Coleta dos Arquivos de Texto com os Microdados

Os dados podem ser adquiridos gratuitamente no domínio do IBGE, assim como sua documentação, descrição dos dados e dicionários.

Link: <http://www.ibge.gov.br/home/estatistica/indicadores/trabalhoerendimento/pme_nova/defaultmicro.shtm>

Leitura dos Microdados

Para a leitura dos Microdados no Software R de forma mais simples, se mostrou necessária a instalação do pacote “dicionariosIBGE”, de tal forma que os códigos e número de casas ocupadas pelos caracteres já está definida e apenas se mostra necessária a aplicação dos mesmos.

|  |
| --- |
| **Código R** |
| install.packages("dicionariosIBGE")  library(dicionariosIBGE)  data(dicPME)  attach(dicPME)  DataPME = read.fwf(“DataPME2012.1.txt”, widths = tamanho)  names(DataPME) = cod |

Tamanho e cod são o nome das colunas com o número de caracteres das variáveis e o código das variáveis, respectivamente.

|  |  |
| --- | --- |
| **Comando** | **O que faz** |
| Install.packages(“”) | Instala o pacote referido |
| data() | Cria o Data.Frame com o dicionário de microdados da PME |
| attach() | Torna possível a utilização do nome da variável |
| Read.fwf(txt, widths = c()) | Lê o arquivo txt de acordo com as especificações de caracteres (widths) da variável tamanho, do dicionário de microdados da PME |
| Names() = cod | Renomeia as colunas de variáveis |

Identificação do Domicílio

Para a identificação única de cada domicílio foram reaproveitadas as informações contidas em “Sobre o Painel da Pesquisa Mensal de Emprego (PME) do IBGE”, de Rafael Perez Ribas e Sergei Suarez Dillon Soares.

|  |  |
| --- | --- |
| **Código da Variável** | **Nome da Variável** |
| V035 | RM |
| V040 | Número de Controle |
| V050 | Número de Série |
| V060 | Painel |
| V063 | Grupo Rotacional |
| **Código R** | |
| DataPME$Domicilio = paste(DataPME$V035, DataPME$V040, DataPME$V050,  DataPME$V060, DataPME$V063, sep = “”) | |

Logo, o código “domicílio” é gerado a partir de outras variáveis disponíveis da pesquisa e torna capaz identificar domicílios de forma única durante todos os seus meses de pesquisa **respondidos**.

Identificação do Indivíduo Respondente

Por sua vez, a identificação do indivíduo também se apresentou como uma variável derivada. A identificação de um indivíduo se dá a partir da variável derivada anterior (Domicílio), juntamente com as variáveis V201 (Número de ordem), V203 (Sexo), V204 (Dia de Nascimento), V214 (Mês de Nascimento) e V224 (Ano do Nascimento).

Dado a possível variação das demais variáveis determinísticas (Como V201 – Número de ordem), essas variáveis se apresentaram como potenciais ID’s.

|  |
| --- |
| **Código R** |
| DataPME$IDpessoa = paste(DataPME$V201, DataPME$V203, DataPME$V204, DataPME$V214, DataPME$V224, sep = “”)  DataPME$ID = paste(DataPME$Domicilio, DataPME$IDpessoa, sep = “”) |

As mesmas variáveis utilizadas em IDpessoa, utilizada em um domicílio durante um único mês, é base para a determinação de gêmeos.

Redução dos Erros

Alguns erros foram encontrados, tais como duplicação de respostas, absurdo de respostas (tais como filhos com a idade dos pais) e falhas na identificação de usuários pela ausência de respostas.

Para reduzir essa possibilidade de erro e assim identificar usuários com determinada certeza, as ausências de respostas nas variáveis de nascimento foram excluídas.

|  |
| --- |
| **Código R** |
| DataPME = subset(DataPME, V204 != 99 & V214 != 99 & V224 != 9999) |

Proporção esperada de respostas

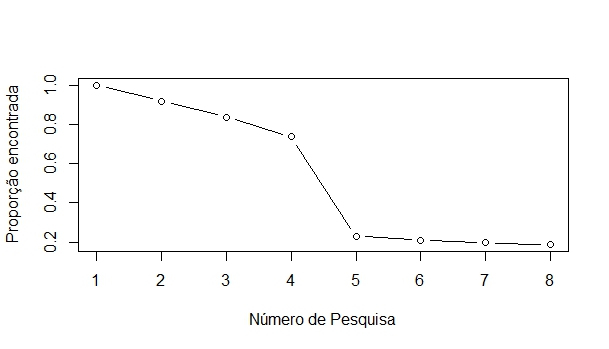
Para conhecer a proporção de respostas encontradas, foram utilizadas respostas do mês de janeiro de 2012 até o mês de abril de 2013, em um total de 16 meses.

