# Xử lý số liệu thống kê - FIFA Data

Nhóm 8

2025-1-10

## Thông tin nhóm

1. Kha Thái Hồ - 22280025 - Leader
2. Nguyễn Văn Trung Chính - 22280007 - Member
3. Trần Chí Hữu - 22280038 - Member
4. Huỳnh Ngọc Hòa - 22280026 - Member
5. Lê Phan Ngọc Hiếu - 22280023 – Member

## Giới thiệu đề tài

Bóng đá từ lâu đã và đang là môn thể thao vua của nhân loại không chỉ trên sân cỏ mà còn trên cả lĩnh vực phân tích và công nghệ. Với sự phát triển của khoa học xử lý và phân tích dữ liệu, việc trích xuất thông tin từ các cơ sở dữ liệu có sẵn đã trở thành 1 xu hướng phổ biến. Dữ liệu fifa_eda_stats là 1 bộ dữ liệu với phong phú và đầy đủ các thông tin về hàng ngàn các cầu thủ đang thi đấu trên toàn thế giới.

Mục tiêu của dự án là phân tích, trích xuất thông tin từ bộ dữ liệu fifa, từ đó rút ra các đặc trưng quan trọng của 1 hay đại đa số cầu thủ để từ đó giúp quản lý các câu lạc bộ có thể lựa chọn được cầu thủ phù hợp với chi phí và nhu cầu của mình.

### Bảng đề xuất phân tích dữ liệu

| Đề xuất | Mô tả | Phương pháp |
|---|---|---|
| Kiểm tra tính độc lập giữa các nhóm dữ liệu | Kiểm tra tính độc lập giữa yếu tố chân thuận đến mức lương cầu thủ hoặcmột số chỉ số quan trọng, giữa các nhóm vị trí trên sân đến lương. | A/B Testing<br>Permutation Test<br>UNOVA |
| Xây dựng các bảng tổng hợp dữ liệu | Khái quát dữ liệu bằng các bảng tóm tắt để có cái nhìn tổng quát đến dữ liệu | Bảng tóm tắt dữ liệu |
| TÌm hiểu các đặc trưng quan trọng ảnh hướng đến lương cầu thủ | Phân tích các đặc trưng quan trọng trong việc chọn lựa 1 cầu thủ tốt | Mô hình hồi quy tuyến tính |

## Bảng đề xuất xử lý dữ liệu

| Đề xuất | Mô tả | Phương pháp |
|---|---|---|
| Xử lý các giá trị sai định dạng | Dùng các lệnh đã được học để đưa các giá trị dạng chuỗi về dạng số | |
| Chuyển đổi đơn vị phù hợp | Dùng kiến thức thực tế và các lệnh để đưa các giá trị về đơn vị phù hợp. | |
| Loại bỏ các biến không cần thiết | Xem xét và loại bỏ các biến không phù hợp. | |
| Kiểm tra và điền khuyết các giá trị null và missing | Dùng biểu đồ Tỉ lệ khuyết để xác định loại khuyết Điền hoặc loại bỏ các giá trị khuyết | Multiple Imputation |
| Tạo các biến mới | Gom nhóm hoặc tách các biến đã có thành các biến mới | |

## Bảng đề xuất xây dựng mô hình

| Đề xuất | Mô tả | Phương pháp |
|---|---|---|
| Mô hình dự đoán Overall | Xây dựng mô hình dự đoán Overall từ các chỉ số chuyên môn | Mô hình hồi quy |
| Mô hình dự đoán mức lương của Thủ môn | Xây dựng mô hình dự đoán Wage của Thủ Môn từ các biến phù hợp cho Thủ môn | Mô hình hồi quy |
| Mô hình dự đoán mức lương của Tiền đạo | Xây dựng mô hình dự đoán Wage của Thủ đạo từ các biến phù hợp cho Tiền đạo | Mô hình hồi quy |
| Mô hình dự đoán mức lương của Tiền vệ | Xây dựng mô hình dự đoán Wage của Tiền vệ từ các biến phù hợp cho Tiền vệ | Mô hình hồi quy |
| Mô hình dự đoán mức lương của Hậu vệ | Xây dựng mô hình dự đoán Wage của Hậu vệ từ các biến phù hợp cho Hậu vệ | Mô hình hồi quy |

## Bảng quy trình xây dựng mô hình

| Giai đoạn | Mô tả | Phương pháp |
|---|---|---|

| | | |
|---|---|---|
| Xử lý dữ liệu | Dùng các phương pháp để lựa chọn và xử lý các định dạng phù hợp | Feature Encoding, Group |
| Lựa chọn mô hình | Tìm kiếm các biến phù hợp để xây dựng mô hình cho từng vị trí. | Hồi quy từng phần |
| Mô hình hồi quy cơ bản | Xây dựng mô hình hồi quy cơ bản bằng các biến đã lựa chọn sẵn | Hồi quy tuyến tính |
| Chuẩn đoán mô hình | Dùng các biểu đồ để kiểm tra các tính chất của mô hình | Tính tuyến tính<br>Đồng nhất phương sai<br>Đa cộng tuyến |
| Đánh giá mô hình | Dùng các thông số thống kê để kiểm tra độ hiệu quả của mô hình | RMSE<br>R2Score<br>k-fold Validation |
| Mở rộng mô hình | Dùng các mô hình bậc 2 với các biến không tuyến tính từng phần với y | Quadratic Regression |

Quy trình cụ thể:

## 1.Làm sạch tên biến

## 2.Tổng quan dữ liệu

**Nhận xét**:

- Dữ liệu có 18207 bản ghi, 57 biến bao gồm 15 biến phân loại và 42 biến dạng số.
- Trong số các biến phân loại, biến bị khuyết nhiều nhất là biến loaned_from với 16943 dữ liệu khuyết.
- Trong số các biến dạng số, cũng tồn tại các giá trị khuyết.
- Trung bình của chỉ số Overall là 71.31 với sai số chuẩn 6.14 (Trên thang điểm 100). Trung bình của international_reputation là 1.11 với sd là 0.39 (Trên thang điểm 5)

## 3.Kiểm tra bản ghi trùng lặp

**Nhận xét**: Không có bản ghi trùng lặp trong dữ liệu.

## 4.Xử lý cho từng biến cụ thể:

### 4.1 Xử lý các giá trị tiền tệ:

Ta kiểm tra thấy các giá trị tiền tệ (Wage, value) chứa các ký tự không hợp lệ

=> Ta sẽ chỉnh sửa các giá trị này (Đơn vị: triệu Euro)

```
## # A tibble: 5 × 2
##    wage value
##   <dbl> <dbl>
## 1 0.565  110.
## 2 0.405   77
## 3 0.29   118.
## 4 0.26    72
## 5 0.355  102
```

## 4.2 Chiều cao và cân nặng:

Chiều cao đang ở đơn vị inch và cân nặng đơn vị lbs,

```
## # A tibble: 5 × 2
##   height weight
##   <chr>  <chr>
## 1 5'7    159lbs
## 2 6'2    183lbs
## 3 5'9    150lbs
## 4 6'4    168lbs
## 5 5'11   154lbs
```

=> Ta sẽ đưa về đơn vị cm và kg để phù hợp hơn.

```
## # A tibble: 5 × 2
##   height weight
##    <dbl>  <dbl>
## 1   1.70   72.1
## 2   1.88   83.0
## 3   1.75   68.0
## 4   1.93   76.2
## 5   1.80   69.9
```

## 4.3 Biến work_rate:

Ta nhận thấy biến work_rate là biến phân loại đang chứa 2 thông tin

```
## [1] "Medium/ Medium" "High/ Low"      "High/ Medium"   "Medium/ Medium"
## [5] "High/ High"
```

=> Ta sẽ tách ra thành attack và defense và encode theo cấp bậc tương ứng.

```
## # A tibble: 5 × 2
##   work_rate_attack work_rate_defense
##              <dbl>             <dbl>
## 1                2                 2
## 2                3                 1
## 3                3                 2
```

```
## 4                2                 2
## 5                3                 3
```

## 4.4 Biến body_type:

Ta số hóa các giá trị dạng type

```
## [1] 5 2 6 4 7
```

## 4.5 Biến loaned_from

Biến loaned_from chứa quá nhiều giá trị rỗng (>80%) và không có cách khôi phục được

```
## [1] 0.9305762
```

=> Ta sẽ loại bỏ biến này

## 4.6 Biến preferred_foot:

Do các dòng bị thiếu ở cột preferred_foot cũng thiếu trên hầu như các biến khác (47/57) do đó nếu dự đoán và điền các giá trị bị thiếu từ các biến còn lại sẽ không chính xác Mà dữ liệu bị khuyết < 5% nên cách giải quyết là loại bỏ các dòng mà preferred_foot bị khuyết

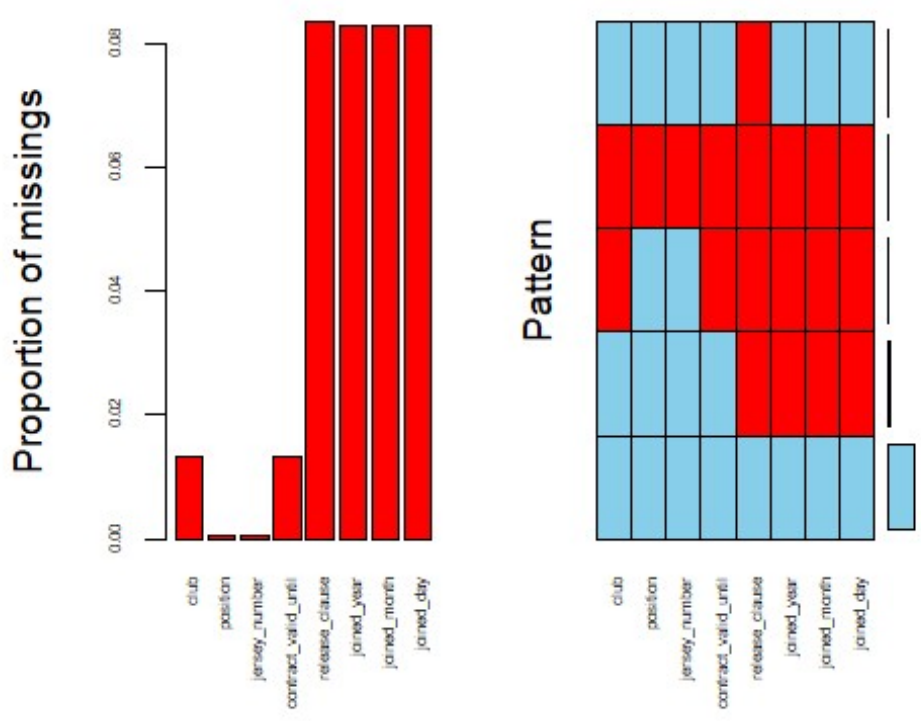
## 4.7 Biến joined:

Biến joined chứa các thông tin về ngày, tháng, năm các cầu thủ gia nhập câu lạc bộ, ta sẽ tiến hành tách theo từng đơn vị thời gian

```
## # A tibble: 5 × 3
##   joined_day joined_month joined_year
##        <int>        <dbl>       <dbl>
## 1          1            7        2004
## 2         10            7        2018
## 3          3            8        2017
## 4          1            7        2011
## 5         30            8        2015
```

# 5.Xử lý dữ liệu khuyết:

## 5.1 Biểu đồ % tỉ lệ khuyết của dữ liệu:




## 5.2 Điền các giá trị khuyết

Ta điền các giá trị khuyết cho biến câu lạc bộ (club) và biến release_clause

- club: Chuyển các giá trị khuyết thành Unknown (Có thể hiểu là chưa có hoặc chưa biết câu lạc bộ)
- release_clause: CHuyển giá trị khuyết thành Unknown và chuyển đổi thành 0

## 5.3 Điền giá trị khuyết bằng multiple imputation

Cuối cùng, điền các giá trị khuyết còn lại bằng Multiple Imputation

# 6.Outliers Remover:

Bằng một vài phương pháp như Histogram hay dùng bảng tóm tắt, ta có thể thấy biến wage và value bị lệch quá lớn vì có nhiều giá trị quá

cao, để phù hợp cho tính toán thống kê, ta sẽ dùng IQR để lọc các giá trị này đi.

Số dòng dữ liệu trước khi xử lý:

```
## [1] 18159
```

Dữ liệu còn lại khi xử lý Outliers:

```
## [1] 13823
```

## 7.Loại bỏ các Feature không cần thiết:

Một số biến không đóng góp nhiều cho dữ liệu hoặc khong cần thiết như: joined_year, joined_month, joined_day

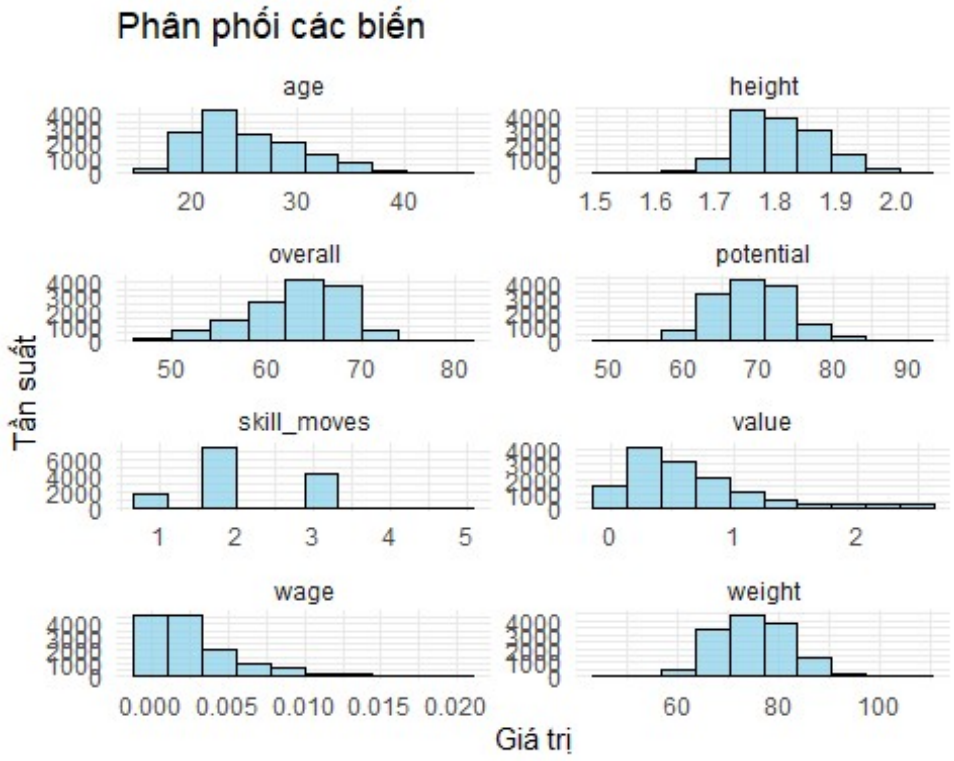
# III. Khám phá dữ liệu

## 1.Bảng tóm tắt

```
##            bien      Min      Max   Median         Mean           SD
## 1           age 16.00000  45.0000 24.00000 24.734717500 4.835704767
## 2        height  1.54940   2.0574  1.80340  1.812117182 0.067013549
## 3       overall 46.00000  82.0000 64.00000 63.585401143 5.192969477
## 4     potential 48.00000  89.0000 69.00000 69.322433625 5.026365973
## 5 releaseClause  0.00000  34.5000  0.86600  1.224713666 1.292484197
## 6    skillMoves  1.00000   5.0000  2.00000  2.219127541 0.666240664
## 7         value  0.01000   2.5000  0.50000  0.654242567 0.543504229
## 8          wage  0.00100   0.0210  0.00200  0.003531867 0.003478708
## 9        weight 49.89512 110.2229 74.84268 75.018761507 7.017914033
```

## 2.Biểu đồ phân phối

Ta dùng biểu đồ Histogram để trực quan hóa phân phối một số biến




**Nhận xét**: Các dữ liệu đã gần như hội tụ về phân phối chuẩn

# IV. Kiểm định giả thuyết

Ta sẽ thực hiện một vài kiểm định giả thuyết để xem xét mối liên hệ giữa một vài thuộc tính

## 1.Chân thuận có ảnh hưởng đến mức lương ?

Ta xét giả thuyết và đối thuyết:

$H_0$: $u_L$ = $u_R$ (Trung bình mức lương của cầu thủ sử dụng chân trái và phải như nhau)

$H_1$: $u_L$ > $u_R$ (Trung bình mức lương của cầu thủ sử dụng chân trái tốt hơn chân phải)

### 1.1 Bảng tóm tắt dữ liệu:

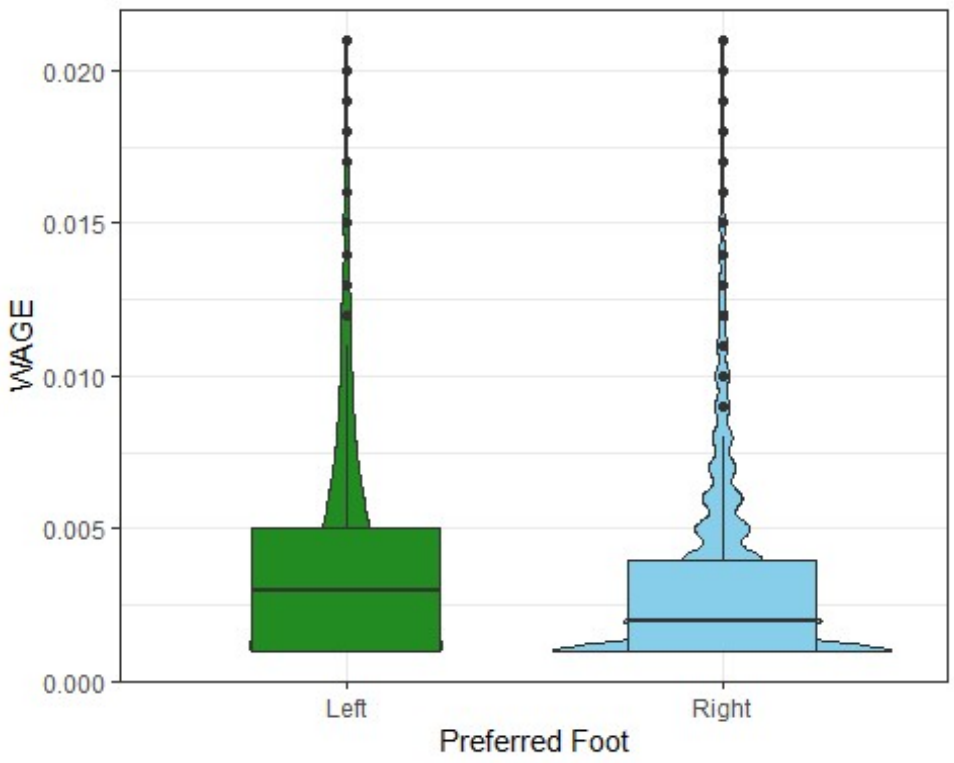
```
## # A tibble: 2 × 4
##   preferred_foot     n    mean      sd
##   <chr>          <int>   <dbl>   <dbl>
## 1 Left            3153 0.00379 0.00359
## 2 Right          10670 0.00346 0.00344
```

