Stats approach

Tom Dalton

08/09/2017

This documents details the approach taken to verify the genalogical populations that we create.

For efficency we have produced five contingency tables which are each concerned with one of the input distrbtions efffecting genalogical structure.

This first dataset contains 9,550,521 individuals.

Lets load these in:

```
path = paste("/Users/tsd4/OneDrive/cs/PhD/code/population-model/validated/src/main",
                "/resources/results/scot-rfs-size/20170922-000418:080/tables/", sep = "")

data.death = read.csv(paste(path, "death-CT.csv", sep = ""), sep = ',', header = T)
data.obirth = read.csv(paste(path, "ob-CT.csv", sep = ""), sep = ',', header = T)
data.mbirth = read.csv(paste(path, "mb-CT.csv", sep = ""), sep = ',', header = T)
data.partner = read.csv(paste(path, "part-CT.csv", sep = ""), sep = ',', header = T)
data.sep = read.csv(paste(path, "sep-CT.csv", sep = ""), sep = ',', header = T)
```

Column abbriviations:

- NPCIAP - Number of previous children in any partnership
- CIY - Children in year (Yes/No)
- NCIY - Number of children in year
- NPA - New partners age
- NCIP - Number of children in partnership

These tables are as follows (data will be cleaned later):

```
head(data.death, 2)
```

```
##   Source    Sex Age Died Date freq
## 1    SIM FEMALE 281   NO 1947    1
## 2    SIM FEMALE 281   NO 1948    1
```

```
head(data.obirth, 2)
```

```
##   Source   Age NPCIAP CIY Date freq
## 1    SIM 40-49      2  NO 1988   70
## 2    SIM 40-49      2  NO 1989   69
```

```
head(data.mbirth, 2)
```

```
##   Source   Age NCIY Date freq
## 1   STAT 30-34    1 1810 3001
## 2    SIM 30-34    1 1810    1
```

```
head(data.partner, 2)
```

```
##   Source   Age   NPA Date freq
## 1   STAT 15-19 45-49 1838    1
## 2    SIM 15-19 45-49 1838    1
```

```
head(data.sep, 2)
```

```
##   Source NCIP Separated Date freq
## 1    SIM    3        NO 1975 1856
## 2    SIM    3        NO 1974 1885
```

Death Analysis

```
# Standardise the data
data.death$freq <- round(data.death$freq)
data.death <- data.death[which(data.death$freq != 0), ]
data.death <- data.death[which(data.death$Date >= 1855) , ]
data.death <- data.death[which(data.death$Date < 2014) , ]

summary(data.death)
```

```
##   Source          Sex               Age             Died
##  SIM :102182   FEMALE:104279   Min.   :  0.00   NO :102163
##  STAT: 66015   MALE  : 63918   1st Qu.: 33.00   YES: 66034
##                                Median : 66.00
##                                Mean   : 88.86
##                                3rd Qu.: 99.00
##                                Max.   :413.00
##       Date          freq
##  Min.   :1855   Min.   :   1
##  1st Qu.:1897   1st Qu.:   2
##  Median :1937   Median :  73
##  Mean   :1936   Mean   :1990
##  3rd Qu.:1976   3rd Qu.:4930
##  Max.   :2013   Max.   :9815
```

```
# Analysis
library("MASS")
model = loglm(freq ~ Date + Sex + Age + Died + Sex:Age + Sex:Died + Age:Died
              + Sex:Age:Died, data = data.death)
model
```

```
## Call:
## loglm(formula = freq ~ Date + Sex + Age + Died + Sex:Age + Sex:Died +
##     Age:Died + Sex:Age:Died, data = data.death)
##
## Statistics:
##                        X^2     df P(> X^2)
## Likelihood Ratio 153043.9 166383        1
## Pearson          152800.9 166383        1
```

Here we see the model created is a good fit for the data and thus that the Source (whether an individual is from the statistics or the simulation) of an indidual has no meaningful effect on the frequency. This is what we want to see.

