Advances in Economics, Business and Management Research, volume 215
Proceedings of the 2022 7th International Conference on Social Sciences and Economic Development (ICSSED 2022)

Prediction of Baseball Average Ticket Price in a Year

Peizhao Li[1,*]

[1]*School of science and engineer, The Chinese university of Hong Kong, China, 450000*
**Corresponding author. Email: 120090622@link.cuhk.edu.cn*

ABSTRACT
Baseball is one of the most popular sports events in the world[1]. Although most people have television now, there are still some people who prefer watching the baseball game at the scene. In this research paper, the research is going to use model to predict the baseball average ticket price. In the procession of prediction, the research is going to use the linear regression model to predict. Also, in the procession of the analyze, there are still some problems in the data. So the research is trying to use several methods to solve it. To make the result as accurate as possible. The research will only predict the result in 2022. The model's prediction is not the real result, so the result's prediction after 2023 might have a little high error. If we want to predict it more exact, we should know the precise ticket price in 2022. So this research, research will only predict the average ticket price in 2022. In this research, the research also considers the error. So the research uses the standard score to find if there is any abnormal data. This prediction result may provide some guidance to consumers. And the research finally give the result is 36.73.

Keywords: *Baseball ticket, Average price, Linear regression, Prediction, Standard score.*

1. INTRODUCTION

Baseball is the world's eighth most popular sport. People buy tickets for it to sense its charm. So the research is interested in the average ticket price for the baseball game. Since 2015, the baseball average ticket price is increasingly higher. But it increases irregularly. This brings a challenge for the consumer to predict its average price every future year. Considering the average price is a great data to predict, factors that affect one play game have no such huge influence on the whole data, so the research is going to use model to predict the baseball average ticket price in a whole year. So in this research, the first thing needs to do is to make sure which model is the most fitted. Then the research needs to use the model to predict the future average ticket price for the baseball game. In the procession, the research may face some errors in the model. So the research will also consider how to deal with these errors. Also, the research will consider if there is any other better way to reach the result. What the reader needs to pay attention to is that the result is just a predicted value, so it can't be to exact as the fact.

2. RESEARCH DESIGN

2.1 Data source

The following is a chart for major league baseball average ticket price from 2006 to 2021[2](in u.s dollars)

Table 1. the average ticket price for baseball game

Year	U.S.Dollars
2006	22.21
2007	22.77
2008	25.43
2009	26.64
2010	26.74
2011	26.91
2012	26.98
2013	27.48
2014	27.93
2015	29.94
2016	31.00
2018	32.44
2019	32.99
2020	34.04
2021	34.21

This chart gives the data from 2006 to 2021, this is vital for our prediction in the future years. The reason why the research choose this chart is because the data gives a long time range for this research. This is a benefit

for the following prediction work. From this chart, we can also find that the data in every year is increasingly higher at about 2% to 13%. There is one thing that we should pay attention to, the average ticket price in 2017 is missing. That means this group of data has a missing value. So whether we consider the missing value becomes a problem. But we have to say that this chart does not even give me a specific value, but also gives me an idea of this prediction model. By the firstly thinking, the linear regression should be a fitted model to predict the data in the future years. But linear regression needs continuous data in group. So we have to think about how to deal with this missing value.

2.2Another example of ticket

One report is from Spatial-Temporal prediction models for active ticket managing in data centers[3]. This reported work gives us different kinds of models for every situation. Reading on this report, this research on predict the ticket for a baseball game may fit with the linear regression. This report also gives many data examples on different situations. It also tells how to use these models and handle these models. What the research needs to do is to choose an accurate model and improve the model to make the prediction more exactly. Also, this report is predicting the data on ticket price which is similar to what the research is going to predict. But it didn't give a specific case to predict. This report is based on these models to give a prediction on a specific case, which may have more value for the consumer to have a standard. Another report is from Airline ticket price and demand prediction: A survey[4]. This report gives a prediction on the airline ticket. Our reports both predict the price of the ticket. But our case has a little different. The price of airline ticket variation is different from the ticket price on baseball. So their prediction model is different. In this report, the author uses linear regression to predict the airline ticket price. But the effect of the factors on the airline has more than two. So the linear regression is multiple linear regression. But in this group of data, there is only a one-factor effect on the data. So in this report, simple linear regression is more fitted with this group of data. This report also reminds me that not all the average price is increasingly getting higher, so different data prediction has no relation. The prediction and the model should base on the situation and the changing of the ticket price.

