -0.27  0.01 -0.10 -0.37 0.99 0.010 4.7
bit  0.69  0.42  0.03  0.38 -0.09  0.13  0.19  0.17 0.89 0.115 2.8
ched 0.80 -0.01  0.24 -0.21 -0.12 -0.25 -0.13  0.01 0.83 0.167 1.7
che  0.86 -0.18  0.04 -0.09 -0.16  0.11  0.23 -0.03 0.87 0.127 1.4
elb  0.87 -0.31  0.02  0.12  0.04  0.10 -0.12  0.07 0.89 0.107 1.4
wri  0.83 -0.31  0.00  0.09  0.14  0.13 -0.08  0.04 0.85 0.153 1.5
kne  0.82  0.02 -0.16  0.22  0.20  0.24 -0.05  0.00 0.84 0.155 1.6
ank  0.80 -0.24  0.05  0.25  0.17  0.02 -0.29  0.04 0.88 0.122 1.8
shg  0.90 -0.23  0.01 -0.19 -0.11  0.05  0.10 -0.02 0.93 0.067 1.3
chg  0.91 -0.15  0.13 -0.25 -0.13  0.03 -0.02 -0.04 0.95 0.046 1.3
wag  0.89  0.08  0.24 -0.21 -0.10 -0.08  0.03 -0.02 0.91 0.091 1.3
nag  0.69  0.53  0.30 -0.13 -0.07 -0.07 -0.15  0.05 0.89 0.109 2.6
hig  0.72  0.62 -0.04 -0.09 -0.05  0.06  0.00  0.18 0.95 0.050 2.2
thg  0.50  0.70 -0.33 -0.22 -0.02  0.12 -0.06  0.09 0.92 0.085 2.7
big  0.90 -0.16  0.00 -0.25 -0.05  0.14 -0.07 -0.09 0.92 0.077 1.3
fog  0.91 -0.24 -0.07 -0.15  0.01  0.10 -0.04 -0.10 0.93 0.067 1.3
kng  0.80  0.30 -0.26  0.04  0.18 -0.12  0.05 -0.01 0.85 0.151 1.7
cag  0.78  0.25 -0.31 -0.03  0.26 -0.10  0.15 -0.16 0.89 0.106 2.1
ang  0.81  0.03 -0.20  0.03  0.29 -0.32  0.06 -0.06 0.88 0.119 1.7
wrg  0.89 -0.30 -0.05 -0.01  0.14  0.01 -0.01 -0.04 0.90 0.103 1.3
wei  0.97  0.08 -0.03 -0.07 -0.09 -0.05 -0.02  0.03 0.97 0.029 1.1
hei  0.73 -0.29 -0.06  0.37 -0.21 -0.26 -0.04  0.22 0.91 0.085 2.7
age  0.26  0.15  0.83  0.06  0.39  0.03  0.17  0.03 0.96 0.036 1.9

                        PC1  PC2  PC3  PC4  PC5  PC6  PC7  PC8
SS loadings           15.00 2.50 1.29 1.06 0.70 0.43 0.40 0.32
Proportion Var         0.62 0.10 0.05 0.04 0.03 0.02 0.02 0.01
Cumulative Var         0.62 0.73 0.78 0.83 0.86 0.87 0.89 0.90
Proportion Explained   0.69 0.11 0.06 0.05 0.03 0.02 0.02 0.01
Cumulative Proportion  0.69 0.81 0.87 0.91 0.95 0.97 0.99 1.00

Mean item complexity =  1.9
Test of the hypothesis that 8 components are sufficient.

The root mean square of the residuals (RMSR) is  0.02
```

Multiple Regression Models

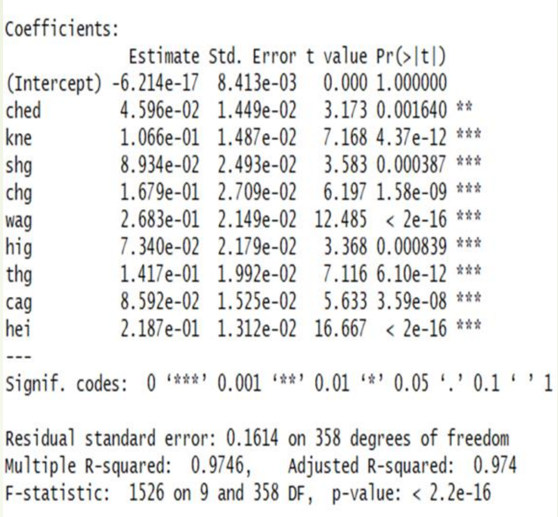
a)Weight= 69.148-1.059bia+0.876pel+1.816bit+4.546ched+4.1686che+1.444wri+2.23kne-0.895age+1.794hei

b)Weight= 69.148+0.73212shg+2.07chg+3.966wag+1.762hig+1.27thg+1.511fog+0.751kng+1.126cag-0.356age+2.997hei

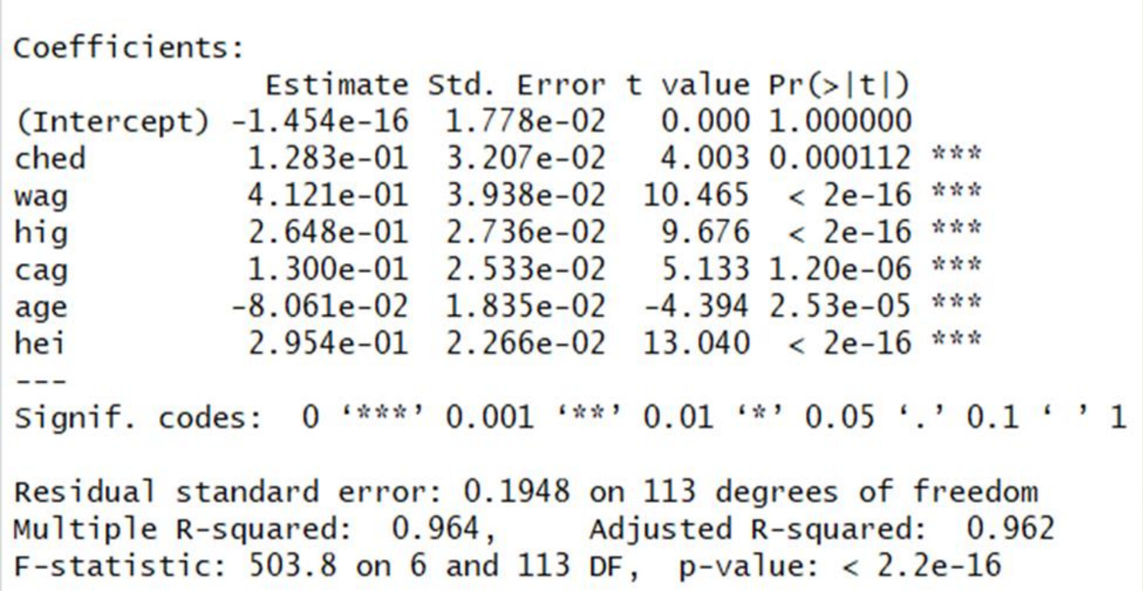
In the young group, the people from age 18-35 don't change too much on the weight with other skeletal and girth variables. It shows that increasing the age from 36 will have negative influence on weight.



