Final Project

Brady Heinig, Hunter Sanders, Eli Ladle, Dallin Draper

Abstract

This report uses Division 1 college basketball statistics from 2009-2021 to predict an NBA rookie's season performance in terms of Wins Above Replacement (WAR) using multiple linear regression. Although we used variable selection methods to choose relevant predictors, the overall predictive power of our model remained low with an adjusted R^2 of 0.0798. Moreover, our model failed to fully meet the assumptions of normality, impacting its reliability. Additionally, our analysis revealed that playing in a high major NCAA conference did not directly impact a rookie's wins above replacement. These findings show how difficult it is to predict a rookie's success using statistical methods alone.

1 Problem and Motivation

As a scout or general manager in the NBA, your job security is determined by how well you sniff out the prospects who end up being successful in the league. Drafting well is an art that few organizations have mastered, and there is no magic method that teams use to evaluate the potential of college and international prospects. With our project, we aim to assist these front offices in their drafting efforts by identifying key statistics and characteristics that can explain how valuable a player will be to their team during their rookie season. The metric we chose to harness in order to evaluate the player's first season is WAR (wins above replacement), because it takes into account both offensive and defensive metrics to provide a holistic view of a player's overall performance while scaling based on minutes played.

1.1 Data Description

This analysis uses data found on Kaggle called College Basketball 2009-2021 + NBA Advanced Stats, which was scraped and cleaned by Aditya Kumar. This dataset includes statistics for every Division 1 college basketball player from 2009-2021, as well as WAR values for all NBA players from the same time period.

The variables that were deemed relevant to this analysis are listed below:

- *war_total*: Total wins above replacement for that player for their entire NBA rookie season.
- *mp*: Minutes played by that player per game during their final college season
- *treb*: Average total rebounds collected by the player per game during their final college season
- *stl*: Average total steals made by the player per game during their final college season
- *ast*: Average total assists made by the player per game during their final college season
- *usg*: Season average Usage percentage of player during their final college season
- *bpm*: Average box plus-minus per game during player's final college season

- *drtg*: Season average defensive rating for each player
- *conf*: 1 if player played in a high major conference, 0 if they did not

1.2 Questions of Interest

The questions that we will look to answer in this project stem from the intended use of this data to evaluate possible NBA prospects. First, we want to discover the collegiate statistics that are most predictive of an NBA rookie's total Wins Above Replacement (WAR). Secondly, we will investigate if a player's participation in a major conference while in college determines their success during their NBA rookie season.

1.3 Regression Methods

The first question will be answered by fitting a linear regression model to our data and determining which statistics give us the highest predictive power. To answer the second question, we will fit a linear model and run tests of significance to determine if including a player's NCAA conference affiliation helps explain their projected wins above replacement in their NBA rookie season.

2 Analyses, Results, and Interpretation

Question 1

We have several variables in our dataset that could be used to predict rookie season WAR, so we used several variable selection methods to decide which predictors to include in our model.

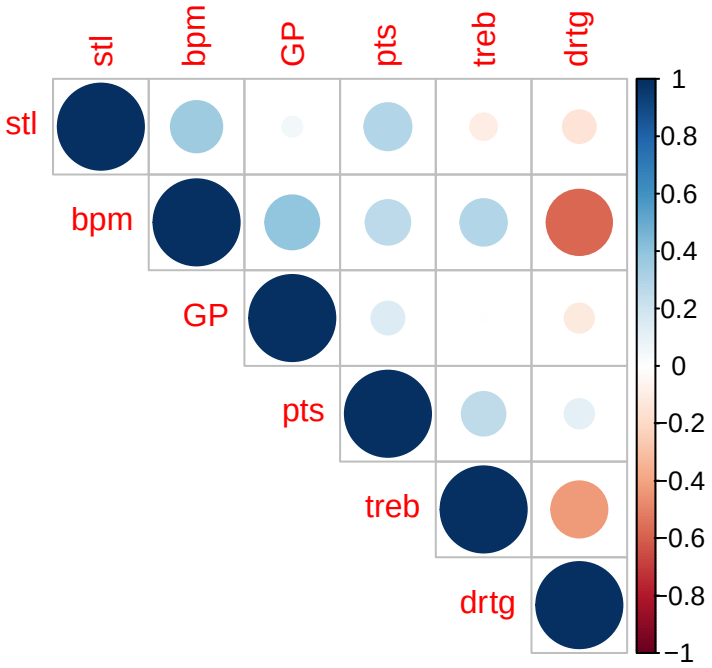
```
Call:
lm(formula = war_total ~ stl + bpm + GP + pts + treb + drtg +
    conf, data = nba)

Residuals:
    Min      1Q  Median      3Q     Max
-7.0760 -0.4715 -0.1135  0.3013  6.8979

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.144265   1.194803  -1.795 0.073250 .
```

```
stl          0.495596   0.106412   4.657 4.01e-06 ***
bpm          0.079398   0.022043   3.602 0.000344 ***
GP          -0.015509   0.009692  -1.600 0.110105
pts         -0.038678   0.014155  -2.732 0.006487 **
treb         0.057516   0.026091   2.204 0.027903 *
drtg         0.021736   0.011515   1.888 0.059597 .
conf1       -0.178678   0.124965  -1.430 0.153329
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.194 on 557 degrees of freedom
Multiple R-squared:  0.09118,   Adjusted R-squared:  0.07976
F-statistic: 7.983 on 7 and 557 DF,  p-value: 2.964e-09
```

We chose to use the step wise selection method with AIC to give us our variables, because the model selected yielded the highest R^2 and adjusted R^2. This is our chosen model:

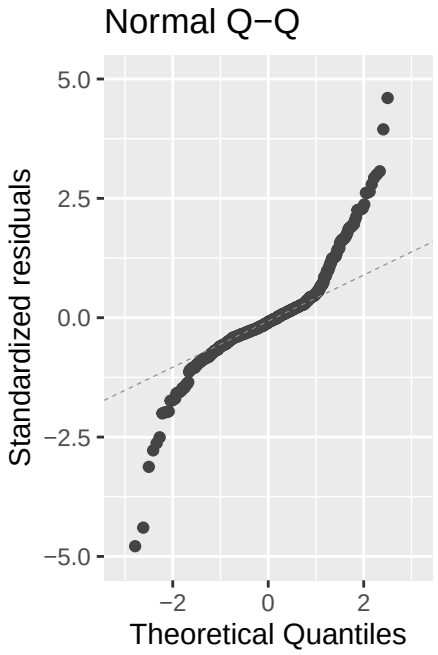
$\hat{war total}_i = 5.248713 + 0.495596 \times \text{stl}_i + 0.079398 \times \text{bpm}_i - 0.015509 \times \text{GP}_i - 0.038678 \times \text{pts}_i + 0.057516 \times \text{treb}_i + 0.021736 \times \text{drtg}_i - 0.178678 \times (conf_i = 1)$

Once the variables were selected, we then wanted to check to make sure there was no multi-collinearity between the chosen predictors.



