Facultad de Ciencias Sociales y Jurídicas

Grado en Estadística y Empresa

UNIVERSIDAD DE JAÉN
Facultad de Ciencias Sociales y Jurídicas

Trabajo Fin de Grado

ANÁLISIS DE DATOS MEDIANTE TÉCNICAS ESTADÍSTICAS

Alumno: Ortega Bonilla, Sara

Julio, 2017

ÍNDICE

1 RESUMEN

En este trabajo he planteado un modelo de regresión de Poisson utilizando la inferencia bayesiana. Se ha considerado la variable respuesta como el número de víctimas por violencia de género en cada una de las provincias de España durante los años 2000-2014 y un conjunto de covariables que influyen en la variable respuesta. Se han utilizado mapas visuales para llevar a cabo un análisis descriptivo de la variable respuesta y para mostrar los resultados obtenidos del modelo.

Palabras clave: regresión de Poisson, inferencia bayesiana, modelo bayesiano, distribución a priori, distribución a posteriori, mapas.

ABSTRACT

In this paper I have proposed a Poisson regression model using Bayesian inference. The response variable has been considered as the number of victims of gender violence in each of the provinces of Spain during the years 2000-2014 and a set of covariates that influence the response variable. Visual maps have been used to carry out a descriptive analysis of the response variable and to show the results obtained from the model.

Keywords : Poisson regression, Bayesian inference, Bayesian model, prior distribution, posterior distribution, maps.

2 INTRODUCCIÓN Y OBJETIVOS

2.1 MOTIVACIÓN

La Organización de Naciones Unidas[1] define la violencia contra la mujer como *"todo acto de violencia de género que resulte, o pueda tener como resultado un daño físico, sexual, psicológico para la mujer, inclusive las amenazas de tales actos, la coacción o la privación arbitraria de la libertad, tanto si se producen en la vía pública como en la privada".*

En diciembre de 2004 se aprobó en España la Ley Integral contra la Violencia de Género, resultado de una lucha constante de la reivindicación de los derechos de la mujer.

La violencia de género, por desgracia, es uno de los temas que más presentes está en nuestro día a día. Son numerosos los medios de comunicación que hacen referencia a este hecho a través de noticias, donde las protagonistas son principalmente víctimas mortales, mujeres que han fallecido a manos de su pareja, marido o exmarido. Si bien el término violencia de género no excluye a la ejercida contra los varones, por razón de su género, la realidad es que este tipo de violencia es prácticamente testimonial.

Hasta el 9 de marzo de este mismo año, son más de 800 mujeres las que han muerto por violencia machista[2]. Para ser más exactos, 855 mujeres han perdido la vida a manos de sus parejas o exparejas en los últimos 15 años.

Me ha resultado interesante realizar un estudio sobre la violencia de género a través de métodos estadísticos. Para ello he utilizado un enfoque diferente a la estadística clásica, el enfoque bayesiano.

2.2 OBJETIVOS

Los objetivos de este trabajo son, por un lado, modelizar los datos referentes a la violencia de género utilizando un modelo de Poisson mediante inferencia bayesiana y, por otro lado, realizar una representación mediante mapas visuales para un primer análisis descriptivo de la variable respuesta y para los resultados obtenidos del modelo bayesiano.

2.3 ESTRUCTURA DE LA MEMORIA

Una vez explicados los dos primeros capítulos de la memoria, se explicará ahora, de manera resumida, el contenido que se abarca en cada uno de los capítulos posteriores.

En el capítulo ESTADO DEL ARTE se desarrollará todo el contenido teórico de la

[1] OMS. Nota descriptiva. "Violencia de pareja y violencia sexual contra la mujer". Noviembre de 2016. Disponible online: http://www.who.int/mediacentre/factsheets/fs239/es/

[2] Abad, J.M. (15 de junio de 2017). "Las mujeres asesinadas por violencia machista de 2017". El País. Disponible on line: http://www.política.elpaís.com/

memoria que será la base de la metodología aplicada. Una vez explicados los fundamentos teóricos, se procederá a la descripción tanto de los datos utilizados en el estudio, como del modelo más apropiado para modelizar nuestros datos en el capítulo METODOLOGÍA. Todos los resultados obtenidos se explicarán y analizarán en el capítulo RESULTADOS y se explicarán las conclusiones extraídas de la aplicación del modelo y de los resultados obtenidos en nuestro estudio en el capítulo CONCLUSIONES. En el capítulo VÍAS FUTURAS DE ESTUDIO se incluirán otros posibles modelos más complejos y completos para el estudio de los datos, considerados interesantes, que se podrían añadir al estudio realizado en la memoria o bien, realizar de manera paralela. En el capítulo BIBLIOGRAFÍA se incluirán todas las referencias bibliográficas que han servido de apoyo para el desarrollo de la memoria. Por último, toda la información necesaria para el desarrollo de la memoria, pero que no se ha creído conveniente incluir en el cuerpo, propiamente dicho de ésta, formará parte del capítulo ANEXOS.

3 ESTADO DEL ARTE

En este apartado se explicarán, desde lo general a lo específico, los fundamentos teóricos de la metodología utilizada en la memoria.

3.1 REGRESIÓN DE POISSON

La variable respuesta de un modelo de regresión de Poisson expresa valores enteros no negativos y representa un número de sucesos, por ejemplo, las llamadas de teléfono, en un intervalo de tiempo fijado.

La función de probabilidad viene dada por:

$$P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!}\ para\ y\ \epsilon\ \{0, 1, 2, 3, \dots\}$$

donde y representa el número de veces que ocurre el suceso que es objeto de estudio y λ es el parámetro positivo que representa el número medio de veces que se espera que ocurra el suceso en un periodo determinado de tiempo.

Si $Y \sim \mathcal{P}(\lambda)$, entonces $E(Y) = \lambda$ y $V(Y) = \lambda$, por lo que una característica de la distribución de Poisson es que su media y su varianza son coincidentes con el parámetro λ

Podemos ajustar un modelo de regresión de Poisson cuando la variable Y es una variable de recuento y, además, se pretende estudiar su relación con otras variables explicativas del modelo, llamadas covariables, para ver si influyen o no en el comportamiento de la variable respuesta y cómo lo hacen.

El modelo de regresión siguiente es muy utilizado en la regresión de Poisson y se denomina modelo log-lineal:

$$ln\,(\lambda_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$$

$$\lambda_i = e^{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}}$$

La función de regresión del modelo de Poisson se expresa como: $\lambda(x, \beta) = e^{x'\beta}$.

La regresión de Poisson forma parte de los modelos lineales generalizados y la estimación de los coeficientes del modelo se puede realizar mediante el método de máxima verosimilitud L, que expresado en términos de logaritmo sería:

$$ln(L) = \sum_{i=1}^{n} (y_i x'_i \beta - e^{x'_i \beta} - log(y_i))$$

En cuanto a la interpretación de las estimaciones de los coeficientes, en un modelo log-lineal:

1. Se interpreta $e^{\widehat{\beta_0}}$ como el valor esperado de la variable respuesta cuando las

variables explicativas valen todas 0.

2. El valor de $\widehat{\beta_J}$ representa el incremento (si $\widehat{\beta_J} > 0$) o decremento (si $\widehat{\beta_J} < 0$) porcentual en la variable respuesta esperada para un incremento unitario de la covariable o variable explicativa correspondiente. Cuando el resto de predictores permanecen constantes, estimamos un incremento o decremento porcentual de $(e^{\widehat{\beta_J}} - 1)\times 100$ por cada unidad adicional de la variable explicativa correspondiente.

3.2 INFERENCIA BAYESIANA

Es ente subapartado se explicarán, de forma resumida, los aspectos más importantes de la inferencia bayesiana. Seguidamente, de una forma más detallada, se explicarán los modelos basados en inferencia bayesiana, base de todo el desarrollo en los apartados posteriores.

3.2.1 LOS MODELOS BAYESIANOS EN EL SIGLO XXI

Al principio del siglo XXI la estadística bayesiana se puso de moda en la ciencia. Pero hasta finales de los años 80, la estadística bayesiana era solamente considerada como una alternativa interesante a la estadística clásica. La principal diferencia entre la teoría estadística clásica y el enfoque bayesiano es que el enfoque bayesiano considera los parámetros como variables aleatorias, caracterizados por una distribución a priori. Esta distribución a priori se combina con la probabilidad tradicional para conseguir obtener la distribución a posteriori del parámetro de interés, sobre la que se basa la inferencia estadística. Aunque la herramienta principal de la teoría bayesiana es la teoría de la probabilidad, durante muchos años los bayesianos han sido considerados como una minoría non grata por diversas razones. El principal argumento de los estadísticos clásicos era el subjetivo punto de vista del enfoque introducido por los bayesianos en el análisis a través de la distribución a priori. Sin embargo, la historia ha demostrado que la razón principal por la cual la teoría bayesiana no fue capaz de establecer un punto de apoyo, así como un enfoque cuantitativo aceptado para el análisis de datos, fue la insolubilidad o intratabilidad de los cálculos implicados en la distribución a posteriori.

La aparición de nuevas técnicas, como la técnica MCMC (Métodos Markov Chain Monte Carlo), junto al avance y desarrollo de los ordenadores hizo posible paliar este problema de cálculo de la inferencia bayesiana.

En la realización de este trabajo se ha utilizado el software estadístico WinBUGS a través de la consola de RStudio. WinBUGS es un software estadístico para el análisis de

modelos bayesianos que utiliza los métodos MCMC, mencionados anteriormente.

3.2.2 ESTADÍSTICA BAYESIANA VERSUS ESTADÍSTICA CLÁSICA.

Aunque en el apartado anterior hemos realizado una breve comparación entre la estadística bayesiana y la clásica, vamos a abordar esta cuestión de una forma más detallada.

