Baseball Level Analysis

Jaden Thomas

2024-02-05

Problem Description

To be able to determine what level of baseball a player plays at by the individuals height and weight, a logistic regression model is fit.

The model formula is:

$$p(X) = \frac{e^{\beta_0 + \beta_1 HT_i + \beta_2 WT_i}}{1 + e^{\beta_0 + HT_i + WT_i}}; i = 1, ..., n$$

Or,

$$\log(\frac{p(X)}{1 - p(X)}) = \beta_0 + \beta_1 HT_i + \beta_2 WT_i; i = 1, ...n$$

Where HT is in inches, WT is in lbs, and the two classes are MLB and Club, with MLB labeled as 1 and Club baseball labeled as 0.

Import Necessary Packages

```
library(tidyverse)
library(ggthemes)
library(verification)
```

Set Seed

```
set.seed(2024)
```

Import Datasets

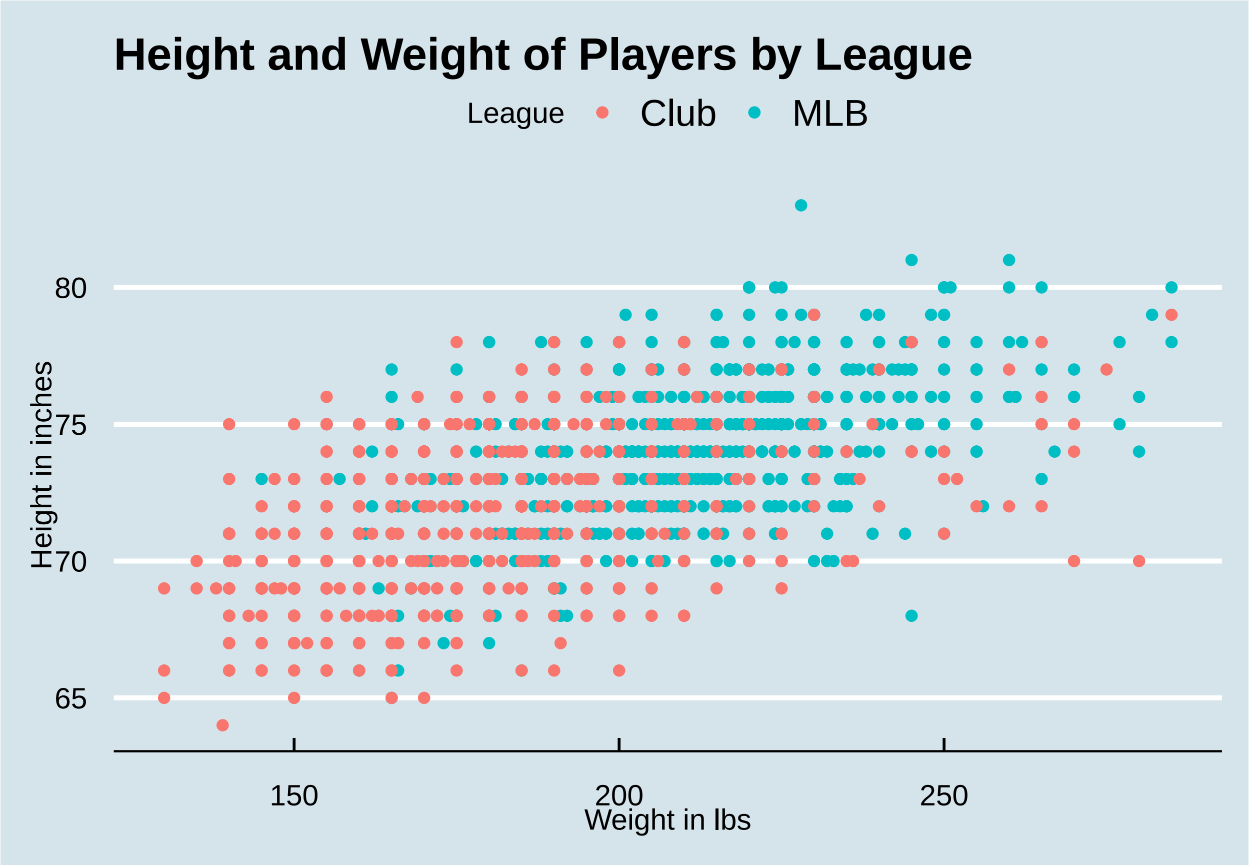
```
mlb_2023 <- read_csv("mlbBaseballPlayers_2023.csv")
club_2024 <- read_csv("clubBaseballPlayers_2024.csv")
```

