HOW HIGH SCHOOL LOCATION AFFECTS YOUR ODDS TO MAKE IT TO THE NBA

by

Rajat Chopra

A thesis submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Master of Science in
Economics

Charlotte

2021

Approved by:

____
Dr. Craig Depken

____
Dr. Artie Zillante

____
Dr. Paul Gaggl

ABSTRACT

RAJAT CHOPRA. How high school location affects your odds of making it to the NBA.
(Under the direction of DR. CRAIG DEPKEN)

Some states or metropolitan statistical areas produce more professional athletes when compared to their representation in the population whereas some areas produce less than their represented population. Figure 1 shows the geographical distribution during the 2014-15 season for the number of active NBA players born in each state compared to the number of active NBA players who went to high school in each state in the United States. Figure 3 also supports my assumption that the parents of the athletes move or send their children to schools in favorable locations during one's formative years to help their child accomplish higher levels in professional sports.

In this study, I assess whether factors related to where an athlete went to high school and the characteristics of the city where they went to high school influence their likelihood of playing professional basketball. I am going to test for a high-school-location bias towards income, education, and poverty levels for a particular metropolitan area. The findings of this paper will help me determine that location of a high school has a positive effect on the likelihood of making it to the NBA due to the environmental factors of the metropolitan area that set an athlete up for success during their formative years.

ACKNOWLEDGMENTS

I would like to thank my thesis committee chair, Dr. Craig Depken, for his continuous support of my study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis from sourcing data to selecting appropriate methods for analysis.

Besides my committee chair, I would like to thank the rest of my thesis committee: Dr. Artie Zillante, and Dr. Paul Gaggl, for their encouragement, insightful comments, and hard questions to spark my interest to pursue this study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

NBA	National Basketball Association
MSA	Metropolitan Statistical Area
U.S.	United States
FPL	Federal Poverty Threshold
PRM	Poisson Regression Model
NBRM	Negative Binomial Regression Model
ZINB	Zero-Inflated Negative Binomial Regression Model
IV	Independent Variable

CHAPTER 1: INTRODUCTION

Comparative analyses suggest that contextual factors associated with the location of high school contribute more towards the likelihood of an athlete playing professional basketball as there are more professional basketball players who went to high school in the U.S. but were not born in the U.S.

There have been prior studies which highlight the variation in individual sport development due to differences in learning opportunities and psychological environment (Côte´, Baker, & Abernethy, 2003). These differences in environmental experiences during the formative years of one's life could be a deciding factor between professional and aspiring athletes, because of differences in performance level, in motivation to practice, and the type of skills acquired (Bloom, 1985; Côte´ et al., 2003). Studies have highlighted that critical incidents that promote a child investing in one sporting activity over others include positive experiences with a coach, encouragement from an older sibling, early success, and/or simple enjoyment of the activity (Carlson, 1988; Côte´, 1999; Kalinowski, 1985; Monsaas, 1985). Kalinowski (1985) describes the early years of 21 elite swimmers as critical for later achievement in swimming as follows: ‘‘Had there been no excitement during the early years, and no sense that the young swimmer was very successful, there would never have been a middle or later period’’.

Another environmental variable that has not received much attention in sport expertise research is the size of the city where elite athletes gain their formative experiences. This variable could have a significant influence on how athletes will first be exposed to sports, which can limit or benefit performance. It is apparent that many children who live in smaller cities have access to facilities that introduce them to sport in different ways than children from larger cities. The children in larger cities usually have access to more resources compared to children from smaller cities (e.g. arenas, or specialized coaching). Urban athletes are also more likely to practice their sport in a structured setting such as a league, which is monitored by coaches with specific practice times and games, whereas individuals in smaller cities are more likely to engage in games without the structure of the urban setting (Côte´ et al., 2003). There might also be greater diversity in player size and ability in small cities, since all the children from the neighborhood gather to play together independent of age and ability. Urban athletes, who live within a more densely populated and structured environment, usually find themselves playing opponents and having team-mates who are all relatively the same age, size, and ability. It has been suggested (Côte´ et al., 2003; Soberlak & Côte´, 2003) that more opportunities to play with older children and adults and experiment with different types of sport and physical activity, such as those found in rural settings, might facilitate the development of sport expertise. However, data on the ‘‘urban – rural’’ debate are limited.

