Examen 2. Métodos Estadísticos Avanzados

Antonio, H. F.

2022-11-27

Objetivo

Reproducir los resultados del Ejemplo 1 (Modelo I y II) de la sección 3 del artículo **Fitzmaurice, G. M. and N. M. Laird (1993). A likelihood - based for analysingg longitudinal binary responses. Biometrika 80 (1), 141 - 151**, utilizando:

1) El método propuesto en el artículo.
2) El enfoque bayesiano con el método computacional de su preferencia.

Datos

Los datos corresponden a un subconjunto tomado del estudio longitudinal de los efectos en la salud de la contaminación del aire reportados por **Ware et al., (1984)**. El subconjunto de datos contiene solo los registros de los niños que fueron analizados durante los 4 años del estudio, esto es 537 niños con cuatro registros, uno por año. La variable respuesta es binaria, 1 si se presenta silibancia, 0 de otro modo. Las covariables que pretenden explicar la probabilidad de silibancia son:

1) Mamá fumadora, codificada como 1 si fuma regularmente, 0 de otro modo.
2) Edad del niño, codificada como 0 cuando el niño tenia 9 años, 1 cuando tenia 10, -1 cuando tenia 8 y -2 cuando tenia 7.

Modelos

La variable respuesta es binaria, por lo tanto para modelar la $E(Y_i)$ se plantean varios escenarios:

1) Las Y_i son independientes y se puede realizar una regresión logistica. Desde un enfoque frequentista se obtienen los coeficientes de regresión o los valores de los parámetros de interés mediante mínimos cuadros ponderados. Desde un enfoque bayesiano, se obtienen las distribuciones marginales aposteriori de cada uno de los parámetros de interés mediante muestreo de Gibs.

2) Las Y_i son dependientes, es decir, que las cuatro mediciones de cada niño tienen una estructura de dependencia que se debe incluir en el modelo. Desde un enfoque frecuentista incluir la dependencia plantea un problema que es más complejo, pero existen métodos como quasi-verosimilitud, verosimilitud restringida, ecuaciones cuasi-score. Desde un enfoque bayesiano se puede resolver de una manera más directa con la aumentación de datos.

Enfoque frecuentista

1) Modelo I: Independencia entre las observaciones

Preparación de los datos

```
data <- read.csv("Ohio.csv",header=T)
data$resp <- as.factor(data$resp)
data$smoke <- as.factor(data$smoke)
```

Modelo

El modelo planteado es $E(Y_{ij}) = logit(\mu_{ij})$, $logit(\mu_{ij}) = \eta_{ij}$, $\eta_{ij} = \beta_0 + \beta_1 age_j + \beta_2 smoke_i + u_{ij}$, $i = 1, ..., 537$, $j = 1, 2, 3, 4$, y como la esperanza en un modelo binomial es igual a la probabilidad de éxito, entonces:

$$P(Y_{ij} = 1) = \frac{exp\{\eta_{ij}\}}{1 + exp\{\eta_{ij}\}}$$

y el término $u_{ij} \sim N(0, \sigma_e^2)$ contiene el efecto aleatorio de los niños, considerando que las observaciones entre cada niño son independientes, debido a esto, el problema se reduce a utilizar la verosimilitud completa y resolver mediante mínimos cuadrados ponderados iterativos.

```
model1 <- resp ~ age + smoke + age*smoke
m1 <- glm(model1, family = binomial(link="logit"),
          data = data)
summary(m1)
```

```
##
## Call:
## glm(formula = model1, family = binomial(link = "logit"), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6503  -0.6014  -0.5636  -0.4940   2.0804
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.90084    0.08874 -21.420   <2e-16 ***
## age         -0.14125    0.06951  -2.032   0.0422 *
## smoke1       0.31395    0.13944   2.252   0.0244 *
## age:smoke1   0.07084    0.11072   0.640   0.5223
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1829.1  on 2147  degrees of freedom
## Residual deviance: 1819.5  on 2144  degrees of freedom
## AIC: 1827.5
##
## Number of Fisher Scoring iterations: 4
```

