An investigation into the prediction of complex movements in relation to sports

Blake VanDuyvenvoorde

Bachelor's thesis
2023:K6

Lund University

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

Centrum Scientiarum Mathematicarum

Bachelor's Theses in Mathematical Sciences 2023:K6
ISSN 1654-6229

LUNFMS-4071-2023

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/

Acknowledgments

I would firstly like to thank my supervisor, Linda Hartman, for her continued guidance, interest and dedication over the course of the study. She provided constant support throughout the process and kept me moving in the right direction. I would also like to thank my friends in Output Sports Limited for providing me with the technology needed to complete this investigation, and my brother, Conor, for inspiring the idea. Finally I wish to thank all the players and teammates who participated in this study, and the coaches for allowing me to take time out of their trainings.

Summary

The influence of science within sports has caused high performance athletes to break through the bounds of older records across multiple disciplines whether it be from technique or technology. Now that athlete testing has been made accessible and portable, we can quantify and collect information regarding a wide range of physical characteristics. I have investigated whether this data, collected firsthand on multiple areas of physical performance from power output to flexibility, can be used to predict the results of a more complex movement.

In this case, Olympic Handball players were used as subjects in order to uncover a prediction model that could accurately predict the throwing velocity of the players. Many methods were used in order to identify this model and, subsequently, which exercise or movement that had been tested was the most influential in the throwing movement. Out of 13 different exercises, only 4 of them were contained in the final model: Squat Jump, Left Armed Internal Shoulder Rotation, Right Armed Internal Shoulder Rotation and Underhand Medicine Ball Squat Throw.

Once the strongest model had been identified using certain comparative values, new and previously unseen data was supplied to the model to test it predictive capabilities. This test resulted in a very accurate prediction of 86.36 km/h which was only 2.5% higher than the observed value average of 84.25 km/h for this new subject.

Overall, a bigger testing pool accompanied with a larger number of exercises to be tested on would greatly increase the accuracy and usability of these results. Though the model provided seemingly accurate predictions, it did so only on one new subject where more would be required for significant findings. If this study were to be improved in such areas, the results could help to innovate the way coaches and trainers implement their training workout programs, improving complex yet integral elements of whatever sport they happen to be a part of.

Abstract

In this project, we attempt to predict a complex movement in space using smaller scale exercises relating to the movement in question. Specifically, to try and predict the speed a handball player can throw the ball by using data collected on the physical abilities of each player.

The predictor data was collected using a sensor produced by Output Sports Limited (ref.1) and the throwing speed was measured with a velocity gun. It was then cleaned and processed before a correlation analysis was run and the model building process began. Some high correlations were found (not unexpectedly) between the "Jump" variables and relationships were confirmed after the Principle Component Analysis.

Table 3 shows the results of the model analysis that led us to include the variables provided by the Stepwise Regression in our Final Model as it provided the best results as well as being one of the more simple models. The Stepwise model showed few issues after a residual analysis was performed and provided quite an accurate prediction for previously unseen data, giving a prediction error of only 2.5% higher than the observed value.

In conclusion, there were some noticeable shortcoming of this study such as the small data set and a narrow range of exercises to test the athletes on. Improvements in these areas would lead to much more accurate and beneficial results. In saying this, the investigation rewarded us with a method and model that provided accurate predictions (albeit on a single test subject, but very accurate nonetheless) and an area of study with a lot of potential to help improve elements of athletes' training.

Description of Variables

Exercise Name	Description	Metric	Class
Bilateral CMJ	Two-footed jump with arms on hips.	Jump Height (cm)	Jump
Squat Jump	Similar to Bi CMJ but starting from squat position.	Jump Height (cm)	Jump
10-5	10 consecutive jumps with legs straight and hands on hips.	Reactive Strength (RSI)	Jump
Unilateral CMJ	One-footed jump with arms on hips. Performed on both legs (L & R)	Jump Height (cm)	Jump
Downward Med Ball Slam	Slam medicine ball downward from above head.	Peak Velocity (m/s)	Power
Forward Underhand Med Ball Squat Throw	Similar to Kettlebell Swing, throw medicine ball forward while pressing up from squat position.	Peak Velocity (m/s)	Power
External Shoulder Rotation	Elbow at shoulder height, arm at 90° and starting from flat, rotate upwards and back. Performed on both arms (L & R).	Range of Motion (∘)	Mobility
Internal Shoulder Rotation	Elbow at shoulder height, arm at 90° and starting from flat, rotate downwards and back. Performed on both arms (L & R).	Range of Motion (∘)	Mobility
20 Meter Sprint	Sprint the 20m after countdown	Duration (s)	Speed
T-Test	Run forward, left, right, left and backwards, tracing out shape of "T"	Duration (s)	Agility
Throwing	Running step shot towards velocity gun at head height	Speed (km/h)	

Table 1: Descriptions of the exercises used in the data collection.

Table of Contents

Contents

1 Introduction **7**

2 Theory and Method **8**
- 2.1 Definitions . . . 8
- 2.2 Data Collection . . . 8
- 2.3 Data Cleaning & Analysis . . . 9
- 2.4 Models and Model Building . . . 9
  - 2.4.1 Full Model . . . 10
  - 2.4.2 Stepwise Selection . . . 10
  - 2.4.3 Lasso Regression . . . 11
  - 2.4.4 Random Forest . . . 11
  - 2.4.5 Principal Component Analysis (PCA) . . . 12
- 2.5 Residual Analysis . . . 12
  - 2.5.1 Selection Metrics . . . 13
  - 2.5.2 Studentized Residuals . . . 14
  - 2.5.3 Leverage . . . 14
  - 2.5.4 Cook's Distance and DFBETAS . . . 14

3 Results **15**
- 3.1 Exploratory Data Analysis . . . 15
- 3.2 Model Selection . . . 18
  - 3.2.1 Full and Null Model . . . 19
  - 3.2.2 Lasso Model . . . 19
  - 3.2.3 Random Forest Model . . . 20
  - 3.2.4 Stepwise Model . . . 21
  - 3.2.5 Principal Component Regression . . . 22
- 3.3 Final Model Analysis . . . 23
  - 3.3.1 Final Model Testing . . . 28

4 Discussion **28**

5 Conclusion **30**

6 References **32**

7 Appendix **32**
- 7.1 A . . . 32
- 7.2 B . . . 34

1 Introduction

The topic of movement prediction in relation to sport was of interest to the author due to his continued involvement in the sport of handball. The original idea for the project was to be a comparative assessment of the physical capabilities between handball and volleyball players, however due to logistical difficulties the testing on the volleyball athletes had to be stopped. Yet this provided an opportunity to focus on the arguably more interesting and relevant topic of complex movement prediction using data collected from simpler, more spatially contained movements. In this study, the complex movement in question will be the throwing speed of a handball player's step shot.

Output Sports Limited (ref.1) is an Ireland based company that produces a portable and easy to use athlete testing and tracking device (Fig.18 in Appendix B). It measures strength, power and speed giving accurate and reliable results. After some discussions with the company, a conclusion was reached to use a set of ten measurements (described in Table 1 above), each believed to have an intrinsic relationship to the throwing movement being studied.

The study comprised of several different elements, beginning with the data collection and data cleaning process, and then moving on into the data analysis, model building and prediction assessment. The data collection was arduous but not very complex. The cleaning, done in Python, required the data to be filtered for unreasonable results and compiled into a single file, allowing for the analysis on correlations and variable distribution to be done in R after.

Several models were created and tested in the investigation, with variables deemed relevant from several different analyses. Principal Component Analysis, Lasso Regression and Stepwise Regression were employed to identify the significant variables and variable combinations used in the models, who's Residuals were then examined and important comparative metrics extracted for conclusions to be drawn later.

2 Theory and Method

2.1 Definitions

We describe *complex movements* as those such that a higher number of degrees of freedom are required in order to perform said movement. *Degrees of freedom* are defined as the mechanisms that are required in order for you to perform a certain action under specific circumstances and coordinate between different parts of the body and the environment (ref.2). For example, a standing two-footed jump requires the leg muscles to push off from the ground and the coordination between the brain and legs to land stably on the flat ground. The individual muscle groups themselves are also broken down into more specific degrees of freedom. However comparing this with the complexity of the step shot movement, we see that the number of degrees of freedom are much higher. The throwing muscles, coordination of the legs and arms with the need to aim the ball in order to avoid the measuring equipment (and so on) are all needed in this action. Both of these concepts are part of the broader subject of non-linear pedagogy and are a current point of interest in the research of sport-scientists.

