**Final project data analysis**

**Abstract.** In this paper, Jiyoon (Clover) Jeong and Jin Kweon are trying to inspect the official FIFA 2017 data. Our goal is to find out how the FIFA ratings on the players were decided. The better rating indicates how valuable the players are. We were curious whether the rating well indicates the players’ stats. Our data contains 17588 players with 53 different variables.

**Executive summary:**

**What should we write here…?**

**Do we need to double-space…?**

**Background:**

Two of us have pretty solid prior knowledge of soccer, and we used some of our intuitions when we do analysis. For example, When we do test out the PCAs, it is possible for us not just test all the quantitative variables, but select some of the variables that we thought would be reasonable to test. (which is more efficient)

Another important thing we decide to do for our analysis is to focus on clubs. Players play either on their national team or clubs. And, it is definitely possible for players to play for different positions and kit number in national and clubs. That way, it makes much harder for us to analyze and draw conclusions if we consider both. As players are selected to represent their countries based on their performance in their teams and players spend much more time in clubs, we decide to let players’ club profiles as our major variables. (however, depends on the situation, we used national information, and we are going to explicitly say it if so.

Also, please refer to appendix for the R codes.

**1.** **Introduction**

**a) Problem (the question I want to address)**

**1. We want to know how the ratings and ages are different based on club positions (goalkeeper, defense, midfielder, attacker) and preferred foot of the players.**

Reason: There has been enough arguments that FIFA has brought more and unfair attentions onto attacker (sometimes, midfielder as well), and other positions are treated/rated unfairly. And, we want to check whether the argument is true. Also, we are just curious how ratings are different amongst the players with different preferred foot. (also at the same time, the variable “Age” is really important to soccer players, so I want to include this for our quantitative variable. And, we assumed ratings and ages are closely related.)

Hypothesis: The means for each group of position-effect, preferred-foot-effect, and interaction are the same.

Aim/objective: Subjects will be 3656 soccer player (but different groups of population). As we stated above under the “Data” section, we filtered out the unspecified positions and modified the data, and 3656 samples are pretty decent size of samples. (around 20% of entire data) The quantitative measurement/variable are ratings and ages. There are 2 factors: club positions and preferred foot. There are 4 levels for club position: goalkeeper, defense, midfielder, and attacker. There are 2 levels for preferred foot: left and right. More than two different populations with two factors implies that I need to use “two-way MANOVA.”

**2. We are going to find out what quantitative variables are related with the rating.**

Reason: It is reasonable to assume that the overall ratings of professional soccer players are proportional to their score variables such as ‘Weak\_foot’, ‘Skill\_Moves’, ‘Ball\_control’, … , ‘GK\_Reflexes’. The assumption that the higher these score variable means the higher overall ratings is reasonable and we want to check which predictors are more stronger than other predictors by linear regression and transformation of variables.

Hypothesis: There should exist some quantitative variables that have a linear relationship with rating.

Aim/objective: Detect the predictors which have strong linear relationship with variable ‘Ratings’ and find proper transformation of variables if needed.

**3. I want to test how much each variable is correlated with the rating. PCA will also help me find co-linearity issue.**

Reason: We aim to find PCs to best summarize the variables, and see how players’ rankings are plotted.

Hypothesis: Different positions have different variables correlated with the rating, and I should find the skills that are important to the position should have high correlation with the rating.

Aim/objective: I hope to make good interpretation of the components to explain the the relationships between variables, and eventually help us how rating can be explained by other variables.

**b) Data (summary of the data, the study design, data collection)**

Please refer to the data dictionary we made: <https://github.com/yjkweon24/public-health-245/blob/master/dictionary.csv>

We collected our data in Kaggle website (please refer to the reference). The original data has 17588 rows and 53 columns.

It is important how we sample the data. Players are selected to play in the national team if they perform well in the club. It is true nowadays in the soccer world, players spend more time playing for the club. So, we are focusing on inspecting players’ club profiles only.

Although national related information is not our major variables, we our not going to take them out, as these information help us as some of the players do not have enough club informations. In this case, we replace national information with club’s. For example, many players do not have specific positions (“Sub”) for their clubs, and we tried to find these missed informations from their national positions, if possible.

