PRONÓSTICO BAYESIANO PARA EL MODELO LOGÍSTICO

Por

Eliana Mangones Cervantes

Tesis sometida en cumplimiento parcial de los requisitos para el grado de:

MAESTRÍA EN CIENCIAS

en

MATEMÁTICAS (ESTADÍSTICA)

UNIVERSIDAD DE PUERTO RICO
RECINTO UNIVERSITARIO DE MAYAGÜEZ

2017

Aprobada por:

Edgardo Lorenzo González, Ph.D Miembro, Comité Graduado	Fecha
Pedro A. Torres Saavedra, Ph.D Miembro, Comité Graduado	Fecha
Dámaris Santana Morant, Ph.D Presidente, Comité Graduado	Fecha
Raúl E. Torres Muñiz, Ph.D Representante de Estudios Graduados	Fecha
Olgamary Rivera Marrero, Ph.D Directora del Departamento	Fecha

Resumen de Tesis Presentado a Escuela Graduada
de la Universidad de Puerto Rico como requisito parcial de los
Requerimientos para el grado de Maestría en Ciencias

PRONÓSTICO BAYESIANO PARA EL MODELO LOGÍSTICO

Por

Eliana Mangones Cervantes

Mayo 2017

Consejero: Dámaris Santana Morant, Ph.D
Departamento: Ciencias Matemáticas

El 28 de enero de 1986, la NASA lanzó al transbordador Challenger a cumplir una misión en el espacio. A 73 segundos del despegue el Challenger explotó dejando siete víctimas fatales. El accidente del Challenger se catalogó como uno de los peores desastres en la historia astronáutica. Es posible que el accidente se debiera a que las bajas temperaturas en la noche anterior y el día del lanzamiento ocasionaron daños en los aros que sellaban las diferentes etapas de los cohetes aceleradores sólidos del transbordador. La noche previa al accidente, los ingenieros que fabricaron el motor sólido del cohete, debatieron junto a los expertos de la NASA, el efecto que podría tener las bajas temperatura con relación al fallo de los aros. La discusión se basó en un conjunto de datos obtenidos de 23 lanzamientos previos al Challenger. Sin embargo, la conclusión de esa discusión fue que los datos que se tenían no eran concluyentes para predecir un posible fallo en los aros, y tomaron la decisión de no detener el lanzamiento. Los 23 lanzamientos previos al Challenger se hicieron en temperaturas entre $53°F$ y $81°F$, pero el Challenger fue

lanzado a una temperatura de $31°F$, esto es, $22°F$ menos de la temperatura mínima reportada en los lanzamientos previos.

Para la comunidad científica ha sido de interés el estimar la probabilidad de que el Challenger tuviera un accidente usando el conjunto de datos de los 23 lanzamientos previos. Se han desarrollado modelos de probabilidad, métodos de extrapolación y de pronóstico. Motivados por el mismo interés proponemos dos métodos con enfoque Bayesiano que tratan el problema como uno de pronóstico y desde un punto de vista de datos faltantes asumiendo que los datos faltantes siguen un patrón aleatorio (MAR). Mediante un estudio de simulación mostramos que ambos métodos son prometedores para analizar problemas de este tipo. Encontramos que el error cuadrático medio de los estimadores del modelo es menor en los dos métodos propuestos comparando con otro método Bayesiano y el método de máxima verosimilitud.

En general, se considera un modelo logístico con parámetro $\theta = (\alpha, \beta)$. El primer método que se propone usa la distribución posterior de $\theta = (\alpha, \beta)$ que se obtiene usando los datos observados para generar los datos faltantes con el fin de generar de la distribución posterior de $\theta^* = (\alpha^*, \beta^*)$ que se obtiene de los datos completos. Los datos completos se componen de los datos faltantes imputados y del remuestreo de los datos observados. Esto, para no usar directamente los datos observados en las dos estimaciones: la de θ y la de θ^*. El segundo método usa la distribución posterior de θ que se obtiene haciendo remuestreo de los datos observados para generar los datos faltantes con el fin de generar de la distribución posterior de θ^* que se obtiene de los datos completos. Los datos completos se componen de los datos faltantes imputados y de los datos observados. En ambos métodos, ya con la distribución posterior de θ^*, se puede estimar la probabilidad de éxito del modelo Binomial tras el modelo logístico.

Abstract of Thesis Presented to the Graduate School
of the University of Puerto Rico in Partial Fulfillment of the
Requirements for the Degree of Master of Sciences

BAYESIAN FORECASTING FOR LOGISTIC MODEL

By

Eliana Mangones Cervantes

May 2017

Chair: Dámaris Santana Morant, Ph.D
Major Department: Mathematical Sciences

On January 28, 1986, NASA launched the Space Shuttle Challenger to accomplish a mission in space. At 73 seconds off the Challenger exploded leaving seven fatal victims. The Challenger accident is listed as one of the worst disasters in astronaut history. The accident was possibly due to the low temperatures on the previous night and of the day of the launching that caused damages in the rings that sealed the different stages of the shuttle rockets. The night before the accident, the engineers who made the rocket's solid engine debated together with NASA experts the effect that low temperatures could have on the failure of the o-rings. The discussion was based on a set of data from 23 shuttles launches previous to the Challenger. The conclusion of that discussion was that the data was not conclusive to predict a possible failure in the o-rings, and they decided not to stop the launch. The 23 launches prior to the Challenger were made at temperatures between $53°F$ and $81°F$, but the Challenger was launched at a temperature of $31°F$, which is $22°F$ less than the lowest temperature reported in previous launches.

For the scientific community it has been of interest to estimate the likelihood of the Challenger accident using the data set from the previous 23 launches. Probability models, methods of extrapolation and pronostic have been developed. With the same motivation, we propose two methods with a Bayesian approach that treat the problem as forecasting and from a missing data point of view, assuming that the missing data follow a random pattern (MAR). Through of a small simulation study we showed that both methods are promising to analyze this type of problems. We found that the mean square error of the model estimators is lower in the two proposed methods compared to the other Bayesian method and the method of maximum likelihood.

In general, a logistic model with $\theta = (\alpha, \beta)$ parameter is considered. The first of the proposed method uses the posterior distribution of $\theta = (\alpha, \beta)$, which is obtained using the observed data to generate the missing data in order to generate the posterior distribution of $\theta^* = (\alpha^*, \beta^*)$, the parameter of the model for the complete data. The complete data consists of the imputed missing data and the resampling of the observed data. This, is done to avoid the use of the observed data in both the estimation of θ and θ^*. The second method uses the posterior distribution of θ which is obtained by resampling the observed data, to generate the missing data in order to generate the posterior distribution of θ^* that is obtained from the complete data. The complete data consists of the imputed missing data and the observed data. In both methods, and having the posterior distribution of θ^*, we can estimate the probability of success of the Binomial model that defines the logistic model.

A mi esposo, por ser mi aliado y mi cómplice de vida.

AGRADECIMIENTOS

Mis más emotivos agradecimientos a mis padres y hermanas por brindarme su apoyo moral y espiritual.

Agradezco a la Dra. Dámaris Santana Morant por su gran aporte, conocimiento, paciencia y dedicación a este trabajo.

A los doctores Pedro A. Torres y Edgardo Lorenzo por ser parte de mi comité graduado, y por contribuir a mi formación académica.

Al Dr. Luis F. Cáceres y todo el equipo de AFAMaC por darme la oportunidad de conocer otra faceta de la vida académica. Adémas, por ayudarme a completar este proceso.

Al Departamento de Ciencias Matemáticas de la Universidad de Puerto Rico del Recinto Universitario de Mayagüez por darme la oportunidad de crecer académicamente y brindarme su apoyo.

A todos mis amigos porque gracias a su compañía hicieron este proceso más llevadero.

Mil gracias a todos.

TABLA DE CONTENIDO

página

LISTA DE TABLAS

Tabla página

LISTA DE FIGURAS

Figura página

CAPÍTULO 1
INTRODUCCIÓN

El 28 de enero de 1986 el transbordador Challenger explotó 73 segundos después del despegue. Se cree que las causas del accidente se debieron a que la baja temperatura el día del lanzamiento provocó fallos en los aros de los motores sólidos de los cohetes.

Para determinar la probabilidad de que al menos uno de los seis aros fallara se ha utilizado un conjunto de datos tomado de 23 vuelos previos al Challenger. El conjunto de datos tiene registrado el número de aros que fallaron y la temperatura en el momento del lanzamiento. El problema central es que para los 23 vuelos previos, los lanzamientos se hicieron en temperaturas entre $53°F$ y $81°F$, mientras que el Challenger fue lanzado a $31°F$, siendo esta una temperatura muy baja respecto a los vuelos anteriores.

Los datos relacionados a los 23 lanzamientos, han sido utilizados para determinar la probabilidad de que cada aro sufriera fallo a $31°F$. Dalal et al.,(1989), usaron el modelo logístico y construyeron un intervalo de confianza para los parámetros del modelo utilizando un procedimiento de bootstrap paramétrico. Lavine,(1991), propuso un análisis de extrapolación con métodos no paramétricos y Maranzano y Krzysztofowicz, (2008), desarrollaron un modelo de pronóstico usando la fórmula de Bayes. Hicieron un análisis de extrapolación que consiste en añadir información sobre el posible número de aros dañados en la temperatura $31°F$.

En este trabajo se propone una metodología de pronóstico con enfoque Bayesiano tomando las características del conjunto de datos del Challenger. El objetivo es determinar como se puede pronosticar un evento futuro dado unas observaciones previas por medio de simulaciones.

Para pronosticar eventos futuros la idea es enfocar el problema como uno de datos faltantes. En el caso del Challenger los datos faltantes serían temperaturas menores a $53°F$ con el número de aros que fallaron a esas temperaturas. Se hace imputación en los datos faltantes y se obtiene un conjunto de datos completos. Con los datos completos se estiman los parámetros del modelo y se estima la probabilidad de que al menos un aro falle a $31°F$.

Para probar la metodología propuesta se elaboró un estudio de simulación y posteriormente se implementó en el conjunto de datos del Challenger.

Por la naturaleza de los datos se ajustó el modelo logístico para datos binomiales, y se estimaron los coeficientes de la regresión mediante el método de máxima verosimilitud, el método propuesto por Gelman et al. (2014) y los dos métodos propuestos en este trabajo. Una vez estimados los coeficientes de la regresión, se estimó la probabilidad de que cada aro fallara a una temperatura de $31°F$ para los datos del Challenger y en $45°F$ para los datos del estudio de simulación.

Los objetivos de este trabajo son:

- Proponer dos métodos de pronóstico bajo un enfoque Bayesiano con datos binomiales.
- Comparar los métodos propuestos con el método de máxima verosimilitud del modelo logístico y el método propuesto por Gelman et al. (2014) en términos de pronóstico

- Pronosticar la probabilidad de que al menos uno de los aros del Challenger fallara a $31°F$.

Este trabajo está organizado de la siguiente manera. En el Capítulo 2 se presenta la revisión de literatura que se utilizó para desarrollar los métodos propuestos. Se describen las características del análisis Bayesiano y el modelo de regresión logística con datos binomiales. En el Capítulo 3 se describen cada uno de los métodos implementados en los conjuntos de datos simulados y en la aplicación. En el Capítulo 4 se presenta el estudio de simulación con sus respectivos resultados y la aplicación de la metodología a los datos del Challenger. En el Capítulo 5 se presentan las conclusiones de este trabajo así como los trabajos futuros.

CAPÍTULO 2
REVISIÓN DE LITERATURA

2.1 Análisis Bayesiano

Dado un modelo $f(y|x,\theta)$ para los datos observados $y = (y_1, y_2, ..., y_n)$ y dado un vector de parámetros desconocidos $\theta \in \Theta$ y x un vector de covariables, el enfoque Bayesiano considera a θ como una variable aleatoria. Esta variable aleatoria adopta una distribución de probabilidad que resume cualquier información que tenemos sobre ella no relacionada con la proporcionada por los datos y que se conoce como distribución a priori $\pi(\theta)$ (Robert, 2007).

La inferencia concerniente a θ está basada en su distribución posterior, dada por

$$\pi(\theta|y) = \frac{f(y|x,\theta)\pi(\theta)}{\int f(y|x,\theta)\pi(\theta)d\theta}, \tag{2.1}$$

donde $\int f(y|x,\theta)\pi(\theta)d\theta$ es la distribución marginal de y. La ecuación 2.1 se conoce como el Teorema de Bayes. Si θ tiene una distribución discreta, la integral es reemplazada por una suma. Como el denominador en esta ecuación no depende de θ, entonces, la distribución posterior $\pi(\theta|y)$ es proporcional al producto de la función de verosimilitud de θ $f(y|x,\theta)$, y la distribución a priori $\pi(\theta)$, es decir

$$\pi(\theta|y) \propto f(y|x,\theta)\pi(\theta). \tag{2.2}$$

Distribución a Priori $\pi(\theta)$

La distribución a priori es una parte importante del enfoque Bayesiano. Representa la información sobre un parámetro incierto que se combina con la distribución de probabilidad de nuevos datos para obtener la distribución posterior, la cual a su vez se utiliza para futuras inferencias y decisiones. Por esto la elección de la distribución a priori es uno de los puntos críticos en el análisis Bayesiano. En la práctica rara vez la información previa que se tiene acerca de los parámetros es precisa, por lo tanto no se llega a una determinada distribución a priori, en el sentido de que muchas distribuciones de probabilidad pueden ser compatibles con la información que se tiene.

Algunos autores consideran que $\pi(\theta)$ debería ser no informativa o difusa, de tal manera que en el análisis los datos sean quienes proporcionen la información. Parecería que la inferencia bayesiana sería inapropiada en tales contextos pero se podría argumentar que toda la información que resulta de la distribución posterior surgió de los datos, y por lo tanto, las inferencias resultantes fueran objetivas en lugar de subjetivas.

Ejemplos de distribuciones a priori son la distribución uniforme que asigna la misma probabilidad a cada valor de los parámetros y la distribución de Jeffreys que está basada en la matriz de información de Fisher. La matriz de información de Fisher es un indicador de la cantidad de información de θ traida por el modelo (Robert, 2007).

Las distribuciones a priori conjugadas son otra alternativa de representar la información previa sobre los parámetros. Si una distribución a priori para θ pertenece a una familia de distribuiones paramétricas F, entonces, la distribución a priori se denomina conjugada si la distribución posterior de θ también pertenece a la familia de distribuciones paramétricas F (Jackman, 2009).

Las distribuciones a priori conjugadas tienen una ventaja práctica, además de una conveniencia computacional.

Distribución Posterior $\pi(\theta|y)$

Una distribución posterior comprende una distribución a priori sobre un parámetro y un modelo de verosimilitud que proporciona información sobre el parámetro basándose en los datos observados. Dependiendo del modelo de distribución verosimilitud elegido, la distribución posterior se puede calcular analíticamente o de forma aproximada, por ejemplo, usando métodos Monte Carlo de Cadenas de Markov (MCMC, por sus siglas en inglés),(Robert, 2007).

La inferencia bayesiana utiliza la distribución posterior para hacer estimaciones de los parámetros, incluyendo estimaciones puntuales tales como medias posteriores, medianas, percentiles y estimaciones de intervalo conocidas como intervalos de credibilidad.

Distribución Predictiva Posterior

Para hacer inferencia sobre un valor desconocido pero observable, se sigue una lógica similar. Antes de considerar los datos y, la distribución de y desconocida pero observable está dada por:

$$f(y) = \int f(y, x, \theta)d\theta = \int f(y|x, \theta)\pi(\theta)d\theta \tag{2.3}$$

A menudo se denomina distribución marginal de y o distribución predictiva a priori: a priori porque no está condicionada a una observación previa del proceso, y predictiva porque es la distribución de una cantidad que es observable.

Una vez que los datos y han sido observados, se puede predecir una cantidad desconocida observable $\tilde{y}$. La distribución de $\tilde{y}$ se conoce como distribución predictiva posterior (Gelman et al., 2014). La distribución predictiva para una nueva observación $\tilde{y}$ está dada por:

$$\begin{aligned} f(\tilde{y}|y) &= \int f(\tilde{y}, \theta|y) d\theta \\ &= \int f(\tilde{y}|\theta, y)\pi(\theta|y) d\theta \\ &= \int f(\tilde{y}|\theta)\pi(\theta|y) d\theta \end{aligned} \tag{2.4}$$

Se asume que $\tilde{y}$ y y son independientes.

2.2 Métodos Monte Carlo de Cadenas Markov (MCMC)

Los métodos Monte Carlo de Cadenas de Markov son métodos de simulación que permiten generar de manera iterativa valores de parámetros, cuya distribución estacionaria es exactamente la distribución posterior que se interesa calcular. Los métodos Monte Carlo son utilizados cuando no es posible o no es eficiente muestrear θ directamente de $\pi(\theta|y)$ ya sea con métodos analíticos o numéricos (Little y Rubin, 1987).

Los valores para θ son generados secuencialmente y cada uno de ellos depende solo del valor anterior que fue generado. La muestra de valores que se obtiene forman una Cadena de Markov.

2.2.1 Metropolis-Hastings

Metropolis-Hasting es un método MCMC que sigue una regla de aceptación y rechazo que converge a una distribución objetivo específica. La idea es muestrear de una distribución candidata, también conocida como distribución de saltos $q_t(\theta^*|\theta^{t-1})$. Supongamos que se quiere generar de una distribución posterior conjunta

$$p(\theta|y) \propto h(\theta) \equiv f(y|\theta)\pi(\theta). \tag{2.5}$$

Inicialmente se específica la distribución candidata $q_t(\theta^*|\theta^{t-1})$ de la cual es fácil simular.

Dado un valor inicial $\theta^{(0)}$ en la iteración $t = 0$, el algoritmo procede de la siguiente manera.

Para $t = 1, \cdots, T$, se repite:

1. Simular un valor θ^* de la distribución candidata $q_t(\theta^*|\theta^{t-1})$.
2. Calcular la razón
$$R = \frac{h(\theta^*)q(\theta^{t-1}|\theta^*)}{h(\theta^{t-1})q(\theta^*|\theta^{t-1})}. \tag{2.6}$$
3. Calcular la probabilidad de aceptación $p = min\{R, 1\}$.
4. Entonces se define
$$\theta^t = \begin{cases} \theta^* & \text{con probabilidad } p, \\ \theta^{t-1} & \text{de otro modo.} \end{cases} \tag{2.7}$$

En caso de que la distribución q sea simétrica, es decir

$$q(\theta^*|\theta^{t-1}) = q(\theta^{t-1}|\theta^*) \tag{2.8}$$

entonces la razón R se simplifica a

$$R = \frac{h(\theta^*)}{h(\theta^{t-1})}. \tag{2.9}$$

Al implementar el algoritmo se genera una secuencia que converge a la distribución posterior de interés. Para evaluar la convergencia, se grafican los promedios estimados de θ. Esta gráfica se conoce como *runnings means*, en ella se observa como la secuencia de promedios se estabiliza a medida que aumenta el número de simulaciones (Robert y Casella, 2009).

Una vez el algoritmo coverge, generalmente se descarta una porción inicial de la secuencia conocida como *burn-in period*, la parte de la secuencia que aún tiene la

distribución de interés. El *burn-in period* se escoge analizando la convergencia en las gráficas de los *runnings means*.

2.3 Modelo de Regresión Logística para Datos Binomiales

Datos cuya variable respuesta se puede modelar con la distribución binomial con una o más covariables, pueden ser analizados a través de un modelo lineal generalizado con una función de enlace específica (Agresti, 1996).

Sea $Y = \{y_1, y_2, \cdots, y_n\}$ variables aleatorias binomiales independientes, es decir, $y_i \sim \text{Bin}(m_i, p_i)$ donde m_i se conoce como el número de ensayos Bernoulli y p_i son las probabilidades de éxito con $0 \leq p_i \leq 1$. Sea $X = \{x_1, x_2, \cdots, x_n\}$ un vector de variables predictoras. La función de probabilidad de y_i viene dada por

$$f(y_i|p_i) = \binom{m_i}{y_i} p_i^{y_i}(1-p_i)^{m_i-y_i}; \quad y_i = 0, 1, \cdots, m_i, \tag{2.10}$$

para $m_i \geq 1$, $i = 1, 2, \cdots, n$.

Para el modelo de regresión logística se define $p_i = g^{-1}(x_i'\theta)$, por lo que $g(p_i) = x_i'\theta = \eta_i$ donde θ es el vector desconocido de coeficientes de la regresión, η_i es el predictor lineal y g es la función de enlace. Usualmente se usa la función de enlace logit que se define como

$$g(p_i) = \log\{p_i/(1-p_i)\} \Rightarrow p_i = \frac{e^{x_i'\theta}}{1+e^{x_i'\theta}}. \tag{2.11}$$

Esta función de enlace es inyectiva y simétrica, con $g : (0,1) \longrightarrow \Re$. También es monótona y diferenciable en el intervalo $(0,1)$. La curva del modelo de regresión logística se muestra en la Figura 2–1. El eje x corresponde a una variable predictora y el eje y es el vector de probabilidades.

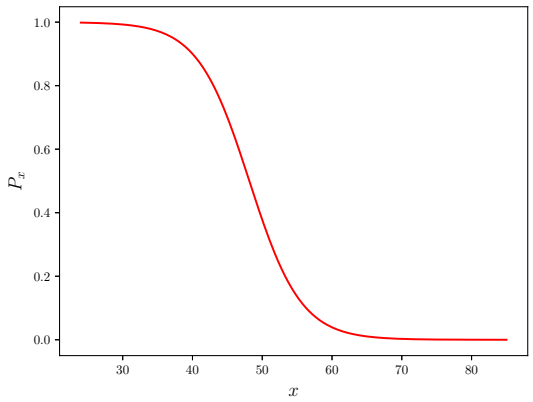


Figura 2–1: Curva del modelo de regresión logística.

