Residency Exams

```r
library(lme4)
```

```
Loading required package: Matrix
```

```
Warning in check_dep_version(): ABI version mismatch:
lme4 was built with Matrix ABI version 2
Current Matrix ABI version is 1
Please re-install lme4 from source or restore original 'Matrix' package
```

```r
library(knitr)
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
```

```
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x tidyr::expand() masks Matrix::expand()
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
x tidyr::pack()   masks Matrix::pack()
x tidyr::unpack() masks Matrix::unpack()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
library(glmmTMB)
```

```
Warning in checkMatrixPackageVersion(getOption("TMB.check.Matrix", TRUE)): Package version in
TMB was built with Matrix ABI version 2
Current Matrix ABI version is 1
Please re-install 'TMB' from source using install.packages('TMB', type = 'source') or ask CRA
```

```r
library(broom)
library(emmeans)
```

```
Welcome to emmeans.
Caution: You lose important information if you filter this package's results.
See '? untidy'
```

Background & Motivation

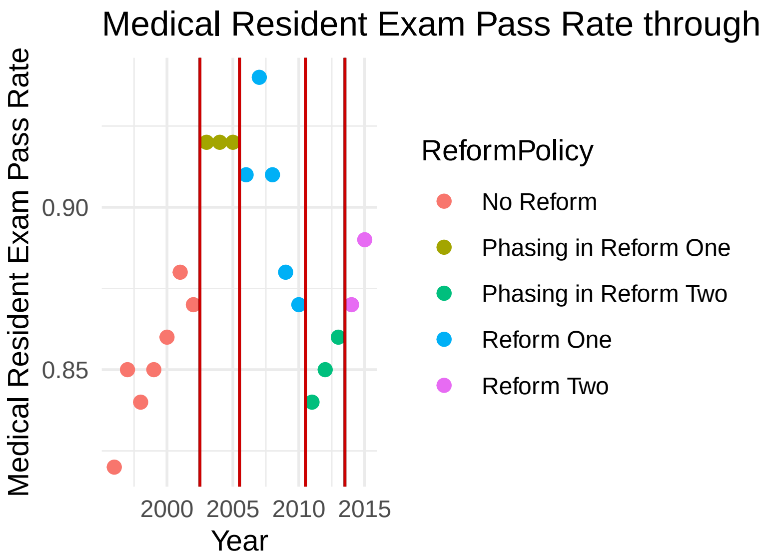
Prior to practicing full-time medicine, internal medicine students must complete a three year training known as a medical residency. At the end of their medical residency, they are required to pass an examination to obtain their MD. These residency programs are known for their extensive work weeks and rigorous workload.

In 2003 and 2011, two reform policies were passed to change the structure of internal medicine residency. The 2003 reform, passed on July 1, sought to limit the shift length and number of hours worked each week by capping it at 30 hours and 80 hours, respectively. The 2011 reform, also passed on July 1, placed stricter limits on students' shift length, capping it at 16 hours for interns and 28 hours for resident students.

These reforms were passed with the goal of improving patient care by decreasing the stress placed on internal medicine residents. In other words, with more time away from the hospital to rest and rejuvenate, students will be more successful at their jobs when they are on their shift. However, limiting the number of hours medical residents can work each week brings about several concerns regarding their performance of the examination at the end of their residency. While some believed a work time limit granted them more time to study, others thought that less hands on work would result in decreased pass rates.

These concerns bring about out motivating question: is there empirical evidence of an association between the reforms and the rate at which the medical residents passed the exam?

Data Manipulation and Exploratory Data Analysis



Model Implementation Details

Model 1 - Binomial Mixture

Rationale

We reasonably assumed that exam pass rates contain unobserved year-to-year variation, such as differences in exam difficulty or cohort quality. To capture this extra source of randomness, we fit a binomial mixture model by including a random intercept for each year. This approach allowed us to separate systematic effects of reform policies from idiosyncratic annual noise.

