Estimating Female Labor Force Participation through Statistical and Machine Learning Methods: A Comparison

Omar Zambrano[a], Claudio M. Rocco S.[b], Marco Muselli[c]

[a] Banco Central de Venezuela and Kennedy School of Government MPA/ID'05
[b]Universidad Central Venezuela, Facultad de Ingeniería
[c] Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni, CNR, Genova, Italy
[a]e-mail: Omar_Zambrano@ksg05.harvard.edu
[b]e-mail: crocco@reacciun.ve
[c]e-mail: marco.muselli@ieiit.cnr.it

Abstract

In this paper, we compare four techniques to estimate Female Labor Force Participation. Two of them, Probit and Logit are from the statistical area, while Support Vector Machines (SVM) and Hamming Clustering (HC) are from the machine learning paradigm. The comparison is performed using data from the Venezuelan Household Survey of the second semester 1999.

Keywords: Female Labor Force Participation, Probit, Logit, Support Vector Machine, Hamming Clustering.

1. INTRODUCTION

The decision of participation in the labor force, as well as its determinants, has traditionally raised wide attention as a relevant theoretical issue in both labor and behavioral economics. Moreover, basic theoretical models can be encountered in the very foundations of a variety of methodologies, including the traditional approach to aggregate supply and the more comprehensive general equilibrium models with micro foundations [1].

The specialized literature on the determinants of the labor force participation decision has put particular emphasis on variables such like wages, educational level, fertility rate or marital status as relevant explanatory variables of the labor force participation (LFP). For instance, in the less complex of the LFP models, each agent has to decide the optimal allocation of time between work and leisure, subject to the fact that every single day has a finite number of hours and any additional hour of leisure is utility improving [1]. In this model, the final allocation of time is derived from a maximization of a convex utility function subject to linear restrictions, all of this consistent with individual rational choice assumptions.

Seminal articles on female labor force participation (FLFP) theory are considered to be those led by Mincer [2] and Cain [3]. Since as a general proposition LFP depends on the relative value of leisure time, it becomes easy to imply that FLFP is affected by circumstances that can be considered exclusive of the feminine genre. Among such variables is maternity, as well as number of children, marital status, partner's income, etc. [4].

Beside traditional associated variables of LFP, there are technological and labor market aspects that affect FLFP on the individual level. In some cases, the availability of technologically superior capital goods at home can affect the woman's domestic productivity and, therefore, stimulate a higher FLFP. Labor market conditions, as part-time job availability or lasting of the daily work journey, can also allow woman to conciliate better out and in domestic work [5].

The empirical evidence regarding FLFP determinants is as diverse as abundant. In general, after controlling for some cultural and socioeconomic particularities, empirical work at local, national and cross-national level, tend to support the theoretical propositions on determinants of FLFP (e.g. wages, education, number of children, etc.).

Regarding methodological aspects, most of the reviewed studies preferred a multivariate limited variable estimation approach (Logit/Probit) to address the finding of such relationship. Additionally, it can be found the use of extreme value estimations and data panel models as complementary techniques [6].

The decision to participate in the labor market is modeled as a dichotomous random variable, which takes the value of +1 if the individual participates in the labor market and –1 if it does not.

To our best knowledge, there are no reports on the use of machine learning techniques such as SVM or HC models as binary classifier applied to the FLFP issues. In addition, also to our best knowledge, there are no researches addressing the issue and particularities of the FLFP phenomena for the Venezuelan case.

Section 2 contains an overview of the classifiers to be compared. The results of the example analyzed are presented in Section 3. Finally, Section 4 presents the conclusions.

2. Classifiers to be compared

2.1. Probit and Logit

Probit and Logit are statistical techniques for estimating the effects of a set of independent variables ($\boldsymbol{X}$) on a binary dependent variable (y). Probit and Logit avoid several statistical problems with linear probability models and generally yield results that make more sense. The inadequacies of the linear probability model suggest that a nonlinear specification is more appropriate. A natural candidate is an S-shaped curved bounded in the interval zero-one. One such curve is the cumulative normal distribution function corresponding to the Probit model. An alternative S-shaped curve is the logistic curve corresponding to the Logit model. Then the goal in Logit and Probit is to model: $\Pr(y=1|\boldsymbol{X}) = F(\boldsymbol{X\beta})$ [7].

