ASOCIACIÓN DEL TIEMPO DE HOSPITALIZACIÓN FRENTE A VARIABLES SOCIODEMOGRÁFICAS, CLÍNICAS Y PARACLÍNICAS DE PACIENTES PEDIÁTRICOS CON INFECCIÓN POR VIRUS EPSTEIN BARR MEDIANTE MODELOS DE REGRESIÓN

Jorge Arturo Baquero Sánchez

01 de junio de 2022

Director de tesis:
MSc. Mario José Pacheco López

Facultad de Ciencias
Departamento de Matemáticas
Bogotá D.C., Colombia

ASOCIACIÓN DEL TIEMPO DE HOSPITALIZACIÓN FRENTE A VARIABLES SOCIODEMOGRÁFICAS, CLÍNICAS Y PARACLÍNICAS DE PACIENTES PEDIÁTRICOS CON INFECCIÓN POR VIRUS EPSTEIN BARR MEDIANTE MODELOS DE REGRESIÓN

Jorge Arturo Baquero Sánchez

Trabajo de grado presentado como requisito parcial para optar al título de Estadístico

Director de tesis:
MSc. Mario José Pacheco López

Universidad el Bosque
Facultad de Ciencias
Departamento de Matemáticas
Bogotá D.C., Colombia

Agradecimientos

A los médicos Maria Alejandra Moreno y Paula Carolina Gómez García por el apoyo en esta investigación.
Al profesor Mario por sus consejos a lo largo de la carrera.
A mis padres por el acompañamiento en este viaje.

Resumen

Recientemente, se han realizado trabajos de investigación sobre la veracidad de la literatura americana de la cual están basadas la gran mayoría de escuelas de medicina en latinoamerica, con el diagnóstico y evolución de enfermedades en cohortes de diferentes países. Un ejemplo de lo anterior, es el trabajo Moreno(2020) donde se caracterizan y difieren ciertos diagnósticos de la enfermedad causada por el virus Epstein Barr en una población pediátrica de una clínica de Bogotá, Colombia, entre los años 2015 y 2019. En el trabajo anterior, se identificó de forma descriptiva, un error en el diagnóstico a causa de esas diferencias con los parámetros de enseñanza, lo que genera ineficiencia en el uso de recursos en salud por tiempos de hospitalización prolongados.

A partir de lo anterior, en este trabajo se realizó una comparación de modelos de regresión que expliquen la asociación de las variables sociodemográficas, clínicas y paraclínicas de los pacientes con el número de días hospitalizados de la cohorte estudiada. Se ajustaron modelos con aproximación frecuentista y bayesiana, apoyados en la selección de variables por métodos como el Step AIC, evaluación de importancia por Random Forest o probabilidad de inclusión para el manejo de sobre ajuste.

Con los modelos ajustados, se identificaron variables como la edad, la presencia de mialgias, la trombocitosis, entre otras, que explican el tiempo de hospitalización de pacientes pediátricos con infección por virus Epstein Barr en la cohorte estudiada. Después de discutir los resultados obtenidos, se concluyó que se utilizarían la totalidad de las variables significativas generadas de los diferentes modelos propuestos ya que, por un lado, se complementan posibles falencias de unos modelos con los otros y, por otro lado, serán la base argumentada del siguiente estudio con una muestra representativa de la cohorte local.

Abstract

Recently, some studies are researching the veracity of the American literature, on which the vast majority of medical schools in Latin America are based, with the diagnosis and evolution of diseases in cohorts from different countries. One example is the work of Moreno (2020) which characterizes and differs in certain diagnoses of the disease caused by the Epstein Barr virus, in a pediatric population of a clinic in Bogotá, Colombia, between the years 2015 and 2019. With the previous work, a possible fault in the diagnosis was identified due to these differences with the teaching parameters, which generates inefficiency in the hospitalization times of the patients.

Therefore, a comparison of regression models that explain the association of the sociodemographic, clinical and paraclinical variables of the patients with the number of hospitalized days in the studied cohort was carried out. Models were made with a frequentist and Bayesian approach, supported by the selection of variables by Step AIC methods, evaluation of importance by Random Forest, or probability of inclusion for handling overfitting.

Variables such as age, presence of myalgia, and thrombocytosis, among others, that explain the hospitalization time of pediatric patients with Epstein Barr virus infection in the studied cohort were identified. After discussing the results obtained, it was concluded that all the variables generated from the different proposed models would be used since, on the one hand, possible shortcomings of some models are complemented with the others and, on the other hand, they will be the basis argued of the following study with a representative sample of the local cohort.

Índice general

Capítulo 1

Introducción

Actualmente existe variabilidad en la literatura médica sobre el diagnóstico, tratamiento y manejo de diferentes enfermedades que afectan a la población general. Esto se ve reflejado en las diferencias tanto clínicas como socioculturales de las cohortes de pacientes estudiados en países industrializados, como Estados Unidos (potencia en medicina basada en la evidencia), y países latinoamericanos (González Saldaña et al., 2012), como Colombia donde los algoritmos de manejo se basan en la literatura estadounidense. Esto manifiesta una problemática en torno a el desenlace heterogéneo de los pacientes y un golpe para los recursos en salud del país. Es por esto, que se ha incrementado la recolección y análisis de variables en las diferentes enfermedades que afectan a la población colombiana, con el objetivo de contrastar y generar criterios que se adapten a las características de nuestra población. En particular, se resalta la alta prevalencia de enfermedades infecciosas en un país tropical como Colombia, donde es importante el comportamiento local de los virus y bacterias responsables de enfermedad.

Un trabajo reciente es el de Moreno (2020), el cual hace un análisis descriptivo de variables clínicas y paraclínicas de pacientes entre el mes de vida y los 17 años de edad, con infección por Virus Epstein Barr (VEB por sus iniciales) durante el periodo de 2015 al 2019 de una clínica de Bogotá.

Este virus, del tipo herpes, afecta a la población general y provoca un espectro clínico inespecífico que puede llegar a confundirse con otras enfermedades infecciosas virales y bacterianas como el Dengue y la Faringoamigdalitis respectivamente, las cuales son más graves y requieren de más recursos en salud para su tratamiento. Esto finalmente afecta el presupuesto en salud del país, ya afectado por las condiciones políticas y culturales de Colombia, lo que conlleva a sistema de salud escaso en recursos que no logra cobijar a toda la población.

Este estudio tuvo como objetivo, a partir de la información del trabajo de Moreno (2020) generar un modelo de regresión estadísticamente significativo, que identificara los factores que explican la mayor variabilidad del tiempo de hospitalización de la cohorte estudiada en Bogotá.

Moreno (2020) concluye que sí hay diferencias entre las variables estudiadas de la población

objetivo y lo descrito en la literatura norteamericana. Esto se resalta en las conclusiones donde se narra "la necesidad de una mayor sospecha diagnóstica de la infección por VEB en la población estudiada. Lo que sugiere estudios adicionales para correlacionar los resultados encontrados" (Moreno, 2020). Es decir, que existe un vacío en el algoritmo diagnóstico de la infección por VEB en la población colombiana por falta de sospecha clínica. Esto teniendo en cuenta que muchos síntomas descritos en la literatura americana difieren de los presentados por los pacientes colombianos, dando como resultado un diagnóstico sesgado por la similitud con otras enfermedades; ya se nombró previamente la infección por el virus del Dengue, que requiere mayor tiempo de hospitalización y un gasto innecesario en los recursos de la salud.

Dado lo anterior, es necesario desarrollar un modelo asociativo de las variables analizadas en el estudio descriptivo de Moreno (2020), tipo modelos de regresión, con el propósito de sentar nueva información argumentada de la enfermedad por VEB en la población colombiana y así, aportar datos base a la comunidad científica que permita realizar estudios y trabajos posteriores con mayor peso estadístico para poder estandarizar el algoritmo de diagnóstico de la enfermedad por VEB y con esto ejercer un manejo eficiente de recursos en el sistema de salud.

Capítulo 2

Objetivos

2.1. Objetivo general

Ajustar un modelo de regresión para el número de días hospitalizados de pacientes de 1 mes a 17 años de vida con infección por virus Epstein Barr, durante el periodo 2015 al 2019 en una clínica Infantil de Bogotá, en función de variables sociodemográficas, clínicas y paraclínicas.

2.2. Objetivos específicos

- Escoger coherentemente las variables explicativas iniciales de la base de datos.
- Identificar modelos candidatos según la naturaleza de la variable y del contexto médico.
- Comprobar que el modelo escogido tenga validez estadística.
- Explicar el modelo seleccionado desde una perspectiva estadística y médica.

