Università degli studi di Milano
MSc: Data Science for Economics

Statistical Learning Module

Body Fat Prediction

Giordano Vitale
Matriculation Number: 14310A

submitted to
Prof. Dr. Silvia Salini

August 8, 2023

Abstract

This report presents the methods and the analysis developed on body fat data for 252 men. The data were generously supplied by Dr. A. Garth Fisher who gave permission to freely distribute the data and use it for non-commercial purposes.

The goal is to investigate the relationships between body characteristics and the body fat score of a man. Furthermore, the goal is to investigate which statistical models yield more effective results in predicting body fat percentage.

The initial phase of the analysis involves data preprocessing, including handling missing values and scaling the data appropriately. Following that, various supervised and unsupervised learning techniques are applied to gain insights, build predictive models, as well as to explore patterns in the data.

The advantages and disadvantages of applying those methods on the data set are discussed and the performance of the different machine learning algorithms is then compared and analyzed.

Contents

1 Introduction

The prevalence of obesity and its associated health risks have led to an increased interest in understanding body fat composition. Body fat percentage is a crucial indicator of overall health and is influenced by various physiological and lifestyle factors.

1.1 Data set

The data set used for this project was originally released by a research journal called "Medicine & Science in Sports & Exercise" [1] from a study conducted by Penrose, K. W.; Nelson, A. G.; Fisher A. G. and is available for download at the well-known website for data sets Kaggle.com [2].

The data set contains 252 observations and each observation represents one man. The variables contained in the data set are shown in the table below:

Name	**Description**	**Type**	**Measurement**
Density	Density determined from underwater weighing	Numeric	0.995,..., 1.109
BodyFat	Percent body fat	Numeric	0,..., 47.5
Age	Age of the person	Integer	22,..., 81
Weight	Weight in lbs	Numeric	lbs
Height	Height in inches	Numeric	inches
Neck	Neck circumference	Numeric	*cm*
Chest	Chest circumference	Numeric	*cm*
Abdomen	Abdomen circumference	Numeric	*cm*
Hip	Hip circumference	Numeric	*cm*
Thigh	Thigh circumference	Numeric	*cm*
Knee	Knee circumference	Numeric	*cm*
Ankle	Ankle circumference	Numeric	cm
Biceps	Biceps circumference	Numeric	cm
Forearm	Forearm circumference	Numeric	cm
Wrist	Wrist circumference	Numeric	cm

As one can see, all the variables are numeric or integers. This data set doesn't have categorical variables. This characteristic makes it perfectly suitable for regression analysis.

1.2 Problem description

The goal of this report is to drive the reader through the methodological approach and all the models adopted, with their relative advantages and drawback, highlighting the trade-offs of each choice.

Furthermore, the goal is to inspect how *BodyFat* for a man may be affected by the other variables in the data set. This goal is hopefully reached by implementing and discussing various statistical modeling methods we studied in class to predict the target variable *BodyFat* using the variables about the measurements of the body, quantifying and interpreting the individual effects on the *BodyFat* and providing possible explanations and intuitions to the findings.

Moreover, a final evaluation of the models will be performed in order to assess the ability of each one of them in predicting new, unseen data.

2 Data Preparation

Before beginning with the application of any unsupervised or supervised statistical learning method the data has to be cleaned, prepared and if necessary transformed.

Multiple steps of preparation are taken to enhance the usability of the data for further analysis and modeling.

2.1 Data preprocessing

Data Classes

The variables data classes were already up to date. For this reason, no change was applied to the data classes.

Missing Values

The data set has been checked for missing values, infinite values, and duplicates but none were found.

Removing Density

The variable *Density* is a variable that is determined from underwater weighing. It is often used as an alternative way to estimate body fat, hence it was removed because it was less easy to interpret than *BodyFat* and the goal is to focus on one dependent variable only.

Even if this variable is completely removed, the goals described in the previous chapter don't suffer any change.

Transformation

As described in the table describing the variables, all body measurements were in *cm*, except for and *Height*. In order to use the same unit of measure, *Height* has been transformed in *cm* according to the formula:

$$\text{HeightCm} = \text{HeightInch} \times 2.54$$

As regards *Weight*, it is preferable to use the *kilograms* unit of measure instead of *lbs*. It was then transformed as follows:

$$\text{WeightKg} = \text{WeightLbs} \times 0.45359237$$

2.2 Exploratory Data Analysis and Descriptive Statistics

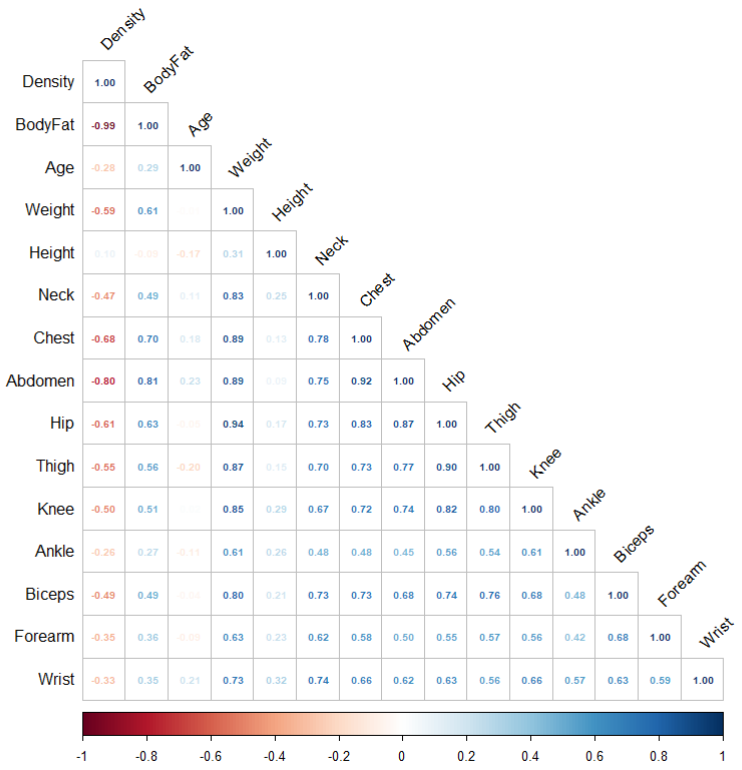
Now that the data set has been cleaned, it is ready for usage. The first step is to start with a descriptive analysis of the variables.

