Kenyan Demographic and Health Survey 2003: Data Preparation and Visualization

Preparing data

The dataset *childrenfinal.dta* is obtained from Kenyan Demographic and Health Survey 2003 and contains various variables sampled in 2003 on the Kenyan children of age between 0 and 5 years. We want to "play" with this data. We will work with the dataset, where we will remove unnecassary columns and visualize some relationships between variables.

```r
library(tidyverse) # to be able to use data visual.tools
library(foreign) #reading data in diff.formats
library(raster) #to manipulate geographic data
library(viridis) #to make plots easier to read
library(ggrepel) #ggrepel provides geoms for ggplot2
```

```r
childrenfinal <- read.dta("childrenfinal.dta") # read the data
head(childrenfinal,2) #quick look at the dataset with 2 rows
```

```
##   hypage deathu5 v001 ruralfacto female tetanusmother breastfeeding wantedchild
## 1      6       0    1          1      0             1             6     no more
## 2     28       0    1          1      0            NA            24        then
##   anetalvisits  placedelivery caesarian birthweight m37a m37c m37f m37h m37l
## 1            5 govt. hospital        no        3500    8    8    8    8    8
## 2           NA govt. hospital        no        3600   NA   NA   NA   NA   NA
##   m37m m37n m37o m37p m37q m37r m37u m37v m37w m37x aidsinfo            vaccbcg
## 1    8    8    8    8    8    8    8    8    8    8      yes vacc. date on card
## 2   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA     <NA> reported by mother
##             vaccdpt1         vaccpolio1           vaccdpt2         vaccpolio2
## 1 vacc. date on card vacc. date on card vacc. date on card vacc. date on card
## 2 reported by mother reported by mother reported by mother reported by mother
##             vaccdpt3         vaccpolio3        vaccmeasles         vaccpolio0
## 1 vacc. date on card vacc. date on card                 no vacc. date on card
## 2 reported by mother                 no reported by mother reported by mother
##   diarrhea1 diarrhea2 diarrhea3 childage childweight childheight zstunt zweight
## 1        no      <NA>      <NA>        6         8.3          73   2.23    0.72
## 2        no      <NA>      <NA>       28        14.4          94   1.67    0.81
##   zwast  sdist s820 s821 s823 v824 married v505 v002 v003    v005 interviewdate
## 1 -1.32 mbeere   no   no   no   no married    0   20    2 1374352          1241
## 2  0.12 mbeere   no   no   no   no married    0   20    2 1374352          1241
##   v012    v024        v103 v104        v105      v106 v107               water
## 1   36 eastern countryside    1 countryside secondary    1 piped into dwelling
## 2   36 eastern countryside    1 countryside secondary    1 piped into dwelling
##   v115       toilet electricity radio television v122 bicycle motor car  floor
## 1   NA flush toilet          no   yes         no   no     yes    no  no cement
## 2   NA flush toilet          no   yes         no   no     yes    no  no cement
##   walls                     roof                       v130 ethnicity
## 1    NA corrugated iron (mabati) protestant/other christian      embu
## 2    NA corrugated iron (mabati) protestant/other christian      embu
##   yearsofedu        v134           v135 v136 numberchildrenbelow5        v141
## 1          8 countryside usual resident    6                    3 countryside
## 2          8 countryside usual resident    6                    3 countryside
##                   v149 relationtohead sexhh agehh awfactt awfactu awfactr
## 1 incomplete secondary           wife     0    40     100     100     100
## 2 incomplete secondary           wife     0    40     100     100     100
##   awfacte         twin birthdate deathdateexact deathdatemonths birthinterval
## 1     100 single birth      1235           <NA>              NA            22
## 2     100 single birth      1213           <NA>              NA            40
##   deadson deaddaughter agefirstbirth numberlivingchild knowledgecontraception
## 1       0            0            21                 7    knows modern method
## 2       0            0            21                 7    knows modern method
##     contraceptionuse           v367 v420 v421 v437 v438 v445 v446
## 1 used modern method wanted no more   NA   NA  741 1649 2725 1653
## 2 used modern method wanted no more   NA   NA  741 1649 2725 1653
##                                   v704          v716                  v739
## 1 other professional & related workers sales workers husband/partner alone
## 2 other professional & related workers sales workers husband/partner alone
##             v742            v743a                          v743b
## 1 less than half respondent alone respondent and husband/partner
## 2 less than half respondent alone respondent and husband/partner
##              v743c                          v743d            v743e v753
## 1 respondent alone respondent and husband/partner respondent alone  yes
## 2 respondent alone respondent and husband/partner respondent alone  yes
##       v754cp                                                       v754dp
## 1 don’t know reduce chance of aids: have 1 sex partnr with no oth partner
## 2 don’t know reduce chance of aids: have 1 sex partnr with no oth partner
##   v754jp v754wp v756 v774  wealth assetindex   motherid deathu1 deathu3 death
## 1     no     no  yes  yes richest    1.04446 1000200200       0       0     0
## 2     no     no  yes  yes richest    1.04446 1000200200       0       0     0
##   periodborn periodborn3 birthage birthorder childorder ff modhypage edumother
## 1          0           0       35          7          1  1         6         3
## 2          0           0       34          6          2  1        28         3
##   yearsofedu2 primary secondary birthorder2 childorder2 ruraljure birthage2
## 1          64       1         0          49           1         1      1225
## 2          64       1         0          36           4         1      1156
##   deadchildren dtwin dbreast   BMI motherunderweight severeunderweight Rohrer
## 1            0    NA       1 27.25                 0                 0  16.53
## 2            0    NA       1 27.25                 0                 0  16.53
##   ai_toiletqual waterquality1 waterquality2 contraknowledge numbvac vacindex
## 1             2             1             0               1       8        1
## 2             2             1             0               1       8        1
##   circumcision health1 health2 health3 cluster  adm2 identifier distance
## 1            0   12393     848     1.5       1 NITHI         34 330.2652
## 2            0   12393     848     1.5       1 NITHI         34 330.2652
##   ddistance1 ddistance2 hivclust hivnumb hivline hiv03 hiv05 hiv
## 1          0          0        1      20       2  <NA>     0  NA
## 2          0          0        1      20       2  <NA>     0  NA
```

