Multiple Linear Regression:

Analyzing the Impact of Game Performance Metrics on Points Scored of Dallas Mavericks

Alessandra Rodriguez

I. Proposal

A. Problem, Variables, & Data Description

For this project, we will investigate the number of points that the Dallas Mavericks score in one game during the 2023-2024 season. Our goal is to see if there is a linear relationship between points scored in a single game and the following variables: Turnovers, Field Goal Percentage, Total rebounds, and Luka Doncic minutes played per game.

We chose these variables because of their impact on basketball. Turnovers are when a member of a team violates a rule or causes the ball to go out of bounds. This results in the other team getting possession of the ball. A high number of turnovers leads to the opposing team having more possessions, thus having more chances to score. Field goal percentage is the ratio of field goals made to field goals attempted. This includes 2- and 3-point shots. Field Goal Percentage shows how efficiently the team shoots; the higher the percentage, the better the chance of scoring more points per game. Rebounds are defined as retrieving the ball after a missed shot attempt from either the Mavs or opponent. More rebounds lead to having more possessions, thus having more scoring opportunities. Lastly, total minutes played by Luka Doncic per game was considered due to his reputation of being a star player and effective shooting. We expect that when Doncic plays a high number of minutes, the Mavericks will score more points.

The data was collected from: https://www.espn.com/nba/team/_/name/dal/dallas-mavericks. Our sample consists of 30 games randomly selected from the 2023-2024 season. ESPN keeps track of the stats from each game, and we compiled them altogether from the beginning of the season. Each observation is reflective of the team’s cumulative stats from the individual game and not specific to one player. The data is also only specific to the Mavericks and the opponent’s stats are not recorded.

B. Scatter Plot & Correlation Discussion

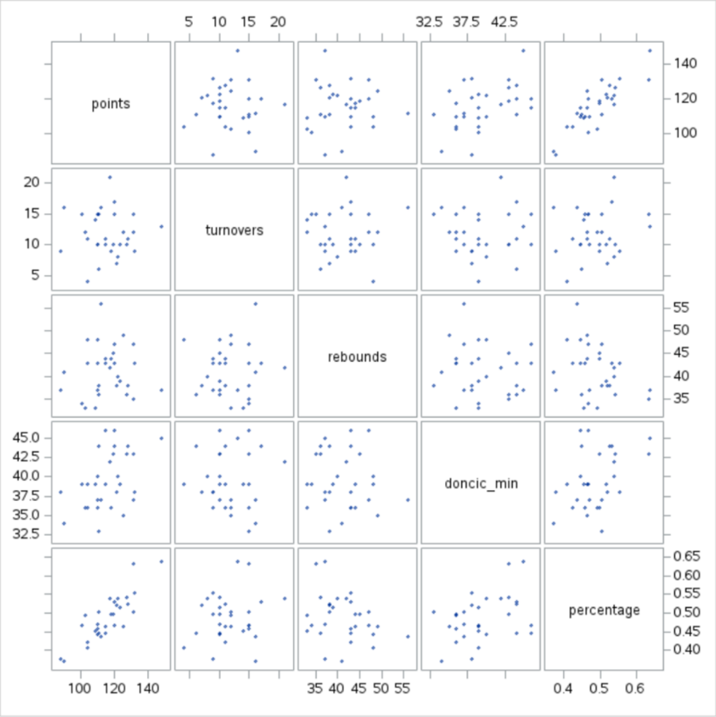
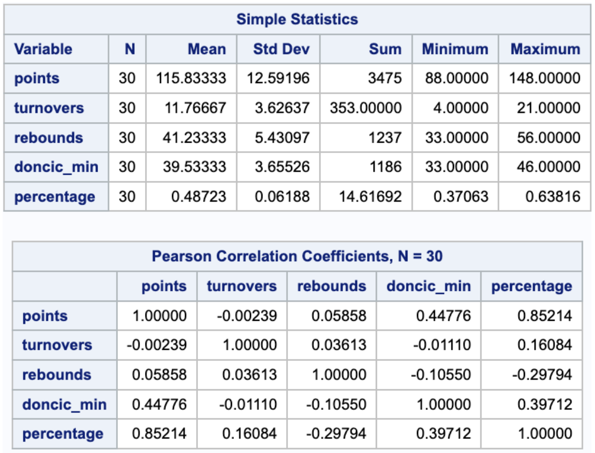


Figure 1: Plots of the response and predictor variables



Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
points	30	115.83333	12.59196	3475	88.00000	148.00000
turnovers	30	11.76667	3.62637	353.00000	4.00000	21.00000
rebounds	30	41.23333	5.43097	1237	33.00000	56.00000
doncic_min	30	39.53333	3.65526	1186	33.00000	46.00000
percentage	30	0.48723	0.06188	14.61692	0.37063	0.63816

Pearson Correlation Coefficients, N = 30					
	points	turnovers	rebounds	doncic_min	percentage
points	1.00000	-0.00239	0.05858	0.44776	0.85214
turnovers	-0.00239	1.00000	0.03613	-0.01110	0.16084
rebounds	0.05858	0.03613	1.00000	-0.10550	-0.29794
doncic_min	0.44776	-0.01110	-0.10550	1.00000	0.39712
percentage	0.85214	0.16084	-0.29794	0.39712	1.00000

Figure 2: Simple Statistics and Correlation Tables

Between the predictor variable, turnovers per game, and the response variable, points per game, there appears to be a weak negative linear relationship. There does not appear to be curvature suggesting a linear model could be appropriate. The downward pattern suggests that the Mavs score less as they turnover the ball more in one game. This relationship is quite weak due to how scattered the points are in the vertical dimension. Turnovers and points have a correlation of -0.002 which also implies that these variables have a weak negative relationship to one another; this is because the correlation is a negative value remarkably close to zero. A game with more than 20 turnovers could be considered an outlier on the horizontal axis.

Between the predictor variable, number of rebounds per game and the response variable, points per game, there appears to be a weak positive linear relationship between the two, with no obvious curvature. This suggests that as the number of rebounds goes up, so does the number of points in one game. This relationship is quite weak due to how scattered the points are in the vertical dimension. Rebounds and points have a correlation of 0.058 which also implies that these variables have a weak positive relationship to one another; this is because the correlation is a positive value remarkably close to zero. The one game with over 55 rebounds may be considered an outlier due to its far-right position on the graph.

The relationship between the predictor variable, total minutes Luka Doncic played per game and the response variable, points per game shows a moderately strong positive relationship between the two variables. Which would suggest that the longer Doncic is in the game, the more points scored in total in most of the games. Doncic minutes and points have a correlation of 0.448 which also implies that these variables have a moderately strong positive relationship to one another; this is because the correlation is a positive value that's halfway between zero and one. But there is a concern that there is possible curvature because there is a drop off around 40 minutes, this could be an issue with the assumption that a linear model is appropriate between the predictor and response variable. There appear to be no obvious outliers.

The relationship between the predictor variable, shot percentage per game and the response variable, points per game has a clear positive relationship between the two with no obvious curvature. Thus, a linear model would be appropriate here. This graph suggests that as the shot percentage per game increases, so do points per game. Shot percentage and points have a correlation of 0.852 which also implies that these variables have a strong positive relationship to one another; this is because the correlation is a positive value that's close one. It's possible that the far 2 right games on the graph are outliers on the horizontal axis.

Between the two predictor variables: turnovers and rebounds, there appears to be no relationship, only scatter, and no outliers. Their correlation is 0.036 which also implies there is no linear relationship between these variables because their correlation is close to zero.

Between the two predictor variables: Doncic minutes played per game and rebounds, similarly there is no linear relationship between the variables and no clear outliers. Their correlation is –0.1 which also implies there is a very weak negative linear relationship between these variables because their correlation is close to zero.

There appears to be a weak linear relationship between the two predictor variables, turnovers and shot percentage, but it has no outliers. As shot percentage increases, the number of turnovers also increases. Their correlation is 0.16 which also implies there is a very weak linear relationship between these variables because their correlation is close to zero.

The two predictor variables, rebounds and the number of minutes Doncic played per game have no clear linear relationship, just random scatter with no outliers. Their correlation is -0.11 which also implies there is a very weak negative linear relationship between these variables because their correlation is close to zero.

For the two predictor variables, rebounds and shot percentage, there also appears to be random scatter with a very weak linear relationship between the variables and no outliers. Their correlation is -0.29 which also implies there is a weak negative linear relationship between these variables because their correlation is close to zero. While the relationship between these two predictor variables is slightly stronger than the other pairs of predictor variables, the relationship is still too weak to be a concern for multicollinearity later.

There appears to be a weak linear relationship between the two predictor variables, Doncic minutes and successful shot percentage, which would suggest that the longer Doncic plays, the higher the shot percentage by game. No clear outliers in this relationship. Their correlation is 0.39 which also implies there is a moderately strong negative linear relationship between these variables because their correlation is closer to one. While the relationship between these two predictor variables is slightly stronger than the other pairs of predictor variables, the relationship is still too weak to be a concern for multicollinearity later.

C. **Potential Complications**

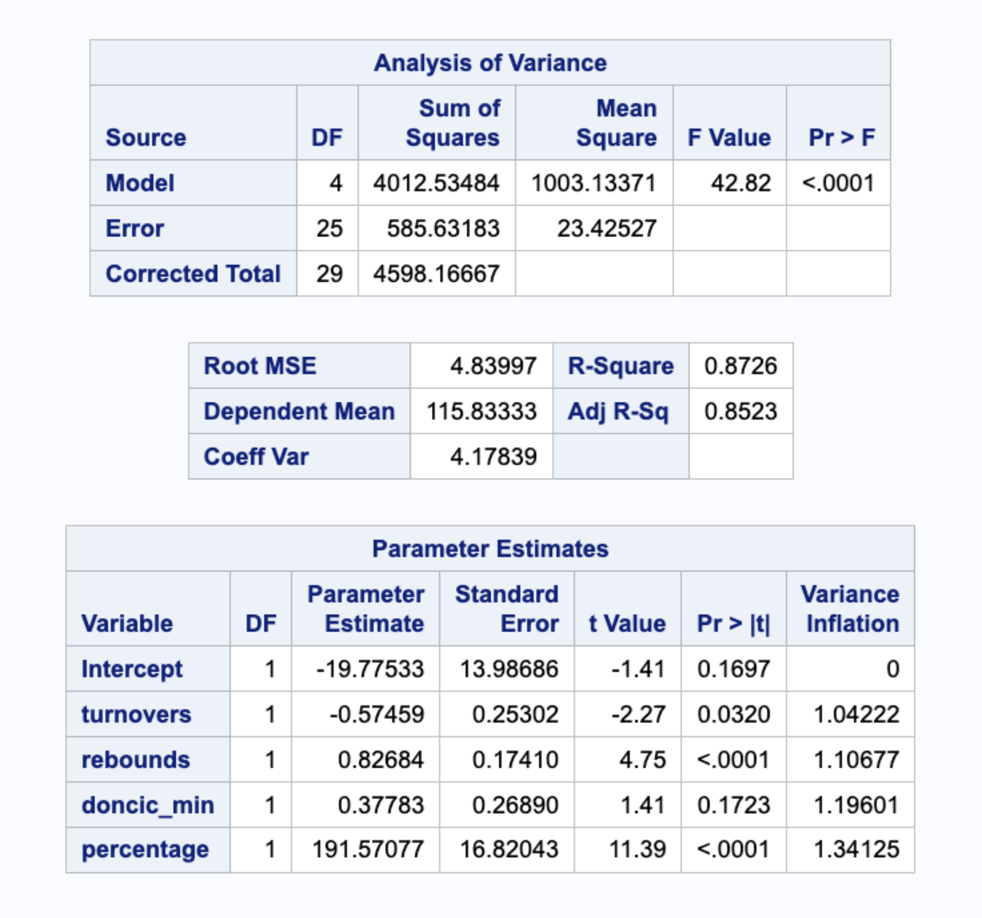
As shown by both the scatter plot and correlation matrices, there appears to be no pair of predictors with a strong linear relationship. This suggests that multicollinearity shouldn’t be a problem while trying to predict points scored in a single game using rebounds, turnovers, Doncic minutes played, and successful shot percentage.

Although multicollinearity was not detected, rebounds, turnovers, and Doncic minutes played have a very weak to moderate correlation with points scored. Therefore, these variables may not be great predictors for points scored per game. Only the successful shot percentage had a strong relationship with points scored with a correlation of 0.85.

II. Preliminary MLR Model Analysis

A. Preliminary Model

We will be investigating the multiple linear regression model that will estimate the number of points scored by the Mavs in a single game using turnovers, rebounds, Doncic game play minutes, and shot percentages.



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	4012.53484	1003.13371	42.82	<.0001
Error	25	585.63183	23.42527		
Corrected Total	29	4598.16667			

Root MSE	4.83997	R-Square	0.8726
Dependent Mean	115.83333	Adj R-Sq	0.8523
Coeff Var	4.17839		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-19.77533	13.98686	-1.41	0.1697	0
turnovers	1	-0.57459	0.25302	-2.27	0.0320	1.04222
rebounds	1	0.82684	0.17410	4.75	<.0001	1.10677
doncic_min	1	0.37783	0.26890	1.41	0.1723	1.19601
percentage	1	191.57077	16.82043	11.39	<.0001	1.34125

Figure 3: The ANOVA and Parameter Estimates Table for the Preliminary Model

Before we can assess the fitted preliminary model with all the predictors, we must investigate if the model assumptions are held.

B. Model Assumptions

For the following figures (fig. 4, fig. 5, fig. 6, fig. 7) residuals are calculated using the multiple linear regression model that uses turnovers, rebounds, Doncic minutes, and shot percentages to model total points scored in one game.

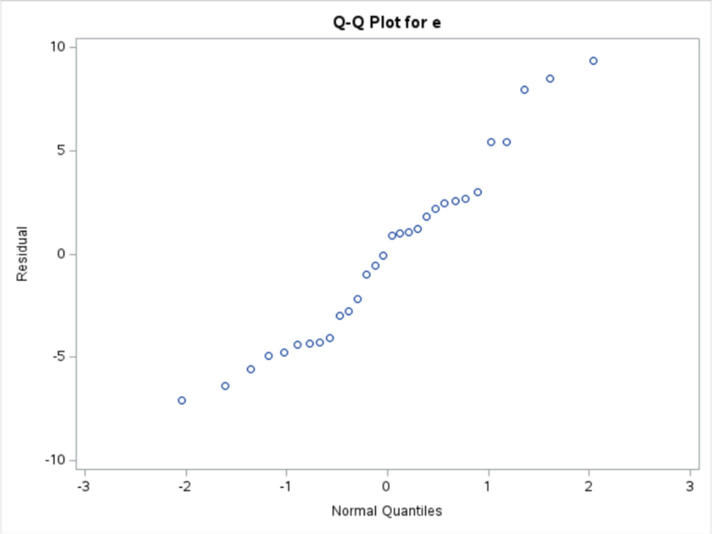


Figure 4: Normal Probability Plot for the Residuals

From figure 4, the normal probability plot appears to have a slight S-shape. This curvature in the plot implies that the distribution of residuals does not follow a normal distribution instead it has shorter tails than the normal distribution. While the residuals are not normally distributed, they do have a symmetrical distribution, like the normal distribution.

