Bayesian Modeling of Student Academic Outcomes

Eromonsele Okojie

April 21, 2025

Abstract

We develop hierarchical Bayesian multinomial logistic regression models to predict students' academic outcomes (Dropout, Enrolled, Graduate) using demographic, socio-economic, and performance features. We compare a flat model, a hierarchical model with program-level varying intercepts and slopes, and an extended model with additional program-level random slopes on unemployment rate. Model fit is assessed via PSIS-LOO, and uncertainty quantification is provided through posterior predictive checks and credible intervals for program- and student-level effects.

1 Introduction

Predicting student dropout and graduation outcomes is crucial for early intervention. We apply Bayesian multinomial logistic regression to a UCI dataset of 4,424 students with 36 predictors, including admission grade, demographics, and first-year performance. We introduce hierarchical structure by allowing program-specific intercepts and admission-grade slopes, and further extend to random slopes on local unemployment rate.

2 Data

2.1 Data Source

Data come from the UCI Machine Learning Repository: Predicting Students' Dropout and Academic Success (Martins *et al.*, 2021). The dataset had no missing values; features include admission grade, age, semester grades, macroeconomic indicators, gender, scholarship status, and degree program (36 categories).

2.2 Preprocessing

Column names were stripped of extra whitespace; the target variable `Target` was trimmed and mapped to integers Dropout=0, Enrolled=1, Graduate=2. We standardized numeric predictors and one-hot encoded binary covariates (Gender, Scholarship holder). Degree programs were factorized to obtain a program index for hierarchical modeling.

3 Model Specification

We model the three-category outcome via a softmax link. Let $y_i \in \{0, 1, 2\}$ be the outcome for student i, and X_i be the vector of predictors (standardized). The baseline category is Dropout ($y = 0$). For $k = 1, 2$ we define linear predictors as follows:

$$\eta_{ik} = a_{0,k} + a_{\mathrm{prog}[i],k} + b_{\mathrm{adm},k}\, x_i^{\mathrm{adm}} + \sum_{j=1}^{J} b_{j,k}\, x_{ij}, \quad \Pr(y_i = k) = \frac{\exp(\eta_{ik})}{\sum_{\ell=0}^{2} \exp(\eta_{i\ell})},$$

where $\eta_{i0} = 0$ and J is the number of predictors besides admission grade. Full model definitions are given in `model.py`.

4 Posterior Inference

We ran NUTS sampling with 4 chains, 2,000 warmup and 3,000 draws each, target acceptance 0.95. All $\hat{R} \approx 1.0$ and effective sample sizes > 200, with no divergent transitions.

4.1 Hyperparameter Summaries

Table 1 reports posterior means, standard deviations, and 94% HDIs for the hierarchical model's hyperparameters $(\alpha_0, \alpha_1, \beta_0, \beta_1, \sigma_a, \sigma_b)$ and key fixed effects.

Parameter	Mean	SD	2.5% HDI	97.5% HDI	$\hat{R}$
α_0	0.01	0.59	-1.12	1.12	1.00
α_1	0.50	0.22	0.07	0.89	1.00
β_0	0.06	0.06	-0.05	0.17	1.00
β_1	0.02	0.06	-0.09	0.12	1.00
σ_a	1.21	0.17	0.88	1.56	1.00
σ_b	0.17	0.05	0.08	0.26	1.00

Table 1: Hierarchical model hyperparameter posteriors.

The posterior mean of α_1 is 0.50 (94% HDI [0.07, 0.89]), indicating that programs with higher average admission grades have substantially higher baseline odds of graduation versus dropout. The partial-pooling standard deviation $\sigma_a = 1.21$ shows substantial variation in baseline odds across programs. By contrast, β_1's 94% HDI overlaps zero, suggesting weak evidence for heterogeneity in admission-grade slopes.

4.2 Trace Plots

Figure 1 shows trace and density plots for all six hyperparameters.