To do the better analysis, we tried to be really careful of choosing the right sample. As we care about club positions more, we replace club positions with national positions if club positions are “Sub,” “Res,” and empty(“ ”), and this will help us a lot when we want to analyze the relationship between positions and ratings. Whenever we do analysis that has to do with positions, we needed to work on the extracted samples of the size 3656, as the others do not show clear positions.

We examined the raw data and found that variable ‘National\_Position’ has 16513 missing values and variable ‘National\_Kit’ has 16513 NA values. Variable ‘Club\_Position’ and ‘Club\_Joining’ has one missing value which is 384th observation, and variable ‘Contract\_Expiry’ and ‘Club\_Kit’ has one NA values which is also 384th observation.

**c) Purpose of the study**

The purpose of this study is to examine how the ratings were decided, and we hope to correct them if some of the players are either over- or under- rated.

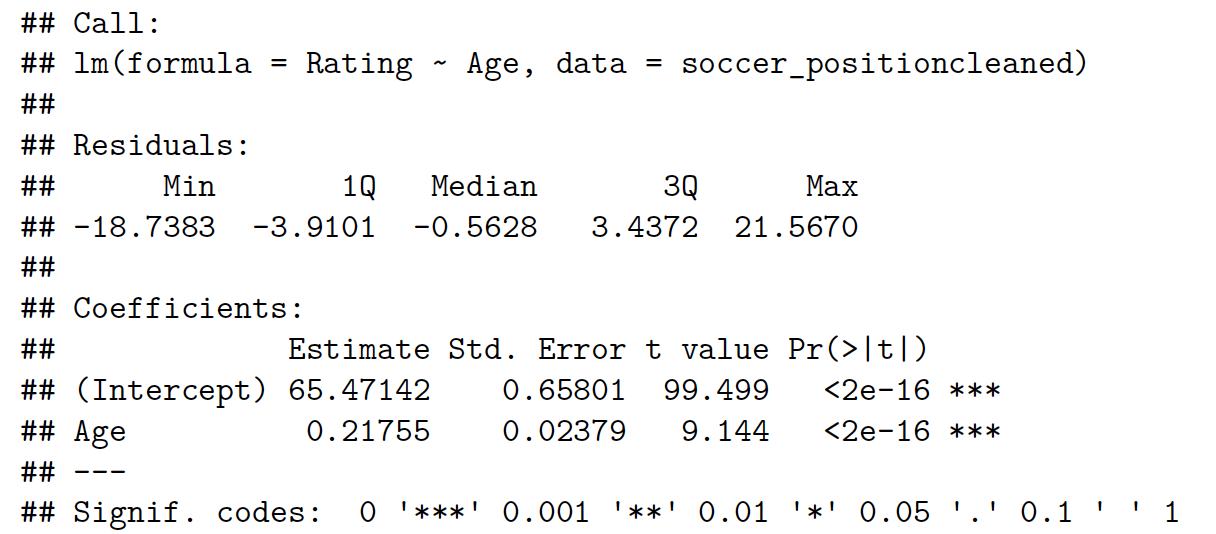
**2.** **Methods**

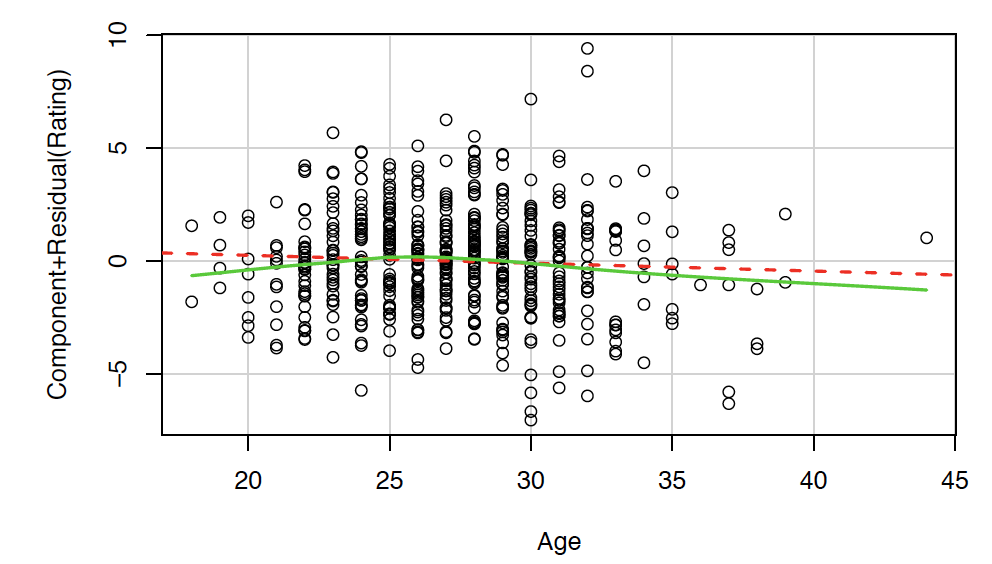
**a) Method (my choice of model, analytic method, why)**

