WMU Second-year Retention Model

William E Stutz

October 30, 2020

Contents

Introduction

During the 2017-18 academic year, the Office of Institutional Research at Western Michigan University initiated an ongoing project to model second-year retention of first-time, full-time students (FTIACs). Second-year retention is an important measure of student success, yet, as many as 25% of WMU FTIACs in the last decade have not returned for their second full year of school[1]. Many of these students transfer out; many more simply jettison higher education altogether, either temporarily or permanently. Rates of second-year retention at WMU are not uniform across the student body: second-year retention varies substantially by ethnicity, Pell-grant eligibility, residency, and campus housing usage, to name just a few. Improving second-year retention rates for all students on campus and reducing the gaps between student groups has been and remains a primary goal of the University.

The primary purpose of this project is to better understand the degree to which pre-entry factors (e.g. high school GPA or Pell-eligibility) are associated with second-year retention. Second-year retention can be viewed as an outcome resulting from attributes that students have or bring with them when they arrive at the beginning of the semester ('pre-entry') and factors that arise or occur once they begin their academic careers ('post-entry'). Many, but certainly not all, of these factors can be measured by the university in ways that can aid our understanding of who is likely to return at WMU, and why.

The decision to focus exclusively on pre-entry variables for this project was made for two primary reasons. First, pre-entry variables are both readily available for every new student and highly reliable in terms of accuracy.[2] Second, post-entry factors are themselves influenced

[1] https://wmich.edu/institutionalresearch/reporting/reports/retention/retention/20190306_second_year_retention_report.html

[2] The Office of Institutional Research audits most these data prior to census each term to ensure the data are present, accurate, and error-free

by and related to pre-entry variables in ways that make modeling them a more complex affair than we wanted to attempt here. To clarify, consider that factors like Fall semester GPA or spring-semester retention both *predict* second-year retention, but is also strongly *predicted by* many of the same pre-entry variables that also predict second-year retention. This is not to say that post-entry variables cannot or should not be investigated, only that we are not (yet) addressing them directly here.

Below we: (1) present an outline our methodology in choosing variables and constructing a predictive model for second-year retention (2) run through the results of fitting the model to the data (3) assess the predictive performance of the model and (4) discuss some general conclusions from the project and suggest some points of departure for future work. Lastly, we intentionally referred to this project as ongoing for a reason. While we believe the model and results presented here capture a lot of what's important for understanding retention as a function of pre-entry factors, it is not a *final* analysis in the traditional sense of the word. Remember, "all models are wrong but some are useful"[3]

Methodology

There are many steps on the road from data to insight. In this section we (1) provide some definitions for key terms, (2) list and describe the variables used in the model, (3) discuss our process and reasoning for missing data and data transformations, (4) outline the model structure and method for fitting the model, and (5) give an overview of how we will present the results of model-fitting and the models predictive ability. All of the data processing and modeling fitting was performed using the R statistical programming language (R Core Team 2018).

Data definitions

FTIACs: This project focuses exclusive on predicting and understanding retention for our *first-time, full-time, degree-seeking students*, otherwise known and referred to as FTIACs. FTIACs are defined in terms of Fall cohorts and FTIAC status is assigned to qualifying students at Fall census each year. Specifically all bachelors-degree-seeking, undergraduate students enrolled for at least 12 credit hours at Fall census and who were either new students in the Fall term *or* the prior Summer II term are considered FTIACS. Transfer and returning students are never counted as FTIACs.

Once a student is determined to be a FTIAC in a particular cohort, they are *always* a FTIAC in that cohort even if they subsequently withdraw without ever taking a course!. A student can also be a FTIAC in more than one cohort if they withdraw from all their courses and re-enroll during a later Fall semester as a new student (these students are not counted as returning students because they did not earn any credits during their first attempt).

[3] An aphorism generally attributed to statistician George Box who outlined a the sentiment in his 1976 paper 'Science and Statistics.'

Returned: If a FTIAC is registered for classes at Fall census of their second year, they are considered as having 'returned'. Unlike for determining FTIAC status, a student does not need to be degree-seeking or full-time when they return, they just need to be registered (for credit) in at least one class. For the purposes of this report, we use 'returned' or 'retained' to refer to a student's actual or predicted status in their second year (either yes or no). For example, we would say "150 out of 200 students in the 2017 cohort returned in 2018" or "This student did not return in 2016".

Probability of Returning: We use the phrase 'probability of returning' or 'probability of retaining' interchangeably to refer to the probability (expressed as a percentage) that a given student will or would have returned for their second year. For example, we would say "Even though we predicted that this student had a 90% probability of returning for their second-year, they didn't come back."

Retention: We use 'retention' to refer to the percentage of students of a given group that returned for their second-year. For example, we might say that "Retention for African American students was 75% in 2017".

Percentage Points: We refer to all *changes* in percentages as changes in *percentage points (pp)*. Percentage point changes are additive (not fractional) changes in a given percentage. For example a change from 50% to 55% is expressed as a 5pp increase and not a 10% increase. We use percentage point changes to describe changes in the probability of retention (e.g. this student's probability of returning went up by 7pp, from 70% to 77%) *and* changes in retention (e.g. African American retention went down by 2pp from 80% to 78%).

Data sources

Most of the data used for this project are derived from extracts taken from the WMU Banner Operational Data System (Banner ODS) that are frozen at census every term. We pulled data for FTIAC cohorts from Fall 2012 through Fall 2017. Not all data used in the model are frozen at census and are subject to change or updating throughout the first year. These include data derived from FAFSA information and AP course credit data. Non-censused data are frozen when the model is fit so that results can be replicated to the same data set. The specific variables that are not frozen at census are noted below.

Variables included in the model

There are a wide range of potential pre-entry variables available that might be used to predict student retention. Data can come from student application materials and transcripts, federal application for financial aid (FAFSA), major and course registration data, and potentially other sources as well.

Given the large pool of potential variables, we applied a 'Continuous Model Expansion' approach to building our model. We started with a relatively simple model that contained only an intercept term and a HSGPA term (the latter of which is almost universally a strong predictor of second-year retention). We then applied a flexible, iterative process to add

variables to the model that either (1) improved the convergence, fit, and/or predictive power of the model or (2) that would otherwise aid our understanding of what does and does not affect retention at Western.

In the current version of our model, we can divide the variables we used into four categories: (1) individual-level continuous, (2) individual-level binary, (3) group, and (4) cohort-level variables. The distinction between these groupings is somewhat arbitrary but aligns with how the variables are grouped together and estimated by the model. All variables listed below are frozen at census unless otherwise indicated.

Individual-level, continuous variables

These are variables measured at the individual student level and which vary (more or less) continuously (i.e. can take a wide range of values):

High school grade point average (`HSGPA`): HSGPA data are taken from high school transcripts submitted by the student prior to admission and standardized to a single 5-point scale by the admissions office. In principal students are not admitted with HSGPA below 2.50, but this is only a soft lower bound and some HSGPA data fall below this value. HSGPA data are not available for all students (see 'Missing Data' below for how these students are handled).

Number of AP or IB courses (`AP`): AP/IB courses are reported by students via test score submission and transcripts and only courses for which credit was granted by WMU are counted and used here. We treat AP and IB courses interchangeably and refer to them both throughout as AP courses. The number of AP/IB courses are *not* frozen at census.

Expected family contribution (`EFC`): EFC is estimated and reported on a student's FAFSA in whole dollar amounts. Only students who submit a FAFSA will have an associated EFC value and many students will have an EFC of zero. (See 'Model Fitting' below for how these students are dealt with). EFC is *not* frozen at census.

Individual-level, binary variables

These variables apply to individual students but can take only two values:

Gender (`GENDER`): students report their gender as either Male (M) or Female (F).

On-Campus Housing (`CAMPUS`): this is an variable indicating whether the student was living 'On' of 'Off' campus.

Group variables

Group variables are variables that can take one of discrete set of values (referred to as levels) for each student. Typically these levels are not ordered. In practice, however, it sometimes makes sense to treat a discrete, ordered variables as though it was not ordered (see student credit hours below for an example).

County (`COUNTY`): the county of origin as of Fall census. Because these are pulled from the permanent address data on file, some international students occasionally have domestic counties assigned as their county rather than the international county code. We elected to assign all international students to the single international county code. This ensures that only domestic students have a domestic county code.

Residency (`RES`): each student is assigned one of three residency levels: 'domestic resident', 'domestic non-resident', and 'international'. All non-international students arriving from a Michigan county of origin are considered 'domestic residents'. All other non-international students are considered 'domestic non-residents'. International students are considered 'international' no matter what county they are arriving from. Note that this categorization largely mirrors–but does not completely match– tuition based residency categorization. Using the geographic definition of residency ensures that county codes are completely nested within and not crossed between residency groupings.

Primary Ethnicity (`ETHNICTY`): each student self-reports one of seven different primary ethnicities (or is assigned an 'unknown' ethnicity). International students are do not report a primary ethnicity.

Pell-Eligibility (`PELL`): Pell-eligibility is determined using the expected family contribution (EFC) the student's FAFSA and can take one of three values: Pell-eligible, Pell-ineligible, and Unknown (students for whom no EFC data are available). Full-time students must have an EFC at or below a given cutoff dollar amount in order to be considered Pell-eligible. The cutoff varies from year-to-year and is set by the Federal government (as does the amount the student receives). Cutoff values have been between $4,995 and $5,328 since 2012. Note that students must meet additional, non-EFC based criteria to be be considered officially eligible for Pell grants. Here we are only looking at EFC. Pell-eligibility is *not* frozen at census.

First-generation Status (`FIRST_GEN`): first-generation status is determined using the data from the student's FAFSA and takes one of three values: 'First-generation', 'Not First-generation', and 'Unknown'. Students are asked on the FAFSA to provide the highest level of education obtained by each parent, if known. We consider students first generation if neither parent is indicated as having finished college *and* at least one parent has some education information provided (e.g. father has high school degree and mother's education level is not provided). Students that did not submit a FAFSA, or who did not indicate any education-level for their parents are considered 'Unknown'. First-generation status is *not* frozen at census.

Credit Hours (`SCH`): the number of credit hours enrolled as of census (minimum 12SCH). Although credit hours are ordered, numeric data, we treat each SCH value as distinct categories. There are two primary reasons we do this. First, our preliminary analysis suggested that the relationship between SCH and retention was likely non-linear, tending to increase only up to a point, and then decrease. By treating SCH levels independently, we can account for this non-linearity[4]. Second, we were more interested in contrasting expected retention rates at different levels of SCH after the model is fit than in estimating an effect size of SCH

[4]That is not to say that adding an additional linear or polynomial term wouldn't also make sense and reduce variance in the estimates somewhat. We just have not yet done so

per se. Thus, keeping credit hours as a group variable kept the model set-up, analysis, and interpretation more straightforward.

Department (`DEPT`): the department associated with a student's major at census.

College (`COLL`): the college associated with a student's major and department at census. Note that this does not necessarily correspond to the ultimate degree-granting college.

Cohort-level Variables

These two variables apply specifically to cohorts:

Year (`YEAR`): this is an integer variable defining the year of the cohort (treated as a continuous variable in the model).

Cohort (`COHORT`): this an unordered variable that indicates the student's FTIAC cohort. The year value of the cohort (2012, 2013, 2014) is treated separately from the cohort effect as described above. Another way to think of the 'cohort' variable is as cohort-level residual variation once the year effect is taken into account.

Variables *not* included in the model

It might seem like there are a few variables where it would make sense to include them in the model, but they are not listed above. Generally speaking if a variable is not in the model either: (1) it's too correlated with another predictor already in the model (2) reliable data for the variable is not available.

As an example of reason (1), we have good data on student ACT scores from student application materials, but we found *in practice* that ACT scores were too correlated with HSGPA data to be useful. Specifically, the estimates for both become much more uncertain when combined than when modeled separately. Honors college membership was excluded because it was redundant with high school GPA scores: virtually all students with HSGPA > 4 were in the college, and all students in the college had HSGPA > 4.

For reason (2) there may be data that could be helpful, but that are not readily available for many or all students. For example, reliable data on a student's intention to transfer when they arrive at WMU would be quite helpful, but it is not widely available.

Missing Data

Of the variables included in the model, only HSGPA had missing data. It's unclear whether for any given student the HSGPA value is truly 'missing' or whether it simply does not exist. Moreover, international students are much more likely to be missing HSGPA data domestic students, meaning that missingness is non-random in the data. Given that we expected HSGPA to be an important predictor in the model, we considered a few options for dealing with the missing HSGPA data:

1. remove all students without HSGPA data
2. assign a HSGPA value equivalent to the average value at WMU
3. assign a HSGPA value based on the average value of some larger group of students (e.g. the entire cohort, the entire residency grouping, etc.)
4. assign a HSGPA through data imputation

Given the number of students (especially international) students without HSGPA data, we decided against option (1). Options (2) and (3) are reasonable and conservative defaults given the nature of our model (a linear model with a global intercept and centered predictors) where the assigned HSGPA values would be at or close to zero for students missing HSGPA and thus have little impact on the HSGPA estimate other than to regress it to the mean.

We ultimately decided on option (4), using the multiple imputation method described in Buuren and Groothuis-Oudshoorn (2011) and implemented in the *mice* R package (Buuren et al. 2019). All the variables ultimately included in the model (plus ACT scores) were used for imputation. Only a single imputation was used performed.

To check the robustness of our imputation we compared the HSGPA parameter estimates of our fitted model to:

1. models fit with three other imputations of the same data
2. a model fit where students with missing data were assigned the overall average HSGPA
3. a model fit with students missing HSPGA removed.

In all cases the model effect estimates were largely unchanged and no different conclusions could would be drawn from a fitted model using a different procedure for missing data. Therefore we believe our general results to be robust to our treatment of missing HSGPA data[5].

Data transformations

We performed a number of data transformations prior to fitting the model:

1. **Expected Family Contribution**: EFC data were strongly positively skewed, with most students bunched towards the lower end of the spectrum and fewer students with extremely large EFC values. We $\log_2$ transformed these data to create a new input variable (`logEFC`) to use in the model. Because we only estimate an EFC effect for Pell-ineligible students, EFC values below the Pell-eligibility cutoffs (including zero values and `NULL` EFC values) are ignored for the purposes of this transformation

[5]The optimal fully Bayesian treatment would be to treat each unobserved HSGPA value as an unknown, apply a a weakly informative prior distribution for missing HSGPA values, and incorporate the uncertainty in the imputation into the joint posterior. In practice this would be lot of coding and computational effort for likely very little difference in the outcome

2. **Binary Variables**: Both binary variables (`GENDER` and `CAMPUS`) were shifted to have a mean of 0 and differ by 1 between the two levels (Gelman et al. 2008). To do this, one value was assigned its proportion in the data set (e.g. 0.52) while the other value was assigned it's negative proportion (e.g. -0.48). This also ensures that the SD will be about 0.5.

3. **Continuous Variables**: All continuous input variables (`HSGPA`, `AP`, `logEFC`, and `YEAR`) were mean-centered and scaled to a standard deviation of 0.5 to be on the same scale as the binary variables. The transformation also ensures that an effect size of 1 on the transformed scale would correspond to the change going from 1 standard deviation below to 1 SD above the mean on the raw data scale.

4. **Rare Counties**: A large number of counties (163) appear only once in the data set (i.e. only one student came from that county). The effects for these counties would be conflated with the unobserved (and unobservable) latent student effects rendering them uninterpretable. Instead, all rare counties were reassigned to and recoded to a new 'Rare Counties' level.[6].

