NHL Skater Development in the NCAA

Trevor Greissinger

Introduction

The modern NHL is unrecognizable from the league of even fifteen years ago, prior to the 2004-05 lockout season. Moving from the Dead Puck era, when scoring was at an all-time low, through the Clutch-and-Grab era, where the league's fastest and most skilled players were interfered with constantly, the game has changed into a competition of speed and skill. Teams rely upon their youngest players to score the most points and contribute the most to the team's success. As such, teams are becoming more and more interested in trying to find and exploit an edge in drafting and developing highly skilled players.

The NHL Entry Draft works differently than the drafts for the other major North American sports leagues. In the NFL and NBA drafts, players must declare themselves eligible for the draft, thereby forfeiting any remaining eligibility to play for an NCAA team. The MLB draft is more complicated, but there are also restrictions on a player's movements after being drafted. In the NHL, however, all players who will be 18 years old on or before September 15th of the draft year are automatically eligible to be drafted. After a team drafts a player, it owns the player's signing rights for the next three years, after which the player becomes an unrestricted free agent, able to sign with any team.

During the three years between being drafted and forfeiting signing rights, a player can choose to play wherever he wants (for a team that wants his services). For North American players, the traditional junior developmental leagues are the Canadian Hockey League (CHL) and NCAA hockey. The CHL is broken up into the Quebec Major Junior Hockey League (QMJHL), Ontario Hockey League (OHL), and Western Hockey League (WHL) and for a long time was the largest supplier of highly skilled players to the draft, especially the OHL and QMJHL. The NCAA has risen in popularity as a developmental league over the years. It had a reputation in the past as being the ideal league to develop the "tough guys" that used to populate every NHL team. Teams found that sending their grinders and enforcers to college for several years gave them a great opportunity to play against older, tougher competition and develop physically in college weightrooms.

In the past fifteen years, however, NCAA hockey is becoming a place where some of the game's brightest young stars are developing. Elite players like Michigan's Zach Werenski and Kyle Connor, Boston University's Jack Eichel and Brady Tkachuk, and Boston College's Johnny Gaudreau are some of the most exciting young players in the NHL today. However, none of these players stayed at school for the full four years of a college education. Connor, Eichel, and Tkachuk left after one year, Werenski after two, and Gaudreau after three.

It's tempting for a high-end player like one of these to sign his standard three year, $800,000/year Entry-Level Contract (ELC) when a team offers it to him. But something that is unknown is how much a player loses in development due to leaving the NCAA early.

In this paper, the effects of staying additional years in college on a drafted player's professional career will be studied. With data measuring drafted players' NCAA and NHL performance, bootstrapped linear regressions will be utilized to estimate how a player's performance in each year of NCAA eligibility predicts future NHL performance. I hypothesize that staying additional years will provide increased but diminishing returns in terms of NHL success. If the data do show this trend, it may help a player make a decision important to his future that has the best impact on them, their NCAA team, and the NHL team that drafted him; if a player really does improve over the course of his college career, it should be in all parties' best interests to pursue further development in the NCAA. If all three members of Michigan's vaunted CCM (Connor-Compher-Motte) line hadn't signed contracts after the 2015-16 season, it isn't unreasonable to say that Michigan would have been the favorite to win the national championship the following year (Coller). Hopefully this research can be used to convince players to stay in college for the benefits of their own development and the NCAA team's strength.

Data

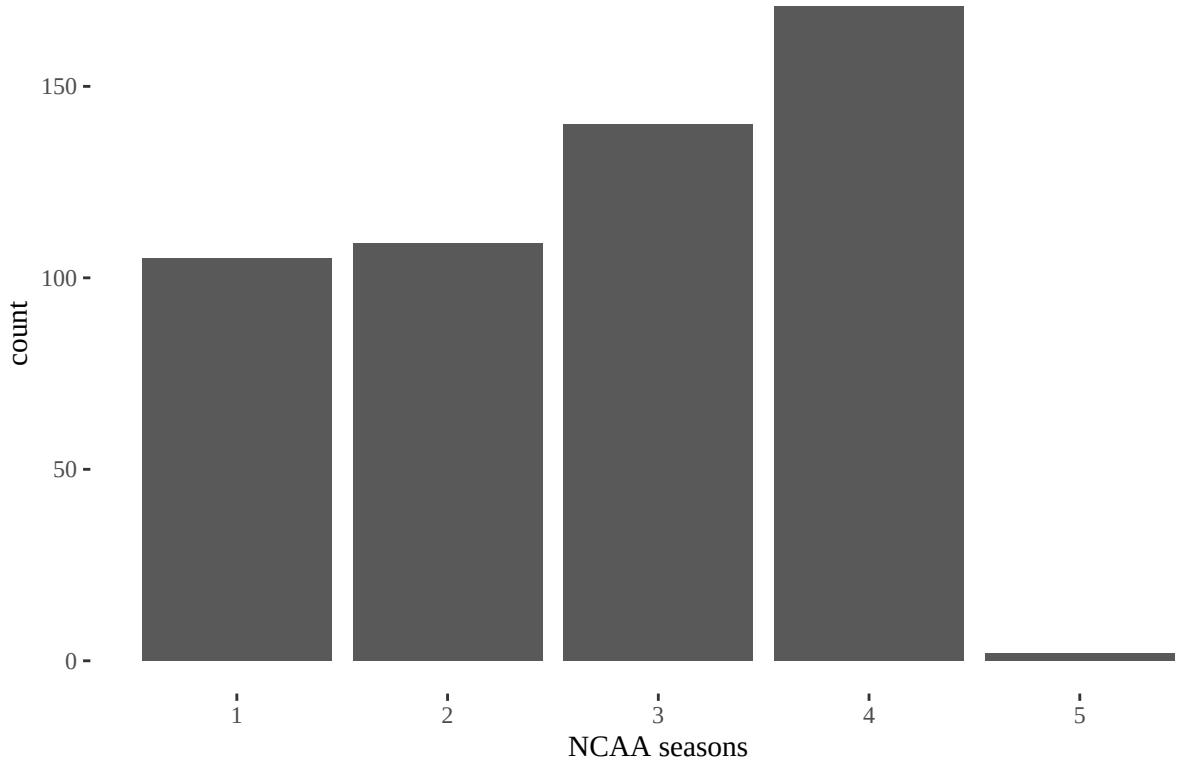
To narrow the focus of this question, only statistics for skaters (forwards and defensemen) will be considered, not goalies. Predicting goalie development is a much more difficult task because a goalie's performance numbers (typically goals against average (GAA) and save percentage) can fluctuate depending on the team he is on, and they aren't really indicative of a goalie's true talent level. There are methods to adjust GAA and save percentage for the performance of the goalie's teammates, but these are beyond the scope of this project.

