Financial Inclusion: Improving Measures in NYC

Louis Nix II

12/12/2019

Introduction

The topic of financial inclusion is of vital importance in the drive to understand the factors that determine the long-term wealth of communities. Over the last two centuries, access to the services offered by traditional banks has enabled individuals to store their excess income, utilize secure payment services, and to invest in financial instruments to build wealth. It appears only rational that individuals should maintain an account with a bank and utilize the services that have proven effective in converting income to long-term financial security. However, many individuals across developed and developing nations live without bank accounts and continue to utilize high-cost, short-term alternative sources of financing (AFS). The first step to understanding why people live without an account or choose high-cost alternatives is to quantify how many people choose to do so and where they are.

This paper builds on previous research into financial inclusion by further tuning models used for estimating unbanked and underbanked rates in the community districts of New York City. In particular, a report commissioned by the City and published in 2015 produced initial estimates of both rates for the 55 community districts of NYC.[1] Ratcliffe et al. used demographic data collected in the FDIC's biennieal 'National Survey of Unbanked and Underbanked Households' to generate rough estimates with two rounds of data from 2011 and 2013. However, Ratcliffe et al. did not publish the technical details of the predictive regression used to generate the estimates. With two more rounds of survey data published, this paper asks the following questions: How good was Ratcliffe et al.'s predictive regression and can the model be tuned further with new data or alternative models?

The hypothesis of this paper is twofold. First, the investigation retains Ratcliffe et al.'s hypothesis that demographic data can be used to predict whether or not an individual is unbanked or underbanked. Such a prediction is useful when implemented through a logistic regression and applied to relevant demographic data to estimate unbanked/underbanked rates in regions, as Ratcliffe et al. showed when predicting NYC rates. Second, this paper contends that a predictive regression should include controls for non-demographic factors that relate to bank account ownership. In particular, Ratcliffe et al. failed to include internet access, mobile phone ownership, and other relevant variables contained in the FDIC survey dataset.

The dependent variable of the investigation is financial inclusion, which measures whether an individual has a bank account as well as what sources of financing the individual has used in the last 12 months. The independent variables are demographic measures of the type of household, educational attainment, age, employment, home ownership status, citizenship, ethnicity, and income. Each of the demographic variables controlled for is a signal for systemic realtionships associated with barriers to account ownership. Educational attainment, ethnicity, type of household (i.e. headed by mother, father, or both), and citizenship represent disparities between various social groups that often exhibit different degrees of financial inclusion. Age, employment, homeownership status, and income control for economic disparities that are associated with different levels of financial education, and thus variation in financial inclusion. The full suite of independent variables should allow for accurate prediction of whether or not an individual is financially included.

In addition to the initial independent variables, the final models produced by this paper will introduce additional demographic as well as non-demographic controls associated with financial inclusion in an increasingly digitized industry. The new demographic controls will account for disability status, generational change, and whether or not Spanish is the only language spoken by the head of the household. Disability status controls for another physical barrier to financial access, while generation and laguage control for additional interface barriers. The non-demographic controls will account for smart/mobile phone ownership and internet access.

[1] Ratcliffe, Caroline, Signe-Mary McKernan, Emma Kalish, and Steven Martin. "Where Are the Unbanked and Underbanked in New York City?," (2015), 25.

The two variables will add additional information to the prediction as many individuals no longer need to leave their home to start an account, access card services, and utilize credit lines offered by traditional banks.

Data and Variables

Dataset

The dataset used for estimation of the regression was gathered by the FDIC through the 'National Survey of Unbanked and Underbanked Households.' The survey is collected biennially and was started in 2009 as part of the FDIC's attempt to understand the primary barriers to account ownership for underserved households. The first two rounds were conducted by the FDIC, after which the FDIC collected subsequent surveys in conjunction with the Census Bureau as a supplement to Census surveys. The survey is collected from over 35,000 households each round and is sampled to be representative of U.S. households. Furthermore, the surveys are collected such that one can subset for representative samples from each of the 50 states. The latest edition of the dataset contains surveys from 2009-2017, collected every other year.

