Contributions

1 Introduction

2 Data Description

The data of players in FIFA20 was originally found on Kaggle https://www.kaggle.com/datasets/stefanoleone992/fifa-20-complete-player-dataset?select=players_20.csv. The data of the players was collected by EA throughout the year based the players' performance in game.

The column selected is overall, shooting, passing, physic, wage and international_reputation. We eliminated all goal keepers and randomly selected 30% of original observation, and replaced 0 salary with 1 to analyse.

Responsive Varible

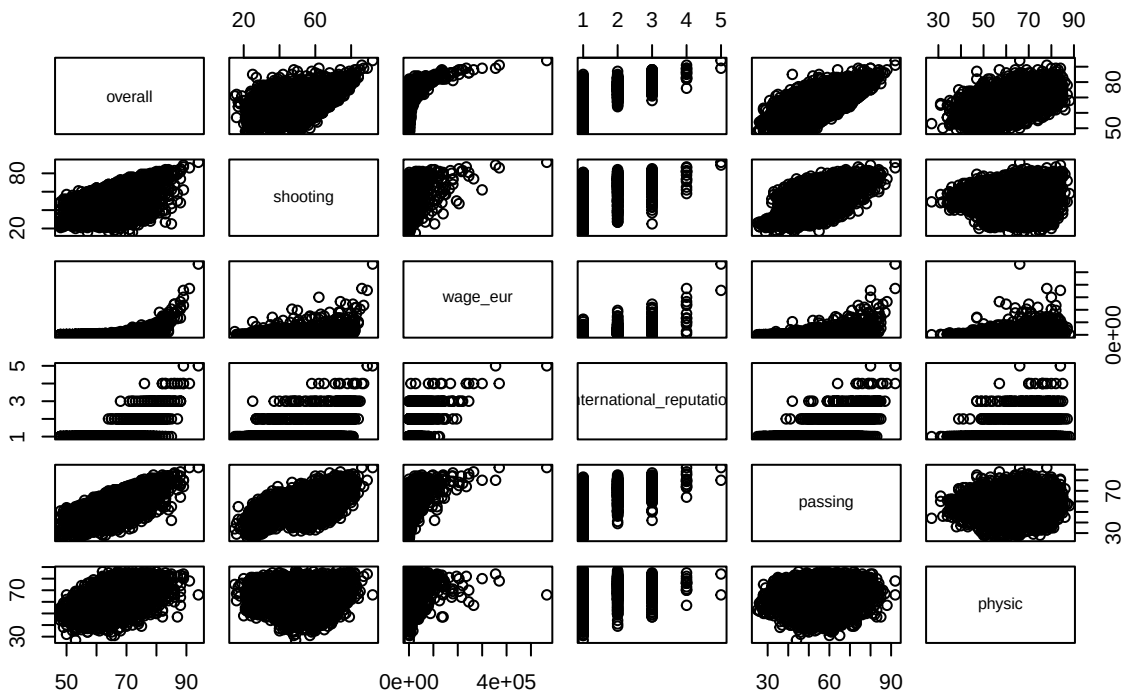
The response variable used in this research was "overall", it is the overall attribute of one player. This variable measures the players overall rating in the game, ranging from 0 to 100.

Table 1: Descriptive Statistic of Response Variable

min	max	mean	median	standard_deviation
48	94	66.43309	66	6.819944

The mean of the overall is 66.43

The responsive variable "overall" is the overall rating of one player. We believe the overall rating of a player is largely influenced by their attributes like shooting, wages, international_reputation, passing, physic. As our understanding to soccer, when the player gets higher score in these attribute, their overall rating should also be higher. We think the changes in predictors should result in proportional changes in the response variable. Therefore, we take overall as the responsive variable is appropriate.



