BMI trends in Australia — population surveys

SDC / BakerIDI
BendixCarstensen.com
November 2013
Version 3.2

Compiled Monday 11th November, 2013, 17:19
from: `C:/Bendix/Melbourne/IDI2012/Peeters/BMI-APC.tex`

Bendix Carstensen
Steno Diabetes Center, Gentofte, Denmark
& Department of Biostatistics, University of Copenhagen
bxc@steno.dk
http://BendixCarstensen.com

Contents

0.1 Introduction and overview

This is a report on trends in BMI in Australia, and how they depend on calendar time and date of birth.

The data base is survey data from Australia, where we have data on sex, BMI, date of survey, date of birth (and educational level).

The general approach will be to model the BMI-distribution by regression as a function sex, age, calendar time and date of birth. This can be seen as a variant of the Age-Period-Cohort model for occurrence rates ([1]), where we instead of rates are modeling mean BMI (or, as will appear later, median BMI as well as other quantiles of the BMI distribution).

Chapter 1

Summary

This chapter gives a summary of the findings documented in subsequent chapters. There is also also a section with a first proposal text to the paper.

1.1 Preliminary observations

- Log-transformed BMI gives a better fit than non-transformed.
- Log-transformed BMI is approximately symmetric for men (as measured by residuals), but right-skewed for women.
- The distribution of BMI is inadequately described by a log-normal distribution.
- Assuming a normal distribution of the log-transformed BMI still understates the variation in BMI among women; in particular it will not capture the very right-skewed distribution of BMI among women
- BMI is smaller for women than for men in ages up to mid 50es, in older ages, mean BMI is similar for men and women.
- The BMI distribution is largely constant after age 60.
- Population variation in BMI is substantially larger for females than for males, and in particular substantially more right-skewed than for males.
- The secular changes in BMI is particularly pronounced for the upper 25% of the population with the highest BMI. Both the absolute increase and the relative increase in BMI is larger, the more obese people are.
- BMI is increasing for the entire population, the smallest (both relative and absolute) increase is for the leanest part of the population, the increase in BMI is steeper, the larger the BMI.

1.2 Proposed article sections

1.2.1 Material and methods

. . .

1.2.1.1 Data preparation

For the National health surveys where no individual date of examination were available we assigned a random date in the interval of the survey.

The analysis dataset had one record per person with complete information on date of birth, date of examination (and hence age at examination) as well as height and weight (and hence BMI). Figure 1 shows the distribution of age and date of examination in the dataset; the 6 surveys are clearly discernible.

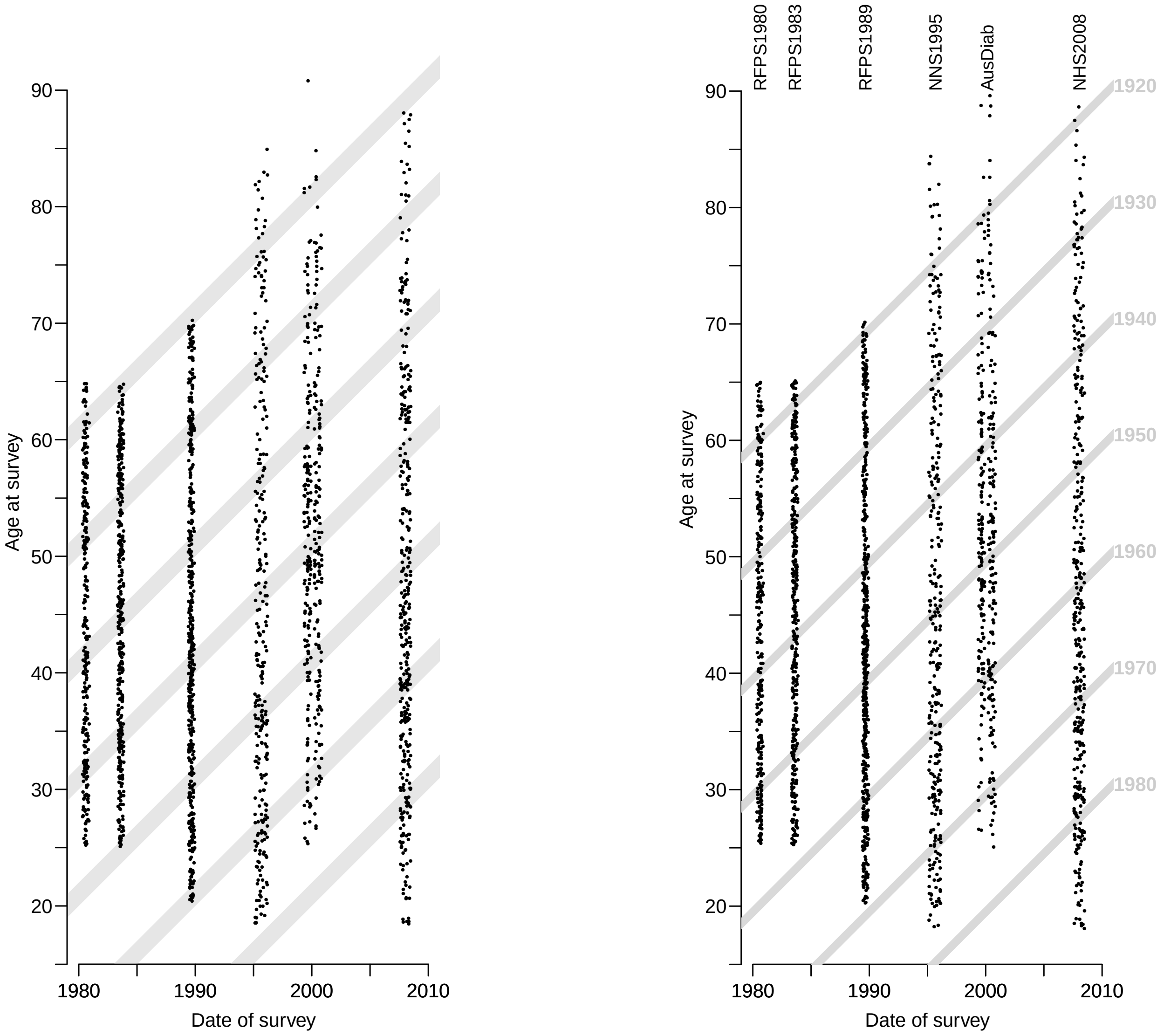

Figure 1.1: *Age and date of survey for a random 5% sample of persons in the analysis. The gray lines indicate the birth cohorts (from the top) 1920, 1930, . . . , 1980.*

Here we read the analysis file used throuighout and create the table 1 of the paper:

```
> library( Epi )
> library( quantreg )
> load( file="./data/abmi.Rdata" )
> abmi$age <- abmi$dos - abmi$dob
> str( abmi )
```

```
'data.frame':        42618 obs. of  11 variables:
 $ sex: Factor w/ 2 levels "M","F": 1 2 1 1 2 2 2 1 1 2 ...
 $ dob: num  1926 1935 1938 1955 1926 ...
 $ dos: num  1980 1980 1980 1981 1980 ...
 $ bmi: num  31.7 24.2 25.2 24.5 32.1 ...
 $ ht : num  1.73 1.7 1.78 1.85 1.72 1.6 1.53 1.73 1.76 1.61 ...
 $ wt : num  95 70 80 84 95 61 50 66 86 56 ...
 $ edu: Factor w/ 2 levels "lo","hi": 2 2 2 2 1 2 1 2 2 2 ...
 $ smk: Factor w/ 2 levels "non","cur": 2 2 1 2 1 1 1 1 1 1 ...
 $ aus: Factor w/ 2 levels "no","aus": 1 2 2 2 2 2 2 2 2 2 ...
 $ srv: Factor w/ 6 levels "RFPS1980","RFPS1983",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ age: num  54.4 45.4 42.2 25.2 54.6 ...
```

```
> t0 <- with( abmi, table( srv, sex ) )
> t1 <- with( abmi, table( srv, cut( age,
+                                    breaks=c(0,25,35,45,55,65,Inf),
+                                    right=FALSE ) ) )
> t2 <- with( abmi, table( srv, aus ) )
> t3 <- with( abmi, table( srv, edu ) )
> t4 <- with( abmi, table( srv, smk ) )
> arttab <- rbind( t( pctab( t0, dec=FALSE ) )[c(4,1),,drop=FALSE],
+                  t( pctab( t1, dec=FALSE ) )[1:6,,drop=FALSE],
+                  t( pctab( t2, dec=FALSE ) )[2,,drop=FALSE],
+                  t( pctab( t3, dec=FALSE ) )[2,,drop=FALSE],
+                  t( pctab( t4, dec=FALSE ) )[2,,drop=FALSE] )
> round( arttab, 1 )
```

```
         RFPS1980 RFPS1983 RFPS1989 NNS1995 AusDiab NHS2008
N          5572.0   7569.0   9134.0  6184.0  6758.0  7401.0
M            49.6     49.3     49.2    47.9    46.0    48.0
[0,25)        0.0      0.1      8.1    11.6     0.0     9.8
[25,35)      27.7     26.7     22.1    21.3    12.5    18.7
[35,45)      23.9     27.4     26.8    19.9    22.3    21.2
[45,55)      25.4     23.1     19.4    16.5    27.9    17.3
[55,65)      22.9     22.6     16.3    13.0    19.2    15.1
[65,Inf)      0.2      0.2      7.2    17.6    18.2    17.9
aus          71.1     70.6     72.7    68.9    70.6    66.7
hi           50.6     57.9     56.5    40.5    51.4    56.1
cur          32.2     29.7     24.0    23.7    14.5    19.8
```

```
> round( arttab )
```

```
         RFPS1980 RFPS1983 RFPS1989 NNS1995 AusDiab NHS2008
N            5572     7569     9134    6184    6758    7401
M              50       49       49      48      46      48
[0,25)          0        0        8      12       0      10
[25,35)        28       27       22      21      12      19
[35,45)        24       27       27      20      22      21
[45,55)        25       23       19      17      28      17
[55,65)        23       23       16      13      19      15
[65,Inf)        0        0        7      18      18      18
aus            71       71       73      69      71      67
hi             51       58       56      40      51      56
cur            32       30       24      24      14      20
```

```
> write.csv( round( arttab, 0 ), file="art-table1r.csv" )
> write.csv( round( arttab, 1 ), file="art-table1.csv" )
```

1.2.1.2 Statistical methods

BMI was analyzed separately for men and women. The basic assumption was that the distribution of BMI varies smoothly by age and date of survey and date of birth (an age-period-cohort model). Even though the BMI-distribution is only recorded at six narrow calendar time periods, we assumed that the surveys represent an overall smooth variation of BMI distribution. How the period from 1980 through 2009 is covered by measurements is illustrated in figure 1.

Initially we modelled mean log-BMI by a linear model with normal errors, but found that this approach that others [4] has taken violated assumptions about variance homogeneity, and assumptions about normality (primarily *symmetry*) of the residual distribution, in particular among women.

Since it was of primary interest to inspect the *distribution* of BMI as it evolves over time, we used quantile regression for the percentiles 5, 10, 25, 50, 75, 90 and 95. These percentiles were modeled (separately) by natural splines in age and and date of birth, both with 6 knots (and hence 5 parameters), and date of survey with 3 knots (one non-linear parameter). Also, a simplified model assuming a linear trend in BMI-quantiles by date of birth (and hence by date of survey too) was fitted in order to provide overall figures of trends in BMI percentiles in more detail. This was done both for absolute and relative BMI (using the log-transform).

The age-period-cohort model estimates are shown for each of the chosen percentiles as age-specific BMI, birth cohort effect centered at 1950 and residual period effect. This allows the age-specific BMIs to be interpreted as how the development will be in a particular birth cohort. Note that we are not considering mortality in this study, so we are modeling the BMI among *survivors* at a given age.

The 5-year change in all BMI-percentiles from 2 to 98, separately for men and women were computed based on quantile regression models with a 6-parameter age-effect and a linear drift (linear effect of date of birth or date of survey).

To show the changes in absolute levels of BMI over time, results from the models were shown as the age-specific percentiles in cohorts 1920, 1930,. . . ,1980 and for the survey dates 1980, 1985,. . . ,2010. These are shown for the 10th, 50th and 90th percentiles.

Finally we inverted the fitted models to estimate the fraction of persons that at a given age and time exceeded a BMI of 30, 35 and 40 kg/m^2. Confidence intervals for these fractions were computed by the bootstrap.

All analyses were done in R[3], version 3.0.1 with the `quantreg` package, [**?**]

```
> print( sessionInfo(), loc=F )
```

```
R version 3.0.2 (2013-09-25)
Platform: i386-w64-mingw32/i386 (32-bit)

attached base packages:
[1] utils     datasets  graphics  grDevices stats     methods   base

other attached packages:
[1] quantreg_4.98   SparseM_0.99   Epi_1.1.59     foreign_0.8-55

loaded via a namespace (and not attached):
[1] tools_3.0.2
```

A complete account of all data reading, transformation and all statistical analyses is available as `http://BendixCarstensen.com/IDI/BMI/BMI-APC.pdf`

1.2.2 Results

There were a total of 43,631 persons in the analysis surveyed between May 1980 and May 2009. Overall there were slightly fewer men than women in the surveys (see table 1).

```
> with( abmi, as.Date.cal.yr( range( dos ) ) )
```

```
[1] "1980-05-03" "2008-07-01"
```

Age-period-cohort modeling of the percentiles 5, 10, 25, 50, 75, 90 and 95, showed that the spread in BMI was very skew, the skewness increasing both by age and birth cohort, see figure 1.2. We found no clinically relevant curvature by period, although the effects were formally statistically significant.

We found that the distribution of BMI is spreading out over time, the increase by calendar time in the higher quantiles of BMI were steeper than among the lower percentiles as shown in figure 1.3. The increase in the 95th percentile was around 1 kg/m^2 per 5 years for women and 0.8 for men, whereas the increase in the median was only about a third of this and for the 10th percentile a fifth of this. It was noticeable that even the leanest 5 percent of the population saw a small increase of about 0.15 kg/m^2/5y. There was a small tendency that the increases were higher for males than females in the lower percentiles and vice versa for the higher percentiles, so there seems to be a tendency that the female distribution in the higher end of the BMI scale is spreading more over time than the male distribution (figure 1.3.

This tendency was also seen in the picture of how the 10th, 50th and 90th percentiles develop by age for different generations in figure 1.4. The tendency was less pronounced when the effects were shown by date of survey, because the general increase in all BMI-percentiles is causing the age-effects within generations to be masked (figure 1.5).

When inverting the model to assess the fraction of the population exceeding BMI 30, 35 or 40, we found that in the latest period more than 30% of the middle aged men and women had BMI over 30 kg/m^2, whereas it was only about 10% at the peak age that exceeded a BMI of 35 kg/m^2 (figure 5.3). However it was remarkable how rapidly the fraction exceeding 30 and 35 kg/m^2 increased by age within cohorts (figure 5.4), only for the threshold 35 there seemed to be a stabilization after age 50. The fraction exceeding 40 kg/m^2 was very small particularly for men, but we saw a weak tendency that about 5% of women already in ages 50 in the more recent cohort could reach this level.

1.2.3 Discussion

To our knowledge this is the first report that has analyzed trends in BMI allowing a description of the increasing spread in the distribution by quantile regression, while allowing a description of how BMI evolve across different generations of Australians.

By looking at the generational patterns we saw the tendency to a potential massive increase in BMI particularly in the younger generations, which were not immediately detectable from the cross-sectional figures. By taking the date of birth into account in the description of trends we were able to foreshadow these trends.

It was noticeable that the major increases were seen among the part of the population with the highest BMI, particularly in the most obese quarter of the population. Among the 20% most obese women, the increase in BMI was more than 0.8 kg/m^2 or 2.5% per 5 years, among men they were more than 0.5 kg/m^2 or 2% per year.

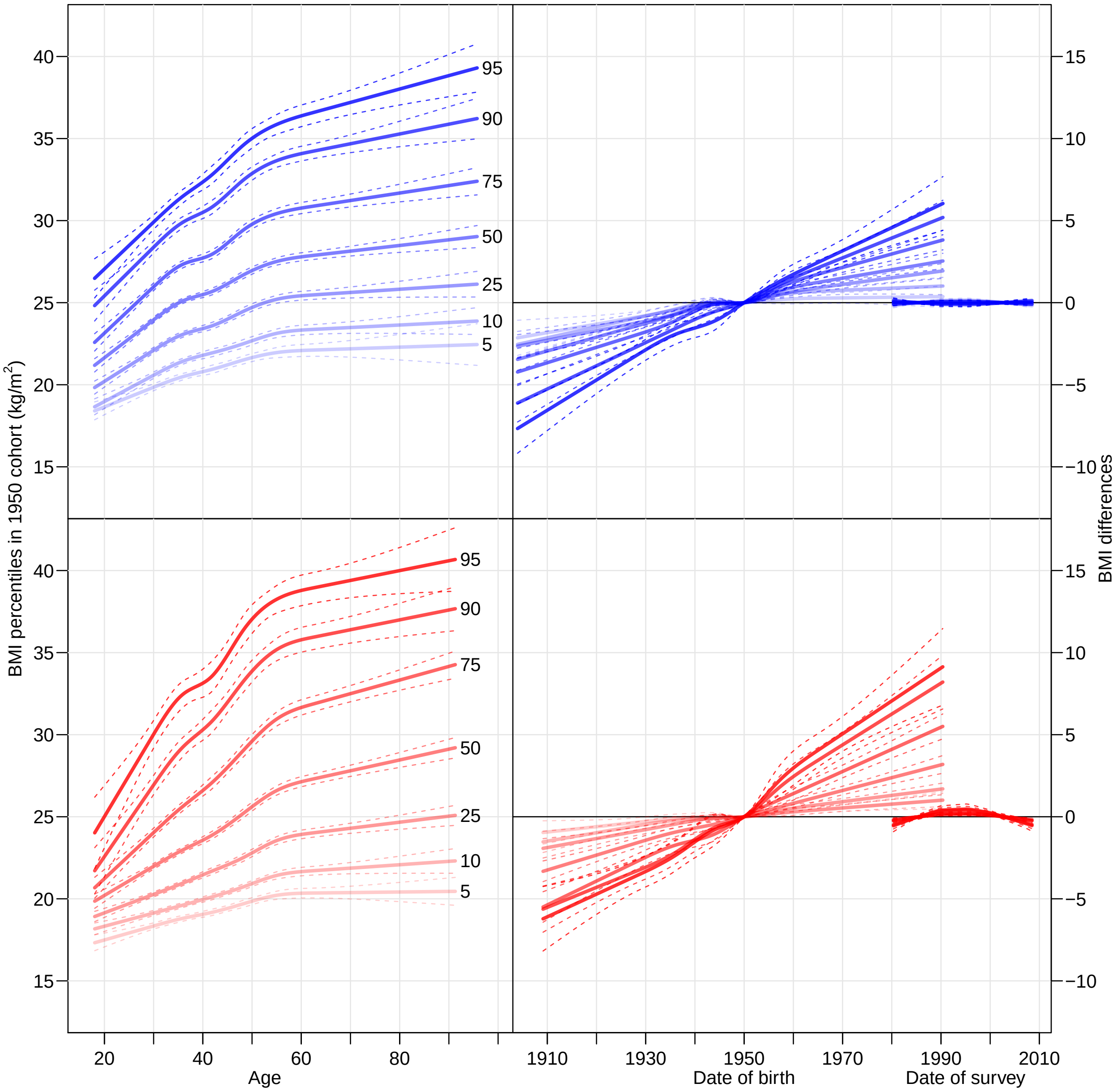

Figure 1.2: *Age, cohort and period effects for the 7 percentiles, separately for men (blue, top) and women (red, bottom). Estimates are from 7 separate models, one for each of the percentiles, each model shown in a separate shade. The dotted lines are 95% confidence intervals.*

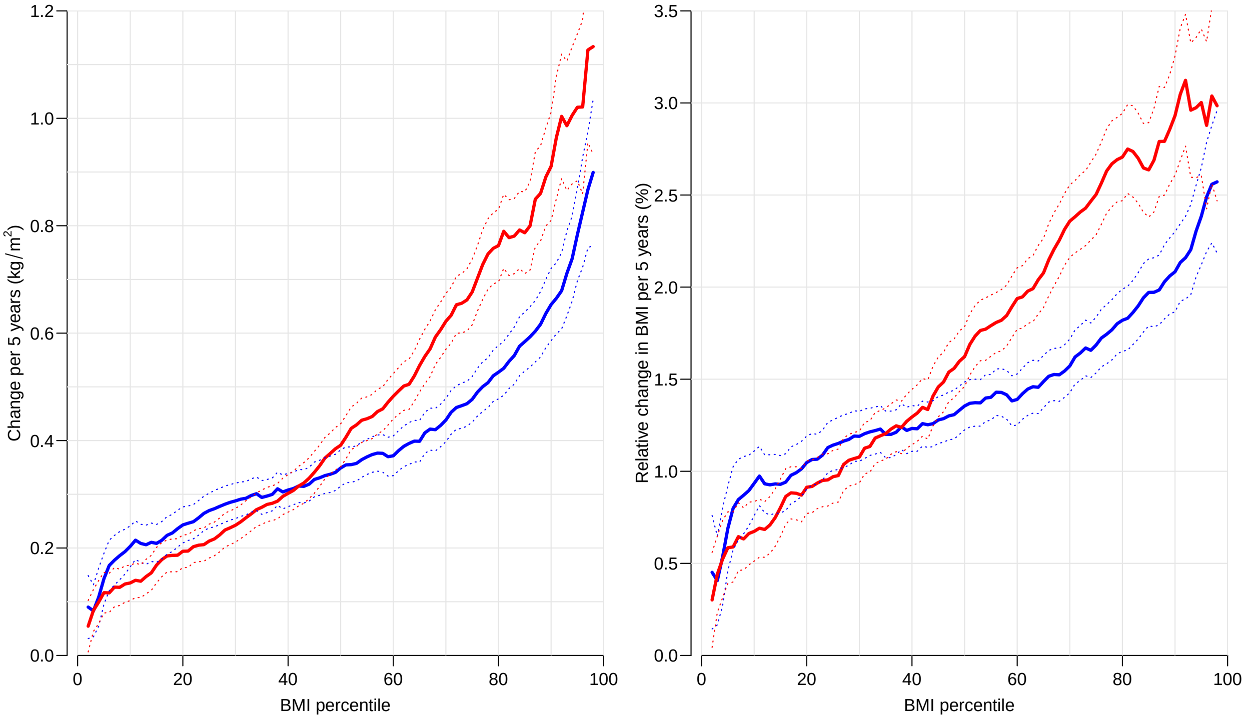

Figure 1.3: *Absolute (left) and relative (right) changes in quantiles of the BMI-distribution by time (date of birth / calendar time). Estimates are from separate models for each percentile from 2 to 98. There are no constraints on the models that require these estimates to be positive, nor increasing by percentile.*

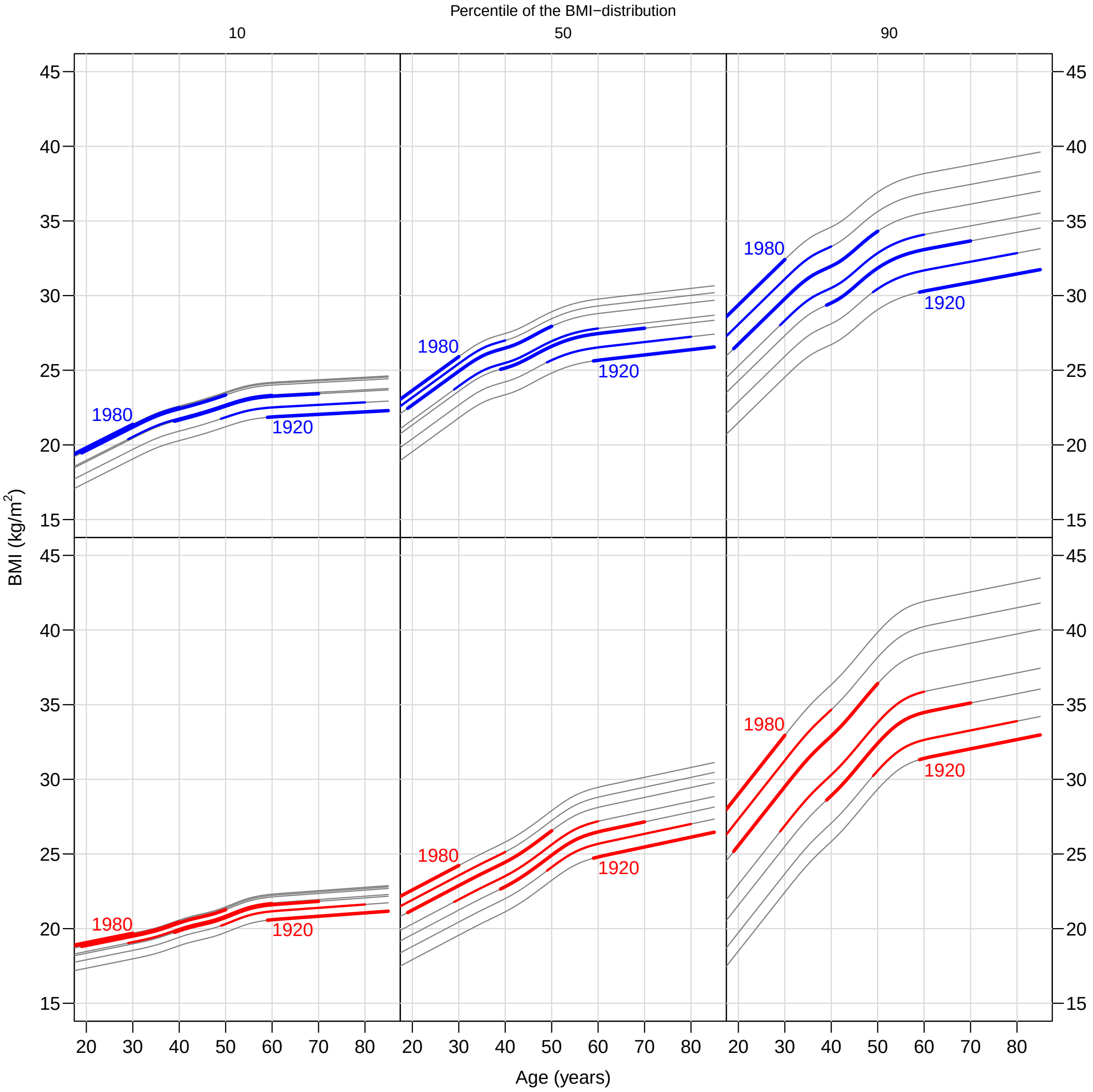

Figure 1.4: *Estimated age-specific BMI-quantiles for cohorts 1920,1930,...,1980.*

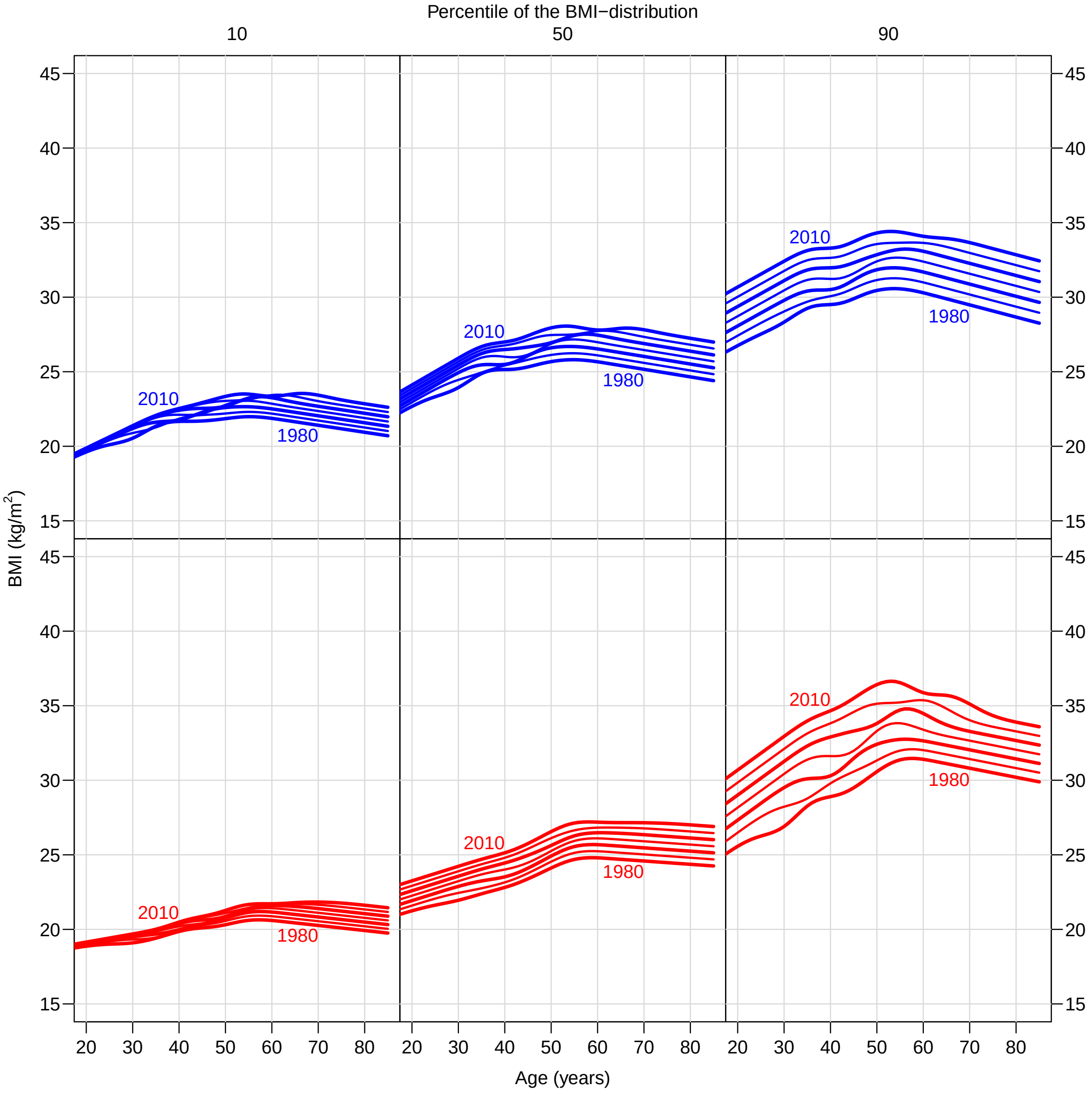

Figure 1.5: *Estimated age-specific BMI-quantiles at dates 1980, 1985,. . . ,2010.*

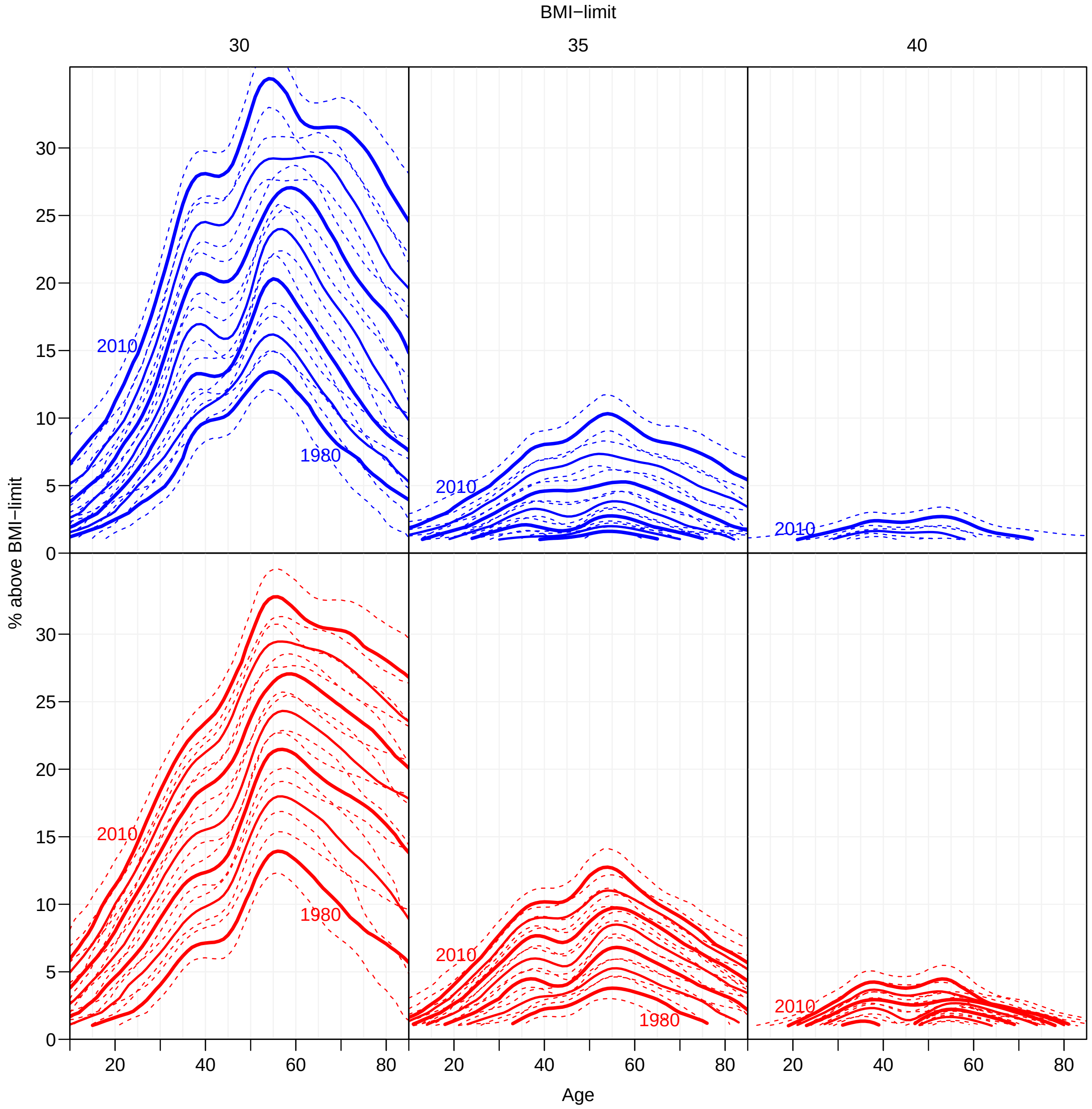

Figure 1.6: *Fraction of persons that exceed BMI 30, 35 and 40 at the dates 1980,1985,...,2010, as a function of age. Blue curves are for males, red for females. Full curves shows the median from a bootstrap sample of 1000, dotted curves the 95% confidence intervals.*

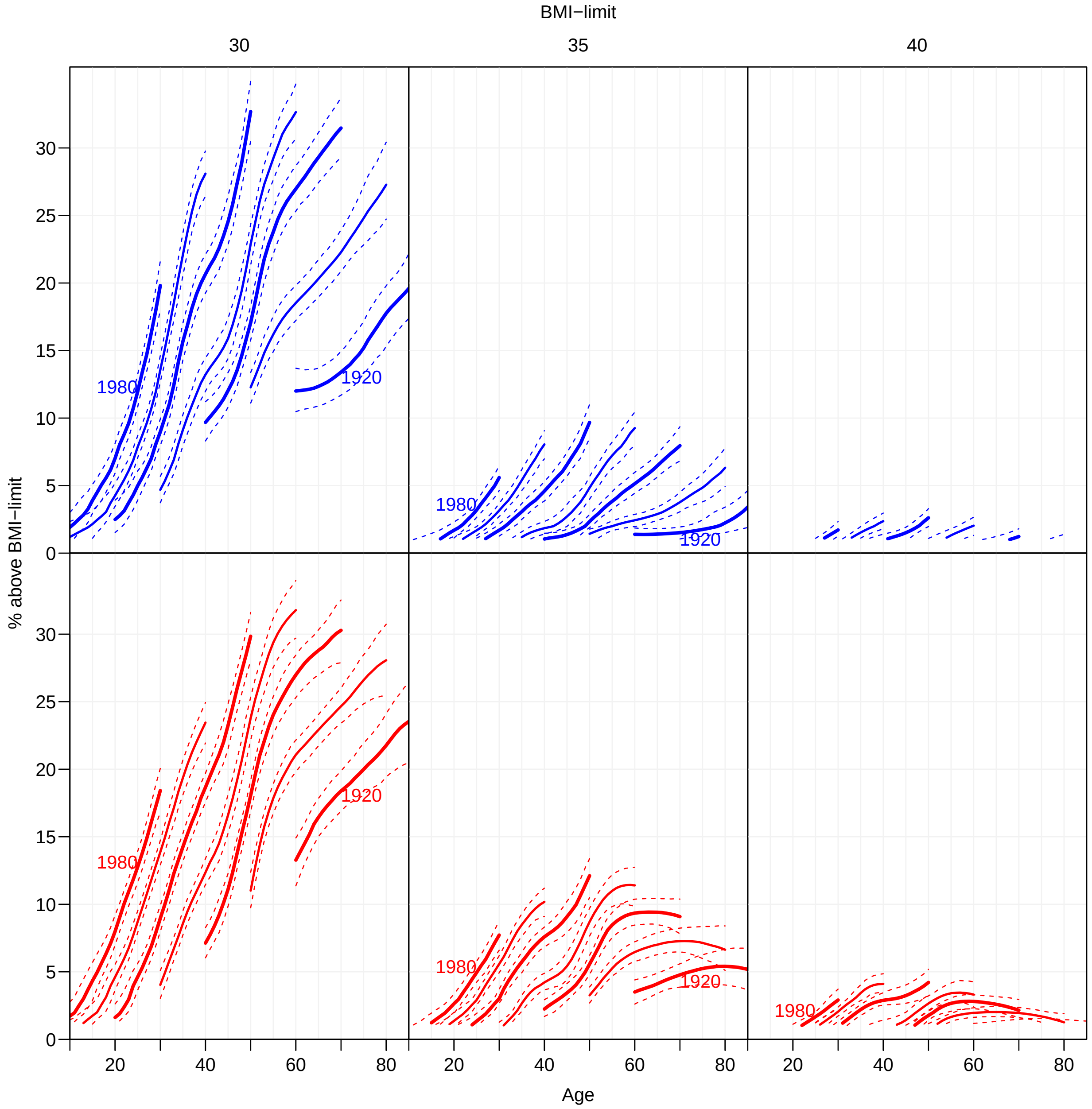

Figure 1.7: *Fraction of persons that exceed BMI 30, 35 and 40 from birth cohorts 1920,1930,. . . ,1980, as a function of age. Blue curves are for males, red for females. Full curves shows the median from a bootstrap sample of 1000, dotted curves the 95% bootstrap confidence intervals.*

The median BMI in the 1950 birth cohort is well below 30 throughout life and the increase in BMI for the leanest half of the population is moderate, but certainly not non-existing. The leanest quarter of the population has a BMI well below 25 throughout life, and the increase is BMI is only slightly more than 1% per 5 years.

Thus it is clear that the obesity increase is primarily in the upper part of the BMI distribution, meaning that the distribution of BMI is becoming increasingly skew, and that obesity problems can be expected to be increasingly severe in the most obese part of the population.

It should also be noticed that the leveling off of the BMI percentiles after age 50–60 to some extent may be attributable to excess mortality among the most obese, thus masking the obesity problems somewhat in older ages.

Chapter 2

Data sources for BMI trends

There are several sources of BMI data, and we shall derive a dataset with the following variables from all of the six surveys:

sex	sex (M,F)
dob	date of birth
dos	date of survey
bmi	measured BMI (kg/m^2)
ht	measured height (m)
wt	measured weight (kg)
edu	education (hi=high school completed, lo=not)
smk	smoking status (cur=current smoker, non=non-smoker)
aus	Australia born (yes, no)

The names of variables and codings of them are slightly different, so reading data from each survey requires its own piece of code, and in the log we will make tables that show how the original variables from the surveys are transformed to the variables that we will be using for the analyses.

We shall exclude persons with weight outside the range 40–140 kg and height outside range 145–200 cm, and since we are generating some random dates, we also explicitly set a seed to make sure that everything is exactly reproducible:

```
> w.lo <- 40
> w.hi <- 140
> h.lo <- 145
> h.hi <- 200
> set.seed( 1328765 )
> options( width=90 )
> library( Epi )
> library( foreign )
```

2.0.4 Random dates

Some of the surveys do not have a survey date for each participant, but instead of assigning everyone in the survey a fixed data in the middle of the survey period, we have chosen to assign survey participants a random date, uniformly distributed over the survey period.

2.1 RFPS - Risk factor prevalence surveys

These are three surveys done in 1980, 1983 and 1990, with different variable names and and -codings.