**1. Multivariate test:**

As I mentioned above under the “Problem,” our subject will be 3656 soccer players (but different groups of population). The quantitative variables are ratings and ages (and as we assume ratings and ages are somewhat closely related, we can say that our measurement is generally just ratings. Intuitively, for most of the cases, it is true for ratings and ages. To prove my points, I include the lm() summary. Again, this is not 100% correlated, but I assumed to be), and there are 2 factors: club positions (4 levels) and preferred foot (2 levels). So, I need to use Two-way MANOVA. Our team decides to conduct this test to see if there any difference of means from many groups.

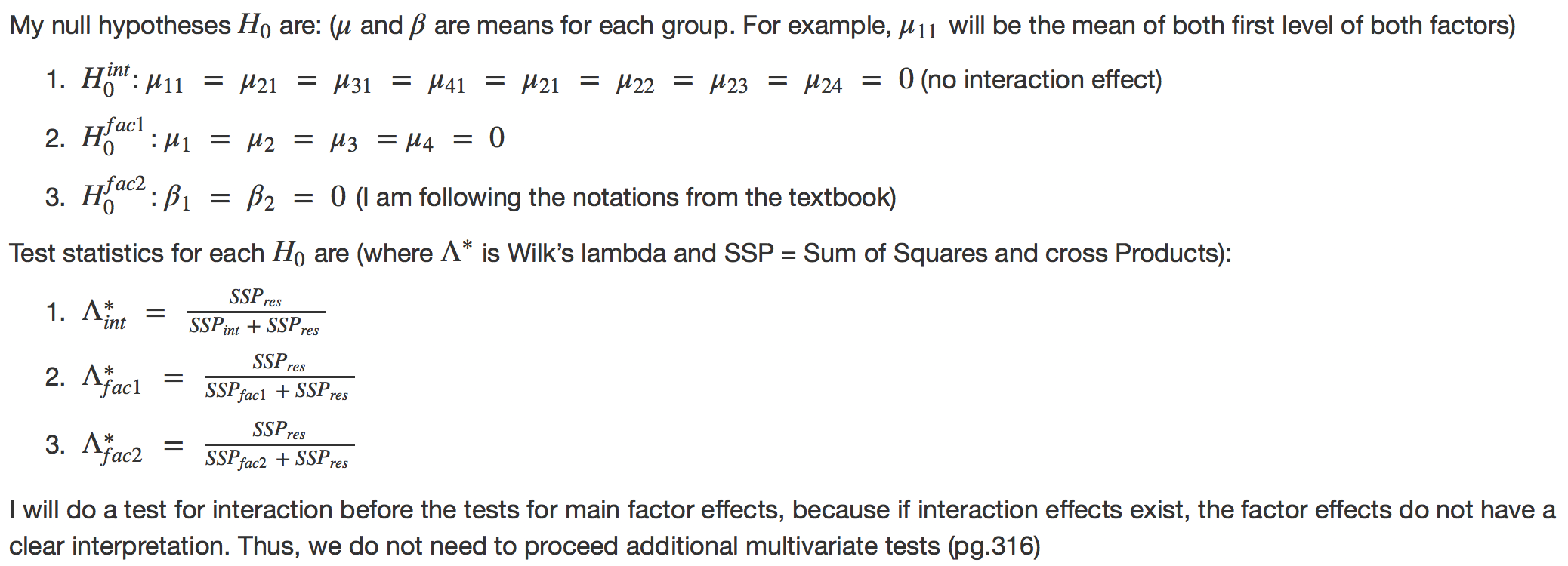
Here is the lm() output, below: (as you can see the variable “Age” is significant, but this is not perfect way, as this is just a linear model with marginal effect. For example, there might be some other variables affecting the relationship between Rating and Age variables, so I decided to Component Plus Residual plot to see better linear relationship between them.)



