**625 Final Project**

**Dan Huang**

**Investigating the Underlying Relationships in Body Dimensions**

**1. Abstract**

A dataset consisting of skeletal diameter measurements and body girth measurements were provided by Grete et al and the other three predictor such as age, height, weight also included in the dataset, which can be used as the identified groups or standards for classifying the different body dimensions. In the process of investigating the relationships in body dimensions, the purpose of this report will target on the various body dimensions with essential statistical analysis to explore the significant variables on influencing the evolution of body structure in women and men.

**2. Introduction**

Separating the body dimensions in different group such as female and male has reasonable way to distinguish while measuring the body build, as well as the body girth and weight, is complexed procedure. In the dataset, the body build (skeletal width and depth) measurements that were measured at nine complete active body spots could be used to analyze the relationship in weight with height. The assumption that including height may be significant and better in the analysis was confirmed in the group using regression. In the regression, the body build variables and height were contained simultaneously in the weight equation, which let me select the vital variables and continue the subsequent analysis.

At twelve body sites, trunk and limb girth measurements were taken that were used to determine the weight equation by regression analysis. Then the girth variables that have larger impact on weight would be chosen into new regression equation, combining with skeletal variables from last step. The estimation of weight through skeletal variables and body build girth variables can be used to decide the design of art analysis such as commercial or military uniforms. In order to analyze the variables efficiently, all variables except gender were scaled in the analysis to determine the best prediction in weight.

There would be challenging issue when discriminant analysis is applied because the variables election that seems to be optimal in the dataset could have the noise to disturb the accuracy of the analysis process. In the data mining for body build dimensions, discriminating gender groups or age groups by body build dimensions or other predictors will be the interesting thing to analyze and compare the results. It was demonstrated by Lohman et al. (1988) that biacromial diameter is "useful ... in the evaluation of sex-associated differences in physique." In this report, I will use gender and age as the classifier variables to determine the difference between groups via principle components and significant skeletal and girth measurements.

|  |  |  |
| --- | --- | --- |
| **Variables** | **Measurement** | **Comments** |
| Biacromial | Diameter/cm | Measure the distance across the shoulders |
| Pelvic Breadth | Diameter/cm | Between the abdomen and the thighs or skeleton embedded |
| Bitrochanteric | Diameter/cm | The distance between the outer points of the hips |
| Chest Depth | Diameter/cm | Between spine and sternum at nipple level, mid-expiration |
| Chest | Diameter/cm | At nipple level, mid-expiration |
| Elbow | Diameter/cm | Sum of two elbows |
| Wrist | Diameter/cm | Sum of two wrists |
| Knee | Diameter/cm | Sum of two knees |
| Ankle | Diameter/cm | Sum of two ankles |
| Shoulder Girth | cm | Over deltoid muscles |
| Chest Girth | cm | Nipple line in males and just above breast tissue in females, mid-expiration |
| Waist Girth | cm | Narrowest part of torso below the rib cage, average of contracted and relaxed position |
| Navel Girth | cm | At umbilicus and iliac crest |
| Hip Girth | cm | At level of bitrochanteric diameter |
| Thigh Girth | cm | Below gluteal fold, average of right and left girths |
| Bicep Girth | cm | Flexed, average of right and left girth |
| Forearm Girth | cm | Extended, palm up, average of right and left girths |
| Knee Girth | cm | Over patella, slightly flexed position, average of right and left girths |
| Calf Maximum Girth | cm | Average of right and left girths |
| Ankle Minimum Girth | cm | Average of right and left girths |
| Wrist Minimum Girth | cm | Average of right and left girths |
| Age | years | Range: 18-67,median 27 |
| Weight | kg | Kg |
| Height | cm | cm |
| Gender | categorial | Males and females |

**3. Methods**

The methods proposed in this project are different, which depend on the purposes of analysis. For testing weight against skeletal measurements and girth measurements multiplied height, multiple regression models were taken in the project. Based on the reduced regression model with significant terms, it’s also useful to apply logistic regression model, linear discriminant analysis, quadratic discriminant analysis, K-Nearest Neighbors, classification regression tree, boosting, bagging and random forest. Results obtained from principle components also were applied in discriminant analysis by using scores.

For a binary response problem, logistic function can model the predictors with response Y either be 1 or 0: p= E(Y|Z) =β0+β’Z with an error term. Robust and parsimonious parameter selection also requires the final fitted model with identified parameters by using logistic regression.

