Big Data: Homework 3

Will Clark & Matthew DeLio
41201-01

April 23, 2015

1 Player Contribution Regression

The model for player contribution is:

$$\log\left[\frac{Pr(y=1)}{1-Pr(y=1)}\right] = \beta_0 + \alpha_{\texttt{team,season}} + \alpha_{\texttt{config}} + \sum_{\texttt{homeplyr}} \beta_j + \sum_{\texttt{awayplyr}} \beta_j$$

For a given player j, the model estimates the odds multiplier that a goal scored while player j is on the ice was scored by the his team. It includes the following two control factors:

- `team,season`: This should control for high- or low-offense years, certain arenas that provide a special home-ice advantage, or coaches that are better/worse than average; and
- `config`: This should control for disproportionate playing time in power plays or end-of-game situations where the goalie has been pulled.

To use one example, the coefficient on Alex Ovechkin is 0.30. This means that a goal scored while Ovechkin is on the ice is $\exp(0.30) = 1.35$ times as likely to be scored by his team, the Washington Capitals, than by their opponents. Put differently, if a goal is scored while Ovi is on the ice, it is 35 percent more likely that it is scored by the Caps than by their opponents.

We can sort the array of player coefficients to determine the 10 most and least valuable players in the data set. We show the results in Table 1 and Table 2. This evaluation metric accords with our intuition about hockey. The list of 10 best players includes some of the conventionally-regarded best players of the last decade, which tells us the model is doing a reasonably good job of quantifying performance. It also includes one player (Tyler Toffoli) who has only played since 2013, although his brief career has been successful to date. Table 1 and Table 2 have a column labeled **G**, displaying the number of goals a player was on the ice for. We can see that Toffoli is indeed an inexperienced player compared to veterans like Joe Thornton and Pavel Datsyuk

By this performance metric, the best and worst players are both outliers and most players have a 0 rating. The sample includes 2439 players and only 646 have non-zero ratings (390 have net positive and 256 have net negative ratings). We can see in Figure 1 that only a small handful of players are significantly better or significantly worse than average.

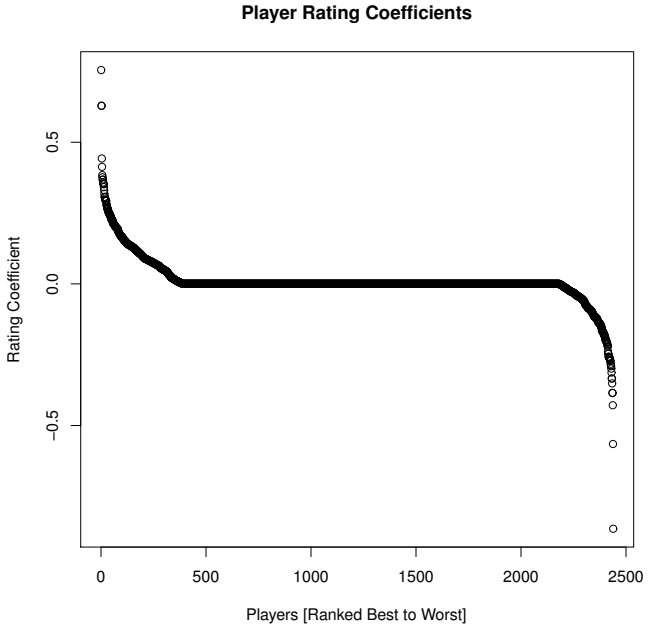


Figure 1: Player Ratings Based on Regression Estimate

2 Gamma-Lasso Regression Standardization

The Gamma-Lasso regression estimates the coefficients β_j by minimizing an objective function similar to:

$$\min_{\lambda} \left(-\frac{2}{n} \log \text{LHD}(\beta) + \lambda \sum_j c(\beta_j) \right)$$

For this discussion, we will assume that the cost function is a simple lasso of the form $c(\beta_j) = |\beta_j|$. In most cases we would scale this penalty function by the standard deviation of x_j, but doing so in this example would penalize players who have been on the ice for a large number of goals.

We ran the Gamma-Lasso regression setting `standardize=TRUE` and the resulting list of the 10 best players had been on the ice for a combined 31 goals between them. We can look more closely at two example players, Pavel Datsyuk and Jeff Toms, to see how the standardized penalty functions led to this outcome.

Datsyuk is a long-tenured player who has been on the ice for 1725 goals; he is ranked 7th by our logit regression model. Toms had a long career in the NHL and minor leagues but played his last NHL game in 2002, seeing action in only eight games and being on the ice for only four goals (which are the only four included in our data set). For Datsyuk, $\text{var}(x_j) = 0.0248$ while for Toms, $\text{var}(x_j) = 5.7597\text{E}{-5}$.

