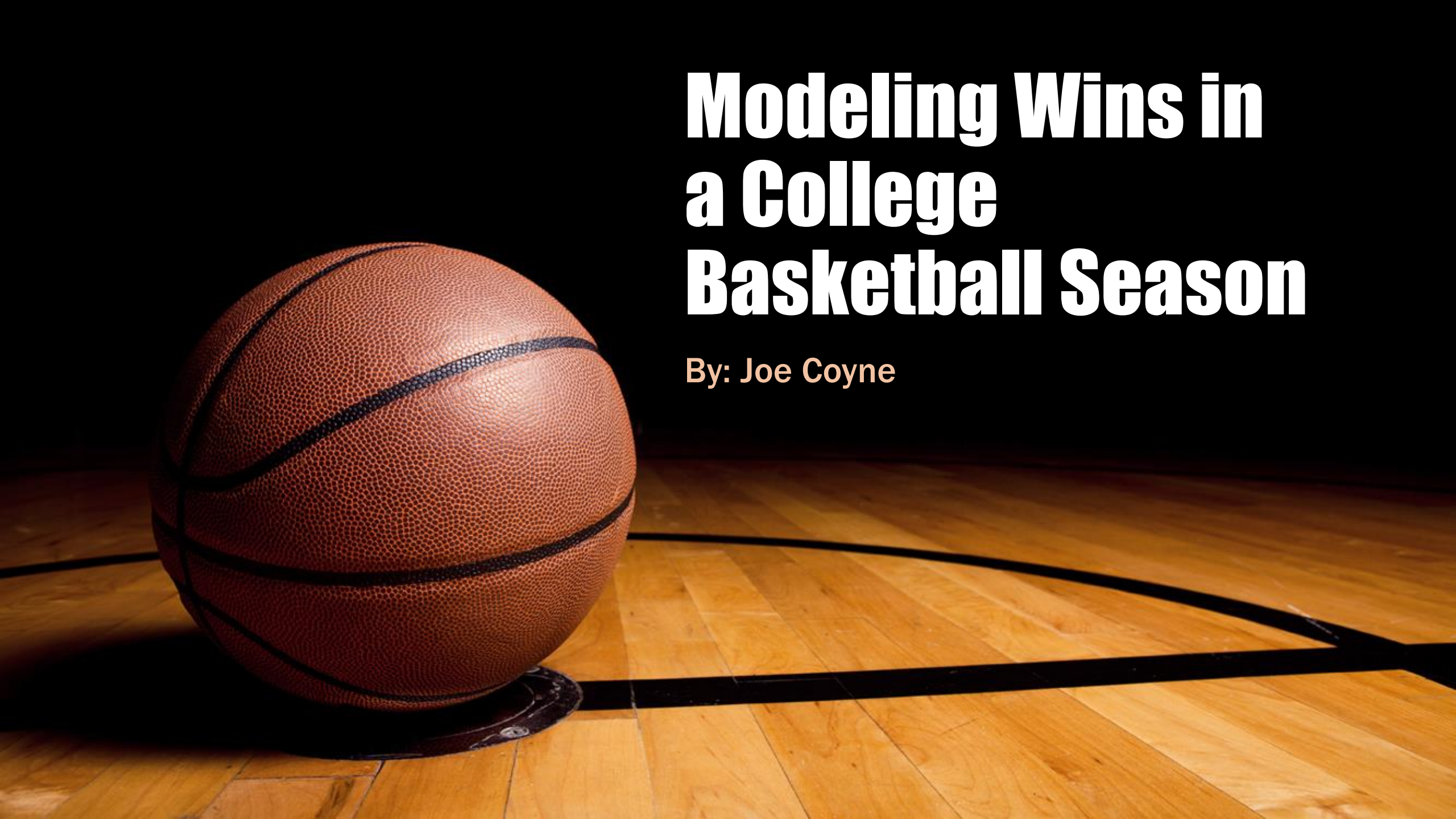


Modeling Wins in a College Basketball Season

By: Joe Coyne

Table of Contents

Introduction

It's often said that defense wins championships, but how true is this really?

Since rosters change so much from year to year, it can be helpful to know what type of play best sets a team up to win

This dataset looks at both standard and advanced basketball statistics

Project Goals/Hypotheses

- Explore what factors are associated with the number of wins a college basketball team has in a given season
- Does the conference a team is in has an impact on the total number of wins?
- Hypothesis: Having a higher scoring percentage than scoring allowed percentage will be important
  - 3 point percentage scored will be important as basketball emphasizes 3 point shooting more than in the past
  - Interaction term between tempo and shooting percentage allowed will be significant

Data Overview

- 3885 observations – Team and Year combinations
- Data spans from 2013-2024 and scrapes data from BartTorvik
- Variables collected:
  - ADJOE/ADJDE (Adjusted Offensive/Defensive Efficiency)
  - EFG_O/EFG_D (Effective Field Goal Percentage Shot/Allowed)

$$eFG\% = \frac{(FGM + (0.5 X 3PTM))}{FGA}$$

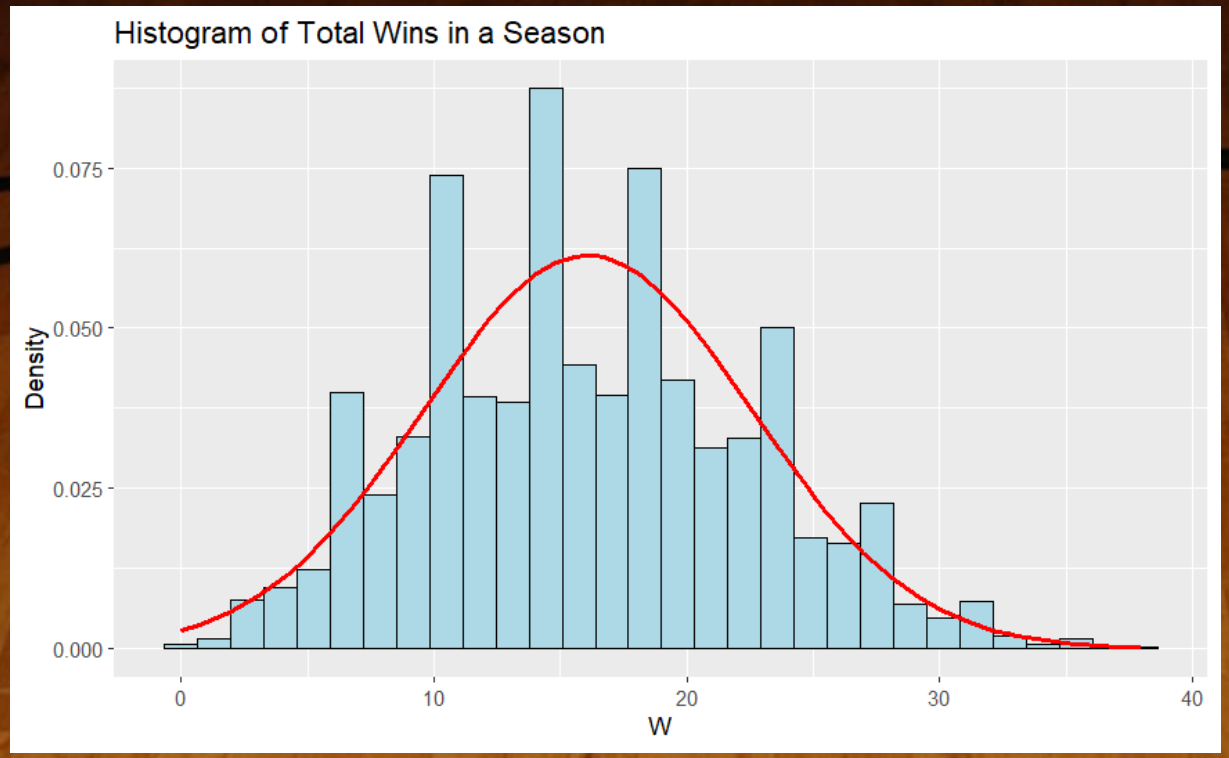
FGM = Field Goals Made (2PT and 3PT)

