An Investigation of Willingness to Provide Genetic Data to Public Schools

STA2201 - Final Research Project

Lindsay Katz

April 18, 2022

Abstract

An interest in individual-level genetic data has led to many advancements in the techniques which may be used to analyze it. Polygenic risk scores (PGSs) are one example, which provide an estimate of individual relative risk for certain disease outcomes. Recently, there has been discussion regarding how these personalized genetic scores may be used in the context of school, to provide "precision education" (Brookman-Byrne, 2021). It is of interest to both researchers and policymakers alike to understand the public opinion on this matter, given its highly controversial nature. Inspired by the work of Zhang et al. (2021) which employs frequentist statistical tools to explore the topic at hand, this analysis in contrast operates within a Bayesian framework. Using replication data from the Harvard Dataverse, demographic factors which effect respondent willingness to provide genetic data to public schools is investigated. Key demographic factors such as age, education, gender, religion, ethnicity, and political alignment are considered in two logistic regression models, one fit with entirely fixed effects, and another fit with a non-nested hierarchical structure. After performing thorough model validation, the results illustrate notable differences in willingness to provide across certain demographic groups. This analysis and subsequent findings provide valuable insight into the demographic differences in opinion toward uses of genetic data beyond the medical field.

1 Introduction

In today's world, the collection and analysis of individual-level data in any domain is of greater interest than ever before. The field of genetics is no exception; genetic data science first emerged in the 1990s with the launch of the Human Genome Project (NHGRI, n.d.). Since this project, there have been major advancements in the generation of genetic data as well as techniques to analyze it, both within the biomedical field and beyond. Along with these advancements have come many ethical questions about how this data should be used in various domains.

One of these data-driven techniques currently being studied is the polygenic risk score (PGS). A PGS can provide an estimate of individual relative risk for certain disease outcomes, including but not limited to coronary heart disease, schizophrenia, and diabetes (Int Common Dis Alliance *et al.*, 2021; Wray *et al.*, 2021). The implementation and reliability of PGSs is a prominent area of research in the field of genetics which is currently being investigated through clinical trials, however it is not yet approved for implementation in standard medical practice (Wray *et al.*, 2021). That being said, PGSs have been implemented by some private healthcare providers and direct-to-consumer genetic testing companies such as 23andMe (Lewis & Vassos, 2020).

Though seemingly premature given the state of the clinical trials, there has been recent discussion of how genetic scoring might be used in schools through the implementation of "precision education"- a highly controversial and ethically sensitive application. In general, precision education entails using individual students' genetic data to classify those who are genetically at risk of developing learning disabilities, and using this information to provide individualized accommodation plans in an effort to improve long-term educational attainment (Hart, 2016). Proponents of precision education assert that this could have positive societal outcomes, while others discuss the numerous risks and ethical concerns associated with precision education (Sabatello *et al.*, 2021). To get a sense of societal views on this contentious matter, Zhang *et al.* (2021) investigated public attitudes toward the use of genetic risk scoring in both medical and non-medical contexts using the first nationally-representative survey of this nature. Using replication data from this work, this analysis will serve to investigate the impact of demographic factors on individual willingness to provide genetic data to public schools.

2 Data

As mentioned, the data set which will be used in this analysis is replication data from Zhang *et al.*'s 2021 article, which was made publicly available through the Harvard Dataverse. The original survey data was collected by the University of Chicago's AmeriSpeak panel, from their Nonpartisan and Objective Research Organization (NORC). The AmeriSpeak panel is unique in that it prioritizes national representativeness of their survey sample, through improved representation of rural households who are characterized as difficult to reach. This is further enhanced through a focus on the representation of Hispanic ethnicity, age, race, gender, and education level in their sample (Zhang *et al.*, 2021).

The dataset obtained contains demographic information for 1457 respondents, as well as their responses to a number of questions related to opinions on the use and distribution of genetic data. All demographic variables will be used in this analysis, however only one binary response variable, willingness to provide genetic information to public schools, will be used. In particular, the demographic variables include age group, education level, gender, race/ethnicity, religious belief, and political party alignment. All of these variables are categorical in nature. There are two levels for gender (male/female), five age categories spanning 18-60+, six race/ethnicity groups, six education levels, as well as five levels each for political alignment and religious belief.

Prior to performing any statistical analysis, data pre-processing and exploration are valuable and necessary practices to employ. In doing so, the presence of 21 respondents with missing data were identified. It is important to determine the nature of these incomplete cases, that is, whether they are missing at random or not, in order to correctly deal with them for subsequent analysis. Through exploratory data analysis (EDA), the distribution of observations was assessed by computing proportions of respondents in each group, for each variable. Subsequently, another check of these proportions with incomplete cases removed illustrated that this removal did not lead to concerning changes in the data distribution. Further, there did not appear to be any patterns of missingness in terms of demographic groups. In summary, EDA illustrated that the missing values were in fact missing at random, and for this reason, have been dropped. As for pre-processing, the outcome variable which was originally coded as TRUE/FALSE was re-coded to take on values 1 and 0. For simplicity, the "Associates Degree" and "Some college, no degree" levels of education were combined into one called "< Bachelors" to group individuals who have some higher education but have not completed a Bachelor's degree.

