

Structuring College Access: The Market Segment Model and College Board Geomarkets

1 Methods

I love doing research. I love teaching. I love teaching students how to do research.

the shit I study is weird. it might strike some people as obscure, but I believe the things I study are fundamental to equality of opportunity in college access. they just haven't been studied before, so they seem obscure.

a riverbed predicts which way the water will flow. that's sorta how I think about Geomarkets.

do you wanna be in the business (Jaquette et al. 2022)

During the 2017 calendar year we collected data on recruiting visits to high schools from a sample of 26 selective private colleges. RQX asks, to what extent do College Board Geomarkets explain which high schools receive recruiting visits from selective private high schools, after controlling for other variables thought to explain visit patterns?

Analyses are cross-sectional. The dependent variable, Y_{ij} indicates whether high school i received at least one recruiting visit from college j . We began by estimating the simplest possible model and then incorporated additional sophistication. In the simplest model, Y_i indicates whether high school i received at least one recruiting visit from a particular college. This model can be run separately for each college, j .

The independent variable of interest is the Geomarket that high school i is located in. Geomarket is a categorical variable that contains 304 categories. There are two broad choices in how to treat Geomarket: (1) include as fixed effects by treating Geomarket as a factor variable; or (2) include as random effects.

Practically, implementing fixed effects for Geomarkets means including a dummy variable for each Geomarket except for a reference group. Fixed effects for Geomarkets would capture **all** variation in high school visit patterns that varies between Geomarkets but is constant within Geomarkets. After including fixed effects, the remaining variation in high school visits is variation within geomarkets because variation between geomarkets has been eliminated by fixed effects. In econometrics, the purpose of implementing fixed effects for Geomarkets would usually be that we are substantively interested in the effect of a different independent variable, but we want to eliminate potential

omitted variable bias from Geomarkets. For example, we might be interested in estimating the causal effect of school enrollment size on the probability of obtaining a visit. However, we are concerned that school enrollment size is correlated with Geomarket and that Geomarket also has a causal effect on high school visits. If so, our coefficient on school enrollment size is biased because it is partially picking up the effect of Geomarket on high school visit. If we include fixed effects for Geomarkets, then we eliminate between Geomarket variation as a source of potential omitted variable bias in the coefficient on school enrollment size.

Attempting to implement fixed effects for 304 Geomarkets introduces several practical problems. First, this factor would add $k - 1 = 303$ extra parameters which would reduce degrees of freedom. Second, Geomarkets with few high schools would lead to unstable coefficient estimates. Third, some Geomarkets would perfectly predict the outcome because college j visits zero high schools in the Geomarket. Perfect prediction causes huge coefficients and standard errors. Third, interpreting 303 coefficients is impracticable. Fourth, models with many parameters may not converge in glm.

EXPLANATION OF RANDOM EFFECT A random intercept model is the alternative to implementing fixed effects for Geomarket. A random intercept model is the simplest and most common case of the broader random effects model.

Random intercept model (simplest type)

- Only the intercept varies across groups.
- Slopes for all other predictors are the same for everyone.

Formula for a logistic version:

$$\text{logit}(p_{ij}) = \beta_0 + u_j + \beta_1 x_{ij} + \dots$$

where $u_j \sim N(0, \sigma_u^2)$.

Random slopes model (more complex)

- The intercept **and** one or more slopes vary by group.
- Example: both baseline visit probability and the effect of school size vary across Geomarkets.

Formula:

$$\text{logit}(p_{ij}) = \beta_0 + u_{0j} + (\beta_1 + u_{1j})x_{ij} + \dots$$

where both u_{0j} and u_{1j} are random draws from a joint distribution.

In a fixed effects model, each Geomarket is associated with its own dummy variable and its own coefficient. By contrast, in a random intercept model each group (e.g., each Geomarket) has its own intercept shift. However, we do not estimate a separate, fixed coefficient for each value of Geomarket. Rather, these extra coefficients for each group are modeled as coming from a shared distribution (such as a normal distribution). We use the word “random” effects because we don’t estimate a fixed coefficient for each group. Instead, we treat the effect for each group as “random draws from a population distribution”

Generally, in a random effects model, coefficients on a variable may vary across clusters or groups, but we don’t estimate the exact value of the coefficient for a particular group. A random intercept model is a special case of the random effects model in which *only the intercept* is allowed to vary by group, while the coefficients on other predictors are the same across groups. A simple random intercept model is shown here:

$$\text{logit}(p_{ij}) = \beta_0 + u_j + \beta_1 x_{ij} + \dots$$

where

- i = individual school, j = EPS region.
- β_0 = overall intercept (fixed part).
- u_j = random intercept for group j — a random effect drawn from a distribution, usually $N(0, \sigma_u^2)$.

