MATH 5900 Project 2: Longitudinal Data Modelin

Angel Barrera — E00775034

April 6th 2025

Part I

1. Describe at least three descriptive statistics that can be used to explore a longitudinal data set. What does each tell you about the data?

1. **Mean Profiles Across Time:** By averaging the response variable at each time point across all individuals, we get a sense of the overall trend over time. This helps us see whether there's a general upward or downward pattern, and it often guides what type of model (linear, nonlinear, etc.) might make sense later.

2. **Within-Subject Standard Deviation:** This tells us how much an individual's values fluctuate over time. If this variation is large, it suggests strong temporal dynamics within individuals and indicates that models need to account for these subject-specific patterns, like with random effects or correlation structures.

3. **Lag or Autocorrelation (e.g., Lag-1):** This measures how strongly each observation relates to its previous value within the same subject. High autocorrelation is a clear sign of dependence over time, which is a core feature of longitudinal data and something we'll need to model explicitly (often using covariance structures or time series elements).

2. Briefly describe the importance of a "random effect" in a conditional longitudinal model.

In a conditional longitudinal model, the inclusion of a **random effect** is crucial because it captures the subject-specific variability that isn't explained by the observed covariates alone. Essentially, it allows us to account for the fact that individuals start at different baseline levels or follow different trajectories over time.

For example, if we're modeling repeated blood pressure measurements across patients, a random intercept would allow each person to have their own starting average. Without it, we'd incorrectly assume that everyone starts from the same baseline, which doesn't usually reflect reality.

Random effects improve the accuracy of our estimates by acknowledging that repeated measures within the same subject are correlated. This is one of the core reasons why mixed-effects models are often preferred for longitudinal data.

3. Briefly describe the differences between "population-averaged" and "subject-specific" effects in a longitudinal model.

The key difference between **population-averaged** and **subject-specific** effects lies in what type of inference we're making.

- **Population-averaged models**, like those estimated using Generalized Estimating Equations (GEE), focus on the average response across the entire population. These models tell us how the average individual is expected to respond to a covariate, without focusing on individual-level variability.
- **Subject-specific models**, on the other hand, incorporate random effects (like in mixed-effects models) and are concerned with how a particular individual responds over time. These models allow each subject to have their own intercept and/or slope, which makes them more flexible when we care about personalized patterns.

So, while both approaches are valid, they answer slightly different questions. Population-averaged models are often used in public health studies where we're interested in average treatment effects, whereas subject-specific models are better suited for personalized medicine or tracking within-person changes.

4. What are the differences between time-independent and time-dependent covariates?

In longitudinal data analysis, it's important to distinguish between **time-independent** and **time-dependent** covariates because they influence the model structure and interpretation.

- **Time-independent covariates** are variables that don't change over time for a subject. These might include things like gender, baseline treatment group, or genetic markers. They help explain between-subject differences and remain constant throughout the repeated measurements.
- **Time-dependent covariates** vary across time points for the same individual. Examples include current medication dosage, time-varying biomarkers, or daily physical activity. These are used to explain within-subject

variation and are essential when we want to capture dynamic relationships over time.

Part II

1. Purpose of the Analysis

For this project, I wanted to study how NBA players' performance changes over the course of their careers. Basketball has always interested me, and this dataset gave me a chance to explore something I enjoy while applying longitudinal analysis techniques. The main goal is to understand how player statistics—especially points per game and usage rate—evolve across seasons, and whether factors like age influence that progression.

Since each player appears in the dataset across multiple seasons, the data has a clear longitudinal structure. However, because some players only appear once or twice, using a subject-specific model like a linear mixed-effects model becomes challenging in practice. For this reason, I decided to use a population-averaged approach with Generalized Estimating Equations (GEE), which still accounts for within-player correlation while estimating average effects across all players.

The goal is to understand how performance changes with age and experience, not just for a single player, but across the league as a whole.