Capítulo 3

Antecedentes

Debido a la reciente necesidad de evaluar los estándares de la literatura americana frente a la evolución y tratamiento de las cohortes locales, no existen estudios que describan las características sociodemográficas, clínicas y paraclínicas de la infección por VEB en Colombia. Es por esto, que el principal trabajo de apoyo fue el de Moreno (2020). Sin embargo, se realizó una búsqueda sistemática de la literatura en bases de datos, donde se encontraron estudios de diferentes países que están elaborando este tipo de análisis en sus respectivas cohortes, de los cuales se resaltan los siguientes trabajos:

- Son y Shin (2011) en el estudio "Clinical features of Epstein-Barr virus-associated infectious mononucleosis in hospitalized Korean Children", realizaron un estudio similar en una cohorte de Corea del Sur. Por medio de análisis de asociación Chi cuadrado y kruskall-Wallis test, identificaron que el virus infecta pacientes en edades menores a 10 años. La incidencia de dolores de cabeza, elevación de enzimas del hígado y la duración de la fiebre, estaban asociadas a la edad en la cual se infecta el paciente.

- Garcia-Peris et al. (2018) en el estudio "Primo infección por el virus de Epstein Barr en niños sanos de Valencia, España", también encontraron diferencias entre la literatura de países desarrollados y la cohorte estudiada en España. Donde con ayuda de análisis de asociación mediante test de Fisher y Kruskall- Wallis, concluyen que el porcentaje de anticuerpos heterófilos positivos ha sido bajo en le muestra del estudio. Así mismo la primoinfección por VEB es frecuente en niños de menor edad, y en ellos predominan las formas oligosintomáticas y fue común detectar coinfección con otros virus.

- Gao et al. (2011) en el estudio "Epidemiologic and clinical characteristics of infectious mononucleosis associated with Epstein Barr virus infection in children in Beijing, China", realizaron un estudio en 418 pacientes con infección por VEB para encontrar las características y asociaciones del virus con la mononucleosis infecciosa; y se encontró que las mayores manifestaciones del virus en los niños estudiados, eran fiebre, linfadenopatías y faringitis. En niños menores de 6 años, la incidencia de hepatomegalias, esplenomegalias y sarpullido fue mayor que en los

niños menores de 6 años. Los estudios de asociación fueron realizados a través de pruebas Chi cuadrado y ANOVAs.

- Topp et al. (2015) en el estudio "Clinical characteristics and laboratory findins in Danish children hospitalized with primary Epstein-Barr virus infection", realizado en Copenhagen, encontraron que el grupo más afectado por el virus fueron niños entre 1 y 2 años de edad, seguido por adolescentes (14-15 años de edad). Las variables clínicas más relacionadas con el virus fueron fiebre, linfadenopatía cervical, amigdalitis y fatiga. Estos resultados fueron analizados con test Chi cuadrado.

- Wu et al. (2020) en el estudio "Clinical manifestations and laboratory results of 61 children with infectious mononucleosis" investigaron sobre las manifestaciones clínicas de la infección en niños de Shanghái. En este estudio, por medio de análisis no paramétrico Kruskall-Wallis y Chi cuadrado de independencia, se encontró que la enfermedad afectó a niños de temprana edad y niños preescolares. Donde los síntomas más asociados fueron fiebre, amigdalitis, linfadenopatías cervicales y hepatomegalia.

- González Saldaña et al. (2012) en el estudio "Clinical and laboratory characteristics of infectious mononucleosis by Epstein-Barr virus in Mexican children", el más cercano frente a la caracterización y correlación de las características de la enfermedad por VEB en una cohorte pediátrica de México. En este estudio, se describe una mayor incidencia en pacientes más jóvenes con respecto a los pacientes descritos en la literatura americana (donde los adolescentes son el grupo más afectado). Lo mismo sucede, con el orden de manifestaciones clínicas, los cuales difieren de cohortes norteamericanas.

- Por otro lado, existen estudios donde se utilizan herramientas más avanzadas como los modelos lineales generalizados para explicar el tiempo hospitalización. Sin embargo, son estudios diferentes a la enfermedad en cuestión, donde resaltan que para encontrar asociación de variable de conteo por modelos de regresión, es importante realizar la escogencia adecuada, ya que la regresión Poisson, puede tener limitaciones cuando se genere sobredispersión, exceso de ceros en los datos o ambos casos combinados (Weaver et al., 2015).

Finalmente, se resaltan 2 conclusiones importantes de los estudios anteriores. La primera, que es notable cómo la edad de incidencia y la presentación clínica varía según la cohorte estudiada frente a la establecida en literatura americana, por lo tanto, es importante identificar estas características y asociaciones de la población colombiana, para poder establecer una base en futuros estudios. La segunda, que existe un vacío en el análisis estadístico de métodos más elaborados para el estudio de las asociaciones de VEB y sus manifestaciones o complicaciones. La mayoría de estudios se realizaron con análisis no paramétricos o análisis univariados en ciertos casos. Lo anterior, puede generar ineficiencia en el análisis o resultados poco verídicos por el análisis variable a variable en bases de datos que cuentan con un mayor número de factores que de observaciones.

Capítulo 4

Marco referencial

4.1. Marco teórico

De la manera más básica, se entiende la regresión linear simple como una herramienta para modelar la relación de dos variables, una respuesta y una explicativa. Para lo anterior se utiliza un modelo de la forma $y = \beta_0 + \beta_1 x + \varepsilon$, en donde la variable "y" se explica dado un punto base β_0 y un múltiplo de un valor en relación a la segunda variable β_1 que son estimados bajo ciertos supuestos de independencia entre variables y un manejo de un error que se produce por información que no se tiene en cuenta en el modelo. Para adicionar más factores, se utiliza la regresión lineal múltiple, la cual toma en cuenta más de una variable predictiva y su modelo se presenta de la forma

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Los parámetros β_i, son denominados coeficientes de regresión y la linealidad del modelo recae en el exponente de estos mismos. Con ayuda de estos modelos de regresión, podemos estimar valores de y basados en datos de nuevos individuos con los mismos parámetros evaluados y modelado descriptivo de la asociación de la variable respuesta frente a sus variables explicativas (Rencher y Schaalje, 2008).

Los modelos de regresión que más se enseñan en la academia están basados en variables de respuesta cuantitativas y estos, se rigen bajo cuatro supuestos importantes sobre los errores: normalidad, varianza constante, independencia y sobre las variables explicativas, ausencia de multicolinealidad (Rencher y Schaalje, 2008). Sin embargo, cuando estos supuestos no se cumplen o la naturaleza de la variable respuesta es de otro tipo, se pueden usar otro tipo de modelos como los modelos lineales generalizados.

Estos modelos lineales generalizados, son modelos que explotan la linealidad de diferentes variables debido a la unificación de diferentes técnicas estadísticas donde la base general, radica en el comportamiento de la variable respuesta, transformaciones lineales o familias de distribución entre otras (McCullagh y Nelder, 1983).

Para variables respuesta de conteo como "días de hospitalización", se pueden utilizar modelos

regresión Poisson, los cuales están dados por la distribución de probabilidad

$$f(y) = \frac{\mu^y e^{-\mu}}{y!}$$

$$y = 0, 1, 2, ...$$

El modelo de regresión Poisson está dado por $y_i = E(y_i) + \varepsilon_i$, con $i = 1, 2, ..., n$, donde y_i son variables aleatorias independientes con distribución Poisson y $\mu_i = E(y_i)$ es una función de

$$g(x_i'\beta) = \beta_0 + \beta_1 x_i 1 + ... + \beta_k x_i k$$

Algunas funciones de enlace usuales de $x_i'\beta$, donde β está dado por el vector de parámetros $\beta = (\beta_o, \beta_i)^t$ son $\mu_i = x_i'\beta$, $\mu_i = e x_i'\beta$, $\mu_i = ln(x_i'\beta)$

Para cualquier función de enlace, los valores de μ_i deben ser positivos y para estimar β, usualmente se usa el método de máxima verosimilitud con la distribución Poisson dada por la siguiente ecuación:

$$\beta'^{(m+1)} = (X'WX)^{-1}X'W\left[X\beta^{(m)} + \frac{y_i - \mu_i}{\mu_i}\right]$$

(Rencher y Schaalje, 2008).

