Estimating the Causal Effect of Injury on Performance

Tyrel Stokes

September 15, 2018

Oscar Klefbom: Impact of Injury on Performance?

- Missed the Last 52 Games of the 2015-2016 season
- Part-way through 2nd full season in NHL

Elective Surgery on the Other foot?

Season	TOI	P1/60	GS/60	CF%	xGF%
2014-2015	1621	0.48	0.96	50.05%	48.07%

Elective Surgery on the Other foot?

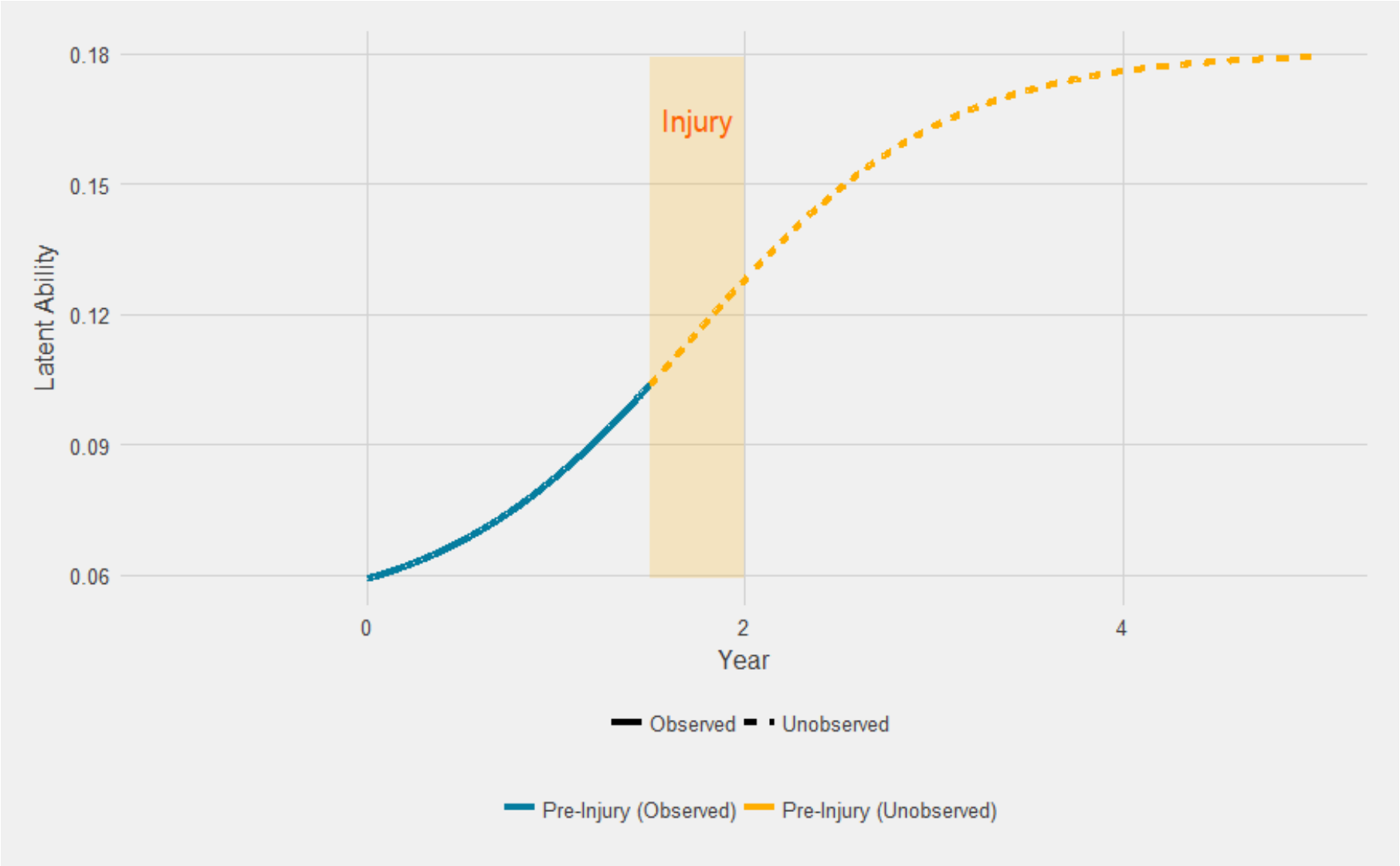
Season	TOI	P1/60	GS/60	CF%	xGF%
2014-2016	1621	0.48	0.96	50.05%	48.07%
2016-2017	1405	0.77	1.44	50.37%	51.13%

Elective Surgery on the Other foot?

Season	TOI	P1/60	GS/60	CF%	xGF%
2014-2016	1621	0.48	0.96	50.05%	48.07%
2016-2017	1405	0.77	1.44	50.37%	51.13%

