Final Project

Federica Sfeir 2025512

Bayesian Modelling a.y. 2024-2025

M.Sc. in Statistical Methods & Applications

Introduction

The educational outcomes are shaped by a complex interaction of institutional characteristics, student demographics, and socioeconomic factors. In order to understand these dynamics is essential for identifying disparities in higher education and informing policies to enhance student success. There are various measures of academic performance, among which graduation rates are often considered as a key indicator of institutional effectiveness, as they strongly correlate with future economic opportunities and social contributions.

In this project we will analyze the College dataset that offers an opportunity to break down these dynamics, offering detailed data on 777 colleges and universities across the US. This project focuses on analyzing graduation rates, aiming to determine how much of the variability can be explained by institutional factors, such as tuition costs, faculty qualifications, and student-to-faculty ratios, as well as individual characteristics like enrollment size and the proportion of students from high-achieving high school backgrounds.

This analysis seeks to answer the following question: "What are the key factors influencing graduation rates, and how do these factors differ between private and public institutions?"

Through the Bayesian mixed-effects model, the study aims to disentangle the effects of institutional and individual characteristics on graduation outcomes, offering insights into how colleges can better support student success and equity in higher education.

College Dataset

The dataset used in this analysis is derived from the College dataset available in the `ISLR2` R package. This dataset contains information on 777 colleges and universities in the United States, providing insights into institutional characteristics and their relationship with graduation rates and other outcomes.

The dataset includes the following 18 variables:

- **Private**: A binary variable indicating the type of institution (1 = Private, 0 = Public).
- **Apps**: Number of applications received.
- **Accept**: Number of applicants accepted.
- **Enroll**: Number of new students enrolled.
- **Top10perc**: Percentage of new students who graduated in the top 10% of their high school class.
- **Top25perc**: Percentage of new students who graduated in the top 25% of their high school class.
- **F.Undergrad**: Number of full-time undergraduate students.
- **P.Undergrad**: Number of part-time undergraduate students.
- **Outstate**: Tuition for out-of-state students (in USD).
- **Room.Board**: Cost of room and board (in USD).
- **Books**: Estimated cost of books (in USD).
- **Personal**: Estimated personal expenses (in USD).
- **PhD**: Percentage of faculty with PhD degrees.

- **Terminal**: Percentage of faculty with terminal degrees (highest degree in their field).
- **S.F.Ratio**: Student-to-faculty ratio.
- **perc.alumni**: Percentage of alumni who donate to the institution.
- **Expend**: Instructional expenditure per student (in USD).
- **Grad.Rate**: Graduation rate of the institution (percentage of students who graduate within a specified time frame).

Data Cleaning

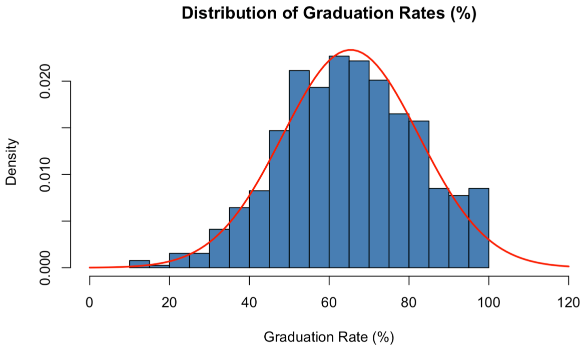
In order to prepare the dataset for reliable analysis, we proceed with a data cleaning process. First of all, no missing values were found in the dataset. Therefore, no observations were removed or imputed for this reason. We identified one observation with an unrealistic **Grad.Rate** greater than 100%. Graduation rates represent percentages and should range between 0 and 100. Thus the observation for Cazenovia College, which had a **Grad.Rate** exceeding 100%, was removed to maintain the validity of the analysis. The continuous variables such as **Outstate** (tuition for out-of-state students), **Expend** (instructional expenditure per student), **S.F.Ratio** (student-to-faculty ratio), and others were standardized. This ensures all variables are on a comparable scale, which is particularly useful for Bayesian modeling and aids in model convergence. The **Private** variable, which indicates whether a college is private or public, was encoded as a binary variable (1 for private, 0 for public) to facilitate its inclusion. After these steps, the dataset was left with 776 observations.

Descriptive Statistics

In this section we provide an overview of the dataset used in this analysis, through a descriptive analysis, highlighting key patterns and relationships between variables.

By examining the distribution of graduation rates and their association with factors such as institutional type (private vs. public), tuition costs, instructional expenditures, and student-to-faculty ratios, this section lays the foundation for understanding the variability in college performance. These insights will guide the subsequent modeling efforts to explore the contributions of institutional characteristics to graduation outcomes.

The first histogram displays the distribution of graduation rates across colleges and universities in the dataset.The distribution of graduation rates appears slightly right-skewed, with a peak around 60-70%. This suggests that most colleges have graduation rates within this range, representing the majority of institutions. The mean graduation rate is 65.4%, and the range is from 10% to 100%. Few colleges achieve very low or very high graduation rates, indicating a clustering of institutions around the median. This initial analysis provides insights into the overall performance of colleges and sets the stage for further exploration of contributing factors.



The following boxplot compares graduation rates between public and private colleges in the dataset. We observe that Public colleges tend to have a wider spread of graduation rates, with many institutions clustered around lower rates. While, Private colleges show a higher median graduation rate and a more compact distribution, indicating less variability compared to public colleges.

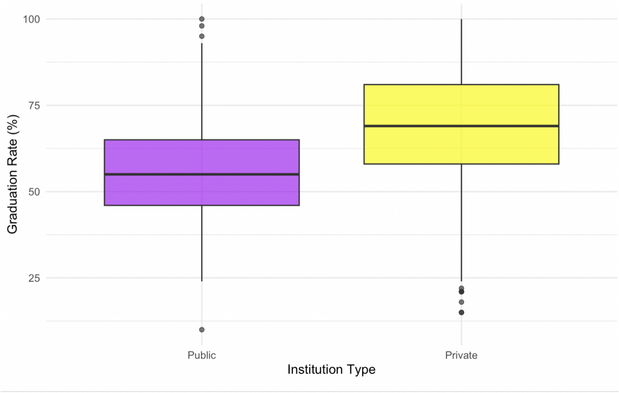
However, it is essential to note that public colleges make up a smaller proportion of the dataset compared to private colleges (approximately 35.5% of the total). This imbalance in representation could influence the overall descriptive statistics and lead to conclusions that reflect private colleges more strongly than public colleges.

Type (0 = Public, 1 = Private)	Mean Graduation Rate (%)	Proportion
0	56.04245	0.2731959
1	68.91135	0.7268041

Tabella 1: Mean Graduation Rate and Proportion by College Type

The table presents the mean graduation rate and the proportion of colleges by type (public or private). As we can observe, the Private colleges (Type = 1) exhibit a higher average graduation rate of 68.91% and make up the majority of the sample (72.68%). While the Public colleges (Type = 0) have a lower average graduation rate of 56.04%, accounting for a smaller proportion of the sample (27.32%). These results suggest a potential disparity in graduation outcomes between public and private colleges. This could reflect differences in institutional characteristics, resources, or the socio-economic backgrounds of students.

In the subsequent analyses, such as Bayesian mixed-effects modeling, this problem might lead to larger uncertainty in parameter estimates related to public colleges. Fewer data points for public colleges may reduce the precision of estimates, highlighting the need to account for this imbalance in the analysis.

