Macalester College
DigitalCommons@Macalester College

Mathematics, Statistics, and Computer Science Honors Projects

Mathematics, Statistics, and Computer Science

5-5-2014

Hurdle Models and Age Effects in the Major League Baseball Draft

Justin Sims
Macalester College, jsims@macalester.edu

Follow this and additional works at: https://digitalcommons.macalester.edu/mathcs_honors

Part of the Computer Sciences Commons, Mathematics Commons, and the Statistics and Probability Commons

Recommended Citation

Sims, Justin, "Hurdle Models and Age Effects in the Major League Baseball Draft" (2014). *Mathematics, Statistics, and Computer Science Honors Projects*. 35.
https://digitalcommons.macalester.edu/mathcs_honors/35

This Honors Project - Open Access is brought to you for free and open access by the Mathematics, Statistics, and Computer Science at DigitalCommons@Macalester College. It has been accepted for inclusion in Mathematics, Statistics, and Computer Science Honors Projects by an authorized administrator of DigitalCommons@Macalester College. For more information, please contact scholarpub@macalester.edu.

Hurdle Models and Age Effects in the Major League Baseball Draft

Justin Sims

Honors in Mathematics, Statistics, and Computer Science

Advisor: Vittorio Addona

May 5, 2014

Abstract

Major League Baseball (MLB) franchises expend an abundance of resources on scouting in preparation for the June Amateur Draft. In addition to the classic "tools" assessed, another factor considered is age: younger players may get selected over older players of equal ability because of anticipated development, whereas college players may get selected over high school players due to a shortened latency before reaching the majors. Additionally, Little League rules in effect until 2006 operated on an August 1-July 31 year, meaning that, in their youth, players born on August 1 were the eldest relative to their cohort. We examine the performance of players selected in the June Draft from 1987-2011. We find that for all draftees, more relatively old players are selected in the Draft. Conversely, for high school (HS) draftees, both relative age and absolute age have a significant negative relationship with the odds of reaching the major leagues. Given that a HS draftee reaches the majors, there is no difference in professional performance based on age or relative age, measured by games played, wins above replacement (WAR) and on-base plus slugging percentage (OPS). For college draftees the results are less clear. We find that age, but not relative age, has a significant negative relationship with the odds of reaching MLB. Given that a college draftee reaches the majors, there is no difference in professional performance based on age or relative age. Had the draft market operated efficiently, neither relative age nor age on draft day would have captured additional variation in performance after controlling for draft position and other factors. We conclude that teams have undervalued both absolutely, and relatively, younger high school players in the draft and have undervalued absolutely young and relatively old college draftees.

Contents

1 Introduction

Do Major League Baseball (MLB) teams always make the "best" selections in the June Amateur Draft based on players future performance? Well, the simple answer is no. Take Mike Piazza for example who was selected in the 62^{nd} round, $1,390^{th}$ overall, by the Los Angeles Dodgers in 1988 who accumulated a career 59.2 WAR (average WAR for all draftees 1987-2011 was .45) over 1,912 career games played at the major league level and is likely headed to the Hall of Fame in the coming years. At the same time, the Dodgers 1^{st} round pick, 5^{th} overall, that year was Bill Bene who spent 9 seasons bouncing around the minor leagues before washing out prior to reaching the majors and eventually getting a 6 month jail sentence for operating a counterfeit karaoke business and failing to pay taxes on the income he earned. That is a pretty stark difference in career outcomes. A more recent example is Wade Townsend who was selected 8^{th} overall by the Tampa Bay Devil Rays (now the Rays) in 2005 only to play in 4 affiliated minor league seasons before being released after an injury riddled career; while pitchers such as Tommy Hanson and Jaime Garcia have enjoyed major league success after being selected in the 22^{nd} round of the same year.

What do these successful players have in common? Perhaps not much at all, but these examples illustrate that teams do not have perfect information in the market about who will succeed and who will fail. This occurs even though MLB franchises expend an abundance of resources on scouting each year in preparation for the June Amateur Draft. Thus, we attempt to find places where we can inform teams on how to better select their players. We attempt to find any inefficiencies in the Draft market based on age effects.

In addition to the classic "tools" (e.g. ability to hit for average, hit for power, throw, field, and run) assessed, another factor considered by teams is age: younger players may get selected over older players of equal ability because of anticipated development; whereas college players may get selected over high school players due to a shortened latency before reaching the majors and less variability in possible career outcomes. Additionally, Little League rules in effect until 2006 operated on an

August 1–July 31 year, meaning that, in their youth, players born on August 1 were the eldest relative to their cohort. We test the hypotheses of whether teams have undervalued certain players based on age differences.

The MLB Draft market is very different from many other professional sports for a variety of reasons. Teams are not drafting for current need as they may in the National Basketball Association (NBA) and the National Football League (NFL), but rather for 3-5 years down the road, as players must climb the ranks of the minor leagues before making a major league roster. This latency leads to many unforeseen outcomes for top draft prospects who may not develop as expected, sustain a severe injury, or become blocked at the major league level by other more experienced players already on the roster. On the positive side this extra time in the minor leagues could lead to greater development than expected, where teams can reap returns on lower round selections. Therefore, there is greater variability in the return on investment in the MLB draft and we wish to shrink the gap by determining whether or not some characteristics of players tend to be correlated with more or less career success.

The paper is organized as follows: In section 2 we present our literature review. This includes papers on the relative age effect in general, the relative age effect in baseball specifically, the relative age effect reversal, and market inefficiencies in the June Amateur Draft. In section 3 we describe our data which includes all top 50 round draftees from 1987-2011. Section 3 also includes our methods and summary statistics. Section 4 presents the results of our analysis. We use logistic regression, zero inflated count regression techniques, generalized linear models and OLS to estimate the relationships between age and relative age with baseball career performance. We also perform other analyses dealing with the traditional relative age effect and possible market inefficiencies in the draft. In Section 5, we discuss our findings, in particular whether we see the relative age effect, the relative age effect reversal and over/under valuation of certain players. Section 6 provides some concluding remarks, while section 7 suggests areas for future research. All figures and tables are provided in the Appendix.

2 Literature Review

Since the seminal paper by Barnsley, Thompson and Barnsley (1985) the relative age effect (RAE) has been a hot topic of study in the sports world. Barnsley, et al. (1985) find that there exists a strong relationship between the month of birth (from January through December) and the proportion of players playing in the National Hockey League (NHL). Using data from the 1982-83 NHL season, they show that nearly twice as many players were born in the first three months of the year compared to the last three months. Additionally, the RAE was more pronounced in two Canadian junior hockey leagues, the Ontario Hockey League and the Western Hockey League, where more than three times as many players were born in the first quarter compared to the last quarter.

To explain this finding in hockey, Barnsley et al. (1985) suggest that the method of grouping used in minor hockey (up to age 20) results in a developmental-age advantage for those born in the early months of the sport year. For minor hockey league players, the age cohorts correspond to the calendar year. The children born later in the calendar year compete for a position against other children who, on average, bigger, stronger and in general more developed physically. Accordingly, the relatively older children will be given further advantages such as better coaching, more time on the ice, placement on All-Star teams (which leads to play against better competition), and more rewards and recognition since they are seen to be of higher status within the local sporting community.

In baseball, Thompson, Barnsley and Stebelsky (1991) examined the birth months of 837 major league baseball (MLB) players to see if there was a relationship between relative age and participation at the MLB level. Additionally they re-examined data from 1985, using 682 players, which had been previously determined to not exhibit the relative age effect (Daniel and Jansen 1987). As opposed to hockey, where the cutoff date for the pee wee level is January 1, the cutoff date in baseball is August 1. Thus we would not expect a disproportionately high number of January, February and March birthdays, but rather a disproportionately high number of August, September and October birthdays. They use Spearman rank-order correlation between birth month relative age rank (i.e. August is assigned a rank of 1 and July a rank of 12) and birth month frequency rank (where the highest frequency of players is assigned a rank of 1, the second highest was assigned a rank of 2, and so on). The findings reflect the expected prior belief, that “there is a significant tendency for professional players to have been born early in the baseball year”.

The reasoning for this result was that players born shortly after the cut-off date were relatively older compared to their peers and thus more physically mature. Such a developmental advantage gained by the relatively older players “when competing against other youngsters who are considerably younger, although they are placed in the same age category for league play” led to real effects where relatively older players were deemed more talented (even though the skill level differences observed were most likely due to age differences), thus receiving preferential treatment growing up from coaches and exposure to opportunities not afforded to those of less physical maturity. They find that the correlation here is not as strong as in hockey and suspect that this is because “Little League baseball starts at a later age than hockey, thus reducing the magnitude of the relative age effect”.

In addition to the initial studies on hockey and baseball, this same effect has been deeply chronicled in minor league and professional hockey, and soccer as well as observed in American football (Daniel and Janssen 1987, Glamser and Marciani 1992, Stanaway and Hines 1995), handball (Ryan 1989), swimming (Baxter-Jones 1995, Ryan 1989), tennis (Baxter-Jones 1995, Dudink 1994) and volleyball (Grondin et al. 1984, Ryan 1989). More recently, Addona and Yates (2010) investigated the RAE using complete data on every player who has ever played in the NHL, a total of 6,407 players (birthday information was available for 6,391 of these players). They made a case for when and why the RAE began to manifest itself in Canada. After accounting for actual birth distributions, they find that “all evidence indicates that the RAE is present for Canadian born players, regardless of whether we use a uniform birth distribution or adjust for the actual birth distribution” confirming the long-held belief that the RAE is present. Next, to pinpoint the year when the RAE began to materialize, Addona and Yates (2010) ran a change point analysis on the yearly difference in proportion of first and last quarter births from 1930 to 1987. The results indicated that the RAE was significantly present for players born since 1951, but not before.

Although a prior study (Daniel and Janssen 1987) suggested that a potential transitional event which led to the RAE was the international hockey series between Canada and the Soviet Union in 1972, Addona and Yates (2010) find that this is in fact not the case. “If the RAE was present in the 1985-86 NHL season, then the vast majority of these players were born in the early 1960s, or earlier, and went through youth hockey before any effects of the 1972 series could have been experienced.” However, what they do find is that “all signs point to the series of events surrounding the Soviets’ emergence on the international hockey scene in the early 1950s as the initial catalyst for the RAE. The circumstances that led to the RAE were meant to lead to the discovery of the best hockey talent in Canada” leading to relatively older players receiving preferential treatment and being selected for the top youth teams, as they were viewed at the time as the more talented players in their age groups. The RAE, as a byproduct, led to a lot of wasted genuine hockey ability in the relatively younger players who did not make it because of the extra training they did not receive.

Additionally, Addona and Yates (2010) find that there is no evidence of the RAE for hall-of-fame status, suggesting that relatively older players, once they reach the NHL do not continue to outperform their relatively younger peers. Therefore, even though relatively older players are more likely to reach the NHL, they are not more likely to enjoy exceptional careers as well. This provides an interesting topic of study which we will explore further: do relatively older players not only make it to the professional ranks more often, but also outperform their relatively younger peers throughout their professional careers? This research question has received less attention, although a recent paper by Deaner (2013) investigates whether older NHL players outperform their younger counterparts even holding draft position constant.