In one qualitative study of 10 Swedish elite tennis players, Carlson (1988) concluded that elite players predominantly came from rural areas, and that these areas provided the athletes more opportunities to participate in sports. Carlson also suggested that coaches in rural areas were more likely to take great care in maintaining the player – coach relationship even if they did not have the technical tennis knowledge of the coaches in urban cities. Curtis and Birch (1987) suggested that ‘‘top ice hockey players are more likely to come from communities large enough to build rinks, but not so large that the demand for ice time outweighs opportunities to skate’’. The qualitative nature of Carlson’s (1988) study and the focus of Curtis and Birch’s (1987) study on ice hockey did not permit the identification of optimal city sizes for sport development in youth. The primary purpose of my study is to examine whether the location and the size of the metropolitan statistical area in which an athlete attends high school influences the likelihood of playing professional sport and to compare the magnitude of any observed influences of high school location effects on the probability of becoming a professional athlete in the NBA. Another aspect in which my study differs from the previous studies is that none of the prior studies have studied the location effect for players not born in the United States whereas my study incorporates those players who were born outside the U.S. but went to high school in the U.S. To maximize generalizability and identify effects that may be time specific, details of NBA athletes from different decades will be captured.

CHAPTER 2: MATERIALS AND METHODS

CHAPTER 2.1: Data

The data for the players on the NBA rosters of each team who were active from 2010 to 2018 are evaluated for this study. There are a total of 960 active players in the NBA who went to high school in the United States from 2010-18. The dataset includes the foreign born players who went to high school in the United States. The player information dataset including the high school name, city, county, and the state is collected from https://www.basketball-reference.com. Figure 1 captures the 960 players that attended high school in the United States and were active during 2010-18 in the NBA. Idaho and North Dakota are the only states that do not have any representation from the players that attended high school in these states.

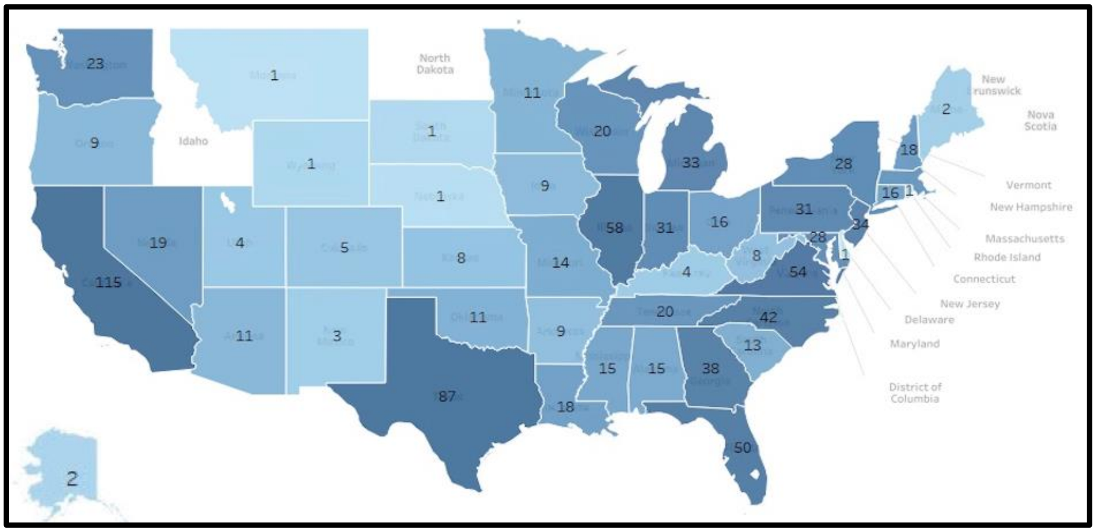


Figure 1: Geographical distribution of number of NBA players active from 2010-18 by state

