Assignment #5: Regularized Regression

Your task is to apply regularized regression on a dataset from the sport of your choice.

Why are you being asked to do this?

The best way to check how well you've learned something is to try doing it yourself. We have already covered the code to implement this method in class. Successfully fitting regularized regression on a new dataset will require you to wrangle the data and make necessary adaptations based on the unique challenges of your data. You will practice two competencies: (a) applying a model to a new dataset and (b) interpreting the results to inform decision-making in the real world.

What (exactly) are you being asked to do?

Find a dataset of event results for the sport of your choice. The data need to include the identity of at least one player for each event. Examples of datasets you could use are:

- **Soccer:** Shots
  Offensive adversary: the shooter
  Defensive adversary: the goalkeeper
  Other variables: shot location, etc.
  Outcome variable: indicator of whether goal is scored

- **Football:** Running plays
  Offensive adversary: the running back
  Defensive adversary: the defensive team
  Other variables: distance to first down, etc.
  Outcome variable: number of yards gained by rusher

- **Basketball:** Missed shots
  Offensive adversaries: the offensive players on the floor
  Defensive adversaries: the defensive players on the floor
  Other variables: indicator of whether shot was a 3-pointer, etc.
  Outcome variable: indicator of whether defense got the rebound

- **Baseball:** Plate appearances
  Offensive adversary: the batter
  Defensive adversary: the pitcher
  Other variables: the stadium, etc.
  Outcome variable: indicator of the batter strikes out

- **Golf:** Holes
  Offensive adversary: the golfer
  Defensive adversary: the hole
  Other variables: weather conditions, etc.
  Outcome variable: number of strokes

For your chosen dataset, describe the offensive and defensive adversaries; what other variables you have chosen to include in the model; and what outcome variable you are modeling. Use the data to fit a regularized regression model (linear mixed-effects regression or ridge regression). Report the top five and bottom five offensive and defensive adversaries from your model fit. To what extent do the results from the model match your intuition? Create a data visualization that tells a story about some aspect of your results.

Finally, **escape from model land** by interpreting your results in the context of the real world. Imagine that a front office executive has asked you to evaluate the offensive and defensive adversaries in your model and that they are going to use your report to make decisions about which players to acquire. What are the blindspots of your model? What other information might you want to consider before making a decision?

Submission Requirements

- A PDF report (max 4 pages) summarizing your findings, including at minimum the following:
  - a ranking of the top and bottom five offensive and defensive adversaries, according to your model

  - a data visualization that tells a story about some aspect of your results
  - an interpretation of your results in the context of the real world

- An R script that contains all of the code you used to perform the analysis

Reminders

- Prepare your report as if your audience is a front office executive who has not seen the assignment prompt. Write clearly and concisely, and format your report in a way that makes it easy to read.
- In this class we value exercising **creativity** on homework assignments! Look for opportunities to put your own personal touch on your work—try to do more than parrot what you've been taught.
- Please **anonymize** your submission by removing any personally identifiable information (including file paths in your R script that contain things like a username!).

Extra Credit

You may earn one percentage point of extra credit tacked on to your final semester grade by submitting (in a separate PDF file) a proof of the following claim made in Section 5.2 of the lecture notes: "One can show that any $\boldsymbol{\beta}$ satisfying equation (2) will also satisfy $\sum_j \beta_j = 0$, meaning that the estimated regression coefficients will be mean zero."

How will your grade be determined?

You will get feedback on your work product based on several criteria. Within each of those criteria, the feedback will be: Missing (0%), Needs Improvement (70%), Good (85%) or Exceeds Expectations (100%). Your grade on the assignment will be the average of the grades across criteria. The criteria are:

1. **Data modeling.** Did you correctly implement regularized regression and report offensive and defensive player rankings? Did you choose appropriate features to include in your model?
2. **Data visualization.** Did you include a plot that tells a worthwhile story about your results? Is that story easy to understand from a quick glance at your plot?
3. **Creative thinking.** Did you bring your own ideas from outside of this class to bear on the assignment?
4. **Critical thinking.** Did you escape from model land? Did you weigh evidence from multiple perspectives in forming your conclusion? Did you provide a thoughtful interpretation of your results?
5. **Written communication.** Did you write clearly and concisely? Did you organize your key ideas with the evidence supporting them? Did you format your report in a way that makes it easy to read?