



Linearly Modeling the Premier League

Marie Meredith

[don't include in final product, just notes]

R Stuff

Not that elaborate, very basic, but it's ok

GOAL

We want to find a the best **linear model*** that uses one or more of our eight chosen variables to predict final rank for a team in the Premier League.

*we understand that it's probably not the case that the best model is a linear one, but for simplicity this is what we're doing



CHOSEN STATISTICS



Goals For

How many goals a team actually scored in the season



Goals Against

How many goals a team gives up in the season



Yellow Cards

Total yellow cards received during the season



Red Cards

Total red cards received during the season



Possession

Percentage of successful passes attempted (over the season)



Age

Average team age, weighted by minutes played



Wages

Annual team wages, millions of British Pounds



XG

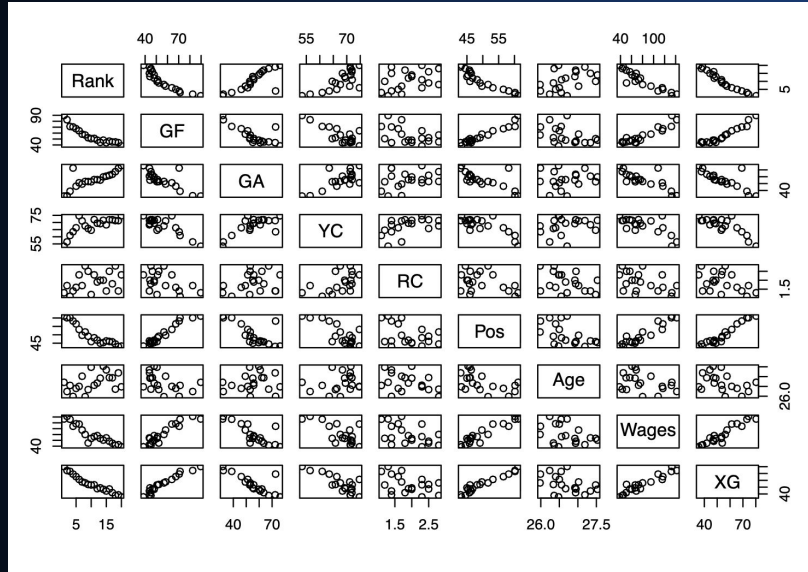
Expected amount of goals a team will score in the season

DATA

- All stats are from FBREF, and XG is from FOTMOB
- Past 10 seasons of Premier League (2015-2016 through 2024-2025)
 - XG was only available for past 5 seasons however
- Premier League Setup:
 - 38 games per season
 - 20 total teams (Rank is 1-20)



