2021

Data & Decisions Term Project

TEAM 2

DREW HELBERG

DYLAN MAY

JASON WOODSON

Table of Contents

I. Executive Summary

In this paper we are taking a deep dive into the advance statistics for the English Premier League. Specifically, we are looking at individual metrics during the 2020-201 season surrounding passing, shooting, possession, and shot creation and how these statistics relate to the team success metric of points per match (PPM) for each player. The rationale behind studying this is to see which parts of the game have the greatest impact on team success and which statistics are overvalued or potentially misleading when it comes to looking at a player's contribution to their team. Before running any analysis, we expected statistics such as Completed Passes, Touches in the Attacking Penalty Area, and Shot Creating Actions to have a positive correlation with PPM.

Our data for this research was pulled directly from https://fbref.com/en/. To begin, we looked at a handful of statistics from the Player Passing, Player Shooting, Player Possession, and Player Goal and Shot Creation tables. Eventually, we narrowed down the number of statistics we would consider for independent variables to the 15 listed below:

	Description	Explanation
Passing		
CmpPass	Passes Completed	Total passes completed
AttPass	Passes Attempted	Total passes attempted
xA	Expected Goals Assisted	Expected goals which follows a pass that assists a shot
KP	Key Passes	Passes that directly lead to a shot (assisted shots)
PPA	Passes into Penalty Area	Completed passes into the 18-yard box (not including set pieces)
CrsPA	Crosses into Penalty Area	Completed crosses into the 18-yard box (not including set pieces)
Shooting		
Sh	Shots	Total shots attempted (not including penalty kicks)
SoT	Shots on Target	Total shots taken that were on target (not including penalty kicks)
npxG	Non-Penalty Expected Goals	Expected goals (not including penalty kicks)
Possession		
Touches	Touches	Number of times a player touched the ball (receiving a pass, then dribbling, then sending a pass counts as one touch)
Touches Att 3rd	Touches in the Attacking 3rd	Number of times a player touched the ball in the attacking 3rd area (receiving a pass, then dribbling, then sending a pass counts as one touch)
Touches Att Pen	Touches in the Attacking PenaltyArea	Number of times a player touched the ball in the attacking penalty area (receiving a pass, then dribbling, then sending a pass counts as one touch)
Carries	Carries	Number of times the player controlled the ball with their feet
PrgDistCarry	Progressive Distance	Total distance, in yards, a player moved the ball while controlling it with their feet towards the opponent's goal
Shot Creation		
SCA	Shot Creating Actions	The two offensive actions directly leading to a shot, such as passes, dribbles and drawing fouls (a single player can receive credit for multiple actions and the shot-taker can also receive credit)
All stats converted to per 90 minutes (p90) for analysis purposes		

Next, we trimmed down the number of observations to 124 by only including players whose position was either attacking forward and/or midfield[1]. This excluded all goalkeepers and defensive forwards from our analysis. Our rationale behind this was that, by focusing on players within a specific position or area on the pitch, the model could better focus on the

[1] Diagram showing positions on a soccer field were included for analysis

metrics which have the greatest influence on the dependent variable. This was primarily due to specific positions requiring different on-the-field actions to yield a positive impact on the team. For example, measuring a goalkeeper's shots per 90-minute timeframe would be worthless because most goalkeepers will go an entire season without registering a shot, but their stats would still contribute positively to PPM. Finally, our last step before beginning analysis was to convert all statistics that we considered for regression to a per 90-minute basis. The rationale behind this was to standardize the metrics to compare players' stats more easily and in a more direct capacity. Since our model analyzes points garnered per match as opposed to over the course of a season, if we included seasonal metrics, our data could've potentially negatively influenced our model. Lastly, we felt using the 90-minute basis would help reduce the effect of inflated stats from players who played minimally each match compared to the stats of the players who typically played either the entire 90-minutes or a majority of the match.

Our regression analysis included 3 rounds of testing. We began by running a backwards, stepwise regression that included all 15 variables. In the end, we were left with 7 of the original 15 variables, with 3 of them being transformed in some way in our final model. The first variable transformed was Touches Att 3rd. A log transformation was performed on this variable in round 1 of testing to mitigate the non-linear tendency it was displaying with the dependent variable, PPM, when plotted against each other. The second and third transformation were performed on CmpPass and Touches in round 2. A reciprocal transformation was applied to both variables to reduce the endogeneity that both residual plots were showing. In round 3 of testing, all assumptions were satisfied, and our final model included the following variables:

Final Model
reciprocal(Touches (p90))
reciprocal(CmpPass (p90))
log(Touches Att 3rd (p90))
npxG (p90)
SoT (p90)
Carries (p90)
SCA (p90)

