UNIVERSIDAD SIMÓN BOLÍVAR
DECANATO DE ESTUDIOS PROFESIONALES
COORDINACIÓN DE MATEMÁTICAS

ADAPTACIÓN DE MODELOS DE REGRESIÓN LASSO BAYESIANO A LOS DATOS DE BROTES DE ETA EN CHILE 2017

Por
Oropeza Oropeza, Keily Marian

PROYECTO DE GRADO
Presentado ante la ilustre Universidad Simón Bolívar
como requisito parcial para optar al título de
Licenciado en Matemáticas, opción Estadística y Matemáticas Computacionales

Sartenejas, Septiembre de 2020

K. OROPEZA
2020

ADAPTACIÓN DE MODELOS DE REGRESIÓN LASSO BAYESIANO A LOS DATOS DE BROTES DE ETA EN CHILE
2017

USB
LICENCIATURA EN MATEMÁTICAS

UNIVERSIDAD SIMÓN BOLÍVAR
DECANATO DE ESTUDIOS PROFESIONALES
COORDINACIÓN DE MATEMÁTICAS

ADAPTACIÓN DE MODELOS DE REGRESIÓN LASSO BAYESIANO A LOS DATOS DE BROTES DE ETA EN CHILE 2017

Por
Oropeza Oropeza, Keily Marian

Realizado con la asesoría de:

Desireé Villalta

PROYECTO DE GRADO
Presentado ante la ilustre Universidad Simón Bolívar
como requisito parcial para optar al título de
Licenciado en Matemáticas, opción Estadística y Matemáticas Computacionales

Sartenejas, Septiembre de 2020

UNIVERSIDAD SIMÓN BOLÍVAR
VICERRECTORADO ACADÉMICO
DECANATO DE ESTUDIOS PROFESIONALES
Coordinación de Matemáticas

ACTA DE EVALUACION DE PROYECTO DE GRADO

Código de la asignatura: EP1509
Modalidad: **Fecha**: 08-10-2020
A distancia (Asincrónica)
Nombre del estudiante: Keily Marian Oropeza Oropeza **Carnet**: 13-11009
Título del proyecto: ADAPTACIÓN DE MODELOS DE REGRESION LASSO BAYESIANO A LOS DATOS DE BROTES DE ETA EN CHILE, 2017.
Tutor: Prof. Desiree Villalta
Jurados: Profesores Zoraida Martínez e Isabel Llatas

X	APROBADO		REPROBADO		INCOMPLETO

Observaciones:

El jurado examinador, **por unanimidad**, considera el proyecto de grado merecedor de la mención especial SOBRESALIENTE:

	SI	X	NO

En caso afirmativo, justifique su decisión en estricto cumplimiento del documento "Criterios para otorgar la mención especial en proyectos de grado y pasantías largas e intermedias"[1], aprobado por el Consejo Plenario 07-2015 del Decanato de Estudios Profesionales de fecha 15 de junio de 2015:

Presidente del Jurado
Prof. Isabel Llatas S.
(CI V-4.767053)

Jurado
Prof. Zoraida Martínez
(CI V-11.165.566)

Tutor
Prof. Desiree Villalta
(CI V-14.964.151)

Nota: Colocar los sellos de los respectivos Departamentos Académicos. Para jurados externos usar el sello de la Coordinación Docente. Este documento debe entregarse en original y sin enmiendas.

*Copia de la comunicación emitida por la **Comisión Permanente de Pedagogía Digital** donde se valida la presentación a distancia bajo **la modalidad Sincrónica**

** Copia de la comunicación emitida por la **Comisión Permanente de Pedagogía Digital** donde se valida la presentación a distancia bajo la **modalidad Asincrónica**

[1] Incluir dirección de la página web del Decanato una vez aprobado

DEDICATORIA

Dedicado a mi familia, profesores y amigos, ya que todos y cada uno de ellos pusieron su granito de arena para hacerme llegar a donde estoy.

En especial, este libro esta dedicado a mis abuelos Lucas y Francisco, mi tío Francisco Javier, mi madre Mariliana y por aquellas personas que ya no están aquí pero que desde muy pequeña me impulsaron a lograr mis sueños y mejorar cada día.

AGRADECIMIENTOS

Iniciar una carrera universitaria fue uno de los sueños que desde pequeña fue impulsado por mi madre, aunque al tomar la decisión de estudiar Matemáticas no fue bien recibida al inicio, mi familia me enseñó a ser firme con mi decisión si era lo que tanto deseaba. Es aquí donde debo agradecer a la Lic. Valentina Sojo quien nos mostró a mi familia y a mi el amplio mundo de posibilidades que tiene un Matemático, lo que ayudo a limar esas asperezas que se tenían con mi decisión. También debo agradecer a Ambrosio Núñez quien nos acompañó a mi madre y a mi desde mucho antes de empezar esta carrera por lo que estuvo lidiando con ambas, ayudando a suavizar esas asperezas y que se reía de las loqueras que iba aprendiendo.

Llegar a tomar a la USB como mi alma mater fue un camino lleno de perseverancia, altibajos, separaciones y sobretodo amor a primera vista por esta casa de estudio. Aunque al comienzo la USB me puso a prueba desde muchos ámbitos, llevándome a considerar que no lograría terminar mi carrera aquí, cada una de estas pruebas fue acompañada por el apoyo de nuevos amigos y de mi familia que me ayudaron a superarlas, pero sobretodo aprender de cada prueba una lección de vida que me llevaron a ser la mujer que hoy soy. Por esto, debo agradecer a la USB todo lo que me ha enseñado desde lo académico hasta la vida.

Con cada prueba superada la Simón se fue convirtiendo en mi segundo hogar y donde conocí mi segunda familia que me acompaña desde los inicios de mi vida universitaria y a la que cada vez se fueron agregando más personas. En un principio esta familia estuvo conformada por Victor, Jorhelis, Christian y Karla, quienes a pesar de ser tan diferentes me dieron su apoyo y estuvieron en las buenas y malas durante este inicio de nuestras carreras universitarias, con quienes disfrute de varios encuentros inolvidables. Luego, con el inicio de las materias de carrera, se unieron a esta familia mis hermanos Matemáticos, entre ellos Gianni, Genesis, Daniel y Andy, quienes me recibieron con una sonrisa, me presentaron este grupo tan cerrado e intimidatorio al inicio y que me han acompañado desde que empecé a estudiar las artes oscuras de las Matemáticas como diría Daniel. Con el pasar de los trimestres a esta familia se le unieron cada vez más personas, Pacheco, Zuil, Yemasu, Piera, Rafa, Jose C., Daniel V., Lin, quienes hicieron que mi vida como estudiante de matemáticas fuera más que solo clases; también Jonathan, Irina, Yerimar y Kevin, quienes con sus ocurrencias alegraron hasta los más estresantes momentos. A cada uno de ellos y aquellos que aunque su nombre no aparece, les agradezco que hicieran que mi vida en la Simón tuviera risas, alegrías y momentos inolvidables.

Por último y no por ello menos importante, debo agradecer a los profesionales que dedicaron largas horas en enseñarme todos sus conocimientos y amor por lo que hacen, entre ellos están la Prof. Maria Teresa Varela, la Prof. Sandra Leal y la Prof. Aurora Olivieri, quienes su amor por enseñar me hicieron enamorarme más de mi carrera; el Prof. Jhonnathan Arteaga, la Prof. Minaya Villasana, el Prof. Jesús Nieto, la Prof. Desireé Villalta, el Prof. Alejandro Bravo, la Prof. Isabel Llatas y el Prof. Pedro Ovalles, quienes me enseñaron las diferentes ramas de las Matemáticas mostrándome lo amplio y emocionante que este mundo. A todos ustedes, muchas gracias.

RESUMEN

Recientemente organismos internacionales como la OMS, FAO, FDA y OPS han insistido en la importancia de vigilar las Enfermedades Transmitidas por Alimentos (ETA) y estudiar de forma interdisciplinaria las mismas con la finalidad de controlar los efectos de estas, que afectan principalmente a niños, personas mayores, embarazadas y personas inmuno-comprometidas. Debido a esto surge el presente trabajo, con la finalidad de encontrar modelos matemáticos que describan los brotes de ETA en Chile durante el año 2017, que permitan predecir la cantidad de personas enfermas en un brote y encontrar las variables de mayor peso e importancia en las cuales enfocarse. Para lograr esto se realizaron estudios tomando dos enfoque, la estadística clásica y análisis Bayesiano, bajo tres casos: El primero incluye las variables de atención tomando los datos de los brotes de E. Coli, Salmonella, Shigella y Campylobacter, el segundo caso estudia estas sin considerar las variables de atención y finalmente, el tercero se enfoca en Salmonella sin considerar las variables de atención. De los estudios realizados se concluye el mejor modelo para el primer caso es el modelo LASSO clásico, para el segundo y tercero el modelo Bayesiano con priori flat; además, el grupo etario de 15 a 44 años es una variable a considerar en los brotes, también, la cantidad de expuestos y la región de notificación son primordiales para los modelos en los casos estudiados. Por otro lado, las variables sintomáticas relevantes son los dolores, meteorismo, rush cutáneo, parestesias y nauseas, que resaltan entre los modelos.

Palabras claves: Enfermedades transmitidas por alimentos, regresión lineal, métodos, modelos lineales generalizados, LASSO, Ridge.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

LISTA DE SÍMBOLOS

$\Re^n$	Espacio vectorial de las n-duplas con coeficientes en los reales.
$E[y]$	Esperanza de la variable aleatoria y.
$\mathcal{N}(\mu, \sigma^2)$	Distribución normal con media μ y desviación estándar σ^2.
$\sim$	Distribuye.
$\mathcal{L}_1$, $\|.\|_1$	Norma 1.
$\mathcal{L}_2$, $\|.\|_2$	Norma 2.
$\bar{y}$	Promedio de los valores observados y.
$\propto$	Proporcionalidad.
$\mathcal{Y}$	Subconjunto de $\Re^n$ el cual no depende de $\boldsymbol{\theta}$.
$\boldsymbol{\theta}$	Parámetro conocido.
$\boldsymbol{\beta}$	Parámetro desconocido de coeficientes.
$P(\boldsymbol{y}\|\boldsymbol{\theta})$	Probabilidad condicional de $\boldsymbol{y}$ dado $\boldsymbol{\theta}$.
$P(\boldsymbol{y}, \boldsymbol{\theta})$	Probabilidad conjunta de $\boldsymbol{y}$ y $\boldsymbol{\theta}$.
$\mathcal{M}_{(m)\times(n)}(\Re)$	Subespacio vectorial de las matrices $(m) \times (n)$ con coeficientes en los reales.
$\boldsymbol{\epsilon}$	Vector de errores.

LISTA DE ACRÓNIMOS

CDC Centros para el Control y Prevención de Enfermedades, en Estados Unidos

CV Validación Cruzada

DEIS Departamento de Estadística e Información de Salud, en Chile

EAM Error Absoluto Medio

ECM Error Cuadrático Medio

ETA Enfermedades Transmitidas por Alimentos

FAO Organización de las Naciones Unidas para la Alimentación y la Agricultura

FDA Administración de Alimentos y Medicamentos, en Estados Unidos

IID independientes e idénticamente distribuidos

LASSO Operador de Selección y Contracción Mínimo Absoluta

MCMC métodos de cadenas de Markov Montecarlo

MINSAL Ministerio de Salud, en Chile

MLG Modelos Lineales Generalizados

OLS Mínimos Cuadrados Ordinarios

OMS Organización Mundial de la Salud

OPS Organización Panamericana de la Salud

SAIA Seguridad Alimentaria y Seguridad del Agua, por sus siglas en catalán

UA Unión Africana

INTRODUCCIÓN

Aunque se puede pensar que los controles de calidad en los alimentos son de reciente aplicación, su origen puede ser determinado hace miles de años. El primero de ellos del que se tiene conocimiento data del 2500 A.C, con las leyes de Moisés y las leyes egipcias los cuales contemplaban la prevención de contaminación en la carne. Esta preocupación también se puede observar en las culturas china, griega y romana, en las cuales existen escritos que mencionan pesos y medidas reglamentadas para alimentos y otros productos. En los siglos XVII y XVIII, se comenzó a usar la química como herramienta de análisis para controlar la calidad de los alimentos, desde entonces sus fundamentos son utilizados hasta la actualidad, solo siendo alterados en aspectos como las técnicas para detectarlo.

Se refiere con el término "calidad de los alimentos"a la capacidad de éstos en satisfacer las necesidades declaradas del consumidor, es también el atributo higiénico–sanitario que por su relación con la salud de las personas resulta ser de valor básico y absoluto, dado que presupone que un alimento no debe causar daño a quien lo consume (Prieto y cols., 2008).

Mientras que al hablar del "control de la calidad de los alimentos"se hace referencia a la utilización de parámetros tecnológicos, físicos, químicos, microbiológicos, nutricionales y sensoriales para lograr que un alimento sea sano y sabroso con el objetivo de proteger al consumidor, tanto del fraude como de su salud (SAIA, 2017).

Este atributo sanitario de calidad ha sido nombrado como "Inocuidad de Alimentos", el cual se define como "la ausencia, o niveles seguros y aceptables, de peligro en los alimentos que pueden dañar la salud de los consumidores"(FAO, 2003). Para el logro de esta meta, se le otorga responsabilidad a toda la cadena alimentaria para mantener esta propiedad hasta el momento de su consumo y por tanto, todas las políticas y actividades deben orientarse hacia este objetivo (OMS, 2019).

Por el contrario, si se tiene la pérdida de inocuidad de los alimentos esto afecta en la salud de quienes los consumen y son una importante causa de morbilidad y mortalidad de millones de personas en el mundo, quienes enferman individual o grupalmente al consumir alimentos contaminados (OMS, 2015b), siendo los grupos más vulnerables las mujeres

embarazadas, niños, personas mayores y personas con alguna alteración en el sistema inmune (FDA, 2018).

Según las estimaciones de la Organización Mundial de la Salud (OMS) en el mundo 600 millones de personas enferman por ingerir alimentos contaminados y 420.000 por la misma causa (OMS, 2015). De los cuales, los niños menores de 5 años forman el 40 % de la carga atribuible a las Enfermedades Transmitidas por Alimentos (ETA) produciendo cada año 125.000 muertes en este grupo. En cuanto a la mortalidad infantil se indican que solo debido a diarreas, mueren 1,9 millones de niños cada año en todo el mundo, ocurriendo más frecuentemente en países menos desarrollados, de las cuales una considerable proporción ocurre a causa de enfermedades transmitidas por alimentos, (OMS, 2008). Por otro lado, según los Centros para el Control y Prevención de Enfermedades, en Estados Unidos (CDC) se pueden describir aproximadamente unas 250 enfermedades transmitidas por los alimentos (CDC, 2018).

Se denomina a las ETA, como cualquier enfermedad de naturaleza infecciosa o tóxica causada por el consumo de alimentos (OMS, 2019, 2008a). En cuanto a los brotes de ETA, se establece que son episodios en los cuales dos o más personas presentan una enfermedad similar después de ingerir alimentos, incluida el agua, del mismo origen y donde la evidencia epidemiológica o el análisis de laboratorio implica a los alimentos o al agua como vehículo de la misma (OPS, 2015).

En el mundo las ETA han ido aumentando a causa de varios factores de tipo ambiental, cultural y socio-económico. Si bien la ocurrencia de ETA ha sido medida por países desarrollados que cuentan con sistemas de vigilancia implementados, los valores reales de esta situación son en general desconocidos y se consideran subestimados. Por esto, la necesidad de detectar, investigar y controlar a tiempo los brotes de ETA, presiona a los países a desarrollar e implementar algún sistema que permita su vigilancia.

Según la Organización Panamericana de la Salud (OPS), la vigilancia epidemiológica de las ETA, es el conjunto de actividades que permiten reunir la información necesaria para conocer la conducta o historia natural de las enfermedades y detectar o prever cambios que ocurran debido a alteraciones en los factores condicionantes o determinantes, con la finalidad de recomendar de manera oportuna y con bases firmes, las medidas indicadas para su prevención y control.

Como se puede observar el problema de salud pública mundial causada por las ETA, es un tema de suma importancia para la especie humana desde hace miles de años, pero el tema de la vigilancia de la ETA es algo reciente que se viene presentando en el mundo y que algunos países como Estados Unidos, Chile y Cuba están implementando.

En la Primera Conferencia Internacional conjunta entre la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO), la OMS y la Unión Africana (UA) sobre Inocuidad Alimentaria que se celebró en febrero de 2019 (FAO, OMS, y UA, 2019), se discutió sobre los sistemas de vigilancia epidemiológica de las ETA y sistemas de alerta en materia de inocuidad de los alimentos, donde se reconoció la necesidad de fortalecer estos sistemas con el fin de detectar e investigar los brotes, fortalecer la colaboración intersectorial y aplicar enfoques multisectoriales, entre otras acciones a tomar, con la finalidad de unificar habilidades para el desarrollo de estrategias de control efectivas para las amenazas a la inocuidad de los alimentos.

A diferencia de la inspección que tienen como fin evidenciar incumplimientos regulatorios, la investigación sanitaria a los alimentos en el contexto de un brote de ETA, es guiada por datos e información generada en la investigación epidemiológica y entrevista de los casos, e intenta esclarecer las condiciones en que los alimentos sospechosos fueron preparados y consumidos (Ulloa Bello, 2016).

Según la OMS, la Escherichia Coli (E. Coli), Campylobacter y la Salmonella figuran entre los principales patógenos bacterianos de transmisión de alimentos más comunes que afectan a la gente cada año en América (OMS, 2015a). El primero de ellos la E. Coli se asocia con el consumo de leche no pasteurizada, carne poco cocida, frutas y vegetales, puede causar una colitis hemorrágica; el segundo es asociado con ingestión de leche cruda, carne de ave poco cocida y agua potable, causando diarrea acuosa; y finalmente la salmonella se encuentra en productos como huevos, carne de ave y otros productos de origen animal, esta produce un cuadro gastrointestinal que puede presentar complicaciones como artritis reactiva, Síndrome de Reiter o sepsis (Adams y cols., 1999).

Por otro lado en las "metas 2011-2020: Elige vivir sano"del gobierno de Chile (Piñera, s.f.), entre los microorganismos más virulentos causantes de ETA se encuentran la Encefalitis Espongiforme Bovina, Campilobacter jejuni, Escherichia Coli, Salmonella y Shigella, de estas cuatro últimas trata este trabajo. Por lo antes mencionado, en el presente trabajo se plantea dar un nuevo enfoque al estudio de los brotes de ETA en Chile durante el año 2017 desde el punto de vista estadístico-matemático, que sirva de base para que otras disciplinas puedan desarrollar mejores estrategias de control de estos brotes en el país. Para lograr esto, se plantea como objetivo general de este trabajo ajustar modelos de regresión lineal como LASSO, Ridge y modelo lineal generalizado con vinculo Poisson (por la naturaleza de los datos) para evaluar el comportamiento de estos 4 patógenos más virulentos que afectan a la población chilena, y así mejorar la estimación de personas enfermas y encontrar las variables más significativas.

Para lograr esto se lleva a cabo los siguientes 3 objetivos específicos, que son determinar las capacidades y desventajas de los modelos, luego clasificar los modelos según se adecuen a los datos y finalmente se procederá a compararlos y estudiarlos.

Hay que resaltar que el uso de modelos como LASSO y Ridge han sido implementados para la *predicción del riesgo de padecer disfunción motora en adultas mayores activas de la ciudad de Valdivia* (Medina, 2018), estudios en cáncer de próstata (Ramos Castillo, 2018), incluso en diabetes (Hans, 2009)(Tibshirani, 1996).

El trabajo se desarrolla de la siguiente manera, el capítulo I se presenta el marco teórico, donde se abordan los conceptos de los modelos lineales desde el punta de vista de la estadística clásica y el enfoque Bayesiano. Además, se presentan las medidas utilizadas para comparar dichos modelos. Luego, en el capítulo II se muestra la metodología implementada para obtener los modelos bajo ambas perceptivas. Finalmente en el capítulo III, se analizan los resultados obtenidos y se mencionan algunas recomendaciones como resultado de este estudio.

CAPÍTULO I

MARCO TEÓRICO

En el presente capítulo se presentan los fundamentos teóricos acerca de los modelos lineales, la regresión por mínimos cuadrados ordinarios, luego se presentan métodos derivados de esta regresión como lo son los métodos de Operador de Selección y Contracción Mínimo Absoluta (LASSO) y Ridge. Posteriormente, se presentarán las versiones de estos métodos vistas desde el enfoque del análisis Bayesiano incluyendo un método implementado en caso de no conocer a priori la distribución del parámetro desconocido $\boldsymbol{\beta}$. Después en el capítulo se muestra un método para calcular el parámetro de penalización, λ, utilizado en los métodos de LASSO y Ridge. Finalmente, se presentan tres medidas para la bondad de ajuste de los modelos, estos son: el error absoluto medio (EAM), error cuadrático medio (ECM) y el coeficiente de determinación (R^2).

1.1. Modelos lineales

A continuación se detallan los aspectos fundamentales de los modelos lineales, algunos casos donde no se pueden usar y las condiciones que cumplen los modelos lineales generalizados.

Durante el proceso de modelaje de una variable respuesta $\boldsymbol{y}$ de tamaño $n \times 1$ dependiente de otra variable X que es la matriz de diseño de $n \times p$, donde la fila i-ésima representa las observaciones del i-ésimo individuo en las p variables explicativas y la columna j-ésima representa las observaciones de la j-ésima variable en los n individuos; se utilizan modelos de regresión que tienen como fin definir dicha dependencia y a partir del modelo obtenido hacer predicciones cuando se tiene que X es conocida. Una forma de regresión es la lineal, que establece la siguiente relación entre las variables:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

donde $\boldsymbol{\beta}$ es un vector de tamaño $p \times 1$ que representa el parámetro a encontrar y $\boldsymbol{\epsilon}$ es el vector $n \times 1$ de errores aleatorios independientes e idénticamente distribuidos (IID), obtenidos de la aproximación por un plano usando el parámetro $\boldsymbol{\beta}$ (Allasia y cols., 2016).

Existen algunos tipos de variable que no se pueden trabajar con modelos lineales, sino que se trabajan como Modelos Lineales Generalizados (MLG). Por ejemplo: Las variables de conteo de casos, las variables expresadas como proporciones y las variables establecidas como binaria. En estos casos el interés recae en variables cuya distribución pertenece a la familia exponencial, la cual se detalla a continuación.

1.1.1. Familia exponencial

Se considera una variable aleatoria $\boldsymbol{y}$ cuya función de probabilidad o distribución, $P(\boldsymbol{y}|\boldsymbol{\theta})$ depende de un único parámetro $\boldsymbol{\theta}$ conocido. Dicha función de probabilidad pertenece a la familia exponencial con r parámetros si $P(\boldsymbol{y}|\boldsymbol{\theta})$ puede escribirse como:

$$P(\boldsymbol{y}|\boldsymbol{\theta}) = a(\boldsymbol{y})\, exp\left\{\sum_{j=1}^{r} U_j(\boldsymbol{y})\phi_j(\boldsymbol{\theta}) + b(\boldsymbol{\theta})\right\},$$

donde a, b, U y ϕ son funciones conocidas y se cumple ademas que las variables $\boldsymbol{y}$ y $\boldsymbol{\theta}$ solo se relacionan por las funciones U y ϕ; por otro lado $\boldsymbol{y} \in \mathcal{Y} \subset \Re^n$ y $\mathcal{Y}$ no depende de $\boldsymbol{\theta}$ (Migon y Gamerman, 1999).

1.1.2. Modelos Lineales Generalizados

Los Modelos Lineales Generalizados (MLG) son una familia de modelos que incluyen el modelo lineal, estos fueron introducidos por Nelder y Wedderburn en 1972 (Diluvi, 2017). Estos modelos son definidos en término de un conjunto de variables aleatorias independientes $y_1, ..., y_n$, cada una con una distribución perteneciente a la familia exponencial y que cumple las siguientes propiedades:

- La distribución de cada y_i está en forma canónica y depende de un único parámetro θ_i (los θ_i no todos tienen que ser iguales).
- Un conjunto de parámetros

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix},$$

y una variable explicativa X

$$X = \begin{bmatrix} \boldsymbol{X_1}^t \\ \vdots \\ \boldsymbol{X_k}^t \end{bmatrix}.$$

que definen los predictores lineales de la forma $\boldsymbol{\eta} = X\boldsymbol{\beta}$.

- Una función monótona, llamada función de vínculo g, tal que $g(\mu_i) = \boldsymbol{X_i}^t\boldsymbol{\beta}$ donde $\mu_i = E[y_i]$ y además, μ_i es alguna función de θi.

Los tres casos más comunes de MLG son el modelo lineal, la regresión binomial y binaria y la regresión de Poisson. A continuación se detallará más acerca de la regresión de Poisson.

1.1.3. Regresión de Poisson

Este modelo lineal generalizado es considerado para datos de conteo (Dobson, 2002), asumiendo:

- La variable respuesta tiene una función de probabilidad perteneciente a la distribución Poisson.
- Se mantiene los predictores lineales del modelo lineal, $\boldsymbol{\eta} = X\boldsymbol{\beta}$.
- La función de enlace esta dada por el logaritmo de μ, tal que $\boldsymbol{\eta} = g(\mu) = log(\mu)$, donde μ es la media de la variable respuesta.

1.2. Regresión por mínimos cuadrados ordinarios

En esta sección se abordará el problema de la estimación del vector de parámetros, $\boldsymbol{\beta}$, desconocido del modelo.

La técnica de Mínimos Cuadrados Ordinarios (OLS) es el método más común usado para estimar los parámetros desconocidos en un modelo de regresión lineal para minimizar algún residual de cuadrados.

