Lab: Regularization & Ridge Regression

1. APM & RAPM to measure each NBA player's offensive and defensive "Impact"

We can measure player offensive skill fairly well using box score data, but the box score doesn't tell the full story (see: Westbrook) and won't tell us which role players are impactful or even whose good defensively for non-bigs.

A statistician wants to turn to statistical modeling to measure each NBA player's impact.

This motivates the Adjusted Plus Minus (APM) Model:

<u>Variables</u> i = index of the i^{th} possession in the dataset

$OP1(i), \ldots, OP5(i)$ are the 5 offensive players of possession i

$DP1(i), \ldots, DP5(i)$ are the 5 defensive players of possession i

y_i = outcome (points) of the i^{th} possession

parameters Each player $j = 1, \ldots, m$ gets an offensive strength parameter β_j and a defensive strength parameter γ_j. Interept is α_0.

Model

$$y_i = \alpha_0 + \beta_{OP1(i)} + \beta_{OP2(i)} + \cdots + \beta_{OP5(i)} - \gamma_{DP1(i)} - \gamma_{DP2(i)} - \cdots - \gamma_{DP5(i)} + \varepsilon_i$$

ε_i is mean 0 noise.

Matrix vector notation

$$X = \begin{array}{c} \begin{array}{ccccccc} \text{intercept} & \text{player 1 (off)} & \cdots & \text{player m (off)} & \text{player 1 (def)} & \cdots & \text{player m (def)} \end{array} \\ \left[\begin{array}{cc} 1 & \underbrace{\bullet \quad \cdots \quad \bullet}_{\substack{\text{1 if player } j \\ \text{is on offense} \\ \text{this possession,} \\ \text{else } 0}} \quad \underbrace{\bullet \quad \cdots \quad \bullet}_{\substack{\text{1 if player } j \\ \text{is on defence} \\ \text{this possession,} \\ \text{else } 0}} \end{array} \right] \end{array}$$

Row i = possession i

$$y = X^T \beta + \varepsilon.$$

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \beta = \begin{pmatrix} \alpha_0 \\ \beta_1 \\ \vdots \\ \beta_m \\ \gamma_1 \\ \vdots \\ \gamma_m \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$y_i = x_i^T \beta + \varepsilon_i$.

$x_i = i^{th}$ Row of X.

If you use OLS to fit the coefficients β, γ you get Adjusted Plus Minus (APM).

This will overfit, particularly for players with few possessions, and due to extreme multicollinearity (teammates who almost always play together).

Regularized Adjusted Plus Minus (RAPM) fits this model using Ridge Regression (glmnet in R) to shrink the coefficients towards zero.

* Your task: fit APM and RAPM for one season's worth of possessions. Visualize and compare.

You'll need to manually make the X matrix!

Ridge Regression Code in R:

```r
m1 = -3; m2 = 3; lambdas = 10^(seq(m1,m2,by=0.2));
lambdas
ridge_model = cv.glmnet(
  x = X_rapm_j, y = y_j, [redacted] nfolds = 5, alpha = 0, family="gaussian", lambda = lambdas, standardize=FALSE
)
lambda = ridge_model$lambda.min
plot(ridge_model)
print(paste0("lambda = ", lambda))
```

automatically tunes lamba over a grid by CRoss validation

make sure to include this so as to Not standardize the X matrix!! glmnet default is to standardize