Inferencia II - Entrega 3

Daniel Czarnievicz

October 2017

Considere nuevamente los datos:

Year	Fatal accidents	Passenger deths	Death rate	Miles flown
1976	24	734	0,19	3.863,16
1977	25	516	0,12	4.300,00
1978	31	754	0,15	5.026,67
1979	31	877	0,16	5.481,25
1980	22	814	0,14	5.814,29
1981	21	362	0,06	6.033,33
1982	26	764	0,13	5.876,92
1983	20	809	0,13	6.223,08
1984	16	223	0,03	7.433,33
1985	22	1.066	0,15	7.106,67

El objetivo es modelar la tasa de accidentes fatales por millas voladas. Para esto consideramos el modelo

$$y_i \overset{ind}{\sim} \text{Poisson}(x_i \lambda_i)$$

donde y_i es el número de accidentes fatales en cada año, x_i es el total de millas voladas en cada año y λ_i es la tasa sobre la que nos interesa realizar inferencia.

Ejercicio 1

Supongamos previas independientes para cada año,

$$\lambda_i \overset{ind}{\sim} \text{Exp}(t)$$

donde t es un valor conocido.

1. Muestra que los λ_i también son independientes en la distribución posterior.

$$p(\lambda|t;\, x;\, y) = \frac{p(\lambda;\, y|t,\, x)}{p(y|\, t,\, x)} \propto p(y|\, \lambda,\, t,\, x)\, p(\lambda|\, t,\, x) =$$

$$= \prod_{i=|}^{10} p(y_i|\, \lambda_i,\, t,\, x_i)\, p(\lambda_i|\, t,\, x_i) = \prod_{i=1}^{10} \frac{e^{-x_i\, \lambda_i} (\lambda_i x_i)^{y_i}}{y_i!} t e^{-t\lambda_i} \text{I}_{[y_i \in \mathbb{N}_0]} \text{I}_{[\lambda_i \geq 0]} \propto$$

$$\propto \prod_{i=1}^{10} e^{-(x_i+t)\lambda_i}\, \lambda_i^{y_i}\, \text{I}_{[\lambda_i \geq 0]} = \prod_{i=1}^{10} \underbrace{e^{-(x_i+t)\lambda_i}\, \lambda_i^{(y_i+1)-1}\, \text{I}_{[\lambda_i \geq 0]}}_{\substack{\text{kernel de una} \\ \text{Gamma}(y_i+1;\, x_i+t)}} \Rightarrow$$

$$\Rightarrow p(\lambda|t;\, x;\, y) = \prod_{i=1}^{10} g_i(\lambda_i|t;\, x_i;\, y_i) \Rightarrow \boxed{\lambda_i|t;\, x_i;\, y_i \;\; \text{indep}}$$