2.2. Support Vector Machines

Suppose we have N training data points $\{(\boldsymbol{X}_1,y_1),\ldots, (\boldsymbol{X}_N,y_N)\}$, where $\boldsymbol{X}_i$, $i = 1, \ldots, N$, is a vector of input variables and y_i is the corresponding participation decision. Denote with S^+ (resp. S^-) the convex hull of the points $\boldsymbol{X}_t$ with output +1 (resp. output –1). Thus, if S^+ and S^- are linearly separable, we can think of constructing the optimal hyperplane $\boldsymbol{w}\cdot\boldsymbol{X}+b=0$, which has maximum distance from these two convex hulls. The quantities $\boldsymbol{w}$ and b are usually referred to as weight vector and bias [8]. The problem can be mathematically formulated as:

$$\begin{aligned} &\text{Min } \tfrac{1}{2}\, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} \\ &\boldsymbol{w},b \\ &\text{s.t.} \quad y_i(\boldsymbol{w}\cdot\boldsymbol{X}+b) \geq 1 \end{aligned}$$

This is a convex, quadratic programming problem in the unknowns ($\boldsymbol{w}$, b). It can be equivalently solved by searching for the values of the Lagrange multipliers α_i in the Wolfe dual problem. In this case we have

$$\boldsymbol{w} = \Sigma_i\, \alpha_i\, y_i\, \boldsymbol{X}_i$$

Only those points, which lie closest to the hyperplane, have $\alpha_i > 0$ and contribute to the above sum. These points are called *support vectors* and capture the essential information about the training set at hand.

Once we have found the optimal hyperplane, we simply determine on which side of the decision boundary a given test pattern $\boldsymbol{X}^*$ lies and assign the corresponding class label, using the function $\text{sgn}(\boldsymbol{w}\cdot\boldsymbol{X}^*+b)$.

If the two convex hulls S^+ and S^- are not linearly separable the optimal hyperplane can still be found by accepting a small number of misclassified points in the training set. A regularization factor C accounts for the trade off between training error and distance from S^+ and S^-.

To adopt non linear separating surfaces between the two classes, we can project the input vectors $\boldsymbol{X}_i$ into another high dimensional feature space through a proper mapping $\Phi(\cdot)$. If we employ the Wolfe dual problem to retrieve the optimal hyperplane in the projected space, it is not necessary to know the explicit form of the mapping Φ. We only need the inner product $K(\boldsymbol{X},\boldsymbol{X'}) = \Phi(\boldsymbol{X}) \cdot \Phi(\boldsymbol{X'})$, which is usually called *kernel function* [8]. Different choices for the kernel function have been suggested; they must verify the Mercer's condition [9]; for example, the Gaussian RBF kernel: $K(\boldsymbol{X},\boldsymbol{X'})=\exp(-\|\boldsymbol{X}-\boldsymbol{X'}\|^2/2\sigma^2)$

Classifiers obtained with this method are called *Support Vector Machines (SVM)*. The need of properly choosing the kernel is a limitation of the support vector approach. In general, the SVM with lower complexity should be selected.

2.3. Hamming Clustering

Hamming Clustering (HC) is a rule generation method, based on Boolean function reconstruction, which is able to achieve performances comparable to those of best classification techniques. The decision function built by HC can be expressed as a collection of intelligible rules in the **if-then** form, underlying the classification problem. In addition, as a byproduct of the training process, HC is able to determine redundant input variables for the analysis at hand, thus allowing a significant simplification in the data acquisition process.

Since HC operates on binary strings, problems involving numerical or nominal variables can be solved by previously coding the values assumed by each variable into a Boolean form. To this aim, ordered inputs have to be previously discretized, by dividing their domain into a collection of adjacent subintervals. The choice of a suboptimal set of cutoffs to be used as boundary values for these subintervals can be performed by adopting proper discretization methods [12]. After the coding phase the training set is

formed by pairs $(\boldsymbol{Z}_i, y_i)$, where $\boldsymbol{Z}_i$ is a binary string obtained from $\boldsymbol{X}_i$.