Skater statistics were scraped from Elite Prospects, the most complete database of hockey statistics for a player's whole career, from junior to professional (*Elite Prospects*). Scraping was performed using functions from Evan Oppenheimer's `elite` R package which I modified to make the process more efficient (Oppenheimer).

Draft data for the ten drafts from 2008 to 2017 was also scraped from Elite Prospects. All players from all NCAA teams from the 2007-08 season to the 2016-17 season (13574 player-seasons in total) were also scraped from Elite Prospects. These lists of players were correlated to produce a set of data with 1437 player-seasons in it, comprising the college careers of 527 different drafted players.

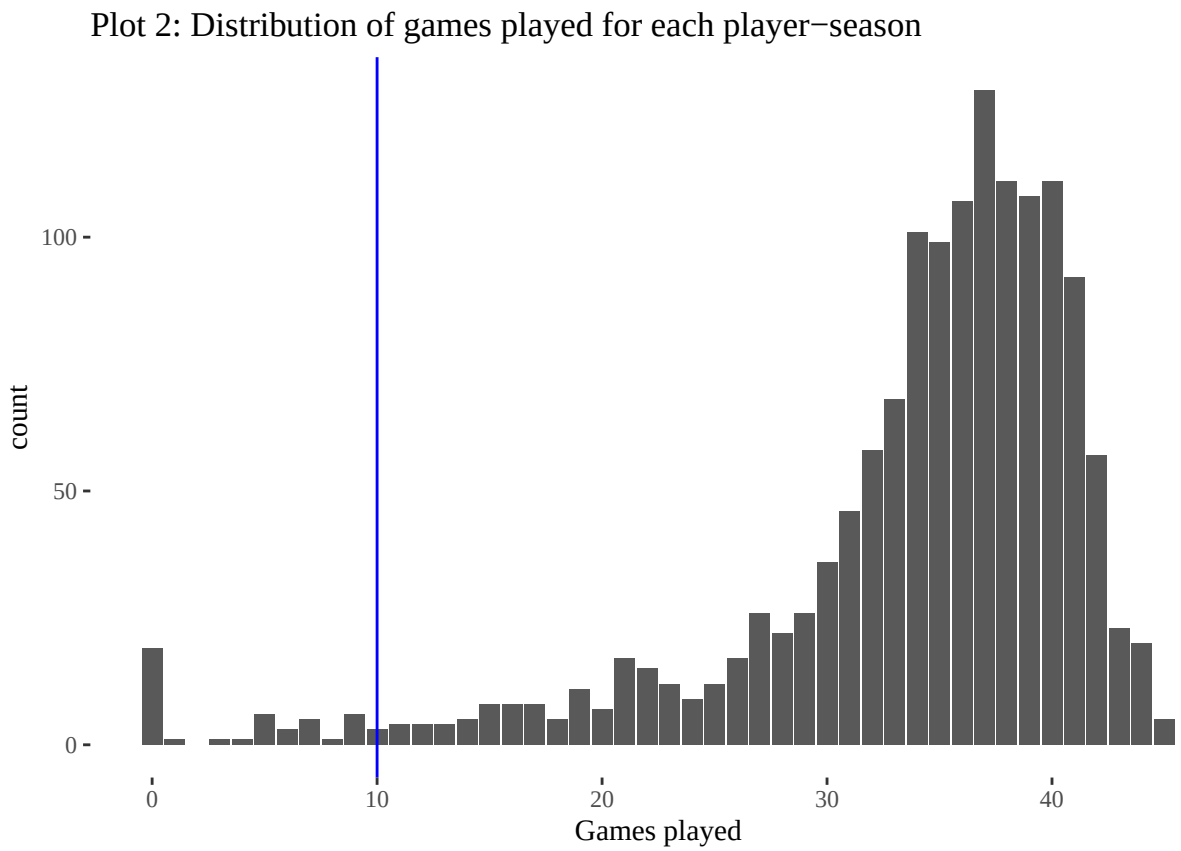
The distribution of lengths of college careers can be examined with a bar graph:

Plot 1: Bar graph of number of NCAA seasons for each player in data



There are two instances of players who played five NCAA seasons. Further inspection of the careers of these players shows that each had a year where he played three or fewer games. One case of this is due to injury, and the other case is due to redshirting. It is reasonable to categorize these players as having playing four full seasons.

There are also a number of instances where players played a small number of games in a season. Because there is no marker in the data for if a player missed most of a season to injury or redshirting, player-seasons with fewer than ten games played were removed. This brings the number of individual players to 501 and the number of player-seasons to 1394. Below is a bar graph showing the distribution of games played and the cutoff:

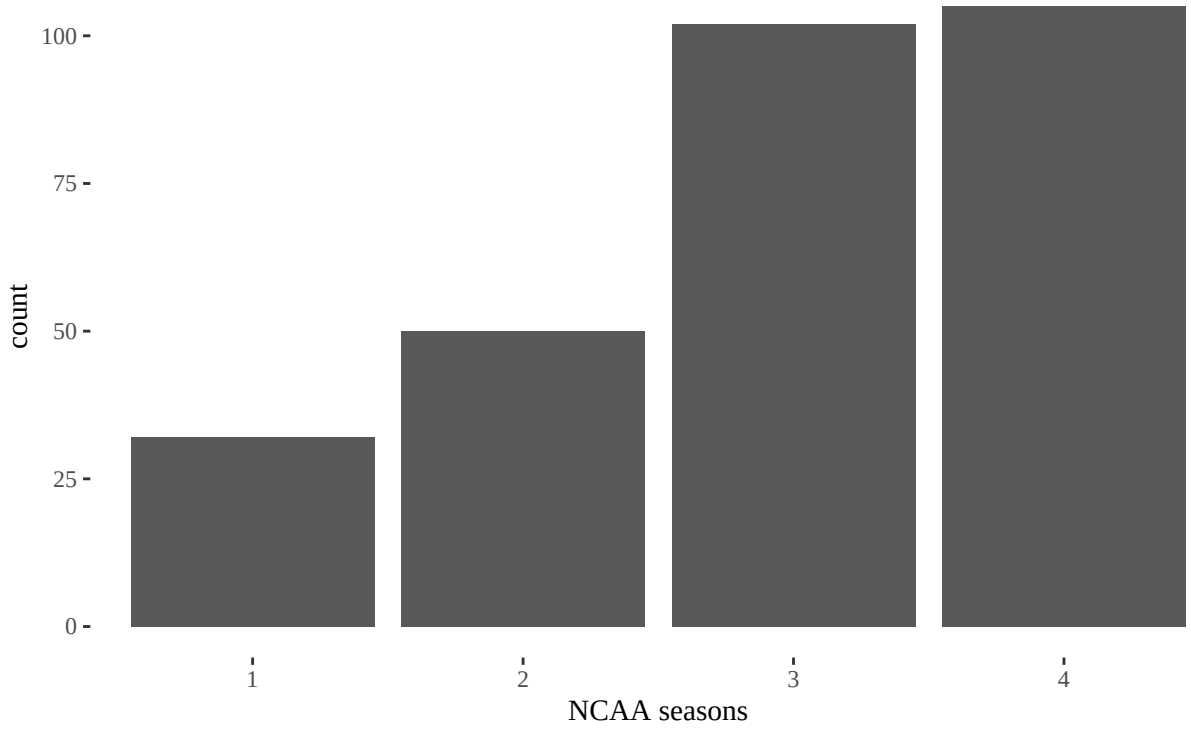


Professional data was scraped from Elite Prospects using the same `elite` package. For the purposes of this project, “professional" experience is defined as NHL or AHL experience, as very few players in the ECHL, the third tier North American professional league, ever move up to the NHL or AHL.

The players’ career professional statistics were aggregated by league (NHL or AHL). Only those players who played twenty or more games in either league will be considered for the following analyses. This was determined to be the optimal cutoff as it eliminates the largest outliers for point per game metrics.

It was found that, in the larger data set, there are 289 players with both NCAA and AHL or NHL experience. We can recreate Plot 1 to see the distribution of the lengths of these players’ NCAA careers:

Plot 3: Bar graph of number of NCAA seasons for players with professional experience



The number of cases for players with one or two seasons in the NCAA had a greater proportional decrease than the other seasons in this narrower data set, but bootstrapping should be effective in combatting the small sample size.

Method

The principal method that will be used in the following simulations is a non-parametric bootstrap from the `boot` package (Rizzo). This will be used to create many linear regressions of different NCAA metrics that may have an impact on professional performance in order to build a confidence interval for the regression coefficients.

