Erasmus
School of
Economics

ERASMUS UNIVERSITEIT ROTTERDAM

Master thesis Economics and Business
Specialization: MSc Economic Policy

Analysis and hierarchy of the determining factors of mammography participation for American women from 1986 to 2016.

Marie Guettier: 559019

Supervisor: Professor Vladimir Karamychev

First reader: TBD

Date of final version: TBD

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam

Abstract

To this date, mammography has been the major screening method proven to significantly reduce mortality from breast cancer. Empirically, the literature mostly relies on odds ratio to analyze relevant factors of mammography's participation prediction. In this thesis, we aim to go further into this kind of analysis, in establishing a hierarchy of these important factors using the Lasso and Classification And Regression Tree (CART) method. Our main finding suggests that being in the 55-74 age group and having at least one child are the strongest predictors of mammography uptake. Further, earning a low income and having done a Clinical Breast Examination (CBE) are the most discouraging variables of mammography participation. Shedding a light on the socio-demographical variables influencing women to get screened is considered of relevant importance in designing and orientating future awareness campaigns and public policies.

Keywords: preventive health behaviors; mammography participation; Lasso model; CART regression; Random Utility Model; Income and Price effects.

Acknowledgements

To my supervisor, Professor Vladimir A. Karamychev, for his guidance, patience, and benevolence.

To Fateh Belaid, for his most precious advice and inspirations.

To my family, for their unconditional support along my study journey.

As a tribute to a beloved family member that died from breast cancer and inspired my research, career, and life.

Table of Content

I. Introduction

The World Health Organization counted up to 2.3 million deaths from breast cancer in 2020. It is the world’s most prevalent cancer, putting women above 40 years old at high risk. Although we are aware of several factors that might trigger an early development (obesity, tobacco use, reproductive history), many unanswered questions remain at this day regarding this disease that has been affecting women since time immemorial. Since its discovery in 1913, mammography is the major screening method proven to significantly reduce mortality from breast cancer for women over 40 or 50 years old (Leitch et al., 1997). Practically, it involves taking an x-ray picture of both breasts, allowing detection of nonpalpable or visible tumors. This technique, although debated and criticized due to risks such as harmful radiation exposure, over-diagnosis, and false-negative results, has proven to be of constant effectiveness, associated with a significantly increased mortality reduction starting for women aged 40-49 years old (Hendrick et al., 1997).

This thesis is aimed at analysing and establishing a hierarchy of the factors determining American women’s mammography participation. One should note that the participation decision is based on physician’s recommendations, national screening programs, as well as spontaneous screening decisions. Rare are the studies trying to establish a hierarchy of these agents, allowing us to quantify their respective magnitudes of importance, as they usually establish a mere enumeration. This thesis is aimed at filling this gap. We will add to the usual odds ratio interpretation the modelling of a Classification And Regression Tree (CART) and a Lasso model to prioritize the factors of importance in mammography decision. We will study the implications of our results in the light of the Income and Price Effects and of the Random Utility Model (RUM). The level of utility derived by the variables of interest might be critical for policymakers to design appropriate health policies, allowing them ultimately to save costs by prioritizing their recommendation target. The expected new insight gained by appling this methodology is a deepened and more precise picture of the mammography’s participation factors.

In an attempt to establish a hierarchy, extend and confirm the current records of women’s characteristics going for breast cancer screenings, our main research question will be defined by the following:

Research question: "What are the factors and the roles they play in women's mammography participation?" Several sub-questions and hypotheses will be considered to answer it.

Sub-question 1: Does income have a significant impact on the propensity of women to get screened?

As we will see in the literature review, although the impact of income level on general health conditions has been clearly depicted, its influence on preventive health care behavior is much less explicit. We will try to understand the influence of high- and low-income levels and their respective weights in mammography decisions. We will first classify both levels via the interpretation of the odds ratio provided by logistic regression, comparing them to the middle-income level. The results will be specified using the Lasso variables selection method. We will compare its outcome with the Classification And Regression Tree (CART) model for variables prediction. In the discussion, we will analyze the end results in light of the income and price effects.

Sub-question 2: What is the most prevalent age at which women start screening?

Medical institutions are still debating what would be the most cost-effective age for women to start screening. From our large and recent sample, we would like to evaluate when women started to feel the need to start screening via classification of the different age groups of importance. This classification will first be done via the interpretation of the odds ratio provided by a logistic regression. The results will be specified using the Lasso variables selection method. We will compare its outcome with the CART model for variables prediction.

Sub-question 3: Does Clinical Breast Examination (CBE) impact future mammography intake ?

While CBE effect on tumors detection and mortality is debated, the literature still contains gaps about its influence on subsequent mammography participation. The result provided by the odds ratio will indicate the positive/negative nature of CBE's impact on mammography participation. Then we'll try to evaluate its hierarchical position among other variables by performing the Lasso and CART models.

Sub-question 4: Does having at least one child impact screening behaviors ?

So far, the impact of having at least one child has been mostly overlooked in the prediction of preventive screening behaviors. In order to accurately try to answer this question, we will add to the performance of the Lasso and CART model, the elaboration of two separated logistic regression models. They will inform us more specifically on the relative importance of this variable and its potential Omitted Variable Bias (OVB).

This thesis will start with a review of the relevant literature, touching upon the preventive behavior classification techniques used so far, the optimal screening age, the impact of income level, CBE, and past pregnancy effect on mammography decision. Following this section, we'll keep on with a description of the data gathered and their preparation for our analysis. Then we will dive into the methodology used, including the statistical techniques and the various models we have drawn. We will display the results and open a discussion to put them in the light of diverse economic theories, as well as answer our research questions, highlighting the limitations of this study and suggestions for future research. Finally, the last section will provide the reader with a conclusion of our findings.

II. Literature Review

Our literature review will follow a thematic structure while comparing the methods being used across studies. Here we identify the themes, controversies, and gaps that the current literature contains.

2.1 Hierarchy and classification of factors determining mammography's participation

The effectiveness of mammography on breast cancer incidence has been demonstrated over the years through the benefits of its frequency. Since its introduction in the US, the number of early-stage breast cancer detected has doubled every year. The proportion of women touched by advanced-stage cancer decreased by 8% (Bleyer & Welch, 2012). Ultimately, mammograms have been proven to reduce breast cancer mortality in a large number of randomized studies (Gabe & Duffy, 2005).

Therefore, identifying the main factors influencing women to get screened is of crucial importance. We would like to explore the different techniques (both qualitative and quantitative) employed to analyze and prioritize them according to their weight in women's decision-making.

2.1.1 Qualitative framework

Surveys and weight attribution represent a straightforward way to establish a hierarchy of mammography's determinants. Salazar (1996) used a two-stage decision model in Hispanic communities: in the first stage, he interviewed women asking what could be the positive/negative factors that could affect their mammography decision. In a second stage, he asked the participants to attribute weights to these factors, for him to draw a hierarchy.

Also, the use of theoretical frameworks is commonly used to analyze psychological determinants, like beliefs and intentions that require a deeper introspection from the population. Secginli and Nahcivan (2005) use the Health Belief Model (HBM) to evaluate the use rate of mammography among Turkish women. The HBM is a psychological model commonly used to highlight health beliefs and behaviors (Rosenstock, 1974). In an attempt to explain preventive health behaviors, it suggests that the latter are based on four axes: (a) perceived susceptibility (perceived personal vulnerability to or subjective risk of a health condition), (b) perceived seriousness (perceived personal harm of the condition), (c) perceived benefits (perceived positive attributes of an action), (d) perceived barriers (perceived negative aspects related to an action), (Champion, 1993). Using this framework to set up a questionnaire and collect data allowed the author to get an overview of participants' beliefs and opinions on mammography intentions. As a result, the low rate of mammography participation among Turkish women (only 20% of women aged 40 years and above) was linked to low perceived benefits, seriousness, and motivation. These suggestions about health beliefs can then be used to orientate future awareness campaigns and policies, motivating women to participate in future national breast cancer screening programs.

In a similar logic, Lechner et al, (1997) tried to evaluate the participation factor between the first and second rounds of a mammography screening program. The questionnaire received by the Dutch population was based on the ASE model (Attitude – Social influence – self-Efficacy). Evaluating how participants in the first round and the second round diverged, the result was that they differed on all of the ASE determinants.

2.1.2 Quantitative framework

Empirical methods can be used to assess predictors of demographic and Socio-Economic Status (SES) characteristics of mammography participation. The most common one is the use of logistic regression, using its related odds ratio to establish factors' priority. Akinyemiju (2012) found that women belonging to a middle SES and living in urban areas had lower odds of receiving a mammography than the ones belonging to the same SES but living in rural areas. Secginli and Nahcivan (2005) were able to assess the significant impact of health insurance status, the location of gynecologists, and others' feedback as key determinants regarding mammography participation.

Also, many studies have based the ranking of their predictors on simple percentage estimations from their sample. For instance, the percentage of past mammography reports per factor (Freeman & Chu, 2005) or the percentage of adherents (Freitas et al, 2012). Einav et al, (2020) have divided their American women population in terms of compliers, always-takers, and never-takers. They calculated the mean characteristic of each variable using the regression coefficient of a simple probit function. The result indicated that a strong predictor of mammography compliance was the past intake of the flu shot. In addition, lower spontaneous health care spending was strongly correlated with the never-taker status. Interestingly, they found no difference between compliers and never-takers on non-medical preventive behaviors (such as alcohol consumption, seatbelt use) and basic demographics.

2.2 Impact of income level on mammography participation.

Higher income has proven to be a leading factor to better health, where a twofold increase of income is associated with a similar effect on health (Ecob & Smith, 1999). But controversies remain regarding the impact of income level on preventive behaviors such as mammography, in particular for developed countries and women not belonging to minorities. Calle et al (1993) showed the under-participation in mammography programs by women living under the poverty level (income below 11 000$ for 4 people family), where 80% of them never had a mammography. Conversely, in the group of women from high-income level households, only 50% of them never had a mammography. Although the literature seems to generally agree on that first point, the second one seems to be much more debated. Gathirua-Mwangi et al (2018) found that being explicitly recommended by a physician to get a mammography predicts women's adherence only in the low-income group, suggesting that high-income women tend not to follow it. Burns et al, (1996) found high-income quintile to have little effect on mammography use for white women. From our sample, we will try to confirm the veracity of the low-income participation trend, and determine the one of the high-income group. In particular, we will test the following hypothesis:

<u>*Hypothesis 1 (H1):*</u> *Income level has no effect on mammography intake.*

2.3 Mammography screening age.

In 2009, the US Preventive Task Force (2009) recommended women start screening at age 50, after publishing experimental data failing to prove significant mortality benefits from screening women in their 40s. The Affordable Care Act explicitly supported that decision by encouraging insurers to refuse financial coverage for younger women (Einav et al, 2020). Soon after this announcement, the American Cancer Society (2009) expressed its disagreement, reinforcing its urge for women to start annual screening from age 40 (Arleo et al, 2017). The National Cancer Institute supported that statement, advising women to even start biennial screening from that age. Science and research encouraged it by quantifying a 15% mortality reduction from early screening for women aged 40-49 years old (Moss et al, 2015).

As one can see, debates remain in the literature in estimating an optimal first screening age. Most studies base their conclusions on the implied mortality rate. However, although mortality should be viewed as one of the most relevant variables to consider early-stage screening, it may not be the only one. Quality-Adjusted Life-Years (QALYs) are often used to measure the costs of a medical technique employed in terms of the number of extra years it gives to people, adjusted for quality, and attributing more weight to better years (Broome, 1993). Regardless of the mortality or recovery rate, many women might care about the quality of the remaining years to live. Considering the modern world we live in, its associated new lifestyle, and environmental challenges (large consumption of processed food, quantity of harmful waves, pollution, etc.), our bodies have become more vulnerable. Women may have felt the urge to adopt preventive health behavior earlier than they did in the past.
From that statement, we will test the following hypothesis:

Hypothesis 2 (H2): Age has no effect on screening decision.

2.4 Impact of Clinical Breast Examination (CBE) on mammography participation.

CBE is defined as the observation of the breast by a health care professional, and is usually performed during regular gynecological and general physician visits. We want to know whether having done a CBE, leads women to perform less mammograms afterwards (consciously or unconsciously). CBE induces fewer financial and psychological costs to the patients compared to mammograms (lower rate of false-positives and over-diagnosis), (Jatoi, 2011). But it might also be a less reliable detection method as it depends on the practice and training of the physician. According to Fletcher et al, (1985) 40% of them failed to perform CBE following a systematic search pattern. However, no reliable evaluation has demonstrated the sufficiency of CBE alone in terms of mortality or detection rate, as it is the least studied breast cancer screening technique (Kearney & Murray, 2009). Women might think that this test is sufficient in itself (which might not be the case, according to White et al, (1993) as the combination of CBE and mammography resulted in a 30% reduction in mortality for women aged 40-64), and could allow them to save costs on other, more expensive cares like mammography. Following this conjecture, we will test the following hypothesis:

Hypothesis 3 (H3):CBE has no effect on women's decision to perform a mammography.

