Estimating the evolution of scoring rates in soccer match: Model selection and variable selection with ???-LASSO

Mladen Laudanovic and Frank Wood

$$\begin{aligned}
\text{BASELINE MODEL 0: } & y_{id} \mid X_{id}, \beta \sim \text{Pois}\left(\mu\alpha_i\right) \\
\text{BASELINE MODEL 1: } & y_{id} \mid X_{id}, \beta \sim \text{Pois}\left(\lambda_{id}\right) \\
\text{BASELINE MODEL 2: } & y_{id} \mid X_{id}, \beta \sim \text{Pois}\left(\alpha_i\lambda_{id}\right) \\
\text{BASELINE MODEL 3: } & y_{id} \mid X_{id}, \beta \sim \text{Pois}\left(\mu\alpha_i\right) \\
\text{MODEL 1: } & y_{id} \mid X_{id}, \beta \sim \text{Pois}\left(\lambda_{id}\right) \\
\text{MODEL 1': } & y_{id} \mid X_{id}, \beta \sim \text{Pois}\left(\mu\lambda_{id}\right) \\
\text{MODEL 2: } & y_{id} \mid X_{id}, \beta \sim \text{Pois}\left(\alpha_i\lambda_{id}\right) \\
\text{MODEL 2': } & y_{id} \mid X_{id}, \beta \sim \text{Pois}\left(\mu\alpha_i\lambda_{id}\right)
\end{aligned}$$

where α_i are market intensities and $\lambda_{id} = g(\beta^T X_{id})$ and also

$$\text{MODEL 3': } y_{id} \mid X_{id}, \beta \sim \text{Pois}\left(\alpha_i\lambda_{id}\right)$$

where α_i are unknown intensities. In this case likelihood is

$$\prod_i \prod_d \alpha_i^{y_{id}} \lambda_{id}^{y_{id}} e^{-\alpha_i\lambda_{id}}/y_{id}! = \prod_i \alpha_i^{\sum_d y_{id}} \prod_d \lambda_{id}^{y_{id}} e^{-\alpha_i\lambda_{id}}/y_{id}!$$

so $n_i = \sum_d y_{id}$ is sufficient statistic for α_i, with the distribution $n_i \sim \text{Pois}\left(\alpha_i \sum_d \lambda_{id}\right)$. Likelihood for n_i is

$$\prod_i \alpha_i^{n_i} \left(\sum_d \lambda_{id}\right)^{n_i} e^{-\alpha_i \sum_d \lambda_{id}}/n_i!$$

hence after conditioning we get:

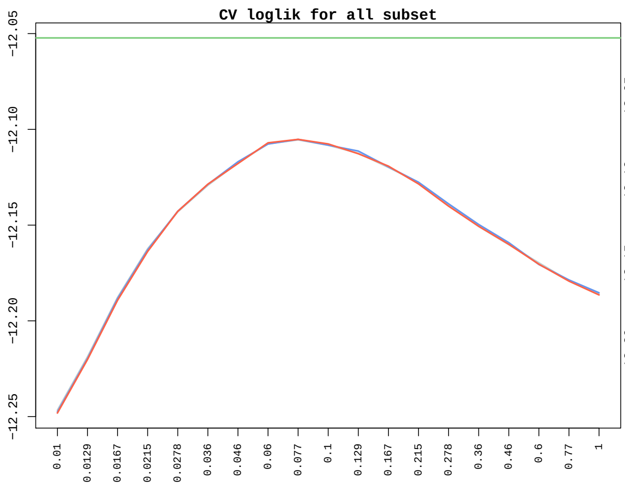
$$\text{MODEL 3: } y_{i.} \mid \beta, n_i, X_{i.} \sim \text{M}\left(n_i, \left\{\frac{\lambda_{id}}{\sum_{d^\star} \lambda_{id^\star}}\right\}_{d=1}^{D}\right)$$

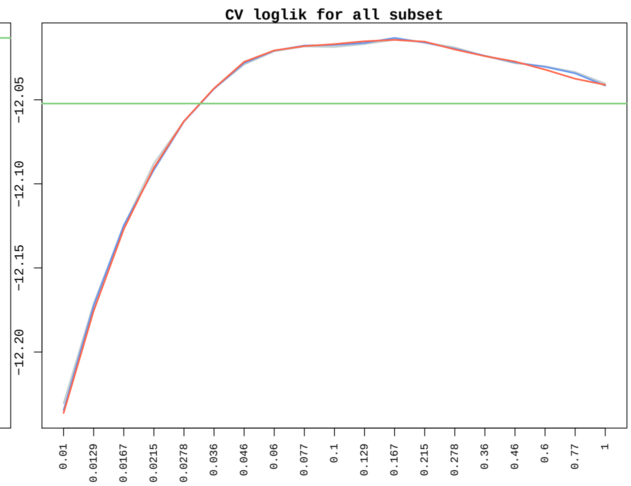
– d are minutes in a match (1-44, 46-89 for both teams; D = 176 per match)
– i are matches ($I = 302$)
– g is a link function. There are three of them: $e^x, \log(1+e^x), \log(\frac{1}{1+e^{-x}})$

