NHL Project Progress Report

4/19/2021

Client: Sam Ventura

Minyue Fan, Steve Kim, Kexiong Shen, Linda Yang

Advisor: Brian MacDonald

Carnegie Mellon University

Agenda

Topic	Slides No.
Overview	3-5
Data	6-9
Method	10-14
Results	15-20
Next Steps & Roadblocks	21
Q&A	22

Introduction: Overview

- There are multiple traditional paths hockey prospects can take to get to the NHL:
  - USHL -> NCAA -> NHL
  - USHL -> (NCAA) -> AHL -> NHL
  - International -> KHL -> NHL
  - Other defined paths
- Most players do not immediately go to the NHL when they are eligible (drafted or not). They stay in or move to some “development leagues” before entering the NHL.
  - Draft eligibility (North America): Players must be 18 years old by 15 September or under 20 years old by 31 December in the year of the draft.
  - Development leagues: USHL, NCAA, etc.

Introduction: Overview

- People have very strong opinions about how players' development paths impact their future in the NHL.
    - Typically, American players who take the NCAA path have higher success rates (e.g. 20% make the NHL, compared to 5% from the USHL path)
    - However, the NCAA player pool are already better in terms of quality. Better players are getting their opportunities in the NCAA.
    - Is there causal impact of taking the NCAA path?

Introduction: Research Question

- Questions: Does taking different development paths matter? How do players' development paths impact their performance and success in the NHL?
- The understanding in the scouting community is that development path does matter.
  - Only anecdotal.
  - We intend to establish grounding on this thought.

Data

- Two datasets:
  - Leagues: NHL, NCAA, USHL and AHL
  - Time period: 2001 - 2020
    - *contains some data earlier than 2001*
  - Players' biographical information
  - Players' performance data each season
    - *box score statistics*

Data Description

- Biographical information:
  - 15786 players

	Player	Position	DateofBirth	Height	Weight	Nation	Shoots
1	Scott May	C	Jan 08, 1982	5'10" / 178 cm	187 lbs / 85 kg	Canada	R
2	Kent Gillings	F	Jun 14, 1979	5'10" / 177 cm	194 lbs / 88 kg	Canada / Ireland	R
3	Tyler Kindle	D	Feb 20, 1978	5'8" / 173 cm	165 lbs / 75 kg	USA	L
4	D'Arcy McConvey	C	Oct 23, 1981	5'10" / 177 cm	185 lbs / 84 kg	Canada	L
5	Lloyd Marks	C	Oct 21, 1977	5'8" / 173 cm	174 lbs / 79 kg	Canada	L
6	Jason Deskins	C	May 06, 1979	5'10" / 178 cm	185 lbs / 84 kg	USA	-
7	Jim Abbott	LW/C	May 03, 1980	6'1" / 186 cm	185 lbs / 84 kg	USA	L

