Team_Essay_2

Sahil, Ariel, Ilya, Lekha, Anthony, Trang

3/14/2021

Introduction

In this essay, our team will perform the process of forecasting height model based on several explanatory variables. The goal of this essay is to model the predicted height based on weight, FGP, FTP and PPG of a basketball team. To find the best model for predicted height, we need to find the model coefficients that minimize the sum of squared errors between the predicted height and the actual height.

Formula and Basics

Multiple linear regression: y = β_0 + $\beta_1 x_1$ + $\beta_2 x_2$ +...+$\beta_k x_k$ + ϵ

- β_0 is value of y when all x_k are equal to zero
- β_k are beta coefficients which measure the correlation between the result and it's predictor variables
- x_k are the independent variables
- ϵ is the error term, the part of y that can be explained through the regression model.

Loading Required R Packages

- readxl - to read data from xl
- tidyverse - for data visualization and manipulation
- Metrics - to compute rmse
- caret - to compute the VIF
- GGally - to graph the correlation between independent variables

```r
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library("readxl")
library("Metrics")
library("caret")
```

```
## Loading required package: lattice

##
## Attaching package: 'caret'
```

```
## The following objects are masked from 'package:Metrics':
##
##     precision, recall
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```r
library("GGally")
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

Data Description

Examples of data and problem:

- height represents the height of basketball players in feet.
- weight represents the weight of basketball players in pounds.
- FGP represents the percentage of successful field goals (out of 100 attempted).
- FTP represents the percentage of successful free throws (out of 100 attempted).
- PPG represents average points scored per game.

```r
bball_data <- read_excel("Basketball.xlsx")
bball_data <- bball_data %>%
  rename(
    "height" = "X1",
    "weight" = "X2",
    "FGP" = "X3",
    "FTP" = "X4",
    "PPG" = "X5"
  )
full_data = bball_data[1:55,]
bball_data
```

```
## # A tibble: 54 x 5
##    height weight   FGP   FTP   PPG
##     <dbl>  <dbl> <dbl> <dbl> <dbl>
##  1    6.8    225 0.442 0.672   9.2
##  2    6.3    180 0.435 0.797  11.7
##  3    6.4    190 0.456 0.761  15.8
##  4    6.2    180 0.416 0.651   8.6
##  5    6.9    205 0.449 0.9    23.2
##  6    6.4    225 0.431 0.78   27.4
##  7    6.3    185 0.487 0.771   9.3
##  8    6.8    235 0.469 0.75   16
##  9    6.9    235 0.435 0.818   4.7
## 10    6.7    210 0.48  0.825  12.5
## # ... with 44 more rows
```

Computation

- Height_Model: `height = b0 + b1*weight + b2*FGP + b3*FTP + b4*PPG`

```r
bball_data <- read_excel("Basketball.xlsx")
bball_data <- bball_data %>%
```

```r
  rename(
    "height" = "X1",
    "weight" = "X2",
    "FGP" = "X3",
    "FTP" = "X4",
    "PPG" = "X5"
  )
train_data = bball_data[1:30,]
test_data = bball_data[31:52,]
```

Interpretation of the Model

```r
Height_Model <- lm(height ~ ., data = bball_data)
#Summary of the model
summary(Height_Model)
```

```
##
## Call:
## lm(formula = height ~ ., data = bball_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49883 -0.19071  0.01065  0.09117  0.78835
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.798050   0.448437   8.470 3.69e-11 ***
## weight       0.011489   0.001454   7.899 2.72e-10 ***
## FGP          1.138890   0.794921   1.433    0.158
## FTP         -0.049316   0.380348  -0.130    0.897
## PPG         -0.008273   0.006660  -1.242    0.220
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2562 on 49 degrees of freedom
## Multiple R-squared:  0.7119, Adjusted R-squared:  0.6883
## F-statistic: 30.26 on 4 and 49 DF,  p-value: 1.062e-12
```

```r
summary(Height_Model)$coefficient
```

```
##                 Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  3.798049837 0.448437080  8.4695267 3.690763e-11
## weight       0.011489032 0.001454441  7.8992770 2.724423e-10
## FGP          1.138889566 0.794921477  1.4327070 1.582910e-01
## FTP         -0.049315550 0.380347508 -0.1296592 8.973669e-01
## PPG         -0.008273347 0.006659883 -1.2422661 2.200511e-01
```

```r
#Confidence Interval of the model coefficients
confint(Height_Model)
```

```
##                    2.5 %      97.5 %
## (Intercept)  2.896881785 4.699217890
## weight       0.008566224 0.014411841
## FGP         -0.458564949 2.736344082
## FTP         -0.813652483 0.715021384
```