2.3The material about the model

The book is from Linear regression analysis[5]. In this book, the author introduces the linear regression model in detail. It also tells the fitting situation for the group of data. By comparison, this report's group of data fitted with the characters in general. But the missing data in 2017 should be made up. By this introduction of the book, the research finally decided that my analyzing should be based on the linear regression model. And the missing data in 2017 can also use this method to make up. Another book is Applied linear regression[6]. In this book, the author introduces how to use the linear regression. It gives this research the formula and which situation is fitted with the linear regression. It list that the linear regression needs a group of continuous data, which makes this research more firmly to use this model to predict the final results.

2.4 Linear regression

The research has decided that in this group of data, the linear regression model is fitted. Because the data is increasing high in general, all the data in the group are continuous except for the data in 2017. So firstly we should solve the missing data in 2017. By solving this, we can also use linear regression to solve it. In this part, the research needs to use the data from 2006 to 2016 to make a linear regression model. Then the research will use this trend line to predict the data in 2017. Although there might be a little error. But this method can reduce the error as possible. Also, it makes up the missing value in 2017. Then the research will use the data from 2006 to 2021 to make a new linear regression model. This time there will be a new trend line. Then we can use this trend line to predict the result in 2022. However, before start building the model, the research should first check if there is any abnormal data. If there do have abnormal data, the research should ignore it and use the same method as 2017 to give it a new result. The research can't let this affect the final result.

Then the research should discuss what we need for the linear regression model. The trend line is Y=AX+B. The $B= \overline{Y}-A\overline{X}$. The $A=(\sum_{i=1}^{n} xi * yi - \overline{x} * \overline{y})/\sum_{i=1}^{n} xi^2 - \overline{x}^2$. From the formula, we know that we should first calculate the average of the ticket price and the average of the year. What's more, the research will pay attention that we need check the abnormal data. The research decide use the standard score to check if there is any abnormal score. The standard score $z=\frac{xi-\mu}{\sigma}$.So then we need to know the standard deviation of the group of data. Considering that the year won't have any abnormal. So this part only calculate the average ticket price in every year. The $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(xi-\bar{x})^2}{n}}$. Then we set the standard score as [-3,3], then if the data' s standard score is out of the range, we will regard it as an abnormal data.

3. ANALYSIS

Firstly, the reason the research why choose the linear regression model is as follows. The model was chosen mainly depended on the chart. So figure 1 is as follows. And the analyze of the data is mainly detected on the trend line.

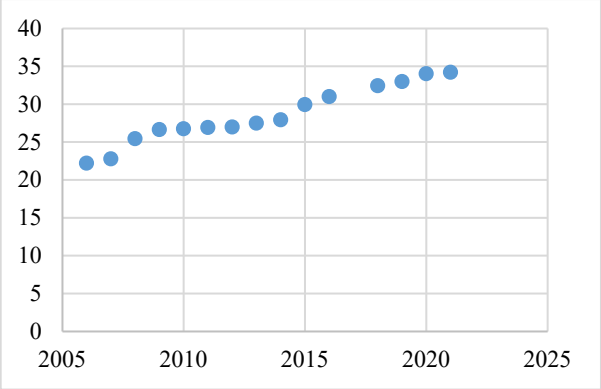

Figure1. Initially scatter diagram

Let the independent variable be the x, dependent variable be the y. From the chart, it can show that the x is the time of the year, the y is the price of the ticket. So there is only one factor influence on the price of the ticket. Also, the chart can show that the price of ticket is increasingly higher by the time going. Also, the data is continuous in general. Although there is a missing data in 2017, but that's a thing the research will solve in the following part. So considering all the things above, the research may use linear regression is the most fitted model to predict the future price firstly.