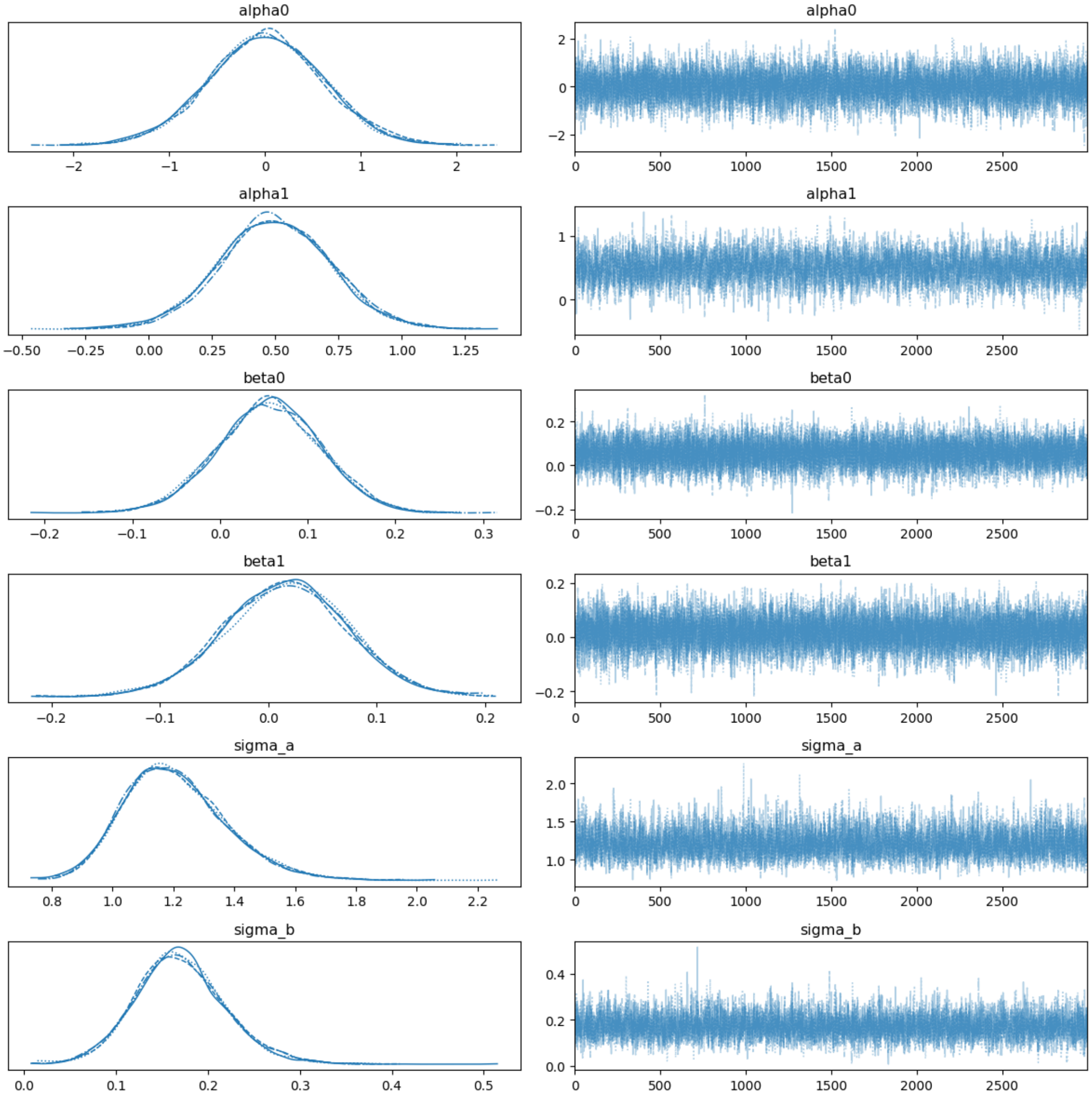

Figure 1: Trace (right) and posterior density (left) for each hyperparameter.