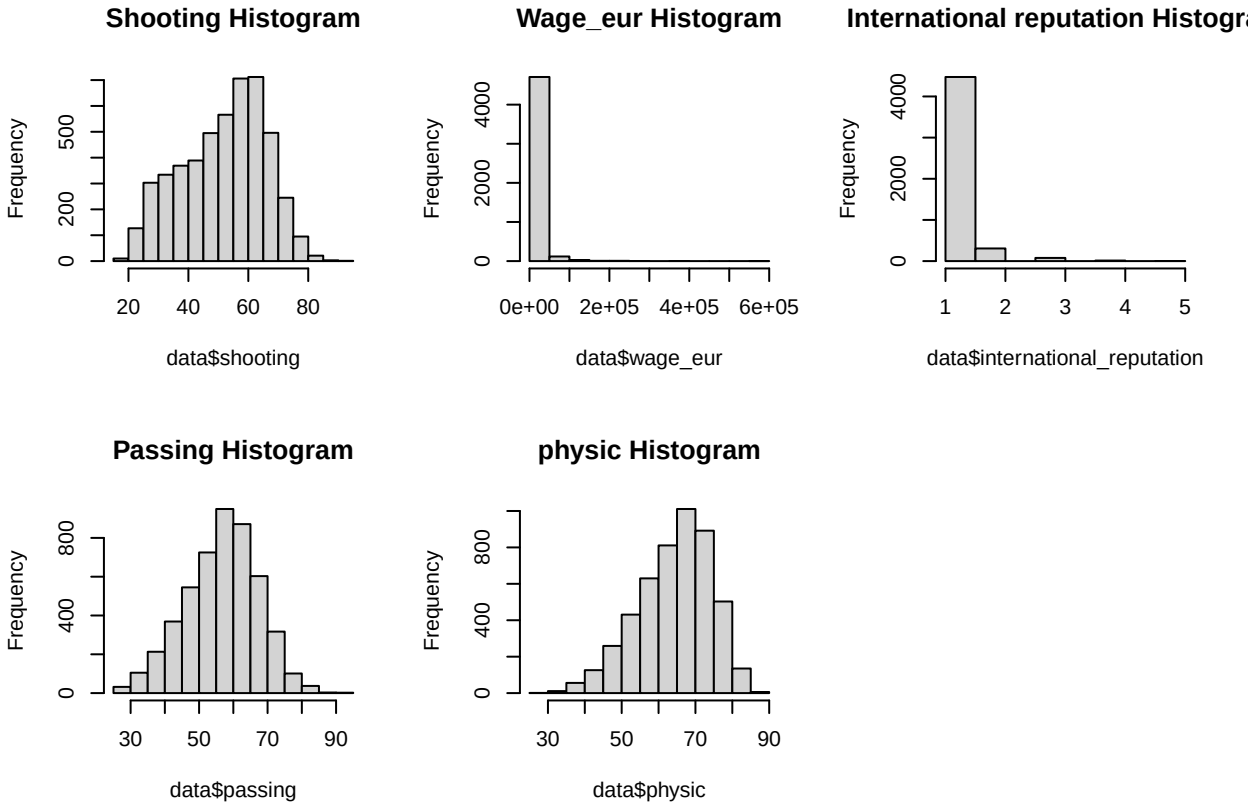

```
## bcPower Transformations to Multinormality
##                            Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## overall                       2.4109        2.41       2.2453       2.5765
## shooting                      1.3688        1.37       1.2734       1.4642
## wage_eur                      0.2149        0.21       0.2049       0.2248
## international_reputation  -15.2877      -15.29     -15.7175     -14.8579
## passing                       1.9380        2.00       1.8260       2.0501
## physic                        2.2949        2.29       2.1432       2.4465
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                       LRT df       pval
## LR test, lambda = (0 0 0 0 0 0) 20116.34  6 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                       LRT df       pval
## LR test, lambda = (1 1 1 1 1 1) 34743.88  6 < 2.22e-16
```

The response variable has an approximately linear relationship with some of the predictor variables. meaning that as the predictors change, the response variable changes in a consistent, straight-line fashion. The international_reputation was originally a continuous predictor ranging from 1 to 5, we changed it to categorical predictor with 4 level(2, 3, 4, 5). By looking at the Box-cox transformation, we noticed that we need to apply transformations to the predictors to improve our model. Here we take wage_eur a fouth-root transformation, take passing to power of 2, physic to the power of 2 for simplicity purpose.

