# Regularized Estimation of Player Performance

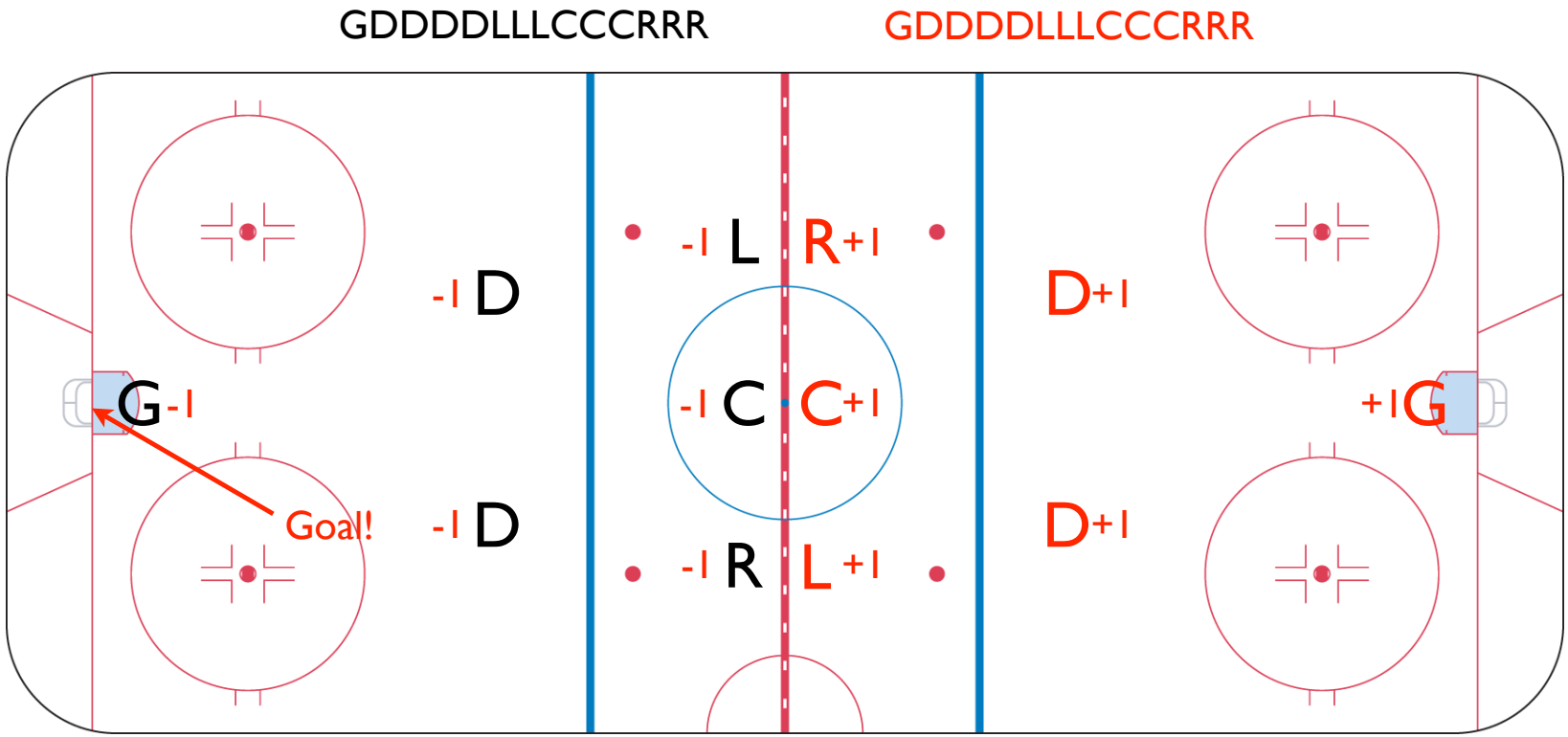
**Bobby Gramacy**, Chicago Booth

**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 $\beta_{HG}$ is Home-Goalie and $\beta_{AR}$ is Away-Right-wing, etc.

