Abstract

The essential objective of this project per se is to investigate the relationship of the height of an individual among other human body measurements. Techniques involving data transformation, variable selection and linear regression model diagnostics will be implemented with a purpose of constructing a multiple linear regression model with decent explanatory power and an acceptable level of multicollinearity, which is able to be interpreted to reveal how a person's height is related to other body figures. The final conclusion, where the height of an individual is modelled to be related to weight, gender, biacromial, pelvic breadth and waist girth, features presentable fit to the provided data, satisfactory significance, as well as properly alleviated multicollinearity between predictors. The methodology, results and general evaluation regarding the project as a whole will be described and discussed in the paper to follow.

1. Introduction

In the project, candidates are provided with a data-set containing different body measurements, describing physical properties of each individual from a myriad of dimensions.

The main goal of the project is to build a model that better represent the data-set, explaining the height of an individual using some or all other body measurements available. In order to do so, one of, if not the most important issue is the balance between explanatory power and model complexity.

This paper seeks to handle the balance by achieving a desirable level of multicollinearity while not sacrificing too much explanatory power. Moreover, significance of the model is always key for a well-rounded result. Techniques such as covariate transformation, variable selection using best subset method, Box-Cox transformation, hierarchical ANOVA and detecting influential points with Cook's distance and high leverage points with hat values will be used to analyse the data-set and construct a model.

1.1 Description of the data-set

The data-set body.dat contains 21 body dimension measurements, together with age, weight, height, and gender on 507 individuals. The 247 men and 260 women were primarily individuals in their twenties and thirties, with a scattering of older men and women, all exercising several hours a week.

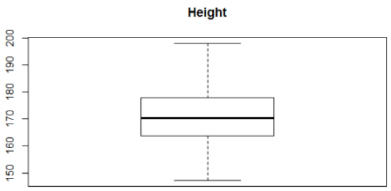


Figure 2.0

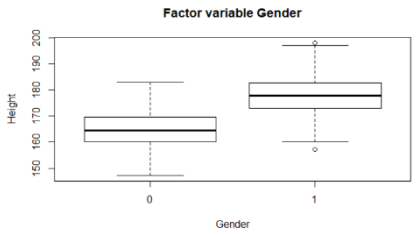


Figure 1.1

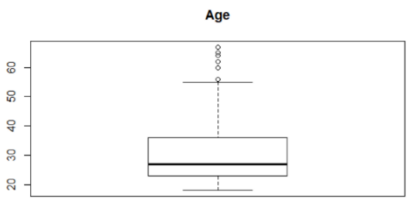


Figure 3.2

We may be interested in an insight into the response variable “height” for the model. Figure 1.0 shows a boxplot of the response variable “height” and it can be observed that the variable displays a satisfactorily symmetric pattern, which could potentially indicate that no transformation needs to be made regarding the variable.

Another variable worth inspection is “gender”, which should be transformed into a factor variable with 2 levels. Figure 1.1 shows a boxplot of height with respect to 2 different genders. The factor levels show significant mean effect from the boxplot, which could potentially indicate that the factor variable “gender” is a significant predictor explaining the height of an individual.

It is worth noticing that the data-set may contain highly skewed variables, which is epitomised by the variable “age” as per Figure 1.2. The overt positively skewness and outliers in the plot might potentially indicate the need for a log-transformation on the variable, as well as the need for outlier detection and removal.

1.2 General brief of the approach:

The methodology adopted in the project involves data transforming and cleaning before implementing variable selection, variable selection according to best subset method, intuition and alleviating multicollinearity, Box-Cox transformation, high Cook’s distance points and high leverage points detection.