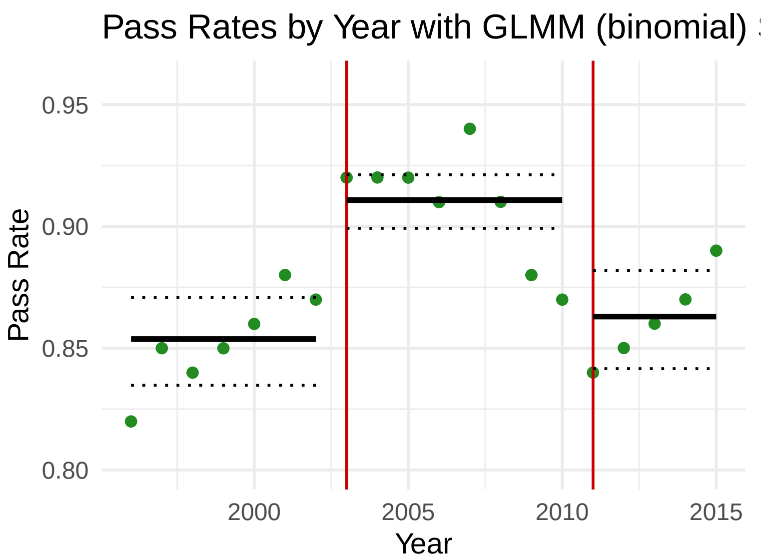
Implementation Details

We fit a generalized linear mixed model (GLMM) with a logit link, specifying pass/fail outcomes as the response and reform time period as the fixed effect, with year as a random effect:

$$\text{Pass}_{iy} \sim \text{Binomial}(n_{iy}, p_{iy}), \quad \text{logit}(p_{iy}) = \alpha + \beta \cdot \text{timeperiod}_{iy} + u_y, \quad u_y \sim N(0, \sigma^2).$$

Result The random effect variance was small (0.037), suggesting year effects contributed little beyond reform periods. Coefficients indicated statistically significant differences: time period 1 and time period 3 both had lower pass rates compared to time period 2 ($p < 0.001$).

```
                 Estimate Std. Error   z value      Pr(>|z|)
(Intercept)     2.3234138 0.06885337 33.744374 1.292817e-249
timeperiodtp1 -0.5591105 0.10041764 -5.567851  2.579000e-08
timeperiodtp3 -0.4831369 0.11034058 -4.378597  1.194459e-05
```



The random effect variance was small (0.037), suggesting year effects contributed little beyond reform periods. Coefficients indicated statistically significant differences: time period 1 and time period 3 both had lower pass rates compared to time period 2 ($p < 0.001$).

Model 2 - Beta-binomial

Rationale

The binomial assumption may underestimate variability in exam outcomes because pass probabilities likely vary within each period. To allow for overdispersion, we modeled the yearly pass probabilities as Beta-distributed. This yields a hierarchical structure where exam outcomes are drawn from a binomial conditional on the latent Beta-distributed rate.

Implementation Details

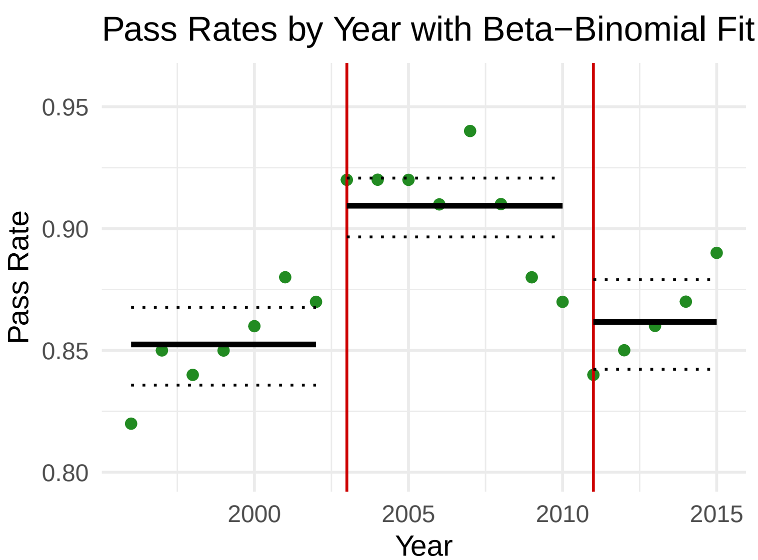
We fit a beta-binomial regression with time period as the predictor using the glmmTMB package with a logit link. The model structure was:

$$\pi_y \sim \text{Beta}(\alpha, \beta), \quad \text{Pass}_y \sim \text{Binomial}(n_y, \pi_y).$$

Here, dispersion was estimated directly from the data, capturing unmodeled heterogeneity.