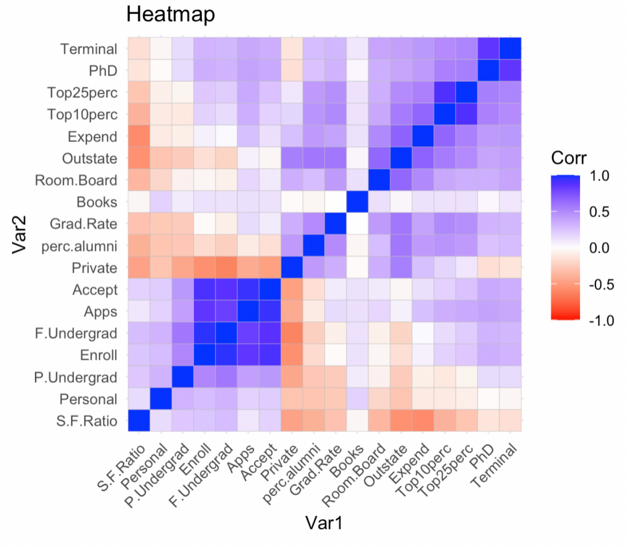


We can also consider the heatmap, which provides a clear and visually appealing representation of the relationships between the numeric variables in the dataset. The heatmap highlights strong positive correlations (dark blue squares), such as: **Grad.Rate** and **Top10perc**, so Colleges with a higher percentage of top-performing students tend to have better graduation rates. **Expend** and **Grad.Rate**, thus institutions that spend more on instructional expenditure generally achieve higher graduation rates. **Top10perc** and **Top25perc**, these variables are highly correlated, as expected, since both represent academic quality measures of incoming students.

While the dark red squares indicate strong negative correlations, including: **S.F.Ratio** and **Grad.Rate**, so we can say that Colleges with lower student-to-faculty ratios (smaller classes) tend to have higher graduation rates, reinforcing the importance of personalized instruction.

The light-colored squares (white or pale shades) show weak or negligible correlations. For example in **Books** and **Grad.Rate** there's little to no relationship between the cost of books and graduation rates. Also in **Personal Expenses** and most other variables do not strongly correlate with institutional or performance metrics.

The hierarchical ordering groups variables with similar patterns of correlation, making it easier to identify clusters of related variables. In fact academic quality measures (**Top10perc** and **Top25perc**) are closely grouped with **Grad.Rate**, reflecting their shared impact on performance. We can also observe that institutional spending metrics (**Expend**, **Room.Board**) cluster together, suggesting they are related in terms of institutional resource allocation.



Bayesian Mixed-Effects Model Specification

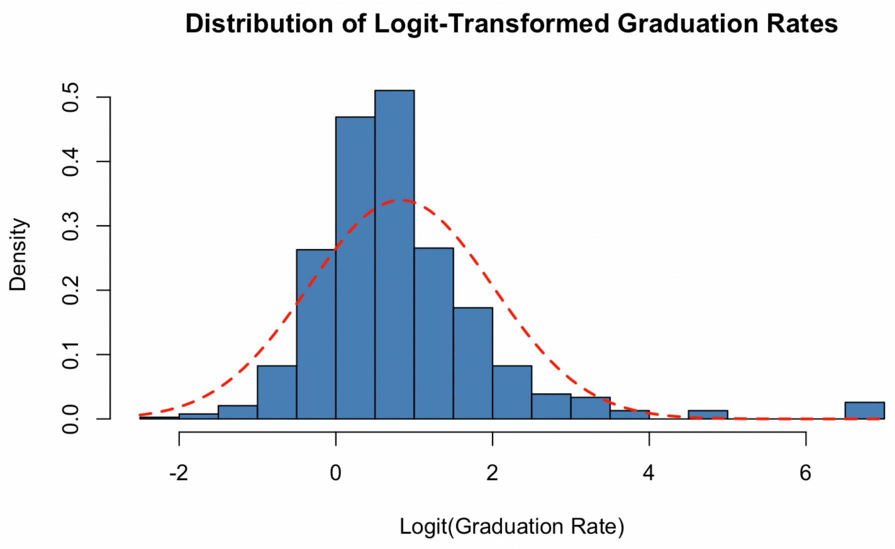
The goal of this project is to develop and specify Bayesian mixed-effects models to analyze the variability in graduation rates and evaluate the effects of institutional and student-level factors. By employing a hierarchical structure, the models account for both institutional-level predictors (e.g., instructional expenditure, student-to-faculty ratio) and the clustering of institutions by type (public vs. private). This approach allows us to partition the variability in graduation rates across these two levels and identify key factors influencing institutional performance.

Linear mixed-effects models are an extension of simple linear models to allow both fixed and random effects, and are particularly used when there is non independence in the data, such as arises from a hierarchical structure. The core of mixed models is that they incorporate fixed and random effects. Fixed effects represent population-level parameters, while random effects capture group level variability, in this case, differences between public and private colleges. The inclusion of random effects allows the model to account for the hierarchical structure of the data, where institutions are nested within types (public vs. private). This framework increases flexibility and improves inference by sharing information across different groups. The Bayesian framework further enhances this analysis by integrating prior knowledge and quantifying uncertainty through posterior distributions.

To ensure the suitability of regression modeling and compliance with the assumptions of the Bayesian framework, the dependent variable **Grad.Rate** was transformed into its logit form. This transformation is particularly relevant as **Grad.Rate** was originally expressed as a percentage ranging between 0 and 100, which makes it bounded and unsuitable for direct modeling with a normal likelihood. The Graduation rates were first converted from percentages to proportions by dividing by 100, then the logit transformation was applied:

$$\text{Logit}(y) = \log\left(\frac{y}{1-y}\right)$$

This transforms the proportion y into a variable that is unbounded and continuous, ranging from $(-\infty, +\infty)$.The logit transformation ensures that the variable maps onto the real number line, removing the boundary constraints while preserving its interpretability in terms of probabilities. After transformation, the coefficients in the model $(\beta_1, \beta_2, ...)$ represent the expected change in the log-odds of graduation rates for a one-unit increase in the corresponding predictor, holding all other predictors constant.



The general structure of a Bayesian mixed-effects model for nested data, as applied to the College dataset, is the following:

$$y_{ij} = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ijk} + u_j + \epsilon_{ij},$$

where the components of the model are defined as follows:

- y_{ij}: The observed outcome (graduation rate) for institution i in group j (public or private). It is assumed to follow a normal distribution with mean μ_{ij} and variance σ^2:

  $$\text{Grad.Rate}_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma^2),$$

  where the mean structure is given by:

  $$\mu_{ij} = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ijk} + u_j.$$

- β_k: Represent the fixed-effects coefficients for institutional-level predictors x_{ijk}, such as `Expend` (instructional expenditure per student), `Outstate` (Out-of-state tuition), `S.F.Ratio` (Student-to-faculty ratio). The fixed effects are assigned weakly informative priors:

$$\beta_k \sim \mathcal{N}(0, \tau_\beta),$$

  where τ_β is a small precision value to reflect a weakly informative prior.

- u_j: random effects, containing deviations for each group j (institution type: public or private) from the population-level mean.Modeled as:

$$u_j \sim \mathcal{N}(0, \sigma_u^2),$$

  where σ_u^2 represents the variance of the random effects across groups.

- ϵ_{ij}: Residual errors, represent the variability within institutions that is not explained by the predictors or random effects:

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2).$$

- σ^2 and σ_u^2): Variance Components, modeled using weakly informative priors:

$$\sigma^2 \sim \text{InverseGamma}(a, b), \quad \sigma_u^2 \sim \text{InverseGamma}(a, b),$$

  where a and b are hyperparameters representing prior beliefs about the variances.

By incorporating both fixed and random effects, the Bayesian framework allows us to quantify uncertainty and capture hierarchical structures in the data.

Model Formulation for the College Dataset

The following four models are specified to analyze the variability in graduation rates ($y_{ij} = \text{Grade.Rate}_{ij}$) and evaluate the effects of institutional characteristics. These models progressively incorporate fixed effects, random effects, and predictors to assess their contributions to explaining the observed variability.

1. Null Model (random intercept only): The null model includes only a random intercept for institution types (public vs. private), with no predictors. This model serves as a baseline to decompose the total variability in graduation rates into within group (within public or private institutions) and between group (public vs. private) components

$$\text{Grad.Rate.Logit}_{ij} = \beta_0 + u_j + \epsilon_{ij},$$

   where: β_0 is the overall mean graduation rate; $u_j \sim \mathcal{N}(0, \sigma_u^2)$ the random intercept for institution type j; $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ the residual variability within institutions.

