Modelos estadísticos para el tiempo de recuperación en pacientes COVID-19 de Colombia 2020-2021

Leidy Vanesa Espitia Cruz

Trabajo de Grado para Optar al Título de Matemático

Director

Tulia Esther Rivera Flórez

Magister en Estadística

Universidad Industrial de Santander

Facultad de Ciencias

Escuela de Matemáticas

Bucaramanga

2022

Dedicatoria

A Dios quien me ha otorgado la vida y las posibilidades para la realización de este trabajo.

A mi madre quien con su amor y dedicación me ha formado y sacado adelante, quien es mi fuerza, mi apoyo y guía, quien nunca me ha abandonado y por quien trabajo para enorgullecer cada día.

A mi hermano quien ha sido un apoyo y un hombro amigo al cual siempre recurrir.

A mis profesores quienes me han enseñado y otorgado conocimientos en diferentes campos, de quienes he recibido apoyo y orientación en todo este proceso formativo.

A mi directora de tesis Tulia Esther Rivera quien me guio en este proceso con la mayor disposición y dedicación, de quien recibí un apoyo incondicional.

A mis compañeros y amigos con quienes he compartido, he crecido, he aprendido y he vivido toda esta primera etapa de formación, de quienes he recibido apoyo, ayuda y motivación cuando la necesite.

A mi abuelita quien ya no está con nosotros, pero sigue siendo un enorme apoyo y confort para mi vida.

Agradecimientos

Agradezco a la Universidad Industrial de Santander por ser mi lugar de formación como persona y profesional.

A mi directora de tesis la profesora Tulia Esther Rivera Flórez por ser mi orientadora, mi ayuda y mi apoyo incondicional durante el desarrollo de este trabajo, quien con su ayuda, consejos y recomendaciones se construyó este proyecto.

A mis docentes y orientadores a lo largo de todo el recorrido en este proceso de formación para optar al título de Matemática.

A mi familia principalmente a mi madre por ser la mayor fuente de apoyo en todos los ámbitos.

A todas aquellas personas que han aportado a mi formación y crecimiento como persona y profesional.

A quienes aportaron con sus correcciones y sugerencias en este proyecto.

Tabla de Contenido

Pág.

Lista de Tablas

Pág.

Lista de Figuras

Pág.

Lista de Apéndices

Pág.

Resumen

Título: Modelos estadísticos para el tiempo de recuperación en pacientes COVID-19 de Colombia 2020-2021[*]

Autor: Leidy Vanesa Espitia Cruz[**]

Palabras Clave: COVID-19, Tiempo de Recuperación, Análisis de Supervivencia, Modelos paramétricos, Modelos no paramétricos, Modelo AFT, Supervivencia Bayesiana, Modelo AFT CGM.

Descripción:

En este trabajo se presenta un análisis de uno de los eventos involucrados y de gran interés en el estudio de cualquier enfermedad como lo es la recuperación de un paciente, en este caso en la pandemia por COVID-19 en el entorno nacional, para esto se realiza un estudio del Tiempo de Recuperación de los pacientes diagnosticados con COVID-19 en Colombia, a partir de la información obtenida de la base de datos abiertos **"**Casos positivos de COVID-19 en Colombia**"** del INS y el uso de paquetes para modelar datos de supervivencia del software estadístico R.

El estudio y análisis del Tiempo de recuperación se realiza a través de análisis de supervivencia para distintos periodos de tiempo establecidos de la pandemia en 2020-2021. Se ajustan modelos de supervivencia paramétricos (Exponencial, Weibull, Logístico, Log-Logístico, Normal, Log-Normal, Gamma generalizado) buscando el mejor ajuste por medio de criterios de información (AIC, BIC) y pruebas gráficas, así mismo se ajustan modelos no paramétricos (Kaplan Meier). Para establecer posibles efectos en la recuperación de un paciente COVID-19 por parte de variables categóricas como el Sexo, la Etnia, el Clima del municipio de residencia y el Grupo de Edad al que pertenece, ajustando Modelos de Tiempo de Falla Acelerado (AFT).

Por otra parte, se realiza también un análisis de supervivencia en un enfoque bayesiano al primer periodo de tiempo considerado de la pandemia (febrero - abril 2020), en primera instancia se modela la curva de supervivencia a posteriori a partir de dos a prioris (Gamma dependiente, Gamma independiente) y se comparan con la alternativa no paramétrica clásica (Kaplan Meier). Se realiza además el ajuste de un Modelo AFT con una mezcla gaussiana clásica como distribución de error (AFT CGM) para estudiar los efectos de las variables categóricas usadas en uno de los modelos AFT clásicos.

---

[*] Trabajo de Grado

[**] Facultad de Ciencias. Escuela de Matemáticas. Director: Tulia Esther Rivera Flórez. Magister en Estadística.

Abstract

Title: Statistical models for the recovery time in COVID-19 patients in Colombia 2020-2021[*]

Author: Leidy Vanesa Espitia Cruz[**]

Key Words: COVID-19, Recovery Time, Survival Analysis, Parametric Models, Nonparametric Models, AFT Models, Bayesian Survival, AFT-CGM Model.

Description:

This paper presents an analysis of one of the events involved and of great interest in the study of any disease such as the recovery of a patient, in this case in the COVID-19 pandemic in the national environment, for this a study of the Recovery Time of patients diagnosed with COVID-19 in Colombia, from the information obtained from the open database **"**Casos positivos de COVID-19 en Colombia**"** of the INS and the use of packages for modeling survival data of the R statistical software.

The study and analysis of the Recovery Time is performed through survival analysis for different established time periods of the pandemic in 2020-2021. Parametric survival models (Exponential, Weibull, Logistic, Log-Logistic, Normal, Log-Normal, Generalized Gamma) are adjusted looking for the best fit by means of information criteria (AIC, BIC) and graphical tests, as well as non-parametric models (Kaplan Meier). To establish possible effects on the recovery of a COVID-19 patient by categorical variables such as Sex, Ethnicity, Climate of the municipality of residence and the Age Group to which belongs, adjusting Accelerated Failure Time Models (AFT).

On the other hand, a survival analysis is also performed in a Bayesian approach to the first period of time considered of the pandemic (February - April 2020), in the first instance the a posteriori survival curve is modeled from two priors (Gamma dependent, Gamma independent) and compared with the classical non-parametric alternative (Kaplan Meier). In addition, an AFT model with a classical Gaussian mixture as error distribution (AFT CGM) is fitted to study the effects of the categorical variables used in one of the classical AFT models.

[*] Degree Work

[**] Science Faculty. School of Mathematics. Director: Tulia Esther Rivera Flórez. Master in Statistics.

Introducción

A través de la historia, el estudio y análisis de datos ha venido ganando protagonismo y cada vez resultan ser de mayor relevancia para las sociedades más desarrolladas o países del primer mundo, con lo cual los procesos de toma de decisiones son cada vez más racionales y se privilegia que estos estén basados en evidencias objetivas; esta nueva forma de entender el mundo ha logrado impactar todos los escenarios, desde el ámbito gubernamental donde se dictan políticas basados en las proyecciones censales hasta decisiones personales. Los métodos que han hecho esto posible provienen mayoritariamente de la denominada estadística clásica, enfoque dominante por años y que de hecho se sigue enseñando casi de manera exclusiva en el mundo académico, no obstante, en los últimos años la Estadística Bayesiana ha venido cobrando fuerza por su forma de manejar y analizar distintos problemas, utilizando toda la información disponible de los expertos respecto del comportamiento de las variables a modelar, yendo más allá de la información otorgada exclusivamente por los datos.

En la actualidad, la pandemia del COVID-19 es el mayor reto enfrentado de manera global, sus implicaciones han sido negativas en distintos ámbitos de la sociedad mundial pero a su vez esta crisis ha precipitado el interés de muchos analistas por estudiar y modelar aspectos sobre el comportamiento de este virus como son: la postulación y análisis de indicadores que resuman el comportamiento de la pandemia, factores asociados a la expansión de la pandemia, formulación de medidas preventivas y de control, métodos de tratamiento, recientemente la evolución y efectividad de los procesos de vacunación. Para el caso de Colombia, el primer caso confirmado fue importado de Italia y se dio el 6 de marzo de 2020, desde entonces se han presentado más de cinco millones de casos positivos para el virus, para cada uno de ellos se cuenta con un registro

oficial consignado en la base de datos *Casos positivos de COVID-19 en Colombia* la cual es administrada por el Ministerio de Salud y permite libre acceso a cualquier usuario; las variables a disposición en este sitio dan cuenta de información relacionada con la situación médica frente al virus e información demográfica en un nivel básico, infortunadamente no se registra otro tipo de información relevante para estudios relacionados con la pandemia como son las comorbilidades del paciente, síntomas presentados, aspectos socioeconómicos, familiares y de cubrimiento en salud de cada paciente, no obstante es importante analizar lo que se tiene en beneficio de nuestra sociedad.

Con base en lo anterior, este trabajo se propone analizar los datos mencionados usando métodos estadísticos explorando dos alternativas: la bayesiana y la clásica, éstas describen grandes diferencias desde su implementación teórica, supuestos que demandan, metodología de análisis, interpretación de resultados y criterios de evaluación en cada contexto específico. Este trabajo pretende evidenciar las diferencias entre los enfoques mencionados a partir de un análisis basado en unos objetivos comunes como son, analizar los datos de los pacientes COVID-19 en Colombia y modelar la variable tiempo de recuperación acorde a las posibilidades que el lenguaje R y sus paquetes específicos nos ofrezcan.

1. Objetivos

1.1 Objetivo General

Comparar el comportamiento de modelos clásicos tanto paramétricos como no paramétricos frente a la alternativa bayesiana para el caso de la variable tiempo de recuperación de pacientes COVID 19.

1.2 Objetivos Específicos

Identificar las variables (factores de riesgo) que describen diferencias significativas para el tiempo de recuperación de pacientes COVID 19.

Ajustar los modelos de supervivencia que permitan explicar el comportamiento de la variable Tiempo de Recuperación acorde a los resultados encontrados en el primer objetivo específico y el marco teórico en consideración.

- Evaluar, utilizando criterios estadísticos, la bondad de ajuste de los modelos paramétricos clásicos propuestos.
- Comparar el comportamiento de los modelos clásicos frente al comportamiento de la alternativa bayesiana.

2. Marco Referencial

2.1 Planteamiento del Problema

Desde el punto de vista de salud pública una epidemia es una situación de máxima complejidad que demanda poner al servicio del control de ésta todo el sistema de salud de un país, además de la gestión administrativa gubernamental en todos los niveles de poder; no obstante, la crisis que hemos tenido que afrontar nos ha mostrado el papel preponderante que juega la ciencia para ampliar el conocimiento de la causa que genera una crisis, los factores que precipitan su diseminación y como atenuar sus nefastas consecuencias. Desafortunadamente, en el contexto nacional son pocos los grupos de investigación en áreas como bioestadística o epidemiología, por tanto, incentivar el estudio de estos temas desde la academia es más que beneficioso para el país porque estimula la formación de capital humano capaz de aportar en estos campos de investigación.

De otro lado, el gobierno colombiano viene promoviendo la cultura del registro de la información y para ello ha creado la base de datos abiertos “https://www.datos.gov.co/” que contiene varias bases de datos a las cuales se puede acceder de forma libre, la base de datos de nuestro interés es la creada por el Instituto Nacional de Salud que viene siendo alimentada desde el inicio de la pandemia en 2020. A partir de esta base de datos que se actualiza diariamente, el Instituto Nacional de Salud (INS) y algunas otras entidades han publicado los reportes diarios que se basan en presentaciones del consolidado de los indicadores sobre: nuevos casos, número de recuperados y número de muertos, no obstante, la base de datos tiene más información que conviene ser analizada. Por lo anterior, el estudio que se plantea realizar en este proyecto centra su interés en una variable poco explorada, el Tiempo de recuperación, la cual se analizará de

manera individual y en relación con factores como el género, la etnia, la edad y la ubicación geográfica acorde a la disponibilidad de información.

Con base en el anterior planteamiento, la técnica estadística a aplicar será un Análisis de supervivencia clásico para el cual existe una amplia literatura que incluye técnicas tanto paramétricas como no paramétricas de uso común. En general, en el modelamiento estadístico clásico de cada técnica viene acompañada de varios supuestos sobre el comportamiento probabilístico de los datos o de relación entre variables, que desde el punto de vista práctico no siempre se pueden validar, así mismo hay dificultad en la interpretación práctica de los resultados a nivel inferencial y de disponibilidad de tamaños de muestra óptimos requeridos. Ante estas dificultades, ha emergido el enfoque bayesiano para hacer posible analizar un contexto específico desde fases tempranas, es decir, con poco tamaño de muestra; para lograr esto lo que se hace es iniciar el análisis a partir de información a priori aportada por una fuente experta y confiable que luego es actualizada en la medida que se van consiguiendo datos. Este trabajo pretende explorar el ajuste de los datos de COVID 19 en Colombia según algunos modelos clásicos y que será confrontado con una alternativa bayesiana ubicándonos en la fase inicial de la pandemia.

2.2 Antecedentes

Durante este lapso de tiempo en el que se ha desarrollado la pandemia por COVID-19, muchos estadísticos, instituciones de salud y analistas de datos se han interesado en el estudio de la pandemia utilizando para ello una amplia gama de enfoques adicional al descriptivo. A continuación, se describen algunos estudios y referencias que presentan análisis estadísticos sobre los datos de la pandemia COVID-19 tanto a nivel nacional como internacional y que guardan relación directa con los objetivos de este proyecto:

En primer lugar, encontramos el análisis que genera diariamente World Health Organization (WHO) (2021), éste es un recurso de dominio público que consiste de un tablero de control (dashboard) el cual presenta información sobre la evolución de la pandemia a nivel mundial discriminando la información por continentes y por países, los datos corresponden a una serie de tiempo con punto de inicio en diciembre de 2019 y detallan información como el nivel de contagios, número de muertes y cifras relacionadas con el avance en materia de vacunación. Adicionalmente, se cuenta con el reporte semanal de la WHO (2021) titulado COVID-19 Weekly Epidemiological Update el cual es un documento que resume las cifras actualizadas sobre la pandemia a nivel mundial, una de las versiones disponibles es la del 5 de diciembre de 2021 donde se reportan una estabilización en las cifras de casos confirmados pero un aumento en un 10% de las muertes respecto a la semana anterior, además presentan un resumen actualizado por continentes.

A nivel nacional, encontramos un estudio similar al de la WHO realizado por el Instituto Nacional de Salud en el cual se presenta información actualizada diariamente discriminando por departamento, el informe se basa en la presentación de indicadores de la pandemia como son: casos positivos, número de recuperados y fallecidos, también se tiene acceso a estas cifras acumuladas hasta la fecha.

De otro lado, en materia de artículos publicados sobre temas relacionados con el COVID-19 hay una amplia variedad de literatura, en esta sección se pretende mostrar sólo aquellos que guardan relación directa con el objetivo del proyecto. Iniciaremos por presentar los artículos encontrados a nivel nacional y luego se hará lo mismo para algunos del contexto internacional.

En el artículo de Diaz et al. (2021) publicado en la revista International Journal of Infectious Diseases, los autores buscaron caracterizar la dinámica de la epidemia COVID-19 con

fines de modelado. Se presenta un análisis de los datos para Colombia durante 5 meses y se exploran e identifican modelos que representan el tiempo desde el inicio de los síntomas hasta la hospitalización, la admisión en la unidad de cuidados intensivos (UCI) y muerte, discriminan estos resultados por las variables edad y sexo. Al final del estudio se encontró que los tiempos de paso o de transición de una etapa de la enfermedad a otra fueron, en palabras de los autores, casi independientes de la edad y el sexo, en contraste con las probabilidades de pasar de una etapa a otra de la enfermedad, donde hallaron una fuerte dependencia con estas variables, el único factor que afectó la duración de la estancia hospitalaria y en la UCI, fue el resultado final de la enfermedad es decir, la supervivencia o la muerte, donde los tiempos fueron significativamente mayores para los sobrevivientes, y presentaban en palabras de los autores una casi independencia del sexo o la edad de los pacientes.

También para datos de Colombia, encontramos el artículo escrito por Cifuentes et al. (2021) donde se presenta el estudio con datos reportados durante los primeros ocho meses de pandemia. Estos autores estudiaron el efecto en la mortalidad de algunos factores como la edad, etnia, lugar de residencia, estrato socioeconómico y tipo de seguro de salud, los cuales están relacionados con la desigualdad socioeconómica en Colombia. La técnica utilizada fue un análisis de supervivencia para el tiempo hasta la muerte o recuperación de cada caso confirmado, además se ajustó un modelo de regresión de Cox dependiente del tiempo multivariable extendido para estimar los riesgos proporcionales (HR) por los grupos estudiados. Este estudio produjo evidencia del efecto que presentan las desigualdades socioeconómicas en la mortalidad por COVID-19 respecto a los factores estudiados.

Por su parte, Yadav y Akhter (2021), presentan una revisión de distintas técnicas de predicción y modelado estadístico para el estudio de enfermedades infecciosas, como es la

afección por COVID-19. Los autores, presentan una amplia información referente a las enfermedades infecciosas y algunos datos a tener en cuenta relacionados con la caracterización y dinámica de estas. Estos autores muestran el ajuste de una distribución para las etapas de crecimiento de la epidemia para ello utilizan el modelado de series tiempo junto con enfoques de monitoreo predictivo y modelamiento epidemiológico, (Series de tiempo estacionarias, Modelo AR, Modelo MA, Modelo ARMA, Modelo ARIMA, Modelo SARIMA, etc). También se muestran referencias del uso y estudio de estos métodos en el estudio de enfermedades infecciosas donde generalmente los principales objetivos son encontrar la manera y la razón de propagación de las infecciones, así mismo como prevenirlas y restringirlas. Además, Yakav y Akhter realizan una revisión de los resultados de los recientes esfuerzos de modelado estadístico para predecir el curso de la propagación de COVID-19, donde en los últimos estudios, los autores han llevado a cabo la predicción y análisis del COVID-19 a través de modelos de series temporales.

Adicionalmente, la búsqueda de estudios relacionados con la variable Tiempo de recuperación de los pacientes COVID-19 nos direccionó hacia el artículo de Mollazehi et al. (2020), publicado en Reshear Square, éste presenta un análisis del tiempo de recuperación de un paciente discriminando por algunos grupos o factores que podrían tener influencia (edad, sexo, nacionalidad) para datos de los tres primeros meses de pandemia en Singapur, la técnica utilizada es un análisis de supervivencia paramétrico que comparó el ajuste conseguido al utilizar diferentes distribuciones, en particular los autores probaron las distribuciones: Exponencial, Weibull, Log-Logística, Normal, Log-Normal y Gamma Generalizada; al final, concluyeron sobre el mejor ajuste utilizando el Criterio de Información de Akaike (AIC) el cual los condujo a optar por la distribución log-logística. Finalmente, muestran un ajuste del modelo de regresión de tiempo de

falla acelerado log-logístico a partir del cual concluyen que la edad y la nacionalidad presentan un efecto significativo en el tiempo de recuperación del paciente, mientras el sexo no.