El enfoque bayesiano utiliza dos tipos de informaciones, la información muestral y la información a priori y las combina utilizando la Regla de Bayes de probabilidad condicionada.

Como información a priori se considera cualquier tipo de información que sea tan valiosa como la muestral, por ejemplo, juicios de expertos o resultados de estudios anteriores.

Como destacamos en la Tabla 1, el enfoque bayesiano no tiene solamente ventajas, sino que también tiene algunas desventajas como la dificultad de cálculo de la distribución a posteriori, que es uno de los problemas principales, así como la incorporación del punto de vista subjetivo a través, por ejemplo, del juicio de expertos y también, la escasez de programas y paquetes estadísticos que permitan aplicar este enfoque bayesiano.

En oposición a este enfoque, nos encontramos con el enfoque frecuentista o clásico, que utiliza solamente información muestral. Hay que resaltar que uno de los contras de este enfoque es que se requiere, en su gran mayoría, de un tamaño muestral suficientemente grande, mientras que con el enfoque bayesiano no hay problemas en cuanto a tamaño de la muestra.

Tabla 1. Ventajas y desventajas del enfoque bayesiano

Ventajas	**Desventajas**
• Permite abordar problemas más complejos y completos. • Información más completa (no sólo información muestral). • No requiere un tamaño muestral suficientemente grande.	• Dificultad de cálculo de la distribución a posteriori. • Incorporación de un punto de vista subjetivo. • Software y paquetes estadísticos de modelos bayesianos disponibles escasos.

Fuente: elaboración propia

3.2.3 MODELOS BASADOS EN INFERENCIA BAYESIANA

Como hemos mencionado en el apartado LOS MODELOS BAYESIANOS EN EL SIGLO XXI, la estadística bayesiana se diferencia de la clásica en el hecho de que los parámetros son consideramos como variables aleatorias. Este motivo hace necesario que las distribuciones a priori deban ser definidas inicialmente. Esta distribución tiene especial interés en el cálculo de la distribución a posteriori $f(\boldsymbol{\theta}|\boldsymbol{y})$ de los parámetros $\boldsymbol{\theta}$ dada la variable observada $\boldsymbol{y}$.

La distribución a posteriori se puede calcular, según el Teorema de Bayes, de la siguiente forma:

$$f(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\, f(\boldsymbol{\theta})}{f(\boldsymbol{y})} \propto f(\boldsymbol{y}|\boldsymbol{\theta})\, f(\boldsymbol{\theta})$$

donde:

$f(\boldsymbol{\theta}|\boldsymbol{y})$ es la densidad a posteriori;

$f(\boldsymbol{y}|\boldsymbol{\theta})$ es la verosimilitud de θ aportada por y ;

$f(\boldsymbol{\theta})$ es la densidad a priori.

La anterior fórmula conduce a la siguiente afirmación: la densidad a posteriori es proporcional a la verosimilitud por la densidad a priori.

La distribución a posteriori reúne tanto información a priori disponible por el investigador, como la información extraída de la observación de los datos, que es expresada por la distribución a priori y la verosimilitud, respectivamente.

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})$$

O lo que es igual, la densidad a posteriori es proporcional a la densidad a priori por la verosimilitud.

Especificar la distribución a priori es importante en la inferencia bayesiana por, como se ha podido observar, su importancia en la distribución a posteriori.

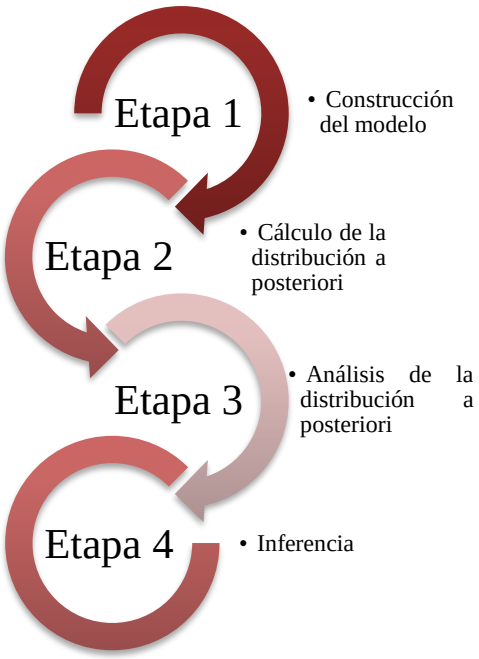
La media a priori proporciona una estimación puntual previa del parámetro de interés, mientras que la varianza a priori expresa nuestra incertidumbre con respecto a dicha estimación. Cuando creemos fuertemente a priori que esta estimación de la media a priori es acertada, estableceremos un valor bajo de la varianza; por el contrario, la gran incertidumbre o la ignorancia relativa a la media a priori suele ser expresada por un gran valor de la varianza. Este procedimiento es llamado elicitación de conocimientos previos. Cuando no disponemos de información previa, que es lo que suele ocurrir generalmente, necesitaremos especificar a priori todo aquello que no va a influir en la distribución a

posteriori y “dejar que los datos hablen por sí solos”. A las distribuciones de este tipo se les suele llamar distribuciones a priori no informativas.

Los momentos de la distribución a posteriori, pueden ser utilizados para hacer inferencia sobre la incertidumbre del vector de parámetros $\boldsymbol{\theta}$. Para ser más específicos, medidas de posición centrales como la media, mediana o moda a posteriori pueden ser utilizadas como estimación puntual, mientras que los cuantiles a posteriori $q/2$ y $1 - q/2$ pueden ser usados como $(1 - q)\ 100\%$ intervalos de confianza a posteriori.

Se puede observar en la Ilustración 1 que el todo el procedimiento relacionado con el modelo bayesiano se ha dividido en cuatro etapas. En una primera etapa consideraremos un modelo, con hipótesis razonables. Calcularemos la distribución a posteriori de interés con un método computacional apropiado en una segunda etapa. Luego realizaremos un análisis utilizando las medidas descriptivas, gráficos e intervalos de confianza. Por último, sacaremos conclusiones relacionadas con el problema que estamos tratando.

Ilustración 1. Etapas del procedimiento de un modelo bayesiano



Fuente: elaboración propia

El procedimiento a seguir en el modelo bayesiano es muy importante. Dada esta importancia, se ha creído conveniente especificar más el procedimiento a seguir en las etapas.

En la primera etapa de construcción del modelo podemos seguir el procedimiento que se describe a continuación:

1. Identificar la variable respuesta Y (la variable principal del problema) y los datos correspondientes y.
2. Encontrar la distribución que describe a la variable Y.
3. Identificar las covariables, las variables explicativas que influyen en el comportamiento de la variable Y.
4. Construir la estructura para los parámetros de la distribución.
5. Especificar los valores iniciales de la distribución a priori.
6. Comprobar la verosimilitud del modelo y, si es posible, realizar una comparación con otros modelos mediante criterios de comparación como, por ejemplo, el criterio DIC.

En la segunda etapa identificaremos, en primer lugar, el método para calcular la distribución y, en segundo lugar, implementaremos el método elegido para estimar la distribución a posteriori. Se puede elegir, por ejemplo, un método analítico o utilizar una técnica de simulación.

Para llevar a cabo el análisis de la distribución a posteriori en la tercera etapa, se proponen algunas medidas a continuación. Si bien, es conveniente aclarar que no todas las medias son necesarias para realizar este análisis, por lo que se pueden elegir todas las medidas propuestas o bien alguna de ellas.

- Se pueden utilizar algunos gráficos para analizar la distribución a posteriori, como el histograma o el gráfico de barras; también se pueden utilizar diagramas de caja y, para estudiar correlación, un gráfico de dos variables.
- Se pueden calcular diferentes medidas a posteriori como la media, mediana, desviación típica, correlaciones y cuantiles o intervalos de confianza a posteriori al 95 o 99%.

En la última etapa llamada inferencia se extraen las conclusiones del modelo que está siendo objeto de nuestro estudio. Se analiza si el modelo es o no apropiado a nuestros datos, si las conclusiones finales a las que se ha llegado son razonables y otras cuestiones relacionadas con el análisis de los resultados. Si fuera necesario, se podría ampliar o modificar el modelo y se tendrían entonces que repetir las tres etapas anteriores.

3.3 INFERENCIA BAYESIANA EN EL MODELO DE REGRESIÓN DE POISSON

Asumimos que nuestros datos siguen una distribución de Poisson. Por lo tanto:

$$y_i \sim \mathcal{P}(\lambda_i), \quad i = 1, , \dots , n.$$

Como se ha comentado en el apartado REGRESIÓN DE POISSON de este capítulo, se usa con mucha frecuencia el modelo log-lineal de Poisson. Este modelo es el que se utilizará en inferencia bayesiana y será el siguiente:

$$\log \lambda_i = \beta_0 + \sum_{i=1}^{n} \beta_j x_{ij} = \boldsymbol{X}_{(i)}\boldsymbol{\beta}$$

$$j = 1, \dots, k$$

donde k representa el número total de covariables del modelo.

Hay que recordar que, en el modelo bayesiano, nuestros parámetros serán variables aleatorias y seguirán una distribución determinada. En nuestro modelo,

$$\beta_j \sim \mathcal{N}\ (\mu, \sigma^2)$$

Es necesario aclarar que el programa en WinBUGS se calcula σ^2 como $1/t$ y, es el valor de t es que tendremos que introducir como valor de σ^2 en la distribución de los parámetros. Si queremos que σ^2 tome un valor muy alto, tenemos que establecer un valor bajo de t; si queremos que, por el contrario, que σ^2 tome un valor muy bajo, tenemos que establecer un valor alto de t. Esto es, a efectos de cálculo, tendremos que introducir en nuestra ventana de RStudio lo siguiente:

$$\beta_j \sim \mathcal{N}\ (\mu, t)$$

Una vez planteado el modelo, tal y como aparece en el ANEXO 5. MODELO, donde se establece gran parte de los pasos a seguir en la primera etapa del procedimiento a seguir en un modelo bayesiano, se completará esta primera etapa y se llevarán a cabo las etapas restantes, tal y como aparece explicado en el subapartado MODELOS BASADOS EN INFERENCIA BAYESIANA y, de forma gráfica, en la Ilustración 1.

