SMGT 432: Plus-Minus Lecture

Announcements

1. No office hours today
2. Last 3 Rice Soccer matches are in next 8 days
3. Clarify what I mean by "two improvements" on Assignment #3 → and normalize by events instead of minutes.

Notation

Definition: A segment is a period of time within a game during which no substitutions are made.

For each segment $s = 1, \dots, S$, we observe:

→ t_s: the time duration (in minutes) of segment s

→ g_s^h: the number of goals scored by the home team

→ g_s^a: the number of goals scored by the away team

→ $h_{s,p} = \begin{cases} 1 & \text{if player } p \text{ is on the field for the home team} \\ 0 & \text{otherwise} \end{cases}$

→ $a_{s,p} = \begin{cases} 1 & \text{if player } p \text{ is on the field for the away team} \\ 0 & \text{otherwise} \end{cases}$

$y_s = (g_s^h - g_s^a)/t_s$

$x_{s,p} = h_{s,p} - a_{s,p}$

Plus-Minus

$$PM_p = \frac{\sum_{s=1}^{S} (h_{s,p} - a_{s,p})(g_s^h - g_s^a)}{\sum_{s=1}^{S} t_s (h_{s,p} - a_{s,p})^2} = \frac{\sum_{s=1}^{S} t_s \cdot x_{s,p} \cdot y_s}{t_p^*}$$

t_p^* ← total minutes played by player p

Adjusted Plus-Minus (Gaussian)

Model: $y_s = \sum_{p=1}^{P} \beta_p x_{s,p} + \varepsilon_s \qquad \varepsilon_s \overset{ind}{\sim} \mathcal{N}(0, \sigma^2/t_s)$

$\boxed{\text{Pause}}$ Does this look familiar? (Bradley-Terry)
How do we estimate β? (Linear Regression)

$$\hat{\beta} = \operatorname{argmin}_{\beta} \sum_{s=1}^{S} t_s \Big(y_s - \sum_{p=1}^{P} \beta_p x_{s,p}\Big)^2$$

Cool Property

$$\hat{\beta}_p = \frac{\sum_{s=1}^{S} t_s \cdot x_{s,p} \big(y_s - \sum_{p' \neq p} \hat{\beta}_{p'} x_{s,p'}\big)}{t_p^*}$$

Regularized Adjusted Plus-Minus

$$\hat{\beta} = \operatorname{argmin}_{\beta} \Big\{ \sum_{s=1}^{S} t_s \Big(y_s - \sum_{p=1}^{P} \beta_p x_{s,p}\Big)^2 + \lambda \sum_{p=1}^{P} \beta_p^2 \Big\} \qquad \lambda > 0$$

$\boxed{\text{Pause}}$ Does this look familiar?

Box Plus-Minus

We observe "box-score" statistics $b_{p1}, \ldots, b_{pK}$
Examples: pass completion rate, aerial duel win rate, tackles per 90, clearances per 90

Model: $\hat{\beta}_p = \sum_{k=1}^{K} \gamma_k b_{pk} + \varepsilon_p \qquad \varepsilon_p \overset{ind}{\sim} \mathcal{N}(0, \sigma^2/t_p^*)$

$$\hat{\gamma} = \operatorname{argmin}_{\gamma} \sum_{p=1}^{P} t_p^* \Big(\hat{\beta}_p - \sum_{k=1}^{K} \gamma_k b_{pk}\Big)^2$$

$\boxed{\text{Pause}}$ Why not directly regress y onto $\vec{b}_1, \ldots, \vec{b}_K$?