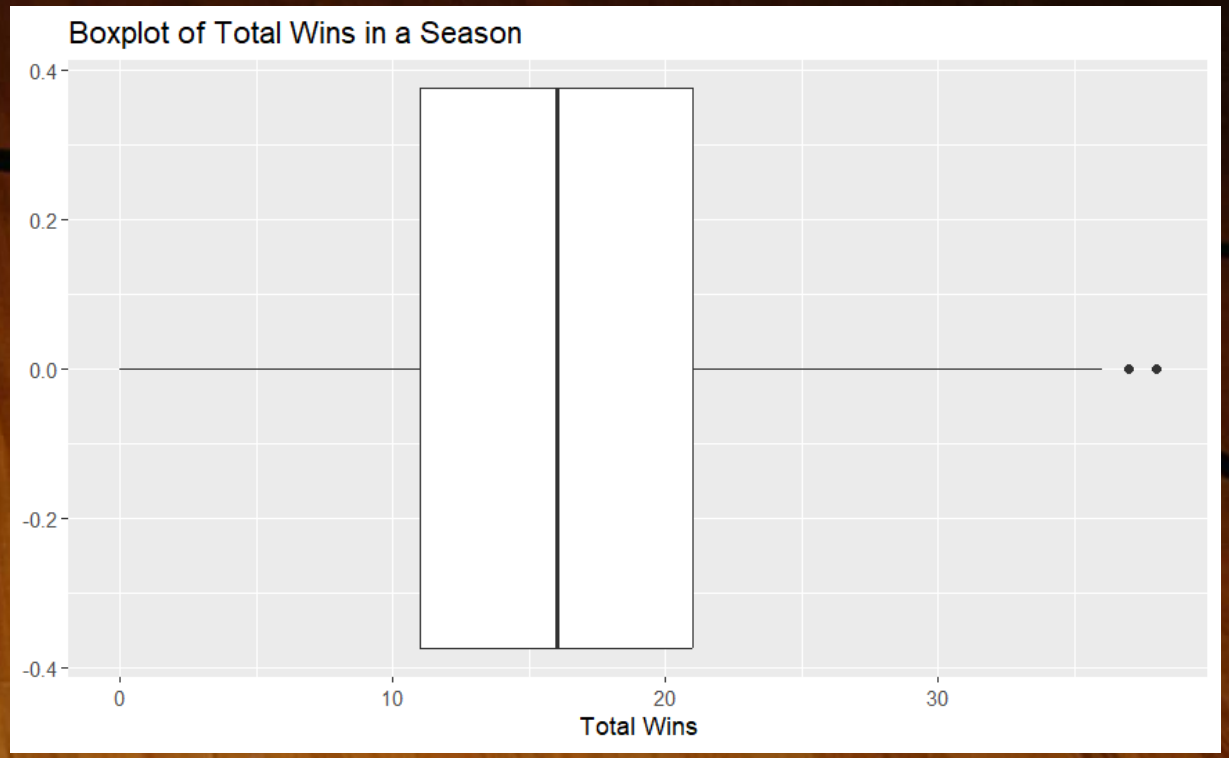
3PTM = Three Point Goals Made

FGA = Field Goals Attempted (2PT and 3PT)

- TOR/TORD (Turnover/Steal Rate)
- ORB/DRB (Offensive Rebound Rate/Rate Allowed)
- FTR/FTRD (Free Throw Rate/Rate Allowed)
- 2P_O/3P_O (2/3 point shooting percentage)
- 2P_D/3P_D (2/3 point shooting percentage allowed)
- ADJ_T (Adjusted Tempo/Possessions per 40 minutes)
- WAB (Wins Above Bubble)
- CONF (Conference a school plays in)

