Implementación de Regresión Lineal Bayesiana En pruebas farmacéuticas

Pedro Hernández Serrano
Examen Parcial - MLG
159549

---

1. Introducción

En el campo de la medicina, se llevan a cabo profundas pruebas antes de lanzar al mercado una nueva droga, debido a que dichos productos tienen que cumplir con legislaciones correspondientes al país donde se fabrican y donde se pondrán en circulación, de igual manera debe contar con los estándares publicados por las organizaciones mundiales de farmaco-medicina.
Dichas pruebas se llevan a cabo en diferentes fases, ya que, si no supera alguna fase, se replantea el diseño del producto.

Las pruebas clínicas fase I son experimentos con personas o animales en donde el objetivo es determinar la dosis más alta de una droga que se le puede dar a un paciente.

Durante el presente documento se hará mención de algunas herramientas las cuales nos ayudarán durante el estudio, el cual se llevó acabo en lenguaje R, utilizando las bondades de la paquetería $rjags$ cuya función principal es la de simular valores de una distribución posterior vía cadenas de markov a partir de distribuciones de probabilidad propuestas con la técnica de muestreo de Gibs.

2. Caso de Estudio

El departamento de Ciencia de Datos llevará a cabo el estudio estadístico del requerimiento de la empresa farmaceutica, el área correspondiente proporciona los datos

Se entiende previamente que se considerarán las variables más importantes para medir la severidad de la droga, el experimento fue llevado a cabo con 20 animales expuestos a 4 distintos niveles de droga, asignados de manera uniforme, es decir, cada dosis se probó en 5 animales.

Con ayuda de técnicas avanzadas de regresión con enfoque Bayesiano, y donde se probaron distintos variaciones de modelos lineales generalizados, se consiguiío un modelo que ajusta muy bien a los datos, acorde con el objetivo de poder hacer estimación con las variables respuesta

El objetivo es modelar el número de animales muertos (Y_i) en función de la cantidad de droga (en escala logartmica) (X_i) y el número de individuos expuestos o tratados (m_i). Este ejercicio se realizará comparando las siguientes opciones:

- Liga logit y aprioris no informativas
- Liga logit y aprioris informativas
- Liga probit y aprioris no informativas
- Liga probit y aprioris informativas

3. Análisis Exploratorio

En la tabla 1 observamos los resultados del ensayo

	Expuestos	**Muertes**	**Dosis**	**LogDosis**
1	5	0	0.423	-0.86
2	5	3	0.951	-0.05
3	5	1	0.740	-0.30
4	5	5	2.075	0.73

Cuadro 1: Resultado de Ensayo Médico

Notamos que el numero de muertes de los animales en los ensayos depende de la cantidad de dosis a la que es tratado, en este punto cabe señalar que se agergaró una columna adicional en la que se re-escaló la información de la dosis, para este experimento cada unidad de la dosis original se refiere a una pildora administrada al animal

En el gráfica 1 se expone la relación entre la variable respuesta Muertes y la variable explicativa Dosis (re-escalado), es claro ver que aumentan el número de muertes conforme la dosis suministrada aumenta, se tiene que tener cuidado con las relaciones de causalidad, Es por ello que el enfoque bayesiano ayuda, ya que es evidente que no existe relación digamos lineal entre las variables.

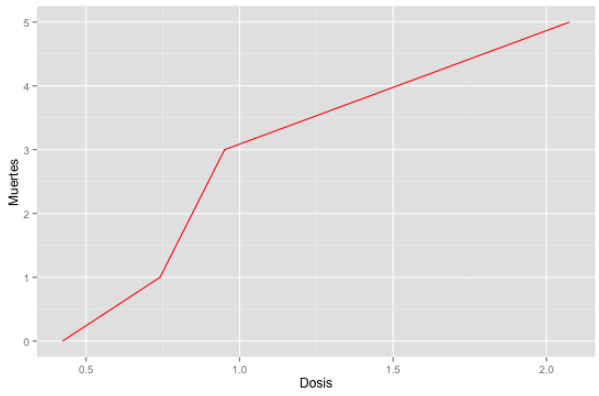

Figura 1: Muertes por nivel administrado

4. Modelo y Metodología

Se eligiío un modelo de regresión (lineal sobre los parametros) para el análisis, ya que permite mayor intuitividad en la interpretación de los resultados.