Figure 2 further explores the representation of players that attended high school in each MSA. It is evident that certain MSAs produce significantly more players than the others. It is worth noting that even in a state like Texas, which produced 87 players, all the players came from the MSAs located in the eastern half of the state. This makes it crucial to study the education, income, and poverty levels at the MSA level. Figure 2 portrays that some MSAs produce a significant number of players whereas the majority of the remaining MSAs did not produce any players.

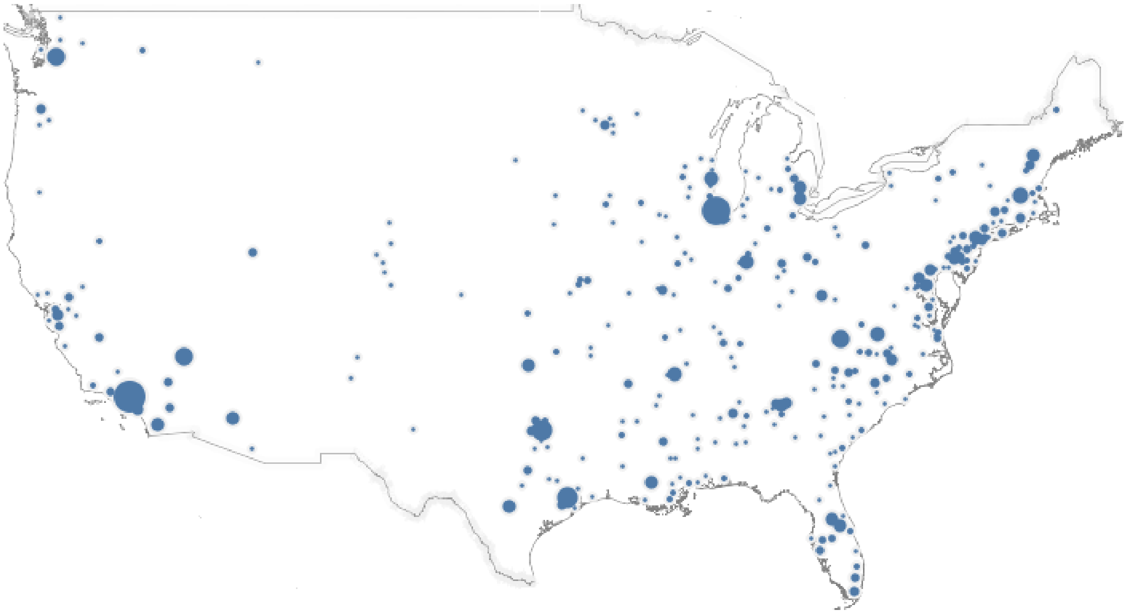


Figure 2 Geographical distribution of number of NBA players active from 2010-18 by MSA

Figure 3 helps us further examine the data. There are a handful of MSAs that produced a significant number of players, namely Los Angeles with 53 players. On the contrary, 2824 MSAs did not produce even a single player which will be crucial in selecting the appropriate methodology explained in Section 2.2.

