SPM 5708 Homework 6 (Multiple Linear Regression)

6.1. Which Stats Matter in Football?

We've spent time looking at stats that are more or less relevant to the success of baseball teams. But what about football? Which basic, *team-level* stats have the largest effect on point margins? In this set of questions, we'll take an initial look.

1. Download the *Gridiron* dataset from Canvas. Import the *NFL* worksheet and keep it named *NFL*. Note that it contains basic team-level stats for all NFL teams from 2010-2017.

2. It has been said that efficiency stats are a better measure of performance than count stats. Test this by seeing if *passing yards per attempt* (*PassYPA*) has a stronger linear relationship with *points for* (*PF*) than total passing yards (*TotPassYds*) does with points for (*PF*). Report and interpret your results below.

   Answer: PassYPA and PF have a correlation of .7577183 while TotPassYds and PF have a correlation of .6402673. Thus, PassYPA and PF have a stronger, linear relationship than TotPassYds and PF.

3. Now, do the same for *passing yards allowed per attempt* (*OPassYPA*) and *points against* (*PA*), and *total pass yards allowed* (*OTotPassYds*) and *points against* (*PA*). Report and interpret your findings.

   Answer: OPassYPA and PA have a correlation of .7116725 while OTotPassYds and PA have a correlation of .4632597. Thus, OPassYPA and PA have a stronger, linear relationship than OTotPassYds

4. Create a new variable called *Margin* by subtracting *PA* from *PF*. This point-margin variable (*Margin*) will now serve as our dependent, *y*-variable.

5. Regress *Margin* on *PassYPA, OPassYPA, RushYPA, ORushYPA,* turnovers committed (*TOC*), turnovers forced (*TOF*), offensive penalty yards (*OffPenYds*), and defensive penalty yards (*DefPenYds*) in a multiple regression model. Type the equation you obtain for the model **and** interpret the coefficients below.

   Margin^ = -33.92459 + 67.90231(PassYPA) - 59.32273 (OPassYPA) + 17.82161(RushYPA) - 21.28884 (ORushYPA) - 4.50917 (TOC) + 4.46022 (TOF) - 0.08656 (OffPenYds) + 0.07869 (DefPenYds)

   For every additional pass yard per attempt, we expect the margin of the score to increase by 67. 90231 points, holding all else constant.

   For every additional pass yard per attempt allowed, we expect the margin of the score to decrease by 59.32273 points, holding all else constant.

   For every additional rush yard per attempt, we expect the margin of the score to increase by 17.82161 points, holding all else constant.

   For every additional rush yard per attempt allowed, we expect the margin of the score to decrease by 21.28884 points, holding all else constant.

For every additional turnover committed, we expect the margin of the score to decrease by 4.50917 points, holding all else constant.

For every additional turnover forced, we expect the margin of the score to increase by 4.46022 points, holding all else constant.

For every additional offensive penalty yard, we expect the margin of the score to decrease by 0.08656 points, holding all else constant.

For every additional defensive penalty yard, we expect the margin of the score to increase by 0.07869 points, holding all else constant.

6. What do these coefficients say about the importance of the rush vs. the importance of the pass? The importance of offense vs. defense?

   Answer: These coefficients suggest that winning the Passing Yard game is so much more important than winning the Rushing Yard game. Each additional passing yard per attempt gained is worth around 67 points, holding all else constant, whereas each additional rushing yard gained per attempt is only worth around 18 points, holding all else constant. When comparing offense vs defense, it appears that offense is slightly more important than defense. If a team has the same number of TOF and TOC, the margin of points would be roughly a wash (0). The same can be seen with OffPenYds and DefPenYds. However, when comparing RushYPA vs ORushYPA, we see that defense is slightly more important. If these were equal, holding all else constant, the margin of points scored would be negative (-4). Lastly, when comparing PassYPA vs OPassYPA, we see that offense is definitely more important. If these were equal, holding all else constant, the margin of points scored would be positive (8). I would also like to point out that the intercept is -33. Essentially, this is saying that in a hypothetical situation, in which all predictors are equal to 0, the teams are dealing with a margin of -33 points. However, it is understood that this intercept is not realistic and should not be considered fully. Therefore, offense seems to be slightly more important, assuming that to goal is to win the game and finish with a positive margin.

