Smoking prevalence preprocessing

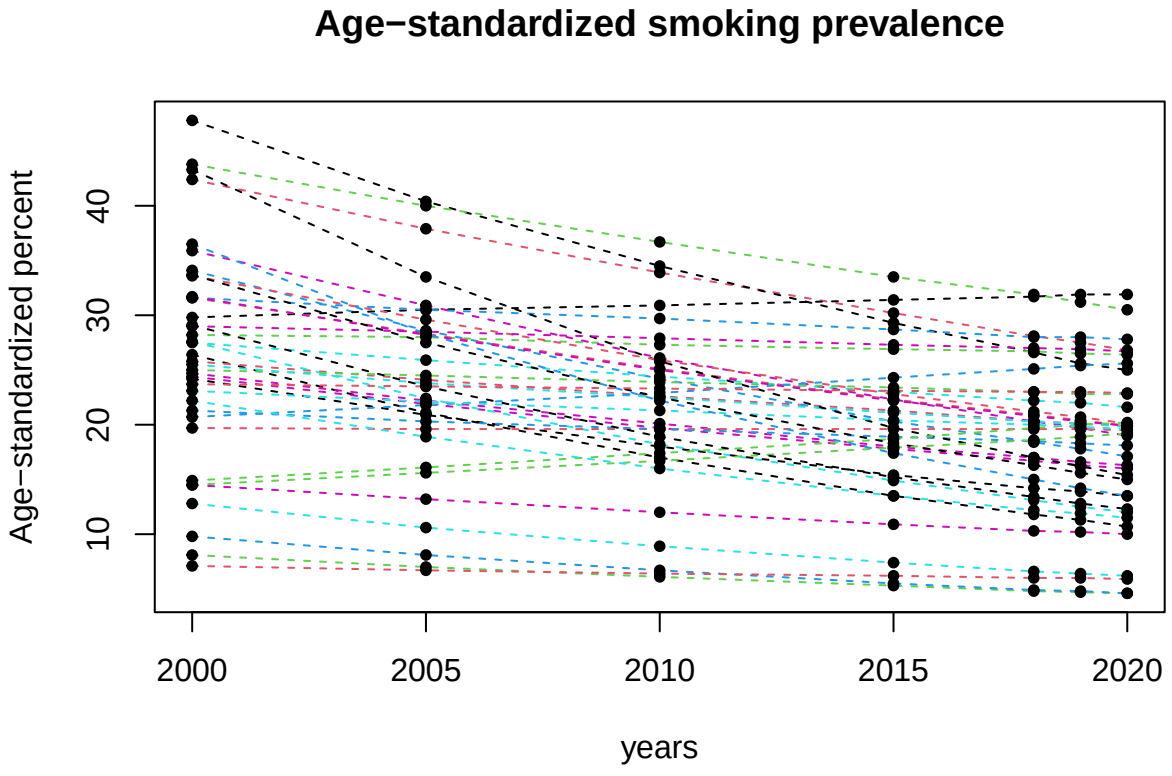
2023-07-02

#Plotting the original values

```
matplot(years,data,
        type = "l",
        lty="dashed",
        xlab="years",
        ylab="Age-standardized percent",
        main="Age-standardized smoking prevalence")

countryrange<-1:38

for (i in countryrange){
  points(years, data[,i], pch = 20)
}
```

