Bayesian Linear Regression

– US air pollution data

庄亮亮

目录


```
#devtools::install_github("julianfaraway/brinla") #可能要翻墙
library(INLA)
library(brinla)
```

```
library(ggplot2)
library(GGally)
library(tidyr)
library(MASS)
library(nlme)
```

1. Bayesian inference for linear regression

空气污染数据：调查美国 41 个城市污染的决定因素。以 SO2 作为因变量，其他六个变量作为潜在解释变量。

在这些潜在解释变量中，有两个与人类生态相关 (pop,manuf)，另外四个与气候相关 (negtemp,wind,precip,days)。变量 negtemp 表示年平均气温的负值。在这里使用负值是因为所有的变量都是这样的，高的值表示一个不太事宜的环境。

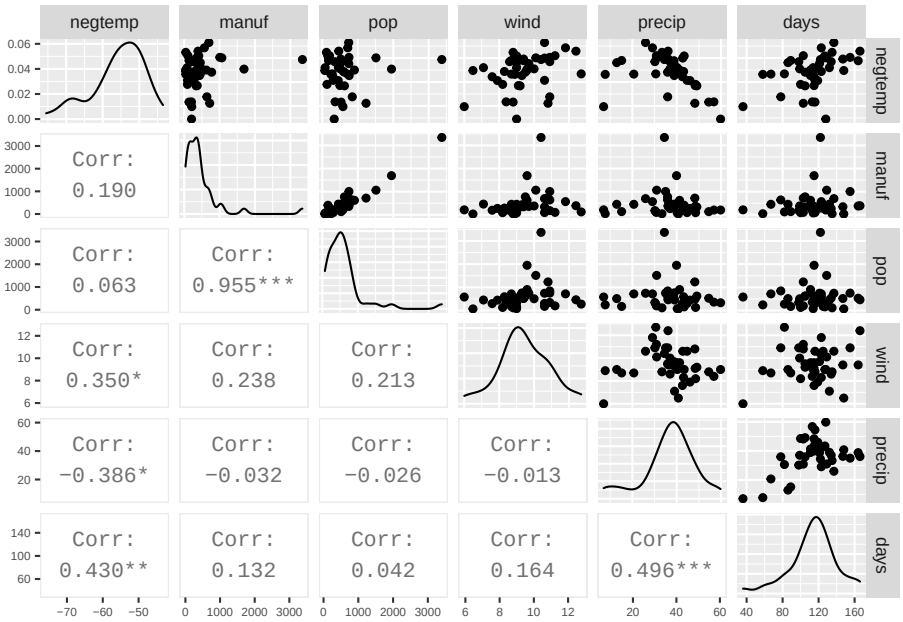
```
data(usair, package = "brinla")
knitr::kable(head(usair),
             caption = 'usair数据',
             align='c')
```

表 1: usair 数据

	SO2	negtemp	manuf	pop	wind	precip	days
Phoenix	10	-70.3	213	582	6.0	7.05	36
Little Rock	13	-61.0	91	132	8.2	48.52	100
San Francisco	12	-56.7	453	716	8.7	20.66	67
Denver	17	-51.9	454	515	9.0	12.95	86
Hartford	56	-49.1	412	158	9.0	43.37	127
Wilmington	36	-54.0	80	80	9.0	40.25	114

```
pairs.chart <- ggpairs(usair[,-1],
    lower = list(continuous = "cor"),
```

```
    upper = list(continuous = "points", combo = "dot")) +
    ggplot2::theme(axis.text = element_text(size = 6))
pairs.chart
```



Manuf 与 Pop 高度相关:r=0.955；我们在模型中只需保留一个。

1.1 Frequentist approach

```
usair.formula1 <-  SO2 ~ negtemp + manuf + wind + precip + days
usair.lm1 <- lm(usair.formula1, data = usair)
knitr::kable(round(coef(summary(usair.lm1)), 4),
             caption = '拟合情况汇总',
             align='c')
```

表 2: 拟合情况汇总

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	135.7714	50.0610	2.7121	0.0103

	Estimate	Std. Error	t value	Pr(>|t|)
negtemp	1.7714	0.6366	2.7824	0.0086
manuf	0.0256	0.0046	5.5544	0.0000
wind	-3.7379	1.9444	-1.9224	0.0627
precip	0.6259	0.3885	1.6111	0.1161
days	-0.0571	0.1748	-0.3265	0.7460

结果表明，`negtemp` 和 `manuf` 是重要的解释变量，而 `wind`,`precip` 和 `days` 不是。

- 标准残差

```
round(summary(usair.lm1)$sigma, 4)
```

```
## [1] 15.79
```

1.2 Bayesian Regression with INLA

默认情况下：

$$\beta_j \sim N\left(0, 10^6\right), \quad j = 0, \ldots, p$$

$$\log(\tau) \sim \log \text{Gamma}\left(1, 10^{-5}\right)$$

```
usair.inla1 <- inla(usair.formula1, data = usair,
    control.compute = list(dic = TRUE, cpo = TRUE))
```

```
knitr::kable(round(usair.inla1$summary.fixed, 4),
             caption = '固定效应信息',
             align='c')
```

表 3: 固定效应信息

	mean	sd	0.025quant	0.5quant	0.975quant	mode	kld
(Intercept)	135.4904	49.8882	36.9784	135.4963	233.8821	135.5120	0
negtemp	1.7690	0.6347	0.5157	1.7690	3.0209	1.7692	0
manuf	0.0256	0.0046	0.0165	0.0256	0.0346	0.0256	0
wind	-3.7230	1.9357	-7.5432	-3.7234	0.0964	-3.7241	0
precip	0.6249	0.3874	-0.1401	0.6249	1.3890	0.6249	0
days	-0.0567	0.1743	-0.4007	-0.0567	0.2872	-0.0567	0

```
knitr::kable(round(usair.inla1$summary.hyperpar, 4),
             caption = '超参数信息',
             align='c')
```