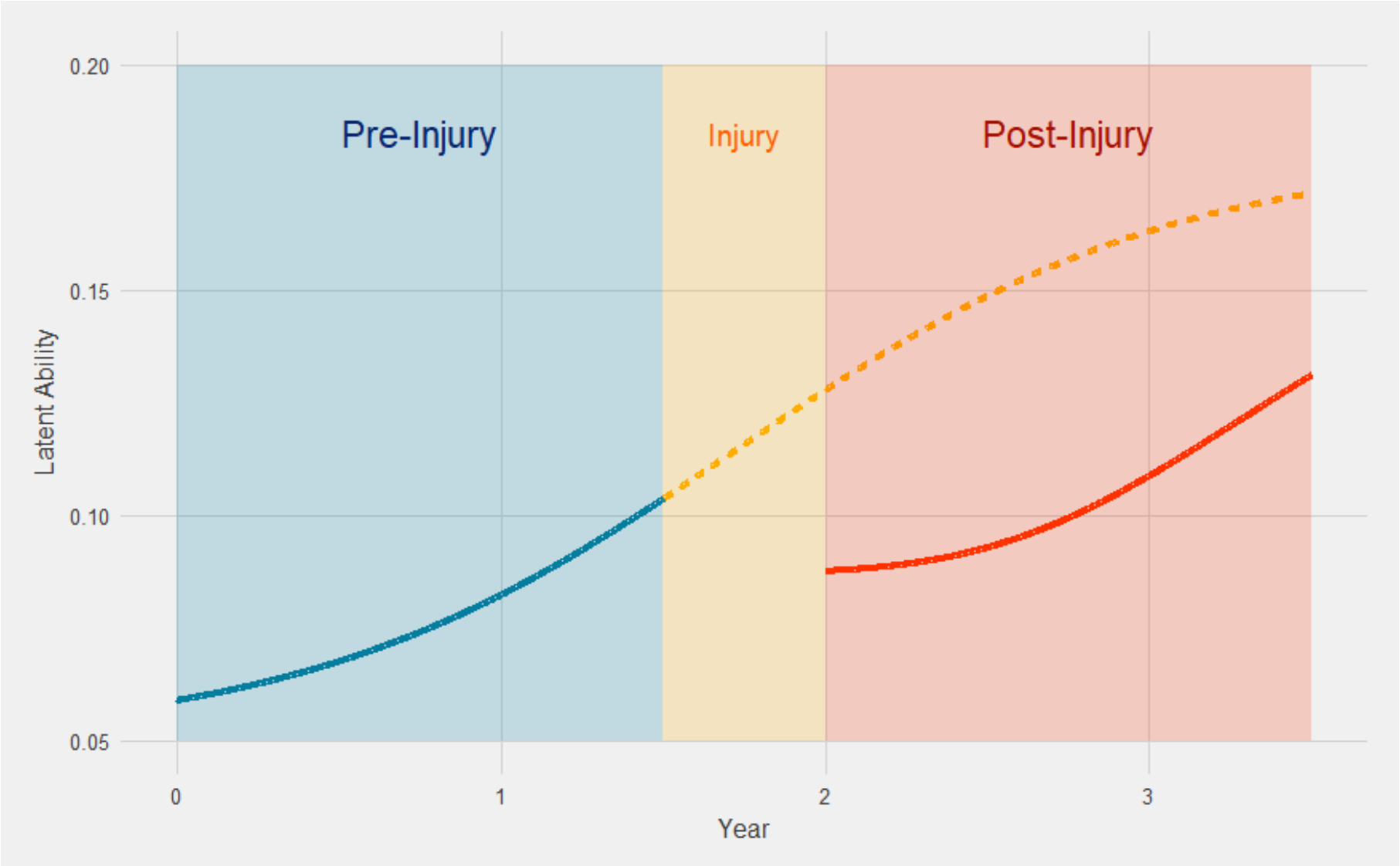
Injury *made* him better?

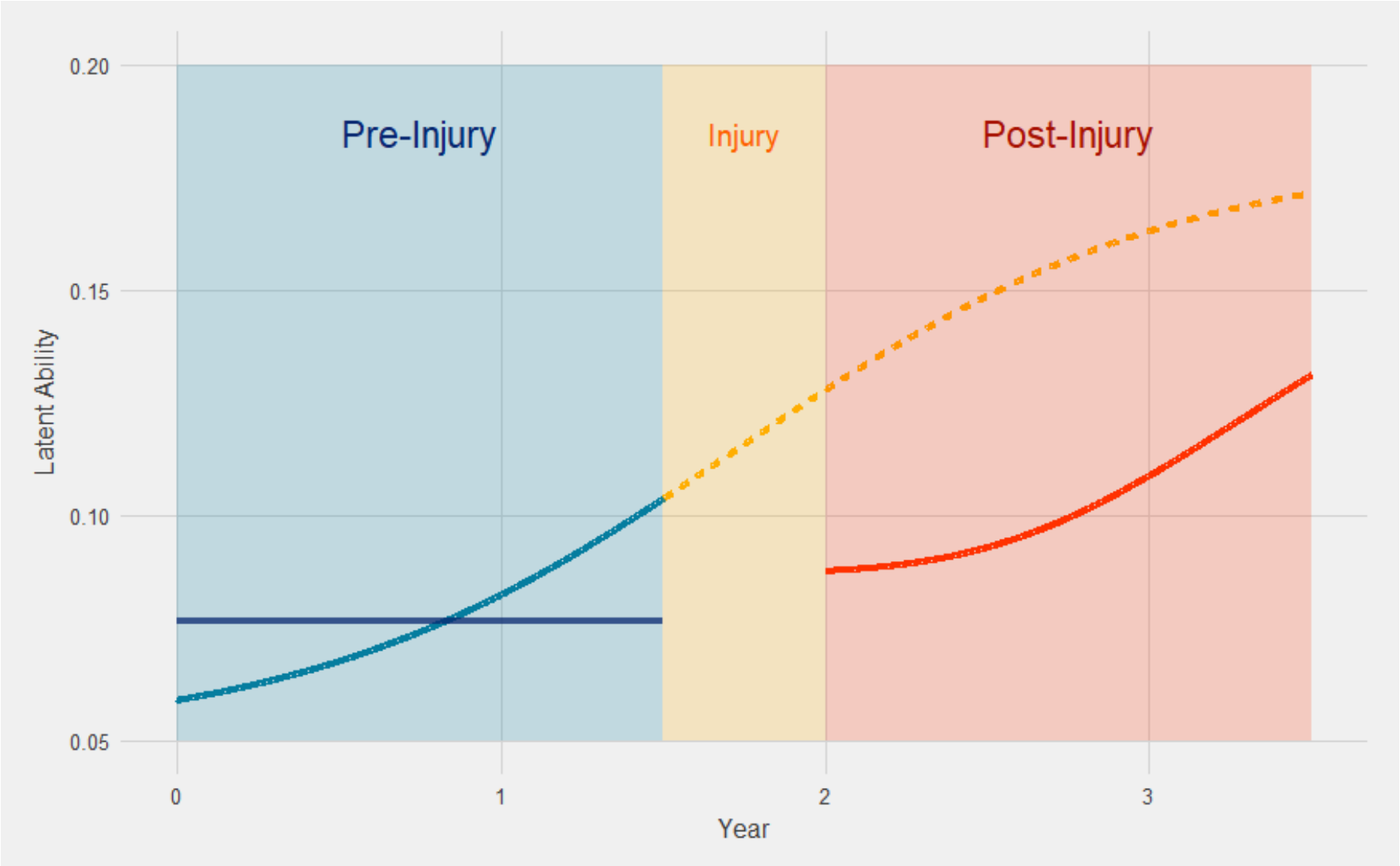
”Oscar Klefbom” Latent Performance Over Time