2.1.1 RFPS 1980

First we read the data from 1980 survey:

```
> rfps <- read.spss( file="rfps/RFPS1980.sav", to.data.frame=TRUE )
> names( rfps )
```

```
  [1] "centre"   "id"       "card1"    "rollsex"  "rollage"  "esd"      "q01day"  
  [8] "q01mnth"  "q01yr"    "q02"      "day"      "mnth"     "yr"       "q03"     
 [15] "q04"      "q05"      "q06"      "q0701"    "q0702"    "q0703"    "q0704"   
 [22] "q0705"    "q0706"    "q0707"    "q0708"    "q08"      "q09"      "q1001"   
 [29] "q1002"    "q1101"    "q1102"    "q1103"    "q1104"    "q1105"    "q1106"   
 [36] "q1107"    "q12"      "q13"      "q14"      "q15"      "q16"      "q17"     
 [43] "q18"      "q19"      "q20"      "q21"      "q22"      "q23"      "q24"     
 [50] "card2"    "q25"      "q2601"    "q2602"    "q2701"    "q2702"    "q2801"   
 [57] "q2802"    "q2803"    "q29"      "q30"      "q31"      "q32"      "q3301"   
 [64] "q3302"    "q3303"    "q3304"    "q3305"    "q3306"    "q3307"    "q3308"   
 [71] "q3401"    "q3402"    "q35"      "q36"      "q37"      "q38"      "q39"     
 [78] "q40"      "q41"      "q42"      "q4301"    "q4302"    "q4303"    "q4304"   
 [85] "q4305"    "q4306"    "q4307"    "q44"      "q45"      "q46"      "q47"     
 [92] "q48"      "card3"    "q49"      "q50"      "q51"      "q5201a"   "q5201b"  
 [99] "q5202a"   "q5202b"   "q5203a"   "q5203b"   "q5204a"   "q5204b"   "q5205a"  
[106] "q5205b"   "q5206a"   "q5206b"   "q5301a"   "q5301b"   "q5302a"   "q5302b"  
[113] "q5303a"   "q5303b"   "q5304a"   "q5304b"   "q5305a"   "q5305b"   "q5306a"  
[120] "q5306b"   "q54"      "card4"    "q55"      "q5601"    "q5602"    "q5603"   
[127] "q5604"    "q5605"    "q5606"    "q5607"    "q5608"    "q5609"    "q5610"   
[134] "q5611"    "q5612"    "height"   "weight"   "sys1"     "sys2"     "dia1"    
[141] "dia2"     "pulse"    "response" "card5"    "rose01"   "rose02"   "rose03"  
[148] "rose04"   "rose05"   "rose0601" "rose0602" "rose0603" "rose0604" "rose0605"
[155] "rose07"   "rose08"   "rose09"   "rose1001" "rose1002" "card6"    "serchol" 
[162] "hdlchol"  "triglyc"  "daystoba" "popn"     "samplen"  "age"      "bthplace"
[169] "fastflag"
```

```
> sex <- factor( rfps$rollsex, labels=c("M","F") )
> table( rollsex=rfps$rollsex, sex, exclude=NULL )
```

```
       sex
rollsex    M    F <NA>
   1    2764    0    0
   2       0 2839    0
   <NA>    0    0    0
```

```
> dob <- with( rfps, cal.yr( as.Date(paste("19",q01yr,"-",q01mnth,"-",q01day,sep="")) ) )
> dos <- with( rfps, cal.yr( as.Date(paste("19",yr,"-",mnth,"-",day,sep="")) ) )
> ht  <- rfps$height/100
> wt  <- rfps$weight
> bmi <- wt/ht^2
> edu <- factor( rfps$q06>=4, labels=c("lo","hi") )
> table( q06=rfps$q06, edu, exclude=NULL )
```

```
      edu
q06      lo   hi <NA>
  1      21    0    0
  2     832    0    0
  3    1914    0    0
  4       0 1703    0
  5       0 1133    0
  <NA>    0    0    0
```

```
> smk <- factor( rfps$q49==1 & rfps$q51==8888, labels=c("non","cur") )
> ftable( table( q51=rfps$q51, smk, q49=rfps$q49 )[-(10:330),,], row.vars=1 )
```

```
     smk  non       cur
     q49    1    2    1    2
q51
34          1    0    0    0
38          2    0    0    0
39          1    0    0    0
40          2    0    0    0
42          1    0    0    0
44          1    0    0    0
45          7    0    0    0
46          2    0    0    0
48          1    0    0    0
1276        4    0    0    0
1277        2    0    0    0
1278        5    0    0    0
1279       10    0    0    0
8888        0    0 1803    0
9999        2    0    0    0
```

```
> aus <- factor( rfps$q04<=8, labels=c("no","aus") )
> table( q04=rfps$q04, aus )
```

```
    aus
q04    no  aus
  1     0 1104
  2     0  591
  3     0  624
  4     0  517
  5     0  513
  6     0  626
  7     0    2
  8     0    4
  9    70    0
  10  657    0
  11  292    0
  12  363    0
  13  170    0
  14   50    0
  15    1    0
  16    8    0
  17    5    0
  18    6    0
```

```
> d1  <- data.frame( sex, dob, dos, bmi, ht, wt, edu, smk, aus, srv="RFPS1980" )
> str( d1 )
```

```
'data.frame':        5603 obs. of  10 variables:
 $ sex: Factor w/ 2 levels "M","F": 1 2 1 1 2 2 2 1 1 2 ...
 $ dob:Classes 'cal.yr', 'numeric'  num [1:5603] 1926 1935 1938 1955 1926 ...
 $ dos:Classes 'cal.yr', 'numeric'  num [1:5603] 1980 1980 1980 1981 1980 ...
 $ bmi: num  31.7 24.2 25.2 24.5 32.1 ...
 $ ht : num  1.73 1.7 1.78 1.85 1.72 1.6 1.53 1.73 1.76 1.61 ...
 $ wt : num  95 70 80 84 95 61 50 66 86 56 ...
 $ edu: Factor w/ 2 levels "lo","hi": 2 2 2 2 1 2 1 2 2 2 ...
 $ smk: Factor w/ 2 levels "non","cur": 2 2 1 2 1 1 1 1 1 1 ...
 $ aus: Factor w/ 2 levels "no","aus": 1 2 2 2 2 2 2 2 2 2 ...
 $ srv: Factor w/ 1 level "RFPS1980": 1 1 1 1 1 1 1 1 1 1 ...
```

```
> head( d1 )
```

```
  sex      dob      dos      bmi   ht wt edu smk aus      srv
1   M 1925.992 1980.368 31.74179 1.73 95  hi cur  no RFPS1980
2   F 1935.027 1980.382 24.22145 1.70 70  hi cur aus RFPS1980
3   M 1938.175 1980.385 25.24934 1.78 80  hi non aus RFPS1980
4   M 1955.328 1980.502 24.54346 1.85 84  hi cur aus RFPS1980
5   F 1925.795 1980.346 32.11195 1.72 95  lo non aus RFPS1980
6   F 1946.315 1980.382 23.82812 1.60 61  hi non aus RFPS1980
```

2.1.2 RFPS1983

```
> rfps <- read.spss( file="rfps/RFPS1983.sav", to.data.frame=TRUE )
> names( rfps )
```

```
  [1] "place"   "id"      "card1"   "sex"     "agegp"   "subdiv"  "q1day"   "q1mth"  
  [9] "q1yr"    "q2"      "dateday" "datemth" "dateyr"  "q3"      "q4"      "q5"     
 [17] "q6"      "q7a"     "q7b"     "q7c"     "q7d"     "q7e"     "q7f"     "q7g"    
 [25] "q7h"     "q8"      "q9"      "q10"     "q11a"    "q11b"    "q11c"    "q11d"   
 [33] "q11e"    "q11f"    "q12"     "q13"     "q14"     "q15"     "q16"     "q17"    
 [41] "q18"     "q19a"    "q19b"    "q20a"    "q20b"    "q21a"    "q21b"    "q21c"   
 [49] "q22"     "q23"     "q24"     "seq1"    "card2"   "q25a"    "q25b"    "q25c"   
 [57] "q26a"    "q26b"    "q27a"    "q27b"    "q28a"    "q28b"    "q28c"    "q29"    
 [65] "q30"     "q31mth"  "q31yr"   "q32a"    "q32b"    "q32c"    "q32d"    "q33a"   
 [73] "q33b"    "q33c"    "q33d"    "q34"     "q35mth"  "q35yr"   "q36"     "q37"    
 [81] "q38"     "q39a"    "q39b"    "q39c"    "q39d"    "q39e"    "q39f"    "q39g"   
 [89] "q39h"    "seq2"    "card3"   "q40a"    "q40b"    "q40c"    "q40d"    "q40e"   
 [97] "q40f"    "q40g"    "q40h"    "q40i"    "q40j"    "q40k"    "q40l"    "weight" 
[105] "height"  "arm1"    "arm2"    "bpobs"   "sphygmo" "sys1"    "dias1"   "sys2"   
[113] "dias2"   "fasted"  "est1982" "obsno"   "var362"  "var363"  "var364"  "var365" 
[121] "var366"  "var367"  "var368"  "var369"  "var370"  "var371"  "var372"  "seq3"   
[129] "card4"   "totchol" "hdlchol" "triglyc" "glucose" "gamma"   "var431"  "var432" 
[137] "var433"  "var434"  "var435"  "var436"  "seq4"    "bmi"     "bmicat" 
```

```
> sex <- factor( 1+(rfps$sex %in% c("FEMALE","CODED MALE,BUT PROB FEMALE")),
+                labels=c("M","F") )
> table( sex=rfps$sex, sex, exclude=NULL )
```

```
                            sex
sex                             M    F <NA>
  MALE                       3735    0    0
  FEMALE                        0 3863    0
  CODED MALE,BUT PROB FEMALE    0   12    0
  CODED FEMALE,BUT PROB MALE    5    0    0
  <NA>                          0    0    0
```

```
> dob <- with( rfps, cal.yr( as.Date(paste("19",q1yr,"-",q1mth,"-",q1day,sep="")) ) )
> dos <- with( rfps, cal.yr( as.Date(paste("19",dateyr,"-",datemth,"-",dateday,sep="")) ) )
> ht  <- rfps$height/100
> wt  <- rfps$weight
> bmi <- wt/ht^2
> edu <- factor( rfps$q6>=4, labels=c("lo","hi") )
> table( q6=rfps$q6, edu, exclude=NULL )
```

```
      edu
q6        lo   hi <NA>
  1       24    0    0
  2      844    0    0
  3     2346    0    0
  4        0 2597    0
  5        0 1803    0
  <NA>     0    0    1
```

```
> smk <- factor( rfps$q29==1 & is.na(rfps$q31yr), labels=c("non","cur") )
> ftable(q31yr=rfps$q31yr, smk, q29=rfps$q29, row.vars=1, exclude=NULL )
```

```
      smk  non            cur             NA
      q29    1    2   NA    1    2   NA    1    2   NA
q31yr
40           3    0    0    0    0    0    0    0    0
42           1    0    0    0    0    0    0    0    0
43           2    0    0    0    0    0    0    0    0
44           4    0    0    0    0    0    0    0    0
45           4    0    0    0    0    0    0    0    0
46           6    0    0    0    0    0    0    0    0
47           3    0    0    0    0    0    0    0    0
48           6    0    0    0    0    0    0    0    0
49           5    0    0    0    0    0    0    0    0
50          18    0    0    0    0    0    0    0    0
51           7    0    0    0    0    0    0    0    0
52           5    0    0    0    0    0    0    0    0
53          13    0    0    0    0    0    0    0    0
54           5    0    0    0    0    0    0    0    0
55          10    0    0    0    0    0    0    0    0
56          16    0    0    0    0    0    0    0    0
57          17    0    0    0    0    0    0    0    0
58          26    0    0    0    0    0    0    0    0
59           6    0    0    0    0    0    0    0    0
60          39    0    0    0    0    0    0    0    0
61          10    0    0    0    0    0    0    0    0
62          22    0    0    0    0    0    0    0    0
63          50    0    0    0    0    0    0    0    0
64          20    0    0    0    0    0    0    0    0
65          34    0    0    0    0    0    0    0    0
66          26    0    0    0    0    0    0    0    0
67          26    0    0    0    0    0    0    0    0
68          47    0    0    0    0    0    0    0    0
69          35    0    0    0    0    0    0    0    0
70          85    0    0    0    0    0    0    0    0
71          49    0    0    0    0    0    0    0    0
72          46    0    0    0    0    0    0    0    0
73          90    0    0    0    0    0    0    0    0
74          59    0    0    0    0    0    0    0    0
75          65    0    0    0    0    0    0    0    0
76          78    0    0    0    0    0    0    0    0
77          68    0    0    0    0    0    0    0    0
78         126    0    0    0    0    0    0    0    0
79          88    0    0    0    0    0    0    0    0
```

```
80          94    0    0    0    0    0    0    0    0
81         107    0    0    0    0    0    0    0    0
82         151    0    0    0    0    0    0    0    0
83         189    0    0    0    0    0    0    0    0
NA           0 3597    0 2257    0    0    0    0    0
```

```
> aus <- factor( rfps$q4<=8, labels=c("no","aus") )
> table( q4=rfps$q4, aus )
```

```
    aus
q4     no  aus
  1     0 1411
  2     0  742
  3     0  771
  4     0  670
  5     0  997
  6     0  770
  7     0    7
  8     0    4
  9     2    0
  11    5    0
  12  107    0
  15  706    0
  16   26    0
  17  135    0
  18   34    0
  19   27    0
  20   23    0
  21   19    0
  22    8    0
  23   22    0
  24   67    0
  25   32    0
  26   64    0
  27   60    0
  28   10    0
  29   38    0
  30  117    0
  31  243    0
  32   31    0
  33    4    0
  34    7    0
  35   88    0
  36   22    0
  40    2    0
  42   17    0
  43   14    0
  44   19    0
  45    6    0
  46   46    0
  47   14    0
  49    2    0
  50    8    0
  51    2    0
  52    1    0
  53    2    0
  55   28    0
  56   23    0
  57    3    0
  58    7    0
  59   15    0
  60   12    0
  61    2    0
  63    3    0
  64    2    0
  65    6    0
  66   13    0
  70   70    0
  80    4    0
  81    4    0
  82   12    0
  83    8    0
```

```
> d2  <- data.frame( sex, dob, dos, bmi, ht, wt, edu, smk, aus, srv="RFPS1983" )
> str( d2 )
```

```
'data.frame':        7615 obs. of  10 variables:
 $ sex: Factor w/ 2 levels "M","F": 2 1 1 1 2 2 1 2 1 2 ...
 $ dob:Classes 'cal.yr', 'numeric'  num [1:7615] 1956 1926 1931 1937 1930 ...
 $ dos:Classes 'cal.yr', 'numeric'  num [1:7615] 1983 1983 1983 1983 1983 ...
 $ bmi: num  19.2 34.9 34 25.7 21.7 ...
 $ ht : num  1.67 1.8 1.77 1.83 1.6 1.77 1.7 1.6 1.75 1.62 ...
 $ wt : num  53.5 113.1 106.4 86.2 55.5 ...
 $ edu: Factor w/ 2 levels "lo","hi": 2 2 2 2 2 1 2 1 2 1 ...
 $ smk: Factor w/ 2 levels "non","cur": 2 1 2 2 1 1 1 1 1 1 ...
 $ aus: Factor w/ 2 levels "no","aus": 2 2 2 1 2 2 2 2 2 2 ...
 $ srv: Factor w/ 1 level "RFPS1983": 1 1 1 1 1 1 1 1 1 1 ...
```

```
> head( d2 )
```

```
  sex      dob      dos      bmi   ht    wt edu smk aus      srv
1   F 1955.832 1983.374 19.18319 1.67  53.5  hi cur aus RFPS1983
2   M 1926.197 1983.358 34.90741 1.80 113.1  hi non aus RFPS1983
3   M 1931.391 1983.377 33.96214 1.77 106.4  hi cur aus RFPS1983
4   M 1936.768 1983.418 25.73980 1.83  86.2  hi cur  no RFPS1983
5   F 1929.797 1983.361 21.67969 1.60  55.5  hi non aus RFPS1983
6   F 1941.526 1983.394 20.58795 1.77  64.5  lo non aus RFPS1983
```

2.1.3 RFPS1989

```
> rfps <- read.spss( file="rfps/RFPS1989.sav", to.data.frame=TRUE )
> names( rfps )
```

```
  [1] "questno"  "ssdata"   "dob"      "sex"      "datecom"  "mstat"    "kid014"  
  [8] "fts1524"  "livearr"  "plobir"   "inoz"     "educlev"  "measrbp"  "measrbc" 
 [15] "highbp"   "angina"   "heart"    "stroke"   "highcol"  "hightri"  "bptabs"  
 [22] "blodfat"  "tangina"  "tolddia"  "yeardia"  "sugurine" "tolduri"  "treatdia"
 [29] "treatyr"  "treatyp"  "ocpill"   "lengpill" "pillnow"  "preg"     "vexer"   
 [36] "nvexer"   "tvexer"   "lvexer"   "nlvexer"  "wexer"    "nwexer"   "vtask"   
 [43] "nvtask"   "timetask" "eversm"   "agesm"    "quitsm"   "brandq"   "rollq"   
 [50] "cigarq"   "pipeq"    "brand"    "roll"     "cigar"    "pipe"     "brandcig"
 [57] "golow"    "alcohol"  "ndrinks"  "salt"     "usualdit" "meatfat"  "fullmilk"
 [64] "skim"     "lowfat"   "yoghurt"  "lowfatyo" "cream"    "icecream" "height"  
 [71] "weight"   "ftorpt"   "mainjob"  "fulltime" "fulltimp" "parttime" "parttimp"
 [78] "nowork"   "noworkp"  "home"     "homep"    "ftstu"    "ftstup"   "ptstu"   
 [85] "ptstup"   "ret"      "retp"     "unable"   "unablep"  "other"    "otherp"  
 [92] "grossinc" "grosincp" "inc"      "incp"     "mweight"  "mheight"  "waista"  
 [99] "waistb"   "hipa"     "hipb"     "bpobs"    "sphyg"    "ambtemp"  "bpsa"    
[106] "bpda"     "bpsb"     "bpdb"     "fast"     "donor"    "donorwen" "iron"    
[113] "totchol"  "hdl"      "tri"      "pcode"    "pop"      "n"        "wt"      
[120] "agegp"    "born"    
```

```
> dob <- cal.yr(as.Date(paste(substr(as.character(rfps$dob),1,4),"19",
+                             substr(as.character(rfps$dob),5,6),sep=""),
+                             format="%d%m%Y"))
> dos <- cal.yr(as.Date(paste(substr(as.character(rfps$datecom),1,4),"19",
+                             substr(as.character(rfps$datecom),5,6),sep=""),
+                             format="%d%m%Y"))
> sex <- factor( 1+(rfps$sex=="Female"), labels=c("M","F") )
> table( sex=rfps$sex, sex )
```

```
        sex
sex         M    F
  Male   4552    0
  Female    0 4727
```

```
> ht  <- rfps$mheight/100
> wt  <- rfps$mweight/10
> bmi <- wt/ht^2
> edu <- factor( as.integer(rfps$educlev)>3, labels=c("lo","hi") )
> table( educlev=rfps$educlev, edu, exclude=NULL )
```

```
                       edu
educlev                   lo   hi <NA>
  Never attended school   26    0    0
  Primary school         796    0    0
  Some high school      3233    0    0
  Completed high school    0 2629    0
  Tertiary institution     0 2595    0
  <NA>                     0    0    0
```

```
> smk <- with( rfps, factor( quitsm==8888 & eversm=="Yes", labels=c("non","cur") ) )
> chktab <- addmargins( table( eversm=rfps$eversm,
+                              smk,
+                              quit=rfps$quitsm,
+                              exclude=NULL ),
+                       margin=c(1,3))
> ftable( chktab[,,c(1:5,dim(chktab)[3]-5:0)], row.vars=2:1 )
```

```
           quit   41   42   44   45   46 1287 1288 1289 8888   NA  Sum
smk eversm
non Yes            1    1    2    1    1   23   20    1    0    0 2397
    No             0    0    0    0    0    0    0    0    0 4634 4634
    NA             0    0    0    0    0    0    0    0    0    0    0
    Sum            1    1    2    1    1   23   20    1    0 4634 7031
cur Yes            0    0    0    0    0    0    0    0 2235    0 2235
    No             0    0    0    0    0    0    0    0    0    0    0
    NA             0    0    0    0    0    0    0    0    0    0    0
    Sum            0    0    0    0    0    0    0    0 2235    0 2235
NA  Yes            0    0    0    0    0    0    0    0    0   13   13
    No             0    0    0    0    0    0    0    0    0    0    0
    NA             0    0    0    0    0    0    0    0    0    0    0
    Sum            0    0    0    0    0    0    0    0    0   13   13
```

```
> aus <- factor( rfps$plobir<=8, labels=c("no","aus") )
> print( addmargins( table( aus, plobir=rfps$plobir ), 2 ), zero.print=".")
```

```
     plobir
aus      1    2    3    4    5    6    7    8    9   11   13   15   16   17   18   19
  no     .    .    .    .    .    .    .    .    1   24    2   39    2   11    1    5
  aus 1858  983  815 1363  646  842  113  113    .    .    .    .    .    .    .    .
     plobir
aus     21   22   24   25   26   27   29   30   31   33   34   35   36   43   45   46
  no     1    2    3    1    3   16    3    1    1    4    1    4    1   12   10    7
  aus    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .
     plobir
aus     48   51   52   54   55   56   57   58   59   60   61   63   64   65   66   67
  no     1    2    2    1    2   10    4   19    2    2    1    4   18    1   33    2
  aus    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .
     plobir
aus     70   81   82   83   91  111  112  113  114  115  116  131  132  133  134  151
  no    19    1    3    1    2   49   32    8    2    4    4    4    2    1    1  178
  aus    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .
     plobir
aus    152  153  154  155  156  161  162  163  164  171  172  173  174  176  181  182
  no   295  156   41    2   11    5   18    4    1   26   50   38    5    1    5   11
  aus    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .
     plobir
aus    183  184  186  191  192  193  201  202  203  204  205  211  212  213  214  221
  no     9    4    1    7    4    1    2    6   17    1    3    2    3    4    1    6
  aus    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .
     plobir
aus    222  223  224  225  231  232  233  241  242  243  244  245  251  252  253  254
  no     4    2    1    1    1    1    7    1   15   55    8    1    2    4   21    5
  aus    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .
     plobir
aus    255  261  262  263  264  271  272  273  274  281  283  284  291  292  293  294
  no     1    4   15   65    1    5   10   18   16    4    3    3   11   11   21   19
  aus    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .
     plobir
aus    296  301  302  303  304  305  306  311  312  313  314  315  316  321  322  323
  no     1   19   92   42    4    1    1   10   58  119    3    5    1    3    3   13
  aus    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .
     plobir
aus    324  331  332  333  341  342  343  351  352  353  354  355  361  362  363  364
  no     2    4    3    1    1    5    1   25   46   17    5    3    1    1    4    4
  aus    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .
     plobir
aus    365  402  411  422  424  431  432  433  434  435  441  442  443  444  445  451
  no     2    2    1    4    2    7    4    5    1    1   15    6    9    1    1    8
  aus    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .
     plobir
aus    452  454  461  462  463  464  471  472  473  481  491  501  504  511  512  521
  no     2    1   23   11    3    1    2    1    4    3    1    1    1    1    1    2
  aus    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .
     plobir
aus    531  541  551  552  553  554  555  561  562  563  565  571  573  574  581  582
  no     1    1   16   11    3    1    1   14    7    4    1    1    1    1    9    4
  aus    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .
     plobir
aus    591  592  601  602  604  611  612  613  621  622  631  632  633  641  642  651
  no     8    1    3    1    1    1    1    1    1    2    1    1    1   20    2    6
  aus    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .
     plobir
aus    652  653  654  661  671  701  702  703  704  801  802  803  805  811  812  813
  no     2    1    1   16    1   25   34    7    2    4    5    1    2    8    5    1
  aus    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .
     plobir
aus    821  822  831  832  833  834  Sum
  no    18    5    2    5    1    2 2546
  aus    .    .    .    .    .    . 6733
```

```
> d3  <- data.frame( sex, dob, dos, bmi, ht, wt, edu, smk, aus, srv="RFPS1989" )
> str( d3 )
```

```
'data.frame':        9279 obs. of  10 variables:
 $ sex: Factor w/ 2 levels "M","F": 1 1 1 1 1 1 1 1 1 1 ...
 $ dob:Classes 'cal.yr', 'numeric'  num [1:9279] 1963 1962 1964 1955 1954 ...
 $ dos:Classes 'cal.yr', 'numeric'  num [1:9279] 1990 1989 1989 1990 1990 ...
 $ bmi: num  23.7 21 23.5 24.8 24.1 ...
 $ ht : num  1.7 1.76 1.85 1.75 1.8 1.8 1.72 1.85 1.76 1.71 ...
 $ wt : num  68.5 64.9 80.5 76 78 ...
 $ edu: Factor w/ 2 levels "lo","hi": 1 2 2 2 2 2 1 2 2 2 ...
 $ smk: Factor w/ 2 levels "non","cur": 2 2 2 1 1 1 2 1 2 1 ...
 $ aus: Factor w/ 2 levels "no","aus": 2 2 2 2 2 2 2 2 2 1 ...
 $ srv: Factor w/ 1 level "RFPS1989": 1 1 1 1 1 1 1 1 1 1 ...
```

```
> head( d3 )
```

```
  sex      dob      dos      bmi   ht   wt edu smk aus      srv
1   M 1962.775 1989.532 23.70242 1.70 68.5  lo cur aus RFPS1989
2   M 1961.502 1989.455 20.95170 1.76 64.9  hi cur aus RFPS1989
3   M 1964.311 1989.425 23.52082 1.85 80.5  hi cur aus RFPS1989
4   M 1955.410 1989.562 24.81633 1.75 76.0  hi non aus RFPS1989
5   M 1953.540 1989.532 24.07407 1.80 78.0  hi non aus RFPS1989
6   M 1950.635 1989.428 23.14815 1.80 75.0  hi non aus RFPS1989
```

2.2 National Health/Nutrition Surveys

2.2.1 1995 Health Survey

We first read the data from the 1995 health survey (the NNS, that is the national nutrition survey):

```
> nns  <- read.spss( file="nns95/nns95.sav", to.data.frame=TRUE )
> dim( nns )
```

```
[1] 13858   158
```

```
> names(nns)
```

```
  [1] "randomid" "persnop"  "idp"      "sidp"     "hhtype"   "sex"      "marstat" 
  [8] "cob"      "yoarr"    "age"      "famrelcp" "ffqflag"  "state"    "rrmaclas"
 [15] "qlowcda"  "typoccp"  "nrperh"   "lansah"   "wcurstdy" "schatt"   "wcomsesc"
 [22] "aglftsch" "empstat"  "umact"    "occa"     "hrswkd"   "nuwklfj"  "nfutempp"
 [29] "incmsrc"  "gainchp"  "equivinc" "obsa"     "pwhrespa" "pwhrespb" "pwhrespc"
 [36] "pwhrespd" "pwhrespe" "pwhrespf" "pwhrespg" "pwhresph" "pwhrespi" "pwhrespj"
 [43] "pwhrespk" "pwhrespl" "obsc"     "diettype" "foodinpd" "wchange"  "reaswcha"
 [50] "reaswchb" "reaswchc" "reaswchd" "reaswche" "reaswchf" "reaswchg" "nofood"  
 [57] "accffq"   "mintake"  "dayofwee" "breastfd" "tenergy"  "tprotein" "ttotfat" 
 [64] "ttotsfat" "ttotmfat" "ttotpfat" "tcholest" "ttotsugr" "tstarch"  "tcarbohy"
 [71] "tdfibre"  "tvitare"  "tvitapre" "tprovita" "tthiamin" "tribofla" "tniacinp"
 [78] "tniacind" "tniacine" "tfolate"  "tvitac"   "tcalcium" "tphospho" "tmagnesi"
 [85] "tiron"    "tzinc"    "tpotassi" "talcohol" "tmoistur" "kprotein" "ktotfat" 
 [92] "ktotsfat" "ktotmfat" "ktotpfat" "kcholest" "ktotsugr" "kstarch"  "kcarbohy"
 [99] "kdfibre"  "kvitare"  "kvitapre" "kprovita" "kthiamin" "kribofla" "kniacinp"
```

```
[106] "kniacind" "kniacine" "kfolate"  "kvitac"   "kcalcium" "kphospho" "kmagnesi"
[113] "kiron"    "kzinc"    "kpotassi" "kalcohol" "kmoistur" "vegserve" "fruitsve"
[120] "selfhlth" "weighth"  "heighth"  "sasbwh"   "rawbmih"  "wcursmk"  "wcursreg"
[127] "whesmreg" "physacti" "pregnant" "medblpr"  "systave"  "diastave" "bpriskgr"
[134] "bprgint"  "notsecbp" "hyperten" "hghtave"  "noheight" "personwt" "adchwt"
[141] "noweight" "waistave" "hipave"   "nowaisth" "whratio"  "bodymixn" "bodymage"
[148] "bmi4aust" "zhtage"   "zwtage"   "zwtht"    "bmr"      "eibmr"    "ifiqwt"
[155] "ffqwt"    "famnop"   "iunop"    "pnop"
```

```
> # nns[1:10,c("sex","age","personwt","hghtave","rrmaclas","wcomsesc")]
```

We only want to include persons from large cities:

```
> levels(nns$rrmaclas)
```

```
[1] "Not available"                "Capital city/other metro"
[3] "Large/small rural areas"      "Other rural areas"
[5] "ACT/NT"                       "Brisbane"
[7] "Other QLD metro/rural centres"
```

```
> ( wh <- levels(nns$rrmaclas)[c(2,6)] )
```

```
[1] "Capital city/other metro" "Brisbane"
```

```
> nns <- subset( nns, rrmaclas %in% wh )
> with( nns, table(rrmaclas,exclude=NULL) )
```

```
rrmaclas
                Not available      Capital city/other metro       Large/small rural areas
                            0                          6968                             0
            Other rural areas                        ACT/NT                      Brisbane
                            0                             0                          1025
Other QLD metro/rural centres                          <NA>
                            0                             0
```

Then we create an age-variable which is uniformly distributed over the intervals:

```
> table( agr <- as.numeric( substr(nns$age,1,2) ), exclude=NULL )
```

```
   2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19
  93  106  119  118  106  100   98  105  101   98   85  109   77   81   78   78   77   85
  20   25   30   35   40   45   50   55   60   65   70   75   80 <NA>
 578  688  711  704  560  563  481  420  407  430  350  209  178    0
```

```
> age <- agr + runif(length(agr),0,1)*(1+4*(agr>19.5))
```

Then sex:

```
> sex <- factor( as.integer(nns$sex), labels=c("M","F") )
> table( nns$sex, sex, exclude=NULL )
```

```
         sex
             M    F <NA>
  Males   3810    0    0
  Females    0 4183    0
  <NA>       0    0    0
```

Then a random survey-date in the absence of a recorded one, the survey was conducted from February 1995 through March 1996, so we take the midpoints of these months as endpoints for the uniform distribution:

```
> dos <- 1995 + runif( length(age), 1.5/12, 14.5/12 )
> dob <- dos - age
```

Then we can compute the BMI and finalize the data frame:

```
> wt <- nns$personwt/10
> # hist(wt,breaks=0:200,col="black")
> ht <- nns$hghtave/10000
> # hist( ht*100, breaks=90:210,col="black")
> bmi <- wt/ht^2
> edu <- factor( as.integer(nns$wcomsesc)==2, labels=c("lo","hi") )
> table( wcomsesc=nns$wcomsesc, edu, exclude=NULL )
```

```
                                   edu
wcomsesc                              lo   hi <NA>
  Not applicable                    3926    0    0
  Completed secondary school           0 2610    0
  Did not complete secondary school 1457    0    0
  <NA>                                 0    0    0
```

Finally we construct the smoking and australian born variables:

```
> smk <- with( nns, factor( wcursmk=="Yes", labels=c("non","cur") ) )
> table( cursmk=nns$wcursmk, smk, exclude=NULL )
```

```
                smk
cursmk            non  cur <NA>
  Not applicable 1568    0    0
  Yes               0 1504    0
  No             4921    0    0
  <NA>              0    0    0
```

```
> aus <- factor( nns$cob=="Australia", labels=c("no","aus") )
> table( cob=nns$cob, aus, exclude=NULL )
```

```
                          aus
cob                          no  aus <NA>
  Australia                   0 5863    0
  New Zealand               151    0    0
  UK & Ireland              847    0    0
  Italy                      87    0    0
  Greece                     45    0    0
  Other Southern Europe     107    0    0
  Western Europe            159    0    0
  Other Europe,USSR,Baltic  116    0    0
  Middle East                43    0    0
  Vietnam                    38    0    0
  Other Southeast Asia      148    0    0
  Southern Asia              81    0    0
  Northeast Asia             92    0    0
  All other                 216    0    0
  <NA>                        0    0    0
```

```
> n95 <- data.frame( sex, dob, dos, bmi, ht, wt, edu, smk, aus, srv="NNS1995" )
> head( n95 )
```

```
  sex      dob      dos      bmi     ht   wt edu smk aus     srv
1   M 1934.988 1995.346 31.38062 1.6765 88.2  lo non aus NNS1995
2   F 1939.488 1995.705 33.51182 1.4960 75.0  lo non aus NNS1995
3   F 1945.225 1995.790 21.62182 1.6350 57.8  lo cur aus NNS1995
4   M 1919.057 1995.313 29.67749 1.7210 87.9  lo non  no NNS1995
5   F 1922.991 1995.732 28.85963 1.6110 74.9  lo non aus NNS1995
6   M 1950.502 1995.878 22.95265 1.6880 65.4  hi non  no NNS1995
```

2.2.2 2008 Health Survey

We then read the data from the 2008 health survey

```
> bsp  <- read.dta( file="nhs/nhs08bhhbspAP.dta" )
> dim( bsp )
```

```
[1] 20788   463
```

```
> # names(bsp)
>
> names(bsp)[grep("RA",names(bsp))]
```

```
 [1] "RA2006CF" "INCRACFA" "INCRACFB" "INCRACFC" "INCRACFD" "INCRACFE" "EXTRAMCF"
 [8] "MWBPSYRA" "RPHICVRA" "EXTRANO"
```

```
> bsp[1:10, c("RA2006CF","SEX", "AGEPRSBN", "PHDKGWCF", "PHDCMHCF",
+              "BMWEIGCF","MEASWEIG","BMHEIGCF","MEASHEIG","EDATTCF")]
```

```
                    RA2006CF    SEX    AGEPRSBN PHDKGWCF PHDCMHCF BMWEIGCF
1  Major cities of Australia Female 50-54 years      997      997      110
2   Inner Regional Australia Female 70-74 years       62      159       58
3  Major cities of Australia Female 60-64 years      997      997       90
4   Inner Regional Australia Female 70-74 years       50      164       48
5  Major cities of Australia Female 50-54 years      997      997       68
6  Major cities of Australia Female 65-69 years       58      161       60
7   Inner Regional Australia Female 60-64 years       71      166       70
8  Major cities of Australia   Male 70-74 years      997      997       85
9  Major cities of Australia   Male 60-64 years      110      170      110
10  Inner Regional Australia Female 50-54 years       75      167       74
              MEASWEIG BMHEIGCF            MEASHEIG
1  Measurement refused      168 Measurement refused
2             Measured      157            Measured
3  Measurement refused      999 Measurement refused
4             Measured      155            Measured
5  Measurement refused      173 Measurement refused
6             Measured      163            Measured
7             Measured      170            Measured
8  Measurement refused      173 Measurement refused
9             Measured      170            Measured
10            Measured      165            Measured
                                           EDATTCF
1  Year 8 or below including never attended school
2                            Year 10 or equivalent
3                             Year 9 or equivalent
4                            Year 12 or equivalent
5                            Year 10 or equivalent
6  Year 8 or below including never attended school
7                            Year 10 or equivalent
8                            Year 10 or equivalent
9                             Year 9 or equivalent
10                           Year 12 or equivalent
```

```
> match( c("RA2006CF","SEX", "AGEPRSBN", "PHDKGWCF", "PHDCMHCF",
+             "BMWEIGCF","MEASWEIG","BMHEIGCF","MEASHEIG","EDATTCF"),
+        names(bsp) )
```

```
 [1]  10 343 114 134 137 136 323 139 324 133
```

```
> # Select persons from the major cities only
> bsp <- subset( bsp, RA2006CF=="Major cities of Australia" )
> dim( bsp )
```

```
[1] 14037   463
```

```
> sex <- factor( as.integer(bsp$SEX), labels=c("M","F") )
> table( bsp$SEX, sex )
```

```
        sex
            M    F
  Male   6638    0
  Female    0 7399
```

We can now tease out the-5-year age-group, and then put in some bogus exact age — note the uneven lengths of the age-classes: . . . , 10–15, 15–18, 18–25, 25–30, . . . :

```
> agr <- as.numeric( substr( gsub("-"," -", bsp$AGEPRSBN ), 1, 2 ) )
> table( agr, exclude=NULL )
```

```
agr
   0    5   10   15   18   25   30   35   40   45   50   55   60   65   70   75   80   85
1052  827  876  640  986  927 1010 1204 1066 1026  886  819  789  596  454  408  288  183
<NA>
   0
```

```
> age <- agr + runif(nrow(bsp)) * (5+2*(agr==18)-2*(agr==15))
```

There are two different sets of height / weight variables, some are missing on some occasions, some on other, the variables `BMWEIGCF` and `BMHEIGCF` are self-reported so we do not use these, but the variables `PHDKGWCF` and `PHDCMHCF`

```
> table( bsp$PHDKGWCF, exclude=NULL )
```

```
   0    1   40   41   42   43   44   45   46   47   48   49   50   51   52   53   54   55
3395    6    5    6   14    8   19   24   31   33   38   39   60   59   77   71  100  109
  56   57   58   59   60   61   62   63   64   65   66   67   68   69   70   71   72   73
 102  125  118  119  150  141  152  167  172  170  151  159  187  148  187  151  184  165
  74   75   76   77   78   79   80   81   82   83   84   85   86   87   88   89   90   91
 161  210  178  179  194  161  173  148  161  150  135  163  111  128  105  115   83   95
  92   93   94   95   96   97   98   99  100  101  102  103  104  105  106  107  108  109
  99   88   88   94   86   58   69   69   66   57   53   40   43   46   40   42   36   19
 110  111  112  113  114  115  116  117  118  119  120  121  122  123  124  125  126  127
  30   21   24   20   21   18   14   12   10    8   16    8    6   11   10   11   10    7
 128  129  130  131  132  133  134  135  136  137  138  139  140  997 <NA>
   1    4    8    6    7    4    4    3    3    5    1    2   27 3120    0
```

```
> table( bsp$PHDCMHCF, exclude=NULL )
```

```
   0    1  145  146  147  148  149  150  151  152  153  154  155  156  157  158  159  160
3395   24   11   13   22   18   34   53   69  109  104   94  151  137  190  190  188  224
 161  162  163  164  165  166  167  168  169  170  171  172  173  174  175  176  177  178
 212  269  291  251  362  264  264  290  245  301  237  283  224  209  267  218  187  257
 179  180  181  182  183  184  185  186  187  188  189  190  191  192  193  194  195  196
 169  177  136  137  115  101  111   81   57   61   37   35   16   25   14   14   11    5
 197  198  199  200  997 <NA>
   3    5    3    2 3065    0
```

```
> addmargins( with( bsp, table( is.na(PHDKGWCF), MEASWEIG ) ) )
```

```
       MEASWEIG
        Not applicable Measured Measurement refused
  FALSE           1052     9087                3693
  Sum             1052     9087                3693
       MEASWEIG
        Measurement not taken for other reason   Sum
  FALSE                                    205 14037
  Sum                                      205 14037
```

```
> addmargins( with( bsp, table( is.na(PHDCMHCF), MEASHEIG ) ) )
```

```
       MEASHEIG
        Not applicable Measured Measurement refused
  FALSE           1052     9159                3680
  Sum             1052     9159                3680
       MEASHEIG
        Measurement not taken for other reason   Sum
  FALSE                                    146 14037
  Sum                                      146 14037
```

```
> ht <- bsp$PHDCMHCF/100
> wt <- bsp$PHDKGWCF
> bmi <- wt/ht^2
> edu <- factor( as.integer(bsp$EDATTCF)==2, labels=c("lo","hi") )
> table( bsp$EDATTCF, age>15, exclude=NULL )
```

```
                                                  FALSE TRUE <NA>
  Not applicable                                   2755    0    0
  Year 12 or equivalent                               0 5935    0
  Year 11 or equivalent                               0 1277    0
  Year 10 or equivalent                               0 2359    0
  Year 9 or equivalent                                0  844    0
  Year 8 or below including never attended school     0  866    0
  Not stated                                          0    1    0
  <NA>                                                0    0    0
```

```
> table( bsp$EDATTCF, edu, exclude=NULL )
```

```
                                                  edu
                                                     lo   hi <NA>
  Not applicable                                   2755    0    0
  Year 12 or equivalent                               0 5935    0
  Year 11 or equivalent                            1277    0    0
  Year 10 or equivalent                            2359    0    0
  Year 9 or equivalent                              844    0    0
  Year 8 or below including never attended school  866    0    0
  Not stated                                          1    0    0
  <NA>                                                0    0    0
```

We construct current smoking status and Australian born:

```
> smk <- with( bsp, factor( SMKSTAT %in% levels(SMKSTAT)[2:4], labels=c("non","cur") ) )
> table( bsp$SMKSTAT, smk, exclude=NULL )
```

```
                                                             smk
                                                               non  cur <NA>
  Not applicable                                              2755    0    0
  Current smoker daily                                           0 1961    0
  Current smoker weekly (at least once a week but not daily)     0  162    0
  Current smoker less than weekly                                0   45    0
  Ex-smoker                                                   3319    0    0
  Never smoked                                                5795    0    0
  <NA>                                                           0    0    0
```

```
> levels( bsp$COBCODBC ) <- substr( levels( bsp$COBCODBC ), 1, 31 )
> aus <- with( bsp, factor( COBCODBC=="Australia", labels=c("no","aus") ) )
> table( bsp$COBCODBC, aus, exclude=NULL )
```

```
                                 aus
                                     no   aus  <NA>
  Australia                           0 10165     0
  Main English speaking countries  1405     0     0
  Other                            2467     0     0
  <NA>                                0     0     0
```

We then make a continuous version of the data for age and date of survey; there is no specific survey date given for the individuals, so we assign a random date in the survey period 15 August 2007 to 15 July 2008:

```
> dos <- cal.yr("2007-08-01") + runif( nrow(bsp), 0, 11/12 )
> dob <- dos - age
> n8 <- data.frame( sex, dob, dos, bmi, ht, wt, edu, smk, aus, srv="NHS2008" )
> head( n8 )
```

```
  sex      dob      dos      bmi   ht  wt edu smk aus     srv
1   F 1955.647 2007.763 10.03009 9.97 997  lo non aus NHS2008
2   F 1947.763 2008.484 10.03009 9.97 997  lo non aus NHS2008
3   F 1953.283 2008.196 10.03009 9.97 997  lo cur aus NHS2008
4   F 1938.950 2008.147 22.37568 1.61  58  lo non  no NHS2008
5   M 1935.839 2008.292 10.03009 9.97 997  lo cur aus NHS2008
6   M 1947.088 2008.166 38.06228 1.70 110  lo non  no NHS2008
```

```
> summary(n8 )
```

```
 sex           dob            dos            bmi                 ht
 M:6638   Min.   :1918   Min.   :2008   Min.   :     0.4   Min.   :0.000
 F:7399   1st Qu.:1953   1st Qu.:2008   1st Qu.:    10.0   1st Qu.:1.490
          Median :1970   Median :2008   Median :    23.8   Median :1.680
          Mean   :1970   Mean   :2008   Mean   :  1277.6   Mean   :3.085
          3rd Qu.:1989   3rd Qu.:2008   3rd Qu.:    28.2   3rd Qu.:1.850
          Max.   :2008   Max.   :2008   Max.   :860000.0   Max.   :9.970
                                        NA's   :3395
       wt         edu        smk         aus             srv
 Min.   :  0.0   lo:8102   non:11869   no : 3872   NHS2008:14037
 1st Qu.: 47.0   hi:5935   cur: 2168   aus:10165
 Median : 75.0
 Mean   :262.8
 3rd Qu.:107.0
 Max.   :997.0
```