Fruto de esa búsqueda específica en relación con análisis del tiempo de supervivencia encontramos que en el Journal of International Medical Research se publicó el artículo de Gayathri et al. (2020), el cual buscaba establecer los factores asociados al tiempo de recuperación (días desde el ingreso hospitalario hasta que es dado de alta) de un paciente COVID-19, ajustando un modelo de tiempo de falla acelerado a datos de COVID-19 entre el 21 de junio y 30 de agosto de 2020 en la India, la conclusión sobre la distribución con mejor ajuste fue la log-logística y las variables que resultaron afectar el tiempo de estancia hospitalaria fueron la saturación de oxígeno, la lactato deshidrogenasa, la relación neutrófilos-linfocitos, dímero D, ferritina, creatinina, recuento total de leucocitos, edad > 80 años y enfermedad de las arterias coronarias.

Otro estudio encontrado se sitúa en Indonesia, allí Linasari (2021) buscó establecer la probabilidad de supervivencia de un paciente con COVID-19 luego de ser admitidos en el Hospital Dustira y RSUD Cibabat, para hallarla utilizó un análisis de supervivencia desde un enfoque no paramétrico Actuarial (Tablas de vida) y Kaplan-Meier. El estudio encontró inicialmente que los síntomas que más aquejaban a los infectados eran fiebre, tos, dolor de garganta, vómito, anosmia, entre otras. Además, aquellos pacientes con comorbilidades presentaron una menor probabilidad de supervivencia en relación con los que no. Se encontró además que enfermedades como la hipertensión aumenta la probabilidad de fallecimiento de un paciente COVID-19 y que pacientes que presentan una neumonía severa tienen una probabilidad mucho mayor de fallecer que un paciente que no.

En el campo de la estadística bayesiana encontramos el artículo escrito por Khan et al. (2021) publicado en PLOS ONE, en este se presenta un estudio de los datos de COVID-19 desde

el 26 de febrero de 2020 al 20 de marzo de 2021 en Pakistán por medio del ajuste de un modelo lineal dinámico bayesiano (BDLM) para el pronóstico, con veinte días de anticipación, de nuevas infecciones diarias, muertes y recuperaciones diarias; para ello realizaron una simulación del modelo a partir del cual sugieren que el promedio diario de casos nuevos será más alto que los valores medios de los datos observados, las muertes diarias muestran que el promedio durante los próximos veinte días será de 52 y para los casos de recuperación diaria, en promedio habrá 1.840 recuperados por día.

Finalmente destacamos el artículo de Xiang Gao y Qunfeng Dong (2020) que presenta una clara descripción de la metodología bayesiana la cual es ilustrada con un ejemplo práctico donde el objetivo es estimar la prevalencia de la muerte por infección con COVID-19 en Islandia y la prevalencia de niños asintomáticos en Estados Unidos. En forma detallada se describe uno de los aspectos más llamativos del enfoque inferencial bayesiano que consiste en el no uso de estimaciones puntuales a través de intervalos de confianza clásicos, a cambio lo que se hace es la estimación de la distribución de probabilidad posterior para la prevalencia, parámetro desconocido en este caso; de esta forma se logra superar las tres grandes limitaciones de los métodos estadísticos tradicionales: en primera instancia, se suele solo obtener estimaciones puntuales de la prevalencia inferidas a partir de los datos disponibles, dichas estimaciones pueden ser los valores más probables de la prevalencia desconocida pero esta forma de estimar no da cuenta de la posibilidad de que otros valores, con probabilidades no despreciables, puedan ser la prevalencia desconocida.

En segundo lugar, al reportar Intervalos de Confianza del 95% es importante tener en cuenta que estos no representan un rango de valores con una probabilidad del 95 % de contener las estimaciones puntuales, sino que son un rango producido por un procedimiento estadístico de muestreo repetido, tiene una efectividad del 95% en cuanto a contenencia del valor verdadero del

parámetro desconocido. Con relación a lo anterior plantean un segundo elemento novedoso de la inferencia bayesiana y es el tratamiento de la incertidumbre, la cual es medida a través de una distribución de probabilidad. Por último, se plantea como las estimaciones y análisis clásico no incorporan conocimiento previo sobre el parámetro desconocido lo cual llega a ser importante cuando no se tiene un tamaño de muestra significativo y la prevalencia es baja.

Finalmente, las dos aplicaciones utilizadas por Xiang Gao y Qunfeng Dong (2020) muestran detalladamente cómo estimar la prevalencia de sus dos eventos de interés, para hallar la distribución de probabilidad posterior parten de una distribución a priori del tipo beta y una función de verosimilitud binomial, esta combinación resulta ideal para una relación analítica entre ellas, la cual técnicamente se denomina a priori conjugada (ver definición 3.12 en la siguiente sección), una vez hallada la distribución de probabilidad posterior proceden a calcular intervalos de credibilidad al 95% que representan los rangos probables de valores para la prevalencia desconocida en cada uno de los dos casos presentados.

2.3 Metodología

El tipo de estudio que se realiza corresponde a un enfoque cuantitativo que incluye el uso de métodos estadísticos descriptivos e inferenciales. Con el fin de alcanzar los objetivos propuestos, se llevaron a cabo las siguientes etapas:

1. Revisión bibliográfica relacionada con el fundamento teórico de la estadística bayesiana, modelos para análisis de supervivencia tanto en el enfoque clásico como en el bayesiano.

2. Exploración y adecuación de la base de datos “Casos positivos de COVID-19 en Colombia”.

3. Selección del soporte computacional, es decir, definir las librerías de R requeridas para los análisis propuestos y programar los códigos necesarios en R para la realización tanto del análisis descriptivo como del análisis de supervivencia para los modelos propuestos.

4. Análisis e interpretación del comportamiento de los modelos propuestos.

5. Formulación de conclusiones y recomendaciones sobre el comportamiento de los modelos construidos bajo el enfoque clásico y su bondad de ajuste frente a la propuesta bayesiana.

2.4 Marco Teórico

A continuación, se presenta de manera general los elementos teóricos que son el soporte para el análisis de datos en este proyecto, considerando los elementos necesarios a partir de la particularidad de los datos y las necesidades según la evaluación de la bondad de ajuste de los modelos a explorar.

2.4.1 Análisis de supervivencia

El análisis de supervivencia también conocido como análisis del tiempo de falla, análisis de duración, análisis de confiabilidad o análisis del historial de eventos, dependiendo de la situación y el campo de estudio, es un análisis de una variable no negativa llamada tiempo, hasta la ocurrencia de un evento. En el campo de la medicina el ejemplo más común es el estudiar el tiempo desde el diagnostico de una enfermedad hasta la muerte, es decir estudia el tiempo de supervivencia, de allí se deriva el nombre de este análisis en el ámbito de la salud.

2.4.1.1 Conceptos básicos análisis de supervivencia. Uno de los propósitos de este proyecto es analizar y modelar el tiempo de recuperación de un paciente diagnosticado con COVID-19 a través de un análisis de supervivencia, técnica que involucra cierta terminología reservada:

Definición 1 (Evento terminal). Es el evento que determina el momento en que termina el seguimiento o estudio en cada uno de los sujetos, generalmente, este evento suele ser la muerte, el evento terminal no siempre ocurre a cada uno de los sujetos del estudio o análisis.

Definición 2 (Tiempo de supervivencia). Se define como el tiempo transcurrido desde el estado inicial hasta la ocurrencia de un evento dado.

Definición 3 (Tiempo de seguimiento). Es el tiempo que transcurre entre la fecha de entrada en el estudio hasta la fecha registrada en la última observación.

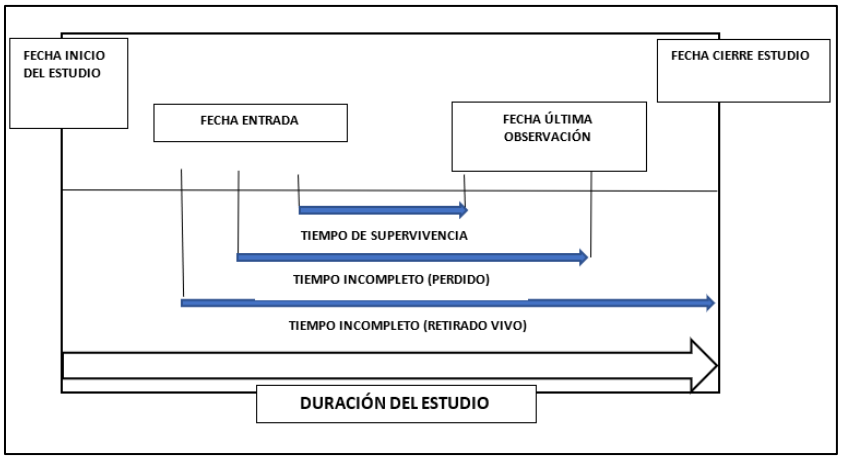
Definición 4 (Censura a derecha). Una de las características que hace diferente el análisis de supervivencia de cualquier estudio de la ocurrencia de un evento son las observaciones censuradas o incompletas, las cuales ocurren cuando no se observa el evento terminal en el paciente durante el tiempo que dura el estudio, el paciente sale o abandona el estudio antes de terminar el período de observación por esto el registro de información es parcial sobre estos pacientes.

Aunque existen otros tipos de censura, para los datos que estamos proponiendo analizar se considera que sólo es posible tener datos censurados a derecha por tanto no se definen los otros tipos.

En la siguiente figura se resume los puntos importantes en un estudio típico del análisis de supervivencia:

Figura 1

Esquema general de un análisis de supervivencia



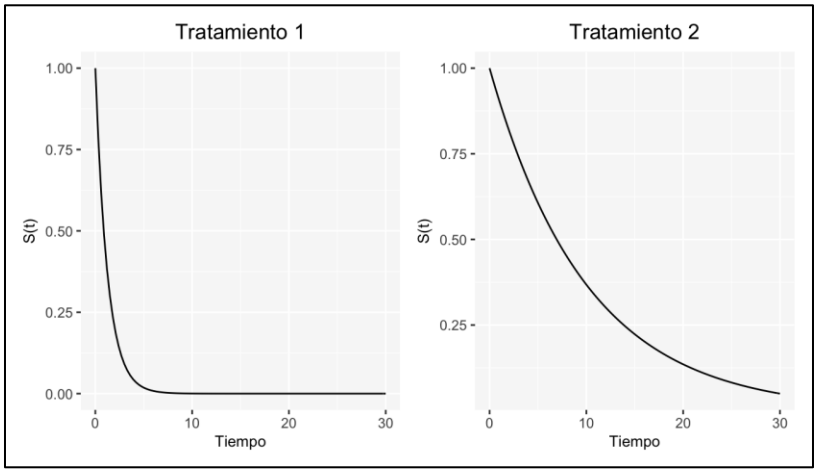
2.4.1.2 Función de supervivencia y función de riesgo.

Definición 5 (Función de supervivencia S(t)). Representa la probabilidad de que un individuo sobreviva ***(no ocurra el evento)*** desde la fecha inicio de seguimiento hasta un momento determinado en el tiempo t, así,

$$S_T(t) = P(T > t), \forall\, t \geq 0$$

Figura 2

Gráficas función de supervivencia



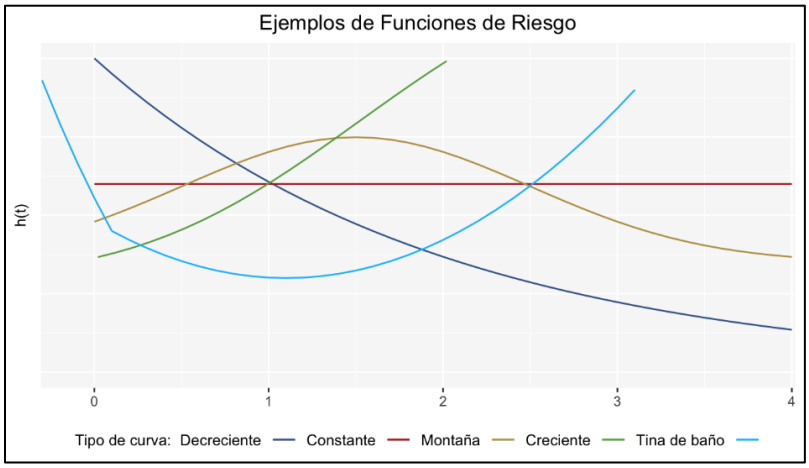
Nota. Las gráficas en la Figura 2 muestran dos ejemplos de gráficas de funciones de supervivencia. Tomado de *Modelos de Supervivencia* (Capítulo Modelos Paramétricos), por Villers et al., 2021.

Definición 6 (Función de riesgo h(t)). representa la probabilidad de que un individuo que está siendo observado en el tiempo t muera ***(suceda el evento).***

$$h(t) = -\frac{d}{dt}\log\left(S(t)\right)$$

Figura 3

Gráficas función de riesgo



Nota. La gráfica muestra algunos ejemplos de comportamientos de gráficas de la función de riesgo. Tomado de *Modelos de Supervivencia* (Capítulo Modelos Paramétricos), por Villers et al., 2021.

2.4.2 Análisis de supervivencia paramétrico

Como su nombre lo sugiere, el análisis de supervivencia implica una distribución de probabilidad donde las principales familias paramétricas son: Exponencial, Weibull, Log-Normal, Log-logística y Gamma, sobre cómo elegir cuál usar, Villers et al. (2021) afirman que “La

motivación para usar un modelo en particular es, por lo general, empírica o bien, con base en la información que proporcione algún modelo no paramétrico.” (Capitulo 5 Modelos Paramétricos).

Es de tener en cuenta que al usar un modelo paramétrico se pueden encontrar beneficios o ganancias para el análisis, dado que con una distribución para los datos que sea absolutamente continua, la función de supervivencia es “suave”, además, que al elegir una S(t) paramétrica se restringe la flexibilidad del modelo, lo cual es bueno cuando se posee una base de datos de tamaño pequeño y se ha ajustado un modelo paramétrico adecuado.

2.4.2.1 Modelo de tiempo de falla acelerado (ATF). Para un tiempo de supervivencia dado T y un vector de covariables X, un modelo ATF supone que el efecto de una covariable es acelerar o desacelerar el curso de vida de una enfermedad en un factor constante.

Se puede formular en la escala logarítmica (similar a la regresión lineal) como

$$Y = \beta_0 + \beta' X + \varepsilon$$

donde $Y = \log(T)$, ε es un término de error aleatorio que se supone sigue alguna distribución paramétrica y β_0 es el intercepto. Para algunas distribuciones (por ejemplo, Weibull) hay un parámetro adicional σ, que escala a ε. En este caso, el modelo AFT se convierte en:

$$Y = \beta_0 + \beta' X + \sigma\varepsilon$$

De lo cual se tiene que:

$$T = e^{(\beta 0 + \beta' X + \sigma\varepsilon)} = e^{(\beta 0)} \times e^{(\beta' X)} \times e^{(\sigma\varepsilon)}.$$

Definición 7 (Distribución gamma generalizada). La distribución gamma generalizada es una distribución de tres parámetros que incluye las distribuciones de Weibull, log-normal y gamma como casos especiales.

La gamma generalizada tiene tres parámetros: $a>0$, $d>0$ y $p>0$. Para X no negativo, la función de densidad de probabilidad de la gamma generalizada se encuentra dada por:

$$f(X;a,d,p)=\frac{\left(\frac{p}{a^d}\right)X^{d-1}e^{-(X/a)^p}}{\Gamma\left(\frac{d}{p}\right)}$$

Donde Γ representa a la función gamma. Donde a *a* es un parámetro de escala mientras que *d* y *p* son parámetros de forma.

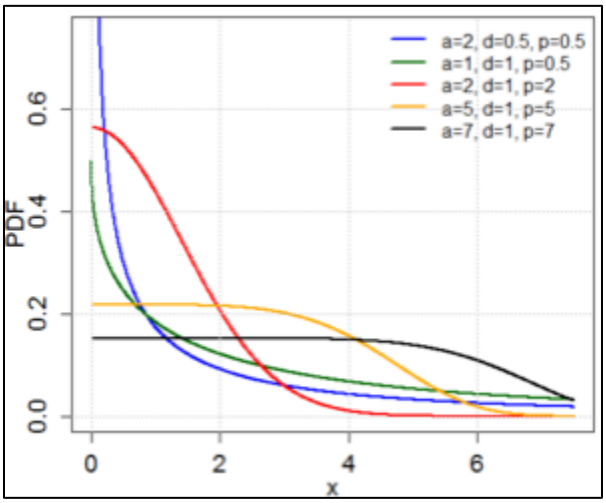
En particular, el lenguaje de programación R y como parte del paquete *flexsurv*, incluye la función "dgengamma" con parametrización:

$$\mu=\ln a+\frac{\ln d-\ln p}{p} \qquad Q=\sqrt{\frac{p}{d}} \qquad \sigma=\frac{1}{\sqrt{pd}}$$

Donde μ es un parámetro de localización.

Figura 4

Curvas gamma generalizada con parámetros distintos



Nota. La gráfica muestra diferentes ajustes de paramétros para la distribución gamma generalizada. Tomado de *Modelo de Tiempo de Falla Acelerado* (Distribución gamma generalizada Caracteristicas y Momentos), por tok.wiki.

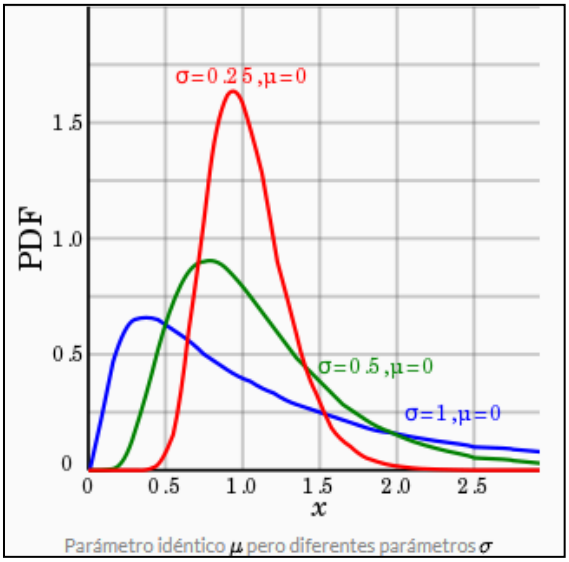
Definición 8 (Distribución log-normal). Es una distribución de probabilidad continua de una variable aleatoria cuyo logaritmo se distribuye normalmente. Por tanto, si la variable aleatoria X tiene una distribución logarítmica normal, $Y = \ln(X)$ tiene una distribución normal luego si Y tiene una distribución normal, así la función exponencial de Y, $X = \exp(Y)$, tiene una distribución logarítmica normal.