2) Modelo II: Dependencia entre las observaciones

El modelo planteado es similar al modelo anterior, solamente agregregamos correlación entre las observaciones de cada niño, y planteamos que esta es la misma entre cada observación, por lo tanto el término $u_{ij} \sim N\left(0, \frac{\sigma_e^2}{(1-\rho^2)}\right)$ contiene el efecto aleatorio de los niños considerando la correlación entre las observaciones.

```
m2 <- gee(resp ~ age + smoke + age*smoke, id = id,
          data=data, family=binomial, corstr = "exchangeable")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27

## running glm to get initial regression estimate

## (Intercept)         age      smoke1  age:smoke1
##  -1.9008426  -0.1412531   0.3139540   0.0708441
```

```
summary(m2)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Logit
##  Variance to Mean Relation: Binomial
##  Correlation Structure:     Exchangeable
##
## Call:
## gee(formula = resp ~ age + smoke + age * smoke, id = id, data = data,
##     family = binomial, corstr = "exchangeable")
##
## Summary of Residuals:
##        Min         1Q     Median         3Q        Max
## -0.1906393 -0.1654776 -0.1468831 -0.1148906  0.8851094
##
##
## Coefficients:
##                Estimate Naive S.E.     Naive z Robust S.E.    Robust z
## (Intercept) -1.90049539 0.11871090 -16.0094430  0.11908696 -15.9588874
## age         -0.14123592 0.05608034  -2.5184570  0.05820089  -2.4266968
## smoke1       0.31382583 0.18719721   1.6764450  0.18784180   1.6706922
## age:smoke1   0.07083185 0.08917757   0.7942788  0.08827886   0.8023647
##
## Estimated Scale Parameter:  1.001273
## Number of Iterations:  1
##
## Working Correlation
##           [,1]      [,2]      [,3]      [,4]
## [1,] 1.0000000 0.3543843 0.3543843 0.3543843
## [2,] 0.3543843 1.0000000 0.3543843 0.3543843
## [3,] 0.3543843 0.3543843 1.0000000 0.3543843
## [4,] 0.3543843 0.3543843 0.3543843 1.0000000
```

Enfoque Bayesiano

1) Modelo I: Independencia entre las observaciones

Preparación de los datos

```
data <- read.csv("Ohio.csv",header=T)
```

Modelo

El modelo planteado es similar al modelo planteado en el enfoque frecuentista, solo agregamos distribuciones apriori para los parámetros de interés, en este caso como no conocemos apriori las distribuciones de los coeficientes de regresión proponemos $\beta_i \sim N(0, 1000)$, $i = 1, 2, 3, 4$ y $\sigma_e^2 \sim \Gamma(1, 10^{-5})$.

```
model3 <- resp ~ age + smoke + age*smoke + f(id, model="iid")
m3 <- inla(model3, family = "binomial", Ntrials = 1,
           data=data)
```

Resumen de los efectos fijos:

```
round(m3$summary.fixed, 4)
```

```
##                mean     sd 0.025quant 0.5quant 0.975quant mode kld
## (Intercept) -3.0191 0.2013    -3.4362  -3.0109    -2.6468   NA   0
## age         -0.2052 0.0804    -0.3639  -0.2049    -0.0485   NA   0
## smoke        0.4505 0.2518    -0.0425   0.4500     0.9464   NA   0
## age:smoke    0.1001 0.1285    -0.1518   0.1001     0.3522   NA   0
```