Figure 2 shows what the lasso penalty functions look like for each player *after* they have been scaled by the standard deviation of x_j. We can see that there is a very high cost for the coefficient on Pavel Datsyuk and almost no cost on the coefficient for Jeff Toms. What this means for the

	Player	Rank	β_j	$\exp(\beta_j)$	G
1	PETER FORSBERG	1	0.7548	2.1272	532
2	TYLER TOFFOLI	2	0.6293	1.8762	93
3	ONDREJ PALAT	3	0.6284	1.8746	140
4	ZIGMUND PALFFY	4	0.4427	1.5569	197
5	SIDNEY CROSBY	5	0.4131	1.5115	1568
6	JOE THORNTON	6	0.3838	1.4678	1740
7	PAVEL DATSYUK	7	0.3762	1.4567	1725
8	LOGAN COUTURE	8	0.3682	1.4451	513
9	ERIC FEHR	9	0.3677	1.4444	369
10	MARTIN GELINAS	10	0.3578	1.4301	460

Table 1: Top 10 NHL Players (2002-2014)

	Player	Rank	β_j	$\exp(\beta_j)$	G
1	RYAN HOLLWEG	2430	-0.2989	0.7417	78
2	RAITIS IVANANS	2431	-0.3129	0.7313	81
3	DARROLL POWE	2432	-0.3340	0.7161	337
4	CHRIS DINGMAN	2433	-0.3342	0.7159	30
5	MATHIEU BIRON	2434	-0.3512	0.7038	203
6	THOMAS POCK	2435	-0.3844	0.6809	131
7	NICLAS HAVELID	2436	-0.3855	0.6801	1041
8	P. J. AXELSSON	2437	-0.4284	0.6516	121
9	JOHN MCCARTHY	2438	-0.5652	0.5683	45
10	TIM TAYLOR	2439	-0.8643	0.4213	148

Table 2: Bottom 10 NHL Players (2002-2014)

estimation is that the coefficient on Datsyuk is very likely to be small, and the coefficient on Toms is very likely to be larger. This is in fact what we see from running the Gamma-Lasso regression setting `standardize=TRUE`; the coefficient on Datsyuk is 0.2908 while the coefficient on Toms is 1.7381.

3 IC and CV Model Selection

For this discussion we are looking at 3 information criteria, AIC, AICc, and BIC, used in model selection. To compare these different criteria, we employ a 5-fold cross-validation to provide an estimate of the Out-of-Sample error (see Figure 3a on page 5). Included on that figure is the mean-squared error for the different folds across many different values for λ. The λ yielding the minimum average mean-squared error as well as the largest one yielding an average MSE no more than 1 standard error away from the minimum. Choosing the value of λ_{1se} will yield a simpler model (i.e. one with less coefficients), but one with, potentially, a higher mean-squared error for any new data samples.

Since cross-validation is expensive (in terms of compute-time), the goal is to find another information criteria that will approximate the λs provided by the cross-validated model. Figure 3b on the

