IST 707 Applied Machine Learning

By Prof. Kelvin King



Time Series Analysis

Submitted by: Samarth Sandesh Mengji SUID: 718473878 NetID: smengji@syr.edu

Date of Submission: 12/3/2023  
Syracuse University – School of Information Studies

**Time Series Analysis**

**Data Exploration:**

A graph of different sizes of houses

Description automatically generated with medium confidence

**Introduction:** Baseball, as a sport, boasts a diverse array of players, each contributing to the dynamic nature of the game. In this report, the first thing we delve into is the demographic and performance characteristics of baseball players, focusing on key metrics such as age, Batting Average (BA), Home Runs (HR), and Runs Batted In (RBI).

**Demographic Overview:** Analysis reveals a predominant concentration of players in their mid to late 20s, reflecting the prime years for athletic performance. This concentration suggests that teams predominantly invest in players who are at their physical peak.

**Performance Metrics: Batting Average (BA)** The analysis of Batting Average (BA) provides insights into the offensive capabilities of players. The distribution of BA is notably skewed to the right, indicating that many players have a batting average below 0.3. This skewness suggests that while there are standout hitters with high averages, the majority of players exhibit batting averages that place them below the league average.

**Performance Metrics: Home Runs (HR)** Home Runs (HR) is a measure of a player's power, exhibits a heavily skewed distribution to the right. This skewness emphasizes that most players hit a relatively small number of home runs, with a select few achieving notable power-hitting status. This distribution aligns with the general trend observed in baseball, where hitting a high number of home runs is a rarity, reserved for elite power hitters.

**Performance Metrics: Runs Batted In (RBI)** Similar to Home Runs, the distribution of Runs Batted In (RBI) is skewed to the right. This skewness implies that the majority of players contribute to a low number of RBIs, while a smaller fraction achieves higher RBI totals. This aligns with the structure of baseball offenses, where a few key players (stars) often drive in the majority of runs.

**Overall Analysis:** In summary, the concentration of players in their mid to late 20s, coupled with skewed distributions in Batting Average, Home Runs, and Runs Batted In, reflects the unique combination of youth, experience, and performance heterogeneity that characterizes the world of baseball. This insight can prove valuable for teams, analysts, and enthusiasts seeking a deeper understanding of the trends within the player landscape.

A graph with a line

Description automatically generated

**Average Batting Average per Season: Unveiling a Slight Downward Trend**

The analysis of the average Batting Average per season reveals a downward pattern. While variations are expected, a downward trend is observed over the years. This trend prompts us to explore potential factors contributing to this phenomenon. Possibilities include improvements in pitching strategies, advancements in training and technology, or shifts in playing strategies that have collectively impacted the overall batting averages in baseball. A deeper examination of these factors could unveil insights into the evolving dynamics between pitchers and batters.

A graph with a line

Description automatically generated

**Average Home Runs per Season:** Contrary to the trend observed in Batting Average, the average number of Home Runs per season showcases a notable upward trajectory. This trend suggests a shift in the game dynamics, with players demonstrating an increased capacity for power hitting in recent years. The reasons behind this trend may be because of change in player training approaches, alterations in bat and ball technologies, or adjustments in ballpark dimensions that favor hitters. The surge in home run production could also be a response to alterations in pitching strategies, as batters adapt to capitalize on favorable conditions.

**AutoRegressive Integrated Moving Average (ARIMA)**

Next, we aimed to conduct a thorough time series analysis of player performances, leveraging the AutoRegressive Integrated Moving Average (ARIMA) model for forecasting. Our approach involved aggregating player data by season and employing statistical tests and visualizations to guide the modeling process. The goal was to provide meaningful insights and predictions for the next seasons.

**Data Inspection and Aggregation:** Upon revisiting the dataset, we identified the 'Season' column as a potential time component. Recognizing that the data contained multiple entries per season due to various players, we opted to aggregate the data by 'Season' and calculate average performance metrics. This step ensured that each row in the new data frame represented a season, providing a foundation for meaningful time series analysis.

