DATA 2025

10 - 12 June, 2025

Bilbao - Spain

14TH INTERNATIONAL CONFERENCE ON DATA SCIENCE, TECHNOLOGY AND APPLICATIONS

Predictors of Freshmen Attrition: A Case Study of Bayesian Methods and Probabilistic Programming

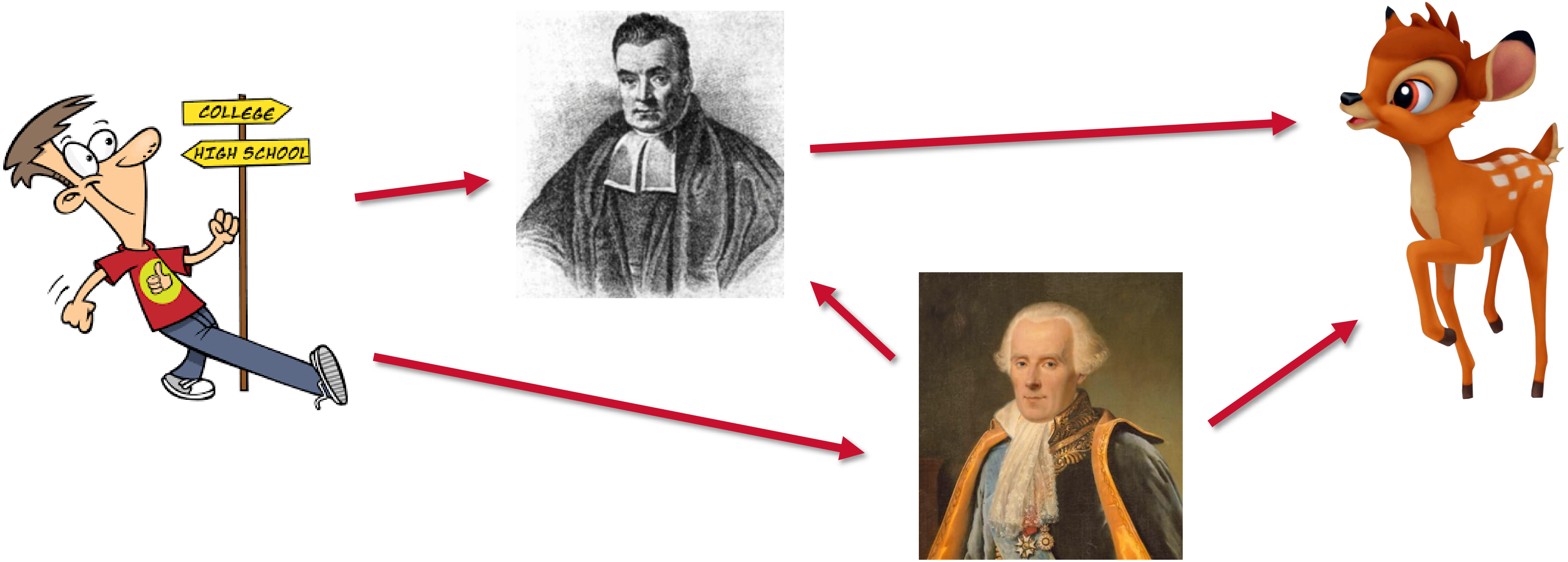
Eitel J.M. Lauría

Poughkeepsie, NY, USA

MARIST UNIVERSITY

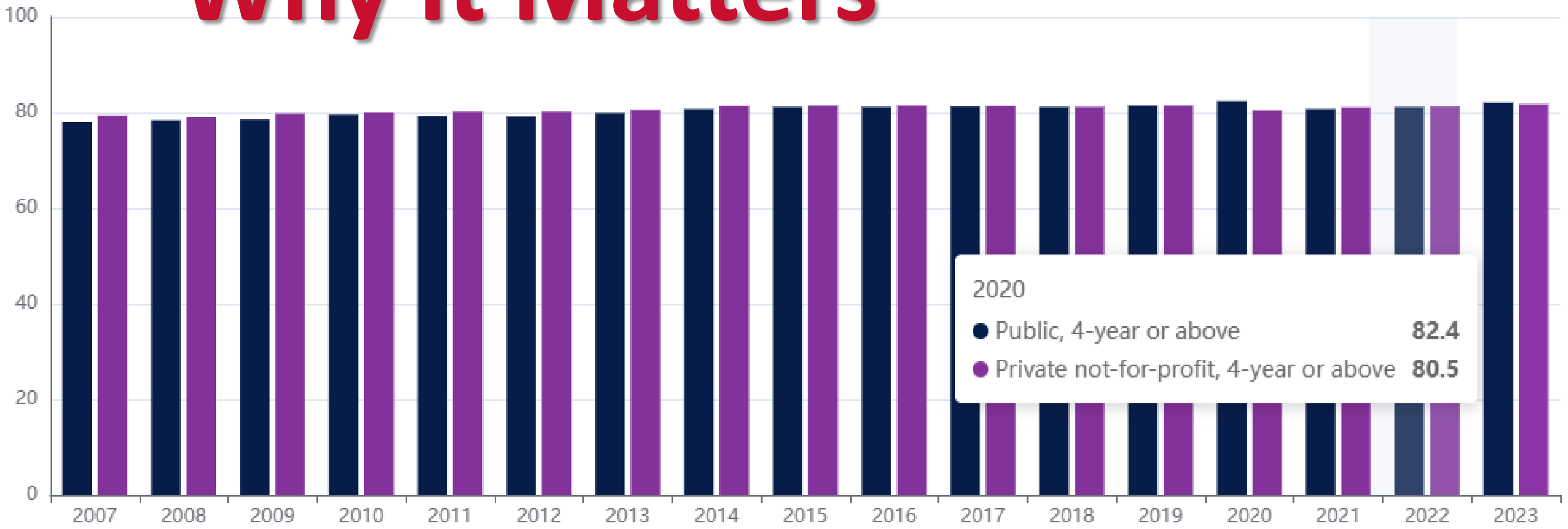
Let’s start with a short quiz…

What do they have in common?