```
## PPG          -0.021656883 0.005110189
```

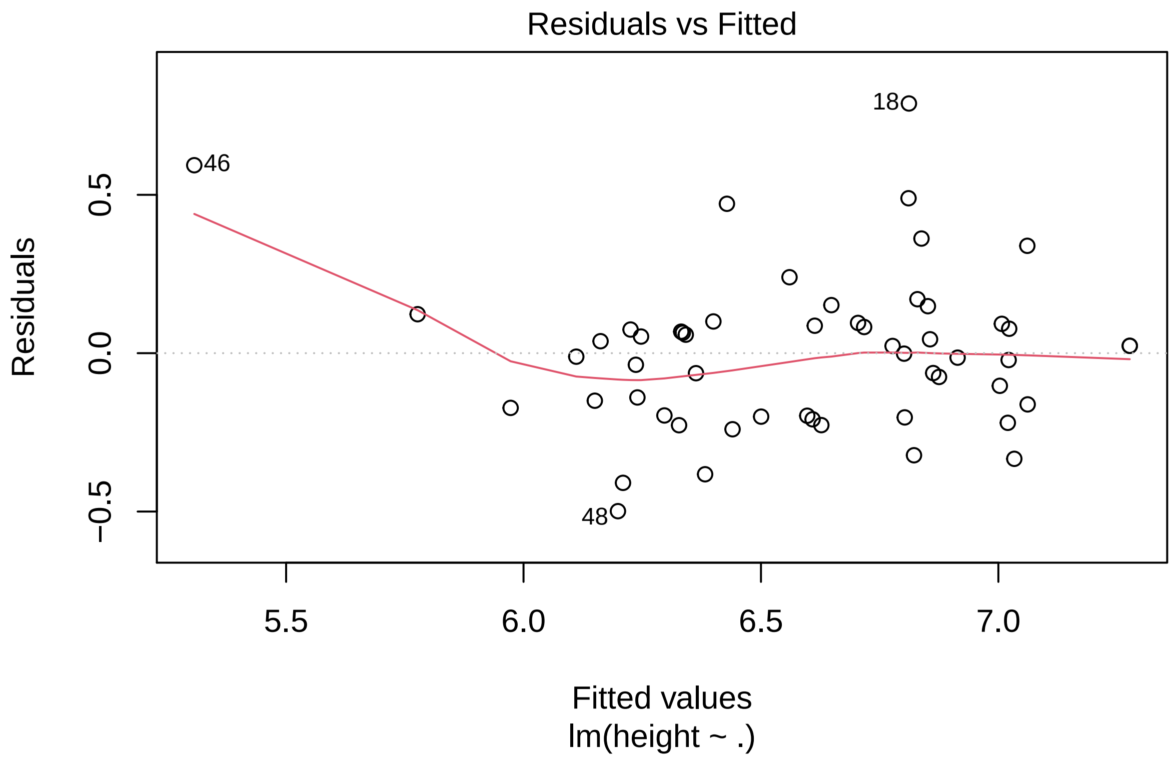
```r
#Coefficient of determination
summary(Height_Model)$r.squared
```

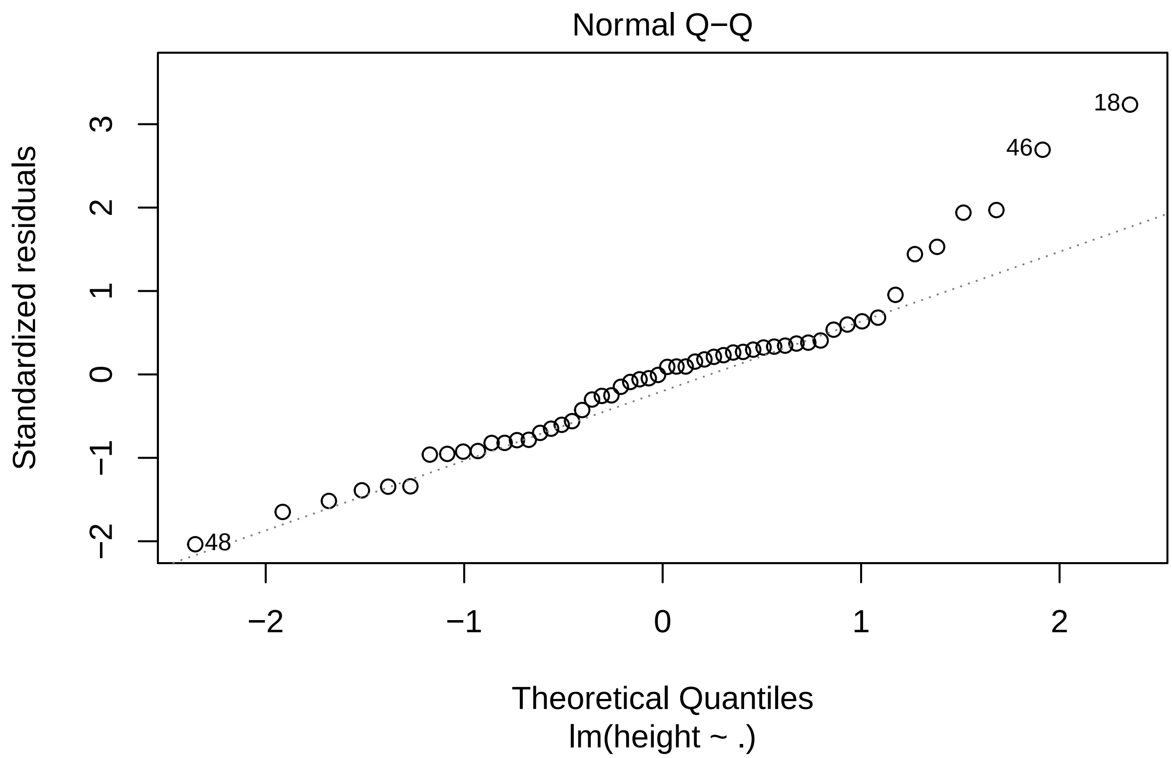
```
## [1] 0.7118583
```

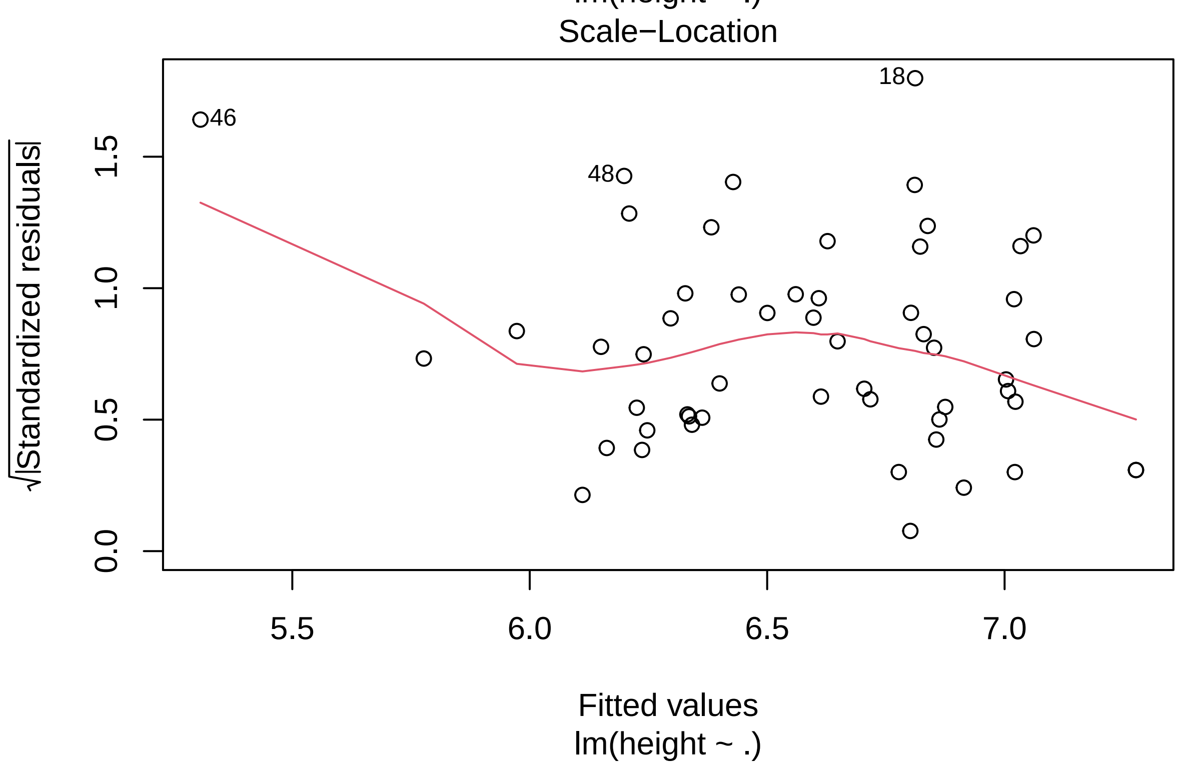
```r
#RSE
sigma(Height_Model)/mean(bball_data$height)
```

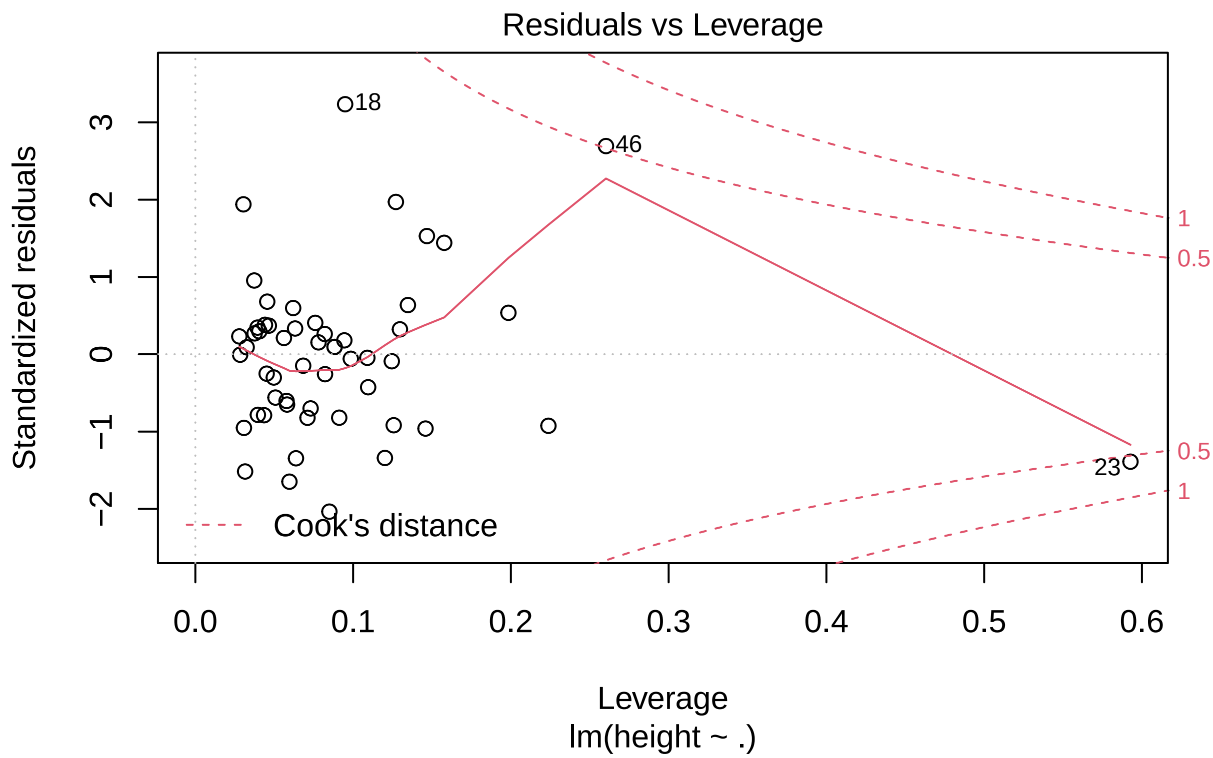
```
## [1] 0.03889242
```

