Factors that Influence NBA Player WinShares

Mohamed Hassan-El Serafi

Abstract

This analysis examines factors that influence an NBA basketball player's WinShare totals. Using specific explanatory variables within the dataset, I explore the linear relationship between WinShares and each selected variable using a multiple linear regression model. Each observation in the dataset are players who were drafted between 1989 and 2021. Independent variables used include career player stats of points, rebounds, assists, field goal percentage, minutes played, plus/minus, and overall draft selection. For purposes of this analysis, players containing missing data were removed.

Because of its skewed distribution, WinShares was log-transformed and created as a separate target variable. ggpairs() was used to check for multicollinearity between each of the chosen independent variables. Field goal percentage, plus/minus, and overall draft pick selection did not show high correlation with other variables, and were used in the multiple linear regression model. The initial regression model using the original WinShare variable produced an Adjusted R-square of 32.63%. Although there were outliers, most of the data points in the Residual vs Fitted Values plot were clustered around the zero line threshold with no distinct pattern. The near normal residuals histogram showed skewness to the right, with its center at approximately zero. The QQ-plot displayed a relatively straight line, with its upper end positively skewed. This indicates that the conditions of linearity, near normal residuals, and constant variability are met. When replacing the target variable with its log-transformed counterpart, the model increased its Adjusted R-Square performance to 46.54%. The Residual vs Fitted Values displayed a cluster of data points around the zero threshold but had less outliers. The near normal residuals histogram showed a symmetrical normal distribution with its center at approximately zero. The line of data points in the QQ-plot was more straight with no discernible skewness. The revised model improved the model's overall performance.

Dataset

- Obtained from Kaggle: https://www.kaggle.com/datasets/mattop/nba-draft-basketball-player-data-19892021
- Contains 1,922 observations and 24 columns.
- Each observation is a NBA player drafted between 1989 and 2021.
- Players with missing data were drafted but did not play in the NBA and were removed from the dataset.

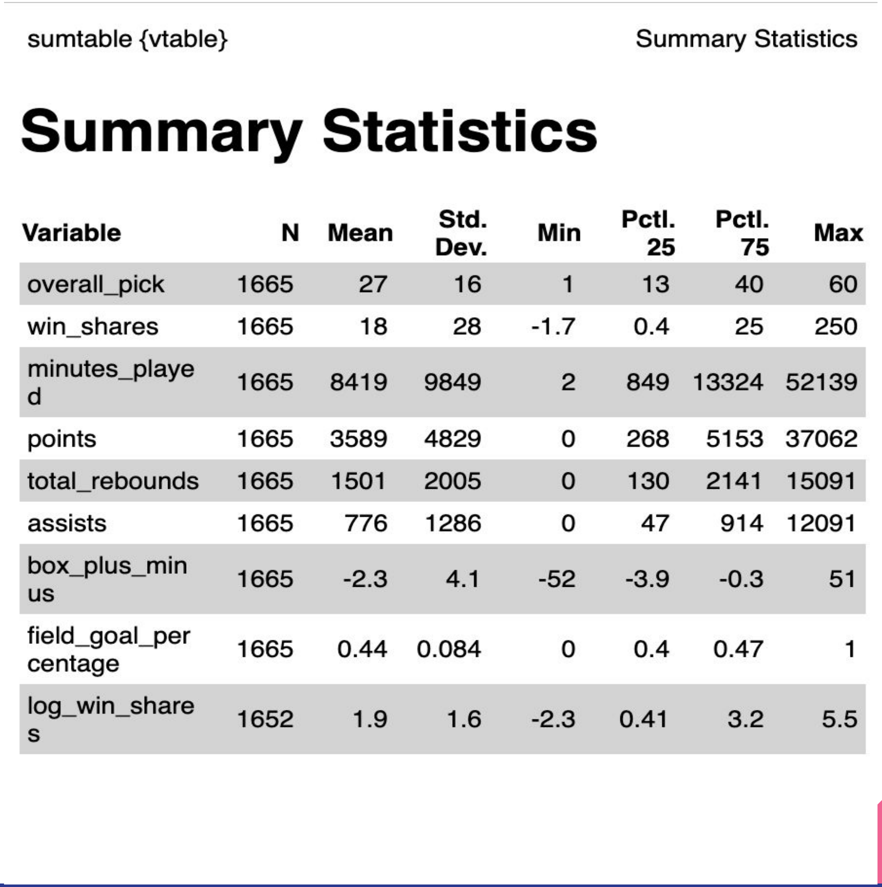
Dependent Variable

- WinShares
- Combination of Offensive and Defensive Win Shares
- Offensive Win Shares: marginal offense/marginal points per win
- Defensive Win Shares: marginal defense/marginal points per win
- Source: https://www.basketball-reference.com/about/ws.html

Independent Variables

- overall_pick: Overall draft selection of each player (categorical variable)
- points: Total career points
- total_rebounds: Total career rebounds
- assists: Total career assists
- field_goal_percentage: Career field goal percentage
- minutes_played: Total career minutes played
- box_plus_minus: Measure of a player's productivity on the court. Positive numbers indicate that the player helped increase their respective team's lead or decrease the deficit. A minus indicates that the deficit increased or the team's lead decreased.