2.2 Data Collection

The data collection process began during the spring semester 2022, with the author contacting the coaches and trainers of the three handball teams from the club he desired to use for testing. In an attempt to get a good spread of results, the three teams were of varying skill levels, with one of them being the clubs women's team also. The initial data gathering commenced in August 2022 and continued throughout the semester since testing could only realistically be done during the respective trainings of each team.

One player at a time was taken out of training in order for them to complete the ten exercises decided upon with the members of Output Sports. The four jumping exercises were tested first, followed by the power output tests, the shoulder mobility and finally the speed and agility tests. It took approximately 15-25 minutes to fully test an athlete, with the goal of testing 45 handball players in total by the end of the semester. Unfortunately, not all of the athletes were able to be tested given time restrictions, cancelled trainings, upcoming matches etc.

Once the players created a profile on the Output mobile app, the testing would begin. The sensor, connected via Bluetooth through the app itself for a stronger connection, was attached using a Velcro strap to different parts of the body depending on which exercise was being tested. For the jumps it was placed on the players' dominant foot (with the exception of the Uni CMJ's as both legs were tested), for the power on the dominant wrist, and each wrist for the mobility exercises as both shoulders were tested. No sensor was needed for the speed and agility tests, nor for the throwing speed readings.

Apart from the Sprint and T-Test exercises, everything was done in free space and only required a stable connection with the sensor. The sprint required a 20 meter straight line, and the T-Test had a 10 meter center line and a 10 meter top line across (5 meters each side from the center line to the ends of the top line). Since these dimensions had to be measured out each time testing was to

be done, it led to some obvious human error. However this was deemed mostly insignificant due to the difference in scale between the (at most) centimeters of error versus the meter or more length of stride the players took. The biggest factor of error would (besides that of improper technique) be found in the sprints. Each player was given a countdown, with the timer on the mobile application being started on "GO" by the author. This error was minimised by the applications ability to stop the timer automatically once the athlete crossed the finish line (which was aligned with the indicator on the camera opened through the app).

The last thing to do was to test the author on the exercises contained in the final model in order to provide the model with new data is has not seen before. Once the authors results had been fed into the Final Model, a percentage error was calculated in order to gauge how well the model did in the prediction.

To briefly mention the ethics of the testing, all players volunteered for the testing of their own volition and their data is covered under the terms and conditions of Output Sports Limited. The only identifying piece of information actually used in the investigation was the player names, of which I make no mention of in this or any report. I am the only non-company person who had access to the raw data and details of the volunteers, all of which was deleted once this study had been completed.

2.3 Data Cleaning & Analysis

Once a portion of the data had been collected, the cleaning process could get underway. The raw data was exported from the online Output Hub onto the authors computer, where the individual ".csv" files had to be filtered and combined. An iterative program was written in Python to read in the 10 exercise files and filter the column of the desired metric for any unreasonable results. The ranges for reasonable values were devised after consultation with colleagues at Output Sports, and also from the practical knowledge gained during the testing process.

Once the data had been filtered, some of the exercises needed to be split into files based on which limb was begin tested (left or right). Once this had been done, the average values for the desired metric from each of the 10 exercise files, as well as the "Throwing" measurements, were calculated and the important information from each was extracted and added to a new singular file. From here we imported the file into R to alter the column names, remove some unfinished observations and begin looking at the data.

Some descriptive values were extracted like the predicted variables mean, standard deviation etc. Following this, a correlation analysis was run in order to identify any significant relationships between variables that may cause us problems in the future. Those pairings with a high correlation were kept in mind when interpreting the results and choosing the final model.

2.4 Models and Model Building

We used several methods over the course of this investigation in order to identify the most effective model, and in turn discern which of the tested exercises had the greatest impact in the prediction of

the throwing speed of an athlete. Once the optimal model has been identified, we examine certain properties, such as its residuals, in order to uncover any possible issues or shortcomings it might have with the fit of the data. In the following section we talk about standard linear regression, The Lasso Method, Random Forests, Stepwise Regression and Principal Component Analysis (PCA). Following this, the discussion leads into an explanation of the Residual Analysis we undertake when examining our final model.

2.4.1 Full Model

To begin, we formed the "Full Model", containing all 13 variables using standard regression. As we can see from Fig.1, with the exception of the tails, the model fits quite well to the centre line, indicating that the model residuals (the difference between the observed values and values predicted by the model) are normally distributed, a necessary condition in order to run linear regression. This bodes well for future analyses of smaller models.

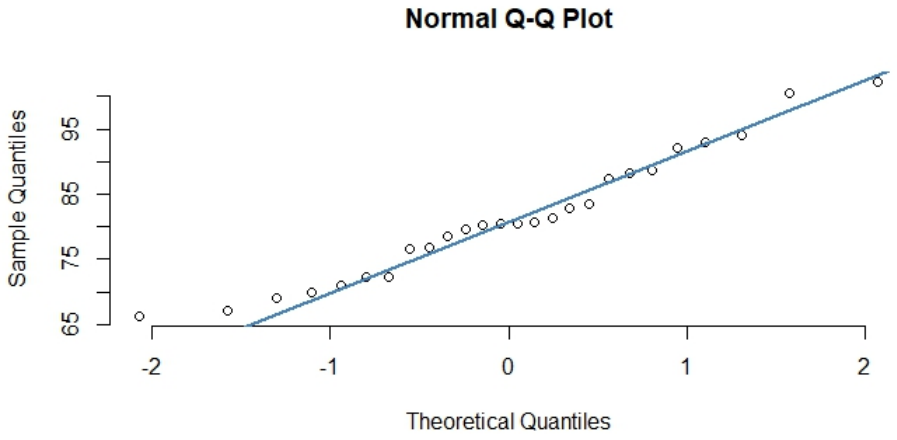


Figure 1: Q-Q Plot for the Full Model

2.4.2 Stepwise Selection

There are several approaches to Stepwise selection, including the how to compare the measure of fit, the stopping criteria and the direction. In this project we performed both Forward and Backward selection two times, once using the models' AIC as the measure of fit, and another using BIC.

In Forward Selection, the algorithm begins with the null model, i.e. no predictor variables, and at each step adds the variable that most improves the accuracy of the model. Since we are using AIC and BIC as measures of accuracy, the variable that lowers the relevant metric the most will be

included in the next selection step. Backwards Selection however begins with the full model and *removes* the variable that effects the model the least until no improvement was found or the Null model was achieved.

2.4.3 Lasso Regression

The Lasso method is a regularisation technique that we utilise in order to prevent overfitting in our model. Since our full model has quite a high number of independent variables relative to the number of observations, it is at risk of such a problem.

The method estimates the coefficients by minimizing the cost function, to which it adds a shrinkage penalty in order to encourage the beta-coefficients to be as small as possible while still maintaining a good fit to the data. The shrinkage penalty is controlled by a tuning parameter, λ, that is proportionate to the absolute value of the coefficients. When λ is large it results in a stronger penalty that reduces the value of the coefficients, even lowering some of them to exactly 0. This feature of the Lasso method is the reason we include it in our model selection process.

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=0}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j| \tag{1}$$

Eq. (1) (ref.4) shows the loss function with the additional penalty term ($\lambda\sum_{j=1}^{p}|\beta_j|$) and tuning parameter λ. Since λ is a hyperparameter and therefore must be chosen before the process begins, we will use Repeated K-Fold Cross Validation with RMSE as the model assessment criteria in order to determine the most optimal value of λ.

2.4.4 Random Forest

Random Forest analysis is an ensemble method (the combination of multiple models) used to split and organise ones data in order to identify the importance of each variable with respect to the predictor variable. It works by creating multiple *decision trees* and combining their results in order to find the model that minimises the RMSE.

In a regression analysis, a decision tree is used to predict a continuous target variable. The algorithm starts at the "root" of the tree, and at each internal "node", it splits the data based on the value of a feature, so that the samples are split into subsets with similar target values. The algorithm continues to build the tree recursively by repeating the process at each internal node until a stopping criterion is met. The final tree can be used to make predictions on new data by traversing the tree from the root to a leaf node.

The number of variables considered at each split is an important parameter in random forest model as it controls the complexity of the decision trees in the model. The random forest algorithm

will still consider all the variables in the data set when building the decision trees, but at each split, it will randomly select a subset of variables from all available variables to consider for the split.

We again used cross validation in order to determine the most optimal structure for the decision trees in our random forest. We can see from Fig.6 that the forest in which 3 variables were considered at each split held the lowest amount of error, with the error increasing steadily for more complex models.

2.4.5 Principal Component Analysis (PCA)

PCA is, amongst other things, a dimension reduction technique that transforms the factors into new, uncorrelated variables such that the most variance is captured by these "principal components" as possible. Geometrically speaking, principal components represent the directions of the data that explain a maximal amount of variance (ref.5). In doing so, it can in some settings provide us with a simpler model, removing a lot of the noise contained in models such as those with lots of predictor variables.

