A Bayesian Hierarchical Approach to Modelling Financial Inclusion in Emerging Economies

Gaurja Newatia

Abstract

Financial Inclusion has a strong positive impact on economic growth and the overall wellbeing of an individual.Using the World Bank Financial Inclusion Database as well as World Development Indicators and Financial Access Survey, we investigate individual level variables and country level variables that impact financial inclusion in Emerging Market Economies; using Bank Account Ownership as as our proxy to measure financial inclusion. Account Ownership is modelled as a binary variable using a Bayesian Hierarchichal Logistic Regression Model. The findings of the report illustrate a significant relationship between demographics and income of an individual in influencing account ownership. We model the variations in country via several macroeconomic and geographical outreach indicators and find differences among emerging markets. We also show that interaction with female and other individual level indicators is an important source to account for financial inclusion in the emerging markets. Ultimately, the model can be used to make better policies that foster financial inclusion so that emerging economies focus not only on economic prosperity but also the overall development of its' population.

Introduction

Access to bank accounts, which is a key indicator of financial inclusion has a strong positive effect on individual's economic and social well being. A number of studies have shown that having a bank account increases savings (Aportela, 1999; Ashraf et al., 2006), consumption (Dupas and Robinson, 2013; Ashraf et al., 2010), productive investment of entrepreneurs (Dupas and Robinson, 2011, 2013), and female empowerment (Ashraf et al., 2010). The aim of this study is to therefore determine individual level factors and the country level indicators that are important to advance financial inclusion in emerging economies.[1]

Data

Description of the Dataset

The Global Financial Inclusion Database 2021 (Global Findex) is the source of individual level data used in the analysis and data for the country level indicators comes from the

[1]Countries selected based on MSCI Emerging Markets Index

International Monetary Fund's Annual Financial Access Survey (FAS)and World Development Indicators (WDI) published by the World Bank. The dependent variable in our case is the variable "account", that =1 if the individual has an account at a financial institution, a mobile money account, or both , = 0 otherwise.

We consider a group of independent variables that the literature identifies as important determinants of bank accounts uptake. For the purpose of our models, we divide age into categories and transform income quintiles into "High Income Group" and "Low/ Middle Income" group if a person falls in >=4 quintile or less than quintile 4 respectively.Education measures the highest completed level of education and is grouped into three categories (1= completion of primary school or less, 2= completion of secondary education or more, 3= completed tertiary education or more). Internet access (1=respondent has access to the Internet,2=if no) is included as a proxy for infrastructure for the delivery of financial services. Female (1=Female, 2=Male) indicates whether the respondent is a woman.

For the purpose of the analysis,We convert the variable age into 8 categories as follows- "Under 18", "19-30", "31-40", "41-50", "51-60", "61-70","71-80" and "Over 80". In addition to controlling for individual-level characteristics, this study also estimates the effect of theoretically relevant country-level variables on financial inclusion. GDP per capita growth(annual %) is the growth rate of gross domestic product (GDP) per capita and Unemployment, total (% of total labor force) (modeled ILO estimate) is the unemployment rate and comes from WDI. Bank branch density denotes the average number of commercial bank branches per 1,000 square kilometres and ATM density denoted average number of ATMs per 1,000 square kilometres serve as a proxy for the availability of financial services. Data for both of these indicators comes from the FAS.

Data Issues and Missing Values

The Global Findex dataset included data on 21 emerging economies, contrary to the 24 in the MSCI Index. Therefore, the analysis was limited to the 21 economies covered by the Global Findex dataset. The variable "educ" contained indicators 4 and 5, and "internet access" included values 3 and 4. Unfortunately, the codebook did not provide explanations for the "educ" values of 4 and 5. For "internet access," the values 3 and 4 were defined as "I don't know" and "refused to answer" respectively in the codebook. Given the lack of clarity regarding the "educ" variable and the ambiguous nature of the responses for "internet access," I chose to exclude rows with these specific indicators. This exercise led to the removal of Egypt from the dataset, that had only 2 rows with the previously mentioned indicators. Finally, 109 rows with missing values in the "age" variable were also removed from the analysis.

Exploratory Data Analysis

Account Ownership Across Countries

In Figure 1, we can see that account ownership varies significantly by countries, with countries such as Korea, Thailand, Poland, and China that exhibit universal ownership while in countries such as Columbia, Indonesia, Mexico and India, there are a large proportion of individuals having no bank accounts.

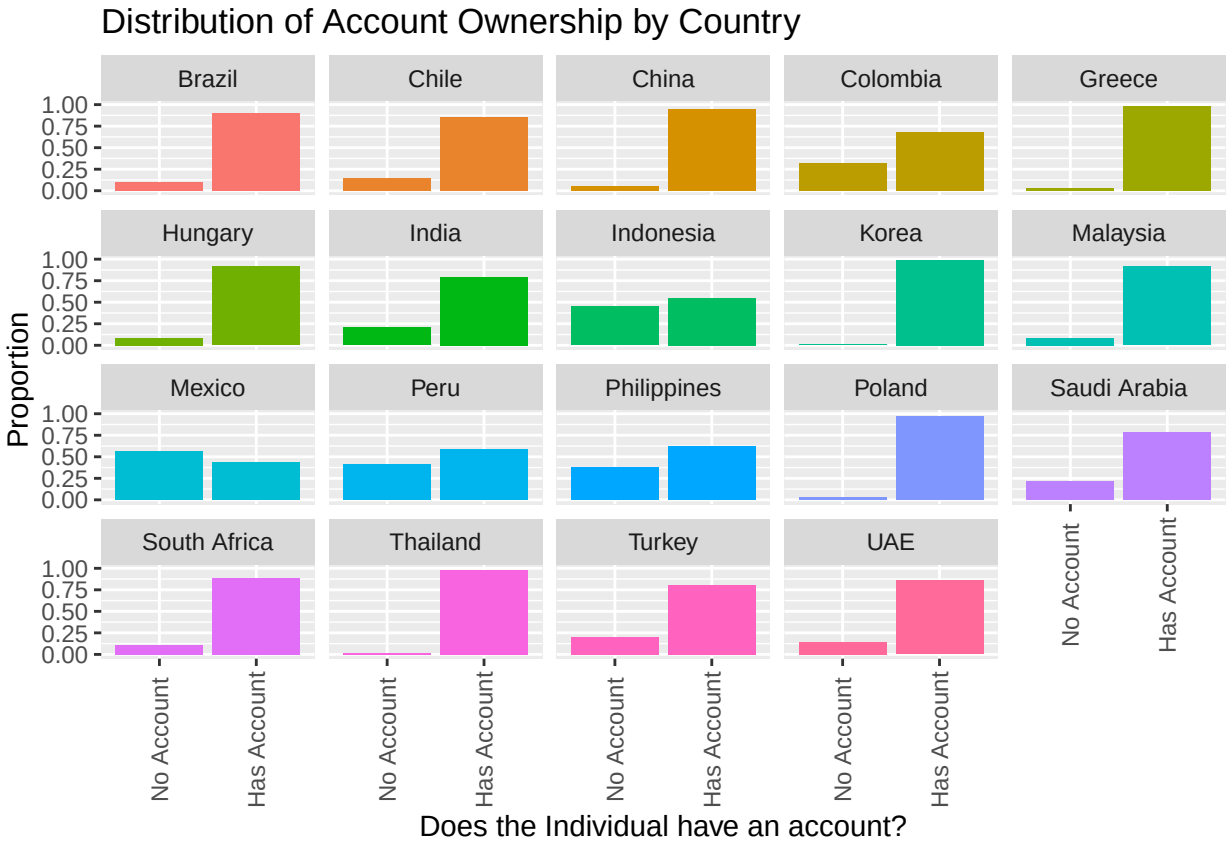

Figure 1: Distribution of Account Ownerhip Across Countries

Age Distribution

Figure 2 plots the age distribution of individuals in our dataset and as we can see, the dataset is skewed towards individuals in the 19-50 year old age groups while over 70 age group is under represented

