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 64,504 individuals.

Lets load these in:

```
path = paste("/Users/tsd4/OneDrive/cs/PhD/code/population-model/validated/src/main",
               "/resources/results/pre-test/20170922-073944:253/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  52   NO 1995   35
## 2   STAT FEMALE  52   NO 1996   35
```

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

```
##   Source   Age NPCIAP CIY Date freq
## 1    SIM 40-49     4+ YES 1931    1
## 2    SIM 40-49     4+ YES 1935    1
```

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

```
##   Source   Age NCIY Date freq
## 1   STAT 30-34    1 1810   20
## 2    SIM 30-34    1 1810    1
```

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

```
##   Source   Age   NPA Date freq
## 1   STAT 15-19 20-24 1930    1
## 2   STAT 15-19 20-24 1931    1
```

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

```
##   Source NCIP Separated Date freq
## 1    SIM    3        NO 1975   12
## 2    SIM    3        NO 1974   17
```

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            Date
##  SIM :44655   FEMALE:43657   Min.   :  0.00   NO :60803   Min.   :1855
##  STAT:41906   MALE  :42904   1st Qu.: 33.00   YES:25758   1st Qu.:1893
##                              Median : 60.00               Median :1932
##                              Mean   : 54.34               Mean   :1933
##                              3rd Qu.: 77.00               3rd Qu.:1972
##                              Max.   :103.00               Max.   :2013
##       freq
##  Min.   : 1.00
##  1st Qu.: 2.00
##  Median :33.00
##  Mean   :26.93
##  3rd Qu.:46.00
##  Max.   :73.00
```

```
# 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 10087.42 85987        1
## Pearson          10199.34 85987        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 3646.358 14898        1
## Pearson          3715.288 14898        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 179.3453 302 1.0000000
## Pearson          200.0426 302 0.9999988
```

Partnering

```
# 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 2921.769 5289        1
## Pearson          2865.926 5289        1
```

Separation

```
# 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 245.2521 1454        1
## Pearson          250.3835 1454        1
```