2. Methodology

The predictors in the data-set are first studied their skewness to inspect whether there are highly skewed variables that are expected to be transformed to reduce their skewness. Best subset variable selection is then implemented to remove some variables. Further variable removals are based on intuition, reduction in VIF and model significance as per hierarchical ANOVA. High Cook’s distance points and high leverage points are then removed to see whether there are improvements in the explanatory power (measured by adjusted R-square) and normality of the residuals. The conclusion features a multiple linear regression model relating the height to weight, gender, biacromial, pelvic breadth and waist girth with a model specification as follows:

$$Height = \beta 0 + \beta 1 \times Weight + \beta 2 \times Gender + \beta 3 \times Biacromial + \beta 4 \times Pelvic.breadth + \beta 5 \times Waist.girth$$

2.1 Data transforming and cleaning

The data-set is found to be a perfect one without missing values.

The skewness of 23 variables from the data-set with the exception of the response variable “height” and a factor variable “gender” is inspected for decision of transformation. Figure 2.0 is the skewness table for the 23 variables.

If the skewness value lies above +1 or below -1, data is considered highly skewed. If it lies between +0.5 to -0.5, it is regarded as moderately skewed. If the value is 0, then the variable is symmetric. It is observed that the “age” variable is highly skewed with a skewness of approximately 1.125 and is therefore log-transformed to alleviate its skewness. The skewness decreases to approximately 0.54 after taking the natural logarithm of the “age” variable.

	Skewness
biacromial	0.155843
pelvic.breadth	-0.416262
bitrochanteric	-0.086988
chest.depth	0.493892
chest.diam	0.258655
elbow.diam	0.052680
wrist.diam	0.047998
knee.diam	0.340130
ankle.diam	0.069420
shoulder.girth	0.139859
chest.girth	0.238129
waist.girth	0.539026
navel.girth	0.448460
hip.girth	0.496789
thigh.girth	0.688936
bicep.girth	0.220367
forearm.girth	0.152586
knee.girth	0.467740
calf.girth	0.276840
ankle.girth	0.403370
wrist.girth	0.151567
age	1.125043
weight	0.400607

Figure 2.0

2.2 Best-subset variable selection using BIC

The best-subset variable selection method fits all the 2^p possible combinations of models and see which model gives the best model under a specific model comparison criterion when the number of

Adj.R2	CP	BIC
18	17	11

Figure 2.1

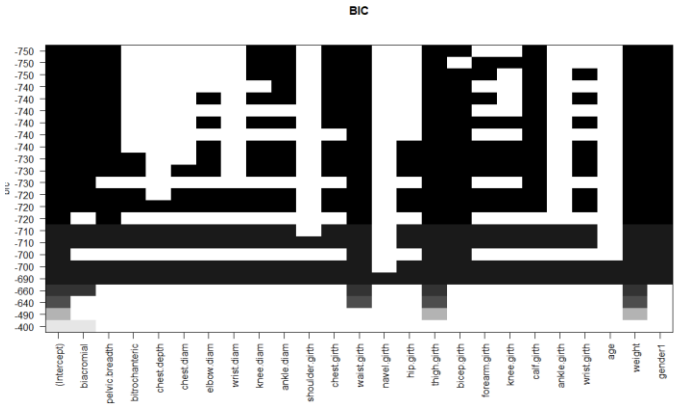


Figure 2.2

variables p is not too large. Such a method always returns the globally optimal model under the pre-specified criterion. However, the method becomes extremely computation-intensive as the number of variables p is large. Stepwise regression could be performed more easily, but could produce results that are far from the global optimum. The LASSO approach does not perform well, doing very little job in deleting variables, which will be discussed in the Appendix section.