La función de distribución de Y está dada por:

$$f(Y|\theta) = \prod_{i=1}^{n} \binom{m_i}{y_i} \left(\frac{e^{x_i'\theta}}{1+e^{x_i'\theta}}\right)^{y_i} \left(\frac{1}{1+e^{x_i'\theta}}\right)^{m_i-y_i} \tag{2.12}$$

2.3.1 Estimador de Máxima Verosimilitud

Una técnica para estimar los parámetros de un modelo es máxima verosimilitud. Consiste en maximizar la función de verosimilitud asociada a una distribución de probabilidad. Sea $Y = \{y_1, \cdots, y_n\}$ una muestra de variables independientes con función de densidad $f(y_i|\theta)$, donde θ es un vector de parámetros. La función de verosimilitud está dada por:

$$L(\theta|Y) = f(Y|\theta) = f(y_1, \cdots y_n) \tag{2.13}$$

$$= \prod_{i=1}^{n} f(y_i|\theta). \tag{2.14}$$

La idea es maximizar la ecuación 2.14 con respecto a θ. En muchos casos resulta más fácil maximizar el logaritmo de la función de verosimilitud (log-verosimilitud)

$$l(\theta|Y) = \log L(\theta|Y) = \sum_{i=1}^{n} \log f(y_i|\theta). \tag{2.15}$$

En el caso del modelo de regresión logística para datos binomiales la función de log-verosimilitud se define como:

$$l(Y|\theta) = \sum_{i=1}^{n} \log \left[\binom{m_i}{y_i} \left(\frac{e^{x_i'\theta}}{1+e^{x_i'\theta}} \right)^{y_i} \left(\frac{1}{1+e^{x_i'\theta}} \right)^{m_i - y_i} \right] \tag{2.16}$$

No hay solución analítica al maximizar la ecuación 2.16, por lo tanto se recurre a métodos numéricos iterativos.

2.3.2 Enfoque Bayesiano para el Modelo Logístico con datos Binomiales

Se considera una distribución a priori uniforme no informativa para los coeficientes de la regresión θ, esto es $\pi(\theta) \propto 1$, entonces la distribución posterior $\pi(\theta|Y, X)$ es proporcional a la función de verosimilitud

$$\pi(\theta|Y, X) \propto \prod_{i=1}^{n} \binom{m_i}{y_i} \left(\frac{e^{x_i'\theta}}{1+e^{x_i'\theta}} \right)^{y_i} \left(\frac{1}{1+e^{x_i'\theta}} \right)^{m_i - y_i}. \tag{2.17}$$

La distribución posterior es propia si y solo si

$$\int f(Y|\theta)\pi(\theta)d\theta < \infty. \tag{2.18}$$

Roy y Kaiser (2013), demuestran que la condición anterior se cumple para diferentes funciones de enlace, entre estas el enlace logit cuando se usa una distribución a priori uniforme no informativa $\pi(\theta) \propto 1$.

Para el modelo de regresión logística con datos binomiales Roy y Kaiser (2013) desarrollan el algoritmo Metropolis-Hasting tomando una distribución normal multivariada escalonada para generar de la distribución posterior de los parámetros del modelo. Esto es

$$q(\theta^*|\theta^{t-1}) = N_d(\theta^*|\theta^{t-1}, c^2\Sigma). \tag{2.19}$$

Gelman et al. (1996), proponen que una escala óptima para el algoritmo Metropolis con esta distribución candidato está dada por $c \approx 2.4/\sqrt{d}$, donde d es el número

de parámetros a estimar. Σ es la matriz de covarianza del estimador de máxima verosimilitud que se escoge como el valor inicial.

El algoritmo Metropolis-Hasting queda, entonces, de la siguiente manera: para la iteración t=0: se calcula el estimador de máxima verosimilitud $\hat{\theta}$ y la matriz de covarianza $\hat{\Sigma}$ correspondiente a $\hat{\theta}$, ajustando el modelo logístico, y sea $\theta^0 = \hat{\theta}$.

Para $t = 1, \cdots, T$, se repite:

1. Generar $\theta^* \sim N_d(\theta^{t-1}, c^2\hat{\Sigma})$ donde $c \approx 2.4/\sqrt{d}$.
2. Calcular la razón
$$R = \frac{\pi(\theta^*|y,x)}{\pi(\theta^{t-1}|y,x)}. \tag{2.20}$$
3. Calcular la probabilidad de aceptación $p = min\{R, 1\}$.
4. Entonces se define
$$\theta^t = \begin{cases} \theta^* & \text{con probabilidad } p, \\ \theta^{t-1} & \text{de otro modo.} \end{cases} \tag{2.21}$$

2.4 Datos Faltantes

Un problema común en el análisis de datos surge cuando se obtienen muestras con datos faltantes. De acuerdo a la razón de la ausencia, los datos faltantes se clasifican como, datos faltantes aleatorios (*Missing at Random*, MAR), datos faltantes completamente aleatorios (*Missing Completely at Random*, MCAR) y datos faltantes no aleatorios (*Missing not at random*, MNAR).

Sea $Y^c = (Y_{obs}, Y_{falt})$ la variable de interés, donde Y_{obs} y Y_{falt} corresponde al vector de datos observados y faltantes respectivamente. Se afirma que un proceso de datos omitidos se genera en forma aleatoria MAR si la distribución de los datos observados Y_{obs} es diferente a la de los datos faltantes Y_{falt}, pero estos pueden explicarse completamente por otras variables observadas. En el proceso en forma completamente aleatoria MCAR, la distribución de los datos observados Y_{obs} luce

igual a la distribución de los datos faltantes Y_{falt}, por lo tanto no hay diferencia entre los datos faltantes y los observados. Por otro parte, MNAR quiere decir que la probabilidad de que una respuesta a una variable sea dato faltante es dependiente de los valores de la variable. Es común referirse a los procesos MAR como mecanismos de no respuesta ignorable, en tanto que MNAR significa que la falta de respuesta no puede ser ignorada (Little y Rubin, 1987).

En la literatura se presentan varios métodos para el análisis estadístico con datos faltantes. Por ejemplo, el análisis de casos completos que consiste en trabajar solamente con las observaciones que tienen la información completa, así que, se ignoran las observaciones que tienen información incompleta. Esto elimina información potencialmente importante, por lo que surgen métodos que retienen toda la información recolectada y reemplazan los datos faltantes por datos sustitutos. Existen dos modalidades para estimar las unidades faltantes en la muestra: imputación simple y múltiple.

2.4.1 Imputación Simple

La imputación simple consiste en reemplazar un valor faltante por otro. Hay dos enfoques para estimar las unidades faltantes, los modelos explícitos y los modelos implícitos (Little y Rubin, 1987). Los métodos que incluyen modelos explícitos son los siguientes:

- Imputación por media no condicionada: consiste en sustituir cada uno de los datos faltantes en la muestra por la media de las unidades observadas en dicha muestra.
- Imputación por media condicionada: si se clasifican los datos faltantes y no faltantes en distintas clases o grupos, se imputa la media de los datos observados para los valores faltantes en cada clase.
- Imputación por regresión: consiste en sustituir los valores faltantes por valores predichos de una regresión de los datos faltantes en las unidades observadas.

- Imputación por regresión estocástica: consiste en añadir al valor predicho por la regresión un ente aleatorio ξ con media 0 y varianza igual a la varianza residual de la regresión basada en los datos completos.
- Método de Buck: este método se aplica cuando existe un patrón en los datos faltantes, para el caso donde las variables faltantes tengan una regresión lineal en las variables observadas. El método primero estima la media y la matriz de covarianza basada en los casos completos y después usa estos estimados para calcular la ecuación lineal por mínimos cuadrados de las variables faltantes en las variables presentes para cada patrón de datos faltantes.

En los modelos implícitos se requiere de un algoritmo para estimar los datos faltantes. Los métodos que implican modelos implícitos son los siguientes:

- Cold Deck: consiste en sustituir el dato faltante por un valor constante obtenido de una fuente externa. Por ejemplo, si nos referimos a una encuesta un valor faltante en esta encuesta puede ser reemplazado por un valor obtenido en una misma encuesta hecha previamente.
- Hot Deck por muestreo aleatorio simple con reemplazamiento: este método consiste en sustituir un dato faltante por otro que es escogido aleatoriamente y con reemplazo de los datos completos.
- Hot Deck en celdas ajustadas: los datos faltantes en cada celda son sustituidos por datos observados en la misma celda.
- Hot Deck por vecino más cercano: para este método es necesario definir una métrica que mida la distancia entre las unidades. El dato faltante se sustituye por otro a partir de la distancia calculada a través de una variable con información completa.
- Hot Deck secuencial ordenado por una covariable: el procedimiento secuencial Hot Deck es usado cuando la muestra tiene algún tipo de orden dentro de cada grupo

de clasificación y para cada uno de los valores faltantes, el valor previo registrado es duplicado.

2.4.2 Imputación Múltiple

Otro método para reemplazar datos faltantes por sustitutos se conoce como imputación múltiple. En lugar de reemplazar cada uno de los datos faltantes por un solo valor, el procedimiento de imputación múltiple reemplaza cada valor faltante por un conjunto de valores.

Si se tienen m conjuntos de datos a imputar en cada valor faltante, entonces, se obtendrán m conjuntos de datos completos, donde cada uno de ellos tiene un valor posible para cada uno de los valores faltantes.

Una vez se desarrolle un análisis estadístico en cada una de las bases de datos completos, se procede a combinar los resultados con el fin de obtener conclusiones finales. La combinación de los resultados sigue las reglas propuestas por Little y Rubin (1987), que se resumen a continuación.

Sea θ el parámetro de interés, $\hat{\theta}$ su estimación puntual y U la varianza estimada de $\hat{\theta}$. Tras analizar los datos imputados tenemos m estimaciones $\hat{\theta}_1, ..., \hat{\theta}_m$ con varianzas estimadas asociadas $\hat{U}_1, ..., \hat{U}_m$.

La estimación puntual para θ basada en imputación múltiple, $\bar{\theta}$, viene dada por:

$$\bar{\theta} = \frac{1}{m} \sum_{i=1}^{m} \hat{\theta}_i. \tag{2.22}$$

La varianza estimada asociada con $\bar{\theta}$ tiene dos componentes. La varianza dentro de la imputación, que es la varianza promedio de las varianzas estimadas de los datos completos,

$$\bar{U} = \frac{1}{m} \sum_{i=1}^{m} \hat{U}_i. \tag{2.23}$$

y la varianza entre imputaciones, que es la varianza de las estimaciones con los datos completos,

$$B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\theta}_i - \bar{\theta})^2. \tag{2.24}$$

Así la varianza total es definida como

$$T = \bar{U} + (1 - m^{-1})B. \tag{2.25}$$

2.5 Bootstrap

Bootstrap es una técnica de remuestreo propuestas por Efron (1979), que se caracteriza por obtener submuestras de una muestra tomada de la población. Permite estimar la distribución muestral de un estadístico y también propiedades de un estimador como el sesgo o el error estándar.

Dada una muestra observada $X = \{x_1, \cdots, x_n\}$ se generan j submuestras $X_1^*, X_2^* \cdots, X_j^*$. Las unidades de las submuestras son seleccionadas aleatoriamente con reemplazo de la muestra X. Cada una de las submuestras se conoce como muestra bootstrap y tendrá el mismo número de unidades de X. Una vez generadas las muestras bootstrap se puede estimar el error estadístico, así como estimaciones de los parámetros de interés.

CAPÍTULO 3
METODOLOGÍA

Dado el conjunto de datos observados (Y_{obs}, X_{obs}), donde $Y_{obs} = \{y_1, y_2, \cdots, y_n\}$ es el vector de variables respuesta que son independientes con $y_i \sim Bin(m_i, p_i)$, y, $X_{obs} = \{x_1, x_2, \cdots, x_n\}$ el vector de variables predictoras. La metodología se centra en pronosticar la probabilidad de éxito p_h para $y_h|x_h$ que no son observados y $x_h < x_i$, $i = 1, \cdots, n$. Como estamos tomando las características del conjunto de datos del Challenger, consideramos $m_i = 6$, $i = 1, \cdots, n$ y X como un vector de temperaturas donde la temperatura mínima es $53°F$.

Para pronosticar la probabilidad p_h, trataremos el problema desde el punto de vista de datos faltantes asumiendo que los datos faltantes siguen un patrón MAR. Usando el modelo logístico y los datos observados la distribución de los datos faltantes depende de los datos observados. A menores temperaturas el número de aros dañados tiende a aumentar.

Sea (Y^c, X^c) el conjunto de datos completos donde $Y^c = (Y_{falt}, Y_{obs})$ y $X^c = (X_{falt}, X_{obs})$, tal que (Y_{falt}, X_{falt}) corresponden a los datos faltantes o no observados. Los datos en el vector X_{falt} son menores a los datos de X_{obs}, o sea menores a $53°F$ pero mayores que $24°F$. La idea es imputar (Y_{new}, X_{new}) en (Y_{falt}, X_{falt}) para obtener un conjunto de datos completos y de esta forma estimar los parámetros del modelo. Una vez estimados los parámetros del modelo se estima la probabilidad p_h.

La Figura 3–1 muestra gráficamente la idea general de la metodología del trabajo.

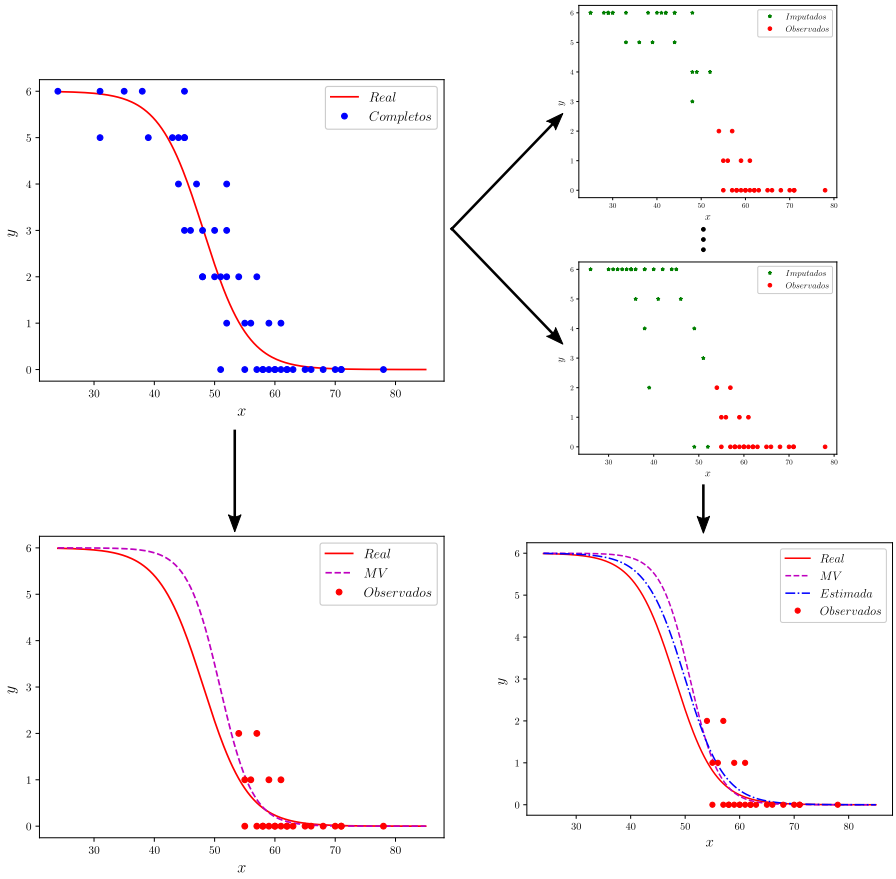


Figura 3–1: Diagrama de la Metodología

En esta figura la gráfica superior izquierda muestra un conjunto de datos completos (puntos azules) y la curva logística real, de la cual fueron generados los datos. Esta es la curva que deseamos estimar cuando no vemos los datos completos. La gráfica inferior izquierda muestra solo parte del conjunto de datos de arriba. Estos son los datos observados, los que tendremos disponibles al momento de hacer la estimación. Por ejemplo, se muestra la curva logística estimada con el método de Máxima Verosimilitud (MV) con estos datos observados. Las dos gráficas superiores a la derecha, presentan dos ejemplos de lo que harían nuestros métodos en los que luego de la imputacón (puntos verdes) se obtienen dos conjuntos de datos completos (puntos rojos y verdes). La gráfica

inferior derecha muestra en azul la curva estimada luego de usar múltiples conjuntos de datos completos como estos. Vemos que nuestra curva estimada está más cerca de la real comparada con la estimada con el método MV.

Proponemos dos métodos, los cuales compararemos con el método de máxima verosimilitud para el modelo de regresión logística donde no se tiene en cuenta el enfoque de datos faltantes y con el propuesto por Gelman et al. (2014). A continuación presentamos cada uno de éstos.

3.1 Método de Máxima Verosimilitud (MV)

Para los datos observados (Y_{obs}, X_{obs}) se ajusta el modelo de regresión logística para datos binomiales. Como solo hay una variable predictora entonces hay dos coeficientes de la regresión para estimar y $\theta = (\alpha, \beta)$. Se tiene α que corresponde al intercepto y β es el coeficiente de la regresión de la variable X_{obs}.

Los parámetros son estimados usando el método de máxima verosimilitud, y se utiliza la función *glm()* del paquete estadístico R para hacer los cálculos. Para máximizar la función de verosimilitud R utiliza el método iterativo de mínimos cuadrados reponderados iterativos.

Una vez se obtienen las estimaciones $\hat{\alpha}$ y $\hat{\beta}$, para un dato no observado x_h se calcula la probabilidad

$$p_h = \frac{e^{\hat{\alpha}+\hat{\beta}x_h}}{1 + e^{\hat{\alpha}+\hat{\beta}x_h}}. \tag{3.1}$$

3.2 Método de Gelman et al. (2014)

Cuando se tiene un patrón de datos faltantes MAR, Gelman et al. (2014) proponen un *Gibbs Sampler* para generar de la distribución posterior de los parámetros de un modelo, dado un conjunto de datos observados Y_{obs}.

En cada iteración del algoritmo *Gibbs Sampler* se siguen los siguientes pasos:

1. Imputar observaciones nuevas Y_{new} en lugar de los datos faltantes Y_{falt} dado (Y_{obs}, θ) donde $\theta = (\alpha, \beta)$ es el vector de parámetros del modelo.
2. Generar estimaciones de los parámetros del modelo dado (Y_{new}, Y_{obs}).

Esto quiere decir que en cada iteración del algoritmo se completan los datos y se estiman los parámetros del modelo de manera iterativa.

Se implementa la idea propuesta por Gelman et al. (2014) para el modelo logístico con datos binomiales. Se genera de la distribución posterior de los parámetros α y β del modelo utilizando el algoritmo Metropolis-Hasting con una distribución normal multivariada como candidato. Cada iteración del algoritmo Metropolis-Hasting queda de la siguiente manera:

Para la iteración t=0: se calcula el estimador de máxima verosimilitud de $\hat{\alpha}$ y $\hat{\beta}$, la matriz de covarianza $\hat{\Sigma}$ correspondiente al modelo logístico y sea $\alpha^0 = \hat{\alpha}$, $\beta^0 = \hat{\beta}$.

Para $t = 1, \cdots, T$, se repite:

1. Generar k valores enteros aleatorios x_k^t, tal que, $x_k^t \sim U(24, min(X_{obs}))$. A estas nuevas observaciones se le denominará X_{new}.
2. Predecir k valores $y_k^t | x_k^t$, tal que $y_k^t | x_k^t \sim Bin(6, p_k^t)$, donde

$$p_k^t = \frac{e^{\alpha^{t-1} + \beta^{t-1} x_k^t}}{1 + e^{\alpha^{t-1} + \beta^{t-1} x_k^t}}. \tag{3.2}$$

   A estas nuevas observaciones se le denominará Y_{new}.
3. Con el conjunto de datos completos (Y^c, X^c) donde $Y^c = (Y_{new}^t, Y_{robs}^t)$ y $X = (X_{new}^t, X_{robs}^t)$, el algoritmo genera α_t^* y β_t^*.

Una vez el algoritmo converge, y ya con la distribución posterior de α^* y β^* y dado x_h un dato no observado se calcula

$$\hat{\alpha} = \frac{1}{T} \sum_{t=1}^{T} \alpha_t^*, \quad \hat{\beta} = \frac{1}{T} \sum_{t=1}^{T} \beta_t^* \quad \text{y} \quad p_h^* = \frac{1}{T} \sum_{t=1}^{T} \frac{e^{\alpha_t^* + \beta_t^* x_h}}{1 + e^{\alpha_t^* + \beta_t^* x_h}} \tag{3.3}$$

Un diagrama del método se muestra en la Figura 3–2.

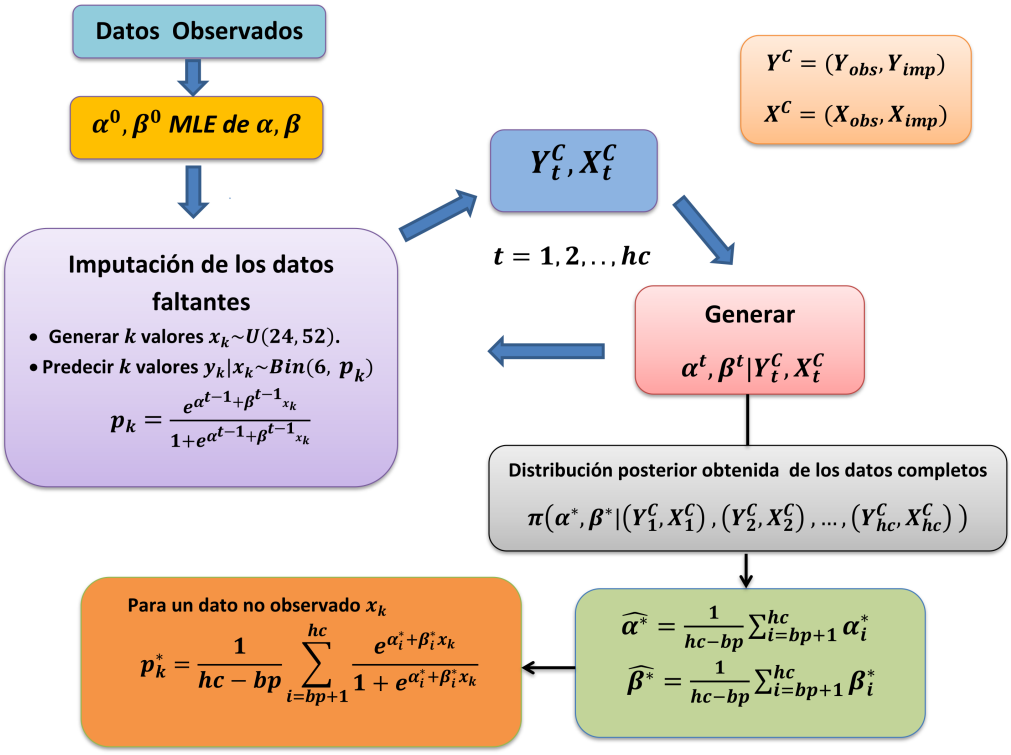


Figura 3–2: Diagrama para el método Gelman et al.