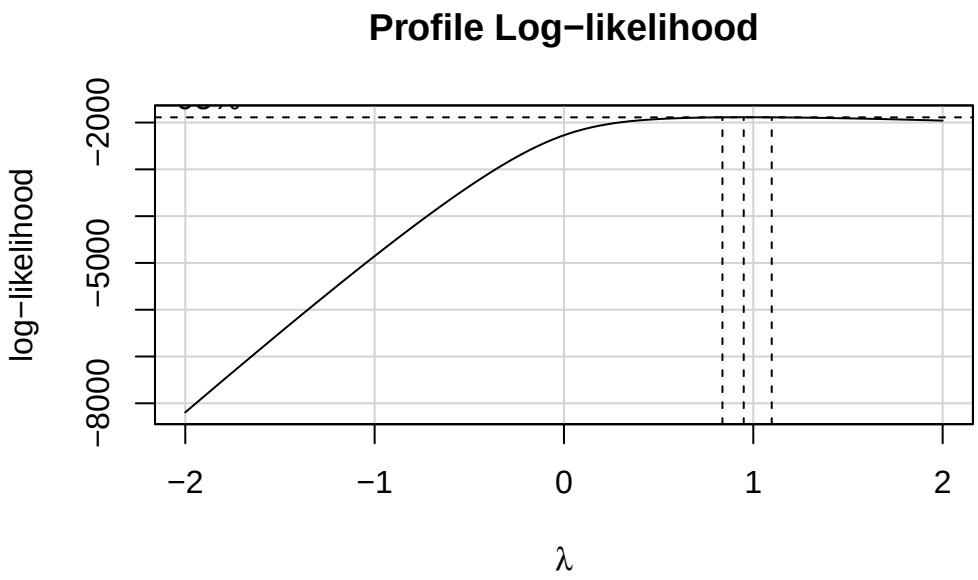
There was no indication that any of the variables were highly correlated, so that assumption was checked. We checked all other assumptions in the appendix, and all were sufficiently met

except for normality of the residuals. Here is the output when checking our model for normality of residuals:




```
    Shapiro-Wilk normality test

data:  hoops_lm$residuals
W = 0.83797, p-value < 2.2e-16
```

Because this assumption was not met, we attempted to transform the response variable of our model. We used the Box-Cox test to discover the optimal lambda value to apply to our response variable, but the result was close enough to 1 so we chose not to apply any type of transformation.




```
[1] 0.9494949
```

Because the adjusted R^2 of 0.079 is so low, and the assumption of normality is not met, we cannot sufficiently say that a player's rookie season WAR in the NBA can be confidently predicted based on their last year of college stats. This makes sense due to the nature of scouting a prospect and how rookies assimilate to an NBA roster. Player evaluation includes detailed film study, interviews, and body measurables, all of which were not included in this dataset. However, for the sake of this project, we will attempt to create a confidence interval for the mean NBA rookie season WAR of BYU star Jaxson Robinson. Using his stats from his most recent year at BYU, we created the following confidence interval:

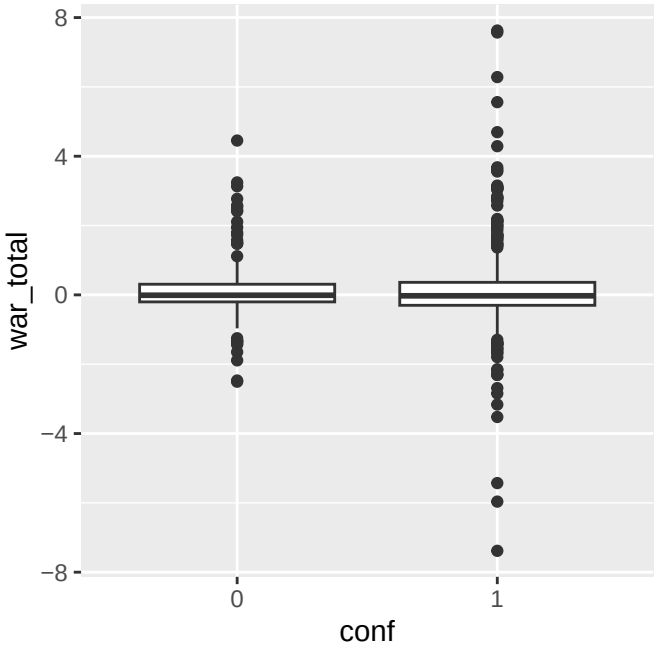
```
predict(hoops_lm,
        newdata = data.frame(stl = 0.7,
                             bpm = 4.9,
                             GP = 33,
                             pts = 14.2,
                             treb = 2.5,
                             drtg = 103.9,
                             conf = '1'),
        interval = "confidence",
        level = 0.95)
```

```
         fit        lwr         upr
1 -0.2458759 -0.4555091 -0.03624284
```

We are 95 % confident that Robinson's rookie season war will be between -0.4555091 and -0.03624284 on average. His wins above replacement project to be slightly below average, but due to the low explainability of our model, other factors could raise or lower this projection.

Question 2

Most of the highest rated high school and international players choose to play college basketball at a school found in one of the high major conferences (Big 12, Big 10, Pac-12, SEC, ACC, Big East). We wanted to see if a player's rookie success in the NBA can be determined by their participation in one of these conferences. In our data, if a player played their final season in a major conference, their 'conf' value is 1, and 0 if they played in a mid major conference.



The mean WAR is similar between players who played in high major conferences vs players who did not, but high major conference players have a wider range of rookie season WAR's, likely due to the higher number of such players in our data. Because of the overlap of box plots, there does not appear to be a significant difference between to the two levels of conference.

To confirm this, we will create a model with 'conf' as the only predictor variable and test it's significance.

```
Call:
lm(formula = war_total ~ conf, data = nba)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5180 -0.4166 -0.1625  0.2071  7.4908

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.15221    0.09817   1.550    0.122
conf1       -0.01715    0.11610  -0.148    0.883

Residual standard error: 1.246 on 563 degrees of freedom
Multiple R-squared:  3.874e-05, Adjusted R-squared:  -0.001737
F-statistic: 0.02181 on 1 and 563 DF,  p-value: 0.8826
```

The p-value for the conf = 1 variable is well above 0.05, and therefore we conclude that by itself, the conference a player plays in is not a good predictor of rookie season WAR.