Model Structure

We used a Bayesian, hierarchical regression model to predict second-year retention for FTI-ACs. There are many reasons to use such a model over other approaches. To highlight just two of the most important: (1) the hierarchical structure and application of (weakly informative) priors allows parameters to appropriately pooled and regularized and (2) using Bayesian probability best allows us to estimate and propagate uncertainty in our estimates and predictions.

The specific model we applied was a binomial generalized linear model with a logistic link function (i.e. a logistic regression model). The model is fully Bayesian in the sense that all parameters in the model have fully defined priors (some with hyper-priors) and the posterior distributions are estimated given a the specified binomial likelihood function and the data available.

The basic version model can be written as follows:

$$y \sim Bernoulli(p)$$

$$logit(p) = X\beta$$

[6]Currently, the model does not distinguish between rare Michigan counties and rare non-Michigan counties

Where $logit(p) = log(p/(1-p))$, X is the full design matrix and β is a vector of all parameters to be estimated. In our case it is possible to breakdown X and β further into sub-matrices and sub-vectors corresponding to each of the predictor variables:

$$y \sim Bernoulli(p)$$

$$logit(p) = \alpha + X^{HSGPA}\beta^{HSGPA} + X^{AP}\beta^{AP} + ... + X^{COHORT}\beta^{COHORT}$$

Here α refers to the intercept, X^{HSGPA} and β^{HSGPA} refer to the design sub-matrix and parameter sub-vector corresponding to `HSGPA`, X^{AP} and β^{AP} refer to the sub-matrix and sub-vector for `AP`, and so on through each variable in the model. In practice, we can condense this notation somewhat by considering how our priors are assigned across parameters:

$$y \sim Bernoulli(p)$$

$$logit(p) = \alpha + X^0\beta + X^1\gamma^1 + X^2\gamma^2 + ... + X^k\gamma^k$$

$$\alpha \sim N(\mu_\alpha, \sigma_\alpha)$$

$$\beta_j \sim t_{\nu_\beta}(0, \sigma_{\beta_j})$$

$$\sigma_{\beta_j} \sim half\ N(0, \tau)$$

$$\gamma^k \sim t_{\nu_\gamma}(0, \sigma_{\gamma_k})$$

$$\sigma_{\gamma_k} \sim half\ t_{\nu_\sigma}(0, \eta)$$

$$\tau \sim half\ N(0, \sigma_\tau)$$

$$\eta \sim half\ N(0, \sigma_\kappa)$$

Here the parameters associated with the J total predictor variables are combined into the j-length vector β. Each β_j parameter is drawn from a unique, centered t-distribution with a scale σ_{β_j}. The σ_{β_j} scales are themselves drawn from a single half-normal distribution with mean 0 and scale τ. In practice this means that each individual-level parameter is assumed to arise from (and be constrained by) semi-independent t-distributions, the scales of which are drawn from (and constrained by) a single half-normal distribution.

Similarly, for each of K group variables (`PELL`, `SCH`, `RES`, etc...) we have a separate vector of parameters γ^k associated with the levels of group K. For each vector γ^k, the parameters are drawn from t-distributions with mean = 0 and scales σ_{γ_k}. The σ_{γ_k} scales are themselves drawn from a single centered, half-t distribution with scale = η. This means, for example, that ethnicity effects are drawn from a single common t-distribution, residency effects are drawn from a single common t-distribution, and so on, and the scales of those t-distributions are drawn from (and constrained by) separate (half) t-distributions, all of which are drawn from (and constrained by) a half-normal distribution with mean = 0 and scale η.

Lastly, the τ and η hyper-prior distributions governing the individual-level and group-level scales are drawn from normal with mean = 0 and scales σ_τ and σ_η

These t-distribution priors on the individual and group level effects are *weakly informative*, giving preference to values of less than 5 on the logit scale, which corresponds to a change from 0.01 to 0.5 or 0.5 to 0.99 on the probability scale (Gelman et al. 2008). In our model this implies that the difference should be be less than 50 percentage points moving from one standard deviation below the mean to one standard deviation above the mean. The half-t and half-normal distributions on the scale parameters are also only weakly informative, constraining the posterior scales away from unreasonably high values when the number of levels is small or when there is little signal arising from the data (Gelman 2006).

Fitting the Model

Testing and training data

Prior to fitting the model, the full data set (2012 - 2017 FTIAC cohorts) was subdivided into a 'training' and a 'test' dataset. Specifically, 5/6 of the students were randomly chosen to be part of the training set, which was used for model fitting. The remaining 1/6 comprised the testing dataset, which we used to calculate out-of-sample accuracy of the model (rather than relying on within-sample approximations). Dividing the data allowed us to compare different iterations of the model and to judge the ability of the model to make predictions given new input data.

Fully nested groups

In some cases, levels in two different groups were reciprocally and uniquely nested within one another. For example, the College of Aviation has only one department (Aviation Sciences), therefore all the students in Aviation Sciences department are also in the College of Aviation and vice versa. This means the parameter for the Aviation Sciences department and the parameter for College of Aviation use the exact same data points and will attempt to estimate the same effect. This can cause the model fitting procedure to misestimate the effects of even fail altogether.

Our general approach to this problem was to use NULL values for one of the two associated levels and estimate the effect only for the other. We performed this procedure for:

- **International Students**: international students have their own unique ethnicity value ('International'), their own residency value ('International') and their own county code, all of which, if included in the model, would be estimate the effect of 'being an international student'. We elected to keep only the international coding for residency; international students had NULL values for both their ethnicity and county.
- **University Curriculum**: these students comprise a single department (INTO) and a single college ('Other'). We elected to keep the college effect and assign University Curriculum students a NULL value for department.
- **Aviation**: As mentioned, Aviation students comprise a single department within a single college. We elected to keep the college effect and assign University Curriculum students a NULL value for department.

Additional parameters

The model contained 357 non-scale parameters: the intercept parameter (α), one HSGPA parameters for each residency level (i.e. a `RES` x `HSGPA` interaction) and five continuous/binary parameters (`AP`, `EFC`, `GENDER`, `CAMPUS`, `YEAR`) all in β, and the 348 parameters contained within the 8 group variable-associated γ vectors. The inclusion of the interaction was spurred by the observation from the raw data that HSGPA was strongly and positively associated with retention for domestic students but slightly negatively associated with retention for International students.

Hyperparameters and constants

As described above, our model has two hyperparameters–τ and η–that dictate the scale of the distributions from which the $\sigma_b eta$ and $\sigma_g amma$ parameters are drawn. We ultimately set both of these parameters to be drawn from half-normal distributions with mean = 0 and SD = 1 (i.e. standard half-normal distributions):

$$\tau \sim halfN(0, 1)$$

$$\eta \sim halfN(0, 1)$$

We also set a very strong, bounded prior distribution for the α parameter (i.e. the global intercept). This was done to improve model-stability, as allowing the alpha parameter to vary widely prevented the MCMC algorithms from converging on a stable posterior distribution. The prior for α was a normal distribution with mean equal to the overall proportion (logit-transformed) of returning students in the data, and a σ_α of 0.01.

$$\alpha \sim N(\mu_\alpha, \sigma_\alpha^2)$$

$$\mu_\alpha = logit(\frac{\sum_1^n y_i}{N})$$

$$\sigma_\alpha = 0.01$$

Prior predictive check

We also performed a prior-predictive check by asking whether the prior predictive distribution included grossly unrealistic values for the predicted individual probabilities p_i. This is, admittedly, a somewhat subjective exercise because (1) the individual probabilities are never directly observed (2) opinion can vary widely about what, for example, the upper-bound or lower-bound probability of returning should be for an individual student. Is it reasonable to say that a student has a 99.9% (999/1000) probability of returning for a second-year? A 99.999% (99999/100000)?

To make sure that our prior-distribution was not too liberal, we calibrated the hyper-prior parameters τ and η above to ensure that the 95% joint prior-probability distribution for p_i was upper-bounded at about $p = 0.9999$. This prior assumes that students puts the upper bound probability of returning at 99.99%. Put another way, if we simulated a school year for a student with $p = 0.9999$ 10,000 times, we would expect them to *not return* only once.

Programming language and MCMC diagnostics

The model was fit using Stan probabilistic modeling platform, interfaced through the *rstan* R package (Guo et al. 2018). The model itself was written using the Stan programming language (Stan Development Team 2017). Stan uses Hamiltonian Monte Carlo (HMC) sampling, a form of Markov chain Monte Carlo sampling.

We ran four Markov chains for 2,000 iterations each. After discarding the first 1,000 iterations of each chain we were left with 4,000 iterations from which to sample the full posterior. During model development, the total number of iterations was increased until the effective sample size (ESS) was > 1,000 for each of the continuous/binary parameters in (β) and each of the group scale parameters (σ_{gamma_k}). Convergence among chains was assumed to have been sufficient if $\hat{R}$ values were < 1.01.

Assessing the model

We performed three different analyses to assess the model fit: (1) posterior predictive checks (2) within-sample goodness-of-fit and (3) out-of-sample predictive accuracy.

Posterior predictive checks

The purpose of performing posterior predictive checks (PPC) is to ask whether the fitted model captures relevant features of the data used to fit the model. A lack of congruency between the posterior prediction and the raw data indicates the model may be under-specified in some important way and suggests avenues for improving or expanding the model. Note that Posterior predictive checks can (and were) used to evaluate features not present in the fitted data (e.g. estimates for groups not included in the model fitting).

Given the nature of binary data, we are somewhat limited in the summary statistics we can check to ensure proper model fit. We used PPC to ensure that the model correctly predicted retention rates for each sub-group or sub-cohort in the model (e.g. every ethnicity, every college, etc..). We compared the posterior predictive distribution to the observed value to ensure that the observed value fell within the 95% probability distribution for $\sim 95\%$ of the parameters we checked. For the model presented here we found no additional discrepancies between the posterior predictive distributions and the observed data.

Goodness-of-fit

For each sample from the posterior distribution we calculated a psuedo-R^2 value using the following formula:

$$R_L^2 = \frac{(D_{null} - D_{fitted})}{D_{null}}$$

where D is the deviance of a given model and is calculated as:

$$D = -2ln\ \frac{likelihood\ of\ the\ given\ model}{likelihood\ of\ the\ saturated\ model}$$

In this case the null model is the 'intercept-only' model, which assigns each student a probability of returning equal to the overall retention rate. R_L^2 varies between 0 and 1, approaching 1 as the fitted model approaches full saturation.

In order to measure both within-sample and out-of-sample goodness of fit, we calculated separate R_L^2 for both the training data set and the testing data set.

Predictive accuracy

We used the testing data to assess three metrics of predictive accuracy defined as follows:

- **Accuracy**: the proportion of *all* students predicted correctly by the model
- **Sensitivity**: the proportion of students who returned that were predicted to return
- **Specificity**: the proportion of students who did not return who were predicted not to return

To calculate these metrics, for each sample we:

1. used the fitted probability p_i for each individual i in the test data set to simulate a new returned value y_i^{rep} for that student
2. compare each y_i^{rep} to y_i:
   - for all y_i, if $y_i^{rep} = y_i$, then $acc = 1$, else $acc = 0$
   - for all $y_i = 1$, if $y_i^{rep} = 1$ then $sens = 1$, else $sens = 0$
   - for all $y_i = 0$, if $y_i^{rep} = 0$ then $spec = 1$, else $spec = 0$
3. calculate accuracy as $\frac{\sum_1^n acc}{N}$, sensitivity as $\frac{\sum_1^{n_{y_i=1}} sens}{N_{y_i=1}}$, and specificity as $\frac{\sum_1^{n_{y_i=0}} spec}{N_{y_i=0}}$

We also calculated the same three metrics for the intercept-only model (assuming each student's probability of returning was equal to the overall retention of the test data) and for a random model (assuming each students probability of returning was 0.5).

Reporting results

Posterior summaries

Unless otherwise specified, parameter estimates are reported or graphed as posterior summaries. Specifically, for non-scale parameters we report the posterior median and 90% credible intervals of the marginal posterior distributions (unless otherwise noted). For continuous effects, effect sizes are reported as the expected change going from one standard deviation below the mean to one standard deviation above the mean.

Effect size scale

We elected to report effect sizes on the probability scale rather than the log-odds scale or odds ratio scale. This means that effect sizes are reported as *changes in percentage points (pp)* and not as a percent change or change in the log-odds. Practically, this means we report a change in probability from 50% to 55% (0.50 to 0.55) as a 5 percentage point increase and not a 10% increase or change of 0.2 in the log-odds.

The downside to reporting effects as changes in percentage points is that the reported effects are non-linear. For example, for a continuous effect, the effect size in percentage points associated with moving from one standard deviation below to one standard deviation above the mean will differ (slightly) than the effect size going from the mean to two standard deviations above the mean. This is not the case on the logit scale, which changes linearly with the input variable.

However, we believe the benefits to clarity (50% to 55% and 55% to 60% are both +5 percentage points) outweigh the non-linearity, primarily because the non-linearity of the effects is very slight over the range of the data and does not qualitatively alter the extrapolation and interpretation of effect sizes beyond +/- one standard deviation around the mean.

Marginal effects

The effect sizes reported below are all *marginal* effects. They show the expected change in the probability of returning for the *average* student. In this case the average student is defined as a student with an average HSGPA, EFC, number of AP courses, entering in the average year, and *with no particular level of any of the group effects*.

In a *linear* model without interactions, the marginal effect is mathematically equivalent to common interpretation of "all else equal". In other words the effect of variable A applies equally to all individuals with different values of the other variables in the model. In a non-linear model (like the logistic model here), the marginal effect is not the same as "all else equal"– the specific estimate effect technically applies *only* to individuals with the average values of the other variables. Mathematically speaking, reporting marginal effects is completely sound–it is the interpretation that can lead to trouble. There are at least two reasons why reporting marginal effects in non-linear models *might* be misleading:

1. The effect of increasing A by 1 unit may be slightly (or dramatically) different depending on the starting value of A (e.g. starting at the mean or starting 2 SD above the mean).

2. In cases where data are bimodal or heavily skewed, the average student might be highly non-representative of the actual underlying data.

One general remedy for countering these issues in non-linear models is to use 'average predictive comparisons' (Gelman and Pardoe 2007). There are various ways of calculating APCs but the general idea is to integrate or summarize the expected effect size when applied to every individual data point (and not just the average). In practice, APCs can be difficult and time-consuming summaries to calculate, especially when the number of parameters of interest is high. Also, it's not entirely clear how best to use APCs to summarize group effects (Gelman and Pardoe 2007).

In building the model, we calculated APCs for all individual and group effects in parallel with marginal effects and found that, in all cases, *the APC estimates were nearly indistinguishable from the marginal effects*. This is likely due to: (1) our input data are centered, scaled, and unimodal, (2) the range of probabilities predicted by the model falls between 0.5 and 0.95 (mostly), meaning the effect sizes remain approximately linear relative to the input data, and (3) there were no large estimated effect sizes (i.e. +30 or more pp) which might have exacerbated any non-linearities. Therefore for convenience we calculate and report only the marginal effects.

Correlations between marginal effects

When comparing marginal effects, especially within-group effects (e.g. Domestic residents and Domestic non-residents) keep in mind that the marginal posterior distributions may be positively correlated in the joint posterior. Consequently, marginal effects whose 95% credible intervals strongly overlap when plotted independently may actually be quite distinct.