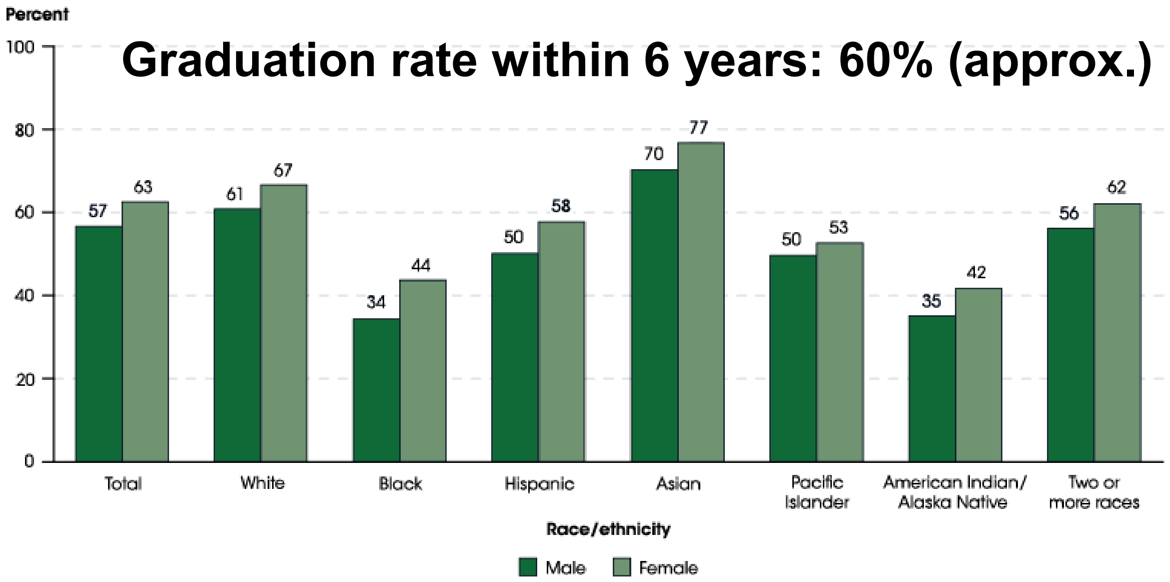
Why It Matters

Freshmen attrition: ~10% nationwide



Source: National Center for Education Statistics, Retrieved 5/5/2025

Graduation rate within 6 years: 60% (approx.)



Source: National Center for Education Statistics. Cohort entry year 2010

Graduation Rate within 4 years

36%

Source: U.S. Dept. of Education, Postsecondary Education Data System (2009)

Research Questions

1. How do student demographics, high school and university academic performance, and student activities affect the odds of freshmen attrition?
2. Is there considerable fluctuation in freshmen attrition across different academic years and among different schools?
3. Bayesian vs. frequentist models: how do they compare?

The Data

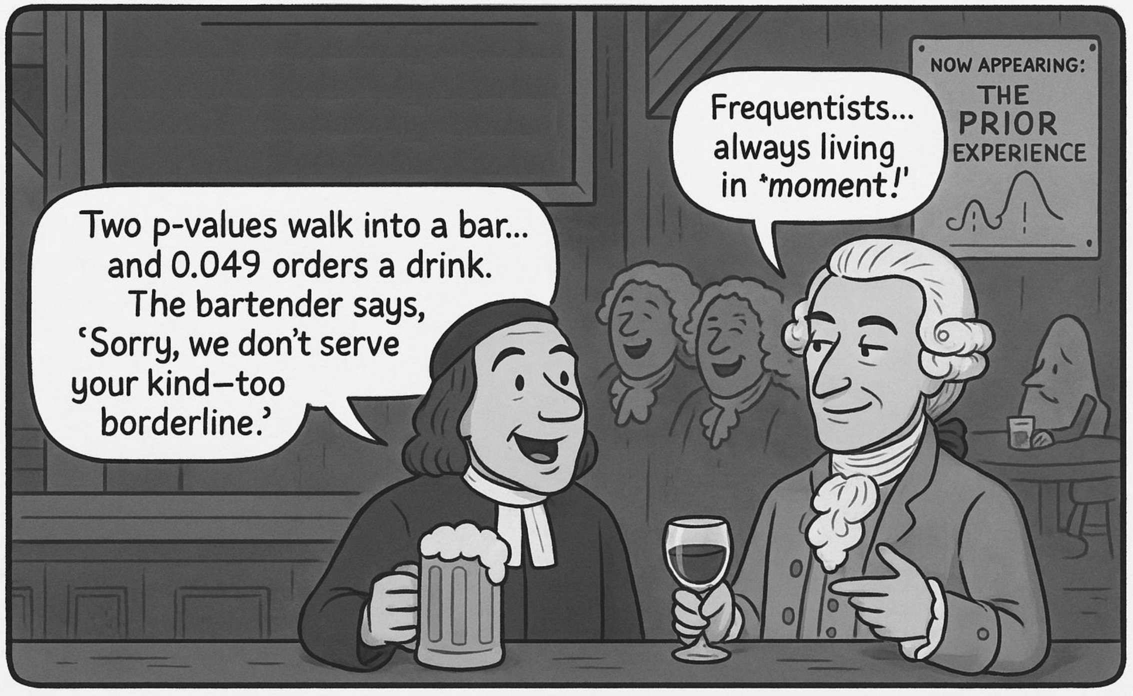
- 9 years of data (2012-2018, and 2021-2022)
- 2019-2020 skipped (COVID)
- Each record: accepted and registered freshman student in the Fall of the corresponding academic year.
- 10921 records from 6 schools, with 1154 instances of attrition.
- 10.6% did not return
- Data was imputed using KNN
- Numeric variables were scaled as z-scores
- Outliers, Correlations and Variance Inflation Factor (VIF) checked.
- Predictors had a VIF < 5, meaning multicollinearity is not a major concern.

