

Tarea 4 - MLG

Blanca Garcia - 118886 Yuner Perez - 199813
Thomas Rudolf - 169293 Mariano Villafuerte - 156057

Spiegelhalter et al. (1995) analiza la mortalidad del escarabajo del trigo en la siguiente tabla, usando BUGS.

Dosis	# muertos	# expuestos
1.6907	6	59
1.7242	13	60
1.7552	18	62
1.7842	28	56
1.8113	52	63
1.8369	53	59
1.8610	61	62
1.8839	60	60

Estos autores usaron una parametrización usual en dos parámetros de la forma $p_i \equiv P(\text{muerte}|w_i)$, pero comparan tres funciones ligas diferentes:

$$\text{logit} : p_i = \frac{\exp(\alpha + \beta z_i)}{1 + \exp(\alpha + \beta z_i)}$$

$$\text{probit} : p_i = \Phi(\alpha + \beta z_i)$$

$$\text{complementario log-log} : p_i = 1 - \exp[-\exp(\alpha + \beta z_i)]$$

en donde se usa la covariada centrada $z_i = w_i - \bar{w}$ para reducir la correlación entre la ordenada α y la pendiente β . En OpenBUGS el código para implementar este modelo es el que sigue:

```

model {
  for (i in 1:k) {
    y[i] ~ dbin(p[i], n[i])
    logit(p[i]) <- alpha + beta * (w[i] - mean(w))
    # probit(p[i]) <- alpha + beta * (w[i] - mean(w))
    # cloglog(p[i]) <- alpha + beta * (w[i] - mean(w))
  } # End of loop over i
  alpha ~ dnorm(0.0, 1.0E-3)
  beta ~ dnorm(0.0, 1.0E-3)
} # End of model

```

Lo que sigue al símbolo `#` es un comentario, así que esta versión corresponde al modelo logit. También `dbin` denota la distribución binomial y `dnorm` denota la distribución normal, donde el segundo argumento denota la precisión, no la varianza (entonces las iniciales normales para α y β tienen precisión 0.001, que son aproximadamente iniciales planas (no informativas)). Se hace el análisis en STAN, se incluye el código en el anexo.

Logit model

variable	mean	sd	q5	q95
lp__	-187.27	1.03	-189.32	-186.29
alpha	0.75	0.14	0.53	0.99
beta	34.69	3.01	29.87	39.78
p[1]	0.06	0.02	0.04	0.09
p[2]	0.16	0.03	0.12	0.21
p[3]	0.36	0.03	0.31	0.42
p[4]	0.61	0.03	0.55	0.66
p[5]	0.80	0.03	0.75	0.84
p[6]	0.90	0.02	0.87	0.93
p[7]	0.96	0.01	0.93	0.97
p[8]	0.98	0.01	0.97	0.99

5. Supongan que se tiene una matriz de 4×4 de variables aleatorias Bernoulli, y la denotamos por $[X_{ij}]$, y sea $N(X)$ el número total de éxitos en X (la suma de X) y $D(X)$ es el total de vecinos de dichos éxitos (horizontales o verticales) que difieren. Por ejemplo:

$$\begin{array}{ccc} & X & N(X) \quad D(X) \\ \left(\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right) & 1 & 2 \\ \\ \left(\begin{array}{ccc} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{array} \right) & 3 & 5 \end{array}$$

Noten que si se escriben los elementos de X en un vector V , entonces existe una matriz M de 24×16 tal que $D(X)$ es la suma de los valores absolutos de MV . El 24 surge porque hay 24 pares de vecinos a revisar. Supongan que $\pi(X)$, la distribución de X , es proporcional a

$$\pi(X) \propto p^{N(X)}(1-p)^{16-N(X)} \exp(-\lambda D(X))$$

Si $\lambda = 0$, las variables son iid Bernoulli (p).

Usar el método de Metropolis-Hastings usando los siguientes kernels de transición.

Hay 216 posibles estados, uno por cada posible valor de X .

- Sea q_1 tal que cada transición es igualmente plausible con probabilidad $1/216$. Esto es, el siguiente estado candidato para X es simplemente un vector de 16 iid Bernoulli (p).
- Sea q_2 tal que se elige una de las 16 entradas en X con probabilidad $1/16$, y luego se determina el valor de la celda a ser 0 o 1 con probabilidad 0.5 en cada caso. Entonces sólo un elemento de X puede cambiar en cada transición a lo más.

Ambas q son simétricas, irreducibles y tienen diagonales positivas. La primera se mueve más rápido que la segunda.

Estamos interesados en la probabilidad de que todos los elementos de la diagonal sean

- Usen las dos q para estimar la probabilidad de 1's en la diagonal para $\lambda = 0, 1, 3$ y $p = 0.5, 0.8$. Esto se puede hacer calculando la fracción acumulada de veces que la cadena tiene 1's sobre la diagonal conforme se muestrea de la cadena. Comparar los resultados de las 2 cadenas. Tienen el mismo valor asintótico, pero ¿una cadena llega más rápido que la otra? ¿Alguna cadena muestra más autocorrelación que la otra (por ejemplo, estimados de la probabilidad en 100 pasos sucesivos muestran más correlación para una cadena que para la otra?).

En este problema, también determinen cuántas simulaciones hay que hacer, para desechar el periodo de calentamiento.

Anexo

STAN Model

```
data {
  int<lower=0> k; // observaciones
  array[k] int<lower=0> n; // expuestos
  array[k] int<lower=0> y; // muertos
  array[k] real w; // dosis
}

transformed data {
  real w_mean = mean(w); // promedio de w para centrar
}

parameters {
  real alpha;
  real beta;
}

transformed parameters {
  array[k] real p; // probability of success

  for (i in 1:k) {
    p[i] = inv_logit(alpha + beta * (w[i] - w_mean)); // logit
    // p[i] = Phi_approx(alpha + beta * (w[i] - w_mean)); // probit
    // p[i] = 1 - exp(-exp(alpha + beta * (w[i] - w_mean))); // cloglog
  }
}

model {
  alpha ~ normal(0, 1000); // inicial de alfa
  beta ~ normal(0, 1000); // inicial de beta

  for (i in 1:k) {
    y[i] ~ binomial(n[i], p[i]);
  }
}
```