Sea Z una variable normal estándar y sea μ y σ dos números reales positivos. Entonces, la distribución de la variable aleatoria $X = e^{\mu+\sigma Z}$ se llama distribución logarítmica normal con parámetros μ y σ, estos son el valor esperado (o media) y la desviación estándar del logaritmo natural de la variable. Para producir una distribución con la media deseada y la varianza deseada μ_X y σ_X^2 se usa

$$\mu = \ln \frac{\mu_X^2}{\sqrt{\mu_X^2 + \sigma_X^2}} \qquad \sigma^2 = \ln 1 + \frac{\sigma_X^2}{\mu_X^2}$$

Figura 5

Curva de una distribución log-normal con parámetros distintos



Nota. La gráfica muestra diferentes ajustes de paramétros para la distribución normal logaritmica (Log normal). Tomado de *Modelo de Tiempo de Falla Acelerado* (Distribución Logaritmica Normal), por tok.wiki.

2.4.3 Análisis de supervivencia no paramétrico

Los métodos en el análisis de supervivencia más utilizados son los no paramétricos, debido a su facilidad de uso al no ser necesario el conocimiento de la distribución que siguen los tiempos de supervivencia. Los usados generalmente, son el método actuarial y el método de método límite-producto de Kaplan-Meier, no obstante, para este proyecto se prevé solo el uso del segundo método el cual será descrito a continuación.

2.4.3.1 Kaplan-Meier (KM). Este método calcula la supervivencia cada vez que un paciente muere o presenta el evento, éste da proporciones exactas de supervivencia debido a que utiliza tiempos de supervivencia precisos.

Sea S(t) la función de supervivencia de una población, y sean $0 \leq t_1 \leq \ldots \leq t_n$ los tiempos hasta la ocurrencia de la muerte o presencia del evento en todos los participantes, también se definen:

- d_j, el número de ocurrencia de la muerte en el momento t_i
- n_j, el número de sujetos en riesgo justo antes de t_i.

De no haber censura, n_j es el número de supervivientes inmediatamente antes del momento t_j. Con censura es el número de supervivientes menos el número de casos censurados. A partir de la notación anterior se plantea el estimador por Kaplan-Meier como se indica en la siguiente expresión:

$$S(t) = \prod_{t_j<t} \frac{n_j - d_j}{n_j}.$$

Aquí es claro que al tener el supuesto que los eventos ocurrirán de manera independiente, las probabilidades de sobrevivir de un intervalo al siguiente se multiplican y así se halla una probabilidad de sobrevivir acumulada.

Sobre el proceso de estimación al usar Kaplan-Meier, la probabilidad de supervivencia se puede estimar de forma no paramétrica a partir de los tiempos de supervivencia observados, tanto censurados como no censurados, utilizando el método KM como se ya se indicó en el apartado anterior, así, más formalmente diremos que la probabilidad de estar vivo en el momento t_j , representada por S(t_j), se calcula a partir de S(t_{j-1}) que representa la probabilidad de estar vivo en el momento t_{j-1}, la expresión viene dada por

$$S(t_j) = S(t_{j-1})\left(1 - \frac{d_j}{n_j}\right)$$

donde $t_0 = 0$ y $S(0) = 1$, y n_j, el número de pacientes vivos justo antes de t_j, y d_j, el número de eventos en t_j,

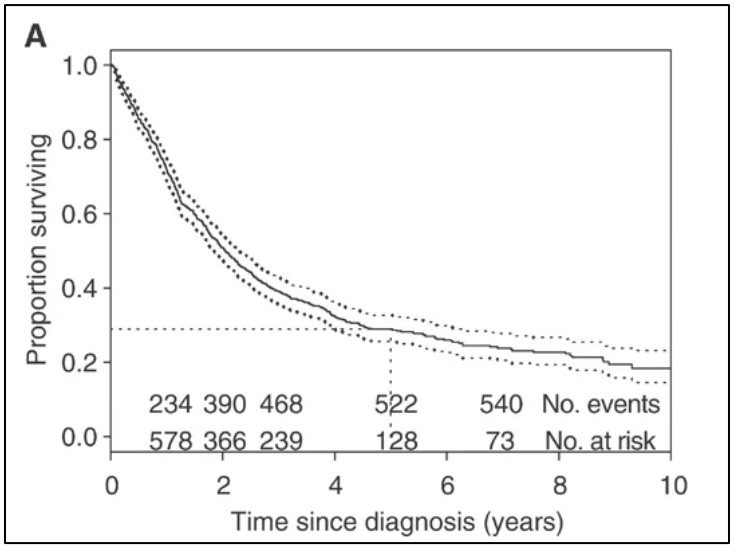
El valor de S(t) es constante entre eventos, por lo tanto, la probabilidad estimada es una función escalonada que cambia de valor solo en el momento de cada evento. Este estimador permite que cada paciente contribuya con información a los cálculos mientras se sepa que no tiene el evento. Si todos los individuos experimentaran el evento (es decir, sin censura), este estimador simplemente se reduciría a la proporción del número de eventos individuales libres en el momento *t* dividido por el número de personas que ingresaron al estudio.

Adicionalmente, se pueden calcular intervalos de confianza para la probabilidad de supervivencia. La curva de supervivencia de KM es un gráfico de la probabilidad de supervivencia

frente al tiempo que proporciona un resumen útil de los datos que se puede utilizar para estimar medidas como el tiempo de supervivencia medio (Ver figura 6).

Figura 6

Gráfica curva de supervivencia y sus intervalos de confianza



Nota. El gráfico muestra un ejemplo de una curva de supervivencia con sus intervalos de confianza al 95%, para un estudio del tiempo desde el diagnóstico de cáncer de ovario hasta la muerte. Tomado de *Survival Analysis Part I: Basic concepts and first analyses* (Survival function of the ovarian data), por Clark et al., 2003, Br J Cancer.

2.4.4 Evaluación del ajuste de los modelos de supervivencia

En el modelamiento estadístico, generalmente se realizan para el estudio de una variable en un grupo de datos, múltiples ajustes de modelos, en este caso de modelos de supervivencia, de los cuales se busca encontrar el que presente el mejor ajuste, existen diversos criterios para la elección de un modelo según sea la técnica empleada, en este trabajo se utilizarán los siguientes criterios:

2.4.4.1 Criterio de información de Aiken (AIC). Es una medida de la bondad de ajuste de un modelo estadístico. Se puede decir que describe la relación entre el sesgo y varianza en la construcción del modelo, o hablando de manera general acerca de la exactitud y complejidad del modelo. En general, se define como:

$$AIC = 2k - 2\ln(L)$$

Donde:

- k es el número de parámetros.
- $\ln(L)$ es la función de log-verosimilitud para el modelo estadístico.

2..4.4.2 Criterio de información bayesiana (BIC) o criterio Schwarz (SIC). Es una medida de bondad de ajuste de un modelo estadístico. Generalmente utilizada para la selección de modelos en un conjunto finito de modelos, se encuentra estrechamente relacionado con el AIC. Es un criterio de evaluación de modelos en términos de sus probabilidades posteriores. En general, se define como:

$$BIC = k\ln(n) - 2\ln(L)$$

- k es el número de parámetros.
- $\ln(L)$ es la función de log-verosimilitud para el modelo estadístico.
- n es el tamaño de muestra.

Tanto en el AIC y el BIC, el modelo que presente el mínimo valor correspondiente en ese criterio será el modelo elegido. Al comparar los criterios anteriores se tiene que la diferencia entre ellos radica en que el BIC penaliza en mayor medida los modelos con mayor número de parámetros.

2.4.5 Fundamentos estadística bayesiana

La estadística bayesiana es una nueva forma de tratar y abordar problemas estadísticos, una mirada a un análisis más completo y real, particularmente, Correa y Barrera (2018) afirman:

"Consideramos que si la estadística bayesiana se diferencia en algo de la estadística tradicional (clásica) es en permitirle al usuario incorporar información disponible de una manera transparente y directa." (p. II).

Este campo de la estadística que durante mucho tiempo fue dejada de lado, con el avance tecnológico y disponibilidad de software estadístico especializado ha cobrado vida y ha comenzado a mostrar su relevancia e importancia para responder a diferentes objetivos de análisis. A continuación, se expondrán algunos de los principales elementos teóricos de Estadística Bayesiana que se usarán en la etapa de ajuste de un modelo para describir los datos de COVID-19 en Colombia.

2.4.5.1 Probabilidad y verosimilitud. Entre los conceptos base que tiene la estadística bayesiana esta la probabilidad condicional y la verosimilitud que se definen como sigue:

Definición 9 (Probabilidad condicional). Sean los eventos A y B definidos sobre un espacio muestral S, la probabilidad condicional de A dado B, *P (A | B)*, mide la probabilidad de observar A, a partir de la información de que B ya ocurrió. Esta probabilidad se define como:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

para $P(B) \neq 0$.

Definición 10 (Función de verosimilitud). Dada una muestra $X = (x_1, \dots, x_n)$ y parámetros, $\theta = (\theta_1, \dots, \theta_n)$ entonces,

$$L(\theta|X) = \prod_{i=1}^{n} f(x_i, \theta).$$

Representa la probabilidad de observar un vector de parámetros dada una muestra de datos observados, aquí $f(x_i, \theta)$ representa una función de densidad.

2.4.5.2 Teorema de Bayes. Dentro de la estadística clásica muchas veces no se le da el estudio o la relevancia que realmente tiene al teorema de Bayes para un buen análisis, en la estadística bayesiana, este teorema cobra gran importancia al extender este al contexto de variables aleatorias.

Teorema 1 (Teorema de Bayes). Sean B_1, B_2, ..., B_k eventos mutuamente excluyentes y exhaustivos, para cualquier evento nuevo A, tenemos:

$$P(B_i|A) = \frac{P(B_i \cap A)}{P(A)} = \frac{P\ (A|B_i)P(B_i)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}$$

$P(A) \neq 0$, $P(B_i) \neq 0$, i = 1, 2, . . ., k.

Teorema 2 (Teorema de Bayes para variables aleatorias). Sean X y θ variables aleatorias con funciones de densidades de probabilidad (fdp's) $f(x|\theta)$ y $\xi(\theta)$.

$$\xi\ (\theta|x) = \frac{f\ (x|\theta)\xi(\theta)}{\int_{\Theta} f\ (x|\theta)\xi(\theta)\ d\theta}.$$

Debe notarse en la anterior formulación que se incorpora θ como variable aleatoria cuando en la notación usual en Inferencia las letras griegas están restrictas a representar parámetros, precisamente eso es lo novedoso en este enfoque, la expresión describe cómo hallar la distribución del parámetro desconocido con lo cual la estimación de θ se hará a través de la distribución $\xi\ (\theta|x)$ que se denominará Distribución a posterior y se describirá en 2.4.5.4.

2.4.5.3 Distribución a priori. Para la estadística bayesiana el concepto de distribución a priori, es uno de los más importantes junto al Teorema de Bayes, esta busca resumir la incertidumbre sobre los parámetros poblacionales en la primera etapa del análisis. Para Ashby (2006, citado por Correa y Barrera, 2018):

> Tres interpretaciones se le pueden dar a las distribuciones a priori: como distribuciones de frecuencia basadas quizá en datos previos, como representaciones normativas y objetivas de lo que es racional creer acerca de un parámetro o como una medida subjetiva de los que un individuo particular realmente cree. (p. 14).

A continuación, se muestra una clasificación de a prioris que presentan Correa y Morales (2018) son:

Definición 11 (Distribución propia). Una distribución propia asigna pesos no negativos y que suman o integran hasta uno, a todos los valores posibles del parámetro.

Mientras que una impropia, suma o integra hasta un valor diferente de uno.

Definición 12 (Distribución no informativa). Una distribución a priori es no informativa cuando presenta información limitada o un desconocimiento total sobre el parámetro de interés. De lo contarios se habla de una a priori informativa.

Definición 13 (Distribución conjugada). Decimos que una distribución a priori es conjugada, si al proceder a su actualización mediante la información muestral, la distribución a posteriori es igual a la a priori, excepto en los hiperparámetros, es decir, en parámetros distintos a los del modelo muestral, si esto no sucede hablamos de una a priori no conjugada.

El proceso que busca la obtención de una distribución a priori subjetiva, y cuantifica la información de expertos u otros medios se le conoce como ***elicitación*** de la distribución a priori.

2.4.5.4 Distribución a posteriori. La inferencia bayesiana se basa y fundamenta en el estudio de la distribución que será llamada a posteriori, esta se deduce a partir del Teorema de Bayes para variables aleatorias:

$$\xi(\theta|x_1,\ldots,x_n) = \frac{f\ (x_1,\ldots,x_n|\theta)\xi(\theta)}{\int_\Theta\ f\ (x_1,\ldots,x_n|\theta)\xi(\theta)d\theta}.$$

Donde, acorde al marco de referencia bayesiano se tiene que:

- *X*: datos (escalar o vector o matriz)
- θ: parámetro desconocido (escalar o vector o matriz)
- $f\ (x_1, \ldots, x_n|\theta)$: verosimilitud de los datos dado el parámetro (desconocido) θ.
- $\xi(\theta)$: distribución a priori de θ.

En la práctica, el denominador de la expresión no requiere ser necesariamente calculado en general, por ello la regla de Bayes suele ser presentada de manera resumida como:

$$\xi\ (\theta|x_1,\ldots,x_n)\ \propto\ f\ (x_1,\ldots,x_n|\theta)\ \xi(\theta).$$

El aprendizaje bayesiano simbólicamente queda expresado en la siguiente secuencia de expresiones:

$$\xi\ (\theta|x_1) \propto f\ (x_1|\theta)\ \xi(\theta),$$

$$\xi\ (\theta|x_1, x_2) \propto f\ (x_2|\theta)\ f\ (x_1|\theta)\ \xi(\theta),$$

$$\propto f\ (x_2|\theta)\ \xi\ (\theta|x_1).$$

Adicionalmente, Correa y Barrera (2018) resaltan sobre este proceso que: “Por lo tanto el teorema de Bayes nos muestra cómo el conocimiento acerca del estado de la naturaleza representada por θ es continuamente modificada a medida que nuevos datos son adquiridos.” (p. 26).

2.4.5.5 Cadenas de Markov Monte Carlo (MCMC). Cuando una posterior bayesiana se encuentra imposible o verdaderamente difícil de hallar analíticamente, se debe buscar aproximar mediante simulación, el medio de aproximación usual es el uso de cadenas de Markov Monte Carlo.

Las cadenas de Markov Monte Carlo son la aplicación de cadenas de Markov para simular modelos de probabilidad utilizando métodos iniciados por el proyecto Monte Carlo. Las muestras obtenidas por MCMC no son tomadas directamente de las pdf´s a posteriori $\xi(\theta| y)$, y cada valor de muestra posterior depende directamente del valor de la anterior (no son independientes). En general, el valor (i+1) de la cadena $\theta^{(i+1)}$ se extrae de un modelo que depende de los datos y el valor de la cadena anterior $\theta^{(i)}$ con la función de probabilidad condicional

$$\xi(\theta^{(i+1)} \mid \theta^{(i)}, y).$$

Hay un par de cosas a tener en cuenta sobre esta dependencia entre los valores de la cadena, primero, por la propiedad de Markov, $\theta^{(i+1)}$ depende de los valores de la cadena anterior solo a través del valor más reciente $\theta^{(i)}$:

$$\xi(\theta^{(i+1)} \mid \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(i)}, y) = \xi(\theta^{(i+1)} \mid \theta^{(i)}, y).$$

(Johnson et al., 2021, Markov chains vía rstan, párr.3).

Además, cada valor de la cadena se puede extraer de un modelo diferente, y ninguno de estos modelos es el objetivo posterior. Es decir, la pdf a partir del cual se simula un valor de la cadena de Markov no es equivalente a la pdf posterior. (Johnson et al., 2021, Markov chains via rstan, párr.4).

Definición 14 (Muestreo de Gibbs). Es el método de tipo MCMC utilizado por defecto en los paquetes de R para modelamiento bayesiano, el muestreo de Gibbs que consiste en:

Muestrear de la distribución $\xi(\theta)$ donde $\theta = \theta_1, ..., \theta_k$. El algoritmo de Gibbs permite generar muestras de la posterior:

$$\xi(\theta_1, \theta_2, \dots, \theta_p | y)$$

siempre y cuando podamos generar valores de todas las distribuciones condicionales:

$$\xi(\theta_k | \theta_1, \theta_2, \dots, \theta_{k-1}, \theta_{k+1}, \dots, \theta_p, y).$$

El algoritmo Gibbs fija un valor inicial $\theta^{(0)}$ y después muestra valores sucesivamente de las distribuciones condicionales. Es una cadena de Markov y tiene distribución del equilibrio igual a $\xi(\theta)$. El proceso del muestreador de Gibbs es una caminata aleatoria a lo largo del espacio de parámetros.

2.4.6 Inferencia bayesiana

En el análisis bayesiano se pueden encontrar algunas diferencias con la estadística clásica en la forma de interpretar ciertos conceptos o realizar procesos como pruebas de hipótesis e interpretación de intervalos de credibilidad.

2.4.6.1 Intervalos de credibilidad. Una de las características que se puede notar en el análisis bayesiano es que ya no vamos a hablar de intervalos de confianza a un cierto nivel de confianza, sino que hablaremos de intervalos de credibilidad.

- La estimación se reduce a un problema de decisión. En situaciones distintas se eligen estimadores diferentes: **se usa la teoría de utilidad para elegirlos.**
- Un intervalo de credibilidad del 95 % para θ es un intervalo tal que la probabilidad de que este contenga a θ igual a 0,95.

Una forma de elegir el intervalo de credibilidad es utilizando la región de densidad posterior más alta (RDPMA). Box y Tiao (1973) definen una región de credibilidad como:

2.4.7 Análisis de supervivencia bayesiano

Para la parte bayesiana de este trabajo se van a tener en cuenta los siguientes conceptos y procesos básicos al hacer el ajuste de un modelo de falla acelerado.

2.4.7.1 Proceso de Dirichlet (DP). El proceso de Dirichlet se define a través de la distribución de Dirichlet, Bogaerts et al. (2017) presenta la noción formal del proceso de Dirichlet adaptado al contexto de supervivencia de la siguiente manera: Sea c > 0 y S(t) una función de supervivencia continua por la derecha definida en [0,1), Un proceso de Dirichlet (DP) con parámetros (c, S (t)) es una función de supervivencia aleatoria S, que para cada partición finita (medible) $\{B_1, B_2, \dots, B_k\} \in \mathbb{R}^+$ asigna probabilidades a los B_j tal que la distribución conjunta del vector de probabilidades:

$$S(B_j)\{i = 1, \dots, k\},$$

$$(S(B_1);\ S(B_2); \dots ;\ S(B_k))$$

es la distribución de Dirichlet

$$Dir\ (cS\ (B_1);\ cS\ (B_2); \dots;\ cS\ (B_k)).$$

Definición 15 (Distribución de Dirichlet). Supongamos que $\theta = (\theta_1, \theta_2, \dots, \theta_J)$ es un vector J-dimensional de parámetros continuos que para $\theta_j > 0$ para j=1, …, J satisfacen:

$$\sum_{j=1}^{J} \theta_j = 1$$

entonces la a priori de Dirichlet **Dir(α)** viene dada por:

$$p(\theta) = \frac{\Gamma(\sum_{j=1}^{J} \alpha_j)}{\prod_{j=1}^{J} \Gamma(\alpha_j)} \prod_{j=1}^{J} \theta_j^{\alpha_{j-1}}.$$

Con $\boldsymbol{\alpha_j = (\alpha_1, \dots, \alpha_J)}$ con $\alpha_j > 0$ para j=1, …, J.

