Regresión probit bayesiana con variable latente

Sergio P.E.

11/16/2021

Longnecker, MP, MA Klebanoff, H Zhou, JW Brock. 2001. Association between maternal serum concentration of the DDT metabolite DDE and preterm and small-for-gestational-age babies at birth. The Lancet 358: 110-114.

Longnecker et al. demonstrate a powerful association between DDE levels in mothers' serum and the likelihood of premature birth. The higher the contamination level, the more likely was preterm birth. They also show that contamination is linked to the baby's size, with babies more likely to be small for their gestational age when born to mothers with higher DDE levels.

According to the lead author, in the US in the 1960s there may have been "an epidemic of pre-term births that we are just now discovering."

In the Lancet paper, the authors write that their results

```
" strongly suggest that DDT use increases preterm births, which is a major contrib
utor to infant mortality. If this association is causal, it should be included in
any assessment of the costs and benefits of vector control with DDT."
```

Interviewed by the British science magazine, New Scientist, Dr. Matthew Longnecker (US National Institute of Environmental Health Sciences), the lead author of the report, estimated that DDT use in the United States could have been responsible for as much as 15% of infant deaths during the 1960s.

What did they do? Longnecker et al. analyzed data and samples from a prospective study conducted of the causes of neurological disorders and other conditions in children born to mothers pregnant in the US between 1959 and 1966 (Niswander and Gordon 1972). The data were compiled in the US Collaborative Perinatal Project (CPP) , which enrolled some 42,000 women over that time span and evaluated 55,000 babies. Of the total sample of women participating, the vast majority deposited sufficient serum into a bank to allow subsequent measurement of DDE levels. Details of the sampling are available in the paper.

Longnecker et al. measured concentrations of p,p'-DDT and p,p'-DDE in serum samples taken from the pregnant mothers. They then evaluated the DDE concentrations in relation to the chances of a pre-term birth and infant size (controlling for gestational age).

What did they find?

The likelihood of preterm birth increased steadily with increasing concentration of DDE in the mother's serum ($p < 0.0001$). The likelihood of the baby being small for its age was highest for babies whose mother had the highest DDE concentration). Similarly, the chances of a baby weighing less than 2500 grams (or 5 pounds, 8 ounces) increased dramatically with increasing DDE concentration. The highest levels of DDE increased that likelihood by over four-fold. ($p < 0.0001$). The authors were able to detect the onset of these DDE impacts at a concentration of 10 micrograms/liter. What does this mean?

DDE blood levels in the US are now well beneath those concentrations detected in this study (made of blood samples obtained in 1959-1966) and beneath the apparent threshold for the effect detected in the authors' statistical analysis (10 micrograms/liter). Hence even though DDE contamination persists in people in the US, similar studies today in the US would be unlikely to detect an effect. This is not the case, however, in countries where DDT continues to be used for insect control.

According to the authors:

```
"In tropical countries, where DDT is used for malaria control, blood concentration
s of DDE can greatly exceed the range observed" in the sample they studied. If DDE
causes premature birth, it is likely to cause increased infant mortality."
```

They go on to observe that “Benefits of vector control with DDT might need to be reassessed in the context of this adverse effect on human beings and the availability of alternative methods of vector management.” This is a cautious way of stating that DDT’s potential contribution to infant mortality should be factored into decisions about phasing out DDT use, increasing the urgency of finding safe and affordable alternatives.

Datos

```
library(mvtnorm)
# read in data
dde<- read.table("/Users/peser/Dropbox/Metodos Estadísticos Avanzados/datos/ddedat
a2.txt",header=T)
# variable key
# x = dde dose
# y = 0/1 indicator of preterm birth
# z1-z5 = possible confounders
dde$xtrue <- dde$x
dde$x<- (dde$x - mean(dde$x))/sqrt(var(dde$x)) # normalize dde
```

Estimación por máxima verosimilitud

```
# Maximum likelihood results
X<- cbind(1,dde$x,dde$z1,dde$z2,dde$z3,dde$z4,dde$z5)
y<- dde$y
dde_mle<- glm(y ~ -1+X, family=binomial("probit"))
# Summary table

summary(dde_mle)
```

```
##
## Call:
## glm(formula = y ~ -1 + X, family = binomial("probit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3854  -0.5952  -0.5208  -0.4496   2.3678
##
## Coefficients:
##    Estimate Std. Error z value Pr(>|z|)
## X1 -1.08068    0.04355 -24.816  < 2e-16 ***
## X2  0.17536    0.02909   6.028 1.67e-09 ***
## X3 -0.12817    0.03528  -3.633 0.000280 ***
## X4  0.11097    0.03366   3.297 0.000978 ***
## X5 -0.01705    0.03405  -0.501 0.616659
## X6 -0.08216    0.03576  -2.298 0.021571 *
## X7  0.05462    0.06473   0.844 0.398721
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3299.4  on 2380  degrees of freedom
## Residual deviance: 1966.9  on 2373  degrees of freedom
## AIC: 1980.9
##
## Number of Fisher Scoring iterations: 4
```

```
# Wald test is highly significant
beta_mle<- dde_mle$coef # maximum likelihood estimates
```

