STAT7630: Bayesian Statistics
Lecture Slides # 15

Hierarchical Models and Pooling
Chapter 15 Hierarchical Models are Exciting

Elvan Ceyhan
Department of Mathematics & Statistics
Auburn University

Fall 2024,
Updated: November, 2024

Outline

Hierarchical Models

- Hierarchical (Grouped) Data
- Modeling with Completely Pooling
- Modeling with No Pooling
- Modeling with Partial Pooling - Hierarchical Models

Outline

Hierarchical (Grouped) Data

- **Hierarchical models** are utilized for data exhibiting a **grouping structure**.
- **Examples from the literature:**
  - A sampled set of schools with data Y collected on multiple individual students within each school (*clustered data*).
  - A sampled set of laboratories with data Y gathered from multiple experiments conducted within each lab (*clustered data*).
  - A sampled set of individuals with repeated measurements of a variable Y recorded over time (*longitudinal data*).
- **Dependency structure:**
  - Observations within the same school, lab, or individual are **not independent** and exhibit **intra-group correlation**.
  - Ignoring the grouped nature of the data may result in **biased estimates** and **misleading inferences**.

Terminology

- **Hierarchical models** are often referred to by other names depending on context and complexity:
  - **Multilevel models**
  - **Mixed-effects models**
  - **Random-effects models**
- **Panel data:** A specific case of *longitudinal data* where the same set of subjects is observed at each time point.
- In general, with longitudinal data, the set of subjects at each time point may vary across time.

Example of Grouped Data

- **Cherry Blossom Race Data:**
    - Dataset includes race times (in minutes) for runners aged in their 50s or 60s, recorded over multiple years.
    - Many runners appear **multiple times** in the dataset, as they participated in multiple races across different years.
- **Grouping structure:**
    - Data are **grouped by runner**, as multiple observations are associated with each individual.
    - Race times from the same runner are **not independent** but exhibit **intra-runner correlation**.

Visualizing the Cherry Blossom Race Data

- **Visualization:**
  - Side-by-side boxplots (see next slide) display the distribution of race times for 36 runners who participated in multiple Cherry Blossom races.
- **Observations:**
  - Runner 10: **Slow performance**.
  - Runner 29: **Fast and consistent performance**.
  - Runner 17: **High variability in performance**.
- **Research Question:**
  - What is the **relationship between a runner's age** and their **race time**?

Visualizing the Cherry Blossom Race Data

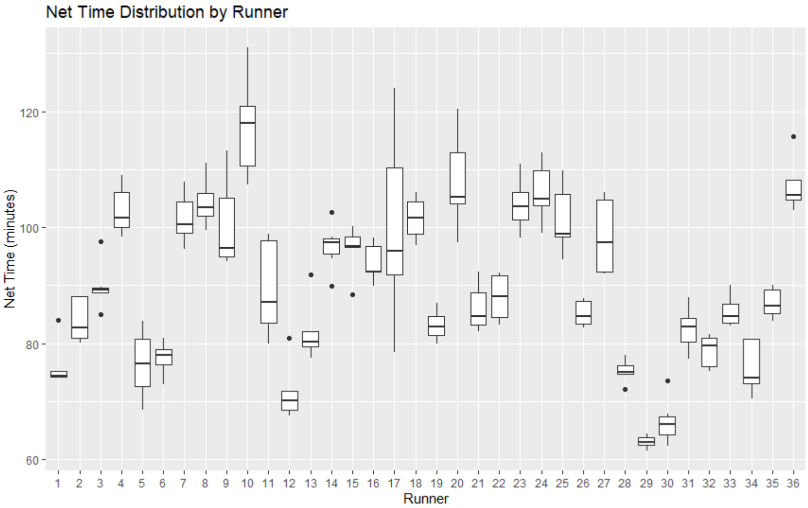

Figure 1: Boxplots of net running times (in minutes) for 36 runners that entered the Cherry Blossom race in multiple years.

Outline

Complete Pooling Analysis

- **Approach:**
  - Pool all data together, disregarding the grouping by runner.
  - Create a scatterplot of **race time** (Y) against **age** (X) (see next slide).
- **Observations:**
  - The scatterplot suggests a **weak relationship** between age and race time.
- **Simple Linear Regression:**
  - Model: $Y =$ net race time, $X =$ age.
  - Results: Age does not appear to be a **significant predictor** of race time.
- **Key Question:**
  - Does it make sense to ignore the grouping structure and conclude age has no significant effect?

