Multiple Linear Regression

# Overview

## Project context

* This report summarises the investigation carried out on the factors that affect the turnovers committed (i.e., number of dispossessions) committed by an NBA player.
* The NBA (National Basketball Association) is the top professional basketball league in North America. It was founded in 1947 and is currently home to the best basketball players in the world.
* The data used for this project is the per-game statistics for players in the NBA spanning the last 31 seasons (1990-2021). It was scraped from the basketballreference.com (see project: [Data-Pipeline](https://github.com/favourumeh/DATA-PIPELINE) project for more details ETL process used)

## Python Modules Used

* statsmodels, sklearn
* seaborn, matplotlib.pyplot
* pandas
* numpy

## Skills showcased

* Utilised Python statistical packages (namely sklearn and statsmodels) to conduct predictive modelling (namely linear regression)
* Validate linear regression model against regression assumptions:

1. Correct functional form: (by observing the scatter plots )
2. Constant variance/no Heteroskedasticity: (by Breush-Pagan Test)
3. No Autocorrelation: (by Durbin Watson test)
4. Normal Residuals (by Omnibus, Jarque-Bera )
5. No Multicollinearity (by correlation heatmap )
6. Exogeneity/no omitted variable bias (by intuition )

* Evaluating linear regression using metric such as: 1) Mean Absolute Error; 2)Mean Squared Error; 3) R2

## Project outcome

* Out of a potential 27 independent variables three were shortlisted (Two-point attempts (2PA) per 36 minutes, Assists(AST) per 36 minutes and Player Position (Pos)).
* Two linear regression models were created to predict TOV:
* Model 1: independent variables: 2PA, AST
* Model 2: independent variables: 2PA, AST and Pos

Table 1: Summary of Model 1 and Model 2

|  |  |  |
| --- | --- | --- |
| Model Result Summary | *Model 1* | *Model 2* |
| Mean absolute error | 0.405 | 0.399 |
| Root mean squared error | 0.529 | 0.525 |
| **R2(adjusted R2)** | **0.464(0.461)** | **0.473(0.471)** |
| Intercept | 0.706 | 0.602 |
| Coefficient of 2PA | 0.087 | 0.047 |
| Coefficient of FTA | 0.123 | 0.121 |
| Coefficient of AST | 0.166 | 0.167 |
| Coefficient of Pos\_C | -- | 0.223 |
| Coefficient of Pos\_PF | -- | 0.132 |
| Coefficient of Pos\_SF | -- | 0.055 |
| Coefficient of Pos\_SG | -- | 0.045 |
| Coefficient of Pos\_PG | -- | 0.147 |

# Problem Formulation

This section outlines how the step taken to better define the independent and dependent variables.

The goal of the project was to predict turnovers (or heading TOV in figure 1), so this was the dependent variable. To narrow down the independent variables some early feature selection and feature engineering was used:

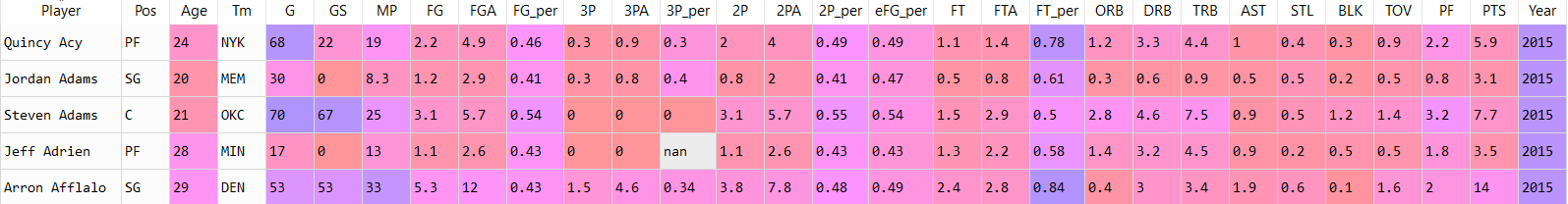


Figure 2: The early feature selection and feature engineering actions undergone

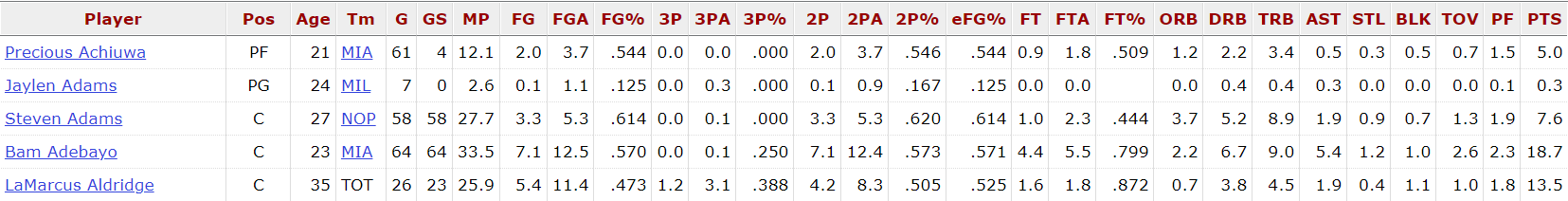
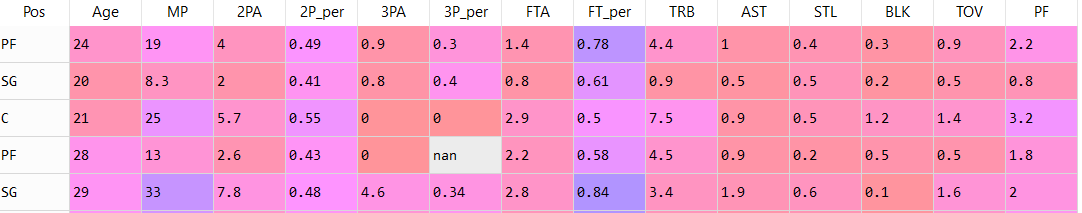
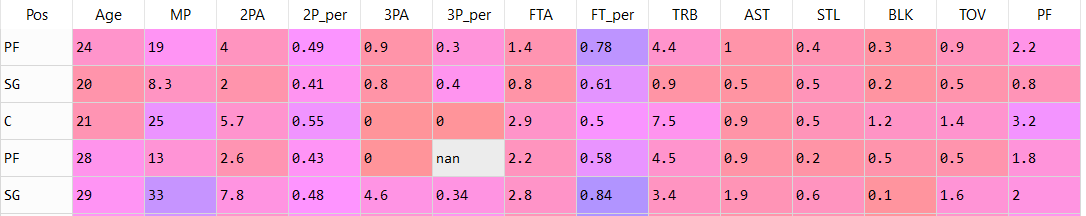


Figure 1: The per-game statistics for 5 players for the 2021 NBA season

1. **Feature Selection 1**: To remove any variable that would correlate with one another:

* The field goal statistics (e.g., FGx) were removed as they can be derived from the Two pointer and Three-pointer statistics (e.g., 2Px, 3Px) thus would be highly correlated
* The shots made statistics were removed (i.e., 2P, 3P and FT) as these highly correlate with shot attempts
* Offensive and Defensive rebound (ORB and DBR) were combined were removed whilst Total Rebounds (TRB) was kept as intuitively these statistics would be highly correlated.
* Points scored (PTS) was removed as this would correlate with the shooting attempts statistics (2PA, 3PA)

1. **Feature Selection 2**: To remove any statistics that intuitively that don’t relate to TOV

* Games (G), Games started (GS) and Team (Tm) were removed as these hardly indicate what a player does in a game

1. **Feature Engineering 1**: Addressing potential for omitted variable bias

**Heading abbreviations Explained:**

* **Pos** = Position; **Tm** = Team;
* **MP** = Minutes Played
* **G** = Games Played; **GS** = Games Started
* **FG** = Field Goals (made);
* **FGA** = Field Goals Attempted
* **FG\_per** = Field Goal percentage;
* **eFG\_per** = Effective Field Goal percentage;
* **3P** = Three pointers (made); **3PA** = …; **3P\_per** = …
* **2P** = Two Pointer(s) (made); **2PA** = …; **2P\_per** = …
* **FT** = Free Throws made ; **FTA** = … ; **FT\_per** = …
* **ORB** = Offensive Rebound; **DRB** = Defensive …
* **TBR** = Total Rebound; **AST** = Assists;
* **BLK** = Blocks; **TOV** = Turnovers;
* **PF** = personal fouls; **PTS** = points scored
* The remaining potential independent variables were:

Pos, Age, MP, 2PA, 2P\_per, 3PA, 3P\_per, FT, FTA, TRB, AST, STL, BLK, PF

* Except for the percentage statistics( 2P\_per, 3P\_per) and Age every independent variable and the dependent variable, TOV, is dependent on a player’s Minutes Played (MP)
* There is, thus, a causal link between MP and most variables in the model. However, MP cannot be removed because the model would suffer from omitted variable bias.
* To remedy this the effect of MP was dampened by converting the per-game statistics to per-36-minute statistics\*

**\*Note: This makes it so that each player plays that same number of minutes per-game by linearly extrapolating their current statics to their per-game statistics had they played 36 minutes. This is common practice amongst NBA statisticians. It assumes a linear relationship between MP and all per-game statistics which is not the case for the most part. However, there is no way to conduct an unbiased model without controlling for minutes played. To ensure the statistics of low-minute players were not exaggerated only players that played over 10 minutes had their statistics extrapolated.**

Outcomes of Problem Formulation

* The problem definition is clearer: The model will aim to predict the turnover committed per-36 minutes (**TOV**)
* The independent variables that will be investigated in the next phase are:

**Pos, Age, 2PA, 2P\_per, 3PA, 3P\_per, FTA, FT\_per, TRB, AST, STL, BLK, PF**

* All obvious cases of multicollinearity have been addressed
* Omitted variable bias has been largely limited. However, ‘Passing’ should probably be used as an independent variable instead of Assists (AST) as there is a causal link between the two, and it is better predictor of TOV than AST. However, this stat is not readily available
* Overall, the independent variables were reduced from 27 to 14 (48% reduction) using feature selection and feature engineering

# Data Pre-processing: Exploratory Data Analysis (EDA)

## Areas explored:

### Null values:

* Only the percentage statistics (2P\_per, 3P\_per, FT\_per) had null values
* The percentage of rows that were null were 0.015%, 10.25% and 0.83% for the rows of 2P\_per, 3P\_per and FT\_per respectively
* The cause of these nulls was a lack of shot attempts (i.e., 2PA, 3PA and FTA are zero) as these statics are the percentage of shots made compared to shots attempted
* To tackle the nulls three options (option 3 chosen) were considered:

1. Removing all rows were shot attempts were equal to zero

* **Pro:** easy fix
* **Con:** loss of at least 10% of row data

1. Setting all instances of nulls to zero

* **Pros:**

1. Easy fix;
2. Contextual argument: It could be argued that if a player does not attempt a specific shot it is because they do not have the license to. This implies that they would be highly inefficient (low shooting percentage) if allowed to take it. This is particularly true for difficult shots like three-point shots (3PA), and these nulls make up the bulk of null rows (10.25%)

* **Cons:**

1. This is an oversimplification. Whilst the player would probably have a low shooting percentage, there is no guarantee that a player’s shooting percentage would be zero
2. The logic used in pro b) is only good for three-point shots as they are the most difficult type of shot. For simpler shots like free throws, it is almost certain that a player’s shooting percentage would not be zero.
3. Setting all instances of nulls percentage and zero shot attempts to the median value of its column

* **Pros:**

1. No loss of data
2. Fairer allocation of player statistic than option 2

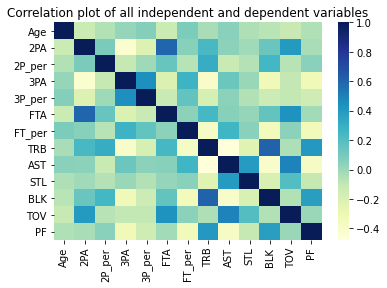
* **Cons:**

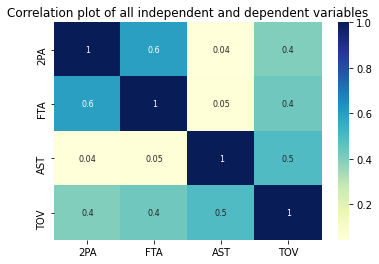
1. Harder to implement than option 1 and 2
2. Not grounded in reality; No way of knowing what a player’s shooting percentages would be; gives some players made up statistics;

### Further feature selection and engineering

* Utilised correlation heatmap to further reduce the number of independent variables
* Any variables that had a correlation coefficient of at least 0.4 (with TOV) was considered. This is low but for the subject matter the highest correlation observed was at least 0.5
* From the heatmap in figure 3 it is decided that all numeric variables except for 2PA, FTA and AST had a weak correlation with TOV thus were removed
* The correlation between FTA and 2PA of 0.6 was amongst the highest observed in the heatmap. Whilst this is not a strong correlation in a general sense it is relatively high for the model, so Multicollinearity was checked using variance inflation factor (VIF)
* VIF(2PA) = 1.55, VIF(FTA) =1.55 and VIF(AST) = 1.00
* All the VIF are below 5 thus the model will not be greatly affected by multicollinearity
* For this reason, FTA was removed from the independent variables

