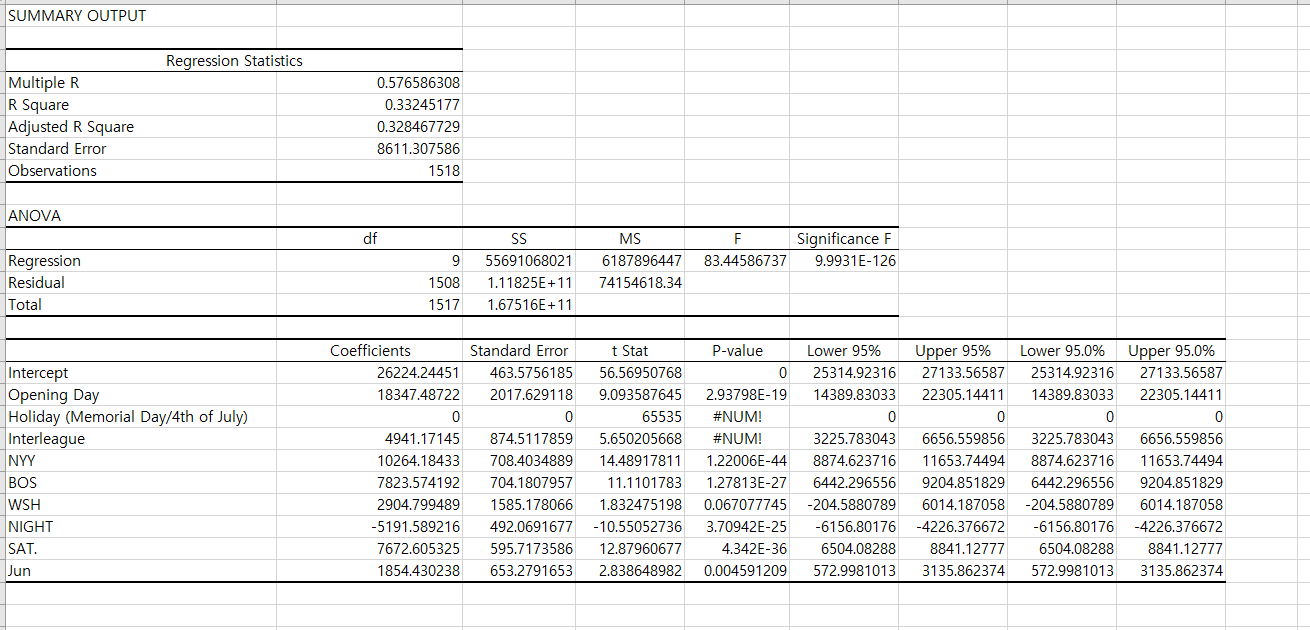
Assignment #1 : Orioles 01-19 Attendance Model/Regression and 22 Forecast

Due by Sunday, November 7, 2021 at 11:50pm ET

**Part A1 :**



It is the regression result from given data. I made the regression model with all variables with all rows rather than cleaning because I wanted to see first of how the current data is set up. From the result above, I was able to see that R-square is only 0.33 which is 33 percent of attendance is affected or explained by Opening day, holiday, interleague, opponent team, day, and month. On top of that, from the result, p-value didn’t come out well. Most of independent variables were large number or very small number that we could assume that as zero. Although most of independent variables weren’t worked well from this regression model, I could see a few independent variables had some value that I can think more of such as WSH, and Jun. These two independent variables had the reasonable value that I could approach for which were around 0.05.