3 Conclusions

Due to the limitations in our data previously mentioned, unfortunately we were unable to find a conclusion that is fully reliable. However, our analysis offers valuable insights for evaluating the potential impact of college players on NBA teams, particularly in predicting their Wins Above Replacement (WAR) scores during their rookie seasons.

From our analysis and testing, we found that out of the various predictor models we started out with, the ones that were most significant were steals, box plus-minus, games played, points, total rebounds, defensive rating, and conference status. We found these variables to be important through a stepwise sequential variable section method starting with our base model and using the AIC as our measurement. With these selected predictor variables, we were able to fit a linear model with WAR score as our response variable. With this model, we were able to find the coefficients for each predictor variable, thus being able to identify their individual impact while holding other variables constant.

With this data, we first and foremost learned that trying to predict a rookie's impact on his NBA team through WAR score based on their final collegiate season statistics is rather difficult to do with much confidence. That being said, we have learned that there is a good chance that the previously mentioned variables are ones that are deserving of attention of scouts or anyone interested in a player's NBA potential. We have also learned it does not appear that

having played in one of the "High Major" conferences in college will boost a player's WAR score for their first year in the NBA.

4 Contributions

Eli contributed by creating the framework of our project. He worked hard going through previous coding exercises from the class and gathering notes, descriptions, and examples of code for all the stages our project would have. This was extremely useful as we were then able to make quick progress once our data was properly set up. He also contributed by taking turns being the one to type as our group discussed what we needed to do for each stage of the project.

Brady contributed to the project by finding our data sets and joining them for us to be able to use. This was vital so that we could look at both stats from each player's final college year as well as their NBA rookie year WAR score. He also worked hard to prepare this final report. Within the format, he was the one to gather the appropriate code needed for the Analysis section and provided the descriptions/interpretations for the various parts within it.

Dallin played an important part in our group project by being the one noting down all that we were doing in our analysis. This enabled us to have a detailed description of what we had done so that we could then organize all of our work properly and know what we had done to get the results we did. Dallin also aided in the organization of this final report by taking the format provided us and organizing the notes he took accordingly. He also was the one to completely put together the Abstract section of this report.

Hunter contributed to the overall project by being involved in the data cleaning and preparation stages. He analyzed the different data sets we had and determined which variables needed to be removed or adjusted in order to perform the analysis we wanted to do and created code to do so, such as the creation of dummy variables. He also took turns being the one to type and code as the group all discussed various stages of the project. As for this final report, he provided the Conclusion and Contributions sections.

Overall, our group worked together very well on this project. We all met together various times to work on it and this allowed us to collaborate with everyone on everything we did. There wasn't anything done or decided in the process of our project without the whole group being a part of it.

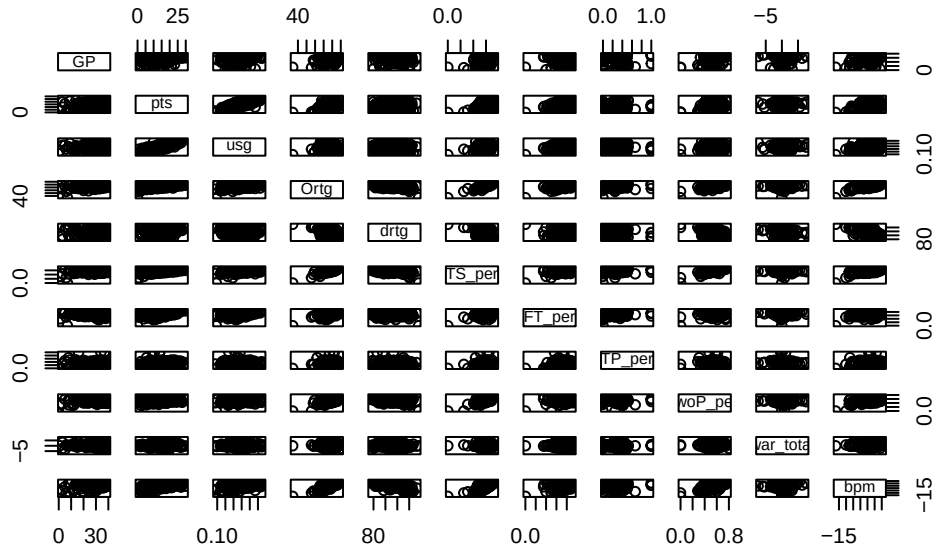
APPENDIX

Data Cleaning

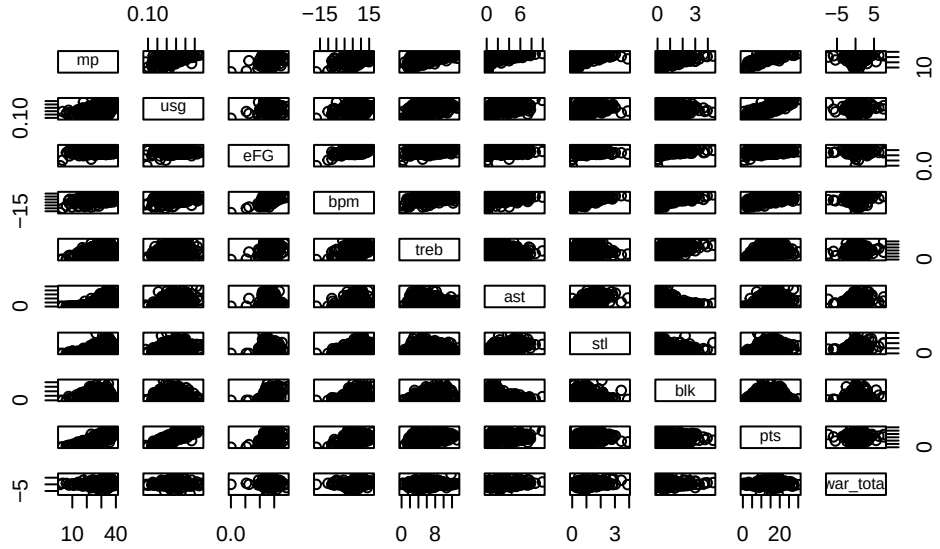
```
nba <- read_csv("nba_df.csv") %>%
  mutate(
    upperclassman = ifelse(yr %in% c("Sr", "Jr"), 1 , 0),
    upperclassman = as.factor(upperclassman),
    conf = factor(conf)
  )

nba <- subset(nba, select = -c(player_name, ast.tov, yr))
nba <- nba[, c(names(nba)[names(nba) != "war_total"], "war_total")]
```