Response Variable – Total Wins in a Season



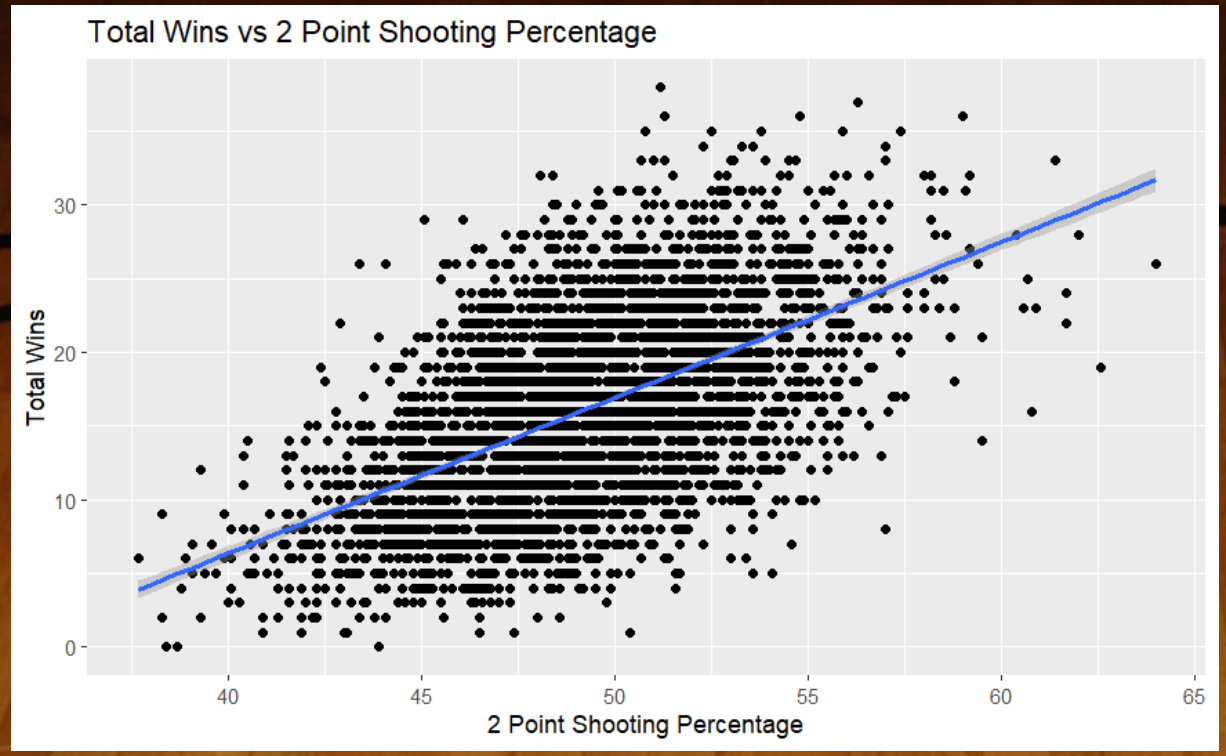


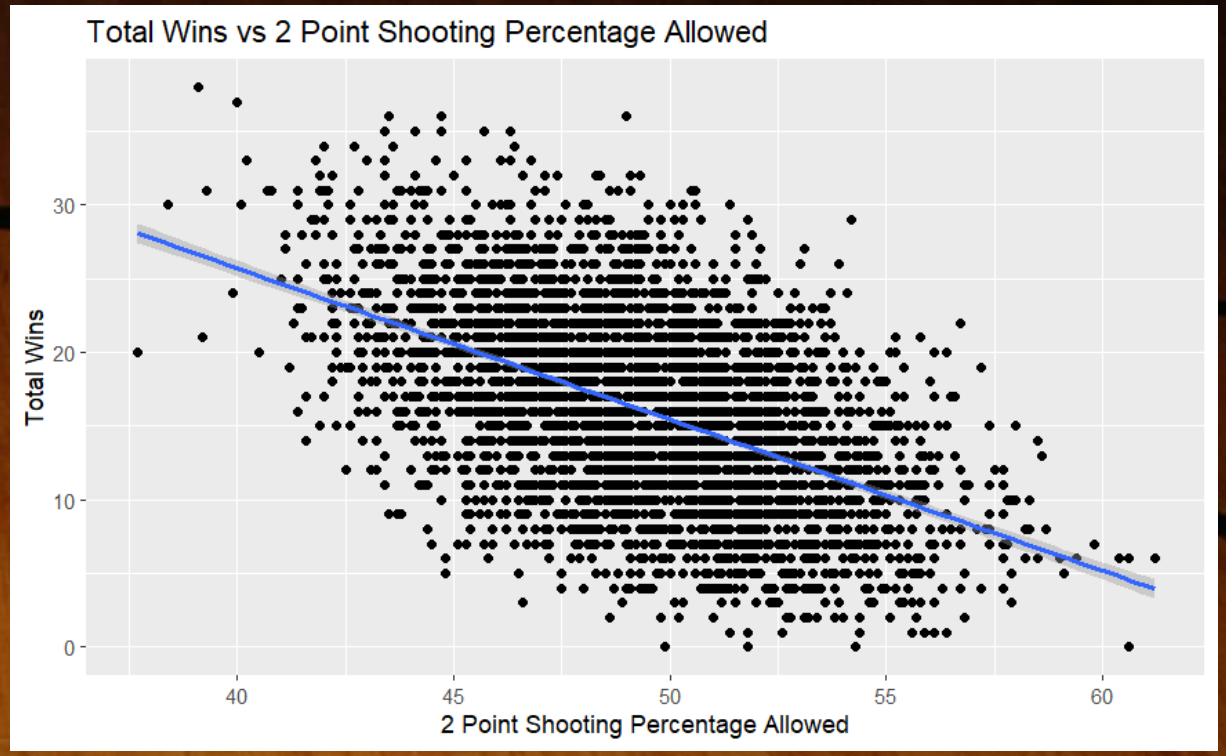
Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
0	11	16	16.08	21	38

Top Variables in Predicting Total Wins

- 2P_O (2 Point Shooting Percentage) → R^2 = .3074
- 2P_D (2 Point Shooting Percentage Allowed) → R^2 = .2693
- 3P_D (3 Point Shooting Percentage Allowed) → R^2 = .2010
- TOR (Turnover Rate) → R^2 = .1943