**Nhận xét**: Nếu chỉ dựa vào bảng số liệu, ta có thể thấy trung bình lương của cầu thủ thuận chân trái cao hơn chân phải

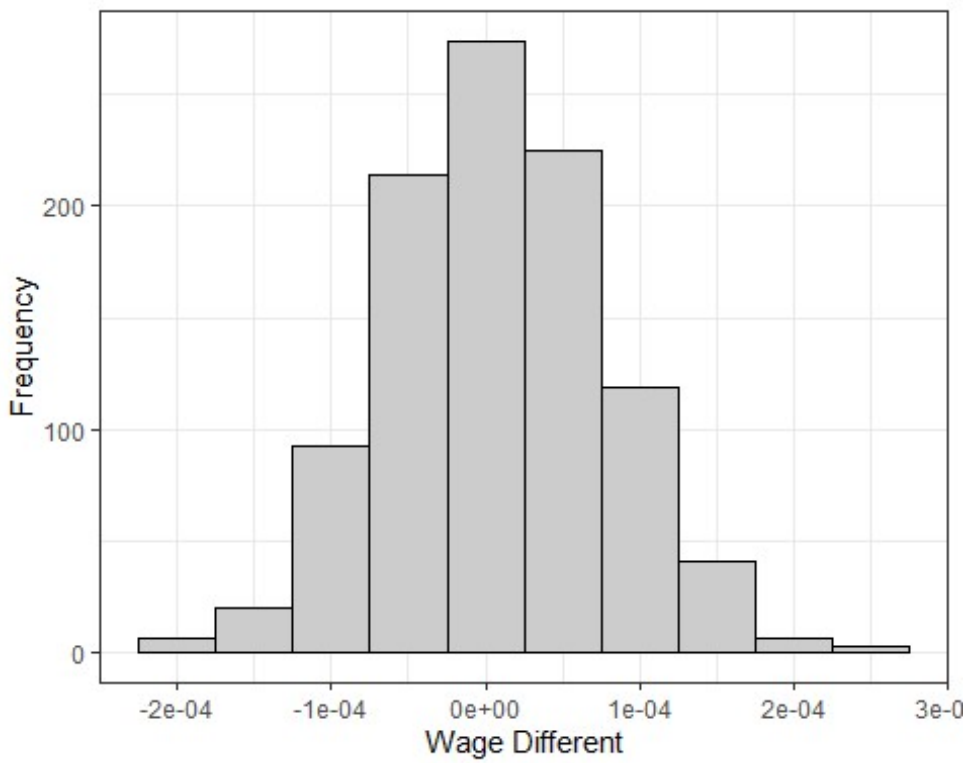
## 1.2 A/B Testing:

Ta thực hiện A/B Testing kết hợp Permutation Test

### *1.2.1 Biểu đồ chênh lệch*



### *1.2.2 Biểu đồ phân phối Histogram*




### *1.2.3 p-value*

```
p_value_wage
```

```
## [1] 0
```

**Nhận xét**: Với mức ý nghĩa $alpha$ = 0.05, $p_{value}$ < alpha => Cầu thủ thuận chân trái có mức lương tốt hơn thuận chân phải

## 2.Chân thuận có ảnh hưởng đến khả năng kiểm soát bóng ?

Ta xét giả thuyết và đối thuyết:

$H_0$: $u_L$ = $u_R$ (Khả năng kiểm soát bóng của cầu thủ sử dụng chân trái và phải như nhau)

$H_1$: $u_L$ > $u_R$ (Khả năng kiểm soát bóng của cầu thủ sử dụng chân trái tốt hơn chân phải)

### 2.1 Bảng tóm tắt dữ liệu:

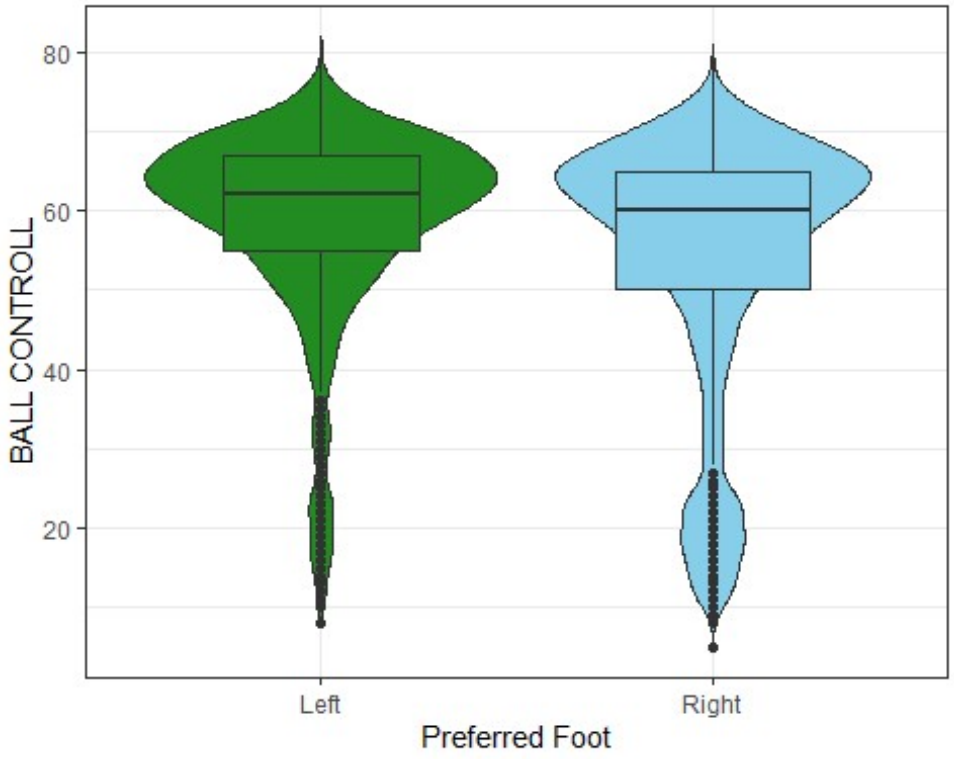
```
## # A tibble: 2 × 4
##   preferred_foot     n  mean    sd
##   <chr>          <int> <dbl> <dbl>
## 1 Left            3153  58.8  12.3
## 2 Right          10670  54.2  16.6
```

**Nhận xét**: Nếu chỉ dựa vào bảng số liệu, ta có thể thấy trung bình khả năng kiểm soát bóng của cầu thủ thuận chân trái cao hơn chân phải

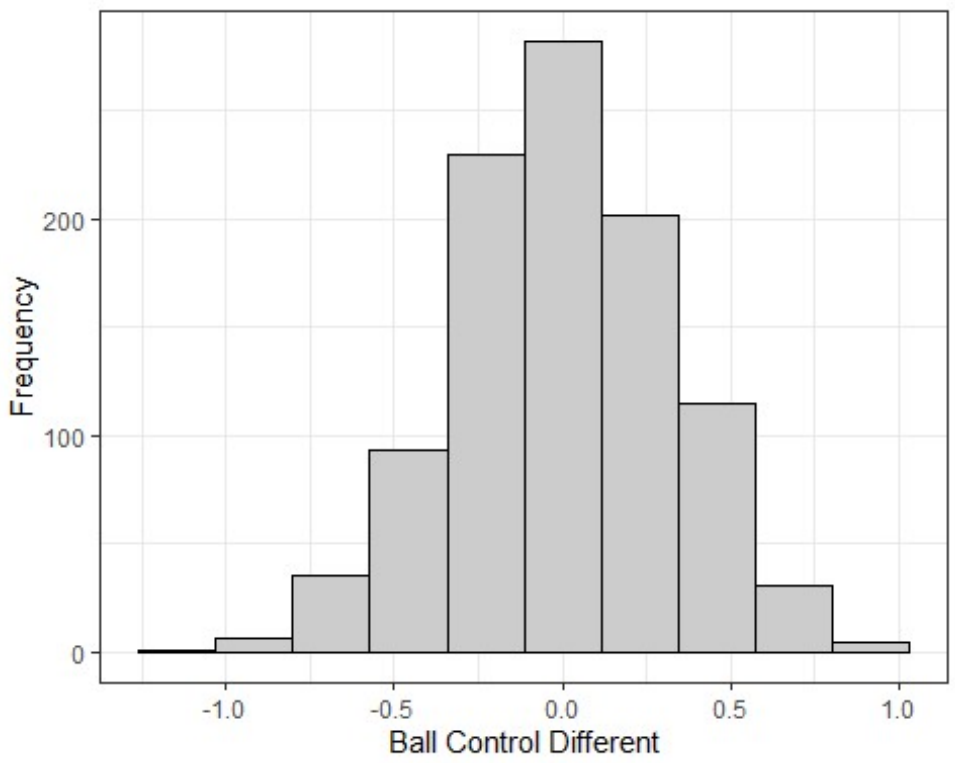
## 2.2 A/B Testing:

Ta thực hiện A/B Testing kết hợp Permutation Test

### *2.2.1 Biểu đồ chênh lệch*



### *2.2.2 Biểu đồ phân phối Histogram*




### *2.2.3 p-value*

```
p_value_bc
```

```
## [1] 0
```

**Nhận xét**: Với mức ý nghĩa $alpha$ = 0.05, $p_{value}$ < alpha => Cầu thủ thuận chân trái kiểm soát bóng tốt hơn thuận chân phải

## 3.Chân thuận có ảnh hưởng đến lực sút hay không ?

Ta xét giả thuyết và đối thuyết:

$H_0$: $u_L$ = $u_R$ (Lực sút của cầu thủ sử dụng chân trái và phải như nhau)

$H_1$: $u_L$ > $u_R$ (Lực sút của cầu thủ sử dụng chân trái tốt hơn chân phải)

## 3.1 Bảng tóm tắt dữ liệu:

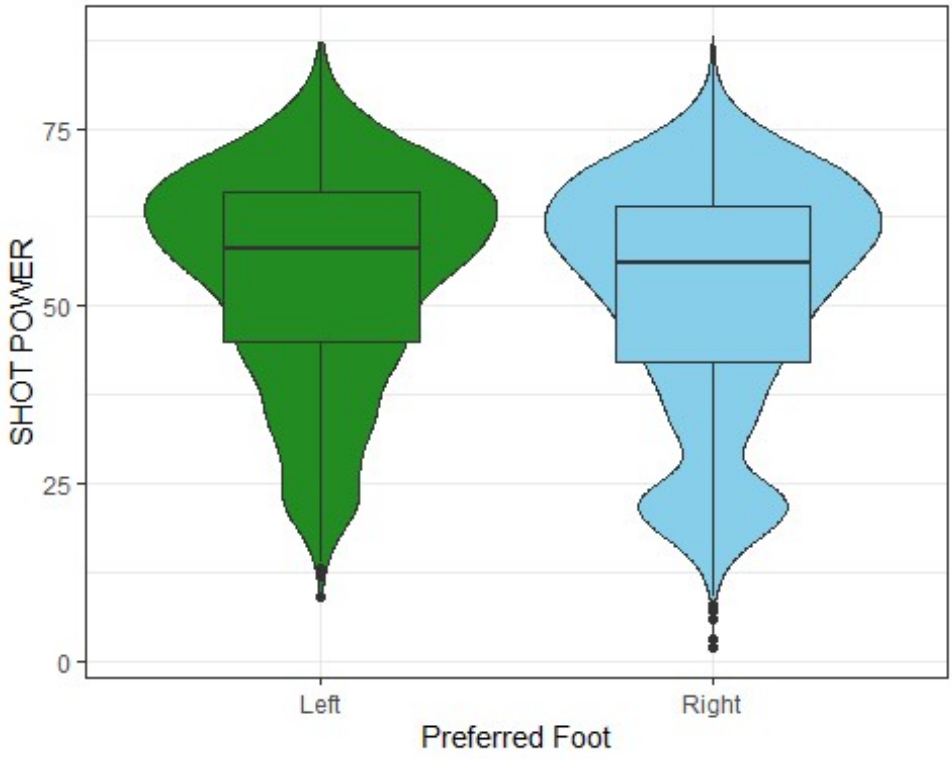
```
## # A tibble: 2 × 4
##   preferred_foot     n  mean    sd
##   <chr>          <int> <dbl> <dbl>
## 1 Left            3153  54.4  15.0
## 2 Right          10670  51.7  16.6
```

**Nhận xét**: Nếu chỉ dựa vào bảng số liệu, ta có thể thấy trung bình lực sút của cầu thủ thuận chân trái cao hơn chân phải

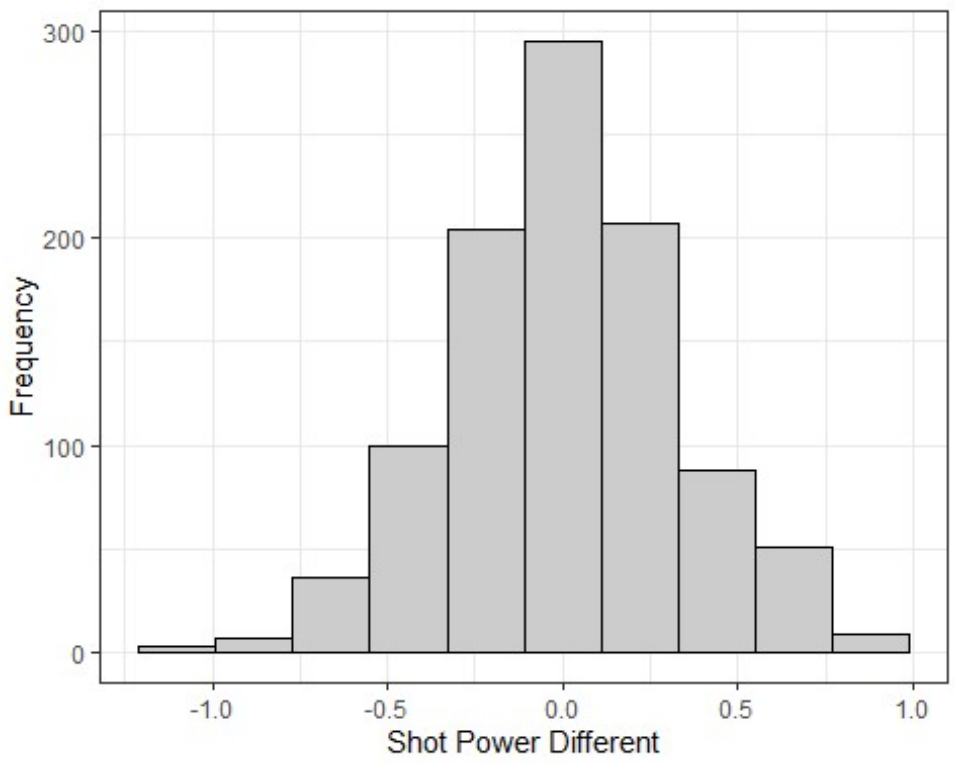
## 3.2 A/B Testing:

Ta thực hiện A/B Testing kết hợp Permutation Test

### *3.2.1 Biểu đồ chênh lệch*



### *3.2.2 Biểu đồ phân phối Histogram*




### *3.2.3 p-value*

```
p_value_shot
```

```
## [1] 0
```

**Nhận xét**: Với mức ý nghĩa $alpha$ = 0.05, $p_{value}$ < alpha => Cầu thủ chân trái có lực sút hơn chân phải

## 4.Vị trí của cầu thủ có ảnh hưởng đến mức lương hay không ?

Ta xét giả thuyết và đối thuyết:

$H_0$: $u_{ThuMo}$ = $u_{TienDao}$ = $u_{TienVe}$ = $u_{HauVe}$ (Lương giữa các vị trí là như nhau)

$H_1$: $u_{ThuM}$ ≠ $u_{TienDao}$ ≠ $u_{TienVe}$ ≠ $u_{HauVe}$ (Lương giữa các vị trí là khác nhau)

### 4.1 Kiểm tra và gom nhóm các vị trí

Ta sẽ kiểm tra xem có bao nhiêu vị trí cụ thể:

```
##    position
## 1        GK
## 2       LCB
## 3        RB
## 4        LB
```

```
## 5        CB
## 6       RDM
## 7       LCM
## 8        RS
## 9       RCB
## 10       CM
## 11      LDM
## 12      CDM
## 13       ST
## 14      RCM
## 15      CAM
## 16       LM
## 17       RW
## 18       LS
## 19      LAM
## 20      RAM
## 21      RWB
## 22       RM
## 23      LWB
## 24       CF
## 25       LF
## 26       LW
## 27       RF
```

Sau đó, ta sẽ gom nhóm theo 4 nhóm chính: Tiền đạo, Tiền vệ, Hậu vệ, Thủ môn

## 4.2 Bảng tóm tắt dữ liệu:

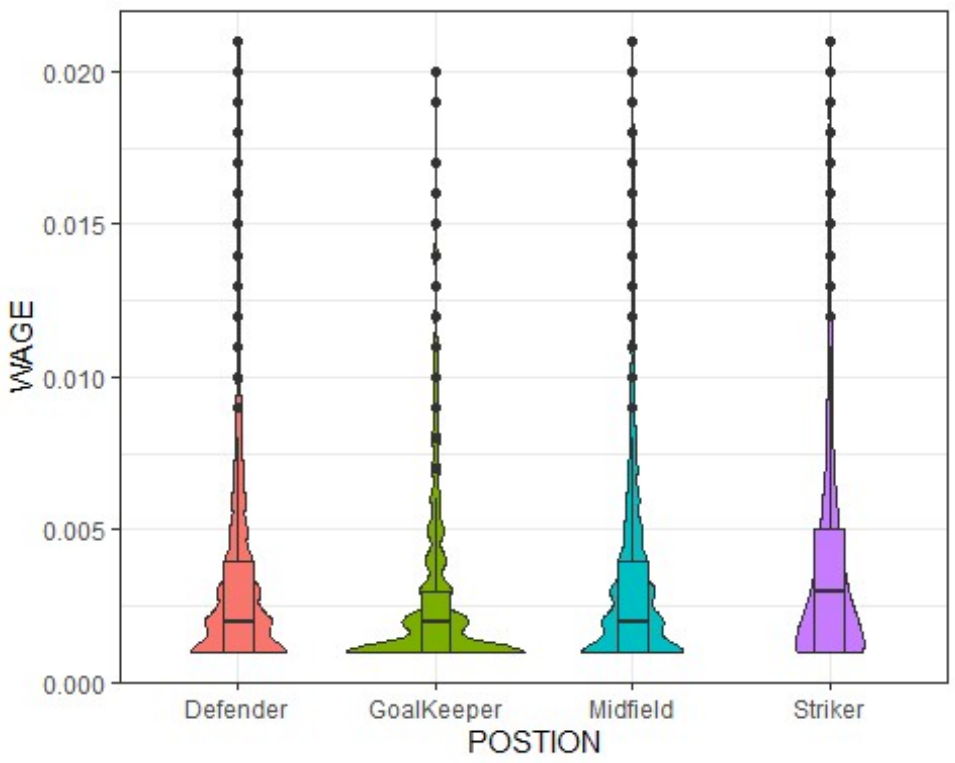
```
## # A tibble: 4 × 4
##   position_group     n    mean      sd
##   <chr>          <int>   <dbl>   <dbl>
## 1 Defender        4589 0.00360 0.00354
## 2 GoalKeeper      1692 0.00290 0.00314
## 3 Midfield        5020 0.00356 0.00347
## 4 Striker         2522 0.00378 0.00357
```

**Nhận xét**: Nếu chỉ dựa vào bảng số liệu, ta có thể thấy trung bình lương của Tiền đạo là cao nhất, sau đó đến lần lượt là Hậu vệ , Tiền vệ và cuối cùng là Thủ môn.