Dependent Variable

The dependent variable of interest in this investigation is a three category measure of financial inclusion. As defined by the FDIC, an individual can be classified as one of the following three:[2]

- Unbanked (1): The individual does not own a bank account with a traditional bank (i.e. a bank backed by the FDIC).
- Underbanked (2):The individual does own a bank account but has utilized alterntive sources of financing in the last 12 months (e.g. payday loans or prepaid cards).
- Fully Banked (3): The individual does own a bank account and has not utilized alternative sources of financing in the last 12 months

The three categories are designed to measure whether or not individuals have access to a bank account as well as the full suite of services offered by traditional banks at sustainable interest rates. Alternative sources of financing have been highlighted as a detriment to the long-term financial health of low-income individuals. As a consequence, it is vital to measure both whether an individual has an account as well as how sustainably they manage their finances.

Unfortunately, the dataset updated the definition listing alternative financial sources that could classify an individual as underbanked in 2011 and 2013. The 2009 version of the question was only asked in that year and was labeled *hbankstat*. The 2011 and 2013 versions were asked in all subsequent years following their introduction. For the greatest analytical consistency, this paper utilizes the 2011 definition for all available years as it is the closest to the 2009 definition. The investigation will use the 2009 definition in place of the 2011 for the first year of data, which will represent a methodological defficiency in the output of the estimated regression. Both definitions are as follows, with the only difference pertaining to the inclusion of 'remittances' as a form of alternatve financing in 2011:

- 2009 AFS definition: check cashing, money order, payday loan, rent-to-own service, pawn shop loan, and refund anticipation loan.
- 2011 AFS definition: AFS: check cashing, money order, payday loan, rent-to-own service, pawn shop loan, refund anticipation loan, and remittance.

[2] UNSGSA FinTech Working Group and CCAF. (2019). Early Lessons on Regulatory Innovations to Enable Inclusive FinTech: Innovation Offices, Regulatory Sandboxes, and RegTech. Office of the UNSGSA and CCAF: New York, NY and Cambridge, UK.

Independent Variables

The independent variables of interest are primarily demographic and were selected by Ratcliffe et al. as useful predictors associated with common barriers to account ownership. Ratcliffe et al. included in the predictive regression the following controls:

- *peducgrp*: educational attainment of household head

  1. no high school diploma
  2. high school diploma
  3. some college
  4. college degree

- *hhtype*: whether house head is a single mother (the original variable was multi-categorical with more household types that Ratcliffe et al. chose to recode to binary)

  0. All other
  1. unmarried w female head

- *pnativ*: citizenship status of household head

  1. U.S.-born
  2. Foreign-born citizen
  3. Foreign- born non-citizen

- *praceeth*: race/ethnicity of the household head

  1. black
  2. hispanic
  3. asian
  4. American Indian
  5. hawaiian/pacific islander
  6. white
  7. other

- *hhincome*: household income

  1. family income less that $15k
  2. 15-30k
  3. 30-50k
  4. 50-75k
  5. at least 75k

- *pagegrp*: if household head is over age 65 (the original variable included multiple age groups but Ratcliffe et al. chose to recode to a binary focusing on homes with an elderly head vs. those without)

  0. All other
  1. 65 yrs or older

- *pempstat*: household employment

  1. employed
  2. unemployed
  3. not in labor force

- *hhtenure*: home ownership

  1. owner
  2. non-owner

Additional Controls

The first additional control investigated by this paper to be added to the final model accounts for additional physical to financial access. Physical barriers of access refer to the inability of individuals to reach access points to financial services such as bank branch locations or ATMs.

