Linear Regression Analysis on Basketball Data

Cameron Scolari

Loyola Marymount University

April 2024

Abstract

For this particular study, I chose to use data pertaining to the NBA player, Nikola Jokić, the reigning 2023 Finals MVP. I chose Jokić because he is the best player on a championship roster. I hope to see if linear regression analysis is viable in understanding a player's impact on their team's overall performance. To do so, I collected data from the 2023-2024 season and constructed different models using different regressor variables in hopes of finding the best one. The models taught me about how a player's performance in specific statistical categories affects their team's overall performance. This study informed me what Jokić can do and prevent from doing to help his team succeed on any given night. This idea can be used to help better understand a player's importance to their team's success and help coaches maximize a player's effectiveness by putting emphasis on the statistical categories that help the team and highlighting the categories a player should avoid.

1 Introduction

The simple linear regression model is

$$y = \beta_0 + \beta_1 x + \varepsilon,\ \varepsilon \overset{iid}{\sim} N(0, \sigma^2), \tag{1}$$

where the intercept β_0 and the slope β_1, also known as regression coefficients (Definition 1), are unknown constants and ε is a random error component. The errors are assumed to have an expected value (Definition 2) of zero and unknown variance σ^2 (Definition 3). The errors are also assumed to be independent.[5]

The regressor (Definiton 4) x is a controlled variable set by the researcher while the response (Definiton 5), y, is a random variable that has a normal probability distribution (Definiton 6). The expected value of equation 1 is $\beta_0 + \beta_1 x$ and the variance is σ^2.

Regression analysis (Definiton 7) utilizes least-squares estimation that relies on techniques from calculus and algebra to minimize an objective function written in terms of the parameters, β_0 and β_1, to create estimates of these parameters.

To explain the least-squares estimation of these parameters, I will use the following example dataset[5] that measures the purity of oxygen produced by a fractional distillation process compared to the percentage of hydrocarbons in the main condensor of the processing unit:

Table 1: Sample Data

	purity	hydrocarbon
1	86.91	1.02
2	89.85	1.11
3	90.28	1.43
4	86.34	1.11
5	92.58	1.01
6	87.33	0.95
7	86.29	1.11
8	91.86	0.87
9	95.61	1.43
10	89.86	1.02
11	96.73	1.46
12	99.42	1.55
13	98.66	1.55
14	96.07	1.55
15	93.65	1.40
16	87.31	1.15
17	95.00	1.01
18	96.85	0.99
19	85.20	0.95
20	90.56	0.98

Least-squares estimation estimates β_0 and β_1 so that the sum of the squares of the differences between y_i and the straight line is a minimum. So, with the equation,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, ..., n, \ \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), \tag{2}$$

we can write an equation in terms of β_0 and β_1 as

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{3}$$

This equation is just the sum of the squares of the differences, which in this case is just ε_i. Now we must take partial derivatives of β_0 and β_1 respectively to find equations that produce the minimum values. After that, we can plug in the example data to find the estimated values for β_0 and β_1. First:

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{4}$$

and

$$\left.\frac{\partial S}{\partial \beta_1}\right|_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0 \tag{5}$$

With a bit of simplifying and algebraic manipulation, you get these two equations for $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{6}$$

and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{7}$$

In many textbooks, you may see equation 7 written differently, but all forms should be equivalent and produce the same result. Now, I can use the sample data from Table 1 to get estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$. First, I must calculate $\overline{x}$ and $\overline{y}$, the sample means of the two variables. Since the purity is what is being predicted, that is the response variable (y variable). To find $\overline{x}$, I add up all the data in the hydrocarbon column and divide by the sample size, which is 20. I do the same for the purity column to find $\overline{y}$. For this example,

$$\overline{x} \approx 1.1825$$

and

$$\overline{y} \approx 91.818$$

With these numbers, we can use them along with each individual data point to find the estimated values for β_0 and β_1:

$$\hat{\beta}_0 \approx 77.867$$

and

$$\hat{\beta}_1 \approx 11.798$$

Then, plugging those estimates in, we find the least squares line,

$$\hat{y} = 77.867 + 11.798x \tag{8}$$

We could expect that given a percentage of 0 hydrocarbons in the main condensor of the processing unit, the estimated mean purity of oxygen produced is 77.867. Additionally, for every 1 percentage of hydrocarbon increase, there is on average a 11.798 increase in the purity of oxygen. This is how we can construct a simple linear regression model using a dataset. For further analysis, we could assess the quality of the model by checking the coefficient of determination (R^2) value and residual plots. I will omit this analysis and save it for the basketball data I actually collected.