There are 4686 observations on 177 variables, most of the variable names are self-explanatory.

Now we remove all variables that start with “s”,“v” and “m”. First of all we look how much are variables in the dataset, which have names starting with these characters:

```
# names() allows us to show the column names from our dataset
# substring() allows us to get first letters from each column name
s <- substring(names(childrenfinal),1,1) # show dataset columns with only 1st letters
s
```

```
##   [1] "h" "d" "v" "r" "f" "t" "b" "w" "a" "p" "c" "b" "m" "m" "m" "m" "m" "m"
##  [19] "m" "m" "m" "m" "m" "m" "m" "m" "m" "a" "v" "v" "v" "v" "v" "v" "v" "v"
##  [37] "v" "d" "d" "d" "c" "c" "c" "z" "z" "z" "s" "s" "s" "s" "v" "m" "v" "v"
##  [55] "v" "v" "i" "v" "v" "v" "v" "v" "v" "v" "w" "v" "t" "e" "r" "t" "v" "b"
##  [73] "m" "c" "f" "w" "r" "v" "e" "y" "v" "v" "v" "n" "v" "v" "r" "s" "a" "a"
##  [91] "a" "a" "a" "t" "b" "d" "d" "b" "d" "d" "a" "n" "k" "c" "v" "v" "v" "v"
## [109] "v" "v" "v" "v" "v" "v" "v" "v" "v" "v" "v" "v" "v" "v" "v" "v" "v" "v"
## [127] "v" "w" "a" "m" "d" "d" "d" "p" "p" "b" "b" "c" "f" "m" "e" "y" "p" "s"
## [145] "b" "c" "r" "b" "d" "d" "d" "B" "m" "s" "R" "a" "w" "w" "c" "n" "v" "c"
## [163] "h" "h" "h" "c" "a" "i" "d" "d" "d" "h" "h" "h" "h" "h" "h"
```

```
# define a list with letters, where columns from dataset must be dropped
dropped <- c("s","v","m")

#we modify our df without columns
#which names start with s,v and m
childrenfinal <- childrenfinal[, !(s %in% dropped)]
# The function above from the right side allows us
# To select us columns without s,v and m in the beginnig of names

head(childrenfinal)#quick look at the modifed df
```

```
##   hypage deathu5 ruralfacto female tetanusmother breastfeeding wantedchild
## 1      6       0          1      0             1             6     no more
## 2     28       0          1      0            NA            24        then
## 3     20       0          1      0             2            20        then
## 4     47       0          1      1            NA            24        then
## 5     14       0          1      1             3            14        then
## 6     15       0          0      0             2            15        then
##   anetalvisits       placedelivery caesarian birthweight aidsinfo
## 1            5      govt. hospital        no        3500      yes
## 2           NA      govt. hospital        no        3600     <NA>
## 3            4 private hosp/clinic       yes        2500      yes
## 4           NA    respondents home        no        2500     <NA>
## 5            4      govt. hospital        no        2900       no
## 6            4 govt. health center        no        2800      yes
##               diarrhea1         diarrhea2 diarrhea3 childage childweight
## 1                  no              <NA>      <NA>        6         8.3
## 2                  no              <NA>      <NA>       28        14.4
## 3                  no              <NA>      <NA>       20        12.4
## 4                  no              <NA>      <NA>       47        12.7
## 5 yes, last two weeks yes: no treatment        no       14         8.1
## 6                  no              <NA>      <NA>       15         9.5
##   childheight zstunt zweight zwast interviewdate                    water
## 1        73.0   2.23    0.72 -1.32          1241      piped into dwelling
## 2        94.0   1.67    0.81  0.12          1241      piped into dwelling
## 3        85.0   0.24    0.44  0.45          1241      piped into dwelling
## 4        91.0  -2.44   -1.82 -0.40          1241 piped into compound/plot
## 5        73.5  -0.84   -1.65 -1.38          1241 piped into compound/plot
## 6        76.9  -0.68   -1.16 -0.91          1242               public tap
##                   toilet electricity radio television bicycle car  floor walls
## 1           flush toilet          no   yes         no     yes  no cement    NA
## 2           flush toilet          no   yes         no     yes  no cement    NA
## 3 traditional pit toilet          no   yes        yes     yes yes cement    NA
## 4           flush toilet          no   yes        yes     yes  no cement    NA
## 5           flush toilet          no   yes        yes     yes  no cement    NA
## 6 traditional pit toilet          no   yes         no      no  no cement    NA
##                       roof    ethnicity yearsofedu numberchildrenbelow5
## 1 corrugated iron (mabati)         embu          8                    3
## 2 corrugated iron (mabati)         embu          8                    3
## 3 corrugated iron (mabati)         meru         15                    1
## 4 corrugated iron (mabati)         embu          8                    2
## 5 corrugated iron (mabati)         embu          8                    2
## 6 corrugated iron (mabati) taita/tavate          8                    1
##   relationtohead agehh awfactt awfactu awfactr awfacte         twin birthdate
## 1           wife    40     100     100     100     100 single birth      1235
## 2           wife    