Para poder realizar un enfoque bayesiano de regresión es necesario definir las distribuciones de las variables que forman funciones de verosimilitud, así como las distribuciones de conocimiento previo o a priori, y de este modo el producto de ambas será proporcional a una distribución posterior.

Necesitamos también definir una función liga que relacione los estimadores de la regresión con los parámetros de la distribución de la variable respuesta o de interés, Es decir, una componente sistemática y otra aleatoria para el modelo.

La variable de interés es el némero de muertes, la cual se busca explicar con la cantidad de dosis y el número de animales estudiados en cada categoría.

Consideramos una distribución binomial para explicar la variable aleatoria donde los parḿetros de probabilidad y de conteo n y p se refieren a la probabilidad de muerte y número de expuestos respectivamente.

Lo anterior hace sentido ya que se trata de una variable aleatoria con dominio en los enteros no negativos lo cual va acorde con el número de muertes en cada experimento médico. De este modo, la verosimilitud de los datos se ve como :

$$Y_i|p_i, ne_i \ Bin(p_i, ne_i)$$

Se utilizaron dos tipos de ligas para el modelo binomial, la logística y la probit, las cuales van muy bien con dicha distribución, más aún, la logística es la liga canónica.

$$logit(p_i) = \frac{p_i}{1-p_i} = \beta_0 + \beta_1 * x$$
$$\Phi(p_i)^-1 = \beta_0 + \beta_1 * x$$

Se utilizaron también dos variantes en las funciones a priori, una opción vaga o no informativa (se construye con varianza grande) y una más sesgada con conocimiento previo, como sigue.

$$\beta_i \ N(0, 0{,}001) \text{ (opción vaga)}$$
$$\beta_0 \ N(-17{,}31, 1/1053{,}72) \text{ (opción informativa)}$$
$$\beta_1 \ N(2{,}57, 1/23{,}24) \text{ (opción informativa)}$$

La varianza se define en terminos de la precisión como $1/\tau$ donde $\tau \ Gamma(0{,}001, 0{,}001)$
No hay que olvidar señalar que para este problema el número de exposiciones es fijo $ne = 5$

5. Cómputo y resultados

Una manera útil de manipular los modelo de tipo *bugs* es escribirlas como funciones en R, y así poder cambiar muy fácil entre las variaciones del modelo con solo cambiar el nombre de la función.

Para simular probabilidades posteriores, utilizamos la función $jags$ la cual está cargada en la paquetería $rjags$ cuyo método de simulación es un proceso recursivo de muestreo de gibs, para el presente experimento se llevaron a cabo corridas con el listado de parámetros previamente mencionados, 1 cadena, 50 mil iteraciones y un proceso de calentamiento de 5 mil iteraciones, en caso de querer replicar la información se utiliz la semilla 159549.

Tenemos los siguientes resultados

	Modelo 1	**Modelo 2**	**Modelo 3**	**Modelo 4**
Liga	Logit	Logit	Probit	Probit
Prior	Vaga	Informativa	Vaga	Informativa
DIC	8.54	7.17	6.99	6.86
LD50	-0.11	-0.09	-0.12	-0.11

Cuadro 2: Variaciones en el Modelo Binomial

Para la opción binomial con liga logística es de interés conocer la dosis letal 50, la cual nos dice la dosis utilizada para que sea letal para la mitad de expuestos en el experimento.

$$p = \frac{\exp(\beta_0+\beta*LD50)}{1+(\exp(\beta_0+\beta*LD50))}$$

$p = log(\frac{0,5}{1-0,5})$, cuando $\eta = \beta_0 + \beta * LD50 = 0$, se tiene $LD50 = \frac{-\beta_0}{\beta}$

Considerando el criterio de informacin de la devianza (DIC) la cual mide la bondad de ajuste, y dado que el modelo con el menor DIC es que mejor se ajusta al conjunto de datos. De aquí que debido al criterio elegimos la variación 4 para utilizarse en las estimaciones como sigue.