When evaluating the variables included in the final model, three distinct groupings of each metric type became apparent. Those were: Goal Impacting Actions, High Value Possessions, and Play Involvements. Goal Impacting includes actions like SoT, SCA, and npxG, due to these metrics evaluating actions dependent on goal scoring opportunities, shots, occurring. High Value Possessions include Touches Att 3rd and Carries as these metrics relate to actions either in dangerous spaces or utilized to get into dangerous spaces. Play Involvements include Touches and CmpPass, as both metrics directly show how often a player is involved in a game. These groupings highlight how the model valued players who were able to convert possessions into shots from dangerous positions when evaluating an attacking player's PPM. In terms of real-world applicability, we can conclude that the most important factor when evaluating an attacking player and whether they will have a positive influence on their team is in essence their efficiency. For teams looking to recruit attacking players to bolster their squad or make tactical changes to get the upper hand in an up incoming match, looking to maximize the efficiency of attacking play should be the primary concern to yield the greatest positive impact.

While our model effectively evaluated the factors with the greatest influence on PPM, there were a few limitations, specifically with multicollinearity between reciprocal(Touches (p90)) and reciprocal(CmpPass (p90)). Whilst building our model we were unable to find an effective method of satisfying the multicollinearity assumption between these variables without drastic changes to the model. Due to this we decided that the best solution was to leave in the highly correlated variables as the model's ability to evaluate PPM would not be damaged, however, in doing this we conceded the ability to evaluate the impact of reciprocal(Touches (p90)) and reciprocal(CmpPass (p90)) individually. Additionally, despite evaluating strictly attacking forwards and/or midfielders, players in these roles could be tasked with a variety of different duties, depending on the specific position they are played in (Left Wing, Striker, False 9, Attacking Midfielder, etc.). Due to this, a larger sample size could aid in better accounting for some of the nuance associated with differences in specific positions.

II. Modeling Steps

We began our analysis by comparing the results of a backward and forward regression with all 15 variables entered under each scenario. We achieved the same results using backward and forward stepwise regression: an overall R-squared value of .53, with all the same variables included[2]. Given these results, we decided to move forward with the results from our backwards stepwise regression. The initial independent variables included in the model are listed below:

Round 1
Touches (p90)
CmpPass (p90)
Touches Att 3rd (p90)
npxG (p90)
SoT (p90)
Carries (p90)
SCA (p90)

We began round 1 with testing the linearity of each independent variable against the dependent variable[3]. While each independent variable didn't show strong signs of linearity, we determined that all but one showed no signs of non-linearity. The variable that suggested non-linearity was Touches Att 3rd (p90). We began to test multiple

[2] Results of the backwards, forwards, and mixed stepwise regression
[3] Scatterplots of independent variables plotted against the dependent variable (PPM)

transformations on Touches Att 3rd (p90). The three transformations that worked the best were square-rooting, cube-rooting, and log with an r-squared of .013, .011, and .009 respectively[4]. Ultimately, we made the decision to pick the log transformation. Although this transformation had the lowest r-squared value, it had the least number of outliers in the scatterplot. Going into round 2 of testing we were left with the independent variables listed below:

Round 2
Touches (p90)
CmpPass (p90)
log(Touches Att 3rd (p90))
npxG (p90)
SoT (p90)
Carries (p90)
SCA (p90)

Since Touches Att 3rd (p90) was the only independent variable that violated the linearity assumption, we now knew that all independent variables satisfied that assumption was the log transformation was applied. We began round 2 by looking at the intercept term to check if the error was unbiased. Since the intercept in this round of testing was significant, we decided to keep it in the model, satisfying the assumption that the error term has a population mean of zero[5]. Next, we checked our model for any signs of endogeneity by looking at each residual plot. We found that the residual plot for both CmpPass (p90) and Touches (p90) violated this assumption by showing some sort of pattern in the plot[6]. To mitigate the endogeneity, we first tried to add an interaction term between CmpPass (p90) and Touches (p90). The interaction term did not prove to be significant. Next, we attempted both a log and reciprocal transformation on CmpPass (p90). We found the reciprocal transformation mitigated the endogeneity very well. Oddly enough, the reciprocal transformation worked best on Touches (p90) as well[7]. This was somewhat expected as these two variables had the most interaction with the dependent variable. Going into round 3 of testing we were left with the independent variables listed below:

Round 3
reciprocal(Touches (p90))
reciprocal(CmpPass (p90))
log(Touches Att 3rd (p90))
npxG (p90)
SoT (p90)
Carries (p90)
SCA (p90)

[4] Scatterplots of Touches Att 3rd (p90) in transformed states
[5] Parameter estimates, including the intercept term, of round 2 testing
[6] Residual plots for each independent variable included in round 2 of testing
[7] Residuals plots for CmpPass (p90) and Touches (p90) in transformed states

