Gender Gap in Account Ownership: A glance into South Asian Economies

by

Muhammad Arbash Malik – 2202071

Introduction

Financial inclusion implies that both individuals and businesses can avail themselves of beneficial and cost-effective financial products and services tailored to their requirements, including transactions, payments, savings, credit, and insurance. According to the United Nations, Financial Inclusion has been identified as an enabler for 7 of the 17 Sustainable Development Goals. According to the World Bank, it is a key enabler to reducing poverty and boosting prosperity. And from the latest World Bank Findex data, close to 25% of adults - 1.3 billion - are still unbanked. About half of unbanked people included women poor households in rural areas of the workforce. Between 2011 and 2021 gender gap in account ownership remained stuck at 6.2 percentage points in developing countries, hindering women from being able to effectively control their financial lives. More than half of the world's unbanked adults live in seven economies and mostly the economies are from the South Asia region.

My research question is whether there is a pattern of association between financial inclusion and gender in the South Asian Economies. Other control variables, such as age, employment, level of education, income quintile, and rurality will also be introduced to examine the association.

Data

The dataset used for this analysis is the Global Financial Inclusion (Global Findex) Dataset for 2021. The data consists of surveys from 139 economies. The survey is done on the adult population, and as per the World Bank definition the age is 15+ for being an adult. The survey collected approximately 1000 observations for every economy, while for bigger countries in terms of population, such as India and China, there are 3000 observations. Each observation is a response to the survey. The data can be downloaded from this link.

The dataset has 143,888 observations and 128 variables. For my analysis I chose South Asia as my region, which consists of 6 economies: Afghanistan, Bangladesh, India, Nepal, Pakistan, and Sri Lanka. I trimmed down the data set, by deleting rows above 1000 for each economy. Since it was done randomly, it ensures that the sample is still representative of India. And ensures that all the economies have the same number of observations in the data set. This trimmed down the dataset to 6000 rows, each economy having 1000 observations.

The primary variable in this analysis is *account* which takes the value of 1 and 0, representing if the respondent has an account or not. The explanatory variable in this analysis is *female* indicating if the respondent is a female or not. The control variables in this analysis are *age*, *educ, emp_in*, *inc_q*, and *urbanity_f2f* representing age of the respondent, education level of the respondent, if respondent is employed, which income quintile does the respondent belong to, and does the respondent is in rural area. The rest of the variables were dropped from the dataset, and the remaining variables were transformed, and some new columns were created to better represent the dataset.

The *female* variable originally had values 1 (female) and 2 (male), which was transformed to take value of 1 (female) and 0 (male). The *employed* variable also originally had values 1 (employed) and 2 (unemployed), which was transformed to take value of 1 (employed) and 0 (unemployed). The column *rural* originally had values 1 (rural) and 2 (urban), which was transformed to take value of 1 (rural) and 0 (urban). Three new columns for education level were created (*primary_edu, secondary_edu,* and *tertiary_edu,* all having values 1 or 0 to identify the education level). Five new columns for income quintile level were also created to represent the quintiles (*poorest_inc, second_inc, middle_inc, fourth_inc, richest_inc* having values 1 or 0 to identify the income level). I explored my data distributions in my data and found out that *age* had one missing value which I replaced with the mean of the column. Except *age,* all variables were binary (see Exhibit 1)

Models

Since my dependent variable is a binary variable, the pattern of association will be modeled with the use of various linear probability models. Every model except the first has country as a control variable so that we can check the gap for each country.

Model 1: Unconditional Gender Gap

The first model was designed to uncover the unconditional gender gap, meaning I did not use any control variables. From the results, 59.4% of male respondents had an account and female participants were 11.8% less likely to have an account. The observed results are also statistically significant at the 1% significance level.

Model 2: Age

The second model incorporated *age* as the control variable as well as the *country*. From the results, respondents who are the from same gender and who are 1 year older are 0.1% more likely to have an account. The model also implies that females are 10.4% less likely to have an account. The obtained outcomes demonstrate statistical significance at the 1% significance level.

Model 3: Employed

The third model included *employed* as another control variable on top of *age.* From the results, when comparing respondents who are employed and who are the same age, females were 7.8% less likely to have an account than their male counterpart. When comparing the effect of employment in females, an employed female is 6.7% more likely to have an account than an unemployed female with the same age. The observed results are statistically significant at the 1% significance level.

Model 4: Rural

The fourth model incorporates rurality, *rural* variable, as the third control variable. Based on the coefficients, females who are from rural areas are 5.4% less likely to have an account than females who are from urban areas. The result is interpreted for the same age and employment status. The model also gives a coefficient of -0.076 on *female* meaning that according to the model, females are 7.6% less likely to have an account than their male counterpart. The results are statistically significant at the 1% significance level.

Model 5: Education

In this model, only two levels of education (*secondary_edu & tertiary_edu)* were added so that the *primary_edu* is the reference level. From the results it is clear, higher levels of education lead to a higher probability of having an account. Comparing respondents of same gender, age, employment status, and rurality; respondents with a secondary level of education are 11.9% more likely to have an account. For respondents with tertiary level of education, they are 29.8% more likely to have a balanced account than their counterparts. When comparing female and male respondents with the same age, employment status, and the level of education, females were 5.6% less likely to have an account than their male counterparts. All the results demonstrate statistical significance at the 1% significance level.