The linear discriminant analysis algorithm gives this data to separate the space of predictor variables into two groups. The difference from principle component analysis is that LDA choose dimensions that maximally distinguish the regions in the transformed space. Given the same multivariate Gaussian distribution for all predictors of each group, LDA takes the coefficients of a discriminant function for each group according to each scaled variable. Generally, we used the highest score in each function to assign a case predicted to a group.

The quadratic discriminant analysis is similar to LDA except containing the second order terms and the covariance matrix is not identical. Defining class which maximize the quadratic functions is the key task in QDA.

As a classification and regression non-parametric method, K-nearest neighbor’s primary idea is to predict the new point with closest distance by presorting the number of training samples and allocating from these. Concerned the simplicity, KNN can be considered a simple machine learning method for both classification and regression.

For predicting modeling problems, classification regression tree is one of the most important algorithm. As the foundation of classification method, bagging, boosting and random forest methods have been derived from CART, which provides us comprehensive model selection. CART is the process of separating the inputs space by creating a binary decision tree with different splitters. Here, CART has been applied to classifying the gender and age groups respectively with and without Gini index function.

Tree Network is a very powerful method for analysists to investigate predictive models with minimum cost. By using stochastic gradient boosting to iterate procedure, the purpose of tree network is to fit the model with small classification trees by comparing the error of all previous trees.

**4. Result**

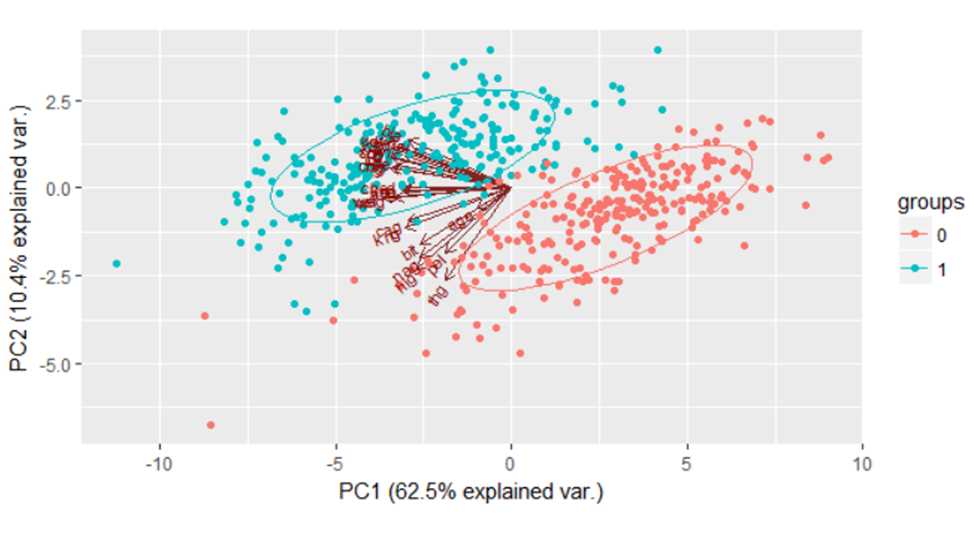
4.1 Multiple Regression

By modeling weight as the linear combination of all skeletal and girth measurements including height and age, I obtained the adjusted R2 value of 97.49% and 2.113 of residual standard error. From the fitted reduced model with skeletal measurements including height and age, all terms are positively significant except biacromial diameter and age, which is interesting. In the fitted reduced model with girth measurement including height and age, all terms are positively significant except age. Then, I took whole body dimensions predictor variables including height and age into consideration, age variable still has negative impact on weight, which means as the age of people increases, the weight would decrease. This result maybe inaccurate from the practice. Therefore, I wonder if age variable can be split into two groups to see how age influence on weight. In the young group, age variable has no significant difference on weight while in elder group, age variable does affect weight negatively, indicating as people grow up starting from 35, the weight will decrease, which makes more sense. Overall, the girth measurements take larger contribution on predicting the weight in the model than skeletal measurements. In summary of all three final reduced models above, pelvic diameter, chest depth, knee diameter, shoulder girth, chest girth, waist girth, hip girth, thigh girth, forearm girth, knee girth, calf girth, age and height are important predictors in weight equation.

