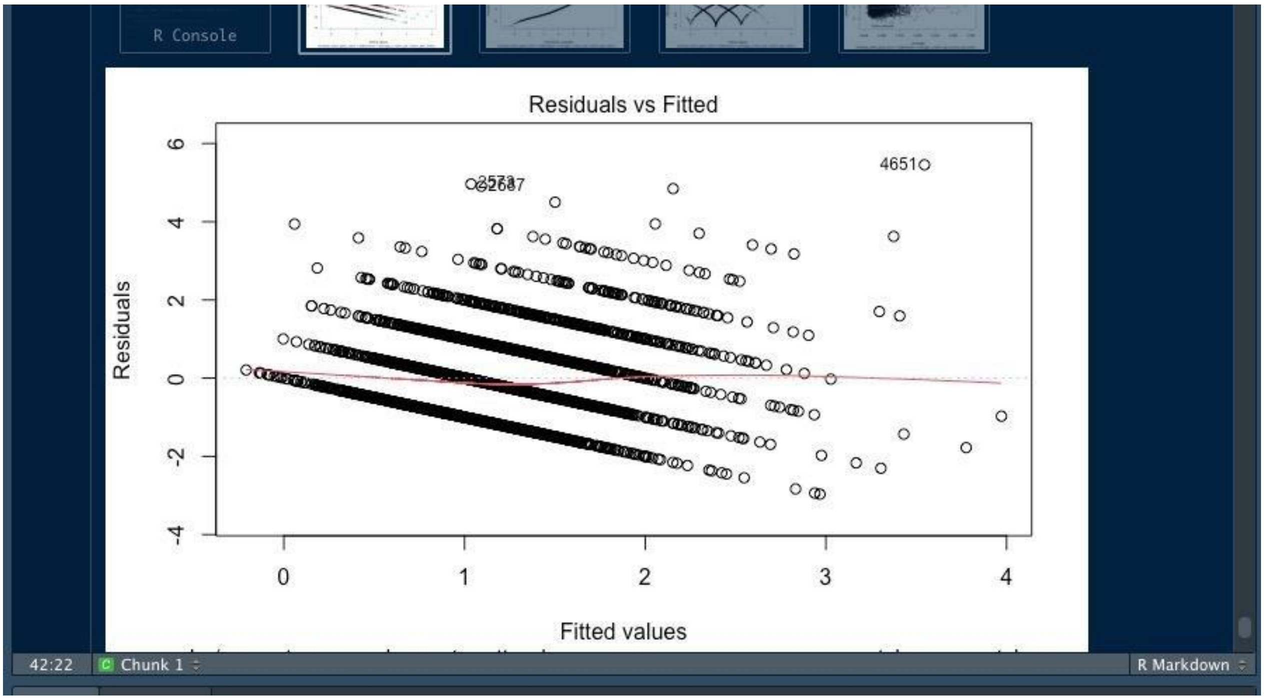
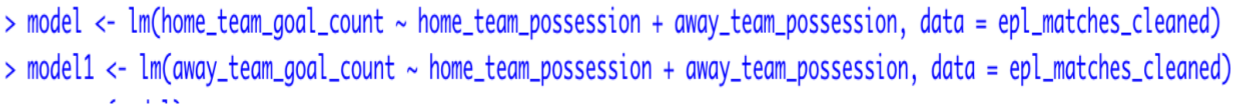
Team Hyperborea : Sports | Individual Milestone 2