Distribución *a priori*

```
# Prior parameters
beta0<- rep(0,7)
Pbeta0<- 0.25*diag(7)
```

Algoritmo de Albert y Chib (Gibbs sampler)

```
# Run Albert & Chib algorithm - output results in beta.out file
beta<- rep(0,7) # starting value of chain
n<- nrow(dde)   # number of subjects
z<- rep(0,n)    # initial values of underlying variables
G<- 10000        # number of MCMC iterations

# Run Gibbs sampler
gibbs.fun <- function(){
  eta<- X%*%beta # linear predictor
  # sample underlying normal variables from truncated normal
  # full conditional posterior distributions
  z[y==0]<- qnorm(runif(sum(1-y),0,pnorm(0,eta[y==0],1)),eta[y==0],1)
  z[y==1]<- qnorm(runif(sum(y),pnorm(0,eta[y==1],1),1),eta[y==1],1)

  # sample betas from normal full conditional posterior distribution
  Vbeta<- solve(Pbeta0 + t(X)%*%X)
  Ebeta<- Vbeta%*%(Pbeta0%*%beta0 + t(X)%*%z)
  beta <<- c(rmvnorm(1,Ebeta,Vbeta))
}

out <- replicate(G,gibbs.fun())
write(out,file="beta.out",ncol=7)
```

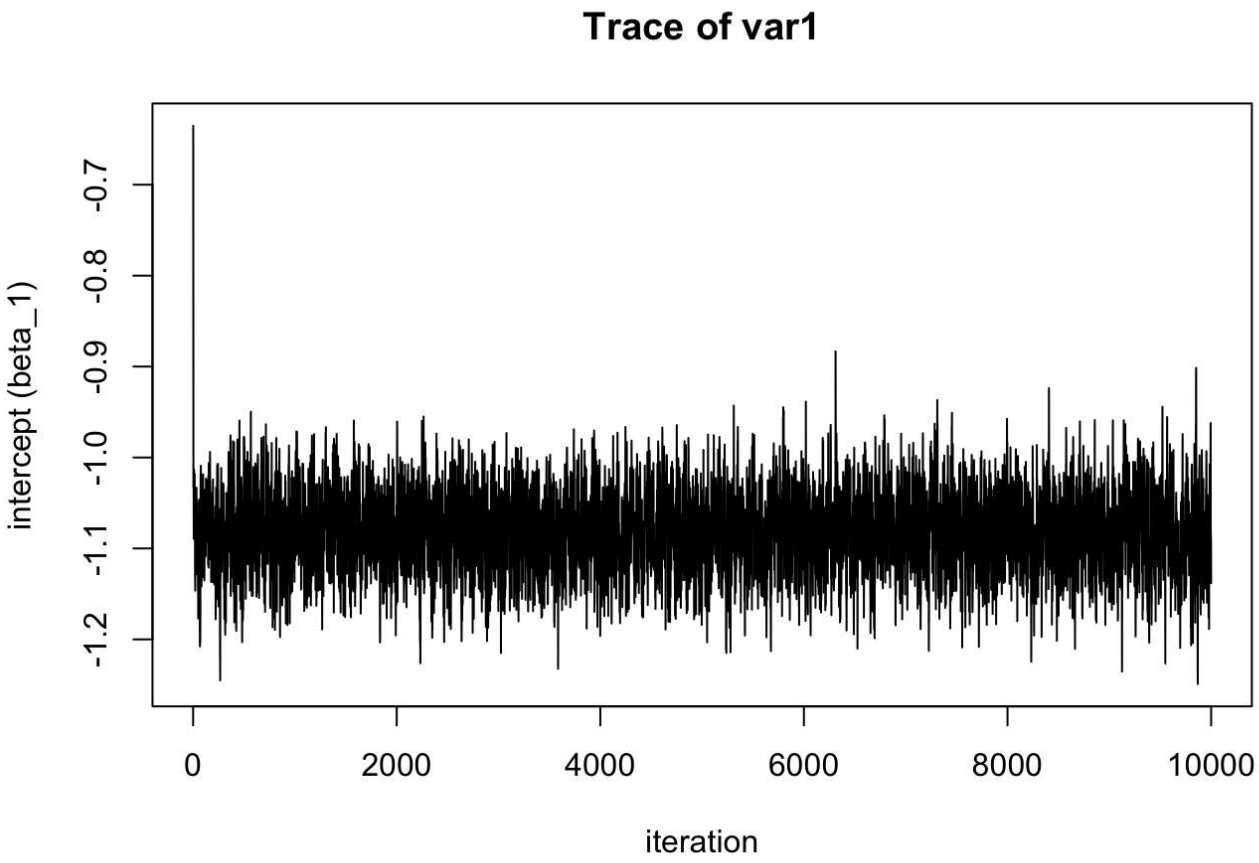
Diagnóstico

```
library(coda)
beta_out<- as.mcmc(matrix(scan("beta.out"), ncol=7, byrow=T))
geweke.diag(beta_out, frac1=0.1, frac2=0.5)
```