4.2 Logistic Regression

4.2.1 take first eight components solution

As the first eight principle components can explain 90.481% of the total variability in the model, I selected the first eight components to model logistic regression in gender. The first three components and the fifth components are significant in the model, which makes sense as the first two components can separate the gender groups well shown below. The first component can explain 62.5% of the total variation and the second principle component can explain 10.4% of the total variation. The girth measurements take higher loadings on the first component, which shows the first component can represent the girth factor. The second principle component can interpret thigh girth factor.



4.2.2 take all significant terms in previous regression models into logistic regression

It’s interesting that all terms included are not significant with scaled data; therefore, I chose four persuasive predictors such as waist girth, forearm girth, height and weight into logistic regression. The fitted reduced model explains well on classifying the two regions with significant variables but the result of APER is not better than that of PCs method.

4.2.3 take two age groups into account

The result indicates age group variables is not statistically significant in the model, which shows there is no difference on age variable to classify males and females.

4.3 Linear Discriminant Analysis

4.3.1 complete regression model

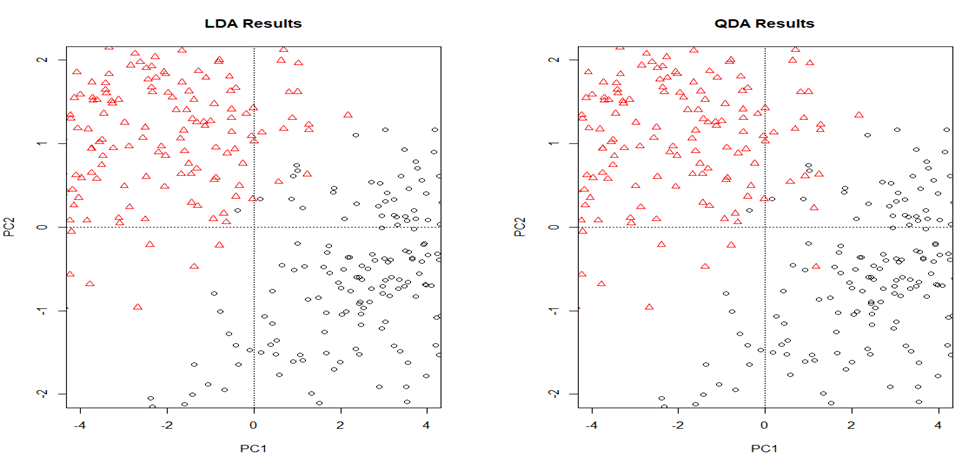
I took the significant predictors from the previous regression models to analyze the LDA and found out weight takes largest associated with group 0 (female) and forearm girth has the largest associated with group 1 (male) in the discriminant function.

4.3.2 Principle Components Method

Applying the first eight principle components to LDA is also good idea. The second principle component has largest associated with group 1 and the first component takes largest associated group 0, which also matches the result that PC1 represents girth factor and PC2 interprets skeletal factor.

4.4 Quadratic Discriminant Analysis

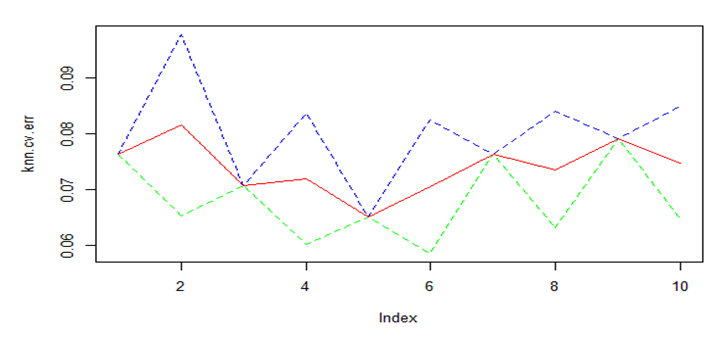
The females and males can be classified well by the combination of significant terms such as waist girth, forearm girth, height and weight and the first eight components, respectively. The outcome of misclassification error rate of PCs is better than that of significant terms in the QDA.



By comparing the result of PCA, LDA and QDA, it’s obvious that the first two components separate males and females well and there are high positive scores loading on PC1 in group 1 and negative scores loading on PC1 in group 0 even though there are some misclassifications in LDA and QDA.

