Term project

Sára Vargha

22/12/2021

Introduction

Financial inclusion is gaining increasing attention as it is among the key enablers of the UN Sustainable Development Goals (SDGs), which will be a key driver of reducing poverty and enhancing shared global economic prosperity. According to World Bank studies, globally around 1.7 billion people, or 31% of the adult population, lack access to financial services - and a significant portion of these unbanked groups are women. Thus, besides its importance in driving economic development, improved financial inclusion can also lead to greater gender equality. The aim of this analysis is (1) to uncover the pattern of association between financial inclusion and gender, and (2) examine the same when other factors, such as age, level of education, employment status and household income quantile are taken into consideration.

Data description

The basis of the analysis is the 2017 Global Findex database drawn from survey data covering almost 150,000 people in 144 economies—representing more than 97 percent of the world's population. The survey was carried out over the 2017 calendar year by Gallup, Inc., as part of its Gallup World Poll, which since 2005 has annually conducted surveys of approximately 1,000 people in each of more than 160 economies and in over 150 languages, using randomly selected, nationally representative samples. The target population is the entire civilian population from the age of 15 and above. The original dataset can be downloaded from the World Bank's Microdata Library. (https://microdata.worldbank.org/index.php/catalog/3324)

Data cleaning

The dataset contained originally 154,923 observations and 105 variables. As the majority of these variables were not necessarily relevant for my analysis, I decided to keep only a set of variables that I intend to incorporate in my models. As a result, I kept 11 variables from the original dataset.

As part of the exploratory data analysis, I continued with examining the most important variables. I originally wanted to include four explanatory variables in my analysis, out of which age was the only numeric variable, while the other three were categorical variables that needed to be converted into binaries for the sake of modeling: (1) gender (0 - male, 1 - female), (2) level of education (three binaries for primary, secondary and tertiary levels), and (3) income quantile category (five binaries for the five quantiles).

Once I had all the variables in the right form, I explored their distributions in my data and did some extra cleaning based on the outcomes. First, "age" had 451 missing values that I replaced with the average value of age in the data. Since the variable had a skewed distribution with a long right tail, I decided to apply log transformation to check which form would lead to a better fit later on. In case of the other four key variables, I detected missing values in case of employment status ("employed") and the level of education ("education"). Since the number of missing values were negligible (<1% of the entire dataset), I dropped them.

After the above steps taken, the remaining number of observations is 153,012. As we can see from Table 1, 63% of the respondents had an account and 54% of them were female. The age of survey participants ranged from 15 to 99 with a median of 39 years. The median respondent had a secondary education and belongs to the thirds income quantile.

Models

As our dependent variable is a binary variable, the pattern of association between gender and having an account will be examined with the use of various linear probability models.

The first model tested (lpm1) aimed to uncover the unconditional gender gap, and thus, contained no control variables. Based on the outcome of the linear probability model, we can conclude that 67.8% of male respondents had an account and female participants were 8.4% less likely to have an account, which means that 59.1% of them had an account at 1% level of significance.

In my second model, I ran two versions with (a) ln_age (lpm2a) and (b) age as control variable to see (lpm2b). While using age variable has a better interpretation, ln_age might be a better due its close-to-normal distribution. Looking at the results of version (a), we can conclude that respondents from the same gender who were "10% older" were 1.816 % points more likely to have an account with a 3.7% R-squared, while version (b) tells us that respondents from the same gender who were 1 year older were 0.4% points more likely to have an account with a 3% R-squared. Although the fit of version (a) was somewhat better, I decided to keep age in my model because of the easier interpretation. Also, both models implied at 1% level of significance that comparing respondents with the same age, females were 8.6% points less likely to have an account than their male counterparts.

In the third model (lpm3), I added employment status on top of gender and age. Based on the coefficients, we can conclude that when comparing respondents both employed and with the same age, females were 5.2% points less likely to have an account than their male counterparts at 1% level of significance. Also, employed females with the same age were 17.5% points more likely to have an account than unemployed females with the same age, indicating that being employed has a positive association with having an account.