Secondly, the data from this chart miss the value of 2017. But the work of the following should make sure if the research do need the missing data on 2017. From the above related work, the linear regression need a group of continuous data. And this research has decided to use the linear regression to give the result of prediction. So if the research wants to get a more accurate value, the following work should fill the value in 2017. To make the error smaller, the research needs to predict the value of 2017 first. To get the value as accurate as possible, the research can use linear regression to predict the value in 2017. So let Y=BX+A. Then we calculate the and $\bar{y}$. The =2011, $\bar{y}$=26.73. By using the formula give before. Then the research can get the B=0.74591, A=-1473.29501. Then substitute 2017 in the Y=BX+A, the research can predict the value on 2017 is 31.20. So the research can get the data and the new chart is as follow:

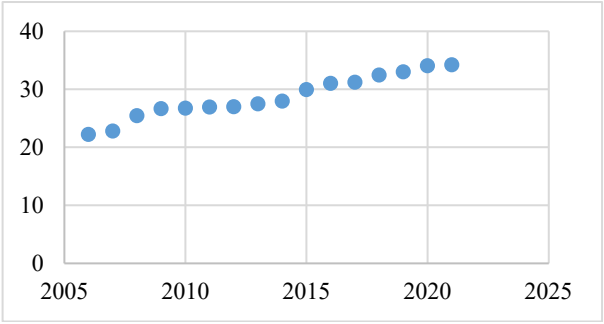

Figure2. The scatter diagram which adds the data of 2017

Thirdly, the research still needs to check if there is any abnormal data. The x represents the year, so there is no abnormal in x. Just as what the research analysis before, the research will use the standard score to check if there is any abnormal data. By using the formula before. The research needs to first calculate the and σ. After calculating, the $\bar{y}$=28.68, the σ=3.61. To make the error smaller, the research need to set a standard score. [-1,1]is too small to exclude the true abnormal data. However, it may exclude the normal float number. [-4,4]is too big for the standard score, which may leave out the abnormal number. Considering all the factors, choosing [-3'3] is a great standard score for the research. Then set the standard score from -3 to 3. Now the research needs to calculate the standard score for each data(z). By using the formula before, the research can get standard score for each data as following. z1=1.79, z2=1.63, z3=0.90, z4=0.56, z5=0.53, z6=0.49, z7=0.47, z8=0.33, z9=-0.34, z10=0.20, z11=-0.64, z12=-0.69, z13=-1.04, z14=-1.19, z15=-1.48, z16=-1.53. All the standard scores are in the range[-3,3], so all the data in the chart are the normal data. It means all year's ticket price in the statistics is increasing at a normal speed. And there is no abnormal ticket price to influence the final result. So the research needn't delete any of them. Last, what the researchers need to do is use the linear regression model to make this prediction. Let y=bx+a, by using the formula before, the research can get $\bar{x}$ =2013.5, $\bar{y}$=28.68. b=0.947, a=-1878.1. So the linear regression is y=0.947x-1878.1. By this linear regression model, substitute the 2022 in x, then the research can predict the price of baseball game ticket in 2022 is about 36.73. the research can also use this model to predict future years. But if the research wants more precise value in 2023 or farther future. The research needs the accurate value in 2022. But the research can't get that data. So the model needs a small adjustment every year.

Add all the elements into the chart, we can get the final chart as follow:

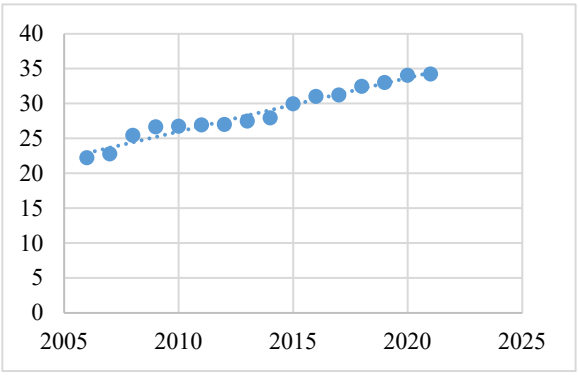

Figure3. The linear regression for these data.