Para este método se escoge como el estimador de $\boldsymbol{\beta}$, el resultado obtenido de resolver un problema de programación cuadrática sin restricciones, como el siguiente:

$$\min_{\boldsymbol{\beta}} \frac{1}{n}\|(\boldsymbol{y} - X\boldsymbol{\beta})\|_2^2, \tag{1.1}$$

de aquí, que el estimador obtenido al resolver la ecuación (1.1) por OLS es:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{OLS}} = argmin_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|(\boldsymbol{y} - X\boldsymbol{\beta})\|_2^2 \right\}. \tag{1.2}$$

Esta técnica presenta un buen comportamiento en el caso de que todos los parámetros del modelo sean significativos. Por tanto, existen dos razones por las que el método de mínimos cuadrados podría no ser adecuado para estimar modelos con variables no relevantes.

1. Precisión de la predicción: Las estimaciones de los parámetros por mínimos cuadrados tienen bajo sesgo pero gran varianza. Por lo que la precisión de la predicción a veces se puede mejorar mediante la reducción o ajuste a cero de algunos coeficientes.

2. Interpretación: En el caso donde se tiene un gran número de predictores o variables explicativas suele ser de interés determinar un subconjunto más pequeño que muestre los efectos más fuertes.

De aquí que se busquen otros métodos que permitan resolver las limitantes del método de mínimos cuadrados ordinarios.

1.3. Selección de variables

En esta parte se exponen los motivos por los cuales se da la necesidad de seleccionar variables en un modelo de regresión.

En ocasiones cuando se estudian variables dependientes se dispone de un conjunto grande de posibles variables explicativas, por lo que surge en la mente del investigador la pregunta acerca de cuales son las variables que debe considerar en el modelo de regresión, para responder dicha interrogante se tiene dos opciones: seleccionar todas las variables explicativas posibles o escoger un subconjunto del universo de variables.

Los métodos de selección de variables se encargan del problema de construir y seleccionar el modelo. En general se tienen dos situaciones al resolver dicho problema, la primera ocurre cuando el investigador selecciona las variables a utilizar para construir un modelo que se ajuste bien y sea sencillo, a partir de la modificación de un modelo inicial. En este caso se tienen dos posibles soluciones: Incluir otras variables o eliminar algunas de las variables presentes en el modelo. En el caso en que se desea incluir cada vez más variables en un modelo de regresión, se da como resultado que el ajuste a los

datos mejora al aumentar la cantidad de parámetros a estimar, pero disminuye su precisión individual (desarrollando mayor varianza) y por lo tanto, en la función de regresión estimada se produce un sobreajuste. Por otro lado, si se encuentra el investigador en el caso donde lo que se desea es excluir algunas variables de un modelo inicial, se tiene como resultado que la eliminación de algunas variables del modelo produce que las varianzas se reducen pero los sesgos aumentarían, produciendo una mala descripción de los datos.

Por otra parte, la segunda situación que se suele presentar ocurre cuando algunas variables predictoras pueden perjudicar la confiabilidad del modelo, esto se produce especialmente cuando están correlacionadas unas variables con otras. De esta manera, el objetivo de los métodos de selección de variables es buscar un modelo que resuelva dichas situaciones, es decir, que se ajuste bien a los datos y que a la vez sea posible tenga un equilibrio entre bondad de ajuste y sencillez.

Al estudiar el método de mínimos cuadrados ordinarios, teniendo como objetivo encontrar un modelo que presente las características antes expuestas, se obtiene un par de razones por las cuales dicho método podría no ser adecuado para seleccionar variables.

La principal razón se produce al estimar los parámetros de regresión por el método de mínimos cuadrados ordinarios, debido a que puede ocurrir que alguna de estas estimaciones sean casi cero y por tanto la variable correspondiente a dicho coeficiente tendría muy poca influencia en el modelo, sin embargo, es poco usual que estas estimaciones sean exactamente el valor cero, por tanto, este método no es adecuado para seleccionar variables. De este modo, se necesitan de otros métodos para lograr tal objetivo.

Uno de los métodos que se desarrolló para resolver este inconveniente al seleccionar variables en el OLS, fue el métodos de mínimos cuadrados penalizados, el cual se basan en los mínimos cuadrados ordinarios pero añadiendo una penalización en la función objetivo, para forzar que alguna componente del vector de parámetros desconocidos $\boldsymbol{\beta}$ sea cero y de esta manera conseguir estimación de los parámetros y selección de variables conjuntamente.

1.4. Métodos de mínimos cuadrados penalizados

A continuación se presentan algunos métodos de regresión basados en mínimos cuadrados penalizados, en particular del método LASSO desde el punto de vista de la estadística clásica y la estadística bayesiana, además del método Ridge.

Algunos métodos pueden resultar inestables o directamente son inaplicables cuando el número de variables p es similar o incluso superior al número de observaciones n. Una

alternativa para estos son los métodos de regresión penalizada. La idea principal es la penalización, para así lograr evitar el sobreajuste debido al gran número de variables explicativas, se impone una penalización o término de penalización, que obligaría que alguna componente del vector de parámetros $\boldsymbol{\beta}$ sea cero.

La elección del parámetro de penalización es de gran importancia: Es necesario un procedimiento que estime el valor de dicho parámetro a partir de los datos. Por tanto, en un intento de seleccionar las variables y de estimar los parámetros de forma automática y simultánea, se propone un enfoque unificado a través de mínimos cuadrados penalizados, que consiste en estimar el vector de parámetros, que resuelva el siguiente problema:

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - X_i^t \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} P_\lambda(|\beta_j|) \right\}, \tag{1.3}$$

donde P_λ es la función de penalización que será diferente para cada método y λ es el parámetro de penalización. Ademas, el parámetro λ debe ser elegido a través de algún procedimiento basado en los datos muestrales.

De este modo, se estima el vector de parámetros $\boldsymbol{\beta}$ como aquel que minimiza la expresión (1.3) y se denotará como $\hat{\boldsymbol{b_n}}$. Naturalmente si $\lambda = 0$ este estimador se corresponde con el estimador de mínimos cuadrados ordinarios, que es denotado por $\hat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}$ descrito en la ecuación (1.2).

A continuación se proponen tres condiciones deseables que un método de penalización debería cumplir:

1. Esparsidad: Realizar selección de variables automáticamente, es decir, tener la capacidad de fijar coeficientes a cero.

2. Continuidad: Ser continuo en los datos para evitar inestabilidad en la predicción.

3. Insesgadez: Tener bajo sesgo, especialmente para valores grandes de los coeficientes β_j.

Ahora se describirán dos métodos de regresión penalizada. El método LASSO y el método Ridge, los cuales se diferencian en el tipo de penalización (P_λ) utilizada.

1.4.1. Método de LASSO clásico

En 1996 es cuando se introduce un nuevo método de análisis de regresión llamado operador de Selección y Contracción Mínimo Absoluta (Least Absolute Shrinkage and Selection Operator, LASSO, por sus siglas en ingles) basado en el método de mínimos cuadrados utilizando penalización (Tibshirani, 1996).

El LASSO en la estadística clásica desarrollado por Robert Tisbshirani es una técnica de regresión y selección de variables por mínimos cuadrados penalizados basado en el método de Garrote no negativo de Leo Breiman (Tibshirani, 1996). Además, LASSO considera restricciones con la norma 1 ($\mathcal{L}_1$) en los coeficientes de regresión estimados, dando como resultado que la función $P_\lambda(|\beta_j|) = |\beta_j|$ de la ecuación (1.3).

Así, siendo $\boldsymbol{\beta}$ el vector de coeficientes, el método LASSO lo estima minimizando el siguiente problema de mínimos cuadrados penalizados:

$$\text{mín}_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|(\boldsymbol{y} - X\boldsymbol{\beta})\|_2^2 \right\} \qquad \text{Sujeto a :} \|\boldsymbol{\beta}\|_1 \leq t \ , \tag{1.4}$$

siendo t el parámetro de regularización o de penalización. De aquí se tiene que el estimador obtenido al resolver el problema dado en la ecuación (1.4), correspondiente a LASSO clásico es:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{LASSO}} = argmin_{\boldsymbol{\beta}} \{ (\tilde{\boldsymbol{y}} - X\boldsymbol{\beta})^{'} (\tilde{\boldsymbol{y}} - X\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1 \}, \tag{1.5}$$

donde, $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \bar{y}\mathbf{1}_{\boldsymbol{n}}$, con $\mathbf{1}_{\boldsymbol{n}}$ el vector de unos de longitud n y $\lambda \geq 0$ el cual determina la influencia de la penalización en la estimación.

Para valores grandes de λ o valores pequeños de t, los coeficientes β_j se contraen hacia cero y alguno se anula, por eso se dice que LASSO produce estimación de parámetros y selección de variables simultáneamente.

Entre las desventajas presentadas por María J. A. Medina en su trabajo de maestría (Medina, 2018), para este modelo se encuentran:

- No es posible seleccionar una cantidad mayor de variables explicativas que el número de observaciones.
- Cuando hay problemas de multicolinealidad el modelo de LASSO selecciona sólo una variable entre muchas variables que se encuentran correlacionadas.

- LASSO no es adecuado usarlo cuando no se cumplen las siguientes propiedades: Consistencia del estimador de parámetros, normalidad asintótica, consistencia del modelo y esparcidad.

1.4.2. Método Ridge

El método de Ridge fue propuesto originalmente por Hoerl y Kennard en 1970, como un método para eludir los efectos adversos del problema de colinealidad en un modelo lineal estimado por OLS, en el contexto $p < n$, esta basado también en el método de mínimos cuadrados penalizados, donde la función de penalización presente en la ecuación (1.3), es $P_\lambda(|\beta_j|) = \beta_j^2$ con los β_j estandarizados (Hoerl y Kennard, 1970), quedando el problema como:

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - X_i^t \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{1.6}$$

donde $\lambda \geq 0$ es el parámetro de contracción que se determinará por separado. De este modo se tiene que el estimador que resuelve la ecuación (1.6) es:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{Ridge}} = argmin_{\boldsymbol{\beta}} \{ (\tilde{\boldsymbol{y}} - X\boldsymbol{\beta})^{'} (\tilde{\boldsymbol{y}} - X\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_2^2 \}, \tag{1.7}$$

En general, la regresión Ridge produce predicciones más precisas que OLS y selección de variables.

Entre las desventajas presentadas por María Carrasco en su tesis de grado (Carrasco C., 2016), para este modelo se encuentran:

- Este método contrae todos los coeficientes hacia cero sin llegar a la nulidad.
- Los modelos suelen ser más complejos y difíciles de interpretar.
- El número de predictores que se relaciona con la respuesta no se conoce a priori para los conjuntos de datos reales.

1.5. Metodos Bayesianos

En esta sección se presentarán los métodos trabajados bajo el enfoque de la estadística Bayesiana. Se mostrará las versiones de los métodos de LASSO y Ridge desde el enfoque Bayesiano.

$$P(y|\mu,\theta) = exp\left\{\frac{\mu y - \frac{1}{2}\mu^2 I}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + log(2\pi\sigma^2)\right)\right\} \tag{1.10}$$

donde I es la matriz identidad.

De esta forma se puede pensar en un MLG bajo el enfoque Bayesiano como aquel donde la matriz de diseño es X de tamaño $n \times p$ la cual se modifica agregando una columna de unos a la izquierda quedando la matriz con dimensión $n \times (p+1)$, y cuya variable de respuesta es de la forma:

$$\boldsymbol{y} \sim \mathcal{N}_n(X\boldsymbol{\beta}, \sigma^2 I_n), \tag{1.11}$$

donde I_n es la matriz identidad de orden n.

Por otro lado, Guitierrez-Peña en su trabajo *Análisis Bayesiano de Modelos Jerárquicos Lineales* (Guitiérrez-Peña, 2016) argumentan que la verosimilitud del modelo esta dada por:

$$P(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\tau}) \propto \tau^{n/2} exp\left\{-\frac{\tau}{2}[(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T X^T X(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + n\hat{\sigma}^2]\right\}, \tag{1.12}$$

donde $\tau = \frac{1}{\sigma^2}$ que se conoce como la precisión, $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$, y $\hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})$, los cuales son los estimadores de máxima verosimilitud de $\boldsymbol{\beta}$ y σ^2. De este modo que la familia para la distribución a priori tiene la forma:

$$P(\boldsymbol{\beta},\tau) = \mathcal{N}_p(\boldsymbol{\beta}|\boldsymbol{b_0}, \tau^{-1}B_0) Gamma\left(\tau|\frac{a}{2},\frac{b}{2}\right), \tag{1.13}$$

donde $\boldsymbol{b_0} \in \Re^{p+1}$, $B_0 \in \mathcal{M}_{(p+1)\times(p+1)}(\Re)$ es una matriz simétrica definida positiva, y $a, d > 0$ son los hiperparámetros del modelo.

Ahora bien, ya se mostró como esta definida para el análisis Bayesiano un modelo de regresión lineal; pero cuando no se tiene información a priori de la distribución de $\boldsymbol{\beta}, \tau$, en este caso se procede al uso de una distribución a priori no informativa.

Inicialmente se propuso el uso de una priori uniforme para representar poca o ninguna información disponible sobre los parámetros (Bayes, 1763), otra opción la propuso Jeffreys en 1961, con la priori no informativa que lleva su nombre y que depende de la información de Fisher (Migon y Gamerman, 1999).

Los métodos Bayesianos aparecen en el siglo XVIII con los trabajos de Bayes y Laplace, pero no es sino hasta finales de la década de 1980 y principios de la década de 1990 cuando la aparición de métodos computacionalmente intensivos basados en simulación, como lo son los métodos de cadenas de Markov Montecarlo (MCMC), que se permite implementar el paradigma del análisis Bayesiana.

El análisis Bayesiano se refiere al proceso de obtener conclusiones, a partir de la información presente en los datos numéricos observados, sobre cantidades no observadas. Este busca tratar de manera unificada la inferencia y la decisión, tomando en consideración la incertidumbre asociada al modelo y a los parámetros, proporcionando de una vez las herramientas para cuantificar el grado de incertidumbre que se tiene. Por otro lado, el tratamiento de las cantidades no observadas como variables aleatorias y el análisis condicional, permiten naturalmente considerar modelos jerárquicos o de variables latentes que son difíciles o imposibles de manejar con la estadística clásica (Bravo y cols., 2008).

Entre las dificultades de la inferencia desde este punto de vista están: la necesidad de establecer una distribución previa (priori) sobre las cantidades no observables, proponer una distribución de muestreo o verosimilitud para las cantidades que se pueden observar y la dificultad de encontrar las varias integrales requeridas este enfoque Bayesiano.

El análisis Bayesiano deriva del teorema de Bayes el cual establece que:

Sea $P(\boldsymbol{\theta}, \boldsymbol{y})$ la distribución de θ y $\boldsymbol{y}$. En general se puede escribir $P(\boldsymbol{\theta}, \boldsymbol{y})$ como $P(\boldsymbol{\theta})P(\boldsymbol{y}|\boldsymbol{\theta})$, con $P(\boldsymbol{\theta})$ la distribución previa o priori y $P(\boldsymbol{y}|\boldsymbol{\theta})$ la distribución condicional de $\boldsymbol{y}$ dado $\boldsymbol{\theta}$ o distribución de muestreo. Entonces la distribución a posteriori de $\boldsymbol{\theta}$, condicionada sobre los valores conocidos de $\boldsymbol{y}$, viene dada por (Bravo y cols., 2008):

$$P(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{P(\boldsymbol{\theta}, \boldsymbol{y})}{P(\boldsymbol{y})} = \frac{P(\boldsymbol{\theta})P(\boldsymbol{y}|\boldsymbol{\theta})}{P(\boldsymbol{y})}. \tag{1.8}$$

Para $\boldsymbol{y}$ fijo, sin pérdida de generalidad la ecuación (1.8) queda como:

$$P(\boldsymbol{\theta}|\boldsymbol{y}) \propto P(\boldsymbol{\theta})P(\boldsymbol{y}|\boldsymbol{\theta}). \tag{1.9}$$

Las distribuciones iniciales comúnmente asignadas para el caso de modelos de regresión lineales clásicos vistos desde el análisis Bayesiano suponen que $y \sim \mathcal{N}(\mu, \sigma^2)$ (Diluvi, 2017), así:

1.5.1. Priori aplanada o flat

En 1763, se propuso cuando la información que se conoce en principio sobre la distribución a priori es escasa o no existe el uso de una distribución uniforme (Migon y Gamerman, 1999), la cual puede ser expresada como:

$$P(\boldsymbol{\theta}) \propto k \,\text{para}\, \boldsymbol{\theta} \in K \subset \Re^n. \tag{1.14}$$

Nótese que esta distribución no es propia si el rango de $\boldsymbol{\theta}$ no es acotado, lo cual va en contra de las reglas básicas de probabilidad, por lo que hay que tener cuidado en el uso puesto que lo que se busca es una distribución a posteriori propia sin importar si la distribución a priori no lo es.

1.5.2. Método de LASSO Bayesiano

En el año 2008, se estudió una variación de este método con enfoque en la estadística Bayesiana, el modelo denominado LASSO Bayesiano, es aquel el cual se utiliza una distribución a priori doble-exponencial (Park y Casella, 2008). Quedando el estimador para el método de LASSO Bayesiano de la siguiente manera:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{LASSO}} = argmax_{\boldsymbol{\beta}} \left\{ f(\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2, \lambda) \right\}, \tag{1.15}$$

donde, f es una función de probabilidad, σ^2 y λ son parámetros de la doble-exponencial (Allasia y cols., 2016). El estimador $\hat{\boldsymbol{\beta}}_{\boldsymbol{LASSO}}$ puede pensarse como la moda de la distribución a posteriori de $\boldsymbol{\beta}$. Además, la priori se representa por:

$$\boldsymbol{\beta} \sim f(\boldsymbol{\beta}|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} e^{\frac{\lambda|\beta_j|}{\sqrt{\sigma^2}}},$$

también se considera una priori marginal no informativa $f(\sigma^2) \sim \frac{1}{\sigma^2}$ para σ^2 o cualquier distribución gamma inversa, para mantener la conjugación (Park y Casella, 2008).

1.5.3. Método de Ridge Bayesiano

Un enfoque Bayesiano del método de regresión de Ridge es obtenido al notar que el minimizador del estimador de Ridge (bajo la estadística clásica) puede ser considerado como la media a posterior de un modelo, donde $\boldsymbol{\beta} \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda} I)$ y que la verosimilitud esta definida por:

$$\boldsymbol{y}|X, \boldsymbol{\beta} \sim \mathcal{N}_n(X\boldsymbol{\beta}, \sigma^2 I_n),$$

donde $\boldsymbol{y}$ es la variable respuesta X la matriz de diseño que contiene los datos observados para las variables explicativas, σ^2 es la desviación estándar de los datos y finalmente I_n es la matriz identidad de tamaño $n \times n$.

Esto permite calcular la distribución a posterior deseada como

$$P(\boldsymbol{\beta}|\boldsymbol{y}, X) \propto P(\boldsymbol{\beta}) P(\boldsymbol{y}|X, \boldsymbol{\beta}),$$

(Roan Veerman, 2018), dando como resultado el estimador:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{Ridge}} = argmax_{\boldsymbol{\beta}}\, exp\left[\frac{-1}{2\sigma^2}(\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) - \frac{-\lambda}{2\sigma^2}||\boldsymbol{\beta}||_2^2\right]. \tag{1.16}$$

1.6. Métodos para seleccionar el parámetro de penalización λ

En esta sección se presentan los métodos para escoger el parámetro de penalización λ en los métodos de mínimos cuadrados penalizados.

Como puede observarse todas las técnicas de mínimos cuadrados penalizados dependen de un parámetro de penalización λ, que controla la importancia dada a dicha penalización en el proceso de optimización. Además, cuanto mayor es λ, mayor es la penalización en los coeficientes de $\boldsymbol{\beta}$ de la regresión y más son contraídos estos hacia cero. Nótese también que si $\lambda = 0$ la estimación coincide con la de mínimos cuadrados ordinarios.

Algunos de los métodos para seleccionar el parámetro λ son los siguientes:

1. Una propuesta inicial y que continúa siendo sugerida por algunos autores es la utilización de una traza Ridge para determinar λ. Esto consiste en gráficar simultáneamente los coeficientes de regresión estimados en función de λ, y elegir el valor más pequeño del parámetro para el cuál se estabilizan dichos coeficientes.

CAPÍTULO II

MARCO METODOLÓGICO

En el siguiente capítulo se explican y presentan los modelos implementados en el software libre estadístico R (R-foundation, s.f.) para predecir el número de enfermos por Enfermedades Transmitidas por Alimentos (ETA) en Chile durante el año 2017.

2.1. Los Datos

Los datos utilizados en esta investigación tratan acerca de los brotes de ETA presentados en Chile durante el año 2017. Esta base de datos fue extraída de la página del gobierno de Chile perteneciente al Departamento de Estadística e Información de Salud, en Chile (DEIS),(DEIS, 2019a). La misma esta compuesta de 133 variables obtenidas a través del sistema de vigilancia de ETA desarrollado en el país, de las cuales se considera solo los siguientes patógenos de Campylobacter spp, Escherichia coli, Salmonella spp y Shigella spp., de acuerdo con los patógenos más virulentos según la Organización Mundial de la Salud (OMS) y el Gobierno de Chile en su publicación *METAS 2011-2020* (Piñera, s.f.) y las siguientes 55 de las 133 variables(DEIS, 2019):

Variables consideradas en el estudio

Nauseas, Vomitos, Diarrea, Dolores, Heces Sang, Parestesias, Otros, Neurológicos, Espasmos, Fiebre, Deshidratacion, Hipotension, Rush Cutaneo, Cefalea, Mialgia, Meteorismo, Otros: Variable lógica sobre la presencia de cada síntoma.

Semana: Comprende las semanas estadísticas de la 1 a la 52.

R.notificación: La región donde se notificó el brote de las 15 regiones del país.

R.consumo: La región donde se dió el consumo del alimento sospechoso de originar el brote de las 15 regiones del país.

Al implementar este coeficiente como una medida de bondad de ajuste del modelo se debe tener en cuenta que entre más cercano sea a 1 mejor sera el modelo.

1.7.4. Nivel de significancia

Se define de la siguiente manera el nivel de significancia, valor p o $p-valor$ (Wackerly y cols., 2008):

"Si W un estadístico de prueba, el valor p, o nivel de significancia alcanzado, es el nivel más pequeño de significancia α para el cual la información observada indica que la hipótesis nula debe ser rechazada".

Para concluir, se tiene que las medidas anteriores indican la bondad del ajuste sobre los propios elementos, pero no dan información sobre la bondad del ajuste para una observación diferente de la muestra.

$$EAM = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|, \tag{1.17}$$

donde N es la cantidad de datos, $\hat{y}_i$ son los valores estimados por el modelo y y_i los valores reales. Lo importante de esta métrica en el calculo de errores es que penaliza los valores grandes de los mismos, por lo que no es tan sensible a los valores atípicos como el error cuadrático medio, por lo cual es considerado como una métrica más robusta.

Utilizando este error se considera un buen modelo aquel cuyo valor de EAM sea más pequeño, por lo cual se obtiene ajuste mejor del modelo.

1.7.2. Error Cuadrático Medio

El error cuadrático medio (ECM) es una forma de evaluar la diferencia entre un estimador y el valor real de la cantidad que se quiere calcular. El ECM mide el promedio del cuadrado del "error", siendo el error el valor en la que el estimador difiere de la cantidad a ser estimada. En otras palabras, se esta construyendo es estimador muestral de $E((\boldsymbol{y} - X\boldsymbol{\beta})^2)$ como:

$$ECM = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}, \tag{1.18}$$

donde N es la cantidad de datos, $\hat{y}_i$ son los valores estimados y y_i los valores reales.

Para considerar un buen modelo según el ECM se debe tomar en cuenta que mientras más pequeña sea esta medida de error, mejor es el ajuste del modelo.

1.7.3. Coeficiente de determinación R^2

El coeficiente de determinación R^2 expresa la proporción de variación de la variable y que es explicada por la variable X (variable predictora o explicativa). Si la proporción es igual a 0, significa que la variable predictora no tiene ninguna capacidad predictiva de la variable respuesta ($\boldsymbol{y}$). Cuanto mayor sea R^2, mejor sería la predicción. Si llegara a ser igual a 1 la variable predictora explicaría perfectamente la variación de $\boldsymbol{y}$, y las predicciones no tendrían error. Este coeficiente se define matemáticamente como:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{Y}_i)^2}, \tag{1.19}$$

donde $\hat{y}_i$ son los valores estimados, y_i son los valores reales y $\bar{Y}_i$ es el promedio de los y_i.

2. Un método más automático, pero intensivo computacionalmente, consiste en estimar λ mediante validación cruzada.

1.6.1. Validación Cruzada (CV)

Uno de los métodos más utilizados para estimar el parámetro λ es el método validación cruzada por k-pliegues (k-fold cross-validation, en ingles).

El método de validación cruzada consiste en dividir el modelo en un grupo de datos de entrenamiento (training set) para ajustar un modelo y un grupo de prueba (test set) para evaluar su capacidad predictiva, mediante el error de predicción u otra medida.

La forma en que se aplica la validación cruzada es mediante la división del conjunto de datos disponibles de manera aleatoria en k subconjuntos o pliegues de igual tamaño y mutuamente excluyentes.

Uno de los subconjuntos se utiliza como datos de prueba y el resto $(k-1)$ como datos de entrenamiento. El proceso de validación cruzada es repetido durante k iteraciones, con cada uno de los posibles subconjuntos de datos de prueba. Finalmente, se realiza la media aritmética de los resultados de cada iteración para obtener un único resultado. De aquí que el valor del parámetro será el que dé el mínimo error. Este método es muy preciso puesto que evaluamos a partir de k combinaciones de datos de entrenamiento y de prueba, pero aún así tiene una desventaja, y es que es lento desde el punto de vista computacional.

1.7. Medidas de bondad de ajuste

En esta sección se explican los métodos utilizados para comparar los ajustes de modelos obtenidos a partir de los métodos de regresión antes expuestos.

1.7.1. Error Absoluto Medio

Según Arias en su libro "Análisis introductorio de métricas básicas para el cálculo de errores de pronósticos", el Error Absoluto Medio (EAM) se calcula como un promedio de diferencias absolutas entre los valores reales y las predicciones (Arias, 2016). Este error es una puntuación lineal, lo que significa que todas las diferencias por observación se ponderan por igual en el promedio. Se define por la ecuación:

R.N-RC: Variable lógica sobre la comparación entre las regiones de consumo y notificación.