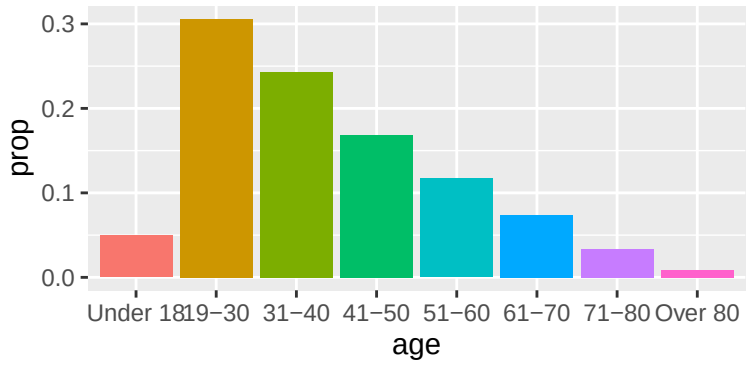

Figure 2: Distribution of Age Groups observed in our dataset

Account Ownership based on Demographic Characteristics

Figure 3 plots the distribution of account ownership our dataset based on demographics of an individual.The disparities are evident, with people having higher income have higher proportion of account holders. Account ownership also increases as the education level of an individual improves and also access to internet improves account ownership. Looking at the age groups, account ownership first increases with age but then falls as age crosses 50 years old. Finally, males show a higher proportion of account ownership compared to females, signalling gender discrimination. These findings point towards the association between demographics and account ownership, which is in line with literature

Account Ownership Variation by Sex and Education across Countries

Figure 4 and Figure 5 plot the distribution of account ownership by education levels and sex respectively. We find that education status while does not improve account ownership in countries such as Greece, UAE and Thailand, it matters a lot in other countries. Account ownership also varies differently by countries by sex. For example, Thailand, India and Indonesia exhibit no or little gender disparity, countries such as Mexico and Saudi Arabia exhibit significant disparities.

Account Ownership and Macroeconomic Indicators

Figure 6 plots the association between various country level indicators comprising of macroeconomic as well as geographical outreach variables. We observe that a positive correlation