2.3 AUSDiab baseline survey (2000)

```
> ad <- read.dta( "ausdiab/AusDiabBase.dta")
> names( ad )
```

```
  [1] "id"                  "drsex_00"            "drdate_00"
  [4] "drage_00"            "agegrp_00"           "drdob_00"
  [7] "drhrs_00"            "drdiab_00"           "drtrt_00"
 [10] "diabstat_00"         "drpreg_00"           "drurine_00"
 [13] "drfastgl_00"         "drgluclo_00"         "drglucti_00"
 [16] "dr2hour_00"          "dr2hourt_00"         "drwhobsi_00"
 [19] "drwaist1_00"         "drhip1_00"           "drwaist2_00"
 [22] "drhip2_00"           "drwaist3_00"         "drhip3_00"
 [25] "drheight_00"         "drweight_00"         "drbioobs_00"
 [28] "bioweigh_00"         "drimped_00"          "drfatper_00"
 [31] "drfatmas_00"         "drlbm_00"            "drtbw_00"
 [34] "drbpobsi_00"         "drpulse_00"          "drsyst1_00"
 [37] "drdiast1_00"         "drsyst2_00"          "drdiast2_00"
 [40] "drsyst3_00"          "drdiast3_00"         "drecg_00"
 [43] "drsf36_00"           "drgenq_00"           "drgenqob_00"
 [46] "drhka_00"            "drhkaobs_00"         "drdietq_00"
 [49] "systolic_00"         "diastoli_00"         "choltabl_00"
 [52] "cd_area_00"          "cdcluste_00"         "state_00"
 [55] "systgrp_00"          "diastgrp_00"         "smokstat_00"
 [58] "waistot_00"          "waistfe_00"          "waistma_00"
 [61] "waist_00"            "hip_00"              "bmi_00"
 [64] "bmigrp_00"           "site_no_00"          "duration_00"
 [67] "agebysex_00"         "type_00"             "obese_00"
 [70] "overweig_00"         "absagegp_00"         "absabys_00"
 [73] "wtdrclin_00"         "stratum_00"          "exercise_00"
 [76] "urbrural_00"         "wsr_00"              "wsrcat_00"
 [79] "wtonsite_05"         "wtehc_05"            "dr_genqu_00"
 [82] "q1_gende_00"         "q2_date__00"         "q3_livin_00"
 [85] "q4_schoo_00"         "q5_schoo_00"         "q6_quali_00"
 [88] "q7_highe_00"         "q8_educa_00"         "q9_natur_00"
 [91] "q10_moth_00"         "q10_don__00"         "q11_moth_00"
 [94] "q11_don__00"         "q12_moth_00"         "q13_diag_00"
 [97] "q14_natu_00"         "q15_fath_00"         "q15_don__00"
[100] "q16_fath_00"         "q16_don__00"         "q17_fath_00"
[103] "q18_diag_00"         "q19_gout_00"         "q20_angi_00"
[106] "q20_coro_00"         "q20_stro_00"         "q21_bloo_00"
[109] "q22_bloo_00"         "q23_tabl_00"         "q24_lose_00"
[112] "q25_chol_00"         "q26_chol_00"         "q27_doc__00"
[115] "q28_tabl_00"         "q29_alco_00"         "q30_heav_00"
[118] "q31_alco_00"         "q32_drin_00"         "q33_alco_00"
[121] "q34_numb_00"         "q35_cut__00"         "q36_crit_00"
```

```
[124] "q37_bad__00"         "q38_hang_00"         "q39_alco_00"
[127] "q40_smok_00"         "q41_smok_00"         "q41smday_00"
[130] "q41smwk_00"          "q42_ciga_00"         "q42_cigd_00"
[133] "q42_cigw_00"         "q43_pipe_00"         "q43_pipd_00"
[136] "q43_pipw_00"         "q44_smok_00"         "q45_smok_00"
[139] "q46_week_00"         "q46_mont_00"         "q46yrold_00"
[142] "q46yrago_00"         "q47_age__00"         "q47norem_00"
[145] "q48_yrsm_00"         "q49_trie_00"         "q50a_age_00"
[148] "q50b_day_00"         "q50b_mon_00"         "q50b_wee_00"
[151] "q50b_yea_00"         "q51_time_00"         "q52_walk_00"
[154] "q52wkmin_00"         "q53_time_00"         "q54_hour_00"
[157] "q54_minu_00"         "q55_time_00"         "q56_hour_00"
[160] "q56_minu_00"         "q57_time_00"         "q58_hour_00"
[163] "q58_minu_00"         "q59_time_00"         "q60_hour_00"
[166] "q60_minu_00"         "q60b_typ_00"         "q61_monf_00"
[169] "q61mfmin_00"         "q61_sats_00"         "q61ssmin_00"
[172] "q62_inco_00"         "q63_inco_00"         "q64_join_00"
[175] "q65_job__00"         "q66_asco_00"         "q68_asco_00"
[178] "asco1_00"            "asco2_00"            "q67a_00"
[181] "q67b_00"             "q67c_00"             "q67d_00"
[184] "q67e_00"             "q67f_00"             "q67g_00"
[187] "q67h_00"             "q67i_00"             "q69_oral_00"
[190] "q70_cont_00"         "q71_taki_00"         "q72_hyst_00"
[193] "q73_age__00"         "q74_remo_00"         "q75_meno_00"
[196] "q76_meno_00"         "q77_estr_00"         "q78_taki_00"
[199] "q79_ever_00"         "q80_firs_00"         "q81_how__00"
[202] "q82_preg_00"         "q83_age__00"         "q84_test_00"
[205] "q85_gest_00"         "q86_gest_00"         "q87_inte_00"
[208] "q88_inte_00"         "q89_inte_00"         "tvtime_00"
[211] "exertime_00"         "sdte_00"             "fbg_00"
[214] "plg_00"              "creat_00"            "fibr_00"
[217] "chol_00"             "hdl_00"              "ldl_00"
[220] "ldl_hdl_00"          "trig_00"             "malb_00"
[223] "malbstr_00"          "ucr_00"              "macr_00"
[226] "macrstr_00"          "uprotein_00"         "uproteinstr_00"
[229] "ua_00"               "egfr00"              "ckd00"
[232] "selfreportdm_05"     "diabstat_05"         "drtrt_05"
[235] "type_05"             "drdiab_05"           "attendeestatus_05"
[238] "ehcq1aco_05"         "ehcq1b_05"           "ehcq1c_05"
[241] "ehcq1d_05"           "ehcq2a_05"           "ehcq2b_05"
[244] "ehcq2c_05"           "ehcq2d_05"           "ehcq2f_05"
[247] "cleaningcomments_05" "drgluclo_05"         "drhr_05"
[250] "sdte_05"             "drhrs_05"            "Gribcm_2"
[253] "Rg_cm_2"             "bp_cm_2"             "hw_cm_2"
[256] "Attendee_status_05"  "sex_05"              "agebysex_05"
[259] "dob"                 "gender"              "drage_05"
[262] "agegrp_05"           "drdate_05"           "Anthobid"
[265] "bmi_05"              "bmigrp_05"           "drheight_05"
[268] "drweight_05"         "drpreg_05"           "CDcluster_05"
[271] "drbiowt_05"          "drbioimp_05"         "drfatmas_05"
[274] "drlbm_05"            "drtbw_05"            "bpobid"
[277] "drpulse_05"          "drgenq_05"           "drhka_05"
[280] "drcompl_05"          "drsf36_05"           "drBWQ_05"
[283] "drDrugq_05"          "drEHQ_05l"           "drdiet_05"
[286] "drUrine_05"          "drecg_05"            "drurbloo_05"
[289] "drurleuk_05"         "drurnitr_05"         "drurprot_05"
[292] "drhip1_05"           "drhip2_05"           "drhip3_05"
[295] "hip_05"              "drwaist1_05"         "drwaist2_05"
[298] "drwaist3_05"         "waist_05"            "waistfe_05"
[301] "waistma_05"          "waistot_05"          "drsyst1_05"
[304] "drsyst2_05"          "drsyst3_05"          "drdiast1_05"
[307] "drdiast2_05"         "drdiast3_05"         "systgrp_05"
[310] "diastgrp_05"         "hypprev_05"          "hypertrt_05"
[313] "systolic_05"         "diastoli_05"         "fbg_05"
[316] "plg_05"              "gribble_05"          "fibr_05"
[319] "creatadj_05"         "ucr_05"              "ua_05"
```

```
[322] "malbstr_05"          "malb_05"             "macrstr_05"
[325] "macr_05"             "crpstr_05"           "crp_05"
[328] "uproteinstr_05"      "uprotein_05"         "egfr_05"
[331] "ckd_05"              "chol_05"             "hdl_05"
[334] "ldl_05"              "trig_05"             "cholprev_05"
[337] "hdlprev_05"          "ldlprev_05"          "trigprev_05"
[340] "lipidtot_05"         "comment_05"          "energy_00"
[343] "all_fat_00"          "satfat_00"           "death_st"
[346] "dod"                 "dead_10"             "fail_10"
[349] "dcod_07"             "fcod_07"             "CVD_D"
[352] "cod"                 "mcod1"               "mcod2"
[355] "mcod3"               "mcod4"               "mcod5"
[358] "mcod6"               "mcod7"               "mcod8"
[361] "mcod9"               "mcod10"              "mcod11"
[364] "mcod12"
```

```
> adet <-  read.dta( "ausdiab/Ethnicity.dta")
> c( dim(ad), dim(adet) )
```

```
[1] 11247   364 11247    12
```

```
> intersect( names(ad), names(adet) )
```

```
[1] "id"
```

```
> ad <- merge( ad, adet )
> dim( ad )
```

```
[1] 11247   375
```

The coding of `q5_schoo_00` is indeed a bit fishy, but the following tabulation shows that it is indeed the education variable as used below — the labelling has just been mixed up. “Overweight” (factor level 2) corresponds to “finished high school” (`hi`):

```
> table( ad$q5_schoo_00, exclude=NULL )
```

```
    Normal Overweight      Obese       <NA>
         0       5181       5997         69
```

So there is no reason to panic over the strange labeling of this variable below.

```
> dim(ad)
```

```
[1] 11247   375
```

```
> table( ad$q5_schoo_00 )
```

```
    Normal Overweight      Obese
         0       5181       5997
```

```
> str(ad[,c("drsex_00","drage_00","drweight_00","drheight_00","drdate_00","urbrural_00","q5_schoo_00
'data.frame':        11247 obs. of  7 variables:
 $ drsex_00   : Factor w/ 2 levels "Male","Female": 1 2 2 1 2 2 1 1 1 2 ...
 $ drage_00   : int  54 52 52 64 66 57 34 39 49 48 ...
 $ drweight_00: num  77.5 61.5 72.5 99.9 81.8 55.2 NA 84.5 70.7 60 ...
 $ drheight_00: num  176 167 164 187 178 ...
 $ drdate_00  : int  14374 14374 14374 14374 14374 14374 14374 14374 14374 14374 ...
 $ urbrural_00: Factor w/ 2 levels "capital city CD",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ q5_schoo_00: Factor w/ 3 levels "Normal","Overweight",..: 2 3 3 2 2 2 3 2 2 3 ...
```

```
> # ad[1:10,c("drsex_00","drage_00","drweight_00","drheight_00","drdate_00","urbrural_00","q5_schoo_
```

We then select only persons in the urban areas for the further analysis.

```
> with( ad, table(urbrural_00) )
```

```
urbrural_00
    capital city CD non-capital city CD
               6911                4336
```

```
> ad <- subset( ad, ad$urbrural_00=="capital city CD")
> dim( ad )
```

```
[1] 6911  375
```

```
> sex <- factor( as.integer( ad$drsex_00 ), labels=c("M","F") )
> table( ad$drsex_00, sex )
```

```
        sex
            M    F
  Male   3150    0
  Female    0 3761
```

```
> dos <- ad[,"drdate_00"]/365.25+1960
> age <- ad[,"drage_00"]+runif(nrow(ad),0,1)
> dob <- dos - age
> ht  <- ad$drheight_00/100
> wt  <- ad$drweight_00
> bmi <- wt/ht^2
> edu <- factor( as.integer(ad$q5_schoo_00)==2, labels=c("lo","hi") )
> table( ad$q5_schoo_00, edu, exclude=NULL )
```

```
            edu
               lo   hi <NA>
  Normal        0    0    0
  Overweight    0 3529    0
  Obese      3340    0    0
  <NA>          0    0   42
```

```
> smk <- with( ad, factor( smokstat_00=="Current Smoker", labels=c("non","cur") ) )
> table( ad$smokstat_00, smk, exclude=NULL )
```

```
                smk
                  non  cur <NA>
  Current Smoker    0  979    0
  Ex-smoker      1984    0    0
  Non-smoker     3819    0    0
  <NA>              0    0  129
```

```
> aus <- Relevel( ad$ethn6, list("aus"=1,"no"=2:6) )
> table( ad$ethn6, aus )
```

```
                        aus
                          aus   no
  Aust/NZ                4870    0
  Other English speaking    0  999
  South Europe              0  295
  Other Europe              0  252
  Asia                      0  417
  Other                     0   78
```

There seems to be no variable of the name ethn6 in the dataset:

```
> names(ad)[grep( "e", names(ad) )]
```

```
  [1] "drsex_00"            "drdate_00"           "drage_00"
  [4] "agegrp_00"           "drpreg_00"           "drurine_00"
  [7] "drheight_00"         "drweight_00"         "bioweigh_00"
 [10] "drimped_00"          "drfatper_00"         "drpulse_00"
 [13] "drecg_00"            "drgenq_00"           "drgenqob_00"
 [16] "drdietq_00"          "cd_area_00"          "cdcluste_00"
 [19] "state_00"            "waistfe_00"          "site_no_00"
 [22] "agebysex_00"         "type_00"             "obese_00"
 [25] "overweig_00"         "absagegp_00"         "exercise_00"
 [28] "wtonsite_05"         "wtehc_05"            "dr_genqu_00"
 [31] "q1_gende_00"         "q2_date__00"         "q7_highe_00"
 [34] "q8_educa_00"         "q24_lose_00"         "q30_heav_00"
 [37] "q43_pipe_00"         "q46_week_00"         "q47_age__00"
 [40] "q47norem_00"         "q49_trie_00"         "q50a_age_00"
 [43] "q50b_wee_00"         "q50b_yea_00"         "q51_time_00"
 [46] "q53_time_00"         "q55_time_00"         "q57_time_00"
 [49] "q59_time_00"         "q67e_00"             "q73_age__00"
 [52] "q74_remo_00"         "q75_meno_00"         "q76_meno_00"
 [55] "q77_estr_00"         "q79_ever_00"         "q82_preg_00"
 [58] "q83_age__00"         "q84_test_00"         "q85_gest_00"
 [61] "q86_gest_00"         "q87_inte_00"         "q88_inte_00"
 [64] "q89_inte_00"         "tvtime_00"           "exertime_00"
 [67] "sdte_00"             "creat_00"            "uprotein_00"
 [70] "uproteinstr_00"      "egfr00"              "selfreportdm_05"
 [73] "type_05"             "attendeestatus_05"   "ehcq1aco_05"
 [76] "ehcq1b_05"           "ehcq1c_05"           "ehcq1d_05"
 [79] "ehcq2a_05"           "ehcq2b_05"           "ehcq2c_05"
 [82] "ehcq2d_05"           "ehcq2f_05"           "cleaningcomments_05"
 [85] "sdte_05"             "Attendee_status_05"  "sex_05"
 [88] "agebysex_05"         "gender"              "drage_05"
 [91] "agegrp_05"           "drdate_05"           "drheight_05"
 [94] "drweight_05"         "drpreg_05"           "CDcluster_05"
 [97] "drpulse_05"          "drgenq_05"           "drdiet_05"
[100] "drUrine_05"          "drecg_05"            "drurleuk_05"
[103] "waistfe_05"          "hypprev_05"          "hypertrt_05"
[106] "gribble_05"          "creatadj_05"         "uproteinstr_05"
[109] "uprotein_05"         "egfr_05"             "cholprev_05"
[112] "hdlprev_05"          "ldlprev_05"          "trigprev_05"
[115] "comment_05"          "energy_00"           "death_st"
[118] "dead_10"             "ethn6"               "europe"
[121] "cdcluste"            "state"
```

```
> names(ad)[grep( "b", names(ad) )]

 [1] "drdob_00"    "drdiab_00"   "diabstat_00" "drwhobsi_00" "drbioobs_00" "bioweigh_00"
 [7] "drlbm_00"    "drtbw_00"    "drbpobsi_00" "drgenqob_00" "drhkaobs_00" "choltabl_00"
[13] "bmi_00"      "bmigrp_00"   "agebysex_00" "obese_00"    "absagegp_00" "absabys_00"
[19] "urbrural_00" "q21_bloo_00" "q22_bloo_00" "q23_tabl_00" "q28_tabl_00" "q34_numb_00"
[25] "q37_bad__00" "q50b_day_00" "q50b_mon_00" "q50b_wee_00" "q50b_yea_00" "q60b_typ_00"
[31] "q65_job__00" "q67b_00"     "fbg_00"      "fibr_00"     "malb_00"     "malbstr_00"
[37] "diabstat_05" "drdiab_05"   "ehcq1b_05"   "ehcq2b_05"   "Gribcm_2"    "bp_cm_2"
[43] "agebysex_05" "dob"         "Anthobid"    "bmi_05"      "bmigrp_05"   "drbiowt_05"
[49] "drbioimp_05" "drlbm_05"    "drtbw_05"    "bpobid"      "drurbloo_05" "fbg_05"
[55] "gribble_05"  "fibr_05"     "malbstr_05"  "malb_05"     "cob_gp2"     "cob_gp4"
[61] "cob_gp5"

> # aus <- factor( ad$ethn6=="Australia", labels=c("no","aus") )
> # table( ad$ethn6, aus, exclude=NULL )
>
> ad  <- data.frame( sex, dob, dos, bmi, ht, wt, edu, smk, aus, srv="AusDiab" )
> head( ad )

  sex      dob      dos      bmi    ht   wt edu smk aus     srv
1   M 1944.646 1999.354 25.01937 1.760 77.5  hi non  no AusDiab
2   F 1946.379 1999.354 22.05170 1.670 61.5  lo non aus AusDiab
3   F 1946.432 1999.354 26.79207 1.645 72.5  lo non aus AusDiab
4   M 1934.874 1999.354 28.56816 1.870 99.9  hi non aus AusDiab
5   F 1932.413 1999.354 25.81745 1.780 81.8  hi non aus AusDiab
6   F 1941.967 1999.354 21.29547 1.610 55.2  hi non  no AusDiab
```

The variable `q5_schoo_00` looks fishy in its coding for using it as an education indicator:

```
> with( ad, tapply( bmi, edu, quantile, probs=0:5/5, na.rm=TRUE ) )

$lo
      0%      20%      40%      60%      80%     100%
15.23118 23.17740 25.51889 27.80099 30.86577 55.91479

$hi
      0%      20%      40%      60%      80%     100%
14.96094 22.41399 24.63795 26.64001 29.35547 59.39646
```

2.4 Merging all surveys

Once we have read in all the surveys we can stack them into one dataset:

```
> ABMI <- rbind( d1, d2, d3, n95, ad, n8 )
> summary( ABMI )

 sex            dob             dos             bmi                   ht
 M:24654   Min.   :1904   Min.   :1980   Min.   :     0.1   Min.   :0.000
 F:26784   1st Qu.:1937   1st Qu.:1984   1st Qu.:    21.8   1st Qu.:1.591
           Median :1950   Median :1996   Median :    24.7   Median :1.672
           Mean   :1953   Mean   :1995   Mean   :   304.9   Mean   :2.062
           3rd Qu.:1965   3rd Qu.:2008   3rd Qu.:    27.9   3rd Qu.:1.760
           Max.   :2008   Max.   :2008   Max.   :860000.0   Max.   :10.000
           NA's   :5      NA's   :6      NA's   :3751       NA's   :134
       wt           edu           smk           aus            srv
 Min.   :  0.0   lo  :26861   non :40350   no  :14453   RFPS1980: 5603
 1st Qu.: 59.3   hi  :24534   cur :10946   aus :36984   RFPS1983: 7615
 Median : 71.0   NA's:   43   NA's:  142   NA's:    1   RFPS1989: 9279
 Mean   :124.2                                          NNS1995 : 7993
 3rd Qu.: 84.0                                          AusDiab : 6911
 Max.   :999.9                                          NHS2008 :14037
 NA's   :203
```

```
> abmi <- ABMI[complete.cases(ABMI[,1:6]),]
> abmi$srv <- factor( abmi$srv )
> summary( abmi )
```

```
 sex            dob            dos            bmi                ht
 M:22866   Min.   :1904   Min.   :1980   Min.   :     0.1   Min.   :0.010
 F:24813   1st Qu.:1936   1st Qu.:1984   1st Qu.:    21.8   1st Qu.:1.610
           Median :1949   Median :1995   Median :    24.7   Median :1.685
           Mean   :1949   Mean   :1994   Mean   :   304.9   Mean   :2.215
           3rd Qu.:1961   3rd Qu.:2001   3rd Qu.:    27.9   3rd Qu.:1.770
           Max.   :1994   Max.   :2008   Max.   :860000.0   Max.   :10.000
       wt          edu          smk          aus              srv
 Min.   :  1.00   lo  :23326   non :36726   no  :14041   RFPS1980: 5600
 1st Qu.: 62.00   hi  :24327   cur :10828   aus :33637   RFPS1983: 7611
 Median : 72.90   NA's:   26   NA's:  125   NA's:    1   RFPS1989: 9172
 Mean   :133.45                                          NNS1995 : 7851
 3rd Qu.: 85.25                                          AusDiab : 6803
 Max.   :999.90                                          NHS2008 :10642
```

2.4.1 Exclusions

We exclude persons under age 20 and outside the height and weight ranges:

```
> hw.ok <- with( abmi, ht > h.lo/100 & wt > w.lo &
+                      ht < h.hi/100 & wt < w.hi )
> a.ok  <- with( abmi, dos-dob > 18 )
> with( abmi, ftable( addmargins( table( srv, hw.ok, a.ok, exclude=NULL ),
+                                 margin=1 ),
+                     row.vars=1 ) )
```

```
         hw.ok FALSE             TRUE                NA
         a.ok  FALSE  TRUE    NA FALSE  TRUE    NA FALSE  TRUE    NA
srv
RFPS1980           0    25     0     0  5575     0     0     0     0
RFPS1983           0    39     0     0  7572     0     0     0     0
RFPS1989           0    36     0     0  9136     0     0     0     0
NNS1995         1013   117     0   531  6190     0     0     0     0
AusDiab            0    43     0     0  6760     0     0     0     0
NHS2008            0  3238     0     0  7404     0     0     0     0
NA                 0     0     0     0     0     0     0     0     0
Sum             1013  3498     0   531 42637     0     0     0     0
```

```
> abmi <- subset( abmi, hw.ok & a.ok )
> str( abmi )
```

```
'data.frame':        42637 obs. of  10 variables:
 $ sex: Factor w/ 2 levels "M","F": 1 2 1 1 2 2 2 1 1 2 ...
 $ dob: num  1926 1935 1938 1955 1926 ...
 $ dos: num  1980 1980 1980 1981 1980 ...
 $ bmi: num  31.7 24.2 25.2 24.5 32.1 ...
 $ ht : num  1.73 1.7 1.78 1.85 1.72 1.6 1.53 1.73 1.76 1.61 ...
 $ wt : num  95 70 80 84 95 61 50 66 86 56 ...
 $ edu: Factor w/ 2 levels "lo","hi": 2 2 2 2 1 2 1 2 2 2 ...
 $ smk: Factor w/ 2 levels "non","cur": 2 2 1 2 1 1 1 1 1 1 ...
 $ aus: Factor w/ 2 levels "no","aus": 1 2 2 2 2 2 2 2 2 2 ...
 $ srv: Factor w/ 6 levels "RFPS1980","RFPS1983",..: 1 1 1 1 1 1 1 1 1 1 ...
```

We then plot BMI as function of age, for a 10% random sample of the entire data set, color-coded by survey:

```
> par( mar=c(3,3,1,1), mgp=c(3,1,0)/1.6 )
> with( subset(abmi,runif(nrow(abmi))<0.10),
+       plot( dos-dob, bmi, ylim=c(10,55),
+             pch=16, cex=0.5, col=rainbow(6)[srv],
+             xlab="Age at survey", ylab="BMI at survey" ) )
> abline( h=c(15,50), col=gray(0.8) )
```

From the figures it seems that something is fishy about some of the BMIs; we have some very high and very low BMI-values:

```
> subset(abmi, bmi<15 )
```

```
      sex      dob      dos      bmi     ht   wt edu smk aus     srv
2656    F 1945.357 1980.612 14.87879 1.6600 41.0  lo non aus RFPS1980
9378    M 1949.228 1983.369 14.73930 1.7800 46.7  hi cur aus RFPS1983
24042   F 1936.078 1995.712 14.86477 1.6445 40.2  lo non aus  NNS1995
```

```
> subset(abmi, bmi>50 )
```

```
      sex      dob      dos      bmi     ht    wt edu smk aus      srv
252     M 1929.447 1980.634 57.84627 1.4700 125.0  lo non  no RFPS1980
4095    F 1943.038 1980.366 50.67825 1.6200 133.0  lo non aus RFPS1980
10001   F 1938.085 1983.895 50.19531 1.6000 128.5  lo cur aus RFPS1983
10359   F 1934.876 1983.388 51.04848 1.6400 137.3  lo non aus RFPS1983
17766   F 1930.427 1989.666 53.55186 1.4800 117.3  hi non aus RFPS1989
20084   F 1921.622 1989.617 53.47701 1.5800 133.5  lo non aus RFPS1989
22875   F 1945.744 1996.028 50.04084 1.5550 121.0  hi non  no  NNS1995
23241   F 1952.791 1995.749 52.53404 1.5755 130.4  hi non aus  NNS1995
27297   F 1968.439 1995.259 51.99247 1.6280 137.8  lo non aus  NNS1995
29174   F 1948.870 1996.130 52.57433 1.6260 139.0  lo non aus  NNS1995
29707   F 1936.597 1995.832 52.27809 1.5745 129.6  lo non aus  NNS1995
35348   F 1964.856 2000.512 55.91479 1.5710 138.0  lo non aus  AusDiab
36702   F 1971.639 2000.704 50.37536 1.6340 134.5  lo cur aus  AusDiab
38431   F 1962.998 2007.711 51.05601 1.6500 139.0  lo cur aus  NHS2008
44608   F 1974.534 2008.399 50.78125 1.6000 130.0  lo cur aus  NHS2008
47552   F 1988.146 2007.633 53.41880 1.5600 130.0  hi non aus  NHS2008
```

However, these only constitute a very small fraction of the data, so we will exclude them in order to avoid undue influence on the results:

```
> abmi <- subset( abmi, bmi>15 & bmi<50 )
> with( abmi, addmargins( table( srv, sex ) ) )
```

```
          sex
srv            M     F   Sum
  RFPS1980  2761  2811  5572
  RFPS1983  3734  3835  7569
  RFPS1989  4494  4640  9134
  NNS1995   2962  3222  6184
  AusDiab   3110  3648  6758
  NHS2008   3554  3847  7401
  Sum      20615 22003 42618
```

To get an overview of how data is distributed by age and date we plot the age and date of survey for the resulting persons:

```
> par( mar=c(3,3,1,1), mgp=c(3,1,0)/1.6 )
> with( subset(abmi,runif(nrow(abmi))<0.05),
+       plot( dos, dos-dob,
+             pch=16, cex=0.4, col=rainbow(6)[srv],
+             xlab="Date of survey", ylab="Age at survey" ) )
```

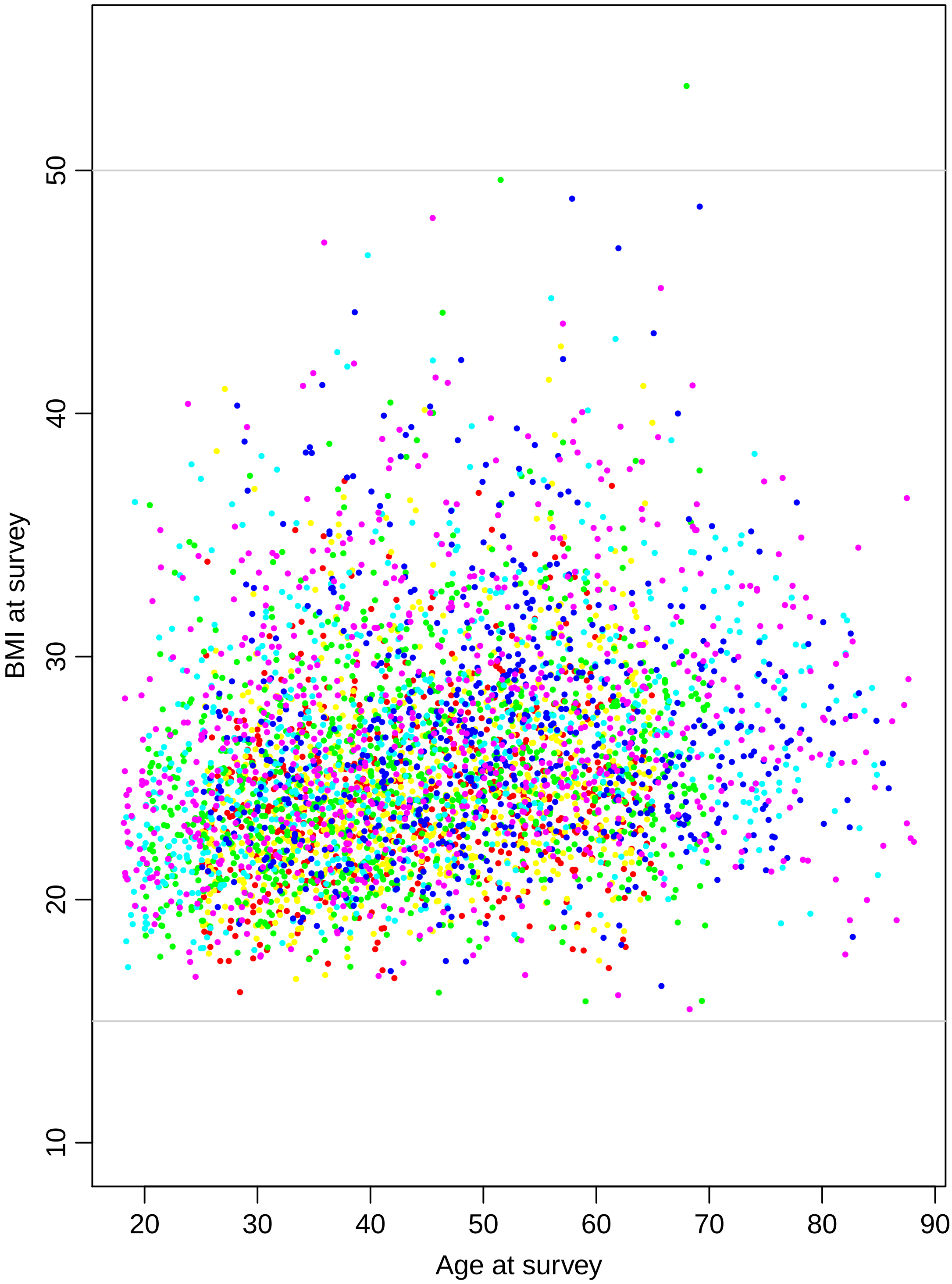

Figure 2.1: *BMI and age at survey as imputed for the 6 surveys, based on a 5% random sample of the final database. Colouring by survey (a proxy for date).*

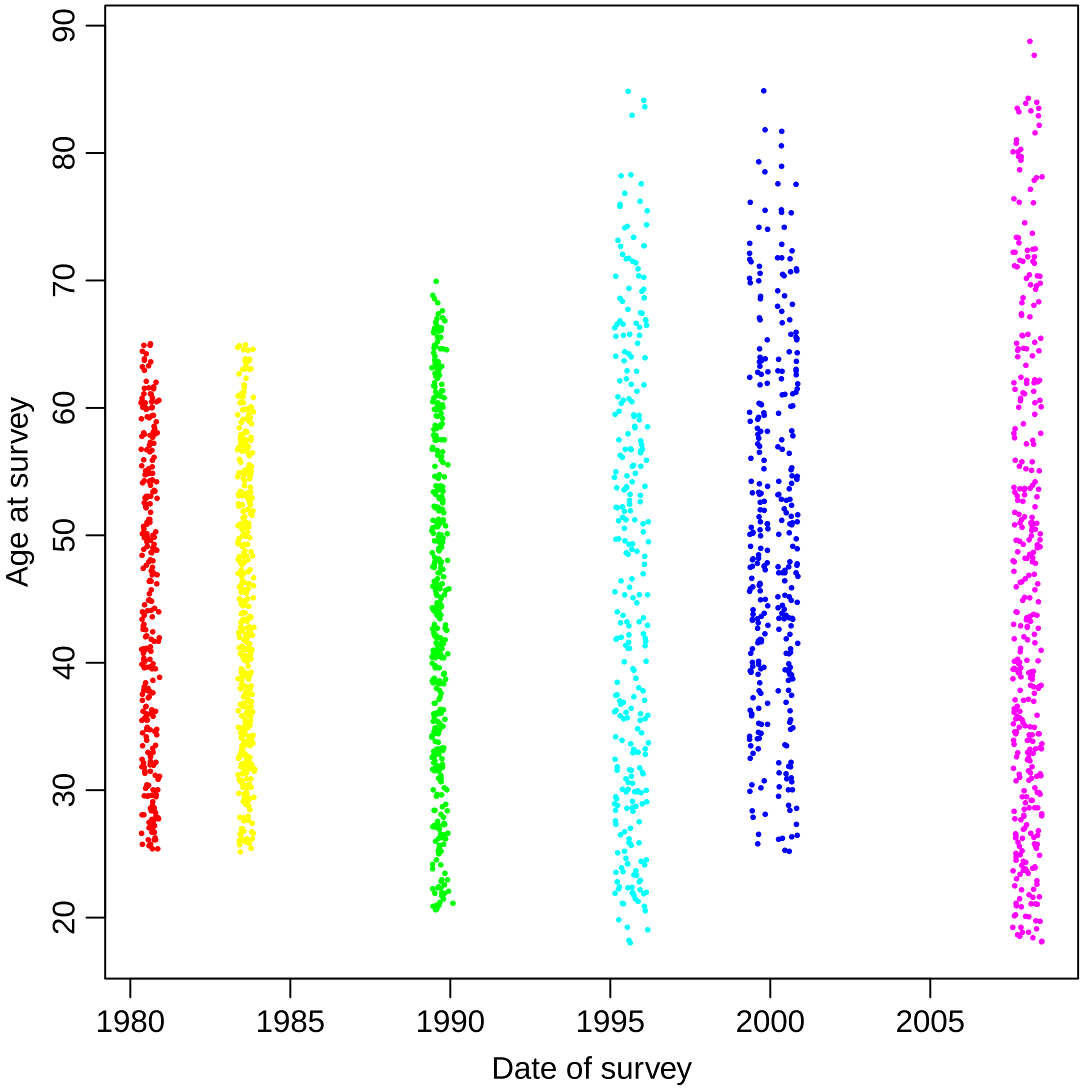

Figure 2.2: *Age and date of survey as imputed for the 6 surveys, based on a 5% random sample of the final database.*

Here is a version for the paper (the argument mm is the margin multiplier - margins have a total width of 8/mm inches bith in horizontal and vertical direction):

```
> temp.graph <-
+ function(mm)
+ {
+ par( mai=c(7,7,1,1)/mm, mgp=c(3,1,0)/1.6, las=1 )
+ plot( NA, bty="n",
+       xaxs="i", xlim=tlim, xlab="Date of survey",
+       yaxs="i", ylim=alim, ylab="Age at survey" )
+ for( da in -3:3*10 )
+ polygon( tlim[c(1,2,2,1,1)],
+          30+c(1,diff(tlim)+c(1,-1),-1,1)+da, col=gray(0.9), border="transparent" )
+ with( subset(abmi,runif(nrow(abmi))<0.05),
+       points( dos, dos-dob, pch=16, cex=0.4 ) )
+ axis( side=1 )
+ axis( side=2, at=seq(10,90,5), labels=FALSE )
+ axis( side=2, at=seq(10,90,20) )
+ }
> tlim <- c(1979,2011)
> alim <- c(15,96)
> ypi  <- 10
> mm   <- 12
> pdf("./graph/BMI-APC-Lexis.pdf", height=8/mm+diff(alim)/ypi,
+                                   width=8/mm+diff(tlim)/ypi ) ; temp.graph(mm) ; dev.off()


null device
          1
```

Here is an annotated version of the same graph:

```
> temp.graph <-
+ function(mm)
+ {
+ par( mai=c(7,7,1,5)/mm, mgp=c(3,1,0)/1.6, las=1 )
+ plot( NA, bty="n", yaxt="n",
+       xaxs="i", xlim=tlim, xlab="Date of survey",
+       yaxs="i", ylim=alim, ylab="Age at survey" )
+ axis( side=1 )
+ axis( side=2, at=seq(10,90,5), labels=FALSE )
+ axis( side=2, at=seq(10,90,10) )
+ for( bc in bcoh )
+ polygon( tlim[c(1,2,2,1,1)],
+          tlim[c(1,2,2,1,1)] - bc + c(0,0,-1,-1,0),
+          col=gray(0.85), border="transparent" )
+ axis( side=4, at=tlim[2]-bcoh-0.5,
+       labels=bcoh, mgp=c(3,0,0)/2,
+       col.axis=gray(0.80), lty=0, font.axis=2 )
+ with( subset(abmi,runif(nrow(abmi))<0.05),
+       points( dos, dos-dob, pch=16, cex=0.4 ) )
+ with( abmi, text( tt <- tapply( dos, srv, mean ), rep(90,6),
+                   names(tt), srt=90, cex=0.9, adj=0 ) )
+ }
> tlim <- c(1979,2011)
> alim <- c(15,100)
> bcoh <- 1920+0:6*10
> ypi  <- 10
> mm   <- 12
> pdf("./graph/BMI-APC-Lexis-ann.pdf", height=8/mm+diff(alim)/ypi,
+                                       width=12/mm+diff(tlim)/ypi ) ; temp.graph(mm) ; dev.off()


null device
          1
```

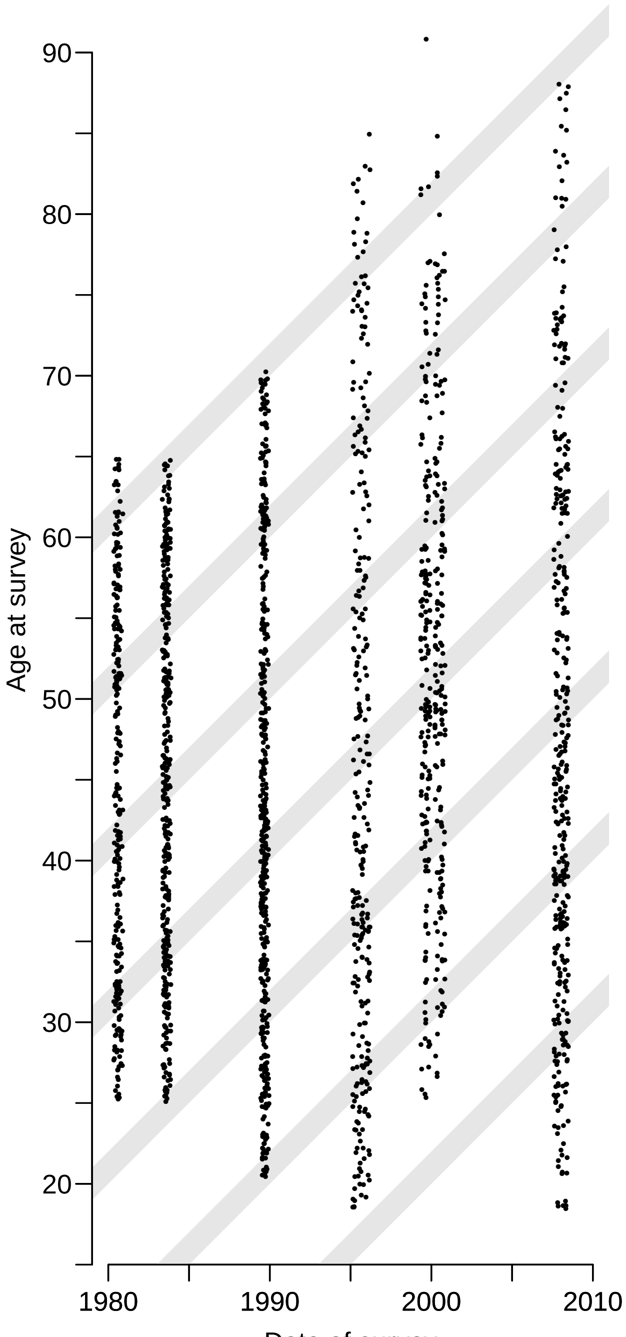

Figure 2.3: *Imputed ages and dates of survey for a 5% sample of the participants. The gray areas show birth-cohorts (top to bottom) 1920,1930,. . . ,1980.*

```
> win.metafile(        "art/fig1.emf", height=8/mm+diff(alim)/ypi,
+                                      width=12/mm+diff(tlim)/ypi ) ; temp.graph(mm) ; dev.off()
```

```
null device
          1
```

```
>   postscript(        "art/fig1.eps", height=8/mm+diff(alim)/ypi,
+                                      width=12/mm+diff(tlim)/ypi ) ; temp.graph(mm) ; dev.off()
```

```
null device
          1
```

```
>          pdf(        "art/fig1.pdf", height=8/mm+diff(alim)/ypi,
+                                      width=12/mm+diff(tlim)/ypi ) ; temp.graph(mm) ; dev.off()
```

```
null device
          1
```

Finally we save the dataset for further analysis

```
> with( abmi, addmargins( table( srv, sex, exclude=NULL ) ) )
```

```
          sex
srv            M     F  <NA>   Sum
  RFPS1980  2761  2811     0  5572
  RFPS1983  3734  3835     0  7569
  RFPS1989  4494  4640     0  9134
  NNS1995   2962  3222     0  6184
  AusDiab   3110  3648     0  6758
  NHS2008   3554  3847     0  7401
  <NA>         0     0     0     0
  Sum      20615 22003     0 42618
```

```
> save( abmi, file="./data/abmi.Rdata" )
```

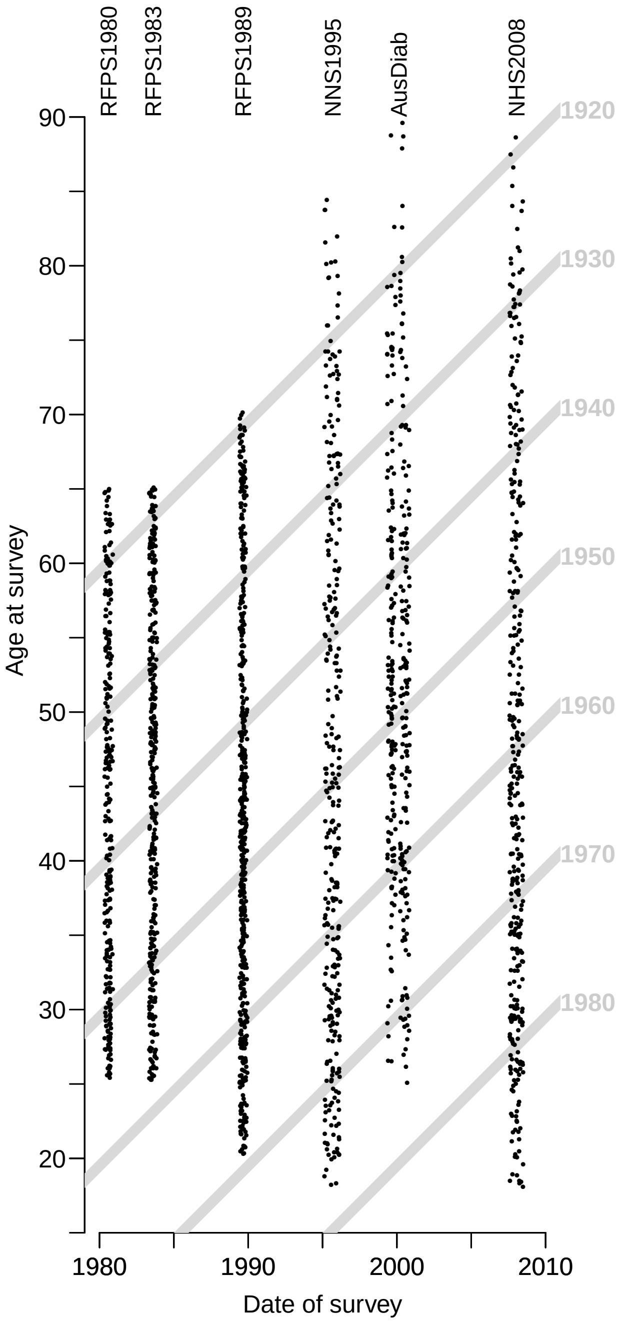

Figure 2.4: *Imputed ages and dates of survey for a 5% sample of the participants. The gray areas show birth-cohorts (top to bottom) 1920,1930,. . . ,1980.*

Chapter 3

Simple regression models for BMI

```
> library( Epi )
> library( splines )
> library( quantreg )
> print( sessionInfo(), l=F )
```

```
R version 3.0.2 (2013-09-25)
Platform: i386-w64-mingw32/i386 (32-bit)

attached base packages:
[1] splines   utils     datasets  graphics  grDevices stats     methods
[8] base

other attached packages:
[1] quantreg_4.98  SparseM_0.99   Epi_1.1.59     foreign_0.8-55

loaded via a namespace (and not attached):
[1] tools_3.0.2
```

```
> options( width=100 )
> load( file="./data/abmi.Rdata" )
> abmi$age <- abmi$dos - abmi$dob
> head( abmi )
```

```
  sex      dob      dos      bmi   ht wt edu smk aus      srv      age
1   M 1925.992 1980.368 31.74179 1.73 95  hi cur  no RFPS1980 54.37645
2   F 1935.027 1980.382 24.22145 1.70 70  hi cur aus RFPS1980 45.35524
3   M 1938.175 1980.385 25.24934 1.78 80  hi non aus RFPS1980 42.20945
4   M 1955.328 1980.502 24.54346 1.85 84  hi cur aus RFPS1980 25.17454
5   F 1925.795 1980.346 32.11195 1.72 95  lo non aus RFPS1980 54.55168
6   F 1946.315 1980.382 23.82812 1.60 61  hi non aus RFPS1980 34.06708
```

We start out by fitting a normal linear regression model to data, but first we find suitable knots for age and date of birth, and the fit separate models for males and females.