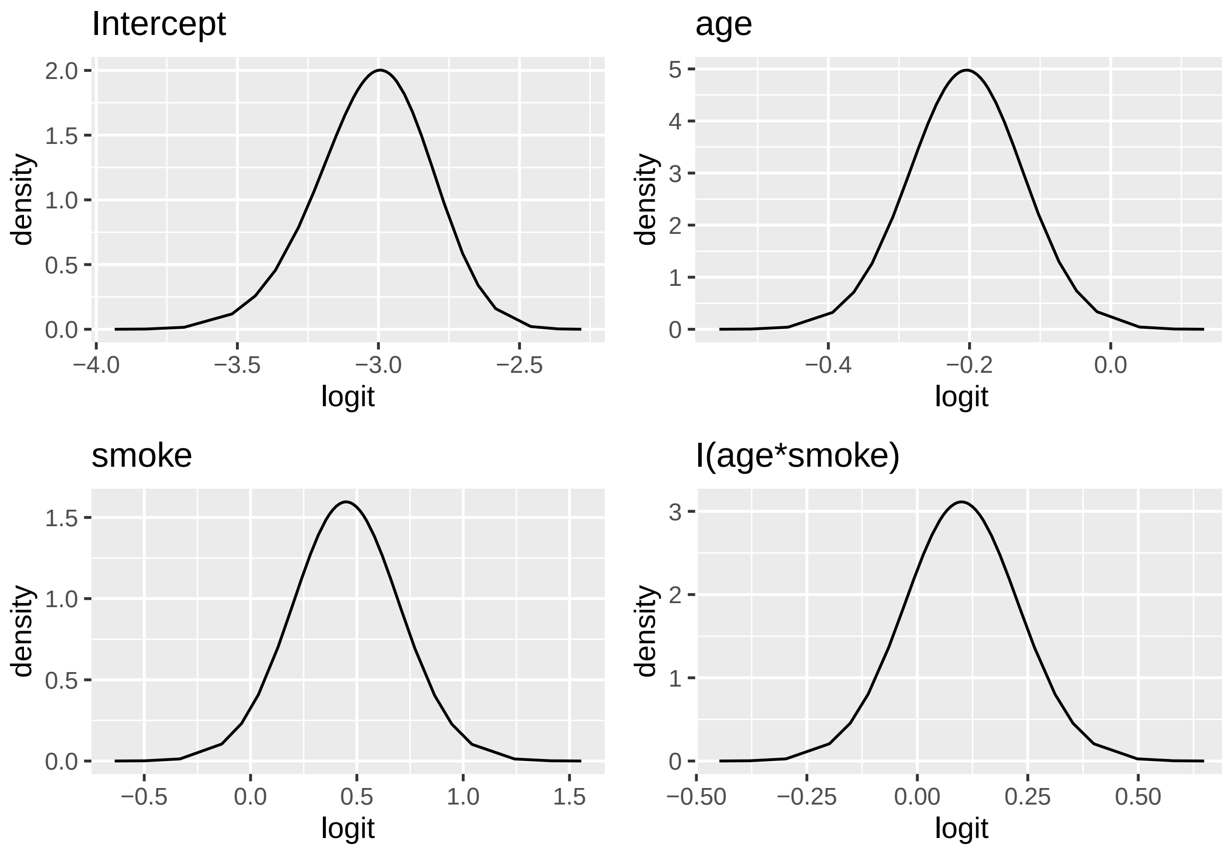
Resumen de los efectos aleatorios:

```
round(bri.hyperpar.summary(m3), 4)
```

```
##             mean     sd q0.025   q0.5 q0.975   mode
## SD for id 1.9273 0.1569 1.6296 1.9117 2.2571 1.8833
```