2.5 Impact of having at least one child on mammography participation.

Various studies have demonstrated that pregnancy was responsible for a 10-30% decrease in breast cancer risk (Nechuta et al, 2010). It has become a common belief, even in educated minds, that pregnancy would decrease women's chances to develop breast cancer later in life as well. It would discourage many women to start mammograms at an early age, especially if they are not insured. But it might also be the case that having had children, women got more aware of cancer risks, predispositions, and need for prevention (as they necessarily engaged more with medical and physician interactions), increasing their propensity to get screened. Until now, the literature did not present enough studies to be able to draw definite conclusions from any of these potential scenarios. In this thesis, we would like to test this hypothesis:

Hypothesis 4 (H4): Having had at least one child has no effect on mammography participation.

III. Data description

3.1 Data collection, randomization, and studied population

Our dataset was obtained from the Behavioral Risk Factor Surveillance System Survey (BRFSS). It is a collaborative project between each state of the US, and the Centers for Disease Control and Prevention (CDC). It is a self-reported data collection of several preventive health-related behaviors and demographics of 4,711,434 American resident women of different age groups, that rather decided to get at least one mammography in their life, or that did not. The data was collected from year 1987 to 2016. Due to the large sample size originally used, observations containing missing values were dropped, as well as outliers, leaving us with a final sample of 2,300,749 observed American women, including a total of 23 variables. Note that the experiment was randomized by a telephone-based survey sampling, where only women in the adult population (18 years or older) were surveyed. The general questionnaire that allowed the creation of this dataset was made of 3 parts: first, the core section (questions about general health-related behaviors, such as exercising routine, alcohol consumption, etc.) Second, the optional modules (set of questions regrouping specific topics, like flu shot intake, potential CBE made, etc.). And lastly, state-added questions. The core and optional portion of the questionnaire (that can be found in the References part), made the phone calls last for an average of 18 minutes, where potentially added state-specific questions required 5 to 10 more minutes. Participants did not receive any material compensation for their contribution. One should note that the current data is a cross-sectional study, where we decided to study the population within a specific time interval. Conversely to longitudinal studies (studying the characteristics' trends of a group of people over an extended period), cross sectional studies are aimed at observing the characteristics of a population at a particular point in time. Therefore, we cannot exclude the potential change in the characteristics observed over the course of this study. However, our goal is to highlight the global nature of the relationship present across different variables, rather than focusing on their potential evolution.

3.2 Dependent variable

We will focus our interest on the binary outcome variable "hadmamyes" that describes whether women have gone for a mammography or not. In particular:

$$y = \begin{cases} 1, & \text{if the women did receive a mammography before,} \\ 0, & \text{if the women did not receive a mammography before.} \end{cases}$$

where “did receive a mammography before” is the target category, and where “did not receive a mammography before” is the reference (baseline) category (Ranganathan et al, 2017).

3.3 Independent Variables

In the next sub-section can be found a detailed summary statistic containing the independent variables of interest used in our study. In order to simplify our predictive models and our interpretation, we created several dichotomous variables. First, the “employed” variable, when equal to 1, gathers all the women employed for wage compensation, and when equal to 0, all the women that were unable or that refused to work. Second, we created three categorical variables to study the impact of the income level, described as follows: high-income level was considered if the annual women’s personnal revenue was above $100,000. Middle-income revenue was characterized by an annual income level between $100,000 and $50,000. Low-income level corresponds to annual revenue of $50,000 or less. Throughout our experiment, we tool the middle-income level as the reference category. These income thresholds are defined by the Pew Research Center. For our CART model, we’ve set a number of drinks threshold, considering a woman to be a heavy drinker if her consumption goes beyond 7 drinks a week (as defined by the National Institute on Alcohol Abuse and Alcoholism (NIAAA)). Moreover, the following variables have been purposefully transformed into dummy variables: the age groups, being married, having had a CBE (“hadprofexam”), having at least one child…

3.4 Summary statistics

In Table 1 (Appendix), can be found the summary and description of our main variables of interest, as we conducted a between-subjects research design.

IV. Methods

4.1 Data Collection

The data set used in this thesis corresponds to an experiment of the Behavioral Risk Factor Surveillance System Survey (BRFSS). For the logistic regression, we used the Stata software. For the CART model, we used the R software.

4.2 Logistic Regression

4.2.1 Introduction to Logistic Regression

Logistic regression analysis is a statistical technique that evaluates the relationship between various predictor variables, that can rather be categorical or continuous, and a binary outcome variable that will reflect here whether a woman received a mammography (“hadmamyes”).

Logistic regression does not use Ordinary Least Square (OLS) for predictors estimation, but rather Maximum Likelihood Estimation (MLE). The logistic regression technique does not require the dependent variable to be normally distributed and may therefore be preferred to linear regression.

Our model is defined by the following equation:

$$g\big(E(y)\big) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n + \varepsilon \qquad (1)$$

- Where g is the logit function and $g(y)$ is the link function, that establishes both the probability of success p, (or the probability that a woman has received a mammography) and the probability of failure $(1-p)$, (or the probability that a woman did not receive a mammography).
- Also, $E(y)$ is the expectation of past mammography intake.
- β_n are the parameters of the model.
- Finally, $\{x_1 + x_2 + \cdots + x_n\}$ represent the set of predictors listed above in part 3.3.

Since a probability is required to be always positive, we will formulate the above equation in exponential form to represent the probability of success:

$$p = \frac{\exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots\beta_n x_n+\varepsilon)}{\exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots\beta_n x_n+\varepsilon)+1} \qquad (2)$$

$$p = \frac{e^y}{1+e^y} \qquad (3)$$

Similarly, the probability of failure can be written:

$$q(x) = 1 - p = 1 - \left(\frac{e^y}{1+e^y}\right) \qquad (4)$$

By breaking down p and q :

$$\frac{p}{1-p} = e^y \qquad (5)$$

$$\log\left(\frac{p}{1-p}\right) = y \qquad (6)$$

Substituting y results in:

$$M = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n + \varepsilon \qquad (7)$$

Here $\frac{p}{1-p}$ represents the odds ratios. In our case, it is the ratio of the probability that a woman had a mammography to the probability that she did not. It is important to notice that the interpretation of the obtained coefficients differs according to whether we analyze the simple logit coefficients or the odds ratio. Indeed, the latter does not represent the predicted change in probability of the targeted group per unit increase of a particular predictor, but rather the predicted change in log odds per unit increase of a given predictor. However, we can make a usual interpretation of the positive (negative) regression coefficient, referring to the likelihood of falling into the target group increases (decreases), resulting from a change on a particular predictor. The interpretation of the odds ratio also differs when compared to coefficients. For now, in order to make the interpretation of results easier, we will focus the analysis of our results on the coefficients delivered by the logistic regression (Table 2), rather than on the odds ratios.

Note that the significance level used in our analysis is $\propto = 0.05$.

4.2.2 Two Models

In this analysis, we are going to run two different models to check for Omitted Variable Bias (OVB) of the “children” variable: the first one excludes the “children” variable, while the second one includes it. Overall, both logistic regression models are going to help us answer our research questions as they will include the impact of CBE on mammography intake, the most prevalent age groups as well as income and price effects.

4.2.3 Least Absolute Shrinkage and Selection Operator (Lasso)

Lasso allows us to select which covariates are important predictors of our dependent outcome variable (having done a mammography). This analysis can be used to compare the predictions that we will obtain with the CART Model in the next section, which will also be used to establish a classification of covariates’ power.

As we have seen in equation (8), the value of our logistic coefficients of interest β_0 and β_1 can be found by minimizing the log-likelihood ratio:

$$M = \log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n + \varepsilon \qquad (8)$$

Lasso is going to add a penalty term defined by $|\beta_1|$, where $|\beta_1|$ is a vector containing as many components as there are predictors. Therefore, our minimization equation becomes:

$$M + \lambda \Sigma \left|\beta_j\right| \qquad (9)$$

Where λ represents the shrinkage parameter. It is chosen so that the final out of sample error is minimized (this restriction does not apply to the intercept β_0). We computed Lasso to retain 3 different models: the first one selects the value of λ having the smallest cross-validation mean, the second selects the one having the lowest BIC, and the third one having the lowest cross-validation mean prediction error. The output of this analysis will provide a ranking of the most influential predictors of mammography participation.

4.2.4 Note on OLS regression model and odds ratio

Although we are dealing with a dichotomous outcome variable, an OLS regression might deliver a satisfactory outcome as well. The literature might be in favor of using it instead of logistic regression, mainly for interpretational reasons. Indeed, interpreting the odds ratio might be tricky and less intuitive than the coefficients obtained with OLS (Von Hippel, 2015). We decided to focus exclusively on the coefficients delivered by the logistic regression for the general interpretation of the results, which follow the very same interpretation as an OLS. Therefore, as we rely on logistic regression coefficient and not odd ratios, performing an OLS would have been both methodologically and mathematically redundant.

However, we will need to refer to the probabilities given by the odds ratios, later on, to analyze our logit model in the light of the Random Utility Model (part 6.1.1). As underlined by the literature, logistic regression has superior accuracy in predicting an attributes' probability (Pohlman & Leitner, 2003). Indeed, when dealing with probabilities, OLS faces three major problems: logic inconsistent choice probability (delivering predicted values that are not between 0 and 1), the lack of sum constraint (as the regression may produce non-complementary probabilities that do not add up to one). Logistic regression can deal with these two issues (Hofacker, 2007).

4.3 Classification And Regression Tree Analysis (CART Model)

Classification trees analysis is a widely used machine learning algorithm when it comes to predicting binary variable changes. The CART Model is a powerful and sophisticated prediction tool that helps to classify the most impactful explanatory variables $\{x_1 + x_2 + \cdots + x_n\}$, by ranking their respective power on a dependent variable of interest, Y (Morgan, 2014). This tree system is largely used in public health contexts to facilitate the reading and interpretation of complex statistical outputs, by providing an intuitive model design.

The output of the analysis gives us a hierarchical order of the most influential factors associated with Y (at the top) to the least influential (at the bottom). Each intermediary factor is referred to a "node", and each terminatory node is referred to as a "leaf".

Before getting to the leaf, each node split is optimally chosen to minimize the Residual Sum of Squares (RSS):

$$RSS = \sum_{i=1}^{n}(y_i - \breve{y_i})^2 \qquad (10)$$

With y_i being the outcome variable to be predicted having had a mammography, $\breve{y_i}$ being the predicted value of y_i (including all the estimated value of our coefficient β). Also, n is the number of observations. After getting to a leaf, we assume that splitting the data further would add nothing to explaining the variance of the response variable. Note that at the beginning of this analysis, we combined our logistic regression results to our first attempt to classify our variables to "prune" our decision tree (that is, the model excludes the least significant variables of our analysis). Pruning allows avoiding the risk of overfitting the decision tree. The model dropped the explanatory variables of having a high income, the weight level, and being obese. We checked whether this pruning made sense by comparing the results of our logistic regression and observed that these variables were also the ones recognized as insignificant by our regression model. To ease the reading of our results, we decided to divide our classification tree into 3 distinct models: the first one focuses on the classification of the different age cohorts. The second one comprises only the variables that had a positive effect on the outcome variable. The third one comprises only variables that had a negative effect on the outcome variable.

In part 5.3, we will analyze the obtained results and compare the classification of the CART model to the prediction of Lasso.