In addition to the relative age effect, the literature presents the idea of the “rise of the underdog" and a relative age effect reversal, where even though players born shortly before the cut-off date are disadvantaged growing up, leading them to be less likely to reach the professional ranks, these relatively younger players turn out to be more successful in the professional ranks compared to their relatively older peers. Deaner (2013) tested whether or not selection bias occurs in the NHL Draft, with the logic that "a player's draft slot serves as a measure of their perceived talent whereas career productivity indicates their realized talent. If selection bias occurs, then, for any given draft slot, relatively younger players will enjoy more productive careers." The data collected for this study included players drafted over a 27 year span, ending in 2007. The main productivity measure used for analysis was career games played and as a second productivity measure they considered career points scored.

Using Tobit regression analysis, Deaner finds that "compared to first quarter draftees, relatively younger draftees played more games" holding constant draft position. According to Deaner, this indicates that selection bias is indeed occurring because even though relatively younger players were drafted somewhat earlier than relatively older players, the regressions that controlled for draft position indicated that, given their future performance, they were not drafted early enough. In grouping players into birth month quartiles, the regressions indicated that a second quarter draftee played the same number of games as a first quarter draftee who was selected 20 slots earlier. Additionally, they tested to determine whether selection bias applies even to first round draft choices, where more is known about players and "decisions are weighed more carefully". The results indicated that the selection bias was maintained, even though the results were not statistically significant due to small sample size.

A cost Deaner (2013) associates with the observed selection bias is what is known as escalation, where a player drafted later, who an organization has not invested as much in, has to perform more than a player drafted earlier to receive equal playing opportunities. They note, "the hockey data yielded evidence consistent with escalation: games played was jointly predicted by points per game and draft slot; crucially the draft slot regression coefficient was negative. Furthermore, when position, height, and plus-minus per game were added as predictors, the effect of draft slot remained substantial." This may have serious consequences for the relatively younger players whose talent is underestimated on draft day, as they may not receive the playing opportunities that their talent warrants.

Regarding the relative age effect reversal (younger players outperforming relatively older players), Gibbs (2011) finds–using data from 2000-2009–that the traditional relative age effect (relatively older players outperforming relatively younger players) exists moderately for the average Canadian NHL player, but reverses when examining All-Star (2007-2009) and Olympic (1998-2010) team rosters. They find that "the percent of all Canadian hockey players in the NHL born in the first three months is a modest 28 percent" compared to previous results of 30-40 percent. Then upon examination of the most elite levels of play in hockey, the RAE dwindles. We see that of NHL All-Star rosters in 2007, 2008, and 2009 respectively, only 20 percent, 15 percent, and 13 percent consist of Canadian-born players with birthdays in the first three months of the year. These roster sizes are a small sample, but suggest that relatively young players may be outperforming their relatively older counterparts. This pattern can also be found among Canadian Olympians, as the 2010 gold medal-winning Canadian Olympic hockey team had 13 percent of its players born in January, February or March. Similar results were seen in 2006, 2002 and 1998.

Gibbs (2011) does not conduct a formal analysis of whether the results are significantly different than the 28 percent observed overall, but rather simply says that "it appears that being born at the start of the year reduces the chance of elite play. Consider the average distribution of players born in the first quarter of the year for the NHL Canadian-born players, 28 percent. The combined average of the All-stars and Olympic rosters is 17 percent. This represents a 40 percent reduction in the distribution of players born in the first three months of the year. If birth month had no effect on elite play, the percentage would remain 28 percent."

Additionally Gibbs (2011) finds that the average career of Canadian born NHL players born later in the year is longer. In fact, they find that those born in the first quarter of the year had a career duration of one season shorter than those born in the last quarter of the year. As opposed to Wattie et al. (2007), which found no relative age effect for NHL players when examining career length, Gibbs' result indicates a reverse relative age effect when examining the length of a player's career. Gibbs (2011) believes that this reverse relative age effect is a byproduct of the relatively younger players continually being challenged by their more advanced relatively older peers, where becoming an elite player comes from being the underdog and overcoming the odds. The question then becomes why it takes until the elite levels of professional hockey for this effect to be observed.

In addition to performance measures (career games and career points) and competing on the most elite level teams, we have observed that relatively younger players have been selected earlier in drafts revealing that the relatively younger "players may show superior performance compared with their relatively older peers" (Baker, 2007). Even though we already noted that relatively younger players outperform their draft position (Deaner, 2013) the evidence presented here provides insight into when the reverse relative age effect may begin. The result here indicates that it may not be only at the most elite level of play, but well before then, specifically for high school and college aged players, as they prepare to be drafted. The fact that we still observe some relative age effect in the NHL may be because of the extreme effect seen in junior hockey where there are so few players born in the later months compared to the early months. It may be that this is the only reason for observing that more players in the NHL come from the early months of the year.

Baker (2007) examined 1,013 North American players drafted to play in the NHL between 2000 and 2005. The results indicated that Canadian born players exhibited a negative correlation between birth quartile and draft round number using Spearman rank-order correlations. "Interestingly, these findings suggest that relatively younger athletes are more likely to be chosen in the earlier rounds of the draft." It is unclear from this paper whether the relatively younger players are selected more often in the earlier rounds of the NHL draft due to greater talent than their relatively older peers in their cohort or due to the perceived ability to improve more in the future than their relatively older peers due to anticipated development (i.e. the existence of an "aging curve"). As Deaner (2013) noted, since draftees can be both absolutely older and relatively younger on draft day, it is likely to be the former.

Furthermore, Baümler (2000) finds that amongst the youngest soccer professionals in Germany 68% are born in the first half of the soccer year. However, amongst the oldest professionals we see that only 49% are born in the first half of the soccer year. This is further evidence of the relative age effect reversal where even though the relatively old make it more often to the professional ranks, the younger players are those who enjoy more career success. Meanwhile, Williams (2010) found that the RAE existed in the FIFA U17 World Cup for all geographical zones, except for Africa, where the rosters of these teams were comprised of many players born in December. They did not have a clear reason for this difference, but suggested that it could have been due to recording or deliberate error, where many birth dates for children in this region are not recorded and birth certificates are not readily available.

The RAE phenomenon has been recognized outside the sporting world. With respect to education, relatively older students perform better in grade school, on cognitive exams and attend college more often (Bedard and Dhuey 2006, Crawford et al. 2007, Mayer and Knutson 1999). However, it is the relatively young individuals who enjoy more academic success by graduating college with better degrees (Pellizzari and Billari 2011, Russell and Startup 1986). Pellizari and Billari (2011) reconcile this finding with the idea that starting school earlier is connected to better results in the long-run (Fredriksson and Ockert 2005, Goodman and Sianesi 2005, Skirbekk 2005, Skirbekk et al. 2004, Black et al. 2009). Cunha and Heckman 2007 suggest that early investment in skills improves the return of future human capital investments. Due to the inverse-U shape for the physiological profile of cognitive development (Salthouse et al. 2004), the youngest in a cohort are penalized at early stages. However, given this development curve, it is conceivable that such a disadvantage levels off, and possibly reverses, at some later age. Some–Salthouse et al. (2004) and Jones (2005)– even "suggest that the turning point in the profile of cognitive development might, in fact, be between age 20 and 25 years," fitting in very nicely with this reversal in academic achievement at the university level.

Age in the MLB Draft has also been researched dealing with a possible market inefficiency in the ability of scouts to forecast the projected talent of MLB Draft prospects. Jazayerli (2011) suggests that players who are young for their draft class have been seemingly undervalued by professional scouts and organizations. Young in this sense does not mean relatively young, but rather absolutely young. Absolute age is the age of a player on a given day, say draft day. Given an average aging curve–one would expect that all else constant–a younger player should outperform an older player in the long run. Jazayerli claims that even though teams take into consideration this predicted development, teams have significantly underestimated it, leading to an inefficiency in the market.

Other potential market inefficiencies in the MLB Draft have been investigated through the years, including whether certain positions have been overvalued relative to others (Burger and Walters 2009, Salaga 2012), whether college players are better choices than high school players (Spurr 2000, Burger and Walters 2009, Lewis 2003, Bradbury 2011, Salaga 2012), whether certain teams are better at evaluating talent than others (Spurr 2000), and whether the area a player is drafted out of has a significant effect on career performance (Salaga 2012). The main results of these studies were that at one time college draftees were undervalued compared to high school draftees and hitters were undervalued relative to pitchers. However, teams were not significantly different in their player evaluations and "players from any specific geographical location do not exhibit characteristics which would significantly alter the odds of them reaching MLB as compared to a player from" New England (Salaga 2012). Additionally Wachter (2012) finds, using data from 2000-2005, that college draftees are not uniformly superior to high school draftees. Specifically, pre-arbitration (first three years in MLB) production, measured by wins above replacement (WAR) is greatest near the center of the defensive spectrum (OF, 3B). Moreover, college draftees average more pre-arbitration production than high school draftees at the more defensively demanding positions (2B, SS, 3B), while high school draftees produce more than college draftees at less defensively demanding positions (1B, OF) and pitcher.

The current work fills a void in the literature, in that it seeks to analyze the effects of both age and relative age on (1) draft selection, (2) chance of reaching MLB once drafted, and (3) different measures of MLB success once drafted, including games played at the major league level and wins above replacement (WAR). A comprehensive analysis such as this one, spanning 25 years (1987-2011) and 50 rounds of draft data per year, with major league statistics for all players who made the major leagues, has not previously been carried out.

3 Data and Methods

3.1 Data

Baseball-Reference.com and TheBaseballCube.com have in depth information on every player to ever be selected in the June Draft. We gathered data for all players drafted in the top 50 rounds of the June Draft from 1987-2011. We gathered information over this period, as 1987 was the first year that the June Draft was the only draft where players were selected from high schools and colleges in the United States and territories. Before 1987, other drafts included the June Secondary, January and January Secondary Drafts. We collected data for players up until 2011, as very few players drafted in 2012 reached the majors by June 2013. All performance data are as of June 12, 2013 and gathered from Baseball-Reference.com. The relevant variables here are Draft Year, Overall Pick Number, Position at the time of Draft, WAR, OPS, Games and Type of School Drafted from (college, high school, junior college or other). The relevant variables collected from TheBaseballCube.com are Draft Year, Overall Draft Selection, Draft Round, Team Drafted by, HS State, Birth Day, Birth Month and Birth Year.

We created a Last Time Drafted variable ($lastDraft$) that is a dummy variable, where 1 indicates that as of the 2011 Draft, this was the last time a player was drafted in the top 50 rounds. We created this variable in order to account for signability issues (e.g. players who either decide to attend college, as opposed to signing out of high school or decide to return to college for another year). We also created an *age* (in years) variable that expresses how old the draft pick was on draft day. Additionally we created a *relativeage* (in days) variable which expresses how relatively old given an August 1 Little League cut-off date a player is, with 1 representing July 31 and 365 representing August 1. We shifted all values of WAR up by the minimum value observed, plus .1, so that all values are positive, as we note that these values are distributed lognormally (see Figure 1). We call this variable Wins Above Worst Observed, but continue to refer to it as WAR throughout the rest of the paper.

We created multiple variables for transformations of overall draft selection, as it is clear that the relationship between overall pick number and production (WAR and Games) is not a linear relationship. The variables that we created are $\frac{1}{pick}$, $\frac{1}{\sqrt{pick}}$, $\log(pick)$ (see Figures 2 and 3). We created a dummy variable called *pitcher* that is a 1 if the player was drafted as a pitcher and 0 if the player was drafted as a position player, to differentiate between the evaluation of pitchers versus hitters.