3.3 Método usando la Distribución Posterior de los Datos Observados ($\mathbf{Post}_{obs}$)

La idea es usar la distribución posterior de $\theta = (\alpha, \beta)$ que se obtiene usando los datos observados para generar los datos faltantes con el fin de generar de la distribución posterior de $\theta^* = (\alpha^*, \beta^*)$ que se obtiene de los datos completos para estimar p_h. Los datos completos se componen de los datos faltantes imputados y del remuestreo de los datos observados. Esto, para no usar los datos observados en las dos estimaciones: la de θ y la de θ^*.

Dado un conjunto de datos observados (Y_{obs}, X_{obs}) se propone la siguiente metodología:

1. **Obtener la distribución posterior de α y β con los datos observados**: Generar de la distribución posterior de α y β del modelo logístico. Esta distribución posterior se genera con los datos observados y el algoritmo Metropolis-Hasting descrito en el Capítulo 2, observando la convergencia. Se realizan 110,000 iteraciones del algoritmo y se eliminan las primeras 10,000. Por lo tanto, quedan $T = 100,000$ iteraciones del algoritmo.
2. Para cada x_k no observados se genera la distribución de probabilidad P_k, usando la distribución posterior de α y β con los datos observados que se obtuvo en el paso anterior. Para esto se calcula

$$P_k = \frac{e^{\alpha+\beta x_k}}{1 + e^{\alpha+\beta x_k}} \tag{3.4}$$

   para $t = 1, \cdots, T$.
3. **Obtener la distribución posterior de α^* y β^* con los datos completos**: Para generar de la distribución posterior con los datos completos cada iteración del algoritmo Metropolis-Hasting queda de la siguiente manera:

   Para la iteración t=0: se calcula el estimador de máxima verosimilitud de $\hat{\alpha}$ y $\hat{\beta}$, la matriz de covarianza $\hat{\Sigma}$ correspondiente al modelo logístico y sea $\alpha^0 = \hat{\alpha}$, $\beta^0 = \hat{\beta}$.

   Para $t = 1, \cdots, T$, se repite:

   (a) Generar k valores enteros aleatorios x_k^t, tal que, $x_k^t \sim U(24, min(X_{obs}))$. A estas nuevas observaciones se le denominará X_{new}^t.

   (b) Predecir k valores $y_k^t|x_k^t$, tal que $y_k^t|x_k^t \sim Bin(6, p_k^t)$, donde p_k^t se selecciona aleatoriamente con reemplazo de la distribución de probabilidades generada en el paso 2. A estas nuevas observaciones se le denominará Y_{new}^t.

   (c) Hacer bootstrap en (Y_{obs}, X_{obs}) para obtener una muestra (Y_{robs}^t, X_{robs}^t)

(d) Con el conjunto de datos completos (Y^c, X^c) donde $Y^c = (Y^t_{new}, Y^t_{robs})$ y $X = (X^t_{new}, X^t_{robs})$, el algoritmo genera α^*_t y β^*_t.

Una vez el algoritmo converge, y ya con la distribución posterior de α^* y β^* y dado x_h un dato no observado se calcula

$$\hat{\alpha} = \frac{1}{T}\sum_{t=1}^{T}\alpha^*_t, \quad \hat{\beta} = \frac{1}{T}\sum_{t=1}^{T}\beta^*_t \quad \text{y} \quad p^*_h = \frac{1}{T}\sum_{t=1}^{T}\frac{e^{\alpha^*_t+\beta^*_t x_h}}{1+e^{\alpha^*_t+\beta^*_t x_h}} \tag{3.5}$$

Un diagrama del método se muestra en la Figura 3–3.

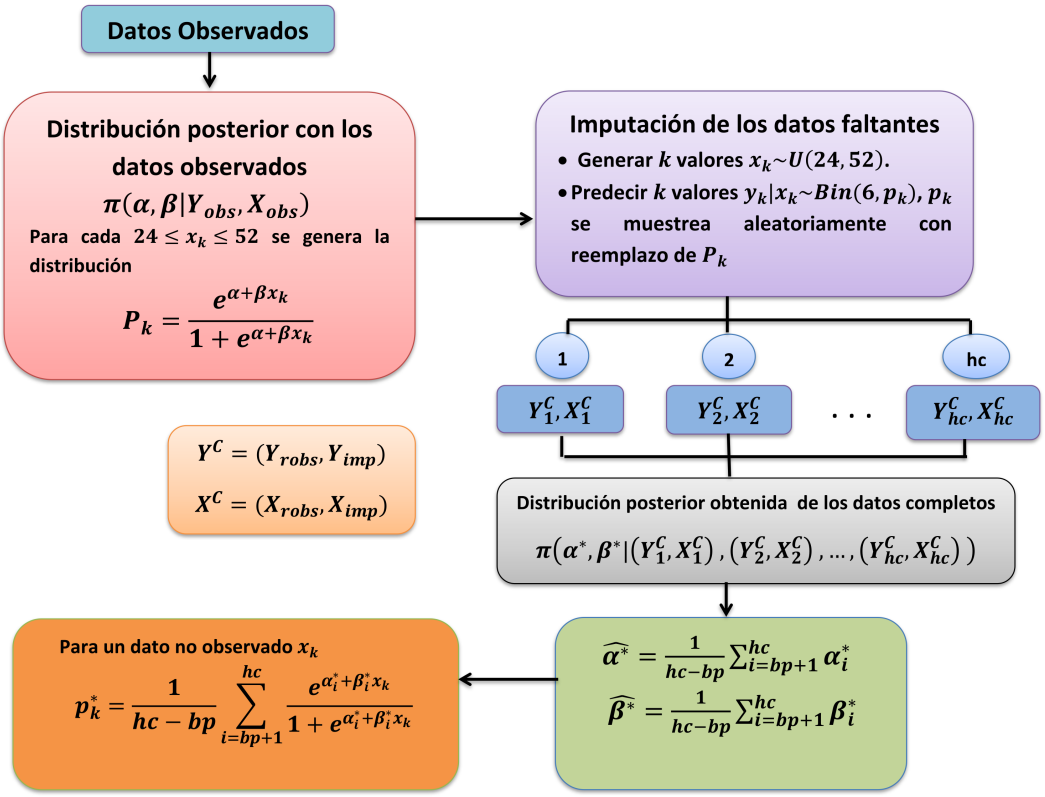


Figura 3–3: Diagrama para el método Post_{obs}.

3.4 Método usando la Distribución Posterior de los Datos Observados haciendo Remuestreo (Post$_{robs}$)

La idea es usar la distribución posterior de $\theta = (\alpha, \beta)$ que se obtiene haciendo remuestreo de los datos observados para generar los datos faltantes con el fin de generar de la distribución posterior de $\theta^* = (\alpha^*, \beta^*)$ que se obtiene de los datos completos para estimar p_h. Los datos completos se componen de los datos faltantes imputados y de los datos observados. Para el conjunto de datos observados (Y_{obs}, X_{obs}) se propone la siguiente metodología:

1. **Obtener la distribución posterior de α y β con los datos observados haciendo remuestreo:** Generar de la distribución posterior de α y β del modelo logístico. Esta distribución posterior se genera con el algoritmo Metropolis-Hasting haciendo remuestreo de los datos observados en cada iteración del algoritmo, observando la convergencia, se realizan 110,000 iteraciones del algoritmo y se eliminan las primeras 10,000. Por lo tanto quedan $T = 100,000$ iteraciones del algoritmo.
2. Para cada x_k no observados se genera la distribución posterior de probabilidad P_k, usando la distribución posterior de α y β con los datos observados que se obtuvo en el paso anterior. Para esto, se calcula

$$P_k = \frac{e^{\alpha+\beta x_k}}{1 + e^{\alpha+\beta x_k}} \tag{3.6}$$

   para $t = 1, \cdots, T$
3. **Obterner la distribución posterior de α^* y β^* con los datos completos**: Para generar de la distribución posterior con los datos completos cada iteración del algoritmo Metropolis-Hasting queda de la siguiente manera:

   Para la iteración t=0: se calcula el estimador de máxima verosimilitud de $\hat{\alpha}$ y $\hat{\beta}$, la matriz de covarianza $\hat{\Sigma}$ correspondiente al modelo logístico y sea $\alpha^0 = \hat{\alpha}$, $\beta^0 = \hat{\beta}$.

Para $t = 1, \cdots, T$, se repite:

(a) Generar k valores enteros aleatorios x_k^t, tal que, $x_k^t \sim U(24, min(X_{obs}))$. A estas nuevas observaciones se le denominará X_{new}^t.

(b) Predecir k valores $y_k^t|x_k^t$, tal que $y_k^t|x_k^t \sim Bin(6, p_k^t)$, donde p_k^t se selecciona aleatoriamente con reemplazo de la distribución de probabilidades generada en el paso 2. A estas nuevas observaciones se le denominará Y_{new}^t.

(c) Con el conjunto de datos completos (Y^c, X^c) donde $Y^c = (Y_{new}^t, Y_{obs}^t)$ y $X = (X_{new}^t, X_{obs}^t)$, el algoritmo genera α_t^* y β_t^*.

Una vez el algoritmo converge, y ya con la distribución posterior de α^* y β^* y dado x_h un dato no observado se calcula

$$\hat{\alpha} = \frac{1}{T}\sum_{t=1}^{T}\alpha_t^*, \quad \hat{\beta} = \frac{1}{T}\sum_{t=1}^{T}\beta_t^* \quad \text{y} \quad p_h^* = \frac{1}{T}\sum_{t=1}^{T}\frac{e^{\alpha_t^*+\beta_t^* x_h}}{1+e^{\alpha_t^*+\beta_t^* x_h}} \tag{3.7}$$

Un diagrama del método se muestra en la Figura 3–4.

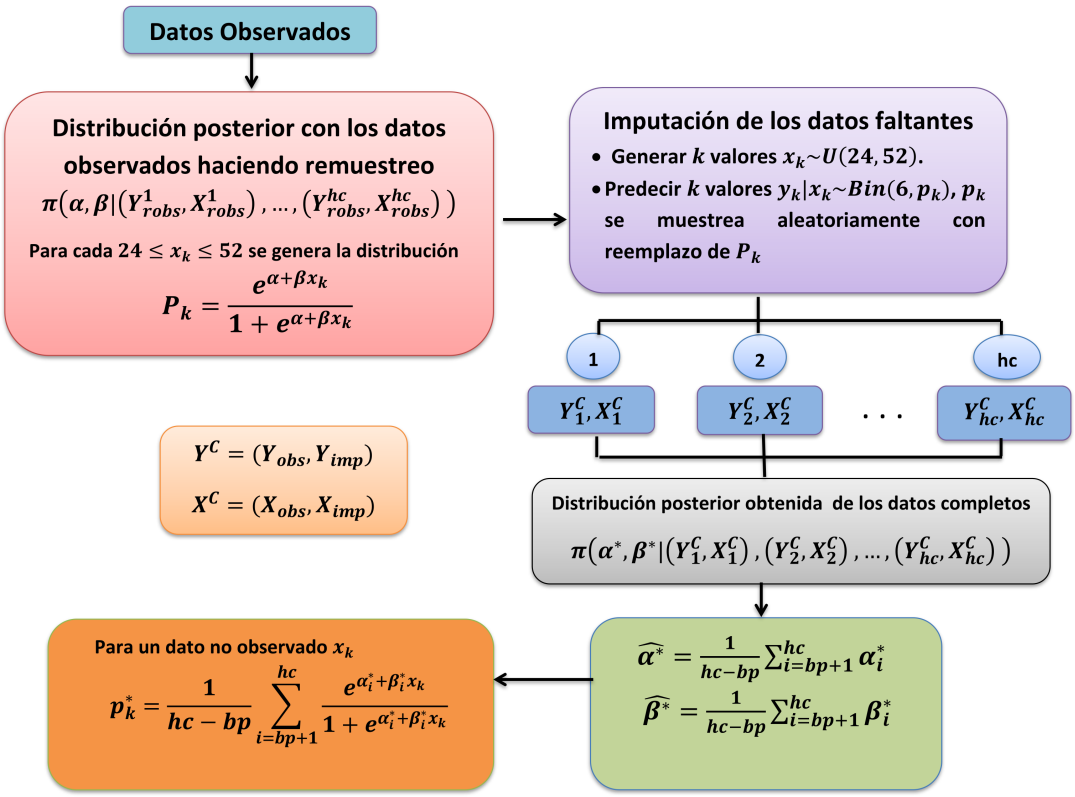


Figura 3–4: Diagrama para el método Post_{robs}.

CAPÍTULO 4
ESTUDIO DE SIMULACIÓN Y RESULTADOS

Para verificar la eficacia de los métodos propuestos en este trabajo, se realizó el siguiente estudio con datos simulados y características similares al conjunto de datos del Challenger.

4.1 Estudio de Simulación

Se simuló la variable predictora $X^c = \{x_1, \cdots, x_{50}\}$ donde $x_i \sim N(55, 11)$ para $i = 1, \cdots, 50$. Con el modelo logístico y con valores definidos para α y β (los que llamaremos reales) se generó la variable respuesta $Y^c = \{y_1, \cdots, y_{50}\}$ donde $y_i \sim Bin(6, p_i)$ son independientes y

$$p_i = \frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}}. \tag{4.1}$$

Los datos completos generados (Y^c, X^c), se ordenaron de forma ascendente y se removieron los primeros 24 datos (los datos faltantes). Los datos restantes son los que se denominan observados (Y_{obs}, X_{obs}).

Usando solo los datos observados, se implementaron los métodos descritos en el Capítulo 3, se compararon los estimadores de los parámetros del modelo obtenidos en cada método, con el α y β real utilizado para generar los datos completos.

Se presenta los resultados del estudio de simulación en dos partes.

- **Parte I:** Se presentan ejemplos particulares de tres conjuntos de datos simulados. Estos tres conjuntos de datos se denominan *similar*, *intermedio* y *diferente*, dependiendo de cuan alejado se encuentra el α y β real utilizado para

generar los datos completos, de los estimados obtenidos por el método de máxima verosimilitud con los datos observados. Para el método de Gelman y los propuestos en este trabajo se generan k observaciones para ser imputadas en lugar de las que fueron eliminadas. Estas k observaciones se generan bajo dos escenarios, en el primero se imputan $k = 25$ observaciones y en el segundo $k = 100$ observaciones. Los cuatro métodos son implementados en cada ejemplo.

- **Parte II:** Se generan treinta conjuntos de datos con diferentes α y β reales. En cada uno de estos conjuntos de datos se eliminaron 24 observaciones, para crear el conjunto de datos observados. Luego, para cada uno de ellos se implementan los cuatro métodos. Como en la Parte I, para las imputaciones se generan k observaciones bajo los escenarios $k = 25$ y $k = 100$. Se calcula el error cuadrático medio (ECM) de los estimadores para todos los conjuntos de datos y para la probabilidad en $x = 45$.

4.2 Resultados

Parte I. De los ejemplos del estudio de simulación se obtienen los siguientes resultados:

Ejemplo 1. El primer conjunto de datos simulado se denomina *similar*. Para este conjunto de datos $x_i \sim N(55, 11)$ para $i = 1, \cdots, 50$, cada x_i se redondeó al entero más cercano y con $\alpha = 9$ y $\beta = -0.2$ se generaron independientemente $y_i \sim Bin(6, p_i)$ donde

$$p_i = \frac{e^{9-0.2*x_i}}{1 + e^{9-0.2*x_i}}. \tag{4.2}$$

En la Figura 4–1 se muestran los datos generados. Los datos observados son los puntos rojos y los removidos son las estrellas azules.

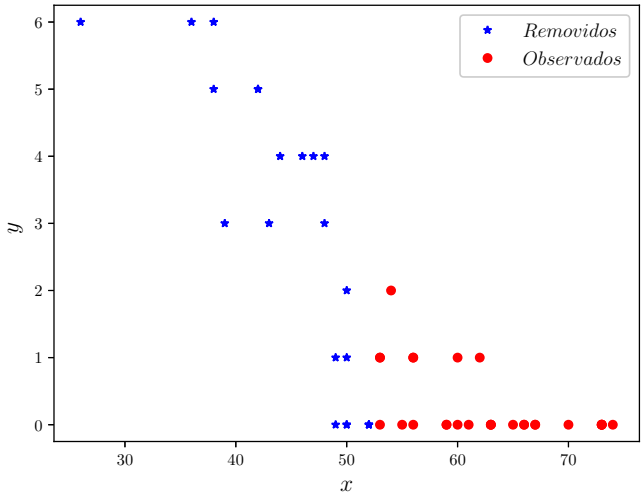
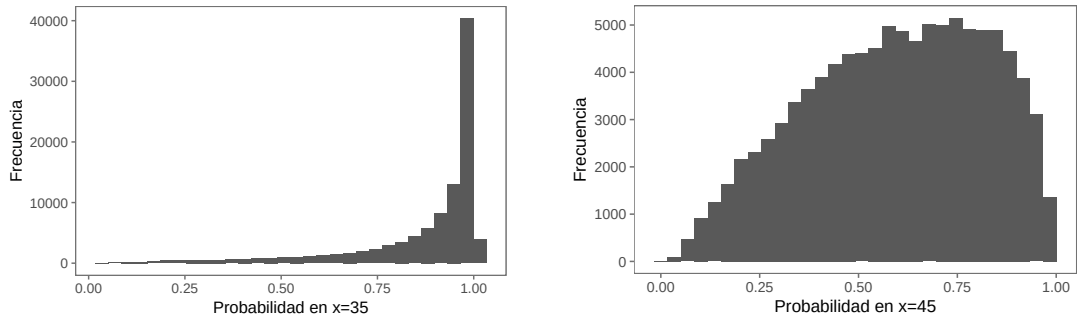


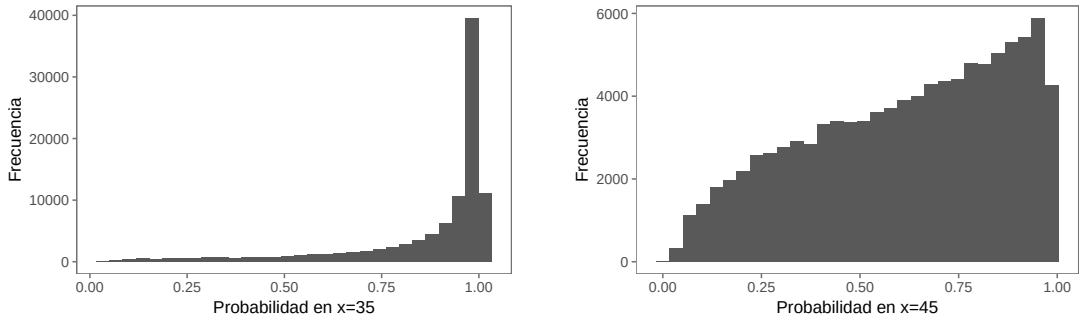
Figura 4–1: Datos observados • y datos removidos ⋆

La estimación de máxima verosimilitud usando los datos observados para $\hat{\alpha}$ es 10.913 y $\hat{\beta}$ es -0.236 siendo estos muy parecidos al α y β reales. Para los métodos propuestos $Post_{obs}$ y $Post_{robs}$, se generó de la distribución posterior de α y β con los datos observados y haciendo bootstrap de los datos observados, respectivamente. En las Figuras 4–2 y 4–3 se muestra la distribución posterior P_{35} y P_{45} descritas en la ecuación 3.4 y 3.6 obtenidas por ambos métodos.



(a) P_{35} (b) P_{45}

Figura 4–2: Distribuciones posteriores de la probabilidad P_{35} y P_{45} para el método $Post_{obs}$ del Ejemplo 1.



(a) P_{35} (b) P_{45}

Figura 4–3: Distribuciones posteriores de la probabilidad P_{35} y P_{45} para el método $Post_{robs}$ del Ejemplo 1.

Para imputar el valor $y^t|x^t = 35$, por ejemplo, en la iteración t del algoritmo, se muestrea de una distribución $Bin(6, p^t)$ donde p^t se muestrea de la distribución de la Figura 4–2(a) o 4–3(a) de acuerdo al método $Post_{obs}$ o $Post_{robs}$, respectivamente.

Las Figuras 4–4 y 4–5 muestran la distribución posterior de α y β con los datos observados y la distribución posterior de α^* y β^* cuando se imputan k=25 y k=100 datos. La recta vertical entrecortada corresponde al valor real de α y β. En general, se observa que la distribución con solo los datos observados tiende a tener mayor variabilidad que las que se obtienen luego de la imputación.

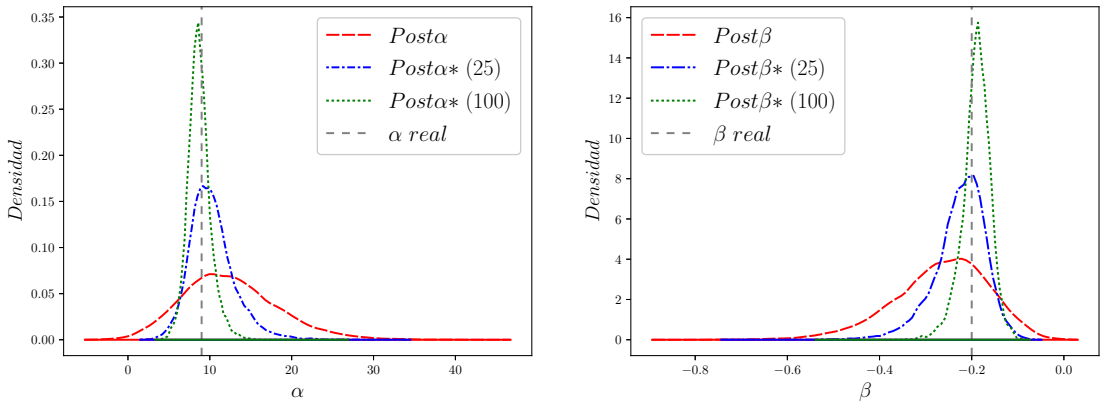


Figura 4–4: Distribuciones posteriores para el método $Post_{obs}$. $Post\alpha$ y $Post\beta$ corresponden a la posterior de los datos observados, $Post\alpha^*$ y $Post\beta^*$ son las posteriores para $k = 25$ y $k = 100$.

