eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

EAP MASTER

FACULTY OF ECONOMICS

Pedro Iraburu Muñoz

Gustavo Julio García Bernal

CULTURAL ECONOMICS:

Analysis of the Survey of Public Participation in the Arts

4th of April, 2020
Marbella & Bilbo

Understanding the assistance to live musical stage plays

Pedro Iraburu Muñoz & Gustavo Julio García Bernal

02/04/2020

Abstract

In this short project we regress the assistance to live musical stage plays in the last 12 months, using the logit methodology, in order to truly understand the relationsips with other key variables, as well to assess the significance of the independent variables.

Keywords: Cultural Economics, musical plays, logit, Public Participation in the Arts.

1.0 Introduction

Since the popularization of compact discs, access to music has grown continuously and at an exponential rate. Nowadays, we can simply look up a song's name on Internet and we'll find plenty of virtual platforms that let us play it for free, whether it is a Beethoven's masterpiece or the latest tune of a pop star. Moreover, we can download it and play it whenever we want, in our PC or in our mobile phone. And even so, we are willing to pay and go to live musical stage plays, to listen to the same songs, but with lower sound quality and more background noise. Are we being irrational? Not at all. It is the experience what we pay for, something we cannot feel when listening to our favourite song while lying on our bed[1] . What are the reasons that make us leave our houses and go into a crowd even though the music content barely differs? In the following pages we carry out some logistic regressions to determine what socioeconomic factors are behind attendance to live musical stage plays.

2.0 Analysis of the Data

US' National Archive of Data on Arts and Culture (NADAC) will be our only data source. More precisely, we will work with the latest version of the Survey of Public Participation in the Arts (SPPA), which includes demographic and economic information at household level, together with responses to questions on public participation in the arts[2]. The data was collected in July 2017. It is a huge database, with 147.629 observations for 639 variables[3]. Therefore we will keep only a few variables, relevant for our study. We will use many variables descriptive of the socioeconomic characteristics of the household and their previous contact with art. The main variable in our analysis is *PEC1Q5A*, a dummy variable that takes value 1 when the subject attended a live musical stage play in the last 12 months. *PTC1Q5B* contains the number of plays that the individual attended, if she or he attended any.

3.0 Methodology

We will compute some descriptive statistics to get a better understanding of our variables, and afterwards we will carry out some logistic regressions attempting to estimate the decision on attendance to live musical stage plays depending on those socioeconomic indicators, which we considered important. We will start by including all the covariates and iteratively those revealed insignificant will be removed in order to develop a more efficient model.

As most variables in the SPPA come with a number of different categories, we have grouped observations in different ways. In the case of income, we gathered all observations into 3 categories: low (below 25.000$), high (above 75.000$) or medium, for all the values in the middle, and which we will take as benchmark. We also created a few dummies: for the gender (1 if male), for children (1 if any), for education level (1 if college education or above), for labour status (1 if employed) and finally, for having received lessons in any arts (music, dance, photography, etc.)[4].

[1] Earl (2001) provides a much more in-detail insight regarding why live music is being barely affected by digitalisation.

[2] More information about the data and how it is collected can be found at NADAC's website: https://www.icpsr.umich.edu/icpsrweb/NADAC/studies/37138/summary

[3] Some of these observations are lost if we take into account the deletion of the NAS within variables.

[4] For the income categories, we used the variable HEFAMINC. For gender, PESEX; for children, PRCHLD; for education, PEEDUCA; for labour status, PEMLR; and for arts lessons, a combination of PEMEQ1A, B, C, D, E, F and G, making use of the logical OR operator.

4.0 Previous Research

Many authors have studied the impact of individual characteristics on participation in the arts [5]. But there are others that have gone beyond that. For example, Kracman (1996) found that having received instruction in the arts in school has a significant influence on adult participation in the arts. Upright (2004) states that "participation in arts events is a function not just of cultural capital but [also] of social capital". He defines this concept of *social capital* as the "attributes of one's close associates", and vindicates that people's social relationships influence their adult participation in the arts. He focuses on the marital relationship. His study demonstrated a strong influence of the spouse. This influence turns out to be greater from women on husband's arts participation than on the opposite way.

There are also predecessors in the task of carrying out econometric procedures in order to estimate the impact of such factors on attendance to cultural events. As of 1983, Louviere and Hensher already estimated demand for cultural events using discrete choice models. However, they focused on unique, large-scale events such as international expositions of different nature. Hand (2009) studied attendance at live music events and introduced an interesting development: he identified different segments of audience for different types of music. For this purpose he takes into account the distinction between *omnivore* and *univore public*[6] that Peterson (1992) established.

[5] Like skin colour (DiMaggio, 1990), social class (DiMaggio, 1978) or income and educational level (O'Hagan and John, 1996).

[6] Following Peterson (1992), omnivores are those that participate in most kinds of leisure activities, whereas univores only get actively involved in one, or just a few, of those.

5.0 Understanding the dependent variable

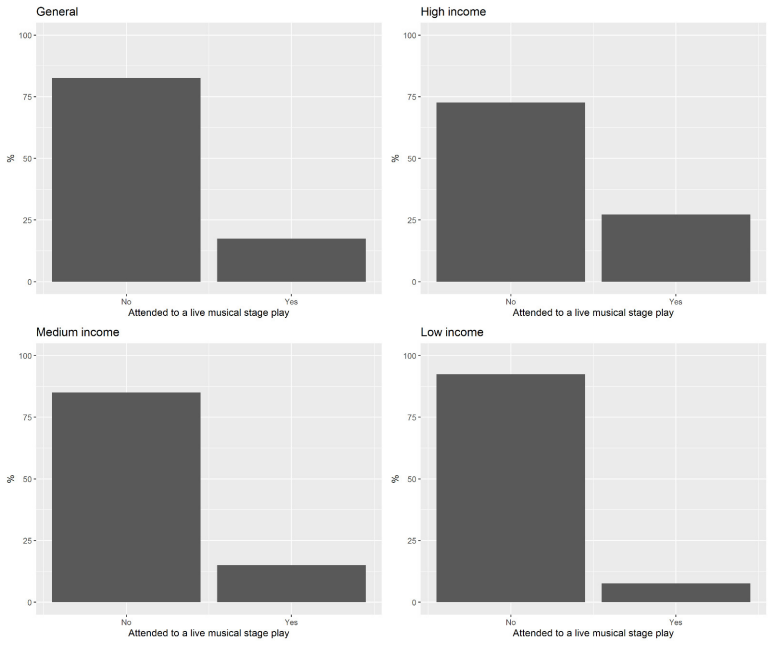
The aim of this study is to provide an accurate econometric model for the choice on participation in the arts. Figure [1] shows some histograms of this variable, without sample restriction and then dividing individuals based on their income[7]. When we look at the whole sample (Table 1 and Figure [1]), roughly a 20% has attended any live musical stage play during the last 12 months. However, this rate increases up to more than 25% when we only consider individuals with high income. Barely 10% of low income individuals has attended any play during the last year, what means a gap of more than 15 percentage points in participation in this type of events between rich and poor.

Table 1: Summary Statistics of the dependent variable

Summary Statistics of Assistance				
Assist	**N**	**N(cum.)**	**%**	**% (cum)**
0	7231	7231	0.824	0.824
1	1524	8755	0.174	0.997
NA	23	8778	0.003	1.000

Source: selfmade with R code.