Scatter Plot of Completely Pooled Data

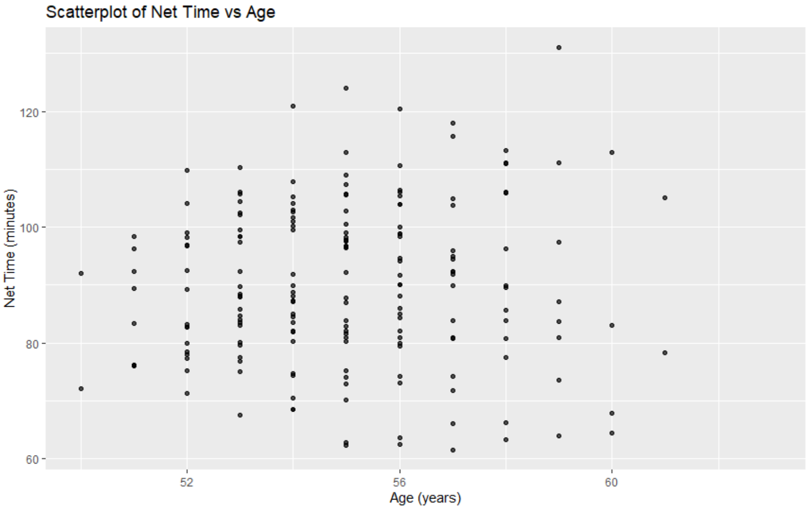

Figure 2: A scatterplot of net running time versus age for every race result.

Looking Further: Regression Analysis by Group

- **Exploring Regression Lines:**
  - Use the **posterior median values** of β_0 and β_1 to plot the pooled regression line (see next slide).
  - Overlay **individual regression lines** (in gray) for each runner, modeling race time against age separately.
- **Key Observations:**
  - The pooled regression line is **almost flat**, indicating a weak relationship overall.
  - Individual regression lines are **steeper**, showing that race times worsen with age.
- **Detailed Examination:**
  - Focus on three runners: 1, 20, and 22 (see two slides ahead).
  - Their aging trends are **highly variable**, illustrating how poorly the pooled regression line captures individual differences.
- **Visual Insight:**
  - R plots highlight the discrepancies between the pooled regression and the individual trends.

The Fitted Regression Lines, Pooled and for Each Runner

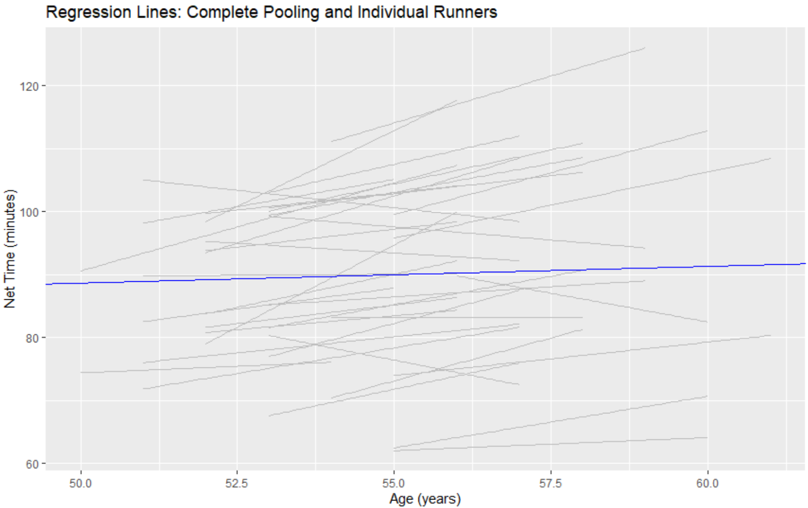

Figure 3: Observed trends in running time versus age for the 36 subjects (gray) along with the posterior median model (blue).

The Fitted Regression Lines, Pooled and for Three Specific Runners

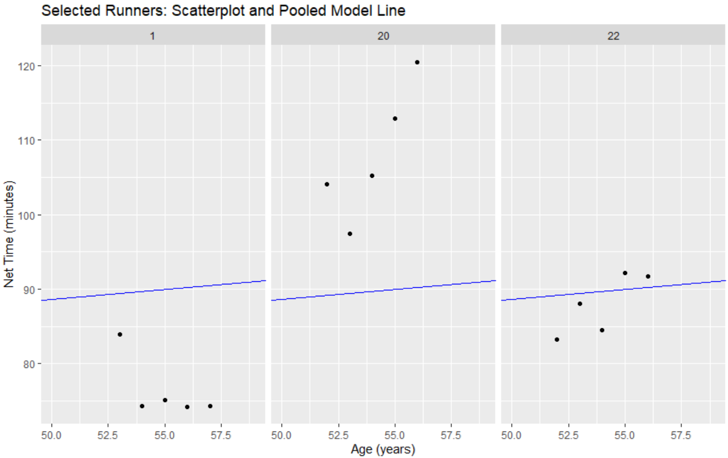

Figure 4: Scatterplots of running time versus age for 3 subjects, along with the posterior median model (blue).