We began round 3 of testing by looking at the two new variables, reciprocal(Touches (p90)) and reciprocal(CmpPass (p90)), and their linear relationship with the dependent variable[8]. After plotting each, we did not see any signs of a non-linear relationship, so we proceeded with testing. In the latest model, the intercept was still significant and therefore included in our model. This satisfies the assumption that the error term has a population mean of zero[9]. Because we know the endogeneity was mitigated with the reciprocal transformation of Touches (p90) and CmpPass (p90), there was no need to test this assumption in round 3 as we are now confident that our model shows exogeneity in all the residual plots. Next, we looked to see that the error term was normally distributed. Although this assumption is optional, we still looked at the residual normal quartile plot and nothing of note stood out, so we moved on with our testing[10]. Next, we looked at the correlation estimates between each independent variable to see if there was any indication of multicollinearity. The only correlation estimate that was somewhat alarming was between reciprocal(Touches (p90)) and reciprocal(CmpPass (p90))[11]. These two variables had a correlation of .77. Naturally, we attempted to mitigate this by looking at the model with one of those variables taken out and we looked at the model with both variables taken out. In all three cases, the impact on the model was severe and the significance became much worse when either or both variables were removed. This led us to make the decision to keep both variables in the model, despite their high degree of multicollinearity. We will discuss later in this paper how these play into interpreting the results. We next looked at the autocorrelation of the model. To test this assumption, we ran a Durbin-Watson test in JMP. The Durbin-Watson test is on a scale of 0-4, with a 2 being perfect no autocorrelation. The result of the Durbin-Watson test on our model was a 1.78[12]. We were comfortable with this level of autocorrelation and moved on to test the final assumption. Lastly, we looked at our model to see if homoscedasticity was present in any of the residual plots. We concluded that all residual plots were heteroscedastic and settled on our final model below:

Final Model
reciprocal(Touches (p90))
reciprocal(CmpPass (p90))
log(Touches Att 3rd (p90))
npxG (p90)
SoT (p90)
Carries (p90)
SCA (p90)

[8] Scatterplots of reciprocal(CmpPass(p90)) and reciprocal(Touches(p90)) plotted against the dependent variable (PPM)
[9] Parameter estimates, including the intercept term, of round 3 testing
[10] Residual normal quantile plot
[11] Correlation of estimates table
[12] Durbin-Watson table

III. Analysis and Discussion

As discussed previously, the goal of our regression model was to determine what factors had the greatest influence on a player's impact on their team's performance shown through the PPM (Points Per Match) metric. The final variables remaining in our model were able to be grouped into three categories:

- Goal Impacting Actions - metrics that those that directly led to a shot
- High Value Possessions - either touches in dangerous areas (the attacking third of the pitch) or carries with the intent of moving the play into more valuable spaces
- Play Involvements - measures how often a player could impact the game through either being on the ball of distributing the ball

Together, these metrics were used to develop our model which found players who were able to consistently convert possessions into shots from dangerous locations to have the highest PPM. In other words, our model values players who are efficient in attacking the goal. One area that was particularly intriguing was how the model values SoT and npxG. It is not surprising to see that SoT is highly significant in our model. This is an understandable significance because SoT tend to be the most direct action that is not a goal being scored. In fact, our model sees a high volume of SoTs per match to be valuable justified by its coefficient of .388. However, npxG, also, has a high coefficient of 2.049; thus, highlighting how strictly getting a shot on frame is less valuable than taking shots from dangerous positions.

Furthermore, looking at our other independent variables. Our model has a coefficient of .479 for log(Touches (p90)). This coefficient builds on the idea that wherever actions occur, each action has a larger impact than the raw frequency of an action. Reading further into log(Touches (p90)), the log transformation is intriguing as this enables the variable to account for players who play on possession-dominant teams. This is an intriguing statistic because it's well known amongst soccer community that teams whose tactics focus on possession often spend a significant time in the attacking third of the field, in turn leading to traditionally high value touches to be less efficient and dangerous as the defending team will typically be more compact and concede this area of the pitch. Next, in our model, the SCA and Carries variables both require no transformations and yield low positive coefficients. This information is particularly interesting when accounting for variables that were discounted for the model, such as xA, PPA, and PrgDistCarry. These discounted variables are typically highly valued since they portray a player's ability to penetrate the opposing team's defensive schemes and create chances; however, these are potentially overrated by the public when valuing players. While it seems obvious that a player being capable of consistently taking and/or creating shots near the opponents' goal position would in turn yield a high PPM, our model which contains metrics that support this claim. Therefore, justifying the variables we have chosen to keep in our model. Essentially, our model highlights that, despite soccer being a complex game, simply being able to take shots as close to the opponent's goal as possible is in invaluable.

IV. Appendices

1.

For the purposes of this study, we looked at players on the field above labeled CM, LW, AM, RW, WF, and CF

2.