## 4.3 A/B Testing:

Ta thực hiện A/B Testing kết hợp Permutation Test

*4.3.1 Biểu đồ chênh lệch*




*4.3.2 p-value*

**Nhận xét**: Với mức ý nghĩa $alpha$ = 0.05, $p_{value}$ < alpha => Các vị trí khác nhau sẽ có mức lương trung bình khác nhau.

## 5.Độ tuổi của cầu thủ có ảnh hưởng đến chỉ số tổng thể hay không ?

Ta xét giả thuyết và đối thuyết:

$H_0$: $u_1 = u_2 = u_3 = u_4$ (Trung bình tổng thể giữa các độ tuổi là như nhau)

$H_1$: $u_1 \neq u_2 \neq u_3 \neq u_4$ (Trung bình tổng thể giữa các độ tuổi là khác nhau)

### 5.1 Kiểm tra và gom nhóm độ tuổi

Ta sẽ chia thành 4 độ tuổi dựa trên mức phân vị của dữ liệu.

```
##     25% 50% 75%
##  16  21  24  28  45
```

### 5.2 Bảng tóm tắt dữ liệu:

```
## # A tibble: 4 × 4
##   age_group     n  mean    sd
##   <chr>     <int> <dbl> <dbl>
## 1 (21, 24]   3040  63.7  4.16
```

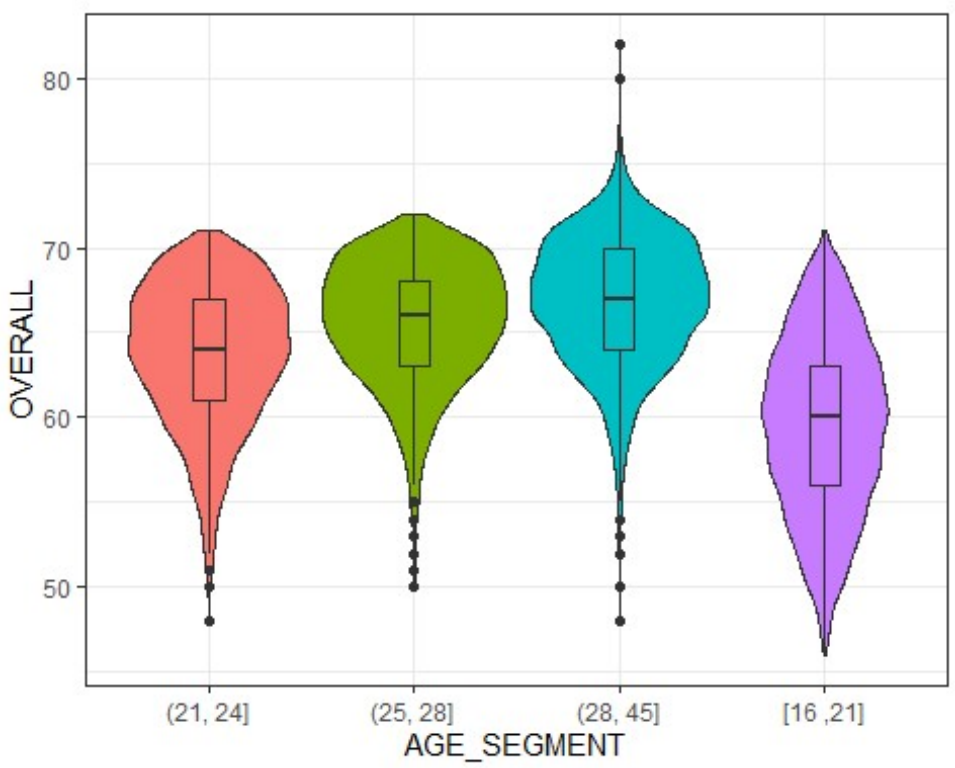
```
## 2 (25, 28]   3330  65.6  3.63
## 3 (28, 45]   3185  66.9  3.81
## 4 [16 ,21]   4268  59.5  5.10
```

**Nhận xét**: Nếu chỉ dựa vào bảng số liệu, ta có thể thấy trung bình chỉ số tổng thể của độ tuổi 28 - 45 là cao nhất, sau đó đến lần lượt là 25 - 28, 21 - 24 và cuối cùng là 16 - 21.

## 5.3 A/B Testing:

Ta thực hiện A/B Testing kết hợp Permutation Test

### *5.3.1 Biểu đồ chênh lệch*




### *5.3.2 p-value*

```
## Component 1 :
##                Df R Sum Sq R Mean Sq Iter  Pr(Prob)
## age_group1      3   118849     39616 5000 < 2.2e-16 ***
## Residuals   13819   253888        18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Nhận xét**: Với mức ý nghĩa $alpha$ = 0.05, $p_{value}$ < alpha => Các độ tuổi khác nhau sẽ có chỉ số tổng thể khác nhau.

# V. Mô hình hồi quy tuyến tính

## 1. Feature Engineering:

### 1.1 Biến Nationality:

Biến Nationality đang ở dạng phân loại, ta sẽ đưa biến về dạng số

```
## [1] "Nationality trước khi biến đổi: "

## [1] "Spain"     "Spain"     "Italy"     "Argentina" "Uruguay"

## [1] "Nationality sau khi biến đổi: "

## [1] 139 139  78   7 156
```

### 1.2 Biến club:

Biến Club đang ở dạng phân loại, ta sẽ đưa biến về dạng số

```
## [1] "Club trước khi biến đổi: "

## [1] "FC Porto"      "RCD Espanyol"  "Chievo Verona" "River Plate"
## [5] "FC Porto"

## [1] "Club sau khi biến đổi: "

## [1] 233 468 139 482 233
```

### 1.3 Biến prefered_foot:

Biến prefered_foot đang ở dạng phân loại, ta sẽ đưa biến về dạng số

```
## [1] "Chân thuận trước khi biến đổi: "

## [1] "Left"  "Right" "Right" "Left"  "Right"

## [1] "Chân thuận sau khi biến đổi: "

## [1] 0 1 1 0 1
```

### 1.4 Gom nhóm các vị trí:

Ta nhận thấy rằng, một vị trí khác nhau trên sân sẽ có những biến đáp ứng phù hợp khác nhau, nên ta sẽ gom nhóm và thực hiện mô hình theo từng vị trí.

## 2. Mô hình hồi quy cho tổng thể (Overall):

### 2.1 Thực hiện mô hình

Ta chỉ dùng các biến chỉ số (stats) của cầu thủ để dự đoán cho mô hình hồi quy dự đoán chỉ số Overall

```r
# In kết quả của mô hình
summary(overall_model)
```

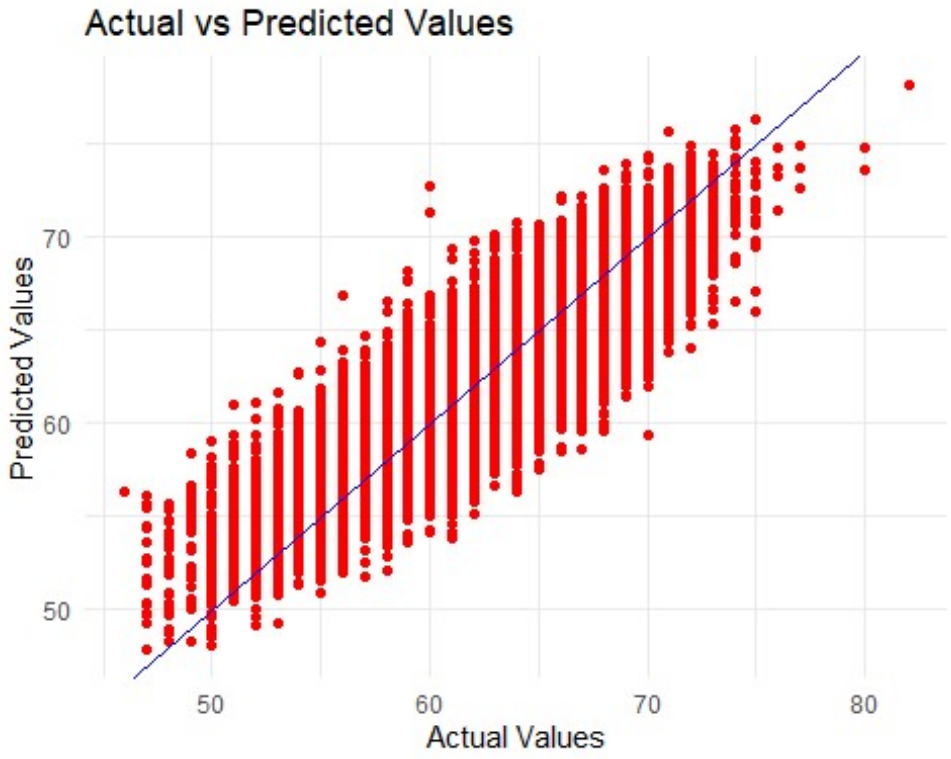
```
##
## Call:
## lm(formula = overall_formula, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.6946  -1.5666   0.0263   1.6186  10.6457
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      15.5482215  0.3116881  49.884  < 2e-16 ***
## crossing          0.0255147  0.0028571   8.930  < 2e-16 ***
## finishing         0.0249122  0.0036442   6.836 8.48e-12 ***
## heading_accuracy  0.0901723  0.0030954  29.132  < 2e-16 ***
## short_passing     0.0633011  0.0049221  12.860  < 2e-16 ***
## volleys           0.0001586  0.0032008   0.050  0.96049
## dribbling         0.0092761  0.0043346   2.140  0.03237 *
## curve             0.0092908  0.0031015   2.996  0.00274 **
## fk_accuracy       0.0042769  0.0028238   1.515  0.12989
## long_passing     -0.0212367  0.0037228  -5.704 1.19e-08 ***
## ball_control      0.1388071  0.0052813  26.283  < 2e-16 ***
## acceleration      0.0245346  0.0041185   5.957 2.63e-09 ***
## sprint_speed      0.0202616  0.0038543   5.257 1.49e-07 ***
## agility           0.0092926  0.0030211   3.076  0.00210 **
## reactions         0.2239427  0.0041292  54.234  < 2e-16 ***
## balance          -0.0202140  0.0027711  -7.295 3.16e-13 ***
## shot_power        0.0210914  0.0031272   6.744 1.60e-11 ***
## jumping           0.0094578  0.0022195   4.261 2.05e-05 ***
## stamina           0.0182397  0.0025002   7.295 3.14e-13 ***
## strength          0.0405743  0.0026049  15.576  < 2e-16 ***
## long_shots       -0.0150189  0.0033795  -4.444 8.89e-06 ***
## aggression        0.0003776  0.0023840   0.158  0.87414
## interceptions     0.0005898  0.0034453   0.171  0.86407
## positioning      -0.0414167  0.0032819 -12.620  < 2e-16 ***
## vision           -0.0187179  0.0030505  -6.136 8.69e-10 ***
## penalties         0.0021793  0.0030231   0.721  0.47100
## composure         0.1023534  0.0032408  31.583  < 2e-16 ***
## marking           0.0337005  0.0027841  12.105  < 2e-16 ***
## standing_tackle   0.0270676  0.0052903   5.116 3.15e-07 ***
## sliding_tackle   -0.0216442  0.0049459  -4.376 1.22e-05 ***
## gk_diving         0.0599779  0.0063456   9.452  < 2e-16 ***
## gk_handling       0.0681598  0.0064758  10.525  < 2e-16 ***
## gk_kicking        0.0321502  0.0059805   5.376 7.75e-08 ***
## gk_positioning    0.0726863  0.0062922  11.552  < 2e-16 ***
## gk_reflexes       0.0784910  0.0063311  12.398  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.48 on 13788 degrees of freedom
## Multiple R-squared:  0.7725, Adjusted R-squared:  0.772
## F-statistic:  1377 on 34 and 13788 DF,  p-value: < 2.2e-16
```

**Nhận xét**: Dựa vào kết quả của mô hình ta có thể đúc kết một vài nhận xét

- Các biến reactions, ball_controll, composure lần lượt là 4 chỉ số có trọng số cao (> 0.1) trong việc quyết định Overall.
- Các biến volleys, aggression, interceptions là các biến không ảnh hưởng đáng kể đến mô hình (p-value > 0.1).
- $R_2$ Score ~ 0.7725 cho thấy mô hình có ý nghĩa giải thích ở mức khá.

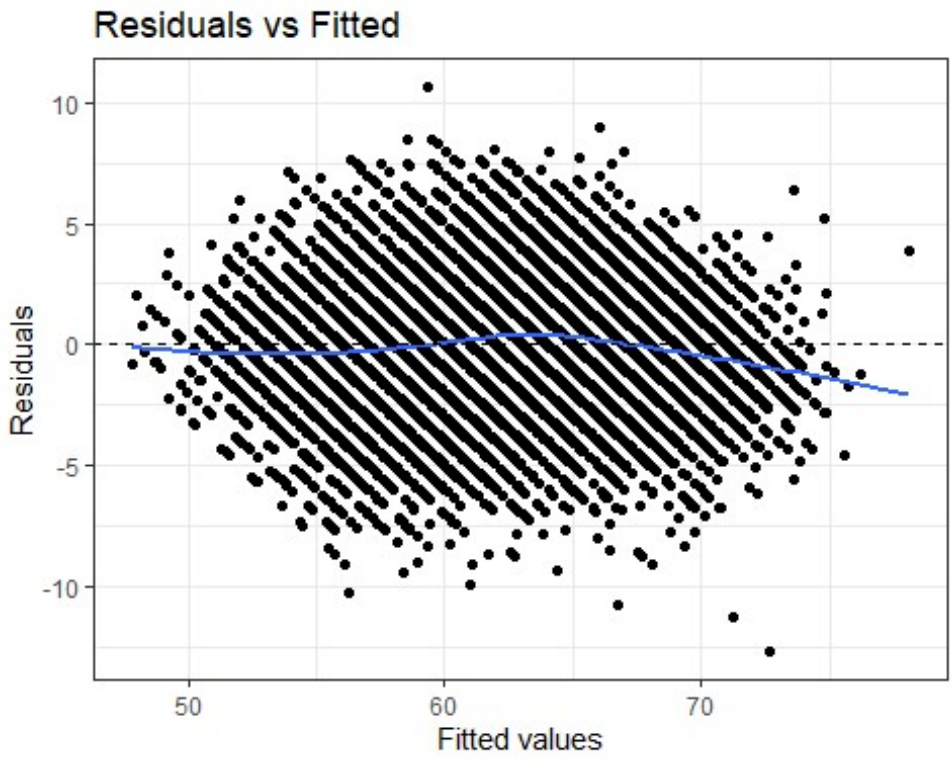
## 2.2 Biểu đồ giá trị thực tế và dự đoán




**Nhận xét**: Đường hồi quy của mô hình di chuyển khá khớp với các điểm dữ liệu

## 2.3 Một số chuẩn đoán cho mô hình

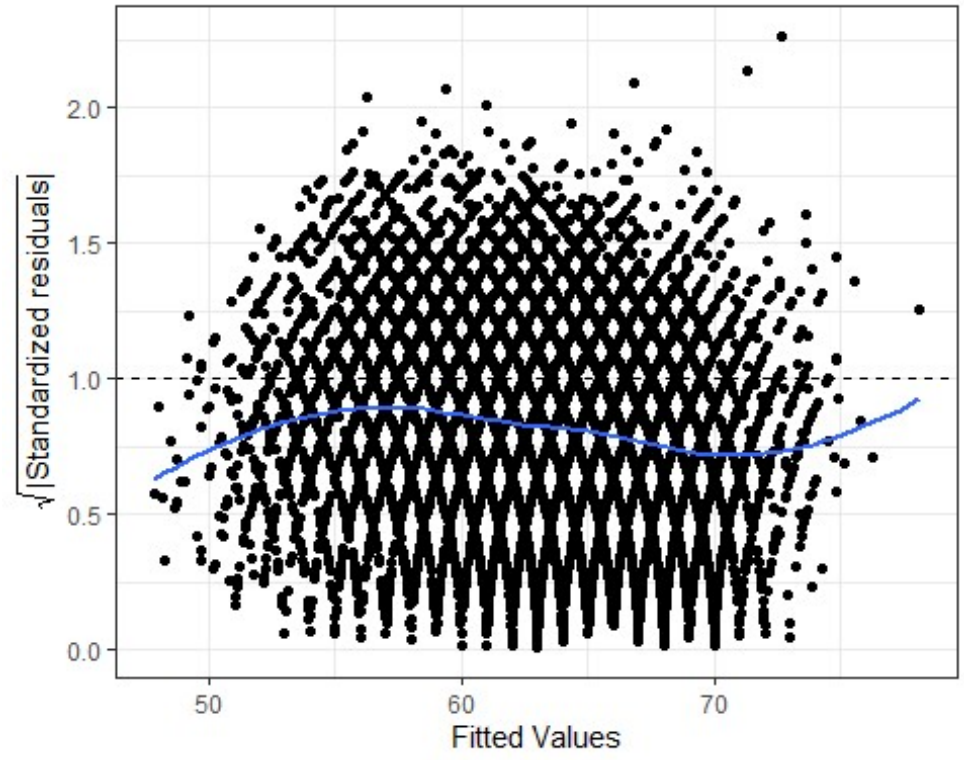
### *2.3.1 Kiểm định tuyến tính mô hình*




**Nhận xét**: Hình vẽ không cho thấy xu hướng một đường cong đáng kể

=> Giả định về tuyến tính của mô hình là phù hợp

### *2.3.2 Kiểm tra tính đồng nhất phương sai*




**Nhận xét**: Đường thẳng không xấp xỉ 1

=> Phương sai không đồng nhất

### *2.3.3 Kiểm tra đa cộng tuyến*