Figure 1: Distribution of the Attendance to a Live Musical Stage Play

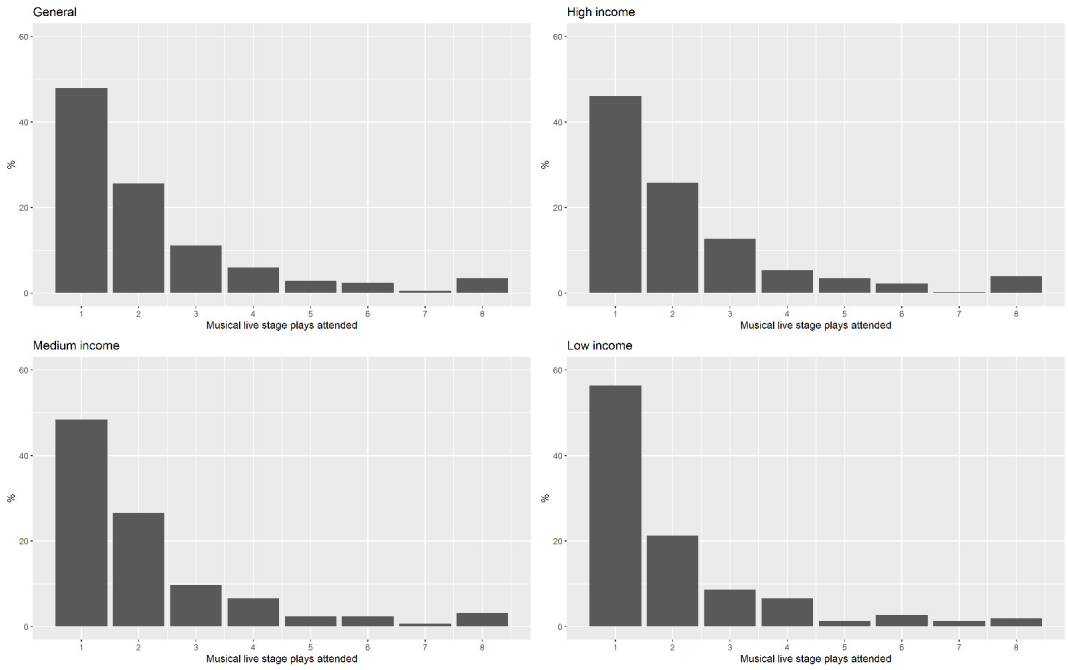


Source: selfmade with R.

[7] We chose income because it turned out to be the one generating the greatest changes in the dependent variable

Moreover, Figure [2] indicates that income is also a cause for large differences in the frequency of attendance. Around 55% of low income attendants went to just one play in the last 12 months, meaning more than 10 percentage points of difference with rich individuals.

Figure 2: Distribution of the Musical Live Stage Plays Attended



Source: selfmade with R.

6.0 Logit Regression: initial stance and improvement

6.1 Initial Stance

As already anticipated, we start the analysis including all covariates into the model and carrying out a logit regression. A logistic regression let us deal with a discrete dependent variable, a dummy in this case. Unlike in linear models, the magnitude of coefficients doesn't offer an immediate interpretation, but still their sign and the ratios between pairs of coefficients can be interpreted in a similar way as those from a linear regression. Results are collected in Table 2. Most variables are relevant, with the exception of age, the squares of age and the dummy that determines whether the individual lives in a big city. We only leave age, which without its squares stays relevant[8]. We also removed from the model the parents' education, given that, as interesting as it would be to take such information into consideration in our model, we found that there were more than 60% missing values in these variables.

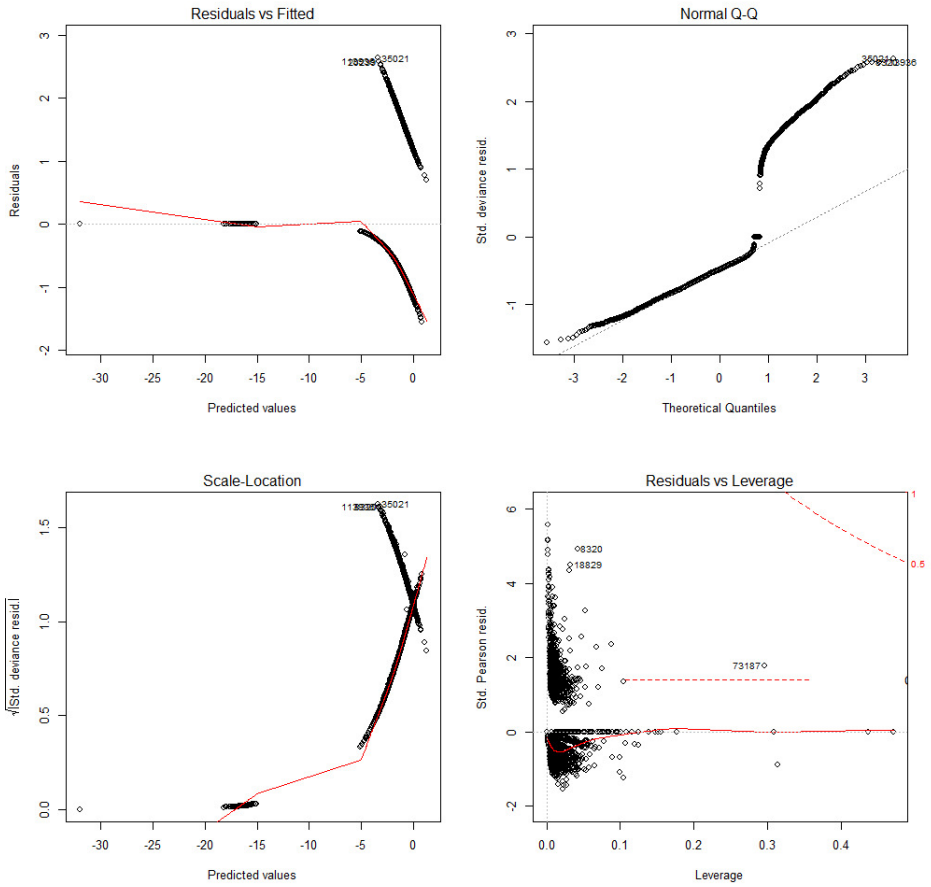
Table 2: Initial Logit Regression

	Dependent variable:
	assist
income1	-0.516^{***} (-0.859, -0.173)
income2	0.584^{***} (0.367, 0.800)
metropoli	0.114 (-0.107, 0.335)
age	0.007 (-0.032, 0.047)
age2	0.0001 (-0.0003, 0.0005)
man1	-0.643^{***} (-0.845, -0.442)
employed	0.239^{*} (-0.011, 0.489)
children	-0.368^{***} (-0.630, -0.106)
cult.lessons	0.603^{**} (0.041, 1.164)
p.educ	0.124^{***} (0.040, 0.208)
m.educ	0.150^{***} (0.066, 0.234)
univ	0.767^{***} (0.546, 0.987)
Constant	-3.362^{***} (-4.394, -2.330)
Observations	2,895
Akaike Inf. Crit.	2,592.140
Note:	*p$<$0.1; **p$<$0.05; ***p$<$0.01

[8]We made trials with age and age squared separately and both turned out to be relevant. However, age showed a lower p.value. Other important reason for the removal of $age2$ is that it induces multicollinearity in the model

From a statistical/methodological point of view, logit models are very different from the standard OLS regression. For example, in your average least squares regression, the model must met a few assumptions (error term uncorrelated with the explained variable, no multicollinearity, etc.). For a probit/logit model[9] the assumptions are very much relaxed. In Fig. [3] we can observe a 2x2 grid containing distinct graphs, a quick analysis of the initial regression. The top left figure showcases the relationship between residuals and the fitted values, a visual way of identifying heterokedasticity: if we manage to find a clear pattern. Without surprise we find that we indeed have a case of heterokedasticity[10]. In logit models, homokedasticity is not a vital assumption for unbiased estimators, even more: it is pretty usual. At the top right corner, we can observe the distribution of the errors: it shows an evident logistic distribution. As explained before, the OLS regressions must have their error terms normally distributed[11].

