第7章

非 线 性 模 型

Moving Beyond Linearity

非线性模型

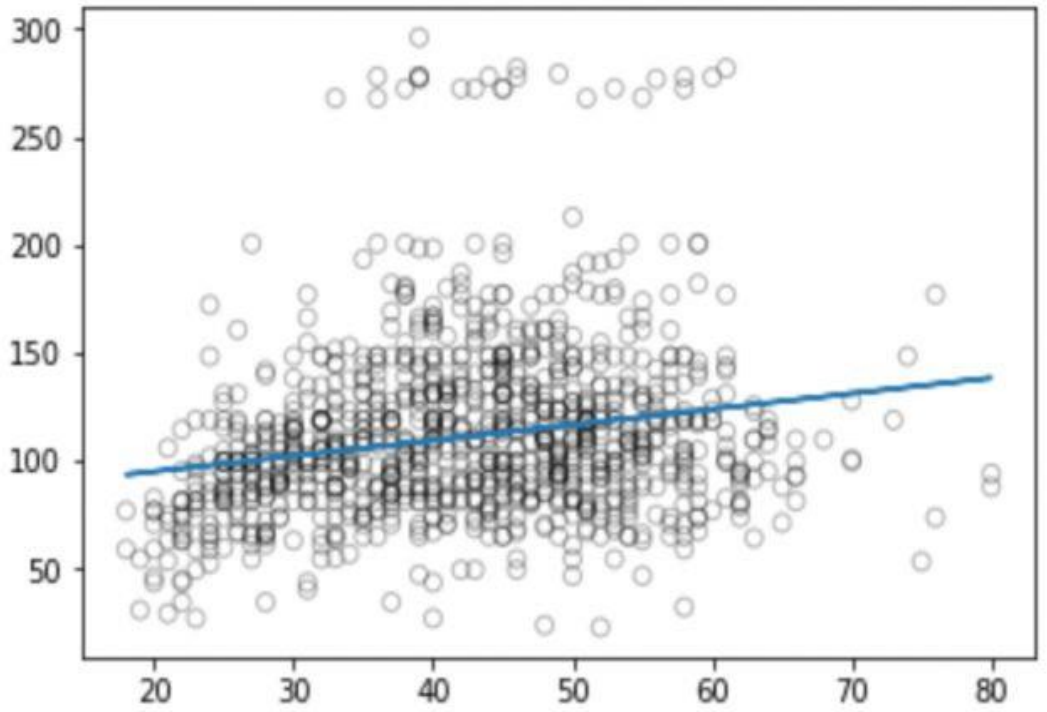
线性模型的线性表达式：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p$$

式中，Y是因变量，X是自变量，也就是特征，β则是分配给特征的权值系数。

线性回归中只有一个特征时：

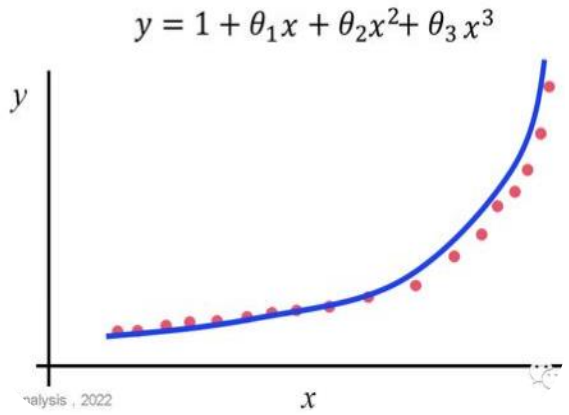
$$Y = \beta_0 + \beta_1 X_1$$



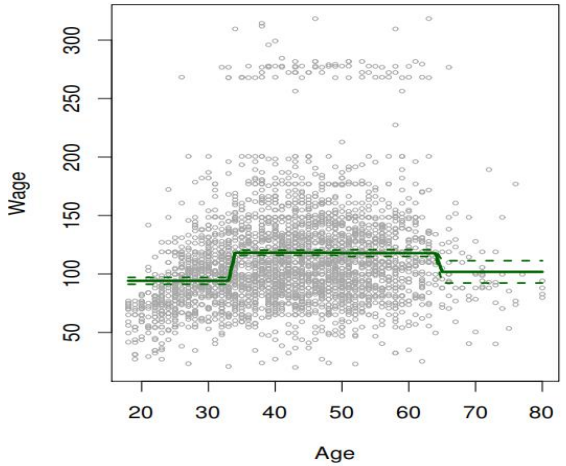
线性模型预测效果并不理想！

非线性模型

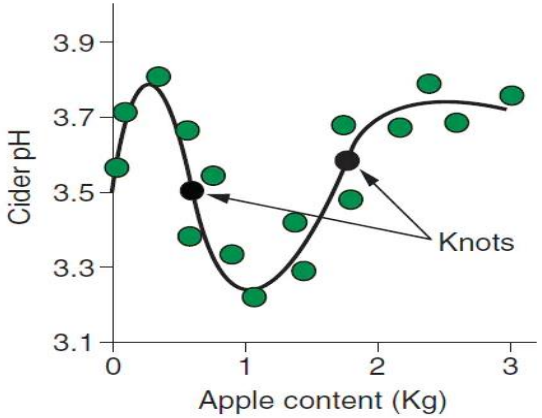
多项式回归



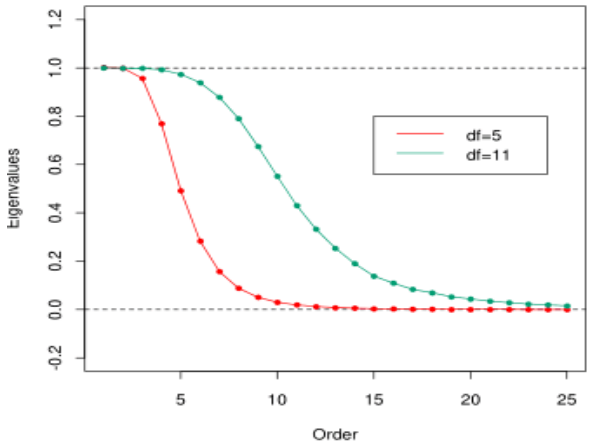
阶梯函数



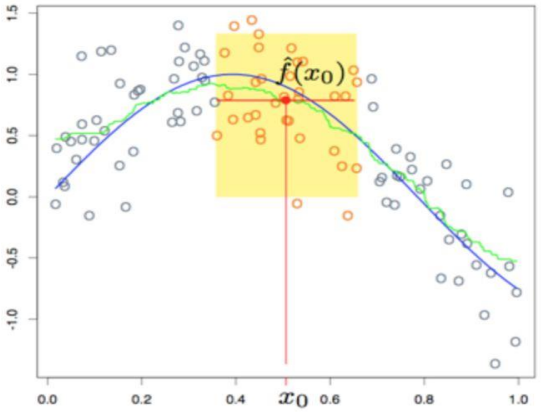
回归样条



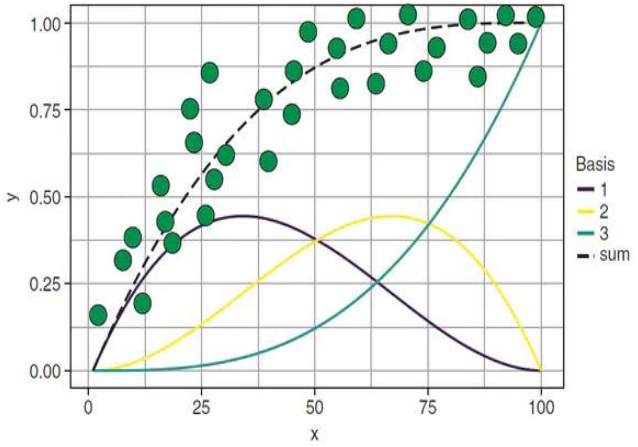
光滑样条



局部回归



广义可加模型



1 多项式回归

多项式回归（polynomial regression）：以预测变量的幂作为新的预测变量以替代原始变量。例如，三次回归模型有三个预测变量X,X^2，X^3，是一种简单实用的表达数据非线性关系的模型。

标准线性模型：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p$$

将标准线性模型换成一个多项式函数：

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \epsilon_i$$

对阶数d选择不宜过大，一般不大于3或4。

1 多项式回归

Wage数据集为例，此数据集是以美国亚特兰大中部地区男员工收入水平为背景的调查数据，尝试分析比较研究工资水平（wage）和年龄（age）之间的关系。

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

	row.names	year	age	maritl	race	education	region	jobclass	health	health_ins	logwage	wage
1	231655	2006	18	1. Never Married	1. White	1. < HS Grad	2. Middle Atlantic	1. Industrial	1. <=Good	2. No	4.318063	75.04315
2	86582	2004	24	1. Never Married	1. White	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	2. No	4.255273	70.47602
3	161300	2003	45	2. Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	1. <=Good	1. Yes	4.875061	130.9822
4	155159	2003	43	2. Married	3. Asian	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	5.041393	154.6853
5	11443	2005	50	4. Divorced	1. White	2. HS Grad	2. Middle Atlantic	2. Information	1. <=Good	1. Yes	4.318063	75.04315
6	376662	2008	54	2. Married	1. White	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	4.845098	127.1157
7	450601	2009	44	2. Married	4. Other	3. Some College	2. Middle Atlantic	1. Industrial	2. >=Very Good	1. Yes	5.133021	169.5285
8	377954	2008	30	1. Never Married	3. Asian	3. Some College	2. Middle Atlantic	2. Information	1. <=Good	1. Yes	4.716003	111.7208
9	228963	2006	41	1. Never Married	2. Black	3. Some College	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	4.778151	118.8844
10	81404	2004	52	2. Married	1. White	2. HS Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	4.857332	128.6805
11	302778	2007	45	4. Divorced	1. White	3. Some College	2. Middle Atlantic	2. Information	1. <=Good	1. Yes	4.763428	117.1468
12	305706	2007	34	2. Married	1. White	2. HS Grad	2. Middle Atlantic	1. Industrial	2. >=Very Good	2. No	4.39794	81.28325