2 Data Collection and Purification

All of the data collected came from basketball-reference.com. For this particular study, I chose to use data pertaining to the NBA player, Nikola Jokić. To access this data, I opened his 2023-2024 gamelog. This log was downloadable as a Comma-Separated Values (CSV) file and it contained every statistic about each individual game he played during the given season. I cleaned the data using pandas, a Python library. This involved me filling in any blank data cells as well as removing columns of statistics that were not useful and removing rows containing no data. Once the data was in the proper format, I could save the data as a CSV file and open it in R.

3 Data Analysis

Table 2 shows a list of the 20 regressor variables and the response variable that were used in this study. These are statistics that an individual player has during each game. The mean and standard deviation (Definiton 8) were calculated using data from the first 60 games of the season.

Table 2: Available Variables

variable	description	mean	standard deviation
MP	Minutes played	33.72	5.21
FG	Number of shots made that are not free throws	10.18	3.41
FGA	Number of shot attempts that are not free throws	17.58	6.26
FG%	FG/FGA	0.60	0.17
3P	Number of three point shots a player makes	1.067	0.97
3PA	Number of three point shots a player attempts	3.05	2.39
3P%	3P/3PA	0.35	0.34
FT	Number of free throws a player makes	4.50	3.49
FTA	Number of free throws a player attempts	5.50	3.96
FT%	FT/FTA	0.77	0.28
ORB	Number of rebounds on opponent's side of court	2.87	2.13
DRB	Number of rebounds on own side of court	9.43	3.17
TRB	ORB+TRB	12.30	3.90
AST	Number of passes that led to a player scoring	9.15	3.47
STL	Number of times player steals ball from opponent	1.20	1.13
BLK	Number of times player blocks opponent's shot	0.92	0.96
TOV	Number of times player gets ball stolen by opponent	2.98	2.00
PF	Number of times player commits a foul	2.57	1.18
PTS	Number of points a player scores	25.93	8.07
+/-	Team's point differential while on floor	7.75	14.07
RESULT	The player's team's final score	114.70	12.23

Originally, I did not have all these regressors and was only analyzing a few. The original model I constructed compiled a total of 60 games from 10/24/23 to 3/5/24. The regressors that I found worked best were 'AST' and 'TOV'. Table 3 displays the estimated coefficients along with other information:

Table 3: Model 1 Summary Statistics

Variable	Estimate	Std. Error	t value	p
(Intercept)	106.84	5.19	20.57	$2*10^{-16}$
AST	1.30	0.43	3.06	0.00337
TOV	-1.37	0.74	-1.86	0.06780

The R^2 for this model was 0.2355 which indicates that only about 24% of the variability in y could be explained by the x variables which is low. Because of this low value, this model is probably missing important regressor variables. The adjusted R^2 for this model was 0.2086. This will be important for comparison with different models. Despite the low R^2, I still tested this model using Jokić's data from the next 10 games to see how well the prediction worked. Table 4 shows the results of the predicted and actual scores of the 10 games.

Table 4: Predicted and actual scores of the 10 games

predicted	actual	|difference|	predicted lower bound	predicted upper bound
120.61	115	5.61	99.05	142.18
119.65	142	22.35	97.78	141.52
116.18	125	8.82	93.80	138.56
115.90	100	15.90	94.41	137.39
114.50	117	2.50	94.45	134.55
115.38	105	10.38	94.83	135.93
106.70	115	8.30	84.38	129.03
129.21	113	16.21	107.60	150.82
119.11	128	8.89	98.83	139.39
109.03	97	12.03	88.54	129.53

The average of the absolute difference between the predicted score and the actual score was 11.099. Across these ten games, the actual game score fell within the 95% prediction interval 9/10 times. Since the R^2 was small, I skipped any in-depth analysis and found more potential regressor variables because important predictors are missing in this initial model. I constructed a model using the same 60 games from 10/24/23 to 3/5/24, but this time with different regressors. For this model, the regressors that produced the largest adjusted R^2 were '+/-', '3PA', 'AST',

'3P%', 'DRB', and 'PF'(Table 1). Table 5 displays the estimated coefficients along with other information.