2.2.1 Correlation Analysis

An analysis of the correlation between the numeric variables allows us to explore potential relationships between the variables. The *Abdomen (0.81)*, *Chest (0.70)* and *Hip (0.62)* seem to have the strongest relationship with *BodyFat*.

The visualization of the correlation matrix and the magnitudes it displays allow us to notice that *Abdomen* and *Chest* are highly correlated (0.92), as well as *Thigh* and *Hip* (0.89): this information will be addressed later in the feature engineering section.

Figure 2.1: Correlation between the variables



The correlation matrix also gives an indication of whether the variables have a positive or a negative effect on body fat. Not surprisingly, all the features are positively correlated with *BodyFat*.

2.2.2 Data visualization

Here only one variable *BodyFat* will be plotted, analyzed and discussed, but a complete visualization and disquisition of all the distributions is available at the cited GitHub repository [3] at the end of this report.

Summary statistics - BodyFat

The summary statistics is helpful to have a first glance at implausible values for the variable checked, *BodyFat* in this case. In fact, a value higher than 100 would have been incompatible, as well as negative values. Fortunately, there wasn't any unlikely value.

Since the mean is only slightly different from the median, we can assume that this variable is very close to being normally distributed.

The same summary statistics were computed for all the variables in the data set and none of them showed concerning values.

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
0	12.47	19.20	19.15	25.30	47.5

Distribution - BodyFat

By plotting the histogram with the kernel density line, we can visualize whether the variable looks like a normal distribution or not. It's possible to see that it is slightly positively skewed, meaning that the majority of the data is concentrated on the left side of the distribution, while there are few extreme values on the right side.

With the aim to carry out a more detailed analysis of the distribution, the Q-Q plot has been produced. Its interpretation is the following: the more the points lie on the diagonal line, the more the data follow the assumed distribution, in this case the normal one.

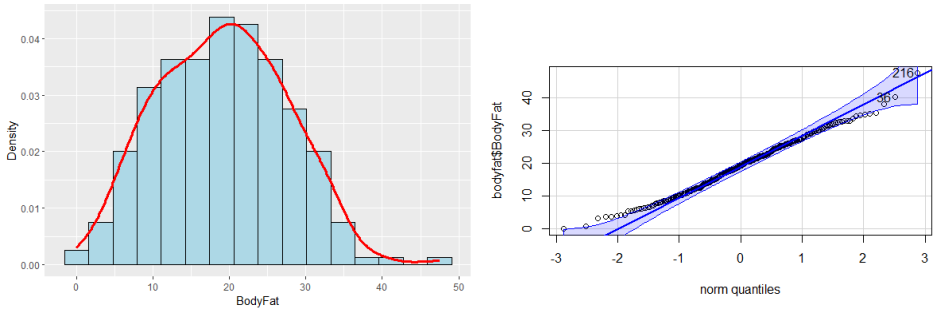
Furthermore, skewness has been computed for *BodyFat*, and its value, 0.14, suggests that the skewness is not significant, since it's close to zero.

Another element that provides support to the assumption of normality distribution is the *Shapiro-Wilk Test*. In fact, this statistical technique is based on the null hypothesis that the data follows a normal distribution, while the alternative hypothesis supports the opposite. After choosing a significance level, commonly $\alpha = 0.05$, if the *p-value* obtained from the test is higher than the α value, then the null hypothesis cannot be rejected. The Shapiro-Wilk test applied to *BodyFat* produces a p-value of 0.16, making it possible to reject the null hypothesis.

On the basis of the visualizations and tests described above, we can conclude that the variable's distribution is approximately symmetrical, with a minor tendency for slightly more extreme values on the right side.

As a matter of completeness, it has to be said that for most statistical modeling methods, it is not strictly needed that the dependent or independent variables are normally distributed. However, it is desirable to deal with non-skewed variables, because skewed ones produce non-normal residuals that compromise the accuracy of the inference of the models' coefficients.

Figure 2.2: Visualization of the distribution



(a) Histogram with density line

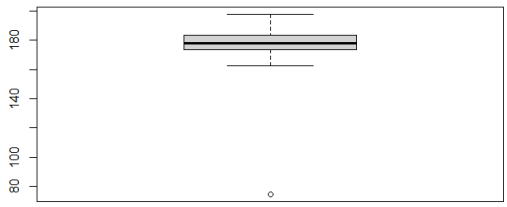
(b) Q-Q plot with 95% C.I.

Box plot - Height

Similarly to summary statistics, *box plot* visualization provides concise insights about the distribution of the data and the potential presence of outliers.

Differently from the previous paragraphs, here the graphical analysis will be developed using the variable *Height*. The motivation for this choice is that precisely this variable had an outlier that needed to be addressed. This allows us to show the strategy implemented for outliers management.