Stepwise Fit for PPM

Stepwise Regression Control

Stopping Rule: Minimum BIC

Direction: Backward

Enter All | Make Model | Remove All | Run Model

Go | Stop | Step

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
13.728079	116	0.3440138	0.5297	0.5013	7.0664318	8	98.57176	122.3753

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
✓	✓	Intercept	0.36681765	1	0	0.000	1
☐	✓	CmpPass(p90)	0.05211263	1	1.922674	16.246	0.0001
☐	☐	AttPass(p90)	0	1	0.005396	0.045	0.83198
☐	☐	xA(p90)	0	1	0.004524	0.038	0.84596
☐	☐	KP(p90)	0	1	0.060697	0.511	0.47628
☐	☐	PPA(p90)	0	1	0.139792	1.183	0.279
☐	☐	CrsPA(p90)	0	1	0.155097	1.314	0.25403
☐	☐	Touches Att Pen(p90)	0	1	0.183038	1.554	0.21508
☐	✓	Carries(p90)	0.02707776	1	0.727051	6.143	0.01463
☐	☐	PrgDistCarry(p90)	0	1	0.013453	0.113	0.73759
☐	✓	SCA(p90)	0.10858717	1	0.635269	5.368	0.02227
☐	☐	Sh(p90)	0	1	0.188808	1.604	0.20794
☐	✓	SoT(p90)	0.40133681	1	0.890269	7.523	0.00706
☐	✓	npxG(p90)	1.44955534	1	1.434013	12.117	0.0007
☐	✓	Touches(p90)	-0.0514051	1	2.200245	18.592	3.42e-5
☐	✓	Touches Att 3rd(p90)	0.00767269	1	1.559481	13.177	0.00042

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	
1	All	Entered	.	.	0.5586	16	16	110.907	153.079	○
2	PrgDistCarry(p90)	Removed	0.9413	0.000649	0.5586	14.005	15	108.224	148.265	○
3	CrsPA(p90)	Removed	0.7175	0.01556	0.5580	12.136	14	105.734	143.594	○
4	KP(p90)	Removed	0.4928	0.055537	0.5561	10.601	13	103.676	139.306	○
5	xA(p90)	Removed	0.5364	0.044896	0.5546	8.9777	12	101.56	134.915	○
6	AttPass(p90)	Removed	0.3933	0.085256	0.5517	7.6923	11	99.8726	130.905	○
7	PPA(p90)	Removed	0.2264	0.171356	0.5458	7.1286	10	99.032	127.698	○
8	Touches Att Pen(p90)	Removed	0.1229	0.281003	0.5362	7.4839	9	99.2225	125.478	○
9	Sh(p90)	Removed	0.2079	0.188808	0.5297	7.0664	8	98.5718	122.375	○
10	SCA(p90)	Removed	0.0223	0.635269	0.5079	10.391	7	101.854	123.164	○
11	SoT(p90)	Removed	0.0078	0.899442	0.4771	15.93	6	107.099	125.876	○
12	CmpPass(p90)	Removed	0.0099	0.888088	0.4467	21.374	5	111.865	128.068	○
13	Touches(p90)	Removed	0.0298	0.656299	0.4242	24.875	4	114.594	128.187	○
14	Touches Att 3rd(p90)	Removed	0.0007	1.678385	0.3667	36.943	3	124.225	135.17	○
15	npxG(p90)	Removed	0.0000	5.303255	0.1850	79.394	2	153.365	161.625	○
16	Carries(p90)	Removed	0.0000	5.401032	-0.000	122.66	1	176.635	182.176	○
17	Best	Specific	.	.	0.5297	7.0664	8	98.5718	122.375	◉

The results of the backwards stepwise regression, the model we ultimately decided to move forward with for assumption testing

Stepwise Fit for PPM

Stepwise Regression Control

Stopping Rule: Minimum BIC

Direction: Forward

Enter All | Make Model | Remove All | Run Model

Go | Stop | Step

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
13.728079	116	0.3440138	0.5297	0.5013	7.0664318	8	98.57176	122.3753