- $pdisabl\_age25to64$: whether or not an individual is age 25 to 64 and is disabled

  1. Disabled age 25-64
  2. not disabled age 25-64

The remaining variables introduced are theorized to control for interface-based barriers to access. As finance becomes increasingly digitized, users interact with services offered by banks through mobile means. Consequently, when individuals struggle to connect to services online or to utilize new technology, they may struggle to navigate financial services. I have also chosen to include a control for interface barriers in the form of languages spoken as some individuals may struggle to obtain services through English-based bank branches.

- *hintacc*: internet access

  1. has access
  2. does not have access

- *hsmphone*: smart/mobile phone access

  1. smartphone
  2. non-smartphone
  3. no mobile phone

- *huspnish*: whether spanish is the only language spoken by the household head

  0. spanish is not the only language spoken
  1. spanish is the only language spoken

- *pgen*: head of the household's generation (using CPS definition for generation: 1946-1964 for boomer, 1982-2000 for millennial, 1965-1981 for GenX, 1928-1945 for silent generation, although age 85+ is top coded in base CPS)

  1. silent or earlier generation
  2. Boomer
  3. GenX
  4. Millenial
  5. Post-millenial

Descriptive Statistics

Dependent Variable

The dependent variable $hbankstatv2$ is quite heavily skewed towards the coded response of 3, with the mean sitting well above 2.5 and the median at 3. The variance of the variable also confirms that the distribution is tightly packed around the upper end of the possible responses. From this information, the respondents are clearly skewed towards fully banked as a status of financial inclusion. However, the standard devitation is larger than one might expect, which is likely a result of the categorical construction of the variable. In other words, an unbanked individual would respond 1, which would drag the standard deviation down quite significantly per unbanked response. Regardless, the skewed nature of the variable is unavoidable since traditional bank accounts are so widely accepted and utilized in a developed economy like that of the United States.

Table 1: Summary statistics - Independent Variables

	hbankstatv2	peducgrp	hhtype	pnativ	praceeth
median	3.0000	3.0000	0.0000	1.0000	6.0000
mean	2.6719	2.8361	0.1158	1.1645	4.9507
SE.mean	0.0014	0.0023	0.0007	0.0011	0.0042
CI.mean.0.95	0.0027	0.0044	0.0014	0.0022	0.0082
var	0.3578	1.0034	0.1024	0.2455	3.4402
std.dev	0.5982	1.0017	0.3200	0.4955	1.8548
coef.var	0.2239	0.3532	2.7635	0.4255	0.3747

Table 2: Summary statistics - Independent Variables (cont.)

	hhincome	pagegrp	pempstat	hhtenure
median	3.0000	0.0000	1.0000	1.0000
mean	3.3347	0.2396	1.7385	1.3239
SE.mean	0.0037	0.0010	0.0021	0.0011
CI.mean.0.95	0.0072	0.0019	0.0042	0.0021
var	2.0258	0.1822	0.8908	0.2190
std.dev	1.4233	0.4268	0.9438	0.4680
coef.var	0.4268	1.7814	0.5429	0.3535

Table 3: Summary statistics - New Control Variables

	hintacc	hsmphone	huspnish	pdisabl_age25to64	pgen
median	1.0000	1.0000	0.0000	2.0000	2.0000
mean	1.1965	1.4295	0.0181	1.8742	2.3910
SE.mean	0.0020	0.0021	0.0003	0.0009	0.0022
CI.mean.0.95	0.0040	0.0041	0.0006	0.0017	0.0042
var	0.1579	0.4609	0.0177	0.1100	0.9161
std.dev	0.3973	0.6789	0.1332	0.3316	0.9571
coef.var	0.3321	0.4749	7.3724	0.1770	0.4003

Independent Variables

Analyzing the descriptive statistics for the explanatory variables involves much more work, by virtue of the number of variables included. The means around *peducgrp*, *praceeth*, and *pnativ* indicate that the respondents are largely white, U.S. citizens with at least a high school diploma and possibly some college education. The standard deviations around each are somewhat large, though especially *praceeth*. As a result, the dataset appears to be in line with the broader U.S. population, which is majority white with most Americans attaining at least a high school diploma and, often, some college experience. Moreover, with somewhat large standard deviations and a large variance for *praceeth* it is likely the dataset does not over represent this group.