表 4: 超参数信息

	mean	sd	0.025quant	0.5quant	0.975quant	mode
Precision for the Gaussian observations	0.0042	0.001	0.0025	0.0042	0.0064	0.004

```
summary(usair.inla1)
```

```
##
## Call:
##    c("inla(formula = usair.formula1, data = usair, control.compute =
##    list(dic = TRUE, ", " cpo = TRUE))")
## Time used:
##     Pre = 1.12, Running = 0.428, Post = 0.119, Total = 1.67
## Fixed effects:
##                mean     sd 0.025quant 0.5quant 0.975quant    mode kld
## (Intercept) 135.490 49.888     36.978  135.496    233.882 135.512   0
## negtemp       1.769  0.635      0.516    1.769      3.021   1.769   0
## manuf         0.026  0.005      0.017    0.026      0.035   0.026   0
```

```
## wind         -3.723  1.936     -7.543   -3.723      0.096  -3.724   0
## precip        0.625  0.387     -0.140    0.625      1.389   0.625   0
## days         -0.057  0.174     -0.401   -0.057      0.287  -0.057   0
##
## Model hyperparameters:
##                                           mean    sd 0.025quant 0.5quant
## Precision for the Gaussian observations 0.004 0.001      0.003    0.004
##                                         0.975quant  mode
## Precision for the Gaussian observations      0.006 0.004
##
## Expected number of effective parameters(stdev): 6.00(0.001)
## Number of equivalent replicates : 6.84
##
## Deviance Information Criterion (DIC) ...............: 350.76
## Deviance Information Criterion (DIC, saturated) ....: 51.32
## Effective number of parameters .....................: 7.21
##
## Marginal log-Likelihood:  -208.73
## CPO and PIT are computed
##
## Posterior marginals for the linear predictor and
##  the fitted values are computed
```

- 计算 σ 的后验估计值

默认情况下，inla 对象输出的是后验信息的精确参数 τ。然而常常我们对后验均值 σ 感兴趣。

bri.hyperpar.summary 用于生成超参数 σ 的汇总统计信息。

```
knitr::kable(round(bri.hyperpar.summary(usair.inla1), 4),
             caption = '超参数信息',
             align='c')
```

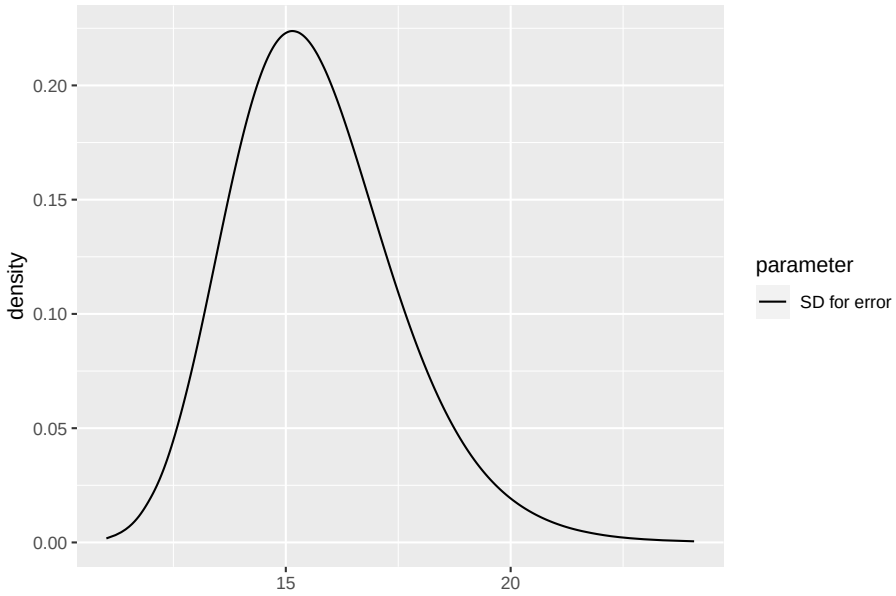
表 5: 超参数信息

	mean	sd	q0.025	q0.5	q0.975	mode
SD for the Gaussian observations	15.67	1.865	12.53	15.49	19.84	15.14

- 绘制 σ 的后验密度

  使用函数 `bri.hyperpar.plot`。

```
bri.hyperpar.plot(usair.inla1)
```



  σ 有略微右偏的后验分布。

- 改变先验

  假设 $\beta_0 \sim N(100, 100)$, $\beta_{negtemp} \sim N(2, 1)$ 和 $\beta_{wind} \sim N(3, 1)$。

  如果想假设 τ 服从对数正态分布 (对数 τ 服从正态分布)，可以使用选项 `control.family` 来指定。

```
usair.inla2 <- inla(usair.formula1, data = usair, control.compute = list(dic = TRUE, cp
    control.fixed = list(mean.intercept = 100, prec.intercept = 10^(-2),
        mean = list(negtemp = 2, wind = -3, default =0), prec = 1),
    control.family = list(hyper = list(prec = list(prior="gaussian", param =c(0,1)))))
summary(usair.inla2)
```

```
##
## Call:
##    c("inla(formula = usair.formula1, data = usair, control.compute =
##    list(dic = TRUE, ", " cpo = TRUE), control.family = list(hyper =
##    list(prec = list(prior = \"gaussian\", ", " param = c(0, 1)))),
##    control.fixed = list(mean.intercept = 100, ", " prec.intercept =
##    10^(-2), mean = list(negtemp = 2, wind = -3, ", " default = 0), prec =
##    1))")
## Time used:
##     Pre = 0.616, Running = 0.475, Post = 0.12, Total = 1.21
## Fixed effects:
##                mean    sd 0.025quant 0.5quant 0.975quant    mode kld
## (Intercept) 102.395 9.562     83.615  102.397    121.149 102.401   0
## negtemp       1.384 0.212      0.966    1.385      1.801   1.385   0
## manuf         0.025 0.004      0.018    0.025      0.033   0.025   0
## wind         -2.952 0.803     -4.529   -2.952     -1.376  -2.952   0
## precip        0.441 0.246     -0.045    0.441      0.925   0.442   0
## days          0.041 0.090     -0.136    0.040      0.218   0.040   0
##
## Model hyperparameters:
##                                           mean    sd 0.025quant 0.5quant
## Precision for the Gaussian observations 0.005 0.001      0.003    0.005
##                                         0.975quant  mode
## Precision for the Gaussian observations      0.008 0.005
##
## Expected number of effective parameters(stdev): 4.33(0.074)
## Number of equivalent replicates : 9.48
```

```
##
## Deviance Information Criterion (DIC) ...............: 347.57
## Deviance Information Criterion (DIC, saturated) ....: 57.38
## Effective number of parameters .....................: 5.26
##
## Marginal log-Likelihood:  -197.50
## CPO and PIT are computed
##
## Posterior marginals for the linear predictor and
##  the fitted values are computed
```