The metric of professional performance that will be used is professional games played, rather than points or goals. While the statistics that measure a player's impact on his team are important from the team's perspective, from the player's perspective, more games played means more money for them. If these results were to be used by an NCAA coach trying to convince a player to stay at school for additional years, the prospect of higher future earnings would be a strong motivator. Games played will be normalized on a per-season basis to account for players with fewer seasons of experience by dividing the number of professional

games played by the number of professional seasons over which those games were played.

Also, unlike baseball, hockey has yet to develop a wins against replacement (WAR) statistic that is generally agreed upon. Emmanuel Perry, a data scientist who runs the advanced statistics site Corsica, developed a WAR metric using neural networks and machine learning that may be interesting to correlate with NCAA performance in future study. It is reasonable to assume that games played per season is an appropriate proxy for this sort of "total value" metric.

The bootstrap was introduced in 1979 by Bradley Efron of Stanford University (Rizzo). The non-parametric bootstrap used in this analysis is a Monte Carlo method that uses our relatively small sample to estimate the population distribution from which it is drawn. Test statistics generated from the bootstrap samples are used to estimate a sampling distribution of the population parameters. The bootstrap can also be used to reduce bias in the sample distribution. Ordinary least squares (OLS) regression coefficients are by definition unbiased, but if we were to study another statistic like variance, the bootstrap would narrow the confidence interval estimating population variance.

The bootstrap method is an appropriate choice for analysis here because of the relatively small sample size in our data (289 players). By sampling with replacement from the data many times and performing regressions on the bootstrapped samples, we can approximate the true regression coefficients with greater confidence than if we performed one regression on the whole data set.

In order to use the bootstrap, we must make certain assumptions about the population and sample. For one, we must assume that the 289 players in our sample are representative of the population of NHL draft picks who have played NCAA hockey. This assumption holds true because our sample is all the players since 2007 who have played college hockey and have been drafted. This also ensures that our sample is an unbiased representation of the population.

Additionally, in order to perform linear regression, we need to make assumptions about our data. There must be a linear relationship between our response variable and the predictor variables. As shown below, scatter plots and Q-Q plots can be used to verify the assumption that the data's residuals are normally distributed with distribution $N(0, \sigma^2)$.

Simulations

Model selection

To properly simulate the data, first we must determine the variables of interest in the data set. Using all the NCAA per game statistics as predictors, along with position and number of seasons, and using professional

games played per season (`gp.per.season`) as the response, all possible models were fit with the `regsubsets()` function from the `leaps` package using exhaustive search. From this and evaluating the models with BIC, it was determined that the best model took into account player position, number of seasons, NCAA games played, and NCAA points per game. Interestingly though, it did not seem like any interaction terms in the model were appropriate. This is suprising because forwards score at a higher rate than defensemen in all levels of hockey.

Table 1: NCAA points per game by position

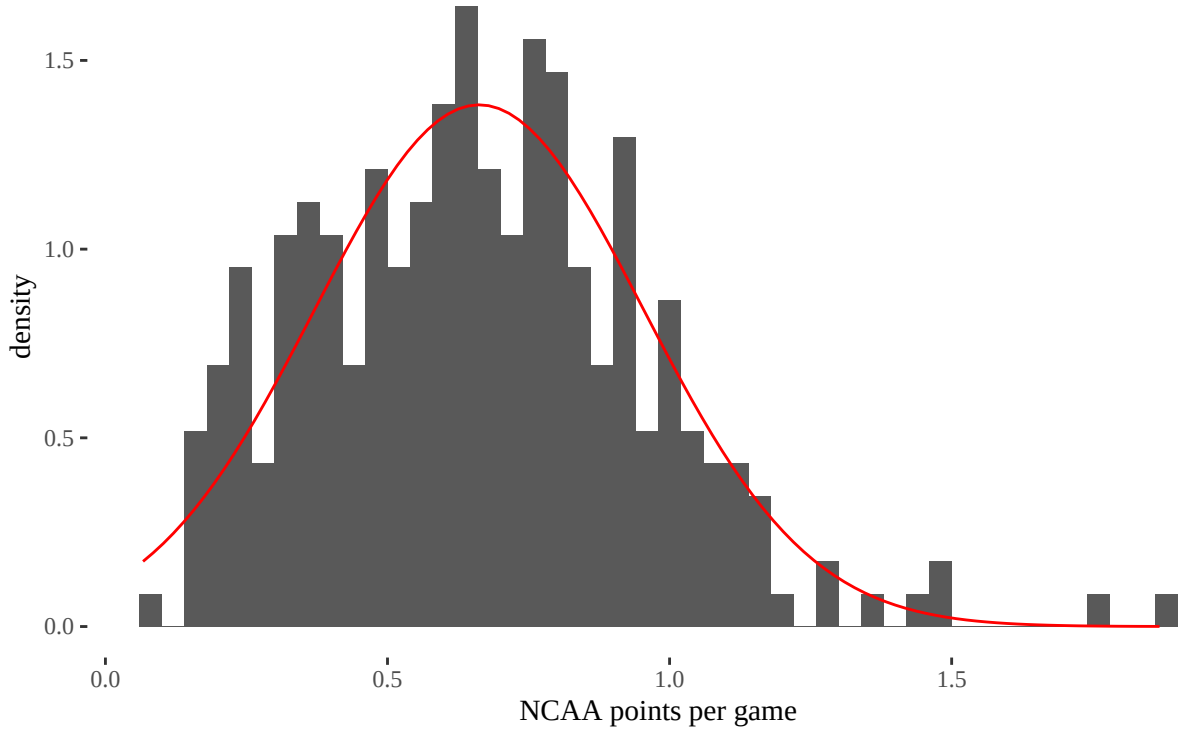
position	mean points per game	standard deviation
F	0.804	0.264
D	0.460	0.196