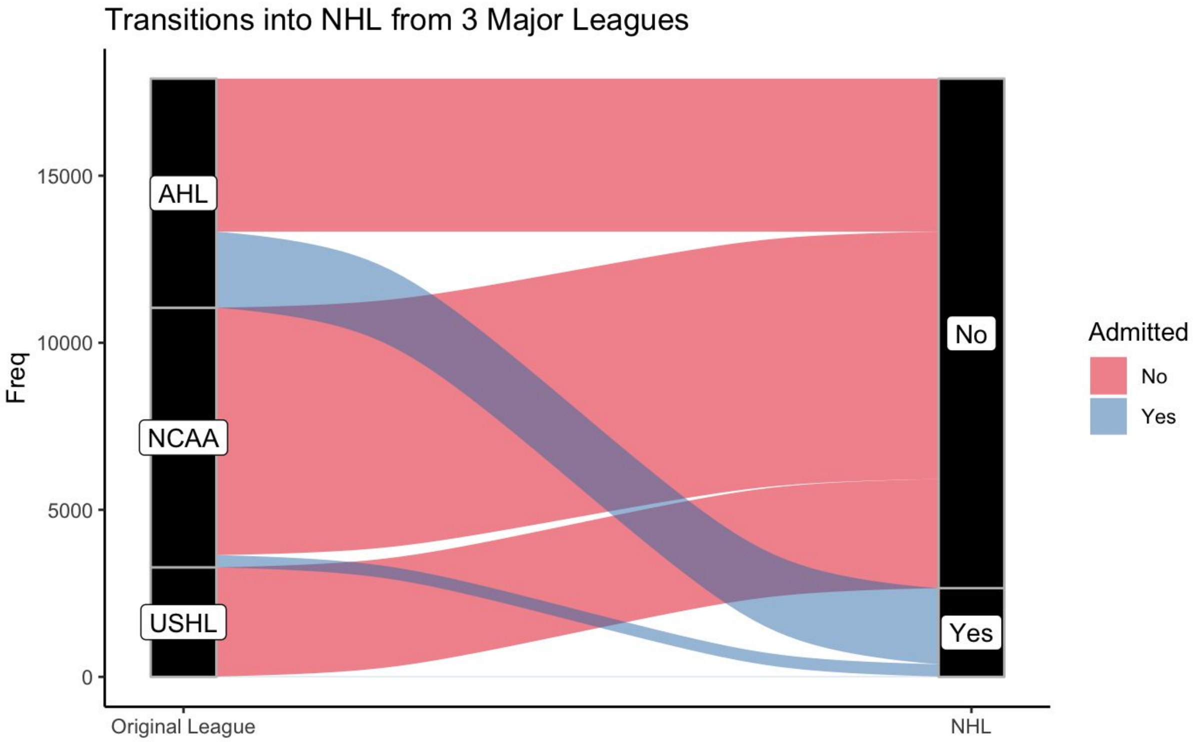
Data Description

- Player performance:
  - 266326 rows (15220 of players * # of seasons they played)

	Player	Season	Team	League	Games	Goals	Assists	TotalPoints	PenaltyMinutes	PlusMinus
1	Scott May	1998-99	South Surrey Eagles	BCHL	45	10	28	38	23	
2	Scott May	1999-00	South Surrey Eagles	BCHL	54	42	42	84	86	
3	Scott May	2000-01	Ohio State Univ.	NCAA	37	9	9	18	26	-3
4	Scott May	2001-02	Ohio State Univ.	NCAA	40	12	17	29	42	4
5	Scott May	2002-03	Ohio State Univ.	NCAA	43	10	25	35	56	5
6	Scott May	2003-04	Ohio State Univ.	NCAA	41	15	19	34	42	4
7	Scott May		St. John's Maple Leafs	AHL	5	1	1	2	2	3
8	Scott May	2004-05	St. John's Maple Leafs	AHL	16	0	1	1	21	-3

Carnegie Mellon University

EDA



Original League	Transition to NHL	Fail to transition to NHL
AHL	2275	4581
NCAA	367	7404
USHL	13	3265

Carnegie Mellon University

Method: Causal Inference

- Goal: determine the causal effect of development paths (Treatment **Z**) on players' future in the NHL (Response **Y**), controlling for player quality etc. (Confounders **X**)
  - Conditional Average Treatment Effect (CATE):

    $$E[Y \mid Z = z1, X] - E[Y \mid Z = z0, X]$$

  - The fundamental problem: our samples are biased (e.g. better prospects are more likely to enter NCAA than USHL)

Method: Two solutions

- Solution 1: Control the treatment assignment mechanism; then estimate causal effect just as in a randomized experiment
  - Method: Propensity Score Matching
- Solution 2: If we can precisely estimate a model for the outcome $Y = f(z, x) + \varepsilon$, then we can calculate CATE
  - Method: Bayesian Additive Regression Trees (BART)