---

2. Prediction

2.1 The prediction in lm

```r
# new observations
new.data <- data.frame(
  negtemp = c(-50, -60, -40),
  manuf = c(150, 100, 400),
  pop = c(200, 100, 300),
  wind = c(6, 7, 8),
  precip = c(10, 30, 20),
  days = c(20, 100, 40))
knitr::kable(head(new.data))
```

negtemp	manuf	pop	wind	precip	days
-50	150	200	6	10	20
-60	100	100	7	30	100
-40	400	300	8	20	40

```
predict(usair.lm1, new.data, se.fit = TRUE)
```

```
## $fit
##     1     2     3
## 33.73 18.95 55.48
##
## $se.fit
##      1      2      3
## 14.937  5.329 17.639
##
## $df
## [1] 35
##
## $residual.scale
## [1] 15.79
```

输出：预测向量（\$fit）、预测均值的标准误差向量（\$se.fit）、残差的自由度（\$df）和残差标准差（$residual.scale）。

2.2 The prediction in INLA

在 INLA 中，与 lm 的函数预测不同。预测可以作为模型拟合的一部分。由于预测和拟合一个有缺失数据的模型是一样的，我们需要设置响应变量“y[i]=NA”表示我们想要预测的值。

```
usair.combined <- rbind(usair, data.frame(SO2 = c(NA, NA, NA), new.data))
knitr::kable(tail(usair.combined))
```

	SO2	negtemp	manuf	pop	wind	precip	days
Seattle	29	-51.1	379	531	9.4	38.79	164
Charleston	31	-55.2	35	71	6.5	40.75	148
Milwaukee	16	-45.7	569	717	11.8	29.07	123
1	NA	-50.0	150	200	6.0	10.00	20
2	NA	-60.0	100	100	7.0	30.00	100
3	NA	-40.0	400	300	8.0	20.00	40

```
usair.link <- c(rep(NA, nrow(usair)), rep(1, nrow(new.data)))
tail(usair.link)
```

```
## [1] NA NA NA  1  1  1
```

```
usair.inla1.pred <- inla(usair.formula1, data = usair.combined, control.predictor = lis

knitr::kable(usair.inla1.pred$summary.fitted.values[(nrow(usair)+1):nrow(usair.combined
    caption = '拟合情况', align='c')
```

表 8: 拟合情况

	mean	sd	0.025quant	0.5quant	0.975quant	mode
fitted.Predictor.42	33.65	14.88	4.316	33.66	63.00	33.66
fitted.Predictor.43	18.93	5.31	8.461	18.93	29.40	18.93
fitted.Predictor.44	55.41	17.58	20.753	55.41	90.07	55.41

3. Model selection and checking

3.1 DIC

- AIC

```
usair.step <- stepAIC(usair.lm1, trace = FALSE)
usair.step$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## SO2 ~ negtemp + manuf + wind + precip + days
##
```

```
## Final Model:
## SO2 ~ negtemp + manuf + wind + precip
##
##
##     Step Df Deviance Resid. Df Resid. Dev   AIC
## 1                           35       8726 231.8
## 2 - days  1    26.57         36       8753 229.9
```

```
# Final multiple regression model
usair.formula2 <- SO2 ~ negtemp + manuf + wind + precip
usair.lm2 <- lm(usair.formula2, data = usair)
knitr::kable(round(coef(summary(usair.lm2)), 4))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	123.1183	31.2907	3.935	0.0004
negtemp	1.6114	0.4014	4.015	0.0003
manuf	0.0255	0.0045	5.615	0.0000
wind	-3.6302	1.8923	-1.918	0.0630
precip	0.5242	0.2294	2.285	0.0283

- DIC

在贝叶斯分析中，DIC 是 AIC 的推广，是最常用的贝叶斯模型比较方法之一，定义为拟合优度度量和模型复杂度度量的总和

```
usair.inla3 <- inla(usair.formula2, data = usair, control.compute = list(dic = TRUE, cp
c(usair.inla1$dic$dic, usair.inla3$dic$dic)
```

```
## [1] 350.8 348.7
```

最佳模型：所有子集回归中最小的 DIC。

3.2 Posterior Predictive Checking

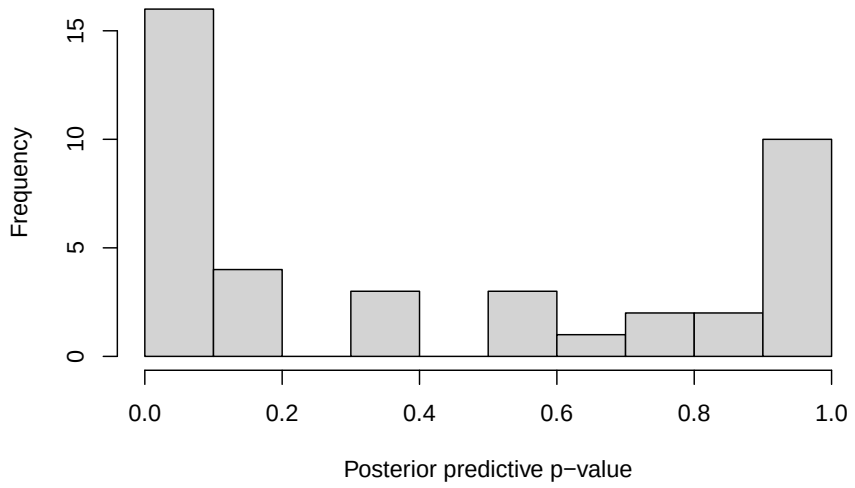
在贝叶斯分析中，模型评估通常是

1. 基于后验预测检查；
2. 留一交叉验证预测检查。

3.2.1 基于后验预测检查

后验预测 p 值可以通过 R 函数 `INLA .pmarginal` 得到