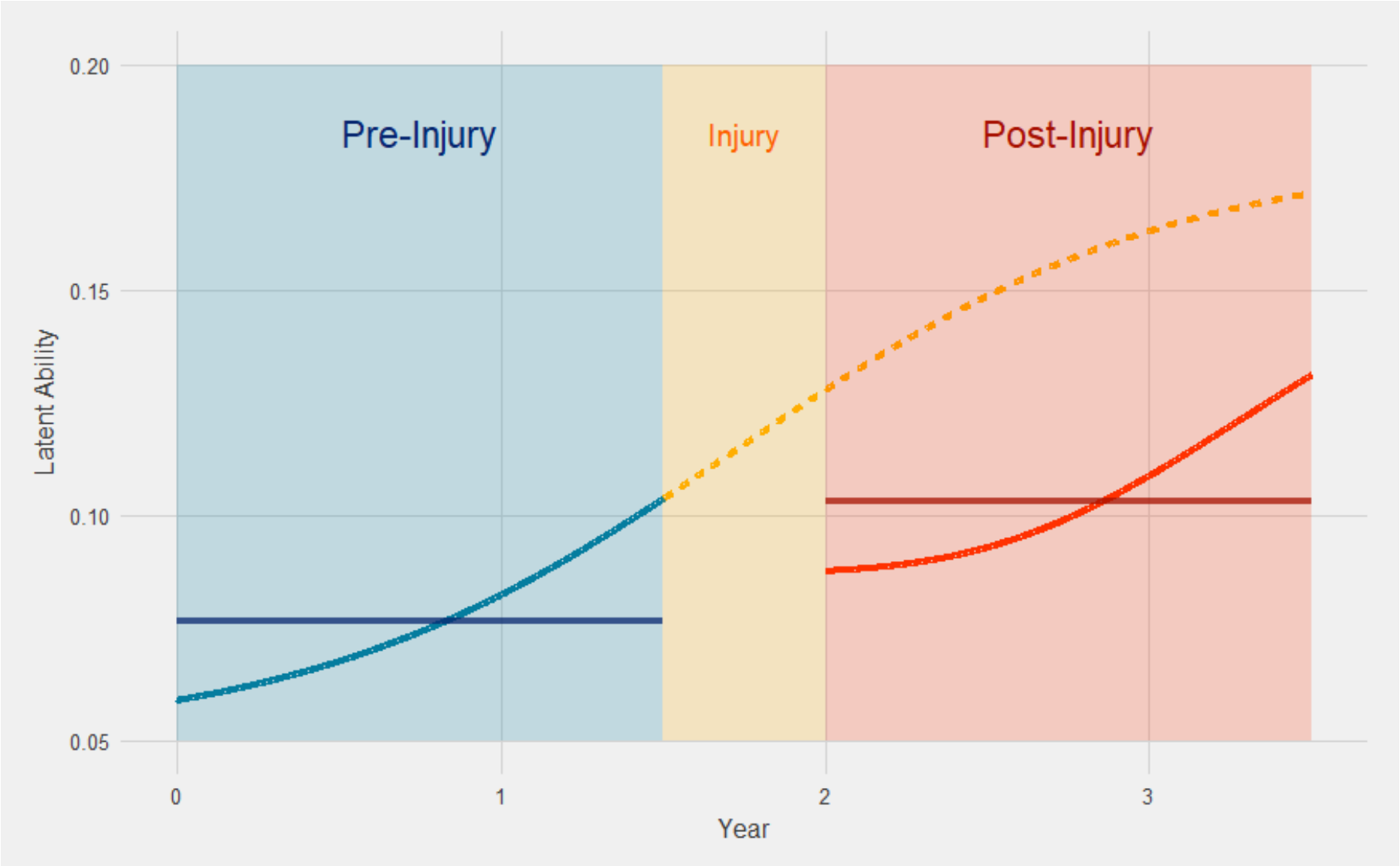


”Oscar Klefbom” Latent Performance Over Time

”Oscar Klefbom” Latent Performance Over Time


Pre-Injury
Injury
Post-Injury
0.20
0.15
0.10
0.05
Latent Ability
0
1
2
3
Year


Pre-Injury
Injury
Post-Injury
0.20
0.15
0.10
0.05
Latent Ability
0
1
2
3
Year


Pre-Injury
Injury
Post-Injury
0.20
0.15
0.10
0.05
Latent Ability
0
1
2
3
Year

Key Points

- Underlying Ability is relatively smooth
- Injuries typically have **large local effects**, small long-term effects

Problem

Regression models for Ability are noisy and typically require at least a season of data for accurate results!!!

