Study of Demographic Information of TV Show

I. Abstract

In this project, our objective is to obtain the demographic information of five TV shows based on their audience's consumption structure. Our analysis is based on two sources of data, collected from two different agencies. One is about the types of merchandise bought by the audience of each show, while the other is about the types of merchandise bought by people from 29 demographic categories, both calculated as percentage of the group. The limitations of these two data sets brought us several obstacles in the process of analysis. First, we only have information about the consumption structure, which is our response, and the two data sets are from different agencies. It is hard to connect the two sources of data based on the response. Also, the demographic information we have is only about marginal distribution of each category, but we want the joint distribution between subgroups. Besides, we find overlap in some categories, some of which are easy to correct, while we are unable to handle the rest. Considering all these situations, instead of fitting a model with the sparse information, we decided to estimate the conditional probability within each demographic subgroup conditioned on watching each TV show. And we interpreted the estimates we got, combined with some plots we made, to have a better understanding of what's going on. We also came up with a way of doing simulation with external data to resolve the problem of no joint distribution. Though we didn't regard this as our final method of this study, it is worth consideration if you want more detailed demographic information of each TV show.

II. Methods and Results

First, we did some data exploration on our data sets. We calculated the correlation between the row for each demographic category and the row for each show, and plotted them to have a brief understanding of how they are correlated.

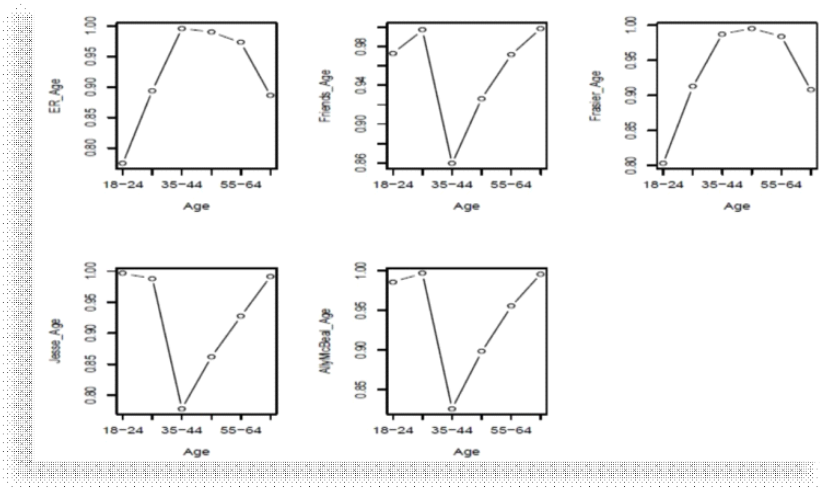

Figure 1: Plots for correlation between Age and five TV shows.

From Figure 1, we can see that almost all age and TV show combinations have a relationship larger than 0.8, which indicate they are highly correlated. However, some combinations still have relatively smaller correlation than others, like Age 35-44 with Friends.

Goal: To obtain the demographic information of five TV shows based on their audience's consumption structure.

Since we only have information about the marginal distribution of demographic categories in proportions, it is hard to fully answer the question above. We instead tried to find the conditional probability within each demographic subgroup conditioned on the people who watch one of the five TV shows and then combined the results with some plots to give a partial description of how their potential audiences look like.

To implement this method, we first did some data preparation.
First, we divided the 29 demographic variables into several subgroups, such as age, education, income, etc.
Then, to eliminate the overlap in each subgroup, we did some calculating. For example, we find the categories 40K-more and 30K-more have some overlap.

SubgroupDescription	SubgroupTotal	Clothing
40K--more	80078	31.56
30K--more	106838	30.62

Then the percentage of Clothing for new category 30K-40K is
(106838*30.62-80078*31.56)/(106838-80078)=27.81

We also divided the employment group into two subgroups, employment by gender (Unemployed_Male, Unemployed_Female, Employed_Male, Employed_Female), and employment-only (Full_Time, Part_Time, Unemployed).