```
vif(overall_model)
```

```
##         crossing        finishing heading_accuracy    short_passing
##         5.552935        10.172675         6.020769        10.520131
##          volleys        dribbling            curve      fk_accuracy
##         6.045662        13.836927         6.214723         4.622659
##     long_passing     ball_control     acceleration     sprint_speed
##         6.547098        15.664924         8.578576         7.293810
##          agility        reactions          balance       shot_power
##         4.434670         2.159674         3.406125         5.821774
##          jumping          stamina         strength       long_shots
##         1.514201         3.536436         2.398477         8.325097
##       aggression    interceptions      positioning           vision
##         3.619193        10.455563         8.247791         3.495460
##        penalties        composure          marking  standing_tackle
##         4.560617         2.442009         6.312105        27.229225
##   sliding_tackle        gk_diving      gk_handling       gk_kicking
##        22.700176        28.267235        26.931141        22.099031
##   gk_positioning      gk_reflexes
##        25.529147        28.870321
```

**Nhận xét**: Ta thấy mô hình xảy ra đa cộng tuyến ở nhiều biến, ta sẽ xử lý và khởi tạo lại mô hình

Ta sẽ loại bỏ biến có VIF cao nhất sau đó tính lại VIF của toàn bộ mô hình cho đến khi VIF của tất cả các biến đều < 10

Ta có danh sách các biến sau khi xử lý

```
##  [1] "crossing"         "finishing"        "heading_accuracy" "short_passi
ng"
##  [5] "volleys"          "curve"            "fk_accuracy"      "long_passin
g"
##  [9] "acceleration"     "sprint_speed"     "agility"          "reactions"
## [13] "balance"          "shot_power"       "jumping"          "stamina"
## [17] "strength"         "long_shots"       "aggression"       "interceptio
ns"
## [21] "positioning"      "vision"           "penalties"        "composure"
## [25] "marking"          "gk_kicking"
```

Khởi tạo lại mô hình

```
summary(overall_model)
```

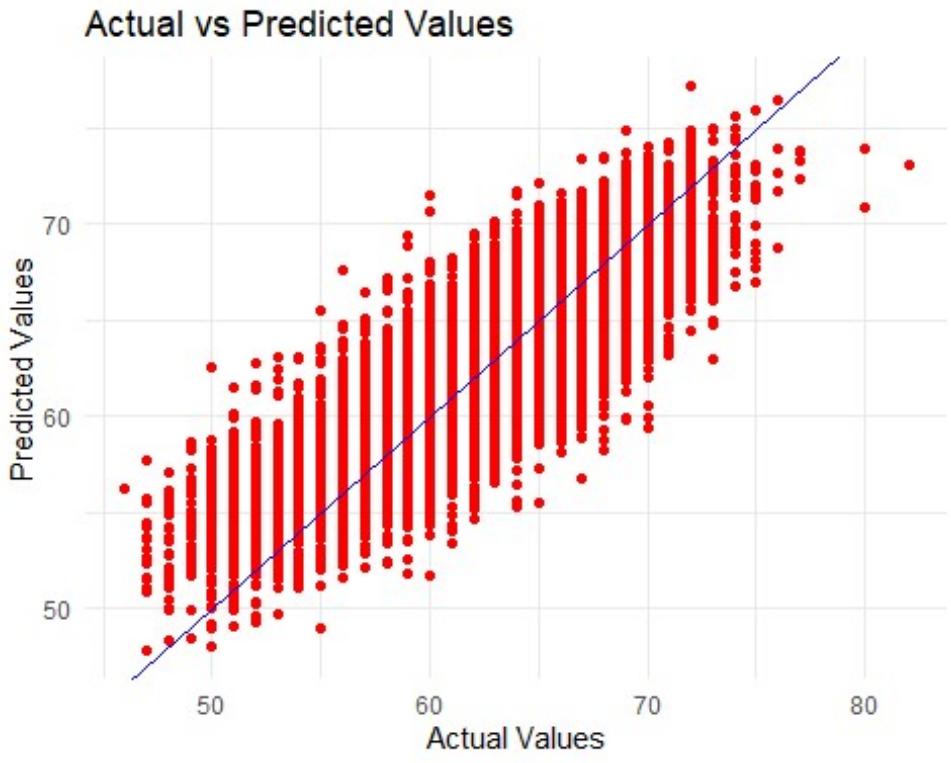
```
##
## Call:
## lm(formula = overall_formula, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.5126  -1.7105   0.0524   1.7540  10.6376
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      19.2760316  0.3207496  60.097  < 2e-16 ***
## crossing          0.0407148  0.0029721  13.699  < 2e-16 ***
## finishing         0.0297898  0.0038878   7.662 1.95e-14 ***
## heading_accuracy  0.0641490  0.0032356  19.826  < 2e-16 ***
## short_passing     0.1002124  0.0049568  20.217  < 2e-16 ***
## volleys          -0.0022428  0.0034697  -0.646 0.518039
## curve             0.0120754  0.0033503   3.604 0.000314 ***
## fk_accuracy      -0.0018164  0.0030515  -0.595 0.551681
## long_passing     -0.0162090  0.0040202  -4.032 5.56e-05 ***
## acceleration      0.0231739  0.0044449   5.214 1.88e-07 ***
## sprint_speed      0.0197895  0.0041772   4.737 2.19e-06 ***
## agility           0.0180110  0.0032491   5.543 3.02e-08 ***
## reactions         0.2796384  0.0042861  65.243  < 2e-16 ***
## balance          -0.0251754  0.0029986  -8.396  < 2e-16 ***
## shot_power        0.0248802  0.0033818   7.357 1.99e-13 ***
## jumping           0.0112348  0.0023877   4.705 2.56e-06 ***
## stamina           0.0122547  0.0027062   4.528 5.99e-06 ***
## strength          0.0486025  0.0028078  17.310  < 2e-16 ***
```

```
## long_shots       -0.0064530  0.0036516  -1.767 0.077224 .
## aggression       -0.0066004  0.0025616  -2.577 0.009984 **
## interceptions    -0.0009187  0.0029859  -0.308 0.758332
## positioning      -0.0399465  0.0034568 -11.556  < 2e-16 ***
## vision           -0.0103764  0.0032928  -3.151 0.001629 **
## penalties        -0.0013957  0.0032706  -0.427 0.669585
## composure         0.1203550  0.0034825  34.560  < 2e-16 ***
## marking           0.0246474  0.0028269   8.719  < 2e-16 ***
## gk_kicking        0.2140908  0.0037661  56.846  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.691 on 13796 degrees of freedom
## Multiple R-squared:  0.7319, Adjusted R-squared:  0.7314
## F-statistic:  1449 on 26 and 13796 DF,  p-value: < 2.2e-16
```

**Nhận xét**:

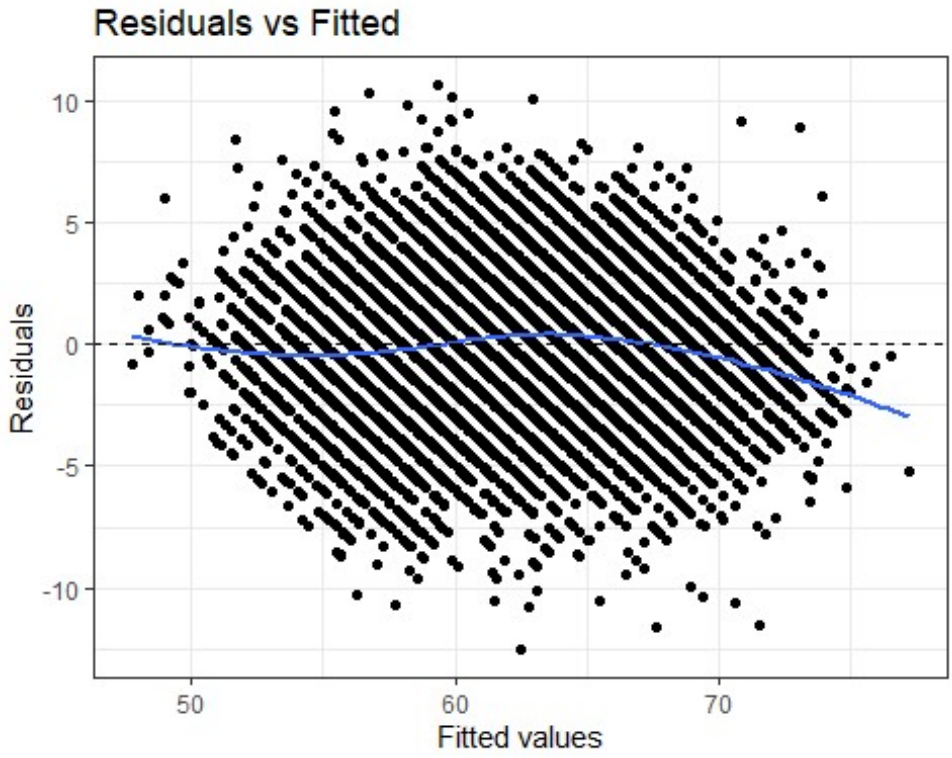
- Mô hình cho ra R2-Score khá cao, chứng tỏ mô hình giải thích tốt cho biến mục tiêu
- Các biến có mối quan hệ tuyến tính.

Biểu đồ đường thẳng hồi quy




**Nhận xét**: Các điểm dữ liệu phân bố khá khớp với đường thẳng hồi quy

### Kiểm định tuyến tính cho mô hình mới

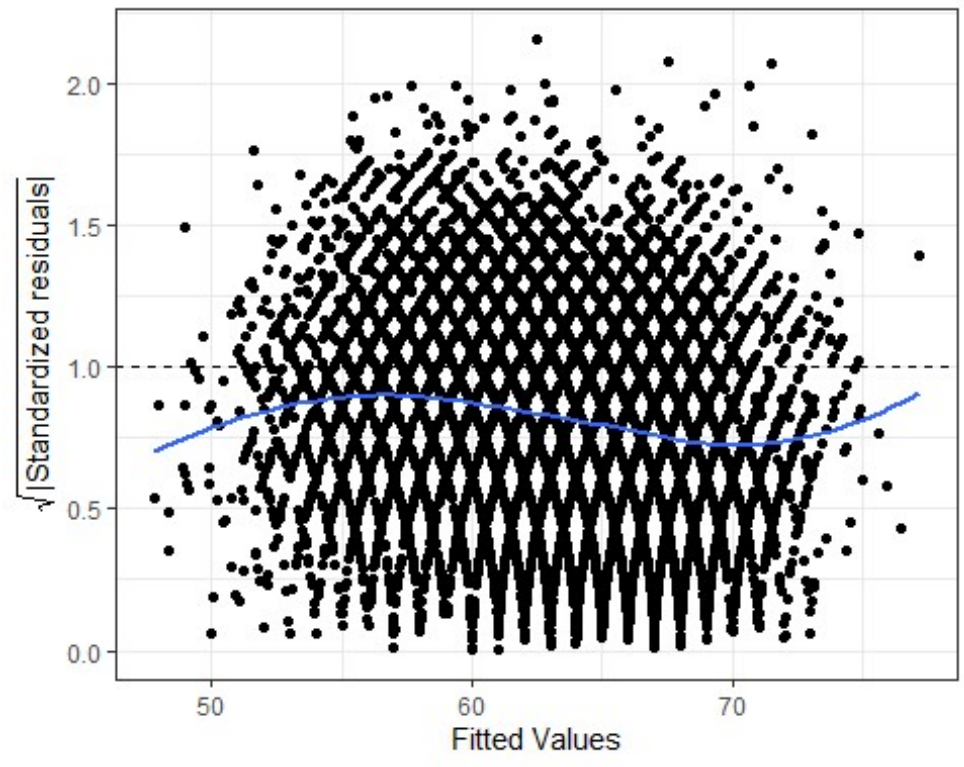



**Nhận xét**: Đồ thị không cho thấy 1 xu hướng cong cụ thể

=> Giả định về mối liên hệ tuyến tính giữa các biến là phù hợp.

### Kiểm định đồng nhất phương sai cho mô hình mới

```
## `geom_smooth()` using formula = 'y ~ x'
```




**Nhận xét**: Đồ thị chưa khớp lắm với đường thẳng 1

=> Giả định về đồng nhất phương sai là chưa phù hợp.

*Đánh giá mô hình*

RMSE của mô hình

```
## [1] "RMSE từ tập kiểm tra:  2.51080556990446"
```

*R2 Score*

```
## [1] "R2 Score của mô hình là: 0.731917085371723"
```

*Áp dụng k-fold validation để tính RMSE*

```
## [1] "RMSE từ K-Fold CV (k = 5 ): 2.98007900900518"
```

## 3. Mô hình hồi quy cho từng vị trí

### Phân chia dữ liệu:

Tách dữ liệu theo từng vị trí:

Các vị trí như Tiền đạo, hậu vệ, tiền vệ sẽ không cần các chỉ số của Thủ môn, nên ta sẽ loại bỏ các biến này

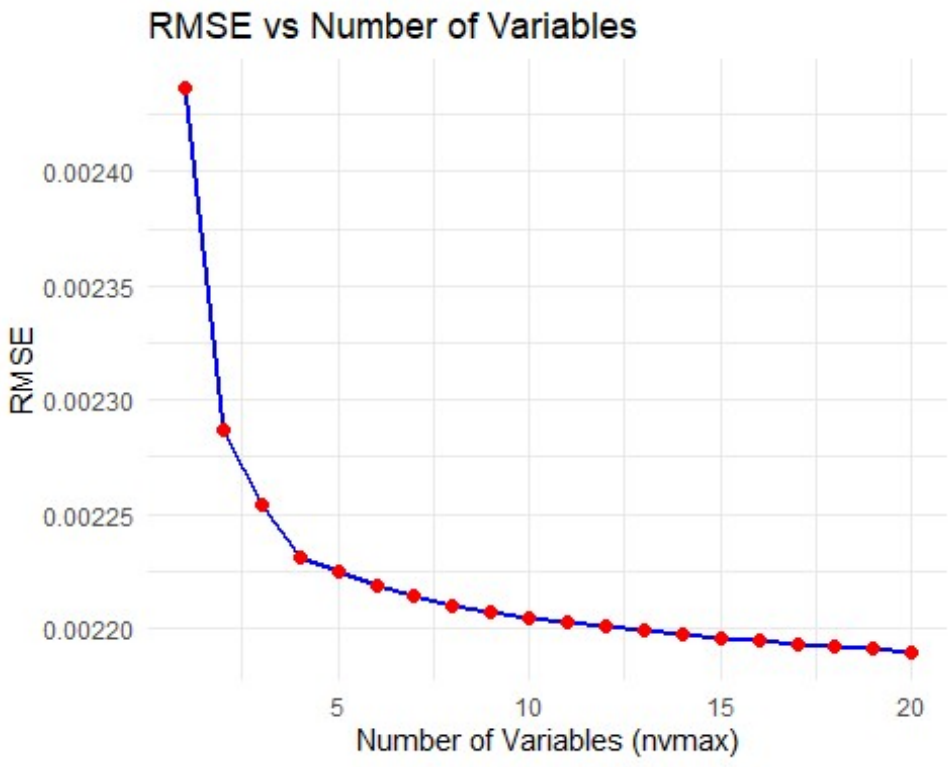
## 3.1 Mô hình hồi quy cho Goal Keeper:

### *3.1.1 Lựa chọn biến phù hợp với mô hình*

Các biến lựa chọn là

```
##  [1] "age"                      "overall"
##  [3] "value"                    "preferred_foot"
##  [5] "international_reputation" "height"
##  [7] "crossing"                 "short_passing"
##  [9] "fk_accuracy"              "long_passing"
## [11] "sprint_speed"             "reactions"
## [13] "jumping"                  "aggression"
## [15] "positioning"              "gk_diving"
## [17] "gk_handling"              "gk_kicking"
## [19] "gk_reflexes"              "release_clause"
```

### *3.1.2 Biểu đồ RMSE theo các biến đã chọn*




**Nhận xét**: Vậy, các biến trên là các biến phù hợp

### *3.1.3 Áp dụng mô hình cho các biến đã chọn*

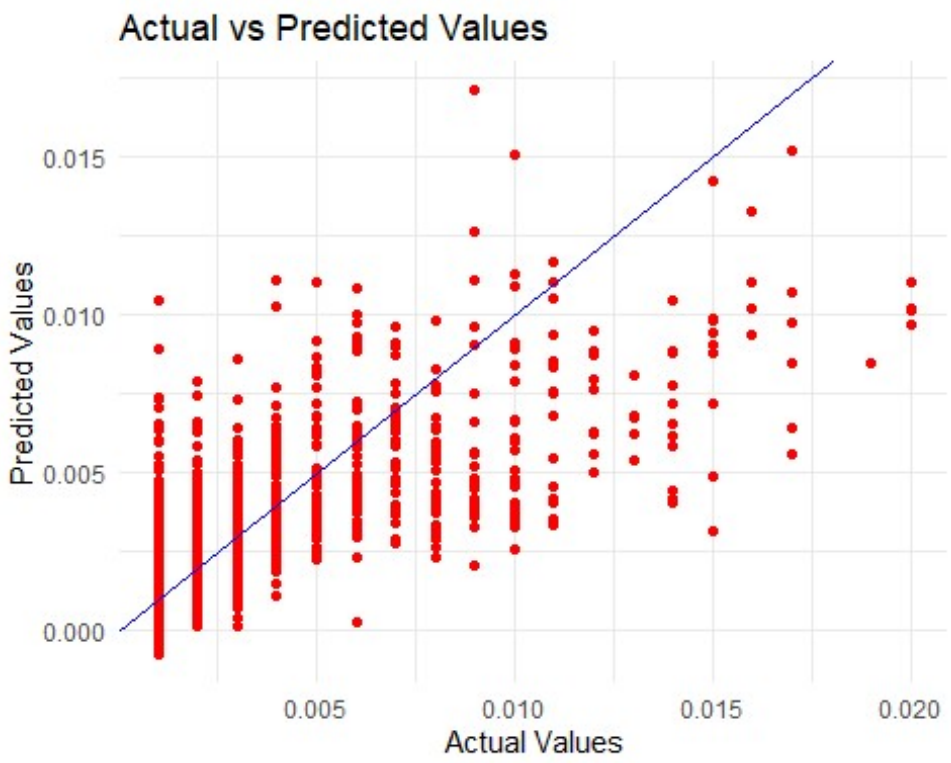
```
##
## Call:
## lm(formula = gk_formula_final, data = df_gk_numeric)
##
## Residuals:
```

