Private Label Demand: Main Analysis

Günter J. Hitsch

January 19, 2017

```
library(bit64)
library(data.table)
library(psych)
library(lfe)
library(ggplot2)
library(stargazer)
```

Prerequisite: Please make sure you already calculated the household-level private label market shares (see Private-Label-Data-Preparation.Rmd/pdf)

Household data preparation

Before we begin with the main analysis we first examine the household data in the *panelists.RData* file more closely. Please consult the official Homescan Data documentation for an exhaustive discussion of the variable and measurement details.

```
load("./Data/panelists.RData")
names(panelists)
```

```
 [1] "household_code"                "panel_year"
 [3] "projection_factor"             "projection_factor_magnet"
 [5] "household_income"              "household_size"
 [7] "type_of_residence"             "household_composition"
 [9] "age_and_presence_of_children"  "female_head_age"
[11] "male_head_age"                 "male_head_employment"
[13] "female_head_employment"        "male_head_education"
[15] "female_head_education"         "male_head_occupation"
[17] "female_head_occupation"        "male_head_birth"
[19] "female_head_birth"             "marital_status"
[21] "race"                          "hispanic_origin"
[23] "panelist_zip_code"             "fips_state_code"
[25] "fips_state_descr"              "fips_county_code"
[27] "fips_county_descr"             "region_code"
[29] "scantrack_market_code"         "scantrack_market_descr"
[31] "dma_code"                      "dma_descr"
[33] "kitchen_appliances"            "tv_items"
[35] "household_internet_connection" "wic_indicator_current"
[37] "wic_indicator_ever_notcurrent" "Member_1_Birth"
[39] "Member_1_Relationship_Sex"     "Member_1_Employment"
[41] "Member_2_Birth"                "Member_2_Relationship_Sex"
[43] "Member_2_Employment"           "Member_3_Birth"
[45] "Member_3_Relationship_Sex"     "Member_3_Employment"
[47] "Member_4_Birth"                "Member_4_Relationship_Sex"
[49] "Member_4_Employment"           "Member_5_Birth"
[51] "Member_5_Relationship_Sex"     "Member_5_Employment"
```

```
[53] "Member_6_Birth"                 "Member_6_Relationship_Sex"
[55] "Member_6_Employment"            "Member_7_Birth"
[57] "Member_7_Relationship_Sex"      "Member_7_Employment"
```

The data are the `household_code`/`panel_year` level. Each observation also has an associated **`projection_factor`**. The purpose of the projection factor is to make the Nielsen *sample* representative of the whole U.S. population. The projection factor represents an estimate of the number of households in the population at large that a single household in the sample represents. The projection factors can then be used to *project* sample statistics from the Homescan sample to the whole population at the national or regional level. You may use the projection factors below when calculating aggregate market shares, although the projection factors are not needed for the main regression analysis.

Household income is one of the most important variables in the analysis. The variable is represented as a *factor* with different *levels* representing an income range:

```
is.factor(panelists$household_income)
```

```
[1] TRUE
```

```
levels(panelists$household_income)
```

```
 [1] "-$5000"              "$5000-$7999"         "$8000-$9999"        
 [4] "$10,000-$11,999"     "$12,000-$14,999"     "$15,000-$19,999"    
 [7] "$20,000-$24,999"     "$25,000-$29,999"     "$30,000-$34,999"    
[10] "$35,000-$39,999"     "$40,000-$44,999"     "$45,000-$49,999"    
[13] "$50,000-$59,999"     "$60,000-$69,999"     "$70,000-$99,999"    
[16] "$100,000 + "         "$100,000 - $124,999" "$125,000 - $149,999"
[19] "$150,000 - $199,999" "$200,000 + "        
```