Principal Components are calculated as follows:

1. Standardize the data set using the standardization formula:
$$x_{new} = \frac{x - \mu}{\sigma}, \tag{2}$$
and form the standardization matrix, S.
2. Calculate the covariance matrix.
3. Calculate and sort the eigenvectors and eigenvalues of the covariance matrix in decreasing order.
4. Choose the top k eigenvalues and form a matrix, P of the respective vectors.
5. Transform the standardized matrix by multiplying it by the eigenvector matrix: $S * P$

The resulting matrix will contain the transformed data that can now be examined on simpler axes of dimension k. Once the above procedure was performed, we subjected our newest model, containing the top 2 PCs, to cross validation in order to determine its predictive abilities.

2.5 Residual Analysis

Once the different models have been formed using the methods outlined above, we identified the most promising model using some residual analysis techniques. We looked at some metrics commonly used in model selection, as well as the Studentized Residuals of the final model. In addition, properties of the data such as Leverage and Cook's Distance were analysed to check for any irregularities within the final model.

2.5.1 Selection Metrics

Root Mean Square Error

The RMSE differs from the Mean Square Error only in the square root seen in Eq. (3). From this equation we can also see that a smaller RMSE value is more desirable.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \tag{3}$$

R^2

R^2 is a measure of the variability of the dependant variable. Traditionally, the R^2_{adj} is generally thought of as a stronger indicator given that it accounts only for the variables that account for some measure of variability and punishes models that contain variables that don't. In our case however, we have access to stronger numerical tools, allowing us to run more complex operations and bypassing the need for the R^2_{adj}. Instead, since repeated k-fold cross validation is used, the R^2 value can be considered a stronger representation of the quality of given model and so we use that.

The R^2 value is calculated by using Eq. (4) (ref.6). The higher the R^2 value the better, as it indicates a higher percentage of variability explained by the data.

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} = \frac{SS_{ex}}{SS_{tot}}, \tag{4}$$

where SS_{ex} is the explained variation and SS_{tot} is the total variation.

Mean Absolute Error

As the name suggests, the MAE calculates the average absolute error over the data in the model. It was of course preferable that the models had a lower MAE, calculated with the following equation:

$$MAE = \frac{\sum_{i=1}^{N}|\hat{y}_i - y_i|}{N} \tag{5}$$

Similar to the RMSE, a lower error value is of course more desirable.

Using these four metrics to evaluate the effectiveness of our models, we chose the most successful method to investigate further and took a deeper look at its characteristic by analysing those in the section below.

2.5.2 Studentized Residuals

Residuals are the difference between the observed values of the dependent variable and the predicted value provided by the model. The difference between Residuals and Studentized Residuals (SRs) is that SRs have been standardized by dividing the residual value by the estimated standard deviations of the residuals, also referred to as the residual standard error. This results in the SRs to be unaffected by the dependant variables scale (km/h) and the size of the model's residuals, leading it to be a more appropriate detector for potential outliers in our model.

The studentized residuals are calculated by using Eq. (6) (ref.7) below.

$$t_i = \frac{|y_i - \hat{y}_i|}{\sqrt{\frac{\sum_{j=1}^{N-1} |y_j - y_{j\hat{(-i)}}|^2}{N-p-1}}} = \frac{|y_i - \hat{y}_i|}{\sqrt{MSE_{(-h_{ii})}(1-h_{ii})}} \tag{6}$$

where $y_{j\hat{(-i)}}$ is the j'th fitted value when $\hat{y}_i$ is removed, h_{ii} is the leverage of the i'th observation and $MSE_{(-h_{ii})}(1-h_{ii})$ is the MSE of the residuals with the i'th leverage removed.

2.5.3 Leverage

Leverage measures the relative influence a given data point has on the fitted model. Leverage was of interest to us because a high leverage can have a significant effect on the intercept and coefficient values of our model. Since we worked with quite a small data set, any strong effects on our data or model wished to be avoided. We also include a plot of the fitted values against the square root of the leverage value to avoid the skewedness that tends to occur when the original value is used. Transforming the leverage values also compresses the range of values, making it easier to see patterns in the graph.

By standard convention, any leverage value greater than three times the average leverage (Eq. (7)) of the data points is considered high and should be monitored. We included points that were two times the average leverage but kept an eye on these points also.

$$\bar{h} = 3\Big(\frac{p+1}{N}\Big) \tag{7}$$

2.5.4 Cook's Distance and DFBETAS

The last methods of analysis that were looked at were the Cook's D and the DFBETAS of the data. Cook's D looks at the impact removing a certain observation has on the regression coefficients, whereas the DFBETAS looks at the impact of removing one of the predictions instead. A large Cook's D or DFBETAS score indicates the point holds a large sway on the data and should be noted and or investigated.

Cook's D and DFBETAS of a point can be calculated using Eq. (8) (ref.8) and Eq. (9) (ref.9) respectively.

$$D_i = \frac{\sum_{j=1}^{N}(\hat{y_j} - y_{j\hat{(-i)}})^2}{p * MSE} \tag{8}$$

$$DFBETAS_i = \frac{\beta_k - \beta_{k(-i)}}{\sqrt{MSE_{(-i)} * c_{kk}}} = \frac{\beta_k - \beta_{k(-i)}}{SE\beta_{(-i)}} \tag{9}$$

where p = number of predictor variables in the model, $y_{j\hat{(-i)}}$ the j'th fitted value when y_i is removed, $\beta_{k(-i)}$ is the k'th coefficient of the model with the i'th observation removed and c_{kk} the k'th diagonal element of the unscaled covariance matrix. The unscaled covariance matrix is a matrix such that multiplying it by an estimate of the error variance produces an estimated covariance matrix for the coefficients (ref.10). Typically a Cook's D and DFBETAS value greater than $\frac{4}{N}$ and $\frac{2}{\sqrt{N}}$ respectively can be considered as a potentially influential observation, however this varies with the size and dimensions of the data set, hence the use of the previous analysis methods.

3 Results

To briefly summarise this section, we first look at the results from the data and correlation analysis, identifying any descriptive values and significant relationships combined with a look at the PCA. Then we present a table with the comparative metric values. Next we look into meaningful results from each model and show the results of the further investigation into the final model. Lastly, the authors own measurements will be used in order to test the final model on its predictive capabilities with previously unseen data.

Repeated k-fold cross validation testing was performed on all models, which we ran over $k = 5$ folds and repeated 20 times. These parameter values were chosen so the data could be best split in order to avoid folds that were too small and therefore diluted the data. In order to compensate for the small number of folds, a high repeat value was chosen, resulting in a higher computation time but more accurate results. The results in Table 3 below are taken from the cross-validated model in each case with the exception of the Null Model, which has no predictor variable values to split.

3.1 Exploratory Data Analysis

We started by checking the marginal distribution of our response variable to see if it is normally distributed or if we need to apply some transformation before we run our regression analysis. As we can see from Fig.2 (ref.3) our variable is nicely distributed with a mean around $81km/h$ with a standard deviation of 9.71. Table 2 shows more descriptive metrics for the Throwing variable. Next we looked into the correlation of our data set, keeping in mind the results of this during the model building process described below.

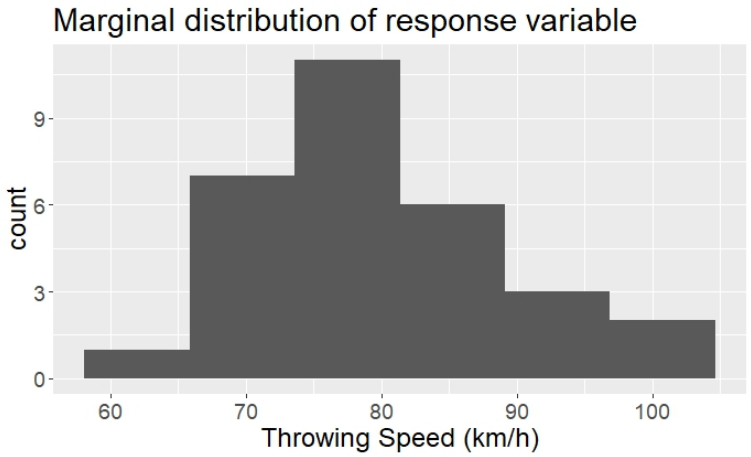


Figure 2: The image shows the marginal distribution of our response variable "Throwing". As we can see, the variable is quite normally distributed (as expected).

	Mean	Median	SD	Min	Max
Throwing	81.1	80.5	9.7	66.3	102.3

Table 2: Descriptive measures of the dependant variable.