The full model, which regresses "height" on all other 23 variables and a categorical factor variable "gender", is applied to best-subset variable selection method. It can be illustrated by Figure 2.1 that the best-subset method returns the globally optimal model with 18, 17 and 11 predictors as per Adjusted R-square, CP and BIC. Therefore, BIC, returning a model with 11 predictors according to Figure 2.1 and 2.2, is adopted as the model comparison criterion as it removes the most variables among the three, which better facilitates our further variable selection with the purpose of reducing multicollinearity and improving hierarchical ANOVA.

```
Call:
lm(formula = height ~ biacromial + pelvic.breadth + knee.diam +
    ankle.diam + chest.girth + waist.girth + thigh.girth + bicep.girth +
    calf.girth + weight + gender, data = body)

Residuals:
     Min       1Q   Median       3Q      Max
-14.8812  -2.9340   0.0289   2.6612  13.8033

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  190.45911    7.34007  25.948  < 2e-16 ***
biacromial     0.58643    0.11636   5.040 6.55e-07 ***
pelvic.breadth   0.36407    0.11029   3.301 0.001032 **
knee.diam     -0.81695    0.25329  -3.225 0.001341 **
ankle.diam     0.95049    0.26945   3.528 0.000458 ***
chest.girth   -0.24205    0.06290  -3.848 0.000135 ***
waist.girth   -0.58715    0.04731 -12.411  < 2e-16 ***
thigh.girth   -0.52097    0.08155  -6.388 3.88e-10 ***
bicep.girth   -0.55565    0.12326  -4.508 8.18e-06 ***
calf.girth    -0.56486    0.11734  -4.814 1.97e-06 ***
weight         1.19417    0.06268  19.052  < 2e-16 ***
gender1        5.16836    0.92149   5.609 3.40e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.221 on 495 degrees of freedom
Multiple R-squared:  0.803,     Adjusted R-squared:  0.7987
F-statistic: 183.5 on 11 and 495 DF,  p-value: < 2.2e-16
```

From the model summary, it can be observed that the 11-predictor model presents a decent adjusted R-squared of 79.87% and all predictors appear to be very significant in individual t-tests. However, when the Variance Inflation Factor (VIF) is obtained. The conventional acceptable level of multicollinearity restricts VIF values of all variables to be less than 10. Therefore, It is not difficult to observe that the model suffers severe multicollinearity. Further variable selection, aiming at reducing multicollinearity, should be performed.

```
> vif(BIC_model)
    biacromial pelvic.breadth      knee.diam     ankle.diam    chest.girth    waist.girth
      3.598530       1.681346       3.308575       3.207833      11.297870       7.708951
   thigh.girth    bicep.girth    calf.girth         weight         gender
      3.756919       7.781937       3.170434      19.871577       6.036482
```

2.3 Multicollinearity reducing variable selection

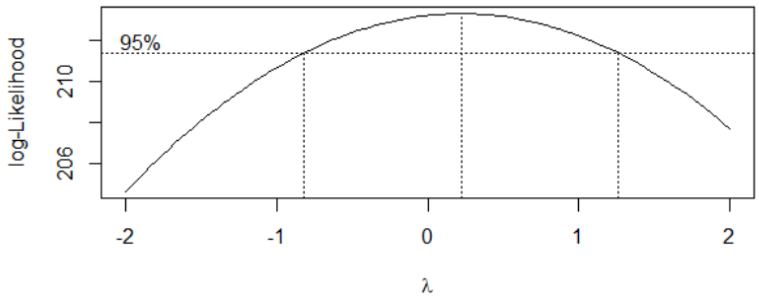
By intuition, further 6 variables, “knee.diam”, “ankle.diam”, “chest.girth”, “thigh.girth”, “bicep.girth” and “calf.girth” are deleted. The 4 deleted girth measurements, “chest.girth”, “thigh.girth”, “bicep.girth” and “calf.girth” are believed to be “nearly” linearly related to “waist.girth”. Intuitively, individuals with large waist girth measurements are very likely to be strong and hefty rather than slim, and therefore feature large values in all girth measurements, vice versa. The 2 deleted skeletal measurements, “knee.diam” and “ankle.diam” are believed to be closely related to “biacromial” and “pelvic.breadth”. Intuitively, individuals with large biacromial diameter and pelvic breadth tend to have rather large skeletons, and therefore feature large values in other skeletal measurements, vice versa.

```
Call:
lm(formula = height ~ weight + gender + biacromial + pelvic.breadth +
    waist.girth, data = body)

Residuals:
     Min       1Q   Median       3Q      Max
-15.6390  -3.3075  -0.0862   3.1029  15.1021

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    126.11028    5.22133  24.153  < 2e-16 ***
weight           0.57013    0.04549  12.533  < 2e-16 ***
gender1          7.43779    0.81156   9.165  < 2e-16 ***
biacromial       0.70746    0.13442   5.263 2.11e-07 ***
pelvic.breadth   0.63249    0.12446   5.082 5.29e-07 ***
waist.girth     -0.55953    0.05020 -11.145  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.023 on 501 degrees of freedom
Multiple R-squared:  0.7177,    Adjusted R-squared:  0.7149
F-statistic: 254.7 on 5 and 501 DF,  p-value: < 2.2e-16
```

```
> vif(model_no_chest.girth_calf.girth_bicep.girth_knee.diam_ankle.diam_thigh.girth)
       weight        gender     biacromial pelvic.breadth    waist.girth
     7.390482      3.306235      3.391083      1.512161      6.129691
```