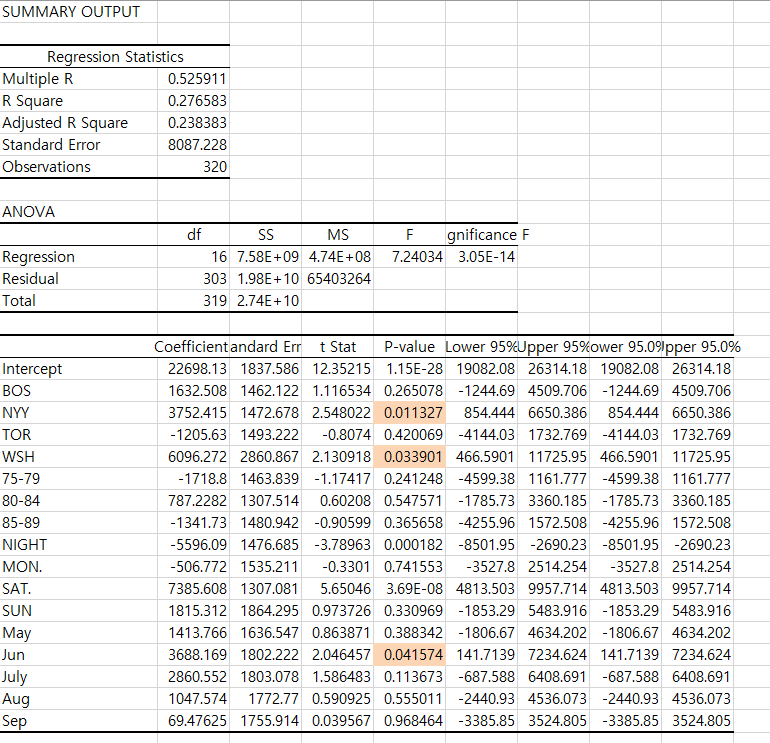
**Part A2:**

To maximize the result, I decided to clean some data. Before cleaned the data, I added all independent variables that were missing in given data. For example, list of all teams (opponents), temperature with grouping in every 5 degree (ex. 35-39,40-44,45-49, and up to 100-104), time (before 6pm, after 6pm), all day (Monday to Sunday), and season month (April to October).

On top of that, the given data was the time range between 2001 to 2019, so I decided to use only between 2016 to 2019. Although it has many independent variables, I decided to use only 16 independent variables that is major variables based on previous regression model.

To select 16 independent variables, I decided to use the opponent team with BOS, TOR, NYY, WSH. Although TOR is not as big name-valued team as BOS, NYY and WSH but the reason of choose TOR was TOR play total 37 games between 2016 to 2019 with Orioles which was the number that played one of the top numbers of playing against to Orioles. Also, I choose three different temperature group which was between 75-79, 80-84, and 85-89 because these were the temperature that most of the games were played. I also choose Night as another independent variables, and Monday, Saturday, and Sunday for independent variables. Lastly, I choose May, Jun, July, August, and September for the independent variables.

From the regression model of given data, I was able to check that WSH which was the independent variable of team opponent and JUN which was the independent variable of month had a valuable result at p-value compared with other independent variables. Because of the result of it, I decided to add one more team and add major months of MLB seasons. On top of that, we were not able to check the temperature from regression model of given data due to given data were contained temperature with word (description of weather).



From the regression model, I observed only 320 with 16 independent variables. I think this cleaned data is stronger model than given data because this data does not contain any old data, and it was cleaned with the purpose of getting the best result. For example, I was focused more on the opponent team and month. Also tried to add some new independent variables that couldn’t be the independent variables in given data.

R-Squared of the cleaned data is 0.276 which is about 28 percent. It is low percentage compared to given data which was 33 percent. However, p-value of independent variables are obtained lower value than given data which means that it is more strong independent variables.

**Part B:**

To predict 2021, I used Coefficient values from Regression Model. If 2021 event that has connection with the 15 independent variables at regression model then I count it as value of 1 (one) and multiply by the value at coefficient. After that, we sum up everything including intercept. The average was 21553. I rounded up everything into integer because people can not be into decimal. Interestingly, the one of highest attendance prediction was 29007 and 28981. These two events were against New York Yankees within Saturday or Sunday in June.