Hodo Duale, Conor Desmond, Garrett Kronenberger

Prof. Beate Klingenberg

DSS Final Report

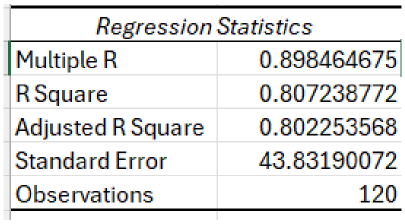
30 April 2024

Major League Baseball, MLB, consists of 30 teams across the United States and Canada, aspiring to become world champions. Navigating the landscape of the MLB requires more than scouting talent for the field. Behind the game of baseball lies a web of financial dynamics, intertwining revenue streams with team performance and fan engagement. From its core, baseball remains a game of statistics, consisting of analyzing players and teams playing statistics to form insightful predictions of your team and your opponents.

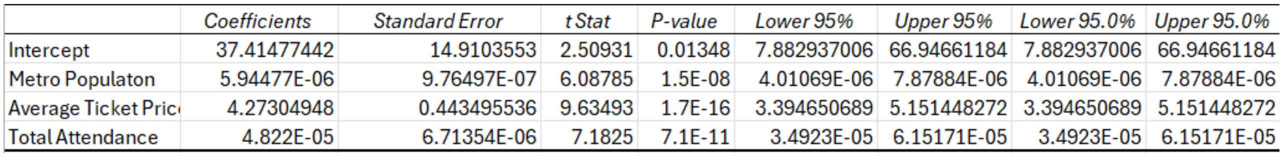
Our project focused on the business intelligence topic of regression analysis, which works to dissect and understand certain variables that can influence an MLB team's revenue. By analyzing data found from past years' valuations, including team performance metrics, market demographics, promotional activities, and other pertinent factors, we aim to uncover the underlying pattern between ticket prices, metro population, stadium attendance, and win percentage for predicting team's revenue in the following year. Moreover, we seek to uncover insights that can enable MLB teams to optimize their revenue strategies and enhance their overall financial viability.

To start our project, gathering data specifically on a team's average ticket price, the metro population of stadiums, state population (used province of Ontario for Toronto and Maryland for Washington, D.C.), total attendance, revenue, and win percentage from 2017 to 2023 create our foundation for using regression analysis. Compiling our data consisted of combing through Forbes MLB Valuation Lists, Statista's MLB teams' sales statistics, USA Facts population data, Macro Trends population data, MLB.com's standings, World Population Review population data, and Baseball Reference's team attendance information, transferring the data into our project's master Excel file, and grouping them by year. Forbes breaks down each team's financial portfolio into overall value, revenue, operating income, expenses, and cost ratio. For our purposes, we focused on the revenue and each season's total attendance for each team. World Population Review and Statista were used to get the metro population of each team. Statista and USA Facts were used to get state/province populations.

The first regression model used the Metro population, Average Ticket Price, and Total Attendance to fit Revenue. Overall, the multiple R and R squared in our output shows that around 81 to 90 percent of our points fall in the linear regression line. The win percentage variable is a negative determinant in predicting next year's revenue as it received -22.14 for the coefficient. The standard error of the model is 44, which is okay for predicting revenue. This means the average error the equation experienced when predicting revenue was 44 million. The standard errors of the Metro population, Average Ticket Price, and Total Attendance all are useful to the prediction of revenue, the most predictive being the average ticket price, which had a standard error of 0.44 and a t-statistic of 9.63. This is generally to be expected, as selling tickets at a higher price will ultimately lead to increased revenue. The t-stats for the variables and intercept are statistically significant (above 1.96) for all variables. The p-values for all variables and the intercept are low and statistically significant.

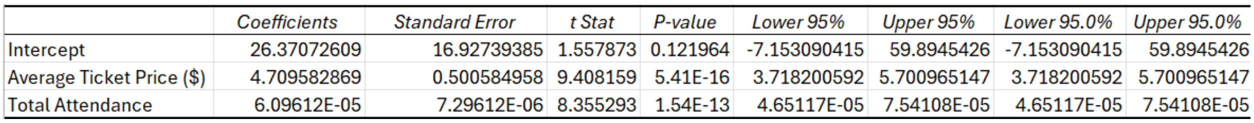


Regression Statistics	
Multiple R	0.898464675
R Square	0.807238772
Adjusted R Square	0.802253568
Standard Error	43.83190072
Observations	120

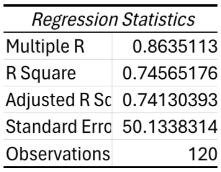


	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	37.41477442	14.9103553	2.50931	0.01348	7.882937006	66.94661184	7.882937006	66.94661184
Metro Populaton	5.94477E-06	9.76497E-07	6.08785	1.5E-08	4.01069E-06	7.87884E-06	4.01069E-06	7.87884E-06
Average Ticket Prici	4.27304948	0.443495536	9.63493	1.7E-16	3.394650689	5.151448272	3.394650689	5.151448272
Total Attendance	4.822E-05	6.71354E-06	7.1825	7.1E-11	3.4923E-05	6.15171E-05	3.4923E-05	6.15171E-05