L.elaboración: Los 5 tipos de locales de elaboración de alimentos.

L.consumo: Los 5 tipos de locales de consumo de alimentos.

P.P.I: Los 8 tipos de procesos de pérdida de inocuidad de los alimentos.

L.P.I: Los 3 tipos de lugares donde se dió la pérdida de inocuidad de los alimentos.

Grupo Sospechoso: Los 14 grupos alimenticios de contaminación.

F.Contaminacion: Los 8 tipos de factores de contaminación de alimentos.

F.Supervivencia: Los 5 tipos de factores de supervivencia de los contaminantes de los alimentos.

F.Proliferacion: Los 8 tipos de factores de Proliferación de los contaminantes de los alimentos.

d.agrupado: Los 4 tipos de diagnosticos agrupados por los patógenos.

CIE-10: Los 32 tipos de diagnosticos de enfermedades.

Expuestos: Cantidad de expuestos al brote.

Enfermos: Cantidad de enfermos por el brote.

P.Incubación: Días que tarda en presentarse los síntomas.

Duración.B: Periodo de duración del brote.

Tasa Ataque: Tasa en la que los expuestos enfermaron.

T.Amb: Cantidad de personas enfermas atendidas en ambulatorios.

T.ambk: Cantidad de personas enfermas atendidas en ambulatorios en los rangos de edad k: menor a 1 año (1), entre 1 y 4 (1-4), entre 5 y 14 (5-14), entre 15 y 44 (15-44), entre 45 y 64 (45-64), y mayores de 65 (65).

T.Hosp: Cantidad de personas enfermas atendidas en hospitales.

T.hospk: Cantidad de personas enfermas atendidas en hospitales en los rangos de edad k: menor a 1 año (1), entre 1 y 4 (1-4), entre 5 y 14 (5-14), entre 15 y 44 (15-44), entre 45 y 64 (45-64), y mayores de 65 (65).

T.s.a: Cantidad de personas enfermas que no recibieron atención medica.

T.s.ak: Cantidad de personas enfermas que no recibieron atención medica en los rangos de edad k: entre 5 y 14 (5-14), entre 15 y 44 (15-44), entre 45 y 64 (45-64), y mayores de 65 (65).

C.Inspec: Variable lógica sobre la necesidad de inspección de un brote de ETA.

Se eliminaron de la base de datos las variables de total de personas que no recibieron atención medica cuyo grupo etario son menores de 1 año y entre 1 y 4 años (T.s.a1 y T.s.a1-4 respectivamente), esto debido a que no se presentaron ningún caso en dichas variables.

2.2. Análisis Exploratorio

En esta sección se presentan los resultados obtenidos del análisis descriptivo de los datos antes mencionados, que comprenden 116 observaciones en sus 55 variables, implementados en R, versión 3.6.3. Para ello se utilizaron tablas, diagramas de barra y de cajas para resumir la información almacenada en la base de datos.

En los diagramas de barra de las variables lógicas (ver Figura 3.7 del Anexo 1) se observa que hay presencia de los síntomas diarrea, dolores, vómitos, nauseas y fiebre, los cuales están resumidos en la Tabla 2.1. Se observa que el síntoma que se presenta en la mayoría de los brotes de la muestra es la diarrea en un 96,55 %, seguido por dolores con un 89,66 % resaltados de rojo, mientras que la fiebre se presenta en menor medida con un 72,41 % resaltado de color azul.

Tabla 2.1: Síntomas presentes: Diarrea, vómito, nauseas, dolores y fiebre. Tabla: Cantidad y porcentaje.

Variable	NO	SI	NO (%)	SI (%)
Diarrea	4	112	3,45	96,55
Dolores	12	104	10,34	89,66
Vómitos	19	97	16,38	83,62
Nauseas	31	85	26,72	73,28
Fiebre	32	84	27,59	72,41

Por otro lado, al resumir la información encontrada en los diagramas de barra de los síntomas que no se presentan en la mayoría de los brotes (ver Figura 3.8 del Anexo 2), se encontró que hay mayor cantidad de respuestas negativas en las variables parestesias, rush cutáneo, meteorismo, espasmos, hipotensión, heces sanguinolentas (Heces Sang), mialgia, deshidratación, cefalea, otros neurológicos(Neurológicos) y otros síntomas (otros). De las variables anteriores se encontró que la menos frecuente es otros neurológicos con una ausencia de 99,14 %, seguida por parestesia, rush cutáneo y meteorismo, todas con un 98,28 % de ausencia en los brotes las cuatro variables resaltadas en azul en la Tabla 2.2.

Tabla 2.2: Síntomas no presentes: Parestesia, rush cutáneo, meteorismo, espasmo, hipotensión, mialgia, heces sanguinolentas, deshidratación, cefalea, otros neurológicos y otros. Tabla: Cantidad y porcentaje.

Variable	NO	SI	NO (%)	SI (%)
Otros Neurológicos	115	1	99,14	0,86
Parestesias	114	2	98,28	1,72
Rush Cutáneo	114	2	98,28	1,72
Meteorismo	114	2	98,28	1,72
Espasmos	111	5	95,69	4,31
Hipotensión	110	6	94,83	5,17
Heces Sanguinolentas	109	7	93,97	6,03
Mialgia	109	7	93,97	6,03
Otros	94	22	81,03	18,97
Deshidratación	77	39	66,38	33,62
Cefalea	71	45	61,21	38,79

Mientras que la cefalea es el síntoma que tiene menor porcentaje de ausencia en los brotes con un 61,21 % resaltado de rojo en la Tabla 2.2.

En cuanto a las semanas del brote (Semana) presentadas en el diagrama de barra de la Figura 2.1, se tiene que las más frecuentes son las semana 10 seguida de 3, 11, 50 y 9, es decir, los meses de marzo, enero y diciembre (colores dorados, aguamarina y amarillo). De lo anterior se puede observar que es más común este tipo de enfermedades durante la estación de verano específicamente al final de la temporada (Figura 2.1).

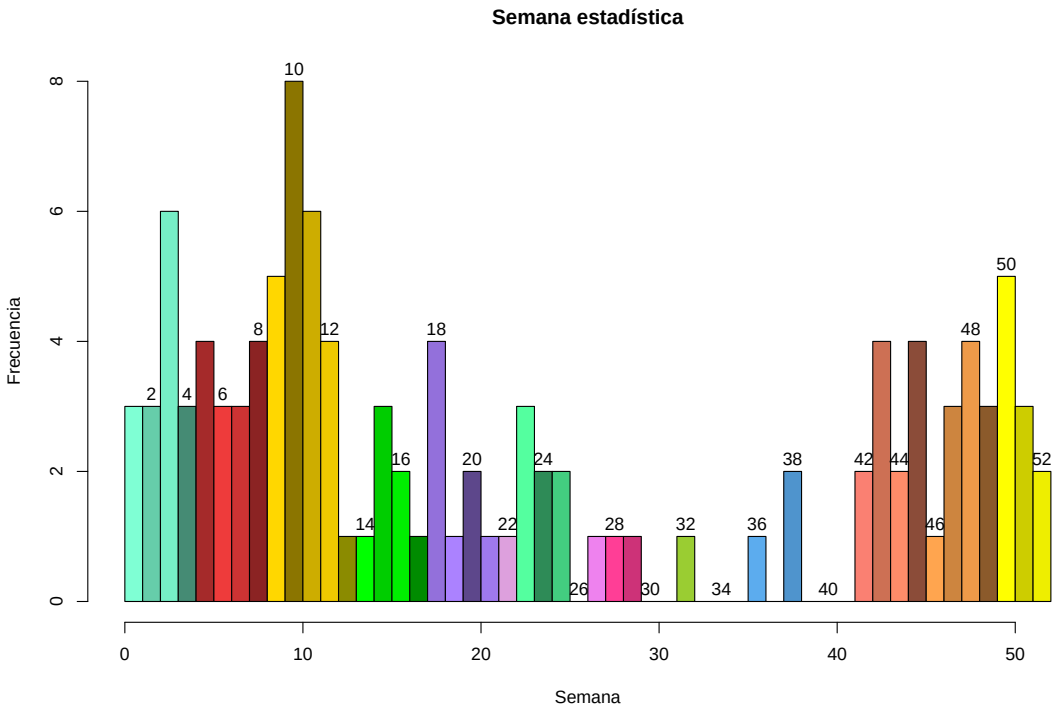
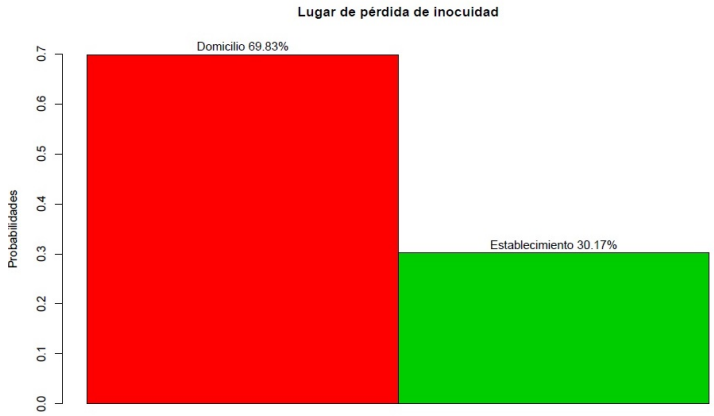


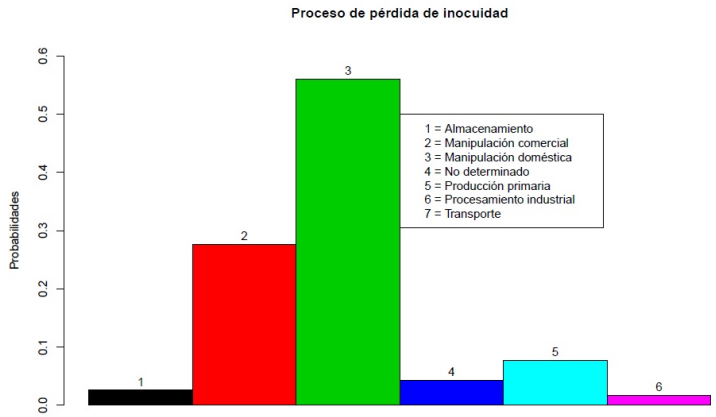
Figura 2.1: Semana de brote. Diagrama de barra

Al observar los diagramas de barra de las regiones de consumo y notificación (R.consumo y R.notificación, respectivamente) de la Figura 2.2, se encontró que la región Metropolitana de Santiago es aquella en la que se notifican los brotes más frecuentemente seguida de Arica y Parinacota (Figura 2.2(a)), igualmente para las de consumo se mantienen estas regiones en el mismo orden (Figura 2.2(b)). Por otro lado, la región con menor porcentaje de brotes de ETA notificados es Los Ríos, barra número 14 en la Figura 2.2(a). Mientras para la de consumo que causó el brote se tienen Los Lagos, Aisén del General C. Ibañez del Campo, Los Ríos y la de Magallanes y de La Antártica chilena, barras 10, 11,12 y 14 en la Figura 2.2(b).

con 27,58 % (barra roja) y finalmente el menos frecuente es el procesamiento industrial con 1,72 % (barra fucsia)(Figura 2.4(b)).



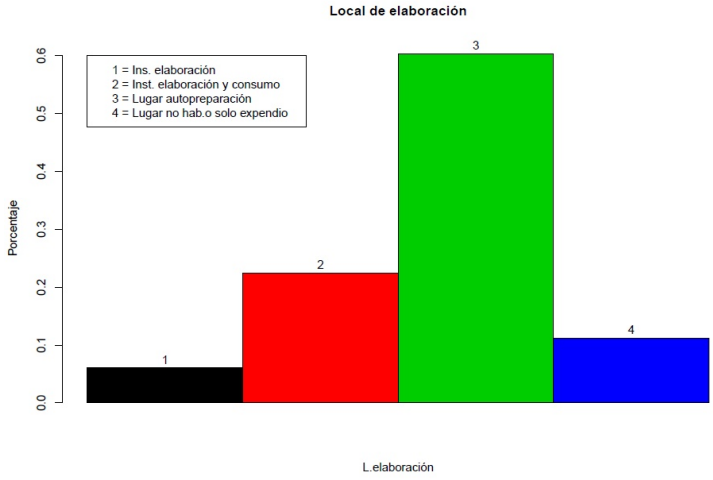
(a) Lugar de pérdida



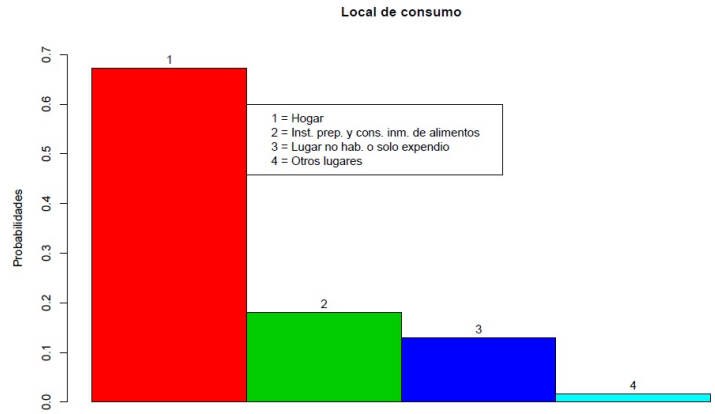
(b) Proceso de pérdida

Figura 2.4: Pérdida de inocuidad. Diagramas de barra. (a) Lugar de pérdida y (b) Proceso de pérdida.

preparación y consumo inmediato de los alimentos (barra verde, número 2) con 18,10 %. En la Figura 2.3(a) se observa que el local de elaboración más frecuente es el lugar de autopreparación con 60,35 % (barra verde), seguido de la instalación de elaboración y consumo con 22,41 % (barra roja).



(a) Local de elaboración



(b) Local de consumo

Figura 2.3: Locales estudiados. Diagramas de barra. (a) Local de elaboración y (b) Local de consumo

Al estudiar el lugar de pérdida de inocuidad (L.P.I) se tiene que predomina con 69,83 % el domicilio, quedando con un 30,17 % los establecimientos(Figura 2.4(a)). Mientras que al analizar el proceso de pérdida de inocuidad (P.P.I) se tiene que el más frecuente es la manipulación domestica con 56,03 % (barra verde), seguida de la manipulación comercial

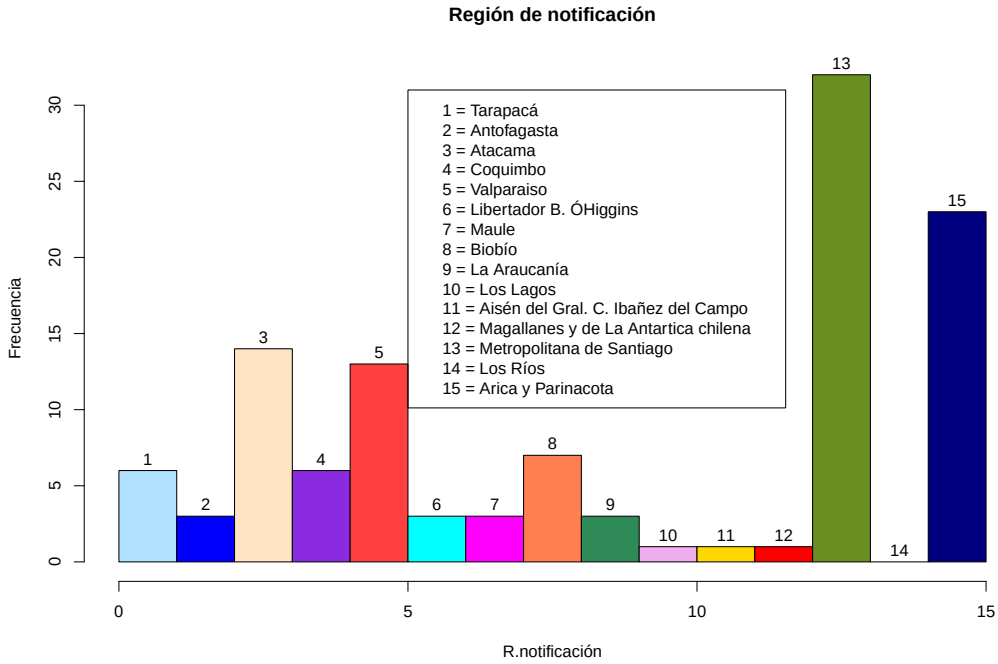
Al resumir la información relacionada con las regiones estudiadas para los brotes de ETA en la Tabla 2.3, se tiene la cantidad de brotes por región y su porcentajes. Nótese que los brotes que representaban el 1,72 % donde no coinciden las regiones de notificación y consumo del alimento asociado a los mismo, están presentes en las regiones Metropolitana de Santiago y la región de Los Ríos, ambas resaltadas de azul, las cuales difieren en un brote entre la zona donde se notifica y donde se consume el alimento. Por otro lado, cuando se observan los porcentajes de estas regiones se encuentra que el cambio de los brotes se realiza entre la región con mayor y aquella de menor porcentaje de casos, lo cual produce una reducción en los brotes causados en la región Metropolitana y un aumento en la región de Los Ríos.

Tabla 2.3: Regiones estudiadas. Tabla: Cantidad de brotes y porcentaje.

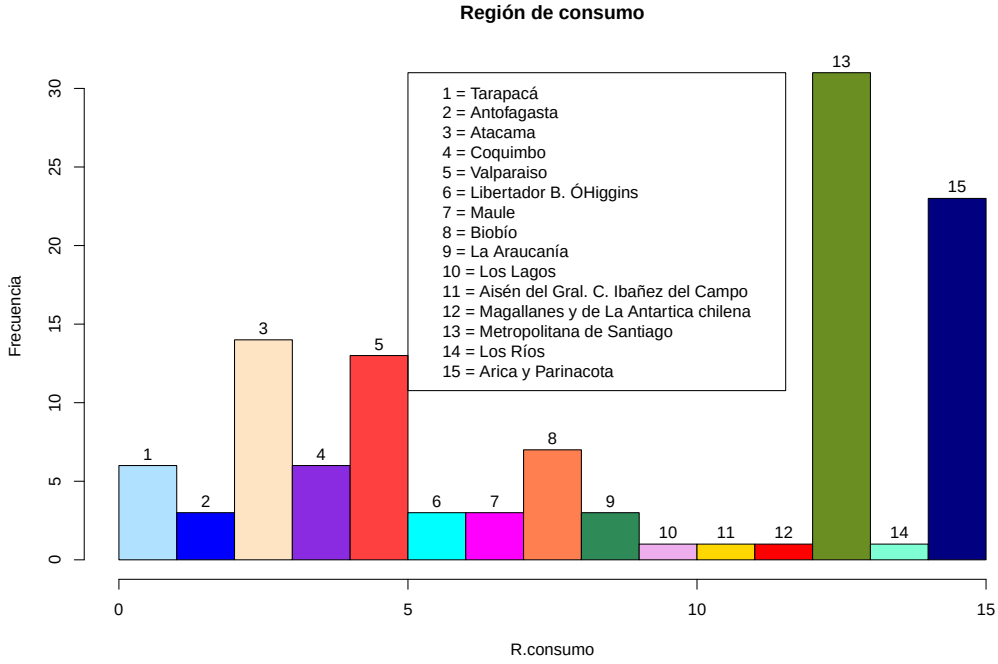
Regiones	Notificación	Consumo	Notificación (%)	Consumo (%)
Metropolitana de Santiago	32	31	27,59	26,72
Arica y Parinacota	23	23	19,83	19,83
Atacama	14	14	12,07	12,07
Valparaiso	13	13	11,21	11,21
Biobío	7	7	6,03	6,03
Tarapacá	6	6	5,17	5,17
Coquimbo	6	6	5,17	5,17
Antofagasta	3	3	2,59	2,59
Libertador B. ÓHiggins	3	3	2,59	2,59
Maule	3	3	2,59	2,59
La Araucanía	3	3	2,59	2,59
Los Lagos	1	1	0,86	0,86
Aisén del Gral. C. Ibañez del Campo	1	1	0,86	0,86
Magallanes y de La Antártica chilena	1	1	0,86	0,86
Los Ríos	0	1	0,00	0,86

Sobre la coincidencia de las regiones de notificación y consumo en se observa que la mayoría de los brotes de ETA presentes en la base de datos estudiada coinciden en un 98,27 % de los casos las regiones donde se notifica y consume el alimento causante.

En cuanto al local de consumo (L.consumo) se puede observar en la Figura 2.3(b) que predomina el hogar con 67,24 % (barra roja, número 1), a este lo sigue la instalación de



(a) Notificación



(b) Consumo

Figura 2.2: Regiones estudiadas. Diagrama de barra

En la Tabla 2.4 los grupos alimenticios sospechosos más frecuentes en ser los causantes de los brotes de ETA son los huevos con 41,38 %, seguido por los platos preparados con 34,48 %, ambos resaltados en rojo. Por otro lado, los menos frecuentes que causen estas enfermedades son los grupos de frutas y hortalizas con 1,72 %, bebidas con 1,72 %, helados con 0,86 % y estimulantes con 0 %, todos resaltados en color azul.

Tabla 2.4: Grupo alimenticio. Tabla: Cantidad y porcentaje.

Grupo alimenticio	Cantidad	Porcentaje(%)
Bebidas	2	1,72
Carnes	6	5,17
Platos preparados	40	34,48
Estimulantes y fruitivos	0	0,00
Frutas y hortalizas	2	1,72
Huevos	48	41,38
Productos lácteos	3	2,59
No identificado	3	2,59
Productos de la pesca	6	5,17
Panadería y pastelería	5	4,31
Helados	1	0,86

El factor contribuyente a la contaminación más frecuente es aquel debido a productos crudos o ingredientes contaminados por patógenos con un 54,31 %, seguido por la contaminación cruzada con 16,37 %, luego están los factores sin identificar 9,48 %, no presentes en la lista 6,89 % y sin definir 6,03 % (Figura 2.5). Se observa que el factor de sustancia tóxica agregada intencionalmente no esta presente en la base de datos (OPS, 2015a).

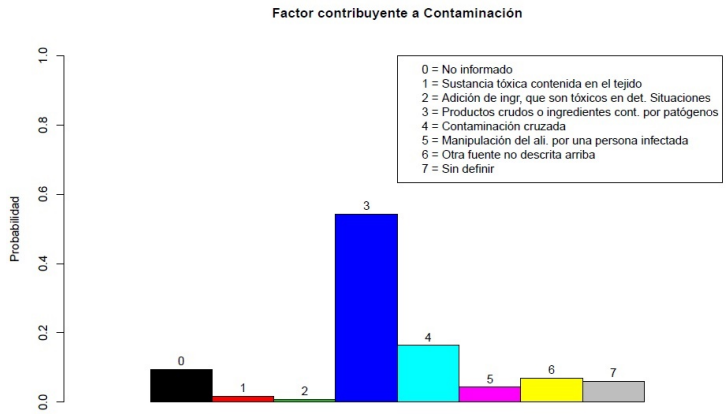


Figura 2.5: Factor contribuyente a la contaminación. Diagrama de barra

Para el factor contribuyente a la proliferación se encontró que es más común no identificarlo en la lista de los factores considerados con un 47,41 %, seguido de no informarlo con 37,93 % y por último debido inadecuada conservación con 14,65 % (Figura 2.6).

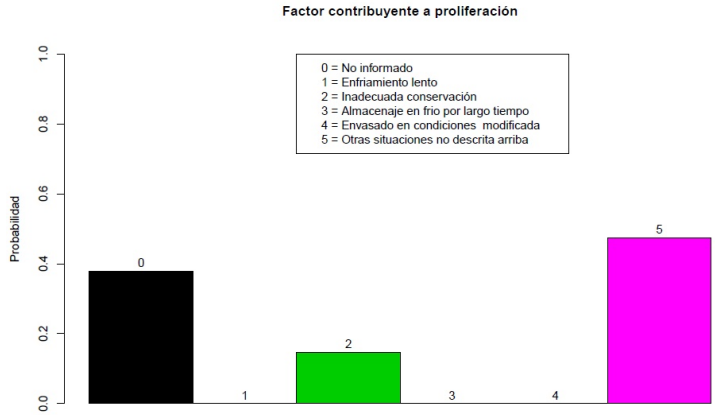


Figura 2.6: Factor contribuyente a la proliferación. Diagrama de barra

Finalmente, para el factor contribuyente a la supervivencia se muestra el mismo comportamiento que en la proliferación siendo más frecuente que no se identifique con 50,86 % o no se notifique con 25 %, seguido de insuficiente tiempo y/o temperatura durante calentamiento o cocción 18,10 % e insuficiente tiempo de descongelamiento y cocción 0,86 % (Figura 2.7). Es de resaltar que el factor de insuficiente acidificación no se presentaba en la base de datos estudiada(OPS, 2015a).

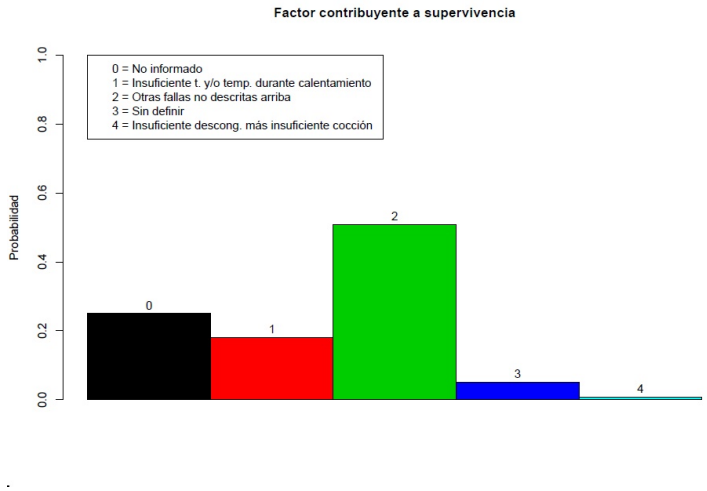


Figura 2.7: Factor contribuyente a la supervivencia. Diagrama de barra.