V. Empirical results

5.1 Logistic regression

5.1.1 Logistic regression output

⇨ *Model 1 (Table 2)*

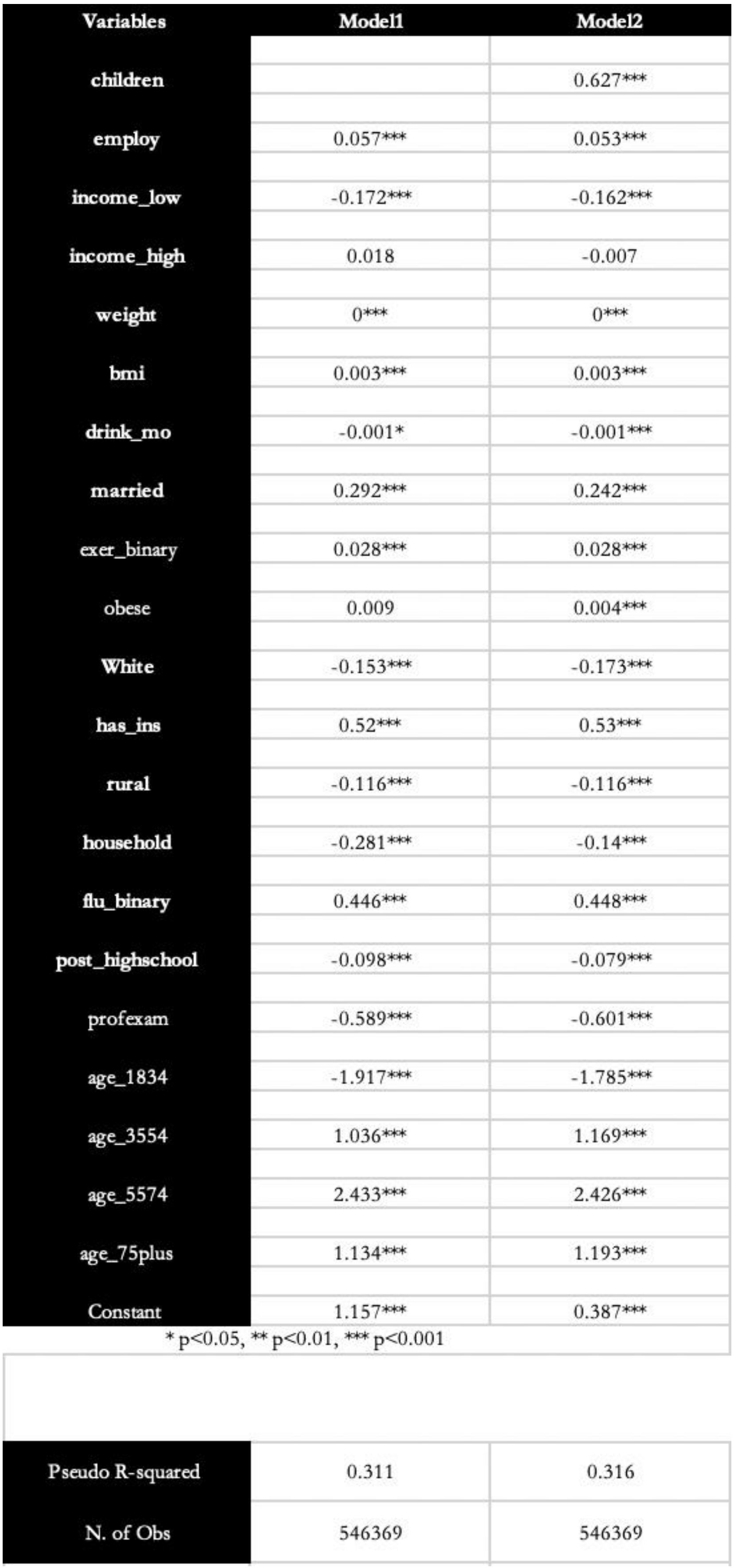
To answer our research questions, we will analyze the coefficient significance of our covariates. These will give us a first glimpse of the most decisive factors in predicting women's mammography participation. As explained in the methodology, we deliberately exclude the "children" variable in this first model. Looking first at the independent variables' effect on the propensity of women to have gone for a mammography, we find out that the following variables have a positive influence in predicting women's mammography participation: being employed, BMI, being married, exercising, having insurance, having received a flu shot. In addition, the age group has a major significance: after 35 years old, women are more likely to have received a mammography. Moreover, we can observe that the following variables have a negative influence in predicting women's mammography participation: having a low income, being white, living in a rural area, having an increasing number of household members, having completed education after high school, having ever had a CBE by a doctor. We can also observe that the age group has a major significance: starting from 18 years old to 34 years old, women are less likely to have received a mammography. We observe that the "high income" and "obese" variables are not statistically significant at the $\propto = 0.05$ significance level. The "drinking" variable is only significant at the $\propto = 0.1$ significance level. (Note that the extensive result of the regression can be found I the Appendix, Table 3).

- ⇨ *Model 2 (Table 2)*

To answer the second research question, we will add up the “children” independent dummy variable (indicating whether the women have at least one child or not). We will wonder whether the results differ. We want to add up this variable in a second and separated stage, to observe its impact on the estimates of the ones already present in the model.

In addition to the same positive covariates found in Model 1, we find the newly added “children” variable. Therefore, we can now assume that having a child has a positive impact on women’s propensity to get mammograms. We do observe the same negative covariates as well, except that the “drinking” variable became significant at the $\propto = 0.05$ significance level. As in the first model, we observe that the “high income” and “obese” variables are not statistically significant at the $\propto = 0.05$ significance level. (Note that the extensive result of the regression can be found in the Appendix, Table 4).

Table 2: Two Models: side-by-side comparison



Variables	Model1	Model2
children		0.627***
employ	0.057***	0.053***
income_low	-0.172***	-0.162***
income_high	0.018	-0.007
weight	0***	0***
bmi	0.003***	0.003***
drink_mo	-0.001*	-0.001***
married	0.292***	0.242***
exer_binary	0.028***	0.028***
obese	0.009	0.004***
White	-0.153***	-0.173***
has_ins	0.52***	0.53***
rural	-0.116***	-0.116***
household	-0.281***	-0.14***
flu_binary	0.446***	0.448***
post_highschool	-0.098***	-0.079***
profexam	-0.589***	-0.601***
age_1834	-1.917***	-1.785***
age_3554	1.036***	1.169***
age_5574	2.433***	2.426***
age_75plus	1.134***	1.193***
Constant	1.157***	0.387***

* p<0.05, ** p<0.01, *** p<0.001

Pseudo R-squared	0.311	0.316
N. of Obs	546369	546369

⇨ *Models comparisons and answers to research questions*

First, let’s compare these two models:

- Without “children”, having an increasing number of drinks per month was not a significant enough factor to influence women’s mammography intake. However, when including this “children” independent variable, we observe that drinking became significant at the $\propto = 0.05$ significance level, and that for one additional drink, the propensity of women to get mammography declines.

- Apart from the inclusion of the drinking variable among the negative coefficients, no other variables changed significantly nor switched from one model to the other.

Secondly, the results obtained can already help us to figure out some answers to our research questions:

A. Having a low income is negatively correlated with breast cancer screening in both models (considering that middle income is our reference group). Earning a high income is apparently not significant in predicting women’s mammography intake. Note that this result will be discussed in the light of “Income and Price effects” in the Discussion part.

B. The significance of the age group is similar across both models, and the inclusion of the “children” independent variable does not significantly change them. Indeed, in both cases, women are more likely to have received a mammography later in their life: from 35-74, and more significantly between 55 and 74 years old.

C. Finally, having had a CBE is negatively correlated to mammography participation in our two models.

5.1.2 Goodness-of-fit tests

Before evaluating the results and conclusions that can be drawn from this logistic regression, we need to make sure that they fit the underlying data appropriately, to confirm that our regression method is appropriate and correctly specified. In order to assess the goodness of fit of our model, we ran a Hosmer and Lemeshow's goodness-of-fit test.

In Hosmer and Lemeshow's goodness-of-fit test, we want to test that the sum of expected probabilities, corresponds to the actual sum of observed probabilities. In the output of the test (see Table 8, Appendix), we can observe that it divides our sample into rank groups of people in deciles of "risks", where the underlying "risk" is having had a mammography. We see that in the riskiest group, (ranked 10), out of a total of 54 636 people, for 53 793 of them, Y=1 (received a mammography); and for 843 of them, Y=0 (did not receive a mammography). For these people, the model expected that 53 285 people would receive the outcome, while 1351 would not. Overall, the expected and observed results line up. However, our regression model faces a goodness-of-fit issue. Indeed, the Hosmer and Lemeshow's goodness-of-fit tests provide a statistically significant result at the 95% confidence level, with a 0.000 p-value, which implies that our regression does not fit the data appropriately, as we can reject the goodness-of-fit. However, one needs to note that we've chosen to restrict our regression only to a very limited number of control variables, compared to the original data set, which can make the goodness-of-fit sink. Moreover, the obtained pseudo-R-square is roughly constant across both models. Also, the structure of the dependent variable that we've chosen to split into two distinct groups by making it binary, can strongly affect the regression fit. Having a large dataset can also affect the significance of any goodness-of-fit test, where even a small departure from the observed model can be considered significant (Lemeshow & Hosmer, 1982). Nonetheless, this does not come as a crucial bias in the interpretation of the logistic regression coefficients, allowing us to make inferences. We can perform a scalar measure of fit to ensure that our second model is complementary to the first.

5.1.3 Scalar measure of fit

The BIC (Bayesian Information Criterion) and AIC (Akaike's Information Criterion) tests can be used to assess the goodness-of-fit of our two compared models and their respective plausibility. In particular, the BIC identifies the model that is more likely to have produced the observed underlying data. Overall, the more negative the value of the BIC, the better fit. In this analysis (Table 9, Appendix), we can observe that there is strong evidence for adding the “children” independent variable (corresponding to the “current” model), the difference in BIC being much greater than 10. Also, the AIC becomes slightly smaller when this single variable is added (as smaller values of AIC are preferred). Another proof is the significance the McFabben R-squared, (considered as the main pseudo-R-squared of reference), but also all the various pseudo-R-squared listed go up, when the “children” variable is added in the “current” model.

5.1.4 Least Absolute Shrinkage and Selection Operator (Lasso)

First, we split our data into two groups. The first one will be our training data set (used to select our model) and the second one will be our testing data set (used to test the final prediction). (See Table 10)

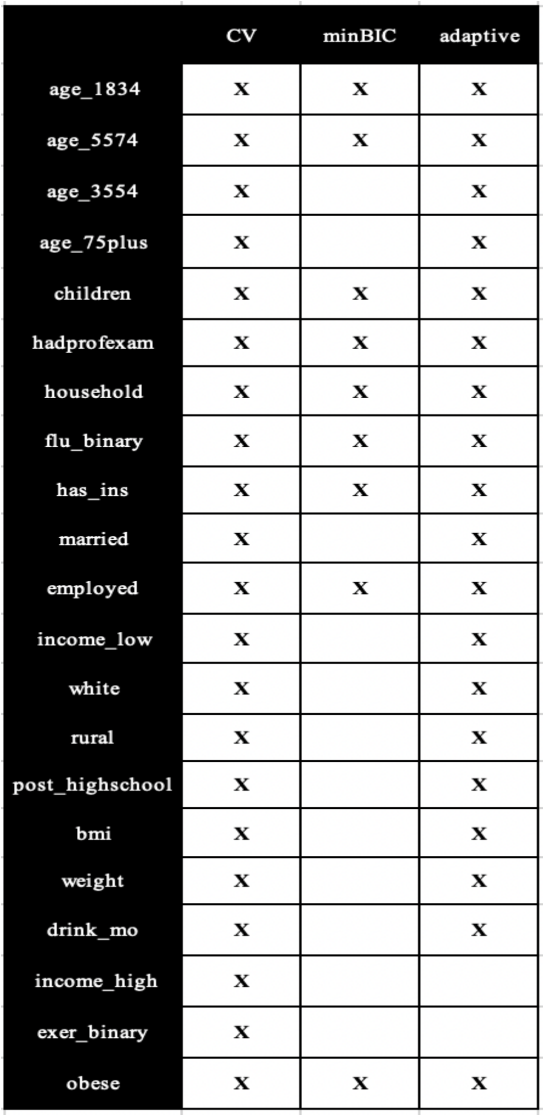
In the first part, we use our testing data set to fit a logit lasso model. It shows that the smallest cross-validation mean (0.11), is obtained with lambda equal to 0.0001, and will be the best one to be used for prediction. (See Table 11, Appendix)
In the cross-validation plot (see Figure 5, Appendix), we can clearly see that the cross-validation function is minimized when lambda is about 0.0001.

Further, we run a second model selecting the lowest BIC, which is obtained when lambda equals 0.013. This model is more parsimonious than the first, including only 8 variables rather than 21 (when lambda equals 0.00012). See Table 12 and Figure 6, Appendix.

Finally, we run a very last, third model (an adaptive Lasso model), that will again generate multiple models to select the best-fitting one, this time according to cross-validation mean prediction error. The smallest one is chosen (where CV mean prediction error equals 0.11). See Table 13, Appendix.
As a result, “Lasso Coef” gives us a table displaying the variables of importance that were selected, using our three different models. For instance, looking at the BIC, Lasso selected the variable (noted with an “x”), if the variable was selected using the minimum BIC method. The variables with the largest standardized coefficient are ranked at the top of Table 15 (the most important first). This ranking order will be our focus to analyze the results.

As we can see in Table 14, the first variables of importance in the prediction of mammography intake are all the age variables, starting with the 18-34 age group. Then comes the "children" variable, followed by the "profexam" (CBE) variable. This ranking emphasizes the accuracy and importance of our research questions, as we can see that the age of the woman, having a child, and having done a CBE are the strongest predictors of mammography intake.