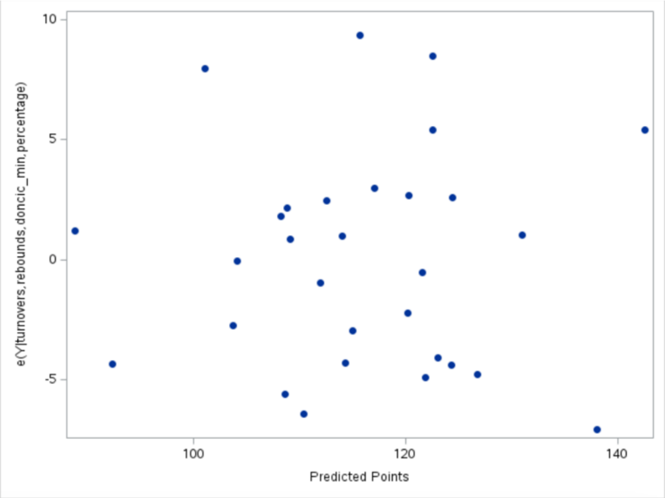


Figure 5: Predicted Points (Y_hat) vs the residuals of the preliminary model

From figure 5, the residuals are randomly scattered around 0. There appears to be no funnel shape which would indicate that we can reasonably assume the residuals have constant variance.

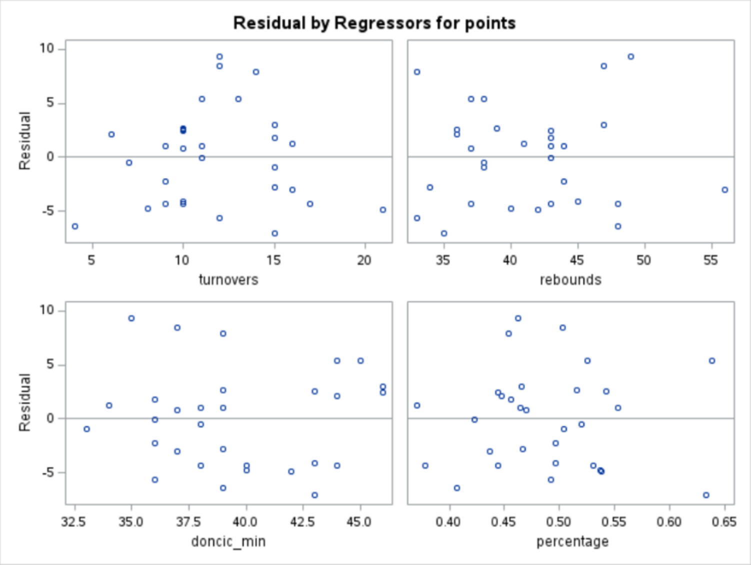


Figure 6: Plots of the predictor variables versus the residuals

From figure 6, each plot represents the residuals versus one predictor variable. For the turnovers plot, the residuals are randomly scattered around zero and there appears to be no curvature or outliers. The same can be said for the percentage plots Since there is no curvature detected amongst these plots, we can reasonably assume that the current model is appropriate to predict the total points scored in a single Mavs game using turnovers and shooting percentage.

For both Doncic minutes played and rebounds, there appears to be possible curvature each plot. The U-shapes found in these plots could possibly be a violation of the linear model assumption. We will later investigate transformations of these variables to see if this issue can be improved or if these are simply odd artifacts.

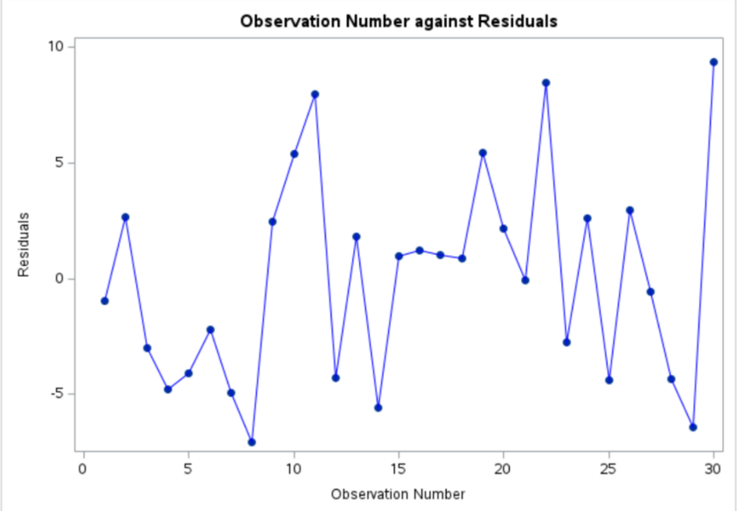


Figure 7: Plot of the Observation Number and Residuals

Lastly, figure 7 shows that there is random jaggedness and no serial correlation between the residuals. Therefore, we can reasonably assume that the residuals are uncorrelated. This is important to evaluate because these games were collected over the same season; we had to ensure there was no increase in the total points per game for the Mavs due to time. If the Mavs were increasing their effectiveness in shooting over time, then regression would've been an inappropriate model.

C. **Transformations**

From figure 1, there is curvature apparent between the number of minutes Doncic plays and points scored. Also, in figure 1, while there seems to be a linear relationship between turnovers and points scored, it is weak with a correlation of –0.002 from figure 2. Due to these factors, we will fit a model with additional transformed variables of both Doncic minutes and turnovers. For each variable, we standardized their values then squared them as a transformation and added them to the preliminary model for a new total of 6 predictors: turnovers, rebounds, Doncic minutes, shot percentage, standardized turnovers squared, and standardized Doncic minutes squared.

We will now investigate the residual plots and see if there are any improvements compared to the original preliminary model.

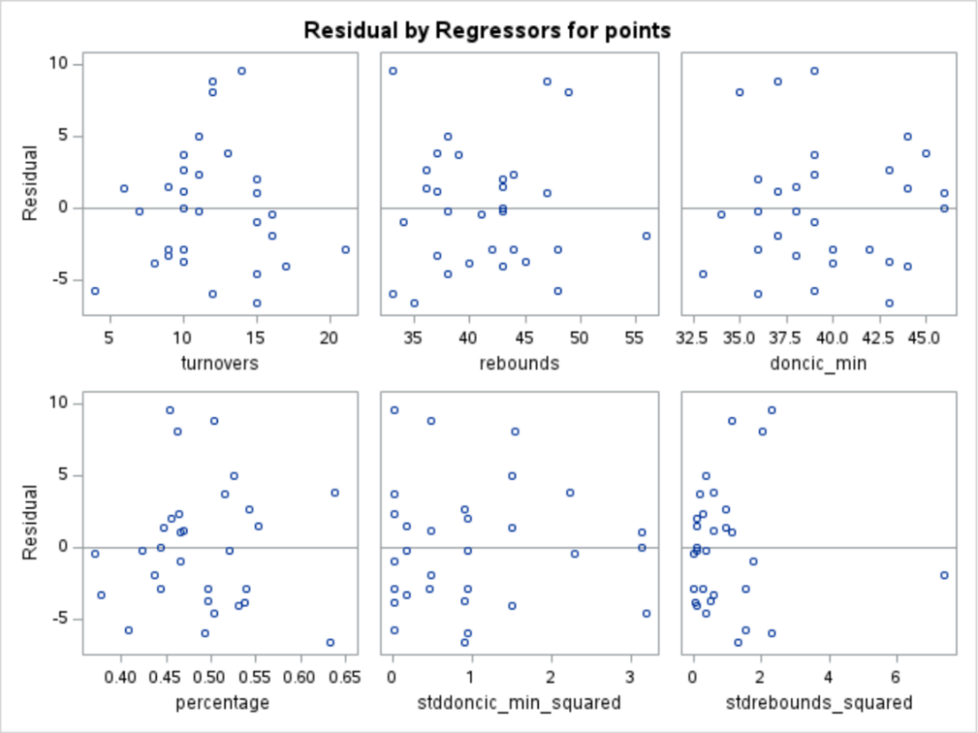
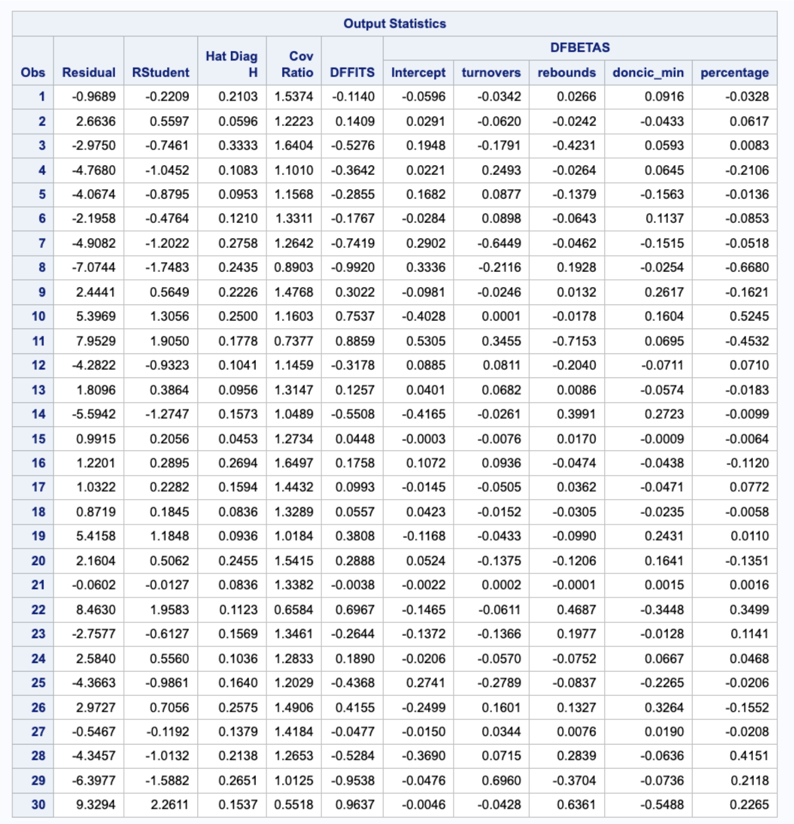


Figure 8: Plots of the predictor variables and standardized predictor variables vs the residuals

Comparing figure 8 to figure 6, there does not seem to be an improvement by including the transformed variables. Since neither rebounds nor Doncic minutes played were improved by their transformations, we can deem the curvature in figure 6 plots as unusual patterns, but they are not to be a concern. Thus, we will continue to estimate points scored per game using only turnovers, rebounds, Doncic minutes, and shot percentage.

D. Diagnostics

Next, we will check for outliers and evaluate their influence and leverage. If an observation is identified as an outlier in the X axis, it may be appropriate to restrict the X range to remove the observation if it strongly impacted the model. For outliers in total points scored in one game, we will flag these observations and study their influence on the model but if it is a valid observation, we must keep it in the model.



Output Statistics

						DFBETAS				
Obs	**Residual**	**RStudent**	**Hat Diag H**	**Cov Ratio**	**DFFITS**	**Intercept**	**turnovers**	**rebounds**	**doncic_min**	**percentage**
1	-0.9689	-0.2209	0.2103	1.5374	-0.1140	-0.0596	-0.0342	0.0266	0.0916	-0.0328
2	2.6636	0.5597	0.0596	1.2223	0.1409	0.0291	-0.0620	-0.0242	-0.0433	0.0617
3	-2.9750	-0.7461	0.3333	1.6404	-0.5276	0.1948	-0.1791	-0.4231	0.0593	0.0083
4	-4.7680	-1.0452	0.1083	1.1010	-0.3642	0.0221	0.2493	-0.0264	0.0645	-0.2106
5	-4.0674	-0.8795	0.0953	1.1568	-0.2855	0.1682	0.0877	-0.1379	-0.1563	-0.0136
6	-2.1958	-0.4764	0.1210	1.3311	-0.1767	-0.0284	0.0898	-0.0643	0.1137	-0.0853
7	-4.9082	-1.2022	0.2758	1.2642	-0.7419	0.2902	-0.6449	-0.0462	-0.1515	-0.0518
8	-7.0744	-1.7483	0.2435	0.8903	-0.9920	0.3336	-0.2116	0.1928	-0.0254	-0.6680
9	2.4441	0.5649	0.2226	1.4768	0.3022	-0.0981	-0.0246	0.0132	0.2617	-0.1621
10	5.3969	1.3056	0.2500	1.1603	0.7537	-0.4028	0.0001	-0.0178	0.1604	0.5245
11	7.9529	1.9050	0.1778	0.7377	0.8859	0.5305	0.3455	-0.7153	0.0695	-0.4532
12	-4.2822	-0.9323	0.1041	1.1459	-0.3178	0.0885	0.0811	-0.2040	-0.0711	0.0710
13	1.8096	0.3864	0.0956	1.3147	0.1257	0.0401	0.0682	0.0086	-0.0574	-0.0183
14	-5.5942	-1.2747	0.1573	1.0489	-0.5508	-0.4165	-0.0261	0.3991	0.2723	-0.0099
15	0.9915	0.2056	0.0453	1.2734	0.0448	-0.0003	-0.0076	0.0170	-0.0009	-0.0064
16	1.2201	0.2895	0.2694	1.6497	0.1758	0.1072	0.0936	-0.0474	-0.0438	-0.1120
17	1.0322	0.2282	0.1594	1.4432	0.0993	-0.0145	-0.0505	0.0362	-0.0471	0.0772
18	0.8719	0.1845	0.0836	1.3289	0.0557	0.0423	-0.0152	-0.0305	-0.0235	-0.0058
19	5.4158	1.1848	0.0936	1.0184	0.3808	-0.1168	-0.0433	-0.0990	0.2431	0.0110
20	2.1604	0.5062	0.2455	1.5415	0.2888	0.0524	-0.1375	-0.1206	0.1641	-0.1351
21	-0.0602	-0.0127	0.0836	1.3382	-0.0038	-0.0022	0.0002	-0.0001	0.0015	0.0016
22	8.4630	1.9583	0.1123	0.6584	0.6967	-0.1465	-0.0611	0.4687	-0.3448	0.3499
23	-2.7577	-0.6127	0.1569	1.3461	-0.2644	-0.1372	-0.1366	0.1977	-0.0128	0.1141
24	2.5840	0.5560	0.1036	1.2833	0.1890	-0.0206	-0.0570	-0.0752	0.0667	0.0468
25	-4.3663	-0.9861	0.1640	1.2029	-0.4368	0.2741	-0.2789	-0.0837	-0.2265	-0.0206
26	2.9727	0.7056	0.2575	1.4906	0.4155	-0.2499	0.1601	0.1327	0.3264	-0.1552
27	-0.5467	-0.1192	0.1379	1.4184	-0.0477	-0.0150	0.0344	0.0076	0.0190	-0.0208
28	-4.3457	-1.0132	0.2138	1.2653	-0.5284	-0.3690	0.0715	0.2839	-0.0636	0.4151
29	-6.3977	-1.5882	0.2651	1.0125	-0.9538	-0.0476	0.6960	-0.3704	-0.0736	0.2118
30	9.3294	2.2611	0.1537	0.5518	0.9637	-0.0046	-0.0428	0.6361	-0.5488	0.2265

Figure 9: Table that displays Outliers, Leverage, Influence

To find X outliers in each predictor, we must determine if an observation's leverage values (h_{ii}) are greater than $\frac{2\cdot p}{n} = \frac{2\cdot 5}{30} = \frac{1}{3}$. The leverage values are labeled as the column 'Hat Diag H' found in figure 9. Only observation 3 is questionable with a leverage value of exactly 1/3, a larger h_{ii} implies a higher influence on the predicted total points scored per game. Since observation 3 is not extremely higher than 1/3, we will continue to investigate this X outlier's influence and determine if it is necessary to remove this from analysis.

