MGEC45 Homework Assignment #1

This problem set uses data from the NHL, NFL and MLB, and will ask you to analyze the way in which goals/points/runs are related to a team's winning percentage in a given season.

Question 1

Use the three spreadsheets with data from the 2015, 2016 and 2017 Major-League Baseball (MLB) seasons. Please refer to the "Readme file for MLB data" file for a description of the data in each spreadsheet. Once you have loaded your data, please complete the following questions:

(a) With the data from each season, run the following regression that estimates the Pythagorean Win Expectancy model:

$$WP = \alpha + \beta\left(\frac{(RS - RA)}{4R_{avg}}\right)$$

In this case, "R_{avg}" is the average number of runs scored by all teams in the season. So, in this case, you will estimate 3 separate regressions: one for the 2015 season, one for the 2016 season and one for the 2017 season. For each season's regression output, formally test the hypothesis that your estimate of β is equal to 2, as James's model would predict. (5 marks)

(b) Now, pool the three seasons of data together, and estimate a fixed effect regression model:

$$WP_{it} = \alpha + \beta\left(\frac{(RS_{it} - RA_{it})}{4R_{avg,t}}\right) + \sum_{i=1}^{(k-1)} \delta_i$$

In this case, the subscript "i" represents the team in the data, the subscript "t" represents the season in the data, and "k" represents the total number of teams in your data set. For this regression, formally test the hypothesis that your estimate of β is equal to 2, as James's model would predict. (10 marks)

(c) Compare the regression estimates of β in parts (a) and (b) – what does the similarity (or dissimilarity) of the estimates suggest about the potential problem of omitted variable bias within the model? (5 marks)

(d) Now, separate the right-hand-side variable into two parts, as follows:

$$WP_{it} = \alpha + \beta_1\left(\frac{RS_{it}}{4R_{avg,t}}\right) + \beta_2\left(\frac{-RA_{it}}{4R_{avg,t}}\right) + \sum_{i=1}^{(k-1)} \delta_i$$

In this case, be sure to specify your second variable as the ***negative*** value of runs allowed over four times the average number of runs scored in a given season. For this regression, formally test the hypothesis that your estimate of β_1 is equal to your estimate of β_2. Do your results suggest that runs scored and runs allowed have an equal impact on win percentage? (10 marks)

Question 2

Use the three spreadsheets with data from the 2016, 2017 and 2018 NFL seasons. Please refer to the "Readme file for NFL data" file for a description of the data in each spreadsheet. Once you have loaded your data, please complete the following questions:

(a) With the data from each season, run the following regression that estimates the Pythagorean Win Expectancy model:

$$WP = \alpha + \beta\left(\frac{(PF - PA)}{4P_{avg}}\right)$$

In this case, "P_{avg}" is the average number of points scored by all teams in the season. So, in this case, you will estimate 3 separate regressions: one for the 2016 season, one for the 2017 season and one for the 2018 season. For each season's regression output, formally test the hypothesis that your estimate of β is equal to 2, as James's model would predict. (5 marks)

(b) Now, pool the three seasons of data together, and estimate a fixed effect regression model:

$$WP_{it} = \alpha + \beta\left(\frac{(PF_{it} - PA_{it})}{4P_{avg,t}}\right) + \sum_{i=1}^{(k-1)} \delta_i$$

In this case, the subscript "i" represents the team in the data, the subscript "t" represents the season in the data, and "k" represents the total number of teams in your data set. For this regression, formally test the hypothesis that your estimate of β is equal to 2, as James's model would predict. (10 marks)

(c) Compare the regression estimates of β in parts (a) and (b) – what does the similarity (or dissimilarity) of the estimates suggest about the potential problem of omitted variable bias within the model? (10 marks)

(d) Now, separate the right-hand-side variable into two parts, as follows:

$$WP_{it} = \alpha + \beta_1\left(\frac{PF_{it}}{4P_{avg,t}}\right) + \beta_2\left(\frac{-PA_{it}}{4P_{avg,t}}\right) + \sum_{i=1}^{(k-1)} \delta_i$$

In this case, be sure to specify your second variable as the ***negative*** value of points allowed over four times the average number of points scored in a given season.

For this regression, formally test the hypothesis that your estimate of β_1 is equal to your estimate of β_2. Do your results suggest that points scored and points allowed have an equal impact on win percentage? Or does it appear that there is one factor that is more important than the other in determining a team's winning percentage. (10 marks)

Question 3

Use the three spreadsheets with data from the 2014-2015, 2015-2016 and 2016-2017 NHL seasons. Please refer to the "Readme file for NHL data" file for a description of the data in each spreadsheet. Once you have loaded your data, please complete the following questions:

(a) With the data from each season, run the following regression that estimates the Pythagorean Win Expectancy model:

$$WP = \alpha + \beta\left(\frac{(GF - GA)}{4G_{avg}}\right)$$

In this case, "G_{avg}" is the average number of goals scored by all teams in the season. So, in this case, you will estimate 3 separate regressions: one for the 2014-2015 season, one for the 2015-2016 season and one for the 2016-2017 season. For each season's regression output, formally test the hypothesis that your estimate of β is equal to 2, as James's model would predict. (5 marks)

(b) Now, pool the three seasons of data together, and estimate a fixed effect regression model:

$$WP_{it} = \alpha + \beta\left(\frac{(GF_{it} - GA_{it})}{4G_{avg,t}}\right) + + \sum_{i=1}^{(k-1)} \delta_i$$

In this case, the subscript "i" represents the team in the data, the subscript "t" represents the season in the data, and "k" represents the total number of teams in your data set. For this regression, formally test the hypothesis that your estimate of β is equal to 2, as James's model would predict. (10 marks)

(c) Compare the regression estimates of β in parts (a) and (b) – what does the similarity (or dissimilarity) of the estimates suggest about the potential problem of omitted variable bias within the model? (10 marks)

(d) Now, separate the right-hand-side variable into two parts, as follows:

$$WP_{it} = \alpha + \beta_1\left(\frac{GF_{it}}{4G_{avg,t}}\right) + \beta_2\left(\frac{-GA_{it}}{4G_{avg,t}}\right) + \sum_i \delta_i$$

In this case, be sure to specify your second variable as the ***negative*** value of goals allowed over four times the average number of goals scored in a given season.

For this regression, formally test the hypothesis that your estimate of β_1 is equal to your estimate of β_2. Do your results suggest that goals scored and goals allowed have an equal impact on win percentage? Or does it appear that there is one factor that is more important than the other in determining a team's winning percentage? (10 marks)