Wage Example – Regression spline

Part 1: Cubic Spline

Wage Example – Regression spline

Basic Implementation

We can fit a cubic spline with knots at age 25, 40 and 60 as follows:

```
#load required library
library(splines)
#Fit a cubic spline
fit=lm(wage~bs(age,knots=c(25,40,60)),data=Wage)
summary(fit)
```

Wage Example – Regression spline

Basic Implementation

The summary of the cubic spline is as follows:

```
Call:
lm(formula = wage ~ bs(age, knots = c(25, 40, 60)), data = Wage)

Residuals:
    Min      1Q  Median      3Q     Max 
-98.832 -24.537  -5.049  15.209 203.207 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                       60.494      9.460   6.394 1.86e-10 ***
bs(age, knots = c(25, 40, 60))1    3.980     12.538   0.317 0.750899    
bs(age, knots = c(25, 40, 60))2   44.631      9.626   4.636 3.70e-06 ***
bs(age, knots = c(25, 40, 60))3   62.839     10.755   5.843 5.69e-09 ***
bs(age, knots = c(25, 40, 60))4   55.991     10.706   5.230 1.81e-07 ***
bs(age, knots = c(25, 40, 60))5   50.688     14.402   3.520 0.000439 ***
bs(age, knots = c(25, 40, 60))6   16.606     19.126   0.868 0.385338    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.92 on 2993 degrees of freedom
Multiple R-squared:  0.08642,	Adjusted R-squared:  0.08459 
F-statistic: 47.19 on 6 and 2993 DF,  p-value: < 2.2e-16
```

The degree of freedom for the cubic spline is K+4=7. (K=3)

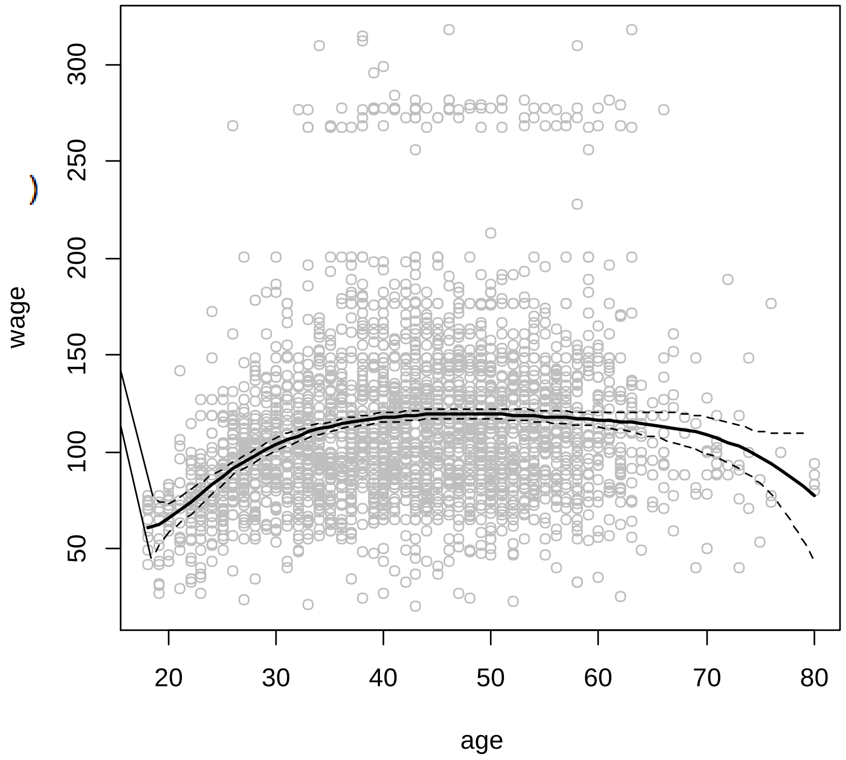
Coeff. cannot be directly interpreted

Wage Example – Regression spline

Basic Implementation

We can visualize the fit with the following code:

```
#Predict the grid
pred <- predict (fit,newdata =list(age = age.grid ) , se = T )
plot(age,wage ,col = " gray ")
#plot the fit
lines (age.grid,pred $fit, lwd = 2)
#plot confidence interval with the existing graph
lines (age.grid,pred $fit+2*pred $se, lty = "dashed")
lines (age.grid,pred$fit-2*pred $se, lty = "dashed")
```



Wage Example – Regression spline

Basic Implementation

We can also define the spline with specified degree of freedom. The function produces a spline with uniform quantile.

