Analysis of the rugby players

Student Number: 2516212c

Introduction

In the sport of rugby there are many different positions a player could play but, in general, players are classified into one of two playing positions, namely 'Forward' and 'Back'. From the data collected on rugby players from the 2015 Men's Rugby World Cup, we want to figure out the relationship between position and heights. 89 rugby players were used in the research and they each had their own position and heights(in cm). This data is analysed in this report.

Exploratory Data Analysis

Summary statistics of the heights of the rugby players are presented in the following table for each kind of position separately.

Table 1: Summary statistics on height(in cm) by position of 89 rugby players.

position	n	Mean	St.Dev	Min	Q1	Median	Q3	Max
Back	33	183	4.4	173	180	183	185	193
Forward	56	181	3.6	173	180	181	183	190

This table shows that there were almost twice as many rugby players in the sample (56 compared to 33) and that the summarises of the heights of the rugby players which position is 'Back' seem has no remarkable difference with these rugby players who played 'Forward'. But from the median, the summarises of 'Back' was greater than 'Forward' although the difference was not large. For example the maximum of 'Back' was taller than the maximum of 'Forward' 3 cm and from the Standard deviation, we could see that the spread of 'Back' seems widely than the spread of 'Forward'. These differences can be seen directly in the following boxplots which summarise the distribution of the heights of each kind of players.

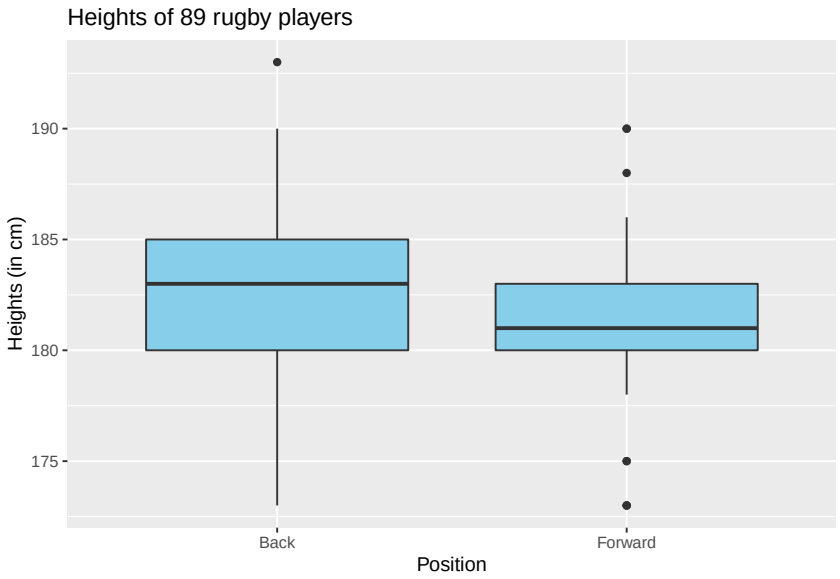

Figure 1: Boxplot:Heights by Position

The boxplot shows that the rugby players' heights of 'Back' is larger than 'Forward', in general, compared to the players' heights of 'Forward' and that the players' heights of 'Back' were more widely distributed. There are also potentially five outliers and most of them belong to 'Forward' which have special heights(quite not same with the mean), as shown by these points far away from the "whiskers" of the boxplots.

Formal Data Analysis

To begin to analyse the heights of players data formally, we fit the following linear model to the data.

$$\widehat{\mathrm{H}} = \widehat{\alpha} + \widehat{\beta}_{\mathrm{f}} \cdot \mathbb{I}_{\mathrm{f}}(x)$$

Where

*$\widehat{\mathrm{H}}$ is the expected value of the height of the ith player in the sample;

*the intercept $\widehat{\alpha}$ is the mean height for the baseline category of 'Back';

*$\widehat{\beta}_f$ is the difference in the mean height of a 'Forward' player relatively to the baseline category 'Back';

*$\mathbb{I}_f(i)$ is an indicator function such that

$$\mathbb{I}_{\mathrm{f}}(i) = \begin{cases} 1 & \text{if position of } i\text{th observation is forward,} \\ 0 & \text{Otherwise.} \end{cases}$$

When this model is fitted to the data, the following estimates of α(intercept) and $\beta_f(Posf)$ are returned:

Table 2: Estimates of the parameters from the fitted linear regression model.

term	estimate
intercept	182.697
positionForward	-1.447

Hence the model estimates the average height of players in 'Back' is 182.697 cms (which agrees with the sample mean reported and that the 'Forward' players' heights are, on average, 1.447 lower than the 'Back' players' heights.).

