Premier League Player Regression Analysis Project

Abstract

This study explores player performance in the 2023–24 Premier League season by analyzing how various advanced metrics predict goal contributions per 90 minutes (G+A_90). Using five predictors (xG, xAG, PrgP, PrgC, and Minutes Played), we applied multiple linear regression and various model selection techniques, including Best Subsets, Forward Selection, Backward Elimination, and Stepwise Selection. The analysis identified xG, xAG, and Minutes Played as significant predictors, resulting in a final model with strong explanatory power (Adjusted $R^2 \approx 0.4627$).

Introduction

Football analytics has expanded beyond traditional metrics to include advanced statistics such as expected goals (xG) and progressive movements. Understanding how these metrics correlate with actual output like goals and assists is crucial for scouting, coaching, and player evaluation. In this project, we investigate how xG, xAG, PrgP, PrgC, and Minutes predict a player's contribution per 90 minutes (G+A_90).

Premier League Player Dataset Preview

Obs	Player	Nation	Pos	Min	xG	xAG	PrgC	PrgP	G+A_90
1	Rodri	es ESP	MF	2931	4.1	3.9	76	376	0.52
2	Phil Foden	eng ENG	FW,MF	2857	10.3	8.4	93	168	0.85
3	Ederson	br BRA	GK	2785	0	0.1	0	4	0
4	Julián Álvarez	ar ARG	MF,FW	2647	13	6.4	64	103	0.65
5	Kyle Walker	eng ENG	DF	2767	0.4	2.6	74	157	0.13
6	Bernardo Silva	pt POR	MF,FW	2578	3.7	7.6	140	177	0.52
7	Erling Haaland	no NOR	FW	2552	29.2	4.3	35	26	1.13
8	Rúben Dias	pt POR	DF	2559	1.4	0.3	34	173	0
9	Manuel Akanji	ch SUI	DF,MF	2511	1.9	0.5	46	148	0.07
10	Joško Gvardiol	hr CRO	DF	2328	3.1	1.4	63	136	0.19

CHECKING FOR ASSUMPTIONS

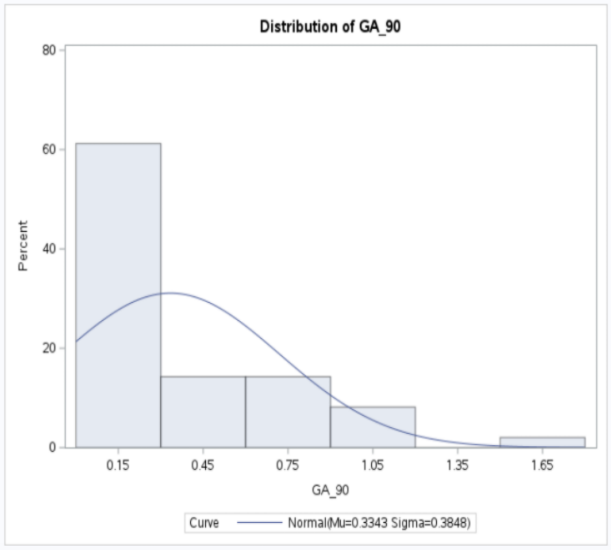
After importing our data, we used proc print to display the data and we could see that we had 49 observations however we couldn't include a whole picture of the 49 observation so we did the first 10

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
GA_90	49	0.33	0.38	0.00	1.70
Min	49	1529.92	956.46	1.00	3177.00
xG	49	3.52	5.83	0.00	29.20
xAG	49	2.56	3.02	0.00	11.80
PrgC	49	41.86	42.65	0.00	218.00
PrgP	49	85.04	76.90	0.00	376.00

After importing we tried to do a descriptive statistics by checking the number of observation for each variable, the mean, standard deviation ,minimum and maximum points by using pro means.

