비모수함수추정을 위한 보충 R 코드와 결과

서울대학교
통계학과

August 18, 2018

Contents

1 준비

1.1 패키지 로딩

```
library(ISLR)
attach(Wage)
```

1.2 다항회귀와 계단함수

```
fit=lm(wage~poly(age,4),data=Wage)
coef(summary(fit))

##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)     111.70     0.7287 153.283 0.000e+00
```

```
## poly(age, 4)1   447.07    39.9148  11.201 1.485e-28
## poly(age, 4)2  -478.32    39.9148 -11.983 2.356e-32
## poly(age, 4)3   125.52    39.9148   3.145 1.679e-03
## poly(age, 4)4   -77.91    39.9148  -1.952 5.104e-02
```

```
fit2=lm(wage~poly(age,4,raw=T),data=Wage)
coef(summary(fit2))
```

```
##                         Estimate Std. Error t value  Pr(>|t|)
## (Intercept)            -1.842e+02  6.004e+01  -3.067 0.0021803
## poly(age, 4, raw = T)1  2.125e+01  5.887e+00   3.609 0.0003124
## poly(age, 4, raw = T)2 -5.639e-01  2.061e-01  -2.736 0.0062606
## poly(age, 4, raw = T)3  6.811e-03  3.066e-03   2.221 0.0263978
## poly(age, 4, raw = T)4 -3.204e-05  1.641e-05  -1.952 0.0510386
```

```
fit2a=lm(wage~age+I(age^2)+I(age^3)+I(age^4),data=Wage)
coef(fit2a)
```

```
## (Intercept)         age    I(age^2)    I(age^3)    I(age^4)
##  -1.842e+02   2.125e+01  -5.639e-01   6.811e-03  -3.204e-05
```

```
fit2b=lm(wage~cbind(age,age^2,age^3,age^4),data=Wage)
agelims=range(age)
age.grid=seq(from=agelims[1],to=agelims[2])
preds=predict(fit,newdata=list(age=age.grid),se=TRUE)
se.bands=cbind(preds$fit+2*preds$se.fit,preds$fit-2*preds$se.fit)
par(mfrow=c(1,2),mar=c(4.5,4.5,1,1),oma=c(0,0,4,0))
plot(age,wage,xlim=agelims,cex=.5,col="darkgrey")
title("Degree-4 Polynomial",outer=T)
lines(age.grid,preds$fit,lwd=2,col="blue")
matlines(age.grid,se.bands,lwd=1,col="blue",lty=3)
preds2=predict(fit2,newdata=list(age=age.grid),se=TRUE)
max(abs(preds$fit-preds2$fit))
```

```
## [1] 7.816e-11
```

```
fit.1=lm(wage~age,data=Wage)
fit.2=lm(wage~poly(age,2),data=Wage)
```

```r
fit.3=lm(wage~poly(age,3),data=Wage)
fit.4=lm(wage~poly(age,4),data=Wage)
fit.5=lm(wage~poly(age,5),data=Wage)
anova(fit.1,fit.2,fit.3,fit.4,fit.5)
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ age
## Model 2: wage ~ poly(age, 2)
## Model 3: wage ~ poly(age, 3)
## Model 4: wage ~ poly(age, 4)
## Model 5: wage ~ poly(age, 5)
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   2998 5022216
## 2   2997 4793430  1    228786 143.59 <2e-16 ***
## 3   2996 4777674  1     15756   9.89 0.0017 **
## 4   2995 4771604  1      6070   3.81 0.0510 .
## 5   2994 4770322  1      1283   0.80 0.3697
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
coef(summary(fit.5))
```

```
##               Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)     111.70     0.7288 153.2780 0.000e+00
## poly(age, 5)1   447.07    39.9161  11.2002 1.491e-28
## poly(age, 5)2  -478.32    39.9161 -11.9830 2.368e-32
## poly(age, 5)3   125.52    39.9161   3.1446 1.679e-03
## poly(age, 5)4   -77.91    39.9161  -1.9519 5.105e-02
## poly(age, 5)5   -35.81    39.9161  -0.8972 3.697e-01
```

```r
(-11.983)^2
```