Model 6: Education

For the last model, all income quintiles were added except the *poorest_inc,* so that a reference level could be set. Based on the coefficients, it is also clear that having a higher income leads to a higher probability of having an account. For the second income quintile and the middle-income quintile, the coefficients are 0.011 and 0.013 respectively, but they are not statistically significant. However, for the fourth income quintile and the richest income quintile, the coefficients are 0.051 and 0.056 respectively, and both are statistically significant. Meaning Comparing respondents of same gender, age, rurality, employment status, and education, respondents from fourth income quintile are 5.1% more likely to have an account while respondents from richest income quintile are 5.6% more likely to have an account.

Analysis

Of all the models, model 6 has the highest R^2 (see Exhibit 3). Although the increase in R^2 from model 5 to model 6 is negligible (0.374 to 0.376), I am keeping the model 6 as my main model to explain the gender gap. Since the Linear Probability Model has a major limitation that there is no restriction on LPM so that it doesn't generate predicted probability greater than 1, as proven in our case (1.21) (See Exhibit 4).

To overcome this limitation, I incorporated Logit and Probit models. From the results, we can see that Logit and Probit models give similar results like the Linear Probability Model, and all the results are statistically significant at the 1% significance level except the second income quintile and the middle-income quintile (See Exhibit 5 and Exhibit 6). Although the coefficients of Logit and Probit models have no straightforward interpretation, their marginal differences have practically the same interpretation as Linear Probability Model (LPM).

The estimates are very similar for LPM, Logit Marginal, and Probit Marginal. For example, if we take the explanatory variable *female* the estimates are -5.2%, -5.4%, -5.2% for LPM, Logit Marginal, and Probit Marginal respectively. When comparing respondents with the same gender and age, employed individuals are 6.9% (Logit marginal difference), 6.9% (Probit marginal difference) more likely to have an account. For education, comparing respondents with the same gender, age, employment status, and rurality, those who had secondary level of education are 10.5% (Logit marginal difference), 10.6% (Probit marginal difference) more likely to have a bank account compared to their counterparts with a primary level or below education, while those with a tertiary or above education level are 26.9% (Logit marginal difference), 25.9% (Probit marginal difference) more likely to have. All the results are statistically significant at 1% significance level.

From Exhibit 7, the plot for both Logit and Probit models, you can see the curve move away from the line at higher predicted values of the Linear Probability Model. For LPM probability values between 0.35 and 0.6, Logit and Probit both predicted higher probabilities for some of the points.

Goodness of Fit and Bias

To compare which model fits better, a histogram for predicted values for Model 1 and Model 6 was made (see Exhibit 8). Hollow bars are where account = 0, and filled bars are where account = 1. We can see that the fit of the prediction for Model 6 is far from perfect: the two distributions mostly overlap. But the histogram for Model 6 has smaller overlap compared to Model 1, so Model 6 has a better fit. Comparing Model 6 with Logit and Probit models, a statistic table of R-squared, Brier-score, Pseudo R-squared, and Log-Loss was created (see Exhibit 9). R-squared here is useless since our dependent variable is not a quantitative variable. So, for comparing Model 6 and Logit and Probit we use Brier score and Log-Loss. From the Brier Score Logit (0.154) and Probit (0.154) models are a better fit than LPM (0.155) but just marginally. The Logit and Probit Models also take the win when we compare Log-Loss, both models have a value of -0.469 while LPM has a value of -0.543.

For comparing the biasedness between the models, calibration curves for the predictions were made (See Exhibit 10). We can see all the models were well calibrated as they all stayed close to the 45-degree line. But LPM had the lowest bias.

Findings and Summary

Throughout all the 6 linear probability models, and the Logit and the Probit model constructed, show that there is a gender gap in financial inclusion in South Asian Economies. Although the gap decreases as I controlled for age, employment, rurality, education, and income, it remains.

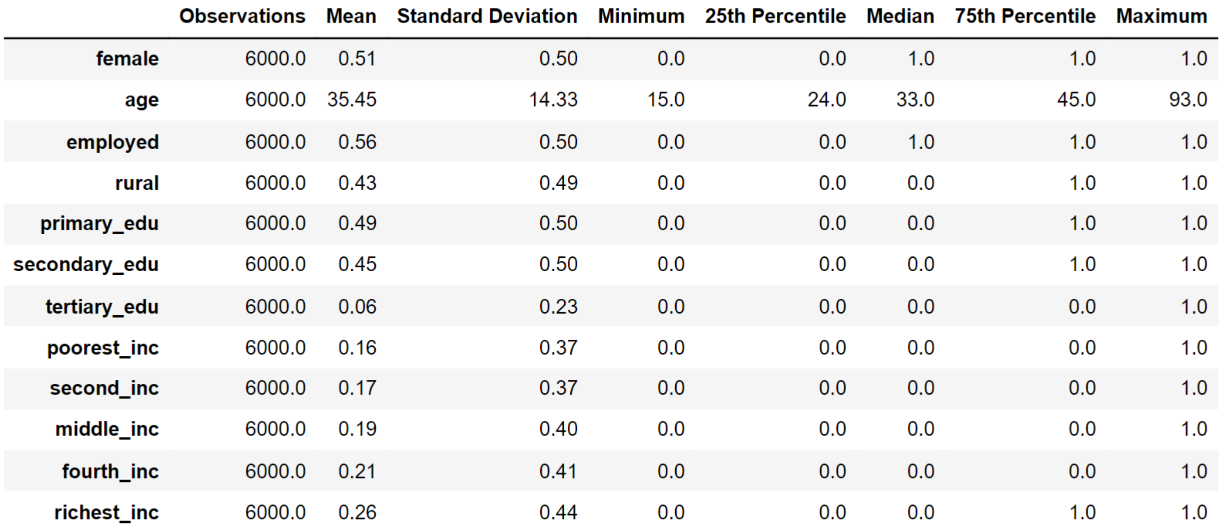
Although the models I created show results that are statistically significant at 1% level, the results might be different when we use a different regression approach that includes other variables as control variables from the dataset. Because as you could see from the different models, enriching the linear probability models with new control variables constantly increased the explaining power of the model. This suggests that the relationships between the control variables and the dependent variable align with what was hypothesized, contributing to the validity of the model.

