CHIS Exercise

Swati Jani Joshi

July 9, 2016

CHIS

This is the solution to the CHIS exercise from Data Camp

Load all packages

```
library(haven)
library(ggplot2)
library(reshape2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggthemes)
library(car)
```

Importing SPSS data file to R

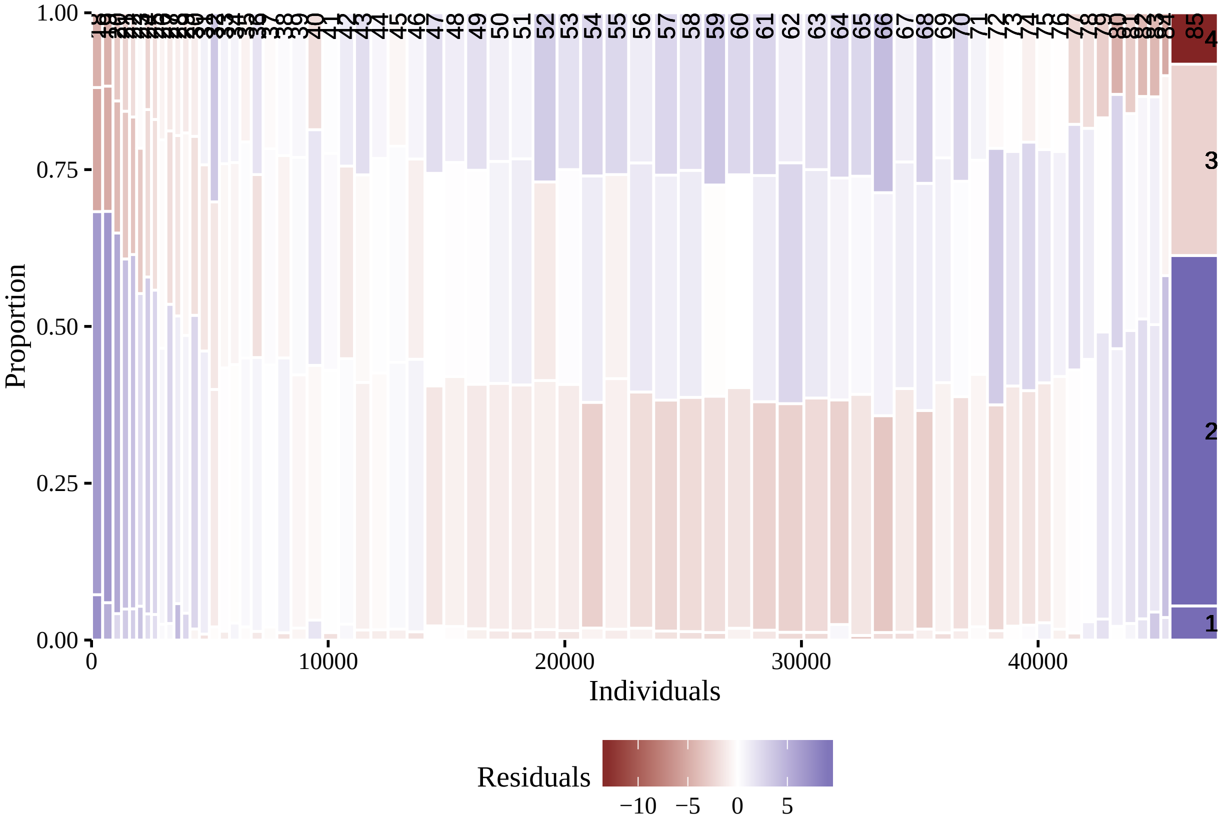
```
adult <- read_spss("~/Documents/Data Visualization/chis09_adult_spss/chis09_adult_spss/ADULT.sav")
```