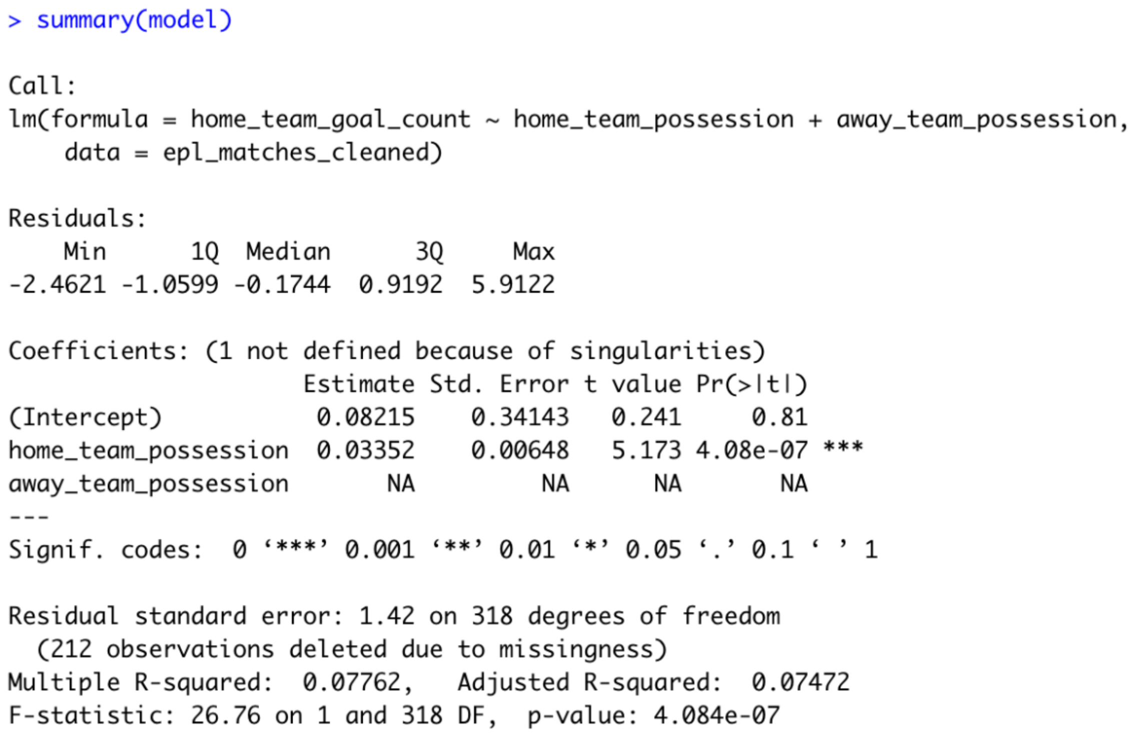
Annapoorani Sundararaj Shanthi

Our team decided to move forward focusing on a goal differential variable home_goal_count and away_goal_count as our dependent variable.