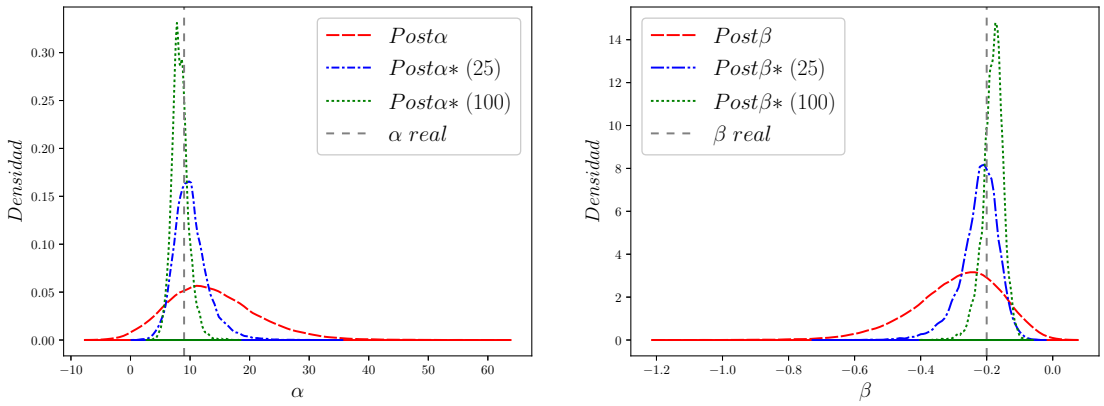


Figura 4–5: Distribuciones posteriores para el método $Post_{robs}$. $Post\alpha$ y $Post\beta$ corresponden a la posterior de los datos observados, $Post\alpha^*$ y $Post\beta^*$ son las posteriores para $k = 25$ y $k = 100$.

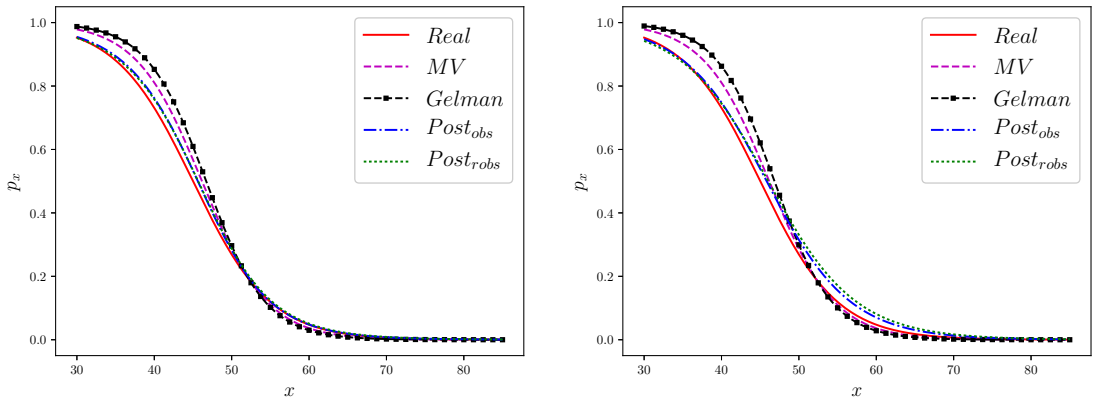
La Tabla 4–1 resume los resultados obtenidos después de implementar los cuatro métodos, se muestran los estimados para α, β y p_{45}. Para el método de Gelman,

$Post_{obs}$, $Post_{robs}$ se calculó el promedio definido en las ecuaciones 3.3, 3.5 y 3.7, respectivamente.

Tabla 4–1: Estimación de α, β y p_{45} para $k = 25$ y $k = 100$ en los cuatro métodos desarrollados.

	k=25			k=100		
Método	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}_{45}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}_{45}$
Real	9.000	-0.200	0.500	9.000	-0.200	0.500
MV	10.913	-0.236	0.567	10.913	-0.236	0.567
Gelman et al	12.225	-0.261	0.582	12.572	-0.268	0.594
Post_{obs}	10.276	-0.225	0.533	8.705	-0.189	0.538
Post_{robs}	10.132	-0.221	0.534	8.350	-0.181	0.545

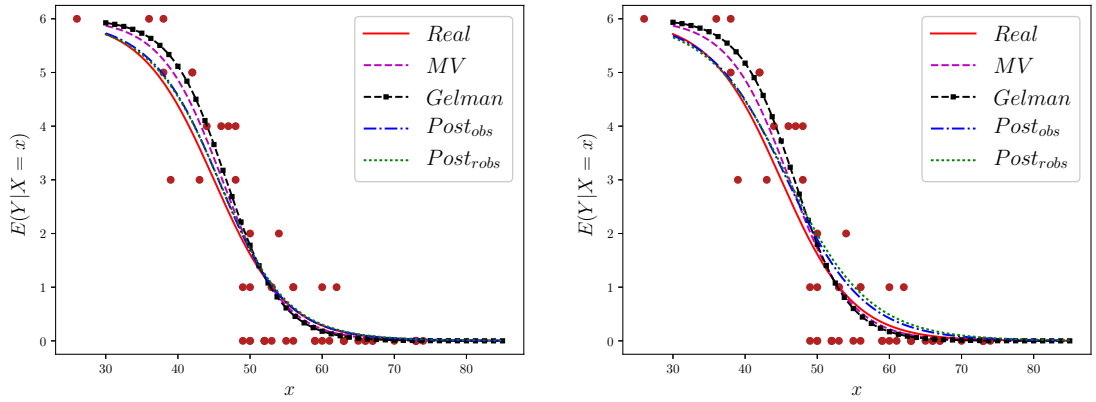
En la Tabla 4–1 observamos que la estimación de α, β y p_{45} con los métodos $Post_{obs}$ y $Post_{robs}$ se acercan más al valor real en los dos escenarios de simulación. En las Figura 4–6 se muestran las gráficas del modelo logístico usando los estimados de la Tabla 4–1 para cada escenario de simulación. En la Figura 4–6 se puede observar que las curvas correspondientes a los métodos propuestos $Post_{obs}$ y $Post_{robs}$ se acercan a la curva real proporcionando una mejor estimación de los parámetros α y β.



(a) k = 25 (b) k = 100

Figura 4–6: Probabilidad p_x, cada curva se graficó con los estimadores obtenidos para cada método y el modelo logístico.

En la Figura 4–7 se muestra la gráfica de $E(Y|X = x)$ que se obtiene multiplicando las probabilidades de la Figura 4–6 por 6.



(a) k = 25 (b) k = 100

Figura 4–7: Valor esperado $E(Y|X = x)$, los puntos rojos corresponden a los datos completos.

La Tabla 4–2 muestra el valor esperado esperado $E(Y|X = 45)$ para los cuatro métodos y los dos escenarios de simulación.

Tabla 4–2: $E(Y|X = 45)$ para k=25 y k=100 en los cuatro métodos desarrollados

	$E(Y\|X = 45)$	
Método	k=25	k=100
Real	3.000	3.000
MV	3.405	3.405
Gelman	3.494	3.565
$Post_{obs}$	3.198	3.233
$Post_{robs}$	3.207	3.275

Ejemplo 2. El segundo conjunto de datos simulado se denomina *intermedio*. Para este conjunto de datos $x_i \sim N(55, 11)$ para $i = 1, \cdots, 50$, cada x_i se redondeó al entero más cercano y con $\alpha = 13$ y $\beta = -0.27$ se generaron independientemente $y_i \sim Bin(6, p_i)$ donde

$$p_i = \frac{e^{13-0.27*x_i}}{1 + e^{13-0.27*x_i}}. \tag{4.3}$$

En la Figura 4–8 se muestran los datos generados. Los datos observados son los puntos rojos y los removidos son las estrellas azules.

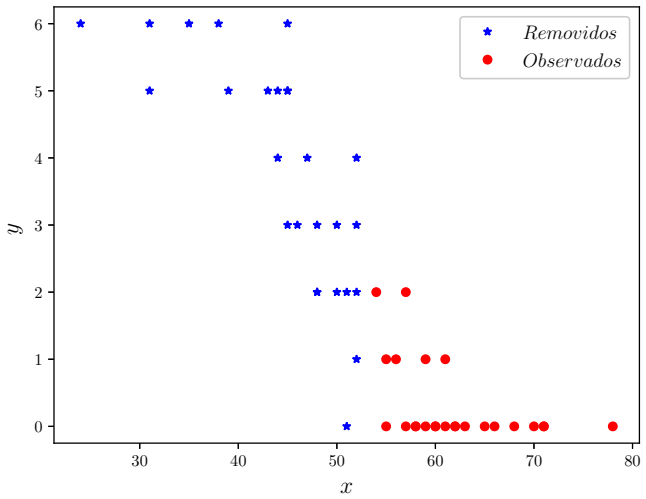
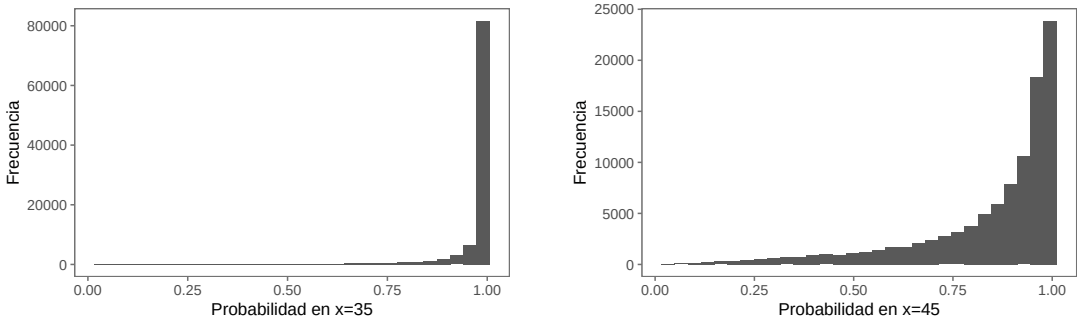


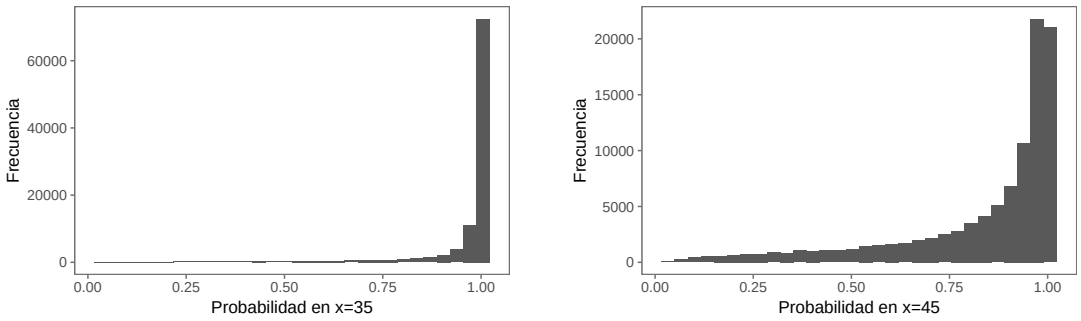
Figura 4–8: Datos observados • y datos removidos ⋆

La estimación de máxima verosimilitud usando los datos observados para $\hat{\alpha}$ es 18.756 y $\hat{\beta}$ es -0.368. Para los métodos propuestos $Post_{obs}$ y $Post_{robs}$, se generó de la distribución posterior de α y β con los datos observados y haciendo bootstrap de los datos observados, respectivamente. En las Figuras 4–9 y 4–10 se muestra la distribución posterior P_{35} y P_{45} descritas en las ecuaciónes 3.4 y 3.6 obtenidas por ambos métodos.



(a) P_{35} (b) P_{45}

Figura 4–9: Distribuciones posteriores de la probabilidad P_{35} y P_{45} para el método $Post_{obs}$ del Ejemplo 2.



(a) P_{35} (b) P_{45}

Figura 4–10: Distribuciones posteriores de la probabilidad P_{35} y P_{45} para el método $Post_{robs}$ del Ejemplo 2.

Para imputar el valor $y^t|x^t = 35$, por ejemplo, en la iteración t del algoritmo, se muestrea de una distribución $Bin(6, p^t)$ donde p^t se muestrea de la distribución en la

Figura 4–9(a) o 4–10(a) de acuerdo al método $Post_{obs}$ o $Post_{robs}$, respectivamente. Las Figuras 4–11 y 4–12 muestran la distribución posterior de α y β con los datos observados y la distribución posterior de α^* y β^* cuando se imputan k=25 y k=100 datos. La recta vertical entrecortada corresponde al valor real de α y β. En general, observamos que la distribución con solo los datos observados tiende a tener mayor variabilidad que la que se obtiene después de la imputación.

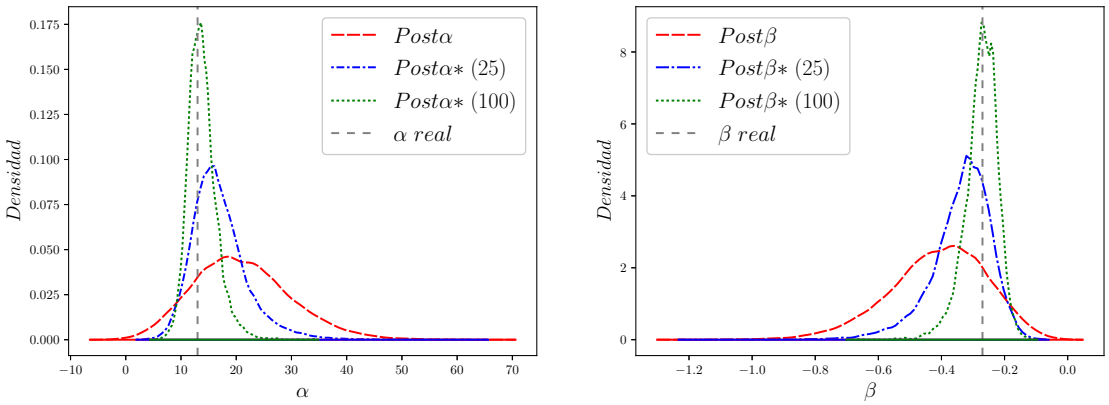


Figura 4–11: Distribuciones posteriores para el método $Post_{obs}$ $Post\alpha$ y $Post\beta$ corresponden a la posterior de los datos observados, $Post\alpha^*$ y $Post\beta^*$ son las posteriores para $k = 25$ y $k = 100$.

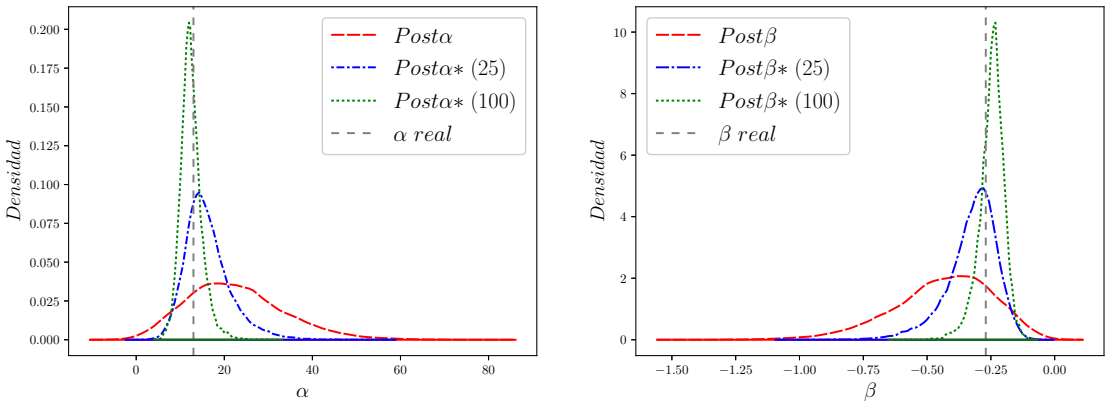


Figura 4–12: Distribuciones posteriores para el método $Post_{robs}$, $Post\alpha$ y $Post\beta$ corresponden a la posterior de los datos observados, $Post\alpha^*$ y $Post\beta^*$ son las posteriores para $k = 25$ y $k = 100$.

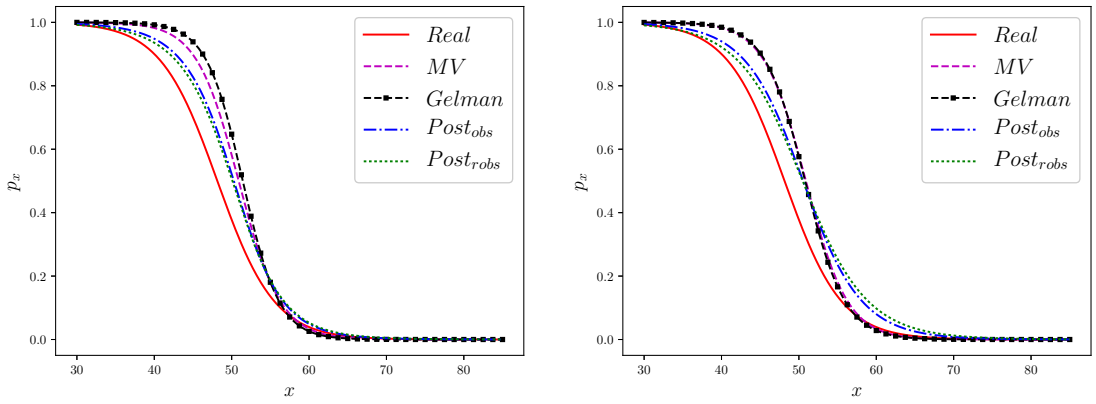
La Tabla 4–3 resume los resultados obtenidos después de implementar los cuatro métodos, se muestran los estimados para α, β y p_{45}. Para el método de Gelman, $Post_{obs}$, $Post_{robs}$ se calculó el promedio definido en las ecuaciones 3.3, 3.5 y 3.7, respectivamente.

Tabla 4–3: Estimación de α, β y p_{45} para $k = 25$ y $k = 100$ en los cuatro métodos desarrollados.

	k=25			k=100		
Método	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}_{45}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}_{45}$
Real	13.000	-0.270	0.701	13.000	-0.270	0.701
MV	18.756	-0.368	0.897	18.756	-0.368	0.897
Gelman	21.810	-0.424	0.911	19.528	-0.384	0.893
$Post_{obs}$	16.946	-0.336	0.827	13.844	-0.273	0.813
$Post_{robs}$	16.239	-0.324	0.806	12.481	-0.247	0.788

En la Tabla 4–3 observamos que la estimación de α, β con los métodos $Post_{obs}$ y $Post_{robs}$ se acercan al valor real en los dos escenarios de simulación. Para el método de Gelman las estimaciones quedan por encima del método de máxima verosimilitud.

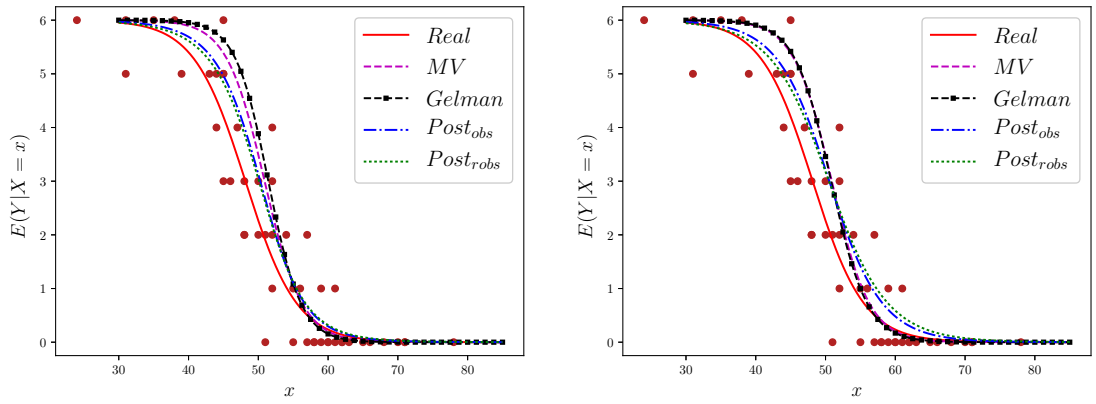
En la Figuras 4–13 se muestran las gráficas del modelo logístico usando los estimados de la Tabla 4–3 para los dos escenarios de simulación. Se puede observar que las curvas correspondientes a los métodos propuestos $Post_{obs}$ y $Post_{robs}$ se acercan a la curva real proporcionando una mejor estimación de los parámetros α y β.



(a) k = 25 (b) k = 100

Figura 4–13: Probabilidad p_x, cada curva se graficó con los estimadores obtenidos para cada método y el modelo logístico

En la Figura 4–14 se muestra la gráfica $E(Y|X = x)$ que se obtiene multiplicando las probabilidades de la Figura 4–13 por 6.



(a) k = 25 (b) k = 100

Figura 4–14: Valor esperado $E(Y|X = x)$, los puntos rojos corresponden a los datos completos.

La Tabla 4–4 muestra el valor esperado esperado $E(Y|X = 45)$ para los cuatro métodos y los dos escenarios de simulación.

Tabla 4–4: $E(Y|X = 45)$ para $k = 25$ y $k = 100$ en los cuatro métodos desarrollados.

	$E(Y\|X=45)$	
Método	k=25	k=100
Real	4.203	4.203
MV	5.387	5.387
Gelman	5.165	4.762
$Post_{obs}$	4.965	4.879
$Post_{robs}$	4.837	4.728

Ejemplo 3. El tercer conjunto de datos simulado se denomina *diferente*. Para este conjunto de datos $x_i \sim N(55, 11)$ para $i = 1, \cdots, 50$, cada x_i se redondeó al entero más cercano y con $\alpha = 12$ y $\beta = -0.25$ se generaron independientemente $y_i \sim Bin(6, p_i)$ donde

$$p_i = \frac{e^{12-0.25*x_i}}{1 + e^{12-0.25*x_i}}. \tag{4.4}$$

En la Figura 4–15 se muestran los datos generados. Los datos observados son los puntos rojos y los removidos son las estrellas azules.