Approximately 21.3% of respondents indicated willingness to provide genetic data to schools. The respondents are 51.9% female and 48.1% male. The distributions of respondents in terms of race/ethnicity, religious belief, and political alignment are summarized in the tables below. Evidently, the majority of respondents are white non-Hispanic (~66%). In regard to religion, most respondents are either Christian (Protestant or Catholic), or do not identify with a religious group. As for political alignment, the majority of respondents are democratic (~38%), independent (~25%), or republican (~25%). As for the distribution of respondents in terms of gender, age group, and education level, the majority of respondents fall into the "< Bachelors" education level, while the minority of respondents are in the "< HS diploma" group. With the exception of "< HS diploma", the majority of respondents in each education group are in the 60+ age category. In terms of gender, the division of respondents is generally consistent across age and education groupings, except for the "< Bachelors" group in which there are 42 more female respondents than male respondents.

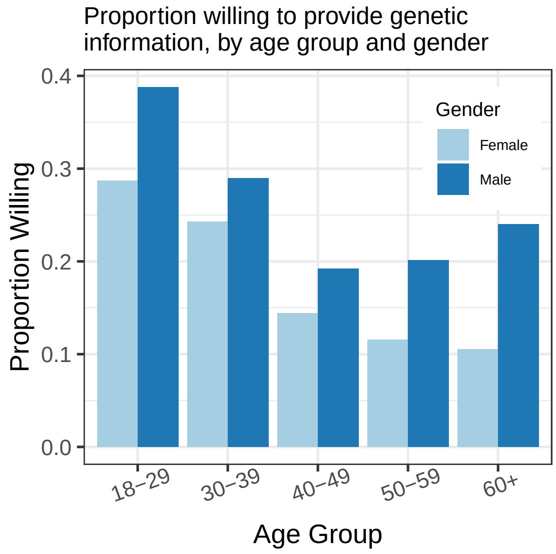
Race/Ethnicity	Prop.
2+, non-Hispanic	0.0404
Asian, non-Hispanic	0.0320
Black, non-Hispanic	0.0975
Hispanic	0.1546
Other, non-Hispanic	0.0146
White, non-Hispanic	0.6609

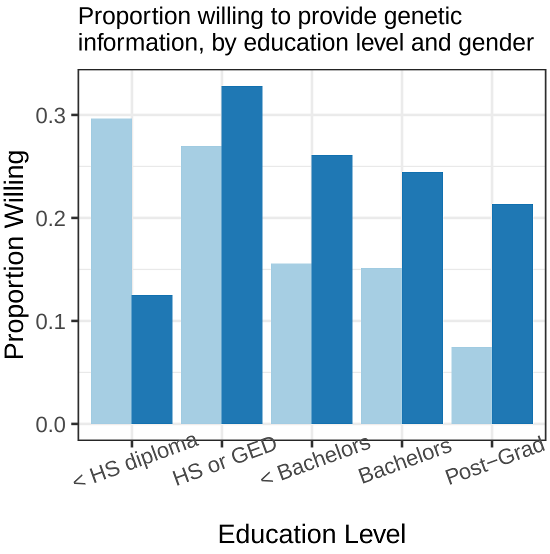
Religious Belief	Prop.
Catholic	0.2047
Jewish	0.0223
None	0.2820
Other	0.2347
Protestant	0.2563

Political Belief	Prop.
Democrat	0.3823
Independent	0.2535
No preference	0.0822
Other	0.0341
Republican	0.2479

Table 1: Proportions of Respondents by Race/Ethnicity, Religious Belief, and Political Belief

The final step of EDA was to visually investigate the relationship between the independent variables as they relate to the research question of interest. As such, the plots below illustrate the proportion of respondents who indicated willingness to provide genetic data to schools, by age group (left), and education level (right). Looking at the plot on the left, it is apparent that for every age group, men have a higher proportion willing to give genetic data to schools than women. The same is true in the right plot for all education levels, with the exception of "< HS diploma". There is a clear decrease in proportion of women willing as age group increases. For male respondents, the proportion of willing respondents decreases from 18-29 until 40-49, and slightly increases as age progresses to 60+. Also, the plot on the right shows that the group with the highest proportion of willingness is the "HS or GED" group, while the "post-grad" group has the smallest proportion of willing respondents. Though not shown, similar plots were created for the remaining predictors of interest which displayed seemingly notable differences in proportions of willingness across various groups.