To identify outliers in the points scored per game, we must use the studentized delete residuals, labeled as RStudent in figure 9. Using the Bonferroni outlier test, if the absolute value of the studentized deleted residuals, $|t_i| > t\left(1 - \frac{\alpha}{2n}; n - p - 1\right) = t\left(1 - \frac{0.10}{2 \cdot 30}; 30 - 5 - 1\right) = 3.2583808824$ then we will flag the game as a Y-outlier. Since there are no studentized deleted residuals that pass this threshold from figure 9, there are no observations that were identified as y-outliers.

Since observation 3 was flagged as being a potential outlier, we need to investigate its influence on the fitted values and the individual least squares estimates of our predictors. If observation 3 has a strong influence on the predicted number of points scored in one game then the third fitted value computed with observation 3 omitted would be flagged if $|DIFFITS| > 2\sqrt{\frac{p}{n}} = 2\sqrt{\frac{5}{30}} = 0.816$. Since $|DIFFITS| = 0.5276$ for observation 3, which does not exceed 0.816, we do not need to flag observation 3 as having a strong influence on the third fitted value of points scored. If observation 3 has a strong influence on any of the least squared estimates of our predictor variables, it would be flagged by $|DFBETAS| > \frac{2}{\sqrt{n}} = \frac{2}{\sqrt{30}} = 0.365$ for any of the individual predictor variables. Since $|DFBETAS| = 0.4231$ for rebounds, this implies that observation 3 has a influence on the rebounds predictor. Considering that observation 3 is only flagged as influential for the rebounds predictor, and no other predictors, it is not considered highly influential overall.

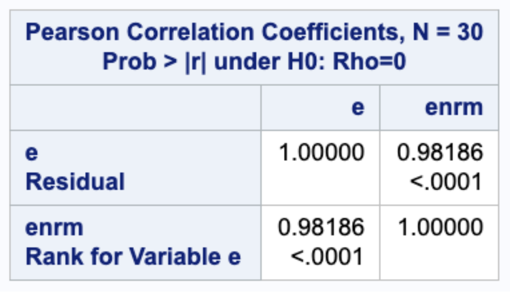
To complete our diagnostics check, we will discuss variance inflation. Variance inflation is an indicator of multicollinearity which is when predictors are highly correlated with one another. Multicollinearity makes it difficult for the model to distinguish the differences between each predictors effects and thus lead to unstable coefficient estimates.

From figure 3, all the variance inflation factors are relatively low. As a rule of thumb, we want our variance inflation factors to be below 5 and this is satisfied in this case. We can now proceed knowing that there is no serious multicollinearity issue between the predictor variables.

E. <u>Assumptions Tests</u>

In combination of the residual plot analysis, we will test for normality and use the modified-Levene test to confirm if our assumptions hold.

From figure 4, we concluded that our residuals do not follow a normal distribution but their distribution was at least symmetric. We will conduct a test for normality to further understand the distribution of the residuals



Pearson Correlation Coefficients, N = 30
Prob > |r| under H0: Rho=0

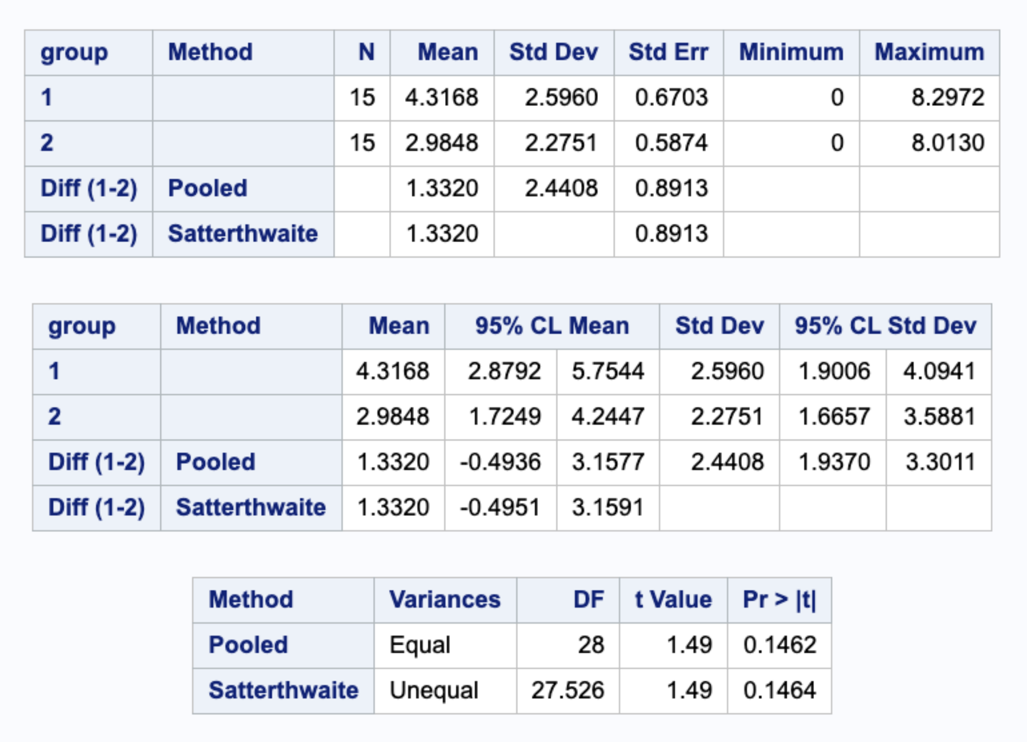
	e	enrm
e Residual	1.00000	0.98186 <.0001
enrm Rank for Variable e	0.98186 <.0001	1.00000

Figure 10

For each residual, we calculated its normal score then found the sample correlation between the residuals and its normal score. If the sample correlation is less than the cutoff $c(\alpha = 0.1, n = 30) = 0.971$, we will reject the null hypothesis that normality is reasonable. Since the sample correlation is 0.98 which is greater than the cutoff, we will fail to reject the null hypothesis that normality is reasonable. This a weak conclusion and considering the S-shaped curvature found in the normal probability plot, we will continue to assume that the residuals do not follow a normal distribution. But due to the symmetry of the distribution of the residuals, this is not an extreme impact of the validity of our model.

We will now investigate if the residuals have constant variance using the modified-Levene test. Using figure 5 for reference, we will split the data into two groups using predicted points. Residuals that are left of the median of the fitted values, which is 115 points, were put into one group and all other residuals were put into another group. Then for each group, the absolute deviations of the residuals and their group medians were obtained. First an F-test is conducted with the null hypothesis that the variances of these groups are equal, the appropriate t-test will be conducted depending on the results. A two-sample t-test was completed on theses deviations of each group. The null hypothesis of the t-test is that the means of the deviations are equal.

Regardless of whether we conclude that the variances of these groups are equal or unqual, the p-values for both t-test is beyond any common alpha level. Thus we will fail to reject the null hypothesis and we can reasonably assume that constant variance in the residuals holds.



group	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
1		15	4.3168	2.5960	0.6703	0	8.2972
2		15	2.9848	2.2751	0.5874	0	8.0130
Diff (1-2)	Pooled		1.3320	2.4408	0.8913		
Diff (1-2)	Satterthwaite		1.3320		0.8913		

group	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		4.3168	2.8792	5.7544	2.5960	1.9006	4.0941
2		2.9848	1.7249	4.2447	2.2751	1.6657	3.5881
Diff (1-2)	Pooled	1.3320	-0.4936	3.1577	2.4408	1.9370	3.3011
Diff (1-2)	Satterthwaite	1.3320	-0.4951	3.1591			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	28	1.49	0.1462
Satterthwaite	Unequal	27.526	1.49	0.1464

Figure 11

F. Preliminary Model Regression Output Analysis

Using both the residual plots and hypothesis tests, we can reasonably assume our model assumptions hold and can now evaluate the preliminary model:

$$\widehat{points} = -19.78 - 0.57(turnovers) + 0.83(rebounds) + 0.38(doncicMin) + 191.57(percentage)$$

An increase of one turnover will result in a 0.57 decrease in predicted total points scored in one game, given all other variables are held constant. An increase of one rebound will result in a 0.83 increase in predicted total points scored in one game, given all other variables are held constant. An increase of one minute in Doncic court time will result in a 0.38 increase in predicted total points scored in one game, given all other variables are held constant. An increase of one percent in successful shot percentage will result in a 191.57 increase in predicted total points scored in one game, given all other variables are held constant.

Using the ANOVA table in figure 3, the total sum of squares is 4598.17, this is total variability in total points scored across all data points when ignoring turnovers, rebounds, Doncic minutes played and shot percentage. The regression sum of squares is 4012.53 and the error sum of squares is 585.63. Thus, the variability in total points scored per game explained by the model is 4012.53 and the unexplained variability in total points scored per game is 585.63. Thus 87% of the total variability in points scored can be explained by the model with turnovers, rebounds, Doncic minutes played and shot percentage as predictors.

Using an F-test, we can evaluate if at least one predictor is significant to estimating points scored. The null hypothesis of the test is the coefficient for all the predictors are equal to 0 and the alternative hypothesis is that at least one is not equal to 0. From the table labeled Analysis of Variance in figure 3, the p-value from the F test is <.0001 which is smaller than an alpha of 0.01. Thus, we can say with 99% confidence that the regression is significant, we reject the null hypothesis that the coefficients for all predictors are equal to 0.

In the parameter estimates table from figure 3, a t-test for each predictor was completed. The null hypothesis of the t-test is that the coefficient for the individual predictor is equal to zero, the alternative hypothesis is that the coefficient for the individual predictor is not equal to zero. A coefficient of 0 means that the predictor has no linear relationship with points scored.

Doncic minutes played has a p value higher than 0.05, we will fail to reject the null. We conclude that Doncic minutes and points scored in one game do not have a linear relationship. This is not surprising since we found curvature in figure 1 in the plot of Doncic minutes and points scored.

However, all other predictors have a p value less than 0.05, meaning we will reject the null hypothesis for the t test. This suggests that turnovers, rebounds, and shot percentage have a significant linear relationship to points score in one game.

III. Exploration of Interaction Terms

We will now determine if the interaction of any two predictors in the preliminary model can be identified as significant factors in predicting total points scored in one game. The following figures will be used to identify if there is a relationship between the errors of the preliminary model and the errors of the model using the same predictors (turnovers, rebounds, Doncic minutes, shot percentage) to fit the interaction term. We will consider adding the interaction term to the model if we believe a relationship exists. These plots are partial regression plots and are beneficial because they will show if an interaction term can be used to predict points scored without the confusion of the interaction term's relationship with other predictors variables; making the interaction term look helpful when it may not be.

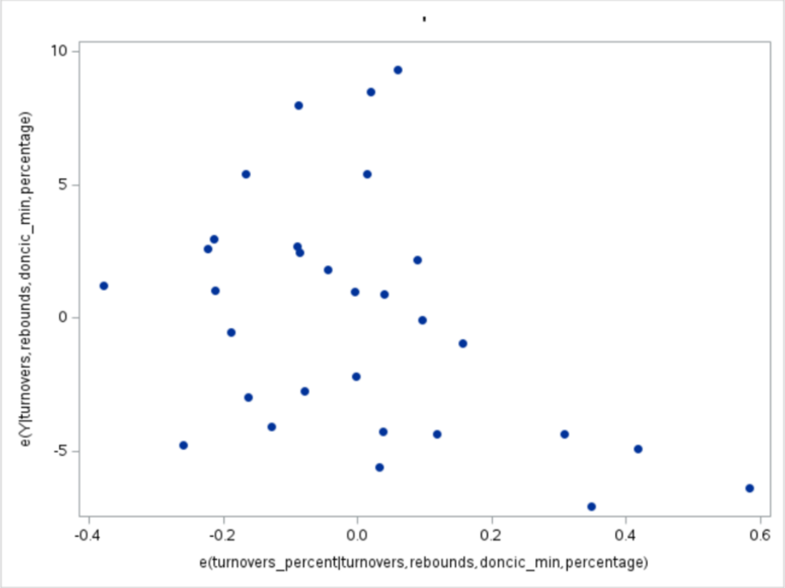


Figure 12:

Figure 12 will determine the interaction between turnovers and shot percentage. This will be helpful in predicting the total amount of points scored per game. Since there is a negative trend between the residuals, we will add the interaction between number of turnovers and shot percentage as a predictor. (note figure 12: y axis- resiudals of prelim model, x axis- residuals of model predicting the interaction term, turnovers&percentage, using same predictors as prelim model)

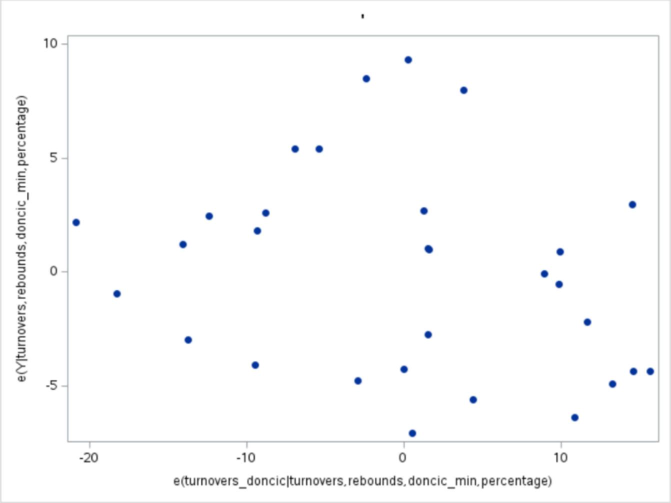


Figure 13

Figure 13 will be used to determine the interaction between the number of turnovers and Doncic minutes played. It can be helpful in predicting total points scored per game. Since there is no clear trend between the residuals and scatter, we will not add the interaction between number of turnovers and Doncic minutes played as a predictor.

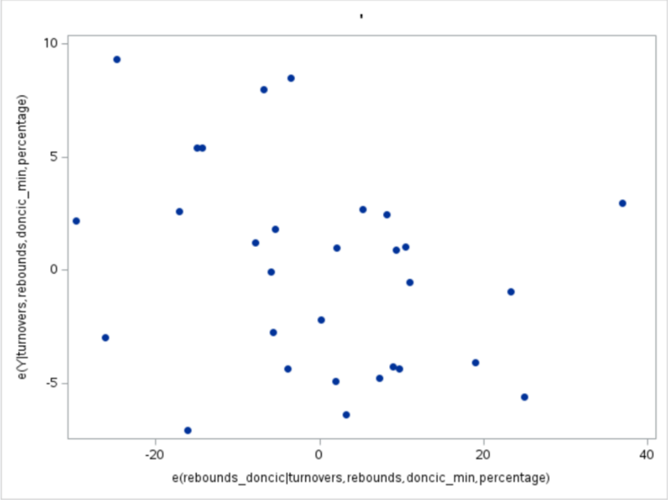


Figure 14

Figure 14 will be used to determine the interaction between the number of rebounds and Doncic minutes to see if it will be helpful in predicting total points scored per game. Since there is a negative trend between the residuals, we will add the interaction between number of rebounds and Doncic minutes as a predictor to the model.