While the findings are promising, they may be considered preliminary or fundamental. There might be room for further refinement or deeper investigation. The analysis could be further enriched by using transformation of variables, for example age to ln age, using piecewise linear splines, quadratic forms, and other interaction terms to check robustness.

Generalization and External Validity

The external validity of the analysis is also low, as the analysis was based on a survey done in 2021, 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 2 years. We also cannot be 100% certain that the pattern that our data presents is the same in the actual population - in our case the adult population of the Earth. One reason for this is that economies differ greatly in the world, as the pattern I uncovered in South Asia will be different from developed economies for e.g. European economies. Another reason for this is the number of observations; a thousand observations per economy may not be sufficient to represent the population of the economy.

Exhibit 1: Descriptive table for variables.



	Observations	Mean	Standard Deviation	Minimum	25th Percentile	Median	75th Percentile	Maximum
female	6000.0	0.51	0.50	0.0	0.0	1.0	1.0	1.0
age	6000.0	35.45	14.33	15.0	24.0	33.0	45.0	93.0
employed	6000.0	0.56	0.50	0.0	0.0	1.0	1.0	1.0
rural	6000.0	0.43	0.49	0.0	0.0	0.0	1.0	1.0
primary_edu	6000.0	0.49	0.50	0.0	0.0	0.0	1.0	1.0
secondary_edu	6000.0	0.45	0.50	0.0	0.0	0.0	1.0	1.0
tertiary_edu	6000.0	0.06	0.23	0.0	0.0	0.0	0.0	1.0
poorest_inc	6000.0	0.16	0.37	0.0	0.0	0.0	0.0	1.0
second_inc	6000.0	0.17	0.37	0.0	0.0	0.0	0.0	1.0
middle_inc	6000.0	0.19	0.40	0.0	0.0	0.0	0.0	1.0
fourth_inc	6000.0	0.21	0.41	0.0	0.0	0.0	0.0	1.0
richest_inc	6000.0	0.26	0.44	0.0	0.0	0.0	1.0	1.0