Figure 2.3: Box plot - Height



As visible in Figure 2.3, there is one observation that falls way distant from the *whiskers* of the box plot. To identify it on the data set, a query has been produced: the result of it was an individual with *Height* = *73cm*. Even if this observation may be

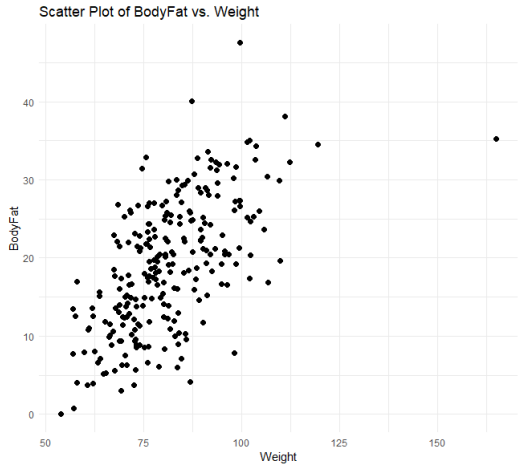
possible in real-world life, it was decided to remove it because of the potential misleading impact this outlier could have on the estimation of coefficients and performance of the models on unseen data.

Scatter plot - BodyFat vs Weight

With the goal to graphically inspect the relationship between the target variable and each independent variable i separately, a scatter plot for each one of all the possible combinations between *BodyFat* and i has been produced.

Even if this might seem a basic task, it still has a valuable application for understanding the relationships between variables and giving intuition about modeling decisions to be made in the next chapters. Here's an example:

Figure 2.4: Scatter plot - BodyFat vs Weight



2.3 Feature engineering

As discussed in the Section 2.2.1, highly correlated variables may be a potential source for *multicollinearity*. Loosely speaking, multicollinearity arises when two or more predictor variables have a strong linear relationship.

This phenomenon can be troublesome because, in its presence, the estimated coefficients of a regression model may not be correctly determined, undermining the statistical efficacy and power of the model. In fact, the model may struggle in disentangle the individual effects of each predictor.

A way to assess this situation is to combine two or more correlated variables into one that summarizes the information of both. The motivation for the following manipulation of the features is strengthened by the fact that, after building a simple model including all the variables (before the transformations), the *VIF (Variance Inflation Factor)* values

of that model were tremendously high for those exact features that will be manipulated in the incoming lines.

Abdomen & Chest

The *Abdomen* and *Chest* variables showed a strong correlation, and since they are both elements of the same area of the body, it has been thought that merging them would be a reasonable approach. The new variable, whose unit of measure is *cm*, is called *ACratio* and it is obtained as follows:

$$\text{ACratio} = \frac{Abdomen}{Chest}$$

Hip & Thigh

As for the above-mentioned case, *Hip* and *Thigh* showed a strong correlation, and since they are both elements of the same area of the body, the same strategy of the previous case was applied. The new variable, whose unit of measure is *cm*, is called *HTratio* and it is obtained as follows:

$$\text{HTratio} = \frac{Hip}{Thigh}$$

After the manipulation

As a consequence of the above manipulations, the variables *Abdomen, Chest, Hip, Thigh* have been removed from the data set.

Applying the *VIF* measure to a new simple model including all the variables in the *transformed* data set, a more encouraging output is obtained. Indeed, now only one variable shows a VIF value higher than 10, which is a usually adopted threshold associated with potential concerns about multicollinearity. Since the problems of multicollinearity have been noticeably limited, but not completely extinguished, it will be a task of model selection - both in supervised and unsupervised techniques - to restrict the number of variables.

3 Unsupervised Learning

This chapter covers topics related to unsupervised machine learning techniques, namely *Principal Component Analysis (PCA)* and *clustering*. Unlike its counterpart, supervised learning, where algorithms are trained on labeled data to make predictions, unsupervised learning operates on unlabeled data, aiming to uncover hidden structures and relationships within the data itself. This means that, in order to perform the above-mentioned techniques, the target variable *BodyFat* has to be omitted, and the explanatory variables only will be part of the analysis.

Throughout this report, popular unsupervised learning methods are employed, such as dimensionality reduction techniques and clustering algorithms.

3.1 Principal Component Analysis

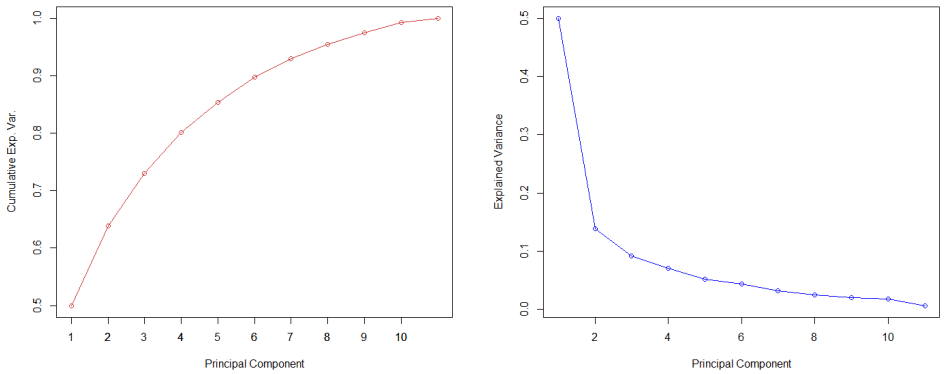
The Principal Component Analysis is a technique whose aim is dimensionality reduction. This tool is particularly useful when dealing with large data sets having a high number of features. The main goal of PCA is then to simplify the input data into a lower-dimension space whose directions are the principal components, i.e. linear combinations of the original features designed to be uncorrelated with each other. The uncorrelated nature of principal components helps in reducing the dimensionality of the data while preserving as much variance as possible and also enables a graphical representation of the data in a more comprehensible space.