3 Methods

To answer the research question of interest, two Bayesian models will be fit in Stan. As the dependent variable for both models is the `provide` outcome, logistic regression will be used to appropriately model the binary nature of this response. Further, since the individual respondent's one-time answer to this survey question is being modeled, the Bernoulli distribution is the appropriate probability distribution in this case. The first model is a simple fixed-effects model, which includes all the demographic variables as predictors in the model. That is, age, education, gender, religion, political belief, and race are all included. Standard normal priors will be placed on all the βs, including the intercept β_0. The model formulation and associated notation are defined below.

$$
\begin{aligned}
y_i \mid \pi_i &\sim \text{Bernoulli}(\pi_i) \\
\pi_i &= \text{logit}^{-1}(\beta_0 + x_i^T \beta) \\
\beta_0, \beta &\sim N(0, 1)
\end{aligned}
$$

where:

- $y_i = 1$ if the respondent indicates willingness to provide genetic information, and 0 otherwise
- x_i is a vector of demographic variables for respondent i, these variables being religious group, political group, gender, race/ethnicity, age group, and education level.

The second model which will be used to answer the research question of interest is a Bayesian non-nested hierarchical model. This model will include fixed effects for gender, race, religion, and political belief, and a hierarchical structure will be placed on age group and education level membership. This hierarchical structure will allow the model to estimate unique predictive effects for respondents after accounting for structured patterns related to education and age. Due to the added complexity in this model, priors must be placed on the variance parameters in addition to the βs. Also, recall from the plot of age group versus proportion willing in the previous section that there appeared to be an ordered trend in proportion willing as age group increased. For this reason, a random walk will be placed on the prior for the age-specific intercept, allowing for the use of information from the previous age group $a - 1$ to estimate the effect of age group a. A standard normal prior will be placed on the first age group, as it is defined slightly differently. The full specification and notation of this hierarchical model are clearly defined below.

$$
\begin{aligned}
y_i \mid \pi_i &\sim \text{Bernoulli}(\pi_i) \\
\pi_i &= \text{logit}^{-1}(\beta_0 + x_i^T \beta + \alpha^{age}_{a[i]} + \alpha^{edu}_{e[i]}) \\
\alpha^{age}_a &\sim N(\alpha^{age}_{a-1}, \sigma^2_{age}), \quad \text{for } a = 2, \dots, 5 \\
\alpha^{edu}_e &\sim N(0, \sigma^2_{edu}), \quad \text{for } e = 1, \dots, 5 \\
\sigma^2_{age}, \sigma^2_{edu} &\sim N^+(0, 1) \\
\beta_0, \beta &\sim N(0, 1)
\end{aligned}
$$

where:

- $y_i = 1$ if the respondent indicates willingness to provide genetic information, and 0 otherwise
- x_i is a vector of demographic variables for respondent i, these variables being religious group, political group, gender, and race/ethnicity
- α^{age}_a is the age-specific intercept, and a is an index for age group
- α^{edu}_e is the education-specific intercept, and e is an index for education level

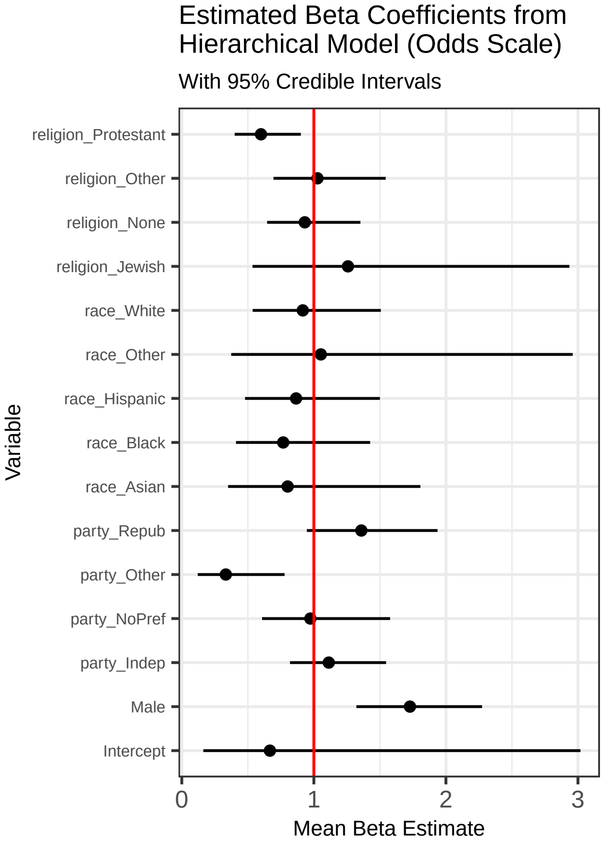
Note that since the predictors going into both models are un-ordered and categorical, a matrix of dummy (indicator) variables for all but one level of each variable will be passed into the model, along with a column of 1's for the intercept. One level will be dropped from each variable to avoid multi-collinearity, and the level dropped will act as the reference category for that variable.

In order to validate each of the models, a visual inspection of sampling behavior will be performed using traceplots of all the parameters. If there are no issues in terms of sampling, chain mixing, and convergence, the traceplot should have good overlap between chains, relative stability over iterations, and some movement to show that there is not too much correlation between the samples. Pairs plots are another validation tool which will be employed in this analysis, which provide univariate histograms and bivarate scatter plots for specified parameters (Gabry & Modrák, 2022). Pairs plots facilitate the identification of issues in convergence, and illustrate whether the sampler has access to the entire parameter space from which to sample.