Exhibit 2: Unconditional Gender Gap – Scatterplot and Regression line

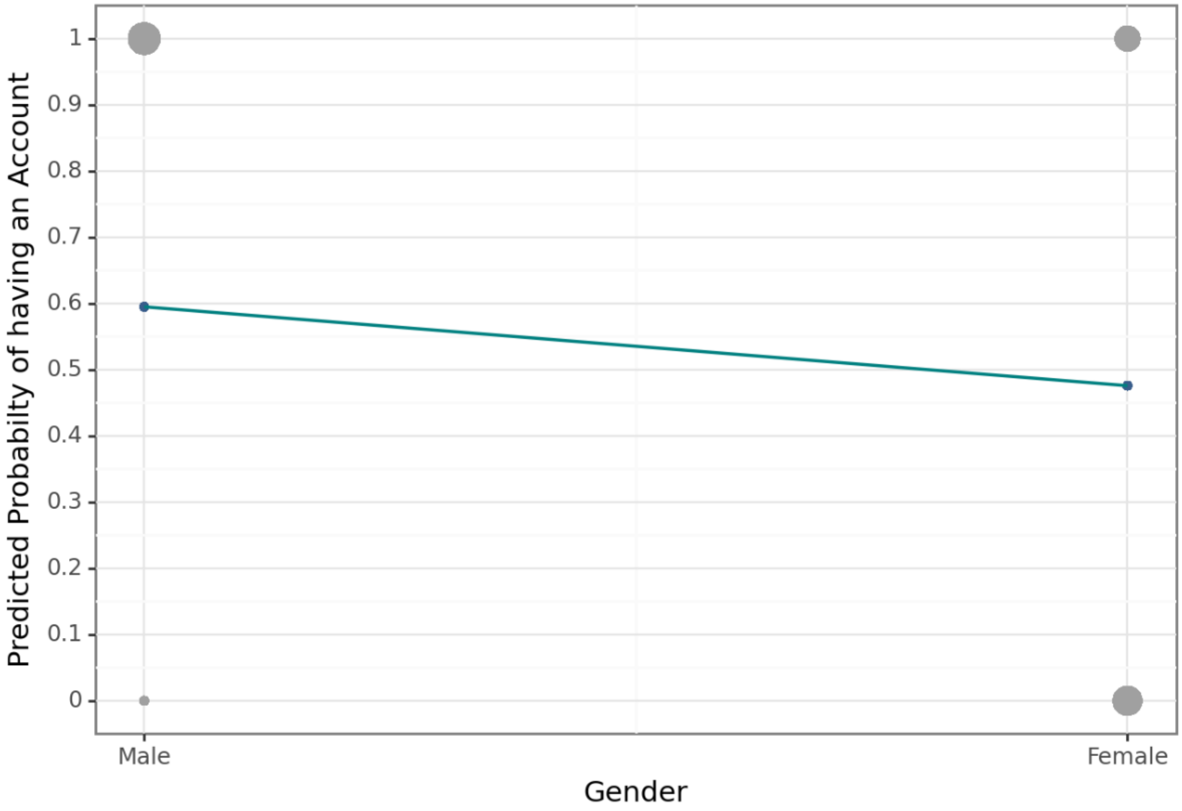
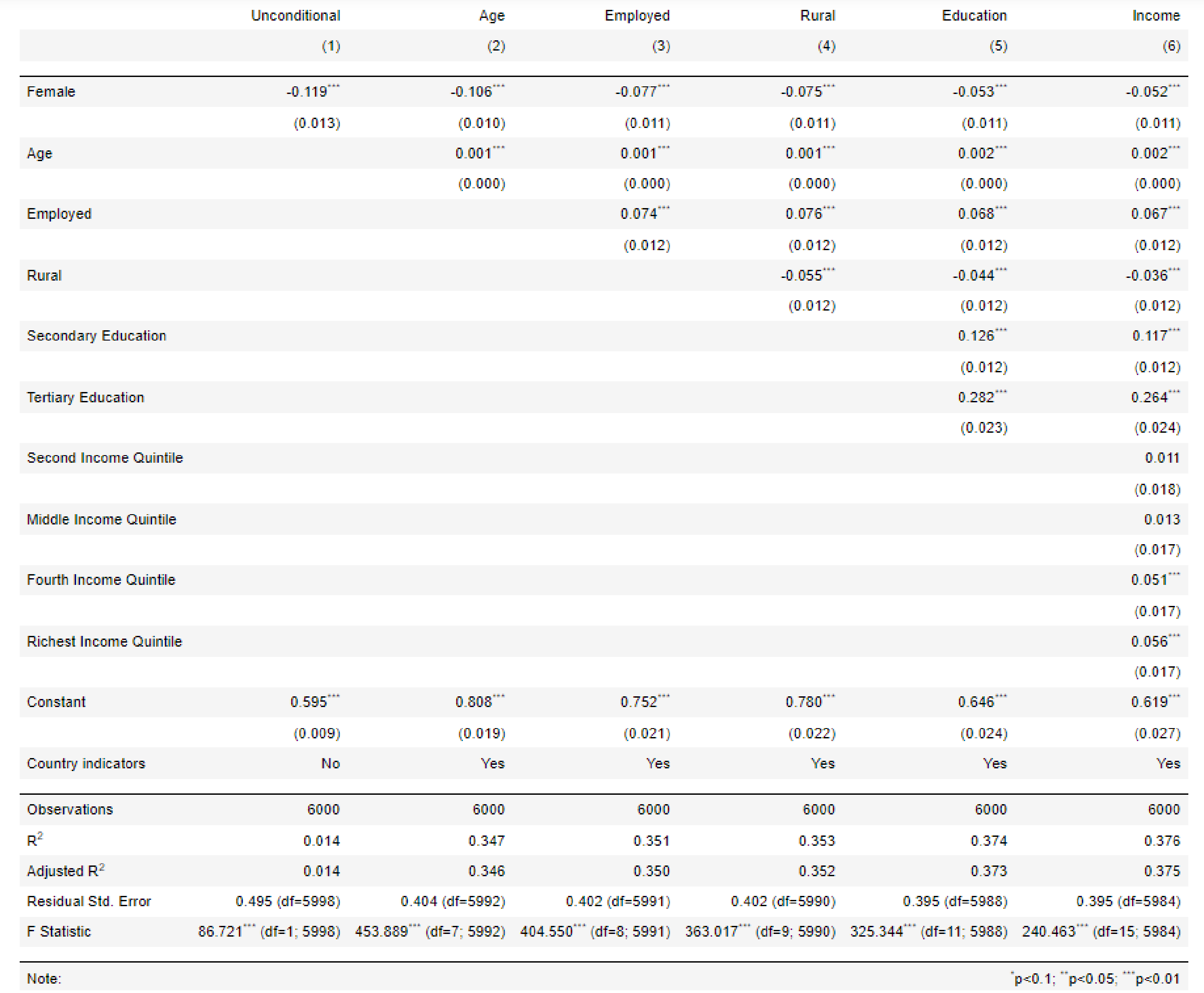
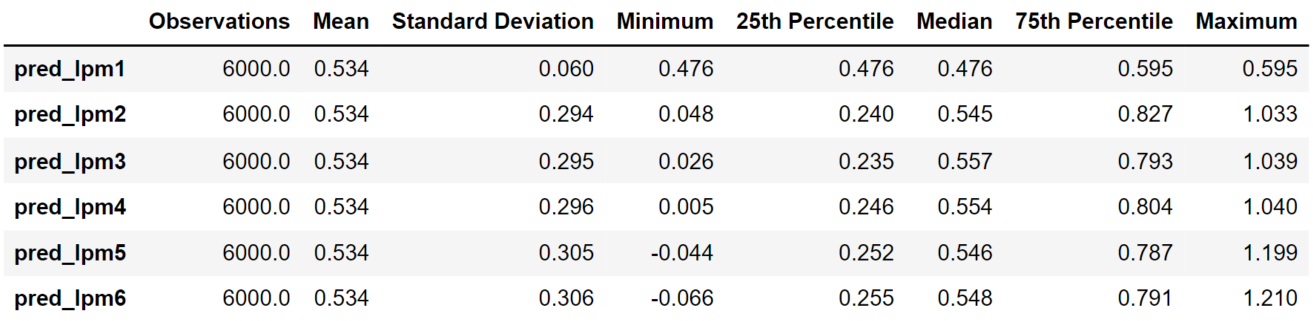