El a priori de Dirichlet es una a priori popular para los parámetros de la verosimilitud multinomial, por su propiedad de conjugación, las a priori de Dirichlet más populares son:

- $Dir(1, \dots, 1)$ (**a priori uniforme**).
- $Dir(0.5, \dots, 0.5)$ (**a priori de Jeffreys**).

2.4.7.2 Mezcla gaussiana clásica (GMM). Para la densidad g(y) de $Y = \log(T)$ un modelo flexible es proporcionado por una mezcla gaussiana (desplazada y escalada):

$$g(y) = g(y;\theta) = \frac{1}{\tau}\sum_{k=1}^{K} w_k \varphi_{\mu k,\sigma_k^2}\left(\frac{y-\alpha}{\tau}\right)$$

donde $\propto \in \mathbb{R}$ es el parámetro de desplazamiento fijo $\tau > 0$ es el parámetro de escala fija, y donde $\varphi_{\mu_k,\sigma_k^2}$ denota la distribución gaussiana $N(\mu_k, \sigma_k^2)$. La densidad de supervivencia f se da entonces como

$$f(t) = f(t;\theta) = t^{-1} g\{\log(t)\,;\theta\}$$

De hecho, teóricamente es suficiente considerar $\alpha = 0$ y $\tau = 1$. Sin embargo, para la estabilidad numérica de los procedimientos MCMC, es util elegir α y τ tal que la distribución

$$\tau^{-1}(Y - \alpha)$$

tenga aproximadamente una media de 0 y una varianza de 1.

El vector de parámetros desconocidos del Modelo AFT con una mezcla gaussiana ajustado en la parte bayesiana está compuesto por K, el número de componentes de la mezcla, $w = (w_1, \dots, w_K)$ los pesos de la mezcla, $\mu = (\mu_1, \dots, \mu_K)$ las medias y $d = (\sigma_1^2, \dots, \sigma_K^2)$ las varianzas, así:

$$\theta = (K, w, \mu, d).$$

2.4.7.3 Modelo AFT con una mezcla gaussiana clásica como distribución de error. En el modelo de falla acelerado con una mezcla gaussiana como distribución de error busca fortalecer el modelo AFT frente a supuestos de distribución mal especificados, así una densidad del término de error, ahora se especifica como una mezcla gaussiana

$$g(e) = \frac{1}{\tau}\sum_{k=1}^{K} w_k \varphi_{\mu_k,\sigma_k^2}\left(\frac{e-\alpha}{\tau}\right).$$

Sin pérdida de generalidad se puede considerar α = 0 y τ = 1, además de los parámetros desconocidos para una GMM el vector de los parámetros desconocidos del modelo CGM AFT contiene adicionalmente el vector β de coeficientes de regresión. Así se tendrá que:

$$\theta = (K, w, \mu, d, \beta).$$

2.4.8 Procesamiento datos en R

El software estadístico utilizado para el procesamiento de los datos en este trabajo es R bajo la interfaz de R-Studio. Los paquetes estadísticos que se usaron para el procesamiento de los datos son:

2.4.8.1 flexsurv(). Supervivencia paramétrica flexible y modelos multiestado. Paquete estadístico que trabaja con modelos paramétricos flexibles para datos de tiempo hasta el evento, incluye la gamma generalizada, la F generalizada y el modelo spline de Royston-Parmar, y extensible a distribuciones definidas por el usuario.

2.4.8.1.1 flexsurvreg(). Se usa para realizar un modelado paramétrico o regresión para datos de tiempo hasta el evento. Tiene disponibles varias distribuciones integradas y los usuarios pueden proporcionar las suyas propias.

Las distribuciones integradas son:

Tabla 1

Distribuciones integradas en flexsurv() y sus parámetros

	Parámetros (localización en gris)	**Función de densidad en R**	**Dist**
Exponencial	**Intercepto**	Exp	"exp"
Weibull (tiempo de falla acelerado)	Forma, **escala**	dweibull	"weibull"
Weibull (riesgos proporcionales)	Forma, **escala**	dweibullPH	"weibullPH"
Gamma	Forma, **intercepto**	dgamma	"gamma"
Log-normal	**Media logarítmica**, sd logarítmica	dlnorm	"lnorm"
Gompertz	Forma, **intercepto**	dgompertz	"gompertz"
Log-logística	Forma, **escala**	dllogis	"llogis"
Gamma generalizada (Prentice 1975)	$\boldsymbol{\mu}, \sigma, Q$	lgengamma	"gengamma"
Gamma generalizada (Prentice 1962)	Forma, **escala**, k	dgengamma.orig	"gengamma.orig"
F generalizada (estable)	$\boldsymbol{\mu}, \sigma, Q, P$	Dgenf	"genf"
F generalizada (original)	$\boldsymbol{\mu}, \sigma, s_1, s_2$	dgenf..orig	"genf.orig"

Nota. La tabla presenta un resumen de las distribuciones que flexsurv resaltando en negrilla los parámetros de localización de cada una, así como su nombramiento en el paquete. Tomada y traducida de *flexsurv: A Platform for Parametric Survival Modelling in R* (Table 1: Built-in parametric survival distributions in flexsurv), Jackson, 2016.

2.4.8.2 survival(). Contiene las rutinas principales de análisis de supervivencia, incluida la definición de objetos Surv, curvas de Kaplan-Meier y Aalen-Johansen (multiestado), modelos de Cox y modelos paramétricos de tiempo de falla acelerada.

2.4.8.2.1 survfit(). Esta función crea curvas de supervivencia a partir de una fórmula por ejemplo, Kaplan-Meier, un modelo de Cox previamente ajustado o un modelo de tiempo de falla acelerado previamente ajustado.

El análisis de supervivencia implementado en R a través de la función surfit() utilizada en el modelamiento por Kaplan-Meier retorna la siguiente lista de variables:

- **tiempo:** los puntos de tiempo en la curva.
- **n.risk:** el número de sujetos en riesgo en el tiempo t.
- **n.event:** el número de eventos que ocurrieron en el tiempo t.
- **n.censor:** el número de sujetos censurados, que salen del conjunto de riesgo, sin evento, en el tiempo t.
- **lower,upper:** límites de los intervalos de confianza inferior y superior para la curva, respectivamente.

2.4.8.3 Survminer(). Es un paquete que presenta funciones de graficado y presentación de información como: la función “ggsurvplot()” para dibujar fácilmente curvas de supervivencia y la tabla de “número en riesgo” y el “gráfico de recuento de censura”. La función de este paquete usada en este trabajo para el gráfico de curvas de supervivencia es:

2.4.8.4 ggsurvplot(). Es una función genérica para trazar curvas de supervivencia. Está formado por una familia de funciones ggsurvplot_xx() como:

- ggsurvplot_list()
- ggsurvplot_facet()
- ggsurvplot_group_by()
- ggsurvplot_add_all()
- ggsurvplot_combine()

2.4.8.4 BayesSurv(). Este paquete contiene implementaciones bayesianas de Modelos de tiempo de falla acelerados de efectos mixtos (MEAFT) para datos censurados. Ya sean no solo censurados por la derecha, sino también censurados por intervalos, censurados por intervalos dobles o clasificados erróneamente como censurados por intervalos.

bayessurvreg1(). Es una función para muestrear de la distribución a posteriori para un modelo de regresión de supervivencia

$$\log(T[i,l]) = \beta x[i,l] + b[i]z[i,l] + \varepsilon[i,l],$$

$$i = 1,\ldots,N, \qquad l = 1,\ldots,n[i]$$

donde la distribución de $\varepsilon[i,l]$ se especifica como una mezcla normal con un número desconocido de componentes y el efecto aleatorio $b[i]$ se distribuye normalmente.

2.4.8.5 BayesSurvival(). Este paquete es usado y adecuado para análisis no ajustados de datos de supervivencia censurados por la derecha, tiene en su base dos distribuciones a priori las cuales son:

Definición 16 (A priori gamma dependiente e independiente). Sea Gamma(α, β), la distribución Gamma con parámetro de forma α y parámetro de escala β. Su función de densidad puede escribirse como

$$f_{\alpha,\beta}(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}.$$

La a priori es constante por tramos y se define en el riesgo λ. El intervalo de tiempo en estudio, denotado por [0, τ], se divide en K intervalos de igual tamaño, denotados por I_k, k=1, …, K. La a priori en λ se puede escribir como

$$\lambda(t) = \sum_{k=1}^{K} \lambda_k 1\{t \in I_k\}$$

donde λ_1 proviene de una distribución Gamma(α_0, β_0) y λ_k, para k=2,…,K, se extrae de una distribución Gamma(α, α / λ_{k-1}). La estructura dependiente de la a priori se refleja en la media y la varianza previas:

$$E[\lambda_k|\lambda_{k-1}, \dots, \lambda_1] = \lambda_{k-1},$$

$$Var[\lambda_k|\lambda_{k-1}, \dots, \lambda_1] = \left(\frac{\lambda_{k-1}}{\alpha}\right)^2$$

para k=2, …, K.

La información de la distribución a priori gamma independiente es la misma a la distribución a priori dependiente con una diferencia clave: cada λ_k se extrae independientemente de una distribución Gamma (α, β).

Las funciones del paquete BayesSurvival utilizadas son:

2.4.8.5.1 BayesSurv(). Calcula cantidades relevantes para un análisis de supervivencia bayesiano de datos de tiempo hasta el evento (posiblemente censurados por la derecha). Comenzando con una a priori exponencial Gamma dependiente o independiente por partes.

2.4.8.5.2 PlotBayesSurv(). Esta función toma la salida de BayesSurv y usa ggplot2 para hacer gráficos de la media posterior de la función de supervivencia con una banda de credibilidad.

3. Resultados

3.1 Análisis exploratorio de la base de datos

La fuente de los datos que sirven de base a este proyecto es la base datos Casos positivos de COVID-19 en Colombia creada por el Instituto Nacional de Salud (INS) que inicia con los datos de la primera paciente reportada con COVID 19 el 6 de marzo de 2020 y se actualiza diariamente, no obstante, para este trabajo de consideraran sólo los datos reportados hasta el 31 de diciembre de 2021, es decir 5.257.543 registros.

3.1.1 Descripción de variables

La base de datos original presenta 23 variables de las cuales se tendrán en cuenta las que se indican a continuación:

Tabla 2

Variables base de datos original

Nombre	Etiqueta	Tipo	Valores
ID caso	ID	Numérica	
Fecha de notificación	F_not	Cualitativa (cadena con año-mes-día)	Año: 2020-2021 Mes: 01-12 Día: 01-31
Fecha inicio de síntomas	F_Inicio_Síntomas	Cualitativa (cadena con año-mes-día)	Año: 2020-2021 Mes: 01-12 Día: 01-31
Fecha de muerte	F_muerte	Cualitativa (cadena con año-mes-día)	Año: 2020-2021 Mes: 01-12 Día: 01-31
Edad	Edad	Numérica	
Unidad de medida de la edad	Unidades_Edad	Categórica	1-Años 2-Meses 3-Días
Sexo	Sexo	Categórica	M- Masculino F- Femenino

Nombre	Etiqueta	Tipo	Valores
Pertenencia étnica	Etnia	Categórica	1- Indígena 2- Rom 3- Raizal 4- Palenquero 5- Negro 6- Otro
Nombre Municipio	Municipio	Nominal	
Ubicación del caso		Categórica	Casa Hospital Hospital UCI Fallecido N/A
Recuperado		Categórica	Recuperado Fallecido N/A (Vacío)
Fecha de recuperación	F_recuperación	Cualitativa (cadena con año-mes-día)	Año: 2020-2021 Mes: 01-12 Día: 01-31
Tipo de recuperación	T_recuperación	Categórica	PCR Tiempo clínico

Nota. La tabla presenta las variables originales de la base "Casos positivos COVID-19" que fueron utilizadas o necesarias para el tratamiento ya análisis de los datos.

La base de datos Casos positivos de COVID-19 en Colombia fue complementada con la base de datos "Clima por municipio" la cual es suministrada por el Ministerio de Vivienda et al. (2020), de allí se tomó la variable Clima leída en una escala que incluye cuatro categorías como se puede ver en la Tabla 3. Adicional al clima, también se reestructuro la variable Pertenencia étnica de la base de datos de acuerdo a la agrupación de grupos étnicos del censo nacional del 2018 suministrada por DANE (2018) y se crearon también algunas variables para complementar el estudio del Tiempo de recuperación las cuales se describen a continuación:

Tabla 3

Variables creadas

Nombre	Etiqueta	Escala	Valores
Tiempo	Tiempo	Días	Tiempo de supervivencia
Delta	Delta	Categórica	0- Censurado 1- No censurado – recuperado por PCR.
Clima	Clima	Categórica	1-Frío 2-Templado 3-Cálido Seco 4-Cálido Húmedo
Etnia	Etnia	Categórica	1- Indígena 2- ROM 3- Negro, afrodescendiente, raizal o palenquero. 4- Otro
Grupo de Edad	Gedad	Categórica	1- [0-15] años 2-(15-25] años 3-(25-50] años 4- más de 50 años

Nota. La tabla presenta las variables que fueron creadas a partir de información de la base de datos original o adquiridas de otras fuentes y que fueron necesarias para el análisis.

3.1.2 Descripción de la muestra

Para este estudio se consideró la base de datos sobre casos positivos de COVID-19 con corte a diciembre de 2021 momento en el cual se tenían 5.257.543 individuos reportados.

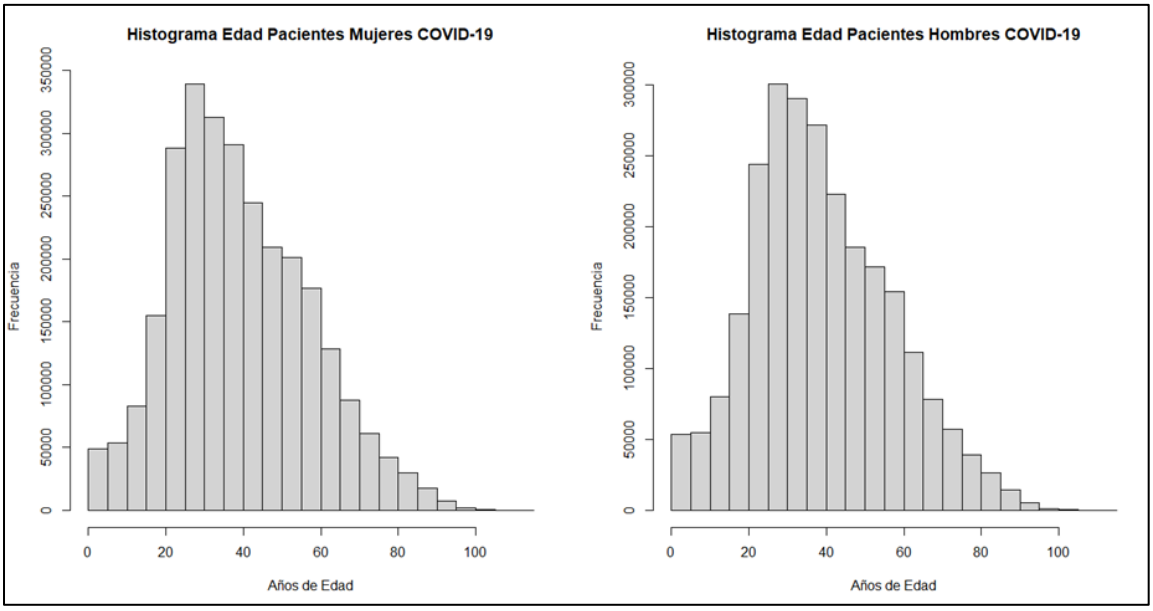
En esta sección presentaremos un análisis descriptivo acorde a las variables disponibles y a los indicadores básicos para este contexto de análisis como son la cantidad de contagiados y el número de muertos por causa del virus comparando su comportamiento al discriminar por las

variables categóricas a disposición que permiten explorar la presencia de un efecto atribuible a estas. No se presentan los consolidados nacionales dado que ha sido información ampliamente difundida en medios de comunicación y el propósito es presentar otros puntos de vista de esta problemática acorde a las variables que integran nuestra base de datos.

3.1.2.1 Sexo. Del total de casos reportados en el período en consideración se tiene que un 47 % son hombres. La Figura 7 muestra que la distribución de la variable Edad para los pacientes positivos para COVID es similar tanto para hombres como para mujeres, en particular muestra como rasgos característicos un comportamiento asimétrico a derecha con una alta concentración de contagios en el rango de edad entre los 20 y 40 años, a partir de esto concluimos que no hay diferencias en la distribución del número de contagios por rango de edad debidas al género.