```r
plot(Height_Model)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(height ~ .)
18
46
48
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
Fitted values
lm(height ~ .)
18
46
48
0.0
0.5
1.0
1.5
5.5
6.0
6.5
7.0



From the output for above Height_Model:

We can see that resulting regression equation is:

- `height = 3.798049837 + 0.011489032*weight + 1.138889566*FGP - 0.049315550*FTP - 0.008273347*PPG`
- b0 = 3.798049837
- b1 = 0.011489032
- b2 = 1.138889566
- b3 = 0.049315550
- b4 = -0.008273347

We also observe that weight is the most significant predictor.

Model Evaluation

Training model

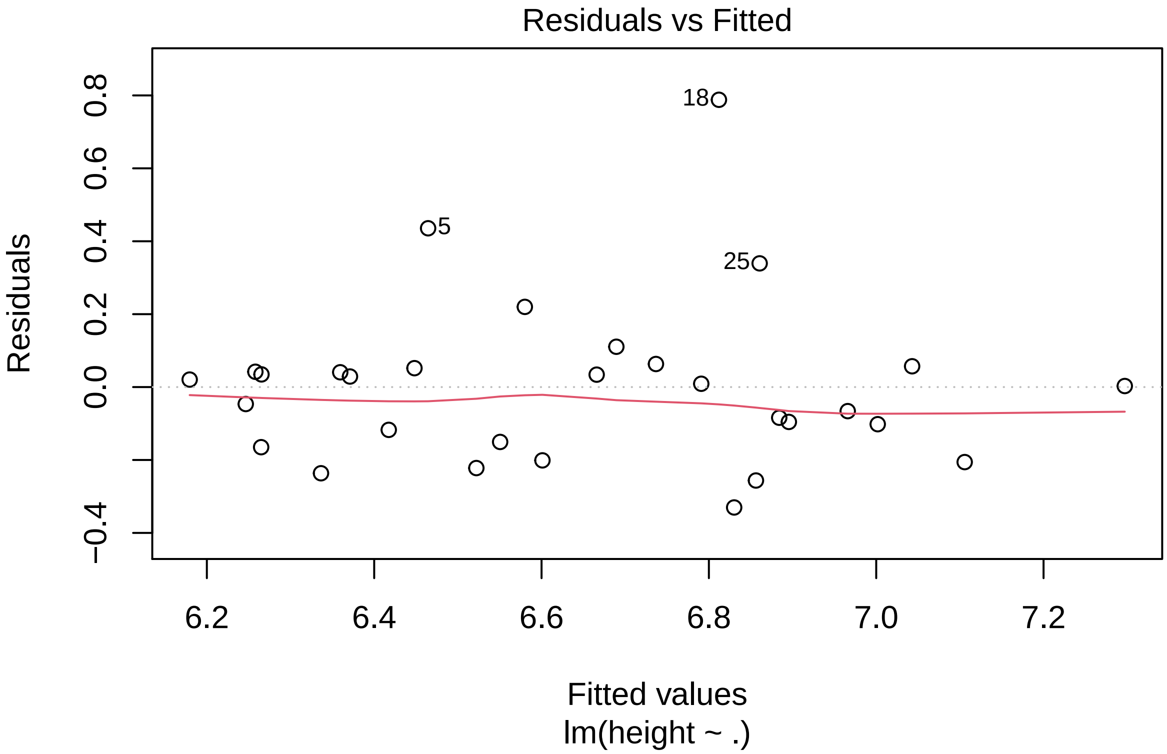
```
train_model <-lm(height ~ ., data = train_data)
train_pred <- predict(train_model, train_data)
rmse(train_pred, train_data$height)
```