Table 5: Model 2 Summary Statistics

Variable	Estimate	Std. Error	t value	p
(Intercept)	113.19	4.86	23.31	$2*10^{-16}$
+/-	0.52	0.09	5.60	$7.75*10^{-7}$
3PA	-1.34	0.51	-2.62	0.01137
AST	0.93	0.34	2.69	0.00943
3P%	4.90	3.08	1.59	0.11791
DRB	-0.57	0.38	-1.52	0.13406
PF	-1.26	0.95	-1.33	0.18995

Table 6 shows the predictions of the same 10 games as before using Model 2.

Table 6: Predicted and actual scores of the 10 games

predicted	actual	|difference|	predicted lower bound	predicted upper bound
113.85	115	1.15	94.36	133.34
114.31	142	27.69	94.88	133.74
116.17	125	8.83	96.67	135.67
114.80	100	14.80	95.34	134.25
115.62	117	1.38	96.13	135.11
114.43	105	9.43	94.94	133.92
113.64	115	1.36	94.18	133.09
125.67	113	12.67	105.79	145.54
119.10	128	8.90	99.59	138.62
108.69	97	11.69	88.99	128.39

The average of the absolute difference between the predicted score and the actual score was 9.79. Across these ten games, the actual game score fell within the 95% prediction interval 9/10 times.

Interestingly, the same game for both models did not lie within the 95% prediction interval. Although this game is likely an outlier because a score of 142 is very high, 9/10 of the scores lying within the 95% prediction intervals (Definition 9) is not too unusual because in the long run, 95% of the intervals will contain the true value.

The average absolute difference was more than a point lower in the second model than the first model. Additionally, the second model had a larger R^2 value of

0.6179. Remember, the initial model had an adjusted R^2 of 0.2086. So, the second model has regressors that account for about 41% more of the variability in the response variable than the regressors in the original model do. We can further prove model 2's superiority by comparing the adjusted R^2 of each model. The adjusted R^2 value for model 2 was 0.5747. Recall, the adjusted R^2 of the first model was only 0.2086. Since the adjusted R^2 value of the second model was greater, the second model is better suited for the data. Another way of comparing regression models is using the AIC value (Definiton 10). The AIC for the second model was 255.73 while the AIC for the first model was 289.3515. The second model having a lower AIC than that of the first model is an indication that the second model is better than the original model for making predictions on the given data. With all of this in mind, it is safe to say that the second model is better.

Here is the equation for the regression line of the second model:

$$\hat{y} = 113.19 + 0.52x_0 - 1.34x_1 + 0.93x_2 + 4.9x_3 - 0.57x_4 - 1.26x_5 \tag{9}$$

where x_0 is the player's +/-, x_1 is the player's three-point shot attempts, x_2 is the number of assists a player has during the game, x_3 is the player's three-point shooting percentage, x_4 is the number of defensive rebounds a player has during the game, and x_5 is the number of fouls a player has during the game.

Based on equation 9, Jokić's team would score around 113 points if all of the x-values were 0. This means Jokić would have to play an entire game without contributing in any of these categories. Now, Jokić playing in a game and not contributing in these categories would be very rare so it is safe to assume that 113 is the expected number of points Jokić's team would score if he sat out of the game. To explain more about the regression line, I provided some personal analysis for each coefficient associated with each variable.

If you recall from Table 5, '+/-' was the most significant variable. This makes sense because it measures how much a player impacts their team[10] by calculating the difference of their team's and the opposing team's points while they are on the floor. I have seen the '+/-' of players reach as high as +30 and as low as -30, so it makes sense that the coefficient is a decimal. If the coefficient were anything greater, this variable would have too much weight.