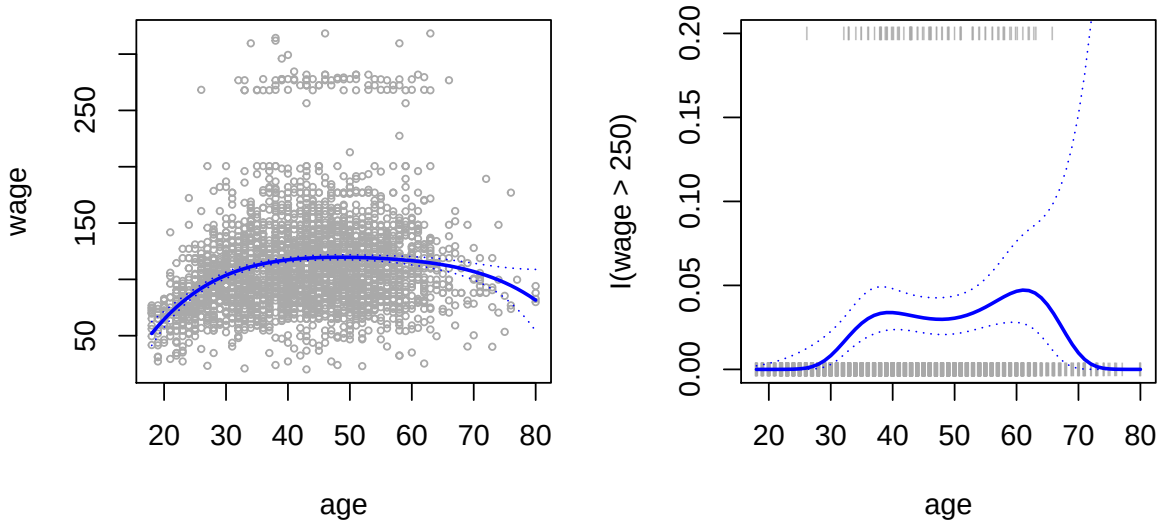
```
## [1] 143.6
```

```r
fit.1=lm(wage~education+age,data=Wage)
fit.2=lm(wage~education+poly(age,2),data=Wage)
fit.3=lm(wage~education+poly(age,3),data=Wage)
```

```r
anova(fit.1,fit.2,fit.3)
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ education + age
## Model 2: wage ~ education + poly(age, 2)
## Model 3: wage ~ education + poly(age, 3)
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   2994 3867992
## 2   2993 3725395  1    142597 114.70 <2e-16 ***
## 3   2992 3719809  1      5587   4.49  0.034 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
fit=glm(I(wage>250)~poly(age,4),data=Wage,family=binomial)
preds=predict(fit,newdata=list(age=age.grid),se=T)
pfit=exp(preds$fit)/(1+exp(preds$fit))
se.bands.logit = cbind(preds$fit+2*preds$se.fit, preds$fit-2*preds$se.fit)
se.bands = exp(se.bands.logit)/(1+exp(se.bands.logit))
preds=predict(fit,newdata=list(age=age.grid),type="response",se=T)
plot(age,I(wage>250),xlim=agelims,type="n",ylim=c(0,.2))
points(jitter(age), I((wage>250)/5),cex=.5,pch="|",col="darkgrey")
lines(age.grid,pfit,lwd=2, col="blue")
matlines(age.grid,se.bands,lwd=1,col="blue",lty=3)
```




```
table(cut(age,4))
```

```
##
## (17.9,33.5]   (33.5,49]   (49,64.5] (64.5,80.1]
##         750        1399         779          72
```

```
fit=lm(wage~cut(age,4),data=Wage)
coef(summary(fit))
```

```
##                         Estimate Std. Error t value  Pr(>|t|)
## (Intercept)               94.158      1.476  63.790 0.000e+00
## cut(age, 4)(33.5,49]      24.053      1.829  13.148 1.982e-38
## cut(age, 4)(49,64.5]      23.665      2.068  11.443 1.041e-29
## cut(age, 4)(64.5,80.1]     7.641      4.987   1.532 1.256e-01
```

1.3 스플라인

```
library(splines)
fit=lm(wage~bs(age,knots=c(25,40,60)),data=Wage)
pred=predict(fit,newdata=list(age=age.grid),se=T)
```

```r
plot(age,wage,col="gray")
lines(age.grid,pred$fit,lwd=2)
lines(age.grid,pred$fit+2*pred$se,lty="dashed")
lines(age.grid,pred$fit-2*pred$se,lty="dashed")
dim(bs(age,knots=c(25,40,60)))
```

```
## [1] 3000    6
```

