Modeling Postsecondary Retention Rates in the United States

Emily Mo

12/2/2019

Why Model Retention Rates?

Retention rates are one of the most commonly used measures of an educational institution's success in educating its students and preparing them for successful careers, and a high retention rate is a badge of honor that institutions tend to strive for. However, an institution's retention rate is a product of not only the institution's accessibility to students, but also external factors that impact students—thus, retention rate as a standalone metric may not be very informative about an institution's educational efficacy. I used data from the Integrated Postsecondary Education System Data System (IPEDS) in order to construct multilevel Binomial models of retention rates based on several institutional characteristics and offerings. Because these models only consider predictors on the institutional level and includes no predictors about institutions' student bodies, they are not intended to—and in fact are not able to—paint a complete picture of the factors that impact retention rates. Instead, the models serve to pinpoint how several helpful educational features within institutions may relate to retention rates across the US, but also to demonstrate that it can be misleading to portray institutional success metrics as a direct product of an institution's efforts, due to the inevitable extra-curricular influences that can be confounded with such educational features.

The data used in these models consisted of institutions in IPEDS records that had retention rate and public/private status information (public, private for-profit, private not-for-profit) available from 2017. All variables used were collected in 2017, which was the most recent year for which all variables of interest were available on the IPEDS data portal. I chose to use distance learning options, recreational options, and life experience credit as predictors because they are offerings that are presumably implemented with the purpose of making college a more welcoming place: distance learning makes educational attainment more accessible to those who may be working or may have difficulty physically attending a class, credit for life experiences acknowledges that accreditable learning opportunities do not all occur within the classroom, and recreational options allow students to take a break from traditional academics while still learning new skills. [1] Additionally, a lower student-faculty ratio may make it easier for students to get individualized attention from faculty, which can make the educational experience feel more personalized.

Models

Using this cleaned and narrowed data set, I fit two Binomial regression models with increasingly complex hierarchical structure. Because a Laplacian fitting algorithm produced nearly unidentifiable parameter estimates [2], I used a Bayesian fitting approach, which allowed for more nuanced evaluation of the parameters' convergence, as well as permitted probabilistic statements about the parameters' true values given the data.

[1] IPEDS defines "credit for life experiences" as "Credit earned by students for what they have learned through independent study, noncredit adult courses, work experience, portfolio demonstration, previous licensure or certification, or completion of other learning opportunities (military, government, or professional).

[2] Results from the Laplacian approach are detailed in Appendix A.)

Model A: Multilevel model with varying intercepts per state and public/private status

Using a Markov Chain Monte Carlo algorithm, I fit a multilevel binomial model with varying intercepts for state and for public/private status. A multilevel structure was appropriate because it would account for the expected correlation of retention rates between colleges within the same state or within the same public/private status. A binomial regression was appropriate because each institution's retention rate reflects a group of individuals, each of whom were retained or not retained—thus, each institution's retention rate can be viewed as a binomial random variable. The institutional characteristics previously mentioned were used as the model's non-varying predictors. The prior for this model was only specified with respect to the residual variance and the covariance structure of the varying intercepts. For all of those parameters, I supplied a weakly informative Inverse Chi-Square prior with a scale parameter of 1 and 2 degrees of freedom, which can be expressed as an Inverse Gamma distribution with a shape parameter of 1 and a scale parameter of 1. [3] The model was fit with 40,000 iterations, with a burn-in period of 6,000 and a thinning factor of 30. These parameters yielded a roughly stationary trace plot and a roughly symmetric distribution plot for each of the fixed coefficients in the model, which suggested that the parameter estimates were trustworthy. The coefficient estimates and between-group variance estimates are as follows [4]:

```
##
##  Iterations = 10001:39971
##  Thinning interval  = 30
##  Sample size  = 1000
##
##  DIC: 11261844
##
##  G-structure:  ~us(1):state
##
##                                post.mean l-95% CI u-95% CI eff.samp
## (Intercept):(Intercept).state   0.08524  0.05062   0.1213     1000
##
##                ~us(1):pubpriv
##
##                                  post.mean l-95% CI u-95% CI eff.samp
## (Intercept):(Intercept).pubpriv     1.023  0.09918     2.54     1000
##
##  R-structure:  ~units
##
##       post.mean l-95% CI u-95% CI eff.samp
## units    0.8829    0.843   0.9202      815
##
##  Location effects: cbind(retcount, notret) ~ dist17 + lifeex17 + recr_offer17 + sfratio_c
##
##              post.mean  l-95% CI  u-95% CI eff.samp  pMCMC
## (Intercept)   1.254295 -0.050227  2.210331   1000.0  0.046 *
## dist17       -0.005937 -0.068010  0.058276    949.5  0.856
## lifeex17     -0.455338 -0.522917 -0.393742   1000.0 <0.001 ***
## recr_offer17 -0.192526 -0.281971 -0.119240   1000.0 <0.001 ***
## sfratio_c    -0.016430 -0.020672 -0.011791   1000.0 <0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the model:

[3] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Analysis, 1(3), 515–534. doi: 10.1214/06-ba117a

[4] full model outputs can be found in Appendix B

- A baseline institution that has no distance learning options, offers no credit for life experiences, offers no recreational or avocational opportunities, and has a typical student-faculty ratio of about 15 is expected to have a log odds retention rate ratio [5] of 1.254, meaning that the estimated retention rate of such an institution would be $logit^{-1}(\ 1.254\ ) = 0.778$.
- An institution that offers distance learning options is expected to have a log odds retention rate ratio that is 0.006 less than that of an institution that does not offer distance learning options, given that credit for life experiences, recreational offerings, and student faculty ratio are held constant. While the corresponding change in retention rate itself can vary depending on the original retention rate, an estimated upper bound on the corresponding change in retention rate is -0.006 / 4 = -0.002.
- An institution that offers credit for life experiences is expected to have a log odds retention rate ratio that is 0.455 less than that of an institution that does not offer credit for life experiences, given that distance learning offerings, recreational offerings, and student faculty ratio are held constant. While the corresponding change in retention rate itself can vary depending on the original retention rate, an estimated upper bound on the corresponding change in retention rate is -0.455 / 4 = -0.114.
- An institution that offers recreational or avocational opportunities is expected to have a log odds retention rate ratio that is 0.193 less than that of an institution that does not offer recreational or avocational opportunities, given that distance learning offerings, credit for life experiences, and student faculty ratio are held constant. While the corresponding change in retention rate itself can vary depending on the original retention rate , an estimated upper bound on the corresponding change in retention rate is -0.193 / 4 = -0.048.
- A unit increase in student-faculty ratio is generally associated with a decrease of 0.016 in the log odds ratio of retention rate, given that distance learning offerings, credit for life experiences, and recreational offerings are held constant. While the corresponding change in retention rate itself can vary depending on the original retention rate, an estimated upper bound on the corresponding change in retention rate is -0.016 / 4 = -0.004.

[5] The log odds retention rate ratio is the natural logarithm of the ratio between retention rate and its complement.