Summary Statistics

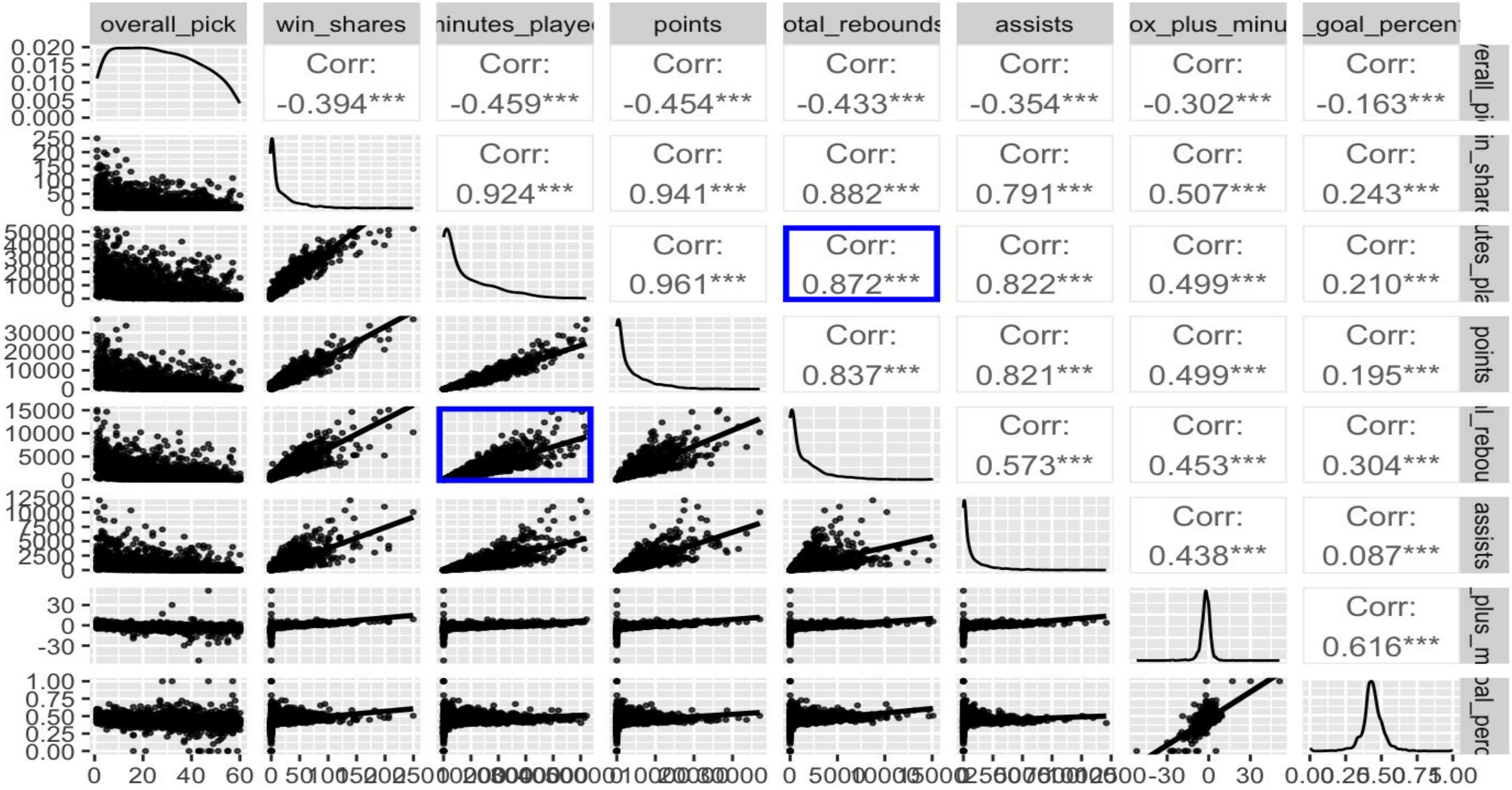


sumtable {vtable} Summary Statistics

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
overall_pick	1665	27	16	1	13	40	60
win_shares	1665	18	28	-1.7	0.4	25	250
minutes_played	1665	8419	9849	2	849	13324	52139
points	1665	3589	4829	0	268	5153	37062
total_rebounds	1665	1501	2005	0	130	2141	15091
assists	1665	776	1286	0	47	914	12091
box_plus_minus	1665	-2.3	4.1	-52	-3.9	-0.3	51
field_goal_percentage	1665	0.44	0.084	0	0.4	0.47	1
log_win_shares	1652	1.9	1.6	-2.3	0.41	3.2	5.5

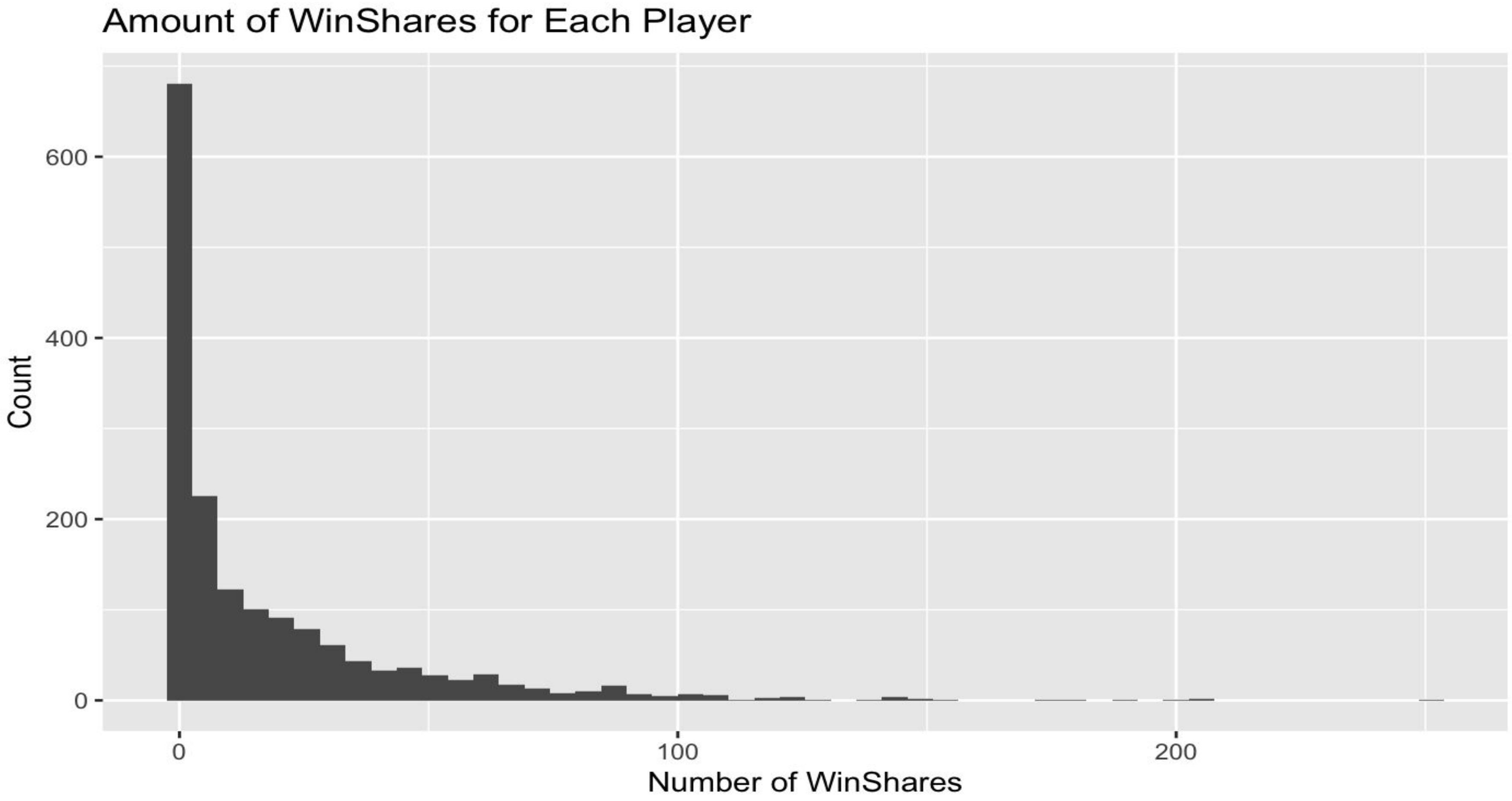
Correlation Coefficients for Each Variable