The means of the recoded variables *pagegrp* and *hhtype* indicate that very few households have a head of house over the age of 65 and even less are classfied as 'headed by a single mother.' This paper assumes that Ratcliffe et al. expected each of these classifications to be outliers in the American population that are tied closely to very specific mechanisms generating financial non-inclusion. Consequently, it makes sense that households that fall into either of these categories are sparse and this investigation expects the outliers of this distribution to be vital to generating an accurate prediction regression later.

For economic variables, *hhincome*, *pempstat*, and *hhtenure* indicate that the average respondent head of household is an employed homeowner making between 30k-50k dollars per year. The distribution for *hhincome* is much tighter than the other two variables, indicating that homeownership is quite common among the respondents. *pempstat* and *hhincome* have larger standard deviations than one might expect. In the case of *pempstat*, the variation is likely influenced by the respondents that are not in the labor force, as the United States has maintained a roughly 62%-63% labor force participation rate in recent years.[3] For *hhincome* it appears the respondents are flatly distributed across the income distribution, with a variation of over 2 categories.

New Controls

Of the new controls to be added to the model, *hintacc* and *hsmphone* indicate that most of the respondents have a smartphone and internet access. Together, the two variables confirm that Americans generally have access to mobile banking in the form of general internet access or specifically through smartphones. The standard deviation and variance for *hintacc* are very small, indicating that nearly all of the sample households had some form of access. *hsmphone* is a bit larger, which is probably due to a not insignificant number of Americans using mobile phones without smart capabilities, though the number is still small as a proportion of the total sample.

The distribution of *huspnish* and $pdisabl\_age25-64$ indicate that nearly all of the respondents speak English to some degree and very few are disabled non-elderly. As expected, these two variables identify outliers in the sample and are predicted to be acutely related to two forms of barriers. As a result, they should help the prediction regression by increasing accuracy around outlier sample points.

Finally, *pgen* indicates that most of the respondents belong to the Boomer or GenX generations. However, *pgen* has the third largest standard deviation and variance of any of the variables. The distribution appears to be quite even accross the generations. In theory, *pgen* should help identify older individuals who are unbanked or underbanked, and who are less knowledgeable in using mobile access to utilize traditional financial services.

Recreating the Initial Model

Multiple Logistic Regression

The primary model specified by Ratcliffe et al. in the 2015 paper was a logistic regression which predicted one's financial inclusion status with the eight original demographic variables. The report does not provide

[3] "Civilian Labor Force Participation Rate." Bureau of Labor and Statistics. Accessed December 13, 2019. https://www.bls.gov/charts/employment-situation/civilian-labor-force-participation-rate.htm.

technical specifics, so this paper assumes from Ratcliffe et al.'s basic descriptions that the regression employed was a multiple logistic to predict the three category dependent variable. The equations recreated by this investigation are summarized as follows:

1. $log(P(y=2)/P(y=0)) = b_0 + b_1 peducgrp + b_2 hhtype + b_3 pnativ + b_4 praceeth + b_5 hhincome + b_6 pagegrp + b_7 pempstat + b_8 hhtenure$
2. $log(P(y=1)/P(y=0)) = b_0 + b_1 peducgrp + b_2 hhtype + b_3 pnativ + b_4 praceeth + b_5 hhincome + b_6 pagegrp + b_7 pempstat + b_8 hhtenure$

The results of the regression are quite promising as all but one category of one of the variables is highly statistically significant. It appears that "Foreign Born Non-US Citizens" as a status is not statistically related to individuals who are underbanked. Since the model is a multiple logistic regression, the potential for analying the results is limited to interpreting the individual coefficients, observing statistical signficance, and comparing to future models through the Akaike Information Criterion (AIC). The remainder of this section will focus on the former two goals while subsequent sections will utilize the AIC to determine model improvements. To some degree, parsimony will be considered when evaluating the models.