From the model summary, it can be seen that all 5 covariates show satisfactory significance in individual t-tests. Adjusted R-square has decreased from 79.87% from last model to 71.49%, which is the unavoidable sacrifice for removing 6 variables from the “already best-subset selected” model. However, the VIF values for 5 variables have all been successfully reduced under 10, which means multicollinearity has been alleviated to an acceptable level within this model. High Cook’s distance points and high leverage points can be deleted to improve the model in terms of Adjusted R-square and normality.

2.4 Box-Cox transformation




```
        Jarque-Bera test for normality

data:  residuals(model_log)
JB = 0.59023, p-value = 0.7315
```

```
        Jarque-Bera test for normality

data:  residuals(model_no_chest.girth_calf.girth_bicep.girth_knee.diam_ankle.diam_thigh.girth)
JB = 0.26269, p-value = 0.879
```

Figure 2.3

According to Figure 2.3, the MLE of the power parameter λ appears to be very close to 0. Therefore, we take the natural logarithm of the response variable “height”. However, it is found that after the

log-transformation of the response variable “height”, the adjusted R-square remains basically unchanged while the normality of residuals reduces. Therefore, “no-transformation” may be a better idea as the value λ=1 is also included in the 95% region of log-likelihood.

2.5 High Cook’s distance points and high leverage points detection

28 high Cook’s distance influential points are first detected and removed in order to improve the adjusted R-square of the model.

```
Call:
lm(formula = height ~ weight + gender + biacromial + pelvic.breadth +
    waist.girth, data = body_no_influential)

Residuals:
     Min       1Q   Median       3Q      Max
-10.2482  -3.0704  -0.1087   3.0943  11.3115

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  128.37094    4.80994  26.689  < 2e-16 ***
weight         0.61853    0.04360  14.185  < 2e-16 ***
gender1        7.34341    0.73042  10.054  < 2e-16 ***
biacromial     0.66717    0.12234   5.454 7.97e-08 ***
pelvic.breadth 0.64148    0.11551   5.553 4.68e-08 ***
waist.girth   -0.61614    0.04784 -12.878  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.399 on 473 degrees of freedom
Multiple R-squared:  0.7627,    Adjusted R-squared:  0.7602
F-statistic:   304 on 5 and 473 DF,  p-value: < 2.2e-16
```

```
        Jarque-Bera test for normality

data:  residuals(model2)
JB = 3.9697, p-value = 0.126
```

