Every Second Counts: Using Big Data to Investigate Referee Accuracy in NBA Games

Virginia York, Leah Kozel, Gavin McDermott, and Matthew Hettleman

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

INTRODUCTION

- All sports fans are deeply interested and committed to referee accuracy
- NBA is the gold standard
- The hope is that in a study, we would not find bias because the system is sophisticated and professional
- Would we find bias in similar data for MLB or the NFL?
- Would collecting 'big data' for the entire length of games change our results?
- How do we use these findings to strengthen and prolong the success?

PREVIOUS LITERATURE

- The data implies there is not a bias on home games or crowd size
  "All estimated coefficients on the interaction term between Home and Fan are not statistically significant, meaning referees do not treat home and away teams differently in foul non-calls when fans are present" (Gong, 2022)
- Prior research indicates bias towards star players, but not different teams
  "There is player-specific bias, but only positive. There is no team-specific bias" (Pelechrinis, 2022)

METHODOLOGY

- 16,873 calls
- 2017-2022 seasons
- Official NBA "Last Two Minute Reports"
- 6 independent variables
- Find which variables influence referee accuracy
- Logistic regression model to predict referee errors

DESCRIPTIVE STATISTICS

Table 1.
Descriptive Statistics and Correlations

Variable	M	SD	Time	Away	Home	Season	Incorrect	Comm
Time in seconds	44.643	35.419						
Away score	111.83	11.857	-.010					
Home score	112.21	12.312	-.009	.889*				
Season	2019.9	1.899	-.028*	.192*	.173**			
Incorrect call	.05	.213	.023*	-.014	-.010	-.037**		
Home commits foul	.49	.500	.017*	.032*	-.050**	.022**	-.006	
Playoff game	.07	.250	.014	-.094*	-.101**	-.064*	.004	.009

** Correlation is significant at the 0.01 level (2-tailed)
* Correlation is significant at the 0.05 level (2-tailed)

LOGISTIC REGRESSION MODEL

Table 2.
Variables in the Equation

Variable	B	S.E.	Wald	df	Sig.	Exp(B)
Time in seconds	.003	.001	8.537	1	.003	1.003
Away score	-.011	.008	2.057	1	.151	.989
Home score	.008	.007	1.116	1	.291	1.008
Season	-.082	.020	16.262	1	<.001	.921
Home commits foul	-.047	.076	.384	1	.535	.954
Playoff game	.040	.151	.071	1	.789	1.041

RESULTS

- Between 0.2% and 0.6% of the pseudo-variance in the dependent variable is explained by the independent variables (Cox & Snell and Nagelkerke)
- Our model exhibits good model fit, as the Hosmer and Lemeshow test does not show statistically significant results (p = .401)
- Our model did not successfully predict any incorrect calls; no systematic bias that the software can detect and utilize in predictions
- Time in seconds (p = .003) and season (p < .001) yielded statistically significant results on the dependent variable

IMPLICATIONS

- There is no major indication of a systematic issue in the NBA
- There was significant correlation in time and seasons, but not in any other of our independent variables
- When should a coach challenge?
- What can other leagues learn?