大数据挖掘软件系列课之：R语言 — 统计建模(第2部分)
案例:逻辑回归模型

肖 韬

	自变量/特征变量/解释变量		
反应变量/响应变量/因变量		连续型	类别型
	连续型	散点图	箱线图，概率密度图（分类别画）
	类别型	脊髓图/条件密度图	脊柱图

	自变量/特征变量/解释变量		
反应变量/响应变量/因变量		连续型	类别型
	连续型		方差分析
	类别型		列联表分析

	自变量/特征变量/解释变量		
反应变量/响应变量/因变量		连续型	类别型
	连续型	线性回归 （连续型自变量）	线性回归 （类别型自变量）
	类别型	逻辑回归 （连续型自变量）	逻辑回归 （类别型自变量）

	自变量/特征变量/解释变量		
反应变量/响应变量/因变量		连续型	类别型
	连续型	线性回归 （连续型自变量）	线性回归 （类别型自变量）
	类别型	逻辑回归 （连续型自变量）	逻辑回归 （类别型自变量）

逻辑回归模型

对于线性回归模型，我们假设响应变量Y（大约）为正态分布。但是，许多分析任务要求对自变量和二元型(0-1型)的响应变量之间的关系进行评估，而二元型的响应变量Y的分布函数是伯努利分布（Bounulli distribution）：

$$P(Y=1)=\pi, P(Y=0)=1-\pi$$

因此，我们不能对以这种二元型(0-1型)变量为响应变量的数据使用线性回归模型。

逻辑回归模型

多元线性回归模型可以用公式描述为：

$y \sim \mathcal{N}(\mu, \sigma^2)$,其中$\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q$.

这清楚表明此模型适用于响应变量为连续型变量，且在解释变量已知的情况下，这个响应变量符合一个方差恒定的正态分布，这个响应变量的数学期望值是回归系数的线性函数。

因此很明显该模型不适合应用在响应变量是二元变量的情况。

逻辑回归模型

但是，在响应变量为二元变量的情况下，我们仍然希望能将响应变量的期望值表示成回归系数的线性函数。如何做到呢？我们可以对响应变量的数学期望值$\pi = P(y = 1)$做如下转换，我们把这种针对π的转换称为π的逻辑函数（*logistic*函数或*logit*函数）：

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q.$$

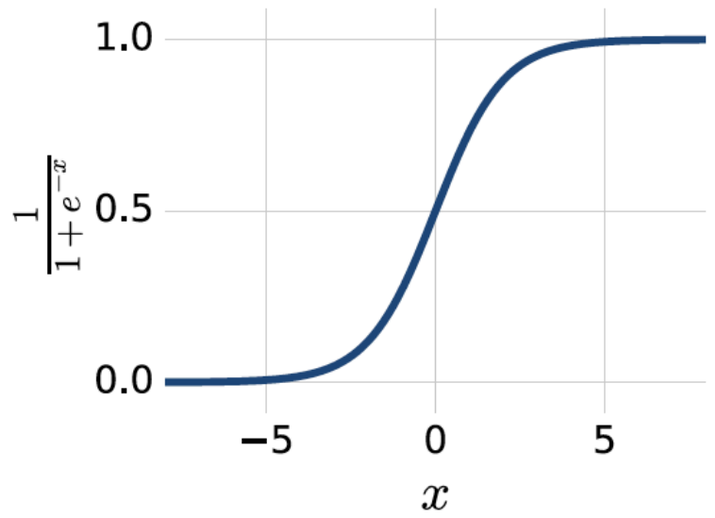
用x表示

"概率$\pi = P(y = 1)$的逻辑函数"就是"Y=1的优势的自然对数"（注："Y=1的优势"的定义为：$\frac{\pi}{1-\pi}$）。

$$\frac{\pi}{1-\pi} = e^x$$

$$\Rightarrow \pi = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_q x_q)}}$$

$$\therefore \hat{\pi} = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x_1+\cdots+\hat{\beta}_q x_q)}}$$



因此，当逻辑回归模型的回归系数$\hat{\beta}_0, \hat{\beta}_1, \cdots \hat{\beta}_q$被估计出来后，我们可以通过上式估计出已知$x_1, \cdots x_q$值的新个体的Y=1的$\pi$(概率)

逻辑回归模型

- 逻辑函数的取值范围可以是任意实数，而这个逻辑函数的输入参数(即概率π)的取值范围总是在$[0,1]$区间中。
- 在一个逻辑回归模型中，与解释变量x_j相对应的参数β_j的意义描述如下：在其他变量的值保持不变的情况下，x_j每增加一个单位，响应变量$Y=1$的优势就会变化为增加前的$\exp(\beta_j)$倍。
- Logistic回归模型的回归系数可由最大似然估计法估计出来。

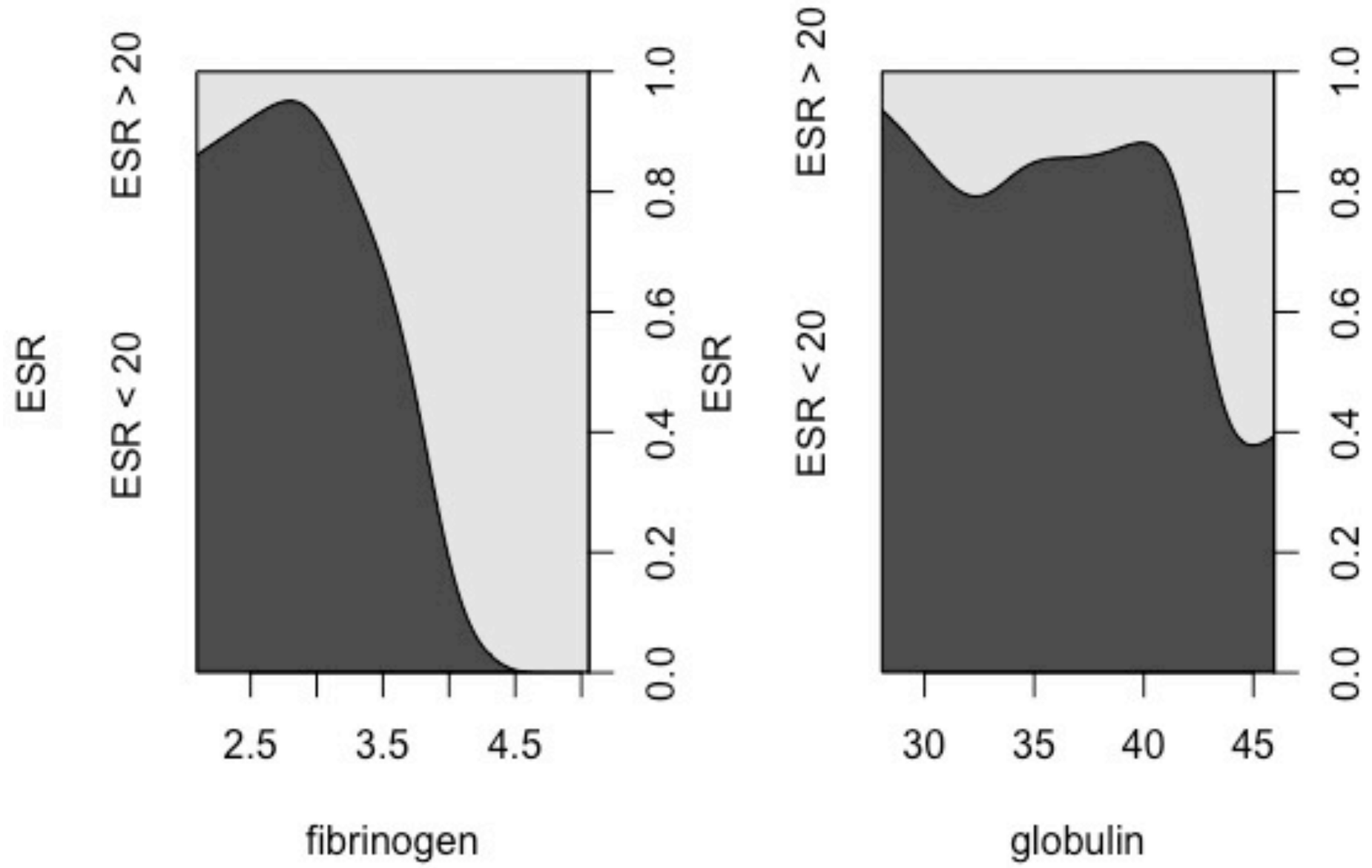
血浆(PLASMA)数据：红细胞沉降率（ESR）

红细胞沉降率（ESR）是指在标准条件下测量血浆时红细胞从血液悬浮液中沉降出来的比例。ESR的绝对数值值不是很重要，但“是否小于20mm / hr”这个阈值却很有意义，如果一个人ESR的值小于20mm / hr则表示这个人是一个“健康”的人，否则，则患风湿性疾病或者其他慢性疾病的风险就会比较高。经过观察，有两种血浆蛋白(纤维蛋白fibrinogen和球蛋白globulin)的含量水平如果升高时，ESR的值好像也会有所升高。

我们感兴趣的问题是：这两种血浆蛋白的水平是否 和“ESR读数大于20mm / hr的概率”之间存在任何关联。如果有关联，那么我们就可以利用相关联的血浆蛋白的水平来辅助对ESR结果的诊断。

血浆(PLASMA)数据：红细胞沉降率（ESR）

```
> load("Datasets/plasma.RData")
> layout(matrix(1:2, ncol = 2))
> cdplot(ESR ~ fibrinogen, data = plasma)
> cdplot(ESR ~ globulin, data = plasma)
```



血浆(PLASMA)数据：红细胞沉降率（ESR）

```
> plasma_fit <- glm(ESR ~ fibrinogen + globulin, data = plasma, family = binomial())
> summary(plasma_fit)

Call:
glm(formula = ESR ~ fibrinogen + globulin, family = binomial(),
    data = plasma)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9683  -0.6122  -0.3458  -0.2116   2.2636

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.7921     5.7963  -2.207   0.0273 *
fibrinogen    1.9104     0.9710   1.967   0.0491 *
globulin      0.1558     0.1195   1.303   0.1925
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 30.885  on 31  degrees of freedom
Residual deviance: 22.971  on 29  degrees of freedom
AIC: 28.971

Number of Fisher Scoring iterations: 5
```

血浆(PLASMA)数据：红细胞沉降率（ESR）

对显著的跟自变量对应的回归系数的解释：

$\hat{\beta}_{fibrinogen} = 1.9104$，Z统计量 = 1.967，P值 = 0.0491 < 0.05，

基于我们的数据，我们在0.05的显著性水平上拒绝 $\beta_{fibrinogen} = 0$ 的原假设。我们推断，在其它变量的值保持不变的情况下，纤维蛋白(fibrinogen)每增加一个单位，"ESR > 20"的优势就会变为原来的 $e^{1.9104} = 6.76$ 倍（或者也可以这样说："ESR > 20"的优势的自然对数会增加1.9104）。

注意：逻辑回归中的回归系数反映的是"优势比的自然对数"。

国产车维修次数数据

	make	foreign	repair
1	AMC Concord	Domestic	2
2	AMC Pacer	Domestic	2
3	Audi 5000	Foreign	3
4	Audi Fox	Foreign	2
5	BMW 320i	Foreign	3
6	Buick Century	Domestic	2
7	Buick Electra	Domestic	3
8	Buick LeSabre	Domestic	2
9	Buick Regal	Domestic	2
10	Buick Riviera	Domestic	2
11	Buick Skylark	Domestic	2
12	Cad. Deville	Domestic	2
13	Cad. Eldorado	Domestic	1
14	Cad. Seville	Domestic	2
15	Chev. Chevette	Domestic	2
16	Chev. Impala	Domestic	3
17	Chev. Malibu	Domestic	2
18	Chev. Monte Carlo	Domestic	1

Showing 1 to 19 of 59 entries

国产车维修次数数据

```
> repairdata = read.csv("Datasets/repair.csv")
> table(repairdata$foreign,repairdata$repair)

                1  2  3
Domestic 10 27  9
Foreign   1  3  9
```

国产车维修次数数据

```
> fit_repairdata=glm(foreign~as.factor(repair),data=repairdata,family=binomial)
> summary(fit_repairdata)

Call:
glm(formula = foreign ~ as.factor(repair), family = binomial,
    data = repairdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1774  -0.4590  -0.4590  -0.4366   2.1899

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -2.3026     1.0486  -2.196   0.0281 *
as.factor(repair)2    0.1054     1.2124   0.087   0.9307
as.factor(repair)3    2.3026     1.1497   2.003   0.0452 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 62.226  on 58  degrees of freedom
Residual deviance: 51.160  on 56  degrees of freedom
AIC: 57.16

Number of Fisher Scoring iterations: 4
```

国产车维修次数数据

对显著的跟自变量对应的回归系数的解释：

$\hat{\beta}_{repair.3} = 2.3026$，$e^{2.3026} = 10$，$Z = 2.003$，P值 $= 0.0452 < 0.05$

基于我们的数据，我们在0.05的显著性水平上拒绝 $\beta_{repair.3} = 0$ 的原假设。

我们推断，修过3次的车为外国车（Foreign）的优势要比只修过1次的车要高大约10倍（或者也可以这样说：修过3次的车为外国车的优势的自然对数比只修过一次的车为外国车的优势的自然对数要高大约2.3）.

注意：对于 $\hat{\beta}_{repair.2}$ 来说，P值 $= 0.9307 > 0.05$，因此我们不能推翻 $\beta_{repair.2} = 0$ 的原假设，即我们不能说修过两次的车和只修过一次的车在"是外国车"的优势上有显著差别。

国产车维修次数数据：跟列联表分析的比较

```
> (table1 = cbind(c(10,1),c(27,3)))
     [,1] [,2]
[1,]   10   27
[2,]    1    3
> chisq.test(table1)

        Pearson's Chi-squared test with Yates' continuity correction

data:  table1
X-squared = 1.1212e-30, df = 1, p-value = 1

Warning message:
In chisq.test(table1) : Chi-squared approximation may be incorrect
> (table2 = cbind(c(10,1),c(9,9)))
     [,1] [,2]
[1,]   10    9
[2,]    1    9
> chisq.test(table2)

        Pearson's Chi-squared test with Yates' continuity correction

data:  table2
X-squared = 3.409, df = 1, p-value = 0.06484

Warning message:
In chisq.test(table2) : Chi-squared approximation may be incorrect
```