S-PLUS

- 因变量取值是分类型的
  - 钢材延展度合格与否
  - 客户流失与否
  - 客户信用等级
- 响应变量的概率的某种变换表示为自变量的线性组合

$$g(\mu)=\beta_0+\sum_{i=1}^{k}\beta_i x_i \qquad (\mu=E[Y\mid x])$$

称函数 $g(.)$ 为联系函数(link fuction)

- Logistic回归是商业上应用最为普遍的一种广义线性模型
- Logistic回归主要针对响应变量具有二项分布的情况
- Logistic回归是当联系函数取logit变换时的广义线性模型

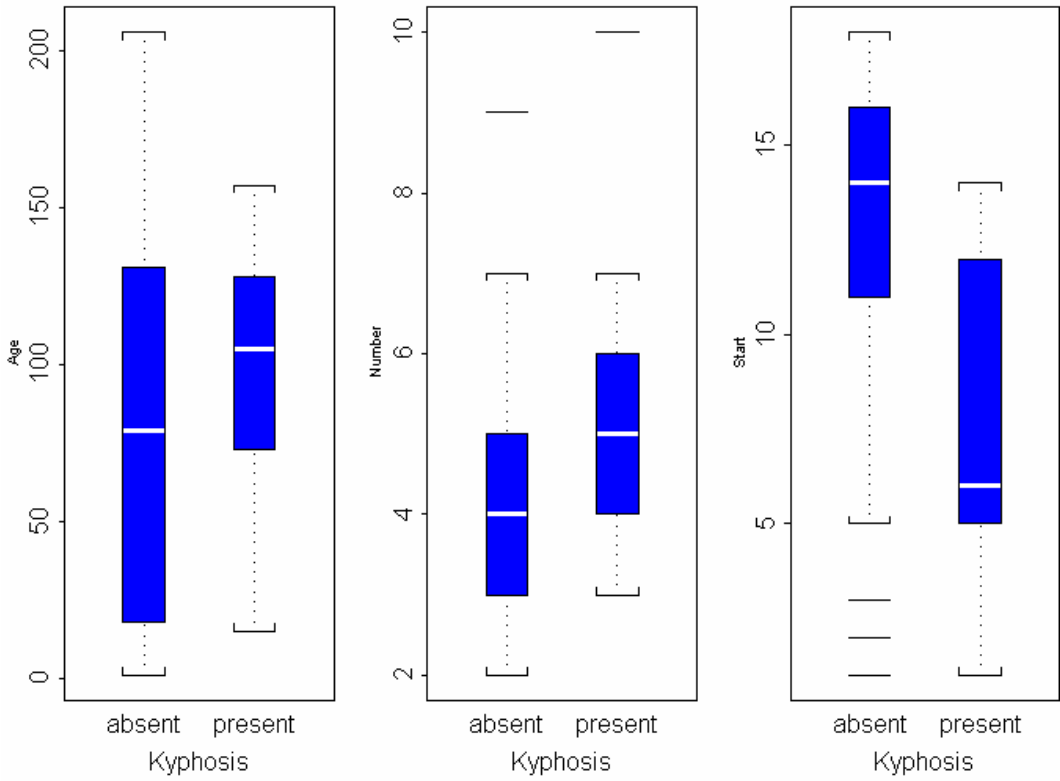
$$\text{logit}(\mu)=\log(\frac{\mu}{1-\mu})$$

案例 内置数据kyphosis，是有关儿童脊椎矫正的临床数据

Kyphosis	手术后是否存在驼背(present表示存在，absent表示治愈)
Age	年龄(出生的月数)
Number	已经手术的椎骨数
Start	手术的椎骨的开始位置

第一步：数据探索

```
> par(mfrow = c(1, 3), cex = 0.7)
> plot.factor(kyphosis, col = "blue")
```



S-PLUS

第二步：模型拟合

```
> kyph.glm.all <- glm(Kyphosis ~ .,  family = binomial(link=logit), data = kyphosis)
> summary(kyphosis.glm.all)
```