Format Data

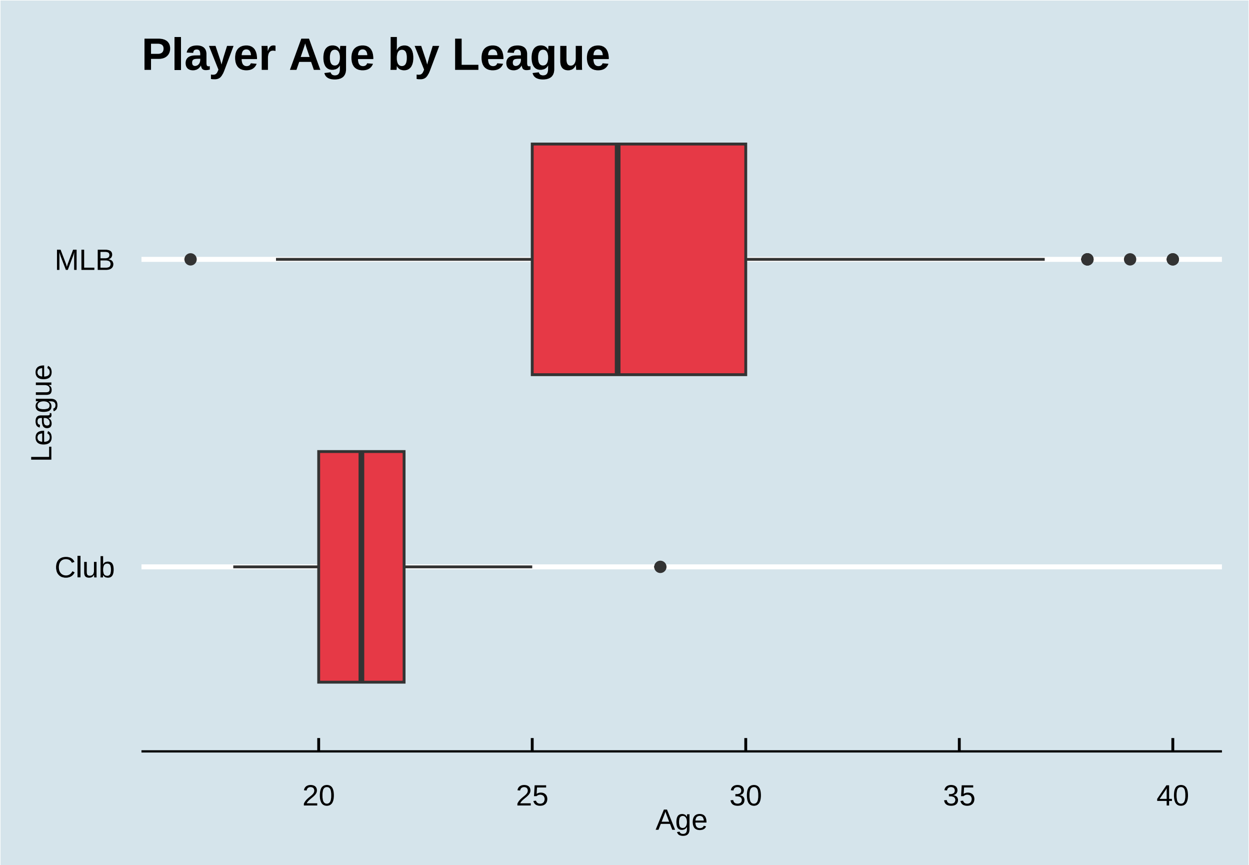
```
club_2024 <- club_2024 %>%
  mutate(Age=(year(Sys.Date())-year(as.Date(paste0("01/", club_2024$DOB), format="%d/%m/%Y"))),
         POS=str_split_i(POS, pattern=" / ", i=1)) %>%
  rename(BAT=Bats, THW=Throws)

full <- bind_rows(mlb_2023, club_2024)

sub <- full %>% dplyr::select(fname, lname, Age, HT, WT, city, state, POS, Team, League) %>%
  mutate(League=as.factor(League)) %>% mutate(MLB=(League=="MLB"))
```