Multicollinearity

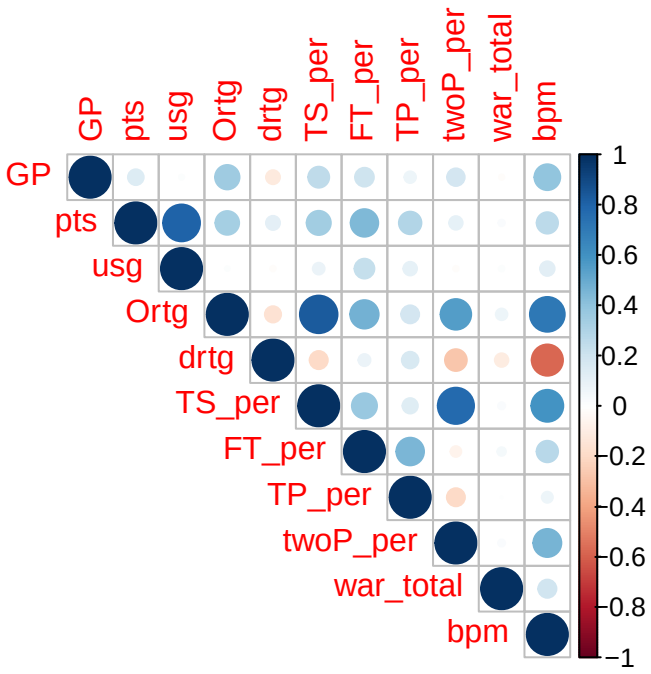
```
first_half <- subset(nba, select = c(GP, pts, usg, Ortg, drtg, TS_per, FT_per,
                                     TP_per, twoP_per, war_total, bpm))
second_half <- subset(nba, select = -c(GP,

                                       conf, Ortg, drtg, TS_per, FT_per,
                                       TP_per, twoP_per, upperclassman))

pairs(first_half)
```



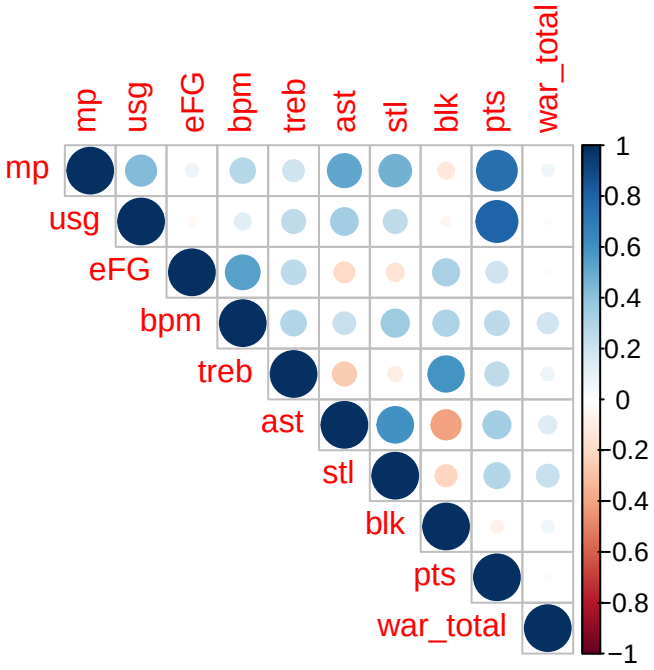

```
pairs(second_half)
```



```
corrplot(cor(first_half), type = "upper")
```




```
corrplot(cor(second_half), type = "upper")
```



Question 1

Variable Selection

```
nba_x <- data.matrix(nba[, 1:13]) # predictors
nba_y <- nba$war_total # response

#Elastic Net
elastic_net <- cv.glmnet(x = nba_x,
                         y = nba_y,
                         type.measure = "mse",
                         alpha = .5)
coef(elastic_net, se = "lambda.1se")
```

```
14 x 1 sparse Matrix of class "dgCMatrix"
                    s1
(Intercept) 0.139954
conf        .
GP          .
mp          .
Ortg        .
```

```
usg         .
eFG         .
TS_per      .
FT_per      .
TP_per      .
twoP_per    .
drtg        .
bpm         .
treb        .
```

```r
#Lasso Regression
lasso_reg <- cv.glmnet(x = nba_x,
                       y = nba_y,
                       type.measure = "mse",
                       alpha = 1)
coef(lasso_reg, se = "lambda.1se")
```

```
14 x 1 sparse Matrix of class "dgCMatrix"
                    s1
(Intercept) 0.139954
conf        .
GP          .
mp          .
Ortg        .
usg         .
eFG         .
TS_per      .
FT_per      .
TP_per      .
twoP_per    .
drtg        .
bpm         .
treb        .
```

```r
base_mod <- lm(war_total ~ 1, data = nba)
full_mod <- lm(war_total ~ ., data = nba)
```

```r
#Backward with AIC
backward_AIC <- step(full_mod,
                     direction = "backward",
```

```
                  scope=list(lower= base_mod, upper= full_mod),
                  trace = 0)
summary(backward_AIC)
```

```
Call:
lm(formula = war_total ~ mp + usg + bpm + treb + ast + stl, data = nba)

Residuals:
    Min      1Q  Median      3Q     Max
-7.1242 -0.4857 -0.1244  0.3244  6.9270

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.09594    0.34421   0.279  0.78056
mp          -0.02793    0.01201  -2.326  0.02037 *
usg         -2.07825    1.36531  -1.522  0.12853
bpm          0.03123    0.01574   1.984  0.04772 *
treb         0.07621    0.02648   2.878  0.00416 **
ast          0.08626    0.04472   1.929  0.05424 .
stl          0.49229    0.12241   4.022 6.58e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.196 on 558 degrees of freedom
Multiple R-squared:  0.08712,   Adjusted R-squared:  0.0773
F-statistic: 8.875 on 6 and 558 DF,  p-value: 2.906e-09
```

```
#Stepwise with AIC
step_base_AIC <- step(base_mod,
                  direction = "both",
                  scope=list(lower= base_mod, upper= full_mod),
                  trace = 0)
summary(step_base_AIC)
```

```
Call:
lm(formula = war_total ~ stl + bpm + GP + pts + treb + drtg +
    conf, data = nba)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-7.0760 -0.4715 -0.1135  0.3013  6.8979

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.144265   1.194803  -1.795 0.073250 .
stl          0.495596   0.106412   4.657 4.01e-06 ***
bpm          0.079398   0.022043   3.602 0.000344 ***
GP          -0.015509   0.009692  -1.600 0.110105
pts         -0.038678   0.014155  -2.732 0.006487 **
treb         0.057516   0.026091   2.204 0.027903 *
drtg         0.021736   0.011515   1.888 0.059597 .
conf1       -0.178678   0.124965  -1.430 0.153329
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.194 on 557 degrees of freedom
Multiple R-squared:  0.09118,   Adjusted R-squared:  0.07976
F-statistic: 7.983 on 7 and 557 DF,  p-value: 2.964e-09
```