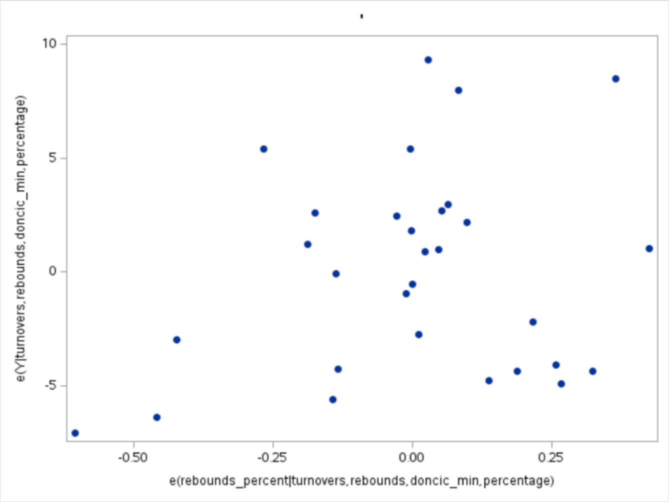


Figure 15

Figure 15 will be used to determine if the interaction between the number of rebounds and shot percentage can be helpful in predicting total points scored per game. Since there is a slight upward trend between the residuals, we will add the interaction between number of rebounds and shot percentage as a predictor to the model.

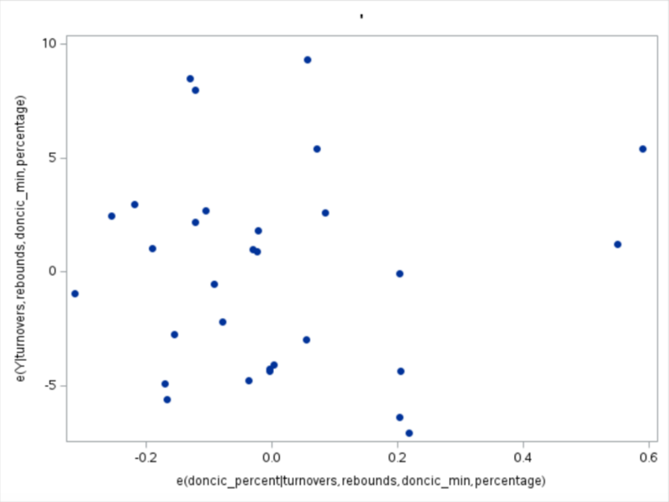


Figure 16

Figure 16 will be used to determine if the interaction between shot percentage and Doncic minutes played can be helpful in predicting total points scored per game. Since there is no clear trend between the residuals and scatter, we will not add the interaction between shot percentage and Doncic minutes played as a predictor.

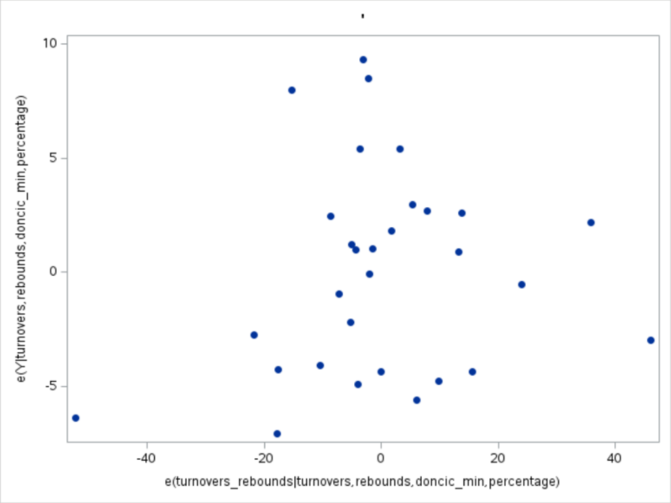
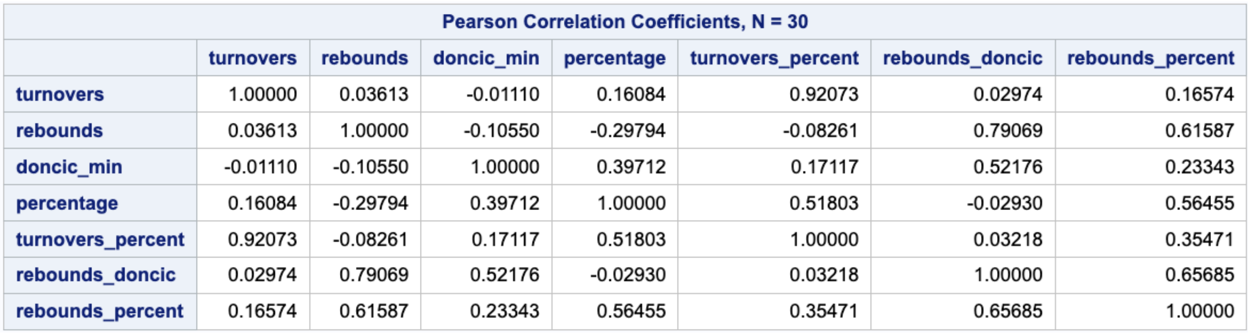


Figure 17

Figure 17 will be used to determine if the interaction between turnovers and rebounds can be helpful in predicting total points scored per game. Since there is no clear trend between the residuals and scatter, we will not add the interaction between turnovers and rebounds as a predictor.

We have determined that the interaction terms of turnovers & shot percentage, rebounds & Doncic minutes, and lastly rebounds & percentage could be beneficial to add to the preliminary model by using the partial regression plots. Next, we will determine if there are high correlations between these interaction terms and the existing predictors which would lead to an issue of multicollinearity

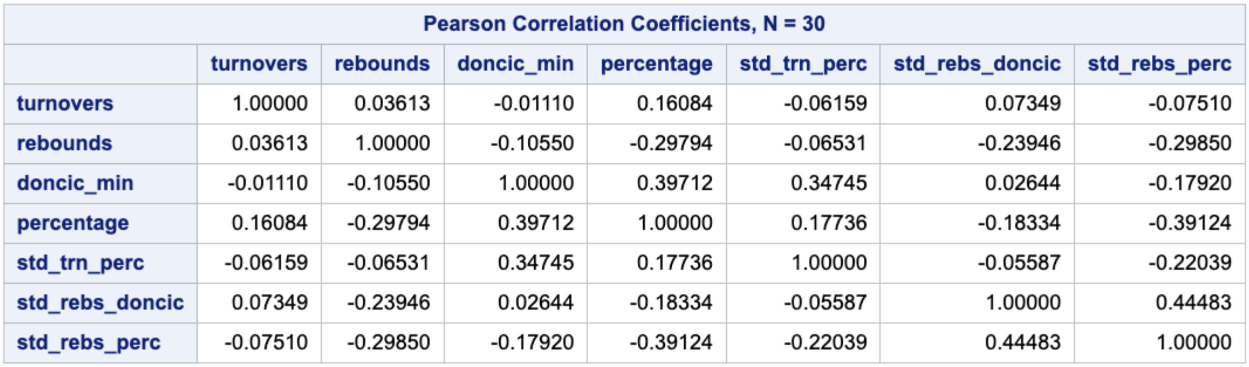


Pearson Correlation Coefficients, N = 30

	turnovers	rebounds	doncic_min	percentage	turnovers_percent	rebounds_doncic	rebounds_percent
turnovers	1.00000	0.03613	-0.01110	0.16084	0.92073	0.02974	0.16574
rebounds	0.03613	1.00000	-0.10550	-0.29794	-0.08261	0.79069	0.61587
doncic_min	-0.01110	-0.10550	1.00000	0.39712	0.17117	0.52176	0.23343
percentage	0.16084	-0.29794	0.39712	1.00000	0.51803	-0.02930	0.56455
turnovers_percent	0.92073	-0.08261	0.17117	0.51803	1.00000	0.03218	0.35471
rebounds_doncic	0.02974	0.79069	0.52176	-0.02930	0.03218	1.00000	0.65685
rebounds_percent	0.16574	0.61587	0.23343	0.56455	0.35471	0.65685	1.00000

Figure 18:

From figure 18, the correlations between turnovers & the turnovers percent interaction is over 0.7, indicating that multicollinearity will be an issue. This also applies to the correlations between rebounds & rebounds Doncic minutes interaction. We will standardize turnovers, rebounds, Doncic minutes, and percentage and use these standardized variables to investigate if the interaction terms still have high correlations with the existing predictors of the preliminary model.



Pearson Correlation Coefficients, N = 30

	turnovers	rebounds	doncic_min	percentage	std_trn_perc	std_rebs_doncic	std_rebs_perc
turnovers	1.00000	0.03613	-0.01110	0.16084	-0.06159	0.07349	-0.07510
rebounds	0.03613	1.00000	-0.10550	-0.29794	-0.06531	-0.23946	-0.29850
doncic_min	-0.01110	-0.10550	1.00000	0.39712	0.34745	0.02644	-0.17920
percentage	0.16084	-0.29794	0.39712	1.00000	0.17736	-0.18334	-0.39124
std_trn_perc	-0.06159	-0.06531	0.34745	0.17736	1.00000	-0.05587	-0.22039
std_rebs_doncic	0.07349	-0.23946	0.02644	-0.18334	-0.05587	1.00000	0.44483
std_rebs_perc	-0.07510	-0.29850	-0.17920	-0.39124	-0.22039	0.44483	1.00000

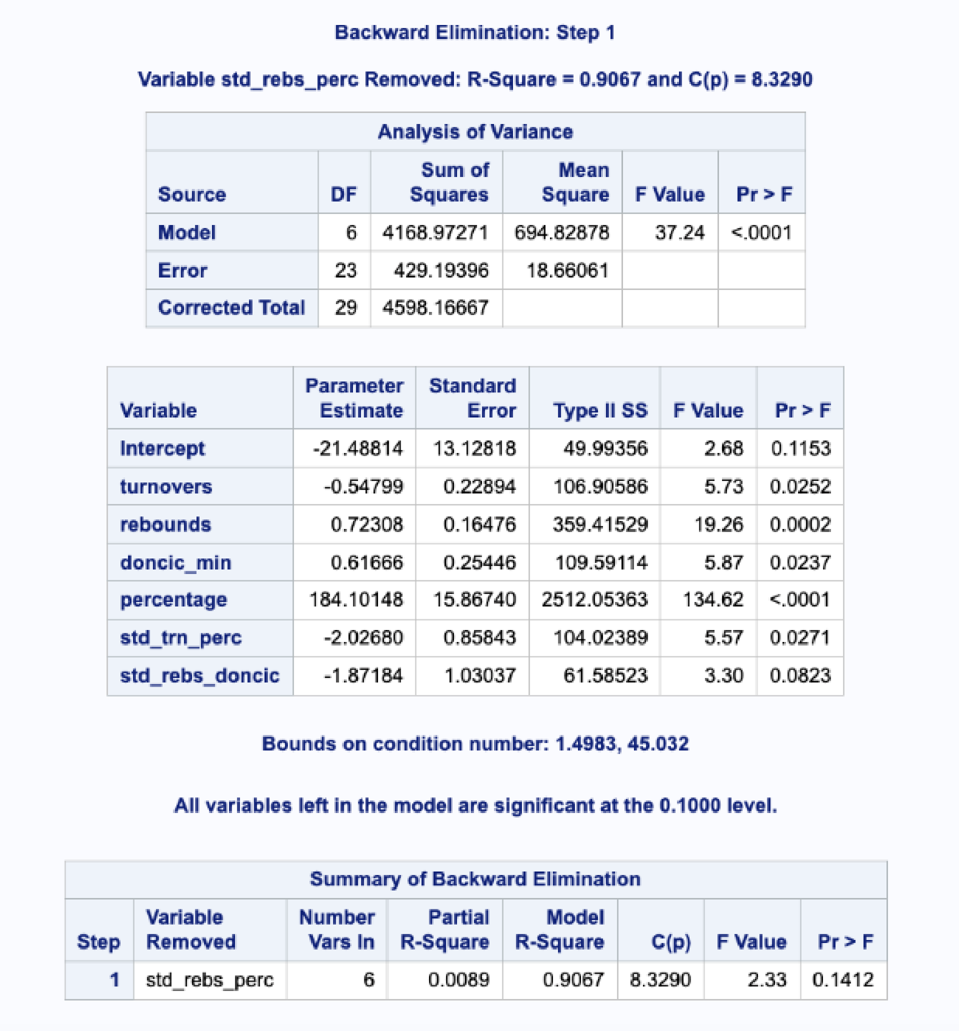
Figure 19

Figure 19 presents the correlations of the existing predictors and the standardized interactions terms. The standardized interaction terms, turnovers & shot percentage, rebounds & Doncic minutes, rebounds & percentage, no longer have high correlations with any of the other predictor variables so multicollinearity will not be an issue. Analysis will be continued by including the standardized interaction terms.

IV. Model Search

A. Backward Deletion

Backward deletion is a model search method in which we begin with the full model including old and new predictor variables: turnovers, rebounds, percentage, Doncic minutes, turnovers & percentage interaction standardized, rebounds & Doncic minutes interaction standardized, and rebounds & percentage interaction standardized. Using this method will help us determine what predictor variables need to be removed based on their p-value. If a predictor's variable is higher than the alpha (which in our case is .10), then it must be removed. When using the Backward deletion method, you can only remove one predictor variable at a time, so the highest p-value that surpasses the alpha value will be eliminated. Once eliminated, we must re-run the regression to get new p-values. This process continues until all remaining predictor variables are significant.



Backward Elimination: Step 1

Variable std_rebs_perc Removed: R-Square = 0.9067 and C(p) = 8.3290

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	4168.97271	694.82878	37.24	<.0001
Error	23	429.19396	18.66061		
Corrected Total	29	4598.16667			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-21.48814	13.12818	49.99356	2.68	0.1153
turnovers	-0.54799	0.22894	106.90586	5.73	0.0252
rebounds	0.72308	0.16476	359.41529	19.26	0.0002
doncic_min	0.61666	0.25446	109.59114	5.87	0.0237
percentage	184.10148	15.86740	2512.05363	134.62	<.0001
std_trn_perc	-2.02680	0.85843	104.02389	5.57	0.0271
std_rebs_doncic	-1.87184	1.03037	61.58523	3.30	0.0823

Bounds on condition number: 1.4983, 45.032

All variables left in the model are significant at the 0.1000 level.

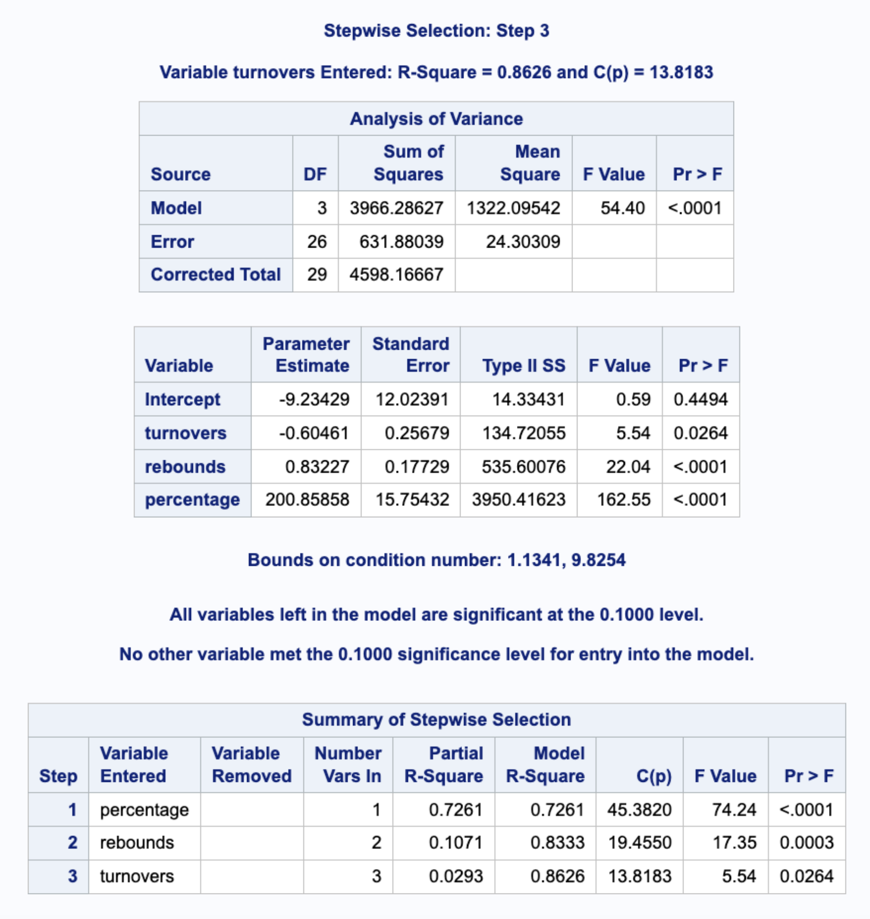
Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	std_rebs_perc	6	0.0089	0.9067	8.3290	2.33	0.1412

Figure 20: Backward Elimination Method

Figure 20 shows that the standardized rebound times field goal percentage predictor variable was removed because its p-value was higher than the alpha (.10). Therefore, the final model chosen by backward deletion will include all variables- turnovers, rebounds, Doncic minutes, percentage, standardized turnovers & percentage, and standardized rebounds & Doncic minutes- will be used to model points scored per game. 91% of the variance in points scored can be explained by the model. The p-value for the F-test is less than 0.1 which suggests that the model is statistically significant. Additionally, the Mallow's C(p)=8.3 which is close to 7, which is the number of predictors including the intercept, suggesting that there are most likely no important predictors missing to model points scored per game.