The data was scaled, centered and the principal components have been computed using singular vector decomposition.

As can be seen in Figure 3.1, the dimensionality reduction implies a significant loss of information. In fact, the lower the number of components considered, the lower the variability explained: this highlights a *trade-off* between the higher interpretability with the lower number of components and the associated lower information explained

Nonetheless, the first 6 components explain almost 90% of the overall variance in the data set, meaning that PCA is not performing poorly. That said, 6 dimensions are still difficult to imagine, for this reason it is common to exploit the first two or three components. Focusing on the first two, we can say that they explain 63% of the overall variance, which is improvable but not extremely bad. A model based on these two principal components will be computed in order to assess their performance on new, unseen data.

Figure 3.1: Cumulative and individual variance explained by the components



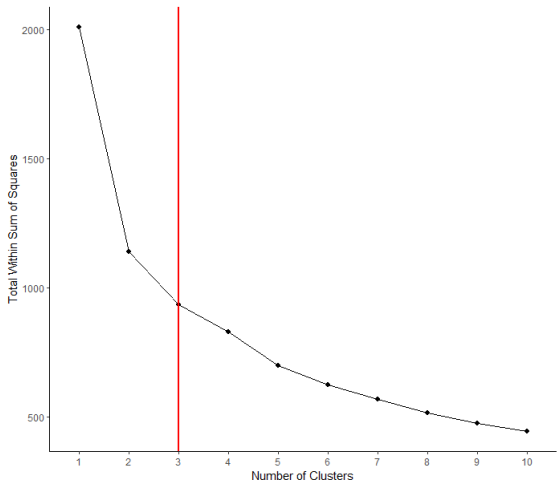
(a) Cumulative (b) Individual

3.2 Clustering

An important implementation of the principal components is to use them as bases to partition the data into clusters, i.e. groups of points that share similar properties and characteristics. To perform clustering analysis, the *K-Means* algorithm will be exploited.

Since there isn't in advance an exact and intuitive number of clusters to display that works for every data set, it is needed to do an investigation of the possible optimal values of clusters to plot. Moreover, in this data set there are no predetermined categories to group the data into, as the target variable is continuous. One way to do it is to visualize the reduction of the *total within sum of squares (WSS)* associated with each number of clusters.

Figure 3.2: Optimal number of clusters - K-Means



Moving from 1 to 2 clusters significantly reduces the within sum of squares, as well as moving from 2 to 3. When reaching 4 clusters or above the *WSS* does not decrease as significantly as needed in order to sustain one additional cluster. Therefore, clustering the data into 3 groups appears to be the most reasonable choice.

The graphical result of the clustering technique based on the optimal number of clusters is represented in the following figure.

Figure 3.3: 3-Means Clustering with first 2 Principal Components

Figure 3.3 shows that the yellow left-side cluster is the most extended one since it includes one extreme value on the left. The other two clusters, light blue and red, seem more defined and less scattered.

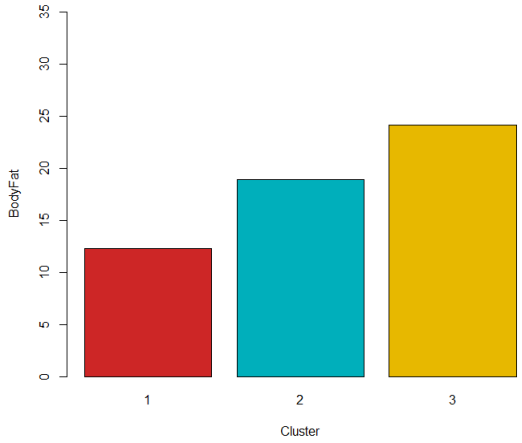
The clear partition of the data provided by the clustering analysis looks encouraging for further analysis. For instance, it can be a good exercise to derive the average *BodyFat* for each cluster. To do so, it is necessary to add a column to the original data frame *bodyfat* containing the classification into the associated cluster for each observation. The result is summarized by the following image

In the above bar plot, the colors are not randomly assigned to each bar: they recall the colors from the scatter plot of Figure 3.3. It's straightforward to note that we can identify three clear classes: one for low values of *BodyFat*, one for regular *BodyFat*, and the last one for high values of *BodyFat*.

Another powerful visualization of the characteristics and properties shared by each cluster is the parallel coordinates plot, where each data point is represented by a line connecting multiple vertical axes, with each axis representing a different variable. The most important advantage of this plot is that it enables us to identify trends among the clusters.

Inspecting Figure 3.5 it is possible to note that cluster *C1* - the one with the lowest average *BodyFat* - has, on average, lower values for each one of the variables. In fact,

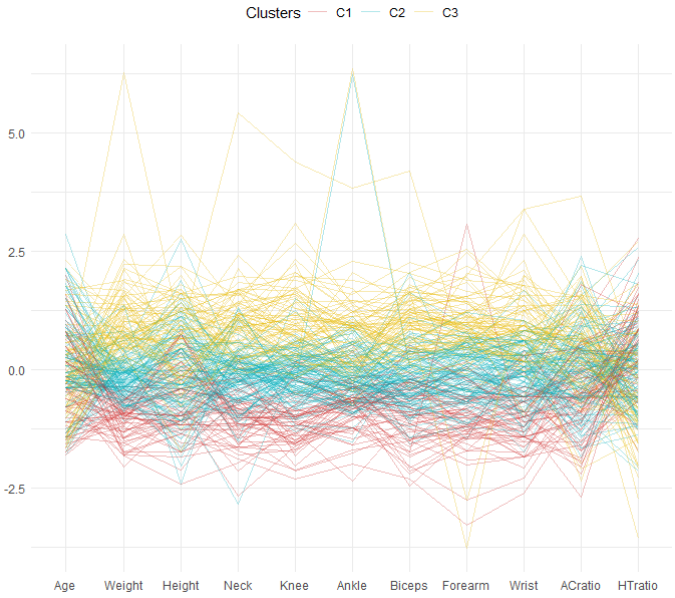
Figure 3.4: Average BodyFat for each cluster



the red lines in the image are under the blue and yellow lines representing the other two clusters. Similarly, we can see that cluster *C3* - the one with the highest average *BodyFat* - has, on average, higher values for each variable, as it is physically above the other two clusters.

