Smoothing and regression splines

Joel Cantero Priego & Ricard Meyerhofer Parra

17/12/2019

Introduction

In this assigment we are working with the bikes Washington dataset that contains information on the bike-sharing rental service in Washington D.C., USA, corresponding to years 2011 and 2012. This file contains only one data frame, bikes, with 731 rows (one for each day of years 2011 and 2012, that was a leap year) and 9 columns:

Vari able name	Description	Values
instant	row index	Integer going from 1 to 731
yr	year	Integer (0: 2011, 1:2012)
dayyr	day of the year	Integer (from 1 to 365 for 2011, and from 1 to 366 for 2012)
weekday	day of the week	Integer (0 for Sunday, 1 for Monday, . . . , 6 for Saturday)
workingday	if day is neither weekend nor holiday	Integer is 1, otherwise is 0
temp	temperature in Celsius	Integer
hum	humidity	Integer in %
windspeed	wind speed	Integer in miles per hour
cnt	count of total rental bikes.	Integer

1. Consider the nonparametric regression of cnt as a function of instant. Estimate the regression function m(instant) of cnt as a function of instant using a cubic regression splines estimated with the R function smooth.splines and choosing the smoothing parameter by Generalized Cross Validation.

```
smoothSpline <- smooth.spline(x = bikes$instant,
                              y = bikes$cnt,
                              cv = FALSE,
                              all.knots = FALSE)
(smoothSpline$lambda)
```

Which is the value of the chosen penalty parameter?

The value of the chosen penalty parameter is 1.005038e-07 by Generalized Cross Validation.

Which is the corresponding equivalent number of degrees of freedom df?

```
(smoothSpline$df)
```

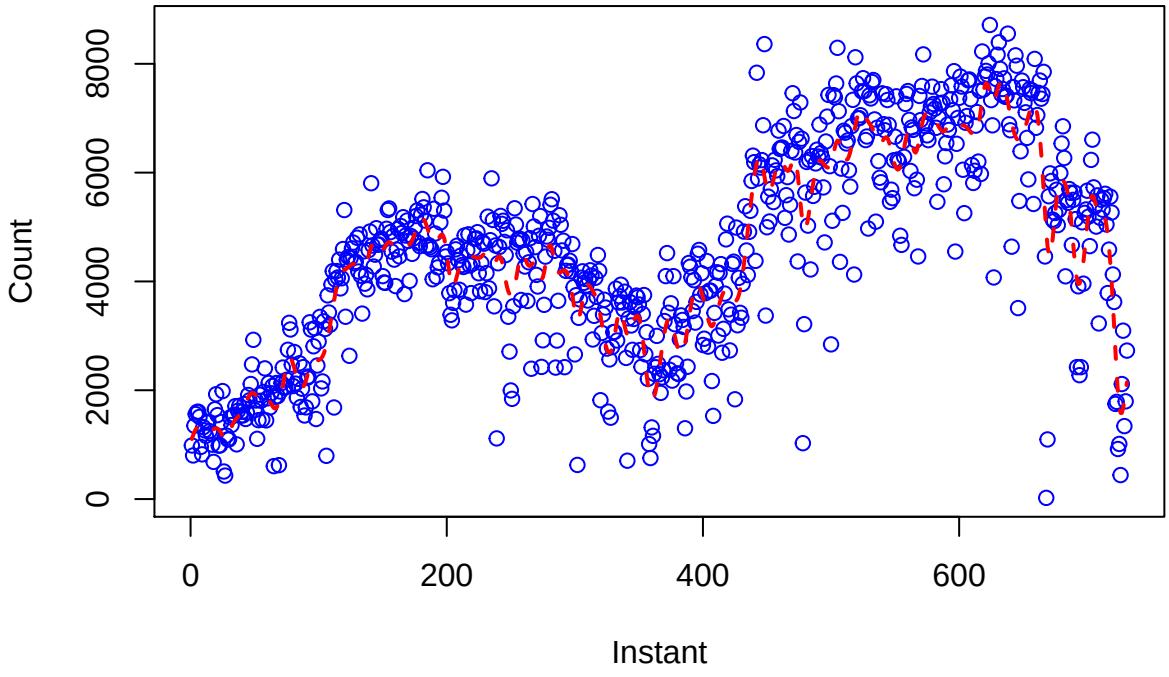
The corresponding equivalent number of degrees of freedom is 93.

How many knots have been used?

```
length(smoothSpline$fit$knot)
```

We have been used 140 knots.