B. **Stepwise Regression**

Stepwise regression is another model search technique. It begins with no predictors and continuously adds or deletes predictor variables using their p-values. Again, we will only be keeping predictors that are significant at the 0.10 level. At each step it will either add variables with the smallest p-value or delete any variables in the model that exceed 0.10.



Stepwise Selection: Step 3

Variable turnovers Entered: R-Square = 0.8626 and C(p) = 13.8183

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3966.28627	1322.09542	54.40	<.0001
Error	26	631.88039	24.30309		
Corrected Total	29	4598.16667			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-9.23429	12.02391	14.33431	0.59	0.4494
turnovers	-0.60461	0.25679	134.72055	5.54	0.0264
rebounds	0.83227	0.17729	535.60076	22.04	<.0001
percentage	200.85858	15.75432	3950.41623	162.55	<.0001

Bounds on condition number: 1.1341, 9.8254

All variables left in the model are significant at the 0.1000 level.

No other variable met the 0.1000 significance level for entry into the model.

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	percentage		1	0.7261	0.7261	45.3820	74.24	<.0001
2	rebounds		2	0.1071	0.8333	19.4550	17.35	0.0003
3	turnovers		3	0.0293	0.8626	13.8183	5.54	0.0264

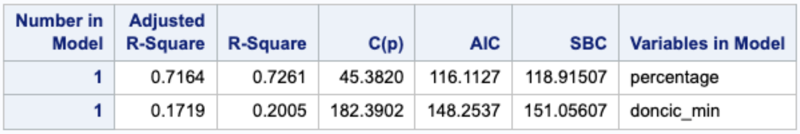
Figure 21: Stepwise Regression Method

The model chose by stepwise regression added percentage, rebounds and turnovers to predict points scored per game. No other variables were added or deleted. The model explains 86% of the variance in points scored per game. The small p-value for the F-test indicates that the regression is statistically significant at the 0.10 level. The Mallow's C(p)=13.8 which is relatively high considering there are only 4 parameters, so there may be important predictors of score missing in this model.

C. <u>Best Subsets</u>

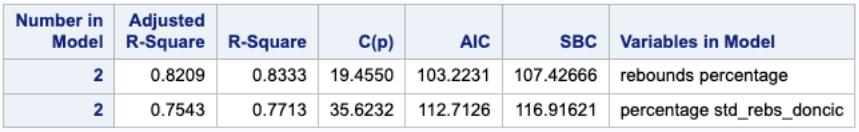
The Best Subsets method, also known as All Possible Regressions method, is like Backwards Deletion, we will start with the full set of data. The difference between the two is that the Best Subset method looks at every possible subset of predictor variables and does the regression on the subset. For example, the regression will be performed on all the 1 predictor models, 2 predictor models, and continues until you reach the maximum number of predictors in a model. The Best Subsets method helps select multiple good models. There are 5 ways we can identify the "best" models. One way is to look at the R^2 and SSE values. We want the R^2 value to be high and the SSE value to be low. Another way that is very similar to the first is to look at the adjusted R^2 and MSE values. As with the first way, we want our adjusted R^2 value to be high but want the MSE value to be low. As we increase the number of parameters in the subset, R^2 and adjusted R^2 will continue to increase and eventually level off.

Figures 22-27 shows the subsets for the 1 predictor models through the 6 predictor models.



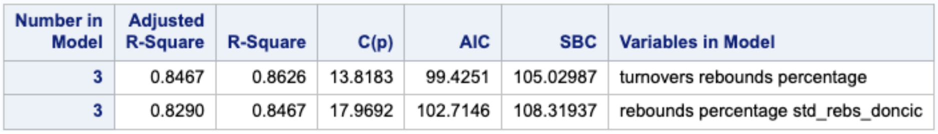
Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
1	0.7164	0.7261	45.3820	116.1127	118.91507	percentage
1	0.1719	0.2005	182.3902	148.2537	151.05607	doncic_min

Figure 22: Best Subsets, 1 Predictor Models (1a &1b)



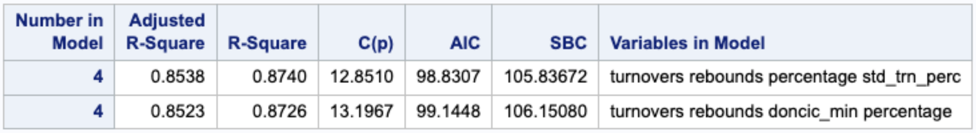
Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
2	0.8209	0.8333	19.4550	103.2231	107.42666	rebounds percentage
2	0.7543	0.7713	35.6232	112.7126	116.91621	percentage std_rebs_doncic

Figure 23: Best Subsets, 2 Predictor Models (2a &2b)



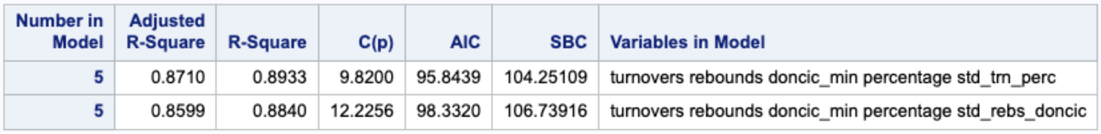
Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
3	0.8467	0.8626	13.8183	99.4251	105.02987	turnovers rebounds percentage
3	0.8290	0.8467	17.9692	102.7146	108.31937	rebounds percentage std_rebs_doncic

Figure 24: Best Subsets, 3 Predictor Models (3a &3b)



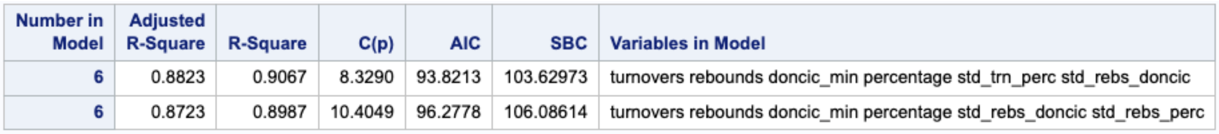
Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
4	0.8538	0.8740	12.8510	98.8307	105.83672	turnovers rebounds percentage std_trn_perc
4	0.8523	0.8726	13.1967	99.1448	106.15080	turnovers rebounds doncic_min percentage

Figure 25: Best Subsets, 4 Predictor Models (4a &4b)



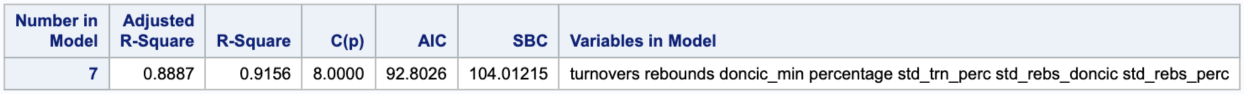
Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
5	0.8710	0.8933	9.8200	95.8439	104.25109	turnovers rebounds doncic_min percentage std_trn_perc
5	0.8599	0.8840	12.2256	98.3320	106.73916	turnovers rebounds doncic_min percentage std_rebs_doncic

Figure 26: Best Subsets, 5 Predictor Models (5a &5b)



Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
6	0.8823	0.9067	8.3290	93.8213	103.62973	turnovers rebounds doncic_min percentage std_trn_perc std_rebs_doncic
6	0.8723	0.8987	10.4049	96.2778	106.08614	turnovers rebounds doncic_min percentage std_rebs_doncic std_rebs_perc

Figure 27: Best Subsets, 6 Predictor Models (6a &6b)



Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
7	0.8887	0.9156	8.0000	92.8026	104.01215	turnovers rebounds doncic_min percentage std_trn_perc std_rebs_doncic std_rebs_perc

Figure 28: Best Subsets, 7 Predictor Model

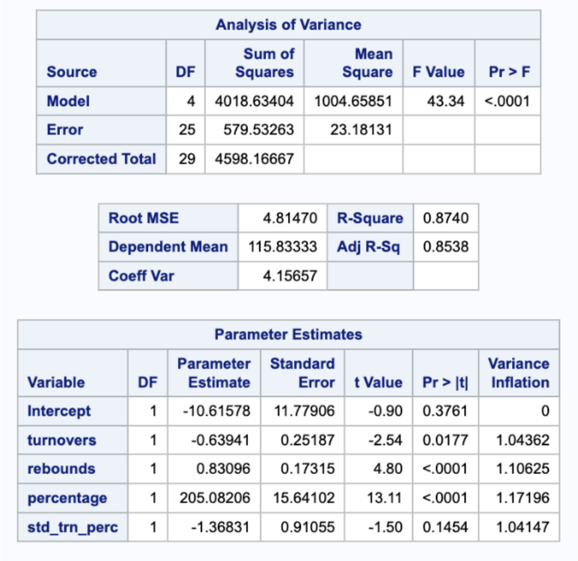
Considering that models presented by best subsets with one predictor, two predictors and three predictors all have high Mallow's C(p) values, this implies that they are missing important predictors of points scored and thus are too simple.

Looking at the 4 predictor models, the adjusted R^2 values have slightly increased. Although, the Mallow's C(p), AIC and SBC values have not decreased very much compared to 3a. For the 5 predictor models, model 5a seems to have a slight improvement compared to the other models that we've observed at this point.

Model 6a has the best Mallow's C(p) value of 8.3 which is close to the number of predictors which is 6. Model 6a also has the lowest AIC and SBC scores, other than the full model. Although the full model in figure 28 has a great Mallow's C(p), SIC, and SBC, the adjusted R^2 does not improve much compared to the 6 predictor models.

D. Candidate Models

We have evaluated many models using backward deletion, stepwise selection and best subsets regression. Now, we must choose what we consider to be potentially the best models in predicting total points scored. These models will be evaluated on a variety of different factors including adjusted R^2, Mallow's C(p), and both their F and t-tests. We will be using adjusted R^2 because we will be comparing models with varying amounts of predictors.



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	4018.63404	1004.65851	43.34	<.0001
Error	25	579.53263	23.18131		
Corrected Total	29	4598.16667			

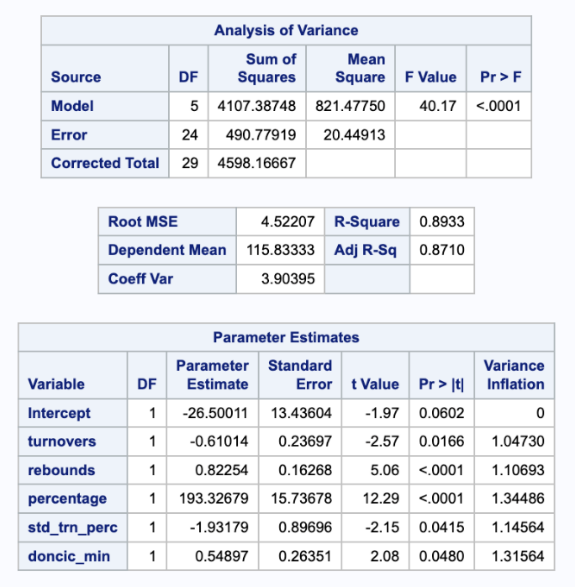
Root MSE	4.81470	R-Square	0.8740
Dependent Mean	115.83333	Adj R-Sq	0.8538
Coeff Var	4.15657		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-10.61578	11.77906	-0.90	0.3761	0
turnovers	1	-0.63941	0.25187	-2.54	0.0177	1.04362
rebounds	1	0.83096	0.17315	4.80	<.0001	1.10625
percentage	1	205.08206	15.64102	13.11	<.0001	1.17196
std_trn_perc	1	-1.36831	0.91055	-1.50	0.1454	1.04147

Figure 29: Model 4a

Using best subsets with 4 predictor variables, figure 25 presented the model using turnovers, rebounds, shot percentage and the standardized interaction between turnovers and percentage (std_trn_perc) as predictors of points scored. Initially we believed this was a strong candidate due to it's adjusted R^2 and slightly improved Mallow's C(p) score compared to the previous model in figure 29. But, after assessing the t-tests of each of the predictors, it was discovered that interaction term between turnovers and percentage was not significant at the 0.1 level. Therefore, we will not be proceeding with this model.



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	4107.38748	821.47750	40.17	<.0001
Error	24	490.77919	20.44913		
Corrected Total	29	4598.16667			

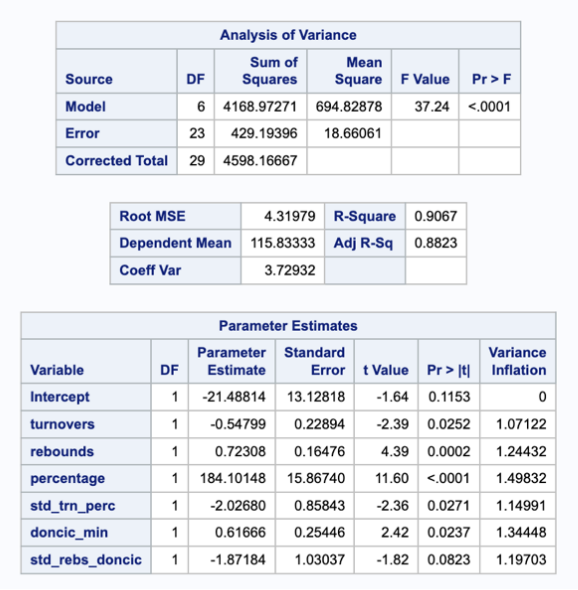
Root MSE	4.52207	R-Square	0.8933
Dependent Mean	115.83333	Adj R-Sq	0.8710
Coeff Var	3.90395		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-26.50011	13.43604	-1.97	0.0602	0
turnovers	1	-0.61014	0.23697	-2.57	0.0166	1.04730
rebounds	1	0.82254	0.16268	5.06	<.0001	1.10693
percentage	1	193.32679	15.73678	12.29	<.0001	1.34486
std_trn_perc	1	-1.93179	0.89696	-2.15	0.0415	1.14564
doncic_min	1	0.54897	0.26351	2.08	0.0480	1.31564

Figure 30: Model 5a

Best subsets with 5 predictors in figure 26 presents the model with turnovers, rebounds, percentage, Doncic minutes and the standardization interaction between turnovers and percentage as predictors of points. 87% of the variability in points is explained by the model with the 5 predictors (accounting for the number predictors and sample size) which implies it could be considered a strong candidate model. Using the F-test, we determined the regression is statistically significant at the 0.1 level. Also, the Mallow's C(p) of 9.8 is fairly good considering there are 5 predictors.