1 多项式回归

给定某个特定的age值x_0，计算得到拟合值：

$$\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_2 x_0^2 + \hat{\beta}_3 x_0^3 + \hat{\beta}_4 x_0^4$$

- 最小二乘法能到每个系数 $\hat{\beta}_j$ 的方差估计以及每一对系数估计之间的协方差。
- 通过这些可得到 $\hat{f}(x_0)$ 的方差，这样 $\hat{f}(x_0)$ 的逐点标准误差就是其方差的平方根。

1 多项式回归

给定某个特定的age值x_0，计算得到拟合值：

$$\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_2 x_0^2 + \hat{\beta}_3 x_0^3 + \hat{\beta}_4 x_0^4$$

R中可采用lm( ) 函数拟合多项式回归模型。以下拟合自由度为4的多项式模型。poly( ) 函数避免了编辑带有age幂项的繁琐公式，其返回的是一个矩阵，每一列都是正交多项式的基，即矩阵每一列都是age，age^2,age^3和age^4的线性组合。

```
> fit=lm(wage~poly(age,4),data=Wage)
> coef(summary(fit))
                Estimate Std. Error    t value     Pr(>|t|)
(Intercept)    111.70361  0.7287409 153.283015 0.000000e+00
poly(age, 4)1  447.06785 39.9147851  11.200558 1.484604e-28
poly(age, 4)2 -478.31581 39.9147851 -11.983424 2.355831e-32
poly(age, 4)3  125.52169 39.9147851   3.144742 1.678622e-03
poly(age, 4)4  -77.91118 39.9147851  -1.951938 5.103865e-02
```

1 多项式回归

poly( ) 函数也可通过内加raw=TRUE参数，实现直接估计age，age^2,age^3和age^4。

```
> fit2=lm(wage~poly(age,4,raw=T),data=Wage)
> coef(summary(fit2))
                        Estimate   Std. Error   t value     Pr(>|t|)
(Intercept)         -1.841542e+02 6.004038e+01 -3.067172 0.0021802539
poly(age, 4, raw = T)1  2.124552e+01 5.886748e+00  3.609042 0.0003123618
poly(age, 4, raw = T)2 -5.638593e-01 2.061083e-01 -2.735743 0.0062606446
poly(age, 4, raw = T)3  6.810688e-03 3.065931e-03  2.221409 0.0263977518
poly(age, 4, raw = T)4 -3.203830e-05 1.641359e-05 -1.951938 0.0510386498
```

等价表示方法：

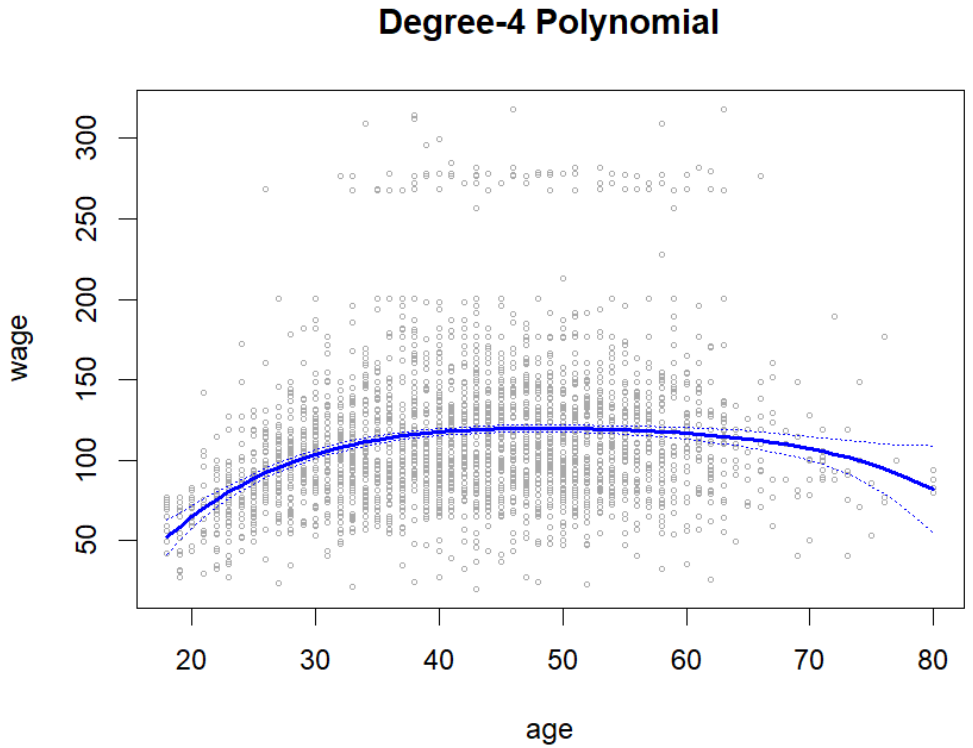
```
> fit2a=lm(wage~age+I(age^2)+I(age^3)+I(age^4),data=Wage)
> coef(fit2a)
  (Intercept)           age      I(age^2)      I(age^3)      I(age^4)
-1.841542e+02  2.124552e+01 -5.638593e-01  6.810688e-03 -3.203830e-05
```

```
> fit2b=lm(wage~cbind(age,age^2,age^3,age^4),data=Wage)
```

1 多项式回归

在每个点计算相应位置的标准误差，然后画出拟合值曲线以及距拟合值曲线的两倍标准误差的曲线。

```
> 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)
```



构造一组age值，并调用predict( ) 函数进行预测，同时给出相应的标准误差。画出数据点散点图以及用4次多项式函数拟合的结果。

```
> 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)
```

图1 实线表示wage关于age的四阶多项式模型曲线，用最小二乘法拟合。虚线为95%置信区间

1 多项式回归

把wage看作一个二元变量就能将数据分成两个组，这样以age的多项式函数作为预测变量的逻辑斯谛回归就能用来预测这个二元响应变量。即拟合的为下列模型：

$$\Pr(y_i > 250|x_i) = \frac{\exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_d x_i^d)}{1 + \exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_d x_i^d)}$$

应用glm( ) 函数以及family=“binomial”参数拟合多项式逻辑斯谛回归模型前，需生成一个响应变量。I（）函数被用于生成一个二元响应变量。

```
> fit=glm(I(wage>250)~poly(age,4),data=Wage,family=binomial)
```

1 多项式回归

绘制地毯图

```
> 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)
```

上部和下部的灰色部分分别对应高收入人群和低收入人群。图中显示右半部分的置信区间很宽？

样本量足够（n=3000），但只有79个高收入的人，导致了估计系数有较大的方差，所以置信区间也较宽。

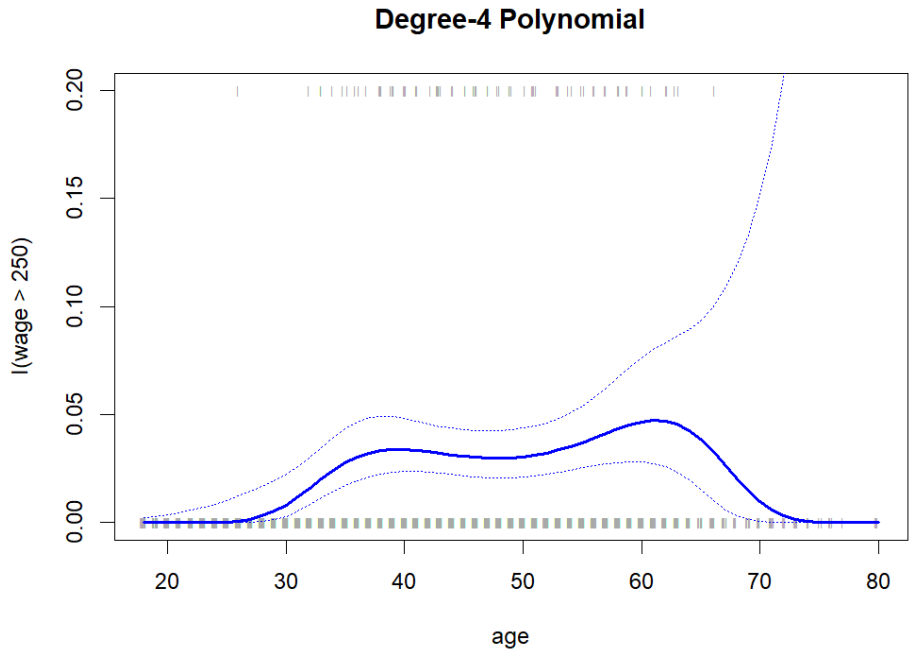


图2 针对二元变量wage>250 逻辑斯谛回归模型建模结果，通常采用4阶多项式.蓝色实线表示wage>250的后验概率，虚线表示95%置信区间。

2 阶梯函数

阶梯函数（step function）拟合是将某个预测变量的取值空间切割成K个不同区域，以此来生成一个新的定性变量，分段拟合一个常量函数。

首先在X取值空间上创建分割点$c_1, c_2, \cdots, c_k$，然后构造K+1个新变量如下：

$$
\begin{aligned}
C_0(X) &= I(X < c_1) \\
C_1(X) &= I(c_1 \le X < c_2) \\
C_2(X) &= I(c_2 \le X < c_3) \\
&\vdots \\
C_{k-1}(X) &= I(c_{k-1} \le X < c_k) \\
C_k(X) &= I(c_k \le X)
\end{aligned}
$$