Figure 3: Correlation heatmaps showing the transition from 12 numeric independent variables to 3

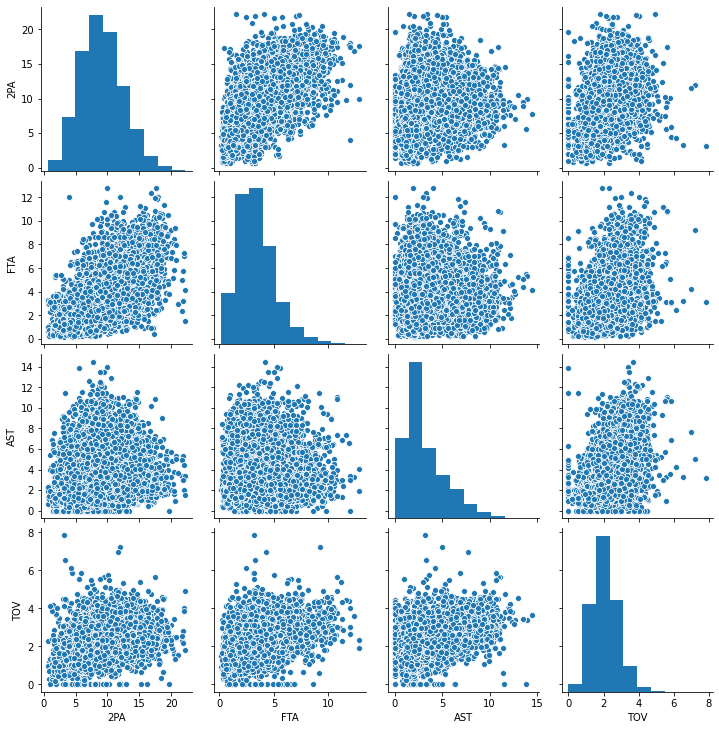




### Data Visualisation: Variable distribution and scatterplot

* A ‘pairplot’ (see figure 5) was used to get a broad sense of the distribution of the variables and the relationship between the finalised numeric independent variables(2PA and AST) and the dependent variable (TOV)
* Key observations:
* The distribution of 2PA and TOV appear somewhat normal but it is unlikely they will be perfectly normal due to the sheer size of data used
* AST and FTP displays a heavy right skew
* For the TOV v 2PA/AST graphs a lot of datapoints have TOV = 0 or AST =0. It is highly unlikely that a player plays commits zero turnovers or has zero assist playing 36mins per-game for a season.
* For the TOV v 2PA/AST graphs a linear fit is possible, but so is a polynomial fit. For the sake of simplicity, a linear fit was used
* For the TOV v 2PA/AST graphs heteroskedasticity may be present, but this will be tested when the independent variables are combined for the multiple linear regression model

Figure 4: Pairplot of the finalised variables in the model



### Categoric Data

* Unlike numeric data, the significance of categoric data (Position, Pos) was determine with several plots. Player position
* The first type of plot was a bar chart to gauge the distribution of positions
* All player position occurrences are between 2300 and 2600 thus no position is underrepresented