Exhibit 3: Linear Probability Models – stargazer



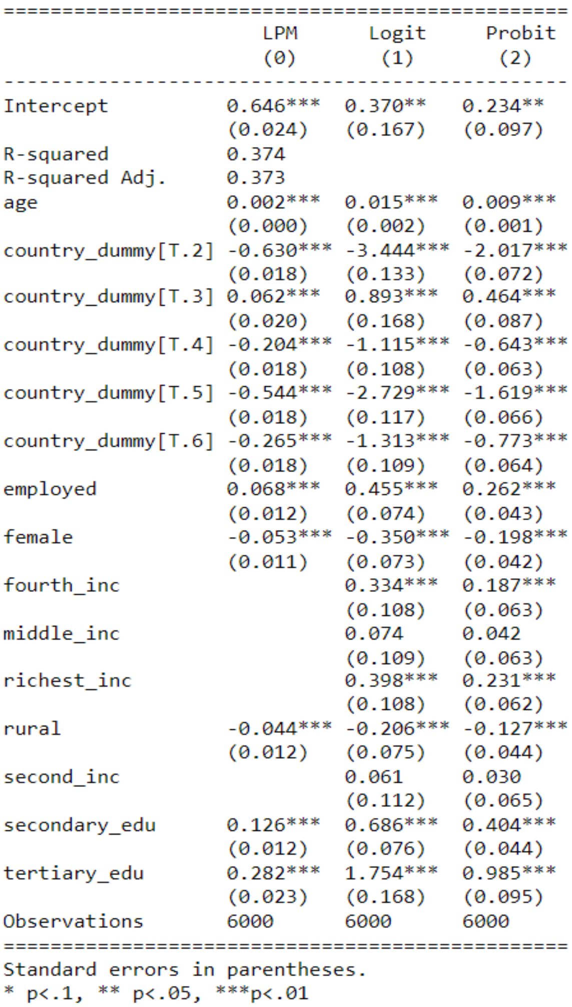
	Unconditional	Age	Employed	Rural	Education	Income
	(1)	(2)	(3)	(4)	(5)	(6)
Female	-0.119***	-0.106***	-0.077***	-0.075***	-0.053***	-0.052***
	(0.013)	(0.010)	(0.011)	(0.011)	(0.011)	(0.011)
Age		0.001***	0.001***	0.001***	0.002***	0.002***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Employed			0.074***	0.076***	0.068***	0.067***
			(0.012)	(0.012)	(0.012)	(0.012)
Rural				-0.055***	-0.044***	-0.036***
				(0.012)	(0.012)	(0.012)
Secondary Education					0.126***	0.117***
					(0.012)	(0.012)
Tertiary Education					0.282***	0.264***
					(0.023)	(0.024)
Second Income Quintile						0.011
						(0.018)
Middle Income Quintile						0.013
						(0.017)
Fourth Income Quintile						0.051***
						(0.017)
Richest Income Quintile						0.056***
						(0.017)
Constant	0.595***	0.808***	0.752***	0.780***	0.646***	0.619***
	(0.009)	(0.019)	(0.021)	(0.022)	(0.024)	(0.027)
Country indicators	No	Yes	Yes	Yes	Yes	Yes
Observations	6000	6000	6000	6000	6000	6000
R^2	0.014	0.347	0.351	0.353	0.374	0.376
Adjusted R^2	0.014	0.346	0.350	0.352	0.373	0.375
Residual Std. Error	0.495 (df=5998)	0.404 (df=5992)	0.402 (df=5991)	0.402 (df=5990)	0.395 (df=5988)	0.395 (df=5984)
F Statistic	86.721*** (df=1; 5998)	453.889*** (df=7; 5992)	404.550*** (df=8; 5991)	363.017*** (df=9; 5990)	325.344*** (df=11; 5988)	240.463*** (df=15; 5984)
Note:						*p<0.1; **p<0.05; ***p<0.01