There were some discrepancies in birth dates between Baseball-Reference and The Baseball Cube. For most players we used the birth date on Baseball Cube, but when it was clear that this birth date was incorrect, we used the Baseball-Reference birth date (if that was determined to be correct). Additionally, if we could not find a clear correct birthdate, we treated this player as missing. We manually checked (and changed if necessary) the top 5 rounds of each year and found approximately 2% of birth dates to be incorrect using The Baseball Cube data, thus we expect around this percent error in our results. We also checked highly potentially incorrect birth dates, i.e. players who were listed as under 17 or older than 24 on draft day and made corrections and omissions here when necessary.

We also used the draft order reported by The Baseball Cube, as there were some discrepancies between this source and Baseball-Reference. For the discrepancies we checked a third party source, http://mlb.mlb.com/mlb/history/draft/draft.jsp, "First-Year Player Draft History: June Amateur Draft" and determined the correct order accordingly.

3.2 Methods

The first stage of this research was to determine whether we observe the traditional relative age effect in the MLB Draft and MLB both graphically and analytically using chi-square goodness-of-fit tests (see Table 2 Figures 4 and 5). Once this was determined, we proceeded to examine the relationship between performance once drafted and age. To do this, we used four measures of baseball career performance: whether the player reached the majors, career number of games played at the major league level, career WAR and career OPS (for players drafted as position players only). We note that this is a two stage process, whereby a player must first reach the majors to be able to accumulate statistics. Thus, we use the generalized linear model, logistic regression, to determine the probability of reaching the majors modeled by age, $relativeage$, $\frac{1}{\sqrt{pick}}$, $pitcher$, factor($year$), and $lastDraft$.

After modeling the probability of reaching the major leagues, we continue on to MLB performance statistics, beginning with games played at the major league level. We model $games$ by the same variables: age, $relativeage$, $\frac{1}{\sqrt{pick}}$, $pitcher$, factor($year$), and $lastDraft$. However, $games$ is not continuous, but rather a special type of discrete random variable known as a count. Thus, we must use count regression techniques. We observe that there are two problems with the data: the presence of overdispersion and excess zeros. Overdispersion is the presence of greater variability in the data than would be expected given a certain distribution. Therefore, we cannot use a Poisson distribution (typically used in count models) which has the property that the mean equals the variance. So, we use a negative binomial distribution, which is similar to the Poisson distribution without the restriction of the mean equaling the variance.

To account for the excess zeros problem we must use zero inflated or zero truncated models, such as the hurdle negative binomial model and the zero inflated negative binomial model. Both of these regression techniques are two stage processes, first estimating the chance of a zero (for the hurdle model the results are identical to the logistic regression discussed above) and then the chance of some non-negative integer. The difference between the models is that the hurdle model conditions its count estimates on the variable not being a zero, whereas the zero inflated model reuses the zeros in its estimates, while controlling for the fact that there are an excess amount of zeros. Therefore, the hurdle model asks the questions, did you reach the majors and then given that you reached the majors how many games did you play? Whereas the zero inflated negative binomial model asks, did you reach the majors and how many games did you play (allowing the possibility of a zero for players reaching the majors)?

Formally, the density of the hurdle model can be written as:

$$f_{hurdle}(y; x, z, \beta, \gamma) = \begin{cases} f_{zero}(0; z, \gamma) & \text{if } y = 0 \\ (1 - f_{zero}(0; z, \gamma))\frac{f_{count}(y;x,\beta)}{(1-f_{count}(0;x,\beta))} & \text{if } y > 0, \end{cases} \tag{1}$$

where y represents games played, f_{zero} is the binomial probability function and f_{count} is the negative binomial probability function. The covariates in the count model, x, can be different from the covariates in the binomial model, z, if the factors influencing getting over the hurdle are different from those influencing the count once over the hurdle. In our analyses, however, we use the same covariates in each model, so $x = z$. The model parameters β and γ are estimated by maximum likelihood, where the hurdle and count components can be maximized separately. Since f_{count} is the negative binomial probability function, a dispersion parameter is estimated which distinguishes this from Poisson regression. This parameter is also estimated by maximum likelihood.

Due to the truncated zeros in the count model, interpretation of the coefficients is difficult compared with standard count regression. We now elaborate on this point briefly, since much of the applied work which employ these models do not make this clear. The Negative Binomial distribution with parameters k and r can be written as follows:

$$P(r) = \frac{(k+r-1)!}{(k-1)!r!}\frac{p^r}{(1+p)^{k+r}} \; (r = 0, 1, \ldots; p > 0; k \geq 1). \tag{2}$$

We observe that the expected value using the parametrization in (2) is kp. We also note that

$$P(0) = \frac{1}{(1+p)^k}, \tag{3}$$

thus we must divide equation (2) by equation (3) to find the truncated negative binomial distribution used in equation (1). First, we let

$$\omega = \frac{1}{1+p} \text{ and } \eta = 1 - \omega.$$

We find that the truncated negative binomial distribution, $P_t(r)$, has the form

$$P_t(r) = \frac{\omega^k}{1-\omega^k} \frac{(k+r-1)!}{(k-1)!r!} \eta^r \ (r = 1, 2, \ldots; k \geq 1; \omega, \eta \in (0,1)), \tag{4}$$

with expected value $\frac{k\eta}{\omega(1-\omega^k)}$.

Negative binomial count models are built for the parameter k. In standard Negative Binomial regression, we observe that if $k = \exp(\mathbf{X}\beta)$, then $E(R) = \exp(\mathbf{X}\beta)p$, which represents a constant multiplicative change in the mean by $\exp(\beta_i)$ for 1 unit increases of $x_i \in \mathbf{X}$. However, the truncated negative binomial model does not possess this property, as

$$E(R_t) = \frac{\exp(\mathbf{X}\beta)\eta}{\omega(1 - \omega^{\exp(\mathbf{X}\beta)})}, \tag{5}$$

which we can easily verify has a rate of change with respect to $x_i \in \mathbf{X}$ that is not independent of the values of x_i. Consequently, there is no simple interpretation of β_i. In our subsequent results, we report coefficient values, and for simplicity, report $\exp(\beta_i)$ as an approximation of the multiplicative change in the mean for 1 unit increases of x_i.

We next consider WAR. Since WAR is distributed lognormally, we cannot employ OLS regression. We observe that we can use generalized linear models using this distribution. Again, we use the same explanatory variables: age, $relativeage$, $\frac{1}{\sqrt{pick}}$, $pitcher$, factor($year$), and $lastDraft$.

OPS, unlike WAR, is distributed normally, thus we can use OLS regression using the same explanatory variables as above. The issue here is that some players who were drafted as hitters were converted to pitchers. This creates the issue where we have many values close to zero which should not be part of the distribution. We also then have the problem of pitchers who were converted to hitters, who are not captured in these regressions, even though they theoretically should be.

We briefly look into the proportion of high school and college players drafted over time (see Figure 6) and whether there exists a breakpoint. Additionally, we examine whether there are differing age and relative age effects over time. We also explore other questions presented in the literature including whether certain teams are better at evaluating talent than others, whether certain positions have been undervalued in the draft, whether certain regions have been undervalued and whether college draft choices are better selections than high school players.

3.3 Summary Statistics

Summary statistics for variables included in our analyses can be found in Table 1 (this does not include variables that we include as categorical such as year). Some stats of note include the low minimum age and high maximum age. Even though the ages seem extreme, we observe that the players on the high end are mostly players drafted out of Cuba, which makes sense, as it is unlikely that a player drafted out of college in the United States would be almost 28 years old. On the young end, we see that these players are mostly from outside of the United States, including, Puerto Rico, Cuba, while some are from high school's in the United States. Since most of these players were not drafted out of high school or college, they are not included in our analyses.

A few other summary statistics of note are that only 12.48% of all drafted players reached the majors, while the average WAR and median WAR were 4.947 and 4.5 respectively. We shifted WAR by 4.5, since the worst observed WAR was -4.4. This was done in order to make all WAR values positive. Thus, we see that the median WAR value was actually no better than replacement level, whereas the average WAR was .4 better than replacement.

4 Results

4.1 Distribution of Birth Months for Drafted Players

Figure 4 presents a histogram for the birth month distribution of all top 50 round MLB draftees from 1987-2011. Moreover, Table 2 illustrates the observed frequencies of birth months for draftees compared with theoretical frequencies based on uniform births and observed births in the United States from 1995-2002 (James, 2005). Using both the uniform distribution and theoretical distribution given observed births in the U.S., the chi-square goodness-of-fit test reveals that the observed distribution of the birth months of draftees was significantly different from both of these ($p < 2.2e-16$).

4.2 High School: Games Results

Using total games played as a measure of career success, we find consistent results between two count regression models, hurdle with negative binomial distribution and zero inflated negative binomial (see Tables 3 and 4). We note that the correlation of $lastDraft$ with games played does not alter the main results over the top 5, 10 and 50 rounds, thus we shall examine the models not including $lastDraft$. Over the top 5 rounds we observe that both age and relative age are negatively, but statistically insignificantly correlated with total games played (see Tables 4 (i) and 3 (i)). We note that using the hurdle model (see Table 3 (i)), a player one year older than another on draft day played in 92.46% ($p = 0.5076$) of the games that the younger one played in over the course of their careers. Additionally, a player 100 days older relatively than another played in 94.38% ($p = 0.2239$) of the games the younger player played in over the course of their careers. For a 1000 game career, this difference is approximately 56 games, only about one third of a full 162 game season.

Furthermore, we note that over the top 5 rounds $\frac{1}{\sqrt{pick}}$ is positively correlated ($p < .001$) with games played (see Table 3 (i)). The results indicate that a .1 unit increase in $\frac{1}{\sqrt{pick}}$ is correlated with an increase in games played by 19.12% (see Table 3 (i)). This means that holding all else constant, a player drafted with the 25 overall pick played in 19.12% more games than a player selected at 100^{th} overall. If the 100^{th} pick played in 1000 career games, then the 25^{th} pick played in 1191 games, over a full season more games. We also see that being as a pitcher has a negative relationship ($p < .001$) with a player's total games played (see Table 3 (i)). Holding all else constant, a player drafted as a pitcher played in 29.89% of the games of a hitter. Over a 162 game season (assuming a position player plays every game), a pitcher would be expected to pitch in approximately 48 games. This makes sense, as the typical starting pitcher will pitch in 32 games in a season, while the typical relief pitcher will throw in 70 games; the average of which is 51.

As aforementioned, the results using the zero inflated negative binomials were concurrent with the results using the hurdle model. We observe that a player one year older than another on draft day played 8.10% ($p = 0.4803$) fewer games over the course of their careers (see Table 4 (i)). For a 1000 game career, a player one year older on draft day would be expected to play in 81 fewer games, exactly one half of a full major league season. Moreover, a player 100 days older relatively played in 5.63% ($p = 0.2277$) fewer games over the course of their careers (see Table 4 (i)). This implies that if a player 100 days relatively younger than another played in 1000 games, the relatively older player would play in 56 fewer games. The results for pitcher and $\frac{1}{\sqrt{pick}}$ did not change from the hurdle model.

For both the hurdle and zero inflated models, we used negative binomial distributions as opposed to Poisson distributions due to overdispersion, which is the presence of greater variability in a data set than would be expected. This was an appropriate choice, as we see the coefficient on Log(theta) is significant ($p < .001$) for both models (see Tables 3 (i) and 4 (i)) indicating the presence of overdispersion given a Poisson distribution. Furthermore, the results for the top 10 and 50 rounds of the draft compared to the top 5 rounds do not change any of the main findings significantly for either the hurdle or zero inflated models (see Tables 3, 4 and 5).