EDA

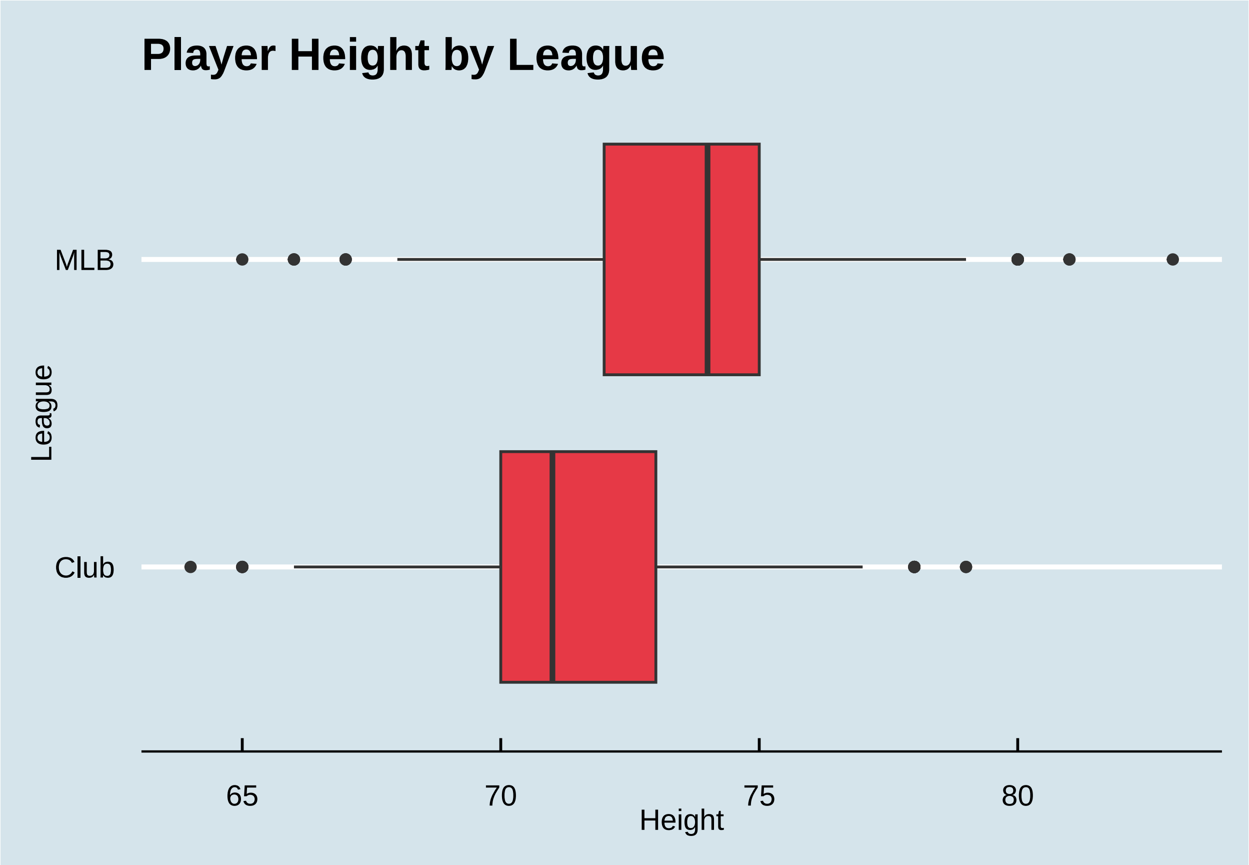
```
ggplot(sub, aes(x=WT, y=HT)) +
  geom_point(aes(color=League)) +
  labs(title="Height and Weight of Players by League", x="Weight in lbs",
       y="Height in inches") +
  theme_economist()
```