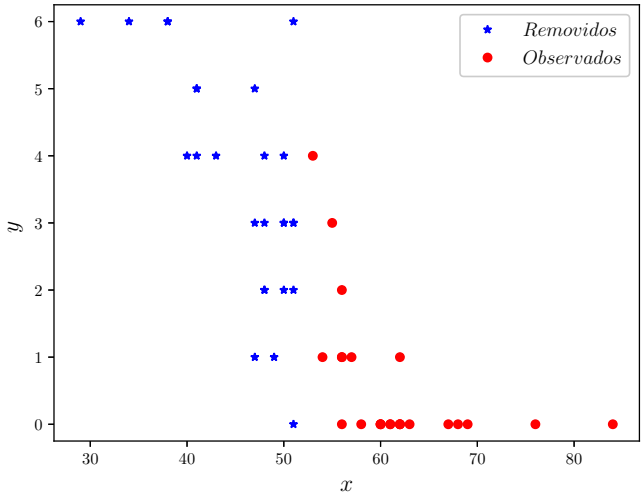
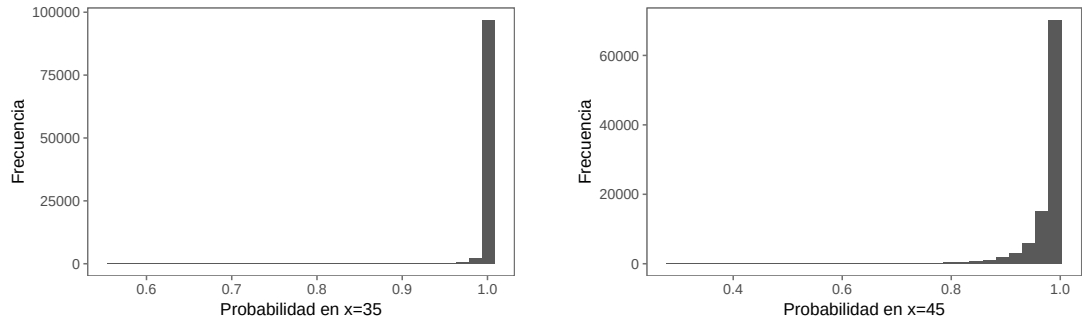


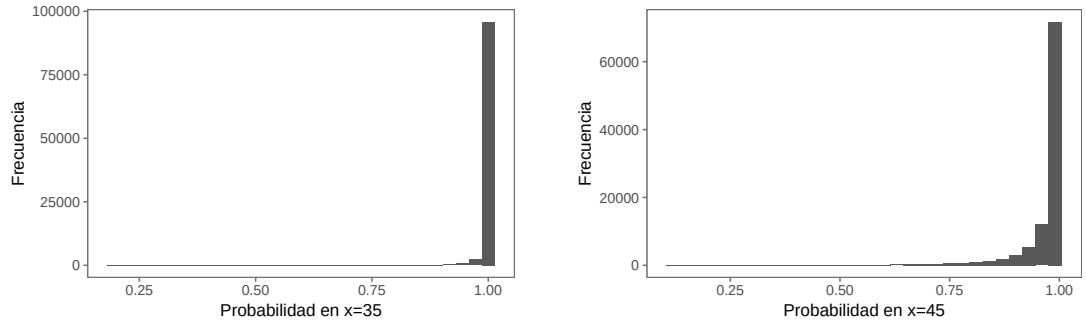
Figura 4–15: Datos observados • y datos removidos ⋆

La estimación de máxima verosimilitud usando los datos observados para $\hat{\alpha}$ es 28.388 y $\hat{\beta}$ es -0.5329, siendo esta muy diferente al α y β reales. Para los métodos propuestos $Post_{obs}$ y $Post_{robs}$, se generó de la distribución posterior de α y β con los datos observados y haciendo bootstrap de los datos observados respectivamente. En las Figuras 4–16 y 4–17 se muestran la distribución posterior P_{35} y P_{45} descritas en las ecuaciónes 3.4 y 3.6 obtenidas por ambos métodos.



(a) P_{35} (b) P_{45}

Figura 4–16: Distribuciones posteriores de la probabilidad P_{35} y P_{45} para el método $Post_{obs}$ del Ejemplo 3.



(a) P_{35} (b) P_{45}

Figura 4–17: Distribuciones posteriores de la probabilidad P_{35} y P_{45} para el método $Post_{robs}$ del Ejemplo 3.

Para imputar el valor $y^t|x^t = 35$, por ejemplo, en la iteración t del algoritmo, se muestrea de una distribución $Bin(6, p^t)$, donde p^t se muestrea de la distribución en la Figura 4–16(a) o 4–17(a) de acuerdo al método $Post_{obs}$ o $Post_{robs}$, respectivamente. Las Figuras 4–18 y 4–19 muestran la distribución posterior de α y β con los datos observados y la distribución posterior de α^* y β^* cuando se imputan k=25 y k=100 datos. La recta vertical entrecortada corresponde al valor real de α y β. En general, observamos que la distribución con solo los datos observados tiende a tener mayor variabilidad que la que se obtiene después de la imputación.

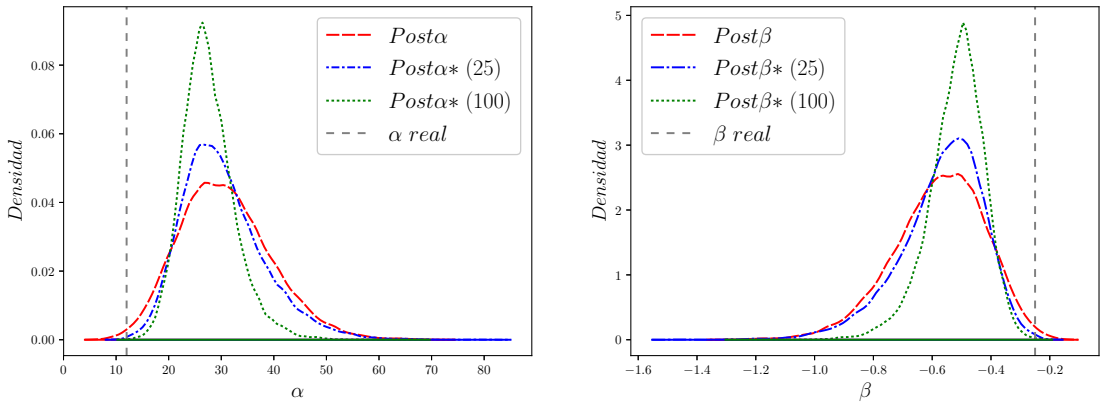


Figura 4–18: Distribuciones posteriores para el método $Post_{obs}$. $Post\alpha$ y $Post\beta$ corresponden a la posterior de los datos observados, $Post\alpha^*$ y $Post\beta^*$ son las posteriores para $k = 25$ y $k = 100$.

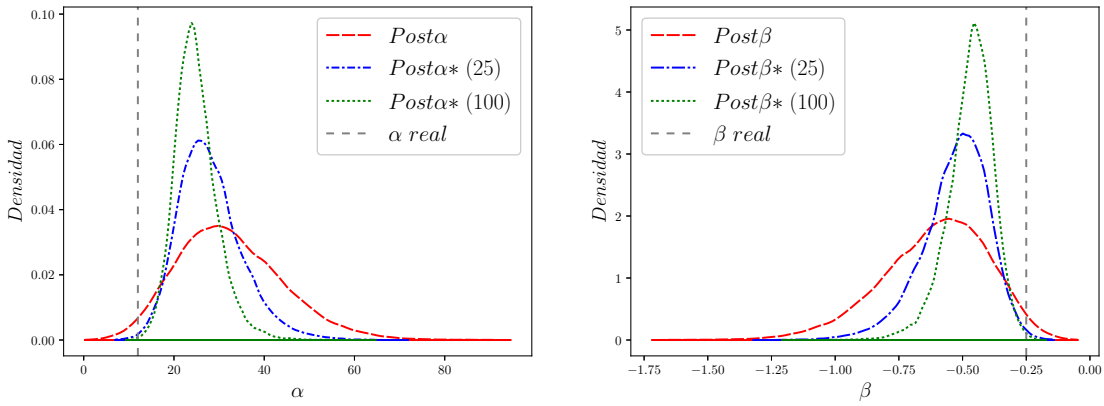


Figura 4–19: Distribuciones posteriores para el método $Post_{robs}$. $Post\alpha$ y $Post\beta$ corresponden a la posterior de los datos observados, $Post\alpha^*$ y $Post\beta^*$ son las posteriores para $k = 25$ y $k = 100$.

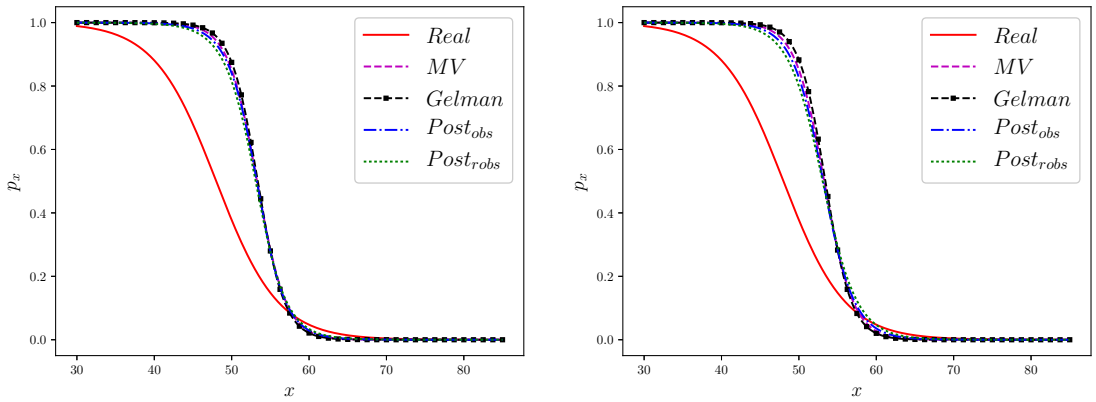
La Tabla 4–5 resume los resultados obtenidos después de implementar los cuatro métodos, se muestran los estimados para α, β y p_{45}. Para el método de Gelman,

$Post_{obs}$, $Post_{robs}$ se calculó el promedio definido en las ecuaciones 3.3, 3.5 y 3.7, respectivamente.

Tabla 4–5: Estimación de α, β y p_{45} para $k = 25$ y $k = 100$ en los cuatro métodos desarrollados.

	k=25			k=100		
Método	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}_{45}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}_{45}$
Real	12.000	-0.250	0.679	12.000	-0.250	0.679
MV	28.388	-0.532	0.987	28.388	-0.532	0.987
Gelman	30.804	-0.577	0.976	31.482	-0.589	0.981
Post_{obs}	30.273	-0.568	0.981	27.544	-0.518	0.981
Post_{robs}	28.007	-0.527	0.973	24.902	-0.469	0.971

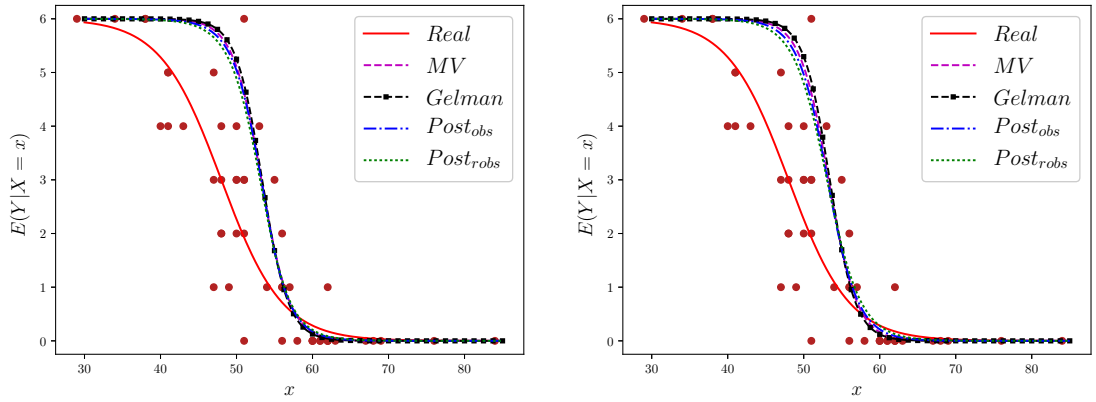
En la Tabla 4–5 observamos que la estimación de α, β con los cuatro métodos se mantiene lejos del los valores reales. El método $Post_{robs}$ para k=100 muestra una estimación de α, β menores al método de máxima verosimilitud. La probabilidad estimada para p_{45} es similar para los cuatro métodos y bajo los dos escenarios de simulación. En la Figura 4–20 se muestran las gráficas del modelo logístico usando los estimados de la Tabla 4–5 para los dos escenarios de simulación. Se puede observar que las curvas correspondientes a los métodos de Gelman, $Post_{obs}$ y $Post_{robs}$ se mantienen junto a la del método de máxima verosimilitud en los dos escenarios de simulación.



(a) k = 25 (b) k = 100

Figura 4–20: Probabilidad p_x, cada curva se graficó con los estimadores obtenidos para cada método y el modelo logístico.

En la Figura 4–21 se muestra la gráfica $E(Y|X = x)$ que se obtiene multiplicando las probabilidades de la Figura 4–20 por 6.



(a) k = 25 (b) k = 100

Figura 4–21: Valor esperado $E(Y|X = x)$, los puntos rojos corresponden a los datos completos.

La Tabla 4–6 muestra el valor esperado esperado $E(Y|X = 45)$ para los cuatro métodos y los dos escenarios de simulación.

Tabla 4–6: $E(Y|X = 45)$ para k=25 y k=100 en los cuatro métodos desarrollados.

	$E(Y\|X = 45)$	
Método	k=25	k=100
Real	4.075	4.075
MV	5.927	5.927
Gelman	5.858	5.890
$Post_{obs}$	5.886	5.886
$Post_{robs}$	5.843	5.831

Se puede concluir que cuando se tiene un conjunto de datos como el *similar*, los métodos propuestos se acercan a los valores reales. Para el ejemplo *intermedio* los métodos propuestos proporcionan una buena la estimación de los valor reales de α y β, y se encuentra muy cerca del promedio de la distribución posterior encontrada. Para el ejemplo *diferente* las estimaciones quedan cerca de la estimación obtenida con solo los datos observados y el método de máxima verosimilitud. Para este ejemplo se puede ver que las distribuciones de probabilidad presentadas en las Figuras 4–16 y 4–17 son más sesgadas a la izquierda. Esto quiere decir que con alta probabilidad los datos imputados que son generados a partir de estas distribuciones no toman valores pequeños, evitando que las curvas se acerquen a la real. Para $k > 100$ los estimados de α y β cambian muy poco y a medida que aumenta se estabilizan. Las figuras en el Apéndice 5.2 muestran las densidades de los métodos propuestos para $k = 15, 25, 100, 150$ y 200 así como las gráficas de los *running means* de la distribución posterior de α y β para $k = 25$ y $k = 100$, con las que evaluamos la convergencia de nuestros métodos. En estas gráficas se observa que los métodos propuestos convergen mientras que el método de Gelman no alcanza la convergencia.

En los tres ejemplos anteriores la curva real está por debajo de la obtenida por el método de máxima verosimilitud y los datos observados, por lo que se propone un

cuarto ejemplo donde la curva obtenida por el método de máxima verosimilitud y los datos observados esta por debajo de la real. El análisis se presenta a continuación.

Ejemplo 4. Para este conjunto de datos se simuló $x_i \sim N(55, 11)$ para $i = 1, \cdots, 50$, cada x_i se redondeo al entero más cercano y con $\alpha = 13$ y $\beta = -0.27$ se generaron independientemente $y_i \sim Bin(6, p_i)$ donde

$$p_i = \frac{e^{13-0.27*x_i}}{1 + e^{13-0.27*x_i}}. \tag{4.5}$$

En la Figura 4–22 se muestran los datos generados. Los datos observados son los puntos rojos y los removidos son las estrellas azules.

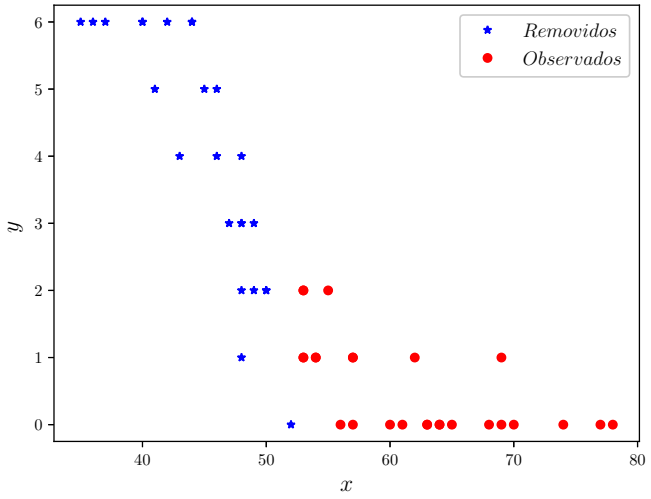
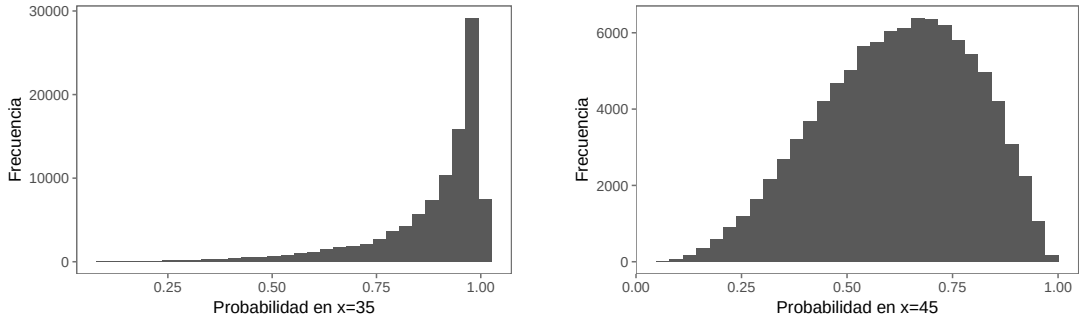


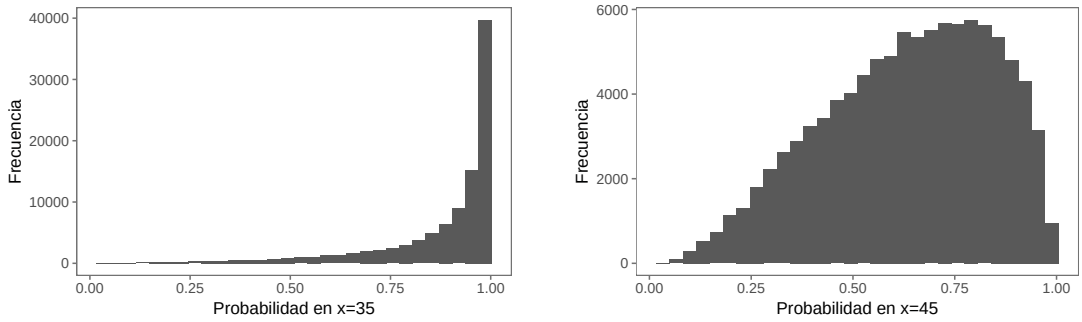
Figura 4–22: Datos observados • y datos removidos ⋆

El estimado de máxima verosimilitud usando los datos observados para $\hat{\alpha}$ es 9.50098 y $\hat{\beta}$ es -0.20232. Para los métodos propuestos $Post_{obs}$ y $Post_{robs}$, se generó de la distribución posterior de α y β con los datos observados y haciendo bootstrap de los datos observados, respectivamente. En las Figuras 4–23 y 4–24 se muestran la distribución posterior P_{35} y P_{45} descritas en las ecuaciónes 3.4 y 3.6 obtenidas por ambos métodos.



(a) P_{35} (b) P_{45}

Figura 4–23: Distribuciones posteriores de la probabilidad P_{35} y P_{45} para el método $Post_{obs}$ del Ejemplo 4.



(a) P_{35} (b) P_{45}

Figura 4–24: Distribuciones posteriores de la probabilidad P_{35} y P_{45} para el método $Post_{robs}$ del Ejemplo 4.

Para imputar el valor $y^t|x^t = 35$, por ejemplo, en la iteración t del algoritmo, se muestrea de una distribución $Bin(6, p^t)$ donde p^t se muestrea de la distribución en la Figura 4–23(a) o 4–24(a) de acuerdo al método $Post_{obs}$ o $Post_{robs}$, respectivamente. Las Figuras 4–25 y 4–26 muestran la distribución posterior de α y β con los datos observados y la distribución posterior de α^* y β^* cuando se imputan k=25 y k=100 datos. La línea vertical entrecortada corresponde al valor real de α y β. En general, observamos que la distribución con solo los datos observados tiende a tener mayor variabilidad que la que se obtiene después de la imputación.

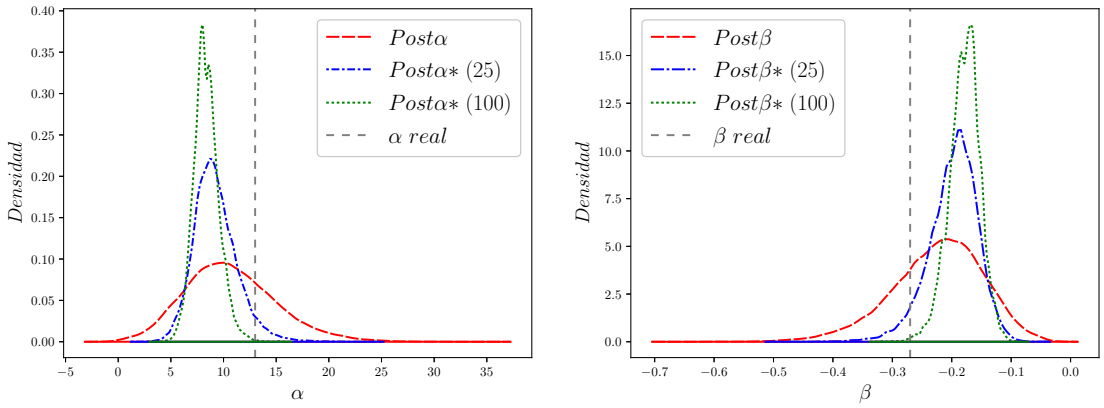


Figura 4–25: Distribuciones posteriores para el método $Post_{obs}$. $Post\alpha$ y $Post\beta$ corresponden a la posterior de los datos observados, $Post\alpha^*$ y $Post\beta^*$ son las posteriores para $k = 25$ y $k = 100$.