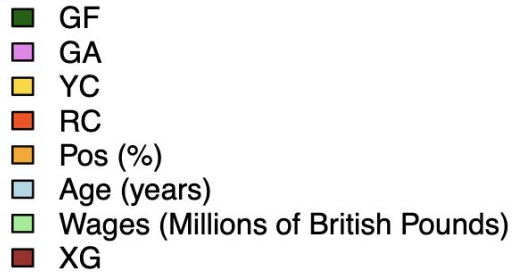
Aggregated Data



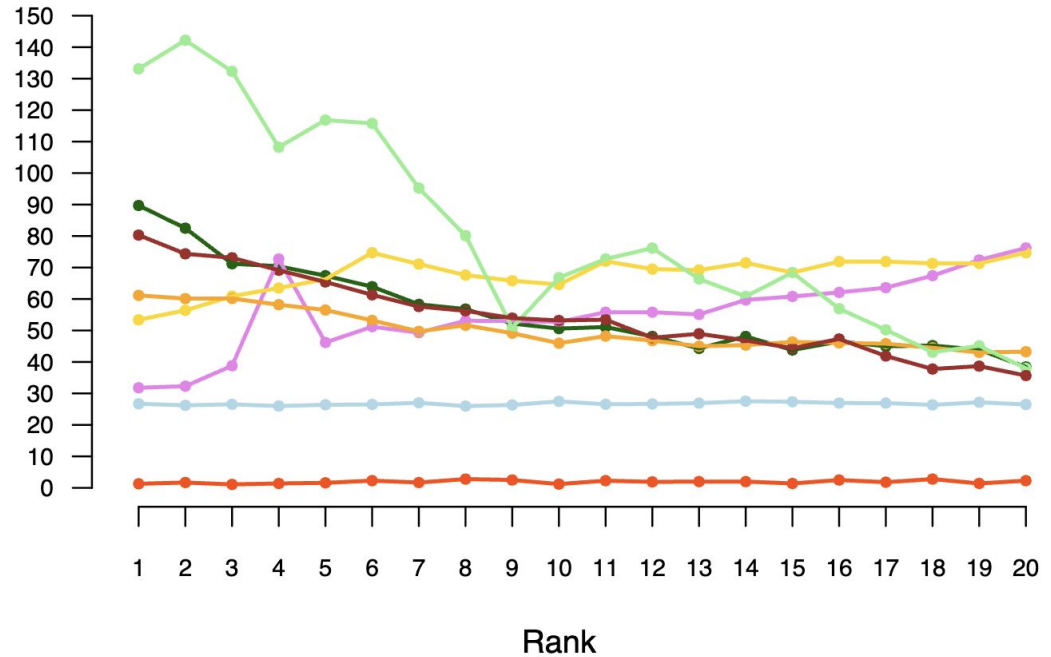
##	Rank	GF	GA	YC	RC	Pos	Age	Wages	XG
## 1	1	89.7	31.8	53.4	1.3	61.15	26.71	133.12810	80.30
## 2	2	82.5	32.3	56.4	1.7	60.13	26.24	142.19550	74.36
## 3	3	71.2	38.8	60.9	1.1	60.17	26.55	132.32040	73.10
## 4	4	70.4	72.7	63.5	1.4	58.18	25.99	108.23490	69.14
## 5	5	67.4	46.2	66.3	1.6	56.50	26.40	116.84980	65.42
## 6	6	63.9	51.2	74.7	2.3	53.19	26.50	115.78160	61.32
## 7	7	58.3	49.3	71.1	1.7	49.72	27.02	95.23860	57.60
## 8	8	56.8	53.1	67.6	2.8	51.67	25.98	80.11820	56.24
## 9	9	52.2	53.0	65.8	2.5	49.13	26.34	50.87384	53.94
## 10	10	50.6	52.7	64.6	1.2	45.96	27.48	66.77740	53.18
## 11	11	51.1	55.8	72.0	2.3	48.22	26.56	72.69650	53.44
## 12	12	48.1	55.8	69.5	1.9	46.79	26.67	76.16390	47.70
## 13	13	44.3	55.1	69.2	2.0	45.00	26.92	66.37968	48.94
## 14	14	48.1	59.7	71.5	2.0	45.37	27.53	60.80670	46.84
## 15	15	43.8	60.8	68.4	1.4	46.35	27.34	68.40810	44.36
## 16	16	46.5	62.1	71.9	2.5	46.04	26.95	56.92319	47.30
## 17	17	45.0	63.6	71.9	1.8	45.80	26.94	50.20500	41.90
## 18	18	45.2	67.4	71.3	2.8	44.51	26.36	43.11058	37.76
## 19	19	43.9	72.4	71.3	1.4	43.04	27.19	45.10492	38.68
## 20	20	38.4	76.2	74.7	2.3	43.26	26.48	37.99732	35.68

Across each Rank, we took the mean of each stat from the 10 (or 5, for XG) seasons of data.