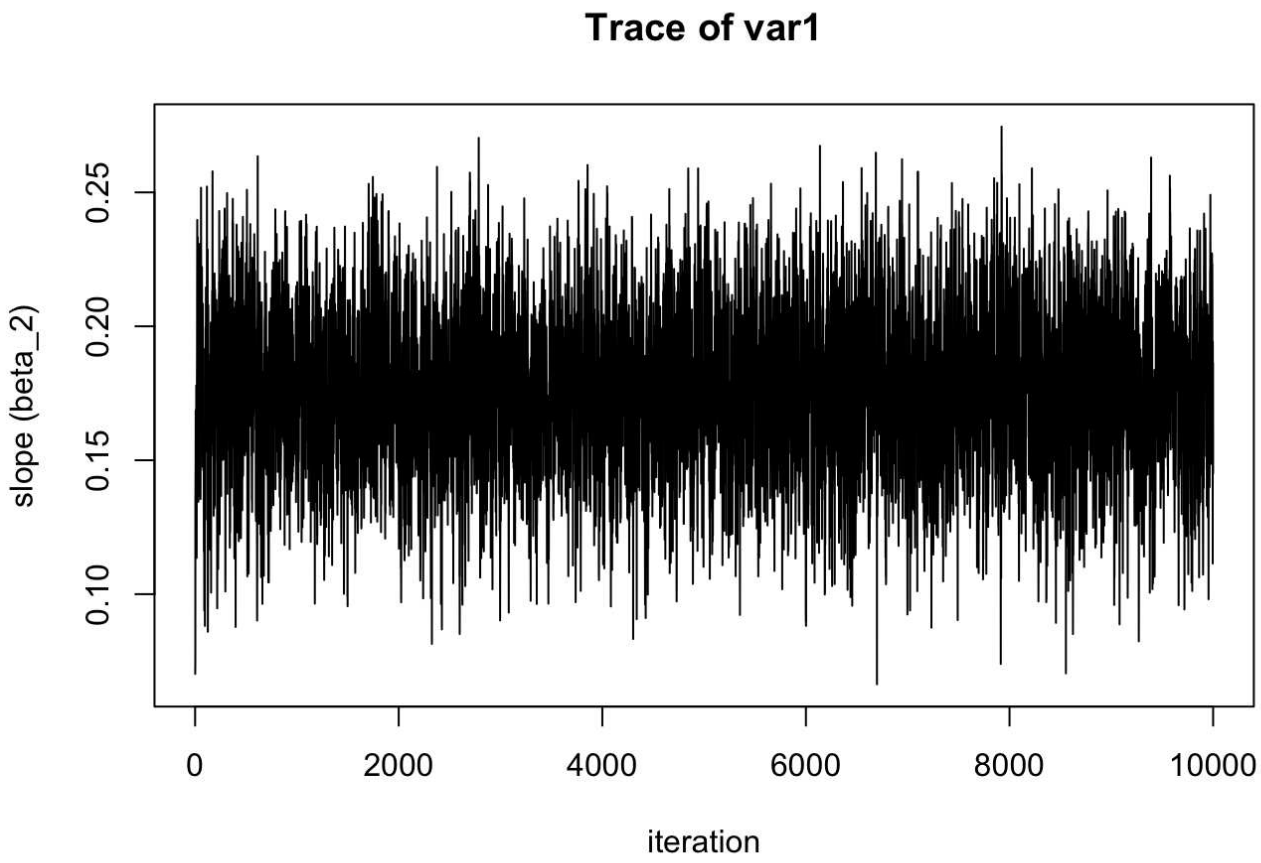
```
##
## Fraction in 1st window = 0.1
## Fraction in 2nd window = 0.5
##
##     var1     var2     var3     var4     var5     var6     var7
##   0.3393 -0.1356   0.5726   0.2662   0.5937   1.2247 -0.6910
```