```
usair.inla3.pred <- inla(usair.formula2, data = usair, control.predictor = list(link =
post.predicted.pval <- vector(mode = "numeric", length = nrow(usair))
for(i in (1:nrow(usair))) {
  post.predicted.pval[i] <- inla.pmarginal(q=usair$SO2[i],
    marginal = usair.inla3.pred$marginals.fitted.values[[i]])
}
hist(post.predicted.pval, main="", breaks = 10, xlab="Posterior predictive p-value")
```



很多后验预测 p 值都接近于 0 或 1。然而，解释后验预测 p 值的一个**缺点**是，即使数据来自真实的模型，它们也不能具有均匀分布。因此，后验预测 p 值的图并不令人满意，我们希望使用其他模型评估方法进一步检验该模型。

3.2.2 留一交叉验证预测检查

评价模型的优劣的两个量：

1. conditional predictive ordinate (CPO)

$$CPO_i = p\left(y_i \mid y_{-i}\right)$$

2. probability integral transform (PIT)

$$\mathrm{PIT}_i = p\left(y_i^* \leq y_i \mid \mathbf{y}_{-i}\right)$$

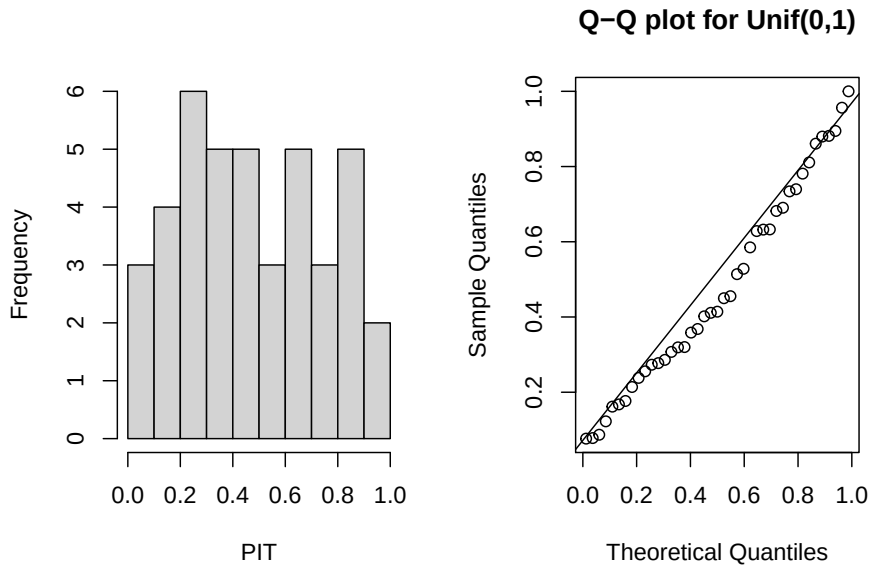
在 INLA 中有针对潜在问题的内部检查，这些问题出现在 `usair.inla3$cpo$failure` 中。它是一个向量，每个观测值都包含 0 或 1。当值为 1 时，表示 CPO 或 PIT 的估计对于相应的观测是不可靠的。在我们的例子中，我们可以通过以下方法检查是否存在故障：

```
sum(usair.inla3$cpo$failure)
```

```
## [1] 0
```

因此，在 `usair.inla3` 中不存在 CPOs 和 PITs 的计算问题。

```
par(mfrow = c(1, 2))
hist(usair.inla3$cpo$pit, main="", breaks = 10, xlab = "PIT")
qqplot(qunif(ppoints(length(usair.inla3$cpo$pit))), usair.inla3$cpo$pit, main = "Q-Q pl
    xlab = "Theoretical Quantiles", ylab = "Sample Quantiles")
qqline(usair.inla3$cpo$pit, distribution = function(p) qunif(p), prob = c(0.1, 0.9))
```



PITs 的分布接近均匀分布，表明模型对数据的拟合较为合理。值得注意的是，PITs 直方图比对应的后验预测直方图更接近均匀分布。

- Compare LPML for the full model and reduce model

如果我们把所有 CPO 值的乘积看作一个 “伪边际似然”，这就给出了一个交叉验证的拟合度度量。Geisser 和 Eddy(1979) 提出的对数拟边际似然 $(LPML)$：

$$LPML = \log\left\{\prod_{i=1}^{n} p\left(y_i \mid \mathbf{y}_{-i}\right)\right\} = \sum_{i=1}^{n} \log p\left(y_i \mid \mathbf{y}_{-i}\right) = \sum_{i=1}^{n} \log \mathrm{CPO}_i$$

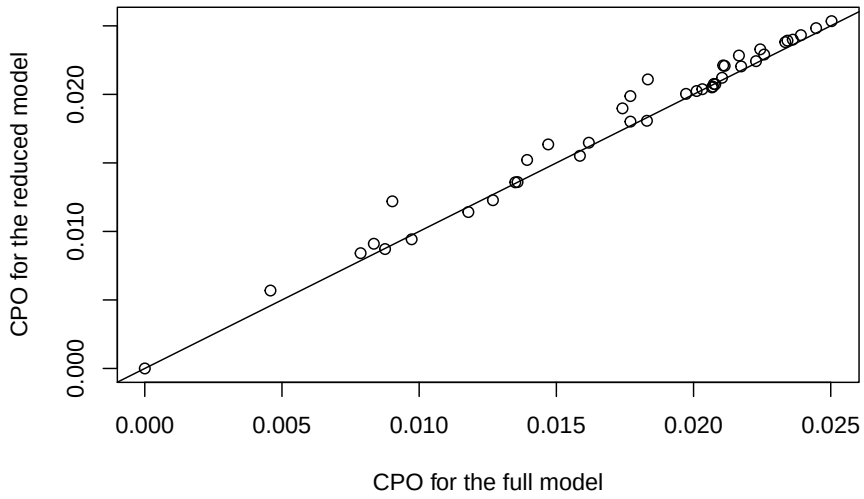
```
LPML1 <- sum(log(usair.inla1$cpo$cpo))
LPML3 <- sum(log(usair.inla3$cpo$cpo))
c(LPML1, LPML3)
```