Figura 7

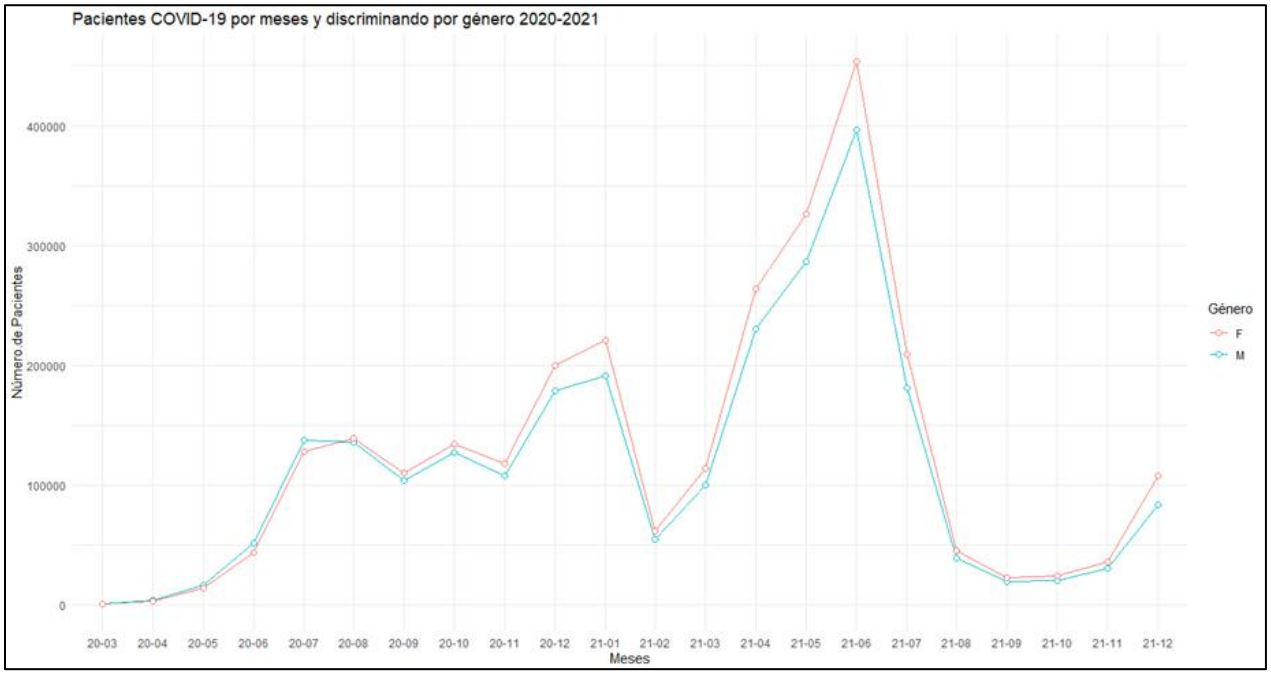
Distribución de la edad contagiados COVID-19 por género



La Figura 8 muestra un gráfico con la evolución en el número de contagios desde marzo del 2020 hasta el 31 de diciembre de 2021 discriminando por mes y por género, se observa que en los primeros meses de la pandemia, marzo-agosto del 2020, la cantidad de contagios en hombres fue levemente mayor que en mujeres pero a partir de septiembre de 2020 la cantidad de contagios en mujeres se muestra levemente por encima en relación con el género masculino; este comportamiento permite afirmar que el número de contagios no ha descrito diferencias significativas debidas al género excepto tal vez por las diferencias que se presentaron en los picos de diciembre de 2020 - enero 2021 y mayo - julio de 2021 donde las diferencias rondaron las 50.000 personas en perjuicio de las mujeres.

Figura 8

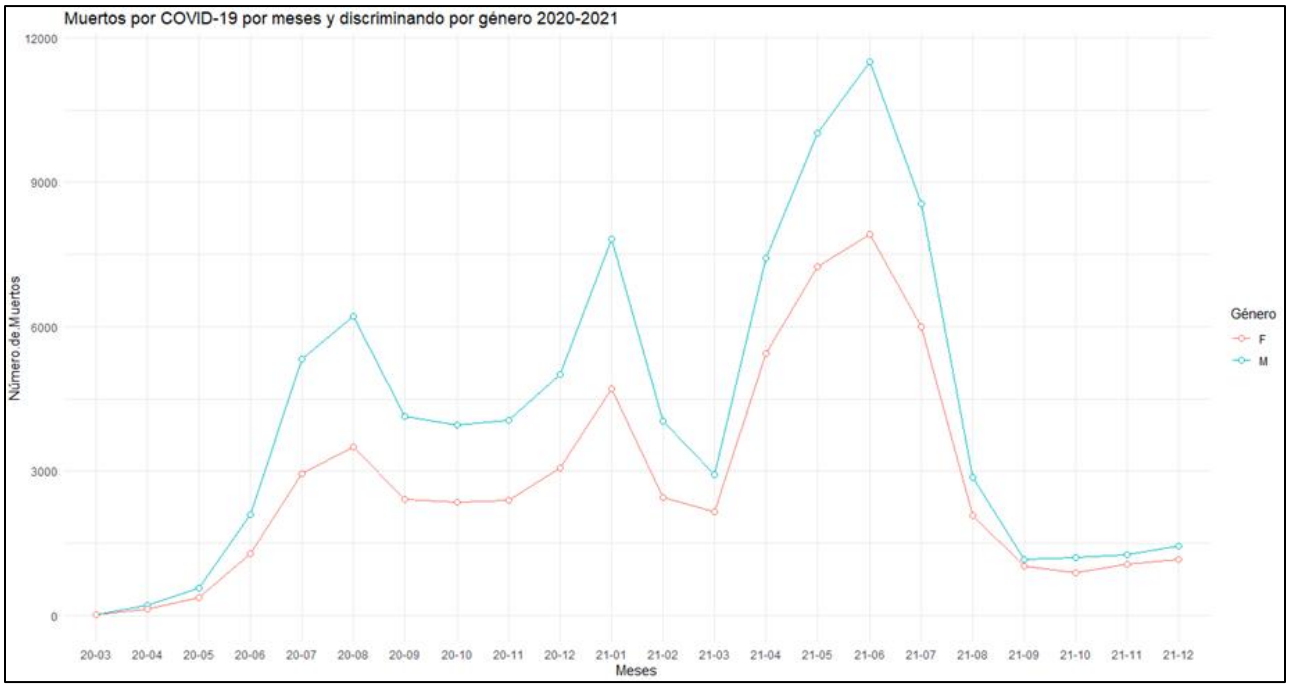
Contagiados COVID-19 por género Colombia 2020-2021



Ahora, al realizar un análisis similar, pero en relación con la cantidad de muertes (Ver Figura 9) se tiene que desde marzo de 2020 hasta diciembre de 2021 la diferencia en el número de fallecimientos en hombres a lo largo de este período de tiempo fue mayor, alcanzando diferencias de hasta 50.000 individuos (junio de 2021). Además, es claro que las mayores diferencias se presentaron en los denominados picos de la pandemia que se presentaron en julio a diciembre de 2020 y enero, mayo-julio de 2021, posterior a septiembre las dos gráficas lucen muy parecidas con lo cual se desvirtúa la existencia de un efecto del género que haya sido consistente.

Figura 9

Muertos COVID-19 por género Colombia 2020-2021



En cuanto a la forma de la distribución de las edades de los fallecidos discriminando por género, en las Figuras 10 y 11 se muestra un comportamiento similar para ambos géneros, pero con un nivel de apuntamiento significativamente diferente en la zona de alta concentración, además notemos que las edades de mayor mortalidad han estado entre los 50 y 90 años.

Figura 10

Distribución de las edades de los fallecidos COVID-19 mujeres

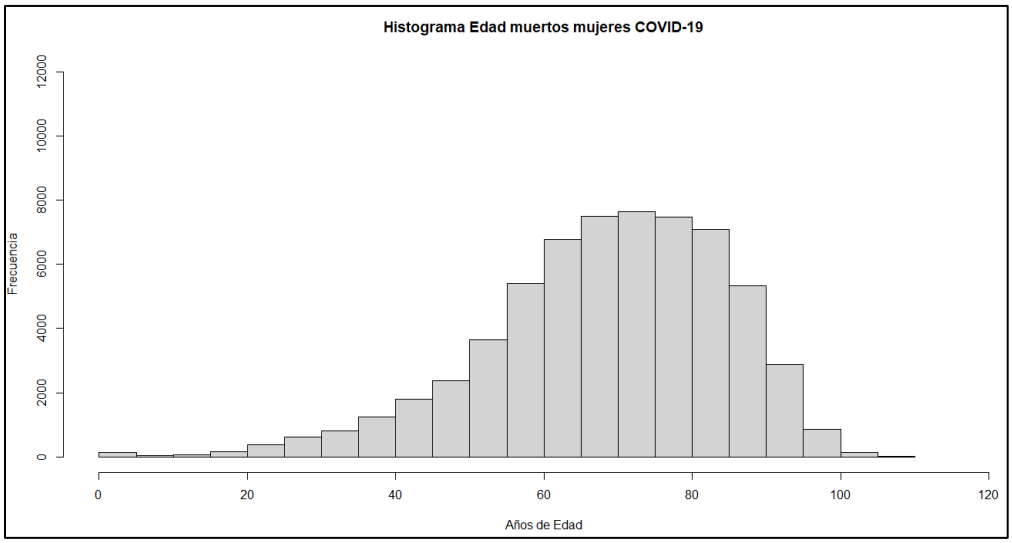
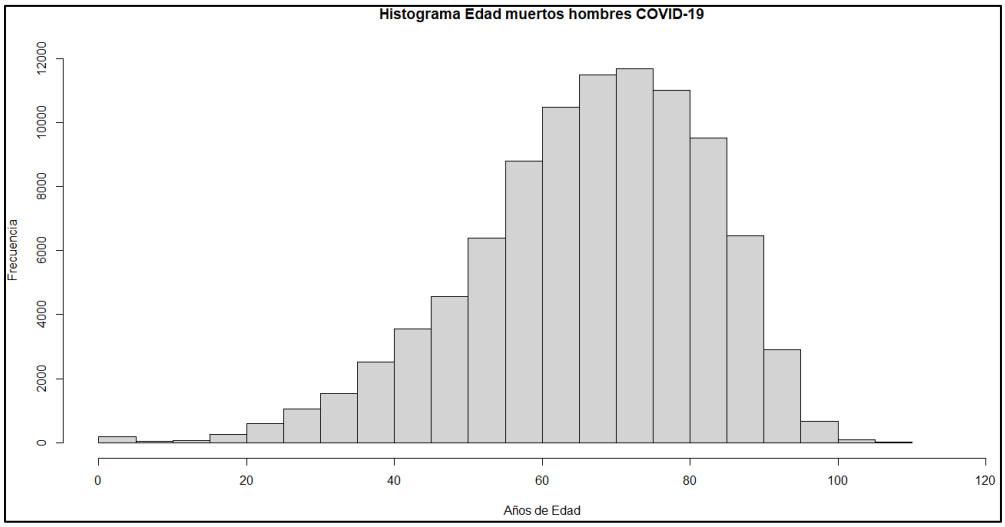


Figura 11

Distribución de las edades de los fallecidos COVID-19 hombres



3.1.2.2 Etnia. La variable sobre pertenencia étnica en otros países ha mostrado tener cierto efecto en el comportamiento de la pandemia especialmente entre la descendencia afro. En

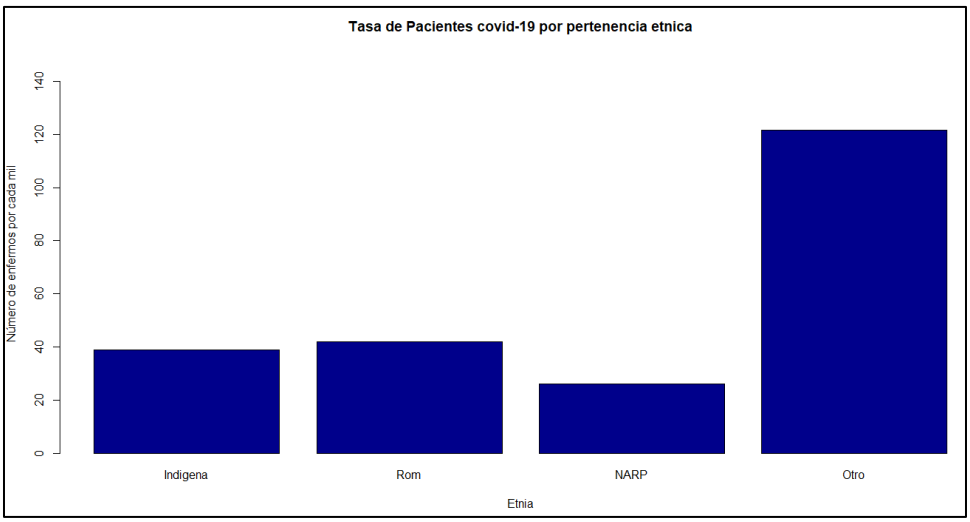
Colombia, país diverso étnicamente hablando, las minorías reportadas por el DANE acorde al censo nacional de 2018 y su participación según la población del país son como se indica a continuación:

- Indígenas (4.4 %)
- Rom (0.006 %)
- Negros, Afrocolombianos, Palenqueros y Raizales (9.34 %) cuyo acrónimo es NARP

En la Figura 12 se observa una comparación del número de contagios discriminando por grupo étnico, el gráfico muestra la tasa de incidencia del virus por cada 1000 personas pertenecientes a cada grupo étnico, como era de esperarse el grupo Otros que incluye a todos quienes no pertenecen a una minoría, exhibe el mayor valor, pero dentro de los grupos étnicos minoritarios la lista la encabezan los Rom seguido de cerca por los indígenas y en un menor nivel el grupo NARP.

Figura 12

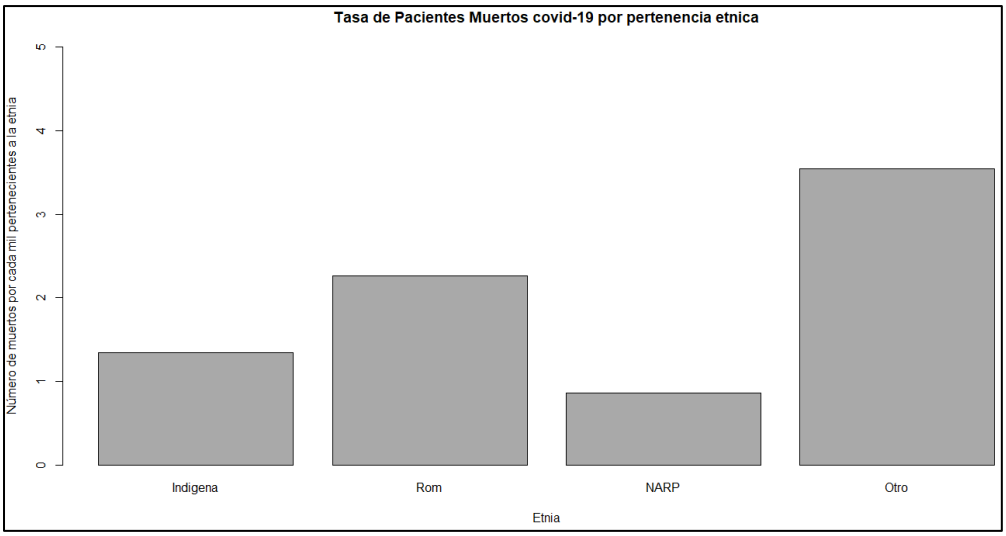
Tasa contagios COVID-19 por grupo étnico



Se tiene además para la comparación de la tasa de mortalidad (Figura 13) que el comportamiento presentado en el párrafo anterior se mantiene, evidenciando que contrario a los hallazgos reportados al inicio de la pandemia en otros países, la comunidad NARP es la que menos se vio afectada en el período 2020-2021 a causa del COVID-19 contrario a lo observado al inicio de la pandemia en otros países como Estados Unidos y Reino Unido; por su parte, la comunidad Rom se perfila como el grupo étnico más afectado en Colombia, siendo preocupante esta situación por su baja presencia dentro de la población en Colombia.

Figura 13

Tasa muertos COVID-19 por grupo étnico



3.1.2.3 Clima. Al inicio de la pandemia se popularizo el supuesto efecto que el clima podría tener en la transmisión del virus, hecho interesante en ese momento porque brindaba una opción para reducir la incidencia del COVID 19, en especial se hablaba de una menor velocidad de reproducción en temperaturas altas y a mayor humedad; nuestro país es un país diverso en cuanto

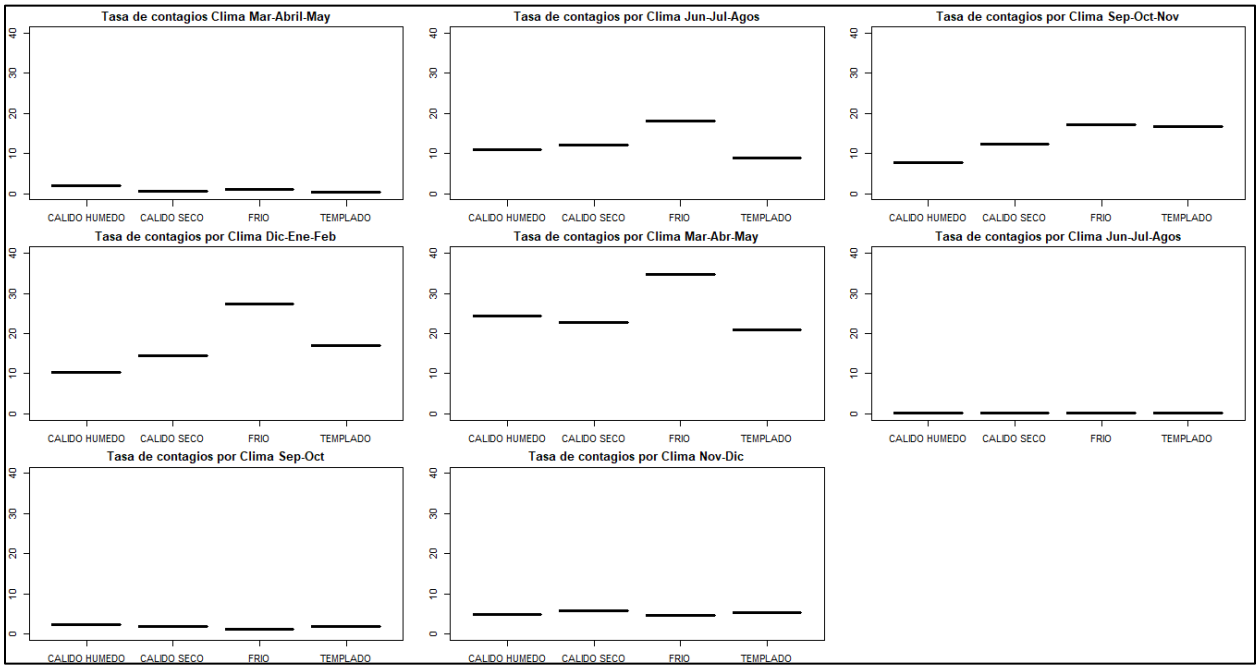
al factor climatológico donde el Ministerio de Vivienda et al. (2020) reporta en Colombia la existencia de 4 tipos de climas:

- Cálido húmedo
- Cálido seco
- Frío
- Templado

La variable clima se registró para cada paciente a partir de la variable municipio reportada en la base de datos, nuevamente para favorecer la comparación a través de este factor se calculan las tasas de contagio por cada 1000 personas que residen en los municipios en cada clima, en la figura a continuación se presenta esta distribución considerando distintos subconjuntos de tiempo definidos al interior del período en estudio, es decir marzo de 2020 hasta diciembre de 2021:

Figura 14

Tasa media de contagios COVID-19 por clima



En el primer, sexto, séptimo y octavo periodos se observa que las tasas son muy bajas y similares a través de los climas en consideración mientras que en los intermedios hay variaciones, en el segundo se presenta una tasa mayor en el clima frío seguido de tasas de contagio levemente mayores la una a la otra para los climas cálido húmedo y cálido seco, en el tercero, el nivel de la tasa de incidencia para frío y templado es mayor respecto a los demás climas, en el periodos cuarto y quinto se nota que las mayores tasas son del clima frío, donde las demás tasas varían de periodo a periodo. Así se puede decir, que en los municipios con un clima frío presentaron mayores tasas de contagio, aunque las diferencias existentes en las tasas realmente no son significativas.