All $\widehat{R} \approx 1.00$ and effective sample sizes exceed 200, with no divergences, indicating good mixing and convergence.

4.3 Posterior Predictive Check

Figure 2 overlays observed counts by outcome with 100 posterior predictive replicates and the predictive mean.

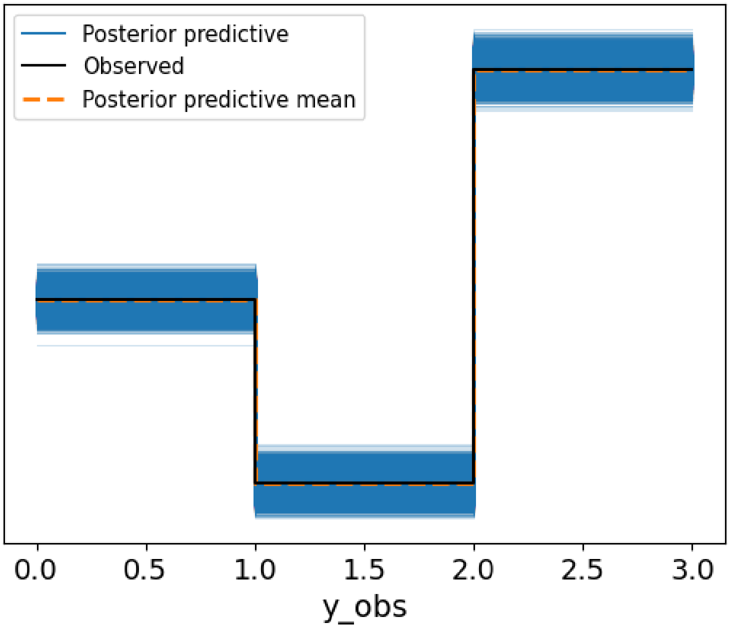

Figure 2: Posterior predictive overlays: blue bands are replicates, dashed orange line is predictive mean, black is observed.

The blue bands are 100 posterior predictive replicates, the dashed orange line is the predictive mean, and black dots show the observed counts. The model captures the overall outcome frequencies well, though the "Enrolled" category shows slightly wider variability, reflecting honest uncertainty.

5 Model Comparison via LOO

We computed PSIS-LOO on the deviance scale. Table 2 shows that the extended model has the lowest LOO deviance, followed by hierarchical, then flat.

Model	elpd_loo	p_loo	elpd_diff	weight
Extended	6405.70	75.48	0.00	0.80
Hierarchical	6411.09	65.41	-5.38	0.10
Flat	6984.36	20.21	-578.66	0.09

Table 2: LOO comparison (deviance scale) of the three models.

On the deviance scale (lower is better), the extended model (elpd_loo=6405.70) outperforms hierarchical (6411.09) and flat (6984.36). The large difference between flat and hierarchical highlights the benefit of partial pooling.

6 Program-Level Effects

Figure 3 gives 95% CIs for each program's admission-grade slope $b_{\text{prog\_adm}}$. Table 3 lists the intercepts a_{prog} for the first 10 programs as an example.