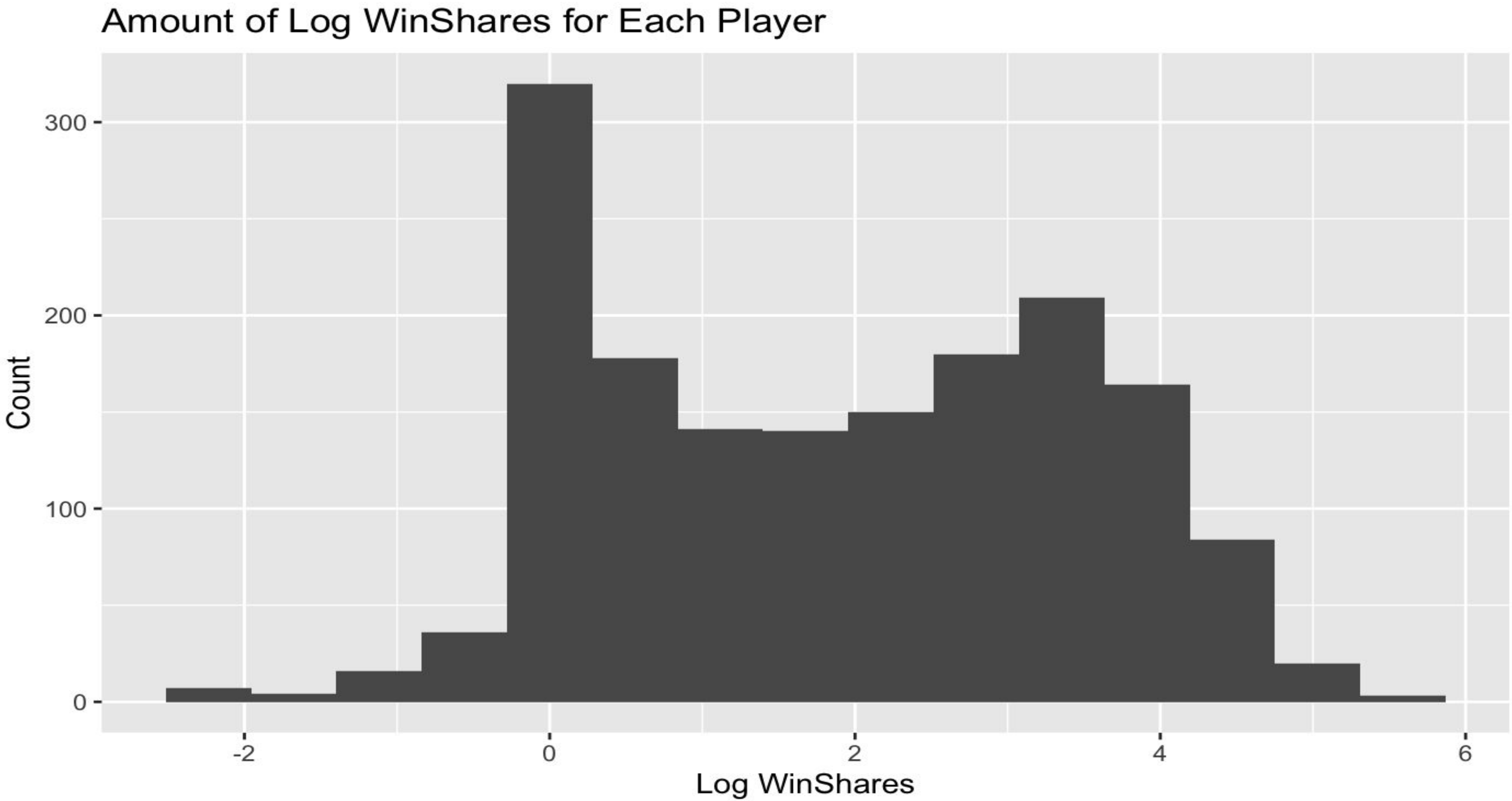
Independent Variables

- Field Goal Percentage
- Box Plus/Minus
- Overall Draft Pick Selection
- Variables not used due to high multicollinearity: points, assists, total rebounds, minutes played.