Every Geomarket, j gets its own intercept shift, u_j , capturing unobserved differences between Geomarkets. In contrast to fixed effects model, the model does not estimate a separate coefficient for each Geomarket. Rather, it estimates the variance the Geomarket intercept σ_u^2) and “shrinks

extreme estimates toward the overall mean, especially for small-N groups” [quotes from GPT5]

ChatGPTs version of my paragraph: “**Random Effects Models**

Random effects are an alternative to fixed effects for Geomarket. In a fixed effects model, each Geomarket has its own dummy variable and its own coefficient. In a random effects model, we still allow each Geomarket to have its own intercept shift, but we do not estimate a completely separate free parameter for each Geomarket. Instead, these Geomarket-specific intercepts are modeled as random draws from a common distribution, usually $N(0, \sigma_u^2)$.

The term “random” reflects the fact that we treat these effects as realizations from a distribution, not as unrelated constants. The model estimates the variance of these effects (σ_u^2) and produces predicted intercepts for each Geomarket, shrinking extreme values toward the overall mean — especially for groups with few observations.”

1.1 subsection

Our goal is to measure the incremental contribution of Geomarket to explaining variation – or improving predictions – after controlling for other variables. Conceptually, this could be achieved with either fixed effects (i.e., a factor with 304 categories) or with random intercepts. However, a factor with 304 categories runs into big practical problems (303 extra parameters » big degrees-of-freedom hit; Sparse groups » unstable estimates, huge SEs, perfect prediction issues; potential convergence problems in glm). By contrast, with random intercepts we:

- Add only 1 variance parameter instead of 303 dummy coefficients
- Get shrinkage for small-N Geomarkets → more stable
- Avoid perfect separation issues in many cases
- Still capture the between-Geomarket variation you’re interested in

The approach to seeing whether Geomarket meaningfully improves fit/prediction proceeds like this:

- fit model with all other predictors except Geomarket
- fit random intercept model with Geomarket + all other predictors

3. Ways to quantify improvement

Model fit metrics

- Pseudo- R^2 : Compare McFadden's or Nagelkerke's pseudo- R^2 between models.
- AIC/BIC: Lower values indicate better fit (with a penalty for extra parameters).

Prediction accuracy

Use cross-validation (CV) to compare:

- Accuracy
- Log-loss
- AUC (area under ROC curve)

This directly tests if Geomarket improves out-of-sample predictions.

Variance explained (random effects only)

If you fit a random intercept model, you can also look at:

- ICC (Intraclass Correlation Coefficient): Proportion of total variance explained by Geomarket grouping.
- Reduction in residual variance after adding Geomarket.

Interpretation for your question. If adding Geomarket substantially improves pseudo- R^2 , reduces AIC/BIC, and/or improves CV metrics, that's evidence that Geomarket explains additional variation beyond other predictors. In a random effects model, a large ICC means Geomarket accounts for a meaningful share of the unexplained variation in visits.

1.2 Comparing model fit

Want to compare model fit between model that has state fixed effects but not Geomarket to model that does not have state fixed effects but has random intercepts for Geomarket

- chatGPT: 3. Compare model fit

Since one model is GLM (fixed effects) and the other is GLMM (random effects), a direct likelihood ratio test only works if they're truly nested, which they're not in this case. Instead, compare using

information criteria:

```
AIC(mod_state, mod_eps)
```

```
BIC(mod_state, mod_eps)
```

Lower AIC/BIC = better fit (tradeoff between fit & complexity).

AIC differences of:

0–2 → models are essentially equivalent.

4–7 → some evidence for the lower-AIC model.

10 → strong evidence for the lower-AIC model.

Compare pseudo- R^2 / variance explained

You can use `MuMIn::r.squaredGLMM()` for both models to get marginal (fixed effects only) and conditional (fixed + random effects) R^2 .

```
library(MuMIn)
```

```
r.squaredGLMM(mod_state)
```

```
r.squaredGLMM(mod_eps)
```

Interpretation:

- If EPS random intercept gives a big jump in conditional R^2 , that means the geomarket grouping is explaining variance that state dummies can't.
- If the state fixed effect gives similar AIC/BIC and R^2 as EPS, then geomarkets aren't adding much beyond state.

1.3 ICC = intraclass correlation

ICC stands for Intra-Class Correlation.

In your context, it's the proportion of the residual variance in visit likelihood that's attributable to differences between geomarkets (EPS codes), after accounting for all your fixed effects.