Next, to illustrate how we got the estimated conditional probability, we take employment-only subgroup for example.
Denote the marginal distribution of the three categories (full time, part time, and unemployed) are D1,D2,D3, respectively.
Each of the three categories have Ni people, and Wi is the weight assigned to each category, Wi=Ni/(N1+N2+N3)
Since there are 6 purchasing categories, Di is a vector of length 6, and so is ER.

	Clothing	Electronics	Home	Housewares	Sporting	Toys
D1	26.65	4.46	9.20	6.31	8.00	5.03
D2	33.51	2.98	10.74	7.78	5.47	11.51
D3	22.73	3.37	7.41	6.01	2.76	6.20

ER	31.65	4.22	10.65	7.22	4.97	9.58

W1	0.594	N1	110363
W3	0.059	N2	11047
W3	0.347	N3	64412

Suppose P1,P2,P3 are the estimated probability we want, and we construct the function as follows:
F=||W1*P1*D1+W2*P2*D2+W3*P3*D3 -(W1+W2+W3)*TV1 ||2
We want to minimize F where P1+P2+P3=1 and P1,P2,P3 are all non-negative. We could solve this problem by using the "optim" function in R.

By minimizing this kind of quadratic target function, we can calculate the estimate of the conditional distribution between every subgroup and every TV show.

Here is our results.

	ER	Friends	Frasier	Jesse	Ally_McBeal"
Age_18--24	0.1428	0.1429	0.1428	0.1427	0.1427
Age_25--34	0.2245	0.1429	0.2239	0.2287	0.2282
Age_35--44	0.2961	0.1429	0.2956	0.3029	0.3011
Age_45--54	0.1791	0.1429	0.1795	0.1827	0.1818
Age_55--64	0.1428	0.1429	0.1428	0.1427	0.1427
Age_65--over"	0.0147	0.2857	0.0155	3e-04	0.0035
College_Grad.	0.44	0.5	0.4984	0.5002	0.5006
Attend_College	0.25	0.2498	0.2497	0.2496	0.2497
High_School	0.25	0.2498	0.2497	0.2496	0.2497
No_High_Sch.	0.0601	5e-04	0.0022	5e-04	0
Unemployed_Male	0	0	0	0	0
Unemployed_Fem."	0	0	0	0	0
Employed_Male"	0	0	0	0	0
Employed_Fem.	1	1	1	1	1
Full_Time	0.6209	0.6639	0.6209	0.702	0.6886
Part_Time	0.3331	0.333	0.3331	0.2974	0.3104
Unemployed	0.046	0.0032	0.046	6e-04	0.001
Single	0.2498	0.2498	0.2498	0.2497	0.2497
Married	0.4938	0.4937	0.4939	0.479	0.4824
Div/Sep/Wid	0.2498	0.2498	0.2498	0.2497	0.2497
Parents	0.0066	0.0068	0.0064	0.0216	0.0181
75K--more	0.1745	0.1686	0.1743	0.1715	0.1743
60K--75K	0.1178	0.1162	0.1177	0.1171	0.1178
50K--60K	0.1292	0.1267	0.129	0.128	0.1291
40K--50K	0.143	0.1395	0.1428	0.1416	0.1428
30K--40K	0.1577	0.1533	0.1575	0.1558	0.1576

20K--30K	0.1465	0.1428	0.1466	0.1449	0.1463
10K--20K	0.1311	0.1285	0.1313	0.1298	0.1315
under_10K	2e-04	0.0246	7e-04	0.0113	7e-04

III. Conclusions and Improvements

To illustrate how to combine the estimate and graphs to do interpretations, we take the results of Age vs ER show as example and other relationships can be explained in the same way.