40     100     100     100     100 single birth      1213
## 3           wife    43     100     100     100     100 single birth      1221
## 4           wife    30     100     100     100     100 single birth      1194
## 5           wife    30     100     100     100     100 single birth      1227
## 6           wife    30     100     100     100     100 single birth      1227
##   deathdateexact deathdatemonths birthinterval deadson deaddaughter
## 1           <NA>              NA            22       0            0
## 2           <NA>              NA            40       0            0
## 3           <NA>              NA            24       0            0
## 4           <NA>              NA            NA       0            0
## 5           <NA>              NA            33       0            0
## 6           <NA>              NA            24       0            0
##   agefirstbirth numberlivingchild knowledgecontraception   contraceptionuse
## 1            21                 7    knows modern method used modern method
## 2            21                 7    knows modern method used modern method
## 3            24                 1    knows modern method used modern method
## 4            18                 2    knows modern method used modern method
## 5            18                 2    knows modern method used modern method
## 6            18                 1    knows modern method used modern method
##    wealth assetindex deathu1 deathu3 death periodborn periodborn3 birthage
## 1 richest    1.04446       0       0     0          0           0       35
## 2 richest    1.04446       0       0     0          0           0       34
## 3 richest    1.05364       0       0     0          0           0       24
## 4 richest    0.98059       0       0     0          0           1       18
## 5 richest    0.98059       0       0     0          0           0       20
## 6 richest    0.93924       0       0     0          0           0       18
##   birthorder childorder ff edumother yearsofedu2 primary birthorder2
## 1          7          1  1         3          64       1          49
## 2          6          2  1         3          64       1          36
## 3          1          1  1         5         225       1           1
## 4          1          2  1         2          64       1           1
## 5          2          1  1         2          64       1           4
## 6          1          1  1         2          64       1           1
##   childorder2 ruraljure birthage2 deadchildren dtwin dbreast   BMI Rohrer
## 1           1         1      1225            0    NA       1 27.25  16.53
## 2           4         1      1156            0    NA       1 27.25  16.53
## 3           1         1       576            0    NA       1 23.00  14.35
## 4           4         1       324            0    NA       1 28.01  18.58
## 5           1         1       400            0    NA       1 28.01  18.58
## 6           1         0       324            0    NA       1 21.14  13.41
##   ai_toiletqual waterquality1 waterquality2 contraknowledge numbvac
## 1             2             1             0               1       8
## 2             2             1             0               1       8
## 3             1             1             0               1       0
## 4             2             1             2               1       9
## 5             2             1             2               1       9
## 6             1             0             0               1       9
##   circumcision health1 health2  health3 cluster   adm2 identifier distance
## 1            0   12393     848 1.500000       1  NITHI         34 330.2652
## 2            0   12393     848 1.500000       1  NITHI         34 330.2652
## 3            0   12393     848 1.500000       1  NITHI         34 330.2652
## 4            0   12393     848 1.500000       1  NITHI         34 330.2652
## 5            0   12393     848 1.500000       1  NITHI         34 330.2652
## 6            0   14022     819 2.433333       2 KILIFI         14 655.0334
##   ddistance1 ddistance2 hivclust hivnumb hivline hiv03 hiv05 hiv
## 1          0          0        1      20       2  <NA>     0  NA
## 2          0          0        1      20       2  <NA>     0  NA
## 3          0          0        1      30       2  <NA>     0  NA
## 4          0          0       NA      NA      NA  <NA>    NA  NA
## 5          0          0       NA      NA      NA  <NA>    NA  NA
## 6          0          0       NA      NA      NA  <NA>    NA  NA
```

```
substring(names(childrenfinal),1, 1)#to be sure that we dropped necessary columns
```

```
##  [1] "h" "d" "r" "f" "t" "b" "w" "a" "p" "c" "b" "a" "d" "d" "d" "c" "c" "c" "z"
## [20] "z" "z" "i" "w" "t" "e" "r" "t" "b" "c" "f" "w" "r" "e" "y" "n" "r" "a" "a"
## [39] "a" "a" "a" "t" "b" "d" "d" "b" "d" "d" "a" "n" "k" "c" "w" "a" "d" "d" "d"
## [58] "p" "p" "b" "b" "c" "f" "e" "y" "p" "b" "c" "r" "b" "d" "d" "d" "B" "R" "a"
## [77] "w" "w" "c" "n" "c" "h" "h" "h" "c" "a" "i" "d" "d" "d" "h" "h" "h" "h" "h"
## [96] "h"
```