Exhibit 4: Predicted Probabilities – All LPMs



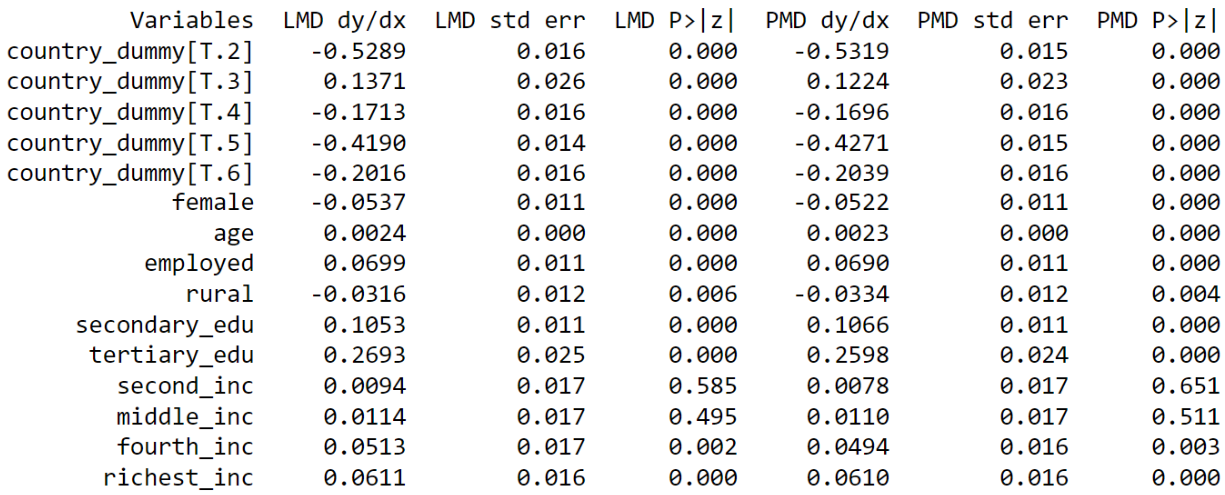
	Observations	Mean	Standard Deviation	Minimum	25th Percentile	Median	75th Percentile	Maximum
pred_lpm1	6000.0	0.534	0.060	0.476	0.476	0.476	0.595	0.595
pred_lpm2	6000.0	0.534	0.294	0.048	0.240	0.545	0.827	1.033
pred_lpm3	6000.0	0.534	0.295	0.026	0.235	0.557	0.793	1.039
pred_lpm4	6000.0	0.534	0.296	0.005	0.246	0.554	0.804	1.040
pred_lpm5	6000.0	0.534	0.305	-0.044	0.252	0.546	0.787	1.199
pred_lpm6	6000.0	0.534	0.306	-0.066	0.255	0.548	0.791	1.210

Exhibit 5: LPM, Logit and Probit Summary



```
================================================
                     LPM      Logit     Probit
                     (0)       (1)        (2)
------------------------------------------------
Intercept          0.646***  0.370**   0.234**
                   (0.024)   (0.167)   (0.097)
R-squared          0.374
R-squared Adj.     0.373
age                0.002***  0.015***  0.009***
                   (0.000)   (0.002)   (0.001)
country_dummy[T.2] -0.630*** -3.444*** -2.017***
                   (0.018)   (0.133)   (0.072)
country_dummy[T.3] 0.062***  0.893***  0.464***
                   (0.020)   (0.168)   (0.087)
country_dummy[T.4] -0.204*** -1.115*** -0.643***
                   (0.018)   (0.108)   (0.063)
country_dummy[T.5] -0.544*** -2.729*** -1.619***
                   (0.018)   (0.117)   (0.066)
country_dummy[T.6] -0.265*** -1.313*** -0.773***
                   (0.018)   (0.109)   (0.064)
employed           0.068***  0.455***  0.262***
                   (0.012)   (0.074)   (0.043)
female             -0.053*** -0.350*** -0.198***
                   (0.011)   (0.073)   (0.042)
fourth_inc                   0.334***  0.187***
                             (0.108)   (0.063)
middle_inc                   0.074     0.042
                             (0.109)   (0.063)
richest_inc                  0.398***  0.231***
                             (0.108)   (0.062)
rural              -0.044*** -0.206*** -0.127***
                   (0.012)   (0.075)   (0.044)
second_inc                   0.061     0.030
                             (0.112)   (0.065)
secondary_edu      0.126***  0.686***  0.404***
                   (0.012)   (0.076)   (0.044)
tertiary_edu       0.282***  1.754***  0.985***
                   (0.023)   (0.168)   (0.095)
Observations       6000      6000      6000
================================================
Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01
```

Exhibit 6: Logit Marginal Difference and Probit Marginal Difference



```
         Variables  LMD dy/dx  LMD std err  LMD P>|z|  PMD dy/dx  PMD std err  PMD P>|z|
country_dummy[T.2]    -0.5289        0.016      0.000    -0.5319        0.015      0.000
country_dummy[T.3]     0.1371        0.026      0.000     0.1224        0.023      0.000
country_dummy[T.4]    -0.1713        0.016      0.000    -0.1696        0.016      0.000
country_dummy[T.5]    -0.4190        0.014      0.000    -0.4271        0.015      0.000
country_dummy[T.6]    -0.2016        0.016      0.000    -0.2039        0.016      0.000
            female    -0.0537        0.011      0.000    -0.0522        0.011      0.000
               age     0.0024        0.000      0.000     0.0023        0.000      0.000
          employed     0.0699        0.011      0.000     0.0690        0.011      0.000
             rural    -0.0316        0.012      0.006    -0.0334        0.012      0.004
     secondary_edu     0.1053        0.011      0.000     0.1066        0.011      0.000
      tertiary_edu     0.2693        0.025      0.000     0.2598        0.024      0.000
        second_inc     0.0094        0.017      0.585     0.0078        0.017      0.651
        middle_inc     0.0114        0.017      0.495     0.0110        0.017      0.511
        fourth_inc     0.0513        0.017      0.002     0.0494        0.016      0.003
       richest_inc     0.0611        0.016      0.000     0.0610        0.016      0.000
```