Three models with three different link functions were tested. There were two subsets of variables, each one having 100 of them (selected by "quality", one without other with market variables)

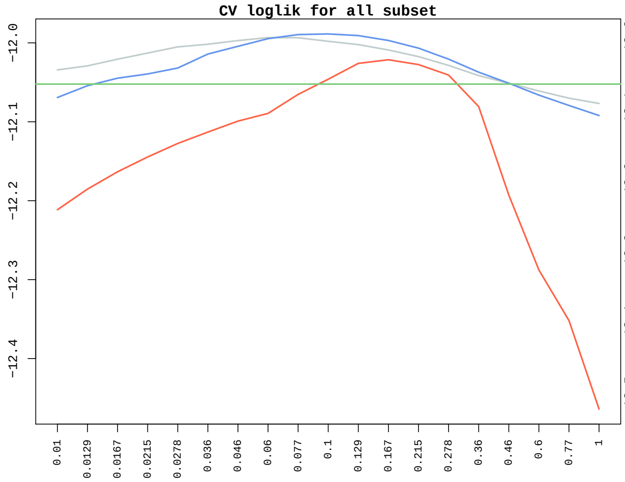
– K number of variables (K=100)

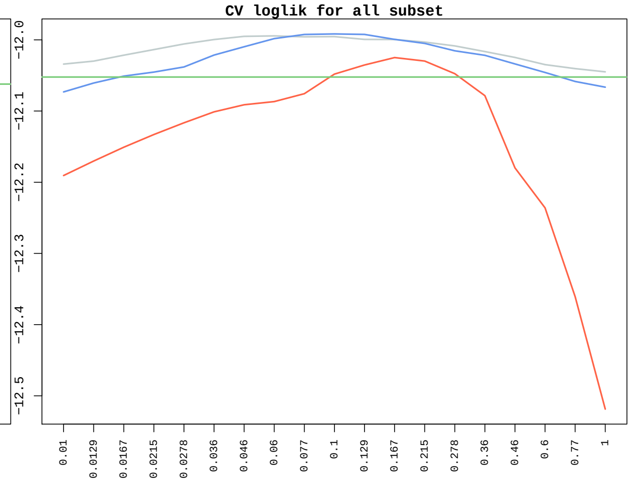
MODEL1: $y_{id} \mid X_{id}, \beta \sim \text{Pois}\left(\lambda_{id}\right)$



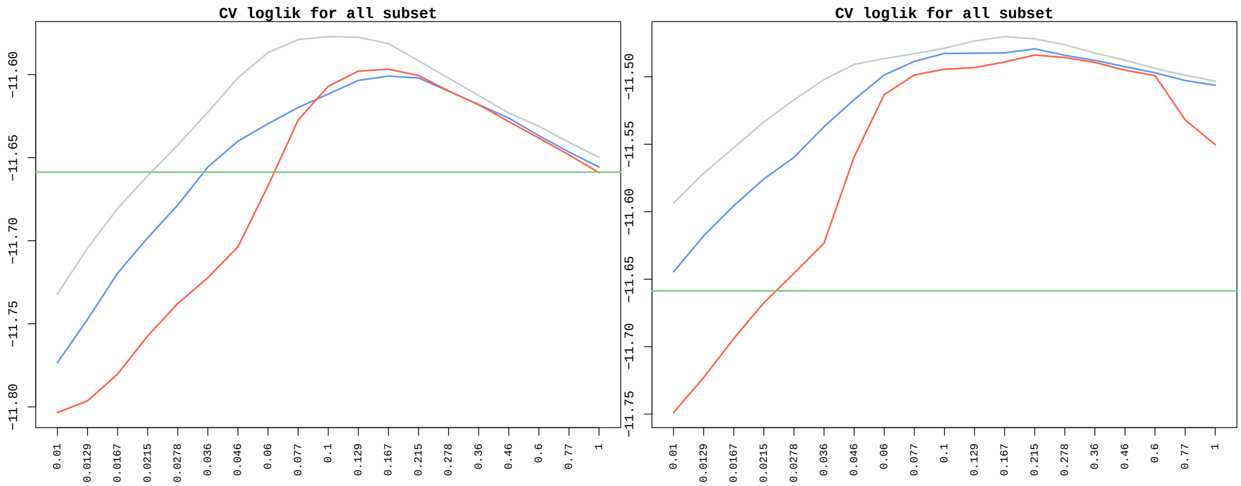


MODEL2: $y_{id} \mid X_{id}, \beta \sim \text{Pois}\left(\alpha_i \lambda_{id}\right)$





$$\text{MODEL3: } y_{i.} \mid \beta, n_i, X_{i.} \sim \mathrm{M}\left(n_i, \left\{\frac{\lambda_{id}}{\sum_{d^\star} \lambda_{id^\star}}\right\}_{d=1}^{D}\right)$$