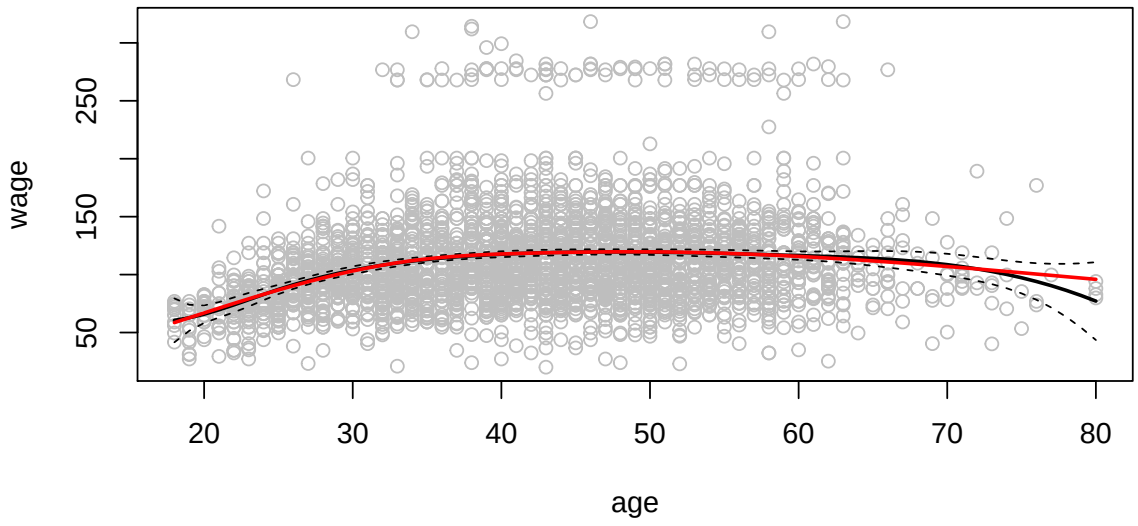
```r
dim(bs(age,df=6))
```

```
## [1] 3000    6
```

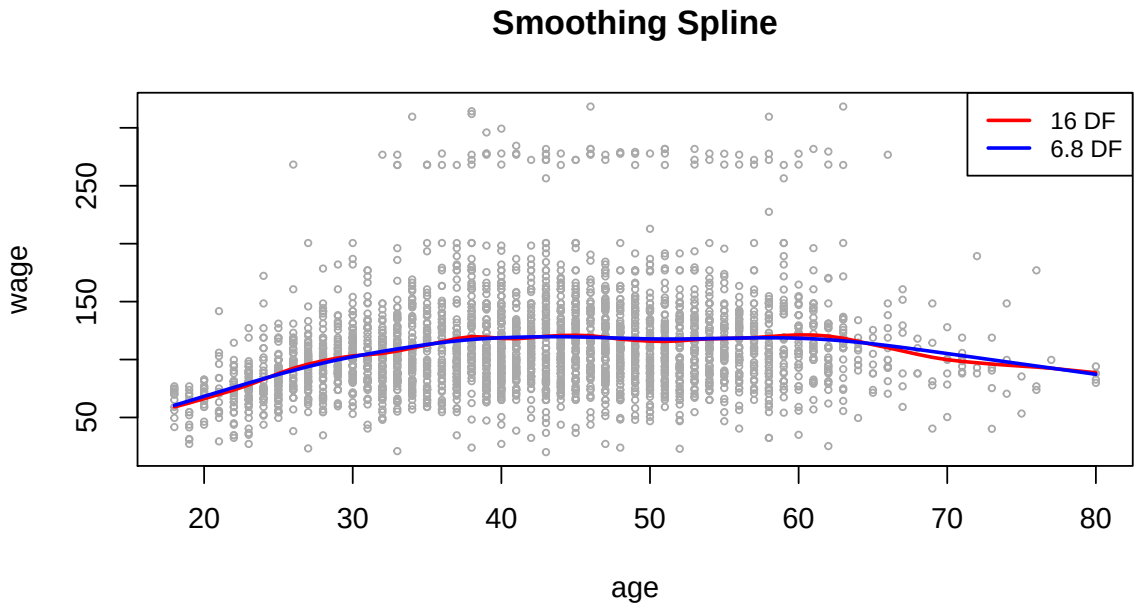
```r
attr(bs(age,df=6),"knots")
```

```
##   25%   50%   75%
## 33.75 42.00 51.00
```