Potential Solution

Specify a flexible performance path that allows us to borrow information from neighboring periods

Dynamic Paired Comparison Model with Stochastic Variances (Glickman 2001)

$$\log(L(\gamma, \sigma^2, \theta, \tau^2, \omega^2 | y, x)) =$$

$$\sum_{\mathbf{i=1}}^{\mathbf{I}} \log(\mathbf{N}(\gamma_{\mathbf{i}}^{(\mathbf{0})} | \mathbf{0}, \omega^{\mathbf{2}})) + \tag{1}$$

(2)

Dynamic Paired Comparison Model with Stochastic Variances (Glickman 2001)

$$\log(L(\gamma, \sigma^2, \theta, \tau^2, \omega^2 | y, x)) =$$

$$\sum_{i=1}^{I} \log(N(\gamma_i^{(0)} | 0, \omega^2)) \tag{3}$$

$$+\sum_{\mathbf{t=1}}^{\mathbf{T}}(\sum_{\mathbf{i<j}}\sum_{\mathbf{k=1}}^{\mathbf{K_{ij}^{(t)}}}[\mathbf{y_{ijk}^{(t)}} \log(\mathbf{F}(\gamma_{\mathbf{i}}^{(\mathbf{t})} - \gamma_{\mathbf{j}}^{(\mathbf{t})}, \mathbf{x_{ijt}}, \beta)) \tag{4}$$

$$+ (\mathbf{1} - \mathbf{y_{ijk}^{(t)}}) \log(\mathbf{1} - \mathbf{F}(\gamma_{\mathbf{i}}^{(\mathbf{t})} - \gamma_{\mathbf{j}}^{(\mathbf{t})}, \mathbf{x_{ijt}}, \beta))]+ \tag{5}$$

Dynamic Paired Comparison Model with Stochastic Variances (Glickman 2001)

$$\log(L(\gamma, \sigma^2, \theta, \tau^2, \omega^2 | y, x)) =$$

$$\sum_{i=1}^{I} \log(N(\gamma_i^{(0)} | 0, \omega^2)) \tag{6}$$

$$+ \sum_{t=1}^{T} (\sum_{i<j} \sum_{k=1}^{K_{ij}^{(t)}} [y_{ijk}^{(t)} \log(F(\gamma_i^{(t)} - \gamma_j^{(t)}, x_{ijt}, \beta)) \tag{7}$$

$$+ (1 - y_{ijk}^{(t)}) \log(1 - F(\gamma_i^{(t)} - \gamma_j^{(t)}, x_{ijt}, \beta))]$$

$$+ \sum_{\mathbf{t=0}}^{\mathbf{T-1}} \sum_{\mathbf{i=1}}^{\mathbf{I}} \log(\mathbf{N}(\gamma_{\mathbf{i}}^{(\mathbf{t+1})} | \gamma_{\mathbf{i}}^{(\mathbf{t})}, \sigma_{\mathbf{i}}^{\mathbf{2}^{(\mathbf{t+1})}})) + \tag{8}$$

$$\tag{9}$$

Dynamic Paired Comparison Model with Stochastic Variances (Glickman 2001)

$$\log(L(\gamma, \sigma^2, \theta, \tau^2, \omega^2 | y, x)) =$$

$$\sum_{i=1}^{I} \log(N(\gamma_i^{(0)} | 0, \omega^2)) \tag{10}$$

$$+ \sum_{t=1}^{T} (\sum_{i<j} \sum_{k=1}^{K_{ij}^{(t)}} [y_{ijk}^{(t)} \log(F(\gamma_i^{(t)} - \gamma_j^{(t)}, x_{ijt}, \beta)) \tag{11}$$

$$+ (1 - y_{ijk}^{(t)}) \log(1 - F(\gamma_i^{(t)} - \gamma_j^{(t)}, x_{ijt}, \beta))]$$

$$+ \sum_{t=0}^{T-1} \sum_{i=1}^{I} \log(N(\gamma_i^{(t+1)} | \gamma_i^{(t)}, \sigma_i^{2^{(t+1)}})) \tag{12}$$

$$+ \sum_{\mathbf{t=0}}^{\mathbf{T-1}} \sum_{\mathbf{i=1}}^{\mathbf{I}} \log(\mathbf{N}(\log \sigma_{\mathbf{i}}^{\mathbf{2}^{(\mathbf{t+1})}} | \log \sigma_{\mathbf{i}}^{\mathbf{2}^{(\mathbf{t})}}, \tau^{\mathbf{2}})) \tag{13}$$

Model Benefits

- Models Players as evolving with Random walk with potential for jumps
- Allows us to estimate player ability over a short period (say 1 month), without introducing too much noise.
- Similar models already in use for a variety of sports at the team level (Glickman and Stern (1998), Lopez, Matthews, and Baumer (2017))

Adaptation to modeling player ability

$$F((\gamma_{i1}^{(t)} + \gamma_{i2}^{(t)} + \gamma_{i3}^{(t)} + \gamma_{i4}^{(t)} + \gamma_{i5}^{(t)}) \\ -(\gamma_{j1}^{(t)} + \gamma_{j2}^{(t)} + \gamma_{j3}^{(t)} + \gamma_{j4}^{(t)} + \gamma_{j5}^{(t)}) - \gamma_{goalie}^{(t)}, x_{ijt}, \beta)$$