INTRODUCTION: BASIC MODEL

$$y_{id} \mid X_{id}, \beta \sim \text{Pois}\left(g(\beta^T X_{id})\right)$$

$$\text{P}(\beta \mid X, \text{Data}) \propto \text{P}(\beta) \prod_i \prod_d g(\beta^T X_{id})^{y_{id}} e^{-g(\beta^T X_{id})} / y_{id}!$$

– d are minutes in a match (1-44, 46-89 for both teams; D = 176 per match)
– i are matches ($I = 302$)
– g is a link function ($\exp(x)$ is default)
– K number of variables (length of beta)

$$\ell(\beta) = \log(\text{P}(\beta)) + \sum_{l=(i,d)} \left[y_l \log\left(g(\beta^T X_l)\right) - g(\beta^T X_l)\right] + \text{const}$$

is it convex with respect to beta and lambda???

In coordinatewise gradient descent algorithm in each step $k \in 1, \dots K$ we are maximizing ℓ as a function of one parameter $z = \beta_k$. It is fine for convex functions.

So goal is to minimize

$$h(z) = \sum_l \left[g(r_l + (z - \beta_k)x_{lk}) - y_l \log\left(g(r_l + (z - \beta_k)x_{lk})\right)\right] + z^2/2\tau$$

where $r_l = \beta^T X_l$ and τ is regularization parameter.

One good way to find the maximum of ℓ [David's paper, at al.] is that in each step we find maximum of 2nd degree Tailor polynomial of h, so suggested shift for β_k will be $\Delta u_k = -h'(\beta_k)/h''(\beta_k)$ where

$$h'(\beta_k) = \sum_l \left[g'(r_l)x_{lk} - y_l x_{lk} g'(r_l)/g(r_l)\right] + \beta_k/\tau$$

$$h''(\beta_k) = \sum_l \left[g''(r_l)x_{lk}^2 - y_l x_{lk}^2 \frac{g''(r_l)g(r_l) - g'(r_l)^2}{g(r_l)^2}\right] + 1/\tau$$

SHAWN'S MODEL

Pros and cons

$$y_{id} \mid X_{id}, \beta \sim \text{Pois}\left(\alpha_i g(\beta^T X_{id})\right)$$

$$P(\beta \mid X, \text{Data}) \propto P(\beta) \prod_i \prod_d \left(\alpha_i g(\beta^T X_{id})\right)^{y_{id}} e^{-\alpha_i g(\beta^T X_{id})} / y_{id}!$$

Since $n_i = \sum_d y_{id}$ is sufficient statistic for α_i we should maximize

$$P(\beta \mid X, \text{Data}, n) = P(\beta) \prod_i \prod_d \left(\frac{g(\beta^T X_{id})}{\sum_{d^\star} g(\beta^T X_{id^\star})} \right)^{y_{id}}$$

$$\ell(\beta) = \log(\text{P}(\beta)) + \sum_i \sum_d y_{id} \log g(\beta^T X_{id}) - \sum_i n_i \log \textstyle\sum_{d^\star} g(\beta^T X_{id^\star}) + \text{const}$$

$$h(z) = \sum_i n_i \log \textstyle\sum_d g(r_{id} + (z - \beta_k) x_{idk}) - \displaystyle\sum_{i,d} y_{id} \log g(r_{id} + (z - \beta_k) x_{idk}) + z^2/2\tau$$

$$h'(\beta_k) = \sum_i n_i \frac{\sum_d g'(r_{id}) x_{idk}}{\sum_d g(r_{id})} - \sum_{i,d} y_{id} \frac{g'(r_{id}) x_{idk}}{g(r_{id})} + \beta_k/\tau$$

$$\begin{aligned} h''(\beta_k) = \sum_i n_i & \frac{\left(\sum_d g''(r_{id}) x_{idk}^2\right)\left(\sum_d g(r_{id})\right) - \left(\sum_d g'(r_{id}) x_{idk}\right)^2}{\left(\sum_d g(r_{id})\right)^2} \\ & - \sum_{i,d} y_{id} x_{idk}^2 \frac{g''(r_{id}) g(r_{id}) - g'(r_{id})^2}{g(r_{id})^2} + 1/\tau \end{aligned}$$