```
##        Min         1Q     Median         3Q        Max
## -0.0094265 -0.0011449 -0.0001792  0.0005829  0.0118551
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -1.223e-02  2.491e-03  -4.912 9.89e-07 ***
## age                       1.140e-04  1.551e-05   7.351 3.05e-13 ***
## overall                  -1.614e-04  6.721e-05  -2.402  0.01641 *
## value                     1.174e-03  2.980e-04   3.938 8.54e-05 ***
## preferred_foot           -4.759e-04  1.813e-04  -2.625  0.00874 **
## international_reputation  2.178e-03  3.284e-04   6.633 4.43e-11 ***
## height                    2.990e-03  1.225e-03   2.442  0.01471 *
## crossing                 -3.279e-05  1.618e-05  -2.026  0.04290 *
## short_passing             1.913e-05  9.419e-06   2.032  0.04235 *
## fk_accuracy              -2.604e-05  1.372e-05  -1.899  0.05777 .
## long_passing             -2.008e-05  9.279e-06  -2.164  0.03059 *
## sprint_speed              9.600e-06  6.594e-06   1.456  0.14560
## reactions                 2.203e-05  1.213e-05   1.816  0.06958 .
## jumping                   1.128e-05  5.899e-06   1.913  0.05591 .
## aggression                1.666e-05  7.934e-06   2.100  0.03588 *
## positioning              -3.150e-05  1.596e-05  -1.973  0.04865 *
## gk_diving                 5.447e-05  2.350e-05   2.318  0.02059 *
## gk_handling               6.590e-05  2.445e-05   2.695  0.00711 **
## gk_kicking                2.768e-05  1.229e-05   2.253  0.02440 *
## gk_reflexes               4.348e-05  2.337e-05   1.861  0.06297 .
## release_clause            1.082e-03  1.261e-04   8.582  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002203 on 1671 degrees of freedom
## Multiple R-squared:  0.5125, Adjusted R-squared:  0.5067
## F-statistic: 87.84 on 20 and 1671 DF,  p-value: < 2.2e-16
```

**Nhận xét**:

- Mô hình Có $R_2$ Score ~ 0.5, mô hình dữ đoán tương đối mức lương của thủ môn.
- Các biến quan trọng ảnh hưởng đến wage của 1 thủ môn: age, value, inter_reputa.
- Một số biến không quan trọng: sprint_speed.

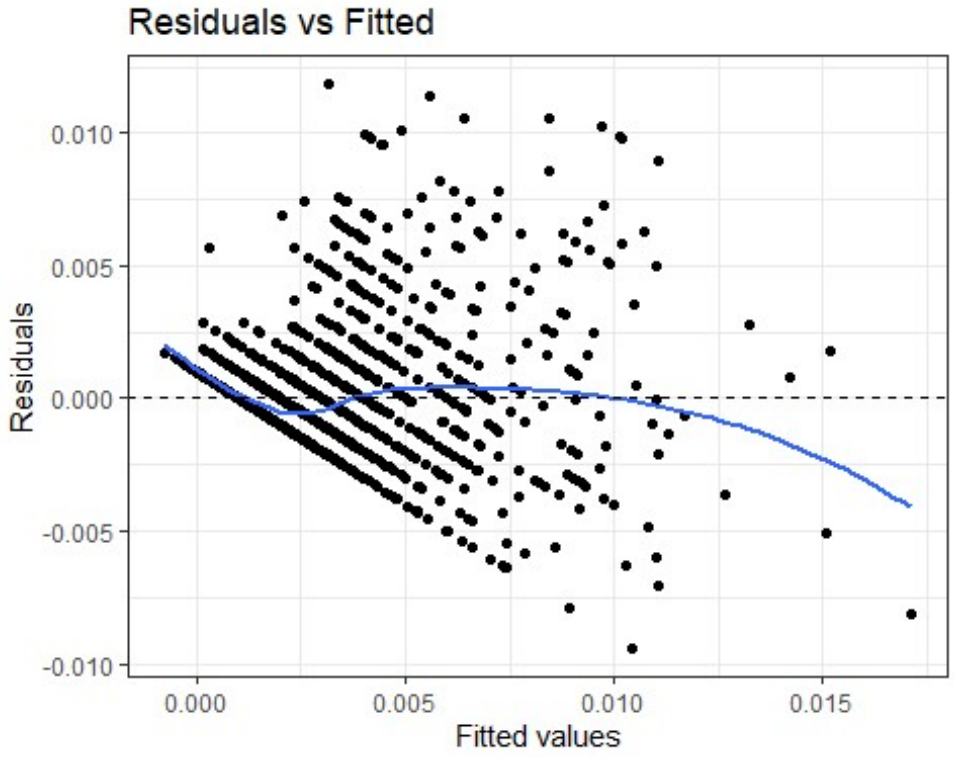
*3.1.4 Biểu đồ hồi quy*




**Nhận xét**: Đường thẳng hồi quy biểu diễn tương đối với các điểm dữ liệu

*3.1.5 Một số chuẩn đoán cho mô hình*

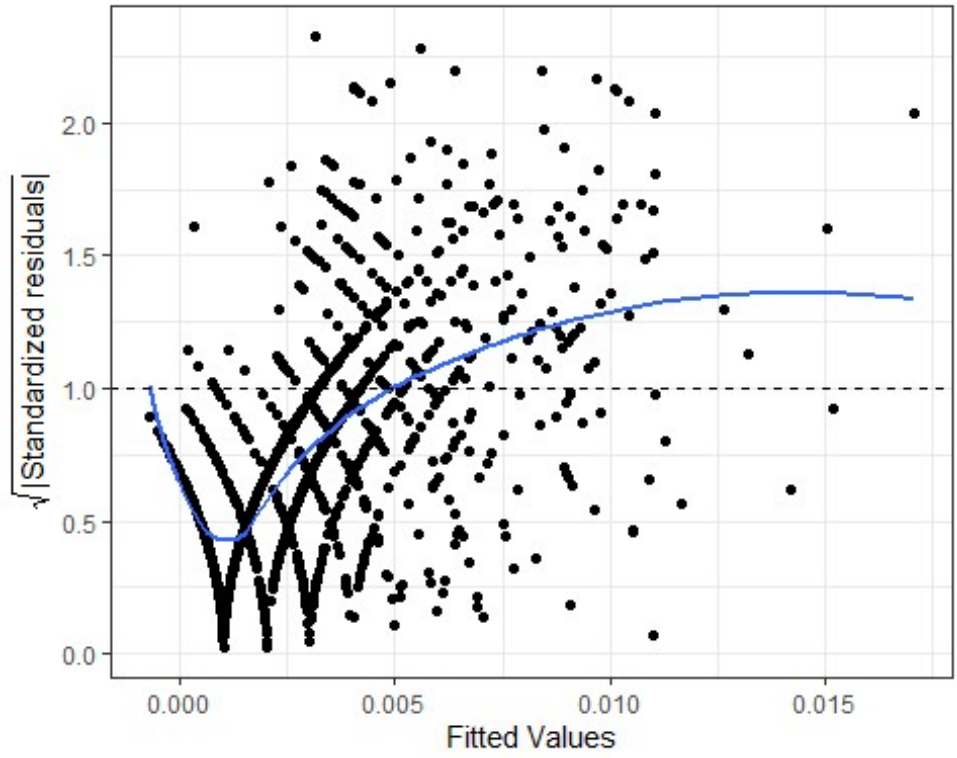
Kiểm định tuyến tính mô hình




**Nhận xét**: Hình vẽ không cho thấy xu hướng đường cong cụ thể

=> Giả định về tuyến tính của mô hình là phù hợp

### Kiểm định đồng nhất phương sai




**Nhận xét**: Đồ thị xấp xỉ đường thẳng 1

=> Giả định về phương sai đồng nhất là hợp lý

### Kiểm tra đa cộng tuyến

```
vif(gk_model_final)
```

```
##                      age                  overall                    value
##                 2.614630                56.518512                 7.560625
##          preferred_foot international_reputation                   height
##                 1.013639                 1.111261                 1.178275
##                crossing            short_passing             fk_accuracy
##                 1.367916                 1.770965                 1.303980
##            long_passing             sprint_speed               reactions
##                 1.807838                 1.712508                 4.371838
##                 jumping               aggression             positioning
##                 1.538795                 1.281977                 1.609332
##              gk_diving             gk_handling              gk_kicking
##                 7.475954                 7.693632                 2.249457
##             gk_reflexes          release_clause
##                 8.449822                 5.903059
```

**Nhận xét**: Mô hình xảy ra đa cộng tuyến ở biến overall.

Ta tiến hành loại bỏ biến overall và kiểm tra lại

```
##                      age                    value           preferred_foot
##                 2.336463                 7.478412                 1.012005
## international_reputation                   height                 crossing
##                 1.108806                 1.172409                 1.367808
##            short_passing             fk_accuracy             long_passing
##                 1.770391                 1.303705                 1.807176
##             sprint_speed                reactions                  jumping
##                 1.710004                 2.560455                 1.531703
##               aggression              positioning               gk_diving
##                 1.281242                 1.608974                 4.274074
##             gk_handling              gk_kicking              gk_reflexes
##                 2.854346                 1.930916                 4.358327
##          release_clause
##                 5.900788
```

**Nhận xét**: Mô hình đã không còn đa cộng tuyến

### *3.1.6 Đánh giá mô hình*

#### RMSE của mô hình

```
## [1] "RMSE từ tập kiểm tra:  0.00258115609438589"
```

#### *R2 Score*

```
## [1] "RMSE từ K-Fold CV (k = 5 ): 0.00223730512926805"
```

### *3.1.7 Mở rộng mô hình*

#### Khởi tạo mô hình

Ta sẽ tiến hành chạy mô hình bậc 2 với các biến có ít độ tuyến tính với mô hình

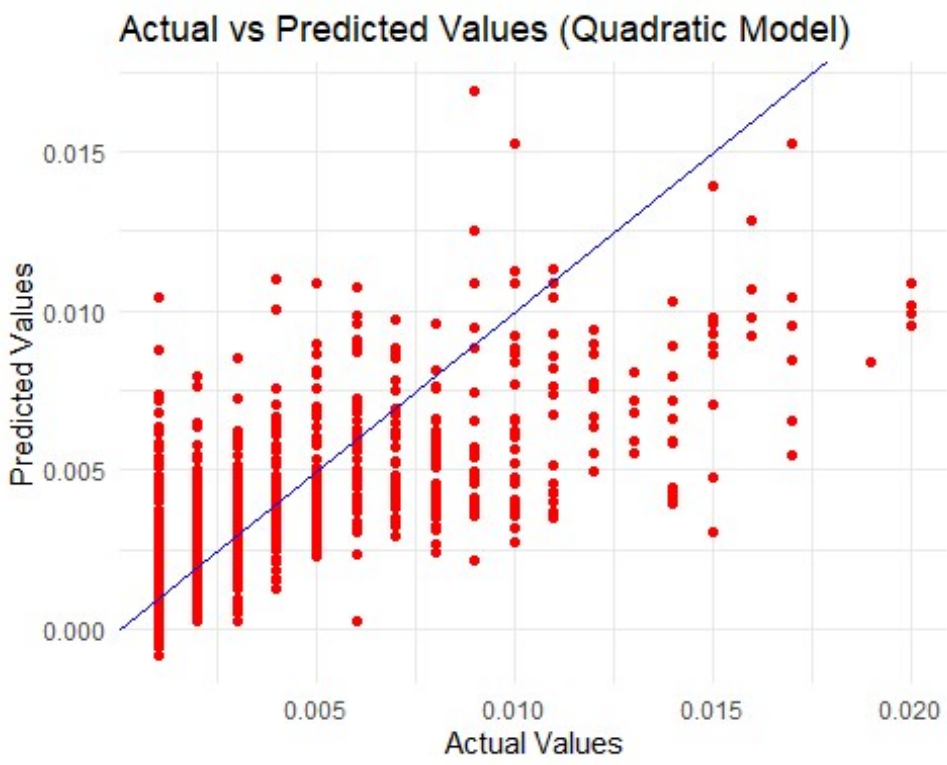
```
summary(gk_model_final)
```

```
##
## Call:
## lm(formula = gk_formula_final, data = df_gk_numeric)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0094518 -0.0011684 -0.0001765  0.0006419  0.0119524
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -1.046e-02  2.426e-03  -4.311 1.72e-05 ***
## age                       9.126e-05  1.493e-05   6.113 1.21e-09 ***
## value                     9.285e-04  3.022e-04   3.072  0.00216 **
## international_reputation  2.071e-03  3.289e-04   6.299 3.83e-10 ***
## height                    2.634e-03  1.220e-03   2.159  0.03102 *
## short_passing             1.436e-05  9.338e-06   1.538  0.12430
## fk_accuracy              -2.522e-05  1.394e-05  -1.809  0.07065 .
## jumping                   1.117e-05  5.871e-06   1.902  0.05729 .
## release_clause            1.079e-03  1.261e-04   8.555  < 2e-16 ***
```

```
## I(preferred_foot^2)      -4.684e-04  1.813e-04  -2.584  0.00986 **
## I(crossing^2)            -9.520e-07  4.537e-07  -2.098  0.03602 *
## I(long_passing^2)        -2.021e-07  1.556e-07  -1.299  0.19417
## I(sprint_speed^2)         1.025e-07  8.534e-08   1.201  0.23002
## I(reactions^2)            8.014e-08  8.486e-08   0.944  0.34511
## I(aggression^2)           2.522e-07  1.379e-07   1.828  0.06767 .
## I(positioning^2)         -1.400e-06  6.562e-07  -2.134  0.03300 *
## I(gk_diving^2)            1.719e-07  1.444e-07   1.190  0.23421
## I(gk_handling^2)          2.156e-07  1.231e-07   1.751  0.08015 .
## I(gk_kicking^2)           1.542e-07  9.520e-08   1.620  0.10539
## I(gk_reflexes^2)          5.409e-08  1.344e-07   0.402  0.68738
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002204 on 1672 degrees of freedom
## Multiple R-squared:  0.5119, Adjusted R-squared:  0.5064
## F-statistic:  92.3 on 19 and 1672 DF,  p-value: < 2.2e-16
```

Nhận xét: R2 Score tăng lên sau khi áp dụng mô hình bậc 2

### Đường thẳng hồi quy




### Đánh giá mô hình

```
## [1] "RMSE từ tập kiểm tra (Quadratic Model): 0.00250867873365324"

## [1] "R2 Score của mô hình bậc hai là: 0.511920338029513"
```

**Nhận xét**: Mô hình bậc 2 có kết quả chạy tốt hơn, với R2 Score cao hơn và RMSE thấp hơn.

## 3.2 Mô hình hồi quy dự đoán mức lương cho vị trí Tiền đạo

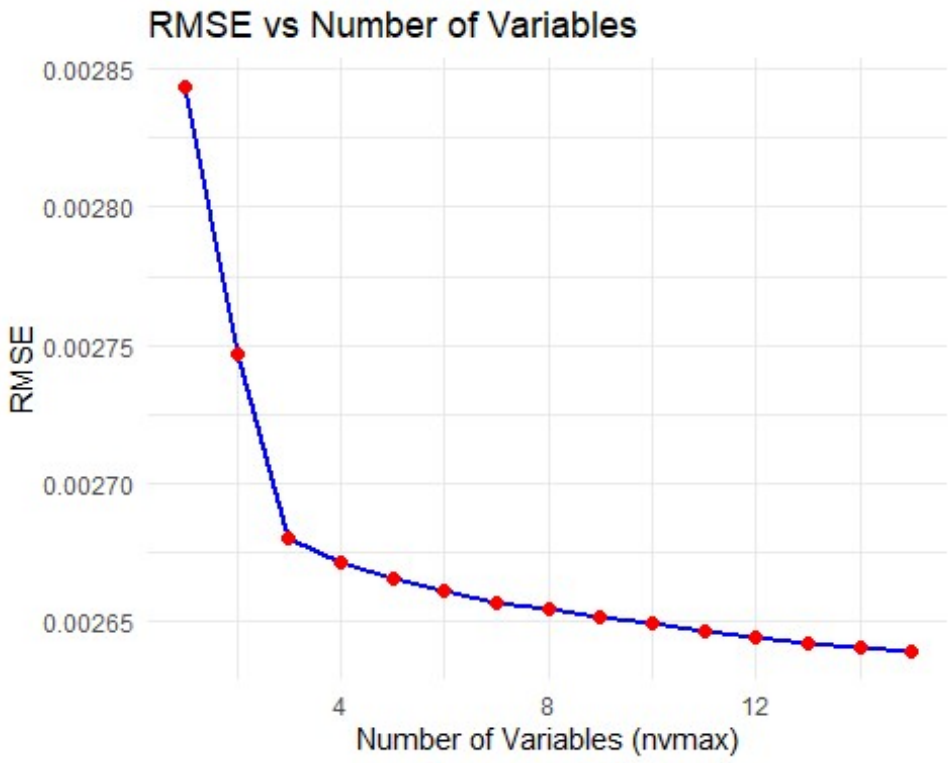
### *3.2.1 Lựa chọn biến phù hợp cho mô hình*

Các biến phù hợp để khởi tạo mô hình là:

```
print(selected_vars)
```

```
##  [1] "age"                    "potential"
##  [3] "value"                  "international_reputation"
##  [5] "body_type"              "crossing"
##  [7] "short_passing"          "volleys"
##  [9] "fk_accuracy"            "sprint_speed"
## [11] "long_shots"             "interceptions"
## [13] "penalties"              "marking"
## [15] "release_clause"
```

### *3.2.2 Biểu đồ RMSE cho các biến đã chọn*




**Nhận xét**: Vậy, các biến trên được chọn để khởi tạo mô hình là phù hợp

### *3.2.3 Áp dụng mô hình cho các biến đã chọn*

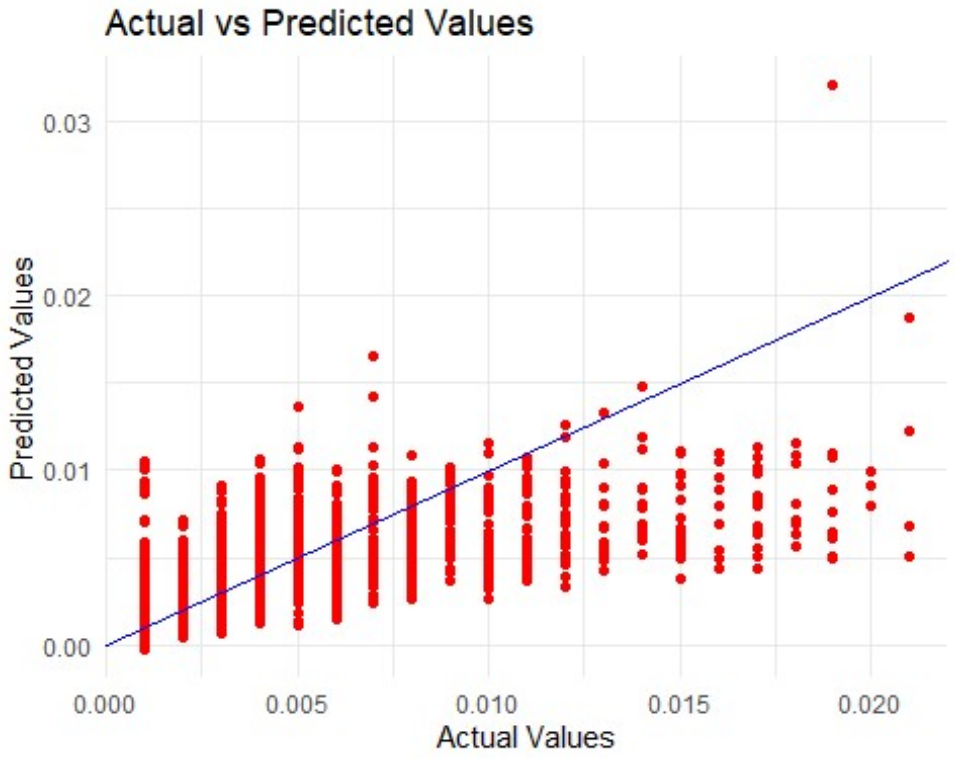
```
##
## Call:
## lm(formula = st_formula_final, data = df_st_numeric)
##
```

