Lab: Multivariable Linear Regression

1. Four Factors in Basketball

Basketball has many dimensions:
shooting accuracy, defense, steals, blocking, rebounding...
Which correlates with (or causes) winning?
Which matters most? Least?
Dean Oliver defined <u>four factors</u>:
scoring, crashing, protecting, attacking, which correspond to scoring accuracy, rebounding efficiency, turnover rates, and free throws.
A big innovation: express these in percentage terms (this normalizes for pace of play).

Your task is to use multivariable regression to assess the relative impacts of the four factors on winning. Let

$$\begin{cases} y = \text{team wins in a season} & \text{(outcome variable)} \\ x_1 = \text{EFG\%} - \text{Opp EFG\%} \\ x_2 = \text{TOV\%} - \text{Opp TOV\%} \\ x_3 = \text{OREB\%} - \text{DREB\%} \\ x_4 = \text{FTRate} - \text{Opp FTRate} \end{cases}$$

Here, EFG% = effective field goal percentage = $\frac{FG + \frac{1}{2}(3PT)}{FGA}$
↳ weight successful three pointers 50% more
TOV% = turnover percentage
OREB% = off. rebounds %
DREB% = def. rebounds %
FTrate = free throw rate.

- First, get to know your data.
  Find means, SDs, and correlations b/t the variables and consider their marginal distributions.
- Fit a multivariable regression model
  $$\mathbb{E}y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4.$$
  Fit a second regression model after standardizing each of the x variables (to have mean 0 and sd $\frac{1}{2}$).
  Which regression tells you about the relative value of each of the four factor to winning?
  Order the factor by importance to winning.
  Which has better predictive performance and why
  (use math; no need to code up out-of-sample prediction calculation if you don't want to)

2. Expected outcome of a punt

We have a dataset consisting of punts,

- row = a punt
- i = index of i^{th} punt
- y_i = outcome (next yardline, from opponent's perspective) of the punt
- ydl_i = yardline (yards from opp. endzone) of i^{th} kick
- pq_i = Punter quality of the i^{th} punter (I made this variable)
- $punter_i$ = name of punter

- use multivariable regression to model outcome (next yardline) as a function of yardline and punter quality.
  Consider linear terms $\beta_1 \cdot ydl_i + \beta_2 \cdot pq_i$
  and also transformations (e.g. quadratic, cubic, splines).
  Select a model (e.g., by out-of-sample predictive performance)
  and visualize that model (e.g. plot expected next yardline versus yardline for various punter quality values (color)).
- Rank the punters by PYOE (punt yards over expected) and visualize the rankings. (e.g. plot punter vs. career PYOE)