Result

```
$cond
                Estimate Std. Error   z value      Pr(>|z|)
(Intercept)    2.3061300 0.07465315 30.891263 1.565055e-209
timeperiodtp1 -0.5521503 0.09862549 -5.598454  2.162715e-08
timeperiodtp3 -0.4769413 0.10817339 -4.409045  1.038276e-05

$zi
NULL

$disp
NULL
```



The estimated dispersion parameter was 280, confirming overdispersion relative to a pure binomial. Both time period 1 and time period 3 had significantly lower pass rates than time period 2 ($p < 0.001$).

Model 3 - Beta-binomial on subset

Rationale

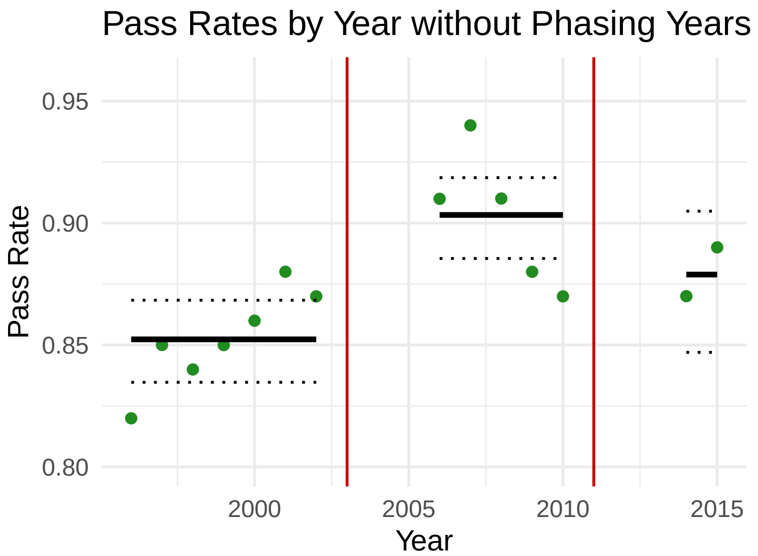
Because policy reforms were phased in gradually, exam cohorts in transition years likely had mixed exposure. To avoid contamination, we fit the beta-binomial model on a restricted dataset excluding these phase-in years. This design isolates the effect of reforms once they were fully implemented.

Implementation Details

We restricted the dataset to years 1996–2002, 2006–2010, and 2014–2015. The same beta-binomial specification was used, with time period as the predictor and a logit link for the mean pass probability.

Result

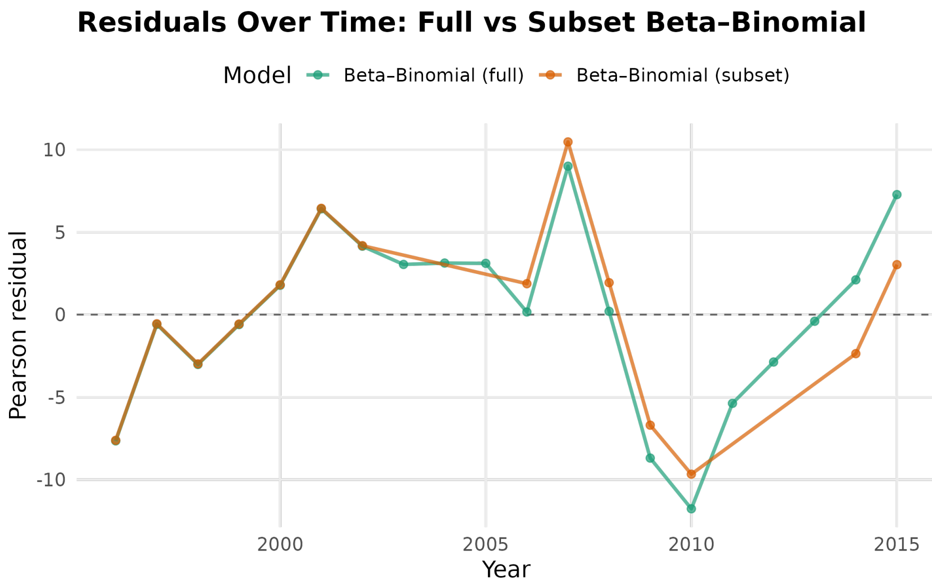
```
$cond
                Estimate Std. Error   z value      Pr(>|z|)
(Intercept)    2.2345867 0.09658272 23.136507 1.987766e-118
timeperiodtp1 -0.4817205 0.11798326 -4.082956  4.446642e-05
timeperiodtp3 -0.2528107 0.16833072 -1.501869  1.331310e-01

$zi
NULL

$disp
NULL
```