Cuando se cuenta con un número considerable de variables explicativas, la escogencia de los modelos se puede realizar mediante la selección previa de variables, en donde se evaluan todas las posibles combinaciones de los predictores y se identifica el modelo con el criterio Akaike menor, es decir el modelo que con menor número de predictores explique la mayor variabilidad (Amat, 2016)

El estadístico *Akaike Information Criterion* (AIC) se aplica en modelos ajustados mediante la máxima verosimilitud. Este criterio es de la forma

$$AIC = 2k - 2ln(\hat{L})$$

donde RSS es la suma del cuadrado de los residuales. Para regresión lineal por mínimos cuadrados el valor es proporcional a otros estadísticos como el Cp de Mallows por lo que seleccionan el mismo modelo (McCullagh et al., 1989). El algoritmo contiene la limitación del requerimiento computacional, por lo tanto se utilizó una selección híbrida (adición y sustracción simultanea de predictores) para disminuir el tiempo de ejecución.

De manera similar, se cuenta con el criterio de información Bayesiano (BIC) el cuál, está altamente relacionado con el criterio de información AIC, pero desde una perspectiva Bayesiana. Este criterio es de la forma

$$BIC = kln(n) - 2ln(\hat{L})$$

Otro método de selección de variables utilizado desde la perspectiva de aprendizaje de máquina es la evaluación de importancia de variables. Este método se concentra en realizar simulaciones de

árboles de decisión para disminuir la varianza y el sobre ajuste de los mismos por medio de modelos lineales múltiples kuhn (2019). Para lo anterior, se generan nuevas bases de datos muestreando las observaciones de la base de datos original con reemplazo. Esto genera bases de datos simuladas por cada variable, donde por medio de bootstrap, se realiza un ranking de variables de mayor asociación a la variable respuesta según la precisión de predicción del árbol de decisión en esa configuración con la información no seleccionada de la muestra inicial de la base de datos. Se realiza el mismo paso con diferentes permutaciones y se promedia el valor de todas las simulaciones. Por último, se registra la diferencia de este poder predictivo con el promedio de las simulaciones y se genera un ranking de importancia estandarizado a 100 para mejor lectura (kuhn, 2019).

Finalmente, desde la perspectiva Bayesiana, se puede evaluar un modelo de regresión a partir del kernell de una distribución Poisson, la cual está dada por la forma:

$$f(y|\mu) = \prod_{i=1}^{n} \frac{\mu^y e^{-\mu}}{y!}$$

$$= exp\left\{-n\mu + n\bar{y}ln(\mu) - \sum_{i=1}^{n} ln(y!)\right\}$$

$$\propto exp\left\{-n\mu + n\bar{y}ln(\mu)\right\}$$

Por lo tanto, para la regresión $y_i|x_i$ se distribuye Poisson con $\mu_i = exp(x'\beta)$, dada por la densidad apriori $\pi(\beta)$ lo que generaría:

$$p(\beta|y, X) \propto exp\left(\sum_{i=1}^{n}[-exp(x_i'\beta) + y_i x_i'\beta]\right) \cdot \pi(\beta)$$

donde $\pi(\beta)$ es la distribución a priori, que en este caso se utilizó la distribución no informativa de Jeffrey dada por la ecuación:

$$J(\theta) = -E\left(\frac{d^2 log(f(y|\theta))}{d\theta}\right)$$

La distribución posterior, sale de la multiplicación la probabilidad a priori por el valor de verosimilitud de los datos, para después ser dividido por la probabilidad de los datos observados, tal cual, como muestra se estructura el teorema de Bayes (Castellano, 2015).

El método bayesiano se apoya de los métodos de simulación para poder generar mayor estabilidad en las distribuciones resultantes. Es por esto, que se apoya en el método de cadenas de Markov de Monte Carlo(MCMC) para generar muestras de la distribución posterior. Este método se basa en la convergencia del algoritmo, donde las Cadenas de Markov muestran que cada valor agregado solo depende del valor anterior generado (Marín et al., 2007).

4.2. Marco conceptual

El Virus Epstein Barr (VEB) pertenece a la superfamilia de los virus del herpes y es también conocido como el virus del herpes humano tipo 4, que se transmite a través de secreciones corporales como la saliva (más frecuentemente) y entre personas que se encuentran en contacto estrecho (Sullivan & Luzuriaga, 2017). Particularmente, puede generar enfermedad en personas que adquieren la infección por primera vez (primo-infección). Sin embargo, este virus puede vivir en el cuerpo de forma inactiva (latente) una vez resuelve la enfermedad aguda, aunque el individuo ya no lo propague. Este aspecto es importante porque, en algunos casos, el virus puede reactivarse y causar síntomas. En este último caso, el virus puede volverse a propagarse desde el individuo infectado a otros sanos.

La infección causada por el VEB tiene un amplio espectro de síntomas que varían según el grupo de edad. Los niños menores de 4 años y adultos mayores de 30 años suelen ser asintomáticos, siendo los adolescentes y adultos jóvenes los más afectados; esto según datos en países industrializados. Los niños pequeños con primo-infección pueden presentar síntomas inespecíficos e indistinguibles de otras enfermedades infecciosas frecuentes en la niñez. Por otro lado, durante la adolescencia y la adultez la infección se manifiesta como una mononucleosis infecciosa (MNI) en aproximadamente en un tercio de los casos(Yang et al., 2015).

Ésta entidad clínica (MNI) es causada por el VEB en más del 90 % de los casos (González Saldaña et al., 2012). Se caracteriza por la triada clínica de: fiebre, adenomegalias (aumento del tamaño de ganglios) y faringitis; aunque también puede estar asociada a otros síntomas como fatiga, ictericia (coloración amarilla de la piel y mucosas), hepato-esplenomegalia (aumento significativo del tamaño del hígado y bazo), erupciones cutáneas, entre otros (Moreno, 2020).

La prevalencia de la infección por VEB varía según la edad, la localización geográfica y la etnia (Dunmire et al., 2018). En la mayoría de países subdesarrollados se considera que el 90-95 % de los adultos ya tienen anticuerpos medibles en sangre contra el VEB, lo que se refleja en el pico de incidencia de la de la primo-infección, que es mayor antes de los 3 años de vida (Yang et al., 2015).

En países desarrollados como Estados Unidos, los niños adquieren la infección de forma más tardía durante la adolescencia, donde la presentación típica de la infección es la MNI (Balfour et al., 2015). Se reportan aproximadamente 500 casos al año de MNI por 100.000 habitantes en E.E.U.U, con mayor incidencia en pacientes entre los 15 y 24 años (González Saldaña et al., 2012). Sin embargo, existen discrepancias en cuanto a la edad de presentación, ya que en cohortes estudiadas de México y Corea, la enfermedad suele ser más frecuente durante la infancia entre los 2 y 6 años de vida (Son & Shin, 2011).

La MNI es una enfermedad autolimitada que resuelve espontáneamente en dos o tres semanas de evolución (Losa García et al., 1998). Existe gran variabilidad clínica a nivel mundial, lo que se ve reflejado en el porcentaje de presentación de los síntomas comunes como la fiebre prolongada, que se presenta en 28 al 95 %, la faringitis presentada en 50 al 85 % y las linfadenopatías en 72 al 100 % de los casos (Moreno Bermeo, 2020).

Otras manifestaciones como la erupción de la piel, llamada rash, generalmente está asociada a la

administración de antibióticos específicos (derivados de penicilina) secundario a alergia transitoria por mecanismos inmunes propios de la infección aguda; esto ocurre en el 95 % de las personas en las cohortes estudiadas (Likic & Kzmanic, 2004). Las alteraciones en la sangre y la inflamación del hígado se presentan en el 75 % de los casos; y la hepatitis con aumento del tamaño hepático e ictericia se reportan en 5 % de los casos, entre otros (Losa García et al., 1998).

Actualmente, no hay tratamiento estandarizado para la infección (Balfour et al., 2015). El manejo de esta entidad consiste principalmente en tratar los síntomas. Se debe iniciar manejo para el control de la fiebre, manejo para el control del malestar general, dolor de garganta y dolor muscular. Médicos recomiendan el reposo durante las primeras 6 semanas posterior a la infección aguda por riesgo de ruptura del bazo, debido a su gran tamaño (urgencia médica) (Balfour et al., 2015).

4.3. Marco legal

Este trabajo, se acoge a políticas de privacidad de la información de la clínica de la cual se extrajo la información de los pacientes analizados. Adicionalmente, se acoge a la resolución 34 8430 de 1993 del Ministerio de Salud de Colombia. Esta resolución, establece la investigación biomédica en seres humanos debe estar realizada solo por personas científicamente calificadas bajo supervisión de un profesional médico competente (art. 395) (Ministerio de Salud, 1993).

Anexo a esto, este trabajo se acoge a las normas de la declaración de Helsinki- 64 Asamblea General, Fortaleza, Brasil, octubre 2013. Estas normas, garantizan la protección de datos recolectados para resguardar la privacidad del individuo (Helsinki, 2013). Lo anterior, se ve reflejado en la omisión de variables de identificación, los cuales no son relevantes en este estudio.