Identifier	**Description**
Academic Performance	
EffectiveGPA (academic year)	Numeric
isDeansList (made it to Dean's list)	Binary (1/0)
TutoringClassCount (classes tutored in)	Numeric
HSGPA (high school GPA)	Numeric
NumAPCourses (taken during high school)	Numeric
Demographics	
UScitizen	Binary (1/0)
Gender	Binary (F, M)
StudentofColor	Binary (1/0)
isFirstGeneration (college student)	Binary (1/0)
DistanceFromHome (miles)	Numeric
Institutional and Enrollment Factors	
isCampusWorkStudy	Binary (1/0)
isDivisionI (athlete)	Binary (1/0)
WaitListed (before admitted)	Binary (1/0)
Financial Aid and Need	
EFC (Expected Family Contribution, in $)	Numeric
UnmetNeed (after financial aid, in $)	Numeric
HasLoans	Binary (1/0)
PellAmount (federal grant, in $)	Numeric
AcademicYear	Discrete
School ((CC, CO, LA, SB, SI, SM)	Discrete
didNotReturnNextFall (response variable)	Binary (1/0)

Why Bayesian?

Bayes' Theorem: $P(\theta \mid X) = P(X \mid \theta) \cdot P(\theta) / P(X)$

- ✓ Embraces and quantifies uncertainty
- ✓ Gives distributions, not just point estimates
- ✓ Prior knowledge = regularization
- ✓ No overreliance on p-values



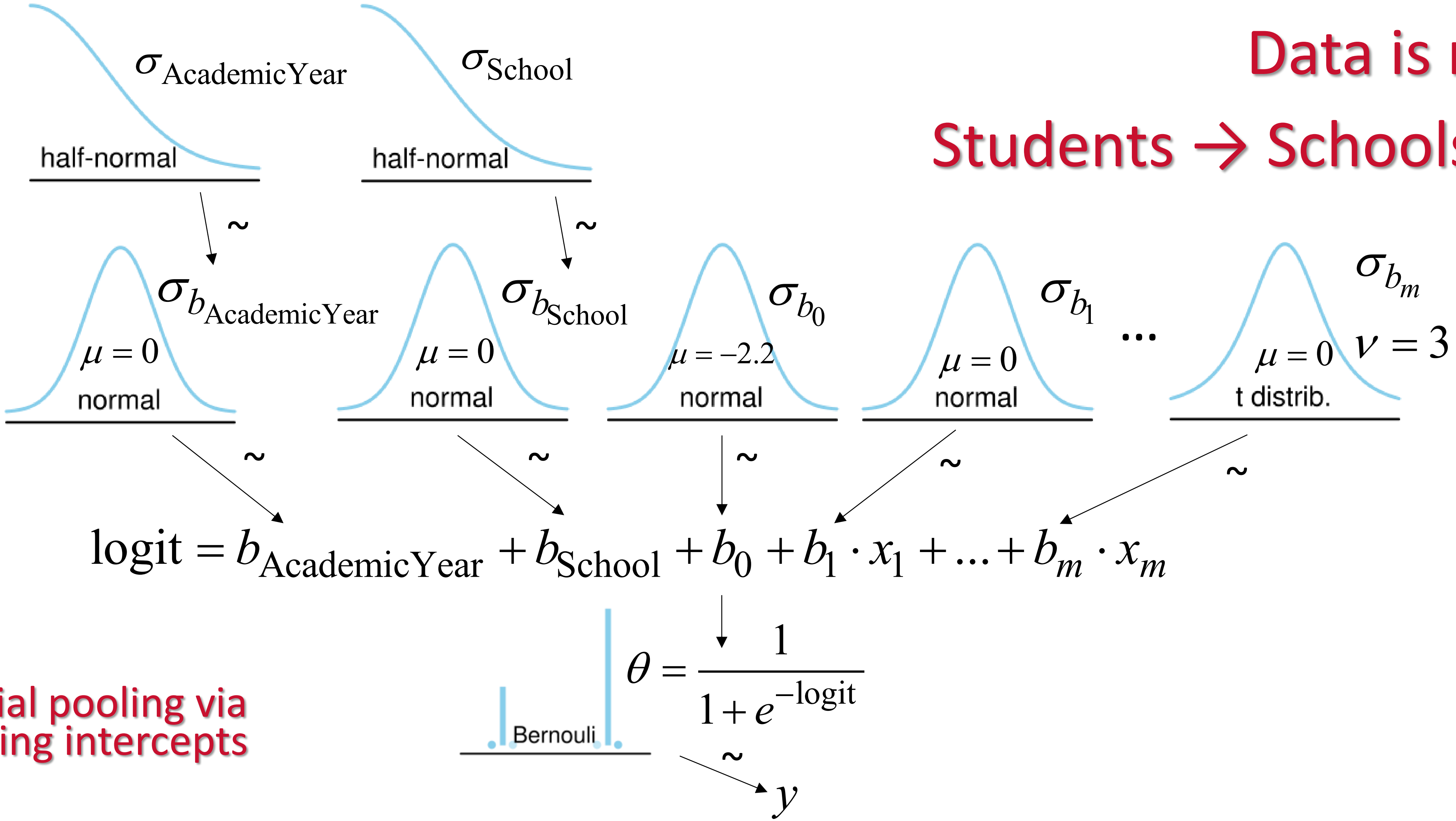
Bayes in Educational Research

- Bayes: underused in education analytics
- Lots of work with ML and traditional stats
- Few studies use Bayesian inference
  - ➢Relegated to certain niches:
    - ✓Bayesian knowledge tracing (in intelligent tutoring systems)
    - ✓NLP and text mining
    - ✓Bayesian nets
- Try searching these keywords in Google Scholar: student retention (or attrition), predictors and Bayesian / MCMC / Markov chain Monte Carlo / variational inference / NUTS