Looking at Fig.3 (ref.3) we can see that there is occasional mild, positive correlations between some of the variable pairs. Almost all of these were between pairs of "Jump" variables, the only exception being "TTest" and "tenfive" with a correlation of -0.70. The strongest correlations were between (BiCMJ, SquatJump), (UniCMJL, UniCMJR), (BiCMJ, UniCMJL and UniCMJR) and (SquatJump, UniCMJL and UniCMJR), with correlations 0.94, 0.95, 0.87, 0.88, 0.84 and 0.81 respectively . This was expected given how similar all of these exercises are to each other, especially the first two pairs which are extremely synonymous. In Fig.17 found in Appendix A, we can see the correlation of each variable pair, the variable distributions and the point plot for each combination of variables.

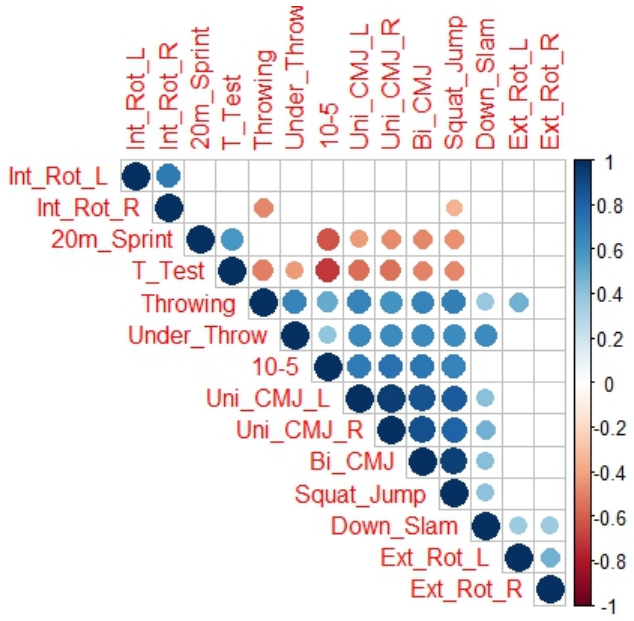


Figure 3: Variable Correlations. Only variables with a correlation of $p > 0.05$ are shown. Lesser correlations are left blank.

Though it may be quite difficult to see in Fig.4, the PCA has clustered together exercises of similar classes, for example all of the Jumping exercises are contained in the red group. cluster 1. Running PCA allows us to have our data contained to a 2 dimensional space, where interpretations and conclusions are more easily drawn, instead of the 13 dimensions if all variables were used.

Later we show that the PC model did not prove to be as effective as originally hoped given by its model results in Table 3, although it did provide some credibility to the correlation analysis and our choice for the final model. As we can see in Fig.4 above, over 50% of the variability of our data is described by just the first two PCs. Having a good spread of variables from both of these components is a good sign of how well our final model might capture the variability found in our observations.

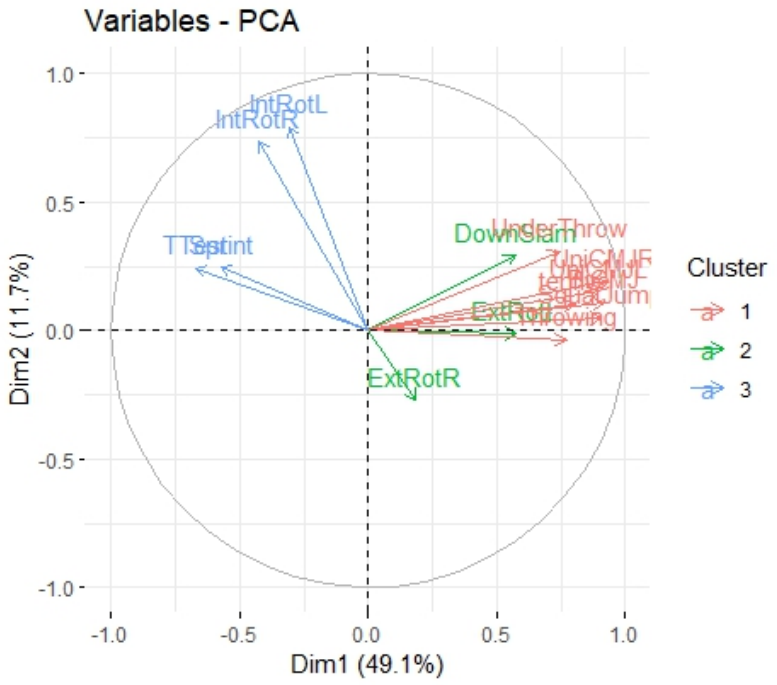


Figure 4: PCA circle graph showing how the transformed variables were grouped. It can be seen on the axis labels that the first PC contained almost 50% of the variability in the data, while the second PC contained less than 12%.

3.2 Model Selection

We now dig deeper into the results from each of the model selection methods outlined in the previous sections. Below is the results table of all the comparative measures we used to identify the most suitable model from all of techniques we utilised over the course of this study.

Model	RMSE	R^2	MAE
Full Model	10.650	0.417	8.979
Null Model	9.530	~ 0	7.540
Lasso Model	6.966	0.597	6.017
Random Forest	5.788	0.707	4.971
Stepwise Model	5.887	0.750	4.853
PCA Model	6.841	0.595	6.039

Table 3: Model selection metrics and results. A lower RMSE and MAE indicate a better fitting model, where as a larger R^2 value indicates the same. The values in this table have been calculated by cross-validation and thus reflect test error or test prediction accuracy (depending on the metric) and not training error.

3.2.1 Full and Null Model

After fitting the Full Model, it was clear that this was not a very reliable model. Overfitting was perhaps present as there were only a single significant variable (Right Arm Internal Rotation). All results for this model were found to be quite poor, with the highest values for RMSE and MAE, and the lowest value for R^2. These results may have been caused by a number of things, from unnecessary variables that only add to any overfitting, or the variable relationships we found during the correlation analysis (Fig.3).

Though the Null Model does not present any interesting results in and of itself, it is used to provide a base level to compare other models to. The Null Model results (Table 3) further justify our dismissal of the Full Model. Other than the near 0 value for the R^2 (presumably due to the lack of predictor variables), both of the other comparative metrics are better in the Null Model than in the Full. This indicates that a more complex model than the Full Model is needed in order to determine any further relationships within the data.

3.2.2 Lasso Model

The Lasso Model proved to garner quite good results in relation to those we have already looked at as well as in general, having a smaller (and more desirable) RMSE value than the Final Model (Table 3). The repeated k-fold cross validation on 50 tuning values (λ) proved to be quite supportive of the final model that was chosen. A $\lambda = 2.56$ was found to provide the best lasso model, having an RMSE of 6.97. After forming a lasso model with this value of λ, it was found to contain the four variables SquatJump, IntRotR, UnderThrow and UniCMJL.

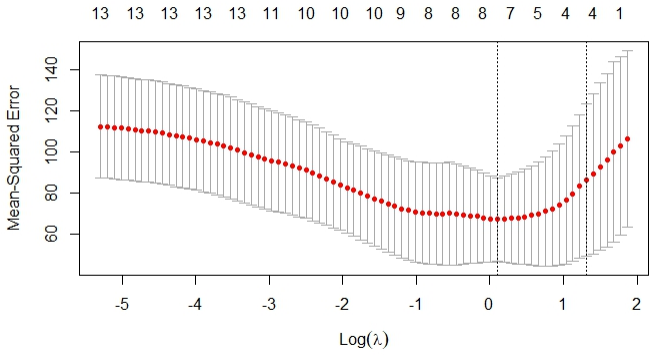


Figure 5: The results of the cross validation process showing the smallest λ (left dotted line) and the lambda.1se (right dotted line).

Fig.5 shows the results of k-fold cross validation from a slightly different R function. The cross validated model it formed provided 2 choices of λ. The leftmost dotted line shows the λ that

minimised the RMSE value, $\lambda = 1.11$. However this is not always the best choice as it can be a result of overfitting. The second dotted line is a more realistic value for λ. The $\lambda = 3.72$ model may have had a slightly higher RMSE value of 9.28, but it results in a simpler model (4 variables, taken from the top axis).

3.2.3 Random Forest Model

One of the first things we found out after running the repeated k-fold cross validation on the random forest model was that the best results occurred when exactly 3 variables were chosen at random to be considered at each split. Fig.6 shows the trend in RMSE for the best model in each choice of the number of variables considered. A very clear minimum can be seen when 3 variables were used.

