Национальный исследовательский университет «Высшая школа экономики»

Домашнее задание 2

Markov chain Monte Carlo version

Мерлов Андрей, группа 71ММАЭ

27.12.2011

Оглавление

Москва, 2011 г.

Общая информация

Выполнение каждого пункта задания структурировано следующим образом:

- Описание того, что происходит в данном пункте.
- Отрывок кода из программы, отвечающий за выполнение описанной задачи (если он содержит что-то интересное, а не рутинные операции).
- Графики, таблицы, замечания, итоги – если это необходимо.

Полный код программы и данные доступны по ссылкам: http://dl.dropbox.com/u/13224753/hse/MCMC/merlov_hw2_mcmc.R, http://dl.dropbox.com/u/13224753/hse/MCMC/merlov_hw2_data.csv. Для подсветки синтаксиса использовался Pretty R syntax highlighter (http://www.inside-r.org/pretty-r/tool).

Пункты задания

Номера пунктов в данном разделе соответствуют номерам пунктов в задании.

1. Выбор стартовых значений и гиперпараметров

В качестве стартовых значений параметров β и дисперсии ошибки σ_ε^2 возьмём средние значения оценок этих параметров по данным 2000-2004 годов. Затем вычислим параметры a, b априорного распределения по известным формулам. Эта работа проделана в рамках ДЗ 2, однако эти же операции воспроизводятся в прилагаемой программе (чтобы, с одной стороны, добиться самодостаточности программы, и, с другой стороны, проверить проделанные в ДЗ 2 расчёты).

```
data = read.csv("D:\\Dropbox\\public\\hse\\MCMC\\merlov_hw2_data.csv")

#
# 1. Выбираем что-то разумное в качестве значений sigma^2, bi и гиперпараметров
#
# Здесь сохраним коэффициенты и параметры точности регрессий по годам 2000-2004
# Порядок столбцов: sigma^2, b1,...,bk, h; строки - года.
store = matrix(NA, ncol=5, nrow=5)

for(i in 1:5) {
  reg = lm(SocCoh ~ Sec + QL, subset(data, Year==1999+i))
  sum = summary(reg)
  store[i,] = c(sum$sigma^2, reg$coefficients, 1/sum$sigma^2)
}

# Считаем средние и стандартные отклонения коэффициентов по годам
means = apply(store, 2, mean)
sd = apply(store, 2, sd)

starting_values = means[-5]
hyper_values = c((means[5]/sd[5])^2, means[5]/(sd[5])^2)
```

Значения на выходе не отличаются от полученных ранее. Они приводятся в следующей таблице.

Стартовые значения		**Значения гиперпараметров**	
Константа	1.045	a	2.472
Sec	0.411	b	0.854
QL	0.375		
Дисперсия ошибки	0.423		

2. Применение MCMC

Метод Монте-Карло по марковской цепи реализован в виде функции в скрипте-примере «2011_mcmc_regression.r». Воспользуемся этой функцией, выбрав число итераций равным 5000, для получения оценок по данным 2005 года.

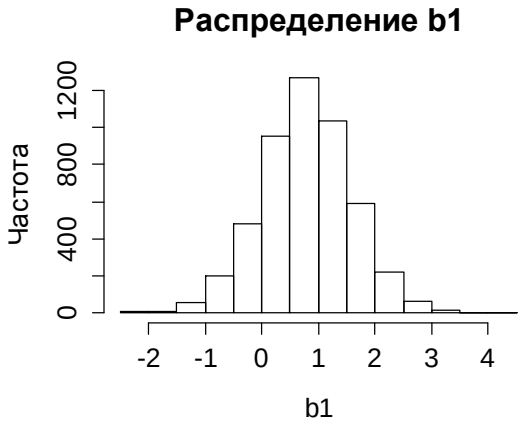
```
#
# 2. Применяем MCMC
#
# подготовим данные
data.2005 = subset(data, Year==2005)
nobs=nrow(data.2005)
x_data=cbind(c(rep(1,nobs)), data.2005$Sec, data.2005$QL)
y_data=data.2005$SocCoh
niterations = 5000
res=mcmc_regression(niterations,y_data,x_data,starting_values,hyper_values)
```

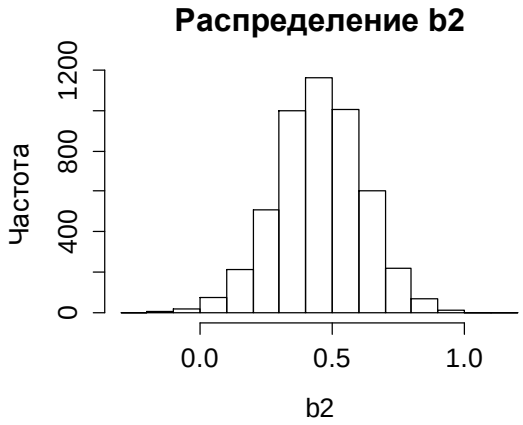
3. Гистограммы распределения для полученных оценок

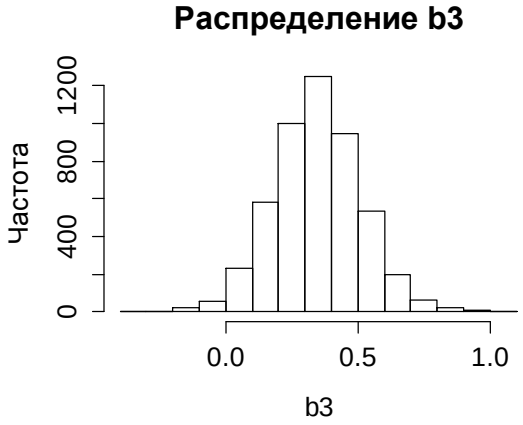
Посчитаем средние значения оценок, полученных на итерациях с 100 по 5000, и сравним с байесовскими оценками и оценками МП, полученными в ДЗ 2. Код в этом разделе не приводится, т.к. не представляет интереса.