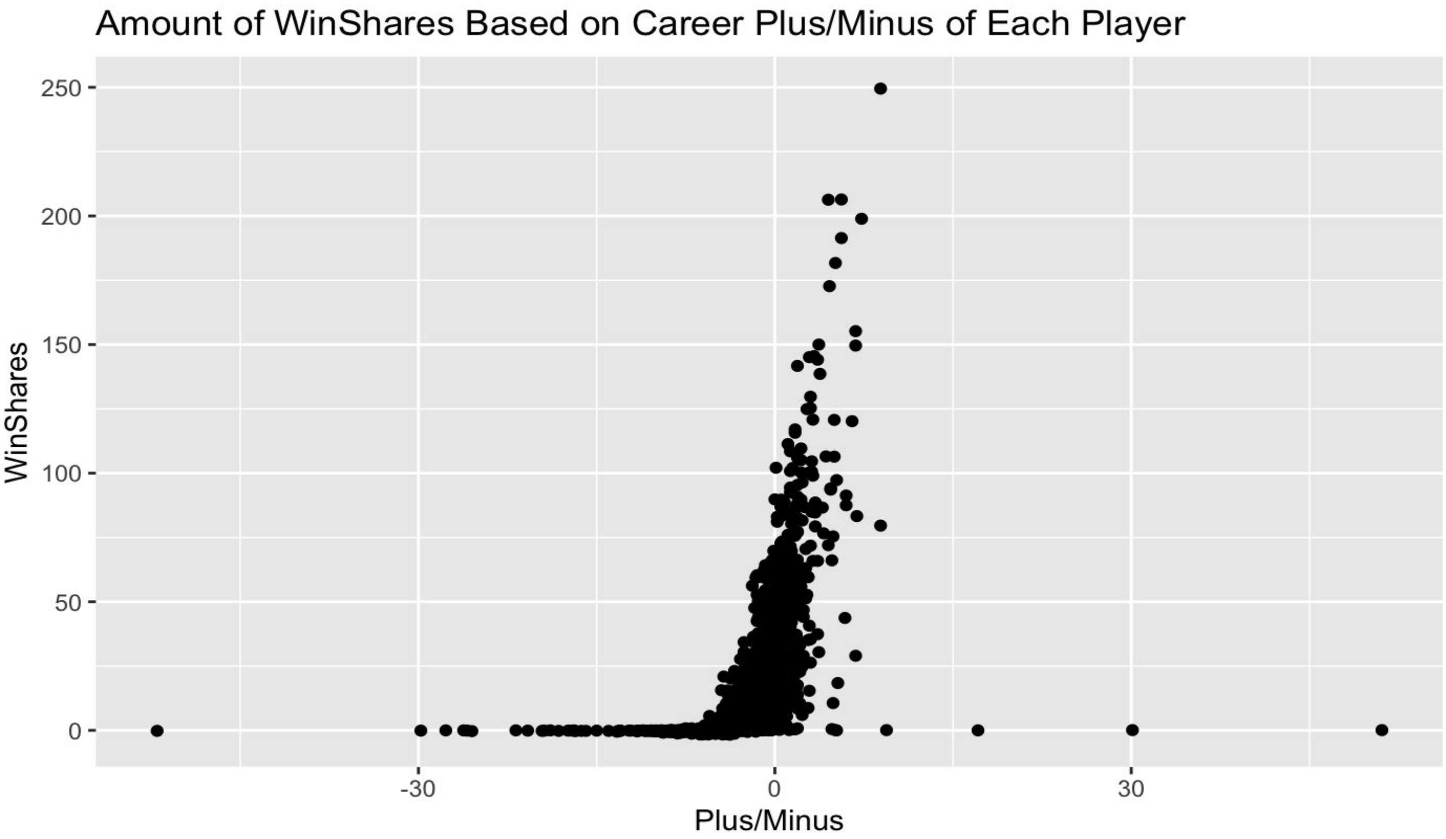
WinShares Distribution



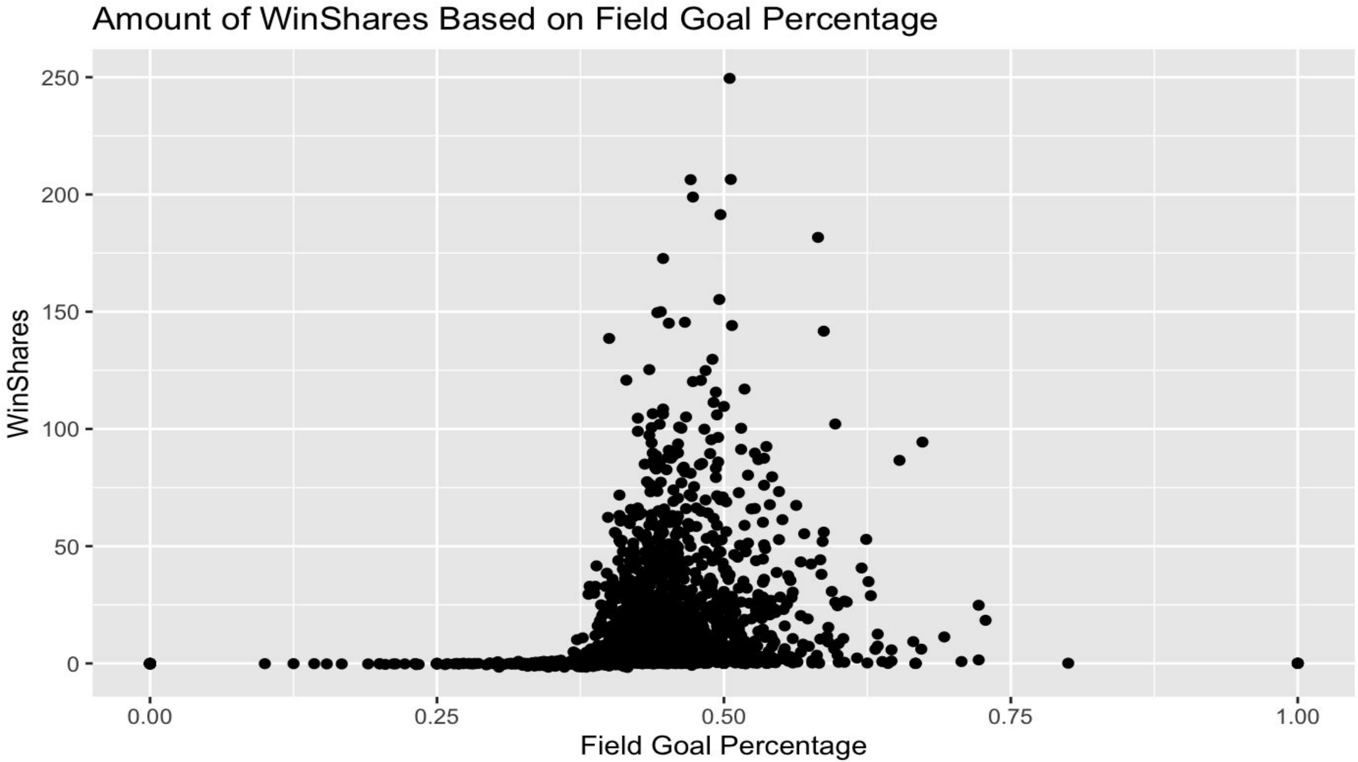
Log-Transformed WinShares



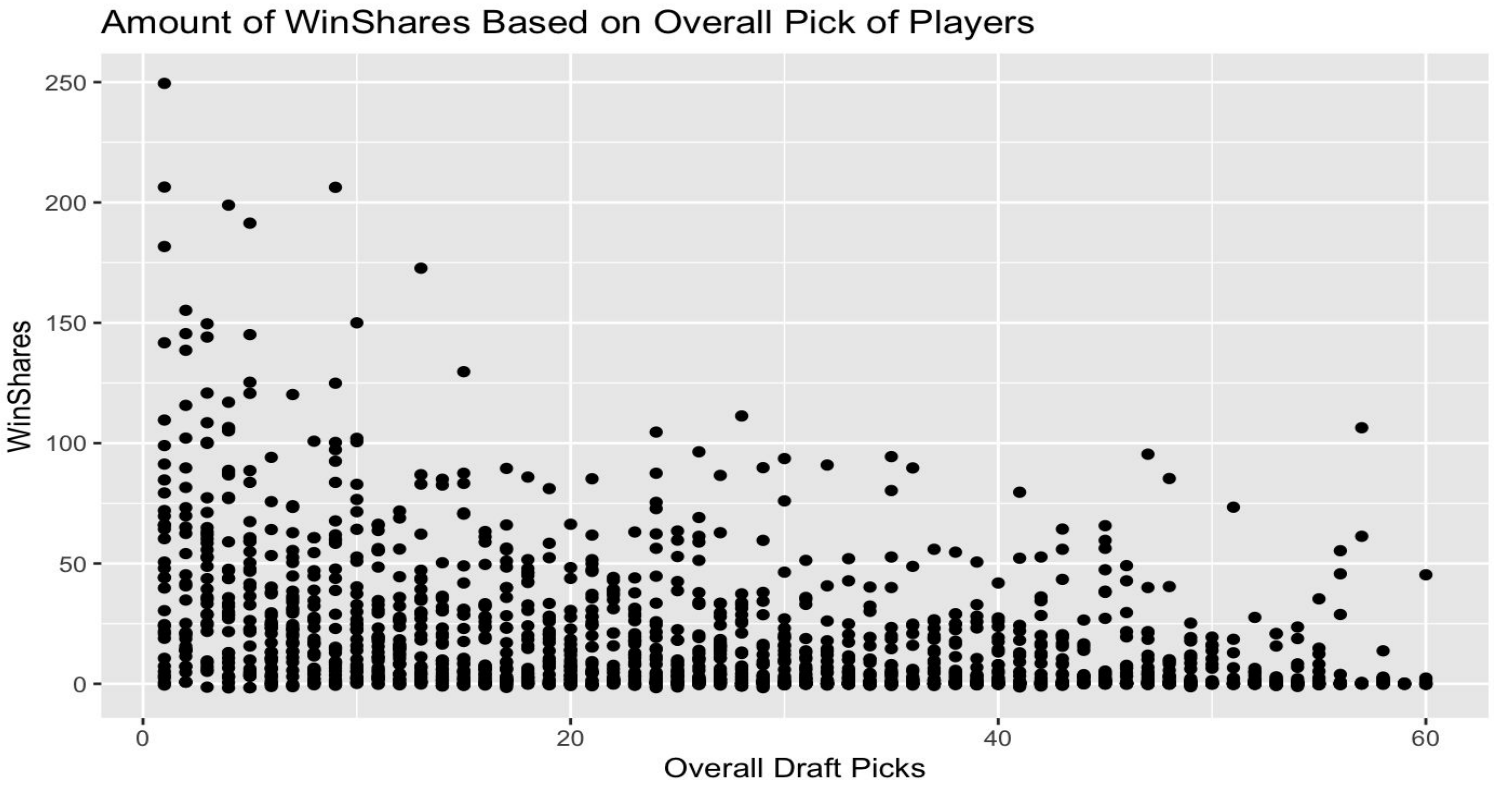
Relationship Between WinShares and Independent Variables: Box Plus/Minus