<u>TABLE 14:</u> Lasso variables prediction



	CV	minBIC	adaptive
age_1834	X	X	X
age_5574	X	X	X
age_3554	X		X
age_75plus	X		X
children	X	X	X
hadprofexam	X	X	X
household	X	X	X
flu_binary	X	X	X
has_ins	X	X	X
married	X		X
employed	X	X	X
income_low	X		X
white	X		X
rural	X		X
post_highschool	X		X
bmi	X		X
weight	X		X
drink_mo	X		X
income_high	X		
exer_binary	X		
obese	X	X	X

X = Estimated

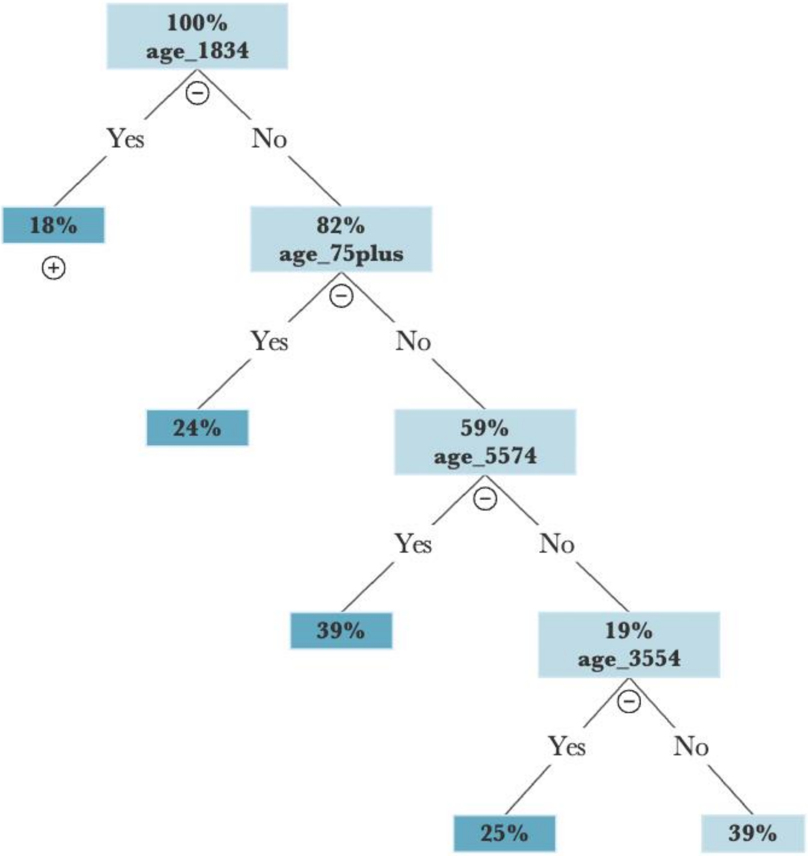
Following our Lasso experiment, we can also use these results to assess the goodness-of-fit of our model (Table 15, Appendix). Where "sample 1" corresponds to our training group and "sample 2" is our testing group that we obtained by creating splits sample. The model with the minimum BIC has also the smallest mean squared error (MSE) and a larger R-squared than the training data set.

5.2 CART Model

We now want to analyze the results of the classification obtained by the CART Model.

Following the CART model interpretation guidelines (Morgan, 2014), we can start to observe our first model that depicts a classification of the age groups (Figure 1). As noted on the tree, the left branch of the node is conditional on the node being true, and the right one on the node being false. The most important cohort associated with mammography intake is the 18-34 years old. The percentages at the bottom of the classification tree show the mean of the mammography intake in each data subset. For instance, a woman belonging to the 18-34 age cohort is associated with 18% average mammography rate, while a woman above this age group has an increased 82% associated rate. The second most important age group to consider according to the CART model is then the 75+ age group, conditional on not belonging to the 18-34 cohort. If the woman is more than 75 years old, she only has a 24% average mammography rate; While if she does not (that is, if she's aged between 35 and 74 years old), (if she is between 35 and 74 years old) her mammogram rate increases to 59%. This interpretation follows for the rest of the classification tree. Overall, the 55-74 age group seems to be the one associated with the highest mammography rate.

FIGURE 1:


CART Model 1 : age cohorts

The second model (Figure 2 and 3) classifies only the explanatory variables positively correlated with our outcome variable, from the logistic regression. As this model includes a lot of variables, we decided to split it into two smaller separated models, which makes the reading of the results easier. We observe in Model 2 (1), that a woman who does not have at least one child has a lower mammography rate (44%) compared to a woman having at least one (56%). Conditional on having at least one child, if a woman did not receive a flu shot, her mammography rate would be higher than if she had received it. Lastly, being insured largely increases women's propensity to get screened. Looking at Model 2 (2), we observe that being employed is classified first, followed by the level of physical activity and marital status.

FIGURE 2:

CART Model 2 (1) : positively impacting explanatory variables

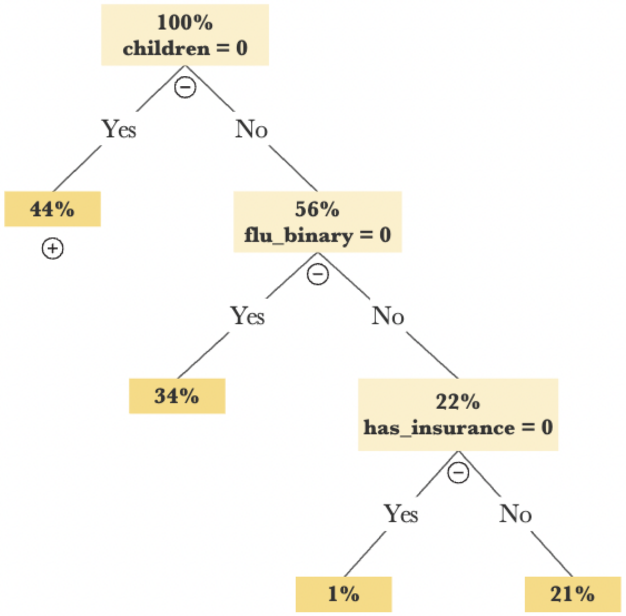
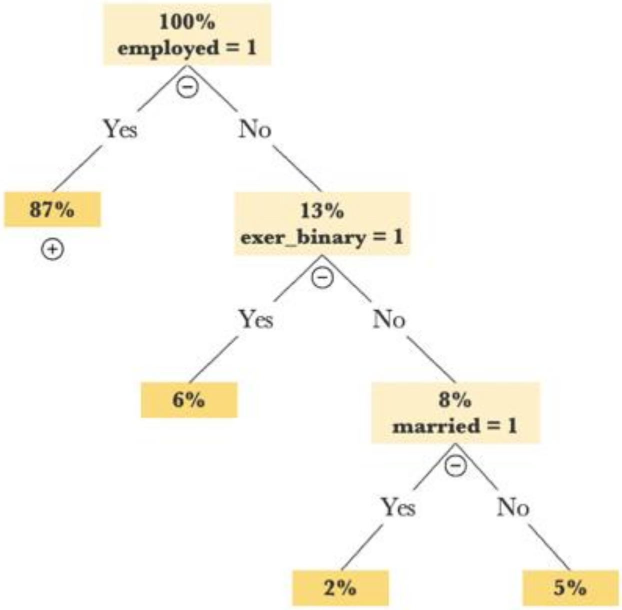


FIGURE 3:

CART Model 2 (2) : positively impacting explanatory variables

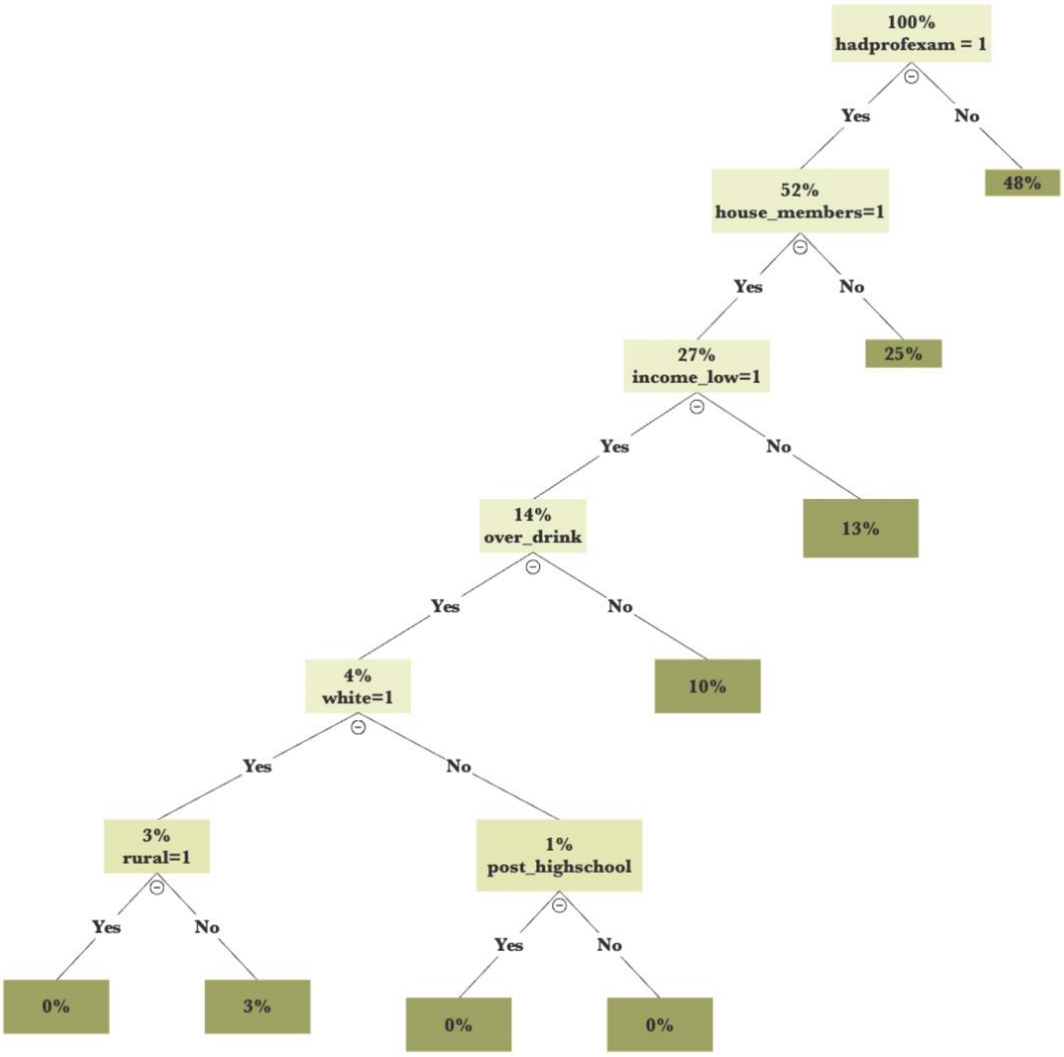


The third model (Figure 4) classifies only the explanatory variables negatively correlated with our outcome variable, from the logistic regression. The CART model classifies having done a CBE ("hadprofexam") at the very top of the tree. Living in a rural area and having completed higher education seem to have an insignificant predictive power on mammography intake.

Overall, just like the Lasso model, the results of the CART model come as a confirmation of the accuracy and importance of our research questions. The age group, having at least one child, having had a CBE, as well as having a low income, all appear as strong factors influencing women's decision to have a mammogram.

FIGURE 4:

CART Model 3 : negatively impacting explanatory variables



5.3 CART model and Lasso results comparison

Observing both the predictions made by the Lasso model and the classification of the CART model, we now want to compare to which extent they differ and come together.

First, regarding the age cohort, both methods seem to meet on the first one to consider, being in the 18-34 groups. As Lasso confirms the classification made by the decision tree, the CART model adds some useful specification, depicting the mammogram's percentage rate. Indeed, as Lasso only ranks the variables according to their predictive power, CART shows how to consider their impact. Lasso tells us that the 18-35 age group is of the utmost importance, while CART shows that it is not women belonging to this age group that have a higher mammography rate, but women that do not belong to it. Also, as the tree shows, the women not belonging to the 18-35 nor 75+ age group (and therefore belonging to the 35-74 age group) have an increased rate of mammography intake. Combining the two models, we see that women aged between 55 and 74 years are overall the most susceptible to have received a mammography.

Second, regarding the positively impacting explanatory variables, we notice that the first predictor ranked by the Lasso model is also the same top variable classified by CART, namely, having at least one child. The ranking is then the same for having received a flu shot, having insurance, and being employed. However, doing physical exercise and being married seem to be more significant in the Lasso model.

Third, regarding the negatively impacting explanatory variables, we observe that the independent variable of primary importance for Lasso was having gone for a CBE, which ranks this variable in the same position as the decision tree. Overall, all of the variables in these groups are attributed a similar ranking for both methods, except for the number of drinks variable. (However, note that we slightly modified this variable for the CART model, setting a threshold of drinks number to double-check its significance).

Comparing classification models has been of increasing importance these past few years, as the number of models and algorithm expanded. Anandhanathan and Gopalan (2021) compared numerous machine learning algorithms, including decisions trees like the CART model and the Lasso model to predict the leading factors of COVID-19-related deaths.

Overall, they concluded that Lasso had both a better fit and accuracy level compared to decision trees. However, as we can observe in our experiment, the mammography rate depicted at the bottom of the classification tree is a solid added value, as it shows us in which proportion a woman with a particular attribute is more or less likely to have had a mammography. Therefore, regardless of their fit or accuracy, we can observe the complementarity of these two methods, while they provided us with similar predictions.