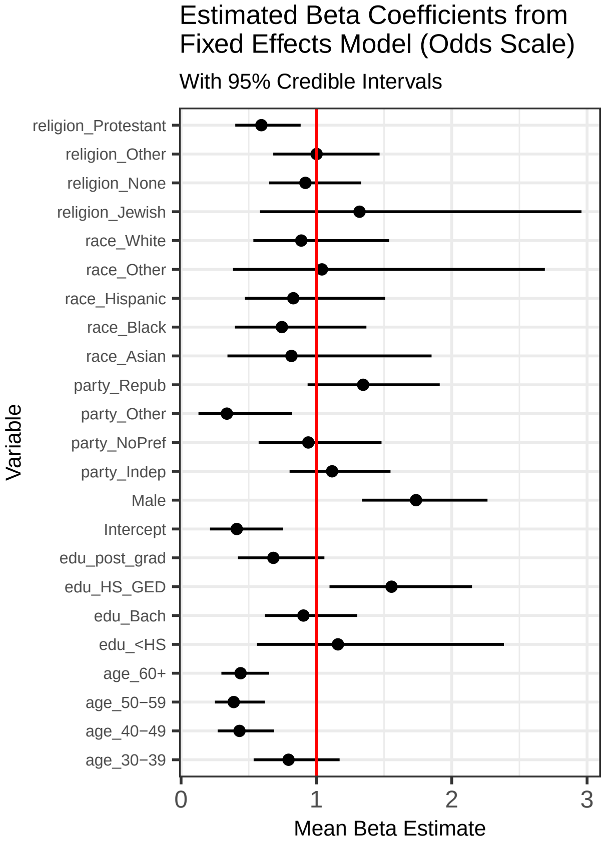
To compare the two models, the difference in expected log-predictive density (ELPD) will be computed using Pareto smoothed leave-one-out (LOO) cross-validation to measure how well each model is performing in terms of out-of-sample prediction. A larger ELPD value implies better out-of-sample predictive performance. The pointwise log-likelihood values will be computed within the generated quantities blocks of the Stan models. Also, a binned residual plot will be produced for each model to assess the validity of the underlying model assumptions.

4 Results

The resulting coefficients of these two models, as well as their corresponding 95% credible interval (CI) bounds are illustrated in the plots below, on the odds scale. Looking at the CI bounds visually this way helps to get a sense of those coefficients with wide CI bounds, and those with CIs that cover 1. Looking at the two plots, it is evident that the CI bounds for the `race` variables are quite similar both in range and magnitude, and all include 1. Also, the CI for the `Male` variable is relatively narrow, and larger than 1 in both models. Another similarity between the model coefficient estimates can be seen by looking at the political belief variables. In both plots, we can see that the CI bounds and positions along the x-axis are similar, and the only variable that doesn't cover 1 is `party_Other` which also has similar CI bound and mean values. In both models, the only religion variable that has a CI which doesn't include 1 is `religion_Protestant`. Also, the intercept estimate for the fixed effects model has much narrower CI bounds than that of the hierarchical model. As for the fixed effects that were not included in the hierarchical model (age and education), the plot shows that the estimates for the coefficients of age groups 40-49, 50-59, and 60+ are all less than 1. The only education level coefficient which does not include 1 is the `HS or GED` group, which has CI bounds larger than 1.

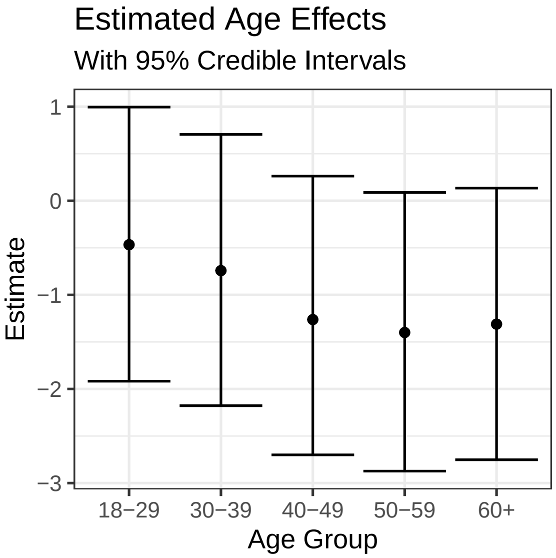

Estimated Beta Coefficients from Hierarchical Model (Odds Scale)

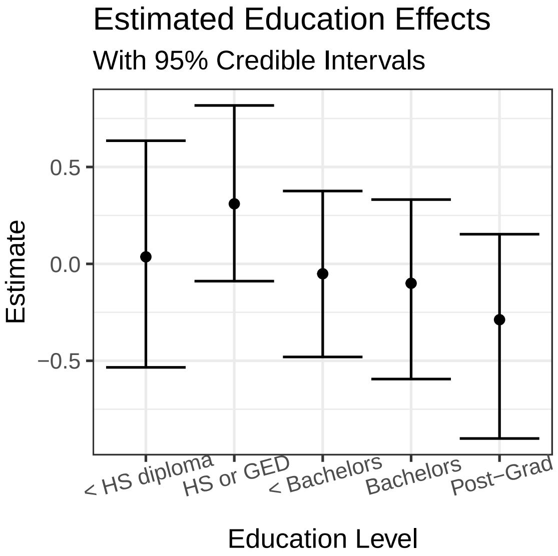
With 95% Credible Intervals


Estimated Beta Coefficients from Fixed Effects Model (Odds Scale)