```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 69.14753    0.09381 737.065  < 2e-16 ***
pel          0.26756    0.12492   2.142 0.032693 *
ched         0.64153    0.17270   3.715 0.000227 ***
kne          0.63014    0.16431   3.835 0.000142 ***
shg          0.83047    0.29212   2.843 0.004656 **
chg          1.80133    0.33994   5.299 1.76e-07 ***
wag          3.78662    0.25754  14.703  < 2e-16 ***
hig          1.60470    0.25214   6.364 4.49e-10 ***
thg          1.23239    0.21297   5.787 1.28e-08 ***
fog          1.31813    0.25277   5.215 2.72e-07 ***
kng          0.54387    0.19357   2.810 0.005156 **
cag          1.00699    0.17612   5.718 1.87e-08 ***
age         -0.49730    0.11377  -4.371 1.51e-05 ***
hei          2.72625    0.14883  18.317  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.112 on 493 degrees of freedom
Multiple R-squared:  0.9756     Adjusted R-squared:  0.9749
```



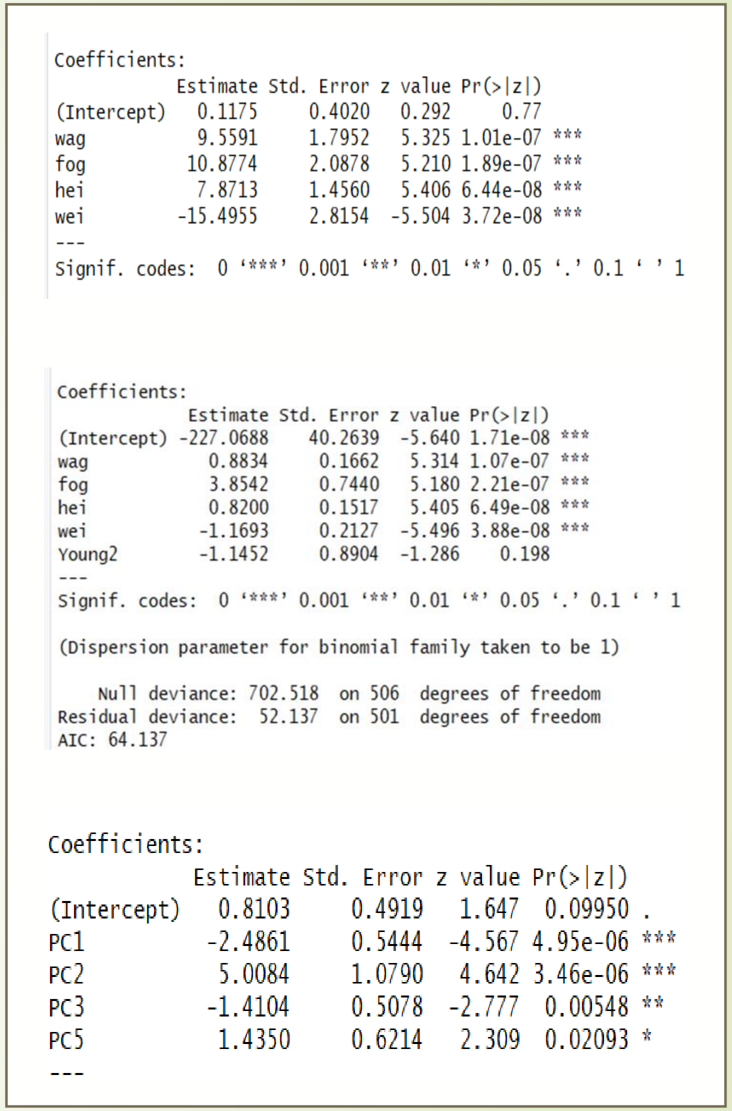
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.214e-17  8.413e-03   0.000 1.000000
ched         4.596e-02  1.449e-02   3.173 0.001640 **
kne          1.066e-01  1.487e-02   7.168 4.37e-12 ***
shg          8.934e-02  2.493e-02   3.583 0.000387 ***
chg          1.679e-01  2.709e-02   6.197 1.58e-09 ***
wag          2.683e-01  2.149e-02  12.485  < 2e-16 ***
hig          7.340e-02  2.179e-02   3.368 0.000839 ***
thg          1.417e-01  1.992e-02   7.116 6.10e-12 ***
cag          8.592e-02  1.525e-02   5.633 3.59e-08 ***
hei          2.187e-01  1.312e-02  16.667  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1614 on 358 degrees of freedom
Multiple R-squared:  0.9746,    Adjusted R-squared:  0.974
F-statistic:  1526 on 9 and 358 DF,  p-value: < 2.2e-16
```