Solving the Model

First Choice

Bayesian Model: Use an MCMC Sampling Algorithm

Problem

To use full data set from 2007-2017, there will be $\approx$ 7,000,000 rows and $>$ 250,000 parameters (where periods $\leq$ 6 weeks)

Second Best

Maximum A posteriori Estimation

Maximum A posteriori Estimation

Model is *Almost Concave*
Can use **Accelerated Stochastic Gradient Descent** with parameters from the constant variance model as initial values to ensure stability

Maximum A posteriori Estimation Interpretation

Like Bradley-Terry with a smoothness penalty (Fahmeir and Tutz 1994)

Rewrite Likelihood

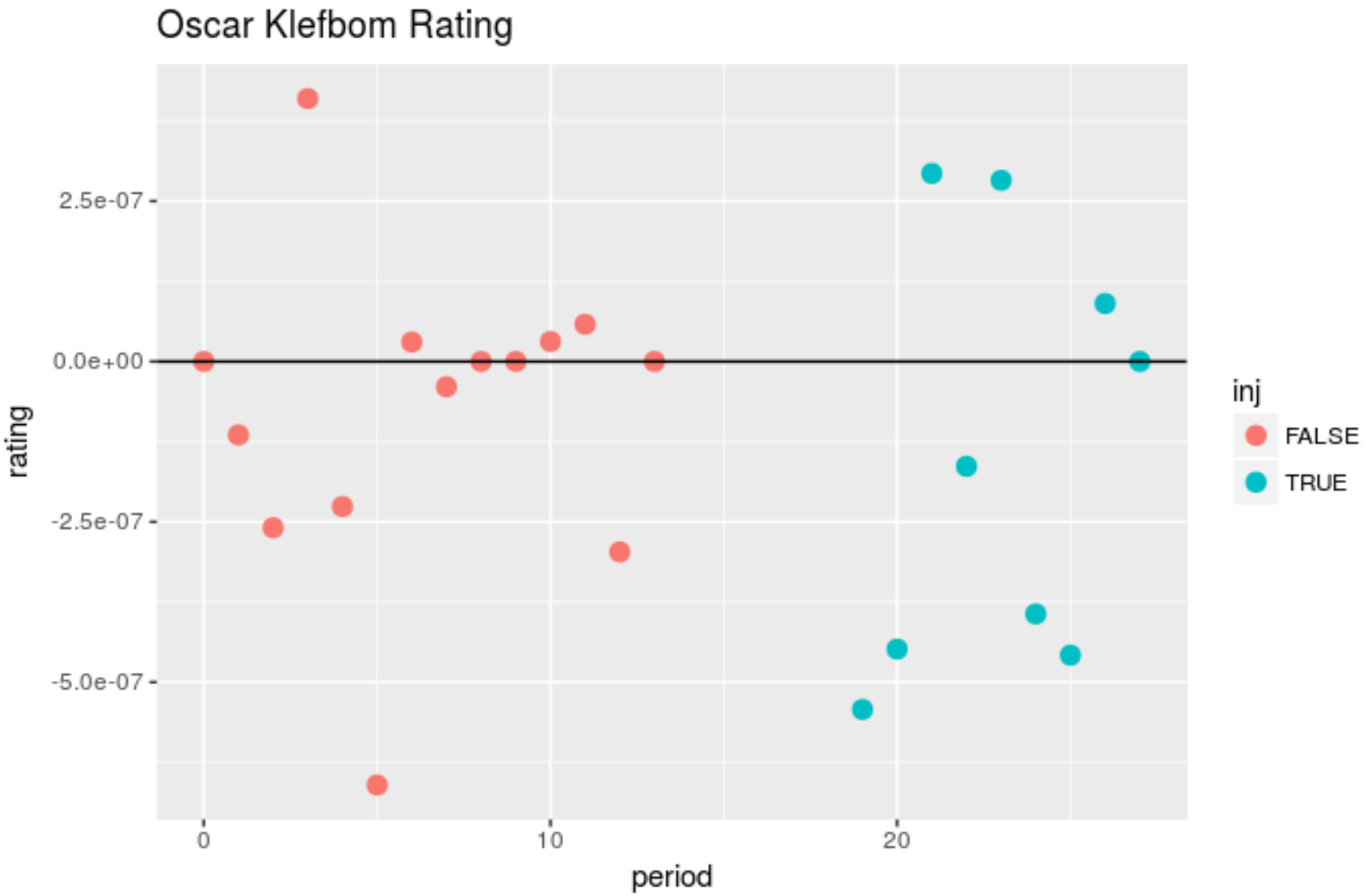
$$\log(L(\gamma, \sigma^2, \theta, \tau^2, \omega^2 | y, x)) =$$

$$\sum_{t=1}^{T}(\sum_{i<j}\sum_{k=1}^{K_{ij}^{(t)}}[y_{ijk}^{(t)} \log(F(\gamma_i^{(t)} - \gamma_j^{(t)}, x_{ijt}, \beta))$$