Drawbacks of the Complete Pooling Model

- **Key Limitations:**
  - Assumes **independence of observations**, ignoring the fact that data from the same individual are **correlated**.
  - Imposes a **uniform aging trend** across all runners, disregarding potential **individual variability**.
- **Consequences:**
  - Leads to **misleading conclusions** about the regression relationship between Y (race time) and X (age).
  - Fails to accurately assess the **significance** of the relationship.

Outline

The No-Pooling Model

- **Approach:**
  - No pooling: Fit separate regressions for each runner in the dataset.
  - Model specification:

$$Y_{ij} \mid \beta_{0j}, \beta_{1j}, \sigma \sim N(\mu_{ij}, \sigma^2)$$

    where $\mu_{ij} = \beta_{0j} + \beta_{1j} X_{ij}$, allowing each runner ($j = 1, \ldots, n$) to have their own **intercept** β_{0j} and **slope** β_{1j}.
- **Complexity:**
  - This model introduces significantly more parameters:

    Instead of 2 regression coefficients, we estimate $2n$ coefficients.
- **Performance:**
  - `R` plots for 3 example runners demonstrate that this model captures individual trends **exceptionally well**.

Drawbacks of the No-Pooling Model

- **Key Issues:**
  - The model is **runner-specific**:
    - Predictions are only valid for the runners the model was fit for.
    - It cannot generalize to predict race times for **new runners**.
  - Lacks **population-level insight**:
    - The model cannot make general statements about the effect of **age on race time** across the population.
    - Individual slopes vary for each runner, making it impossible to derive a **single population-level trend**.

Additional Drawbacks of the No-Pooling Model

- **Limited Generalizability:**
  - Group-specific (runner-specific) models cannot be reliably **generalized** to groups (runners) outside the sample.
- **Information Loss:**
  - Assumes that one group contains **no relevant information** about another.
  - Ignores **shared patterns or trends**, potentially discarding valuable insights that could improve predictions or inferences.

Outline

Other Examples of Hierarchical Data

- **Recall earlier examples of multilevel data:**
  - Students in multiple schools taking the same achievement test.
  - **Dependency structure:** Test scores of students within the same school are likely to be **correlated**.
- **Grouping:**
  - Schools act as **groups**, analogous to runner-specific models.
  - School-specific models would face similar **drawbacks**, such as limited generalizability.
- **Complex Hierarchical Structures:**
  - Students grouped within classrooms $\rightarrow$ schools $\rightarrow$ districts $\rightarrow$ states.
  - Such nested structures can become highly **complex**.
- **Practicality:**
  - In practice, the number of hierarchical levels is typically limited to **2 or 3 levels**.

Why Use Hierarchical Data?

- **Practical Advantages:**
  - Collecting hierarchical data can be more **feasible** and **efficient** in practice.
- **Example 1: pH Levels in Rainfall**
  - Option 1: Take **1 measurement** from each of 30 rainfalls (*independent data*).
  - Option 2: Take **6 measurements** from each of 5 rainfalls (*correlated data within rainfall*).
- **Example 2: Soil Measurements**
  - Collect **10 measurements per field** at 8 fields (*correlated within field*) vs.
  - Collect **1 measurement from each of 80 fields** (*independent data*).
- **Key Consideration:**
  - Statistical models must account for **correlation** within groups.
  - Ignoring correlation and treating data as independent leads to **biased estimates and invalid inferences**.

A Happy Medium: Partial Pooling

- **Partial Pooling in Hierarchical Models:**
  - Balances between complete pooling and no pooling.
  - Results are **partially influenced** by both individual group information and shared information across groups.
- **Core Idea:**
  - Each group is **unique**, so retain group-specific information.
  - Borrow **shared information** across groups to improve parameter estimation.
- **Advantages:**
  - Allows for assessment of:
    - **Within-group variability:** How similar are observations within a group?
    - **Between-group variability:** How different are the groups?
- **Bayesian Suitability:**
  - The **Bayesian framework** is particularly effective for modeling hierarchical structures, leveraging priors to account for variability across levels.