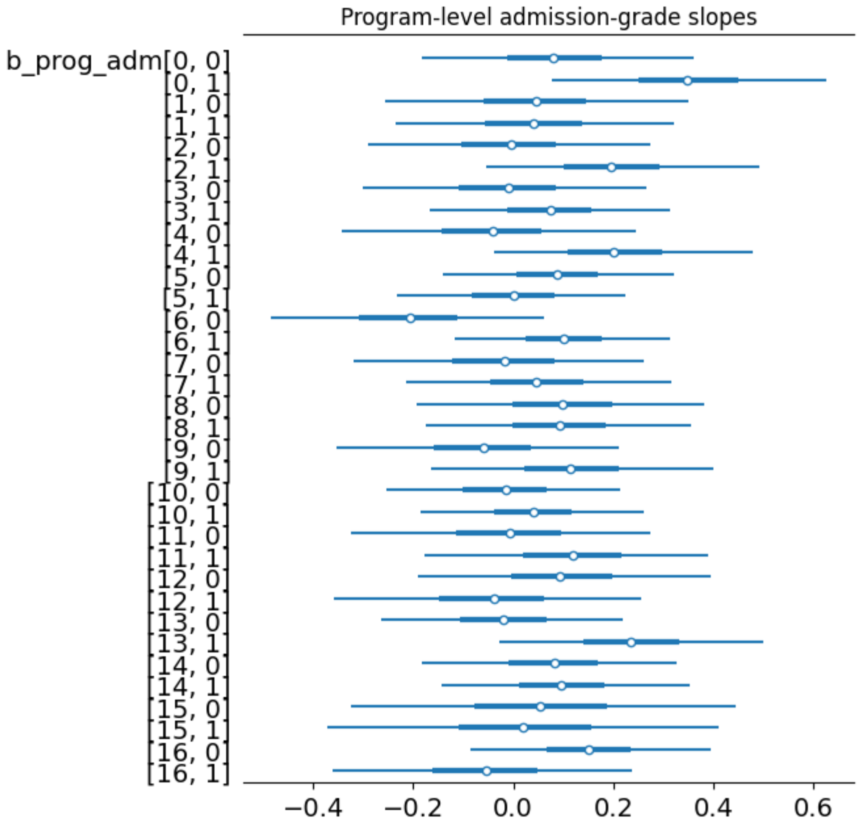

Figure 3: Forest plot of program-specific admission-grade slopes.

Some programs (e.g., Program 0) show strong positive slopes of admission grade on graduation odds (95% CI excludes zero), while others are effectively flat, indicating heterogeneity in grade-slope effects across majors.

Program	Mean	SD	2.5%	97.5%	$\hat{R}$
$a_{\text{prog}}[0,0]$	1.594	0.663	0.308	2.889	1.00
$a_{\text{prog}}[0,1]$	5.657	0.757	4.145	7.119	1.00
$a_{\text{prog}}[1,0]$	-0.314	0.646	-1.592	0.912	1.00
$a_{\text{prog}}[1,1]$	-0.304	0.644	-1.559	0.961	1.00
$a_{\text{prog}}[2,0]$	0.254	0.653	-0.961	1.589	1.00
$a_{\text{prog}}[2,1]$	0.149	0.650	-1.093	1.468	1.00
$a_{\text{prog}}[3,0]$	-0.708	0.649	-1.975	0.560	1.00
$a_{\text{prog}}[3,1]$	-0.279	0.638	-1.463	1.037	1.00
$a_{\text{prog}}[4,0]$	0.057	0.674	-1.295	1.324	1.00
$a_{\text{prog}}[4,1]$	0.992	0.658	-0.287	2.287	1.00

Table 3: Posterior summaries for program-level intercepts (a_{prog}).

7 Discussion and Conclusion

The extended hierarchical model—with random slopes on unemployment rate—achieved the best LOO score, indicating improved fit while accounting for program heterogeneity. Hierarchical partial pooling stabilizes estimates for smaller programs, and Bayesian inference provides full uncertainty quantification at both student and program levels.

Limitations We did not perform a prior predictive check, and our grouping is limited to degree programs; future work could explore alternative groupings (e.g. by department), additional interactions, or more informative priors.

Code Availability

All of the analysis code for this project are publicly available at:

`https://github.com/seleokojie/bayes-student-outcome`

References

M. V. Martins, D. Tolledo, J. Machado, L. M. T. Baptista, and V. Realinho. Early prediction of student's performance in higher education: a case study. In *Trends and Applications in Information Systems and Technologies*, Advances in Intelligent Systems and Computing, vol. 1. Springer, 2021. doi:10.1007/978-3-030-72657-7_16.