2. Description of the Data

The dataset used in this project contains season-by-season statistics for NBA players from the 1996–97 season through 2022. Each row represents a player's performance during a single season, which makes this a strongly longitudinal structure where players appear in multiple rows across time. This allows us to observe how individual players evolve over the course of their careers.

Each player is identified using the column `player_id`, which acts as a subject ID. The `season` variable indicates the year for each observation, effectively serving as our time variable. Together, these allow us to track each player's performance over time.

The dataset includes a mix of demographic information, career history, and performance metrics:

- **player_id**: Unique player identifier (numeric)
- **player_name**: Full name of the player (character)
- **team_abbreviation**: Team abbreviation (character)
- **season**: NBA season (e.g., 1996-97, 2005-06) (character)
- **age**: Player age in years at the time of the season (numeric)

- **player_height**: Height in centimeters (numeric)
- **player_weight**: Weight in kilograms (numeric)
- **college**: College attended (if any) (character)
- **country**: Country of birth or citizenship (character)
- **draft_year**, **draft_round**, **draft_number**: Draft information indicating when and how the player entered the NBA (numeric/character)
- **gp**: Games played in that season (numeric)
- **pts**: Points scored per game (numeric)

3. Proposed Analyses

To analyze how player performance evolves over time, I'm focusing on **points per game (pts)** as the main outcome. This metric captures offensive productivity and is consistently available across seasons. The main predictors in the model are **age** and **season**, both treated as continuous variables.

Since many players have only a few repeated measures, a subject-specific model like a mixed-effects model wasn't ideal for convergence. Instead, I'm using a **Generalized Estimating Equations (GEE)** approach via `PROC GENMOD` in SAS. GEE accounts for within-player correlation and estimates population-averaged effects — which is aligned with the goal of understanding overall league trends rather than individualized predictions.

The model I fit has the form:

$$E(pts_{it}) = \beta_0 + \beta_1 \cdot age_{it} + \beta_2 \cdot season_{it}$$

Where:

- β_0 is the intercept (overall average scoring baseline),
- β_1 reflects the average change in points per game with each additional year of age,
- β_2 captures the trend across seasons.

The `player_id` variable was used to define clusters in the repeated statement, and I specified an exchangeable correlation structure to account for correlation between repeated observations within the same player.

Before modeling, I also filtered out incomplete cases and created a numeric version of the season variable to simplify interpretation. This setup helps answer the research question: *On average, how does NBA scoring performance change with age and over time?*

4. Analysis Results

The Generalized Estimating Equations (GEE) model successfully converged, using an exchangeable correlation structure to account for repeated observations within each player. A total of 12,844 player-season records were used in the analysis, each uniquely identified by the `player_id` variable. The results are summarized in Figure 1, which shows the parameter estimates along with standard errors, confidence intervals, and significance levels from the GEE model.

The final model included **age** and **season (numeric)** as predictors of **points per game (pts)**. The results are presented in Table 1.

Table 1: GEE Parameter Estimates for Points per Game

Parameter	**Estimate**	**Std. Error**	**95% CI**	**p-value**
Intercept	-67.26	14.03	[-94.75, -39.77]	$< .0001$
Age	0.0250	0.0117	[0.0021, 0.0478]	0.0322
Season (numeric)	0.0372	0.0070	[0.0236, 0.0508]	$< .0001$

Both **age** and **season** had statistically significant effects on scoring. Specifically, the coefficient for age ($\hat{\beta}_{age} = 0.025$) suggests that, on average, a one-year increase in age is associated with a 0.025 point increase in scoring per game. Similarly, the positive coefficient for season ($\hat{\beta}_{season} = 0.0372$) indicates that scoring tends to increase slightly across later seasons, even after adjusting for age.

This observed trend is also supported by the visualization in Figure 2, which shows a gradual upward shift in average points per game across seasons. While the effect sizes in the model are small, they align with the broader trend of increasing league-wide scoring over the last two decades.

Fit statistics from the model include a QIC of 12846.86, confirming the model's stability and suitability for drawing population-level inferences.