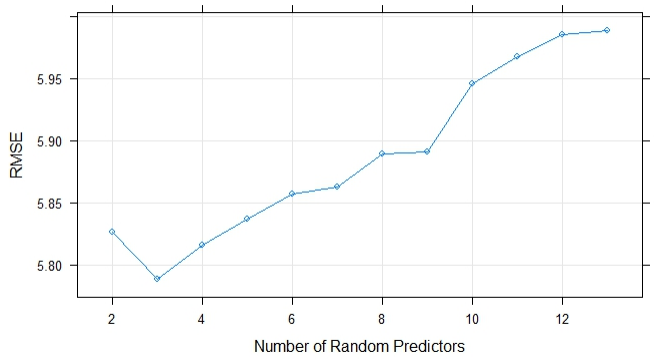


Figure 6: Number of random predictor variables considered at each split against the RMSE of the Random Forest.

Though the random forest doesn't give us a specific regression model, it does allow us to gain some insight into the "Importance" of the variables in our data set. The repeated cross validation revealed that the decision trees showed the best results when sets of 3 variables were chosen to be the filters at each split (Fig.6). The smaller the number of variables used here the more decorrelated the trees in the forest get from each other..

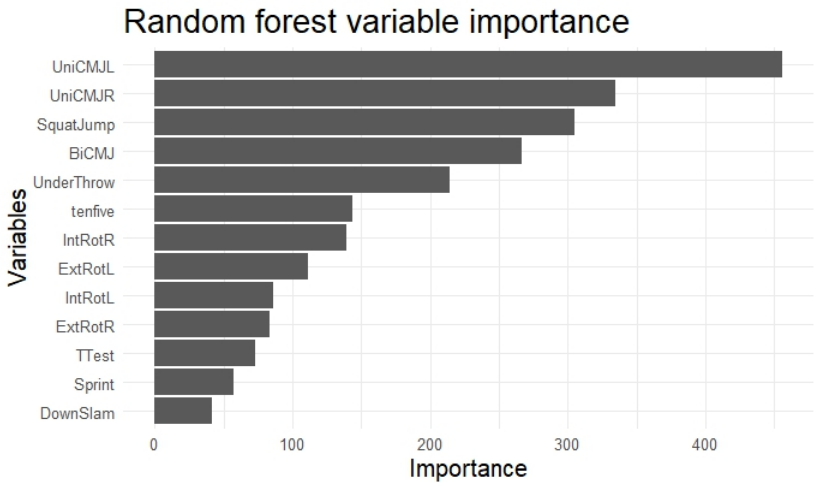


Figure 7: Importance values for each variable after repeated cross validation.

Fig.7 shows the Importance of the variables in our model. Found by measuring the relative change in accuracy between models when the variable is removed, importance is a very good indicator of the significance of predictors. We notice that the jumping exercises are actually the most important for predicting the throwing speed. We look into the relation of this result with that of our final model later in the Discussion section.

3.2.4 Stepwise Model

In our stepwise regression analysis we formed 4 models, exhausting all combinations of forward, backward, AIC criteria and BIC criteria. We decided to go with the AIC criteria for both forward and backward models since they garnered the same model summary results and AIC tends to be a better facilitator of balancing goodness of fit and model complexity.

Forward AIC			Backward AIC		
Step	AIC	# of Variables	Step	AIC	# of Variables
1	123.74	0	1	106.10	13
2	108.83	1	2	104.17	12
3	105.94	2	...	...	...
4	99.20	3	9	94.95	5
5	94.28	4	10	94.28	4

Table 4: Stepwise Regression results. BIC criteria was also used in the testing, though no differences could be seen.

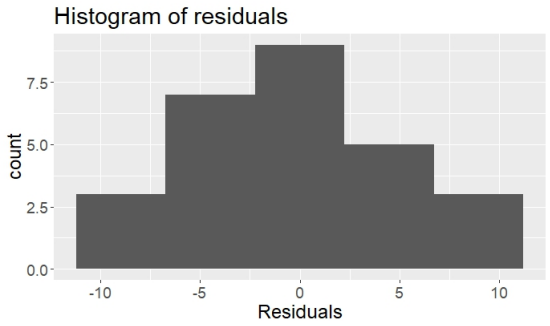


Figure 8: A histogram of the residuals from the Stepwise Model. They appear to be quite well normally distributed

After running the stepwise algorithms we arrived at quite an interesting conclusion. Both the forward and backward methods resulted in the same model containing the SquatJump, IntRotR, UnderThrow and IntRotL variables, and with an AIC of 94.28. This is a very positive result in favour of this model since both directions ended up at the same conclusion even with different accuracy criteria. With a look at our results table we can that the Stepwise Model is slightly better than the next best model, the Random Forest, on all fronts, labelling stepwise regression as the most successful method and therefore our choice as the Final Model.

3.2.5 Principal Component Regression

Lastly we formed a model using the first two PCs to see if there were any significant results, although we did not have many expectations given the amount of variability described by the number of principle components we used (just over 60%). Though not the worst model, it only really surpassed the Full and the Null Model in its performance. It actually performed quite similarly to the Lasso Model, with a slightly better RMSE, almost the same R^2 value, and a slightly worse MAE.

3.3 Final Model Analysis

Model Summary				
Min -8.0	1Q -3.0	Median -0.2	3Q 2.7	Max 10.0
RSE 5.275	R^2 0.7503	R^2_{adj} 0.7049	p-value $2.2x10^{-6}$	
	Coefficient Estimates			
Intercept 41.12	SquatJump 0.34	IntRotR -0.52	UnderThrow 8.77	IntRotL 0.40
	Coefficient Confidence Intervals (2.5%, 97.5%)			
Intercept (10.2, 72.0)	SquatJump (−0.13, 0.81)	IntRotR −(0.78, 0.26)	UnderThrow (3.88, 13.66)	IntRotL (0.07, 0.72)

Table 5: Summary of the Final Model containing some descriptive measures and the estimates for the Intercept and the four predictor variables.

Variable	SquatJump	UnderThrow	IntRotR	IntRotL
Importance	290	193	151	85

Table 6: Table of the variables from the Final Model with the Importance values taken from the Random Forest investigation after repeated cross validation. The highest importance found between all 13 variables in the Random Forest was 380, while the lowest was 54.

After the insight provided by all of the techniques outlined above we finally arrived at the decision to set the Stepwise Model as our Final Model. Containing 4 variables: SquatJump, IntRotR, UnderThrow and IntRotL. it had the best R^2 and MAE value, and was only 0.1 higher than the RMSE value of the Random Forest. Once the Final Model had been chosen we took a closer look at the finer details of the collection of variables used and how the data was affecting the results. To start, the Studentized residuals were calculated and plotted against the fitted values.

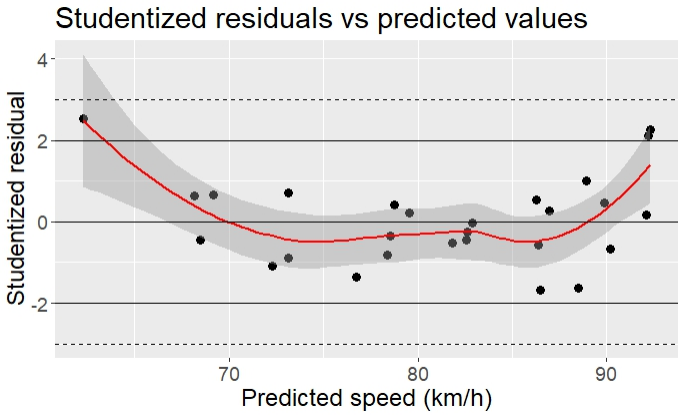
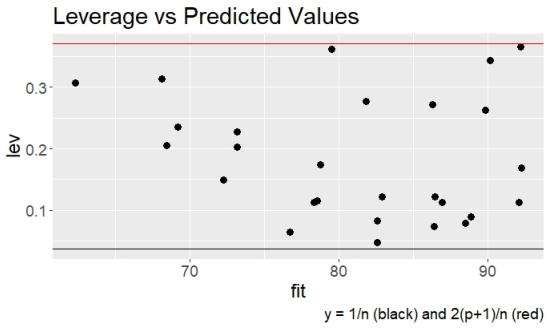


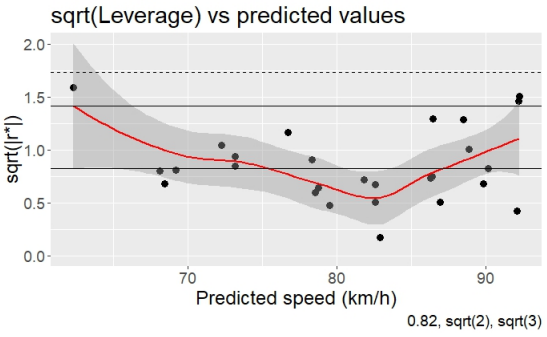
Figure 9: Studentized Residuals vs Fitted values of Throwing variable. Points above or below the solid lines at ($|y| = 2$) are considered unusually large, while points above or below the broken lines at ($|y| = 3$) are considered suspiciously large

Fig.9 reveals 3 data points above the lower limit we indicated with the solid lines. Since all of these values occur at the tails of the graph and the remainder of the data is spread quite evenly around 0 (with only a few points even coming close to the lower limit), the plot looks quite well fitted.