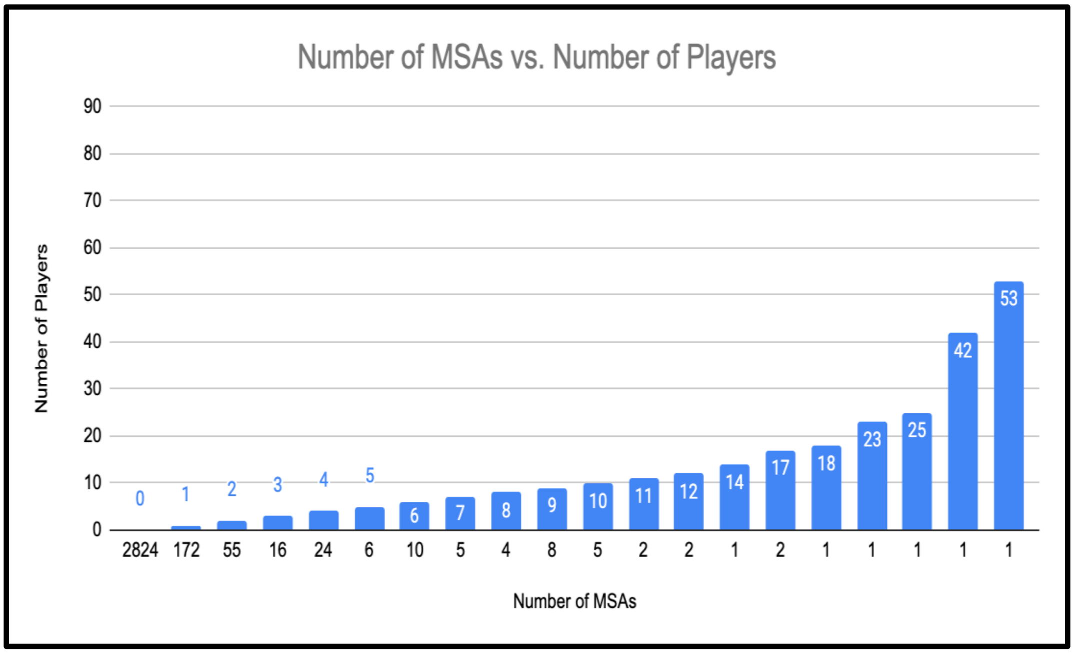


Figure 3 Number of players vs. number of MSAs

The data for the poverty, income, unemployment, and education levels, is collected from https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/. The county characteristics referenced in my study is for the year 2017. Table 1 shows the summary of the data used in this study. Median household income is in dollars, and education levels, unemployment rate, and poverty levels are in percentage form. The education levels are measured as the percentage of population having 1) less than high school diploma 2) only high school diploma 3) associates or some college degree 4) bachelors or higher degree. Poverty levels that are evaluated in my study are measured as the percentage of population that is 1) below poverty threshold for all age groups 2) below poverty threshold for ages 0 to 17. The FPL is an economic measure that is used to decide whether the income level of an individual or family qualifies them for certain federal benefits and programs.

Table 1 Summary statistics

	<high school	**highschoolonly**	**some college**	**bachelorsorhigher**	**UErate**	**medianhh income**	**povertyall**	**poverty 0to17**
Min:	1.10	7.30	8.80	4.70	1.60	22679	3.00	2.70
1st Qu:	9.00	30.00	27.00	14.70	3.50	42275	10.90	14.80
Median:	12.40	34.80	30.60	19.00	4.40	48885	14.40	20.50
Mean:	13.81	34.42	0.56	21.21	4.62	51091	15.38	21.55
3rd Qu:	17.70	39.40	34.00	25.30	5.40	56696	18.40	26.70
Max:	58.70	54.90	46.70	78.10	20.10	136191	56.70	74.70

The high school location of athletes provides a proxy for the location in which children spent their developmental years. It is important to recognize that the place of birth does not always coincide with the place of development. For example, athletes born in large urban centers might have moved to smaller communities during their development or, conversely, athletes born in small towns might have moved to larger cities. Figure 4 highlights the active NBA players in the 2014-15 season that were born in each state compared to the number of players that attended high school there. The larger states, such as California, Texas, and Florida, do not show much deviance between the two; however, states like, Nevada and North Carolina, do stand out as they show significant representation from players that went to high school in these states compared to players who were born there.