Hierarchical Models

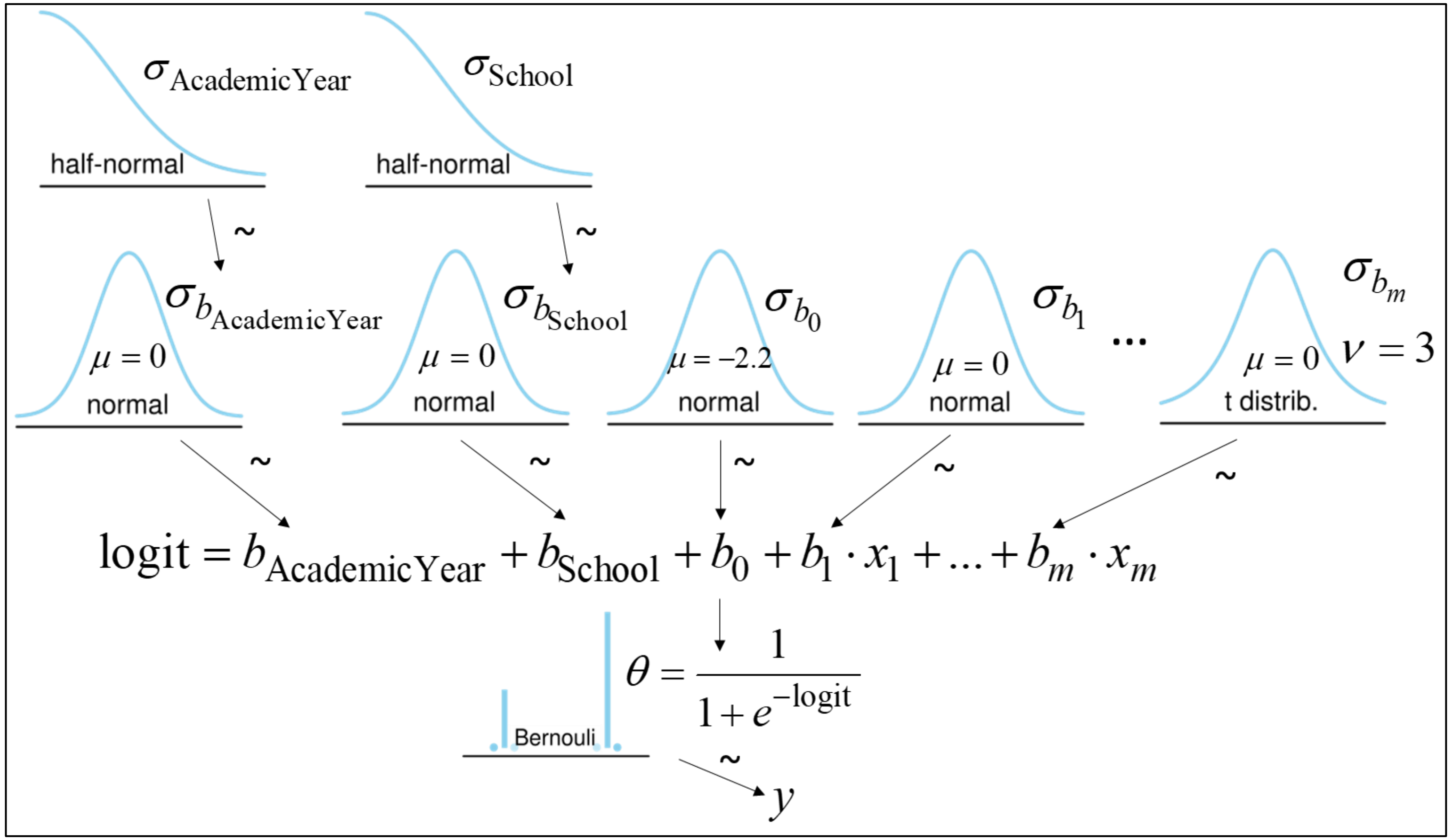
Data is nested:

Students → Schools, Years



Partial pooling via varying intercepts

Hierarchical Models



$$\sigma_{b_{\text{AcademicYear}}} \sim \text{HalfNormal}(\sigma_{\text{AcademicYear}})$$

$$\sigma_{b_{\text{School}}} \sim \text{HalfNormal}(\sigma_{\text{School}})$$

$$b_{\text{AcademicYear}} \sim \text{Normal}(0, \sigma_{b_{\text{AcademicYear}}})$$

$$b_{\text{School}} \sim \text{Normal}(0, \sigma_{b_{\text{School}}})$$

$$b_0 \sim \text{Normal}(\mu_{b_0} = -2.2, \sigma_{b_0} = 1.0)$$

$$b_j \sim P_\beta \quad \text{for} \quad j = 1, 2, \ldots, m$$

$$\text{logit} = b_{\text{AcademicYear}} + b_{\text{School}} + b_0 + \sum_{j=1}^{m} b_j x_j$$

$$\theta = \frac{1}{1 + \exp(-\text{logit})}$$

$$y \sim \text{Bernoulli}(p = \theta)$$

Setting the Priors

- The choice of the Intercept's prior as Normal(-2.2, 1.0) is based on the approximate 10% attrition rate.

$$P(y=1) = \frac{e^{Intercept}}{1+e^{Intercept}} \approx\sim 0.1 \Rightarrow Intercept \approx\sim -2.2$$

- StudentT(ν=3, ν=0, σ=2.5) on correlated predictors or predictors with moderate outliers.
  - ν=3 allows for some large deviations and σ=2.5 keeps the prior weakly informative.
- Normal weakly informative priors -Normal(0,2.5) for all other predictors.
- For group effects (School and AcademicYear), HalfNormal(2.5).
  - A σ = 2.5 allows for moderate variation while keeping the group-specific intercepts within a rather similar scale as the fixed-effects intercept.

$$\begin{aligned}
\sigma_{b_{\text{AcademicYear}}} &\sim \text{HalfNormal}(\sigma_{\text{AcademicYear}}) \\
\sigma_{b_{\text{School}}} &\sim \text{HalfNormal}(\sigma_{\text{School}}) \\
b_{\text{AcademicYear}} &\sim \text{Normal}(0, \sigma_{b_{\text{AcademicYear}}}) \\
b_{\text{School}} &\sim \text{Normal}(0, \sigma_{b_{\text{School}}}) \\
b_0 &\sim \text{Normal}(\mu_{b_0} = -2.2, \sigma_{b_0} = 1.0) \\
b_j &\sim P_\beta \quad \text{for} \quad j = 1, 2, \ldots, m \\
\text{logit} &= b_{\text{AcademicYear}} + b_{\text{School}} + b_0 + \sum_{j=1}^{m} b_j x_j \\
\theta &= \frac{1}{1+\exp(-\text{logit})} \\
y &\sim \text{Bernoulli}(p = \theta)
\end{aligned}$$