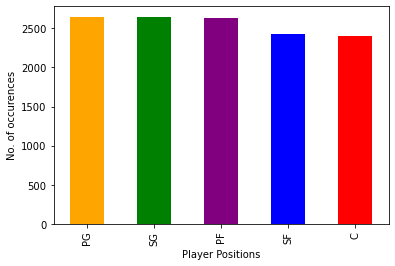
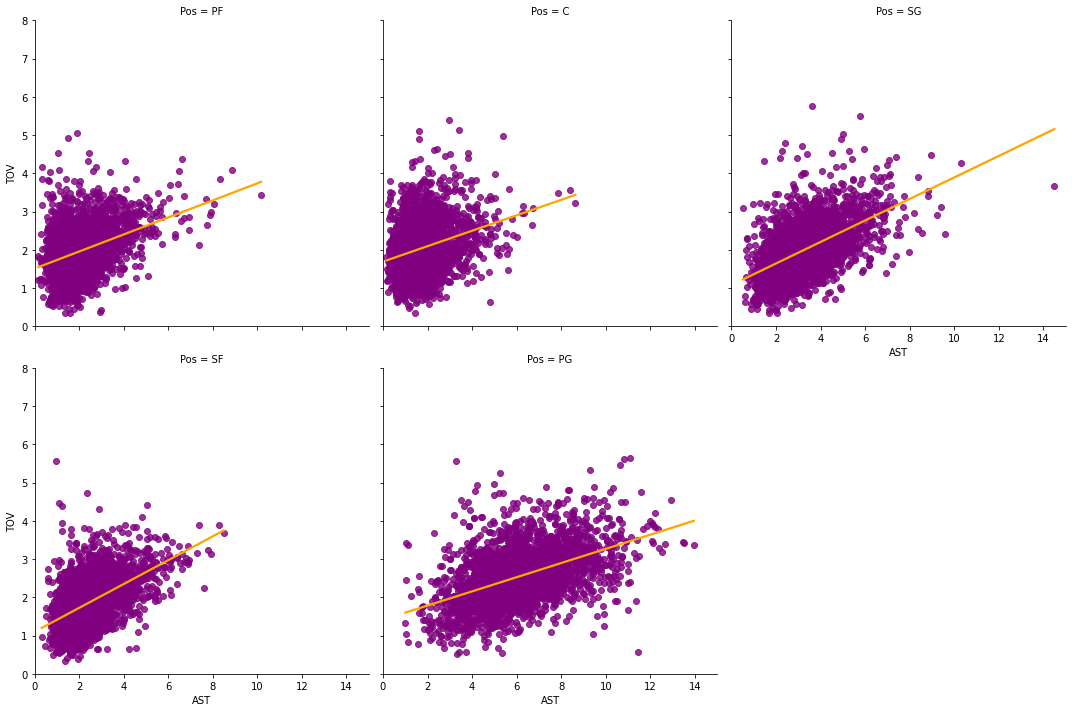


Figure 5: The occurrences of different player positions

* The second type of plot was ‘lmplot’ to gauge how a player’s position might affect the relationship between TOV and AST/2PA/FTA.
* For the TOV v AST relationship:
* Shooting Guards(SG) and Point Guards(PG) display a clear positive linear relationship
* Power Forwards(PF), Centres (C) and Small Forwards (SF) appear to cluster at lower AST thus indicating that these types of player(especially C and PF) are unlikely to make many assists
* A polynomial best fit curve might be better for PF, C and SF and a different model could be conducted for these types of player. However, for the sake of simplicity a linear curve was chosen.

Figure 6: The relationship between Turnover(TOV) and Assists(AST) for the different player positions (Pos)



* TOV v 2PA and TOV v FTA have similar trends. Both graphs show a clear linear relationship for all positions
* The third type of plot was a box and whisker plot. It was created to spot any severe outliers in the numeric variables and to determine if player position impact turnovers in any way not explained by 2PA and AST. From the figures 9-11
* No severe outliers were spotted
* From the boxplots there is some evidence that player position influences TOV in ways that is not explained by 2PA and AST:
* E.g.: Centres (C) had the 2nd highest median Tov despite having a median 2PA that was just 11% more other player positions and a median AST that was 47% less than the median for the other player positions.

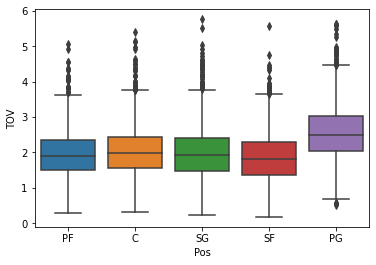


Figure 8: The spread TOV across all positions

Figure 7: The relationship between Turnover(TOV) and Two-point shots attempts (2PA) for the different player positions (Pos)