Figure 3: Initial Analysis



Source: selfmade with R.

[9] Where the difference between these two lies in the assumption on the distribution of the error term: logit errors follow a logistic form, whereas the probit a normal one.

[10] A Breusch-Pagan test has been also been made, whose results further supports this statement, you can find it in the annex.

[11] We also made a Normal test for the errors, see annex.

Earlier, I stated that logit regressions did not share the same assumptions for unbiasedness with the standard OLS regression. In spite of this, there is one they share, and that one is absence of perfect multicollinearity. In order to test if we do not have efficient estimator, we use the VIF method, whose results are displayed at Table 3. The VIF is a measure for correlations within the independent variables. If we have high values (more than 10), there is multicollinearity in the model. As seen in the results, only age and $age2$ are correlated (very correlated, logically). This is why we did not include $age2$ in the second instance of the logit regression.

Table 3: VIF test for Multicollinearity

	VIF	Df	VIFˆ(1/(2*Df))
income	1.173452	2	1.040798
metropoli	1.039357	1	1.019489
age	50.461689	1	7.103639
age2	50.380170	1	7.097899
man	1.019570	1	1.009738
employed	1.526841	1	1.235654
children	1.273084	1	1.128310
cult.lessons	1.579935	1	1.256955
p.educ	1.658319	1	1.287757
m.educ	1.644038	1	1.282200
univ	1.230714	1	1.109375

Source: selfmade with R code.

6.2 Improvement of the model

Table 4 contains the results of our second model. In this attempt, all our variables are clearly relevant. The signs are in line with our expectations. High income individuals participate more in live musical stage plays than medium income individuals, whereas low income individuals participate less. Men are less prone to attend these events than women, and the negative effect is of similar magnitude as of being poor. Having children under 18 years old is, as could be anticipated, another hindrance to attend these plays. However, the effect isn't as remarkable as those previously commented. Probably due to the availability of solutions: it's easier to leave our children with a babysitter than to change our economic status or the preferences associated to genders by society or by nature.

Having received lessons in arts of any kind is also relevant and introduces a positive effect. We already warned in a previous section that we should be cautious with these variable, as other individual and family characteristics, like income, may have some significant influence on this variable. However, the analysis of multicolinearity was successful and we discarded any kind of endogeneity among our regressors. The next coefficient shows that people in possession of college education or above are more likely to attend these events. Finally, employment shows also a positive sign, suggesting that the income effect overtakes the larger time endowment of those unemployed.

Table 4: Final Logit Regression

	Dependent variable:
	assist
income1	-0.591^{***} (-0.786, -0.396)
income2	0.560^{***} (0.431, 0.689)
age	0.006^{***} (0.002, 0.010)
man1	-0.618^{***} (-0.739, -0.497)
children	-0.260^{***} (-0.411, -0.110)
cult.lessons	0.725^{***} (0.281, 1.168)
univ	1.070^{***} (0.946, 1.194)
employed	0.186^{**} (0.041, 0.330)
Constant	-2.325^{***} (-2.627, -2.023)
Observations	8,755
Log Likelihood	$-3{,}650.367$
Akaike Inf. Crit.	7,318.734
Note:	*p$<$0.1; **p$<$0.05; ***p$<$0.01