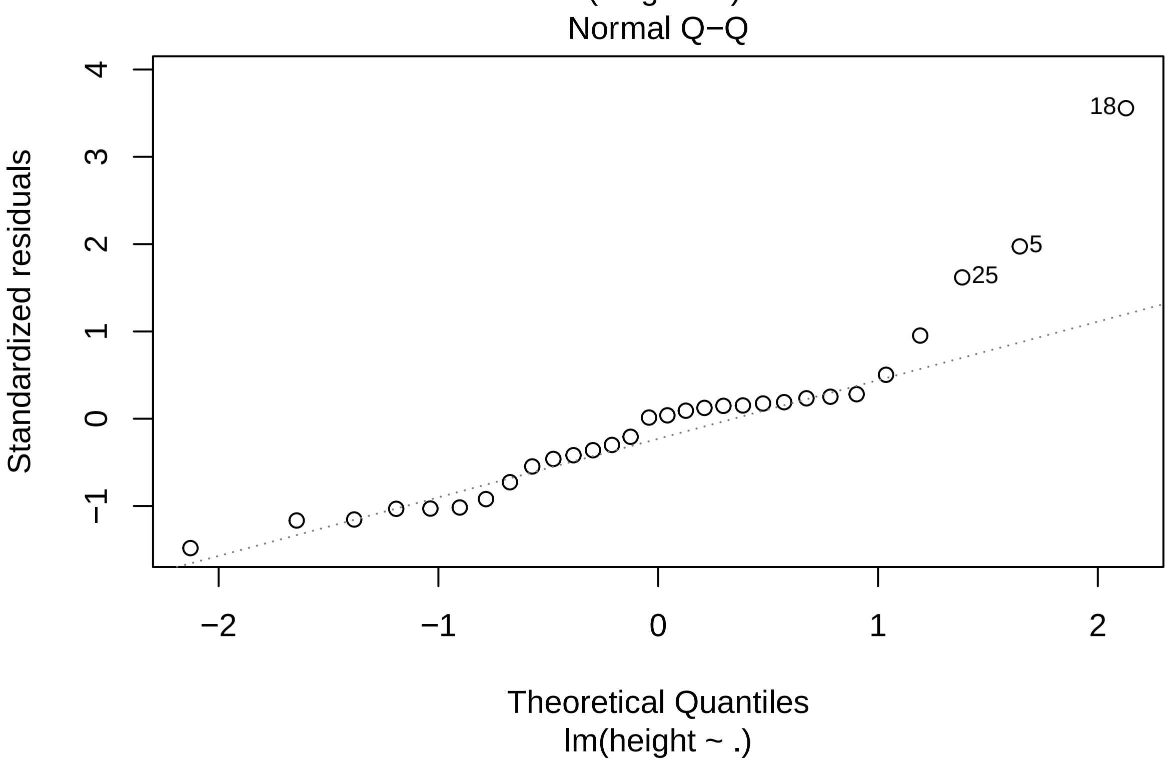
```
## [1] 0.2208217
```

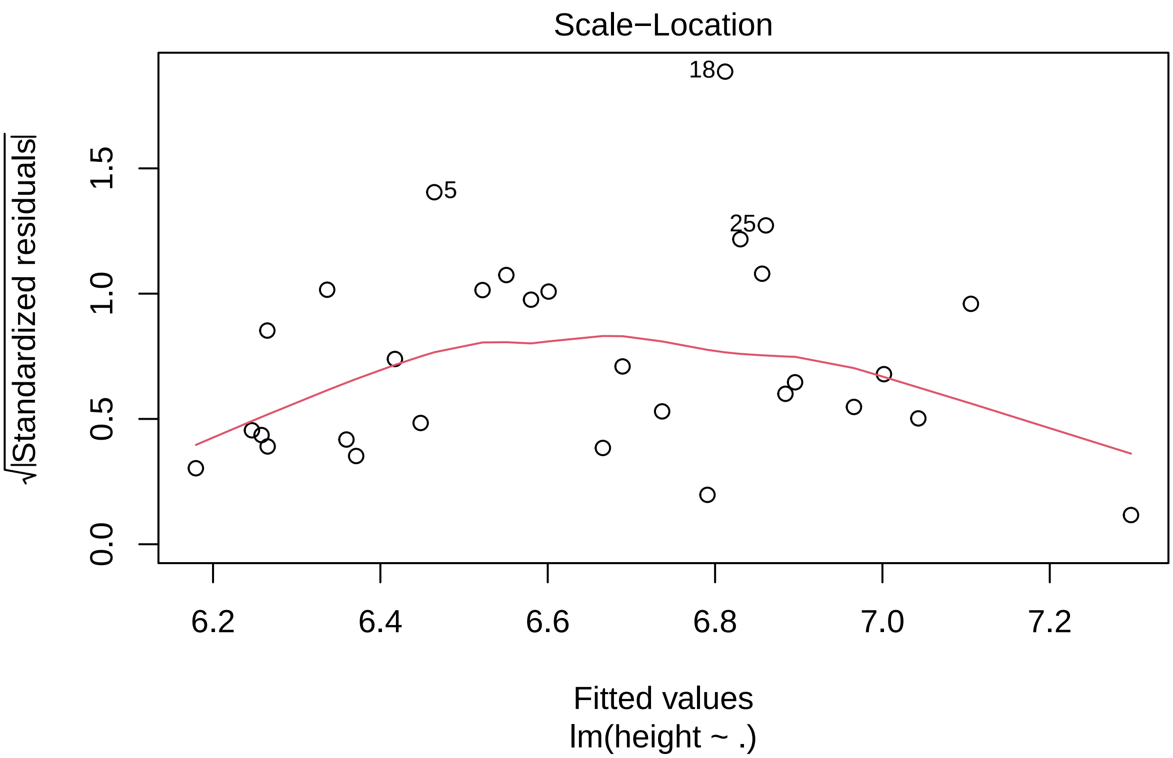
```
res <- train_pred - train_data$height
sum(res)
```