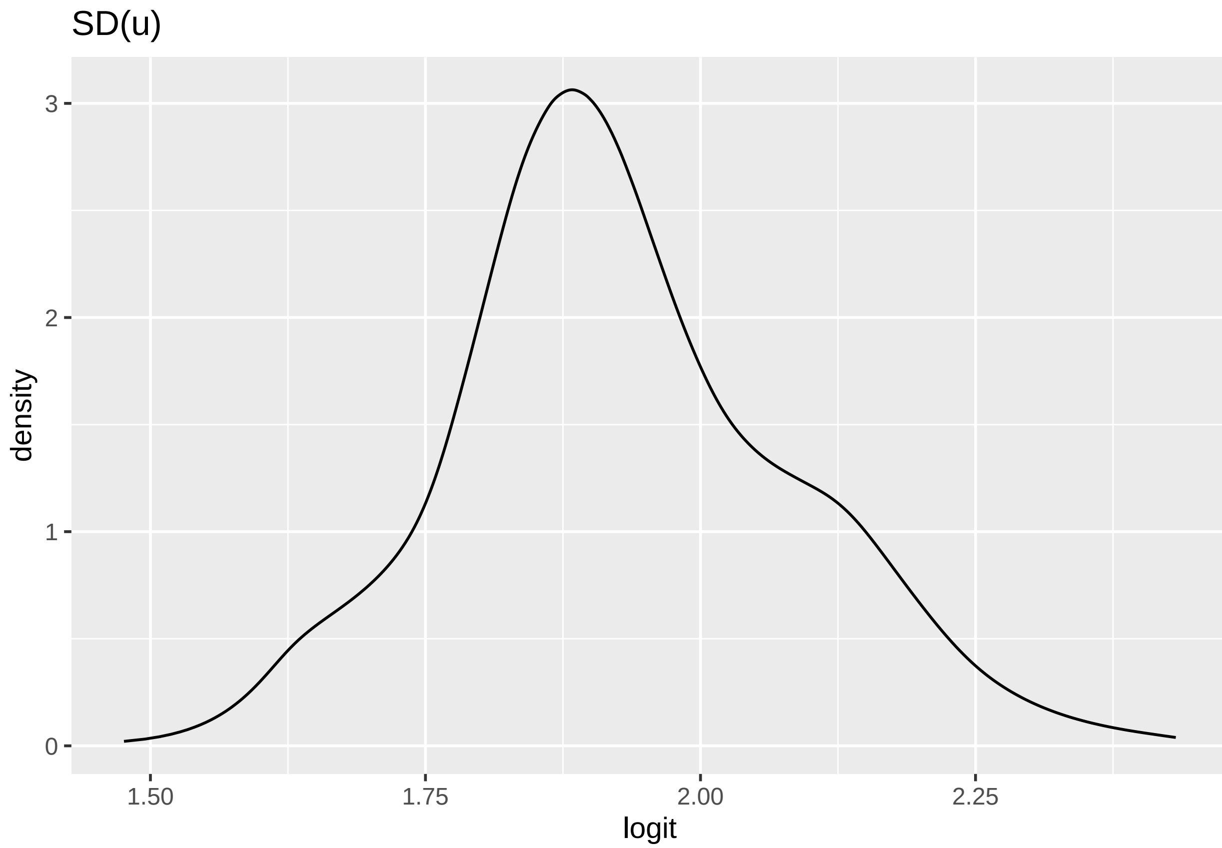
Graficas de las distribuciones a posteriori de los parámetros de interés:

```
p1 <- ggplot(data.frame(m3$marginals.fixed[[1]]),aes(x,y)) +
geom_line()+xlab("logit")+ylab("density")+ggtitle("Intercept")
p2 <- ggplot(data.frame(m3$marginals.fixed[[2]]),aes(x,y)) +
geom_line()+xlab("logit")+ylab("density")+ggtitle("age")
p3 <- ggplot(data.frame(m3$marginals.fixed[[3]]),aes(x,y)) +
geom_line()+xlab("logit")+ylab("density")+ggtitle("smoke")
p4 <- ggplot(data.frame(m3$marginals.fixed[[4]]),aes(x,y)) +
geom_line()+xlab("logit")+ylab("density")+ggtitle("I(age*smoke)")
grid.arrange(p1,p2,p3,p4,ncol=2)
```



Graficas de las distribuciones a posteriori de los hiperparámetros de interés:

```
sden <- data.frame(bri.hyper.sd(m3$marginals.hyperpar[[1]]))
p1 <- ggplot(sden,aes(x,y)) + geom_line() + xlab("logit") +
ylab("density")+ggtitle("SD(u)")
grid.arrange(p1,ncol=1)
```



2) Modelo II: Dependencia entre las observaciones

Modelo

El modelo planteado es similar al modelo anterior, solamente agregregamos correlación entre las observaciones de cada niño, y planteamos que esta es la misma entre cada observación, por lo tanto agregamos una distribución apriori para $\sigma_u^2 \sim \Gamma(1, 10^{-5})$.

```
data$group <- rep(1:4,537)
model4 <- resp ~ age + smoke + I(age*smoke) +
f(id, group=group, control.group = list(model="exchangeable"))
m4 <- inla(model4, family = "binomial", Ntrials = 1,
           data=data)
```

```
##
##  *** inla.core.safe:  rerun to try to solve negative eigenvalue(s) in the Hessian
```

Resumen de los efectos fijos:

```
round(m4$summary.fixed, 4)
```

```
##                    mean     sd 0.025quant 0.5quant 0.975quant mode kld
## (Intercept)     -3.0508 0.2112    -3.4866  -3.0427    -2.6604   NA   0
## age             -0.2075 0.0811    -0.3677  -0.2071    -0.0495   NA   0
## smoke            0.4548 0.2530    -0.0402   0.4542     0.9536   NA   0
## I(age * smoke)   0.1014 0.1296    -0.1527   0.1013     0.3557   NA   0
```