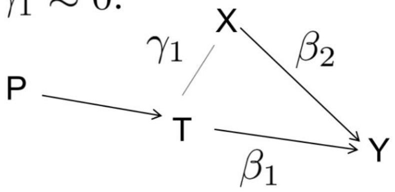
Method

Propensity Score Matching

- Propensity scores are used to rearrange the data so that we don't have any selection effect or bias in our treatment
- Reflecting back to the research question, we are interested in assessing the causal effect of development path on a players success in the NHL
    - Treatment and control groups (at draft year 1): NCAA and USHL
    - Predictors (at draft year 0): goals per game, plus minus per game, penalty minutes per game, position, height, weight
    - Outcome: if the player played in 10 or more games after being drafted into the NHL,
- We use logistic regression to predict the treatment T as well as possible from all of the predictors. P(T = NCAA) is our **propensity score**
- For each player that was in the NCAA in their draft year 1 (when they were 19), I matched it to a unit that was in USHL during their draft year 1 with the same or similar propensity score
    - Non-marching units were not used in the modeling

Propensity score matching
Makes **λ_1 = 0.** This ensures that we are simulating a experiment in each player is randomly assigned into the USHL and NCAA during their draft year 1

$\gamma_1 \approx 0.$



Method

BART

- Estimate Y = f (z, x) + ε using a **sum-of-trees** model
- The idea is to fit a bunch of **weak-learning** (small) trees, each fitting to the residuals of the previous trees. Then, **additively** combine these trees to reduce bias, similar to boosting.
- Introduce a **regularization prior** to avoid overfitting. The prior controls the size of the trees (T), the magnitude of the outputs of the trees (M), and the value of σ.
- Compute the **posterior** using **Markov Chain Monte Carlo** (MCMC). At each iteration of MCMC, (T, M) and σ are redrawn to seek a good f.
- After estimating Y using BART, we can calculate CATE by

$$\frac{1}{n}\sum_{i=1}^{n} f(1, x_i) - f(0, x_i)$$

- Select players who is not in NHL at Draft Year 1:
- Predictors:
  - League, Games, Goals, Assists, PenaltyMinutes, PlusMinus, Position, Nation, Shoots, Performance, Height_cm, Weight_kg
- Response variable:
  - How many games played in NHL?

```
bm <- bartMachine(X, Y, verbose = FALSE,
                  serialize = TRUE, use_missing_data = TRUE)
```

Carnegie Mellon University

Method

Bayesian Additive Regression Trees (BART)

- Selected 11,637 players who played in some developmental league in the following season

- Predictors:
  - League, Games, Goals, Assists, PenaltyMinutes, PlusMinus, Position, Nation, Shoots, Performance, Height_cm, Weight_kg

- Response variable:
  - The average number of games played in the NHL per player

Results

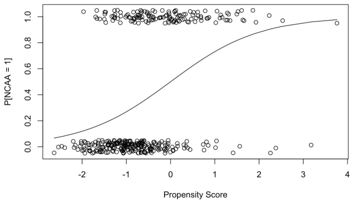
Propensity Score Matching: Forward Positioned Players

Matching Process

- The result of the matching process should give us zero coefficient estimates when predicting propensity score (P(NCAA = 1))
- For all of the coefficient estimates, we do not observe enough evidence to reject the null hypothesis that the coefficient estimates are zero.
- While we observe non-negative coefficients, we cannot conclude that they are non-zero based on the large SE's of the estimates

	coef.est	coef.se
(Intercept)	-3.22	4.34
penaltymin_pergame_1	0.01	0.15
plusminus_pergame_1	0.60	0.48
assists_pergame_1	0.82	0.87
goals_pergame_1	1.65	1.06
Weight	-0.01	0.01
height	0.02	0.03

glm(NCAA ~ penaltymin_pergame + plusminus_pergame + assists_pergame_1 + goals_pergame + Weight + height , family = 'binomial', data = matched))



Matching Results

- Development path (NCAA vs USHL) remains an insignificant predictor in our model even after matching players based on propensity scores

Treatment Variable	Estimate	Standard Error
USHL (Before matching)	-0.513	0.61
USHL (After matching)	-0.458	0.59

glm(more_than_10_games ~ penaltymin_pergame + plusminus_pergame + assists_pergame + goals_pergame + Weight + height + development path,, family = 'binomial')