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.454e-16  1.778e-02   0.000 1.000000
ched         1.283e-01  3.207e-02   4.003 0.000112 ***
wag          4.121e-01  3.938e-02  10.465  < 2e-16 ***
hig          2.648e-01  2.736e-02   9.676  < 2e-16 ***
cag          1.300e-01  2.533e-02   5.133 1.20e-06 ***
age         -8.061e-02  1.835e-02  -4.394 2.53e-05 ***
hei          2.954e-01  2.266e-02  13.040  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1948 on 113 degrees of freedom
Multiple R-squared:  0.964,     Adjusted R-squared:  0.962
F-statistic: 503.8 on 6 and 113 DF,  p-value: < 2.2e-16
```

Logistic Regression Models

- PCA: good, 1.78% misclassification error rate
- Reduced the regression: all insignificant, 2.17% misclassification error rate
- Now I take two age groups into account to classifications in the model. The misclassification error rate is 2.17%, which is the same in the previous model because the Young variable is not significant in the model.
- Classification by age groups: APER=41.42%



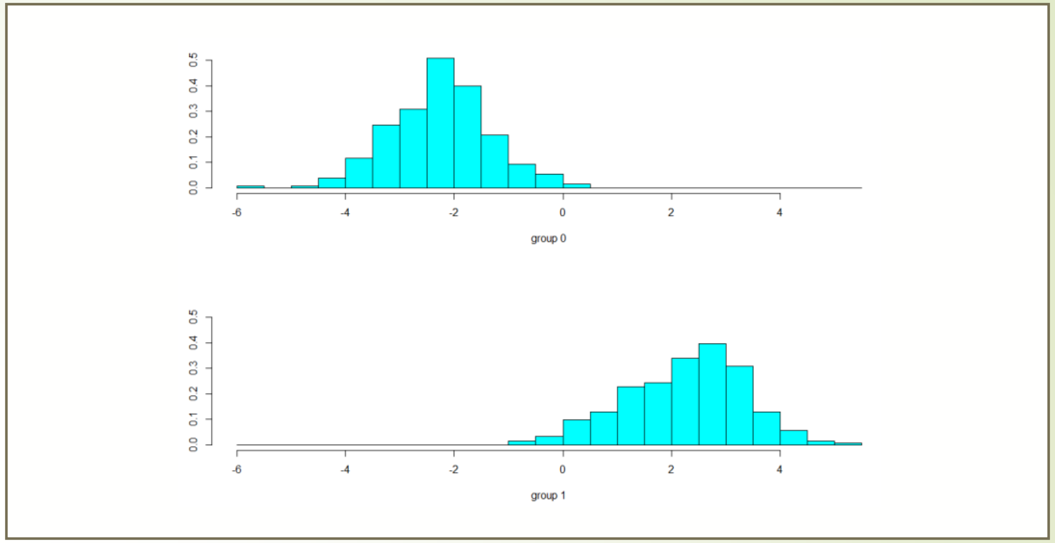
```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)   0.1175     0.4020   0.292     0.77    
wag           9.5591     1.7952   5.325 1.01e-07 ***
fog          10.8774     2.0878   5.210 1.89e-07 ***
hei           7.8713     1.4560   5.406 6.44e-08 ***
wei         -15.4955     2.8154  -5.504 3.72e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

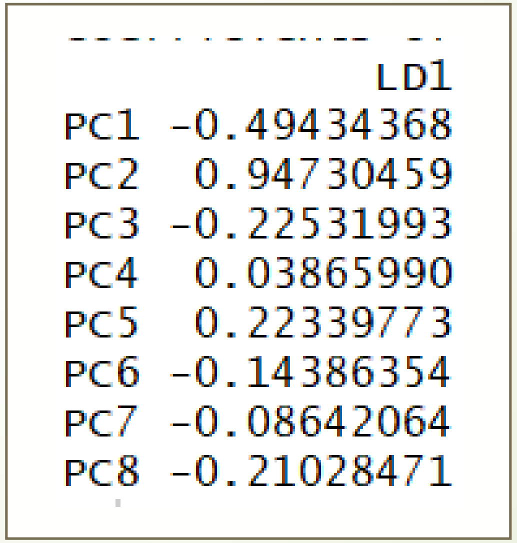
```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -227.0688    40.2639  -5.640 1.71e-08 ***
wag            0.8834     0.1662   5.314 1.07e-07 ***
fog            3.8542     0.7440   5.180 2.21e-07 ***
hei            0.8200     0.1517   5.405 6.49e-08 ***
wei           -1.1693     0.2127  -5.496 3.88e-08 ***
Young2        -1.1452     0.8904  -1.286    0.198    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 702.518  on 506  degrees of freedom
Residual deviance:  52.137  on 501  degrees of freedom
AIC: 64.137
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)   0.8103     0.4919   1.647  0.09950 .  
PC1          -2.4861     0.5444  -4.567 4.95e-06 ***
PC2           5.0084     1.0790   4.642 3.46e-06 ***
PC3          -1.4104     0.5078  -2.777  0.00548 ** 
PC5           1.4350     0.6214   2.309  0.02093 *  
---
```