```
## Residuals:
##        Min          1Q      Median          3Q         Max
## -0.0130889 -0.0013881 -0.0003789  0.0006981  0.0158854
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -4.352e-03  1.451e-03  -2.999 0.002731 **
## age                       1.002e-04  2.140e-05   4.681    3e-06 ***
## potential                 2.702e-05  1.744e-05   1.549 0.121401
## value                     1.918e-03  1.756e-04  10.918  < 2e-16 ***
## international_reputation  1.089e-03  3.781e-04   2.880 0.004016 **
## body_type                -5.506e-05  3.040e-05  -1.811 0.070213 .
## crossing                  2.436e-05  6.580e-06   3.702 0.000218 ***
## short_passing            -2.770e-05  1.068e-05  -2.593 0.009572 **
## volleys                   1.311e-05  8.076e-06   1.623 0.104738
## fk_accuracy              -1.507e-05  5.841e-06  -2.580 0.009946 **
## sprint_speed             -2.141e-05  5.738e-06  -3.730 0.000195 ***
## long_shots                2.135e-05  9.599e-06   2.225 0.026200 *
## interceptions             1.938e-05  5.863e-06   3.305 0.000963 ***
## penalties                 1.862e-05  7.818e-06   2.381 0.017336 *
## marking                  -1.235e-05  5.413e-06  -2.281 0.022624 *
## release_clause            7.688e-04  5.099e-05  15.079  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002648 on 2506 degrees of freedom
## Multiple R-squared:  0.453,  Adjusted R-squared:  0.4497
## F-statistic: 138.4 on 15 and 2506 DF,  p-value: < 2.2e-16
```

**Nhận xét**:

- R2 Score: mô hình giải thích tương đối cho biến mục tiêu
- Một số biến ảnh hưởng mạnh đến wage của vị trí Tiền đạo: age, value, crossing.
- Một số biến không ảnh hưởng nhiều đến mô hình: body_type, penalties.

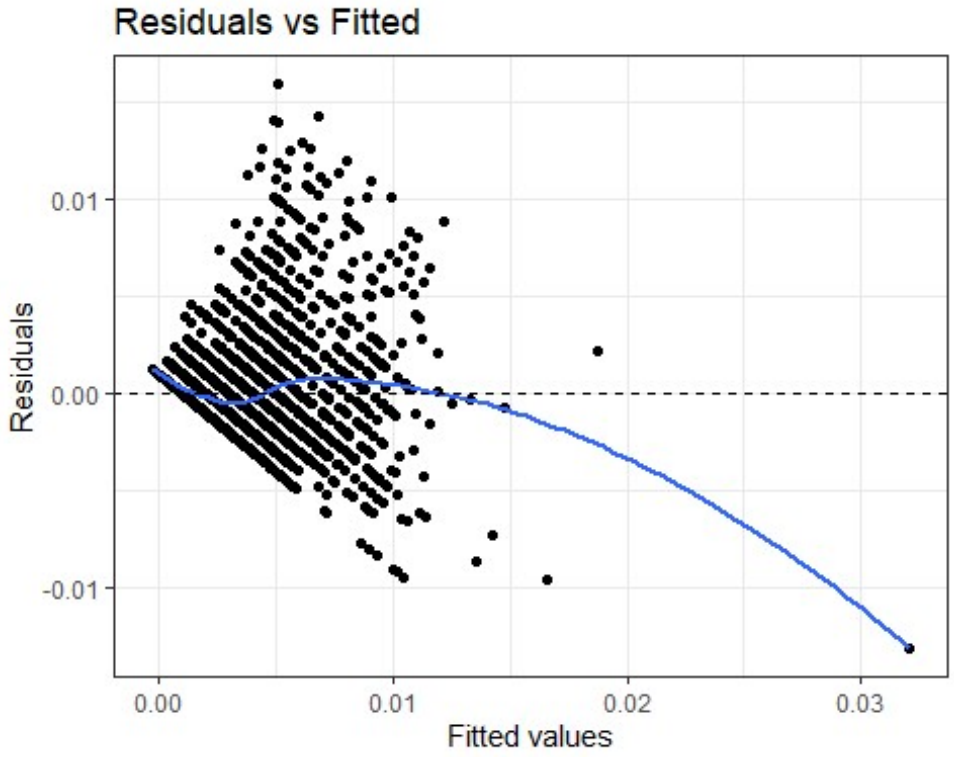
*3.2.4 Biểu đồ hồi quy*




**Nhận xét**: Đường thẳng hồi quy không quá khớp với dữ liệu.

### *3.2.5 Một số chuẩn đoán cho mô hình*

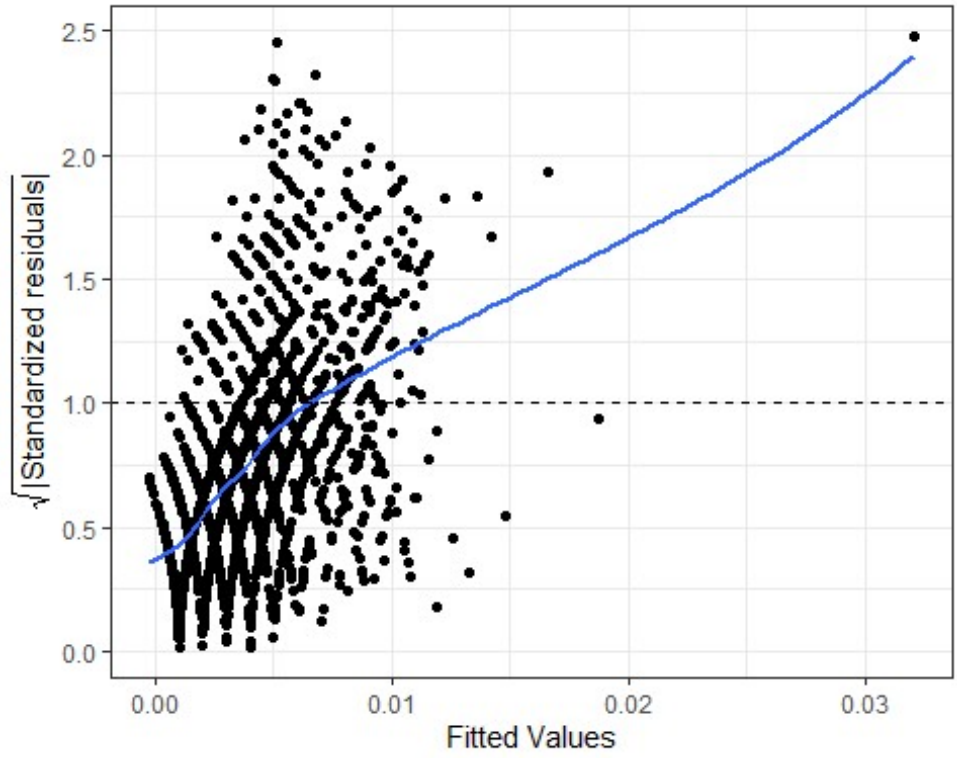
Kiểm định tuyến tính mô hình




**Nhận xét**: Hình vẽ cho thấy xu hướng đường cong

=> Giả định về tuyến tính của mô hình là chưa phù hợp

Kiểm đinh đồng nhất phương sai




**Nhận xét**: Đồ thị không xấp xỉ đường thẳng 1

=> Giả định về phương sai đồng nhất là chưa hợp lý

Kiểm tra đa cộng tuyến

```
vif(st_model_final)
```

```
##                       age                 potential                     value
##                  3.653687                  2.904254                  3.442385
## international_reputation                 body_type                  crossing
##                  1.139488                  1.070764                  2.446190
##             short_passing                   volleys               fk_accuracy
##                  2.588161                  2.187751                  1.826849
##              sprint_speed                long_shots             interceptions
##                  1.299768                  1.977285                  1.598253
##                 penalties                   marking             release_clause
##                  1.496617                  1.271898                  2.126271
```

**Nhận xét**: Mô hình không xảy ra đa cộng tuyến.

### *3.2.6 Đánh giá mô hình*

RMSE của mô hình

```
## [1] "RMSE từ tập kiểm tra:  0.00297938308297955"
```

*R2 Score*

```
## [1] "R2 Score của mô hình là: 0.453015728877717"
```

*Áp dụng k-fold validation để tính RMSE*

```
## [1] "RMSE từ K-Fold CV (k = 5 ): 0.00252403927699496"
```

### *3.2.7 Mở rộng mô hình*

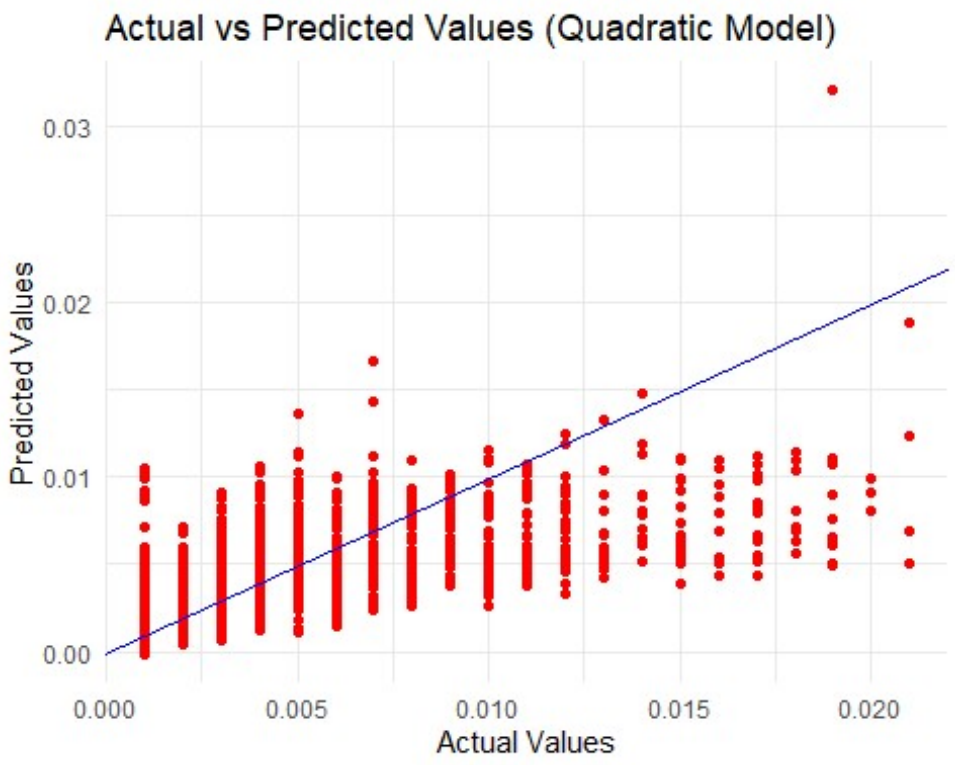
#### Khởi tạo mô hình

Ta sẽ tiến hành chạy mô hình bậc 2 với các biến có ít độ tuyến tính với mô hình

```
summary(st_model_final)
```

```
##
## Call:
## lm(formula = st_formula_final, data = df_st_numeric)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0131295 -0.0013897 -0.0003798  0.0007080  0.0159614
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -2.485e-03  1.014e-03  -2.450 0.014354 *
## age                        9.301e-05  2.129e-05   4.369  1.3e-05 ***
## value                      1.891e-03  1.762e-04  10.731  < 2e-16 ***
## international_reputation   1.047e-03  3.779e-04   2.771 0.005623 **
## crossing                   2.430e-05  6.574e-06   3.697 0.000223 ***
## short_passing             -2.723e-05  1.064e-05  -2.560 0.010539 *
## fk_accuracy               -1.614e-05  5.859e-06  -2.755 0.005905 **
## sprint_speed              -2.079e-05  5.742e-06  -3.620 0.000300 ***
## interceptions              1.943e-05  5.854e-06   3.320 0.000914 ***
## penalties                  1.806e-05  7.800e-06   2.316 0.020657 *
## marking                   -1.213e-05  5.411e-06  -2.242 0.025038 *
## release_clause             7.684e-04  5.098e-05  15.073  < 2e-16 ***
## I(potential^2)             1.651e-07  1.225e-07   1.348 0.177901
## I(body_type^2)            -4.572e-06  2.364e-06  -1.934 0.053217 .
## I(volleys^2)               1.632e-07  7.613e-08   2.143 0.032200 *
## I(long_shots^2)            2.216e-07  8.720e-08   2.541 0.011110 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002646 on 2506 degrees of freedom
## Multiple R-squared:  0.4538, Adjusted R-squared:  0.4506
## F-statistic: 138.8 on 15 and 2506 DF,  p-value: < 2.2e-16
```

Đường thẳng hồi quy




Đánh giá mô hình

```
## [1] "RMSE từ tập kiểm tra (Quadratic Model):  0.00313692065779201"

## [1] "R2 Score của mô hình bậc hai là: 0.453831605265703"
```

**Nhận xét**: Mô hình bậc 2 có kết quả chạy tốt hơn, với R2 Score cao hơn và RMSE thấp hơn.

## 3.3 Mô hình hồi quy dự đoán mức lương cho vị trí Tiền vệ

### *3.3.1 Lựa chọn biến phù hợp cho mô hình*

Dùng Stepwise để lựa chọn các biến phù hợp:

```
selected_vars <- names(coef(step_md))[-1]
```

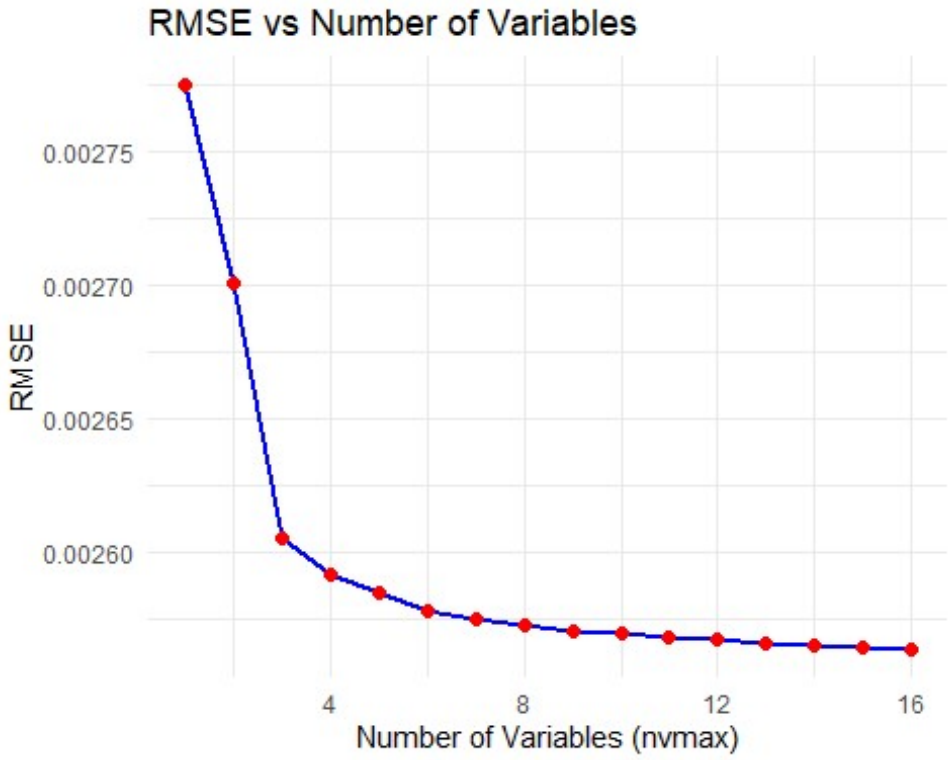
Các biến phù hợp để khởi tạo mô hình là:

```
print(selected_vars)
```

```
##  [1] "age"                      "value"
##  [3] "preferred_foot"           "international_reputation"
##  [5] "body_type"                "crossing"
##  [7] "heading_accuracy"         "short_passing"
##  [9] "sprint_speed"             "stamina"
## [11] "aggression"               "positioning"
```

