CHIC601 Group Project

greyhypotheses

30/11/2021

Exploratory Data Analysis

After inspecting/preparing Social Contact Survey data, in-line with Dr. Read's data preparation suggestions, the outline is

```
'data.frame':	4217 obs. of  14 variables:
 $ id               : int  1 2 3 4 5 6 7 8 ...
 $ postal           : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 ...
 $ unmatched_postcode: Factor w/ 3 levels "no","yes","unspecified": 1 1 1 1 1 1 1 1 ...
 $ web              : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 ...
 $ age              : int  51 62 36 27 35 61 41 73 ...
 $ date             : Date, format: "2009-05-28" "2009-05-28" ...
 $ day_of_week      : Ord.factor w/ 7 levels "Monday"<"Tuesday"<..: 5 5 5 5 5 5 5 5 ...
 $ postcode         : chr  "KT11 2JF" NA "GU34 2BG" ...
 $ sex              : Factor w/ 3 levels "female","male",..: 1 2 2 1 1 1 1 1 ...
 $ household_size   : Factor w/ 7 levels "2","1","3","4",..: 3 5 3 2 4 2 3 2 ...
 $ occupation       : Factor w/ 17 levels "retired","office",..: 3 10 4 2 7 1 9 1 ...
 $ total_contacts   : int  106 20 7 13 44 30 16 1 ...
 $ method           : Factor w/ 2 levels "postal","online": 1 1 1 1 1 1 1 1 ...
 $ agegroup         : Factor w/ 19 levels "[60,65)","[55,60)",..: 7 1 6 8 6 1 4 10 ...
```

The Variables

Age Groups & Survey Method

Disaggregation of responses by survey method

Table 1: The number of survey reponses per survey method

method	frequency
postal	3091
online	1126
NA	0

The age groups, based on the *age* field. [ref. SurveyData() function]