Además, nótese que el grupo del patógeno de salmonella es el más frecuente con 81,03 %, específicamente por la enfermedad de enteritis debido a salmonella que representa un 70.68 %, resaltado en rojo (Tabla 2.5), por mucha diferencia con las demás enfermedades tanto en su diagnostico por patógeno agrupado como por enfermedad específica (Figura 2.8).

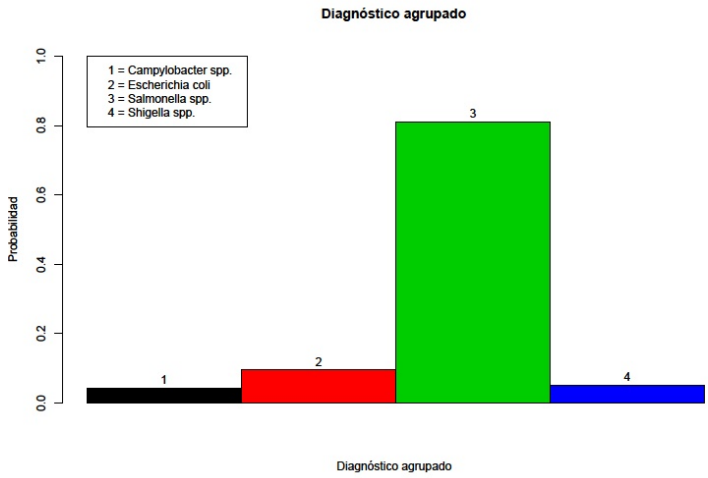


Figura 2.8: Diagnóstico agrupado. Diagrama de barra.

En la Tabla 2.5 también se observa que las menos frecuentes son la shigelosis de tipo no específica y la fiebre tifoidea de las cuales se presentaron un solo caso durante 2017, ambas resaltadas de color azul.

Tabla 2.5: Tabla de enfermedades CIE-10. Cantidad y porcentaje.

CIE.10	Cantidad	Porcentaje (%)
Enteritis debida a salmonella	82	70,69
Infección debida a salmonella no específica	11	9,48
Otras infecciones intestinales debidas a escherichia coli	5	4,31
Enteritis debida a campylobacter	5	4,31
Infección debida a escherichia coli enterotoxigena	2	1,72
Infección debida a escherichia coli enteropatogena	4	3,45
Shigelosis de tipo no específica	1	0,86
Shigelosis debida a shigella sonnei	2	1,72
Shigelosis debida a shigella flexneri	3	2,59
Fiebre tifoidea	1	0,86

Nótese además que la enteritis debido a salmonella es la enfermedad específica que tiene mayor porcentaje de brotes con una diferencia significativa con las demás de 61,21 % respecto a la más cercana que es la infección debida a salmonella no específica, la cual representa un 9,48 % de los brotes.

Tabla 2.7: Tabla resumen para variables de atención

Variable	Mínimo	Máximo	Desviación	Varianza
T.Amb	0	58	8,6455	74,7439
T.amb1	0	1	0,0929	0,0086
T.amb1-4	0	5	0,6605	0,4363
T.amb5-14	0	3	0,8072	0,6515
T.amb15-44	0	40	6,3669	40,5376
T.amb45-64	0	18	2,5645	6,5766
T.amb65	0	6	0,6623	0,4386
T.Hosp	0	5	0,9911	0,9823
T.hosp1	0	1	0,0929	0,0086
T.hosp1-4	0	1	0,2040	0,0416
T.hosp5-14	0	2	0,3852	0,1484
T.hosp15-44	0	4	0,6736	0,4537
T.hosp45-64	0	2	0,2586	0,0669
T.hosp65	0	1	0,1594	0,0254
T.s.a	0	37	3,5579	12,6588
T.s.a5-14	0	2	0,2615	0,0684
T.s.a15-44	0	29	2,7536	7,5822
T.s.a45-64	0	8	0,8087	0,6540
T.s.a65	0	1	0,0929	0,0086

Finalmente, al estudiar si se contempla inspección en los brotes, se observó que para la variable C.Inspec, la mayoría de los brotes no contemplan inspección en un 68,9 % de los casos.

2.3. Metodología

Se procedió a implementar el método de LASSO y Ridge con los paquetes de software libre R (R-foundation, s.f.), "glmnet"y "BGLR", para los enfoques de estadística clásica y Bayesiana respectivamente. Además, se definieron dos grupos, el primero con el 70 % de datos con los cuales realizar los ajustes a los modelos y un segundo grupo con el 30 % de los datos con los cuales realizar la comparación para los cálculos de predicciones realizados por cada brote. También se utilizó una grilla de 100 puntos para λ entre 10^{-2} y 10^{10}, la cual permite reducir el efecto del intercepto en el modelo, esto permite enfocarse en las variables explicativas.

Por otro lado, el periodo de incubación (P.Incubacion) y la duración del brote (Duracion.B) tienen desviaciones y varianzas pequeñas de alrededor de 1 o máximo 2 unidades, además estas variables tienen valores mínimos de 0 días (Tabla 2.6, azul). También, se puede observar que el valor mínimo de la tasa de ataque es de 0,46 % la cual es una tasa bastante baja (Tabla 2.6, azul).

Tabla 2.6: Tabla resumen para las variables: P.Incubacion, Duracion.B, Expuestos, Enfermos y Tasa.Ataque

Variable	Mínimo	Máximo	Desviación	Varianza
P.Incubacion	0	10	1,1626	1,3517
Duracion.B	0	11	1,4399	2,0734
Expuestos	2	658	89,8207	8067,7565
Enfermos	2	58	9,6301	92,7385
Tasa.Ataque	0,46	100	27,2845	744,4423

Las variables de atención fueron resumidas en la Tabla 2.7, observe que las variables con valores altos de desviaciones estándar y varianza, resaltados en rojo, son el total de personas atendidas en ambulatorios (T.Amb), las personas con edades entre 15 y 44 años (T.amb15-44) y entre 45 y 64 atendidos en este centro (Tamb45-64), también el total de personas que no recibieron atención médica (T.s.a) y aquellas con edades entre 15 y 44 años que cumplen con esta condición (T.s.a15-44).

Nótese además que las variables antes mencionadas también son aquellas que tienen los valores máximos más altos, estos son: 58 personas para el total atendido en ambulatorio, para las atendidas en este centro con edades entre 15 y 44 años se tiene 40 personas, 18 para aquellas que tienen entre 45 y 64 años, en el total de personas sin atención médica se encuentran 37 personas y para las personas en esta condición con edades entre 15 y 44 años hay 29 personas (Tabla 2.7, rojo).

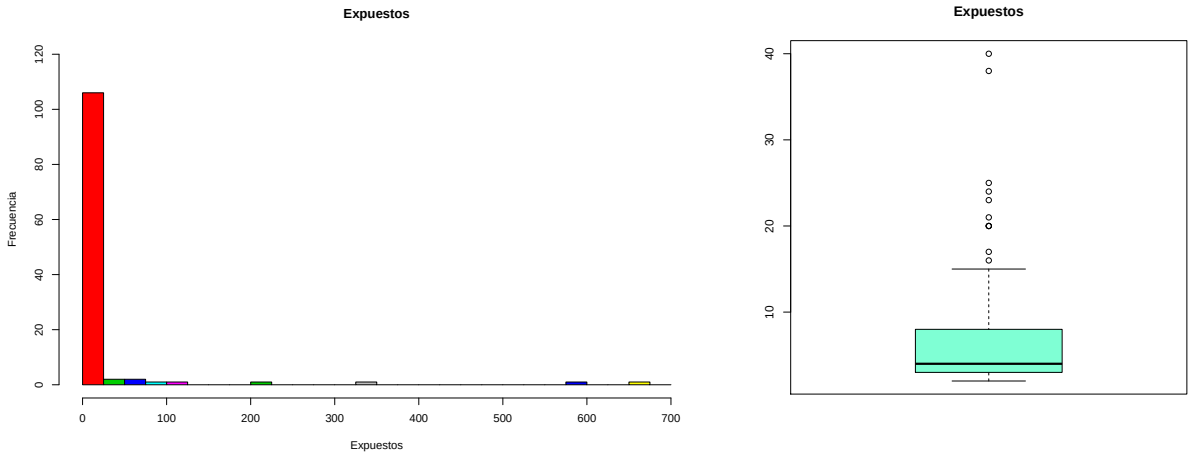
Además, hay que destacar que el grupo etario de 15 a 44 años se repite como el grupo con mayores valores máximos en todos los tipos de atención, incluyendo el caso de atendidos en hospitales donde el máximo para atención es 5 personas y para el grupo etario de 15 a 44 años es de 4 personas (Tabla 2.7, azul).

Además, hay mayor dispersión entre el segundo y tercer cuartil con 2 unidades mientras que entre el primer y segundo cuartil es de 1. Por otro lado, la media de los enfermos esta en 6,97 y de los expuestos 25.

Al observar con más detalle el comportamiento de las personas expuestas mostradas en la Figura 2.10(a), se encontró que entre 0 y 25 personas se encuentra la mayor parte de éstas con más de 100 casos, mientras que los puntos atípicos están ubicados entre 15 y 675 en 5 conjunto de datos disjuntos, en específico, el conjunto

$$\left\{(15, 125] \bigcup [200, 225] \bigcup [325, 350] \bigcup [575, 600] \bigcup [650, 675]\right\}.$$

Además, se observa el mismo comportamiento que en enfermos para la distribución de los datos, siendo mayor entre el segundo y tercer cuartil (Figura 2.10(b)).



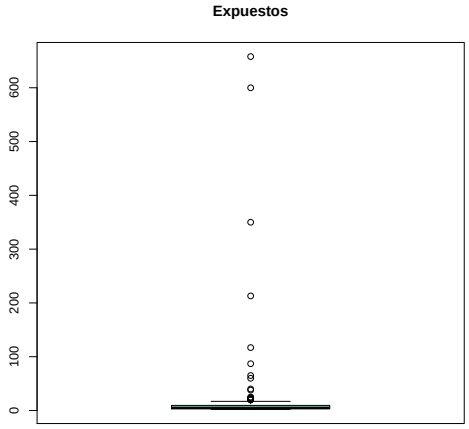
(a) Acercamiento al diagrama de barra de la cantidad de personas expuestas.

(b) Acercamiento al diagrama de caja de la cantidad de personas expuestas.

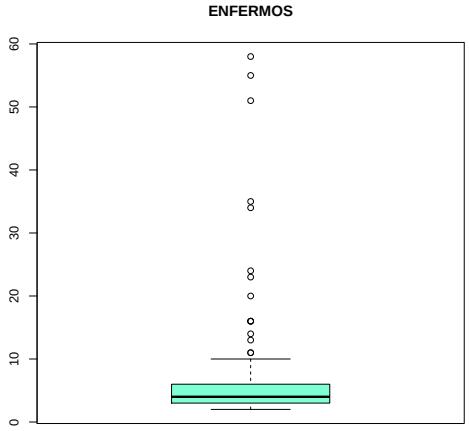
Figura 2.10: Estudio detallado de los pacientes expuestos. (a) Acercamiento al diagrama de barra y (b) Acercamiento al diagrama de caja.

En la Tabla 2.6 se puede observar resaltado en rojo que los valores de desviación estándar y varianza de la cantidad de expuestos son altos, al igual que los de la tasa de ataque (Tasa.Ataque) que es la proporción de las personas expuestas que enfermó, además nótese que el máximo de la tasa de ataque es del 100 % lo cual es un valor alto.

Por otro lado, se puede observar los diagramas de caja en la Figura 2.9 varios puntos atípicos en la variable respuesta enfermos, al igual que en la variable expuestos. Nótese que en los expuestos están concentrados entre 0 y 100 personas casi la totalidad de los datos incluyendo el bigote superior y algunos datos atípicos, pero el máximo de expuesto esta alrededor de 600 personas con dos datos atípicos (Figura 2.9(a)), los cuales son brotes confirmados que contemplaron inspección. Por otro lado, en los enfermos, el 75 % de los datos esta en el rango de 0 a 10 personas incluyendo el bigote superior por lo que los valores superiores a 10 personas se consideran atípicos (Figura 2.9(b)).



(a) Expuestos



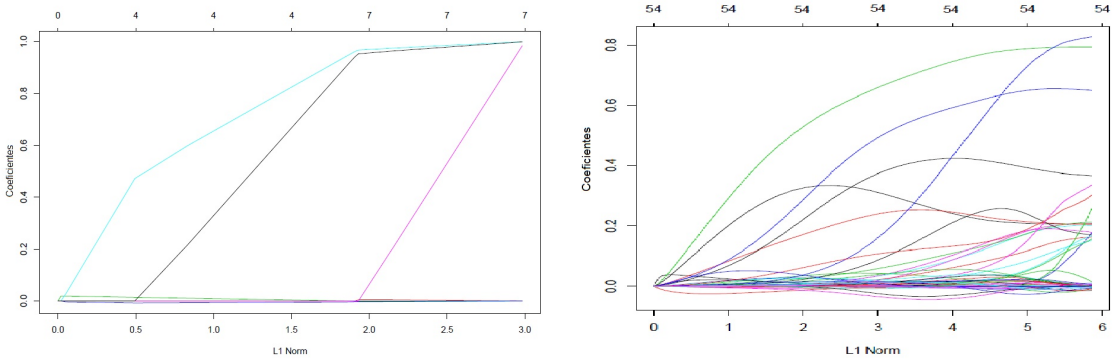
(b) Enfermos

Figura 2.9: Expuestos y enfermos. Diagramas de caja

En este trabajo se estudiaron 3 casos, el primero con todas las variables (con 54 variables explicativas) para observar el comportamiento de los grupos etarios y el tipo de atención, para el segundo caso se excluyeron las variables de atención, es decir, totales ambulatorios, hospitalarios y sin atención (T.Amb, T.Hosp, T.s.a, T.ambk,T.hospk y T.s.ak). El tercer caso es el estudio de salmonella sin variables de atención, por ser el patógeno más frecuente.

Los modelos clásicos entrenados con esta grilla presentaron los siguientes comportamientos en los coeficientes de cada variable y la norma de λ, mostrando como se van igualando a cero los coeficientes (eliminando las variables) según como varía el valor de dicha norma. Hay que resaltar la función glmnet trabaja con los datos estandarizados.

En el primer caso de estudio con todas las variables, se observa en la Figura 2.11 que cada coeficiente de una de éstas variables son representadas con una curva de un color distinto, al estudiar el modelo LASSO se encuentra que con una norma igual a 1 se puede notar que solo se tienen 4 variables con valores diferentes de 0, es decir, que no fueron eliminadas (Figura 2.11(a)). Además, se puede observar que a medida que la norma se hace más pequeña y se acerca a 0 se van eliminado estas variables, reduciéndose así su número en el modelo. En el modelo Ridge (Figura 2.11(b)) se pueden ver que son más las variables que faltan por eliminar (en específico 54 variables), donde la mayoría fluctúa cerca de 0 sin ser igual a éste. Por otro lado, nótese que en el modelo Ridge para ningún valor de la norma se iguala a 0 alguna variable.

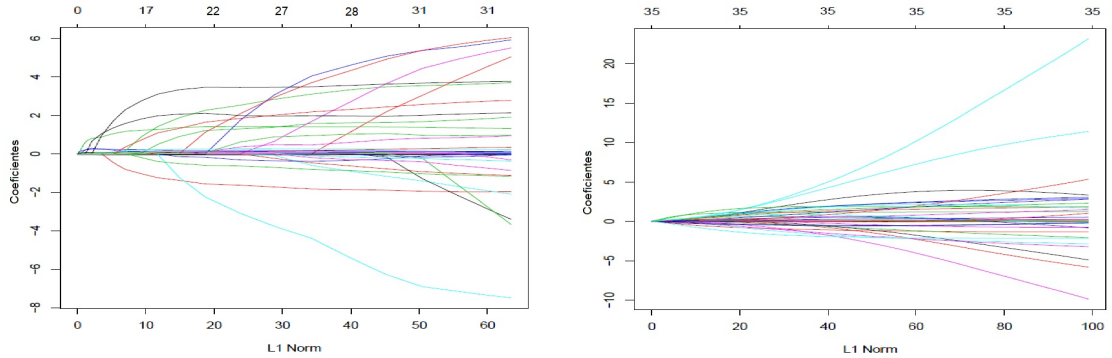


(a) Modelo LASSO clásico

(b) Modelo Ridge clásico

Figura 2.11: Coeficientes versus norma L1 de las variables totales usando el modelo Lasso clásico (a) y Ridge clásico (b). Caso 1: Todas las variables.

Para el segundo caso se estudian los datos sin incluir las variables totales de atención, reduciendo de esta manera las variables de 55 a 36. En la Figura 2.12 se presentan las variaciones en los coeficientes dependiendo de los valores de la norma de λ. En este caso en los modelos se observa un comportamiento casi lineal en el modelo Ridge (Figura 2.12(b)) de manera que los valores cercanos a 0 son más pequeños. Por otro lado, se observa la similitud al caso anterior en que ninguna de las variables su coeficiente es igualado a 0. Mientras que en la Figura 2.12(a), en el modelo LASSO, se encontró que a diferencia del anterior caso con todas las variables, existen más variables sin eliminar y en algunas sus coeficientes son negativos.

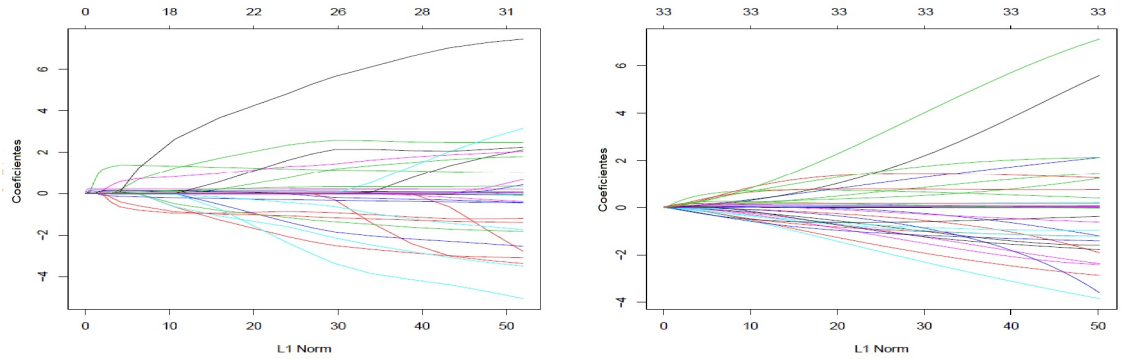


(a) Modelo LASSO clásico

(b) Modelo Ridge clásico

Figura 2.12: Coeficientes versus norma L1 de las variables totales usando el modelo Lasso clásico (a) y Ridge clásico (b). Caso 2: Sin variables de atención.

Para tercer caso con los modelos sin variables de atención y considerando solo los casos de salmonella se reducen los datos de 116 a 94 y se mantienen las 33 variables del caso anterior; en la Figura 2.13 se observa en este caso un comportamiento similar a los modelos anteriores, donde es casi lineal en el modelo Ridge (Figura 2.13(b)) sin eliminar las variables para ningún valor de la norma y el modelo LASSO (Figura 2.13(a)) tiene más variables a eliminar y en algunas sus coeficientes son negativos, pero en comparación con el segundo modelo LASSO solo se tiene dos variables cuyos coeficientes son en valor absoluto mayores a 4 para la norma igual a 50 y además cuando esta es igual a 10 se tiene una variable más sin eliminar, pasando de 17 a 18 variables.

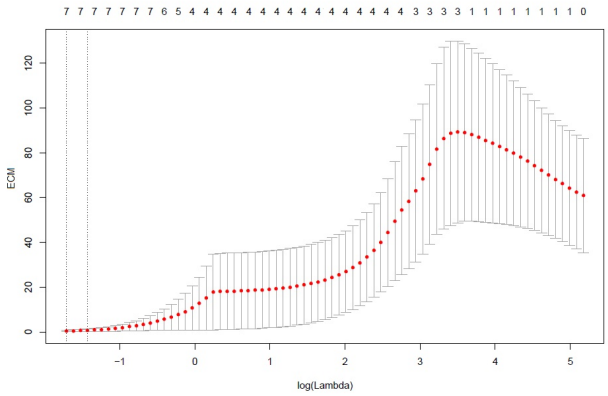


(a) Modelo LASSO clásico. Salmonella (b) Modelo Ridge clásico. Salmonella

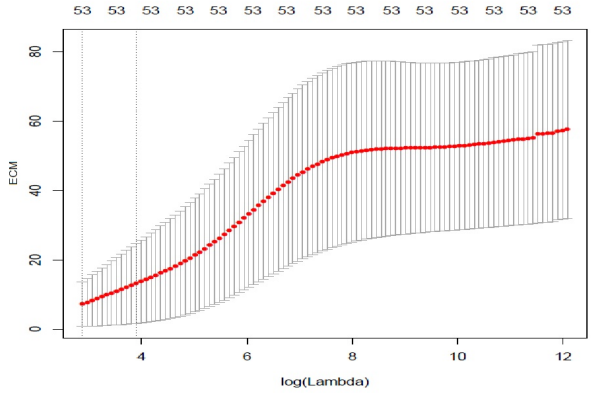
Figura 2.13: Coeficientes versus norma L1 de las variables totales usando el modelo Lasso clásico (a) y Ridge clásico (b). Caso 3: Solo salmonella y sin variables de atención.

Para los casos del enfoque clásico se realizaron una validación cruzada para cada uno, con la función "cv.glmnet"del software libre R con el fin de obtener el mejor valor de penalización, λ, para cada caso.

En el primer caso de estudio se obtuvo la Figura 2.14, para los intervalos de confianza del Error Cuadrático Medio (ECM) estimado para cada valor del $log(\lambda)$, donde el punto rojo representa el valor estimado de error. En la Figura 2.14(a), se observa que para el modelo LASSO el menor ECM esta para valores de $log(\lambda)$ menores a -1, mientras que en el caso de Ridge (Figura 2.14(b)) esto ocurre para valores de $log(\lambda)$ entre 2 y 4. Además, nótese que en el modelo LASSO el error aumenta para los valores de $log(\lambda)$ entre -1 y 3,5 pero para los valores superiores disminuye, mientras en el modelo Ridge se observa que para valores mayores de $log(\lambda)$ aumenta el error.



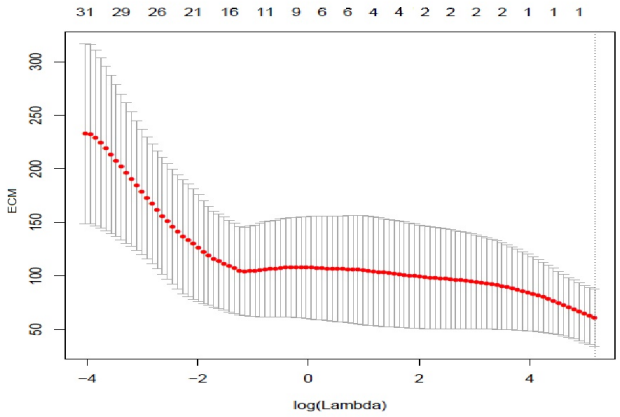
(a) Modelo LASSO clásico



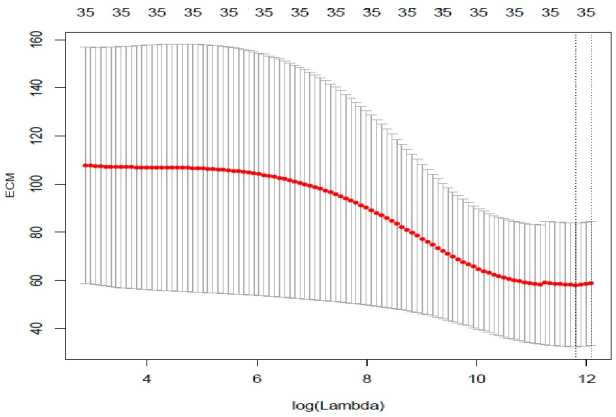
(b) Modelo Ridge clásico

Figura 2.14: Validación Cruzada. Caso 1: Todas las variables

Para el segundo caso sin variables totales de ambulatorio, hospitalización y sin atención, se tiene que el mejor λ se encuentra donde el $log(\lambda)$ es alrededor de 6 para el modelo LASSO (Figura 2.15(a)) y de 12 para el modelo Ridge (Figura 2.15(b)). Además, en este caso el error parece disminuir en ambos modelos a medida que $log(\lambda)$ aumenta, contrario a lo ocurrido en el caso anterior (Figura 2.15).



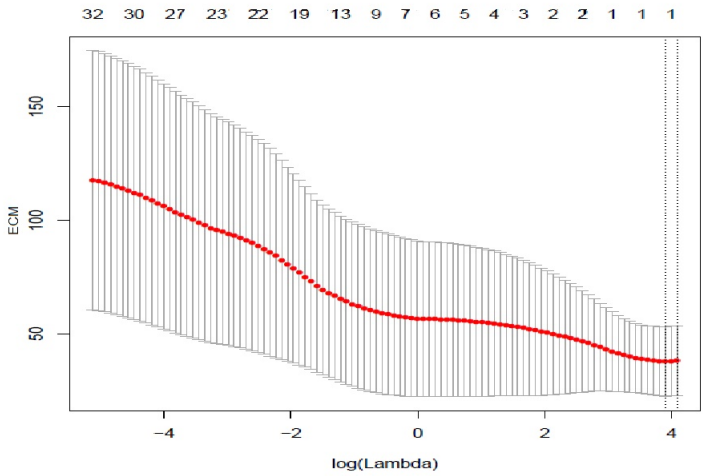
(a) Modelo LASSO clásico



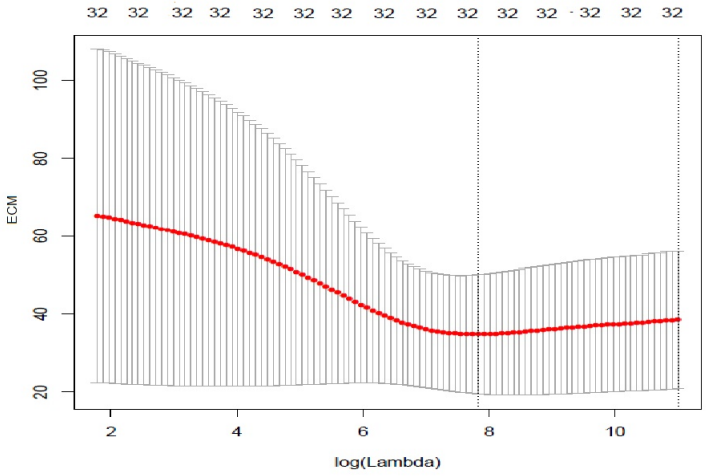
(b) Modelo Ridge clásico

Figura 2.15: Validación Cruzada. Caso 2: Sin variables de atención

En el tercer caso, solo de la enfermedades producidas por la salmonella y sin las variables de atención, se tiene que el mejor λ se encuentra ubicado donde el $log(\lambda)$ es alrededor de 4 para el modelo LASSO (Figura 2.16(a)) y entre 7 y 11 para el modelo Ridge (Figura 2.16(b)). Nótese que en este caso se encuentra un comportamiento similar al anterior donde el error disminuye a medida que aumenta $log(\lambda)$ en ambos modelos salvo por una leve aumento del error entre los valores de 8 a 11 para el modelo Ridge (Figura 2.16(b)).