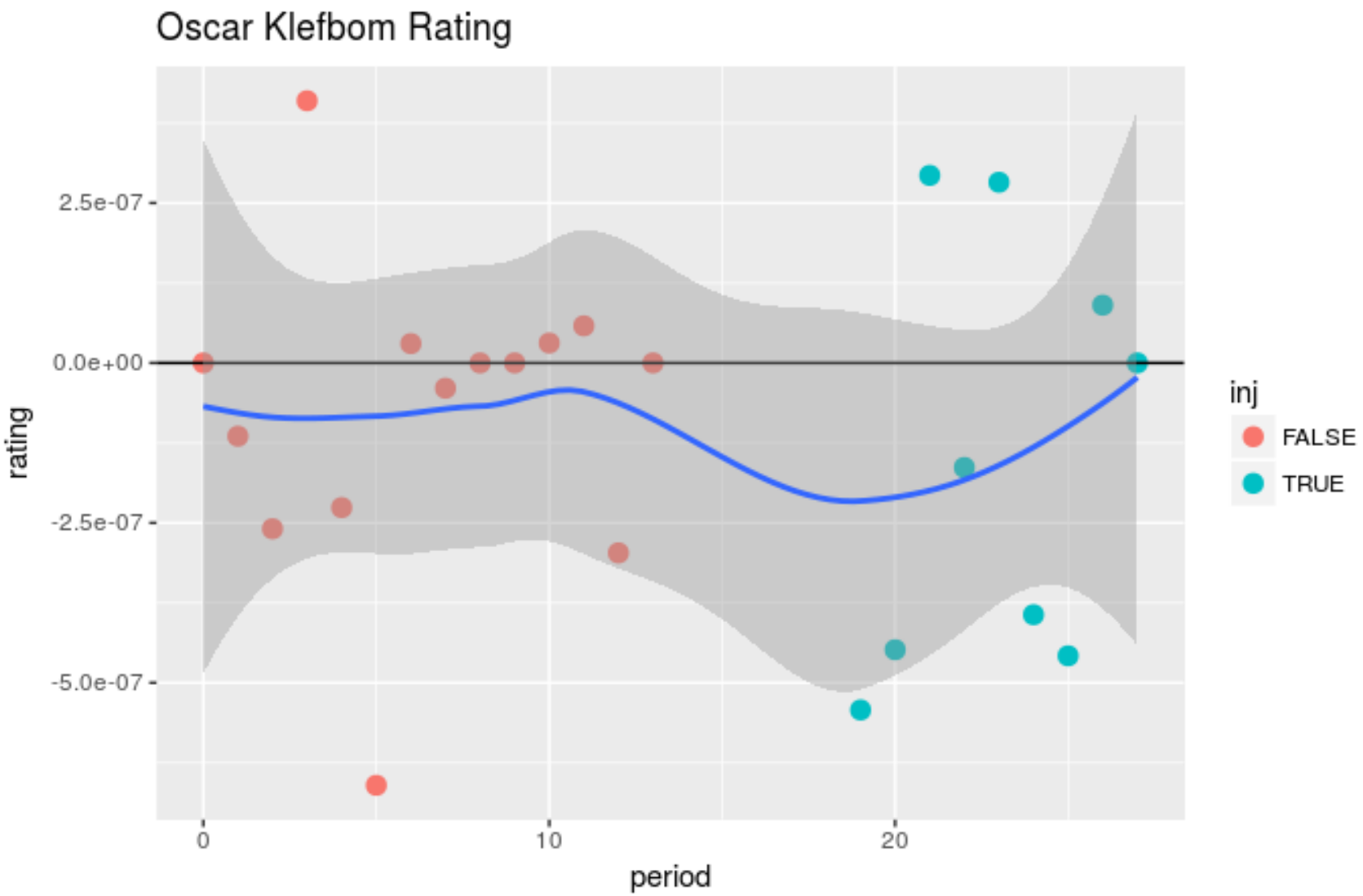
$$+ (1 - y_{ijk}^{(t)}) \log(1 - F(\gamma_i^{(t)} - \gamma_j^{(t)}, x_{ijt}, \beta))] + a(\gamma, \sigma^2)$$

Connects the Paired Comparison Litterature with previous regularized regression work in hockey (MacDonald 2012, Schuckers and Curro 2013 etc.)