The results for the probability of reaching the majors for players drafted in the top 5 and 10 rounds using the hurdle and zero inflated negative binomial models are reported in Tables 6 and 7 respectively. We note that the results are the same for the two types of models and thus we examine on the hurdles models in depth here. We see that relative age and age both have negative and significant relationships with the odds of making the majors. Over the top 5 rounds, the results suggest that a player one year older than another on draft day had 30.88% lower odds of reaching the major leagues ($p = 0.0038$). We also find that a player 100 days relatively older than another had 10.78% lower odds of playing in MLB ($p = 0.0326$). Additionally, we note that a .1 unit increase of $\frac{1}{\sqrt{pick}}$ (i.e. the 100^{th} pick to the 25^{th} pick) was correlated with a 70% increase in the odds of reaching MLB. Also, of note we see that the coefficient on pitcher flipped from negative in the count model to positive in the probability model, suggesting that even though pitchers play in fewer games at the major league level, they are more likely to reach the majors.

Again, we note that the main results do not change if we look at the top 10 rounds compared with the top 5 rounds, except that the coefficients on age and pitcher become more significant over the top 10 rounds ($p_{age} < .001, p_{pitcher} < .001$) (see Table 6 (ii)). Furthermore, the top 50 rounds results are identical to the top 10 rounds, suggesting that the results are robust to changes in the draft rounds considered (see Tables 6 and 8). In addition, the main results are robust to the inclusion of $lastDraft$ (see Tables 6 (iii) and 6 (iv)).

4.3 High School: WAR Results

We have already estimated the probability of reaching the majors and now that we have used games played as a measure of career success, we will also use WAR to estimate the relationships between age and relative age and MLB performance given that you have made the majors. We observe that the correlation of $lastDraft$ with WAR does not change the main results over the top 5, 10 and 50 rounds, thus we only examine the models not including $lastDraft$. Over the top 5 rounds of the draft (see Table 9 (i)) we note that a player 100 days relatively older than another produced 95.14% of the WAR of the younger player ($p = .106$). Furthermore, a player one year older than another on draft day produced 89.28% of the WAR of the younger player ($p = .141$). We note that a .1 unit increase of $\frac{1}{\sqrt{pick}}$ (i.e. the 100^{th} pick to the 25^{th} pick) was correlated with a 18.81% increase in WAR ($p < .001$). We observe that pitchers produced 92.89% of the WAR that position players produced ($p = .242$).

Over the top 10 rounds (see Table 9 (ii)) we see that a player 100 days relatively older than another produced 94.88% of the WAR of the younger player ($p = .0436$). In addition, a player one year older than another on draft day produced 93.56% of the WAR of the younger player ($p = .264$). Similar to our other results, we note that a .1 unit increase of $\frac{1}{\sqrt{pick}}$ (i.e. the 100^{th} pick to the 25^{th} pick) was correlated with a 17.57% increase in WAR ($p < .001$). We also see that pitchers produced 91.49% of the WAR that position players produced ($p = .100$).

Over the top 50 rounds (see Table 9 (iii)) we observe that a player 100 days relatively older than another produced 97.63% of the WAR of the younger player ($p = .149$). Additionally, a player one year older than another on draft day produced 95.80% of the WAR of the younger player ($p = .269$). Moreover, we observe that a .1 unit increase of $\frac{1}{\sqrt{pick}}$ (i.e. the 100^{th} pick to the 25^{th} pick) was correlated with a 16.75% increase in WAR ($p < .001$). We note that pitchers produced 91.95% of the WAR that position players produced ($p = .0163$).

4.4 High School: OPS Results (Position Players Only)

Thus far we have only examined career total statistics. Here we take a closer look at a career average statistic: on-base plus slugging percentage (OPS). We weight the observationss by the number of career at-bats. We note that the data are normally distributed (see Figure 7) so we may simply use OLS regression. Additionally, OPS is not a good measure of success for pitchers, thus we limit ourselves to only examining hitters. The results for these regressions are reported in Table 10. The main results for the regressions that include $lastDraft$ are the same as those which do not (see Table 10), thus we only discuss the regressions which do not include $lastDraft$. We also note that the results do not change much over the top 5, 10 and 50 rounds, thus we only comment on the top 5 round regression here. We observe that a player 100 days relatively older than another produced .00038 less OPS ($p = 0.9303$) during their careers. Additionally, we note that a player one year older than another on draft day produced .008411 less OPS ($p = 0.4646$) over their careers. We also observe that a .1 unit increase in $\frac{1}{\sqrt{pick}}$ (i.e. the 100^{th} to the 25^{th} pick) was correlated with a .0122 unit increase in OPS ($p < .001$).

4.5 College: Games Results

Using total games played as the response variable, we again find consistent results between two count regression models, hurdle with negative binomial distribution and zero inflated negative binomial (see Tables 11, 12 and 13). Given this finding, we shall only report here on the hurdle model results (see Tables 11, 12 and 13 for full results). We note that the relationship between $lastDraft$ and games played has no significant effect on the main regression coefficients. Thus, we only report here on those regressions not including $lastDraft$.

There is no significant relationship between relative age and games played. Interestingly, the coefficient on relative age is consistently positive (see Table 11). This is contrary to the results found for high school players, where the coefficients on relative age were consistently negative (see Table 3). In neither case were the coefficients significant; however it is of note that the coefficients have different signs based on *type* (i.e. High School vs. College) and that these signs are consistent throughout changes to the number of draft rounds considered. For college draftees selected in the top 5 rounds we see that a player 100 days relatively older than another played in .17% more games than the relatively younger player (see Table 11). For a 1000 game career, this difference is almost negligible at approximately 2 games.

We note that the coefficient on age is consistently negative, similar to our count regression results for high school players. Additionally, we see that there is no significant relationship between age and games played for the top 5 and 10 rounds. Over the top 5 rounds, a player one year older than another played in 99.48% of the games the younger player played. Over a 1000 game career for the younger player, this would result in approximately a 5 game difference, which is negligible. If we consider the top 10 rounds we see that this difference jumps to 89 games, which is fairly large, at over a half a season. In addition, we see borderline significance ($.01 < p < .05$) on age in all top 50 rounds regressions with a difference of 96 games over a 1000 game career (see Table 13).

Similar to the results for high school players, the coefficients on $\frac{1}{\sqrt{pick}}$ were consistently positive ($p < .001$). We note that for top 5 round draft choices, a .1 unit increase in $\frac{1}{\sqrt{pick}}$ (i.e. the 100^{th} pick to the 25^{th} pick) was correlated with a 13.72% increase in games played. This implies that if a player selected at 100^{th} overall played in 1000 games, a player selected at 25^{th} overall would be expected to play in 1137 games, nearly a full season more.

Additionally, we observe that being a pitcher was negatively ($p < .001$) correlated with games played. Considering top 5 round draft choices, pitchers played in 32.92% of the games of position players. Over a 162 game season (assuming the position player competed in every game) a pitcher would be expected to play in 53 games, which is again roughly the average number of games in which starting and relief pitchers appear. Given this result, it appears as though college pitchers on average play in more games than high school pitchers (see Tables 3 and 11). These results are consistent if we consider the top 10 and 50 rounds as well (see Tables 5 and 11). Furthermore, due to overdispersion, we observe that it was appropriate to use the negative binomial distribution as opposed to the poisson distribution, as the coefficient on log(theta) was significant on all regressions ($p < .001$) (see Tables 11, 12 and 13).

The results for the probability of reaching the majors for players drafted in the top 5 and 10 rounds using the hurdle and zero inflated negative binomial models are reported in Tables 14 and 15 respectively. The results are identical across the type of model used, thus we report here on the hurdle model results. Moreover, the results do not change with the inclusion of $lastDraft$, thus we only report on regressions not including $lastDraft$. Over the top 5 rounds we see negative coefficients and borderline significance for both the age and relative age variables. We note that a player one year older on draft day had 85.13% of the odds of reaching the major leagues of the younger player. We also observe that a player one year older on draft day had 91.95% of the odds of reaching the major leagues of the relatively younger player. Over the top 5 rounds we also note that a .1 unit increase in $\frac{1}{\sqrt{pick}}$ (i.e. the 100^{th} pick to the 25^{th} pick) is associated with a 132.88% increase in the odds of making the majors ($p < .001$). Further, we observe that pitchers had 72.02% of the odds of reaching the majors of a position player.

Over the top 10 rounds we see that the effects of age and pitcher change moderately, while the remainder of the results remain consistent. The magnitude of the coefficient on age increased, revealing that a player one year older on draft day had 82.17% of the odds of reaching the majors of the one year younger player ($p = .0015$). Additionally, pitcher is now only borderline significant ($p = .074$) with pitchers having 87.69% of the odds of reaching the majors of position players.

Over the top 50 rounds we again see an increase in the magnitude of the coefficient on age, as a player one year older than another on draft day had 70.87% ($p < .001$) the odds of the younger player of making the majors. Another interesting result we see over the top 50 rounds is that being drafted as a pitcher is significantly associated with an improvement of a draftee's odds of reaching the major leagues. We previously observed this result for high school players, but did not for college players over the top 5 and 10 rounds. For college players over the top 50 rounds we note that a pitcher was 13.52% more likely to reach the majors than a position player ($p = .014$).

4.6 College: WAR Results

Using WAR as the response variable for college players, we note that the main results do not change with the inclusion of $lastDraft$. Accordingly, we only consider regressions that do not include $lastDraft$. Additionally, the results are the same over the top 5 and 10 rounds and only differ over the top 50 rounds in an increased significance of *pitcher*. Thus, we only examine the top 5 rounds regression. Our regressions (see Table 17) show that a player 100 days relatively older than another produced 100.84% of the WAR of the younger player ($p = .688$). Furthermore, a player one year older than another on draft day produced 96.97% of the WAR of the younger player ($p = .490$). Additionally, we note that a .1 unit increase of $\frac{1}{\sqrt{pick}}$ (i.e. the 100^{th} pick to the 25^{th} pick) was correlated with a 12.98% increase in WAR ($p < .001$). We note that pitchers produced 90.12% of the WAR that position players produced ($p = .0196$). .

4.7 College: OPS Results (Position Players Only)

Using OPS weighted by at-bats as a measure of career success, we observe that the regression results do not change with inclusion of $lastDraft$ (see Table 18); thus, we only report on regressions that do not include $lastDraft$. There are also only small differences when examining the top 5, 10 and 50 rounds, so we only look at the top 5 rounds here. The regression results suggest that a player 100 days relatively older than another produced .01126 more OPS over their careers ($p = .00411$). The results show that a player one year older than another on draft day produced .00458 less OPS ($p = .61365$) over the course of their careers. We also find that a .1 unit increase in $\frac{1}{\sqrt{pick}}$ (i.e. the 100^{th} pick to the 25^{th} pick) was correlated with a .0105 unit increase in OPS ($p < .001$).

4.8 Other Results

We note that the proportion of high school and college draftees was pretty similar from the late 1980s until 2000, but then beginning in 2001 we begin to see a large increase in the proportion of college players selected, with 2003 and 2004 representing large discrepancies in this proportion. We note that this trend has continued to date, however the effect has dampened over time. Using breakpoint analysis, we find that the most likely break is 2000, suggesting that after this year the results were different than before, meaning significantly more college players were selected over high school players in the 2000s.

We looked into the possibility of changing relative age and age effects over time and found no consistent evidence corresponding to such effects. Some other results we looked into were whether certain teams were better at evaluating talent than others, whether certain regions produce better talent than others relative to their draft position and whether college or high school draft selections were better and we find no significant results across all of these tests. If we do not hold draft position constant, we note that *age* and *relativeage* are negative and borderline significant; we also find no difference in draft position for young and old (absolutely or relatively) players.