	Mean	**SD**
$\hat{y_1}$	0.025	0.1685331
$\hat{y_2}$	0.961	1.0404622
$\hat{y_3}$	3.077	1.2832450
$\hat{y_4}$	4.959	0.2671898

Cuadro 3: Tabla de Estimaciones

	Mean	**2.25 %**	**97.5 %**
β_0	0.585	-0.359	1.700
β_1	5.223	1.844	10.07

Cuadro 4: Tabla de Regresores

Continuando con la misma temática en las siguientes gráficas observamos la estabilización y convergencia de los parámetros monitoreados, es claro observar que depués de la primera cuarta parte de las iteraciones, los valores arrojados por la cadena oscilan al rededor de la media, esto gracias al teorema ergódico de convergencia, como se ve en la figura 2

Con respecto a la dosis letal es deseable administrar paciente una dosis entre 0.71 y 1.15 ya que con probabilidad de 95 % ocurrirá en dicho intervalo, es decir $P(LD50 < 1{,}15) = 97{,}5$

	Mean	**2.25 %**	**97.5 %**
$LD50$	-0.1468	-0.314	1.435
$\exp^{log(X)}$	0.898	0.712	1.15

Cuadro 5: Monitoreo Dosis Letal

Notamos incluso que entre ms dosis se administre aumentará el número de muertes, por lo que como se platicó al inicio el objetivo de la prueba es encontrar el menor número de muertes.

Si consideramos $log = 0$ es decir una unidad de medida (una pastilla en este caso), superará la media del intervalo de probabilidad mencionado con anterioridad. Se recomienda entonces administrar una dosis de 1.15 unidades

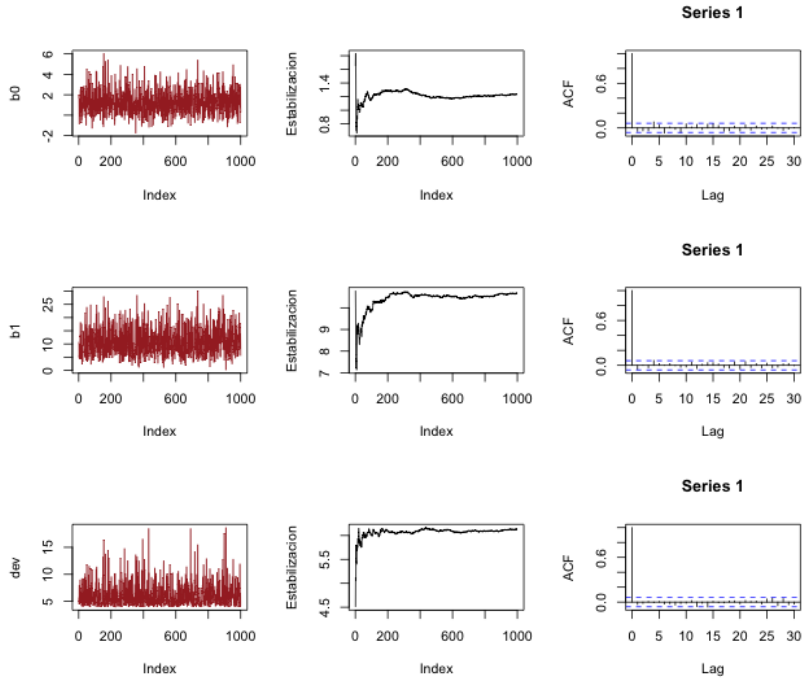

Figura 2: Convergencia

6. Conclusiones

Después de observar los resultados del computo, notamos que todos los modelos son parecidos en realidad, aunque el ajuste es mejor para la última prueba ya que aproxima mejor la dosis letal necesaria para matar a los animales. De cualquier forma la interpretación de los coeficientes es más clara utilizando una liga logit.

Dado que se selecciono aquel con menor DIC es decir el probit con aprioris informativas nos deja un claro escenario, de aquí que se recomiende una dosis mayor a 1.15 si se quiere maximizar la posibilidad de cura de los animales

En la figura 3 notamos la distribución que modela el número de muertes y la figura 4 nos da una perspectiva del ajuste del modelo, comparando los el número de muertes reales contra las predicciones.