Then, for player $j$ and given a goal was scored, $e_j^{\beta}$ 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:

$$\log\left[\frac{q_i}{1-q_i}\right] = \alpha + \mathbf{u}_i'\gamma + \mathbf{v}_i'\varphi + \mathbf{x}_i'\beta_0 + (\mathbf{x}_i \circ \mathbf{s}_i)'(\beta_s + p_i\beta_p)$$

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,

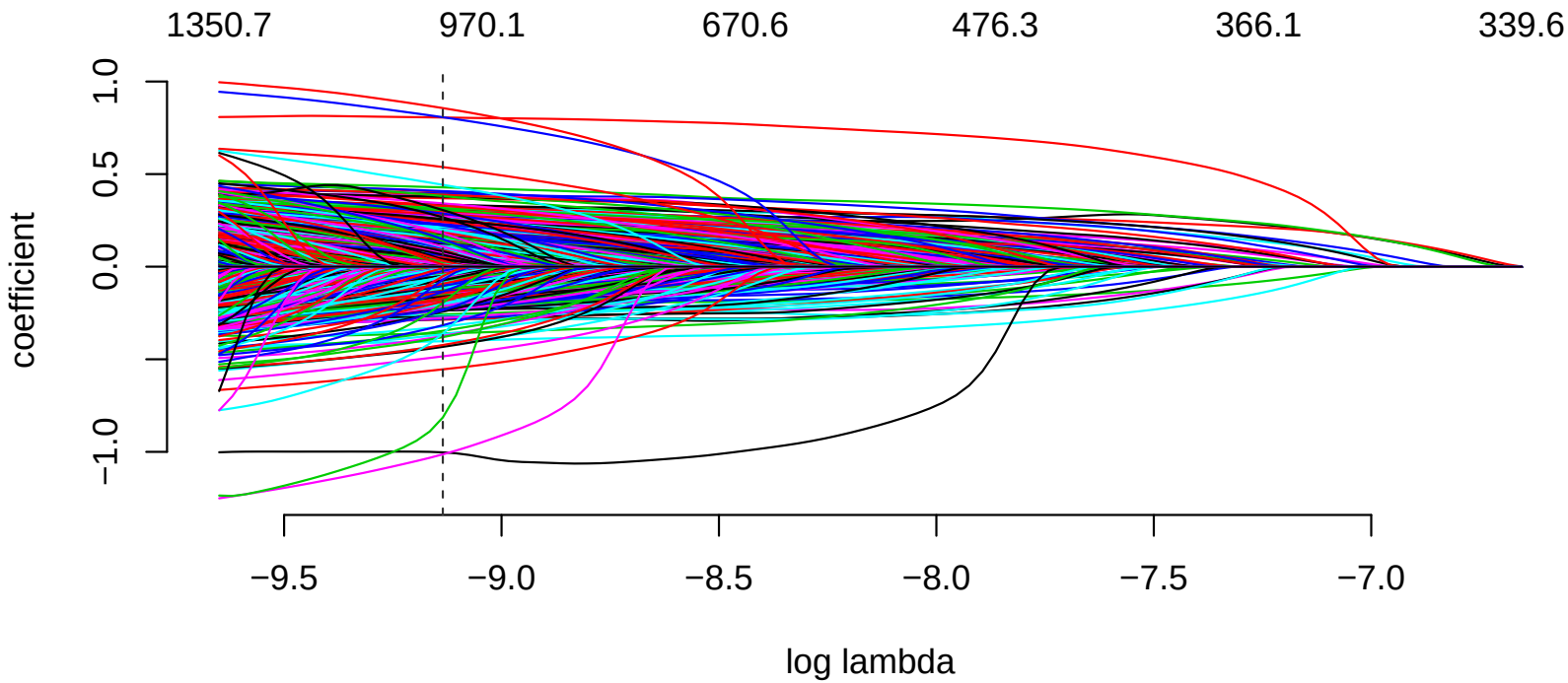
All of these indicators are $+1$ for home and -1 for away.

Then $\beta_j$ 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* $\beta_j$ *coefficient*. This is called the LASSO.




Enumerate a 'path' of models for different $\lambda$, and use the one that predicts best out-of-sample. This is how modern statistics works.

See the `gamlr` package for R, and `help(hockey)` for this example.

## Partial PM and FP

We use the estimated logistic regression model to produce 'partial' (i.e., without effect of confounders) versions of standard statistics.

- For-Percent: $PFP_j = \dfrac{e^{\beta_j}}{(1 + e^{\beta_j})}$
- Plus-Minus: $PPM_j = G_j PFP_j - G_j(1 - PFP_j)$
  where $G_j$ is the total goals with player $j$ on-ice.

PFP measures the player's average contribution, and PPM scales this up by # of goals (which is a rough surrogate for time-on-ice).

We also fit models for non-goal response $y_i$, like shots or corsi events. This gives partial versions of stats based on those metrics.