Regardless, the model that will be used from this point is given by the formula `gp.per.season ~ position + number of NCAA seasons + NCAA points per game`. NCAA games played will not be included because of a high degree of collinearity with NCAA seasons played. It also introduces some error due to players who missed the majority of one season due to injury or ineligiility.

Simulation parameters

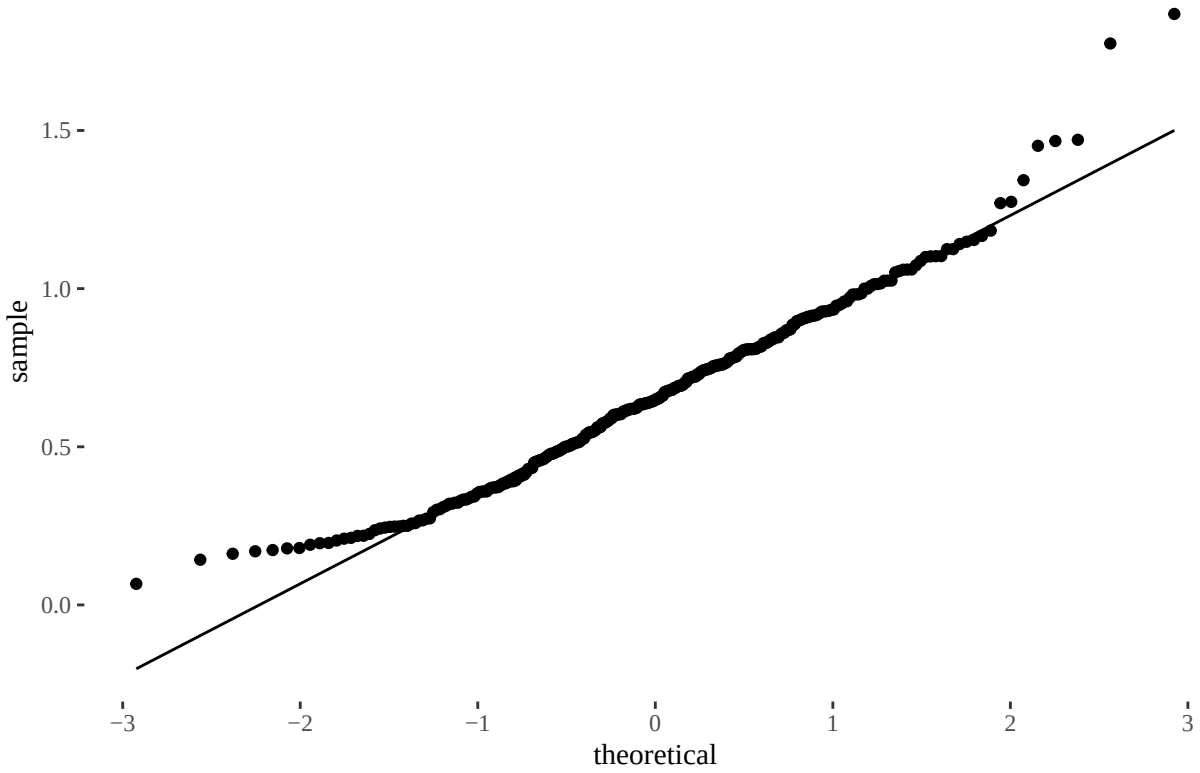
In order to simulate the methods listed above, we must generate random data that fits the distribution of the statistic of choice in the data set. Examining the density of NCAA points per game with the following histogram, it seems like that statistic can be reasonably modeled by a truncated normal distribution with the same mean and variance as the distribution of the statistic itself and a minimum value of 0.

Plot 4: Histogram of NCAA points per game and comparable truncated normal distribution



This normality can also be verified using a Quantile-Quantile (Q-Q) plot:

Plot 5: Quantile−quantile plot of NCAA points per game



While there is some deviation at the extreme quantiles, the general linearity shows that the assumption of normality is valid.

Despite the model selection not showing that an interaction between position and NCAA points per game is significant, it will still improve the accuracy of the simulation to sample data for each position based on the distribution in the larger data set. Using the parameters from Table 1, truncated normal distributions for forwards' and defensemen's NCAA points per game were created that matched the sample data.