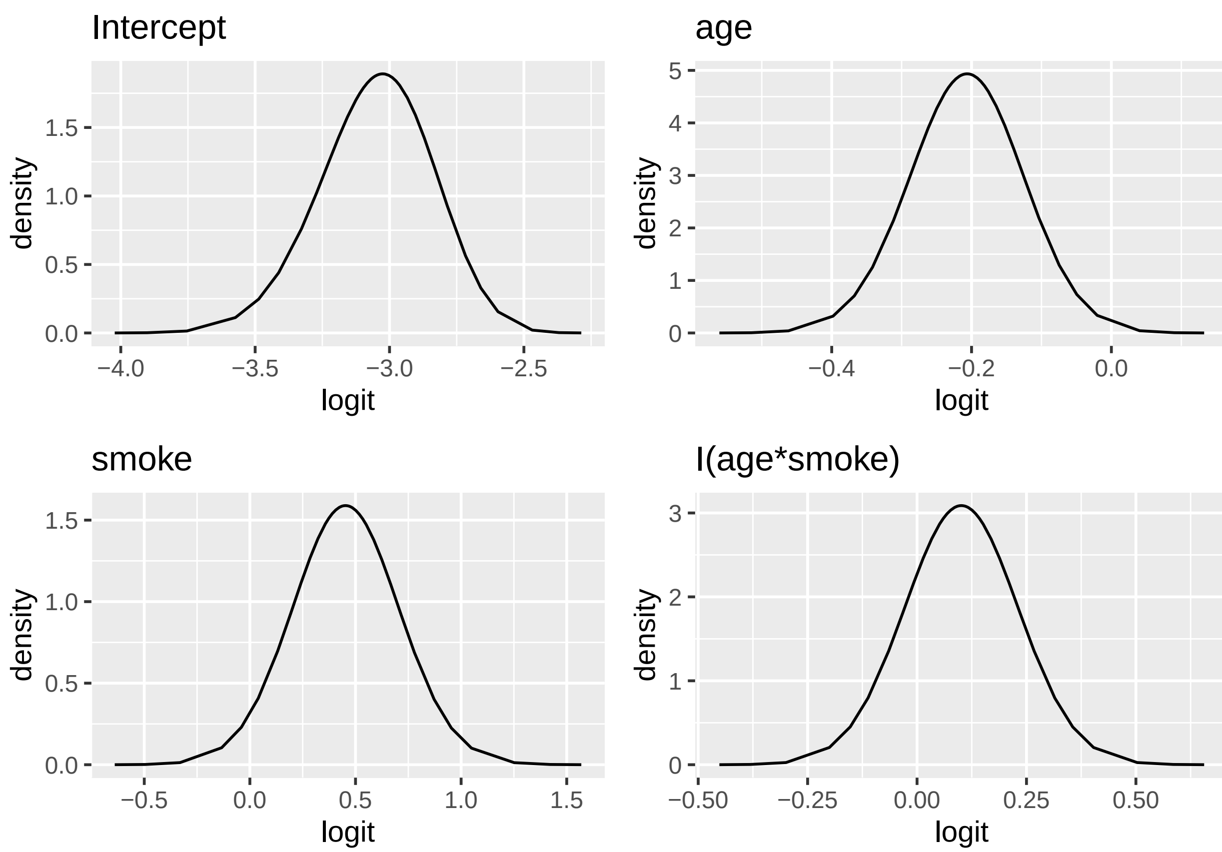
Resumen de los efectos aleatorios:

```r
round(bri.hyperpar.summary(m4), 4)
```

```
##                    mean     sd q0.025   q0.5 q0.975   mode
## SD for id        1.9427 0.1307 1.6956 1.9393 2.2087 1.9347
## GroupRho for id 0.9715 0.0237 0.9085 0.9779 0.9955 0.9883
```