Aggregated Data



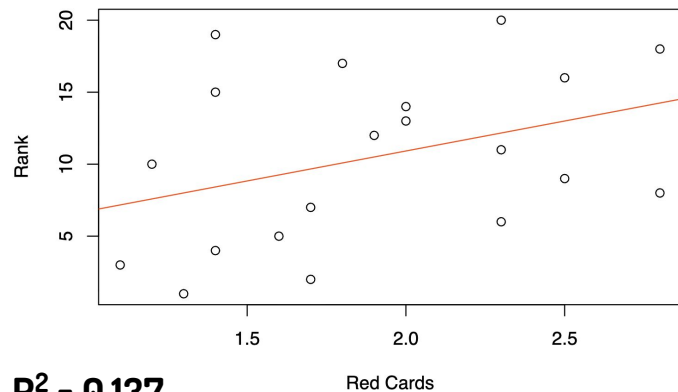
Graph of Chosen Stats



Single Variable Models Overview

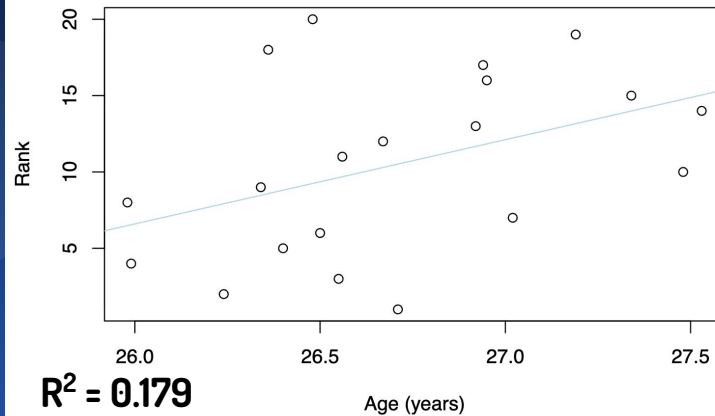


Scatterplot of Red Cards



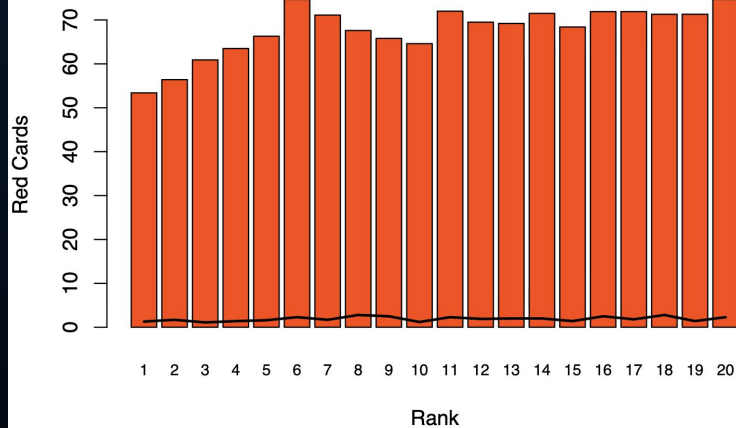
$R^2 = 0.137$

Scatterplot of Age

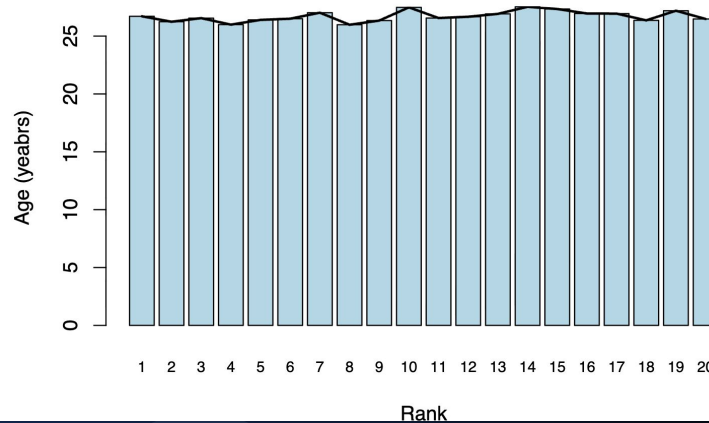


$R^2 = 0.179$

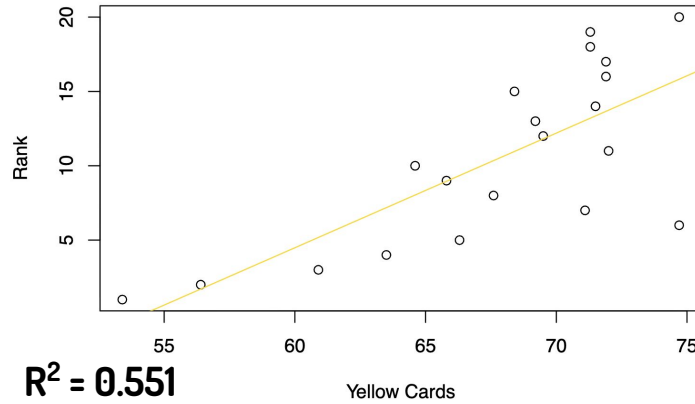
Bargraph of Red Cards



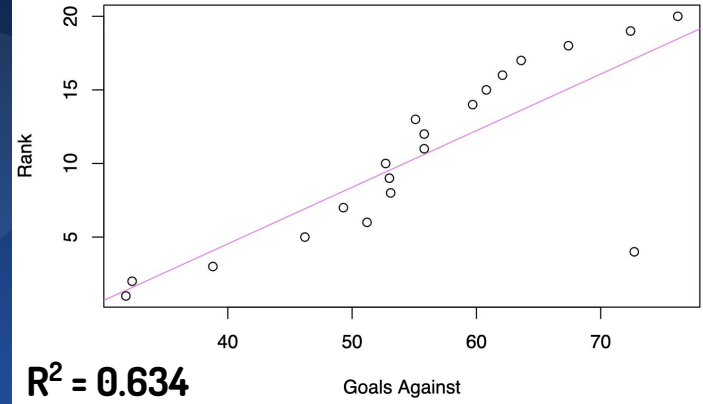
Bargraph of Age



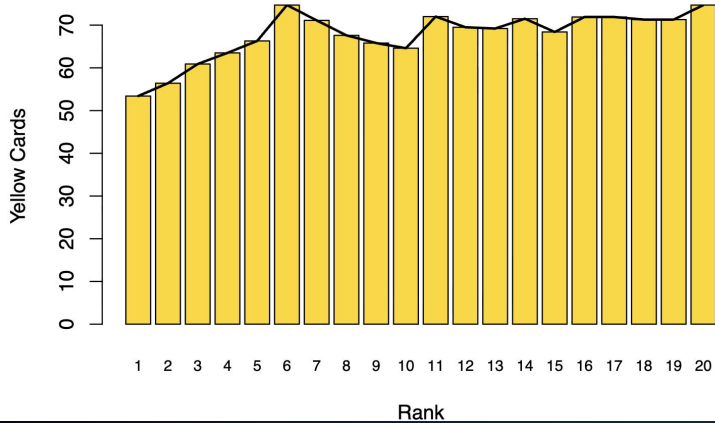
Scatterplot of Yellow Cards