Current Estimates

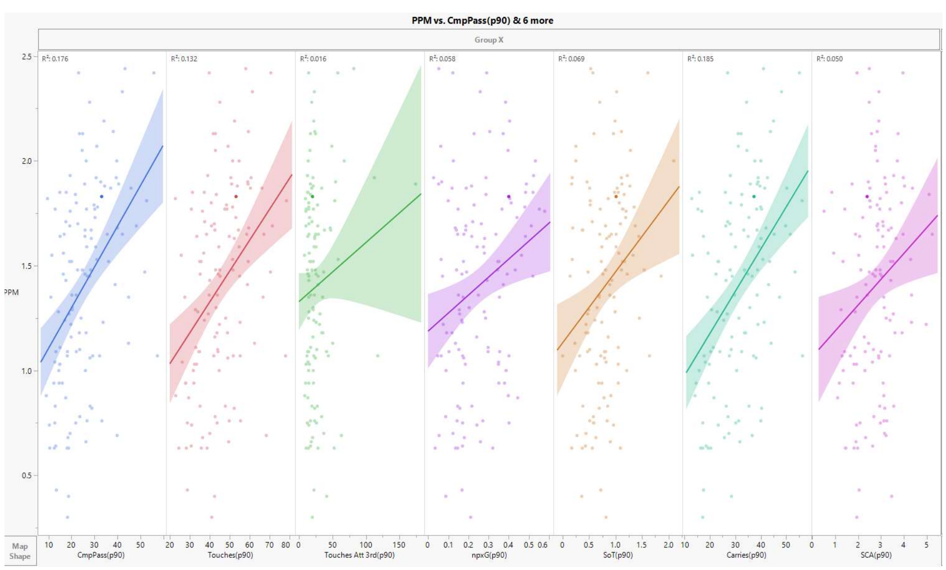
Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
✓	✓	Intercept	0.36681765	1	0	0.000	1
☐	✓	CmpPass(p90)	0.05211263	1	1.922674	16.246	0.0001
☐	☐	AttPass(p90)	0	1	0.005396	0.045	0.83198
☐	☐	xA(p90)	0	1	0.004524	0.038	0.84596
☐	☐	KP(p90)	0	1	0.060697	0.511	0.47628
☐	☐	PPA(p90)	0	1	0.139792	1.183	0.279
☐	☐	CrsPA(p90)	0	1	0.155097	1.314	0.25403
☐	☐	Sh(p90)	0	1	0.188808	1.604	0.20794
☐	✓	SoT(p90)	0.40133681	1	0.890269	7.523	0.00706
☐	☐	Touches Att Pen(p90)	0	1	0.183038	1.554	0.21508
☐	✓	Carries(p90)	0.02707776	1	0.727051	6.143	0.01463
☐	☐	PrgDistCarry(p90)	0	1	0.013453	0.113	0.73759
☐	✓	SCA(p90)	0.10858717	1	0.635269	5.368	0.02227
☐	✓	npxG(p90)	1.44955534	1	1.434013	12.117	0.0007
☐	✓	Touches(p90)	-0.0514051	1	2.200245	18.592	3.42e-5
☐	✓	Touches Att 3rd(p90)	0.00767269	1	1.559481	13.177	0.00042

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC
1	Carries(p90)	Entered	0.0000	5.401032	0.1850	79.394	2	153.365	161.625
2	npxG(p90)	Entered	0.0000	5.303255	0.3667	36.943	3	124.225	135.17
3	Touches Att 3rd(p90)	Entered	0.0007	1.678385	0.4242	24.875	4	114.594	128.187
4	Touches(p90)	Entered	0.0298	0.656299	0.4467	21.374	5	111.865	128.068
5	CmpPass(p90)	Entered	0.0099	0.888088	0.4771	15.93	6	107.099	125.876
6	SoT(p90)	Entered	0.0078	0.899442	0.5079	10.391	7	101.854	123.164
7	SCA(p90)	Entered	0.0223	0.635269	0.5297	7.0664	8	98.5718	122.375
8	Sh(p90)	Entered	0.2079	0.188808	0.5362	7.4839	9	99.2225	125.478
9	Touches Att Pen(p90)	Entered	0.1229	0.281003	0.5458	7.1286	10	99.032	127.698
10	PPA(p90)	Entered	0.2264	0.171356	0.5517	7.6923	11	99.8726	130.905
11	AttPass(p90)	Entered	0.3933	0.085256	0.5546	8.9777	12	101.56	134.915
12	xA(p90)	Entered	0.5364	0.044896	0.5561	10.601	13	103.676	139.306
13	KP(p90)	Entered	0.4928	0.055537	0.5580	12.136	14	105.734	143.594
14	CrsPA(p90)	Entered	0.7175	0.01556	0.5586	14.005	15	108.224	148.265
15	PrgDistCarry(p90)	Entered	0.9413	0.000649	0.5586	16	16	110.907	153.079
16	Best	Specific	.	.	0.5297	7.0664	8	98.5718	122.375