2. Calcula la probabilidad que la tasa en 1985 sea mayor a los años anteriores (son 9 probabilidades en total).

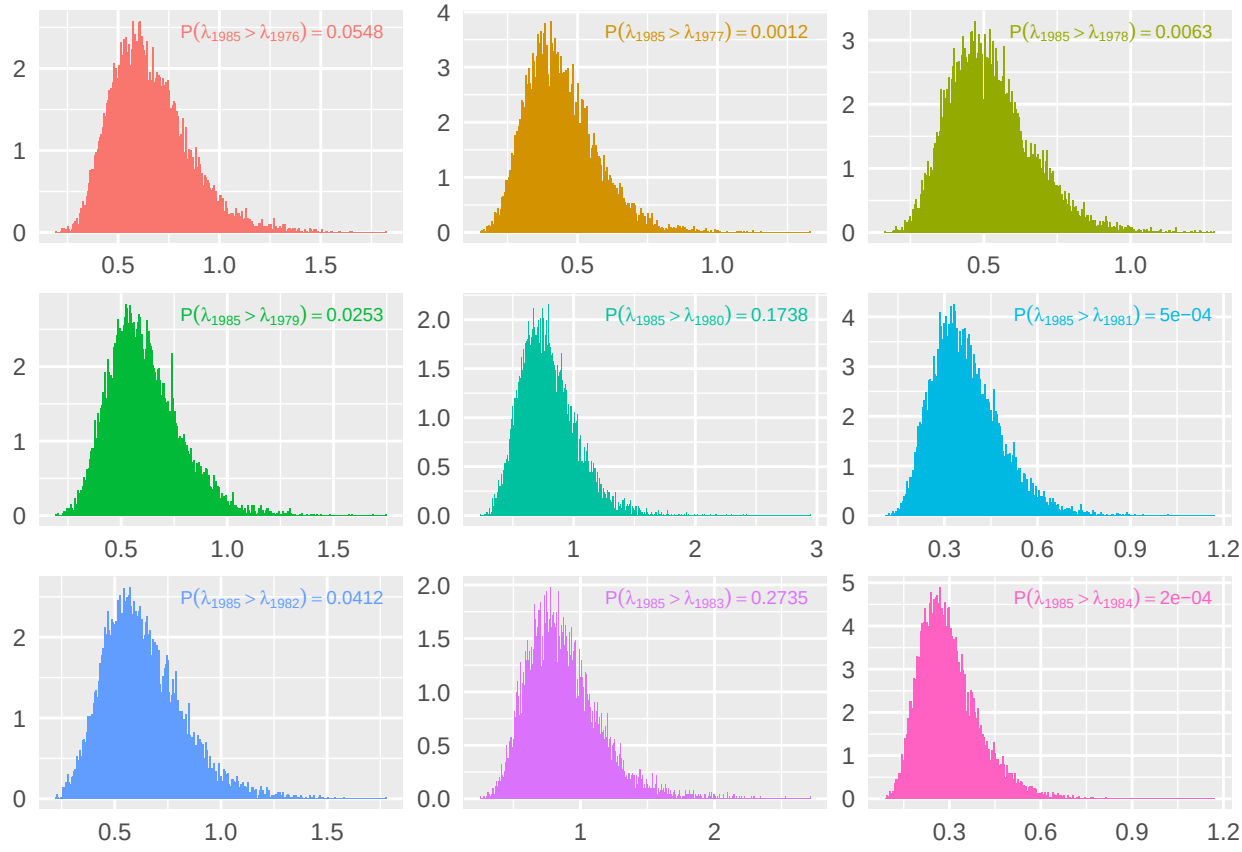
Dado que las posteriores marginales para cada λ_j son independientes, podemos simularlas por separado.

```
n.iter <- 10000
t <- 1
lambdas <- matrix(ncol=dim(d)[1], nrow=n.iter, dimnames=list(paste0("iter", 1:n.iter),
                                                             paste0("lambda", 1:10)))
h <- 0.005

set.seed(123456789)
for(i in 1:10){
   lambdas[,i] <- rgamma(n.iter, shape=d$fatal_accidents[i] + 1, rate=d$passenger_deaths[i] + t)
}
```

Una vez simulados los valores de $\lambda_j \ \forall j = 1976; \ldots; 1985$ (los valores de j fueron indexados como $1; \ldots; 10$) obtenemos simulaciones del cociente $\frac{\lambda_{1985}}{\lambda_j}$ los cuales podemos utilizar para aproximar la probabilidad de interés: $P(\lambda_{1985} > \lambda_j \,|\, \cdot) = P\left(\frac{\lambda_{1985}}{\lambda_j} > 1 \,\middle|\, \cdot\right)$. A continuación se muestran los resultados.

```
cocientes <- matrix(ncol=dim(d)[1]-1, nrow=n.iter, dimnames=list(paste0("iter", 1:n.iter),
                                                                 paste("Cociente", 1:9)))
for(j in 1:9){
   cocientes[,j] <- lambdas[,10]/lambdas[,j]
}
cocientes <- as_tibble(cocientes) %>%
   gather(key="lambda", value="valores") %>%
   mutate(mayora1 = ifelse(valores>1, 1, 0))
```



Ejercicio 2

Consideremos el siguiente modelo **jerárquico**

$$\begin{aligned} y_i \,|\, x_i\, \lambda_i &\overset{ind}{\sim} \text{Poisson}(x_i\, \lambda_i) \\ \lambda_i \,|\, \tau &\overset{ind}{\sim} \text{Exp}(\tau) \\ \tau &\sim \text{Gamma}(a;\, b) \end{aligned}$$

1. Plantea la condicional conjunta de todos los parámetros del modelo.

$$p(\lambda_1; \ldots; \lambda_{10};\, \tau \,|\, y;\, x) \propto p(y \,|\, \lambda_1; \ldots; \lambda_{10};\, \tau;\, x)\, p(\lambda_1; \ldots; \lambda_{10};\, \tau \,|\, x) =$$

$$= p(y \,|\, \lambda_1; \ldots; \lambda_{10};\, \tau;\, x)\, p(\lambda_1; \ldots; \lambda_{10} \,|\, \tau;\, x)\, p(\tau \,|\, x) =$$

