reducedPMEprototypeAnalysis

Pedro Girardi

December 11, 2018

PME

```
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------------------------- tidyverse 1.2
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.2.1     v forcats 0.3.0
```

```
## -- Conflicts -------------------------------------------------------------------- tidyverse_conflicts
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)
library(readxl)
library(dicionariosIBGE)
```

```
setwd("~/Comps/preliminaryPMEdata/")
reduced_pme = read.csv("pme_reduced.csv")[,-1]
reduced_pme_original = reduced_pme
colnames(reduced_pme_original) = sub('.*\\.', '', colnames(reduced_pme_original))
```

```
# PME relabeling (depends on what variables I imported)
names(reduced_pme) = c("Gender", "Month", "Household", "MetropolitanArea", "Income")
# find a way to also relabel the categories themselves
```

```
# manipulation of income variable
mean(is.na(reduced_pme$Income))
```

```
## [1] 0.6499104
```

```
mean(reduced_pme$Income==0, na.rm=T)
```

```
## [1] 0.02045483
```

```
#there are people with income of 0. I will add $1 to all incomes so we don't have issues with log
# only getting the observations with income
percentiles_income = quantile(reduced_pme$Income, probs = 1:100/100, na.rm=T)# some percentiles are off
reduced_pme = reduced_pme %>% filter(Income<1e6)
reduced_pme = reduced_pme[complete.cases(reduced_pme),]
reduced_pme$logIncome = log(reduced_pme$Income+1)
reduced_pme$MetropolitanArea = as.factor(reduced_pme$MetropolitanArea)
# summary(reduced_pme)
# hist(reduced_pme$logIncome) # I should probably remove the 0s and they will be overly influential
reduced_pme = reduced_pme %>% filter(Income>0)
```

```
dictionaryPME = read_xls("~/Comps/PME/documentacao/Layout/dicionario.xls", skip = 7, col_names = F)
names(dictionaryPME) = c("inicio", "tamanho", "cod", "desc", "valor", "rotulo")
# fixing NAs
```

```
for (i in 1:(length(dictionaryPME$cod)-1)){
  if (is.na(dictionaryPME$cod[i+1])) {
    dictionaryPME$cod[i+1] = dictionaryPME$cod[i]
  }
}
dictionaryPME = dictionaryPME %>%
  group_by(cod) %>%
  mutate(inicio = max(inicio, na.rm = T))

# relabeling and re-organizing so I can use the dicionariosIBGE package
rotPME = dictionaryPME[,c(3,5,6)]
dictionaryPME = dictionaryPME[,c(1,3,2,4)]
dictionaryPME = dictionaryPME[complete.cases(dictionaryPME),]

# relabeling factor variables
rotPME = rotPME[!is.na(rotPME$valor),] #taking NAs out
rotPME = rotPME[rotPME$valor>=0&rotPME$valor<=99,] # keeping only factor variables
originalColNames = str_to_upper(colnames(reduced_pme_original))
rotPME = rotPME %>% filter(cod %in% originalColNames)
# for (i in 1:length(reduced_pme_original[,-5])){
#   currentRotPME = rotPME %>% filter(cod==originalColNames[i])
#   currentPMEindicator = reduced_pme_original[,i]
#   for (j in 1:length(currentPMEindicator)){
#     for (k in 1:nrow(currentRotPME)){
#       if (currentPMEindicator[j]==currentRotPME$cod[k]) {
#         currentPMEindicator[j] = currentRotPME$rotulo[k]
#       }
#     }
#   }
# }

# relabeling current PME dataset (colnames only)
for (j in 1:length(originalColNames)){
  for (i in 1:nrow(dictionaryPME)){
    if (originalColNames[j] == dictionaryPME$cod[i]) {
      originalColNames[j] = dictionaryPME$desc[i]
    }
  }
}
originalColNames = str_replace_all(originalColNames, " ", "")
```