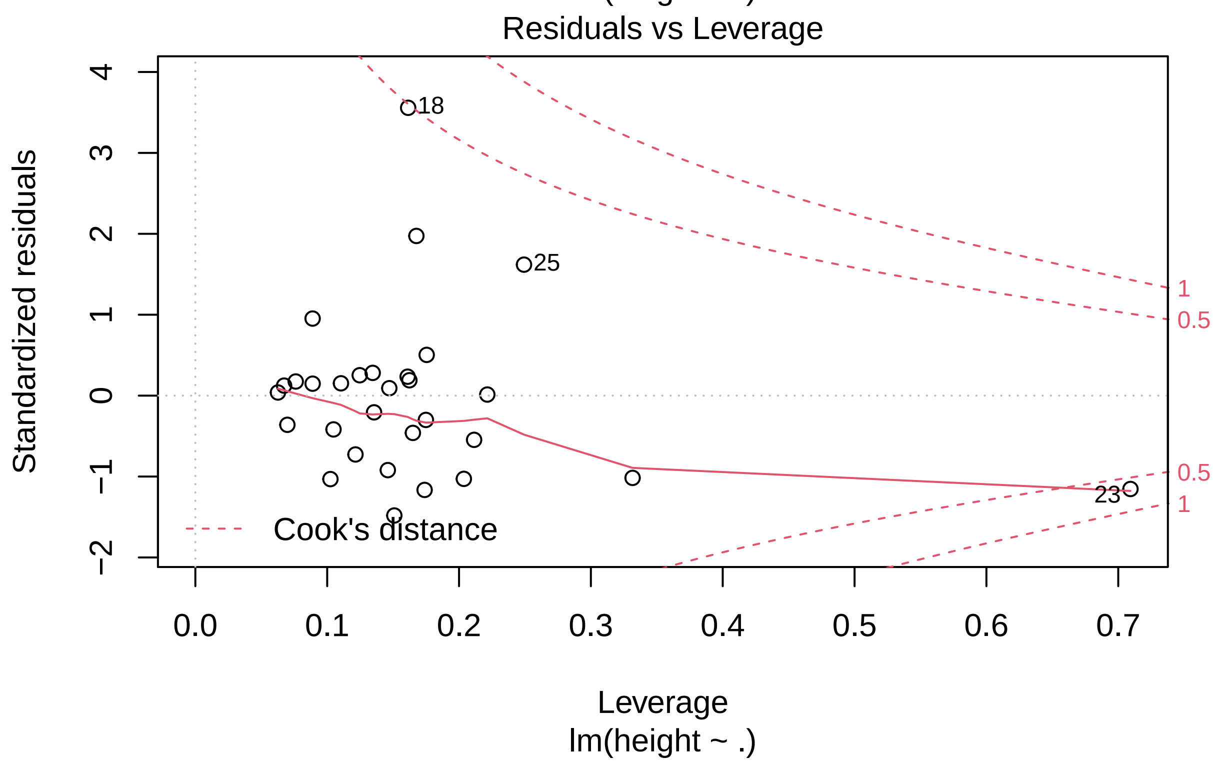
```
## [1] -1.865175e-14
```

```
plot(train_model)
```


Residuals vs Fitted
Residuals
0.8
0.6
0.4
0.2
0.0
−0.4
18
5
25
6.2
6.4
6.6
6.8
7.0
7.2
Fitted values
lm(height ~ .)


Normal Q−Q
Standardized residuals
4
3
2
1
0
−1
18
5
25
−2
−1
0
1
2
Theoretical Quantiles
lm(height ~ .)





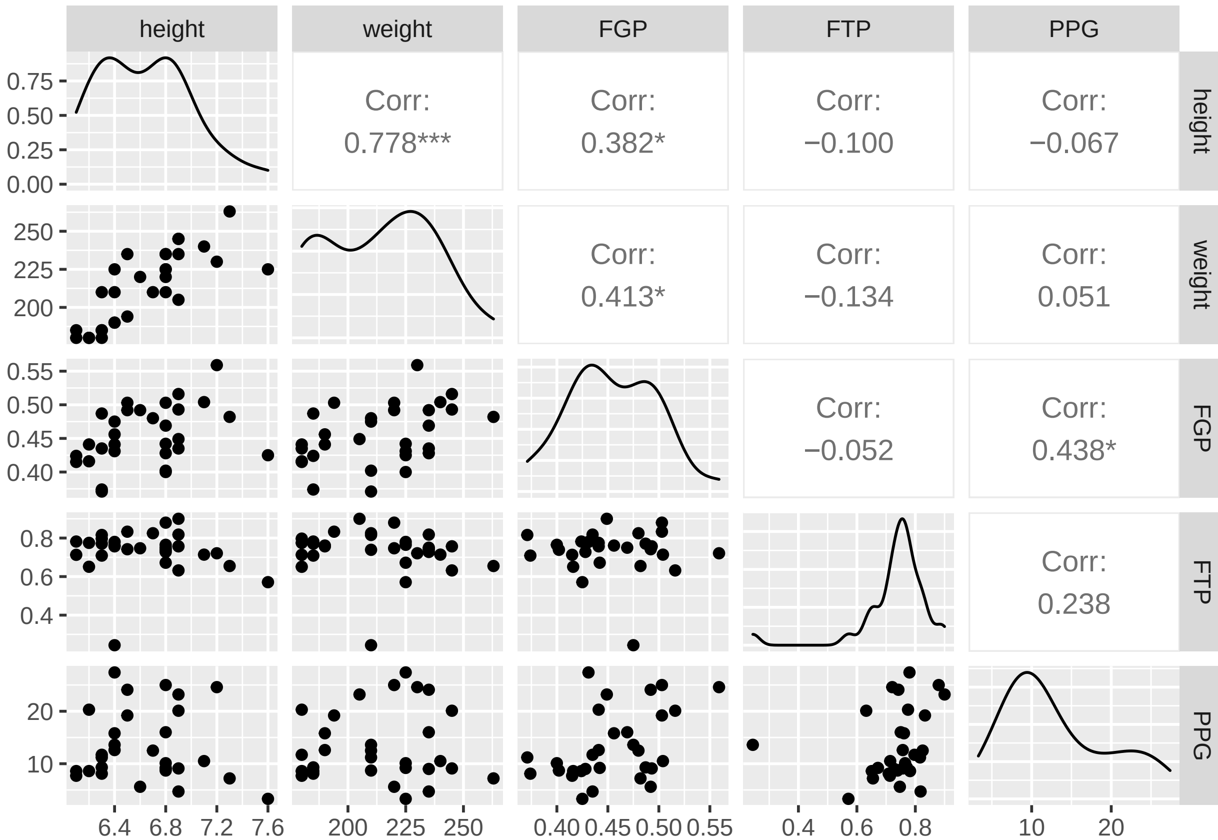

```r
car::vif(train_model)
```

```
##   weight      FGP      FTP      PPG
## 1.245640 1.551091 1.102578 1.373456
```

- The RMSE for training model is 0.2208217.
- The sum of residuals for training model is -1.865175e-14.

- The VIF values weight, FGP, FTP, and PPG are 1.25, 1.55, 1.10, 1.37, respectively. These values are all around 1 which signifies that there is no collinearity between the variables.

```
ggpairs(train_data)
```



- These charts demonstrate that there is no correlation between our independent variables.