Marginal Relationships with Total Wins

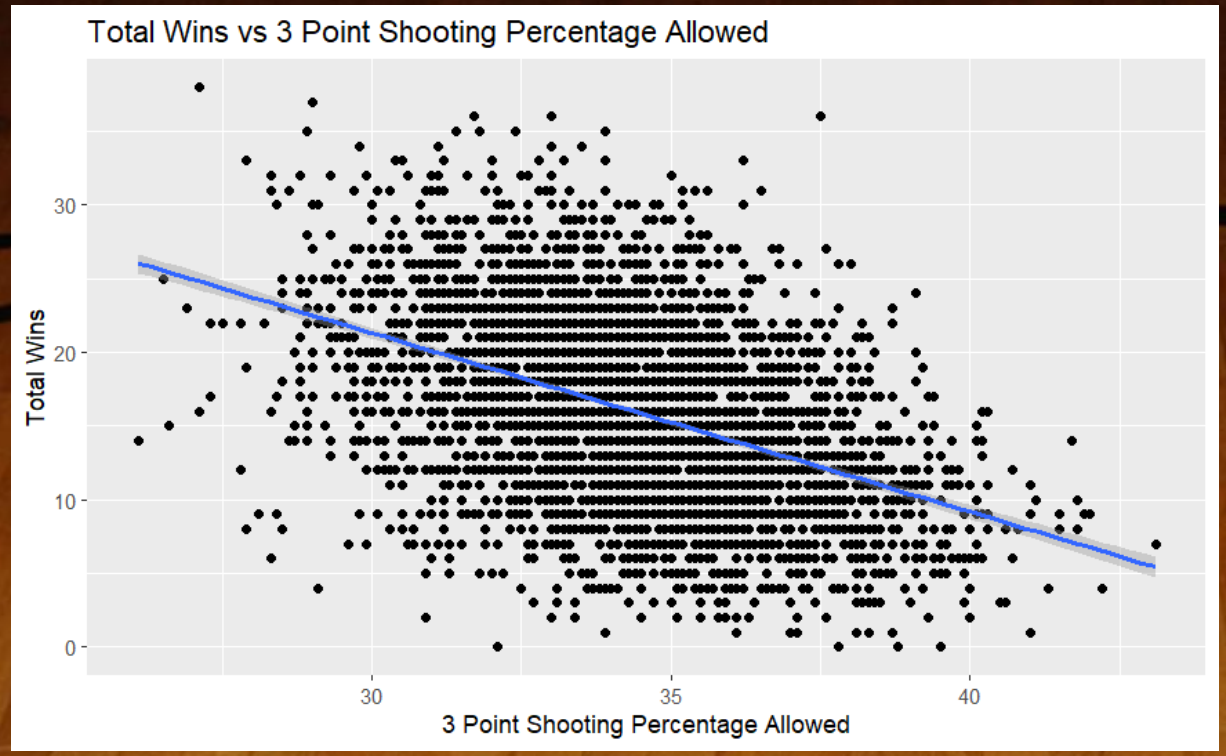


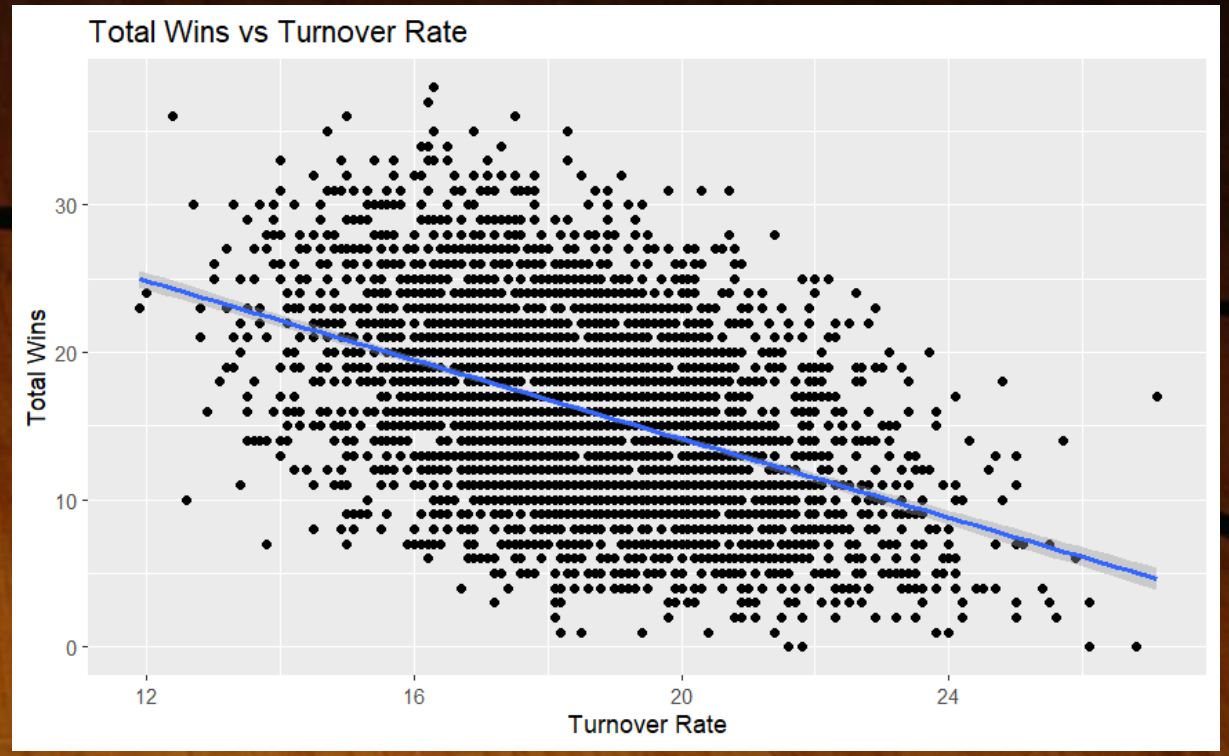


$$\widehat{Total\ Wins} = -35.9 + 1.056 * (2\ Point\ Shooting\ Percentage)$$

$$\widehat{Total\ Wins} = 66.75 - 1.026 * (2\ Point\ Shooting\ Percentage\ Allowed)$$

Marginal Relationships with Total Wins

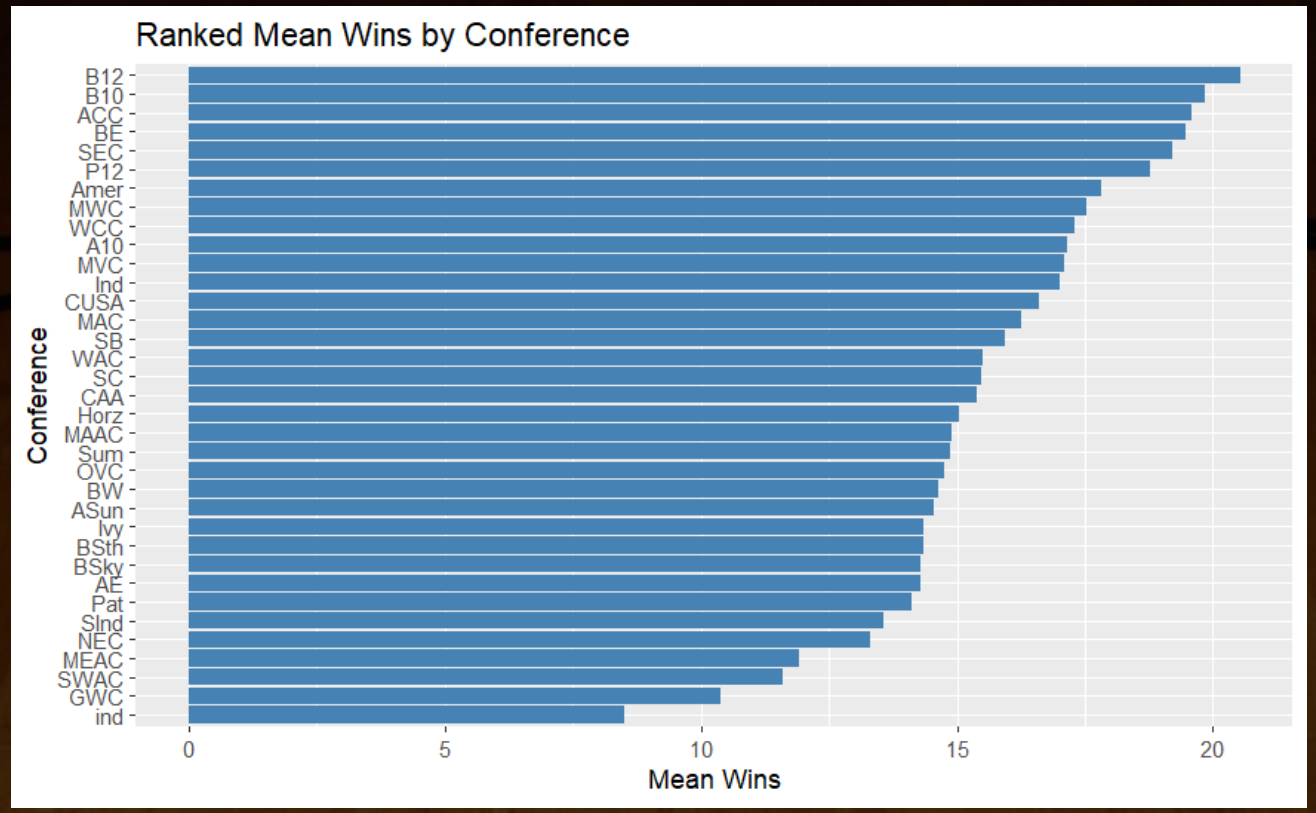




$\widehat{Total\ Wins} = 57.68 - 1.212 * (3\ Point\ Shooting\ Percentage\ Allowed)$

$\widehat{Total\ Wins} = 40.88 - 1.337 * (Turnover\ Rate)$

Marginal Relationships with Total Wins



Conference

- Adjusted R^2 = .3074

Conferences with Positive Win Coefficients:

- ACC
- American*
- Big 10
- Big 12
- Big East
- Mountain West*
- PAC 12
- SEC
- WCC*