The only exception to this pattern is represented by *Age*, which seems to be irrelevant in clustering the groups, as well as *HTratio*. In fact, in both cases, the lines' colors are not as precisely disentangled as for the other variables.

Figure 3.5: Parallel coordinates for the 3-means clusters



The main takeaway from this analysis is that individuals with high *BodyFat* are associated with high values of all the body measurements, except for the above-mentioned two variables. As a consequence, men with lower values of body measurements are associated with lower values of *BodyFat*. Even if this result may look pretty intuitive and trivial, establishing it with a thorough approach makes it more reliable and justified.

4 Supervised Learning

This chapter will drive the reader through all the supervised learning techniques applied to the data set. Unlike unsupervised learning, we do not deal with the whole number of observations: the cleaned data set, including the target variable, will be *randomly* split into two parts, namely the *training set* and the *test set*.

The first one will serve as the foundation for training the various models, while the test set will be vital in assessing how well each model fits the data and its ability to make accurate predictions on unseen data. The training set will consist of 75% of the data, while the test set will contain the remaining 25%.

A variety of models will be deployed - both parametric and non-parametric techniques - with the aim to answer the research questions introduced in the abstract.

As regards the tools used for the evaluation of the models, the measure used for assessing the goodness of the predictions will be the *MSE*, while for assessing the goodness of fit will be the *BIC* as well as the $\bar{R}^2$ when possible.

4.1 Linear models

Linear models are a simple and widely used statistical tool in data analysis and machine learning. At its core, a linear model represents a linear relationship between a response variable and one or more predictor variables.

The goal of a linear model is to estimate the coefficients and intercept that best describe the linear relationship between the response and predictors. Linear regression methods have the benefit of simplicity and interpretability.

4.1.1 Principal component regression

One of the possible applications of principal components is to use them as regressors to fit a linear model, and then to evaluate it on unseen data.

Since the previous analysis of PCA (Section 3.1) was performed on the whole data set (except the target variable, of course), it would be problematic if we were using those principal components for prediction: we wouldn't have a set of points with which to assess the goodness of the prediction, and it would raise concerns about *data leakage*. For this reason, a new PCA was performed, this time on the training set only. After computing the principal components, a linear model with the first three of them is fit on 174 observations and then evaluated on the test set.

$$\widehat{BodyFat}_i = \hat{\beta}_0 + \hat{\beta}_1 PC1_i + \hat{\beta}_2 PC2_i + \hat{\beta}_3 PC3_i \tag{4.1}$$

The model outputs an $\bar{R}^2 = 0.57$ which can not be considered a bad result once taken into account that only the first three principal components have been used to train the model. All three coefficients appear to be statistically significant. The BIC of this model is 1253.29, while the MSE on the test set data is 117.70.

A drawback of this model is that, since it is built exploiting the principal components, it's not equipped with intuitive interpretability of the coefficients' estimates.

In summary, the model appears to be explaining a substantial amount of the target's variability with just three principal components, but it may still have room for improvement.

Figure 4.1: PC Regression summary

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.0929     0.4114  46.410  < 2e-16 ***
pc1_train     2.0363     0.1797  11.334  < 2e-16 ***
pc2_train    -3.0771     0.3319  -9.270  < 2e-16 ***
pc3_train     2.4676     0.4044   6.102 6.37e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.55 on 178 degrees of freedom
Multiple R-squared:  0.5857,    Adjusted R-squared:  0.5787
F-statistic: 83.88 on 3 and 178 DF,  p-value: < 2.2e-16
```

4.1.2 Multiple Linear Regression

The following fitted model exploits all possible variables as regressors to explain the target variable *BodyFat*:

$$\begin{aligned}\widehat{BodyFat}_i =& \hat{\beta}_0 + \hat{\beta}_1 Age_i + \hat{\beta}_2 Weight_i + \hat{\beta}_3 Height_i + \hat{\beta}_4 Neck_i + \\ & \hat{\beta}_5 Knee_i + \hat{\beta}_6 Ankle_i + \hat{\beta}_7 Biceps_i + \hat{\beta}_8 Forearm_i + \\ & \hat{\beta}_9 Wrist_i + \hat{\beta}_{10} ACratio_i + \hat{\beta}_{11} HTratio_i\end{aligned} \tag{4.2}$$

The model outputs an $\bar{R}^2 = 0.67$ which considerably improves the analog of the previous model. This doesn't surprise us, since a higher number of features - namely, all of them - has been included in the model. According to this model, *Age, Weight and Height*, as well as *Wrist and ACratio*, are all statistically significant at a confidence level of $\alpha = 0.001$, while *Forearm* at a level $\alpha = 0.01$

From this model, we can highlight two curious findings. An interesting and counter-intuitive observation we can draw from this summary is that, focusing on the significant

Figure 4.2: Linear Regression Output with all Variables

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 31.48420   23.50237   1.340 0.181995
Age          0.16983    0.03526   4.816 3.02e-06 ***
Weight       0.49116    0.08198   5.991 1.05e-08 ***
Height      -0.30892    0.06607  -4.676 5.60e-06 ***
Neck        -0.31268    0.28774  -1.087 0.278579
Knee        -0.06704    0.29833  -0.225 0.822439
Ankle       -0.13930    0.30695  -0.454 0.650483
Biceps      -0.12191    0.20921  -0.583 0.560792
Forearm      0.69827    0.23155   3.016 0.002921 **
Wrist       -2.47232    0.64778  -3.817 0.000184 ***
ACratio     59.24668    9.60533   6.168 4.18e-09 ***
HTratio     -7.25599    6.27994  -1.155 0.249391
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.739 on 187 degrees of freedom
Multiple R-squared:  0.6951,    Adjusted R-squared:  0.6771
F-statistic: 38.75 on 11 and 187 DF,  p-value: < 2.2e-16
```