(a) Modelo LASSO clásico



(b) Modelo Ridge clásico

Figura 2.16: Validación Cruzada. Caso 3: Solo salmonella y sin variables de atención

Por lo tanto, utilizando los resultados obtenidos en la validación cruzada se obtuvieron los siguientes mejores λ para cada caso minimizando el ECM del $log(\lambda)$:

- Caso 1. Todas las variables:
  Para Ridge $\lambda = 17,6875$ y LASSO $\lambda = 0,0181$
- Caso 2. Sin variables de atención:
  Para Ridge $\lambda = 133799,6000$ y LASSO $\lambda = 176,8752$
- Caso 3. Solo salmonella y sin variables de atención:
  Para Ridge $\lambda = 2516,4650$ y LASSO $\lambda = 49,3992$

Para el tercer modelo clásico se utilizó un modelo lineal generalizado (GLM) con vínculo Poisson debido a que los datos son de conteo y porque se tiene como variable respuesta la cantidad de enfermos que se modela mediante la distribución Poisson.

Para los modelos bajo el enfoque Bayesiano se estandarizó la matriz de diseño definida en el capítulo anterior, además, se utilizó la función BGLR para ajustar los modelos, para ello se llevaron a cabo 5000 iteraciones por modelo de las cuales se quemaron las primeras 2000.

En el modelo donde se utilizó una distribución a priori flat, la constante k a la cual es proporcional la priori en la ecuación 1.14 es 1, no se considera hiperparámetros.

Para el modelo de LASSO Bayesiano se empleó una distribución a priori doble exponencial con función de distribución para el hiperparámetro σ^2 definida como gamma y el parámetro de escala es asignado por una exponencial con tasa $\lambda^2/2$.

En el modelo Ridge Bayesiano se utilizo una priori normal multivariada con media 0 y varianza σ^2 cuya función de probabilidad esta dada por una χ^-2 con 5 grados de libertada por default y el parámetro de escala se resuelve para encajar con el R^2 del modelo.

Se considera que los hiperparametros deben ajustar un $R^2 = 0.5$ de la varianza de la varaible respuesta atribuida a los predictores lineales

CAPÍTULO III

ANÁLISIS DE LOS RESULTADOS

En el siguiente capítulo se analizan y comparan los resultados obtenidos para los modelos: LASSO, Ridge y GLM en el enfoque clásico, y LASSO, Ridge y flat en el análisis Bayesiano con respecto a información real y en los tres casos considerados anteriormente.

3.1. Predicciones

En esta sección se presentan los resultados obtenidos con los 3 modelos desde el enfoque clásico y los 3 modelos bajo el análisis Bayesiano, para los 3 casos de estudio.

3.1.1. Modelos clásicos con todas las variables

Para las predicciones clásicas se presentan los siguientes resultados en el caso con todas las variables para ajustar la distribución del número de enfermos por brote:

En el modelo LASSO en la Figura 3.1(a), utilizando el valor de penalización $\lambda = 0,0181$ descrito en el capítulo anterior para este caso, se observa que la relación entre los datos predichos y los observados se logro casi una perfecta alineación entre ambos valores (círculos negros), siendo la referencia visual la línea azul sobre la igualdad de ambos valores.

En el modelo de Ridge en la Figura 3.1(b), se utilizó $\lambda = 17,6875$ y se observa en la gráfica de predicciones versus valores reales que pocos valores difieren de los reales y se aglomeran alrededor del cero.

En el modelo de GLM en la Figura 3.1(c), se utilizó un modelo lineal generalizado (GLM) con vínculo Poisson, nótese, que en la gráfica se observa que hay mayor dispersión de valores que difieren de los reales y se aglomeran alrededor del cero comparada con los dos modelos anteriores.

Nótese que meteorismo tiene el segundo porcentaje más alto de respuesta negativa y espasmos el tercero (Tabla 2.2), además el total de las personas sin atención y el total de estos con edades entre 15 y 44 años tienen el tercer y cuarto lugar de los valores más altos de máximos, desviación estándar y varianza, respectivamente (Tabla 2.7).

3.1.2. Modelos clásicos sin incluir las variables de atención

Para el segundo caso con los modelos sin considerar las variables de atención se encontraron las siguientes predicciones:

Para el modelo LASSO se utilizó como valor de penalización $\lambda = 176,8752$. En la Figura 3.2(a), se observa a diferencia del caso anterior un comportamiento lineal vertical con los valores predichos cercanos a 5 lo que implica no hay un buen ajuste.

personas sin atención entre 15 y 44 años (T.s.a15.44) y total de pacientes ambulatorios con edades entre 15 y 44 años (T.amb15.44), son las que tienen mayor peso y en este orden, de color azul en la columna Ridge. Mientras para el modelo GLM las variables de mayor peso son meteorismo, intercepto, T.s.a15.44, T.s.a y total de pacientes hospitalarios con edades entre 1 y 4 años (T.hosp1.4), respectivamente en orden, de color azul en dicha columna.

Tabla 3.3: Coeficientes $\boldsymbol{\beta}$ con mayor peso en Ridge y GLM clásicos. Caso 1: Todas las variables.

Variables	LASSO	Ridge	GLM
T.Amb	0,9962	0,6547	-0,0033
Intercepto	0,1042	0,5708	1,3684
T.s.a	0,9801	0,4853	0,7466
T.s.a15.44	0	0,3684	-0,7716
T.amb15.44	0	0,3139	0,1040
T.hosp1.4	0	0,0081	-0,6509
Meteorismo	0	0,0016	-1,6763

De este modo se observa que se repite el peso de las variables donde no hubo atención a los pacientes y donde la atención no se le suministró a las personas entre 15 y 44 años. Se puede observar que este grupo etario de 15 a 44 años también aparece entre las 5 variables con mayor peso en su coeficiente en el modelo Ridge pero esta vez con la población que fue atendida en ambulatorios, por lo que esta es una población resaltante.

Al estudiar los $p-valores$ en el modelo GLM (Tabla 3.4), se observa que la variable meteorismo es significativa a nivel $\alpha = 0.001$ y las variables presencia de espasmos, el total de personas sin atención médica y total de personas entre 15 y 44 años que no tuvieron atención médica son significativas a nivel $\alpha = 0.05$ (Tabla 3.23, Anexo 6).

Tabla 3.4: Tabla de $p-valores$ modelo GLM clásico. Caso 1: Todas las variables.

Variables	$P-valor$	Significancia
T.s.a	0,0545	.
T.s.a15.44	0,0685	.
Espasmos	0,0522	.
Meteorismo	0,0020	**

Tabla 3.1: Bondad de ajuste modelos clásicos. Caso 1: Todas las variables.

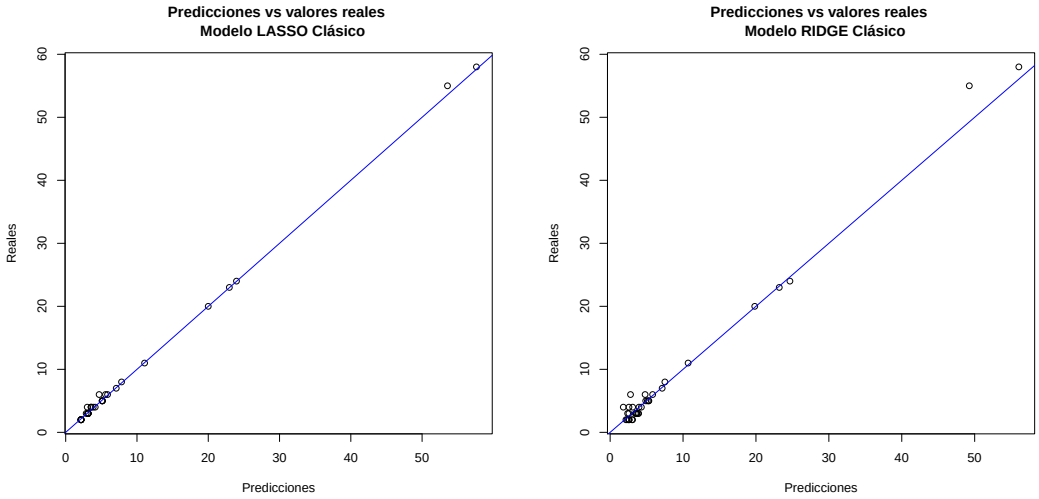
Método	LASSO	Ridge	GLM
EAM	8,8513	29,2022	99,4136
ECM	0,1691	1,8242	1,3964
R^2	0,9992	0,9926	0,9850

En cuanto a los valores del parámetro $\boldsymbol{\beta}$ para cada modelo, los cuales fueron descritos en las ecuaciones (1.2), (1.5) y (1.7) (Tabla 3.20, Anexo 3). A continuación en la Tabla 3.2 se muestran los valores de las variables que no se eliminaron en el modelo LASSO, nótese que las variables con mayor peso son los totales por atención lo cual tiene sentido, dado que en ellos se almacena el 100 % de los casos, además, de estas la que tiene menor peso es la atención en hospitales. Lo interesante es que las otras variables que no se eliminaron fueron el intercepto, la región de notificación, la tasa de ataque y cantidad de expuestos, todas resaltadas con azul, note la única con efecto negativo es la tasa de ataque (Tabla 3.2). Además, se debe resaltar que se eliminaron del modelo de GLM las variables: Total de pacientes atendidos en ambulatorios mayores de 65 años (T.amb65), total de pacientes atendidos en ambulatorios menores de 1 años (T.Hosp.1) y total de personas sin atención médica mayores de 65 años (T.s.a65) debido a que nunca eran significativas ($P-valor = 1$).

Tabla 3.2: Coeficientes $\boldsymbol{\beta}$ de las variables no eliminadas por LASSO clásico. Caso 1: Todas las variables.

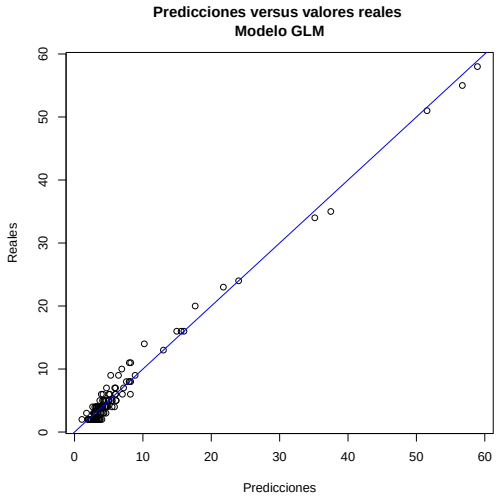
Variables	LASSO	Ridge	GLM
T.Amb	0,9962	0,6547	-0,0033
T.s.a	0,9801	0,4853	0,7466
T.Hosp	0,8013	0,1497	0,5247
Intercepto	0,1042	0,5708	1,3684
R.notificación	0,0023	0,0172	-0,0837
Expuestos	0,0006	0,0046	-0,0011
Tasa.Ataque	-0,0001	-0,0035	-0,0041

Por otro lado, al estudiar los coeficiente con mayor valor absoluto en los modelos Ridge y GLM, y extraer las 5 variables cuyo coeficiente cumple esta condición se presenta en la Tabla 3.3. De aquí, se tiene que para el modelo Ridge las variables: Total de pacientes en ambulatorios (T.Amb), intercepto, total de personas sin atención médica (T.s.a), total de



(a) Modelo LASSO clásico

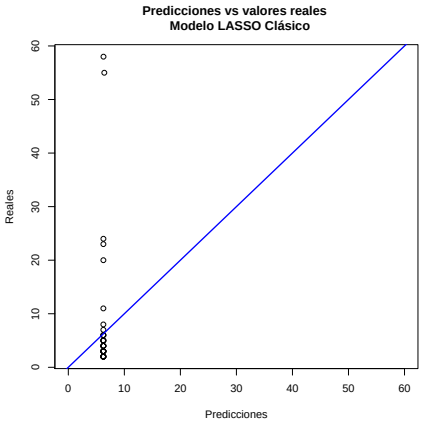
(b) Modelo Ridge clásico



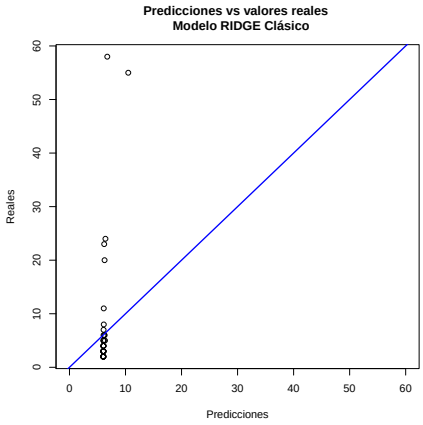
(c) Modelo GLM

Figura 3.1: Predicciones versus valores reales de los modelos clásicos. Caso 1: Todas las variables.(a) Modelo LASSO Clásico ,(b) Modelo Ridge Clásico y (c) Modelo GLM.

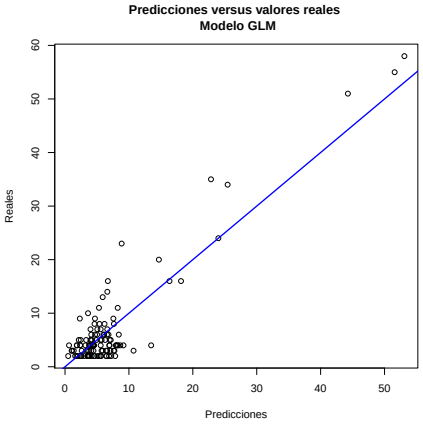
Ahora al observar las medidas de bondad de ajuste para este caso como se presenta en la Tabla 3.1, se tiene que según el método Error Absoluto Medio (EAM) y para el método Error Cuadrático Medio (ECM) el modelo LASSO es el mejor por un amplio margen, mientras en el método R^2 el modelo de LASSO es el mejor pero no hay una diferencia significativa con los demás modelos. De lo anterior, se tiene que para el caso con todas las variables y bajo el enfoque clásico se considera que **el mejor modelo es el modelo LASSO**.



(a) Modelo LASSO clásico



(b) Modelo Ridge clásico



(c) Modelo GLM

Figura 3.2: Predicciones versus valores reales de los modelos clásicos. Caso 2: Sin las variables de atención. (a) Modelo LASSO Clásico ,(b) Modelo Ridge Clásico y (c) Modelo GLM.

El modelo Ridge utiliza $\lambda = 133799,6000$ como valor de penalización. Se observa en la Figura 3.2(b) un comportamiento similar al obtenido en el modelo de LASSO para este segundo caso, es decir, un comportamiento lineal vertical con los valores predichos cercanos a 6.

En el modelo GLM se realizó nuevamente un modelo lineal generalizado con vínculo Poisson por cumplirse las mismas condiciones expuestas para el caso 1. En la Figura 3.2(c) se observa un comportamiento similar al obtenido en los modelos del caso 1 con todas las variables, con una dispersión de los puntos en los alrededores del 5.

Ahora al observar las medidas de bondad de ajuste para este caso (Tabla 3.5), se tiene que según el método EAM el modelo Ridge es el mejor pero por poca diferencia con respecto a los demás modelos, para el método ECM y por el método R^2 el modelo GLM es el mejor con una diferencia significativa sobre los demás modelos. A continuación en la Tabla 3.5 se observa en azul los mejores modelos por métodos.

Tabla 3.5: Bondad de ajuste modelos clásicos. Caso 2: Sin variables de atención.

Método	LASSO	Ridge	GLM
EAM	228,5988	219,6595	310,7356
ECM	173,9398	160,9044	13,7288
R^2	0,5250	0,5418	0,8572

De aquí se tiene que bajo el enfoque clásico para el segundo caso, sin las variables de atención, se considera que **el mejor modelo es el modelo GLM**.

En cuanto a los valores de $\boldsymbol{\beta}$ para cada modelo, los cuales fueron descritos en las ecuaciones (1.2), (1.5) y (1.7) (Tabla 3.21, Anexo 4). A continuación en la Tabla 3.6 se muestran los valores de los coeficientes de las variables que no se eliminaron en el modelo LASSO, nótese que las variables con mayor peso para el modelo LASSO son el intercepto y la cantidad de expuestos, lo cual se mantiene del caso con todas las variables.

Tabla 3.6: Coeficientes $\boldsymbol{\beta}$ de las variables no eliminadas por LASSO clásico. Caso 2: Sin variables de atención.

Variables	LASSO	Ridge	GLM
Intercepto	6,2539	6,8386	0,6372
Expuestos	0,0288	0,0182	0,0030

Al estudiar los coeficientes del modelo Ridge, se tiene que entre las 5 variables cuyos coeficientes en valor absoluto son mayores se encuentran el intercepto, la cantidad de expuestos, la tasa de ataque (Tasa.Ataque), la semana (Semana) y las enfermedades (CIE-10). Mientras que para el modelo GLM se tienen las variables diarrea, meteorismo, la coincidencia de regiones de notificación y consumo (R.N-RC), el rush cutáneo (Rush.Cutaneo) y el intercepto (Tabla 3.7).

Tabla 3.7: Coeficientes $\boldsymbol{\beta}$ con mayor peso en Ridge y GLM clásico. Caso 2: Sin las variables de atención.

Variables	LASSO	Ridge	GLM
Intercepto	6,2539	6,8386	0,6372
Expuestos	0,0288	0,0182	0,0030
Diarrea	0	0	1,8198
Meteorismo	0	0	0,9606
R.N-RC	0	0	-0,8716
Rush.Cutaneo	0	0	0,7814
CIE.10	0	0,0014	0,0072
Tasa.Ataque	0	-0,0045	-0,0036
Semana	0	0,0017	0,0015

De aquí, se puede observar que el modelos que tiene mayor ajuste (el modelo GLM) prioriza las variables sintomáticas, mientras que en el modelo Ridge se le otorga prioridad a variables más globales como la semana, la cantidad de expuestos y la enfermedad. Además, se puede observar que las únicas variables cuyo coeficiente tiene efecto negativo son la tasa de ataque y si hay coincidencia entre las regiones de notificación y consumo.

Al observar el $p - valor$ en el modelo GLM el nivel de significancia (Tabla 3.8), se tiene que siempre son significativas las variables duración del brote (Duración.B), región de notificación (R.notificación), región de consumo (R.consumo), expuestos, diarrea

y factor contribuyente a la supervivencia (F.Supervivencia), resaltados de azul; luego a nivel $\alpha = 0.001$ tenemos las variables periodo de incubación (P.Incubación), coincidencia de las regiones (R.N-RC), vómitos, el rush cutáneo, la cefalea, el meteorismo y otros síntomas (Otros).

Tabla 3.8: Tabla de $p - valores$ modelo GLM clásico. Caso 2: Sin las variables de atención.

Variables	$P - valor$	Significancia
Dolores	0,0272	*
Hipotensión	0,0277	*
Mialgia	0,0332	*
P.Incubación	0,0012	**
R.N-RC	0,0039	**
Vómitos	0,0061	**
Rush.Cutaneo	0,0041	**
Cefalea	0,0015	**
Meteorismo	0,0020	**
Otros	0,0023	**
Duración.B	0	***
R.notificación	0	***
R.consumo	0,0001	***
Expuestos	0	***
Diarrea	0	***
F.Supervivencia	0	***

Finalmente, a nivel 90 % de significancia se encuentran incluidas las variables dolores, hipotensión y mialgia (Tabla 3.24, Anexo 7).

3.1.3. Modelos clásicos sin incluir las variables de atención y enfocado en la salmonella

En el tercer caso se tienen los modelos para salmonella y sin variables de atención, para estos modelos se encontraron los siguientes resultados:

En la Figura 3.3(a) en el modelo LASSO utilizando el valor de penalización $\lambda = 49,3992$, se observa se mantiene el comportamiento del modelo en el segundo caso sin las variables totales, donde se veía linealidad vertical alrededor del 6 para los valores predichos.

Tabla 3.12: Coeficientes $\boldsymbol{\beta}$ con mayor peso en GLM clásico. Caso 3: Solo salmonella y sin variables de atención.

Variables	LASSO	Ridge	GLM
Intercepto	5,2004	5,3623	2,5773
Meteorismo	0	-0,0001	-1,1527
Parestesias	0	-0,0001	-1,0925
Mialgia	0	0,0008	0,8432
Rush Cutáneo	0	0,0002	0,7826
Otros Neurológicos	0	-0,0001	-0,6037
Dolores	0	-0,0003	-0,4552
Espasmos	0	0,0003	0,4428
Nauseas	0	-0,0004	-0,4201
C.Inspe	0	0,0010	0,3981

Nótese, que según el $p - valor$ en el modelo GLM (Tabla 3.13), se tiene que siempre es significativa la mialgia, después con un nivel de significancia de $\alpha = 0.001$ se tienen las variables de los factores de supervivencia, el rush cutáneo, la nauseas y la cantidad de expuestos. Luego, con $\alpha = 0.01$, las variables de el CIE-10, la fiebre, los dolores y el intercepto son significativas. Finalmente, para $\alpha = 0.05$ se tienen que el periodo de incubación, el proceso de pérdida de inocuidad (P.P.I) y la contemplación de inspección son significativas (C.Inspec) (Tabla 3.25, Anexo 8). Por otro lado, observe que las variables dolores, nauseas, rush cutáneo, mialgia y la contemplación de inspección se presentan tanto en las variables significativas a nivel 95 % como en las 10 con coeficientes más pesados en el mejor modelo considerado para este caso.

se encuentran entre las 5 primeras. Por otro lado, se puede observar que las regiones de notificación tomaron mayor peso en el caso de salmonella que el CIE-10.

Tabla 3.11: Coeficientes $\boldsymbol{\beta}$ con mayor peso en el modelo de Ridge clásico. Caso 3: Solo salmonella y sin variables de atención.

Variables	LASSO	Ridge	GLM
Intercepto	5,2004	5,3623	2,5773
Expuestos	0,0240	0,0416	0,0086
Semana	0	0,0128	0,0058
Tasa.Ataque	0	-0,0096	-0,0013
R.notificación	0	0,0072	0
R.consumo	0	0,0072	0
CIE.10	0	0,0043	0,0244
Duración.B	0	0,0041	0,0544
F.Contaminacion	0	0,0023	-0,0027
L.consumo	0	0,0016	-0,0008

Además, nótese que la única variable cuyo efecto es negativo entre las 10 más pesadas variables del modelo de Ridge para salmonella es la tasa de ataque comportamiento que se mantiene de los casos anteriores.

Para el modelo GLM se extrajeron las 10 variables cuyo coeficiente en valor absoluto es mayor (Tabla 3.12), se encontró que nuevamente aparecen meteorismo y rush cutáneo como en el caso anterior entre los 5 primeros y el último de estos se mantiene desde el primer caso estudiado. Además, se mantiene la predominancia de las variables sintomáticas sobre las demás tal como se expuso en el segundo caso. Por otro lado, se puede observar que las variables de mialgia, el rush cutáneo, los espasmos, la contemplación de inspección y el intercepto son las únicas variables de este grupo que tienen efecto positivo en el modelo.

Ahora, al observar las medidas de bondad de ajuste para este caso (Tabla 3.9), se encontró que según el método EAM el modelo Ridge es el mejor pero por poca diferencia con respecto a los demás modelos, para el método ECM y por el método R^2 el modelo GLM es el mejor con una diferencia significativa con respecto a los demás modelos. De lo anterior, se tiene que se considera que **el mejor modelo es el modelo GLM**.

Tabla 3.9: Bondad de ajuste modelos clásicos. Caso 3: Solo salmonella y sin variables de atención.

Método	LASSO	Ridge	GLM
EAM	2,5121	2,8421	1,9333
ECM	8,6755	17,7806	8,3278
R^2	0,4262	0,4301	0,7336

Además, se puede notar que el comportamiento de Ridge y LASSO son similares en los dos últimos casos expuestos.

Nótese que al extraer las variables no eliminadas por el modelo de LASSO (Tabla 3.10), se tiene que las variables con mayor peso son el intercepto y la cantidad de expuestos, lo cual se mantiene de los casos anteriores.

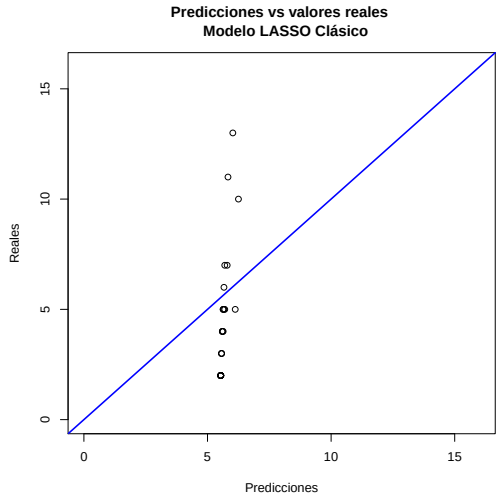
Tabla 3.10: Coeficientes $\boldsymbol{\beta}$ de las variables no eliminadas por LASSO clásico. Caso 3: Solo salmonella y sin variables de atención.