7. Did any of the resulting coefficients surprise you? If so, which ones and why?

   Answer: Although it is an unmeaningful extrapolation, I was surprised that the coefficient on the intercept was -33. That seems like a large, negative margin to overcome despite all the predictors being 0. I was also surprised by the sheer magnitude of the PassYPA and OPassYPA coefficients, especially when compared to the other variables. Controlling the pass game on both ends of the ball seems to be vital in finishing the game with a positive margin.

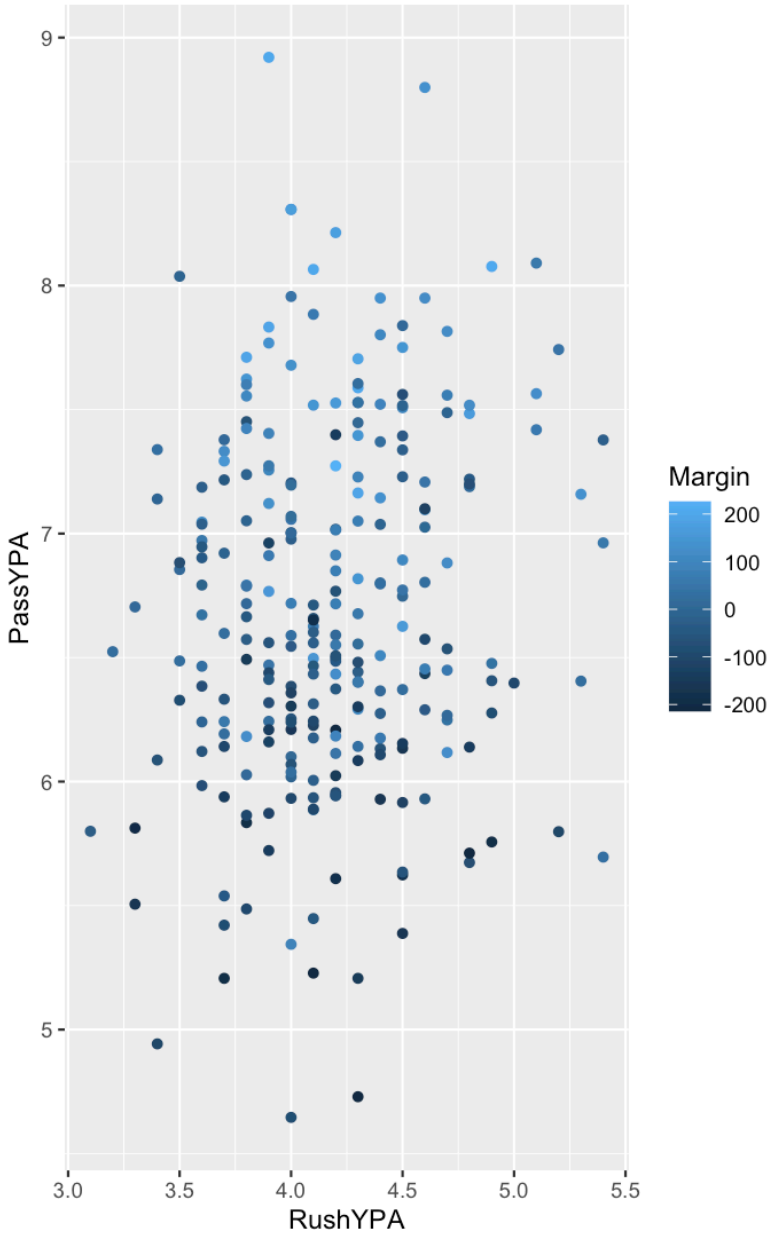
8. Look at the average salaries by position in the NFL. How do these seem to align with the results we just obtained?

   Answer: With the passing game being so important, it checks out that the highest paid positions per player seem to be QBs and WRs. These are both included in the offense, so it also makes sense that the highest paid sides of the ball is on offense as well.

9. It is often said that a good run offense is needed to set up a good pass offense. Test the strength between this relationship using the appropriate variables from the *Gridiron* data and

create a scatterplot to visualize this relationship. Does there appear to be any relationship between efficient passing teams and efficient rushing teams?