With 95% Credible Intervals

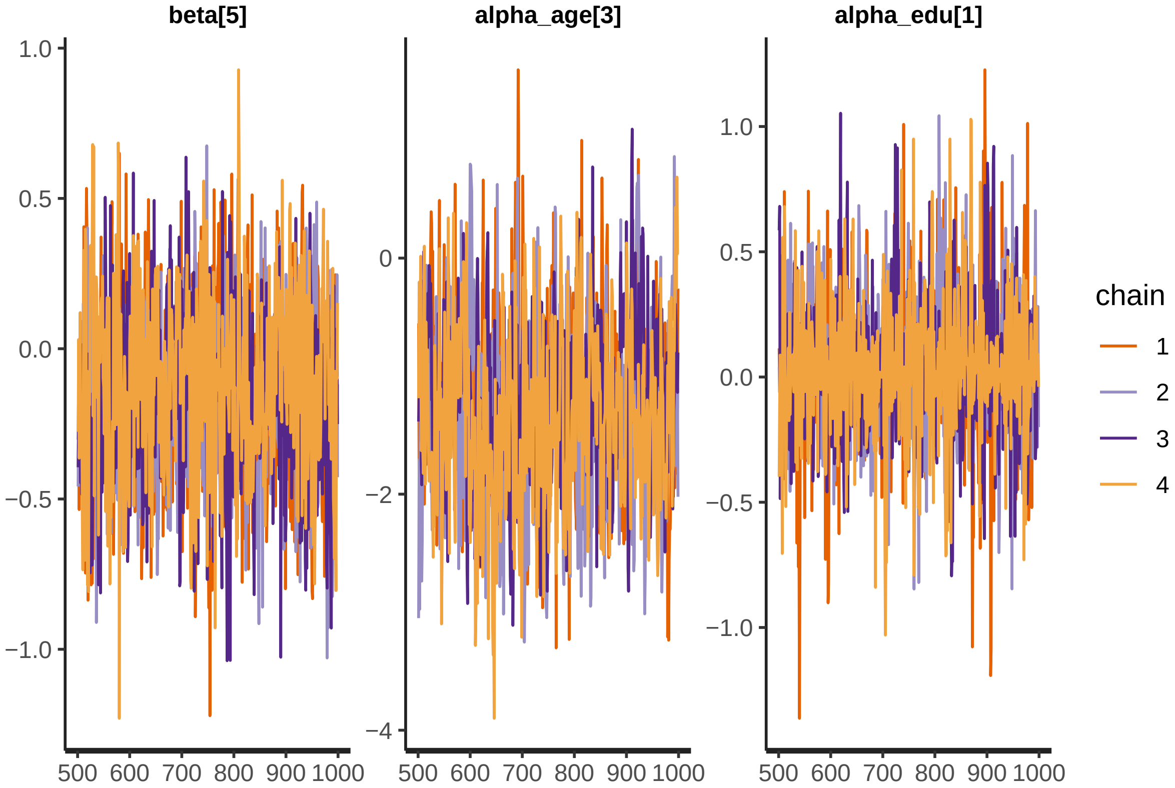
The estimates and 95% CIs for the age and education specific intercepts from the hierarchical model are illustrated in the two plots below. It is apparent that the credible interval bounds are quite wide across the education effects, and are even wider across the age effects. The CI bounds across all levels in both plots include 0. Looking at the plot on the left, it can be seen that all of the age-specific intercepts have mean estimates less than 0. As for the education specific intercepts, the plot on the right shows that the `< HS diploma` and `HS or GED` groups have mean estimates above 0, while the rest of the education levels have mean estimates below 0. These estimated effect plots correspond well to the bar plots shown in the section 2, created through exploratory data analysis.