Field Goal Percentage



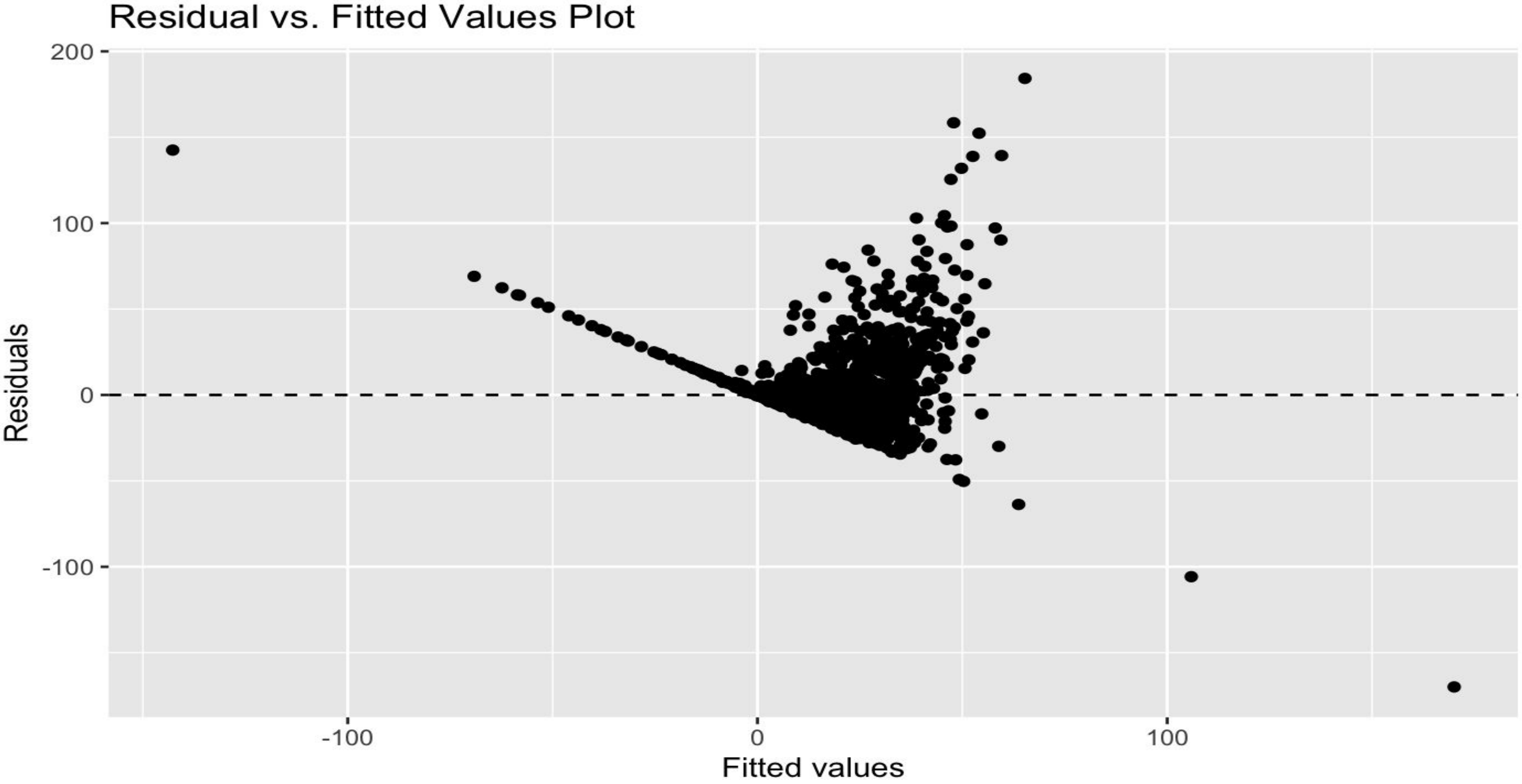
Overall Draft Pick Selection



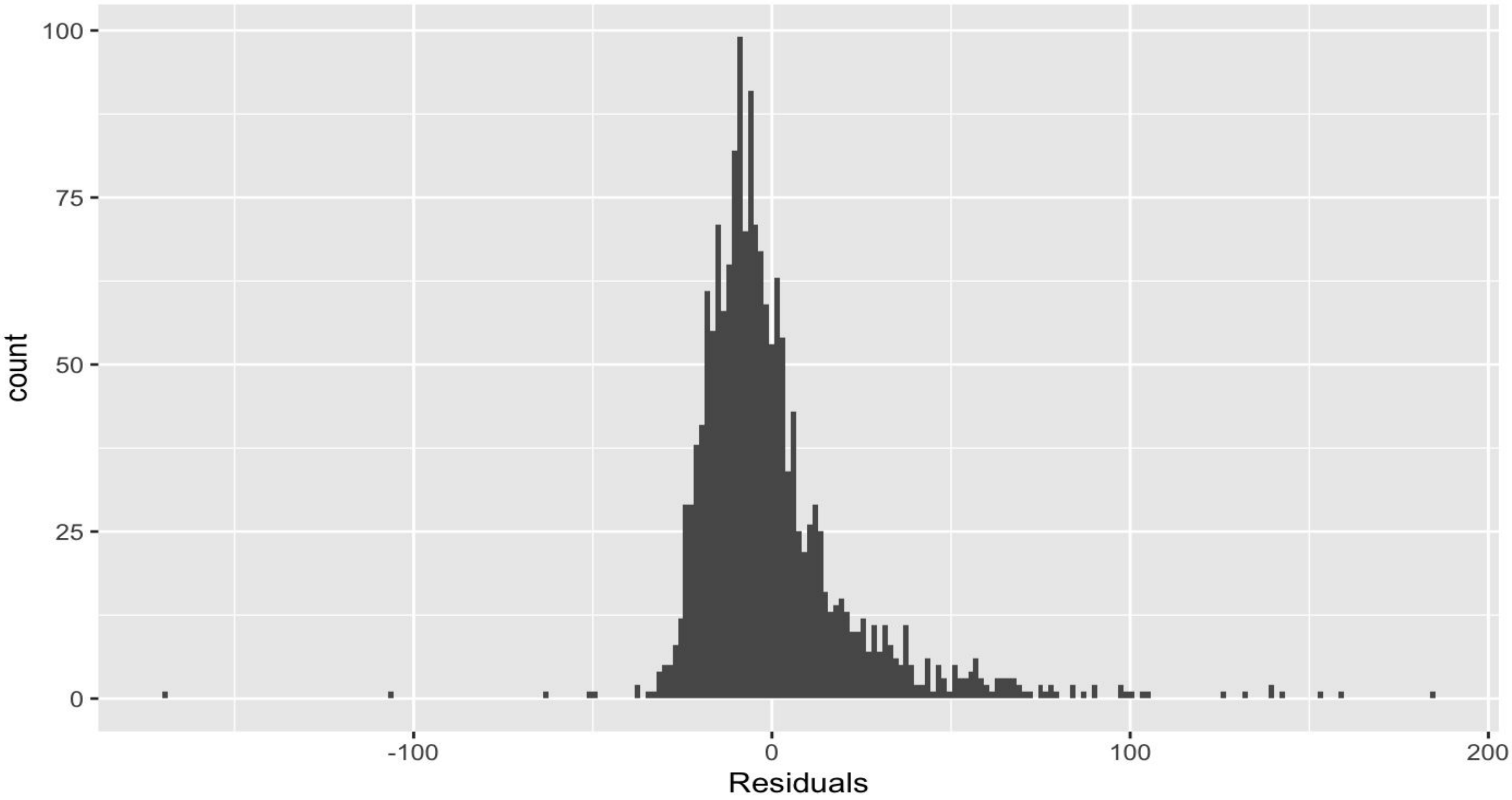
Initial Multiple Linear Regression Model

- Adjusted R-Square: 32.63%
- Residual vs Fitted Values: Clustered around zero threshold.
- Good amount of outliers.
- Near Normal residuals: shows skewness to the right, center approximately zero, narrow distribution
- QQ-plot: upper end of line positively skewed