6.3 Likelihood Ratio test

To test a few of our self-made dummy variables (using the original categorical variables), we will make use of the Likelihood ratio test:

```
# Likelihood ratio test
Model 1: assist ~ income + age + man + children + cult.lessons + univ +
employed
Model 2: assist ~ income + age + man + children + cult.lessons
#Df  LogLik Df  Chisq Pr(>Chisq)
1   9 -3650.4
2   7 -3806.3 -2 311.91  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The Likelihood Ratio test works by comparing two models, and estimating the difference between the log likelihood. If the difference is large, we reject the null hypothesis and thus, the larger model is a significant improvement over the smaller one.

We also included a table with the two models we used. In this case we wanted to test if the variables $univ$, a dummy that takes value one when the individual has at least a bachelor degree, and the variable $employed$, improved the model. As we can see, the p-value is well below the level of significance: $\alpha = 0.05$. We can conclude that those two variables better our model[12]

Table 5: Regression Results from the LR test

	Dependent variable:	
	assist	
	(1)	(2)
income1	-0.591^{***} (0.100)	-0.826^{***} (0.096)
income2	0.560^{***} (0.066)	0.824^{***} (0.063)
age	0.006^{***} (0.002)	0.003 (0.002)
man1	-0.618^{***} (0.062)	-0.612^{***} (0.060)
children	-0.260^{***} (0.077)	-0.230^{***} (0.075)
cult.lessons	0.725^{***} (0.226)	0.475^{**} (0.222)
univ	1.070^{***} (0.063)	
employed	0.186^{**} (0.074)	
Constant	-2.325^{***} (0.154)	-1.586^{***} (0.117)
Observations	8,755	8,755
Log Likelihood	−3,650.367	−3,806.322
Akaike Inf. Crit.	7,318.734	7,626.643
Note:		*p<0.1; **p<0.05; ***p<0.01

[12] This is also easily concluded looking at the significance level, this was made to showcase the test as a measure of the goodness of fit, and also to have doubly robust results.

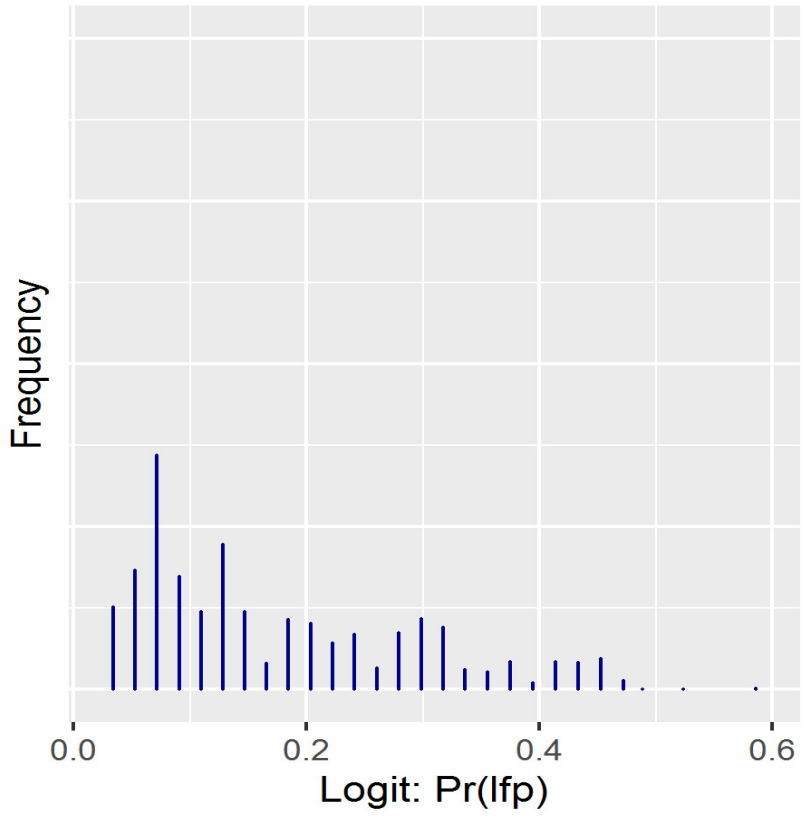
7.0 Results of our study

In order to understand properly the results we have obtained in the model, we propose and explain three approaches:

7.1 Plot of the distribution of probabilities

First of all we compute all of the predicted probabilities for the whole sample, and we plot them using R. As you can appreciate, the y-axis does not have any measurable values. This is because for the $ggplot$ function from R, displaying the count, or even the relative frequency, is pretty difficult. Fig [4] serves as an intuitive analysis of the distribution of the estimated probabilities. As seen, the majority of the individuals would have their estimated probabilities between the 0.00-0.25 values for the probability of assistance.

Figure 4: Distribution of Probabilities



Source: selfmade with R.

7.2 CDF of two given scenarios

The second approach we used to analyze the results, was computing the CDF of two different hypothetical individuals: one rich woman, educated, with no children, who assists/has assisted to cultural lessons, and is employed; and a poor woman, not educated, without children, and unemployed. We chose the age as the x-axis variable for the CDF as is, conveniently, a continuous variable.

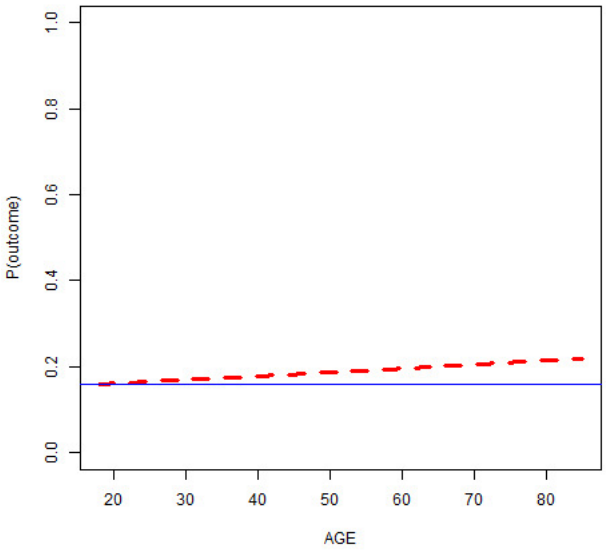


Figure 5: CDF of a Rich Individual

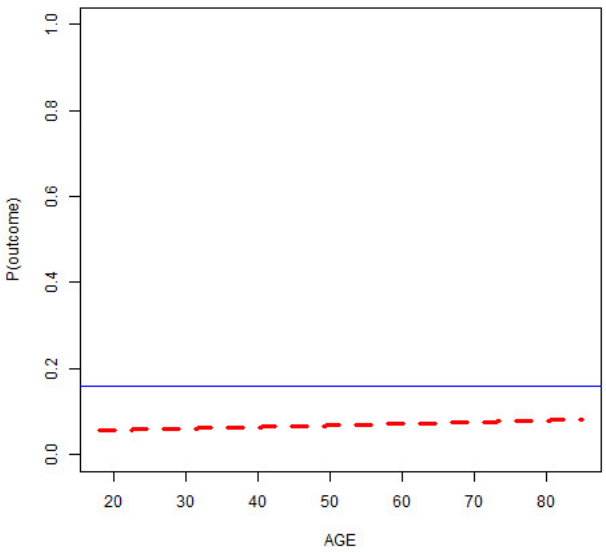


Figure 6: CDF of a Poor Individual

As we can appreciate, not only the intercept of the CDF changes drastically within the individuals, but also the slope changes drastically. We can also induct from the fig [5] and [6], that age has a direct relationship with the probability of assisting to live musical stage plays. This can be explained by the higher monetary stability of older individuals, as well as by the fact that they have a bigger portion of free time than young/employed persons.

7.3 Discrete Changes in Probability

Finally, we computed the discrete changes in the probability of assisting to these live concerts by comparing to a benchmark, different individuals. In order to properly asses the effects of these changes in the variables, only one of them was altered when compared to the benchmark.

The benchmark is a female individual of average age, neither rich nor poor, with children, not educated but employed, that has not received any cultural lessons. We compared this benchmark to individuals with the same characteristics but one: a higher income, a lesser income, one who assisted cultural lessons, unemployed and finally, one that is male. This effects are displayed at the Table 6. All of the effects are consistent with the sign and value of the estimated coefficients of the same variables.

Table 6: Results

	Discrete Change in Probability
Being Poor	-0.04573018
Being Rich	0.06757878
Being a Man	-0.0473474
Attending Lessons	0.09293273
Not Employed	-0.01683791