Then, HC proceeds by grouping together binary strings that belong to the same class and are close to each other according to the Hamming distance. A basic concept in the procedure followed by HC is the notion of *cluster*. A cluster is the collection of all the binary strings having the same values in a fixed subset of components; as an example, the four binary strings '01001', '01101', '11001', '11101' form a cluster since all of them only have the values 1, 0, and 1 in the second, the fourth and the fifth component, respectively.

The procedure employed by HC consists of the following four steps:

1. Choose at random an example $(\boldsymbol{Z}_i, y_i)$ in the training set.
2. Build a cluster of points including $\boldsymbol{Z}_i$ and associate that cluster with the class y_i.
3. Remove the example $(\boldsymbol{Z}_i, y_i)$from the training set. If the construction is not complete, go to Step 1.
4. Simplify the set of clusters generated and build the corresponding Boolean function.

The execution of HC does not involve the tuning of any parameter.

2.4. Performance of a classifier

The performance of a binary classifier (BC) is measured using sensitivity, specificity and accuracy [13]:

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}; \qquad \text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where:

TP=Number of True Positive classified cases (BC correctly classifies)

TN=Number of True Negative classified cases (BC correctly classifies)

FP=Number of False Positive classified cases (BC labels a case as positive while it is a negative)

FN=Number of False Negative classified cases (BC labels a case as negative while it is a positive)

Sensitivity gives the percentage of correctly classified participation events, whereas specificity provides the percentage of correctly classified non-participation events. In order to select the best model, we used the Noise/Signal Ratio (NSR) [15]:

NSR = (1-sensitivity)/specificity

This index measures the false signals as a ratio of the good signals issued. A good selection rule amounts to choose the model that minimizes the value of NSR.

3. EXAMPLE

We used data from the Venezuelan Household Survey (VHS) of the second semester 1999. This survey publishes official data on main socioeconomic indicators, labor market characteristics, household structure and human development variables, from a set of stratified representatives of Venezuelan households.

We choose a data set of 2497 women from the economically active labor force, between 15 and 64 years old and from to the higher quintile of the income distribution. From this data set, 1876 samples were used during the training phase, and the rest were used for the testing phase. We selected the following explanatory variables:

AGE (discrete): Number of years old.

AGE2 (discrete): **AGE** squared.

SINGLE (Boolean): women's marital status: 1 when woman has no domestic partner, 0 otherwise.

HOUSE_HEAD (Boolean): 1 when the individual woman is the head of her household, 0 otherwise.

SCHOOL (discrete): Last school year achieved and approved.

SCHOOL2 (discrete): **SCHOOL** squared.

SCHOOL_LEVEL (discrete): Last educational level achieved: 1 = less than primary, 2 = primary, 3 = secondary, 4 = college and 5 = graduate school level completed.

WAGE_PARTNER (continuous): wages and income from other sources gained by the woman's household domestic partner.

WAGE_OPP (continuous): Estimated opportunity cost for woman relative to domestic work as a point estimate of a Mincer regression [2]. It can be interpreted as an indicator of the economy wage conditions.

Even if the number of children under six years old is a widely suggested candidate as a determinant of female labor force participation, it was not included in the study. The target subset of women in the study is that belonging to the higher quintile in the income distribution, so there might be close substitutes (probably domestic workers) for the child care chores of the mothers among this group. All variables were scaled.

3.1. Logit/Probit models results

The results of the econometric models show that all the estimated coefficients for the broadest specification that included all the variables mentioned above, for both Probit and Logit estimations, resulted statistically significantly different from zero and with the expected sign. The relative magnitude as well as the sign of the estimated coefficients is a good proxy

of the relevance of each of the explanatory variables. In this particular example, wage variables were found to be a high explanatory power. Table 1 shows the performance of Probit and Logit classifiers, during both the training and testing phase. Since the dependent variable is binary, the results of testing are evaluated as follows: If the predicted probability is ≥ 0.5, then output is +1, –1 otherwise.