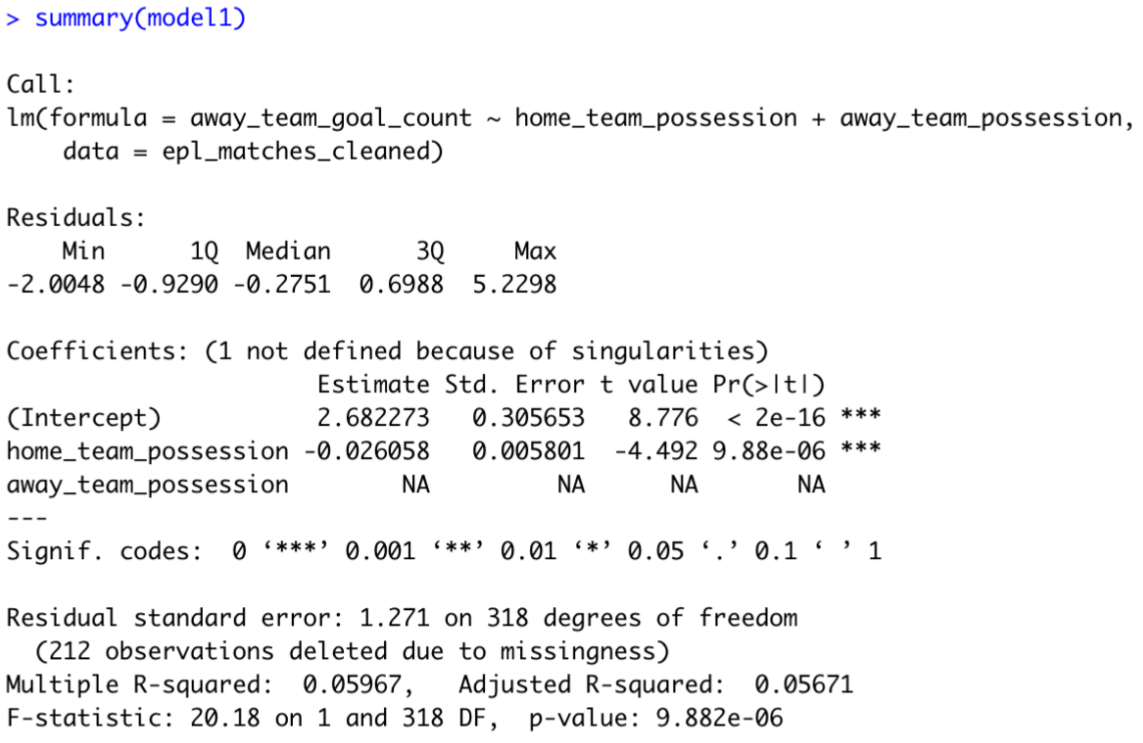




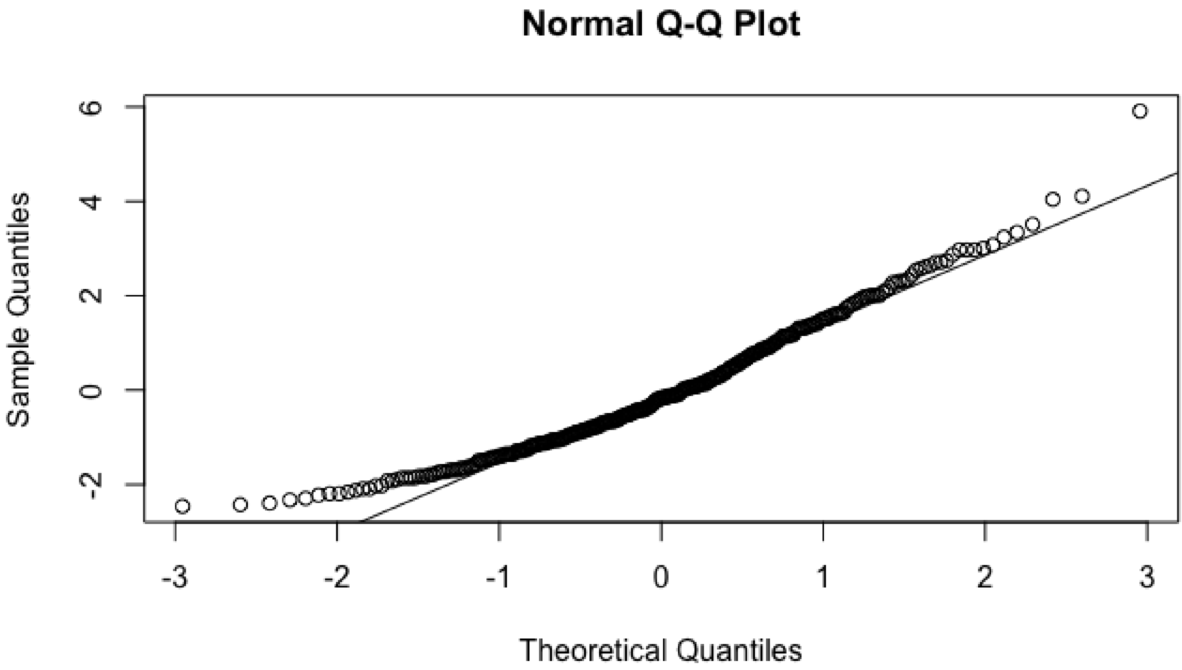
```
> model <- lm(home_team_goal_count ~ home_team_possession + away_team_possession, data = epl_matches_cleaned)
> model1 <- lm(away_team_goal_count ~ home_team_possession + away_team_possession, data = epl_matches_cleaned)
```



```
> summary(model)

Call:
lm(formula = home_team_goal_count ~ home_team_possession + away_team_possession, 
    data = epl_matches_cleaned)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.4621 -1.0599 -0.1744  0.9192  5.9122 

Coefficients: (1 not defined because of singularities)
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           0.08215    0.34143   0.241     0.81    
home_team_possession  0.03352    0.00648   5.173 4.08e-07 ***
away_team_possession       NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.42 on 318 degrees of freedom
  (212 observations deleted due to missingness)
Multiple R-squared:  0.07762,	Adjusted R-squared:  0.07472 
F-statistic: 26.76 on 1 and 318 DF,  p-value: 4.084e-07
```