$I(\cdot)$ 是示性函数，当条件成立时返回1否则返回0。

2 阶梯函数

X只能落在K+1个区间中的某一个，于是对任意X的取值，$C_0(X)+C_1(X)+\cdots+C_k(X)=1$

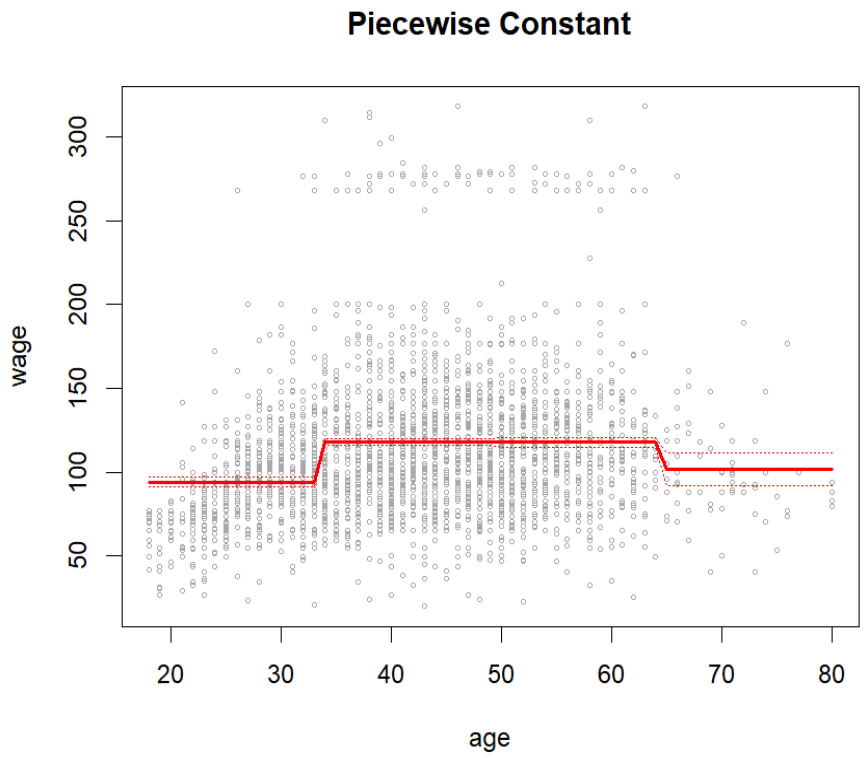
以 $C_1(X),C_2(X),\cdots,C_K(X)$ 为预测变量用最小二乘法来拟合线性模型：

$$y_i=\beta_0+\beta_1C_1(x_i)+\beta_2C_2(x_i)+\cdots+\beta_KC_K(x_i)+\varepsilon_i$$

- 对于X的一个给定值，$C_1(X),C_2(X),\cdots,C_K(X)$ 中至多只有一项系数非零。
- 若X<c_1时，式中每个预测变量均为零，所以β_0即为X<c_1时的Y的平均值。
- 当c_j<X<c_{j+1}时，预测值为$\beta_0+\beta_j$，这样βj为当X由X<c_1增至c_j<X<c_{j+1}时响应变量的平均增量。

2 阶梯函数

以wage数据拟合阶梯函数的效果。



拟合阶梯函数前，需对数据先用cut( ) 函数处理，可自动对年龄变量进行分割点的选择，cut( ) 函数返回的实际是一个有序变量，之后lm( ) 据此生成一系列回归中的哑变量。

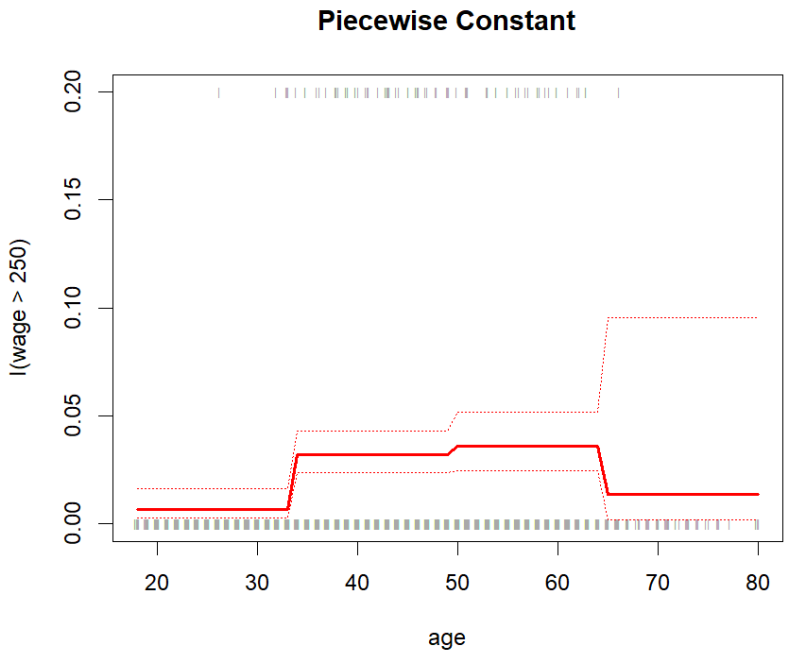
```
> table(cut(age,4))

(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)            94.158392   1.476069 63.789970 0.000000e+00
cut(age, 4)(33.5,49]   24.053491   1.829431 13.148074 1.982315e-38
cut(age, 4)(49,64.5]   23.664559   2.067958 11.443444 1.040750e-29
cut(age, 4)(64.5,80.1]  7.640592   4.987424  1.531972 1.256350e-01
```

图3 Wage数据集。实线表示用阶梯函数拟合wage对age的最小二乘回归的拟合结果，虚线为相应的95%置信区间。

2 阶梯函数

以wage数据拟合阶梯函数的效果。用wage对age拟合逻辑斯谛回归如下：

$$\Pr(y_i > 250 \mid x_i) = \frac{\exp(\beta_0 + \beta_1 C_1(x_i) + \cdots + \beta_K C_K(x_i))}{1 + \exp(\beta_0 + \beta_1 C_1(x_i) + \cdots + \beta_K C_K(x_i))}$$




```
> fit=glm(I(wage>250)~cut(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="red")
> matlines(age.grid,se.bands,lwd=1,col="red",lty=3)
> title("Piecewise Constant")
```

图4 使用阶梯函数，对二元变量wage>250建立逻辑斯谛回归模型。虚线表示95%置信区间。

3 基函数

多项式和阶梯函数回归模型实际上是特殊的基函数方法。基本原理是对变量X的函数或变换 $b_1(X), b_2(X), \cdots, b_k(X)$ 进行建模 。用以下模型来替代线性模型。

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \beta_3 b_3(x_i) + \cdots + \beta_K b_K(x_i) + \varepsilon$$

多项式回归：

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \epsilon_i \longrightarrow b_i(x_i) = x_i^j$$

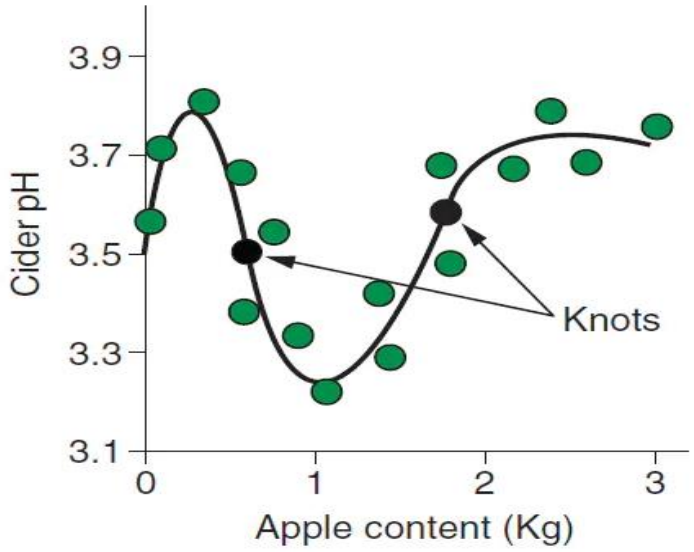
阶梯函数：

$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \cdots + \beta_K C_K(x_i) + \varepsilon_i \longrightarrow b_j(x_i) = I(c_j \le x_i < c_{j+1})$$

对应基函数

4 回归样条

回归样条（regression spline）将X的取值范围切割成K个区域，在每个区域分别独立拟合一个多项式函数。回归样条的多项式一般有一些限制以保证在区域边界或称为结点的位置，使这些多项式得到光滑的连接。