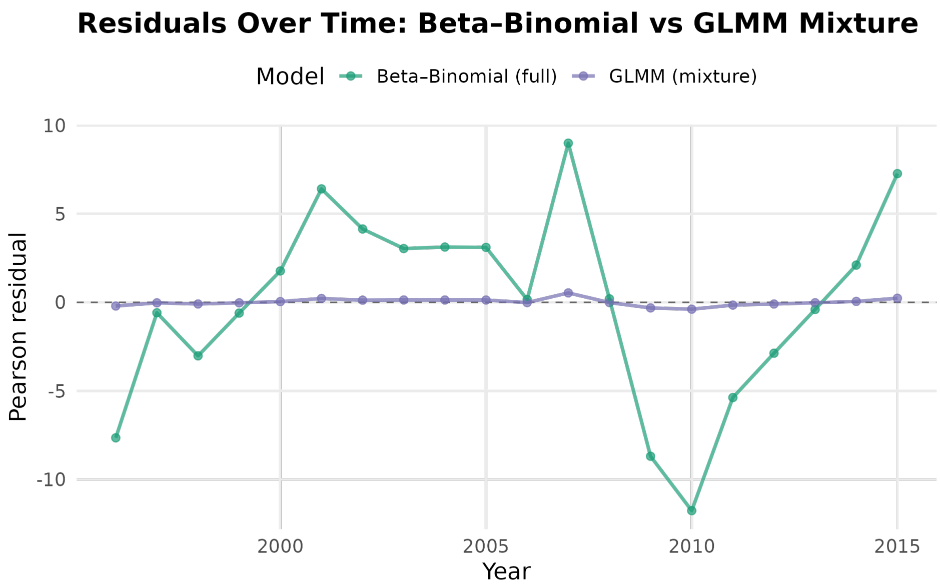


The dispersion parameter was estimated at 251, again supporting overdispersion. Results showed that time period 1 had significantly lower pass rates than time period 2 ($p < 0.001$). However, time period 3 was no longer significantly different from time period 2 ($p = 0.133$).

Model Evaluation

To check goodness of fit, we plotted Pearson residuals over time. We appear to minimize residuals over time in the mixture model, as seen below. Putting this together, we concluded that the full beta-binomial mixture model is the best fit.





Shortcomings

An issue came with finding good evaluation metrics that compare across models. We initially attempted AIC across all models, but soon learned this was not comparable for any of the models because the mixture modeling package and beta binomial packages use different formulations of AIC. In addition, AIC is dependent on the likelihood and because of this comparing AIC across a full or subset model since they utilize variable number of observations.

Looking ahead, a next step would be to use Bayesian models, which would allow us to incorporate auxiliary information into the analysis.

Conclusion

We concluded that the binomial mixture model is the best fit. With that model, we found that the first reform increased the odds of passing, while the second reform was associated with a decline in performance. Both of these were confirmed by significant coefficients.

Citations

https://www.bmj.com/content/366/bmj.l4134#:~:text=The%20first%20reform%2C%20in%202003,the%20peri
https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/1672284