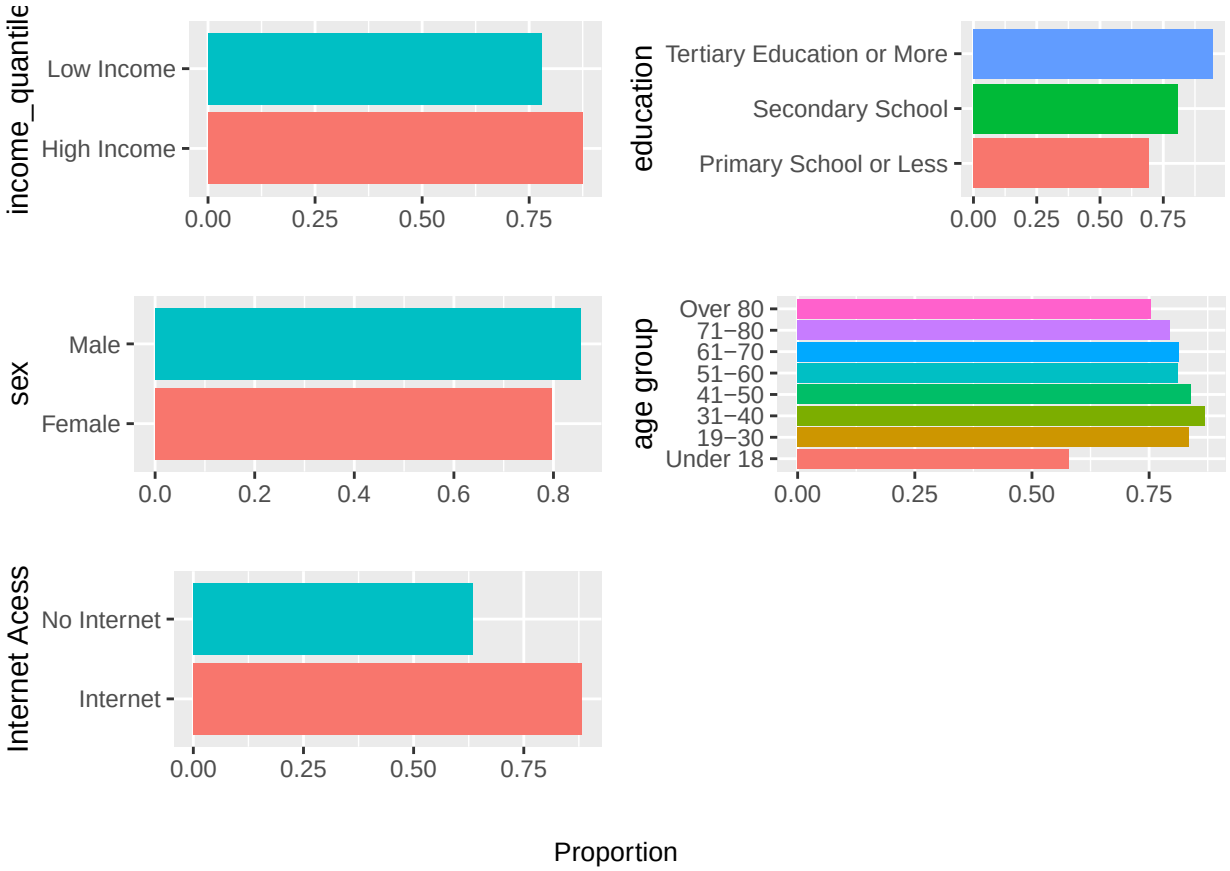

Figure 3: Account Ownership based on Individual Characteristics

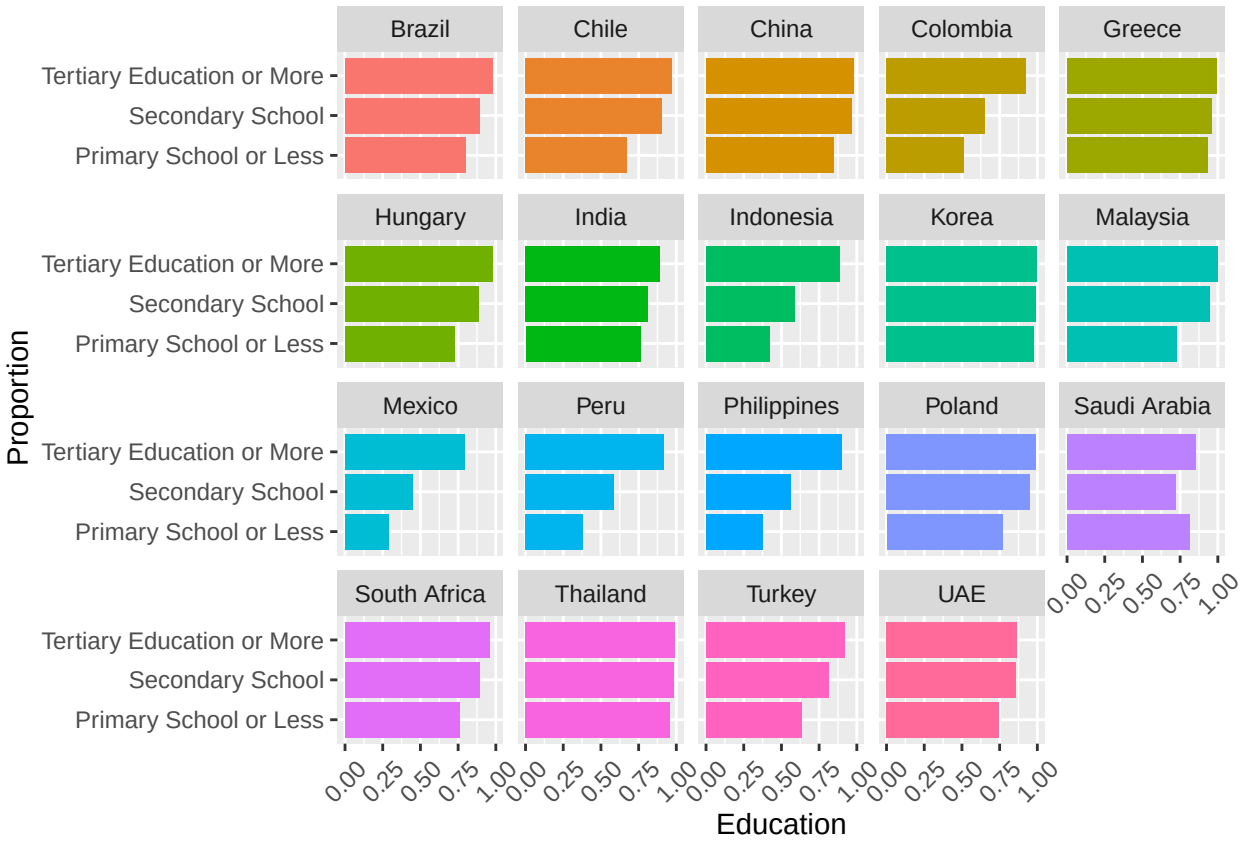

Figure 4: Distribution of Account Ownership within Countries by Education Levels

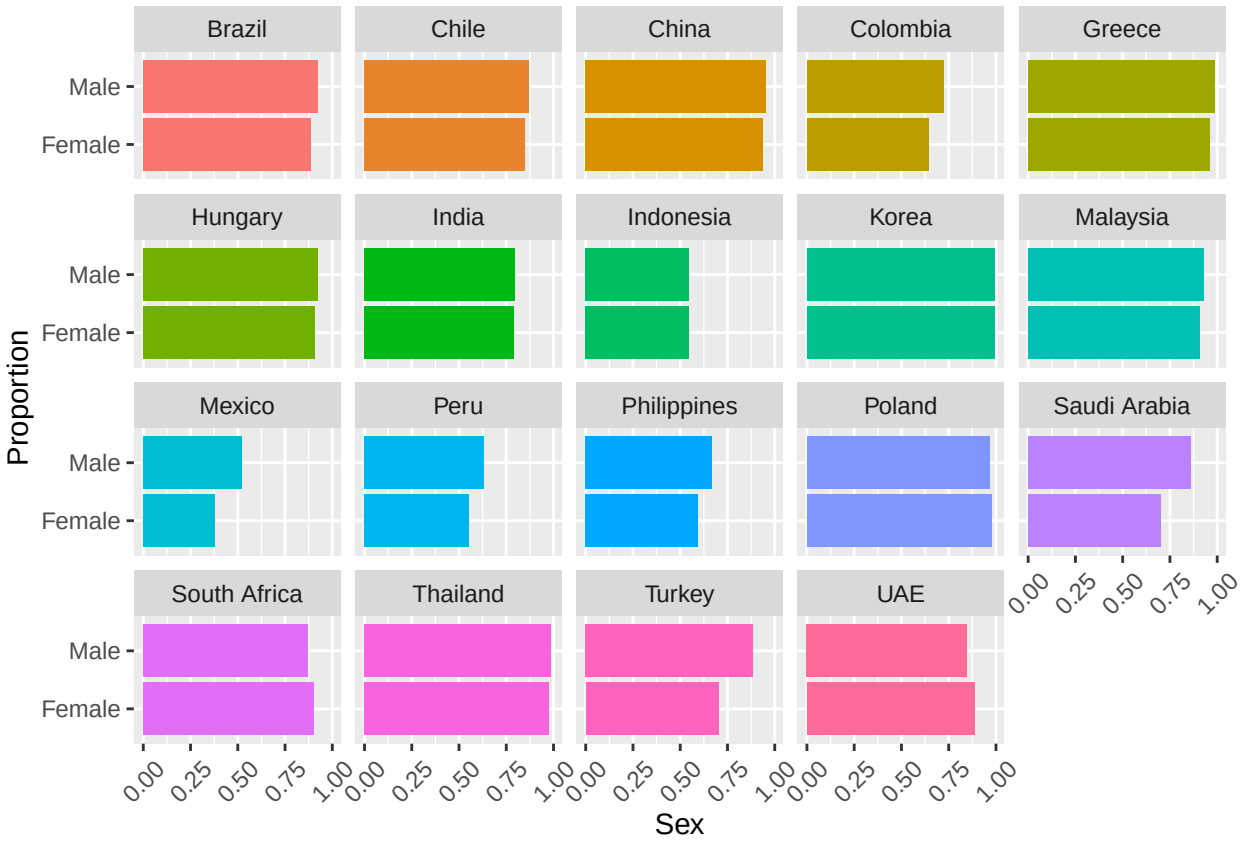

Figure 5: Distribution of Account Ownership within Countries by Sex

emerges between the proportion of account owners and metrics such as Bank Branch Density, ATM Density, and GDP growth rates. On the other hand, higher unemployment rates correlate negatively with account ownership.