5. Conclusions

This project gave me the opportunity to explore NBA player performance through a longitudinal lens, using season-level data across more than two decades. By modeling points per game over time, I was able to better understand how age and experience relate to scoring at the population level.

The results showed that both age and season were associated with small but statistically significant increases in scoring. This aligns with the trend seen in Figure 2, where average points per game have gradually risen

over time. This might reflect improvements in player training, offensive strategies, or shifts in league dynamics favoring higher scoring games.

Even though the effects were modest, the analysis helped demonstrate that simple variables like age and time can still reveal important patterns when analyzed within a longitudinal framework. Switching to a population-averaged model turned out to be a practical and effective choice for this dataset, especially given the high number of players with only one or two seasons.

Overall, this was a meaningful learning experience that combined my interest in sports with the statistical methods we've learned throughout the course. It reinforced how important it is to let the structure of the data guide the model selection and showed me how longitudinal analysis can uncover real-world trends in performance and behavior.

References

– Justinas. (2023). *NBA Players Data (1996–2022)*. Retrieved from Kaggle: `https://www.kaggle.com/datasets/justinas/nba-players-data?resource=download`

Appendix

GEE Model Information	
Correlation Structure	Exchangeable
Subject Effect	player_id (12844 levels)
Number of Clusters	12844
Correlation Matrix Dimension	1
Maximum Cluster Size	1
Minimum Cluster Size	1

Algorithm converged.

GEE Fit Criteria	
QIC	12846.8553
QICu	12847.0000

Analysis Of GEE Parameter Estimates						
Empirical Standard Error Estimates						
Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	-67.2565	14.0263	-94.7476	-39.7654	-4.80	<.0001
age	0.0250	0.0117	0.0021	0.0478	2.14	0.0322
season_numeric	0.0372	0.0070	0.0236	0.0508	5.35	<.0001

Figure 1: GEE model output from SAS, including parameter estimates and model fit statistics.

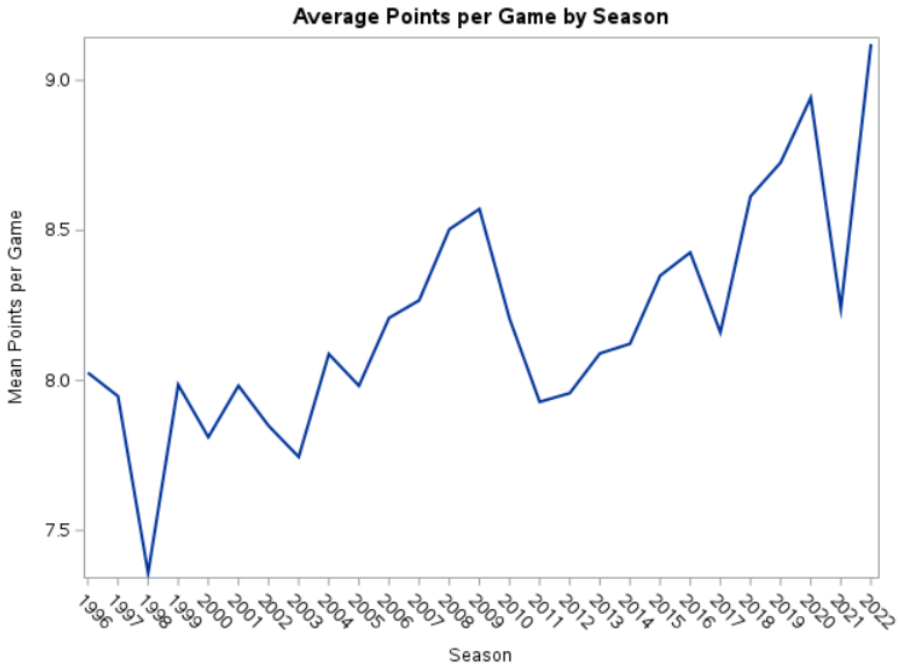

Figure 2: Average points per game by season, showing a gradual increase in scoring over time.