Capítulo 5

Metodología

Esta investigación cuantitativa se basa en los datos recolectados del estudio observacional descriptivo de serie de casos de Moreno (2020), los cuales fueron recolectados de la población accesible de paciente de 1 mes a 17 años de vida con diagnóstico por infección del virus Epstein Barr de una clínica de Bogotá durante los años 2015-2019.

El estudio de Moreno (2020) al ser retrospectivo por conveniencia, no manejó una muestra, sino la población con la característica de inclusión (pacientes positivos de IgM para antígeno capsular (Ag VCA) del Virus del Epstein Barr de la clínica en el periodo anteriormente mencionado.

La base de datos inicial contaba con 91 observaciones y 101 variables, de las cuales se eliminaron variables bajo 3 criterios principales:

- Información sensible o irrelevante para el estudio de la asociación. Un ejemplo de esto, son las variables ID, nombres, fechas, etc.

- Exámenes propuestos a investigar que no se realizaron según el historial clínico analizado.

- Variables de segundo nivel frente a la población estudiada. Es decir, varaibles que componen una variable principal, por ejemplo: Bilirrubina total se compone de Bilirrubina directa más bilirrubina indirecta, por lo tanto se tiene en cuenta la interpretación de la Bilirrubina total según grupo etario.

Cabe aclarar, que las reglas de filtro y la consolidación del grupo de variables finales, se realizaron bajo la supervición de las médicas que acompañaron el proceso.

Por lo tanto, la base de datos a trabajar, cuenta con 91 observaciones, 36 variables. Estas variables están divididas en 5 grupos que conforman parte del proceso médico en relación a los niños infectados. El primer grupo de variables (5 variables), describe características sociodemográficas de los niños analizados. El segundo grupo de variables (14 variables), habla sobre factores clínicos de los niños, es decir, síntomas con los que llegan al día de la consulta con el médico. El tercer grupo de variables (8 variables), son de tipo paraclínico, es decir, resultados de los exámenes que el médico requirió del paciente. El cuarto grupo de variables (7 variables), son de tipo desenlace del paciente en hospitalización. Por último, el quinto grupo de variables (2 variables), son de tipo de complicaciones o desenlace presentado por el paciente. Para este estudio, la última variable "Número de días hospitalizado"se escogió como la variable dependiente para los modelos a trabajar.

Variables sociodemográficas

- *Edad:* Categórica por grupo etário (Lactante menor, Lactante mayor, Preescolar, Escolar y Adolescente).
- *Sexo:* Categórica por condición biológica y genética que define a un individuo (Masculino, Femenino).
- *Nivel educativo del menor:* Categórica por nivel educativo actual del menor (Ninguno, Guardería, Jardín, Primaria o Secundaria).
- *Comorbilidades:* Categórica por presencia de condiciones patológicas previas a la consulta (si, no, desconocido).
- *Antibiótico 7 días previos a la consulta:* Categórica por administración de antibióticos la semana previa a la consulta del paciente (si, no, desconocido).

Variables clínicas

- *Primer síntoma:* Categórica por primer síntoma manifestado de la infección por VEB (fiebre,malestar general, emesis, odinofagia, dolor abdominal, rinorrea, tos, adenopatías, otros).
- *Malestar general:* Categórica por presencia de malestar general al momento de la consulta (si, no, desconocido).
- *Mialgias:* Categórica por presencia de mialgias al momento de la consulta (si, no, desconocido).
- *Tiempo de fiebre:* Continua discretizada por número de días desde el inicio de la fiebre hasta la consulta.
- *Cervicales:* Categórica por presencia de aumento de tamaño de ganglios cervicales (si, no, desconocido).
- *Abscedación adenopatía cervical:* Categórica por crecimiento asimetrico de adenopatías con presencia de imágenes sujestivas de absceso por ecografía (si, no, desconocido).
- *Linfadenopatías generalizadas:* Categórica por presencia de adenopatías en más de una región diferente a la cervical (si, no, desconocido).
- *Faringitis:* Categórica por presencia de exudado faringeo al examen físico (si, no, desconocido).
- *Esplenomegalias:* Categórica por presencia de aumento de tamaño del bazo al examen físico (si, no, desconocido).
- *Hepatomegalias:* Categórica por presencia de aumento de tamaño del hígado al examen físico (si, no, desconocido).
- *Edema periorbital:* Categórica por presencia de edema periorbital al examen físico (si, no, desconocido).
- *Ictericia:* Categórica por presencia de ictericia al examen físico (si, no, desconocido).
- *Enantema palatino:* Categórica por presencia de enantema en paladar al examen físico (si, no, desconocido).

- *Exantema:* Categórica por presencia de exantema maculopapular o petequial (si, no, desconocido).

Variables paraclínicas

- *Interpretación leucocitos:* Categórica por número de leucocitos en el cuadro hemático ingreso células/mm^3 según grupo etario (normal, leucopenia, leucocitosis).
- *Interpretación neutrófilos:* Categórica por número de neutrófilos en el cuadro hemático ingreso células/mm^3 según grupo etario (normal, neutropenia, neutrofilia).
- *Interpretación linfocitos:* Categórica por número de linfocitos en el cuadro hemático ingreso células/mm^3 según grupo etario (normal, linfopenia, linfocitosis).
- *Linfocitos atípicos:* Continua por prensencia de linfocitos atípicos en sangre %.
- *Interpretación monocitos:* Categórica por número de monocitos en el cuadro hemático ingreso células/mm^3 según grupo etario (normal, monopenia, monocitosis).
- *Interpretación plaquetas:* Categórica por número de plaquetas en el cuadro hemático ingreso células/mm^3 según grupo etario (normal, trombocitopenia, trombocitosis).
- *Interpretación PCR:* Categórica por proteína c reactiva según grupo etario (normal, elevada, desconocida).
- *Ecografía abdominal total:* Categórica por exámen de ecografía abdominal total (si, no).

Variables de evolución

- *Número de antibióticos utilizados:* Discreta por número de antibióticos utilizados durante la estancia de hospitalización del paciente.
- *Número de días totales con antibiótico:* Continua discretizada por el número de días totales en que el paciente recibió antibiótico, durante su hospitalización.
- *Infección bacteriana documentada:* Categórica por tipo de infección bacteriana presentada durante la hospitalización (Sobreinfección cavidad oral, Absceso periamigdalino, OMA, Sinusitis, adenitis abscedada, Neumonía, EDA bacteriana, otra).
- *Valoración por oncología:* Categórica por evaluación de oncología durante la hospitalización (si, no).
- *Sospecha síndrome linfoproliferativo:* Categórica por sospecha de patología oncológica (si, no).
- *TAC:* Categórica por realización de algún tipo de tomografía (si, no).
- *Aspirado de médula ósea:* Categórica por realización de aspirado de médula ósea por sospecha de patología (si, no).

Variables de desenlace

- *Neumonía:* Categórica por presentar complicación de neumonía como complicación del cuadro agudo evidencia durante la hospitalización.
- *Número de días hospitalizado:* Continua discretizada por el número de días totales que el paciente duro hospitalizado.

Esta investigación, al ser una continuidad del trabajo descriptivo de Moreno (2020), se centró en la parte asociativa de las variables bajo métodos de regresión. Para esto, se realizaron 4 modelos de regresión lineal (lineal múltiple, Poisson con selección Step AIC, Poisson con selección por Random Forest y regresión bayesiana con selección por probabilidades de inclusión) que se ajustaron a la variable dependiente de número de días hospitalizado y se comprobaron sus supuestos y significancia estadística con ayuda del software R y Rstudio.

Para la selección de variables por el método AIC, se utilizó la función *StepAIC* de la librería "MASS". Después de seleccionar el mejor modelo, se utilizó el resultado para ajustar un modelo lineal múltiple y otro modelo de regresión Poisson. Para la evaluación de importancia por Random Forest, se utilizó la función *VarImp* del paquete RandomForest y después, se ajustó un modelo de regresión Poisson con las 10 variables con mayor importancia. Finalmente, para la selección de variables del método bayesiano, se utilizó la función *bas.glm* donde arrojó un modelo con las variables de mayor probabilidad de inclusión para el modelo bayesiano Poisson. Después, con las variables seleccionadas se ajustó un modelo de regresión Poisson con la función *stan.glm* del paquete rstanarm.

Después de analizar y escoger los modelos con mejor desempeño estadístico, se presentaron los resultados de las variables independientes de mayor asociación al grupo médico que apoya la investigación sin decirles el método de regresión utilizado (estudio ciego), para así, poder escoger el grupo de variables con mayor significancia estadística y médica para la fase en evaluación del proyecto de investigación.