In the fourth model (lpm4), the levels of education were added with primary or below (edu_primary) being the reference level. Based on the outcome, we can conclude that higher levels of education lead to an increased probability of having a bank account: when comparing respondents with the same gender, age and employment status, those who had secondary level of education were 30.4% points more likely to have a bank account compared to their counterparts with a primary or below level of education, while those with a tertiary or above level education level were 47.1% points more likely to have an account at 1% level of significance. When comparing female and male respondents with the same age, employment status and level of education, females were 3.4 % points less likely to have an account than their male counterparts at 1% level of significance. This model reached an R-squared of 18.6%.

In the last and fifth model (lpm5), the income quantiles were added to the model with the first quantile being the reference level (inc_poorest). However, although results were significant at <1% level of significance, I decided not to go further as adding these variables only sightly increased our R-squared (19.1% from 18.6%) while bringing significant complexity to the model.

In the final part of the analysis I used the same set of explanatory variables as in case of lpm4 to run logit and probit probability models. The summary of the outcomes of these models (including the marginal differences of our logit and probit models) were stored in Table 3.

From the results, we can indeed see that LPM, logit and probit lead to very similar results, with gender, age, employment status and the different education levels all being significant explanatory variables at 1% and logit and probit having the smallest standard errors. Although the outcome of logit and probit models has no straightforward interpretation, their marginal differences have practically the interpretation as LPM:

- When comparing respondents with the same gender, age and employment status, those who had secondary level of education were 26.6% points (logit)/26.8% points (probit) more likely to have a

bank account compared to their counterparts with a primary or below level of education, while those with a tertiary or above level education level were 37.4% points (logit)/37.8% percentage points (probit) more likely to have an account at 1% level of significance
- When comparing female and male respondents with the same age, employment status and level of education, females were 3.4 % points less likely to have an account than their male counterparts at 1% level of significance.

It was interesting to see, that when it came to "employed" and "edu_secondary", the coefficients of the probit and logit models were slightly below the LPM models, while in case of "edu_tertiary" the difference was almost 10% points. When we consider the plot that compares the fit of the three models, we can see that the probit and logit curves are moving away from with higher values of predicted probabilities of LPM.

Looking at our histogram where we plotted one histogram for observations with actual y=1 (has an account) and one for observations with actual y=0 (doesn't have an account) based on our LPM model (lpm4), we can see that the fit of the prediction is far from perfect: the two distributions overlap to a large extent and a larger part of the distribution covers higher predicted values among individuals who have an account than it should.

When we consider biasedness of our models, all of our three models (LPM, logit, probit) are slightly biased with LPM having the lowest level of bias. As for calibration, based on the calibration figures all of our models seem well-calibrated, as they stay close to the 45 degree line.

Robustness check

Although the above discussed models are showing convincing results at 1% level of significance, the results might be different when we use a different regression approach and the external validity of the analysis also can be low, as we cannot be 100% certain that the three dimensions (time, space, other groups of respondents) of the general pattern that are data represents is the same in the actual population - in our case adult population of the Earth - that we are interested in. To address external validity, the analysis could be repeated at a different time - using the survey results from a previous year or taking the geographic data of respondents into account (which was not part of the current analysis).

As for using various regression approaches, the introduction of new control variables in the above discussed linear probability models is one fairly simple way to proceed. However, the analysis could be further enriched by using piecewise linear splines, quadratic forms and interaction terms to check robustness.

Nevertheless, as the analysis was based on a survey conducted using randomly selected, nationally representative samples and included approximately 1,000 people in each of more than 160 countries in 2017, the covered sample can be a good approximation of the actual adult population of the world when we assume that its composition hasn't changed drastically in the past 4 years. Also, enriching the linear probability models with new control variables constantly increased the explaining power of the model and supported the expected pattern of association, therefore the result of this analysis can be regarded at a basic level.