Em primeira instancia, foi analisado a proporção dos respondentes que entraram no mês de janeiro de 2012 e foram encontrados no mês de fevereiro de 2012, depois entraram em janeiro e foram encontrados até março de 2012 e assim por diante. Os resultados encontrados foram os seguintes:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Espaço das respostas** | **Número de meses respondidos** | **Número de pessoas encontrada** | **Número de pessoas esperada** | **Proporção Encontrada** |
| Janeiro – Fevereiro 2012 | 2 | 1139 | 1241 | 0,9178 |
| Janeiro – Março 2012 | 3 | 1039 | 1241 | 0,8372 |
| Janeiro – Abril  2012 | 4 | 913 | 1241 | 0,7357 |
| Janeiro – Janeiro  2013 | 5 | 286 | 1241 | 0,2305 |
| Janeiro – Fevereiro  2013 | 6 | 260 | 1241 | 0,2595 |
| Janeiro – Março  2013 | 7 | 244 | 1241 | 0,1966 |
| Janeiro – Abril  2013 | 8 | 232 | 1241 | 0,1869 |

O número de pessoas esperada, representa a quantidade de pessoas que entraram no mês de janeiro de 2012 e estavam respondendo pela primeira vez a pesquisa. Por sua vez, o número de pessoas encontrada, são os respondentes do mês de janeiro que também responderam a todos os meses subsequentes delimitados pelo espaço das respostas.

Graficamente, a proporção encontrada foi decrescente e com um impacto maior na queda após o intervalo de 8 meses e o retorno às pesquisas da Pesquisa Mensal de Emprego.



Os valores foram encontrados a partir da seleção especifica de indivíduos a partir do mês de pesquisa e do número de pesquisa do domicílio.

Ex.: O número esperado de respondentes se deu pela seleção específica dos indivíduos que entraram no mês de janeiro de 2012 e estavam respondendo ao primeiro mês de pesquisa do domicílio. A proporção esperada se deu a partir da seleção específica dos respondentes no mês de fevereiro que estavam respondendo ao segundo mês de pesquisa MAIS o primeiro mês de pesquisa, e então foram obtidos os resultados duplicados. Logo, os respondentes que participaram dos dois primeiros meses de pesquisa, e assim por diante.

|  |
| --- |
| **Código R** |
| DataPME$cod = paste(DataPME$V070, DataPME$V072, DataPME$V075, sep = "")  DataPME.usual = subset(DataPME, V204 != 99 & V214 != 99 & V224 != 9999)  Data = rbind(DataPME.usual[which(DataPME.usual$cod == "112012"),],  DataPME.usual[which(DataPME.usual$cod == "222012"),],  DataPME.usual[which(DataPME.usual$cod == "332012"),],  DataPME.usual[which(DataPME.usual$cod == "442012"),],  DataPME.usual[which(DataPME.usual$cod == "152013"),],  DataPME.usual[which(DataPME.usual$cod == "262013"),],  DataPME.usual[which(DataPME.usual$cod == "372013"),],  DataPME.usual[which(DataPME.usual$cod == "482013"),])  DataCod = names(c(table(Data$ID))[c(table(Data$ID)) == "nº de resp desejado"]) |

Logo, o seguinte algoritmo determina o número de respondentes que é possível encontrar a partir do ID utilizado. Tal que ele sempre encontra as respostas em sequência (que respondeu todos até determinada pesquisa no domicilio).

Obtendo o comportamento de respostas a partir do ID utilizado

Uma variável booleana foi utilizada para determinar resposta e não resposta do questionário nos meses indicados. Tal que 1 representa que foi respondido naquele mês de pesquisa e 0 que não foi respondido.

A partir disso temos todas as possíveis respostas de todos os ID’s reconhecidos na pesquisa. Assim como suas frequências.

|  |
| --- |
| **Código R** |
| ID = c(Data$ID)  ID = unique(ID)  y = matrix(0, ncol = 9, nrow = length(ID))  for(i in 1:length(ID)){  x = Data[which(Data$ID == ID[i]),]  aux = sort(x$V072)  aux = c(aux, rep(99,(8-length(aux))))  for(j in 1:8){  if(aux[j] != 99){  y[i,aux[j]] = 1  y[i,9] = ID[i]  }  }  }  y = as.data.frame(y)  y$respostas = paste0("(", y$V1,",", y$V2,",", y$V3,",", y$V4,",", y$V5,",", y$V6,",", y$V7,",", y$V8,")")  freq(y$respostas) |