First we define a convenience function that easier allows us specify natural splines:

```
> source("c:/stat/r/bxc/library.sources/useful/R/Ns.R" )
> Ns
```

```
function (x, df = NULL, knots = NULL, intercept = FALSE, Boundary.knots = NULL)
{
    if (is.null(Boundary.knots)) {
        if (!is.null(knots)) {
            knots <- sort(unique(knots))
            ok <- c(1, length(knots))
            Boundary.knots <- knots[ok]
            knots <- knots[-ok]
        }
    }
    ns(x, df = df, knots = knots, intercept = intercept, Boundary.knots = Boundary.knots)
}
```

as well as a function that allows quicker specification of the arrays we need to store the various results:

```
> Array <- function( ll ) array( NA, dimnames=ll, dim=sapply(ll,length) )
```

```
> a.kn <- with( abmi, quantile( dos-dob, probs=1:9/10 ) )
> b.kn <- with( abmi, quantile(     dob, probs=1:9/10 ) )
> m1M <- lm( bmi ~ Ns( dos-dob, kn=a.kn, intercept=TRUE ) - 1 +
+                  Ns( dob,     kn=b.kn ),
+             data=subset(abmi,sex=="M") )
> m1F <- update( m1M, data=subset(abmi,sex=="F") )
```

We also fit the same model for log-transformed data (using natural log):

```
> l1M <- update( m1M, log(bmi) ~ . )
> l1F <- update( m1F, log(bmi) ~ . )
```

Just to get a feeling for how the models fits we do a residual plot:

```
> res4 <- function( m1, col="black" )
+ {
+ n1 <- length(residuals(m1))
+ abmi <- eval(as.list(m1$call)$data) # This is how to retrieve the dataframe used
+ par( mfrow=c(2,2), mar=c(3,3,1,1), mgp=c(3,1,0)/1.6 )
+ with( abmi, plot( dos-dob, residuals(m1), pch=".", col=col,
+                    xlab="Age at survey", ylab="Residuals" ) )
+ with( abmi, plot( dob    , residuals(m1), pch=".", col=col,
+                    xlab="Date of birth", ylab="Residuals" ) )
+ hist( residuals(m1), breaks=100, col=col, border=col, main="", xlab="Residuals" )
+ plot( pnorm( sort( residuals(m1)/sd(residuals(m1)) ) ), (1:n1-0.5)/n1,
+         pch=".", col=col, ylab="Uniform", xlab=expression(Phi^-1*"[sort(std. res.)]") )
+ abline(0,1)
+ }
> res4( m1M, col="blue" )
```

```
> res4( m1F, col="red" )
```

We also see if the log-transformed response produces a better fit:

```
> res4( l1M, col="blue" )
```

```
> res4( l1F, col="red" )
```

Looking at figures 3.1, 3.2, 3.3 and 3.4 there is an indication that the normality assumption, and in particular the symmetry assumption is much better met for the log-transformed data.

But for the sake of completeness we shall also report the results from the traditional linear model.

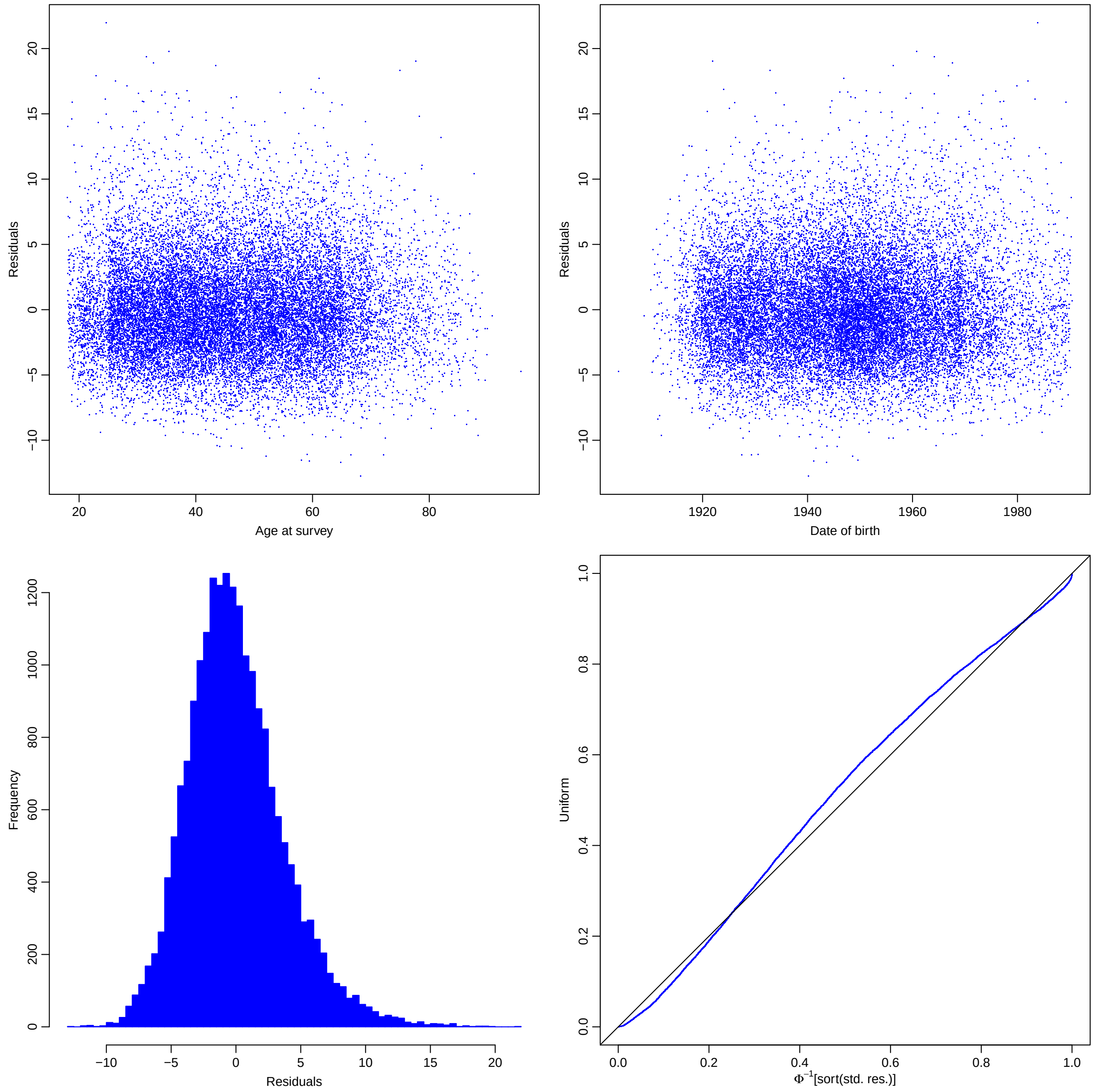

Figure 3.1: *Residual plot for an age-cohort model for untransformed BMI-response for men.*

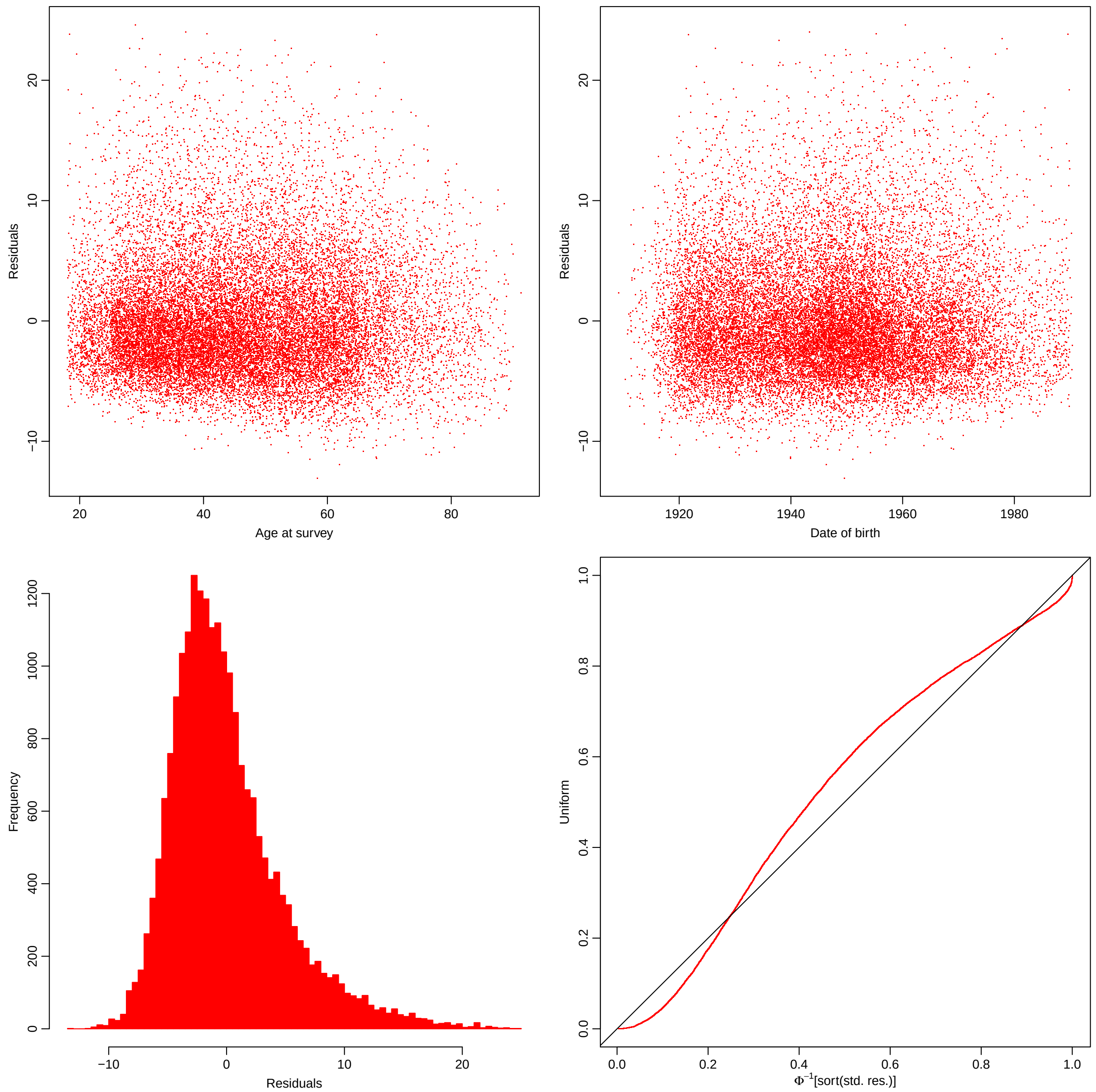

Figure 3.2: *Residual plot for an age-cohort model for untransformed BMI-response for women.*

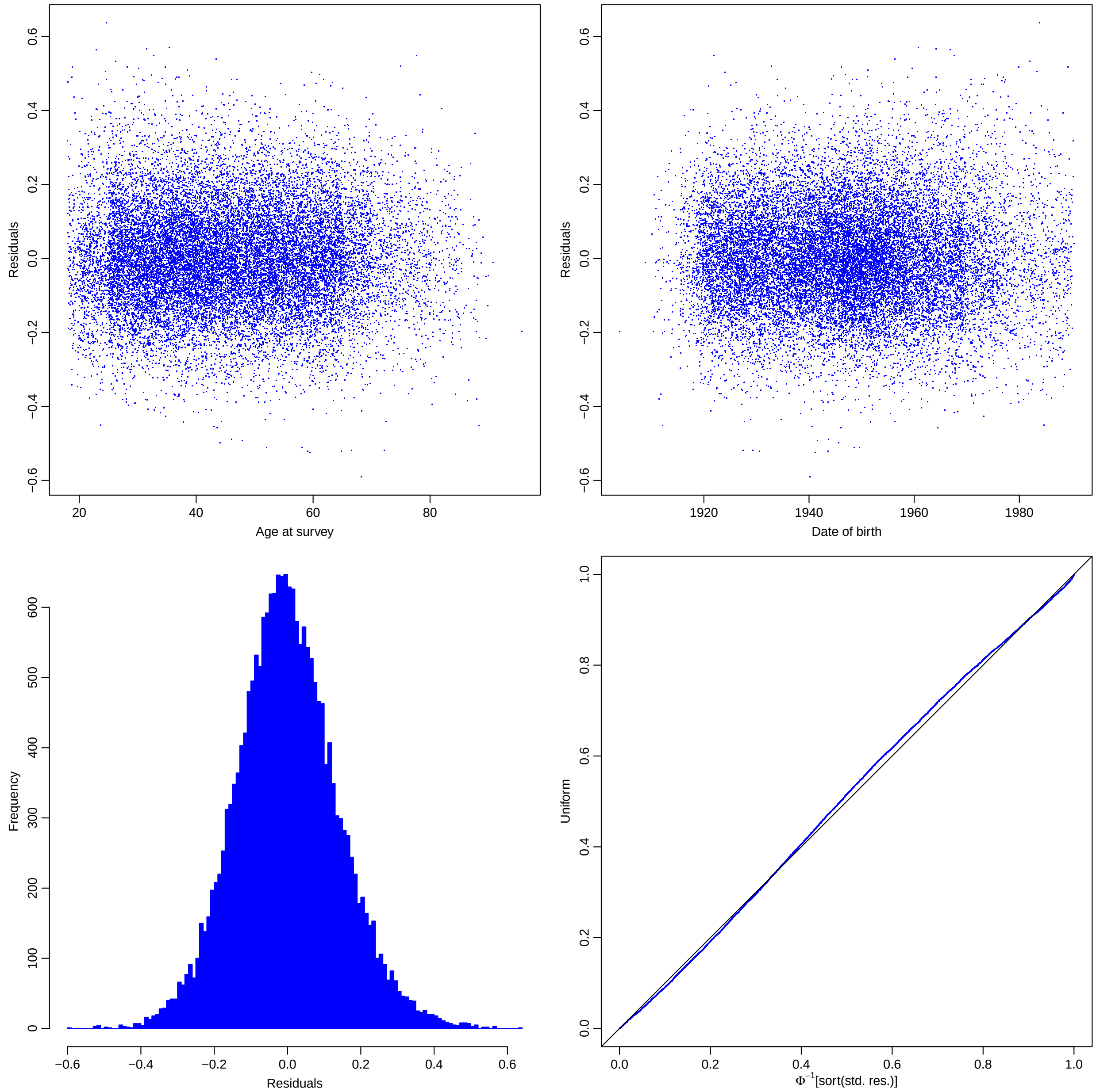

Figure 3.3: *Residual plot for an age-cohort model for log-transformed BMI-response for men.*

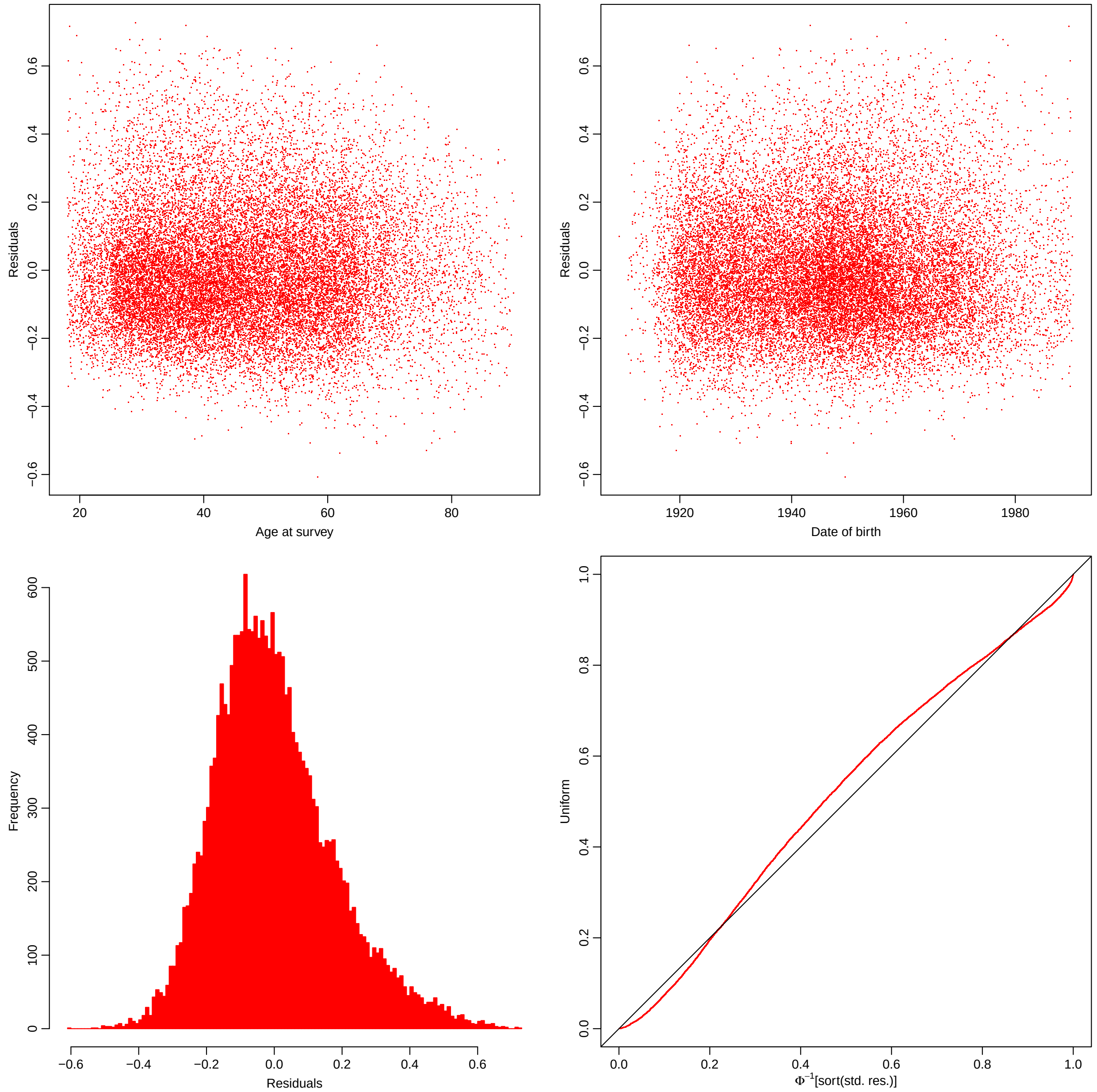

Figure 3.4: *Residual plot for an age-cohort model for log-transformed BMI-response for women.*

3.1 Estimates for BMI by age and date of birth

In order to extract estimates we need prediction points for and and date of birth, as well as a decision on the reference point on the date-of-birth scale, as well as matrices to multiply with the estimates from the spline models:

```
> a.pt <- seq(10,90,,200)
> b.pt <- seq(1915,1985,,200)
> b.ref <- 1950
> Ca <- Ns( a.pt, kn=a.kn, intercept=TRUE )
> Cb <- Ns( b.pt, kn=b.kn )
> Cb.ref <- Ns( rep(b.ref,length(a.pt)), kn=b.kn )
```

3.1.1 Linear model

The linear model does not fit so well as the model for log-transformed data, and the results we will get are estimates of the age-specific BMIs for a given birth cohort in casu the 1950 birth cohort.

```
> par( mfrow=c(1,2), mar=c(3,3,1,1), mgp=c(3,1,0)/1.6 )
> a.linM <- ci.lin( m1M, ctr.mat=cbind(Ca,Cb.ref) )[,c(1,5,6)]
> b.linM <- ci.lin( m1M, subset="b.kn", ctr.mat=Cb-Cb.ref )[,c(1,5,6)]
> a.linF <- ci.lin( m1F, ctr.mat=cbind(Ca,Cb.ref) )[,c(1,5,6)]
> b.linF <- ci.lin( m1F, subset="b.kn", ctr.mat=Cb-Cb.ref )[,c(1,5,6)]
> matplot( a.pt, cbind(a.linM,a.linF), lwd=c(3,1,1), type="l", lty=1,
+          col=rep(c("blue","red"),each=3),
+          xlab="Age", ylab=paste("BMI in", b.ref, "cohort"),
+          ylim=c(15,30) )
> matplot( b.pt, cbind(b.linM,b.linF), lwd=c(3,1,1), type="l", lty=1,
+          col=rep(c("blue","red"),each=3),
+          xlab="Date of birth", ylab=paste("BMI difference from", b.ref, "cohort"),
+          ylim=7.5*c(-1,1) )
```

From figure 3.5 it seems reasonably safe to assume that the increase in BMI has been pretty constant from generation to generation, a test for linearity is obtained by comparing a model where the spline effect of date of birth is replaced by a linear term:

```
> m1lM <- update( m1M, . ~ . - Ns( dob, kn=b.kn ) + I((dob-b.ref)/10) )
> m1lF <- update( m1F, . ~ . - Ns( dob, kn=b.kn ) + I((dob-b.ref)/10) )
> anova( m1lM, m1M )

Analysis of Variance Table

Model 1: bmi ~ Ns(dos - dob, kn = a.kn, intercept = TRUE) + I((dob - b.ref)/10) -
    1
Model 2: bmi ~ Ns(dos - dob, kn = a.kn, intercept = TRUE) - 1 + Ns(dob,
    kn = b.kn)
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1  20605 288773
2  20598 288451  7    322.14 3.2862 0.001709

> anova( m1lF, m1F )

Analysis of Variance Table

Model 1: bmi ~ Ns(dos - dob, kn = a.kn, intercept = TRUE) + I((dob - b.ref)/10) -
    1
Model 2: bmi ~ Ns(dos - dob, kn = a.kn, intercept = TRUE) - 1 + Ns(dob,
    kn = b.kn)
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1  21993 505207
2  21986 504776  7    431.22 2.6832 0.008918
```

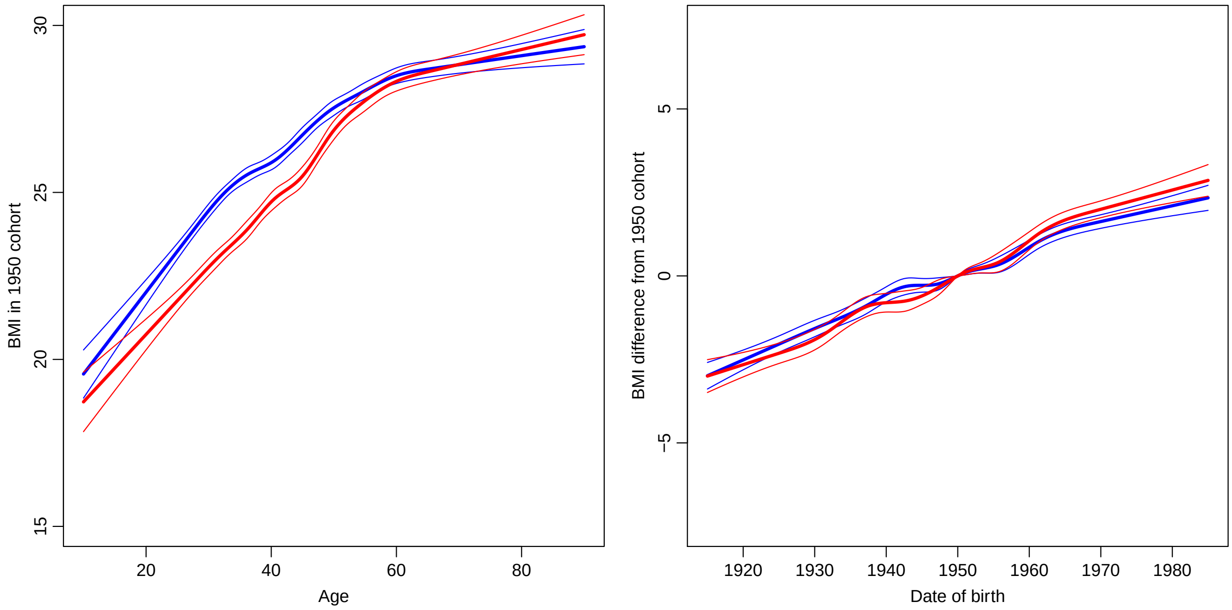

Figure 3.5: *Estimates of mean BMI in the Australian population in the 1950 birth cohort, and the change in mean BMI over cohorts. Blue curves are for men, red curves for women.*

```
> lchg <-
+ rbind( ci.lin( m1lM, subset="ref"),
+        ci.lin( m1lF, subset="ref") )[,c(1,5,6)]
> rownames(lchg) <- c("M","F")
> lchg
```

```
   Estimate      2.5%     97.5%
M 0.7837478 0.7267196 0.8407759
F 0.9184447 0.8474640 0.9894253
```

thus there is *formally* a non-linear effect for men, but given the massive amount of data a p-value of 0.006 is not particularly convincing. The estimated mean increase in BMI by birth cohort is 0.87 kg/m^2 per 10 years (0.22-0.92) for men and 1.08 kg/m^2 (1.01,1.14) for women.

3.1.2 Model for log-transformed data

The linear model does not fit so well as the model for log-transformed data, and the results we will get are estimates of the age-specific BMIs for a given birth cohort in casu the 1950 birth cohort.

```
> par( mfrow=c(1,2), mar=c(3,3,1,1), mgp=c(3,1,0)/1.6, las=1 )
> a.linM <- ci.exp( l1M, ctr.mat=cbind(Ca,Cb.ref) )
> b.linM <- ci.exp( l1M, subset="b.kn", ctr.mat=Cb-Cb.ref )
> a.linF <- ci.exp( l1F, ctr.mat=cbind(Ca,Cb.ref) )
> b.linF <- ci.exp( l1F, subset="b.kn", ctr.mat=Cb-Cb.ref )
> matplot( a.pt, cbind(a.linM,a.linF), lwd=c(3,1,1), type="l", lty=1,
+          col=rep(c("blue","red"),each=3),
+          xlab="Age", ylab=paste("BMI in", b.ref, "cohort"),
+          ylim=c(15,30) )
> matplot( b.pt, cbind(b.linM,b.linF), lwd=c(3,1,1), type="l", lty=1,
+          col=rep(c("blue","red"),each=3),log="y",
+          xlab="Date of birth", ylab=paste("Ratio of BMI relative to", b.ref, "cohort"),
+          ylim=c(1/1.4,1.4) )
> abline( h= 1 )
```

As for the linear model we also see a linear effect (that is a constant percent-wise increase from generation to generation) in the model for log-transformed data, see figure 3.6. As previously, the test for men is formally significant.

```
> l1lM <- update( l1M, . ~ . - Ns( dob, kn=b.kn ) + I((dob-b.ref)/10) )
> anova( l1lM, l1M )
```

```
Analysis of Variance Table

Model 1: log(bmi) ~ Ns(dos - dob, kn = a.kn, intercept = TRUE) + I((dob -
    b.ref)/10) - 1
Model 2: log(bmi) ~ Ns(dos - dob, kn = a.kn, intercept = TRUE) + Ns(dob,
    kn = b.kn) - 1
  Res.Df    RSS Df Sum of Sq     F   Pr(>F)
1  20605 398.98
2  20598 398.51  7   0.47075 3.476 0.001001
```

```
> l1lF <- update( l1F, . ~ . - Ns( dob, kn=b.kn ) + I((dob-b.ref)/10) )
> anova( l1lF, l1F )
```

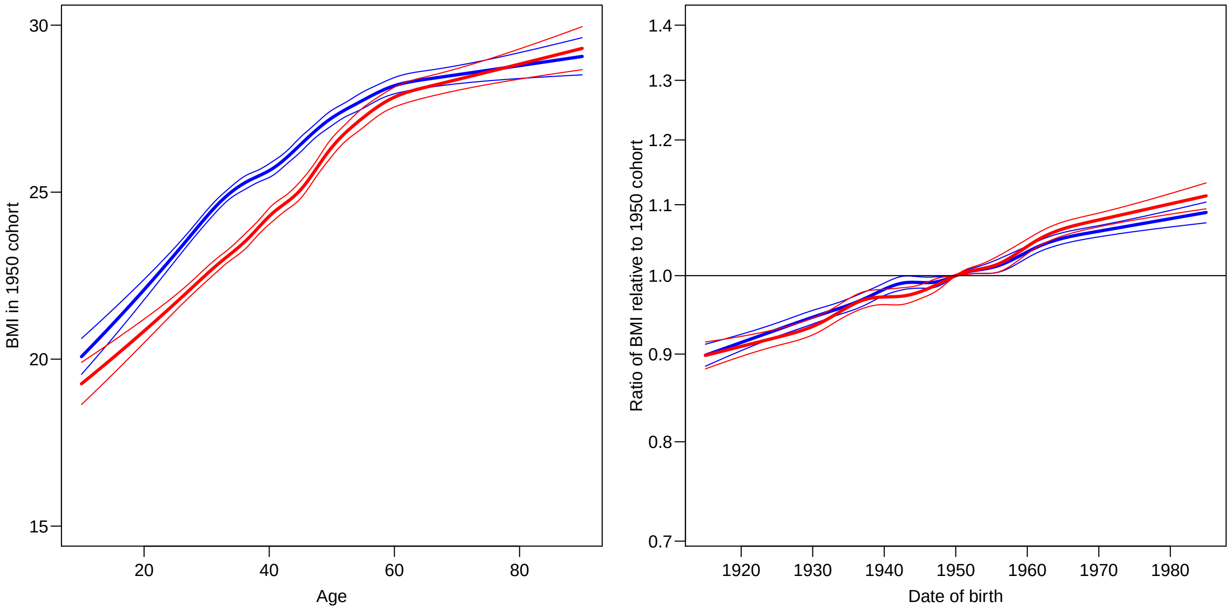

Figure 3.6: *Estimates of mean BMI in the Australian population in the 1950 birth cohort, and the relative change in mean BMI over cohorts. Blue curves are for men, red curves for women.*

```
Analysis of Variance Table

Model 1: log(bmi) ~ Ns(dos - dob, kn = a.kn, intercept = TRUE) + I((dob -
    b.ref)/10) - 1
Model 2: log(bmi) ~ Ns(dos - dob, kn = a.kn, intercept = TRUE) + Ns(dob,
    kn = b.kn) - 1
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1  21993 682.02
2  21986 681.36  7   0.65726 3.0298 0.003484
```

```
> rchg <- ( exp(rbind(ci.lin(l1lM,subset="ref"),
+                     ci.lin(l1lF,subset="ref"))[,c(1,5,6)]) - 1 ) * 100
> rownames(rchg) <- c("M","F")
> round(rchg,2)
```

```
  Estimate 2.5% 97.5%
M     2.86 2.65  3.08
F     3.42 3.15  3.69
```

The estimate of the relative change is a 3.2% increase in BMI per 10 years (95% c.i.: 3.0–3.4%) for men and 4.0% (3.7–4.2%) for women.

3.2 Prediction interval / population quantiles

Finally, we can look at the residual variation from the log-transformed model, which is interpretable as the coefficient of variation of BMI in the population (cf. [2], chapter 9):

```
> res <- rbind(
+ c( summary( l1lM )$sigma,
+    sd( residuals( l1lM ) ) ),
+ c( summary( l1lF )$sigma,
+    sd( residuals( l1lF ) ) ) )
> colnames( res ) <- c("sigma","sd(res)")
> rownames( res ) <- c("M","F")
> res
```

```
      sigma   sd(res)
M 0.1391519 0.1391215
F 0.1760985 0.1760624
```

Hence we see that the coefficient of variation for men is about 14%, whereas it for women is 18%.

This means that 95% prediction intervals are computed from the mean estimate by multiplying / dividing by exp(1.96 × 0.14) (for men). However we should also take the estimation error of the mean into account. Doing this we can make population quantiles separately for men and women.

We first set up an array to hold the prediction intervals (population quantiles) as a function of age, sex, cohort and quantile

```
> qnt <- c(2.5,5,10,25,50,75,90,95,97.5)/100
> pr.arr <- Array( list( sex = c("M","F"),
+                        coh = 1930 + 0:2*20,
+                        age = a.pt,
+                        qnt = qnt ) )
> str( pr.arr )
```

```
 logi [1:2, 1:3, 1:200, 1:9] NA NA NA NA NA NA ...
 - attr(*, "dimnames")=List of 4
  ..$ sex: chr [1:2] "M" "F"
  ..$ coh: chr [1:3] "1930" "1950" "1970"
  ..$ age: chr [1:200] "10" "10.4020100502513" "10.8040201005025" "11.2060301507538" ...
  ..$ qnt: chr [1:9] "0.025" "0.05" "0.1" "0.25" ...
```

```
> for( ic in dimnames(pr.arr)[[2]] )
+    {
+    ib <- as.numeric( ic )
+    a.linM <- ci.lin( l1lM, ctr.mat=cbind(Ca,(ib-b.ref)/10) )[,1:2]
+    a.linM[,2] <- sqrt( a.linM[,2]^2+summary(l1lM)$sigma^2)
+    pr.arr["M",ic,,] <- a.linM %*% rbind( 1, qnorm(qnt) )
+    a.linF <- ci.lin( l1lF, ctr.mat=cbind(Ca,(ib-b.ref)/10) )[,1:2]
+    a.linF[,2] <- sqrt( a.linF[,2]^2+summary(l1lF)$sigma^2)
+    pr.arr["F",ic,,] <- a.linF %*% rbind( 1, qnorm(qnt) )
+    }
> pr.arr <- exp( pr.arr )
> range( pr.arr )
```

```
[1] 12.69292 44.51355
```

We want to plot the predicted quantiles for three different generations of persons, separately for men and women.

3.3 Using the empirical distribution of residuals

We saw that the model for men fitted very nicely to the normal distribution of the residuals, whereas the model for women produced quite skewed residuals. So in producing prediction intervals we should not just use a normal distribution with variance equal to the residual variance, but the empirical distribution of the residuals instead.

But we should incorporate the estimation variance, too. Now recall the classical normal-based theory for generation of prediction intervals, for a linear model:

$$\mathbf{y} = \mathbf{X}\beta + e, \qquad e \sim \mathcal{N}(0, \sigma^2\mathbf{I})$$

The prediction variance for a new observation y_n at the (new) covariate point $\mathbf{x}_n$ ($< \times 1$) is:

$$\sigma^2\big(1 + \mathbf{x}_n(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_n'\big)$$

Thus, carrying this over to the use of the quantiles of the empirical residuals, just means that we compute the estimated mean under the model, and then add the empirical quantiles of the residuals multiplied by:

$$\sqrt{1 + \mathbf{x}_n(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_n'}$$

Recall from the theory of linear models that:

$$\mathrm{var}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

The matrix $(\mathbf{X}'\mathbf{X})^{-1}$ is available from a fitted `lm` object as `summary(m)$cov.unscaled`, so it is easy to set up a function that derives the required multiplier for the empirical residuals (`e`mpirical `r`esiduals `m`ultiplier), which takes the model and a matrix (with each row a prediction point) as the two arguments, and returns the fitted values plus the quantiles of the empirical residuals multiplied by this factor:

```
> erm <-
+ function( mm, xx )
+ as.vector( xx %*% coef(mm) ) +
+ outer( sqrt( 1 + diag(xx%*%summary(mm)$cov.unscaled%*%t(xx)) ),
+        quantile( residuals( mm ), probs=qnt ) )
```

With this settled we can now make a variant of the predictions from before:

```
> pr.emp <- pr.arr * 0
> for( ic in dimnames(pr.emp)[[2]] )
+    {
+    ib <- as.numeric( ic )
+    xx <- cbind( Ns(a.pt, kn = a.kn, intercept = TRUE), (ib-b.ref)/10 )
+    pr.emp["M",ic,,] <- erm( lllM, xx )
+    pr.emp["F",ic,,] <- erm( lllF, xx )
+    }
> pr.emp <- exp( pr.emp )
> range( pr.emp )
```

```
[1] 13.4011 47.5601
```

```
> onefr <-
+ function( ints, col, lty, lwd=c(1:4,6,4:1), ann=FALSE, grid=TRUE )
+ {
+ pt <- as.numeric(rownames(ints))
+ plot( NA, xlab="Age", ylab=paste("BMI in", b.ref, "cohort"),
+           ylim=c(12,50), xlim=range(pt)*c(1,1.05),
+           xaxt="n", yaxt="n")
+ if( grid )
+   {
+ abline( h=10:50, col=gray(0.9) )
+ abline( h=seq(10,50,5), v=seq(0,100,5), col=gray(0.8) )
+   }
+ matlines( pt, ints, lwd=lwd, type="l", lty=lty,
+           col=col )
+ if( ann ) {
+ text( max(pt)*1.06, ints[nrow(ints),], floor(qnt*100), adj=1 )
+ text( max(pt)*1.06, ints[nrow(ints),c(1,9)], ".5", adj=0 )
+           }
+ box()
+ }
```

```
> par( mfcol=c(3,2), mar=c(0,0,0,0),
+      oma=c(4,4,1,2), mgp=c(3,1,0)/1.6, las=1 )
> for( sx in 1:2 )
+ for( ic in 1:3 )
+ {
+ onefr( pr.emp[sx,ic,,], col=c("blue","red")[sx], lty=1, ann=TRUE )
+ par( new=TRUE )
+ onefr( pr.arr[sx,ic,,], col=c(rgb(50,50,100,max=100),rgb(100,50,50,max=100))[sx], lty=2, lwd=1, gr
+ if( ic==3 ) axis( side=1 )
+ if( sx==1 ) axis( side=2 )
+ if( sx==2 ) mtext( paste( dimnames(pr.arr)[["coh"]][ic], "generation" ),
+                    side=4, line=0.5, cex=0.8, las=0 )
+ }
> mtext( expression("BMI (kg/"*m^2*")"), side=2, line=2, outer=TRUE,las=0,cex=0.8 )
> mtext( "Age (years)", side=1, line=2, outer=TRUE,las=0,cex=0.8 )
> par( mfcol=c(3,2), mar=c(0,0,0,0),
+      oma=c(4,4,1,2), mgp=c(3,1,0)/1.6, las=1 )
> for( sx in 1:2 )
+ for( ic in 1:3 )
+ {
+ onefr( pr.emp[sx,ic,,], col=c("blue","red")[sx], lty=1, ann=TRUE )
+ if( ic==3 ) axis( side=1 )
+ if( sx==1 ) axis( side=2 )
+ if( sx==2 ) mtext( paste( dimnames(pr.arr)[["coh"]][ic], "generation" ),
+                    side=4, line=0.5, cex=0.8, las=0 )
+ }
> mtext( expression("BMI (kg/"*m^2*")"), side=2, line=2, outer=TRUE,las=0,cex=0.8 )
> mtext( "Age (years)", side=1, line=2, outer=TRUE,las=0,cex=0.8 )
```

```
> par( mfrow=c(2,3), mar=c(0,0,0,0),
+      oma=c(4,4,2,1), mgp=c(3,1,0)/1.6, las=1 )
> for( sx in 1:2 )
+ for( ic in 1:3 )
+ {
+ onefr( pr.emp[sx,ic,,], col=c("blue","red")[sx], lty=1, ann=TRUE )
+ if( ic==1 ) axis( side=2 )
+ if( sx==1 ) mtext( paste( dimnames(pr.arr)[["coh"]][ic], "generation" ),
+                    side=3, line=0.5, cex=0.8 )
+ if( sx==2 ) axis( side=1 )
+ }
> mtext( expression("BMI (kg/"*m^2*")"), side=2, line=2, outer=TRUE,las=0,cex=0.8 )
> mtext( "Age (years)", side=1, line=2, outer=TRUE,las=0,cex=0.8 )
```

It is clear from figure 3.7 that the distribution is somewhat skew, even after log-transformation, so that a simple model-based prediction would underestimate the higher quantiles of the BMI, particularly among women. Therefore, the relevant figures to consider are those based on the empirical distribution of the residuals, namely figures 3.8 and 3.9.

The figures 3.8 and 3.9 contain precisely the same curves, only arranged differently. From the figures it is seen that the middle 50% of the 1950 male generation in age 65 (that is in 2015) has BMI in the range 26–31, while for females the range is 25–32. The middle 90% of the males are in the range 22–34, for females that is 21–38.

3.4 Differential generational patterns of age-effect

Asking the question of whether the age-profile of BMI is the same across generations is statistically speaking one of an interaction between age and date of birth.

However, since the calendar time span of the data acquired is fairly narrow, a traditional approach of subdividing date of birth in a number of categories is not going to work, because different birth cohorts will span different age-ranges. Hence it will be more useful (and technically simpler too) to define a continuous interaction.

Since we expect curved deviations in age-specific deviations between birth cohorts, we simply introduce a quadratic in age which depends on date of birth, that is we add the terms

$$(\texttt{dos} - \texttt{dob}) \times \texttt{dob} \quad \text{and} \quad (\texttt{dos} - \texttt{dob})^2 \times \texttt{dob}$$

The coefficients to these two added terms will have no meaning, but we can show how the BMI relates to age for different birth cohorts by plotting the fitted values for select dates of birth.