3.4 MANEJO DE DATOS ESPACIALES

En este apartado se explicarán los pasos a seguir para poder realizar la representación de datos en mapas visuales en el programa RStudio.

La Real Academia Española[3] define el término mapa como *"representación geográfica de una parte de la superficie terrestre, en la que se da información relativa a una ciencia determinada"*. Esto es, nuestro objetivo a cumplir con la utilización de los mapas es representar ciertas variables a estudiar, con el fin de poderlas visualizar de una forma rápida y sencilla.

Se utilizará el programa RStudio para el tratamiento y manejo de datos espaciales y por

[3] Real Academia Española. (2001). *Diccionario de la lengua española* (22.a ed.). Consultado en http://www.rae.es/rae.html

este motivo, se ha creído conveniente explicar en este apartado todo lo referente al manejo de datos espaciales en este programa.

En primer lugar, tenemos que instalar y cargar los paquetes que aparecen en la Tabla 2 para poder empezar a trabajar con los datos espaciales, tal y como se especifica en el ANEXO 1. PAQUETES Y DATOS.

Tabla 2. Paquetes a instalar y cargar en RStudio

Paquetes de RStudio	
✓ sp	✓ maptools
✓ rgeos	✓ png
✓ rgdal	✓ raster
✓ pbapply	✓ geosphere
✓ FNN	✓ RColorBrewer

Fuente: elaboración propia

Una vez instalados y cargados los paquetes anteriores, tenemos que descargar el mapa de una página web. En este trabajo, se ha descargado el mapa de España con las poligonales por provincias de la página web del Instituto de Estadística y Cartografía de Andalucía (IECA). La poligonal hará referencia al área que queremos representar en el mapa y sus respectivos límites; en este caso, nuestra poligonal serán las provincias de España que queremos representar en nuestro mapa. Por este motivo, se ha procedido a descargar en las bases cartográficas de referencia del IECA el archivo zip *G19 Contexto España*, que contiene diversas capas de información geográfica de datos diversos, incluida la capa geográfica con información referente a las divisiones administrativas. La capa que contiene esta información la encontramos en el archivo con el nombre de *ctx01_limites* y será la que importaremos bajo el formato shapefile (shp). Para realizar la importación utilizaremos la función readShapeSpatial que permite la lectura de los datos espaciales de un archivo shapefile.

El proceso descrito anteriormente se encuentra especificado para RStudio en el ANEXO 2. MAPA ESPAÑA.

En la Tabla 3 aparecen algunas de las funciones que se pueden utilizar para la representación de mapas, así como una breve descripción de cada una de ellas. Es necesario aclarar que no es necesario utilizarlas todas y que también faltan por nombrar

y explicar otras funciones. La utilización de unas u otras dependerá de la persona que realice el mapa y del formato y aspecto visual que quiera darle.

Tabla 3. Funciones de RStudio para representar mapas y descripciones

Funciones	**Descripción**
jpeg	Guardar el gráfico en el formato elegido. Además del formato JPEG se puede elegir guardar en otros formatos (BMP, PNG o TIFF). Permite dar nombre de guardado al gráfico, definir calidad, altura o anchura del gráfico, entre otros.
plot	Representar los datos (data.frame). Permite, por ejemplo, elegir los colores de representación o el color del borde de los límites del mapa.
title	Dar nombre al título del mapa. Permite definir el tamaño de la letra y el color del título, entre otros.
legend	Establecer la legenda del mapa. Se puede definir la posición en la que se quiere que aparezca la leyenda en el mapa, nombres a mostrar en la leyenda y colores de relleno, así como el tamaño del texto.
text	Añadir texto al mapa. Permite establecer el tamaño del texto y color.

Fuente: elaboración propia

4 METODOLOGÍA

En este capítulo se realizará una descripción detallada de los datos utilizados en el estudio, así como el planteamiento del modelo apropiado para nuestros datos.

4.1 DESCRIPCIÓN DE LOS DATOS UTILIZADOS

En este apartado se proporcionará toda la información que concierne a los datos: variables que se incluyen, período al que hacen referencia y fuentes.

- En nuestro modelo la variable observada va a ser *víctimas* que representa el número de víctimas mortales en cada una de las provincias de España por violencia de género desde el año 2000 hasta el 2014. Los datos referentes a esta variable se han extraído del Portal Estadístico de Violencia de Género del Ministerio de Sanidad, Servicios Sociales e Igualdad del Gobierno de España.
- Las covariables son las siguientes:
    - *separaciones* representa el número de separaciones que se han producido en cada una de las provincias de España.
    - *matrimonios* representa el número de matrimonios celebrados en cada una de las provincias de España.
    - *denuncias* representa el número de denuncias por violencia de género que se han producido en cada una de las provincias de España.
    - *divorcios* representa el número de divorcios que se han producido en cada una de las provincias de España.
    - *llamadas* representa el número de llamadas al 016 que se han producido en cada una de las provincias de España.
    - *nacionalidad* representa el número de personas extranjeras, tanto hombres como mujeres, que tienen su residencia habitual en España y adquieren la nacionalidad española en cada una de las provincias de España.
    - *ordenes* representa el número de órdenes de protección por violencia de género concedidas en cada una de las provincias de España.

    Por la dificultad de acceso a los datos, el año al que hacen referencia todas estas variables es al año 2014.

    En cuanto a las fuentes de extracción, las covariables separaciones, matrimonios, divorcios y nacionalidad se han extraído del Instituto Nacional de Estadística (INE). El resto han sido extraídas del Portal Estadístico de Violencia de Género.

Hemos transformado estas covariables en tasas para eliminar el efecto del tamaño de la población. En la Tabla 4 podemos observar los anteriores nombres de las covariables sin

tasa y su nombre actual, una vez calculada la tasa. Para dicho cálculo se ha dividido, para cada una de las provincias, cada covariable entre la población y se ha multiplicado por 100; por tanto, las tasas representan el número de eventos por cada 100 habitantes.

Tabla 4. Nombres de las variables calculadas como tasas

Nombre variable	**Nombre variable (tasa)**
✓ *separaciones*	✓ *pseparaciones*
✓ *matrimonios*	✓ *pmatrimonios*
✓ *denuncias*	✓ *pdenuncias*
✓ *divorcios*	✓ *pdivorcios*
✓ *llamadas*	✓ *pllamadas*
✓ *nacionalidad*	✓ *pnacionalidad*
✓ *ordenes*	✓ *pordenes*

Fuente: elaboración propia

Además, en nuestro modelo de regresión consideramos la variable *población*, que representará el número de habitantes en cada una de las provincias de España en el año 2014 y ha sido extraída del Instituto Nacional de Estadística (INE).

En nuestros datos aparecen dos variables más referentes a las provincias. Una de ellas es la variable *provincia* que es el nombre de cada una de las provincias de España; la otra variable es *cod_provincia* que relaciona cada una de las provincias con su código de provincia. Ambas variables han sido extraídas del Instituto Nacional de Estadística (INE).

Todas las variables anteriormente definidas formarán parte, por columnas, del fichero de texto *violencia_genero_tasas* que será utilizado para importar los datos a RStudio, tal y como aparece en el ANEXO 1. PAQUETES Y DATOS.

Se ha creído conveniente, en un primer análisis descriptivo de la variable respuesta, realizar un mapa, puesto que es la mejor forma de poder visualizar la variable respuesta *víctimas*. Para ello, se representarán dos mapas, uno referente al número de víctimas por violencia de género en las distintas provincias de España desde el año 2000 al 2014 y otro, hará referencia al número de víctimas por violencia de género por cada 100.000 habitantes en cada una de las provincias de España para el mismo periodo de tiempo que el anterior. Ambos serán representados y analizados en el capítulo RESULTADOS.

4.2 PLANTEAMIENTO DEL MODELO

Una vez explicada la metodología a seguir y las variables a usar, se va a proceder a plantear el modelo como un modelo de regresión de Poisson utilizando la inferencia bayesiana. Para ello, seguiremos los pasos para la primera etapa del modelo bayesiano explicados en subapartado MODELOS BASADOS EN INFERENCIA BAYESIANA.

1. Nuestra variable respuesta Y es *victimas* que representa el número de víctimas por violencia de género en cada una de las provincias de España y los datos correspondientes a nuestra variable han sido descargados del Portal Estadístico de Violencia de Género, tal y como se explicó en el apartado DESCRIPCIÓN DE LOS DATOS UTILIZADOS.
2. $y_i \sim \mathcal{P}(\lambda_i), \quad i = 1, 2, \dots, 52$

   donde i representa el número de provincias, siendo un total de 52 provincias las existentes en España.
3. Las covariables que influyen en la variable respuesta son:
   - *pseparaciones*
   - *pmatrimonios*
   - *pdenuncias*
   - *pdivorcios*
   - *pllamadas*
   - *pnacionalidad*
   - *pordenes*

   Cada una de ellas han sido explicadas en el apartado DESCRIPCIÓN DE LOS DATOS UTILIZADOS.

Hasta este paso, nuestro modelo de regresión es el siguiente:

$$log(\lambda_i) = log(poblacion_i) + \beta_0 + \beta_1 * pseparaciones_i + \beta_2 * pmatrimonios_i + \beta_3 * pdenuncias_i + \beta_4 * pdivorcios_i + \beta_5 * pllamadas_i + \beta_6 * pnacionalidad_i + \beta_7 * pordenes_i$$

4. Los parámetros del modelo van a seguir la siguiente distribución a priori:

   $$\beta_j \sim \mathcal{N}(0.0, 1.0\text{E}-6), \quad j = 0, 1, \dots, 7$$

   donde

   $$\beta_j \in (-100, 100)$$
5. Especificar los valores iniciales de la distribución a priori: se ha seleccionado la distribución normal para generar los valores iniciales de los parámetros a priori

que, en este caso, son los β_j .