Predictors

Table 2: Descriptive Statistic of Predictors

predictors	min	max	mean	median	standard_deviation
shooting	15	92	52.176108	54	1.399594e+01
wage_eur	1	565000	9769.100780	3000	2.238749e+04
international_reputation	1	5	1.104269	1	3.820632e-01
passing	25	92	57.128284	58	1.055442e+01
physic	27	88	64.792693	66	9.850584e+00



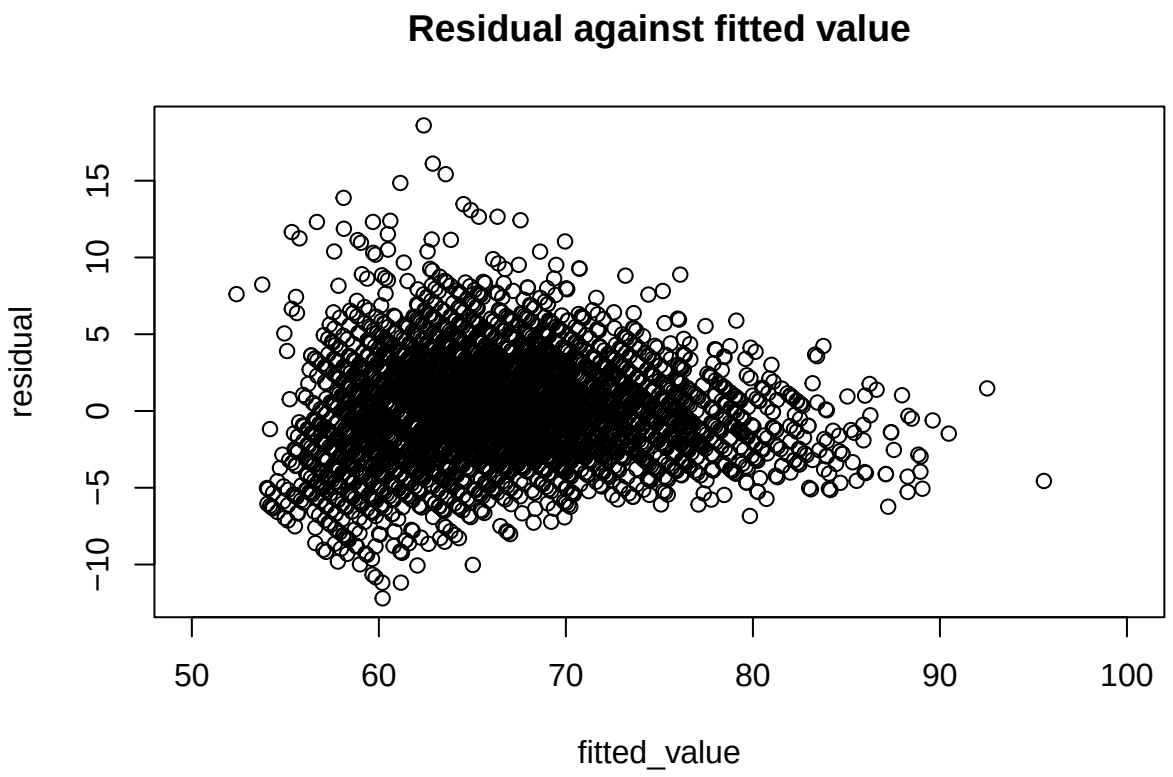
- shooting measures the player's shooting ability, include strength, accuracy etc.
- wage_eur measures the player's weekly salary in Euro.
- international_reputation measures the player's popularity in the world.
- passing measures the player's passing ability, include strength, accuracy etc.
- physic measures the player's physical condition, include stamina, jumping etc.