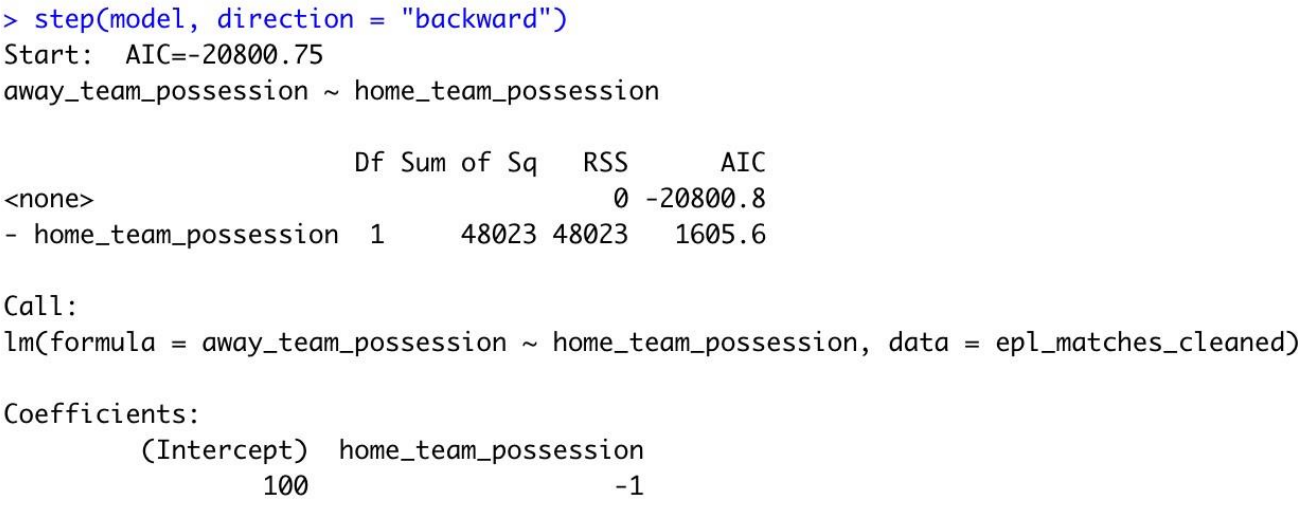


```
> summary(model1)

Call:
lm(formula = away_team_goal_count ~ home_team_possession + away_team_possession, 
    data = epl_matches_cleaned)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.0048 -0.9290 -0.2751  0.6988  5.2298 

Coefficients: (1 not defined because of singularities)
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           2.682273   0.305653   8.776  < 2e-16 ***
home_team_possession -0.026058   0.005801  -4.492 9.88e-06 ***
away_team_possession        NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.271 on 318 degrees of freedom
  (212 observations deleted due to missingness)
Multiple R-squared:  0.05967,	Adjusted R-squared:  0.05671 
F-statistic: 20.18 on 1 and 318 DF,  p-value: 9.882e-06
```

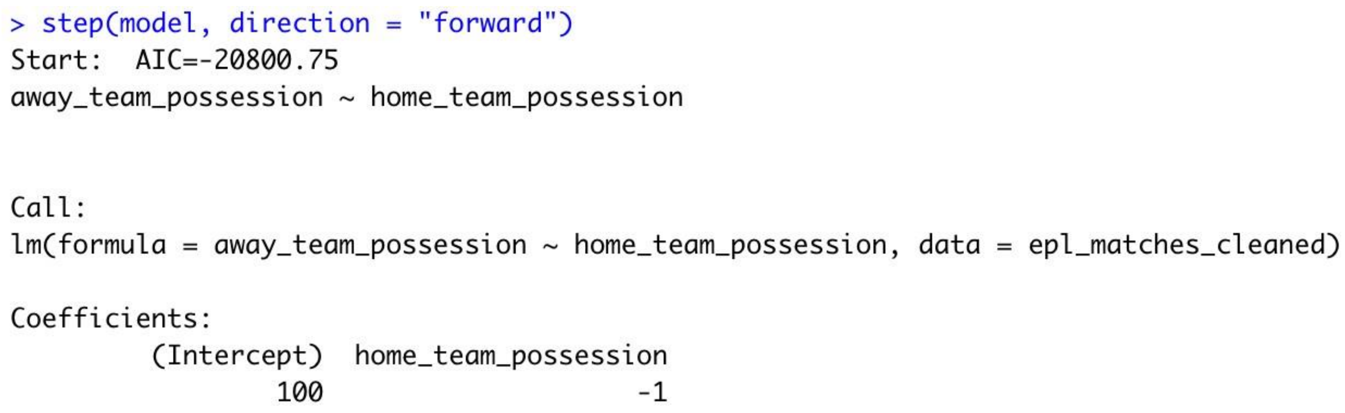


This a normal QQ plot where the dependent variables are home_team_goal_count and away_team_goal_count and the independent variables home_team_possession and away_team_possession.