Software Platform

Sorry…

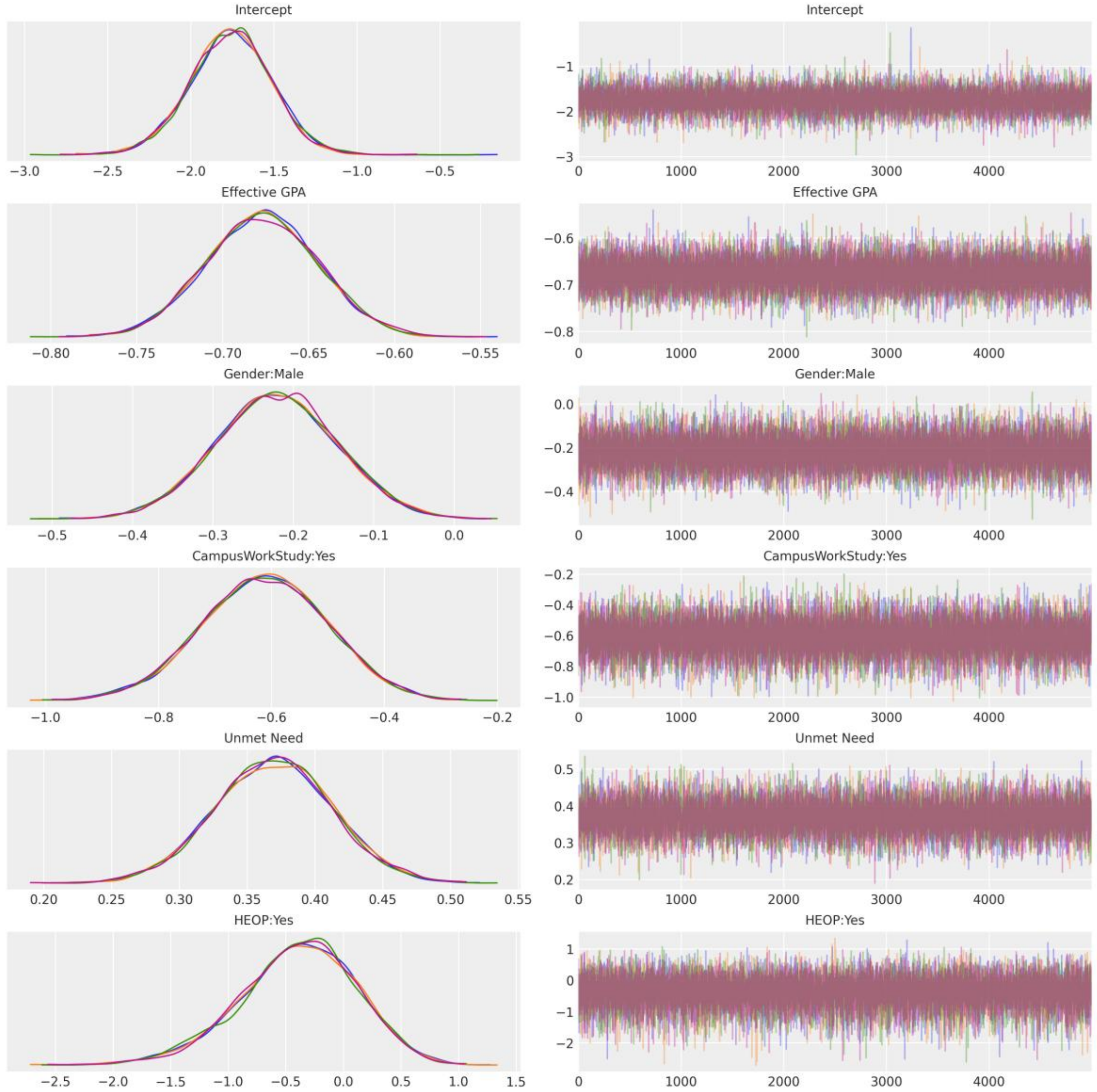
- Bambi (**Ba**yesian **M**odel **B**uilding **I**nterface), a Python package for generalized linear models built on top of PyMC to model the hierarchical logistic regression.
- The ArviZ package for exploratory analysis, diagnostics and to produce visualizations.
- Bayesian models used the No U-Turn (NUTS) algorithm to obtain the posterior distribution samples of the regression parameters.
- We chose the NumPyro backend implementation of the NUTS sampler.

Not this one.

