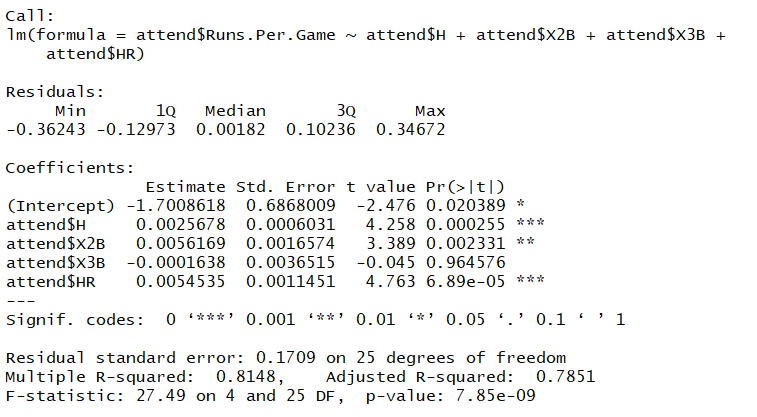
**Step 1.)** Grab some data and estimate a simple multivariate regression like you did in the prior discussion board. You can use [DASL (Links to an external site.)](https://dasl.datadescription.com/datafiles/) again if you like. Tell us your model (i.e. happiness on the LHS and age and income on the RHS)

For the purpose of this discussion I am re-examining my previous multireg model. I was attempting to determine the number of runs a team would score per game based on the number of singles, doubles, triples, or home runs they would hit.



The intercept of -1.7 doesn’t really make sense since if you didn’t get a hit you wouldn’t score any runs

The coefficients say that for every single you can expect .002 more runs per game, every double is .005, every triple is -.0001 and every home run is .005 more runs per game.

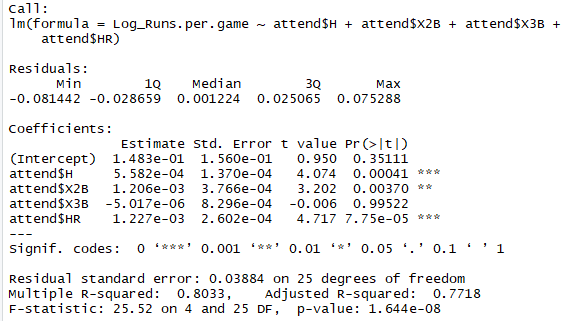
**Step 2.)** Identify 2 ways you can change the regression model to incorporate new techniques learned in Module 5 and explain why the changes from Module 5 might be better (or not). (i.e. income might have a non-linear effect, and marital status matters a lot, and age and marital status should have an interaction since being single at 20 is really different than being single at 45)

One way to improve the model is to take the log of runs scored / game and perform a log level analysis. This would be to take into account that for doubles, triples, and home runs these could have more affect on the bases gained than a single would. We don’t want to understate the affect of any one type of hit

The second way would be to use quadratic terms. This would give us a better idea of what type of hit would be best to go for. Should we just hit for singles or is having a lot more power hitters on your team better?

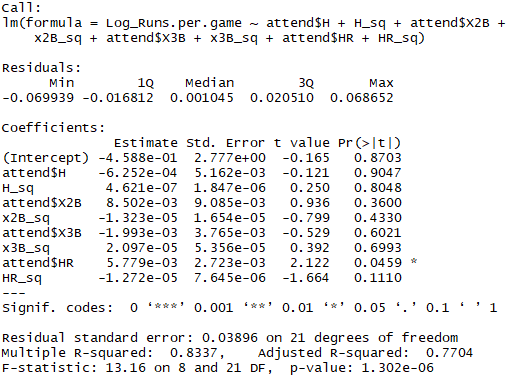
**Step 3.)** Use the two (or more) new methods from module 5 to estimate a new regression model. Show (probably with a screenshot) how it differs from the model you estimated in step 2. Explain why you think the models are different, and which you prefer and why.

Option 1 – Log-Level



The betas are quite different and show a much higher value for a single. I believe this would be due to the fact that more singles are hit, but it shows they are more impactful with a 500% increase in runs / game for each additional hit.

Option 2 – Quadratic



The coefficients differ greatly but by looking at the quadratic terms we can see that H is positive and x2b and HR are negative. Since they are negative that means there is diminishing returns for these types of hits. Yes, HRs will score runs but trying to hit HRs all the time is not guaranteed and more often than not you will get out.

**Step 4.)** Comment on how you think your results are different, or the same, as what you would get from randomized assignment of one RHS variable of interest. In other words, relate your output and model to what we learned in Module 6 and how that informs your understanding of your model.

Since I looked at total game data for my model there is already randomization built in. All of the games account for different pitchers on different days, changes in days of rest and travel, as well as weather conditions. Since all of those variables would be difficult to make truly random in a lab environment based on the size of the sample there is already randomization built in.