3.2 Análisis de Supervivencia para el Tiempo de Recuperación

Un aspecto importante en el estudio de una enfermedad, infección o en este caso una pandemia, es la recuperación o no de aquellos que la presentan o se ven afectados por ella, una forma de hacerlo es a través del estudio de la variable Tiempo de recuperación de un paciente en nuestro caso contagiado de COVID-19 en Colombia; nosotros haremos uso de un análisis de supervivencia para los datos de pacientes que cuentan con una fecha de inicio de síntomas en la base de datos del INS, el Tiempo de recuperación se calculó para todos los pacientes, los recuperados son aquellos que aparecían registrados como PCR en la variable Tipo de recuperación lo cual indica la presencia de una prueba PCR con resultado negativo hecho que permita dar de alta al paciente.

Dado que los datos no obedecen a una recolección en el marco de un ensayo clínico sino en el contexto de una pandemia se cuenta con un tamaño de muestra considerable, en consecuencia el análisis del Tiempo de recuperación se presentará dividiendo la base de datos en períodos de tiempo que se obtuvieron al agrupar los pacientes positivos para COVID-19 que registran inicio

de síntomas al interior del lapso de tiempo que define cada periodo tal como se indica a continuación, para cada uno de estos se mostrará un análisis de supervivencia más adelante:

2020

- Febrero-marzo-abril (02-04)
- Mayo-junio (05-06)
- Julio-agosto (07-08)
- Septiembre-octubre (09-10)
- Noviembre-diciembre (11-12)

2021

- Enero-febrero (01-02)
- Marzo-abril (03-04)
- Mayo-junio (05-06)
- Julio-agosto (07-08)
- Septiembre-octubre (09-10)
- Noviembre-diciembre (11-12)

En cada análisis de supervivencia los datos censurados serán aquellos que no presentan recuperación dentro de cada período de tiempo, la variable Tiempo se calculó para cada paciente de la siguiente manera:

- Para un paciente que se recupera dentro del periodo de tiempo, el Tiempo de recuperación será el número de días desde el inicio de síntomas hasta su recuperación.
- Para un paciente muerto en el periodo estudiado, el Tiempo será el tiempo desde el inicio de síntomas hasta este desenlace.
- Para un paciente que no presente ni recuperación ni muerte en el periodo de tiempo, el Tiempo será los días transcurridos desde el inicio de síntomas hasta el último día del estudio.

A continuación, en la Tabla 4 se presenta la distribución de los pacientes por las diferentes condiciones discriminando por periodos de tiempo:

Tabla 4

Censura, muerte, recuperados 2020-2021

			2020			
Periodo	**02-04**	**05-06**	**07-08**	**09-10**	**11-12**	
Recuperados	2.116	25.965	52.362	29.412	32.255	
Muertos	376	4009	14838	7843	9110	
Censurados	7.879	111.265	454.900	408.239	549.038	
Total casos	9.995	137.230	507.262	437.651	581.293	
			2021			
Periodo	**01-02**	**03-04**	**05-06**	**07-08**	**09-10**	**11-12**
Recuperados	68.968	72.725	102.554	33.987	4652	4.146
Muertos	10.669	13.186	24.620	7.089	1.200	2.043
Censurados	381.043	621.162	1.787.468	332.391	79.392	326.798
Total casos	450.011	693.887	1.290.022	366.378	84.044	330.944

Nota. La tabla presenta un resumen de los recuperados, fallecidos y censurados en cada periodo de tiempo.

3.2.1 Modelos no paramétricos

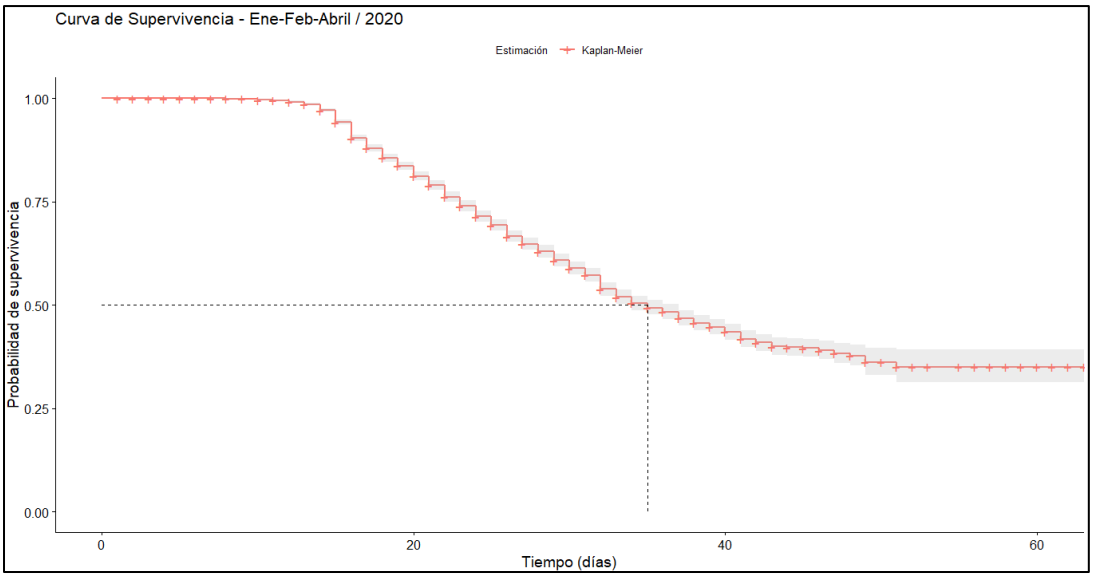
En primera instancia, se ajusta a los datos de cada período de tiempo el modelo no paramétrico de Kaplan-Maier cuyo principal resultado es la curva de supervivencia que en nuestro caso permite identificar la probabilidad de recuperarse después de *t* días; esta curva de supervivencia describe un comportamiento característico, inicia ubicándose en una probabilidad de 1 y ésta va decreciendo a través del tiempo indicando la disminución en la probabilidad de seguir enfermo conforme el tiempo aumenta.

3.2.1.1 Modelos no paramétricos 2020. A continuación, se presenta el ajuste de Kaplan Meier para el primer trimestre de la pandemia, después se hará lo propio con los demás períodos en consideración:

Al inicio de la pandemia, febrero-abril de 2020, se observa en la figura 15 que la probabilidad de recuperarse en un mayor tiempo comienza a decrecer a los 15 días de la enfermedad, el tiempo mediano de supervivencia es de 35 días y cerca del 75% de los pacientes siguen enfermos a los 23 días de haber iniciado la infección por el virus, es de notar que el tercer cuartil no es calculado, esto se debe a que como se observa en la curva de supervivencia todos los pacientes tienen una probabilidad de supervivencia mayor a 0,25 también se observan en la gráfica los intervalos de confianza al 95% más anchos en la cola de la curva.

Figura 15

Ajuste Kaplan-Meier enero-febrero 2021



A continuación, se muestra una parte de la salida producida por la función *survfit()* de R, como se ve el programa aporta información detallada del proceso de estimación en cada punto que se presenta el evento de interés, de esta resaltaríamos la precisión con que se está realizando la estimación a través de Intervalos de confianza, razón por la cual en la Figura 16 no es visible la banda de confianza alrededor de la curva en rojo.

Figura 16

Salida R survfit( )

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
7	7644	2	1.000	0.000185	0.999	1.000
8	7336	5	0.999	0.000356	0.998	1.000
9	7036	3	0.999	0.000433	0.998	0.999
10	6784	14	0.997	0.000699	0.995	0.998
11	6521	12	0.995	0.000876	0.993	0.996
12	6137	18	0.992	0.001111	0.990	0.994
13	5882	37	0.986	0.001505	0.983	0.989
14	5580	81	0.971	0.002165	0.967	0.976

Nota. La imagen muestra la salida de R al aplicar la función *surfit* para la estimación por Kaplan Meier en el primer periodo de tiempo, mostrando para cada tiempo el numero de pacientes en riesgo, número de pacientes que presentaron recuperación así mismo los intervalos de confianza a un nivel de significancia del 0.05.

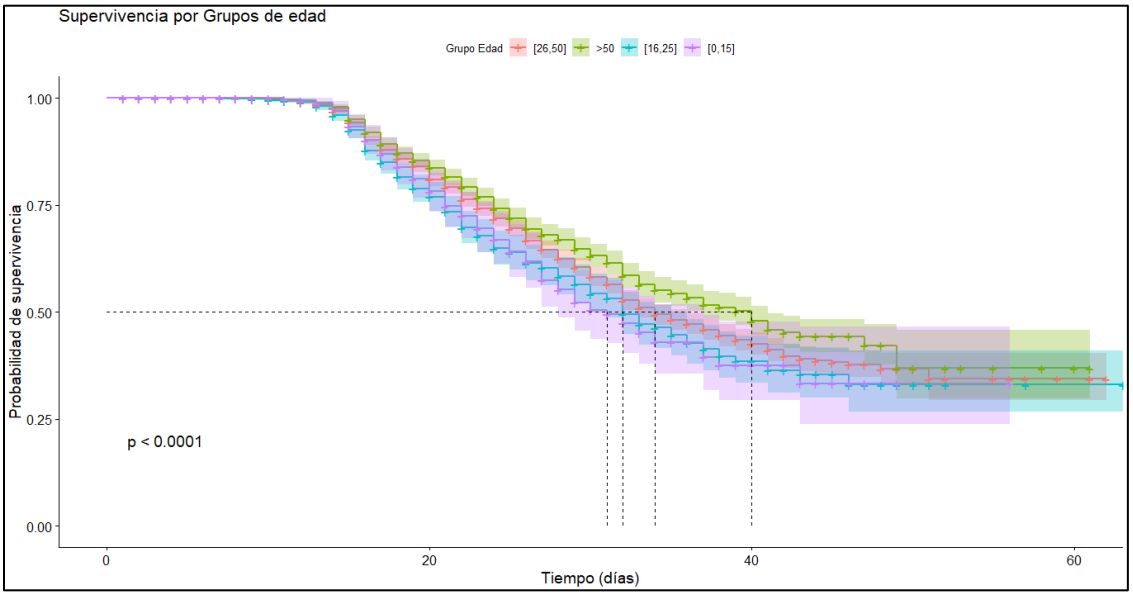
En este periodo en específico se realiza un ajuste del modelo implicando las covariables Grupo edad y Etnia, Sexo y Grupo edad para analizar el comportamiento de la curva de supervivencia estimada por Kaplan Meier en diferentes grupos poblacionales, durante este comienzo de la pandemia donde se presentaban muchos supuestos de comportamiento.

Observando las curvas de supervivencia por grupos de edades mostrada en la Figura 17 se pueden apreciar en distintos colores la curva de supervivencia asociada para cada grupo, en primera instancia por medio de una prueba Log-Rank se concluye que las curvas son significativamente diferentes (Valor P valor << 0.0001). Ahora, se aprecia que las curva que se encuentran por debajo de las demás son las correspondientes a las edades de quienes tienen 15 años o menos y quienes están entre los 16 y los 25, evidenciando una disminución de la probabilidad de seguir enfermo en *t* días más rápido que las demás, podría decirse que tienen mayores probabilidades de recuperarse en menor tiempo dentro del periodo estudiado. Así, mismo

la curva que presenta una velocidad de decrecimiento menor es la curva de los que tienen más de 50 años y solo denota que para este grupo la probabilidad de seguir enfermos a través del tiempo es superior a la de los demás.

Figura 17

Curvas de supervivencia por grupos de edad



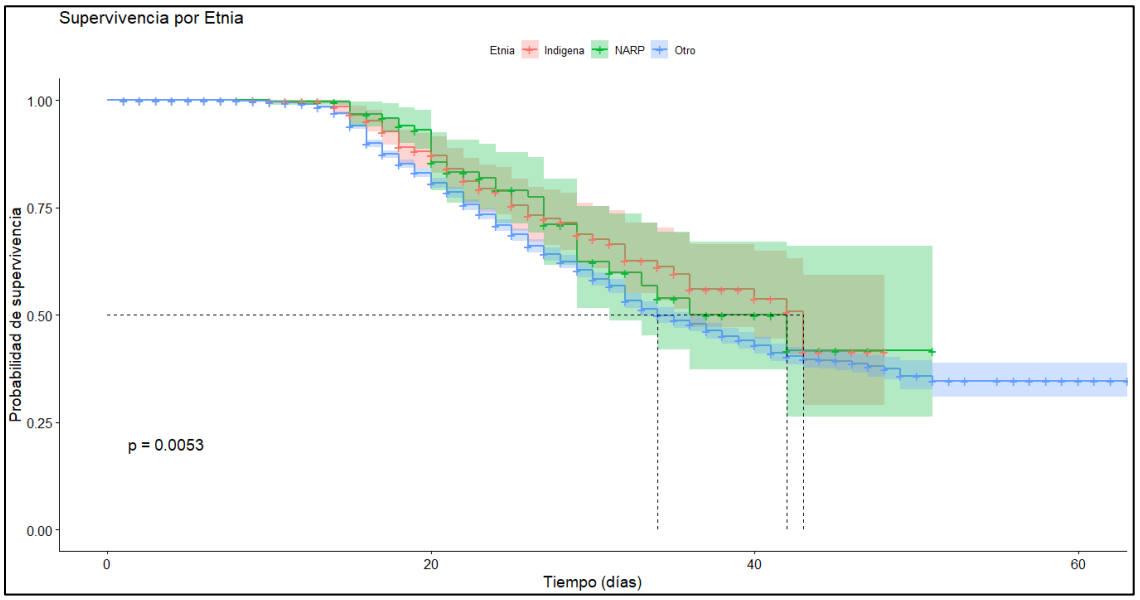
Nota. Curvas de supervivencia por grupos de edad donde el color purpura es [0,15], color azul [16-25], color rosa es [26-50] y color verde más de 50 años.

Por otro lado, en la Figura 18 se aprecian las curvas de supervivencia para los tres grupos de pertenencia étnica presentes del periodo en estudio dado que en este periodo no hay pacientes pertenecientes a la comunidad ROM, con un valor p de 0.0053 se tiene que las curvas entre los grupos presentan diferencias significativas, donde la que presenta una mayor diferencia con las demás es la del grupo de quienes no tienen una pertenencia étnica (Otros), quienes presentan una

menor probabilidad de recuperarse en los 62 días en lo que abarca el periodo por otro lado se puede apreciar cierta semejanza entre las curvas de los grupos indígena y NARP, además que sus intervalos de confianza se superponen en gran parte de la curva, pudiendo suponer igualdad entre estas dos curvas .

Figura 18

Curvas de supervivencia por grupos étnicos

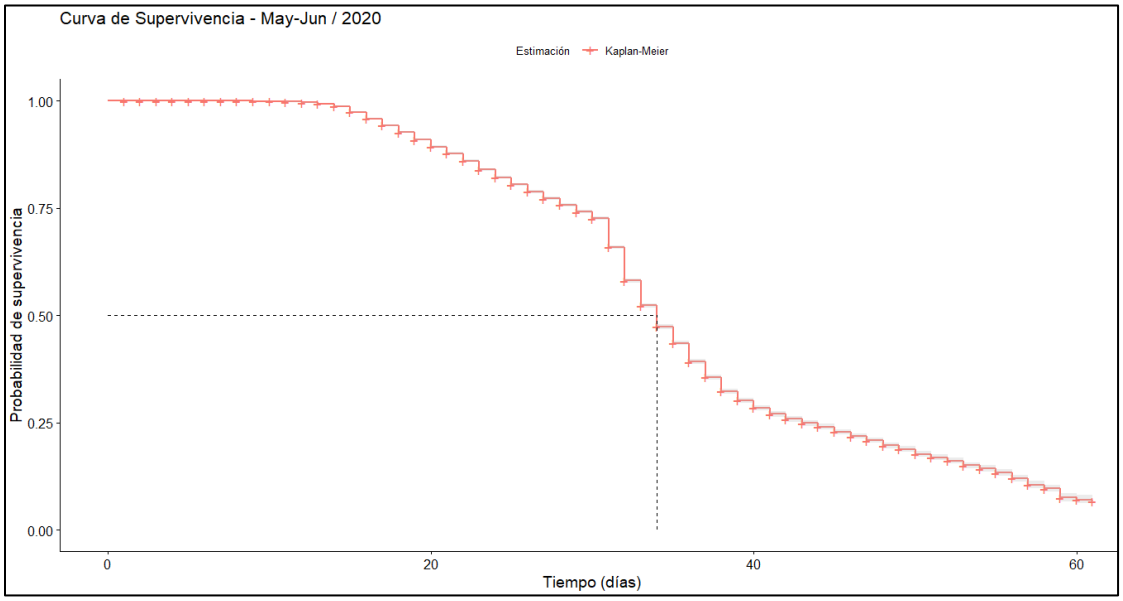


Nota. Curvas de supervivencia por grupos étnicos donde el color azul es Otro, color rosa es Indigena y color verde NARP.

Otro ajuste interesante, es el realizado al periodo mayo-junio de 2020, se observa en primera instancia diferencias en el comportamiento con el primer periodo. En este periodo, se tiene un tiempo mediano de supervivencia de 34 días, con el 75% de los pacientes enfermos durante al menos 29 días, y el 25% por lo menos 43 días. Esta curva, contrario a la de primer periodo no se estabiliza antes del 0,25 sino que la tendencia es a decrecer hacia cero a medida que pasa el tiempo.