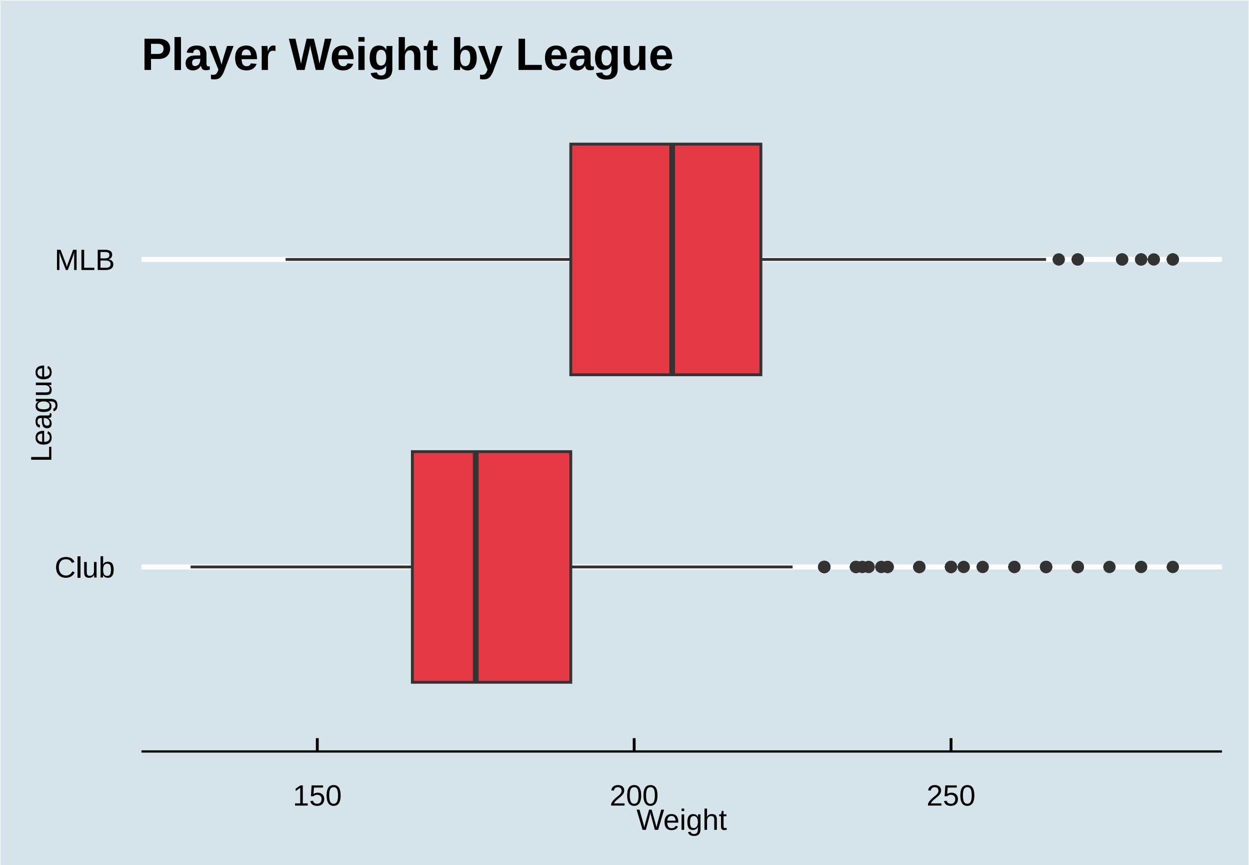
```
sub %>% ggplot(aes(x=Age, y=League)) +
  geom_boxplot(fill="#e63946") +
  theme_economist() +
  labs(title="Player Age by League", x="Age")
```



```
sub %>% ggplot(aes(x=HT, y=League)) +
  geom_boxplot(fill="#e63946") +
  theme_economist() +
  labs(title="Player Height by League", x="Height")
```



```
sub %>% ggplot(aes(x=WT, y=League)) +
  geom_boxplot(fill="#e63946") +
  theme_economist() +
  labs(title="Player Weight by League", x="Weight")
```



```
summary(sub)
```

```
##     fname              lname                Age              HT       
##  Length:2463        Length:2463        Min.   :17.00   Min.   :64.00  
##  Class :character   Class :character   1st Qu.:21.00   1st Qu.:71.00  
##  Mode  :character   Mode  :character   Median :23.00   Median :73.00  
##                                        Mean   :24.21   Mean   :72.61  
##                                        3rd Qu.:27.00   3rd Qu.:74.00  
##                                        Max.   :40.00   Max.   :83.00  
##        WT          city              state               POS           
##  Min.   :130   Length:2463        Length:2463        Length:2463       
##  1st Qu.:175   Class :character   Class :character   Class :character  
##  Median :190   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :193                                                           
##  3rd Qu.:210                                                           
##  Max.   :285                                                           
##      Team             League        MLB         
##  Length:2463        Club:1242   Mode :logical  
##  Class :character   MLB :1221   FALSE:1242     
##  Mode  :character               TRUE :1221     
##                                                
##                                                
## 
```

```
sub %>% group_by(League) %>% summarise(meanHT=mean(HT), meanWT=mean(WT), meanAGE=mean(Age), n=n())
```

```
## # A tibble: 2 x 5
##   League meanHT meanWT meanAGE     n
##   <fct>   <dbl>  <dbl>   <dbl> <int>
## 1 Club     71.5   179.    21.1  1242
## 2 MLB      73.7   207.    27.4  1221
```

As seen by the graphs and numerical summary, the average height and weight for MLB players seems to be higher than that of Club baseball players.