Through the graph we can see that the distribution of passing is approximately normal, physic and shooting are slightly left skewed, international reputation and wage are extremely right skewed.

Each of these predictor contributes to the construction of the overall rating, a higher score means the player is better at this aspect, which should make the overall rating higher.

Preliminary Results

Model Assumptions




```
##
## Call:
## lm(formula = overall ~ shooting + wage_eur_transformed + international_reputation +
##     passing_transformed + physic_transformed, data = transform_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.202  -2.118  -0.043   2.015  18.601
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                4.223e+01  2.540e-01 166.240  < 2e-16 ***
## shooting                   3.694e-02  4.446e-03   8.309  < 2e-16 ***
## wage_eur_transformed       7.816e-01  2.141e-02  36.499  < 2e-16 ***
## international_reputation2  1.463e+00  2.125e-01   6.885 6.51e-12 ***
## international_reputation3  2.575e+00  3.999e-01   6.440 1.31e-10 ***
## international_reputation4  3.028e-01  9.283e-01   0.326    0.744
## international_reputation5 -1.853e+00  2.326e+00  -0.797    0.426
## passing_transformed        2.261e-03  5.772e-05  39.178  < 2e-16 ***
## physic_transformed         1.880e-03  4.050e-05  46.413  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3.254 on 4863 degrees of freedom
## Multiple R-squared:  0.7728, Adjusted R-squared:  0.7724
## F-statistic:  2067 on 8 and 4863 DF,  p-value: < 2.2e-16
```

The residual plot has some outlier on the right side, but overall the plot doesn't show a clear patern. This implies the population satisifies Linearity, constant variance, and uncorrelated errors assumptions.