Script generalized into a function

```
mosaicGG <- function(data, X, FILL) {
#### Proportions in raw data
  DF <- as.data.frame.matrix(table(data[[X]], data[[FILL]]))
  DF$groupSum <- rowSums(DF)
  DF$xmax <- cumsum(DF$groupSum)
  DF$xmin <- DF$xmax - DF$groupSum
  DF$X <- row.names(DF)
  DF$groupSum <- NULL
  DF_melted <- melt(DF, id = c("X", "xmin", "xmax"), variable.name = "FILL")
  DF_melted <- DF_melted %>%
    group_by(X) %>%
    mutate(ymax = cumsum(value/sum(value)),
           ymin = ymax - value/sum(value))

#### Chi-sq test
  results <- chisq.test(table(data[[FILL]], data[[X]])) # fill and then x
  resid <- melt(results$residuals)
  names(resid) <- c("FILL", "X", "residual")

#### Merge data
  DF_all <- merge(DF_melted, resid)

#### Positions for labels
  DF_all$xtext <- DF_all$xmin + (DF_all$xmax - DF_all$xmin)/2
  index <- DF_all$xmax == max(DF_all$xmax)
  DF_all$ytext <- DF_all$ymin[index] + (DF_all$ymax[index] - DF_all$ymin[index])/2

#### Plot:
  g <- ggplot(DF_all, aes(ymin = ymin,  ymax = ymax, xmin = xmin,
                          xmax = xmax, fill = residual)) +
    geom_rect(col = "white") +
    geom_text(aes(x = xtext, label = X),
              y = 1, size = 3, angle = 90, hjust = 1, show.legend = FALSE) +
    geom_text(aes(x = max(xmax),  y = ytext, label = FILL),
              size = 3, hjust = 1, show.legend = FALSE) +
    scale_fill_gradient2("Residuals") +
    scale_x_continuous("Individuals", expand = c(0,0)) +
    scale_y_continuous("Proportion", expand = c(0,0)) +
    theme_tufte() +
    theme(legend.position = "bottom")
  print(g)
}
```

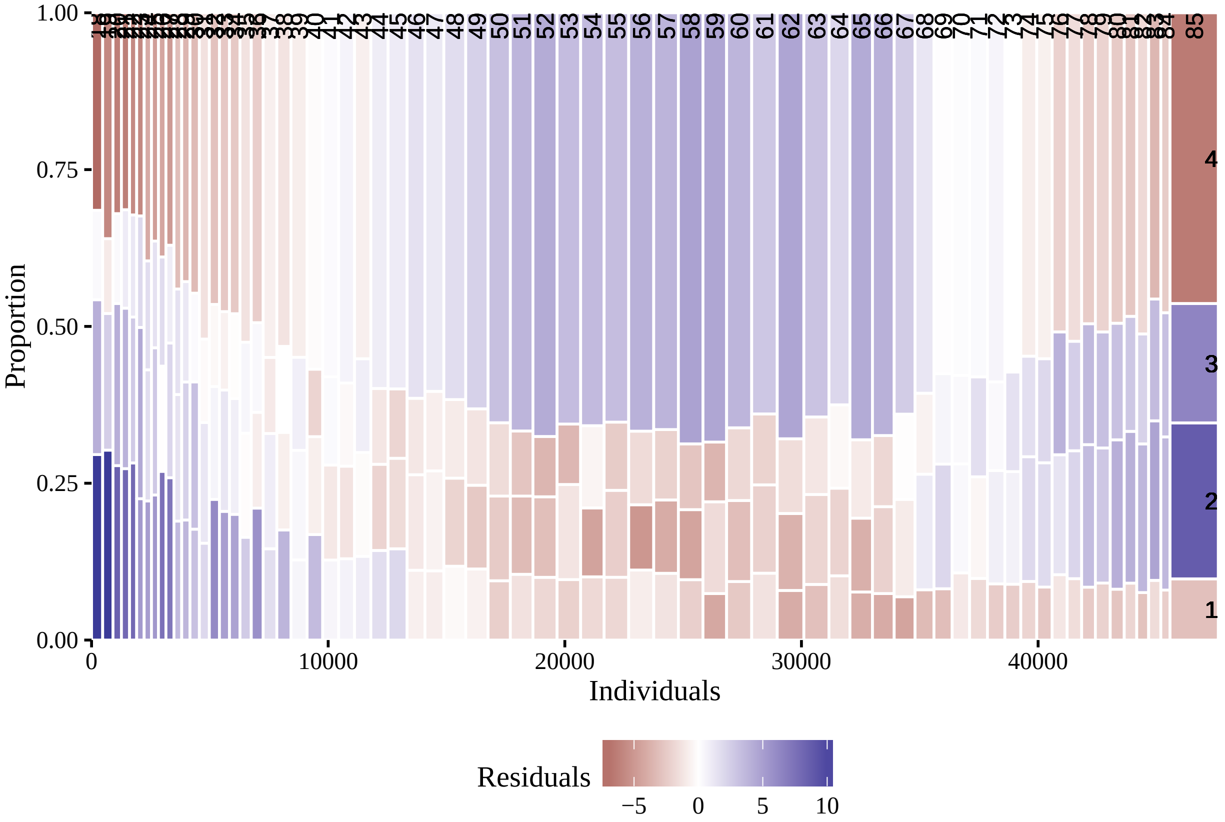
BMI described by age

```
mosaicGG(adult, "SRAGE_P", "RBMI")
```