One way to see differences between the marginal effects is to plot within-group effects against a common baseline. For example instead of viewing plotting the effects for Domestic residents, Domestic non-residents, and International students against the residency group average, we may use Domestic residents (or any other group[7]) as the *baseline* for comparison. This allows us to visualize the posterior distribution of the *difference* between the baseline effect and the others. This has the upshot of taking the positive correlation between effects into account, and can help to see where there are, in fact, differences between the estimated marginal effects.

Results

Individual-level parameters

We estimated seven individual-level parameters corresponding to seven different predictor variables (`YEAR` is not included): `GENDER`, `HSGPA[domestic residents]`, `HSGPA[domestic non-residents]`, `HSGPA[international]`, `AP`, `EFC` and `CAMPUS`. For ease of interpretation we report the binary effects as one level (i.e. female and living on campus) compared to another 'baseline' level (i.e. male and living off campus),

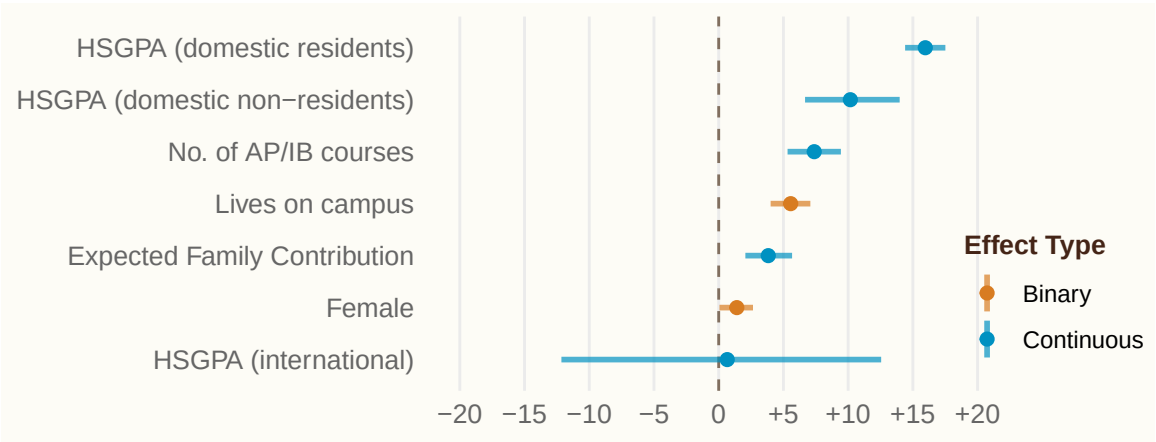

Figure 1: Posterior summaries (median +/- 90%) for all individual-level parameters. Estimates are reported as the effect on the probability (in percentage points) of a FTIAC returning for a second year relative to an average student (dotted line). Continuous effects are associated with moving from 1 SD below to 1SD above the mean.

Figure 1 shows the estimated effect sizes for each of the seven individual-level parameters. Every individual effect that we included in the model is associated with increased retention

[7]the choice is arbitrary but in practice we like to use the group that makes up the largest percentage of students in the data

(maxing out at about 16pp). The lone exception is the HSGPA effect for international students, of which the direction and magnitude are highly uncertain. Note that the trend towards *positive* effects is a result of the variables we included (we chose variables that are likely to increase retention as they go up) and an artefact of choosing to use male students and off-campus living students as the baselines for the gender and living-on-campus predictors.

Note that the posterior distribution for the gender effect suggests does not rule out the possibility that the difference between female and male students is extremely small on average ($<$ 1pp).

High School GPA

We can plot the marginal HSGPA effects for each residency group against the raw retention data to illustrate the differences between residency groups and to highlight the relative magnitude each effect:

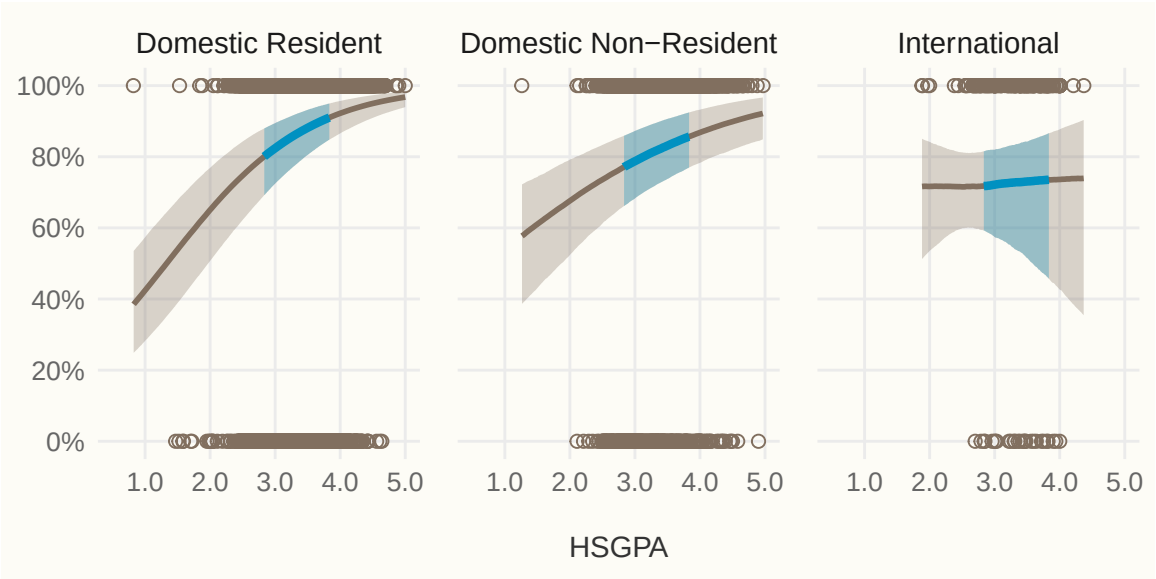

Figure 2: Estimated marginal probability (%) of a FTIAC returning for a second year over the range of observed HSGPA values. The brown ribbon around the trendline illustrates the estimated uncertainty in the effect size plus the uncertainty in residency group estimate. The magnitude of the effect associated with the +/- 1 SD change in HSGPA is highlighted in blue. Raw retention data are plotted as circles at the top and bottom of each panel.

HSGPA is clearly strongly associated with retention for *non-international* students at western. Non-international students with a HSGPA that is one standard deviation higher than average HSGPA (3.83) have anywhere between a 10 and 16 percentage point higher probability of returning for their second year than students on standard deviation below the aveage HSGPA (2.83).

By extending out the marginal effect beyond 1 SD (note the non-linearity as HSGPA decreases or increases), we can how the marginal effect of HSGPA impacts students at the more extreme ends of the HSGPA range. Following the domestic student trendlines to the far right, we can see that students with extremely high HSGPA values very are very likely to return, while students at the low end are only about 60% likely to return (on average).

Lastly note that while the overall effect of HSGPA for domestic students is quite large, the difference in effects between the residents and non-residents is more moderate, with residents students recieving more of boost from increased HSGPA than non-residents (see Figure 1).

Number of AP/IB courses

We can also plot the effect of the number of AP/IB courses against the raw data to highlight the magnitude of the marginal `AP` effect (note we did not estimate variation across residency groups, so only one panel here):

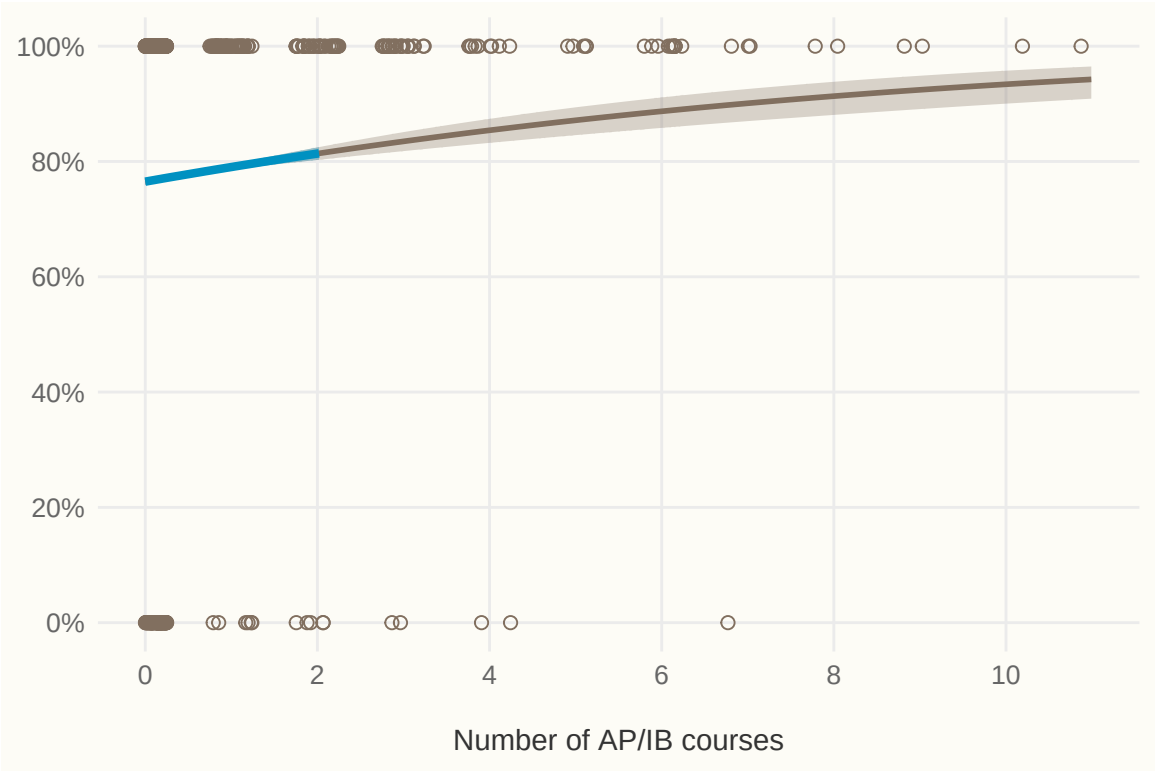

Figure 3: Estimated marginal probability (%) of a FTIAC returning for a second year over the full range of AP courses taken. The brown ribbon around the trendline illustrates the estimated uncertainty in the effect size. The magnitude of the effect associated with the +/- 1 SD change in the number of AP courses is highlighted in blue. Raw retention data are plotted as circles at the top and bottom of each panel.

The estimated `AP` effect shown in Figure 1 was about +8 pp, which represents the change in expected probability for an increase from 0 to 2 courses taken. Unlike HSGPA, the distribution of the number of AP courses taken by FTIACs between 2012 and 2017 was highly right-skewed. The average number of AP/IB courses taken by FTIACs in the data set was about 0.6 (or 1/2 course) and the majority of students arrived having taken zero AP courses. Because the average number of courses was so low, students who take no AP courses have only slightly smaller marginal probabilities of returning compared to the average student. Students at with extremely high `AP` values (10+ courses taken) are still very likely to return, however.

Expected Family Contribution

The effect of expected family contribution (`EFC`) was estimated for Pell-ineligible students only.

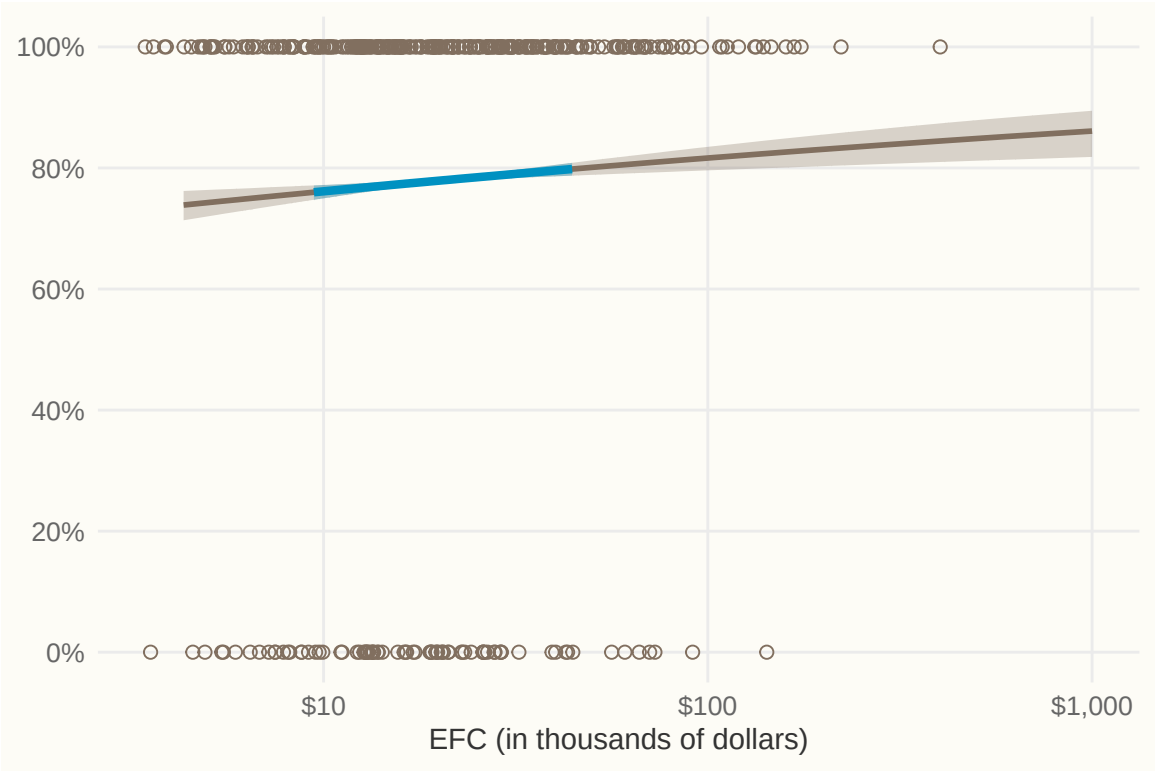

Figure 4: Estimated marginal probability (%) of a Pell-ineligible FTIAC returning for a second year over the full range of expectef family contribution (EFC). Pell-eligible students and students without FAFSA information were not used to calculate the effect and are not shown in the plot. The brown ribbon around the trendline illustrates the estimated uncertainty in the effect size. The magnitude of the effect associated with the +/- 1 SD change in logEFC is highlighted in blue. Raw retention data are plotted as circles at the top and bottom of each panel.

Note that because EFC was transformed to a $\log_2$ scale, the `logEFC` effect is aassociatedwith *multiplicative* changes in EFC. For example, the estimated effect size shown in Figure 1 for EFC was about +3.8 percentage points. This 5 pp increase was for an increase from $ 9,410 to $ 44,453, or an increase by a factor of about 4.7. In other words, for every doubling of EFC we would expect an approximate increase of 2 pp in the expected probability of returning.

Group parameters

Relative importance

Although we can estimate the marginal effects of different levels within each group, it is useful to compare the within-group variances associated with each group predictor. In particular, comparing the within-group variances among the gives us some idea of the relative *importance* of each group in the model. The more variation there is among the effects within each group, the more the inclusion of that group in the model makes a difference for the predicted outcome. In contrast as the within-group variance approaches zero, the 'importance' of that variable in the model also approaches zero.[8]

We can convert the group scale parameters σ_γ into variances using the general formula for the variances of a scaled t-distribution:

$$var(X) = \sigma^2 \frac{\nu}{\nu - 2}$$

[8]This analysis is related to an ANOVA-type analysis of a standard linear regression model. The primary difference is that the variance in the underlying individual-level probabilities p_i is unidentifiable in a logistic model. Thus the ratio of explained variance/total variance cannot be calculated as it would be in a standard linear regression.