New linear model based off stepwise, AIC

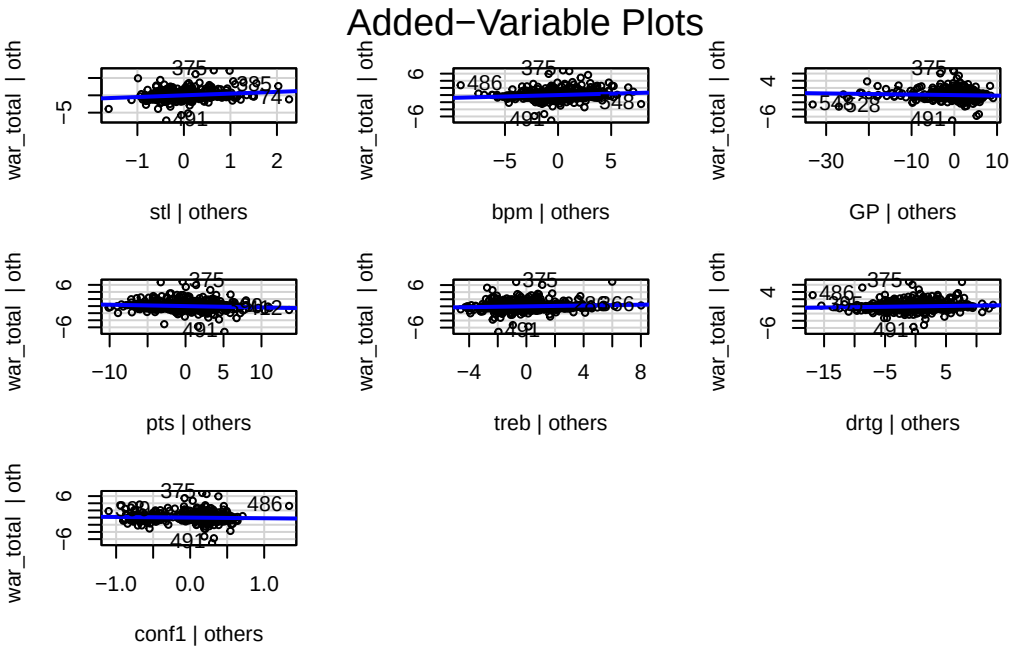
```
hoops_lm <- lm(war_total ~ stl + bpm + GP + pts + treb + drtg +
    conf, data = nba)

summary(hoops_lm)
```

```
Call:
lm(formula = war_total ~ stl + bpm + GP + pts + treb + drtg +
    conf, data = nba)

Residuals:
    Min      1Q  Median      3Q     Max
-7.0760 -0.4715 -0.1135  0.3013  6.8979

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.144265   1.194803  -1.795 0.073250 .
stl          0.495596   0.106412   4.657 4.01e-06 ***
```

```
bpm          0.079398   0.022043   3.602 0.000344 ***
GP          -0.015509   0.009692  -1.600 0.110105
pts         -0.038678   0.014155  -2.732 0.006487 **
treb         0.057516   0.026091   2.204 0.027903 *
drtg         0.021736   0.011515   1.888 0.059597 .
conf1       -0.178678   0.124965  -1.430 0.153329
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.194 on 557 degrees of freedom
Multiple R-squared:  0.09118,   Adjusted R-squared:  0.07976
F-statistic: 7.983 on 7 and 557 DF,  p-value: 2.964e-09
```

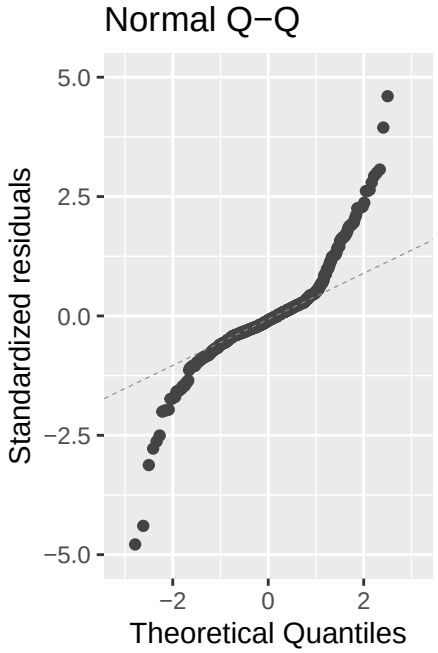
```r
#Add residuals
nba$residuals <- hoops_lm$residuals
```