**Goal-based performance analysis, ordered by PPM: Studs**

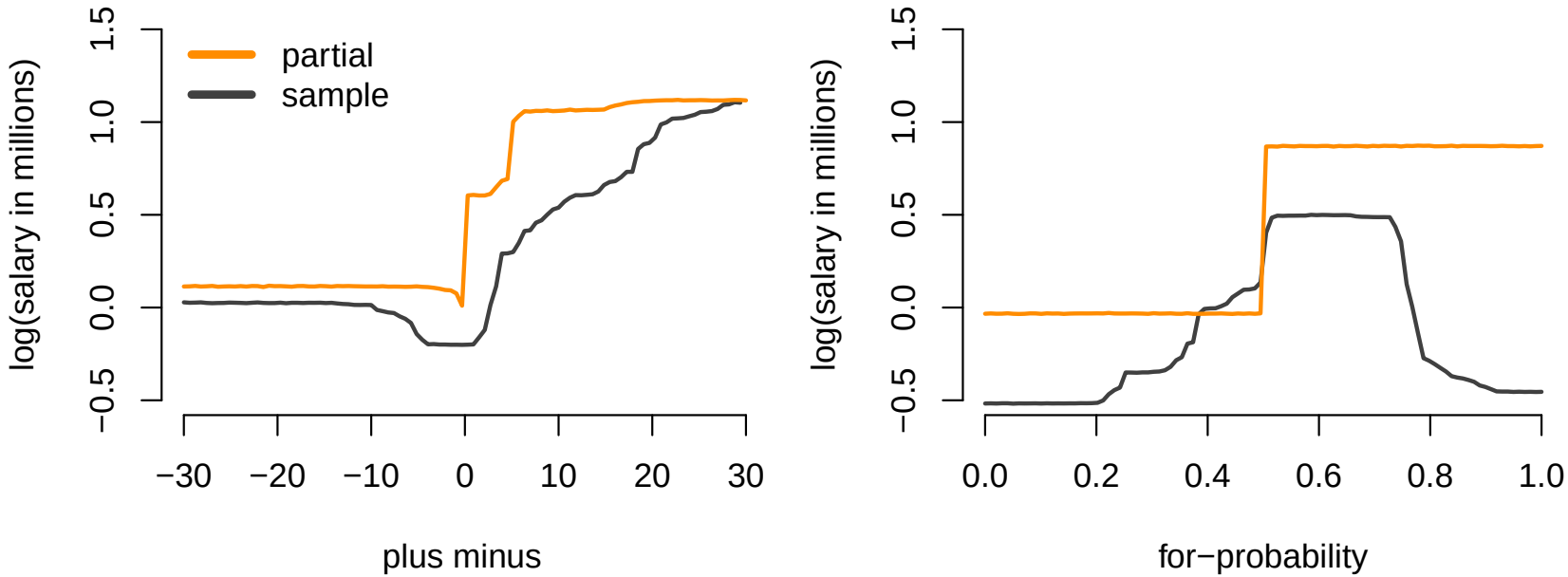
| Rank | Player | Season | Team | PFP | FP | PPM | PM |
|---|---|---|---|---|---|---|---|
| 1 | PETER FORSBERG | 2002-2003 | COL | 0.68 | 0.77 | 55.52 | 85 |
| 2 | SIDNEY CROSBY | 2009-2010 | PIT | 0.60 | 0.64 | 43.47 | 60 |
| 3 | DOMINIK HASEK | 2005-2006 | OTT | 0.59 | 0.67 | 42.45 | 80 |
| 4 | SIDNEY CROSBY | 2008-2009 | PIT | 0.60 | 0.61 | 42.26 | 48 |
| 5 | SIDNEY CROSBY | 2005-2006 | PIT | 0.60 | 0.62 | 41.86 | 52 |
| 6 | PETER FORSBERG | 2005-2006 | PHI | 0.68 | 0.77 | 40.67 | 61 |
| 7 | PAVEL DATSYUK | 2007-2008 | DET | 0.60 | 0.72 | 39.49 | 87 |
| 8 | PAVEL DATSYUK | 2008-2009 | DET | 0.60 | 0.67 | 39.49 | 69 |
| 9 | SIDNEY CROSBY | 2006-2007 | PIT | 0.60 | 0.72 | 35.62 | 79 |
| 10 | MARK STREIT | 2008-2009 | NYI | 0.59 | 0.56 | 35.08 | 24 |
| 11 | MATT MOULSON | 2011-2012 | NYI | 0.60 | 0.61 | 34.92 | 37 |
| 12 | LUBOMIR VISNOVSKY | 2010-2011 | ANA | 0.58 | 0.66 | 34.52 | 70 |
| 13 | ALEX OVECHKIN | 2008-2009 | WAS | 0.57 | 0.66 | 34.46 | 80 |
| 14 | JOE THORNTON | 2009-2010 | SJS | 0.60 | 0.65 | 33.91 | 52 |
| 15 | JOE THORNTON | 2010-2011 | SJS | 0.60 | 0.64 | 33.91 | 48 |
| 16 | ONDREJ PALAT | 2013-2014 | TAM | 0.64 | 0.66 | 32.75 | 37 |
| 17 | PAVEL DATSYUK | 2006-2007 | DET | 0.60 | 0.71 | 32.61 | 70 |
| 18 | JOE THORNTON | 2002-2003 | BOS | 0.60 | 0.64 | 32.17 | 47 |
| 19 | JOE THORNTON | 2007-2008 | SJS | 0.60 | 0.71 | 32.17 | 69 |
| 20 | ANDREI MARKOV | 2007-2008 | MON | 0.57 | 0.60 | 31.9 | 47 |