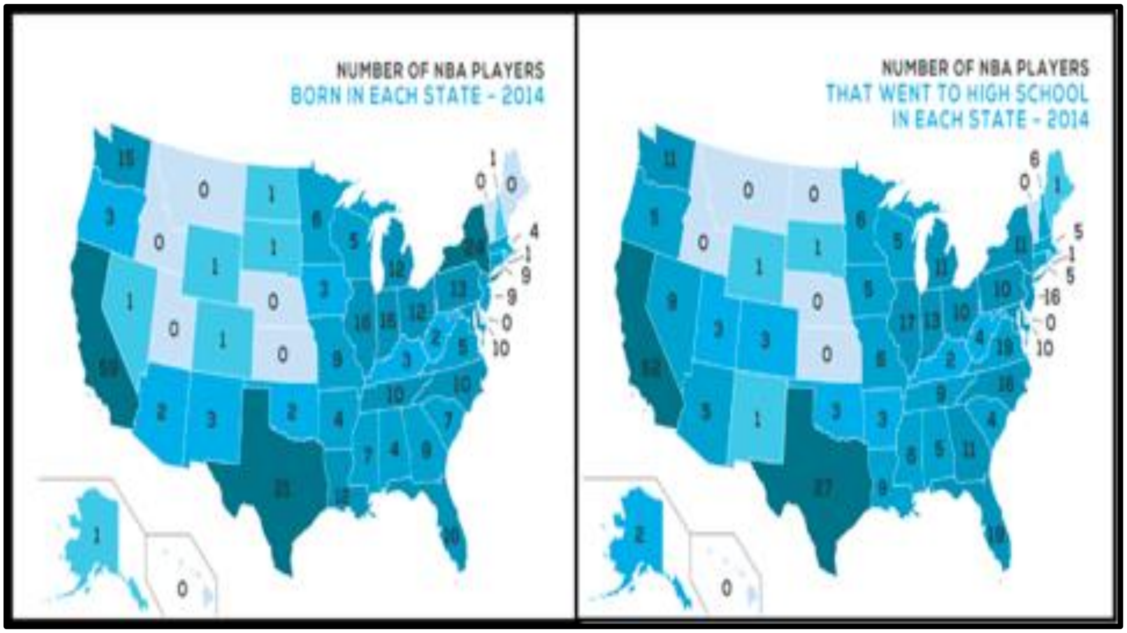
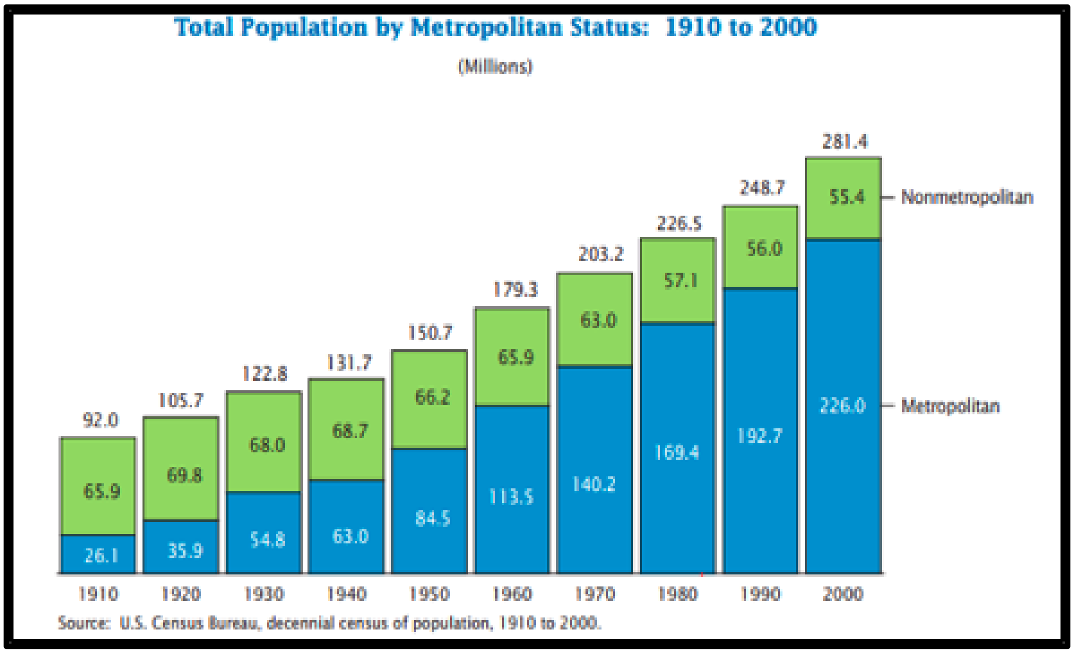