Source: selfmade with R code.

8.0 Conclusion and potential improvements

In our study, we provided a model for the estimation of probabilities to attend a live music play considering several socioeconomic factors. We verified that such elements have a determinant influence on this variable. Income, being a woman and possessing college education stand out as the most determinant factors tipping the scales in favour of attending a play. By contrast, although less significantly, having children affects negatively the probability of our dependent variable. Parents' limitation is not as influential as one might think probably due to the availability of solutions (babysitters, other family members, etc.). Our findings can be complemented with those of Upright (2004), who found that the wife was more influential on husband's participation in arts than the other way around. Our results cannot support (or reject) this statement, but we can assert that such influence will be, on average, positive, as women are generally more prone to attend live musical stage plays.

It would be also interesting to control for heterogeneity between different states and the SPPA provides useful information for such analysis[13]. However, such a thorough analysis was beyond the scope of our study. So is the estimation of the number of sessions attended, by those who attended any. A multinomial model could be estimated for this purpose.

We could also incorporate the aforementioned findings in Upright (2004) to the model, although some information about the spouse would be needed. The educational level of the subject's husband or wife may be of greater relevance than that of the parents. This insight also applies to having received lessons on arts and other variables that are relevant to determine the spouse's own participation in arts. We do have taken previous instruction in arts into consideration, following Kracman (1996) results. However, our variable measures courses out of school, whose probability -as the author points out in the paper- is considerably more affected by other characteristics as race or parents' formation than the probability of receiving in-school arts lessons. That is, our variable should be replaced with school-based arts instruction in order to control for possible endogeneity of this covariate. Finally, we couldn't split the demand for different type of music as in Hand (2009) due to the lack of information on the events' nature.

[13] Variables GEREG, GEDIV and GCFIP

References

[1] DiMaggio, P. and Ostrower, F. (1990): *"Participation in the arts by black and white Americans."* Social Forces, 68, 753–778.

[2] DiMaggio, P. and Useem, M. (1978): *"Social class and arts consumption: the origins of class differences in exposure to the arts in America."* Theory and Society, 5, 141–161.

[3] Earl, P. E. (2001): *"Simon's travel theorem and the demand for live music."* Journal of economic psychology, 22(3), 335-358.

[4] Hand, C. (2009): *"Modelling patterns of attendance at performing arts events: The case of music in the United Kingdom."* Creative Industries Journal, 2(3), 259-271.

[5] Kracman, K. (1996): *"Modelling patterns of attendance at performing arts events: The case of music in the United Kingdom."* Poetics, 24(2-4), 203-218.

[6] Louviere, J. J., and Hensher, D. A. (1983): *"Using discrete choice models with experimental design data to forecast consumer demand for a unique cultural event."* Journal of Consumer research, 10(3), 348-361.

[7] O'Hagan and John W. (1996): *"Access to and participation in the arts: the case of those with low incomes/educational attainment."* Journal of Cultural Economics, 20, 260–282.

[8] ICPSR (2020): *"National Archive of Data on Arts and Culture."* University of Michigan.

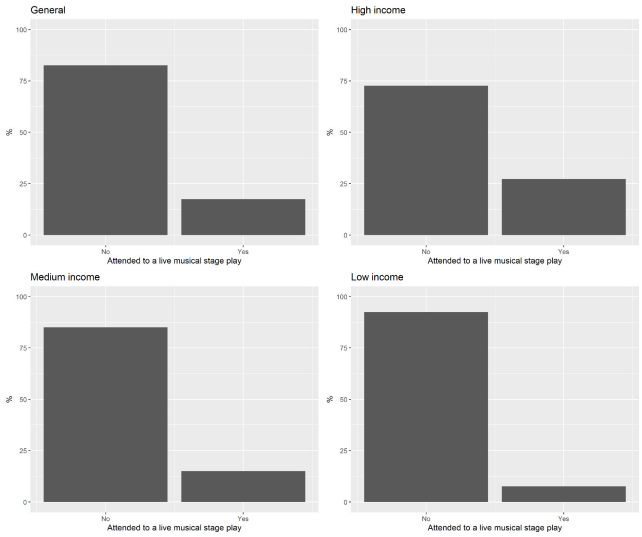
[9] Peterson, R. A. (1992). : *"Understanding audience segmentation: From elite and mass to omnivore and univore."* Poetics, 21(4), 243-258.

[10] Upright, C. B. (2004): *"Social capital and cultural participation: spousal influences on attendance at arts events."* Poetics, 32(2), 129-143.

9.0 ANNEX

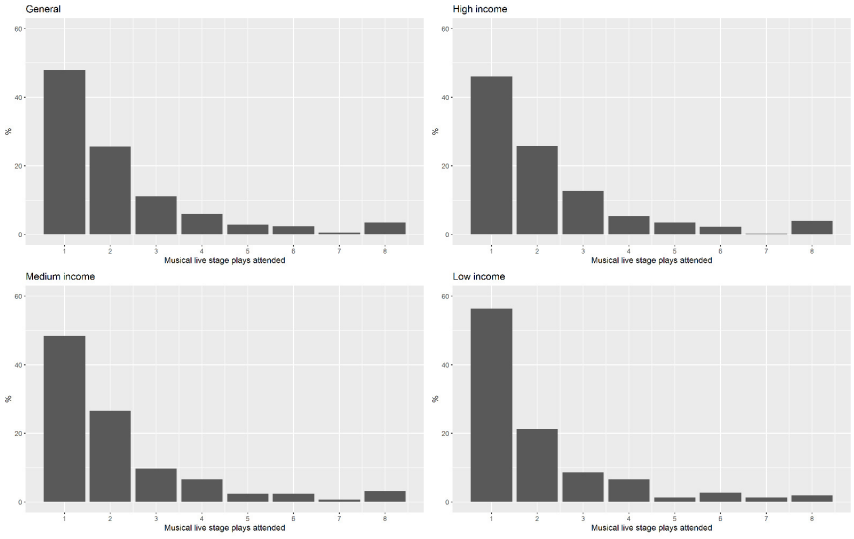
9.1 Figures

Distribution of the Attendance to a Live Musical Stage Play



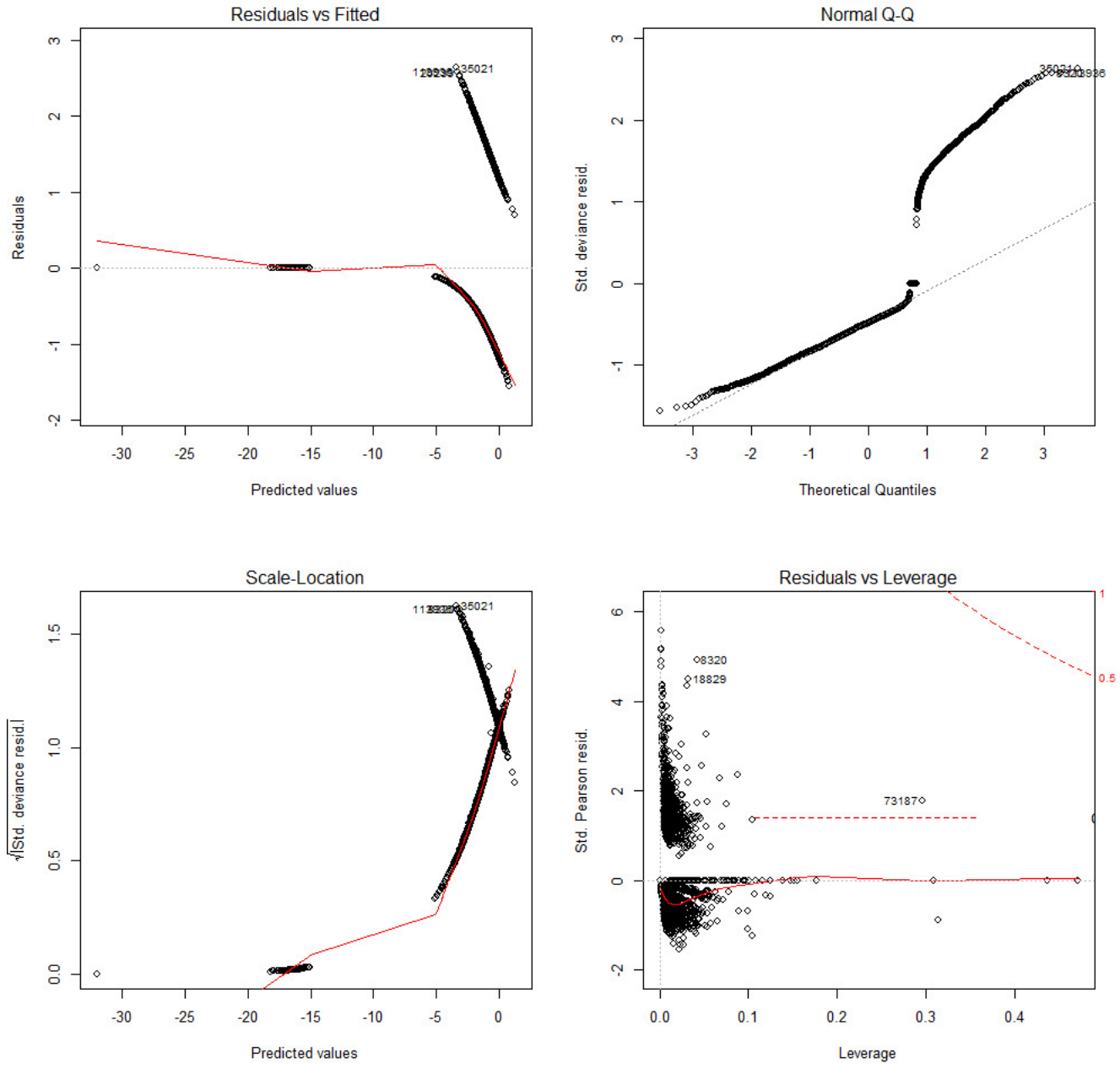
Source: selfmade with R.

Distribution of the Musical Live Stage Plays Attended



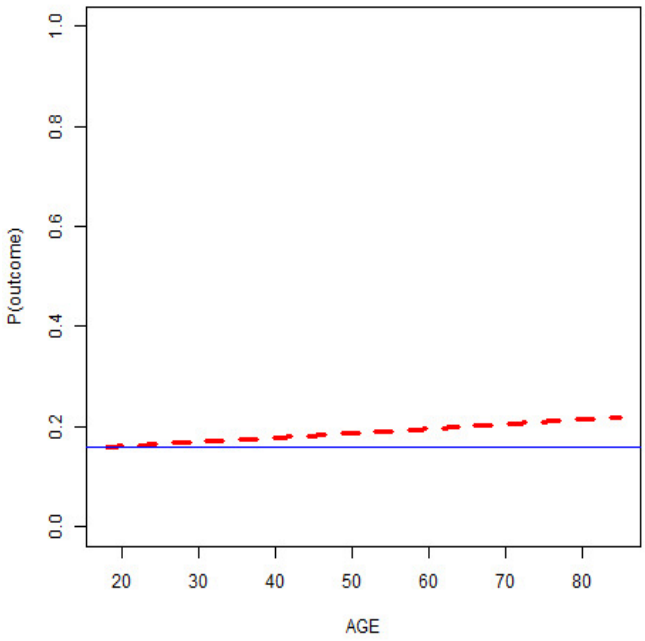
Source: selfmade with R.

Initial Analysis



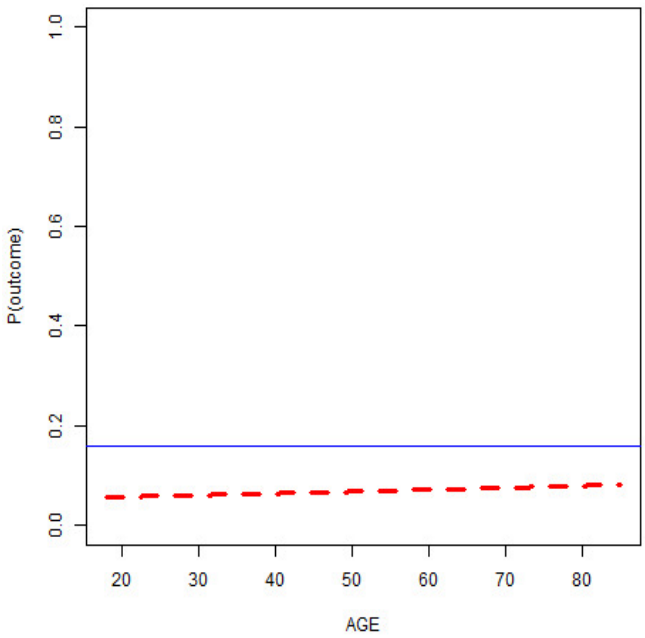
Source: selfmade with R.

CDF of a Rich Individual



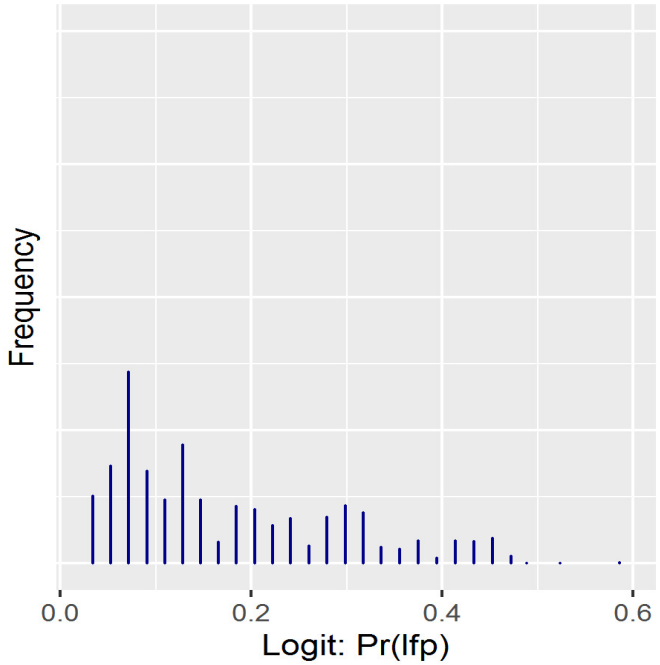
Source: selfmade with R.

CDF of a Poor Individual



Source: selfmade with R.

Distribution of Probabilities



Source: selfmade with R.

9.2 Tests and Tables

Breusch-Pagan test

```
# BP TEST FOR THE INITIAL LOGIT MODEL
studentized Breusch-Pagan test

data:  initial.logit
BP = 348, df = 12, p-value < 2.2e-16
```

Normal distribution test for the errors

```
# NORMALITY TEST FOR THE ERRORS
# OF THE INITIAL LOGIT MODEL

Shapiro-Wilk normality test

data:  initial.logit$residuals
W = 0.54693, p-value < 2.2e-16
```

Likelihood Ratio test

```
# Likelihood ratio test
Model 1: assist ~ income + age + man + children + cult.lessons + univ +
employed
Model 2: assist ~ income + age + man + children + cult.lessons
#Df  LogLik Df  Chisq Pr(>Chisq)
1   9 -3650.4
2   7 -3806.3 -2 311.91  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Summary Statistics of the dependent variable

Summary Statistics of Assistance				
Assist	**N**	**N(cum.)**	**%**	**% (cum)**
0	7231	7231	0.824	0.824
1	1524	8755	0.174	0.997
NA	23	8778	0.003	1.000

Source: selfmade with R.

Initial Logit Regression

	Dependent variable:
	assist
income1	-0.516^{***} (-0.859, -0.173)
income2	0.584^{***} (0.367, 0.800)
metropoli	0.114 (-0.107, 0.335)
age	0.007 (-0.032, 0.047)
age2	0.0001 (-0.0003, 0.0005)
man1	-0.643^{***} (-0.845, -0.442)
employed	0.239^{*} (-0.011, 0.489)
children	-0.368^{***} (-0.630, -0.106)
cult.lessons	0.603^{**} (0.041, 1.164)
p.educ	0.124^{***} (0.040, 0.208)
m.educ	0.150^{***} (0.066, 0.234)
univ	0.767^{***} (0.546, 0.987)
Constant	-3.362^{***} (-4.394, -2.330)
Observations	2,895
Akaike Inf. Crit.	2,592.140
Note:	*p$<$0.1; **p$<$0.05; ***p$<$0.01

VIF test for Multicollinearity

	VIF	Df	VIF^(1/(2*Df))
income	1.173452	2	1.040798
metropoli	1.039357	1	1.019489
age	50.461689	1	7.103639
age2	50.380170	1	7.097899
man	1.019570	1	1.009738
employed	1.526841	1	1.235654
children	1.273084	1	1.128310
cult.lessons	1.579935	1	1.256955
p.educ	1.658319	1	1.287757
m.educ	1.644038	1	1.282200
univ	1.230714	1	1.109375

Final Logit Regression

	Dependent variable:
	assist
income1	-0.591^{***} (-0.786, -0.396)
income2	0.560^{***} (0.431, 0.689)
age	0.006^{***} (0.002, 0.010)
man1	-0.618^{***} (-0.739, -0.497)
children	-0.260^{***} (-0.411, -0.110)
cult.lessons	0.725^{***} (0.281, 1.168)
univ	1.070^{***} (0.946, 1.194)
employed	0.186^{**} (0.041, 0.330)
Constant	-2.325^{***} (-2.627, -2.023)
Observations	8,755
Log Likelihood	$-3{,}650.367$
Akaike Inf. Crit.	7,318.734
Note:	*p$<$0.1; **p$<$0.05; ***p$<$0.01

Regression Results from the LR test

	Dependent variable:	
	assist	
	(1)	(2)
income1	−0.591*** (0.100)	−0.826*** (0.096)
income2	0.560*** (0.066)	0.824*** (0.063)
age	0.006*** (0.002)	0.003 (0.002)
man1	−0.618*** (0.062)	−0.612*** (0.060)
children	−0.260*** (0.077)	−0.230*** (0.075)
cult.lessons	0.725*** (0.226)	0.475** (0.222)
univ	1.070*** (0.063)	
employed	0.186** (0.074)	
Constant	−2.325*** (0.154)	−1.586*** (0.117)
Observations	8,755	8,755
Log Likelihood	−3,650.367	−3,806.322
Akaike Inf. Crit.	7,318.734	7,626.643
Note:		*p<0.1; **p<0.05; ***p<0.01

Results

	Discrete Change in Probability
Being Poor	-0.04573018
Being Rich	0.06757878
Being a Man	-0.0473474
Attending Lessons	0.09293273
Not Employed	-0.01683791

9.3 R code

```
# Master in Economics: Empirical Applications and Policies (EAP)                #
# Growth and Development                                                        #
# Homework 1                                                                    #
# Pedro & Gustavo                                                               #


#### DATA MANIP                                                                ####
# Preliminaries                                                                 #
rm(list = ls())                                                  # Clear workspace
setwd("C:/Users/pimpeter/Desktop/Cultural/hw3")    # Setting directory
library(dplyr)                                    # na_if() included in this library
library(ggplot2)                                 # ggplot() included in this library
data <- read.csv("notanreduc.csv", header=T, sep=";", dec=",")     # Loading data
#View(data)
#summary(data)
#dim(data)

# Renombrar variables
colnames(data) <- c("id", "assist", "funciones", "income", "tipociudad", "CSAcod",
"age", "est.civil", "man", "raza", "famtipo", "pais",
"est.lab", "horas.trab", "trab.tipo", "trab.ind", "trab.ocu",
"major.ind", "major.ocu", "weight", "menores", "libros",
"clases.mus", "clases.foto", "clases.artes.vis", "clases.act",
"clases.danza", "clases.esc", "clases.arte.apr", "p.educ",
"m.educ", "educ")

# Mantener las observaciones válidas y arreglar variables:
# Variable: asiste. ¿Ha asistido a algún espectáculo en el último año?
# quitar -9 (no response), -3 (refused) y -1 (not in universe)
data <- data[data$asiste != -9 & data$asiste != -3 & data$asiste != -1, ]
data$asiste <- na_if(data$asiste, "#N/A")
#class(data$asiste)
data$asiste <- as.numeric(as.character(data$asiste))
# sum(as.numeric(data$asiste==0))     # 1524 yes and 7231 no, as in codebook ¡BIEN!
# Variable: funciones. ¿A cuántas funciones ha asistido?
```

```
# quitar -1 (not in universe), pero conservar aquellos para los que tenemos
# datos de asistencia
data$funciones <- na_if(data$funciones, "#N/A")
#class(data$funciones)
data$funciones <- as.numeric(as.character(data$funciones))
data[is.na(data$funciones),]$funciones <- -1
data$funciones <- ifelse(data$asiste==0, 0, data$funciones)
data$funciones <- na_if(data$funciones, -1)
# sum(as.numeric(data$funciones==5), na.rm=T)     # 727 once,  44 went to 5, etc. ¡BIEN!
# Variables de educación: p.educ y m.educ Convertir <0 en NA
data[data$p.educ<0,]$p.educ <- rep(-1, length(data[data$p.educ<0,]$p.educ))
data$p.educ <- na_if(data$p.educ, -1)
data[data$p.educ<0,]$p.educ <- -1
data$p.mduc <- na_if(data$m.educ, -1)

# Nuevas variables:
# age al cuadrado
data$age2 <- data$age^2
# gran ciudad, dummy
data$metropoli <- as.numeric(data$tipociudad==1)
# clases cultura
data$clases.cult <- as.numeric(data$clases.mus==1 | data$clases.foto==1 |
data$clases.artes.vis==1 | data$clases.act==1 |
data$clases.danza==1 | data$clases.esc==1 |
data$clases.arte.apr==1)



# Recolocando variables:
# Variable de income: reducir categorías
data$income <- as.factor(ifelse(data$income < 8, 1, ifelse(data$income > 13,2,0)))
# Variable de educación: reducir categorías
data$univ <- as.numeric(data$educ>41)
# Variable de empleo: reducir categorías
data$employed <- as.numeric(data$est.lab<3)
# Horas trab. como factor (inital.logit)
data$horas.trab <- as.factor(data$horas.trab)
# Educ. como factor (inital.logit)
```

```
data$educ <- as.factor(data$educ)
# Variable menores: dummy =1 si hay children <14 años
data$children <- as.numeric(data$menores >0)
# Menores como factor
data$menores <- as.factor(data$menores)


# Descripción de las variables #####
library(scales)
summary(data$asiste)
summary(data$funciones)
summary(data[data$asiste==1,]$funciones)
data %>%
group_by(asiste) %>%
summarise(n = n()) %>%
mutate(totalN = (cumsum(n)),
percent = round((n / sum(n)), 3),
cumpercent = round(cumsum(freq = n / sum(n)),3))
data[data$asiste==1,] %>%
group_by(funciones) %>%
summarise(n = n()) %>%
mutate(totalN = (cumsum(n)),
percent = round((n / sum(n)), 3),
cumpercent = round(cumsum(freq = n / sum(n)),3))
panel1 <- ggplot(data) +
geom_bar(aes(asiste, 100*(..count..)/sum(..count..))) +
ylab("%") + labs(title="General") +
scale_x_continuous("Attended to a live musical stage play", breaks=c(0,1), labels=c("No","Yes")) +
ylim(0,100)
panel2 <- ggplot(data[data$income==2,]) +
geom_bar(aes(asiste, 100*(..count..)/sum(..count..))) +
ylab("%") + labs(title="High income") +
scale_x_continuous("Attended to a live musical stage play", breaks=c(0,1), labels=c("No","Yes")) +
ylim(0,100)
panel3 <- ggplot(data[data$income==0,]) +
geom_bar(aes(asiste, 100*(..count..)/sum(..count..))) +
ylab("%") + labs(title="Medium income") +
scale_x_continuous("Attended to a live musical stage play", breaks=c(0,1), labels=c("No","Yes")) +
```

```
ylim(0,100)
panel4 <- ggplot(data[data$income==1,]) +
geom_bar(aes(asiste, 100*(..count..)/sum(..count..))) +
ylab("%") + labs(title="Low income") +
scale_x_continuous("Attended to a live musical stage play", breaks=c(0,1), labels=c("No","Yes")) +
ylim(0,100)
png(file="asiste_hist.png", pointsize=22, width=12, height=10, units='in', res=300)
grid.arrange(panel1, panel2, panel3, panel4, ncol=2, nrow=2)
dev.off()

panelA <- ggplot(data[data$asiste==1,]) +
geom_bar(aes(funciones, 100*(..count..)/sum(..count..))) +
ylab("%") + ylim(0,60) + labs(title="General") +
scale_x_continuous("Musical live stage plays attended", breaks=c(1:8))
panelB <- ggplot(data[data$asiste==1 & data$income==2,]) +
geom_bar(aes(funciones, 100*(..count..)/sum(..count..))) +
ylab("%") + ylim(0,60) + labs(title="High income") +
scale_x_continuous("Musical live stage plays attended", breaks=c(1:8))
panelC <- ggplot(data[data$asiste==1 & data$income==0,]) +
geom_bar(aes(funciones, 100*(..count..)/sum(..count..))) +
ylab("%") + ylim(0,60) + labs(title="Medium income") +
scale_x_continuous("Musical live stage plays attended", breaks=c(1:8))
panelD <- ggplot(data[data$asiste==1 & data$income==1,]) +
geom_bar(aes(funciones, 100*(..count..)/sum(..count..))) +
ylab("%") + ylim(0,60) + labs(title="Low income") +
scale_x_continuous("Musical live stage plays attended", breaks=c(1:8))
png(file="funciones_hist.png", pointsize=22, width=16, height=10, units='in', res=300)
grid.arrange(panelA, panelB, panelC, panelD, ncol=2, nrow=2)
dev.off()
# Modelo logit #####
sapply(data,class)

library(stargazer)
initial.logit <- glm(asiste ~ income + metropoli + age + age2 + man + employed +
children + clases.cult + p.educ + m.educ + univ,
data = data, family = "binomial"(link = "logit"))
summary(initial.logit)
stargazer(initial.logit, omit.stat=c("LL","ser","f"), no.space=TRUE, single.row = T
```

```
,ci=TRUE, ci.level=0.95)

# ANALYSIS OF LOGIT REGRESSIONS #####
# Let's check for problems: heterokedasticity, normality of error term
par(mfrow=c(2,2))
jpeg(filename="analysis.jpeg")
plot(initial.logit)
dev.off()


# It looks that there is heterokedasticity: BREUSCH PAGAN TEST

library(lmtest)
bptest(initial.logit)
# We reject the null hypothesis of homocedasticity: NO PROBLEMO
# Heterokedasticity happens in logistic regressions.

# As we can also see, the errors are distributed almost normally
shapiro.test(initial.logit$residuals)
# The logistic regression does not need to fulfill normality of errors

# THE MOST IMPORTANT ASSUMPTION IN LOGISTIC REGRESSIONS
# IS THE NON EXISTANCE OF MC.

# ANALYSIS OF LOGIT REGRESSIONS 1)
sum(!is.na(data$p.educ))
sum(!is.na(data$m.educ))
sum(!is.na(data$horas.trab))
# TOO MANY NAS: DELETING THIS VARS.

# ANALYSIS OF LOGIT REGRESSIONS 2)

# CORR AND VIF TEST: CHECKING FOR MC
# THE CORRELATIONS BETWEEN INDEP VAR SHOW THAT THE VAR P.EDUC HAS MANY NAS
# WE SHOULD DELETE IT

subset   <- data[,c("income","metropoli","age" ,"age2"  , "horas.trab",
"menores" , "clases.cult", "m.educ" , "educ")]
```

```
cor(subset, method = "pearson")

# VIF TEST: age AND age2 CORRELATED
library(car)
vif(initial.logit)
# Checking if all vars are relevant: we eliminate metropoli and worked.hours
initial.logit <- glm(asiste ~ income + metropoli + age + age2 + man + horas.trab +
menores + clases.cult + p.educ + m.educ + educ,
data = data, family = "binomial"(link = "logit"))
summary(initial.logit)

# FINAL MODEL ######
final.logit <- glm(asiste ~ income  + age  + man  +
children + clases.cult  + univ+ employed,
data = data, family = "binomial"(link = "logit"))
summary(final.logit)
stargazer(final.logit, no.space=TRUE, single.row = T)

# ALL VAR ARE RELEVANT AT ALPHA=0.05
# MOST RELEVANT EFFECTS SEX,EDUCATION, CLASES CULT.
# LETS DO A LRTEST TO CHECK IF THE MODEL HAS IMPROVED



a.logit <- glm(asiste ~ income  + age  + man  +
children + clases.cult +univ+ employed ,
data = data, family = "binomial"(link = "logit"))

b.logit <- glm(asiste ~ income  + age  + man  +
children + clases.cult ,
data = a.logit$model, family = "binomial"(link = "logit"))

stargazer(a.logit, b.logit, title="Regression Results", no.space=TRUE, single.row = T)
library(lmtest)
# LR test #
lrtest.default(a.logit, b.logit)
# AS WE CAN SEE THE PR(CHI)>0.05: THE SMALLER MODEL IS A SIGNIFICANT IMPROVEMENT
# OVER THE NEW ONE
```

```
# GRAPHICAL REPRESENTATION: #########

# Plot of the  distribution of probabilities
prlogit <- final.logit$fitted.values
summary(final.logit)
dev.off()
ggplot(final.logit, aes(x = prlogit)) +geom_dotplot(dotsize = 0.0080, stackdir = "up")
histogram(prlogit)

ggplot(final.logit, aes(x = prlogit)) +geom_dotplot(dotsize = 0.0090, stackdir = "up", color="darkblue") +
plot.title = element_text(hjust = 0.5),axis.text.y = element_blank(), axis.ticks.y = element_blank()
)
# saving as jpeg
ggsave("distrib_prob.jpeg")
dev.off()
# dotplot that showcases the frequency of the predicted probabilities of our sample

# CDF : PROBABILITY OF ASSISTANCE GIVEN AGE FOR A WELL OFF INDIVIDUAL

#  CDF

# Coefficients
b0          <- final.logit$coef[1] # intercept
income       <- final.logit$coef[3]
age         <- final.logit$coef[4]
man      <- final.logit$coef[5]
children       <- final.logit$coef[6]
clases.cult <- final.logit$coef[7]
univ        <- final.logit$coef[8]
employed    <- final.logit$coef[9]



# Range of the indep. variable
X_range <- seq(from=min(data$age ), to=max(data$age ), by=1)

# Values (WE CONSIDER DUMMIES=0)
a_logits = b0 + income + age *X_range
```

```
+ clases.cult + univ +employed


#Probabilities
a_probs <- exp(a_logits)/(1 + exp(a_logits))

# Plotting CDF
plot(X_range, a_probs,
ylim=c(0,1),
type="l",
lwd=3,
lty=2,
col="red",
xlab="AGE", ylab="P(outcome)", main="Probability of Assistance: rich individual")
abline(h=0.16, col="blue")
# saving as jpeg
ggsave("cdf_rich.jpeg")
dev.off()


# Coefficients
b0          <- final.logit$coef[1] # intercept
income       <- final.logit$coef[2]
age         <- final.logit$coef[4]
man      <- final.logit$coef[5]
children       <- final.logit$coef[6]
clases.cult <- final.logit$coef[7]
univ        <- final.logit$coef[8]
employed    <- final.logit$coef[9]


# Range of the indep. variable
X_range <- seq(from=min(data$age ), to=max(data$age ), by=1)

# Values (WE CONSIDER DUMMIES=0)
a_logits = b0 + income + age *X_range
```

```
#Probabilities
a_probs <- exp(a_logits)/(1 + exp(a_logits))

# Plotting CDF
plot(X_range, a_probs,
ylim=c(0,1),
type="l",
lwd=3,
lty=2,
col="red",
xlab="AGE", ylab="P(outcome)", main="Probability of Assistance: poor individual")
abline(h=0.16, col="blue")
# saving as jpeg
ggsave("cdf_poor.jpeg")



# INTERPRETATIONS ####
# DISCRETE CHANGE IN PROBABILITY: if income changes for example

# Benchmark clase media:
bm <- data.frame(
income            <- as.factor(0),
age              <- median(data$age ),
man           <- as.factor(0),
children            <- 1,
clases.cult      <- 0,
univ             <- 0,
employed         <- 1
)

benchmark_prob  <- predict(final.logit, bm,type="response")


#income (pobre)
```

```
rp <- data.frame(
income            <- as.factor(1),
age              <- median(data$age ),
man           <- as.factor(0),
children            <- 1,
clases.cult      <- 0,
univ             <- 0,
employed         <- 1
)

rp_prob  <- predict(final.logit, rp,type="response")

#income (rico)

rr <- data.frame(
income            <- as.factor(2),
age              <- median(data$age ),
man           <- as.factor(0),
children            <- 1,
clases.cult      <- 0,
univ             <- 0,
employed         <- 1
)

rr_prob  <- predict(final.logit, rr,type="response")

#ser man

h  <- data.frame(
income            <- as.factor(0),
age              <- median(data$age ),
man           <- as.factor(1),
children            <- 1,
clases.cult      <- 0,
univ             <- 0,
employed         <- 1
)
```

```
h_prob  <- predict(final.logit, h,type="response")

#clases.clut

c  <- data.frame(
income            <- as.factor(0),
age              <- median(data$age ),
man           <- as.factor(0),
children            <- 1,
clases.cult      <- 1,
univ             <- 0,
employed         <- 1
)

c_prob  <- predict(final.logit, c,type="response")


# no employed

ne  <- data.frame(
income            <- as.factor(0),
age             <- median(data$age ),
man           <- as.factor(0),
children            <- 1,
clases.cult      <- 0,
univ             <- 0,
employed         <- 0
)
cult.lessons
ne_prob  <- predict(final.logit, ne,type="response")


# Changes in probabilities:

# if you are poor
rp_prob  - benchmark_prob
# if you are rich
```

```
rr_prob    - benchmark_prob
# if you are a man
h_prob    - benchmark_prob
# with lessons
c_prob - benchmark_prob
# not employed
ne_prob - benchmark_prob
```