Capítulo 6

Resultados

6.1. Modelo de regresión lineal múltiple

Primero, se generó un modelo lineal normal para la variable respuesta *Número de días de hospitalización*, con selección híbrida por pasos de variables independientes bajo el criterio AIC. Esta algoritmo filtró a 25 variables explicativas con un AIC de 251.69. Después, se ajustó el modelo de regresión lineal múltiple con las variables anteriormente nombradas y se realizó la comprobación de los supuestos, los cuales dieron los siguiente resultados:

- *Test Shapiro-Wilks:* Se utilizó esta prueba para comprobar la normalidad de los residuales del modelo ajustado. Este test generó un p valor de 0.6686, con hipótesis nula de distribución diferente a la normal. Es decir, los residuales del modelo propuesto se distribuyen normalmente.
- *Test Breusch-Pagan:* Esta prueba se utilizó para evaluar la varianza constante del modelo ajustado. Este test generó un valor p de 0.5954, con hipótesis nula de varianza no constante. Es decir, los residuales del modelo propuesto presentan homocedasticidad.
- *Test Durbin-Watson:* Este test se realizó para evaluar el supuesto de independencia del modelo. Este test generó un valor p de 0.864, con hipótesis nula de autocorrelación. Es decir, los residuales del modelo propuesto presentan independencia.
- *Factores de inflación (VIF):* Esta prueba busca multicolinealidad entre las variables independientes. En el resultado, ninguna variable supero el valor límite de múlticolinealidad de 10. Por lo tanto, no existe evidencia estadística significativa de multicolinealidad en las variables anteriormente seleccionadas para el modelo.

Los resultados anteriores se pueden identificar en el Cuadro 6.1

	Shapiro-Wilks	Breusch-Pagan	Durbin-Watson	Factores inflación (VIF)
Valor p	0.6686	0.5954	0.864	<10
Cumple supuesto	Sí	Sí	Sí	Sí

Cuadro 6.1: Supuesto m. lineal múltiple

Además de comprobar los supuestos, se identificaron las siguientes variables estadísticamente significativas que explican el número de días hospitalizado por infección VEB, según un modelo lineal múltiple:

- *Mialgias*
- *Interpretación leucocitos*
- *Número de atibióticos*
- *Aspirado de médula ósea*
- *Neumonía*

Sin embargo, al revisar el ajuste del modelo, las variables independientes solo explican el 48.2 % de la variabilidad del *Número de días hospitalizados*, lo que índica que el modelo no explica eficientemente la variable respuesta. Adicionalmente, la naturaleza de la variable respuesta de estos modelos son para variables continuas, por lo tanto, este modelo quedó descartado para el objetivo del trabajo.

6.2. Modelo de regresión Poisson

Al tener una variable dependiente continua discretizada, como número de días hospitalizado por VEB, un modelo candidato óptimo es el modelo de regresión Poisson. Este modelo tiene encuenta la naturaleza de la variable y por medio de una función de enlace, que en este caso es el logaritmo, logra generar un modelo que se ajusta mejor a los datos.

En modelos lineales generalizados, existen otros modelos que se ajustan a este tipo de variables de conteo como el modelo de regresión binomial negativo, el modelo de regresión Poisson cero inflado y el modelo de regresión binomial negativo cero inflado. Sin embargo, la variable dependiente no presenta ceros en sus registros, por lo que no se toman los modelos cero inflados. Por otro lado, el modelo de regresión binomial negativo se utiliza cuando la variable respuesta de conteo presenta sobredispersión. Para el caso puntual de los datos, se realizó el test de sobredispersión del paquete AER de R, el cual dió como resultado con un p valor de 1, que los datos no presentan sobredispersión. Dado lo anterior, solo se ajustó el modelo Poisson.

Para ajustar este modelo, se utilizó la selección de variables híbrida de la función stepAIC, la cual generó 18 variables explicativas. Una vez se ajustó el modelo, se calculó el valor p de la

Valor p Deviance	Pseudo R2	V. atípicos P/D	AIC
0.00	0.6737	4/3	480.51

Cuadro 6.2: Modelo Poisson

Sociodemográficas	Clínicas	Paraclínicas	Evolución	Desenlace
Edad	Mialgias Tiempo de fiebre Cervicales Hepatomegalias	Inter. Leucocitos Inter. Neutrófilos Inter. Plaquetas	No. Días antibiótico Infección bacteriana Aspirado médula ósea	Neumonía

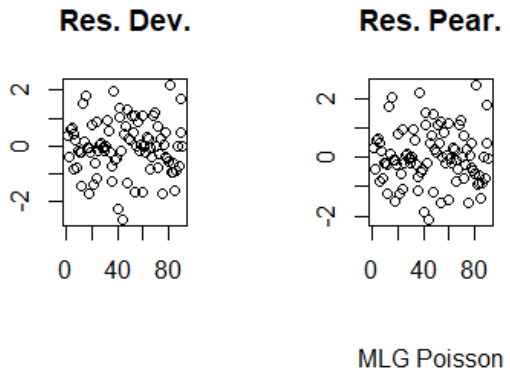
Cuadro 6.3: Variables significativas modelo Poisson

deviance. Este valor se contrasta con la hipótesis nula donde el modelo no se ajusta a los datos. Con un p valor de $2{,}15e^{-11}$, existe evidencia estadística que los datos se ajustan a un modelo lineal generalizado con variable respuesta Poisson.

Adicionalmente, se calculó el pseudo R^2 ajustado, el cuál con un valor de 0.6737 muestra que existe una muy buena explicación del número de días hospitalizados por VEB, en términos de las variables explicativas anteriormente nombradas, con un modelo lineal generalizado tipo Poisson.

Después, se comprobó la significancia de los betas estimados, los cuales mostraron las variables significativas (valor p<0.05) del modelo en el cuadro 6.3.

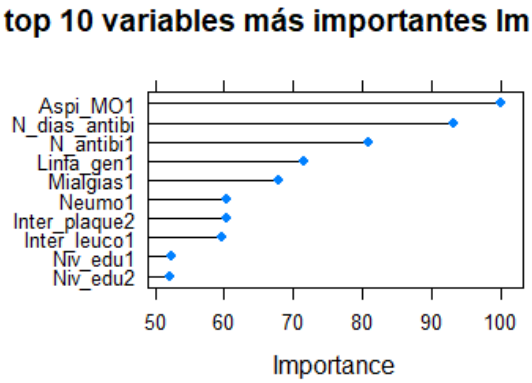
Finalmente, se comprobaron posibles valores atípicos en los registros, los cuales según el criterio de la deviance existen 3 registros atípicos y por el criterio de Pearson existen 4. En general, los valores se encontraban bastante cerca a la frontera del criterio de decisión (-2,2), lo que generó sospecha si los registros realmente se consideran atípicos. Al presentar estos registros al experto, no identificó algún detalle o variable característico para resaltarlos. Por lo tanto, no se consideraron registros atípicos desde la perspectiva técnica. Lo anterior, se evidencia en la imagen de residuales del modelo Poisson.



6.3. Random Forest

Otra manera de afrontar el sobreajuste, es desde una perspectiva de machine learning con la herramienta de evaluación de importancia de variables por random forest. Esta técnica, utiliza métodos de boostrap para simular diferente árboles de decisión y así poder disminuir la varianza y el sobre ajuste (Kuhn, 2019).

Se ajustó un modelo con la variable respuesta *Número de días hospitalizado* en función de todas las variables independientes, utilizando la función train del paquete Caret, con el método de random forest por asociación de modelo lineal. Después, con ayuda de la función varImp del mismo paquete, anteriormente nombrado, se generó una evaluación de las variables de mayor importancia frente a la variable respuesta. Las 10 variables más importantes se ven en la siguiente gráfica:



Según el análisis de la evaluación de importancia por Random Forest, se identificó el aspirado de médula ósea como la variable más influyente en el número de días hospitalizado. La segunda variable más importante es el número de días con antibiótico, seguido del número de antibióticos administrados al paciente, luego linfadenopatías genéticas, mialgias, neumonía, plaquetas altas, leucocitosis y nivel educativo respectivamente.

Una vez realizada la evaluación de importancia de variables, se ajustó un modelo lineal generalizado Poisson con estas 10 variables más importantes para el número de días hospitalizados por random forest. El valor p de la deviance con un valor de 0.00842, muestra que existe evidencia estadística que los datos se ajustan a un modelo lineal generalizado con variable respuesta Poisson.