Now we have 96 columns.

Now we remove all but the variables *hypage, ruralfacto,breastfeeding, birthweight, yearsofedu, female, zstunt, zweight, zwast, adm2.*

```
#define a list with names, which want to have in our df
nec.vars <- c("hypage","ruralfacto","breastfeeding", "birthweight",
  "yearsofedu", "female", "zstunt", "zweight", "zwast", "adm2")

#we make same procedure as above
#but without ! in the right side
#it means that we leave only necessary columns in the df
childrenfinal <- childrenfinal[, (names(childrenfinal) %in% nec.vars)]
head(childrenfinal,2)
```

```
##   hypage ruralfacto female breastfeeding birthweight zstunt zweight zwast
## 1      6          1      0             6        3500   2.23    0.72 -1.32
## 2     28          1      0            24        3600   1.67    0.81  0.12
##   yearsofedu  adm2
## 1          8 NITHI
## 2          8 NITHI
```

So, now our dataset has 10 columns, what will allow us to work further. Although, we have to be sure, that all remaining variables have reasonable variable type:

```
str(childrenfinal)#quick review of variable types in df
```

```
## 'data.frame':    4686 obs. of  10 variables:
##  $ hypage       : num  6 28 20 47 14 15 50 14 10 18 ...
##  $ ruralfacto   : num  1 1 1 1 1 0 0 0 0 0 ...
##  $ female       : num  0 0 0 1 1 0 1 0 0 1 ...
##  $ breastfeeding: int  6 24 20 24 14 15 17 14 10 18 ...
##  $ birthweight  : int  3500 3600 2500 2500 2900 2800 3000 4000 3000 3500 ...
##  $ zstunt       : num  2.23 1.67 0.24 -2.44 -0.84 ...
##  $ zweight      : num  0.72 0.81 0.44 -1.82 -1.65 ...
##  $ zwast        : num  -1.32 0.12 0.45 -0.4 -1.38 ...
##  $ yearsofedu   : int  8 8 15 8 8 8 0 0 8 7 ...
##  $ adm2         : chr  "NITHI" "NITHI" "NITHI" "NITHI" ...
```