2. Random Effects with Key Predictors: This model incorporates key institutional level predictors, including instructional expenditure (Expend), out-of-state tuition (Outstate), and student-to-faculty ratio (S.F.Ratio). These variables provide a foundation for exploring the relationships between institutional characteristics and graduation rates before extending the model to include additional predictors. This allows us to evaluate how much of the within-group variability can be explained by these predictors:

$$\text{Grad.Rate.Logit}_{ij} = \beta_0 + \beta_1 \cdot \text{Expend}_{ij} + \beta_2 \cdot \text{Outstate}_{ij} + \beta_3 \cdot \text{S.F.Ratio}_{ij} + u_j + \epsilon_{ij},$$

   where: β_1, β_2, β_3 represent the fixed effects for predictors, $u_j \sim \mathcal{N}(0, \sigma_u^2)$ is the random intercept for institution type, $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ is the residual variability within institutions.

3. Full Random Effects Model: This model extends the previous model by including additional predictors to account for institutional characteristics, such as:

- Percentage of alumni who donate to the institution (**perc.alumni**), which highlights its role as a proxy for alumni satisfaction, institutional reputation, and potential resource availability.

- Top 10% student (**Top10perc**) reflects the quality or preparedness of incoming students, which could strongly influence graduation rates.
- The number of new students enrolled (**Terminal**), which reflects institutional scale and may impact resource allocation and academic outcomes.

$$\text{Grad.Rate.Logit}_{ij} = \beta_0 + \beta_1 \cdot \text{Expend}_{ij} + \beta_2 \cdot \text{Outstate}_{ij} + \beta_3 \cdot \text{S.F.Ratio}_{ij} + \\ \beta_4 \cdot \text{perc.alumni}_{ij} + \beta_5 \cdot \text{Top10perc}_{ij} + \beta_6 \cdot \text{Terminal}_{ij} + u_j + \epsilon_{ij}.$$

where the fixed effects (β_4, β_5, β_6) represent the contributions of these additional predictors.

4. Fixed Effects Only (No Random Effects): this model is specified without random intercepts, in order to assess the importance of random effects. It assumes no between-group variability (public vs. private):

$$\text{Grad.Rate.Logit}_{ij} = \beta_0 + \beta_1 \cdot \text{Expend}_{ij} + \beta_2 \cdot \text{Outstate}_{ij} + \beta_3 \cdot \text{S.F.Ratio}_{ij} + \\ \beta_4 \cdot \text{perc.alumni}_{ij} + \beta_5 \cdot \text{Top10perc}_{ij} + \beta_6 \cdot \text{Terminal}_{ij} + \epsilon_{ij}.$$

where $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ is the residual variability.

Assessing Model Fit: DIC Comparison

In order to choose the best model, we compared the models using the Deviance Information Criterion (DIC), where lower DIC values indicate a better fit to the data. This comparison allows us to evaluate the result of incorporating random effects and additional predictors on model performance.

Observing the results, summarized in the table below, we note that Model 4 provides the best fit to the data, as evidenced by its lowest DIC value (2154.562). This model includes all predictors but excludes random effects, suggesting that fixed effects alone sufficiently explain the variability in graduation rates without accounting for group-level differences (public vs. private institutions). However, Model 3, which incorporates random effects, has a slightly higher DIC value (2156.189). The small difference in DIC between these models indicates that random effects provide minimal additional explanatory power when all predictors are included.

Model	**DIC**
Model 1	2394.236
Model 2	2244.750
Model 3	2156.189
Model 4	2154.562

Tabella 2: DIC Values

Models 1 and 2, which include fewer predictors, show significantly higher DIC values (2394.236 and 2244.750, respectively), highlighting their limited ability to capture the variability in graduation rates compared to the more comprehensive models.

These results indicate that incorporating a full set of predictors significantly improves model performance. However, the small difference in DIC values between Models 3 and 4 suggests that random effects contribute minimally when all predictors are included. As a result, we selected Model 4 for further interpretation and diagnostic analysis.

```
# Inference for Bugs model at "8", fit using jags,
# 1 chains, each with 10000 iterations (first 1000 discarded), n.thin = 10
# n.sims = 900 iterations saved. Running time = 1.848 secs
#           mu.vect sd.vect    2.5%     25%     50%     75%   97.5%
# beta0       0.829   0.033   0.767   0.807   0.828   0.852   0.900
# beta1      -0.019   0.056  -0.127  -0.055  -0.022   0.017   0.089
# beta2       0.317   0.055   0.207   0.283   0.319   0.355   0.422
# beta3       0.079   0.045  -0.002   0.047   0.077   0.111   0.165
# beta4       0.018   0.004    0.011    0.016    0.018    0.021    0.025
# beta5       0.366   0.050    0.268    0.332    0.365    0.401    0.461
# beta6      -0.005   0.003   -0.010   -0.007   -0.005   -0.003    0.001
# sigma       0.967   0.025    0.920    0.950    0.966    0.984    1.019
# deviance 2147.890   3.653 2142.158 2145.100 2147.414 2150.149 2155.862

# DIC info (using the rule: pV = var(deviance)/2)
# pV = 6.7 and DIC = 2154.6
# DIC is an estimate of expected predictive error (lower deviance is better).
```

Results of the Bayesian Fixed-Effects Model (Model 4)

Using the Bayesian fixed-effects model, the results provide valuable insights into the factors influencing graduation rates in colleges. Below is a detailed interpretation of the results, focusing on the fixed effects, variance components, and overall model fit.

- Intercept (β_0 = 0.829, 95% CI: [0.767, 0.900]): The intercept represents the baseline graduation rate for a college when all predictors are at their reference or mean levels. This suggests that colleges with average characteristics across all predictors are expected to have a moderate baseline graduation rate. In terms of proportions, this corresponds to:

$$y = \frac{\exp(0.829)}{1 + \exp(0.829)} \approx 0.696 \text{ (or 69.6\% graduation rate).}$$

Considering the key predictors, we get:

- Instructional Expenditure (β_1 = -0.019, 95% CI: [-0.127, 0.089]): The coefficient is small and its credible interval includes zero, indicating that instructional expenditure does not have a statistically significant effect on graduation rates in this model. This suggests that simply increasing instructional expenditure does not directly translate to better graduation outcomes. So a one-unit increase in instructional expenditure (e.g., $1,000) is associated with a 0.019 decrease in the logit of graduation rates.
- Out-of-State Tuition (β_2 = 0.317, 95% CI: [0.207, 0.422]): Higher out-of-state tuition is positively associated with graduation rates. This statistically significant effect suggests that institutions charging higher tuition may have better resources or attract students who are more likely to graduate. A one-unit increase in out-of-state tuition (e.g., $1,000) is associated with a 0.317 increase in the logit of graduation rates.
- Student-to-Faculty Ratio (β_3 = 0.079, 95% CI: [-0.003, 0.165]): The coefficient is small, and the credible interval is slightly below zero. This indicates that the student-to-faculty ratio might have a slight positive association with graduation rates, but the effect is not strongly supported by the data.
- Percentage of alumni who donate to the institution (β_4 = 0.018, 95% CI: [0.011, 0.025]): A higher percentage of alumni donations is positively and significantly associated with graduation rates. This reflects positively on the institution's reputation and its ability to provide a high-quality educational experience. For every 1% increase in alumni donations, the logit of graduation rates increases by 0.018.
- Top 10% students (β_5 = 0.366, 95% CI: [0.268, 0.461]): A higher percentage of students who graduated in the top 10% of their high school class is strongly and positively associated with higher graduation rates. This statistically significant effect emphasizes the importance of student quality as a major driver of institutional success.
- Terminal (β_6 = -0.005, 95% CI: [-0.010, 0.001]): The percentage of faculty with terminal degrees has a small negative coefficient, but the credible interval includes zero. This suggests that this variable is not a statistically significant predictor of graduation rates in this model. While a high percentage of faculty with terminal degrees (e.g., PhDs) often reflects academic expertise, it may not directly impact students'

graduation outcomes. However, this effect is very weak. A 1% increase in the percentage of faculty with terminal degrees corresponds to a 0.005 decrease in the logit of graduation rates.