Adicionalmente, se calculó el pseudo R^2, el cuál con un valor de 0.4445 muestra que existe una muy buena explicación del número de días hospitalizados por VEB, en términos de las variables explicativas anteriormente nombradas, con un modelo lineal generalizado tipo Poisson. Lo anterior, se evidencia en el Cuadro 6.4.

Valor p Deviance	Pseudo R2	V. atípicos P/D	AIC
0.00842	0.4445	11/9	505.3

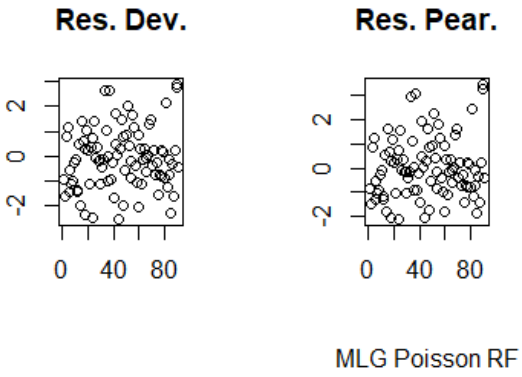
Cuadro 6.4: Modelo Poisson selección Random Forest

Sociodemográficas	Clínicas	Paraclínicas	Evolución	Desenlace
	Mialgias	Inter. Leucocitos Inter. Plaquetas Linfadenopatías generalizadas	No. Días antibiótico Aspirado médula ósea	Neumonía

Cuadro 6.5: Variables significativas modelo Poisson selección Random Forest

Después, se comprobó la significancia de los betas estimados, los cuales mostraron que las variables significativas (valor p<0.05) que aparecen en el Cuadro 6.5.

Finalmente, se comprobaron posibles valores atípicos en los registros, los cuales según el criterio de la deviance existen 9 registros atípicos y por el criterio de Pearson existen 11. De igual manera al modelo realizado con stepAIC, los valores atípicos se encontraban bastante cerca a la frontera del criterio de decisión (-2,2), lo que generó sospecha si los registros realmente se consideran atípicos.



6.4. Modelo de regresión Bayesiano

El cuarto y último modelo de regresión, se realizó bajo una aproximación bayesiana. Se utilizó la función bas.glm del paquete Bayesian Adaptative Sampling(BAS), la cual permite seleccionar los mejores 5 modelos dada las probabilidades de inclusión en la comparativa de modelos por menor AIC de modelos poisson con función de enlace Logarítmica (Clyde et al., 2010).

Valor p Deviance	Pseudo R2	V. atípicos P/D	AIC
0.00	0.45	8/4	486.9

Cuadro 6.6: Modelo Poisson Bayesiano

Sociodemográficas	Clínicas	Paraclínicas	Evolución	Desenlace
Edad	Primer síntoma	Inter. Plaquetas	No. Días antibiótico Aspirado médula ósea	Neumonía

Cuadro 6.7: Variables significativas modelo Poisson Bayesiano

Una vez realizado este complejo proceso computacional, se identificó el mejor modelo construído por las variables edad, primer síntoma presentado, interpretación de plaquetas, número de días con antibiótico, infección bacteriana, aspirado de médula ósea y neumonía.

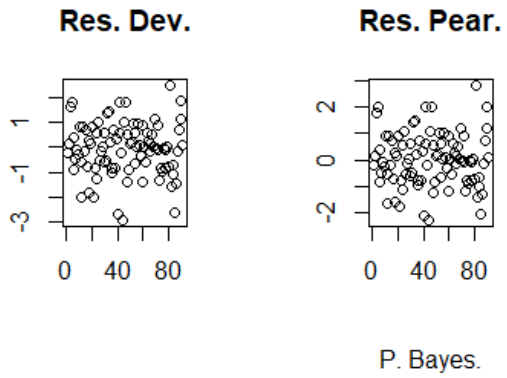
En el paso siguiente, se ajustó una regresión Poisson con enfoque bayesiano gracias a la función $stan_glm$ de la librería rstanarm. Dicha función, deja modificar la familia de la distribución, la cual en el caso particular de la variable respuesta, al ser conteo, se utilizó la distribución Poisson en la fórmula y se especificó el algoritmo Markov chain Monte Carlo (MCMC) con 2000 iteraciones.

Para la evaluación de los parámetros, se utilizó la función $describe_posterior()$ de la librería bayestestR. Esta función arroja, para cada variable, el intervalo de credibilidad, donde se cuantifica la incertidumbre sobre los coeficientes de regresión; La probabilidad de dirección, la más asemejada al valor p frecuentista; ROPE o región de equivalencia práctica, entre otras métricas (Chaloner et al., 1988).

Se comprobó la significancia de los coeficientes verificando que el intervalo de credibilidad no contuviera el cero. Además, se comparó con las variables que estuvieran con valores fuera del intervalo ROPE al mismo tiempo.

En el Cuadro 6.6 se encuentras los resultados del modelo bayesiano.

En cuanto a los residuales del modelo, si se evidencia una mayor dispersión lo que se ve reflejado en la imagen de los residuales por la Deviance y los residuales por Pearson. Lo que indica que a pesar de ser el modelo más parsimonioso, también es el modelo con mayor número de datos atípicos significativos (mayores o menores de 2,-2) de los modelos ajustados.



Dados los resultados anteriores, se evidenció cómo el enfoque Bayesiano puede generar diferentes aproximaciones al mismo conjunto de datos. Sin embargo, al igual que los procesos del enfoque frecuentista, hay que tener cuidado con ciertas limitantes como los supuestos en los modelos lineales generalizados o la conformación de la distribución a priori en el método Bayesiano.

6.5. Comparativo

Los modelos frecuentistas como la regresión lineal y regresión Poisson, son los modelos con mayor comparabilidad si y solo sí, se ajustaron con la misma función de enlace, en este caso el logaritmo. Y aún así, los resultados descartan al modelo lineal múltiple por su mala explicación (48.18 %) de la variable independiente.

El modelo Poisson con StepAIC generó un pseudo R^2 de 0.6737, lo que demuestra una muy buena explicación de la variabilidad. Por otro lado el modelo lineal generalizado Poisson con selección por Random Forest muestra una buena explicación con un Pseudo R^2 de 0.4445. Finalmente, se calculó el Pseudo R^2 del modelo bayeasion el cual dió 0.45.

Otro comparativo es por la medida de calidad del criterio de información Akaike (AIC), el cual en el modelo Poisson por StepAIC dió un valor de 480, mientras que el modelo Poisson por random forest, generó un valor de 505.28. Finalmente, con la función $extractAIC()$ el modelo bayesiano generó un AIC de 487 aproximadamente. En este escenario, el mejor modelo sería el Poisson por pasos (StepAIC), sin embargo, hay que tener en cuenta que es el modelo menos parsimonioso.

Teniendo en cuenta lo anterior, en vez de escoger un *mejor* modelo sin tener un criterio fijo como la predicción, ya que no es el objetivo de esta etapa de investigación, se decidió presentar el Cuadro 6.8 a la médico experta en el tema, sin mencionar el nómbre del procedimiento (evaluación ciega), para no sesgar la escogencia de las variables.

El objetivo de este paso, fue generar diferentes grupos que explican el tiempo de hospitalización según diferentes herramientas estadísticas, así la profesional pudiera darse una idea sobre la asociación estadística desde diferentes aproximaciones y con esto, poder generar un formulario con

Variables significativas			
R.L. Múltiple	Poisson StepAIC	Poisson RF	Bayesiano P.
Mialgias	Lac. menor & mayor	Mialgias	Lac. menor & mayor
Leucocitos	Mialgias	Leucocitosis	**trombocitosis**
No. antibióticos	Tiempo fiebre	**trombocitosis**	**No. días antibiótico**
Asp. med. os.	Cervicales	Linfadenopatías generalizadas	P.S. emesis & dolor ab.
Neumonía	Hepatomegalias	**No. días antibiótico**	**Aspirado médula ósea**
	Leucocitosis	**Aspirado médula ósea**	**Neumonía**
	Neutrofilia	**Neumonía**	
	Trombocitosis		
	No. días antibiótico		
	Infección neumonía		
	Aspirado médula ósea		
	Neumonía		

Cuadro 6.8: Variables significativas

	Poisson StepAIC	Poisson RF	Bayesiano P.
AIC aprox.	480	505	487
Pseudo R^2	0.67	0.44	0.45

Cuadro 6.9: medidas de calidad

más eficiente para la siguiente fase del proceso, donde se buscará tener una muestra representativa de los pacientes con infección por VEB y pacientes control para mejorar el aporte al diagnóstico y evolución de este virus en la cohorte Colombiana.

En el Cuadro 6.8, se presentan las variables significativas de cada modelo ajustado en función del tiempo de hospitalización.