```
## [1] -177.4 -176.1
```

简化模型的 LPML 比完整模型的 LPML 大，这表明简化模型是首选的。

- CPOs 的逐点比较

```
par(mfrow=c(1,1))
plot(usair.inla1$cpo$cpo, usair.inla3$cpo$cpo,
    xlab="CPO for the full model", ylab="CPO for the reduced model")
abline(0,1)
```

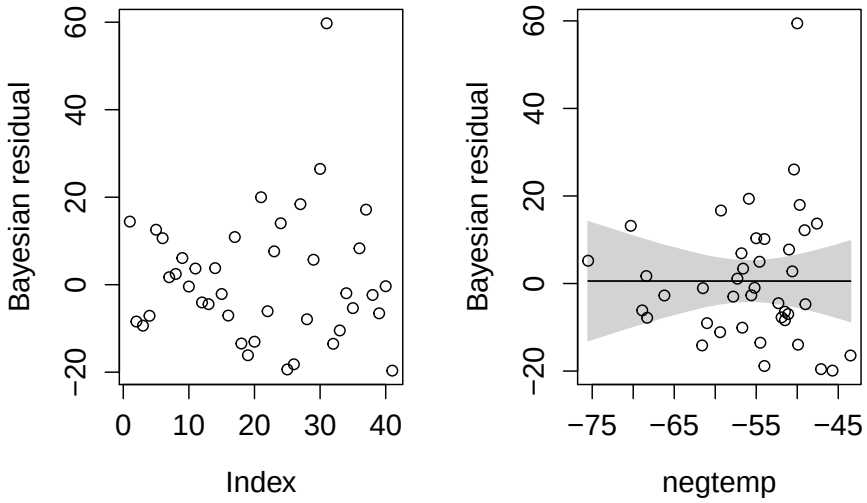


CPO 值表明模型拟合较好，参考线以上的点的优势意味着对简化模型的偏好。这与我们之前使用 DIC 和 LPML 标准的研究结果一致。

3.4 Bayesian residual plots

使用 `bri.lmresid.plot` 生成贝叶斯残差图:

```
par(mfrow=c(1,2))
bri.lmresid.plot(usair.inla3)
bri.lmresid.plot(usair.inla3, usair$negtemp,
    xlab = "negtemp", smooth =TRUE)
```



贝叶斯残差一般都在零附近呈现随机模式。但我们发现观测数 31 似乎是一个离群值，其贝叶斯残差高达 59.6927。

---

4. Robust Linear regression with t-distribution

在 INLA 中，通过指定 `family="T"` 来实现。

```
## Usair data example
usair.inla4 <- inla(usair.formula2, family = "T", data = usair,
    control.compute = list(dic = TRUE, cpo = TRUE))
# summary(usair.inla4)

knitr::kable(round(usair.inla4$summary.fixed, 4),
    caption = '固定效应情况', align='c')
```

表 10: 固定效应情况

	mean	sd	0.025quant	0.5quant	0.975quant	mode	kld
(Intercept)	118.4034	26.0576	67.4027	118.2333	170.3176	117.9043	0
negtemp	1.4297	0.3447	0.7729	1.4216	2.1329	1.4051	0
manuf	0.0267	0.0036	0.0192	0.0268	0.0335	0.0270	0
wind	-4.0148	1.5948	-7.1428	-4.0229	-0.8415	-4.0363	0
precip	0.4257	0.1909	0.0682	0.4186	0.8223	0.4038	0

```
knitr::kable(round(usair.inla4$summary.hyperpar, 4),
caption = '
    超参数情况', align='c')
```

表 11: 超参数情况

	mean	sd	0.025quant	0.5quant	0.975quant	mod
precision for the student-t observations	0.0053	0.0016	0.003	0.0051	0.0092	0.004
degrees of freedom for student-t	10.9806	8.7815	3.567	8.3861	34.1006	5.609

5. Analysis of Variance

下面的例子，我们将只关注**固定效应**的方差分析模型。关于随机效应模型，将在第 5 章详细介绍。

来自一项研究可待因和针灸对男性患者术后牙痛的影响的实验 (Kutner 等，2004 年)。

该研究采用随机区组设计，在一个因子结构中出现两个治疗因素。两种治疗因素都有两个层次。

反应变量缓解是疼痛缓解评分 (分数越高，患者缓解程度越好)。根据对疼痛耐受性的评估，32 名受试者被分配到 8 个区块，每个区块 4 名受试者。

```
# PainRelief data
data(painrelief, package = "brinla")

painrelief$PainLevel <- as.factor(painrelief$PainLevel)
painrelief$Codeine <- as.factor(painrelief$Codeine)
painrelief$Acupuncture <- as.factor(painrelief$Acupuncture)

knitr::kable(head(painrelief),
caption = 'painrelief数据', align='c')
```

表 12: painrelief 数据

PainLevel	Codeine	Acupuncture	Relief
1	1	1	0.0
1	2	1	0.5
1	1	2	0.6
1	2	2	1.2
2	1	1	0.3
2	2	1	0.6