Table 4: Original Model (Exponentiated Coefficients)

	Dependent variable:	
	2	3
	(1)	(2)
peducgrp	1.547***	1.916***
	(0.015)	(0.014)
hhtype1	0.853***	0.555***
	(0.030)	(0.029)
pnativ2	1.256***	1.048
	(0.046)	(0.046)
pnativ3	0.604***	0.396***
	(0.036)	(0.036)
praceeth2	0.999	1.581***
	(0.041)	(0.041)
praceeth3	2.125***	6.226***
	(0.093)	(0.091)
praceeth4	1.056	1.310***
	(0.073)	(0.073)
praceeth5	2.224***	4.385***
	(0.200)	(0.197)
praceeth6	1.688***	4.550***
	(0.031)	(0.031)
praceeth7	421.827***	553.779***
	(0.424)	(0.424)

hhincome	1.847***	2.139***
	(0.013)	(0.012)
pagegrp1	2.070***	4.238***
	(0.039)	(0.037)
pempstat	0.776***	0.795***
	(0.015)	(0.014)
hhtenure2	0.577***	0.327***
	(0.029)	(0.027)
Constant	0.551***	0.324***
	(0.051)	(0.050)
Akaike Inf. Crit.	180,098.700	180,098.700
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 4 and all remaining tables report the results of ordered and unordered logistic regressions as exponentiated coefficients to ease interpretation of the models. The interpretation of any individual coefficient is done in net terms of all other independent variables and their categories.

The demographic variables *peducgrp*, *hhtype*, and *pagegrp* indicate very interesting and useful results. For each interpretation, the improvement in odds is relative to being unbanked. In the case of *peducgrp*, the coefficients indicate that an additional degree improves one's odds of being underbanked, as opposed to unbanked, by 54.7% and their odds of being fully banked by 91.6%, net of all other factors accounted for. The *hhtype* indicates that single-mother households are 14.7% more likely to be unbanked than underbanked and 45.5% more likely to be unbanked than fully banked. Single-mother households appear to have significant barriers to financial inclusion. *pagegrp*, recoded to identify homes with an elderly head of household, indicates that homes with a head over 65 years of age are 107% more likely to be underbanked than unbanked, and 323.8% more likely to be fully banked than unbanked. Having a head of household over 65 appears to be a strong predictor of a fully banked household.

The last two demographic variables, *pnativ* and *praceeth*, are more involved in their interpretations as they are non-binary factors. *pnativ*2 gives the odds for foreign-born U.S. citizens, indicating they are 25.6% and 4.8% more likely to be underbanked and fully banked, respectively. For foreign-born non-U.S. citizens, the relationship is much weaker in magnitude, indicating they are 60.4% more likely to be underbanked and 39.6% more likely to be fully banked than unbanked.

For *praceeth*, each instance in the table refers to a specific racial group with individuals identifying as 'black' serving as the reference group. As compared to the reference group, hispanic individuals are as likely to be underbanked as unbanked (though this coefficient is the only one that is not statistically significant in the entire table), and 58.1% more likely to be fully banked than unbanked. Individuals identifying as Native American, are 5.6% more likely to be underbanked than unbanked, but 31% more likely to be fully banked. The remaining groups, asian, hawaiian/pacific islander, white, and other exhorbantly high probabilities that they are fully banked, or at minimum, underbanked instead of being unbanked.

For the economic controls *hhincome*, *pempstat*, and *hhtenure*, the interpretations are much more efficient. *hhincome* indicates that net of all other factors, every one income band an individual moves up makes them 84.7% more likely to be underbanked than unbanked and 113.9% more likely to be fully banked than unbanked. *pempstat* indicates that for every degree less attached one is to the economy (i.e. unemployed or not in the labor force), that person becomes 22.4% more likely to be unbanked than underbanked and 20.5% more likely to be unbanked than fully banked. Lastly, *hhtenure* indicates that individuals who do not own their home are 44.9% more lkely to be unbanked than underbanked and 67.3% more likely to be unbanked than fully banked.