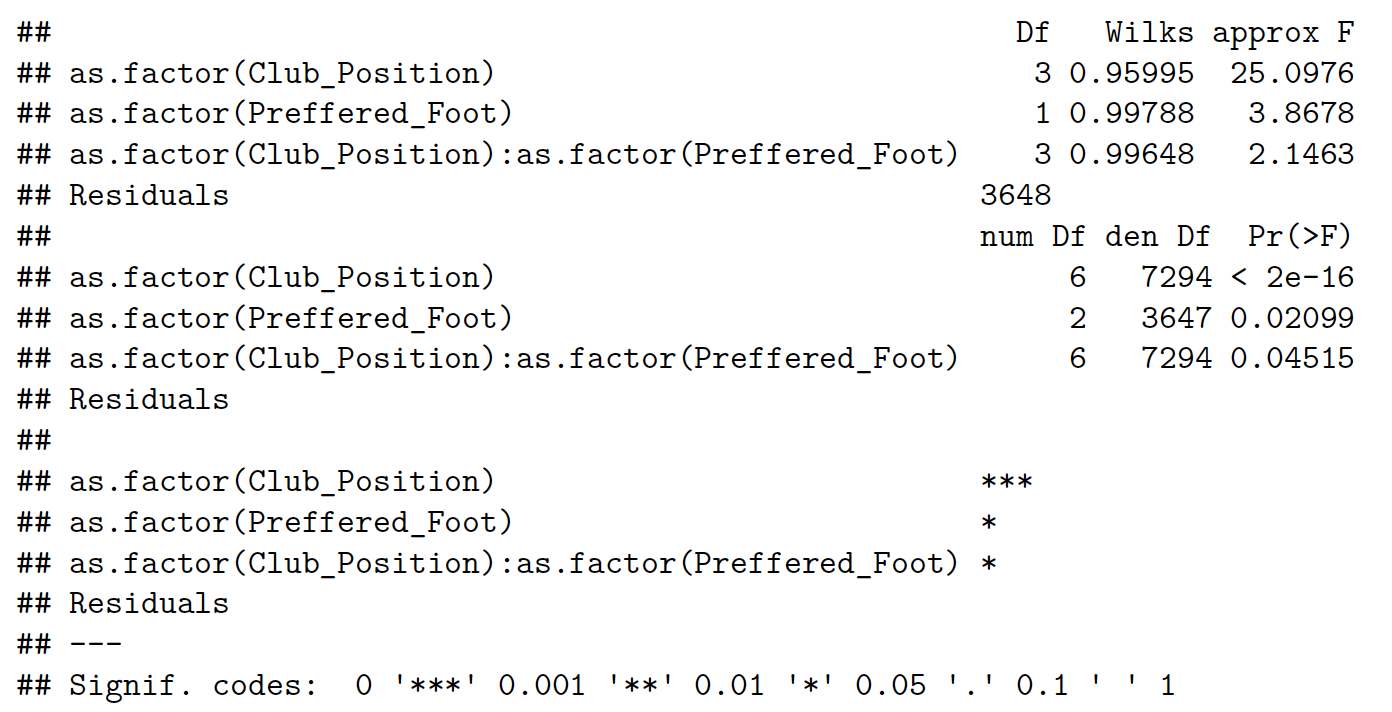


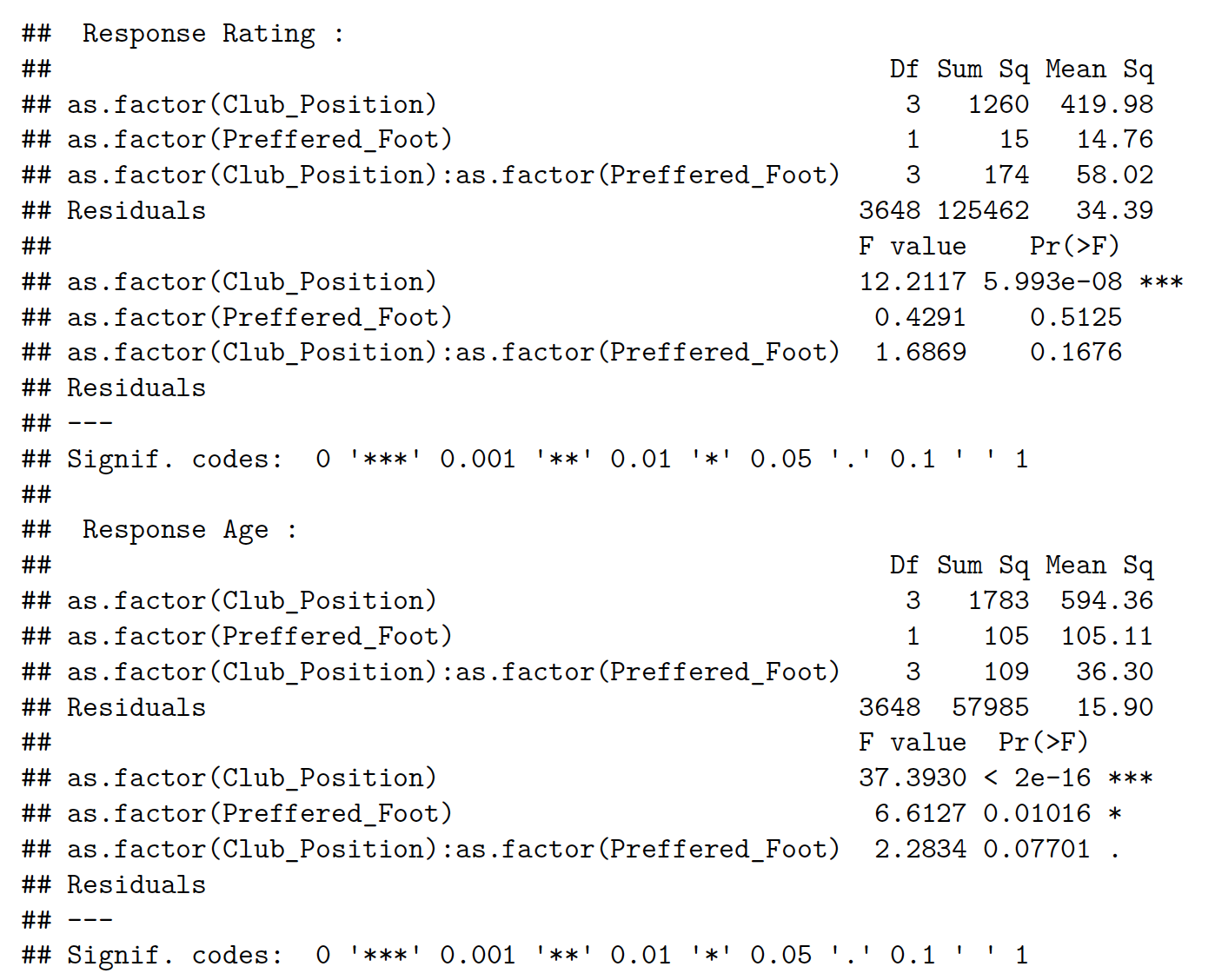
As we can see the test output, we can ambiguously argue that ratings and ages are kind-of related.

Here is our test, below:



Here is our test summary outputs, below:



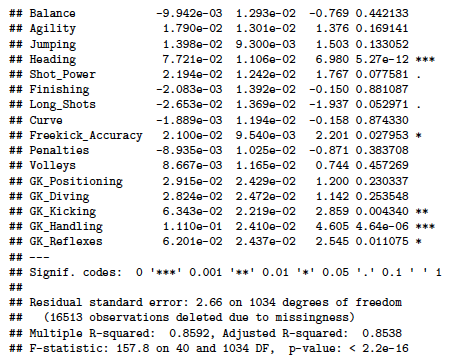
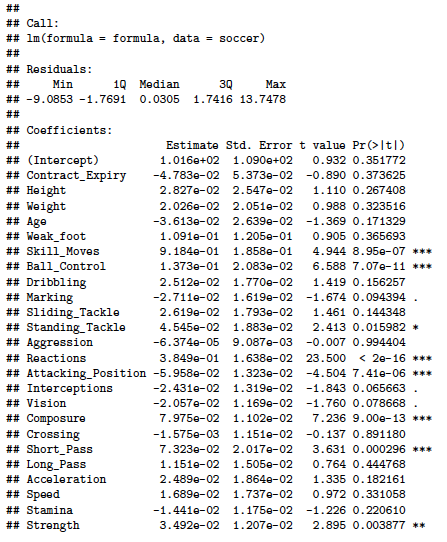


Our conclusion is that we reject all of three null hypothesis. Thus, there is no position effect, preferred foot effect, and position-preferred foot interaction effects on ratings (based on the assumption that ratings and ages are correlated well enough).

**2. Linear regression:**

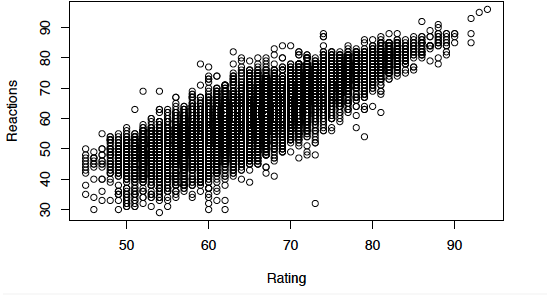
We will use linear regression coefficients, T-test, and F-test to find which predictors(quantitative and qualitative variables in factors) are significant by regressing YYYYY on some XXXXX (you can fill in -Jin).

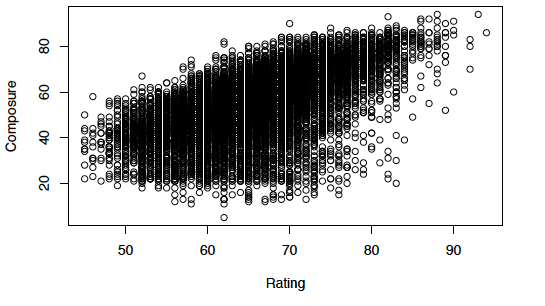
1. Fit a linear regression model with percent body fat using rating as the response and other variable ‘ ‘ as the predictors. Check significant predictors & t-test for significance of each variable



We converted Height and Weight variable into numeric variable and fit a linear regression model using Ratings as the response and the all other quantitative variables as the predictors. The summary of this fit shows interesting result. We expected that most of the score variables(‘Weak\_foot’, ‘Skill\_moves’, …. , ‘GK\_Reflexes’) will have at least 0.01 significance in the beginning. However, the lm results clearly shows that only 10 variables are strongly related to their ratings. (T test) If we set significant level as 0.001, variable ‘Skill\_Moves’, ‘Ball\_Control’, ‘Reactions’, ‘Attacking\_Position’, ‘Composure’, ‘Short\_Pass’, ‘Heading’, ‘GK\_Handling’ are significant among 41 predictors.

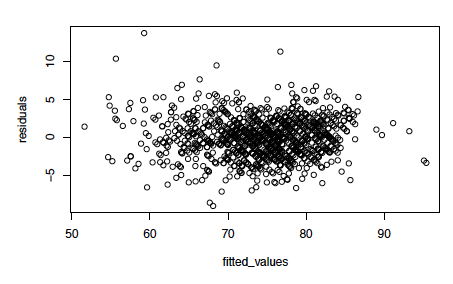
1. Data visualization - shows several plots of significant variables in part (a) VS ratings