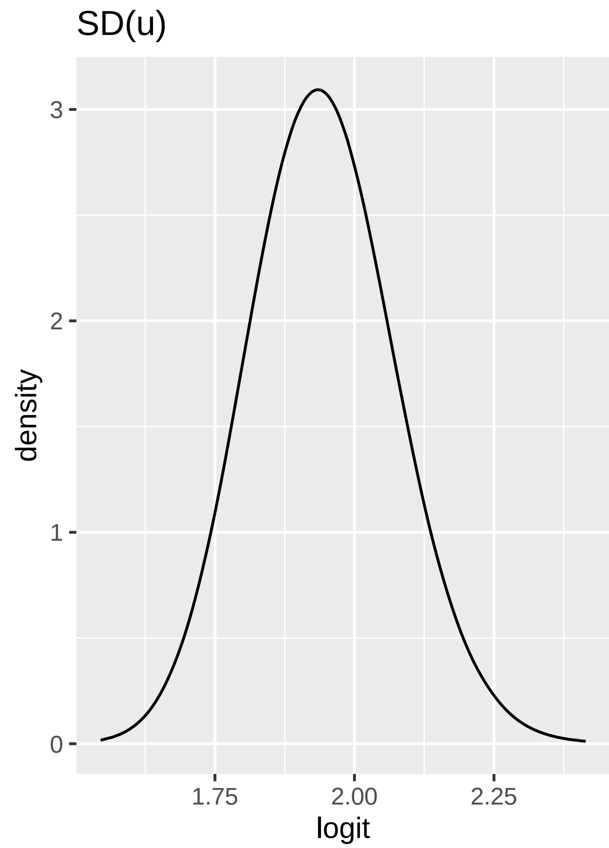
Graficas de las distribuciones a posteriori de los parámetros de interés:

```r
p1 <- ggplot(data.frame(m4$marginals.fixed[[1]]),aes(x,y)) +
geom_line()+xlab("logit")+ylab("density")+ggtitle("Intercept")
p2 <- ggplot(data.frame(m4$marginals.fixed[[2]]),aes(x,y)) +
geom_line()+xlab("logit")+ylab("density")+ggtitle("age")
p3 <- ggplot(data.frame(m4$marginals.fixed[[3]]),aes(x,y)) +
geom_line()+xlab("logit")+ylab("density")+ggtitle("smoke")
p4 <- ggplot(data.frame(m4$marginals.fixed[[4]]),aes(x,y)) +
geom_line()+xlab("logit")+ylab("density")+ggtitle("I(age*smoke)")
grid.arrange(p1,p2,p3,p4,ncol=2)
```



Graficas de las distribuciones a posteriori de los hiperparámetros de interés:

```r
sden <- data.frame(bri.hyper.sd(m4$marginals.hyperpar[[1]]))
rho <- data.frame(bri.hyper.sd(m4$marginals.hyperpar[[2]]))
p1 <- ggplot(sden,aes(x,y)) + geom_line() + xlab("logit") +
ylab("density")+ggtitle("SD(u)")
#p2 <- ggplot(rho,aes(x,y)) + geom_line() + xlab("logit") +
#ylab("density")+ggtitle("rho")
grid.arrange(p1,ncol=2)
```



Resumen

Se puede observar en la siguiente gráfica que tanto en el enfoque frecuentista y el enfoque bayesiano no hay diferencia significativa al considerar la correlación entre las observaciones de los niños dentro del efecto aleatorio. Por otra parte, las media de los parámetros de los efectos fijos relacionados con las covariables son similares en ambos enfoques y que solamente cambió el valor del intercepto. Los valores de la desvición estandar (en paréntesis) son muy similares dentro de cada enfoque.

Parámetros	Frecuentista		Bayesiano	
	Obs. Ind.	Obs. Dep	Obs. Ind.	Obs. Dep
β_1	-1.9008 (0.0887)	-1.9005 (0.1187)	-3.0386 (0.1976)	-3.0501 (0.2113)
β_2	-0.1413 (0.0695)	-0.1412 (0.0561)	-0.2062 (0.0805)	-0.2075 (0.0811)
β_3	0.3140 (0.1394)	0.3138 (0.1872)	0.4530 (0.2538)	0.4547 (0.2529)
β_4	0.0708 (0.1107)	0.0708 (0.0892)	0.1006 (0.0887)	0.1014 (0.1296)
ρ		0.3544		0.9715