Table 2: Professional games played per season by years of NCAA experience

ncaa.seasons	mean pro games played per season	standard deviation
1	37.180	15.707
2	32.845	15.226
3	34.928	14.252
4	33.893	14.070

These distributions were combined with all effects accounted for in order to create a simulated data set for 1000 player with a measure of professional games played per season under the null hypothesis that additional NCAA seasons have no impact on future games played, as well as under the alternative hypothesis that there is an impact on future games played. These were calculated with a baseline of 36 games per season for those players who only played one year of NCAA hockey, increased by roughly 0.25 standard deviations, or four games, per additional season played.

Simulations

First, bootstrapped confidence intervals were created for the data generated under the null hypothesis using the `boot()` and `boot.ci()` functions from the `boot` package. The 95% confidence intervals for the coefficients appear in the table below.

Table 3: 95% bootstrapped confidence intervals for bootstrapped null regression coefficients

	2.5 %	97.5%
(Intercept)	34.81	39.97
positionF	-2.40	1.91
ncaa.ppg	-5.32	2.63
ncaa.seasons2	-4.66	0.17
ncaa.seasons3	-2.48	2.44
ncaa.seasons4	-4.05	0.78

Next, the same procedure was applied to the data simulated under the null hypothesis.

Table 4: 95% bootstrapped confidence intervals for bootstrapped alternative regression coefficients

	2.5 %	97.5 %
(Intercept)	32.35	37.89
positionF	-1.10	3.57
ncaa.ppg	-3.72	4.67
ncaa.seasons2	1.76	6.92

	2.5 %	97.5 %
ncaa.seasons3	7.13	12.16
ncaa.seasons4	7.92	13.05

The confidence intervals for the non-intercept coefficients generated for data simulated under the null hypothesis all contain zero, meaning at a 95% confidence level, there is no correlation between each predictor and the response. On the other hand, the confdience intervals for the bootstrapped regressions under the alternative hypothesis show a positive correlation between number of NCAA seasons and professional games played at a 95% confidence level.

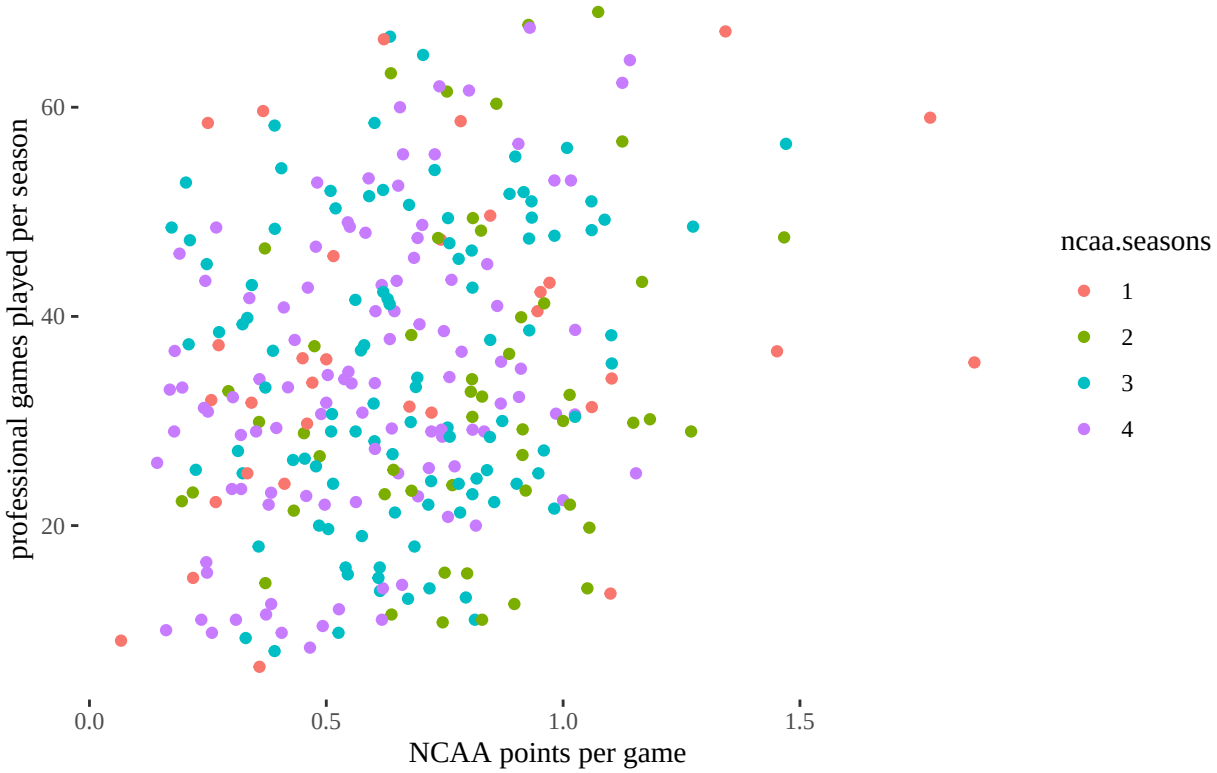
In comparison with the confidence intervals generated from the simple linear regression on the null data, the bootstrapped confidence intervals are narrower. This is expected because of the role bootstrapping plays in minimizing the variance of calculated statistics.