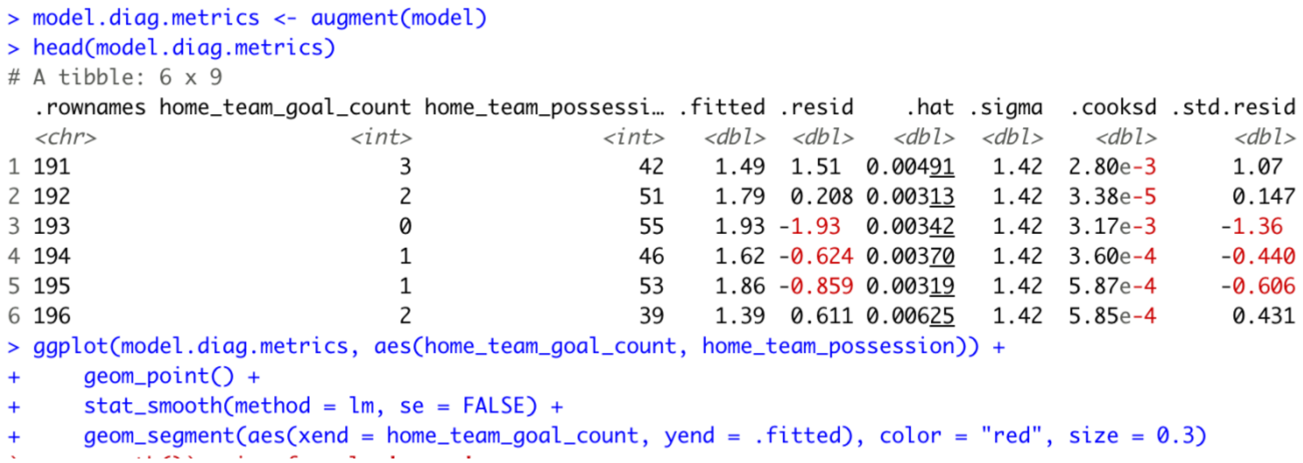
I've done forward and backward selection,



```
> step(model, direction = "backward")
Start:  AIC=-20800.75
away_team_possession ~ home_team_possession

                        Df Sum of Sq   RSS      AIC
<none>                                   0 -20800.8
- home_team_possession  1     48023 48023   1605.6

Call:
lm(formula = away_team_possession ~ home_team_possession, data = epl_matches_cleaned)

Coefficients:
         (Intercept)  home_team_possession  
                 100                    -1  
```