6. Este paso de realizar una comparación mediante un criterio con otros modelos no lo podemos llevar a cabo, puesto que solamente tenemos un modelo.

El modelo ha sido ajustado mediante la función bugs de la librería R2WinBUGS de R, tal y como se describe en el ANEXO 6. MODELO POISSON.

5 RESULTADOS

En este capítulo se analizarán, por un lado, los resultados referentes al análisis descriptivo de la variable respuesta y, en segundo lugar, todos los resultados relacionados con el modelo bayesiano planteado en el anterior apartado PLANTEAMIENTO DEL MODELO. Se utilizarán mapas donde se representarán los resultados obtenidos para una mejor visualización.

En lo que concierne a la representación de los mapas, es conveniente aclarar tres aspectos:

1. Para dar nombre a las provincias en el mapa se han utilizado abreviaturas, tal y como se muestra en la Tabla 5, donde se puede observar la asignación a cada una de las provincias de España de su correspondiente abreviatura.
2. En el mapa de España aparece, dentro de la provincia de Álava, un área en color blanco y esto puede llevar a confusión al pensar que podría ser una provincia que no se ha tenido en cuenta. Este área es el Condado de Treviño o isla administrativa de Treviño que, a pesar de pertenecer administrativamente a Burgos, se encuentra rodeada de la capital vasca. Desde finales de 1999 se declaró, por motivos políticos y territoriales, en "indefinición administrativa".[4]
3. Se ha creído conveniente realizar una partición en intervalos de los datos obtenidos para la variable respuesta analizada con el fin de poder visualizar las diferencias entre las provincias de una forma más sencilla. Para los valores de los intervalos, el extremo inferior es abierto, mientras que el superior se ha considerado cerrado. Por ello, cuando se haga referencia al intervalo, el valor del extremo inferior no se considerará incluido en este.

En un primer análisis descriptivo de la variable respuesta *victimas* se ha realizado un mapa representado en la Ilustración 2, donde aparece el número de víctimas por violencia de género en cada una de las provincias españolas desde el año 2000 hasta el 2014.

En la Ilustración 2 se observa que las provincias que han tenido un mayor número de víctimas entre los años 2000 y 2014 son Madrid y Barcelona, con un número de víctimas entre 73 y 91. Valencia y Alicante se encuentran en el intervalo medio con un número de víctimas comprendido entre 37 y 55. De las provincias restantes, la mayoría tienen un número de víctimas comprendido entre 2 y 19, a excepción de las provincias de Pontevedra, La Coruña, Asturias, Bizkaia, Gerona, Tarragona, Murcia, Islas Baleares, Gran Canaria, Tenerife y las provincias andaluzas de Sevilla, Málaga, Granada y Almería,

[4] Martínez, I. (22 de febrero de 2000). "Treviño, una isla en el mapa de Álava". El País. Disponible on line: http://www.elpaís.com/diario/

con un número de víctimas entre 19 y 37. No encontramos provincias con un número de víctimas igual a 0.

La realización de este mapa en RStudio se describe en el ANEXO 3. MAPA VÍCTIMAS.

Tabla 5.Abreviaturas de las provincias para su representación en el mapa

VI	Álava	**AB**	Albacete	**A**	Alicante
AL	Almería	**O**	Asturias	**AV**	Ávila
BA	Badajoz	**IB**	Islas Baleares	**B**	Barcelona
BU	Burgos	**CC**	Cáceres	**CA**	Cádiz
S	Cantabria	**CS**	Castellón	**CE**	Ceuta
CR	Ciudad Real	**CO**	Córdoba	**C**	La Coruña
CU	Cuenca	**GI**	Gerona	**GR**	Granada
GU	Guadalajara	**SS**	Gipuzkoa	**H**	Huelva
HU	Huesca	**J**	Jaén	**LE**	León
L	Lérida	**LU**	Lugo	**M**	Madrid
MA	Málaga	**ML**	Melilla	**MU**	Murcia
NA	Navarra	**OR**	Orense	**P**	Palencia
GC	Las Palmas	**PO**	Pontevedra	**LO**	La Rioja
SA	Salamanca	**SG**	Segovia	**SE**	Sevilla
SO	Soria	**T**	Tarragona	**TF**	Tenerife
TE	Teruel	**TO**	Toledo	**V**	Valencia
VA	Valladolid	**BI**	Bizkaia	**ZA**	Zamora
Z	Zaragoza				

Fuente: elaboración propia

Se puede pensar que los resultados del mapa de Ilustración 2 en relación a la población son lógicos; a más población más víctimas por violencia de género. Por este motivo, como se comentó en el apartado DESCRIPCIÓN DE LOS DATOS UTILIZADOS, se ha realizado también otro mapa en la Ilustración 3 donde se representa el número de víctimas en cada una de las provincias de España por cada 100.000 habitantes desde el año 2000 hasta el 2014.

Se observa en la Ilustración 3 que Melilla tiene la tasa más alta de víctimas por violencia de género, con una tasa entre 4.88 y 5.92, por cada 100.000 habitantes. Le sigue la provincia de Almería con una tasa comprendida entre 3.83 y 4.88. Las provincias de Granada, Cuenca, Tarragona, Islas Baleares y Tenerife se sitúan por debajo por una tasa

que oscila entre 2.79 y 3.83. Las provincias restantes quedarían representadas en el mapa en los dos primeros intervalos, con una tasa más baja, comprendida entre 0.694 y 1.74 para el primer intervalo y de 1.74 a 2.79 para el segundo intervalo.

La realización de este mapa en RStudio aparece detallada en el ANEXO 4. MAPA VÍCTIMAS POR CADA 100.000 HABITANTES.

Una vez comentados los resultados del análisis descriptivo para la variable respuesta, pasaremos ahora a comentar los resultados extraídos del modelo.

En la Tabla 6 se muestran la media, la desviación típica y los diferentes cuantiles para cada uno de nuestros parámetros del modelo. Los asteriscos se sitúan en los parámetros que son significativos, aquellos para los que su intervalo de confianza no contiene al 0. Los parámetros β_2, β_3 y β_5 son significativos y, por tanto, las variables *pmatrimonios, pdenuncias* y *pllamadas*, respectivamente, son significativas. Las variables *pmatrimonios* y *pllamadas* tienen un efecto significativo negativo en la variable respuesta, mientras que la variable *pdenuncias* tiene un efecto significativo positivo. El modelo indica que:

- Las provincias con un mayor número de matrimonios por cada 100 habitantes tienen un número de víctimas por violencia de género menor.
- En aquellas provincias donde el número de denuncias por cada 100 habitantes es mayor se encuentra un número de muertes por violencia de género mayor.
- Las provincias donde el número de llamadas por cada 100 habitantes es mayor, el número de víctimas por violencia de género es menor.

Tabla 6. Resultados del modelo para los parámetros

$\boldsymbol{\beta_j}$	**mean**	**sd**	**2.5%**	**25%**	**50%**	**75%**	**97.5%**
β_0	- 10.3	0.4	- 11.2	- 10.6	- 10.3	- 10.1	- 9.6
β_1	7.4	14.9	- 21.9	- 2.7	7.4	17.4	36.9
β_2 *	- 3.9	1.2	- 6.3	- 4.7	- 3.9	- 3.2	- 1.4
β_3 *	3.0	0.6	1.8	2.6	3.0	3.4	4.2
β_4	1.5	1.3	- 1.2	0.6	1.5	2.3	4.1
β_5 *	- 3.4	0.9	- 5.2	- 4.0	- 3.4	- 2.8	- 1.7
β_6	- 0.1	0.2	- 0.4	- 0.2	- 0.1	0.0	0.2
β_7	1.0	2.5	- 3.7	- 0.7	1.0	2.6	5.6

Fuente: elaboración propia

Los valores medianos estimados del modelo para los λ_i de cada provincia, es decir, para

el número mediano estimado de víctimas por violencia de género para cada una de las provincias, se muestran en la Ilustración 4.

Se observa en este mapa que Madrid y Barcelona son las provincias con un valor mediano estimado de λ_i mayor, con un valor comprendido entre 74.4 y 92.5. Se sitúan por detrás, con un menor valor mediano estimado de λ_i, entre 37.7 y 56 las provincias de Valencia y Alicante. La mayoría de las provincias restantes tienen el valor mediano estimado de λ_i más bajo, comprendido entre 1.19 y 19.5, a excepción de Asturias, Zaragoza, Tarragona, Murcia, Sevilla, Cádiz, Málaga, Granada, Islas Baleares, Tenerife y Gran Canaria, cuyo valor mediano estimado está comprendido entre 19.5 y 37.7.