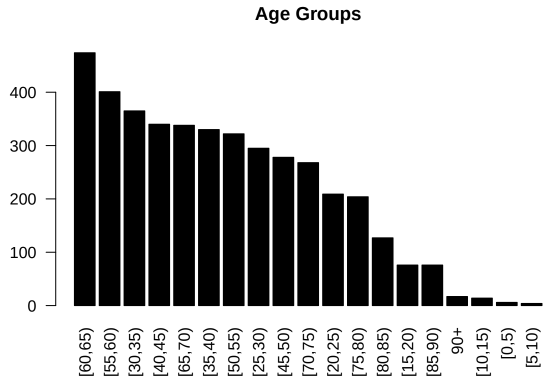

Figure 1: The distribution of age groups

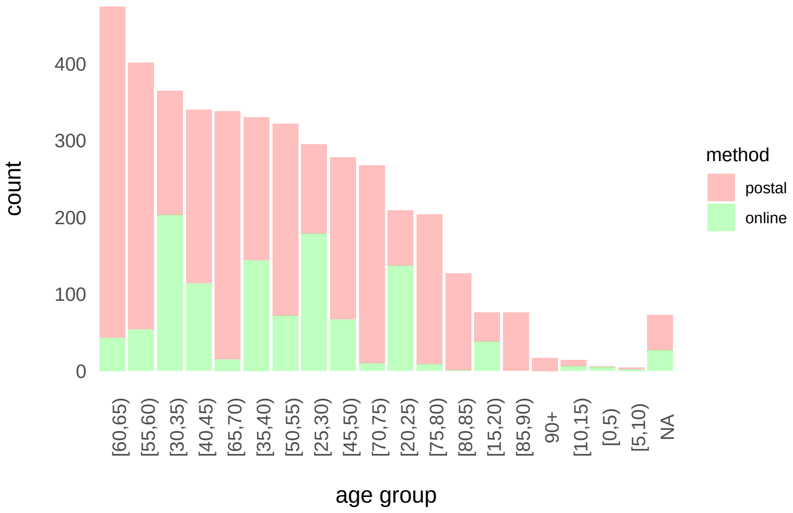

Figure 2: Age group and survey method

Day of Week & Survey Method

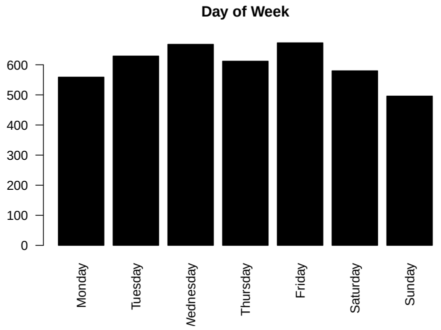

Figure 3: The distribution of responses by day of week

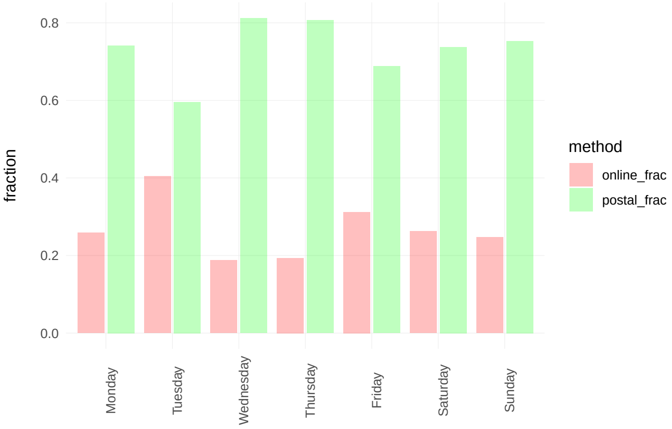

Figure 4: The distribution of responses by day of week and survey method

Occupations

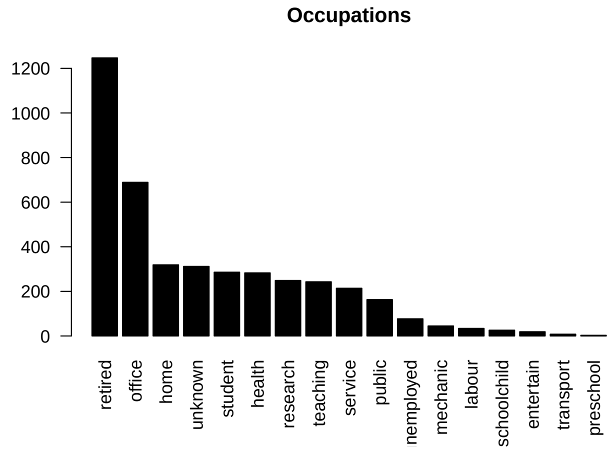

Figure 5: The distribution of occupations

Age Group & Sex

Table 2: The sex distribution

sex	frequency
female	2757
male	1393
unknown	67
NA	0

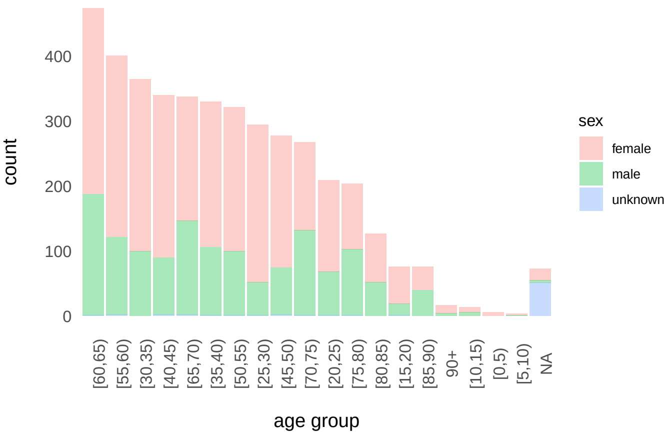

Figure 6: Age group and sex

Tests

```
##
## Attaching package: ’data.table’

## The following objects are masked from ’package:dplyr’:
##
##     between, first, last

##
## Attaching package: ’MASS’

## The following object is masked _by_ ’.GlobalEnv’:
##
##     survey

## The following object is masked from ’package:dplyr’:
##
##     select
```

Data

Modelling

```
##
## Call:
## glm.nb(formula = total_contacts ~ agegroup + ru11ind + occupation +
##     household_size, data = focus, init.theta = 0.8644334923,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2629  -1.0236  -0.5870   0.0081   8.4011
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           2.99989    0.07211  41.601  < 2e-16 ***
## agegroup[55,60)      -0.02435    0.08445  -0.288 0.773074
## agegroup[40,45)       0.10277    0.09741   1.055 0.291414
## agegroup[65,70)       0.17119    0.08429   2.031 0.042253 *
## agegroup[35,40)       0.14759    0.09938   1.485 0.137525
## agegroup[50,55)       0.28525    0.09484   3.008 0.002634 **
## agegroup[45,50)       0.06672    0.10282   0.649 0.516413
## agegroup[70,75)       0.10124    0.09161   1.105 0.269139
## agegroup[75,80)      -0.08665    0.10185  -0.851 0.394893
## agegroup[80,85)      -0.28407    0.11808  -2.406 0.016137 *
## agegroup[85,90)      -0.35534    0.14785  -2.403 0.016242 *
## agegroup90+          -1.19989    0.32740  -3.665 0.000247 ***
## ru11indB1             0.88958    0.13976   6.365 1.95e-10 ***
## ru11indC1            -0.10326    0.05192  -1.989 0.046717 *
## ru11indC2            -0.96428    0.35737  -2.698 0.006969 **
## ru11indD1            -0.01294    0.07483  -0.173 0.862737
## ru11indD2             0.13432    0.27843   0.482 0.629510
## ru11indE1            -0.11165    0.08524  -1.310 0.190240
## ru11indE2             0.57734    0.24872   2.321 0.020273 *
## ru11indF1            -0.32751    0.10841  -3.021 0.002519 **
## ru11indF2             0.13323    0.33258   0.401 0.688724
## occupationoffice     -0.05073    0.08468  -0.599 0.549116
## occupationhome        0.21281    0.09313   2.285 0.022311 *
## occupationunknown     0.07929    0.09941   0.798 0.425136
## occupationstudent    -0.49255    0.26006  -1.894 0.058226 .
## occupationhealth      0.54761    0.10157   5.391 6.99e-08 ***
## occupationresearch    0.33193    0.11710   2.835 0.004588 **
## occupationteaching    1.13247    0.10697  10.587  < 2e-16 ***
## occupationservice     0.68278    0.10921   6.252 4.06e-10 ***
## occupationpublic      0.42990    0.11939   3.601 0.000317 ***
## occupationunemployed -0.87481    0.18604  -4.702 2.57e-06 ***
## occupationmechanic    0.52528    0.20674   2.541 0.011060 *
## occupationlabour      0.42715    0.23503   1.817 0.069144 .
## occupationentertain   0.66849    0.29762   2.246 0.024699 *
## occupationtransport   1.55466    0.44595   3.486 0.000490 ***
## household_size1      -0.01783    0.05255  -0.339 0.734378
## household_size3      -0.25249    0.07102  -3.555 0.000377 ***
## household_size4       0.20903    0.08064   2.592 0.009536 **
## household_size5       0.34111    0.12175   2.802 0.005082 **
## household_size6+      0.32157    0.18923   1.699 0.089255 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
## (Dispersion parameter for Negative Binomial(0.8644) family taken to be 1)
##
##     Null deviance: 3750.1  on 2781  degrees of freedom
## Residual deviance: 3148.0  on 2742  degrees of freedom
## AIC: 23483
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.8644
##           Std. Err.:  0.0213
##
##  2 x log-likelihood:  -23400.9280
```