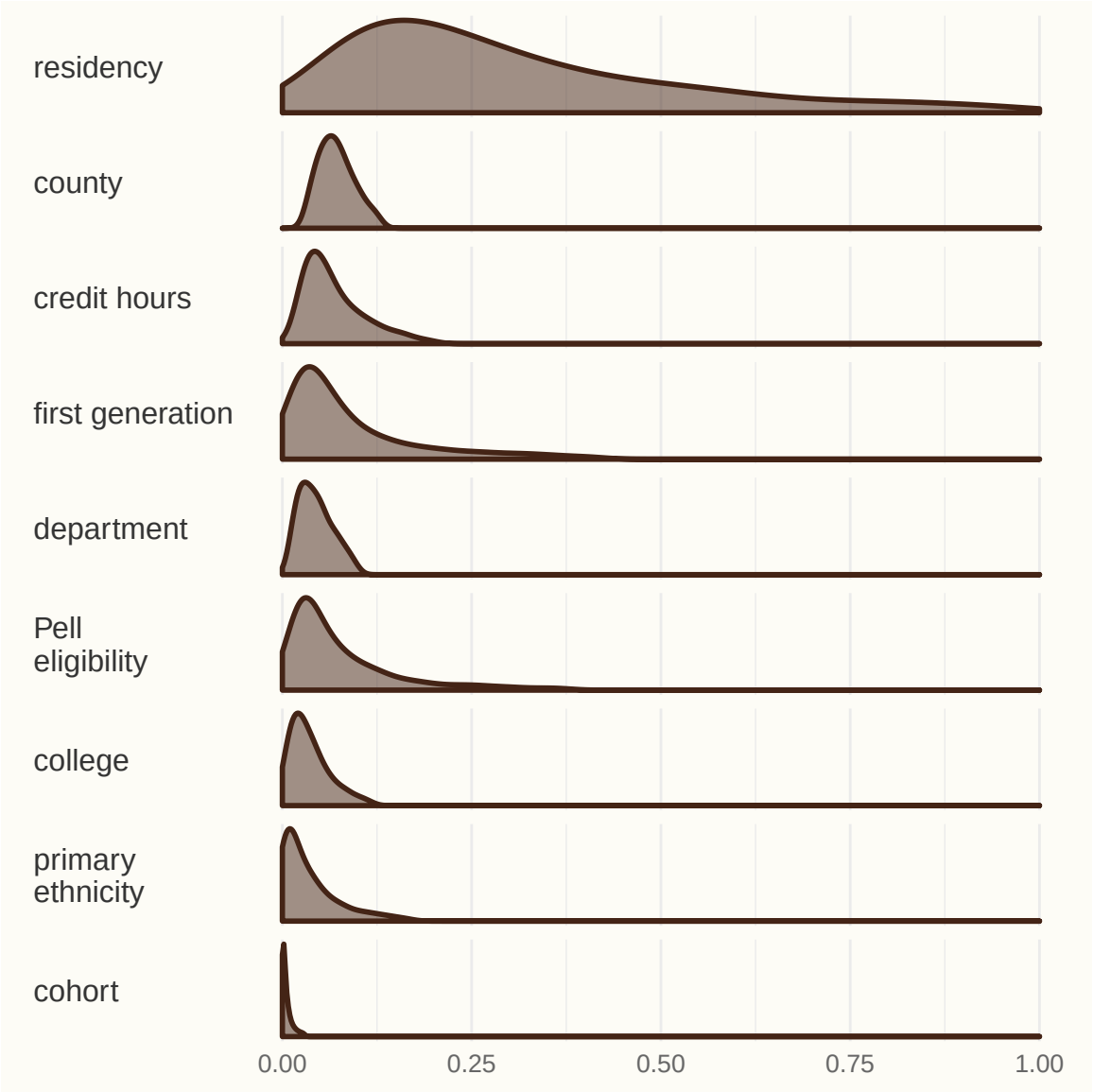

Figure 5: Posterior density plots of the model-estimated, super-population variance for each group effect in the model. Super-population variance is the variance among both observed and unobserved levels for each group, estimated from the data available. Posterior densities are ordered by their modes and right-truncated at 1 to aid comparisons.

Three things stand out from this graphical comparison. First, even when ignoring the long right-hand tail, residency has a stronger effect on the probability of returning than any other group variable in the model. As we will see below, this is largely due the large difference in the marginal effects between international and domestic students.

Second, the variances for county, department, Pell-eligibility, first-generation status, credit hours, and college are all roughly equivalent. These group variables are more or less equally important for predicting the probability of a student returning.

Lastly, unlike for the other group variables, the distributions for primary ethnicity and cohort level variance have a posterior modes that approach zero. Relative to other group variables, there is likely little remaining variance in retention left to be 'explained' by primary ethnicity once the other variables in the model are accounted for (this is *not* to say that ethnicity doesn't matter–it simply plays a more indirect role in retention via variables like Pell eligibility, EFC, and county, and first-generation status).

Residency

We estimated three separate effects for each for each of the three residency levels in our data: Domestic Resident, Domestic Non-Resident, and International.

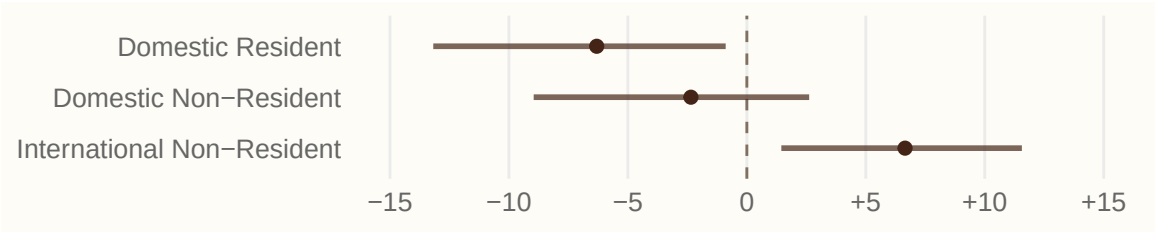

Figure 6: Posterior summaries (median +/- 90%) of the marginal effects for residency. Estimates are reported as the effect on the probability (in percentage points) of a FTIAC returning for a second year relative to an average student (dotted line).

Figure 6 shows that international students are expected to have a much higher probability of returning that domestic students.

We can see this difference more clearly by replotting the effects using the domestic residents as the baseline:

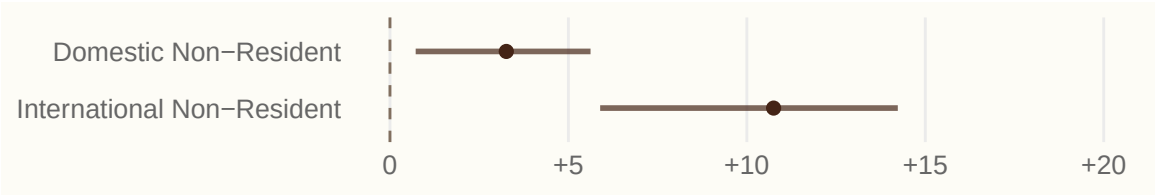

Figure 7: Posterior summaries (median +/- 90%) of the marginal effects for residency. Estimates are reported as the effect on the probability (in percentage points) of a FTIAC returning for a second year relative to 'Domestic Resident' students (dotted line).

In figure 7 the dotted line now represents domestic residents. Both domestic non-residents and international students have higher probabilities of returning in their second year, with

the former between 1 and 6 and the latter between 10 and 15 percentage points more likely to come back. The substantial differences between residency levels shown here account for the high variance in residency effects shown in figure 5 relative to other group variables.

County

Layered underneath variation in residency groups is variation at the county level. We estimated effects for 254 different domestic counties nested within just the two domestic residency groups (recall that international students do not have a separate county effect).

Because counties are nested within residency groups, the county effect estimates must be interpreted as relative respective residency categories and not the university at large. In other words, a large positive county effect means that students from that county have a higher probability of returning than students from counties *at the same residency level*. This is important because given the differences between domestic residency groups noted in figure 7.

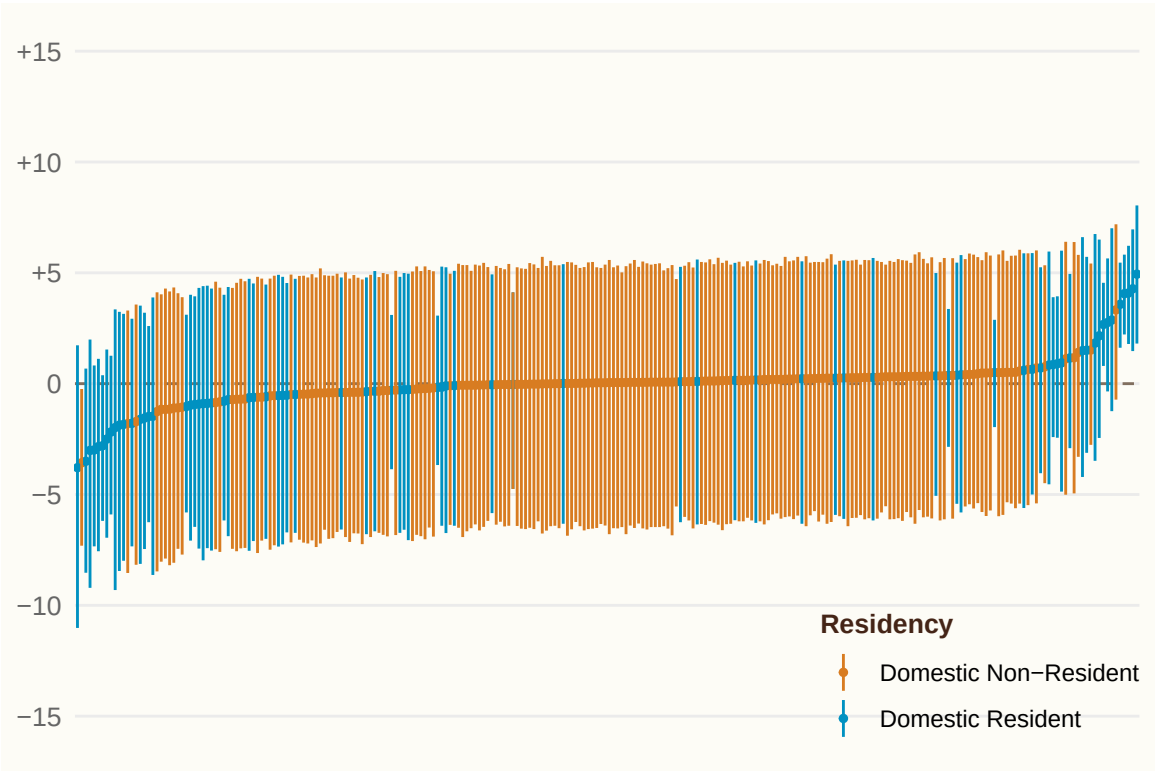

Figure 8: Posterior summaries (median +/- 90%) of the marginal effects for counties. Estimates are reported as the effect on the probability (in percentage points) of a FTIAC returning for a second year relative to an average student at the given residency level (dotted line). Effects are ordered by the posterior median.

The model was unable to identify strong effects for most of the 254 counties included in the data set. This is not surprising, as most counties have very few students, leading to the high uncertainty in the magnitude and direction of their effects. However, there are handful of counties (primarily Michigan counties) at the tails which appear to have positive or negative marginal effects on the probability of returning.

We can zoom in on the counties for which there were at least 50 students:

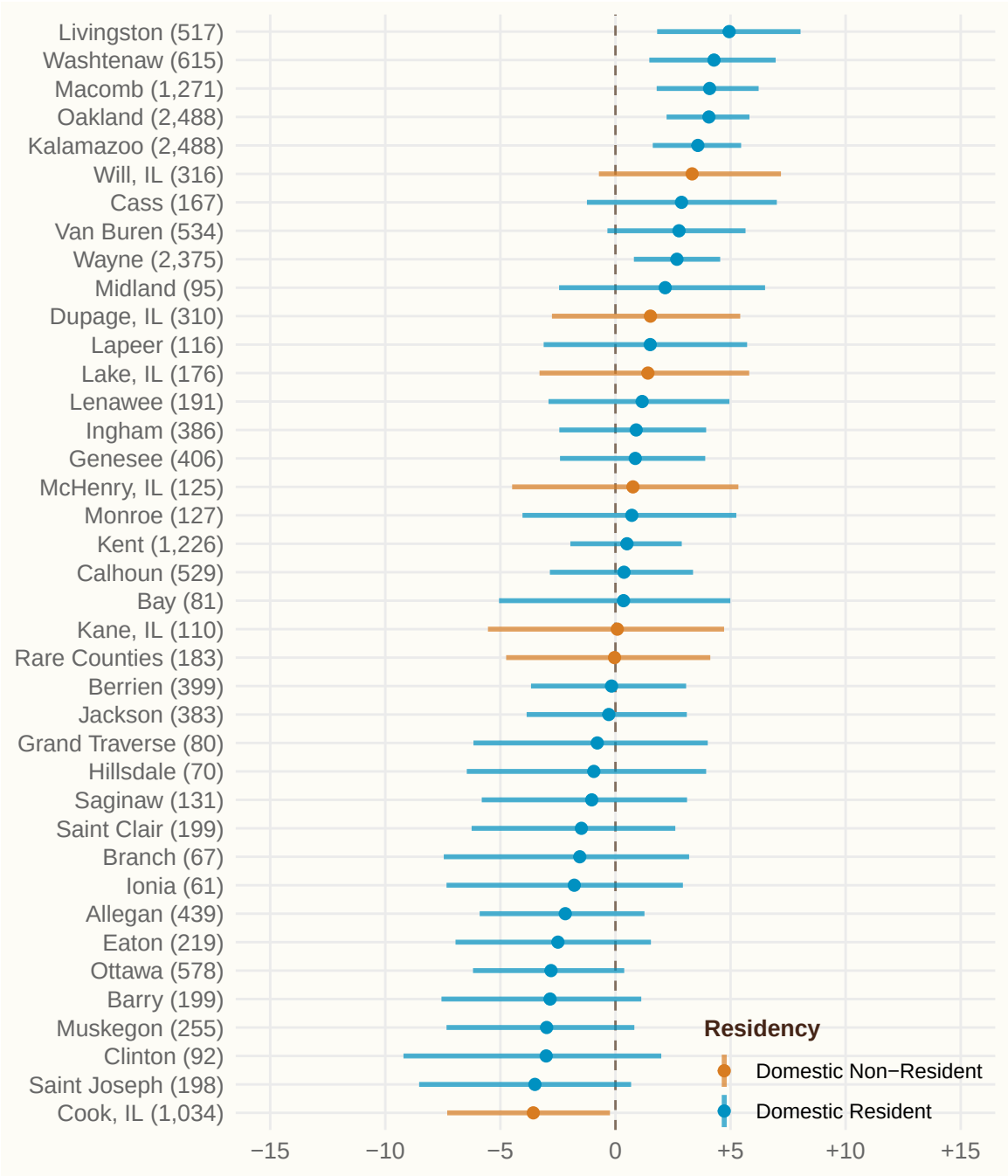

Figure 9: Posterior summaries (median +/- 90%) of the marginal effects for counties with at least 50 students. Estimates are reported as the effect on the probability (in percentage

points) of a FTIAC returning for a second year relative to an average student at the given residency level (dotted line). Effects are ordered by the posterior median.