$$= \prod_{i=1}^{10} \left[\frac{e^{-(x_i\, \lambda_i)}\, (x_i\, \lambda_i)^{y_i}\, \tau e^{-(\tau \lambda_i)}}{y_i!}\, \mathrm{I}_{[y_i \in \mathbb{N}_0]}\, \mathrm{I}_{[\lambda_i \geq 0]} \right] \frac{b^a}{\Gamma(a)}\, \tau^{a-1} e^{-b\tau}\, \mathrm{I}_{[\tau \geq 0]} \propto$$

$$\propto \prod_{i=1}^{10} \left[e^{-(x_i \lambda_i)}\, \lambda_i^{y_i}\, \tau\, e^{-(\tau \lambda_i)} \mathrm{I}_{[\lambda_i \geq 0]} \right] \tau^{a-1}\, e^{-b\,\tau}\, \mathrm{I}_{[\tau \geq 0]} = \prod_{i=1}^{10} \left[\lambda_i^{y_i}\, \exp\left\{ -(x_i + \tau) \lambda_i \right\}\, \mathrm{I}_{[\lambda_i \geq 0]} \right] \tau^{a+10-1}\, e^{-b\,\tau}\, \mathrm{I}_{[\tau \geq 0]}$$

2. Para implementar un algoritmo de Gibbs es necesario obtener las posteriores condicionales. Encuentra $p(\lambda_i \,|\, \lambda_{-i};\, \tau; y)$ y $p(\tau \,|\, \lambda_1; \ldots; \lambda_{10};\, y)$.

$$p(\lambda_i \,|\, \lambda_{-i};\, \tau;\, y;\, x) \propto \lambda_i^{y_i}\, \exp\Big\{ -(x_i + \tau) \lambda_i \Big\}\, \mathrm{I}_{[\lambda_i \geq 0]} \Rightarrow \lambda_i \,|\, \cdot \sim \text{Gamma}(y_i + 1;\, x_i + \tau)$$

$$p(\tau \,|\, \lambda;\; y;\, x) \propto \tau^{a+10-1}\, \exp\left\{ -\left(b + \sum_{i=1}^{10} \lambda_i \right) \tau \right\}\, \mathrm{I}_{[\tau \geq 0]} \Rightarrow \tau \,|\, \cdot \sim \text{Gamma}\left(a + 10;\, b + \sum_{i=1}^{10} \lambda_i \right)$$

3. Escriba el paso iterativo de un algoritmo de Gibbs para este modelo. Es decir, dado $(\lambda;\, \tau)^k$, ¿cómo se obtiene $(\lambda;\, \tau)^{k+1}$? (k representa el numero de iteración).

```
set.seed(1234)
a <- b <- 1
n.iter <- 100
titas <- matrix(nrow=n.iter, ncol=dim(d)[1]+1, byrow=TRUE,
                dimnames=list(paste0("iter", 1:n.iter), c("tau", paste0("lambda", 1:dim(d)[1]))))
titas["iter1", "tau"] <- rgamma(1, shape=a, rate=b)
titas["iter1", 2:(dim(d)[1]+1)] <- rexp(dim(d)[1], rate=titas["iter1", "tau"])
for(i in 2:dim(titas)[1]){
   titas[i, "tau"] <- rgamma(1, shape=a+dim(d)[1], rate=b+sum(titas[i-1, 2:dim(d)[1]+1]))
   for(k in 1:dim(d)[1]){
      titas[i, k+1] <- rgamma(1, shape=d$fatal_accidents[k] + 1,
                              scale=d$passenger_deaths[k] + titas[i, "tau"])
   }
}
```

Ejercicio 3

Consideremos otro modelo **jerárquico**

$$\begin{array}{rcl} y_i & \overset{ind}{\sim} & \text{Poisson}(x_i\,\lambda_i) \\ \lambda_i & \overset{ind}{\sim} & \text{Gamma}(\alpha;\,\beta) \\ (\alpha;\,\beta) & \sim & p(\alpha;\,\beta) \end{array}$$

1. Proponga una previa para $(\alpha;\,\beta)$. Cualquier opción puede ser válida. También se puede reparametrizar Gamma$(\alpha;\,\beta)$ y trabajar con otros dos parámetros que sea más fácil de modelizar. Escribe una breve justificación de tu elección.