The two plots for the leverage against fitted values appear to be quite normal. A couple values approach the red line in Fig.10a but most remain quite low. As desired, most of the data varies around 0.82 in Fig.10b, with a similar occurrence as the studentized residuals where we notice the tails spike slightly due to a few singular points above the lower indicator line.



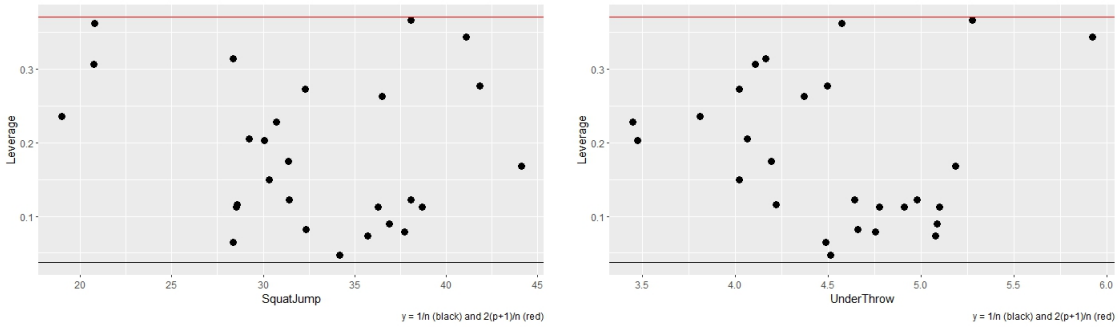
(a) Leverage plotted against the fitted values. Notice the red line to indicate extreme values.



(b) Square root of the leverage value vs fitted values. Data should vary randomly around 0.82 (lower solid black) without systematic trends.

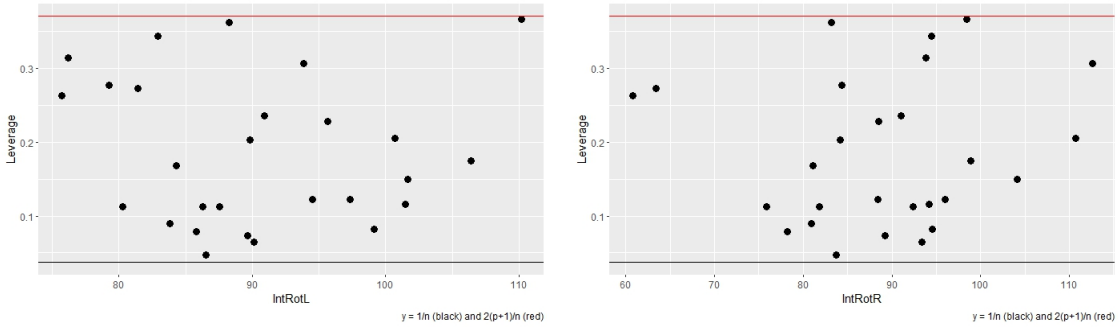
Figure 10: Leverage plots of the Final Model.

Similar results can be seen in figures Fig.11 and Fig.12. A small number of observations approach close to the line for high leverage, with none going over. The data is spread quite far out but this is to be expected with such a physically diverse group and such a small data set.



(a) Leverage vs SquatJump. **(b)** Leverage vs UnderThrow.

Figure 11: Leverage plots for the SquatJump and UnderThrow variables with an upper threshold for concern at $y = 0.37$ in red.



(a) Leverage vs IntRotL. **(b)** Leverage vs IntRotR.

Figure 12: Leverage plots for the IntRotL and IntRotR variables with an upper threshold for concern at $y = 0.37$ in red.

When looking at the Cook's D plot in Fig.13, we see the same three points above the lower bound for suspicious observations as those in the Studentized residual plot. Though not an extreme distance from the commonly accepted norm, however in such a small data set this pushed us to keep an eye on these specific observations to avoid any significant influence.

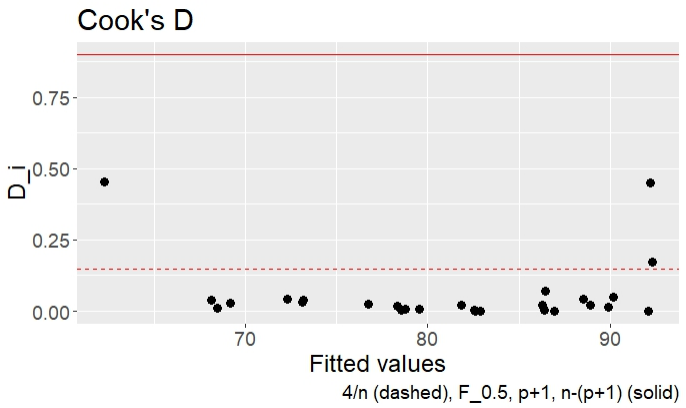


Figure 13: Cook's D vs Fitted values of Throwing variable. Points above the broken line ($y = 0.15$) are considered unusually large, while points above the solid line ($y = 0.90$) are considered suspiciously large

Noticing the single point past the higher threshold of the first DFBETAS plot in Fig.14 prompted us to investigate this subject further. Also a part of the group of three found in the earlier analyses, it was found to be the subject with the highest result for our predicted variable, which is thought to be a strong cause of the high influence it has over the data in the model.

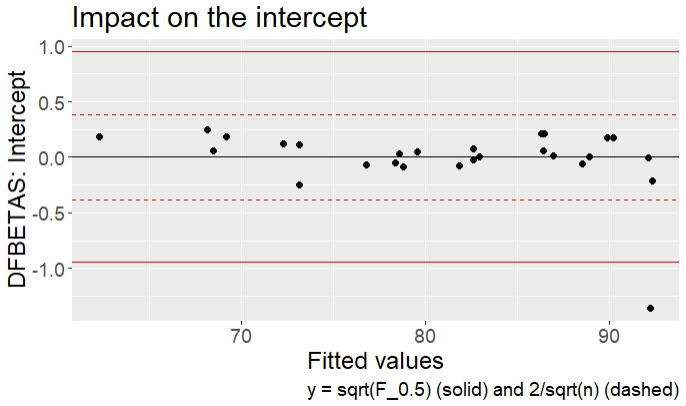
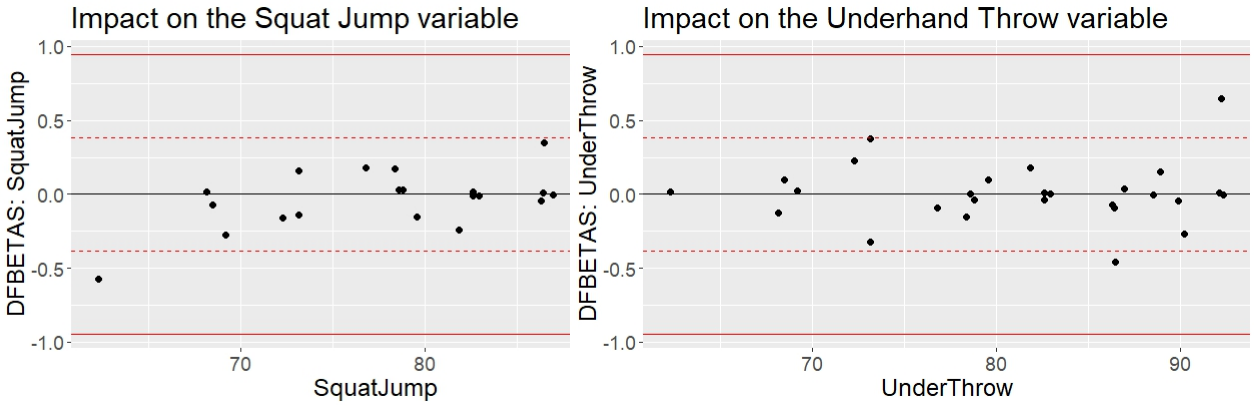


Figure 14: DFBETAS vs Fitted values of Throwing variable. Points above or below the broken lines at ($|y| = 0.39$) are considered unusually large, while points above or below the solid lines at ($|y| = 0.95$) are considered suspiciously large.

Again we see in the DFBETAS plots for each of the variables in the final model that almost all points lie within the lower threshold of "normal" influence with the exception of one or more of the three points noted earlier lying in between the two threshold bands (Fig.15a, Fig.15b and Fig.16a).