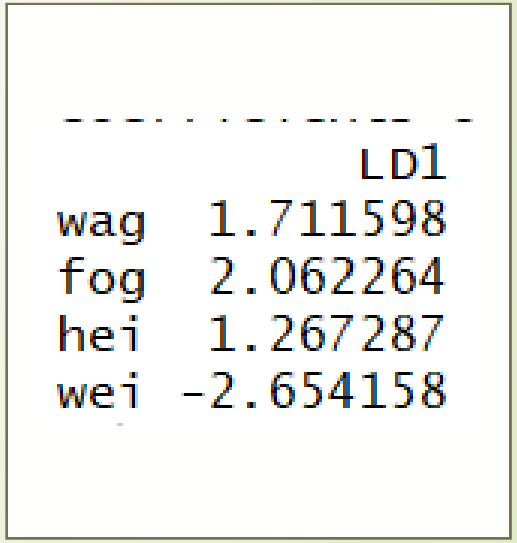
Linear Discriminant Analysis and Quadratic Discriminant Analysis

- LDA: with significant terms to fit, misclassification error rate is 2.76%
- with PCA method to fit, misclassification error rate is 1.58%
- QDA: Fitting the model with significant terms, the misclassification rate is 2.96%
- Fitting the model with PCA method, the misclassification error rate is 1.78%.





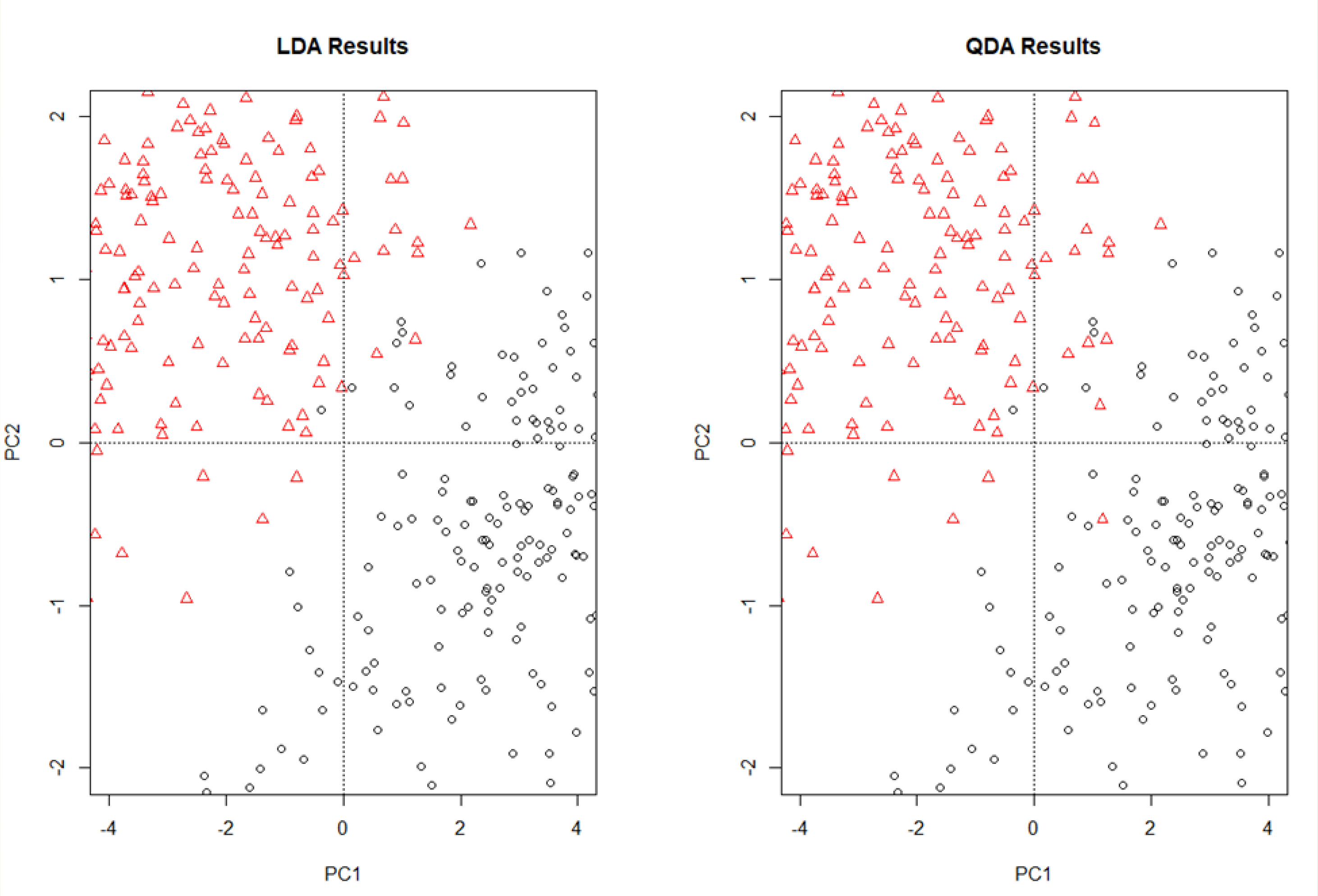
```
          LD1
PC1 -0.49434368
PC2  0.94730459
PC3 -0.22531993
PC4  0.03865990
PC5  0.22339773
PC6 -0.14386354
PC7 -0.08642064
PC8 -0.21028471
```



```
          LD1
wag  1.711598
fog  2.062264
hei  1.267287
wei -2.654158
```