The AIC, which will be important in comparing to the three models to follow in the next section, is 180,098.7. When comparing to subsequent models, this paper will treat a decrease of seven or more in the AIC as an improvement over the model presented by Ratcliffe et al.

Towards an Improved Model

The process of evaluting and improving upon Ratcliffe et al.'s model involves three steps. First, this paper looks at new controls to determine if Ratcliffe et al. could have improved their model by making use of other important controls available in the FDIC dataset. In addition, Ratcliffe et al. did not specify the logistic model used, which was assumed thus far to be a non-ordinal multinomial logistic regression. This paper contends that the dependent variable is better modeled as an ordered set of categories, and therefore an ordinal logistic regression would produce a more efficient use of information. Finally, the new ordinal model is to be applied to the full suite of new controls to test whether the two changes together generate an overall improvement in the model.

Original Model with New Controls

Table 5 presents the result of the same multinomial regression used in the original model, but with the following added controls: *hsmphone*, *huspnish*, $pdisabl\_age25to64$, and *pgen*.

Table 5: Original Model w/ New Controls (Exponentiated Coefficients)

	Dependent variable:	
	2	3
	(1)	(2)
peducgrp	1.419***	1.883***
	(0.035)	(0.035)
hhtype1	0.766***	0.558***
	(0.069)	(0.069)
pnativ2	2.461***	1.652***
	(0.117)	(0.117)
pnativ3	1.245**	0.739***
	(0.097)	(0.097)
praceeth2	1.109	1.934***
	(0.099)	(0.101)
praceeth3	2.963***	8.482***
	(0.242)	(0.240)
praceeth4	1.328	1.657***
	(0.177)	(0.181)
praceeth5	4.216***	9.380***
	(0.481)	(0.478)
praceeth6	1.609***	4.483***
	(0.073)	(0.075)

praceeth7	1.000	1.000
	(0.000)	(0.000)
hhincome	1.749***	2.133***
	(0.031)	(0.030)
pagegrp1	1.000	1.000
		(0.000)
pempstat	0.801***	0.843***
	(0.039)	(0.038)
hhtenure2	0.603***	0.349***
	(0.069)	(0.067)
hintacc2	0.386***	0.437***
	(0.073)	(0.071)
hsmphone2	0.714***	0.817***
	(0.072)	(0.072)
hsmphone3	0.436***	0.685***
	(0.094)	(0.090)
huspnish1	0.624***	0.722**
	(0.150)	(0.153)
pdisabl_age25to642	0.714***	1.062
	(0.086)	(0.084)
pgen	0.681***	0.609***
	(0.044)	(0.044)
Constant	3.356***	1.229
	(0.166)	(0.164)
Akaike Inf. Crit.	34,154.450	34,154.450
Note:	*p<0.1; **p<0.05; ***p<0.01	

From the table, it is immediately clear that most of the old controls are unaffected by the addition of the new variables. However, *pnativ*2, *pnativ*3, and *praceeth*4 saw siginificant increases in the odds that individuals in each of the categories are not unbanked, net of other categories for each. *praceeth*2 also saw a large increase in the probability hispanic individuals are fully banked as opposed to unbanked.

Additionally, *praceeth*7 and *pagegrp* were both rendered insignificant and predictionless by the addition of the controls. In the case of *praceeth*7, the change is likely related to the small count of individuals identifying as 'other' for their race. The small number of data points for this group may be overwhelmed by the number of variables account for, rendering the effect of one identifying as 'other' meaningless. For *pagegrp* the change is more peculiar. The variable was highly predictive in the original model. Regardless, this investigation was unable to diagnose the issue, initially.