4.9 Summary of Main Results

1. More relatively old players are drafted by MLB teams.

2. More relatively old players play in MLB because more are drafted: the relatively young players reach MLB at significantly higher rate.

3. The results using Games, WAR, and OPS suggest there are no significant differences in MLB performance based on age or relative age for high school draftees.

4. Young and relatively young college draftees reach MLB at a higher rate.

5. Once in MLB there is no difference in performance as measured by Games and WAR for college draftees based on age or relative age.

6. There is evidence that the traditional relative age effect occurs at the MLB level with respect to OPS for college draftees.

7. College and high school players were selected at the same rates until the early 2000s, when college players began to be selected more often.

5 Discussion

We find that there exists strong evidence for the relative age effect in the MLB draft. We see the majority of draftees have birth days in first few months of the Little League year, which began on August 1 up until 2006. This result is consistent for high school as well as college draftees across rounds. Aside from being consistent throughout our results, the results follow prior literature on the topic, suggesting that the relative age effect exists in MLB (Thompson, Barnsley and Stebelsky, 1991).

With respect to our regression results, we note that for high school players there appears to be a significant relative age effect reversal (and for college players this result is borderline significant) on the odds of reaching the major leagues. It is possible that this is due to what has been described as the underdog effect, where the relatively young players have had to work harder their whole lives to be as good as the relatively older players and this propels the player to higher talent levels.

Another possible explanation is one more related to the relative age effect. Take for example a relatively young player who struggles to compete against his relatively older counterparts; he may drop out of playing ball at a young age, which is one of the explanations for the relative age effect. But take that same relatively young player and say he enjoys success at a young age playing against his relatively older peers. Now, this player will be unlikely to drop out. Given that he enjoyed success as a younger player it is likely that he is more naturally talented. Thus, he continues to play and is more talented, so he is more likely to make the majors.

Then, the problem arises as to why this does not continue at the major league level. As we observe from our games, WAR and OPS regressions there does not appear to be a relative age effect reversal at the major league level. This means that given that a player has reached the majors, he is about equally as talented as all other players. Why would this be the case? We will explore this question in short order.

Furthermore, not only do relatively young players make the majors more often, but absolutely young players as well. This result can be more easily explained. If a player is drafted as an 18 year old out of high school and begins his pro career in rookie ball–

which is a fairly regular occurrence–he is likely to be young for his level. If he enjoys success at this level, great, he can be promoted to A ball. Now, take for example the same 18 year old who struggles in rookie ball. Since he is young for his level he may be given an opportunity the next year to try to prove himself worthy by repeating the level–getting a second chance. If another player were 19 years old when drafted and placed in rookie ball and struggled, the team would be less willing to let the player have another opportunity and he may quickly move down the organizational depth chart at his position. Thus, a player who is younger when drafted may be given extra opportunities in the minors, enabling them to reach the majors more often.

Moreover, we again see that absolute age has no significant relationship with performance once a player has reached the majors. So, again we ask, why would this be the case? It seems to be that there is some talent level that a player must have to reach the majors, but then after the player has reached the majors everyone is fairly similar and differences in age and relative age do not make much of a difference anymore.

Additionally, with respect to the original Moneyball drafting philosophy of college over high school players, we note that during the Moneyball era (early 2000s), the proportion of high school to college draftees dropped considerably. However, it appears as though there was an over correction in the market–too many college players were selected–as the difference has come down some in more recent drafts.

6 Conclusion

We find that the relative age effect does indeed exist in the MLB Draft. Further, we can conclude that even though more relatively older players reach the major leagues this is only due to the larger proportion of relatively old players drafted. We see that a statistically significant proportion of those who are drafted and reach the majors are relatively young, suggesting evidence for the RAE reversal. However, given that a player reaches the majors we find no evidence that relatively old players perform differently than relatively young players. Similarly, we find that given that a player has been drafted, the younger the player is, the more likely he is to reach the majors. But, given that the player has reached the majors, young and old players appear to perform on similar levels. This is similar to other results that suggest that relative age effects disappear at the highest levels of athletics. Addona and Yates (2010) find that even though the RAE was present in the NHL across all positions, the RAE was not evident for hall of fame players. Here we note that the RAE is present for drafted players, there is an RAE reversal for the odds of reaching the majors, but no difference at the MLB level.

7 Future Research

Given our current work and previous research on these topics, we believe that there are a few key questions that remained unanswered and deserve further consideration. Namely, we would have liked to have a status variable to indicate whether a player was still active, to be able to take advantage of survival analysis regression techniques.

We do not have a variable that indicates whether a player was drafted as a college junior or senior, making it so that the *age* variable could get a little tricky, as we are comparing across classes. Older players in college could be college seniors, who are generally less talented than college juniors drafted, but we also control for being drafted again and there is no difference. So, if you are drafted late as a junior and return to school and then get selected earlier as a senior due to signability, this variation is captured by $lastDraft$. Since, the age coefficient is not significant anyway, this is not much of an issue.

Additionally, we would like to look into different transformations of WAR instead of the shift that we use. There are estimation issues with using simply WAR or shifting all the negative values to 0 or taking the logarithm of WAR, which were some of the transformations considered.

We would also like a variable to compare pitchers besides WAR. We found that

WHIP was promising, but we do not have a total career innings variable to account for the difference between pitchers who pitched for a long time compared to those who only got a “cup of coffee” in the majors. We also do not have a games started variable to compare starters and relievers.

Furthermore, we would like to use an ordered probit type of analysis to determine when exactly the relative age effect reversal occurs between draft day and reaching the majors. At present, we only observe that it occurs sometime between draft day and the MLB level. However, it may not truly be at the MLB level, as it could occur at AA, but we only observe it at the MLB level in our results. This is similar to how one would believe that the relative age effect is occurring at the MLB level if we did not have data on players drafted, where we clearly see the effect is much larger.

Finally, we would like to track the changes in the composition of draftees, as new Little League rules–as of 2006–have an age cutoff of May 1. We would expect over the next few years to see an influx of May, June and July draftees, compared to what we observe here, i.e. that the majority of players have birth days in August, September and October.

References

[1] Addona, V. and P. A. Yates. "A Closer Look at the Relative Age Effect in the National Hockey League." *Journal of Quantitative Analysis in Sports*, **6**, No.4, (2010).

[2] "Amateur Draft History." The Baseball Cube.

[3] Baker, J., and A. J. Logan. "Developmental Contexts and Sporting Success: Birth Date and Birthplace Effects in National Hockey League Draftees 2000-2005." *British Journal of Sports Medicine*, **41**, No.8, (2007): 515-17.

[4] Barnsley, R. H., A. H. Thompson, and P. E. Barnsley. "Hockey Success and Birthdate: The Relative Age Effect." *Canadian Association of Health, Physical Education and Recreation Journal*, **51**, (1985): 23-28.

[5] Baumler, G. "The Relative Age Effect in Soccer and Its Interaction with Chronological Age." *Sportonomics*, **6**, (2000): 25-30.

[6] Baxter-Jones, A. DG. "Growth and Development of Young Athletes. Should Competition Levels Be Age Related?" *Sports Medicine*, **20**, No.2, (1995): 59-64.

[7] Bedard, K., and E. Dhuey. "The Persistence of Early Childhood Maturity: International Evidence of Long-Run Age Effects." *The Quarterly Journal of Economics*, **121**, No.4, (2006): 1437-472.

[8] Billari, F. C., and M. Pellizzari. "The Younger, the Better? Relative Age Effects at University." *IZA Discussion Paper*, No. 3795, (2008).

[9] Black, S. E., P. J. Devereux, and K. G. Salvanes. "Too Young to Leave the Nest? The Effects of School Starting Age." *Review of Economics and Statistics*, **93**, No.2, (2011): 455-67.

[10] Bradbury, J. C. "Hot Stove Economics: Understanding Baseball's Second Season." New York: *Copernicus*, 2011.

[11] Burger, J. D. and S. JK Walters. "Uncertain Prospects: Rates of Return in the Baseball Draft." *Journal of Sports Economics*, (2009): 1-17.

[12] Crawford, C., L. Dearden, and C. Meghir. "When You Are Born Matters: The Impact of Date of Birth on Child Cognitive Outcomes in England." *The Institute for Fiscal Studies*, (2007).

[13] Cunha, F., and J. J. Heckman. "Investing in Our Young People." *NATIONAL BUREAU OF ECONOMIC RESEARCH*, (2010).

[14] Daniel, T. E. and C. T. Janssen. "More on the Relative Age Effect." *Canadian Association of Health, Physical Education, and Recreation Journal*, **53**, (1987): 21-24.

[15] Deaner, R. O., A. Lowen, and S. Cobley. "Born at the Wrong Time: Selection Bias in the NHL Draft." *PLOS One* **8**, No.2, (2013).

[16] Dudink, A. "Birth Date and Sporting Success." *Nature*, **368**, No.6472, (1994): 592.

[17] Fredriksson, P., and B. Ockert. "Is Early Learning Really More Productive? The Effect of School Starting Age on School and Labor Market Performance." *IZA Discussion Paper*, No. 1659, (2005).

[18] Gibbs, B. G., J. A. Jarvis, and M. J. Dufur. "The Rise of the Underdog? The Relative Age Effect Reversal among Canadian-born NHL Hockey Players: A Reply to Nolan and Howell." *International Review of the Sociology of Sport*, (2011): 1-6.

[19] Glamser, F. D. and L. M. Marciani. "The Birthdate Effect and College Athletic Participation: Some Comparisons." *Journal of Sport Behavior*, **15**, No.3, (1992): 227-38.

[20] Goodman, Alissa, and Barbara Sianesi. "Early Education and Children's Outcomes: How Long Do the Impacts Last?" *Fiscal Studies*, **26**, No.4, (2005): 513-48.

[21] Grondin, S., P. Deshaies, and L. P. Nault. “Trimestres De Naissance Et Participation Au Hockey Et Au Volleyball.” *La Revue Quebecoise De LActivite Physique*, **2**, (1984): 97-103.

[22] James, M. S. “Tables: Births and Deaths by Month, 1995-2002.” ABC News. *ABC News Network*, 15 Aug. 2005.

[23] Jazayerli, R. “Doctoring The Numbers: Starting Them Young, Part One.” *Baseball Prospectus*, 13 Oct. 2011.

[24] Jones, B. F. “Age and Great Invention.” *NBER Working Paper*, No. 11359, (2005).

[25] Knutson, D. “Does the Timing of School Affect How Children Learn. Earning and Learning: How Schools Matter.” *S. E. Mayer. Brookings Institution*, (1999): 79-102.

[26] Lewis, Michael. “Moneyball: The Art of Winning an Unfair Game.” New York: Ŵ.W. Norton, (2003).

[27] “MLB Draft History — MLB.com: History.” *Major League Baseball.*

[28] “MLB June Amateur Draft - Baseball-Reference.com.” *Baseball-Reference.com.*

[29] Russell, R. JH, and M. J. Startup. “Month of Birth and Academic Achievement.” *Personality and Individual Differences*, **7**, (1986): 839-46.

[30] Ryan, P. “The Relative Age Effect on Minor Sport Participation.” Diss. *McGill University*, 1989.

[31] Salaga, S. “Empirical Essays in Sport Management.” Diss. *University of Michigan*, 2012.

[32] Salthouse, T. A., and D. H. Schroeder. “Estimating Retest Effects in Longitudinal Assessments of Cognitive Functioning in Adults Between 18 and 60 Years of Age.” *Developmental Psychology*, **40**, No.5, (2004): 813-22.