```
painrelief.inla <- inla(Relief ~ PainLevel + Codeine*Acupuncture, data = painrelief)
#summary(painrelief.inla)
knitr::kable(round(painrelief.inla$summary.fixed, 4))
```

	mean	sd	0.025quant	0.5quant	0.975quant	mode	kld
(Intercept)	0.0188	0.0702	-0.1202	0.0188	0.1577	0.0188	1e-04
PainLevel2	0.1500	0.0846	-0.0176	0.1500	0.3175	0.1500	1e-04
PainLevel3	0.3250	0.0846	0.1574	0.3250	0.4925	0.3250	1e-04
PainLevel4	0.3000	0.0846	0.1324	0.3000	0.4675	0.3000	1e-04
PainLevel5	0.6750	0.0846	0.5074	0.6750	0.8425	0.6750	1e-04
PainLevel6	0.9750	0.0846	0.8074	0.9750	1.1425	0.9750	1e-04
PainLevel7	1.0750	0.0846	0.9074	1.0750	1.2425	1.0750	1e-04

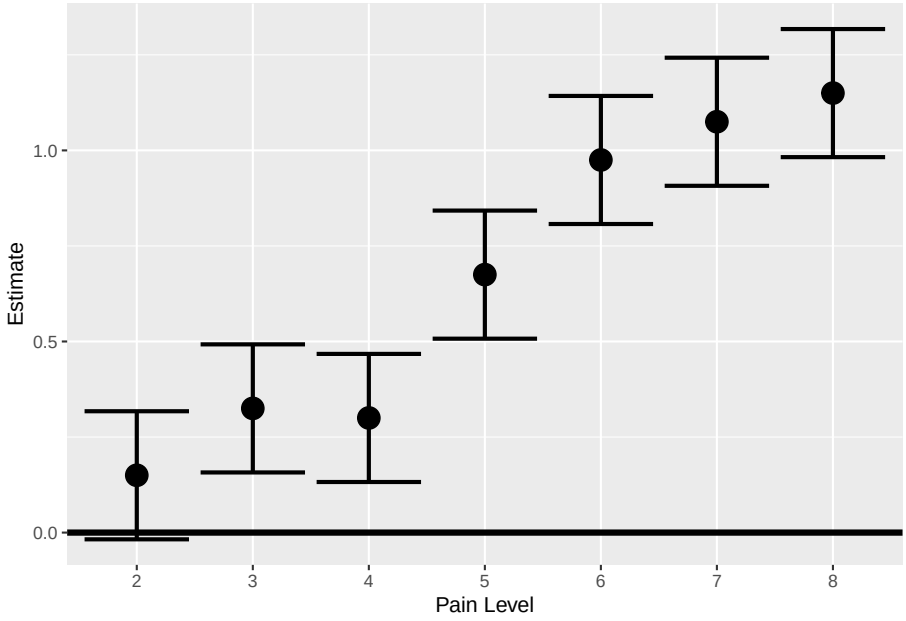
	mean	sd	0.025quant	0.5quant	0.975quant	mode	kld
PainLevel8	1.1500	0.0846	0.9824	1.1500	1.3175	1.1500	1e-04
Codeine2	0.4625	0.0598	0.3440	0.4625	0.5810	0.4625	1e-04
Acupuncture2	0.5750	0.0598	0.4565	0.5750	0.6935	0.5750	1e-04
Codeine2:Acupuncture2	0.1500	0.0846	-0.0176	0.1500	0.3175	0.1500	1e-04

治疗因素可待因和针灸的主要作用在贝叶斯意义上都是高度显著的。但在 95% 可信水平上，两者之间的交互作用不显著，说明两者之间不存在交互作用。

```
est1 <- data.frame(x =c("2", "3", "4", "5", "6", "7", "8"),
           Estimate=painrelief.inla$summary.fixed[c(2:8),1],
           L = painrelief.inla$summary.fixed[c(2:8),3],
         U = painrelief.inla$summary.fixed[c(2:8),5])
```

为了更好地理解影响的重要性，我们生成了不同疼痛水平的后验均值估计和 95% 可信水平的图。水平线是疼痛级别 1 的参考线，它被设置为 0。疼痛程度越高，受试者的疼痛缓解评分就越高 (疼痛程度 4 除外)。显然，疼痛程度是模型中需要考虑的一个显著的混杂因素。

```
p1 <- ggplot(est1, aes(x = x, y = Estimate)) + geom_point(size = 5) + geom_errorbar(aes
p1
```



6. Ridge regression

法国经济进口活动有关数据：因变量为进口（import）、国内生产（DOPROD）、股票形式（stock）和国内消费（CONSUM）。

- 样本相关性

```
data(frencheconomy, package = "brinla")
head(frencheconomy)
```

```
##   YEAR IMPORT DOPROD STOCK CONSUM
## 1   49   15.9  149.3   4.2  108.1
## 2   50   16.4  161.2   4.1  114.8
## 3   51   19.0  171.5   3.1  123.2
## 4   52   19.1  175.5   3.1  126.9
## 5   53   18.8  180.8   1.1  132.1
## 6   54   20.4  190.7   2.2  137.7
```

```
knitr::kable(round(cor(frencheconomy[,-1]),4), caption = '
    frencheconomy数据', align='c')
```

表 14: frencheconomy 数据

	IMPORT	DOPROD	STOCK	CONSUM
IMPORT	1.0000	0.9842	0.2659	0.9848
DOPROD	0.9842	1.0000	0.2154	0.9989
STOCK	0.2659	0.2154	1.0000	0.2137
CONSUM	0.9848	0.9989	0.2137	1.0000

- 数据标准化

贝叶斯岭回归假设所有预测因子的系数元素 $(\beta_1, \ldots, \beta_p)$ 都是从一个标准正态密度中提取的。

```
fe.scaled <- cbind(frencheconomy[, 1:2], scale(frencheconomy[, c(-1,-2)]))
```

```
# set priors
n <- nrow(frencheconomy)

fe.scaled$beta1 <- rep(1,n)
fe.scaled$beta2 <- rep(2,n)
fe.scaled$beta3 <- rep(3,n)
```

对于岭回归，参数的先验有一个共同的未知方差，要实现这一点，我们必须使用副本，并改变数据集