Exhibit 7: Logit and Probit Graph

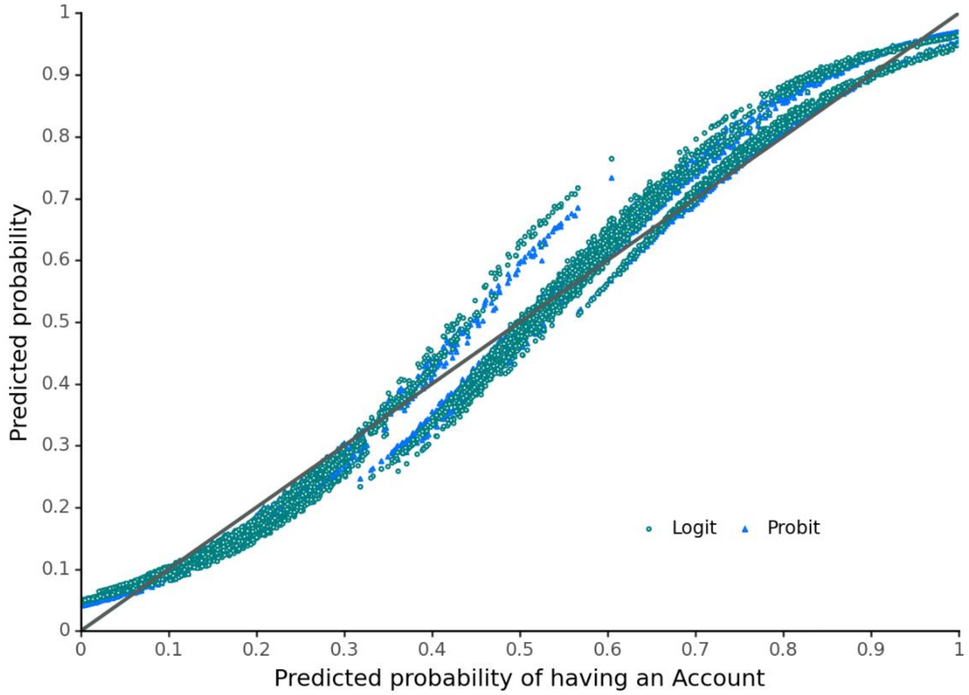
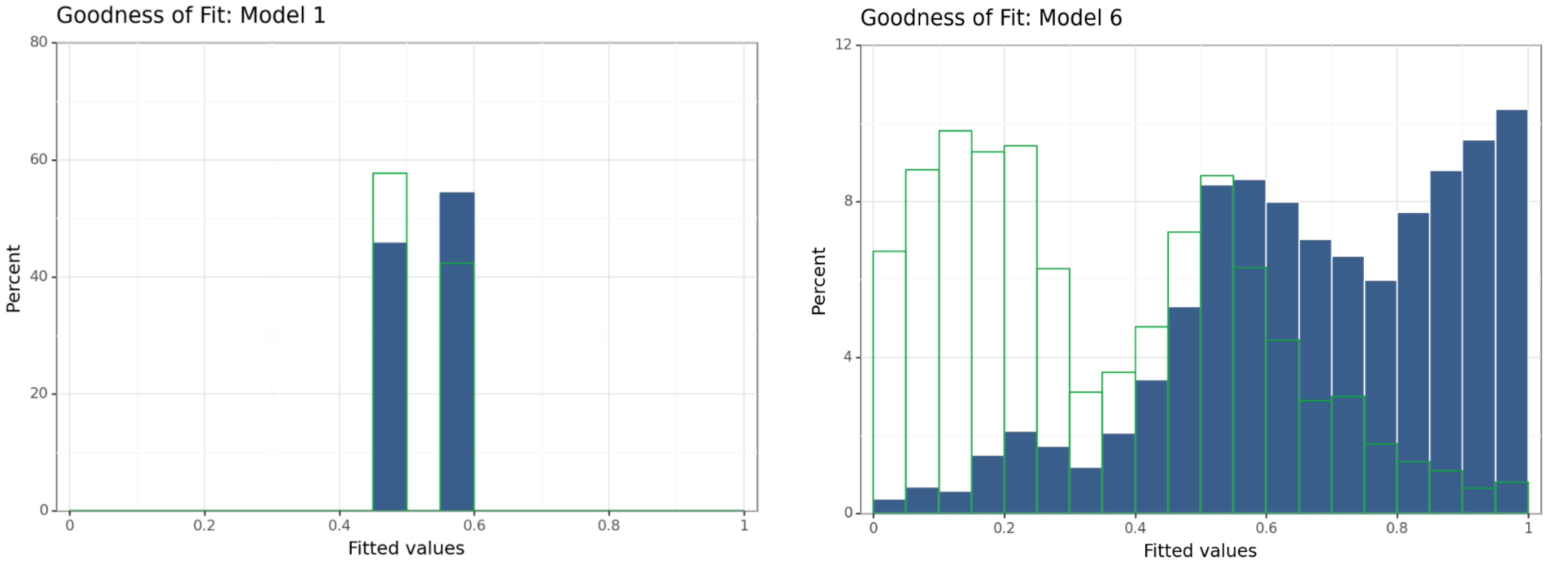