```
## [13] "penalties"                "marking"
## [15] "sliding_tackle"           "release_clause"
```

### *3.3.2 Biểu đồ RMSE cho các biến đã chọn*




**Nhận xét**: Vậy, các biến trên được chọn để khởi tạo mô hình là phù hợp

### *3.3.3 Áp dụng mô hình cho các biến đã chọn*

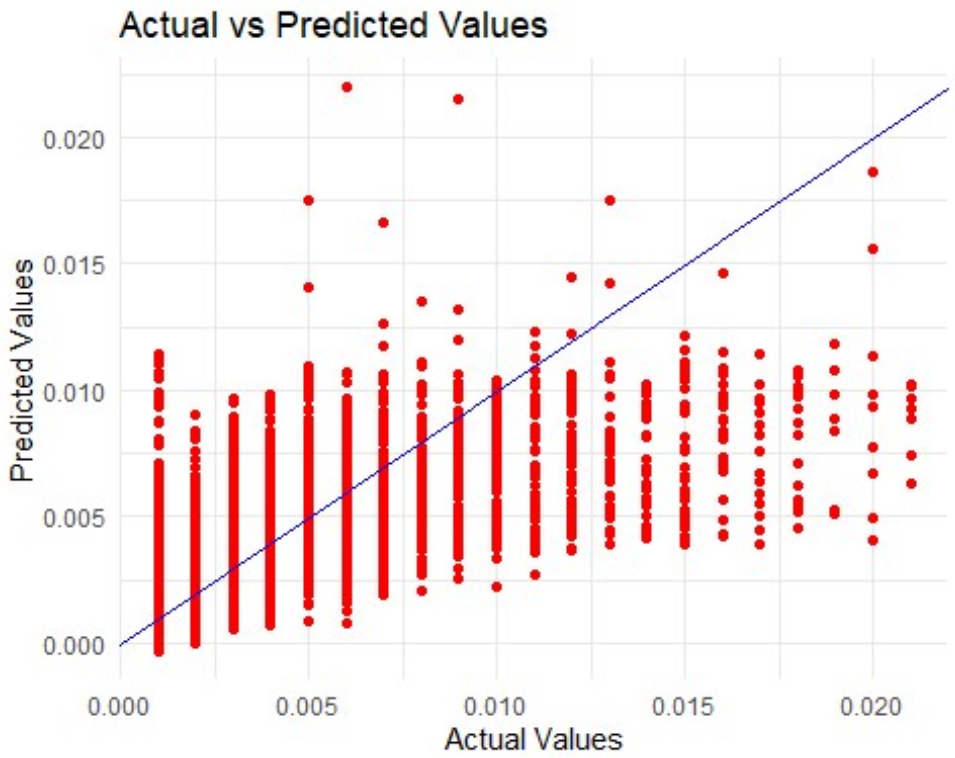
```
##
## Call:
## lm(formula = mid_formula_final, data = df_mid_numeric)
##
## Residuals:
##       Min         1Q     Median         3Q        Max
## -0.0159923 -0.0013575 -0.0003090  0.0007257  0.0159109
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -3.168e-03  7.101e-04  -4.461 8.32e-06 ***
## age                       1.399e-04  1.076e-05  12.997  < 2e-16 ***
## value                     2.132e-03  1.204e-04  17.714  < 2e-16 ***
## preferred_foot            1.659e-04  9.054e-05   1.833 0.066930 .
## international_reputation  1.498e-03  2.789e-04   5.371 8.17e-08 ***
## body_type                -6.651e-05  2.188e-05  -3.040 0.002375 **
## crossing                  1.580e-05  5.172e-06   3.055 0.002266 **
## heading_accuracy          1.682e-05  4.430e-06   3.795 0.000149 ***
## short_passing            -1.700e-05  8.629e-06  -1.970 0.048872 *
```

```
## sprint_speed              -9.153e-06  4.519e-06  -2.025 0.042899 *
## stamina                   -1.903e-05  3.877e-06  -4.907 9.54e-07 ***
## aggression                 1.016e-05  3.931e-06   2.585 0.009778 **
## positioning               -8.148e-06  5.326e-06  -1.530 0.126095
## penalties                  1.578e-05  4.720e-06   3.344 0.000832 ***
## marking                   -1.209e-05  4.029e-06  -3.001 0.002705 **
## sliding_tackle             9.188e-06  4.245e-06   2.164 0.030494 *
## release_clause             7.610e-04  3.997e-05  19.042  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002568 on 5003 degrees of freedom
## Multiple R-squared:  0.4529, Adjusted R-squared:  0.4511
## F-statistic: 258.8 on 16 and 5003 DF,  p-value: < 2.2e-16
```

**Nhận xét**:

- R2 Score: các biến giải thích không quá tốt cho biến mục tiêu
- Một số biến ảnh hưởng mạnh đến wage của vị trí Tiền vệ: age, value, inter_repu, heading_accuracy, stamina.
- Một số biến không ảnh hưởng nhiều đến mô hình: standing_tackle.

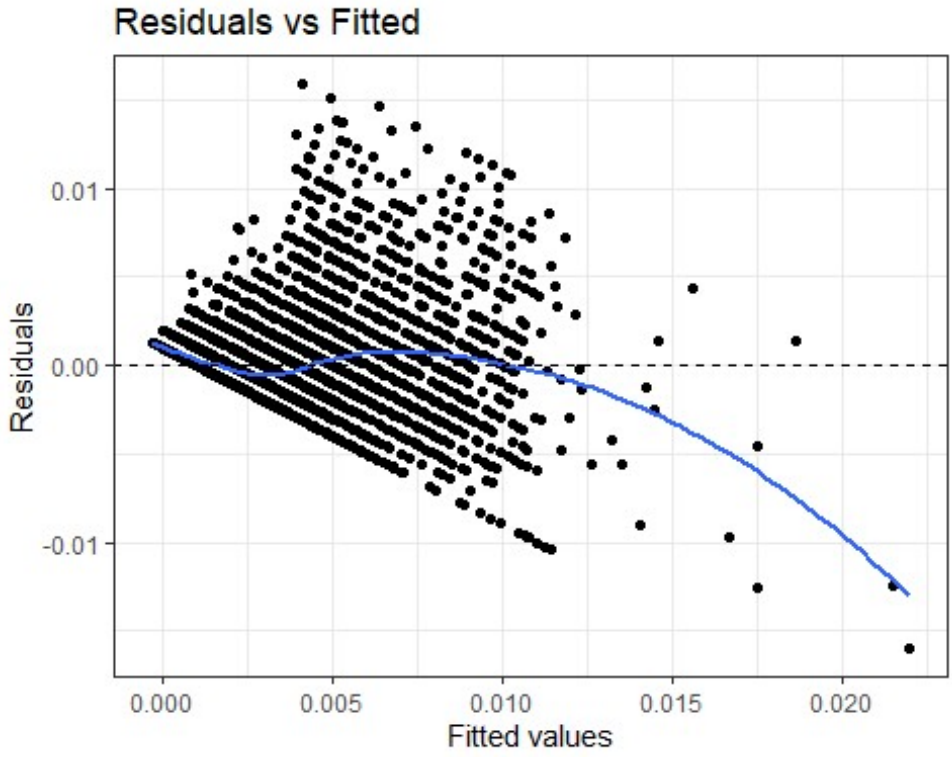
### *3.3.4 Biểu đồ hồi quy*




**Nhận xét**: Đường thẳng hồi quy không quá khớp với dữ liệu.

### *3.3.5 Một số chuẩn đoán cho mô hình*

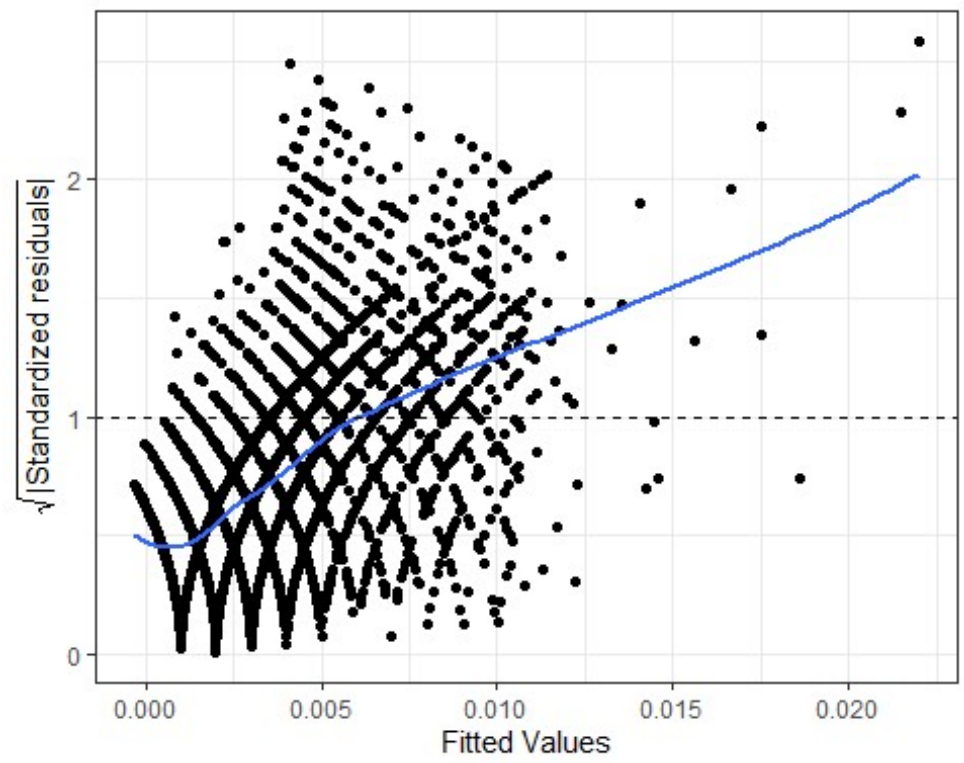
Kiểm định tuyến tính mô hình




**Nhận xét**: Hình vẽ cho thấy xu hướng đường cong

=> Giả định về tuyến tính của mô hình là chưa phù hợp

Kiểm đinh đồng nhất phương sai




**Nhận xét**: Đồ thị không xấp xỉ đường thẳng 1

=> Giả định về phương sai đồng nhất là chưa hợp lý

Kiểm tra đa cộng tuyến

```
vif(mid_model_final)
```

```
##                      age                    value           preferred_foot
##                 1.942179                 3.347891                 1.056237
## international_reputation                body_type                 crossing
##                 1.113094                 1.080256                 1.948800
##         heading_accuracy           short_passing             sprint_speed
##                 1.491797                 2.188567                 1.629152
##                  stamina              aggression              positioning
##                 1.613723                 2.309149                 1.872825
##                penalties                  marking           sliding_tackle
##                 1.595357                 2.486109                 3.057128
##           release_clause
##                 2.358634
```

**Nhận xét**: Mô hình không xảy ra đa cộng tuyến.

*3.3.6 Đánh giá mô hình*

RMSE của mô hình

```
## [1] "RMSE từ tập kiểm tra:  0.00242488609317167"
```

R2 Score

```
## [1] "R2 Score của mô hình là: 0.452865354010294"
```

Áp dụng k-fold validation để tính RMSE

```
## [1] "RMSE từ K-Fold CV (k = 5 ): 0.00245414783576511"
```

### *3.3.7 Mở rộng mô hình*

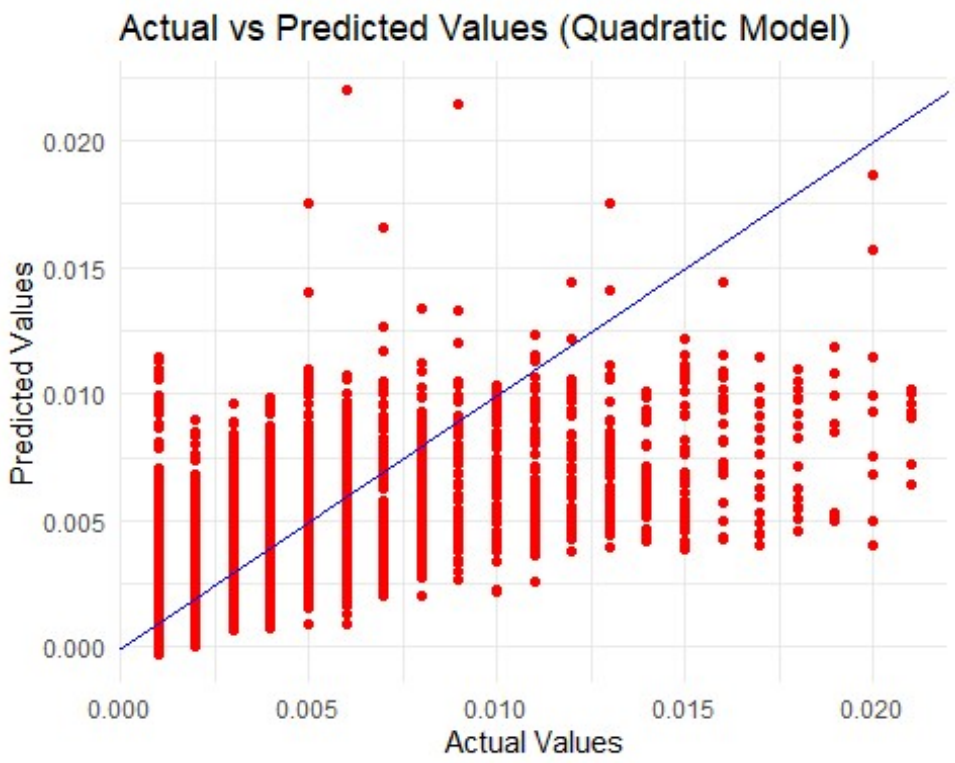
Khởi tạo mô hình

Ta sẽ tiến hành chạy mô hình bậc 2 với các biến có ít độ tuyến tính với mô hình

```
summary(mid_model_final)
```

```
##
## Call:
## lm(formula = mid_formula_final, data = df_mid_numeric)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0160041 -0.0013496 -0.0003071  0.0007204  0.0159465
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -3.254e-03  7.214e-04  -4.511 6.61e-06 ***
## age                       1.388e-04  1.083e-05  12.810  < 2e-16 ***
## value                     2.129e-03  1.210e-04  17.598  < 2e-16 ***
## preferred_foot            1.649e-04  9.058e-05   1.821 0.068687 .
## international_reputation  1.483e-03  2.791e-04   5.313 1.12e-07 ***
## body_type                -6.694e-05  2.189e-05  -3.059 0.002235 **
## crossing                  1.554e-05  5.159e-06   3.013 0.002600 **
## heading_accuracy          1.552e-05  4.476e-06   3.467 0.000530 ***
## short_passing            -2.065e-05  8.543e-06  -2.418 0.015653 *
## sprint_speed             -8.113e-06  4.521e-06  -1.795 0.072794 .
## stamina                  -2.072e-05  3.873e-06  -5.350 9.19e-08 ***
## aggression                8.190e-06  3.914e-06   2.092 0.036458 *
## penalties                 1.657e-05  4.712e-06   3.516 0.000442 ***
## release_clause            7.614e-04  3.998e-05  19.044  < 2e-16 ***
## I(positioning^2)         -4.781e-08  5.183e-08  -0.922 0.356367
## I(marking^2)             -6.914e-08  4.498e-08  -1.537 0.124381
## I(sliding_tackle^2)       1.139e-07  4.962e-08   2.294 0.021805 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.00257 on 5003 degrees of freedom
## Multiple R-squared:  0.4523, Adjusted R-squared:  0.4506
## F-statistic: 258.2 on 16 and 5003 DF,  p-value: < 2.2e-16
```

Đường thẳng hồi quy




Đánh giá mô hình

```
## [1] "RMSE từ tập kiểm tra (Quadratic Model):  0.00238227273610224"

## [1] "R2 Score của mô hình bậc hai là: 0.452322713961506"
```

**Nhận xét**: Mô hình bậc 2 có kết quả chạy tương đương.

## 3.4 Mô hình hồi quy dự đoán mức lương cho vị trí Hậu vệ

### *3.4.1 Lựa chọn biến phù hợp cho mô hình*

```
df_defender_numeric <- df_df_numeric |> select_if(is.numeric)
```

Dùng Stepwise để lựa chọn các biến phù hợp:

```
selected_vars <- names(coef(step_md))[-1]
```

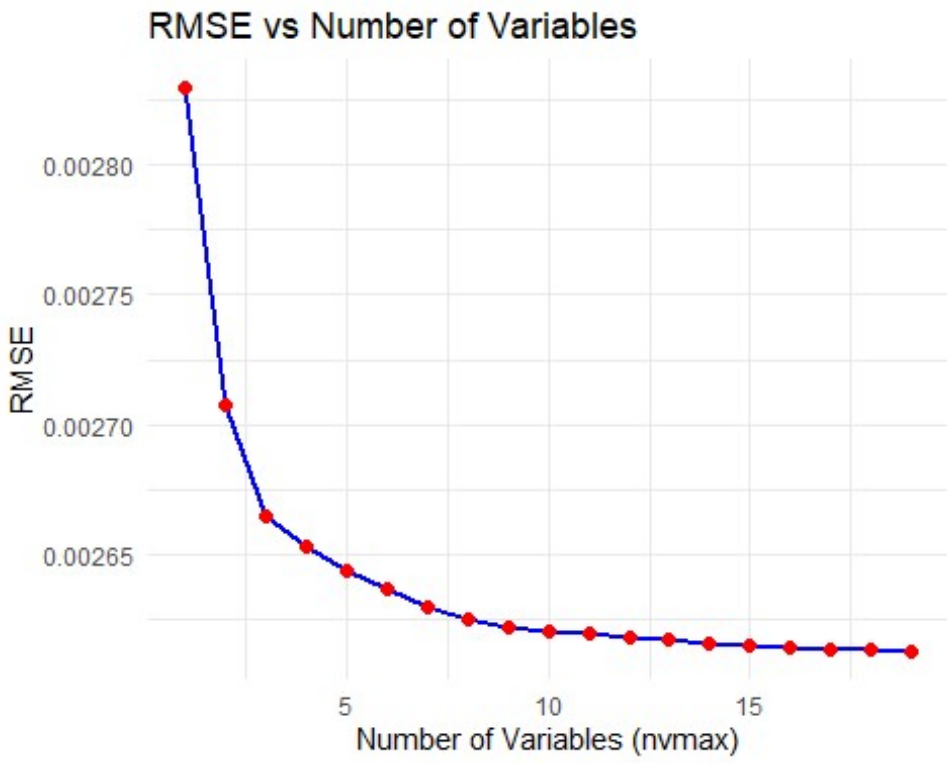
Các biến phù hợp để khởi tạo mô hình là:

```
print(selected_vars)

##  [1] "age"                      "potential"
##  [3] "value"                    "international_reputation"
##  [5] "skill_moves"              "body_type"
##  [7] "crossing"                 "heading_accuracy"
```