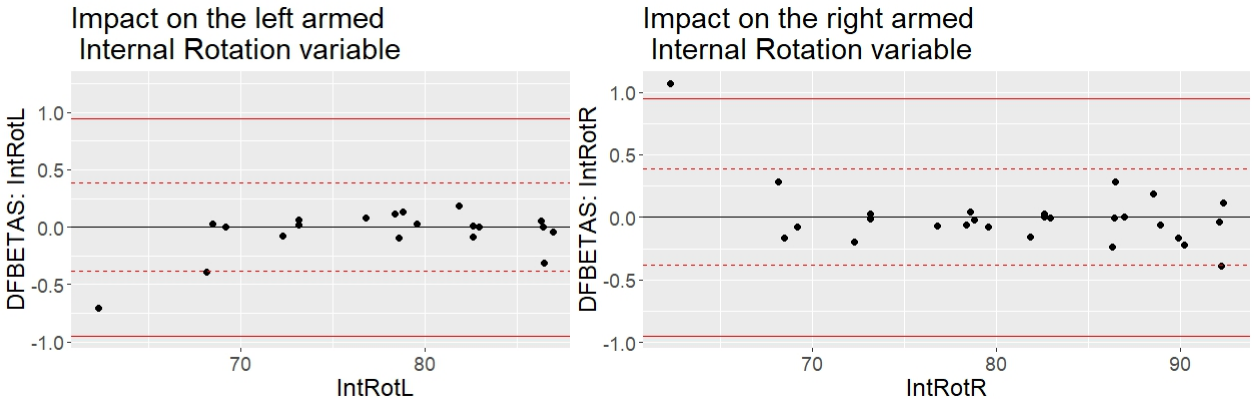
However in Fig.16b the same outlier as in the Cook's D plot lies just outside the edge of the upper threshold.



(a) DFBETAS vs SquatJump.

(b) DFBETAS vs UnderThrow.

Figure 15: DFBETAS plots for the SquatJump and UnderThrow variables. Points above or below the broken lines at ($|y| = 0.39$) are considered unusually large, while points above or below the solid lines at ($|y| = 0.95$) are considered suspiciously large.



(a) DFBETAS vs IntRotL

(b) DFBETAS vs IntRotR

Figure 16: DFBETAS plots for the IntRotL and IntRotR variables. Points above or below the broken lines at ($|y| = 0.39$) are considered unusually large, while points above or below the solid lines at ($|y| = 0.95$) are considered suspiciously large.

3.3.1 Final Model Testing

Once we confirmed that the Final Model was not hiding any extreme underlying disturbances, we proceeded to supply some new, unseen data into the model for prediction. The authors results vector, (SquatJump = 37, IntRotR = 92, UnderThrow = 4.72, IntRotL = 100), was fed into the model. The resulting prediction for throwing speed was 86.36 km/h, which turned out to be very close to the observed value of 84.25 km/h. The prediction on unseen data for a new subject had an error of approximately 2.5%. We deemed this quite reasonable since the observed value we used to compare with the predicted value was an average of multiple measurements, of which the minimum was 80 km/h and the maximum was 87 km/h. As we can see the predicted value was well within the set of observed measurements.

4 Discussion

We take a small step away from the statistical results now to have a critical look at certain aspects of the study itself. To start at the beginning, the data collection of course had some room for improvement, as does any process that involves direct contact with the tester. The tests needed for variables like Sprint and T-Test involved laying out markers a specific distance from each other and the "Start" button for the timer to be pressed as the athlete began their run. Both of these run a high risk of human error that is made more relevant given the accuracy the sensors measure in (to three decimal places for the variables mentioned above).

Another unfortunate aspect was the size of the data set. With only 27 complete observations eligible for the regression analysis, it was quite hard to form any meaningful significance or conclusions without having access to a larger spread of athlete results. This was due to time constraints and athlete availability, as testing could only be done during the trainings of the handball players, and not everyone present wished to be involved. However the authors position in the local handball club allowed him to take measurements from players of different levels throughout the different associated teams. This diversity was very important in order to help compensate for the sample size.

Something that should be noted is the effectiveness of the velocity gun used to take the Throwing variables measurements, which gave results much lower than was to be expected. Even though this was not the most accurate or reliable of tools, it provided a consistent and comparable space in which all of the subjects' results would be relative to the same technology, allowing the study to proceed without any major adjustments to the data. Given the accuracy of the authors predicted throwing speed, we can comfortably say that more accurate methods for gathering this variables information should lead to similar outcomes.

One of the things that was considered when choosing the model building methods was model reduction. Due to the small sample size coupled with the large number of predictor variables, our models ran a high risk of overfitting. The need to combat this was what led us to choose lasso regression over ridge regression. While ridge regression lowers the variable coefficients somewhat like what happens in lasso regression, the lasso method has the ability to set the coefficient to

exactly 0, thus holding the power of feature selection. We also chose to look at PCA for the same reason, as it would allow us to transform the data onto a different space where we could use a much smaller amount of principle components to assess the data instead.

The model results occasionally lined up quite nicely with each other. For example, the forward and backward stepwise regression both suggested the same model as the best option which was a very good sign for the strength of the Final Model. In addition, the Lasso model contained almost all of the same variables as those in the Final Model with the only difference being that the UniCMJL variable was in the place of IntRotL. The λ that resulted in this model also suggested that a four-variable model was the most successful and had the strongest prediction power, a result in favour of both the Lasso and Final Model.

Another positive that was found from another model was the results from the Importance testing done when forming the Random Forest. The results in Fig.7 show that SquatJump and UnderThrow had quite high importance, with both variables used in the Lasso and Final Model. In addition to this, we can accommodate the fact that SquatJump was chosen in place of some other "Jump" variable due to the high correlation found between all the Jump variables (as can be expected). Notice also that SquatJump also lies in the middle of the rest of the Jump predictors in terms of Importance, which could imply that it was chosen because it contains the best spread of information from all Jump options. The inclusion of the IntRot variable was to be somewhat expected in a study testing for a "Throwing" movement. Their inclusion also accounts for a good spread of information from the lower end of the table.

Something interesting that came up later in the study was the changes that occurred after the late addition of a new data subject. The subject had been missing some variable results that forced them to not be included in the analyses until they were filled in. When this missing information was added however, most models showed significant changes to their final model suggestion. For example, previously the stepwise methods provided distinct models, one with four variables and the other with six. Afterwards we see that forward and backward selection aligned to give the same model. This could be because the new subject had good results and therefore held a large influence over the data, although most likely the changes were due to the size of the data set giving each individual a stronger pull on the data than they would have normally had in a larger sample.

The size of the data set impacted the study in other ways. In anything that involves the physical abilities of subjects, there will be a large diversity in their capabilities, especially if testing was done over different performance levels. In a set of this size this diversity was allowed to permeate into the results, as can be seen in some of the model analysis plots in the previous section. Fig.9 for example should ideally be monotone around 0, however the moving average shows the tails to be pulled up slightly from the distinctly low or high predicted values. In practice, this could be caused by the athletes prowess in certain areas given their position or physique. Normally this would be compensated for if the data set was sufficiently large and covered a wider range of physical capabilities.

Using the Final Model to predict the throwing speed based on new data revealed the models surprising predictive capabilities, only differing very slightly from the observed value. In saying this we of course have to mention that it could have simply been a fortunate result given that only a single previously unseen subject's data was tested on. To get a better idea of how well the model

predicts the throwing speed more subjects would be required. Nonetheless, the final result bodes well for future studies into this field.

Since the original purpose of the measurement equipment used in the testing is that of high performance athlete training and tracking, the implications of a more refined study could be quite significant in that field. If further testing reveals the ability to identify the most important exercises or movements in larger scale and more complex movements for different sports, it could sharply refine the focuses of the training and strength and conditioning of athletes in the future. In the specific case of handball in this study and which of the 13 variables tested hold the most influence on the Throwing result, the above results should be taken with a grain of salt given the shortcomings and possible improvements of the project, like including more exercises to test for example.

5 Conclusion

We briefly review some of the more important or interesting insights that were learned over the course of this study.

It was found that the Jump variables held an unexpected influence on the throwing action that was being tested. Most of the variables in the final model regularly ranked high in significance within different models as well as in the Importance analysis observed in the Random Forest, SquatJump being third highest overall.

The Final Model was backed by several results found from other models, like the Lasso model showing the ideal model contained 4 variables and the levels of Importance mentioned in the previous paragraph. In the overall case as well, Stepwise regression had the most promising R^2 and MAE values, as well as a very close second when it came to the RMSE.