Figure 4 Geographical distribution of number of NBA players in the 2014-15 season

As per the United States Department of Agriculture's (USDA) Economic Research Service, the net movements between the two are geared towards metropolitan areas as the size of the rural communities has been shrinking over time since the mid – 1990s (www.ers.usda.gov). Per the census data, 85% of the population resides in 382 MSAs.

Figure 5 shows the shrinkage of non-MSA population and increase in MSA population from 1910 – 2000 (U.S. Census Bureau)



Source: U.S. Census Bureau, decennial census of population, 1910 to 2000.

Figure 5 Histogram for the total MSA population compared to the total non-MSA population from 1910 - 2000

2.2. Methods

The number of active NBA players during 2010-18 that attended high school in each MSA is going to be the dependent variable in this analysis. The dependent variable is a count variable; thus, count data models are used to study the characteristics of the MSA that produced players. I examined Poisson Regression Model (PRM) and Negative Binomial Regression Model (NBRM) for my analysis before choosing a Zero-Inflated Negative Binomial (ZINB). The independent variables chosen for this analysis are the Unemployment rate of the MSA, percentage of population with less than high school education to examine the education level, poverty levels for ages 5 through 17 to target the economic conditions of the players in the high schools, and the rural urban continuum code to study the prevalence of a particular urban influence code. The rural urban continuum forms a classification scheme that distinguishes metropolitan counties by the population size of their metro area, and nonmetropolitan counties by degree of urbanization and adjacency to a metro area. This variable in this database groups the 2013 Rural-Urban Continuum Codes (also referred to as the Beale Codes) into 3 categories: metropolitan counties (rural-urban continuum codes 1–3), nonmetropolitan counties (rural-urban continuum codes 4–9), and unavailable (blank or unknown). All of these variables have been converted to their natural log for this analysis. Other variables, such as median household income and poverty levels for the entire population were found to be insignificant.

Here are the results for each of the methods examined for my analysis:

1) Poisson Regression Model

The probability of a count is determined by a Poisson distribution in PRM, where the mean of the distribution is a function of the IVs. The conditional mean of the outcome is equal to the conditional variance.

Table 2 Poisson Regression Model results

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	12.20222	0.50388	-24.217	<2e-16 ***
ln_UErate	0.35207	0.16363	2.152	0.0314
ln_bachelorsorhigher	2.35753	0.09827	23.898	<2e-16 ***
ln_ruralurbancontcd	-1.54835	0.05949	-26.025	<2e-16 ***
ln_poverty0to17	1.48379	0.08858	16.750	<2e-16 ***

I used the residual deviance to perform a goodness of fit test for the Poisson Regression Model. The residual deviance is the difference between the deviance of the current model and the maximum deviance of the ideal model where the predicted values are identical to the observed. Therefore, if the residual difference is small enough, the goodness of fit test will not be significant, indicating that the model fits the data. I can conclude that the model does not fit well because the goodness-of-fit chi-squared test is statistically significant. It can be due to omitted predictor variables, and/or due to over-dispersion.