Interestingly, the four of the five domestic counties with largest number of students (Kalamazoo, Oakland, Wayne, and Macomb) all have moderate positive effects on the probability of returning (as do Livingstone and Washtenaw counties). Indeed, nearly all the clearly positive or negative effects are from Michigan counties. Effects for non-Michigan counties tend to be centered closer to zero with higher uncertainty.

Credit hours

We estimated marginal effects for students registered at every credit hour level between 12 and 23 credits (recorded at Fall census). However, most students registered for a total between 12 and 15 credit hours. Only 7% of students registered for more than 15 credit hours and very few students (less than 0.01%) were registered for more than 17 credit hours.

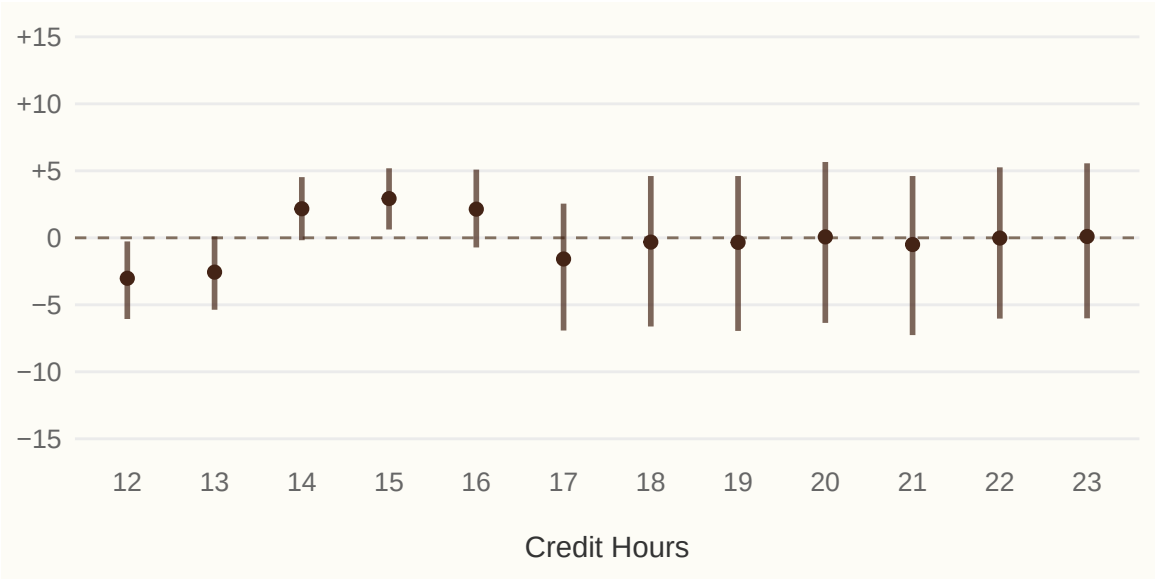

Figure 10: Posterior summaries (median +/- 90%) of the marginal effects for student credit hours. Estimates are reported as the effect on the probability (in percentage points) of a FTIAC returning for a second year relative to an average student (dotted line).

It's clear from figure 10 that the marginal probability of returning is about 4 percentage points higher for students taking 14 or 15 credit hours relative to students taking only 12 or 13. Beyond 16 credit hours there is no evidence that students were more likely to return than those taking 12 credit hours. Given how few students registered for more than 16 credit hours, the effects sizes above 16 are largely regressed to the mean through the

shrinkage introduced by the hierarchical prior (i.e. there is very little information coming from the data).

First Generation

We estimated three separate effects for each for each of the three first-generation status levels in our data: First-generation, Not first-generation, and Generation unknown.

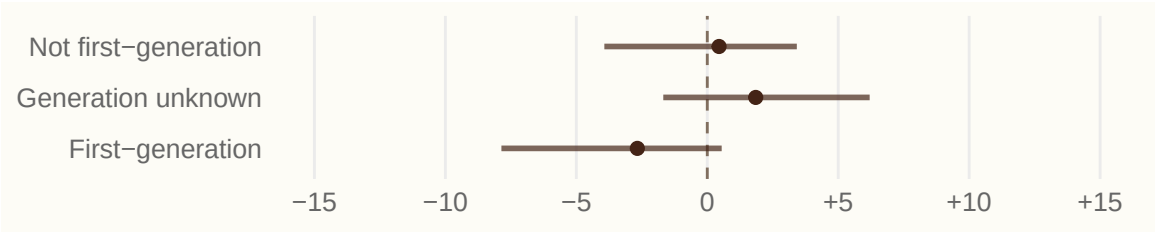

Figure 11: Posterior summaries (median +/- 90%) of the marginal effects for first-generation status. Estimates are reported as the effect on the probability (in percentage points) of a FTIAC returning for a second year relative to an average student (dotted line).

Figure 11 shows substantial variation among and with the different marginal effects. Students who are first generation appear to be around 3 percentage points less likely to return that students who are not first-generation.

We can see this difference more clearly by replotting the effects using the students who are *not* first-generation as the baseline:

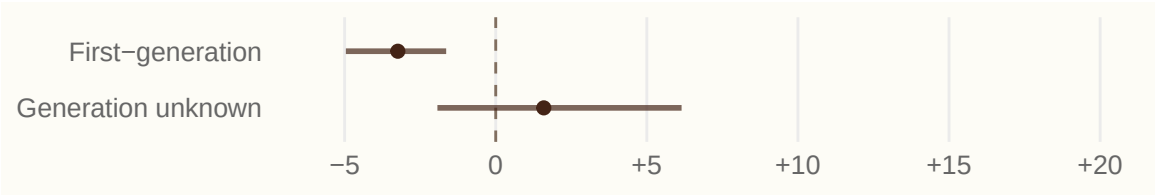

Figure 12: Posterior summaries (median +/- 90%) of the marginal effects for first-generation status. Estimates are reported as the effect on the probability (in percentage points) of a FTIAC returning for a second year relative to 'Non First-generation' students (dotted line).

In figure 12 the dotted line now represents 'Not first-generation'. Here it's clear that first-generation students are somewhere between 2 and 5 percentage points less likely to return than non first-generation students. Students without parental education data are more or less equivalent to the non-first-generation students.

College

We estimated effects for each of the seven colleges plus University Curriculum students treated as a separate college:

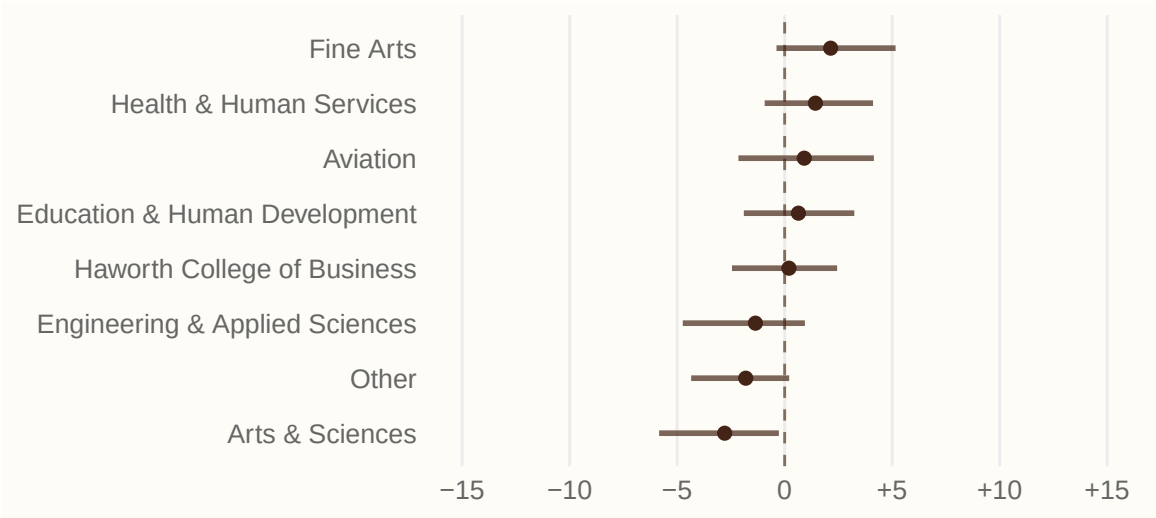

Figure 13: Posterior summaries (median +/- 90%) of the marginal effects for colleges. Estimates are reported as the effect on the probability (in percentage points) of a FTIAC returning for a second year relative to an average student (dotted line). Colleges are associated with a student's primary major (not degree-granting college).

There exists some variation among colleges in their marginal effects on retention. Students enrolled in Fine Arts majors are between 0 and 2 percentage points more likely to return than the average students, whereas Arts & Sciences and University Curriculum students are between 0 and 2 percentage points less likely to return than the average student.

Department

We estimated effects for 51 different departments nested within 6 colleges (University Curriculum and Aviation Sciences students do not have separate department effects.). Because of this nesting, department effects estimate the marginal change in probability relative to the college average. The plot below shows the departmental effects grouped by college:

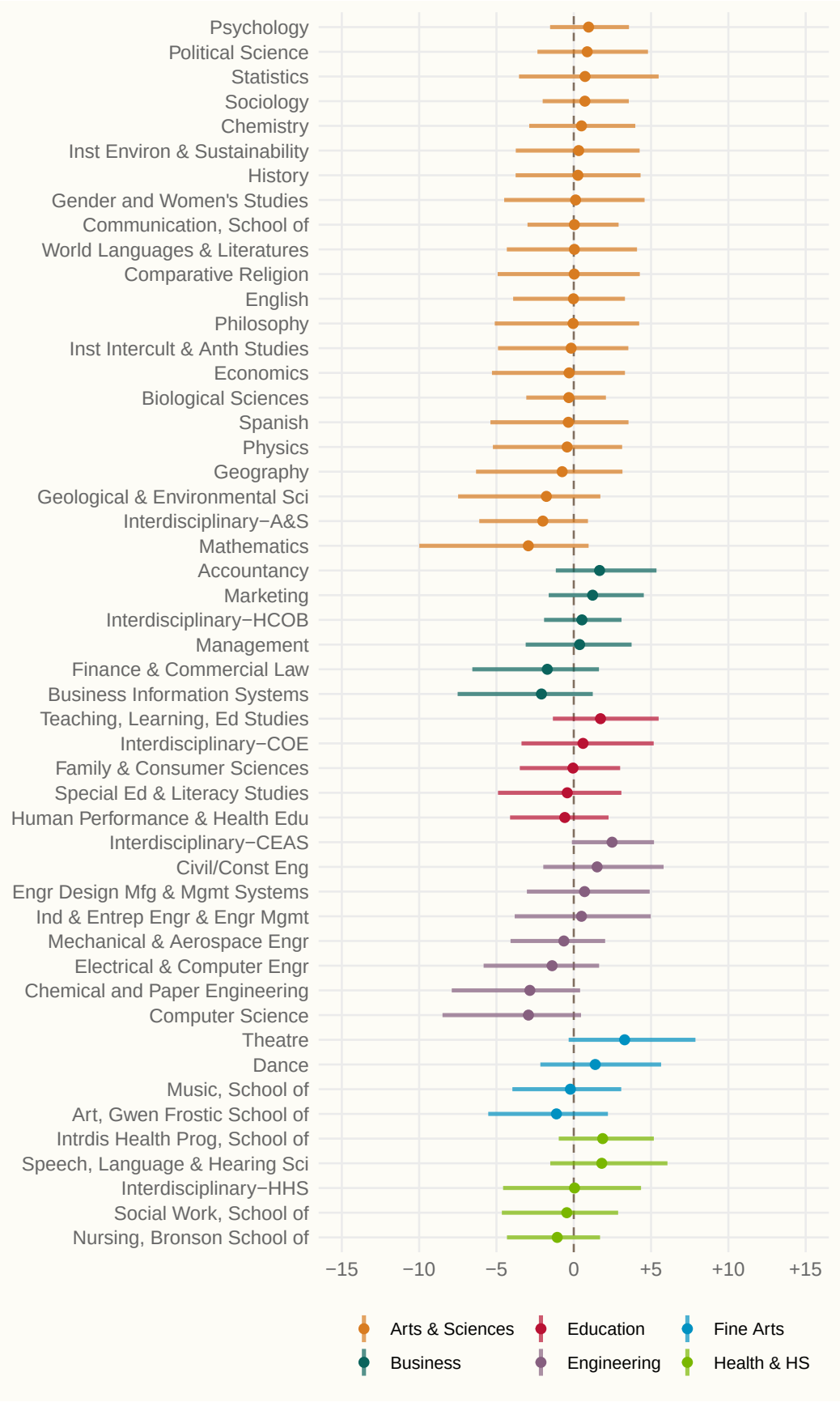
Psychology
Political Science
Statistics
Sociology
Chemistry
Inst Environ & Sustainability
History
Gender and Women's Studies
Communication, School of
World Languages & Literatures
Comparative Religion
English
Philosophy
Inst Intercult & Anth Studies
Economics
Biological Sciences
Spanish
Physics
Geography
Geological & Environmental Sci
Interdisciplinary−A&S
Mathematics
Accountancy
Marketing
Interdisciplinary−HCOB
Management
Finance & Commercial Law
Business Information Systems
Teaching, Learning, Ed Studies
Interdisciplinary−COE
Family & Consumer Sciences
Special Ed & Literacy Studies
Human Performance & Health Edu
Interdisciplinary−CEAS
Civil/Const Eng
Engr Design Mfg & Mgmt Systems
Ind & Entrep Engr & Engr Mgmt
Mechanical & Aerospace Engr
Electrical & Computer Engr
Chemical and Paper Engineering
Computer Science
Theatre
Dance
Music, School of
Art, Gwen Frostic School of
Intrdis Health Prog, School of
Speech, Language & Hearing Sci
Interdisciplinary−HHS
Social Work, School of
Nursing, Bronson School of
−15
−10
−5
0
+5
+10
+15
Arts & Sciences
Education
Fine Arts
Business
Engineering
Health & HS

Figure 14: Posterior summaries (median +/- 90%) of the marginal effects for departments. Estimates are reported as the effect on the probability (in percentage points) of a FTIAC returning for a second year relative to an average student in the associated college (dotted line). Departments are associated with a student's primary major.

All colleges show a comparable degree of among-department variation in effect sizes. Only a handful of individual departments have clear positive or negative effects on retention relative to their college averages (e.g. Interdisciplinary-A&S).