Variables	LASSO	Ridge	GLM
Intercepto	5,2004	5,3623	2,5773
Expuestos	0,0240	0,0416	0,0086

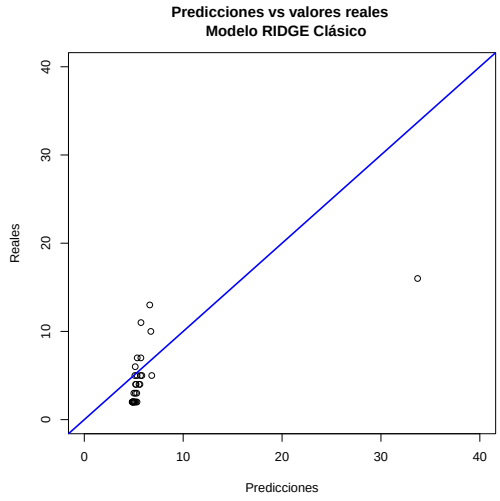
Por otro lado, se puede observar que las variables que fueron eliminadas en los modelos Ridge y GLM fueron: la coincidencia de la región de consumo y notificación, diagnóstico agrupado (d.agrupado) y solo en el caso de GLM la región de consumo (Tabla 3.22, Anexo 5).

Observe que cuando se extraen las 10 variables cuyo coeficiente en valor absoluto es mayor en el modelo Ridge (Tabla 3.11), se tiene que la cantidad de expuestos, la semana, la tasa de ataque y las enfermedades (CIE-10) vuelven a aparecer al igual que en caso anterior, aunque a excepción de CIE-10, todas las variables mencionadas anteriormente

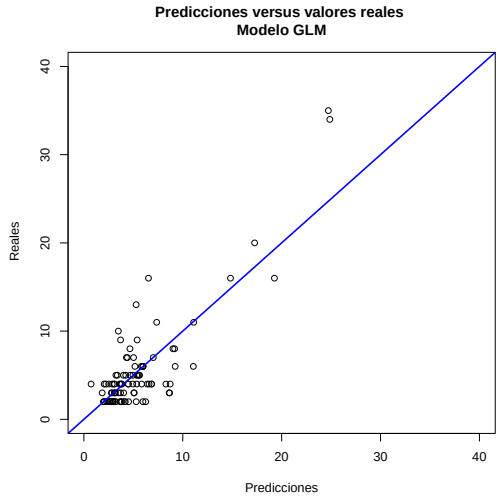
Para el modelo Ridge en la Figura 3.3(b) utilizando como valor de penalización $\lambda = 2516,4650$, se observa un comportamiento similar al encontrado en el modelo LASSO para este caso con la diferencia que se observa un dato mayor a 30 en los valores predichos.



(a) Modelo LASSO clásico



(b) Modelo Ridge clásico



(c) Modelo GLM

Figura 3.3: Predicciones versus valores reales de los modelos clásicos. Caso 3: Solo salmonella y sin las variables de atención.(a) Modelo LASSO Clásico ,(b) Modelo Ridge Clásico y (c) Modelo GLM.

En la Figura 3.3(c) en el modelo lineal generalizado (GLM) con vínculo Poisson. Se observa que se mantiene un comportamiento con dispersión focalizada en los alrededores de 5 en su mayoría similar a lo observado en los 2 casos expuestos anteriormente.

Tabla 3.13: Tabla de $p - valores$ modelo GLM clásico. Caso 3: Solo salmonella y sin las variables de atención.

Variables	$P - valor$	Significancia
Intercepto	0,0199	*
Dolores	0,0345	*
Fiebre	0,0152	*
CIE.10	0,0351	*
Expuestos	0,0080	**
Nauseas	0,0026	**
Rush.Cutaneo	0,0089	**
F.Supervivencia	0,0060	**
Mialgia	0,0003	***
P.Incubación	0,0540	.
P.P.I	0,0710	.
C.Inspec	0,0528	.

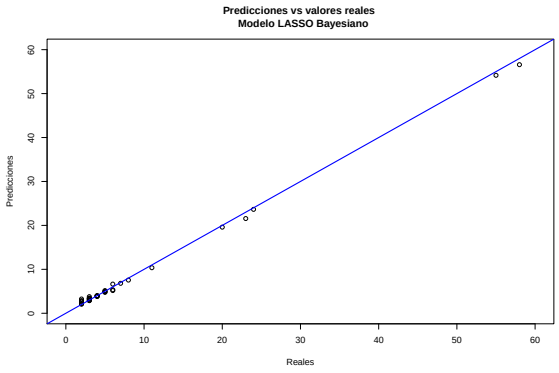
3.1.4. Modelos Bayesianos con todas las variables

A continuación se presentan los 3 casos anteriores pero desde el enfoque Bayesiano. En el primer caso que se tiene es aquel con todas las variables explicativas, al observar las predicciones de las medias estimadas a posterior versus los valores reales de los modelos de LASSO, Ridge y utilizando una distribución a priori flat.

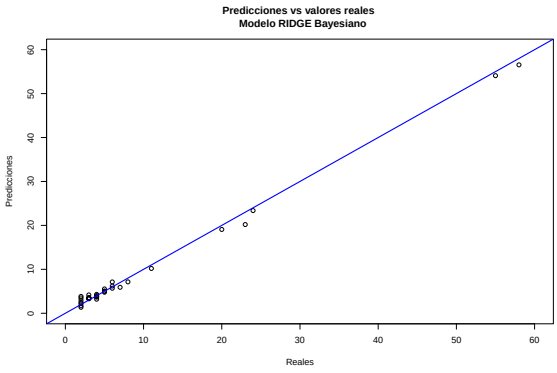
Observe que en el modelo LASSO en la Figura 3.4(a), se valida la linealidad entre los datos predichos y reales, es decir, los círculos negros están casi por completo en la línea azul lo que significa que la mayoría es igual al valor real.

Al observar el modelo Ridge en la Figura 3.4(b) se obtiene un comportamiento donde la mayoría de los círculos están alrededor de la línea azul pero no son iguales a esta, por lo cual el modelo difiere en mayor medida que el modelo LASSO.

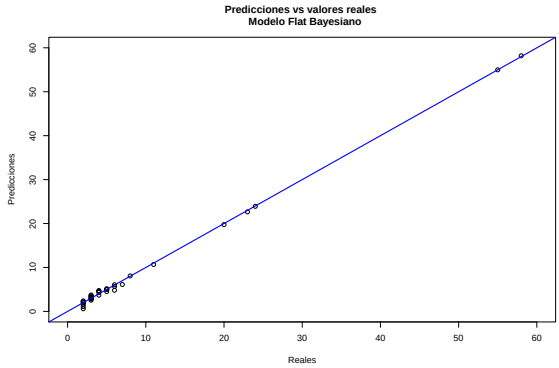
Para el tercer modelo se utilizó una distibución a priori flat o aplanada; para este se observa que en la Figura 3.4(c) la mayoría de las observaciones están sobre la línea y más cerca de la misma en comparación con los anteriores modelos.



(a) Modelo LASSO Bayesiano



(b) Modelo Ridge Bayesiano



(c) Modelo flat Bayesiano

Figura 3.4: Predicciones versus valores reales de los modelos desde el enfoque Bayesiano. Caso 1: Todas las variables.(a) Modelo LASSO Bayesiano ,(b) Modelo Ridge Bayesiano y (c) Modelo flat Bayesiano.

Nótese en la Tabla 3.14 al estudiar la bondad de ajuste de los modelos se tiene que según los métodos EAM y ECM el mejor modelo es el obtenido usando la priori flat o aplanada por una diferencia poco significativa con respecto a los demás modelos, en especial al modelo LASSO; mientras en el método R^2 los tres modelos son buenos ajustándose pero el mejor de todos por poco es el modelo LASSO. De lo anterior expuesto se considera que **el mejor modelo es aquel que utiliza una distribución a priori flat**.

Tabla 3.14: Bondad de ajuste modelos Bayesianos. Caso 1: Todas las variables.

Método	Ridge	LASSO	flat
EAM	0,6754	0,4328	0,3771
ECM	0,7801	0,3396	0,2538
R^2	0,9966	0,9988	0,9985

Se observa que ningún modelo Bayesiano realizó selección de variable puesto que no se igualó el coeficiente de ninguna variable a 0 (Tabla 3.26, Anexo 9).

Luego, al extraer las 6 primeras variables cuyos coeficientes en valor absoluto son mayores para cada modelo se obtuvo la Tabla 3.15. Al observar los valores absolutos de los coeficientes $\boldsymbol{\beta}$ para el modelo obtenido usando la priori flat descrita en la ecuación (1.13), se obtuvo que las variables con mayor peso son los totales de atención en hospitales y ambulatorios, el total de niños menores de 1 año atendidos en hospitales, el total de persona que no tuvieron atención médica y se encuentran entre las edades de los 5 a 14 años, la presencia de otros síntomas neurológicos y la contemplación de inspección del brote. Nótese que ninguna de estas variables se encontraba entre las 5 primeras de este modelo. Además, las únicas de estas variables cuyo efecto sobre el modelo es negativo son el total de personas atendidas en ambulatorios y el total de niños menores de 1 año atendidos en hospitales.

Tabla 3.15: Coeficientes $\boldsymbol{\beta}$ con mayor peso en LASSO, Ridge y flat Bayesianos. Caso 1: Todas las variables.

Variables	LASSO	Ridge	flat
Expuestos	0,4398	0,8358	60,3373
T.amb15.44	1,9858	2,6165	30,7723
Meteorismo	0,3948	1,0181	23,0169
T.amb65	0,3065	0,5954	8,1566
T.amb45.64	0,7669	1,4341	-0,9758
T.s.a5.14	0,0092	0,0564	123,5362
Otros.Neurológicos	0,0234	0,0314	122,3317
T.hosp.1	0,0845	-0,0519	-91,2921
C.Inspec	0,0218	0,1659	79,4991
T.Hosp	0,2739	0,1823	73,6739
T.Amb	4,8706	2,3410	-66,6235

Para el modelo LASSO bajo el enfoque Bayesiano los valores absolutos de los coeficientes $\boldsymbol{\beta}$ descritos en la ecuación (1.15), se tiene que las variables con mayor peso son los totales de personas atendidas en ambulatorio, el total de personas entre 15 y 44 años, 45 y 64 años, y mayores de 65 años que fueron atendidos en ambulatorios, la presencia de meteorismo y la cantidad de expuestos (Tabla 3.15). Nótese que de las variables anteriores solo el total de personas atendidas en ambulatorios y la cantidad de expuestos aparecen tanto en el modelo LASSO en el enfoque clásico y Bayesiano.

Finalmente, se observa en la Tabla 3.15 que para el caso de los coeficientes del modelo Ridge descritos en la ecuación (1.16) para el caso Bayesiano se obtuvieron las mismas variables del modelo LASSO con valores diferentes en ambos modelos. Nótese que las variables total de personas atendidas en ambulatorios y el total de personas entre 15 y 44 años atendidas en estos centros de salud, aparecen tanto en el modelo bajo el enfoque clásico y Bayesiano. En cuanto a los efectos de las variables en los modelos de LASSO y Ridge se tiene que todos los efectos son positivos para los modelo.

Nótese que se repite la variable de total de personas atendidas en ambulatorios en los 3 modelos estudiados en este caso por lo que es una variable que debe ser observada con más atención en futuros estudios. Además, hay que resaltar la recurrencia del grupo etario de 15 a 44 años tanto en el caso clásico como bajo el enfoque Bayesiano.

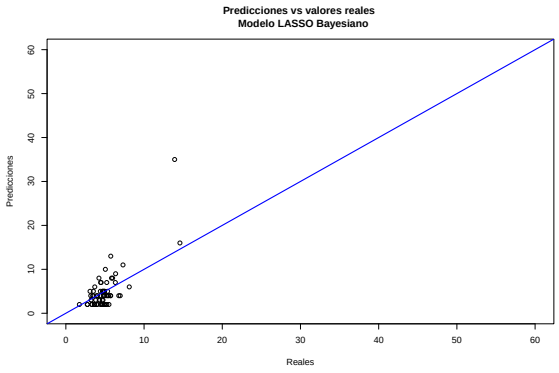
3.1.5. Modelos Bayesianos sin las variables de atención

En el segundo caso se excluyeron las variables de atención, es decir, totales de atención ambulatorio, atención hospitalaria y sin atención junto con sus subvariables divididas en grupos etarios y se realizaron los modelos bajo el enfoque Bayesiano.

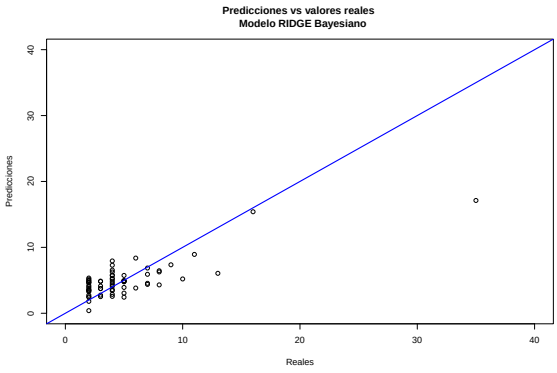
Al observar la gráfica de los valores predichos con los reales en el modelo LASSO Bayesiano en la Figura 3.5(a) se observa los datos predichos están entre 0 y 10 y hay una alta concentración entre los valores 0 y 10 de los reales, y la mayoría no se encuentra sobre la línea azul.

En el modelo Ridge Bayesiano en la Figura 3.5(b) se observa los valores predichos están ubicados de forma similar al modelo LASSO, es decir, la mayoría de los puntos se concentra entre 0 y 10 de los reales sin ubicarse sobre la línea azul.

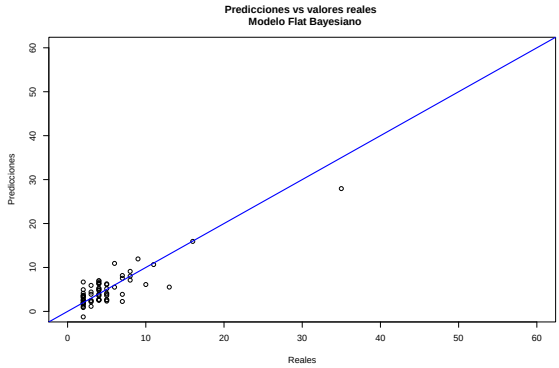
Utilizando una distribución a priori flat se observa en la Figura 3.5(c) una mayor concentración alrededor de la línea azul, en especial en los alrededores de 0 y 10 de los predichos y observados, por lo que parece ajustarse más que los modelos anteriores.



(a) Modelo LASSO Bayesiano



(b) Modelo Ridge Bayesiano



(c) Modelo flat Bayesiano

Figura 3.6: Predicciones versus valores reales de los modelos desde el enfoque Bayesiano. Caso 3: Solo salmonella y sin las variables de atención.(a) Modelo LASSO Bayesiano ,(b) Modelo Ridge Bayesiano y (c) Modelo flat Bayesiano.

Al analizar la tabla 3.18 se observa que el mejor modelo según el método EAM y el coeficiente de R^2 es el modelo donde se utilizó una distribución a priori flat pero por una diferencia poco significativa con los otros, mientras que según el método ECM el mejor modelo sigue siendo el que utiliza una distribución a priori flat con una diferencia mayor que en los otros métodos, pero hay que acotar que según el mismo método los tres modelos se ajuntas entre un 59 % y 76 % a los datos. De lo anteriormente expuesto se tiene que se considera **el mejor modelo para este caso al obtenido utilizando una distribución a priori flat**.

En la Tabla 3.17 al estudiar el modelo utilizando una distribución a priori flat las 6 primeras variables son las regiones de notificación y la de consumo, el lugar de pérdida de inocuidad (L.P.I), la presencia de otros síntomas neurológicos (Otros.Neurológicos), rush cutáneo y parestesias; de estas variables solo tienen efecto negativo sobre el modelo las variables región de consumo y la presencia de rush cutáneo.

Nótese que las variables tasa de ataque y expuestos son las únicas que coinciden con las principales en el mismo caso visto con el enfoque clásico para el modelo Ridge, mientras que se repite la variable cantidad de expuestos en los modelos LASSO clásico y LASSO Bayesiano de este caso. Por otro lado, tanto en el modelo GLM clásico como el flat Bayesiano coinciden la variable rush cutáneo.

Es importante resaltar que el modelo flat da mayor importancia a las regiones mientras que los modelos lineales Ridge y LASSO le dan más importancia a algunos síntomas y la cantidad de expuestos.

3.1.6. Modelos Bayesianos sin considerar las variables de atención y enfocado en la salmonella

Para finalizar el capítulo se presentan los resultados del tercer caso de estudio donde se considera solo las enfermedades producidas por salmonella sin variables de atención vista desde el enfoque Bayesiano.

En la Figura 3.6(a) en el modelo LASSO Bayesiano se observa que los datos observados se encuentran concentrados entre los valores 0 y 10 alrededor de la línea azul, salvo 2 puntos entre 10 y 20 de los valores reales.

Para el modelo Ridge Bayesiano en la Figura 3.6(b) se observa que los datos predichos y observados están concentrados entre los valores 0 y 10 alrededor de la línea azul, además se puede observar una cierta distribución en líneas verticales en los primeros valores de los reales.

Utilizando la distribución a priori flat se observa en la Figura 3.6(c) una esparcimiento de los puntos similar a la obtenida en el modelo Ridge pero con mayor concentración alrededor de la línea azul.

Tabla 3.16: Bondad de ajuste modelos Bayesianos. Caso 2: Sin las variables de atención.

Método	Ridge	LASSO	flat
EAM	2,3610	2,2224	0,7217
ECM	8,6948	8,8120	0,8135
R^2	0,9554	0,9537	0,9952

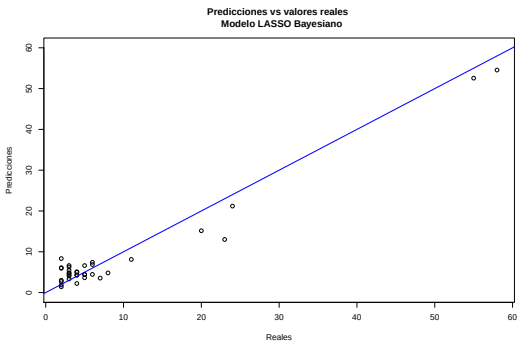
Se observa que ningún modelo realizó selección de variable puesto que no se igualó el coeficiente de ninguna variable a 0 (Tabla 3.27, Anexo 10).

Al extraer las 6 variables cuyos coeficientes en valor absoluto con más grandes (Tabla 3.17), se tiene que para el modelo LASSO y Ridge las primeras 6 son la cantidad de expuestos, la presencia de meteorismo, hipotensión y dolores, la coincidencia de las regiones de consumo y notificación, en cuanto a la 6ta variable se tiene que para el caso del modelo LASSO es duración del brote mientras para el modelo Ridge es la tasa de ataque.

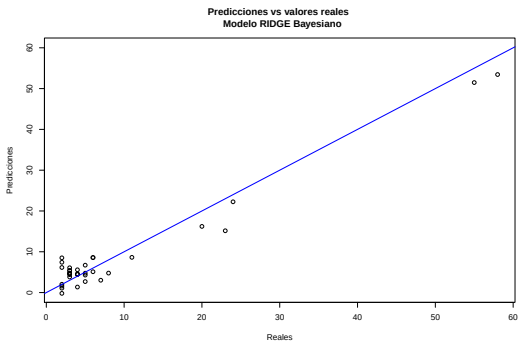
Tabla 3.17: Coeficientes $\boldsymbol{\beta}$ con mayor peso en LASSO, Ridge y flat Bayesiano. Caso 2: Sin las variables de atención.

Variables	LASSO	Ridge	flat
Otros.Neurológicos	0,1108	-0,3166	30,9819
R.notificación	0,3342	0,5322	19,3592
R.consumo	0,2347	0,3507	-18,5799
Rush Cutáneo	-0,3234	-0,1752	-13,1333
L.P.I	0,6358	1,1181	10,2121
Parestesias	0,2896	0,1714	9,9297
Expuestos	5,5951	3,9672	9,3378
Meteorismo	4,8392	3,9690	6,9566
Dolores	-0,9305	-1,6183	-4,7038
Duración.B	0,9060	1,3330	2,5988
Hipotensión	1,0711	2,0222	-2,5800
R.N.RC	-1,3711	-1,4995	-0,6582
Tasa.Ataque	-0,7308	-1,4109	0,2680

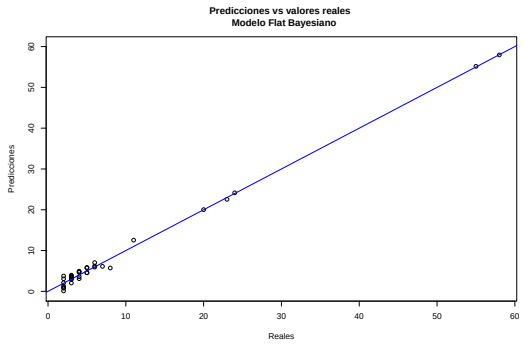
Al observar cuales variables tienen efecto negativo sobre los modelo se encontró que para el modelo LASSO son la coincidencia de regiones y la presencia de dolores, mientras que para modelo Ridge se tiene las mismas que en LASSO pero se incluye la tasa de ataque.



(a) Modelo LASSO Bayesiano



(b) Modelo Ridge Bayesiano



(c) Modelo flat Bayesiano

Figura 3.5: Predicciones versus valores reales de los modelos desde el enfoque Bayesiano. Caso 2: Sin las variables de atención.(a) Modelo LASSO Bayesiano ,(b) Modelo Ridge Bayesiano y (c) Modelo flat Bayesiano.

Al estudiar la bondad de ajuste de los modelos se observa en la Tabla 3.16 que el mejor modelo según el método EAM y el método ECM el mejor modelo es el obtenido con una priori flat con una diferencia significativa respecto a los demás modelos, mientras por el coeficiente R^2 el mejor modelo sigue siendo el obtenido con una priori flat pero la diferencia con los demás modelos no es tan grande como con los otros métodos. Observe que los 3 modelos tienen un ajuste superior al 95 %, lo que significa son modelos con buen ajuste. De lo anterior expuesto se tiene que se considera **el mejor modelo es aquel obtenido utilizando una distribución a priori flat**.

Tabla 3.18: Bondad de ajuste modelos Bayesianos. Caso 3: Solo salmonella y sin variables de atención.

Método	Ridge	LASSO	flat
EAM	1,8919	1,9901	1,6592
ECM	9,0897	11,2079	5,1130
R^2	0,6434	0,5971	0,7627

Al extraer las 6 variables cuyo coeficientes en valor absoluto son mayores para los 3 modelos se creó la Tabla 3.19. En esta tabla se observa que los principales coeficientes para el modelo Ridge y LASSO son la presencia de los síntomas dolores, nauseas, meteorismo y cefalea, la cantidad de expuestos y la duración del brote. Además, se tiene que de las 6 variables anteriormente expuestas los dolores, nauseas y meteorismos tienen efecto negativo sobre ambos modelos.

Cuando se estudia el modelo flat las 6 variables con mayor peso son la presencia de dolores y parestesias, las regiones de consumo y notificación, el local de elaboración (L.elaboración) y la cantidad de expuestos, donde las únicas variables con efecto positivo son la región de notificación y la cantidad de expuesto (Table 3.19).

Tabla 3.19: Coeficientes $\boldsymbol{\beta}$ modelos LASSO, Ridge y flat Bayesianos. Caso 3: Solo salmonella y sin variables de atención.

Variables	LASSO	Ridge	flat
R.consumo	0,1814	0,2646	-50,0297
R.notificación	0,1952	0,2483	50,0083
Expuestos	0,8657	0,8605	5,8588
Parestesias	0,0789	0,1027	-5,1325
L.elaboración	-0,1937	-0,3323	-2,4085
Dolores	-0,8822	-1,1327	-2,1685
Nauseas	-0,5978	-0,8281	-1,8005
Meteorismo	-0,3721	-0,6615	-1,3851
Duración.B	0,5183	0,7863	0,9001
Cefalea	0,3447	0,5268	0,5792

Nótese que en el modelo flat Bayesiano y el GLM clásico se encuentran entre las 6 primeras variables la cantidad de expuestos y las regiones de notificación y consumo. Mientras para el modelo Ridge tanto en el enfoque clásico como en el Bayesiano entre las 10 primeras se encuentran meteorismo, nauseas y dolores.

Por otro lado, es de considerar que bajo el enfoque Bayesiano ningún modelo en este caso hizo selección de variable o igualó el coeficiente de alguna variable a 0 (Tabla 3.28, Anexo 11).

3.1.7. Modelos matemáticos

En esta sección se muestran los mejores modelos para cada caso estudiado y bajo los dos enfoques

- Enfoque clásico:

  Caso 1: Todas las variables. Modelo LASSO.

$$\begin{aligned}\widehat{Enfermos} = \quad & 0,1042 + 0,9962 \times T.Amb + 0,9801 \times T.s.a \\ & + 0,8013 \times T.Hosp + 0,0023 \times R.notificacion \\ & + 0,0006 \times Expuestos - 0.0001 \times Tasa.Ataque.\end{aligned} \tag{3.1}$$

  donde $\widehat{Enfermos}$ es el valor esperado de la cantidad de enfermos.