6.6. Otros modelos

Es importante mencionar que se ajustaron otros modelos que aplicaban al estudio, pero que luego de ciertas pruebas, no se ajustaban a los supuestos necesarios para ser relevantes. Un ejemplo de lo anterior fue el ajuste de un modelo binomial negativo, sin embargo, este modelo no pasó la prueba de sobre dispersión.

Otro modelo que se intentó ajustar, fue con la aproximación de la asociación por medio de modelos de regresión de reducción, los cuales están más enfocados en forzar los coeficientes de las variables mediante diferentes métodos, para obtener un manejo eficiente de modelos de regresión con grandes cantidades de registros y variables. Estos modelos, están más enfocados en el poder predictivo, sin embargo, se experimentó para la situación puntual del problema, para generar otro posible grupo de variables explicativas del tiempo de hospitalización de pacientes con infección por VEB.

Como se realizó con los modelos anteriores, se utilizaron todas las variables finales de la base de datos, con el objetivo de mirar otras posibles variables candidatas para ajustar el modelo. Sin embargo, al verificar los coeficientes del modelo forzados, todas las variables explicativas tenían el valor de cero. Esto generó sospecha sobre el mínimo de información requerido para poder utilizar estos modelos de reducción o la afectación de variables categóricas en la eficiencia del modelo por el valor de sus coeficientes.

Capítulo 7

Discusión

Una vez presentados los resultados a la médico de apoyo, se procedió a registrar la asociación técnica de cada una de las variables con el número de días de hospitalización de los pacientes pediátricos por VEB, junto con la opinión de los resultados. Cabe aclarar, que se omitieron las variables estadísticamente significativas del modelo de regresión lineal múltiple, a pesar de haber cumplido con todos los supuestos, por dos motivos. La mala explicación que generó el modelo en general y las 5 variables significativas, aparecen en los otros tres modelos.

Inicialmente, se revisaron los resultados del modelo Poisson por StepAIC. En general, llamó la atención el número de variables y la composición de estas, ya que el modelo muestra asociación de variables sociodemográficas, clínicas y paraclínicas en relación al tiempo de hospitalización. En cuanto a la parte médica, se resaltaron los siguiente comentarios de la médico:

En cuanto a la variable *Edad*, usualmente los pacientes de menor edad suelen tener mayor tiempo de hospitalización debido a su sistema inmune en desarrollo en los primeros años de vida. Esto se ve reflejado en la importancia de categorizar las diferentes reacciones de un virus como el VEB frente a otras enfermedades confusoras, por ejemplo el virus del Dengue.

La variable *Mialgias* genera curiosidad debido a que aparece en 3 modelos anteriormente realizados. En especial, por que es un síntoma que puede generar confusión debido a la alta frecuencia con la que se presenta en diferentes enfermedades. Sin embargo, vale la pena indagar este síntoma en población pediátrica de un siguiente estudio con población control, infección VEB y otras infecciones con cuadros similares.

En cuestión a la variables *Número de días con fiebre*, dado que la población de estudio es pediátrica, se necesita mayor control de la fiebre debido a su sistema inmunológico, lo que prolonga el tiempo de hospitalización.

Cervicales o aumento del tamaño de ganglios cervicales, tienden a ser suceptibles de cambios debido que el virus principalmente infecta estos ganglios ubicados en la vía aérea lo que puede generar sospecha de un síndrome mielodisplásico o cancer, por lo que aumenta el tiempo de hospitalización al momento de indagar este aspecto.

En cuanto a las *Hepatomegalias*, se refiere a enfermedades del hígado, lo que genera sospecha de falla hepática y prolonga el tiempo de hospitalización mientras se realizan más estudios de

alteración en ese sistema. Cabe resaltar, que este síntoma hace parte del cuadro de la enfermedad causada por VEB según la literatura americana.

Referente a la *Leucocitosis*, al ser una infección viral aumenta el conteo de leucocitos lo que indica, mayores estudios para identificar el tipo de infección que presenta el paciente.

La variable *Neutrofilia* genera curiosidad, ya que al ser una infección viral, deberían tender al aumento de linfocitos. Sin embargo, es interesante ver si el aumento de neutrófilos hace parte de una reacción viral o bacteriana en población pediátrica con diferentes enfermedades.

En cuanto a la *Trombocitosis*, Es una variable que se mira mucho dada su reacción a la infección severa. Esto genera sospecha de sepsis lo que prolonga el tiempo de hospitalización mientras se estabiliza el paciente y se identifica la causa de la infección severa. Adicionalmente, esta condición puede generar trombos, lo que puede causar un potencial daño al paciente pediátrico.

El *Número de días con antibiótico* refleja el esquema utilizado en el paciente según la literatura y su evolución en la enfermedad.

En relación a *Aspirado de médula ósea*, esta variable aparece en los 4 modelos realizados, debido a que es un proceso invasivo que busca la morfología de las células sanguíneas que nacen en la médula ósea cuando no fue clara la causa de la enfermedad del paciente. Dada su complejidad y morbilidad, muestra una relación fuerte en el tiempo de hospitalización.

La última variable del modelo Poisson, también aparece en los 4 modelos realizados, *Neumonía*, desde la causa viral, que desencadena inflamación pulmonar y se debe apoyar en tratamientos de oxigenación u otros mecanismos que prolongan el tiempo de hospitalización del paciente.

Con respescto a la evaluación de importancia generado por random forest, se resaltan las variables *número de días con antibiótico, mialgias, leucocitosis, trombocitosis, aspirado de médula ósea y neumonía*, compartidas con el modelo Poisson y que anteriormente se explicaron en su relación con el tiempo de hospitalización de los pacientes analizados.

Adicionalmente, el algoritmo resaltó 1 variable adicional significativa del modelo y que solo aparece en este escenario. *Linfadenopatías generalizadas*, la cual, puede explicar el aumento del tiempo de hospitalización debido a los estudios de biopsias para esclarecer, posibles causas de este síntoma.

Hablando del método bayesiano, este resultó con las variables *edad, trombocitosis, número de días con antibiótico, aspirado de médula ósea y neumonía* como variables relevantes al igual que los métodos anteriores. Adicionalmente, se recalca la variable de *primer síntoma* en dos categorías, como variable significativa en asociación al tiempo hospitalario de los pacientes:

Dolor abdominal como primer síntoma, es una variable muy inespecífica que puede variar en diferentes resultados como dolor quirúrgico, gástrico, etc. Lo que dilata más el tiempo de observación para obtener un resultado más acertado.

En cuestión de *Emesis como primer síntoma*, al igual que el dolor abdominal, es un primer síntoma muy inespecífico ya que pertenece al cuadro clínico de varias enfermedades o infecciones, lo que dificulta el diagnóstico apropiado a tiempo.

Dadas las variables anteriores, se sospecha que factores como Mialgias, Emesis y otras ante-

riormente nombradas como síntomas inespecíficos, son causa de la confusión en el diagnóstico que desencadena mayor tiempo de hospitalización, al no ser parte del cuadro común de la enfermedad causada por VEB.

Al cabo de discutir los resultados, se estipuló utilizar la totalidad de las variables asociaciadas de los 4 métodos utilizados (*Mialgias, Leucocitos, Número de días con antibióticos, Aspirado de médula ósea, Neumonía, Edad, Tiempo de fiebre, Cervicales, Hepatomegalias, Neutrófilos, Trombocitos, Infección neumonía, Linfadenopatías generalizadas y Primer síntoma*), con el fin de continuar con un siguiente estudio que cuente con población control, pacientes infectados por VEB y población con infección por otros virus con diagnóstico parecido al VEB, como el Dengue o Chikunguña. Lo anterior, con el objetivo de identificar el comportamiento y evolución de virus con sintomatología similar, según la literatura, de la cohorte pediátrica local y así, poder generar resultados que ayuden a la distinción en el diagnóstico diferencial y ayudar a mejorar la eficiencia en el servicio de salud local.

Sin embargo, desde la parte estadística si se quiso escoger un modelo que representara la conclusión del trabajo. Es por esto, que al entender la finalidad de la fase de investigación, que es identificar las variables que afectan el tiempo de hospitalización de pacientes pediátricos con VEB, se escogió el modelo de regresión Poisson con selección paso a paso AIC, a pesar de ser el modelo menos parsimonioso. Si la fase de investigación tuviera un enfoque más predictivo, el modelo de regresión Poisson con selección por Random Forest sería el más adecuado. Si por otro lado, fase de investigación estuviera en una fase final para dar resultados de la población Colombiana, el modelo de regresión Bayesiano muestra una buena explicación de la mejor variable de cada grupo de la base de datos.