The residual variance ($\sigma = 0.967$, 95% CI: [0.920,1.019]) captures the variability in graduation rates that remains unexplained by the predictors included in the model. In other words, it measures how much of the outcome (graduation rates) cannot be attributed to the effects of the fixed predictors. In this case, a residual variance of 0.967 indicates a moderate level of unexplained variability in the graduation rates after accounting for the predictors.

The model fit suggests that this model effectively captures the influence of institutional and student-level characteristics on graduation rates. Overall, the findings emphasize the importance of considering institutional factors, such as out-of-state tuition and alumni engagement, as well as student quality, represented by the percentage of students in the top 10% of their high school class. These predictors play significant roles in shaping graduation outcomes, highlighting the value of both financial resources and academic excellence in driving institutional success. While the model provides a reasonable explanation of graduation rates, some variability remains unexplained, suggesting the potential contribution of additional unmeasured factors.

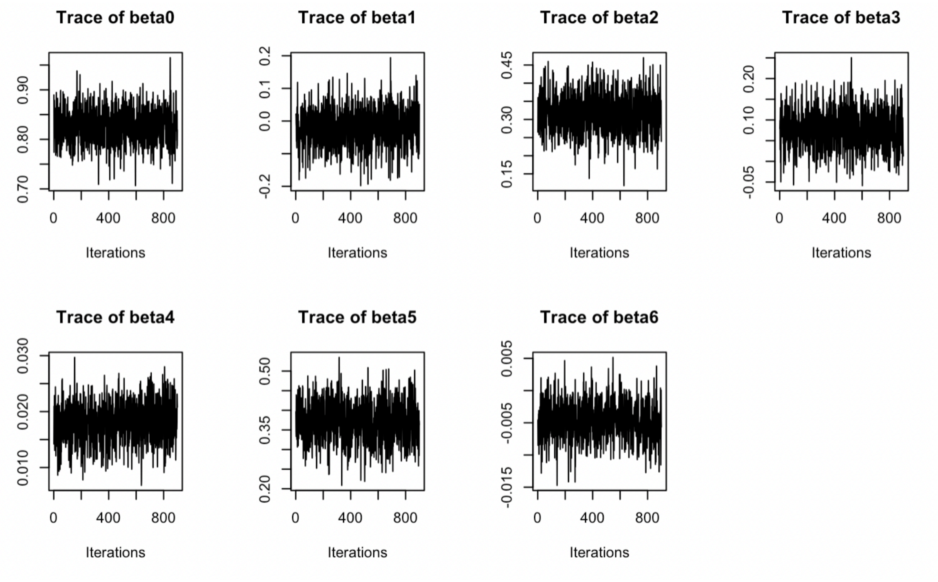
Diagnostic

In this section we have conducted several diagnostic checks to assess the convergence, the goodness of adaptation and the model hypotheses, to be able to guarantee the validity and reliability of the Bayesian model of fixed effects.

The application of these diagnostics are essential for interpreting the results and ensuring that the model accurately captures the relationships between the predictors and graduation rates. In this case , since Model 4 has the lowest Deviance Information Criterion (DIC), all diagnostics were performed on this model to assess its performance and robustness.

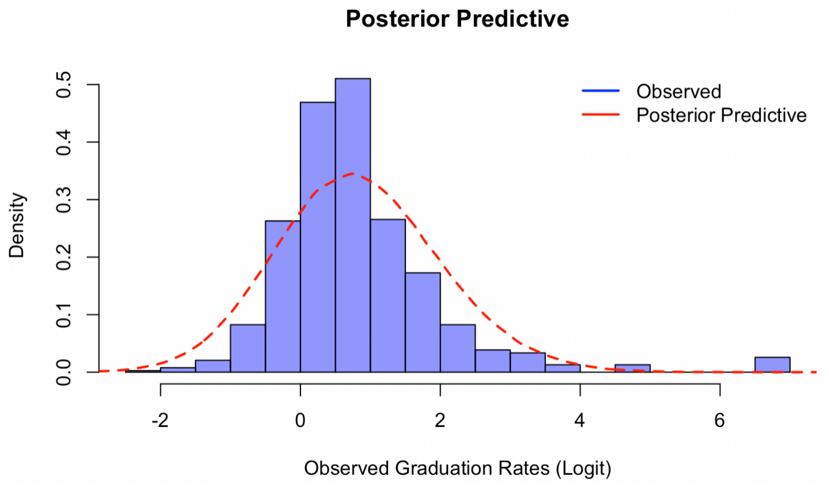
We examined the convergence of the Markov Chain Monte Carlo (MCMC) chains for the model parameters using the trace plots for the fixed-effects-only Model 4. These plots depict the sampling process for each parameter across iterations. The results are as follows:

- The trace plots for all parameters (β_0 to β_6) exhibit stability, as the chains oscillate around a consistent mean value without noticeable trends or systematic deviations.
- The chains demonstrate good mixing, with no obvious patterns or signs of autocorrelation. This suggests that the sampler effectively converged to the target posterior distribution.
- There is no apparent "burn-in" phase in the plots, indicating that the sampler attained the stationary distribution early in the process. The initial burn-in period of 1,000 iterations proved sufficient to ensure convergence.



Posterior Predictive

Moreover, we did posterior predictive checks to assess how well the model captures the observed data. The following graph illustrates the Posterior Predictive distribution for graduation rates (logit-transformed), comparing the observed data distribution (blue histogram) with the posterior predictive distribution (red density curve) generated by the Bayesian model.



The posterior predictive distribution closely aligns with the observed data in terms of the general shape and density. This alignment suggests that the model captures the central tendencies and overall variability in the data reasonably well.

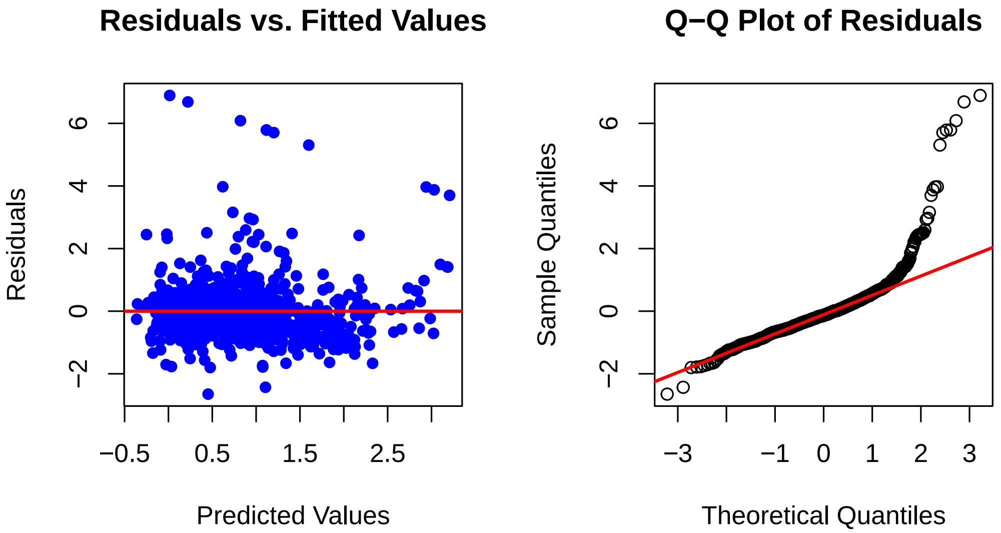
However there appears to be a slight discrepancy in the extreme tails (e.g., above 4 or below -1), where the observed data has more extreme values compared to the predictive distribution. This may indicate that the model does not fully account for the extreme values in the data.

Also around the peak of the observed data (approximately between 0 and 2 in logit space), the posterior predictive distribution closely follows the observed data. This indicates that the model performs well in capturing the most frequent values in the dataset.