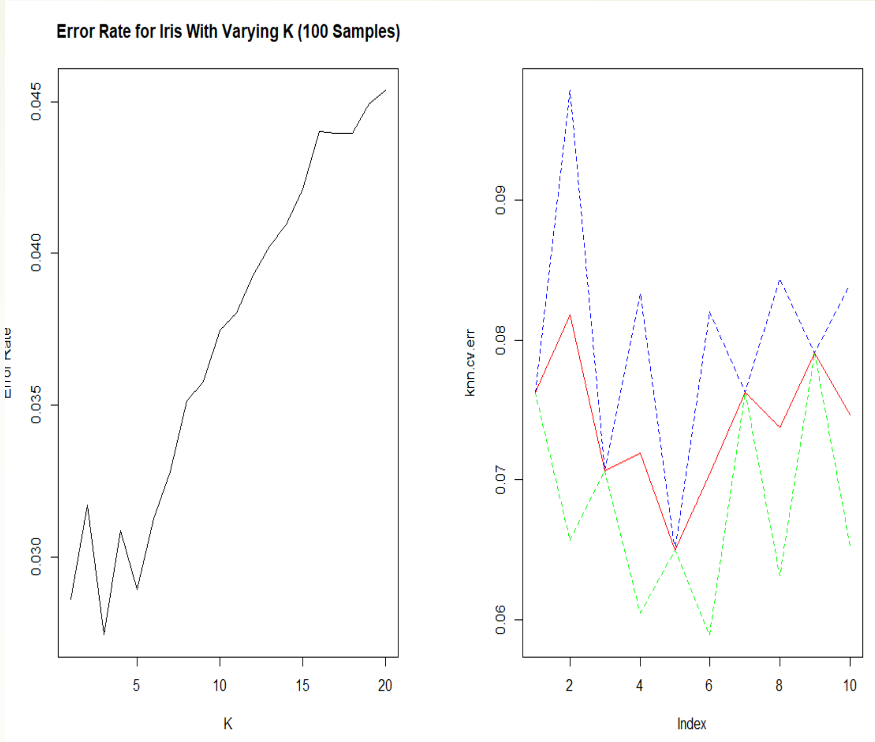
PCA, LDA and QDA

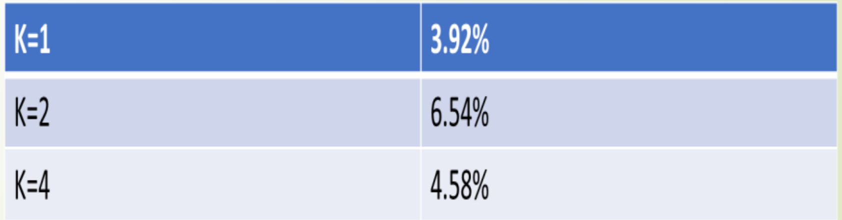
- Obviously, the LDA and QDA results demonstrate the separation in males and females well even though there are few observations misclassified from the plots.
- PC1 is good discriminant.



K-Nearest Neighbors

- Take 70% of the sample as the train dataset, 30% of the sample as the test dataset.
- Test the significant terms: waist girth, forearm girth, weight and height
- When K=5, the misclassification error rate is 5.23%.





K=1	3.92%
K=2	6.54%
K=4	4.58%

Classification Tree

Fitting the complete model by using standardized data, APER is 1.38%. After pruning the tree with size 8, APER is 2.17%.

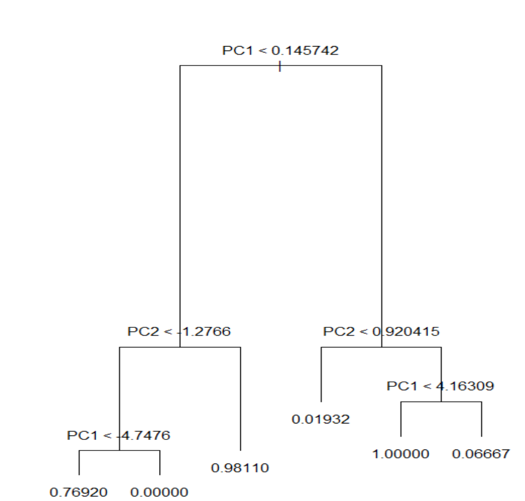
Gini method: APER is 3.75%.

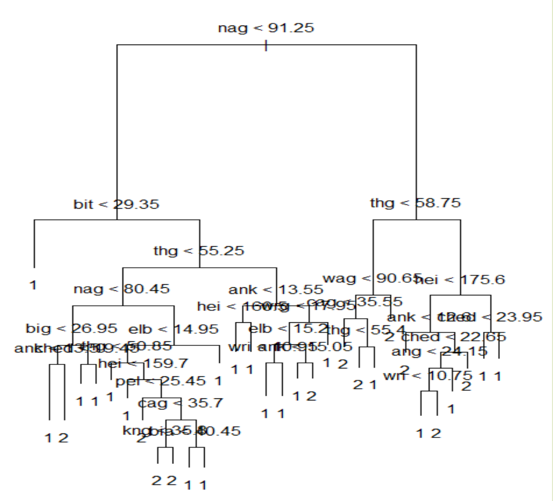
Take two age groups into account, APER is 9.67%.

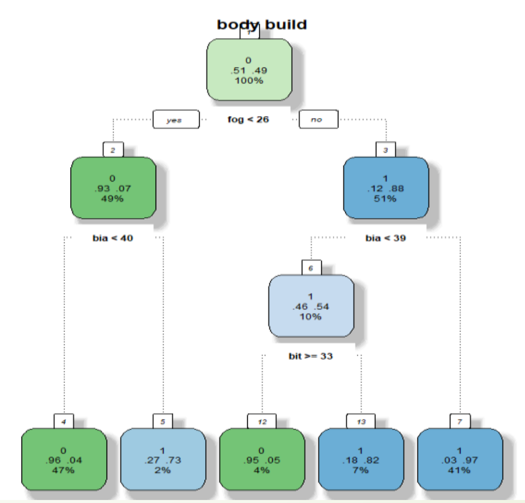
Gini method: APER=11.24% taking age groups.