Exhibit 8: Predicted Values Distributions


Goodness of Fit: Model 1

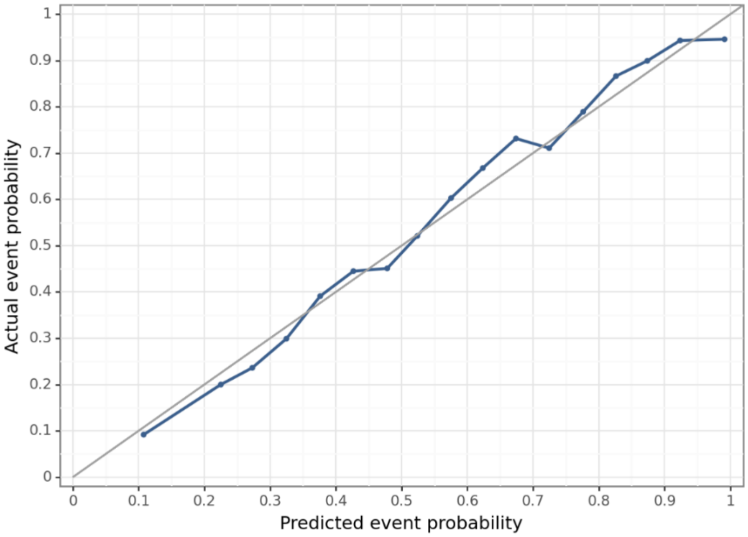
Goodness of Fit: Model 6

Exhibit 9: Goodness of Fit – Comparison

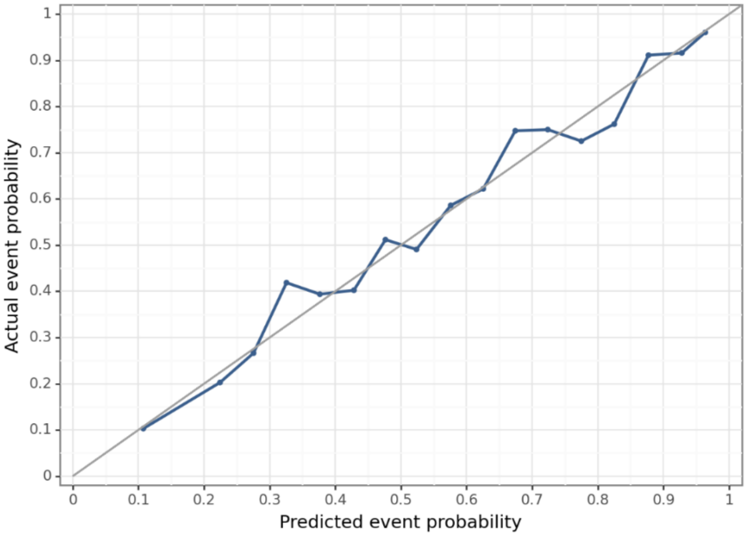
	LPM	Logit	Probit
R-squared	0.351	0.383	0.382
Brier-score	0.155	0.154	0.154
Pseudo R-squared	NaN	0.321	0.321
Log-loss	-0.543	-0.469	-0.469

Exhibit 10: Biasedness – Comparison

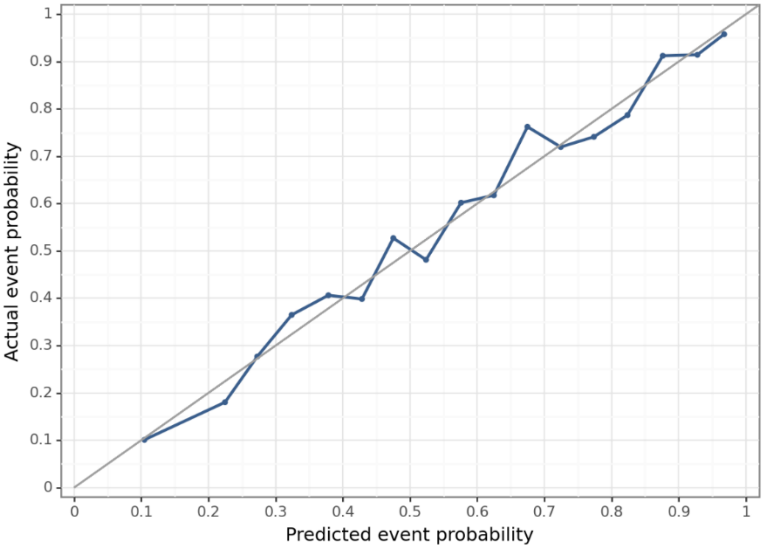
LPM Calibration Curve



Logit Calibration Curve



Probit Calibration Curve


Actual event probability
Predicted event probability