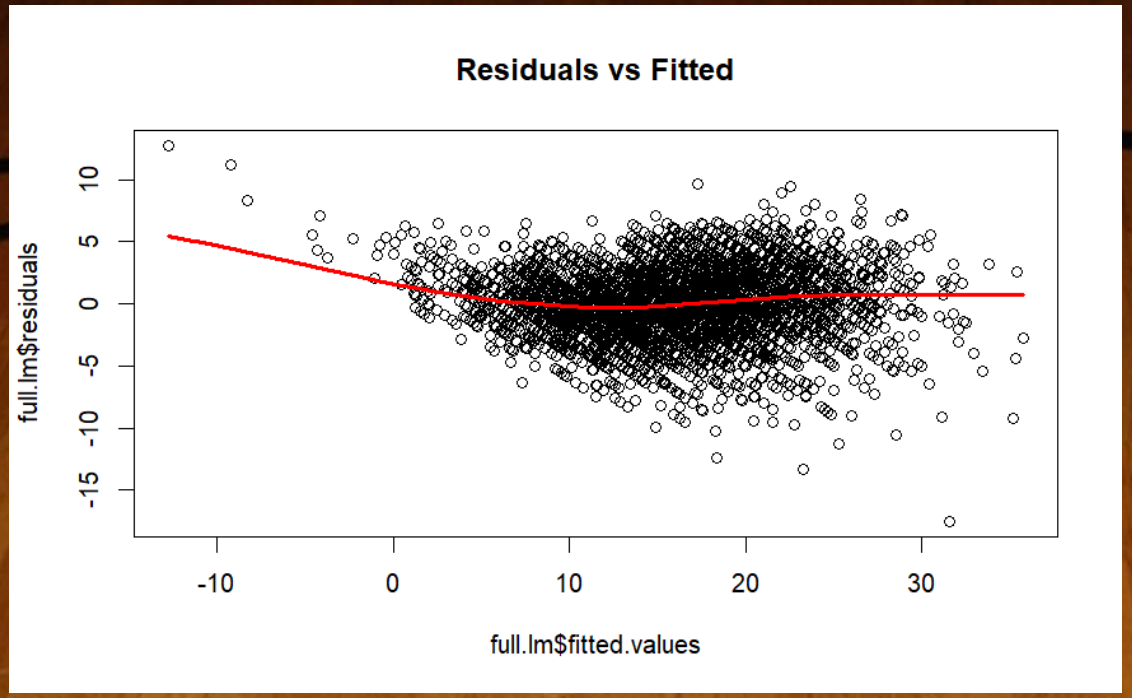
Multiple Linear Regression: Model 1

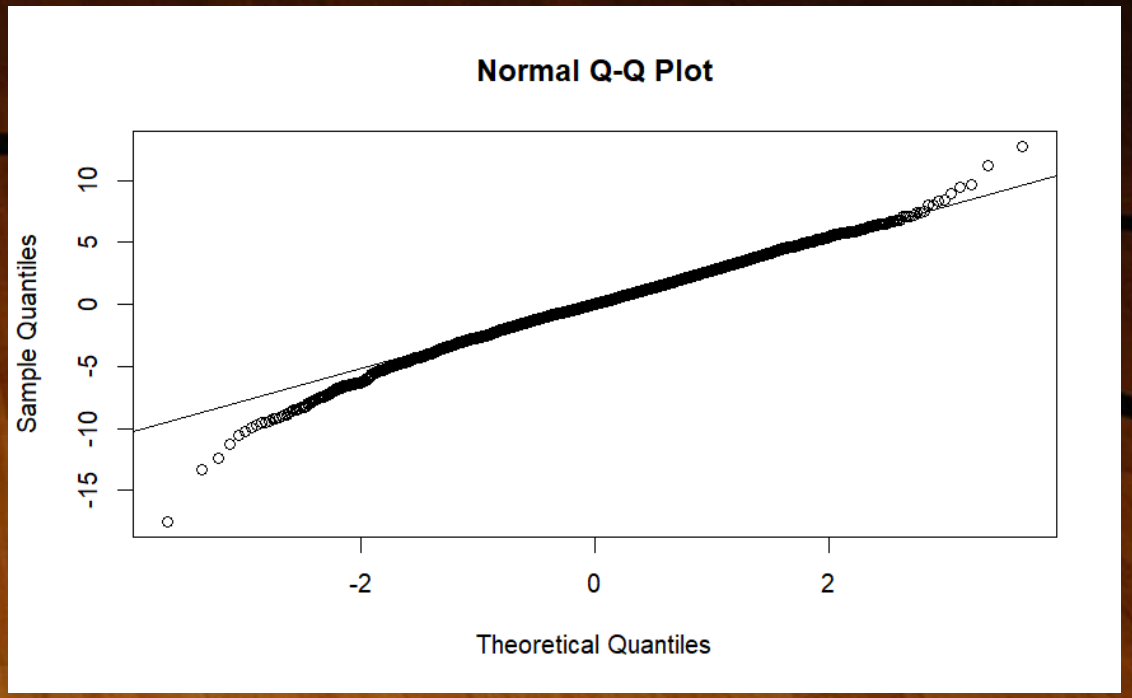
$$\widehat{Total\ Wins} = 10.586 - 0.912\,(TOR) + 0.734\,(TORD) + 0.379\,(ORB) - 0.281\,(DRB) + 0.142\,(FTR) - 0.130\,(FTRD) + 0.548\,(2P_O) - 0.499\,(2P_D) + 0.537\,(3P_O) - 0.535\,(3P_D) + 0.050\,(ADJ_T)$$