Assumptions

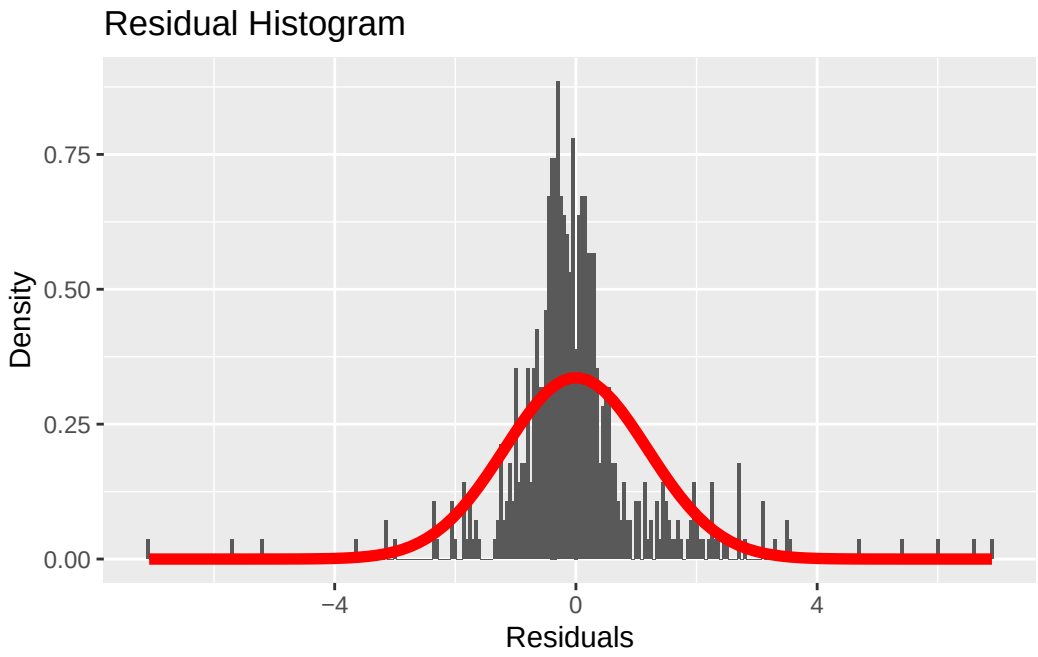


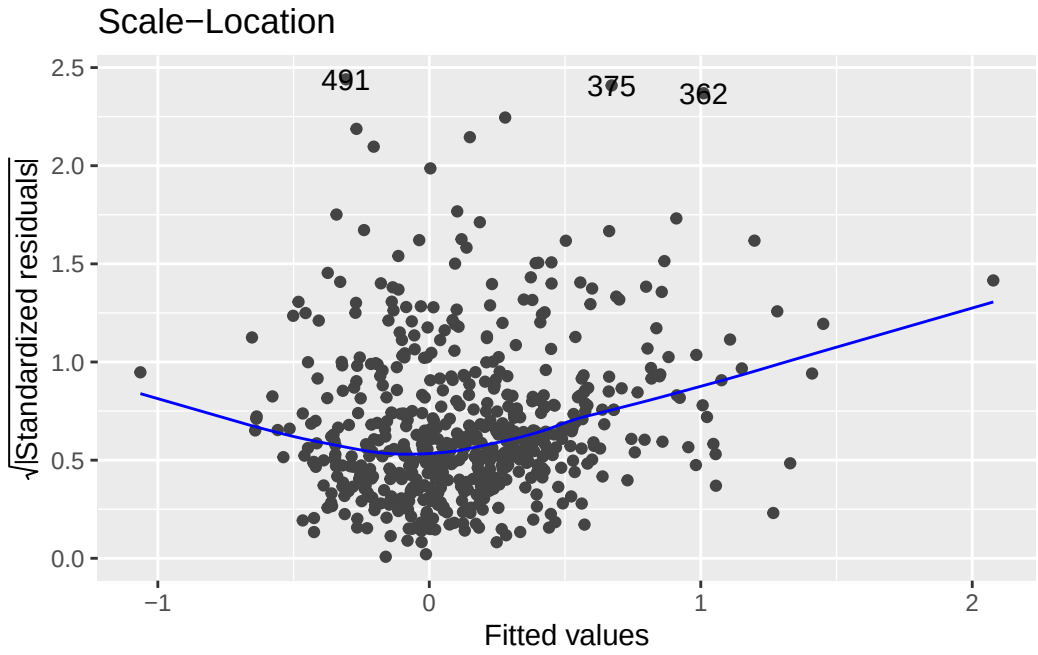

```
Warning: Removed 4 rows containing missing values or values outside the scale range
(`geom_point()`).
```

```
Warning: Removed 3 rows containing missing values or values outside the scale range
(`geom_text()`).
```



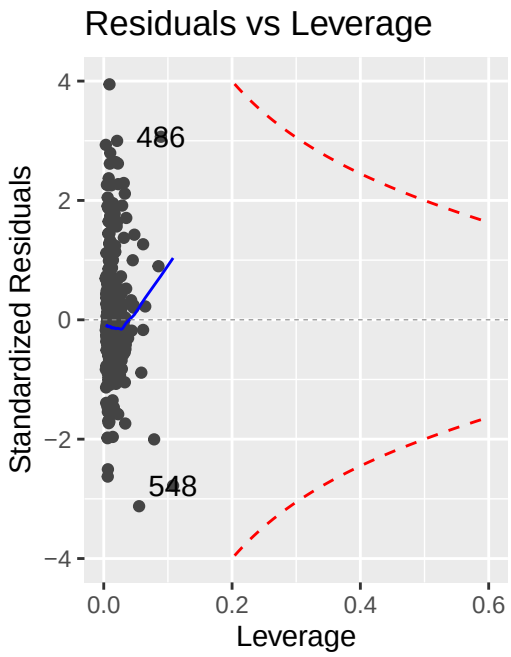

```
    Shapiro-Wilk normality test

data:  hoops_lm$residuals
W = 0.83797, p-value < 2.2e-16

Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.

Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
i Please use `after_stat(density)` instead.
```



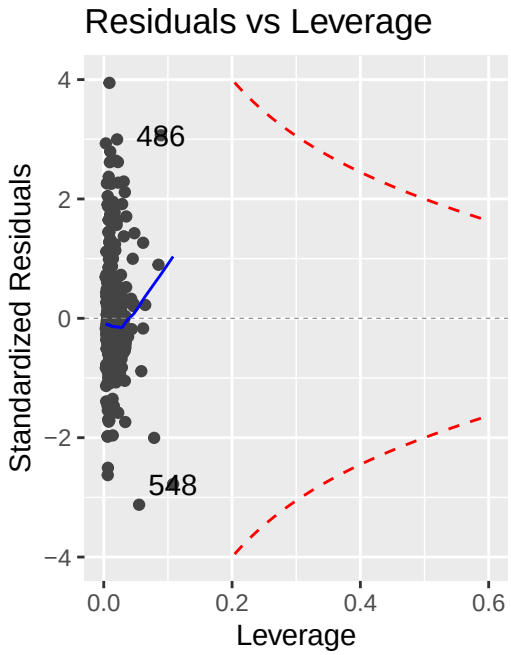



```
Warning: Removed 7 rows containing missing values or values outside the scale range
(`geom_point()`).

Warning: Removed 1 row containing missing values or values outside the scale range
```

(`geom_text()`).



Warning: Removed 7 rows containing missing values or values outside the scale range (`geom_point()`).
Removed 1 row containing missing values or values outside the scale range (`geom_text()`).