Primero reparametrizamos la distribución de λ_i de forma tal que:

$$\left\{ \begin{array}{rcccl} \mu & = & E(\lambda_i) & = & \frac{\alpha}{\beta} \\ \tau & = & V(\lambda_i) & = & \frac{\alpha}{\beta^2} \end{array} \right. \Rightarrow \left\{ \begin{array}{rcl} \alpha & = & \frac{\mu^2}{\tau} \\ \beta & = & \frac{\mu}{\tau} \end{array} \right. \Rightarrow \lambda_i \sim \text{Gamma}\left(\frac{\mu^2}{\tau};\,\frac{\mu}{\tau}\right)$$

Eligo como previa para $(\mu;\,\tau)$ distribución cuyas márgenes se distribuyen Gamma, y son independientes.

$$(\mu;\,\tau) \sim \text{Gamma}(a;\,b) \times \text{Gamma}(c;\,d)$$

Justificación:

a. Why not?

b. Respeta el recorrido de $(\alpha;\,\beta)$, y por lo tanto, el de $(\mu;\,\tau)$

c. Dada la reparametrización, esta elección permite modelar la media y la varianza de λ_i, de forma de elegir previas consistentes con el conocimiento previo del investigador.

2. Escribe el modelo seleccionado en STAN (`modelo.stan`).

```
modelo.stan <- "
data {
      int<lower=1> n; // n?mero de observaciones
      int y[n];       // cantidad de acidentes fatales
      int x[n];       // cantidad de los muertos
      real a;         //
      real b;         //
      real c;         //
      real d;         //
}
parameters{
      real<lower=0> lambdas[n];
      real<lower=0> mu;
      real<lower=0> tau;
}
transformed parameters {
      real alpha;
      real beta;
      alpha = mu^2/tau;
      beta = mu/tau;
}
model {
      mu ~ gamma(a, b);
      tau ~ gamma(c, d);
      for (i in 1:n) lambdas[i] ~ gamma(alpha, beta);
```

```
      for (i in 1:n) y[i] ~ poisson(x[i]*lambdas[i]);
}
"
```

3. **(Opcional)** Utiliza STAN para obtener simulaciones posteriores de tu modelo. Comenta sobre la convergencia, presenta un resumen de la inferencia posterior y realiza el mismo ejercicio que en el deber anterior (histograma de las pendientes).

```
# Compila el modelo en C++
modelo <- stan_model(model_code=modelo.stan)
# Obtiene simulaciones para el modelo
sim <- sampling(modelo, iter=20000, data=list(n=dim(d)[1], y=d$fatal_accidents,
                                              x=d$passenger_deaths, a=1, b=1, c=1, d=1))
```