## Goal-based performance analysis, ordered by PPM: Duds

| Rank | Player | Season | Team | PFP | FP | PPM | PM |
|---|---|---|---|---|---|---|---|
| 10184 | PATRICK LALIME | 2008-2009 | BUF | 0.43 | 0.44 | -15.79 | -15 |
| 10185 | JACK JOHNSON | 2007-2008 | LOS | 0.45 | 0.39 | -15.82 | -34 |
| 10186 | BRETT CLARK | 2011-2012 | TAM | 0.44 | 0.35 | -16.93 | -47 |
| 10187 | NICLAS HAVELID | 2008-2009 | ATL | 0.45 | 0.39 | -16.97 | -40 |
| 10188 | JACK JOHNSON | 2010-2011 | LOS | 0.45 | 0.53 | -17.21 | 9 |
| 10189 | JACK JOHNSON | 2011-2012 | LOS | 0.45 | 0.5 | -17.21 | -1 |
| 10190 | P. J. AXELSSON | 2008-2009 | BOS | 0.41 | 0.49 | -17.35 | -1 |
| 10191 | BRYAN ALLEN | 2006-2007 | FLA | 0.45 | 0.45 | -17.9 | -17 |
| 10192 | JACK JOHNSON | 2009-2010 | LOS | 0.45 | 0.49 | -19.46 | -4 |
| 10193 | PATRICK LALIME | 2005-2006 | STL | 0.43 | 0.40 | -19.77 | -29 |
| 10194 | ALEXANDER EDLER | 2013-2014 | VAN | 0.37 | 0.27 | -20.49 | -35 |
| 10195 | PATRICK LALIME | 2007-2008 | CHI | 0.43 | 0.49 | -22.29 | -4 |
| 10196 | TIM THOMAS | 2009-2010 | BOS | 0.43 | 0.46 | -24.22 | -16 |
| 10197 | ANDREJ MESZAROS | 2006-2007 | OTT | 0.42 | 0.48 | -27.32 | -6 |
| 10198 | BRYCE SALVADOR | 2008-2009 | NJD | 0.35 | 0.37 | -34.4 | -31 |
| 10199 | PATRICK LALIME | 2002-2003 | OTT | 0.43 | 0.58 | -37.81 | 47 |
| 10200 | PATRICK LALIME | 2003-2004 | OTT | 0.43 | 0.56 | -37.81 | 37 |
| 10201 | NICLAS HAVELID | 2006-2007 | ATL | 0.34 | 0.44 | -62.64 | -22 |
| 10202 | NICLAS HAVELID | 2005-2006 | ATL | 0.33 | 0.40 | -65.94 | -41 |
| 10203 | JAY BOUWMEESTER | 2005-2006 | FLA | 0.33 | 0.42 | -69.62 | -32 |

For abundant detail and analysis, see our handbook chapter
Hockey Performance via Regression by Gramacy, Taddy, and Tian.

e.g., look at average player salary by standard and partial statistics



## Bargains!

Top-15 undervalued players in 2013-2014

| Rank | Player | Team | Goals per million |
|---|---|---|---|
| 1 | ONDREJ PALAT | TAM | 58.27 |
| 2 | RYAN NUGENT-HOPKINS | EDM | 19.81 |
| 3 | GABRIEL LANDESKOG | COL | 16.74 |
| 4 | TYLER TOFFOLI | LOS | 16.72 |
| 5 | GUSTAV NYQUIST | DET | 9.08 |
| 6 | JADEN SCHWARTZ | STL | 8.43 |
| 7 | ERIC FEHR | WAS | 7.51 |
| 8 | ANDREW MACDONALD | NYI | 7.48 |
| 9 | BENOIT POULIOT | NYR | 6.43 |
| 10 | BRAD BOYES | FLA | 6.01 |
| 11 | TOMAS TATAR | DET | 5.83 |
| 12 | AL MONTOYA | WPG | 5.79 |
| 13 | BRANDON SAAD | CHI | 5.5 |
| 14 | FRANS NIELSEN | NYI | 5.5 |
| 15 | JAROMIR JAGR | NJD | 4.73 |