Call: glm(formula = Kyphosis ~ Age + Number + Start, family = binomial(link = logit),data = kyphosis)

```
Deviance Residuals:
      Min         1Q     Median         3Q      Max
 -2.312363 -0.5484307 -0.3631874 -0.1658652 2.161331
Coefficients:
                 Value  Std. Error   t value
(Intercept) -2.03693352 1.449574526 -1.405194
        Age  0.01093048 0.006446256  1.695633
     Number  0.41060119 0.224860819  1.826024
      Start -0.20651005 0.067698863 -3.050421
(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 83.23447 on 80 degrees of freedom
Residual Deviance: 61.37993 on 77 degrees of freedom
Number of Fisher Scoring Iterations: 5
Correlation of Coefficients:
       (Intercept)         Age     Number
   Age -0.4635592
Number -0.8481136   0.2323701
 Start -0.3780260  -0.2851666  0.1104551
```

S-PLUS

第二步：模型拟合

模型主要包括以下5个部分：

① 提交的函数调用
② 离差描述
③ 回归参数估计表、它们的标准误以及相应的 t 统计量
④ 零模型估计以及相应离差
⑤ 系数估计的相关阵

从所估计系数的 t 统计量也可以看出Start变量最为重要

S-PLUS

第三步：离差分析

```
> anova(kyphosis.glm.all, test = "Chi")
Analysis of Deviance Table
Binomial model
Response: Kyphosis
Terms added sequentially (first to last)
       Df Deviance Resid. Df Resid. Dev   Pr(Chi)
  NULL                    80   83.23447
   Age  1  1.30198         79   81.93249 0.2538510
Number  1 10.30593         78   71.62656 0.0013260
 Start  1 10.24663         77   61.37993 0.0013693

> anova(glm(formula = Kyphosis ~ Start + Number + Age, family = binomial(link =
    logit), data = kyphosis), test = "Chi")
Analysis of Deviance Table
Binomial model
Response: Kyphosis
Terms added sequentially (first to last)
       Df Deviance Resid. Df Resid. Dev    Pr(Chi)
  NULL                    80   83.23447
 Start  1 15.16229         79   68.07218 0.00009865
Number  1  3.53571         78   64.53647 0.06006055
   Age  1  3.15654         77   61.37993 0.07562326
```

统计分析与建模——广义加性模型(GAM)

S-PLUS

- 线性模型假定响应变量与自变量之间具有线性性质
- 广义线性模型依然假定了预测器具有线性性质
- 广义加性模型拟合非参数函数来估计响应变量与解释变量之间关系的模型
- 广义加性模型的一般形式为:

$$g(\mu) = \beta_0 + \sum_{i=1}^{k} f_i(x_i) \qquad (\mu = E[Y \mid x])$$

称函数 $g(.)$ 为联系函数(link fuction)，β_0 为常数截距项

$f_i(.)$ 是用来描述 $g(\mu)$ 与第 i 个解释变量关系的非参数函数

- S-PLUS用两类方法拟合GAM模型：s—B样条，lo—局部加权回归

局部加权回归：

1. 给定一点 x_0，取 x_0 最近的 k 个点，它们构成 x_0 的一个临域 $N(x_0)$，其中的 k 是由一个叫做带宽(span)的参数确定，在局部加权最小二乘中，参数span是指占总样本点的百分比。
2. 计算临域 $N(x_0)$ 中距 x_0 的最大距离：$\Delta(x_0)=\max_{x\in N(x_0)}|x-x_0|$
3. 使用一个加权函数设定每个点的加权系数，加权函数要满足：1）是对称的、单峰的、以 x_0 为中心的；2）在 x_0 临域 $N(x_0)$ 的边界上为零或接近零。通常使用的加权函数为：

$$W(u)=\begin{cases}(1-u^3)^3 & 0\le u\le 1\\ 0 & u<0\text{或}u>1\end{cases}$$

这样，对 x_0 临域 $N(x_0)$ 中的点 x 的加权系数为：$W_{x_0}(x)=W\left(\frac{|x-x_0|}{\Delta(x_0)}\right)$