In this case, The quantiles are 25%, 50% and 75%.

```
> dim(bs(age,df=6))
[1] 3000    6
> attr(bs(age,df = 6),"knots")
  25%   50%   75%
33.75 42.00 51.00
```

In R, intercept NOT counted as 1 df

So df (in R) = 7 – 1 = 6

Wage Example – Regression spline

Part 2: Natural Spline

Wage Example – Regression spline

Basic Implementation

Natural spline is stable at the boundary because the function has to be linear at boundaries. Therefore, natural spline is considered in practice.

The natural spline can be created by the **ns function** as follows:

```
> dim(ns(age,df=4))
[1] 3000    4
> attr(ns(age,df = 4),"knots")
  25%   50%   75%
33.75 42.00 51.00
```

Wage Example – Regression spline

Basic Implementation

Summary of ns

```
# In order to instead fit a natural spline, we use the ns() function
fit2=lm(wage~ns(age,df=4),data=Wage)
summary(fit2)

Call:
lm(formula = wage ~ ns(age, df = 4), data = Wage)

Residuals:
    Min      1Q  Median      3Q     Max 
-98.737 -24.477  -5.083  15.371 204.874 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)        58.556      4.235  13.827   <2e-16 ***
ns(age, df = 4)1   60.462      4.190  14.430   <2e-16 ***
ns(age, df = 4)2   41.963      4.372   9.597   <2e-16 ***
ns(age, df = 4)3   97.020     10.386   9.341   <2e-16 ***
ns(age, df = 4)4    9.773      8.657   1.129    0.259    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.92 on 2995 degrees of freedom
Multiple R-squared:  0.08598,	Adjusted R-squared:  0.08476 
F-statistic: 70.43 on 4 and 2995 DF,  p-value: < 2.2e-16
```

Again in R, intercept NOT counted as 1 df
So df (in R) = 3 + 1 = 4

Coeff cannot be directly interpreted

Wage Example – Regression spline

Basic Implementation

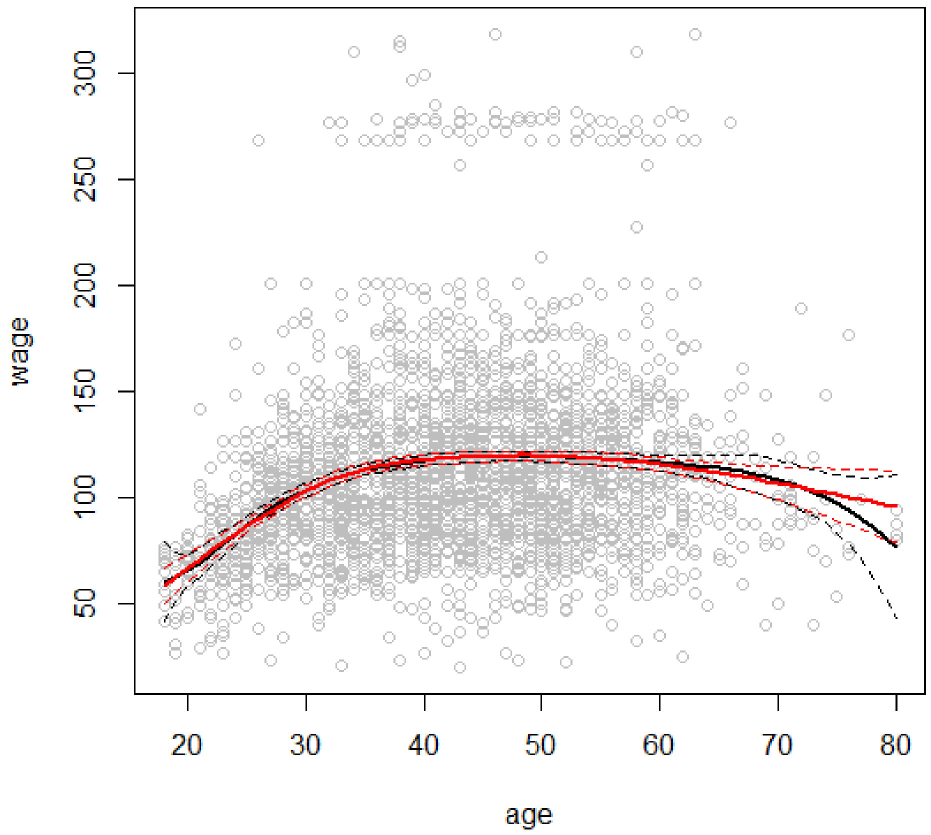
We can produce the plot with the following code to compare the natural cubic spline with cubic spline:

```
fit2 <- lm(wage~ns(age,df = 4),data = Wage )
#Predict the grid
pred2 <- predict (fit2,newdata =list(age = age.grid ) , se = T )
#plot the fit
lines (age.grid,pred2$fit, lwd = 2,col="red")
#plot confidence interval with the existing graph
lines (age.grid,pred2$fit+2*pred2$se, lty = "dashed",col="red")
lines (age.grid,pred2$fit-2*pred2$se, lty = "dashed",col="red")
```