VIF

```
vifs <- vif(hoops_lm)
max(vifs)
```

```
[1] 2.631187
```

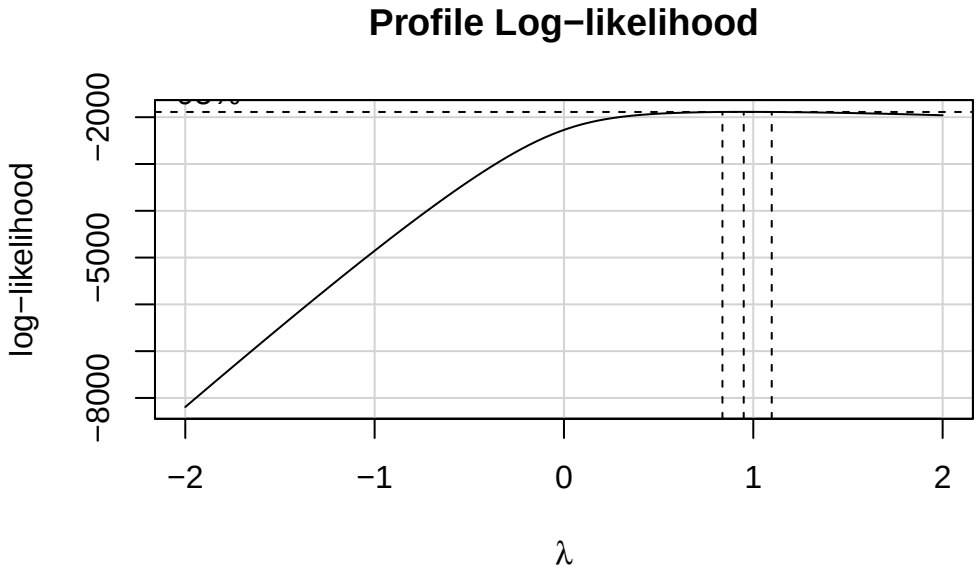
```
mean(vifs)
```

```
[1] 1.703283
```

Log Transformations

```
Call:
lm(formula = war_total ~ stl + bpm + GP + pts + treb + drtg +
    conf, data = log_nba)

Residuals:
```

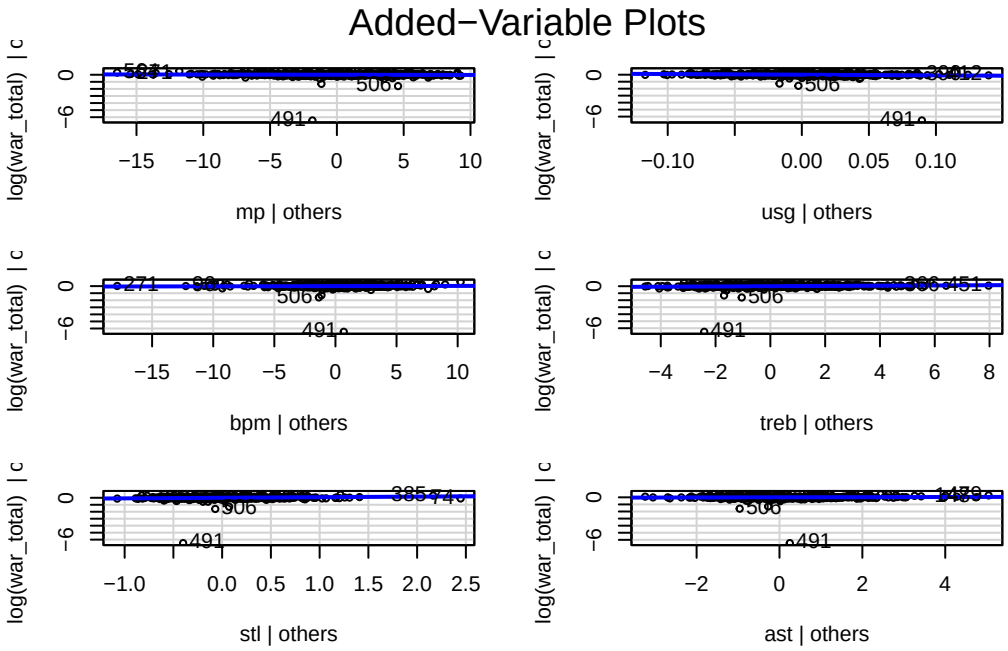
```
    Min      1Q  Median      3Q     Max
-7.0760 -0.4715 -0.1135  0.3013  6.8979

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.248713   1.194803   4.393 1.34e-05 ***
stl          0.495596   0.106412   4.657 4.01e-06 ***
bpm          0.079398   0.022043   3.602 0.000344 ***
GP          -0.015509   0.009692  -1.600 0.110105
pts         -0.038678   0.014155  -2.732 0.006487 **
treb         0.057516   0.026091   2.204 0.027903 *
drtg         0.021736   0.011515   1.888 0.059597 .
conf1       -0.178678   0.124965  -1.430 0.153329
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.194 on 557 degrees of freedom
Multiple R-squared:  0.09118,   Adjusted R-squared:  0.07976
F-statistic: 7.983 on 7 and 557 DF,  p-value: 2.964e-09
```




```
[1] 0.9494949
```

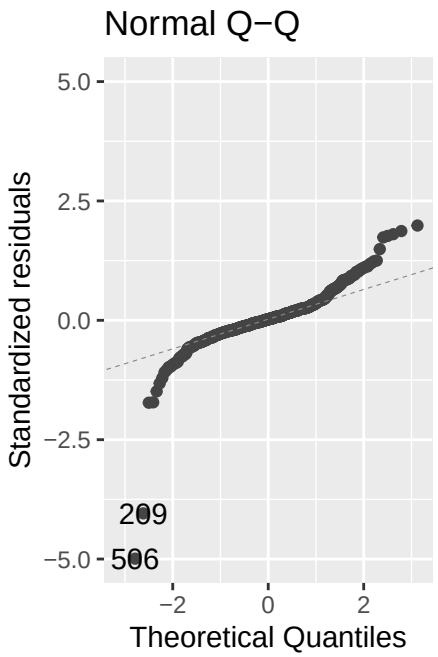
```
Call:
lm(formula = log(war_total) ~ mp + usg + bpm + treb + stl + ast,
    data = log_nba)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4687 -0.0592  0.0072  0.0744  0.6317

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.065847   0.092137  22.421  < 2e-16 ***
mp          -0.002786   0.003214  -0.867  0.38646
usg         -0.958873   0.365465  -2.624  0.00894 **
bpm          0.003000   0.004214   0.712  0.47684
treb         0.018073   0.007089   2.549  0.01106 *
stl          0.084218   0.032767   2.570  0.01042 *
ast          0.011013   0.011970   0.920  0.35792
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.32 on 558 degrees of freedom
Multiple R-squared:  0.03914,   Adjusted R-squared:  0.02881
F-statistic: 3.788 on 6 and 558 DF,  p-value: 0.001042
```