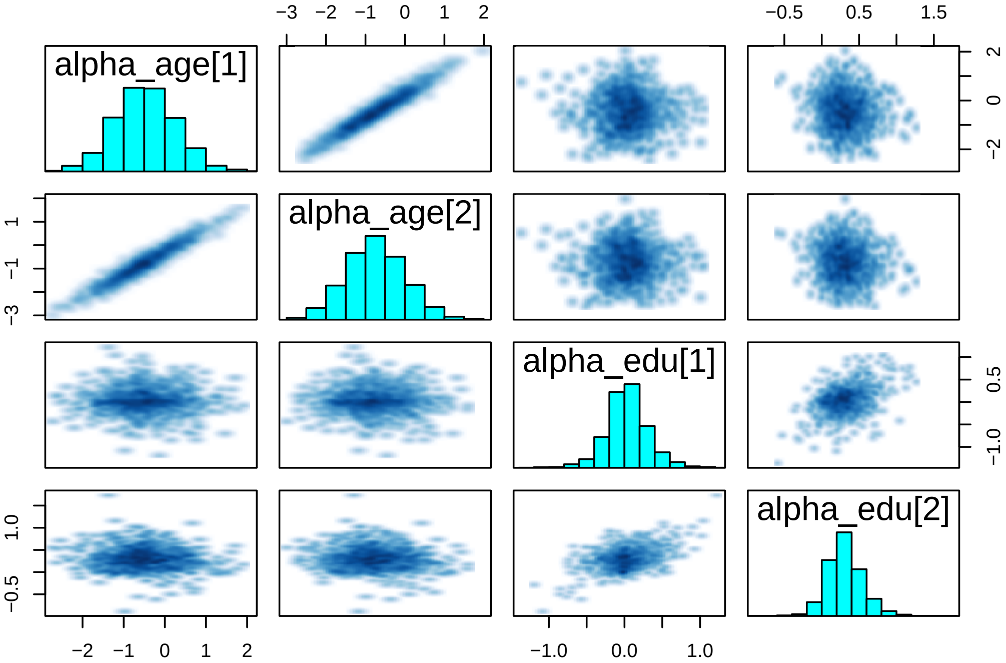


In terms of coefficient interpretation, since logistic regression was performed, the coefficient estimates must be interpreted on the log-odds scale, or alternatively they can be interpreted on the odds scale once exponentiated. For simplicity, coefficient interpretations will be provided based on the hierarchical model results, and one interpretation will be provided for each categorical variable. First, the exponentiated `Male` coefficient can be interpreted as follows: The odds of a male respondent who is Catholic, Democratic and of mixed ethnicity responding that they are willing to given their genetic data to a public school is ≈ 1.73 times that of a female respondent who is also Catholic, Democratic and of mixed ethnicity. Next, the odds of a Hispanic respondent who is female, Democratic, and Catholic indicating willingness to provide is ≈ 0.87 times that of a non-Hispanic mixed respondent (who is also female, Democratic and Catholic). In terms of political belief, the odds of a Republican respondent who is female, Catholic, and non-Hispanic mixed ethnicity indicating that they are willing to provide their genetic information to a public school is ≈ 1.36 times that of a Democratic respondent (who is also female, Catholic, and non-Hispanic mixed ethnicity). Finally, the odds of a Protestant respondent who is female, of non-Hispanic mixed ethnicity, and Democratic, being willing to provide is ≈ 0.60 times that of a Catholic respondent who is also female, non-Hispanic mixed, and Democratic.

As mentioned in the methodology section, visually inspecting the sampling behavior of the models and their resulting estimates can be done using traceplots, and pairs plots. First, note that these plots were inspected for all the estimates of both models, however for simplicity, only a small number will be shown below. For both the fixed effects model and the hierarchical model, the traceplots of all estimates showed no signs of issues in terms of mixing or convergence. The traceplots below belong to the hierarchical model, and illustrate that the chains seem to be mixing quite well for the β, α_a^{age} and α_e^{edu} parameters. All the traceplots showed good mixing and overlap of chains, with some movement across iterations but relative stability overall.