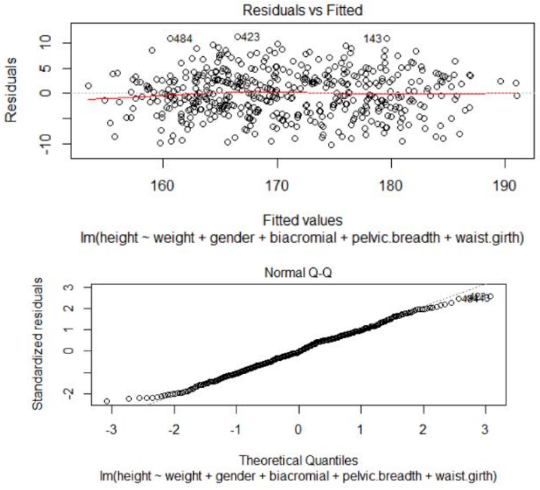
It can be observed that after removing points with Cook’s distance higher than 4/n, the model’s adjusted R-square has increased from 71.49% to 76.02%. The Jarque-Bera normality test remains past with a p-value of 12.6%. However, after deleting high leverage points with hat values higher than 2p/n, adjusted R-square falls and there is not a significant improvement on the normality of residuals. Therefore, points with high hat values are not removed.

3. Results and conclusion

The final model features a multiple linear regression model relating the height to weight, gender, biacromial, pelvic breadth and waist girth with a model specification as follows with coefficient estimates displayed in the model summary:

$$Height = \beta 0 + \beta 1 \times Weight + \beta 2 \times Gender + \beta 3 \times Biacromial + \beta 4 \times Pelvic.breadth + \beta 5 \times Waist.girth$$

```
Call:
lm(formula = height ~ weight + gender + biacromial + pelvic.breadth +
    waist.girth, data = body_no_influential)

Residuals:
     Min       1Q   Median       3Q      Max
-10.2482  -3.0704  -0.1087   3.0943  11.3115

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  128.37094    4.80994  26.689  < 2e-16 ***
weight         0.61853    0.04360  14.185  < 2e-16 ***
gender1        7.34341    0.73042  10.054  < 2e-16 ***
biacromial     0.66717    0.12234   5.454 7.97e-08 ***
pelvic.breadth 0.64148    0.11551   5.553 4.68e-08 ***
waist.girth   -0.61614    0.04784 -12.878  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.399 on 473 degrees of freedom
Multiple R-squared:  0.7627,    Adjusted R-squared:  0.7602
F-statistic:   304 on 5 and 473 DF,  p-value: < 2.2e-16

               Df Sum Sq Mean Sq F value    Pr(>F)
weight          1  20953   20953 1082.84  < 2e-16 ***
gender          1   2995    2995  154.77  < 2e-16 ***
biacromial      1   1790    1790   92.48  < 2e-16 ***
pelvic.breadth  1    469     469   24.24 1.18e-06 ***
waist.girth     1   3209    3209  165.85  < 2e-16 ***
Residuals     473   9153      19
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```




```
> vif(model2)
      weight       gender   biacromial pelvic.breadth   waist.girth
    7.767559     3.294061     3.440653     1.445140     6.328860
```