Our results will be easier to interpret if we convert the factor representation of income to a dollar measure. The conversion is somewhat tedious but straightforward:

```
panelists[household_income == "-$5000",              income := 2500]
panelists[household_income == "$5000-$7999",         income := 6500]
panelists[household_income == "$8000-$9999",         income := 9000]
panelists[household_income == "$10,000-$11,999",     income := 11000]
panelists[household_income == "$12,000-$14,999",     income := 13500]
panelists[household_income == "$15,000-$19,999",     income := 17500]
panelists[household_income == "$20,000-$24,999",     income := 22500]
panelists[household_income == "$25,000-$29,999",     income := 27500]
panelists[household_income == "$30,000-$34,999",     income := 32500]
panelists[household_income == "$35,000-$39,999",     income := 37500]
panelists[household_income == "$40,000-$44,999",     income := 42500]
panelists[household_income == "$45,000-$49,999",     income := 47500]
panelists[household_income == "$50,000-$59,999",     income := 55000]
panelists[household_income == "$60,000-$69,999",     income := 65000]
panelists[household_income == "$70,000-$99,999",     income := 85000]
panelists[household_income == "$100,000 - $124,999", income := 112500]
panelists[household_income == "$125,000 - $149,999", income := 132500]
panelists[household_income == "$150,000 - $199,999", income := 175000]
panelists[household_income == "$200,000 + ",         income := 250000]
```

An alternative, slightly less tedious but less literal approach is shown in the code chunk below. This is *for reference only*—if you run into a situation where you have to convert many factor variables the code will hopefully be helpful.

```
# Numeric values to replace the current factor levels
income_levels = c(2500, 6500, 9000, 11000, 13500, 17500, 22500, 27500, 32500, 37500, 42500,
                  47500, 55000, 65000, 85000, 100000, 112500, 137500, 175000, 200000)

# Create desired numeric variable based on existing factor, then replace the
# current factors with the new levels
panelists[, income := household_income]
levels(panelists$income) = income_levels

# Now convert to numeric values
panelists[, income := as.numeric(levels(income))[income]]

# Confirm that the new variable matches up with the original factors
panelists[, head(.SD, 1), keyby = household_income, .SDcols = "income"]
```

The detailed income levels above $100,000 were only recorded for the 2006-2009 panel years (see the Homescan documentation). For consistency in how income is measured across all years, replace these values with 112,500 dollars:

```
panelists[income >= 100000, income := 112500]
```

Furthermore, the Homescan data documentation warns us that household income represents "ranges of total household income for the full year that is 2 years prior to the Panel Year." This is due to the survey methodology that Nielsen uses to obtain the annual income data. A few months prior to the current `panel_year` Nielsen obtains updated income information from each household. In the survey, Nielsen asks about total annual income during the prior year. Hence, the 2011 `panel_year` data contain information in the 2009 household income.

To **correctly date the income data** we therefore need to associate the income reported two years in the future with the current panel year. This can be achieved using the `shift` operator in data.table, as explained in the data.table *Additional Topics* notes. In particular, first make sure that the `panelists` table is correctly keyed along household code and panel year, and then replace income for year `y` with the lead of income two years ahead.

```
key(panelists)
```

```
[1] "household_code" "panel_year"
```

```
panelists[, income := shift(income, n = 2, type = "lead"), by = household_code]
```

Some of the key demographic variables, in particular, age or birth year, employment, and education, are available both for a *male head* and a *female head* in each household. However, depending on the household composition, only one head may be present, and more generally we may want to simplify this information to make our results more easily interpretable.

A simple solution is to use information on employment, etc., for the male head by default, and use information for the female head if no male head is present in a household. The choice of the male head reflects that the incidence of the employment status `Not Employed for Pay` is higher for the female head than for the male head in the data. Confirm this!

Now code new `age`, `unemployed`, and `education` variables:

```
panelists[, female_head := male_head_age == "No Such Head"]

panelists[, age := male_head_birth]
```

```
panelists[female_head == TRUE,
            age := female_head_birth]
panelists[, age := panel_year - as.numeric(substr(age, 1, 4))]

panelists[, unemployed := male_head_employment == "Not Employed for Pay"]
panelists[female_head == TRUE,
            unemployed := female_head_employment == "Not Employed for Pay"]

panelists[, education := male_head_education]
panelists[female_head == TRUE,
            education := female_head_education]
```

Note the coding of `age`. The birth year is a character string:

```
head(panelists$female_head_birth)
```

```
[1] "1928-08" "1928-08" "1928-08" "1928-08" "1928-08" "1928-08"
```

Therefore, we use the `substr` (sub-string) function to extract the first four characters, and then we converted them to a number using the `as.numeric` function.