3.2. SVM and HC model results

In our simulations we used LIBSVM [14], an integrated software that implements SVM allowing to choose among different kernels. We tried different kernels and best results were achieved using a Gaussian kernel. The performances of the best SVM model (with C = 5000 and σ = 2.247, number of support vectors = 612) are shown in Table 1. These values are better than those obtained by Logit or Probit, but SVM does not provide any information about the importance of the explanatory variables.

Performance results for the HC model are also shown in Table 1. During the training phase HC coded the input vector into a binary string with length 65; then, it produced **52** rules explaining the FLFP and **24** rules for the non-participation. For example, one of such rules is:

IF SINGLE = 1 AND HOUSE_HEAD = 1 AND WAGE_OPP > 110000 Then Participation is 1

Table 1: Performance results for considered classifiers

TRAINING	**LOGIT**	**PROBIT**	**SVM**	**HC**
Accuracy %	86.39	84.24	88.90	**91.25**
Sensitivity %	90.05	87.63	93.31	**95.44**
Specificity %	75.69	79.78	74.31	**85.61**
NSR	0.1315	0.1551	0.0900	**0.0533**
TESTING	**LOGIT**	**PROBIT**	**SVM**	**HC**
Accuracy %	89.39	88.59	88.79	**91.80**
Sensitivity %	93.51	92.05	94.98	**97.28**
Specificity %	75.69	77.08	80.49	**73.61**
NSR	0.0857	0.1031	0.0624	**0.0370**

4. CONCLUSIONS

This paper has presented an approach to compare classifiers for the FLFP problem, belonging to the statistical and machine learning paradigms. The Logit specification performed slightly better than the Probit specification in terms of predictive ability. Both statistical models provide relevance indication for each of the explanatory variables. The SVM model produces better performance values than Probit/Logit, even if it is very difficult to interpret the information included in support vectors. Moreover, a trial-and-error procedure is required to select the best parameters. Finally, HC does not require any parameter tuning and gives the best performance values. It is also able to provide useful rules for explaining the role of the explanatory variables.

5. REFERENCES

[1] Sachs, J., Larraín, F.: *Macroeconomics in the Global Economy*. Prentice Hall, 1994.

[2] Mincer, J.: Labor Force Participation of Married Woman. In *Aspects of Labor Economics*, Universities NBER Studies Conference, No. 14, Princeton University Press, 1962, 63–97.

[3] Cain, G.: *Married Woman in the Labor Force*. University of Chicago Press, 1965.

[4] Blau, F., Ferber, M.: *The Economics of Women, Men and Work*. Prentice-Hall, 1986.

[5] McConnel, C., Brue, S.: *Labor Economics*. McGraw-Hill, 1997.

[6] Beaudry, P., Lemieux, T.: Evolution of the Female Labor Force Participation Rate in Canada, 1976-1994. Applied Research Branch, Human Resources Development Canada, W994E, 1999.

[7] Gujarati D.: Basic *Econometrics, 4th Ed*. McGraw-Hill/Irwin, 2002.

[8] Cristianini N., Shawe-Taylor J.: *An introduction to Support Vector Machines*. Cambridge University Press, 2000.

[9] Campbell C.: An Introduction to Kernel Methods, In R.J. Howlett and L.C. Jain, editors, *Radial Basis Function Networks: Design and Applications*, Springer Verlag, Berlin, 2000, 31.

[10] Muselli M., Liberati D.: Binary Rule Generation via Hamming Clustering, *IEEE Transactions on Knowledge and Data Engineering*, 14 (2002), 1258–1268.

[11] Muselli M., Liberati D.: Training Digital Circuits with Hamming Clustering. *IEEE Transactions on Circuits and Systems I* 47 (2000), 513–527.

[12] Liu, H., Setiono, R.: Feature Selection via Discretization. *IEEE Transactions on Knowledge and Data Engineering*, 9 (1997), 642–645.

[13] Veropoulos K., Campbell C., Cristianini N.: Controlling the Sensitivity of Support Vector Machines. *Proceedings of the International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 1999 (IJCAI99)*, 55–60.

[14] Chang Ch., Lin Ch.: LIBSVM: a library for support vector machines, 2001. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[15] Kaminsky G., Lizondo S., Reinhart C.: "Leading Indicators of Currency Crisis", IMF Staff Paper, No. 45, 1998