```
# this is the prior for the precision of beta
param.beta = list(prec = list(param = c(1.0e-3, 1.0e-3)))

formula.ridge  = IMPORT ~ f(beta1, DOPROD, model="iid", values = c(1,2,3), hyper = para
frencheconomy.ridge <- inla(formula.ridge, data=fe.scaled)
ridge.est <- rbind(frencheconomy.ridge$summary.fixed, frencheconomy.ridge$summary.rando
```

```
knitr::kable(round(ridge.est,4))
```

	mean	sd	0.025quant	0.5quant	0.975quant	mode	kld
(Intercept)	30.0778	0.5198	29.0449	30.0778	31.109	30.0778	0e+00
1	5.1131	5.4237	-6.7892	5.3431	15.668	5.6205	3e-04
2	0.7205	0.5451	-0.3618	0.7202	1.802	0.7199	0e+00
3	6.9769	5.4265	-3.5327	6.7312	18.923	6.4159	3e-04

- Comparing with Standard Bayesian Linear regression

```
formula <- IMPORT ~  DOPROD + STOCK + CONSUM
frencheconomy.inla <- inla(formula, data = fe.scaled, control.fixed = list(prec = 1.0e-
knitr::kable(round(frencheconomy.inla$summary.fixed, 4))
```

	mean	sd	0.025quant	0.5quant	0.975quant	mode	kld
(Intercept)	30.0778	0.5729	28.9378	30.0778	31.216	30.0778	0
DOPROD	3.0052	10.9358	-18.5325	2.9662	24.743	2.8961	0
STOCK	0.7197	0.6037	-0.4819	0.7198	1.919	0.7199	0
CONSUM	9.1322	10.9315	-12.6218	9.1707	30.635	9.2416	0

- Comparing with Ridge regression frequentist approach

```
reg2 <- lm.ridge(IMPORT ~  DOPROD + STOCK + CONSUM, data = fe.scaled, lambda = seq(0, 1

reg2.final <- lm.ridge(IMPORT ~  DOPROD + STOCK + CONSUM, data = fe.scaled, lambda = re
reg2.final
```

```
##           DOPROD   STOCK  CONSUM
## 30.0778  4.9560  0.7176  7.1669
```

7. Linear regression with autoregressive errors

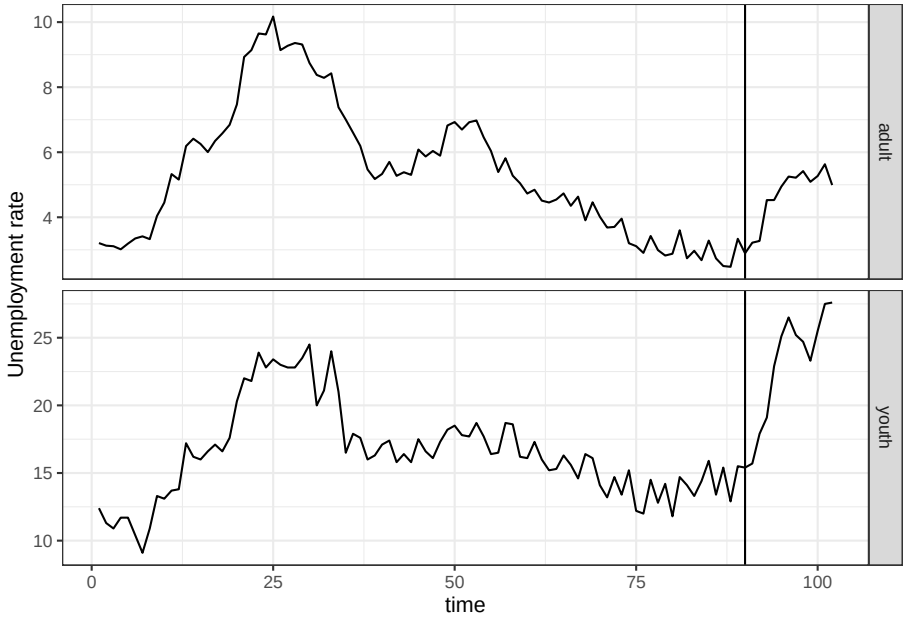
时间序列数据：新西兰的失业数据包括青年 (15-19 岁) 和成人 (19 岁以上) 的季度失业率。

自 2008 年 6 月起，新西兰政府废除了《最低工资法》。在此，我们想研究在该法案废除之前和之后，成年人和年轻人失业率之间的关系。

```
# Read the data
data(nzunemploy, package = "brinla")
nzunemploy$time <- 1:nrow(nzunemploy)
```

绘制成人和青年的时间序列数据

```
#library(tidyr)
qplot(time, value, data = gather(nzunemploy[,c(2,3,5)], variable, value, -time), geom =
```



为了使系数更容易理解，我们将成人失业率集中在时间序列的平均值上。

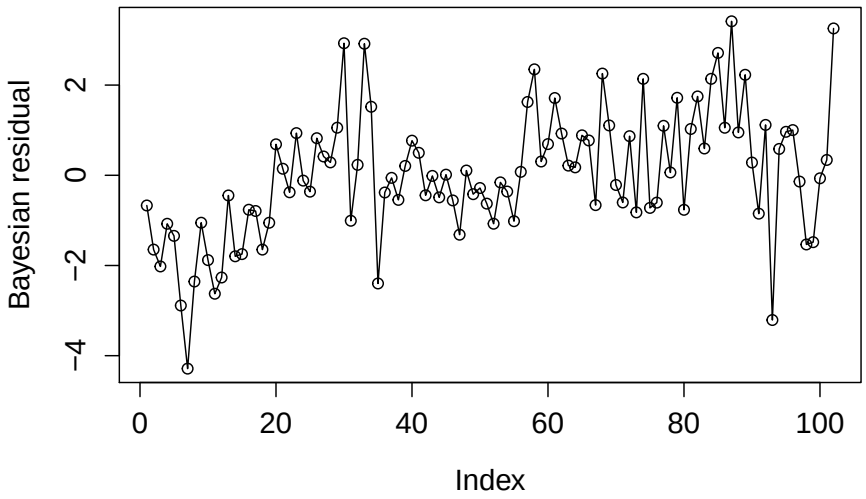
```
# Centering predictor
nzunemploy$centeredadult = with(nzunemploy, adult - mean(adult))
```

- 拟合一个具有独立误差的标准线性回归