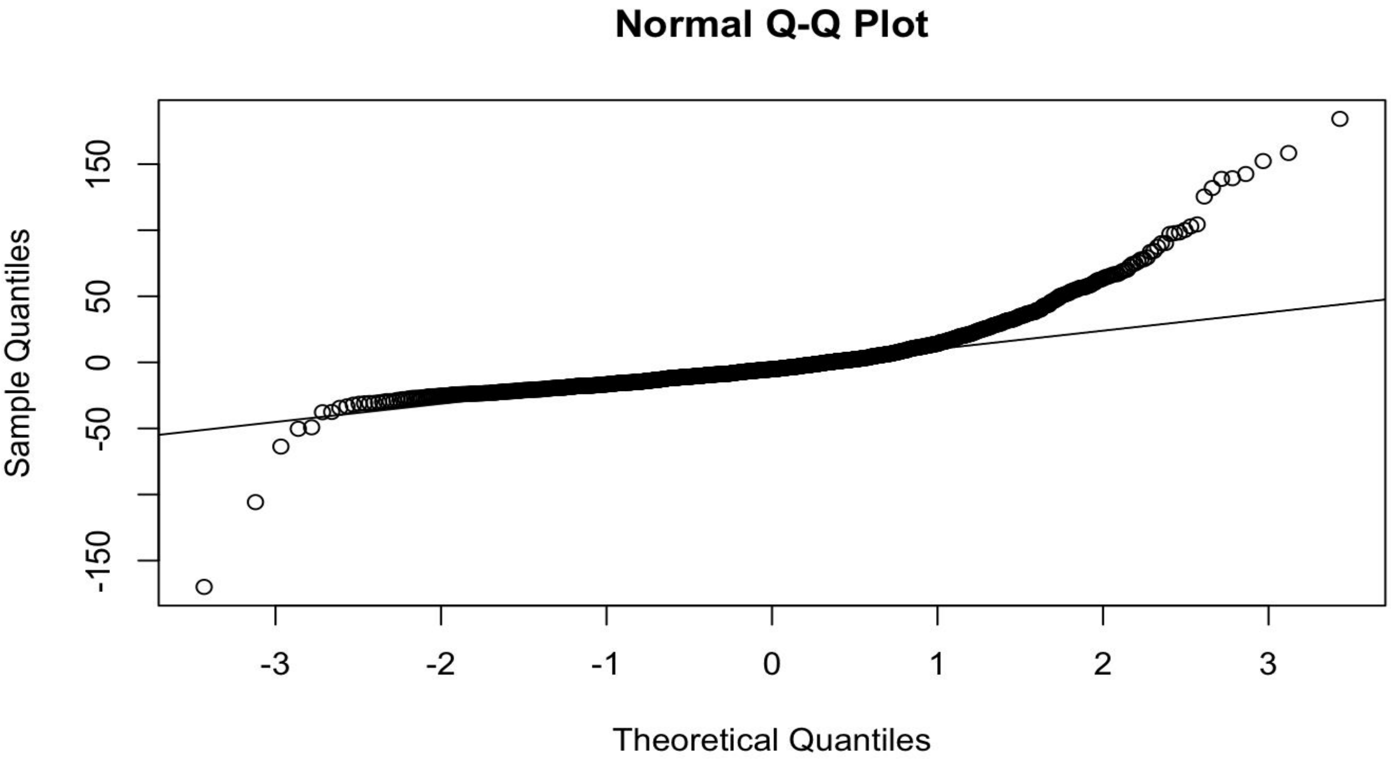
Residual vs Fitted Values



Near Normal Residual Histogram



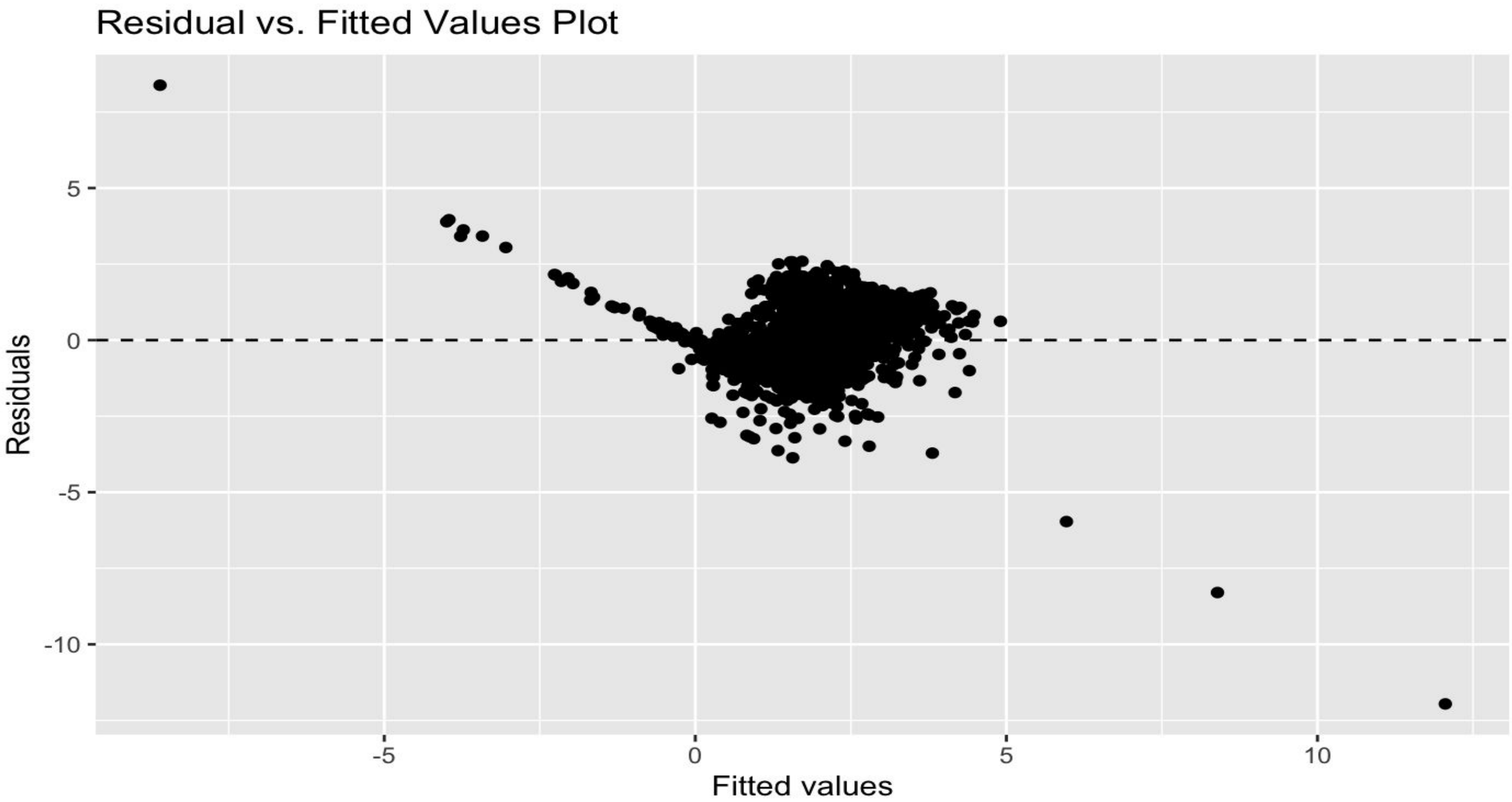
QQ-Plot


Normal Q-Q Plot