Caso 2: Sin las variables de atención. Modelo GLM con vínculo Poisson.

$$\begin{aligned}
\widehat{Enfermos} = \quad & 0,6372 + 0,0030 \times Expuestos \\
& + 0,0015 \times Semana - 0,2025 \times P.Incubacion \\
& + 0,1423 \times Duracion.B - 0,8716 \times R.N.RC \\
& + 0,2815 \times R.notificacion - 0,2618 \times R.consumo \\
& - 0,0036 \times Tasa.Ataque - 0,0955 \times Nauseas \\
& + 0,4087 \times Vomitos + 1,8198 \times Diarrea \\
& + 0,2310 \times Fiebre - 0,3270 \times Dolores \\
& + 0,2926 \times Heces.Sang - 0,4979 \times Parestesias \\
& - 0,4488 \times Otros.Neurologicos + 0,2260 \times Espasmos \\
& + 0,9606 \times Meteorismo - 0,1425 \times Deshidratacion \\
& - 0,6118 \times Hipotension + 0,7814 \times Rush.Cutaneo \\
& + 0,3011 \times Cefalea + 0,3910 \times Mialgia \\
& - 0,3908 \times Otros + 0,0345 \times Grupo.Sospechoso \\
& + 0,0884 \times L.elaboracion + 0,1504 \times L.consumo \\
& + 0,0619 \times F.Contaminacion - 0,3278 \times F.Supervivencia \\
& + 0,0382 \times F.Proliferacion - 0,1207 \times P.P.I \\
& - 0,0550 \times L.P.I - 0,0626 \times d.agrupado \\
& + 0,0072 \times CIE.10 + 0,2118 \times C.Inspec.
\end{aligned} \tag{3.2}$$

Caso 3: Solo salmonella y sin las variables de atención. Modelo GLM con vínculo Poisson.

$$\begin{aligned}\widehat{Enfermos} = \quad & 2,5773 + 0,0086 \times Expuestos \\ & + 0,0058 \times Semana - 0,1682 \times P.Incubacion \\ & + 0,0544 \times Duracion.B + 0,0220 \times R.N - RC \\ & - 0,0013 \times Tasa.Ataque - 1,1527 \times Meteorismo \\ & - 1,0925 \times Parestesias - 0,4201 \times Nauseas \\ & + 0,2562 \times Vomitos - 0,2536 \times Diarrea \\ & - 0,4552 \times Dolores - 0,3661 \times Heces.Sang \\ & - 0,6037 \times Otros.Neurologicos + 0,4428 \times Espasmos \\ & + 0,3940 \times Fiebre - 0,1328 \times Deshidratacion \\ & - 0,3024 \times Hipotension + 0,7826 \times Rush.Cutaneo \\ & + 0,1550 \times Cefalea + 0,8432 \times Mialgia \\ & - 0,1839 \times Otros + 0,0246 \times Grupo.Sospechoso \\ & + 0,0106 \times L.elaboracion - 0,0008 \times L.consumo \\ & - 0,0027 \times F.Contaminacion + 0,0173 \times F.Supervivencia \\ & - 0,2453 \times F.Proliferacion - 0,1390 \times P.P.I \\ & - 0,3113 \times L.P.I + 0,0244 \times CIE.10 + 0,3981 \times C.Inspec.\end{aligned} \tag{3.3}$$

- Enfoque bayesiano

  Caso 1: Todas las variables. Modelo utilizando una priori flat.

$$
\begin{aligned}
\widehat{Enfermos} = & -66,6235 \times T.Amb + 73,6739 \times T.Hosp - 40,1214 \times T.s.a \\
& + 9,0740 \times Semana - 19,1182 \times R.notificacion \\
& + 27,9033 \times R.consumo + 6,2744 \times P.Incubacion \\
& - 13,9512 \times Tasa.Ataque - 44,2015 \times Duracion.B \\
& - 28,1213 \times L.consumo - 17,6073 \times L.elaboracion \\
& + 60,3373 \times Expuestos - 19,4227 \times R.N - RC \\
& - 27,0590 \times Hipotension - 2,7761 \times Deshidratacion \\
& + 14,2378 \times Heces.Sang + 18,6274 \times Mialgia \\
& + 6,3171 \times Parestesias - 1,6069 \times Diarrea \\
& + 1,4279 \times Fiebre + 5,9502 \times Nauseas \\
& - 8,3553 \times Cefalea - 9,1914 \times Rush.Cutaneo \\
& + 6,6295 \times Vomitos - 10,6022 \times Dolores \\
& + 0,1601 \times Otros + 23,0169 \times Meteorismo \\
& + 122,3317 \times Otros.Neurologicos + 26,4942 \times Espasmos \\
& - 15,1007 \times Grupo.Sospechoso + 10,3058 \times d.agrupado \\
& + 8,9712 \times CIE.10 - 11,4677 \times P.P.I - 50,2603 \times L.P.I \\
& + 26,1903 \times F.Proliferacion - 27,3757 \times F.Contaminacion \\
& - 5,1780 \times F.Supervivencia + 79,4991 \times C.Inspec \\
& + 14,0039 \times T.amb1 - 91,2921 \times T.hosp.1 \\
& + 30,1644 \times T.amb1.4 - 19,8052 \times T.hosp1.4 \\
& + 25,7184 \times T.amb5.14 - 23,7111 \times T.hosp5.14 \\
& + 123,5362 \times T.s.a5.14 - 4,7452 \times T.s.a15.44 \\
& + 30,7723 \times T.amb15.44 - 51,7447 \times T.hosp15.44 \\
& - 0,9758 \times T.amb45.64 - 17,0389 \times T.hosp45.64 \\
& - 32,7297 \times T.s.a45.64 + 8,1566 \times T.amb65 \\
& - 12,3819 \times T.hosp65 + 23,5813 \times T.s.a65.
\end{aligned}
\tag{3.4}
$$

Caso 2: Sin las variables de atención. Modelo utilizando una priori flat.

$$\begin{aligned}
\widehat{Enfermos} = \quad & 0,2595 \times Semana + 9,3378 \times Expuestos \\
& + 0,2680 \times Tasa.Ataque + 1,5997 \times P.Incubacion \\
& + 19,3592 \times R.notificacion - 18,5799 \times R.consumo \\
& - 2,5800 \times Hipotension + 6,9566 \times Meteorismo \\
& + 2,5988 \times Duracion.B - 4,7038 \times Dolores \\
& - 0,3800 \times Nauseas + 0,1126 \times Mialgia \\
& + 9,9297 \times Parestesias + 0,4631 \times Deshidratacion \\
& - 0,2093 \times Espasmos - 3,7942 \times Heces.Sang \\
& - 0,1924 \times Otros + 30,9819 \times Otros.Neurologicos \\
& + 0,2503 \times Diarrea + 0,3904 \times Fiebre \\
& + 1,4699 \times Vomitos - 13,1333 \times Rush.Cutaneo \\
& - 1,4288 \times Cefalea + 0,1033 \times d.agrupado \\
& + 5,3762 \times P.P.I + 10,2121 \times L.P.I + 2,1246 \times CIE.10 \\
& - 1,4364 \times L.consumo + 1,2848 \times L.elaboracion \\
& - 5,1316 \times F.Supervivencia + 5,9413 \times F.Proliferacion \\
& + 2,7351 \times F.Contaminacion + 0,2664 \times Grupo.Sospechoso \\
& - 0,6582 \times R.N - RC - 5,5296 \times C.Inspec.
\end{aligned} \tag{3.5}$$

Caso 3: Solo salmonella y sin las variables de atención. Modelo utilizando una priori flat.

$$\begin{aligned}\widehat{Enfermos} = \quad & 0,2639 \times Semana + 5,8588 \times Expuestos \\ & + 0,5377 \times Tasa.Ataque + 0,9001 \times Duracion.B \\ & - 50,0297 \times R.consumo + 50,0083 \times R.notificacion \\ & - 0,9872 \times P.Incubacion - 2,1685 \times Dolores \\ & - 1,8005 \times Nauseas - 1,3851 \times Meteorismo \\ & + 1,8768 \times Vomitos - 0,0026 \times Deshidratacion \\ & + 0,3690 \times Mialgia + 1,1553 \times Rush.Cutaneo \\ & + 0,5792 \times Cefalea - 0,2841 \times Hipotension \\ & - 0,3537 \times Otros.Neurologicos - 5,1325 \times Parestesias \\ & - 0,6061 \times Heces.Sang - 0,3363 \times Diarrea \\ & + 1,0369 \times Otros + 0,8919 \times Espasmos \\ & + 0,7394 \times Fiebre + 0,0707 \times F.Contaminacion \\ & + 0,0907 \times F.Proliferacion - 0,8884 \times F.Supervivencia \\ & + 1,2022 \times L.consumo - 2,4085 \times L.elaboracion \\ & + 0,4607 \times CIE.10 - 1,0180 \times L.P.I - 0,4620 \times P.P.I \\ & - 0,7594 \times Grupo.Sospechoso - 2,1148 \times C.Inspec.\end{aligned} \tag{3.6}$$

3.1.8. Comparaciones entre enfoques clásico y Bayesiano

Al comparar los resultados obtenidos para los 3 casos propuestos en este trabajo y bajo los enfoques de la estadística clásica y el análisis Bayesiano se obtuvieron las siguientes observaciones.

En el primer caso, con todas las variables a estudiar sobre los brotes de Salmonella, E. Coli, Shigella y Campylobacter en Chile durante el año 2017, se encontró que en los 3 modelos bajo el enfoque clásicos se repetía la presencia de el intercepto y el total de pacientes que no tuvieron atención médica, mientras para el enfoque Bayesiano se repite en los 3 modelos solamente el total de pacientes atendidos en ambulatorios. Al estudiar ambos enfoques se encontró que ellos coincidían en que dentro de las variables principales de los modelos se encuentran la cantidad de expuestos, la presencia de meteorismo y el

total de pacientes ambulatorios con edades entre 15 y 44 años. Es de resaltar que este este grupo etario en el enfoque clásico se repetía con el total de personas que no tuvo atención médica por lo que se considera un grupo relevante, mientras en el enfoque Bayesiano se encontró un comportamiento que le daba relavancia a los pacientes ambulatorios con edades superiores a 15 años.

En el segundo caso se tenía el estudio del caso anterior sin las variables de atención. En éste, se encontró que para el enfoque clásico únicamente se repetía el intercepto entre los tres modelos, mientras en el enfoque Bayesiano no se repetían variables entre las principales cuyos coeficiente tenían mayor peso. Por otro lado, cuando se estudiaron las coincidencia entre ambos enfoques se encontró que, la cantidad de expuestos, la tasa de ataque y meteorismo aparecían en los modelos como variables principales, mientras que se repetían con respecto a las más significativas en el modelo GLM clásico las variables de presencia de dolores, rush cutáneo y meteorismo, las regiones de notificación y consumo, la coincidencia de estas, la duración del brote y la cantidad de expuestos. Nótese, que en este caso las variables sintomáticas tienen mayor peso tanto entre las variables más significativas junto con las regiones.

Para el tercer caso, se estudiaron solo los brotes de Salmonella sin considerar las variables de atención. Se encontró en el enfoque Bayesiano que las variables sobre la presencia de dolores y la cantidad de expuestos aparecían entre las principales de todos los modelos, mientras para el enfoque clásico se repite el intercepto y la cantidad de expuestos en los tres modelos estudiados. Al comparar los dos enfoques se tiene que en ambos se puede encontrar las variables cantidad de expuestos, las regiones de consumo y notificación, la duración del brote y la presencia de meteorismo, parestesia, dolores y nauseas entre las principales variables de sus modelos. Además, las variables más significativas según el modelo GLM clásico sobre la presencia de dolores y nauseas, junto con la cantidad de expuestos se repiten en los modelos Bayesianos. Nótese que en este caso las variables sintomáticas tienen mayor peso tanto entre las variables principales como más significativas.

Por otro lado cuando se estudian los métodos de bondad de ajuste de los modelos LASSO, GLM y Ridge, se tiene que para el primer caso y bajo el enfoque clásico el modelo de LASSO es el mejor entre los 3 métodos estudiados, es decir, es el modelo cuyo valor de EAM y ECM es el más bajo y el coeficiente R^2 es el mayor, con un 99,92 % de ajuste sobre los datos. Bajo el enfoque Bayesiano se tiene que el mejor modelo es el que utiliza una distribución a priori flat el cual es el modelo cuyo valores de ECM y EAM es el más bajo con una diferencia significativa respecto a los valores de Ridge y LASSO; seguidamente

según el coeficiente de R^2 el mejor modelo es el LASSO Bayesiano con un 99,88 % por poca diferencia con el de flat que es 99,85 %, por lo anterior expuesto se considera el mejor modelo para este caso el que utiliza una distribución a priori flat, puesto que es el mejor en 2 de los 3 métodos y el segundo mejor en el faltante.

En el segundo caso, en el enfoque clásico el modelo GLM es el mejor bajo dos métodos con una diferencia significativa con respecto al resto, el ECM y R^2, ajustando un 85,72 % de los datos, seguido por el modelo Ridge que es el mejor dado que su EAM es menor que al de LASSO y GLM; por lo antes mencionado se considera el mejor modelo al modelo GLM, puesto que es el mejor bajo 2 de 3 métodos. Mientras, para el enfoque Bayesiano el mejor modelo es el que utiliza una priori flat, al tener los menores valores de EAM y ECM, y el mayor coeficiente $R2$, ajustando un 99,52 % de los datos.

Para el tercer caso se tiene que el mejor modelo según los métodos de bondad de ajuste de el EAM, ECM y el coeficiente R^2 que para el enfoque Bayesiano el modelo que utiliza una distribución a priori flat es el mejor y para el enfoque clásico el modelo GLM al ser los mejores según sus valores en los 3 métodos.

Al observar la relación entre las variables principales y los mejores modelos se tiene que para el primer caso se encuentra que las variables de total de personas atendidas en ambulatorios y en hospitales aparece en ambos modelos. Por otro lado, para el segundo caso se tiene que la presencia de rush cutáneo y las regiones de consumo y notificación aparecen entre las variables principales y más significativas entre los mejores modelos. Para en el tercer caso entre los mejores modelo se encontró la cantidad de expuestos y la presencia de dolores y parestesia como principales variables y más significativas.

CONCLUSIONES

Como se puede observar durante el estudio de las enfermedades transmitidas por alimentos (ETA), se tiene que para el primer caso, donde incluimos las variables de atención, las variables más importantes debido a su peso fueron los totales de personas atendidas en hospitales y ambulatorios, puesto que ambas aparecen entre las 6 primeras de los mejores modelos tanto del enfoque clásico con el modelo LASSO y del análisis Bayesiano con el modelo que utiliza una distribución a priori flat, los cuales tienen un ajuste del 99,92 % y 99,85 % respectivamente. También, hay resaltar la recurrencia del grupo etario de 15 a 44 años que aparece en los modelos entre las principales variables y el cual presentaba los mayores valores máximos, de desviación estándar y varianza, aún cuando no es un grupo vulnerable.

Para el segundo caso, se estudiaron los datos sin considerar las variables de atención, en éste se encontró que la presencia de rush cutáneo aparece entre las principales variables de los mejores modelos, el modelo lineal generalizado con vínculo Poisson para el caso clásico y el modelo con distribución a priori flat. Además, se observó que son significativas siempre las variables sobre las regiones de consumo y notificación, cantidad de expuesto, duración del brote, el factor de supervivencia y la presencia de diarrea, de los cuales se conoció que la duración tenía los valores más bajos de desviación y varianza, mientras que los expuestos tenía los más altos, además la presencia de diarrea era la variable que presenta mayor porcentaje de respuestas positivas, mientras que el rush cutáneo posee el segundo mayor porcentaje de respuestas negativas. Por otro lado, hay que resaltar que la cantidad de expuestos es variable que se repite entre las principales de los modelos en ambos enfoques.

En el tercer caso se realizó un estudio enfocado en la salmonella sin considerar las variables de atención, en este caso se encontró que los mejores modelos fueron el modelo GLM con vínculo Poisson para el enfoque clásico y el modelo utilizando una distribución a priori flat para el Bayesiano, para los cuales se repite la presencia de parestesia y dolores entre las principales variables de ambos modelos. Por otro lado, la variable que siempre es significativa para el caso clásico en el modelo GLM es la mialgia la cual presentaba un 93,97 % de respuestas negativas; además, se observó que entre las variables significa-

tivas se encuentran también la presencia de dolores, nauseas, fiebre, rush cutáneo, factor contribuyente a la supervivencia y expuestos, donde las 3 primeras son síntomas con alto porcentaje de aparición y los 2 últimos vienen heredados del caso de estudio anterior. Por otro lado, hay que resaltar que las variables intercepto, la presencia de dolores y la cantidad de expuestos son aquellas que se repiten entre las principales de los modelos en ambos enfoques.

Por lo antes expuesto, la cantidad de expuestos es la variable de gran importancia en los modelos para los 3 casos; además, la presencia de dolores, parestesias, rush cutáneo y nauseas son principales y significativas para los modelos cuando no se tiene los totales de las personas que fueron o no atendidas médicamente, junto con las variables globales sobre la región de consumo y notificación.

Se recomienda para futuros estudios que se ahonde más en la relevancia del grupo etario de 15 a 44 años, dado que este no es un grupo que para organismos internacionales se considere vulnerable, además, se podría estudiar ya que la atención ambulatoria tiene mayor peso para el modelo LASSO Bayesiano con todas las variables. Finalmente, hay que resaltar que bajo el enfoque Bayesiano el modelo flat predominó sobre los demás en todos los casos, mientras para el enfoque clásico el modelo GLM con vínculo Poisson predominó en los casos sin considerar los totales de atención y el modelo LASSO solo lo hizo con todas las variables.

Referencias

Adams, M., Motarjemi, Y., y Santé, O. (1999). *Basic food safety for health workers.* (Disponible en : `http://libdoc.who.int/hq/1999/WHO_SDE_PHE_FOS_99.1.pdf`)

Allasia, M. B., Branco, M., y Quaglino, M. (2016). Regresión lasso bayesiana. ajuste de modelos lineales penalizados mediante la asignación de priores normales con mezcla de escala. *Vigesimo primeras Jornadas "Investigaciones en la Facultad "de Ciencias Económicas y Estadística.*

Arias, M. (2016). *Análisis introductorio de métricas básicas para el cálculo de errores de pronósticos.* (`https://www.macrologistica.co.cr/images/docs/analisis\_introductorio.pdf`)

Bayes, T. (1763). An essay towards solving in the doctrine of chances. *Philosophical Transactions of the Royal Society London.*

Bravo, L., Llatas, I., y Peréz, M. E. (2008). *Análisis de datos con técnicas bayesianas.* Caracas, Venezuela: XXI Escuela Venezolana de Matemáticas (EVM), Escuela Matemática de América Latina y el Caribe (EMALCA).

Carrasco C., M. (2016). *Técnicas de regularización en represión: Implementación y aplicaciones* (Tesis de Master no publicada). Universidad de Sevilla, Sevilla, España.

CDC. (2018). *Seguridad de los alimentos.* (Disponible en: `http://www.cdc.gov/foofsafety/es/foodborne-germs-es.html`)

DEIS. (2019). *Brotes de enfermedades transmitidas por alimentos. chile, periodos año 2011-2018.* (Disponible en: `http://public.tableau.com/profile/deis4231#!/vizhome/BrotesdeEnfermedadesTransmitidasporAlimentoETA_Aos2011-2017/BrotesETAChile2011-2017`)

DEIS. (2019a). *Brotes de enfermedades transmitidas por alimentos. chile, periodos año 2011.* (Disponible en: `http://www.deis.cl/estadisticas-eta/`)

Diluvi, G. C. (2017). Modelos lineales generalizados: Un enfoque bayesiano. laberintos e infinitos. , *45*(1), 36-45.

Dobson, A. (2002). *An introduction to generalized linear models* (2nd ed.). Chapman y Hall/CRC texts in statistical science series.

FAO. (2003). *Principios generales de higiene de los alimentos cac/rcp 1-1969. revisión 4.*

FAO, OMS, y UA. (2019). *Primera conferencia internacional fao/oms/ua sobre inocuidad alimentaria: Resumen del presidente.* Addis Abeba, Etiopía. (Disponible en: `https://www.who.int/docs/default-source/resources/chairpersons-summary-addis-ababa-es.pdf?sfvrsn=8033cd47_4`)

FDA. (2018). *Las enfermedades transmitidas por alimentos son especialmente peligrosas para las personas vulnerables.* (Disponible en:`http://www.fda.gov/consumers/articulos-en-espaol/las-enfermedades-transmitidas-por-alimentos-son-especialmente-peligrosas\-para-las-personas`)

Guitiérrez-Peña, E. (2016). Análisis bayesiano de modelos jerárquicos lineales. *Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas, Universidad Nacional Autónoma de México..*

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, *96*(4), 835–845. doi: 10.1093/biomet/asp047

Hoerl, A., y Kennard, R. (1970). Ridge regression: biased estimation for non-orthogonal problems. *Technometrics*, *12*(1), 55-67.

Medina, M. J. (2018). *Métodos de regularización para la selección de variables aplicados a la predicción del riesgo de padecer disfunción motora en adultas mayores activas de la ciudad de valdivia* (Tesis Maestría en Estadística). Universidad de Concepción, Chile.

Migon, H., y Gamerman, D. (1999). *Statistical inference: An integrated approach.* Gran Bretaña, Londres: Hodder Arnold.

OMS. (2008). *Initiative to estimate the global burden of foodborne diseases. a summary document.* (Disponible en: http://www.who.int/foodsafety/foodborne_disease/Summary_Doc.pdf)

OMS. (2008a). *Foodborne disease outbreaks: Guidelines for investigation and control. 1–146 p.* (https://apps.who.int/iris/bitstream/handle/10665/43771/9789241547222_eng.pdf?sequence=1)

OMS. (2015). *Carga mundial de enfermedades de transmisión alimentaria: estimaciones de la oms.* (Disponible en: https://www.who.int/foodsafety/areas_work/foodborne-diseases/fergonepager_es.pdf?ua=1)

OMS. (2015a). *Las enfermedades de transmisión alimentaria (eta) en la región de las américas de la oms .* (Disponible en: https://www.who.int/foodsafety/areas_work/foodborne-diseases/amro_es.pdf?ua=1)

OMS. (2015b). *Estimaciones de la oms sobre la carga mundial de enfermedades de transmisiÓn alimentaria.* (Disponible en : http://www.who.int/foodsafety/areas_work/foodborne-diseases/ferg/en/)

OMS. (2019). *Inocuidad de los alimentos.* (Disponible en: https://www.who.int/es/news-room/fact-sheets/detail/food-safety)

OPS. (2015). *Enfermedades transmitidas por alimentos (eta).* (Disponible en: http://www.paho.org/hq/index.php?option=com_content&view=article&id=10836:2015-enfermedades-transmitidas-por-alimentos-eta&Itemid=41432&lang=es)

OPS. (2015a). *Anexo g: Factores determinantes de las enfermedades transmitidas por alimentos. factores de contaminación, supervivencia y multiplicación.* (Disponible en: https://www.paho.org/hq/index.php?option=com_content&view=article&id=10808:2015-anexo-g-factores-determinantes-alimentos&Itemid=41421&lang=es)

Park, t., y Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 681-686. doi: 10.1198/016214508000000337

Piñera, S. (s.f.). *Metas 2011-2020: Elige vivir sano.* (Disponible en: http://www.ispch.cl/objetivossanitarios)

Prieto, M., Mouwen, J., López, S., y Cerdeño, A. (2008). Concepto de calidad en la industria agroalimentaria. http://www.redalyc.org/pdf/339/33933405.pdf. *Prisma*, *33*(4), 258–264.

Ramos Castillo, L. (2018). *Regresion lasso* (Tesis de Grado en Matemáticas). Universidad de Sevilla, Sevilla, España.

R-foundation. (s.f.). *The r project for statistical computing.* (Disponible en: http://www.r-project.org)

Roan Veerman, J. (2018). *Estimating error and prior variance in a high-dimensional ridge regression models* (Masters Thesis in Statistical Science). Universidad de Leiden, Paises Bajo.

SAIA. (2017). *El control de calidad en los alimentos: qué es y de dónde viene.* (Disponible en: `http://saia.es/control-calidad-alimentos/`)

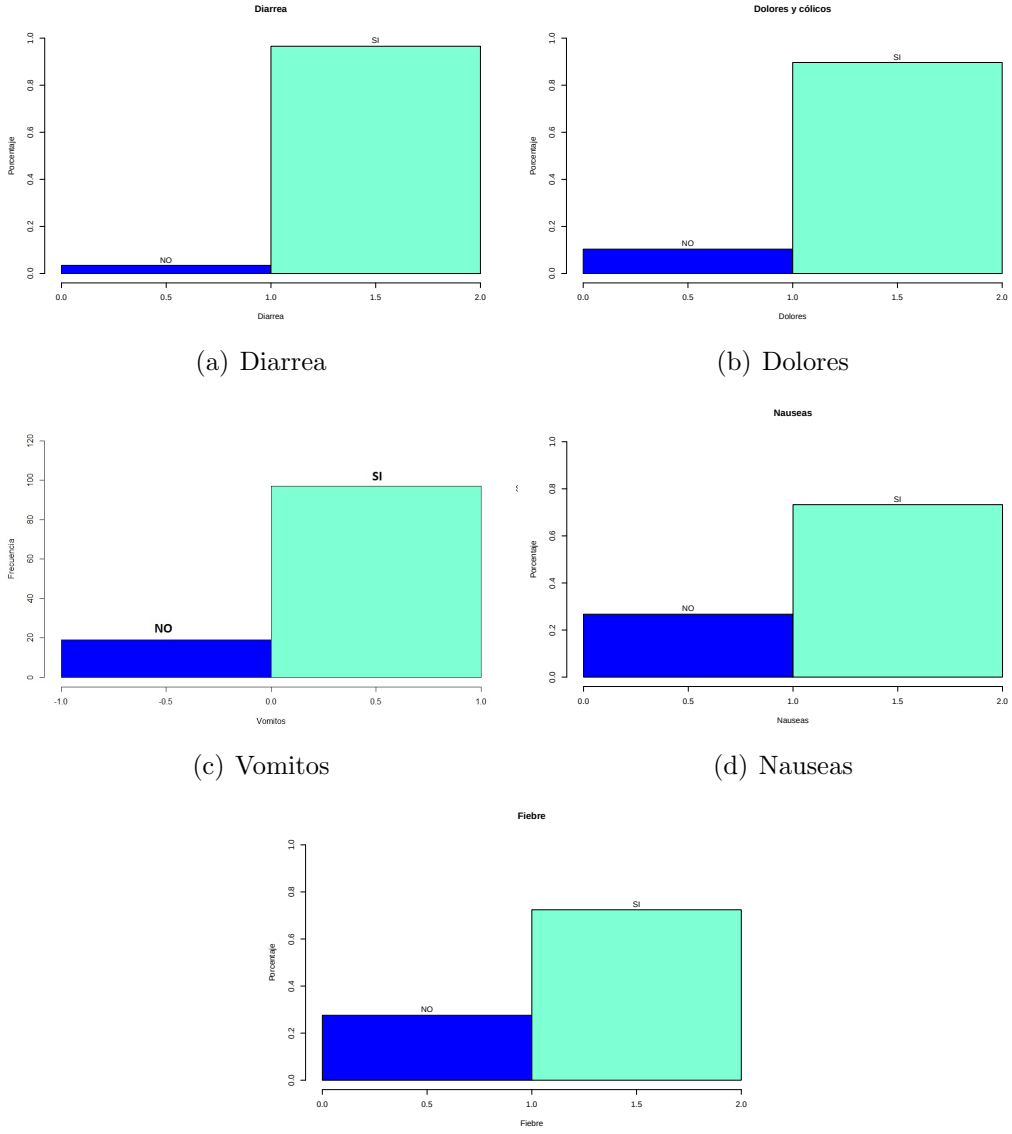
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267-288.