In the second regression model, we analyzed only the ticket price and total attendance. Starting at the regression coefficients, the multiple R is 86% and the r squared is 75%, which still shows our dependent variable tested is highly likely explained by our independent variables. The standard error of the model is 50, so the average error was 50 million, which is okay for predicting revenue. Both the independent variables' coefficients analyzed are positive. The standard errors of both variables tested are relatively low and the t-statistics are statistically significant (above 1.96), concluding that our data containing the ticket price and the stadium attendance adheres to our regression line and is accurate in predicting our results. The p-values of average ticket price and total attendance are low indicating they are statistically significant.

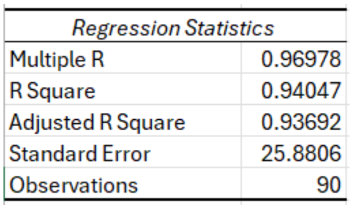


	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	26.37072609	16.92739385	1.557873	0.121964	-7.153090415	59.8945426	-7.153090415	59.8945426
Average Ticket Price ($)	4.709582869	0.500584958	9.408159	5.41E-16	3.718200592	5.700965147	3.718200592	5.700965147
Total Attendance	6.09612E-05	7.29612E-06	8.355293	1.54E-13	4.65117E-05	7.54108E-05	4.65117E-05	7.54108E-05

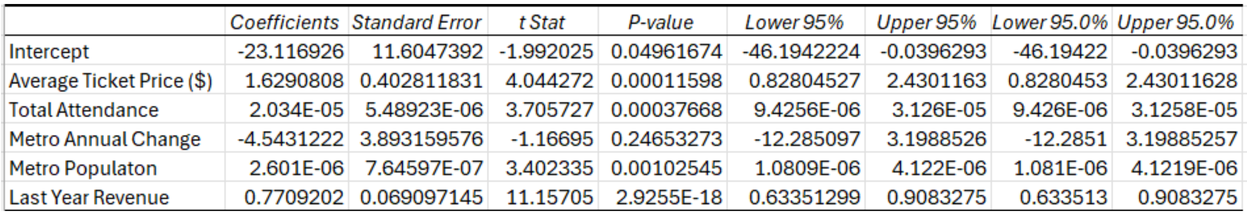


Regression Statistics	
Multiple R	0.8635113
R Square	0.74565176
Adjusted R Sc	0.74130393
Standard Errc	50.1338314
Observations	120

In our third regression model, we used only years 2018, 2019, and 2023 since those were the only years we had last year's revenue for. We used Avg. Ticket Price, Total Attendance, Metro Annual Change, Metro Population, and Last Year's Revenue to fit revenue. Starting at the regression coefficients, the multiple R is 97% and the r squared is 94%, which shows our dependent variable tested is highly explained by our independent variables and around 94 to 97 percent of our points fall in the linear regression line. The standard error of the regression is 26, so the average error was 26 million, which is pretty good for predicting revenue. The absolute values of the t-stats are statistically significant (above 1.96) for all the variables and the intercept except for Metro Annual Change. The p-values for the variables and intercept are low except for Metro Annual Change being .25, indicating all the variables and the intercept are statistically significant except for metro annual change. But removing Metro Annual Change hurt the adjusted r squared and standard error of the regression so it is left in.

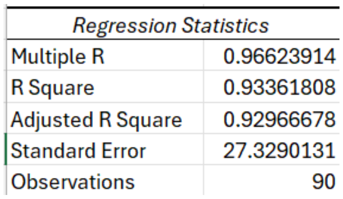


Regression Statistics	
Multiple R	0.96978
R Square	0.94047
Adjusted R Square	0.93692
Standard Error	25.8806
Observations	90