Gráficos de las cadenas

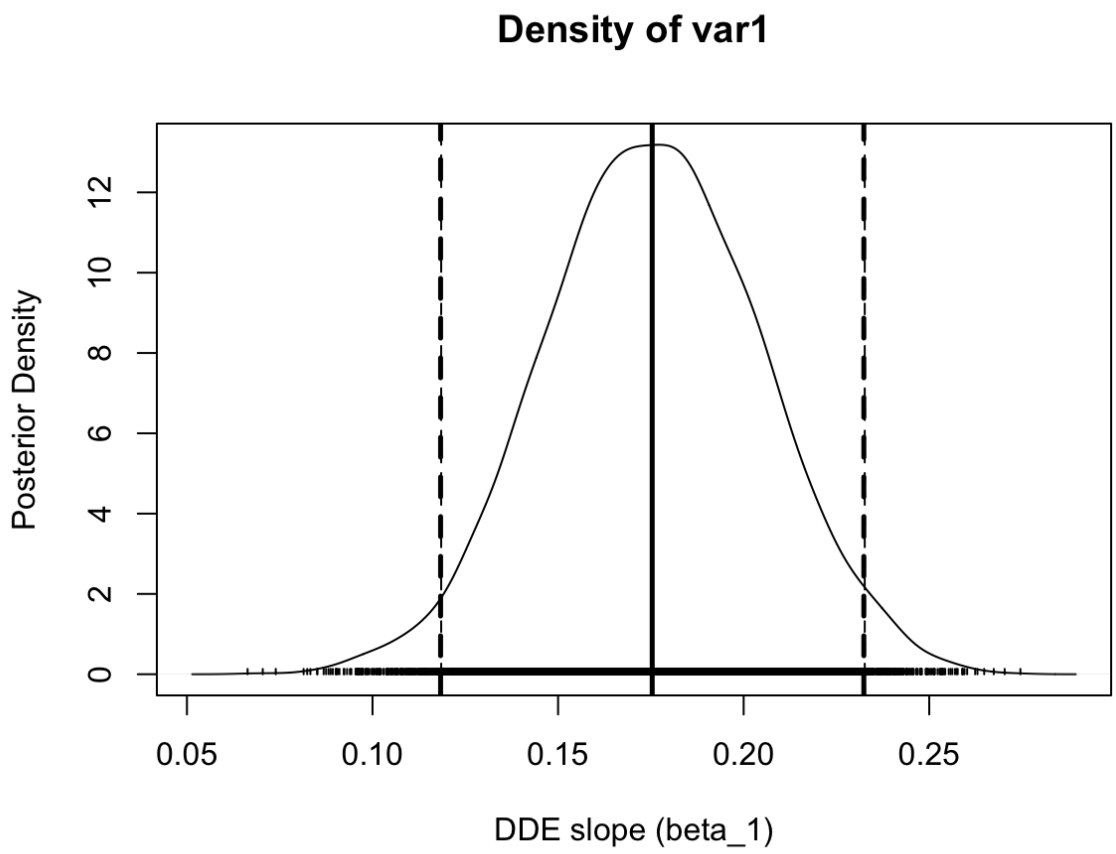
```
# Plot chains for beta[1] and beta[2]
par(mfrow=c(2,1))

plot(beta_out[,1],type="l",xlab="iteration",
     ylab="intercept (beta_1)",density = FALSE)
```