[33] Sampford, M. R. “The Truncated Negative Binomial Distribution.” *Biometrika*, **42**, No.1/2, (1955): 58-69.

[34] Skirbekk, V., H-P Kohler, and A. Prskawetz. “Birth Month, School Graduation, and the Timing of Births and Marriages.” *Demography*, **41**, No.3, (2004): 547-68.

[35] Skirbekk, V. “Why Not Start Younger?: Implications of the Timing and Duration of Schooling for Fertility, Human Capital, Productivity, and Public Pensions.” *International Institute for Applied Systems Analysis*, (2005).

[36] Spurr, S. J. “The Baseball Draft: A Study of the Ability to Find Talent.” *Journal of Sports Economics*, **1**, No.1, (2000): 66-85.

[37] Stanaway, K. B. and T. M. Hines. “Lack of a Season of Birth Effect among American Athletes.” *Perceptual and Motor Skills*, **81**, No.3, (1995): 952-54.

[38] Thompson, A. H., R. H. Barnsley, and G. Stebelsky. “‘Born to Play Ball’: The Relative Age Effect and Major League Baseball.” *Sociology of Sport Journal*, **8**, (1991): 146-51.

[39] Wachter, D. “Investigating MLB Draft Outcomes, 2002-2005.” *Society for American Baseball Research*, (2012).

[40] Wattie, N., J. Baker, S. Cobley, and W. J. Montelpare. “Tracking Relative Age Effects over Time in Canadian NHL Players.” *International Journal of Sport Psychology*, **38**, (2007): 1-9.

[41] Williams, J. H. “Relative Age Effect in Youth Soccer: Analysis of the FIFA U17 World Cup Competition.” *Scandinavian Journal of Science & Medicine in Sports*, **20**, No.3, (2010): 502-08.

[42] Zeileis, A., C. Kleiber, and S. Jackman. “Regression Models for Count Data in R”. *Comprehensive R Archive Network-Project.*

8 Figures and Tables

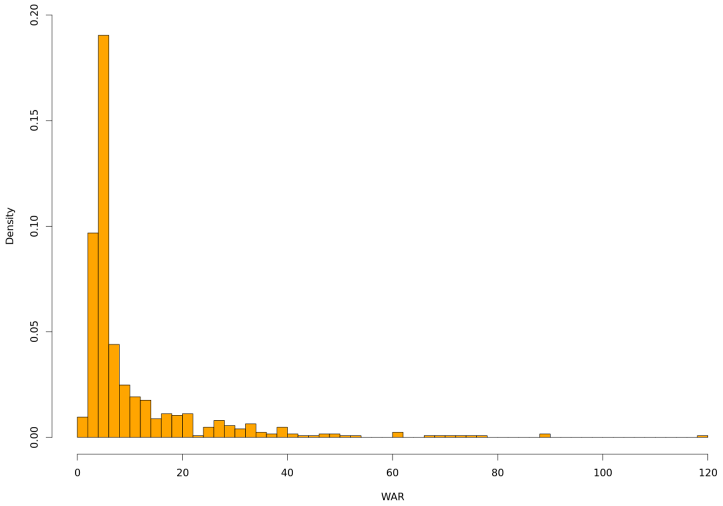

Figure 1: Histogram of WAR for players who played at least 1 game in the Majors and were drafted in the top 5 rounds out of HS

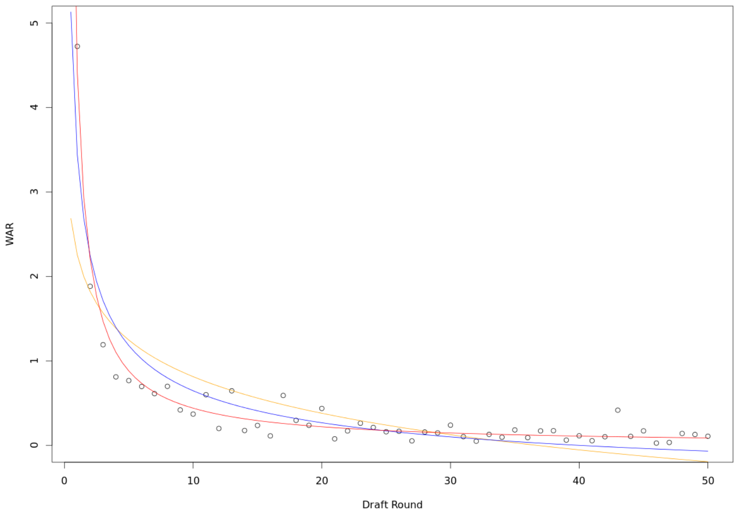

Figure 2: WAR by Draft Round, blue = $\frac{1}{\sqrt{pick}}$, orange = $\log(pick)$, red = $\frac{1}{pick}$

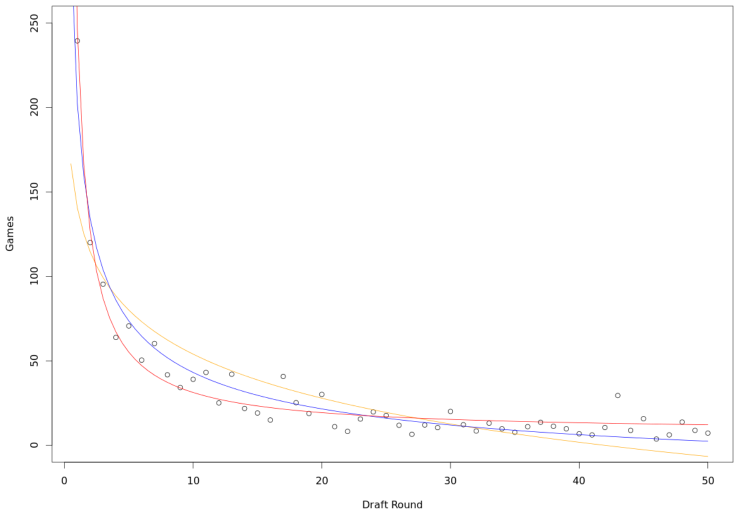

Figure 3: Games Played by Draft Round, blue = $\frac{1}{\sqrt{pick}}$, orange = $\log(pick)$, red = $\frac{1}{pick}$

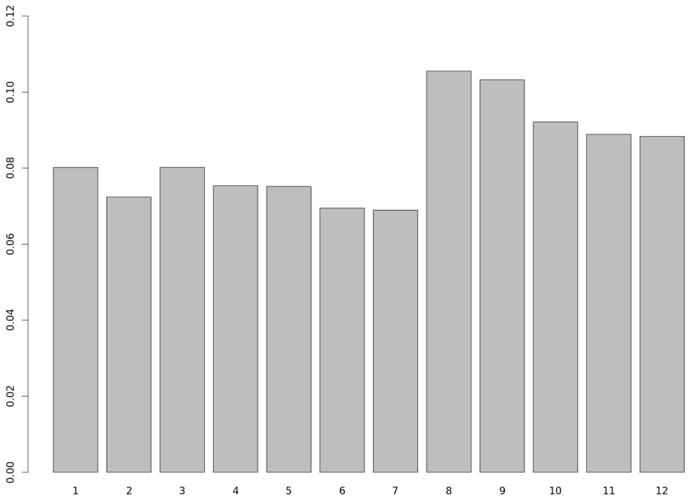

Figure 4: Histogram of Birth Months for top 50 round draftees from 1987-2011

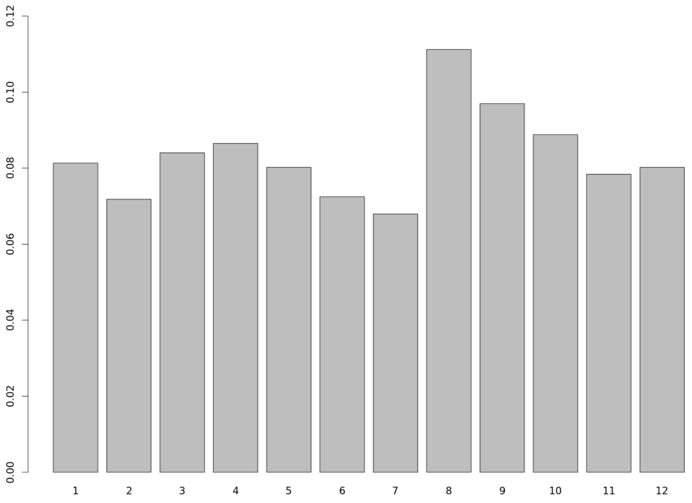

Figure 5: Histogram of Birth Months for top 50 round picks who reached the Major Leagues

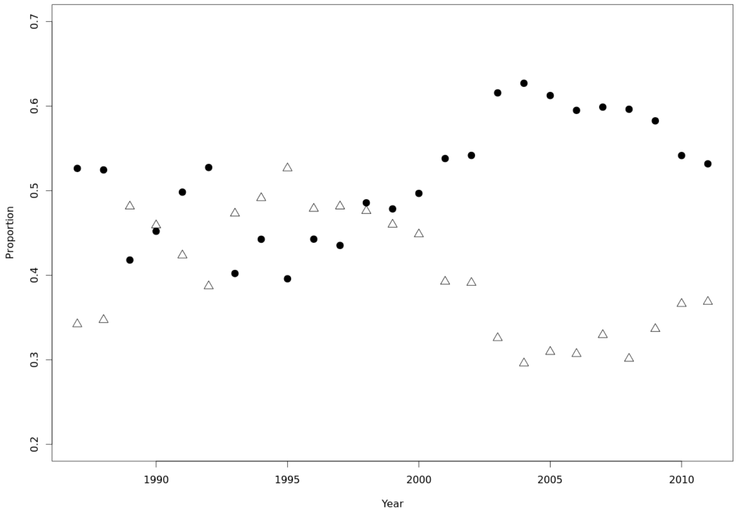

Figure 6: Proportion of High School (triangles) and College (solid circles) draft selections

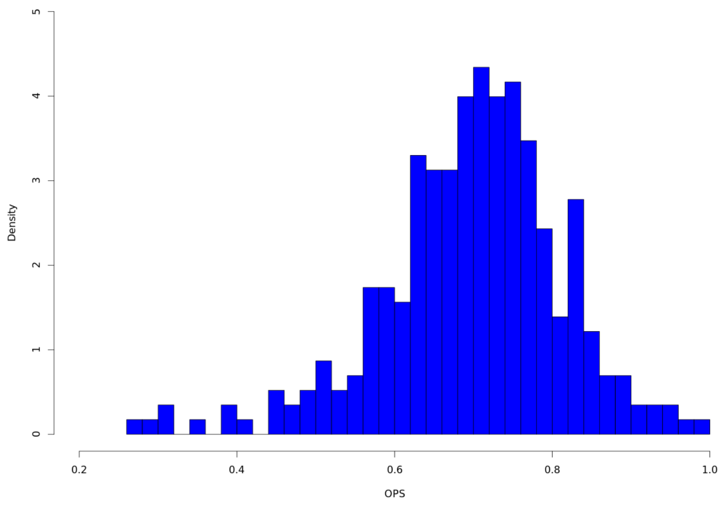

Figure 7: OPS for top 5 round draftees, selected as position players who played more than 20 games at the major league level