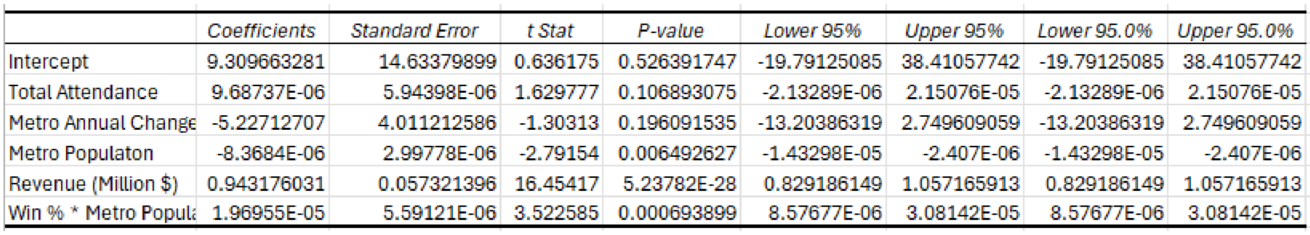


	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-23.116926	11.6047392	-1.992025	0.04961674	-46.1942224	-0.0396293	-46.19422	-0.0396293
Average Ticket Price ($)	1.6290808	0.402811831	4.044272	0.00011598	0.82804527	2.4301163	0.8280453	2.43011628
Total Attendance	2.034E-05	5.48923E-06	3.705727	0.00037668	9.4256E-06	3.126E-05	9.426E-06	3.1258E-05
Metro Annual Change	-4.5431222	3.893159576	-1.16695	0.24653273	-12.285097	3.1988526	-12.2851	3.19885257
Metro Populaton	2.601E-06	7.64597E-07	3.402335	0.00102545	1.0809E-06	4.122E-06	1.081E-06	4.1219E-06
Last Year Revenue	0.7709202	0.069097145	11.15705	2.9255E-18	0.63351299	0.9083275	0.633513	0.9083275

In our fourth regression model, we only used the years 2022, 2018, and 2017 since we did not have the next year's revenue for the other years. We used Total Attendance, Metro Annual Change, Metro Population, Revenue, and Win% * Metro Population to fit next year's revenue. The multiple R is 97% and the R squared is 93%, which shows our dependent variable tested is highly explained by our independent variables and around 93 to 97 percent of our points fall in the linear regression line. The standard error of the regression is 27, so the average error was 27 million, which is pretty good for predicting next year's revenue. The p-values for the variables are low except for Metro Annual Change and Total Attendance both being above .10, indicating most of the variables are statistically significant. The absolute value of the t-stats for the variables are statistically significant (above 1.96) for Metro Population, Revenue, and Win% * Metro Population, but the absolute values of the t-stats for total attendance and metro annual change are not statistically significant (below 1.96). Removing Metro Annual Change and Total Attendance hurt the adjusted r squared and standard error of the model so those are left in.



Regression Statistics	
Multiple R	0.96623914
R Square	0.93361808
Adjusted R Square	0.92966678
Standard Error	27.3290131
Observations	90



	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	9.309663281	14.63379899	0.636175	0.526391747	-19.79125085	38.41057742	-19.79125085	38.41057742
Total Attendance	9.68737E-06	5.94398E-06	1.629777	0.106893075	-2.13289E-06	2.15076E-05	-2.13289E-06	2.15076E-05
Metro Annual Change	-5.22712707	4.011212586	-1.30313	0.196091535	-13.20386319	2.749609059	-13.20386319	2.749609059
Metro Populaton	-8.3684E-06	2.99778E-06	-2.79154	0.006492627	-1.43298E-05	-2.407E-06	-1.43298E-05	-2.407E-06
Revenue (Million $)	0.943176031	0.057321396	16.45417	5.23782E-28	0.829186149	1.057165913	0.829186149	1.057165913
Win % * Metro Popul	1.96955E-05	5.59121E-06	3.522585	0.000693899	8.57677E-06	3.08142E-05	8.57677E-06	3.08142E-05

The first three regressions could be used to predict a team's revenue for the current year with the best being the third regression based on standard error and adjusted r squared. This could be useful to a team that has estimates of the values of the input variables before the season and wants to know its revenue to plan financially. It's also useful for near the end of the season when the inputs should be close to the value they will be at the end of the season. The last regression can be used to predict the revenue a team will generate next year, which can help them make financial decisions in the off-season. They can see the prediction for next year's revenue and decide if they want to make a big signing to try to generate more revenue, and thus sell more tickets, which would allow them to raise the price. More financial data could be added in the future to improve the performance of these regressions.

Works Cited

1. “Oakland, California Population 2024.” *World Population Review*, World Population Review, worldpopulationreview.com/us-cities/oakland-ca-population. Accessed 16 Apr. 2024.
2. “Population Estimates for Ontario, Canada 2023.” *Statista*, statista, 11 Mar. 2024, www.statista.com/statistics/569874/population-estimates-ontario-canada/.
3. Ragsdale, Cliff T. *Spreadsheet Modeling & Decision Analysis: A Practical Introduction to Management Science*. 6th ed., South-Western College Pub., 2007.
4. “Select a State.” *Rank List: States in Profile*, US Economic Development Administration , www.statsamerica.org/sip/rank_list.aspx?rank_label=pop1. Accessed 16 Apr. 2024.
5. “The Business of Baseball.” *Forbes*, Forbes Magazine, www.forbes.com/mlb-valuations/list/.
6. “The Long Term Perspective on Markets.” *Macrotrends*, Macrotrends, www.macrotrends.net/. Accessed 16 Apr. 2024.
7. “US Population over Time.” *USAFacts*, USAFacts, usafacts.org/data/topics/people-society/population-and-demographics/population-data/population/. Accessed 16 Apr. 2024.
8. “2023 Major League Baseball Attendance & Team Age.” *Baseball Reference*, www.baseball-reference.com/leagues/majors/2023-misc.shtml.
9. “2024 MLB Standings and Records: Regular Season.” *MLB.Com*, MLB, www.mlb.com/standings/mlb Accessed 16 Apr. 2024.