Adjusted R^2 = .8099

Global Hypothesis Test: p-value < $2.2 * 10^{-16}$

Multiple Linear Regression: Model 1

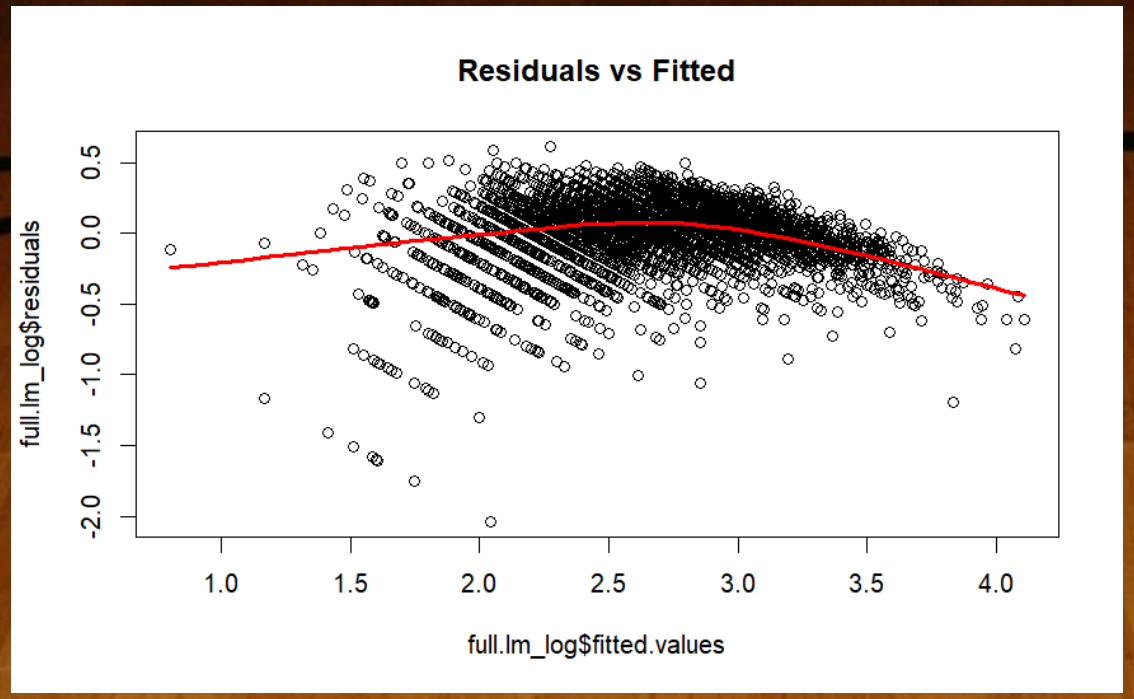


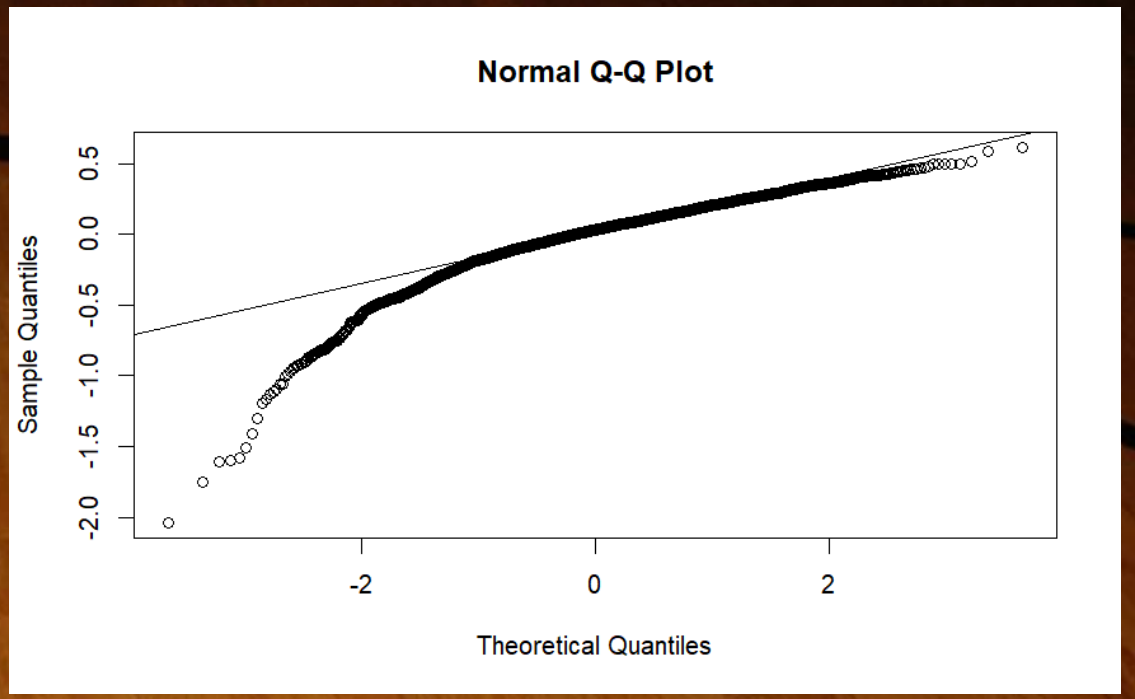


Studentized Breusch - Pagan Test:
p-value = $4.029 * 10^{-16}$

Shapiro - Wilks Normality Test:
p-value = $1.998 * 10^{-14}$

Multiple Linear Regression: Log Model

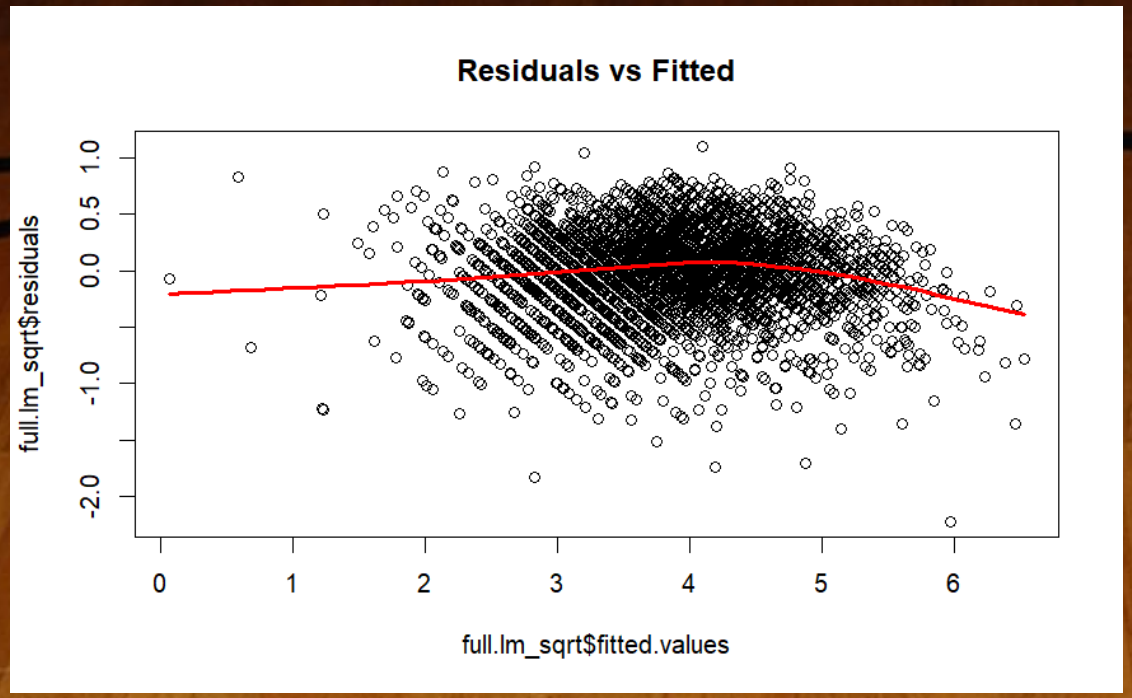