Figura 19

Ajuste Kaplan-Meier mayo-junio 2020

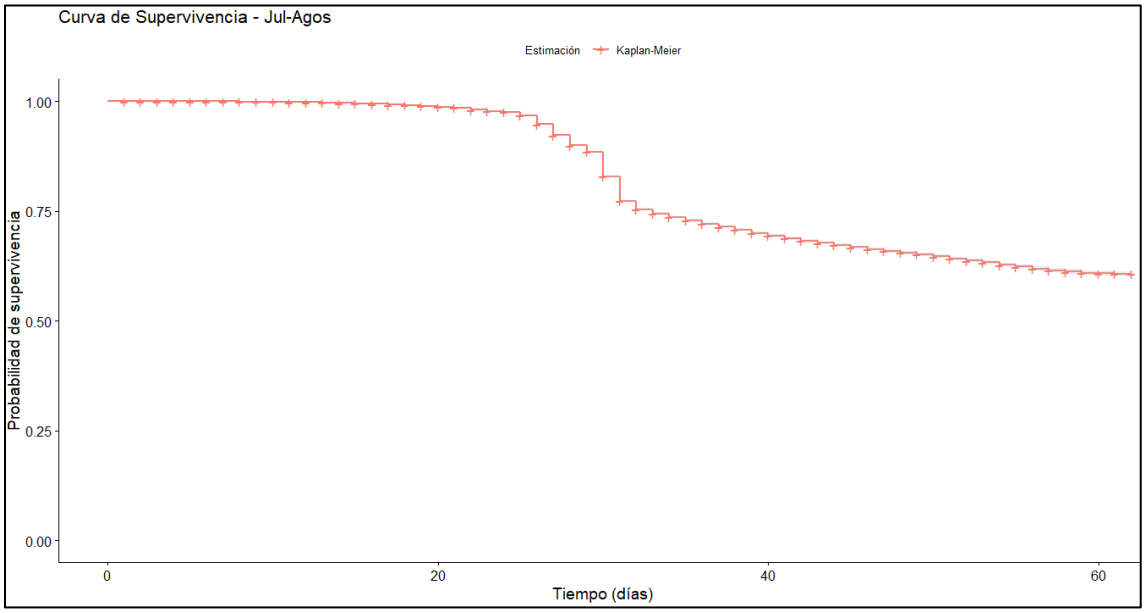


A partir, del periodo julio-agosto de 2020 el comportamiento de las curvas de supervivencia es similar, notando que estas presentan una probabilidad de supervivencia mayor a 0.5 para todo tiempo *t* indicando para cada período de tiempo una alta probabilidad de seguir enfermo después de los 62 días (2 meses aproximadamente). El periodo julio-agosto muestra un decrecimiento lento en la probabilidad de supervivencia que es menor al de los periodos septiembre-octubre y noviembre-diciembre en los cuales se presenta una velocidad de decrecimiento similar, además se diferencian en el cuartil 1, en particular estos son 33,46 y 41 respectivamente.

Para las gráficas de los periodos de tiempo siguientes que se presentan en los Apéndices A y B, no se detalla su análisis porque no describen cambios importantes al que aquí se presenta.

Figura 20

Ajuste Kaplan-Meier julio-agosto 2020



3.2.1.2 Modelos no paramétricos 2021. En 2021 se observa que el comportamiento de las curvas de supervivencia en cada periodo es muy similar, en primera instancia se observa que todas las curvas estabilizan su probabilidad antes de 0.5, e incluso particularmente en el periodo noviembre-diciembre del 2021 antes del 0,75, esto indica que la probabilidad de recuperarse en un tiempo menor o igual a 62 días (dos meses) desde el inicio de síntomas de COVID-19 sigue siendo muy baja. Las diferencias fundamentales se encuentran en la velocidad en que decrece la curva y la ubicación del primer cuartil (ver Tabla 5). En 2021 el periodo con menor probabilidad de recuperación fue el de noviembre-diciembre, la Figura 23 muestra esta curva que luce prácticamente constante a partir de 30 días en un nivel de probabilidad de 0.8.

Figura 21

Ajuste Kaplan-Meier enero-febrero 2021

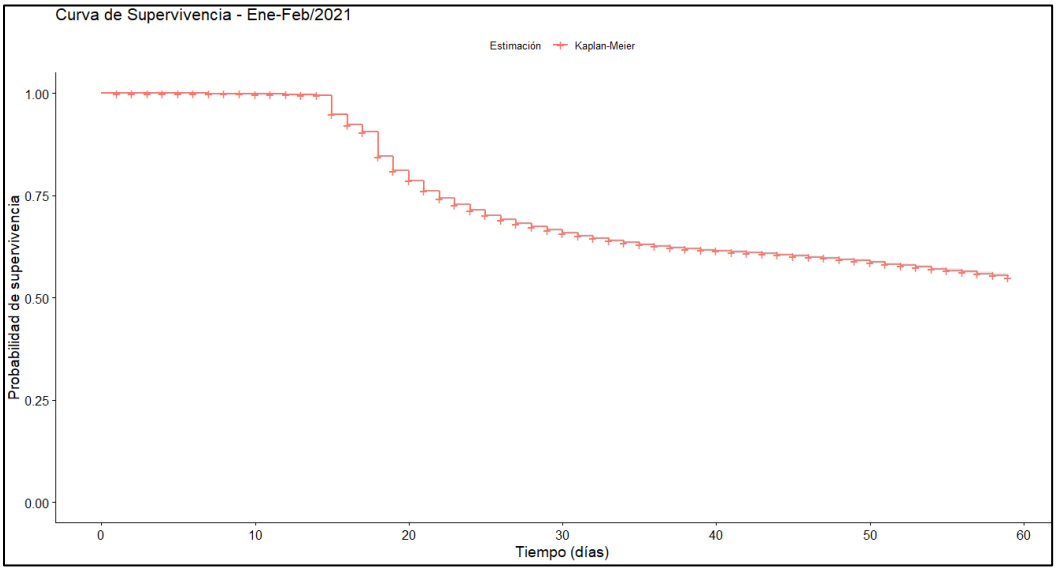
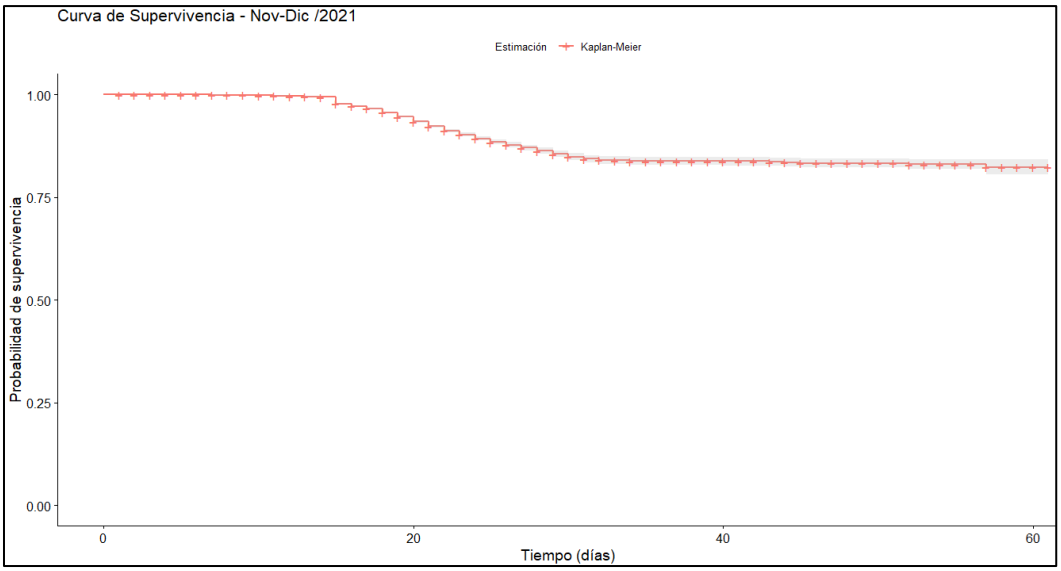


Figura 22

Ajuste Kaplan-Meier noviembre-diciembre 2021



A continuación, en la Tabla 5 se presenta un resumen con los valores que vienen a complementar el análisis gráfico, en este caso no debe desconocerse que el porcentaje de censura es alto, este hecho se explica por el costo que implica una prueba PCR en Colombia y la crisis en materia de salud pública que fue escalando al pasar el tiempo llegando al punto de limitar el acceso a insumos de atención básica en salud, cuidado y prevención:

Tabla 5

Descriptiva ajuste Kaplan-Meier 2020-2021

Año	Periodo	Datos	Eventos (Recuperado según PCR)	% Censura	1er cuartil	Mediana	3er cuartil
2020	Febrero- abril	9.995	2116	78%	23	35	NA
	Mayo-junio	137.230	25.965	81%	29	34	43
	Julio-agosto	507.262	52.362	89%	33	NA	NA
	Septiembre – octubre	437.651	29.412	93%	46	NA	NA
	Noviembre-diciembre	581.293	32.255	94%	41	NA	NA
2021	Enero-febrero	450.011	68.968	84%	22	NA	NA
	Marzo-abril	693.887	72.725	90%	23	NA	NA
	Mayo-junio	1.290.022	102.554	92%	28	NA	NA
	Julio-agosto	366.378	33.987	91%	32	NA	NA
	Septiembre-octubre	84.044	4652	94%	45	NA	NA
	Noviembre-diciembre	330.944	4146	98%	NA	NA	NA

Nota. Esta tabla incluye la información básica de forma de los modelos ajustados Kaplan-Meier a los periodos de tiempo de 2020-2021.

Las gráficas de los ajustes de los periodos no mostrados en esta parte se encuentran en los Apéndices C al F.

3.2.2 Modelos paramétricos

Además de realizar un ajuste no paramétrico por Kaplan Meier, se realiza un ajuste paramétrico, con ayuda de la función *flexsurv()* de R, y se explora el ajuste de cinco distribuciones de uso común para un análisis de supervivencia paramétrico como son:

- Gamma generalizada
- Log-logística
- Log-normal
- Weibull
- Exponencial

Para cada periodo tanto de 2020 y 2021 se realiza el ajuste de las anteriores distribuciones, y a partir del criterio de información de Aiken (AIC por sus siglas en ingles) y el criterio de información bayesiano (BIC por sus siglas en inglés) se escoge el modelo con la distribución de mejor ajuste como se puede observar en el Apéndice G.

3.2.2.1 Modelos paramétricos 2020. De los modelos ajustados de cada periodo, la distribución que dio el mejor modelo en el año 2020 fue la distribución gamma generalizada, donde los parámetros de la distribución μ de localización, σ y Q de forma son cercanos en cada periodo, presentando una discrepancia a partir de alrededor de los 15-21 días con el ajuste no paramétrico de los datos.

En los otros modelos, se puede observar para los periodos febrero-abril y mayo-junio, un ajuste paramétrico que coincide en gran medida con el no paramétrico, excepto quizás en las colas. Mientras en los periodos julio-agosto, septiembre-octubre y noviembre-diciembre en el ajuste se evidencia una mayor discrepancia a partir de los 20 días y sobre todo en las colas de las curvas desde los 35 y 40 días respectivamente (ver Figuras 23-25).

Figura 23

Comparación Kaplan Meier y modelo paramétrico febrero-abril 2020

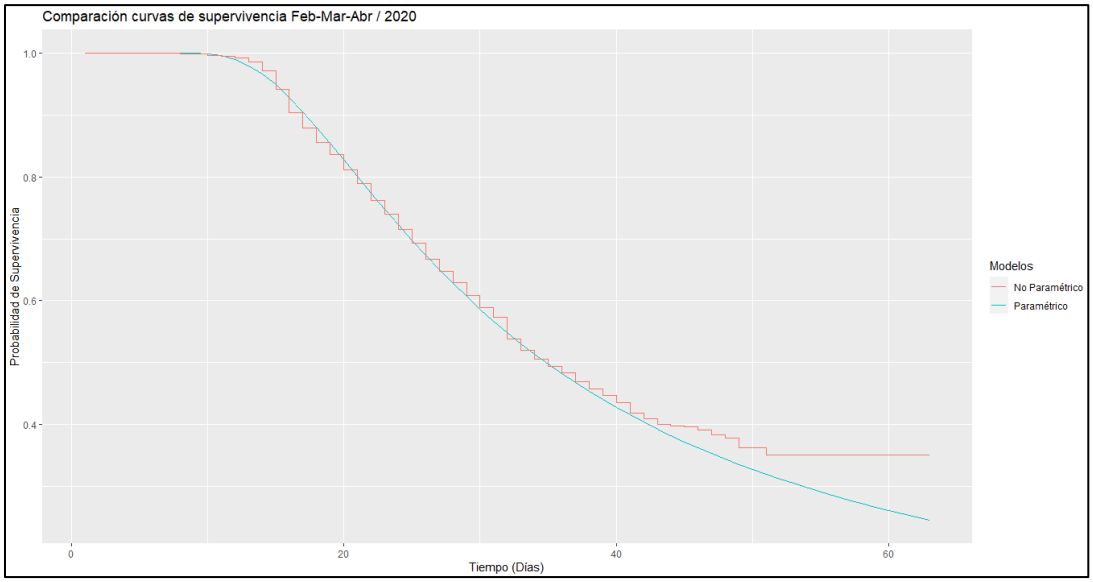


Figura 24

Comparación Kaplan Meier y modelo paramétrico mayo-junio 2020

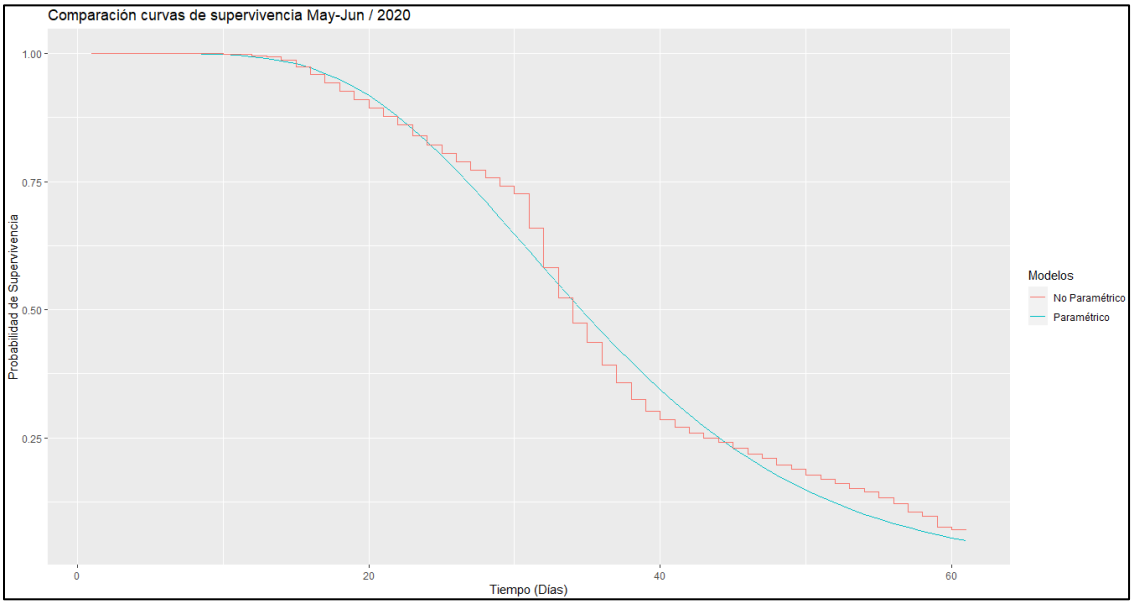
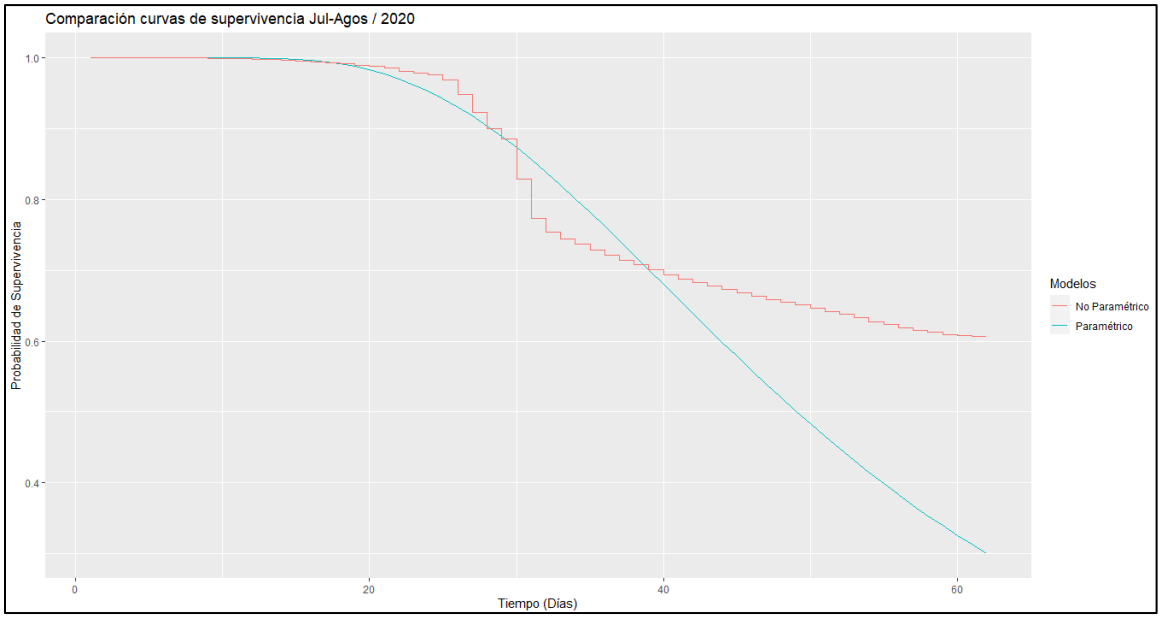


Figura 25

Comparación Kaplan Meier y modelo paramétrico julio-agosto 2020

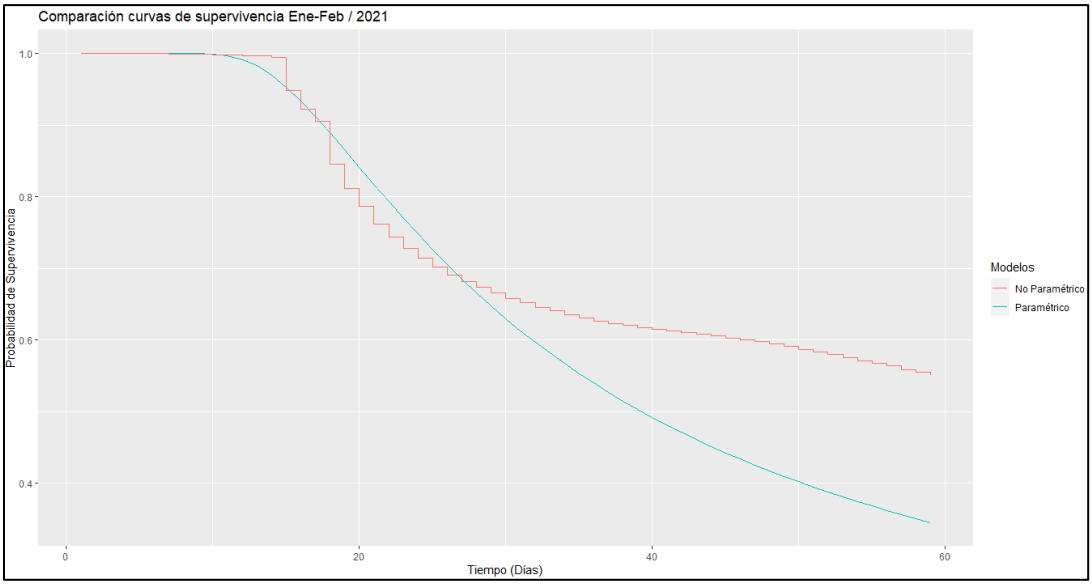


3.2.2.2 Modelos paramétricos 2021. En el 2021 se observa primero que contrario al 2020 no todos los periodos presentaron el mejor ajuste por la distribución gamma generalizada, se tiene que para el periodo mayo-junio el menor AIC lo tiene el modelo que ajusta la distribución log-normal, con parámetros mostrados en la Tabla 6. Los demás periodos, presentaron el mejor ajuste con la distribución la gamma generalizada, al igual que en el año 2020 los parámetros ajustados de la curva presentan mucha similitud presentando las mayores diferencias entre el ajuste paramétrico y el no paramétrico en lo que se constituye como las colas, a partir de los 20-22 días aproximadamente.