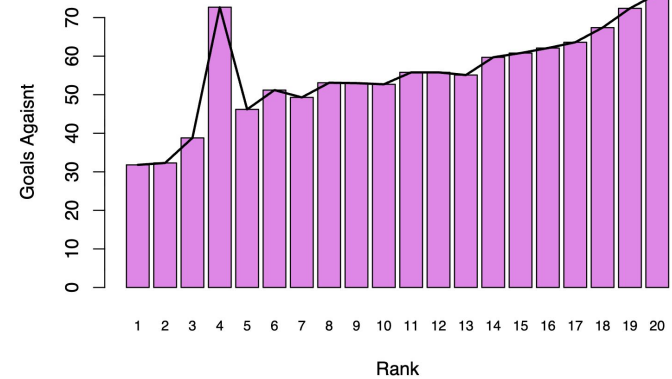
Scatterplot of Goals Against



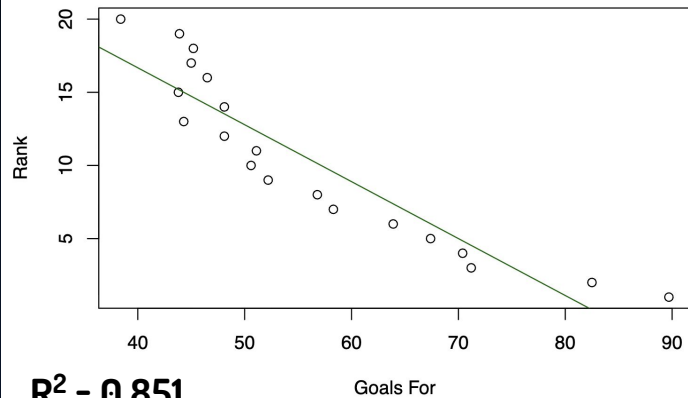
Bargraph of Yellow Cards



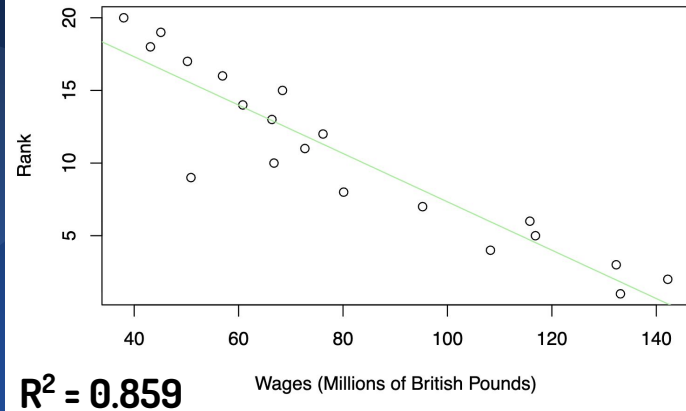
Bargraph of Goals Against



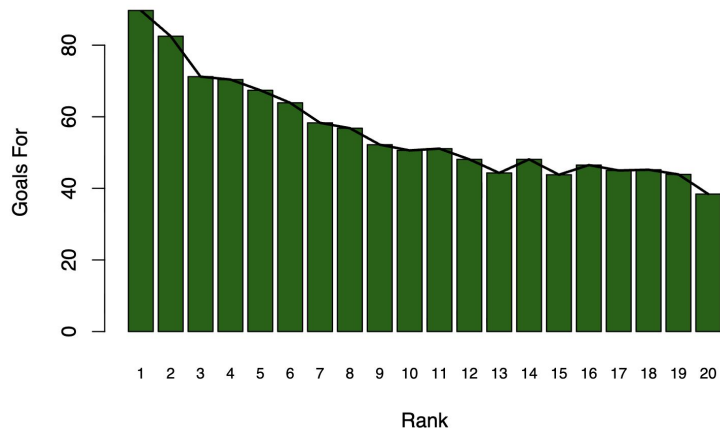
Scatterplot of Goals For



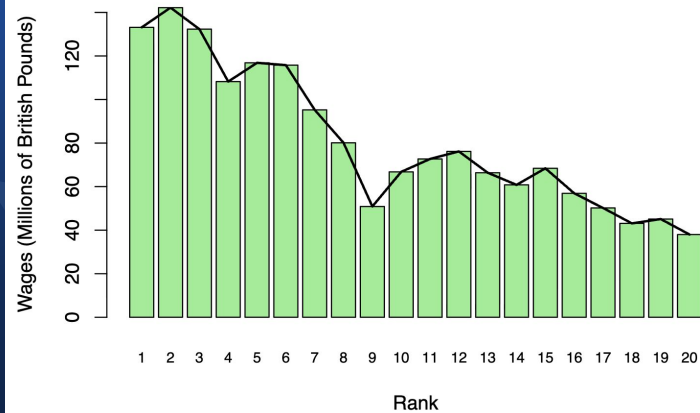
Scatterplot of Wages



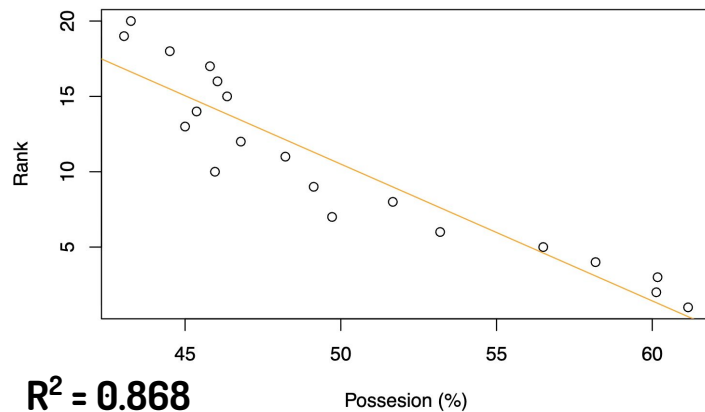
Bargraph of Goals For



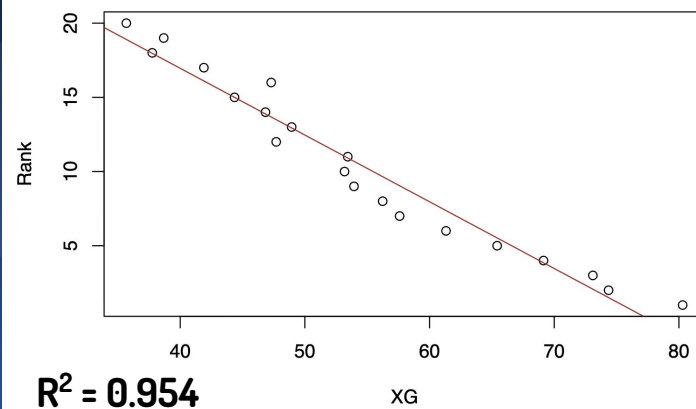
Bargraph of Wages



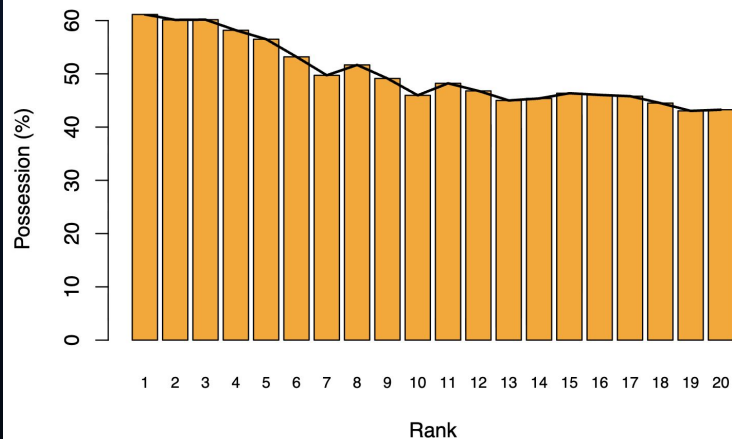
Scatterplot of Possession