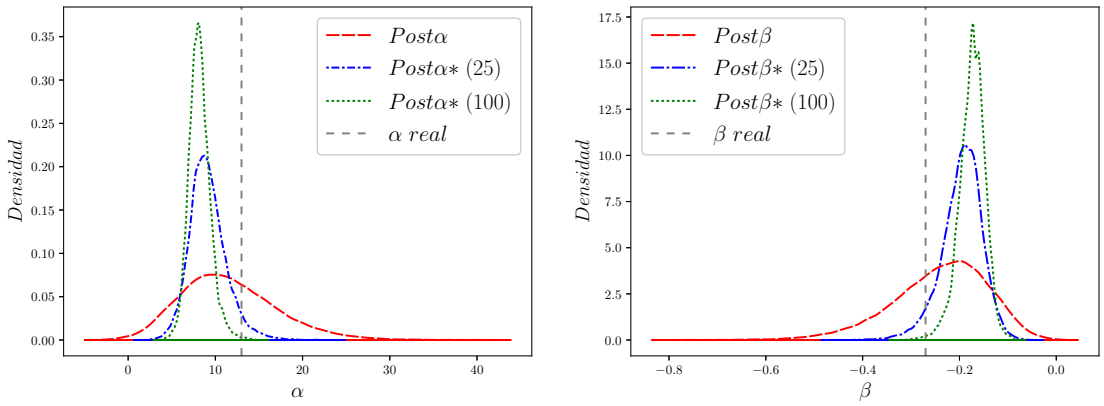


Figura 4–26: Distribuciones posteriores para el método $Post_{robs}$ $Post\alpha$ y $Post\beta$ corresponden a la posterior de los datos observados, $Post\alpha^*$ y $Post\beta^*$ son las posteriores para $k = 25$ y $k = 100$.

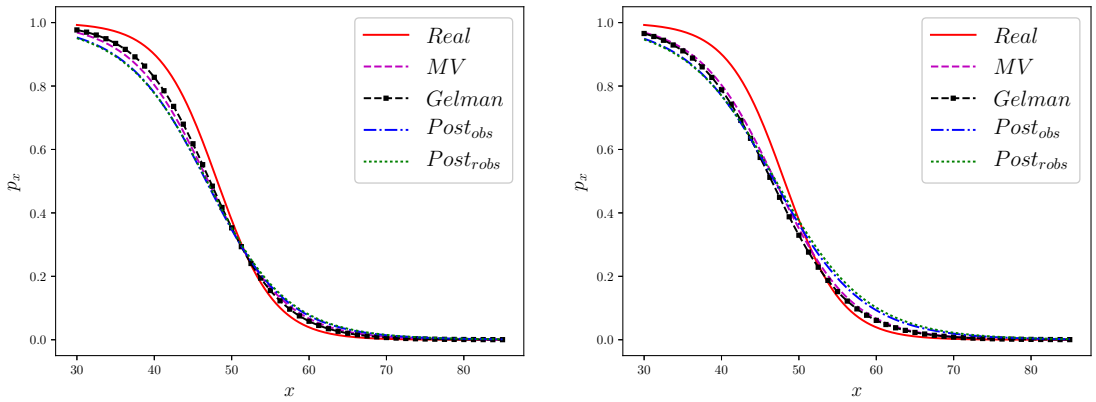
La Tabla 4–7 resume los resultados obtenidos después de implementar los cuatro métodos, se muestran los estimados para α, β y p_{45}. Para el método de Gelman,

$Post_{obs}$, $Post_{robs}$ se calculó el promedio definido en las ecuaciones 3.3, 3.5 y 3.7 respectivamente.

Tabla 4–7: Estimación de α, β y p_{45} para $k = 25$ y $k = 100$ en los cuatro métodos desarrollados.

	k=25			k=100		
Método	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}_{45}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}_{45}$
Real	13.000	-0.270	0.7005	13.000	-0.270	0.701
MV	9.501	-0.202	0.597	9.501	-0.202	0.597
Gelman	10.266	-0.217	0.598	9.404	-0.202	0.564
Post_{obs}	9.233	-0.197	0.581	8.378	-0.178	0.581
Post_{robs}	9.135	-0.195	0.583	8.191	-0.1741	0.587

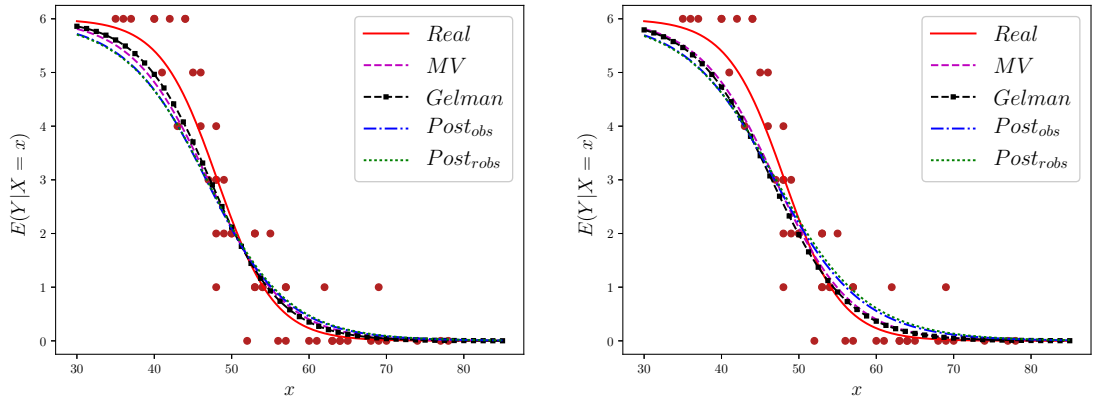
En la Tabla 4–7 observamos que la estimación de α, β con los cuatro métodos se mantiene lejos del los valores reales. Los métodos $Post_{obs}$ y $Post_{robs}$ para k=100 muestra una estimación de α, β menores al método de máxima verosimilitud. La probabilidad estimada para p_{45} es similar para los cuatro métodos y bajo los dos escenarios de simulación. En la Figura 4–27 se muestran las gráficas del modelo logístico usando los estimados de la Tabla 4–7 para los dos escenarios de simulación. Se puede observar que las curvas correspondientes a los métodos de Gelman, $Post_{obs}$ y $Post_{robs}$ se mantienen junto a la del método de máxima verosimilitud en los dos escenarios de simulación.



(a) k = 25 (b) k = 100

Figura 4–27: Probabilidad p_x, cada curva se graficó con los estimadores obtenidos para cada método y el modelo logístico.

En la Figura 4–28 se muestra la gráfica $E(Y|X = x)$ que se obtiene multiplicando las probabilidades de la Figura 4–27 por 6.



(a) k = 25 (b) k = 100

Figura 4–28: Valor esperado $E(Y|X = x)$, los puntos rojos corresponden a los datos completos.

La Tabla 4–8 muestra el valor esperado esperado $E(Y|X = 45)$ para los cuatro métodos y los dos escenarios de simulación.

Tabla 4–8: $E(Y|X = 45)$ para k=25 y k=100 en los cuatro métodos desarrollados.

	$E(Y\|X = 45)$	
Método	k=25	k=100
Real	4.2034	4.2034
MV	3.5874	3.5874
Gelman	3.5916	3.3855
$Post_{obs}$	3.4867	3.4862
$Post_{robs}$	3.5014	3.5221

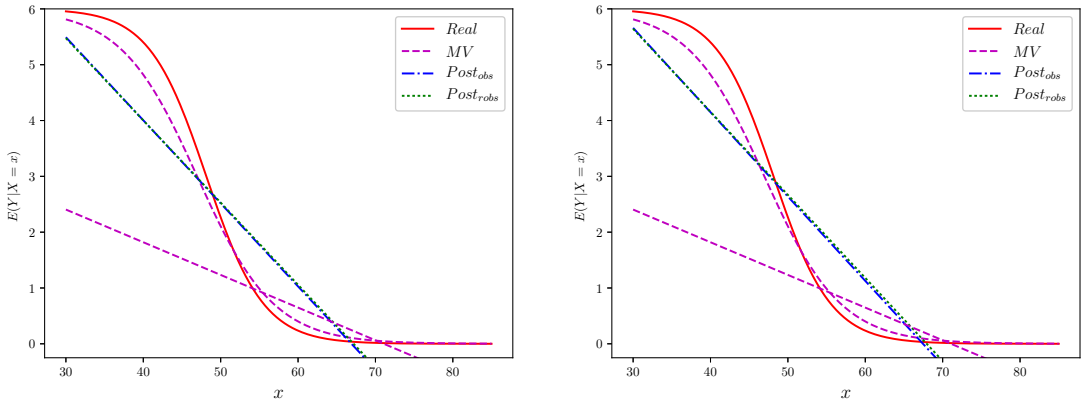
Se puede observar que los métodos propuestos no logran acercarse a los valores reales de α y β, incluso para $k = 100$ las estimaciones de α y β son menores a las del método de máxima verosimilitud. Las curvas estimadas se mantienen cerca de la que se obtiene con el método de máxima verosimilitud bajo los dos escenarios de simulación.

Para investigar sobre la manera como los métodos propuestos trabajan en este ejemplo particular, se ajustaron los datos observados a un modelo lineal y se generó de la distribución posterior para un modelo lineal. Se sabe que el modelo lineal no es apropiado para ajustar estos datos, pero hacemos este ejercicio con el propósito de ver si el hecho de que la estimación no se acerque a lo real en este ejemplo es un problema con los métodos propuestos o podría estar asociado al modelo logístico.

Para generar de la distribución posterior del modelo lineal, en cada iteración del algoritmo se completaron los datos utilizando el procedimiento que se propone en los métodos $Post_{obs}$ y $Post_{robs}$, pero ajustando un modelo lineal.

La Figura 4–29 muestra la curva real y la obtenida con el método de máxima verosimilitud. La recta entrecortada corresponde a la ajustada con el modelo lineal para los datos observados. Las otras dos rectas son las estimadas con los métodos $Post_{obs}$ y $Post_{robs}$ utilizando un modelo lineal. La simulación también se hizó bajo

los dos escenarios $k = 25$ y $k = 100$. Se puede observar que las rectas estimadas con los métodos $Post_{obs}$ y $Post_{robs}$ se alejan de la ajustada con el método de máxima verosimilitud, acercándose más a lo real.



(a) k = 25 (b) k = 100

Figura 4–29: Análisis del modelo lineal para $k = 25$ y $k = 100$ en el Ejemplo 4.

Parte II. La Tabla 4–9 contiene un resumen de los resultados para los treinta conjuntos de datos simulados. En la tabla aparecen los cuatro métodos desarrollados en este trabajo. Se muestran los dos escenarios de simulación que depende del número, k, de observaciones que se imputaron en lugar de las que fueron removidas. Para cada método se calcula el error cuadrático medio (ECM) para $\hat{\alpha}$ y $\hat{\beta}$ y $\hat{p}_{45}$. El error cuadrático medio se define como el promedio de la diferencia entre el estimador y el valor real al cuadrado. Es decir para los valores de α, β y p_{45} reales en cada

conjunto de datos se tiene que

$$\text{ECM}(\hat{\alpha}) = \frac{1}{30}\sum_{i=1}^{30}(\hat{\alpha}_i - \alpha_i)^2 \tag{4.6}$$

$$\text{ECM}(\hat{\beta}) = \frac{1}{30}\sum_{i=1}^{30}(\hat{\beta}_i - \beta_i)^2 \tag{4.7}$$

$$\text{ECM}(\hat{p_{45}}) = \frac{1}{30}\sum_{i=1}^{30}(\hat{p}_{45i} - p_{45i})^2 \tag{4.8}$$

Tabla 4–9: Error cuadrático medio de $\hat{\alpha}$, $\hat{\beta}$ y $\hat{p}_{45}$ con los métodos desarrollados.

	k= 25			k=100		
Método	ECM($\hat{\alpha}$)	ECM($\hat{\beta}$)	ECM($\hat{p}_{45}$)	ECM($\hat{\alpha}$)	ECM($\hat{\beta}$)	ECM($\hat{p}_{45}$)
MV	50.893	0.0181	0.0215	50.893	0.018	0.021
Gelman	104.975	0.036	0.067	106.188	0.036	0.017
Post_{obs}	51.683	0.018	0.016	21.667	0.008	0.017
Post_{robs}	21.667	0.007	0.016	16.889	0.006	0.016

La Tabla 4–9 muestra que el ECM de los estimadores para los métodos Post_{obs} y Post_{robs} propuestos en este trabajo cuando k=100, es menor al del método de máxima verosimilitud y al del método de Gelman. Para k=25 el ECM de los estimadores para Post_{robs} sigue siendo menor al del método de máxima verosimilitud y al método de Gelman, mientras que para Post_{obs} el ECM de los estimadores es similar al del método de máxima verosimilitud. El ECM para los estimadores en el método de Gelman para k=25 se mantiene por encima de los otros métodos, sin embargo, para k=100 el ECM($\hat{p}_{45}$) es menor que el del método de máxima verosimilitud.

También se calculó la varianza de las diferencias entre el estimador y el valor real. Es decir, para $i = 1, \cdots, 30$, definimos esas diferencias como

$$d_{\alpha_i} = \hat{\alpha}_i - \alpha_i, \quad d_{\beta_i} = \hat{\beta}_i - \beta_i \ \text{ y } \ d_{p_{45i}} = \hat{p}_{45i} - p_{45i} \tag{4.9}$$

y los respectivos promedios

$$\bar{d}_\alpha = \frac{1}{30}\sum_{i=1}^{30}(\hat{\alpha}_i - \alpha_i), \quad \bar{d}_\beta = \frac{1}{30}\sum_{i=1}^{30}(\hat{\beta}_i - \beta_i) \text{ y } \quad \bar{d}_{p_{45}} = \frac{1}{30}\sum_{i=1}^{30}(\hat{p}_{45i} - p_{45i}). \quad (4.10)$$

Las varianzas de las diferencias se definen de la siguiente manera

$$\text{var}(d_\alpha) = \frac{1}{29}\sum_{i=1}^{30}(d_{\alpha_i} - \bar{d}_\alpha)^2 \quad (4.11)$$

$$\text{var}(d_\beta) = \frac{1}{29}\sum_{i=1}^{30}(d_{\beta_i} - \bar{d}_\beta)^2 \quad (4.12)$$

$$\text{var}(d_{p_{45}}) = \frac{1}{29}\sum_{i=1}^{30}(d_{p_{45i}} - \bar{d}_{p_{45}})^2 \quad (4.13)$$

Los resultados se resumen en la Tabla 4–10. Note que para los métodos Post_{obs} y Post_{robs} se obtiene una menor varianza que con los métodos de Gelman y máxima verosimilitud cuando $k = 100$. Cuando $k = 25$ la varianza de las diferencia para $\hat{\alpha}$ en los métodos Post_{obs} y Post_{robs} es muy parecida a la obtenida con el método de máxima verosimilitud. Para el método de Gelman las varianzas obtenidas son mayores que las obtenidas en los otros métodos.

Tabla 4–10: Varianza de las diferencias d_α, d_β y $d_{p_{45}}$ con los métodos desarrollados.

	k= 25			k=100		
Método	$\text{var}(d_\alpha)$	$\text{var}(d_\beta)$	$\text{var}(d_{p_{45}})$	$\text{var}(d_\alpha)$	$\text{var}(d_\beta)$	$\text{var}(d_{p_{45}})$
MV	32.249	0.109	0.016	32.249	0.109	0.016
Gelman	64.738	0.021	0.069	66.897	0.022	0.016
Post_{obs}	35.985	0.012	0.015	18.174	0.006	0.015
Post_{robs}	32.667	0.011	0.015	15.998	0.005	0.015

4.3 Aplicación

El 28 de enero de 1986, la NASA lanzó al transbordador Challenger a cumplir una misión en el espacio. A 73 segundos del despegue el Challenger explotó dejando siete víctimas fatales. Entre los tripulantes se encontraba la maestra de escuela

Sharon Christa McAuliffe, quien habia ganado el concurso nacional “Un maestro en el espacio”, convirtiéndose en la primera civil en una misión espacial.

El accidente del Challenger se catalogó como uno de los peores desastres en la historia astronáutica. Es posible que el accidente se debiera a que las bajas temperaturas en la noche anterior y el día del lanzamiento ocasionaron daños en los aros que sellaban las diferentes etapas de los cohetes aceleradores sólidos del transbordador.

La noche previa al accidente los ingenieros que fabricaron el motor sólido del cohete, debatieron junto a los expertos de la NASA, el efecto que podría tener las bajas temperatura con relación al fallo de los aros. La discusión se basó en un conjunto de datos obtenidos de 23 lanzamientos previos al Challenger. Sin embargo, la conclusión de esa discusión fue que los datos que se tenían no eran concluyentes para predecir un posible fallo en los aros, y tomaron la decisión de no detener el lanzamiento.

Los 23 lanzamientos previos al Challenger se hicieron en temperaturas entre $53°F$ y $81°F$, pero el Challenger fue lanzado a una temperatura de $31°F$, esto es, $22°F$ menos de la mínima temperatura reportada en los lanzamientos previos. La Figura 4–30 muestra el número de aros dañados versus la temperatura para los 23 vuelos lanzados.

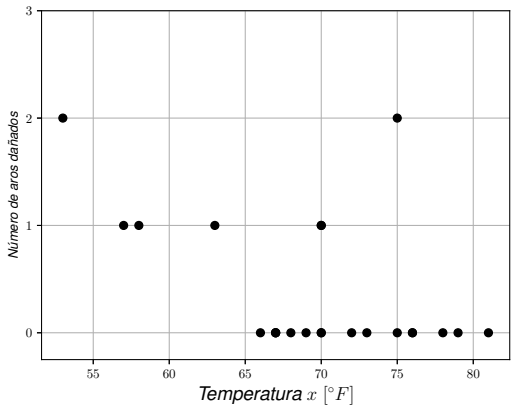


Figura 4–30: Temperatura vs número de aros dañados en los 23 lanzamientos previos al Challenger.

Los datos relacionados a los 23 lanzamientos, han sido utilizados para determinar la probabilidad de que los aros sufrieran daños a $31°F$. Dalal et al. (1989) usan el modelo de regresión logística para determinar la probabilidad de que haya daño en al menos uno de los 6 aros dada la temperatura en el momento del lanzamiento. Con el modelo logístico calcularon esta probabilidad en $31°F$ y construyeron intervalos de confianza del 90% para los parámetros del modelo utilizando un procedimiento de bootstrap paramétrico. El intervalo de confianza estimado para el número esperado de aros dañados en $30°F$ fue de (1,6) aros. También desarrollaron un modelo para analizar el riesgo probabilístico que implicaba el lanzamiento a $31°F$ incluyendo otras covariables, como la presión, entre otras.

Por otro lado, Lavine (1991) argumenta que el problema se debe tratar como un problema de extrapolación por lo que propone un análisis de extrapolación a $31°F$ basandose en las consideraciones de ingenieros expertos que consiste en que hay una relación monótona entre la temperatura y la probabilidad de fallo. Lavine

calculó un intervalo de confianza con métodos no paramétricos y concluyó que la probabilidad de fallo en $31°F$, $p_{31} \in [1/3, 1]$.

Maranzano y Krzysztofowicz (2008) desarrollaron un modelo de pronóstico usando la fórmula de Bayes. Calcularon la probabilidad a priori y unas distribuciones condicionales para calcular la probabilidad posterior de fallo en los aros en la temperatura $31°F$. Luego, calcularon nuevamente las distribuciones condicionales, la probabilidad a priori y posterior haciendo un análisis de extrapolación que consiste en añadir información sobre el posible número de aros dañados en la temperatura $31°F$. Concluyeron, por ejemplo que si en la temperatura $31°F$ hubiera un vuelo con 3 aros dañados la probabilidad de fallo $31°F$, $p_{31} \in (0.64, 0.78)$.

4.4 Resultados de la Aplicación

En este trabajo se implementaron los cuatro métodos descritos en el Capítulo 3 con los datos del Challenger, y se hizo un pronóstico de la probabilidad de fallo de cada aro a una temperatura de $31°F$. La implementación se desarrolló bajo dos escenarios $k = 25$ y $k = 100$, que indican el número de temperaturas que se imputaron entre $24°F$ y $53°F$.

Las Figuras 4–31 y 4–32 muestran la distribución posterior de los datos observados y las distribuciónes posteriores de α^* y β^* para los métodos $Post_{obs}$ y $Post_{robs}$, respectivamente, en los dos escenarios de simulación.

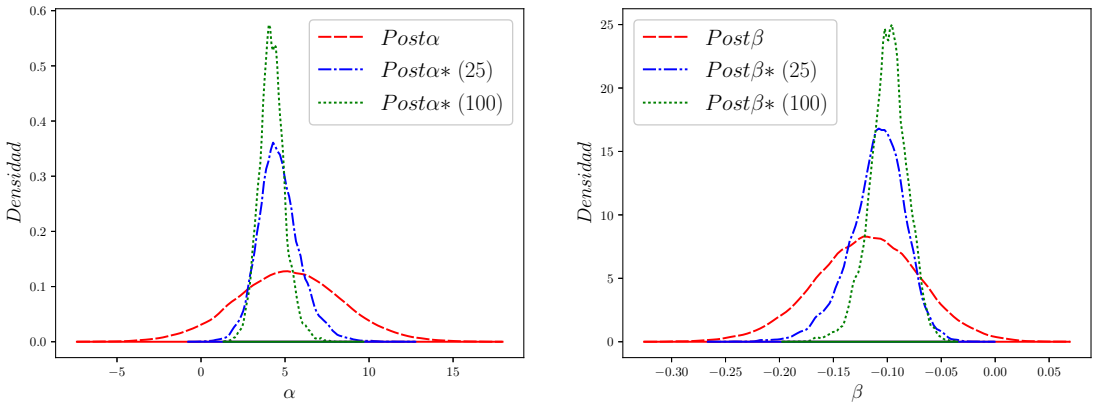


Figura 4–31: Distribuciones posteriores para el método $Post_{obs}$, $Post\alpha$ y $Post\beta$ corresponden a la posterior de los datos observados, $Post\alpha^*$ y $Post\beta^*$ son las posteriores para $k = 25$ y $k = 100$.