VI. Discussion

6.1 Main findings

Overall, using Lasso and CART to rank our explanatory variables according to their relative predictive power, we found relatively similar outcomes. They both emphasized the importance of the research questions we decided to focus on. In the next parts, we are going to try to analyze these results in the light of the Random Utility Model, and of the income and price effects.

6.1.1 Random Utility Model (RUM)

We will analyze the results obtained from our logit regression in the light of the Random Utility Model (RUM), developed by McFadden for discrete choice models. In the health care context, "utilities" are usually considered as weights to measure the overall health status of a patient, both considering the quality and quantity of life (often described in the literature as "QALYs", Quality Adjusted Life Years), (Bakker & Van der Linden, 1995). However, note that analyzing and comparing the impact of breast cancer screening methods on QALYs is beyond the scope of this thesis. RUMs assume that individuals are "rational" and that preferences for mammography screening can be modelled by a utility function that they will aim to maximize (Manski, 1977). The utility of the available alternatives (getting mammography versus not getting it) depends on women's respective attributes, composed of observable (e.g., age, marital and income status, etc.) and unobservable characteristics (e.g., fear of mammogram and its results, social pressure, etc.) These unobservable determinants are represented by random variables. RUM can provide the probability related to the choice of each alternative (Horowitz et al, 1994):

$$U_j = V(\beta,\, x_j) + \varepsilon_j \qquad (11)$$

- Where U_j is the utility function for an individual choosing among j alternatives.
- V is a function by the attribute levels for alternatives (denoted j).
- Where β represents a vector of estimate coefficients.
- Where x_j represents the set of observable factors described in Part 3.3.
- Where ε_j is a random variable accounting for the effects of individual preferences and unobserved attributes on mammography preferences.

The estimated parameters of our logit model generate a valuable perception on utility of choice alternatives by revealing individuals' preferences. Lancaster (1966) calls it « indirect utility », where utility is derived from:

- The characteristics of decision-makers (the women observed in our dataset), which are described by our set of independent variables.
- Attributes of the choice options, described by our dependent variable "hadmamyes".

From our logit model, women's utility can be deduced from their observed choice of getting mammography or not. Of course, the accuracy of this deducted utility can only be as good as the indirect information we have in hand, containing inevitably measurement errors, unobservable characteristics, etc. (Kroh et al, 2003).

In this analysis, the estimated utility will be presented considering the odds ratio from our logit model. Odds ratios can be interpreted as the percentage change in the odds to get a mammography, produced by a unit change in one of the variables of interest, holding the other constant. The interpretation of odds ratio differs from the one of the coefficients from the logistic regression, as a result above one accounts for an increase in the odds to get screened, whereas a result below one accounts for a decrease in the odds to apply. Again, we assume that the information collected in this data set reflects women's preferences and rationality (Horstschräer, 2012).

Firstly, earning a low income induces a decrease in the perceived utility from getting screened, whereas earning a high income seems to have no significant impact on it.

The odds that a woman aged between 55 and 74 years old has done a mammography are approximately 4 times higher than for a woman aged between 35 and 54 years. The level of utility obtained from mammograms increases gradually from this age group and reaches a peak for the 55-74 age cohort, before dropping down for women aged more than 75 years. (Table 7, can be found in the Appendix part)

In Table 7 we observe that the odds of having gone for a mammography for a woman who had a CBE exam is 1.5 time less likely than for a woman who did not. This result implies that a woman's utility from getting a breast cancer screening exam such as a mammography is decreased if she has had a CBE.

From Table 7 we see that a woman having at least one child is 1.8 times more likely to have undergone a mammography, increasing her underlying perceived utility from the examination. Also, we can notice the impact of this "children" variable on the other ones previously described. In particular, we see that the derived utility of a mammography increases for women between 35 and 54 years old.

6.1.2 Income and Price Effects

As our set of regressors includes three different income levels, we will be able to deduce the related income and price effects on breast cancer screening behaviors. On the macro-level, Wang (2018) has shown that countries having a higher GDP are also the ones spending the most on health care services, which can be considered as luxury goods. We will therefore analyze whether increasing wealth and changes in mammography prices are positively or negatively correlated with screening intake among American women.

We will first look at the direct income effects, defined as the change in consumption (here, preventive healthcare service) experienced when someone's income changes. According to our results, having a low income is associated with a 0.16 decrease in the odds of getting a mammogram, compared to a woman earning a middle income-level (Table 6). This result is not affected by having had a child. The high-income variable appears to be insignificant for any of the conventional significance level when compared to earning a middle-income level. The inclusion of the "children" variable does not make any difference.

Moreover, both the Lasso and the CART models predict low-income status to be a strong predictor of mammography intake, and the high-income status to be an insignificant factor. Therefore, we can conclude that having a low income (compared to earning a middle-level income) has a negative impact on preventive healthcare service consumption, such as a mammography), whereas earning a high income (compared to middle income) has no significant impact. In other words, wealth is only impactful on mammography intake below a certain income level. This result is consistent with the literature, where other cancer screening techniques (such as Pap Smear, colorectal cancer screening, PSA testing) were all associated with a decreased use for low-income groups (Swan et al, 2003).

Considering the variables that we have, evaluating the price effect directly is not possible, as it would require detailed data on the share of deductibles and cost-sharing that have to be paid by the individuals present in our sample. However, considering the insurance level of our population can be a good general proxy to evaluate the medical costs women have to bear, as having insurance boils down to a reduction of mammogram costs. The insurance status may have an important interaction magnitude with the income status. Melvin et al, (2016), have demonstrated that uninsured women were associated with a 3.37 increase in the odds of not getting a mammography. (Note that in our dataset, no distinction is made between private and public insurances). To verify how income and insurance statuses interfere in the mammography decision process, we have built an interaction term. In Table 16 (Appendix), we observe that the interaction term between low-income level and the insurance status is significant at the 0.1 significant level and is associated with a 0.949 decrease in the odds of getting a mammography. The interaction between high income and insurance level is highly significant. Being a high-income earner and having a positive insurance status induces a 1.12 increase in the odds of getting screened (Table 17, Appendix), whereas earning a high income but not having enough insurance engenders a decrease of 0.11 in the odds of having received a mammography (Table 18, Appendix).

6.2 Limitations of this study and suggestions for future research

In this section, we want to highlight some significant limitations of our work. First, our set of variables might suffer from possible OVB, such as the use of contraception, current pregnancy status, past cancer history, family record of breast cancer, etc. Our dataset did not include any control variable. Some potentially useful controls could have been the measurement of breast cancer screening centers' availability and accessibility per state, or per-state breast cancer survival/mortality rate. Second, a major concern of having self-reported data is the misreporting of real screening participation. The individuals who received the call and questionnaire of the agent might have been scared to declare their true status. For that reason, one must consider that the actual mammogram rate is likely to be lower than the one estimated in our data (Einav et al, 2020). Third, regarding our outcome variable, our dataset does not allow us to disentangle spontaneous from prescribed mammograms. Indeed, some women may have participated in it after a CBE or a self-palpation that conducted the physician or themselves to doubt about the nature of a lump. Some others may have been influenced by external factors, like some breast cancer awareness campaigns (that are common in precarious areas), allowing them to get a free screening. Others may have just felt the need to get a mammography even without having to do an annual check but feeling more at risk. Lastly, after performing a CBE, some women may have been advised by their physician not to undergo a mammography that year, as he did not detect any trace of harmful lump and did not consider the need to combine both techniques. According to official medical guidelines, we can’t be sure that this phenomenon happened in a significant proportion, but this is an additional element to take into account, implying that our results are probably over-estimated.

Taking into consideration these limitations, we would like to provide some suggestions for future research. First, future studies on preventive health care behaviors should prioritize the need to question individuals on the nature of their decision (whether voluntary or compelled). This would allow them to draw clear conclusions regarding their motivation and decision process. Also, some research has to be done on the sufficiency of CBE alone compared to its combination to mammography, in terms of psychological and medical costs, as well as its induced mortality and QALYs rate. This information could drastically change the recommendations made to women, affecting all of the variables tested in this thesis. Finally, going further in the analysis of the determinant characteristics of mammography participation, it could be interesting to identify the importance of individual and community-level effects.

VII. Conclusions and policy implications

In this part, we are going to draw conclusions on our main research question and each of the four above-mentioned hypotheses, while observing whether these elements correspond to what has been found in the literature.

First of all, we established an empirical classification of mammography determinants, going beyond the usual odds ratio interpretation conducted by most studies. The Lasso and the CART model provided consistent predictions that could efficiently be used to classify with precision patients characteristics and health-behavior determinants. As we have demonstrated, both of these methods complement each other in their predictions, and come as a necessary precision of the logistic regression. Second, answering H1, we found that women living under the poverty level have a lower propensity to get a mammography, which confirms what has been found in the literature (Calle et al, 1993). However, contrary to previous studies (Gathirua-Mwangi et al, 2018; Burns et al, 1996), we found that earning a high-level income was not a significant factor of mammography screening predictions. Therefore, we reject the null hypothesis that income-level has no impact on mammography participation. Third, answering H2, our results suggest that women started screening from 35 years old, and more significantly from 55 years old, suggesting that many of them followed the US Preventive Task Force recommendation. Some of them may have been discouraged to start screening before their 50s after the decision from the Affordable Care Act to stop reimbursing screening from that age. We reject the null hypothesis that age has no effect on screening participation. Fourth, answering H3, we found that the practice of CBE had a negative impact on screening behaviors, allowing us to reject the null hypothesis that CBE has no effect on mammography's participation decision. We can suggest that women probably think CBE is enough in itself and do not feel like going further, regardless of the recommendation of participating in screening at least once or twice a year. Fifth, answering H4, we observed that having at least one child induced an increase in mammography intake, which allow us to reject the null hypothesis that having at least one child has no impact on mammography participation. Again, although the literature does not provide explicit responses in that regard, there could be many explanations for that phenomenon. We can suggest that having had children, women got more aware of cancer risks, predispositions, and need for prevention (as they necessarily engage more in medical and physician interactions), increasing their propensity to get screened. Regarding practical implications, the results from this work could be used in designing and orientating policies to minimize mammography's barriers and increase awareness. In order to increase recovery rates, policy-makers need to make mammograms affordable and accessible to women living with a low-income level while keeping high-income level patients in their target, adjusting and customizing their approach to both of these groups. We observed that early screening (before age 50) is likely to be discouraged by the non-reimbursement of care, even for women having insurance. Policy-makers should keep it in mind if the screening age was to be advanced in the next few years. Our findings regarding CBE imply that a lot of women can be false-negative after this test, leaving room for a potential tumor to grow, engendering both lower recovery rates and extra costs due to more intensive treatments. Physicians should then reconsider the use of this practice for routine consultations. Also, knowing that mothers are more exposed to medical interactions, children care visits could be seen as occasions for recommendations. Policy-makers would then have to emphasize their reach-out with childless women.

VIII. References

Akinyemiju, T. F. (2012). Socio-economic and health access determinants of breast and cervical cancer screening in low-income countries: analysis of the World Health Survey. *PloS one*, *7*(11), e48834.

American Cancer Society. (2009). Cancer facts and figures for Hispanics/Latinos 2009–2011.

Anandhanathan, P., & Gopalan, P. (2021). Comparison of Machine Learning algorithm for COVID-19 Death Risk Prediction.

Arleo, E. K., Hendrick, R. E., Helvie, M. A., & Sickles, E. A. (2017). Comparison of recommendations for screening mammography using CISNET models. *Cancer*, *123*(19), 3673-3680.

Bakker, C., & Van der Linden, S. M. J. P. (1995). Health related utility measurement: an introduction. *The Journal of rheumatology*, *22*(6), 1197-1199.

Bleyer, A., & Welch, H. G. (2012). Effect of three decades of screening mammography on breast-cancer incidence. *New England Journal of Medicine*, *367*(21), 1998-2005.

Broome, J. (1993). Qalys. *Journal of public economics*, *50*(2), 149-167.

Burns, R. B., McCarthy, E. P., Freund, K. M., Marwill, S. L., Shwartz, M., Ash, A., & Moskowitz, M. A. (1996). Black women receive less mammography even with similar use of primary care. *Annals of internal medicine*, *125*(3), 173-182.

Calle, E. E., Flanders, W. D., Thun, M. J., & Martin, L. M. (1993). Demographic predictors of mammography and Pap smear screening in US women. *American journal of public health*, *83*(1), 53-60.

Champion, V. L. (1993). Instrument refinement for breast cancer screening behaviors. *Nursing research*.

Ecob, R., & Smith, G. D. (1999). Income and health: what is the nature of the relationship?. *Social science & medicine*, *48*(5), 693-705.

Einav, L., Finkelstein, A., Oostrom, T., Ostriker, A., & Williams, H. (2020). Screening and selection: The case of mammograms. *American Economic Review*, *110*(12), 3836-70.