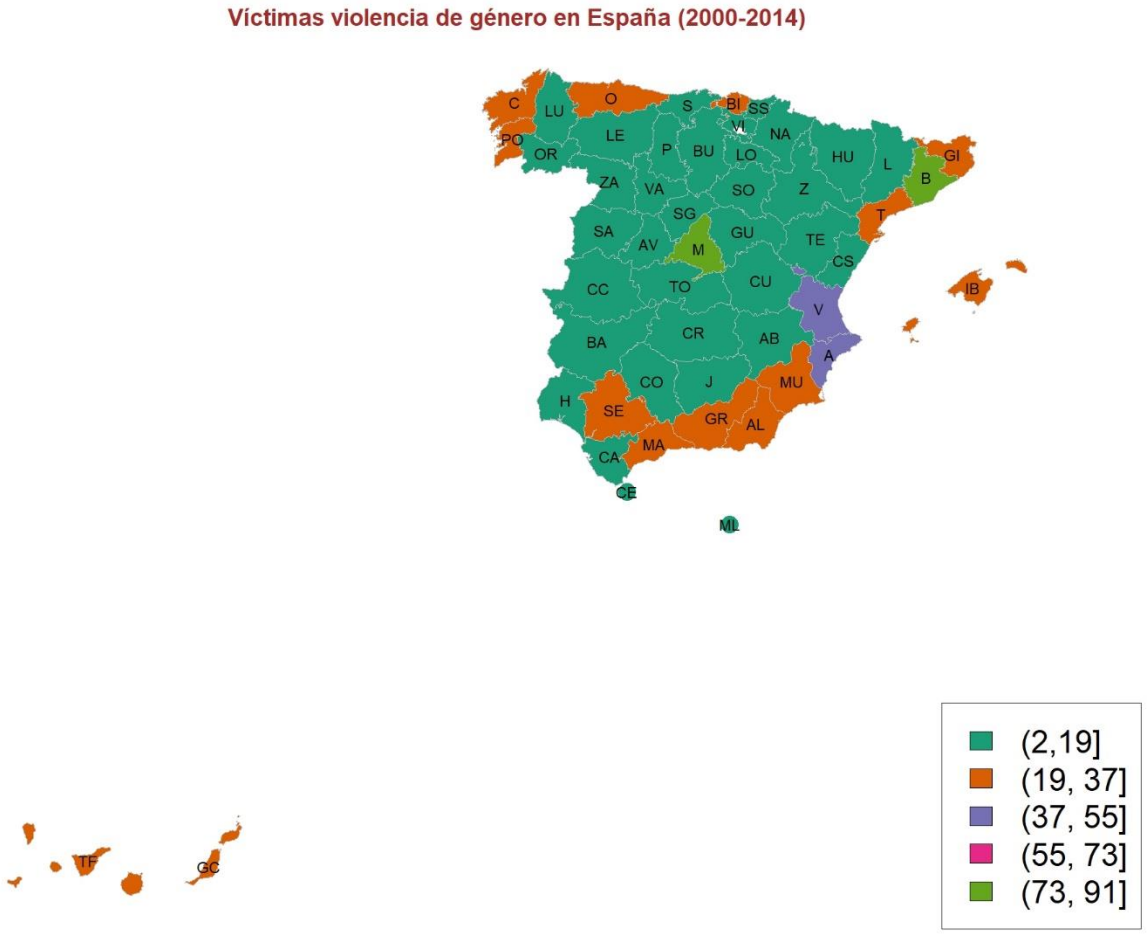
La realización de este mapa en RStudio se encuentra descrita en el ANEXO 7. MAPA LAMBDAS.

Se ha creído conveniente representar en la Ilustración 5 los valores medianos estimados del modelo para los λ_i de cada provincia española por cada 100.000 habitantes, con el fin de poder establecer una comparación entre el valor observado de la variable respuesta y la mediana de la distribución esperada del modelo. Esta vez, los resultados obtenidos se han representado en cuatros intervalos, en lugar de cinco intervalos como en el resto de mapas, puesto que los valores estimados medianos oscilaban entre 1 y 4, aproximadamente.

En este mapa de la Ilustración 5 se observa que la provincia que tiene un valor estimado mediano mayor de λ_i por cada 100.000 habitantes, entre 3.11 y 3.76, es la provincia de Tarragona. Le siguen, con un valor un poco más pequeño, entre 2.45 y 3.11, Murcia y las Islas Baleares. El resto de provincias se encuentran divididas en el mapa en los dos intervalos restantes, con unos valores medianos estimados para λ_i por cada 100.000 habitantes, de mayor a menor, de 1.8 a 2.45 y de 1.14 a 1.8, respectivamente.

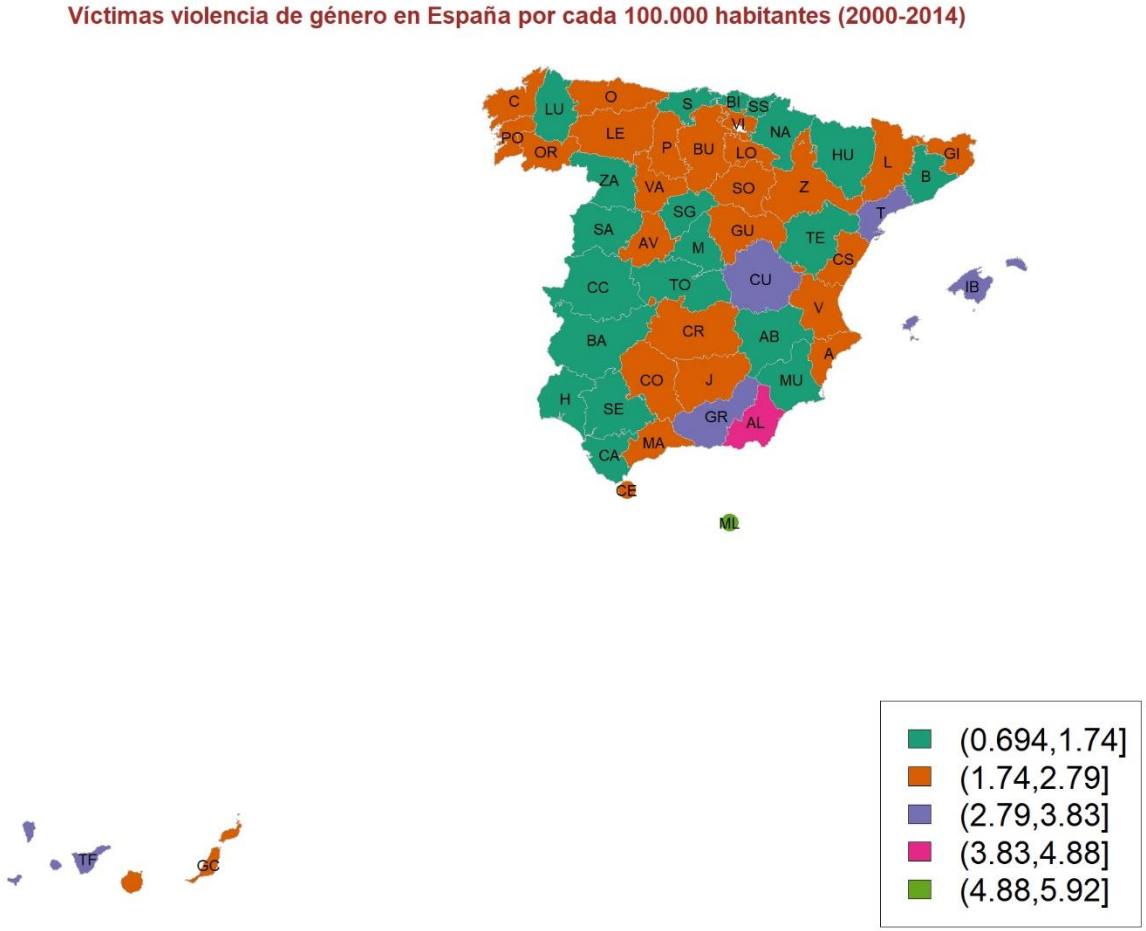
La realización de este mapa en RStudio se encuentra descrita en el ANEXO 8. MAPA LAMBDAS POR CADA 100.000 HABITANTES.

Ilustración 2. Mapa víctimas violencia de género en España (2000-2014)



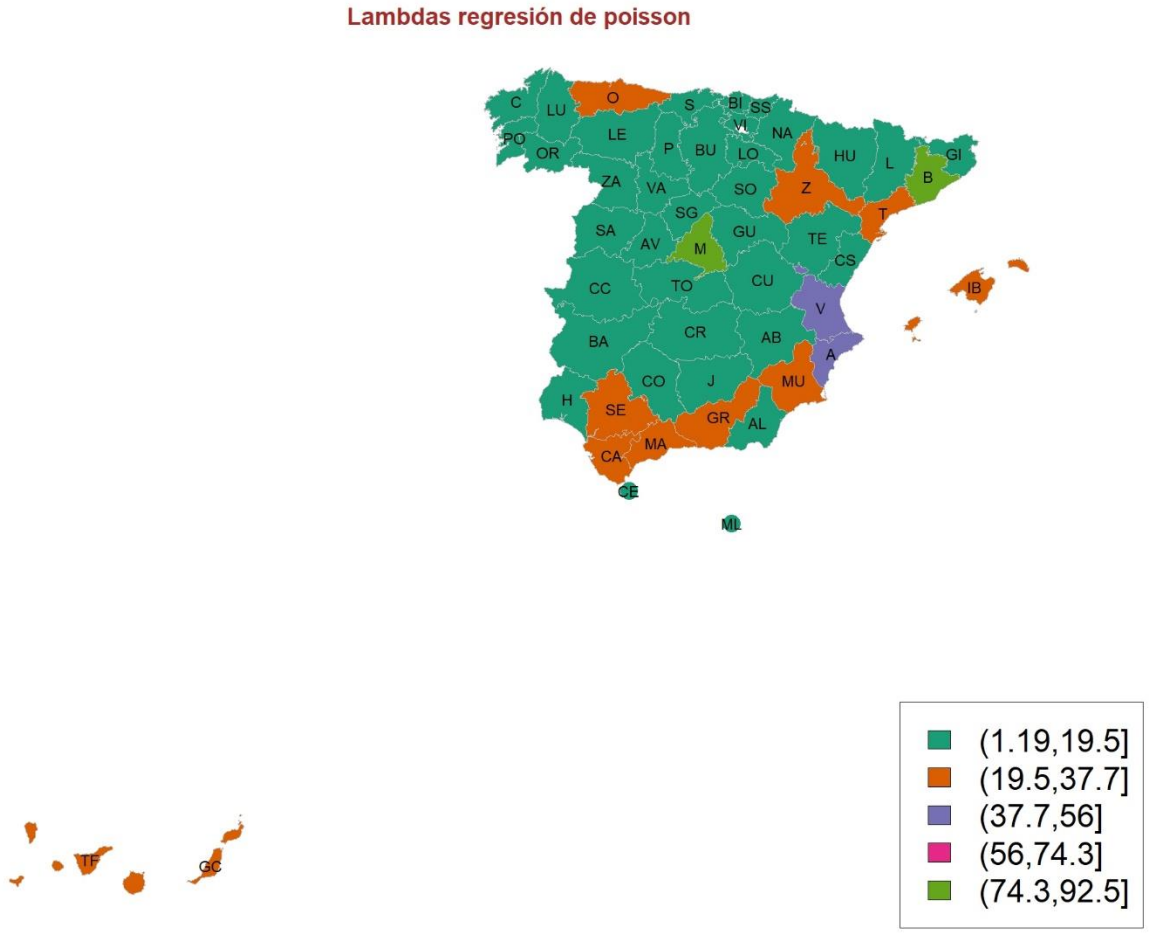
Fuente: elaboración propia

Ilustración 3. Mapa víctimas violencia de género en España por cada 100.000 habitantes (2000-2014)