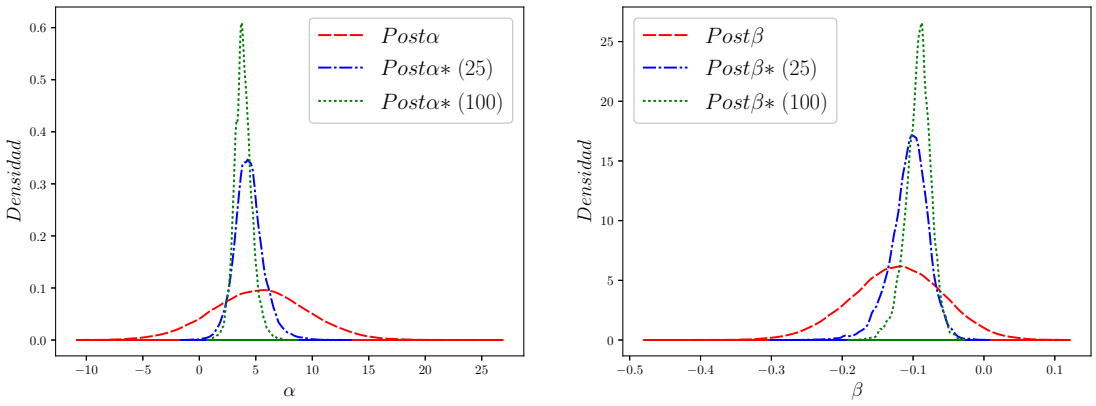
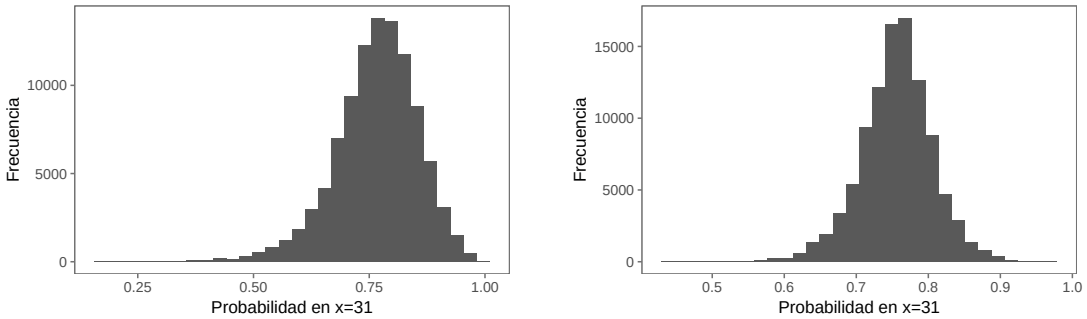


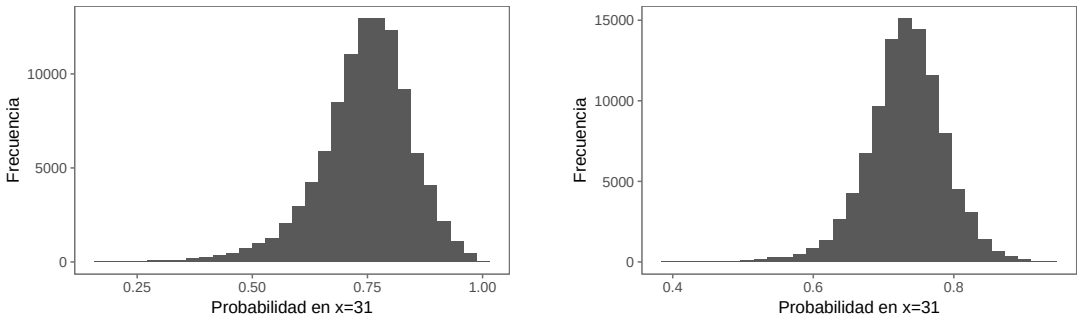
Figura 4–32: Distribuciones posteriores para el método $Post_{robs}$, $Post\alpha$ y $Post\beta$ corresponden a la posterior de los datos observados, $Post\alpha^*$ y $Post\beta^*$ son las posteriores para $k = 25$ y $k = 100$.

Las Figuras 4–33 y 4–34, muestran las distribuciones de probabilidades en $x = 31$ para los métodos $Post_{obs}$ y $Post_{robs}$.



(a) $k = 25$ (b) $k = 100$

Figura 4–33: Distribuciones de probabilidades en $x = 31$ para el método $Post_{obs}$, con $k = 25$ y $k = 100$.



(a) $k = 25$ (b) $k = 100$

Figura 4–34: Distribuciones de probabilidades en $x = 31$ para el método $Post_{robs}$, con $k = 25$ y $k = 100$.

La Tabla 4–11 muestra la estimación para α, β y la probabilidad p_{31} usando el método MV. Para p_{31} se muestra un intervalo de confianza del 95%.

Tabla 4–11: Estimación para el método MV.

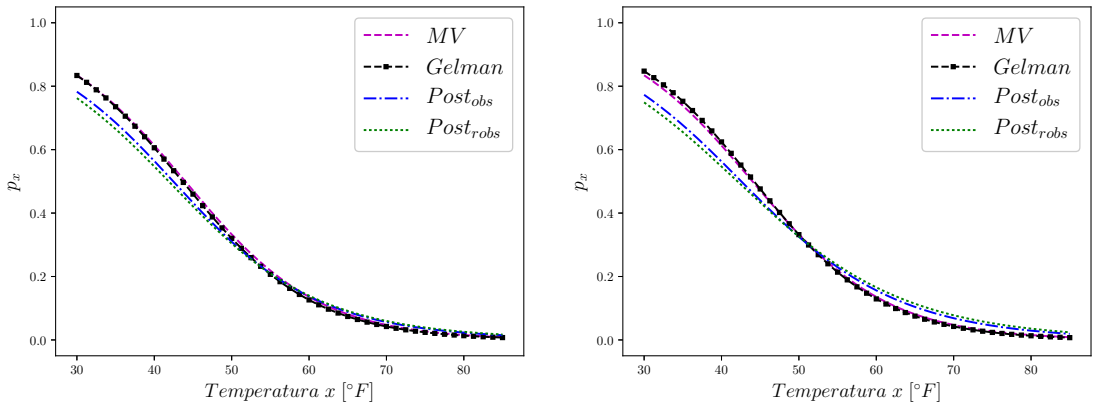
Método	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}_{31}$
MV	5.084	-0.115	0.817 (0.159,0.991)

La Tabla 4–12 muestra los estimados para α, β y la probabilidad p_{31} usando los otros tres métodos. Para p_{31} se muestran intervalos de predicción del 95%.

Tabla 4–12: Estimación de α, β, $\hat{p}_{31}$ y para $k = 25$ y $k = 100$ para los datos del Challenger.

	k=25			k=100		
Método	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}_{31}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{p}_{31}$
Gelman	5.160	-0.118	0.8178 (0.177,0.991)	5.337	-0.120	0.8343 (0.158,0.998)
$Post_{obs}$	4.647	-0.109	0.765 (0.561,0.919)	4.228	-0.099	0.755 (0.648,0.851)
$Post_{robs}$	4.405	-0.105	0.744 (0.648,0.851)	3.891	-0.092	0.731 (0.612,0.835)

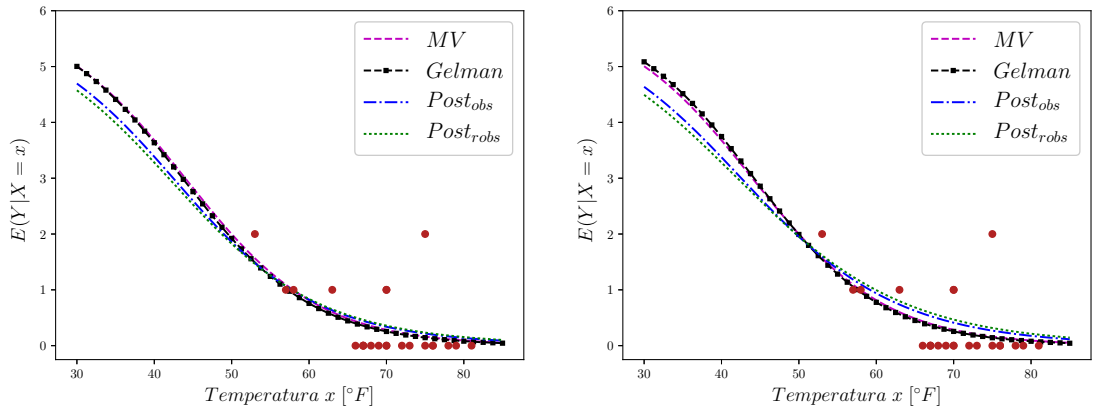
La probabilidad $P(Y \geq 1|X = 31)$ de que al menos un aro falle a $31°F$ es 0.999 para los cuatro métodos. En la Figura 4–35 se muestran las gráficas del modelo logístico usando los estimados de la Tabla 4–12 para los dos escenarios de simulación.



(a) k = 25 (b) k = 100

Figura 4–35: Probabilidad p_x vs temperatura.

En la Figura 4–36 se muestra la gráfica $E(Y|X = x)$ que se obtiene multiplicando las probabilidades de la Figura 4–35 por 6.



(a) k = 25 (b) k = 100

Figura 4–36: Número esperado de aros con fallos vs temperatura.

La Tabla 4–13 muestra el número esperado de aros que sufrirían un fallo a $31°F$, para los cuatro métodos en los dos escenarios de simulación.

Tabla 4–13: $E(Y|X = 31°F)$ para k=25 y k=100 en los cuatro métodos desarrollados.

	$E(Y\|X = 31°F)$	
Método	k=25	k=100
MV	4.920	4.920
Gelman	4.906	5.005
$Post_{obs}$	4.680	4.560
$Post_{robs}$	4.464	4.386

Observamos que según los métodos MV y Gelman, a $31°F$ se espera que alrededor de cinco de los seis aros sufrirán fallos. Para los métodos $Post_{obs}$ y $Post_{robs}$ con $k = 100$ se espera que 4.560 y 4.386 aros sufrieran fallos, respectivamente. Dalal et al. (1989) afirmaron que el número esperado de aros que sufrirían un fallo en $31°F$ está en un intervalo de (1,6) aros.

CAPÍTULO 5
CONCLUSIONES Y TRABAJOS FUTUROS

5.1 Conclusiones

En este trabajo se analizó una metodología de pronóstico con enfoque Bayesiano para el modelo logístico con datos binomiales, tomando las características del conjunto de datos del Challenger. Concluimos que:

- Los dos métodos propuestos parecen ser prometedores como posibles métodos para el análisis con datos faltantes.
- En el estudio de simulación se observa que en general, se obtiene un menor error cuadrático medio cuando se compara con el Método de Máxima Verosimilitud.
- Para los datos del Challenger se obtuvo una estimación de p_{31} de 0.755 con un intervalo de predicción del 95% de (0.648,0.851) y el valor esperado para el número de fallos en los aros de 4.560 cuando se imputan 100 temperaturas con el método $Post_{obs}$. En el método $Post_{robs}$ la estimación de p_{31} fue de 0.731 con un intervalo de predicción del 95% de (0.612,0.835) y el valor esperado para el número de fallos en los aros de 4.386.

5.2 Trabajos Futuros

Para los dos métodos propuestos en este trabajo, se proponen los siguientes trabajos futuros:

- Desarrollar un marco teórico para explicar su eficiencia.
- Estudiar su eficiencia en modelos no logísticos.

- Estudiar el efecto del número de imputaciones en la estimación de los parámetros. En este trabajo solo se imputaron $k = 25$ y $k = 100$ observaciones.
- Usando simulaciones, analizar el efecto que tiene la cantidad de datos que se remueven para luego hacer la imputación.
- Explorar más en detalle la construcción de intervalos de credibilidad para la estimación de los parámetros.
- Para el Challenger: incorporar la distribución del clima en la imputación de las temperaturas.

REFERENCIAS BIBLIOGRÁFICAS

Agresti, A. (1996). *Categorical data analysis*, volumen 990. New York: John Wiley & Sons.

Dalal, S. R., Fowlkes, E. B., y Hoadley, B. (1989). Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure. *Journal of the American Statistical Association*, 84(408):945–957.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The annals of Statistics*, pp. 1–26.

Gelman, A., Carlin, J. B., Stern, H. S., y Rubin, D. B. (2014). *Bayesian data analysis*, volumen 2. Chapman & Hall/CRC Boca Raton, FL, USA.

Gelman, A., Roberts, G. O., y Gilks, W. R. (1996). Efficient metropolis jumping rules. *Bayesian statistics*, 5(599-608):42.

Jackman, S. (2009). *Bayesian analysis for the social sciences*, volumen 846. John Wiley & Sons.

Lavine, M. (1991). Problems in extrapolation illustrated with space shuttle o-ring data. *Journal of the American Statistical Association*, 86(416):919–921.

Little, R. J. A. y Rubin, D. B. (1987). *Statistical Analysis with Missing Data, 2nd Edition*. New York: John Wiley & Sons.

Maranzano, C. J. y Krzysztofowicz, R. (2008). Bayesian reanalysis of the challenger o-ring data. *Risk Analysis*, 28(4):1053–1067.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

Robert, C. y Casella, G. (2009). *Introducing Monte Carlo Methods with R*. Springer Science & Business Media.

Roy, V. y Kaiser, M. S. (2013). Posterior propriety for bayesian binomial regression models with a parametric family of link functions. *Statistical Methodology*, 13:25–41.

APÉNDICES

APÉNDICE A
FIGURAS

A.1 Distribuciones posteriores para los método $Post_{obs}$ y $Post_{robs}$.

Ejemplo 1.

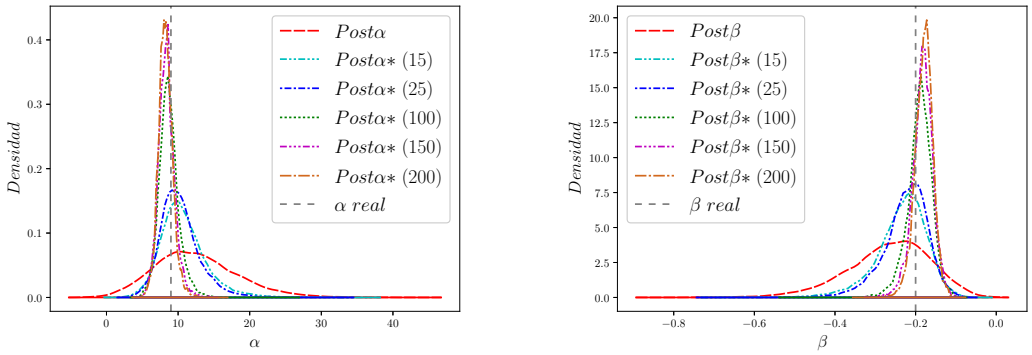


Figura A–1: Distribuciones posteriores para el método $Post_{obs}$ $Post\alpha$ y $Post\beta$ corresponden a la posterior de los datos observados, $Post\alpha^*$ y $Post\beta^*$ son las posteriores para $k = 15, 25, 100, 150, 200$.

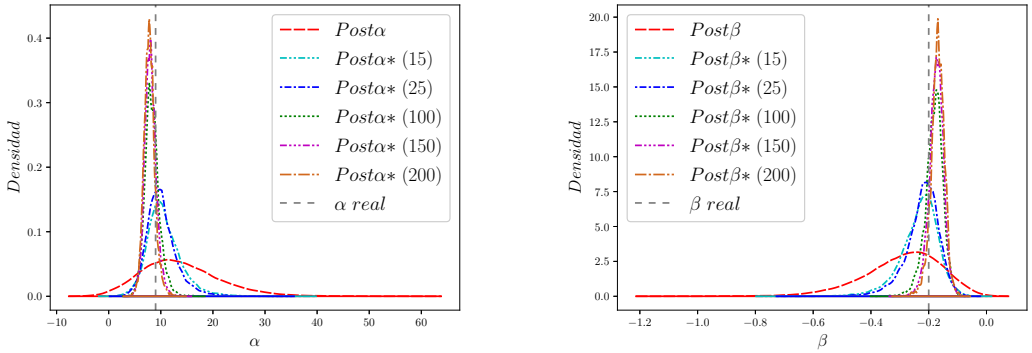


Figura A–2: Distribuciones posteriores para el método $Post_{obs}$ $Post\alpha$ y $Post\beta$ corresponden a la posterior de los datos observados, $Post\alpha^*$ y $Post\beta^*$ son las posteriores para $k = 15, 25, 100, 150, 200$.

Ejemplo 2.

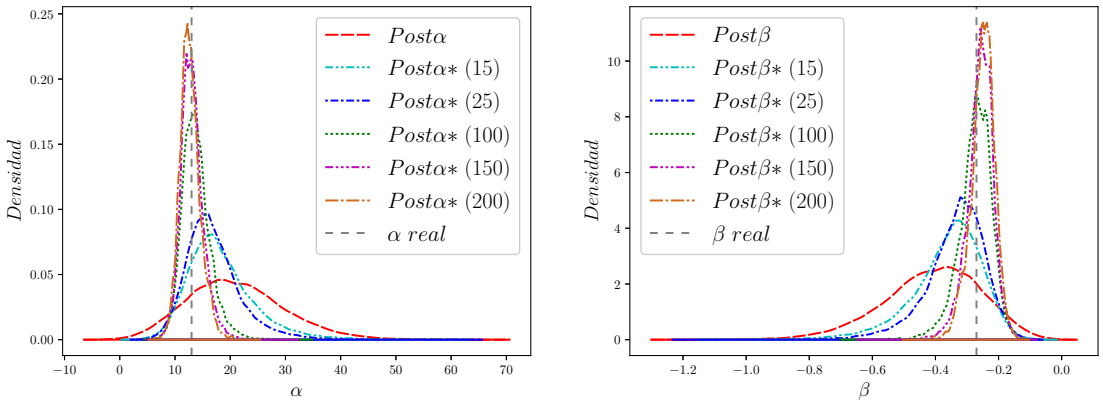


Figura A–3: Distribuciones posteriores para el método $Post_{obs}$ $Post\alpha$ y $Post\beta$ corresponden a la posterior de los datos observados, $Post\alpha^*$ y $Post\beta^*$ son las posteriores para $k = 15, 25, 100, 150, 200$.

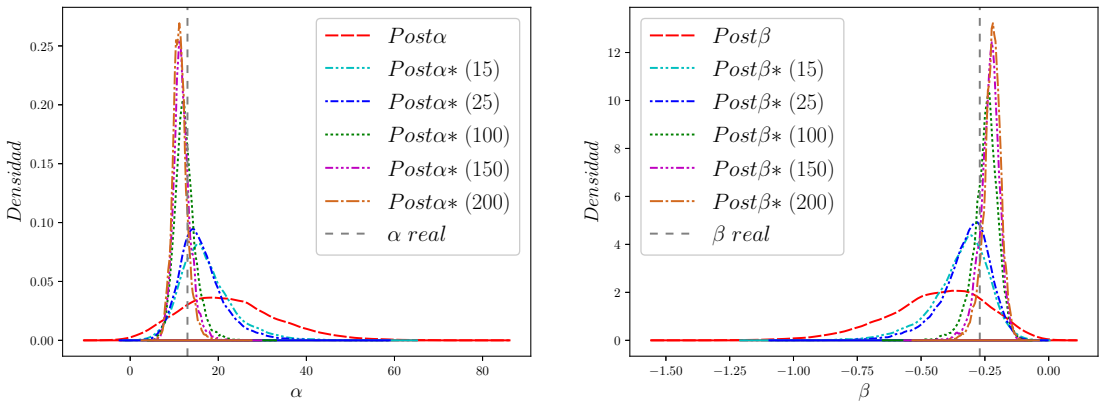


Figura A–4: Distribuciones posteriores para el método $Post_{obs}$ $Post\alpha$ y $Post\beta$ corresponden a la posterior de los datos observados, $Post\alpha^*$ y $Post\beta^*$ son las posteriores para $k = 15, 25, 100, 150, 200$.

Ejemplo 3.

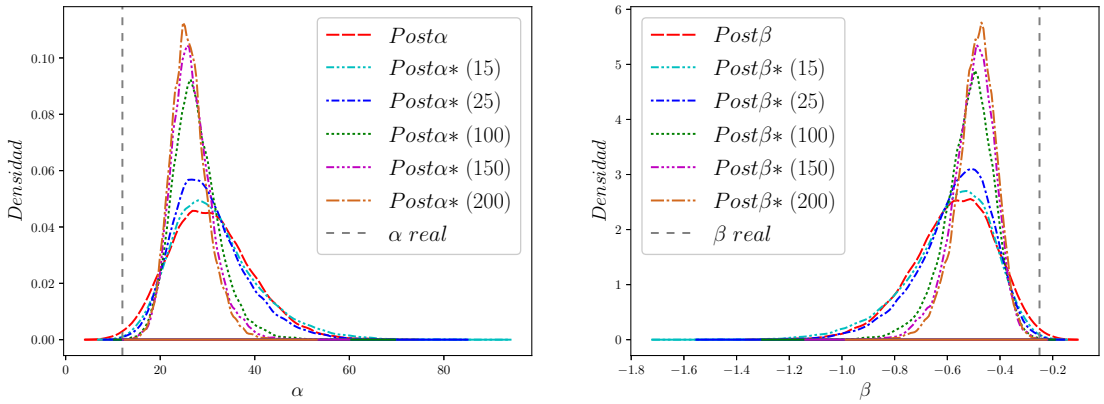


Figura A–5: Distribuciones posteriores para el método $Post_{obs}$ $Post\alpha$ y $Post\beta$ corresponden a la posterior de los datos observados, $Post\alpha^*$ y $Post\beta^*$ son las posteriores para $k = 15, 25, 100, 150, 200$.

