Rady | UC San Diego School of Management

The Case of the Hot Hand[†]

Anyone who plays or watches basketball has heard of the *hot hand*. You might have heard sportscasters say something like "Steph's heating up!", "he's on fire!", or even "he's en fuego". You may have seen headlines such as "Warriors' sizzling backcourt tandem sinks 72 points". Why might that backcourt be sizzling? Well probably because Steph Curry is on fire because he lives with a permanently hot hand. If a team has a player on the court who suddenly gets a hot hand, typically the teammates keep feeding that player the basketball in order to score a lot of points. Watch ESPN Sportscenter and you can occasionally see highlights of an NBA player with a hot hand. On the other hand, if you would like to see what a cold hand looks like, go to RIMAC and watch your professor play basketball. But, that's not as exciting!

In this case, we are going to think about whether or not this notion of a hot hand really exists. Some good questions might include: what data would you need to collect to demonstrate a hot hand exists? How would you collect that data? Would you run an experiment? What information would be relevant? How much data would you need? In fact, there are many questions you would have to dig into, and researchers have been doing so for decades. To this day, the debate on the existence of the hot hand lives on.

Regression Analysis

In order to get at that question, let's first try to simply get familiar with basketball data and with examining the relationships between variables found in such datasets. Load the datafile called `nba_pgdata.parquet` and read carefully through the data description. This dataset has NBA player salary and performance data for the 2012/2013 NBA season. Your job is to come up with a linear model relating variables in the dataset to player salary. In other words, you should develop a model that you feel provides insights into what predictors explain a player's salary level. Further, you may have a goal to use this model to make predictions. Below we provide some highly recommended guidance for your exploration.

- **Visualization**: Make histograms of `Salary`, `Age`, `FG`, `RB`, `AST`, `STL`, `BLK`, and `PTS`.

What do the distributions look like? Take note of any skewness. Make scatterplots of Salary versus each of the predictors and examine the relationships.

- **Regression A**: Run an initial regression of Salary on the predictors: Age, FG, RB, AST, STL, and BLK. Leave out PTS for now.
  - Interpret the estimated coefficients:
    - Which predictors are significant and which ones are not?
    - Do these regression results make sense?
    - What hypothesis test is underlying the output seen for each predictor?
    - What do the estimated values for the coefficients mean?
    - Which predictor seems the most impactful?
  - How does the R^2 value look? How do you interpret it?
  - Take a look at the *Dashboard* regression plots. Think carefully about what is being plotted in each of these plots. Are there any problems standing out to you? In particular, examine closely the *Residuals vs Fitted* plot.

- **Regression B**: Transform Salary using the log function to create a new variable called log_Salary is created. Visualize a histogram of log_Salary. Compare it to the histogram created previously for Salary. Then, run the same regression but using log_Salary as the response variable. Repeat the steps above. What is the correct interpretation of the coefficients in this case?

- **Regression C**: Keeping log_Salary in the model, add the predictor PTS to the model and run another regression. What happens to the regression results? Try to think about why this might occur and explain.

- **Regression D**: Again, keeping log_Salary in the model, drop FG and use PTS instead. Convert to *Standardized coefficients*:

  ```
  from sklearn.preprocessing import StandardScaler

  df = data.copy()
  df[['log_Salary', 'Age', 'RB', 'AST', 'STL', 'BLK', 'PTS']] =
       StandardScaler().fit_transform(df[['log_Salary', 'Age', 'RB',
        'AST', 'STL', 'BLK', 'PTS']])
  ```

How does using standardized coefficients change the interpretation of the regression results?

Hot Hand Hypothesis Testing

Conditional Probability Approach

A few decades ago, there was some interesting research conducted by some very well-known psychologists that tested this hot hand theory. They collected data from an NBA team for each of the 9 primary players. They then analyzed the data to determine what was the probability a player would “hit” (i.e., make) the next shot conditional on having missed the last shot. Similarly, they computed the probability a player would hit the next shot conditional on having made the last shot. The probabilities taken from their study are provided below in Table 1.