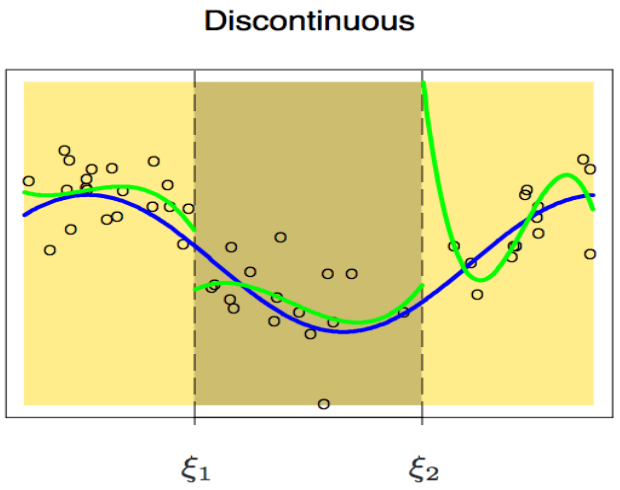
4 回归样条

（1）分段多项式回归

在X的不同区域拟合独立的低阶多项式函数，以此取代在X全部取值范围内拟合高阶多项式。

系数发生变化的临界点称为结点（knot）。

$$y_i = \begin{cases} \beta_{01} + \beta_{11}x_i + \beta_{21}x_i^2 + \beta_{31}x_i^3 + \epsilon_i & \text{if } x_i < c \\ \beta_{02} + \beta_{12}x_i + \beta_{22}x_i^2 + \beta_{32}x_i^3 + \epsilon_i & \text{if } x_i \geq c \end{cases}$$

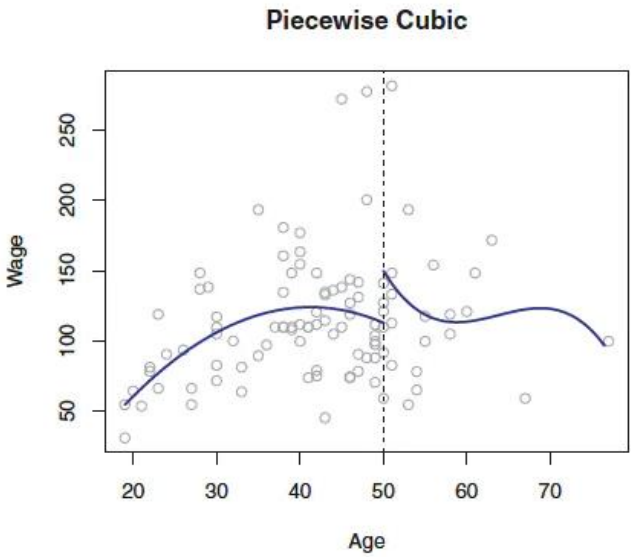


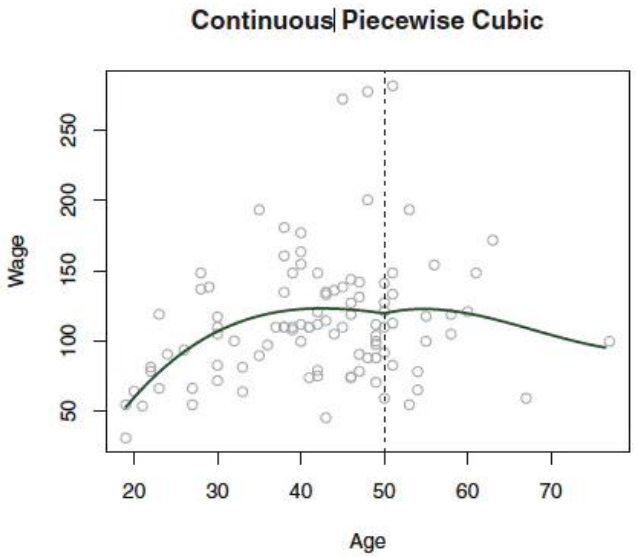
- 不连续的现象。
- 必要的额外限制：任一侧的多项式在节点上应该是连续的。

4 回归样条

（2）约束条件与样条

无限制条件的三次多项式。

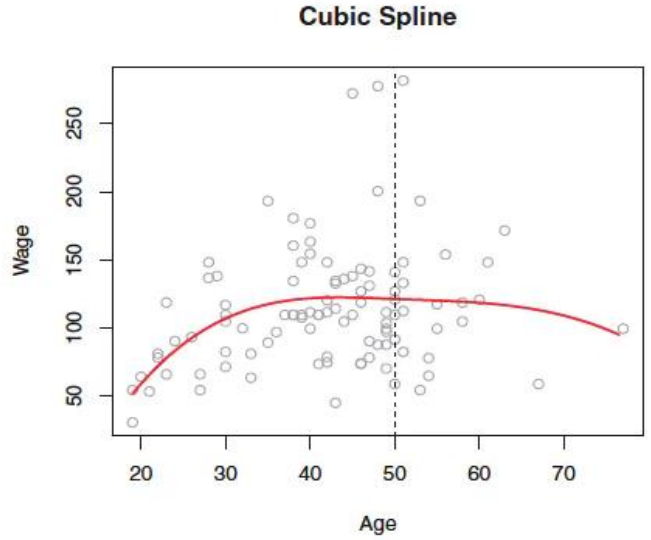


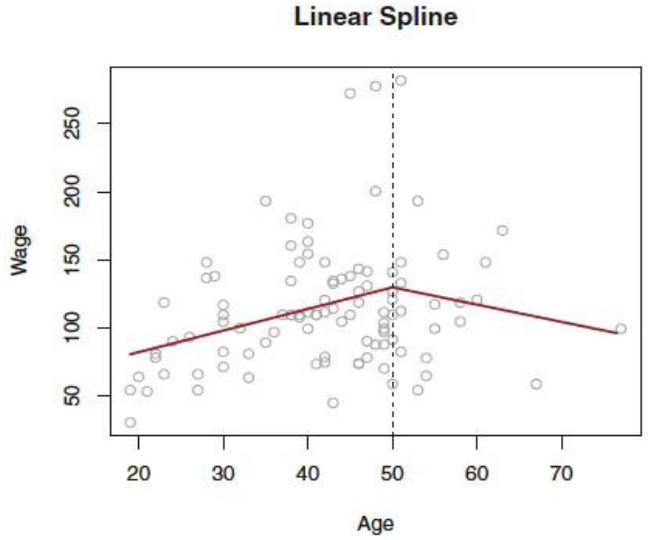


限定多项式在age=50处连续的三次多项式。

限定多项式在age=50连续，且在age的一阶导数和二阶导数都是连续的三次多项式。

三次样条





线性样条，限定在age=50连续。

图5 在Wage数据的一个子集上拟合多个分段多项式模型。

4 回归样条

（3）样条基函数

① 线性样条：由在每个区域内拟合一条直线，同时要求在各结点处满足连续性获得。即一个节点为ξ_k, k = 1，…K的线性样条，是在结点连续的分段线性多项式。

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+1} b_{K+1}(x_i) + \epsilon_i$$

其中b_k是基函数：

$$\begin{aligned} b_1(x_i) &= x_i \\ b_{k+1}(x_i) &= (x_i - \xi_k)_+, \quad k = 1, \ldots, K \end{aligned}$$

这里的$()_+$表示正的部分，ξ_k是结点。

$$(x_i - \xi_k)_+ = \begin{cases} x_i - \xi_k & \text{if } x_i > \xi_k \\ 0 & \text{otherwise} \end{cases}$$

（3）样条基函数

②三次样条：先以三次多项式的基为基础，即x、x^2、x^3，然后在每个节点添加一个截断幂基：

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+3} b_{K+3}(x_i) + \epsilon_i$$

$$\begin{aligned} b_1(x_i) &= x_i \\ b_2(x_i) &= x_i^2 \\ b_3(x_i) &= x_i^3 \\ b_{k+3}(x_i) &= (x_i - \xi_k)_+^3, \quad k = 1, \ldots, K \end{aligned}$$

截断幂基：

$$(x_i - \xi_k)_+^3 = \begin{cases} (x_i - \xi_k)^3 & \text{if } x_i > \xi_k \\ 0 & \text{otherwise} \end{cases}$$

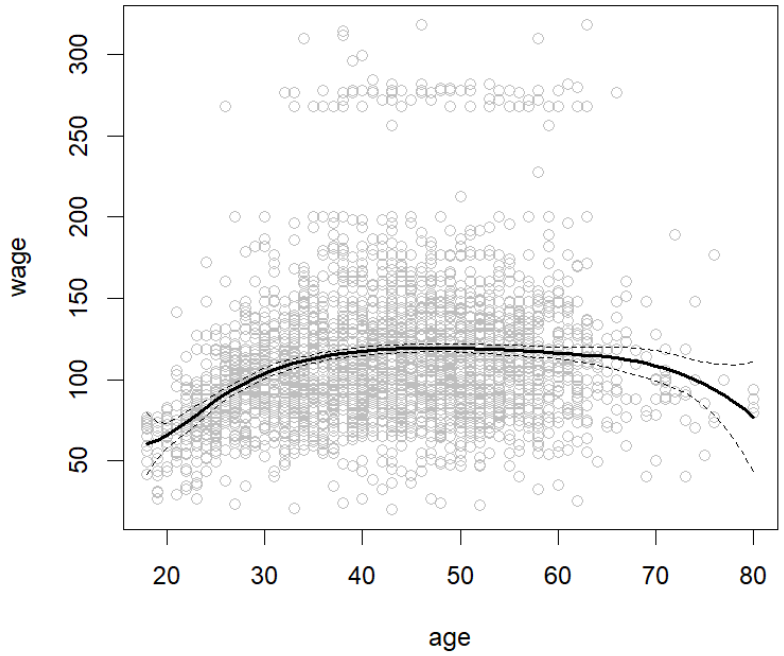
需估计K+4个系数，因此拟合三次样条总共需要K+4个自由度。

4 回归样条

（3）样条基函数

R中splines包拟合回归样条的函数。bs( ) 函数能用来产生针对给定结点的所有样条基函数的矩阵。bs( ) 默认生成三次样条。

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



bs( ) 函数的degree参数方便模型自由选择样条的阶数。

4 回归样条

（3）样条基函数

③自然样条：是附加了边界约束的回归样条，函数在边界区域应该是线性的，这里的边界区域指的是X的值比最小的结点处的值小或比最大的结点处的值大。附加的约束条件使自然样条的估计在边界处更稳定。