We proceeded with this approach and created a new data frame where each row represents a season, and the columns are the average performance metrics for that season.

**Stationarity Check:** Before applying time series models, it is imperative to confirm stationarity, ensuring that the statistical properties of the time series remain constant over time. The Augmented Dickey-Fuller (ADF) test was employed on the 'Runs' (R) metric, yielding a p-value of 0.0086. The significance of this p-value (< 0.05) indicated that the 'R' time series is stationary, setting the stage for further analysis.

The Augmented Dickey-Fuller test is commonly used to determine whether a time series is stationary or not. Stationarity is a crucial assumption in many time series models. A stationary time series is one whose statistical properties, such as mean and variance, remain constant over time. Non-stationary time series, on the other hand, may have trends or seasonality that can affect the reliability of statistical analysis.

The regression parameter in the ADF test allows you to control for deterministic components in the time series that can influence the results. Here's a breakdown of the options:

* **'c' (constant):** This includes an intercept term. This is used when we suspect a constant term in the time series.
* **'ct' (constant and trend):** This includes both an intercept and a trend term. This is used when we suspect both a constant and a linear trend in the time series.

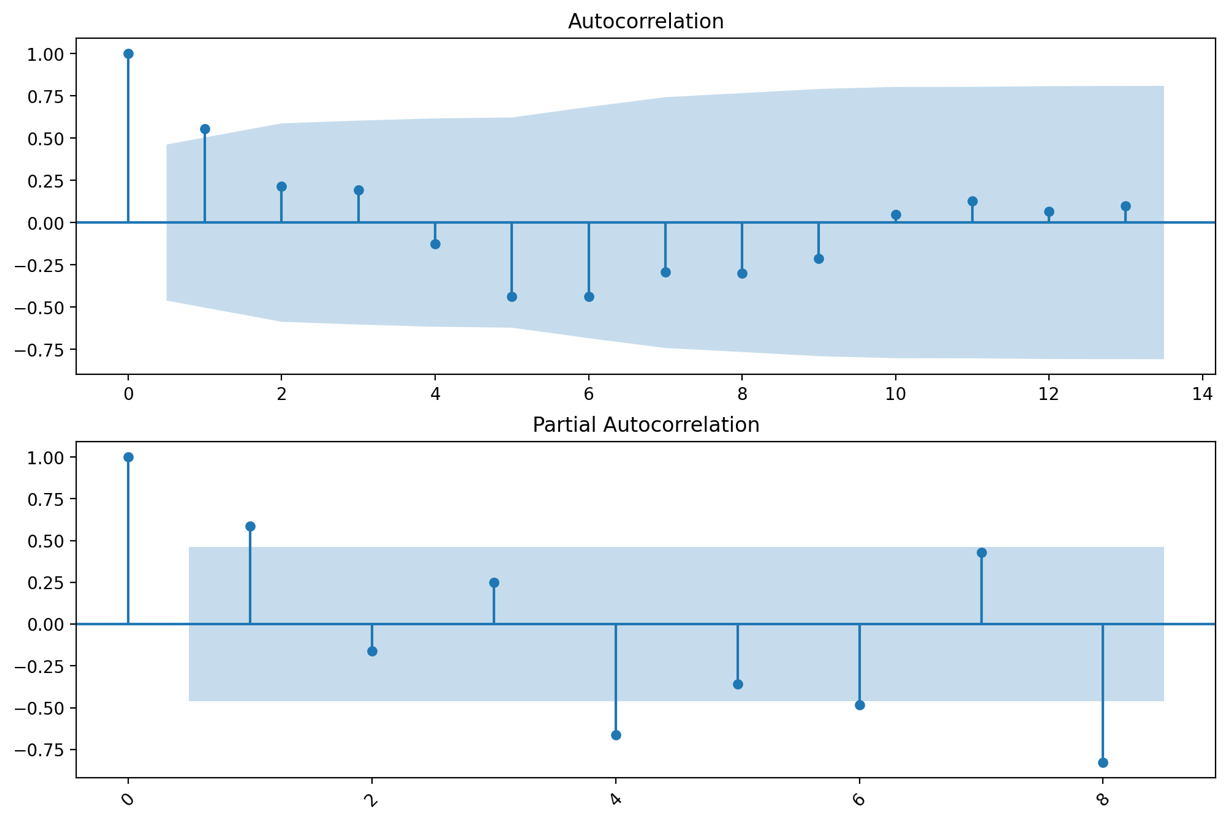
By default, the **adfuller** function in **statsmodels** uses **regression='ct'**, meaning it tests for the presence of both a constant and a trend in the data and we have used the default setting for our project.