In addition, create two more variables:

- `size`, a variable that provides a numeric measure of the `household_size` of a panelist
- `has_children`, an indicator (dummy) variable that equals 1 is children are present in the household

Merge household and Zillow home value index information with the share data

We need to merge the household data and Zillow home indices with the private label share data that we already calculated.

Before merging the household data, note that the year variable is called `year` in the share data. Hence, we need to rename the `panel_year` variable in the household data before we merge the tables. Also, we need to rename the `panelist_zip_code` in the household data to merge with the `zip_code` in the Zillow data.

I recommend to merge the following demographic variables with the share data and use them below in the regression analysis:

```
income, unemployed, education, age, size, has_children,
female_head, marital_status, race, hispanic_origin
```

Also, merge the ZIP code (`zip_code`), DMA code (`dma_code`), and the household projection factors (`projection_factor`) with the share data. A DMA (Designated Market Area) provides a local market definition, similar to an MSA (Metropolitan Statistical Area).

The Zillow data are constructed from publicly available information on the Zillow website. See Private-Label-Zillow-Data-Construction.Rmd. If you run the code in the R Markdown file the data will be directly read from the Zillow website using data.table's `fread` function. The data are then processed and saved in Zillow-Data.RData (this data file is already on Canvas, so you do not need to create it). The resulting data set, `zillow_DT`, contains the local, 5-digit ZIP code average home value, `zillow_index`, which we will use as the measure of housing wealth. The data are available at the `zip_code/year/month` level, and hence you will need to key the data appropriately to merge the shares data with the Zillow data.

Note that the Zillow data do not include home values for all ZIP codes. Hence, in the merge the option `all.x = TRUE` is needed to retain all the original share data.

After the merge, set the key for the share data to return them to the original order.

Finally, for better scaling, it also makes sense to represent the shares on a 0-100 percent scale.

Data description

Distribution of private label shares across households

First, provide an overview of the distribution of private label shares across households. To average away the randomness in the month-to-month shares, create annual private label shares for each household based on a simple average (mean).

Provide summary statistics and a histogram of the distribution of the private label shares across households.

Evolution of private label shares over time

To plot the evolution of private label shares over time, **provide a time-series graph of mean private label shares (across households) by month**. You will first need to calculate the average (across households) private label share in each month (year/month) in the data. If you want to provide a weighted mean with weights based on the household `projection_factor` you can use the `weighted.mean(x, w, ...)` function, that allows for weights provided using the `w` argument.

Providing a time-series plot of the resulting average private label shares is straightforward, but to be able to properly format the time variable on the x-axis it is easiest to supply the time variable as an R Date object. We already have year and month variables. Here is an easy way (if you know the function exists!) to combine `year` and `month` and create a Date variable:

```r
shares_month[, date := as.Date(ISOdate(year, month, 1))]
```

Inspect the `date` variable!

Generally, the third argument in `ISOdate` is a day variable. Here, because the data are at the monthly level, we pick (arbitrarily) the first day of each month.

When plotting the mean private label share over time in ggplot, you will use a sequence of layers that has roughly the following format:

```r
ggplot(...) +
   annotate(...) +
   geom_line(...) +
   geom_point(...) +
   scale_x_date(...) + ...
```

The added symbols for each data point provided by `geom_point` are of course unnecessary. I like them because adding the symbols makes the plot easier to read, but this is large an aesthetic choice.

There are two layers in the plot structure above that require some explanations.