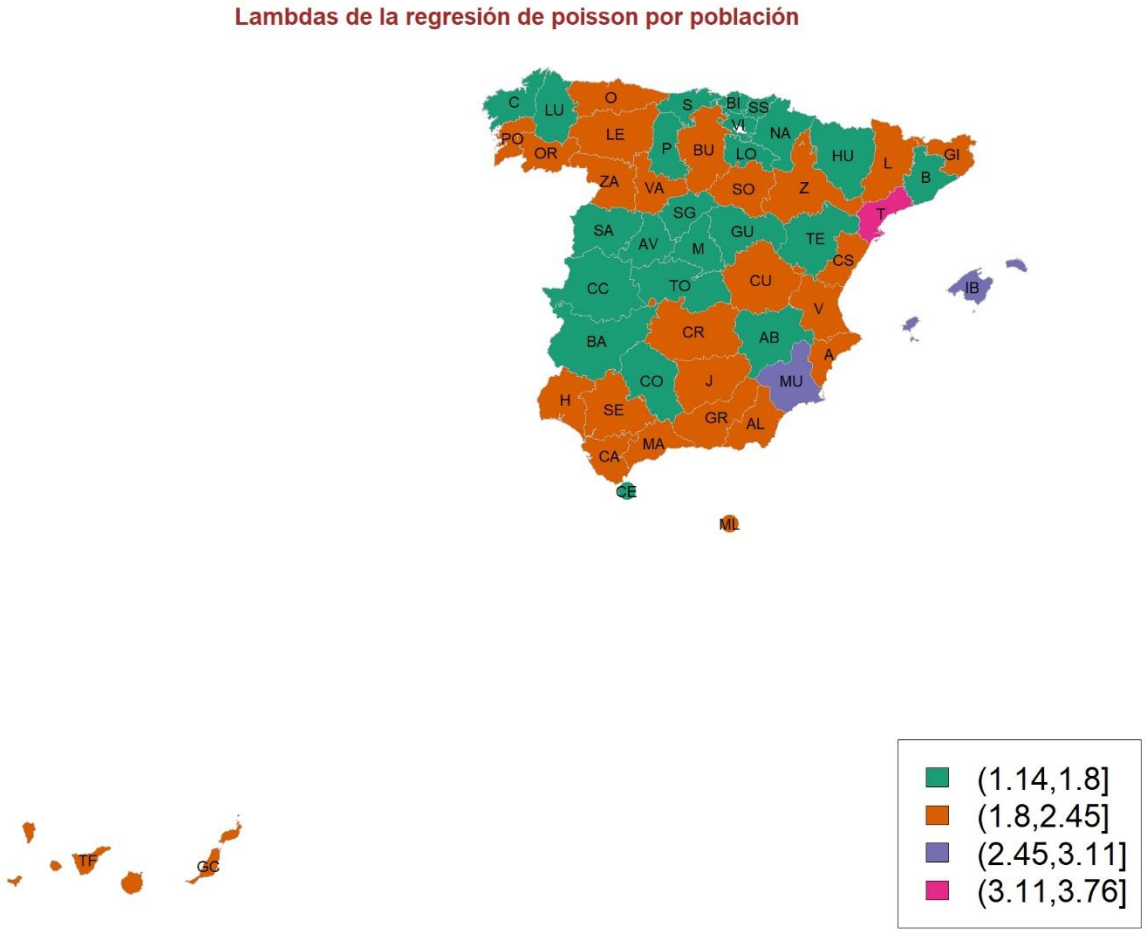
Fuente: elaboración propia

Ilustración 4. Mapa de los valores medianos estimados de λ_i en la regresión de Poisson



Fuente: elaboración propia

Ilustración 5. Mapa de los valores medianos estimados de λ_i por cada 100.000 habitantes en la regresión de Poisson.



Fuente: elaboración propia

6 CONCLUSIONES

Una vez ajustado el modelo, se comentarán en este capítulo las conclusiones extraídas del mismo y se realizará una comparación entre los resultados obtenidos para los valores medianos estimados del modelo y los valores observados de la variable respuesta.

En nuestro modelo, la variable respuesta *victimas* representa el número de víctimas por violencia de género en cada una de las provincias españolas durante los años 2000-2014. Las covariables de nuestro modelo son *pseparaciones*, *pmatrimonios*, *pdenuncias*, *pdivorcios*, *pllamadas*, *pnacionalidad* y *pordenes*.

Una vez ajustado el modelo, las covariables que tienen un efecto significativo sobre la variable respuesta son:

- *pmatrimonios* con un efecto significativo negativo en la variable respuesta.
- *pdenuncias* con un efecto significativo positivo en la variable respuesta.
- *pllamadas* con un efecto significativo negativo en la variable respuesta.

Las covariables que son no significativas en el modelo son:

- *pseparaciones*
- *pdivorcios*
- *pnacionalidad*
- *pordenes*

Si realizamos una comparación entre los resultados obtenidos del ajuste del modelo para valores medianos estimados de los λ_i por cada 100.000, representados en la Ilustración 5 y las víctimas en cada una de las provincias españolas por cada 100.000 habitantes, representados en la Ilustración 3, encontramos algunas diferencias entre las dos ilustraciones.

En primer lugar, hay que tener en cuenta que los resultados en ambos mapas están representados en un número de intervalos diferentes, con valores y colores distintos. Esto puede hacer la comparación a simple vista más complicada.

En segundo lugar, estamos comparando, por un lado, en la Ilustración 3 el valor de la variable respuesta, que hace referencia al número de víctimas por violencia de género, con la mediana de la distribución esperada en la Ilustración 5. En general, los datos de una variable no tienen porqué parecerse ni ser iguales a la mediana.

Las diferencias encontradas en esta comparación podrían indicar una desviación del modelo frente a los datos observados. Si bien, es necesario recordar que, debido a la igualdad de media y varianza, el modelo de Poisson para datos de recuento es restrictivo. Por ello, en el siguiente capítulo VÍAS FUTURAS DE ESTUDIO se tratará, de manera

resumida, este problema y se propondrán otros modelos como solución a este problema del modelo de Poisson.

7 VÍAS FUTURAS DE ESTUDIO

En este capítulo se incluirán otros estudios para la variable respuesta. Para ello, se ha considerado, por un lado, utilizar la distribución binomial negativa para posibles problemas de sobredispersión de los datos y, por otro lado, introducir efectos especiales aleatorios ya que el modelo bayesiano permite introducir fácilmente una estructura de correlación espacial.

Uno de los problemas que encontramos en el modelo de Poisson para datos de recuento es la sobredispersión de los datos. La distribución binomial negativa y el modelo de Poisson generalizado son modelos donde, con la introducción de un nuevo parámetro, se consigue modelizar esta sobredispersión.

El modelo de Poisson es muy popular para los datos de recuento, sin embargo, su igualdad de media y varianza es muy restrictiva para datos que tienen una varianza muy superior a la esperada en el modelo, es decir, datos con sobredispersión. En estos casos es conveniente utilizar un modelo de regresión binomial negativo para datos de recuento.

En lo relativo al efecto espacial, se podría introducir una variable para medir la correlación espacial entre nuestra variable respuesta *victimas* y las provincias que son limítrofes para ver si en las provincias con un mayor número de víctimas se produce un contagio espacial en las provincias colindantes.

8 BIBLIOGRAFÍA

Gschlößl, S. and Czado, C. (2006). “Modelling count data with overdispersion and spatial effects”. *Statistical Papers*, 49(3), pp.531-552.

Miaou, S. (1994). “The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions”. *Accident Analysis & Prevention*, 26(4), pp.471-482.

Ntzoufras, I. (2013). *Bayesian Modeling Using WinBUGS*. Hoboken, N.J.:Wiley.

9 ANEXOS

En este capítulo se han recopilado todos los scripts de RStudio utilizados para la extracción de todos los resultados comentados en la memoria. Todos los apartados que aparecen en este capítulo son scripts, a excepción del ANEXO 5. MODELO que es un fichero de texto necesario para ejecutar el modelo en el ANEXO 6. MODELO POISSON.