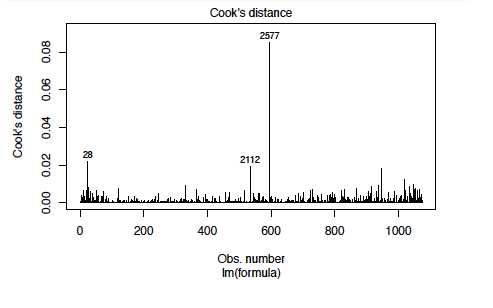
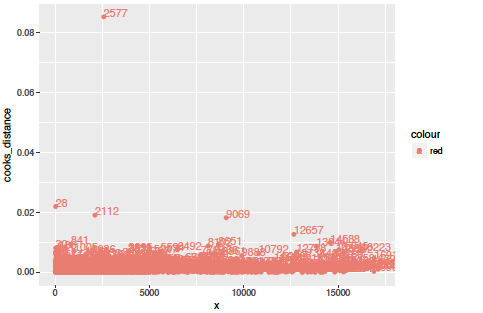
These variable shows clear linear relationship between variable and the response (ratings)

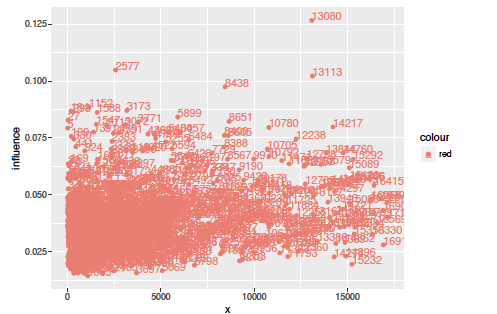
1. Draw a residual plot, with the fitted values on the x-axis, and the residuals on the y-axis to check if there is any violation of assumptions of the linear model. (linear or not, variance constant, normal or not)



The residual plot shows that it does not particularly violate linear model assumption since the residuals are symmetric to y = 0 axis and does not show specific patterns. Also, variance seems like a constant too.

1. Plots of Cook’s distance and influential points to detect outliers





24 534 596 1024

Let’s remove 24th, 2112th, and 2577th players and fit again.

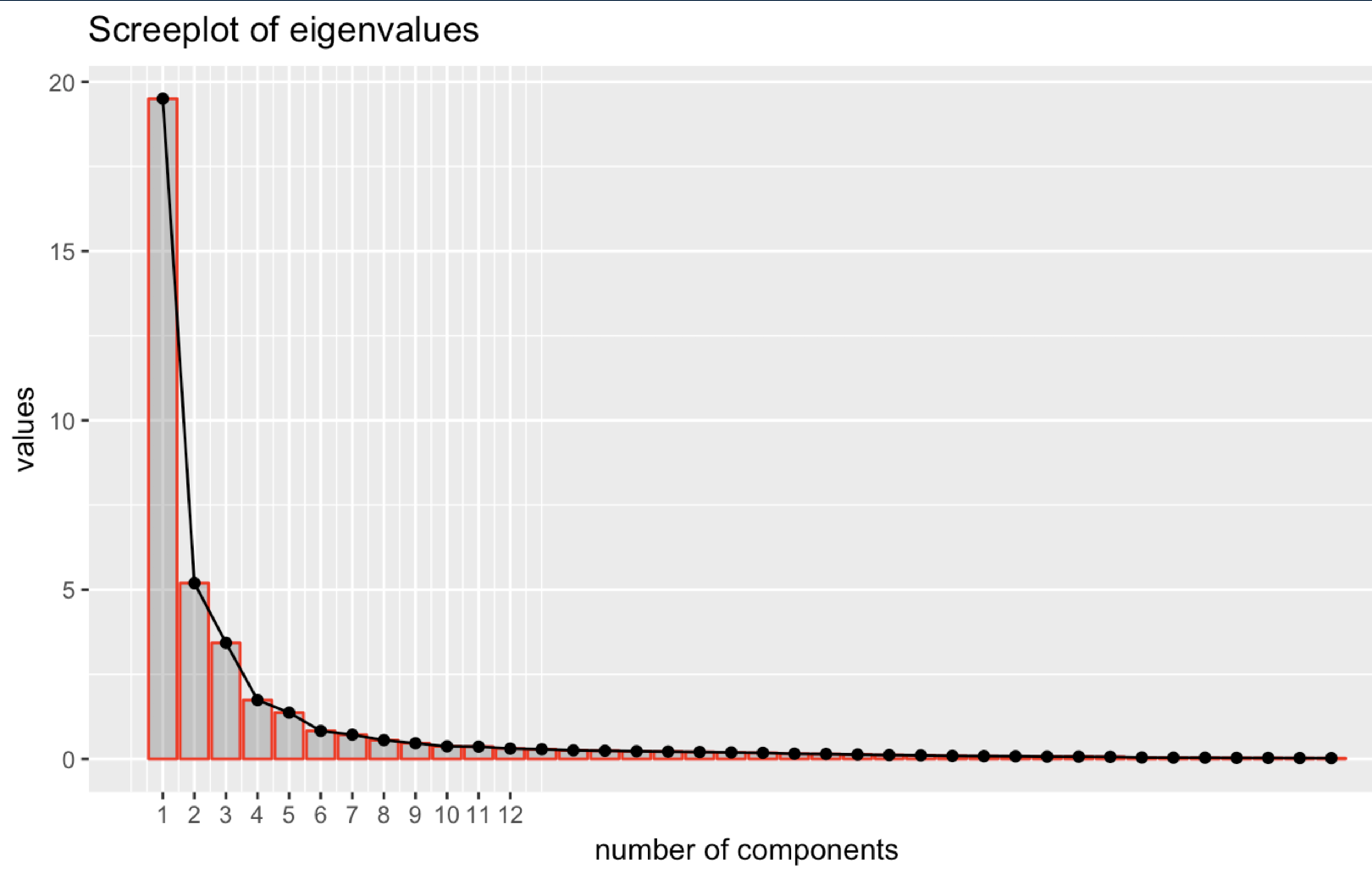
1. Lasso for variable selections…
2. Transform variables using square root and polynomials and find if relationship other than linear is appropriate. Check with residual plot and Component plus residual plot.
3. Compare the model that we build in with the original model → F - statistics
4. Conclude association…

