非参数曲线拟合

本模块调用 R 的 np 程序包的 kernel smoothing 方法，对用户指定的结果变量 Y 与一组自变量进行曲线拟合。详见：1. Tristen Hayfield 和 Jeffrey S. Racine， The np package. 2. Angelo Mazza et al. KernSmoothIRT: An R Package for Kernel Smoothing in Item Response Theory. Journal of Statistical Software June 2014, Volume 58, Issue 6.

自变量可以是多种类型（连续性、分类型）。应变量 Y 可以是连续型或两分类型。模块自动根据数据计算最佳 bandwidths （邻域宽度）。该模块计算时间较长。

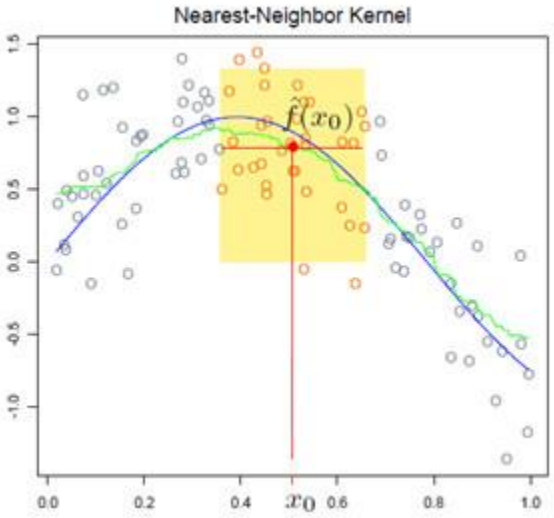
关于 kernel smoothing:

简单起见，我们这里的讨论针对 X 是一维的情况，多维数据的处理是类似的。对于 KNN(K Nearest Neigbor)，我们知道

$$\hat{f}(x) = \mathrm{Ave}(y_i | x_i \in N_k(x))$$

，

下面是一个用 KNN 拟合一条有噪声的曲线的示意图。蓝线表示真正的曲线，绿线表示用 KNN 拟合的效果。



可以看到，这样拟合出来的曲线存在很多问题，比如不连续和不光滑。一种改进就是对于这 K 个近邻，我们给予它们不同的权重，距离越近，权重越大，距离越远，权重越小。其中一种叫做 Nadaraya－Watson kernel-weighted 的方法可表示为:

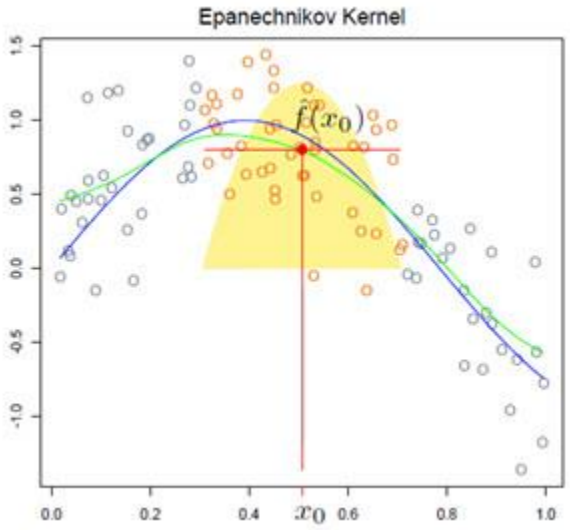
$$\hat{f}(x_0) = \frac{\sum_{i=1}^{N} K_\lambda(x_0, x_i) y_i}{\sum_{i=1}^{N} K_\lambda(x_0, x_i)}$$

其中 K_λ 一种核函数，叫做 Epanechnikov quadratic kernel，可表示为

$$K_\lambda(x_0, x) = D\left(\frac{|x - x_0|}{\lambda}\right)$$

$$D(t) = \begin{cases} \frac{3}{4}(1 - t^2) & \text{if } |t| \le 1; \\ 0 & \text{otherwise.} \end{cases}$$

改进后的示意图如下，可以看到拟合的效果有很大的改进



当然，我们可以使用更加通用的核函数

$$K_\lambda(x_0, x) = D\left(\frac{|x - x_0|}{h_\lambda(x_0)}\right)$$

其中$h_\lambda(x_0)$ 决定了在样本点x_0 处邻域的宽度。在上面的 Epanechnikov quadratic kernel 中，$h_\lambda(x_0) = \lambda$ ，λ 是其辐射的半径，对于高斯核函数来说，λ就是标准差。Smoothing parameter λ 控制着邻域的宽度，λ的值越大，其领域越宽广，所包含的数据越多，这样暗示着 variance 会越小，但是 bias 可能会比较大。可以看到，对于 Epanechnikov quadratic kernel，其 bias 是固定的，但是 variance 跟 local density 呈现出反比例的关系，因为这时跟邻域数目 K 没有多少关系，而跟目标样本点本身周围的密度有关。而对于 KNN 则恰好相反，其 bias 是跟 local density 成反比例的，而 variance 总是不变的。因为 KNN 总是取 K 个紧邻，所起 variance 不受影响，但是如果 local density 比较小，所取的 K 个近邻其实都相隔比较远，那么其 bias 自然就会比较大了。

在自变量 X 是多维的情形时，核函数把x, x_0 改成相应的向量$\mathbf{x}, \mathbf{x_0}$ 即可。但是在多维的情况下会存在一些问题，比如在每个维度上样本间的距离可能相差很大，这时用一个球形的邻

域就不太合适。一种方法是在每个维度上对数据进行标准化，另一种方法是用一个半正定的矩阵 $\mathbf{A}$ 去调整在各个维度上距离的贡献。

$$K_{\lambda,A}(x_0,x)=D\left(\frac{(x-x_0)^T\mathbf{A}(x-x_0)}{\lambda}\right)$$

如果 $\mathbf{A}$ 是对角阵，那么增加或者减少 $\mathbf{A}_{jj}$ 就会增加或减少特征 X_j 的贡献。

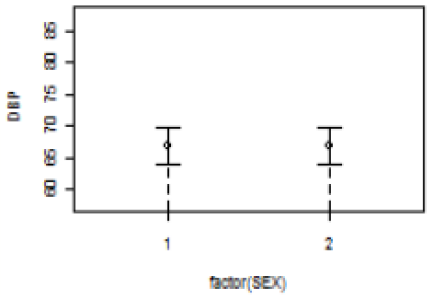
例1. 易侕软件自带的练习项目“demo”，对 DBP 的非参数曲线拟合分析，输入界面如下：

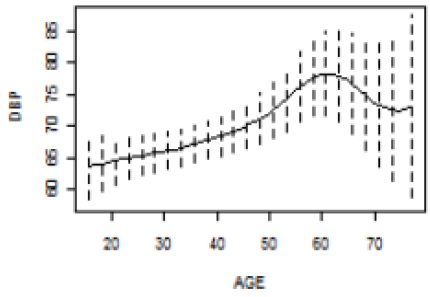
输出结果如下：

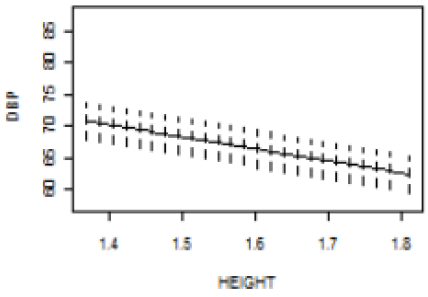
Multivariate nonparametric (Kernel) smoothing regression

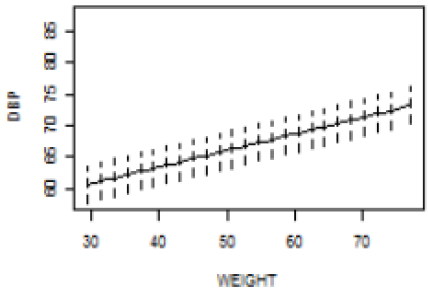
Outcome: Diastolic BP, mmhg
Regression Data: 784 training points, in 8 variable(s)

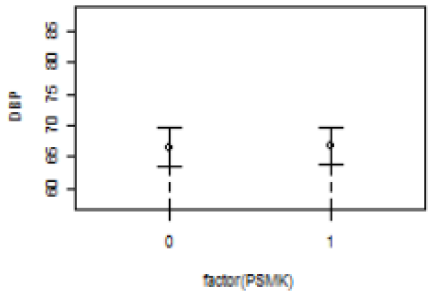
```
No. Complete Observations:  784 No. NA Observations:  48
Observations omitted:  10 56 114 117 152 157 186 187 193 197 198 205 219 247 249 300
394 422 427 461 464 498 511 521 540 543 553 600 604 640 645 650 670 678 712 719 721
732 734 744 754 778 793 806 811 821 823 827
              factor(SEX)      AGE   HEIGHT  WEIGHT factor(PSMK) factor(ALH)
Bandwidth(s):         0.5 5.701958 105128.8 2224151    0.2567232   0.2085370
              factor(SMOKE) factor(EDU.NEW)
Bandwidth(s):     0.2548993       0.1465249

Kernel Regression Estimator: Local-Linear
Bandwidth Type: Fixed
Residual standard error: 86.72913
R-squared: 0.3348089

Continuous Kernel Type: Second-Order Gaussian
No. Continuous Explanatory Vars.: 3

Unordered Categorical Kernel Type: Aitchison and Aitken
No. Unordered Categorical Explanatory Vars.: 5

NULL

Kernel Regression Significance Test
Type I Test with IID Bootstrap (399 replications)
Explanatory variables tested for significance:
factor(SEX) (1), AGE (2), HEIGHT (3), WEIGHT (4), factor(PSMK) (5), factor(ALH) (6),
factor(SMOKE) (7), factor(EDU.NEW) (8)

              factor(SEX)      AGE   HEIGHT  WEIGHT factor(PSMK) factor(ALH)
Bandwidth(s):         0.5 5.701958 105128.8 2224151    0.2567232   0.2085370
              factor(SMOKE) factor(EDU.NEW)
Bandwidth(s):     0.2548993       0.1465249

Significance Tests
P Value:
factor(SEX)     0.5363409
AGE             < 2.22e-16 ***
HEIGHT          0.0050125 **
WEIGHT          0.1428571
factor(PSMK)    0.0651629 .
factor(ALH)     0.0175439 *
factor(SMOKE)   0.0526316 .
factor(EDU.NEW) 0.0150376 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

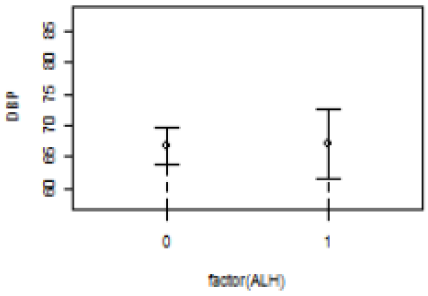

DBP
60 65 70 75 80 85
1
2
factor(SEX)

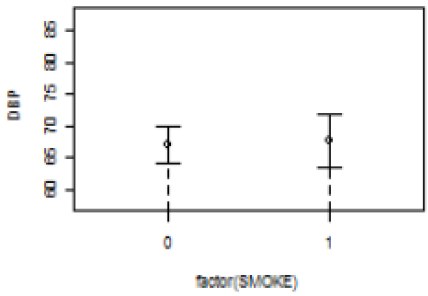

DBP
60 65 70 75 80 85
20 30 40 50 60 70
AGE

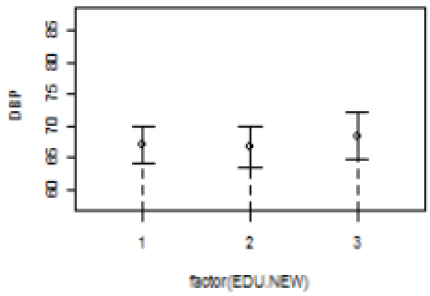

DBP
60 65 70 75 80 85
1.4 1.5 1.6 1.7 1.8
HEIGHT


DBP
60 65 70 75 80 85
30 40 50 60 70
WEIGHT


DBP
60 65 70 75 80 85
0
1
factor(PSMK)


DBP
60 65 70 75 80 85
0
1
factor(ALH)


DBP
60 65 70 75 80 85
0
1
factor(SMOKE)


DBP
60 65 70 75 80 85
1
2
3
factor(EDU.NEW)