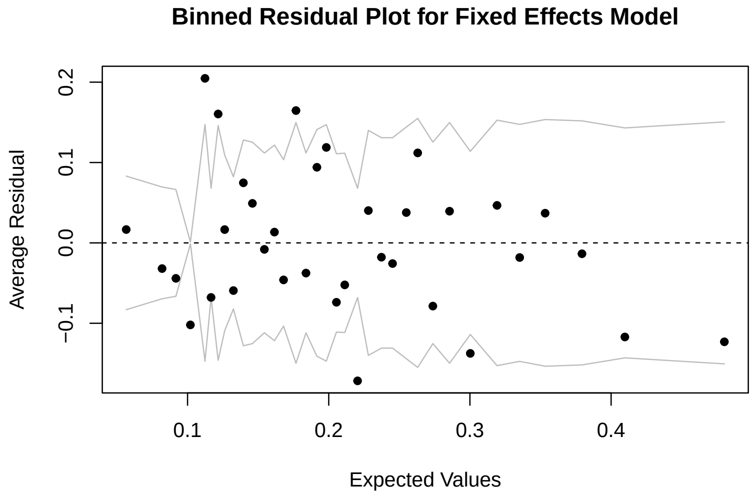


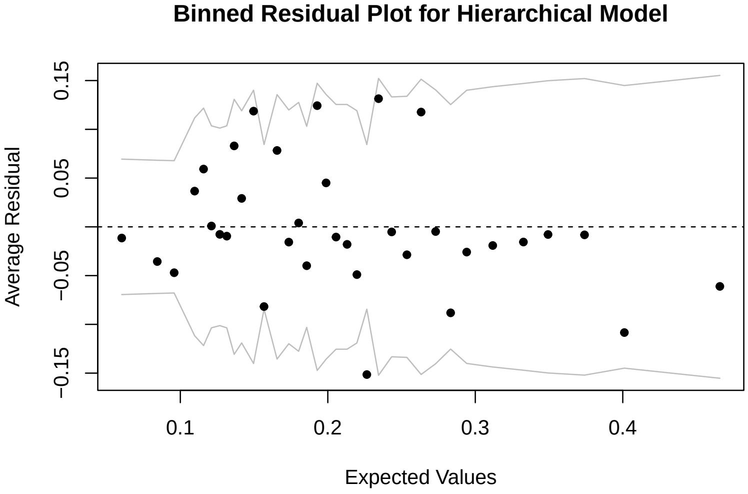
For the fixed effects model, none of the pairs plots provided any cause for concern in terms of convergence or access to the entire sampling space. The histograms all displayed a unimodal distribution of observations about some central point, and the bivariate scatterplots did not have a concerning, narrow distribution of points meaning that the sampler had good access to the entire sampling space. As for the hierarchical model, the same can be said for all the estimates' pairs plots except for those of α_a^{age} and α_e^{edu}. Below is a matrix of pairs plots for two α_a^{age} and α_e^{edu} parameters. Though the histograms are not problematic, some of the bivariate scatterplots between the two α_a^{age} parameters have a very narrow, elliptical-shaped space. This is cause for concern, because it implies the sampler does not have access to the entire parameter space to sample from. The narrow, elliptical shape is less pronounced among the other scatterplots, for example that between `alpha_age[2]` and `alpha_edu[1]`, however some of them are still relatively narrow.



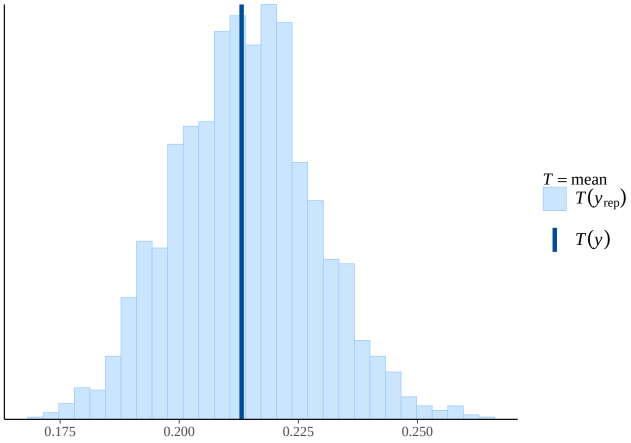
It is also worth noting that the R-hat values for all of the estimates of both models were inspected, and the largest one out of both models was ≈ 1.008. This further indicates that there are no pronounced convergence issues in either model.

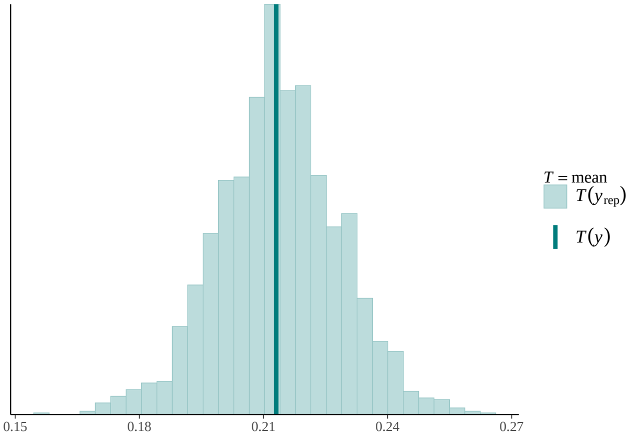
As for the ELPD comparison, the expected predictive accuracy of the models were quite close in value. More specifically, the hierarchical model's ELPD was larger than that of the fixed effects model by 0.7 units. Since the standard error associated with this difference is 1.0, and the difference in ELPD's is less than that, it can be concluded that there is not a strong preference for the hierarchical model over the fixed effects only model. Further, the binned residual plots below provide a useful illustration for the assessment of overall model fit. Looking at the two plots, it is clear that there is no distinct directional relationship or clustering of points, which is an indication that the models are fitting the data well. Further, 1 point is outside the ±2 standard error bounds for the hierarchical model, while 6 are outside those bounds for the fixed effects only model. In terms of binned average residuals, the hierarchical model seems to be doing a slightly better job at fitting the data.




Binned Residual Plot for Hierarchical Model

For a more granular posterior predictive check, the `ppc_stat_grouped` function was used to assess the distribution of the mean outcome from the replicated data in comparison to the mean outcome from the original data, for each level of education, and each age group. This was performed for both models. In looking at the resulting histograms for both models, the distributions were well centered about the true mean outcome for both age groups, and education levels. For simplicity, and to provide an overall visualization of this posterior predictive check, the ungrouped histograms, created using the same method explained for the `ppc_stat_grouped` function, are shown below. Note that the blue histogram on the left is based on the hierarchical model, while the green histogram on the right corresponds to the fixed effects model. Overall, for both models, the predicted mean proportion of respondents willing to provide genetic data matches that of the observed data well.