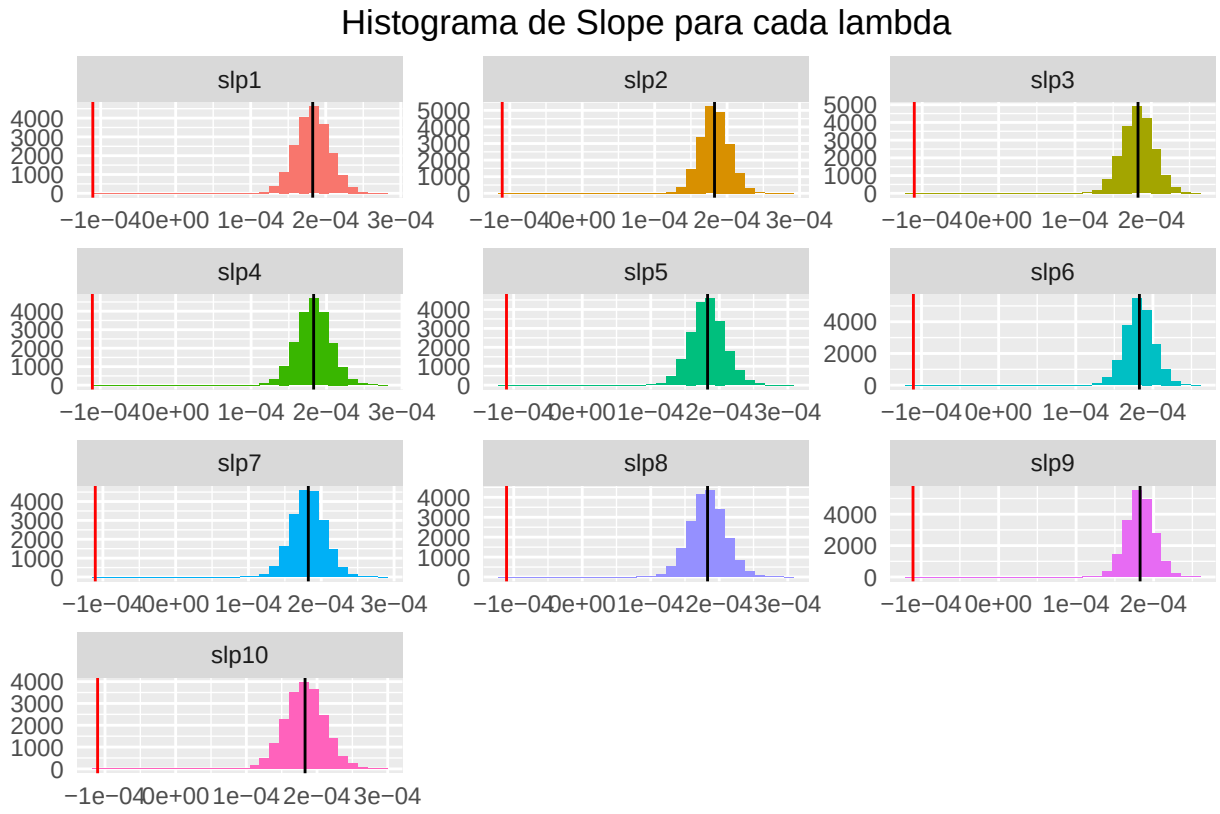
	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
lambdas[1]	0,03	0,00	0,01	0,02	0,03	0,03	0,04	0,05	54.422,54	1
lambdas[2]	0,05	0,00	0,01	0,03	0,04	0,05	0,05	0,06	38.718,42	1
lambdas[3]	0,04	0,00	0,01	0,03	0,04	0,04	0,04	0,05	46.966,91	1
lambdas[4]	0,04	0,00	0,01	0,03	0,03	0,04	0,04	0,05	52.664,65	1
lambdas[5]	0,03	0,00	0,01	0,02	0,03	0,03	0,03	0,04	49.300,94	1
lambdas[6]	0,05	0,00	0,01	0,03	0,04	0,05	0,06	0,08	33.134,79	1
lambdas[7]	0,04	0,00	0,01	0,02	0,03	0,03	0,04	0,05	54.487,55	1
lambdas[8]	0,03	0,00	0,01	0,02	0,02	0,03	0,03	0,04	45.849,99	1
lambdas[9]	0,06	0,00	0,01	0,03	0,05	0,06	0,07	0,09	28.145,81	1
lambdas[10]	0,02	0,00	0,00	0,02	0,02	0,02	0,03	0,03	33.856,89	1
mu	0,04	0,00	0,01	0,03	0,04	0,04	0,04	0,06	8.207,20	1
tau	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	7.858,79	1
alpha	9,98	0,32	15,70	1,40	4,04	6,70	11,21	35,90	2.417,20	1
beta	261,99	9,80	465,19	25,87	95,30	167,49	291,68	993,53	2.253,13	1
lp___	499,89	0,02	2,75	493,65	498,26	500,21	501,87	504,27	12.754,22	1

Tal como puede verse en la tabla de resumen, no hay indicios de falta de convergencia dado que los valores de los `Rhat` son todos muy cercanos a 1.

```
set.seed(1234)
slope_fn <- Vectorize(
   function(lam, xs) {
      yrep <<- rpois(length(xs), lambda=lam*xs)
      m <<- lm(log(yrep) ~ xs)
      coef(m)[2]
   },
   vectorize.args = 'lam')
slp.sims <- matrix(nrow=20000, ncol=20)
colnames(slp.sims) <- paste0(c("lambda", "slp"), rep(1:10, each=2))
for(i in 1:10){
   slp.sims[,paste0("lambda", i)] <- unlist(((sim@sim[["samples"]])[[4]])[i])
   slp.sims[,paste0("slp",i)] <- slope_fn(lam=slp.sims[,paste0("lambda",i)], xs=d$miles_flown)
}
slp.sims.pl <- as_tibble(slp.sims) %>% dplyr::select(matches("slp[0-9]+")) %>%
   gather(key="cadena", value="slp")
slp.sims.pl$cadena <- factor(slp.sims.pl$cadena, levels=paste0("slp", 1:10))

ggplot(slp.sims.pl) +
   geom_histogram(aes(slp, fill=cadena), show.legend=F) +
   labs(x="Slope", y=element_blank()) +
```

```
  ggtitle(label="Histograma de Slope para cada lambda") +
  theme(axis.ticks=element_blank(),
        axis.title.x=element_text(face="bold"),
        plot.title=element_text(hjust=0.5)) +
  geom_vline(xintercept=coef(lm(log(fatal_accidents) ~ miles_flown, data=d))[2], color="red") +
  geom_vline(data=(slp.sims.pl %>% group_by(cadena) %>% summarise(media=mean(slp))),
             aes(xintercept=media), color="black") +
  facet_wrap(~cadena, ncol=3, scales="free")
```



En el gráfico anterior, la linea roja representa el `slope` del modelo, mientras que la línea negra representa la media de las simulaciones para cada `lambda`.