Table 5: 95% confidence intervals for simple linear regression coefficients (null data)

	2.5 %	97.5 %
(Intercept)	32.57	37.92
positionF	-2.53	2.06
ncaa.ppg	-2.90	4.95
ncaa.seasons2	-2.20	2.82
ncaa.seasons3	-2.43	2.59
ncaa.seasons4	-4.02	1.00

An initial proposal for this project included the idea of using k-means clustering to find clusters corresponding to number of NCAA seasons on a graph of NCAA points per game vs. professional games played. The below graph shows that the effects between NCAA seasons and these other variables are too complicated to be sorted into clusters. For this reason, cluster analysis is not appropriate for these data.

Plot 6: Cluster analysis for regression formula



Analysis

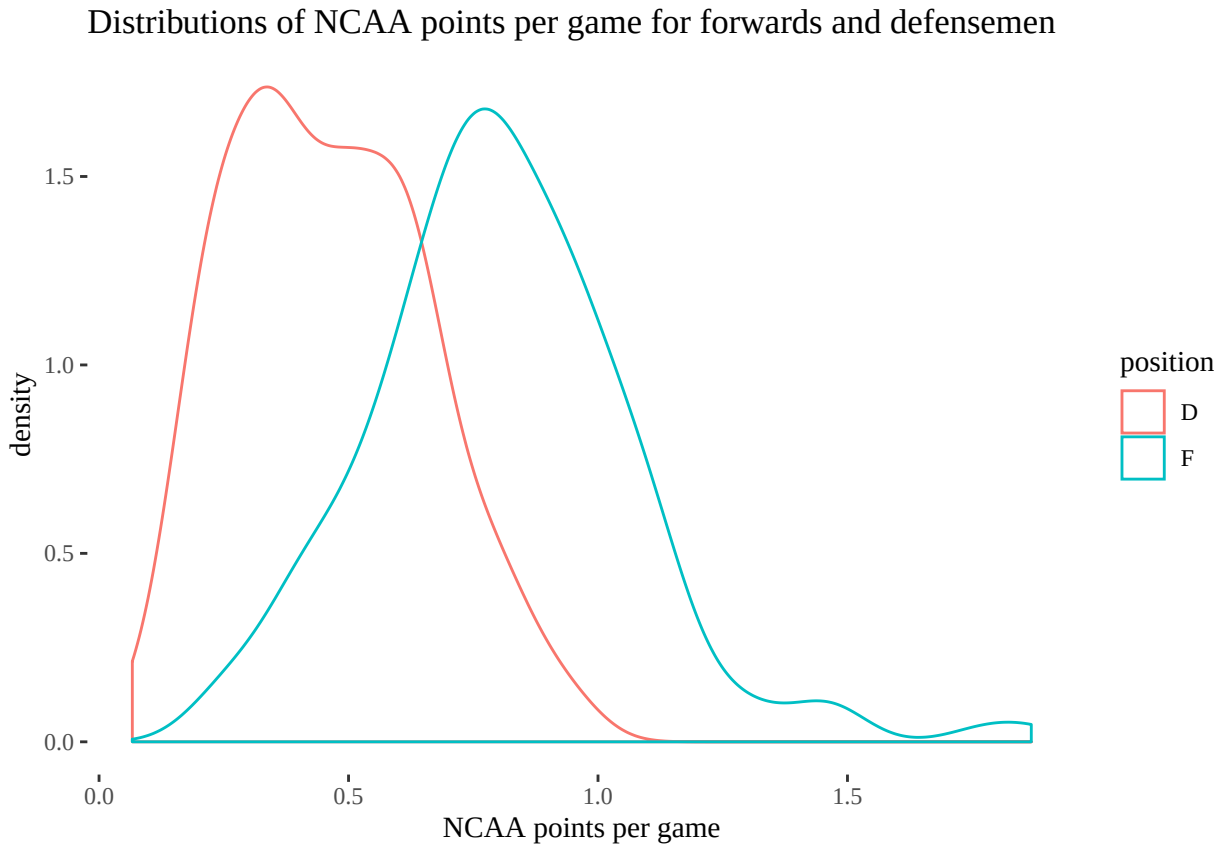
The bootstrap procedure used on the simulated data above is now applied to the real data set to estimate the regression coefficients and their 95% confidence intervals. As before, 1000 bootstrap samples were generated and used to fit linear regression models. The distributions of these models' coefficients were used to construct the confidence intervals with the `boot.ci()` function.

Table 6: Bootstrapped regression coefficients with 95% confidence intervals

coefficient	estimate	2.5 %	97.5 %
(Intercept)	26.924	20.680	33.340
positionF	-6.613	-10.673	-2.453
ncaa.ppg	20.168	12.800	27.070
ncaa.seasons2	-5.150	-11.443	1.230
ncaa.seasons3	-1.900	-7.388	3.666

coefficient	estimate	2.5 %	97.5 %
ncaa.seasons4	-1.076	-6.674	4.656

Despite the model selection procedure showing that an interaction term between position and NCAA points per game was not appropriate, the data from Table 1 and the graph below show that the distributions for point production from forwards and defensemen are quite different.