So,the model performs well overall, as indicated by the strong alignment between the observed data and the posterior predictive distribution. This supports the validity of the model for predicting graduation rates based on the included predictors.

Residuals Diagnostic

By analyzing residuals, we can identify potential issues such as non-linearity, heteroscedasticity, outliers, or model misspecification, which may affect the validity and reliability of the model's predictions.

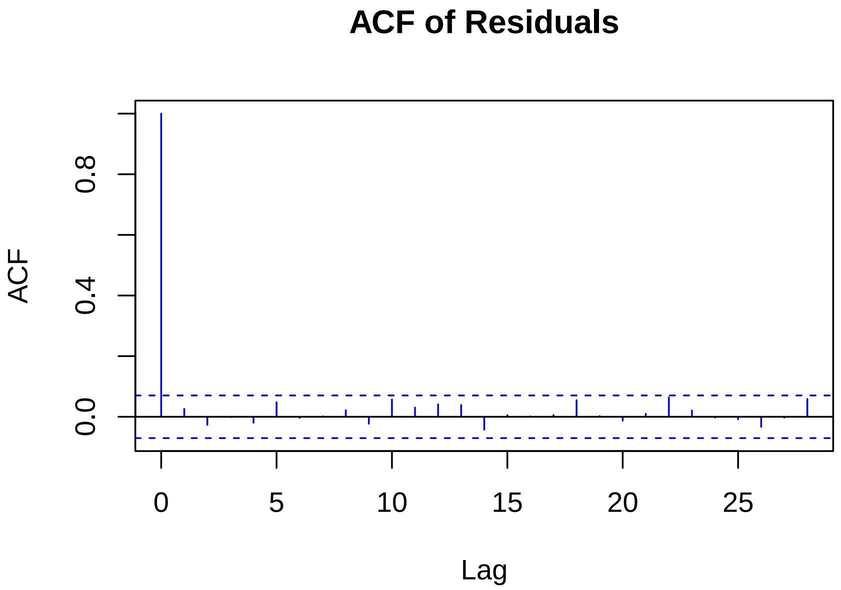


In the first plot we can observe that the residuals appear to be scattered randomly around zero, which is a good indication that the model captures the relationship between predictors and the response variable reasonably well. The spread of residuals is relatively consistent (no visible funnel shape), suggesting that the assumption of homoscedasticity is not violated. However, there are some outliers with large positive residuals at higher fitted values. These outliers could indicate specific institutions with unusual graduation rates not explained by the predictors.

The Q-Q plot compares the quantiles of the residuals to the quantiles of a theoretical normal distribution. In this analysis most of the points lie close to the red line, indicating that the residuals are approximately normally distributed for the central part of the data. However, there are noticeable deviations in the tails, particularly in the upper range, where the residuals deviate significantly from the theoretical line. This suggests potential issues with extreme values (outliers).

Moreover, the ACF plot was used to assess the presence of autocorrelation in the residuals of a model. Autocorrelation measures whether the residuals at one observation are related to residuals at another observation across lags. Ideally, residuals should not exhibit significant autocorrelation, as this would indicate model misspecification or the presence of unmodeled structure in the data.

We can observe that the first spike at lag 0 represents the correlation of residuals with themselves, which is always equal to 1. The spikes at lags greater than 0 are mostly within the blue dashed lines, representing the 95% confidence intervals for no significant autocorrelation. This indicates that the residuals do not exhibit statistically significant autocorrelation at these lags. There is no visible systematic pattern (e.g., a cyclical or decaying structure), which further supports the assumption of no significant autocorrelation. The ACF plot indicates that the residuals are uncorrelated and independent, supporting the adequacy of the model.



The Effective Sample Size (ESS)

The Effective Sample Size (ESS) measures the number of effectively independent samples obtained from the Markov Chain Monte Carlo (MCMC) process, accounting for potential autocorrelation within the chains. A higher ESS indicates that the chains are mixing well and that the posterior estimates are reliable.

We can note that for most parameters (e.g., $\beta_0, \beta_1, \beta_2, \beta_4, \beta_6$), the ESS is either 900 or slightly above 900. These values indicate excellent mixing and sufficient sampling, ensuring the precision of posterior estimates for these parameters. While parameters like β_3 and β_5 show slightly lower ESS values (e.g., $\beta_5 = 771.13$). While these values are lower than 900, they are still acceptable for most inferential purposes.

The ESS for deviance is approximately 938. This indicates reliable sampling for evaluating the overall fit of the model, as deviance measures how well the model explains the data.

```
#    beta0     beta1     beta2     beta3     beta4     beta5     beta6  deviance
# 900.0000 1033.2556  900.0000  858.1638  900.0000  771.1304  900.0000  938.3557
# deviance     mu[1]
# 938.3557  900.0000
```

Comparative analysis with frequentist inference

In this section, we examine and contrast the outcomes of three distinct models applied to the dataset: a basic linear regression model (LM), a frequentist mixed-effects linear model (LMM), and a Bayesian mixed-effects model (BMM). While all three models include the same set of predictors, they differ in their key assumptions and approaches to addressing hierarchical structures within the data.

Simple Linear Regression Model

The linear model (LM) assumes independence of observations and does not account for the nested structure of the data. In this case, the dependent variable is transformed using the logit function, meaning the coefficients represent the change in the log-odds of graduation rates associated with a one-unit change in each predictor. This transformation ensures that the bounded nature of the original outcome variable (graduation rate as a percentage) is properly handled.

The model explains approximately 32.3% of the variability in the log-odds of graduation rates, as indicated by the adjusted R^2. Key findings include: - Outstate Tuition ($\beta_1 = 0.317, p < 0.001$): A one-unit increase in tuition increases the log-odds of graduation rates by 0.317. In odds ratio terms, this is a 37.3% increase in the odds of graduation. - Alumni Donations ($\beta_4 = 0.018, p < 0.001$): For each 1% increase in alumni donation rates, the log-odds of graduation rates increase by 0.018. This corresponds to an odds ratio of a 1.8% increase in the odds of graduation for each additional percentage point of alumni donations. - Top 10% Students ($\beta_5 = 0.367, p < 0.001$): A 1% increase in the percentage of top-performing students increases the log-odds of graduation rates by 0.367, or the odds by 44.4% ($e^{0.367} \approx 1.444$).

Other predictors, such as Expenditure and Student-to-Faculty Ratio, show weaker or non-significant effects. The model explains 32.3% of the variability in the log-odds of graduation rates but leaves considerable residual variability ($RSE = 0.965$), suggesting the need for a mixed-effects model to account for hierarchical effects. In this model, the RSE is 0.965. This means that, on average, the predicted log-odds of graduation rates deviate from the observed log-odds by about 0.965.

```
# lm(formula = Grad.Rate.Logit ~ Outstate + Expend + S.F.Ratio +
#    perc.alumni + Top10perc + Terminal, data = College)

# Residuals:
#     Min      1Q  Median      3Q     Max
# -2.6552 -0.5260 -0.1336  0.3054  6.8858

# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept)  0.829020   0.034643  23.930  < 2e-16 ***
# Outstate     0.316850   0.053886   5.880 6.12e-09 ***
# Expend      -0.018347   0.056874  -0.323   0.7471
# S.F.Ratio    0.079251   0.045527   1.741   0.0821 .
# perc.alumni  0.018101   0.003495   5.179 2.86e-07 ***
# Top10perc    0.367067   0.050079   7.330 5.85e-13 ***
# Terminal    -0.004675   0.002814  -1.661   0.0970 .
---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Residual standard error: 0.965 on 769 degrees of freedom
# Multiple R-squared:  0.3286,  Adjusted R-squared:  0.3234
# F-statistic: 62.73 on 6 and 769 DF,  p-value: < 2.2e-16
```

Frequentist Linear Mixed-Effects Model

The linear mixed-effects model (LMM) incorporates a random intercept to account for variability at the institution type level (public vs. private). By allowing intercepts to vary across groups, the model better captures the hierarchical structure of the data compared to the simple linear regression model.