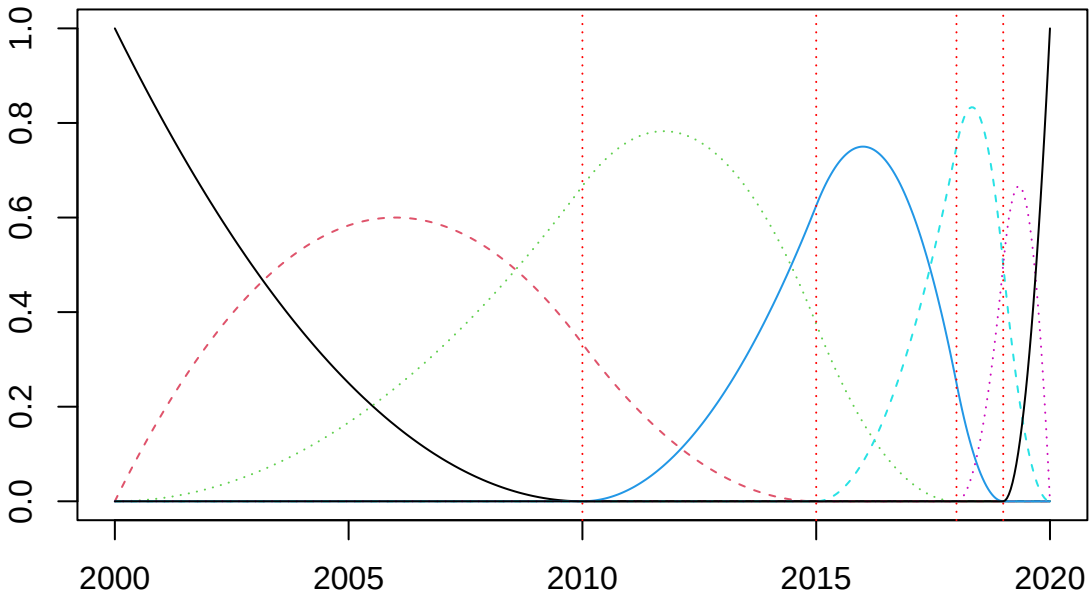


Starting with the interpolation approach

```
# Create bspline basis imposing the passage by the points
myrange<-c(2000,2020)
```

```
myknots = c(2000,2010, 2015, 2018, 2019, 2020)
myorder=3
nbasis = myorder + length(myknots) -2

basis <- create.bspline.basis(rangeval=myrange,
                              breaks=myknots,
                              nbasis=nbasis,
                              norder=myorder)
#Plot basis
plot(basis)
```



Evaluate bspline basis on abscissa

```
basismat <- eval.basis(years[1:length(years)], basis)
#Dimensional check - c(number of timepoints,  number of basis)
#c(dim(data)[1], nbasis)
```

Estimate coefficients

```
est_coef = lsfit(basismat, data, intercept=FALSE)$coef
```

Compute the estiamted values just to chech that they are of course the same of the real ones, and plot

```
smoothed <- basismat %*% est_coef
dim(smoothed)
```

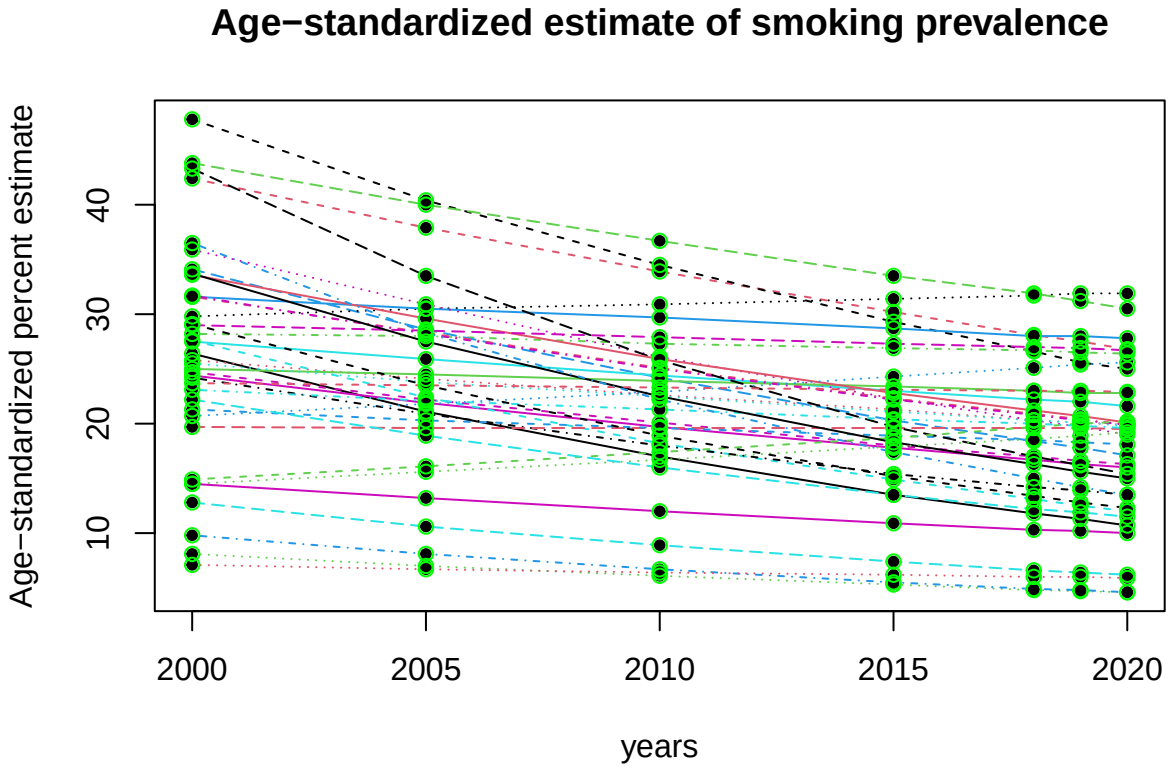
```
## [1]  7 38
```

```
row.names(smoothed)<-years

#Plot smoothed function
matplot(years,smoothed,
```

```
        type="l",
        xlab="years",
        ylab="Age-standardized percent estimate",
        main="Age-standardized estimate of smoking prevalence")

for (i in countryrange){
  points(years[1:length(years)], data[1:length(years),i], pch = 20) # add blue points to second line
}
#Double visual check that the smoothing points exactly superimpose the original ones
for (i in countryrange){
  points(years[1:length(years)], smoothed[1:length(years),i],
         col='green',
         pch = 1)
}
```