Now that we have a stationary time series, we can proceed with the ARIMA model for forecasting. However, ARIMA requires us to specify three parameters: the order of autoregression (p), the degree of differencing (d), and the order of moving average (q). These parameters can be determined using plots of the autocorrelation function (ACF) and partial autocorrelation function (PACF), or by using grid search.

**Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF)**

We are first going to try by ACF and PACF. Let's plot the ACF and PACF for 'R' to help determine the parameters for the ARIMA model.

Here are the **Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF)** plots for the 'R' (Runs) time series:



**ARIMA Model Parameter Selection:** Determining the appropriate parameters (p, d, q) involves analyzing the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots. Our decision to initiate with an ARIMA (1,0,0) model for 'R' was justified by the ACF and PACF plots, where a significant PACF spike at lag 1 guided the autoregressive order.

Based on these plots, we can start with an initial model of ARIMA (1,0,0), where:

* p (order of autoregression) = 1, as suggested by the PACF plot
* d (degree of differencing) = 0, as our time series is already stationary
* q (order of moving average) = 0, as suggested by the ACF plot

Let's fit this ARIMA model to our 'R' time series and evaluate its performance.

**ARMIA (1,0,0) Model Summary:**

A screenshot of a data sheet

Description automatically generated

**Model Evaluation:**

The **p-value for the AR (1) term is 0.034,** which is less than 0.05, indicating that it is statistically significant.

The **AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion)** are measures of the goodness of fit of the model, with lower values indicating a better fit. The AIC and BIC for this model are **99.511 and 102.182**, respectively.

The **Ljung-Box** test checks for autocorrelation in the residuals, with a null hypothesis that the residuals are independently distributed. The **p-value of the Ljung-Box test is 0.70**, which is greater than 0.05, suggesting that we cannot reject the null hypothesis and the residuals are independently distributed.

The **Jarque-Bera** test checks for normality in the residuals, with a null hypothesis that the residuals are normally distributed. The **p-value of the Jarque-Bera test is 0.83**, which is greater than 0.05, suggesting that we cannot reject the null hypothesis and the residuals are normally distributed.

**Residual Analysis:**

Overall, the ARIMA (1,0,0) model seems to be a good fit for the 'R' time series. However, we should also check the residuals to ensure that there are no patterns left unexplained by the model. Let's plot the residuals.

Here is the plot of the residuals of the ARIMA (1,0,0) model:

A graph with a line

Description automatically generated

The residuals seem to be randomly distributed around zero, with no obvious patterns. This suggests that the ARIMA (1,0,0) model has captured the underlying process of the 'R' (Runs) time series well.

Now, let's use this model to make forecasts for the next 5 seasons. We can do this by calling the get\_forecast() method of the fitted model. This method returns an object that contains the forecasted values, the standard error of the forecasts, and the confidence intervals of the forecasts. Let's plot the forecasts along with the confidence intervals.

**Forecasting**

Here is the forecast for the next 5 seasons, along with the original 'R' (Runs) time series:

A graph with blue lines

Description automatically generated

The blue line represents the original 'R' time series, the orange line represents the forecasted values, and the shaded area represents the confidence intervals of the forecasts.

As we can see, the ARIMA (1,0,0) model predicts a constant line in 'R' over the next 5 seasons. There is a considerable amount of uncertainty in these forecasts, as indicated by the wide confidence intervals.