Results

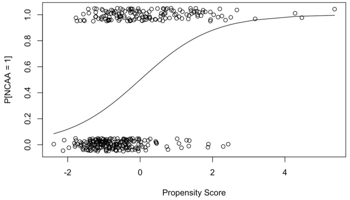
Propensity Score Matching: Backward Positioned Players

Matching Process

- The result of the matching process should give us zero coefficient estimates when predicting propensity score (P(NCAA = 1))
- For all of the coefficient estimates, we do not observe enough evidence to reject the null hypothesis that the coefficient estimates are zero.
- While we observe non-zero coefficients, we cannot conclude that they are non-zero based on the large SE's of the estimates

```
                    coef.est coef.se
(Intercept)         -7.54     4.66
penaltymin_pergame_1 -0.03    0.15
plusminus_pergame_1  0.27     0.43
assists_pergame_1    2.16     0.75
goals_pergame_1      2.75     1.07
Weight              -0.01     0.01
height               0.04     0.03
```

glm(NCAA ~ penaltymin_pergame + plusminus_pergame + assists_pergame_1 + goals_pergame + Weight + height , family = 'binomial', data = matched))



Matching Results

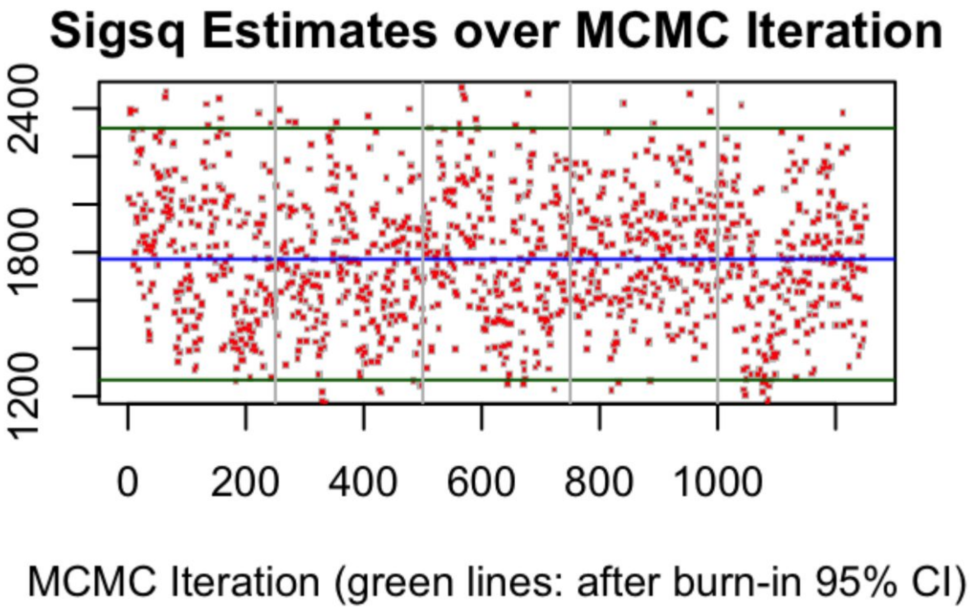
- Development path (NCAA vs USHL) is a significant predictor before and after matching
- After matching, we observe a decrease in the coefficient estimate for development path
- After the removal of selection bias, we observe a decrease in the log odds of success in the NHL when going from the NCAA to the USHL

Treatment Variable	Estimate	Standard Error
USHL (Before matching)	-0.8889	0.373
USHL (After matching)	-1.047	0.409

glm(more_than_10_games ~ penaltymin_pergame + plusminus_pergame + assists_pergame + goals_pergame + Weight + height + development path,, family = 'binomial')