RECHECKING ASSUMPTIONS

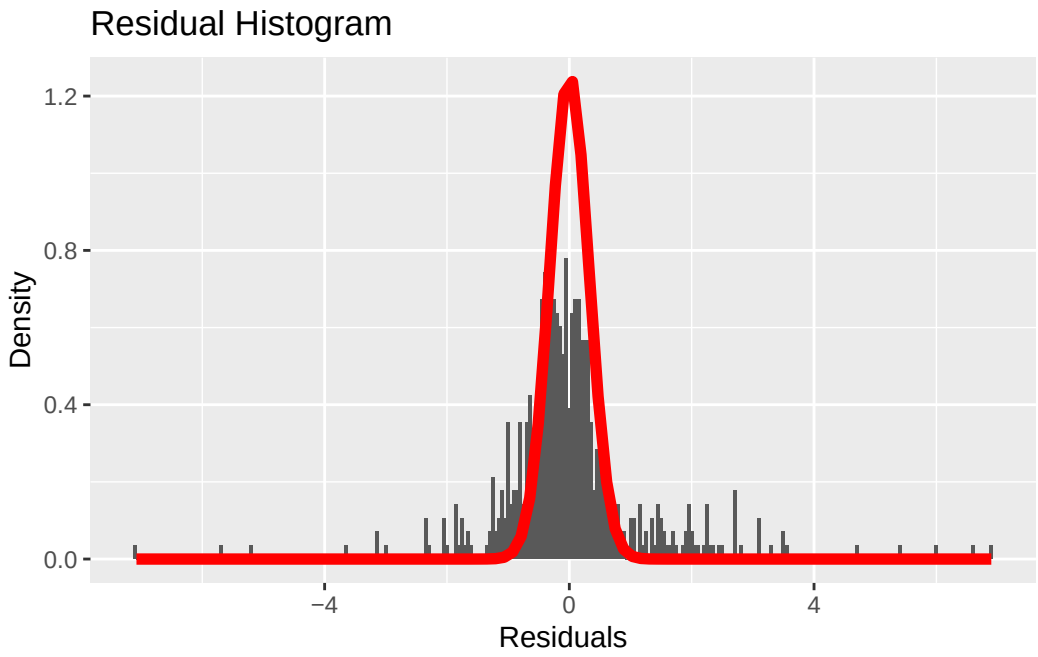


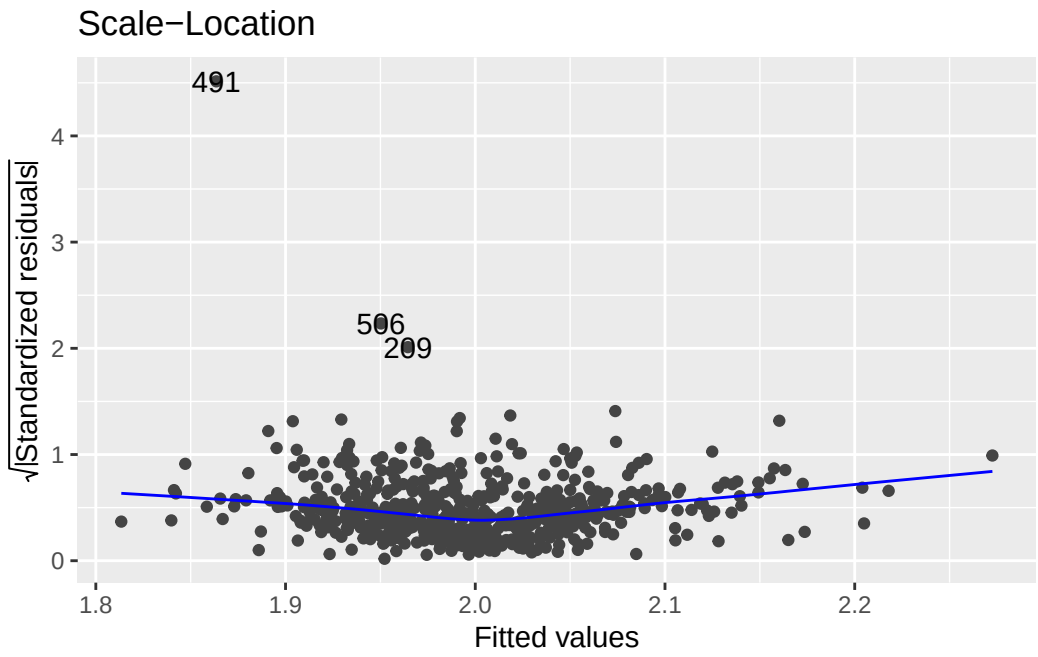

```
Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).
```

```
Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_text()`).
```



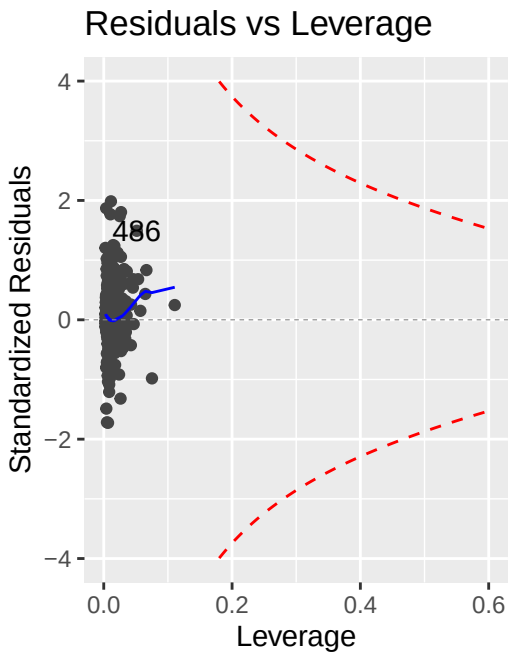

```
    Shapiro-Wilk normality test

data:  transform_lm$residuals
W = 0.33051, p-value < 2.2e-16
```






```
Warning: Removed 3 rows containing missing values or values outside the scale range
(`geom_point()`).

Warning: Removed 2 rows containing missing values or values outside the scale range
```

(`geom_text()`).


Residuals vs Leverage
Standardized Residuals
Leverage
4
2
0
−2
−4
0.0
0.2
0.4
0.6
486