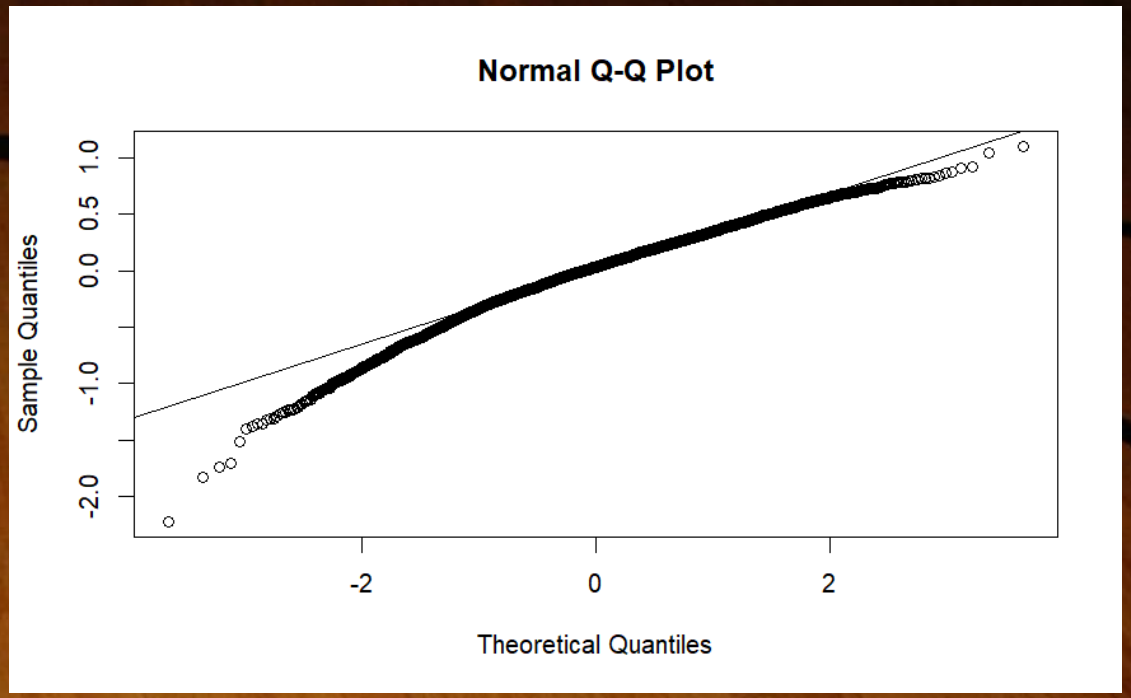


Studentized Breusch - Pagan Test:
p-value < $2.2 * 10^{-16}$

Shapiro - Wilks Normality Test:
p-value < $2.2 * 10^{-16}$

Multiple Linear Regression: Square Root Model

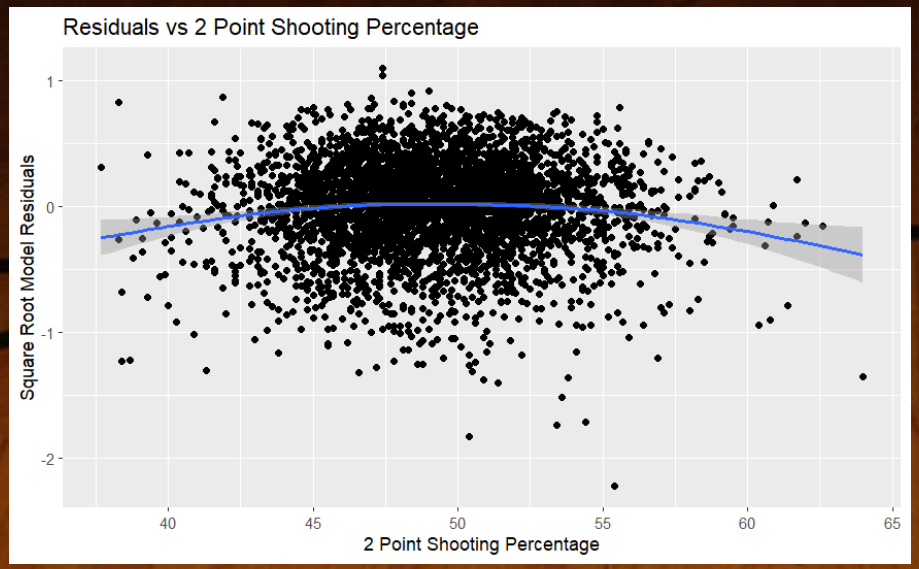


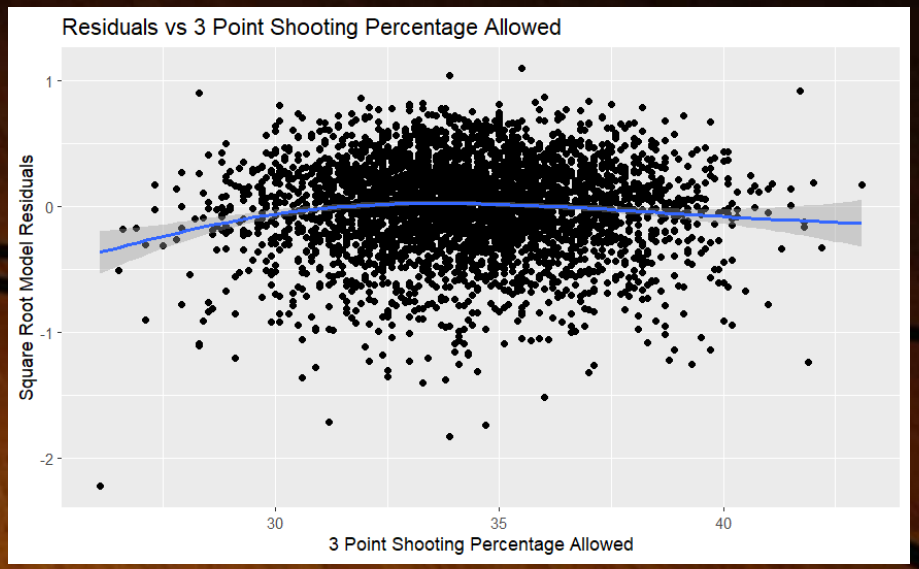


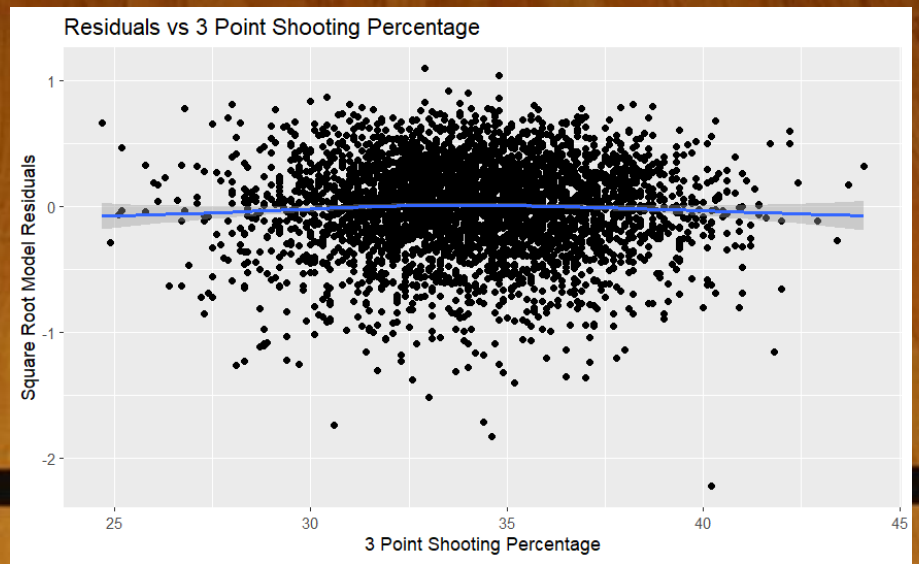
Studentized Breusch - Pagan Test:
p-value = $2.18 * 10^{-8}$