For numerical stability we center the date of birth around `b.ref` before multiplying:

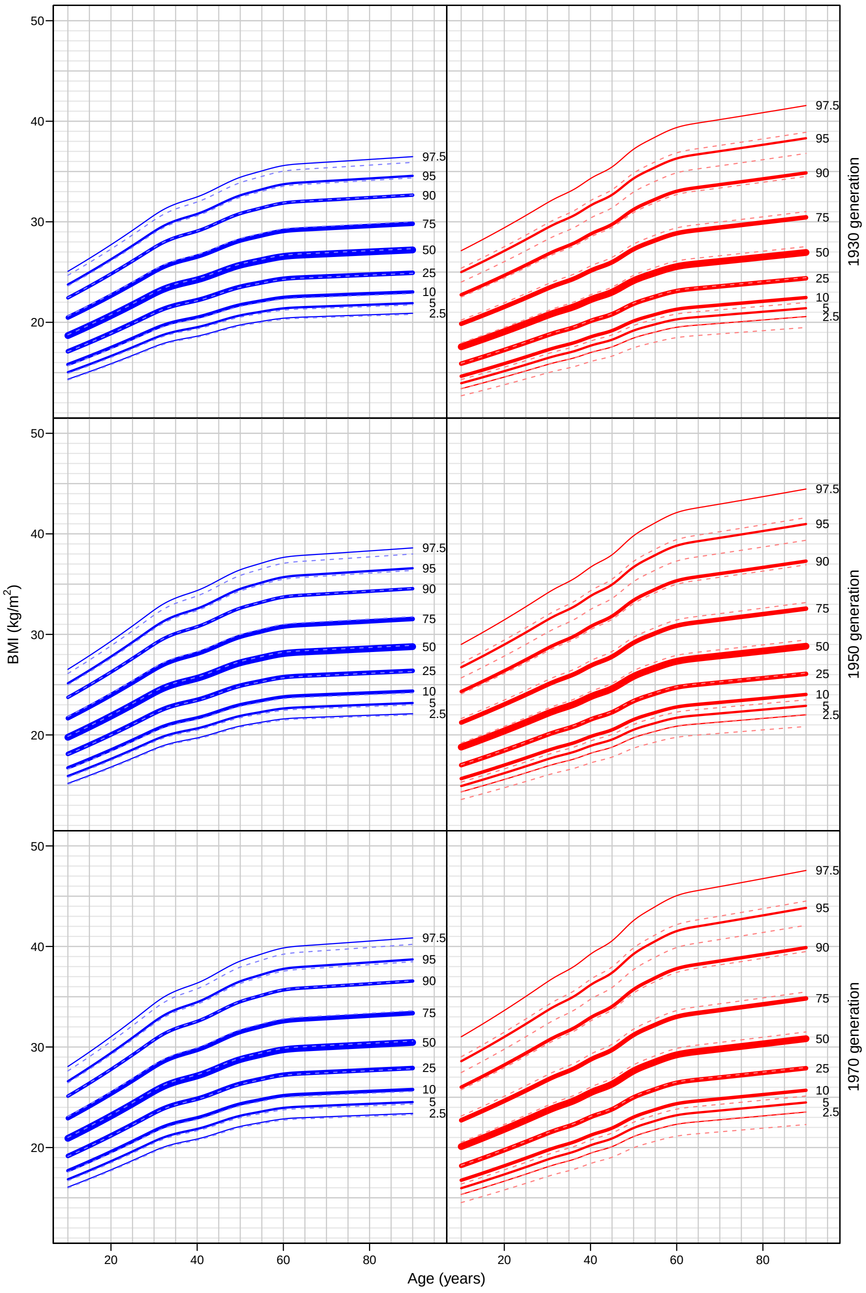

Figure 3.7: *Estimated population percentiles in three generations, based on a model for log-BMI, assuming constant coefficient of variation across the age-range (broken lines), compared to the prediction limits based on the empirical distribution of the residuals from the model. Blue curves for men, red curves for women.*

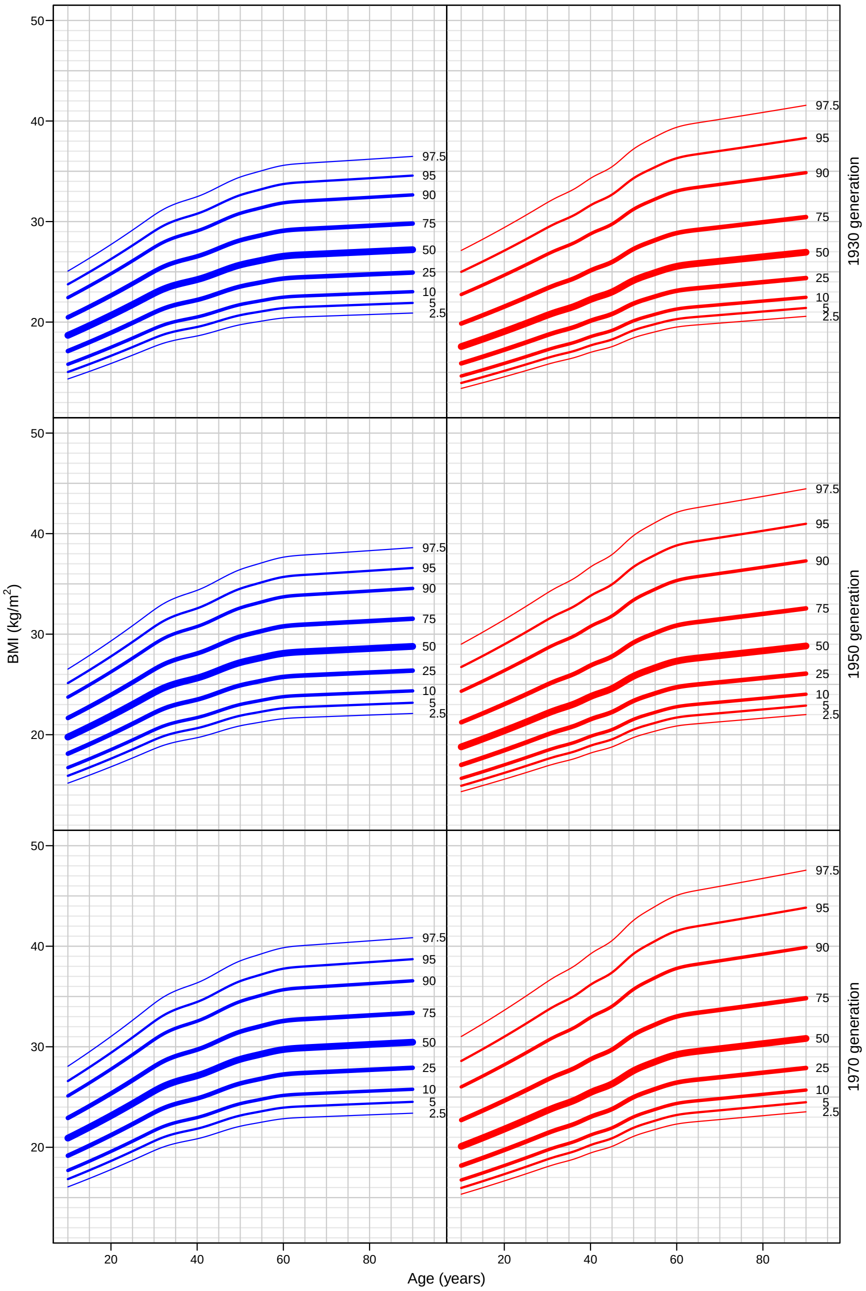

Figure 3.8: *Estimated population percentiles in three generations, based on a model for log-BMI, using the empirical distribution of residuals to construct prediction intervals for BMI. Blue curves for men, red curves for women.*

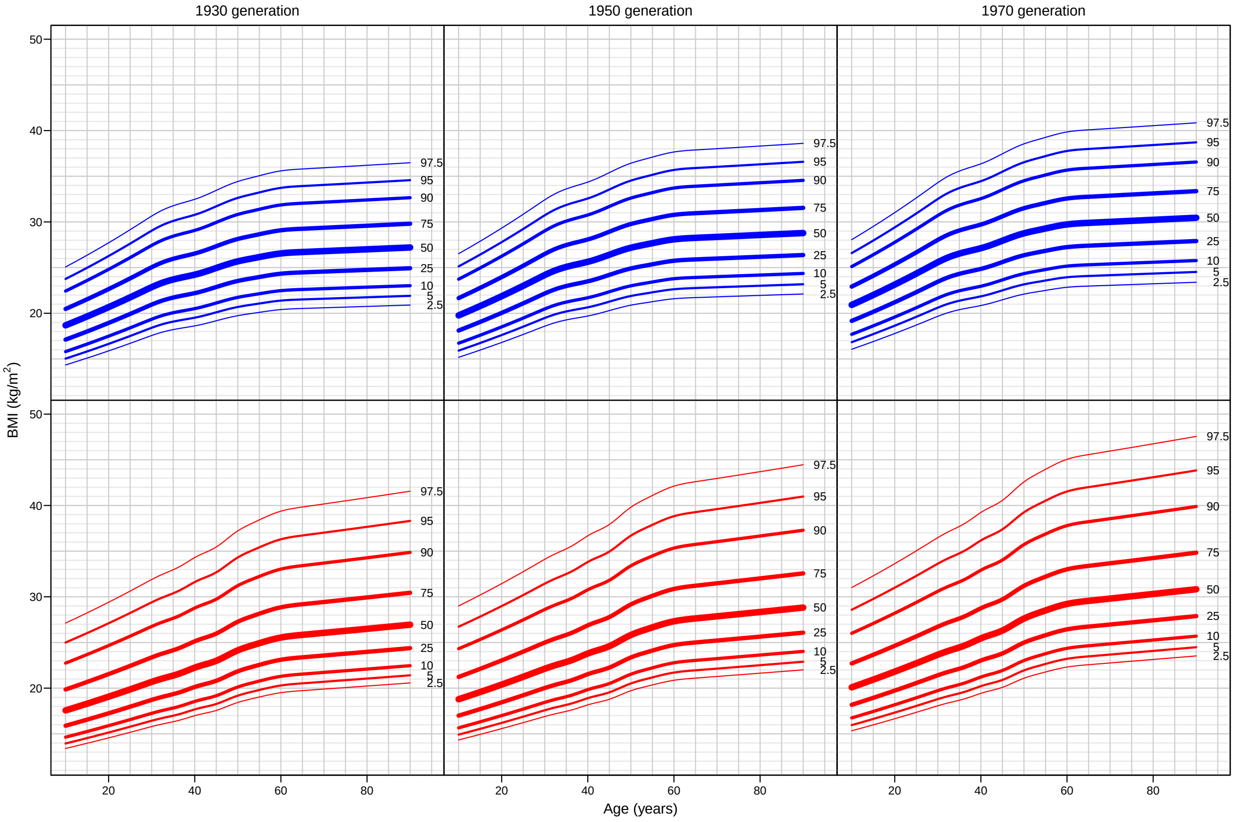

Figure 3.9: *Estimated population percentiles in three generations, based on a model for log-BMI, using the empirical distribution of residuals to construct prediction intervals for BMI. Blue curves for men, red curves for women.*

```
> summary(l1lM)


Call:
lm(formula = log(bmi) ~ Ns(dos - dob, kn = a.kn, intercept = TRUE) +
    I((dob - b.ref)/10) - 1, data = subset(abmi, sex == "M"))

Residuals:
     Min       1Q   Median       3Q      Max
-0.57996 -0.09165 -0.00450  0.08668  0.62673

Coefficients:
                                            Estimate Std. Error t value Pr(>|t|)
Ns(dos - dob, kn = a.kn, intercept = TRUE)1 2.641428   0.003651  723.55   <2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)2 3.233977   0.005848  553.01   <2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)3 3.249296   0.006355  511.32   <2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)4 3.277834   0.006599  496.68   <2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)5 3.306802   0.006503  508.49   <2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)6 3.321951   0.005917  561.45   <2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)7 1.999558   0.004395  454.99   <2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)8 7.238208   0.003954 1830.49   <2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)9 0.794445   0.004512  176.07   <2e-16
I((dob - b.ref)/10)                         0.028233   0.001082   26.11   <2e-16

Residual standard error: 0.1392 on 20605 degrees of freedom
Multiple R-squared:  0.9982,        Adjusted R-squared:  0.9982
F-statistic: 1.131e+06 on 10 and 20605 DF,  p-value: < 2.2e-16


> l1iM <- update( l1lM, . ~ . + I((dob-b.ref)/10 * (dos-dob) )
+                             + I((dob-b.ref)/10 * (dos-dob)^2 ) )
> summary( l1iM )


Call:
lm(formula = log(bmi) ~ Ns(dos - dob, kn = a.kn, intercept = TRUE) +
    I((dob - b.ref)/10) + I((dob - b.ref)/10 * (dos - dob)) +
    I((dob - b.ref)/10 * (dos - dob)^2) - 1, data = subset(abmi,
    sex == "M"))

Residuals:
     Min       1Q   Median       3Q      Max
-0.58879 -0.09174 -0.00432  0.08674  0.62998

Coefficients:
                                              Estimate Std. Error  t value Pr(>|t|)
Ns(dos - dob, kn = a.kn, intercept = TRUE)1  2.643e+00  3.876e-03  681.905  < 2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)2  3.235e+00  5.898e-03  548.507  < 2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)3  3.249e+00  6.367e-03  510.322  < 2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)4  3.278e+00  6.601e-03  496.620  < 2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)5  3.306e+00  6.571e-03  503.052  < 2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)6  3.325e+00  6.017e-03  552.549  < 2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)7  2.000e+00  4.722e-03  423.613  < 2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)8  7.249e+00  6.471e-03 1120.267  < 2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)9  8.023e-01  5.500e-03  145.860  < 2e-16
I((dob - b.ref)/10)                          3.061e-02  9.315e-03    3.286  0.00102
I((dob - b.ref)/10 * (dos - dob))           -3.189e-04  4.062e-04   -0.785  0.43242
I((dob - b.ref)/10 * (dos - dob)^2)          5.446e-06  4.260e-06    1.278  0.20117

Residual standard error: 0.1391 on 20603 degrees of freedom
Multiple R-squared:  0.9982,        Adjusted R-squared:  0.9982
F-statistic: 9.427e+05 on 12 and 20603 DF,  p-value: < 2.2e-16


> anova( l1iM, l1lM )
```

```
Analysis of Variance Table

Model 1: log(bmi) ~ Ns(dos - dob, kn = a.kn, intercept = TRUE) + I((dob -
    b.ref)/10) + I((dob - b.ref)/10 * (dos - dob)) + I((dob -
    b.ref)/10 * (dos - dob)^2) - 1
Model 2: log(bmi) ~ Ns(dos - dob, kn = a.kn, intercept = TRUE) + I((dob -
    b.ref)/10) - 1
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1  20603 398.84
2  20605 398.98 -2  -0.13859 3.5795 0.02791
```

```
> l1iF <- update( l1iM, data = subset(abmi, sex=="F") )
> summary( l1iF )
```

```
Call:
lm(formula = log(bmi) ~ Ns(dos - dob, kn = a.kn, intercept = TRUE) +
    I((dob - b.ref)/10) + I((dob - b.ref)/10 * (dos - dob)) +
    I((dob - b.ref)/10 * (dos - dob)^2) - 1, data = subset(abmi,
    sex == "F"))

Residuals:
     Min       1Q   Median       3Q      Max
-0.60551 -0.12139 -0.02127  0.10047  0.73261

Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
Ns(dos - dob, kn = a.kn, intercept = TRUE)1  2.577e+00  4.689e-03 549.688  < 2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)2  3.156e+00  7.248e-03 435.380  < 2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)3  3.200e+00  7.814e-03 409.510  < 2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)4  3.217e+00  8.133e-03 395.584  < 2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)5  3.273e+00  8.017e-03 408.296  < 2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)6  3.298e+00  7.300e-03 451.722  < 2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)7  2.016e+00  5.786e-03 348.453  < 2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)8  7.135e+00  8.043e-03 887.103  < 2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)9  8.403e-01  6.731e-03 124.846  < 2e-16
I((dob - b.ref)/10)                          4.125e-02  1.173e-02   3.517 0.000437
I((dob - b.ref)/10 * (dos - dob))           -1.508e-04  5.083e-04  -0.297 0.766793
I((dob - b.ref)/10 * (dos - dob)^2)         -3.900e-07  5.296e-06  -0.074 0.941292

Residual standard error: 0.1761 on 21991 degrees of freedom
Multiple R-squared:  0.997,	Adjusted R-squared:  0.997
F-statistic: 6.119e+05 on 12 and 21991 DF,  p-value: < 2.2e-16
```

```
> anova( l1iF, l1lF )
```

```
Analysis of Variance Table

Model 1: log(bmi) ~ Ns(dos - dob, kn = a.kn, intercept = TRUE) + I((dob -
    b.ref)/10) + I((dob - b.ref)/10 * (dos - dob)) + I((dob -
    b.ref)/10 * (dos - dob)^2) - 1
Model 2: log(bmi) ~ Ns(dos - dob, kn = a.kn, intercept = TRUE) + I((dob -
    b.ref)/10) - 1
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1  21991 681.91
2  21993 682.02 -2  -0.10948 1.7654 0.1711
```

From the ANOVA we see that there is no significant interaction for men, but certainly there is for women. Incidentally, neither of the two separate terms added arr significant for women, but the joint effect is. This is an example of collinear effects, where each of the two terms will be significant if the other is removed:

```
> summary( update( l11F, . ~ . + I((dob-b.ref)/10 * (dos-dob) ) ) )


Call:
lm(formula = log(bmi) ~ Ns(dos - dob, kn = a.kn, intercept = TRUE) +
    I((dob - b.ref)/10) + I((dob - b.ref)/10 * (dos - dob)) -
    1, data = subset(abmi, sex == "F"))

Residuals:
     Min       1Q   Median       3Q      Max
-0.60542 -0.12137 -0.02126  0.10046  0.73263

Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
Ns(dos - dob, kn = a.kn, intercept = TRUE)1  2.578e+00  4.444e-03 580.075   <2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)2  3.156e+00  7.239e-03 435.935   <2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)3  3.200e+00  7.802e-03 410.152   <2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)4  3.217e+00  8.132e-03 395.616   <2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)5  3.273e+00  7.936e-03 412.464   <2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)6  3.298e+00  7.300e-03 451.760   <2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)7  2.016e+00  5.526e-03 364.842   <2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)8  7.134e+00  7.950e-03 897.469   <2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)9  8.406e-01  6.042e-03 139.124   <2e-16
I((dob - b.ref)/10)                          4.204e-02  4.678e-03   8.988   <2e-16
I((dob - b.ref)/10 * (dos - dob))           -1.875e-04  9.984e-05  -1.878   0.0604

Residual standard error: 0.1761 on 21992 degrees of freedom
Multiple R-squared:  0.997,        Adjusted R-squared:  0.997
F-statistic: 6.675e+05 on 11 and 21992 DF,  p-value: < 2.2e-16
```

```
> summary( update( l11F, . ~ . + I((dob-b.ref)/10 * (dos-dob)^2 ) ) )


Call:
lm(formula = log(bmi) ~ Ns(dos - dob, kn = a.kn, intercept = TRUE) +
    I((dob - b.ref)/10) + I((dob - b.ref)/10 * (dos - dob)^2) -
    1, data = subset(abmi, sex == "F"))

Residuals:
     Min       1Q   Median       3Q      Max
-0.60598 -0.12141 -0.02139  0.10036  0.73253

Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
Ns(dos - dob, kn = a.kn, intercept = TRUE)1  2.577e+00  4.450e-03 579.139   <2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)2  3.156e+00  7.228e-03 436.594   <2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)3  3.200e+00  7.800e-03 410.241   <2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)4  3.217e+00  8.132e-03 395.635   <2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)5  3.274e+00  7.938e-03 412.380   <2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)6  3.298e+00  7.298e-03 451.872   <2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)7  2.017e+00  5.454e-03 369.785   <2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)8  7.135e+00  7.665e-03 930.845   <2e-16
Ns(dos - dob, kn = a.kn, intercept = TRUE)9  8.396e-01  6.271e-03 133.885   <2e-16
I((dob - b.ref)/10)                          3.786e-02  2.644e-03  14.320   <2e-16
I((dob - b.ref)/10 * (dos - dob)^2)         -1.930e-06  1.040e-06  -1.856   0.0635

Residual standard error: 0.1761 on 21992 degrees of freedom
Multiple R-squared:  0.997,        Adjusted R-squared:  0.997
F-statistic: 6.675e+05 on 11 and 21992 DF,  p-value: < 2.2e-16
```

The latter of these two models is not relevant; it violates the principle of marginality, so we might actually argue that the model with the linear interaction is sufficient for the interaction, but the *a priori* assumption was a curved interaction, so we maintain the

2-parameter interactions, fit them to data and show the resulting BMI-trajectories for select birth cohorts.

So we simply re-use the machinery from the previous section to tease out the estimates of the mean BMI as a function of age for different birth cohorts. To see what birth cohorts would be relevant to use, we make a quick plot of the distribution of the date of birth and date of survey.

```
> par( mfrow=c(1,2) )
> with( abmi, hist( dob, breaks=100 ) )
> with( abmi, hist( dos, breaks=100 ) )
```

To get the predicted *mean* BMI in different cohort using the interaction model, we use the `predict` method, first we construct a data-frame where the only relevant combinations of age and birth cohort are present (that is, we do not plot outside the survey data combinations of age and date of birth):

```
> dob.pt <- seq(1920,1980,10)
> age.pt <- seq(10,90,0.2)
> nd <- data.frame( dob = rep(dob.pt, each=length(age.pt)),
+                    age = rep(age.pt,      length(dob.pt)) )
> nd$dos <- nd$dob + nd$age
> str( nd )
```

```
'data.frame':        2807 obs. of  3 variables:
 $ dob: num  1920 1920 1920 1920 1920 1920 1920 1920 1920 1920 ...
 $ age: num  10 10.2 10.4 10.6 10.8 11 11.2 11.4 11.6 11.8 ...
 $ dos: num  1930 1930 1930 1931 1931 ...
```

```
> nr <- subset( nd, dos>1975 & dos<2015 )
> str( nd )
```

```
'data.frame':        2807 obs. of  3 variables:
 $ dob: num  1920 1920 1920 1920 1920 1920 1920 1920 1920 1920 ...
 $ age: num  10 10.2 10.4 10.6 10.8 11 11.2 11.4 11.6 11.8 ...
 $ dos: num  1930 1930 1930 1931 1931 ...
```

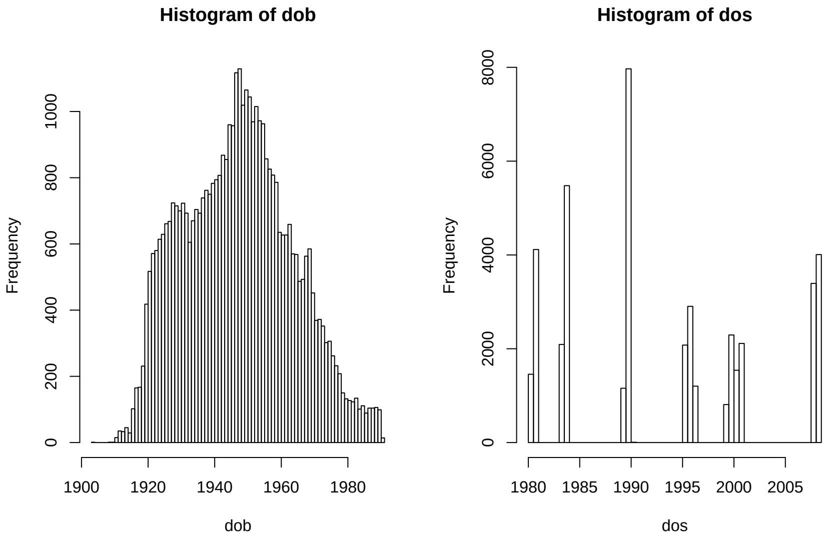

Figure 3.10: *Histogram of data of birth in all included surveys*

```
> prM <- predict( l1iM, newdata=nr )
> prF <- predict( l1iF, newdata=nr )
> pfM <- predict( l1iM, newdata=nd )
> pfF <- predict( l1iF, newdata=nd )
```

With this in place we can now plot the predicted BMI under the interaction model for different ages. The dataframe `nd` was used for prediction, and hence we use the age-column from this (`nd$age`) when plotting against age.

```
> par( mfrow=c(1,2), mar=c(0,0,0,0), oma=c(3,3,1,1), mgp=c(3,1,0)/1.6 )
> plot( NA, xlab="Age", ylab="BMI",
+           ylim=c(15,35), xlim=c(10,90) )
> abline( h=10:50, col=gray(0.9) )
> abline( h=seq(10,50,5), v=seq(0,100,5), col=gray(0.8) )
> points( nd$age, exp(pfM), pch=16, col="gray", cex=0.5 )
> points( nr$age, exp(prM), pch=16, col="blue", cex=0.5 )
> box()
> plot( NA, xlab="Age", ylab="BMI",
+           ylim=c(15,35), xlim=c(10,90),
+           yaxt="n")
> abline( h=10:50, col=gray(0.9) )
> abline( h=seq(10,50,5), v=seq(0,100,5), col=gray(0.8) )
> points( nd$age, exp(pfF), pch=16, col="gray", cex=0.5 )
> points( nr$age, exp(prF), pch=16, col="red", cex=0.5 )
```

From figure 3.11 it is difficult to say much about differential trends in bmi between generations, but there is a tendency that for men generations get further apart in old age than do women.

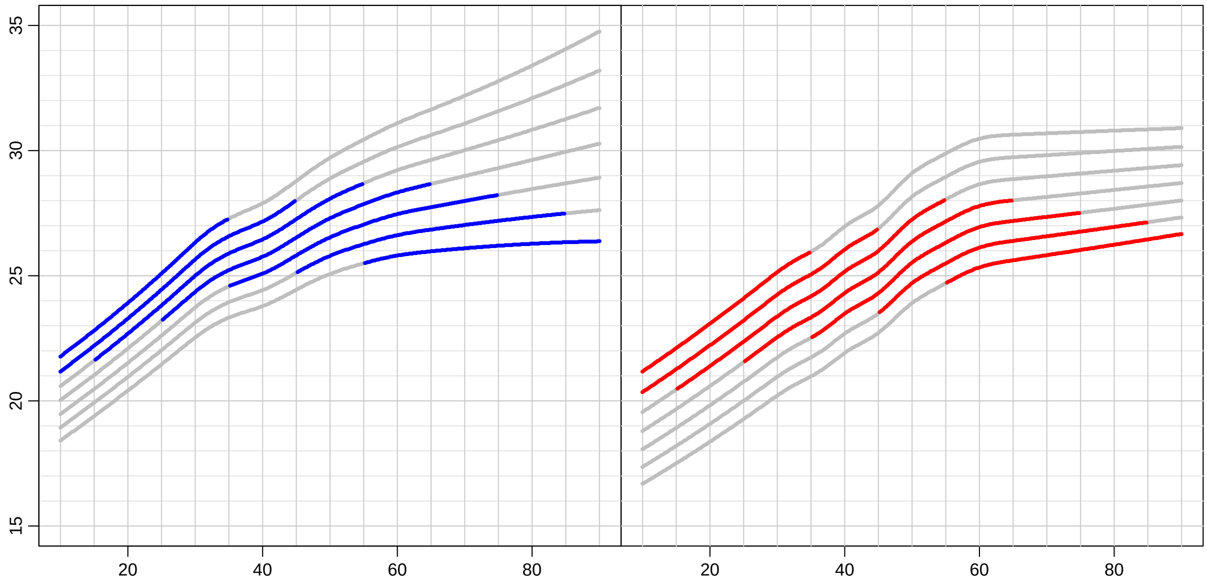

Figure 3.11: *Predicted mean BMI under the logarithmic interaction model for birth cohorts 1920, 1930,...,1980. The colored part of the curves are where data are available from the surveys.*

Chapter 4

Analysis by quantile regression

We first reload data etc.:

```
> library( Epi )
> library( splines )
> library( quantreg )
> print( sessionInfo(), l=F )

R version 3.0.2 (2013-09-25)
Platform: i386-w64-mingw32/i386 (32-bit)

attached base packages:
[1] splines   utils     datasets  graphics  grDevices stats     methods  
[8] base     

other attached packages:
[1] quantreg_4.98  SparseM_0.99   Epi_1.1.59     foreign_0.8-55

loaded via a namespace (and not attached):
[1] tools_3.0.2

> options( width=100 )
> load( file="./data/abmi.Rdata" )
> abmi$age <- abmi$dos - abmi$dob
> nk <- 6
> a.kn <- with( abmi, quantile( dos-dob, probs=1:nk/(nk+1) ) )
> b.kn <- with( abmi, quantile(     dob, probs=1:nk/(nk+1) ) )
> qnt <- c(5,10,25,50,75,90,95)/100
> save( nk, a.kn, b.kn, file="./data/knots.Rdata" )
```

In the light of the previous findings, the logical description of generational trends would therefore be to do plots as those in figure 3.11 for selected percentiles of the distribution, such as 5, 10, 25, 50, 75, 90 and 95.

This can be done by *quantile regression*. Basically this means that for a given selected quantile we make a model similar to the ones above for the mean. This is then done for different quantiles, and the shapes of the age-specific BMI-quantiles then shown graphically. This will allow us to see to what extent the distribution of BMI in the population depends on age, calendar time and birth cohort.

4.1 Age-Period-Cohort effects

We first fit a traditional age-period-cohort model to each of 7 different quantiles, 5, 10, 25, 50, 75, 90 and 95 %. Referring to the discussion of parametrization of rates in an

Age-Period-Cohort model by Carstensen [1], we can make a similar parametrization for each of the quantiles, by properly defining the design-matrices for the age, period and cohort effects.

4.1.1 APC parametrization

We want to use a parametrization where the period effect is constrained to be 0 on average with an average slope of 0. Since we have entered the period effect with a one-parameter non-linear term (a quadratic), we should construct a version of this that is orthogonal to the linear (and constant). Similarly we should do this for the cohort effect, but later add the linear trend in cohort to this, but also subtract the term corresponding to a reference cohort.

What we do is to set up a function that does this for a particular data frame, specifically

```
> age <- "age"
> coh <- "dob"
> c.kn <- b.kn
> ( p.kn <- quantile( abmi$dos, (1:3-0.5)/3 ) )
```

```
16.66667%       50% 83.33333%
 1983.465  1989.811  2007.614
```

```
> c0 <- 1950
> qnt <- c(5,10,25,50,75,90,95)/100
> Gen.APC <-
+ function( dfr )
+ {
+ nr <- nrow( dfr )
+ # Base variables in the model
+ Age <- dfr[,age]
+ Coh <- dfr[,coh]
+ Per <- Coh+Age
+ # Design matrices for the models
+ Ma <- Ns( Age, knots=a.kn, intercept=TRUE )
+ Mc <- Ns(               Coh  , knots=c.kn ) -
+       Ns( rep(c0,length(Coh)), knots=c.kn )
+ mc <- detrend( Mc, Coh )
+ Mp <- detrend( Ns( Per, knots=p.kn ), Per )
+ # Points of prediction
+ prA <- sort(Age)[floor(seq(1,nr,,200))]
+ prC <- sort(Coh)[floor(seq(1,nr,,200))]
+ prP <- sort(Per)[floor(seq(1,nr,,200))]
+ # Contrast matrices to be used - use the relevant 200 rows of the
+ # design matrices
+ Ca <- Ma[match(prA,Age),]
+ Cc <- Mc[match(prC,Coh),]
+ Cp <- Mp[match(prP,Per),,drop=FALSE]
+ # Array for the tests for drift model
+ tarr <- NArray( list( qnt=qnt,
+                       c("lin P?","lin C?",
+                         "Dr.ext","lo","hi",
+                         "Dr.raw","lo","hi") ) )
+ # Fit the relevant model
+ mod <- list()
+ lef <- list( A = prA,
+              C = prC,
+              P = prP )
+ for( q in qnt )
+    {
```

```
+    mod <- rq( bmi ~ Ma + Mc + Mp - 1,
+               tau = q,
+               data = dfr )
+    lef$A <- cbind( lef$A, ci.lin( mod, subset="Ma", ctr.mat=Ca )[,c(1,5,6)] )
+    lef$C <- cbind( lef$C, ci.lin( mod, subset="Mc", ctr.mat=Cc )[,c(1,5,6)] )
+    lef$P <- cbind( lef$P, ci.lin( mod, subset="Mp", ctr.mat=Cp )[,c(1,5,6)] )
+    mod <- rq( bmi ~ Ma + I((dob-1950)/5) + mc + Mp - 1,
+               tau = q,
+               data = dfr )
+    mdc <- update( mod, . ~ . - Mp )
+    mdr <- update( mdc, . ~ . - mc )
+    tarr[paste(q),1:2] <- anova( mod, mdc, mdr )$table[,"pvalue"]
+    tarr[paste(q),3:5] <- ci.exp( mod, subset="dob", Exp=FALSE )
+    tarr[paste(q),6:8] <- ci.exp( mdr, subset="dob", Exp=FALSE )
+    }
+ list( lef=lef, tarr=tarr )
+ }
```

With this we can now make separate models for males and females, collect the estimates and plot the age-, cohort- and period effects

```
> system.time( mm <- Gen.APC( subset(abmi,sex=="M") ) )
```

```
   user  system elapsed
  45.88    0.39   46.43
```

```
> system.time( ff <- Gen.APC( subset(abmi,sex=="F") ) )
```

```
   user  system elapsed
  47.05    0.09   47.32
```

```
> round( mt <- mm$tarr, 3 )
```

```
qnt    lin P? lin C? Dr.ext    lo    hi Dr.raw    lo    hi
  0.05  0.004  0.000  0.141 0.088 0.193  0.143 0.096 0.190
  0.1   0.081  0.000  0.200 0.160 0.241  0.203 0.165 0.241
  0.25  0.025  0.000  0.273 0.238 0.308  0.270 0.237 0.302
  0.5   0.049  0.001  0.343 0.308 0.379  0.349 0.315 0.383
  0.75  0.472  0.025  0.478 0.434 0.522  0.477 0.434 0.519
  0.9   0.123  0.498  0.645 0.580 0.710  0.653 0.586 0.720
  0.95  0.791  0.651  0.771 0.670 0.872  0.784 0.698 0.869
```

```
> round( ft <- ff$tarr, 3 )
```

```
qnt    lin P? lin C? Dr.ext    lo    hi Dr.raw    lo    hi
  0.05  0.000  0.000  0.106 0.069 0.143  0.117 0.079 0.155
  0.1   0.000  0.000  0.135 0.100 0.169  0.135 0.102 0.168
  0.25  0.000  0.001  0.214 0.182 0.247  0.213 0.181 0.244
  0.5   0.000  0.001  0.385 0.343 0.427  0.391 0.351 0.431
  0.75  0.001  0.006  0.665 0.602 0.727  0.677 0.615 0.738
  0.9   0.000  0.001  0.895 0.789 1.001  0.911 0.810 1.011
  0.95  0.003  0.008  1.008 0.867 1.148  1.021 0.884 1.157
```

```
> mm <- mm$lef
> ff <- ff$lef
```

We see that there are significant non-linear effects of both period and cohort, so we plot these for each of the selected quantiles. To get a maximally informative graph, we plot the effects using the same physical units for both cohort and period effects:

```
> xoff <- 1800
> xl   <- c(20+xoff,2005)
> yoff <- 25
> yl   <- c(13,42)
> temp.graph <-
+ function()
+ {
+ par( mfrow=c(2,1), mar=c(0,0,0,0), oma=c(3,3,1,3), mgp=c(3,1,0)/1.6,
+     las=1, cex=1 )
+ # Men
+ clr <- rep(rgb(0,0,1,2:8/10),each=3)
+ plot( NA, xlim=xl, xaxt="n", ylim=yl )
+ abline( h=seq(0,50,5), v=seq(1800,2020,10), col=gray(0.9) )
+ matlines( mm$A[,1]+xoff, mm$A[,-1], type="l", lty=c(1,2,2), lwd=c(3,1,1), col=clr )
+ lines( c(1903,2030), rep(yoff,2) )
+ matlines( mm$C[,1], mm$C[,-1]+yoff, type="l", lty=c(1,2,2), lwd=c(3,1,1), col=clr )
+ matlines( mm$P[,1], mm$P[,-1]+yoff, type="l", lty=c(1,2,2), lwd=c(3,1,1), col=clr )
+ axis( side=4, at=seq(15,40,5), labels=seq(15,40,5)-yoff )
+ # abline( v=1903 )
+ text( rep(max(mm$A[,1]),7)+xoff+1, mm$A[dim(mm$A)[1],2+0:6*3],
+       paste(c(5,10,25,50,75,90,95)), adj=0 )
+ # Women
+ clr <- rep(rgb(1,0,0,2:8/10),each=3)
+ plot( NA, xlim=xl, xaxt="n", ylim=yl )
+ abline( h=seq(0,50,5), v=seq(1800,2020,10), col=gray(0.9) )
+ matlines( ff$A[,1]+xoff, ff$A[,-1], type="l", lty=c(1,2,2), lwd=c(3,1,1), col=clr )
+ lines( c(1903,2030), rep(yoff,2) )
+ matlines( ff$C[,1], ff$C[,-1]+yoff, type="l", lty=c(1,2,2), lwd=c(3,1,1), col=clr )
+ matlines( ff$P[,1], ff$P[,-1]+yoff, type="l", lty=c(1,2,2), lwd=c(3,1,1), col=clr )
+ axis( side=1, at=xoff+seq(20,80,20), labels=seq(20,80,20) )
+ axis( side=1, at=seq(1910,2010,20) )
+ axis( side=1, at=(xx<-seq(xoff,2010,10)), labels=rep("",length(xx)) )
+ axis( side=4, at=seq(15,40,5), labels=seq(15,40,5)-yoff )
+ # abline( v=1903 )
+ text( rep(max(ff$A[,1]),7)+xoff+1, ff$A[dim(ff$A)[1],2+0:6*3],
+       paste(c(5,10,25,50,75,90,95)), adj=0 )
+
+ mtext( side=2, expression("BMI percentiles in 1950 cohort (kg/"*m^2*")"),
+        line=1.7, outer=TRUE, las=0 )
+ mtext( "BMI differences" , side=4, outer=TRUE, line=1.9, las=0 )
+ mtext( side=1, "Age", at=40/diff(par("usr")[1:2]), line=1.5, outer=TRUE )
+ mtext( side=1, "Date of birth" , at=(1950-par("usr")[1])/diff(par("usr")[1:2]), line=1.5, outer=TR
+ mtext( side=1, "Date of survey", at=(1995-par("usr")[1])/diff(par("usr")[1:2]), line=1.5, outer=TR
+ }
> temp.graph()
> win.metafile( "art/fig2.emf", height=10, width=10) ; temp.graph() ; dev.off()
```

```
pdf
  2
```

```
>   postscript( "art/fig2.eps", height=10, width=10) ; temp.graph() ; dev.off()
```

```
pdf
  2
```

```
>          pdf( "art/fig2.pdf", height=10, width=10) ; temp.graph() ; dev.off()
```

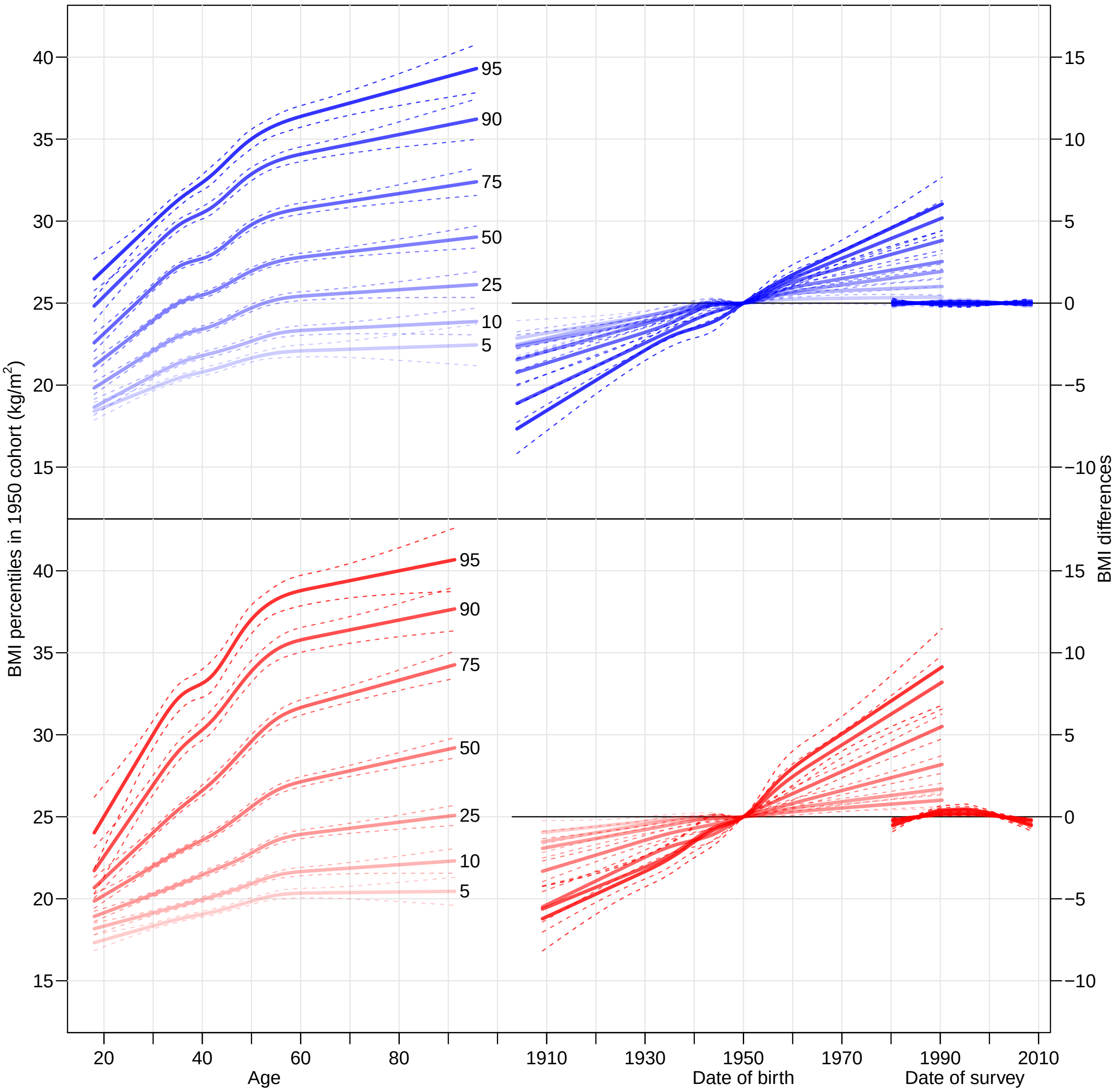

Figure 4.1: *Estimated age-, cohort- and period effects for percentiles, with 95% confidence intervals (broekn lines), using the 1950 birth cohort as referece, and constraining the period effects to be 0 on average.*

```
pdf
  2
```

Figure **??** shows `longitudinal` quantiles of BMI in the population, that is how the quantile in the population alive evolves by age. Since we only have a span of slightly less than 30 years of surveys, these curves are “patched” together from observation in different ages and periods.

It is seen from figure **??** that:

- Within a cohort, the major increase in BMI is before age 50, after that age the increase is moderate. However this may to some extent be due to the effect of mortality, where the most obese are known to have the highest mortality.
- The spread among women is considerably larger for women than for men, at age 60 the 5 and 95th percentiles are 20 and 38 kg/m^2 for women, but 22 and 37 men, while the median is 27 for both sexes.
- The increase across birth cohorts is much stronger in the highest quantiles, than for the smallest quantiles.
- This effect is more pronounced among women than among men, the annual average increase in the 5th and 95th percentiles BMI are 0.12 are 1.02 kg/m^2 for women but 0.14 and 0.78 for men.
- There is no detectable period effect (apart from the linear trend that cannot be separated from the cohort effect).
- For most practical purposes the cohort effects are linear for each percentile, despite the fact that the formal tests of linearity are highly significant (except for the highest percentiles for men, which are not significant).
- As a summary of the secular trends in BMI it is therefore sufficient to quote constant change in each of the quantiles. This can be interpreted both as the change per calendar time or the change per birth cohort.

4.2 Trend across quantiles

The age-period-cohort models we just fitted produce results that are quite close to an age-drift model — virtually no period effect and only a slight curvature in the cohort effects are seen, despite the formally significant tests for linearity.

We therefore estimate the drift parameter for each of the percentiles between 2 and 98 and plot them:

```
> qnt <- 2:98/100
> tdr <- NArray( list( qnt=qnt, sex=levels(abmi$sex), c("Est","lo","hi") ) )
> system.time(
+ for( iq in dimnames(tdr)[["qnt"]] )
+ for( sx in dimnames(tdr)[["sex"]] )
+ {
+ tdr[iq,sx,] <- ci.exp( rq( bmi ~ Ns(age,kn=a.kn) + I((dob-1950)/5),
+                            tau = as.numeric(iq),
+                            data=subset( abmi, sex==sx ) ),
+                         subset="dob",
+                         Exp=FALSE )
+ }
+ )
```

```
   user  system elapsed 
 158.16    0.00  158.44 
```

We can now put these estimated 5-year-changes into a graph:

```
> temp.abs <-
+ function(trsp=TRUE)
+ {
+ par( mar=c(3,3.2,1,1), mgp=c(3,1,0)/1.6 )
+ plot( NA, bty="n", las=1,
+       xlim=c(-2,100), xlab="BMI percentile", xaxs="i",
+       ylim=c(0,1.2), yaxs="i",
+       ylab=expression("Change per 5 years ("*kg/m^2*")" ) )
+ abline( h=seq(0.1,1.4,0.1), v=seq(0,100,10), col=gray(0.9) )
+ matlines( qq <- as.numeric(dimnames(tdr)[["qnt"]])*100,
+           cbind( tdr[,"M",], tdr[,"F",] ),
+           lty=c(1,0,0), lwd=c(3,1,1), col=rep( c("blue","red"), each=3 ) )
+ if( trsp ) {
+ polygon( c(qq,rev(qq)), c(tdr[,"M",2], rev(tdr[,"M",3]) ),
+          col=rgb(0,0,1,0.2), border="transparent" )
+ polygon( c(qq,rev(qq)), c(tdr[,"F",2], rev(tdr[,"F",3]) ),
+          col=rgb(1,0,0,0.2), border="transparent" )
+            }
+ else
+ matlines( qq, cbind( tdr[,"M",-1], tdr[,"F",-1] ),
+           lty="23", lwd=1, col=rep( c("blue","red"), each=2 ) )
+ }
> win.metafile(       "art-BMI-chg.emf", height=7, width=7) ; temp.abs(trsp=FALSE) ; dev.off()
```

```
null device 
          1 
```

```
> pdf("./graph/BMI-APC-art-BMI-chg.pdf", height=7, width=7) ; temp.abs(trsp=FALSE) ; dev.off()
```

```
null device 
          1 
```

```
> pdf("./graph/BMI-APC-art-BMI-cht.pdf", height=7, width=7) ; temp.abs(          ) ; dev.off()
```

```
null device 
          1 
```

Figure **??** (and **??**) shows the changes in each of the BMI-percentiles over time (assuming that the change is constant over the period 1980–2010). The 90th BMI percentile among women has changed by 1 kg/m^2 per 5 years and the higher percentiles even more; meaning that the 10% most obese women have a bmi more than 4 kg/m^2 above that 20 year ago.

It should be noted that there are no inherent assumptions that these curves should necessarily be increasing; the increase is a feature of the data — the most obese parts of the population grow faster.