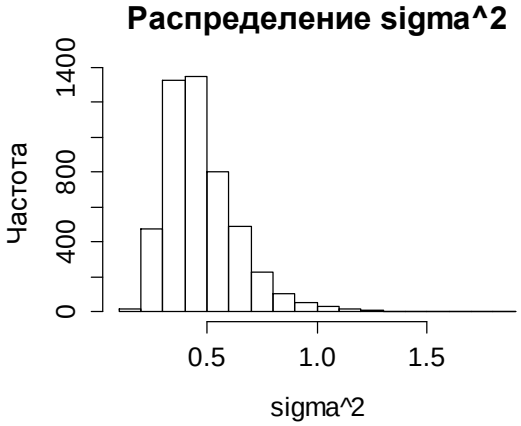
Переменная	MCMC оценки	Байесовские оценки	ММП оценки
Константа (b1)	0.768	0.990	0.791
Sec (b2)	0.462	0.412	0.456
QL (b3)	0.341	0.360	0.343
Дисперсия ошибки	0.468	0.342	0.455

Как видно из таблицы, оценки MCMC ближе к оценкам МП. Ниже приведены гистограммы распределения оценок MCMC.







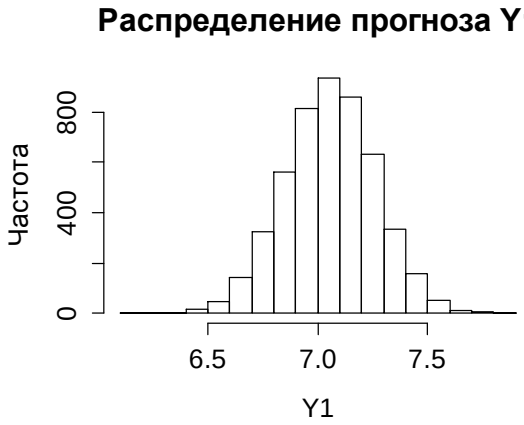


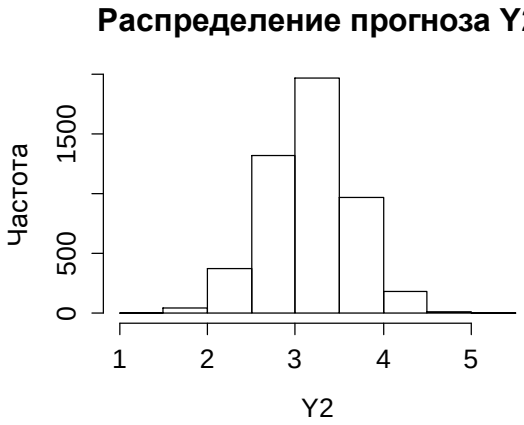
4. Гистограммы для прогнозных значений Y1, Y2

Прогноз строился для двух стран – для Ирландии (номер один) и для Польши (номер два). Сравним средние прогнозные значения с прогнозами, полученными в ДЗ 2.

Страна	MCMC	Прогноз из ДЗ 2
Ирландия	7.065	7.035
Польша	3.155	3.291

Как видно, прогнозные значения довольно близки. Ниже приведены гистограммы распределения для прогнозных значений Y1 (Ирландия) и Y2 (Польша).





5. Сравнение диаграммы рассеяния и доверительной области

Двумерная доверительная область с уровнем доверия α для прогнозных значений Y1, Y2 описывается следующим неравенством (уже подставлены конкретные цифры из ДЗ 2):

$$\frac{1}{2}\begin{pmatrix} Y1-7,03 \\ Y2-3,29 \end{pmatrix}^T \begin{pmatrix} 2,747 & -0,053 \\ -0,053 & 2,673 \end{pmatrix}\begin{pmatrix} Y1-7,03 \\ Y2-3,29 \end{pmatrix} < F_{1-\alpha}(2,15).$$

Совместим диаграмму рассеяния и контуры доверительных областей для следующих уровней доверия:

Уровень доверия	Значение F-статистики
90%	6.359
95%	3.682
99%	2.695

Ниже приведён код, реализующий построение контуров доверительных областей.

```
#
# 5. Диаграмма рассеяния для прогнозов, сравнение с графиком дов.области
#
# строим график доверительной области. Конкретные числа - из ДЗ2
ngrid = 100
B = cbind(c(2.747, -0.053), c(-0.053, 2.673))
cr.x = seq(5, 9, length.out = ngrid)
cr.y = seq(1, 5, length.out = ngrid)
cr.z = matrix(NA, ngrid, ngrid)
for (i in 1:ngrid) {
  for (j in 1:ngrid) {
    cr.v = c(cr.x[i] - 7.03, cr.y[j] - 3.29)
    cr.z[i,j] = 0.5*t(cr.v) %*% B %*% cr.v
  }
}

# изображаем на одном графике контур дов.области и диаграмму рассеяния
# показаны контуры ДО 99%, 95% и 90%
plot(forecast[100:niterations,], xlim=c(5.5,8.8), ylim=c(1.1,5.1),
     main="Диаграмма рассеяния и ДО", xlab="Y1", ylab="Y2")
contour(cr.x,cr.y,cr.z, levels=c(2.695, 3.682, 6.359), add=TRUE)

# конец.
```

Как видно из следующего рисунка, почти все точки $(Y1,Y2)$ попали в 95% доверительную область. Ни одна точка не вышла за пределы 99% доверительной области.