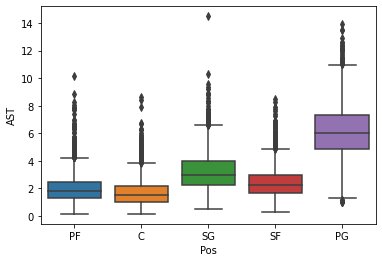
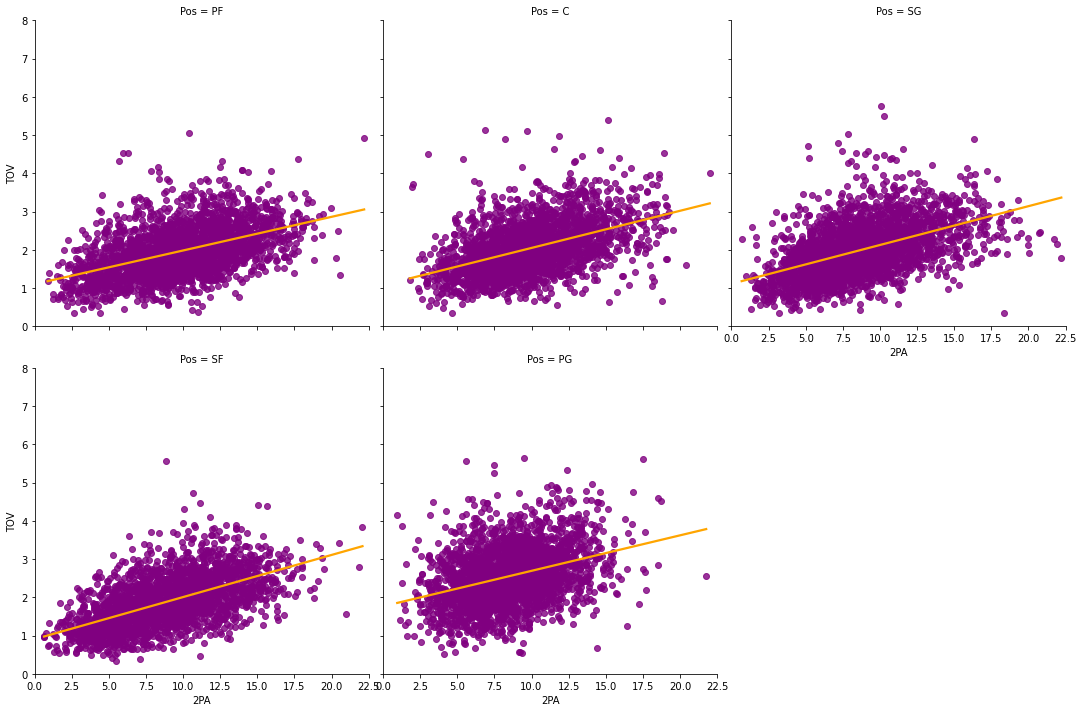


Figure 9: The spread AST across all positions

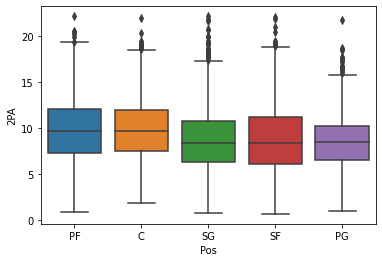


Figure 10: The spread 2PA across all positions

## Summary of data pre-processing phase

* The number of numeric independent variables was reduced from 12 to 3 (75% reduction) based on an analysis of correlation heatmap
* Using data visualisation extreme outliers were spotted and removed
* It was found that Player Position (Pos) has some effect on the TOV committed by a player though the extent of this effect was unknown from simple data exploration.
* To quantify the effectiveness of Pos two models were created. The first model used the independent variables: 2PA , FTA and AST to determine TOV whilst the second model used independent variables: 2PA, FTA, AST and Pos to determine TOV.

# Model Creation

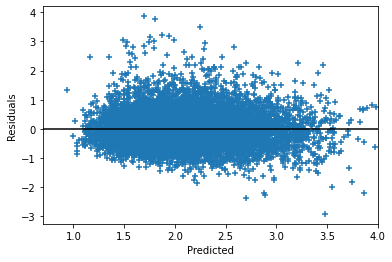


Figure 12: Residual plot for Model 1

* Two models were created to predict turnovers, TOV:

**Models:**

* Model 1: Independent variables = 2PA, AST, FTA
* Model 2: Independent variables = 2PA, AST, FTA, Pos
* The data was split into training-test splits of 80-20%

Table 2: Model Summary

|  |  |  |
| --- | --- | --- |
| Model Result Summary | *Model 1* | *Model 2* |
| Mean absolute error | 0.405 | 0.399 |
| Root mean squared error | 0.529 | 0.525 |
| **R2(adjusted R2)** | **0.464(0.461)** | **0.473(0.471)** |
| Intercept | 0.706 | 0.602 |
| Coefficient of 2PA | 0.087 | 0.047 |
| Coefficient of FTA | 0.123 | 0.121 |
| Coefficient of AST | 0.166 | 0.167 |
| Coefficient of Pos\_C | -- | 0.223 |
| Coefficient of Pos\_PF | -- | 0.132 |
| Coefficient of Pos\_SF | -- | 0.055 |
| Coefficient of Pos\_SG | -- | 0.045 |
| Coefficient of Pos\_PG | -- | 0.147 |