Wage Example – Regression spline

Basic Implementation

The output shows that the natural cubic spline has a narrower confidence interval.



Red: natural cubic spline

Black: cubic spline

Wage Example – Smoothing spline for regression

Part 3.1: Smoothing Spline For Regression

Wage Example – Smoothing spline for regression

Basic Implementation

We can fit a smoothing spline after specifying the degree of freedom or cross-validation.

Provide x, y, df

```
#Specify degree of freedom(df)
fit <- smooth.spline(age,wage,df = 16)
#Or Set cross-validation to TRUE and search the optimal df
fit2 <- smooth.spline (age,wage,cv = TRUE )
```

Provide x, y, use CV to determine df

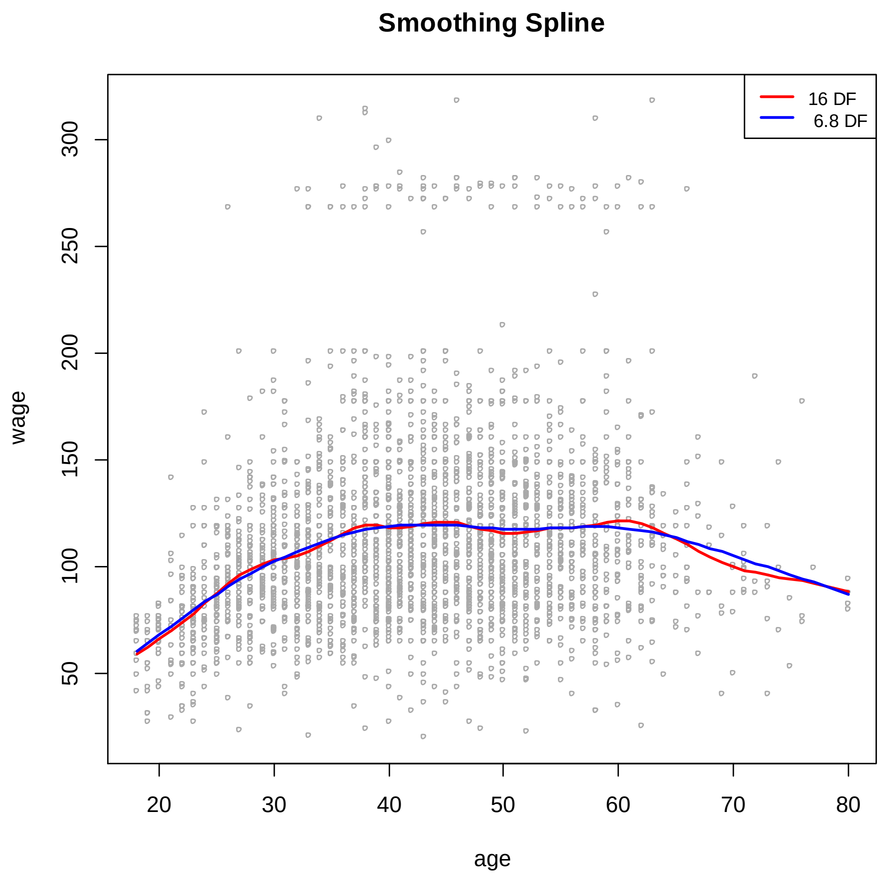
```
> fit2$df
[1] 6.794596
```

Wage Example – Smoothing spline for regression

Basic Implementation

We can plot the graphs with the following code:

```
plot(age,wage,xlim = agelims,cex = .5,col ="darkgrey")
title ("Smoothing Spline")
lines ( fit , col = "red ", lwd = 2)
lines ( fit2 , col = " blue ", lwd = 2)
legend ("topright", legend = c("16 DF", " 6.8 DF") ,
col = c("red", "blue") , lty = 1, lwd = 2 , cex = .8)
```



