NBA Progress Tracker

* Having trouble finding the data that I initially want-
* NBA athletes with their sponsorship amount or top paid NBA athletes each year

Settling for Steph Curry Data and will cross reference with Under Armor stock data to create framework

Once I have dates and lists of regression parameters, I can use better selected data in pipeline.

8/27/2016 3:04 PM – Cleaned data into new file- had to remove extra lines, and games where steph curry did not play, change W/L with spread to W=1 and L= 0, changed spread to be positive or negative depending on W or loss (previously was just the number) – used **Excel** to clean data

8/27/2016 3:04 PM- Thinking about re stating the question

I can use the returns to make a regression of returns vs curry game score

Or a regression of an index + curry game score

Or a log odds regression of up or down stock price with win loss, and gamer score (and maybe index)

Game Date | Gamer Score| W.L| Return Date | Return | +/- Return

Starting to arrange data, change date format, prepare the final data set that has the correct stock returns for the correct games.

8/27/2016 3:04 PM

So I had a regression with the returns on the S&P 500 and Game Score. The R score is 7.5% with the adjusted R squared down to 1.5%. The t scores are pretty high also- the S&P 500 p value is 0.124 and the Game Score is higher at 0.758. For reference, we like these numbers to be below 10% for significance.

So I will try using another benchmark. I googled ETFs that Under Armour is part of and found that UA is only 0.04% of the S&P. I will try the XLF- financial select sector- where UA is 0.33% of the portfolio. Hopefully this portfolio will provide a better baseline for underarmour and then the Game Score will add more value to the regression.

* Import XLF data, find similar dates and make sure it’s in the same format as SP data, run regression again UA and then multiple regression with Game Score on Returns

Tried scaling the data and the regression looks better.

Committed a sin of data science and just removed data points without looking into causes. The first one was the first point which could be some shady math but there is another point, 22, which I am not sure why it was extreme. Removed it and the plot of the regression looks much better.

8/27/2016 3:04 PM

Started a new file to condense some of the goals from the last file.

Used the function for data producing log returns. Had some trouble using the %in% command. I had to switch the order to make it work nice.

Made another clean data set with new cleaning methods.

Had to add a package to make the MP work nicer- turned minutes played into seconds played

Got rid of age column ( its from his birthday, don’t think it will be that important)

When he doesn’t shoot free throws, they leave his FT% out (0/0) – I might impute the mean into those spaces

Also started doing some PCA analysis, Found an interesting link with the IRIS data set. Next is the PCA Regression and maybe Partial least squares after.

Partial least squares is like PCA but minimizes the variance in the predictors and the response variables which can help reduce the dimensionality.

PCA stuff is pretty much from ISLR- applied on pages 256- need to review the theory though