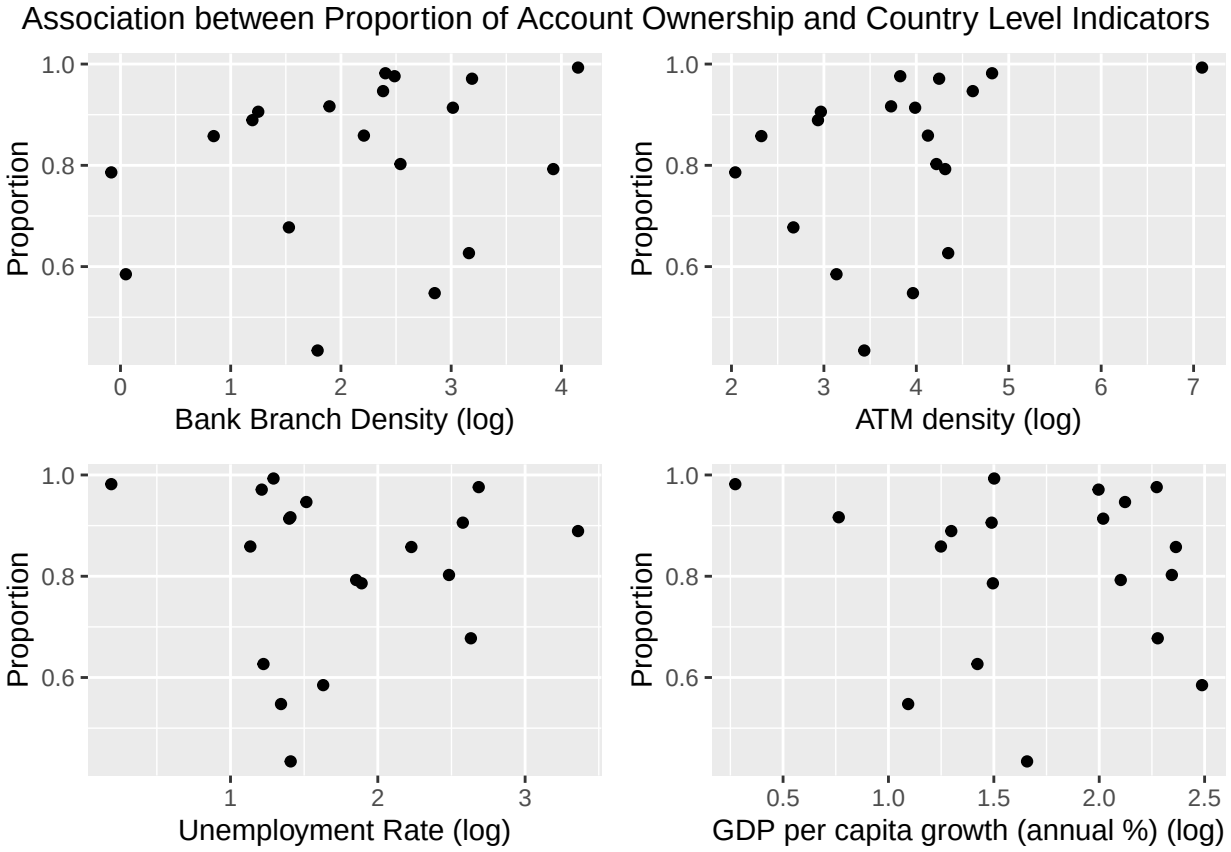

Figure 6: Macroeconomic Indicators and their Association with Account Ownership. Except unemployment, all indicators show a positive relationship as also indicated by literature

Methods

Since we are working with data with multiple levels, the proposed model is a Bayesian hierarchical logistic regression model. Specifically, we fit hierarchical variables for the variable country particularly because account ownership varies significantly by country. For example, in some countries we have negligible number of individuals without bank accounts. We model the hierarchical intercept for countries as a combination of country level variables. The individual level variables are fitted non-hierarchically. Weakly informative priors are set on all our regression parameters using a $N(0,1)$.This was done to incorporate prior knowledge from the literature i.e. the former has a positive impact on account ownership while the latter has a negative impact. This was also done to speed up the convergence process of MCMC. any undue influence or information about the parameters to the prior The following notation will be used to describe the models: let y_i represent the account ownership variable (1=has an account, 0 otherwise), X is a matrix of individual level variables.In Model 2, the matrix X includes interaction terms. The model will be fit using an R package, RStan which allows us the capability to fit a Bayesian model. Three variations of the model were considered which is discussed below. For each model, we validate the results using posterior predictive checks (PPC) to see if our model is able to capture specific test statistics. In particular, we are interested in the Proportion of Account Ownership captured by our Models which will then allow us to compare our observed data to the replicated data from our model to determine which model is the ideal fit. Furthermore, for each model, we will calculate the expected log point-wise predictive density (ELPD) and determine if it is an appropriate criterion for our setting.

Model-1

Bayesian Hierarchical Model with a random intercept by country

$$
\begin{aligned}
y_i|\alpha_{j[i]},\beta &\sim \text{Bernoulli}\left(\text{logit}^{-1}\left(\alpha_{j[i]}+X_i\cdot\beta\right)\right), && \text{for } i=1,2,\dots,N\\
\alpha_{j[i]} &\sim \text{N}\left(\gamma_0+\gamma_1\cdot\text{gdppercapita}_j+\gamma_2\cdot\text{atmdensity}_j+\gamma_3\cdot\text{bankbranchdensity}_j\right.\\
&\quad \left.+\gamma_4\cdot\ \text{UnemploymentRate}_j,\sigma_\alpha^2\right), && \text{for } j=1,2,\dots,19
\end{aligned}
$$

Model-2

Bayesian Hierarchical Model with a random intercept and interaction effect(Matrix X) of female with individual level variables

$$
\begin{aligned}
y_i|\alpha_{j[i]},\beta &\sim \text{Bernoulli}\left(\text{logit}^{-1}\left(\alpha_{j[i]}+X_i\cdot\beta\right)\right), && \text{for } i=1,2,\dots,N\\
\alpha_{j[i]} &\sim \text{N}\left(\gamma_0+\gamma_1\cdot\text{gdppercapita}_j+\gamma_2\cdot\text{atmdensity}_j+\gamma_3\cdot\text{bankbranchdensity}_j\right.\\
&\quad \left.+\gamma_4\cdot\ \text{UnemploymentRate}_j,\sigma_\alpha^2\right), && \text{for } j=1,2,\dots,19
\end{aligned}
$$

Model-3

Bayesian Hierarchical Model with a random intercept and slope by country

$$
\begin{aligned}
y_i \,|\, \alpha_{j[i]}, \beta_{j[i]} &\sim \text{Bernoulli}\left(\text{logit}^{-1}\left(\alpha_{j[i]} + X_i \cdot \beta_{j[i]}\right)\right), && \text{for } i = \\
\alpha_{j[i]} &\sim \text{Normal}\Big(\gamma_0 + \gamma_{1j} \cdot \text{gdppercapita}_j + \gamma_{2j} \cdot \text{atmdensity}_j + \gamma_{3j} \cdot \text{bankbranchdensity}_j \\
&\quad + \gamma_{4j} \cdot \text{Unemployment Rate}_j, \sigma_\alpha^2\Big), && \text{for } j =
\end{aligned}
$$

Results

To ensure our models have converged we look at the $\hat{R}$ and n_{eff} for each parameter. $\hat{R}$ compares the between-and within-chain estimates for each parameter; thus, since it is close to 1 for all parameters in each model, this ensures that the chains have mixed well. Furthermore, our measure of effective sample size, n_{eff} , is also quite large for each parameter. Lastly, the trace plots for the first 10 parameters in each model appeared to have mixed well.

Posterior Predictive Checks

We use 100 samples from our posterior predictive distribution from each model to assess how it fits the distribution of our observed data

Proportion of Account Holders

Figure 7 plots the proportion of account holders in our posterior samples across the three models. In Model 1,most posterior datasets exhibit proportions slightly lower than the actual proportion of account holders while Model 2's posterior datasets most frequently match the true proportion. In contrast, Model 3's posterior datasets exhibit a mix, with some proportions slightly lower and others slightly higher than the actual proportion of account holders.

Distribution of Account Holders by Demographics and Income

Distribution of Account Holders by Sex In Figure 8, we plot the Proportion of Account Holders by Sex across the three models and find that majority of the posterior simulated datasets in Model 1 and Model 3 have higher proportion of male account holders and slightly lower proportion of female account holders than the actual proportion. In contrast, Model 2 has the same proportion as actual for Male account holders while slightly higher proportion for females.