This one ☺

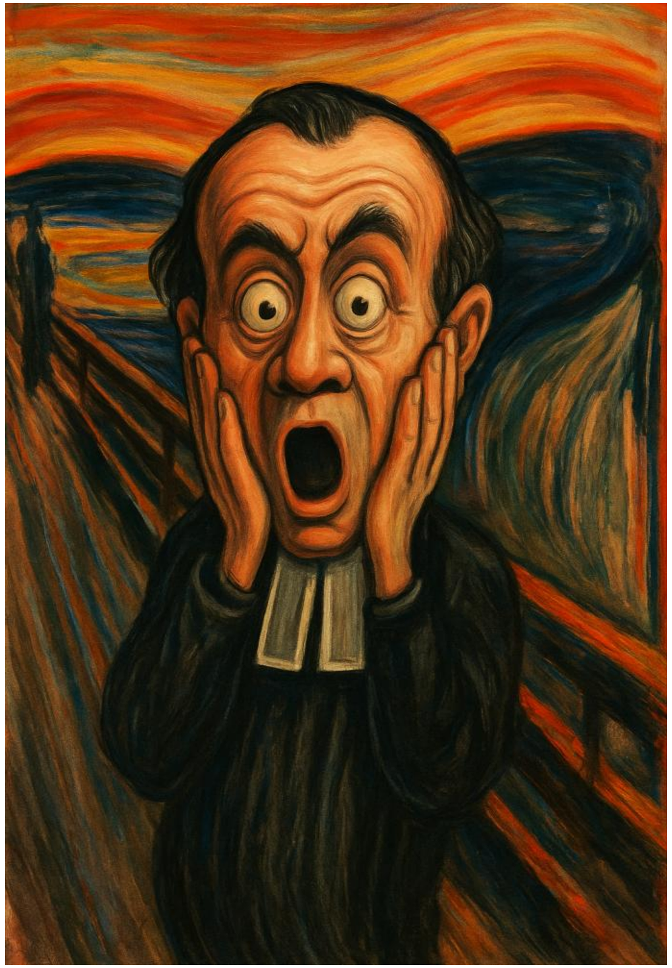
MCMC Chains and Convergence

- Samples of the posterior distributions for the logistic regression parameters were computed using 4 chains of 5000 samples each, with a warm-up period of 1000 samples for each of the chains.
- Chains were tested and no divergences were found (they mix well).
- Convergence of the model was assessed using $\hat{R} < 1.01$ as the threshold for acceptable convergence.



Posterior Metrics and Model Quality

95% HDI	The high-density interval summarizes the range of most credible values of a parameter within a certain probability mass. When 95% HDI includes zero, the regression coefficient is not statistically significant.
% of 95%HDI within ROPE	A measure of the practical significance of the regression coefficients. ROPE: region of practical equivalence. ROPE range = [-0.2,0.2]
$\hat{R} = \sqrt{\dfrac{\frac{n-1}{n}\sigma_W + \frac{1}{n}\sigma_B}{\sigma_W}}$	A measure of MCMC convergence. σ_B is the between-variance (the average of the variances of each of the chains), and σ_W is the within-variance, measuring the variability between the means of the chains
$\mathrm{WAIC} = -2 \cdot (\mathrm{LPPD} - \mathrm{penalty}),$ $\mathrm{LPPD} = \sum_{i=1}^{n} \log \frac{1}{S} \sum_{s=1}^{S} \mathrm{Var}(P(y_i \mid \theta^s))$ $\mathrm{penalty} = \sum_{i=1}^{n} \mathrm{Var}(P(y_i \mid \theta^s))$	Widely applicable information criterion: it is used both for model comparison and to measure the model's predictive performance (how well the model performs when making predictions on new data)
$\mathrm{LOO} = \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} \frac{P(y_i \mid \theta_{-i}^{s})}{\hat{w}_i^s} \right)$	Pareto-smoothed importance sampling leave-one-out cross-validation measures out-of-sample prediction accuracy from a fitted model.



Bayesian version of "The Scream," Edvard Munch, 1893

Fixed Effects Results

Component	mean	sd	hdi_2.5%	hdi_97.5%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat	% 95% HDI within ROPE
Intercept	-2.341	0.270	-2.863	-1.808	0.002	0.002	11857.0	11173.0	1.0	0.000
EFC	0.010	0.036	-0.063	0.078	0.000	0.000	20261.0	11400.0	1.0	100.000
EffectiveGPA	-0.771	0.037	-0.842	-0.698	0.000	0.000	18953.0	13041.0	1.0	0.000
HSGPA	0.129	0.043	0.044	0.214	0.000	0.000	20415.0	12165.0	1.0	91.765
isDeansList[1.0]	0.571	0.095	0.383	0.756	0.001	0.000	20163.0	12725.0	1.0	0.000
NumAPCourses	-0.053	0.039	-0.129	0.024	0.000	0.000	24348.0	12380.0	1.0	100.000
USCitizen[1.0]	-0.271	0.205	-0.678	0.127	0.001	0.001	26992.0	10809.0	1.0	40.621
UnmetNeed	0.370	0.044	0.285	0.456	0.000	0.000	16588.0	12386.0	1.0	0.000
WaitListed[1.0]	0.110	0.116	-0.118	0.331	0.001	0.001	29026.0	11945.0	1.0	70.824
DistanceFromHome	0.122	0.028	0.067	0.176	0.000	0.000	29011.0	12249.0	1.0	100.000
Gender[M]	-0.234	0.075	-0.382	-0.087	0.001	0.000	22534.0	12650.0	1.0	38.305
HasLoans[1.0]	-0.210	0.077	-0.357	-0.056	0.001	0.000	21872.0	12397.0	1.0	47.841
isCampusWorkStudy[1.0]	-0.612	0.114	-0.833	-0.389	0.001	0.000	29701.0	11619.0	1.0	0.000
isDivisionI[1.0]	-0.014	0.099	-0.204	0.179	0.001	0.001	29319.0	11552.0	1.0	98.956
isFirstGeneration[1.0]	0.097	0.104	-0.105	0.300	0.001	0.001	24489.0	12078.0	1.0	75.309
PellAmount	-0.181	0.041	-0.262	-0.103	0.000	0.000	22582.0	13008.0	1.0	61.006
StudentOfColor[1.0]	0.165	0.101	-0.037	0.360	0.001	0.001	22907.0	12177.0	1.0	59.698
TutoringClassCount	-1.802	0.466	-2.732	-0.990	0.004	0.003	18164.0	9474.0	1.0	0.000

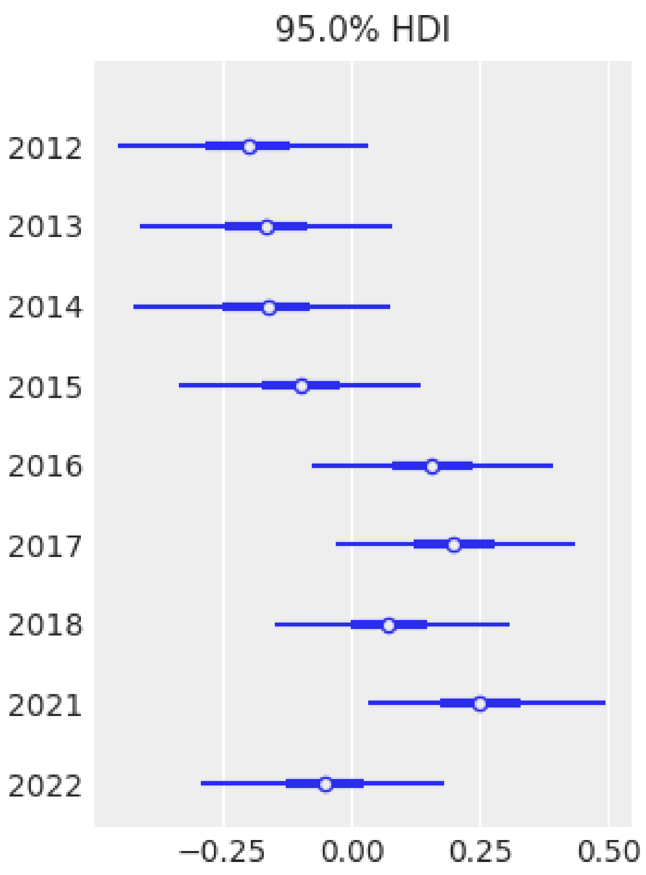
Strong predictors:

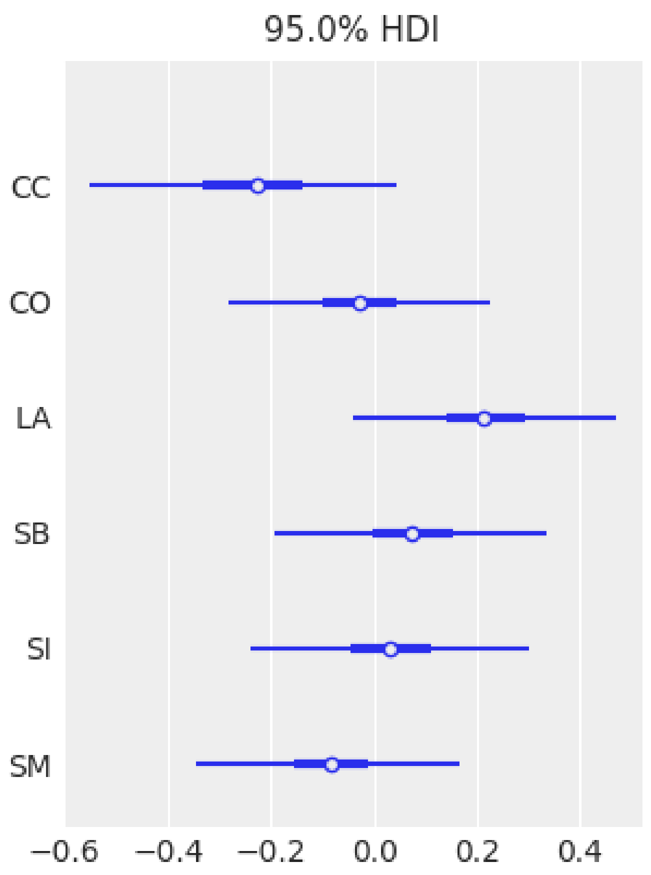
- *GPA* (-), *Unmet Need* (+)
- *Tutoring* (-), *Work-study* (-)
- *Gender* (male = less attrition)
- *Dean's List* (+) (???)

Group Effects

Random intercepts:

- Academic Year (COVID bump)
- School (LA > CC in attrition)
- Odds ratio LA vs. CC ≈ 1.57





Bayes vs. Frequentist

Component	Estimate	SE	2.5% CI	97.5% CI	OR	OR 2.5% CI	OR 97.5% CI	Z-stat	P-val	Significance
Intercept	-3.153	10.675	-24.075	17.769	0.043	0.000	5.214e+07	-0.295	0.768	
EFC	0.013	0.035	-0.056	0.083	1.013	0.946	1.086	0.376	0.707	
EffectiveGPA	-0.769	0.037	-0.841	-0.696	0.464	0.431	0.498	-20.854	0.000	***
HSGPA	0.127	0.043	0.042	0.212	1.136	1.043	1.237	2.926	0.003	**
isDeansList [1.0]	0.573	0.096	0.385	0.761	1.774	1.470	2.140	5.985	0.000	***
NumAPCourses	-0.054	0.039	-0.130	0.022	0.947	0.878	1.022	-1.405	0.160	
USCitizen [1.0]	-0.281	0.204	-0.680	0.118	0.755	0.507	1.126	-1.379	0.168	
UnmetNeed	0.370	0.044	0.285	0.456	1.448	1.330	1.577	8.507	0.000	***
WaitListed [1.0]	0.113	0.116	-0.114	0.340	1.120	0.893	1.405	0.978	0.328	
DistanceFromHome	0.124	0.028	0.069	0.179	1.132	1.072	1.196	4.431	0.000	***
Gender [M]	-0.240	0.076	-0.389	-0.091	0.787	0.678	0.913	-3.154	0.002	**
HasLoans [1.0]	-0.211	0.077	-0.362	-0.060	0.810	0.697	0.942	-2.741	0.006	**
isCampusWorkStudy [1.0]	-0.612	0.113	-0.833	-0.391	0.542	0.435	0.677	-5.425	0.000	***
isDivisionI [1.0]	-0.011	0.099	-0.206	0.184	0.989	0.814	1.202	-0.107	0.915	
isFirstGeneration [1.0]	0.105	0.103	-0.097	0.306	1.110	0.907	1.358	1.016	0.310	
PellAmount	-0.180	0.040	-0.258	-0.102	0.835	0.773	0.903	-4.536	0.000	***
StudentOfColor [1.0]	0.166	0.101	-0.033	0.365	1.180	0.968	1.440	1.635	0.102	
TutoringClassCount	-5.744	50.898	-105.501	94.014	0.003	0.000	6.756e+40	-0.113	0.910	
Random Effects	Academic Year: Var = 0.034, Std = 0.184		School: Var = 0.023, Std = 0.152							
Evaluation metrics	Log-likelihood: -3227.289		AIC: 6494.577							

$$\text{logit} = b_0 + b_{\text{AcademicYear}} + b_{\text{School}} + \sum_{j=1}^{m} \beta_j x_j$$