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
Age (in Years)	16.52	18.53	20.43	20.2	21.7	27.83	3327
Relative Age	1	107	204	196.9	291	365	3327
Pitcher	0	0	1	0.5042	1	1	–
$\frac{1}{\sqrt{pick}}$	0.0256	0.0306	0.0375	0.0520	0.053	1	29
lastDraft	0	1	1	0.8262	1	1	–
Games	0	0	0	31.56	0	2850	–
WAR	0.1	4.5	4.5	4.947	4.5	120	–
OPS	0	0.25	0.55	0.49	0.71	4	31932
Reach MLB	0	0	0	0.1248	0	1	0

Table 1: Summary Statistics for top 50 Round Draftees from 1987-2011

Month	Absolute Frequency	Relative Frequency	Theoretical Frequency (Uniform)	Theoretical Frequency (Observed U.S. Births)
January	2582	0.0802	0.085	0.0815
February	2332	0.0724	0.077	0.0761
March	2583	0.0802	0.085	0.0835
April	2427	0.0754	0.082	0.0801
May	2421	0.0752	0.085	0.0844
June	2238	0.0695	0.082	0.0830
July	2220	0.0689	0.085	0.0880
August	3398	0.106	0.085	0.0888
September	3324	0.103	0.082	0.0865
October	2967	0.0921	0.085	0.0851
November	2863	0.0889	0.082	0.0799
December	2845	0.0884	0.085	0.0831

Table 2: Birth Month Frequencies for Top 50 Round Draftees from 1987-2011

Variable	(i)	(ii)	(iii)	(iv)
Intercept	$7.408***$ (2.183)	$8.163***$ (1.911)	$7.317***$ (2.192)	$8.051***$ (1.918)
Relative Age	-0.000578 (0.000475)	-0.000670 (0.000423)	-0.000570 (0.000475)	-0.000674 (0.000422)
Age (in Years)	-0.0784 (0.118)	-0.117 (0.105)	-0.0774 (0.118)	-0.116 (0.105)
$\frac{1}{\sqrt{pick}}$	$1.749***$ (0.374)	$1.744***$ (0.354)	$1.724***$ (0.377)	$1.677***$ (0.358)
Pitcher	$-1.208***$ (0.100)	$-1.238***$ (0.0898)	$-1.209***$ (0.101)	$-1.245***$ (0.0900)
lastDraft	– –	– –	0.0800 (0.173)	0.135 (0.130)
Log(theta)	$-0.325***$ (0.0592)	$-0.388***$ (0.0529)	$-0.325***$ (0.0592)	$-0.386***$ (0.0529)

Table 3: High School Games Hurdle Count Models: (i) Top 5 rounds (ii) Top 10 rounds (iii) Top 5 rounds with lastDraft (iv) Top 10 rounds with lastDraft. Standard Errors reported in parentheses.

Variable	(i)	(ii)	(iii)	(iv)
Intercept	$7.541***$ (2.211)	$8.298***$ (1.952)	$7.446***$ (2.222)	$8.190***$ (1.961)
Relative Age	-0.000579 (0.000480)	-0.000689 (0.000433)	-0.000570 (0.000480)	-0.000696 (0.000432)
Age (in Years)	-0.0845 (0.120)	-0.122 (0.107)	-0.0839 (0.120)	-0.123 (0.107)
$\frac{1}{\sqrt{pick}}$	$1.648***$ (0.382)	$1.504***$ (0.350)	$1.613***$ (0.385)	$1.421***$ (0.352)
Pitcher	$-1.212***$ (0.102)	$-1.251***$ (0.0921)	$-1.214***$ (0.102)	$-1.260***$ (0.0924)
lastDraft	– –	– –	0.0929 (0.175)	0.156 (0.133)
Log(theta)	$-0.345***$ (0.0625172)	$-0.436***$ (0.0564109)	$-0.347***$ (0.0626689)	$-0.438***$ (0.0565554)

Table 4: High School Games Zero Inflated Negative Binomial Count Models: (i) Top 5 rounds (ii) Top 10 rounds (iii) Top 5 rounds with lastDraft (iv) Top 10 rounds with lastDraft. Standard Errors reported in parentheses.

Variable	(i)	(ii)	(iii)	(iv)
Intercept	$8.023***$ (1.369)	$8.020***$ (1.369)	$8.226***$ (1.429)	$8.389***$ (1.463)
Relative Age	-0.000266 (0.000315)	-0.000266 (0.000315)	-0.000333 (0.000329)	-0.000358 (0.000337)
Age (in Years)	-0.102 (0.0752)	-0.102 (0.0752)	-0.112 (0.0785)	-0.121 (0.0804)
$\frac{1}{\sqrt{pick}}$	$1.844***$ (0.316)	$1.855***$ (0.343)	$1.401***$ (0.305)	$1.217***$ (0.324)
Pitcher	$-1.111***$ (0.0673)	$-1.111***$ (0.0673)	$-1.142***$ (0.0706)	$-1.156***$ (0.0723)
lastDraft	– –	-0.00637 (0.0740)	– –	0.0345 (0.0790)
Log(theta)	$-0.517***$ (0.0401)	$-0.517***$ (0.0401)	$-0.622***$ (0.0463)	$-0.680***$ (0.0483)

Table 5: High School Top 50 Round Games Count Models: (i) Hurdle Model (ii) Hurdle Model with lastDraft (iii) Zero Inflated Negative Binomial Model (iv) Zero Inflated Negative Binomial Model with lastDraft. Standard Errors reported in parentheses.

Variable	(i)	(ii)	(iii)	(iv)
Intercept	$5.903*$ (2.339)	$5.366**$ (1.814)	$5.953*$ (2.342)	$5.515**$ (1.818)
Relative Age	$-0.00114*$ (0.000534)	$-0.000921*$ (0.000429)	$-0.00114*$ (0.000534)	$-0.000888*$ (0.000429)
Age (in Years)	$-0.369**$ (0.128)	$-0.361***$ (0.0988)	$-0.367**$ (0.128)	$-0.357***$ (0.0990)
$\frac{1}{\sqrt{pick}}$	$5.317***$ (0.643)	$7.261***$ (0.609)	$5.362***$ (0.648)	$7.455***$ (0.618)
Pitcher	$0.242*$ (0.107)	$0.318***$ (0.0886)	$0.241*$ (0.107)	$0.317***$ (0.0887)
lastDraft	– –	– –	-0.113 (0.183)	$-0.314*$ (0.130)

Table 6: High School Games Hurdle Binomial Models: (i) Top 5 rounds (ii) Top 10 rounds (iii) Top 5 rounds with lastDraft (iv) Top 10 rounds with lastDraft. Standard Errors reported in parentheses.

Variable	(i)	(ii)	(iii)	(iv)
Intercept	5.804*	5.233 * *	5.858*	5.395 * *
	(2.388)	(1.858)	(2.392)	(1.864)
Relative Age	−0.00117*	−0.000918*	−0.00117*	−0.000877*
	(0.000545)	(0.000441)	(0.000546)	(0.000442)
Age (in Years)	−0.367 * *	−0.361 * **	−0.364 * *	−0.355 * **
	(0.130)	(0.101)	(0.130)	(0.101)
$\frac{1}{\sqrt{pick}}$	5.917 * **	8.773 * **	6.005 * **	9.084 * **
	(0.843)	(0.795)	(0.859)	(0.811)
Pitcher	0.278*	0.368 * **	0.276*	0.368 * **
	(0.110)	(0.0915)	(0.110)	(0.0917)
lastDraft	–	–	−0.137	−0.361 * *
	–	–	(0.187)	(0.134)

Table 7: High School Games Zero Inflated Negative Binomial Probability Models: (i) Top 5 rounds (ii) Top 10 rounds (iii) Top 5 rounds with lastDraft (iv) Top 10 rounds with lastDraft. Standard Errors reported in parentheses.

Variable	(i)	(ii)	(iii)	(iv)
Intercept	$6.418***$ (1.173)	$6.226***$ (1.191)	$6.558***$ (1.212)	$6.368***$ (1.247)
Relative Age	-0.00040549 (0.000275)	-0.000307 (0.000278)	-0.000370 (0.000285)	-0.000236 (0.000292)
Age (in Years)	$-0.455***$ (0.0643)	$-0.425***$ (0.0653)	$-0.471***$ (0.0664)	$-0.445***$ (0.0684)
$\frac{1}{\sqrt{pick}}$	$12.222***$ (0.529)	$14.212***$ (0.581)	$15.66***$ (0.796)	$20.01***$ (0.945)
Pitcher	$0.380***$ (0.0585)	$0.351***$ (0.0590)	$0.453***$ (0.0610)	$0.444***$ (0.0625)
lastDraft	– –	$-0.783***$ (0.0633)	– –	$-0.947***$ (0.0686)

Table 8: High School Top 50 Round Games Probability Models: (i) Hurdle Model (ii) Hurdle Model with lastDraft (iii) Zero Inflated Negative Binomial Model (iv) Zero Inflated Negative Binomial Model with lastDraft. Standard Errors reported in parentheses.

Variable	(i)	(ii)	(iii)	(iv)	(v)	(vi)
Intercept	3.583*	2.806*	2.540 * **	3.625*	2.796*	2.506 * **
	(1.419)	(1.092)	(0.709)	(1.422)	(1.092)	(0.709)
Relative Age	−0.000498	−0.000526*	−0.00024	−0.000497	−0.00053*	−0.000234
	(0.000308)	(0.000261)	(0.000166)	(0.000308)	(0.000261)	(0.000166)
Age (in Years)	−0.113	−0.0666	−0.0429	−0.113	−0.0672	−0.0402
	(0.0771)	(0.0596)	(0.0388)	(0.0771)	(0.0597)	(0.0389)
$\frac{1}{\sqrt{pick}}$	1.723 * **	1.618 * **	1.548 * **	1.736 * **	1.606 * **	1.610 * **
	(0.217)	(0.197)	(0.153)	(0.219)	(0.200)	(0.164)
Pitcher	−0.0737	−0.0889	−0.0839*	−0.0729	−0.0894	−0.0833*
	(0.0631)	(0.0541)	(0.0349)	(0.0631)	(0.0541)	(0.0349)
lastDraft	–	–	–	−0.0482	0.0286	−0.0399
	–	–	–	(0.111)	(0.0801)	(0.0383)
Log(scale)	−0.270 * **	−0.288 * **	−0.384 * **	−0.270 * **	−0.288 * **	−0.385 * **
	(0.0283)	(0.0247)	(0.0178)	(0.0283)	(0.0247)	(0.0178)

Table 9: High School WAR Regressions: (i) Top 5 Rounds (ii) Top 10 Rounds (iii) Top 50 Rounds (iv) Top 5 Rounds (v) Top 10 Rounds (vi) Top 50 Rounds. Standard Errors reported in parentheses.

Variable	(i)	(ii)	(iii)	(iv)	(v)	(vi)
Intercept	$0.890***$ (0.210)	$0.910***$ (0.187)	$0.748***$ (0.132)	$0.927***$ (0.207)	$0.930***$ (0.186)	$0.746***$ (0.132)
Relative Age	−0.00000381 (0.0000435)	−0.0000539 (0.0000389)	−0.0000345 (0.0000293)	−0.0000189 (0.0000430)	−0.0000545 (0.0000387)	−0.0000344 (0.0000293)
Age (in Years)	−0.00841 (0.0115)	−0.00907 (0.0103)	−0.000747 (0.00725)	−0.00749 (0.0113)	−0.00874 (0.0103)	−0.000567 (0.00726)
$\frac{1}{\sqrt{pick}}$	$0.122***$ (0.0185)	$0.112***$ (0.0171)	$0.131***$ (0.0138)	$0.130***$ (0.0184)	$0.119***$ (0.0173)	$0.134***$ (0.0149)
lastDraft	– –	– –	– –	$-0.0545***$ (0.0159)	$-0.0307*$ (0.0128)	−0.00343 (0.00698)

Table 10: High School OPS Regressions Weighted by At-Bats: (i) Top 5 Rounds (ii) Top 10 Rounds (iii) Top 50 Rounds (iv) Top 5 Rounds with lastDraft (v) Top 10 Rounds with lastDraft (vi) Top 50 Rounds with lastDraft. Standard Errors reported in parentheses.