The variance attributable to differences between public and private institutions ($\sigma_u^2 = 0.001544$) is very small (Std.Dev. = 0.03929), indicating minimal variability in graduation rates explained by institution type. The residual variance ($\sigma^2 = 0.931020$) represents the unexplained variability within institutions. This value is lower than in the simple linear regression model, suggesting that the inclusion of random effects has improved the model's fit by accounting for clustering.

The RSE is derived from the residual variance ($\sigma^2 = 0.931020$), resulting in an RSE of 0.965. This model splits variability into residual variance and random effect variance ($\sigma^2 u$), accounting for institution-level clustering.

The LM has a smaller RSE (0.3234) because it assumes all variability is residual, without accounting for grouping by institutions. The LMM, with an RSE of 0.965, captures additional variability explained by random effects (institution type). This better partitions the variance and provides a more realistic estimate of residual error.

The fixed-effect estimates remain similar to those in the LM but are adjusted for the grouping effect, leading to more precise estimates. For instance, Outstate Tuition ($\beta_1 = 0.310$) and Top 10% Students ($\beta_5 = 0.368$) remain significant, showing consistent positive effects on graduation rates. The REML criterion (2182) indicates better overall model fit compared to the LM, further supporting the value of incorporating random effects to capture between-institution variability.

```
# Linear mixed model fit by REML ['lmerMod']
# Formula: Grad.Rate.Logit ~ Outstate + Expend + S.F.Ratio + perc.alumni +
#          Top10perc + Terminal + (1 | Private)
#   Data: College

# REML criterion at convergence: 2182

# Scaled residuals:
#     Min      1Q  Median      3Q     Max
# -2.7369 -0.5484 -0.1426  0.3127  7.1467

# Random effects:
# Groups   Name        Variance Std.Dev.
# Private  (Intercept) 0.001544 0.03929
# Residual             0.931020 0.96489
# Number of obs: 776, groups:  Private, 2

# Fixed effects:
#              Estimate Std. Error t value
# (Intercept)  0.823419   0.045815  17.972
# Outstate     0.309965   0.055635   5.571
# Expend      -0.016477   0.056990  -0.289
# S.F.Ratio    0.081925   0.045838   1.787
# perc.alumni  0.017959   0.003506   5.122
# Top10perc    0.367776   0.050092   7.342
# Terminal    -0.004406   0.002865  -1.538

# Correlation of Fixed Effects:
#             (Intr) Outstt Expend S.F.Rt prc.lm Tp10pr
# Outstate     0.061
# Expend      -0.016 -0.327
# S.F.Ratio   -0.029  0.193  0.361
# perc.alumni  0.020 -0.314  0.073  0.119
# Top10perc   -0.007 -0.090 -0.397 -0.034 -0.198
# Terminal    -0.047 -0.189 -0.139 -0.149 -0.019 -0.248
```

Bayesian Mixed-Effects Model

Considering the Bayesian Mixed-Effects Model (BMM), estimated using JAGS, provides probabilistic insights into the relationships between predictors and graduation rates. The key predictors, such as Instructional Expenditure ($\beta_2 = 0.293$, 95% CI:[0.169,0.408]) and Top 10% Students ($\beta_5 = 0.364$,95% CI: [0.267,0.463]), show strong positive associations with graduation rates, emphasizing their significance. The credible intervals offer a clear understanding of uncertainty, which integrates frequentist interpretations.

The inclusion of random effects is reflected in the posterior mean estimate for the institution-level standard deviation ($\sigma_u = 0.539$, 95% CI: [0.026,2.785]), capturing variability between public and private institutions. The residual variability ($\sigma = 0.966$, 95% CI: [0.922,1.014]) remains moderate, indicating a well-fitted model.

The Deviance Information Criterion (DIC = 2156.2) has a slightly better fit than the LMM (REML = 2182), suggesting that the Bayesian model provides a marginal improvement in capturing variability in the data. While computationally intensive, the Bayesian approach provides richer insights and flexibility compared to the other methods, making it particularly useful for hierarchical data.

```
# Inference for Bugs model at "6", fit using jags,
#  1 chains, each with 10000 iterations (first 1000 discarded), n.thin = 10
#  n.sims = 900 iterations saved. Running time = 4.233 secs
#           mu.vect sd.vect     2.5%      25%      50%      75%    97.5%
# beta0       0.752   1.178   -0.359    0.749    0.824    0.889    1.442
# beta1      -0.010   0.056   -0.118   -0.051   -0.010    0.028    0.102
# beta2       0.293   0.061    0.169    0.253    0.293    0.335    0.408
# beta3       0.089   0.047    0.002    0.056    0.088    0.124    0.177
# beta4       0.018   0.004    0.011    0.015    0.018    0.020    0.025
# beta5       0.364   0.050    0.267    0.329    0.363    0.397    0.463
# beta6      -0.004   0.003   -0.010   -0.006   -0.004   -0.002    0.002
# sigma       0.966   0.024    0.922    0.950    0.966    0.982    1.014
# sigma_u     0.539   3.080    0.026    0.065    0.121    0.280    2.785
# deviance 2147.659   4.131 2141.566 2144.738 2147.058 2150.136 2156.789

# DIC info (using the rule: pV = var(deviance)/2)
# pV = 8.5 and DIC = 2156.2
# DIC is an estimate of expected predictive error (lower deviance is better).
```

Conclusions

The primary objective of this study was to explore the factors influencing graduation rates among U.S. colleges and universities, emphasizing the differences between private and public institutions. Through a combination of descriptive statistics, frequentist methods, and Bayesian mixed-effects modeling, we analyzed the effects of the predictors at the institution and student level, taking into account the hierarchical structure of the data.

The Bayesian mixed-effects model proved to be a powerful approach for addressing the complexities of the dataset. It incorporated both fixed effects, capturing key predictors like out-of-state tuition, alumni donations, and the percentage of top-performing students, and random effects to account for variability between institution types (private vs. public). This hierarchical structure allowed for a better understanding of the factors driving graduation rates, particularly the role of institutional characteristics.

In addressing the research question, we conclude that the key findings from the analysis include: - Institutional Predictors: Out-of-state tuition and alumni engagement emerged as significant predictors of graduation rates, suggesting that financial and reputational factors are critical in shaping student outcomes. - Student Quality: The proportion of top-performing high school students was strongly associated with higher graduation rates, emphasizing the importance of incoming student preparedness. - Institutional Differences: While the random effects in the Bayesian model accounted for variability between private and public institutions, the relatively small variance component suggests that much of the variation in graduation rates is explained by individual institutional characteristics rather than the binary distinction between public and private colleges.

The model comparison using DIC values indicated that while Model 4 (fixed effects only) had the lowest DIC, Model 3 (with random effects) provided richer insights into between-institution variability. The Bayesian framework, with its probabilistic interpretation of parameter estimates and credible intervals, offered a more intuitive understanding of uncertainty compared to frequentist confidence intervals.

