

41201-01: Data Mining

Professor Taddy

Joanne Chen

James Hardiman

Evan Johnson

Honor code: We pledge our honor that we have not violated the Honor Code during the completion of this assignment.

**[1] Interpret AICc selected model from my nhlreg lasso. Just tell some stories about what the model tells you.**

There is a positive--albeit small--intercept inferred that there is a tendency for goals to be scored than not. All uneven play has strong positive influence on the probability to score a goal with 2 person advantages tending to be stronger than 1 person advantages. There are coefficients associated with each team (both positive and negative) although these tend to be small (between -.1 and +.1). These show how consistently good (bad) teams are over time and capture team specific factors such as coaching, ability to attract free agents, etc. All season coefficients are zero implying that there are not seasons where more goals are scored on average (that cannot not be captured by other factors in the model). There are team-season interactions showing the relative strength of teams over time. There are player specific coefficients ranging from ~-.9 to ~.8, with more than half being 0 and an average of ~0.008 (see *exhibit A*). This implies that most players have little influence on whether a goal is score, but some players have a strong influence.

**[2] The gamlr run for nhlreg uses standardize=FALSE. Why did I do this? What happens if you do standardize?**

You don't want to standardize because all the variables are already on the same scale. In this case the sd is only a measure of the spread of the data (i.e., is there a disproportionate number of 0s or 1s) which would cause the sd to shrink. This is unnecessarily punishing the variables with a more even distribution of 0s and 1s in the data set. Also, the underlying data is binomial so standard deviation does not have the same interpretation. There is no need to standardize again because all data is already on the same scale.

If I standardize, the player specific coefficients will range from ~-3.16 to ~1.69, with more than half being 0 and an average of ~-0.03 (see *exhibit B*).

When looking at the distribution of coefficients for standard vs nonstandard across lambdas, nonstandard results in a tighter distribution (see *exhibit C*).

**[3] Compare model selection methods for the nhlreg lasso. Consider both IC and CV (you’ll want to create cv.nhlreg).**

Looking at Exhibit D, BIC has the worst model selection. AICc is slightly better than AIC. Comparing IC vs AICc, they picked similar models with AICc picking a slightly more complex model with similar performance because the average deviance is similar.

**[4] We’ve controlled our estimates for confounding information from team effects and special play configuration. How do things change if we ignored this info (i.e., fit a player-only model)? Which scheme is better (interpretability, CV, and IC)?**

It does make a difference. Freeing the non-player variables is better because then we get coefficients for teams and team:season interactions, which we would expect not freeing these variables we obtain no team:season interactions. From a CV perspective the best performing non-freed model performs worse than the best freed model (see Exhibit F).

In the standard model and player-only model, CV and IC performance is similar. In the model with no team effects and special play configuration, CV and IC choose different models, CV chooses a model with a lower mean deviance (see exhibit E).

**Exhibit A –** player coefficients

****

**Exhibit B**

****

**Exhibit C**

****

**Exhibit D**

****

*1 = AIC, 2 = AICc, 3 = BIC*

****

*1 = AIC, 2 = AICc*

****

Green = AICc, Red = BIC

**Exhibit E**

****

**Exhibit F**

****