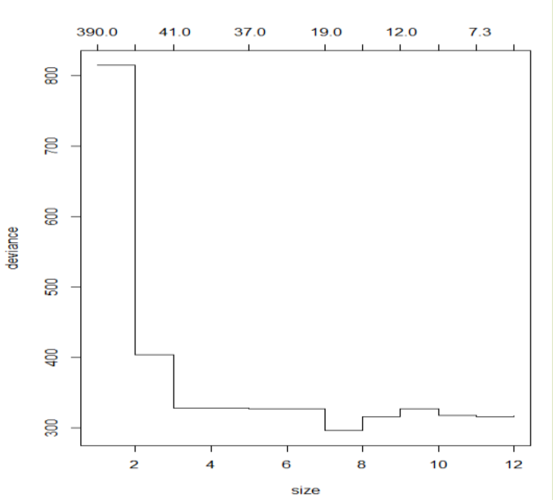
Take PCA method: APER=2.213%.

PC1 and PC2 are the most important splitters.



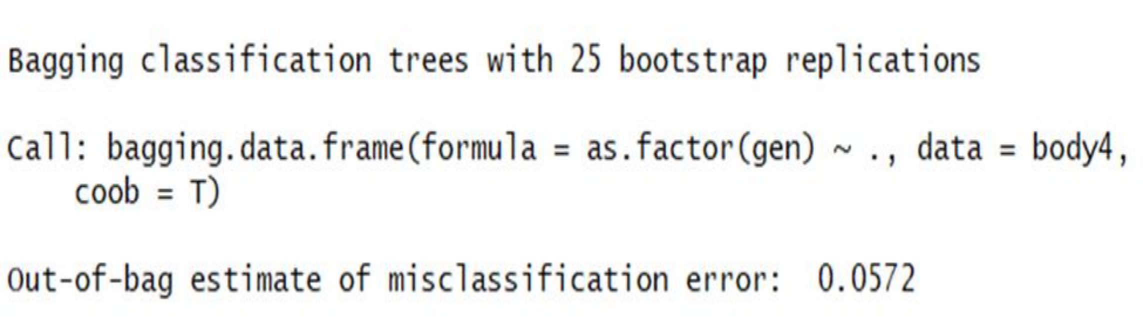




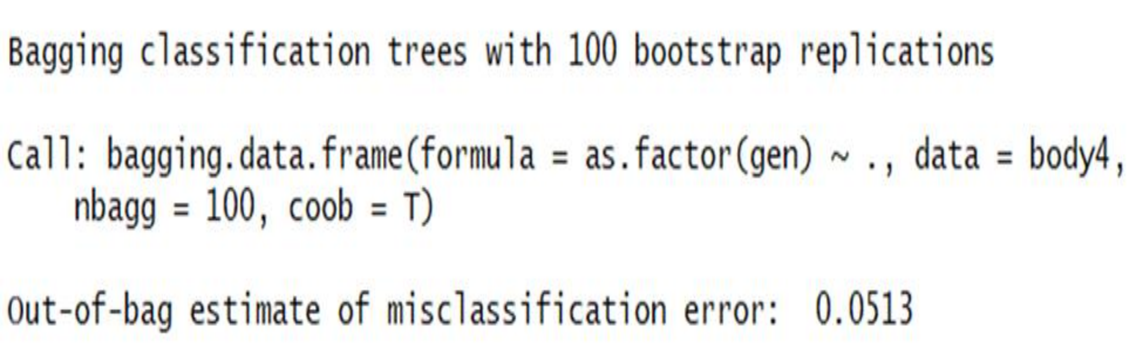


Bagging Analysis

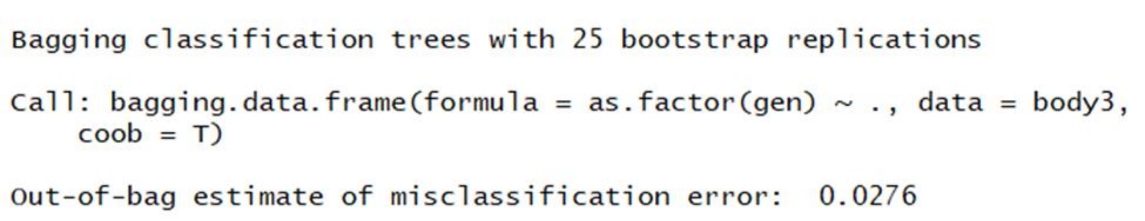
- N=25, OOB=5.72%
- N=100,OOB=5.13%
- Take PCA method, OOB=2.76%.



```
Bagging classification trees with 25 bootstrap replications

Call: bagging.data.frame(formula = as.factor(gen) ~ ., data = body4,
    coob = T)

Out-of-bag estimate of misclassification error:  0.0572
```



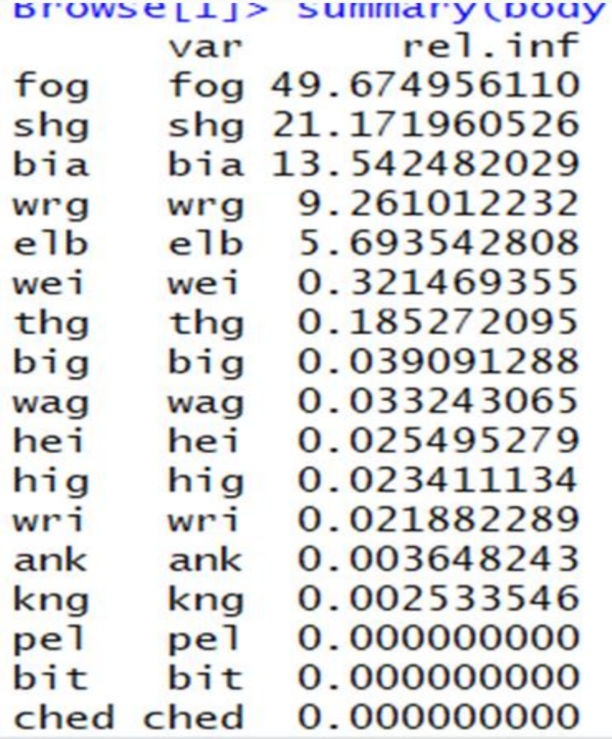
```
Bagging classification trees with 100 bootstrap replications

Call: bagging.data.frame(formula = as.factor(gen) ~ ., data = body4,
    nbagg = 100, coob = T)

Out-of-bag estimate of misclassification error:  0.0513
```