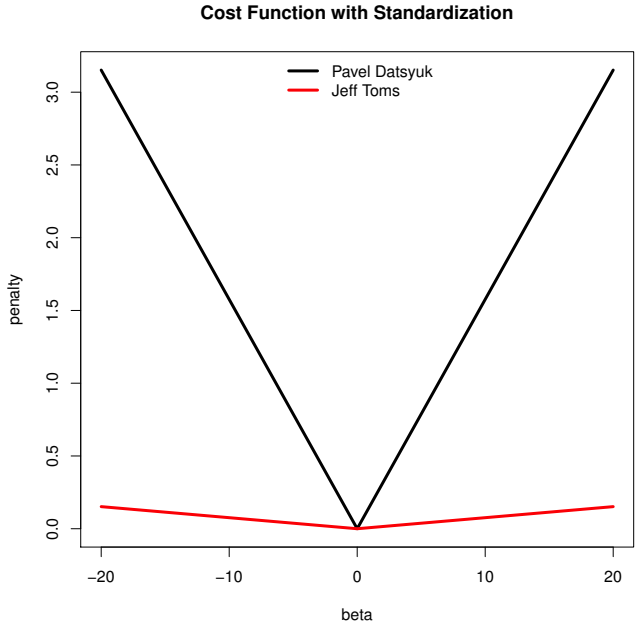


Figure 2: Lasso Cost Functions

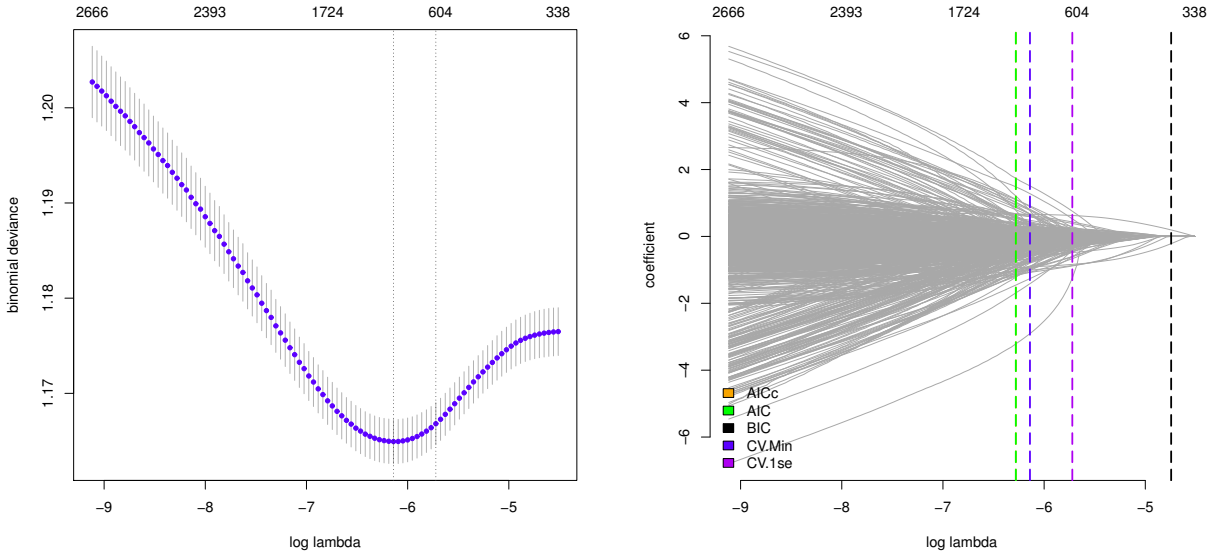
next page shows the 'gamlr' coefficient path plot with varying λs. The λs that minimize AIC, AICc, and BIC as well as CV.min and CV.1se are also shown in the verticals on the plot (see Table 3 for the numerical values). *Note: In Figure 3b on the next page that AIC and AICc are almost identical therefore one is obscured by the other.* From these data we see that both AIC/AICc do a reasonable job at estimating CV.min, opting for a more complex model at the expense of slightly worse MSE for new samples. The BIC chooses a much simpler model, with far fewer coefficients than even the one prescribed by CV.1se. Because the MSE for new samples will likely be higher than that chosen by AIC and AICc, BIC is likely inferior to these other information criteria.

	$log(\lambda)$
AICc	-6.28
AIC	-6.28
BIC	-4.74
CV.Min	-6.14
CV.1se	-5.72

Table 3: ICs for NHL Data

4 Removing Team/Special Play From the Regression

In this section we explore what happens when the team, season and special play effects are removed from the model. Very similar to the original model, the new model is therefore governed by:



(a) (b)

Figure 3

$$\log\left[\frac{Pr(y=1)}{1-Pr(y=1)}\right] = \beta_0 + \sum_{\texttt{homeplyr}} \beta_j + \sum_{\texttt{awayplyr}} \beta_j$$

With this new model, we run the Gamma-Lasso regression as before as well as the 5-fold CV lasso. With the default choice of *lambda.min.ratio* we noticed that the CV average MSE never displays a minima. To satisfy our curiosity, we lowered this parameter to $\exp -10$ and re-ran the CV lasso. After re-running it, we found a minima at $log(\lambda) = -9.46$, a value which includes a majority of the covariates in the final model (see Figure 4a on page 7). Next we ran the Gamma-Lasso regression with the same *lambda.min.ratio* and found the AIC, AICc, and BIC information criteria for the run (see Figure 4b on page 7 and Table 4 on the following page).

We find that the AIC and AICc never reach a minima; therefore with these selection criteria, they choose all covariates for the model. ¡TRANSITION¿ Deviance always decreases as more degrees of freedom are added. ¡Explain why minima occurs¿, ¡Explain why no minima occurs¿. ¡Explain that these results show that our model is either not idea, or the ideal model really does include all covariates¿ ¡Since we have the other model, it is clear that we have become victim of Omitted variable bias¿.

$$AICc = \texttt{Deviance} + 2df\frac{n}{n-df-1} \tag{1}$$

$$AIC = \texttt{Deviance} + 2df \tag{2}$$

$$BIC = \texttt{Deviance} + df * ln(n) \tag{3}$$

1. The AIC and AICc have no minima (see Figure 5 on the next page) selecting all of the covariates for the model.
2. With the exception of the BIC, the CV lasso and the other ICs select almost all of the covariates for the model.