4. 在 x_0 临域 $N(x_0)$ 上拟合因变量：

$$\min_{\beta}\sum_{x\in N(x_0)}[y_x-\beta_{x_0}-\beta_{1x_0}x]^2$$

并用拟合的函数进行平滑估计：$y_{x_0}=\hat{f}(x_0)$

统计分析与建模——广义加性模型(GAM)

平滑样条： 平滑样条拟合是一种特殊的局部多项式拟合，样条函数整体要求一阶连续可导、二阶导数可积，并且连接点光滑的特性。它的拟合惩罚函数也与一般回归不同：设一元样条的区间分割点为 $x_1,\cdots,x_m$

$$\min\{\sum [y_i - f(x_i)]^2 + \lambda \int_{\min(x_i)}^{\max(x_i)} [f''(x)]^2 dx$$

其中，λ 粗糙惩罚度，通常也叫平滑参数，它作用类似于loess中的span；如果没有粗糙惩罚项，上式就是工程中通常的样条插值。尽管平滑样条拟合类似于局部多项式回归，但它在数学上有完备的理论基础，尤其是B样条，有很完美的数学性质，同时还有高效的算法。

案例(续) 内置数据kyphosis，是有关儿童脊椎矫正的临床数据

拟合具有平滑估计的加性模型，使用三次B样条平滑：

```
> kyph.gam.all <- gam(Kyphosis ~ s(Age) + s(Number) + s(Start),family = binomial,
  data = kyphosis)
> summary(kyph.gam.all)
Call: gam(formula = Kyphosis ~ s(Age) + s(Number) + s(Start), family = binomial,
                                    data = kyphosis)
Deviance Residuals:
       Min         1Q     Median         3Q      Max
 -1.348291 -0.4438907 -0.1649084 -0.01053841 2.116355

(Dispersion Parameter for Binomial family taken to be 1 )
    Null Deviance: 83.23447 on 80 degrees of freedom
Residual Deviance: 40.88053 on 68.20125 degrees of freedom

Number of Local Scoring Iterations: 8
DF for Terms and Chi-squares for Nonparametric Effects

            Df Npar Df Npar Chisq    P(Chi)
(Intercept)  1
     s(Age)  1     2.9   5.799710 0.1174531
  s(Number)  1     2.9   5.474503 0.1329755
   s(Start)  1     2.9   5.864877 0.1132223
```

统计分析与建模——广义加性模型(GAM)

S-PLUS

✧ 函数gam的输出类似于glm函数，

✧ 一点不同的是离差表的分析：它的主要应用是遴选模型包含的项

✧ 对于单变量模型，等价于检验线性拟合与平滑拟合的差异

```
> kyph.gam.start <- gam(Kyphosis ~ s(Start), family = binomial, data = kyphosis)
> kyph.gam.start.age <- gam(Kyphosis ~ s(Start) + s(Age), family = binomial,
        data = kyphosis)
> kyph.gam.start.number <- gam(Kyphosis ~ s(Start) + s(Number), family =
        binomial, data = kyphosis)
```

进行离差分析：

```
> anova(kyph.gam.start, kyph.gam.start.age, test = "Chi")
Analysis of Deviance Table
Response: Kyphosis
       Terms Resid.  Df Resid.  Dev              Test     Df Deviance  Pr(Chi)
1          s(Start)  76.08701   58.77385
2 s(Start) + s(Age)  72.14926   48.50280 +s(Age) 3.937745 10.27105   0.03454752

> anova(kyph.gam.start, kyph.gam.start.number, test = "Chi")
Analysis of Deviance Table
Response: Kyphosis
             Terms Resid.  Df Resid.  Dev              Test    Df Deviance     Pr(Chi)
1          s(Start)  76.08701   58.77385
2 s(Start) + s(Number)  72.15558   54.12140 +s(Number) 3.931429 4.652447   0.3157556
```

图形分析：使用gam类的plot.gam函数绘制拟合图

(resid=T是画出离差点、scale=8 是为了比较变量的重要性而调整纵轴到同比例尺度):

```
> par(mfrow = c(2, 2))
> plot(kyph.gam.start.age, resid = T, scale = 8, main = "gam.start.age",col = "blue")
> plot(kyph.gam.start.number, resid = T, scale = 8, main = "gam.start.number",col = "blue")
```