Poverty described by age

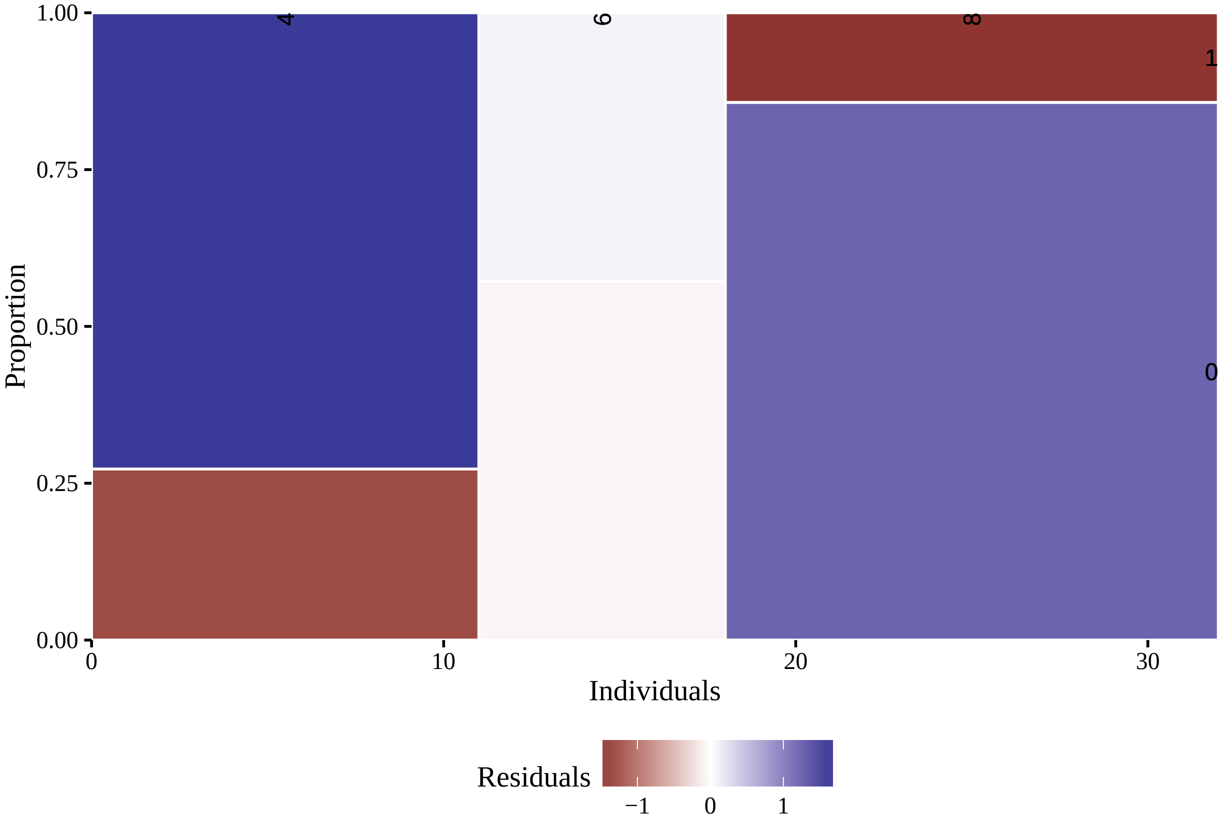
```
mosaicGG(adult, "SRAGE_P","POVLL")
```