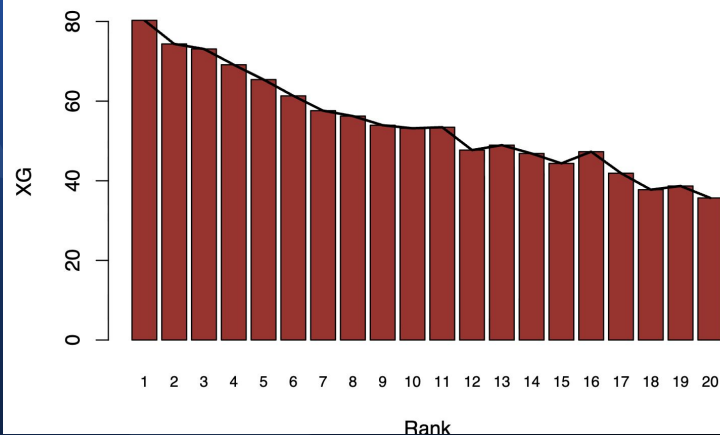
Scatterplot of XG



Bargraph of Possession



Bargraph of XG



Finding Best Model



ANALYSIS

- used program in R to find the best model using the data we have
- Best: XG-only model
 - Close 2nd: XG + GF model
 - Note: correlation is high
 - Scoring is important!

BEST LINEAR MODELS:

Rank = 34.9522309 + -0.4498203 * XG

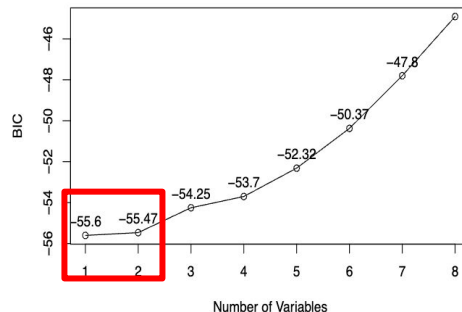
Rank = 35.1298371 + 0.1262715 * GF + -0.5828665 * XG

```
cor(Predictors.by.Rank$XG, Predictors.by.Rank$GF)
```

```
## [1] 0.9651953
```