After that, we used pro Univariate to check for normal distribution for each of our predictive variables, and it shows that some of our variables are skewed right, but it does not influence our data as much, and you can see that min is approximately normally distributed

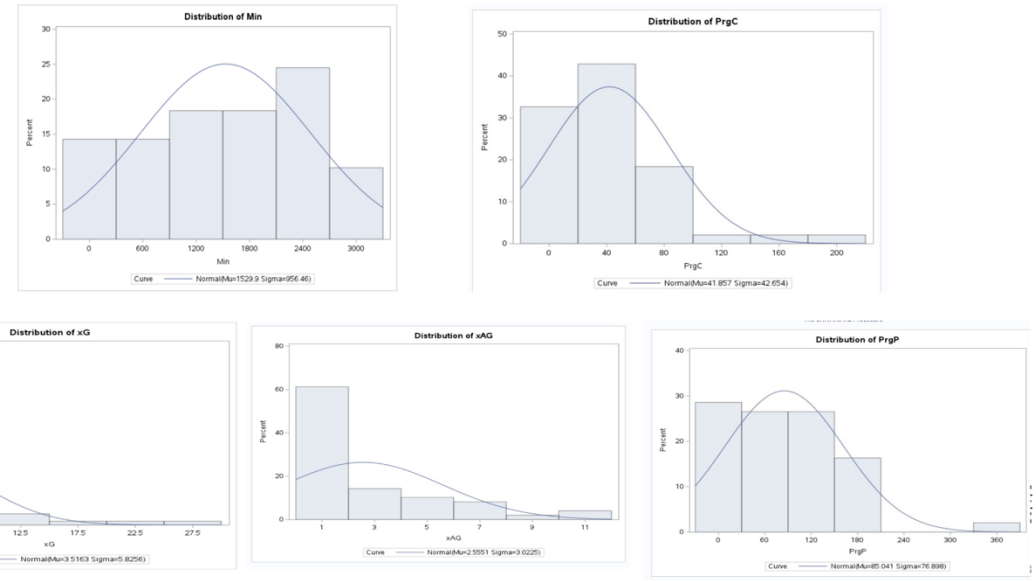


The UNIVARIATE Procedure
Fitted Normal Distribution for GA_90

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	0.334286
Std Dev	Sigma	0.384838

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.19252178	Pr > D	<0.010
Cramer-von Mises	W-Sq	0.43654009	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	2.58901227	Pr > A-Sq	<0.005

HISTOGRAMS CHECKED NORMALITY



Part of checking the assumptions was checking if there are any multicollinearity issues and ran the proc reg for the full model to check our VIF and. it shows that all our VIF is below 10 so we do not have any multicollinearity issues

INITIAL VIF VALUES CHECKED MULTICOLLINEARITY

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.25941	0.07944	3.27	0.0021	0
Min	1	-0.00013310	0.00007372	-1.81	0.0780	2.88872
xG	1	0.02963	0.01021	2.90	0.0058	2.05401
xAG	1	0.05359	0.02240	2.39	0.0212	2.66249
PrgC	1	0.00077195	0.00132	0.58	0.5626	1.85030
PrgP	1	0.00005974	0.00089193	0.07	0.9469	2.73321

We ran a simple linear regression in a multi linear regression, and they both give us the same answers because it both shows the same P values for each predictor variable.

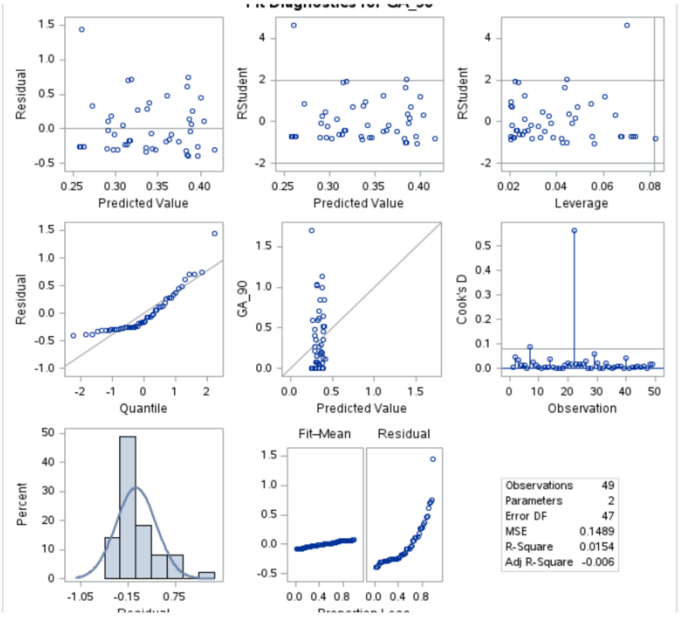
Looking at the QQ plot for each predictive variable, we can conclude that all of them are drawn from a normal distributed population. And as well as the residual is normally distributed to constant variance can be assumed