'3PA' having a negative coefficient is interesting because it suggests that the more three pointers Jokić attempts, the less his team scores. In Jokić's career, he has a 35% three-point percentage. So, it makes sense why the coefficient is negative because Jokić is more likely to miss the three-point shot than make it when he shoots it. Therefore, whenever he shoots a three-point shot, it is most likely a wasted possession for his team. This is because Jokić is one of the taller players on

his team, and when he misses a three-point shot from around 23 feet away from the basket, he is not there to get the rebound for his team.

It makes sense for 'AST' to have a positive coefficient, but I would have expected it to be closer to 2 than 1. An assist is rewarded when a player gives a pass to another player that leads to them scoring. So, whenever Jokić gets an assist, his team is guaranteed to score at least 2 points if not 3.

'3P%' has the largest coefficient by far. However, this makes sense because it is the only variable that is a percentage. It also makes sense for it to be positive because if there were a three-pointer to be made, the percentage would have a value other than 0 and points would be accounted for in the predictive model. If Jokić fails to make any three-pointers, the percentage would be 0 and no points would be deducted.

It is interesting that the coefficient for 'DRB' is negative. I would expect this to be a positive number because when Jokić gains a defensive rebound, his team gains possession of the ball which means they have an opportunity to score.

I would not expect 'PF' to have much of an impact in the linear regression analysis, but it is the 6th best regressor variable. I think 'PF' has a negative coefficient because it is likely the case that when Jokić fouls someone, he gets substituted out of the game. This is because players have a maximum of 6 fouls to commit in a game of basketball and coaches do not want to keep players on the floor when they are fouling. When Jokić is off the floor, his team is likely not scoring as much.

The final step of analysis is to see if our residuals (Definiton 11) follow a normal distribution. To do this, we can look at some graphs that R created for us.

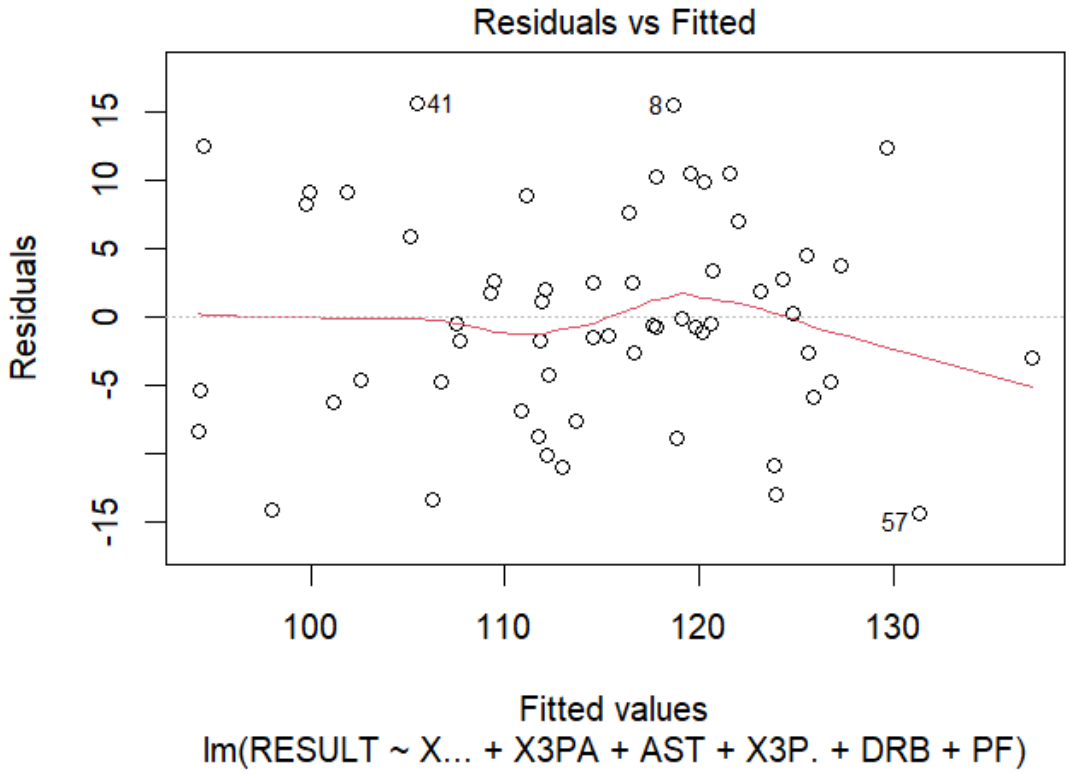

Figure 1: Plot of residuals vs. fitted values from the model.