However, XG and GF are pretty correlated (makes sense).

```
## Subset selection object
## Call: regsubsets.formula(Rank ~ ., data = Predictors.by.Rank, nbest = 1,
##       nvmax = 8)
## 8 Variables (and intercept)
##      Forced in Forced out
## GF      FALSE      FALSE
## GA      FALSE      FALSE
## YC      FALSE      FALSE
## RC      FALSE      FALSE
## Pos     FALSE      FALSE
## Age     FALSE      FALSE
## Wages   FALSE      FALSE
## XG      FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##      GF  GA  YC  RC  Pos Age Wages XG
## 1 ( 1) " " " " " " " " " " " "*"
## 2 ( 1) "*" " " " " " " " " " " "*"
## 3 ( 1) " " " " " " " " "*" "*" " "*"
## 4 ( 1) " " " " " " " " "*" "*" " "*"
## 5 ( 1) "*" " " " " " " " "*" "*" " "*"
## 6 ( 1) " " "*" "*" "*" "*" "*" " " "*"
## 7 ( 1) "*" "*" "*" "*" "*" "*" "*" " " "*"
## 8 ( 1) "*" "*" "*" "*" "*" "*" "*" "*" " "
```

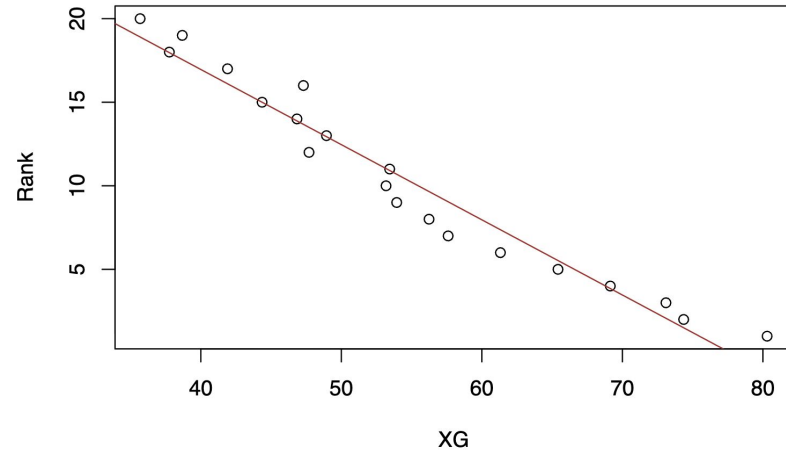


XG Model

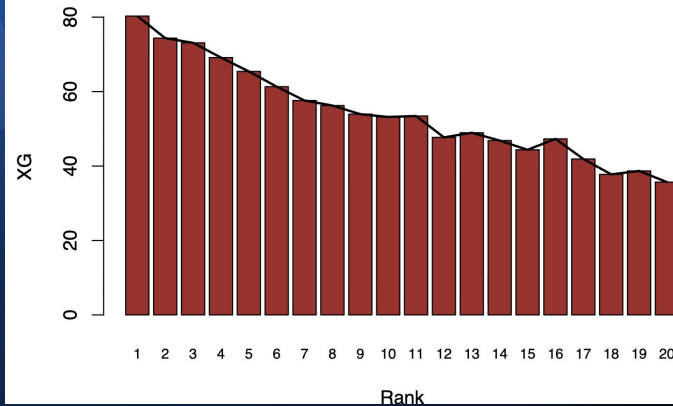
```
##
## Call:
## lm(formula = Rank ~ XG, data = Predictors.by.Rank)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.04258 -1.11540  0.07407  0.90384  2.32427
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.95223    1.29834   26.92 5.41e-16 ***
## XG          -0.44982    0.02327  -19.33 1.74e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.303 on 18 degrees of freedom
## Multiple R-squared:  0.954, Adjusted R-squared:  0.9515
## F-statistic: 373.5 on 1 and 18 DF, p-value: 1.738e-13

## Analysis of Variance Table
##
## Response: Rank
##      Df Sum Sq Mean Sq F value    Pr(>F)
## XG      1  634.43   634.43   373.52 1.738e-13 ***
## Residuals 18    30.57    1.70
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