## Conclusion:

* The R2 only increased by 2% from model 1 to 2 thus adding player position does little to explain the variance between the best fit curve and the data points.
* R2 = 0.46 suggests that the best-fit hyperplane explains its variation with the actual data points only 46% better than the mean TOV line. There are a few reasons for this:
* The data may require a non-linear fit
* Key independent variables have been omitted such as ‘passes made per 36mins’
* The models created do not factor in a player’s skill (e.g. passing skill) which will ultimately determine how likely a player is commit turnovers. Including player position was an (unsuccessful) attempt at modelling skill which is abstract and hard to quantify. The idea was that different player positions will have different skillsets which affect how likely they are to commit TOVs. However, the results show that there is variance in skill level even within player categories.
* The errors are small. RMSE 0.55 which is only larger than 0.4% of all TOV observations
* The error metrics show that there is low absolute error (i.e. small RMSE) between the best-fit hyperplane and the datapoints. However, when analysing the error in relative terms it is notable.
* For context, the RMSE of model 1 is 0.529 which is a small value, but it is roughly 25% of the mean TOV is observed.
* It should be noted though that the RMSE is only larger than 0.27% of all TOV observations (i.e. 35 out of 12738 ) which agrees with the leptokurtic TOV distribution observed in figure 4.
* Both models have low variance. For Model 1 the RMSE for the training set is 0.535, whilst for the test set it is 0.529.

## Model Evaluation

1. Testing for heteroskedasticity

Using the Breush-Pagan test it was found that at 95% confidence interval NO heteroscedasticity was present in either model (p-value = 0.000). The residual plots also agreed with this conclusion (see figure 12)

1. Testing for normality of residual

As the number of observations was great standards tests such as the Shapiro-Wilks test or Omnibus could not be relied upon because for large datasets as they are highly unlikely that to display perfectly normal distributions. As a result, histogram plots and QQ-plots were used to gauge normality.

* The mean of the residuals was -1.876e-16 for model 1.
* From figure 14 it is evident that the residuals have a moderate right skew and are leptokurtic
* This is not a severe issue as the dataset used is fairly large and the degree of non-normality is low

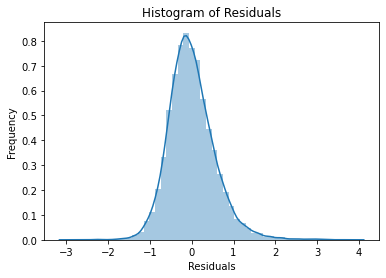


Figure 11: Histogram of residuals for Model 1

Skew = 0.836

Kurtosis (Pearson) = 5.957

1. Predicted V Actual TOV

* The model tends to overpredict when TOV committed is low and underpredict when the TOV committed is high
* This could be an indication that the linear fit used is bias.

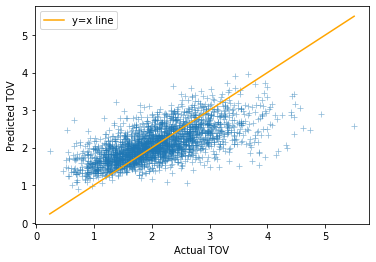


Figure 13: A comparison of predicted v actual for TOV Model 1

1. Cross Validation

* To assess the validity of the results generated from the train-test split (i.e. R2, RMSE, MAE ) 5-fold cross validation was used.
* The means for all metrics used for cross validation were similar to the resulting metrics from the train-test split

Table 3: Result Summary for the Cross Validation of model 1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Metric (Model 1) | *Fold 1* | *Fold 2* | *Fold 3* | *Fold 4* | *Fold 5* | *Mean* |
| Mean absolute error (MAE) | 0. 413 | 0. 405 | 0.386 | 0.403 | 0.422 | 0406 |
| Root mean squared error (RMSE) | 0. 541 | 0.529 | 0.508 | 0.536 | 0.565 | 0.536 |
| R2 | 0.530 | 0. 515 | 0.423 | 0.367 | 0.366 | 0.440 |