The t-tests conducted on each of the three predictors indicate they are valuable when modeling points scored since they are statistically significant at the 0.1 level. This is an interesting result considering that the previous model has the standardized interaction between turnovers and percentage as an insignificant predictor. The variance inflation factors being less than 5 indicate that multicollinearity will not be an issue.



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	4168.97271	694.82878	37.24	<.0001
Error	23	429.19396	18.66061		
Corrected Total	29	4598.16667			

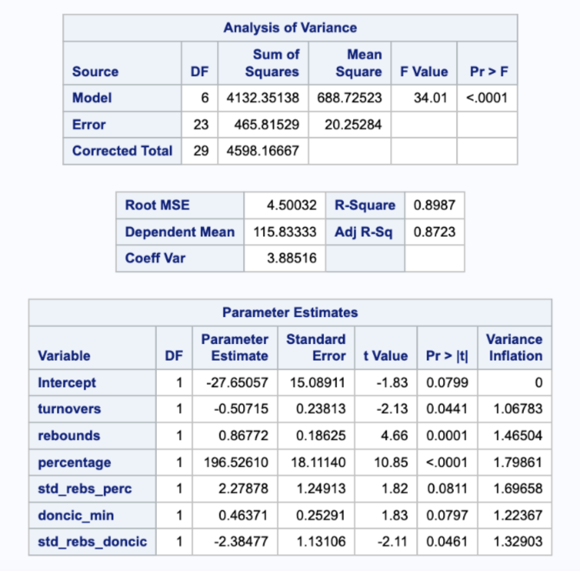
Root MSE	4.31979	R-Square	0.9067
Dependent Mean	115.83333	Adj R-Sq	0.8823
Coeff Var	3.72932		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-21.48814	13.12818	-1.64	0.1153	0
turnovers	1	-0.54799	0.22894	-2.39	0.0252	1.07122
rebounds	1	0.72308	0.16476	4.39	0.0002	1.24432
percentage	1	184.10148	15.86740	11.60	<.0001	1.49832
std_trn_perc	1	-2.02680	0.85843	-2.36	0.0271	1.14991
doncic_min	1	0.61666	0.25446	2.42	0.0237	1.34448
std_rebs_doncic	1	-1.87184	1.03037	-1.82	0.0823	1.19703

Figure 31: Model 6a

We will evaluate Model 6a that uses turnovers, rebounds, percentage, Doncic minutes, the standardized interaction of turnovers and percentage (std_trn_perc), and the standardized interaction of rebounds and Doncic minutes (std_rebs_doncic) to predict total points scored. This model was presented by the best subsets method for models with 6 predictors, seen in figure 27 and backward deletion in figure 20. 88% of the variability in points is explained by the model with turnovers, rebounds, percentage, Doncic minutes, std_trn_perc, and std_rebs_doncic (accounting for the number predictors and sample size) which implies it could be considered a strong candidate model. Using the F-test, we determined the regression is statistically significant at the 0.1 level. The Mallow's C(p)=8.3 is also the strongest yet since it is very close to 6, which is the number of predictors.

Additionally, the t-tests conducted on each of the six predictors indicate they are valuable to modeling points scored since they are statistically significant at the 0.1 level. This further suggests that std_trn_perc is on the borderline of being a helpful predictor of points scored. The variance inflation factors being less than 5 indicate that multicollinearity will not be an issue. The biggest concern for this model is its complexity.



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	4132.35138	688.72523	34.01	<.0001
Error	23	465.81529	20.25284		
Corrected Total	29	4598.16667			

Root MSE	4.50032	R-Square	0.8987
Dependent Mean	115.83333	Adj R-Sq	0.8723
Coeff Var	3.88516		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-27.65057	15.08911	-1.83	0.0799	0
turnovers	1	-0.50715	0.23813	-2.13	0.0441	1.06783
rebounds	1	0.86772	0.18625	4.66	0.0001	1.46504
percentage	1	196.52610	18.11140	10.85	<.0001	1.79861
std_rebs_perc	1	2.27878	1.24913	1.82	0.0811	1.69658
doncic_min	1	0.46371	0.25291	1.83	0.0797	1.22367
std_rebs_doncic	1	-2.38477	1.13106	-2.11	0.0461	1.32903

Figure 32: Model 6b

We will evaluate the model that uses turnovers, rebounds, percentage, Doncic minutes, the standardized interaction of rebounds and percentage (std_rebs_perc), and the standardized interaction of rebounds and Doncic minutes (std_rebs_doncic) to predict total points scored. This model was presented by the best subsets method for models with 6 predictors, seen in figure 27. 87% of the variability in points is explained by the model with turnovers, rebounds, percentage, Doncic minutes, std_rebs_perc, and std_rebs_doncic (accounting for the number predictors and sample size) which implies it could be considered a strong candidate model. Using the F-test, we determined the regression is statistically significant at the 0.1 level. The Mallow's C(p)=10.4 which is not as good as model 6a but still good for having 6 predictors.

The t-tests conducted on each of the six predictors indicate they are valuable to modeling points scored since they are statistically significant at the 0.1 level. The variance inflation factors being less than 5 indicate that multicollinearity will not be an issue. The biggest concern for this model is its complexity.

Going forward we will be proceeding with model 6a and 5a. Model 6b will not be considered going forward because it does not outperform 6a's Mallow's C(p) or adjusted R^2 even with an additional predictor variable.

V. Model Selection

Two candidate models have been chosen as arguably the best models of total points scored per game by the Dallas Mavericks. Before selecting the final model, we must check if the model assumptions are satisfied for each model.

A. <u>Model 5a Assumptions</u>

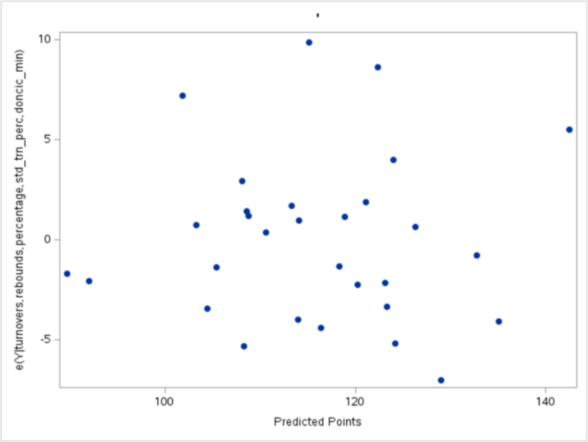


Figure 33:

From figure 33, the residuals are randomly scattered around 0. There does not appear to be a funnel shape which would indicate that we can reasonably assume the residuals have constant variance.

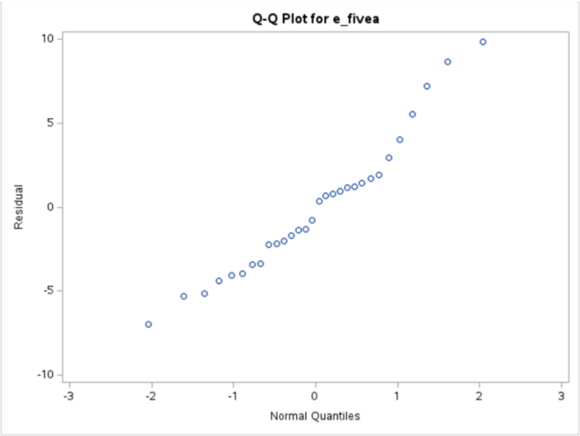


Figure 34: Normal Probability Plot of Model 5a

From figure 34, the normal probability plot appears to have a slight S-shape. This curvature in the plot implies that the distribution of residuals does not follow a normal distribution instead it has shorter tails than the normal distribution. While the residuals are not normally distributed, they do have a symmetrical distribution, like the normal distribution.

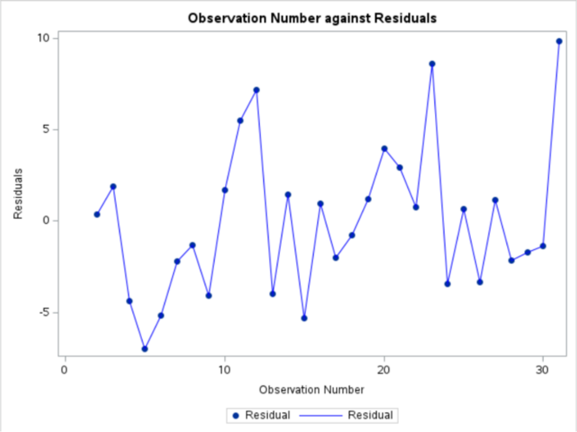


Figure 35

Figure 35 shows that there is random jaggedness and no serial correlation between the residuals. Therefore, we can reasonably assume that the residuals are uncorrelated. There is no increase in the total points per game for the Mavs due to time.

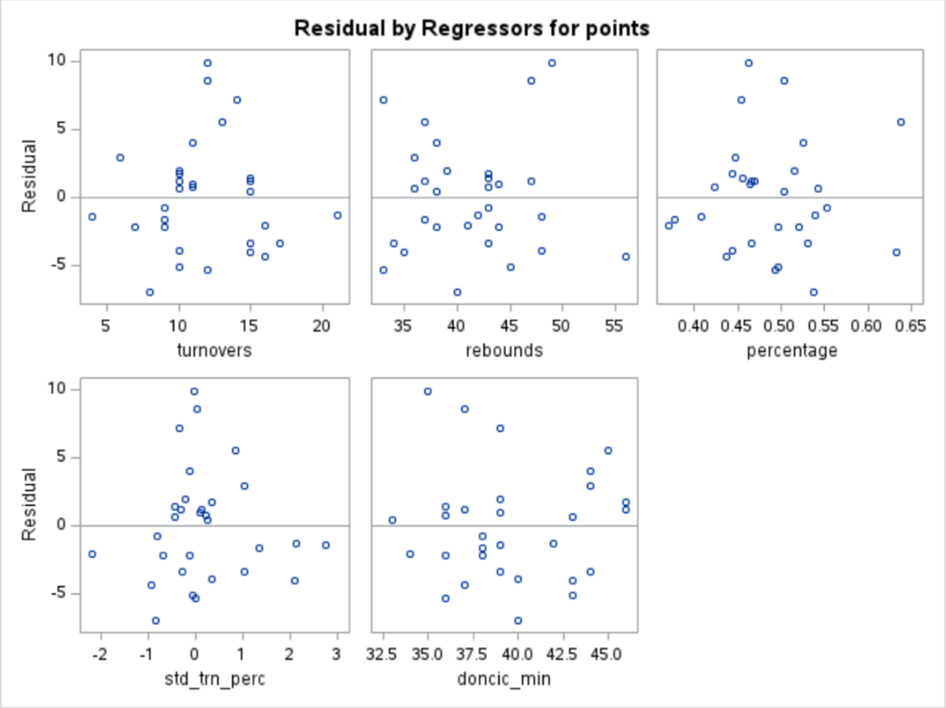
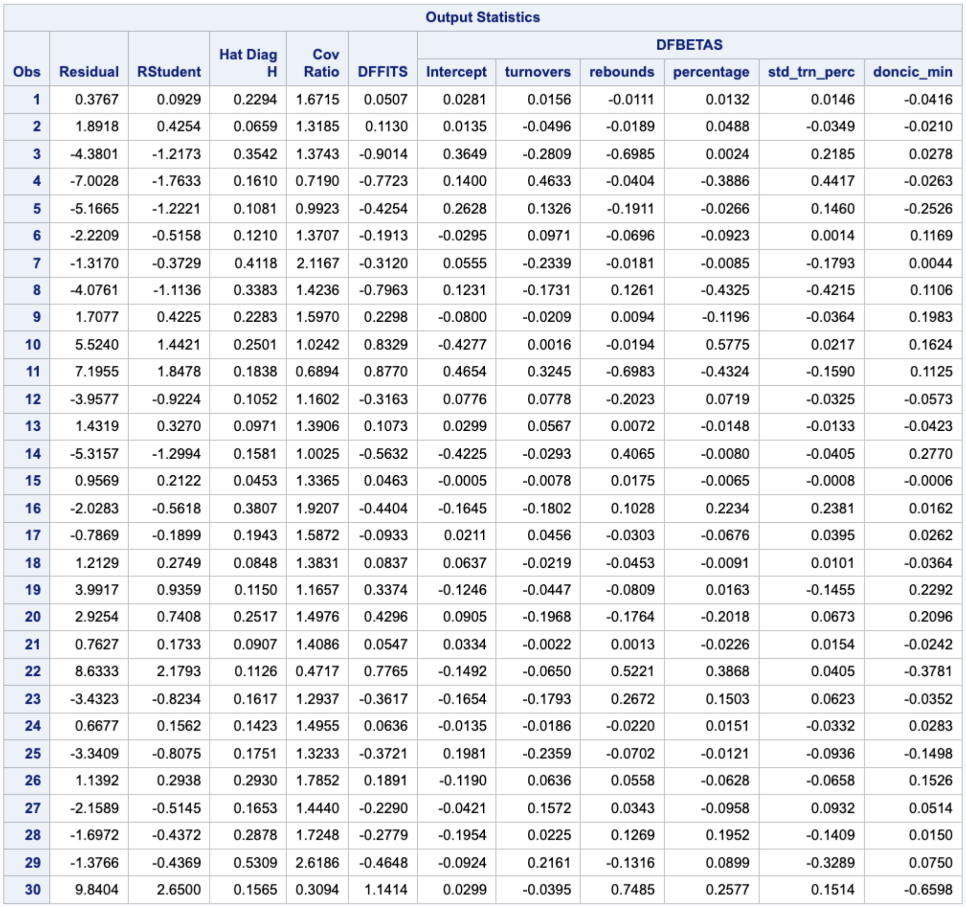


Figure 36

We must check that the multiple linear regression model form is not violated. Using the plots in figure 36, if there is curvature between any of the predictor variables and the residuals of model 5a, this would be a violation of this assumption.

The only concern would be the residuals versus Doncic minutes plot, there is possible curvature with a U-shape around 40 minutes. This curvature is nearly identical to the curvature found in the plot of residuals versus Doncic minutes for the preliminary model in figure 6. In the preliminary model, we attempted a transformation to fix this curvature, but it did not improve this issue as seen in figure 8. We can consider this dip or U-shape this to be an odd artifact as we did for the preliminary model. There does not appear to be curvature in any of the plots in figure 36, thus we can reasonably assume that the multiple linear regression model form is not violated.