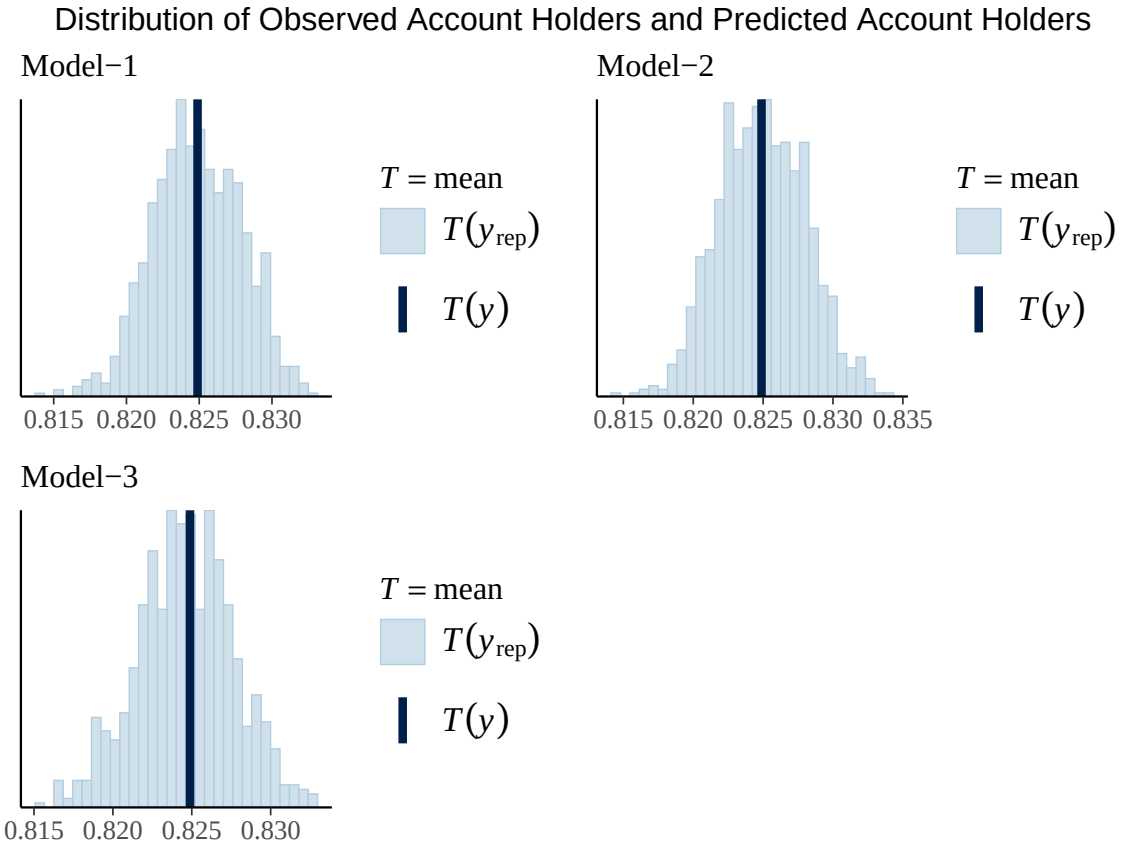

Figure 7: Proportion of Account Ownership in Posterior Samples

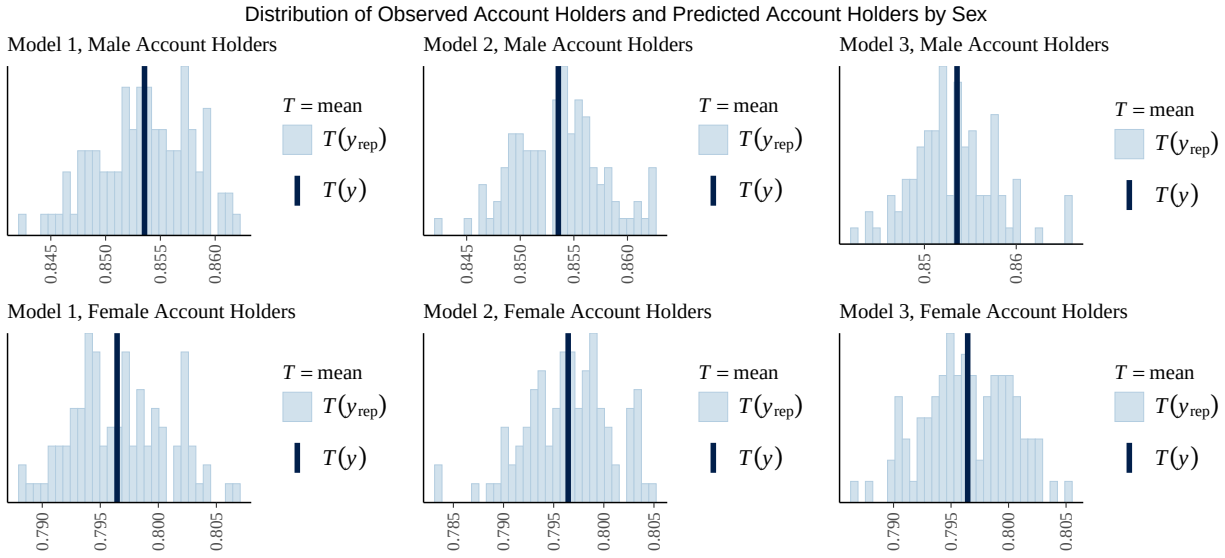

Figure 8: Distribution of Account Ownership by Sex in 100 Samples of Posterior Datasets

Distribution of Account Holders by Internet Access From Figure 9, we observe that Model-1 closely captures the proportion of account holders among those who have internet access but predicts significantly higher proportion of account holders among non internet users.All the three models fail to capture the actual proportion of account holders among the non internet users, this could be due to the less proportion of non internet users in our total dataset.

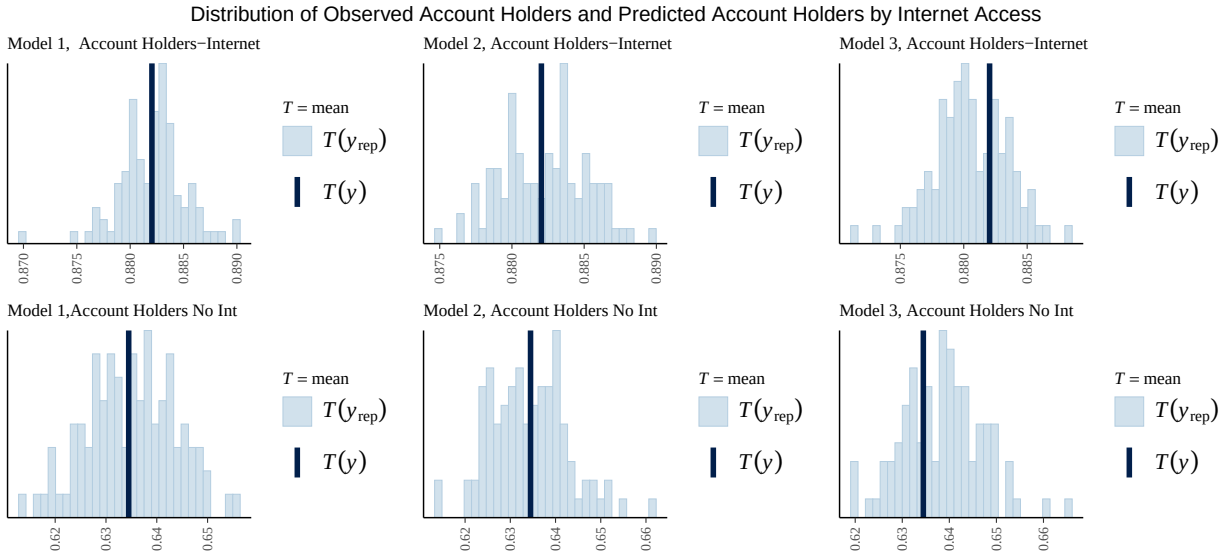

Figure 9: Distribution of Account Ownership by Access to the Internet in 100 Samples of Posterior Datasets

Distribution of Account Holders by Education Levels From Figure 10, it is evident that Model 2 closely matches the proportion of actual account holders observed across the three education levels while Model 3 gives bad results, with higher proportion in primary and secondary education holders but lower proportion for tertiary education holders.