Testing Model

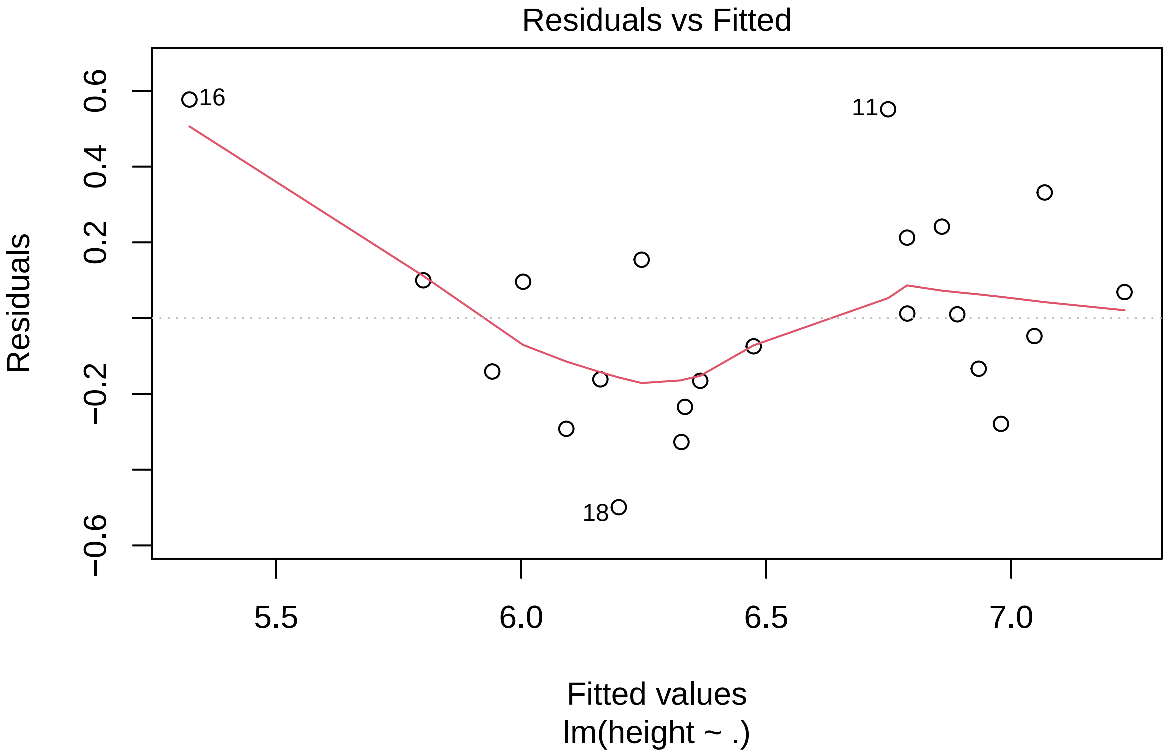
```
test_model <-lm(height ~ ., data = test_data)
test_pred <- predict(train_model, test_data)
rmse(test_pred, test_data$height)
```

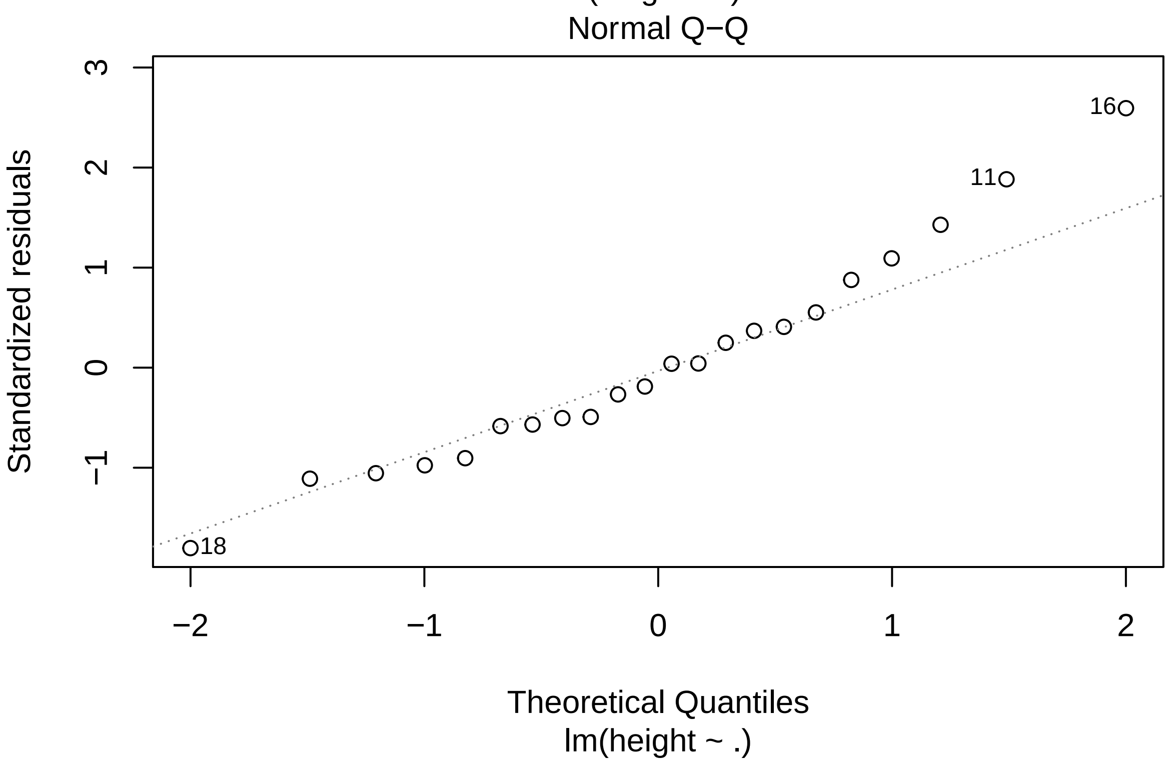
```
## [1] 0.2863395
```

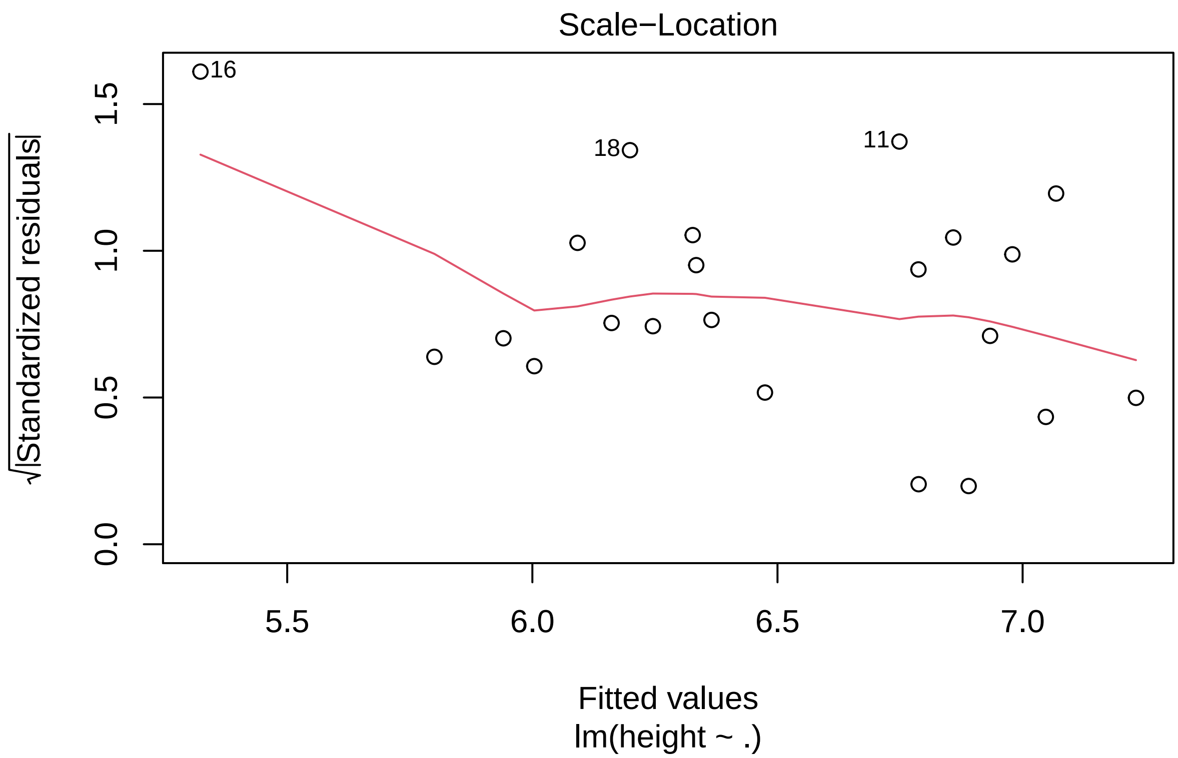
```
res2 <- test_pred - test_data$height
sum(res2)
```