mtcars: am described by cyl

```
mosaicGG(mtcars, "cyl", "am")
```

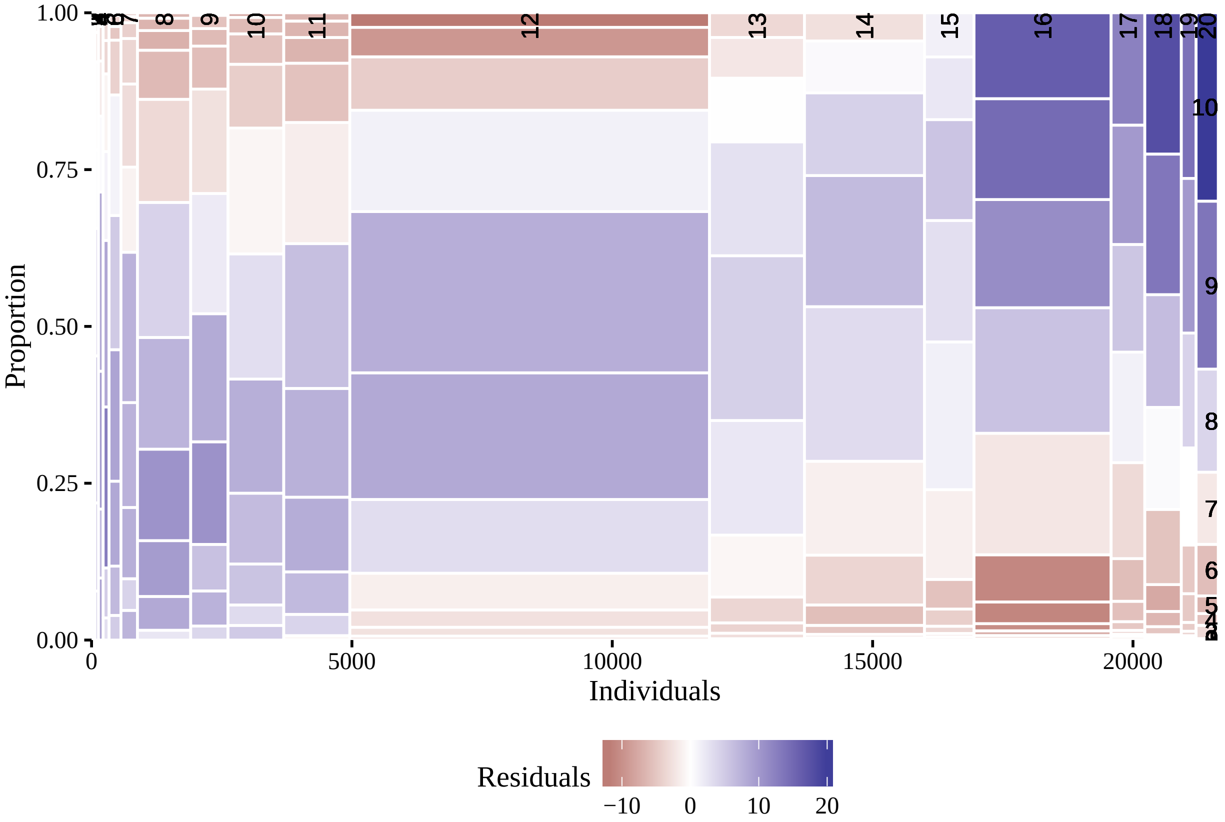
```
## Warning in chisq.test(table(data[[FILL]], data[[X]])): Chi-squared
## approximation may be incorrect
```



Vocab: vocabulary described by education

```
mosaicGG(Vocab, "education", "vocabulary" )
```

```
## Warning in chisq.test(table(data[[FILL]], data[[X]])): Chi-squared
## approximation may be incorrect
```


Proportion
1.00
0.75
0.50
0.25
0.00
0
5000
10000
15000
20000
Individuals
7
8
9
10
11
12
13
14
15
16
17
18
19
20
10
9
8
7
6
5
4
3
Residuals
−10
0
10
20