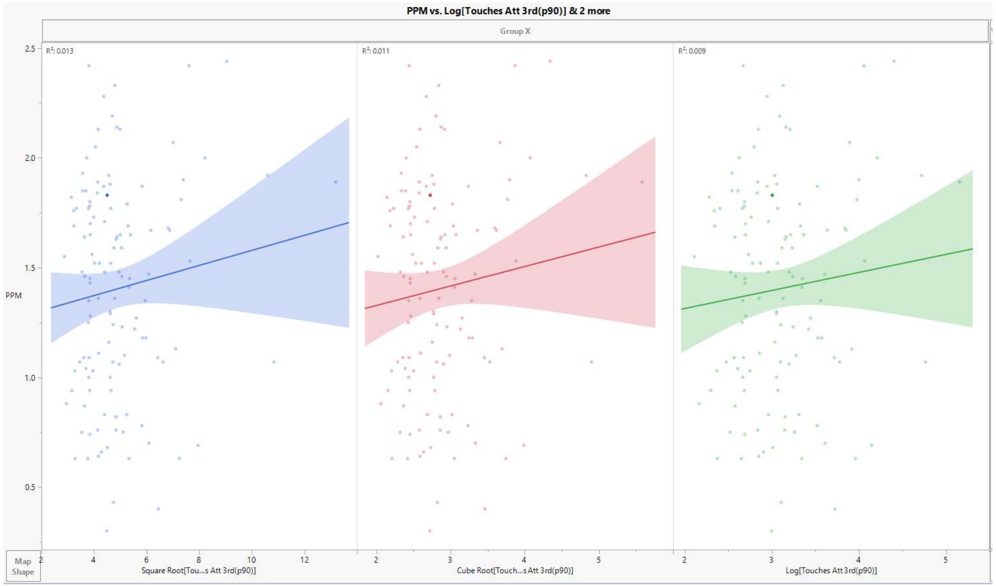
The results of the forward stepwise regression. This model has the same R-squared and included all the same variable as the backward stepwise regression above

3.



The results of the independents variables plotted against the dependent variable showed us that a linear relationship was present in all plots except for Touches Att 3rd (p90).

4.



The results of the three transformations performed on Touches Att 3rd (p90). Although the log transformation had the lowest R-squared value, it had fewer outliers which is why it was chosen.

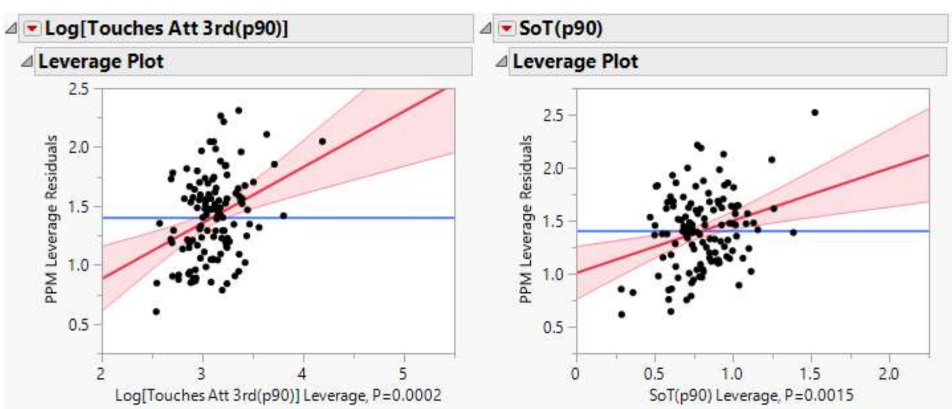
5.

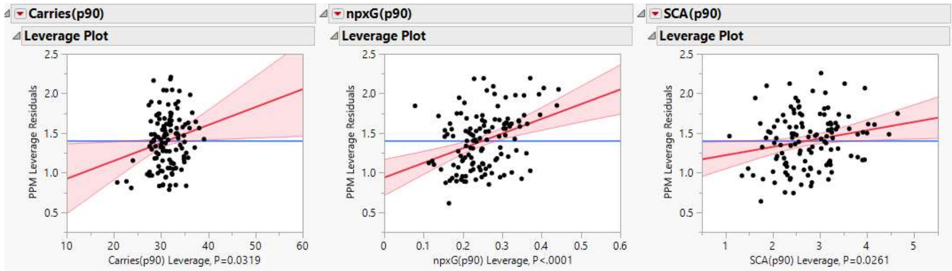
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.872853	0.432506	-2.02	0.0459*
Log[Touches Att 3rd(p90)]	0.4738692	0.123006	3.85	0.0002*
SoT(p90)	0.4941162	0.1523	3.24	0.0015*
Carries(p90)	0.0226271	0.010417	2.17	0.0319*
npxG(p90)	1.8531439	0.442775	4.19	<.0001*
SCA(p90)	0.1051193	0.046655	2.25	0.0261*
Touches(p90)	-0.053571	0.01193	-4.49	<.0001*
CmpPass(p90)	0.0538877	0.01272	4.24	<.0001*

