第九章 广义线性回归模型

Wang Shujia

Contents

1 广义线性回归模型（GLM）

1.1 GLM 基本概念

例 1：飞机损伤数据

Montgomery et al. (2006) 分析了越战时 30 次空中打击中，飞机所受到的损伤数据。研究目的是飞机损伤情况受哪些因素的影响。

因变量 Damage，每次飞机返航检查发现的损伤数

Type 飞机型号，A4：$Type = 0$，A6：$Type = 1$

Bombload 飞机装载炸弹重量（单位：吨）

Airexp 机组人员总飞行经验（单位：月）

建立线性回归模型：

$$Damage_i = \beta_0 + \beta_1 Type_i + \beta_2 Bombload_i + \beta_3 Airexp_i + \varepsilon_i$$

线性回归模型：lm

```
>  damage = c(0, 1, 0, 0, 0, 0, 1, 0, 0, 2, 1, 1, 1, 1, 2, 3,
+             1, 1, 1, 2, 0, 1, 1, 2, 5, 1, 1, 5, 5, 7 )
>  type = c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
+             1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1 )
>  bombload = c(4, 4, 4, 5, 5, 5, 6, 6, 6, 7, 7, 7, 8, 8, 8,
+       7, 7, 7, 10, 10, 10, 12, 12, 12, 8, 8, 8, 14, 14, 14 )
>  airexp = c(91.5, 84, 76.5, 69, 61.5, 80, 72.5, 65, 57.5,
+       50, 103, 95.5, 88, 80.5, 73, 116.1, 100.6, 85, 69.4,
+       53.9, 112.3, 96.7, 81.1, 65.6, 50, 120, 104.4, 88.9,
+       73.7, 57.8 )
>  out_lm <- lm(damage ~ type + bombload + airexp)
```

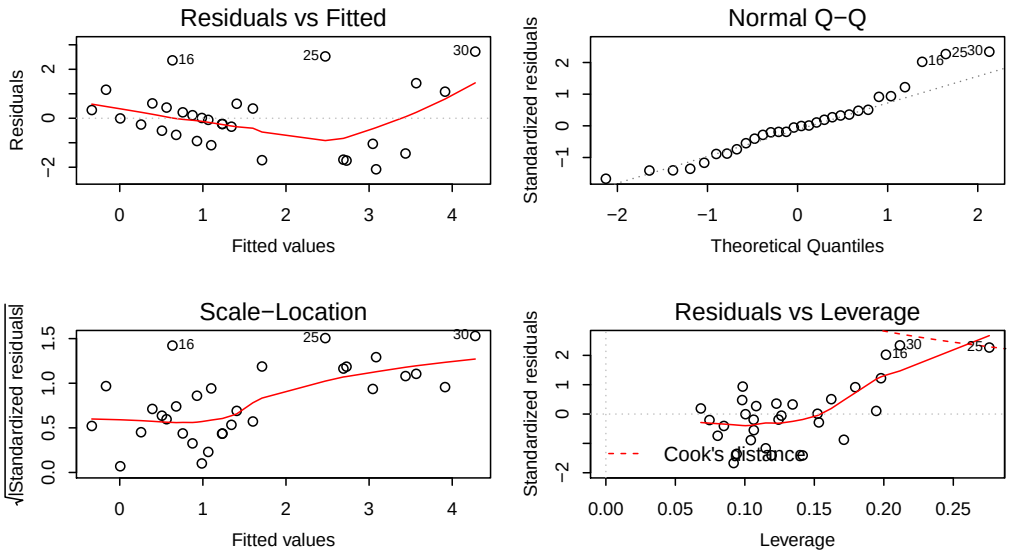
模型结果：lm

```
> summary(out_lm)

##
## Call:
## lm(formula = damage ~ type + bombload + airexp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.08436 -0.86655 -0.03542  0.55377  2.72315
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.43090    1.44788   0.298   0.7684
## type         0.54075    0.72795   0.743   0.4642
## bombload     0.33032    0.12042   2.743   0.0109 *
```

```
## airexp      -0.02283    0.01337  -1.707   0.0997 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.31 on 26 degrees of freedom
## Multiple R-squared:  0.5122,Adjusted R-squared:  0.4559
## F-statistic:   9.1 on 3 and 26 DF,  p-value: 0.0002739
```

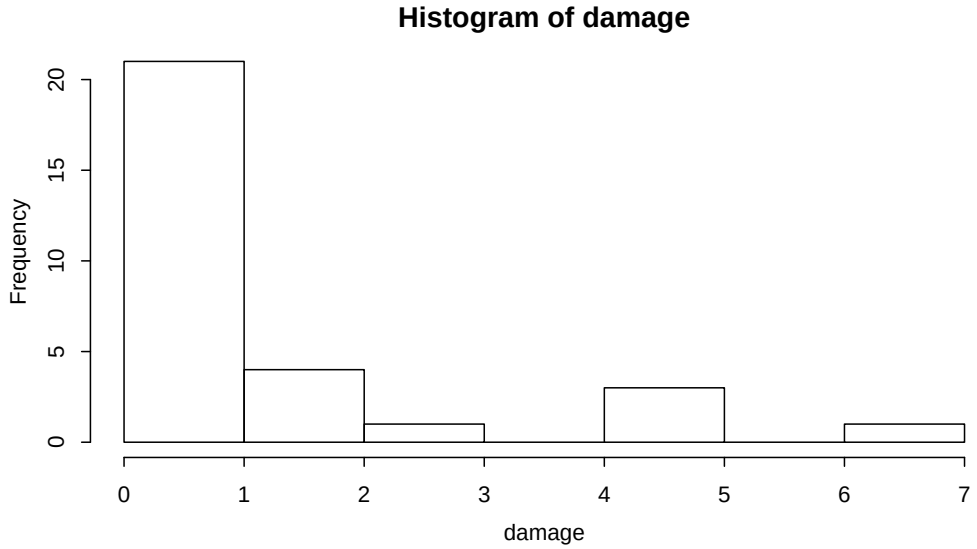
模型诊断：lm

```
>  opar <- par(no.readonly = TRUE); par(mfrow = c(2,2))
>  plot(out_lm); par(opar)
```



因变量是什么分布？

```
> hist(damage)
```



回顾：多元线性回归模型

模型：

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i (i = 1, 2, \ldots, n)$$

$$\mu_i = \mathrm{E}(y_i | \boldsymbol{x}) = \boldsymbol{x}_i^T \boldsymbol{\beta},\ \varepsilon_i \sim N(0, \sigma^2)$$

因变量 y_i 为连续型，服从正态分布。**在如下情形，不能直接建立线性回归模型：**

- 因变量是离散分布
  - 0-1 变量 (Binary)
  - 定序变量 (Ordinal)
  - 计数数据 (Count data)
- 因变量是有界变量
- 因变量与自变量不是线性关系
- 误差项不服从正态分布

泊松回归模型的建立

因变量 $Damage_i \sim \mathrm{Poisson}(\lambda_i)$ $(i = 1, 2, \ldots, n = 30)$，其分布被均值决定。我们想要研究 Damage 的均值如何受解释变量的影响。

但是泊松分布的均值必须大于 0，线性模型

$$\lambda_i = \beta_0 + \beta_1 Type_i + \beta_2 Bombload_i + \beta_3 Airexp_i$$

右边对某些参数可能为负值。因此可先做变换：$\eta_i = \log(\lambda_i)$（称为连接函数），然后建立模型（广义线性模型）：

$$\log(\lambda_i) = \beta_0 + \beta_1 Type_i + \beta_2 Bombload_i + \beta_3 Airexp_i$$

广义线性回归模型的定义

设因变量 $y_1, y_2, \ldots, y_n$ 为来自指数分布族的独立样本，我们要探讨其均值 $\mathrm{E}(y_i|\boldsymbol{x}) = \mu_i$ 与解释变量之间的线性关系。

广义线性回归模型 (Generalized Linear Models，GLM)：

分布 (Distribution) 因变量 y 的分布。通常为**指数分布族**，如正态分布，对数正态分布，泊松分布，二项分布，伽玛分布，超几何分布等

连接函数 (Link function) 因变量期望值 (μ) 的函数：$\eta = g(\mu)$

线性预测因子 (Linear predictor) 对 η 建立线性模型

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

离散参数 (Dispersion parameter,ϕ) 描述分布离散程度的一个参数，因变量 y 的方差表示为 ϕ 与均值 μ 的函数的乘积

$$\mathrm{D}(y) = \phi h(\mu)$$

常用 GLM

分布	$\eta = g(\mu)$	$\mu = g^{-1}(\eta)$
Poisson	$\log \mu$	$\exp(\eta)$
Normal	μ	η
Gamma	$-\mu^{-1}$	$-\eta^{-1}$
Negative binomial	$\log(1-\mu)$	$1-\exp(\eta)$
Binomial logit	$\log\left(\frac{\mu}{1-\mu}\right)$	$\frac{\exp(\eta)}{1+\exp(\eta)}$
Binomial probit	$\Phi^{-1}(\mu)$	$\Phi(\eta)$
Complementary log-log	$\log(-\log(1-\mu))$	$1-\exp(-\exp(\eta))$

GLM 特例：正态线性回归模型

1. 分布 (Distribution)：$y \sim \mathrm{N}(\mu, \sigma^2)$
2. 连接函数 (Link function)：$\eta = \mu$
3. 线性预测因子 (Linear predictor)：$\eta = \boldsymbol{X}^T\boldsymbol{\beta}$
4. 离散参数 (Dispersion parameter,ϕ)：$\phi = \sigma^2, h(\mu) = 1$

R 函数:glm()

```
glm(formula, family = gaussian, data, weights, subset,
    na.action, ...)

family(object, ...):

binomial(link = "logit")
gaussian(link = "identity")
Gamma(link = "inverse")
inverse.gaussian(link = "1/mu^2")
poisson(link = "log")
quasi(link = "identity", variance = "constant")
quasibinomial(link = "logit")
quasipoisson(link = "log")
```

1.2 Binary 因变量：Logit 和 Probit 模型

Logit 和 Probit 模型

假设因变量只取 0 和 1 两个值，则

$$E(y|x) = 0 \times P(y = 0|x) + 1 \times P(y = 1|x) = P(y = 1|x) = p$$

1. **线性概率模型** (Linear Probabilty Model, LPM)：

$$P(y = 1|x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

2. **Logit 模型** (也叫 Logistic 回归模型)：

$$\text{logit}(p) = \log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

```
glm: family=bionomial(link=logit)
```

3. **Probit 模型**：

$$\Phi^{-1}(p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

```
glm: family=bionomial(link=probit)
```

例 2：“挑战者”号太空火箭

1986 年 1 月 28 日，佛罗里达州的肯尼迪航天中心上空万里无云，美国航空航天局 (NASA) 正准备发射“挑战者”号航天飞机。在离发射现场 6.4 公里的看台上，聚集了 1000 多名观众，其中有 19 名中学生代表，他们既是来观看航天飞机发射的，又是来欢送他们心爱的老师 Christa McAuliffe。McAuliffe 是新罕布什尔州康科德中学的老师，美国航空航天局从 11000 多名教师中把她精心挑选出来，计划让她在太空为全国中小学生讲授两节有关太空和飞行的科普课，学生们还可以通过专线向她提问。当孩子们看到航天飞机载着他们的老师升空的壮观场面时，激动得又是吹喇叭，又是敲鼓。然而，航天飞机升空 73 秒后，一个油箱发生爆炸，机上 7 名宇航员全部丧生。

根据调查这一事故的总统委员会的报告，爆炸是一个小小的 O 型封环 (O-ring) 失效所致。这个封环位于右侧固体火箭推进器的两个低层部件之间, 失效的封环使炽热的气体点燃了外部燃料罐中的燃料。O 型封环会在低温下失效，天气预报当地气温在发射时是 31 法式度，在发射前夕有个叫 Rockwell 的工程师警告不要在冷天发射，但是由于发射已被推迟了 5 次，而且因为有教师 McAuliffe 在飞船上，公众特别关注，所以未能引起足够重视，最终导致惨剧发生。

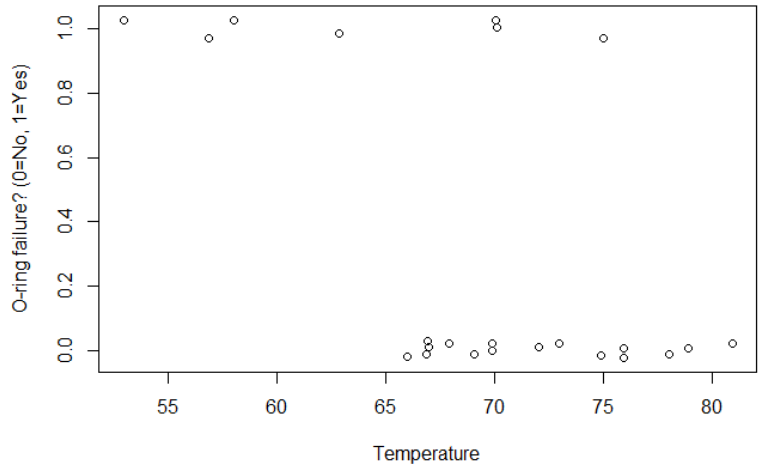
数据

$Y_i = 1$ 表示至少有一个 O-环部件失效，否则 $Y_i = 0$（第 4 次火箭坠入大海）

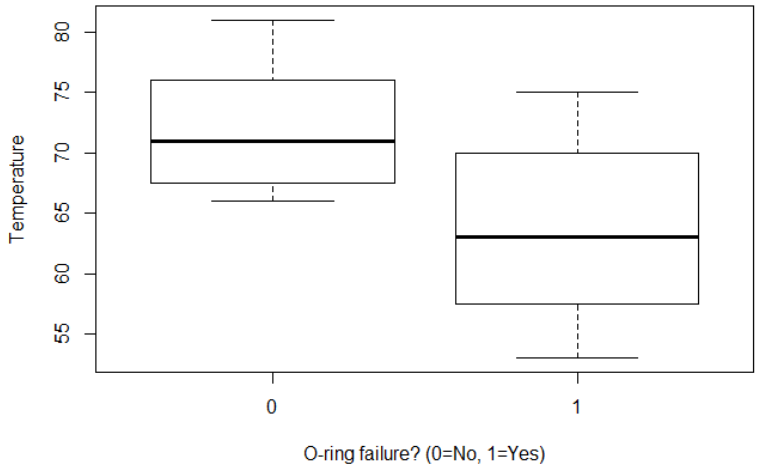
表 7-22 美国 24 次发射航天飞机的 O-环失效情况

发射编号	发射日期	温度(Temp,℉)	O-环(y)	气压(Press)
1	04/12/81	66	0	50
2	11/12/81	70	1	50
3	03/22/82	69	0	50
4	06/27/82	80	*	*
5	11/11/82	68	0	50
6	04/04/83	67	0	50
7	06/18/83	72	0	50
8	08/30/83	73	0	100
9	11/28/83	70	0	100
10	02/03/84	57	1	200
11	04/06/84	63	1	200
12	08/30/84	70	1	200
13	10/05/84	78	0	200
14	11/08/84	67	0	200
15	01/24/85	53	1	200
16	04/12/85	67	0	200
17	04/29/85	75	0	200
18	06/17/85	70	0	200
19	07/29/85	81	0	200
20	08/27/85	76	0	200
21	10/03/85	79	0	200
22	10/30/85	75	1	200
23	11/26/85	76	0	200
24	01/12/86	58	1	200

数据散点图



数据箱线图



Logistic 回归模型

目的：研究 O 环失效概率与发射温度之间的关系

模型：$Y_i \sim \text{Bernoulli}(p_i)$，每次发射和运行中，至少有一个 O-环部件失效的概率 (p_i) 与发射时的温度有关。

连接函数 (Link function)：logit 函数

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

线性预测因子 (Linear predictor)：$\eta = \beta_0 + \beta_1 T$

离散参数 (Dispersion parameter,ϕ)：$\phi = 1$

模型参数含义

$Y = 1$ 表示“成功”，$Y = 0$ 表示“失败”，

1. 优势 (odds)：$odds = p/q$ 表示“成功”的概率与“失败”相比的优势
2. logit 变换：优势的对数，$\text{logit}(p) = \log(p/q)$
3. Logistic 方程：$\text{logit}(p) = \beta_0 + \beta_1 T$,

$$p = \frac{e^{\beta_0+\beta_1 T}}{1+e^{\beta_0+\beta_1 T}}$$

4. 优势比（odds ratio，OR）：自变量 T 增加 1 个单位，对应的优势之比。

$$\beta_1 = \log(odds|T+1) - \log(odds|T) = \log(OR)$$

5. 回归系数与优势比的关系：β_1= 优势比的对数，或：$OR = e^{\beta_1}$

运用 R 估计 Logistic 模型

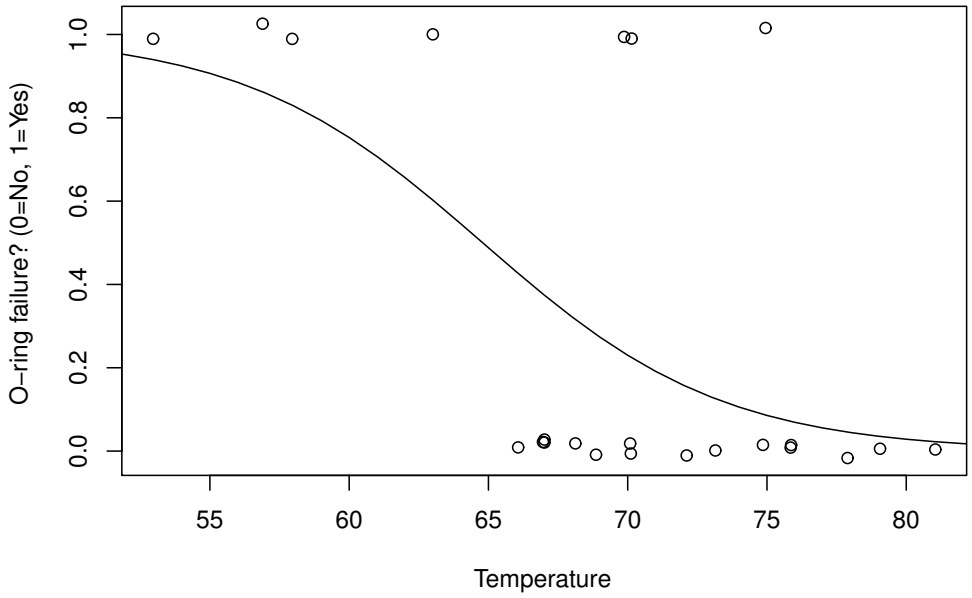
```
data<-list(T=c(53,57,58,63,66,67,67,67,68,69,70,70,70,
70,72,73,75,75,76,76,78,79,81),
y= c(1,1,1,1,0,0,0,0,0,0,0,0,1,1,0,0,0,1,0,0,0,0,0),n=23)
y<-data$y
T<-data$T
oring.glm<-glm(y~T,binomial(link='logit'))
summary.glm(oring.glm)
```

Logistic 模型结果

```
Call:
glm(formula = y ~ T, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0611  -0.7613  -0.3783   0.4524   2.2175

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  15.0429     7.3786   2.039   0.0415 *
T            -0.2322     0.1082  -2.145   0.0320 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.267  on 22  degrees of freedom
Residual deviance: 20.315  on 21  degrees of freedom
AIC: 24.315

Number of Fisher Scoring iterations: 5
```

Logistic 拟合曲线

曲线为 O 环失效概率随温度 T 变化的拟合曲线



预测

```
> times <- c(31,55,75)
> p <- 1/(1 + exp(-1 * (oring.glm$coeff[1] +
          oring.glm$coeff[2] * times)))
> p
[1] 0.99960878 0.90669655 0.08554356
```

1.3 计数数据：泊松回归

泊松回归

因变量是计数数据，用泊松回归模型。

1. 分布 (Distribution)：$y \sim \text{Poisson}(\lambda)$
2. 连接函数 (Link function)：$\eta = \log(\lambda)$
3. 线性预测因子 (Linear predictor)：$\eta = \boldsymbol{X}^T\boldsymbol{\beta}$
4. 离散参数 (Dispersion parameter,ϕ)：$\phi = 1,\ h(\lambda) = \lambda$。

```
glm: family=poisson or
     family=quasipoisson（估计标准误更稳健）
```

泊松回归可能存在两个问题：

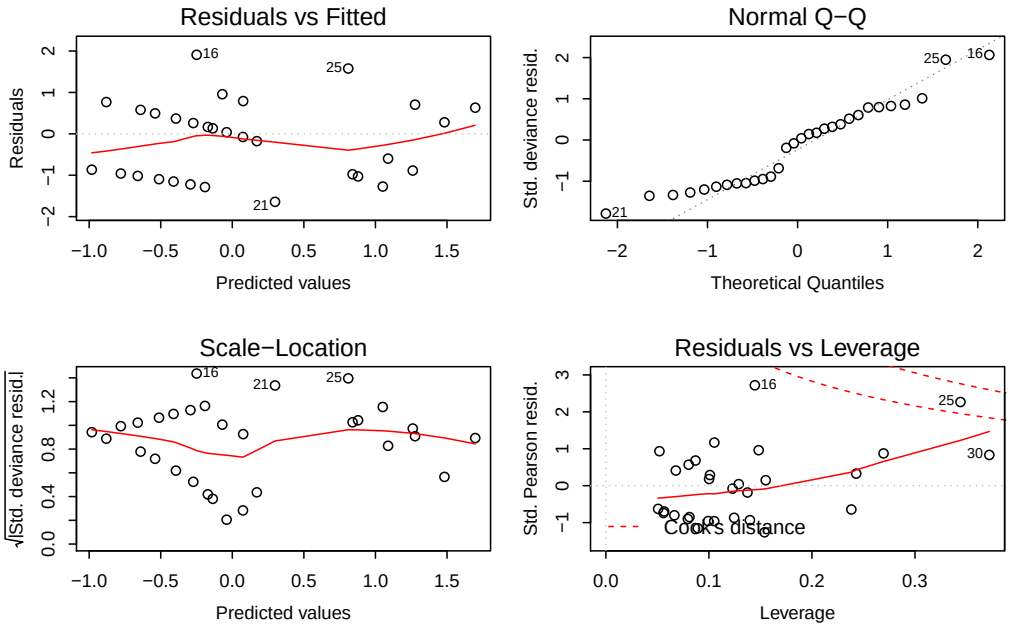
1. 如果 ϕ 大于 1，表明方差大于期望值（泊松分布的方差等于期望值），称为**过度扩散** (Overdispersion);
2. 如果样本观察值为 0 的比重过大（超过泊松分布中 0 的概率），则称为**零膨胀** (Zero-inflated) 现象。

泊松回归模型：glm

```
>  out_glm <- glm(damage~type+bombload+airexp,family=poisson)
>  summary(out_glm)

##
## Call:
## glm(formula = damage ~ type + bombload + airexp, family = poisson)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6418  -1.0064  -0.0180   0.5581   1.9094
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.406023   0.877489  -0.463   0.6436
## type         0.568772   0.504372   1.128   0.2595
## bombload     0.165425   0.067541   2.449   0.0143 *
## airexp      -0.013522   0.008281  -1.633   0.1025
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 53.883  on 29  degrees of freedom
## Residual deviance: 25.953  on 26  degrees of freedom
## AIC: 87.649
##
## Number of Fisher Scoring iterations: 5
```

模型诊断图

```
>  opar <- par(no.readonly = TRUE); par(mfrow = c(2,2))
>  plot(out_glm); par(opar)
```



2 贝叶斯广义线性回归模型

2.1 贝叶斯泊松回归

2.1.1 模型建立及 WinBUGS 计算

例 1：飞机损伤数据

Montgomery et al. (2006) 分析了越战时 30 次空中打击中，飞机所受到的损伤数据，研究飞机损伤数的影响因素。

因变量 Damage，每次飞机返航检查发现的损伤数

Type 飞机型号，A4：$Type = 0$，A6：$Type = 1$

Bombload 飞机装载炸弹重量（单位：吨）

Airexp 机组人员总飞行经验（单位：月）

建立泊松回归贝叶斯模型

因变量 $Damage_i \sim \text{Poisson}(\lambda_i)$ $(i = 1, 2, \ldots, n = 30)$，对 Damage 均值的对数建立线性模型：

$$\log(\lambda_i) = \beta_0 + \beta_1 Type_i + \beta_2 Bombload_i + \beta_3 Airexp_i$$

无信息先验分布：

$$\beta_i \sim dflat(), i = 0, 1, 2, 3$$

系数 β_1 的含义：$\beta_1 = \log(\lambda|Type = 1) - \log(\lambda|Type = 0) = \log(\lambda_1/\lambda_0)$

$$\frac{\lambda_1 - \lambda_0}{\lambda_0} \times 100\% = (e^{\beta_1} - 1) \times 100\%$$

记 $B_k = e^{\beta_k}$，解释：在其它变量保持相同的条件下，自变量 k 增加 1 个单位，因变量的均值增加（减少）$100(B_k - 1)\%$。

运用 R2WinBUGS

1. 写出模型代码文件 (.txt 文件）
2. 准备数据（list 数据结构）
3. 准备初始值（list 数据结构）
4. 指定参数
5. 调用 R2WinBUGS，运行 bugs(包括数据、初始值、参数，迭代次数等等）
6. 有效性和收敛性判断

WinBUGS 模型代码

```
model{
# Poisson model likelihood
for (i in 1:30){
  damage[i] ~ dpois( lambda[i] )
  log(lambda[i])<-beta[1]+beta[2]*type[i]+beta[3]*bombload[i]
                 +beta[4]*airexp[i]
  }
# prior
for (j in 1:4){
  beta[j]~dnorm( 0.0, 0.001 )
  B[j] <- exp( beta[j] )
  }
}
```

准备数据（list 结构）

```
craftdata<-list(
damage=c(0, 1, 0, 0, 0, 0, 1, 0, 0, 2, 1, 1, 1,
          1, 2, 3, 1, 1, 1, 2, 0, 1, 1, 2, 5, 1, 1, 5, 5, 7),
type=c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1,
         1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1),
bombload = c(4, 4, 4, 5, 5, 5, 6, 6, 6, 7, 7, 7, 8, 8, 8, 7, 7,
         7, 10, 10, 10, 12, 12, 12, 8, 8, 8, 14, 14, 14),
airexp = c(91.5, 84, 76.5, 69, 61.5, 80, 72.5, 65, 57.5, 50,
        103,95.5, 88, 80.5, 73, 116.1, 100.6, 85, 69.4, 53.9,
      112.3,96.7, 81.1, 65.6, 50, 120, 104.4, 88.9, 73.7, 57.8)
    )
```

准备初始值和指定参数

```
#Prepare initials
set.seed(1234)
inits <- function(){
list(beta=rnorm(4,0,5))
}
#Specify parameters
parameters <- list("beta","B")
```

运行 R2WinBUGS

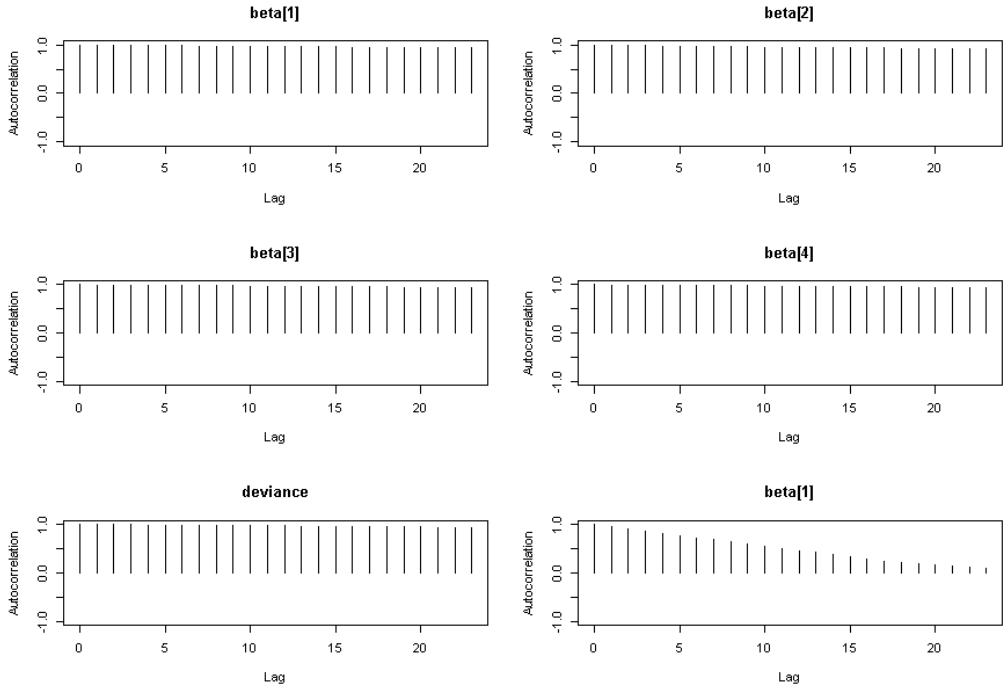
```
output<-bugs(
  craftdata,
  inits,
  parameters,
  n.chains=3,
  n.iter=2000,
  n.burnin=1000,
  n.thin=1,
  debug=FALSE,
  codaPkg=FALSE,
  model.file="F:\\BaiduYun\\Teaching\\Rdata\\Ch6craftmodel.txt",
  bugs.directory="D:\\WinBUGS\\",
  working.directory="F:\\Simulation\\"
  )
```

运行结果：结果是否可靠？

```
> print(output,digit=3)
Inference for Bugs model at "F:\BaiduYun\Teaching\Rdata\Ch6craftmodel.txt", fit using WinBUGS,
 3 chains, each with 2000 iterations (first 1000 discarded)
 n.sims = 3000 iterations saved
             mean       sd   2.5%    25%    50%    75%    97.5%  Rhat n.eff
beta[1]    18.717   45.844 -2.089 -0.927 -0.179  0.781  169.805 1.768     7
beta[2]    -0.414    2.057 -5.988 -0.172  0.361  0.759    1.397 1.885     6
beta[3]     0.811    1.487  0.071  0.144  0.194  0.280    5.569 1.840     6
beta[4]    -0.477    1.108 -4.118 -0.027 -0.016 -0.010    0.001 1.764     7
deviance 1009.021 2229.497 80.210 81.880 83.790 87.990 8369.225 2.033     5

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = Dbar-Dhat)
pD = 69.8 and DIC = 1078.8
DIC is an estimate of expected predictive error (lower deviance is better).
```

1. 参数估计？
2. 变量是否显著？
3. 模型拟合好坏程度？

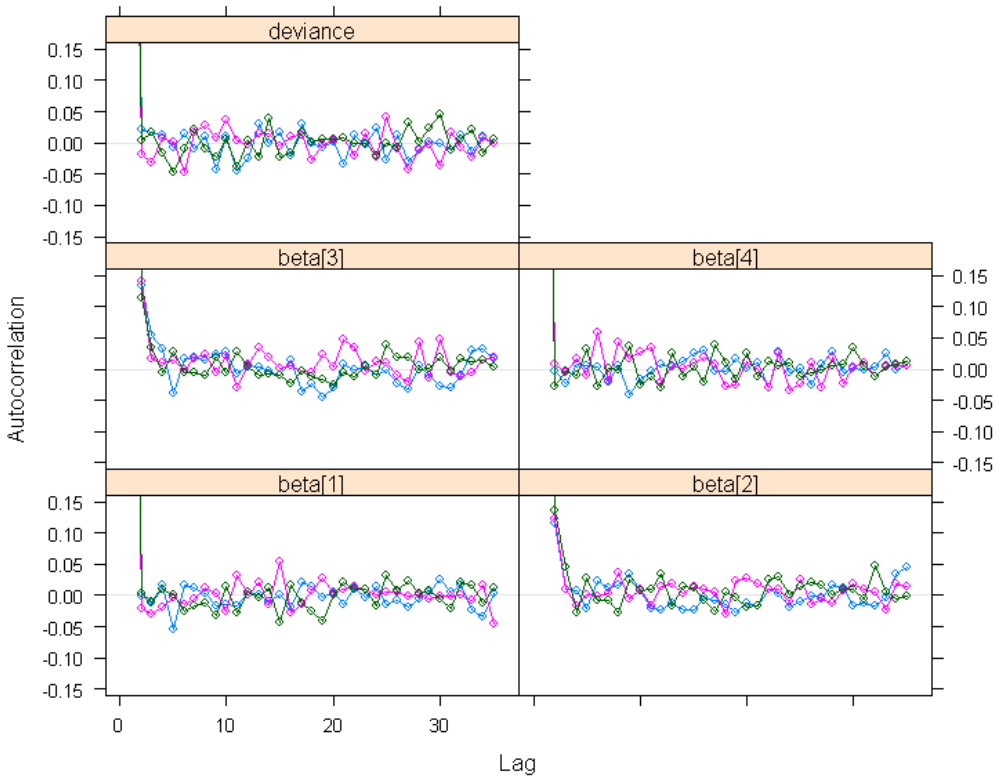
自相关：autocorr.plot()



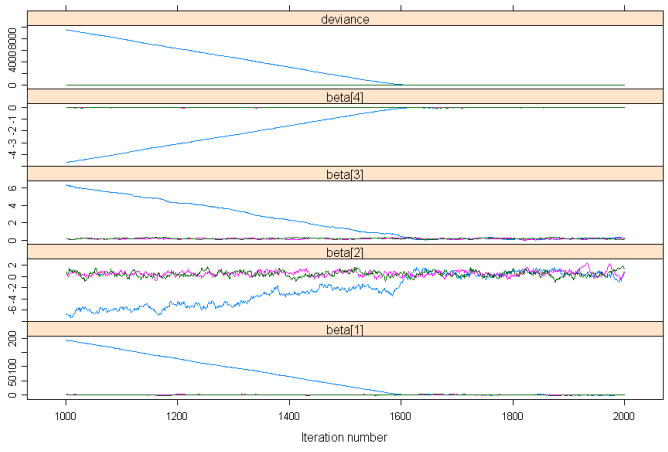
自相关：**autocorr.diag()**

```
> autocorr.diag(A)
          beta[1]   beta[2]   beta[3]   beta[4]  deviance
Lag 0   1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
Lag 1   0.9898863 0.9708728 0.9845537 0.9885256 0.9800317
Lag 5   0.9503494 0.8758036 0.9268341 0.9411988 0.9257937
Lag 10  0.9008789 0.7851445 0.8569282 0.8906172 0.8658388
Lag 50  0.6097393 0.6306516 0.6285145 0.6236613 0.5684869
```

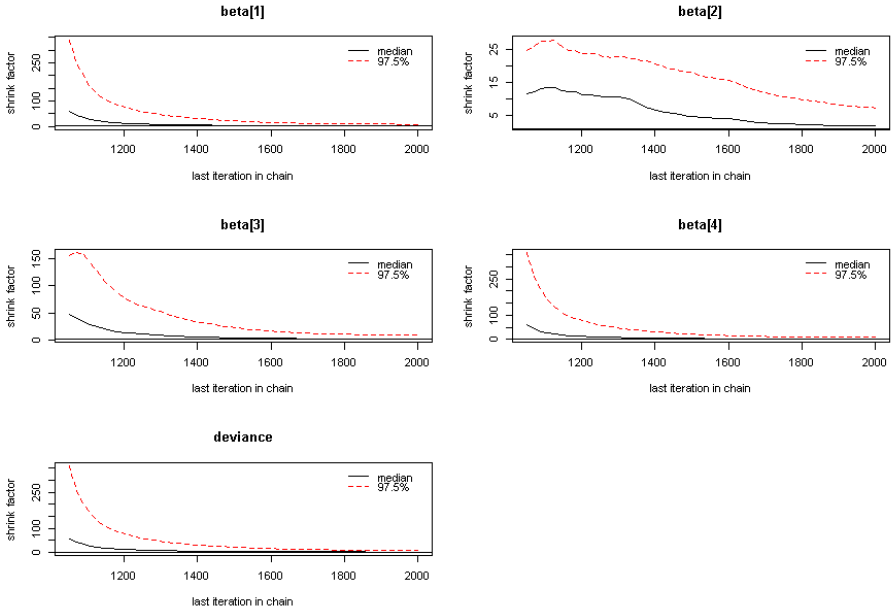
自相关：**acfplot()**



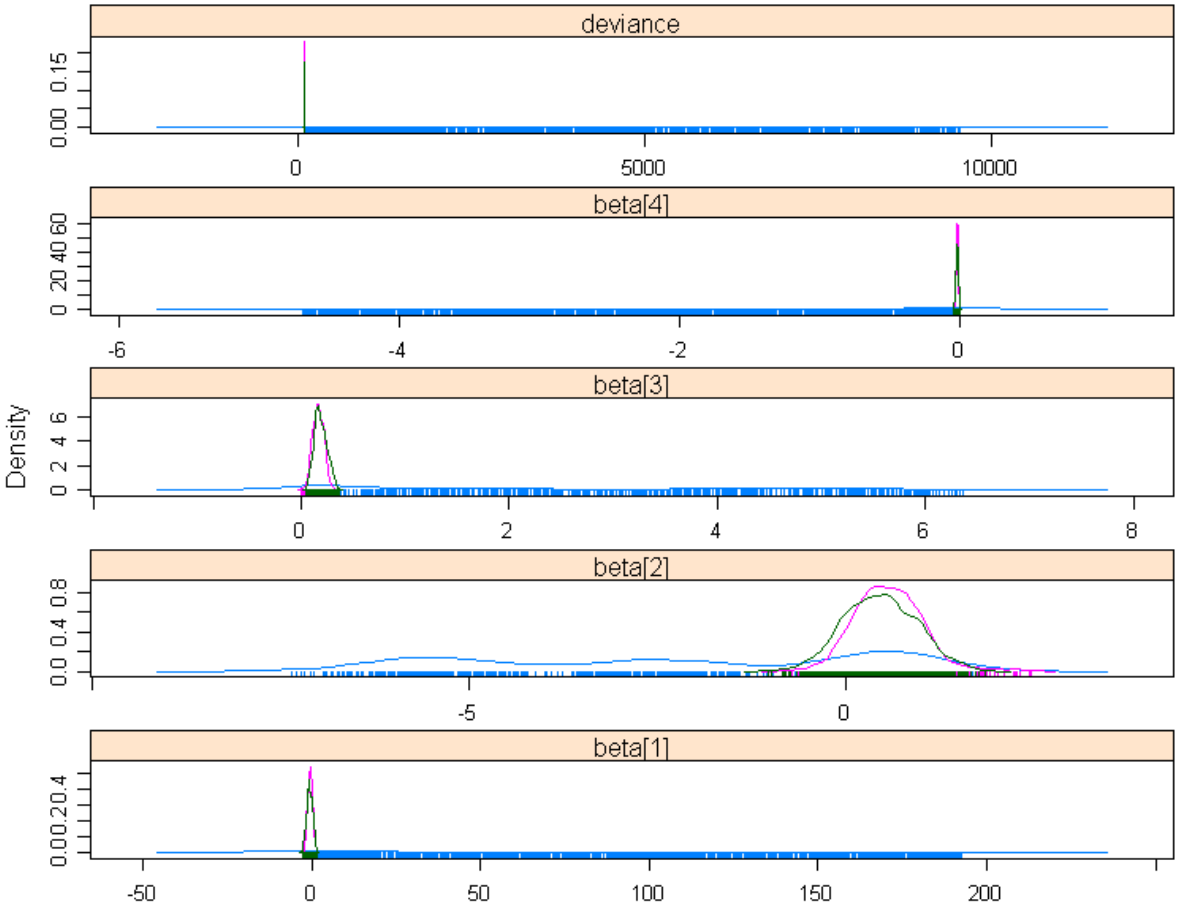
Traceplot



Gelman Plot



densityplot(A)



2.1.2 改进 MCMC 收敛性

模型改进

1. 自变量中心化
2. 增大抽样
3. 稀疏抽样

模型中心化（保存为.txt 文件）

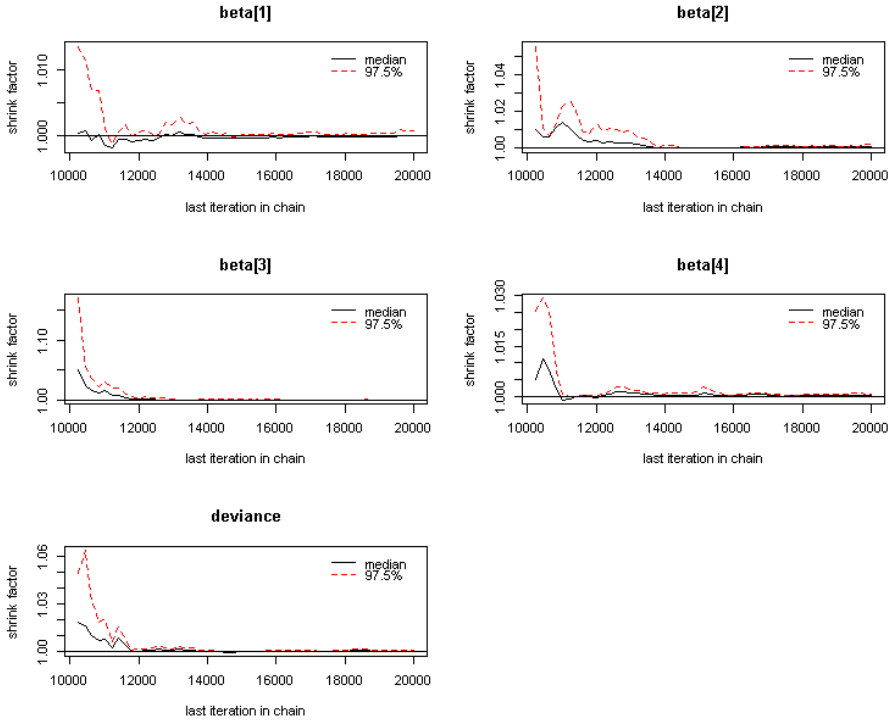
```
model{
# Poisson model likelihood
for (i in 1:30){
```

```
  damage[i] ~ dpois( lambda[i] )
  log(lambda[i])<-beta[1]+beta[2]*(type[i]-mean(type[]))
      + beta[3]*（bombload[i]-mean(bombload[]))
      + beta[4]*(airexp[i]-mean(airexp[]))
  }
# prior
for (j in 1:4){
  beta[j]~dnorm( 0.0, 0.001 )
  B[j] <- exp( beta[j] )
  }
}
```

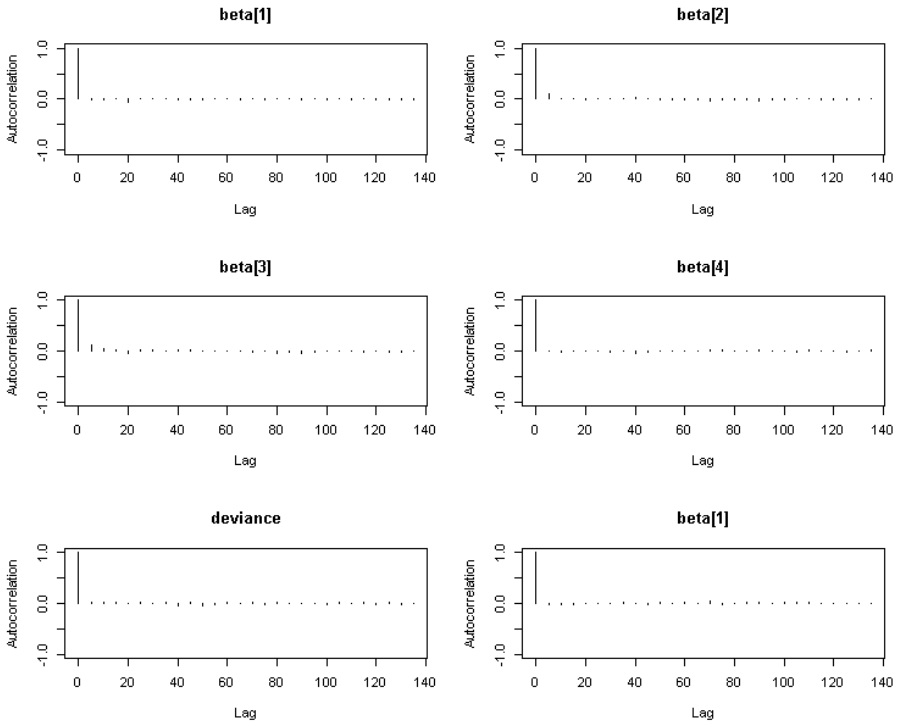
运行 R2WinBUGS

```
output<-bugs(
  data,
  inits,
  parameters,
  n.chains=3,
  n.iter=30000,
  n.burnin=15000,
  n.thin=5,
  debug=FALSE,
  codaPkg=FALSE,
  model.file="F:\\BaiduYun\\Teaching\\Rdata\\Ch6craftmodel_c.txt",
  bugs.directory="D:\\WinBUGS\\",
  working.directory="F:\\Simulation\\"
  )
```

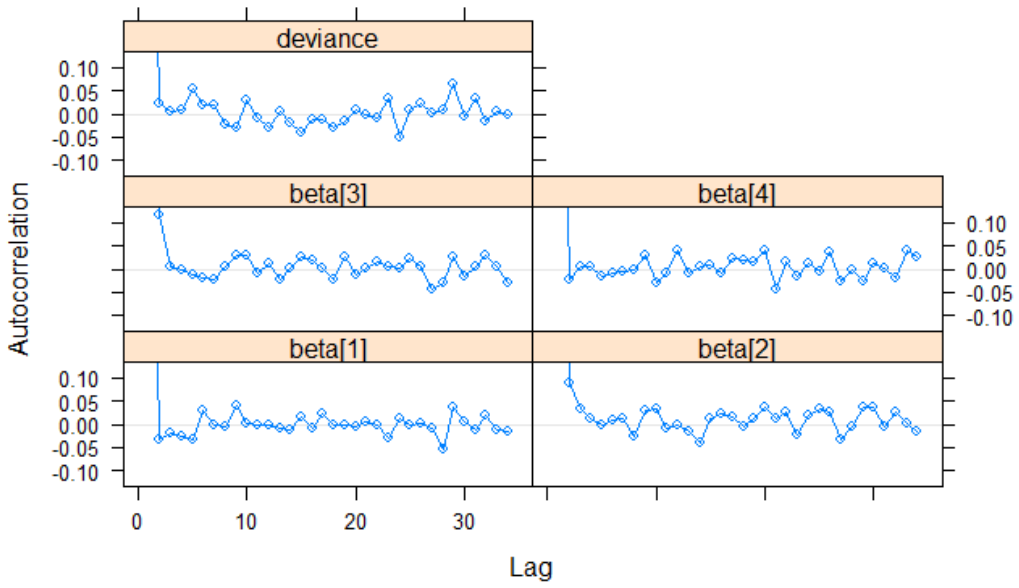
Gelman plot：小于 1.10



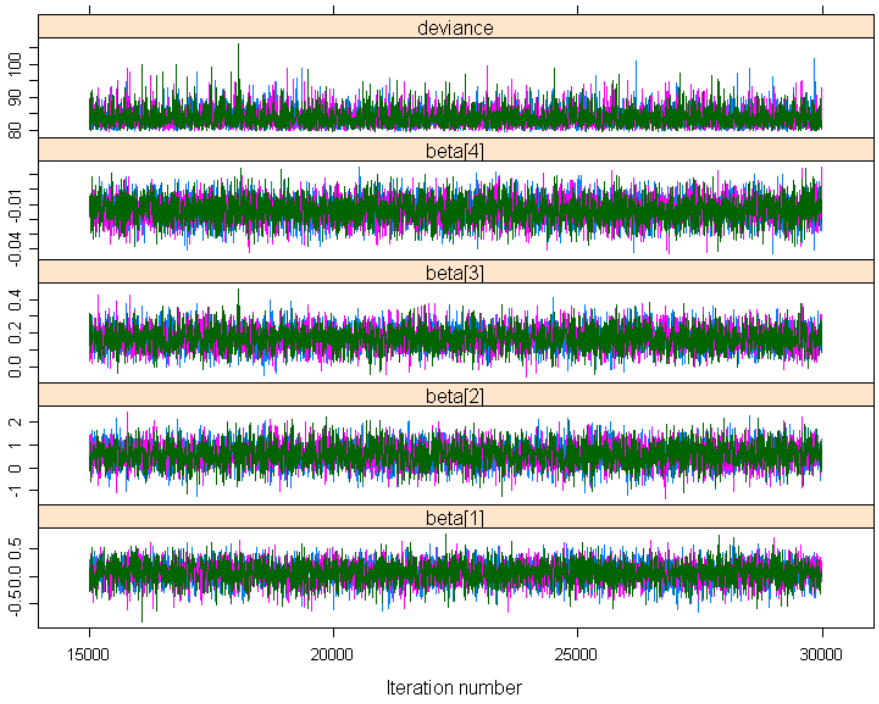
自相关：**autocorr.plot()**


beta[1]
beta[2]
beta[3]
beta[4]
deviance
beta[1]
Autocorrelation
1.0
0.0
-1.0
0
20
40
60
80
100
120
140
Lag

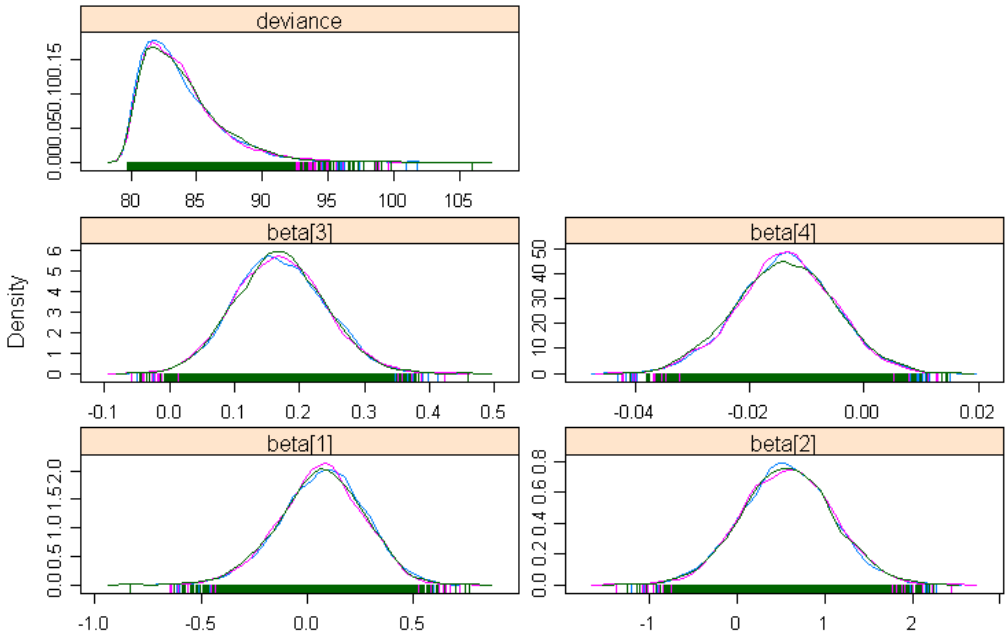
自相关：**acfplot()**


deviance
beta[3]
beta[4]
beta[1]
beta[2]
Autocorrelation
0.10
0.05
0.00
-0.05
-0.10
0
10
20
30
Lag

Traceplot



densityplot(A)



2.1.3 最终模型的结果及解释

运行结果

```
> print(output,digit=3)
Inference for Bugs model at "F:\BaiduYun\Teaching\Rdata\Ch6craftmodel_c.txt", fit using WinBUGS,
 3 chains, each with 30000 iterations (first 15000 discarded), n.thin = 5
 n.sims = 9000 iterations saved
           mean    sd   2.5%    25%    50%    75%  97.5%  Rhat n.eff
beta[1]   0.067 0.197 -0.334 -0.060  0.073  0.202  0.430 1.001  4300
beta[2]   0.572 0.512 -0.426  0.224  0.572  0.919  1.575 1.001  9000
beta[3]   0.170 0.068  0.041  0.123  0.168  0.216  0.306 1.001  9000
beta[4]  -0.014 0.008 -0.031 -0.019 -0.014 -0.008  0.003 1.001  9000
B[1]      1.090 0.212  0.716  0.941  1.075  1.224  1.537 1.001  4300
B[2]      2.020 1.103  0.653  1.251  1.771  2.506  4.832 1.001  9000
B[3]      1.188 0.081  1.042  1.131  1.183  1.241  1.358 1.001  9000
B[4]      0.986 0.008  0.970  0.981  0.986  0.992  1.003 1.001  9000
deviance 83.738 2.874 80.140 81.620 83.110 85.120 90.880 1.001  3400

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = Dbar-Dhat)
pD = 4.0 and DIC = 87.7
DIC is an estimate of expected predictive error (lower deviance is better).
```

比较 lm 和 glm 的系数估计

```
> coefficients(out_glm)

## (Intercept)        type    bombload      airexp
## -0.40602269  0.56877242  0.16542540 -0.01352232

> coefficients(out_lm)

## (Intercept)        type    bombload      airexp
##   0.4308974   0.5407536   0.3303238  -0.0228258
```

系数解释

泊松回归系数的含义：

1. B[2]=2.02: 在机组人员飞行经验和炸弹载重相同的条件下，A6 型轰炸机的平均损伤数是 A4 型的两倍；
2. B[3]=1.188: 在机组人员飞行经验和机型相同的条件下，载重每增加 1 吨，飞机的平均损伤数增加 18.8%。
3. B[4]=0.986: 在飞机载重和机型相同的条件下，机组人员飞行经验每增加 1 年，飞机的平均损伤数减少 1.4%。
4. 但是，从 95%CI 看，只有 beta[3] 的 CI 不包括 0（显著），其它变量都不显著。

问题: 如果自变量进行标准化，各变量系数如何解释？

2.2 贝叶斯 Logistic 回归: 无信息先验

Logistic 回归：贝叶斯模型

贝叶斯模型：

$$Y_i \sim \text{Bernoulli}(p_i)$$

$$\text{logit}(p_i) = \beta_0 + \beta_1 T$$

先验分布（无信息）：

$$\beta_0 \sim dflat(), \beta_1 \sim dflat()$$

计算：$T = 55$ 和 $T = 75$ 时，O-环失效的概率，公式：

$$p = \frac{e^{\beta_0+\beta_1 T}}{1 + e^{\beta_0+\beta_1 T}}$$

WinBUGS model

```
model {
mut<-mean(T[])
   for(i in 1:n){
   y[i]~dbern(p[i])
   logit(p[i])<-beta[1]+beta[2]*(T[i]-mut)
   }
Prob1<-exp(beta[1]+beta[2]*(55-mut))/(1+exp(beta[1]
       +beta[2]*(55-mut)))
Prob2<-exp(beta[1]+beta[2]*(75-mut))/(1+exp(beta[1]
       +beta[2]*(75-mut)))
  for(j in 1:2){
  beta[j]~dflat()
  }}
```

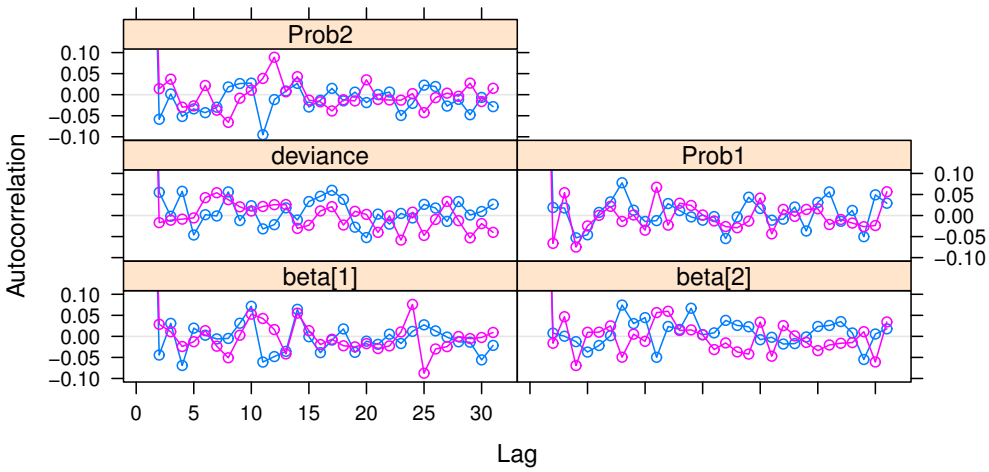
initial values & parameters

```
# Define inits
inits1<-list(beta=c(0,0))
inits2<-list(beta=c(-1,1))
inits=list(inits1,inits2)
# List Parameters that will be monitored
parameters<-c("beta","Prob1","Prob2")
```

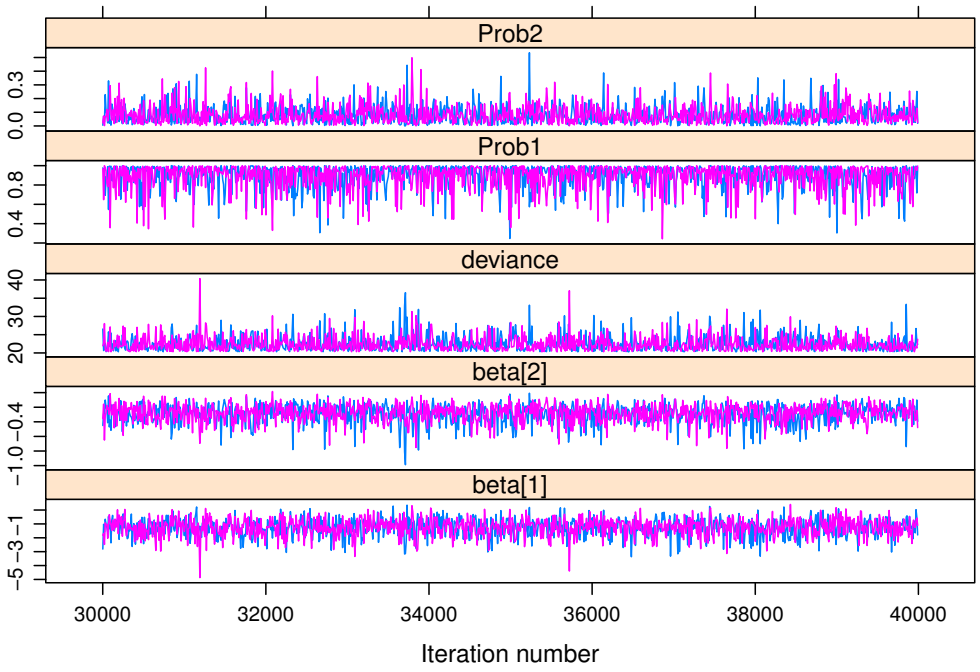
Call R2WinBUGS

```
library(R2WinBUGS)
oringout<-bugs(data,
inits,
parameters,
n.chains=2,
n.iter=40000,
n.burnin=30000,
n.thin=10,
debug=FALSE,
codaPkg=FALSE,
model.file="F:\\BaiduYun\\Teaching\\Rdata\\ch6oringmodel_c.txt",
bugs.directory="D:\\WinBUGS\\",
working.directory="F:\\Simulation\\"
)
```

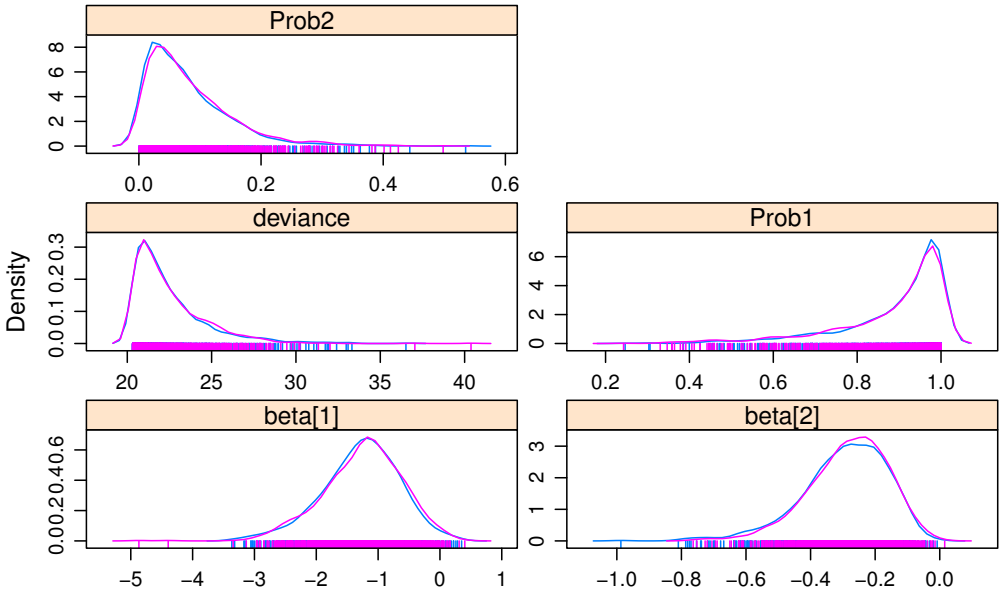
检查收敛性:ACF


Prob2
deviance
Prob1
beta[1]
beta[2]
Autocorrelation
0.10
0.05
0.00
−0.05
−0.10
0
5
10
15
20
25
30
Lag

Trace


Prob2
Prob1
deviance
beta[2]
beta[1]
0.0 0.3
0.4 0.8
20 30 40
−1.0 −0.4
−5 −3 −1
30000
32000
34000
36000
38000
40000
Iteration number

Density



模型结果

```
> print(oringout,digit=3)
Inference for Bugs model at "F:\BaiduYun\Teaching\Rdata\ch6oringmodel_c.txt", fit using WinBUGS,
 2 chains, each with 40000 iterations (first 30000 discarded), n.thin = 10
 n.sims = 2000 iterations saved
           mean    sd   2.5%    25%    50%    75%  97.5%  Rhat n.eff
beta[1]  -1.263 0.636 -2.635 -1.644 -1.205 -0.834 -0.115 1.001  1800
beta[2]  -0.289 0.128 -0.586 -0.362 -0.277 -0.196 -0.085 1.007   390
Prob1     0.889 0.131  0.507  0.844  0.939  0.982  0.999 1.007   750
Prob2     0.081 0.069  0.005  0.030  0.063  0.113  0.269 1.004   460
deviance 22.477 2.146 20.370 20.970 21.830 23.280 28.040 1.001  2000

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = Dbar-Dhat)
pD = 1.9 and DIC = 24.4
DIC is an estimate of expected predictive error (lower deviance is better).
```

2.3 贝叶斯 Logistic 回归：弱信息先验

有先验信息

“挑战者”号发射前夕，有科学家告诉宇航局领导者，地面温度太低，发射有风险。记 P_1 和 P_2 为分别在 $T = 55$ 和 $T = 75$ 时 O-环失效的概率，对一些专家进行初步调查，有

$$P(P_1 > 0.5) = 2/3,\ P(P_2 < 0.5) = 2/3 \tag{1}$$

据此先验信息建立贝叶斯模型。

$$Y_i \sim \text{Bernoulli}(p_i)$$

$$\text{logit}(p_i) = \beta_0 + \beta_1 T$$

先验分布：

1. 给出 P_1 和 P_2 的先验分布：$P_1 \sim \text{Beta}(1, 1.6)$, $P_1 \sim \text{Beta}(1.6, 1)$（超参数根据公式（1）得出）
2. 通过 $\text{logit}(P_1) = \beta_0 + 55\beta_1$, $\text{logit}(P_2) = \beta_0 + 75\beta_1$，反解出：

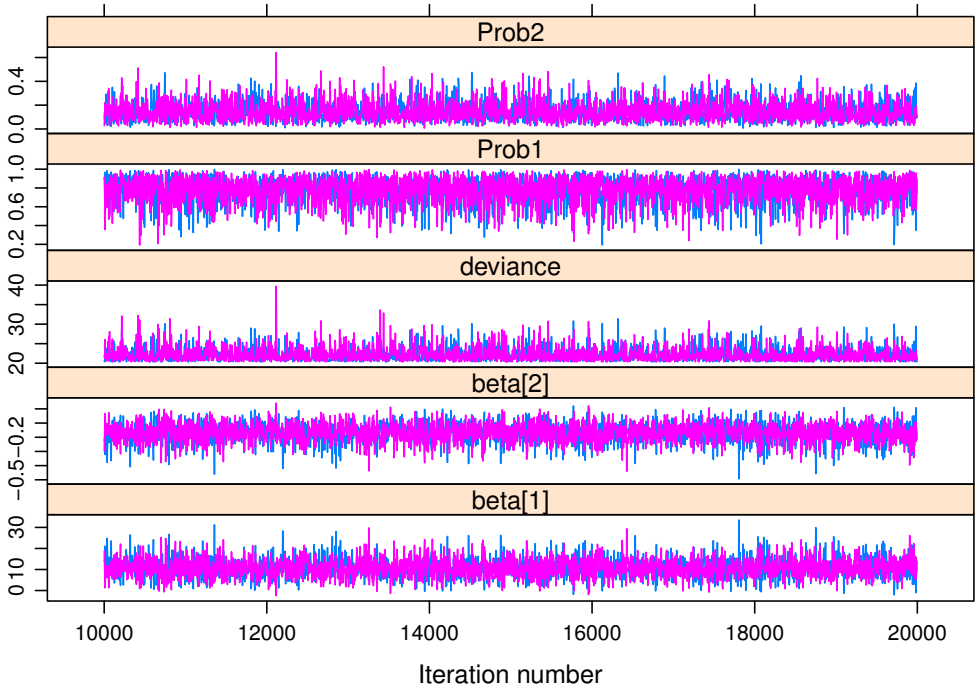
$$\beta_0 = (75/20)\text{logit}(P_1) + (-55/20)\text{logit}(P_2)$$

$$\beta_1 = (-1/20)\text{logit}(P_1) + (1/20)\text{logit}(P_2)$$

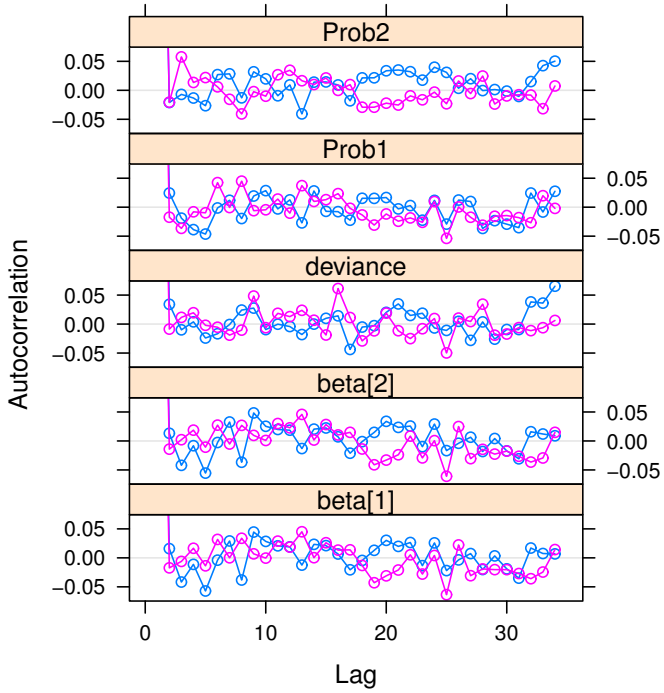
Model Codes

```
model {
   for(i in 1:n){
   y[i]~dbern(p[i])
   logit(p[i])<-beta[1]+beta[2]*T[i]
   }
Prob1<-exp(beta[1]+beta[2]*55)/(1+exp(beta[1]+beta[2]*55))
Prob2<-exp(beta[1]+beta[2]*75)/(1+exp(beta[1]+beta[2]*75))
 P55~dbeta(1.6,1)
 P75~dbeta(1,1.6)
 beta[1]<-(75/20)* logit(P55)+(-55/20)*logit(P75)
 beta[2]<-(-1/20)*logit(P55)+(1/20)*logit(P75)
}
#initial values for informative prior.
inits1.p<-list(P55=0.5,P75=0.5)
inits2.p<-list(P55=0.1,P75=0.9)
inits.p=list(inits1.p,inits2.p)
```

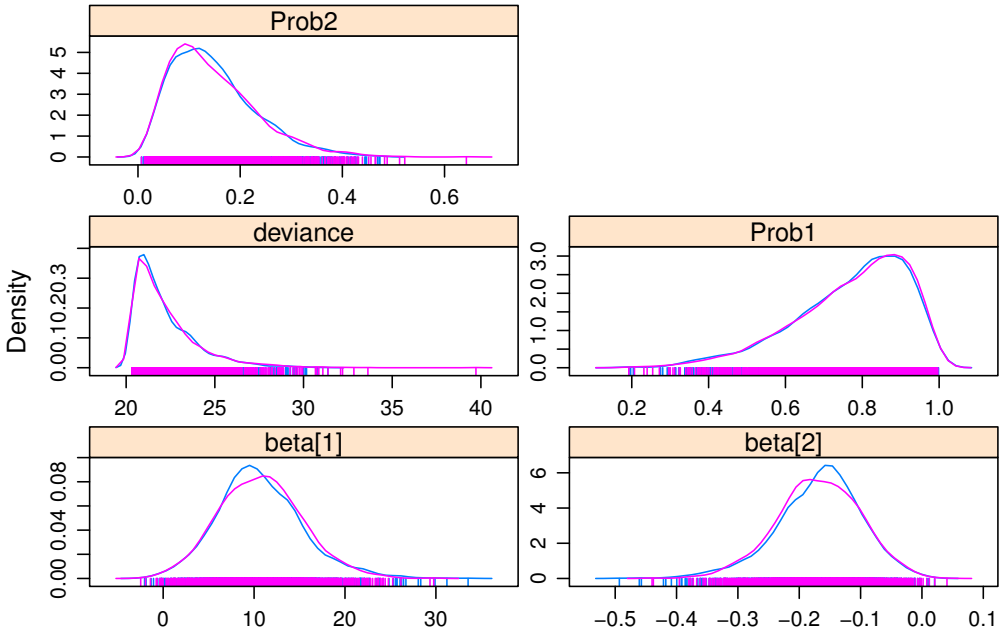
MCMC Trace



ACF Autocorrelation


Prob2
Prob1
deviance
beta[2]
beta[1]
Autocorrelation
0.05
0.00
−0.05
0
10
20
30
Lag

Density



Outputs

```
> print(oringout,digit=3)
Inference for Bugs model at "F:\BaiduYun\Teaching\Rdata\ch6oringmodel_p.txt", fit using WinBUGS,
 2 chains, each with 20000 iterations (first 10000 discarded), n.thin = 5
 n.sims = 4000 iterations saved
           mean    sd   2.5%    25%    50%    75%  97.5%  Rhat n.eff
beta[1]  10.700 4.663  2.110  7.533 10.445 13.590 20.880 1.001  4000
beta[2]  -0.168 0.069 -0.319 -0.211 -0.165 -0.122 -0.041 1.001  4000
Prob1     0.772 0.146  0.421  0.684  0.803  0.887  0.973 1.001  4000
Prob2     0.148 0.084  0.031  0.084  0.132  0.195  0.348 1.001  4000
deviance 22.210 1.820 20.360 20.920 21.670 22.940 27.080 1.003  3000

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = Dbar-Dhat)
pD = 1.2 and DIC = 23.4
DIC is an estimate of expected predictive error (lower deviance is better).
```

预测

温度：	31	55	75
逻辑斯特模型：	0.99960878	0.907	0.086
贝叶斯估计（无信息先验）：		0.669	0.081
贝叶斯估计（弱信息先验）：		0.772	0.148