The new demographic variables *huspnish* and *pdisabl*25*to*64 *show very useful results. Both variables are statistically significant, with the exception of pdisabl*25*to*64 for predicting odds of identifying individuals of being fully banked. As expected, language and physical barriers are associated with individuals identifying as unbanked. Respondents who do not speak some degree of English are 37.6% more likely to be unbanked than underbanked and 27.8% more likely to be unbanked than fully banked. Non-elderly disabled individuals are 28.6% more likely to be unbanked than underbanked, but the coefficient for predicting odds of being fully banked is insignificant.

pgen also shows very interesting demographic results that run contrary to the expectation. This paper expected to find that younger individuals are more adept at mobile banking, and thus younger generations would have better odds of being fully banked. In contrast, the data shows that older generations are more established and have greater odds of being banked. In fact, for every generation younger an individual is, they are 31.9% more likely to be unbanked than underbanked and 39.1% more likely to be unbanked than fully banked.

Finally, the non-demographic variables *hintacc* and *hsmphone* show that having a smart phone or internet access drastically increases one's odds of being fully banked. *hintacc* shows that individuals without internet access are 61.4% more likely to be unbanked that underbanked and 56.3% more likely to be unbanked than fully banked. *hsmphone* shows a simlar trend in more detail. *hsmphone*2 shows that households where the head has a non-smart mobile phone are 28.6% and 18.3% more likely to be unbanked than underbanked and fully banked, respectively. Individuals without a mobile phone are more more negatively affected, with the respective rates sitting at 56.4% and 31.5%.

Despite the challenges associated with two variables losing predictive ability entirely, the added controls reduce the AIC of the model drastically to 34,154.45. Tentatively, it appears that the new controls have improved the model provided by Ratcliffe et al., though the issues with the *praceeth*7 and *pagegrp* are somewhat worrying.

New Model with Original Controls

Running the new model with the original set of variables from Ratcliffe et al.'s report generates a baseline of the ordinal logistic regression. Immediately, it is clear from table 6 that interepting the new model requires much less work.

Table 6: New Model w/ Old Original Controls(Exponentiated Coefficients)

	Dependent variable:
	hbankstatv2
peducgrp	1.392***
	(0.007)
hhtype1	0.629***
	(0.018)
pnativ2	1.614***
	(0.034)
praceeth2	1.553***
	(0.025)
praceeth3	3.385***
	(0.040)
praceeth4	1.247***
	(0.046)
praceeth5	2.589***

	(0.098)
praceeth6	3.151***
	(0.018)
praceeth7	1.836
	(0.838)
hhincome	1.333***
	(0.005)
pagegrp1	2.712***
	(0.019)
pempstat	0.909***
	(0.008)
hhtenure2	0.492***
	(0.014)
Akaike Inf. Crit.	185,031.39
Note:	*p<0.1; **p<0.05; ***p<0.01

peducgrp maintains a positive relationship with financial inclusion, indicating that another category rise in education is associated with a 39.2% chance that the head of household moves from unbanked to a status of greater financial inclusion. Again, *hhtype* shows that single-mother households face large barriers to financial inclusion, indicating that they are 37.1% more likely to be unbanked than to move to a more financial included status. *peducgrp* is also statitsically signficant and predictive again and indicates that households with an elderly head are far more likely to be financially included.

pnativ and *praceeth* show interesting results as well. *pnativ* now indicates that, net of other controls, the more detached from U.S. birth and citizenship a head of house is, the more likely they are to be banked. This could be due to foreign born heads of household being more financially established before moving to the U.S. or considering citizenship, resutling in a selection bias. *praceeth* shows that every racial group, net of other controls, is far more likely to be underbanked or fully banked as opposed to unbanked, when compared to the reference group of individuals identifying as black.

Collectively, *hhincome*, *pempstat*, and *hhtenure* indicate that employment, home ownership, and greater income are associated with financial inclusion. In particular, moving up one income band and owning one's own home predict large percentage changes in the probability that one is financial included. The relationship is weaker, though still statistically significant for *pempstat*.