What ICC means in a GLMM

In a linear model, ICC is defined as:

$$\text{ICC} = \frac{\sigma_{\text{group}}^2}{\sigma_{\text{group}}^2 + \sigma_{\text{residual}}^2}$$

For a logistic mixed model, the residual variance is not estimated — it is fixed at:

$$\sigma_{\text{residual}}^2 = \frac{\pi^2}{3} \approx 3.29$$

So in your case:

$$\text{ICC} = \frac{\sigma_{\text{EPS}}^2}{\sigma_{\text{EPS}}^2 + 3.29}$$

Why ICC is important

The ICC tells you the proportion of unexplained variance that lies between groups (here, geomarkets).

- **High ICC** → Group membership is a big factor; cases in the same group are more similar than cases in different groups.
- **Low ICC** → Little residual clustering by group; most remaining variance is at the individual level.

Interpreting ICC (based on a model for visits by Middlebury College to all high schools)

Summary for you: Your Conditional $R^2 = 0.621$ says fixed + random effects explain 62.1% of total variance. Your ICC = 0.17 says: of the remaining 37.9% unexplained variance, about 17% of it is due to systematic differences between geomarkets within states.

An ICC of 0.17 in the kind of social/educational setting you're working in is moderately large.

Here's the general rule of thumb often cited (though context matters a lot):

- less than 0.05: low clustering (grouping doesn't matter much)
- 0.05 – 0.15: moderate clustering
- greater than 0.15: substantial clustering

At 0.17, you're above that “substantial” threshold — meaning a non-trivial share of the unexplained variation in visits is attributable to differences between geomarkets (within states).

In practical terms:

Ignoring geomarket clustering could noticeably bias your standard errors and understate uncertainty.

It also means that geomarket-level strategies or characteristics could be an important lever if you were thinking about interventions.

Relationship between Marginal R², Conditional R², and ICC

Marginal R² is the proportion of variance explained by the fixed effects only.

Conditional R² is the proportion of variance explained by the fixed effects **plus** the random effects.

The difference between them represents the variance explained **only** by the random effects.

The ICC (Intra-Class Correlation) is the proportion of the *residual variance* that is between groups.

For logistic mixed models, the residual variance is fixed at:

$$\sigma_{\text{residual}}^2 = \frac{\pi^2}{3} \approx 3.29$$

The ICC is then:

$$\text{ICC} = \frac{\sigma_{\text{group}}^2}{\sigma_{\text{group}}^2 + 3.29}$$

Why a large ICC but small R² gain can happen

It is possible to have a **high ICC** (meaning strong clustering in the residuals) but only a modest

increase from marginal to conditional R^2 .

This happens when the fixed effects already explain most of the variance, leaving little residual variance for the random effects to improve upon.

For example:

- Marginal $R^2 = 0.856$
- Conditional $R^2 = 0.881$
- ICC = 0.175

Here, the fixed effects (including state fixed effects) explain **85.6%** of the variance in the latent log-odds of the outcome.

Only **14.4%** of the variance remains unexplained. The ICC tells us that **17.5% of that residual variance** is between groups (here, geomarkets).

Although that is a large share of the residual variance, in absolute terms it is a small share of the total variance, so the total R^2 only increases by 2.5 percentage points.

Why a high ICC with a very small R^2 gap would be surprising

If marginal $R^2 = 0.40$, conditional $R^2 = 0.43$, and ICC = 0.17, this would be unusual.

Reason:

- Marginal $R^2 = 0.40 \rightarrow$ residual variance = 0.60 of the total.
- ICC = 0.17 \rightarrow about $0.60 \times 0.17 = 0.102$ (10.2% of total variance) is between groups.
- If the random intercept explains most of that between-group variance, the conditional R^2 should be closer to $0.40 + 0.10 = 0.50$, not 0.43.

A large ICC implies that the grouping variable explains a substantial portion of the residual variance.

This should translate into a noticeable jump from marginal to conditional R^2 . If the jump is very small, it suggests: 1. The between-group variance is not being fully explained by the random intercept (unlikely for a random intercept model),

2. The ICC and R^2 are being calculated on different scales or definitions, or
3. There is a mismatch in the model specification used for each calculation.

Key point:

- Marginal vs. conditional R^2 compares total variance explained by fixed vs. fixed + random effects.
- ICC looks only at the residual variance and asks how much of it is between groups.
- A high ICC but small R^2 difference usually means the total residual variance is already very small, so even strong clustering changes the total R^2 only slightly.

1.4 reporting measures of model fit

prompt: OK, my interest is in describing how much variation and prediction is explained by geo-markets after controlling for other factors (modeled as fixed effects). does it make sense to report Marginal R^2 , Conditional R^2 and ICC?

Yes — reporting **Marginal R^2** , **Conditional R^2** , and **ICC** together makes perfect sense for your goal, because each one answers a slightly different part of your question:

1. Marginal R^2

- **What it tells you:**

The proportion of variance explained **only** by your fixed effects (all covariates, including state FE).