Fletcher, S. W., O'Malley, M. S., & Bunce, L. A. (1985). Physicians' abilities to detect lumps in silicone breast models. *Jama*, *253*(15), 2224-2228.

Freeman, H. P., & Chu, K. C. (2005). Determinants of cancer disparities: barriers to cancer screening, diagnosis, and treatment. *Surgical oncology clinics of North America*, *14*(4), 655-670.

Freitas, C., Tura, L. F. R., Costa, N., & Duarte, J. (2012). A population-based breast cancer screening programme: conducting a comprehensive survey to explore adherence determinants. *European Journal of Cancer Care*, *21*(3), 349-359.

Gabe, R., & Duffy, S. W. (2005). Evaluation of service screening mammography in practice: the impact on breast cancer mortality. *Annals of oncology*, *16*, ii153-ii162.

Gathirua-Mwangi, W., Cohee, A., Tarver, W. L., Marley, A., Biederman, E., Stump, T. & Champion, V. L. (2018). Factors associated with adherence to mammography screening among insured women differ by income levels. *Women's Health Issues*, *28*(5), 462-469.

Henderson, L. M., O'Meara, E. S., Haas, J. S., Lee, C. I., Kerlikowske, K., Sprague, B. L., ... & Onega, T. (2020). The role of social determinants of health in self-reported access to health care among women undergoing screening mammography. *Journal of Women's Health*, *29*(11), 1437-1446.

Hendrick, R. E., Smith, R. A., Rutledge III, J. H., & Smart, C. R. (1997). Benefit of screening mammography in women aged 40-49: a new meta-analysis of randomized controlled trials. *JNCI Monographs*, *1997*(22), 87-92.

Hofacker, C. (2007). Chapter 13: Random Utility Models. *MAthematical Marketing*, 168-167.

Horowitz, J. L., Bolduc, D., Divakar, S., Geweke, J., Gönül, F., Hajivassiliou, V., ... & Ruud, P. (1994). Advances in random utility models report of the workshop on advances in random utility models duke invitational symposium on choice modeling behavior. *Marketing Letters*, *5*(4), 311-322.

Horstschräer, J. (2012). University rankings in action? The importance of rankings and an excellence competition for university choice of high-ability students. *Economics of Education Review*, *31*(6), 1162-1176.

Jatoi, I. (2011). The impact of advances in treatment on the efficacy of mammography screening. *Preventive medicine*, *53*(3), 103-104.

Kearney, A. J., & Murray, M. (2009). Breast cancer screening recommendations: Is mammography the only answer? *Journal of midwifery & women's health*, *54*(5), 393-400.

Kroh, M., van der Eijk, C., & Amsterdam, D. L. (2003, March). Utilities, Preferences and Choice. In *joint sessions of workshops of the ECPR in Edinburgh*.

Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of political economy*, *74*(2), 132-157.

Lechner, L., De Vries, H., & Offermans, N. (1997). Participation in a breast cancer screening program: influence of past behavior and determinants on future screening participation. *Preventive Medicine*, *26*(4), 473-482.

Leitch, A. M., Dodd, G. D., Costanza, M., Linver, M., Pressman, P., McGinnis, L., & Smith, R. A. (1997). American Cancer Society guidelines for the early detection of breast cancer: update 1997. *CA: A cancer Journal for Clinicians*, *47*(3), 150-153.

Lemeshow, S., & Hosmer Jr, D. W. (1982). A review of goodness of fit statistics for use in the development of logistic regression models. *American journal of epidemiology*, *115*(1), 92-106.

Manski, C. F. (1977). The structure of random utility models. *Theory and decision*, *8*(3), 229.

Melvin, C. L., Jefferson, M. S., Rice, L. J., Cartmell, K. B., & Halbert, C. H. (2016). Predictors of participation in mammography screening among non-hispanic black, non-hispanic white, and hispanic women. *Frontiers in public health*, *4*, 188.

Morgan, J. (2014). Classification and regression tree analysis. *Boston: Boston University, 298*.

Moss, S. M., Wale, C., Smith, R., Evans, A., Cuckle, H., & Duffy, S. W. (2015). Effect of mammographic screening from age 40 years on breast cancer mortality in the UK Age trial at 17 years' follow-up: a randomised controlled trial. *The Lancet Oncology*, *16*(9), 1123-1132.

National Cancer Institute. (2009). Fact sheet: Mammograms.

Nechuta, S., Paneth, N., & Velie, E. M. (2010). Pregnancy characteristics and maternal breast cancer risk: a review of the epidemiologic literature. *Cancer Causes & Control*, *21*(7), 967-989.

Pohlman, J. T., & Leitner, D. W. (2003). A comparison of ordinary least squares and logistic regression.

Rosenstock, I. M. (1974). The health belief model and preventive health behavior. *Health education monographs*, *2*(4), 354-386.

Salazar, M. K. (1996). Hispanic women's beliefs about breast cancer and mammography. *Cancer nursing*, *19*(6), 437-446.

Secginli, S., & Nahcivan, N. O. (2006). Factors associated with breast cancer screening behaviours in a sample of Turkish women: a questionnaire survey. *International journal of nursing studies*, *43*(2), 161-171.

Swan, J., Breen, N., Coates, R. J., Rimer, B. K., & Lee, N. C. (2003). Progress in cancer screening practices in the United States: results from the 2000 National Health Interview Survey. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, *97*(6), 1528-1540.

U.S. Preventive Services Task Force (USPSTF). (2009). Screening for breast cancer: U.S. preventive services task force recommendation statement. Annals of Internal Medicine, 151, 716–726.

US Preventive Services Task Force. (2009). Screening for breast cancer: US Preventive Services Task Force recommendation statement. *Annals of internal medicine*, *151*(10), 716-236.

Von Hippel, P. (2015). Linear vs. logistic probability models: Which is better, and when. *Statistical Horizons*.

Wang, F. (2018). The roles of preventive and curative health care in economic development. *PloS one*, *13*(11), e0206808.

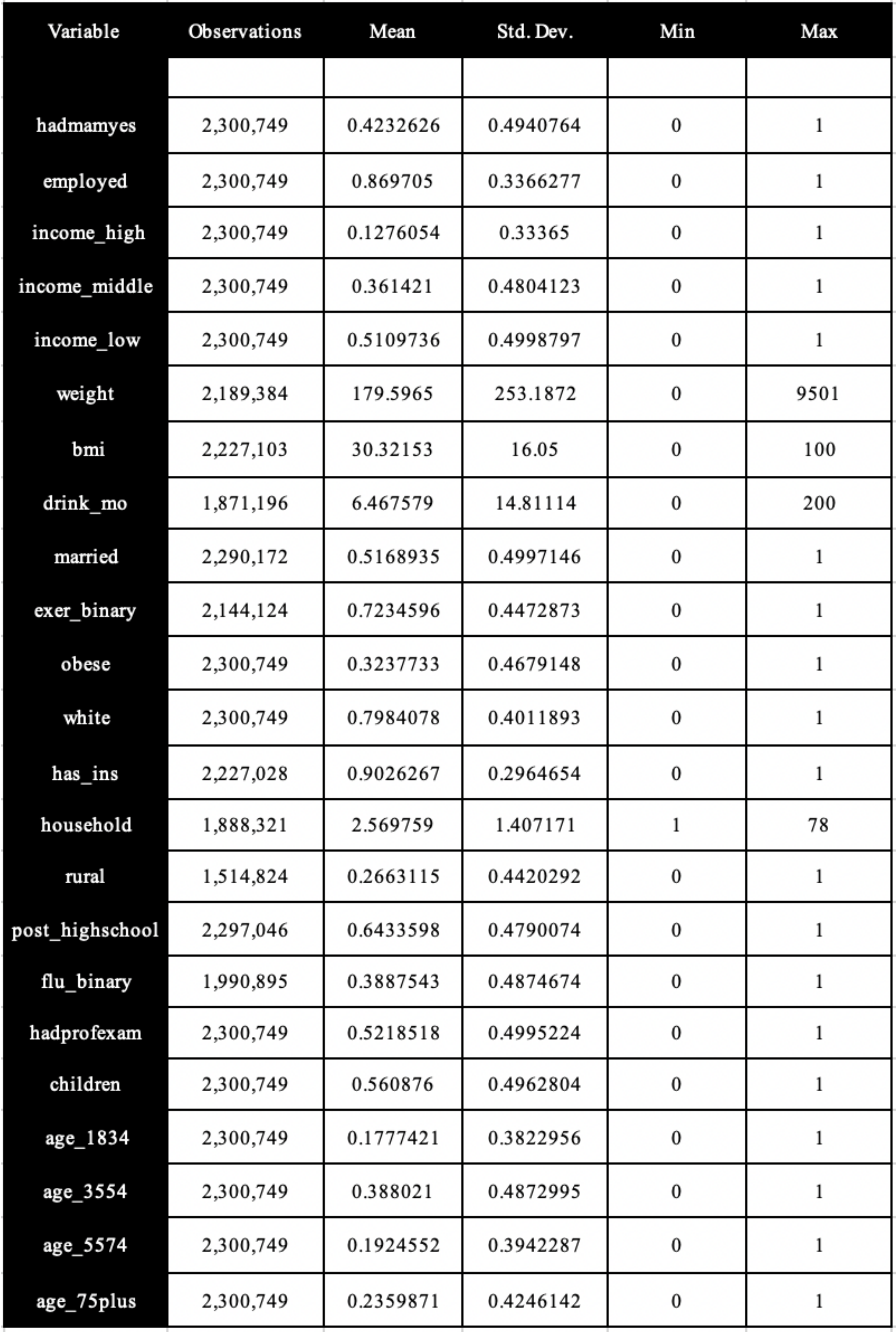
White, E., Urban, N., & Taylor, V. (1993). Mammography utilization, public health impact, and cost-effectiveness in the United States. *Annual Review of Public Health*, *14*(1), 605-633.

Websites

- Byrne, J. (2022, January 6). Vaccine designed to prevent triple-negative breast cancer undergoes phase 1 trial. *Healio, Medical News.* Retrieved from https://www.healio.com/news/hematology-oncology/20220106/vaccine-designed-to-prevent-triplenegative-breast-cancer-undergoes-phase-1-trial
- Behavioral Risk Factor Surveillance System (BRFSS): https://www.cdc.gov/brfss/annual_data/2016/pdf/overview_2016.pdf
- National Cancer Institute. Breast cancer and mammograms. Retrieved from: https://www.cancer.gov/types/breast/mammograms-fact-sheet#what-is-the-best-method-ofnbspscreening-fornbspbreast-cancer
- Pew Research Center: https://www.pewresearch.org/
- Questionnaire:https://www.cdc.gov/brfss/questionnaires/index.htm?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fbrfss%2Fquestionnaires.htm
- World Health Organization (2021, September 21). Cancer facts sheet. Retrieved from: https://www.who.int/news-room/fact-sheets/detail/cancer
- World Health Organization (2021, March 26). Breast cancer facts sheet. Retrieved from: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

IX. Appendix

TABLE 1: Summary Statistics



Variable	Observations	Mean	Std. Dev.	Min	Max
hadmamyes	2,300,749	0.4232626	0.4940764	0	1
employed	2,300,749	0.869705	0.3366277	0	1
income_high	2,300,749	0.1276054	0.33365	0	1
income_middle	2,300,749	0.361421	0.4804123	0	1
income_low	2,300,749	0.5109736	0.4998797	0	1
weight	2,189,384	179.5965	253.1872	0	9501
bmi	2,227,103	30.32153	16.05	0	100
drink_mo	1,871,196	6.467579	14.81114	0	200
married	2,290,172	0.5168935	0.4997146	0	1
exer_binary	2,144,124	0.7234596	0.4472873	0	1
obese	2,300,749	0.3237733	0.4679148	0	1
white	2,300,749	0.7984078	0.4011893	0	1
has_ins	2,227,028	0.9026267	0.2964654	0	1
household	1,888,321	2.569759	1.407171	1	78
rural	1,514,824	0.2663115	0.4420292	0	1
post_highschool	2,297,046	0.6433598	0.4790074	0	1
flu_binary	1,990,895	0.3887543	0.4874674	0	1
hadprofexam	2,300,749	0.5218518	0.4995224	0	1
children	2,300,749	0.560876	0.4962804	0	1
age_1834	2,300,749	0.1777421	0.3822956	0	1
age_3554	2,300,749	0.388021	0.4872995	0	1
age_5574	2,300,749	0.1924552	0.3942287	0	1
age_75plus	2,300,749	0.2359871	0.4246142	0	1