While the new model type does generate statistically significant results across the board (with the notable exception of the coefficient for individuals identifying their race as 'other'), the AIC is roughly 5,000 higher than the original model with the same variables. It appears that Ratcliffe et al.'s model may not have suffered from misspecification of the model.

New Model with New Controls

In evaluating the new ordinal model with the new controls, most of the relationships observed in the three previous models are maintained. However, table 7 shows three interesting developments.

Table 7: New Model w/ New Controls (Exponentiated Coefficients)

	Dependent variable:
	hbankstatv2
peducgrp	1.429***
	(0.017)
hhtype1	0.667***
	(0.040)
pnativ2	1.342***
	(0.076)
praceeth2	1.779***
	(0.057)
praceeth3	3.233***
	(0.087)
praceeth4	1.350***
	(0.106)
praceeth5	3.333***
	(0.215)
praceeth6	3.099***
	(0.042)
hhincome	1.351***
	(0.013)
pempstat	0.947***
	(0.020)
hhtenure2	0.516***
	(0.032)
hintacc2	0.647***
	(0.044)
hsmphone2	1.056
	(0.037)
hsmphone3	1.052
	(0.055)
huspnish1	0.740***
	(0.103)
pdisabl_age25to642	1.306***
	(0.047)
pgen	0.806***
	(0.022)

Akaike Inf. Crit.	35,432.26
Note:	*p<0.1; **p<0.05; ***p<0.01

The first two changes come as a result of the package used to evaluate the ordinal logistic model. When running the model, the package automatically dropped *praceeth*7 (the race/ethnicity code for individuals identifying as 'other') and *pagegrp* as rank deficient. In both cases, it appears that the lack of datapoints from respondents identifying to each of the two categories rendered them inefficient when adding the extra controls. As a result, the new model appears to have handled the issue more effectively than the previous unordered logistic regression.

The third change is in statistical significance as both *hsmphone* is no longer statistically significant for any of the subcategories. *intacc* has remained highly statistically significant with a large magnitude in change indicated by the coefficient. As a result, this paper concludes that non-demographic variables are important indicators to be used in the prediction, but *hintacc* may be the only non-demographic control necessary when utilizing an ordinal logistic regression.

The AIC on the model also gives suprising results as the added controls significantly improve the ordinal logistic regression, but the model as a whole lags behing the unordered logistic regression with the full suite of new variables. The AIC is roughly ~1,300 higher than the unordered logit with all controls. However, parsimony is a factor to be considered. Utilizing and interpreting the ordinal logistic regression is more efficient than the undordered. The unordered logistic regression also handles rank deficiency in the data more effectively.

Conclusion

The investigation presented in this paper accomplished two goals. First, the paper recreated the model developed by Ratcliffe et al. based on the limited technical information presented in their 2015 paper and with updated data. In particular, they appear to have used a multiple logistic regression in which a three category measure of financial inclusion was regressed on eight demographic predictors associated with barriers to financial inclusion. The model performed quite well, producing highly statistically significant coefficients across the board.

The investigation also sought to improve upon Ratfcliffe et al.'s model through investigating additional control variables in the FDIC dataset as well as an alternative theory for modeling the dependent variable. In the case of both the multinomial and ordinal modeling of the dependent variable, adding new controls available in the FDIC improved the AIC score vastly. The primary variables driving the improvement were measures of internet access, disability, linguistic barriers, and which generation the head of household belonged to.

On the other hand, changing the modeling of the dependent variable to ordinal did not improve the AIC for the original set of variables or the new suite with added controls. The improvement introduced by the ordinal model was primarily in parsimony as it handled rank deficiency better and offered more succint interpretation.

In sum, Ratcliffe et al. produced a useful and generally accurate model. However, while Ratcliffe et al. controlled for many variables and specified their model correctly, they failed to make use of a small number of highly predictive variables in the FDIC dataset that were available to them. When choosing a final model, the ordinal logistic regression is likely better for the simple fact of parsimony, though the original multinomial model is more efficient in its use of information provided by the dataset.