Ulloa Bello, M. A. (2016). *Enfermedades transmitidas por los alimentos en chile: agentes causantes y factores contribuyentes asociados a brotes ocurridos durante el año 2013.* (Tesis Maestría en Alimentos mención Gestión, Calidad e Inocuidad de los Alimentos). Universidad de Chile, Santiago, Chile.

Wackerly, D., Mendenhall, W., y Scheaffer, R. (2008). *Estadística matemática con aplicaciones* (7ma. ed.). México: Cengage Learning Editores.

ANEXOS

Anexo 1. Diagrama de barra de los síntomas presentes.



(a) Diarrea

(b) Dolores

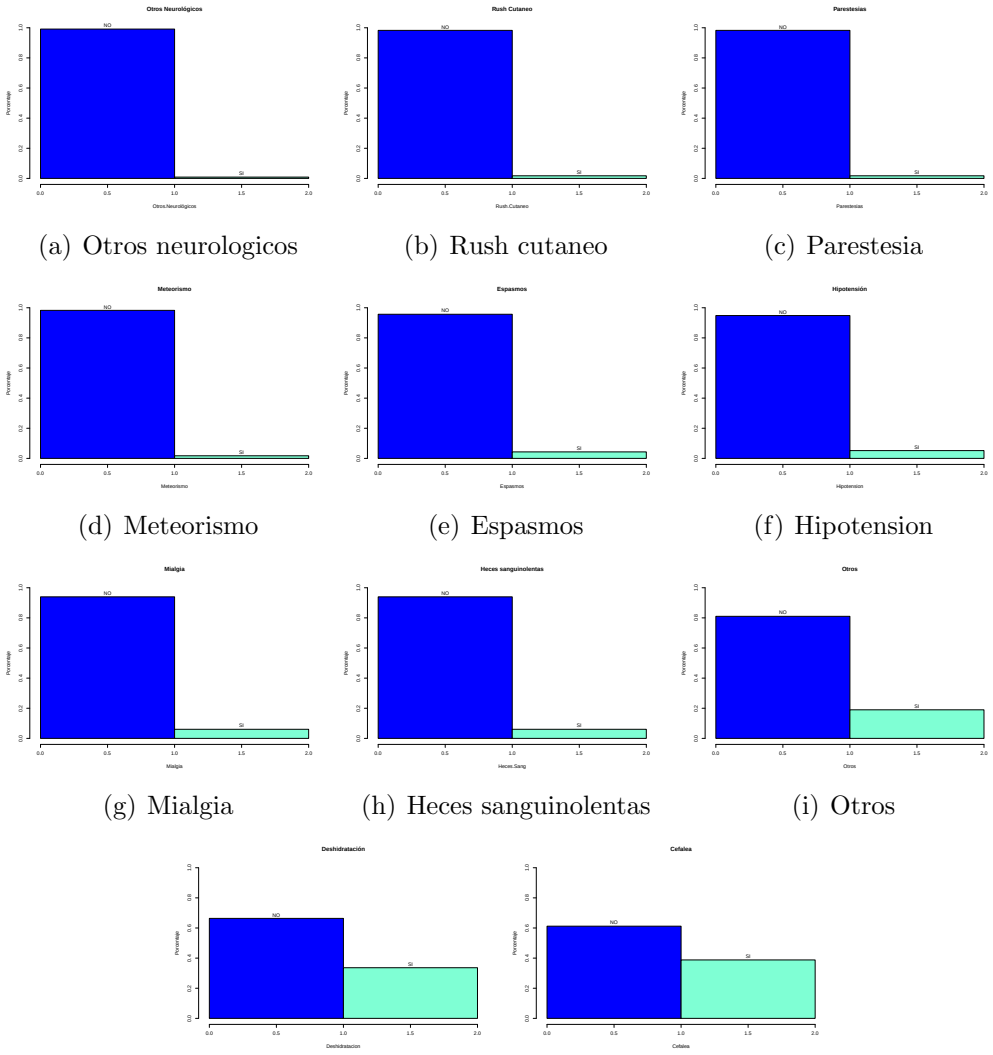
(c) Vomitos

(d) Nauseas

(e) Fiebre

Figura 3.7: Diagrama de barra de los síntomas presentes. (a) Diarrea, (b) Dolores y cólicos, (c) Vómitos, (d) Nauseas, (e) Fiebre.

Anexo 2. Diagrama de barra de los síntomas no presentes.



(a) Otros neurologicos (b) Rush cutaneo (c) Parestesia

(d) Meteorismo (e) Espasmos (f) Hipotension

(g) Mialgia (h) Heces sanguinolentas (i) Otros

(j) Deshidratacion (k) Cefalea

Figura 3.8: Diagrama de barra síntomas no presentes. (a) Otros neurológicos, (b) Rush cutáneo, (c) Parestesias, (d) Meteorismo, (e) Espasmos, (f) Hipotensión, (g) Mialgia, (h) Heces sanguinolentas, (i) Otros, (j) Deshidratación, (k) Cefalea.

Anexo 3. Coeficientes de $\boldsymbol{\beta}$ para el caso clásico con todas las variables.

Tabla 3.20: Tabla de coeficientes de $\boldsymbol{\beta}$ para el enfoque clásico. Caso 1: Todas las variables.

Variable	LASSO	RIDGE	GLM
T.Amb	0,9962	0,6547	-0,0033
T.s.a	0,9801	0,4853	0,7466
T.Hosp	0,8013	0,1497	0,5247
Intercepto	0,1042	0,5708	1,3684
R.notificación	0,0023	0,0172	-0,0837
Expuestos	0,0006	0,0046	-0,0011
Tasa.Ataque	-0,0001	-0,0035	-0,0041
Semana	0	0,0043	-0,0010
P.Incubación	0	-0,0056	-0,0231
Duración.B	0	0,0435	-0,0038
R.consumo	0	0,0217	0,0931
R.N-RC	0	0,0025	0,2793
T.amb1	0	-0,0001	0,4076
T.amb1.4	0	0,0317	0,2104
T.amb5.14	0	0,0162	0,1389
T.amb15.44	0	0,3139	0,1040
T.amb45.64	0	0,2366	0,0499
T.amb65	0	0,0564	0
T.hosp1.4	0	0,0081	-0,6509
T.hosp5.14	0	0,0328	-0,2145
T.hosp15.44	0	0,0818	-0,5521
T.hosp45.64	0	0,0207	-0,1424
T.hosp65	0	0,0024	-0,3307
T.hosp.1	0	0,0039	0
T.s.a5.14	0	0,0110	-0,5476
T.s.a15.44	0	0,3684	-0,7716
T.s.a45.64	0	0,1058	-0,4116
T.s.a65	0	0	0
Nauseas	0	0,0117	0,0437

Continuación.

Variable	LASSO	RIDGE	GLM
Vomitos	0	0,0198	-0,0534
Diarrea	0	0,0099	0,0415
Dolores	0	0,0119	0,0912
Heces.Sang	0	0,0077	0,0561
Parestesias	0	-0,0025	0,2551
Otros.Neurológicos	0	-0,0017	-0,1315
Espasmos	0	-0,0028	0,5657
Fiebre	0	0,0006	0,0106
Deshidratacion	0	0,0047	-0,0934
Hipotension	0	0,0009	-0,0562
Rush.Cutaneo	0	0,0031	0,2639
Cefalea	0	0,0109	0,0232
Mialgia	0	0,0041	-0,1564
Meteorismo	0	0,0016	-1,6763
Otros	0	0,0104	0,0150
Grupo.Sospechoso	0	0,0124	0,0150
L.elaboración	0	0,0377	0,0225
L.consumo	0	0,0158	-0,0828
F.Contaminacion	0	-0,0083	-0,0345
F.Supervivencia	0	-0,0333	-0,0214
F.Proliferacion	0	-0,0229	0,0275
P.P.I	0	0,0042	-0,0297
L.P.I	0	-0,0025	-0,0501
CIE.10	0	0,0080	0,0022
d.agrupado	0	0,0115	-0,1140
C.Inspec	0	-0,0022	0,0798

Anexo 4. Coeficientes de $\boldsymbol{\beta}$ para el caso clásico sin las variables de atención.

Tabla 3.21: Tabla de coeficientes de $\boldsymbol{\beta}$ para el enfoque clásico. Caso 2: Sin las variables de atención.

Variable	LASSO	RIDGE	GLM
Intercepto(Intercept)	6,2539	6,8386	0,6372
Expuestos	0,0288	0,0182	0,0030
Semana	0	0,0017	0,0015
P.Incubación	0	-0,0001	-0,2025
Duración.B	0	0,0002	0,1423
R.notificación	0	0,0003	0,2815
R.consumo	0	0,0002	-0,2618
R.N-RC	0	0	-0,8716
Tasa.Ataque	0	-0,0045	-0,0036
Nauseas	0	0	-0,0955
Vomitos	0	0	0,4087
Diarrea	0	0	1,8198
Dolores	0	0	-0,3270
Heces.Sang	0	0	0,2926
Parestesias	0	0	-0,4979
Otros.Neurológicos	0	0	-0,4488
Espasmos	0	0	0,2260
Fiebre	0	0	0,2310
Deshidratacion	0	0	-0,1425
Hipotension	0	0	-0,6118
Rush.Cutaneo	0	0	0,7814
Cefalea	0	0	0,3011
Mialgia	0	0	0,3910
Meteorismo	0	0	0,9606
Otros	0	0	-0,3908
Grupo.Sospechoso	0	-0,0001	0,0345
L.elaboración	0	0	0,0884
L.consumo	0	0,0001	0,1504

Continuación

Variable	LASSO	RIDGE	GLM
F.Contaminacion	0	0,0001	0,0619
F.Supervivencia	0	-0,0001	-0,3278
F.Proliferacion	0	-0,0001	0,0382
P.P.I	0	-0,0001	-0,1207
L.P.I	0	0	-0,0550
CIE.10	0	0,0014	0,0072
d.agrupado	0	0	-0,0626
C.Inspec	0	0,0001	0,2118

Anexo 5. Coeficientes de $\boldsymbol{\beta}$ para el caso clásico de salmonella sin las variables de atención.

Tabla 3.22: Tabla de coeficientes de $\boldsymbol{\beta}$ para el caso clásico. Caso 3: Solo salmonella y sin las variables de atención.

Variable	LASSO	RIDGE	GLM
Intercepto	5,2004	5,3623	2,5773
Expuestos	0,0240	0,0416	0,0086
Semana	0	0,0128	0,0058
P.Incubación	0	0,0001	-0,1682
Duración.B	0	0,0041	0,0544
R.notificación	0	0,0072	0
R.consumo	0	0,0072	0
R.N-RC	0	0	0,0220
Tasa.Ataque	0	-0,0096	-0,0013
Nauseas	0	-0,0004	-0,4201
Vomitos	0	0,0001	0,2562
Diarrea	0	0,0001	-0,2536
Dolores	0	-0,0003	-0,4552
Heces.Sang	0	-0,0001	-0,3661
Parestesias	0	-0,0001	-1,0925
Otros.Neurológicos	0	-0,0001	-0,6037

Continuación.

Variable	LASSO	RIDGE	GLM
Espasmos	0	0,0003	0,4428
Fiebre	0	0,0007	0,3940
Deshidratacion	0	-0,0002	-0,1328
Hipotension	0	-0,0001	-0,3024
Rush.Cutaneo	0	0,0002	0,7826
Cefalea	0	0,0009	0,1550
Mialgia	0	0,0008	0,8432
Meteorismo	0	-0,0001	-1,1527
Otros	0	-0,0004	-0,1839
Grupo.Sospechoso	0	0,0004	0,0246
L.elaboración	0	-0,0010	0,0106
L.consumo	0	0,0016	-0,0008
F.Contaminacion	0	0,0023	-0,0027
F.Supervivencia	0	-0,0010	-0,2453
F.Proliferacion	0	0	0,0173
P.P.I	0	-0,0001	-0,1390
L.P.I	0	0,0003	-0,3113
CIE.10	0	0,0043	0,0244
d.agrupado	0	0	0
C.Inspec	0	0,0010	0,3981

Anexo 6. Nivel de significancia de las variables del modelo GLM para el caso con todas las variables.

Tabla 3.23: Tabla de $p - valores$ para el caso clásico. Caso 1: Todas las variables.

Variables	$P - valor$	Significancia
Intercepto	0,1908	
Semana	0,7941	
P.Incubación	0,7350	
Duración.B	0,9370	
R.notificación	0,3507	
R.consumo	0,3023	
R.N.RC	0,4617	
Expuestos	0,3264	
Tasa.Ataque	0,1129	
T.Amb	0,9793	
T.amb1	0,4573	
T.amb1.4	0,2149	
T.amb5.14	0,3474	
T.amb15.44	0,4221	
T.amb45.64	0,7161	
T.Hosp	0,5787	
T.hosp1.4	0,5259	
T.hosp5.14	0,8270	
T.hosp15.44	0,5680	
T.hosp45.64	0,8831	
T.hosp65	0,7797	
T.s.a	0,0545	.
T.s.a5.14	0,3301	
T.s.a15.44	0,0685	.
T.s.a45.64	0,2078	
Nauseas	0,7643	
Vomitos	0,7476	
Diarrea	0,8993	
Dolores	0,6669	

Continuación

Variables	$P - valor$	Significancia
Heces.Sang	0,8647	
Parestesias	0,5511	
Otros.Neurológicos	0,8687	
Espasmos	0,0522	.
Fiebre	0,9406	
Deshidratacion	0,5031	
Hipotension	0,8711	
Rush.Cutaneo	0,6473	
Cefalea	0,8539	
Mialgia	0,5366	
Meteorismo	0,0020	**
Otros	0,9278	
Grupo.Sospechoso	0,5923	
L.elaboración	0,8692	
L.consumo	0,4758	
F.Contaminacion	0,4341	
F.Supervivencia	0,8056	
F.Proliferacion	0,4738	
P.P.I	0,6879	
L.P.I	0,8415	
CIE.10	0,7240	
d.agrupado	0,3541	
C.Inspec	0,7040	

Anexo 9. Coeficientes de $\boldsymbol{\beta}$ para el caso con todas las variables en el enfoque Bayesiano.

Tabla 3.26: Tabla de coeficientes de $\boldsymbol{\beta}$ para el caso Bayesiano. Caso 1: Todas las variables.

Variables	RIDGE	LASSO	flat
T.Amb	4,8706	2,3410	-66,6235
T.amb15.44	1,9858	2,6165	30,7723
T.amb45.64	0,7669	1,4341	-0,9758
Expuestos	0,4398	0,8358	60,3373
Meteorismo	0,3948	1,0181	23,0169
T.amb65	0,3065	0,5954	8,1566
T.Hosp	0,2739	0,1823	73,6739
T.hosp15.44	0,2659	0,2622	-51,7447
Tasa.Ataque	-0,2055	-0,4702	-13,9512
T.hosp5.14	0,1981	0,3608	-23,7111
Duración.B	0,1419	0,2057	-44,2015
Heces.Sang	0,1337	0,3419	14,2378
Mialgia	0,1297	0,2511	18,6274
CIE.10	0,1197	0,3083	8,9712
Parestesias	-0,1042	-0,0254	6,3171
d.agrupado	-0,1024	-0,2053	10,3058
Hipotension	0,0882	0,3509	-27,0590
R.notificación	0,0859	0,1722	-19,1182
T.hosp.1	0,0845	-0,0519	-91,2921
Semana	0,0821	0,2604	9,0740
Deshidratacion	0,0776	0,1461	-2,7761
Fiebre	-0,0723	-0,0048	1,4279
L.consumo	0,0649	0,0941	-28,1213
R.N-RC	-0,0646	-0,4826	-19,4227
L.elaboración	0,0632	0,0195	-17,6073
Diarrea	0,0615	0,1391	-1,6069
T.s.a45.64	0,0583	-0,0161	-32,7297
T.amb5.14	0,0568	0,1462	25,7184

Continuación.

Variables	$P - valor$	Significancia
Vomitos	0,1308	
Diarrea	0,5761	
Heces.Sang	0,4825	
Parestesias	0,2240	
Otros.Neurológicos	0,4439	
Espasmos	0,1238	
Deshidratacion	0,3340	
Hipotension	0,4028	
Cefalea	0,1701	
Meteorismo	0,1967	
Otros	0,2093	
Grupo.Sospechoso	0,3265	
L.elaboración	0,9464	
L.consumo	0,9934	
F.Contaminacion	0,9485	
F.Proliferacion	0,6156	
L.P.I	0,2400	

Continuación.

Variables	$P-valor$	Significancia
F.Contaminacion	0,0782	.
F.Supervivencia	0	***
F.Proliferacion	0,2070	
P.P.I	0,0564	.
L.P.I	0,7745	
CIE.10	0,1117	
d.agrupado	0,5294	
C.Inspec	0,2065	

Anexo 8. Nivel de significancia de las variables del modelo GLM para el caso de solo salmonella.

Tabla 3.25: Tabla de $p-valores$ para el caso clásico. Caso 3: Solo salmonella y sin las variables de atención.

Variables	$P-valor$	Significancia
Intercepto	0,0199	*
Dolores	0,0345	*
Fiebre	0,0152	*
CIE.10	0,0351	*
Expuestos	0,0080	**
Nauseas	0,0026	**
Rush.Cutaneo	0,0089	**
F.Supervivencia	0,0060	**
Mialgia	0,0003	***
P.Incubación	0,0540	.
P.P.I	0,0710	.
C.Inspec	0,0528	.
Semana	0,1368	
Duración.B	0,2183	
R.notificación	0,1006	
Tasa.Ataque	0,6507	

Anexo 7. Nivel de significancia de las variables del modelo GLM para el caso sin las variables de atención.

Tabla 3.24: Tabla de $p - valores$ para el caso clásico. Caso 2: Sin las variables de atención.

Variables	$P - valor$	Significancia
Intercepto	0,3905	
Semana	0,6090	
P.Incubación	0,0012	**
Duración.B	0	***
R.notificación	0	***
R.consumo	0,0001	***
R.N-RC	0,0039	**
Expuestos	0	***
Tasa.Ataque	0,0807	.
Nauseas	0,4247	
Vomitos	0,0061	**
Diarrea	0	***
Dolores	0,0272	*
Heces.Sang	0,2644	
Parestesias	0,1920	
Otros.Neurológicos	0,5614	
Espasmos	0,3706	
Fiebre	0,0532	.
Deshidratacion	0,2005	
Hipotension	0,0277	*
Rush.Cutaneo	0,0041	**
Cefalea	0,0015	**
Mialgia	0,0332	*
Meteorismo	0,0020	**
Otros	0,0023	**
Grupo.Sospechoso	0,0942	.
L.elaboración	0,3757	
L.consumo	0,0679	.

Continuación.

Variables	RIDGE	LASSO	flat
T.hosp45.64	0,0525	0,0851	-17,0389
Grupo.Sospechoso	-0,0510	0,1190	-15,1007
Nauseas	0,0474	0,2648	5,9502
F.Supervivencia	-0,0470	-0,1222	-5,1780
T.s.a15.44	0,0415	-0,0098	-4,7452
P.Incubación	-0,0414	-0,0760	6,2744
Rush.Cutaneo	-0,0399	-0,0182	-9,1914
T.amb1.4	0,0365	0,1914	30,1644
T.s.a	0,0353	0,0643	-40,1214
Cefalea	-0,0327	-0,3222	-8,3553
P.P.I	0,0302	0,2075	-11,4677
L.P.I	0,0291	0,1347	-50,2603
Dolores	-0,0284	-0,4556	-10,6022
F.Proliferacion	-0,0279	0,1223	26,1903
T.hosp1.4	0,0276	-0,0954	-19,8052
T.amb1	0,0255	0,1644	14,0039
R.consumo	0,0252	0,0409	27,9033
Vomitos	0,0246	0,0253	6,6295
Otros	0,0241	-0,0872	0,1601
Otros.Neurológicos	0,0234	0,0314	122,3317
C.Inspec	0,0218	0,1659	79,4991
F.Contaminacion	-0,0124	0,1967	-27,3757
T.s.a5.14	0,0092	0,0564	123,5362
Espasmos	0,0088	-0,0082	26,4942
T.hosp65	-0,0050	0,0425	-12,3819
T.s.a65	0,0024	0,1508	23,5813

Anexo 10. Coeficientes de $\boldsymbol{\beta}$ para el caso sin las variables de atención en el enfoque Bayesiano.

Tabla 3.27: Tabla de coeficientes de $\boldsymbol{\beta}$ para el caso Bayesiano. Caso 2: Sin las variables de atención.

Variables	RIDGE	LASSO	flat
Expuestos	5,5951	3,9672	9,3378
Meteorismo	4,8392	3,9690	6,9566
R.N-RC	-1,3711	-1,4995	-0,6582
Hipotension	1,0711	2,0222	-2,5800
Dolores	-0,9305	-1,6183	-4,7038
Duración.B	0,9060	1,3330	2,5988
CIE.10	0,8083	0,9373	2,1247
Tasa.Ataque	-0,7308	-1,4109	0,2680
L.P.I	0,6358	1,1181	10,2121
Semana	0,6010	1,0853	0,2595
Cefalea	-0,4807	-1,0029	-1,4288
P.P.I	0,4462	0,3763	5,3762
Nauseas	0,4173	0,7376	-0,3800
Mialgia	0,4149	0,4821	0,1126
Espasmos	0,4146	0,5220	-0,2093
Heces.Sang	0,3602	1,1318	-3,7942
d.agrupado	-0,3491	-0,3339	0,1033
L.consumo	0,3422	0,1930	-1,4364
R.notificación	0,3342	0,5322	19,3592
Rush.Cutaneo	-0,3234	-0,1752	-13,1333
Parestesias	0,2896	0,1714	9,9297
C.Inspec	0,2692	0,4152	-5,5296
Otros	-0,2585	-0,4206	-0,1924
F.Supervivencia	-0,2524	-0,8663	-5,1316
Deshidratacion	-0,2472	-0,0377	0,4631
Fiebre	0,2357	0,0777	0,3904
R.consumo	0,2347	0,3507	-18,5799
F.Proliferacion	0,2282	0,7113	5,9413

Continuación.

Variables	RIDGE	LASSO	flat
Diarrea	0,2136	0,3227	0,2503
P.Incubación	-0,1988	-0,4200	1,5997
L.elaboración	-0,1316	0,0068	1,2848
Otros.Neurológicos	0,1108	-0,3166	30,9819
Vomitos	0,0899	0,1374	1,4699
F.Contaminacion	0,0861	-0,0158	2,7351
Grupo.Sospechoso	-0,0189	-0,1905	0,2664

Anexo 11. Coeficientes de $\boldsymbol{\beta}$ para el caso de salmonella sin las variables de atención en el enfoque Bayesiano.

Tabla 3.28: Tabla de coeficientes de $\boldsymbol{\beta}$ para el caso Bayesiano. Caso 3: Solo salmonella y sin las variables de atención.

Variables	RIDGE	LASSO	flat
Dolores	-0,8822	-1,1327	-2,1685
Expuestos	0,8657	0,8605	5,8588
Nauseas	-0,5978	-0,8281	-1,8005
Duración.B	0,5183	0,7863	0,9001
Meteorismo	-0,3721	-0,6615	-1,3851
Cefalea	0,3447	0,5268	0,5792
CIE.10	0,3372	0,5203	0,4607
Tasa.Ataque	-0,3228	-0,5218	0,5377
Vomitos	0,2742	0,4882	1,8768
Deshidratacion	0,1990	0,2876	-0,0026
R.notificación	0,1952	0,2483	50,0083
L.elaboración	-0,1937	-0,3323	-2,4085
Semana	0,1912	0,2985	0,2639
R.consumo	0,1814	0,2646	-50,0297
Rush.Cutaneo	0,1680	0,2436	1,1553
P.P.I	-0,1521	-0,2439	-0,4620

Continuación.

Variables	RIDGE	LASSO	flat
P.Incubación	-0,1471	-0,3288	-0,9872
Mialgia	-0,1225	-0,1437	0,3690
L.P.I	0,1163	0,1553	-1,0180
Fiebre	0,1094	0,1739	0,7394
F.Contaminacion	-0,1008	-0,1179	0,0707
F.Proliferacion	-0,0992	-0,2240	0,0907
Hipotension	-0,0953	-0,1982	-0,2841
Otros.Neurológicos	-0,0921	-0,1660	-0,3537
Parestesias	0,0789	0,1027	-5,1325
L.consumo	0,0764	0,1310	1,2022
Otros	0,0598	0,1610	1,0369
F.Supervivencia	-0,0528	-0,1251	-0,8884
Espasmos	0,0501	0,2369	0,8919
Grupo.Sospechoso	-0,0463	-0,0932	-0,7594
C.Inspec	0,0298	-0,0106	-2,1148
Heces.Sang	-0,0197	-0,0633	-0,6061
Diarrea	-0,0181	-0,1089	-0,3363