Logistic Regression Model

Train Test Split

```
train <- sample(c(TRUE, FALSE), nrow(sub), replace=T, prob=c(0.7, 0.3))
sub.train <- sub[train,]
sub.test <- sub[!train,]
Y.test <- sub.test$MLB
```

Model Fitting

```
m.fit <- glm(MLB~HT+WT, data=sub.train, family="binomial")
```

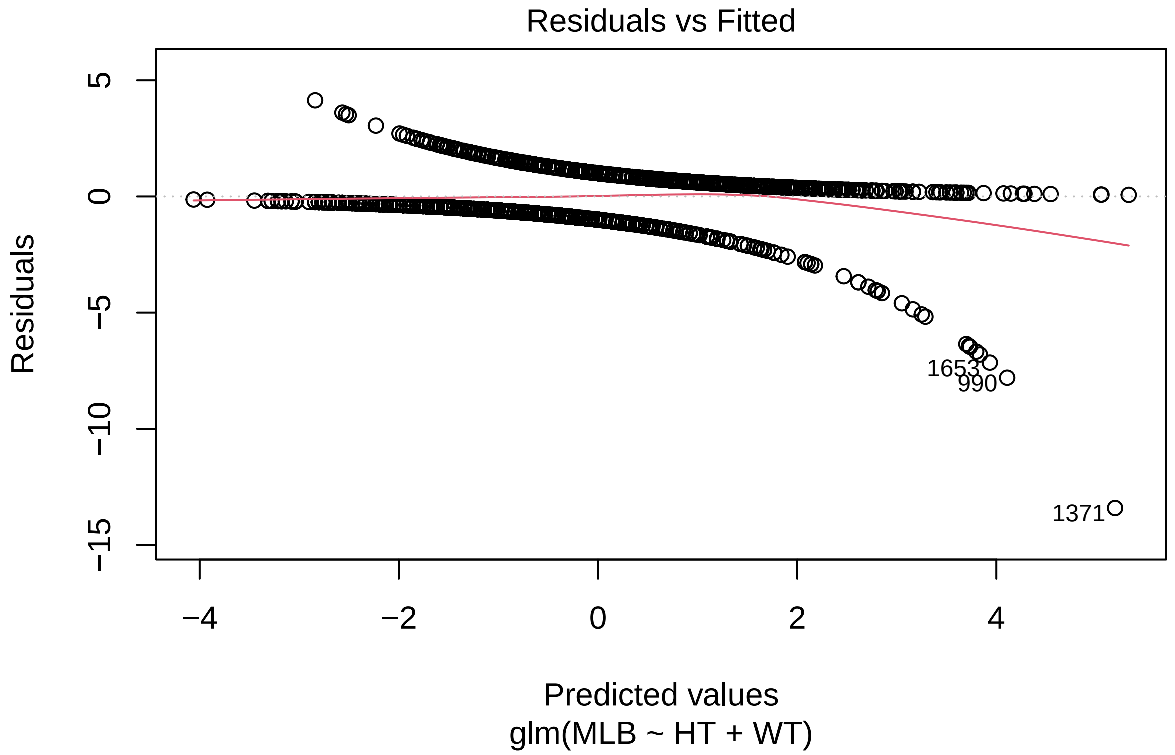
Model Summary

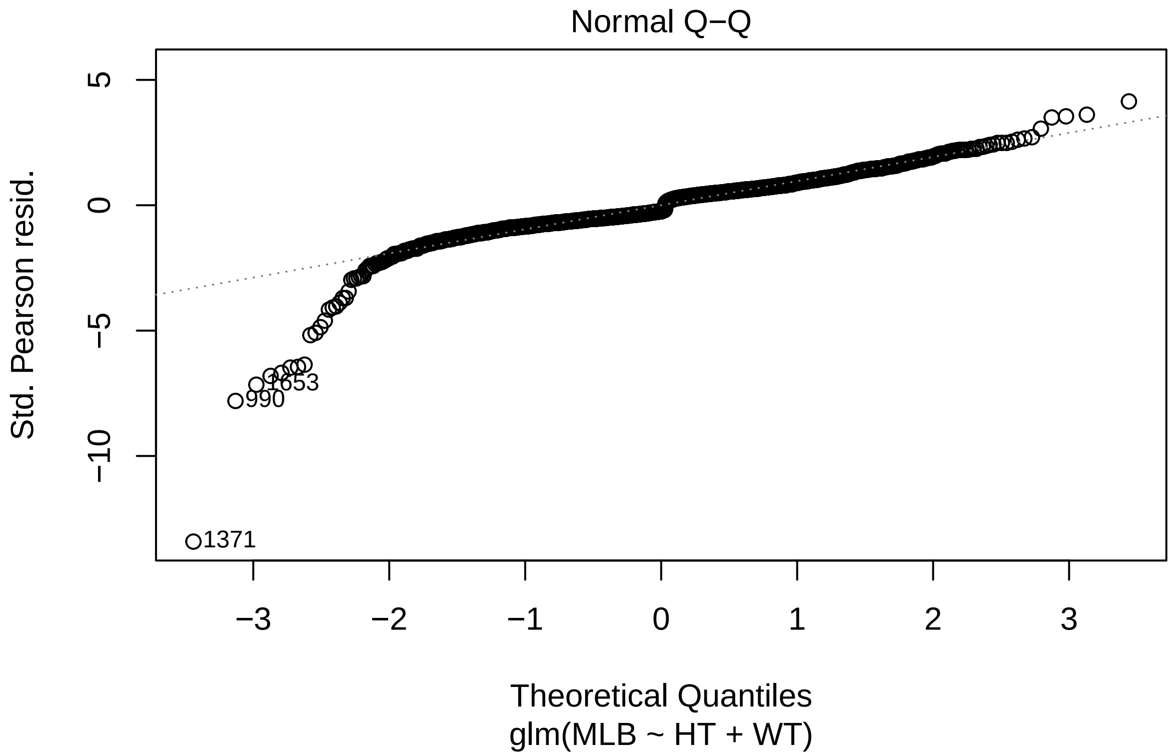
```
summary(m.fit)
```

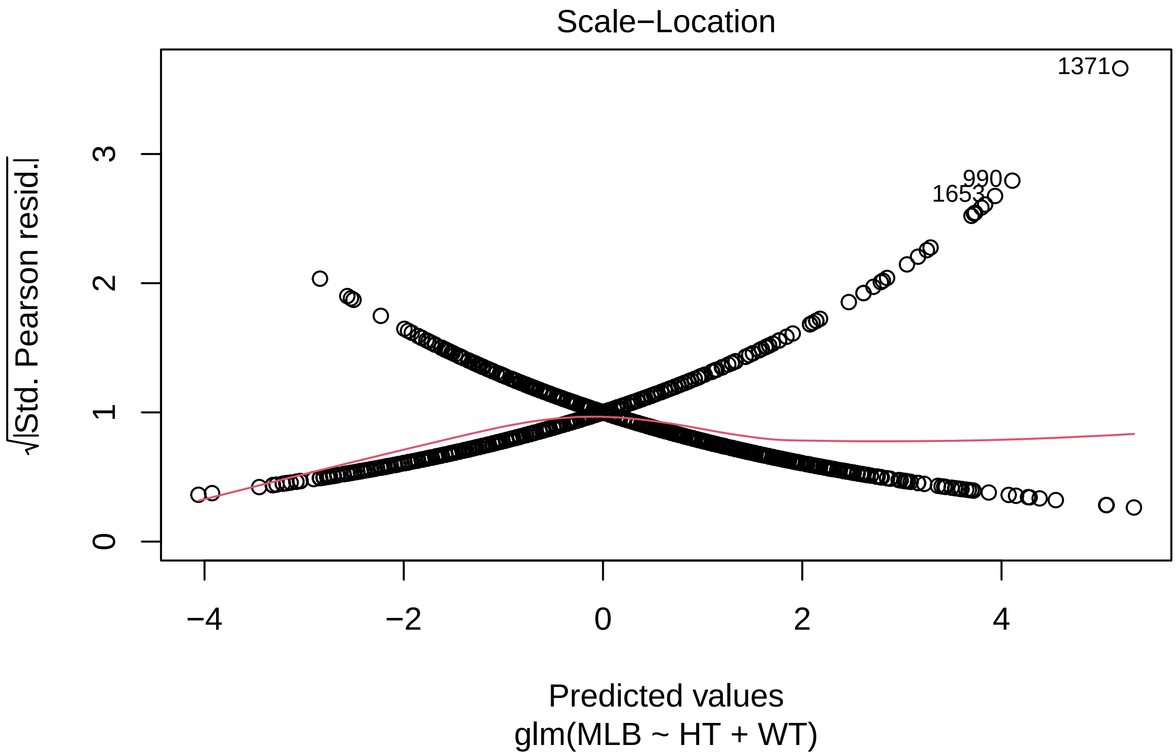
```
##
## Call:
## glm(formula = MLB ~ HT + WT, family = "binomial", data = sub.