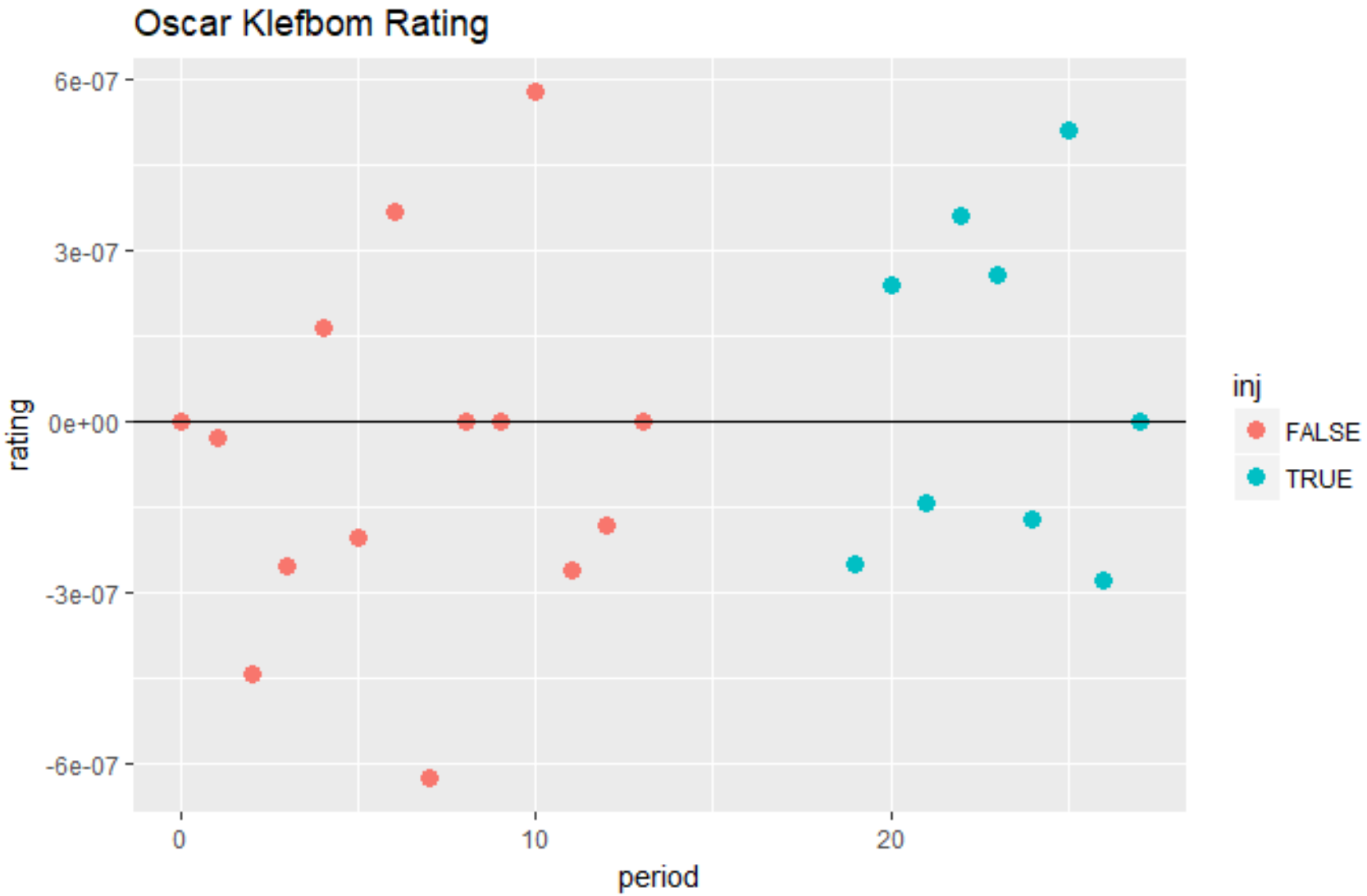
Results: Shots as outcome



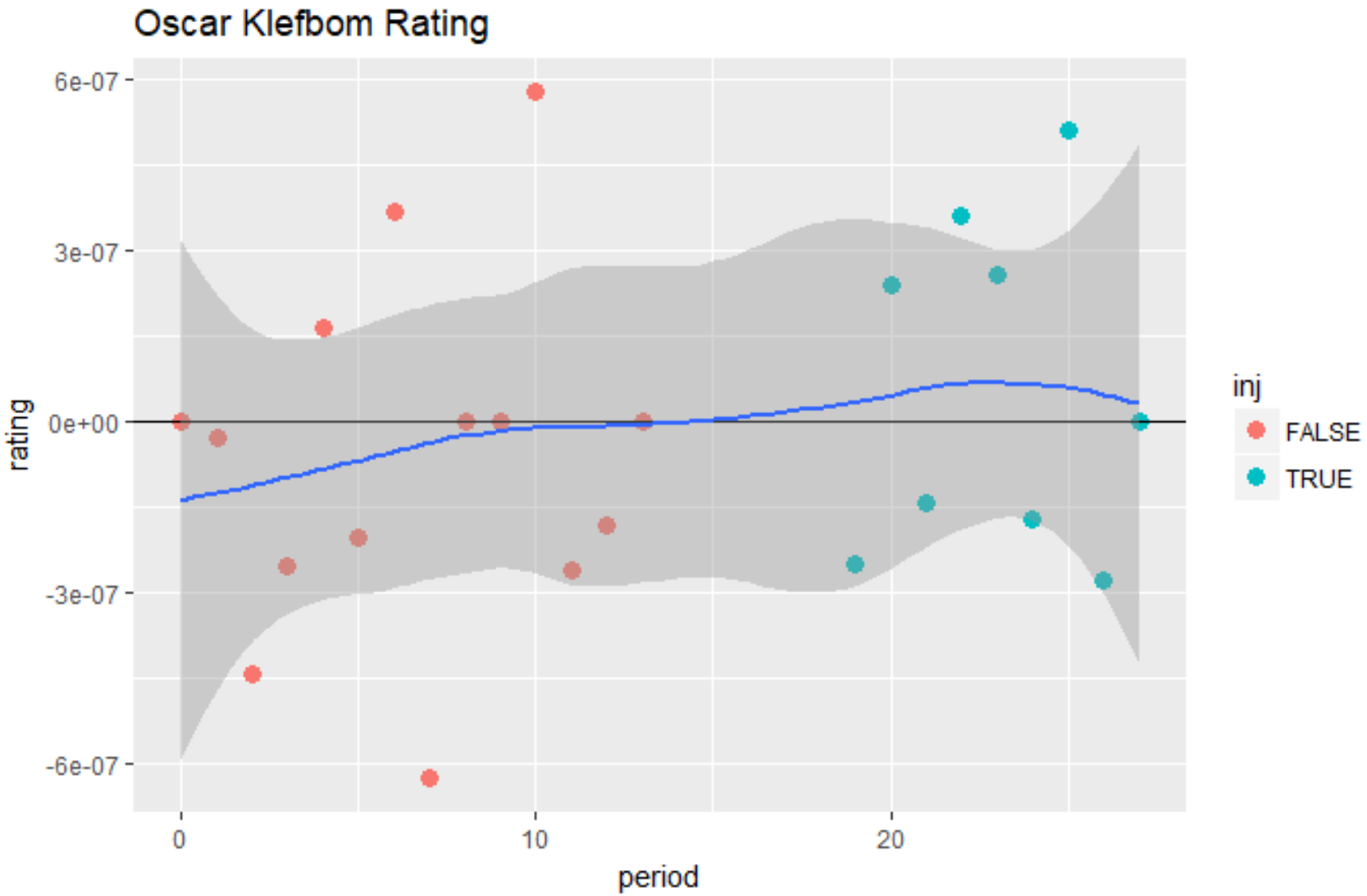
Results: Shots as outcome



Results: Goals as outcome



Results: Goals as outcome



Probability above Positional Average (PAPA)