Thus, the bootstrap procedure was used on separate data sets for forwards and defenseman to see if a positional effect lingered.

Table 7: Regression coefficients separated by position

coefficients	forwards	defense
(Intercept)	19.476	27.701
ncaa.ppg	20.012	21.618
ncaa.seasons2	-5.141	-3.981
ncaa.seasons3	-1.771	-2.442

coefficients	forwards	defense
ncaa.seasons4	1.526	-4.654

Clearly, there is a huge difference in the intercept term for regressions performed on the positional data. However,the tradeoffs for discovering this positional effect are much wider confidence intervals for each coefficient; only at a low confidence level can we reject the hypothesis that there is no distinction between forwards and defensemen in the regression.

The regression shows a significant positive correlation between NCAA point production and professional games played per season. This result is expected because it's extremely rare for NCAA players, whether forwards or defensemen, to score at low rates in the NCAA and then become valuable to a professional team.

There is also a significant negative effect associated with forwards, who appear to play between six and seven fewer professional games per season than defensemen. This result makes sense in the context of hockey coaching. Typically, a hockey roster has twelve forwards and six defensemen. A defenseman with the same value added to the team by his play on the ice as a forward therefore has more intrinsic value from a roster-management perspective. Whether or not this hypothesis bears out in the data is inconclusive, but could be an interesting topic for future study.

There is not sufficient evidence to conclude that additional seasons spent in college hockey have any impact on future games played per season. Using one year (the minimum) in college as the baseline for the regression, further seasons played all have negative coefficient estimates with confidence intervals containing zero (see Table 6). It is interesting that the regression estimates are not monotonic with increasing years in the NCAA–there is more of a negative impact for the second year than for any other year. This could be an indication of systemic bias in the sample (see "Problems" in the discussion).

Discussion

Results and impact

The results of the study do not match the initial hypothesis that additional years in college hockey have a positive impact on a player's professional career. Even when controlling for the player's position and point production at the NCAA level, there was no significant result that showed this hypothesis was true.

This result is disappointing and surprising. A number of problems listed below provide possible explanations for this result if the initial hypothesis is actually true. They are also reasons why this should not be evidence that leaving the NCAA early has a *positive* impact on future careers. There is neither a significant correlation

nor a significant causation between the two factors, at least as far as the scope of this project.

Problems

Data limitations were one of the chief reasons I believe a significant conclusion was not found. Bootstrapping can only ameliorate a small sample size so much, and 289 players, despite being a large proportion of the population, is still a very small sample. Only so many NCAA players matriculate into the professional leagues every year. That number is continuing to grow, but the total number of players in this modern era is still small compared to the number of players coming from other developmental leagues. Additionally, the publicly available statistics for NCAA players are quite limited when compared to the NHL statistics available. Ideally, metrics like time on ice could be used to better evaluate each players individual contributions and create less biased statistics (i.e. points per 60 minutes played) than points per game.

There's also a large deal of selection bias in these data. One of the assumptions necessary for this analysis was that the choice of a player to leave NCAA hockey early was independent of factors like his performance and how his draft team valued him. Upon further reflection, this assumption doesn't really hold up. Regardless of performance in the season prior, players who were drafted higher and therefore valued more by their draft team will be under greater pressure to sign than if they were a lower-end prospect.

The chosen response variable of points per game played was a necessary proxy to use in the absence of a comprehensive statistic like wins above replacement (WAR). At a very high-level, it shows which players were used by their coaches and which players weren't. But it fails to take into account if a player missed games due to injury, which would skew the results. It also doesn't take into account the length of games possible for a player to play in a season. It's common for players to sign their entry-level contracts after the NCAA season finishes around April and then play for an NHL or AHL team for the remainder of its season. This once again highlights the need for a metric that evaluates a player's total performance at a more granular level (Ventura and Thomas, Perry).

Future study

As stated above, the biggest limitation in this research is the data available. The small size of the NHL (31 teams with 21 players on a roster) means that dividing the players into groups makes for an even smaller sample. This research question likely doesn't have an attainable answer without many more years of data and better statistics for measuring player value (see above).

References

Coller, James. *JT Compher signs entry-level contract with Colorado Avalanche.* Apr. 2016, https://www.michigandaily.com/section/ice-hockey/jt-compher-signs-entry-level-contract-colorado-avalanche.

Elite Prospects. 2018, https://www.eliteprospects.com.

Oppenheimer, Evan. *elite: A Package for Scraping Hockey Data from EliteProspects.* 2018, https://github.com/eoppe1022/elite.

Perry, Emmanuel. *The Art of WAR.* 2017, http://www.corsica.hockey/blog/2017/05/20/the-art-of-war/.

Rizzo, Maria L. *Statistical Computing with R.* Chapman & Hall/CRC, 2008.

Ventura, Sam, and A.C. Thomas. *The Road to WAR (for Hockey).* 2014, http://blog.war-on-ice.com/index.html{\%}3Fp=37.html.