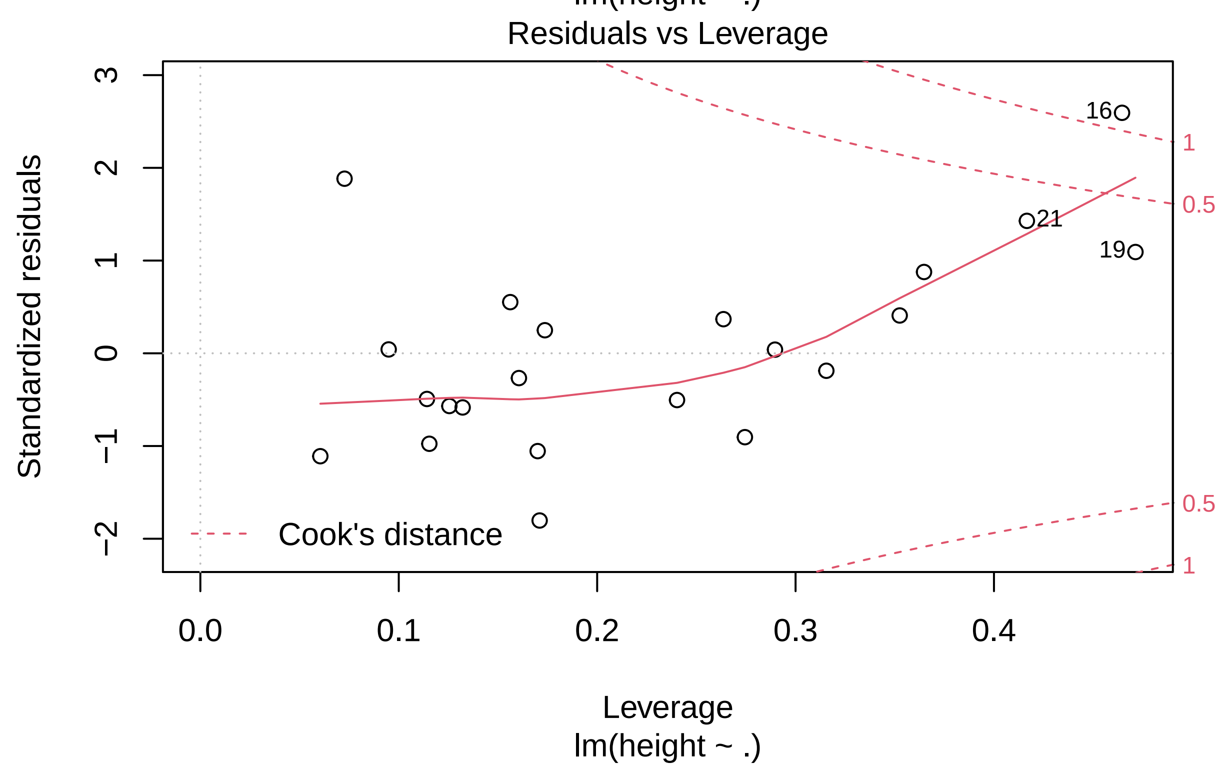
```
## [1] 1.664502
```

```
plot(test_model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(height ~ .)
16
11
18
0.6
0.4
0.2
−0.2
−0.6
5.5
6.0
6.5
7.0


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(height ~ .)
16
11
18
3
2
1
0
−1
−2
−1
0
1
2





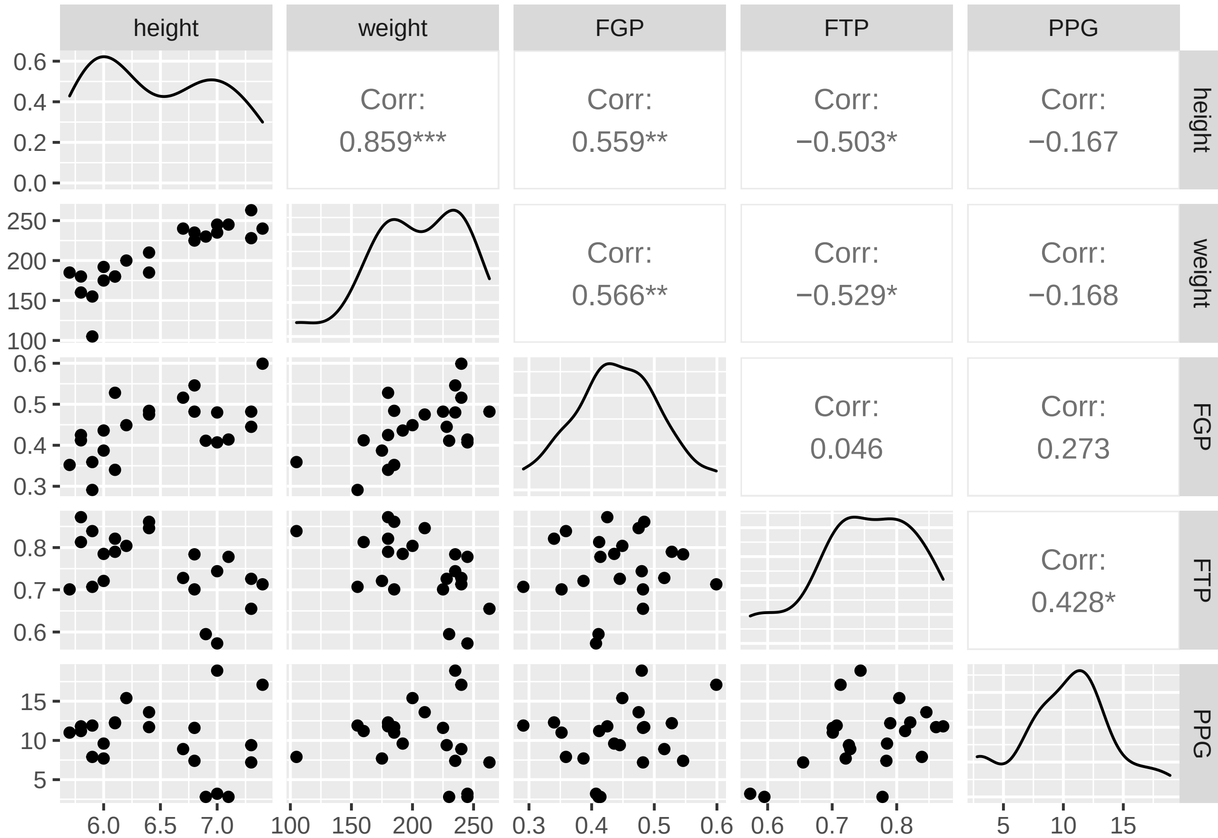

```r
car::vif(test_model)
```

```
##   weight      FGP      FTP      PPG
## 2.780699 2.165573 1.943623 1.370597
```

- The RMSE for testing model is 0.2863395. (This is pretty good)
- The sum of residuals for testing model is 1.664502. (This could have been better)

- The VIF values weight, FGP, FTP, and PPG are 2.78, 2.17, 1.94, 1.37, respectively. These values are all around 2 which signifies that there is no collinearity between the variables.