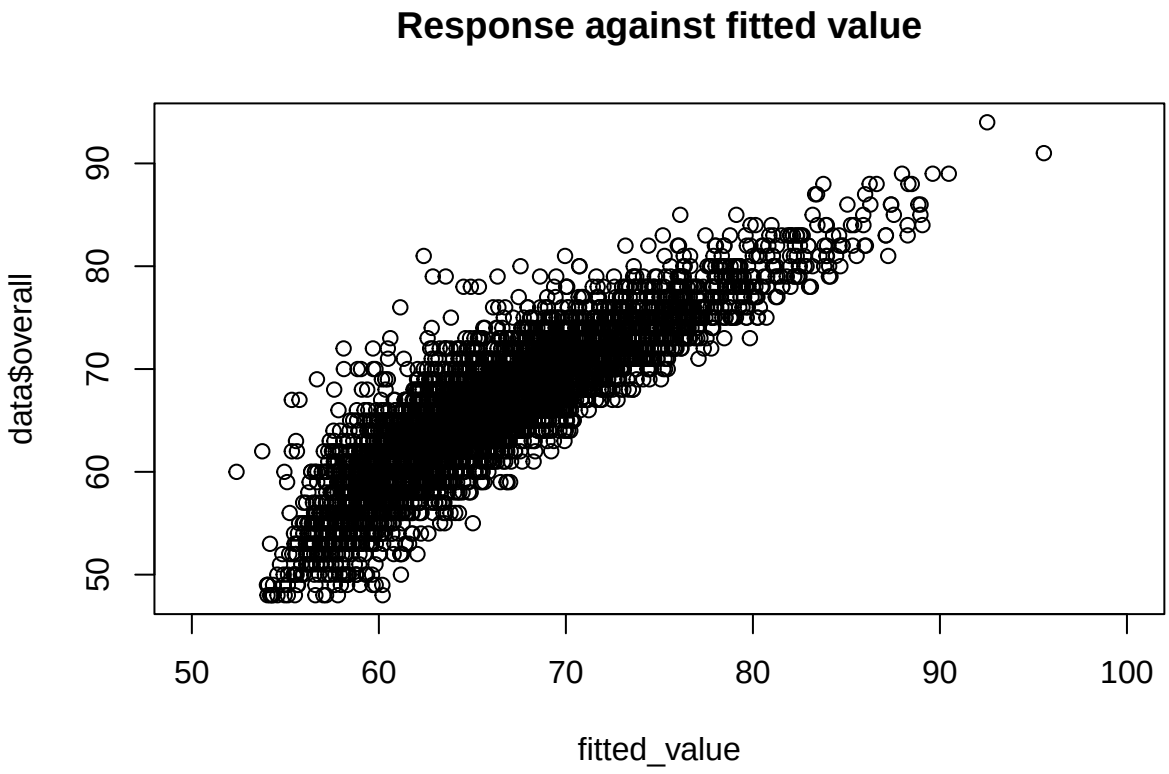
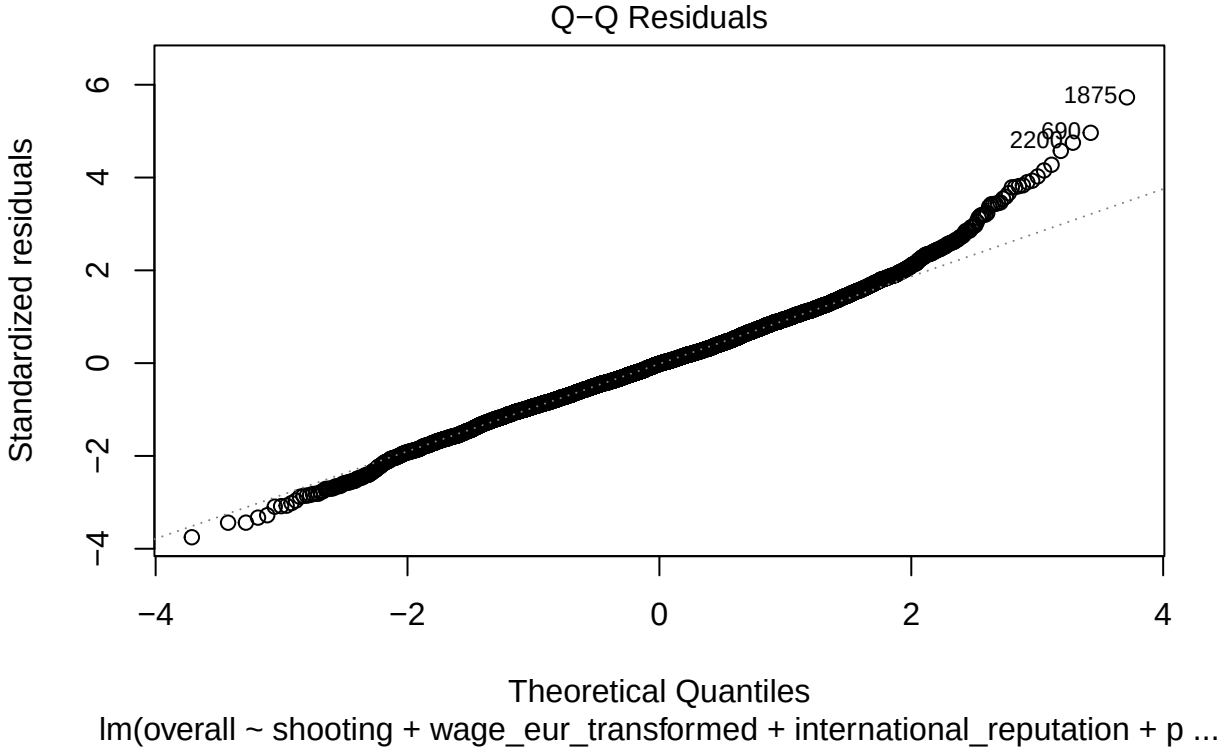


Figure 1: Response against fitted value test

Points closely follow the 45-degree line (i.e., observed ≈ predicted), the model is likely performing well. The mean responses are a single function of a linear combination involving coefficients. This satisfies Linear assumption.



The most points on the QQ plot are on a straight line, with some outliers on each side. Overall this implies the population satisfies the normal errors assumption.

References