STAT 528 Final

your name

Due on May 10th at 11:59 PM

This is your take-home final exam for STAT 528. It is due on May 10th at 11:59 PM. All work must be your own, and you are not allowed to consult with other students in the course. You are allowed to use Chat GPT, or any other large language model, but know that if your conclusions conflict with the work that you have done then there will be point deductions. Submit your exam in a final exam directory called `final`. This exam will consist of three problems:

Problem 1 [100 points]: One of the main interests of baseball fans, historians, and statisticians alike is to compare the performance of players across eras. In this analysis you are going to compare observed batting averages (see `raw_stats.csv`) to batting averages which are adjusted (see `era_adjusted_stats.csv`) for various factors including the stadium in which a player played in and the era (time period) that a player played. Most notably, the adjusted batting averages take in to account the competitiveness of each time period where, generally speaking, the competitiveness has greatly increased with time.

Do the following:

- **part a** [50 points]: Report a final model to both data sets which investigates the batting averages of players across time. Your model should include a fixed effect for year and it should account for differences in the quality of players as well as the aging curves of players which is typically quadratic, players get better as they gain experience and then get worse as they get older. You will need to justify your model choices for both data sets, check that modeling assumptions are not grossly misspecified, and compare to other sensible candidate models.

- **part b** [25 points]: Compare the results obtained by your final models fit to each data set. Specific interest is in an investigation on whether or not players are expected to have a higher batting average as time increases.

- **part c** [25 points]: Report the top 25 highest career batting averages (total career hits divided by total career at bats) according to both data sets and comment on the differences in these two ranking lists. Your commentary should include something about how each list fits with the notion of an increasingly competitive talent pool.

Problem 2 [50 points]: Let Y be a random variable from a full regular exponential family whose log likelihood can be written in canonical form

$$l(\theta) = \langle Y, \theta \rangle - c(\theta),$$

where $\theta \in \mathbb{R}$ is the canonical parameter and $c(\theta)$ is the cumulant function. Do the following:

- **part a** [10 points]: Show that the cumulant function $c(\theta)$ is a convex function.
- **part b** [10 points]: Part a is an important result for maximum likelihood estimation, why?
- **part c** [10 points]: Show that the canonical parameter space of this exponential family is a convex set.
- **part d** [10 points]: Let $Y_1, \ldots, Y_n$ be an iid sample of data with the same distribution as Y. It was shown in the notes that the maximum likelihood estimator $\hat{\theta}$ obeys sub-exponential concentration. Let $g : \mathbb{R} \to \mathbb{R}$ be a function of θ. Show that $g(\hat{\theta})$ also obeys sub-exponential concentration under suitable assumptions.
- **part e** [10 points]: State the assumptions on g that you needed in part d.

Problem 3 [50 points]: Suppose that we have 6 data points and want to fit a logistic regression model with response y, model matrix M, and unknown regression coefficient vector β given below:

$$y = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}; \qquad M = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}; \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

Do the following by hand, you are not allowed to report `glm` or `glmdr` fits but you can use R as a calculator:

- **part a** [20 points]: Write the log-likelihood for this model and find the maximum likelihood estimator (MLE) for β.

- **part b** [10 points]: Does the MLE for β exist in the support $\mathbb{R}^2$? Why or why not? Your answer should be concise but detailed.

- **part c** [20 points]: Construct a valid 95% confidence interval for the mean-value parameters (conditional success probabilities) when $x = (1\ 1)^T$ and when $x = (1\ 0)^T$.