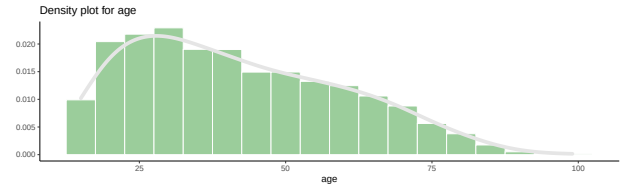
Conclusion

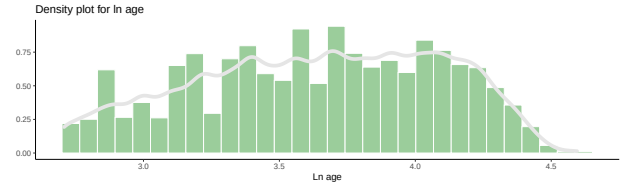
Based on the analysis, we can conclude that there is indeed a gender gap in financial inclusion: females tend to have a lower probability of having an account than their male counterparts. Although the difference seems to increase with higher age and higher levels of education, as well as with being employed, it never fully disappears.

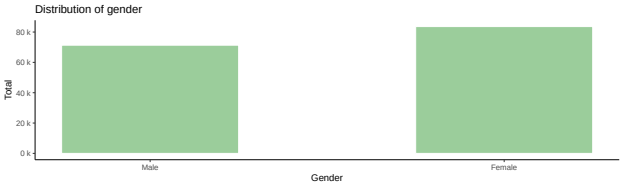
The scope of this analysis did not cover some important aspect, such as geographic differences, but provides initial evidence for the worldwide gender gap in financial inclusion. For policymakers and decision makers of the financial sector this results might serve as motivation for more detailed research to understand why the gap exists and what its consequences might be in our society. In the longer run, financial education programs and/or financial products tailored for women could be potential ways to address and close the gap.

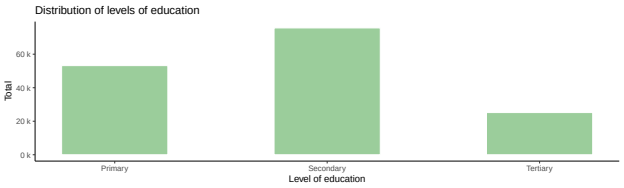
Appendix

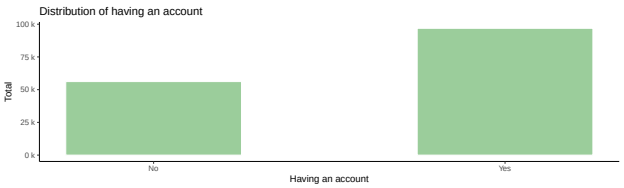
Distribution of key variables











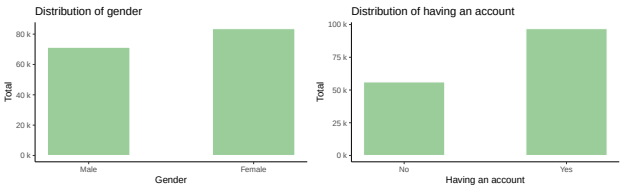


Table 1: Descriptive statistics of variables

Table 1: Descriptive statistics

	Mean	Median	SD	Min	Max	P05	P95
"Having an account"	0.63	1.00	0.48	0	1	0.00	1.00
"Gender"	0.54	1.00	0.50	0	1	0.00	1.00
"Age"	41.79	39.00	17.87	15.00	99.00	18.00	74.00
"Level of education"	1.82	2.00	0.69	1	3	1.00	3.00
"Employment status"	0.63	1.00	0.48	0	1	0.00	1.00
"Income quantile"	3.19	3.00	1.42	1	5	1.00	5.00