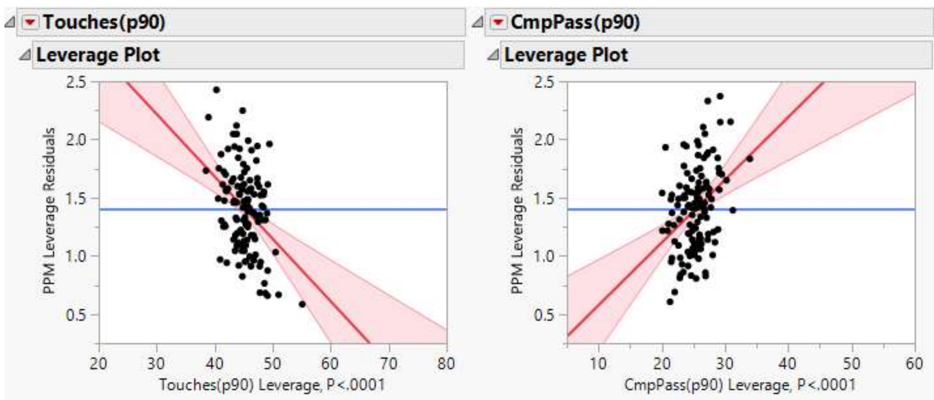
Parameter estimates for round 2 of testing, specifically showing that the intercept term is significant and therefore included in our model, satisfying the assumption that the error term have a population mean of zero

6.



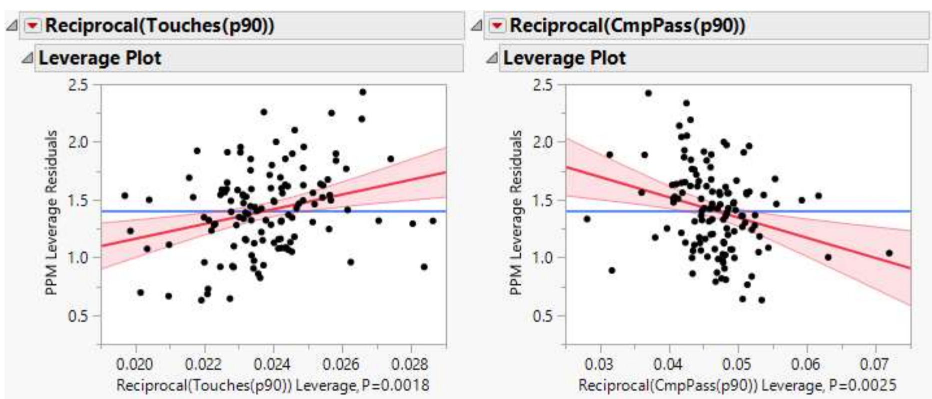


Residual plots for independent variables that we exogenous: log(Touches Att 3rd (p90)), SoT (p90), Carries (p90), npxG (p90), and SCA (p90)



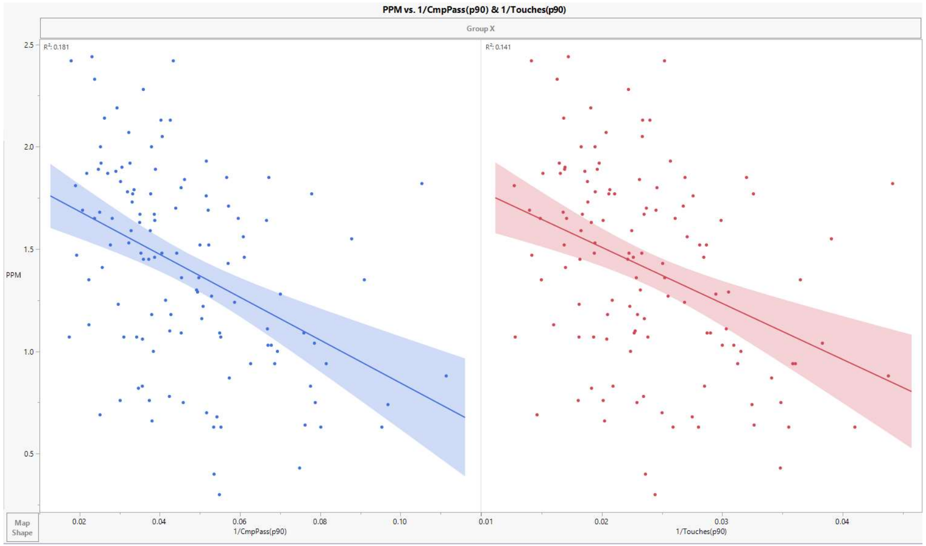
Residual plots for the two independent variables that displayed patterns of endogeneity: Touches (p90) and CmpPass (p90). Transformations were performed on these two variables to mitigate this.

7.



Residual plots for the new, transformed variables: reciprocal(Touches(p90)) and reciprocal(CmpPass(p90)). The new residual pots are much more random, and the degree of endogeneity is greatly diminished with is transformation

8.



The results of reciprocal(CmpPass(p90)) and reciprocal(Touches(p90)) in their new, transformed state. Both plots suggest a linear relationship is present and thus satisfy that assumption.