R中使用ns( ) 函数拟合自然样条，拟合自由度为4的自然样条时，df=4。knots选项也可手动确定结点.

```
> fit2=lm(wage~ns(age,df=4),data=Wage)
```

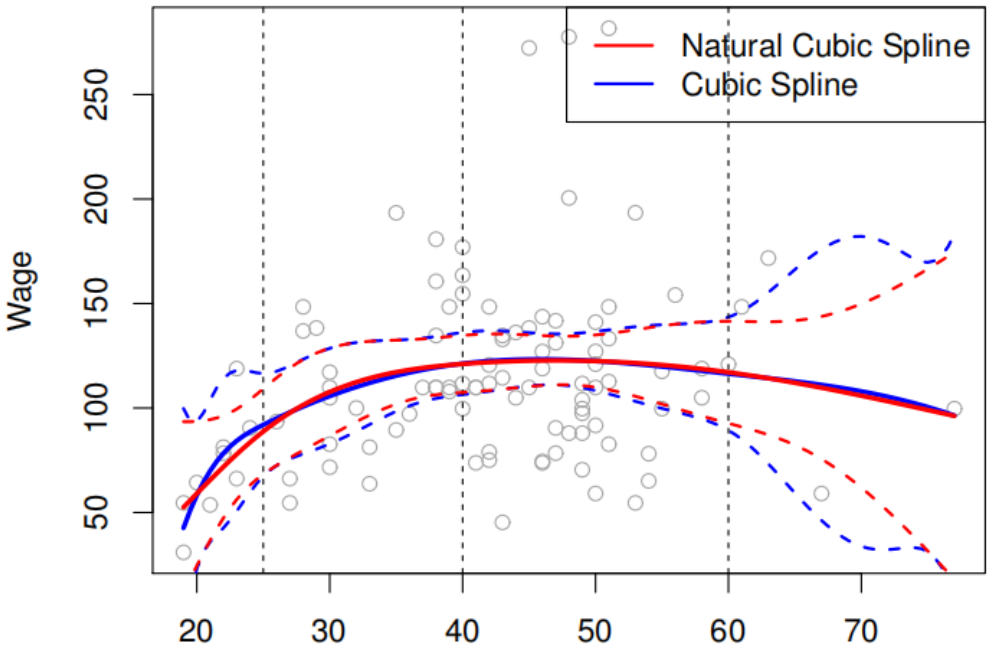


图6 应用三个节点的三次样条和三次自然样条拟合Wage数据的一个子集

4 回归样条

（4）结点选择方式：

让结点在数据上呈现均匀分布：首先确定自由度，然后靠软件自动在数据得均匀分位数点上设置相应个数的结点。具有K个节的自然样条有K个自由度。具有K个结点的三次样条有K + 4个参数或自由度。

- 尝试多个不同结点个数，选择最优曲线。
- 交叉验证方法：首先移除部分数据，用剩余的数据拟合样条函数，接着用拟合的样条函数来对移除的那部分数据做预测，不断重复，最后计算整体交叉验证RSS。

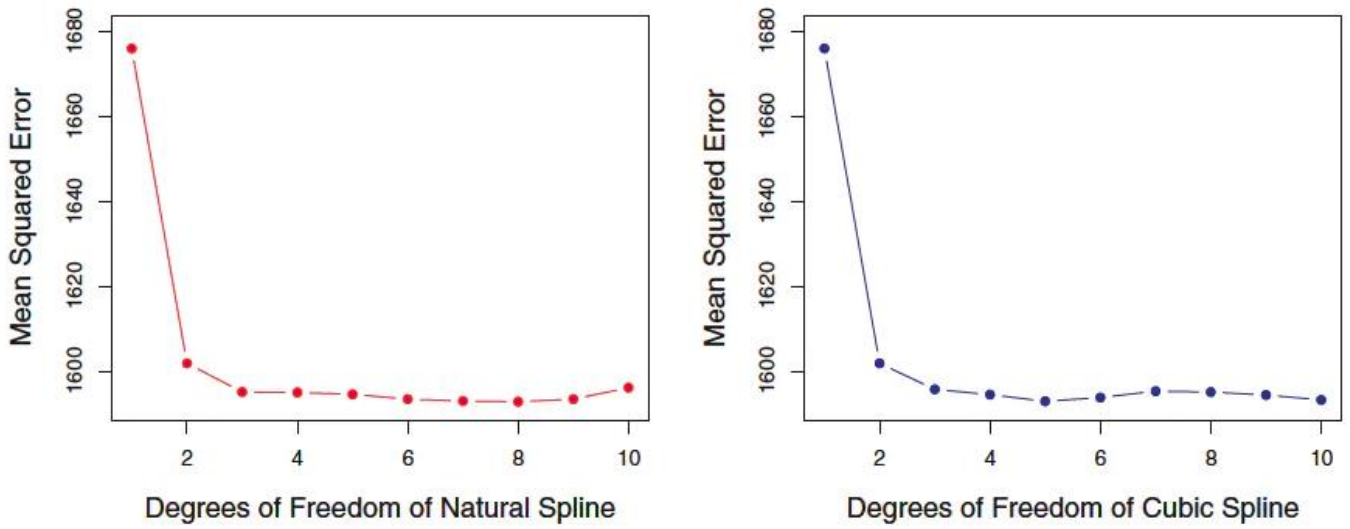


图7 Wage 数据不同自由度下的10折交叉验证的均方误差。左：自然三次样条。右：三次样条

4 回归样条

（5）与多项式回归作对比：

回归样条通常得到的结果比多项式回归更好。多项式回归需设定较高幂次，样条函数通过增加结点个数但保持自由度固定。如下图中多项式回归在尾部区域结果不好。

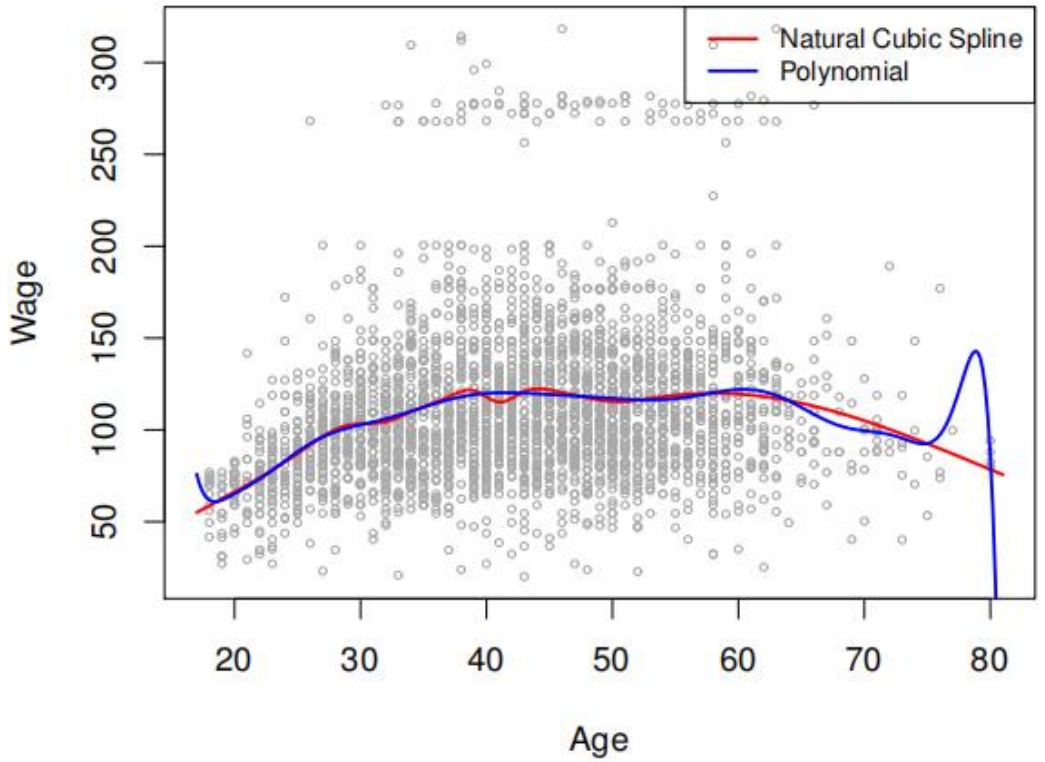


图8 自由度为15的自然三次样条和最高幂次为15的多项式回归在Wage数据的比较结果

5 光滑样条

光滑样条（smoothing spline）： 一般是通过最小化一个带光滑惩罚项的残差平方和的式子来得到光滑样条的结果。

以下是光滑函数g(x)拟合到某些数据的准则，其中λ是一个非负的调节参数：

$$\underset{g\in\mathcal{S}}{\text{minimize}} \sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