Pattern of associaton between having an account and gender

Table 2: Comparison of LPM models

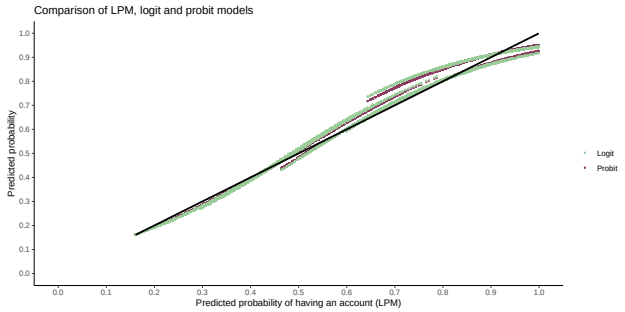
	LPM1	LPM2a	LPM2b	LPM3	LPM4	LPM5
Constant	0.678**	0.004	0.512**	0.344**	0.113**	0.070**
	(0.002)	(0.010)	(0.003)	(0.004)	(0.004)	(0.005)
gender	−0.084**	−0.086**	−0.085**	−0.052**	−0.034**	−0.030**
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
ln_age		0.186**				
		(0.003)				
age			0.004**	0.005**	0.005**	0.005**
			(0.000)	(0.000)	(0.000)	(0.000)
employed				0.175**	0.134**	0.131**
				(0.003)	(0.002)	(0.002)
edu_secondary					0.304**	0.293**
					(0.003)	(0.003)
edu_tertiary					0.471**	0.446**
					(0.003)	(0.003)
inc_second						0.030**
						(0.004)
inc_middle						0.047**
						(0.004)
inc_fourth						0.069**
						(0.004)
inc_richest						0.106**
						(0.004)
Num.Obs.	153 012	153 012	153 012	153 012	153 012	153 012
R2	0.007	0.037	0.030	0.058	0.186	0.191
R2 Adj.	0.007	0.037	0.030	0.058	0.186	0.191
Std.Errors	Heteroskedasticity-robust	Heteroskedasticity-robust	Heteroskedasticity-robust	Heteroskedasticity-robust	Heteroskedasticity-robust	Heteroskedasticity-robust

* $p < 0.05$, ** $p < 0.01$

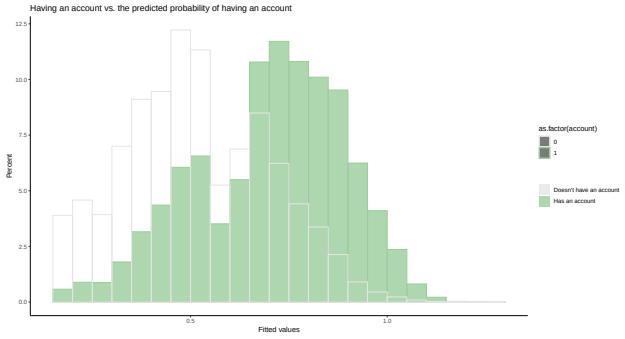
Table 3: Comparison of LPM, Logit and Probit models

	LPM4	Logit	Logit marg.	Probit	Probit marg.
Constant	0.113**	−1.875**		−1.128**	
	(0.004)	(0.023)		(0.013)	
gender	−0.034**	−0.181**	−0.034**	−0.106**	−0.034**
	(0.002)	(0.012)	(0.002)	(0.007)	(0.002)
age	0.005**	0.028**	0.005**	0.016**	0.005**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
employed	0.134**	0.651**	0.126**	0.394**	0.128**
	(0.002)	(0.013)	(0.002)	(0.008)	(0.002)
edu_secondary	0.304**	1.368**	0.266**	0.835**	0.268**
	(0.003)	(0.013)	(0.002)	(0.008)	(0.002)
edu_tertiary	0.471**	2.612**	0.374**	1.530**	0.378**
	(0.003)	(0.024)	(0.002)	(0.013)	(0.002)
Num.Obs.	153 012	153 012	153 012	153 012	153 012
Std.Errors	Heteroskedasticity-robust				

* p < 0.05, ** p < 0.01



Goodness of fit



Calibration curves (LPM, logit, probit)