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2242  -0.8355  -0.3369   0.8395   2.4072
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -19.109369   1.824871  -10.47  < 2e-16 ***
## HT            0.136833   0.027871    4.91 9.13e-07 ***
## WT            0.047340   0.003286   14.41  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

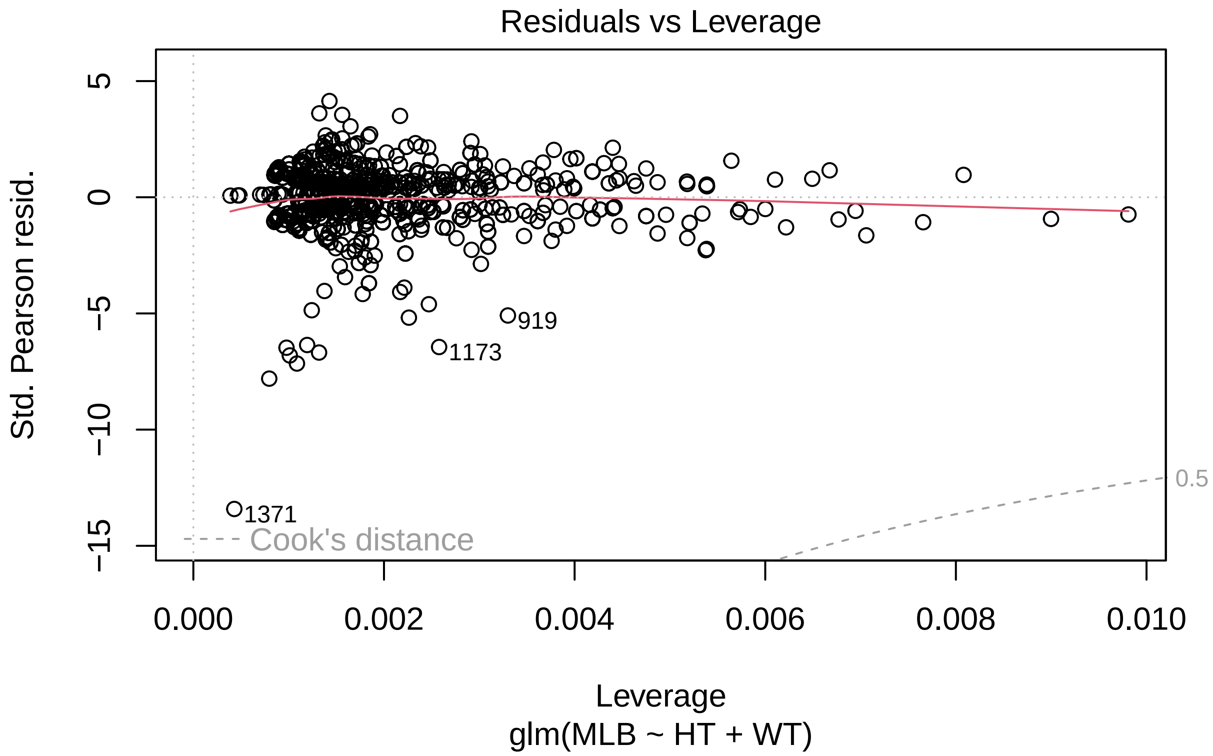
```
##
##     Null deviance: 2383.4  on 1719  degrees of freedom
## Residual deviance: 1802.9  on 1717  degrees of freedom
## AIC: 1808.9
##
## Number of Fisher Scoring iterations: 4
```

```
plot(m.fit)
```




Normal Q−Q
Std. Pearson resid.
5
0
−5
−10
1653
990
1371
−3
−2
−1
0
1
2
3
Theoretical Quantiles
glm(MLB ~ HT + WT)


Scale−Location
√|Std. Pearson resid.|
0
1
2
3
1371
990
1653
−4
−2
0
2
4
Predicted values
glm(MLB ~ HT + WT)



The fitted model is $\hat{p_i} = logit(-19.109369 + 0.136833 * HT_i + 0.047340 * WT_i); i = 1, ..., n$. Both predictors HT and WT are significant also with the null hypothesis $H_0 : \beta_i = 0$ being rejected for all i=(1,2,3).

Model Prediction