```
ggpairs(test_data)
```



- These charts demonstrate that there is no correlation between our independent variables.

Model Assessment

```
summary(train_model)
```

```
##
## Call:
## lm(formula = height ~ ., data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33022 -0.14224  0.00600  0.04941  0.78798
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.593446   0.651928   5.512 9.97e-06 ***
## weight       0.011334   0.002108   5.376 1.41e-05 ***
## FGP          1.431500   1.274968   1.123    0.272
## FTP          0.166296   0.406133   0.409    0.686
## PPG         -0.010552   0.007830  -1.348    0.190
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.2419 on 25 degrees of freedom
## Multiple R-squared:  0.6355, Adjusted R-squared:  0.5772
## F-statistic:  10.9 on 4 and 25 DF,  p-value: 2.967e-05
```

```
sigma(train_model)/mean(train_data$height)
```

```
## [1] 0.03641216
```

We can see that our model has the following indicators:

- Residual standard error: 0.2419 on 25 degrees of freedom (The more it is closer to 1 the better)
- Multiple R-squared: 0.6355 (Higher the better)
- Adjusted R-squared: 0.5772 (Higher the better)
- F-statistic: 10.9 on 4 and 25 DF (Higher the better)
- p-value: 2.967e-05 (Lower the better)
- RSE: 0.03641216 (Lower the better)

Interpretation

Height_Model: `height = 3.798049837 + 0.011489032*weight + 1.138889566*FGP - 0.049315550*FTP - 0.008273347*PPG`

- We can say that height = 3.798049837, when all the other predictors are set to 0.
- In our model, it can be seen that p-value of the F-statistic is < 1.41e-05, which is highly significant. This means that, at least, one of the predictor variables is significantly related to the outcome variable.
- To see which predictor variables are significant, you can examine the coefficients table, which shows the estimate of regression beta coefficients and the associated t-statistic p-values:

```
summary(train_model)$coefficient
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)  3.59344599 0.651927816  5.512030 9.969781e-06
## weight       0.01133357 0.002108088  5.376232 1.412384e-05
## FGP          1.43149956 1.274967590  1.122773 2.722012e-01
## FTP          0.16629595 0.406132756  0.409462 6.856878e-01
## PPG         -0.01055218 0.007829857 -1.347685 1.898469e-01
```

- For a given the predictor, the t-statistic evaluates whether or not there is significant association between the predictor and the outcome variable, that is whether the beta coefficient of the predictor is significantly different from zero.
- It can be seen that, changing weight and PPG are significantly associated to changes in height while changes in FTP and FGP are not significantly associated with height.

```
model2 <- lm(height ~ weight + PPG, data = bball_data)
summary(model2)
```

```
##
## Call:
## lm(formula = height ~ weight + PPG, data = bball_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.57591 -0.14373  0.01017  0.11901  0.78208
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.989189   0.257184  15.511  < 2e-16 ***
```

```
## weight       0.012641   0.001163  10.866 7.05e-15 ***
## PPG         -0.004722   0.005969  -0.791    0.433
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2563 on 51 degrees of freedom
## Multiple R-squared:  0.6998, Adjusted R-squared:  0.688
## F-statistic: 59.44 on 2 and 51 DF,  p-value: 4.728e-14
```

Removing FTP and FGP we get the following:

- Residual standard error: 0.2563 on 51 degrees of freedom
- Multiple R-squared: 0.6998
- Adjusted R-squared: 0.688
- F-statistic: 59.44 on 2 and 51 DF
- p-value: 4.728e-14
- There are few better changes but not as much as we expected.

Model accuracy assessment

Height_Model: `height = 3.798049837 + 0.011489032*weight + 1.138889566*FGP - 0.049315550*FTP - 0.008273347*PPG`

1. weight = 225, FGP = 0.482, FTP = 0.701, PPG = 11.6

- Estimate of Height: 6.8 feet
- Actual: 6.8 feet

2. weight = 215, FGP = 0.457, FTP = 0.734, PPG = 5.8

- Estimate of Height: 6.7 feet
- Actual: 6.8 feet

3. weight = 230, FGP = 0.435, FTP = 0.764, PPG = 8.3

- Estimate of Height: 6.82 feet
- Actual: 7 feet

Conclusion

All in all, our task was to test a number of models, each varying by the response(dependent) variable used, and found that the model with height as the response(dependent) variable gave the most notable result. Hence, this model was used for regression analysis, and we conducted the pertaining steps to find which of the following predictor variables were associated the strongest with height:

- weight
- free-throw percentage (FTP)
- field goal percentage (FGP)
- average points scored per game (PPG)

After we observed their corresponding beta coefficient values and compared the t-values (three asterisks next to the data of the predictor variable with highest t-value, in concerned R code output), it was clear that weight was a significant predictor of height (it's t-value was 7.899, beta coefficient value = 0.011489), followed by the points per game (t-value was -1.242, beta coefficient value = -1.242). Our data containing 52 players was then split up into two parts, the first 30 players going to the training set and the next 22 players going to the testing set. The training set model had a RMSE (Root Mean Square Error) value of 0.22 with a sum of residuals of -1.865175 * 10^-14. A small RMSE and sum of residual values signified that our model predicted the height of basketball players quite accurately. Additionally, the testing set model showed further positive results with a RMSE value of around 0.28; its residual values did prove otherwise at a sum total of 1.664502, though it was only due to the number of outliers in the data. Looking at our model's initial indicators, most of them (also called as the metrics) had already satisfied to a reasonable extent the conditions for what makes a reliable model . We saw improvements (though not as much as ideally wanted) once we rebuilt the model without variables FTP and FGP; we saw:

- residual standard value closer to 1 (0.2419 to 0.2563)
- higher multiple and adjusted R^2 value (0.6355 to 0.6998, 0.5772 to 0.688)
- much higher F-statistic (10.9 to 59.44)
- MUCH lower p-value (2.967e-05 to 4.728e-14)

Assessing the accuracy of the model, the calculated model estimates are observed to be notably close to the actual value.

References

http://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/