4.5 K-Nearest Neighbors Analysis

Starting from here, I chose 70% of the dataset as the training sample and 30% of the dataset as the test sample with all variables to predict the new points. Next step, the model testing with waist girth, forearm girth, weight and height predicts the new observation well in gender. The red curve indicates the best K may be at 5 with the smallest misclassification error rate. The APER is 5.23%, indicating only 5.23% of the observations in the dataset that are misclassified. There are 2 observations from group 0 are misclassified in group 1; 6 observations from group 1 are misclassified in group 0.



4.6 Classification Regression Tree Analysis

4.6.1 Full Classification Tree

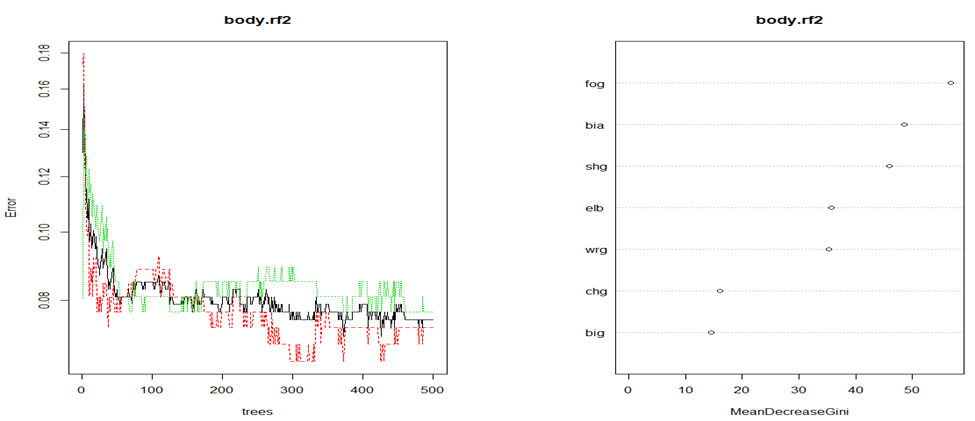
At first, I took all variables into account with standardized data by applying the classification tree and found out forearm girth, shoulder girth and biacromial diameter are important splitters in the tree with the misclassification error rate 1.38%.

4.6.2 Principle Components Method

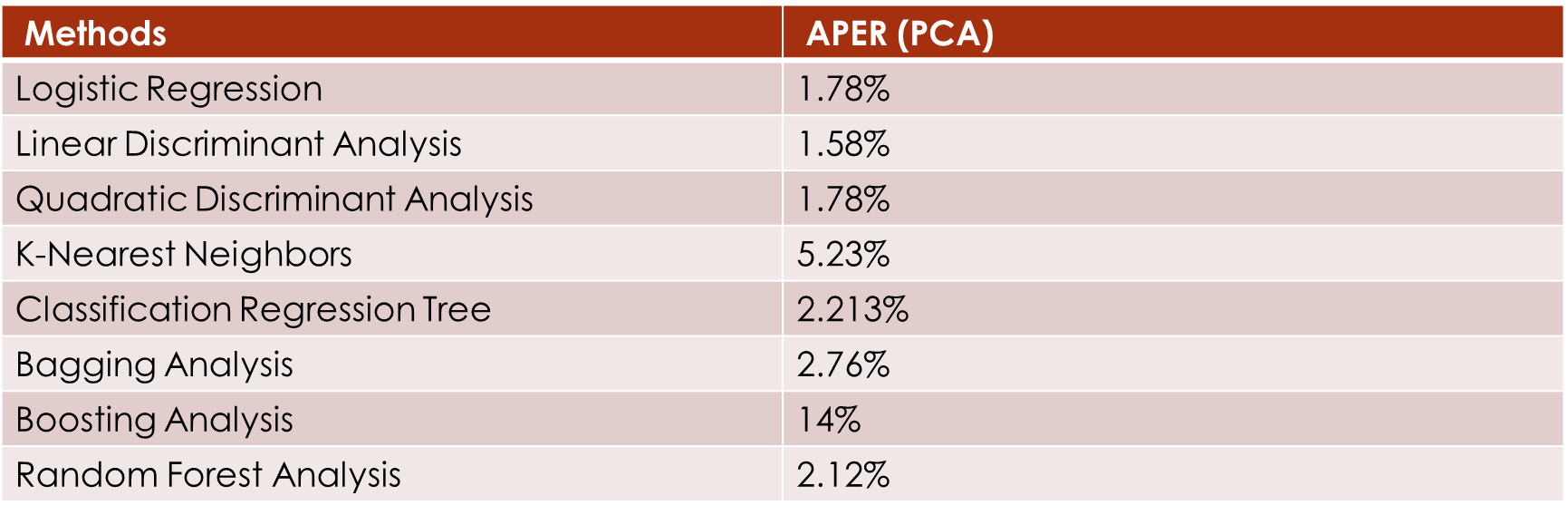
Principle components method is also good way to be applied in CART since the first eight principle components explain the 90.481% of the total variability. In the CART, the first principle and second principle components are the most important splitters. The first principle component loaded highly on girth measurements, which could represent girth body build factor; the second principle component loaded highly on skeletal body build factors, which interprets the skeletal factor.