Sample Quantiles
150
50
0
-50
-150
Theoretical Quantiles
-3
-2
-1
0
1
2
3

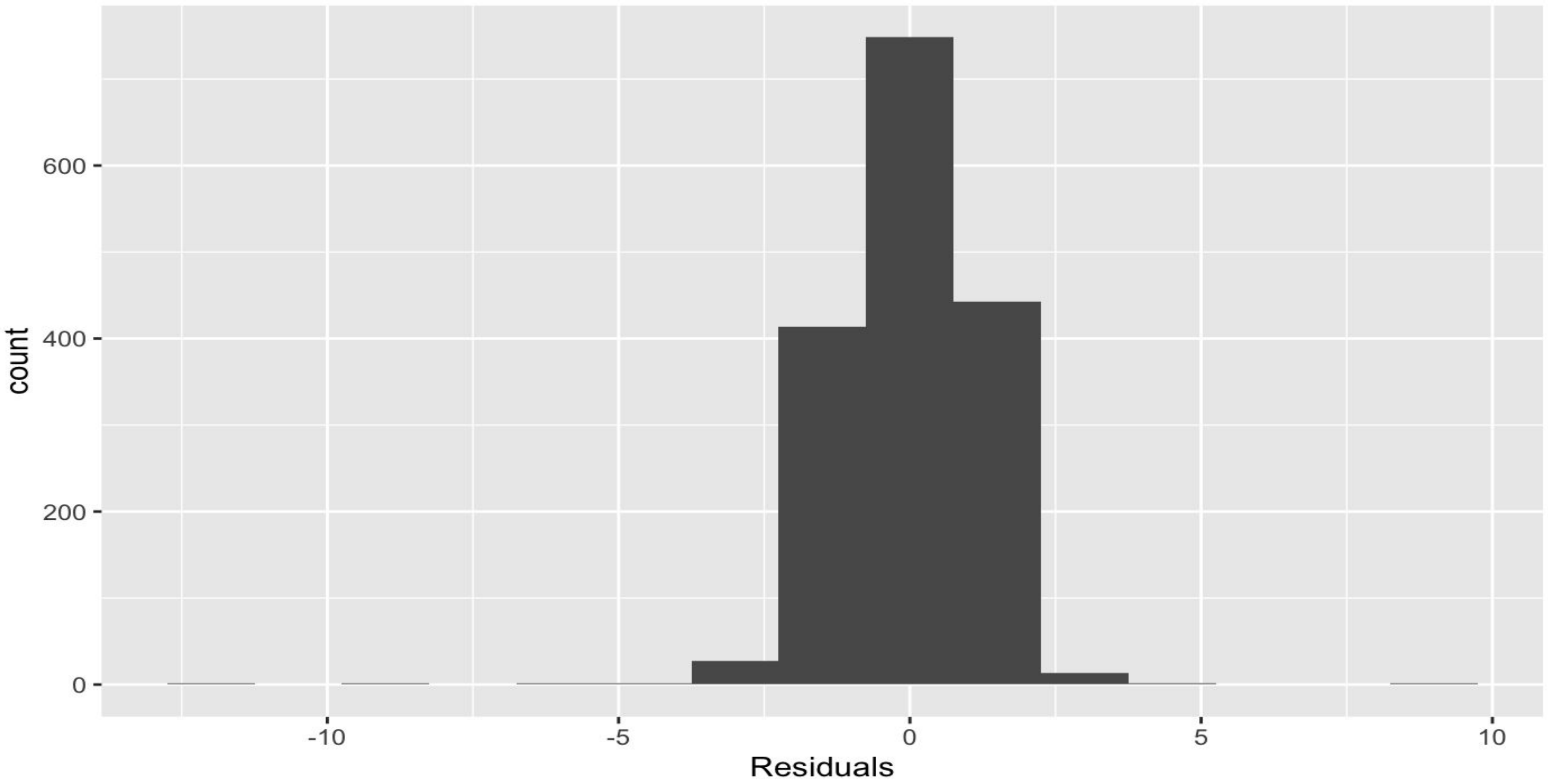
Multiple Linear Regression Model with Log WinShares

