# Assignment: Regression to the Mean

Your task is to apply regression to the mean on a dataset 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 performing regression to the mean 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 individual athlete performance with varying sample sizes, and focus on one particular stat (e.g. pass completion percentage in soccer). Perform regression to the mean to estimate the true talent level in this stat for all of the athletes in the dataset. Report the estimated population *standard deviation* you obtained in your implementation. Create a plot with observed performance on the $x$-axis and estimated true talent on the $y$-axis; include the diagonal line $y = x$ for reference.

Finally, **escape from model land** by interpreting your results in the context of the real world. You need NOT address all of these, but here are examples of questions you may consider addressing: *Why might this model be an inadequate representation of reality? What decision in sport management might be affected by this analysis? How might this analysis change the way fans think about the sport?*

### Submission Requirements

- A PDF report (max 2 pages) summarizing your findings, including at minimum the following:
  - an explanation of the dataset and stat you chose (and where you found it)
  - an estimate of the population *standard deviation* (in true talent) of this stat
  - a plot of estimated true talent v. observed performance
  - 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!).

## 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. **Regression and standard deviation.** Did you correctly implement regression to the mean? Did you provide the population standard deviation and contextualize it? (Is it big? Is it small?)
2. **Data visualization.** Did you include a readable plot and briefly comment on the relationship between the scatterplot and the $y = x$ line?

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?