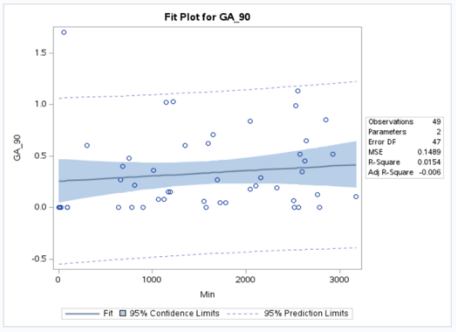
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.10964	0.10964	0.74	0.3952
Error	47	6.99916	0.14892		
Corrected Total	48	7.10880			

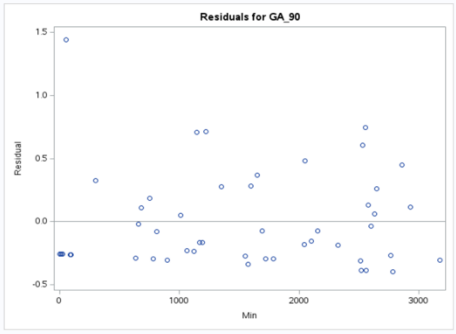
Root MSE	0.38590	R-Square	0.0154
Dependent Mean	0.33429	Adj R-Sq	-0.0055
Coeff Var	115.43994		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.25784	0.10477	2.46	0.0176
Min	1	0.00004997	0.00005824	0.86	0.3952

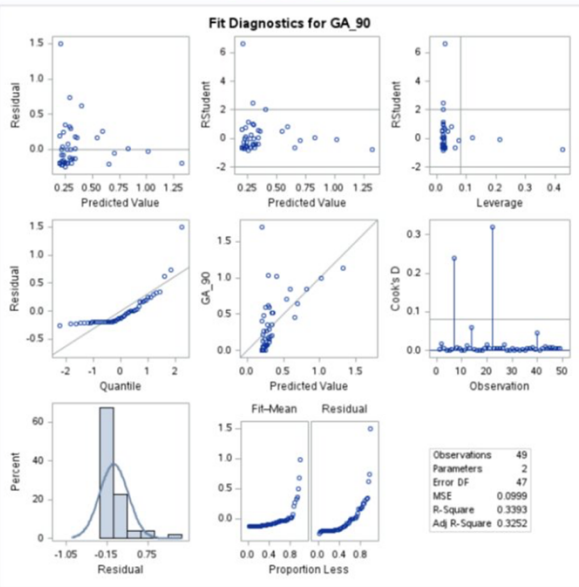
MODEL 1: FOR MINUTES PLAYED

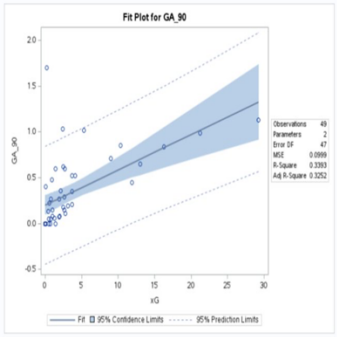






MODEL 2: EXPECTED GOALS





Dependent Variable: GA_90

Number of Observations Read	49
Number of Observations Used	49

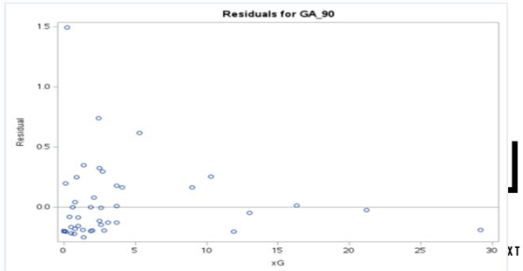
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.41196	2.41196	24.14	<.0001
Error	47	4.69684	0.09993		
Corrected Total	48	7.10880			