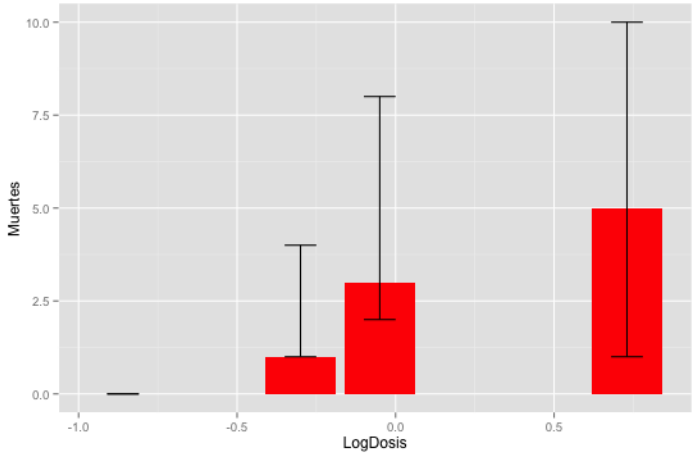

Figura 3: Distribución Binomial de Muertes

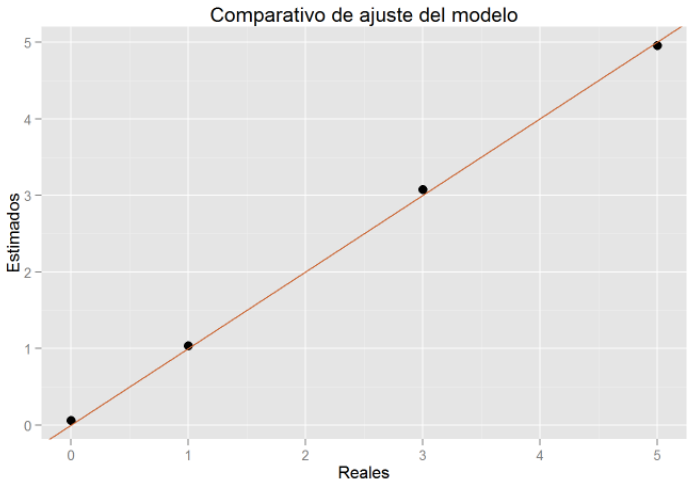

Figura 4: Ajuste del Modelo

```
bin_model1 <- function(){
  #Likelihood
  for (i in 1:n) {
    y[i] ~ dbin(p[i],ne[i])
    logit(p[i]) <- beta0+beta1*x[i]
  }
  #Priors vagas
  beta0 ~ dnorm(0,0.001)
  beta1 ~ dnorm(0,0.001)
  #Estimacion y_hat
  for (i in 1:n) {
    y_hat[i] ~ dbin(p[i],ne[i])
  }
}

bin_model2 <- function(){
  #Likelihood
  for (i in 1:n) {
    y[i] ~ dbin(p[i],ne[i])
    logit(p[i]) <- beta0+beta1*x[i]
  }
  #Priors informativas (conocimiento previo)
  beta0 ~ dnorm(-17.31,0.0009490187)
  beta1 ~ dnorm(2.57,0.04302926)
  #Estimacion y_hat
  for (i in 1:n) {
    y_hat[i] ~ dbin(p[i],ne[i])
  }
}
```

Figura 5: Liga Logit

```
bin_model3 <- function(){
  #Likelihood
  for (i in 1:n) {
    y[i] ~ dbin(p[i],ne[i])
    p[i] <- phi(beta0+beta1*x[i])
  }
  #Priors vagas
  beta0 ~ dnorm(0,0.001)
  beta1 ~ dnorm(0,0.001)
  #Estimacion y_hat
  for (i in 1:n) {
    y_hat[i] ~ dbin(p[i],ne[i])
  }
}

bin_model4 <- function(){
  #Likelihood
  for (i in 1:n) {
    y[i] ~ dbin(p[i],ne[i])
    p[i] <- phi(beta0+beta1*x[i])
  }
  #Priors informativas (conocimiento previo)
  beta0 ~ dnorm(-17.31,0.0009490187)
  beta1 ~ dnorm(2.57,0.04302926)
  #Estimacion y_hat
  for (i in 1:n) {
    y_hat[i] ~ dbin(p[i],ne[i])
  }
}
```

Figura 6: Liga Probit