Variable	(i)	(ii)	(iii)	(iv)
Intercept	$6.392***$ (1.680)	$8.081***$ (1.303)	$6.485***$ (1.691)	$8.053***$ (1.304)
Relative Age	0.0000174 (0.000371)	0.000165 (0.000321)	0.0000189 (0.000371)	0.000165 (0.000321)
Age (in Years)	-0.00520 (0.0789)	-0.0936 (0.0608)	-0.00454 (0.0790)	-0.0951 (0.0609)
$\frac{1}{\sqrt{pick}}$	$1.286***$ (0.290)	$1.654***$ (0.290)	$1.287***$ (0.290)	$1.650***$ (0.290)
Pitcher	$-1.111***$ (0.0804)	$-1.082***$ (0.0691)	$-1.114***$ (0.0806)	$-1.080***$ (0.0694)
lastDraft	– –	– –	-0.109 (0.222)	0.0642 (0.162)
Log(theta)	$-0.348***$ (0.0487)	$-0.465***$ (0.04232)	$-0.347***$ (0.0487)	$-0.465***$ (0.0423)

Table 11: College Games Hurdle Count Models: (i) Top 5 rounds (ii) Top 10 rounds (iii) Top 5 rounds with lastDraft (iv) Top 10 rounds with lastDraft. Standard Errors reported in parentheses.

Variable	(i)	(ii)	(iii)	(iv)
Intercept	$6.377***$ (1.693)	$8.089***$ (1.330)	$6.469***$ (1.704)	$8.059***$ (1.333)
Relative Age	0.0000153 (0.000373)	0.000168 (0.000327)	0.0000166 (0.000373)	0.000168 (0.000327)
Age (in Years)	-0.00376 (0.0795)	-0.0929 (0.0621)	-0.00314 (0.0796)	-0.0943 (0.0622)
$\frac{1}{\sqrt{pick}}$	$1.201***$ (0.289)	$1.487***$ (0.286)	$1.202***$ (0.289)	$1.482***$ (0.286)
Pitcher	$-1.116***$ (0.0810)	$-1.092***$ (0.0705)	$-1.119***$ (0.0813)	$-1.090***$ (0.0708)
lastDraft	– –	– –	-0.107 (0.224)	0.0630 (0.166)
Log(theta)	$-0.363***$ (0.0499)	$-0.511***$ (0.0446)	$-0.363***$ (0.0499)	$-0.512***$ (0.0447)

Table 12: College Games Zero Inflated Negative Binomial Count Models: (i) Top 5 rounds (ii) Top 10 rounds (iii) Top 5 rounds with lastDraft (iv) Top 10 rounds with lastDraft. Standard Errors reported in parentheses.

Variable	(i)	(ii)	(iii)	(iv)
Intercept	$8.163***$ (0.956)	$8.186***$ (0.958)	$8.023***$ (1.011)	$8.026***$ (1.016)
Relative Age	0.0000683 (0.000261)	0.0000678 (0.000261)	0.0000634 (0.000274)	0.0000615 (0.000275)
Age (in Years)	$-0.101*$ (0.0442)	$-0.104*$ (0.0447)	$-0.0937*$ (0.0468)	$-0.0957*$ (0.0474)
$\frac{1}{\sqrt{pick}}$	$2.038***$ (0.282)	$2.019***$ (0.285)	$1.738***$ (0.280)	$1.712***$ (0.283)
Pitcher	$-1.118***$ (0.0579)	$-1.116***$ (0.0581)	$-1.142***$ (0.0609)	$-1.141***$ (0.0614)
lastDraft	– –	0.0409 (0.0972)	– –	0.0445 (0.103)
Log(theta)	$-0.580***$ (0.0354)	$-0.580***$ (0.0354)	$-0.706***$ (0.0400)	$-0.715***$ (0.0405)

Table 13: College Top 50 Round Games Count Models: (i) Hurdle Model (ii) Hurdle Model with lastDraft (iii) Zero Inflated Negative Binomial Model (iv) Zero Inflated Negative Binomial Model with lastDraft. Standard Errors reported in parentheses.

Variable	(i)	(ii)	(iii)	(iv)
Intercept	2.966	2.231	2.975	2.232
	(2.084)	(1.403)	(2.091)	(1.404)
Relative Age	-0.000838	-0.000461	-0.000838	-0.000470
	(0.000497)	(0.000365)	(0.000497)	(0.000365)
Age (in Years)	-0.156	$-0.171**$	-0.156	$-0.164*$
	(0.0966)	(0.0642)	(0.0967)	(0.0647)
$\frac{1}{\sqrt{pick}}$	$9.990***$	$14.95***$	$9.990***$	$15.06***$
	(1.069)	(1.093)	(1.070)	(1.104)
Pitcher	$-0.293**$	-0.0868	$-0.293**$	-0.0897
	(0.104)	(0.0767)	(0.104)	(0.0768)
lastDraft	–	–	-0.0190	-0.170
	–	–	(0.297)	(0.179)

Table 14: College Games Zero Inflated Negative Binomial Probability Models: (i) Top 5 rounds (ii) Top 10 rounds (iii) Top 5 rounds with lastDraft (iv) Top 10 rounds with lastDraft. Standard Errors reported in parentheses.

Variable	(i)	(ii)	(iii)	(iv)
Intercept	3.237	3.057$*$	3.238	3.062$*$
	(2.016)	(1.341)	(2.023)	(1.342)
Relative Age	-0.000840	-0.000420	-0.000840	-0.000425
	(0.000481)	(0.000349)	(0.000481)	(0.000349)
Age (in Years)	-0.161	$-0.196**$	-0.161	$-0.192**$
	(0.0937)	(0.0617)	(0.0938)	(0.0621)
$\frac{1}{\sqrt{pick}}$	$8.454***$	$11.593***$	$8.454***$	$11.629***$
	(0.782)	(0.727)	(0.782)	(0.730)
Pitcher	$-0.328**$	-0.131	$-0.328**$	-0.133
	(0.101)	(0.0735)	(0.101)	(0.0736)
lastDraft	–	–	-0.00118	-0.107
	–	–	(0.287)	(0.172)

Table 15: College Games Hurdle Probability Models: (i) Top 5 rounds (ii) Top 10 rounds (iii) Top 5 rounds with lastDraft (iv) Top 10 rounds with lastDraft. Standard Errors reported in parentheses.

Variable	(i)	(ii)	(iii)	(iv)
Intercept	$4.931***$	$4.434***$	$3.806***$	$3.064***$
	(0.747)	(0.758)	(0.797)	(0.814)
Relative Age	-0.0000955	-0.0000875	-0.000116	-0.000107
	(0.000242)	(0.000243)	(0.000256)	(0.000257)
Age (in Years)	$-0.344***$	$-0.307***$	$-0.311***$	$-0.258***$
	(0.0339)	(0.0351)	(0.0360)	(0.0374)
$\frac{1}{\sqrt{pick}}$	$20.056***$	$20.59***$	$27.34***$	$28.55***$
	(0.662)	(0.679)	(1.043)	(1.097)
Pitcher	$0.127*$	$0.119*$	$0.205***$	$0.195***$
	(0.0517)	(0.0518)	(0.0547)	(0.0550)
lastDraft	–	$-0.404***$	–	$-0.549***$
	–	(0.0871)	–	(0.0913)

Table 16: College Top 50 Round Games Probability Models: (i) Hurdle Model (ii) Hurdle Model with lastDraft (iii) Zero Inflated Negative Binomial Model (iv) Zero Inflated Negative Binomial Model with lastDraft. Standard Errors reported in parentheses.

Variable	(i)	(ii)	(iii)	(iv)	(v)	(vi)
Intercept	$2.45 * *$	$2.57 * **$	$2.35 * **$	$2.57 * *$	$2.580 * **$	$2.35 * **$
	(0.951)	(0.669)	(0.415)	(0.952)	(0.670)	(0.416)
Relative Age	0.0000836	0.0000967	0.0000385	0.0000837	0.000097	0.0000383
	(0.000208)	(0.000162)	(0.000116)	(0.000208)	(0.000162)	(0.000116)
Age (in Years)	-0.0308	-0.0388	-0.0292	-0.0275	-0.03806	-0.0287
	(0.0446)	(0.0312)	(0.0192)	(0.0446)	(0.0313)	(0.0194)
$\frac{1}{\sqrt{pick}}$	$1.22 * **$	$1.27 * **$	$1.29 * **$	$1.22 * **$	$1.271 * **$	$1.29 * **$
	(0.149)	(0.128)	(0.105)	(0.148)	(0.128)	(0.107)
Pitcher	$-0.104*$	$-0.0841*$	$-0.102 * **$	$-0.109*$	$-0.08518*$	$-0.102 * **$
	(0.0447)	(0.0344)	(0.0248)	(0.0448)	(0.03458)	(0.0249)
lastDraft	–	–	–	-0.201	-0.0276	-0.00628
	–	–	–	(0.125)	(0.0812)	(0.0422)
Log(scale)	$-0.393 * **$	$-0.458 * **$	$-0.548 * **$	$-0.395 * **$	$-0.458 * **$	$-0.548 * **$
	(0.0231)	(0.0191)	(0.0150)	(0.0231)	(0.0191)	(0.0150)

Table 17: College WAR Regressions: (i) Top 5 Rounds (ii) Top 10 Rounds (iii) Top 50 Rounds (iv) Top 5 Rounds with lastDraft (v) Top 10 Rounds with lastDraft (vi) Top 50 Rounds with lastDraft. Standard Errors reported in parentheses.

Variable	(i)	(ii)	(iii)	(iv)	(v)	(vi)
Intercept	0.818 * **	0.600 * **	0.734 * **	0.814 * **	0.580 * **	0.733 * **
	(0.190)	(0.133)	(0.0928)	(0.193)	(0.134)	(0.0929)
Relative Age	0.000113 * *	0.0000958 * *	0.0000521*	0.000112 * *	0.0000920 * *	0.0000523*
	(0.0000386)	(0.0000308)	(0.0000231)	(0.0000389)	(0.0000310)	(0.0000231)
Age (in Years)	−0.00458	0.00538	−0.000464	−0.00453	0.00548	−0.000620
	(0.00895)	(0.00626)	(0.00435)	(0.00896)	(0.00625)	(0.00437)
$\frac{1}{\sqrt{pick}}$	0.105 * **	0.123 * **	0.128 * **	0.106 * **	0.124 * **	0.127 * **
	(0.0218)	(0.0186)	(0.0154)	(0.0219)	(0.0186)	(0.0154)
lastDraft	–	–	–	0.00358	0.0211	0.00475
	–	–	–	(0.0263)	(0.0191)	(0.0101)

Table 18: College OPS Regressions Weighted by At-Bats: (i) Top 5 Rounds (ii) Top 10 Rounds (iii) Top 50 Rounds (iv) Top 5 Rounds with lastDraft (v) Top 10 Rounds with lastDraft (vi) Top 50 Rounds with lastDraft. Standard Errors reported in parentheses.