Root MSE	0.31612	R-Square	0.3393
Dependent Mean	0.33429	Adj R-Sq	0.3252
Coeff Var	94.56621		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.19898	0.05290	3.76	0.0005
xG	1	0.03848	0.00783	4.91	<.0001

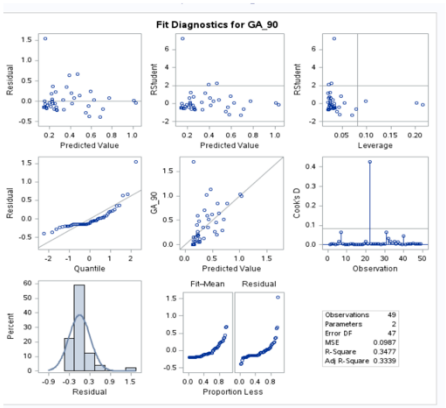


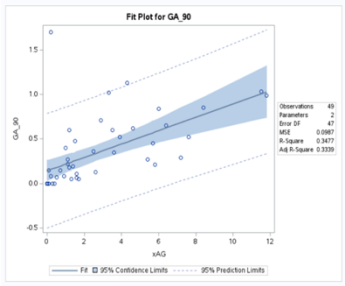
MODEL 3:ASSISTS GOALS

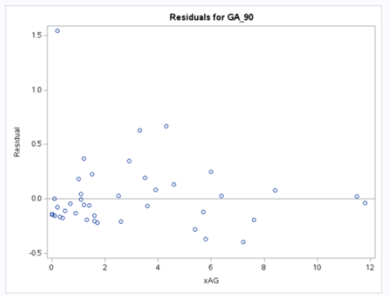
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.47195	2.47195	25.06	<.0001
Error	47	4.63685	0.09866		
Corrected Total	48	7.10880			

Root MSE	0.31410	R-Square	0.3477
Dependent Mean	0.33429	Adj R-Sq	0.3339
Coeff Var	93.96040		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.14244	0.05901	2.41	0.0197
xAG	1	0.07508	0.01500	5.01	<.0001







MODEL 4:PROGRESSIVE PASSES

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.05462	0.05462	0.36	0.5492
Error	47	7.05418	0.15009		
Corrected Total	48	7.10880			

Root MSE	0.38741	R-Square	0.0077
Dependent Mean	0.33429	Adj R-Sq	-0.0134
Coeff Var	115.89283		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.29698	0.08299	3.58	0.0008
PrgP	1	0.00043866	0.00072717	0.60	0.5492

MODEL 5:PROGRESSIVE CARRIES

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.92455	0.92455	7.03	0.0109
Error	47	6.18425	0.13158		
Corrected Total	48	7.10880			

Root MSE	0.36274	R-Square	0.1301
Dependent Mean	0.33429	Adj R-Sq	0.1115
Coeff Var	108.51172		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.19809	0.07297	2.71	0.0092
PrgC	1	0.00325	0.00123	2.65	0.0109

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	3.55632	0.71126	8.61	<.0001
Error	43	3.55248	0.08262		
Corrected Total	48	7.10880			

Root MSE	0.28743	R-Square	0.5003
Dependent Mean	0.33429	Adj R-Sq	0.4422
Coeff Var	85.98329		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.25941	0.07944	3.27	0.0021	0.09920	0.41961
Min	1	-0.00013310	0.00007372	-1.81	0.0780	-0.00028177	0.00001558
xG	1	0.02963	0.01021	2.90	0.0058	0.00905	0.05021
xAG	1	0.05359	0.02240	2.39	0.0212	0.00842	0.09876
PrgC	1	0.00077195	0.00132	0.58	0.5626	-0.00190	0.00344
PrgP	1	0.00005974	0.00089193	0.07	0.9469	-0.00174	0.00186