What readers need to pay attention to is that the data in 2017 is missing. So the data is a predicted value. So the value predicted in 2022 may have a little error. But it's a great way to predict the average ticket price when losing one data. Now the reader can now use this trend line to predict the new average price for the baseball

game ticket. Also, you can predict it in 2023 or the future, but it may have a little gap between the prediction and the fact.

4. CONCLUSION

This research uses linear regression to predict the average ticket price for the baseball game. In the procession, the research uses linear regression to fill up the missing data in 2017. Also, the research uses the standard score to check if there is any abnormal data. Finally, the research uses linear regression to predict the data on 2022 is 36.73. The average ticket price for a baseball game is increasingly higher every year. This could be due to the overall price increase in the country. Because this is the average ticket price, the other reason which affects the ticket price will have little effect on the average price. Because of this, the prediction of the price will only be related to the year and the error will be small. However, the error still exists, and in this analysis, I didn't consider it. So in future work, the research still needs to consider the error, and calculate it. And the research will try to make a new model make the error smaller. Also, there still exists one problem. The data in 2017 is missing, and this research's method is using the trend line from 2006 to 2016 to predict it. This predicted result may not be the best. So in future work, the research will try to use other methods to predict it and compare their error. The prediction in 2017 will also affect the final results for the trend line. From the final result of the average ticket price, the custom's passion for this sports event has greatly increased this year. Also, it means that the major league of baseball will earn much more money in 2022 than in 2021.

So in future work, the research will try to use a model to calculate the error and try to use a new model to reduce this error. Also, the research will try to predict a more accurate result for the 2017 average ticket price. What's more, the research still need to look up for more information on the website to search for better way to make the model more accurate.

The average ticket price for baseball match is increase. This shows that people do have more and more passion on this sports event. The research has looked up for many other prediction on the price ticket. The model will change based on the variation on the data. So the firstly analyzing on the trend of data is important. This research in order to predict the average ticket price for the baseball match. The research first analyze the trend and the characteristic of the data. Then the research has decided the linear regression to predict the data. What's more, the research make up the missing data on 2017 by using the trend line from 2006 to 2016. Then the research use the model to predict the trend line and give a predicted result on 2022. In the procession, the research also use the standard score and find that the data for average ticket price are all normal for the changing. There are no abnormal data in this chart. To make this model more accurate, the research will consider more error in the future. Also, if the research want to predict more accurate result on 2023, it is better to wait for the precise result on 2022 and then refresh the model. Then the research can predict the result more exact.

REFERENCES

[1] Sylvester Devano(2021)Most Watched Sports in the world in 2021. https://sportsbrowser.net/most-watched-sports/

[2] Statista(2021) major league baseball average ticket price from 2006 to 2021 https://www.statista.com/statistics/193426/average-ticket-price-in-the-mlb-since-2006/

[3] Ji Xue, Brake, Y.Chen, Smirni(2018)Spatial-Temporal prediction models for active ticket managing in data centers https://ieeexplore.ieee.org/abstract/document/8260855

[4] Abdella, Zaki, Shuaib, Khan(2021)Airline ticket price and demand prediction: A survey https://www.sciencedirect.com/science/article/pii/S131915781830884X

[5] George A. F. Seber, Alan J. Lee Linear regression analysis (2012)A JOHN WILEY&SONS PUBLICATION. https://books.google.com.hk/books?hl=zh-CN&lr=&id=X2Y6OkXl8ysC&oi=fnd&pg=PR5&dq=linear+regression&ots=selUy3lOmy&sig=2DxQDd-Eae8LouXwxelVTm4On0M&redir_esc=y&hl=zh-CN&sourceid=cndr#v=onepage&q=linear%20regression&f=false

[6] Sanford Weisberg Applied Linear regression(2005)A JOHN WILEY&SONS PUBLICATION https://books.google.com.hk/books?hl=zh-CN&lr=&id=xd0tNdFOOjcC&oi=fnd&pg=PR7&dq=linear+regression&ots=dV4uzuFzLS&sig=UgRniS00943CKMR5_RApOl8wPUM&redir_esc=y&hl=zh-CN&sourceid=cndr#v=onepage&q=linear%20regression&f=false