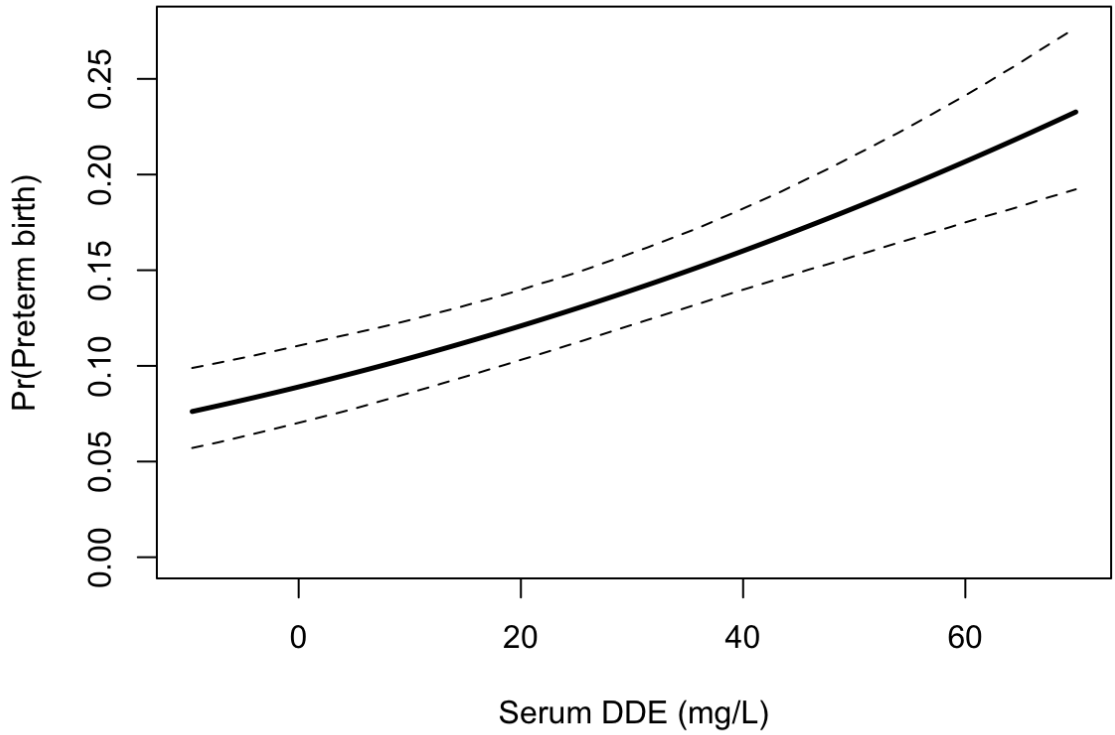

```
plot(beta_out[,2],type="l",xlab="iteration",ylab="slope (beta_2)",density = FALSE)
```




```
# Plot marginal posterior density of slope
slp<- as.mcmc(beta_out[1001:10000,2])
par(mfrow=c(1,1))
par(mar=c(5,5,5,5))
plot(slp,type="l",xlab="DDE slope (beta_1)",ylab="Posterior Density",
     cex=1.2,trace = FALSE)
abline(v=mean(slp))
abline(v=0.17536139,lwd=2.5) # MLE
abline(v=0.17536139 + 1.96*0.02909*c(-1,1),lwd=2.5,lty=2)
abline(v=quantile(slp,probs=c(0.025,0.975)),lty=2)
```




```
# Plot estimated dose response curve for preterm birth
par(mfrow=c(1,1))
par(mar=c(5,5,5,5))
xg<- seq(-2,2,length=100)  # grid of dde values
beta<- beta_out[1001:10000,] # discard burn-in
post<- matrix(0,100,4)
for(i in 1:100){
  post[i,1]<- mean(pnorm(beta[,1] + xg[i]*beta[,2]))
  post[i,2:3]<- quantile(pnorm(beta[,1] + xg[i]*beta[,2]),probs=c(0.025,0.975))
  post[i,4]<- pnorm(-1.08068 + xg[i]*0.17536139)
}
xtrue<- xg*sd(dde$xtrue) + mean(dde$xtrue) # back transform
plot(xtrue,post[,1],xlab="Serum DDE (mg/L)",ylab="Pr(Preterm birth)",cex=1.2,
     ylim=c(0,max(post)), type="l")
lines(xtrue,post[,2],lty=2)
lines(xtrue,post[,3],lty=2)
lines(xtrue,post[,4],lwd=2.5)
```




```
# calculate posterior summaries of regression coefficients
table1<- matrix(0,7,5)
for(i in 1:7){
  table1[i,]<- c(mean(beta[,i]),median(beta[,i]),sqrt(var(beta[,i])),
                 quantile(beta[,i],probs=c(0.025,0.975)))
}
table1<- round(table1*100)/100
colnames(table1) <- c("Media","Mediana","Desv. Est.", "Inf.","Sup")
table1
```

```
##      Media Mediana Desv. Est.  Inf.   Sup
## [1,] -1.08   -1.08       0.04 -1.17 -1.00
## [2,]  0.18    0.18       0.03  0.12  0.23
## [3,] -0.13   -0.13       0.03 -0.20 -0.06
## [4,]  0.11    0.11       0.03  0.05  0.18
## [5,] -0.02   -0.02       0.03 -0.08  0.05
## [6,] -0.08   -0.08       0.04 -0.15 -0.01
## [7,]  0.05    0.05       0.07 -0.07  0.18
```