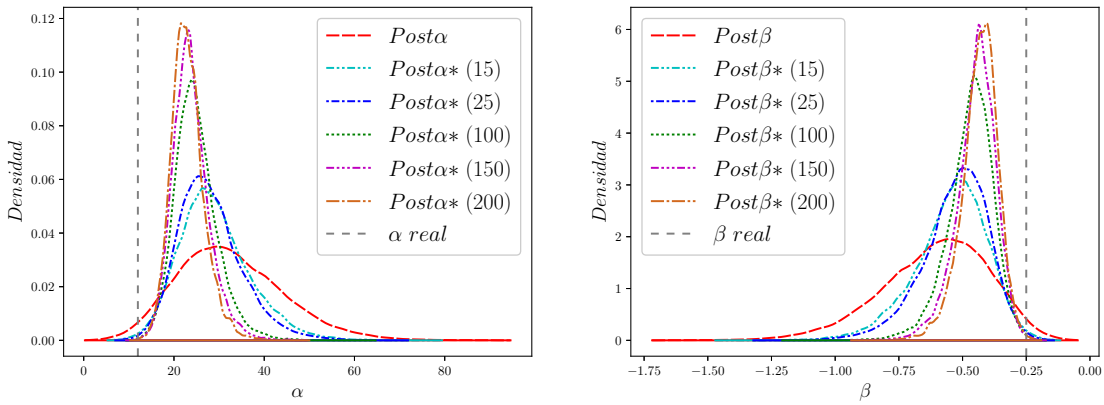
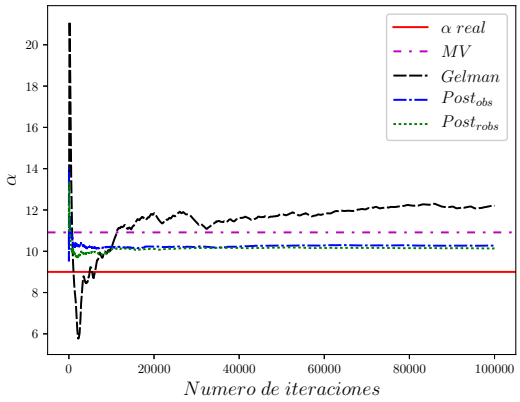


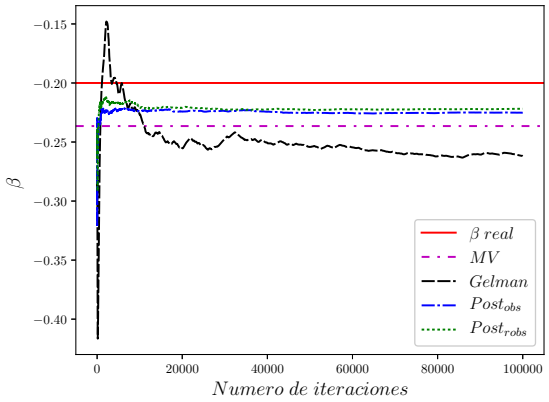
Figura A–6: Distribuciones posteriores para el método $Post_{obs}$ $Post\alpha$ y $Post\beta$ corresponden a la posterior de los datos observados, $Post\alpha^*$ y $Post\beta^*$ son las posteriores para $k = 15, 25, 100, 150, 200$.

A.2 Gráficas de *"running means"* de α y β.

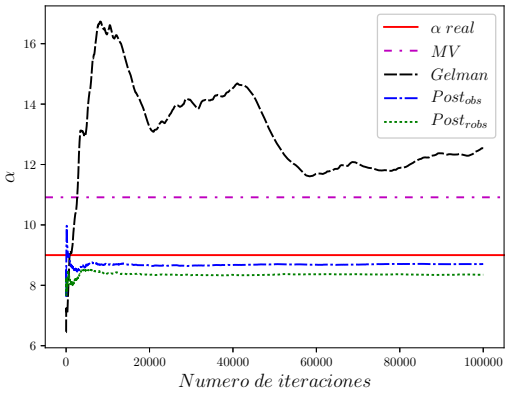
Ejemplo 1.



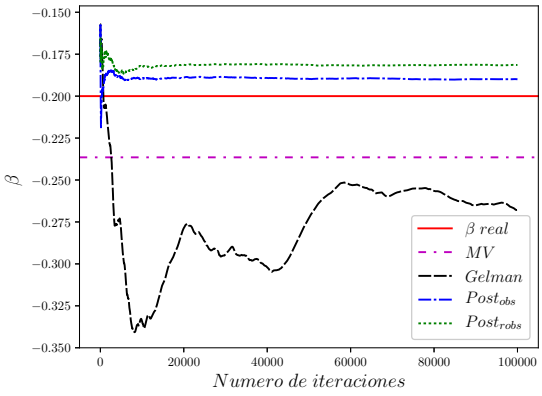
(a) k = 25



(b) k = 25



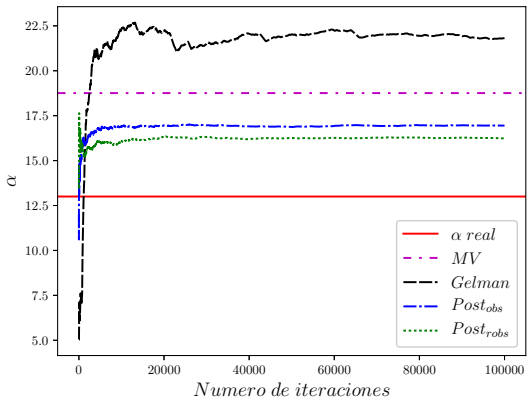
(c) k = 100



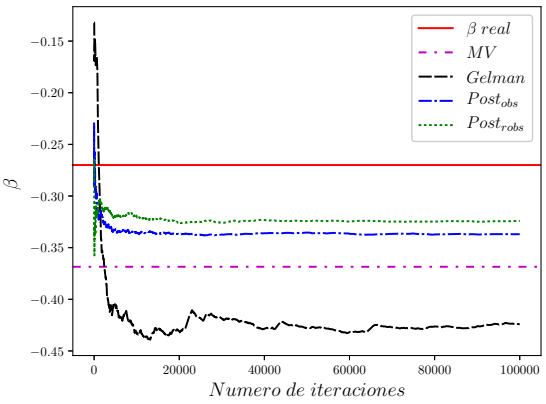
(d) k = 100

Figura A–7: *"Running means"* de α y β

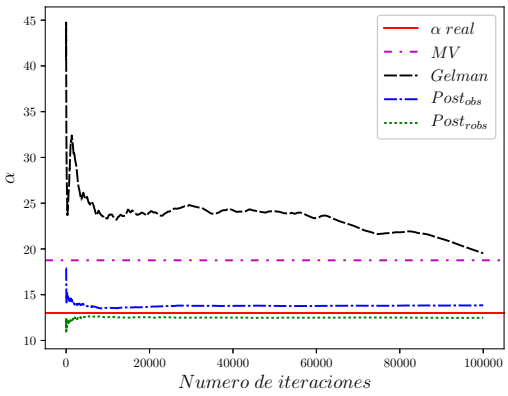
Ejemplo 2.



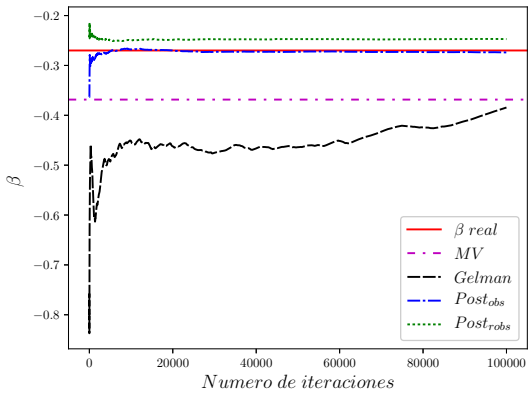
(a) k = 25



(b) k = 25



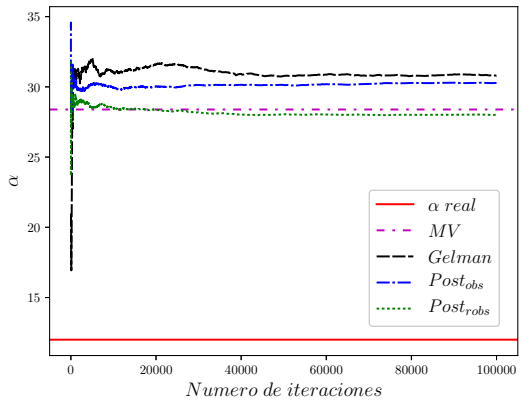
(c) k = 100



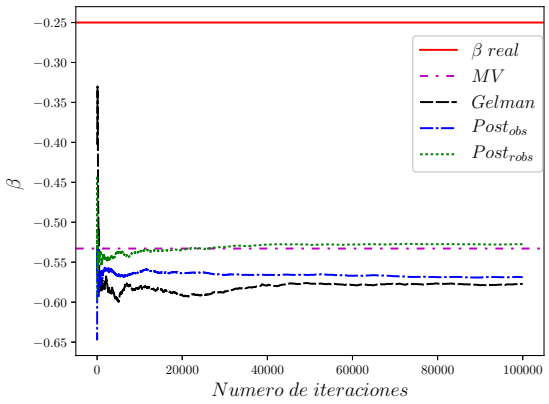
(d) k = 100

Figura A–8: *"Running means"* de α y β

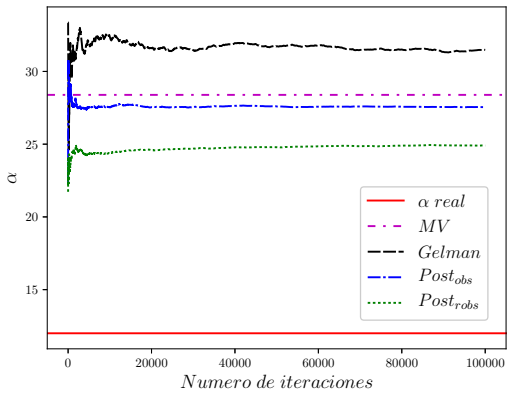
Ejemplo 3.



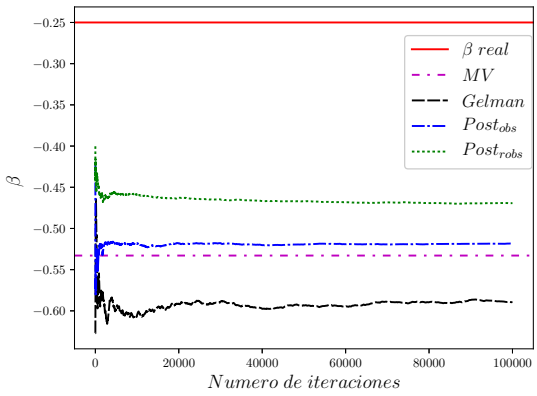
(a) k = 25



(b) k = 25



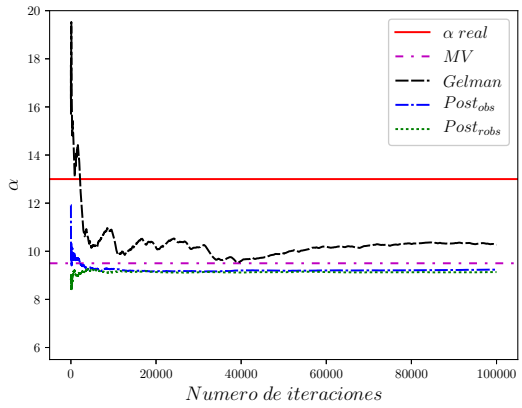
(c) k = 100



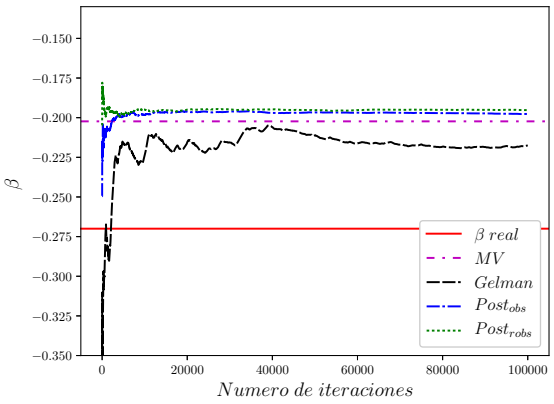
(d) k = 100

Figura A–9: *"Running means"* de α y β

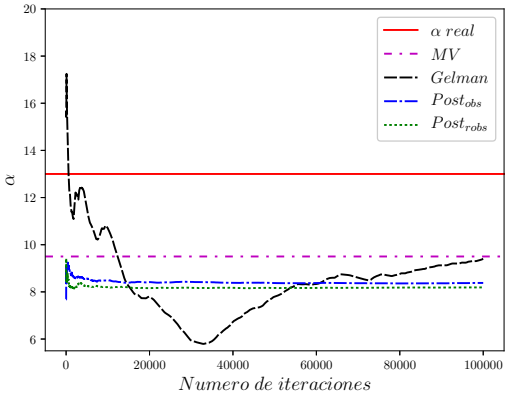
Ejemplo 4.



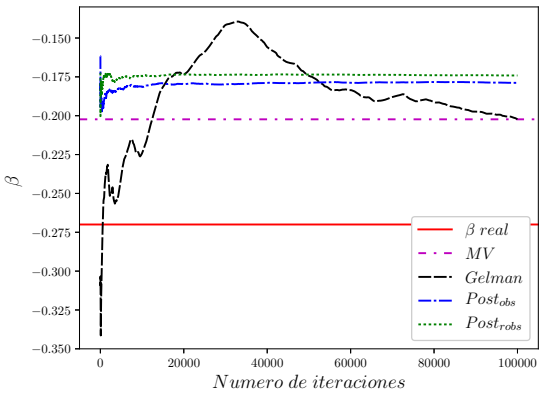
(a) k = 25



(b) k = 25



(c) k = 100



(d) k = 100

Figura A–10: *"Running means"* de α y β

APÉNDICE B
CÓDIGOS

B.1 Código para el Método Post$_{obs}$

1. Generar la distribución posterior de $\alpha =$ a y $\beta =$ b con los datos observados.

```
datos=read.table("datos observados.txt")
x=datos$x
y=datos$y
Nsim        # número de simulaciones
modelo=summary(glm(cbind(y,6-y)~x, family = binomial(link="logit")))
beta=as.vector(modelo$coef[,1])
sigma = modelo$cov.unscaled
scale=2.4/sqrt(2)
a=b=rep(0,Nsim)
a[1]=beta[1]          # valor inicial de a
b[1]=beta[2]          # valor inicial de b
loglike = function(a, b) {
  sum(log(factorial(6))-log(factorial(y)*factorial(6-y))
  +y*(a+b*x)-6*log(1+exp(a+b*x)))}
for (i in 2:Nsim) {
  propa = mvrnorm(1,c(a[i - 1],b[i - 1]),(scale^2)*sigma)
  u=runif(1)
  pi = min((loglike(propa[1], propa[2]) - loglike(a[i - 1], b[i - 1])) , 1)
```

```
  if(log(u)<=pi) {
    a[i] <-propa[1]
    b[i] <-propa[2]
   }
    else {
    a[i] <-a[i - 1]
    b[i] <-b[i - 1]}
}
post=data.frame(a,b)
# eliminar el burning-period
```

2. Generar la distribución posterior de $\alpha^* =$ `newa` y $\beta^* =$ `newb` con los datos completos.

```
# Generar la distribución prob para 24<temp<min(x).
sup=min(x)-1
alpha=post$a
betha=post$b
temp=seq(24,sup)
prob=matrix(0,(sup-23),length(a))
for (t in 1:(sup-23)) {
  z=alpha+betha*temp[t]
  prob[t,]=1/(1+exp(-z))}
#Generar la distribución posterior de newa y newb
Nsim    # número de simulaciones
imp     # cantidad de datos a imputar
modelo=summary(glm(cbind(y,6-y)~x, family = binomial(link="logit")))
beta=as.vector(modelo$coef[,1])
```

```
sigma = modelo$cov.unscaled
newa=newb=rep(0,Nsim)
newa[1]=beta[1]
newb[1]=beta[2]
n=length(x)
scale=2.4/sqrt(2)
loglike = function(a, b,x,y) {
    sum(log(factorial(6))-log(factorial(y)*factorial(6-y))
    +y*(a+b*x)-6*log(1+exp(a+b*x)))}
sim=matrix(c(x,y),n,2)
for (i in 2:Nsim) {
  propa= mvrnorm(1,c(newa[i - 1],newb[i - 1]),((scale^2)*sigma))
  newsim=matrix(0,n,2)
  ind=sample(seq(1:n),n,replace=TRUE)
   for (k in 1:n){
    newsim[k,]=sim[ind[k],]}  # hacer bootstrap con los datos observados
  newx=newsim[,1]
  newy=newsim[,2]
  iind=round(runif(imp,1,(sup-23)))
  jind=sample(seq(1,length(a)),imp,replace = TRUE)
  h=23+iind
  xnew=c(newx,h)           # x completos
  pr=array(0,imp)
  for (j in 1:imp) {
    pr[j]=prob[iind[j],jind[j]]}
  g=rbinom(imp,6,pr)
```

```
  ynew=c(newy,g)           # y completos
  u=runif(1)
  pi = min((loglike(propa[1], propa[2],xnew,ynew)
       - loglike(newa[i - 1], newb[i - 1],xnew,ynew)) , 1)
   if(log(u)<=pi) {
     newa[i] <-propa[1]
     newb[i] <-propa[2]}
     else {
     newa[i] <-newa[i - 1]
     newb[i] <-newb[i - 1]}
}
postcom=data.frame(newa,newb)
# eliminar el burning-period
```

B.2 Código para el Método Post$_{robs}$

1. Generar la distribución posterior de α = a y β = b con los datos observados haciendo bootstrap.

```
datos=read.table("datos observados.txt")
x=datos$x
y=datos$y
Nsim        # número de simulaciones
modelo=summary(glm(cbind(y,6-y)~x, family = binomial(link="logit")))
beta=as.vector(modelo$coef[,1])
sigma = modelo$cov.unscaled
scale=2.4/sqrt(2)
a=b=rep(0,Nsim)
a[1]=beta[1]         # valor inicial de a
```

```
b[1]=beta[2]          # valor inicial de b
n=length(x)
loglike = function(a, b,x,y) {
  sum(log(factorial(6))-log(factorial(y)*factorial(6-y))
  +y*(a+b*x)-6*log(1+exp(a+b*x)))}
sim=matrix(c(x,y),n,2)
for (i in 2:Nsim) {
  propa = mvrnorm(1,c(a[i - 1],b[i - 1]),(scale^2)*sigma)
  newsim=matrix(0,n,2)
  ind=sample(seq(1:n),n,replace=TRUE)
  for (k in 1:n){
    newsim[k,]=sim[ind[k],]}
  xnew=newsim[,1]
  ynew=newsim[,2]
  u=runif(1)
  pi = min((loglike(propa[1], propa[2],xnew,ynew)
       - loglike(a[i - 1], b[i - 1],xnew,ynew)) , 1)
    if(log(u)<=pi) {
     a[i] <-propa[1]
     b[i] <-propa[2]}
     else {
     a[i] <-a[i - 1]
     b[i] <-b[i - 1]}
}
post=data.frame(a,b)
# eliminar el burning-period
```

2. Generar la distribución posterior de $\alpha^* =$ `newa` y $\beta^* =$ `newb` con los datos completos.

```
# Generar la distribución prob para 24<temp<min(x).
sup=min(x)-1
alpha=post$a
betha=post$b
temp=seq(24,sup)
prob=matrix(0,(sup-23),length(a))
for (t in 1:(sup-23)) {
  z=alpha+betha*temp[t]
  prob[t,]=1/(1+exp(-z))}
#Generar la distribución posterior de newa y newb
Nsim    # número de simulaciones
imp     # cantidad de datos a imputar
modelo=summary(glm(cbind(y,6-y)~x, family = binomial(link="logit")))
beta=as.vector(modelo$coef[,1])
sigma = modelo$cov.unscaled
newa=newb=rep(0,Nsim)
newa[1]=beta[1]
newb[1]=beta[2]
n=length(x)
scale=2.4/sqrt(2)
loglike = function(a, b,x,y) {
  sum(log(factorial(6))-log(factorial(y)*factorial(6-y))
  +y*(a+b*x)-6*log(1+exp(a+b*x)))}
for (i in 2:Nsim) {
```

```
  propa= mvrnorm(1,c(newa[i - 1],newb[i - 1]),((scale^2)*sigma))
  iind=round(runif(imp,1,(sup-23)))
  jind=sample(seq(1,length(a)),imp,replace = TRUE)
  h=23+iind
  xnew=c(x,h)       # x completos
  pr=array(0,imp)
  for (j in 1:imp) {
    pr[j]=prob[iind[j],jind[j]]
  }
  g=rbinom(imp,6,pr)
  ynew=c(y,g)       # y completos
  u=runif(1)
  pi = min((loglike(propa[1], propa[2],xnew,ynew)
      - loglike(newa[i - 1], newb[i - 1],xnew,ynew)) , 1)
    if(log(u)<=pi) {
     newa[i] <-propa[1]
     newb[i] <-propa[2]}
     else {
     newa[i] <-newa[i - 1]
     newb[i] <-newb[i - 1]}
}
postcom=data.frame(newa,newb)
# eliminar el burning-period
```

B.3 Código para el Método de Gelman et al. (2014)

Generar la distribución posterior de $\alpha^* =$ `newa` y $\beta^* =$ `newb` con los datos completos.

```
datos=read.table("datos observados.txt")
x=datos$x
y=datos$y
Nsim        # número de simulaciones
imp         # cantidad de datos a imputar
modelo=summary(glm(cbind(y,6-y)~x, family = binomial(link="logit")))
beta=as.vector(modelo$coef[,1])
sigma = modelo$cov.unscaled
scale=2.4/sqrt(2)
newa=newb=rep(0,Nsim)
newa[1]=beta[1]    # valor inicial de newa
newb[1]=beta[2]    # valor inicial de newb
loglike = function(a, b,x,y) {
  sum(log(factorial(6))-log(factorial(y)*factorial(6-y))
  +y*(a+b*x)-6*log(1+exp(a+b*x)))}
for (i in 2:Nsim) {
  propa= mvrnorm(1,c(newa[i - 1],newb[i - 1]),(scale^2)*sigma)
  h=round(runif(imp,24,min(x)))
  xnew=c(x,h)      # x completos
  z=newa[i - 1]+newb[i - 1]*h
  pr=1/(1+exp(-z))
  g=rbinom(imp,6,pr)
  ynew=c(y,g)      # y completos
  u=runif(1)
  pi = min((loglike(propa[1], propa[2],xnew,ynew)
       - loglike(newa[i - 1], newb[i - 1],xnew,ynew)) , 1)
```

```
    if(log(u)<=pi) {
     newa[i] <-propa[1]
     newb[i] <-propa[2]}
     else {
     newa[i] <-newa[i - 1]
     newb[i] <-newb[i - 1]}
}
postcom=data.frame(newa,newb)
# eliminar el burning-period
```