Inflation

```
inflation = read_csv("inflationCleaned.csv")[,-1] %>%
  filter(Region != "Brasil") %>%
  mutate(Region = str_sub(string = Region,
                          end = -6)) #removing Brazil data and abbreviation of cities
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   Region = col_character(),
```

```
##   Month = col_double(),
##   Inflation = col_double(),
##   year = col_double()
## )
```

```
inflation = inflation %>%
  mutate(Region = as.factor(Region))

inflation = inflation %>%
  group_by(Region) %>%
  mutate(PriceIndex = cumprod(1+Inflation/100)) #creates CPI


# matching inflation with cases at PME
rotPMEstates = rotPME%>% filter(cod=="V035", valor%in%levels(reduced_pme$MetropolitanArea))

reduced_pme$MetropolitanArea = plyr::mapvalues(reduced_pme$MetropolitanArea,
                                               from = levels(reduced_pme$MetropolitanArea),
                                               to = levels(as.factor(rotPMEstates$rotulo)))

# joining the datasets
complete.df = left_join(reduced_pme, inflation, by=c("MetropolitanArea" = "Region", "Month" = "Month"))
```

```
## Warning: Column `MetropolitanArea`/`Region` joining factors with different
## levels, coercing to character vector
```

Adding Growth

```
growthData = read_csv(file = "growthData.csv", col_types = c("cd"))
colnames(growthData)=c("Month", "GDP_Nominal")
growthData = growthData %>% mutate(Year = as.integer(str_sub(Month, end = 4)),
                                   Month = as.integer(str_sub(Month, start = 6)))
# need to get the REAL GDP, not Nominal
inflationBR = read_csv("inflationCleaned.csv")[,-1] %>%
  filter(Region == "Brasil")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   Region = col_character(),
##   Month = col_double(),
##   Inflation = col_double(),
##   year = col_double()
## )
```

```
inflationBR = inflationBR %>%
  mutate(Region = as.factor(Region))
inflationBR = inflationBR %>%
  mutate(PriceIndex = cumprod(1+Inflation/100)) #creates CPI

growthData = growthData %>% left_join(inflationBR, by = c("Month", "Year" = "year"))
growthData = growthData[complete.cases(growthData),]
growthData = growthData %>% mutate(RealGDP = GDP_Nominal/PriceIndex)
```

```
cleanGrowthData = growthData %>% select(Month, Year, RealGDP) %>% mutate(ChangeInGDP = RealGDP-lag(RealG
# merging
complete.df = complete.df %>% left_join(cleanGrowthData)
```

```
## Joining, by = "Month"
```

Because I am looking at the power of inflation on real income, I NEED TO account for the positive impact of inflation on NOMINAL terms. So I should adjust income such that if there was no change in real terms,

$$Y_{t+1} = Y_t \cdot (\pi_t + 1)$$

. If not, hyperinflation would automatically trigger a false positive relation between the variables. This will be particularly relevant when looking at the final longitudinal analysis.

```
complete.df = complete.df %>% mutate(RealIncome = Income/PriceIndex,
                                      logRealIncome = log(RealIncome))
```

```
# selecting only people in the bottom quintile of the income distribution at any given period
# keep in mind this WONT be the final method (NOT TRACKING OVER TIME, REPEATED OBS OF INDIVIDUALS)

complete.df = complete.df %>% group_by(Month) %>% mutate(Poor = ifelse(RealIncome <= quantile(RealIncome
poor.df = complete.df %>% filter(Poor==T)
```