损失函数　　波动性惩罚

- 试图使g(x)匹配每个xi上的数据，使RSS尽量小。
- 粗糙度惩罚，控制g(x)的摆动程度。它是由调谐参数调制的λ≥0，λ值越大，函数g越光滑。

  - λ=0时，惩罚项不起作用，函数g会很跳跃。
  - λ→∞时，g会变得非常平稳，接近于直线。

5 光滑样条

选择光滑参数λ：调节参数λ控制光滑样条的粗糙度，同时也控制着**有效自由度**。

$$df_\lambda = \sum_{i=1}^{n} \{\mathbf{S}_\lambda\}_{ii}$$

S_λ是一个n × n矩阵(由x_i和λ决定)。有效自由度时矩阵S_λ的对角线之和。

使用交叉验证方法确定λ的值，选择能够使得交叉验证的RSS尽量小的λ作为解。

$$\mathrm{RSS}_{cv}(\lambda) = \sum_{i=1}^{n} (y_i - \hat{g}_\lambda^{(-i)}(x_i))^2 = \sum_{i=1}^{n} \left[\frac{y_i - \hat{g}_\lambda(x_i)}{1 - \{\mathbf{S}_\lambda\}_{ii}} \right]^2$$

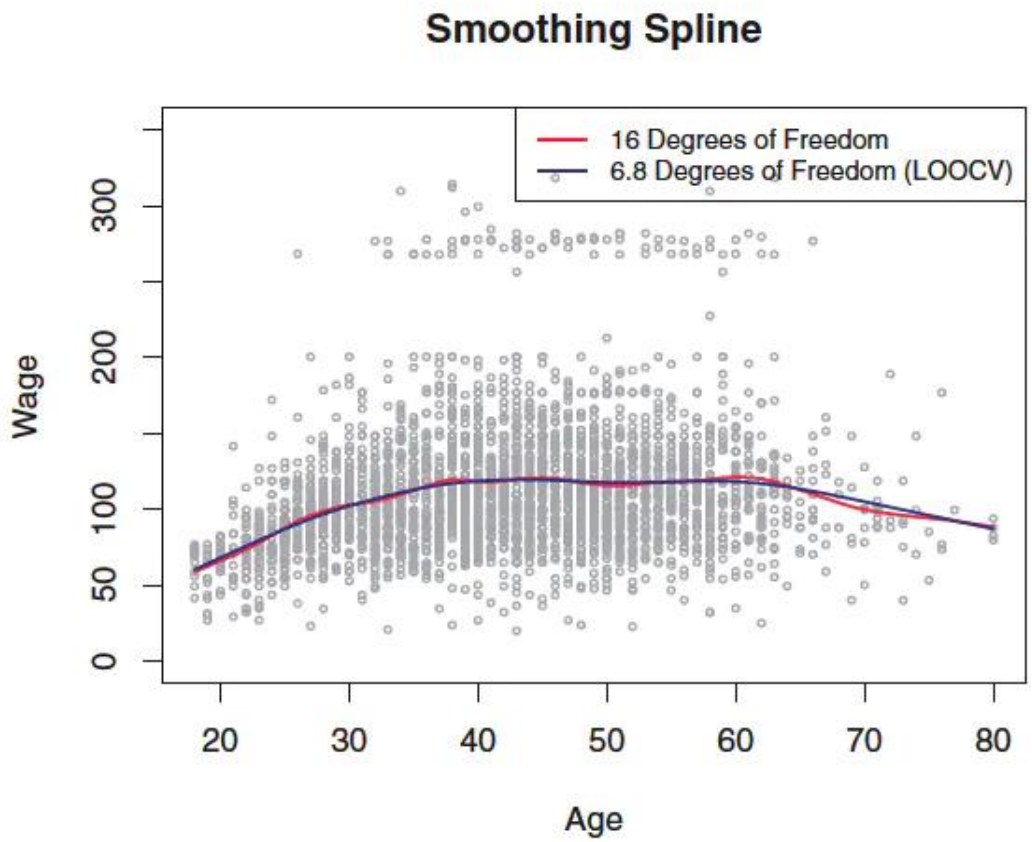
光滑样条在x_i处的拟合值

光滑样条用所有的训练数据集拟合的结果

5 光滑样条

R中光滑样条的拟合采用smooth.splines( )

```
> 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)

> 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)
> plot(age,wage,xlim=agelims,cex=.5,col="darkgrey")
```