**Best parameters for ARIMA model using Grid Search**

As mentioned above in the report, the ARIMA model is defined by three parameters: p (order of autoregression), d (degree of differencing), and q (order of moving average). We performed a grid search over all possible combinations of p, d, and q in the range [0, 2]. For each combination, an ARIMA model was fitted, and the Akaike Information Criterion (AIC) was used to assess model performance.

**Results:** The best-fitting ARIMA model was found to have parameters (2, 2, 0), indicating a second-order autoregressive component, a second-order differencing, and no moving average component. The corresponding **AIC and BIC** for this model were **92.94 and 95.26**, respectively. These metrics are measures of the model's goodness of fit, with lower values indicating better performance.

A screenshot of a cell phone

Description automatically generated

**Model Fitting:** The best fitting ARIMA model, determined to be ARIMA (2, 2, 0), was fitted to the 'R' (Runs) time series data for player performance across seasons. This specific configuration indicates a second-order autoregressive component and a second-order differencing.

**ARMIA (2,2,0) Model Summary:**

**A screenshot of a data

Description automatically generated**

**Model Specification: ARIMA (2, 2, 0)**

**Introduction:** The model is an AutoRegressive Integrated Moving Average (ARIMA) model designed to analyze and forecast the time series of variable R.

**Dependent Variable:** R (Number of observations: 18)

**Model Order:** ARIMA (2, 2, 0): The model includes two autoregressive (AR) terms, two differencing (I) operations, and zero moving average (MA) terms.

**Model Estimation:** The model is using the maximum likelihood method which is the default of fit method in ARIMA.

**Model Coefficients:**

AR.L1: Coefficient = -0.5991, Standard Error = 0.148, z-value = -4.054, p-value < 0.001

AR.L2: Coefficient = -0.7771, Standard Error = 0.235, z-value = -3.308, p-value = 0.001

**Residual Variance:**

Residual variance (sigma2) = 11.8524, Standard Error = 5.503

**Diagnostic Tests:**

* Ljung-Box (L1) (Q): 0.15, p-value = 0.70
* Jarque-Bera (JB): 0.31, p-value = 0.86
* Heteroskedasticity (H): 2.63, p-value = 0.31

**Model Fit Statistics:**

* Log Likelihood: -43.470
* Akaike Information Criterion (AIC): 92.940
* Bayesian Information Criterion (BIC): 95.258
* Hannan-Quinn Information Criterion (HQIC): 93.058