This study highlights the utility of Bayesian mixed-effects models in educational research, particularly for datasets with hierarchical structures.

```r
library("ISLR2")
data("College")
#Data cleaning
# Convert Private to a binary variable
College$Private <- ifelse(College$Private == "Yes", 1, 0)

# Handle missing values
sum(is.na(College)) # Check for missing values
College <- na.omit(College) # Remove rows with missing values

# Check for outliers in Grad.Rate
summary(College$Grad.Rate)
College[College$Grad.Rate > 100, ]
College <- College[College$Grad.Rate <= 100, ] # Remove unrealistic graduation rates > 100%

# Standardize continuous variables
# Variables to standardize
scale_vars <- c("Outstate", "Expend", "S.F.Ratio", "Top10perc", "Top25perc",
  "Apps", "Accept", "Enroll", "F.Undergrad", "P.Undergrad",
  "Room.Board", "Books", "Personal"
)

# Apply standardization
College[scale_vars] <- scale(College[scale_vars])

#HIST Grad.Rate
hist(College$Grad.Rate,
     breaks = 20,
     probability = TRUE,
     col = "steelblue",
     main = "Distribution of Graduation Rates (%)",
     xlab = "Graduation Rate (%)",
     border = "black",
     xlim = c(0, 120))

curve(dnorm(x, mean = mean(College$Grad.Rate, na.rm = TRUE),
            sd = sd(College$Grad.Rate, na.rm = TRUE)),
      col = "red",
      lwd = 2,
      add = TRUE)
```

```r
#Transformation of Grad.Rate
# Convert Graduation Rate to proportion
College$Grad.Rate.Prop <- College$Grad.Rate / 100
# Replace 0 and 1 with a small epsilon (e.g., 0.001 and 0.999)
epsilon <- 0.001
College$Grad.Rate.Prop <- pmax(epsilon, pmin(College$Grad.Rate.Prop, 1 - epsilon))
# Apply logit transformation
College$Grad.Rate.Logit <- log(College$Grad.Rate.Prop / (1 - College$Grad.Rate.Prop))

#HIST Grad.Rate.Logit
# Calcolare media e deviazione standard della variabile logit trasformata
mean_logit <- mean(College$Grad.Rate.Logit, na.rm = TRUE)
sd_logit <- sd(College$Grad.Rate.Logit, na.rm = TRUE)

hist(College$Grad.Rate.Logit,
     breaks = 30,
     probability = TRUE,
     main = "Distribution of Logit-Transformed Graduation Rates",
     xlab = "Logit(Graduation Rate)",
     col = "steelblue",
     border = "black")

curve(dnorm(x, mean = mean_logit, sd = sd_logit),
      col = "red",
      lwd = 2,
      lty = 2,
      add = TRUE)



library(ggplot2)
#Boxplot
ggplot(College, aes(x = factor(Private, labels = c("Public", "Private")), y = Grad.Rate)) +
  geom_boxplot(fill = c("purple", "yellow"), alpha = 0.7) +
  labs(
    title = NULL,
    x = "Institution Type",
    y = "Graduation Rate (%)"
  ) +
  theme_minimal()


library(ggcorrplot)
library(dplyr)
table_summary <- College %>%
  group_by(Private) %>%
  summarise(
    Mean_Grad_Rate = mean(Grad.Rate, na.rm = TRUE),
    Proportion = n() / nrow(College)
  )
```

```r
kable(
  table_summary,
  col.names = c("Type", "Mean Graduation Rate", "Proportion"),
  caption = "Mean Graduation Rate and Proportion by College Type"
)


# correlation matrix for all numeric variables
cor_matrix <- cor(College[, sapply(College, is.numeric)], use = "complete.obs")

# heatmap
ggcorrplot(cor_matrix,
           method = "square",
           lab = FALSE,
           colors = c("red", "white", "blue"),
           title = "Heatmap",
           hc.order = TRUE) +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))



#####
library(knitr)
library(ggplot2)
library(mice)
library(patchwork)
library(tidyr)
library(rjags)
library(R2jags)
library(coda)
library(lme4)
library(readr)
# Center the continuous predictors
College$Expend <- as.vector(scale(College$Expend, center = TRUE, scale = FALSE))
College$Outstate <- as.vector(scale(College$Outstate, center = TRUE, scale = FALSE))
College$S.F.Ratio <- as.vector(scale(College$S.F.Ratio, center = TRUE, scale = FALSE))
College$perc.alumni <- as.vector(scale(College$perc.alumni, center = TRUE, scale = FALSE))
College$Top10perc <- as.vector(scale(College$Top10perc, center = TRUE, scale = FALSE))
College$Terminal <- as.vector(scale(College$Terminal, center = TRUE, scale = FALSE))
College$F.Undergrad <- as.vector(scale(College$F.Undergrad, center = TRUE, scale = FALSE))
College$Books <- as.vector(scale(College$Books, center = TRUE, scale = FALSE))
College$Personal <- as.vector(scale(College$Personal, center = TRUE, scale = FALSE))
College$Enroll <- as.vector(scale(College$Enroll, center = TRUE, scale = FALSE))

# data1 for JAGS
jags_data1 <- list(
  y = College$Grad.Rate.Logit,
  group = as.numeric(as.factor(College$Private)), #college grouping (1-public, 2-private)
  N = nrow(College),    #number of college
  J = length(unique(College$Private))     #number of group
```

```
)

model_code1 <- "
model {
  for (i in 1:N) {
    y[i] ~ dnorm(mu[i], tau)
    mu[i] <- beta0 + u[group[i]]

    # Posterior predictive distribution
    yrep[i] ~ dnorm(mu[i], tau)  # Posterior predictions for y

  }

  #random effects
  for (j in 1:J) {
    u[j] ~ dnorm(0, tau_u)
  }

  # Priors for fixed effects
  beta0 ~ dnorm(0, 0.001)

  # Priors for variances
  tau ~ dgamma(0.001, 0.001)
  sigma <- 1 / sqrt(tau)

  tau_u ~ dgamma(0.001, 0.001)
  sigma_u <- 1 / sqrt(tau_u)
}
"

# Parameters to monitor
params1 <- c("beta0", "sigma", "sigma_u")

# Run the JAGS model1
set.seed(123)
jags_model1 <- jags(
data = jags_data1,
parameters.to.save = params1,
model.file = textConnection(model_code1),
n.chains = 1, # Number of chains
n.iter = 10000, # Total iterations
n.burnin = 1000, # Burn-in iterations
n.thin = 10, # Thinning factor
DIC = TRUE # Calculate DIC for model comparison
)


# data2 for JAGS
jags_data2 <- list(
  y = College$Grad.Rate.Logit,   #outcome
  x1 = College$Expend,     #predictor 1
  x2 = College$Outstate,   #predictor 2
  x3 = College$S.F.Ratio,  #predictor 3
  group = as.numeric(as.factor(College$Private)), #college grouping (1-public, 2-private)
  N = nrow(College),   #number of college
  J = length(unique(College$Private))    #number of group
)


model_code2 <- "
model {
  for (i in 1:N) {
    y[i] ~ dnorm(mu[i], tau)
    mu[i] <- beta0 + beta1 * x1[i] + beta2 * x2[i] + beta3 * x3[i] + u[group[i]]

    # Posterior predictive distribution
    yrep[i] ~ dnorm(mu[i], tau)  # Posterior predictions for y
  }

  for (j in 1:J) {
    u[j] ~ dnorm(0, tau_u)
  }

  # Priors for fixed effects
  beta0 ~ dnorm(0, 0.001)
  beta1 ~ dnorm(0, 0.001)
  beta2 ~ dnorm(0, 0.001)
  beta3 ~ dnorm(0, 0.001)

  # Priors for variances
  tau ~ dgamma(0.001, 0.001)
  sigma <- 1 / sqrt(tau)

  tau_u ~ dgamma(0.001, 0.001)
  sigma_u <- 1 / sqrt(tau_u)
}
"
# Parameters to monitor
params2 <- c("beta0", "beta1", "beta2", "beta3", "sigma", "sigma_u")

# Run the JAGS model2
set.seed(123)
jags_model2 <- jags(
data = jags_data2,
parameters.to.save = params2,
model.file = textConnection(model_code2),
n.chains = 1, # Number of chains