It seems that it would be better to format **female**, **ruralfacto** and **adm2** into factor type to easily categorize the data. We can do it with function "as.factor()":

```
#one approach it to index with the $ sign and the as.factor function

#convert gender column into factor
childrenfinal$female <- as.factor(childrenfinal$female)

#convert territory column into factor
childrenfinal$ruralfacto <- as.factor(childrenfinal$ruralfacto)

#convert provinces column into factor
childrenfinal$adm2 <- as.factor(childrenfinal$adm2)
```

Plots with Z-score

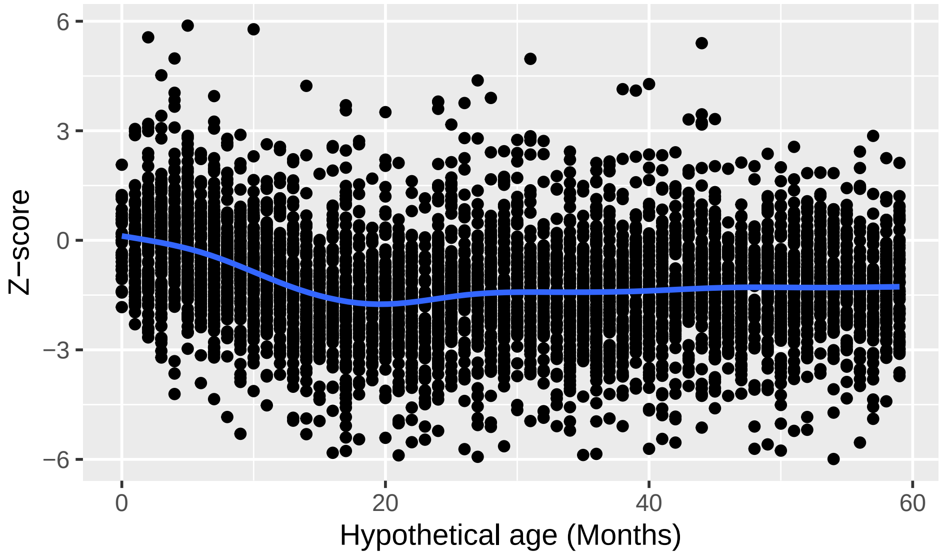
Variable **zstunt** is the so-called *Z-score* for stunting and is defined as the height of a child standardised with the median and standard deviation of heights of children at the same age from a healthy population. Children with Z-score less than -2 are defined to be stunted. We will make a scatter plot of **zstunt** against **hypage** with a smooth line to the plot, without confidence bands.

Lets plot the data. First, we initiate a ggplot2-object:

```
plt <- ggplot(data = childrenfinal)
```

Now, we can plot the relationship between **zstunt** and **hypage** using this object:

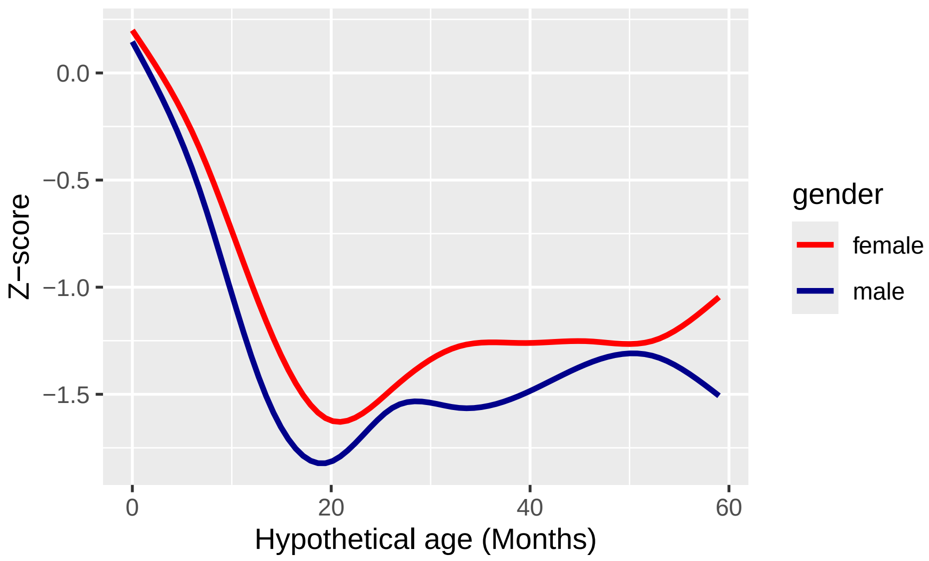
```
plt + geom_point(aes(x = hypage, y = zstunt))+
  geom_smooth(aes(x = hypage, y = zstunt ),se = F)+
  labs(x = "Hypothetical age (Months)",y = "Z-score")#add labels
```