$$b_{\text{AcademicYear}} \sim \text{Normal}(0, \sigma_{\text{AcademicYear}})$$

$$b_{\text{School}} \sim \text{Normal}(0, \sigma_{\text{School}})$$

$$p = \frac{1}{1 + \exp(-\text{logit})}$$

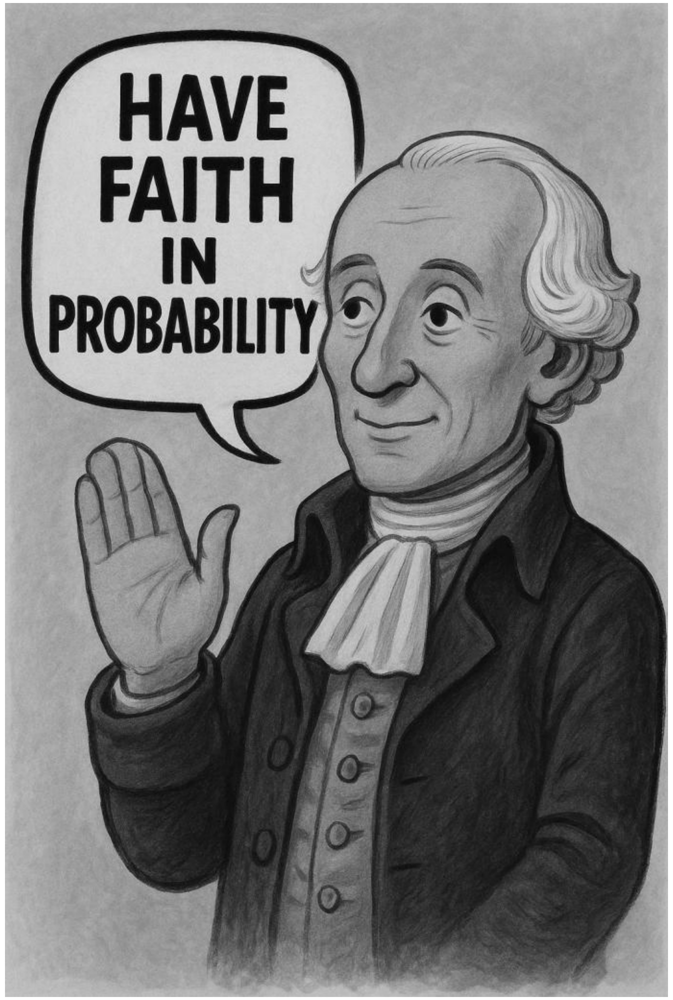
- Used pymer4 (GLMM)
- Check *Intercept* and *Tutoring*
- Bayes shrinks extreme estimates, better with small groups, and uncertainty is built in

Takeaways

- A guideline on how to analyze and report findings in the context of Bayesian methods and probabilistic programming.
- College academic performance, financial need, gender, tutoring, and work-study program participation have a significant effect on the likelihood of freshmen attrition.
- Fluctuations across time and schools may require potential customized intervention strategies.
- Actionable findings to stakeholders, administrators, and decision-makers in higher education.

Takeaways (and call to action)

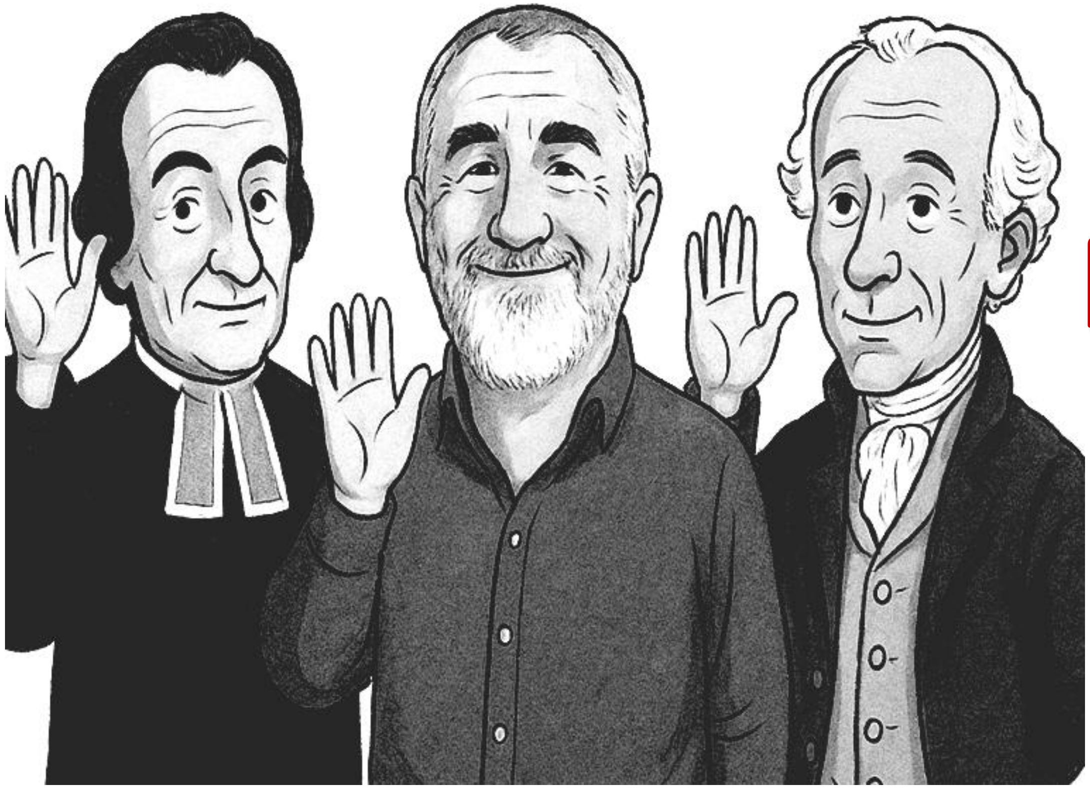
- Bayesian models = transparent, flexible
- Probabilistic programming = scalable
- More than prediction → better decisions
- Try Bayesian tools (e.g. Bambi/PyMC)
- Think probabilistically



Questions?

Eskerrik asko!

Thank you!



Merci!

¡Muchas Gracias!