And

```
equatiomatic::extract_eq(model = modelglmnbstep, wrap = TRUE)
```

$$
\begin{aligned}
\log (E(\operatorname{total_{c}ontacts})) &= \alpha + \beta_{1}(\operatorname{agegroup}_{\operatorname{[55,60)}}) + \beta_{2}(\operatorname{agegroup}_{\operatorname{[40,45)}}) + \beta_{3}(\operatorname{agegroup}_{\operatorname{[65,70)}})\ + \\
&\quad \beta_{4}(\operatorname{agegroup}_{\operatorname{[35,40)}}) + \beta_{5}(\operatorname{agegroup}_{\operatorname{[50,55)}}) + \beta_{6}(\operatorname{agegroup}_{\operatorname{[45,50)}}) + \beta_{7}(\operatorname{agegroup}_{\operatorname{[70,75)}})\ + \\
&\quad \beta_{8}(\operatorname{agegroup}_{\operatorname{[75,80)}}) + \beta_{9}(\operatorname{agegroup}_{\operatorname{[80,85)}}) + \beta_{10}(\operatorname{agegroup}_{\operatorname{[85,90)}}) + \beta_{11}(\operatorname{agegroup}_{\operatorname{90+}})\ + \\
&\quad \beta_{12}(\operatorname{ru11ind}_{\operatorname{B1}}) + \beta_{13}(\operatorname{ru11ind}_{\operatorname{C1}}) + \beta_{14}(\operatorname{ru11ind}_{\operatorname{C2}}) + \beta_{15}(\operatorname{ru11ind}_{\operatorname{D1}})\ + \\
&\quad \beta_{16}(\operatorname{ru11ind}_{\operatorname{D2}}) + \beta_{17}(\operatorname{ru11ind}_{\operatorname{E1}}) + \beta_{18}(\operatorname{ru11ind}_{\operatorname{E2}}) + \beta_{19}(\operatorname{ru11ind}_{\operatorname{F1}})\ + \\
&\quad \beta_{20}(\operatorname{ru11ind}_{\operatorname{F2}}) + \beta_{21}(\operatorname{occupation}_{\operatorname{office}}) + \beta_{22}(\operatorname{occupation}_{\operatorname{home}}) + \beta_{23}(\operatorname{occupation}_{\operatorname{unknown}})\ + \\
&\quad \beta_{24}(\operatorname{occupation}_{\operatorname{student}}) + \beta_{25}(\operatorname{occupation}_{\operatorname{health}}) + \beta_{26}(\operatorname{occupation}_{\operatorname{research}}) + \beta_{27}(\operatorname{occupation}_{\operatorname{teaching}})\ + \\
&\quad \beta_{28}(\operatorname{occupation}_{\operatorname{service}}) + \beta_{29}(\operatorname{occupation}_{\operatorname{public}}) + \beta_{30}(\operatorname{occupation}_{\operatorname{unemployed}}) + \beta_{31}(\operatorname{occupation}_{\operatorname{mechanic}})\ + \\
&\quad \beta_{32}(\operatorname{occupation}_{\operatorname{labour}}) + \beta_{33}(\operatorname{occupation}_{\operatorname{entertain}}) + \beta_{34}(\operatorname{occupation}_{\operatorname{transport}}) + \beta_{35}(\operatorname{household\_size}_{\operatorname{1}})\ + \\
&\quad \beta_{36}(\operatorname{household\_size}_{\operatorname{3}}) + \beta_{37}(\operatorname{household\_size}_{\operatorname{4}}) + \beta_{38}(\operatorname{household\_size}_{\operatorname{5}}) + \beta_{39}(\operatorname{household\_size}_{\operatorname{6+}})
\end{aligned}
\tag{1}
$$