5 Discussion

In summary, two Bayesian logistic regression models were fit, one with entirely fixed effects, and the other with a non-nested hierarchical structure. Using a number of model validation tools including trace plots and pairs plots, the fixed-effects model showed no signs of convergence issues or concerning sampling behavior. As for the hierarchical model, the pairs plots for the age and education specific intercept parameters provided some cause for concern as some had bivariate scatterplots with a narrow, elliptical shape, implying insufficient access to the entire sample space. That said, there were no signs of convergence issues, and the binned residual plots showed that both models were an overall good fit for the data. As well, ELPD-LOO comparison did not show strong preference for one model over the other. Further, the posterior predictive plots using replicated data from each model illustrated that both models predicted the mean proportion of respondents willing to provide their genetic data to schools quite well. All of these findings indicated that there is no strong preference for one model over the other, however the concerning pairs plots mentioned associated with the hierarchical model are definitely something to take into consideration when comparing the two models.

Based on the analysis performed and the subsequent findings, it is evident that variation in demographic characteristics are associated with different degrees of willingness to provide genetic data to schools. Note that the following interpretations operate under the assumption that all other demographic variables are equal to the reference category, which are: female, mixed ethnicity, Catholic, and Democratic. Some notable differences in the odds of willingness in both models include being Protestant compared to Catholic which decreases odds of willingness, or being male compared to female which increases odds of willingness. Another interesting difference is that being Republican is associated with a higher odds of willingness to provide than being Democratic, and aligning with some other political party is associated with a lower odds of willingness compared to a Democrat. While the estimated age and education effects from the hierarchical model had quite wide credible interval bounds, the coefficient estimates for age and education in the fixed effects model did not all follow this trend. As shown in the right plot on page 5, being in the HS or GED education level is associated with an increased odds of willingness compared to the < Bachelors level (the reference category for education), all other variables equal to their reference categories (where age group is 18-29). Under the same assumptions, a respondent in post-grad is associated with a lower odds of willingness to provide than their < Bachelors counterpart. As for age, it is evident that all age groups plotted are associated with lower odds of willingness to respond compared to the 18-29 age group (all other variables equal to reference categories). These results are not surprising, based on what was found in the EDA in terms of proportions of respondents willing to provide their data across levels of demographic variables.

When performing an investigative analysis such as this, it is valuable to consider future directions in which this work might be taken. Time permitting, it would be of interest to make use of the data on other survey questions, which relate to the one used in this analysis. The original data set contains responses to questions about willingness to provide genetic information to institutions other than public schools, such as a life insurance provider or a police forensic database. Inspired by the work of Zhang et al. (2021), it would be interesting to fit Bayesian models as done here on these other responses, and compare the resulting estimates. As well, further investigation into the best performing model specification would be valuable, in terms of modifications to the predictors included in the model, the hierarchical structure, and the priors.

6 References

1. Brookman-Byrne, A. (2021, November 15). Precision education. BOLD. Retrieved April 3, 2022, from https://bold.expert/precision-education/

2. Gabry, J., & Modrák, M. (2022). Visual MCMC diagnostics using the bayesplot package. Retrieved April 15, 2022, from https://cran.r-project.org/web/packages/bayesplot/vignettes/visual-mcmc-diagnostics.html

3. Hart, S. A. (2016). Precision education initiative: Moving toward personalized education. Mind, Brain, and Education, 10(4), 209-211.

4. Int Common Dis Alliance, Adeyemo, A., Balaconis, M. K., Darnes, D. R., Ripatti, S., Widen, E., & Zhou, A. (2021). Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. Nature Medicine, 27(11), 1876-1884. https://doi.org/10.1038/s41591-021-01549-6

5. Lewis, C. M., & Vassos, E. (2020). Polygenic risk scores: from research tools to clinical instruments. Genome medicine, 12(1), 1-11.

6. National Human Genome Research Institute. (n.d.). Genomic Data Science Fact Sheet. Retrieved April 14, 2022 from https://www.genome.gov/about-genomics/fact-sheets/Genomic-Data-Science

7. Sabatello, M., Insel, B. J., Corbeil, T., Link, B. G., & Appelbaum, P. S. (2021). The double helix at school: Behavioral genetics, disability, and precision education. Social Science & Medicine, 278, 113924.

8. Wray, N. R., Lin, T., Austin, J., McGrath, J. J., Hickie, I. B., Murray, G. K., & Visscher, P. M. (2021). From basic science to clinical application of polygenic risk scores: a primer. JAMA psychiatry, 78(1), 101-109.

9. Zhang, Simone; Johnson, Rebecca A.; Novembre, John; Freeland, Edward; Conley, Dalton, 2021, "Replication Data for: Public attitudes toward genetic risk scoring in medicine and beyond", https://doi.org/10.7910/DVN/CL5XCF, Harvard Dataverse, V1, UNF:6:3KufPpI/2BYGTabJ367dMQ== [fileUNF]