Before we can proceed to use the fitted model(for example to perform statistical inference) we must check the assumptions of the model. These are best considered in light of the residual plots in Figure 2.

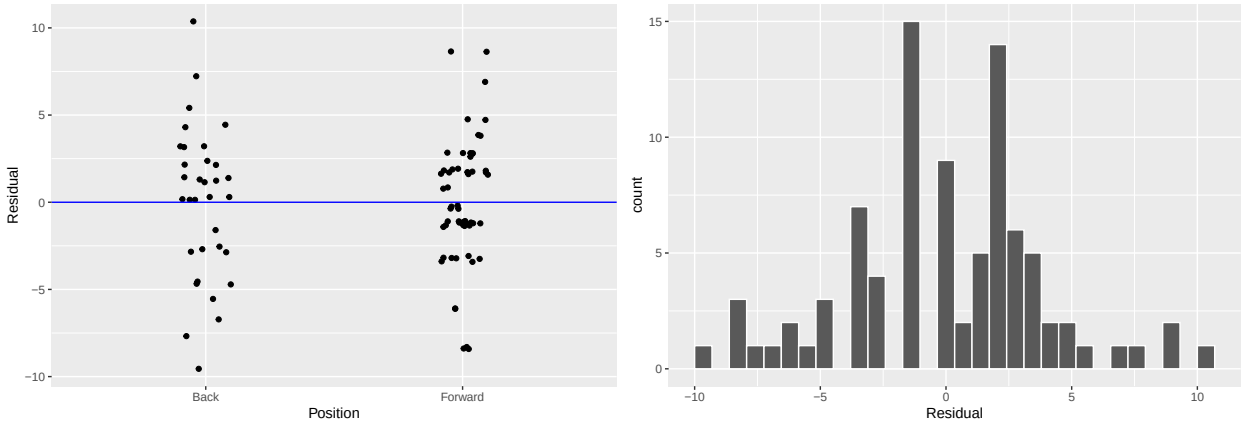

Figure 2: Scatterplots of the residuals by Position(left) and a histogram of the residuals (right)

The scatterplots show an approximately even spread of the residuals above and below the zero line for each position, therefore the assumption that the residuals have mean zero seems valid and the assumption that constant variance within the two positions seems available too in terms of the first scatterplot. The histogram shows the distribution of residuals are nomally distributed errors in the model, which the exception of a potential outliers.

Conclusions

In summary, we have estimated that, on average the 'Back' players have heights which is 2 centimeters taller than the 'Forward' players. In particular, we estimate the average height of 'Back' players is 182.697 centimeters and the average height of 'Forward' players is 181.25 centimeters.

In addition to the centers of the distributions of 'Back' and 'Forward' players being different, we have also observed that the spread of the both position seems familiar and without remarkable difference. So based on the fitted linear model, there is no statistically significant difference in the heights between these two positions.

FURTHER TASK

a)

```
rugby <- read.csv("rugby_full.csv")
rugby <- rugby %>%
  mutate(games_per_year = round(caps/years_playing,digits = 1))
```

b)

```
rugby_tidy <- gather(data = rugby,
key = Type_of_Age, value = Age,
c(age,age_debut))
```

c)

```
rugby_tidy %>%
  group_by(team) %>%
  ggplot(aes(x = caps,y = years_playing))+
  geom_point(mapping = aes(pch = position,color = position))+
  labs(x = "Captitals", y = "Years for playing", title = 'The relationship between caps and year_playing
  facet_grid(.~team)
```