Output Statistics

						DFBETAS					
Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	Intercept	turnovers	rebounds	percentage	std_trn_perc	doncic_min
1	0.3767	0.0929	0.2294	1.6715	0.0507	0.0281	0.0156	-0.0111	0.0132	0.0146	-0.0416
2	1.8918	0.4254	0.0659	1.3185	0.1130	0.0135	-0.0496	-0.0189	0.0488	-0.0349	-0.0210
3	-4.3801	-1.2173	0.3542	1.3743	-0.9014	0.3649	-0.2809	-0.6985	0.0024	0.2185	0.0278
4	-7.0028	-1.7633	0.1610	0.7190	-0.7723	0.1400	0.4633	-0.0404	-0.3886	0.4417	-0.0263
5	-5.1665	-1.2221	0.1081	0.9923	-0.4254	0.2628	0.1326	-0.1911	-0.0266	0.1460	-0.2526
6	-2.2209	-0.5158	0.1210	1.3707	-0.1913	-0.0295	0.0971	-0.0696	-0.0923	0.0014	0.1169
7	-1.3170	-0.3729	0.4118	2.1167	-0.3120	0.0555	-0.2339	-0.0181	-0.0085	-0.1793	0.0044
8	-4.0761	-1.1136	0.3383	1.4236	-0.7963	0.1231	-0.1731	0.1261	-0.4325	-0.4215	0.1106
9	1.7077	0.4225	0.2283	1.5970	0.2298	-0.0800	-0.0209	0.0094	-0.1196	-0.0364	0.1983
10	5.5240	1.4421	0.2501	1.0242	0.8329	-0.4277	0.0016	-0.0194	0.5775	0.0217	0.1624
11	7.1955	1.8478	0.1838	0.6894	0.8770	0.4654	0.3245	-0.6983	-0.4324	-0.1590	0.1125
12	-3.9577	-0.9224	0.1052	1.1602	-0.3163	0.0776	0.0778	-0.2023	0.0719	-0.0325	-0.0573
13	1.4319	0.3270	0.0971	1.3906	0.1073	0.0299	0.0567	0.0072	-0.0148	-0.0133	-0.0423
14	-5.3157	-1.2994	0.1581	1.0025	-0.5632	-0.4225	-0.0293	0.4065	-0.0080	-0.0405	0.2770
15	0.9569	0.2122	0.0453	1.3365	0.0463	-0.0005	-0.0078	0.0175	-0.0065	-0.0008	-0.0006
16	-2.0283	-0.5618	0.3807	1.9207	-0.4404	-0.1645	-0.1802	0.1028	0.2234	0.2381	0.0162
17	-0.7869	-0.1899	0.1943	1.5872	-0.0933	0.0211	0.0456	-0.0303	-0.0676	0.0395	0.0262
18	1.2129	0.2749	0.0848	1.3831	0.0837	0.0637	-0.0219	-0.0453	-0.0091	0.0101	-0.0364
19	3.9917	0.9359	0.1150	1.1657	0.3374	-0.1246	-0.0447	-0.0809	0.0163	-0.1455	0.2292
20	2.9254	0.7408	0.2517	1.4976	0.4296	0.0905	-0.1968	-0.1764	-0.2018	0.0673	0.2096
21	0.7627	0.1733	0.0907	1.4086	0.0547	0.0334	-0.0022	0.0013	-0.0226	0.0154	-0.0242
22	8.6333	2.1793	0.1126	0.4717	0.7765	-0.1492	-0.0650	0.5221	0.3868	0.0405	-0.3781
23	-3.4323	-0.8234	0.1617	1.2937	-0.3617	-0.1654	-0.1793	0.2672	0.1503	0.0623	-0.0352
24	0.6677	0.1562	0.1423	1.4955	0.0636	-0.0135	-0.0186	-0.0220	0.0151	-0.0332	0.0283
25	-3.3409	-0.8075	0.1751	1.3233	-0.3721	0.1981	-0.2359	-0.0702	-0.0121	-0.0936	-0.1498
26	1.1392	0.2938	0.2930	1.7852	0.1891	-0.1190	0.0636	0.0558	-0.0628	-0.0658	0.1526
27	-2.1589	-0.5145	0.1653	1.4440	-0.2290	-0.0421	0.1572	0.0343	-0.0958	0.0932	0.0514
28	-1.6972	-0.4372	0.2878	1.7248	-0.2779	-0.1954	0.0225	0.1269	0.1952	-0.1409	0.0150
29	-1.3766	-0.4369	0.5309	2.6186	-0.4648	-0.0924	0.2161	-0.1316	0.0899	-0.3289	0.0750
30	9.8404	2.6500	0.1565	0.3094	1.1414	0.0299	-0.0395	0.7485	0.2577	0.1514	-0.6598

Figure 37

Using figure 37, we will first asses if we can detect any x-outliers. The leverage values from figure 37 are labeled in the column 'H Diag H', if any of these values exceed $\frac{2p}{n} = \frac{2*6}{30} = 0.4$ then that observation will be considered an x-outlier. Observation 7 has an leverage score of 0.4118 and observation 29 has a leverage scored of 0.5309, now we need to evaluate their influence.

If the absolute value of DIFFITS for observation 7 or observation 29 exceed $2\sqrt{\frac{p}{n}} = 2\sqrt{\frac{6}{30}} =$ 0.894 then we will consider it having high influence on the fitted values of total points scored. Since neither has a DIFFITS value that exceeds 0.894, thus we say that neither observation has a high influence on the fitted values.

If the absolute value of DFBETAs for observation 7 or observation 29 exceed $\frac{2}{\sqrt{n}} = \frac{2}{\sqrt{30}} = 0.365$ then it has a high influence on the corresponding least squares estimate for a particular parameter. For both observations, their DFBETA values do not exceed 0.365 and we can say that they are not influential on any of the least square estimates of the parameters.

Figure 37 can also be used to detect any outliers in total points scored using the Bonferroni outlier test. Using this outlier test, if the absolute value of numbers in figure 37 in the column labeled 'RStudent' are greater than $t\left(1-\frac{\alpha}{2n}; n-p-1\right) = t\left(1-\frac{0.1}{2*30}; 30-6-1\right) =$ $t(0.99833{,}23) = 3.274$ then the observation is considered a y-outlier. Since none of the absolute values of the studentized deleted residuals in figure 27 exceed 3.274, we did not detect any y-outliers.

B. Model 6a Assumptions

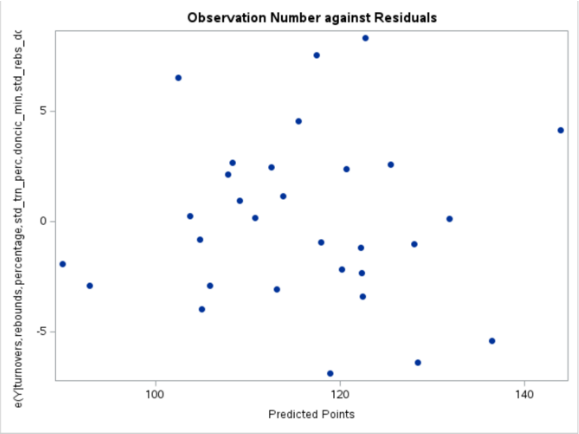


Figure 38

From figure 38, the residuals are randomly scattered around 0. There appears to be no funnel shape which would indicate that we can reasonably assume the residuals have constant variance.

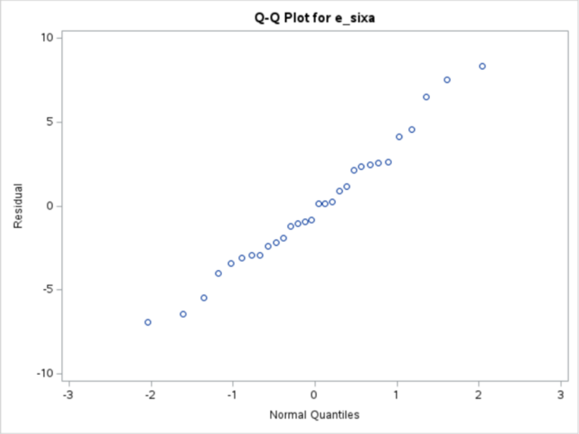


Figure 39

Figure 39 shows a mostly straight line, which suggests that residuals do follow a normal distribution. This is an improvement from model 5a.

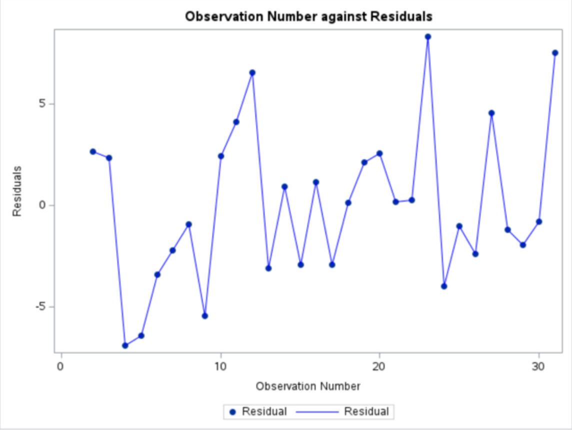


Figure 40

Figure 40 shows that there is random jaggedness and no serial correlation between the residuals. Therefore, we can reasonably assume that the residuals are uncorrelated. There is no increase in the total points per game for the Mavs due to time.

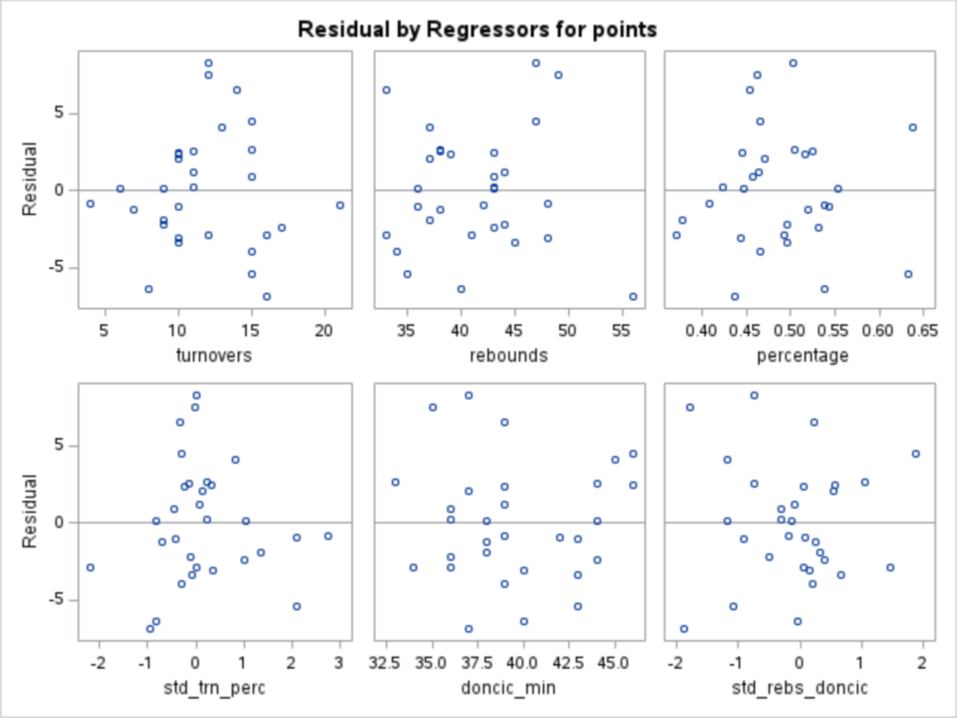
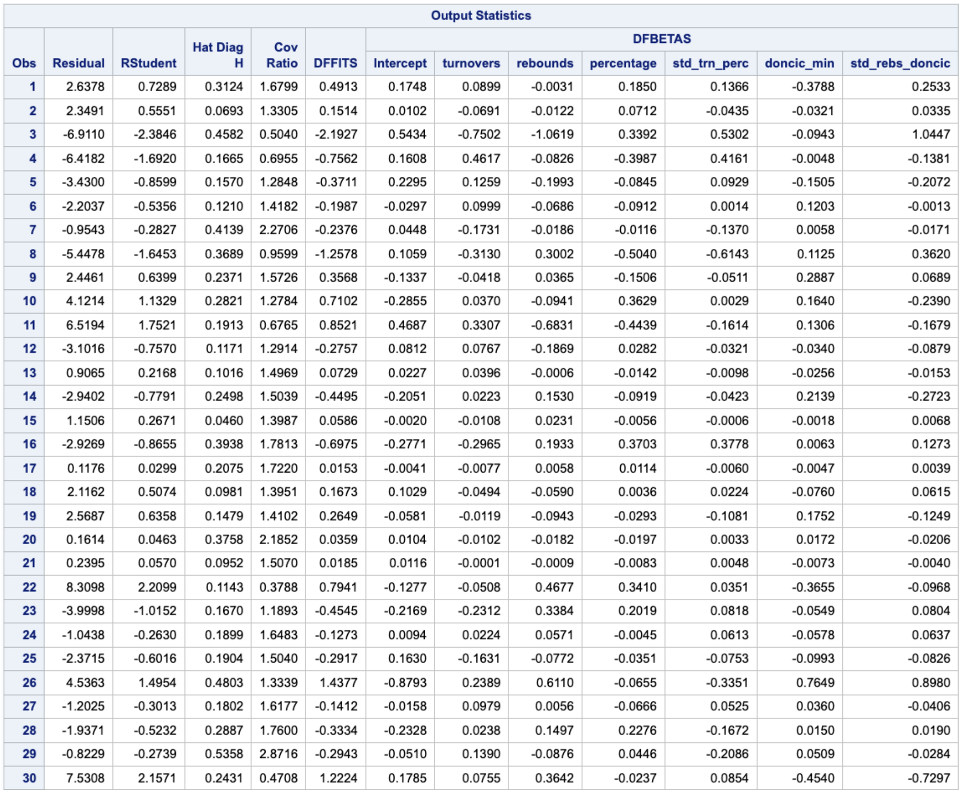


Figure 41

We must check that the multiple linear regression model form is not violated. Using the plots in figure 41, if there is curvature between any of the predictor variables and the residuals of model 6a, this would be a violation of this assumption.

The only concern would be the residuals versus Doncic minutes plot, there is possible curvature with a U-shape around 40 minutes. This curvature is nearly identical to the curvature found in the plot of residuals versus Doncic minutes for the preliminary model in figure 6 and model 5a. We can consider this dip or U-shape this to be an odd artifact as we did for the preliminary model and model 5a since transformations did not improve the U-shape. There does not appear to be curvature in any of the other plots in figure 36, thus we can reasonably assume that the multiple linear regression model form is not violated.