```
# T means that, yes the individual is in the MLB
m.probs <- predict(m.fit, sub.test, type="response")
m.pred <- rep(F, length(m.probs))
m.pred[m.probs>.5] <- T
```

Model Evaluation

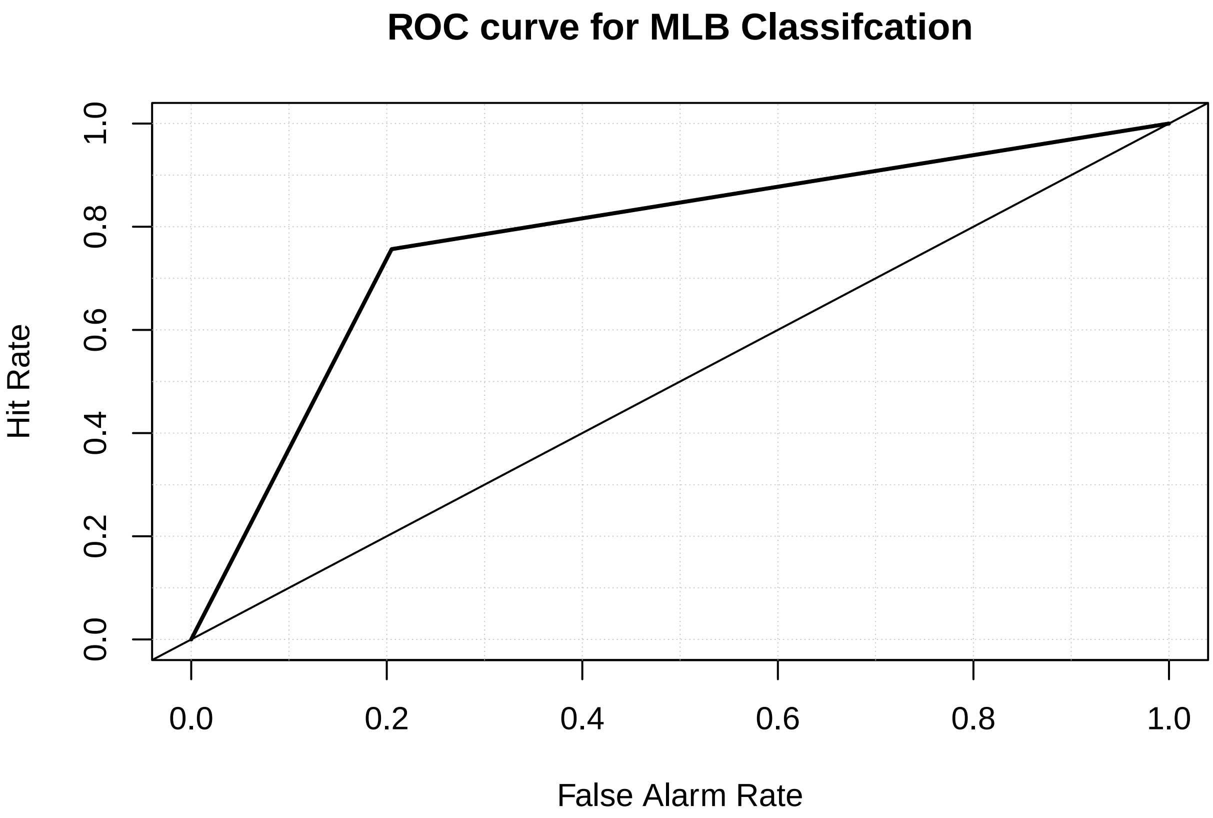
```
table(m.pred, Y.test, dnn=c("Predicted MLB", "Actual MLB"))
```

```
##                Actual MLB
## Predicted MLB FALSE TRUE
##         FALSE   287   93
##         TRUE     74  289
```

```
mean(m.pred==Y.test)
```

```
## [1] 0.7752355
```

```
roc.plot(x=as.numeric(Y.test), pred=as.numeric(m.pred), main="ROC curve for MLB Classifcation", plot.th
```



The model performs with a 79.43925% accuracy, a sensitivity of 77.211796% and a specificity of 81.648936.

Conclusion

As the prediction results from the logistic regression show, predicting the level at which an individual play baseball at, either MLB or Club, can be done at a relatively high rate with just the height and weight of the players as independent variables using logistic regression.