Pearson Correlation Coefficients, N = 49						
	GA_90	Min	xG	xAG	PrgC	PrgP
GA_90	1.00000	0.12419	0.58249	0.58969	0.36063	0.08765
Min	0.12419	1.00000	0.45135	0.48035	0.48808	0.72140
xG	0.58249	0.45135	1.00000	0.59543	0.35946	0.15989
xAG	0.58969	0.48035	0.59543	1.00000	0.64113	0.48359
PrgC	0.36063	0.48808	0.35946	0.64113	1.00000	0.45694
PrgP	0.08765	0.72140	0.15989	0.48359	0.45694	1.00000

Parameter Estimates									
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS	Squared Partial Corr Type I	Squared Partial Corr Type II
Intercept	1	0.25941	0.07944	3.27	0.0021	5.47560	0.88098	.	.
Min	1	-0.00013310	0.00007372	-1.81	0.0780	0.10964	0.26928	0.01542	0.07046
xG	1	0.02963	0.01021	2.90	0.0058	2.47410	0.69621	0.35348	0.16386
xAG	1	0.05359	0.02240	2.39	0.0212	0.94394	0.47299	0.20860	0.11750
PrgC	1	0.00077195	0.00132	0.58	0.5626	0.02827	0.02813	0.00789	0.00786
PrgP	1	0.00005974	0.00089193	0.07	0.9469	0.00037062	0.00037062	0.00010432	0.00010432

Type I error:

- Min entered the first and explain 0.10964 of total variation
- xG entered second and explain an additional 2.47 of the total variation after Min has been added
- xAG entered third and explain an additional 0.94 of the total variation after Min and xG has been added
- PrgC entered fourth and explain an additional 0.028 of the total variation after Min, xG and xAG have been added
- PrgP entered last and explain an additional 0 of the total variation after Min, xG, xAG and PrgP have been added.

Type II error:

- Min explains 0.26928 of the variation after all four of other variables have been added
- xG explains 0.69621 of the variation after all four of other variables have been added
- xAG explains 0.47299 of the variation after all four of other variables have been added
- PrgC explains 0.028 of the variation after all four of other variables have been added
- PrgP explains 0 of the variation after all four of other variables have been added

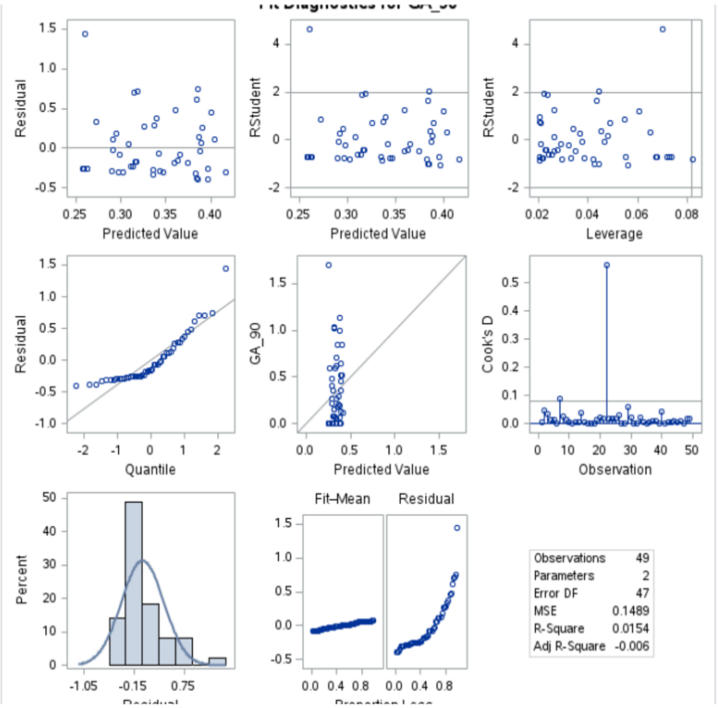
We use two data analysis methods in selecting our best models for predicting goals for 90 minutes and one of them was the best subset method and the criteria were used in selecting the best was the adjusted R squared in the CP value. We were looking for the highest adjusted our square and as well as a CP value that is less than five. And we picked out two models that included minutes, expected goals, expected goals, and assist. In the other model included minutes expected goals, expected goals, and assist and progressive passes, however, earlier it showed us that minutes is not significant, but it's part of the best model selected.