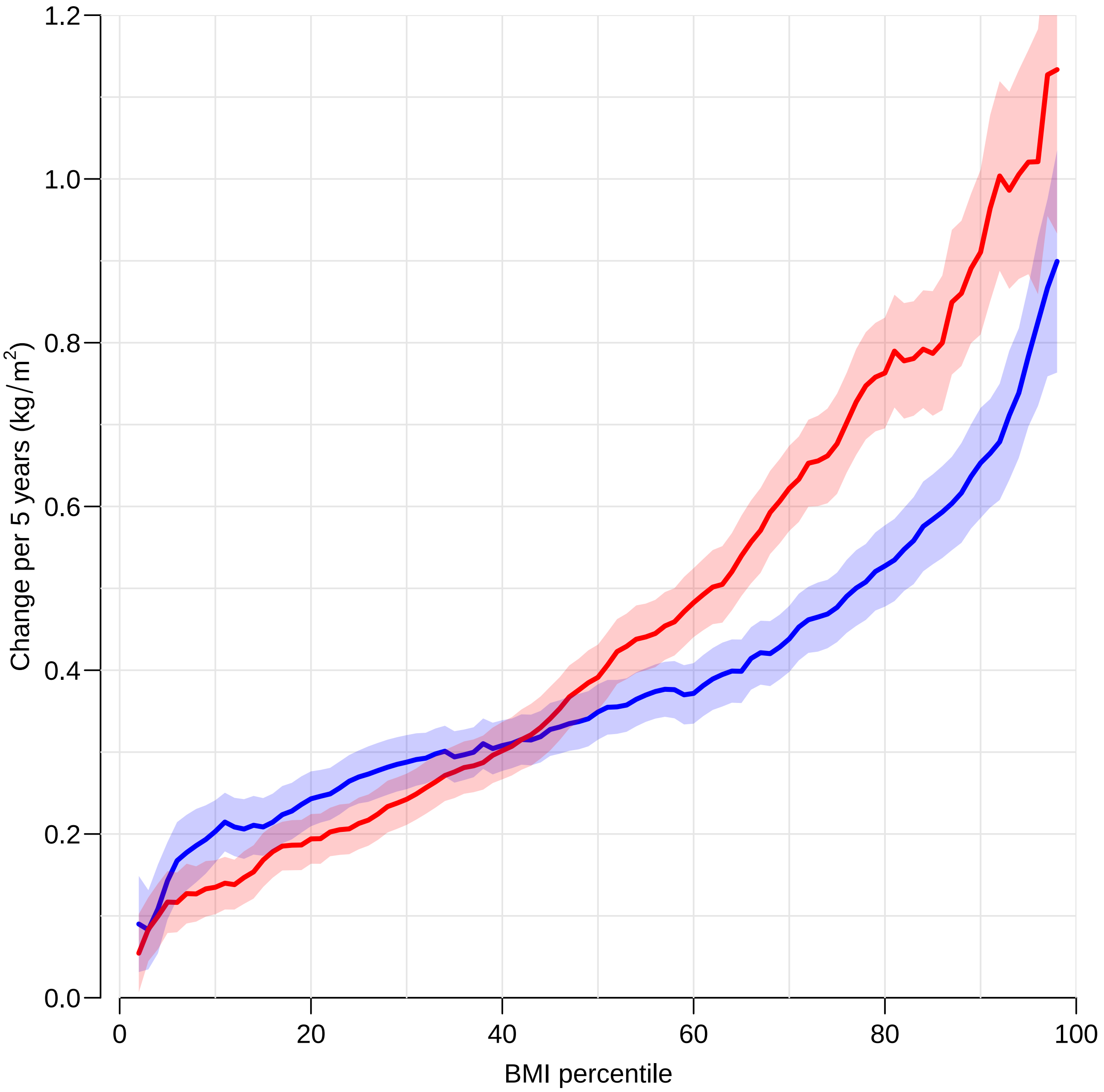

Figure 4.2: *Changes in each percentile of the BMI-distribution per 5 years. Thus, for women the 80th percentile has on average changed by 0.75 kg/m^2 every 5 years, the 90th by 0.90 kg/m^2. For men the corresponding figures are 0.52 and 0.65. Moreover, the 40% leanest in the population has seen an increase in BMI of less than 0.3 kg/m^2 per 5 years; significant bt moderate.*

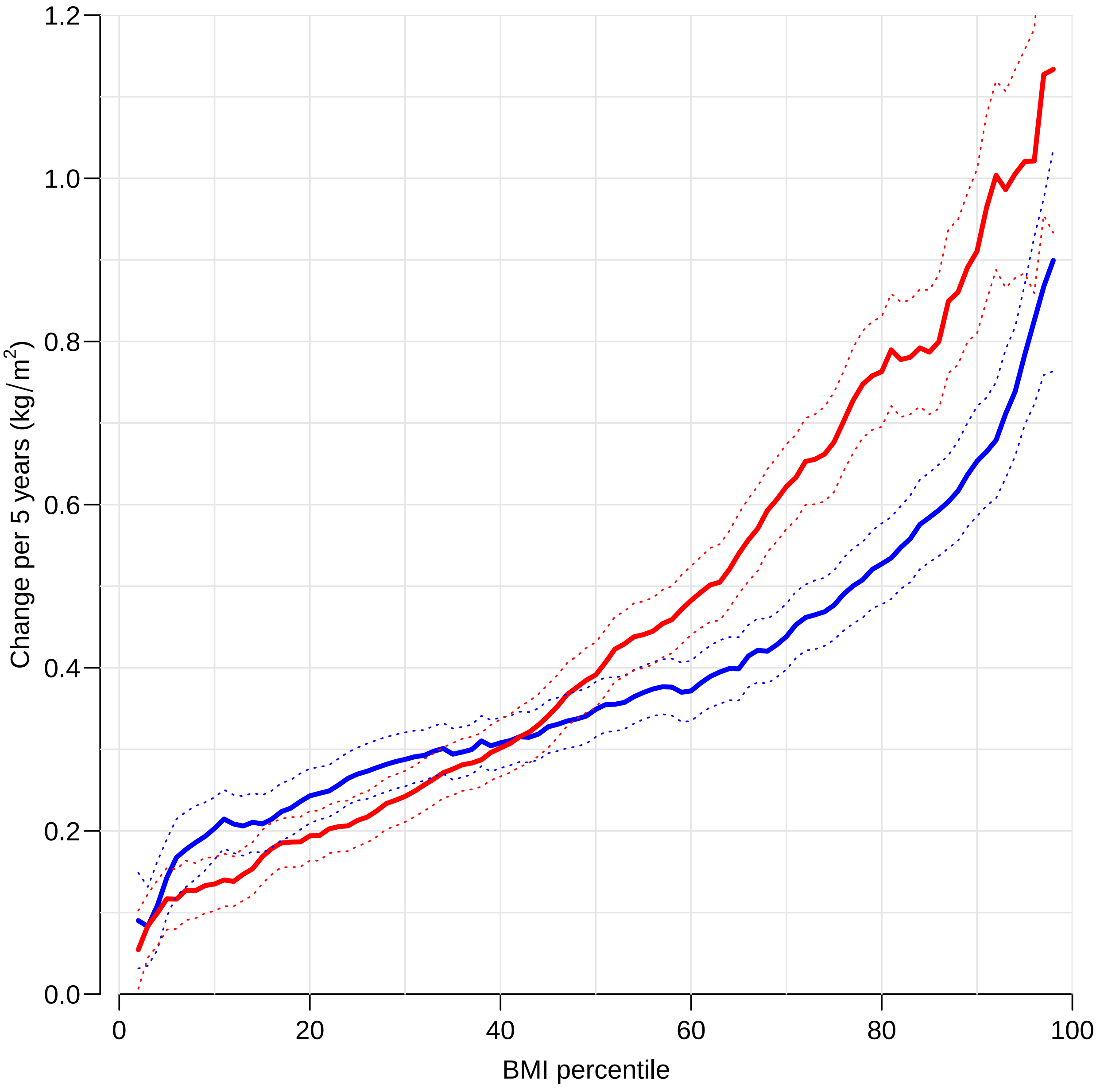

Figure 4.3: *Changes in each percentile of the BMI-distribution per 5 years. Thus, for women the 80th percentile has on average changed by 0.75 kg/m^2 every 5 years, the 90th by 0.90 kg/m^2. For men the corresponding figures are 0.52 and 0.65. Moreover, the 40% leanest in the population has seen an increase in BMI of less than 0.3 kg/m^2 per 5 years; significant bt moderate.*

4.2.1 Relative increase in BMI

It would of course be of interest to see whether the *relative* increase in BMI was largest for the most obese people too. This amounts to fitting a model where the response is now log(BMI), and then back-transforming the parameter to a percentage change by $100(\exp(\beta) - 1)$, otherwise it is all the same all over again:

```
> qnt <- 2:98/100
> pdr <- NArray( list( qnt=qnt, sex=levels(abmi$sex), c("Est","lo","hi") ) )
> system.time(
+ for( iq in dimnames(tdr)[["qnt"]] )
+ for( sx in dimnames(tdr)[["sex"]] )
+ pdr[iq,sx,] <- ( ci.exp( rq( log(bmi) ~ Ns(age,kn=a.kn) + I((dob-150)/5),
+                              tau = as.numeric(iq),
+                              data=subset( abmi, sex==sx ) ),
+                          subset="dob" ) - 1 ) * 100
+           )
```

```
   user  system elapsed
 164.02    0.01  164.19
```

```
> temp.rel <-
+ function(trsp=TRUE)
+ {
+ par( mar=c(3,3.2,1,1), mgp=c(3,1,0)/1.6 )
+ plot( NA, bty="n", las=1,
+       xlim=c(-2,100), xlab="BMI percentile", xaxs="i",
+       ylim=c(0,3.5), yaxs="i",
+       ylab="Relative change per 5 years (%)" )
+ abline( h=seq(0.5,5,0.5), v=seq(0,100,10), col=gray(0.9) )
+ matlines( qq <- as.numeric(dimnames(pdr)[["qnt"]])*100,
+           cbind( pdr[,"M",], pdr[,"F",] ),
+           lty=c(1,0,0), lwd=c(3,1,1), col=rep( c("blue","red"), each=3 ) )
+ if( trsp ) {
+ polygon( c(qq,rev(qq)), c(pdr[,"M",2], rev(pdr[,"M",3]) ),
+          col=rgb(0,0,1,0.2), border="transparent" )
+ polygon( c(qq,rev(qq)), c(pdr[,"F",2], rev(pdr[,"F",3]) ),
+          col=rgb(1,0,0,0.2), border="transparent" )
+            }
+ else
+ matlines( qq, cbind( pdr[,"M",-1], pdr[,"F",-1] ),
+           lty="23", lwd=1, col=rep( c("blue","red"), each=2 ) )
+ }
> win.metafile(       "art-BMI-relchg.emf", height=7, width=7) ; temp.rel(trsp=FALSE) ; dev.off()
```

```
null device
          1
```

```
> pdf("./graph/BMI-APC-art-BMI-relcht.pdf", height=7, width=7) ; temp.rel(        ) ; dev.off()
```

```
null device
          1
```

```
> pdf("./graph/BMI-APC-art-BMI-relchg.pdf", height=7, width=7) ; temp.rel(trsp=FALSE) ; dev.off()
```

```
null device
          1
```

```
> win.metafile( "art/fig3.emf", height=6, width=10) ;
> par( mfrow=c(1,2), mgp=c(3,1,0)/1.6, las=1 )
> temp.abs(trsp=FALSE) ; temp.rel(trsp=FALSE) ; dev.off()
```

```
null device
          1
```

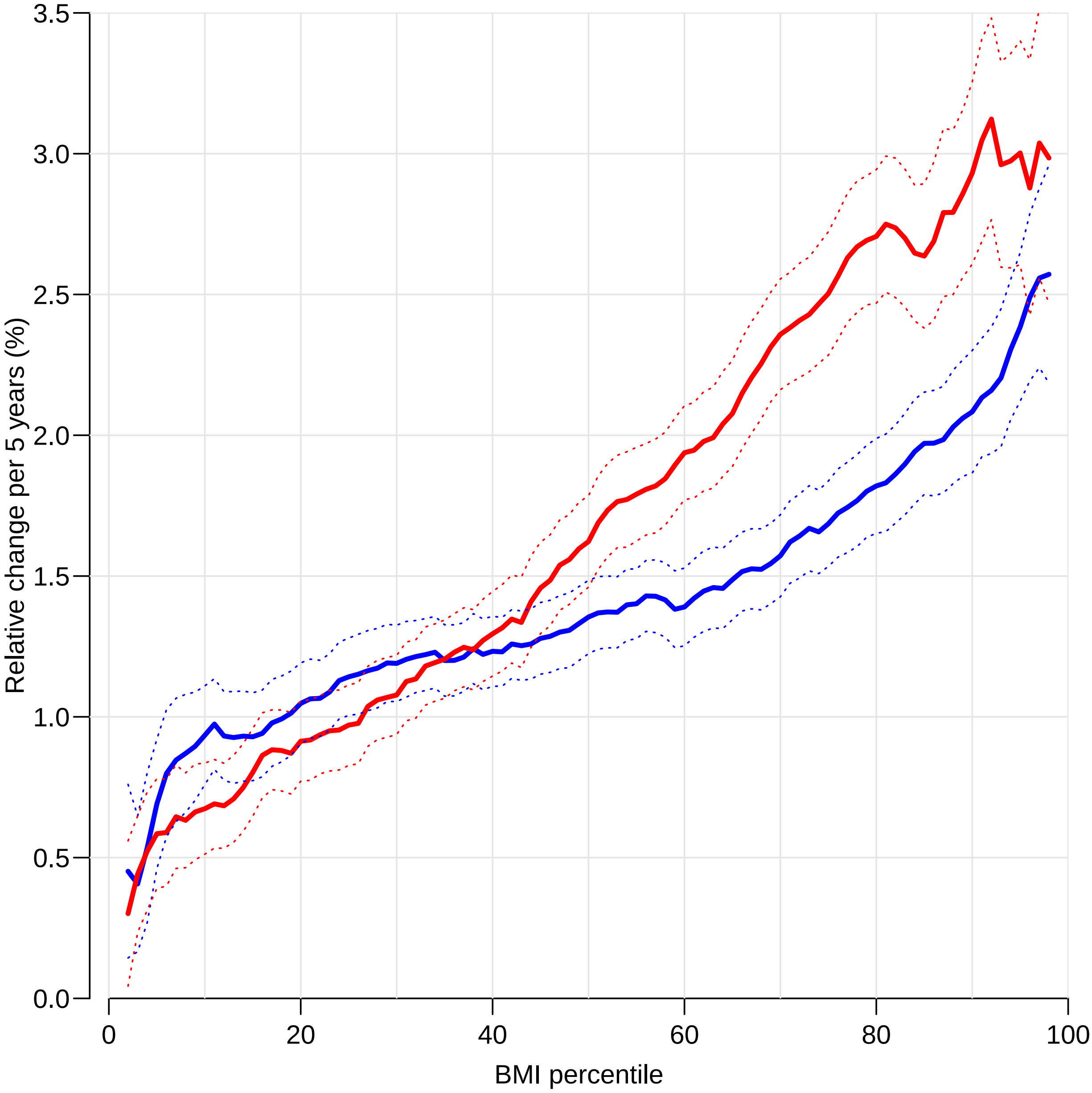

Figure 4.4: *Relative changes in each percentile of the BMI-distribution per 5 years. Thus, for women the 80th percentile has on average changed by 2.8% every 5 years, the 90th by 3%. For men the corresponding figures are 2.0 and 2.5%. Moreover, the 40% leanest in the population has seen an increase in BMI of less than 1.5% per 5 years; significant different from 0, but moderate.*

```
>    postscript("art/fig3.eps", height=6, width=10) ;
> par( mfrow=c(1,2), mgp=c(3,1,0)/1.6, las=1 )
> temp.abs(trsp=FALSE) ; temp.rel(trsp=FALSE) ; dev.off()
```

```
null device
          1
```

```
>          pdf( "art/fig3.pdf", height=6, width=10) ;
> par( mfrow=c(1,2), mgp=c(3,1,0)/1.6, las=1 )
> temp.abs(trsp=FALSE) ; temp.rel(trsp=FALSE) ; dev.off()
```

```
null device
          1
```

```
>          pdf( "art/fig3t.pdf", height=6, width=10) ;
> par( mfrow=c(1,2), mgp=c(3,1,0)/1.6, las=1 )
> temp.abs(trsp=TRUE) ; temp.rel(trsp=TRUE) ; dev.off()
```

```
null device
          1
```

4.3 Models

We now model BMI letting BMI depend additively on age and date of birth, slightly more detailed than in the APC-model before. We could in principle include a more detailed period effect too, by the very nature of data with essentially only 6 different dates of survey (see figure 2.2), this would essentially be modeling survey-specific departures from the overall (linear) trend already included in the models when age and cohort are there. Moreover, there is indeed very little period effct in the data.

4.3.1 Reporting

In order to show the evolution of BMI in the population in more detail than by the estimates from the age-period-cohort model above we will show age-specific BMI quantiles for different birth cohorts, and for different time-points.

The cohort reporting will to a large extent be extrapolations showing how the distribution of BMI will be in a given generation, even if we only have a span of 30 years observation. Hence for each generation we will only have a limited age-range covered. The period reporting will be based on the same models, but when reporting the cross-sectional BMI-distribution we will actually have the entire age-range covered.

So we define arrays to hold the predicted quantiles from regressions involving age, cohort and possibly an interaction. For convenience we create two different sets of arrays, one with a period-classification and one with a cohort classification, but first we get a few utility functions and define the relevant knots for use in the analyses:

Here is the definition of the two arrays:

```
> qnt <- c(5,10,25,50,75,90,95)/100
> q.per <- NArray( list( mod = c("lin","add","per","int"),
+                        sex = levels( abmi$sex ),
+                        qnt = qnt,
+                        age = 10:85,
+                        dos = seq(1980,2010,5) ) )
> q.coh <- NArray( list( mod = c("lin","add","per","int"),
+                        sex = levels( abmi$sex ),
+                        qnt = qnt,
+                        age = 10:85,
+                        dob = seq(1920,1980,10) ) )
> str( q.per )
```

```
 logi [1:4, 1:2, 1:7, 1:76, 1:7] NA NA NA NA NA NA ...
 - attr(*, "dimnames")=List of 5
  ..$ mod: chr [1:4] "lin" "add" "per" "int"
  ..$ sex: chr [1:2] "M" "F"
  ..$ qnt: chr [1:7] "0.05" "0.1" "0.25" "0.5" ...
  ..$ age: chr [1:76] "10" "11" "12" "13" ...
  ..$ dos: chr [1:7] "1980" "1985" "1990" "1995" ...
```

```
> str( q.coh )
```

```
 logi [1:4, 1:2, 1:7, 1:76, 1:7] NA NA NA NA NA NA ...
 - attr(*, "dimnames")=List of 5
  ..$ mod: chr [1:4] "lin" "add" "per" "int"
  ..$ sex: chr [1:2] "M" "F"
  ..$ qnt: chr [1:7] "0.05" "0.1" "0.25" "0.5" ...
  ..$ age: chr [1:76] "10" "11" "12" "13" ...
  ..$ dob: chr [1:7] "1920" "1930" "1940" "1950" ...
```

We also define an array of exactly the same structure as the cohort-classified, with a 1 for those entries (combinations of age an birth cohort) where there are actually survey data.

```
> o.coh <- q.coh*0
> str( o.coh )
```

```
 num [1:4, 1:2, 1:7, 1:76, 1:7] NA NA NA NA NA NA NA NA NA NA ...
 - attr(*, "dimnames")=List of 5
  ..$ mod: chr [1:4] "lin" "add" "per" "int"
  ..$ sex: chr [1:2] "M" "F"
  ..$ qnt: chr [1:7] "0.05" "0.1" "0.25" "0.5" ...
  ..$ age: chr [1:76] "10" "11" "12" "13" ...
  ..$ dob: chr [1:7] "1920" "1930" "1940" "1950" ...
```

```
> for( ia in dimnames( o.coh )[["age"]] )
+ for( ic in dimnames( o.coh )[["dob"]] )
+      o.coh[,,,ia,ic] <- ( as.numeric(ic)+as.numeric(ia)>1978 &
+                           as.numeric(ic)+as.numeric(ia)<2011 )
> o.coh[o.coh==0] <- NA
> table(o.coh,exclude=NULL)
```

```
o.coh
    1  <NA>
11592 18200
```

Finally we define an array to hold the estimated trends in the BMI-quantiles, separately for men and women:

```
> Trend <- NArray( list( est = c("BMI/5y","lower","upper"),
+                        sex = levels( abmi$sex ),
+                        qnt = qnt ) )
> Curve <- NArray( list( est = c("per^2","lower","upper"),
+                        sex = levels( abmi$sex ),
+                        qnt = qnt ) )
```

So now we can fill the arrays with the predictions from the quantile regression models:

```
> aa <- as.numeric(dimnames(q.coh)[["age"]])
> bb <- as.numeric(dimnames(q.coh)[["dob"]])
> ss <- as.numeric(dimnames(q.per)[["dos"]])
> nd.coh <- data.frame( age = rep(aa,     length(bb)),
+                       dob = rep(bb,each=length(aa)) )
> nd.coh <- transform( nd.coh, dos = dob+age )
> nd.per <- data.frame( age = rep(aa,     length(ss)),
+                       dos = rep(ss,each=length(aa)) )
> nd.per <- transform( nd.per, dob = dos-age )
> for( sx in dimnames( q.coh )[["sex"]] )
+ for( iq in dimnames( q.coh )[["qnt"]] )
+    {
+ m.lin <-  rq( bmi ~ Ns(age,kn=a.kn) + dob,
+               tau=as.numeric(iq),
+               data=subset( abmi, sex==sx ) )
+ m.add <-  rq( bmi ~ Ns(age,kn=a.kn) +
+                     Ns(dob,kn=b.kn),
+               tau=as.numeric(iq),
+               data=subset( abmi, sex==sx ) )
+ m.per <-  rq( bmi ~ Ns(age,kn=a.kn) +
+                     Ns(dob,kn=b.kn) +
+                     I(((dos-1990)/10)^2),
+               tau=as.numeric(iq),
+               data=subset( abmi, sex==sx ) )
+ m.int <-  rq( bmi ~ Ns(age,kn=a.kn) +
+                     Ns(dob,kn=b.kn) +
+                     I((age-50)  *(dob-1950)) +
+                     I((age-50)^2*(dob-1950)),
+               tau=as.numeric(iq),
+               data=subset( abmi, sex==sx ) )
+ cf <- summary(m.lin)$coef[grep("dob",names(coef(m.lin))),1:2,drop=F]
+ Trend[,sx,iq] <- (cf*5)  %*% ci.mat()
+ cf <- summary(m.per)$coef[grep("dos",names(coef(m.per))),1:2,drop=F]
+ Curve[,sx,iq] <- cf %*% ci.mat()
+ q.coh["lin",sx,iq,,] <- predict( m.lin, newdata = nd.coh )
+ q.coh["add",sx,iq,,] <- predict( m.add, newdata = nd.coh )
+ q.coh["per",sx,iq,,] <- predict( m.per, newdata = nd.coh )
+ q.coh["int",sx,iq,,] <- predict( m.int, newdata = nd.coh )
+ q.per["lin",sx,iq,,] <- predict( m.lin, newdata = nd.per )
+ q.per["add",sx,iq,,] <- predict( m.add, newdata = nd.per )
+ q.per["per",sx,iq,,] <- predict( m.per, newdata = nd.per )
+ q.per["int",sx,iq,,] <- predict( m.int, newdata = nd.per )
+    }
```

We see that also in the simple model setting, the changes in BMI-quantiles is larger for the higher quantiles, but the increases in the lower ones are also there.

We also checked to see if there was a curvature component (a quadratic) to the period effect, and checked how this changed estimates in the plots:

```
> round( ftable( Curve, col.vars=2:1 ), 3 )
```

```
     sex      M                    F
     est  per^2  lower  upper  per^2  lower  upper
qnt
0.05     -0.179 -0.305 -0.054 -0.238 -0.336 -0.140
```

```
0.1      -0.093 -0.193  0.007 -0.219 -0.303 -0.135
0.25     -0.098 -0.185 -0.012 -0.150 -0.228 -0.072
0.5      -0.085 -0.169 -0.001 -0.204 -0.301 -0.107
0.75      0.044 -0.064  0.152 -0.241 -0.383 -0.098
0.9       0.132 -0.029  0.293 -0.471 -0.715 -0.228
0.95      0.035 -0.179  0.248 -0.526 -0.864 -0.187
```

For the lower quantiles among males and for all quantiles among women there seems to be a negative curvature (steeper increase early, flatter later) but opposite for the quantiles above the median for males. These effects are just detectable in the rightmost panels of figure **??**, but from the figure it is also seen that the effects are very small and of no practical importance.

Once we have filled the arrays `q.coh` and `q.per` with the BMI-quantiles, and we have an indicator of the relevant range of the cohort data in the array `o.coh`, we can plot the quantile predictions and show what parts of the predictions that are directly supported by the data.

```
> par( mfrow=dim(q.coh)[c("sex","qnt")],
+      mar=c(0,0,0,0), mgp=c(3,1,0)/1.6, oma=c(3,3,3,1), las=1 )
> for( sx in 1:dim(q.coh)[["sex"]] )
+ for( iq in 1:dim(q.coh)[["qnt"]] )
+ {
+ plot( NA, xlim=c(20,85), ylim=c(15,45),
+       xaxt="n", yaxt="n", xlab="", ylab="" )
+ if( sx==2 ) axis( side=1 )
+ if( sx==1 ) mtext( 100*as.numeric(dimnames(q.coh)[["qnt"]][iq]),
+                    side=3, line=0.5, cex=0.8 )
+ if( iq==1 ) axis( side=2 )
+ abline( h=seq(0,50,5), v=seq(0,100,10), col=gray(0.85) )
+ matlines( aa, q.coh["add",sx,iq,,]                     , lwd=1, lty=1, col=gray(0.5) )
+ matlines( aa, q.coh["add",sx,iq,,]*o.coh["add",sx,iq,,], lwd=2, lty=1,
+           col=c("blue","red")[sx] )
+ }
> mtext( expression("BMI (kg/"*m^2*")"), side=2, line=2, outer=TRUE,las=0,cex=0.8 )
> mtext( "Age (years)", side=1, line=2, outer=TRUE,las=0,cex=0.8 )
> mtext( "Percentile of the BMI-distribution",
+        side=3, line=1.5, cex=0.8, outer=TRUE )
```

Here is the code that produces the plot for the article:

```
> temp.graph <-
+ function()
+ {
+ par( mfrow=c(2,3), mar=c(0,0,0,0), mgp=c(3,1,0)/1.6, oma=c(3.5,3.5,3,2.5),
+      las=1, cex=1 )
+ for( sx in 1:2 )
+ for( iq in c(2,4,6) )
+ {
+ plot( NA, xlim=c(20,85), ylim=c(15,45),
+       xaxt="n", yaxt="n", xlab="", ylab="" )
+ if( sx==2 ) axis( side=1 )
+ if( sx==1 ) mtext( 100*as.numeric(dimnames(q.coh)[["qnt"]][iq]),
+                    side=3, line=0.5, cex=0.8 )
+ if( iq==2 ) axis( side=2 )
+ if( iq==6 ) axis( side=4 )
+ abline( h=seq(0,50,5), v=seq(0,100,10), col=gray(0.85) )
+ matlines( aa, q.coh["add",sx,iq,,]                     , lwd=1  , lty=1, col=gray(0.5) )
+ matlines( aa, q.coh["add",sx,iq,,]*o.coh["add",sx,iq,,], lwd=3:2, lty=1,
+           col=c("blue","red")[sx] )
+ text( 60, q.coh["add",sx,iq,"60","1920"]*0.99, "1920", adj=c(0,1), col=c("blue","red")[sx] )
+ text( 30, q.coh["add",sx,iq,"30","1980"]*1.01, "1980", adj=c(1,0), col=c("blue","red")[sx] )
+ }
```

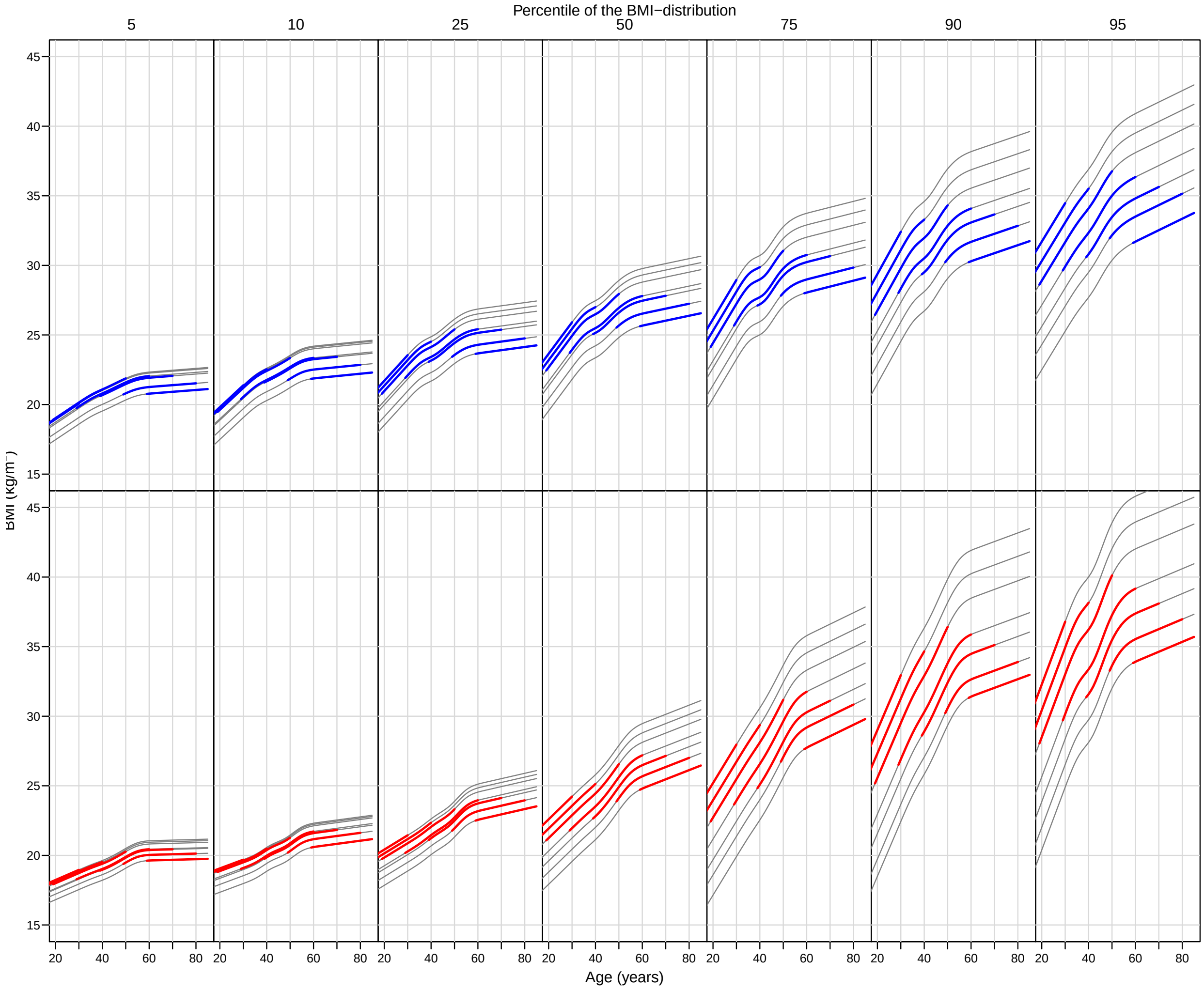

Figure 4.5: *Age-specific percentiles of BMI by sex. Each curve represents a birth cohort (1920, 1930,. . . ,1980). Gray lines are estimates from quantile regression models, the colored part of the lines indicate the ares in which data is available.*
Note that the curves in each panel come from the same model-fit (namely that for the corresponding quantile) and are assumed to be parallel, that is to have the same shape. This is a model which allows a non-linear cohort effect; the model with linear cohort effect would assume that the curves were equidistant.

```
+ mtext( expression("BMI (kg/"*m^2*")"), side=2, line=2, outer=TRUE,las=0 )
+ mtext( "Age (years)", side=1, line=2, outer=TRUE,las=0 )
+ mtext( "Percentile of the BMI-distribution",
+        side=3, line=1.5, cex=0.8, outer=TRUE )
+ }
> temp.graph()
> win.metafile( "art/fig4a.emf", height=10, width=10) ; temp.graph() ; dev.off()
```

```
pdf
  2
```

```
>   postscript( "art/fig4a.eps", height=10, width=10) ; temp.graph() ; dev.off()
```

```
pdf
  2
```

```
>          pdf( "art/fig4a.pdf", height=10, width=10) ; temp.graph() ; dev.off()
```

```
pdf
  2
```

```
> par( mfrow=dim(q.coh)[c("sex","qnt")],
+      mar=c(0,0,0,0), mgp=c(3,1,0)/1.6, oma=c(3,3,3,1), las=1 )
> for( sx in 1:dim(q.coh)[["sex"]] )
+ for( iq in 1:dim(q.coh)[["qnt"]] )
+ {
+ plot( NA, xlim=c(20,85), ylim=c(15,45),
+       xaxt="n", yaxt="n", xlab="", ylab="" )
+ if( sx==2 ) axis( side=1 )
+ if( sx==1 ) mtext( 100*as.numeric(dimnames(q.coh)[["qnt"]][iq]),
+                    side=3, line=0.5, cex=0.8 )
+ if( iq==1 ) axis( side=2 )
+ abline( h=seq(0,50,5), v=seq(0,100,10), col=gray(0.85) )
+ matlines( aa, q.per["add",sx,iq,,], lwd=2, lty=1,
+           col=c("blue","red")[sx] )
+ }
> mtext( expression("BMI (kg/"*m^2*")"), side=2, line=2, outer=TRUE,las=0,cex=0.8 )
> mtext( "Age (years)", side=1, line=2, outer=TRUE,las=0,cex=0.8 )
> mtext( "Percentile of the BMI-distribution",
+        side=3, line=1.5, cex=0.8, outer=TRUE )
```

Here is the code for the plot in the paper

```
> temp.graph <- function()
+ {
+ par( mfrow=c(2,3), mar=c(0,0,0,0), mgp=c(3,1,0)/1.6,
+     oma=c(3.5,3.5,3,2.5), las=1, cex=1 )
+ for( sx in 1:2 )
+ for( iq in c(2,4,6) )
+ {
+ plot( NA, xlim=c(20,85), ylim=c(15,45),
+       xaxt="n", yaxt="n", xlab="", ylab="" )
+ if( sx==2 ) axis( side=1 )
+ if( sx==1 ) mtext( 100*as.numeric(dimnames(q.coh)[["qnt"]][iq]),
+                      side=3, line=0.5 )
+ if( iq==2 ) axis( side=2 )
+ if( iq==6 ) axis( side=4 )
+ abline( h=seq(0,50,5), v=seq(0,100,10), col=gray(0.85) )
```

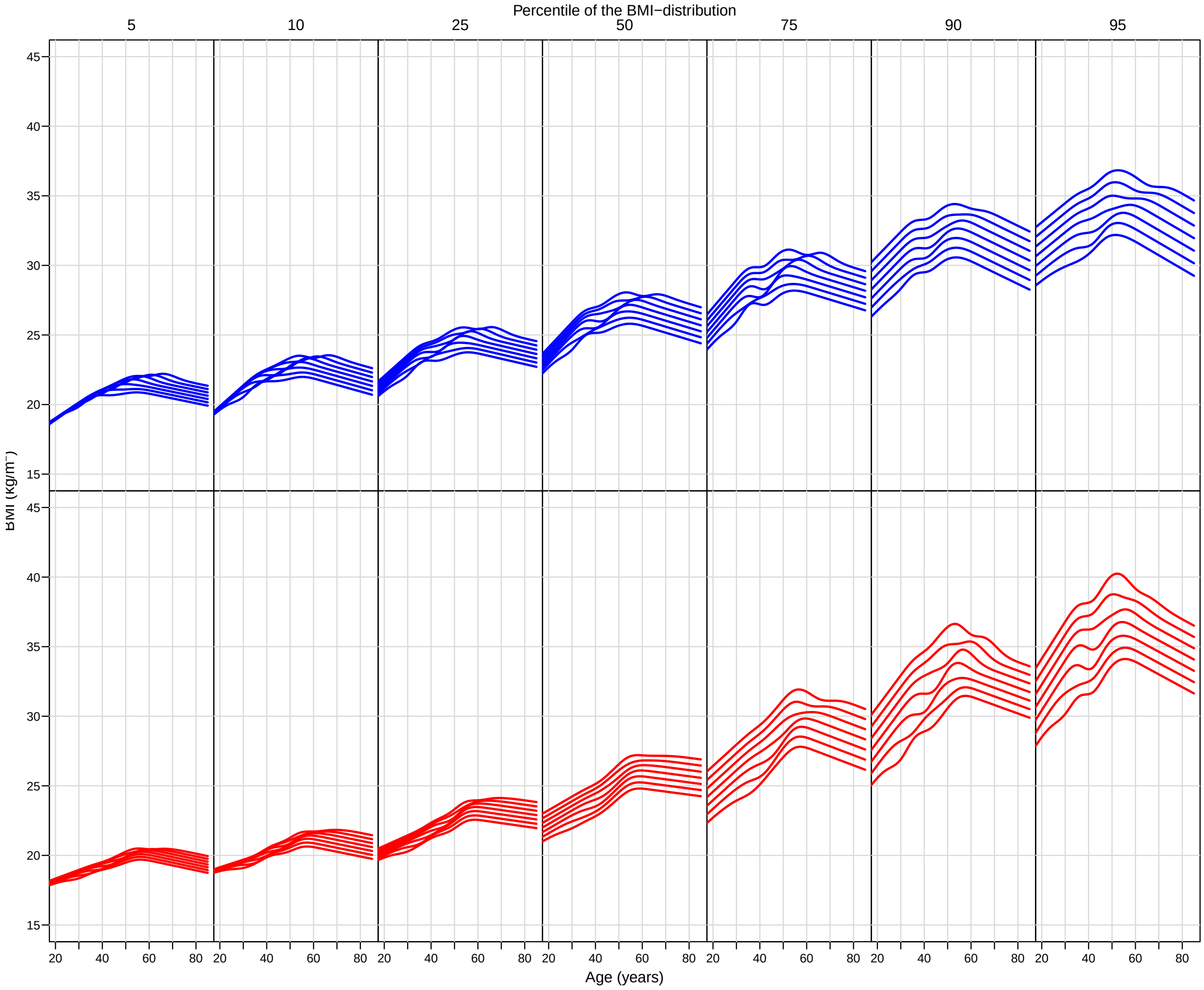

Figure 4.6: *Age-specific percentiles of BMI by sex. Each curve represents a date (1980, 1985,...,2010).*
Note that the curves in each panel come from the same model-fit (namely that for the corresponding quantile) and are assumed to be parallel, that is to have the same shape between cohorts*, but not necessarily between periods.*

```
+ matlines( aa, q.per["add",sx,iq,,], lwd=3:2, lty=1, col=c("blue","red")[sx] )
+ text( 70, q.per["add",sx,iq,"70","1980"]*0.99, "1980", adj=c(1,1), col=c("red","blue")[3-sx] )
+ text( 40, q.per["add",sx,iq,"40","2010"]*1.01, "2010", adj=c(1,0), col=c("red","blue")[3-sx] )
+ }
+ mtext( expression("BMI (kg/"*m^2*")"), side=2, line=2, outer=TRUE,las=0 )
+ mtext( "Age (years)", side=1, line=2, outer=TRUE, las=0)
+ mtext( "Percentile of the BMI-distribution", side=3, line=1.5, outer=TRUE )
+ }
> temp.graph()
> win.metafile( "art/fig4b.emf", height=10, width=10) ; temp.graph() ; dev.off()


windows
      2


>   postscript( "art/fig4b.eps", height=10, width=10) ; temp.graph() ; dev.off()


windows
      2


>          pdf( "art/fig4b.pdf", height=10, width=10) ; temp.graph() ; dev.off()


windows
      2


> par( mfrow=dim(q.coh)[c("sex","dob")],
+      mar=c(0,0,0,0), mgp=c(3,1,0)/1.6, oma=c(4,4,3,1), las=1 )
> for( sx in 1:dim(q.coh)[["sex"]] )
+ for( ib in 1:dim(q.coh)[["dob"]] )
+ {
+ plot( NA, xlim=c(20,85), ylim=c(15,45),
+       xaxt="n", yaxt="n", xlab="", ylab="" )
+ if( sx==2 ) axis( side=1 )
+ if( sx==1 ) mtext( side=3, line=0.5, dimnames(q.coh)[["dob"]][ib], cex=0.8 )
+ if( ib==1 ) axis( side=2 )
+ abline( h=seq(0,50,5), v=seq(0,100,10), col=gray(0.85) )
+ matlines( aa, t(q.coh["add",sx,,,ib])                       , lwd=1, lty=1, col=gray(0.5) )
+ matlines( aa, t(q.coh["add",sx,,,ib]*o.coh["add",sx,,,ib]), lwd=2, lty=1,
+           col=c("blue","red")[sx] )
+ }
> mtext( expression("BMI (kg/"*m^2*")"), side=2, line=2, outer=TRUE,las=0,cex=0.8 )
> mtext( "Age (years)", side=1, line=2, outer=TRUE,las=0,cex=0.8 )
> mtext( "Birth cohort", side=3, line=1.5, cex=0.8, outer=TRUE )


> par( mfrow=dim(q.coh)[c("sex","dob")],
+      mar=c(0,0,0,0), mgp=c(3,1,0)/1.6, oma=c(4,4,3,1), las=1 )
> for( sx in 1:dim(q.coh)[["sex"]] )
+ for( ib in 1:dim(q.coh)[["dob"]] )
+ {
+ plot( NA, xlim=c(20,85), ylim=c(15,45),
+       xaxt="n", yaxt="n", xlab="", ylab="" )
+ if( sx==2 ) axis( side=1 )
+ if( sx==1 ) mtext( side=3, line=0.5, dimnames(q.coh)[["dob"]][ib], cex=0.8 )
+ if( ib==1 ) axis( side=2 )
+ abline( h=seq(0,50,5), v=seq(0,100,10), col=gray(0.85) )
+ matlines( aa, t(q.coh["add",sx,,,ib])                       , lwd=1, lty=1, col=gray(0.5) )
+ matlines( aa, t(q.coh["int",sx,,,ib]*o.coh["int",sx,,,ib]), lwd=2, lty=1,
+           col=c("blue","red")[sx] )
+ }
> mtext( expression("BMI (kg/"*m^2*")"), side=2, line=2, outer=TRUE,las=0,cex=0.8 )
> mtext( "Age (years)", side=1, line=2, outer=TRUE,las=0,cex=0.8 )
> mtext( "Birth cohort", side=3, line=1.5, cex=0.8, outer=TRUE )
```

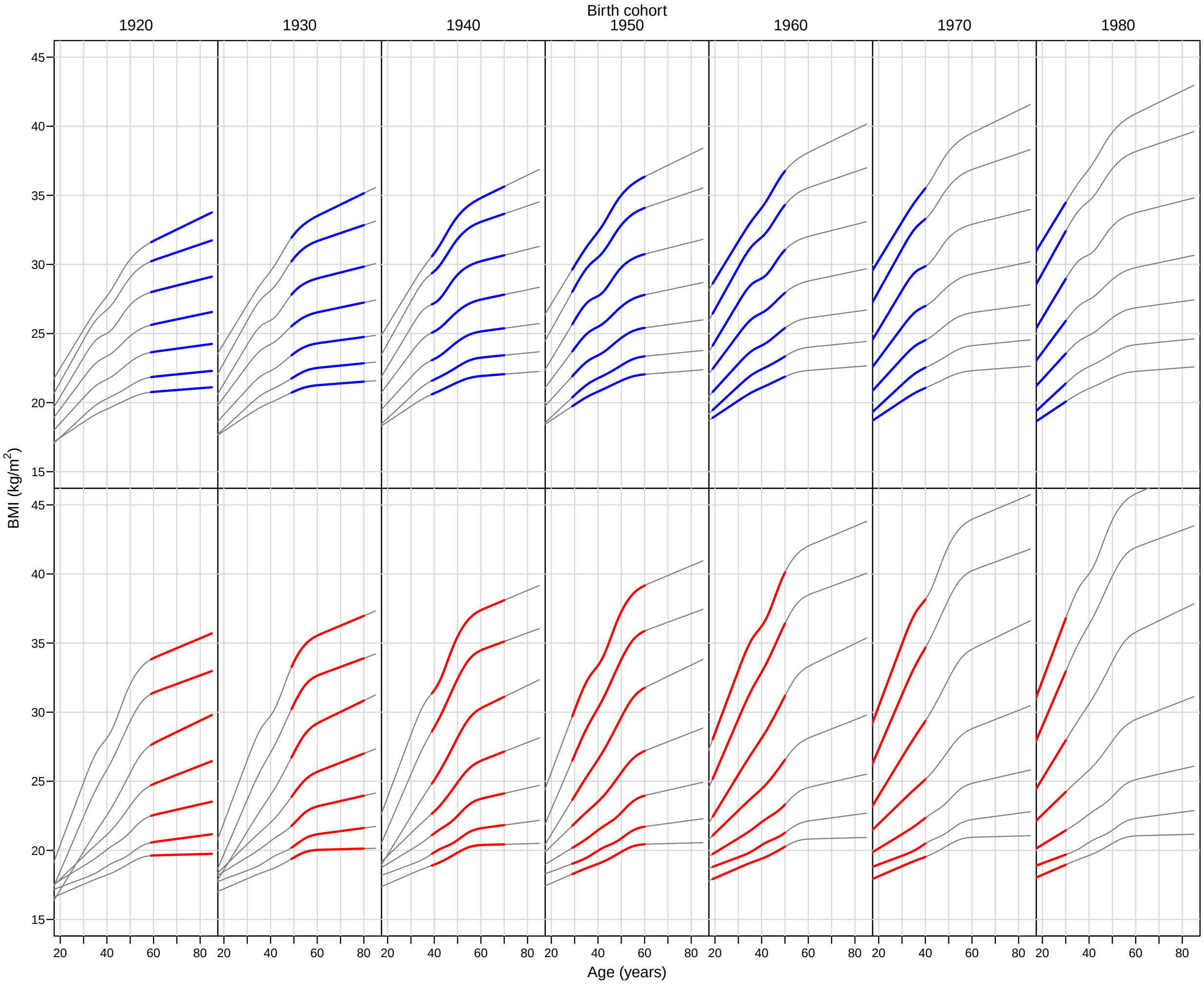

Figure 4.7: *Age-specific percentiles (5,10,25,50,75,90,95) of BMI, by sex and birth cohort. Gray lines are estimates from quantile regression models, the colored part of the lines indicate the ares in which data is available.*
The different lines in each panel are from different models (one model for each quantile).