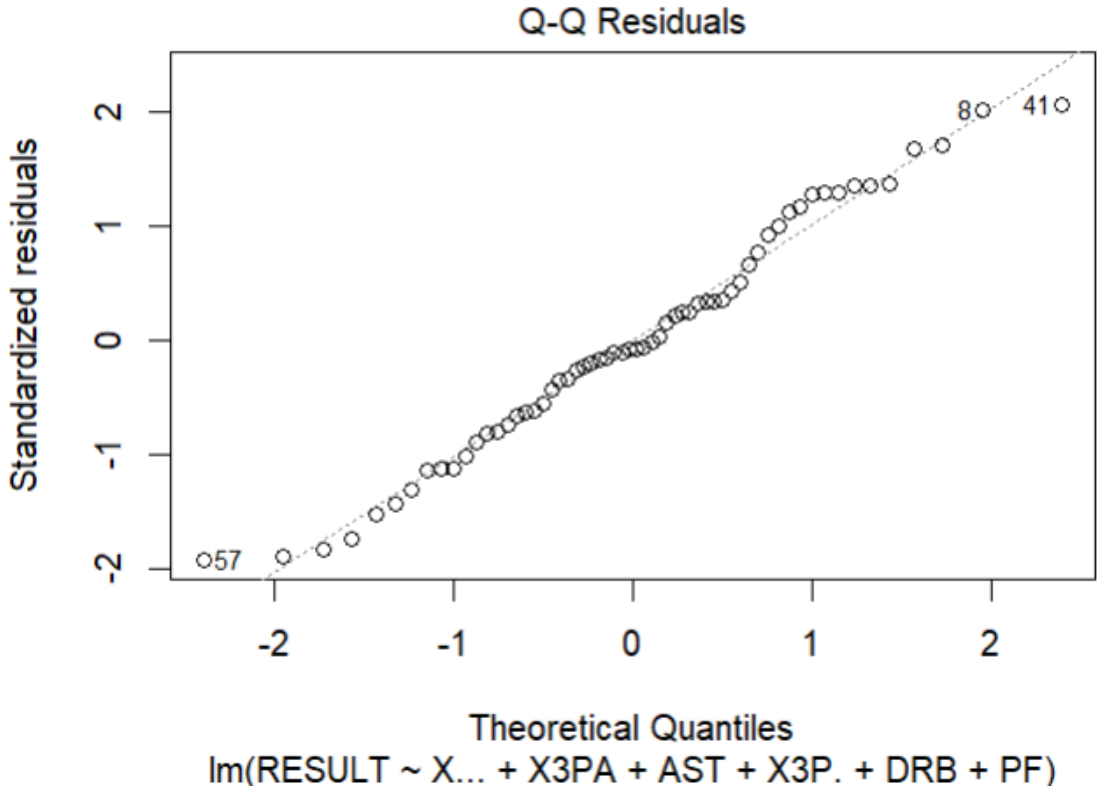

Figure 2: Q-Q plot of standardized residuals.