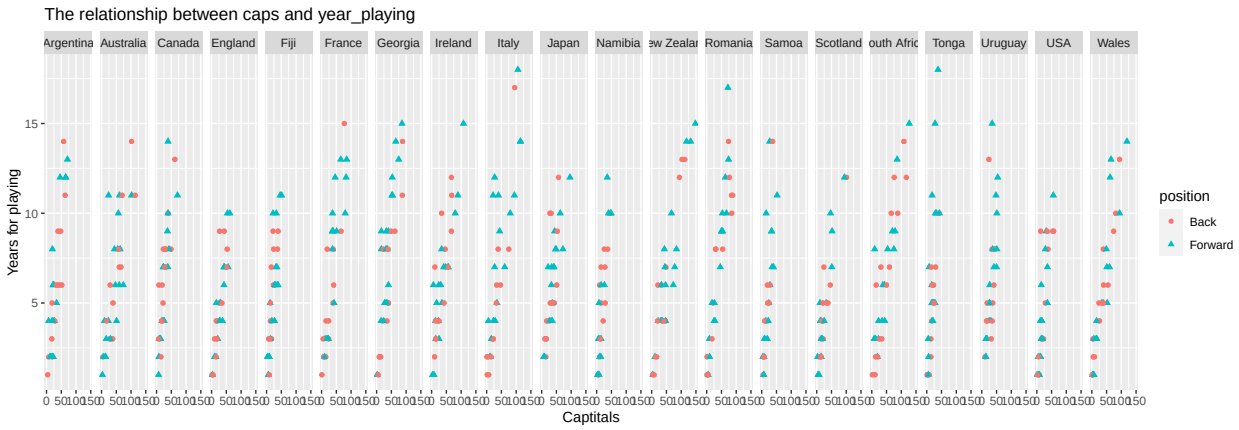

Figure 3: Scatterplots of the working years caps

d)

```
rugby_tidy %>%
  filter(team == c("England","Ireland","Scotland","Wales")) %>%
  group_by(position) %>%
  ggplot(aes(x = height_cm, y = weight_kg))+
  geom_point(mapping = aes(color=position,pch = position),size = 3)+
  geom_smooth(mapping = aes(lty = position,color=position),lwd = 1.5, se = FALSE,method = "lm")
```

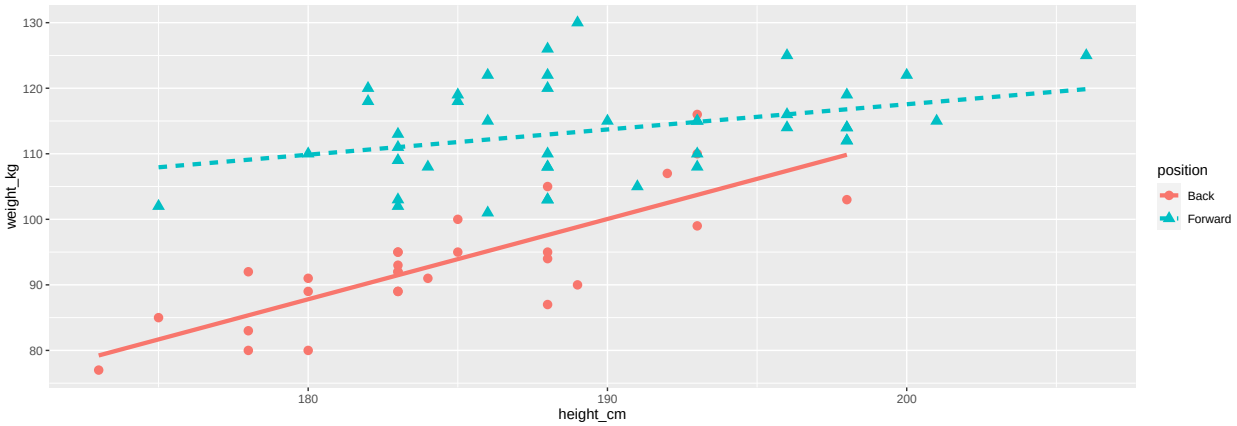

Figure 4: Scatterplots of the Weight(kg) by Height(cm)