Ordered Birth

```
largestBirthLabel = "50+"
```

```r
# Standardise the data
data.obirth$freq <- round(data.obirth$freq)
data.obirth <- data.obirth[which(data.obirth$freq != 0), ]
data.obirth <- data.obirth[which(data.obirth$Date >= 1855) , ]
data.obirth <- data.obirth[which(data.obirth$Date < 2014) , ]
data.obirth <- data.obirth[which(data.obirth$Age != "0to14"), ]
data.obirth <- data.obirth[which(data.obirth$Age != largestBirthLabel), ]
#data.obirth <- data.obirth[which(data.obirth$CIY == "YES"), ]


# Analysis
library("MASS")
model = loglm(freq ~ Age + NPCIAP + CIY + Date + Age:NPCIAP + Age:CIY + NPCIAP:CIY + Age:NPCIAP:CIY, da
model
```

```
## Call:
## loglm(formula = freq ~ Age + NPCIAP + CIY + Date + Age:NPCIAP +
##     Age:CIY + NPCIAP:CIY + Age:NPCIAP:CIY, data = data.obirth)
##
## Statistics:
##                       X^2    df P(> X^2)
## Likelihood Ratio 75852.15 17888        0
## Pearson          74189.21 17888        0
```

Multiple Birth

```r
data.mbirth$freq <- round(data.mbirth$freq)
data.mbirth <- data.mbirth[which(data.mbirth$freq != 0), ]
data.mbirth <- data.mbirth[which(data.mbirth$Date >= 1855) , ]
data.mbirth <- data.mbirth[which(data.mbirth$Date < 2014) , ]
data.mbirth <- data.mbirth[which(data.mbirth$Age != "0to14"), ]
data.mbirth <- data.mbirth[which(data.mbirth$Age != largestBirthLabel), ]
data.mbirth <- data.mbirth[which(data.mbirth$NCIY != "0"), ]

# Analysis
library("MASS")
model = loglm(freq ~ Date + NCIY + Age + Date:NCIY + Date:Age, data = data.mbirth)
model
```

```
## Call:
## loglm(formula = freq ~ Date + NCIY + Age + Date:NCIY + Date:Age,
##     data = data.mbirth)
##
## Statistics:
##                       X^2   df P(> X^2)
## Likelihood Ratio 26631.71 3199        0
## Pearson          61268.72 3199        0
```

Partnering

```r
# Standardise the data
data.partner$freq <- round(data.partner$freq)
data.partner <- data.partner[which(data.partner$freq != 0), ]
data.partner <- data.partner[which(data.partner$Date >= 1855) , ]
data.partner <- data.partner[which(data.partner$Date < 2014) , ]
data.partner <- data.partner[which(data.partner$NPA != "na") , ]

# Analysis
library("MASS")

model = loglm(freq ~ Date + NPA + Age + NPA:Age, data = data.partner)
model
```

```
## Call:
## loglm(formula = freq ~ Date + NPA + Age + NPA:Age, data = data.partner)
##
## Statistics:
##                        X^2    df P(> X^2)
## Likelihood Ratio 497451.5 16471        0
## Pearson          454563.4 16471        0
```

Separation

```r
# Standardise the data
data.sep$freq <- round(data.sep$freq)
data.sep <- data.sep[which(data.sep$freq != 0), ]
data.sep <- data.sep[which(data.sep$Date >= 1855) , ]
data.sep <- data.sep[which(data.sep$Date < 2014) , ]
data.sep <- data.sep[which(data.sep$Separated != "NA") , ]

# Analysis
library("MASS")
model = loglm(freq ~ Date + NCIP + Separated + NCIP:Separated, data = data.sep)
model
```

```
## Call:
## loglm(formula = freq ~ Date + NCIP + Separated + NCIP:Separated,
##     data = data.sep)
##
## Statistics:
##                        X^2   df P(> X^2)
## Likelihood Ratio 818.0541 2396        1
## Pearson          747.2367 2396        1
```