Distribution of Account Holders by Income From Figure 11, we observe that Model 2 closely captures the actual proportion of account holders among both high and low income individuals while Model 3 underestimates the proportion across both the groups.

Distribution of Account Holders for Certain Age Groups In Figure 12, we try to visualize the performance of the three models across the underrepresented age groups in our dataset as indicated in our exploratory data analysis i.e. individuals less than 18 years and above 70 years in age. We find that Model 3 fails to capture the proportion of account holders among less than 18 year olds i.e. it significantly overestimates the proportion when compared to Model-1 and Model-2. When it comes to older age groups, all models fail to capture the true proportion except Model-2 that gives closer to the actual proportion of account holders among people aged 80 or more.

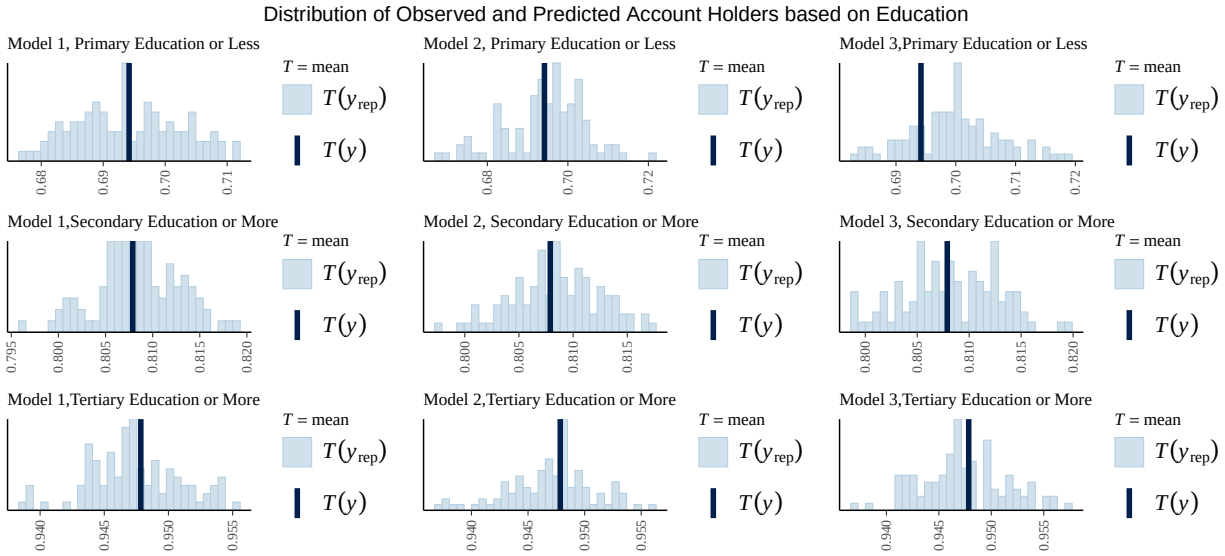

Figure 10: Distribution of Account Ownership by Education Levels in 100 Samples of Posterior Datasets

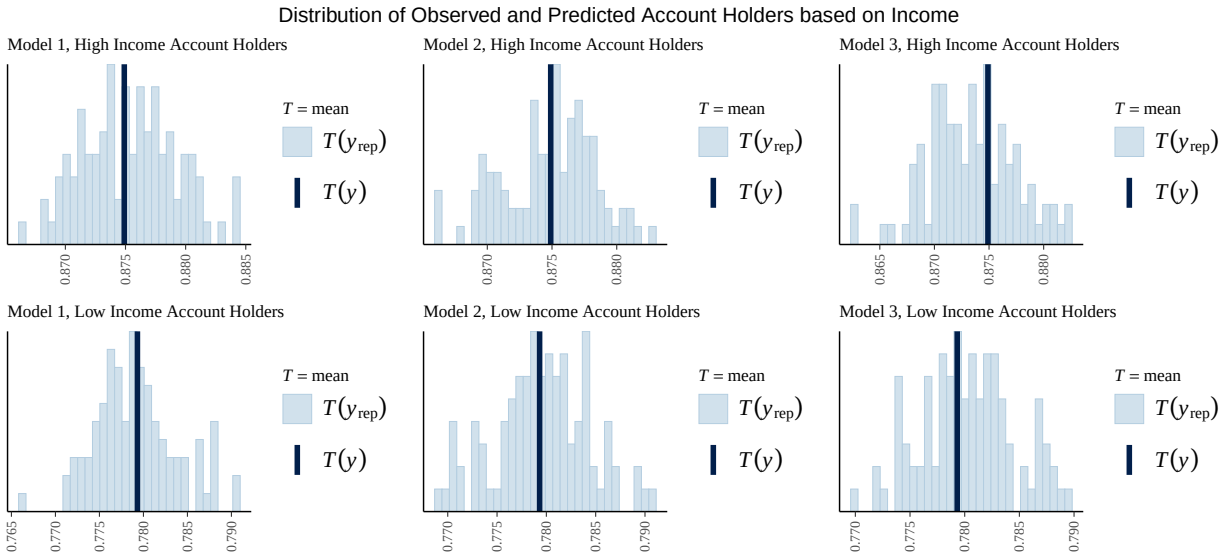

Figure 11: Distribution of Account Ownership by Income Levels in 100 Samples of Posterior Datasets

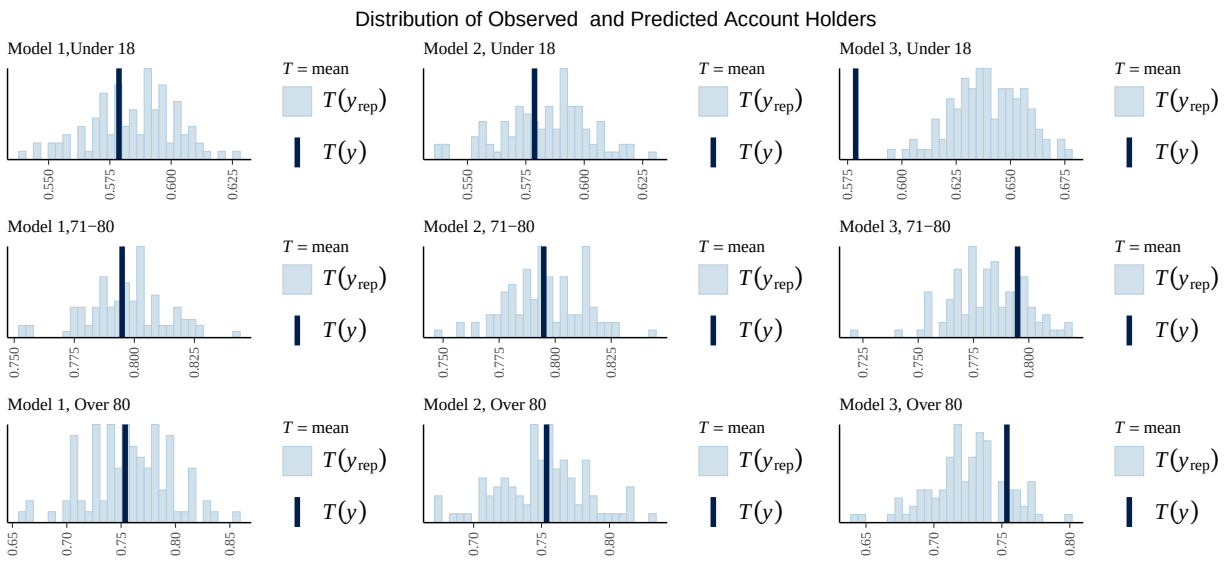

Figure 12: Distribution of Account Ownership by Age in 100 Samples of Posterior Datasets

We also evaluate the Expected Log Predictive Density (ELPD) for the three models to choose an optimal model, While Model 3 has the largest ELPD value among the three models; it clearly performs worse in capturing the test statistic for a lot of variables such as age, sex and education levels. Hence, instead of choosing Model-3, we choose Model-2 which has a second highest ELPD and also performs significantly better in our PPC. The remaining interpretations on the estimated effects will be done solely for this model.