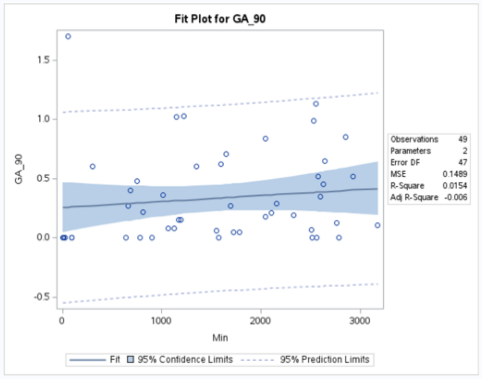
Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.26360	0.07711	3.42	0.0013	0
Min	1	-0.00012085	0.00004993	-2.42	0.0196	1.37567
xG	1	0.02876	0.00895	3.21	0.0024	1.63955
xAG	1	0.06044	0.01755	3.44	0.0013	1.69713

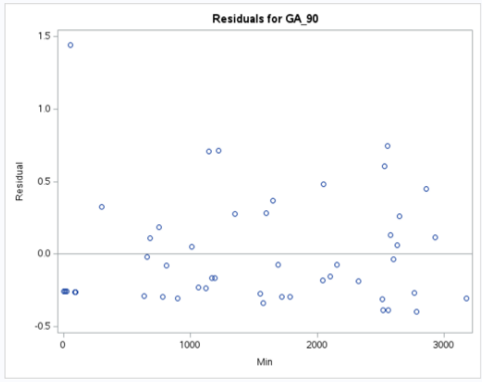
Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.26426	0.07841	3.37	0.0016	0
Min	1	-0.00012491	0.00007183	-1.74	0.0890	2.78403
xG	1	0.02912	0.01009	2.89	0.0060	2.03884
xAG	1	0.05979	0.01957	3.05	0.0038	2.06379
PrgP	1	0.00007033	0.00088504	0.08	0.9370	2.73208

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.12862	0.05598	2.30	0.0262	0
xG	1	0.02368	0.00915	2.59	0.0129	1.54929
xAG	1	0.04791	0.01763	2.72	0.0093	1.54929

MODEL 1: FOR MINUTES PLAYED







We used forward selection, backward elimination, and stepwise selection to help us get the best predictor variables for predicting goals and assists for 90 minutes

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	xAG	1	0.3477	0.3477	11.1254	25.06	<.0001
2	xG	2	0.0829	0.4307	5.9891	6.70	0.0129
3	Min	3	0.0656	0.4962	2.3467	5.86	0.0196

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.26360	0.07711	0.93009	11.69	0.0013
Min	-0.00012085	0.00004993	0.46615	5.86	0.0196
xG	0.02876	0.00895	0.82192	10.33	0.0024
xAG	0.06044	0.01755	0.94394	11.86	0.0013

Bounds on condition number: 1.6971, 14.137

All variables left in the model are significant at the 0.1500 level.

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	PrgP	4	0.0001	0.5002	4.0045	0.00	0.9469
2	PrgC	3	0.0040	0.4962	2.3467	0.35	0.5571

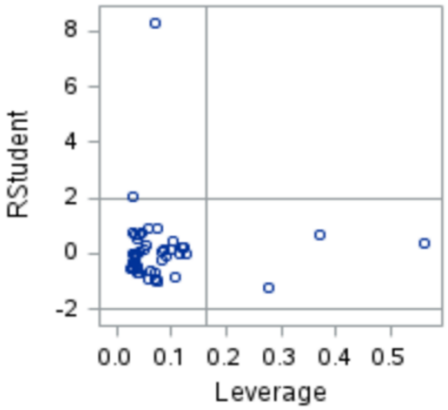
Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	xAG		1	0.3477	0.3477	11.1254	25.06	<.0001
2	xG		2	0.0829	0.4307	5.9891	6.70	0.0129
3	Min		3	0.0656	0.4962	2.3467	5.86	0.0196