Predict values for new years, and plot

```
new_years <- c(2007, 2008, 2012, 2014, 2016)
new_years_basismat <- eval.basis(new_years, basis)
predicted_values <- new_years_basismat %*% est_coef
dim(new_years_basismat)
```

```
## [1] 5 7
```

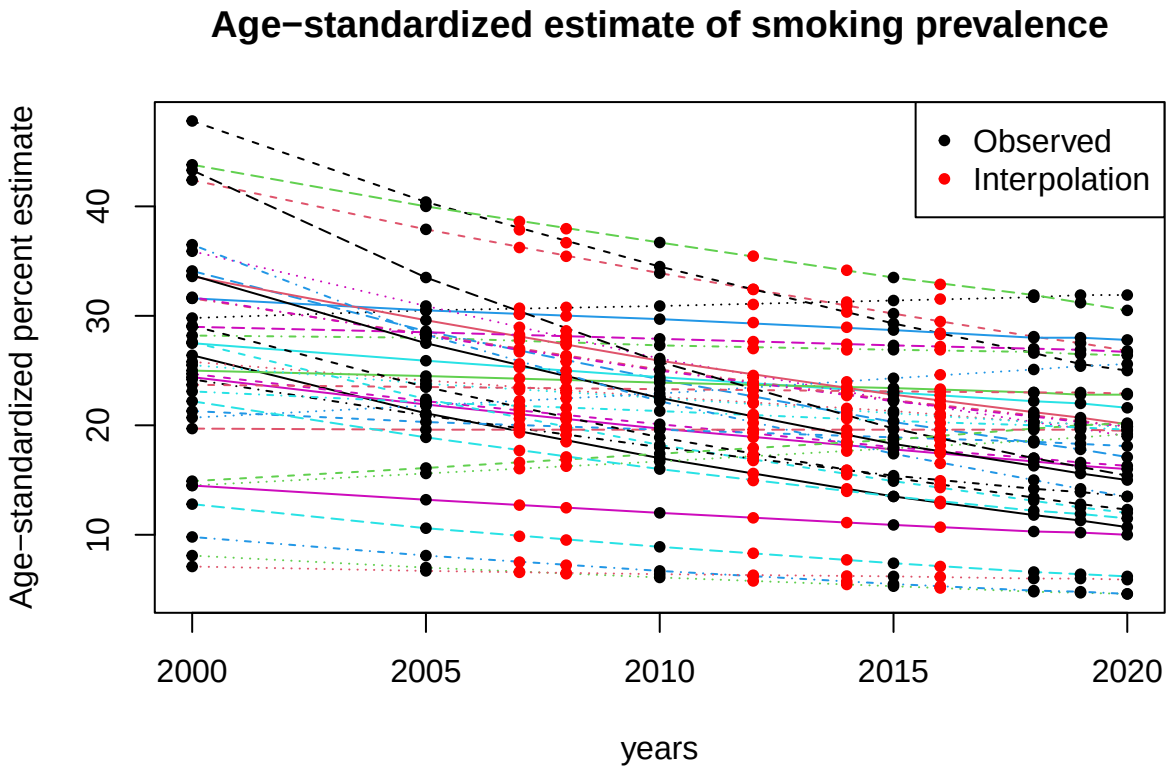
```
dim(est_coef)
```

```
## [1]  7 38
```

```r
# Assign correct column name and print predicted values
row.names(predicted_values)<-new_years

#Plot smoothed function
matplot(years,smoothed,
        type="l",
        xlab="years",
        ylab="Age-standardized percent estimate",
        main="Age-standardized estimate of smoking prevalence",
        cex.main=1.2,
        cex.lab=1)

for (i in countryrange){
  points(years[1:length(years)], data[1:length(years),i],
         pch = 20) # add blue points to second line
}
for (i in countryrange){
  points(new_years[1:length(new_years)], predicted_values[1:length(new_years),i],
         pch = 20,col='red') # add blue points to second line
}
legend("topright",
       legend = c("Observed", "Interpolation"),
       pch = c(20, 20),
       col = c("black", "red"))
```



Comparing interpolation with smoothing obtained with smoothing splines with lambda = 100

```r
abscissa<-years
data.fd.1 <- Data2fd(y = as.matrix(data),
```

```
                    argvals = abscissa,
                    lambda = 100)
#default for the basis is create.bspline.basis(argvals)
```