```
formula1 <- youth ~ centeredadult*policy
nzunemploy.inla1 <- inla(formula1, data= nzunemploy)
# summary(nzunemploy.inla1)
round(nzunemploy.inla1$summary.fixed, 4)
```

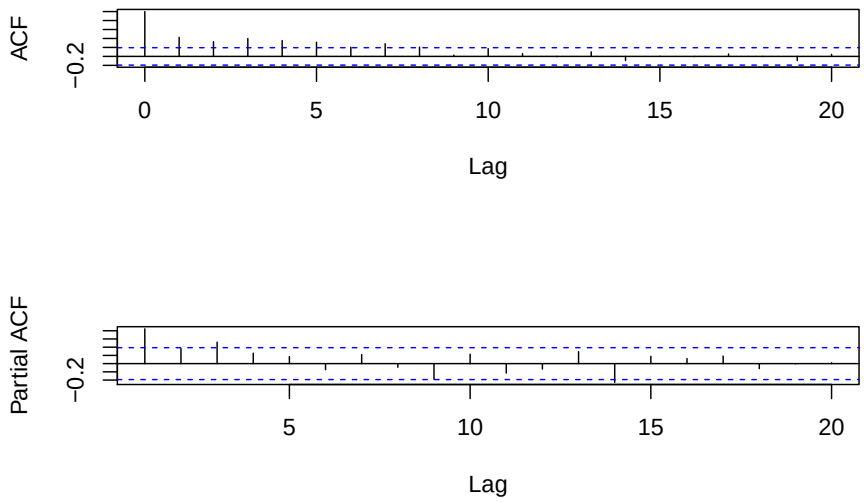
```
##                              mean     sd 0.025quant 0.5quant 0.975quant   mode
## (Intercept)                16.282 0.1534     15.980   16.282     16.584 16.282
## centeredadult               1.533 0.0750      1.386    1.533      1.681  1.533
## policyEqual                 9.442 0.5258      8.407    9.442     10.475  9.442
## centeredadult:policyEqual   2.853 0.4616      1.945    2.853      3.761  2.853
##                              kld
## (Intercept)                    0
## centeredadult                  0
## policyEqual                    0
## centeredadult:policyEqual      0
```

```
# Plot the Bayesian residuals
par(mfrow=c(1,1))
nzunemploy.res1 <- bri.lmresid.plot(nzunemploy.inla1, type="o")
```



它们之间存在一定程度的自相关。

```
# Plot the autocorrelation and partial autocorrelation
par(mfrow = c(2,1))
acf(nzunemploy.res1$resid, main = "")
acf(nzunemploy.res1$resid, type = "partial", main="")
```



图上的虚线对应 95% 置信带。自相关函数的形式呈指数衰减，而偏自相关函数的形式在滞后 1 处有很高的峰值。这些表明 AR(1) 过程将适合于回归模型中的误差项。

- 拟合一个有 AR(1) 误差的线性回归

```
formula2 <- youth ~ centeredadult*policy + f(time, model = "ar1")
nzunemploy.inla2 <- inla(formula2, data = nzunemploy, control.family = list(hyper = lis
# summary(nzunemploy.inla2)
knitr::kable(round(nzunemploy.inla2$summary.fixed, 4))
```

	mean	sd	0.025quant	0.5quant	0.975quant	mode	kld
(Intercept)	16.344	0.3029	15.768	16.334	16.975	16.322	5e-04
centeredadult	1.520	0.1354	1.246	1.521	1.785	1.524	1e-04
policyEqual	8.981	0.9724	6.876	9.036	10.746	9.117	8e-04
centeredadult:policyEqual	2.516	0.6009	1.302	2.525	3.672	2.543	1e-04

```
knitr::kable(round(nzunemploy.inla2$summary.hyperpar, 4))
```

	mean	sd	0.025quant	0.5quant	0.975quant	mode
Precision for time	0.4464	0.0859	0.293	0.4420	0.6282	0.4353
Rho for time	0.5128	0.0898	0.329	0.5151	0.6799	0.5163

注意，回归模型的精度 prec 固定在 $\tau = \exp(15)$

- 与频率派方法进行比对

```
#library(nlme)
nzunemploy.gls <- gls(youth ~ centeredadult*policy, correlation = corAR1(form=~1), data
summary(nzunemploy.gls)
```

```
## Generalized least squares fit by REML
##   Model: youth ~ centeredadult * policy
##   Data: nzunemploy
##   AIC   BIC logLik
##   353 368.5 -170.5
##
## Correlation Structure: AR(1)
##  Formula: ~1
##  Parameter estimate(s):
##    Phi
## 0.5012
##
## Coefficients:
##                              Value Std.Error t-value p-value
## (Intercept)                 16.329    0.2733   59.74       0
## centeredadult                1.522    0.1274   11.94       0
## policyEqual                  9.083    0.8614   10.54       0
## centeredadult:policyEqual    2.545    0.5772    4.41       0
##
```

```
##  Correlation:
##                              (Intr) cntrdd plcyEq
## centeredadult                -0.020
## policyEqual                  -0.318  0.007
## centeredadult:policyEqual -0.067 -0.155  0.583
##
## Standardized residuals:
##     Min      Q1     Med      Q3     Max
## -2.8923 -0.5546 -0.0242  0.5545  2.2957
##
## Residual standard error: 1.505
## Degrees of freedom: 102 total; 98 residual
```

```
# plot the fitted lines
ggplot(nzunemploy, aes(centeredadult, youth)) + geom_point(aes(shape = factor(policy)),
```