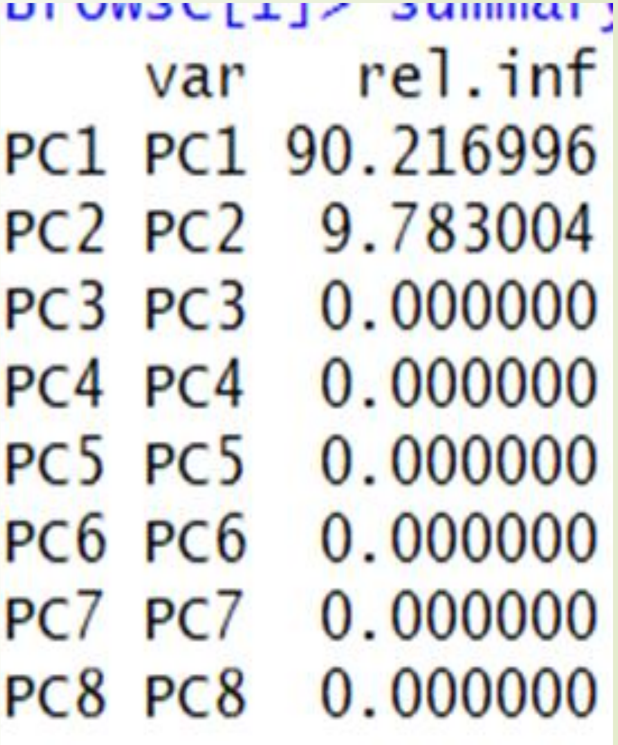
```
Bagging classification trees with 25 bootstrap replications

Call: bagging.data.frame(formula = as.factor(gen) ~ ., data = body3,
    coob = T)

Out-of-bag estimate of misclassification error:  0.0276
```

Boosting Analysis

- N=100, APER=48.72%
- N=500, APER=48.72%
- Take PCA method,
- N=100, APER=14.398%
- N=500, APER=14%



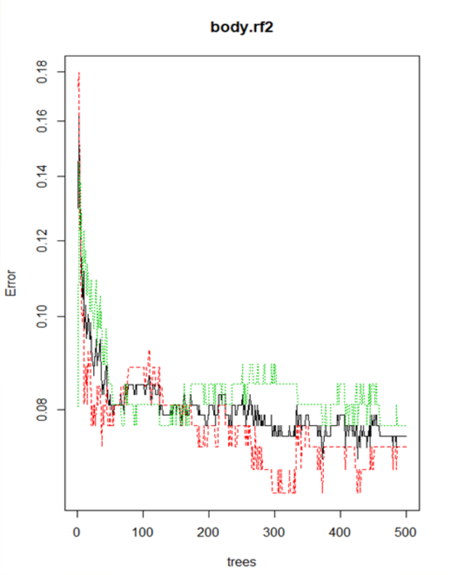
```
Browse[1]> summary(body
          var       rel.inf
fog       fog  49.674956110
shg       shg  21.171960526
bia       bia  13.542482029
wrg       wrg   9.261012232
elb       elb   5.693542808
wei       wei   0.321469355
thg       thg   0.185272095
big       big   0.039091288
wag       wag   0.033243065
hei       hei   0.025495279
hig       hig   0.023411134
wri       wri   0.021882289
ank       ank   0.003648243
kng       kng   0.002533546
pel       pel   0.000000000
bit       bit   0.000000000
ched     ched   0.000000000
```

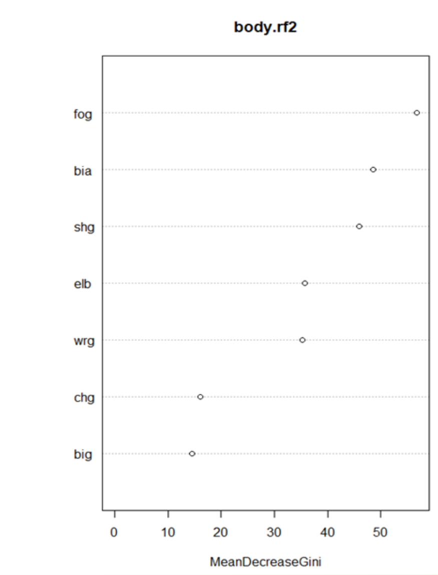


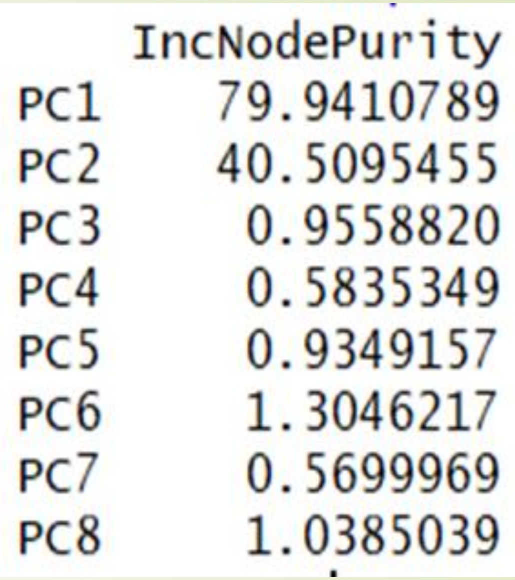
```
Browse[1]> summary
    var   rel.inf
PC1 PC1 90.216996
PC2 PC2  9.783004
PC3 PC3  0.000000
PC4 PC4  0.000000
PC5 PC5  0.000000
PC6 PC6  0.000000
PC7 PC7  0.000000
PC8 PC8  0.000000
```