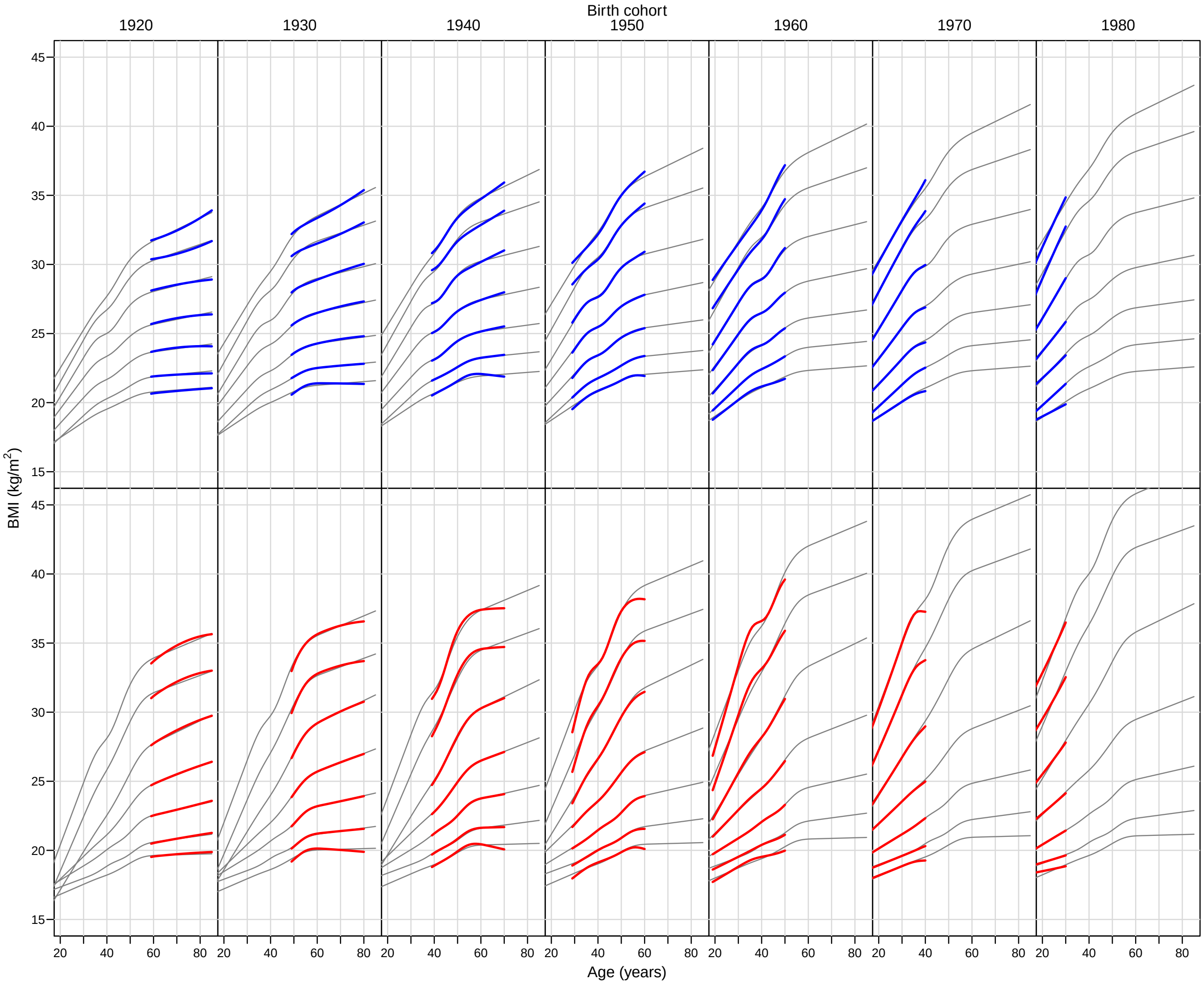

Figure 4.8: *Age-specific percentiles (5,10,25,50,75,90,95) of BMI, by sex and birth cohort. Gray lines are estimates from age-cohort quantile regression models. The colored part of the lines indicate where data is available and show estimates from models with 2-parameter interaction between age and birth cohort.*

```
> par( mfrow=dim(q.coh)[c("sex","dob")],
+      mar=c(0,0,0,0), mgp=c(3,1,0)/1.6, oma=c(4,4,3,1), las=1 )
> for( sx in 1:dim(q.coh)[["sex"]] )
+ for( ib in 1:dim(q.coh)[["dob"]] )
+ {
+ plot( NA, xlim=c(20,85), ylim=c(15,45),
+       xaxt="n", yaxt="n", xlab="", ylab="" )
+ if( sx==2 ) axis( side=1 )
+ if( sx==1 ) mtext( side=3, line=0.5, dimnames(q.coh)[["dob"]][ib], cex=0.8 )
+ if( ib==1 ) axis( side=2 )
+ abline( h=seq(0,50,5), v=seq(0,100,10), col=gray(0.85) )
+ matlines( aa, t(q.coh["add",sx,,,ib])                       , lwd=1, lty=1, col=gray(0.5) )
+ matlines( aa, t(q.coh["per",sx,,,ib]*o.coh["per",sx,,,ib]), lwd=2, lty=1,
+           col=c("blue","red")[sx] )
+ }
> mtext( expression("BMI (kg/"*m^2*")"), side=2, line=2, outer=TRUE,las=0,cex=0.8 )
> mtext( "Age (years)", side=1, line=2, outer=TRUE,las=0,cex=0.8 )
> mtext( "Birth cohort", side=3, line=1.5, cex=0.8, outer=TRUE )
```

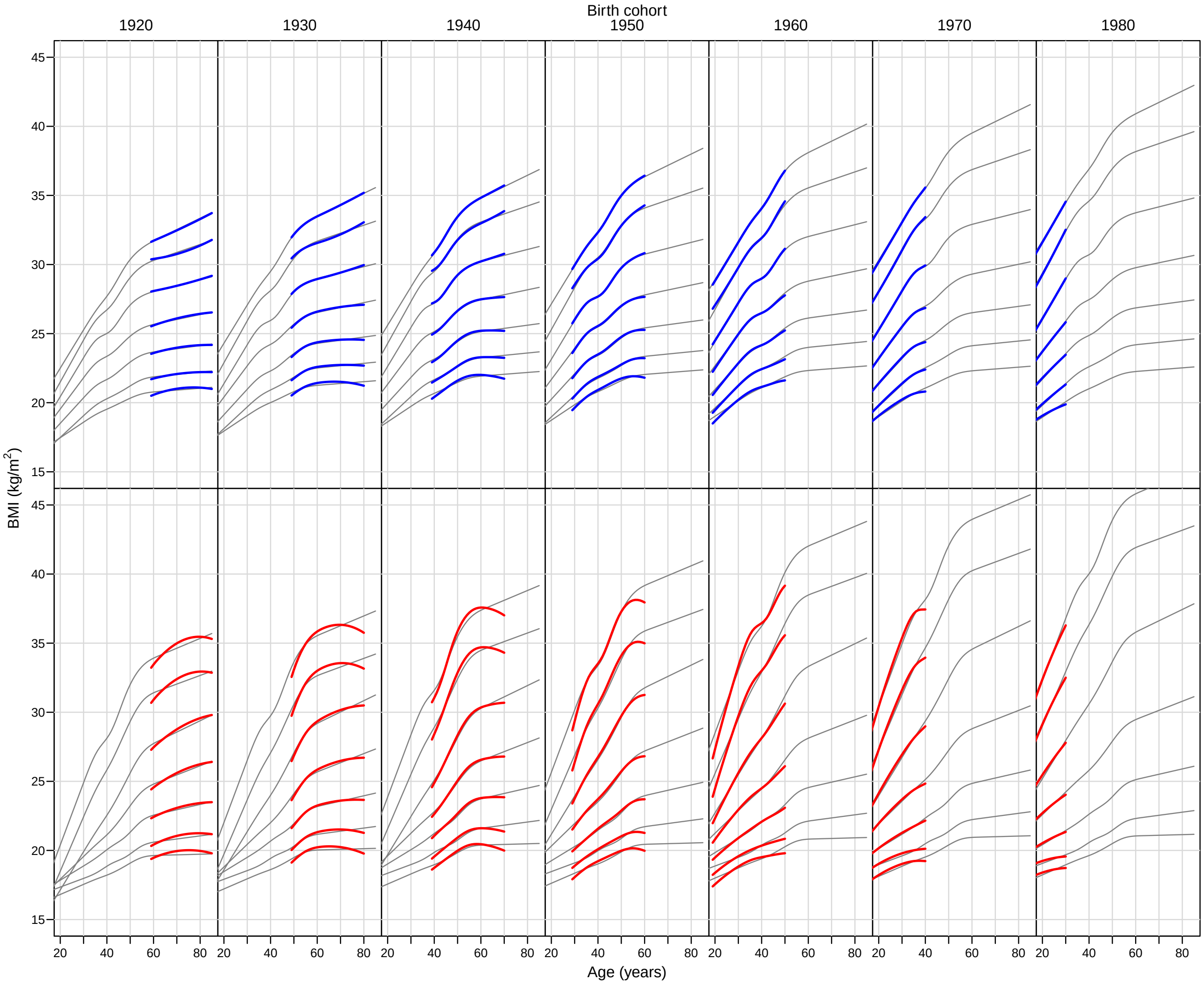

Figure 4.9: *Age-specific percentiles (5,10,25,50,75,90,95) of BMI, by sex and birth cohort. Gray lines are estimates from age-cohort quantile regression models. The colored part of the lines indicate where data is available and show estimates from models with an extra 1-parameter quadratic term in date of survey (period).*

```
> par( mfrow=dim(q.per)[c("sex","dos")],
+      mar=c(0,0,0,0), mgp=c(3,1,0)/1.6, oma=c(4,4,3,1), las=1 )
> for( sx in 1:dim(q.per)[["sex"]] )
+ for( ip in 1:dim(q.per)[["dos"]] )
+ {
+ plot( NA, xlim=c(20,85), ylim=c(15,45),
+       xaxt="n", yaxt="n", xlab="", ylab="" )
+ if( sx==2 ) axis( side=1 )
+ if( sx==1 ) mtext( side=3, line=0.5, dimnames(q.per)[["dos"]][ip], cex=0.8 )
+ if( ip==1 ) axis( side=2 )
+ abline( h=seq(0,50,5), v=seq(0,100,10), col=gray(0.85) )
+ matlines( aa, t(q.per["add",sx,,ip]), lwd=2, lty=1, col=c("blue","red")[sx] )
+ }
> mtext( expression("BMI (kg/"*m^2*")"), side=2, line=2, outer=TRUE,las=0,cex=0.8 )
> mtext( "Age (years)", side=1, line=2, outer=TRUE,las=0,cex=0.8 )
> mtext( "Survey date", side=3, line=1.5, cex=0.8, outer=TRUE )
```

```
> par( mfrow=dim(q.per)[c("sex","dos")],
+      mar=c(0,0,0,0), mgp=c(3,1,0)/1.6, oma=c(4,4,3,1), las=1 )
> for( sx in 1:dim(q.per)[["sex"]] )
+ for( ip in 1:dim(q.per)[["dos"]] )
+ {
```

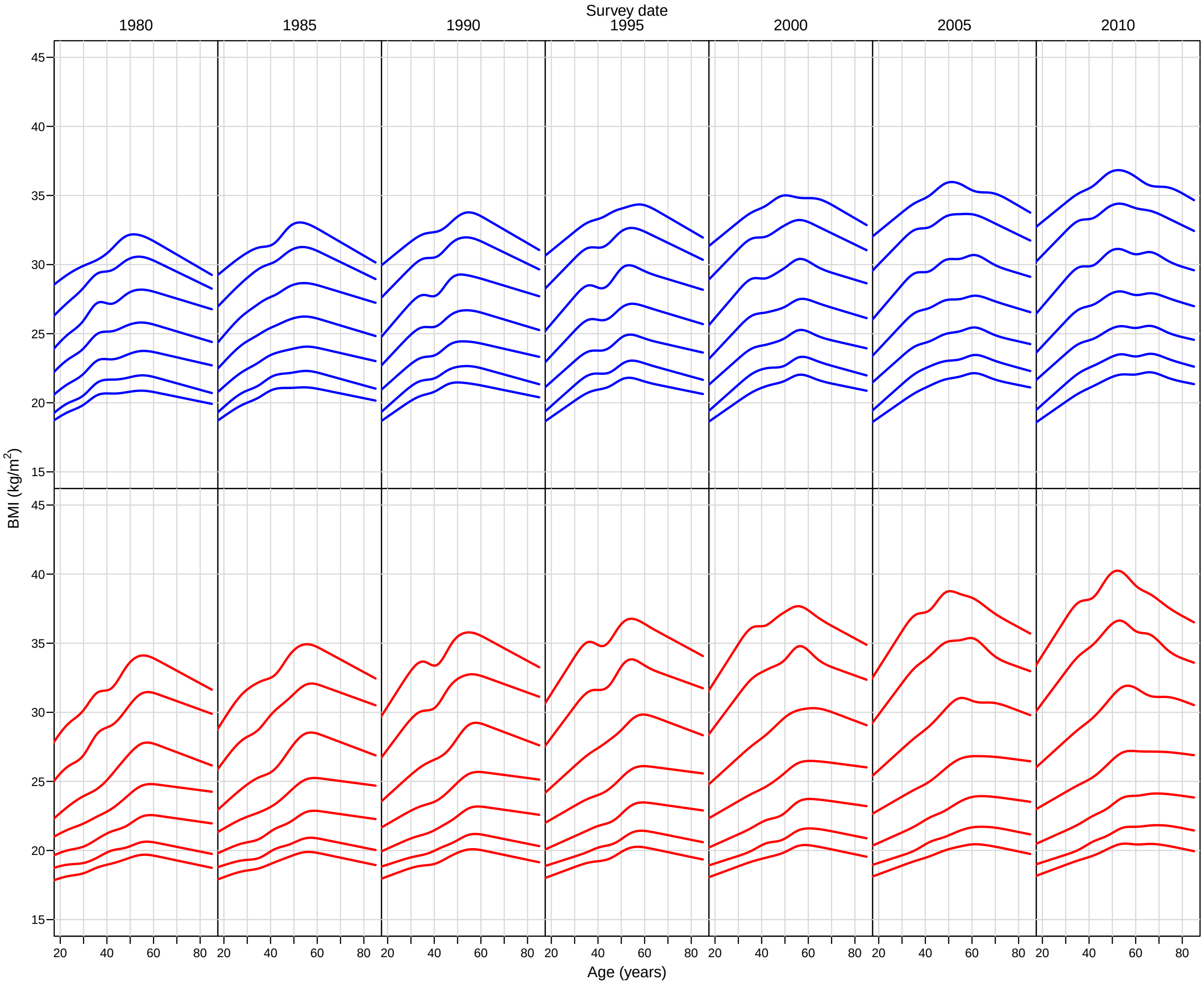

Figure 4.10: *Age-specific percentiles (5,10,25,50,75,90,95) of BMI, by sex and date of survey. The different lines in each panel are from different models (one model for each quantile).*

```
+ plot( NA, xlim=c(20,85), ylim=c(15,45),
+       xaxt="n", yaxt="n", xlab="", ylab="" )
+ if( sx==2 ) axis( side=1 )
+ if( sx==1 ) mtext( side=3, line=0.5, dimnames(q.per)[["dos"]][ip], cex=0.8 )
+ if( ip==1 ) axis( side=2 )
+ abline( h=seq(0,50,5), v=seq(0,100,10), col=gray(0.85) )
+ matlines( aa, t(q.per["add",sx,,ip]), lwd=2, lty=1, col=c("blue","red")[sx] )
+ matlines( aa, t(q.per["per",sx,,ip]), lwd=1, lty=2, col=c("blue","red")[sx] )
+ }
> mtext( expression("BMI (kg/"*m^2*")"), side=2, line=2, outer=TRUE,las=0,cex=0.8 )
> mtext( "Age (years)", side=1, line=2, outer=TRUE,las=0,cex=0.8 )
> mtext( "Survey date", side=3, line=1.5, cex=0.8, outer=TRUE )
```

It is seen from the plots in figure 4.10 that the BMI-quantiles are decreasing after age 60. This is however a phenomenon that is attributable to the fact that BMI is increasing from generation to generation, and we therefore when comparing different generations in a cross-sectional study see a mixture of the "true" age-effect (what happens in a cohort traveling through life) and the cohort effect (change from one cohort to the next).

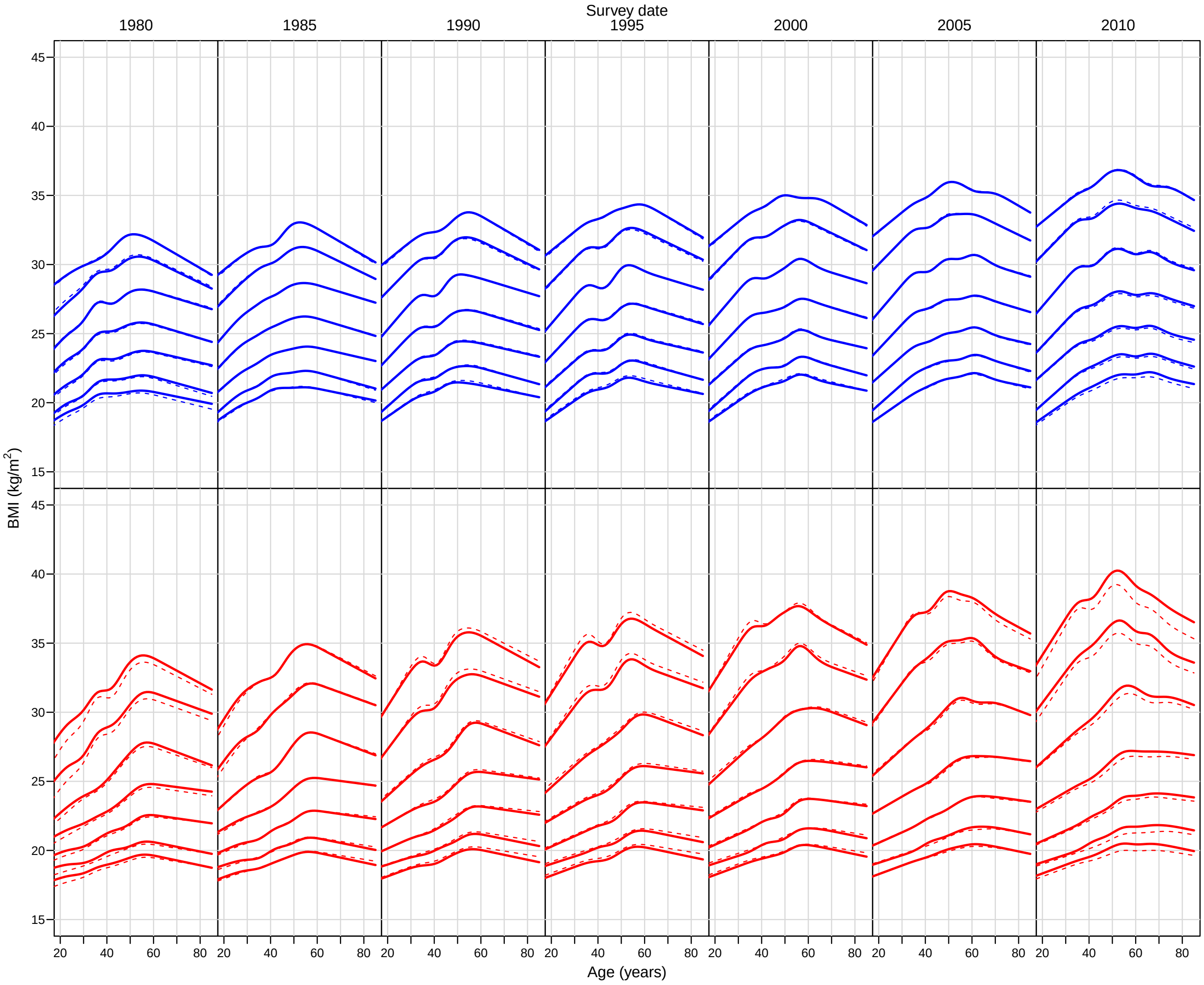

Figure 4.11: *Age-specific percentiles (5,10,25,50,75,90,95) of BMI, by sex and date of survey. The different lines in each panel are from different models (one model for each quantile). The broken lines are the model fit from models with an extra quadratic term in date of survey (period).*

Chapter 5

The prevalence of obesity

The "inverse" question to that answered in the previous chapter where we modelled prespecified quantiles of BMI, is to ask the question of how large a fraction of a given cohort will have a BMI above a certain cutpoint (say 30, 35 or 40 kg/m^2).

This is essentially an interpolation problem, that could be solved by making quantile regression for percentiles 99,98,. . . ,85, say. For a given age and date of birth we could then find out which of these percentiles were above the cutpoint and take that as the estimated percentage. Or slightly more sophisticated, make an interpolation to find the precise probability of being above the cutpoint.

We will do the latter for the age-period-cohort model with the quadratic period term in it. So we set up an array as before to hold the estimated quantiles in the range 99 to 60%, as this is the likely range in which we shall see BMI 30.

5.1 Models

First we load the data ad the knots and the necessary

```
> load( file="./data/knots.Rdata" )
```

Here is the definition of the array to hold the estimated fractions:

```
> qnt <- 60:99/100
> qx.per <- NArray( list( sex = levels( abmi$sex ),
+                         age = 10:85,
+                         dos = 1980:2010,
+                         qnt = qnt ) )
> length( qx.per )


[1] 188480
```

So now we can fill the array with the predictions from the quantile regression models:

```
> aa <- as.numeric(dimnames(qx.per)[["age"]])
> ss <- as.numeric(dimnames(qx.per)[["dos"]])
> nd.per <- data.frame( age = rep(aa,     length(ss)),
+                       dos = rep(ss,each=length(aa)) )
> nd.per <- transform( nd.per, dob = dos-age )
> dim( nd.per )


[1] 2356    3
```

```
> system.time(
+ for( sx in dimnames( qx.per )[["sex"]] )
+    {
+ m.apc <-  rq( bmi ~ Ns(age,kn=a.kn) +
+                     Ns(dob,kn=b.kn) +
+                     I(((dos-1990)/10)^2),
+               tau=qnt,
+               data=subset( abmi, sex==sx ) )
+ qx.per[sx,,,] <- predict( m.apc, newdata=nd.per )
+ } )
```

```
   user  system elapsed
  52.90    0.45   54.60
```

This was in essence fitting the same model as in the previous chapter, however here for a much tighter grid of quantiles (60th–99th percentile).

In order to get estimates of the proportions exceeding some cutpoint, we set up a similar array, however not classified by the quantiles but by the cutpoints on the BMI scale:

```
> cut <- c(30,35,40)
> cx.per <- NArray( list( sex = levels( abmi$sex ),
+                         age = 10:85,
+                         dos = 1980:2010,
+                         cut = cut ) )
> length( cx.per )
```

```
[1] 14136
```

We can then compute the fractions by interpolating between quantiles from `qx.per`. But first we just check that the quantiles are actually monotonely increasing:

```
> table( apply( qx.per, 1:3, FUN=function(x) any(diff(x))<0 ) )
```

```
FALSE
 4712
```

Then we can make the interpolation between the estimated quantiles to derive the fraction below each of the cutpoints by age and calendar time:

```
> int.qn <-
+ function( x, cut )
+ {
+ bl <- max( x[x<cut] )
+ bu <- min( x[x>cut] )
+ ql <- max( qnt[x<cut] )
+ qu <- min( qnt[x>cut] )
+ ql + (qu-ql) * (cut-bl)/(bu-bl)
+ }
> for( ic in dimnames(cx.per)[["cut"]] )
+    cx.per[,,,ic] <- apply( qx.per, 1:3, FUN=int.qn, as.numeric(ic) )
```

Now we have the fractions of persons in a given age and year that are below any one of the three given cutpoints, which enables us to show how these fractions evolve with age, either cross-sectionally or by birth cohort:

```
> temp.graph <-
+ function()
+ {
+ par( mfrow=c(2,3), mar=c(0,0,0,0), oma=c(3,3,3,1), mgp=c(3,1,0)/1.6,
+     las=1, cex=1 )
+ for( sx in  dimnames(cx.per)[["sex"]] )
+ for( ct in  dimnames(cx.per)[["cut"]] )
+ {
+ plot( NA,
+       xlim=c(10,85), xaxs="i", xaxt="n",
+       ylim=c( 0,36), yaxs="i", yaxt="n", bty="o" )
+ abline( h=seq(5,30,5), v=seq(5,85,5), col=gray(0.95) )
+ matlines( as.numeric( dimnames(cx.per)[["age"]] ),
+           prv <- 100*(1-cx.per[sx,,paste(seq(1980,2010,5)),ct]),
+           lty=1, lwd=c(3,2), col=c("red","blue")[1+(sx=="M")] )
+ if( sx=="F"  ) axis( side=1, at=seq(10,80,10), labels=FALSE )
+ if( sx=="F"  ) axis( side=1, at=seq(20,80,20) )
+ if( ct=="30" ) axis( side=2, at=seq(0,30,5) )
+ if( sx=="M" ) mtext( ct, side=3, line=0.5 )
+ text( 70, prv["70","1980"]*0.99, "1980", adj=c(1,1), col=c("red","blue")[1+(sx=="M")] )
+ text( 25, prv["25","2010"]*1.01, "2010", adj=c(1,0), col=c("red","blue")[1+(sx=="M")] )
+ box()
+ }
+ mtext( "Age"              , side=1, outer=TRUE, line=2 )
+ mtext( "% above BMI-limit", side=2, outer=TRUE, line=2, las=0 )
+ mtext( "BMI-limit"        , side=3, outer=TRUE, line=2, las=0 )
+ }
> temp.graph()
> win.metafile( "art/fig5b.emf", height=10, width=10) ; temp.graph() ; dev.off()
```

```
pdf
  2
```

```
>   postscript( "art/fig5b.eps", height=10, width=10) ; temp.graph() ; dev.off()
```

```
pdf
  2
```

```
>          pdf( "art/fig5b.pdf", height=10, width=10) ; temp.graph() ; dev.off()
```

```
pdf
  2
```

```
> temp.graph <-
+ function()
+ {
+ par( mfrow=c(2,3), mar=c(0,0,0,0), oma=c(3,3,3,1), mgp=c(3,1,0)/1.6,
+     las=1, cex=1 )
+ aa <- as.numeric( dimnames(cx.per)[["age"]] )
+ pp <- as.numeric( dimnames(cx.per)[["dos"]] )
+ cc <- seq(1920,1980,10)
+ for( sx in  dimnames(cx.per)[["sex"]] )
+ for( ct in  dimnames(cx.per)[["cut"]] )
+ {
+ # Cohort array initialized to NA
+ cA <- NArray( list( age = aa,
+                     coh = cc ) )
+ # Tease out the probabilities by cohort:
+ for( ic in cc )
+ for( ia in aa )
+    {
+    ip <- ic+ia
+    AA <- paste(ia)
+    PP <- paste(ip)
+    CC <- paste(ic)
+    if( ip<2011 & ip>1979 )
+        cA[AA,CC] <- 100*(1-cx.per[sx,AA,PP,ct])
+    }
+ plot( NA,
+       xlim=c(10,85), xaxs="i", xaxt="n",
+       ylim=c( 0,36), yaxs="i", yaxt="n", bty="o" )
+ abline( h=seq(5,30,5), v=seq(5,85,5), col=gray(0.95) )
+ matlines( aa, cA, lty=1, lwd=3:2, col=c("red","blue")[1+(sx=="M")] )
+ if( sx=="F" )
+   {
+   axis( side=1, at=seq(10,80,10), labels=FALSE )
+   axis( side=1, at=seq(20,80,20) )
+   }
+ if( sx=="M" ) mtext( ct, side=3, line=0.5 )
+ if( ct=="30" ) axis( side=2, at=seq(0,30,5) )
+ text( 70, cA["70","1920"]*0.99, "1920", adj=c(0,1), col=c("red","blue")[1+(sx=="M")] )
+ text( 25, cA["25","1980"]*1.01, "1980", adj=c(1,0), col=c("red","blue")[1+(sx=="M")] )
+ box()
+ }
+ mtext( "Age"              , side=1, outer=TRUE, line=2,       )
+ mtext( "% above BMI-limit", side=2, outer=TRUE, line=2, las=0 )
+ mtext( "BMI-limit"        , side=3, outer=TRUE, line=2, las=0 )
+ }
> temp.graph()
> win.metafile( "art/fig5a.emf", height=10, width=10) ; temp.graph() ; dev.off()
```

```
pdf
  2
```

```
>   postscript( "art/fig5a.eps", height=10, width=10) ; temp.graph() ; dev.off()
```

```
pdf
  2
```

```
>          pdf( "art/fig5a.pdf", height=10, width=10) ; temp.graph() ; dev.off()
```

```
pdf
  2
```

5.2 Bootstrapping confidence intervals

In order to get a handle on the uncertainly of the estimates we set up an array like `cx.per` expanded with an extra dimension of 1000 to hold the results based on each of 1000 bootstrap samples. There are severe memory problems in handling arrays of this size, so we set it up to hold 100 bootstrap samples, fill them in and store on disc. This is repeated 10 times, and the results then retrieved an used for construction of the confidence intervals.

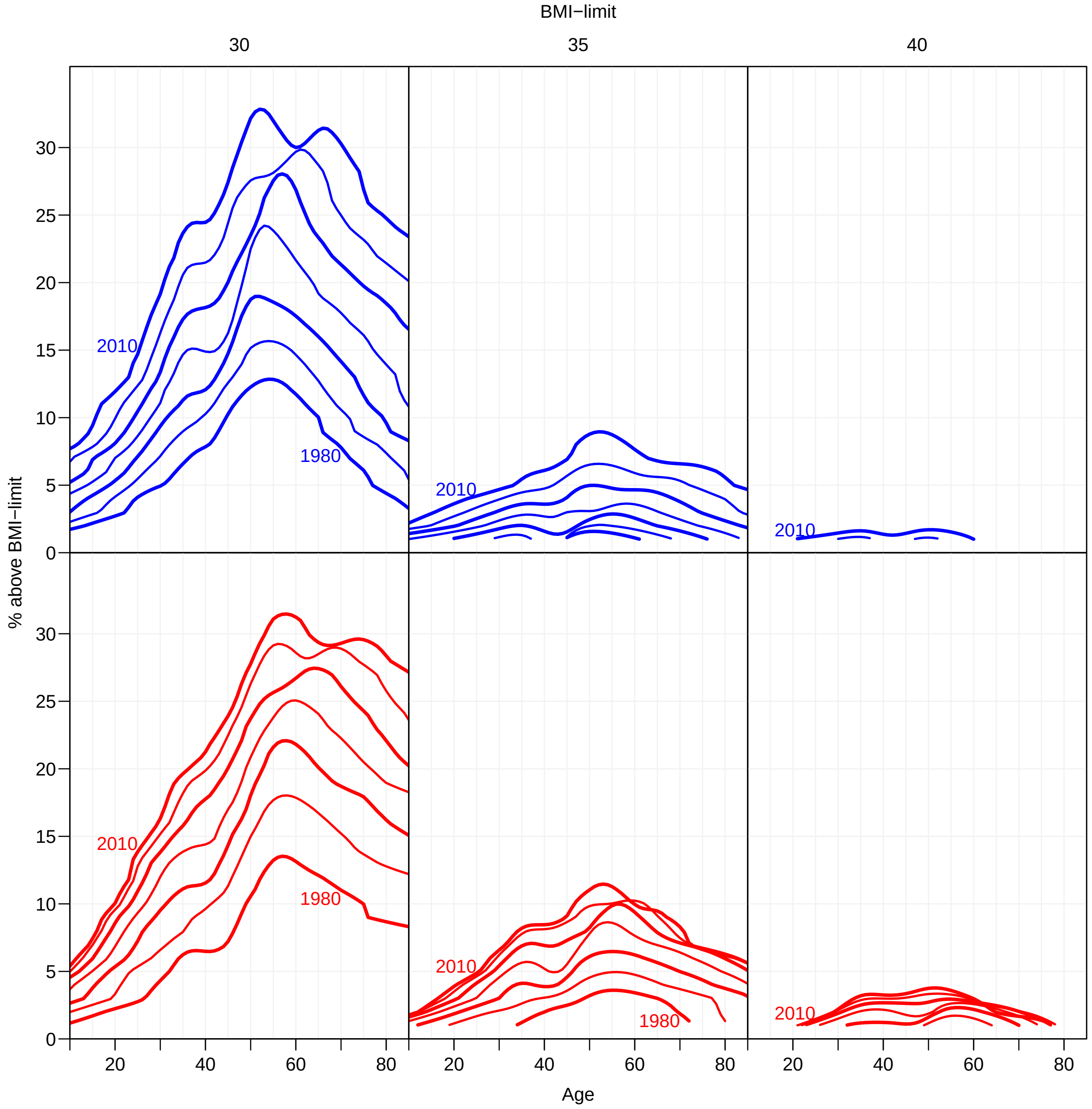

Figure 5.1: *Fraction of persons that exceed BMI 30, 35 and 40 at the dates 1980,1985,...,2010, as a function of age. Blue curves are for males, red for females*

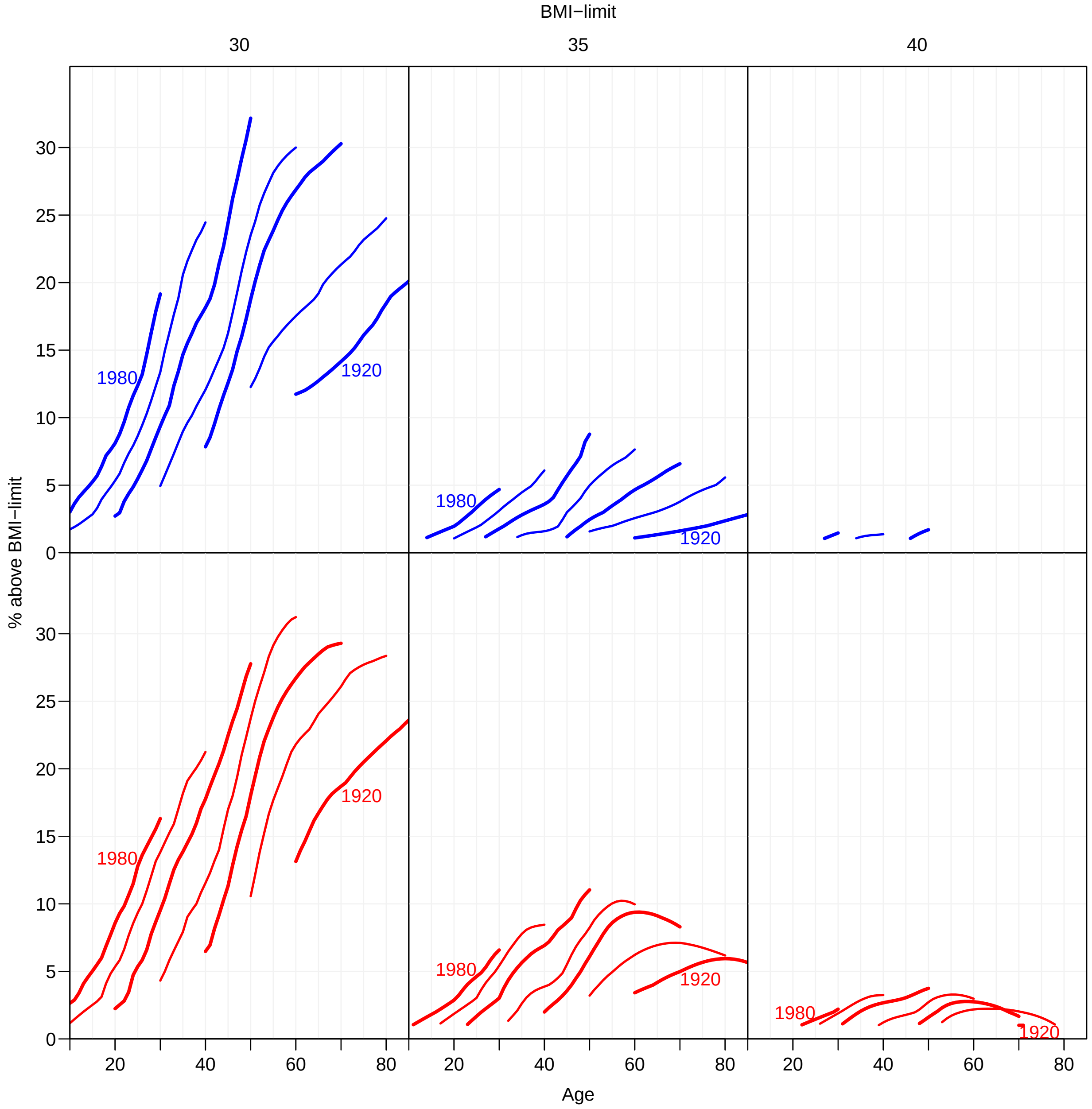

Figure 5.2: *Fraction of persons that exceed BMI 30, 35 and 40 from birth cohorts 1920,1930,...,1980, as a function of age. Blue curves are for males, red for females*

```
> bx.per <- NArray( list( bsm = 1:100,
+                         sex = levels( abmi$sex ),
+                         age = 10:85,
+                         dos = 1980:2010,
+                         cut = cut ) )
> length( bx.per )
```

```
[1] 1413600
```

In order to generate a bootstrap sample of the estimated predictions, we must resample from the datasets, hence we start by defining the function that generates a resampled dataset:

```
> smpl <-
+ function( dfr )
+ dfr[sample( 1:nrow(dfr), nrow(dfr), replace=TRUE ),]
```

So now we can estimate, predict and interpolate for this sample:

```
> snum <- 1:2 #100
> for( bs in snum )
+    {
+ cat( "Bootstrap sample", bs, "start at", format(Sys.time(), "%d-%m-%Y %X"), "\n" )
+ flush.console()
+ for( sx in dimnames( qx.per )[["sex"]] )
+    {
+ m.apc <-  rq( bmi ~ Ns(age,kn=a.kn) +
+                     Ns(dob,kn=b.kn) +
+                     I(((dos-1990)/10)^2),
+               tau=qnt,
+               data = smpl( subset(abmi,sex==sx) ) )
+ qx.per[sx,,,] <- predict( m.apc, newdata=nd.per )
+    }
+ for( ic in dimnames(bx.per)[["cut"]] )
+    {
+    ct <- as.numeric( ic )
+    bx.per[bs,,,,ic] <- apply( qx.per, 1:3, FUN=int.qn, ct )
+    }
+ cat( "Bootstrap sample", bs, "done  at", format(Sys.time(), "%d-%m-%Y %X"), "\n" )
+ flush.console()
+    }
```

```
Bootstrap sample 1 start at 11-11-2013 17:03:05
Bootstrap sample 1 done  at 11-11-2013 17:03:59
Bootstrap sample 2 start at 11-11-2013 17:03:59
Bootstrap sample 2 done  at 11-11-2013 17:04:52
```

```
> save( bx.per, file="./data/bx.perx.Rdata" )
> # save( bx.per, file="./data/bx.per9.Rdata" )
> # save( bx.per, file="./data/bx.per8.Rdata" )
> # save( bx.per, file="./data/bx.per7.Rdata" )
> # save( bx.per, file="./data/bx.per6.Rdata" )
> # save( bx.per, file="./data/bx.per5.Rdata" )
> # save( bx.per, file="./data/bx.per4.Rdata" )
> # save( bx.per, file="./data/bx.per3.Rdata" )
> # save( bx.per, file="./data/bx.per2.Rdata" )
> # save( bx.per, file="./data/bx.per1.Rdata" )
> # save( bx.per, file="./data/bx.per0.Rdata" )
```

We have done the bootstrapping in chunks of 100 resamplings, so now we need to assemble these in one array.

```
> Bx.per <- NArray( list( bsm = 1:1000,
+                         sex = levels( abmi$sex ),
+                         age = 10:85,
+                         dos = 1980:2010,
+                         cut = cut ) )
> length( Bx.per )
```

```
[1] 14136000
```

```
> for( i in 0:9 )
+    {
+    load( file=paste("./data/bx.per",i,".Rdata",sep="") )
+    bx.per[is.na(bx.per)] <- 1
+    Bx.per[i*100+1:100,,,,] <- 1-bx.per
+    }
> table( Bx.per==0, exclude=NULL )
```

```
   FALSE     TRUE     <NA>
10263671  3872329        0
```

Now we have all bootstrap samples in one array, and we now want the upper and lower bootstrap confidence limits. For the sake of completeness we also compute the median of the bootstrap samples to check whether this is in reasonable accordance with the estimate:

```
> Cx.per <- NArray( list( lim = c("Est","med","lo","hi"),
+                         sex = levels( abmi$sex ),
+                         age = 10:85,
+                         dos = 1980:2010,
+                         cut = cut ) )
> length( Cx.per )
```

```
[1] 56544
```

```
> Cx.per["Est",,,,] <- 1-cx.per
> Cx.per[-1,,,,] <- apply( Bx.per[,,,,],
+                          2:5,
+                          quantile,
+                          probs=c(500,25,975)/1000,
+                          na.rm=TRUE )
> # Remove estimates that are outside the interpolation range
> Cx.per[Cx.per<0.01] <- NA
```

