Econ 573: Problem Set 3

Due date: 10/4

Fall 2021

The problem set is due **before** 11 p.m. the day before the class. Instructions: give a justification to all of your answers. Make sure to submit all the relevant files. For the empirical exercise you can find the data following this link:
https://www.statlearning.com/resources-second-edition.

Part I

Ex 2, 3, 10, 11 from Chapter 6 of ISL.

Part II

Ex. 4, 8, 11, 12, 13 from Chapter 4 of ISL.

Part III: Hockey Regression

NB. This part is optional. You can get up to extra 25% of the credit.

A stat for player performance is the 'plus-minus' (PM). PM is a function of goals scored while that player is on the ice: the number of goals for his team, minus the number against. There is no accounting for teammates or opponents. Due to 'line matching' this could make a big difference. Can we build a better performance metric with regression?

Response is a binary 1 if home goal, 0 if away goal. Home players get an x-value of +1, and away players -1. Everyone off the ice is zero.

Our logistic regression plus-minus model is

$$\log \frac{\Pr(Y=1|X)}{1-\Pr(Y=1|X)} = \beta_0 + \sum_{\text{homeplayers}} \beta_j - \sum_{\text{awayplayers}} \beta_j$$

β_j is j^{th} player's partial effect: When a goal is scored and player j is on ice, odds are multiplied by e^{β_j} that his team scored.

In addition to 'controlling' for the effect of who else is on the ice, we also want to control for things unrelated to player ability. (crowd, coach, schedule, ...)

We'll use a fixed effect' for each team-season, $\alpha_{\text{team,season}}$. Also, special configurations (e.g., 5 on 4 power play) get α_{config}. So the full model has 'log odds that a goal was by home team'

$$\beta_0 + \alpha_{\text{team,season}} + \alpha_{\text{config}} + \sum_{\text{homeplayers}} \beta_j - \sum_{\text{awayplayers}} \beta_j$$

`gamlr` includes data on NHL goals from 2002/03-2012/13. The code to design and fit this model is in `hockey_start.R`. Via the `free`, only player β_k's are penalized.

1. Interpret AICc selected model from my `nhlreg` lasso. Just tell some stories about what the model tells you.
2. The `gamlr` run for `nhlreg` uses standardize=FALSE. Why did I do this? What happens if you do standardize?
3. Compare model selection methods for the `nhlreg` lasso. Consider both IC and CV (you'll want to create `cv.nhlreg`).
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)?
5. Can you translate player β_k effects into something comparable to classic Plus-Minus? How do things compare?