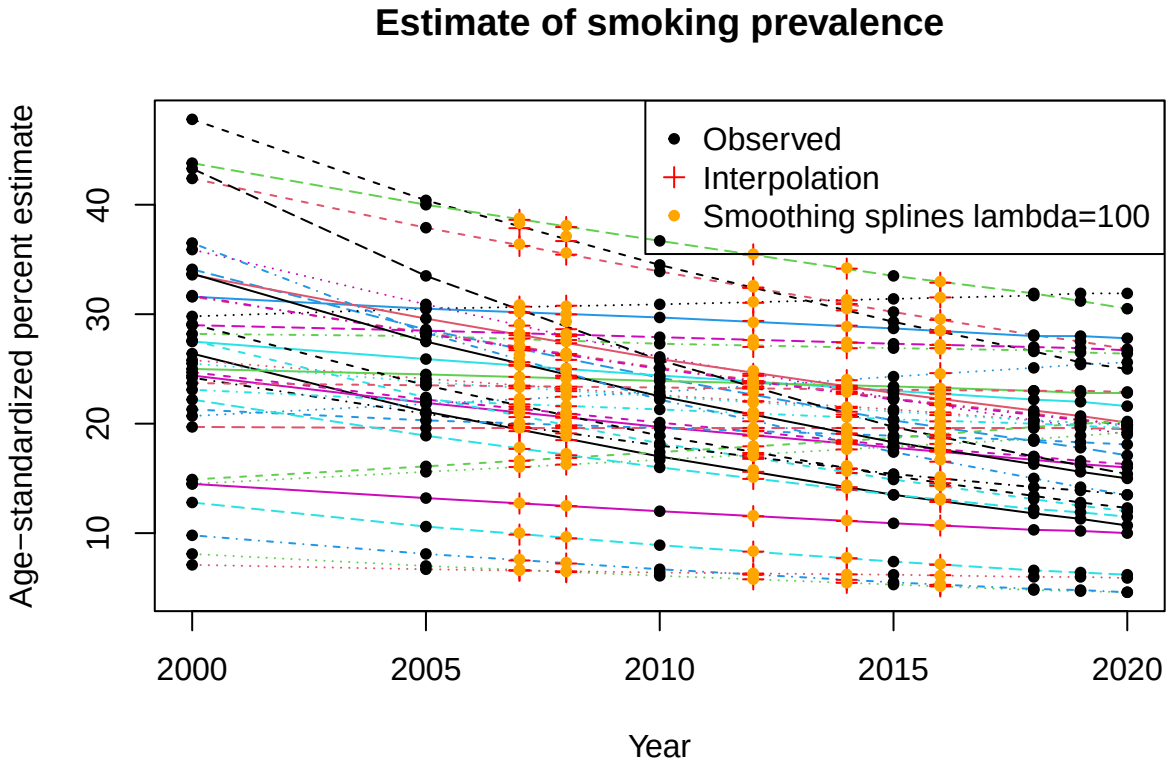
Predict new years values using this approach, and plot

```
basis=create.bspline.basis(abscissa) #same used by default in the functional datum
new_years_basismat_2 <- eval.basis(new_years, basis)
dim(new_years_basismat_2)
```

```
## [1] 5 9
```

```
predicted_values_smooth <- new_years_basismat_2 %*% data.fd.1$coefs


#Visualization and comparison with previous estimates
matplot(years,(data),
        type="l",
        ylab="Age-standardized percent estimate",
        xlab="Year",
        main="Estimate of smoking prevalence",
        cex.main=1.2,
        cex.lab=1)
for (i in countryrange){
  points(years[1:length(years)], data[1:length(years),i],
         pch = 20)
}
for (i in countryrange){
  points(new_years[1:length(new_years)], predicted_values[1:length(new_years),i],
         pch = 3,col='red')
}
for (i in countryrange){
  points(new_years[1:length(new_years)], predicted_values_smooth[1:length(new_years),i],
         pch = 20,col='orange')
}
legend("topright",
       legend = c("Observed", "Interpolation", "Smoothing splines lambda=100"),
       pch = c(20, 3, 20),
       col = c("black", "red", "orange"))
```



I see that the precitions are extremely similar. I also compute the squared Frobenious norm to have a numerical confirm, knowing that my prevalences range from 0 to 60% and I am computing a diff matrix of 38x5

```
norm(predicted_values_smooth-predicted_values, type="F")^2
```

```
## [1] 7.009144
```

```
# 7.486604 both
# 8.835872 males
# 7.009144 females
```

Due to the extreme smoothness of the data, the two predictions are very similar, as also confirmed by the squared frobenious norm and by visual inspectionhence we consider the two estimates interchangeable and proceed by using the first estimate

(I repeat the same analysis for both sexes and, males with similar results)