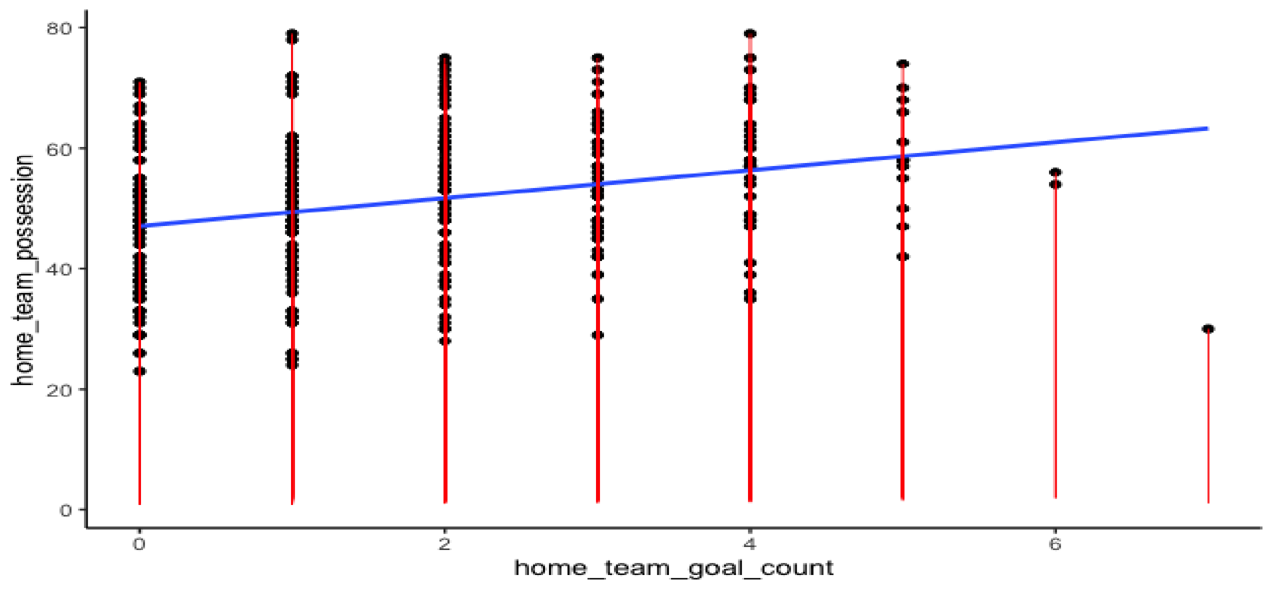


```
> step(model, direction = "forward")
Start:  AIC=-20800.75
away_team_possession ~ home_team_possession


Call:
lm(formula = away_team_possession ~ home_team_possession, data = epl_matches_cleaned)

Coefficients:
         (Intercept)  home_team_possession  
                 100                    -1  
```



```
> model.diag.metrics <- augment(model)
> head(model.diag.metrics)
# A tibble: 6 x 9
  .rownames home_team_goal_count home_team_possessi… .fitted .resid    .hat .sigma  .cooksd .std.resid
  <chr>                    <int>               <int>   <dbl>  <dbl>   <dbl>  <dbl>    <dbl>      <dbl>
1 191                          3                  42    1.49  1.51  0.00491   1.42 2.80e-3       1.07 
2 192                          2                  51    1.79  0.208 0.00313   1.42 3.38e-5       0.147
3 193                          0                  55    1.93 -1.93  0.00342   1.42 3.17e-3      -1.36 
4 194                          1                  46    1.62 -0.624 0.00370   1.42 3.60e-4      -0.440
5 195                          1                  53    1.86 -0.859 0.00319   1.42 5.87e-4      -0.606
6 196                          2                  39    1.39  0.611 0.00625   1.42 5.85e-4       0.431
> ggplot(model.diag.metrics, aes(home_team_goal_count, home_team_possession)) +
+     geom_point() +
+     stat_smooth(method = lm, se = FALSE) +
+     geom_segment(aes(xend = home_team_goal_count, yend = .fitted), color = "red", size = 0.3)
```

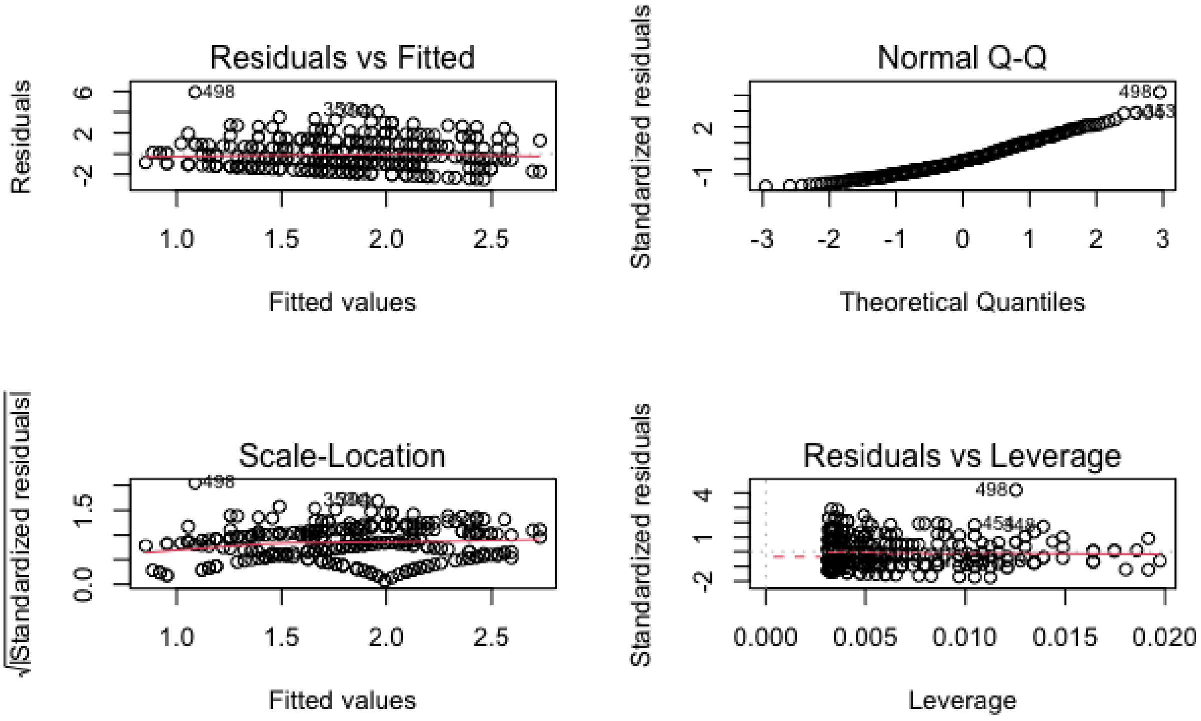
The fitted (or predicted) values are the y-values that you would expect for the given x-values according to the built regression mode.