Answer: Due to the correlation between PassYPA and RushYPA being 0.1251568, there does not appear to be much of a relationship between efficient passing teams and efficient rushing teams. There is a weak, positive relationship, if any. From the scatter plot, we can also see that there is a much stronger relationship between higher PassYPA and winning than between RushYPA and winning.



10. Import the *NFL2018* sheet into R. Using this new (test) data, generate predictions for each 2018 NFL team's point margin. Then, use these predictions to calculate the mean absolute error (MAE) for the model you created in question No. 5.

Answer: The mean absolute error for Margin is 49.55999. On average, our Margin model was off by around 50 points.

11. Import the *NCAAF* sheet into R. Note that this sheet contains identical information for NCAA FBS teams between 2010 and 2017. Create an identical model to the one we used for the NFL data and type the resulting linear regression equation below.

    Margin^ = - 36.84270 + 49.04852 (PassYPA) - 46.61304 (OPassYPA) + 46.82719 (RushYPA) - 58.48172 (ORushYPA) - 2.94967 (TOC) + 3.74849 (TOF) - 0.03469 (OffPenYds) + 0.12688 (DefPenYds)

12. How do the effects (coefficients) of these statistics vary in the NCAA compared to the NFL?

Answer: In the NCAAF model, it appears that controlling the rush game is just as, if not more, important than controlling the passing game. It also appears that defense is just as, if not more, valuable than offense in having a preferable margin.

6.2. Estimating Attendance in Major League Baseball

Not every sport manager is a general manager. Many of you plan to work in other areas of the industry, such as marketing, operations, or event management. Thankfully, regression analysis can readily be applied to problems in these settings, too. For this set of questions, you will model the relationship between MLB attendance levels and a variety of quantitative factors believed to affect them.

1. Download the *MLB_Attendance* Excel file from Canvas and import the first worksheet labeled *MLB_Attendance*. This dataset contains game and attendance information from every regular-season game played during the 2014-2018 MLB seasons.

2. The *y*-variable of interest in this data is *PctCap*, a measure representing the percentage of the venue (stadium) capacity that was filled during a game. Why might we want to look at *PctCap* as the *y*-variable instead of the raw attendance figures (totals) recorded in the *Att* column?

   Answer: We might want to look at the percent of capacity instead of the raw attendance total because these values are standardized based off the number of people each stadium can hold. If we were only looking at raw attendance totals, teams with larger/smaller stadiums would be accounted differently and so the variables would not be weighted equally between each stadium. This way, we can look at what factors that truly drive demand / attendance up.

3. Regress *PctCap* on the following *x*-variables recorded in the data: *GB_Pre*, *RS_PG*, *RA_PG*, *WPct_Pre*, *Pre_Streak*, *Temp*, *Wind*, *Start_Time*, *Opp_RS_PG*, *Opp_RA_PG*, *OppWPct_Pre*, and *Opp_Pre_Streak*. Run the *summary* function to view the results of the analysis (see the *Appendix* on the last page for *x*-variable descriptions).

4. Does every *x*-variable have a significant effect on our measure of attendance (*PctCap*)? Which variables are not significant ($p > .05$)?

   Answer: No, not every variable in the model is significant. The following variables are NOT significant at the alpha = .05 significance level :

WPct_Pre (p-value = .6492)

Pre_Streak (p-value = .5274)

OppWPct_Pre (p-value = .0599)

Opp_Pre_Streak (p-value = .9201)

5. Take the *x*-variable with the *highest p*-value and remove it from the regression equation. Then, re-run the model. Repeat this process (removing the non-significant *x*-variable with the highest *p*-value) until all remaining x-variables are significant ($p < .05$). Which x-variables were dropped from the model?

   Answer: First, I dropped Opp_Pre_Streak and reran the model. The following variables were still not significant at the 5% significance level:

   WPct_Pre (p-value = .6488)

   Pre_Streak (p-value = .4384)

   OppWPct_Pre (p-value = .0572)

   Next, I dropped WPct_Pre and reran the model. The following variables were still not significant at the 5% significant level:

   Pre_Streak (p-value = ..4675)

   OppWPct_Pre (p-value = .0603)

   Then, I dropped Pre-Streak and reran the model. The following variables were still not significant at the 5%significance level:

   OppWPct_Pre (p-value = .0537)

   Lastly, I dropped OppWPct_Pre and reran the model. All of the variables remaining are significant at the 5% significance level.