However, another way compare departments is to combine the college and department effects together, giving the departmental effects relative to the university at large:

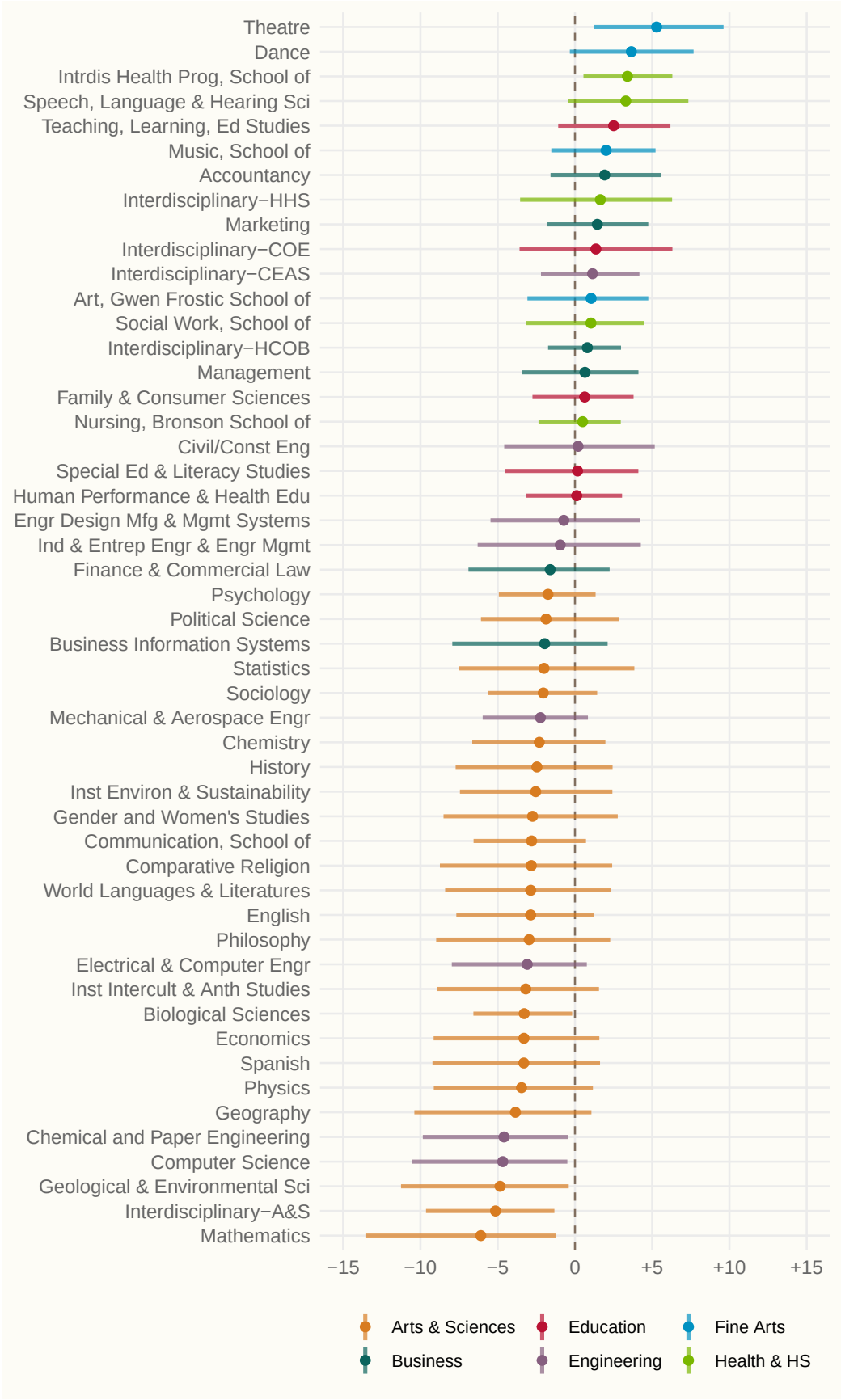
Theatre
Dance
Intrdis Health Prog, School of
Speech, Language & Hearing Sci
Teaching, Learning, Ed Studies
Music, School of
Accountancy
Interdisciplinary−HHS
Marketing
Interdisciplinary−COE
Interdisciplinary−CEAS
Art, Gwen Frostic School of
Social Work, School of
Interdisciplinary−HCOB
Management
Family & Consumer Sciences
Nursing, Bronson School of
Civil/Const Eng
Special Ed & Literacy Studies
Human Performance & Health Edu
Engr Design Mfg & Mgmt Systems
Ind & Entrep Engr & Engr Mgmt
Finance & Commercial Law
Psychology
Political Science
Business Information Systems
Statistics
Sociology
Mechanical & Aerospace Engr
Chemistry
History
Inst Environ & Sustainability
Gender and Women's Studies
Communication, School of
Comparative Religion
World Languages & Literatures
English
Philosophy
Electrical & Computer Engr
Inst Intercult & Anth Studies
Biological Sciences
Economics
Spanish
Physics
Geography
Chemical and Paper Engineering
Computer Science
Geological & Environmental Sci
Interdisciplinary−A&S
Mathematics
−15
−10
−5
0
+5
+10
+15
Arts & Sciences
Education
Fine Arts
Business
Engineering
Health & HS

Figure 15: Posterior summaries (median +/- 90%) of the marginal effects for departments. Estimates are reported as the effect on the probability (in percentage points) of a FTIAC returning for a second year relative to an average student (dotted line). Departments are associated with a student's primary major.

By combining effects it can be seen that a handful of departments have clear negative marginal effects on retention relative to the university at large. For example, students in the Interdisciplinary Arts & Sciences, Mathematics, Geological and Environmental Sciences, Computer Science, and Biological Science all are expected to retain between 0 and 10 (or higher) percentage points lower than average. In contrast, students in the Theatre department, for example, are expected to retain between 1 and 9 percentage points higher than average.

In all cases the uncertainty in the effect sizes is fairly large, which is attributably to the fact that (1) the number of students in any given department is *relatively* small (some more than others) and (2) the uncertainty shown above is the combined college and departmental level uncertainty.

Pell-eligibility

Students were classified into three different Pell-eligibility levels based on expected family contribution (EFC). We can plot the marginal effect of each of these three levels relative to an average student:

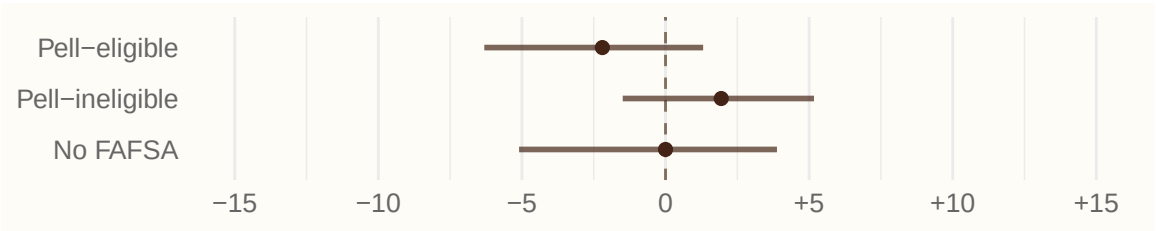

Figure 16: Posterior summaries (median +/- 90%) of the marginal effects for pell eligibility. Estimates are reported as the effect on the probability (in percentage points) of a FTIAC returning for a second year relative to an average student (dotted line).

From figure 16 it appears that eligible students are about 5 percentage points less likely to return than the average ineligible student or student without a FAFSA. There also appears to be little difference between the ineligible and no-FAFSA groups. We can clarify these contrasts by replotting the marginal effects relative to the ineligible students:

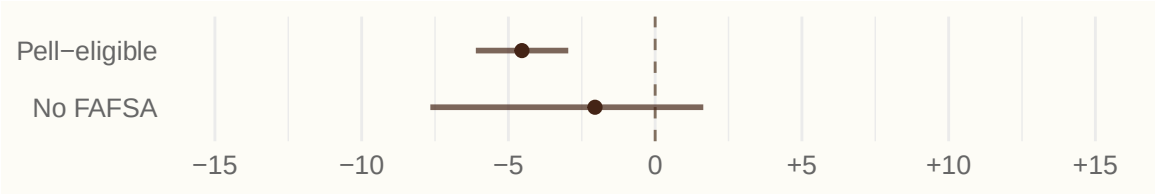

Figure 17: Posterior summaries (median +/- 90%) of the marginal effects for pell eligibility. Estimates are reported as the effect on the probability (in percentage points) of a FTIAC returning for a second year relative to an average Pell-ineligible student (dotted line).

This confirms that eligible students are between 4 and 6 percentage points less likely to return than the average Pell-ineligible student.

Further, keep in mind that the Pell-ineligible effect is estimated for a student an average EFC ($20,453), meaning that ineligible students with higher EFC's will be even more likely to return when compared with Pell-eligible students. For example, if we consider a student with an EFC of $40,906, then we would expect the percentage point margin to go up by around 2 percentage points (about half the effect size shown for EFC in figure 1).

Primary ethnicity

We estimated 8 different primary ethnicity effects (recall that no ethnicity effect is estimated for international students). Of all the group effects estimated by the model (Fig. 5) primary ethnicity accounted for the least amount of variation in retention.

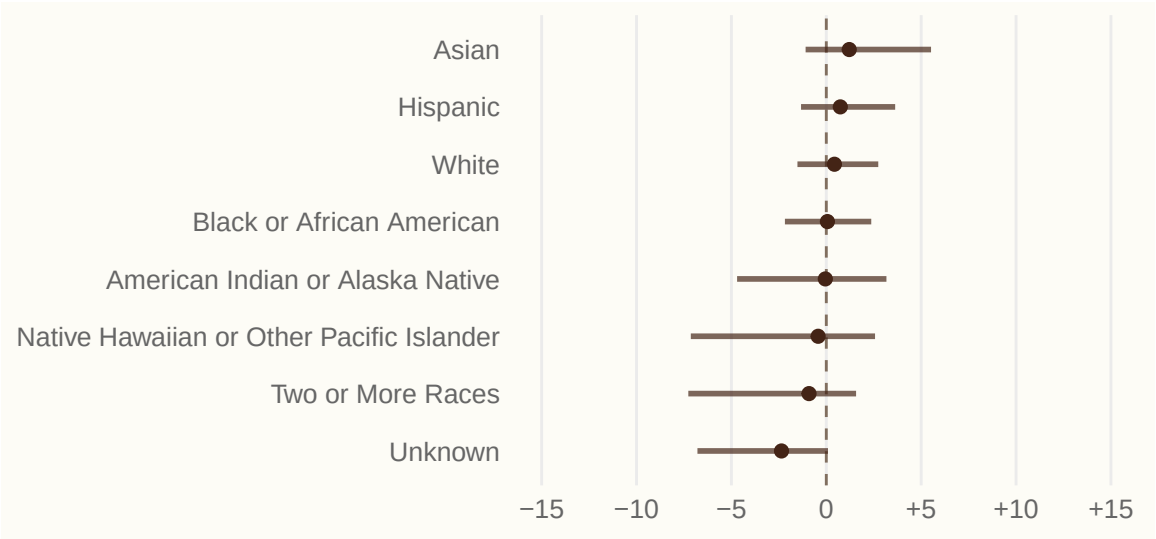

Figure 18: Posterior summaries (median +/- 90%) of the marginal effects for primary ethnicities. Estimates are reported as the effect on the probability (in percentage points) of

a FTIAC returning for a second year relative to an average student (dotted line).

There is little evidence of any appreciable effects of ethnicity on the probability of returning. In all cases the median marginal effects vary between -1 and +1 percentage points relative to the average student. There is a suggestion of negative effects associated with 'Unknown' students and positive effects associated with both white (non-Hispanic) and Asian students

To focus more specifically on the effects for minority ethnicities we can also plot the primary ethnicity effects relative to the effect for white students[9] :

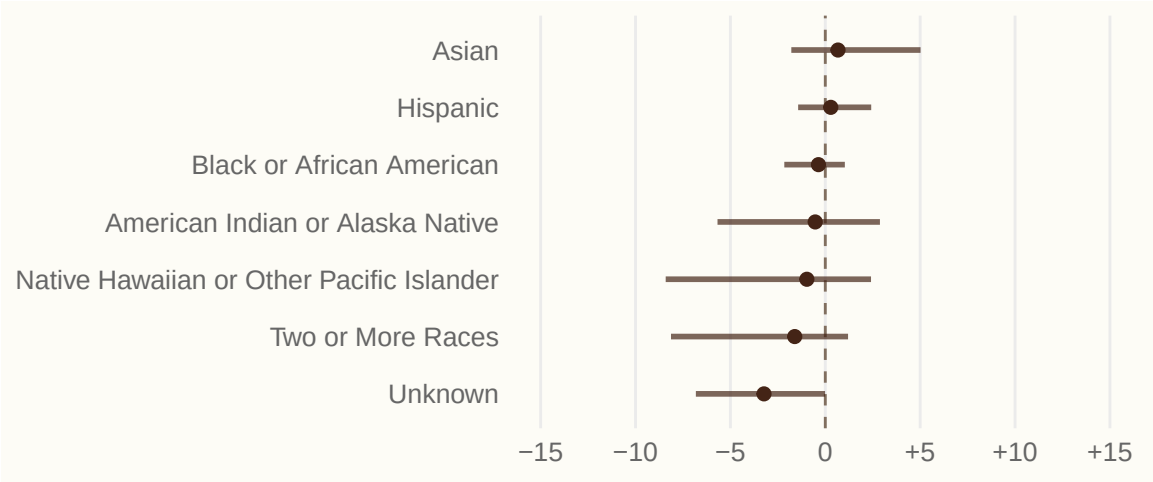

Figure 19: Posterior summaries (median +/- 90%) of the marginal effects for primary ethnicities. Estimates are reported as the effect on the probability (in percentage points) of a FTIAC returning for a second year relative to students reporting as 'White' (non-hispanic).

Figure 19 confirms that there are no strong differences between students reporting as white and other ethnicity groups. However, 'Black or African American' and 'Unknown' students may have between 0 and 3 percentage point lower probabilities of returning than white students, although extremely small effects («1) cannot be ruled out in either case. It is important to point out that these effects are in addition to any differences between ethnic groups accounted for by HSGPA, EFC, etc.

Model Assessment

We performed three primary assessments of our model's predictive ability: (1) posterior predictive checks, (2) psuedo-R^2 as a measure of goodness-of-fit and (3) predictive accuracy.

[9] White students make up between 66% and 72% of each cohort in the data set. https://wmich.edu/institutionalresearch/reporting/reports/retention/retention/20190306_second_year_retention_report.html

Posterior predictive checks were used primarily in the process of model building and all showed that the model could consistently replicate the retention levels for all the sub-groups we tested in the original data. We report goodness-of-fit and predictive accuracy metrics below.

Goodness-of-fit

We used the proportional reduction in deviance ($R_L{}^2$) of the full model as our pseudo R^2 metric for goodness-of-fit. We calculated $R_L{}^2$ when applied to both the training and testing data sets. The $R_L{}^2$ value *should* be lower for the out-of-sample testing data set (because the model will always fit the data used to estimate the parameters than it will out-of-sample data).