Running the analysis

```
lm1 = lm(data=poor.df, logRealIncome ~ Gender + Month +
           MetropolitanArea + Inflation + ChangeInGDP)
lm2 = update(lm1, . ~ . + Gender*Inflation + Inflation*MetropolitanArea)

summary(lm1)
```

```
##
## Call:
## lm(formula = logRealIncome ~ Gender + Month + MetropolitanArea +
##     Inflation + ChangeInGDP, data = poor.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1307 -0.0652  0.1687  0.2316  0.3154
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    6.138e+00  7.862e-03 780.712  < 2e-16 ***
## Gender                        -7.807e-02  3.209e-03 -24.326  < 2e-16 ***
## Month                          1.089e-03  6.413e-04   1.698  0.08956 .
## MetropolitanAreaPorto Alegre  -1.531e-02  4.787e-03  -3.199  0.00138 **
## MetropolitanAreaRecife        -5.503e-02  5.030e-03 -10.941  < 2e-16 ***
## MetropolitanAreaRio de Janeiro 2.871e-02  5.262e-03   5.456 4.89e-08 ***
## MetropolitanAreaSalvador      -1.329e-02  5.468e-03  -2.430  0.01511 *
## MetropolitanAreaSão Paulo     -1.787e-02  5.777e-03  -3.093  0.00198 **
## Inflation                     -4.021e-02  4.431e-03  -9.075  < 2e-16 ***
## ChangeInGDP                    3.056e-07  1.719e-07   1.777  0.07550 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.4023 on 68623 degrees of freedom
##   (6660 observations deleted due to missingness)
## Multiple R-squared:  0.01326,    Adjusted R-squared:  0.01313
## F-statistic: 102.5 on 9 and 68623 DF,  p-value: < 2.2e-16
```

```
summary(lm2)
```

```
##
## Call:
## lm(formula = logRealIncome ~ Gender + Month + MetropolitanArea +
##     Inflation + ChangeInGDP + Gender:Inflation + MetropolitanArea:Inflation,
##     data = poor.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1416 -0.0657  0.1680  0.2315  0.4058
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                              6.118e+00  9.919e-03 616.816
## Gender                                  -7.465e-02  4.462e-03 -16.730
## Month                                    3.065e-03  6.551e-04   4.679
## MetropolitanAreaPorto Alegre            -3.290e-02  7.425e-03  -4.430
## MetropolitanAreaRecife                  -2.729e-02  6.674e-03  -4.089
## MetropolitanAreaRio de Janeiro          -8.836e-03  8.195e-03  -1.078
## MetropolitanAreaSalvador                -4.094e-02  8.153e-03  -5.021
## MetropolitanAreaSão Paulo               -3.484e-02  9.015e-03  -3.864
## Inflation                               -2.221e-02  1.626e-02  -1.366
## ChangeInGDP                              4.391e-07  1.731e-07   2.538
## Gender:Inflation                        -9.014e-03  8.471e-03  -1.064
## MetropolitanAreaPorto Alegre:Inflation   5.042e-02  1.412e-02   3.571
## MetropolitanAreaRecife:Inflation        -1.573e-01  1.284e-02 -12.247
## MetropolitanAreaRio de Janeiro:Inflation 9.471e-02  1.489e-02   6.360
## MetropolitanAreaSalvador:Inflation       6.168e-02  1.332e-02   4.632
## MetropolitanAreaSão Paulo:Inflation      4.682e-02  1.754e-02   2.668
##                                         Pr(>|t|)
## (Intercept)                              < 2e-16 ***
## Gender                                   < 2e-16 ***
## Month                                   2.89e-06 ***
## MetropolitanAreaPorto Alegre            9.43e-06 ***
## MetropolitanAreaRecife                  4.34e-05 ***
## MetropolitanAreaRio de Janeiro          0.280957
## MetropolitanAreaSalvador                5.14e-07 ***
## MetropolitanAreaSão Paulo               0.000111 ***
## Inflation                               0.171926
## ChangeInGDP                             0.011163 *
## Gender:Inflation                        0.287275
## MetropolitanAreaPorto Alegre:Inflation  0.000355 ***
## MetropolitanAreaRecife:Inflation         < 2e-16 ***
## MetropolitanAreaRio de Janeiro:Inflation 2.03e-10 ***
## MetropolitanAreaSalvador:Inflation      3.62e-06 ***
## MetropolitanAreaSão Paulo:Inflation     0.007624 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.401 on 68617 degrees of freedom
##   (6660 observations deleted due to missingness)
## Multiple R-squared:  0.01981,    Adjusted R-squared:  0.0196
## F-statistic: 92.46 on 15 and 68617 DF,  p-value: < 2.2e-16
```

```
anova(lm1, lm2, test="Chisq")
```

```
## Analysis of Variance Table
##
## Model 1: logRealIncome ~ Gender + Month + MetropolitanArea + Inflation +
##     ChangeInGDP
## Model 2: logRealIncome ~ Gender + Month + MetropolitanArea + Inflation +
##     ChangeInGDP + Gender:Inflation + MetropolitanArea:Inflation
##   Res.Df   RSS Df Sum of Sq  Pr(>Chi)
## 1  68623 11106
## 2  68617 11032  6    73.702 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```