6. Type out the equation for your final model and interpret the coefficient (i.e., the effect of *x* on *y*) from at least two of the *x*-variables below.

   Answer: PctCap^ = .7108 + .004249(GB_Pre) + .03199(RS_PG) - .05653(RA_PG) + .002434(Temp) + .006509(Wind) - .0001712(Start_Time) + .01435(Opp_RS_PG) - .02713(Opp_RA_PG).

   .7108 – If all of the independent variables were 0, we would expect the stadium to be at 71.08% capacity.

   .004249(GB_Pre) – Holding all else constant, we would expect the stadium's capacity to increase by about .4% for every additional game back that the home team was from leading the division.

   .03199(RS_PG) – Holding all else constant, we would expect the stadium's capacity to increase by about 3.2% for every additional run scored by the home team per game.

-.05653(RA_PG) – Holding all else constant, we would expect the stadium's capacity to decrease by about 5.6% for every additional run allowed by the home team per game.

.002434(Temp) – Holding all else constant, we would expect the stadium's capacity to increase by about .2% for every additional ºF that the temperature was at first pitch.

.006509(Wind) – Holding all else constant, we would expect the stadium's capacity to increase by about .7% for every additional mph that the average wind speed was recorded at.

-.0001712(Start_Time) – Holding all else constant, we would expect the stadium's capacity to decrease by about .02% for every additional minute after 12:00 pm that the game started.

.01435(Opp_RS_PG) – Holding all else constant, we would expect the stadium's capacity to increase by about 1.4% for every additional run scored by the away team per game.

-.02713(Opp_RA_PG) – Holding all else constant, we would expect the stadium's capacity to decrease by about 2.7% for every additional run allowed by the away team per game.

7. Report the model's R^2. Are you surprised at this value? Why or why not?

   Answer: The model's R^2 value is .1679. This means that roughly 16.79% of the change in percentage of a stadium's capacity can be explained by our independent variables. I am slightly surprised at how low this value is, considering that all of our remaining variables are significant. However, I understand that there are a lot more meaningful factors that go into predicting attendance percentage.

8. Now, import the *Att_2021* worksheet from the *MLB_Attendance* file. As the name implies, this contains game and attendance data from the 2021 MLB season. Use it to calculate your model's *mean absolute error (MAE)*. Report *and interpret* this value below.

   Answer: 0.2847573

   On average, we were off by about 28%.

9. Which variables could potentially be added to improve this model? Do you think certain coefficients/effects might change if these variables were added?

   Answer: I think we could benefit by adding how many games away from the 81st game this is. I do not have any proof, but I would imagine that stadiums are closer to capacity in the beginning/end of the season, especially if they are a playoff team. Speaking of which, we could also include a variable to see if the team is in playoff contention, although this may provide multicollinearity with WPct. We could also include a binary, indicator variable to see if there is a special promotion scheduled or not. We could also include another binary, indicator variable to see if this is part of a double-header or not. Despite them being rare, I would imagine that attendance is low in most double-headers. If we did add all of these variables, I would imagine that the coefficients on some of the variables correlated with winning would decrease.

SUBMIT YOUR COMPLETED WORD DOC AND ASSOCIATED R SCRIPT TO THE CANVAS DROP BOX.

Appendix: 6.2. X-Variable Descriptions

GB_Pre: the number of games back the home team was from the division leader entering the game.

RS_PG: the average number of runs scored per game by the home team entering the game.

RA_PG: the average number of runs allowed per game by the home team entering the game.

WPct_Pre: the home team's winning percentage entering the game.

Pre_Streak: the winning (positive value) or losing (negative value) streak the home team was on entering the game.

Temp: the recorded temperature (in degrees Fahrenheit) at the time the first pitch was thrown in the observed game.

Wind: the average wind speed (in miles per hour) recorded at the venue during the game.

Start_Time: the number of minutes before (negative) or after (positive) noon that the game started.

Opp_RS_PG: the average number of runs scored per game by the visiting team entering the game.

Opp_RA_PG: the average number of runs allowed per game by the visiting team entering the game.

OppWPct_Pre: the visiting team's winning percentage entering the game.

Opp_Pre_Streak: the winning (positive value) or losing (negative value) streak the visiting team was on entering the game.