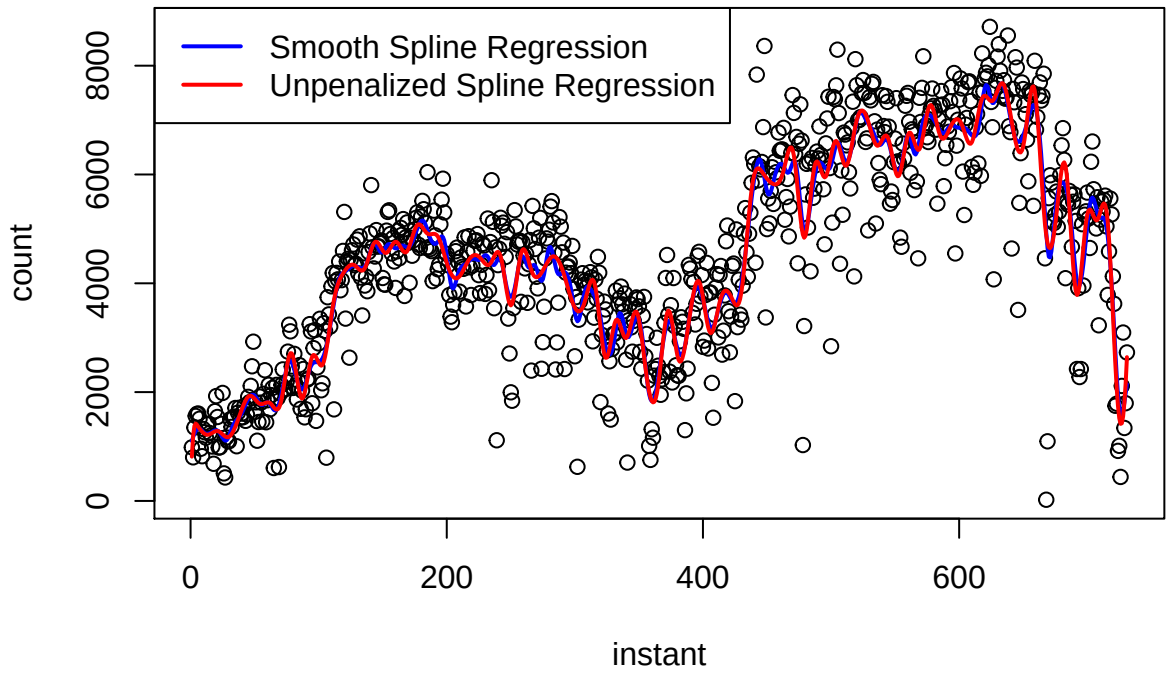
Give a graphic with the scatter plot and the estimated regression function ^m(instant).



Estimate now m(instant) by unpenalized regression splines combining the R functions bs and lm, using the knots my.knots <- quantile(instant,((1:n.knots)-.5)/n.knots) where n:knots is the previous value of df minus 4.

```
linearModel <- lm(bikes$cnt ~ bs(bikes$instant,
                          knots=(quantile(bikes$instant,
                                          ((1:(smoothSpline$df-4))-.5)/(smoothSpline$df-4))),
                          Boundary.knots = (range(bikes$instant)+c(-1,1)*.1*(diff(range(bikes$instant))
```

Give a graphic with the scatter plot and the two estimated regression functions.

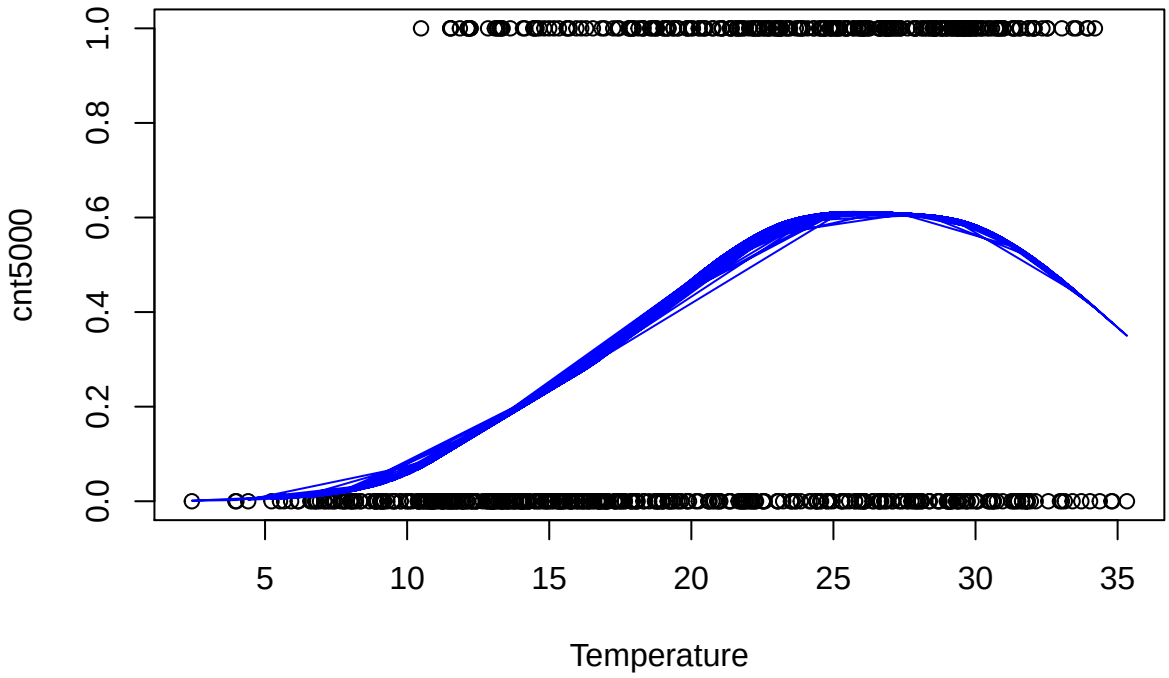


2

We define a new variable cnt.5000 taking the value 1 for days such that the number of total rental bikes is larger than or equal to 5000, on 0 otherwise.

Use the function logistic.IRWLS.splines to fit the non-parametric binary regression cnt.5000 as a function of the temperature, using df=6. In which range of temperatures is Pr(cnt >= 5000jtemp) larger than 0,5?

```
binaryRegression  <- logistic.IRWLS.splines(x = bikes$temp,
                                             y = bikes$cnt.5000,
                                             df= 6,
                                             all.knots=TRUE)
plot(bikes$temp,bikes$cnt.5000,
     col=1,
     xlab="Temperature",
     ylab="cnt5000")
lines(bikes$temp,binaryRegression$fitted.values,col="blue")
```




```
(round(min(bikes$temp[binaryRegression$fitted.values>0.5])))
```

```
## [1] 21
```

```
(round(max(bikes$temp[binaryRegression$fitted.values>0.5])))
```

```
## [1] 32
```

The range of temperatures is Pr(cnt >= 5000jtemp) larger than 0,5 is from 21ºC to 32ºC.