Once we have the estimated prevalences with confidence intervals, we can repeat the plots we did previously, but now including confidence limits to allow us to see how precise the prevalence predictions are:

```
> temp.graph <-
+ function(cl=2)
+ {
+ par( mfrow=c(2,3), mar=c(0,0,0,0), oma=c(3,3,3,1), mgp=c(3,1,0)/1.6,
+      las=1, cex=1 )
+ for( sx in  dimnames(cx.per)[["sex"]] )
+ for( ct in  dimnames(cx.per)[["cut"]] )
+ {
+ plot( NA,
+       xlim=c(10,85), xaxs="i", xaxt="n",
+       ylim=c( 0,36), yaxs="i", yaxt="n", bty="o" )
+ abline( h=seq(5,30,5), v=seq(5,85,5), col=gray(0.95) )
+ matlines( as.numeric( dimnames(Cx.per)[["age"]] ),
+           100*cbind(t(Cx.per[-1,sx,,"1980",ct]),
+                     t(Cx.per[-1,sx,,"1985",ct]),
+                     t(Cx.per[-1,sx,,"1990",ct]),
+                     t(Cx.per[-1,sx,,"1995",ct]),
+                     t(Cx.per[-1,sx,,"2000",ct]),
+                     t(Cx.per[-1,sx,,"2005",ct]),
+                     t(Cx.per[-1,sx,,"2010",ct])),
+           lty=c(1,cl,cl), lwd=c(3,1,1,2,1,1), col=c("red","blue")[1+(sx=="M")] )
+ if( sx=="F"  ) axis( side=1, at=seq(10,80,10), labels=FALSE )
+ if( sx=="F"  ) axis( side=1, at=seq(20,80,20) )
+ if( ct=="30" ) axis( side=2, at=seq(0,30,5) )
+ if( sx=="M" ) mtext( ct, side=3, line=0.5 )
+ text( 70, Cx.per["med",sx,"70","1980",ct]* 99, "1980", adj=c(1,1), col=c("red","blue")[1+(sx=="M")
+ text( 25, Cx.per["med",sx,"25","2010",ct]*101, "2010", adj=c(1,0), col=c("red","blue")[1+(sx=="M")
+ box()
+ }
+ mtext( "Age"              , side=1, outer=TRUE, line=2 )
+ mtext( "% above BMI-limit", side=2, outer=TRUE, line=2, las=0 )
+ mtext( "BMI-limit"        , side=3, outer=TRUE, line=2, las=0 )
+ }
> temp.graph()
> win.metafile( "art/fig5b.ci.emf", height=10, width=10) ; temp.graph() ; dev.off()
```

```
pdf
  2
```

```
>   postscript( "art/fig5b.ci.eps", height=10, width=10) ; temp.graph() ; dev.off()
```

```
pdf
  2
```

```
>          pdf( "art/fig5b.ci.pdf", height=10, width=10) ; temp.graph() ; dev.off()
```

```
pdf
  2
```

```
> temp.graph <-
+ function(cl=2)
+ {
+ par( mfrow=c(2,3), mar=c(0,0,0,0), oma=c(3,3,3,1), mgp=c(3,1,0)/1.6,
+     las=1, cex=1 )
+ aa <- as.numeric( dimnames(cx.per)[["age"]] )
+ pp <- as.numeric( dimnames(cx.per)[["dos"]] )
+ cc <- seq(1920,1980,10)
+ for( sx in  dimnames(cx.per)[["sex"]] )
+ for( ct in  dimnames(cx.per)[["cut"]] )
+ {
+ # Cohort array initialized to NA
+ cA <- NArray( list( est = c("Est","med","lo","hi"),
+                     age = aa,
+                     coh = cc ) )
+ # Tease out the probabilities by cohort:
+ for( ic in cc )
+ for( ia in aa )
+    {
+    ip <- ic+ia
+    AA <- paste(ia)
+    PP <- paste(ip)
+    CC <- paste(ic)
+    if( ip<2011 & ip>1979 )
+        cA[,AA,CC] <- 100*Cx.per[,sx,AA,PP,ct]
+    }
+ plot( NA,
+       xlim=c(10,85), xaxs="i", xaxt="n",
+       ylim=c( 0,36), yaxs="i", yaxt="n", bty="o" )
+ abline( h=seq(5,30,5), v=seq(5,85,5), col=gray(0.95) )
+ for( i in c(2,3,4) )
+ matlines( aa, cA[i,,], lty=if(i>2.1) cl else 1,
+                        lwd=if(i>2.1) 1 else 3:2, col=c("red","blue")[1+(sx=="M")] )
+ if( sx=="F"  ) axis( side=1, at=seq(10,80,10), labels=FALSE )
+ if( sx=="F"  ) axis( side=1, at=seq(20,80,20) )
+ if( sx=="M" ) mtext( ct, side=3, line=0.5 )
+ if( ct=="30" ) axis( side=2, at=seq(0,30,5) )
+ text( 25, cA[2,"25","1980"]*0.99, "1980", adj=c(1,0), col=c("red","blue")[1+(sx=="M")] )
+ text( 70, cA[2,"70","1920"]*1.01, "1920", adj=c(0,1), col=c("red","blue")[1+(sx=="M")] )
+ box()
+ }
+ mtext( "Age"              , side=1, outer=TRUE, line=2 )
+ mtext( "% above BMI-limit", side=2, outer=TRUE, line=2, las=0 )
+ mtext( "BMI-limit"        , side=3, outer=TRUE, line=2, las=0 )
+ }
> temp.graph()
> win.metafile( "art/fig5a.ci.emf", height=10, width=10) ; temp.graph() ; dev.off()
```

```
pdf
  2
```

```
>   postscript( "art/fig5a.ci.eps", height=10, width=10) ; temp.graph() ; dev.off()
```

```
pdf
  2
```

```
>          pdf( "art/fig5a.ci.pdf", height=10, width=10) ; temp.graph() ; dev.off()
```

```
pdf
  2
```

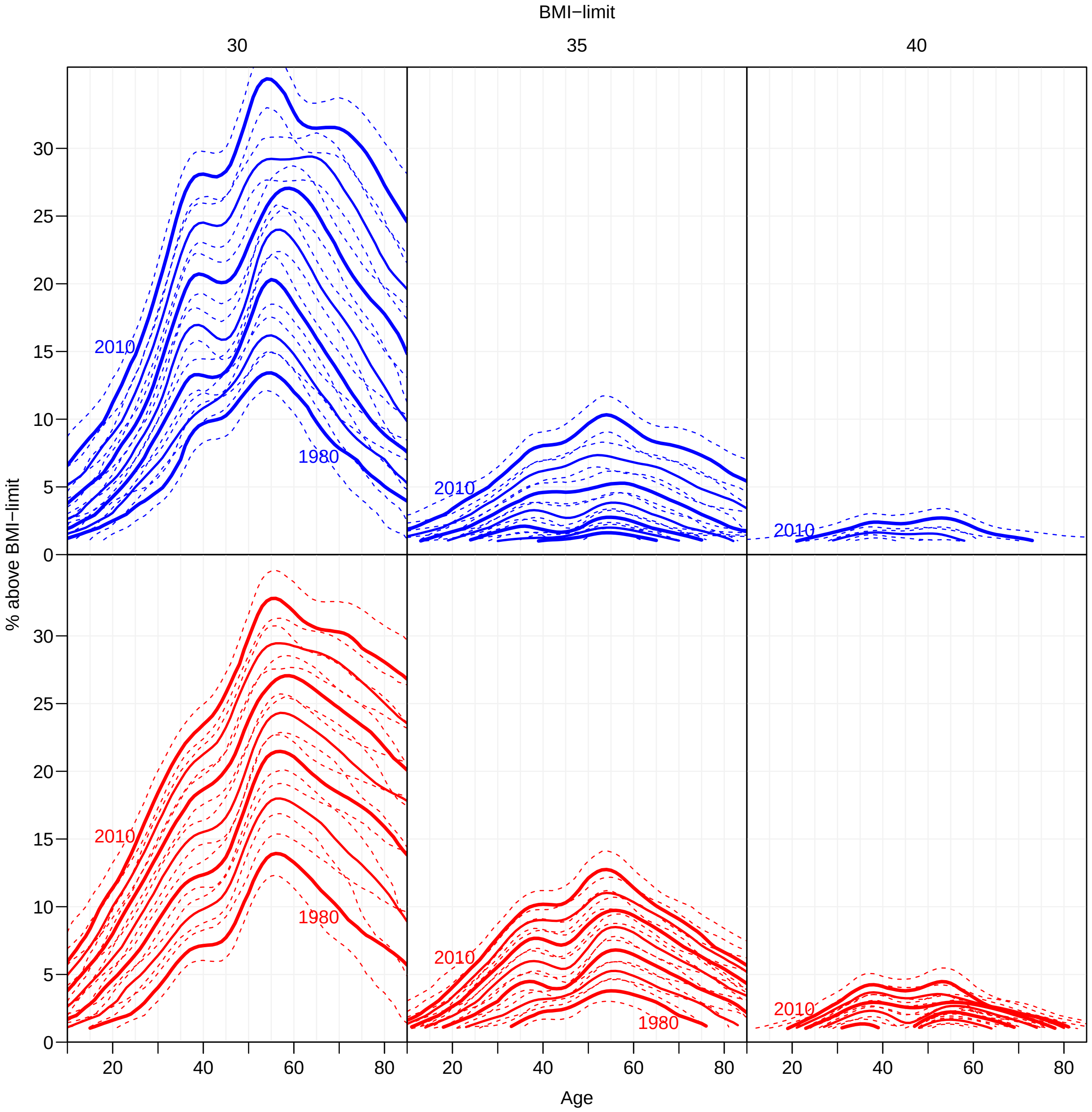

Figure 5.3: *Fraction of persons that exceed BMI 30, 35 and 40 at the dates 1980,1985,...,2010, as a function of age. Blue curves are for males, red for females. Full curves shows the median from a bootstrap sample of 1000, dotted curves the 2.5th and 97.5th percentiles.*

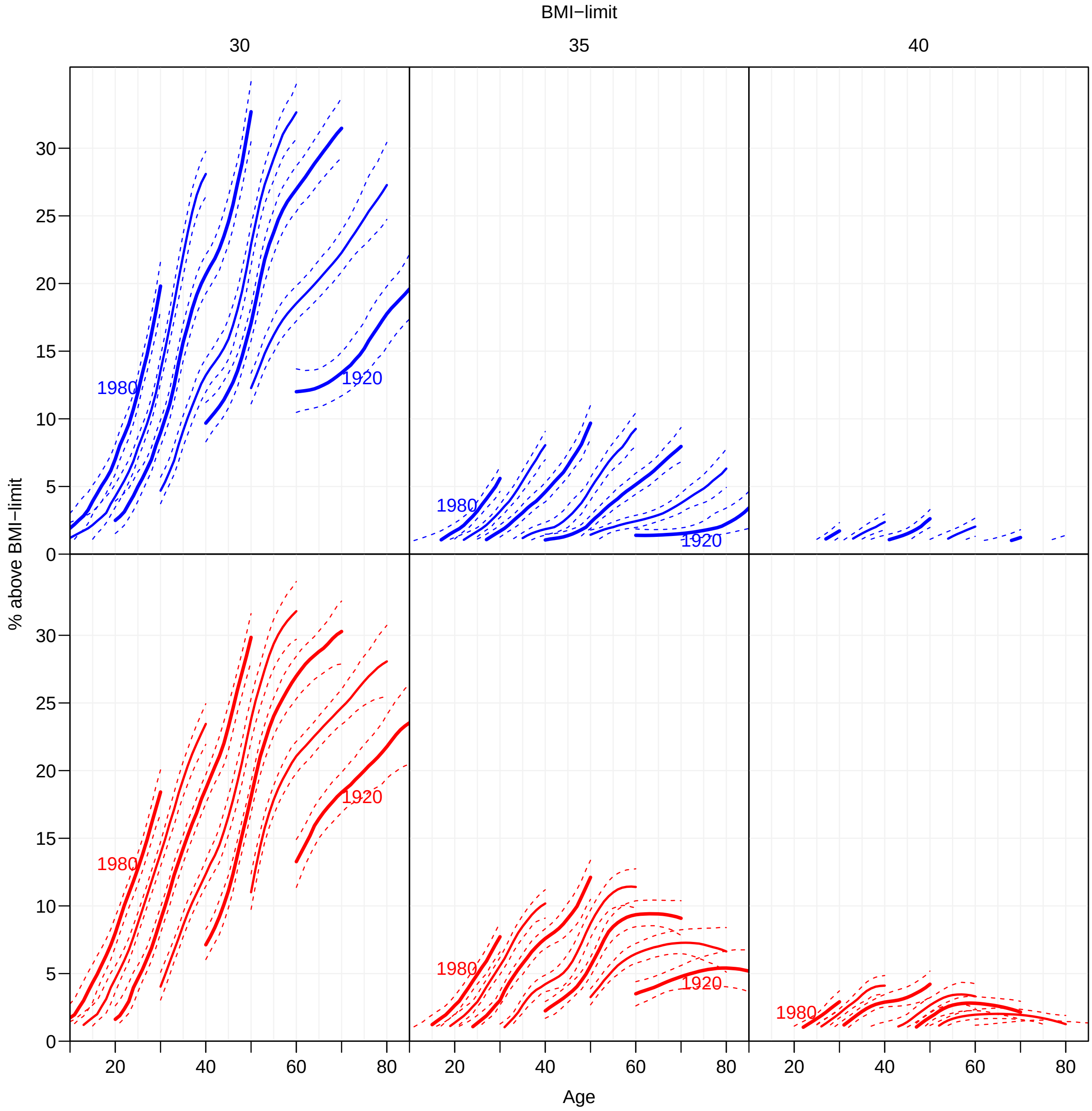

Figure 5.4: *Fraction of persons that exceed BMI 30, 35 and 40 from birth cohorts 1920,1930,. . . ,1980, as a function of age. Blue curves are for males, red for females. Full curves shows the median from a bootstrap sample of 1000, dotted curves the 2.5th and 97.5th percentiles.*

Chapter 6

Quantile regression by education status

Here we repeat the analyses and reporting in the previous chapter, but subdivided not only by sex but also by education, high (at least completed high school or low (less than this). Since this variable is only meaningful after age approximately 20, we restrict these analyses to ages at survey over 20:

```
> abmi <- subset( abmi, age>20 )
```

The analysis now procedes pretty much as the previous, but with an extra dimension to all arrays, so here is the definition of the two arrays:

```
> qnt <- c(5,10,25,50,75,90,95)/100
> q.per <- NArray( list( mod = c("lin","add","per","int"),
+                         sex = levels( abmi$sex ),
+                         edu = levels( abmi$edu ),
+                         qnt = qnt,
+                         age = 10:85,
+                         dos = seq(1980,2010,5) ) )
> q.coh <- NArray( list( mod = c("lin","add","per","int"),
+                         sex = levels( abmi$sex ),
+                         edu = levels( abmi$edu ),
+                         qnt = qnt,
+                         age = 10:85,
+                         dob = seq(1920,1980,10) ) )
> str( q.per )

 logi [1:4, 1:2, 1:2, 1:7, 1:76, 1:7] NA NA NA NA NA NA ...
 - attr(*, "dimnames")=List of 6
  ..$ mod: chr [1:4] "lin" "add" "per" "int"
  ..$ sex: chr [1:2] "M" "F"
  ..$ edu: chr [1:2] "lo" "hi"
  ..$ qnt: chr [1:7] "0.05" "0.1" "0.25" "0.5" ...
  ..$ age: chr [1:76] "10" "11" "12" "13" ...
  ..$ dos: chr [1:7] "1980" "1985" "1990" "1995" ...

> str( q.coh )

 logi [1:4, 1:2, 1:2, 1:7, 1:76, 1:7] NA NA NA NA NA NA ...
 - attr(*, "dimnames")=List of 6
  ..$ mod: chr [1:4] "lin" "add" "per" "int"
  ..$ sex: chr [1:2] "M" "F"
  ..$ edu: chr [1:2] "lo" "hi"
  ..$ qnt: chr [1:7] "0.05" "0.1" "0.25" "0.5" ...
  ..$ age: chr [1:76] "10" "11" "12" "13" ...
  ..$ dob: chr [1:7] "1920" "1930" "1940" "1950" ...
```

We also define an array of exactly the same structure as the cohort-classified, with a 1 for those entries (combinations of age and birth cohort) where there are actually survey data.

```
> o.coh <- q.coh*0
> str( o.coh )
```

```
 num [1:4, 1:2, 1:2, 1:7, 1:76, 1:7] NA NA NA NA NA NA NA NA NA NA ...
 - attr(*, "dimnames")=List of 6
  ..$ mod: chr [1:4] "lin" "add" "per" "int"
  ..$ sex: chr [1:2] "M" "F"
  ..$ edu: chr [1:2] "lo" "hi"
  ..$ qnt: chr [1:7] "0.05" "0.1" "0.25" "0.5" ...
  ..$ age: chr [1:76] "10" "11" "12" "13" ...
  ..$ dob: chr [1:7] "1920" "1930" "1940" "1950" ...
```

```
> for( ia in dimnames( o.coh )[["age"]] )
+ for( ic in dimnames( o.coh )[["dob"]] )
+      o.coh[,,,,ia,ic] <- ( as.numeric(ic)+as.numeric(ia)>1979 &
+                            as.numeric(ic)+as.numeric(ia)<2011 )
> o.coh[o.coh==0] <- NA
> table(o.coh,exclude=NULL)
```

```
o.coh
    1  <NA>
22624 36960
```

We also define an array to hold the estimated trends in the BMI-quantiles, separately for men and women:

```
> Trend <- NArray( list( est = c("BMI/5y","lower","upper"),
+                        sex = levels( abmi$sex ),
+                        edu = levels( abmi$edu ),
+                        qnt = qnt ) )
> Curve <- NArray( list( est = c("per^2","lower","upper"),
+                        sex = levels( abmi$sex ),
+                        edu = levels( abmi$edu ),
+                        qnt = qnt ) )
```

So now we can fill the arrays with the predictions from the quantile regression models:

```
> aa <- as.numeric(dimnames(q.coh)[["age"]])
> bb <- as.numeric(dimnames(q.coh)[["dob"]])
> ss <- as.numeric(dimnames(q.per)[["dos"]])
> nd.coh <- data.frame( age = rep(aa,     length(bb)),
+                       dob = rep(bb,each=length(aa)) )
> nd.coh <- transform( nd.coh, dos = dob+age )
> nd.per <- data.frame( age = rep(aa,     length(ss)),
+                       dos = rep(ss,each=length(aa)) )
> nd.per <- transform( nd.per, dob = dos-age )
> for( sx in dimnames( q.coh )[["sex"]] )
+ for( ed in dimnames( q.coh )[["edu"]] )
+ for( iq in dimnames( q.coh )[["qnt"]] )
+    {
+ m.lin <-  rq( bmi ~ Ns(age,kn=a.kn) + dob,
+               tau=as.numeric(iq),
+               data=subset( abmi, sex==sx & edu==ed ) )
+ m.add <-  rq( bmi ~ Ns(age,kn=a.kn) +
+                     Ns(dob,kn=b.kn),
+               tau=as.numeric(iq),
+               data=subset( abmi, sex==sx & edu==ed ) )
+ m.per <-  rq( bmi ~ Ns(age,kn=a.kn) +
+                     Ns(dob,kn=b.kn) +
+                     I(((dos-1990)/5)^2),
+                 tau=as.numeric(iq),
+                 data=subset( abmi, sex==sx & edu==ed ) )
+ m.int <-  rq( bmi ~ Ns(age,kn=a.kn) +
+                     Ns(dob,kn=b.kn) +
+                     I((age-50)  *(dob-1950)) +
+                     I((age-50)^2*(dob-1950)),
+                 tau=as.numeric(iq),
+                 data=subset( abmi, sex==sx & edu==ed ) )
+ cf <- summary(m.lin)$coef[grep("dob",names(coef(m.lin))),1:2,drop=F]
+ Trend[,sx,ed,iq] <- (cf*5) %*% ci.mat()
+ cf <- summary(m.per)$coef[grep("dos",names(coef(m.per))),1:2,drop=F]
+ Curve[,sx,ed,iq] <- cf %*% ci.mat()
+ q.coh["lin",sx,ed,iq,,] <- predict( m.lin, newdata = nd.coh )
+ q.coh["add",sx,ed,iq,,] <- predict( m.add, newdata = nd.coh )
+ q.coh["per",sx,ed,iq,,] <- predict( m.per, newdata = nd.coh )
+ q.coh["int",sx,ed,iq,,] <- predict( m.int, newdata = nd.coh )
+ q.per["lin",sx,ed,iq,,] <- predict( m.lin, newdata = nd.per )
+ q.per["add",sx,ed,iq,,] <- predict( m.add, newdata = nd.per )
+ q.per["per",sx,ed,iq,,] <- predict( m.per, newdata = nd.per )
+ q.per["int",sx,ed,iq,,] <- predict( m.int, newdata = nd.per )
+    }
```

We see that in the simple model setting, the changes in BMI-quantiles is larger for the higher quantiles, but the increases in the lower ones are also there. This table shows increases in kg/m^2/5 years subdivided by sex and educational status, and the second one the *ratio* of the annual changes between those with low relative to thse with a higher education:

```
> round( ftable( Trend, col.vars=2:1, row.vars=4:3 ), 2 )
```

```
         sex      M                  F
         est BMI/5y lower upper BMI/5y lower upper
qnt  edu
0.05 lo        0.18  0.11  0.24   0.15  0.10  0.21
     hi        0.11  0.05  0.17   0.09  0.05  0.14
0.1  lo        0.24  0.19  0.30   0.18  0.13  0.23
     hi        0.18  0.13  0.23   0.12  0.09  0.16
0.25 lo        0.31  0.26  0.36   0.28  0.24  0.33
     hi        0.24  0.21  0.28   0.19  0.16  0.23
0.5  lo        0.38  0.34  0.43   0.50  0.45  0.56
     hi        0.32  0.28  0.35   0.34  0.30  0.39
0.75 lo        0.59  0.52  0.65   0.75  0.66  0.84
     hi        0.45  0.40  0.50   0.67  0.60  0.75
0.9  lo        0.82  0.72  0.91   0.87  0.73  1.01
     hi        0.62  0.53  0.71   0.96  0.83  1.08
0.95 lo        1.05  0.93  1.17   1.00  0.81  1.19
     hi        0.68  0.57  0.79   1.20  1.04  1.36
```

```
> round( t( Trend[1,,"lo",]/Trend[1,,"hi",] ), 2 )
```

```
      sex
qnt       M    F
  0.05 1.54 1.64
  0.1  1.34 1.46
  0.25 1.27 1.45
  0.5  1.21 1.47
  0.75 1.29 1.11
  0.9  1.32 0.91
  0.95 1.54 0.84
```

We see that the overall increase in BMI is larger among those with lower education than among those with higher education. We can make a forest plot for the 5-year increases in BMI by sex, education and quantile:

```
> dimnames(Trend)[["qnt"]] <-
+ paste( as.numeric(dimnames(Trend)[["qnt"]])*100, "%", sep="" )
> par( mar=c(3,1,1,1), mgp=c(3,1,0)/1.6 )
> plotEst( t(Trend[,"F","hi",]), txtpos=7:1, y=7:1+0.21, col=rgb(1,0,0,0.4),
+          xlab=expression("BMI change per 5 years (kg/"*m^2*")"),
+          vref=0, grid=TRUE )
> linesEst( t(Trend[,"F","lo",]), txtpos=7:1, y=7:1+0.07, col=rgb(1,0,0,1.0) )
> linesEst( t(Trend[,"M","hi",]), txtpos=7:1, y=7:1-0.07, col=rgb(0,0,1,0.4) )
> linesEst( t(Trend[,"M","lo",]), txtpos=7:1, y=7:1-0.21, col=rgb(0,0,1,1.0) )
```

From figure 6.1 it is pretty obvious that the increase is stronger for the higher percentiles, particularly for the most obese 25% of the population. Interstingly, high educated men have smaller rates of increase that do highly educated, whereas the patttern seems to be the opposite among women.

We also checked to see if there was a curvature component (a quadratic) to the period effect:

```
> round( ftable( Curve, col.vars=2:1, row.vars=4:3 ), 3 )
```

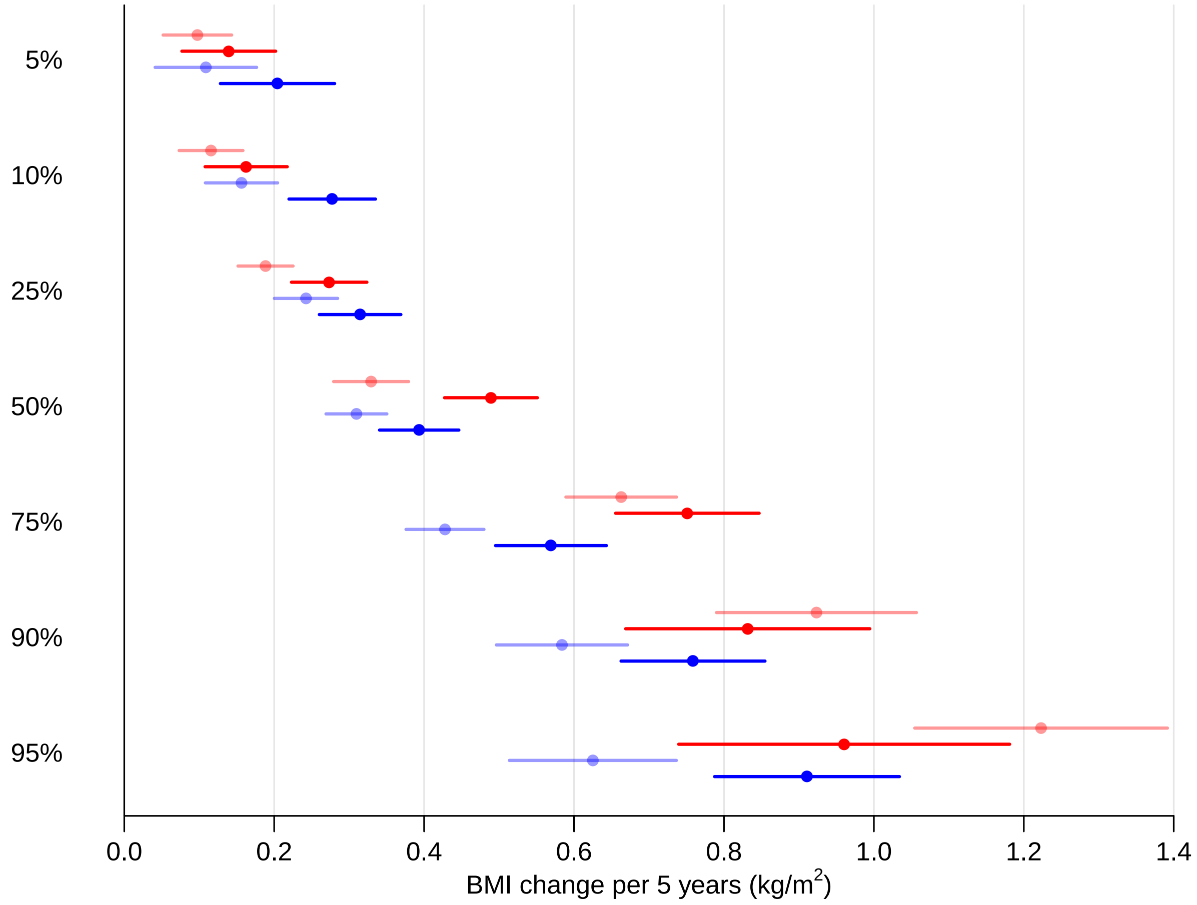

Figure 6.1: *Average change in BMI percebtile per 5 years for men (blue) and women (red), by education (high — pale; low — full).*

```
         sex      M                    F
         est  per^2  lower  upper  per^2  lower  upper
qnt  edu
0.05 lo      -0.067 -0.108 -0.026 -0.066 -0.100 -0.031
     hi      -0.021 -0.057  0.015 -0.038 -0.069 -0.007
0.1  lo      -0.046 -0.080 -0.012 -0.057 -0.084 -0.030
     hi      -0.014 -0.043  0.016 -0.041 -0.069 -0.013
0.25 lo      -0.038 -0.068 -0.009 -0.023 -0.051  0.004
     hi      -0.021 -0.046  0.004 -0.030 -0.053 -0.006
0.5  lo      -0.028 -0.056  0.000 -0.041 -0.071 -0.010
     hi      -0.008 -0.031  0.016 -0.035 -0.065 -0.005
0.75 lo       0.007 -0.031  0.045 -0.070 -0.117 -0.023
     hi       0.035  0.003  0.067 -0.008 -0.052  0.037
0.9  lo       0.046 -0.011  0.102 -0.123 -0.198 -0.047
     hi       0.049  0.005  0.092 -0.065 -0.138  0.007
0.95 lo       0.021 -0.058  0.100 -0.110 -0.217 -0.002
     hi       0.043 -0.031  0.116 -0.114 -0.208 -0.021
```

For the lower quantiles among males and for all qauntiels aming women there seems to be a negative curvature (steeper increase early, flatter later) but opposite for the quantiles above the median for males. But there seems to be no consistent pattern in curvatures between high and low edication.

Once we have filled the arrays `q.coh` and `q.per` with the BMI-quantiles, and we have an indicator of the relevant range of the cohort data in the array `o.coh`, we can plot the quantile predictions and show what parts of the predictions that are directly supported by the data.

```
> par( mfrow=dim(q.coh)[c("sex","qnt")],
+      mar=c(0,0,0,0), mgp=c(3,1,0)/1.6, oma=c(3,3,3,1), las=1 )
> for( sx in 1:dim(q.coh)[["sex"]] )
+ for( iq in 1:dim(q.coh)[["qnt"]] )
+ {
+ plot( NA, xlim=c(20,85), ylim=c(15,45),
+       xaxt="n", yaxt="n", xlab="", ylab="" )
+ if( sx==2 ) axis( side=1 )
+ if( sx==1 ) mtext( 100*as.numeric(dimnames(q.coh)[["qnt"]][iq]),
+                    side=3, line=0.5, cex=0.8 )
+ if( iq==1 ) axis( side=2 )
+ abline( h=seq(0,50,5), v=seq(0,100,10), col=gray(0.85) )
+ matlines( aa, q.coh["add",sx,"lo",iq,,]                          , lwd=1, lty=1, col=gray(0.5) )
+ matlines( aa, q.coh["add",sx,"lo",iq,,]*o.coh["add",sx,"lo",iq,,], lwd=2, lty=1,
+           col=c("blue","red")[sx] )
+ matlines( aa, q.coh["add",sx,"hi",iq,,]                          , lwd=1, lty="11", col=gray(0.5)
+ matlines( aa, q.coh["add",sx,"hi",iq,,]*o.coh["add",sx,"hi",iq,,], lwd=2, lty="11",
+           col=c("blue","red")[sx] )
+ }
> mtext( expression("BMI (kg/"*m^2*")"), side=2, line=2, outer=TRUE,las=0,cex=0.8 )
> mtext( "Age (years)", side=1, line=2, outer=TRUE,las=0,cex=0.8 )
> mtext( "Percentile of the BMI-distribution",
+        side=3, line=1.5, cex=0.8, outer=TRUE )
```

```
> par( mfrow=dim(q.coh)[c("sex","qnt")],
+      mar=c(0,0,0,0), mgp=c(3,1,0)/1.6, oma=c(3,3,3,1), las=1 )
> for( sx in 1:dim(q.coh)[["sex"]] )
+ for( iq in 1:dim(q.coh)[["qnt"]] )
+ {
+ plot( NA, xlim=c(20,85), ylim=c(15,45),
+       xaxt="n", yaxt="n", xlab="", ylab="" )
+ if( sx==2 ) axis( side=1 )
+ if( sx==1 ) mtext( 100*as.numeric(dimnames(q.coh)[["qnt"]][iq]),
+                    side=3, line=0.5, cex=0.8 )
```

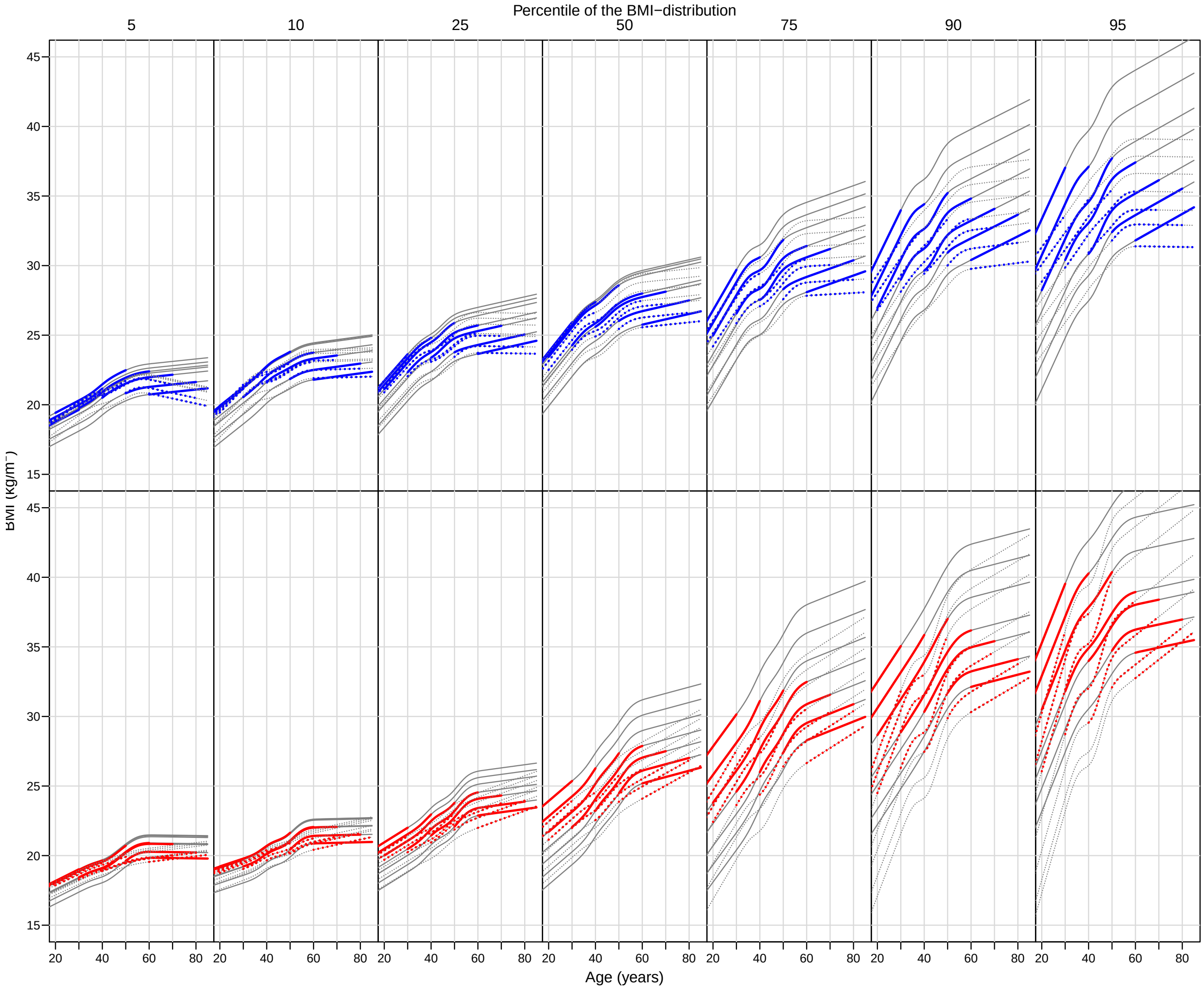

Figure 6.2: *Age-specific percentiles of BMI by sex and education. Each curve represents a birth cohort (1920, 1930,. . . ,1980). Gray lines are estimates from quantile regression models, the colored part of the lines indicate the ares in which data is available. Full lines are low education, dotted lines high.*

Note that the curves in each panel come from the same model-fit (namely that for the corresponding quantile) and are assumed to be parallel, that is to have the same shape. This is a model which allows a non-linear cohort effect; the model with linear cohort effect would assume that the curves were equidistant.

```
+ if( iq==1 ) axis( side=2 )
+ abline( h=seq(0,50,5), v=seq(0,100,10), col=gray(0.85) )
+ matlines( aa, q.per["add",sx,"lo",iq,,], lwd=2, lty=1   , col=c("blue","red")[sx] )
+ matlines( aa, q.per["add",sx,"hi",iq,,], lwd=2, lty="11", col=c("blue","red")[sx] )
+ }
> mtext( expression("BMI (kg/"*m^2*")"), side=2, line=2, outer=TRUE,las=0,cex=0.8 )
> mtext( "Age (years)", side=1, line=2, outer=TRUE,las=0,cex=0.8 )
> mtext( "Percentile of the BMI-distribution",
+        side=3, line=1.5, cex=0.8, outer=TRUE )
```

Here are the figures that we use in the paper; they are only for the 10, 50 and 90th percentiles:

```
> temp.graph <- function()
+ {
+ par( mfrow=c(2,3), mar=c(0,0,0,0), mgp=c(3,1,0)/1.6, oma=c(3.5,3.5,3,2.5), las=1 )
```

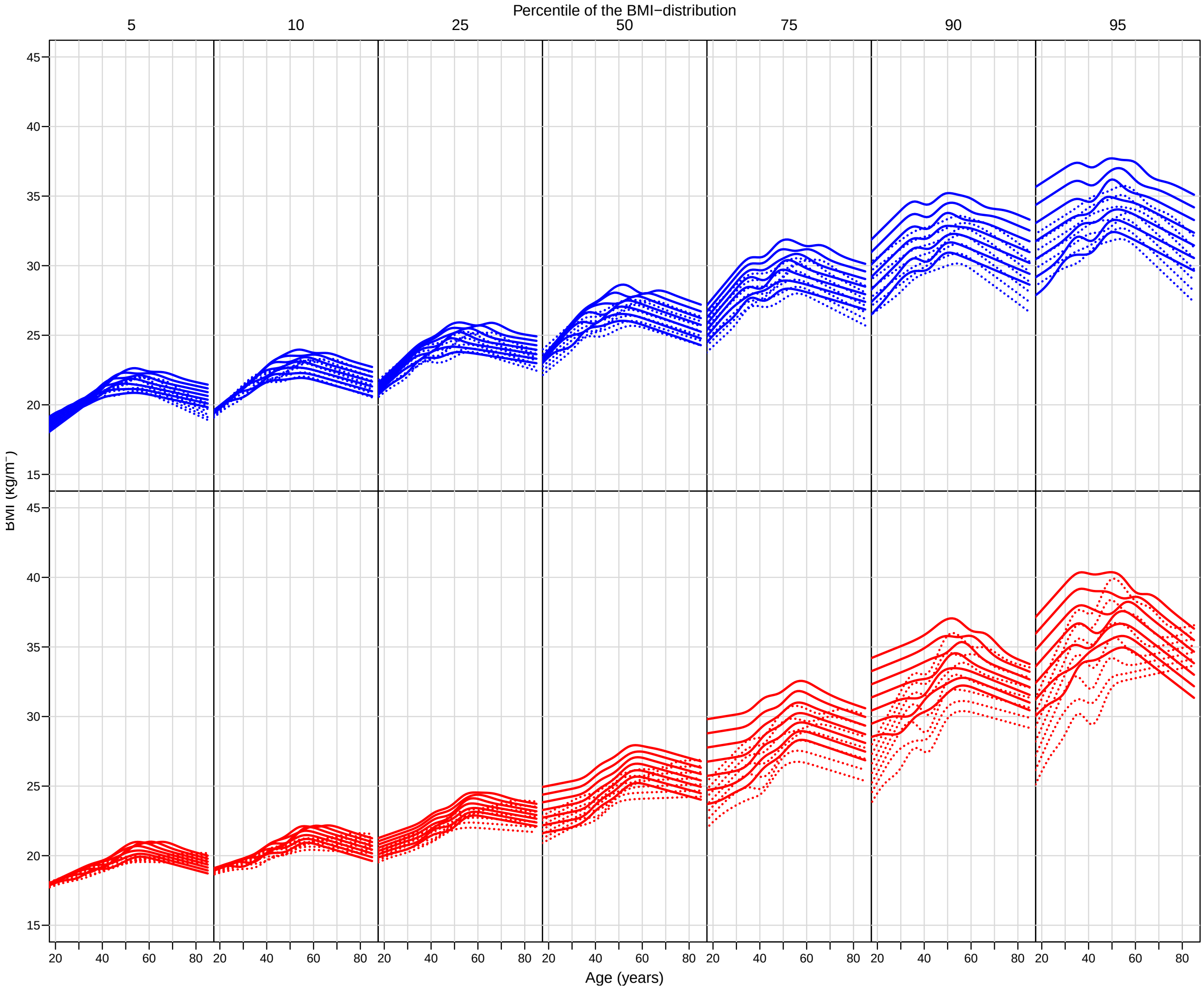

Figure 6.3: *Age-specific percentiles of BMI by sex and education. Each curve represents a date (1980, 1985,. . . ,2010). Men are blue, women red, low edication are full lines, high education broken lines.*
Note that the curves in each panel come from the same model-fit (namely that for the corresponding quantile) and are assumed to be parallel, that is to have the same shape between cohorts, *but not necessarily between periods.*

```
+ for( sx in 1:dim(q.coh)[["sex"]] )
+ for( iq in c(2,4,6) )
+ {
+ plot( NA, xlim=c(20,85), ylim=c(15,45),
+       xaxt="n", yaxt="n", xlab="", ylab="" )
+ if( sx==2 ) axis( side=1 )
+ if( sx==1 ) mtext( 100*as.numeric(dimnames(q.coh)[["qnt"]][iq]),
+                    side=3, line=0.5, cex=0.8 )
+ if( iq==2 ) axis( side=2 )
+ if( iq==6 ) axis( side=4 )
+ abline( h=seq(0,50,5), v=seq(0,100,10), col=gray(0.85) )
+ matlines( aa, q.coh["add",sx,"lo",iq,,]                          , lwd=1, lty=1, col=gray(0.5) )
+ matlines( aa, q.coh["add",sx,"lo",iq,,]*o.coh["add",sx,"lo",iq,,], lwd=2, lty=1,
+           col=c("blue","red")[sx] )
+ matlines( aa, q.coh["add",sx,"hi",iq,,]                          , lwd=1, lty="11", col=gray(0.5)
+ matlines( aa, q.coh["add",sx,"hi",iq,,]*o.coh["add",sx,"hi",iq,,], lwd=2, lty="11",
+           col=c("blue","red")[sx] )
+ }
+ mtext( expression("BMI (kg/"*m^2*")"), side=2, line=2, outer=TRUE,las=0,cex=0.8 )
+ mtext( "Age (years)", side=1, line=2, outer=TRUE,las=0,cex=0.8 )
+ mtext( "Percentile of the BMI-distribution",
+        side=3, line=1.5, cex=0.8, outer=TRUE )
+ }
> win.metafile(       "art-coh-edu.emf", height=7, width=7 ) ; temp.graph() ; dev.off()

null device
          1

> pdf("./graph/BMI-APC-art-coh-edu.pdf", height=7, width=7 ) ; temp.graph() ; dev.off()

null device
          1

> temp.graph <- function()
+ {
+ par( mfrow=c(2,3), mar=c(0,0,0,0), mgp=c(3,1,0)/1.6, oma=c(3.5,3.5,3,3), las=1 )
+ for( sx in 1:dim(q.coh)[["sex"]] )
+ for( iq in c(2,4,6) )
+ {
+ plot( NA, xlim=c(20,85), ylim=c(15,45),
+       xaxt="n", yaxt="n", xlab="", ylab="" )
+ if( sx==2 ) axis( side=1 )
+ if( sx==1 ) mtext( 100*as.numeric(dimnames(q.coh)[["qnt"]][iq]),
+                    side=3, line=0.5, cex=0.8 )
+ if( iq==2 ) axis( side=2 )
+ if( iq==6 ) axis( side=4 )
+ abline( h=seq(0,50,5), v=seq(0,100,10), col=gray(0.85) )
+ matlines( aa, q.per["add",sx,"lo",iq,,], lwd=2, lty=1   , col=c("blue","red")[sx] )
+ matlines( aa, q.per["add",sx,"hi",iq,,], lwd=2, lty="11", col=c("blue","red")[sx] )
+ }
+ mtext( expression("BMI (kg/"*m^2*")"), side=2, line=2, outer=TRUE,las=0,cex=0.8 )
+ mtext( "Age (years)", side=1, line=2, outer=TRUE,las=0,cex=0.8 )
+ mtext( "Percentile of the BMI-distribution",
+        side=3, line=1.5, cex=0.8, outer=TRUE )
+ }
> win.metafile(       "art-per-edu.emf", height=7, width=7 ) ; temp.graph() ; dev.off()

null device
          1

> pdf("./graph/BMI-APC-art-per-edu.pdf", height=7, width=7 ) ; temp.graph() ; dev.off()

null device
          1
```

Bibliography

[1] B Carstensen. Age-Period-Cohort models for the Lexis diagram. *Statistics in Medicine*, 26(15):3018–3045, July 2007.

[2] B. Carstensen. *Comparing Clinical Measurement Methods: A practical guide.* Wiley, 2010.

[3] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[4] H. L. Walls, R. Wolfe, M. M. Haby, D. J. Magliano, M. de Courten, C. M. Reid, J. J. McNeil, J. Shaw, and A. Peeters. Trends in BMI of urban Australian adults, 1980-2000. *Public Health Nutr*, 13(5):631–638, May 2010.