Player	$\mathbf{P\{hit \mid 1\ miss\}}$	$\mathbf{P\{hit\}}$	$\mathbf{P\{hit \mid 1\ hit\}}$
Player A	0.56	0.50	0.49
Player B	0.51	0.52	0.53
Player C	0.46	0.46	0.46
Player D	0.60	0.56	0.55
Player E	0.47	0.47	0.45
Player F	0.51	0.46	0.43
Player G	0.58	0.54	0.53
Player H	0.52	0.52	0.51
Player I	0.71	0.62	0.57

Table 1: Conditional shooting probabilities for NBA players

Does the data suggest that a hot hand exists? Here are some questions/comments that you may want to consider in answering this question.

- From a bird’s eye point of view:
    - What is your impression on what the data is saying?
    - Just from examination, which players seem to exhibit the hot hand and which ones do not?
    - What is your logical reasoning?

- For a given player, think about values you might compare perhaps using difference as a measure of interest. More formally, develop a statistical test on this measure across the 9 players (Hint: Compare the average of this measure across players to something else):
  - What is the null hypothesis ("the something else")? What are you hoping to show (alternative hypothesis)?
  - How much data do you have here?
  - What distribution is appropriate?
  - What conclusions can you draw from your test?

Streaks Approach
In the above analysis, you compared conditional probabilities. Another way to think about the hot hand is examining runs of made shots in a row. Suppose that hits are denoted by `H` and misses by `M`. Then a run could be represented by `HHMHMMHHH`. In this example, there is one run of 2 hits, one run of 1 hit, and one run of 3 hits. There is also one run of 1 miss and one run of 2 misses. So, overall there are 5 runs in that sequence. Now one way to think about *streakiness* is that perhaps there should be fewer (thus longer) runs. For example, perhaps a run such as `HHMMMHHHH` might be more reflective of a person with a sometimes hot hand. Here, this sequence has only 3 runs.

In class, we talked a lot about how theory (models) and data come together. This is another great example because if you told me that a player has a probability p of making any given shot, then we could compute the exact probability of a run occurring just as in either of the sequences above. In fact, we could formulate a model where `H` occurs with probability p and `M` occurs with probability $1-p$, and we could let a random variable R represent the number of runs in a sequence of n shots. For example, if $n = 3$ (i.e., 3 shots), and $p = .4$ (i.e., 40% chance of making any shot), then what would be the probability of having only 1 run? Well, there are only two ways that can happen: either you get `HHH` or `MMM`. The probability of the first one occurring is 0.4^3 and the probability of the second sequence occuring is 0.6^3 for a total probability of $0.064 + 0.216 = 0.280$. Using probability theory, we could *eventually* compute the probability associated with two runs and three runs as well. Furthermore, we could generalize this to any n, rather than just $n = 3$. We could also compute measures such as $E[R]$ and $\sigma_R = \sqrt{Var(R)}$. This would tell us a little bit about the distribution over the different run lengths, i.e., what is the average number of run lengths and how spread out is the distribution over the possibilities.

If we did all of that work, then we could put it to good use because Table 1 tells us $P\{hit\}$ which is the probability a player makes any given shot and, below in the second column of Table 2, we have the number of shots that player took. So using such a model, we could compute the expected number of runs for this player and its standard deviation which is shown in the fourth and fifth columns below.

	⟵DATA⟶		**⟵MODEL⟶**	
Player	**Number of Shots**	**Actual Number of Runs**	**Expected Number of Runs**	**Standard Deviation of Runs**
Player A	248	128	125	7.9
Player B	884	431	?	?
Player C	419	203	?	?
Player D	339	172	?	?
Player E	272	134	?	?
Player F	451	245	?	?
Player G	433	227	?	?
Player H	351	176	?	?
Player I	403	220	?	?

Table 2: Shot runs for NBA players and a comparison model

Develop a simulation model for streaks in order to fill in the rest of the values in the fourth and fifth columns of Table 2.

Using the results from your simulation model, make a hypothesis test for each player comparing what gets actually observed in their shooting (column 3) to what theory tells us. If what we observe is drastically different (or perhaps better stated: *quite unlikely*), then maybe we have collected evidence that the hot hand exists. What do you think? Is the runs data suggestive that any player has a hot hand?