Shapiro - Wilks Normality Test:
p-value < $2.2 * 10^{-16}$

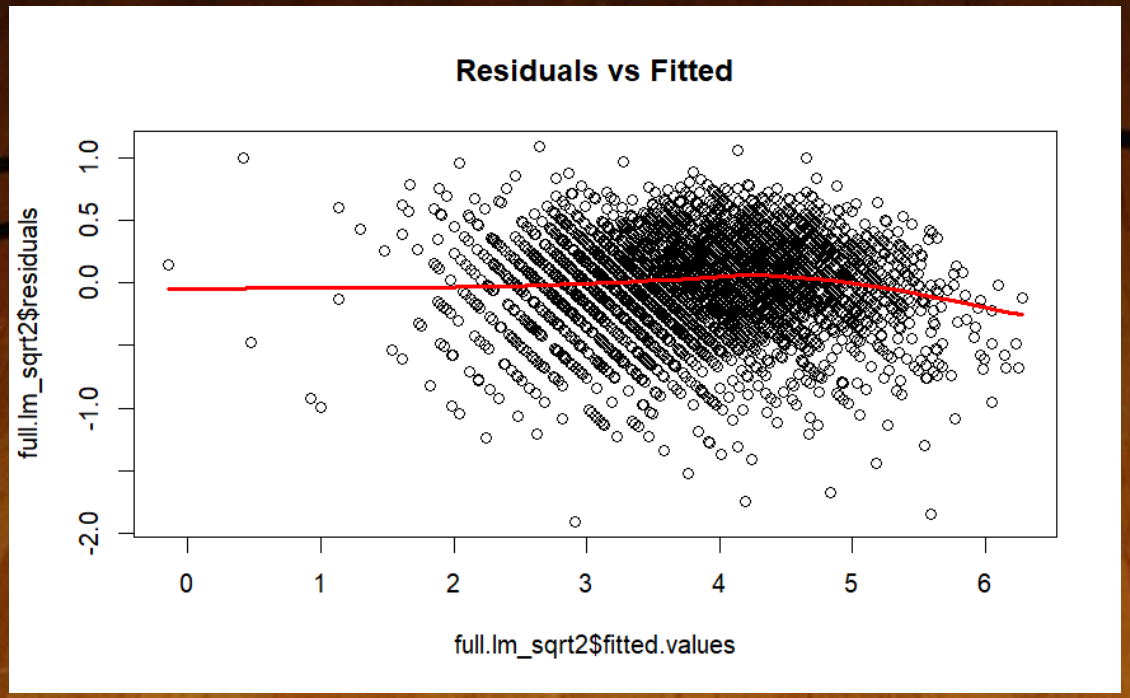
Multiple Linear Regression: Square Root Model

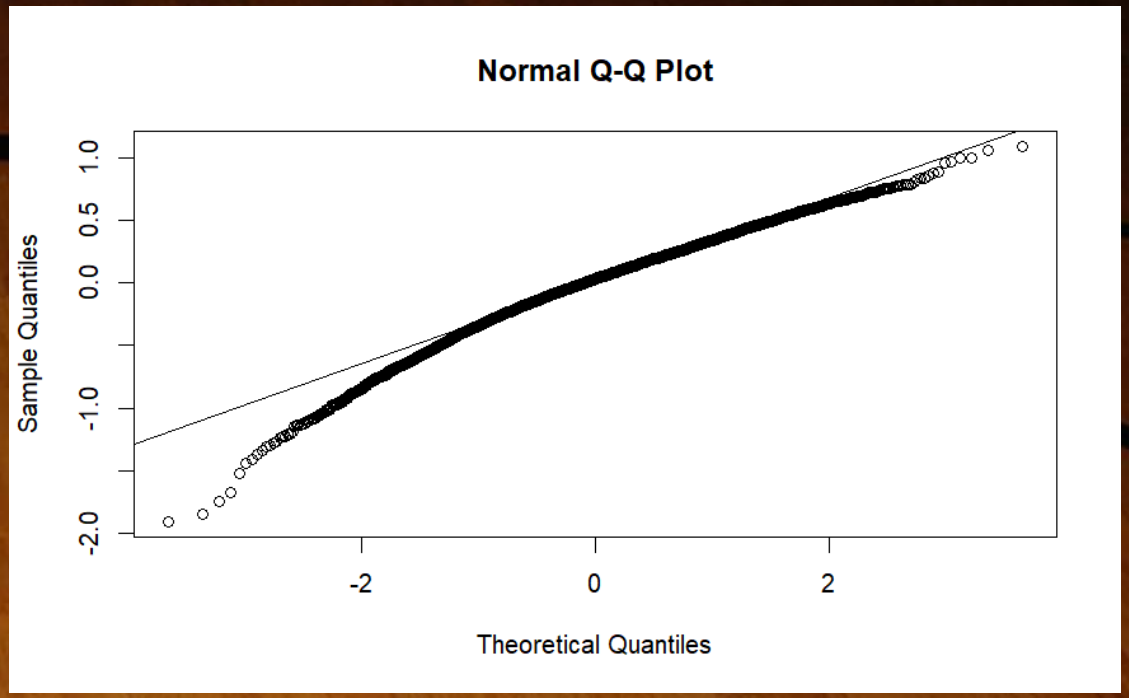






Multiple Linear Regression: Updated Square Root Model





Studentized Breusch - Pagan Test:
p-value = $9.88 * 10^{-10}$

Shapiro - Wilks Normality Test:
p-value < $2.2 * 10^{-16}$

Multiple Linear Regression: Final Model

$$\sqrt{(\widehat{Total\ Wins})} = 3.588 - 0.1298\,(TOR) + 0.097(TORD) + 0.049\,(ORB) - 0.042\,(DRB) + 0.0198\,(FTR) - 0.0171\,(FTRD) + 0.073\,(2P_O) - 0.0686\,(2P_D) + 0.0708\,(3P_O) - 0.0694\,(3P_D) + 0.005\,(ADJ_T) + I(CONF)$$

Adjusted R^2 = .8156

Global Hypothesis Test: p-value < $2.2 * 10^{-16}$

Multiple Linear Regression: Final Model

Coefficient Table for Conference						
A10: Reference	ACC: -0.0435	American East: -0.0023	American: -0.066	Atlantic Sun: 0.162	Big 10: -0.0288	Big 12: -0.0167
Big East: -0.0055	Big Sky: 0.0504	Big South: 0.0703	Big West: 0.0158	CAA: 0.0314	Conference USA: 0.0935	GWC: -0.1013
Horizon: 0.0539	Independent: 0.2148	Ivy: -0.1163	MAAC: 0.0248	MAC: 0.0821	MEAC: 0.1331	MVC: 0.0821
MWC: 0.0059	NEC: 0.0681	OVC: 0.0864	Pac 12: 0.0079	Patriot: -0.004	Sun Belt: 0.1009	SOCON: 0.105
SEC: 0.0011	Southland: 0.1365	Summit: 0.0435	SWAC: 0.1671	WAC: 0.1109	WCC: -0.0132	

Prediction: 2025 Villanova

- TOR = 15.3
- TORD = 15.8
- ORB = 30.3
- DRB = 27.7
- FTR = 29.2
- FTRD = 30.4
- 2P% = 51.3
- 2P%D = 50.5
- 3P% = 39.0
- 3P%D = 34.5
- ADJ T = 63.5

$\widehat{Total\ Villanova\ Wins}$ = 20.344

Actual Villanova Wins = 21

Residual: $e_{Villanova} = 0.6564$

Limitations and Further Considerations

- Only had team statistics and no individual player data
- Lots of multicollinearity between variables
- Lack of normality
- Observations are not fully independent because the teams will play each other and impact each other's scores
- Players may stay on the same team for multiple years, so the year-to-year data for a particular team may not be independent
  - Lots of turnover in college basketball now, so independence may not be incredibly unreasonable

Limitations and Further Considerations

- Further considerations:
  - Would like to compare yearly models to see what variables remain important to predicting wins and which change (get a sense of how the game is changing)
  - Do Ridge Regression or LASSO models better limit non-normality or non-constant variance?
  - How does this data do in predicting if a team will make the NCAA Tournament, and if so, how far they will go
  - How much does conference matter in predicting how far a team will advance in the NCAA Tournament?

Conclusion

- Of all the standard statistics, turnover and steal rate changed the expected number of wins the most
- Tempo has a very minimal impact on total expected wins
- While conference has a small impact on total expected wins, being in a power conference is associated with a decrease in total expected wins

Thank You

Any questions?