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 1,048,194 individuals.

Lets load these in:

```
path = paste("/Users/tsd4/OneDrive/cs/PhD/code/population-model/validated/src/main",
                "/resources/results/review/20170908-120243:448/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   STAT FEMALE 109   NO 1800 2.896647e-02
## 2   STAT FEMALE  52   NO 1995 1.067993e+03
```

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

```
##   Source    Age NPCIAP CIY Date     freq
## 1   STAT 20to24     0+ YES 1760 710.3867
## 2   STAT 20to24     0+ YES 1755 710.4510
```

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

```
##   Source    Age NCIY Date     freq
## 1   STAT 15to49    0 1757 41045.67
## 2   STAT 15to49    0 1756 41065.84
```

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

```
##   Source    Age    NPA Date     freq
## 1   STAT 25to29 25to29 1869 254.6565
## 2   STAT 25to29 25to29 1868 254.7067
```

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

```
##   Source NCIP Separated Date freq
## 1    SIM    3        NO 1975  398
## 2    SIM    3        NO 1974  369
```

Death Analysis

```r
# 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           Date
##  SIM :59788   FEMALE:59297   Min.   :  0.00   NO :64255   Min.   :1855
##  STAT:59047   MALE  :59538   1st Qu.: 30.00   YES:54580   1st Qu.:1894
##                              Median : 53.00               Median :1934
##                              Mean   : 52.89               Mean   :1934
##                              3rd Qu.: 77.00               3rd Qu.:1974
##                              Max.   :159.00               Max.   :2013
##       freq
##  Min.   :   1.0
##  1st Qu.:   7.0
##  Median :  42.0
##  Mean   : 470.7
##  3rd Qu.:1145.0
##  Max.   :1245.0
```

```r
# 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 25457.79 118037        1
## Pearson          25492.74 118037        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

```r
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 2638.919 3626        1
## Pearson          2638.759 3626        1
```

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 0.9747400 -159        1
## Pearson          0.9747394 -159        1
```

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 114714.8 11830        0
## Pearson          103206.6 11830        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 10203.375 2127        0
## Pearson           9597.207 2127        0
```