```
        Jarque-Bera test for normality

data:  residuals(model2)
JB = 3.9697, p-value = 0.126
```

As per results presented above, the final model features decent explanatory power (an adjusted R-square of 76.02%), covariates with good significance in both individual t-tests and hierarchical ANOVA, acceptable level of multicollinearity (VIF values lower than 10 for all covariates) and residuals with acceptable "NICE" property (passed JB test, independence, constant variance and expected value of 0 shown by the Residuals vs Fitted graph). Comparing with the full model and the BIC-selected model, the final model features "enough" explanatory power but much more interpretable results, much less multicollinearity and better significance in both individual t-tests and hierarchical ANOVA.

In the final model, "weight", "biacromial" and "pelvic.breadth" are positively correlated with "height" while "waist.girth" is negatively correlated. For the categorical factor variable "gender", the level "male" shows a positive mean effect over the reference group "female".

Interpreting the coefficients, the final model provides properly interpretable and intuitive results:

- Other variables being held constant, if an individual is 1 kg heavier in weight than others, he is then, on average about 0.62 cm taller in height. (consistent with the intuition people who are heavier are taller)
- Other variables being held constant, if an individual is 1 cm longer in biacromial diameter, he is on average about 0.67 cm taller in height. (consistent with the intuition people with larger skeletons are taller)
- Other variables being held constant, if an individual is 1 cm wider in pelvic breadth, he is on average about 0.64 cm taller in height. (consistent with the intuition people with larger skeletons are taller)
- Other variables being held constant, if an individual is 1 cm longer in waist girth, he is on average about 0.62 cm shorter in height. (consistent with the intuition people who look stronger and heftier, if of the same weight, are shorter than those who are more slim)

Appendix

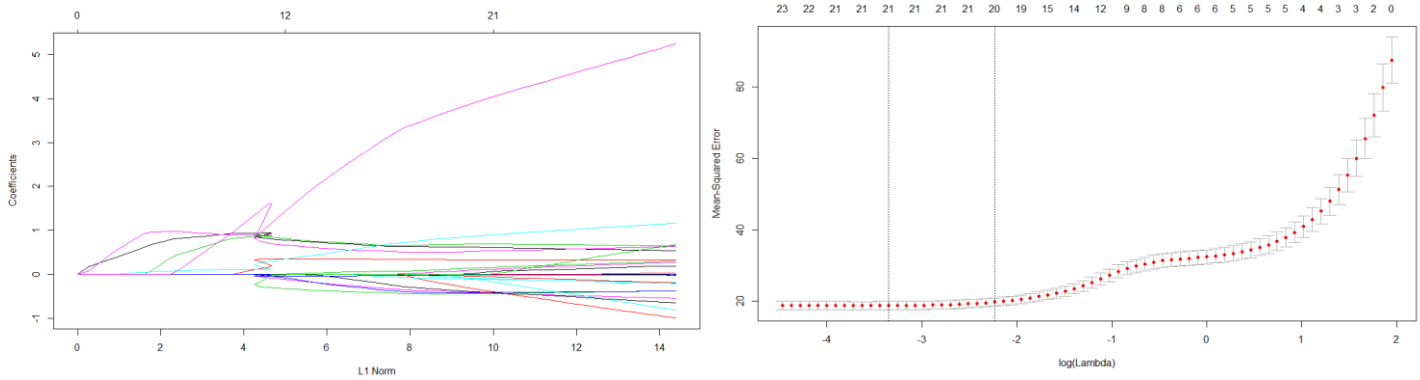
In appendix, two interesting but not appropriate methods of variable selection are included.

1. The LASSO

In the project, LASSO was attempted as the method to do variable selection. However, the approach did little job at removing variables, resulting in deleting only a couple of variables.

The LASSO method puts a constraint on the sum of the absolute values of the model parameters, an upper bound is set for the sum. A shrinking process, where it penalizes the coefficients of the regression variables shrinking some of them to zero, is implemented in order to impose such an upper bound. After the shrinking process, variables that are left with non-zero coefficients are selected to enter the model.

λ, also known as the tuning parameter, dictates the strength of penalty for the shrinking process. When the tuning parameter λ is large enough, the coefficients are shrunk to exactly 0. The larger is the parameter λ, the more number of coefficients are shrunk to zero, the more variables are discarded during the LASSO variable selection.



As displayed by the graph on the left hand side, λ_{min} (line on the left) is the value of the tuning parameter that gives minimum mean cross-validated error and λ_{1se} (line on the right) is the tuning parameter that gives a model such that the MSE is within one standard error of the minimum. λ_{min} returns a model with 21 variables while λ_{1se} returns one with 20 variables.

Therefore, the LASSO approach is not an opportune method for variable selection in this scenario. It only becomes a valid option when the number of covariates p is very large and the best-subset variable selection approach becomes too computationally intensive.

2. Best-subset variable selection using K-fold cross-validation

Instead of using one of adjusted R-square, CP and BIC as the model comparison criterion, a more rigorous approach, which refers to selecting the best model based on the prediction error computed using k-fold cross-validation.

The k-fold cross-validation consists of first dividing the data into k subsets, with k usually set to 5 or 10. Each subset serves successively as test data set and the complementary as training data. The average cross-validation error is computed as the model prediction error.

This method in the project returns a model with 18 predictor variables, which may cause more difficulties in further selection.