Random Forest Analysis

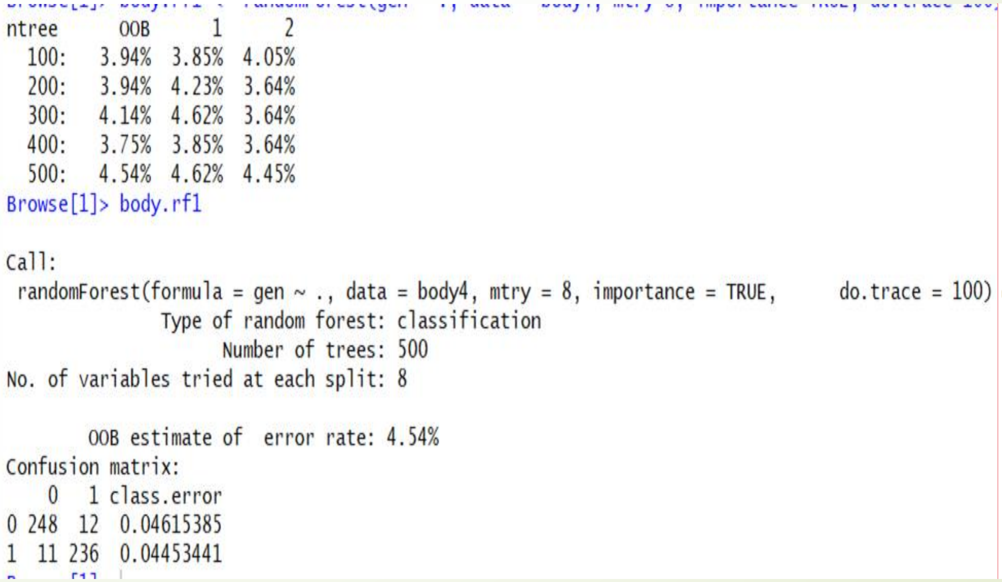
- N=100, OOB=4.54%
- N=400, OOB=3.75%
- Take PCA method, APER=2.12%.



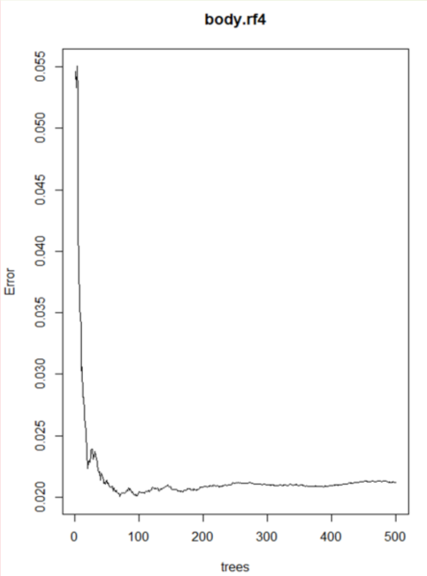




	IncNodePurity
PC1	79.9410789
PC2	40.5095455
PC3	0.9558820
PC4	0.5835349
PC5	0.9349157
PC6	1.3046217
PC7	0.5699969
PC8	1.0385039



```
Browse[1]> body.rf1 <- randomForest(gen ~ ., data = body4, mtry = 8, importance = TRUE, do.trace = 100)
ntree      OOB      1      2
  100:   3.94%  3.85%  4.05%
  200:   3.94%  4.23%  3.64%
  300:   4.14%  4.62%  3.64%
  400:   3.75%  3.85%  3.64%
  500:   4.54%  4.62%  4.45%
Browse[1]> body.rf1

Call:
 randomForest(formula = gen ~ ., data = body4, mtry = 8, importance = TRUE,      do.trace = 100)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 8

        OOB estimate of  error rate: 4.54%
Confusion matrix:
    0   1 class.error
0 248  12  0.04615385
1  11 236  0.04453441
```



Compare all results

Methods	APER (PCA)
Logistic Regression	1.78%
Linear Discriminant Analysis	1.58%
Quadratic Discriminant Analysis	1.78%
K-Nearest Neighbors	5.23%
Classification Regression Tree	2.213%
Bagging Analysis	2.76%
Boosting Analysis	14%
Random Forest Analysis	2.12%

The linear discriminant analysis does perform the best of all models considered to the structure of this data in this kind of prediction.

Conclusion

- Even though linear discriminant analysis, logistic regression and quadratic regression perform better than classification regression tree based on APER, CART model has the advantage of being much more interpretable with 8 components consisting of all 24 inputs. It's obvious the PC1 and PC2 are the most significant splitter, which makes perfect sense. Given the key difference between males and females, the young adults have no much impact on the skeletal and girth measurements but the middle-aged or older adults do have some influence on body build dimensions in males and females.
- As we expected, height variable affects the weight significantly and positively in males and females. In complete regression model, girth measurements account for larger percentage on weight as chest girth, waist girth, hip girth et al. are main parts in body build for designing. In logistic regression model, all input are no so statistically significant but waist girth, forearm girth, height and weight contribute much larger difference in two groups.
- Moreover, with PCA method, the PC1,PC2,PC3 and PC5 take more contribution to separation in males and females where body build dimensions are different. Also, taking PCA method is more comprehensive for demonstrating the differentiation in males and females.

Thank You!