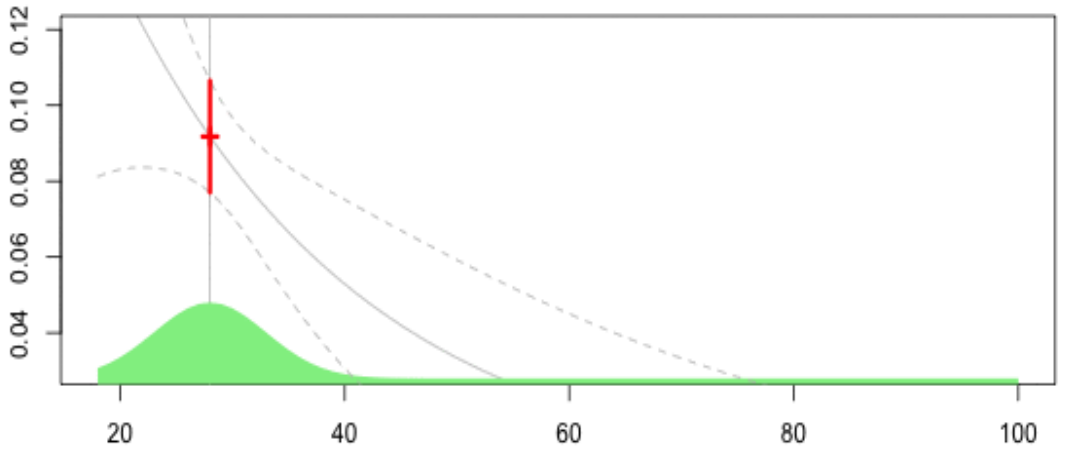
自由度为16的波动会更大一点，并且模型更为复杂。留一交叉验证可得到更好结果。

图9 光滑样条拟合Wage数据。红线表示有效自由度为16的结果。蓝线是λ自动根据留一交叉验证得到的结果，有效自由度为6.8

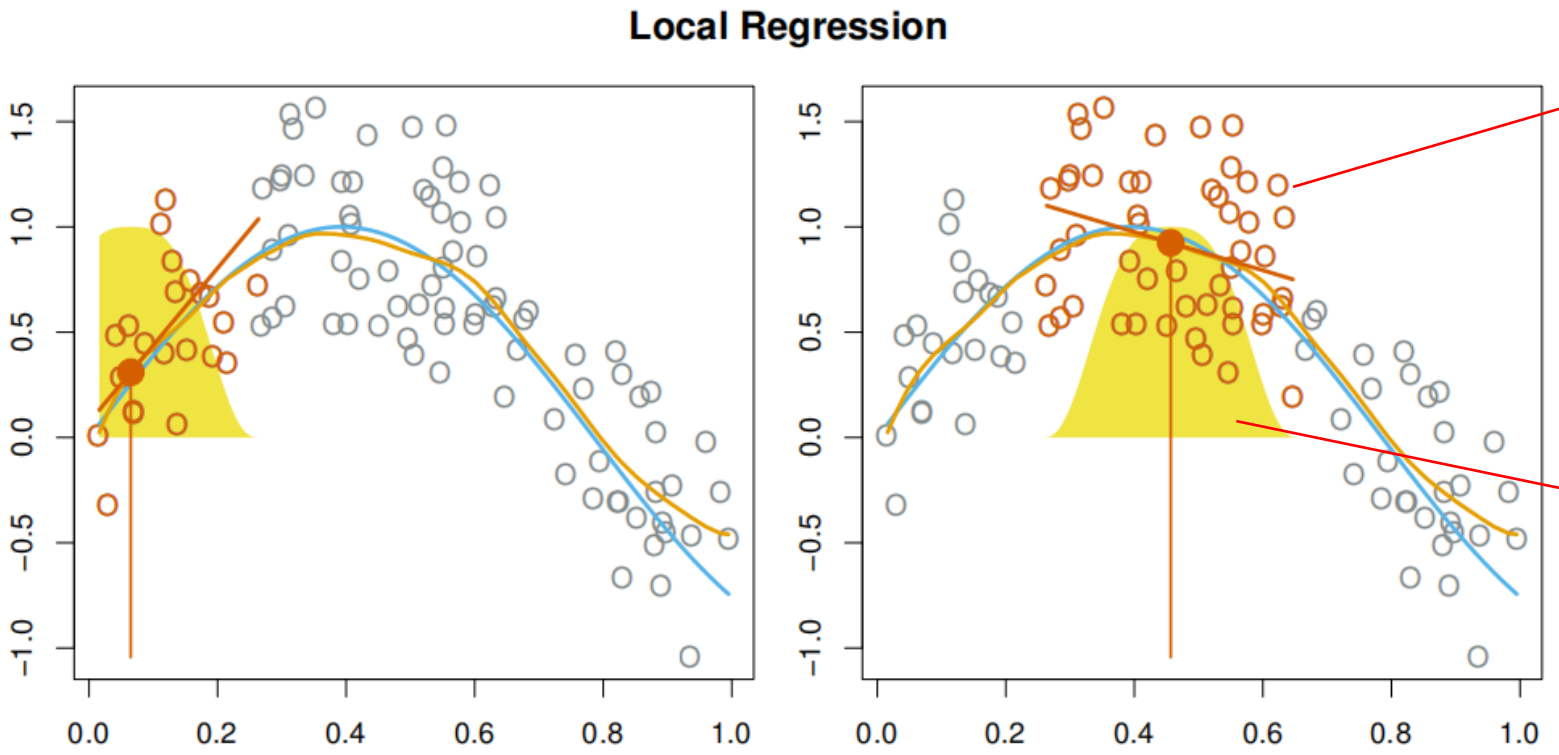
6 局部回归

局部回归（local regression）与样条结果比较相近，最大的差别在于局部回归中的区域之间是可以重叠的，这个条件保证了局部回归整体光滑的拟合结果

- 局部回归涉及仅使用附近的训练观测值来计算目标点x_0 处的拟合度 。
- 可以通过各种方式执行局部回归，某些变量可以全局拟合，而某些局部拟合。



6 局部回归



表示x_0的邻居

赋值给x_0的邻居的权重，随据x_0的距离越远越降至0

图10 模拟数据上的局部回归模型，数据由蓝线对应函数f(x)产生根据浅橘黄色曲线是根据回归拟合得到的结果。

6 局部回归

算法1　在X=x_0处的局部回归模型

1. 选取占所有数据 $s = k/n$ 比例的最靠近 x_0 的数据 x_i。

2. 对选出的数据点赋予其权重 $K_{i0} = K(x_i, x_0)$。离 x_0 最远的点的权重为 0，而最近的点权重最高。那些没有被选中的数据点的权重为 0。

3. 用定义好的权重在 x_i 处拟合加权最小二乘回归，也就是对下式最小化

$$\sum_{i=1}^{n} K_{i0}(y_i - \beta_0 - \beta_1 x_i)^2$$

4. 根据 $\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$ 得到 x_0 的拟合结果.

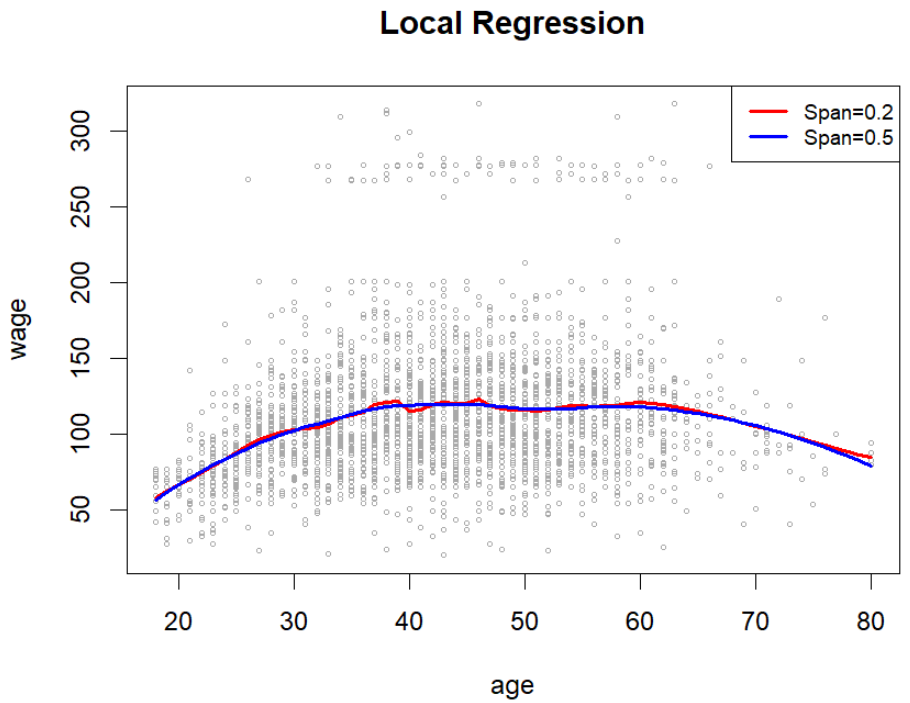
- 权重函数K；
- 间距 s（span）控制非线性拟合的光滑性。

6 局部回归

R中局部回归采用loess( ) 函数

```
> 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)
```

- 设定参数span=0.2和span=0.5 局部回归时使用了预测点的邻域分别囊括20%和50%的数据，得到两个线性模型。
- span越大拟合越光滑。
- R中locifit包也可以用于拟合局部回归模型。



7 广义可加模型

广义可加模型（generalized additive model，GAM）提供一种对标准线性模型进行推广的框架，框架中，每个变量用一个非线性函数替换，同时保持着模型的整体可加性。可用于响应变量定性与定量的情形。

用于回归问题

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \epsilon_i$$

多元线性回归

$$\downarrow$$

$$y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i$$

用于分类问题

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

逻辑斯谛回归

$$\downarrow$$

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)$$

7 广义可加模型

用于回归问题

对多元线性回归模型用一个光滑的非线性函数 $f_j(x_{ij})$ 替代 $\beta_j x_{ij}$

$$y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i$$

这个模型称为可加模型，对每一个X_j对应一个独立的计算单元f_j，然后将分项结果加在一起。

以自然样条为例，假设要对Wage数据拟合这样一个模型：

$$\text{wage} = \beta_0 + f_1(\text{year}) + f_2(\text{age}) + f_3(\text{education})$$

自然样条拟合　　阶梯函数

7 广义可加模型

用于回归问题

应用year和age的自然样条函数和education作为预测变量拟合广义可加模型。

$$\text{wage} = \beta_0 + f_1(\text{year}) + f_2(\text{age}) + f_3(\text{education})$$

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

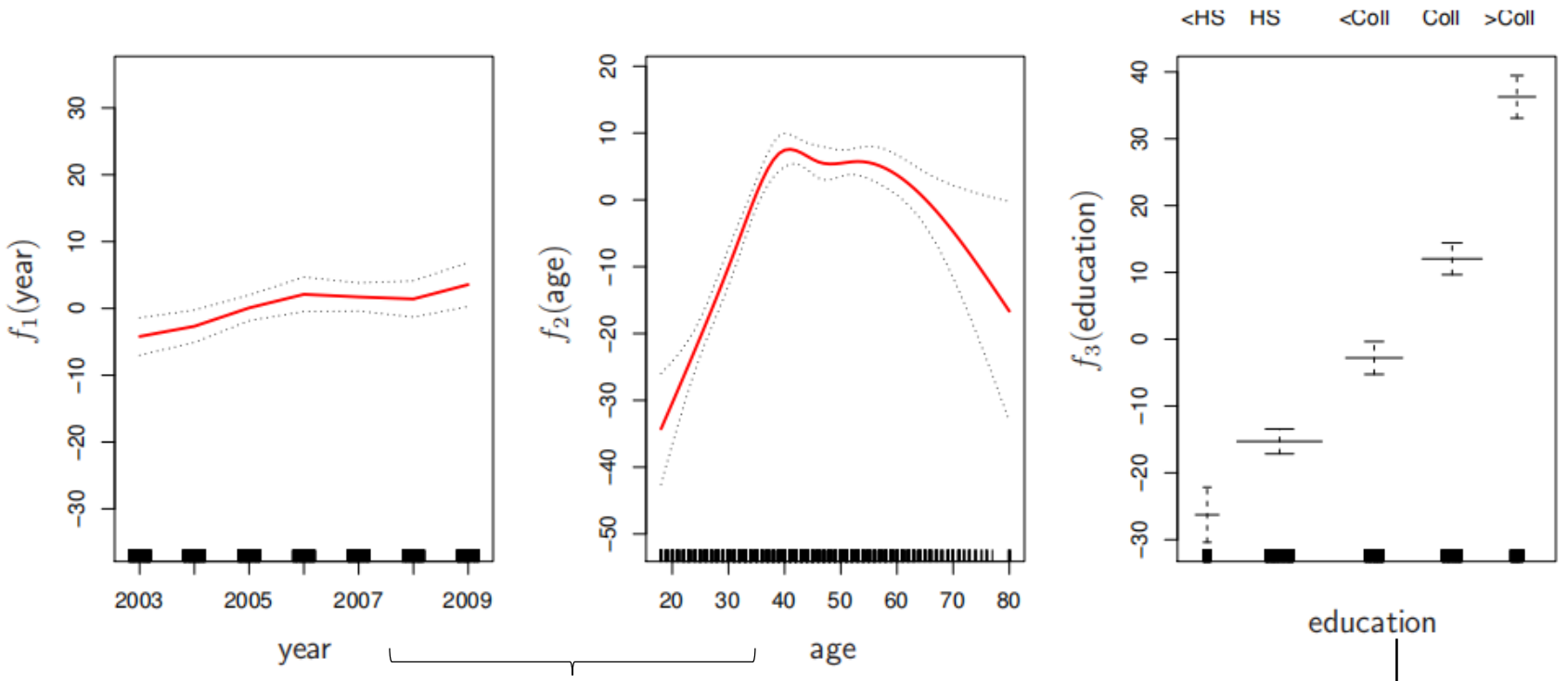
接着改用光滑样条来拟合模型，采用gam库中的s()函数拟合光滑样条。

设置year自由度为4，age为5。

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

7 广义可加模型

$$\text{wage} = \beta_0 + f_1(\text{year}) + f_2(\text{age}) + f_3(\text{education})$$



对应year和age自由度分别为4与5的自然样条

对定性变量education拟合的阶梯函数

图11 对于Wage数据，响应变量wage与其他预测变量的关系图。

7 广义可加模型

GAM的优点与不足：

- GAM 模型可允许对每一个X_j都拟合一个非线性f_i，可自动地对被标准的线性回归模型所忽略的非线性关系进行建模，不需手动为每一个变量设置不同的变形方式。
- 非线性拟合模型能将响应变量预测得更精准。
- 模型是可加的，能在保持其他变量不变的情况下看每个变量X_j对Y单独的影响效果。
- 针对变量X_j的函数f_i 的光滑性也能通过对自由度的分析得到。
- GAM 模型的形式被限定为可加形式。在多变量的情况下，会忽略有意义的交互项。但可以通过增加形式为$X_j \times X_k$使得 GAM 也能够表达交互效应。另外可以增加形式为 $f_{jk}(X_j, X_k)$ 的低维交互项，可应用一些二维光滑方法如局部回归或者二维样条来拟合。

7 广义可加模型

用于分类问题

GAM也可以用于当Y为定性变量的时候，假设Y只取0或者1.令p(x)=Pr(Y=1|X)是给定预测变量值时响应变量为1 的条件概率。下式为逻辑斯谛回归模型：

$$\log\left(\frac{p(X)}{1-p(X)}\right)=\beta_0+f_1(X_1)+f_2(X_2)+\cdots+f_p(X_p)$$

优势比P(Y=1|X)/P(Y=0|X)的对数称为分对数（logit），由上式线性模型推广可得：

$$log(\frac{p(X)}{1-p(X)})=\beta_0+\beta_1\times(\text{year})+f_2(\text{age})+f_3(\text{education})$$

$f_2(\text{age})$ → 光滑样条

$f_3(\text{education})$ → 阶梯函数

7 广义可加模型

用于分类问题

GAM也可以用于当Y为定性变量的时候，假设Y只取0或者1.令p(x)=Pr(Y=1|X)是给定预测变量值时响应变量为1 的条件概率。下式为逻辑斯谛回归模型：

$$\log\left(\frac{p(X)}{1-p(X)}\right)=\beta_0+f_1(X_1)+f_2(X_2)+\cdots+f_p(X_p)$$

优势比P(Y=1|X)/P(Y=0|X)的对数称为分对数（logit），由上式线性模型推广可得：

$$log(\frac{p(X)}{1-p(X)})=\beta_0+\beta_1\times(\text{year})+f_2(\text{age})+f_3(\text{education})$$

R中采用I( ) 函数生成二元相应变量，设置family=binomial

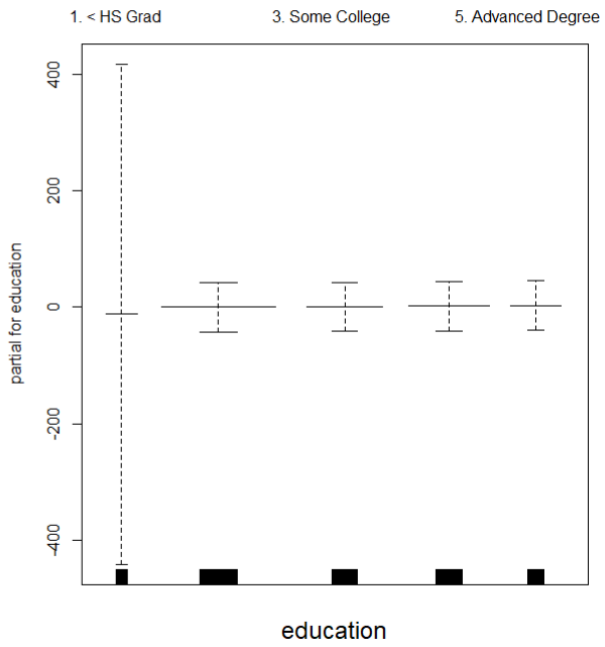
```
> gam.lr=gam(I(wage>250)~year+s(age,df=5)+education,family=binomial,data=Wage)
```

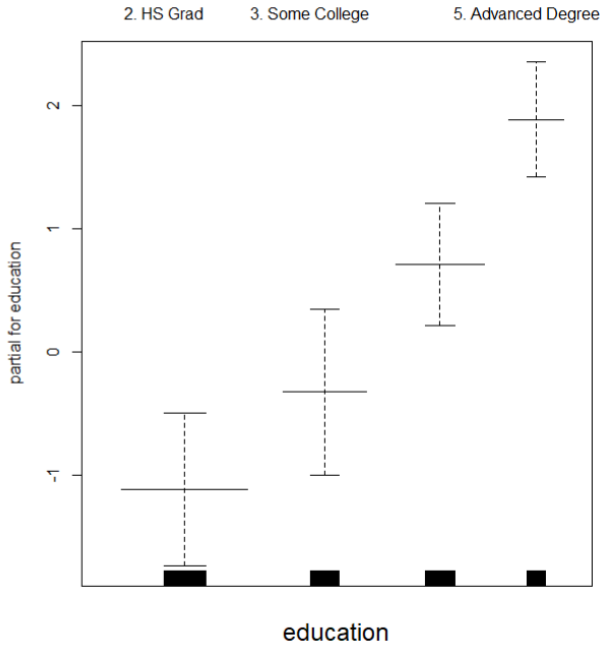
7 广义可加模型

```
> 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

结果表明<HS类没有高收入者





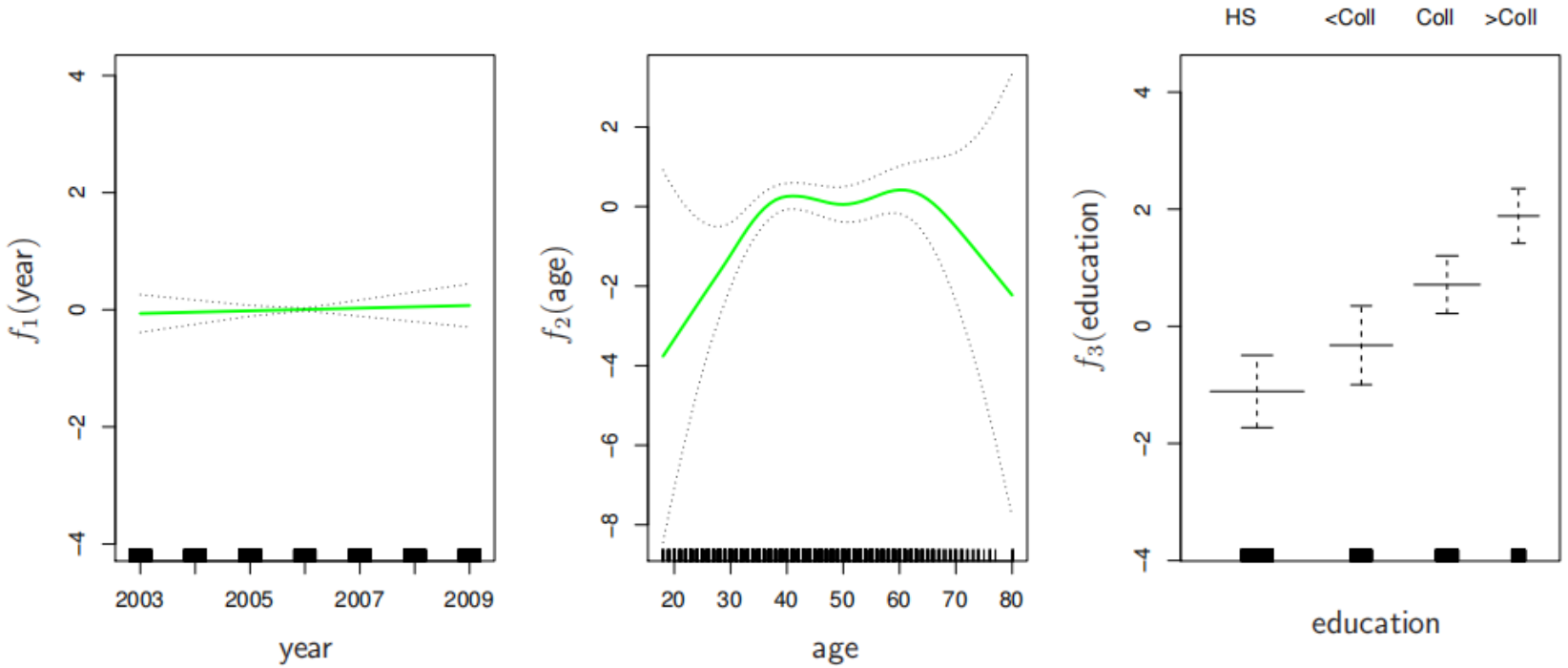
因此，拟合逻辑斯谛回归GAM时需剔除这一类，才可得到有意义的结果。

```
> gam.lr.s=gam(I(wage>250)~year+s(age,df=5)+education,family=binomial,data=Wage,
subset=(education!="1. < HS Grad"))
> plot(gam.lr.s,se=T,col="green")
```

7 广义可加模型

用GAM模型预测Wage数据中个人收入超过250000美元的可能性,最终得到如下图。

$$log(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 \times (\text{year}) + f_2(\text{age}) + f_3(\text{education})$$



线性函数　　自由度为5的光滑样条　　阶梯函数

图12 Wage数据，对二元响应变量I（wage>250）拟合GAM的逻辑斯谛回归。每一张图都绘制了拟合曲线合标准误差。