Coefficient Effects

Using the preferred model, we investigate the coefficient estimates for each of our factors.We will exponentiate our effects so it can be interpreted in terms of the odd ratio.

Effects of Demograhics and Income

In Figure 13, we plot the estimated effect of an Individual's Demographics and Income on Account Ownership. The influence of an individual's sex is lower than 1, with males being 0.77 times more likely to possess account ownership compared to females. However, education has a substantial impact, as individuals with secondary school education are 1.45 times more likely to have account ownership than those with primary education, while tertiary educated individuals are 3.41 times more likely. In contrast, access to the internet and income levels show relatively smaller but still positive effects. This could be attributed to the widespread availability of affordable internet options in many countries, along with targeted government policies aimed at encouraging low-income individuals to open bank accounts.

For Age Effects, as shown in Figure 14, the impact of an individual's age on account ownership first increases then eventually falls. Individuals aged 31-60 are notably more than twice

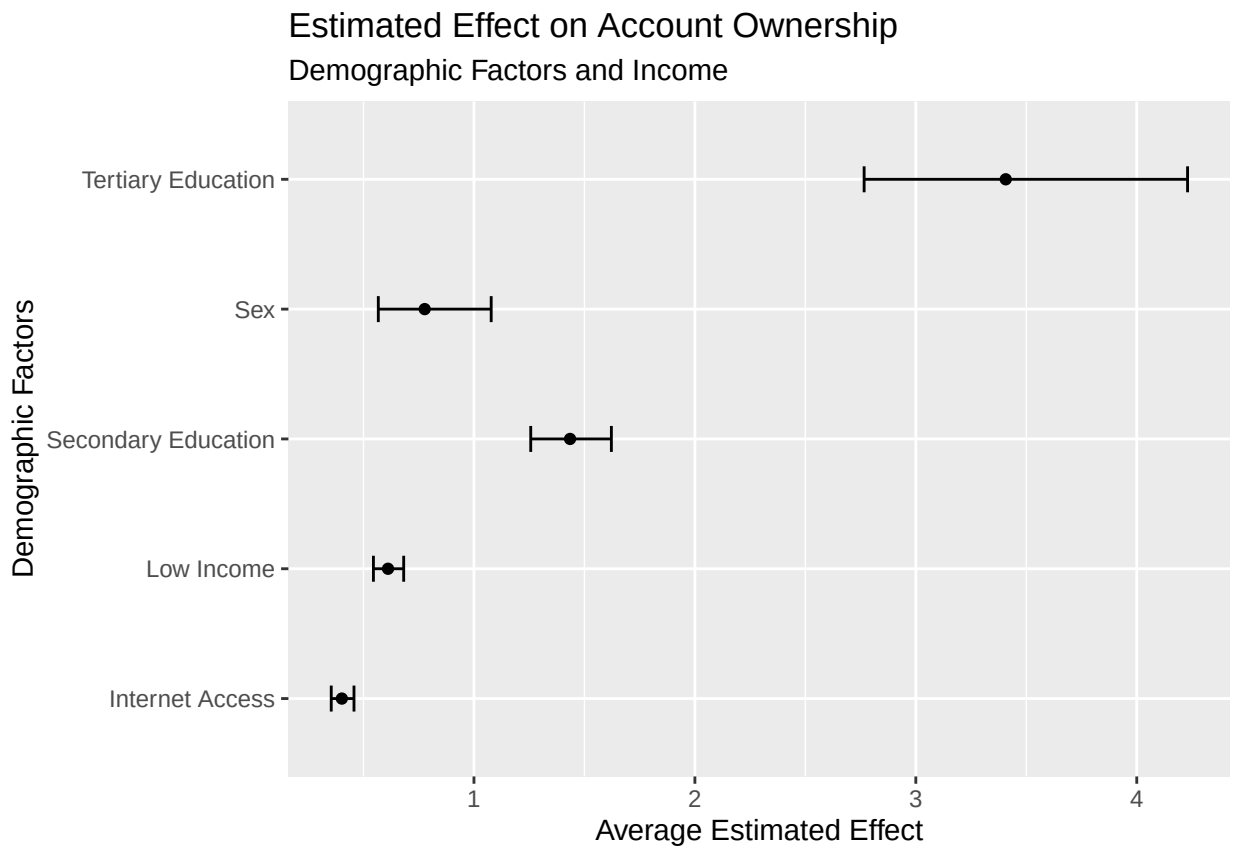

Figure 13: Estimated effect of an Individual's Demographics and Income on Account Ownership.Error Bars Represent 95% Credible Intervals for the Mean Posterior Odds Ratio

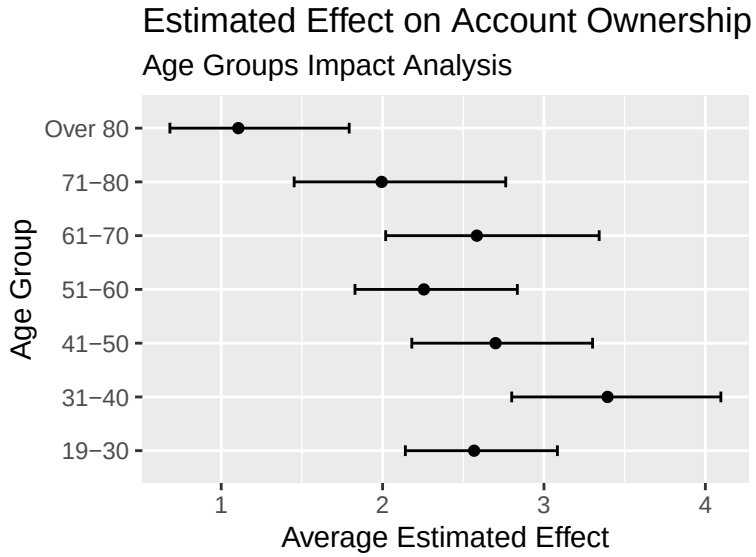

Figure 14: Estimated Effect on Account Ownership Based on Age. Error Bars Represent 95% Credible Intervals for the Mean Posterior Odds Ratio

as likely to have account ownership compared to those under 18 years old. However, this propensity appears to level off among senior citizens, who are only approximately 1 times more likely to possess account ownership.

Interaction Effects

In Figure 15, we present the interaction effects of Sex with certain key attributes like Income and Education Levels. It's evident that among individuals with low income, males are approximately 1.2 times more likely to possess a bank account compared to females. Similarly, when considering education levels, males are generally about 1 times more likely to have account ownership than females, with a range between 0.5 to 1.5. The credible intervals however are large across the three interaction terms, which denotes large uncertainty in the estimates. This could be due to the fact that in some countries, women have made fair advancements while in a lot of countries due to societal norms where, despite being highly educated, women in many countries are discouraged from managing their own finances through personal bank accounts.

