NBA Action, it’s FANtastic
(for data analysis)

Ryan Elmore
17 April 2023
Ryan.Elmore@du.edu
A Problem to Avoid

*The Far Side*®

2007

AUGUST 14

Tuesday

Math/Stats/CS Professor during seminar

Audience during seminar
Topics

1. Timeouts in the NBA
2. NBA Shot Analysis
3. Concluding remarks
   1. Summary of findings
   2. Teaching
Collaborators
Problem Background

• Prior to the 2017-8 NBA season, the NBA reduced the number of timeouts from 18 to 14
• Understanding the value of each timeout in various in-game scenarios became more important.
• Question from a friend in Steamboat Springs, CO: Do timeouts (TOs) impact the game?
Game Five of the 2019 NBA Finals

WHAT was Nick Nurse THINKING with this timeout? | 2019 NBA Finals | Game 5
SlyGlobal • 14K views • 3 months ago
Nick Nurse took a timeout with 3:05 left in the 4th quarter, up by 6, killing his team's momentum. The Warriors went on to score 9...

Stephen A.: Nick Nurse let the Warriors off the hook in Game 5 | SC with SVP
ESPN © 200K views • 3 months ago
Stephen A. Smith reacts to the Golden State Warriors' victory over the Toronto Raptors in Game 5 of the 2019 NBA Finals and...

NBA Finals: Nick Nurse defends calling time out near end of Game 5
Global News © 1.6K views • 3 months ago
Courtesy: NBA TV | Toronto Raptors coach Nick Nurse on Wednesday, June 12, 2019 defended calling a time out near the end of...

Criticizing Nick Nurse's late timeout is 'foolish, stupid and asinine' - Charles Barkley | Get Up
NBA on ESPN © 206K views • 3 months ago
Charles Barkley defends Nick Nurse's decision to give Kawhi Leonard a breather and call a timeout with 3:05 remaining in Game...
To Call a Timeout or Not

Camp One: Call a TO
Allow your team to rest, regroup, diagram a play

Camp Two: Let them Play
Your team will figure things out on the fly

You don’t ever really run out of timeouts; they just have a price.

There are no analytics that really proves that if you call a timeout, it stops the run of the other team.
Research Hypothesis and Data

• If an NBA team is on a run, a timeout (by either team) will stymie that run.
  • What’s a run?
  • What’s an appropriate outcome variable to measure?
  • How do we test something like this?

• Data
  • 2017-18 and 2018-19 regular season games
  • 1.4+ million events; 778,828 “plays”, 26,052 plays w/ a TO, and 1,149 with a run
Definitions
Definitions (cont.)
Outcomes

- Change in score difference? No
- Max score difference? No
- Change in win probability? No
- Integrated centered-score difference? Yes!

\[ y_i = -\text{sign}(s(t_i)) \int_{t_i}^{t_i+1} [\Delta(x) - \Delta(t_i)] \, dx \]
Hypothetical Outcomes
How should we measure the effect?

- **Naïve Estimate**: Use a simple t-test to assess the difference in means between two groups
- **Randomized Experiment**: One of the key benefits of randomized experiments for estimating causal effects is that the treated and control groups are guaranteed to be only **randomly** different from one another on all background covariates, both observed and unobserved.
  
  - Elizabeth Stuart (2010), Statistical Science
How will we measure the effect?

• **Causal Inference:**
  - Use Rubin’s Causal Model
  - Try to replicate a randomized experiment
  - Create matched observations (controls) with similar covariate distributions as the treatments
  - We use the matching software in Sekhon (2011)
    - 2023 Update: I’d probably recommend the MatchIt software now (Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011))
Potential Outcomes Model

- Potential Outcomes:

\[ Y_i = \begin{cases} 
Y_i(0), & T_i = 0 \\
Y_i(1), & T_i = 1 
\end{cases} \]

- Average Causal Treatment Effect on the Treated:

\[
\text{ATT} = \mathbb{E}_{x | T=1} [\mathbb{E} [Y_i(1) - Y_i(0) | T_i = 1, X_i = x]]
\]
GAM Covariates

- Team that is not on the run
- Opposing team
- Run point total
- Run duration
- Time left
- Win probability
- Signed score difference at beginning/end of run
- Possession indicator
- Home indicator
- Week in season
- Spread
- Over/Under
- Money line
Matching

Treated Units

Pool of Potential Controls
Matched Example

NOR @ DEN: March 02, 2019

WAS @ MIL: February 06, 2019
Covariate Balance (Propensity Scores)

Before Match

After Match

a) No Timeout  
   Timeout

b)
Covariate Balance (Standardized Bias)
Franchise Effects

- 2018-19 Season
- The average treatment effect on the treated for a specific franchise is:

\[
ATT_f = \mathbb{E} [Y_i(1) - Y_i(0) \mid T_i = 1, B_i = 1]
\]
To Call a Timeout or Not

Camp One: Call a TO
Allow your team to rest, regroup, diagram a play

You don’t ever really run out of timeouts; they just have a price.

Camp Two: Let them Play
Your team will figure things out on the fly

There are no analytics that really proves that if you call a timeout, it stops the run of the other team.
NBA Shots

Research Goals

• Model-based classification of NBA teams and players based on their shot taking propensities and their ability to make shots
• Post-hoc evaluation of specific players using novel metrics
NBA Shots

We assume each team belongs to a latent shot selection cluster, $w_i$, and a latent accuracy cluster, $z_i$, and that these cluster memberships are independent. Let $L$ and $J$ denote the number of latent shot selection and shot accuracy clusters, respectively, such that $w_i \in \{1, \ldots, L\}$ and $z_i \in \{1, \ldots, J\}$ for the $i^{th}$ team. Note that both $L$ and $J$ are specified a priori.

For generic shot selection cluster $w$, define $p^k_w$ to be the probability of a shot being taken from region $k$ where $\sum_{k=1}^K p^k_w = 1$. Similarly, for generic shot accuracy cluster $z$, let $q^k_z \in [0,1]$ denote the accuracy probability for region $k$. Note that the sum of the individual shot make probabilities for each region have no sum restriction. Given the $i^{th}$ team’s shot selection cluster membership, $w_i$, we model the vector of shot counts across the $K$ regions using a Multinomial distribution where

$$(N^1_i, \ldots, N^K_i)|w_i \sim \text{Multinomial} \left( N_i, (p^1_{w_i}, \ldots, p^K_{w_i}) \right).$$

Next, conditional on the team’s shot accuracy cluster membership, $z_i$, the number of shots made in each region $k = 1, \ldots, K$ follows a Binomial distribution where

$$M^k_i|z_i \sim \text{Binomial}(N^k_i, q^k_{z_i}).$$

- We adopt a Bayesian hierarchical framework
- Use MCMC methods (Gibbs sampler) to sample from the posteriors
- Inference on $p$’s, $q$’s, cluster membership, etc.
A Note on Sports Analytics and Education

• It’s an ideal medium for teaching analytical (often deep) concepts

• If you’re interested, look for *An Introduction to Sports Analytics using R* (by Elmore and Urbaczewski, 2023)