coefficients only, the estimated coefficient associated with *Wrist* is negative: this would imply that a higher Wrist circumference is related to a lower BodyFat. More precisely, one additional centimeter in Wrist circumference would entail a decrease of 2.47 in the BodyFat percentage. This finding seems to be inaccurate and incompatible with a reasonable intuition of the relationship between BodyFat and Wrist.

The second remark can be done with respect to *Height*: according to its estimated coefficient, higher Height is related to lower BodyFat. A possible interpretation is that, for a given amount of BodyFat, the higher the man, the more the BodyFat is distributed along the body, producing a lower value of BodyFat. Even if this finding may be supported by reasonable intuition, it will deserve higher attention in the next models.

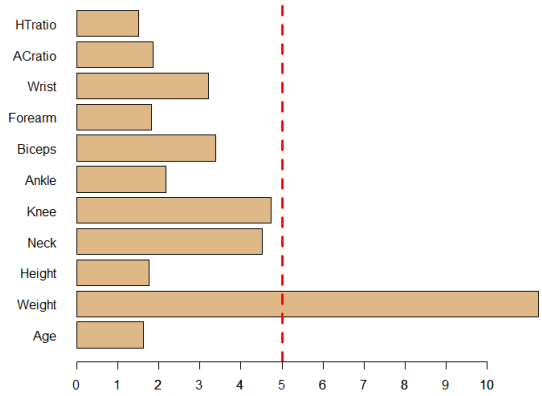
One possible explanation for this contradicting finding is that the model may suffer from over-specification, i.e. the regression equation (4.1) may contain redundant regressors that produce an inaccurate estimate of the true coefficient.

An additional explanation of the counter-intuitive coefficients is that the model may suffer from multicollinearity. In its presence, in fact, coefficients' estimates might have unexpected signs, undermining the model's reliability.

In order to assess the problem of multicollinearity, the VIF can be computed. This measure will provide information about the degree of multicollinearity in the current data set.

Figure 4.3 shows that, as a matter of fact, this model suffers from multicollinearity. The VIF value associated with Weight is above the threshold depicted by the red dashed vertical line. In fact, according to a rule of thumb, every VIF value which exceeds 5 or 10 should raise concerns about multicollinearity. Here the value of Weight not only exceeds 5, but it exceeds 10, highlighting an evident issue that requires adequate attention. The next model will introduce a way to fix this issue.

Figure 4.3: VIF for complete linear model



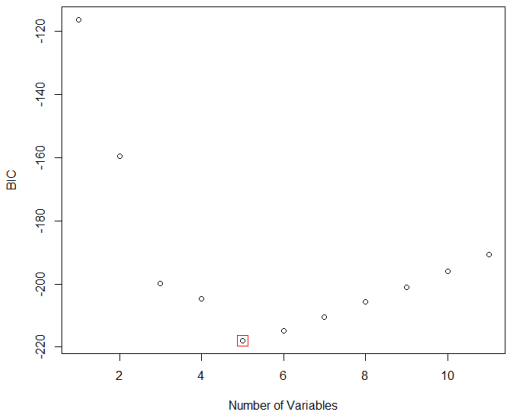
4.1.3 Best Subset Selection for linear model

This approach consists of identifying a subset of the predictors and using this smaller group, namely the most relevant ones, to fit a linear model.

One advantage of this procedure is the interpretability of the model, since the estimated relationship is still linear. Moreover, it also has the benefit of removing redundant variables that may cause excessive noise and thus overfitting.

Subset selection helps to identify a more interpretable and simpler model while preserving the essential relationships between the predictors and the response variable. In summary, it allows us to find a balance between prediction accuracy and model simplicity.

Figure 4.4: Best Subset Selection with BIC



One way to carry out this procedure may be to manually remove the regressors for which the previous model has yielded problematic or counter-intuitive results. But the most complete and meticulous approach is the *best subset selection* procedure.

To perform *best subset selection*, we fit a separate least squares regression for each possible combination of the predictors: they will be eventually compared using the *BIC* evaluation metric. Since Model 4.2 included 11 variables, there are $2^{11} = 2048$ possible combinations.