There is a bit complicated to say something about linear dependence between variables, it seems that Z-score is negative in average. We can make an assumption that children in Kenya have quite serious problem with stunt.

Now we make smooth plots of **zstunt** against **hypage** for **females** and **males**. Here help us *filter*() function, which allows us elegantly to access specific information within data set, as in our case different plots for male/female. We will drop the scatter plot, that doesn't help to vizualize in any way:

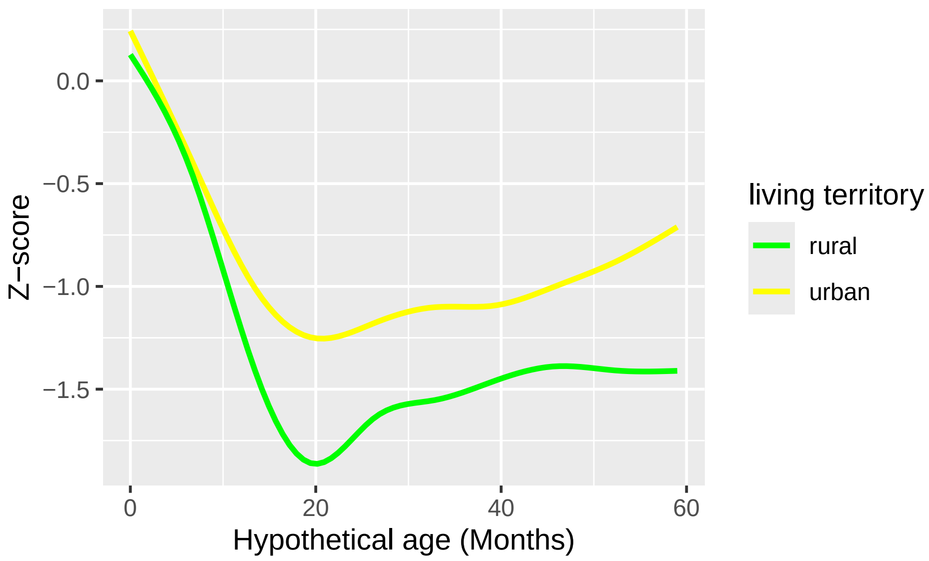
```
#Plot male and female smooth lines
#with filter(), where female == 0 is for male data
ggplot() +
 geom_smooth(data = filter(childrenfinal, female == 0),
             aes(x = hypage, y = zstunt, colour = "male" ),se = F) +
 geom_smooth(data = filter(childrenfinal, female == 1),
             aes(x = hypage, y = zstunt, colour = "female"), se = F)+
  labs(x = "Hypothetical age (Months)",y = "Z-score")+
  scale_colour_manual(name="gender", values=c("red","darkblue"))#set legend
```



We notice that female have smaller Z-score in average comparing to male, it means that in Kenya girls have less problems with stunt than boys.

Similarly, we plot **zstunt** against **hypage** for *urban* and *rural* children. We use the identical code from above:

```r
#Plot urban and rural smooth lines
#with filter(), where ruralfacto == 1 is the data for rural children
ggplot() +
 geom_smooth(data = filter(childrenfinal, ruralfacto == 0),
             aes(x = hypage, y = zstunt, colour = "urban" ),se = F) +
 geom_smooth(data = filter(childrenfinal, ruralfacto == 1),
             aes(x = hypage, y = zstunt, colour = "rural"), se = F)+
  labs(x = "Hypothetical age (Months)",y = "Z-score")+
  scale_colour_manual(name="living territory", values=c("green","yellow"))
```



Here we see a big deviation of urban territory from rural. We can conclude that children, who live in urban area, have less problems with stunt comparing to children living in rural area.