9.1 ANEXO 1. PAQUETES Y DATOS

```
setwd ("C:/Users/sarii/Desktop/ESTADISTICA Y EMPRESA/CUARTO/TFG/scripts")
# INSTALACIÓN PAQUETES MAPAS ----------------------------------------
if (!"sp" %in% installed.packages()) install.packages("sp")
library (sp)
library (rgeos)
library (rgdal)
library (pbapply)
library (FNN)
library (maptools)
library (png)
library (raster)
library (geosphere)
library(RColorBrewer)
library ( gpclib )
dep.pkg <- c ( "pbapply", "sp", "FNN", "rgeos", "rgdal", "maptools", "png", "raster")
pkgs.not.installed <- dep.pkg [!sapply ( dep.pkg, function (p) require (p, character.only=T))]
#install.packages (pkgs.not.installed, dependencies=TRUE)
## Windows
if (!require (gpclib)) install.packages ("gpclib", type="source")
library (gpclib)
library (maptools)
rm (list=ls () )
# IMPORTACIÓN DE DATOS ----------------------------------------------
violencia <- read.delim2 ("C:/Users/sarii/Desktop/ESTADISTICA Y EMPRESA/CUARTO/TFG/datos/violencia_genero_tasas.txt")
#names (violencia)
save.image ("paquetesydatos.RData")
```

9.2 ANEXO 2. MAPA ESPAÑA

```
setwd ("C:/Users/sarii/Desktop/ESTADISTICA Y EMPRESA/CUARTO/TFG/scripts" )
load ("paquetesydatos.RData")
library (sp)
library (rgeos)
library (rgdal)
library (pbapply)
library (FNN)
library (maptools)
library (png)
library (raster)
library (geosphere)
library(RColorBrewer)
library ( gpclib )
# DESCARGA E IMPORTACIÓN DEL MAPA DE ESPAÑA------------------------------
ub_shp<-"C:/Users/sarii/Desktop/ESTADISTICAY
EMPRESA/CUARTO/TFG/datos/ctx01_limites.shp"
provincias <- readShapeSpatial (ub_shp)
#plot (provincias)
#names (provincias)
# FILTRADO DE MAPA ESPAÑA (PAÍS = ESPAÑA Y DEMOGRAFÍA =
PROVINCIAS) ------------------------
#summary (provincias$PAIS)
# Para quedarme con país= España
#datosespana <- provincias [provincias$PAIS == "EspaÃ±a",]
#plot (datosespana)
#summary (datosespana)
# Tenemos varias clasificaciones en demografía (provincia, comunidad autónoma, ciudad
autónoma...) como nos interesa sólo quedarnos con provincias vamos a realizar otro filtro
#Para quedarme con tipo de demografía = provincia
#datosespanaprovincia <- provincias [provincias$TIPO_DEM == "Provincia",]
#plot (datosespanaprovincia)
# De forma resumida, para no tener tantos filtros, he decidido meterlo todo lo explicado
anteriormente en una sola línea
```

```
Espana <- provincias [provincias$PAIS == "EspaÃ±a"& provincias$TIPO_DEM == "Provincia",]
#summary (Espana)
#names (Espana)
Espanadata <- as.data.frame (Espana)
# MAPA ESPAÑA----------------------------------------------
# Añadimos nombres de provincias para representarlas en el mapa
#Espanadata$NOMBRE
nombresprovincias <- c (
  "NA","HU","GI","BU","BI","S","O","LU","C","PO","OR","ZA","L","Z","P","LE",
  "VA","SO","LO","VI","SA","B","T","TE","GU","SG","AV","IB","CS","V","A",
  "MU","GR","J","CO","SE","H","BA","CC","TO","CR","CU","AB","MA","AL",
  "TF","GC","CE","ML","CA","SS","M"
)
# Asignamos el nombre de las provincias a cada una de las filas del polígono espacial de Espana.
row.names (Espana) <- nombresprovincias
# Sobreescribimos el data.frame Espanadata para añadir los nombres de provincias por filas
Espanadata <- as.data.frame (Espana)
# He comprobado que "Espanadata" tiene 52 variables correspondientes a las 52 provincias, para que posteriormente no haya problemas al unir hojas de datos.
# Calculamos los centroides
centroides <- coordinates (Espana)
ceuta_mel <- centroides [48:49,  ]
# Representamos el mapa de España
plot (Espana)
# Añadimos nombres de provincias (vector definido anteriormente) al mapa
text (centroides, nombresprovincias, cex=0.6, col= "black"  )
save.image("mapaEspana.RData")
```

9.3 ANEXO 3. MAPA VÍCTIMAS

```
setwd ("C:/Users/sarii/Desktop/ESTADISTICA Y EMPRESA/CUARTO/TFG/scripts" )
load ("mapaEspana.RData")
```

```
# MAPA VÍCTIMAS----------------------------------------------
##################PREPARACIÓN MAPA VÍCTIMAS
library (RColorBrewer)
library (sp)
# Para representar nuestros datos, tenemos que unir la hoja de datos del mapa de España y de violencia de género a través del código del municipio del INE. En este caso, las variables que hacen referencia al código de la provincia tanto del mapa de España como de nuestra hoja de datos "violencia" tienen el mismo orden.
# Unimos la variable "cod_provincia" de nuestra hoja de datos "violencia" a través de la variable "COD_INE" de nuestro mapa de España
violencia$cod_provincia <- Espanadata$COD_INE
# De este modo quedan unidos los dos data.frame por el código de provincia.
# Ahora tomamos los datos que queremos representar en el mapa:
# Extraemos los datos:
victimas <- as.matrix (violencia$victimas)
# Insertamos como nombre de cada fila el nombre abreviado de la provincia:
row.names (victimas) <- nombresprovincias
# Convertimos el vector anterior en un en data.frame
victimas <- as.data.frame (victimas)
names (victimas) <- "victimas"
row.names (victimas) <- row.names (Espana)
#prov <- as.vector (Espana$COD_INE)
Espana.data <- SpatialPolygonsDataFrame (Espana, victimas)
plotvar <- Espana.data$victimas
# Determinamos un nº de cortes para clasificar el nº de víctimas por intervalos y poder asignar a cada uno de los intervalos un código de color.
ncortes <- 5
cortes <- cut (plotvar, ncortes)
levels(cortes)
summary (cortes)
niveles<- c("(2,19]","(19, 37]","(37, 55]","(55, 73]","(73, 91]" )
# Realizamos un bucle para asociar cada uno de los intervalos del número de víctimas con un código de color
color <- numeric (length (plotvar))
```

```
for (i in 1: length (plotvar)){
  if (cortes[i] == levels(cortes)[1]) color[i] <- 1
  if (cortes[i] == levels(cortes)[2]) color[i] <- 2
  if (cortes[i] == levels(cortes)[3]) color[i] <- 3
  if (cortes[i] == levels(cortes)[4]) color[i] <- 4
  if (cortes[i] == levels(cortes)[5]) color[i] <- 5
}
# Elegimos la paleta de colores
# Como tenemos que asociar 5 intervalos a 5 códigos de colores de una paleta, comprobamos que la paleta de colores tiene su máximo de colores superior al número de intervalos
brewer.pal.info ["Dark2",]
# Efectivamente, tiene 9 colores como máximo, por lo que podemos utilizar esta paleta
plotclr <- brewer.pal (ncortes, "Dark2")
#cbind(color, plotclr[color])
################## MAPA VÍCTIMAS JPEG
jpeg ("MAPAVICTIMAS.jpeg", quality=100, height=1500, width=2000)
plot (Espana.data, col = plotclr[color], border = "Gray")
title (main= "Víctimas violencia de género en España (2000-2014)", cex.main=3, col.main="brown")
legend ("bottomright", legend = niveles, fill= plotclr, cex=4 )
points (ceuta_mel, col = plotclr[color[48:49]], cex = 5, pch = 19)
text (centroides, nombresprovincias, cex=2, col= "black"  )
dev.off ()
save.image("mapavictimas.RData")
```

9.4 ANEXO 4. MAPA VÍCTIMAS POR CADA 100.000 HABITANTES

```
setwd ("C:/Users/sarii/Desktop/ESTADISTICA Y EMPRESA/CUARTO/TFG/scripts" )
load("mapavictimas.RData")
# MAPA VÍCTIMAS/POBLACIÓN ----------------------------------------------
##################PREPARACIÓN MAPA VÍCTIMAS/POBLACIÓN
library (RColorBrewer)
violencia$cod_provincia <- Espanadata$COD_INE
# Extraemos los datos:
violenciatasa <- as.matrix (violencia[, 3:11])
```

```
# Insertamos como nombre de cada fila el nombre abreviado de la provincia:
row.names (violenciatasa) <- nombresprovincias
# He decidido mostar el número de víctimas como la tasa de víctimas que hace referencia al número de víctimas en cada provincia por cada 100.000 habitantes.
victimast <- (violenciatasa [, "victimas"] / violenciatasa [, "poblacion"]) *100000
# Convertimos el vector anterior en un en data.frame
victimast <- as.data.frame (victimast)
row.names (victimast) <- row.names (Espana)
names (victimast) <- "victimast"
Espana.data2 <- SpatialPolygonsDataFrame (Espana, victimast)
plotvart <- Espana.data2$victimast
# Determinamos un nº de cortes para clasificar la tasa de víctimas por intervalos y poder asignar a cada uno de los intervalos un código de color.
ncortest <- 5
cortest <- cut (plotvart, ncortest)
summary (cortest)
# Realizamos un bucle para asociar cada uno de los intervalos del número de víctimas con un código de color
colort <- numeric (length (plotvart))
for (i in 1:length (plotvart)){
  if (cortest[i] == levels(cortest)[1]) colort[i] <- 1
  if (cortest[i] == levels(cortest)[2]) colort[i] <- 2
  if (cortest[i] == levels(cortest)[3]) colort[i] <- 3
  if (cortest[i] == levels(cortest)[4]) colort[i] <- 4
  if (cortest[i] == levels(cortest)[5]) colort[i] <- 5
}
# Elegimos la paleta de colores
# Como tenemos que asociar 5 intervalos a 5 códigos de colores de una paleta, comprobamos que la paleta de colores tiene su máximo de colores superior al número de intervalos
#brewer.pal.info["Dark2",]
# Efectivamente, tiene 9 colores como máximo, por lo que podemos utilizar esta paleta
plotclrt <- brewer.pal (ncortest, "Dark2")
```

```
################## MAPA VÍCTIMAS/POBLACIÓN JPEG
jpeg ("MAPAVICTIMAS_POBLACION.jpeg", quality=100, height=1500, width=2000)
plot (Espana.data2, col = plotclrt [colort], border = "Grey")
title (main= "Víctimas violencia de género en España por cada 100.000 habitantes (2000-2014)", cex.main=3, col.main="brown")
legend ("bottomright", legend = levels(cortest), fill= plotclrt, cex=4 )
points (ceuta_mel, col = plotclrt[colort[48:49]], cex = 5, pch = 19)
text (centroides, nombresprovincias, cex=2, col= "black"  )
dev.off ()
save.image("mapavictimaspoblacion.RData")
```