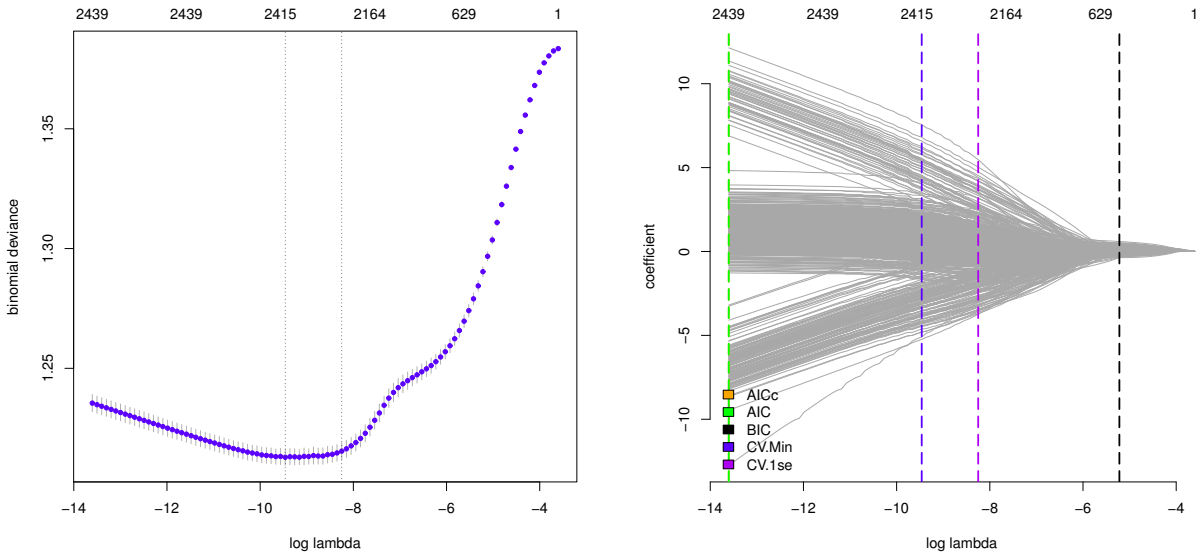
Will wrap this up tomorrow morning, but basically will say:

1. we had to lower lambda
2. we think that the model is bad and the various ICs show it (and a discussion of what we noted today)
3. the AICs rightly indicate that we should just include all of the covariates

1. Bad Model
2. $AICc = DEV + \frac{n}{n-df-1} 2 * df$
3. As we add degrees of freedom always expect deviance to go down
4. As we add degress of freedom expect rhs to go up
5. In a good model, the deviance decreases at a slower rate than the IC "penalty" before the minimum
6. In a bad model, the deviance decreases at a faster rate than the penalty (each degree of freedom helps that much more)

	$log(\lambda)$
AICc	-13.60
AIC	-13.60
BIC	-5.22
CV.Min	-9.46
CV.1se	-8.25

Table 4: ICs for Player-Only Data



(a) (b)

Figure 4

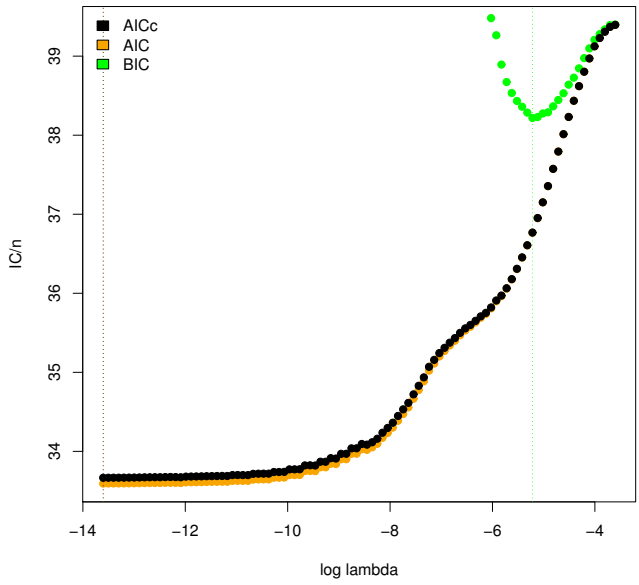


Figure 5