- **Why it matters for your question:**

This is the “baseline” predictive power *before* you give geomarkets any role in the model.

2. Conditional R^2

- **What it tells you:**

The proportion of variance explained by **fixed effects + geomarket random intercepts**.

- **Why it matters:**

Shows the **total** predictive power when you let geomarkets explain whatever systematic variation is left over after the fixed effects.

3. Difference between Conditional and Marginal R^2

- **Interpretation for your question:**

This difference is the *incremental gain in explained variance* attributable to geomarkets.

- Example: Marginal $R^2 = 0.618$, Conditional $R^2 = 0.652 \rightarrow +0.034$ means geomarkets account for ~3.4 percentage points of total variance, on top of the fixed effects.
-

4. ICC (Intra-Class Correlation)

- **What it tells you:**

The proportion of **residual variance** (variance *not* explained by the fixed effects) that is attributable to differences between geomarkets.

- **Why it matters:**

This gives a **relative** measure of how important geomarket clustering is among the unexplained variation — even if the total residual variance is small.

Why all three are worth reporting

- **Marginal R^2** \rightarrow How well the fixed effects perform on their own.
- **Conditional R^2** \rightarrow How well the model performs when geomarkets are allowed to explain residual differences.

- **Difference** → The *absolute* boost in explanatory power from geomarkets.
- **ICC** → The *relative* share of the remaining variance that is due to geomarkets.

Together, they let you say both:

1. “Geomarkets improve prediction by X percentage points beyond fixed effects” (absolute gain)
2. “Geomarkets account for Y% of the remaining unexplained variation” (relative importance in clustering).

other statistics to consider reporting

based on prompt I asked chatGPT

Given your broad goal — **describing how much variation and prediction is explained by geomarkets after controlling for other factors** — Marginal R^2 , Conditional R^2 , and ICC already address the core “variance explained” question.

However, you could strengthen your results by adding **complementary fit and predictive accuracy measures** that answer different but related questions.

1.5 1. Likelihood-based measures

These come directly from the model likelihood and allow comparison of model fit:

- **AIC (Akaike Information Criterion)** and **BIC (Bayesian Information Criterion)**
 - **Why:** AIC favors predictive accuracy, BIC favors parsimony.
 - **Use:** Compare your state+geomarket model to a state-only model. If AIC/BIC drop when you add geomarkets, it’s evidence they improve fit.

1.6 2. Pseudo- R^2 variants

You already use Marginal/Conditional R^2 , but you could also report:

- **McFadden’s R^2** (from `pscl::pR2()`)
 - **Why:** Common in logistic regression literature, easier to benchmark against other studies that don’t use mixed models.
 - **Tjur’s R^2** (already in `performance::r2()` for GLMs)
 - **Why:** Intuitive “mean predicted probability difference” between outcomes.
-

1.7 3. Predictive accuracy metrics

These measure how well the model predicts visits, not just how much variance it explains:

- **AUC (Area Under the ROC Curve)**
 - **Why:** Measures discrimination — ability to separate visited vs. non-visited schools.
 - A higher AUC for the geomarket model compared to the state-only model shows predictive gain.
 - **Brier Score**
 - **Why:** Measures calibration — how close predicted probabilities are to actual outcomes.
 - Lower is better; improvement after adding geomarkets means better-calibrated predictions.
-

1.8 4. Variance partitioning

- **Why:** Break down variance into fixed effects, between-geomarket variance, and residual variance (using `performance::variance_decomposition()`).
- **Use:** Lets you say “X% of total variance is explained by fixed effects, Y% by geomarket, Z% is residual.”

1.9 5. Effect size summaries

- **Median odds ratio (MOR)** for geomarket random effect
 - **Why:** Converts random intercept variance into an odds ratio scale — highly interpretable.
 - **Interpretation:** “Moving from a low-visit to a high-visit geomarket (all else equal) changes the odds of a visit by a factor of MOR.”
-

Why this helps

By combining: - **Variance explained metrics** (Marginal R^2 , Conditional R^2 , ICC)

- **Fit criteria** (AIC/BIC)

- **Prediction accuracy** (AUC, Brier)

- **Interpretability metrics** (MOR)

...you give readers **multiple complementary perspectives**: 1. How much variance is explained (R^2 family, ICC)

2. Whether model fit improves with geomarkets (AIC/BIC)

3. Whether prediction actually improves (AUC/Brier)

4. How large the geomarket effect is on the outcome scale (MOR)

2 References

Jaquette, Ozan, Karina G. Salazar, and Patricia Martin. 2022. *The Student List Business: Primer and Market Dynamics*. The Institute for College Access and Success. https://ticas.org/wp-content/uploads/2022/09/The-Student-List-Business_-_Primer-and-Market-Dynamics.pdf.