4.7 Random Forest Analysis, Bagging and Boosting

As a useful machine learning algorithm, Random Forest algorithm creates a small twist to Bagging and is more elaborate than Bagging since Bagging only has to change one parameter that is the number of the tree while Random Forest has two parameters that contain the number of trees and how features creation form the best feature that splits the data in the tree. Random Forest decreases the correlation between trees to improve the variance by selecting feature-subset randomly for splitting the nodes while boosting and bagging generate multiple fully-grown trees of train sets to improve the variance. In this case, Random Forest performs best at the least misclassification error rate by taking PCs method. From the graph below, the red curve, green curve and black curve almost merge and when n=400 or nearly, the error rate is smallest. The forearm girth takes the largest importance in the classification.



**5. Conclusion**



Even though linear discriminant analysis, logistic regression and quadratic regression perform better than classification regression tree based on APER by applying PCs method, CART model has the advantage of being much more interpretable with 8 components consisting of all 24 inputs. It’s obvious the first principle component and the second principle component are the most significant splitter, which makes perfect sense since the first two components explained 72.89% of the total variability. Given the key difference between males and females, the young adults have no much impact on the skeletal and girth measurement but the middle-aged or elder adults do have some influence on body build dimensions in males and females.

It’s obvious from the data structure that weight could be predicted by the linear combination of skeletal measurements and girth measurements including height. First, from the regression model of weight and skeletal measurements adding height and age, the model indicates that biacromial diameter, pelvic diameter, bitrochanteric diameter, chest depth, chest diameter, wrist diameter, knee diameter and height affect the prediction of weight significantly and positively while age influences the weight negatively. Next, predicting the weight by the model against the linear combination of girth measurements including height and age, all significant terms positively affect the weight except age. Finally, taking all measurements into account plus height and age, I found out only age variable has negative influence on weight. Considering all regression model, weight is predicted by age negatively, which means the weight will decrease as age increases and this is not understandable. Therefore, I filtered the age into two groups and the age on young group is not significant to the prediction of weight while the age variable on elder group has the impact on weight negatively, which shows the weight goes down as the middle-aged becomes elder.

As 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 inputs are no so statistically significant but waist girth, forearm girth, height and weight contribute much larger difference in two groups.

Taking the age groups into account and predicting the gender by significant terms, the logistic regression shows there is no significant impact on the prediction of gender by age group variables. 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 prediction of males and females from the body build dimensions.

Looking at the classification of age groups by body build dimensions, navel girth, bitrochanteric diameter and thigh girth are the most important splitters in the tree with 9.67% misclassification error rate. In contrast, forearm girth, shoulder girth and biacromial diameter are the most significant splitters in the tree with only 1.38% misclassification error rate and less number of terminal nodes in classifying gender by body build dimensions.

**6. Recommendation**

Starting from the principle component analysis, it’s easy to observe the separation of males and females by all skeletal measurements and girth measurements but there is interesting thing that splitting people with different age into two groups is kind of cautious work. Both logistic regression and tree network results suggest that forearm girth is essential for some attribution to classify the gender. Further the work of this report mainly concentrates on identifying the regression models for gender prediction while paying more attention on age classification could a potential challenge to preform data mining analysis in the subsequent work.

**Reference:**

Friedman, J. H. , Greedy Function Approximation: a Gradient Boosting Machine, Technical report, Dept. of Statistics, Stanford University, 1999

Heinz, G., Johnson, R.W., Kerk, C.J., and Peterson, L. J. (2003), “Exploring Relationships in Body Dimensions,”, Journal of Statistics Education Volume 11, (ww2.amstat.org/publications/jse/v11n2/datasets.heinz.html)

Lohman, T., Roche, A., and Martorell, R., (eds.) (1988), Anthropometric Standardization Reference Manual, Champaign, IL: Human Kinetics Books.