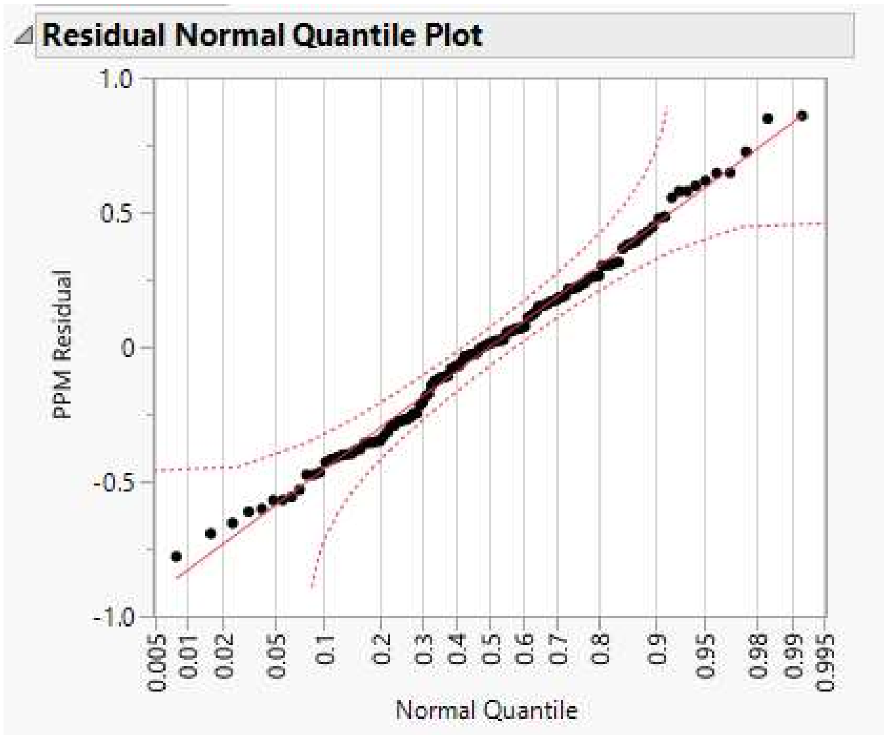
9.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-2.57123	0.79708	-3.23	0.0016*
Reciprocal(CmpPass(p90))	-17.57358	5.676404	-3.10	0.0025*
Carries(p90)	0.0234778	0.007478	3.14	0.0021*
SCA(p90)	0.0816953	0.042109	1.94	0.0548
SoT(p90)	0.3878719	0.162057	2.39	0.0183*
npxG(p90)	2.0488439	0.457122	4.48	<.0001*
Reciprocal(Touches(p90))	64.011593	19.98448	3.20	0.0018*
Log[Touches Att 3rd(p90)]	0.4787032	0.1246	3.84	0.0002*

Parameter estimates for round 3 of testing, specifically showing that the intercept term is still significant and therefore included in our model, satisfying the assumption that the error term have a population mean of zero

10.



Residual normal quantile plot showing the error term to be normally distributed. Although this assumption is considered to be optional, we still included it in our testing to enhance the validity of our model.

11.

Correlation of Estimates

Corr

	Intercept	Reciprocal(CmpPass(p90))	Carries(p90)	SCA(p90)	SoT(p90)	npxG(p90)	Reciprocal(Touches(p90))	Log[Touches Att 3rd(p90)]
Intercept	1.0000	0.0855	-0.6053	-0.0542	-0.5050	-0.1771	-0.5997	-0.7735
Reciprocal(CmpPass(p90))	0.0855	1.0000	0.0695	-0.1570	-0.1154	0.1637	-0.7775	0.0902
Carries(p90)	-0.6053	0.0695	1.0000	-0.2979	0.1567	-0.0600	0.4632	0.0795
SCA(p90)	-0.0542	-0.1570	-0.2979	1.0000	-0.0442	0.1216	0.0581	0.0075
SoT(p90)	-0.5050	-0.1154	0.1567	-0.0442	1.0000	-0.5112	0.3320	0.4454
npxG(p90)	-0.1771	0.1637	-0.0600	0.1216	-0.5112	1.0000	-0.1848	0.3548
Reciprocal(Touches(p90))	-0.5997	-0.7775	0.4632	0.0581	0.3320	-0.1848	1.0000	0.1842
Log[Touches Att 3rd(p90)]	-0.7735	0.0902	0.0795	0.0075	0.4454	0.3548	0.1842	1.0000

Correlation of estimates table showing the correlation between the independent variables. As discussed in the paper, CmpPass and Touches has a correlation estimate of .77, but ultimately was left in the model. However, these two variables were not heavily weighted when discussing the real-world impact of our final model.

12.

Durbin-Watson

Durbin-Watson	Number of Obs.	AutoCorrelation	Prob<DW
1.783798	124	0.1049	0.1109

Durbin-Watson test to identify any signs of auto-correlation. The statistic is on a scale of 0-4, with 2 being perfect no auto-correlation. A score of 1 or below indicates signs of auto-correlation, but our score of 1.78 suggests that auto-correlation is not present.