El modelo escogido se apreciaría de la siguiente manera:

$$\text{Log(No. Días Hospitalizado)}=2e^{-16}+(4{,}7e^{-05})X_{Lact.Men.}+(0{,}0038)X_{Lact.May.}+(0{,}0057)X_{Mial.}+(0{,}0487)X_{D.fieb.}+(0{,}0391)X_{Cerv.}+(0{,}035)X_{Hepa.}+(0{,}0063)X_{Leuc.}+(0{,}0054)X_{Neut.}+(0{,}0036)X_{Trom.}+(0{,}0442)X_{N.d.ant.}+(0{,}0052)X_{Inf.bac.}+(2{,}05e^{-08})X_{A.M.O.}+(0{,}00029)X_{Neum.}$$

Capítulo 8

Conclusiones

Se logró ajustar un modelo de regresión para el número de días hospitalizado de pacientes de 1 mes a 17 años de vida con infección por virus Epstein Barr, durante el periodo 2015 a 2019, en una clínica infantil de Bogotá, en función de sus variables sociodemográficas, clínicas y paraclínicas.

Se entendió que la naturaleza del problema, las variables y el contexto, son fundamentales a la hora de realizar un aporte estadístico en otras ramas de investigación. Específicamente en este proyecto, asimilar el contexto del conocimiento médico de la enfermedad, de la cohorte estudiada y de la fase de investigación, fue esencial para proponer una aproximación de los datos desde la variable de tiempo hospitalizado, dado la imposibilidad de obtener los datos de pacientes control.

Se comprendió que el objetivo de la fase de investigación, busca encontrar asociaciones de las variables sociodemográficas, clínicas y paraclínicas que expliquen el tiempo prolongado de hospitalización de estos pacientes y así, poder argumentar variables de interés para un siguiente estudio que genere mayor conocimiento sobre enfermedades de diagnóstico similar en cohortes locales. Lo anterior, argumenta ajustar modelos por su capacidad de explicación más que por su capacidad de predicción. Esta última, será fundamental, más adelante en la investigación donde se tengan resultados de mayor peso en una muestra representativa.

Se entendió que es importante complementar los resultados obtenidos con diferentes aproximaciones estadísticas que sean acordes a la naturaleza de la variable. En este proyecto, las aproximaciones frecuentistas, bayesianas y por machine learning, presentaron ciertas variables similares asociadas al tiempo de hospitalización, lo que da mayor peso en su relación con la variable dependiente y mayor interés desde el punto de vista médico.

Finalmente, se resalta la cantidad de herramientas y métodos que puede generar otras aproximaciones al mismo problema, lo que le da mayor cantidad de caminos al profesional estadístico para, primero, seguir indagando en los avances para el manejo de datos y segundo, para abordar los problemas desde diferentes perspectivas estadísticas, con el objetivo de poder tener mayor peso en los análisis a realizar.

Capítulo 9

Bibliografía

Balfour, H., Dunmire, S., & Hogquist, K. (2015). Infectious mononucleosis.

Dunmire, S., Verghese, P., & Balfour, H. (2018). Primary Epstein-Barr virus infection.

Gao, L.-W., Xie, Z.-D., Liu, Y.-Y., Wang, Y., & Shen, K. (2011). Epidemiologic and clinical characteristics of infectious mononucleosis associated with Epstein-Barr virus infection in children in Beijing, China. World J Pediatr, 7, 45–49.

Garcia-Peris, M., Jimenez Candel, M. I., Mañes Jimenez, Y., Pariente Marti, M., González Granda, D., & Calvo Rigual, F. (2018). Primoinfección por el virus de Epstein-Barr en niños sanos. An Pediatr(Barc).

González Saldaña, N., Monroy Colín, V. A., Piña Ruiz, G., & Juárez Olguín, H. (2012). Clinical and laboratory characteristics of infectious mononucleosis by Epstein-Barr virus in Mexican children. BMC Research Notes, 5, 361.

Helsinki, D. (2013). Principios éticos para las investigaciones médicas en seres humanos. 4–8.

Likic, R., & Kzmanic, D. (2004). Severe thrombocytopenia as a complication of acute Epstein-Barr virus infection. Wien Klin Wochenschr, 116(2), 47–50.

Losa García, J., Miró Meda, J., Alcaide García, F., & Gatell Artigas, J. (1998). Síndrome mononucleósico. Medicine (Baltimore), 7(82), 7.

Ministerio de Salud. Resolución 8430 de 1993. Ministerio Salud y protección social, Republica Colombia. , (1993).

Moreno Bermeo, M. A. (2020). Características sociodemográficas, clínicas y paraclínicas de pacientes pediátricas con infección por Virus del Epstein Barr.

Son, K. H., & Shin, M. Y. (2011). Clinical features of Epstein-Barr virus-associated infectious mononucleosis in hospitalized Korean children. The Korean Pediatric Society, 54, 409–413.

Sullivan, J., & Luzuriaga, K. (2017). Virology of Epstein-Barr virus.

Topp, S., Rosenfeldt, V., Vestergaard, H., Christiansen, C. B., & Von LInstow, M.-L. (2015). Clinical characteristics and laboratory findings in Danish children hospitalized with primary Epstein-Barr virus infection. Infectious Diseases, 47(12), 908–914.

Wu, Y., Ma, S., Zhang, L., Zu, D., Gu, F., Ding, X., & Zhang, L. (2020). Clinical manifestations and laboratory results of 61 children with infectious mononucleosis. Journal of International Medical

Research, 48(10), 1–8.

Yang, X., Nishida, N., Zhao, X., & Kanegane, H. (2015). Advances in Understanding the Pathogenesis of Epstein-Barr-Virus-Associated rLymphoproliferative Disorders.

Colin G. Weaver, Pietro Ravani, Matthew J. Oliver, Peter C. Austin, Robert R. Quinn, Analyzing hospitalization data: potential limitations of Poisson regression, Nephrology Dialysis Transplantation, Volume 30, Issue 8, August 2015, Pages 1244–1249, https://doi.org/10.1093/ndt/gfv071

Lever, J., Krzywinski, M. & Altman, N. Regularization. Nat Methods 13, 803–804 (2016). https://doi.org/10.1038/nmeth.4014

VanderPlas, J. (2016). Python data science handbook : essential tools for working with data. Sebastopol, CA: O'Reilly Media, Inc. ISBN: 978-1491912058

Rodrigo, J. Regularización Ridge, Lasso y Elastic Net con Python, available under a Attribution 4.0 International (CC BY 4.0) at https://www.cienciadedatos.net/documentos/py14-ridge-lasso-elastic-net-python.html

Regularization and variable selection via the elastic net, Hui Zou and Trevor Hastie, J. R. Statist. Soc.B (2005)

McCullagh, P., Nelder, J. A. (1989). Generalized Linear Models. London: Chapman & Hall / CRC.

Murphy,K. Machine Learning: A Probabilistic Perspective, 2012, page 589

Castellano, D. (2015). Introducción a la Estadística Bayesiana(tesis de licenciatura) (0 ed.). Universidad Autonoma de Barcelona.

Marin, J.M., Robert, C.P. (2007). Bayesian Core: A Practical Approach to Computational Bayesian Statistics.

Hoff, Peter D. 2009. A First Course in Bayesian Statistical Methods. Springer Science & Business Media.

Chaloner, Kathryn, and Rollin Brant. 1988. "A Bayesian Approach to Outlier Detection and Residual Analysis." Biometrika 75 (4): 651–59.

Hoff, Peter D. 2009. A First Course in Bayesian Statistical Methods. Springer Science & Business Media. Jeffreys, Sir Harold. 1961. Theory of Probability: 3rd Edition. Clarendon Press.

Kass, Robert E, and Adrian E Raftery. 1995. "Bayes Factors." Journal of the American Statistical Association 90 (430): 773–95.

Venables, William N, and Brian D Ripley. 2013. Modern Applied Statistics with s-PLUS. Springer Science & Business Media.

Kuhn, M. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2016).

Kuhn, M. May, (2019). The caret package. http://topepo.github.io/caret/

Casas, P. Libro Vivo Ciencia de Datos (2019). https://librovivodecienciadedatos.ai/

Clyde, M. Ghosh, J. and Littman, M. (2010) Bayesian Adaptive Sampling for Variable Selection and Model Averaging. Journal of Computational Graphics and Statistics. 20:80-101

Li, Y. and Clyde, M. (2019) Mixtures of g-priors in Generalized Linear Models. Journal of the American Statistical Association. 113:1828-1845

Raftery, A.E, Madigan, D. and Hoeting, J.A. (1997) Bayesian Model Averaging for Linear Regression Models. Journal of the American Statistical Association.