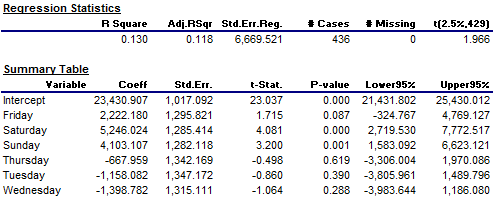
**Assignment #1: Orioles ’01-’19 Attendance Model/Regression and ’22 Forecast**

**Due by Sunday, November 7, 2021 at 11:59 pm ET**

**PART A1**: Using the Orioles 2001-2019 historical attendance, perform a regression/analysis of all 20 seasons listed of the following dependent variable:

1. ATTENDANCE

The historical data attendance file provides the following independent variables: Opponent, Month, Start Time, Day of Week, Weather, Opening Day, Holidays, Interleague (vs. National League teams), top opponent draws (NYY, BOS, WSH) and Promo Giveaways. You are encouraged to many/all of independent variables provided.

You may use any analysis tool of your choice (Excel, SPSS, SAS JMP, RStudio, Python, H20.ai etc.) to run the model/regression (multi-linear, logistic, etc.).

In your results you must show the following type of report to show overall stats along with a summary table for the independent variables.

Then, interpret/summarize your results – how strong is the overall correlation? What independent variables are stronger/more predictive than others?

**PART A2**: Develop a stronger/more predictive model: Omit or add independent variables from the original model/regression and test if it improves the correlation of the model. Omit earliest seasons in the data and similarly test if that improves the correlation. For example, run the previous 5-6 seasons. Look for outliers that might not be helpful for predicting 2022.

You are also welcome to add data/other independent variables to the data set such as teams that made the playoffs the previous season, winning %, holidays, etc. Similarly, describe your results – why do you believe that your new model is the strongest one possible? Note that any season which does not have 81 unique games means that at least one game was postponed by rain and made up as a doubleheader with another game. In 2015, six games are noted in red that were impacted by riots. These games are at the bottom of the data set and should be excluded from your analysis (April 27-May 3, 2015).

Tips on data set up if using Excel: consider grouping promotions into categories (e.g. A-B-C) to signify item appeal (High-Medium-Low), temp ranges (e.g. 50-54, 56-59, 60-64, etc.). Programs such as excel or SPSS may also require breaking out into separate columns with a 0/1 binary designation (e.g. “Interleague” column in the data set). As an example, some of the data has been set up in this format.

**PART B**: Using the results from your **most accurate projection model** in Part A, populate the forecasted attendance for each game on the 2022 game schedule provided. Game times, promotions and average high temps have been listed. Note: Ballpark capacity is 46,500 and no game should have a forecast above that sellout #. For the sake of the assignment, assume no impacts from the Coronavirus/lost 2020-capacity constrained 2021 seasons on game attendance. For this assignment, those two seasons have been excluded from the data set.

**SUBMIT THE FOLLOWING**:

\* Data Output file: Full data file with models/regressions/coefficients run. **Complete the 2022 attendance forecast page by game. There are 80 games, as one game is being played at Williamsport (Little League World Series)**

\* Word: Up to 2 pages to summarize your findings

\* Up to 5 pages of screenshots/exhibits – must use Tableau and must reference all Tableau charts/dashboards within the Word summary

Excel:

For tips on how to download the Analysis Tool Pack in Excel, visit <http://www.wikihow.com/Run-Regression-Analysis-in-Microsoft-Excel>

Regression Definitions:

R Square: Amount of variation in dependent variable that is explained by independent variables (Scale -1 to 1)

Ex: An R- Square of 0.5 shows that Total Paid Attendance (dependent variable) is 50% explained by the independent variables (day of game, opponent, etc.).

Coefficient: Amount that each independent variable (ex. Day of game, opponent, etc.) increases or decreases the dependent variable (attendance).

P-value: % chance that the answer you arrived at was by chance

Ex: A P-value of 0.85 for games vs. the Astros would show that the coefficient has an 85% chance of being different if the model was run again. The closer to 0.00, the stronger the variable. Anything below 0.05 is considered strong.

Lower 95%/ Upper 95%: Actual results are 95% likely to fall within this range. For example, if playing the Yankees has an upper 95% of 15,000, and a lower 95% of 5,000, then you can be 95% sure that playing any game against the Yankees will add between 5,000-15,000 fans.

Standard Error of Regression- Amount of variation in the prediction. To be 95% sure of your prediction, multiply the SE of Regression by 1.96, and add and subtract from your coefficient. The resulting range will give you 95% assurance that results will fall within your prediction.