TABLE 3: **Model A (Coefficients results without children variable)**

Logistic regression

hadmamyes	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
employ	.057	.002	28.54	0	.053	.061	***
income_low	-.172	.01	-16.93	0	-.192	-.152	***
income_high	.018	.017	1.10	.273	-.014	.051	
weight	0	0	-4.97	0	0	0	***
bmi	.003	.001	4.38	0	.002	.005	***
drink_mo	-.001	0	-1.93	.054	-.001	0	*
married	.292	.01	30.45	0	.274	.311	***
exer_binary	.028	.01	2.91	.004	.009	.048	***
obese	.009	.013	0.70	.487	-.016	.033	
white	-.153	.01	-14.55	0	-.173	-.132	***
has_ins	.52	.013	39.27	0	.494	.546	***
rural	-.116	.009	-13.24	0	-.133	-.098	***
household	-.281	.003	-93.13	0	-.286	-.275	***
flu_binary	.446	.009	51.55	0	.429	.463	***
post_highschool	-.098	.01	-10.33	0	-.117	-.08	***
profexam	-.589	.01	-61.07	0	-.608	-.57	***
age_1834	-1.917	.042	-45.20	0	-2	-1.834	***
age_3554	1.036	.041	24.98	0	.955	1.117	***
age_5574	2.433	.043	56.72	0	2.349	2.517	***
age_75plus	1.134	.042	26.95	0	1.051	1.216	***
Constant	1.157	.051	22.53	0	1.057	1.258	***

Mean dependent var	0.771	SD dependent var	0.420
Pseudo r-squared	0.311	Number of obs	546369.000
Chi-square	183047.994	Prob > chi2	0.000
Akaike crit. (AIC)	405238.587	Bayesian crit. (BIC)	405474.019

*** *p<.01*, ** *p<.05*, * *p<.1*

TABLE 4: **Model B (Coefficient results with children variable)**

Logistic regression

hadmamyes	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
employ	.053	.002	26.49	0	.049	.057	***
income_low	-.162	.01	-15.87	0	-.182	-.142	***
income_high	-.007	.017	-0.44	.657	-.04	.025	
weight	0	0	-4.94	0	0	0	***
bmi	.003	.001	4.05	0	.002	.004	***
drink_mo	-.001	0	-3.26	.001	-.001	0	***
married	.242	.01	25.01	0	.223	.261	***
exer_binary	.028	.01	2.88	.004	.009	.048	***
obese	.004	.013	0.33	.739	-.021	.029	
white	-.173	.011	-16.44	0	-.194	-.152	***
has_ins	.531	.013	39.98	0	.505	.557	***
rural	-.116	.009	-13.25	0	-.133	-.099	***
household	-.14	.004	-35.26	0	-.147	-.132	***
flu_binary	.448	.009	51.60	0	.431	.465	***
post_highschool	-.079	.01	-8.29	0	-.098	-.061	***
profexam	-.601	.01	-61.95	0	-.62	-.582	***
children	.627	.011	55.10	0	.605	.649	***
age_1834	-1.785	.043	-41.98	0	-1.868	-1.702	***
age_3554	1.169	.042	28.09	0	1.088	1.251	***
age_5574	2.426	.043	56.49	0	2.342	2.51	***
age_75plus	1.193	.042	28.31	0	1.111	1.276	***
Constant	.387	.053	7.27	0	.283	.492	***

Mean dependent var	0.771	SD dependent var	0.420
Pseudo r-squared	0.316	Number of obs	546369.000
Chi-square	186090.837	Prob > chi2	0.000
Akaike crit. (AIC)	402197.744	Bayesian crit. (BIC)	402444.387

**** p<.01, ** p<.05, * p<.1*

TABLE 6: **Model A *bis* (Odds Ratios without children variable)**

Logistic regression

hadmamyes	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
employ	1.058	.002	28.54	0	1.054	1.063	***
income_low	.842	.009	-16.93	0	.825	.859	***
income_high	1.018	.017	1.10	.273	.986	1.052	
weight	1	0	-4.97	0	1	1	***
bmi	1.003	.001	4.38	0	1.002	1.005	***
drink_mo	.999	0	-1.93	.054	.999	1	*
married	1.34	.013	30.45	0	1.315	1.365	***
exer_binary	1.029	.01	2.91	.004	1.009	1.049	***
obese	1.009	.013	0.70	.487	.984	1.034	
white	.858	.009	-14.55	0	.841	.876	***
has_ins	1.683	.022	39.27	0	1.639	1.727	***
rural	.891	.008	-13.24	0	.876	.906	***
household	.755	.002	-93.13	0	.751	.76	***
flu_binary	1.562	.014	51.55	0	1.535	1.588	***
post_highschool	.906	.009	-10.33	0	.89	.923	***
profexam	.555	.005	-61.07	0	.544	.565	***
age_1834	.147	.006	-45.20	0	.135	.16	***
age_3554	2.818	.117	24.98	0	2.598	3.057	***
age_5574	11.394	.489	56.72	0	10.475	12.393	***
age_75plus	3.107	.131	26.95	0	2.861	3.374	***
Constant	3.181	.163	22.53	0	2.877	3.518	***

Mean dependent var	0.771	SD dependent var	0.420
Pseudo r-squared	0.311	Number of obs	546369.000
Chi-square	183047.994	Prob > chi2	0.000
Akaike crit. (AIC)	405238.587	Bayesian crit. (BIC)	405474.019

**** p<.01, ** p<.05, * p<.1*

<u>TABLE 7:</u> Model B *bis* (Odds Ratios with children variable)

Logistic regression

hadmamyes	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
employ	1.054	.002	26.49	0	1.05	1.059	***
income_low	.85	.009	-15.87	0	.833	.867	***
income_high	.993	.016	-0.44	.657	.961	1.025	
weight	1	0	-4.94	0	1	1	***
bmi	1.003	.001	4.05	0	1.002	1.004	***
drink_mo	.999	0	-3.26	.001	.999	1	***
married	1.274	.012	25.01	0	1.25	1.298	***
exer_binary	1.029	.01	2.88	.004	1.009	1.049	***
obese	1.004	.013	0.33	.739	.98	1.029	
white	.841	.009	-16.44	0	.824	.859	***
has_ins	1.701	.023	39.98	0	1.657	1.746	***
rural	.89	.008	-13.25	0	.875	.906	***
household	.87	.003	-35.26	0	.863	.877	***
flu_binary	1.565	.014	51.60	0	1.539	1.592	***
post_highschool	.924	.009	-8.29	0	.906	.941	***
profexam	.548	.005	-61.95	0	.538	.559	***
children	1.872	.021	55.10	0	1.831	1.914	***
age_1834	.168	.007	-41.98	0	.154	.182	***
age_3554	3.219	.134	28.09	0	2.967	3.493	***
age_5574	11.314	.486	56.49	0	10.4	12.307	***
age_75plus	3.298	.139	28.31	0	3.036	3.582	***
Constant	1.473	.079	7.27	0	1.327	1.635	***

Mean dependent var	0.771	SD dependent var	0.420
Pseudo r-squared	0.316	Number of obs	546369.000
Chi-square	186090.837	Prob > chi2	0.000
Akaike crit. (AIC)	402197.744	Bayesian crit. (BIC)	402444.387

*** *p<.01*, ** *p<.05*, * *p<.1*

<u>TABLE 8:</u> (Hosmer and Lemeshow's goodness-of-fit test)

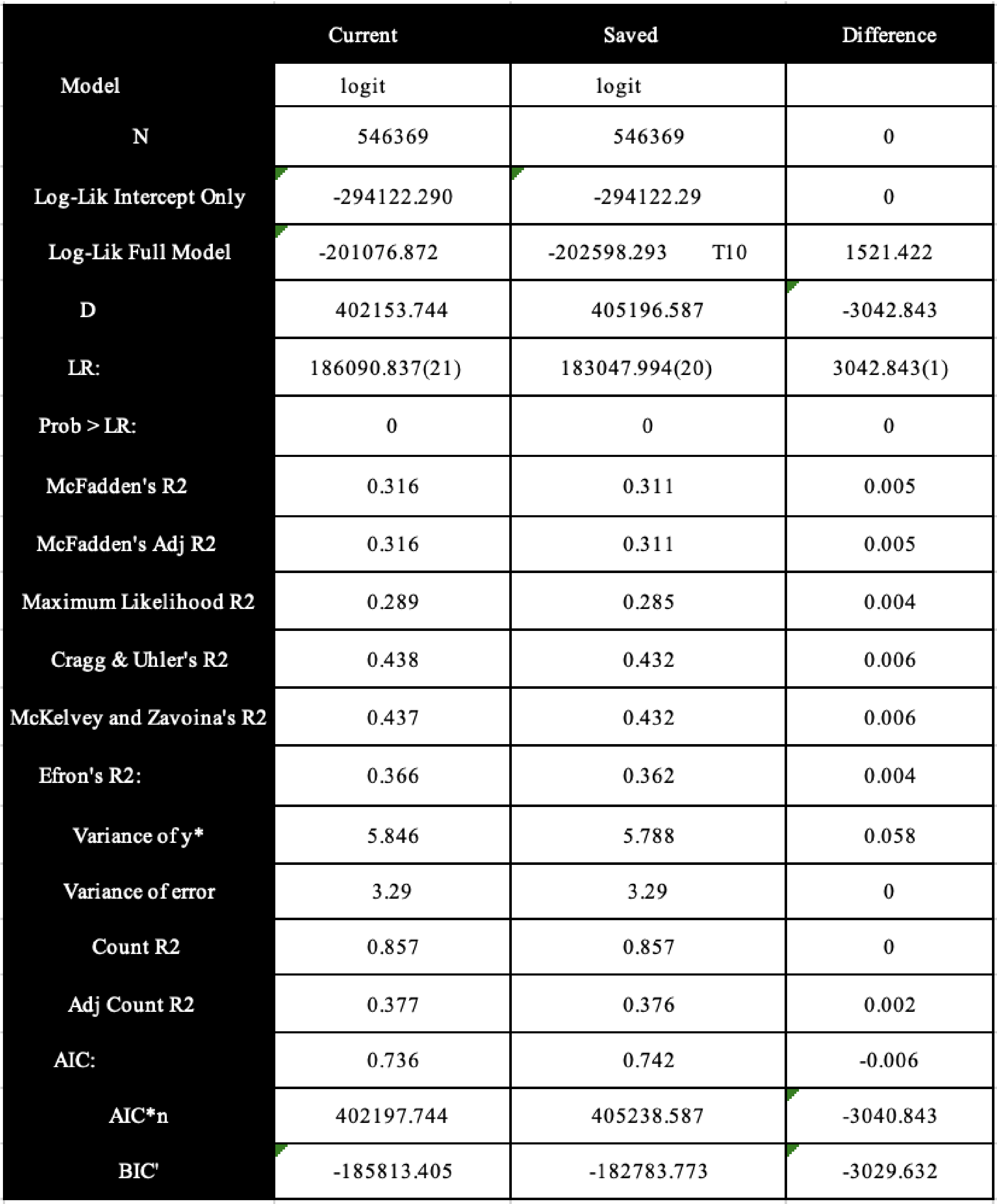
Logistic model for "hadmamyes", goodness-of-fit test
(Table collapsed on quantiles of estimated probabilities)

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0
1	0.220	9556	7082.300	45081	47554.700
2	0.708	25235	28883.900	29402	25753.100
3	0.767	39834	40319.400	14803	14317.600
4	0.824	42169	43587.800	12468	11049.200
5	0.865	46405	46229.200	8239	8414.800
6	0.896	49011	48050.900	5619	6579.100
7	0.921	50892	49672.300	3745	4964.700
8	0.956	51297	51381.400	3340	3255.600
9	0.970	52943	52643.100	1694	1993.900
10	0.991	53793	53284.700	843	1351.300

Number of observations = 546369
Number of groups = 10
Hosmer-Lemeshow chi2(8) = 2959.76
Prob > chi2 = 0.0000

TABLE 9: Scalar Measure of Fits

Measures of Fit for logit of hadmamyes

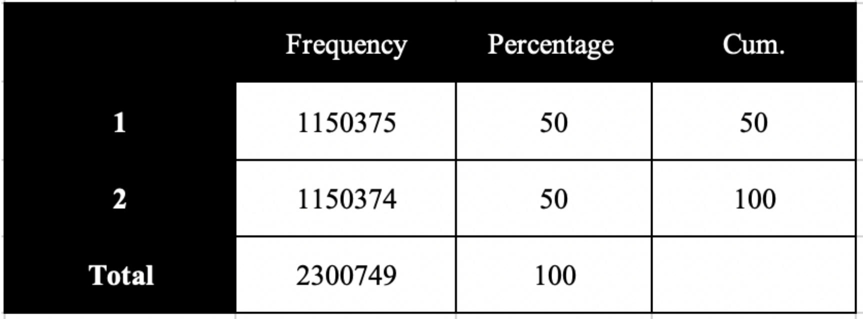


	Current	Saved	Difference
Model	logit	logit	
N	546369	546369	0
Log-Lik Intercept Only	-294122.290	-294122.29	0
Log-Lik Full Model	-201076.872	-202598.293 T10	1521.422
D	402153.744	405196.587	-3042.843
LR:	186090.837(21)	183047.994(20)	3042.843(1)
Prob > LR:	0	0	0
McFadden's R2	0.316	0.311	0.005
McFadden's Adj R2	0.316	0.311	0.005
Maximum Likelihood R2	0.289	0.285	0.004
Cragg & Uhler's R2	0.438	0.432	0.006
McKelvey and Zavoina's R2	0.437	0.432	0.006
Efron's R2:	0.366	0.362	0.004
Variance of y*	5.846	5.788	0.058
Variance of error	3.29	3.29	0
Count R2	0.857	0.857	0
Adj Count R2	0.377	0.376	0.002
AIC:	0.736	0.742	-0.006
AIC*n	402197.744	405238.587	-3040.843
BIC'	-185813.405	-182783.773	-3029.632

Difference of 3029.632 in BIC' provides very strong support for current model.

