Bayesian techniques in the SALO shot model

Gordon Arsenoff

2017-10-19

In 2017, why present a model of only *some* hockey events?

SALO

- Shot rates

Corsica WAR

- Shot rates
- Shot quality
- Penalty rates
- Zone changes

Zooming in on one event to highlight useful methods

1. Using more known knowns (informed priors)
2. Using more known unknowns (don't just optimize; sample!)
3. Putting them together in applications

Priors: *what we already know*, as math

Players are average on average

- Regularized models (e.g. Corsica WAR) penalize large estimates
- Penalty shrinks small-N estimates back toward the mean
- Penalty is a Bayesian prior distribution of ability

Priors: *what we already know*, as math

Players are average on average

- Regularized models (e.g. Corsica WAR) penalize large estimates
- Penalty shrinks small-N estimates back toward average
- Penalty is a Bayesian prior distribution of ability

Better players get more games

- Robinson (2016): regulars are better than replacements!
- Small-N estimates should be shrunk toward **below** average
- SALO adds a model of games played by ability level

Executive summary of the SALO model

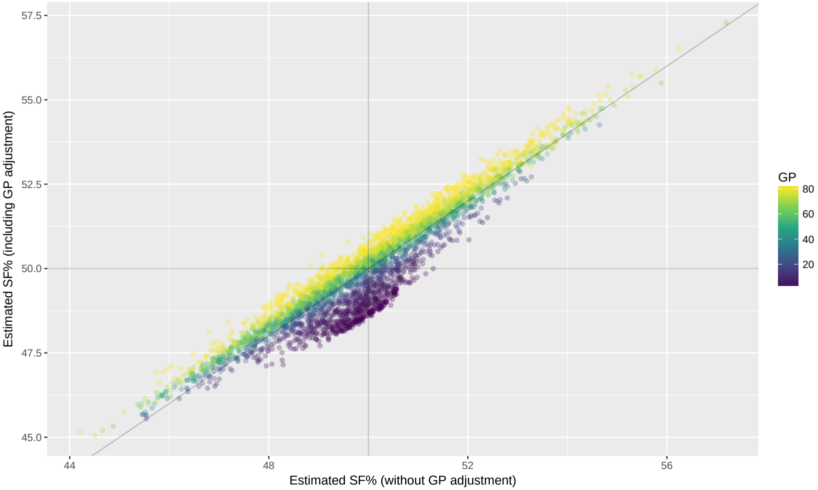
Model terms

- Ordered logit for (net) SoG each second vs. on-ice players
- Gaussian prior on player coefficients (L2 regularization)
- **Beta-binomial regression for games played vs. ability**

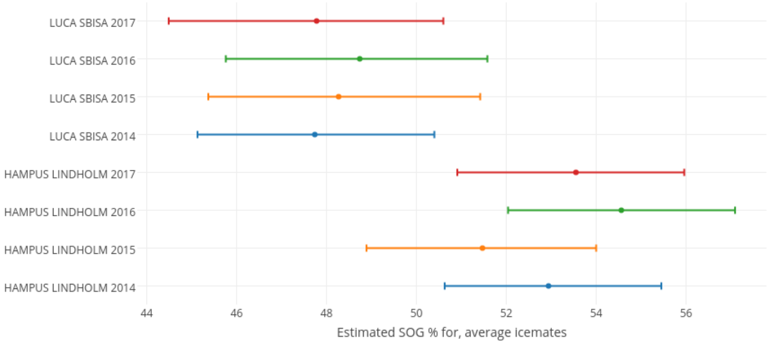
Algorithm

- Fitted with a Monte Carlo method (`http://mc-stan.org`)

Games-played term passes sanity checks



Results presented at `https://www.salohockey.net`



Retaining uncertainty about estimates in applications

- Estimates have error, but applications use numbers, not ranges
- Sample many **plausible** parameter values instead of just one (King, Tomz, and Wittenberg (2000))
- Monte Carlo methods directly yield plausible values
- Other methods permit post hoc sampling, depending on model

About how good is a typical NHL skater with 0 GP?

- SALO has an idea of NHL skater ability at any GP
- What if we plug in zero games played?
- Easy to draw plausible values with **rejection sampling**

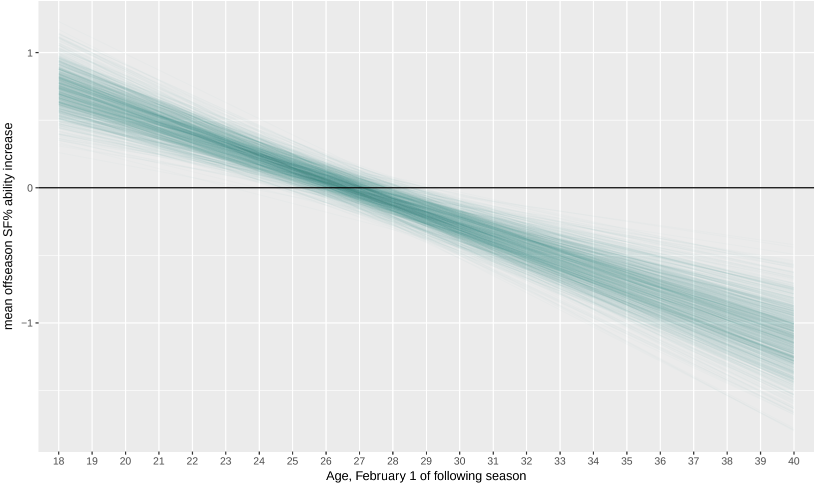
Value above model-based (less arbitrary) replacement

- Natural replacement player: a typical NHL skater with 0 GP
- For every Monte Carlo sample:
  1. Draw a plausible replacement player ability
  2. Figure expected wins for players and the replacement

Aging curves with survivor bias correction

- Parametric delta method: regress ability change on age
- Solberg (2017): delta method suffers survivor bias
- Give non-survivors **phantom** ability values to mitigate bias
- Natural phantom player: a typical NHL skater with 0 GP

Aging curves with survivor bias correction



Future work: faster, deeper, broader

SALO isn't fast enough for prime time

- Re-code as survival model, not ordered logit?
- Alternatives to Stan for efficient MC fitting?