Wage Example – Smoothing spline for logistic regression

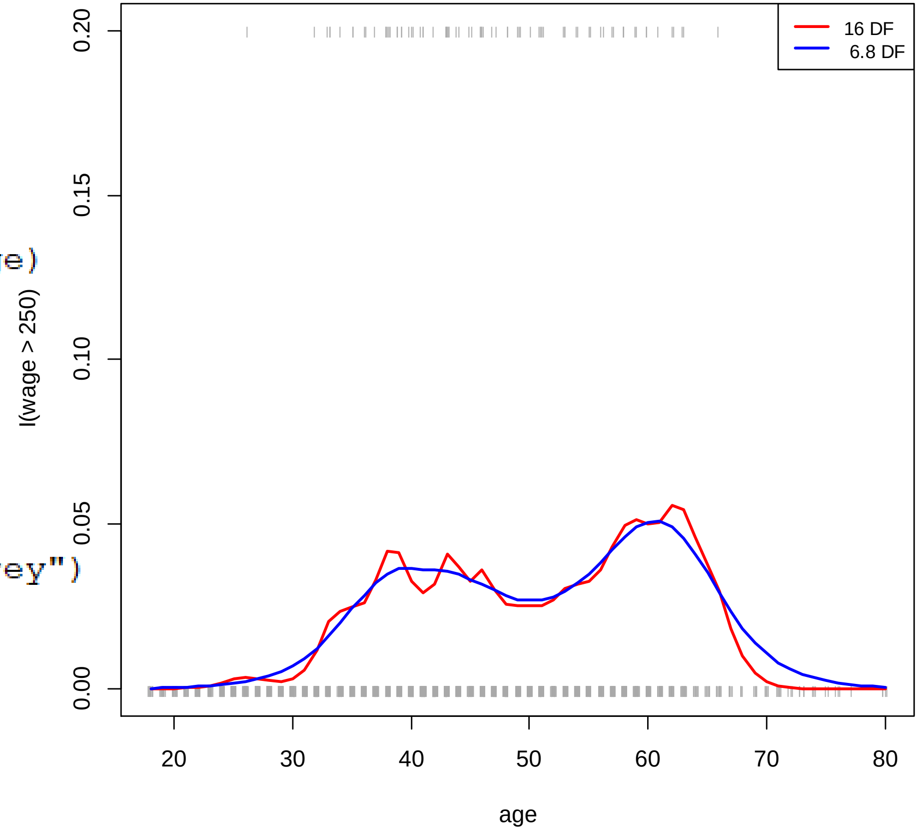
Part 3.2: Smoothing Spline For Logistic Regression

Wage Example – Smoothing spline for logistic regression

Basic Implementation

After loading gam package, we can do the same for logistic regression with a smoothing spline.

```
#load neccessary package
library(gam)
#fit different nonparametric logistic regression models.
fit <- gam(I(wage>250)~s(age,df=16),family=binomial,data=Wage)
fit2 <- gam(I(wage>250)~s(age,df=6.8),family=binomial,data=Wage)
#Predict the grid
preds<-predict(fit,newdata=list(age=age.grid))
pfit<-exp(preds)/(exp(preds)+1)
preds2<-predict(fit2,newdata=list(age=age.grid))
pfit2<-exp(preds2)/(exp(preds2)+1)
#Plot the graph
plot(age,I(wage>250),xlim=agelims,type="n",ylim=c(0,0.2))
points(jitter(age),I((wage>250)/5),cex=0.5,pch="|",col="darkgrey")
lines(age.grid,pfit,lwd=2,col="red")
lines(age.grid,pfit2,lwd=2,col="blue")
legend ("topright", legend = c("16 DF", " 6.8 DF") ,
col = c("red", "blue") , lty = 1, lwd = 2 , cex = .8)
```



Wage Example – Generalized additive model

Part 4: Generalized additive model

Wage Example – Generalized additive model

Basic Implementation

For generalized additive models, we use gam package.

We can choose different combination of covariates.