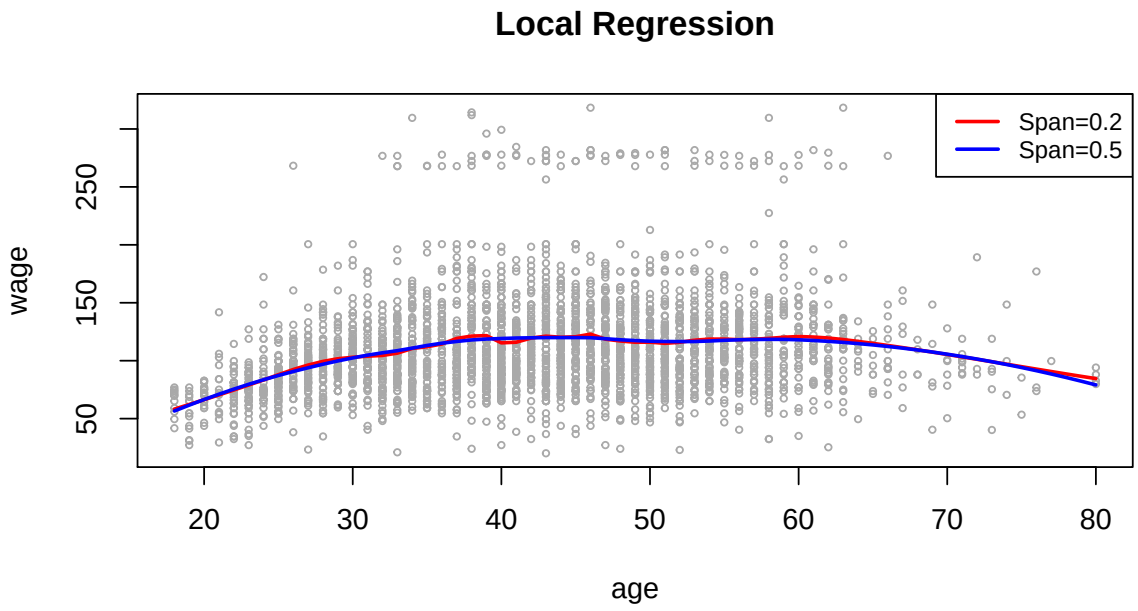
```r
fit2=lm(wage~ns(age,df=4),data=Wage)
pred2=predict(fit2,newdata=list(age=age.grid),se=T)
lines(age.grid, pred2$fit,col="red",lwd=2)
```



```r
plot(age,wage,xlim=agelims,cex=.5,col="darkgrey")
title("Smoothing Spline")
fit=smooth.spline(age,wage,df=16)
fit2=smooth.spline(age,wage,cv=TRUE)

## Warning: cross-validation with non-unique 'x' values seems doubtful

fit2$df

## [1] 6.795

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)
```




```r
plot(age,wage,xlim=agelims,cex=.5,col="darkgrey")
title("Local Regression")
fit=loess(wage~age,span=.2,data=Wage)
fit2=loess(wage~age,span=.5,data=Wage)
lines(age.grid,predict(fit,data.frame(age=age.grid)),col="red",lwd=2)
lines(age.grid,predict(fit2,data.frame(age=age.grid)),col="blue",lwd=2)
legend("topright",legend=c("Span=0.2","Span=0.5"),col=c("red","blue"),lty=1,lwd=2,cex=.8)
```



1.4 가법회귀모형

```
gam1=lm(wage~ns(year,4)+ns(age,5)+education,data=Wage)
library(gam)

gam.m3=gam(wage~s(year,4)+s(age,5)+education,data=Wage)

par(mfrow=c(1,3))
plot(gam.m3, se=TRUE,col="blue")

plot.gam(gam1, se=TRUE, col="red")

gam.m1=gam(wage~s(age,5)+education,data=Wage)

gam.m2=gam(wage~year+s(age,5)+education,data=Wage)

anova(gam.m1,gam.m2,gam.m3,test="F")

summary(gam.m3)

preds=predict(gam.m2,newdata=Wage)

gam.lo=gam(wage~s(year,df=4)+lo(age,span=0.7)+education,data=Wage)
```

```r
plot.gam(gam.lo, se=TRUE, col="green")

gam.lo.i=gam(wage~lo(year,age,span=0.5)+education,data=Wage)

library(akima)

plot(gam.lo.i)

gam.lr=gam(I(wage>250)~year+s(age,df=5)+education,family=binomial,data=Wage)

par(mfrow=c(1,3))
plot(gam.lr,se=T,col="green")

table(education,I(wage>250))

##
## education             FALSE TRUE
##   1. < HS Grad          268    0
##   2. HS Grad            966    5
##   3. Some College       643    7
##   4. College Grad       663   22
##   5. Advanced Degree    381   45

gam.lr.s=gam(I(wage>250)~year+s(age,df=5)+education,family=binomial,data=Wage,subset=(education!="1. < I

plot(gam.lr.s,se=T,col="green")
```