Result

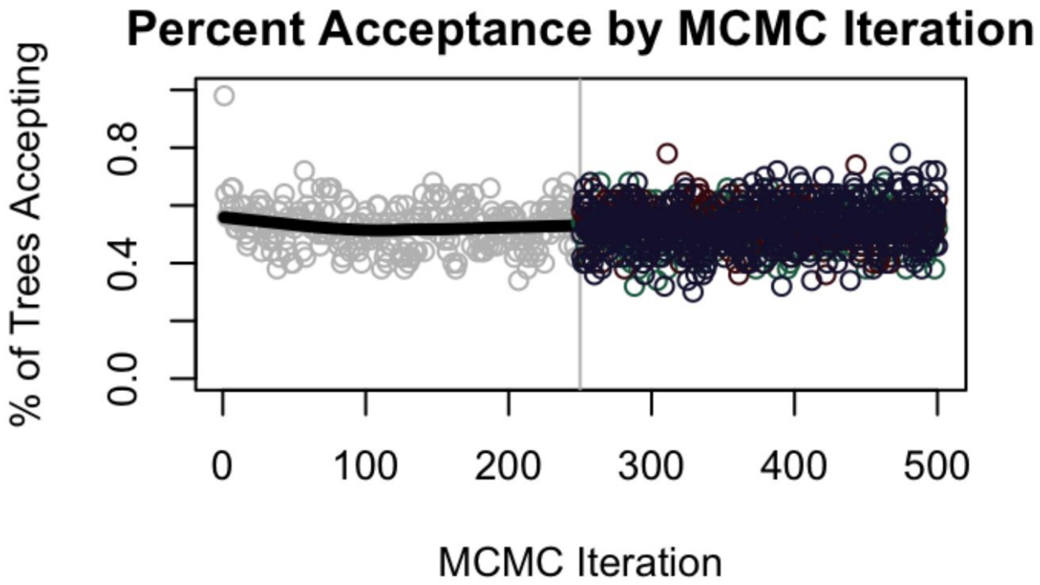
BART



Posterior error variance estimates:

Result

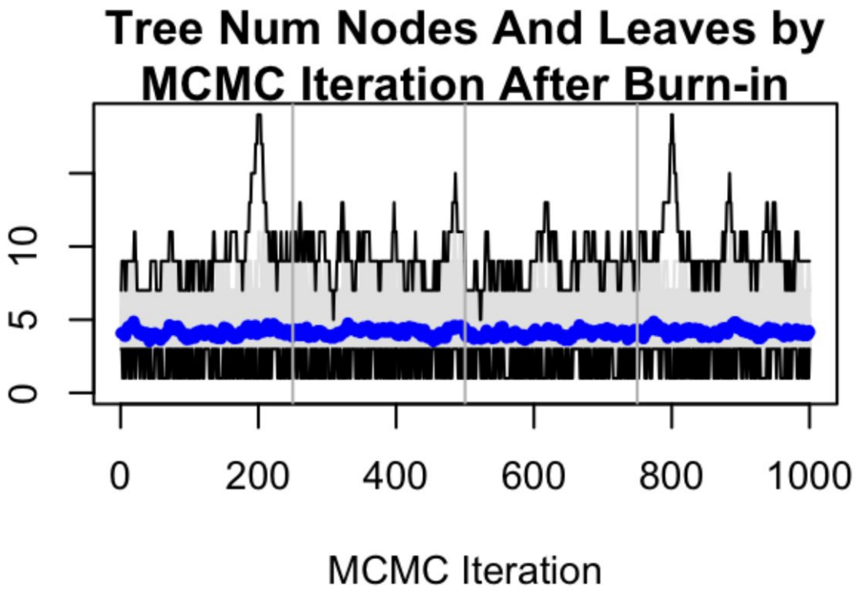
BART



accepted divided by # of trees:
About 50% of the trees was accepted

Result

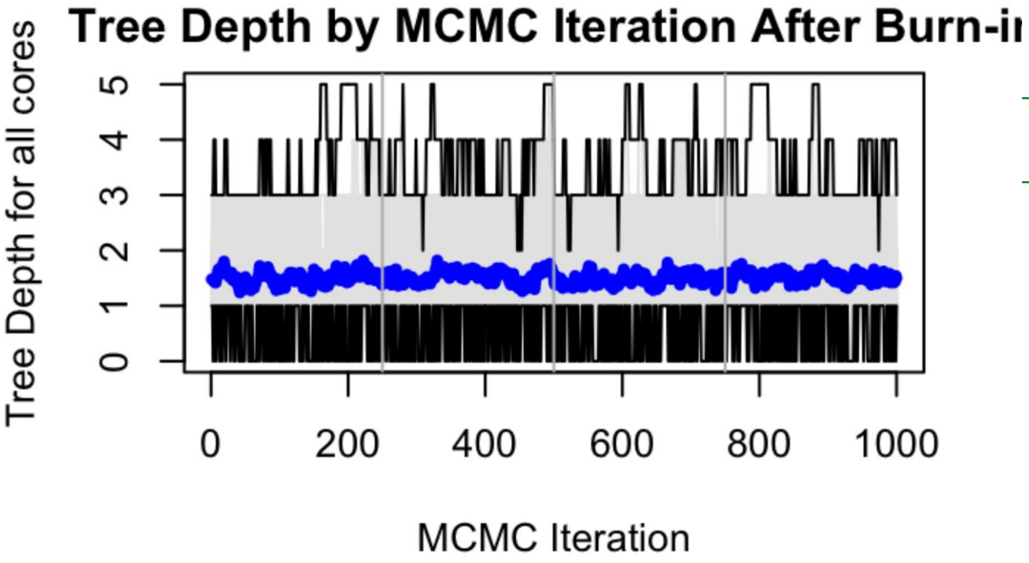
BART



Average ***number*** of nodes across each tree

Result

BART



Average ***depth*** of nodes across each tree

Next Steps & Roadblocks

- Propensity Scores
  - Add additional treatment groups/developmental paths to analysis such as WHL, OHL
  - Add additional predictors to better estimate treatment effect

- BART:
  - Implement separate modes for forward/defence players from different league

Q&A

Thank You!

Appendix

Results

Propensity Score Matching: Predictors at draft year 0 and treatment/control group at draft year 1

When we regress the predictors on the treatment effect for only the matched data, we achieve an ok desirable result. The coefficient estimates appear to not be zero at some statistical significance indicating that we were unsuccessful at removing selection bias

	coef.est	coef.se
(Intercept)	-1.15	1.77
penaltymin_pergame_1	-0.27	0.07
plusminus_pergame_1	0.43	0.20
position_newforward	-0.35	0.11
goals_pergame_1	2.41	0.38
Weight	0.01	0.00
height	0.00	0.01

(Intercept)	penaltymin_pergame_1	plusminus_pergame_1	position_newforward	goals_pergame_1
-3.976201	0.045283	0.149124	-1.137084	2.483174
Weight	height	new_leagueUSHL		
0.038613	-0.027379	-1.989396		

(Intercept)	penaltymin_pergame_1	plusminus_pergame_1	position_newforward	goals_pergame_1
-4.585015	0.057535	0.042629	-1.070950	2.368416
Weight	height	new_leagueUSHL		
0.039696	-0.025307	-1.990116		

We run the `glm(formula = more_than_10_games ~ plusminus_pergame + position_new + PenaltyMin_pergame + Weight + height + League)` call on the matched data and compare the coefficient estimate for our treatment effect with the glm model on the original unmatched data. The coefficient estimate on the unmatched data is -1.98 for treatment: USHL while we observe an estimate of -1.99 on the matched data

Carnegie Mellon University

Results

Propensity Score Matching: Predictors (including league) at draft year 0 and treatment/control group at draft year 1

	coef.est	coef.se
(Intercept)	18.78	2797.12
penaltymin_pergame_1	-0.07	0.09
plusminus_pergame_1	0.37	0.29
lag.valueAHL	-16.10	2797.12
lag.valueAsia League	0.21	4845.12
lag.valueBCHL	-19.58	2797.12
lag.valueDEL	0.28	4845.12
lag.valueDenmark	0.09	4845.12
lag.valueDenmark2	0.32	4845.12
lag.valueECHL	-16.76	2797.12
lag.valueEJHL	-19.60	2797.12
lag.valueNAHL	-20.37	2797.12
lag.valueNCAA	-16.95	2797.12
lag.valueNCAA III	-17.60	2797.12
lag.valueNHL	-0.04	3301.91
lag.valueOHL	-17.88	2797.12
lag.valueQMJHL	-0.02	4845.12
lag.valueUHL	-17.23	2797.12
lag.valueUSHL	-19.45	2797.12
lag.valueUSPHL Premier	-16.67	2797.12
lag.valueVHL	0.27	4845.12
lag.valueWCup	-0.17	4845.12
lag.valueWHL	0.22	3955.94
lag.valueWJC-20	0.14	2835.26
position_newforward	-0.15	0.14
goals_pergame_1	0.74	0.50
Weight	0.00	0.01
height	-0.01	0.02