Recall: education categorical, no need for nonlinear transformation

```
library(gam)
#ns(...) is a natural spline function
gam1<-lm( wage~ ns(year,4) +ns(age ,5) + education ,
data = Wage)
#s(...) is a smoothing spline function
gam.m3<-gam(wage~s(year,4)+s(age,5)+education,data = Wage)
```

Wage Example – Generalized additive model

Basic Implementation

Plot the corresponding fits

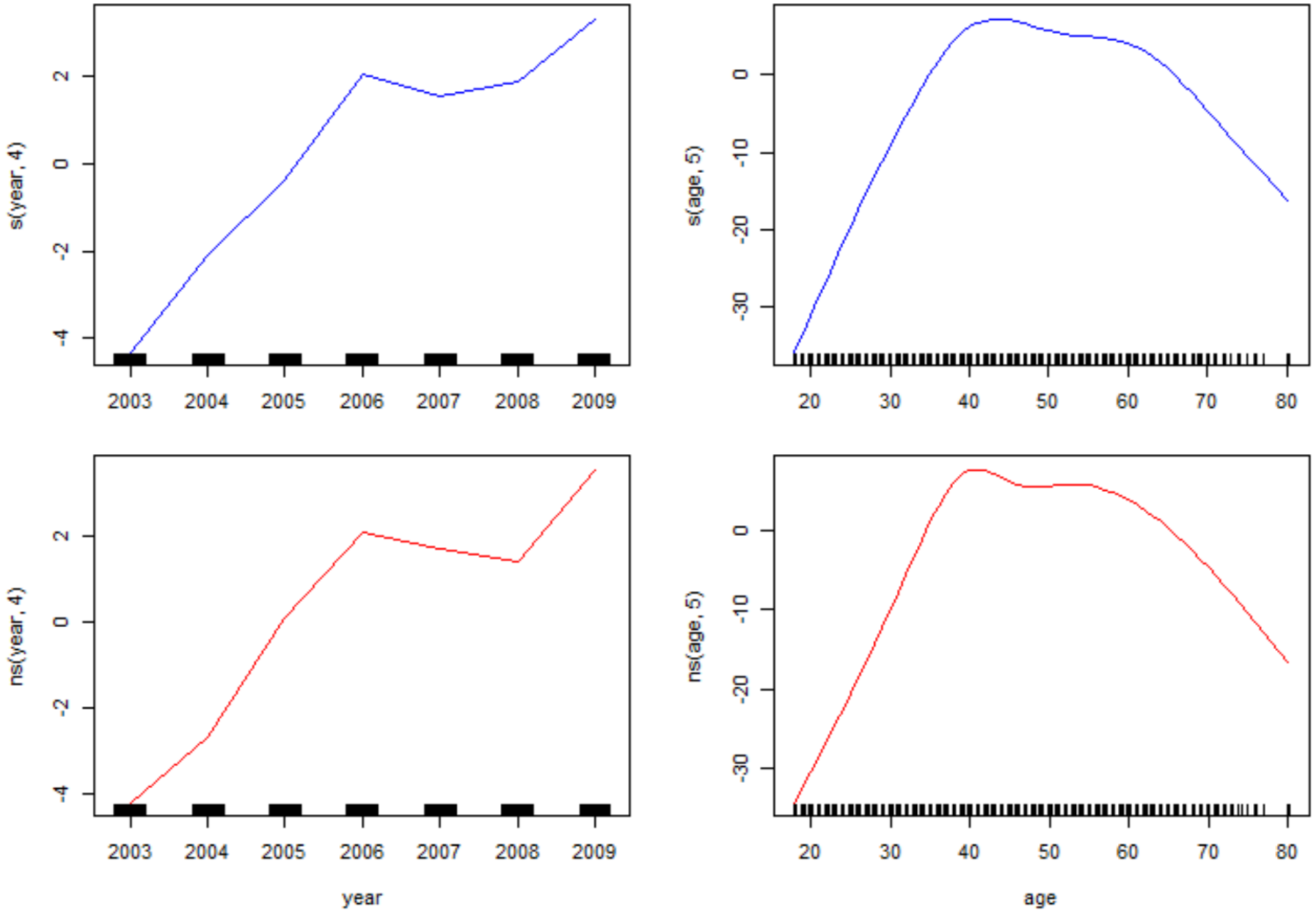
```
par(mfrow = c(1 , 3) )
plot.Gam(gam.m3 , col = " blue ")
plot.Gam(gam1 , col = "red ")
```