```
##  [9] "curve"                    "balance"
## [11] "shot_power"               "jumping"
## [13] "stamina"                  "strength"
## [15] "long_shots"               "interceptions"
## [17] "penalties"                "sliding_tackle"
## [19] "release_clause"
```

### *3.4.2 Biểu đồ RMSE cho các biến đã chọn*




**Nhận xét**: Vậy, các biến trên được chọn để khởi tạo mô hình là phù hợp

### *3.4.3 Áp dụng mô hình cho các biến đã chọn*

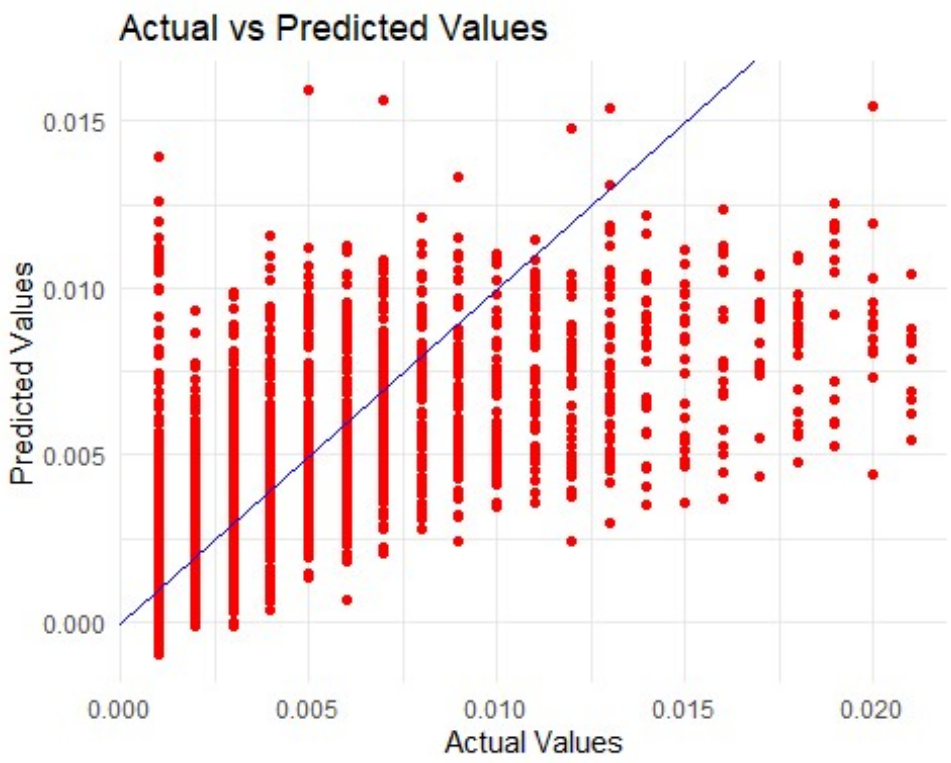
```
##
## Call:
## lm(formula = defender_formula_final, data = df_defender_numeric)
##
## Residuals:
##       Min         1Q     Median         3Q        Max
## -0.0129055 -0.0013748 -0.0003532  0.0008394  0.0155615
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -7.676e-03  1.147e-03  -6.694 2.43e-11 ***
## age                       1.690e-04  1.606e-05  10.520  < 2e-16 ***
## potential                 2.625e-05  1.442e-05   1.819 0.068902 .
## value                     1.647e-03  1.734e-04   9.500  < 2e-16 ***
## international_reputation  1.232e-03  2.737e-04   4.501 6.94e-06 ***
```

```
## skill_moves                4.830e-04  1.305e-04   3.702 0.000216 ***
## body_type                 -4.192e-05  2.406e-05  -1.742 0.081610 .
## crossing                   7.792e-06  4.259e-06   1.830 0.067363 .
## heading_accuracy           1.576e-05  6.657e-06   2.368 0.017929 *
## curve                      8.116e-06  4.652e-06   1.745 0.081133 .
## balance                   -7.879e-06  4.898e-06  -1.609 0.107777
## shot_power                -1.210e-05  4.681e-06  -2.584 0.009786 **
## jumping                    7.456e-06  4.100e-06   1.819 0.069033 .
## stamina                   -2.369e-05  4.646e-06  -5.099 3.56e-07 ***
## strength                  -8.175e-06  5.761e-06  -1.419 0.155973
## long_shots                 1.123e-05  4.949e-06   2.270 0.023274 *
## interceptions             -2.389e-05  9.338e-06  -2.559 0.010544 *
## penalties                  1.321e-05  4.977e-06   2.654 0.007989 **
## sliding_tackle             4.079e-05  1.000e-05   4.079 4.60e-05 ***
## release_clause             9.850e-04  7.436e-05  13.246  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002618 on 4569 degrees of freedom
## Multiple R-squared:  0.4537, Adjusted R-squared:  0.4515
## F-statistic: 199.7 on 19 and 4569 DF,  p-value: < 2.2e-16
```

**Nhận xét**:

- R2 Score: các biến giải thích không quá tốt cho biến mục tiêu.
- Một số biến ảnh hưởng mạnh đến wage của vị trí Hậu vệ: age, value, inter_repu, skill_moves, stamina, sliding_tackle.
- Một số biến không ảnh hưởng nhiều đến mô hình: curve, jumping.

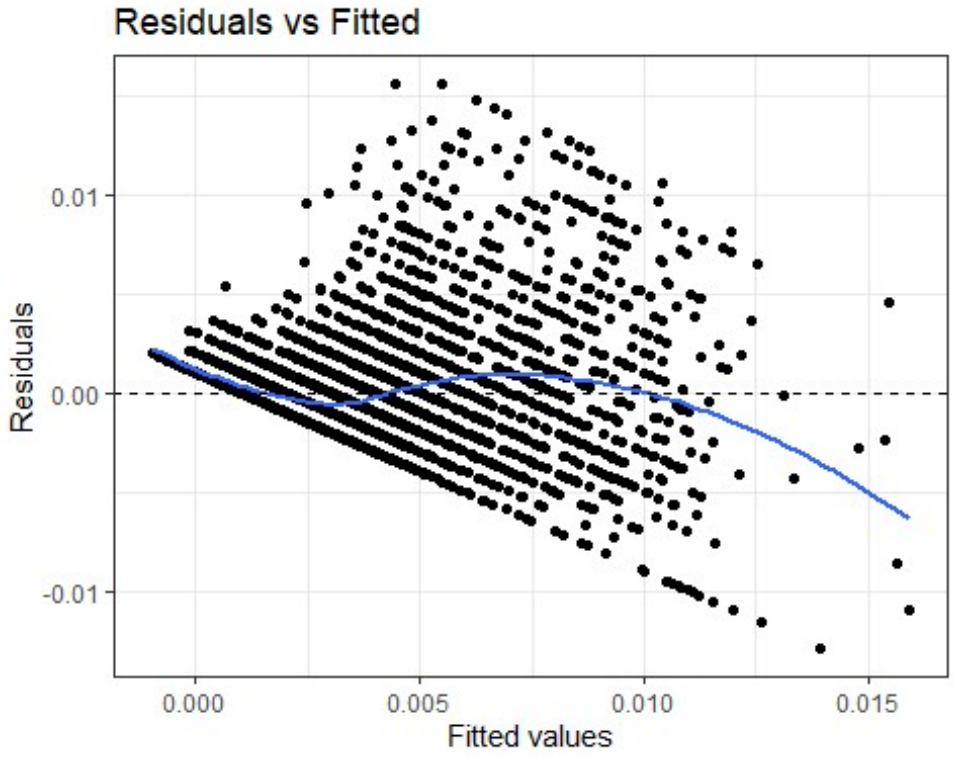
*3.4.4 Biểu đồ hồi quy*




**Nhận xét**: Đường thẳng hồi quy không quá khớp với dữ liệu.

### *3.4.5 Một số chuẩn đoán cho mô hình*

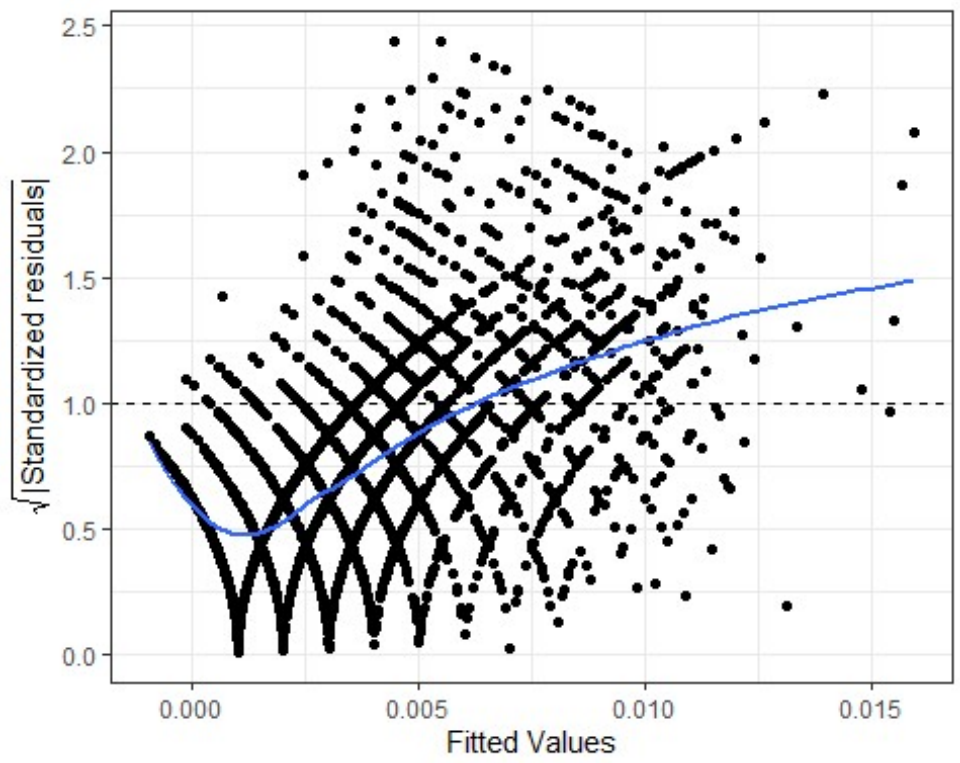
Kiểm định tuyến tính mô hình




**Nhận xét**: Hình vẽ cho thấy xu hướng đường cong

=> Giả định về tuyến tính của mô hình là chưa phù hợp

### Kiểm đinh đồng nhất phương sai




**Nhận xét**: Đồ thị dường như không xấp xỉ đường thẳng 1

=> Giả định về phương sai đồng nhất là chưa hợp lý

### Kiểm tra đa cộng tuyến

```
vif(defender_model_final)
```

```
##                      age                potential                    value
##                 3.752469                 3.104618                 5.594201
## international_reputation             skill_moves                body_type
##                 1.135212                 1.468206                 1.118014
##                 crossing         heading_accuracy                    curve
##                 2.838829                 2.494322                 2.886393
##                  balance              shot_power                  jumping
##                 2.424752                 2.625764                 1.230981
##                  stamina                 strength              long_shots
##                 1.500509                 2.506151                 3.150186
##            interceptions                penalties           sliding_tackle
##                 2.584046                 1.574699                 2.290784
##           release_clause
##                 4.390541
```

**Nhận xét**: Mô hình không xảy ra đa cộng tuyến.

### *3.4.6 Đánh giá mô hình*

#### RMSE của mô hình

```
## [1] "RMSE từ tập kiểm tra:  0.00251712324370953"
```

#### *R2 Score*

```
## [1] "R^2 từ tập kiểm tra:  0.466845548987242"
```

#### *Áp dụng k-fold validation để tính RMSE*

```
## [1] "RMSE từ K-Fold CV (k = 5 ): 0.00256393544868218"
```

### *3.4.7 Mở rộng mô hình*

#### Khởi tạo mô hình

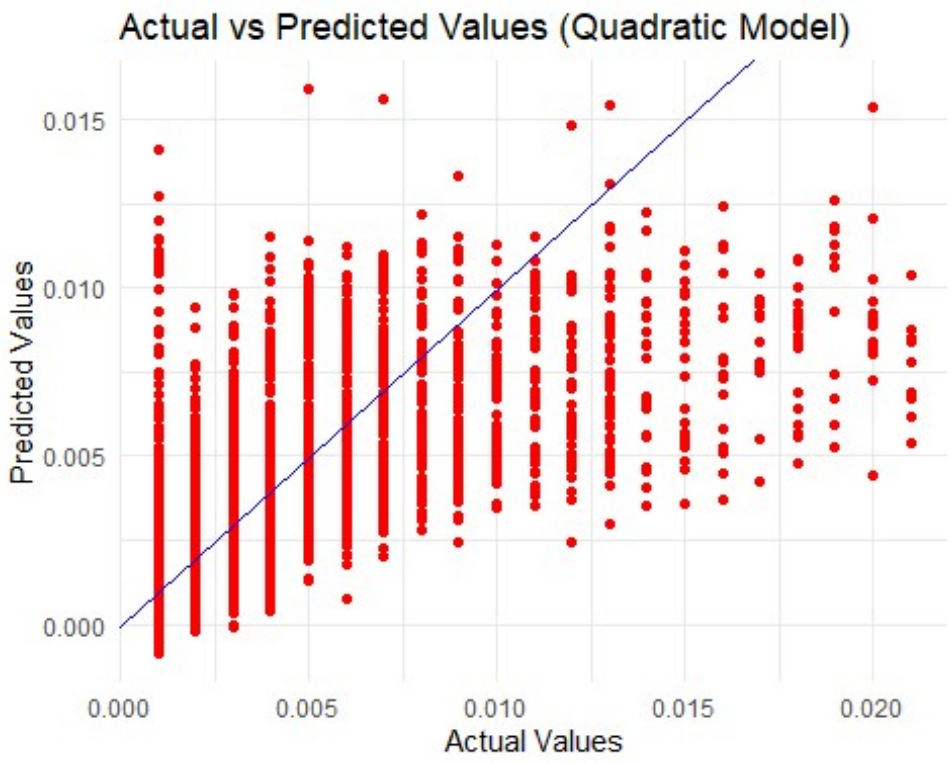
Ta sẽ tiến hành chạy mô hình bậc 2 với các biến có ít độ tuyến tính với mô hình

```
# In kết quả của mô hình
summary(defender_model_final)

##
## Call:
## lm(formula = defender_formula_final, data = df_defender_numeric)
##
## Residuals:
##        Min          1Q      Median          3Q         Max
## -0.0131130 -0.0013765 -0.0003597  0.0008242  0.0156033
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -7.393e-03  1.118e-03  -6.612 4.22e-11 ***
## age                        1.671e-04  1.624e-05  10.288  < 2e-16 ***
## potential                  2.677e-05  1.444e-05   1.854  0.06374 .
## value                      1.644e-03  1.738e-04   9.458  < 2e-16 ***
## international_reputation   1.209e-03  2.737e-04   4.415 1.03e-05 ***
## skill_moves                4.363e-04  1.326e-04   3.289  0.00101 **
## heading_accuracy           1.608e-05  6.697e-06   2.401  0.01639 *
## shot_power                -1.313e-05  4.622e-06  -2.840  0.00452 **
## stamina                   -2.398e-05  4.628e-06  -5.181 2.30e-07 ***
## interceptions             -2.391e-05  9.336e-06  -2.560  0.01049 *
## penalties                  1.281e-05  4.968e-06   2.578  0.00996 **
## sliding_tackle             4.048e-05  1.001e-05   4.043 5.36e-05 ***
## release_clause             9.852e-04  7.436e-05  13.248  < 2e-16 ***
## I(body_type^2)            -3.125e-06  1.907e-06  -1.639  0.10128
## I(crossing^2)              9.031e-08  4.521e-08   1.998  0.04583 *
## I(curve^2)                 8.276e-08  5.071e-08   1.632  0.10271
## I(balance^2)              -5.859e-08  4.111e-08  -1.425  0.15419
## I(jumping^2)               5.365e-08  3.132e-08   1.713  0.08674 .
## I(strength^2)             -5.526e-08  4.357e-08  -1.268  0.20481
## I(long_shots^2)            1.612e-07  5.823e-08   2.769  0.00564 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002618 on 4569 degrees of freedom
## Multiple R-squared:  0.454,  Adjusted R-squared:  0.4517
## F-statistic:   200 on 19 and 4569 DF,  p-value: < 2.2e-16
```

Đường thẳng hồi quy




Đánh giá mô hình

```
## [1] "RMSE từ tập kiểm tra (Quadratic Model):  0.0024801183382002"

## [1] "R2 Score từ tập kiểm tra (Quadratic Model):  0.482406413587549"
```

**Nhận xét**: Mô hình bậc 2 có kết quả chạy tốt hơn, với R2 Score cao hơn và RMSE thấp hơn.

# VI. Tổng kết:

## 1.Các kiểm định:

- Cầu thủ thuận chân trái sẽ có mức lương tốt hơn cầu thủ thuận chân phải.

- Cầu thủ thuận chân trái sẽ điều khiển bóng và có lực sút tốt hơn.

- Mức lương của cầu thủ khác nhau tùy theo vị trí của cầu thủ trên sân.

- Chỉ số tổng thể có bị ảnh hưởng bởi từng độ tuổi khác nhau.

## 2. Các mô hình tuyến tính:

### 2.1 Mô hình dự đoán thông số Overall:

- Mô hình dự đoán khá chính xác với R2 Score ~ 0.7, RMSE ~ 2.5.

- Nếu dựa vào các thông số chuyên môn, mỗi khi đánh giá Overall một cầu thủ thì các thông số reactions (Phản ứng), composure (Điềm tĩnh), short_passing (Chuyển ngắn) được đánh giá cao.

### 2.2 Mô hình dự đoán mức lương vị trí Thủ môn:

- Mô hình dự đoán tương đối R2 Score ~ 0.5 và RMSE ~ 0.0022.

- Với vị trí thủ môn, các thông số về gk_handling, prefered_foot (Chân thuận), height (Chiều cao) sẽ quyết định phần lớn mức lương.

### 2.3 Mô hình dự đoán mức lương vị trí Hậu vệ:

- Mô hình dự đoán tương đối R2 Score ~ 0.48 và RMSE ~ 0.0023.

- Với vị trí thủ môn, các thông số về stamina (Thể lực) value (Giá trị), sliding_tackle (Xoạc banh) sẽ quyết định phần lớn mức lương.

### 2.4 Mô hình dự đoán mức lương vị trí Tiền vệ:

- Mô hình dự đoán tương đối với R2 Score ~ 0.47 và RSME ~ 0.0024.

- Với vị trí tiền vệ, các thông số về stamina (Sức bền), marking (Kèm người) sẽ quyết định phần lớn mức lương.

### 2.5 Mô hình dự đoán mức lương vị trí Tiền đạo:

- Mô hình dự đoán tương đối với R2 Score ~ 0.47 và RSME ~ 0.0025.

- Với vị trí tiền đạo, các thông số về sprint_speed (Tốc độ tối đa), crossing (Tạt cánh), age (Độ tuổi) sẽ quyết định phần lớn mức lương.

Ngoài ra, các thông số như age (Độ tuổi), international_reputation (Danh tiếng quốc tế), release_clause (Chi phí giải phóng hợp đồng) luôn ảnh hưởng mạnh đến lương của cầu thủ

## 3. Các giải pháp thực tế:

- Dùng mô hình dự đoán mức lương để dự đoán mức lương phù hợp cho 1 cầu thủ mới, từ đó có thể tối ưu chi phí cho câu lạc bộ.
- Với từng vị trí khác nhau, nên cân nhắc lựa chọn các cầu thủ có khả năng phù hợp với vị trí đó.
- Xây dựng và bồi dưỡng các cầu thủ trẻ về mặt danh tiếng trên thị trường quốc tế, từ đó có thể mang lại thu nhập khi cho thuê hoặc giao dịch cầu thủ.