**Covariance Type:** OPG (Outer Product of Gradient)

**Conclusion:**

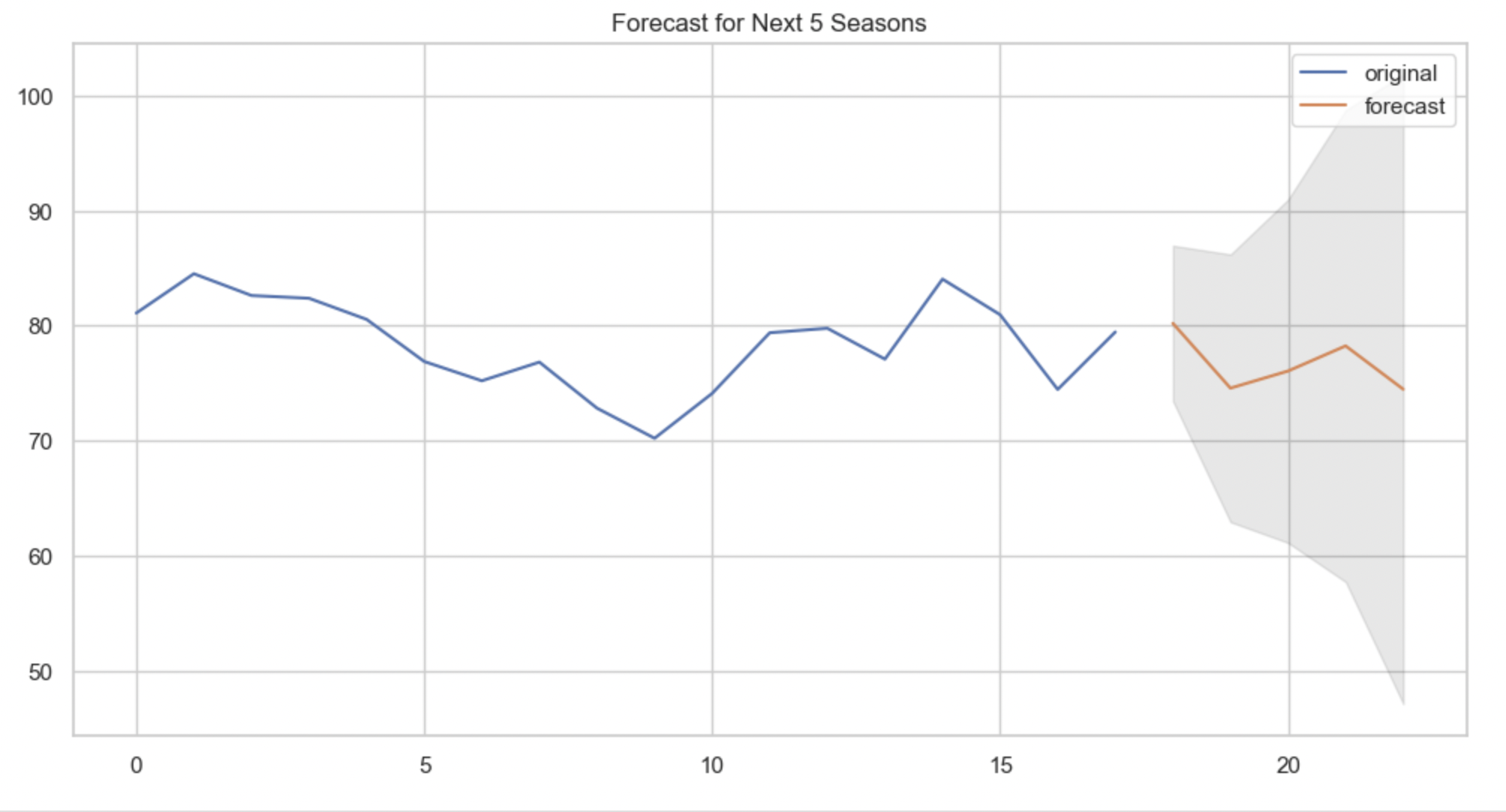
The ARIMA (2, 2, 0) model captures the temporal dependencies in the time series R. The diagnostic tests indicate a reasonably good fit, with no significant autocorrelation in the residuals (Ljung-Box Q-test) and no evidence of non-normality (Jarque-Bera test). The model's AIC, BIC, and HQIC values provide a basis for model comparison and selection, with lower values indicating a better fit.

**Model Evaluation:**

1. The **p-value for the AR (1) term is 0.00,** which is less than 0.05, indicating that it is statistically significant.
2. The AIC and BIC for this model are **92.940 and 95.26**, respectively.
3. The **p-value of the Ljung-Box test is 0.15**, which is greater than 0.05, suggesting that we cannot reject the null hypothesis and the residuals are independently distributed.
4. The **p-value of the Jarque-Bera test is 0.31**, which is greater than 0.05, suggesting that we cannot reject the null hypothesis and the residuals are normally distributed.

**Forecasting for ARIMA (2,2,0)**

Here is the forecast for the next 5 seasons, along with the original 'R' (Runs) time series:



The blue line represents the original 'R' time series, the orange line represents the forecasted values, and the shaded area represents the confidence intervals of the forecasts.

As we can see, the ARIMA (2,2,0) model predicts in 'R' over the next 5 seasons.

We can see that the forecasted runs will be around interval of 70 to 90.

Based on these forecasts can decide whether to buy a player or renew a contract.

**Recommendation and Conclusion:** This time series forecasting approach, incorporating model fitting, summary analysis, and visualization, equips stakeholders with a robust tool for anticipating player performance in the upcoming seasons. The ARIMA model, validated through detailed statistical analysis, contributes to informed decision-making in the dynamic context of sports analytics. Continued monitoring and refinement of the model may further enhance its predictive accuracy over time.