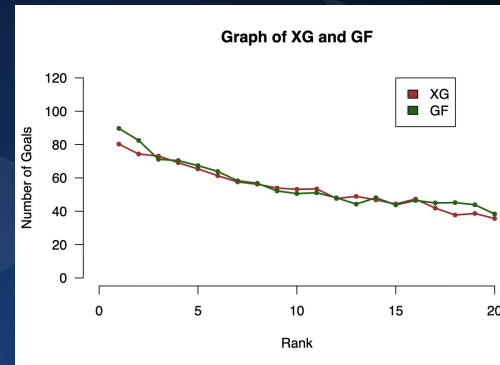
Scatterplot of XG



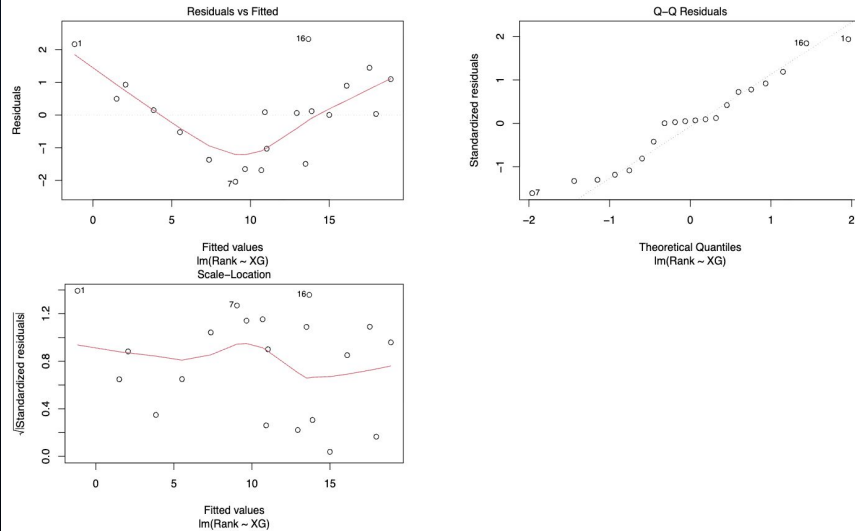
Bargraph of XG



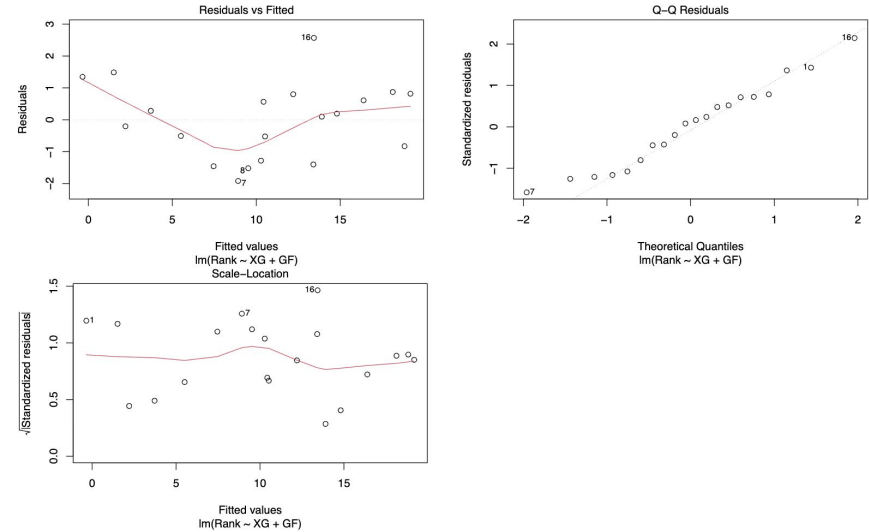
Are these models good?



XG-model



XG and GF Model



WHAT IS XG IN SOCCER?

- Not a “counting stat” like GF or YC
 - Much more complex
- XG as a stat itself is a prediction model
 - Our findings articulate this
- 2012; Opta; ~ one million shots studied
- Factors: distance, angels, goalie position, positions of other players, shot type, play pattern, previous actions
- Looked at by coaches, players, goalies, recruiters



CONCLUSIONS & REMARKS

- While these are the best *linear* models, the fit seems to be better if nonlinear
 - More analysis can be done to find an even better prediction model
- This is only for the eight stats we chose
- Have changes in the Premier League affected these stats at all?
 - e.g. VAR introduced in 2019/2020 season
- Would the best model even be right?
 - Injuries? Sickness? Outside factors?
- Much, much more we can do with this data
 - This is just a quick glance!

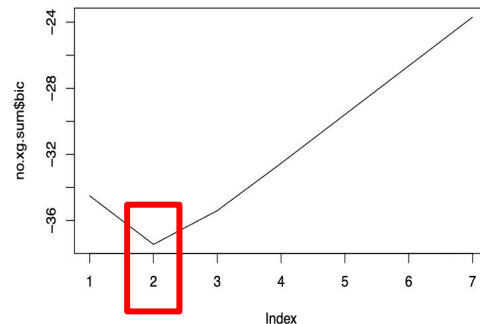
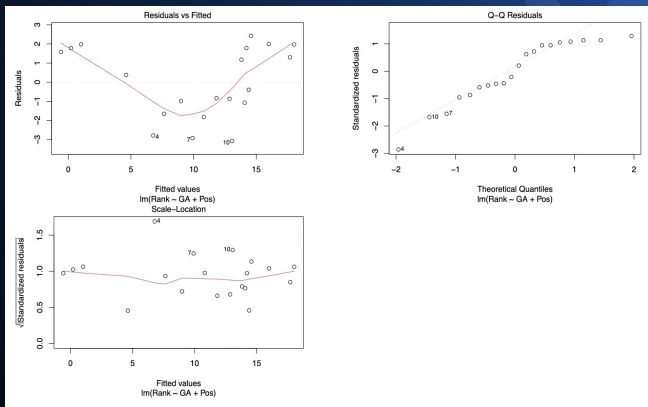


BONUS: EXCLUDING XG

What predictor(s) are used in the best linear model using all our stats, excluding XG?

- 2-variable model using
 - **Goals Against**
 - **Possession**

The model is: $\text{Rank} = 39.6006933 + 0.1275689 * \text{GA} + -0.7234851 * \text{Pos}$