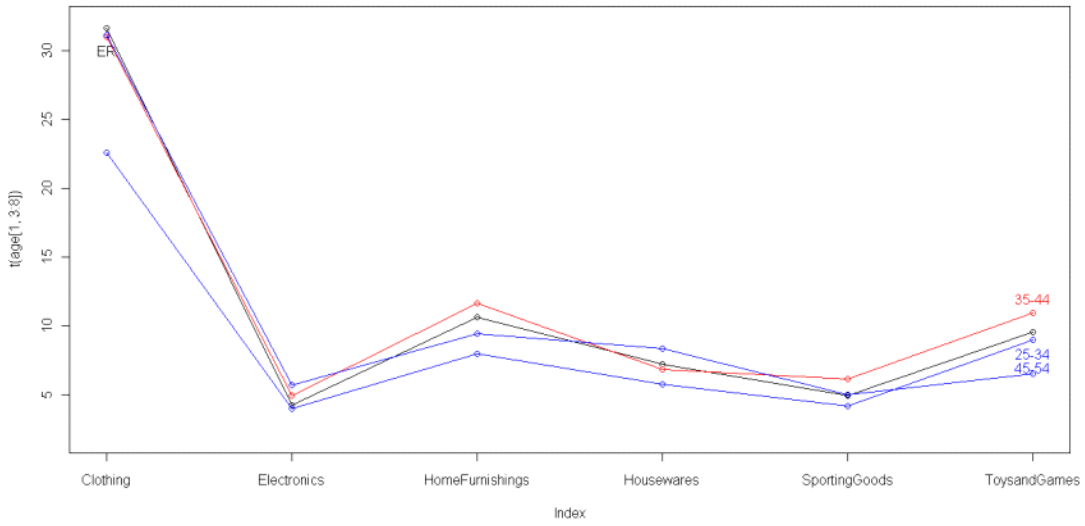

Figure 2: Age vs ER

In Figure 2, the black line is for ER show and the red line is for age 35-44, which has the largest estimated probability. As we can see, the two lines are very close to each other, and consumption pattern of people aged 35-44 is also very similar to that of the audience of ER show.

The two blue lines are for age 25-34 and age 45-54, respectively, which have the second and third largest estimated probability. It is mentioned that, though they have relatively large probability than other categories in this subgroup, and their probabilities are quite close, they have a little different consumption structure. For people aged 25-34, their shopping pattern is somewhat different from the audience of ER show. However, their consumption level is similar to the ER show audience, and in some types of merchandise, their discrepancies are opposite, so as to make up for each other and lead to a small overall gap. For people aged 45-54, the situation is just the opposite. People at this age have similar consumption structure with ER show audience, but their consumption level is a little below.

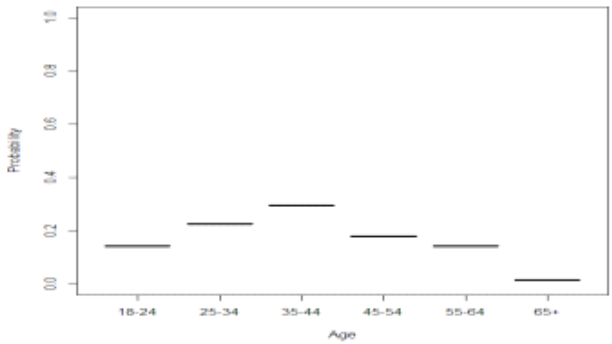

Figure 3: Estimated probability for Age vs ER.

Figure 3 provides a visual description of which category is more likely to watch ER show, and we also make plots like this for each category and TV show combination.

In this study, we only computed the conditional probability in each subgroup. Usually, we may often want to know the probability of falling in the combination of subgroups. In order to do this, we may need external data about the joint distribution of 29 categories. Though incomplete, the idea is like this. We could use external data set from American Community Survey, which contains variables for each demographic category, so joint distributions could be constructed from the individual-level records (Marginal distributions in ACS data do not exactly match our demographic data). We created data set with one record for each combination of demographic categories, then created a weight for each record. We gave initial 'guess' for each weight based on ACS distribution and applied to the Nielsen data. Now, we could assign purchase frequencies to each category as given in the data. After assigning probabilities across multiple categories, the marginal distributions were not maintained. As before, we performed iterative process to calibrate probabilities across each marginal distribution. Now, we had some population distribution and joint probability distribution based on observed data. Repeat this process several times to generate a point estimate and standard errors. However, this method took too much time to run and it is still needed to be improved. Thus we left it for further study.