First, the `annotate` layer can be used to shade a rectangular area. For this application, try:

```r
annotate("rect", xmin = as.Date("2007-12-1"), xmax = as.Date("2009-6-1"),
         ymin = -Inf, ymax = Inf, fill = "lightblue1", alpha = 0.4)
```

This layer will draw a shaded rectangle according to the x and y coordinates provided. The lower x-value corresponds to December 2007 and the upper x-value corresponds to June 2009, the official NBER (National Bureau of Economic Research) start and end dates of the Great Recession in the U.S.

Second, the `scale_x_date` layer allows us to format date values:

```r
scale_x_date("Year", date_labels = "%Y", date_breaks = "1 years", minor_breaks = NULL)
```

This specific layer will format the date values (which are at the day level) according to the corresponding year (`%Y`), and plot the labels in one year intervals. You can find more on how to format date variables here: https://www.r-bloggers.com/date-formats-in-r/

Now provide a time-series graph of the evolution of private label shares between 2004 and 2014.

Change in home values

To document the change in home values between June 2006 and June 2009 (the last month of the Great Recession), calculate the percentage change in the `zillow_index` at the ZIP code level.

Hint: The easiest way to do this is to use the `shift` operator. First, calculate the percentage change in the Zillow home value for a year/month combination relative to the period 36 months ago.

Then provide a histogram of the percentage home value changes for the percentage changes calculated for June 2009.

Main analysis

We now estimate income and wealth effects on private label demand using regression analysis.

Model specifications

Estimate the following models:

1. **Base model**: Use the main variables, `log(income)`, the `unemployed` indicator, and the wealth measure, `log(zillow_index)`, as covariates in the regression. The dependent variable is the household-level private label share in each month. In this base model we *pool* across all units in the panel.
2. **Demographics**: Add all the demographic controls. This is an attempt to control for heterogeneity across the units using observed information.
3. **Household fixed effects**: Don't use the demographics, but use household-level fixed effects instead. This provides within-estimates of the main variables of interest.
4. **Time controls**: Add different time controls,
   (a) Add a trend variable
   (b) Trend plus recession indicator (dummy)
   (c) Year/month fixed effects
5. Use the most flexible or preferred model and cluster standard errors at the market/year level. More on cluster standard errors below.

Notes on the estimation

Recall how to calculate a time trend variable from the `year` and `month` variables.

```
shares_DT[, month_index := 12*(year - min(year)) + month]
```

Cluster standard errors are used when it is likely that the error terms in the regression are *correlated within specific groups* in the data. If the error terms are correlated at the group-level we likely *understate* the true standard errors of the estimates. Cluster standard errors alleviate this problem and predict the correct statistical uncertainty about the true parameter value. In this application, we hypothesize that the error terms are correlated at the local market level over a medium time horizon such as a year. The `dma_code` is an indicator for local markets, and the data contains the `year`.

To provide cluster standard errors using `felm` use the following extension of the model formula:

```
fit = felm(y ~ x | fixed_effects_vars | 0 | cluster_var_1 + cluster_var_2 + ..., data = ...)
```

In our application use `dma_code + year` as cluster variables. Make sure to put "`| 0 |`" between the fixed effect variable(s) and the variables that define the clusters!

Save memory

To save memory keep only the observations that will be used in the regression analysis,

```
shares_DT = shares_DT[complete.cases(shares_DT)]
```

Estimation

Create a recession dummy that takes the value `1` between December 2007 and June 2009, and `0` for all other months.

Estimate the sequence of models discussed above, and use the stargazer package to combine and inspect the results.

Discussion of the results

Discuss and compare the results across the models.

Focus on the following key issues:

(a) As you add additional controls, some of the estimates (or standard errors) change. Can you provide an explanation for the largest observed change(s)?

(b) Which of the estimates have a credible *causal interpretation*? Which controls do you think are necessary before you are convinced that the estimates indeed represent causal effects?

(c) What are the implied economic magnitudes of the income, wealth, and employment effects? Consider the effect on the private label share of:

- A 50 percent reduction in income
- A 50 percent reduction in housing wealth
- Becoming unemployed