```
## Subset selection object
## Call: regsubsets.formula(Rank ~ GF + GA + YC + RC + Pos + Age + Wages,
## data = Predictors.by.Rank, nbest = 1, nvmax = 7)
## 7 Variables (and intercept)
##      Forced in Forced out
## GF      FALSE      FALSE
## GA      FALSE      FALSE
## YC      FALSE      FALSE
## RC      FALSE      FALSE
## Pos     FALSE      FALSE
## Age     FALSE      FALSE
## Wages   FALSE      FALSE
## 1 subsets of each size up to 7
## Selection Algorithm: exhaustive
##      GF  GA  YC  RC  Pos Age Wages
## 1 ( 1) " " " " " " " " " " " "
## 2 ( 1) " " " " " " " " " " " "
## 3 ( 1) " " " " " " " " " " " "
## 4 ( 1) " " " " " " " " " " " "
## 5 ( 1) " " " " " " " " " " " "
## 6 ( 1) " " " " " " " " " " " "
## 7 ( 1) " " " " " " " " " " " "
```

Who Will Win 2025-2026 Season?

- Season isn't over, but our model uses data from a complete 38-game season
- To compensate, do

$$XG_{\text{current}} * 38 / \# \text{ matches played}$$

- Is this really accurate?
 - now we're predicting XG too, which is calculated continuously...

Squad	Poss
Manchester City	60.4
Liverpool	59.6
Chelsea	58.1
Arsenal	56.0
Aston Villa	53.9
Newcastle United	52.9
Brighton	52.8
Manchester Utd	52.8

#		PL	W	D	L	+/-	GD	PTS	Form	Next
1	Arsenal	34	22	7	5	64-26	+38	73	W W L L W	
2	Manchester City	33	21	7	5	66-29	+37	70	D D W W W	
3	Manchester United	34	17	10	7	60-46	+14	61	W D L W W	
4	Liverpool	34	17	7	10	57-44	+13	58	D L W W W	
5	Aston Villa	34	17	7	10	47-42	+5	58	L W D W L	
6	Brighton & Hove Albion	34	13	11	10	48-39	+9	50	W W W D W	
7	AFC Bournemouth	34	11	16	7	52-52	0	49	D D W W D	
8	Chelsea	34	13	9	12	53-45	+8	48	L L L L L	
9	Brentford	34	13	9	12	49-46	+3	48	D D D D L	
10	Fulham	34	14	6	14	44-46	-2	48	D W L D W	

#		PL	xG	xGA	xPTS
1	0 * Arsenal	34	57.4	25.6 +0.4	69 +4
2	0 * Manchester City	33	62.3	36.5 -7.5	63 +7
3	+1 * Liverpool	34	55.4	41.2 +2.8	56 +2
4	+4 * Chelsea	34	62.4	46.5 -1.5	56 -8
5	-2 * Manchester United	34	56.2	44.0 +2.0	54 +7
6	+3 * Brentford	34	54.7	45.7 +0.3	53 -5
7	-1 * Brighton & Hove Albion	34	50.9	43.4 -4.4	52 -2
8	+5 * Crystal Palace	33	51.3	42.3 -3.3	51 -8
9	-2 * AFC Bournemouth	34	55.8	51.3 +0.7	51 -2
10	+4 * Newcastle United	34	50.5	45.9 +4.1	50 -8
11	+4 * Leeds United	34	49.3	48.1 +2.9	49 -9
12	-7 * Aston Villa	34	42.8	48.3 -6.3	43 +15

Current Top 5 (+ Chelsea)

XG-Only Model

Arsenal → 6.09

Man City → 2.68

Man United → 6.70

Liverpool → 7.10

Aston Villa → 13.43

Chelsea* → 3.58

XG & GF Model

Arsenal → 6.77

Man City → 2.92

Man United → 6.99

Liverpool → 7.08

Aston Villa → 13.88

Chelsea* → 9.41

Bonus Model (GA & Pos)

Arsenal → 2.79

Man City → 0.16

Man United → 7.96

Liverpool → 2.75

Aston Villa → 6.59

Chelsea* → 3.98

*current top 5 for XG, but not in top 5 of standings

Some interesting
finds

Verdict?

Across our models: **Manchester City**, then **Arsenal**

- Checks out with current point standings & even Opta's model
 - Exact numbers may be off, but general idea is nearly there
- Interesting to note that 3rd model uses counting stats, not predictive stats, so lines up a bit more to reality

Opta Supercomputer - 2025-26 Premier League Projections

 Opta Analyst

POS	TEAM	P	W	D	L	GF	GA	GD	PTS	WINNER
1	 Arsenal	34	22	7	5	64	26	+38	73	70.86%
2	 Man City	33	21	7	5	66	29	+37	70	29.14%
3	 Man Utd	34	17	10	7	60	46	+14	61	0.00%
4	 Liverpool	34	17	7	10	57	44	+13	58	0.00%
5	 Aston Villa	34	17	7	10	47	42	+5	58	0.00%
6	 Brighton	34	13	11	10	48	39	+9	50	0.00%
7	 Bournemouth	34	11	16	7	52	52	0	49	0.00%
8	 Chelsea	34	13	9	12	53	45	+8	48	0.00%
9	 Brentford	34	13	9	12	49	46	+3	48	0.00%
10	 Fulham	34	14	6	14	44	46	-2	48	0.00%



Thank You!