The obtained result suggests that the optimal set of predictors is formed by only 5 variables (see Figure 4.4). The resulting model is plotted in the following figure:

Figure 4.5: Subset Selection Regression Output

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.65051   14.84506   1.391    0.166
Age          0.13730    0.02921   4.701 4.92e-06 ***
Weight       0.55140    0.04894  11.266  < 2e-16 ***
Height      -0.37061    0.05913  -6.267 2.35e-09 ***
Wrist       -2.36268    0.54496  -4.336 2.34e-05 ***
ACratio     62.08420    8.49221   7.311 6.88e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.428 on 193 degrees of freedom
Multiple R-squared:  0.7148,    Adjusted R-squared:  0.7074
F-statistic: 96.74 on 5 and 193 DF,  p-value: < 2.2e-16
```

The regression output shows that now all variables are highly significant.

One satisfactory result is that now the *VIF* values computed on this model are all below the threshold of 5, meaning that the multicollinearity problem has been addressed correctly.

4.2 Moving beyond linearity

So far we focused our analysis on linear models. As we said, they are relatively simple to implement, explain and interpret. However, standard linear regression models can have significant limitations in terms of prediction accuracy.

In this section, a variety of models will be applied and discussed, where each model will relax the assumption of linearity while still trying to keep as much interpretability as possible.

4.2.1 Polynomial model

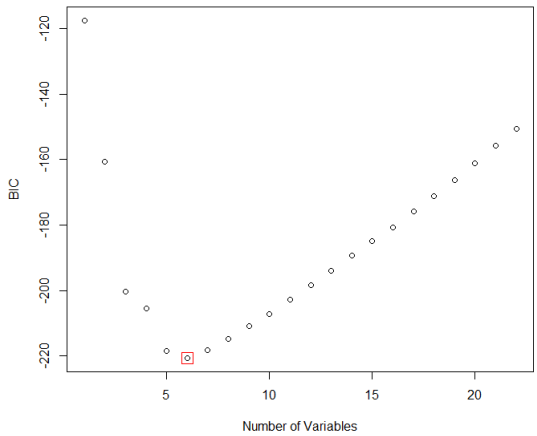
Polynomial regression extends the structure of the linear model by adding extra predictors that are obtained by raising each of the original ones to a power. For example, a quadratic regression uses the original variable, X, and its squared version, X^2.

The degree of the polynomial determines the complexity of the model and the flexibility in capturing nonlinear patterns. The advantages of polynomial regression include its ability to capture complex relationships and its flexibility in modeling a wide range of data patterns. It enables us to uncover hidden patterns that may not be apparent in linear models and provides more accurate predictions for nonlinear data.

In summary, there is a trade-off between model complexity - with the resulting higher fitting ability - and prediction accuracy.

After a careful evaluation, it has been decided to build a polynomial model of degree 2, where the output variable *BodyFat* is explained by all the regressors and the associated squared versions of each one of them. Hence, this "saturated" model has $2 \times 11 = 22$ predictors.

Figure 4.6: Best subset selection with BIC



The "saturated" model was not intended to be the final version of the polynomial

model, since it includes a lot of variables and some of them may be redundant and correlated, as thoroughly discussed previously. The main idea was to introduce the saturated model as specified above and then to run the *best subset selection* algorithm to it, in order to make the non-linear characteristics of the model coexistent with an easier dimension.

Even if the procedure may be computationally intensive due to the high number of possible combinations, the output is quite gratifying. In fact, now the subset selection procedure keeps 6 variables (see Figure 4.6), where two of them are with power 2, capturing the aimed non-linear relationship between regressor and regressand.

Now all the estimated coefficients are extremely significant, meaning that this model appears to be more reliable with respect to the previous models. This implies that the assumption of a non-linear relation in the data is confirmed.

Figure 4.7: Best subset selection for polynomial model

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  28.541355  14.495353   1.969   0.0504 .
Age           0.133916   0.028710   4.664 5.79e-06 ***
Weight        1.265401   0.267550   4.730 4.35e-06 ***
I(Weight^2)  -0.004175   0.001534  -2.721   0.0071 **
Height       -0.401252   0.059426  -6.752 1.68e-10 ***
Wrist        -2.499917   0.538729  -4.640 6.43e-06 ***
I(ACratio^2) 32.469632   4.620428   7.027 3.57e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.351 on 192 degrees of freedom
Multiple R-squared:  0.7261,    Adjusted R-squared:  0.7175
F-statistic: 84.81 on 6 and 192 DF,  p-value: < 2.2e-16
```

There appears to be a quadratic effect for *Weight*, meaning that the predicted BodyFat first increases more than proportionally when Weight increases, but then it increases less than proportionally.

As regards the goodness of fit, the subset polynomial model exhibits a positive $\bar{R}^2 = 0.72$ and an even better result for the $BIC = 1185.17$.

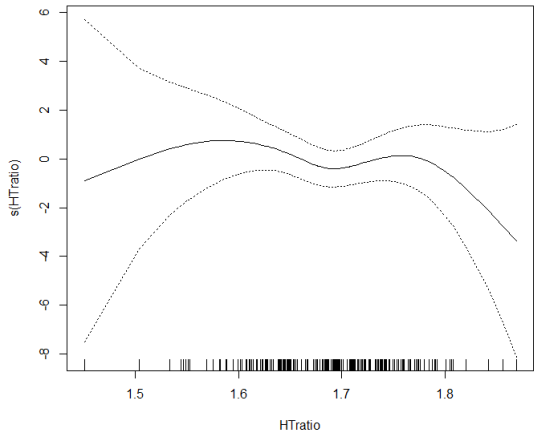
4.2.2 GAM

In GAMs, the relationship between the response variable and predictors is modeled through smooth functions which allow us to capture complex and intricate patterns that linear models might miss. Since it is still an additive model, this approach maintains the possibility of analyzing the individual effect of the single variable under attention. Moreover, it automatically models non-linear fits, without the need of manually trying out different transformations.

Figure 4.8 shows an example of the estimated fitting line for HTratio. We can see that the estimated relationship in the data is quite flexible and thus GAM may positively contribute to capturing the true pattern.