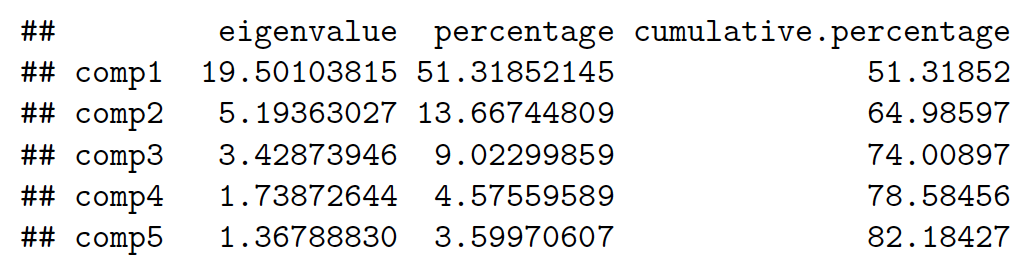
i) also lm onto each position

**3. PCA:**

First of all, we performed the PCA with the correlation matrix. And, we used the 38 variables out of 53 variables, as PCA performs on continuous variables. (again, please refer to the data dictionary for more details)

Here is a scree-plot and percentage of the total sample variance explained by the top (in terms of variance explained) 5 components, we got:





It seems like (based on the elbow on scree-plot and Kaiser or Jolliffe's rules, it is recommended to keep up to around 7 or 8 components, but I will mostly use up to three components for our analysis.

**3. I want to test how much each variable is correlated with the rating. PCA will also help me find co-linearity issue.**

Reason: We aim to find PCs to best summarize the variables, and see how players’ rankings are plotted.

Hypothesis: Different positions have different variables correlated with the rating, and I should find the skills that are important to the position should have high correlation with the rating.

Aim/objective: I hope to make good interpretation of the components to explain the the relationships between variables, and eventually help us how rating can be explained by other variables.

**3.** **Result and Summary**

**a) Results (summary of numerical analysis, interpretation, assumption check)**

**b) Conclusion**

**c) How it can be further developed**

We separate data into four big positions, and this is pretty decent categorization. However, this might need further improvement for better analysis. It is because every team has different tactic and philosophy, and player positions can be more complex than what it is. For example, RWB (right wing back) players are required to play as midfielder and even attacker for some teams. One great example you could find if you are interested in: <https://en.wikipedia.org/wiki/Total_Football>. So, our grouping is arbitrary, and this can be different.

**References**

<https://www.kaggle.com/artimous/complete-fifa-2017-player-dataset-global>

**Appendix**

Please refer to the link for the R codes: