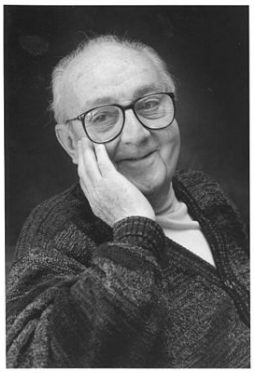
第八章 模型的评价

Wang Shujia

Contents

"All models are wrong but some are useful"



George E. P. Box
(1919 -2013)

模型质量的评价

1. 模型检查 (Model Checking)：模型是否合适？
   - 模型的敏感性
     - 不同的先验分布
     - 不同的总体模型
     - 模型自变量的不同
   - 后验预测检查：模型的预测能力
     - Posterior Predictive Checking(PPC)
2. 模型比较 (Model Comparison)：选择哪个模型？
   - 交叉验证
   - 信息准则
   - 贝叶斯因子

1 模型检查 (Model Checking)

1.1 模型检查为什么重要？

例：线性回归模型真的可靠吗？

- 考虑模型：$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon,\ \varepsilon \sim N(0, \sigma^2)$
- 真实模型为：$y = x_1 + x_2 + \varepsilon,\ \varepsilon \sim N(0, 5)$
- 即真实参数：$\beta_0 = \beta_3 = 0, \beta_1 = \beta_2 = 1, \sigma^2 = 5$
- 从真实模型中随机模拟 N 个样本数据

```
N=30
set.seed(123456)
x1=rnorm(N,0,1)
x2=rnorm(N,0,1)
x3=rnorm(N,0,1)
y=x1+x2+rnorm(N,0,5)
output<-lm(y~x1+x2+x3)
summary(output)
```

30 个样本值未能正确识别模型变量

模型：$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon,\ \varepsilon \sim N(0, \sigma^2)$

```
Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3210 -2.3343 -0.5798  3.3388  7.4133

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.0582     0.9387   1.127    0.270
x1            0.5465     0.9332   0.586    0.563
x2            1.3732     0.8612   1.594    0.123
x3           -0.6307     0.8318  -0.758    0.455

Residual standard error: 4.462 on 26 degrees of freedom
Multiple R-squared:  0.1231,    Adjusted R-squared:  0.02188
F-statistic: 1.216 on 3 and 26 DF,  p-value: 0.3236
```

什么条件下才能正确识别模型？

1. $N = 30$
   - $\sigma^2 = 5$：x_1, x_2 都不显著，R^2 =0.022
     - 未能识别真实模型，数据与模型拟合差
   - $\sigma^2 = 3$：仅 x_2 显著，$R^2 = 0.169$
     - 部分识别，数据与模型拟合较差
   - $\sigma^2 = 1$：x_1, x_2 显著，$R^2 = 0.72$

  - 完全识别，数据与模型拟合较好

2. $N = 300, \sigma^2 = 5$：x_1, x_2 显著，$R^2 = 0.048$
   - 虽然能正确识别显著变量，但拟合质量较差
3. $N = 10000, \sigma^2 = 5$：x_1, x_2 显著，$R^2 = 0.082$
   - 即使样本量很大，拟合质量仍然较差

结论

只有误差的方差不太大，模型才能正确识别。

- 误差方差大
  - 数据少：模型不能识别
  - 数据多：模型能识别，但预测能力差
- 误差方差小
  - 数据即使较少也能正确识别模型，而且预测能力强

例：线性回归模型的模拟结果

一个简单模拟例子（Freedman，1983）

- 模拟产生**相互独立**的数据（100 行，51 列），前 50 列作为自变量 $X_1, X_2, \ldots, X_{50}$，最后一个作为因变量 Y
- 按照模型设计，进行回归分析时，应该：
  - 整个方程的 F-检验不显著；
  - 各个系数的 t-检验不显著
  - 可决系数 R^2 很小

实际模拟的结果

First pass：Y 对所有 50 个自变量回归

- $R^2 = 0.60$，p-值 $= 0.00001$
- 在 $\alpha = 0.25$ 水平下，有 21 个系数显著
- 在 $\alpha = 0.05$ 水平下，有 7 个系数显著

Second pass：仅对 21 个系数显著（$\alpha = 0.25$）的变量回归

- $R^2 = 0.50$，p-值 $= 0.09$
- 在 $\alpha = 0.25$ 水平下，有 20 个系数显著
- 在 $\alpha = 0.05$ 水平下，有 14 个系数显著
- 在 $\alpha = 0.01$ 水平下，有 6 个系数显著

结论：

1. 传统 LRM 中，用 p-值判断自变量的显著性不一定可靠（从而计量经济分析中的 LRM 影响因素分析不一定可靠）；
2. 模型的可靠性、稳健性需要进一步研究。

1.2 如何进行模型检查？

贝叶斯模型检查的逻辑和方法

1. 线性回归模型的诊断：如果模型 $f(\boldsymbol{y}|X,\boldsymbol{\theta})$ 是“好”的，那么**残差** (Residuals) 应该不大。
2. 贝叶斯模型检查的基本逻辑：如果一个贝叶斯模型（包含总体模型和先验分布）是“好”的，那么从它模拟产生的数据**应该**与实际观察到的数据很类似，即**后验预测分布**中随机抽取的**重复样本集**（记为 $\boldsymbol{y}^{rep}=(y_1^{rep},y_2^{rep},\ldots,y_n^{rep})^T$）与**实际**数据集（$\boldsymbol{y}=(y_1,y_2,\ldots,y_n)^T$）应该区别不大（两者容量相同）。
3. 几点说明：
   - **后验预测分布**（Posterior predictive distribution）
   $$p(\tilde{y}|\boldsymbol{y})=\int f(\tilde{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$$
   - 基于新的自变量观察值 x_0 预测未来值 $\tilde{y}$，称为预测值；
   - 基于所有自变量数据 X，从后验预测分布中重复抽取的样本，称为**重复样本集** $\boldsymbol{y}^{rep}$。
   - 模型检查：比较 $\boldsymbol{y}^{rep}$ 与 $\boldsymbol{y}$（难点是高维）
   - 办法：比较低维数的距离函数 $T(\boldsymbol{y},\boldsymbol{\theta})$ 和 $T(\boldsymbol{y}^{rep},\boldsymbol{\theta})$
   - 评价的是**整个**概率模型，包括似然函数和先验分布

如何抽取重复样本集 $\boldsymbol{y}^{rep}$？

设观察值 $y_1,y_2,\ldots,y_n$ 来自总体 $Y\sim f(y|\boldsymbol{\theta})$，参数的先验分布为 $\boldsymbol{\theta}\sim\pi(\boldsymbol{\theta})$，后验分布 $p(\boldsymbol{\theta}|\boldsymbol{y})\propto\pi(\boldsymbol{\theta})f(\boldsymbol{y}|\boldsymbol{\theta})$。

如何模拟产生后验预测分布的重复样本 $\boldsymbol{y}^{rep}$ 并计算 $T(\boldsymbol{y}^{rep},\boldsymbol{\theta})$？

Obtain L samples from $p(\boldsymbol{y}^{rep}|\boldsymbol{y})$

For $i=1,2,\ldots,L$,

1. Simulate a parameter $\boldsymbol{\theta}^{(i)}\sim p(\boldsymbol{\theta}|\boldsymbol{y})$,
2. Simulate $\boldsymbol{y}_i^{rep}\sim f(\boldsymbol{y}^{rep}|\boldsymbol{\theta}^{(i)})$(dimension n)
3. $T^{(i)}=T(\boldsymbol{y}_i^{rep},\boldsymbol{\theta}^{(i)})$

贝叶斯模型检查的基本步骤

假设样本 $y_1, y_2, \ldots, y_n$ 来自总体 $Y \sim f(y|\boldsymbol{\theta})$，参数的先验分布为 $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$，后验分布 $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta}) f(\boldsymbol{y}|\boldsymbol{\theta})$

1. 确定距离函数（Discrepancy Measure）$T(\boldsymbol{y}, \boldsymbol{\theta})$。这个函数必须：
   - 由数据和参数值可以计算；
   - 确实能反映某个模型假设成立与否所导致的数据差异
   - 常用残差 $T(\boldsymbol{y}, \boldsymbol{\theta}) = \boldsymbol{y} - E(\boldsymbol{y}|\boldsymbol{\theta})$ 和标准化残差 $T(\boldsymbol{y}, \boldsymbol{\theta}) = [\boldsymbol{y} - E(\boldsymbol{y}|\boldsymbol{\theta})]/\mathrm{sd}(\boldsymbol{y}|\boldsymbol{\theta})$
2. 由后验预测分布产生 L 个容量相同（都是 n）的重复数据集，记为 $\boldsymbol{y}_1^{rep}, \ldots, \boldsymbol{y}_L^{rep}$
3. 计算各重复数据集以及实际数据集的距离函数：$T(\boldsymbol{y}_i^{rep}, \boldsymbol{\theta})$ 和 $T(\boldsymbol{y}, \boldsymbol{\theta})$
4. 比较它们的大小（常用图形展示或贝叶斯 p-值判断）。如果差距大，则说明模型不能拟合实际数据

贝叶斯 p-值

1. 贝叶斯后验预测 p-值（Bayesian posterior predictive p-value ）

$$p_B = \mathrm{P}(T(\boldsymbol{y}^{rep}, \boldsymbol{\theta}) > T(\boldsymbol{y}, \boldsymbol{\theta})|\boldsymbol{y})$$

   - 传统 p-值（给定 $\boldsymbol{\theta}$）：

$$p - \mathrm{value}(\boldsymbol{y}|\boldsymbol{\theta}) = \mathrm{P}(T(\boldsymbol{y}^{rep}, \boldsymbol{\theta}) \geq T(\boldsymbol{y}, \boldsymbol{\theta})|\boldsymbol{\theta}, \boldsymbol{y})$$

   - 贝叶斯 p-值：

$$\begin{aligned} p_B &= \mathrm{P}(T(\boldsymbol{y}^{rep}, \boldsymbol{\theta}) > T(\boldsymbol{y}, \boldsymbol{\theta})|\boldsymbol{y}) \\ &= \int \mathrm{P}(T(\boldsymbol{y}^{rep}, \boldsymbol{\theta}) \geq T(\boldsymbol{y}, \boldsymbol{\theta})|\boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{\theta}|\boldsymbol{y}) \mathrm{d}\boldsymbol{\theta} \end{aligned}$$

   - 计算：

$$p_B \approx \frac{1}{L} \sum_{i=1}^{L} I_{[T(\boldsymbol{y}_i^{rep}, \boldsymbol{\theta}^{(i)}) \geq T(\boldsymbol{y}, \boldsymbol{\theta}^{(i)})]}$$

2. 如果 p_B 非常接近 0 或者 1，则意味着观察数据与由模型产生的重复数据显著偏离。如果模型正确，p_B 应该接近 0.5
   - 如果是双侧检验，贝叶斯 p-值为 $2\min(p, 1-p)$

例：二项试验中的独立性检验

- 假设 $y_1, y_2, \ldots, y_n$ 为 n 次 0-1 试验的观察结果，θ 表示结果为 1 的概率。
  - 观察数据依次为：1,1,0,0,0,0,0,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0 (n=20,s=7)
  - 目的：检验各次实验之间的独立性
- 贝叶斯模型：

  $y_i \sim \mathrm{Bin}(1, \theta)$,$\theta \sim U(0, 1)$；
- 后验分布（满足独立性假设条件下）：

  $\theta|\boldsymbol{y} \propto \theta^s (1-\theta)^{n-s} \sim Beta(s+1, n-s+1)$, $s = \sum y_i$

如何选取距离函数 $T(\boldsymbol{y}, \boldsymbol{\theta})$?

在具体问题中，应如何选取距离函数？

- 能反映模型与数据之间的系统性差异
- 能反映模型特征的量（如独立性、单调性、偏度峰度等）
- 要有清晰含义
- 距离函数最好不要与模型参数重合（观察数据重复使用了）

二项试验的独立性检验

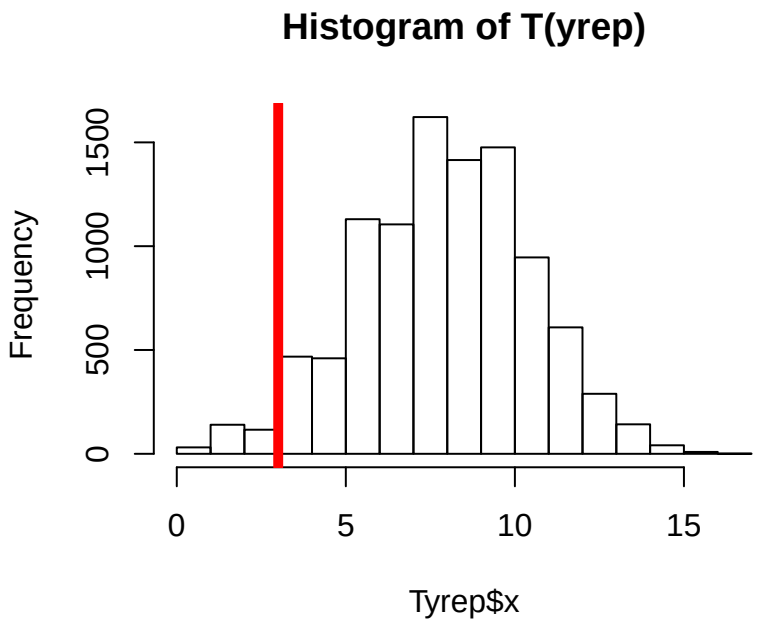
距离函数：$T=$ 数据在 0 和 1 之间转换的次数
观察值：$T(y) = 3$
检查步骤：

1. 抽取重复样本集 $\boldsymbol{y}_1^{rep}, \ldots, \boldsymbol{y}_L^{rep}$：
   (a) 产生后验样本：$\theta_l|\boldsymbol{y} \sim \text{rbeta}(1, 8, 14)$
   (b) 抽取样本集：$\boldsymbol{y}_l^{rep} \sim \text{rbinomial}(n = 20, 1, \theta_l)$
2. 计算 $T(\boldsymbol{y}^{rep}, \theta)$ 及贝叶斯 p-值，做出 T 的直方图

Codes

```
y <- c(1,1,0,0,0,0,0,1,1,1,1,1,0,0,0,0,0,0,0,0)
n <- length(y)
# num of success
s <- sum(y)
# test quantity for real data
Ty <- sum(diff(y) != 0) + 0.0
# 定义函数
Trep <- function(s, n) {
 p <- rbeta(1, s+1, n-s+1) # 从后验分布抽取一个参数 p 的值
 yrep <- rbinom(n, 1, p) # 以该 p 值为概率模拟抽取 n 个样本值
 sum(diff(yrep) != 0) + 0.0 # 输出 0-1 转换次数
 }
# 重复数据集
Tyrep <- data.frame(x = replicate(10000, Trep(s, n)))
# 计算贝叶斯 p 值，直方图
mean(Tyrep<=Ty) # Bayesian p-value
hist(Tyrep$x,main="Histogram of T(yrep)")
abline(v=3, lwd = 5, col = 2)
```

距离函数直方图（p-值 $=0.0272$）



1.3 后验预测检查 (PPC)

基于 bayesplot 的 PPC 图示

软件包 bayesplot 包含 PPC 模块，可以图示比较从后验预测分布中抽取的重复样本与实际观察值之间的差异。

Distributions Histograms, kernel density estimates, boxplots 等，比较 y 和 yrep.

Test-statistics 比较 yrep 和 y 的距离函数.

Intervals Interval estimates of yrep with y overlaid.

Predictive-errors Plots of predictive errors (y - yrep) computed from y and replicated datasets (rows) in yrep.

Scatterplots Scatterplots of y vs. yrep, or vs. the average value of the distributions of each data point (columns) in yrep.

Plots-for-discrete-outcomes PPC functions that can only be used if y and yrep are discrete.

LOO-predictive-checks PPC functions for predictive checks based on (approximate) leave-oneout (LOO) cross-validation.

例：汽车油耗的简单线性回归模型

- 模型一（正态误差）：$y_i = \alpha + \beta x_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$
- 模型二（t 分布误差）：$y_i = \alpha + \beta x_i + \varepsilon_i, \varepsilon_i \sim t(\nu = 2, 0, \sigma^2)$
- α, β, σ 为弱信息先验分布且相互独立

模型 PP Check 的计算

- 模型一可以直接运用 Rstanarm 中的函数类 pp_check()，也可以先用 posterior_predict() 抽取后验预测分布的样本 y^{rep}，然后用 bayesplot 的 PPC 函数展示。
- 模型二要用 Rstan 运行，而 RStan 不会自动计算和记录 log-likelihood 和 y^{rep}。
- RStan 如何抽取后验预测分布样本？要在模型代码中设 generated quantities 模块，然后抽取 y^{rep} 并用 bayesplot 的 PPC 函数展示。

```
generated quantities {
vector[N] log_lik; //pointwise log-likelihood for LOO
vector[N] y_rep; //replications from posterior predictive dist
  for (n in 1:N) {
    real y_hat_i = alpha + beta * x[n];
    log_lik[n] = student_t_lpdf(y[n] |2, y_hat_i, sigma);
    y_rep[n] = student_t_rng(2, y_hat_i, sigma);
  }
}
```

100 个重复样本的分布密度

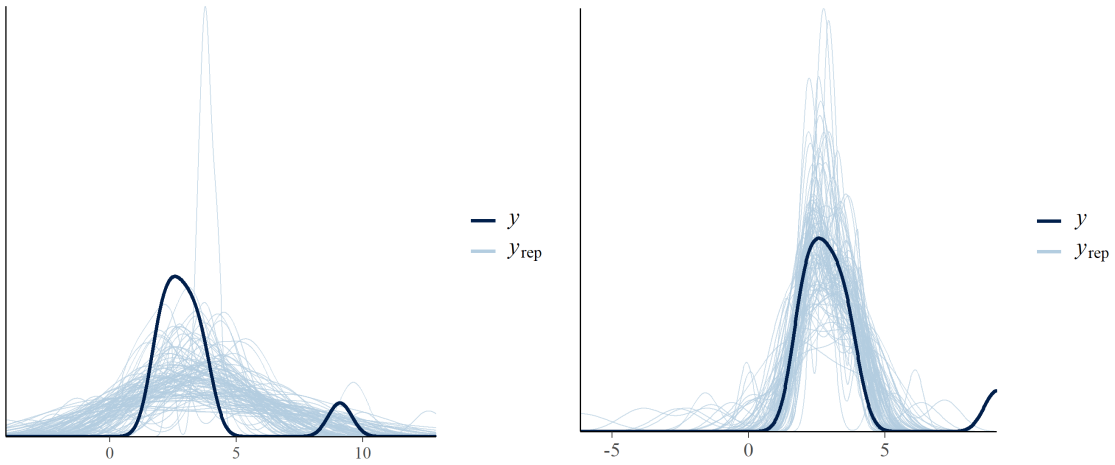


Figure 1: 左边为模型一，右边为模型二

20 个重复样本的 boxplot

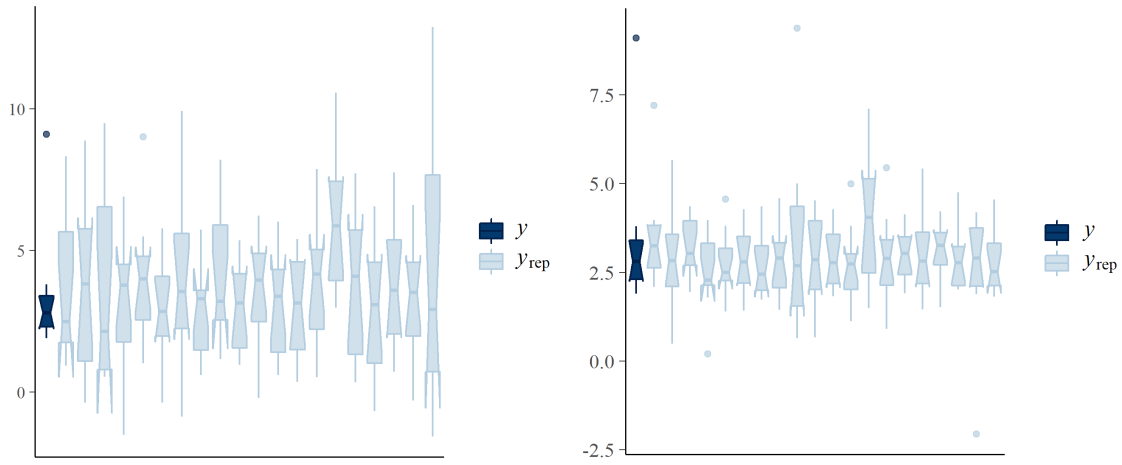


Figure 2: 左边为模型一，右边为模型二

重复样本的 Interval

距离函数的比较：median

距离函数的比较：sd

距离函数的比较：max

距离函数的比较：min

距离函数的比较：skew

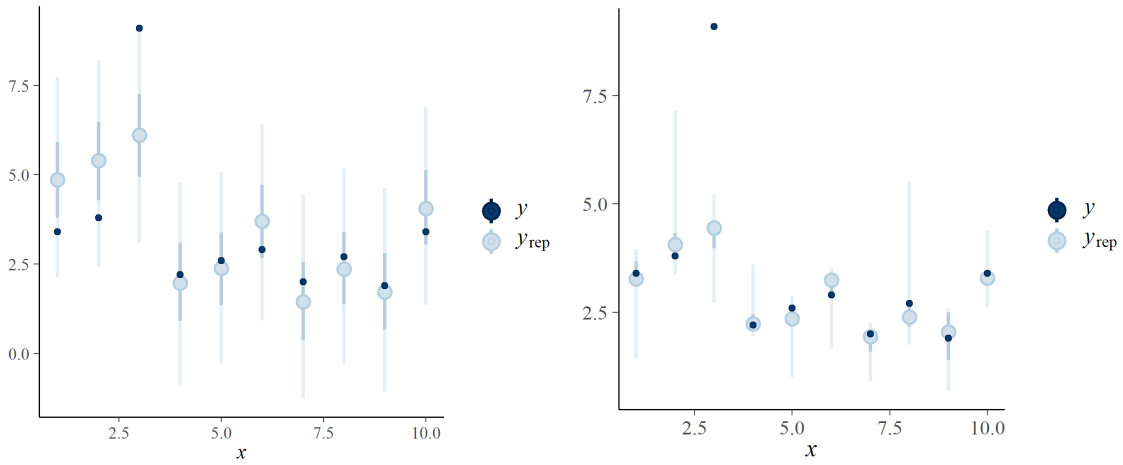


Figure 3: 左边为模型一，右边为模型二

2 模型比较 (Model Comparison)

奥卡姆剃刀法则

假设对同一个问题/数据，有两个模型（对数据的解释同样好），应该选择哪个模型呢？

奥卡姆剃刀法则 (Ockham's razor)：对同一种现象有两种不同的假说，我们应该采取比较简单的那一种。

1. 汽车油耗的例子
   (a) 模型 1：5 阶多项式模型
   (b) 模型 2：3 阶多项式模型
   (c) 模型 3：1 阶多项式模型
   (d) 模型 4：0 阶多项式模型，即常数
2. 简洁性与准确性
   - 高偏差：参数太少，模型不能很好拟合数据，导致欠拟合
   - 高方差：参数太多，模型对数据中的噪声敏感，导致过拟合
3. 如何权衡？
   (a) 正则化先验：让模型不要对数据反应过大（传统方法：目标函数加上惩罚项）
   (b) 信息准则

模型的评估与比较

1. 信息准则 (Information Criterion)
   - 从信息的视角衡量模型的解释能力和模型的复杂度

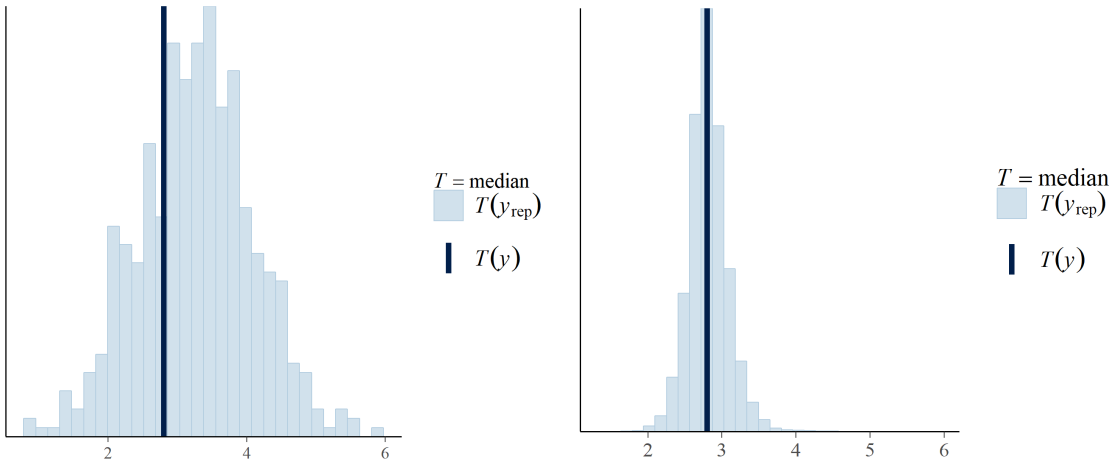


Figure 4: 左边为模型一，右边为模型二

- 常用 DIC、AIC、BIC、WAIC

2. 交叉验证 (LOO)
   - 将数据分成多个子集，然后轮流将其中一个子集作为测试集，将剩余的子集作为训练集训练模型，比较模型的预测能力
   - 如果数据集分成 K 份，称为 K 折交叉验证
   - 如果分成 n 份（n 为样本数据总数），则称为留一交叉验证 (Leave-One-Out Cross Validation, LOO)
3. 贝叶斯因子 (Bayes Factor)
   - 按贝叶斯理论对两个模型进行假设检验
4. LOO 和 WAIC 比 DIC、AIC、BIC 更具优势（但是以前因为计算问题很少应用）

2.1 交叉验证

用对数得分衡量模型的准确性

假设数据模型为 $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T \sim f(\boldsymbol{y}|\boldsymbol{\theta})$，先验分布为 $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)^T \sim \pi(\boldsymbol{\theta})$，后验分布为 $p(\boldsymbol{\theta}|\boldsymbol{y})$，后验预测分布为

$$p(\tilde{y}|\boldsymbol{y}) = \int f(\tilde{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}$$

常用对数得分 (Log Score) **衡量模型的预测能力** (log-pointwise-predictive-density)：

$$\text{lppd} = \sum_{i=1}^{n} \log p(y_i|\boldsymbol{y}) = \sum_{i=1}^{n} \log \int f(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}$$

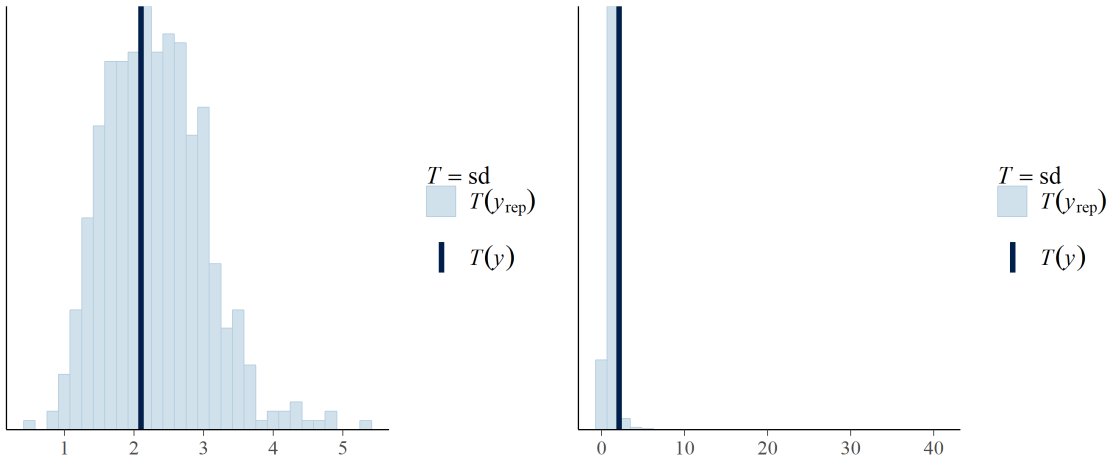


Figure 5: 左边为模型一，右边为模型二

对数得分 (lppd) 的计算：从后验分布中抽取样本 $\theta^s (s = 1, 2, \ldots, S)$ 后，

$$\widehat{\text{lppd}} = \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(y_i|\theta^s) \right).$$

- 对数得分越大，模型的预测能力越好
- 对数得分在一定条件下等价于 Kullback-Leibler 信息

留一交叉验证 (Leave-one-out cross-validation, LOO)

留一交叉验证 (LOO)：比较模型拟合好坏程度加上模型复杂度的修正

$$\text{elpd}_{\text{loo}} = \sum_{i=1}^{n} \log p(y_i|\boldsymbol{y}_{-i}) - p_{\text{loo}}$$

其中 $p(y_i|\boldsymbol{y}_{-i})$ 表示剔除第 i 个数据后，基于剩余的 $(n-1)$ 个数据 (记为 $\boldsymbol{y}_{-i}$) 所得出的后验预测分布，即

$$p(y_i|\boldsymbol{y}_{-i}) = \int f(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}_{-i}) d\boldsymbol{\theta}$$

有效参数个数 (Effective number of parameters, p_{loo})：非交叉验证的后验预测分布密度的对数与 elpd_{loo} 之差，即

$$\hat{p}_{\text{loo}} = \widehat{\text{lppd}} - \widehat{\text{elpd}}_{\text{loo}}$$

LOO 信息准则 (Loo information criterion, LOOIC)：

$$\text{LOOIC} = -2\widehat{\text{elpd}}_{\text{loo}} + 2\hat{p}_{\text{loo}}$$

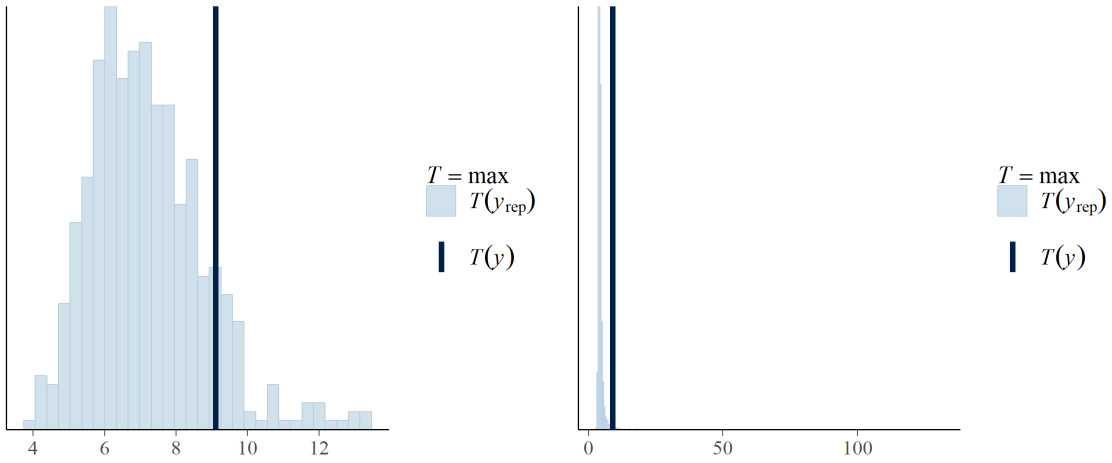


Figure 6: 左边为模型一，右边为模型二

LOO 的计算

计算：R 软件包 loo

计算方法：Pareto-smoothed importance sampling (PSIS)

- The Pareto k diagnostic estimates how far an individual leave-one-out distribution is from the full distribution
- WAIC is asymptotically equal to LOO, but PSIS-LOO is more robust

LOO 在 R 和 RStan 中如何计算？

- 如果模型由 Rstanarm 建立，可以直接调用 loo
- 如果模型由 RStan 建立，则在模型代码的 generated quantities 模块，给出 log_lik（与 y_rep 一起计算）
- 模型比较：
  - loo1 <- loo(fit1)
  - loo_compare(loo1, loo2, loo3...)

例：汽车油耗正态模型（应用 Rstanarm）

```
>fit_n<-stan_glm(y~x,family=gaussian(link="identity"),data=data)
>loo_n <- loo(fit_n)
>print(loo_n)
Computed from 4000 by 10 log-likelihood matrix
         Estimate   SE
elpd_loo    -21.6  5.2
p_loo         5.0  3.6
looic        43.3 10.3
------
Monte Carlo SE of elpd_loo is NA.
Pareto k diagnostic values:
```

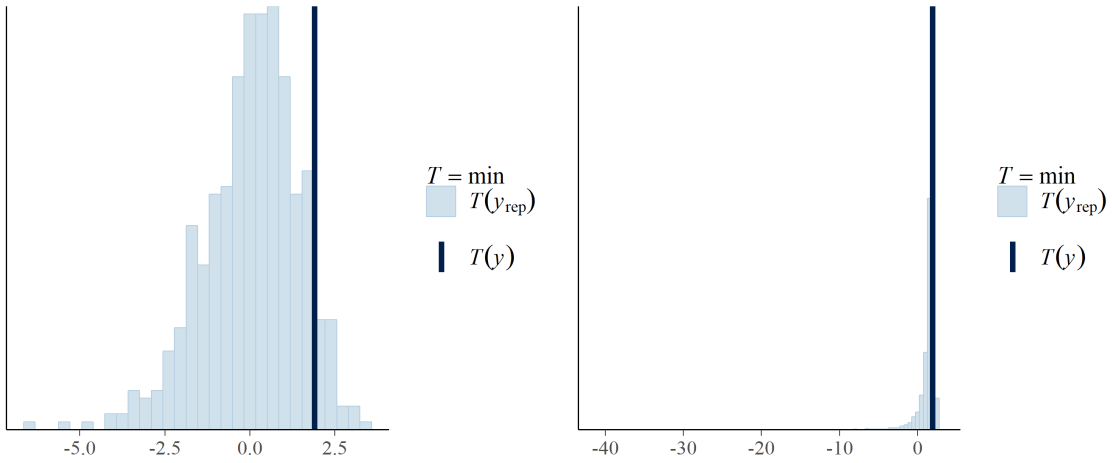


Figure 7: 左边为模型一，右边为模型二

```
                        Count  Pct.    Min. n_eff
(-Inf, 0.5]   (good)     7     70.0%   1434
 (0.5, 0.7]   (ok)       2     20.0%   846
   (0.7, 1]   (bad)      0      0.0%   <NA>
   (1, Inf)   (very bad) 1     10.0%   4
See help('pareto-k-diagnostic') for details.
```

例：汽车油耗稳健模型（应用 RStan）

```
# Extract pointwise log-likelihood and compute LOO
> log_lik_t<-extract_log_lik(fit_t, merge_chains=FALSE)
# PSIS effective sample sizes
> r_eff <- relative_eff(exp(log_lik_t))
loo_t <- loo(log_lik_t, r_eff = r_eff, cores = 2)
> print(loo_t) # (100% good and OK)
> loo_compare(loo_n, loo_t) # 比较两个模型
       elpd_diff se_diff
model2    0.0       0.0   # 最好的模型排在第一行
model1  -11.3       2.9
```

例：比较工资的影响因素模型 (应用 Rstanarm)

```
model1 <- stan_glm(lwage~educ+exper+female+married,
          data=wage_data,family=gaussian(link="identity"))
model2 <- update(model1,
                 formula = lwage ~ educ + exper + female)
model3 <- update(model1, formula = lwage ~ educ + exper)
loo1 <- loo(model1)
loo2 <- loo(model2)
loo3 <- loo(model3)
loo_compare(loo1, loo2, loo3)
          elpd_diff se_diff
model1    0.0       0.0
model2   -0.5       1.7
model3 -111.7      14.5
```

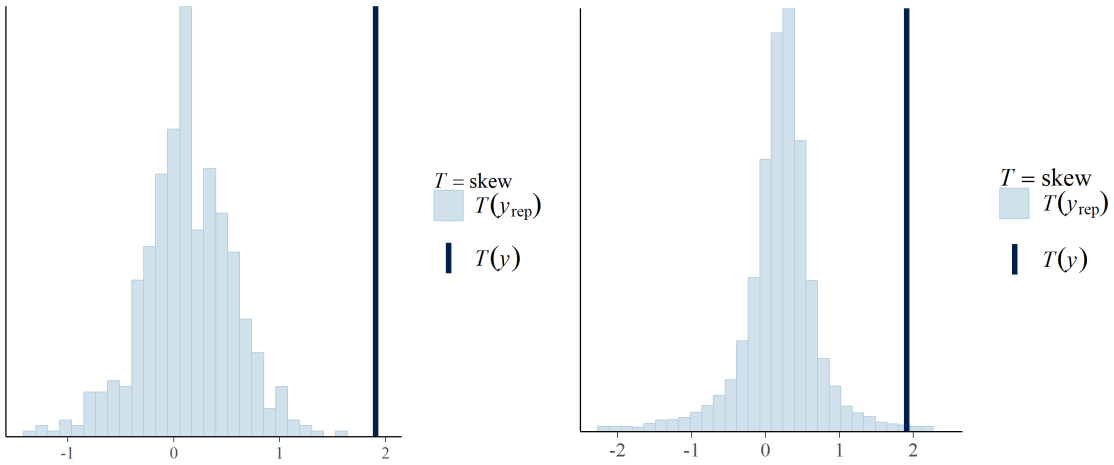


Figure 8: 左边为模型一，右边为模型二

2.2 信息准则

赤池信息准则（AIC）

模型：$\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T \sim f(\boldsymbol{y}|\boldsymbol{\theta}), \boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^T \sim \pi(\boldsymbol{\theta})$，

偏离度 (Deviance)：

$$D(\boldsymbol{y}|\boldsymbol{\theta}) = -2\log f(\boldsymbol{y}|\boldsymbol{\theta})$$

定义 1 (AIC: Akaike Information Criterion)**.**

$$\text{AIC} = D(\boldsymbol{y}|\hat{\boldsymbol{\theta}}) + 2p$$

其中 $\hat{\boldsymbol{\theta}}$ 为 $\boldsymbol{\theta}$ 的最大似然估计，p 为模型中未知参数的个数。

- 赤池信息准则不是贝叶斯方法（或相当于无信息先验）。
- *AIC* 越小，模型拟合越好
- 参数个数 p 为模型复杂度的惩罚项

偏差信息准则 (DIC)

定义 2 (DIC: Deviance Information Criterion)**.**

$$\text{DIC} = \hat{D} + 2p_{DIC}$$

其中

$\hat{D} = D(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_B)$：$\hat{\boldsymbol{\theta}}_B$ 为 $\boldsymbol{\theta}$ 的后验均值（点估计）

$p_{DIC} = \bar{D} - \hat{D}$：有效参数个数，其中 $\bar{D} = E_{\boldsymbol{\theta}}[D(\boldsymbol{y}|\boldsymbol{\theta})]$（可用 *MCMC* 平均计算）。

AIC 与 DIC 的区别：

1. AIC 是 $D(\boldsymbol{y}|\boldsymbol{\theta})$ 在最大似然估计 $\hat{\boldsymbol{\theta}}$ 取值，DIC 是在 $\hat{\boldsymbol{\theta}}_B$（后验均值）取值，所以一定程度上看作 AIC 的贝叶斯形式；
2. AIC 直接加上参数个数，DIC 加上有效参数；
3. DIC 在 WinBUGS 直接输出，AIC 和 BIC 要自行计算。

贝叶斯信息准则（BIC）

定义 3 (BIC:Bayesian Information Criterion)**.**

$$\text{BIC} = D(\boldsymbol{y}|\hat{\boldsymbol{\theta}}) + p\log n$$

其中 p 为模型中未知参数的个数，n 为样本容量。BIC 也是 $D(\boldsymbol{y}|\boldsymbol{\theta})$ 在最大似然估计 $\hat{\boldsymbol{\theta}}$ 取值。

BIC 实际上并不是贝叶斯方法（名称误导哈），是对 AIC 的某种修正：

- 惩罚项除了参数个数 p，还包含样本大小 n
- 对越大的数据集 n 值，对每个参数的惩罚也越大

通用信息准则 (WAIC)

定义 4 (WAIC: Widely Available Information Criterion)**.**

$$\text{WAIC} = -2\widehat{\text{lppd}} + 2\hat{p}_{\text{waic}},\ \widehat{\text{elpd}}_{\text{waic}} = \widehat{\text{lppd}} - \hat{p}_{\text{waic}}$$

其中

$$\widehat{\text{lppd}} = \sum_{i=1}^{n}\log\left(\frac{1}{S}\sum_{s=1}^{S}f(y_i|\theta^s)\right),\ p_{\text{waic}} = \sum_{i=1}^{n}\text{Var}(\log f(y_i|\theta))$$

$$\hat{p}_{\text{waic}} = \sum_{i=1}^{n}V_{s=1}^{S}(\log f(y_i|\theta^s)),\ V_{s=1}^{S}(a_s) = \frac{1}{S-1}\sum_{s=1}^{S}(a_s-\bar{a})^2$$

- p_{WAIC} *为模型有效参数的个数，可以作为模型复杂度的一种度量，避免过拟合。*
- *WAIC 是贝叶斯方法：对后验预测分布密度函数的对数进行平均*
- *WAIC 比 DIC、AIC、BIC 更具优势*

信息准则的计算

- 模型比较：用 LOO 和 WAIC
- AIC 和 BIC：不是贝叶斯方法，可以利用 lm()，直接由 AIC(fit), BIC(fit) 给出
- DIC：WinBUGS 自动给出，RStan 需要计算
- WAIC：在 LOO 软件包中，waic1 <- waic(fit1)
- WAIC 模型比较：loo_compare(waic1, waic2, waic3...)

例：汽车油耗模型的信息准则比较

```
waic_n <- waic(fit_n)
log_lik_t<-extract_log_lik(fit_t, merge_chains=FALSE)
waic_t <- waic(log_lik_t, r_eff = r_eff)
loo_compare(waic_n, waic_t) # 比较两个模型
```

```
            elpd_diff se_diff
      model2   0.0       0.0
      model1 -10.4       3.7
```

例：AIC 和 BIC 的计算

```
lm1<-lm(lwage ~ educ + exper + female + married,
          data = wage_data)
lm2<-lm(lwage~educ + exper + female, data = wage_data)
lm3 <-lm(lwage ~ educ + exper, data = wage_data)
AIC(lm1,lm2,lm3)
BIC(lm1,lm2,lm3)
```

例：AIC 和 BIC 的计算结果

```
model   df         AIC
lm1     6     1764.002
lm2     5     1765.063
lm3     4     1987.654
----------------------
model   df         BIC
lm1     6     1794.835
lm2     5     1790.757
lm3     4     2008.210
```

3 贝叶斯假设检验及贝叶斯因子 (Bayes Factor)

3.1 贝叶斯假设检验

传统假设检验

给定来自总体 $f(x|\theta)$ 的 iid 样本 $x_1, x_2, \ldots, x_n$，用统计量 $T(\boldsymbol{x})$ 检验

$$H_0 : \theta \leq 0, H_1 : \theta > 0$$

假如统计量 $T(\boldsymbol{x})$ 的观察值为 $T(\boldsymbol{x}) = T^*$，则假设检验的 p-值为

$$p-\text{value} = P(T(\boldsymbol{X}) \geq T^* | \theta = 0) = \int_{T^*}^{+\infty} f_T(t|\theta = 0)\mathrm{d}t$$

其中 $f_T(t|\theta)$ 是统计量 $T(\boldsymbol{X})$ 的 pdf 函数。

对传统假设检验的质疑

1. p-值是基于无限次重复的“假设”的数据 $\boldsymbol{X}$，而不是实际观察到的数据
   - 违背似然原理：推论必须来自观察数据，而不是“假设”的数据
2. p-值是这些“假设”的数据在区间 (T^*, ∞) 的平均，而其实大部分在现实中是不太可能出现的（如接近无穷大的值）
   - 因此，在贝叶斯模型中，先验分布**不建议**非正常均匀分布
3. p-值的计算仅基于 $\theta = 0$ 一个点，却要得出关于整个 θ 的结论
4. 第一类错误和第二类错误的概率计算不符合实际
5. 预先给定显著性水平 α 并以星号 (*) 标示显著程度的做法有时候是不恰当的
6. 总而言之，贝叶斯假设检验之所以争议较少，是因为所有未知参数都看作随机分布，更符合科学直观

贝叶斯假设检验

- 贝叶斯方法不太关注**参数**的假设检验，因为参数的后验分布已经提供了全面信息
  - 关于参数，在实际应用中人们更关心它的区间估计
- 对于 $H_0 : \theta \leq 0, H_1 : \theta > 0$，只需要简单比较后验概率 $P(\theta > 0|\boldsymbol{x})$ 即可得出结论
- 对于 $H_0 : \theta = 0, H_1 : \theta \neq 0$，对连续型变量，这个假设不符合常理，因为 $P(\theta = 0|\boldsymbol{x}) = 0$
  - 检验该假设只需给出 θ 的区间估计即可判断
- 贝叶斯方法关注更广泛而复杂的问题：模型的比较
- 比较规范的方法是：**贝叶斯因子**

3.2 贝叶斯因子的概念

贝叶斯检验

假设检验问题可以归结为两个模型的比较：

$$H_0: \text{数据来自模型} M_0；\ H_1: \text{数据来自模型} M_1$$

贝叶斯定理：

$$P(M_j|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|M_j)P(M_j)}{p(\boldsymbol{y})}$$

其中

- $P(M_j|\boldsymbol{y})$ 是模型 j 的后验概率 (posterior probability of model j)
- $P(\boldsymbol{y}|M_j)$ 是在模型 j 下的边缘似然（密度）函数 (marginal likelihood under model j)
- $P(M_j)$ 是模型 j 的先验概率
- $p(\boldsymbol{y})$ 是边缘似然函数或先验预测分布，

$$p(\boldsymbol{y}) = \sum_{j=0}^{J} P(M_j)P(\boldsymbol{y}|M_j)$$

贝叶斯因子

定义 5. *模型偏向于 M_1 的*贝叶斯因子 *(Bayes Factor) 定义为后验机会 (Posterior Odds) 与先验机会 (Prior Odds) 之比：*

$$B_{10} = \frac{P(M_1|\boldsymbol{y})/P(M_0|\boldsymbol{y})}{P(M_1)/P(M_0)} = \frac{\text{Posterior Odds}}{\text{Prior Odds}}$$

含义解释：贝叶斯因子是得到观察数据后，对先验机会比的一个修正系数，即

$$\frac{P(M_1|\boldsymbol{y})}{P(M_0|\boldsymbol{y})} = \frac{P(M_1)}{P(M_0)} \times B_{10}$$

贝叶斯因子的假设检验

优点：

1. 贝叶斯方法为模型的假设检验问题提供统一的处理方式
2. 贝叶斯因子反映了两个假设支持强度的比率。相反，p-值大小没有实际意义。
3. 应用范围更广泛，如

- 对未知参数的假设：$H_0: \mu = \mu_0$ 可看作模型 $M_0: y = \mu_0 + \varepsilon, \varepsilon \sim N(0, \sigma^2)$
- 对模型的检验：
  - 允许嵌套模型：M_0 包括所有解释变量，M_1 删除部分解释变量

  - 也允许非嵌套模型：
    * M_0 误差项为正态，M_1 误差项为 t-分布；
    * M_0 为 logit 模型，M_1 为 probit 模型等

缺点：

1. 对先验分布敏感。即使在大样本情况下，先验分布对后验分布没什么影响，但对贝叶斯因子可能影响很大
2. 非正常先验 (Improper prior) 下，先验机会比无意义

贝叶斯因子计算公式

两个贝叶斯公式对比，得到

$$\frac{P(\boldsymbol{y}|M_1)P(M_1)}{P(\boldsymbol{y}|M_0)P(M_0)} = \frac{P(M_1|\boldsymbol{y})}{P(M_0|\boldsymbol{y})}$$

因此，贝叶斯因子

$$B_{10} = \frac{P(\boldsymbol{y}|M_1)}{P(\boldsymbol{y}|M_0)}$$

其中 $P(\boldsymbol{y}|M_1)$ 为在模型 M_1 下 $\boldsymbol{y}$ 的**边缘密度（似然）函数**：

$$P(y|M_1) = \int \pi(\theta_1|M_1) f(y|\theta_1, M_1) d\theta_1$$

其中 $\pi(\theta_1|M_1)$ 是模型 M_1 下的先验 pdf，$f(\boldsymbol{y}|\theta_1, M_1)$ 是模型 M_1 下的似然函数。

- 如果 $M_0 : \theta = \theta_0, M_1 : \theta = \theta_1$ 各只有一个点，则贝叶斯因子等于似然比统计量。

例：离散型参数的贝叶斯因子

两位政治家争论民众是否支持死刑，约翰认为 40% 的人支持，而琼斯认为 60% 的人支持。现在随机调查 1000 人，结果有 490 人投票支持死刑。

贝叶斯因子为：

$$B_{10} = \frac{P(\boldsymbol{y}|M_1)}{P(\boldsymbol{y}|M_0)} = \frac{\binom{1000}{490}\theta_1^{490}(1-\theta_1)^{1000-490}}{\binom{1000}{490}\theta_0^{490}(1-\theta_0)^{1000-490}}$$

例：离散型参数的贝叶斯因子

```
# compute Bayes factor for John vs. Jones
BF <-
  dbinom(
    x = 490,
    size = 1000,
    prob = 0.4 #John's hypothesis
  ) / dbinom(
    x = 490,
    size = 1000,
    prob = 0.6 #Jones' hypothesis
  )
paste("Bayes factor = ", round(bf,3))
[1] "Bayes factor =  3325.257"
# 把 John 的先验概率改为 0.41，则 BF=113215.911！
```

贝叶斯因子判断标准（Jeffreys，1961）

$\log_{10}(B_{10})$	贝叶斯因子 (B_{10})	拒绝 H_0
<0	<1	Negative
0~0.5	1~3.2	Weak
0.5~1	3.2~10	Substantial
1~2	10~100	Strong
>2	>100	Decisive

贝叶斯因子判断标准（Raftery，1995）

BIC difference	贝叶斯因子 (B_{10})	$P(M_1\|D)(\%)$	拒绝 H_0
0-2	1-3	50-75	Weak
2-6	3-20	75-95	Positive
6-10	20-150	95-99	Strong
>10	>150	>99	Very Strong

Raftery(1995)：BIC 之差近似于贝叶斯因子：$\mathrm{BIC}_1 - \mathrm{BIC}_0 \approx -2\log B_{10}$

3.3 贝叶斯因子的计算

贝叶斯因子软件包：BayesFactor

ttestBF Bayes factors for one- and two- sample designs

anovaBF Bayes factors comparing many ANOVA models

regressionBF Bayes factors comparing many linear regression models

generalTestBF Bayes factors for all restrictions on a full model (0.9.4+)

lmBF Bayes factors for specific linear models (ANOVA or regression)

correlationBF Bayes factors for linear correlations

proportionBF Bayes factors for tests of single proportions

posterior Sample from the posterior distribution of the numerator of a Bayes factor object

recompute Recompute a Bayes factor or MCMC chain, possibly increasing the precision of the estimate

compare Compare two models; typically used to compare two models in BayesFactor MCMC objects

例：比较安眠药的效果

随机抽取 10 个人，分别服用两种安眠药。

数据集 sleep 包含 3 个变量（配对数据）：

extra numeric, increase in hours of sleep

group factor, drug given

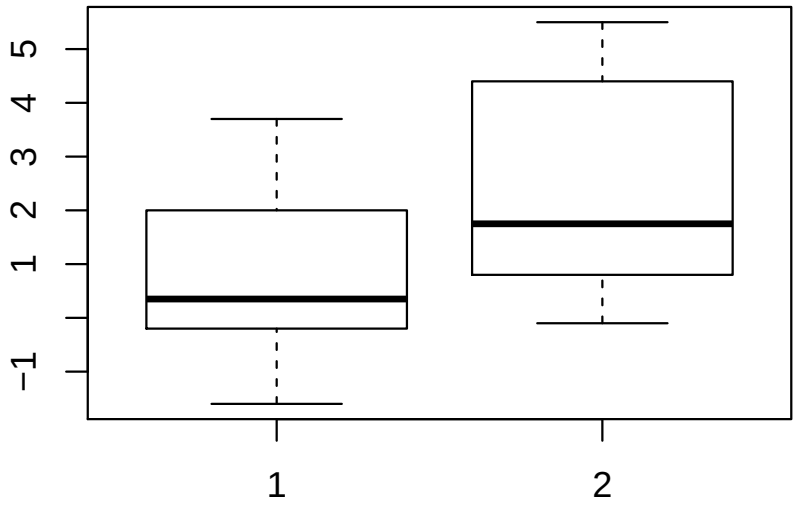
ID factor patient ID

试比较两种安眠药的效果是否有显著差异。

```
> head(sleep)
  extra group ID
1   0.7     1  1
2  -1.6     1  2
3  -0.2     1  3
```

例：Sleep 数据箱线图

```
plot(extra ~ group, data = sleep)
```



例：传统假设检验比较

$H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$

```
>data(sleep)
>diff = sleep$extra[1:10] - sleep$extra[11:20]
>t.test(diff)
One Sample t-test
data:  diffScores
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -2.4598858 -0.7001142
sample estimates:
mean of x
    -1.58
```

结论：?

用 BayesFactor 进行贝叶斯因子比较

假设 样本$x_1, x_2, \ldots, x_n$ 来自总体 $N(\mu, \sigma^2)$，需要检验 $H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$。

记标准化效应为

$$\delta = \frac{\mu - \mu_0}{\sigma}$$

如果是双样本，则

$$\delta = \frac{\mu_2 - \mu_1}{\sigma}$$

先验分布：

- σ^2 用 Jeffreys 先验，$\sigma^2 \propto \sigma^{-2}$
- $\delta \sim \text{Cauchy}(\text{location} = 0, \text{scale} = \text{r})$（r 一般取 $1/\sqrt{2}$，1，和 $\sqrt{2}$）

例：用 ttestBF 进行单样本双侧检验

$H_0 : \delta = 0, H_1 : \delta \neq 0$

```
ttestBF(x = diff)
Bayes factor analysis
--------------
[1] Alt., r=0.707 : 17.25888 ±0%
Against denominator:
  Null, mu = 0
---
Bayes factor type: BFoneSample, JZS
```

- 结论：BF=17.26，表示以较强（Strong）的证据拒绝原假设（与传统 t 检验结论一致）
- 如果贝叶斯因子与 p-值结论矛盾怎么办？

例：单样本单侧检验

$H_0 : \delta = 0, H_1 : \delta < 0$

```
> bf <- ttestBF(x = diff, nullInterval = c(-Inf, 0))
> bf
Bayes factor analysis
--------------
[1] Alt., r=0.707 -Inf<d<0    : 34.41694  ±0%
[2] Alt., r=0.707 !(-Inf<d<0) : 0.1008246 ±0.06%
Against denominator:
  Null, mu = 0
---
Bayes factor type: BFoneSample, JZS
```

结论：?

例：单样本复合单侧检验

$H_0 : \delta < 0, H_1 : \delta \geq 0$

```
> bf[1]/bf[2]
Bayes factor analysis
--------------
[1] Alt., r=0.707 -Inf<d<0 : 341.3547 ±0.06%
Against denominator:
  Alternative, r = 0.707106781186548, mu =/= 0 !(-Inf<d<0)
---
Bayes factor type: BFoneSample, JZS
```

结论：支持 H_0 的强度是 H_1 的 341 倍。

3.4 贝叶斯因子模型比较：回归模型

回归模型的贝叶斯因子比较

软件包 BayesFactor 中函数 regressionBF 计算线性回归模型所有自变量组合的贝叶斯因子，检验的原假设是某斜率等于 0，对立假设是所有斜率均不为 0.

贝叶斯模型：

$$\boldsymbol{y} \sim \text{Normal}(\alpha 1 + \beta \boldsymbol{X}, \sigma^2 I)$$

先验分布：$(\alpha, \sigma^2) \propto \sigma^{-2}$，$\beta \propto \text{Normal}(0, Ng\sigma^2 \left(\boldsymbol{X}^{'} \boldsymbol{X}\right)^{-1})$，其中 $g \sim \text{IG}(1/2, r/2)$，$r$ 取值为 $\sqrt{2}/4, 1/2$，及 $\sqrt{2}/2$。

例：工资模型的 BF 比较

```
>bf<-regressionBF(lwage-educ+exper+female+married,
                  data = wage_data)
>length(bf)
  [1] 15
>bf["educ + exper + female + married"] # 某指定模型的贝叶斯因子
Bayes factor analysis
--------------
[1] educ + exper + female + married : 3.06372e+105 ±0.01%
Against denominator:
  Intercept only
---
Bayes factor type: BFlinearModel, JZS
```

例：所有模型的 **BF** 比较

```
>bf
Bayes factor analysis
--------------
[1] educ                              : 4.783669e+18  ±0%
[2] exper                             : 3.333343e+25  ±0%
[3] female                            : 2.841551e+56  ±0%
[4] married                           : 137980603775  ±0%
[5] educ + exper                      : 6.8911e+58    ±0%
[6] educ + female                     : 7.76965e+80   ±0%
[7] educ + married                    : 3.964887e+32  ±0.01%
[8] exper + female                    : 4.625953e+69  ±0%
[9] exper + married                   : 1.214065e+30  ±0.01%
[10] female + married                 : 1.37707e+57   ±0%
[11] educ + exper + female            : 1.001935e+106 ±0%
[12] educ + exper + married           : 1.532133e+64  ±0.01%
[13] educ + female + married          : 3.887306e+82  ±0%
[14] exper + female + married         : 9.536846e+68  ±0.01%
[15] educ + exper + female + married : 3.06372e+105  ±0.01%
Against denominator:
  Intercept only
---
```

例：**BF** 值排在前 **6** 的模型

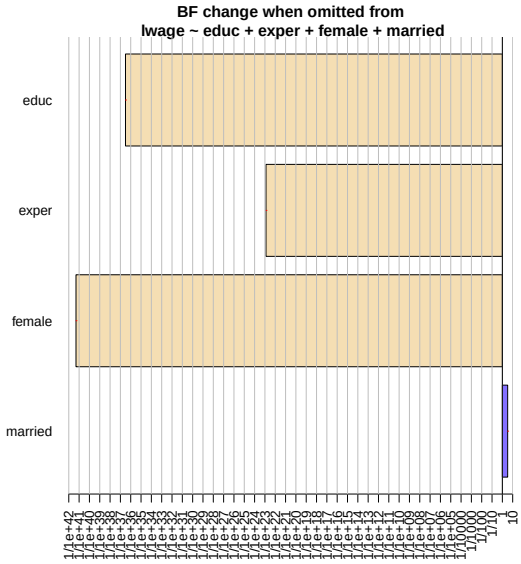
```
>head(bf)
Bayes factor analysis
--------------
[1] educ + exper + female           : 1.001935e+106 ±0%
[2] educ + exper + female + married : 3.06372e+105  ±0.01%
[3] educ + female + married         : 3.887306e+82  ±0%
[4] educ + female                   : 7.76965e+80   ±0%
[5] exper + female                  : 4.625953e+69  ±0%
[6] exper + female + married        : 9.536846e+68  ±0.01%
Against denominator:
  Intercept only
---
Bayes factor type: BFlinearModel, JZS
```

例：排前 **6** 的模型与 **BF** 最大模型的比较

```
>which.max(bf)
educ + exper + female  11
>bf2 = head(bf) / max(bf)
>bf2
Bayes factor analysis
--------------
[1] educ + exper + female           : 1            ±0%
[2] educ + exper + female + married : 0.3057803    ±0.01%
[3] educ + female + married         : 3.879799e-24 ±0%
[4] educ + female                   : 7.754645e-26 ±0%
[5] exper + female                  : 4.617019e-37 ±0%
[6] exper + female + married        : 9.518428e-38 ±0.01%
Against denominator:
  lwage ~ educ + exper + female
---
```

例：删去解释变量时 **BF** 的改变量（图示）

```
>plot(bf)
```



例：用 lmBF() 计算指定模型的 BF

```
best <- lmBF(lwage ~ educ + exper + female,
             data = wage_data)
best
Bayes factor analysis
--------------
[1] educ + exper + female : 1.001935e+106 ±0%
Against denominator:
  Intercept only
---
Bayes factor type: BFlinearModel, JZS
```

例：用 lmBF() 抽取指定模型的后验样本

```
>chains = posterior(best, iterations = 10000)
>summary(chains)
Iterations = 1:10000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000
1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:
           Mean       SD  Naive SE Time-series SE
mu      1.65875 0.013836 1.384e-04      1.384e-04
educ    0.07231 0.005300 5.300e-05      5.300e-05
exper   0.01356 0.001208 1.208e-05      1.208e-05
female -0.46329 0.030237 3.024e-04      3.042e-04
sig2    0.23699 0.009549 9.549e-05      9.408e-05
g       0.32573 0.919790 9.198e-03      9.198e-03
2. Quantiles for each variable:
           2.5%      25%      50%      75%    97.5%
mu      1.63146  1.64946  1.65880  1.66826  1.68568
educ    0.06182  0.06880  0.07235  0.07587  0.08268
exper   0.01117  0.01274  0.01356  0.01437  0.01590
female -0.52205 -0.48361 -0.46333 -0.44297 -0.40375
sig2    0.21899  0.23038  0.23669  0.24319  0.25639
g       0.05483  0.11658  0.18708  0.32886  1.41079
```

本章小结

模型检查 (Model Checking)

- 模型的敏感性
- 用后验预测分布进行模型检查 (Posterior Predictive Checking, PPC)
- 用 Bayesplot 进行 PPC 的计算和图示

模型比较 (Model Comparison)

1. 基于交叉验证：软件包 loo
2. 基于信息准则：
    - AIC、BIC：通过 lm() 的 AIC()、BIC() 计算
    - DIC、WAIC：通过 loo 的 waic() 计算
3. 贝叶斯假设检验及贝叶斯因子
    - 贝叶斯因子的概念、优点和缺点
    - 贝叶斯因子的计算：基于软件包 BayesFactor