Leverage tells us how far away a data point value for the independent variable is from the average of all data. These are X outliers and points with high leverage and big error are very influential, so we calculated the leverage values and any number greater than that has leverage, and these are results from the SAS output

CHECKING FOR X AND Y OUTLIER

Any leverage values > 2*3/49 =0.122 have leverage

- Darwin Nunez
- Mohamed salah
- Kevin de Bruyne
- Erling Haaland
- All have leverage values



Obs	Player	Nation	Pos	Min	xG	xAG	PrgC	PrgP	GA_90	cooksd	leverage	studdelresid	dffits
1	Mateo Kovačić	hr CRO	MF	1551	1.3	1.6	37	100	0.06	0.00187	0.02511	-0.53518	-0.08589
2	Harvey Elliott	eng ENG	MF,FW	1352	2.7	3.5	44	128	0.6	0.00408	0.02771	0.75339	0.12718
3	Ryan Gravenberch	nl NED	MF	1121	2	0.9	39	78	0.08	0.00236	0.02774	-0.57107	-0.09646
4	Jack Grealish	eng ENG	FW,MF	1009	2.1	2.5	87	50	0.36	0.00000	0.02866	0.02429	0.00417
44	Ederson	br BRA	GK	2785	0	0.1	0	4	0	0.00220	0.12070	0.25032	0.09274
45	Virgil van Dijk	nl NED	DF	3177	2.8	1.6	32	200	0.11	0.00139	0.12147	0.19864	0.07386
46	Darwin Núñez	uy URU	FW	2047	16.3	6	58	54	0.84	0.00003	0.12610	-0.02886	-0.01096
47	Mohamed Salah	eg EGY	FW	2534	21.2	11.8	98	147	0.99	0.14141	0.27820	-1.21793	-0.75612
48	Kevin De Bruyne	be BEL	MF	1221	2.4	11.5	47	111	1.03	0.06592	0.37044	0.66525	0.51030
49	Erling Haaland	no NOR	FW	2552	29.2	4.3	35	26	1.13	0.05201	0.56249	0.39847	0.45182

Output Statistics											
Obs	Min	xG	xAG	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	2931	4.1	3.9	0.52	0.2630	0.0764	0.1092	0.4169	-0.3256	0.8517	0.2570
2	2857	10.3	8.4	0.85	0.7223	0.0911	0.5388	0.9058	0.1252	1.3194	0.1277
3	2785	0	0.1	0.00	-0.0669	0.0980	-0.2643	0.1305	-0.6684	0.5346	0.0669
4	2647	13	6.4	0.65	0.7045	0.0807	0.5420	0.8669	0.1135	1.2954	-0.0545
5	2767	0.4	2.6	0.13	0.0979	0.0838	-0.0709	0.2666	-0.4948	0.6906	0.0321
45	753	0.9	1.5	0.48	0.2892	0.0524	0.1835	0.3948	-0.2888	0.8671	0.1908
46	781	0.5	0.1	0.00	0.1896	0.0553	0.0783	0.3010	-0.3893	0.7686	-0.1896
47	679	0.1	1	0.40	0.2449	0.0548	0.1344	0.3553	-0.3340	0.8237	0.1551
48	89	0.1	0	0.00	0.2557	0.0740	0.1066	0.4049	-0.3317	0.8431	-0.2557
49	26	0	0	0.00	0.2605	0.0762	0.1070	0.4139	-0.3281	0.8490	-0.2605
50	3188	30	13	.	1.5269	0.1951	1.1340	1.9198	0.8362	2.2177	.

- Created new data point: Min = 3188, xG = 30, xAG = 13
- Used model GA_90 ~ Min + xG + xAG to predict outcome
- Computed confidence and prediction intervals
- Interpreted predicted GA_90 from regression output

CONCLUSION:

- Final model: GA_90 ~ Min + xG + xAG
- These three variables best explain goal contribution per 90 mins
- Progressive pass and Progressive Carries (PrgC, PrgP) not significant