We also include comparison of shot and goal-based metrics:

**PFP player rankings**

| | goal-based | | | | Corsi-based | | | |
|---|---|---|---|---|---|---|---|---|
| Rank | Player | Season | Team | PFP | Player | Season | Team | PFP |
| 1 | PETER FORSBERG | 2002-2003 | COL | 0.68 | DAVID VAN DER GULIK | 2010-2011 | COL | 0.64 |
| 2 | PETER FORSBERG | 2005-2006 | PHI | 0.68 | DAVID BOOTH | 2012-2013 | VAN | 0.63 |
| 3 | PETER FORSBERG | 2003-2004 | COL | 0.68 | DANIEL SEDIN | 2012-2013 | VAN | 0.62 |
| 4 | PETER FORSBERG | 2006-2007 | PHI | 0.68 | ALEXANDER SEMIN | 2003-2004 | WAS | 0.61 |
| 5 | PETER FORSBERG | 2007-2008 | COL | 0.68 | DANIEL SEDIN | 2010-2011 | VAN | 0.60 |
| 6 | PETER FORSBERG | 2010-2011 | COL | 0.68 | MIKHAIL GRABOVSKI | 2010-2011 | TOR | 0.60 |
| 7 | ONDREJ PALAT | 2013-2014 | TAM | 0.64 | DANIEL SEDIN | 2007-2008 | VAN | 0.60 |
| 8 | ONDREJ PALAT | 2012-2013 | TAM | 0.64 | DANIEL SEDIN | 2008-2009 | VAN | 0.60 |
| 9 | TYLER TOFFOLI | 2013-2014 | LOS | 0.63 | DANIEL SEDIN | 2011-2012 | VAN | 0.60 |
| 10 | TYLER TOFFOLI | 2012-2013 | LOS | 0.63 | PATRIK ELIAS | 2010-2011 | NJD | 0.60 |
| 11 | VINCENT LECAVALIER | 2006-2007 | TAM | 0.61 | SIDNEY CROSBY | 2013-2014 | PIT | 0.60 |
| 12 | VINCENT LECAVALIER | 2003-2004 | TAM | 0.61 | DANIEL SEDIN | 2009-2010 | VAN | 0.60 |
| 13 | SIDNEY CROSBY | 2009-2010 | PIT | 0.60 | JUSTIN WILLIAMS | 2010-2011 | LOS | 0.60 |
| 14 | SIDNEY CROSBY | 2008-2009 | PIT | 0.60 | DANIEL SEDIN | 2013-2014 | VAN | 0.60 |
| 15 | SIDNEY CROSBY | 2005-2006 | PIT | 0.60 | PATRIC HORNQVIST | 2013-2014 | NSH | 0.60 |
| 16 | PAVEL DATSYUK | 2007-2008 | DET | 0.60 | PAVEL DATSYUK | 2012-2013 | DET | 0.60 |
| 17 | PAVEL DATSYUK | 2008-2009 | DET | 0.60 | ALEX STEEN | 2011-2012 | STL | 0.60 |
| 18 | SIDNEY CROSBY | 2006-2007 | PIT | 0.60 | BRAD RICHARDSON | 2011-2012 | LOS | 0.60 |
| 19 | MATT MOULSON | 2011-2012 | NYI | 0.60 | ERIC FEHR | 2008-2009 | WAS | 0.60 |
| 20 | JOE THORNTON | 2009-2010 | SJS | 0.60 | TYLER TOFFOLI | 2013-2014 | LOS | 0.60 |

and much else.

## Wrap-up

The model is very transparent and easy to estimate.

Check out `gamlr` and the `hockey` example:

```
data(hockey)
x <- cBind(config,team,player)
y <- goal$homegoal
fit <- gamlr(x, y, free=1:(ncol(config)+ncol(team)),
                   standardize=FALSE, family="binomial")
```

We take no stand about what stats lead to wins,
and don't attempt to model full game action.

But this simple regression tells you a lot about who is contributing.

**Thanks!**