Wage Example – Generalized additive model

Basic Implementation

first row: smoothing spline
second row: natural spline

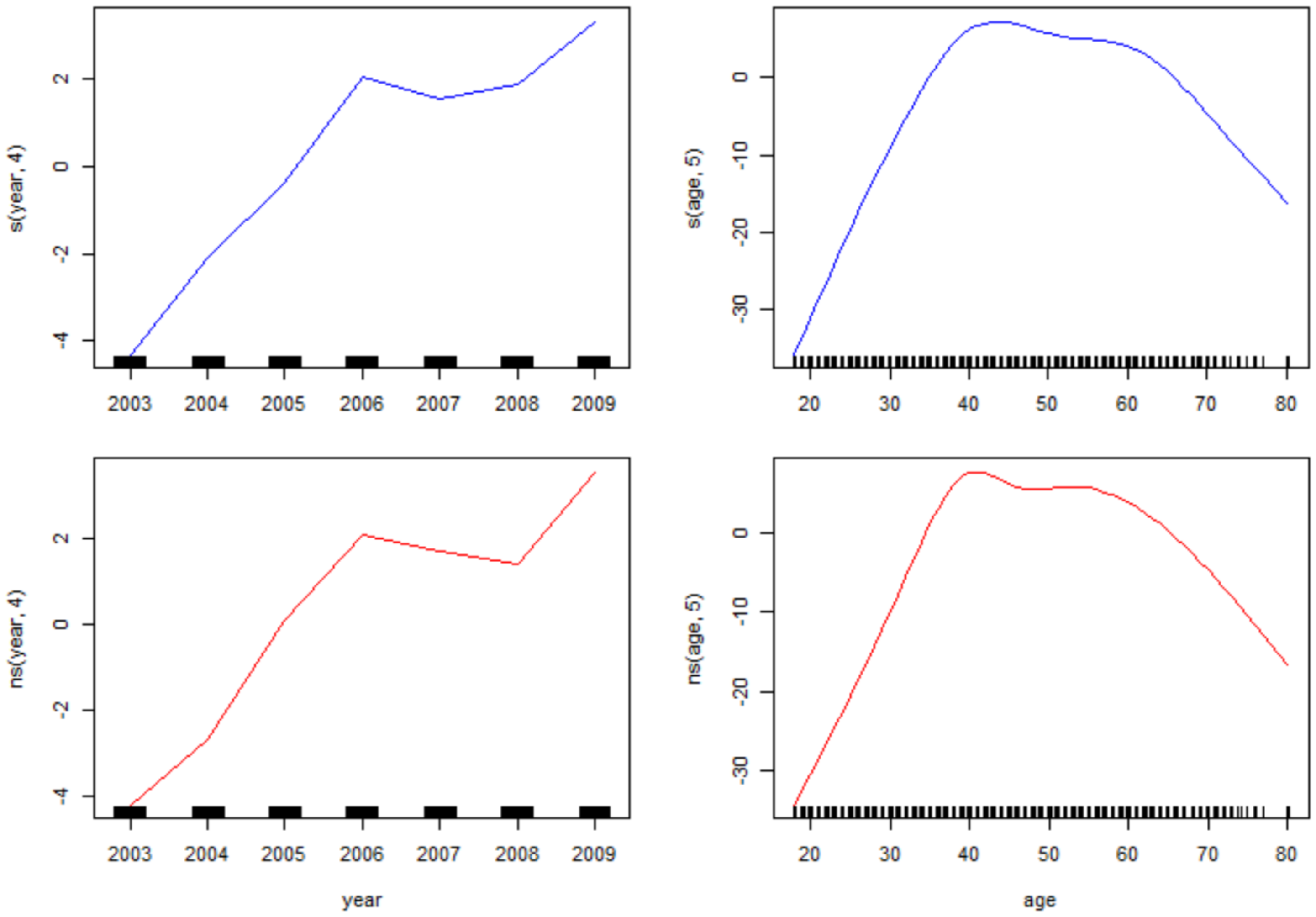
Two splines similar in shape



Easily interpreted: holding age and education constant, wage ↑ slightly with year, etc

Wage Example – Generalized additive model

Basic Implementation



first row: smoothing spline
second row: natural spline

Two splines similar in shape

Age seems nonlinear
year may be linear

Suggest can have an alternative model: `gam.m2 <- gam(wage~year+s(age,5)+education ,`