We can plot the posterior distributions for $R_L{}^2$ on a scale ranging from $R_L{}^2 = 0$ (equivalent to an 'intercept-only' model) to $R_L{}^2 = 1$ (equivalent to a fully saturated model where every student is predicted correctly):

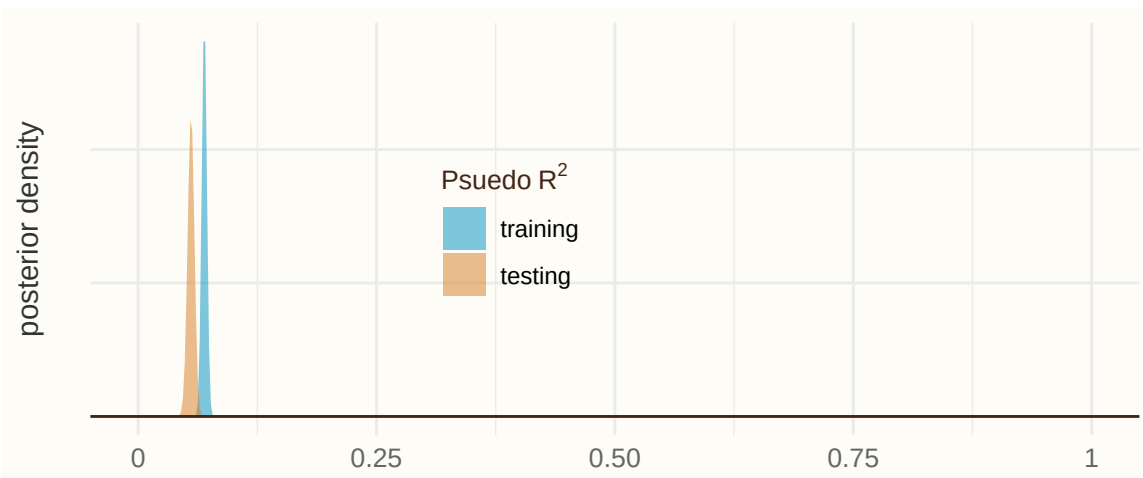

Figure 20: Posterior distribution of the $R_L{}^2$ statistic for the full model applied to the training and testing data subsets. $R_L{}^2$ is calculated as: (Null Deviance - Fitted Deviance)/(Null Deviance). An $R_L{}^2$ value of 1 is equivalent to a fully saturated model while a value of zero is equivalent to an intercept-only model.

As expected the $R_L{}^2$ value for the testing data is slightly lower than the value for the training data, confirming that the model is 'overfit' slightly. Both values are close to zero, accounting for only about 5-7% of the deviance difference between the intercept and saturated model[10]. The intercept-only model is actually quite informative relative to a random model, and our model improves on it slightly via the inclusion of the additional explanatory variables.

[10]This is not surprising considering that the intercept-only model assigns a probability of returning equivalent to the overall retention rate in the original data set

Predictive accuracy

We also assessed the model's out-of-sample predictive accuracy by using it to predict retention data for the testing data set. The higher the out-of-sample predictive accuracy, the better the model (generally). We calculated the full joint posteriors for accuracy, sensitivity, and specificity.

In addition to calculating accuracy metrics for the full model presented here, we also calculated them for the 'intercept-only' model and for a random (p = 0.5) model. The full model should definitely be better than a random model and hopefully should be an improvement over the 'intercept-only' one as well (given the non-zero R^2 values reported above).

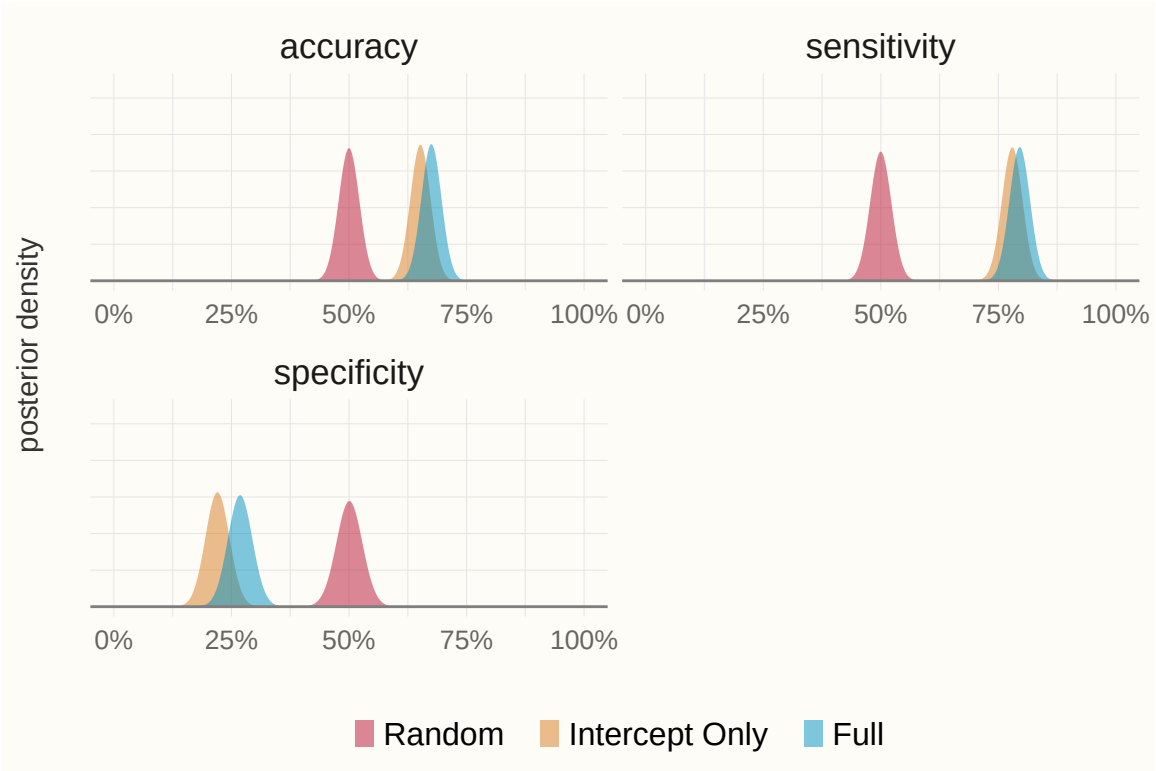

Figure 21: Posterior accuracy, sensitivity, and specificity of the random, intercept-only, and the full fitted model measured against the testing data subset.

The predictive accuracy of the full model out-performs the random model by a substantial amount (about 20 percentage points) but it only outperforms the intercept-only model by about 2 percentage points. This is consistent with the ${R_L}^2$ value of 0.07, which indicated only a slight improvement in the model's predictive capability over the intercept-only model. When we break down accuracy into it's sensitivity and specificity components we an see that most of the advantage of the full model comes from increased specificity (i.e. the ability to correctly predict students who did not come back). Sensitivities of the full and intercept-only models were nearly identical.

One might ask how it is that the random model has greater specificity than either the full or intercept-only model. There are two reasons. First, three-quarters of the students return every year, so an overall more accurate model will prioritize sensitivity over specificity. Second, there is a small negative tradeoff between increased specificity and increased sensitivity in the model. We can see these relationships most clearly by plotting the posterior distributions for all three metrics against one another:

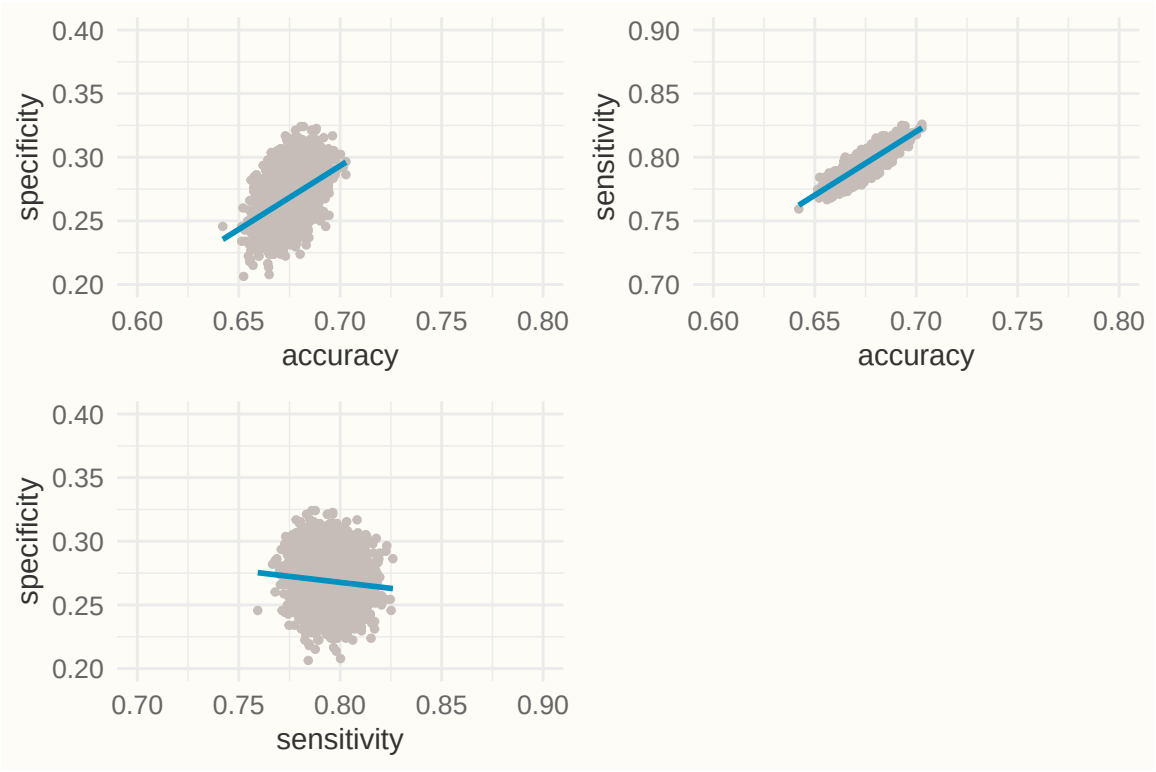

Figure 22: Posterior scatterplots of metrics of model classification success. Each point represents a calculation from a single draw from the joint posterior distribuion. Blue lines indicate best-fit linear regressions.

The upper-left and upper-right panels show that accuracy is more tightly correlated with sensitivity then when specificity goes up. In contrast, the lower-left panel shows that there is a slight negative relationship between specificity and sensitivity in the posterior. This indicates that sensitivity and specificity are largely uncoupled in our model and that increasing one does not necessarily increase the other.

Conclusions

Our stated purpose for this model is to better understand the relationship between pre-entry variables and student retention. In this section we first look at variable effects on both

individual and overall retention. We then discuss some issues regarding prediction from our model. Lastly we highlight some future directions for investigating retention and student success more generally at WMU.

Variable effects on retention

One of the primary reasons for creating an inferential regression model is to understand how the outcome of interest (in this case, student retention) varies as a function of a set of pre-defined variables. Below we run though the variables in our model in (rough) order of decreasing effects on *individual* retention. For each variable we highlight the relative magnitude of the effect, how important is for determining overall WMU retention rates, and potential impacts for increasing future retention[11].

Residency: One of the two variables most strongly associated with student retention is residency: specifically, the distinction between international students and domestic students. International students were between 10 and 15 percentage points more likely to return than domestic students on average (Figure 6). This result should not be surprising given that international students typically travel quite far to attend WMU and they are frequently funded by third-party sources. However, it is important to keep in mind that international students have historically[12] made up between 1 and 2% of all FTIACs, and so the effect of residency on overall retention may be limited to the distinction between in-state and out-of-state students (between whom there is less difference) (Figure 7).

High School GPA: High school GPA is the second variable most strongly associated with student retention. Moving from one standard deviation below to one standard deviation above the mean HSGPA was associated with a 12 to 15 percentage point increase in the probability of returning for domestic students (Figure 1). This effect is roughly equivalent to the difference between international students and domestic residents. Moreover, HSGPA likely has a *much* stronger impact on determining overall retention because any moderate shifts in the underlying HSGPA distribution are going to have noticeable affects on retention.

Living on Campus: The effect of living on campus (versus living off campus) was about half the HSGPA effect but still substantial (Figure 1). It's effect on retention overall (and retention in the future) is mitigated somewhat by the fact that 90% of FTIACs already live on campus. Consequently, large increases in retention (i.e. greater than 1 percentage point) based on increases in the number of students living on campus are unlikely.

Pell-eligibility and EFC: Both Pell-eligibility and expected family contribution were relatively important variables. Pell-ineligible students were around 5 percentage points more

[11] A second benefit of a regression model that we don't explore here is that it can be used to simulate future outcomes given some limiting assumptions. For example, what would our retention look like if we increased out-of-state enrollment to 50%? Or, what would retention be if our out-of-state students came from the same socioeconomic brackets that our in-state students did? While the certainty of these estimates will be limited by the uncertainty inherent in the model and the dependence on prior years data, they can be informative for exploring whether hypothetical changes to the freshman profile would make retention look substantially different (or not).

[12] the last decade

likely to return than the average Pell-eligible student (Figure 17). Combined with the effect of EFC itself (~ 2 percentage point increase every time EFC doubles), and there could be as much as a 10 to 12 percentage point difference between the poorest and wealthiest students. Moreover, Pell-eligibility likely has a large affect on overall retention because 30% of WMU undergraduates are Pell-eligible, and many eligible students have EFC values well under $20,000 (Figure 4).

Counties: The county effects–and their interaction with residency effects– are some of the more challenging to interpret. While it is true that some counties that showed distinct positive or negative effects *relative to their residency level*, the net effects of county and residency are more complex.

If we combine the results from figures 6 and 8 together, we can see that the average student from the best-represented Michigan counties have nearly equivalent marginal probabilities of returning as the average domestic non-Michigan resident. On the other hand, students from Michigan counties that send relatively few students return at much lower rates than both resident students from well represented counties and non-residents.

This gap between the few the best-represented Michigan counties and the less well represented appears to account for a large part of the observed difference between domestic residents and domestic non-residents in retention rates (which has been about four percentage points in favor of non-residents over the last decade). *Put another way, if all of our resident students came from the five counties in Michigan from which we currently obtain most of our new students, the difference we observe between domestic residents and domestic non-resident students would be much smaller.*

It is important to note that the single non-Michigan country from whom we receive the *most* non-resident students – Cook County, Illinois – has the most negative marginal affect of individual retention out of all counties from where we received at least 50 students of the time-period included in the model. This suggests (1) there maybe some specific factors shared by students from Urban Chicago that negatively affect retention and (2) the effect of domestic non-resident retention would likely be higher without the large contingent of students from Cook County.

Credit Hours: We found some evidence for an association between increased credit hour registration and increased individual retention–up to a point. Students who enrolled for 14 or 15 credit hours were between 3 to 6 percentage points more likely to return than students who only enrolled for 12 (Figure 10). However, as the number of registered hours increased beyond 16 there appeared to be no discernable difference from students registered for only 12 hours.

Although the above relationship suggests there may be a 14-15 credit hour sweet spot, there are some caveats. First, we do not have enough data to establish relationships between SCH and retention beyond 16 credit hours, so these effects are heavily regressed to the mean. Second, some students may intentionally take more than 15 credits to establish their credentials prior to transferring to another institution after the semester or year is finished. While these students count as non-returners, they would be fundamentally different than students that transfer to community colleges or that drop-out entirely. Lastly it's possible

that students who take too many credits experience burn out. This last caveat leads to the additional point that not all students could handle more than 12 or 13 credits in their first semester, and there may be diminishing returns for increased retention merely from having students register for more classes[13].

First-generation Status: We found evidence that a first-generation student is between 2 and 5 percentage points less likely to return for a second-year than a student who is not-first generation. First-generation students made up about 16% of incoming FTIACS in 2018[14], meaning that any improvement in this difference would likely to have at least a small impact on overall retention rates.

College and Departments: The model results indicate that there is some variation among colleges and departments in retention and in handful of cases departmental effects may be somewhat strong (Figure 15). Taken all together, colleges in the Arts and Sciences and Engineering colleges tend to retain at lower rates than students in the Fine Arts and Health and Human Services colleges. These differences are more descriptive the proscriptive, as our model cannot truly distinguish between effects due the departments themselves from the effects associated with students who tend to enroll in those departments. Put another way, department affinity is a non-random process and the best we can do is say that being in a certain department or college is *associated* with increased or decreased probability of returning, relative to the average.