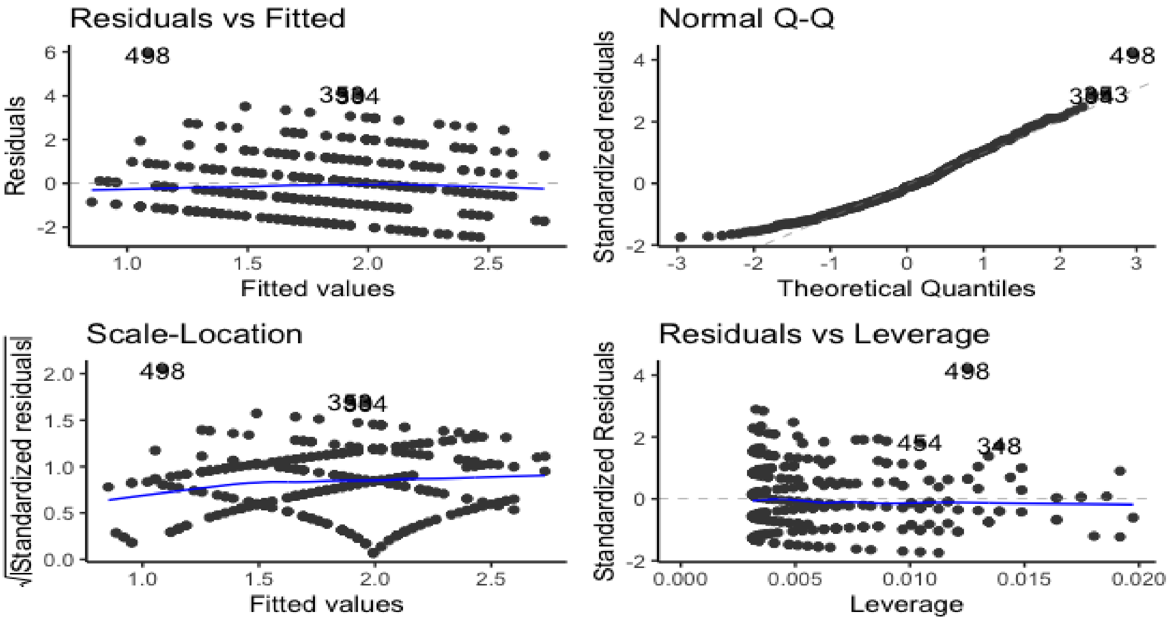
Regression diagnostics plots can be created using the R base function plot() or the autoplot() function [ggfortify package], which creates a ggplot2-based graphics.

```
> par(mfrow = c(2, 2))
> plot(model)
```

diagnostic plots with the R base function,



diagnostic plots using ggfortify:


Residuals vs Fitted
Residuals
Fitted values
498
Normal Q-Q
Standardized residuals
Theoretical Quantiles
498
Scale-Location
√|Standardized residuals|
Fitted values
498
Residuals vs Leverage
Standardized Residuals
Leverage
498
454
348