TABLE 10: Split sample

Tabulation of sample



	Frequency	Percentage	Cum.
1	1150375	50	50
2	1150374	50	100
Total	2300749	100	

TABLE 11: Model (1): lambda and cross-validation mean

ID	*Description*	*lambda*	*No. of nonzero coef.*	*Out-of-sample R-squared*	*CV mean prediction error*
1	First lambda	.2307438	0	0.0002	.1763485
81	Lambda before	.0001351	21	0.3612	.1126768
*82**	Selected lambda	.0001231	21	0.3612	.1126762

* lambda selected by cross-validation.

FIGURE 5: **Cross-validation plot Model (1)**

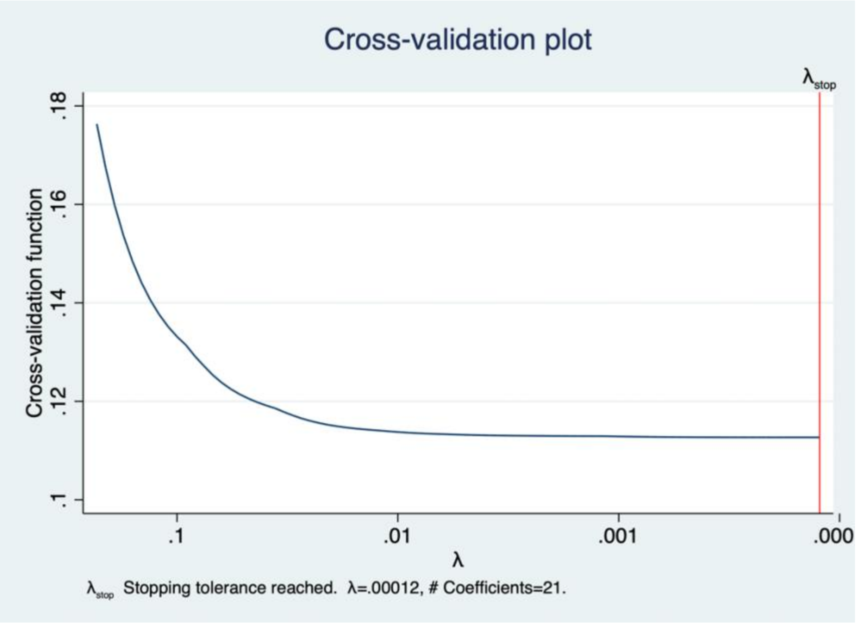


TABLE 12: **Second Model: minimum BIC**

ID	lambda	No. nonzero coef.	Out-of-sample R-squared	BIC
2	0.210	1	0.051	2.87e+05
12	0.083	2	0.268	2.16e+05
13	0.076	3	0.279	2.12e+05
19	0.043	4	0.321	1.95e+05
21	0.036	5	0.328	1.93e+05
22	0.033	6	0.332	1.91e+05
23	0.030	7	0.336	1.89e+05
** 32	0.013	8	0.353	182279
34	0.011	9	0.354	1.82e+05
35	0.010	10	0.355	1.81e+05
40	0.006	11	0.357	1.80e+05
41	0.006	12	0.358	1.80e+05
44	0.004	13	0.359	1.80e+05
49	0.003	15	0.359	1.80e+05
51	0.002	16	0.359	1.80e+05
56	0.001	17	0.360	1.79e+05
59	0.001	18	0.360	1.79e+05
66	0.001	19	0.361	1.79e+05
68	0.000	21	0.361	1.79e+05
* 82	0.000	21	0.361	1.79e+05

** lambda selected by minimum BIC

FIGURE 6: Cross-validation plot Model (2)

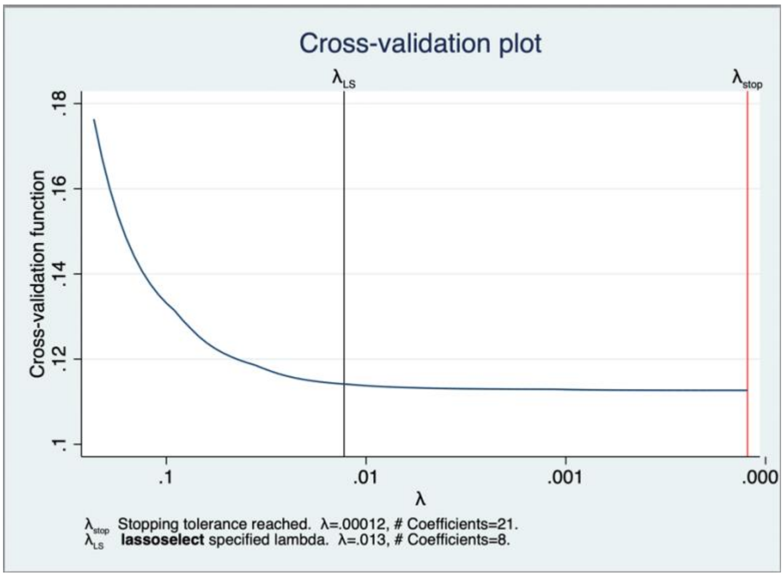


TABLE 13: Third Model: CV mean prediction error

ID	*Description*	*lambda*	*No. of nonzero coef.*	*Out-of-sample R-squared*	*CV mean prediction error*
83	First lambda	17.94511	0	0.0002	.1763485
181	Lambda before	.0019695	17	0.3612	.1126783
*182**	Selected lambda	.0017945	17	0.3612	.1126775

* lambda selected by cross-validation in final adaptive step.

<u>TABLE 15:</u> LASSO goodness-of-fit

Postselection coefficients

Name	sample	MSE	R-squared	Obs
cv				
	1	0.113	0.361	272,885
	2	0.113	0.364	273,484
minBIC				
	1	0.129	0.376	488,763
	2	0.128	0.378	489,716
adaptive				
	1	0.113	0.360	277,101
	2	0.113	0.363	277,623

<u>TABLE 16:</u> (Odds ratios: income_low##has_ins)

Logistic regression

hadmamyes	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
0b.income_low	1	.	.	.	.	.	
1.income_low	.893	.026	-3.81	0	.843	.947	***
0b.has_ins	1	.	.	.	.	.	
1.has_ins	1.773	.048	21.03	0	1.681	1.87	***
0b.income_low #0b.h~s	1	.	.	.	.	.	
0b.income_low #1o.h~s	1	.	.	.	.	.	
1o.income_low #0b.h~s	1	.	.	.	.	.	
1.income_low# 1.has~s	.949	.029	-1.71	.087	.893	1.008	*
employ	1.054	.002	26.63	0	1.05	1.059	***
weight	1	0	-4.94	0	1	1	***
bmi	1.003	.001	4.05	0	1.002	1.004	***
drink_mo	.999	0	-3.26	.001	.999	1	***
married	1.274	.012	25.08	0	1.25	1.298	***
exer_binary	1.029	.01	2.88	.004	1.009	1.049	***
obese	1.004	.013	0.34	.732	.98	1.03	
white	.841	.009	-16.45	0	.824	.859	***
rural	.89	.008	-13.24	0	.875	.906	***
household	.87	.003	-35.24	0	.863	.877	***
flu_binary	1.565	.014	51.60	0	1.539	1.592	***
post_highschool	.924	.009	-8.34	0	.907	.941	***
profexam	.548	.005	-62.06	0	.538	.559	***
children	1.872	.021	55.13	0	1.831	1.915	***
age_1834	.168	.007	-42.02	0	.154	.182	***
age_3554	3.219	.134	28.14	0	2.967	3.492	***
age_5574	11.314	.485	56.55	0	10.402	12.306	***
age_75plus	3.298	.139	28.34	0	3.037	3.582	***
Constant	1.415	.081	6.06	0	1.265	1.583	***

Mean dependent var	0.771	SD dependent var	0.420
Pseudo r-squared	0.316	Number of obs	546369.000
Chi-square	186093.564	Prob > chi2	0.000
Akaike crit. (AIC)	402195.017	Bayesian crit. (BIC)	402441.660

**** p<.01, ** p<.05, * p<.1*

<u>TABLE 17:</u> (Odds ratios: income_high##has_ins)

Logistic regression

hadmamyes	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
0b.income_high	1	.	.	.	.	.	
1.income_high	.987	.04	-0.32	.75	.912	1.069	
0b.has_ins	1	.	.	.	.	.	
1.has_ins	1.746	.024	40.69	0	1.7	1.794	***
0b.income_high#0b.~s	1	.	.	.	.	.	
0b.income_high#1o.~s	1	.	.	.	.	.	
1o.income_high#0b.~s	1	.	.	.	.	.	
1.income_high #1.ha~s	1.126	.049	2.72	.007	1.034	1.227	***
employ	1.05	.002	24.51	0	1.046	1.054	***
weight	1	0	-5.00	0	1	1	***
bmi	1.003	.001	3.69	0	1.001	1.004	***
drink_mo	.999	0	-2.30	.022	.999	1	**
married	1.338	.012	31.66	0	1.314	1.362	***
exer_binary	1.041	.01	4.06	0	1.021	1.061	***
obese	.999	.013	-0.05	.96	.975	1.024	
white	.851	.009	-15.35	0	.834	.869	***
rural	.877	.008	-15.11	0	.862	.892	***
household	.871	.003	-34.88	0	.864	.878	***
flu_binary	1.575	.014	52.37	0	1.548	1.602	***
post_highschool	.955	.009	-4.92	0	.938	.973	***
profexam	.546	.005	-62.17	0	.536	.556	***
children	1.878	.021	55.42	0	1.837	1.921	***
age_1834	.167	.007	-41.91	0	.154	.182	***
age_3554	3.236	.135	28.13	0	2.982	3.512	***
age_5574	11.28	.486	56.26	0	10.367	12.274	***
age_75plus	3.308	.14	28.30	0	3.045	3.594	***
Constant	1.255	.066	4.33	0	1.132	1.391	***

Mean dependent var	0.771	SD dependent var	0.420
Pseudo r-squared	0.316	Number of obs	546369.000
Chi-square	185846.443	Prob > chi2	0.000
Akaike crit. (AIC)	402442.138	Bayesian crit. (BIC)	402688.781

**** p<.01, ** p<.05, * p<.1*

TABLE 18: (Odds ratios: income_high##ins_0)

Logistic regression

hadmamyes	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
0b.income_high	1	.	.	.	.	.	
1.income_high	1.106	.018	6.05	0	1.071	1.143	***
0b.ins_0	1	.	.	.	.	.	
1.ins_0	.573	.008	-40.65	0	.558	.589	***
0b.income_high#0b.~0	1	.	.	.	.	.	
0b.income_high#1o.~0	1	.	.	.	.	.	
1o.income_high#0b.~0	1	.	.	.	.	.	
1.income_high#1.in~0	.893	.039	-2.60	.009	.82	.973	***
employ	1.05	.002	24.59	0	1.046	1.054	***
weight	1	0	-4.90	0	1	1	***
bmi	1.003	.001	3.68	0	1.001	1.004	***
drink_mo	.999	0	-2.32	.02	.999	1	**
married	1.338	.012	31.73	0	1.314	1.362	***
exer_binary	1.04	.01	4.05	0	1.021	1.061	***
obese	.999	.013	-0.04	.965	.975	1.024	
white	.851	.009	-15.43	0	.833	.868	***
rural	.877	.008	-15.14	0	.862	.892	***
household	.871	.003	-34.96	0	.864	.878	***
flu_binary	1.576	.014	52.49	0	1.549	1.603	***
post_highschool	.956	.009	-4.83	0	.938	.974	***
profexam	.546	.005	-62.39	0	.535	.556	***
children	1.877	.021	55.41	0	1.836	1.92	***
age_1834	.17	.007	-41.92	0	.156	.185	***
age_3554	3.284	.136	28.74	0	3.028	3.562	***
age_5574	11.449	.489	57.08	0	10.53	12.449	***
age_75plus	3.354	.141	28.88	0	3.09	3.641	***
Constant	2.161	.109	15.21	0	1.956	2.386	***

Mean dependent var	0.771	SD dependent var	0.420
Pseudo r-squared	0.316	Number of obs	547076.000
Chi-square	186184.697	Prob > chi2	0.000
Akaike crit. (AIC)	403066.192	Bayesian crit. (BIC)	403312.863

*** *p<.01*, ** *p<.05*, * *p<.1*