- Adjusted R-Square: 46.56%
- Field Goal Percentage > 0.05 (p-value)
- New Adjusted R-Square: 46.54%
- Residual vs Fitted Values: Less outliers, more clustered.
- Near normal residuals: symmetrical shape, wider distribution, center approximately zero.
- QQ-plot: line of data points are straight; no discernible skewness.

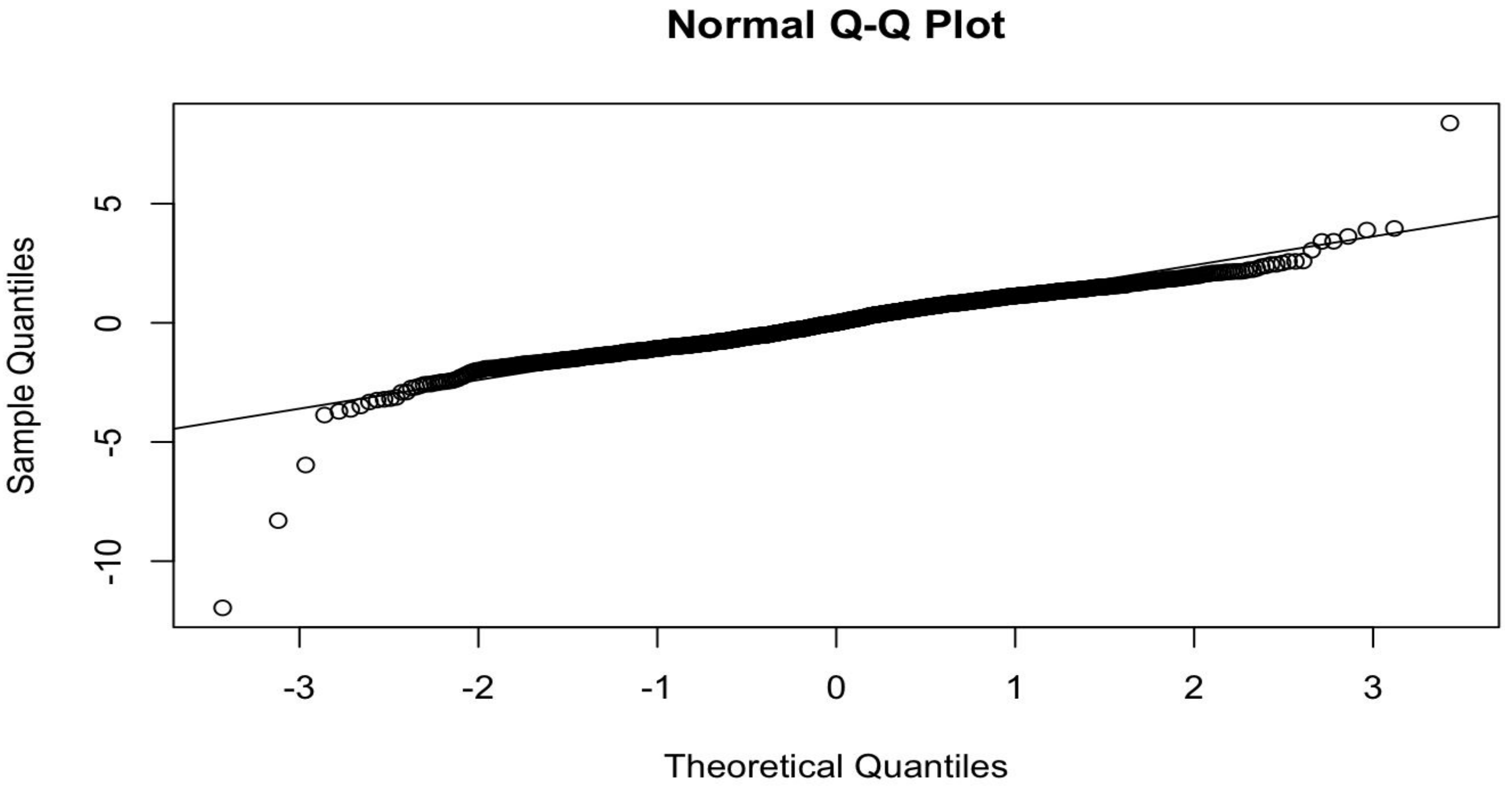
Residual vs Fitted Values



Near Normal Residuals



QQ-Plot


Normal Q-Q Plot
Sample Quantiles
5
0
-5
-10
Theoretical Quantiles
-3
-2
-1
0
1
2
3

Conclusion

- Log-transforming target variable (WinShares) improved overall model performance.
- Revised model improved Adjusted R-Square from 32.63% to 46.54%
- Improved Residual vs Fitted Values, Near Normal Residuals, and QQ-plot.
- Conditions of linearity, near normal residuals, and constant variability are met.
- 46.54% of the variance in WinShares can be explained by what draft position a player is selected and their plus/minus value.
- Possible next steps to further improve model: identify and remove additional outliers, log-transform independent variables.

Thank You!

Rpubs link: https://rpubs.com/moham6839/1042265

GitHub link: https://github.com/moham6839/Data606_Final_Project