It seems important to reiterate the importance of a larger data set for future studies however, given how much the small selection of participants affected the study. Certain elements would preferably be closer to the standard level of acceptance, such as the residuals be closer to the ideal, or the data to have larger dimensions to avoid any unwanted influence from individuals and to get more reliable and interpretable results. Having a larger testing pool removes a lot of the challenges in presenting more concrete results, but adding more variables should not be overlooked accompanied by this. If the sample size was much larger, we could afford to test more exercises without as much a worry on overfitting or diluting the modelling. Doing so would provide a deeper insight into which movements should be valued over others.

Given the previously discussed criticisms, the study yielded a surprisingly accurate final result with the predicted throwing speed only 2.5% larger than the observed value. Since the equipment used when taking these measurements only provided a relative measure of the athletes performance, a more accurate method would be needed in order to get results that would provide sufficient insight into the players performance. With this insight, trainers and coaches could streamline and tailor their training programs in order to maximise certain elements of the game that would otherwise be difficult to identify or focus on.

Further investigations would take these two recommendations, in addition to improving the method of collecting the results for the dependant variable, into account during the experiment building process in order to come to the most accurate conclusions. The conclusions themselves can provide a deep insight into the training and improvements of individuals in general and athletes as a whole if this is expanded into other sports.

6 References

[1] Output Sports Limited. Producer of the sensor used in the data collection.

[2] "Nonlinear Pedagogy in Skill Acquisition; An Introduction" - Jia Yi Chow, Keith Davids, Chris Button, Ian Renshaw; 978-0-367-42377-3. Page 10/11 "Humans as Complex Adaptive Systems".

[3] Online resource for calculating, analysing and visualising correlation of data.

[4] "An Introduction to Statistical Learning with Applications in R; Second Edition" - Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani; 978-1071614174. Page 241 Equation (6.7)

[5] Explanation of Principal Component Analysis.

[6] R^2 formula.

[7] Formula for Studentized residuals.

[8] Cook's D formula.

[9] DFBETAS formula.

[10] Definition of an unscaled covariance matrix.

7 Appendix

7.1 A

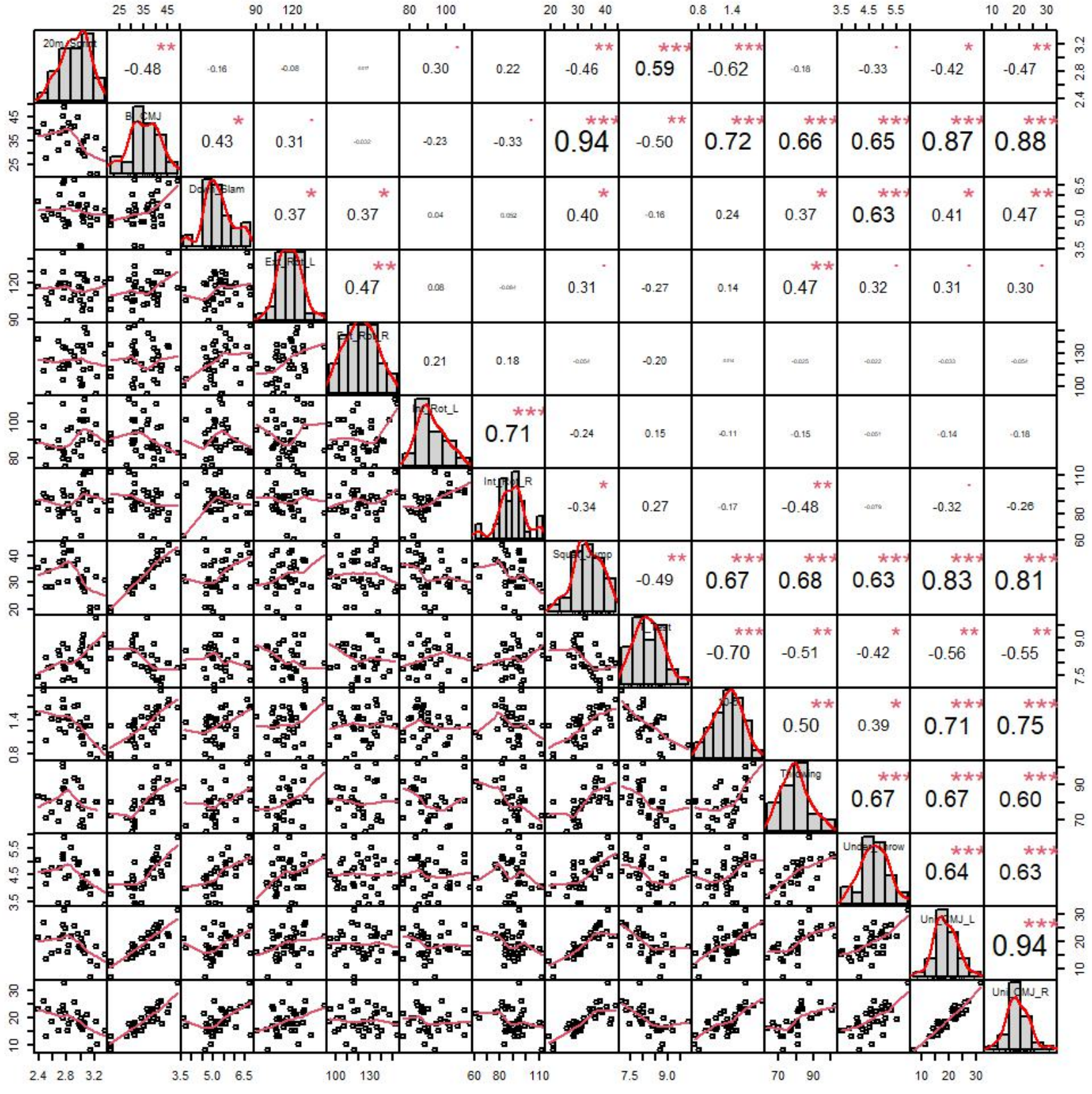


Figure 17: This graph shows the individual variable distributions along the diagonal, the pair plots of each variable pair in the bottom half and the p-value of each variable pair in the top half, highlighting the significant pairs of variables.

7.2 B

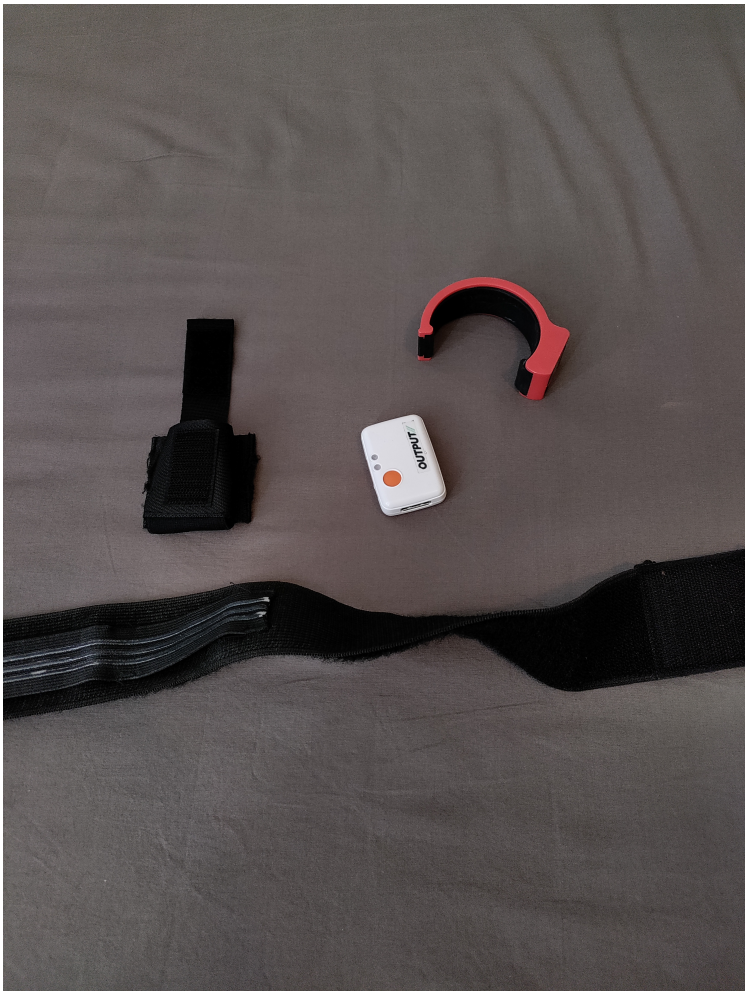


Figure 18: Sensor (white) and straps from Output Sports used in the data collection procedure.

Figure 19: Velocity gun used to measure the Throwing velocity of the handball players' step shot.