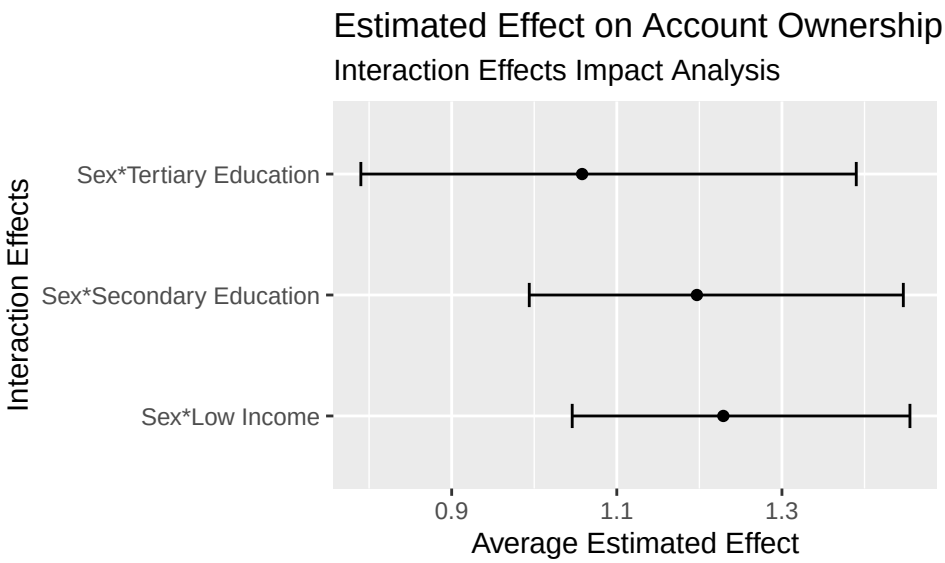

Figure 15: Estimated Effect on Account Ownership Based on Selected Interaction Effects. Error Bars Represent 95% Credible Intervals for the Mean Posterior Odds Ratio

Country Effects

In Figure 16, we plot the log-odds to visualize country effects for the ease of visual understanding. There appears to be varying effects in the baseline log odds of each country. For example, an individual in South Korea, the average baseline log odds of account ownership is 3.85 which ~47 times while for an individual in Saudi Africa, the average baseline log odds is -0.57 which is only 0.56 times. However, we cannot ignore the uncertainty in South Korea's estimates, since our data for South Korea had 100% account ownership that explains

the huge baseline average odds ratio. Based on our posterior estimates for Model-2, we can say that countries such as South Korea, India, Poland, Hungary and China are leading the emerging markets in terms of financial inclusion while Saudi Arabia, Brazil and South Africa need to work harder to promote financial inclusion within their country

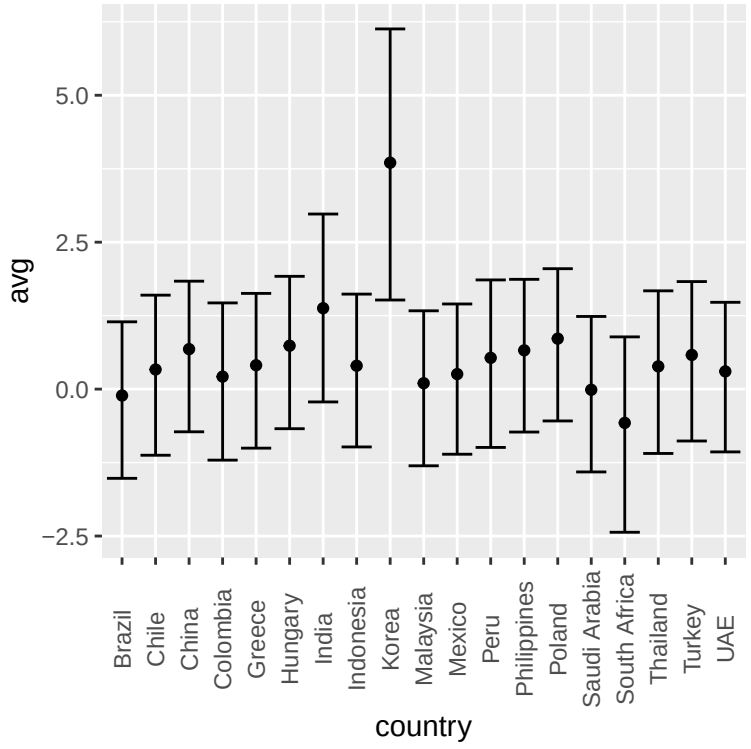

Figure 16: Estimated Effect on Account Ownership Based on Country. Error Bars Represent 95% Credible Intervals for the Mean Posterior Log-Odds Ratio

Additional Models

Alternative models with a hierarchical effect for age was also considered since our dataset has a lesser representation for certain age groups. However, due to computational constraints, we decided to stick to the three models as described throughout our study

Discussion

The focus of this study was to investigate the effects on Financial Inclusion from demographic, income and country level factors. While previous literature focuses on a frequentist framework to achieve this, we use a Bayesian Hierarchichal Approach since it allows us to pool information from multiple levels in order to derive estimates for groups with limited information. For example, the Global Findex Databse doesn't publish indicators for several

emerging markets. This model could therefore be used to model financial inclusion for which Findex doesn't publish.

From the model analysis and validation, we found that the model with interaction term between female and all the individual level was able to capture the data better even if it had a lower ELPD. Demographic analysis can be done by examining the effects from age, sex, access to the internet, education levels and Income. We observed that as individual get older, there are less likely to own a bank account. However, as indicated by our EDA, a lot of countries had missing data for certain age groups, hence future work should concentrate on modelling age hierarchically. As for effects of sex, male individuals are likely to have higher odds of owning a bank account which agrees with previous literature. Our effects based on sex and demographics such as Income and Education when grouped together also showed similar results. Hence, this points towards the need to empower females in emerging markets in addition to focusing on economic growth.

The country level mean baseline odds are modeled using a combination of macroeconomic and geographical outreach variables taken from various data sources as well as our review of literature. While emerging markets are similar in nature when it comes to economic growth, nevertheless we found signifacnt variation in the country level mean baseline odds wherein based on the data, countries such as South Korea, India, China and Poland are performing a lot better in the advancement of Financial Inclusion.

While our model is able to capture the effects and variation in our data, we recommend exploration of more suitable priors as this can have an effect on model fit and performance. Currently, weakly informative priors are being used, so more informative priors using information from previous literature on the effects could provide a better fit. Future work could also consider modelling country level variation by including variables that signify the strength of bankruptcy laws in a nation as good laws instill confidence among borrowers and lenders that further bolsters financial health. In addition, future studies could also include developed countries in this model and draw comparisons on the progress of emerging markets. In conclusion, although the model had room for improvement, this analysis has provided insights into key factors that promote financial inclusion, which paves the way for policyholders to develop policies that focus on such an important indicator of the overall well being in a society.

References

- Al Khub A, Saeudy M, Gerged AM. Digital Financial Inclusion in Emerging Economies: Evidence from Jordan. *Journal of Risk and Financial Management*. 2024; 17(2):66. https://doi.org/10.3390/jrfm17020066
- Determinants of financial inclusion: results of multilevel analyses Bermeo , E. (Author). 1 Oct 2019 Student thesis: Doctoral Thesis › Doctor of Philosophy (PhD)