9.5 ANEXO 5. MODELO

```
model{
        for(i in 1:I){
    y[i] ~ dpois(lambda[i])
    log(lambda[i]) <- log(poblacion[i])
    + beta0
    + beta1 * pseparaciones[i]
    + beta2 * pmatrimonios[i]
    + beta3 * pdenuncias[i]
    + beta4 * pdivorcios[i]
    + beta5 * pllamadas[i]
    + beta6 * pnacionalidad[i]
    + beta7 * pordenes[i]
    }
beta0 ~ dnorm(0.0,1.0E-6)I(-100,100)
beta1 ~ dnorm(0.0,1.0E-6)I(-100,100)
beta2 ~ dnorm(0.0,1.0E-6)I(-100,100)
beta3 ~ dnorm(0.0,1.0E-6)I(-100,100)
beta4 ~ dnorm(0.0,1.0E-6)I(-100,100)
beta5 ~ dnorm(0.0,1.0E-6)I(-100,100)
beta6 ~ dnorm(0.0,1.0E-6)I(-100,100)
beta7 ~ dnorm(0.0,1.0E-6)I(-100,100)
}
```

9.6 ANEXO 6. MODELO POISSON

```
setwd ("C:/Users/sarii/Desktop/ESTADISTICA Y EMPRESA/CUARTO/TFG/scripts" )
load ("paquetesydatos.RData")
# MODELO POISSON ----------------------------------------------------------
library (coda)
library (lattice)
library (R2WinBUGS)
names (violencia)
I <- nrow (violencia)
y <- violencia$victimas
poblacion <- violencia$poblacion
pseparaciones <- violencia$pseparaciones
pmatrimonios <- violencia$pmatrimonios
pdenuncias <- violencia$pdenuncias
pdivorcios <- violencia$pdivorcios
pllamadas <- violencia$pllamadas
pnacionalidad <- violencia$pnacionalidad
pordenes <- violencia$pordenes
data <- list ("I", "y", "poblacion",
              "pseparaciones",
              "pmatrimonios",
              "pdenuncias",
              "pdivorcios",
              "pllamadas",
              "pnacionalidad",
              "pordenes")
inits <- function (){
  t <- 1/0.1#
  list(beta0 = rnorm (1, 0.0, t),
       beta1 = rnorm (1, 0.0, t),
       beta2 = rnorm (1, 0.0, t),
       beta3 = rnorm (1, 0.0, t),
       beta4 = rnorm (1, 0.0, t),
       beta5 = rnorm (1, 0.0, t),
```

```
      beta6 = rnorm (1, 0.0, t),
      beta7 = rnorm (1, 0.0, t))
}
inits ()
parametros <- c ("beta0", "beta1", "beta2", "beta3", "beta4", "beta5", "beta6", "beta7",
"lambda")
modelo <- bugs (data,
            inits,
            model.file = "C:/Users/sarii/Desktop/ESTADISTICA Y
EMPRESA/CUARTO/TFG/scripts/modelo.txt",
            parameters.to.save = parametros,
            n.thin = 1,
            n.chains = 3,
            n.iter = 10000,
            n.burnin = 1000,
            #debug = TRUE,
            bugs.directory=
"C:/Users/sarii/Downloads/winbugs14_unrestricted/WinBUGS14")
# results
print (modelo)
names(modelo)
print(modelo$median)
#plot(modelo)
save.image ("modeloPoisson.RData")
```

9.7 ANEXO 7. MAPA LAMBDAS

```
setwd ("C:/Users/sarii/Desktop/ESTADISTICA Y EMPRESA/CUARTO/TFG/scripts" )
load ("modeloPoisson.RData")
load ("mapaEspana.RData")
# MAPA LAMBDAS ------------------------------------------------------------
library (RColorBrewer)
lambdas <- as.matrix (modelo$median$lambda)
row.names (lambdas) <- nombresprovincias
lambdas<- as.data.frame (lambdas)
row.names (lambdas) <- row.names (Espana)
```

```
names (lambdas) <- "lambdas"
Espana.data3 <- SpatialPolygonsDataFrame (Espana, lambdas)
plotvarlam <- Espana.data3$lambdas
# Determinamos un nº de cortes para poder asignar a cada uno de los intervalos un código de color.
ncorteslam <- 5
corteslam <- cut(plotvarlam, ncorteslam)
summary (corteslam)
levels (corteslam)
# Realizamos un bucle para asociar a cada uno de los intervalos un código de color
colorlam <- numeric (length (plotvarlam))
for (i in 1:length (plotvarlam)){
  if (corteslam[i] == levels(corteslam)[1]) colorlam[i] <- 1
  if (corteslam[i] == levels(corteslam)[2]) colorlam[i] <- 2
  if (corteslam[i] == levels(corteslam)[3]) colorlam[i] <- 3
  if (corteslam[i] == levels(corteslam)[4]) colorlam[i] <- 4
  if (corteslam[i] == levels(corteslam)[5]) colorlam[i] <- 5
}
# Elegimos la paleta de colores
# Como tenemos que asociar 5 intervalos a 5 códigos de colores de una paleta, comprobamos que la paleta de colores tiene su máximo de colores superior al número de intervalos
#brewer.pal.info["Dark2"]
# Efectivamente, tiene 9 colores como máximo, por lo que podemos utilizar esta paleta
plotclrlam <- brewer.pal (ncorteslam, "Dark2")
#cbind (colorlam, plotclrlam[colorlam])
# MAPA LAMBDAS JPEG ----------------------------------------------------
jpeg ("MAPALAMBDAS.jpeg", quality=100, height=1500, width=2000)
plot (Espana.data3, col = plotclrlam[colorlam], border = "Gray")
title (main= "Lambdas regresión de poisson", cex.main=3, col.main="brown")
legend ("bottomright", legend = levels(corteslam), fill=plotclrlam, cex=4 )
points (ceuta_mel, col = plotclrlam[colorlam[48:49]], cex = 5, pch = 19)
text (centroides, nombresprovincias, cex=2, col= "black"  )
dev.off ()
```

```
save.image("mapalambdas.RData")
```

9.8 ANEXO 8. MAPA LAMBDAS POR CADA 100.000 HABITANTES

```
setwd ("C:/Users/sarii/Desktop/ESTADISTICA Y EMPRESA/CUARTO/TFG/scripts" )
load ("modeloPoisson.RData")
load ("mapaEspana.RData")
# MAPA LAMBDAS/POBLACIÓN ------------------------------------------------------------
library (RColorBrewer)
lambdas <- as.matrix (modelo$median$lambda)
lambdas<- as.data.frame (lambdas)
row.names (lambdas) <- row.names (Espana)
names (lambdas) <- "lambdas"
lambdast <- lambdas$lambdas / violencia$poblacion
lambdast <- lambdast*100000
lambdast <- as.data.frame (lambdast)
row.names (lambdast) <- nombresprovincias
row.names (lambdast) <- row.names (Espana)
names (lambdast) <- "lambdast"
Espana.data4 <- SpatialPolygonsDataFrame (Espana, lambdast)
plotvarlamt <- Espana.data4$lambdast
# Determinamos un nº de cortes para realizar la clasificación por intervalos y poder asignar a cada uno de los intervalos un código de color.
ncorteslamt <- 4
corteslamt <- cut(plotvarlamt, ncorteslamt)
summary (corteslamt)
levels (corteslamt)
# Realizamos un bucle para asociar a cada uno de los intervalos un código de color
colorlamt <- numeric (length (plotvarlamt))
for (i in 1:length (plotvarlamt)){
  if (corteslamt[i] == levels(corteslamt)[1]) colorlamt[i] <- 1
  if (corteslamt[i] == levels(corteslamt)[2]) colorlamt[i] <- 2
  if (corteslamt[i] == levels(corteslamt)[3]) colorlamt[i] <- 3
  if (corteslamt[i] == levels(corteslamt)[4]) colorlamt[i] <- 4
}
# Elegimos la paleta de colores
```

```
# Como tenemos que asociar 4 intervalos a 4 códigos de colores de una paleta, comprobamos que la paleta de colores tiene su máximo de colores superior al número de intervalos
#brewer.pal.info["Dark2"]
# Efectivamente, tiene 9 colores como máximo, por lo que podemos utilizar esta paleta
plotclrlamt <- brewer.pal (ncorteslamt, "Dark2")
#cbind (colorlam, plotclrlam[colorlam])
# MAPA LAMBDAS/POBLACIÓN JPEG ----------
jpeg ("MAPALAMBDAS_POBLACION.jpeg", quality=100, height=1500, width=2000)
plot (Espana.data4, col = plotclrlamt[colorlamt], border = "Gray")
title (main= "Lambdas de la regresión de poisson por población", cex.main=3, col.main="brown")
legend ("bottomright", legend = levels(corteslamt), fill=plotclrlamt, cex=4 )
points (ceuta_mel, col = plotclrlamt[colorlamt[48:49]], cex = 5, pch = 19)
text (centroides, nombresprovincias, cex=2, col= "black"  )
dev.off ()
save.image("mapalambdaspoblacion.RData")
```