The way this model is built is very similar to the polynomial model, thus we may expect that during the prediction tasks they will perform similarly. This assumption will be discussed and tested later.

Figure 4.8: Example of GAM fitted line - HTratio

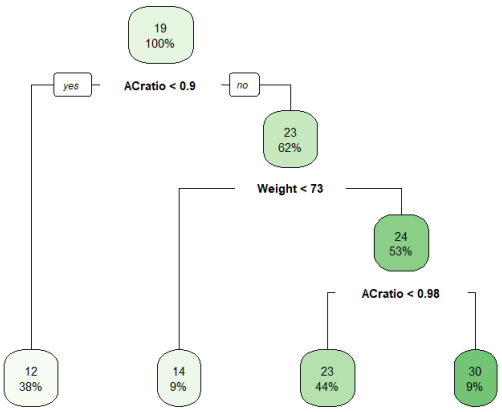


4.2.3 Regression Tree

Regression trees divide the feature space into regions and assign a constant value (usually the average of the target values) to each region, making them suitable for problems involving non-linear relationships between predictors and the target.

The primary goal of a regression tree is to partition the data into subsets that are as homogeneous as possible in terms of the target variable. This partitioning process involves recursively splitting the data into smaller groups based on the values of the input features. At each step, the algorithm selects the feature and the split point that results in the greatest reduction in variability (in this example, using the Cp measure).

Figure 4.9: Pruned tree



The key advantage of tree-based models is the easy interpretability and clear visualization of the splitting procedure. However, to mitigate overfitting, techniques like pruning are applied, which involve simplifying the tree by removing branches that do not provide significant improvement in predictive accuracy.

According to the pruned tree depicted in Figure 4.9, the most important variables in describing the determination of the *BodyFat* appear to be ACratio and Weight. In fact, the first predictor variable at the top of the tree is the most influential in predicting the value of the response variable.

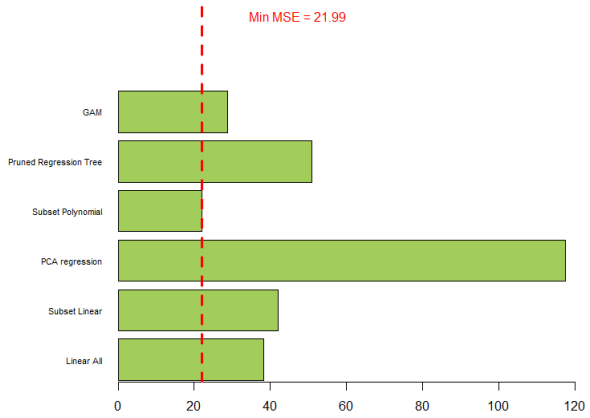
5 Model Evaluation

To evaluate the predictive accuracy of the supervised and unsupervised models described in the previous chapters, their performance is tested on new and unseen test data. We already said that the test set contains 25% of the original data.

Our goal is now to determine which model yields the best prediction. The adopted metric for answering this question is MSE. In terms of goodness of fit, instead, the BIC measure is adopted.

Model	**MSE**	**BIC**
Linear All	38.26	1215.04
Best subset linear	42.01	1187.90
PC regression	117.43	1253.40
Best subset polynomial	21.99	1185.17
Pruned tree	50.80	*NA*
GAM	28.78	1221.79

Figure 5.1: Models Evaluation



Starting from the worst model, we have to highlight the bad performance achieved by *Principal Component regression* in the prediction task, since the MSE is at least twice larger than all the other models. Moreover, it scored the highest BIC, meaning that this model fails in both making accurate predictions and fitting the data. This result is not a surprise because it is built on the first three components only, thus a significant amount of information is cut out.

The *best subset linear model* registers a significant improvement with respect to its counterpart, the *complete linear model*, both for the prediction and the fitting performances. Namely, it is the third-best model of those experimented within this project, overall.

Jumping to the best-performing model in terms of prediction accuracy and fitting accuracy, the *best subset polynomial* model produces an abundantly smaller MSE when compared to all the other models. It also scores the lowest BIC, even if in this case it is not considerably different from the BIC of the best subset linear model.

One last insight that can be grasped is that the second-best performing model is *GAM*. This may suggest that non-linear models are best suited for this data set.

6 Conclusion

The linear regression models yielded decent results in regards to interpretability as well as predictive accuracy. However it is easy to miss underlying multicollinearity, especially if there is a large training set to fit the model. This can lead to completely wrong estimations, as in the case of the *temperature* variable getting a negative coefficient, even though there was statistical significance. In this case the best subset selection was able to remove the problematic predictors but that does not always have to be the case.

The shrinkage methods appear to be more reliable methods to tackle the problems of multicollinearity, especially when opting for a high shrinkage penalty. The down side is that almost all interpretability gets lost, compared to the standard linear regression, because of the standardisation of the variables. Only the direction of the effects, positive or negative, of the variables can be observed.

Performance-wise the GAM yielded the best results, as it is the only modeling technique among the ones discussed, that takes possible non-linear effects into consideration. The down side is again, that there is little interpretability of the smoothed variables, except for the possible graphical analysis.

Bibliography

[1] Generalized Body Composition Prediction Equation for Men Using Simple Measurement Techniques, 1985. URL http://dx.doi.org/10.1249/00005768-198504000-00037.

[2] Fedesoriano @Kaggle. Body Fat Dataser, 2022. URL https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset?datasetId=1408058&sortBy=voteCount.

[3] Giordano Vitale. Github - BodyFatPrediction, 2023. URL https://github.com/giordanovitale/BodyFatPrediction.