Table 3 Goodness of fit results

	res.deviance	df	p
[1,]	3249.067	3136	0.077910537

The results of the analysis came out to be significant; however, the model did not fit well due to over-dispersion in the data as evidenced by the results from the Pearson Test. The dispersion ratio came out to be 10.85113 which suggests that data are highly dispersed.

2) Negative Binomial Regression Method

The NBRM overcomes one limitation of the PRM as it does not make the equidispersion assumption about the data; therefore, it is suitable for over-dispersed data.

Steps:

y = the vector of player counts from each MSA

X = the matrix of predictors or independent variables.

size of matrix X is a (n x m)

λ = the vector of event rates. The vector λ is a primary characteristic of count based data sets. λ is a vector of size (n x 1). It contains n rates [λ_0, λ_1, λ_2,…,λ_n], corresponding to the n observed counts in the counts vector y. The rate λ_i for observation 'i' is assumed to drive the actual observed count y_i in the counts vector y. The λ column is not present in the input data. Instead, λ vector is a deduced variable that is calculated by the regression model during the training phase.

Table 4 Negative Binomial Regression Model results

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-12.7273001	1.132344	-11.240	< 2e-16 ***
ln_UErate	-0.008367	0.285901	-0.029	0.977
ln_bachelorsorhigher	2.528270	0.220622	11.460	<2e-16 ***
ln_ruralurbancontcd	-1.298286	0.098657	-13.160	<2e-16 ***
ln_poverty0to17	1.566746	0.208444	7.516	5.63e-14 ***

NBRM fits the model well as the goodness-of-fit chi-squared test is not statistically significant; however, there is still over-dispersion of data as the over-dispersion is 7.342201.

3) Zero-Inflated Negative Binomial Regression Method

ZINB a two part model. However, both parts predict zero counts. The count model predicts some zero counts, and on top of that the zero-inflation binary model part adds zero counts, thus, the name zero “inflation”. This is more suitable for situations where there are two types of zeros, i.e., structural zeros and sampling zeros. The observations with the structural zeros are those that can only have zeros. Those with sampling zeros are those which can experience non-zero outcome, but by chance experienced zeros. I created dummy variables for non_mtero_urban areas (ruralurbancontcd 4-8) and rural areas (ruralurbancontcd 9-10) to account for zeroes to be considered as a predictor for the logistic part of the zero inflated model.

Table 5 Zero-Inflated Negative Binomial Regression Model results

Count model coefficients (negbin with log link):				
	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-21.34374	1.53072	-13.944	< 2e-16 ***
ln_UErate	1.02841	0.38493	2.672	0.007547
ln_bachelorsorhigher	4.28063	0.29813	14.358	< 2e-16 ***
non_metro_urban	1.10469	0.33253	3.322	0.000894***
rural	3.55693	0.68531	5.190	2.10e-07***
ln_poverty5to17	1.82424	0.22768	8.012	1.13e-15***
log(theta)	-1.31263	0.09607	-13.663	< 2e-16 ***

Zero-inflation model coefficients (binomial with logit link):				
	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-14.02	259.39	-0.054	0.957
non_metro_urban	15.73	259.39	0.061	0.952
rural	17.86	259.39	0.069	0.945

In order to further examine the best model between NBRM and ZINB, I compared the two models using the Likelihood Ratio test. The results of the test show that the Zero-Inflated Negative Binomial model is the best model for this analysis out of these two models.

```
lrtest(NBRM, ZINB)

  #Df  LogLik Df  Chisq Pr(>Chisq)

1   6 -1361.9

2   10 -1400.4  4 76.877   7.984e-16 ***
```

This null hypothesis would be rejected at nearly every significance level; thus, the ZINB model is preferred as it increases the accuracy of our model by a substantial amount.

In order to interpret the results of the ZINB, I exponentiated the coefficients, which places the coefficients in an odds-ratio scale. The logistic part of the model predicts non-occurrence of the outcome.

