Regularized Estimation of Player Performance

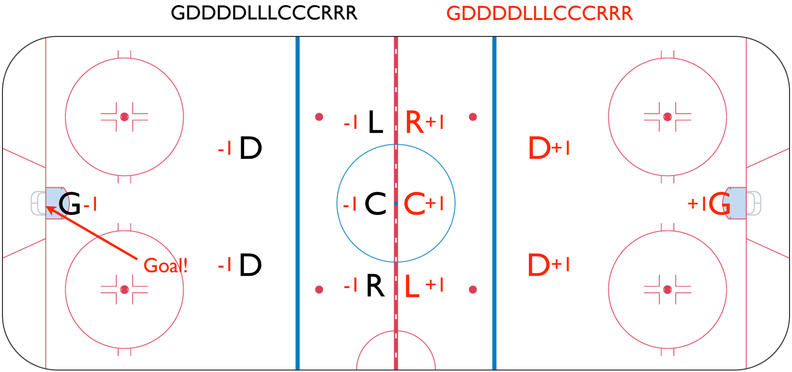
Bobby Gramacy, Chicago Booth

Matt Taddy, Microsoft Research and Chicago Booth

with Sen Tian (NYU) and Shane Jenson (Wharton)

Plus-Minus

PM is a running count of, for every goal, a +1 for those on the scoring team and -1 for those on the team scored upon.



It doesn't quite measure player quality, as we haven't controlled for the effects of teammate and opponent quality (or anything else).

A Regression version of PM

Set up a 'response' variable:

$y_i = +1$ for a *home* team goal,

$y_i = -1$ for an *away* team goal.

We're interested in how individual players affect

$$q_i = \mathrm{p}(y_i = 1) = \mathrm{p}(\text{home team scored goal } i)$$

The standard model for such problems is logistic regression, say

$$\log\left[\frac{q_i}{1-q_i}\right] = \alpha + \beta_{HG} + \beta_{HD} \ldots + \beta_{HR} - \beta_{AG} - \ldots - \beta_{AR}$$

where β_{HG} is Home-Goalie and β_{AR} is Away-Right-wing, etc.

Then, for player j and given a goal was scored, e_j^{β} is the multiplier on odds that it was scored by his team if he's on the ice.

We actually use a larger regression model:

$$\log\left[\frac{q_i}{1-q_i}\right] = \alpha + \mathbf{u}_i'\gamma + \mathbf{v}_i'\varphi + \mathbf{x}_i'\boldsymbol{\beta}_0 + (\mathbf{x}_i \circ \mathbf{s}_i)'(\boldsymbol{\beta}_s + p_i\boldsymbol{\beta}_p)$$

where

- $\mathbf{u}_i$ holds indicators for each team-season,
- $\mathbf{v}_i$ holds indicators for various special-teams scenarios,
- $\mathbf{x}_i$ contains player-presence indicator,

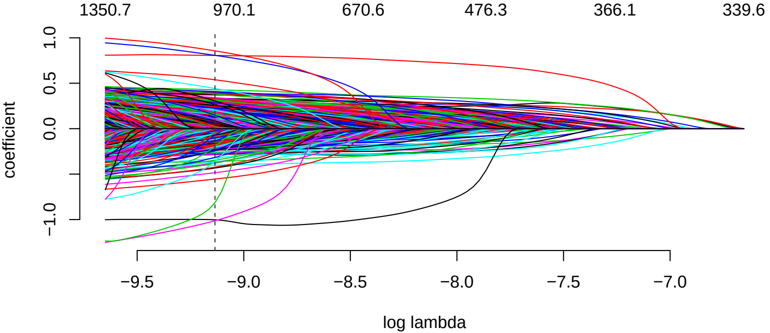
All of these indicators are $+1$ for home and -1 for away.

Then β_j measures player effect after controlling for team strength (e.g., coach or schedule) and on-ice scenarios (e.g., PP or PK).

We also allow deviations in the player effects for specific seasons (s_{it}) and in the playoffs (p_{it}), but these are seldom 'significant'.

Regularization

Instead of minimizing deviance, we minimize deviance *plus penalty* $\lambda|\beta_j|$ *on the size of each* β_j *estimates*.



Enumerate a 'path' of models for different λ, and use the one that predicts best out-of-sample. This is how modern stats/ML works.