If our model was suitable for the data, we would expect Figure 1 to have data points that look randomly distributed in a horizontal band around 0 with no clear pattern. This is what Figure 1 looks like, so we have strong evidence that our model is adequate. Figure 2 is a plot of standardized residuals. Each standardized residual is calculated by taking the raw residual and dividing by the estimated standard deviation of the residuals.[8] A Q-Q plot is used to assess if our residuals are normally distributed. If the data seems to follow the line, then the data is normally distributed. Figure 2 shows data that strongly follows the line indicating normally distributed residuals. Furthermore, Figure 2 also highlights the lack of outliers in the data because the absolute values of all standardized residuals are less than 3.[3]

4 Conclusion

This study was able to highlight some of the strengths of Jokić's game and some of his weaknesses. It would be interesting to see how factoring in more data from previous seasons changes this model. Next season, there may be a better model, with different regressors, that has more accuracy. This model should only be valid for this season. Additionally, this model cannot be used on other players. It would be interesting to see how models vary between different players though. It would also be informative to see what a model would look like if statistics from each player on the team were taken into account. I predict that that model would give a better indication of what role each player plays for their team. This study, however, did highlight how important Jokić is for his team. The R^2 value of 0.6179 suggests that roughly 62% of the variance in Jokić's team's score is explained by the statistics that make up the regressor variables in this model.

5 Appendix

Definition 1. Regression Coefficients - β_0 and β_1. If the range of the data on x includes x=0, then β_0 is the mean of the distribution of y when x=0. There is no practical interpretation for it if the range does not include x=0. β_1 is the slope that can be interpreted as the change in the mean of y for a unit change in x.

Definition 2. Expected Value - An expected value is calculated by multiplying the value of each event by each event's particular probability and summing those numbers. The expected value of a normal distribution is μ. It is also known as the mean.[6]

Definition 3. Variance - The variance provides a quantitative measure of how closely the data set is spread around its center. A variance of smaller magnitude (closer to zero) implies that the set of numbers is quite tightly clustered around the center.

Definition 4. Regressor Variable - Typically referred to as x or the predictor variable.

Definition 5. Response Variable - Typically referred to as y. Responds to the regressor variables and the error component.

Definition 6. Normal Probability Distribution - A distribution is considered "normal" if the graph follows a bell-shaped curve with a population mean μ and population variance σ^2.[4]

Definition 7. Regression Analysis - Statistical technique for investigating and modeling the relationship between variables.

Definition 8. Standard Deviation - Measures the extent of scattering in a set of values, typically compared to the mean value of the set. The standard deviation is the square root of the variation.[9]

Definition 9. Confidence Interval - Confidence Intervals indicate the variability of the estimates. The confidence interval uses the sample to estimate the interval of probable values of the population. The confidence interval shows the range of values you expect the true parameter to fall between if you redo the study many times.[1]

Definition 10. AIC - An AIC score is a number used to determine which machine learning model is best for a given data set in situations where one can't easily test a data set. The lower the AIC score the better.[2]

Definition 11. Residual - Difference between the observed value and the expected value predicted by the model.[7]

6 Acknowledgements

Thank you to Dr. Wang for assisting me in this research and teaching me all I need to know about linear regresssion analysis.

Thank you Dr. Eubanks-Turner for teaching me techniques for writing this paper and helping me understand LaTex better.

References

[1] A. Hazra, *Using the confidence interval confidently*, J Thorac Dis., **9** , 4125-4130, 2017

[2] A. Zajic, *What Is Akaike Information Criterion (AIC)?*, Builtin, 2022.

[3] C. Borse, *Statistical Methods for Identifying Outliers(Regression Analysis Approach)(Parti II)*, Medium, 2020.

[4] C. Tsokos & R. Wooten, *Normal Probability Distribution*, The Joy of Finite Mathematics, Academic Press, 231-263, 2016.

[5] D. Montgomery, E. Peck, & G. Vining, *Introduction to Linear Regression Analysis*, John Wiley & Sons, Inc., **5**, 12-100, 2012 Introducion to Linear Algebra (5th edition) by Montgomery, Peck, Vining

[6] D. Palmer, *Expected Utility*, Encyclopedia Britannica, 2023.

[7] J. Frost, *Residuals*, StatisticsByJim.

[8] K. Feldman, *Standardized Residuals: Insights into Calculations, Interpretations, and Applications*, ISixSigma, 2023.

[9] S. El Omda & S. Sergent, *Standard Deviation*, StatPearls Publishing, 2023.

[10] S. Girotra, *Advanced Stats for Basketball*, Gamechanger, 2021.