Output Statistics

						DFBETAS						
Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	Intercept	turnovers	rebounds	percentage	std_trn_perc	doncic_min	std_rebs_doncic
1	2.6378	0.7289	0.3124	1.6799	0.4913	0.1748	0.0899	-0.0031	0.1850	0.1366	-0.3788	0.2533
2	2.3491	0.5551	0.0693	1.3305	0.1514	0.0102	-0.0691	-0.0122	0.0712	-0.0435	-0.0321	0.0335
3	-6.9110	-2.3846	0.4582	0.5040	-2.1927	0.5434	-0.7502	-1.0619	0.3392	0.5302	-0.0943	1.0447
4	-6.4182	-1.6920	0.1665	0.6955	-0.7562	0.1608	0.4617	-0.0826	-0.3987	0.4161	-0.0048	-0.1381
5	-3.4300	-0.8599	0.1570	1.2848	-0.3711	0.2295	0.1259	-0.1993	-0.0845	0.0929	-0.1505	-0.2072
6	-2.2037	-0.5356	0.1210	1.4182	-0.1987	-0.0297	0.0999	-0.0686	-0.0912	0.0014	0.1203	-0.0013
7	-0.9543	-0.2827	0.4139	2.2706	-0.2376	0.0448	-0.1731	-0.0186	-0.0116	-0.1370	0.0058	-0.0171
8	-5.4478	-1.6453	0.3689	0.9599	-1.2578	0.1059	-0.3130	0.3002	-0.5040	-0.6143	0.1125	0.3620
9	2.4461	0.6399	0.2371	1.5726	0.3568	-0.1337	-0.0418	0.0365	-0.1506	-0.0511	0.2887	0.0689
10	4.1214	1.1329	0.2821	1.2784	0.7102	-0.2855	0.0370	-0.0941	0.3629	0.0029	0.1640	-0.2390
11	6.5194	1.7521	0.1913	0.6765	0.8521	0.4687	0.3307	-0.6831	-0.4439	-0.1614	0.1306	-0.1679
12	-3.1016	-0.7570	0.1171	1.2914	-0.2757	0.0812	0.0767	-0.1869	0.0282	-0.0321	-0.0340	-0.0879
13	0.9065	0.2168	0.1016	1.4969	0.0729	0.0227	0.0396	-0.0006	-0.0142	-0.0098	-0.0256	-0.0153
14	-2.9402	-0.7791	0.2498	1.5039	-0.4495	-0.2051	0.0223	0.1530	-0.0919	-0.0423	0.2139	-0.2723
15	1.1506	0.2671	0.0460	1.3987	0.0586	-0.0020	-0.0108	0.0231	-0.0056	-0.0006	-0.0018	0.0068
16	-2.9269	-0.8655	0.3938	1.7813	-0.6975	-0.2771	-0.2965	0.1933	0.3703	0.3778	0.0063	0.1273
17	0.1176	0.0299	0.2075	1.7220	0.0153	-0.0041	-0.0077	0.0058	0.0114	-0.0060	-0.0047	0.0039
18	2.1162	0.5074	0.0981	1.3951	0.1673	0.1029	-0.0494	-0.0590	0.0036	0.0224	-0.0760	0.0615
19	2.5687	0.6358	0.1479	1.4102	0.2649	-0.0581	-0.0119	-0.0943	-0.0293	-0.1081	0.1752	-0.1249
20	0.1614	0.0463	0.3758	2.1852	0.0359	0.0104	-0.0102	-0.0182	-0.0197	0.0033	0.0172	-0.0206
21	0.2395	0.0570	0.0952	1.5070	0.0185	0.0116	-0.0001	-0.0009	-0.0083	0.0048	-0.0073	-0.0040
22	8.3098	2.2099	0.1143	0.3788	0.7941	-0.1277	-0.0508	0.4677	0.3410	0.0351	-0.3655	-0.0968
23	-3.9998	-1.0152	0.1670	1.1893	-0.4545	-0.2169	-0.2312	0.3384	0.2019	0.0818	-0.0549	0.0804
24	-1.0438	-0.2630	0.1899	1.6483	-0.1273	0.0094	0.0224	0.0571	-0.0045	0.0613	-0.0578	0.0637
25	-2.3715	-0.6016	0.1904	1.5040	-0.2917	0.1630	-0.1631	-0.0772	-0.0351	-0.0753	-0.0993	-0.0826
26	4.5363	1.4954	0.4803	1.3339	1.4377	-0.8793	0.2389	0.6110	-0.0655	-0.3351	0.7649	0.8980
27	-1.2025	-0.3013	0.1802	1.6177	-0.1412	-0.0158	0.0979	0.0056	-0.0666	0.0525	0.0360	-0.0406
28	-1.9371	-0.5232	0.2887	1.7600	-0.3334	-0.2328	0.0238	0.1497	0.2276	-0.1672	0.0150	0.0190
29	-0.8229	-0.2739	0.5358	2.8716	-0.2943	-0.0510	0.1390	-0.0876	0.0446	-0.2086	0.0509	-0.0284
30	7.5308	2.1571	0.2431	0.4708	1.2224	0.1785	0.0755	0.3642	-0.0237	0.0854	-0.4540	-0.7297

Figure 42

We will repeat the same procedure for finding outliers as we did for the previous model, 5a. Starting with x-outliers, our cutoff for ‘H Diag H’ is $\frac{2p}{n} = \frac{2*7}{30} = 0.46667$. Observation 26 has a leverage score of 0.4803 and observation 29 has a leverage scored of 0.5358, these both exceed our cutoff of 0.46667, now we need to evaluate their influence.

If the absolute value of DIFFITS for observation 26 or observation 29 exceed $2\sqrt{\frac{p}{n}} = 2\sqrt{\frac{7}{30}} = 0.966$ then we will consider it having high influence on the fitted values of total points scored. Observation 26 has a DIFFITS value of 1.4377 which exceeds 0.966, thus we say this observation has high influence on the fitted values but observation 29 does not.

If the absolute value of DFBETAs for observation 26 or observation 29 exceed $\frac{2}{\sqrt{n}} = \frac{2}{\sqrt{30}} = 0.365$ then it has a high influence on the corresponding least squares estimate for a particular parameter. Observation 26's DFBETAS exceeds 0.365 for both rebounds and std_rebs_doncic. Although observation 29 is not influential on the least square estimates for any of the parameters.

Using the Bonferroni outlier test if the absolute value of the studentized deleted residuals in the 'RStudent' column exceed $t\left(1 - \frac{\alpha}{2n}; n - p - 1\right) = t\left(1 - \frac{0.1}{2*30}; 30 - 7 - 1\right) = t(0.99833; 22) = 3.291$ it will be considered a y-outlier. Since none of the absolute value of the studentized deleted residuals exceed 3.291, we did not detect any y-outliers.

C. **Choose Final Model**

After assessing the model assumptions for both model 5a and 6a, we must choose the model that performs the 'best'. There is no clear 'best' model so we must compare both using a variety of factors.

Model 6a has a better adjusted R^2, Mallow's C(p), AIC, SBC than model 5a. Also, model 6a satisfies normality in the residuals, and model 5a doesn't. Although model 6a is more complex, it does appear to outperform model 5a.

In the comparative analysis between Model 5a and Model 6a for predicting the Dallas Mavericks' scoring outcomes, several statistical metrics were considered to ascertain which model performs optimally. Model 6a exhibits a higher adjusted R-squared value, indicating superior efficiency in explaining the variability of the scores relative to the number of predictors used. Additionally, it presents a lower Mallow's C(p) value, suggesting a better balance between model complexity and accuracy, as this criterion assesses the trade-off between the model's complexity and its goodness of fit. The Akaike Information Criterion (AIC) and Schwarz Bayesian Criterion (SBC) also favor Model 6a, with both criteria returning lower values compared to Model 5a, signifying a more desirable model by penalizing excessive complexity while rewarding goodness of fit.

Furthermore, the assumption of normality in the residuals—which is critical for validating the inferences made from the model—shows that Model 6a conforms more closely to this statistical prerequisite. The residuals of Model 6a align more accurately with a normal distribution, enhancing the reliability of the regression analysis and the statistical tests applied, such as the F-test. In contrast, Model 5a struggles with this assumption, as evidenced by its residuals which deviate from normality. Although Model 6a is inherently more complex due to the inclusion of additional predictors and interaction terms, its and adherence to critical assumptions underline its superiority over Model 5a, making it the preferable choice for analytical and predictive applications concerning the Mavericks' performance.

VI. Final MLR Model

Our final model to predict total points scored by the Mavs in a single game is the following:

$$\begin{aligned}\widehat{points} = -21.49 - 0.55(turnovers) + 0.72(rebounds) + 184.1(percentage) \\ - 2.03(std\_trn\_perc) + 0.62(Doncic\_min) - 1.87(std\_rebs\_Doncic)\end{aligned}$$

For each turnover, the predicted total points decreases by 0.55 points, assuming all other variables are held constant. For each rebound, the predicted total points increases by 0.72 points, assuming all other variables are held constant. For a 1% increase in shot percentage, the predicted total points decreases by 184.1 points, assuming all other variables are held constant. Each additional minute Doncic plays the predicted total points increases by 0.62 points, assuming all other variables are held constant. The effect of turnovers on points decreases by -2.03 points for every 1% increase in shot percentage. The effect of rebounds on points decreases by -1.87 points for every minute Doncic plays.

The final model explains 90.67% of the variability in points using turnovers, rebounds, percentage, std_trn_perc, Doncic minutes, and std_rebs_doncic as predictors. The F-test in figure 31 confirms that the model is statistically significant in predicting points so we can conclude that at least one of the coefficients is not equal to zero. Additionally using the t-tests in figure 31, we confirmed that each of the 6 predictors are statistically significant at the 0.10 level, so we concluded that their coefficients are not equal to 0. Having a Mallow's C(p) of 8.3 which is very close to 7 (the number of parameters) suggests that there are no important predictors of points missing. Considering all these factors and after verifying it meets the linear regression model assumptions, we felt it was the strongest candidate model in predicting total points scored.

A. Parameter Inferences

We will now conduct the Bonferroni joint confidence intervals for each of the coefficients of the g = 6 predictors. Consider the following confidence interval equation: $b_k \pm B \cdot s\{b_k\}$ such that k = 1, 2, 3, 4, 5, 6 for each predictor and $B = t\left(1 - \frac{\alpha}{2*g}; n - p\right) = t\left(1 - \frac{0.1}{2*6}; 30 - 7\right) = t(0.991667; 23) = 2.582$.

For turnovers, $b_1 \pm 2.582 \cdot s\{b_1\} = -0.55 \pm 2.582(0.23) = (-1.14\,, 0.04)$. We are $100\left(1 - \frac{\alpha}{g}\right)\% = 98.33\%$ confident that β_1 is in $(-1.14\,, 0.04)$.

For rebounds, $b_2 \pm 2.582 \cdot s\{b_2\} = 0.72 \pm 2.582(0.16) = (0.31\,, 1.13)$. We are 98.33% confident that β_2 is in $(0.31\,, 1.13)$.

For percentage, $b_3 \pm 2.582 \cdot s\{b_3\} = 184.1 \pm 2.582(15.87) = (143.12\,, 225.08)$. We are 98.33% confident that β_3 is in $(143.12\,, 225.08)$.

For std_trn_perc, $b_4 \pm 2.582 \cdot s\{b_4\} = -2.03 \pm 2.582(0.86) = (-4.25\,, 0.19)$. We are 98.33% confident that β_4 is in $(-4.25\,, 0.19)$.

For Doncic minutes, $b_5 \pm 2.582 \cdot s\{b_5\} = 0.62 \pm 2.582(0.25) = (-0.03\,, 1.27)$. We are 98.33% confident that β_5 is in $(-0.03\,, 1.27)$.

For std_rebs_doncic, $b_6 \pm 2.582 \cdot s\{b_6\} = -1.87 \pm 2.582(1.03) = (-4.53\,, 0.79)$. We are 98.33% confident that β_6 is in $(-4.53\,, 0.79)$.

We are 90% confident that β_1 is in $(-1.14\,, 0.04)$, β_2 is in $(0.31\,, 1.13)$, β_3 is in $(143.12\,, 225.08)$, β_4 is in $(-4.25\,, 0.19)$, β_5 is in $(-0.03\,, 1.27)$, β_6 is in $(-4.53\,, 0.79)$ simultaneously.

B. <u>Mean Response and Prediction Inferences</u>

Earlier this season, the Mavs beat Orlando Magic with a score of 131 to 129, this is a high scoring game with a close score. We felt that it would be interesting to use similar stats from this game and see what our model predicts the total points scored for the Mavs would be. In this game, the Mavs had 15 turnovers, 35 rebounds, a 63% shot percentage and Doncic played for 43 minutes in total, making std_trn_perc=2.1 and std_rebs_doncic=-1.09.

We will find the 90% confidence interval for the mean response of a game with the stats described above (x_h) using the following: $\widehat{points}_h \pm t\left(1-\frac{0.10}{2}; n-p\right) s\{\widehat{points}_h\}$. The point estimator for x_h is $-21.49 - 0.55(15) + 0.72(35) + 184.1(0.63) - 2.03(2.1) + 0.62(43) - 1.87(-1.09) = 135.88$ points. The standard error of $\widehat{points}_h$is $\sqrt{\boldsymbol{x}_h{}^T \cdot s^2\{\boldsymbol{b}\} \cdot \boldsymbol{x}_h} =$

$$\sqrt{\begin{bmatrix} 1 & 15 & 35 & 0.63 & 2.1 & 43 & -1.09 \end{bmatrix} \cdot s^2\{\boldsymbol{b}\} \cdot \begin{bmatrix} 1 \\ 15 \\ 35 \\ 0.63 \\ 2.1 \\ 43 \\ -1.09 \end{bmatrix}} = \sqrt{6.78} = 2.6$$

and $t\left(1-\frac{0.10}{2}; n-p\right) = t(0.95; 30-7) = 1.714$. We are 90% confident that the expected total points scored for game with the same metrics as the Mavs vs Magic game will be between $135.88 \pm 1.714 * 2.6 = (131.42\,, 140.34)$.

To find the 90% Confidence region interval for the mean response of the Mavericks versus Magic game stated above, we will use the equation: $\widehat{points}_h \pm \sqrt{w^2} * s\{\widehat{points}_h\}$ where $w^2 = p \cdot F(1-\alpha; p, n-p) = 7 \cdot F(1-0.1; 7,30-7) = 7 \cdot F(0.9; 7,33) = 7 \cdot 1.907 = 13.349$. We are 90% confident that the entire regression surface lies within $135.88 \pm \sqrt{13.349} \cdot 2.6 = (126.38, 145.38)$ points for a game with 15 turnovers, 35 rebounds, a 63% shot percentage, 43 Doncic Play minutes, std_trn_perc=2.1 and std_rebs_doncic=-1.09.

To find the 90% Prediction Interval for the mean response of the Mavericks versus Magic game, we will first solve for the prediction error. The equation for the prediction error is $s\{pred\} = \sqrt{(s\{\widehat{points}_h\})^2 + \boldsymbol{MSE}} = \sqrt{(2.6)^2 + 18.66061} = 5.04$. Once we calculate the prediction error, we can solve for the prediction interval using this equation: $\widehat{points}_h \pm t(0.95; 30 - 7)s\{pred\} = 135.88 \pm 1.714 \cdot 5.04 = (127.24, 144.52)$. We are 90% confident that the points scored by the Mavericks in the next game 15 turnovers, 35 rebounds, a 63% shot percentage, 43 Doncic Play minutes, std_trn_perc=2.1 and std_rebs_doncic=-1.09 falls between 127.24 and 144.52 points.

VII. Final Discussion

Model 6a is selected over Model 5a primarily due to its superior performance in explaining the variability in the total points scored by the Dallas Mavericks, as reflected in its higher adjusted R-squared value. This indicates that Model 6a, despite its complexity, captures a larger proportion of variance when adjusted for the number of predictors used, demonstrating effective utilization of its additional predictors, including significant interaction terms. These terms interactions between turnovers and shot percentage and between rebounds and player minutes provide critical insights into how combinations of game dynamics affect game scores. This level of detail is not captured by Model 5a, which lacks these interaction terms.

Furthermore, all predictors in Model 6a, including the interaction terms, were statistically significant with p-values below the significance level, indicating a robust and nuanced understanding of factors influencing game scores. This model also adheres more closely to the assumptions of linear regression. Both models had similar diagnostics in reference to outliers. The residuals from Model 6a exhibit consistent behavior regarding normality and homoscedasticity, additionally we believe that follows the multiple linear regression model form. Model 5a clearly violates the normality in the residuals assumption so model 6a is stronger in this regard. This ensures the reliability of regression analysis and the validity of predictions and inferences drawn from the model.

Although Model 6a is more complex due to additional terms, this complexity is justified by a significant increase in explanatory power and the ability to capture additional nuances in the relationship between predictors and the response variable. The complexity does not lead to overfitting but rather enhances the model's capability to depict a detailed and accurate picture of game dynamics. Thus, model 6a is not only a statistically sound choice but also a more informative tool for strategic decision-making regarding game outcomes.