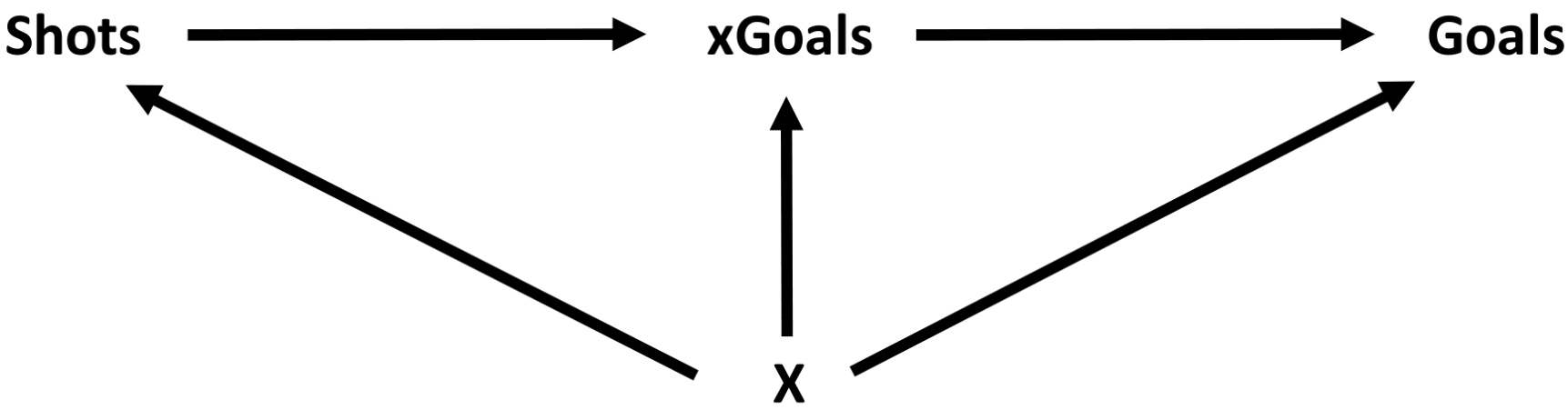
ex: Oscar Klefbom (LD)

PAPA for LD

$$\frac{1}{N}\sum_{i=1}^{T}\sum_{i<j}\sum_{k=1}^{K_{ij}^{(t)}}(E[F(\gamma_{i^*}^{(t)}+\gamma'_{\mathbf{i}_{-\mathbf{i}\mathbf{LD}}}+\gamma'_{\mathbf{j}}+\theta)]-E[F(\gamma_{\mathbf{i}'}+\gamma'_{\mathbf{j}}+\theta)])$$

$$\gamma_{i'}=(\gamma_{i1}^{(t)}+\gamma_{i2}^{(t)}+\gamma_{i3}^{(t)}+\gamma_{i4}^{(t)}+\gamma_{i5}^{(t)})$$
$$\gamma_{j'}=(\gamma_{j1}^{(t)}+\gamma_{j2}^{(t)}+\gamma_{j3}^{(t)}+\gamma_{j4}^{(t)}+\gamma_{j5}^{(t)})$$
$$\theta=\gamma_{goalie}^{(t)}+x_{ijt}\beta$$

Causal Approach Multivariate Regression



Future Work

- Connection between SSM and rating schemes, i.e elo, glicko, glicko 2 etc
- Can we use connection to test our dynamic specification?
- How to model the inherited state independent of players?
- Model hierarchy, multivariate extensions, team structure etc.
- R package (eta: eventually)

Data

- Corsica Hockey
- Man-Games Lost

References

Aldous, David, and others. 2017. "Elo Ratings and the Sports Model: A Neglected Topic in Applied Probability?" *Statistical Science* 32 (4). Institute of Mathematical Statistics: 616–29.

Fahrmeir, Ludwig, and Gerhard Tutz. 1994. "Dynamic Stochastic Models for Time-Dependent Ordered Paired Comparison Systems." *Journal of the American Statistical Association* 89 (428). Taylor & Francis: 1438–49.

Glickman, Mark E. 1999. "Parameter Estimation in Large Dynamic Paired Comparison Experiments." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 48 (3). Wiley Online Library: 377–94.

———. 2001. "Dynamic Paired Comparison Models with Stochastic Variances." *Journal of Applied Statistics* 28 (6). Taylor & Francis: 673–89.

Glickman, Mark E, and Hal S Stern. 1998. "A State-Space Model for National Football League Scores." *Journal of the American Statistical Association* 93 (441). Taylor & Francis: 25–35.

Király, Franz J, and Zhaozhi Qian. 2017. "Modelling Competitive Sports: Bradley-Terry-\'{E} L\H {O} Models for Supervised and on-Line Learning of Paired Competition Outcomes." *arXiv Preprint arXiv:1701.08055*.

Lopez, Michael J, Gregory J Matthews, and Benjamin S Baumer. 2017. "How Often Does the Best Team Win? A Unified Approach to Understanding Randomness in North American Sport." *arXiv Preprint arXiv:1701.05976*.