More event types and context features

- Terms in the likelihood: score effects, event effects
- Distinguish games injured from healthy games not played
- Prior on year-to-year ability change (built-in aging curve)
- Probabilities of more events: penalties, zone changes

Takeaways

1. Use more of what we already know via priors
2. Use more of what we don't via plausible values
3. Derive useful applications from the model itself

Thanks!

Data and modeling choices

- All regular-season games from 2013-14 to 2016-17
- All situations; dummy variables for man advantages
- All data fit at once (i.e. prior constant across years)
- Outcome is SoG, but others (e.g. Corsi) would work fine

Ordered logit vs. (usual) survival model for shot rate

Advantages of ordered logit

- Far simpler in concept to code from scratch
- One parameter per player for both offence and defense
- Trivial to convert to readable shots-for percentage
- Greatly simplifies certain future applications

Disadvantages of ordered logit

- One data point per second demands lots of time and RAM!

Beta-binomial model

The beta-binomial distribution

- Like unto binomial distribution with overdispersion
- For correlated outcomes (e.g. playing yesterday and tomorrow)
- Probability of success assumed distributed Beta(α, β)
- Mean $\mu = \alpha/(\alpha+\beta)$ and precision $\phi = \alpha + \beta$
- Approaches binomial distribution as $\phi \to \infty$

Beta-binomial regression

- Logistic link: $\mu = \Lambda(\gamma + X\delta)$
- Caveat: in SALO, covariates X are ability parameters, not data!

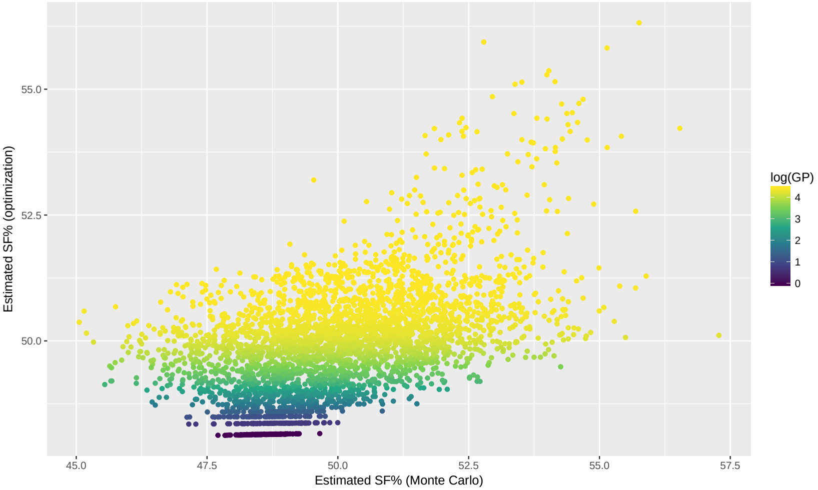
Basic model findings

- Standard deviation of SF% talent about 2.4%
- NHL-average player expects about 49.8 games (skewed much?)
- Player 1 SD above average expects about 57.8 games
- Intraplayer correlation of lineup inclusion about 0.457
- Home teams expect about 1.14 more SoG per 60

Top player-years

player	season	age	SF%	SF% sd
PAVEL DATSYUK	2016	37	57.28	1.38
PATRICE BERGERON	2015	29	56.53	1.48
PATRICE BERGERON	2016	30	55.89	1.49
JOE THORNTON	2014	34	55.75	1.33
ARTEMI PANARIN	2017	25	55.69	1.59
PATRICE BERGERON	2017	31	55.68	1.58
LOGAN COUTURE	2014	24	55.49	1.43
JORDAN STAAL	2016	27	55.41	1.47
CARL HAGELIN	2016	27	55.38	1.29
DANIEL SEDIN	2014	33	55.28	1.47

Optimization of weird models gives weird results



Rejection sampling the prior for a 0 GP skater

Given plausible parameters of the prior terms:

1. Draw a random value from the Gaussian distribution
2. Accept with probability given by the beta-binomial term

Algorithm for projection of next year's abilities

1. Identify all non-survivors
   - Players with 0 GP in a year with > 0 GP the year before
   - Players with 0 GP in a year with > 0 GP the year after
2. For each Monte Carlo sample:
   - 2.1 Rejection sample a phantom value for every non-survivor
   - 2.2 Regress year-to-year ability change on age
   - 2.3 Draw a plausible value of the regression coefficients
   - 2.4 Draw a plausible ability change for each current-year player

References I

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science*, April. `http://www.jstor.org/stable/2669316`.

Robinson, David. 2016. "Understanding Beta Binomial Regression (Using Baseball Statistics)." *Variance Explained*. `http://varianceexplained.org/r/beta_binomial_baseball`.

Solberg, Luke. 2017. "A New Look at Aging Curves for NHL Skaters, Part 2." *Hockey Graphs*. `https://hockey-graphs.com/2017/04/10/a-new-look-at-aging-curves-for-nhl-skaters-part-2`.