Table 6 Zero-Inflated Negative Binomial Regression Model results (Exponentiated coefficients)

	Count Model	**Zero Inflation Model**
(Intercept)	5.376869e-10	8.175304e-07
ln_UErate	2.796608e+00	
ln_bachelorsorhigher	7.22863e+01	
non_metro_urban	3.018299e+00	6.754888e+06
rural	3.505527e+01	5.707735e+07
ln_poverty5to17	6.198092e+00	

CHAPTER 3: DISCUSSION

Zero-Inflation model: The baseline odds of being amongst the MSAs that never produce a player is 0.0000008. Count model: The baseline number for producing a player is 0.15 amongst those MSA that have a chance of producing a player. A unit increase in unemployment rate increases the odds by 2.79 times. A unit increase in bachelorsorhigher education level increases the odds of producing a player by 72 times for areas that have a chance of producing a player.

The results of this study provide support for the view that environmental factors do play a role in development of an emerging player and thereby, helping them make it to the professional level. It seems intuitive that an area with a low education level will have low median household income and therefore, making it less likely to have adequate training and coaching facilities. The low income levels of the population will also leave less on the table for families to invest in additional training for their kids.

On the other hand, an area with higher education levels will comparatively have higher paying jobs, and consequently, families residing in those areas will have more funds to spare for additional training for their children. There are always exceptions to the rule where a player emerges from the poorest of towns based on sheer determination and hard work. There are also some other outliers to this analysis because of the presence of some schools, such as Oak Hill Academy, a private school located in Mouth of Wilson in Grayson County. This school has produced a disproportionate number of professional basketball players in the past few decades compared to the size of its population. That can

be attributed to the coaching facilities provided at the school and its ability to attract emerging talent from across the country. However, there is a relative shift in population growth for areas with better schools which ultimately affects the education and unemployment levels in that area as well. The high school location effect found in this study reinforces the conclusion
that certain environmental factors play a crucial role in determining who makes it to the highest level in the world of basketball.

REFERENCES

Baker, J., Côte´, J., & Abernethy, B. (2003). Sport specific training, deliberate practice and the development of expertise in team ball sports. Journal of Applied Sport Psychology, 15: 12 – 25.

Baker, J., Côte´, J., & Abernethy, B., & MacDonald, D. (2005). When ‘‘where’’ is more important than ‘‘when’’: Birthplace and birthdate effects on the achievement of sporting expertise. Journal of Sports Sciences, 24(10): 1065 – 1073

Bloom, B. S. (1985). Developing talent in young people. New York: Ballantine. Boucher, J.

Côte´, J. (1999). The influence of the family in the development of talent in sports. The Sports Psychologist, 13: 395 – 417.

Coˆ te´, J., Baker, J., & Abernethy, B. (2003). From play to practice: A developmental framework for the acquisition of expertise in team sports. In J. Starkes & K. A. Ericsson (Eds.), Expert performance in sports: Advances in research on sport expertise (pp. 89 – 110). Champaign, IL: Human Kinetics.

Kalinowski, A. G. (1985). The development of Olympic swimmers. In B. S. Bloom (Ed.), Developing talent in young people (pp. 139 – 192). New York: Ballantine.

Carlson, R. C. (1988). The socialization of elite tennis players in Sweden: An analysis of the players' backgrounds and development. Sociology of Sport Journal, 5: 241 – 256.

Monsaas, J. A. (1985). Learning to be a world-class tennis player. In B. S. Bloom (Ed.), Developing talent in young people (pp. 211 – 269). New York: Ballantine

Curtis, J. E., & Birch, J. S. (1987). Size of community of origin and recruitment to professional and Olympic hockey in North America. Sociology of Sport Journal, 4: 229 – 244.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1986). Numerical recipes. New York: Cambridge University Press.

Soberlak, P., & Côte´, J. (2003). The developmental activities of elite ice hockey players. Journal of Applied Sport Psychology, 15: 41 – 49.

United States Department of education (www.ers.usda.gov)

US Bureau of the Census (www.census.gov)