Primary Ethnicity: Of the seven group variables we included in the model (not counting the `COHORT` effect), primary ethnicity appears to have the least direct impact on retention. There was no evidence that membership in any particular ethnicity had substantially effects on individual retention relative to the average student. However, if we narrow our focus to differences between white students and other ethnicities, there was some evidence for small negative effects for Black students (between 0 and 3 pp lower) and students who do not report ethnicity (between 0 and 6 pp lower). More data could potentially help to narrow down the magnitude of these effects and make them more precise.

It is important to keep in mind that while there may be little or no *direct* effect of ethnicity on retention, it is clear that *indirect* effects are potentially quite strong. Non-white students, *on average*, arrive on campus with lower HSGPA, fewer AP courses, lower EFC, and a higher % of Pell-eligibility, all of which are strongly associated lower probability or retention. Further our current model also does not look at how ethnicity might *modify* the effects of other variables like HSGPA (i.e. interactions) where, for example, the effects of lower HSGPA could be even stronger for certain ethnicities than others.

One additional caveat to the analysis that has not been addressed fully here are the potential strong correlations between ethnicity and county of origin. If students from some counties primarily report as a single ethnicity (e.g. Black or African American), and a large proportion of students of that ethnicity come from those counties, then there will be strong entanglement in the model estimates for those counties and that ethnicity. In practice this means that some

[13]The 2018 FTIAC cohort (not included in this analysis) showed a jump form 33 to 50% of FTIACs taking 15 SCH so it will be interesting to how this may or may not effect retention in Fall 2019

[14]This depends slightly on the denominator. If we ignore students for whom we have no parental education information, this jumps up to 18.5% of incoming FTIACS in 2018

of the variation that *should* be attributable to ethnicity effects may instead get subsumed into county effects. In those cases county acts as a proxy for ethnicity[15]. Further investigations will be needed to determine whether this entanglement plays a role in the model presented here.

Inferential models and prediction

From a pure prediction stand-point, the inferential model presented here is mediocre at best. However, there is a distinction between predicting student-by-student retention (where the model is not so great) and 'predicting' overall and group level retention (where the model is actually quite accurate if not terribly precise). What we lose in predictive power at the student level is regained by our ability to say something of use about the variables we are interested in, either for their own sake or for how they might influence university policy and decision-making.

None the less it is useful to understand why prediction is compromised in our inferential model, as it sheds light on a number of the issues in making statistical inference from retention data. We go though these reasons below in more detail.

Single binary data points

Student retention data have two distinct properties that can make inference and prediction difficult: the outcome is binary (yes/no) and the sample size for each student is one.

Binary data are a problem because the unknown variable that we want to predict about *future* students must be couched in terms of probability. That is, when we say "will this student return or not?" in practice we mean "what is the probability this student will return?"[16]. We must therefore infer probabilities from data where only yes/no outcomes exist.

Compounding this issue is the fact that we only have a sample size of one for each student. If we could measure each student's second-year retention repeatedly, we could get a much better estimate of their underlying probability of retention. For example, if we could measure retention for the same student 1,000 times, and they returned 750 of those times, we would be confident in saying that the student's underlying probability of returning was 0.75 with high precision.

The issue of estimating individual probabilities even extends to groups of students. Imagine that out of a group of 100 students (say all the students in a given department) 80 students returned for their second year. What were the probabilities that each of the 100 students was going to return? At one extreme, maybe every student had an 80% chance of returning and 80 out of 100 was what we got. On the other extreme, maybe 80 students had 100%

[15]This geographic entanglement is one of the reasons why predictive or determinative models that *explicitly* exclude race and ethnicity for egalitarian reasons can still be *implicitly* biased via the inclusion of variables like county.

[16]Not unlike the distinction between "Will this candidate win the election?" and "What is the probability this candidate will win the election?"

change of returning and 20 students had 0%. Both scenarios will lead to the same general outcome–80 out of 100 students coming back. What a model like ours is doing is trying to distinguish the latter scenario (are some students closer to 0 or 1?) from the former scenario (where we just assign everyone a probability of 0.8) to the degree possible from the data at hand. Having only one observation per student makes this a more challenging task.

In terms of prediction, the consequence of having single, binary data points will be low precision for–and high uncertainty around–individual probability estimates. Practically, this means that the model's goodness-of-fit and overall accuracy is not much better than a weighted coin flip unless variables (or combinations of variables) in the data clearly delineate returners and non-returners.

Multiple real-life outcomes

Another big issue is that the observed yes/no data encapsulate a number of potential real-life outcomes. Consider the following, all of which are counted as a student not returning:

- a student drops out of higher education and enters the job market
- a student stops out for a semester because of financial reasons to work for year before coming back to WMU
- a student transfers to a community college to get an associates degree before coming back to WMU
- a student transfers to a community college to get an associates degree and then enter the labor force
- a student decides to withdraw from all their classes after census and reenroll the next fall after taking a year off
- a student really wanted to be at a different institution but wasn't able to get in on their first try and transfers out after a year
- a student enrolled at WMU for a reason that no longer exists (girlfriend/boyfriend, program of study, etc. . .) and transfers to another school

In each of those scenarios, the primary reason why a student does not come back is different. In some cases, the primary reason might be unique to that student alone. This means that no single pre-entry variable or combination of variables that we have available is going to be a good at predicting more than one or two of these outcomes with any degree of certainty. In fact, some outcomes (a student transfers to a preferred 4-year institution) are going to be really hard to predict because we don't (yet) have widespread reliable measures of students intentions.

Multiple outcomes are the primary reason why the specificity of the model is disappointingly low (Figure 22). A model that uses pre-entry variables to understand retention will tell us a great deal about how those variables are associated with retention overall, but it won't be good at picking out the reasons why *individual* students choose to stay or particularly why they choose to leave.

Variables are proxies

Although we all intuitively understand it, it's important to reiterate that variables are just proxies for the underlying factors that we are ultimately interested in. Some variables are more closely related to those underlying factors that others, and we need to be cautious about extrapolating from a relationship between a variable and retention to the underlying factor. Consider three variables used in this model which act as proxies of differing reliability: HSGPA, EFC, and county:

- **HSGPA**: HSGPA is a proxy for number of underlying but related factors: obvious things like academic aptitude and academic motivation but also the number of advanced courses taken (because HSGPA is adjusted by the admissions office to put it on a common scale). HSGPA is likely a reliable but not perfect measure of these underlying factors. Consequently, the degree to which HSGPA is associated with retention is at least somewhat indicative of differences academic aptitude or motivation affecting a students likelihood of returning.
- **Expected Family Contribution**: The reliability of EFC as a proxy for something like socioeconomic status is murkier. EFC measures how much (according to the federal government) a student's family might be able to pay for their education. EFC can be considered a proxy for wealth and socio-economic status, but it's also related to parents education and occupation, student's family situation (zero vs. single vs. two parents), family size, and even more subtle with factors like health and well-being. All of these are captured to a greater and lesser degree by EFC, and not all of them will influence retention in the same way. Thus, while we might *say* that EFC is a proxy-variable for socioeconomic status, it encapsulates so much more.
- **County**: It's clear that students from different counties return, or are expected to return, at different rates. But what does it mean to be 'from a county' or to say that a county has a certain effect? In truth it could mean any number of things, and it can mean different things for different counties. In some cases county may act as a proxy for urban/suburban/rural distinctions. It's quite possible that county may act as a proxy for ethnicity. It's not hard to imagine it acts as a proxy for things like socioeconomic status, parents education, quality of high schools, opinion of WMU, and any other of a host of variables.

When we include county in the model what we are doing is accounting for variation that we know exists between counties, whether or not we are able to pin down the underlying factors driving that variation. Attributing or inferring possible causal mechanisms is much, much more difficult. By looking at patterns among counties in a post-hoc fashion, we may be able to make some inference about what the underlying factors driving variation among counties. We can also incorporate county level variation in the model. For example, if the directions and magnitudes of county effects are associated with the population of the county, then including county population size in the model could tell us something about how students from differently populated areas perform at WMU.

These examples illustrate that the connection between underlying explanatory factors, the variables we have, and the outcome of interest are not always as linear or straightforward as we would like them to be. In all cases, moderating the language of causal effects to the degree warranted by the reliability of variables as proxies is always advised.

Future directions

This modeling effort represents just a first step in understanding second-year retention for a subset of Western students (FTIACs). As we move forward, there are number of directions our research efforts can expand to take a more comprehensive view of student success at WMU. Below we present a road-map for possible new avenues.

Pre- and post-entry variables

This analysis focused exclusively on the role that pre-entry variables have in determining second year retention. This approach makes sense. Many of the efforts for improving retention rates on campus focus on narrowing the gap between students with different pre-entry characteristics: between Pell-eligible and ineligible students, minority and non-minority students, high-achieving students and under-prepared students, etc. Further, this sort of modeling is required for predicting future changes in retention based on shifts in the underlying pre-entry variables (e.g. more non-resident students, more minority students, etc.).

However, we might also be interested in extending the analysis to thinking about how retention is related to post-entry data, especially data on academic performance, student engagement, and financial aid related issues. This would give the university some insight into the more proximal-factors that lead to students dropping out, and also how the more proximal post-entry factors are or are not associated with pre-entry variables.

Student persistence

Although second-year retention is an important metric of student success for a university, it makes sense to expand the model both temporally and conceptually to consider year to year student persistence and eventual graduation. Expanding the scope of the model would allow us to address the more fundamental questions by addressing the major factors impede or promote student success. For example, we could get a better understanding of *when* and *why* students drop out or transfer out at various points in their academic career. We could also start to address factors that slow student progress towards graduation and that, consequently, have large impacts on our 4- and 6- year graduation rates.

More than just FTIACs

In some ways it makes sense to focus an analysis on full-time, first-time students. Universities have typically been measured and assessed by the success of these so-called 'traditional'

students. FTIAC students are also more 'interchangeable' in ways that transfer or reentry students are not, making them more amenable to statistical inferential analysis.

However, full-time, first-time students make up only about 60% of the newly enrolled students every Fall semester and 50% of newly enrolled students for a fiscal year. FTIAC students are also becoming less 'interchangeable' every year as a larger fraction of first-time students arrive on campus as sophomores and juniors, having earned a non-trivial amount of college credits while in high school. If we really want to understand the bigger picture of student success at WMU, we need to take a more comprehensive look at all our students, understanding the ways in which the paths to success differ (or not) for the various categories of students that arrive on campus every semester..

Multi-outcome models

By definition, student retention is based on a binary outcome: did the student return or not return. As discussed above, however, this dichotomy hides a wide variety of real-life outcomes that both add noise to the data and represent meaningful differences for WMU's evaluation of its students, policies, and efforts aimed at improving retention and graduation.

It is possible, though technically challenging, to model retention as a multi-outcome process. For example, instead of two categories we might start by considering three: returned, transferred, and dropped out. This would differentiate the students who are leaving (at least temporarily) high-education altogether versus those who are transferring to another college or university. We might even expand our outcomes by dividing transfers into 2- and 4-year schools, as we might expect the factors driving students to transfer to a community college as distinctly different than those driving students to transfer to another university.

There are some downsides to expanding the outcomes in this way. First, multi-outcome models will be more complex than binary models, with many parameters to estimate and more consideration given to how the variables in the model might effect outcomes differently. By increasing the number of outcomes we also decrease the sample-sizes within each outcome, which can lead to more uncertain estimates. This can be offset by using more data, but data from further back in time may obscure trends that have emerged more recently. In the end, expanding the model complexity always represent a balancing act between signal and noise, maximizing our explanatory power without getting lost in the weeds.

Financial aid and cost-of-attendance

Two related elements that were not explicitly included in the model, but which likely have large effects on student retention, are financial aid and cost-of-attendance. We might reasonably expect that as the gap between the cost-of-attendance and the ability to pay widens, the likelihood that a student won't return to school increases. This problem will be more pronounced for less wealthy students that rely on financial aid or student loans to pay for their education.

Financial aid is implicitly incorporated in the current model to some degree; the Pell-eligibility and EFC effects represent *net* effects after financial aid, awards, and loans have

been applied. One way to think about this implicit inclusion is to consider two extremes. First, imagine what the model would look like if no-one received any financial aid (not even Pell grants) and students were not allowed to take out loans[17]. We would see *extremely* large effects of both Pell-eligibility and EFC on retention, with only the most wealthy students continuing to attend the university. Now imagine the opposite case where every student has all their academic and living expenses completely paid for indefinitely by a third-party. In that situation we might easily imagine that the difference between low- and high-EFC students would shrink to near zero (of course their may some remaining differences associated with EFC that are not related to simply paying for school). In practice, the model will be somewhere in the middle, with the magnitude[18] of the Pell and EFC effects indicating the degree to which financial aid helps overcome socioeconomic differences between students.

However, we may want to say more about how each students' individual financial situation affects their retention. We could consider a number of additional factors: how much the student is paying to attend WMU; the amount, sources, and types of their financial assistance; whether they have financial holds; whether or not they have or are taking out student loans; and what is the gap between what they can afford and what they are paying. These variables can also interact with other variables to affect persistence in non-intuitive ways. A student who is close to graduating might be more willing to take out a loan to cover a gap in funding than a first-year student. A student who is failing academically might not bother spending money to override a financial hold. A student with financial problems might not bother attending courses or working to get non-failing grades. Incorporating detailed data about each students' financial and academic situation will allow us to better understand the role of financial factors in determining student persistence.

Acknowledgements

The author would like to thank: Fen Yu, Jason Jach and the full Office of Institutional Research staff for advice, comments, and support; Jeremy Glefke, Jodie Laraway, and Rachel Collingsworth for help with understanding and utilizing financial aid information; and the WMU Retention Action Team for acting as a sounding board for earlier iterations of model.

References

Buuren, Stef van, and Karin Groothuis-Oudshoorn. 2011. "Mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software* 45 (1): 1–67. https://doi.org/10.18637/jss.v045.i03.

Buuren, Stef van, Karin Groothuis-Oudshoorn, Alexander Robitzsch, Gerko Vink, Lisa Doove, Shahab Jolani, Rianne Schouten, Philipp Gaffert, Florian Meinfelder, and Bernie

[17] of course, many students wouldn't bother applying or registering, but imagine they were forced to

[18] technically inverse of the magnitude

Gray. 2019. "Mice: Multivariate Imputation by Chained Equations." https://CRAN.R-project.org/package=mice.

Gelman, Andrew. 2006. "Prior Distributions for Variance Parameters in Hierarchical Models (Comment on Article by Browne and Draper)." *Bayesian Analysis* 1 (3): 515–34. https://doi.org/10.1214/06-BA117A.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. 2008. "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models." *The Annals of Applied Statistics* 2 (4): 1360–83. https://doi.org/10.1214/08-AOAS191.

Gelman, Andrew, and Iain Pardoe. 2007. "Average Predictive Comparisons for Models with Nonlinearity, Interactions, and Variance Components." *Sociological Methodology* 37 (1): 23–51. https://doi.org/10.1111/j.1467-9531.2007.00181.x.

Guo, Jiqiang, Jonah Gabry, Ben Goodrich, Daniel Lee, Krzysztof Sakrejda, Modrak Martin, Trustees of Columbia University, et al. 2018. "Rstan: R Interface to Stan." https://CRAN.R-project.org/package=rstan.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Stan Development Team. 2017. *Stan Modeling Language User's Guide and Reference Manual (Version 2.17.0)*. https://mc-stan.org/docs/2_18/stan-users-guide/index.html.