n.iter = 10000, # Total iterations
n.burnin = 1000, # Burn-in iterations
n.thin = 10, # Thinning factor
DIC = TRUE # Calculate DIC for model comparison
)
```

```
# Data for JAGS Model 3
jags_data3 <- list(
  y = College$Grad.Rate.Logit,    # Outcome variable
  x1 = College$Expend,           # Predictor 1
  x2 = College$Outstate,         # Predictor 2
  x3 = College$S.F.Ratio,        # Predictor 3
  x4 = College$perc.alumni,      # Predictor 4
  x5 = College$Top10perc,        # Predictor 5
  x6 = College$Terminal,         # Predictor 6
  group = as.numeric(as.factor(College$Private)),  # Grouping variable (1-public, 2-private)
  N = nrow(College),             # Number of observations
  J = length(unique(College$Private))  # Number of groups
)

# JAGS Model 3 Code
model_code3 <- "
model {
  for (i in 1:N) {
    y[i] ~ dnorm(mu[i], tau)
    mu[i] <- beta0 + beta1 * x1[i] + beta2 * x2[i] + beta3 * x3[i] +
             beta4 * x4[i] + beta5 * x5[i] + beta6 * x6[i] + u[group[i]]
    # Posterior predictive distribution
    yrep[i] ~ dnorm(mu[i], tau)
  }

  # Random effects for groups
  for (j in 1:J) {
    u[j] ~ dnorm(0, tau_u)
  }

  # Priors for fixed effects
  beta0 ~ dnorm(0, 0.001)
  beta1 ~ dnorm(0, 0.001)
  beta2 ~ dnorm(0, 0.001)
  beta3 ~ dnorm(0, 0.001)
  beta4 ~ dnorm(0, 0.001)
  beta5 ~ dnorm(0, 0.001)
  beta6 ~ dnorm(0, 0.001)


  # Priors for variances
  tau ~ dgamma(0.001, 0.001)
  sigma <- 1 / sqrt(tau)

  tau_u ~ dgamma(0.001, 0.001)
  sigma_u <- 1 / sqrt(tau_u)
}
"

# Parameters to monitor
```

```
params3 <- c("beta0", "beta1", "beta2", "beta3", "beta4", "beta5", "beta6", "sigma", "sigma_u")

# Run JAGS Model 3
set.seed(123)
jags_model3 <- jags(
  data = jags_data3,
  parameters.to.save = params3,
  model.file = textConnection(model_code3),
  n.chains = 1, # Number of chains
  n.iter = 10000, # Total iterations
  n.burnin = 1000, # Burn-in iterations
  n.thin = 10, # Thinning factor
  DIC = TRUE # Calculate DIC for model comparison
)


# Data for JAGS Model 4
jags_data4 <- list(
  y = College$Grad.Rate.Logit,     # Outcome variable
  x1 = College$Expend,            # Predictor 1
  x2 = College$Outstate,          # Predictor 2
  x3 = College$S.F.Ratio,         # Predictor 3
  x4 = College$perc.alumni,       # Predictor 4
  x5 = College$Top10perc,         # Predictor 5
  x6 = College$Terminal,          # Predictor 6
  N = nrow(College)               # Number of observations
)

# JAGS Model 4 Code
model_code4 <- "
model {
  for (i in 1:N) {
    y[i] ~ dnorm(mu[i], tau)
    mu[i] <- beta0 + beta1 * x1[i] + beta2 * x2[i] + beta3 * x3[i] +
             beta4 * x4[i] + beta5 * x5[i] + beta6 * x6[i]
    # Posterior predictive distribution
    yrep[i] ~ dnorm(mu[i], tau)  # Posterior predictions for y
  }

  # Priors for fixed effects
  beta0 ~ dnorm(0, 0.001)
  beta1 ~ dnorm(0, 0.001)
  beta2 ~ dnorm(0, 0.001)
  beta3 ~ dnorm(0, 0.001)
  beta4 ~ dnorm(0, 0.001)
  beta5 ~ dnorm(0, 0.001)
  beta6 ~ dnorm(0, 0.001)


  # Priors for variance
  tau ~ dgamma(0.001, 0.001)
  sigma <- 1 / sqrt(tau)
}
"

# Parameters to monitor
params4 <- c("beta0", "beta1", "beta2", "beta3", "beta4", "beta5", "beta6", "sigma")

# Run JAGS Model 4
set.seed(123)
jags_model4 <- jags(
  data = jags_data4,
  parameters.to.save = params4,
  model.file = textConnection(model_code4),
  n.chains = 1, # Number of chains
  n.iter = 10000, # Total iterations
  n.burnin = 1000, # Burn-in iterations
  n.thin = 10, # Thinning factor
  DIC = TRUE # Calculate DIC for model comparison
)



# DIC
dic_values <- c(
model1 = jags_model1$BUGSoutput$DIC,
model2 = jags_model2$BUGSoutput$DIC,
model3 = jags_model3$BUGSoutput$DIC,
model4 = jags_model4$BUGSoutput$DIC
)
dic_table <- data.frame( Model = names(dic_values), DIC = as.numeric(dic_values)
)
kable(dic_table, caption = "DIC Values")

#model4
print(jags_model3)



# JAGS 4 WITH MU AND U FOR DIAGNOSTICS
# Define parameters to monitor
params4diag <- c("beta0", "beta1", "beta2", "beta3", "beta4", "beta5",
            "beta6", "sigma", "yrep", "mu")
set.seed(123)
jags_model4diag <- jags(
data = jags_data4,
parameters.to.save = params4diag,
model.file = textConnection(model_code4),
n.chains = 1, # Number of chains
n.iter = 10000, # Total iterations
n.burnin = 1000, # Burn-in iterations
n.thin = 10, # Thinning factor
```

```
DIC = TRUE # Calculate DIC for model comparison
)

# TRACEPOLOTS FOR CONVERGENCE
mcmc_samples <- as.mcmc(jags_model4diag$BUGSoutput$sims.matrix)
par(mfrow=c(2,4))
traceplot(mcmc_samples[, c("beta0", "beta1", "beta2", "beta3",
                           "beta4", "beta5", "beta6")])



#POSTERIOR PREDICTIVE CHECK
# Extract posterior predictive samples
posterior_pred <- jags_model4diag$BUGSoutput$sims.list$yrep
observed <- jags_data4$y
# histogram for observed data
hist(observed, breaks = 30, col = rgb(0, 0, 1, 0.5), prob = TRUE,
     main = "Posterior Predictive",
     xlab = "Observed Graduation Rates (Logit)")
lines(density(as.vector(posterior_pred)), col = "red", lwd = 2, lty=2)
legend("topright", legend = c("Observed", "Posterior Predictive"),
       col = c("blue", "red"), lwd = 2, bty = "n")



#DIAGNOSTIC RESIDUALS
fitted_values <- jags_model4diag$BUGSoutput$sims.list$mu  # Valori predetti
residuals <- observed - colMeans(fitted_values)           # Residui

# Residuals vs. Predicted values
par(mfrow = c(1, 2))
plot(colMeans(fitted_values), residuals,
     main = "Residuals vs. Fitted Values",
     xlab = "Predicted Values",
     ylab = "Residuals",
     pch = 16, col = "blue")
abline(h = 0, col = "red", lwd = 2)

# Q-Q Plot
qqnorm(residuals, main = "Q-Q Plot of Residuals")
qqline(residuals, col = "red", lwd=2)
par(mfrow = c(1, 1))

# Hist residuals
hist(residuals, breaks = 30, col = "lightblue",
     main = "Histogram of Residuals",
     xlab = "Residuals", prob = TRUE)
lines(density(residuals), col = "red", lwd = 2)

# Autocorrelation
acf(residuals, main = "ACF of Residuals", col = "blue")
```

```r
# EFFECTIVE SAMPLE SIZE (ESS)
head(effectiveSize(mcmc_samples), 10) # Show the first 10 rows


## FREQUENTIST / BAYESIAN
# Simple linear regression (Frequentist)
simple_lm <- lm(Grad.Rate.Logit ~ Outstate + Expend + S.F.Ratio + perc.alumni + Top10perc
                +Terminal,  data = College)
print("Linear model summary:")
summary(simple_lm)

# Frequentist mixed-effects model
frequentist_model <- lmer(Grad.Rate.Logit ~ Outstate + Expend + S.F.Ratio + perc.alumni +
                            Top10perc + Terminal  + (1 | Private), data = College)
print("Frequentist mixed-effects model summary:")
summary(frequentist_model)

# Bayesian mixed-effects model (Model 3 using JAGS)
# Assuming jags_model3 is already fitted and available
print("Bayesian mixed-effects model summary:")
print(jags_model3)
```

Last update by CM+LT: Sun Jan 26 20:07:57 2025