When we regress the predictors on the treatment effect for only the matched data, we achieve an ok desirable result. The coefficient estimates appear to all be zero at some statistical significance indicating that we were successful in removing selection bias

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-17.9608	4560.8677	0.00	0.99686	
penaltymin_pergame_1	0.0167	0.1328	0.13	0.90013	
plusminus_pergame_1	-0.1035	0.3905	-0.27	0.79098	
lag.valueACHA	2.1086	5911.3844	0.00	0.99972	
lag.valueACHA II	2.3195	6319.0200	0.00	0.99971	
lag.valueAHL	17.5031	4560.8659	0.00	0.99694	
lag.valueAWHL	3.1574	5403.1504	0.00	0.99953	
lag.valueBCHL	2.3076	4939.5912	0.00	0.99963	
lag.valueCIS	1.1207	7959.0397	0.00	0.99989	
lag.valueDEL	0.2718	7959.0397	0.00	0.99997	
lag.valueDenmark	-0.2898	7959.0396	0.00	0.99997	
lag.valueDenmark2	1.9787	7959.0396	0.00	0.99980	
lag.valueECHL	0.6262	4921.1462	0.00	0.99990	
lag.valueEJHL	2.0572	4935.5234	0.00	0.99967	
lag.valueFHL	1.7992	7959.0397	0.00	0.99982	
lag.valueMWHL	2.2227	7959.0396	0.00	0.99978	
lag.valueNA3HL	3.2854	7959.0397	0.00	0.99967	
lag.valueNAHL	1.8563	4629.6607	0.00	0.99968	
lag.valueNCAA	17.2397	4560.8659	0.00	0.99698	
lag.valueNCAA III	1.4885	4743.7737	0.00	0.99975	
lag.valueNHL	19.5277	4560.8660	0.00	0.99658	
lag.valueNOJHL	2.6805	7959.0397	0.00	0.99973	
lag.valueOHL	0.6889	5780.3520	0.00	0.99990	
lag.valueQMJHL	0.1570	7959.0397	0.00	0.99998	
lag.valueUHL	1.5129	6372.6515	0.00	0.99981	
lag.valueUSHL	15.2703	4560.8660	0.00	0.99733	
lag.valueUSHS-Prep	2.2285	7959.0397	0.00	0.99978	
lag.valueUSPHL Premier	0.3694	5842.1216	0.00	0.99995	
lag.valueVHL	1.1994	7959.0396	0.00	0.99988	
lag.valueWC	38.5980	7959.0397	0.00	0.99613	
lag.valueWCup	38.4477	7959.0397	0.00	0.99615	
lag.valueWHL	0.0371	6481.2072	0.00	1.00000	
lag.valueWJC-20	19.3864	4560.8659	0.00	0.99661	
lag.valueWSHL	3.5752	7959.0397	0.00	0.99964	
position_newforward	-1.4733	0.2632	-5.60	2.2e-08	***
goals_pergame_1	2.6289	0.6722	3.91	9.2e-05	***
Weight	0.0364	0.0115	3.16	0.00159	**
height	-0.0438	0.0301	-1.45	0.14630	
new_leagueUSHL	-1.0630	0.2952	-3.60	0.00032	***

We run the `glm(formula = more_than_10_games ~ plusminus_pergame + position_new + PenaltyMin_pergame + Weight + height + League)` call on the matched data and compare the coefficient estimate for our treatment effect with the glm model on the original unmatched data. The coefficient estimate on the unmatched data for treatment: USHL is the same as the estimate on the matched data

Carnegie Mellon University