Mostrando un mejor ajuste en la que podemos catalogar como la primera parte de la enfermedad.

Figura 26

Comparación Kaplan Meier y modelo paramétrico febrero-abril 2020



En la Tabla 6 se encuentran relacionados los distintos periodos de 2020 y 2021 con el modelo de la distribución mejor ajustada y las respectivas estimaciones de parámetros y valores relevantes como los cuartiles.

Tabla 6

Ajustes mejor modelo ajustado periodos de tiempo 2020-2021

Año	Periodo	Distribución ajustada	Nº parámetro	Parámetros	1er Cuartil	Mediana	3er Cuartil
2020	**02-04**	Gamma generalizada	3	**μ:** 3.2593 **σ:** 0.5666 **Q:** -1.3276	16,25	31,5	46,75
	05-06	Gamma generalizada	3	**μ:** 3.5762 **σ:** 0.3656 **Q:** 0.2706	16	31	46
	07-08	Gamma generalizada	3	**μ:** 3.8775 **σ:** 0.4404 **Q:** -0.1110	46,25	31,5	46,75
	09-10	Gamma generalizada	3	**μ:** 3.6439 **σ:** 0.7596 **Q:** -1.5879	16	31	46

Año	Periodo	Distribución ajustada	Nº parámetro	Parámetros	1er Cuartil	Mediana	3er Cuartil
2020	**11-12**	Gamma generalizada	3	**μ:** 3.5663 **σ:** 0.7787 **Q:** -1.5624	16	31	46
2021	**01-02**	Gamma generalizada	3	**μ:** 3.2298 **σ:** 0.5994 **Q:** -1.7576	15,5	30	44,5
	03-04	Gamma generalizada	3	**μ:** 3.20942 **σ:** 0.54431 **Q:** -1.5079	16	31	46
	05-06	Log Normal	2	**Meanlog:** 3.6069 **Sdlog:** 0.4876	16	31	46
	07-08	Gamma generalizada	3	**μ:** 3.3975 **σ:** 0.6771 **Q:** -1.4399	16,25	31,5	46,75
	09-10	Gamma generalizada	3	**μ:** 3.5413 **σ:** 0.7308 **Q:** -1.0940	16	31	46
	11-12	Gamma generalizada	3	**μ:** 3.7972 **σ:** 0.6701 **Q:** -0.5510	16	31	46

3.2.3 Modelo de falla acelerado

Un modelo de tiempo de falla acelerado que identifique si alguna de las variables que se encuentran en la base de datos presentan algún tipo de influencia en términos de aceleración o desaceleración de la recuperación del paciente presenta su mayor importancia dentro del contexto del inicio de la pandemia donde lo que se tenía eran suposiciones e hipótesis, por esta razón vamos a ajustar un modelo de falla acelerado gamma generalizado (mejor distribución ajustada) al periodo que corresponde a los meses de febrero a marzo de 2020.

Las variables que se buscan analizar son: el Sexo, la Etnia, el Clima y Grupos de edad, para ello se ajustan 4 modelos que involucran estas, la Tabla 7 muestra las estimaciones y ajuste del modelo.

Tabla 7

Modelos de falla acelerado (AFT) gamma generalizado periodo 02-04 2020

	ATF	Parámetro	Variables ajustadas		Estimación (est)	exp(est)	II 95%	ID 95%
1	Gamma generalizada AIC=19512.4	μ: 3.3824 σ: 0.5661 Q: -1.3323	**Sexo**	**M**	0.0001	1.0001	0.9702	1.0308
			Etnia	**3**	-0.0198	0.9804	0.8701	1.1046
				4	-0.1325	0.8759	0.7935	0.9668
2	Gamma generalizada AIC=19513.8	μ: 3.2589 σ: 0.5653 Q: -1.3133	**Sexo**	**M**	-0.0022	0.9978	0.9681	1.0284
			G. Edad	**1**	-0.0158	0.9843	0.9211	1.0519
				2	-0.0574	0.9443	0.9011	0.9894
				4	0.0414	1.0423	1.0069	1.0519
3	Gamma generalizada AIC=19500.4	μ: 3.3841 σ: 0.5649 Q: -1.3138	**Etnia**	**3**	-0.0246	0.9757	0.8658	0.9915
				4	-0.1352	0.8735	0.7912	0.9645
			G. Edad	**1**	-0.0161	0.9840	0.9207	1.0518
				2	-0.0554	0.9462	0.9029	0.9915
				4	0.0427	1.0437	1.0082	1.0803
4	Gamma generalizada AIC=19435.5	μ: 3.2773 σ: 0.5585 Q: -1.2798	**G. Edad**	**1**	-0.0111	0.0339	0.9254	1.0565
				2	-0.0571	0.9445	0.9017	0.9893
				4	0.0433	1.0443	1.0091	1.0807
			Clima	**1**	0.0003	1.0003	0.9571	1.0455
				2	-0.1585	0.8534	0.8083	0.9011
				3	0.0610	1.0632	1.0092	1.0455
5	Gamma generalizada AIC=19427.5	μ: 3.3919 σ: 0.5579 Q: -1.2822	**Sexo**	**M**	-0.0052	0.9948	0.9653	1.0252
			Etnia	**3**	-0.0260	0.9743	0.8652	1.0973
				4	-0.1294	0.8786	0.7962	0.9695
			Clima	**1**	0.0105	1.0105	0.9663	0.9110
				2	-0.1480	0.8624	0.8164	1.0568
				3	0.0704	1.0729	1.0180	1.1308
			G. Edad	**1**	-0.0553	0.9879	0.9245	1.0558
				2	-0.0121	0.9462	0.9033	0.9912
				4	0.0444	1.0454	1.0101	1.0820

Nota. La tabla muestra los modelos AFT gamma generalizados (la mejor distribución ajustada) en el periodo febrero – abril de 2020, relacionando en cada modelo un par de variables categóricas Sexo, Etnia, Clima o Grupo de Edad (Gedad), y el último las 4 juntas.

Al ajustar los modelos de tiempo de falla acelerado gamma generalizada en el primer periodo de tiempo de la pandemia (febrero-marzo 2020), es de notar en primera instancia que para cada modelo ajustado las covariables involucradas son significativas a un nivel de significancia del 5% ya que los intervalos de confianza no contienen el cero (Ver Tabla 7). Para los modelos que involucran la variable Etnia en el periodo en estudio no existen pacientes pertenecientes a la etnia ROM, por lo cual la variable Etnia asume sólo 3 categorías.

El primer modelo tiene como covariables Etnia y Sexo, el modelo ajustado según la salida de R sería:

$$\boldsymbol{Log(T)= 3.38+0.0001\ Masculino\ -0.0198\ NARP\ -\ 0.1325\ Otros}$$

La interpretación de este modelo estimado implica dar sentido a los coeficientes en términos del efecto que describen, así por ejemplo un signo positivo indica aceleración y negativo desaceleración, no obstante aquí se debe tener en cuenta que estas estimaciones se les debe aplicar la exponencial para tenerlas en la escala adecuada y poder hacer la interpretación del modelo en términos de la variable T, con esto desaparece el signo y la discriminación anterior se hace considerando que los exponentes negativos producen valores entre cero y uno en la función exponencial. En ese orden de ideas, para este primer modelo se tiene que en los pacientes COVID-19 hombres se acelera la recuperación 1.0001 veces

respecto a las mujeres, más no se presenta un efecto en el Tiempo de recuperación significativamente más corto o largo que las pacientes de sexo femenino (Efecto estimado = 0.0001; $e^{0.0001}$= 1.0001), además el pertenecer a la comunidad NARP o el no pertenecer a una minoría étnica disminuye el tiempo de recuperación pero muy levemente, efectos estimados de 0.98 y 0.88 respectivamente, en relación a aquellos que pertenecen a la etnia indígena. Para los modelos presentados en la Tabla 7 y según los Intervalos de confianza (ninguno contiene al cero) se tiene que las covariables incluidas en estos resultan ser significativas pero la magnitud de los efectos que describen no permite identificar una variable que realmente pueda acelerar o desacelerar la presentación del evento, en este caso la recuperación.

Como complemento técnico para las interpretaciones que tienen lugar al ajustar un modelo AFT, el valor obtenido al aplicar la exponencial, ejemplo si para una variable que discrimina entre dos grupos A ó B y para B se tiene que $e^{0.7}$= 2, a este valor se le denomina Razón de tiempo (TR) la cual se interpreta como que la probabilidad de presentar el evento se da dos veces más rápido para el grupo B que para el grupo A, si fijamos los otros factores, la magnitud encontrada aquí indica que el grupo B tiene su tiempo acelerado con lo cual si el evento de interés es la muerte, el tiempo de sobrevida para el grupo B es menor.

3.2.4 Modelo bayesiano

En esta sección busca ilustrar el modelamiento para el tiempo de recuperación a través de un análisis de supervivencia bayesiano, para ello se analizará solo el periodo correspondiente a los primeros meses de la pandemia en Colombia, febrero, marzo y abril de 2020 dado que era el punto donde más limitaciones técnicas y desconocimiento se tenía frente a la evolución de la condición médica de los pacientes. A continuación, se describe el proceso

de ajuste de una curva de supervivencia a partir de los primeros datos que se recolectaron en Colombia, el periodo correspondiente a febrero-abril 2020.

Dado que no se tiene información especializada sobre el comportamiento de los parámetros se hará uso de unas a priori no informativas, y las especificaciones fueron tomadas a partir de las sugerencias dadas por los paquetes estadísticos usados.

El primer paso es indicar al software una distribución a priori que permita iniciar el modelo para esto se utilizan en la primera parte dos a prioris contenidas en el paquete "BayesSurvival" de R y a partir de estas se estima la curva de supervivencia de los pacientes de COVID-19 para el tiempo de recuperación. Las dos a prioris ajustadas son: la a priori gamma dependiente y la a priori gamma independiente.

Al ajustar en esta ocasión se utilizan los siguientes valores predeterminados y recomendados por el paquete estadístico de R necesarios en las a priori gamma dependiente e independiente:

$$K = \left[\frac{n}{\log(n)}^{\frac{1}{2}}\right], \alpha_0 = 1.5, \alpha = 1$$

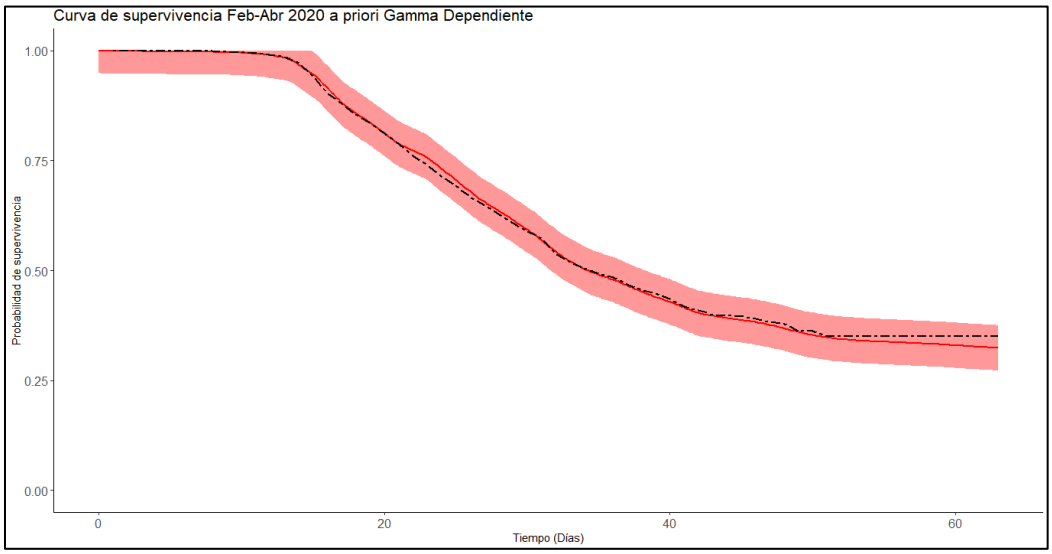
Donde n es el tamaño de la muestra. Así, la curva de supervivencia resultante para la a priori gamma dependiente se encuentra en la Figura 27 donde la curva de color rojo es la dada por la a priori y la curva negra la estimación por Kaplan-Meier, así se nota una gran similitud entre las curvas, con muy leves diferencias, además sombreado en rojo se muestran los intervalos de credibilidad del 95%.

Mientras, en el ajuste con a priori gamma independiente mostrada en la Figura 28 en los primeros 40 días presentan gran similitud, sólo se encuentra una gran diferencia en el

ajuste en la parte final al superar los 43 días, allí Kaplan Meier se estabiliza alrededor de la probabilidad 0.30 y la estimación del posteriori sigue decreciendo.

Figura 27

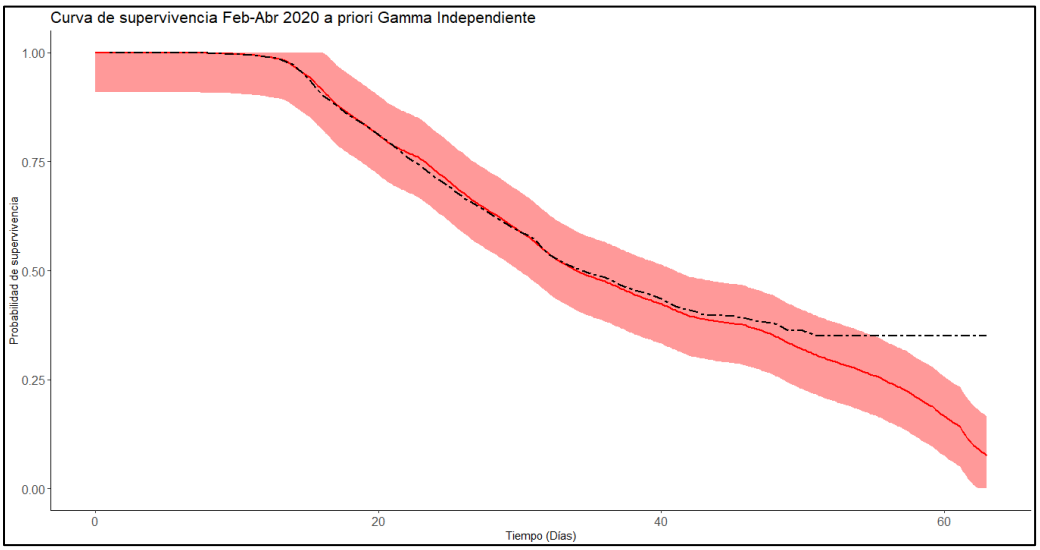
Ajuste bayesiano a priori gamma dependiente



Nota. La gráfica presenta el ajuste bayesiano con a priori gamma dependiente en color rojo, la estimación por Kaplan Meier en una curva negra punteada.

Figura 28

Ajuste bayesiano a priori gamma independiente



Nota. La gráfica presenta el ajuste bayesiano con a priori gamma independiente en color rojo, la estimación por Kaplan Meier en una curva negra punteada y los intervalos de credibilidad sombreado en rojo.

Ahora, en un enfoque bayesiano se trabaja un modelo de tiempo de falla acelerado (AFT por sus siglas en inglés) con una mezcla gaussiana clásica (CGM) como distribución de errores, la cual en R se trabaja a partir del paquete estadístico “bayesSurv”.

En un modelo AFT como el nombrado donde la distribución de los errores corresponde a una mezcla gaussiana clásica es llamado un CGM AFT. Se analizará uno de los modelos ajustados en el AFT clásico que incorpora las variables sexo y etnia.

Para este modelado se utilizan distribuciones y datos a priori no informativos, donde las a priores para el número de componentes de la mezcla K y las medias μ de la mezcla se especifican como sigue:

- La distribución del parámetro K esta dado por una distribución uniforme y con un máximo de 20.

- El hiperparámetro a priori de la distribución de Dirichlet es 1.
- La media a priori será tomada como la media obtenida en el modelo AFT clásico y tomaremos una "alta" varianza de esta de 10^2.
- La a priori beta que define los bloques de parámetros beta que se van a actualizar juntos tiene una media a priori de 0 y una varianza alta de 10^8, se especifica el número de parámetros de regresión, en este caso 4 parámetros.

Además, para comenzar la simulación por MCMC que por defecto en R es por el método de muestreo de Gibbs, se deben dar los valores iniciales para cada parámetro asociado al modelo, para ello se usa en esta ocasión las estimaciones encontradas para los parámetros en el modelado AFT clásico, así:

- El valor inicial para K será 1.
- El valor inicial del hiper-parámetro de la distribución de Dirichlet es 1, usada en el llamado proceso de Dirichlet para la estimación de funciones de supervivencia.
- La media y la varianza inicial será la media estimada y el cuadrado de σ estimado en el modelo 1 de la Tabla 7.
- Los valores iniciales de los betas (β) que corresponden a los coeficientes de regresión serán los coeficientes estimados en la Tabla 7 en el modelo 1.

Así, en este enfoque para el modelo que relaciona las variables de Sexo y Etnia se tienen las a priori:

$$K \sim \mathcal{U}\{1, \dots, 20\}$$

$$w|K \sim Dir_k(1, \dots, 1), \quad \mu|K \sim \prod_{k=1}^{K} N(3.38, 10^2)$$

$$\beta \sim N_3\big(0, diag(10^2, 10^2, 10^2, 10^2)\big).$$

Y el ajuste del modelo nos resulta en:

Tabla 8

CGM AFT Sexo y Etnia en febrero-abril 2020

Variables	Mu	SD	Exp(mu)	Exp(SD)	I95%	D95%
Sexo M	0.0026	0.0125	1.0026	0.01264	0.9789	1.0294
Etnia Rom	0	0	1	1	1	1
Etnia NARP	-0.1034	0.0557	0.9032	0.05166	0.8228	1.0329
Etnia Otros	-0.1927	0.04023	0.8254	0.03383	0.7706	0.9086
K	6	3.626			1	14
Intercepto	3.9098	0.1201			3.624	4.0912
Escala	0.7649	0.1467			0.5172	0.9926

Nota. La tabla presenta las estimaciones realizadas por el modelo CGM AFT, para el ajuste de las variables Sexo y Etnia en el periodo febrero – abril 2020.

Esta salida permite identificar los factores de aceleración o desaceleración de las variables incorporadas respecto al Tiempo de recuperación de un paciente COVID-19, evidenciada en el aumento o disminución del Tiempo de recuperación.

En este contexto se tiene que el ser un paciente de género masculino aumenta el Tiempo de recuperación en un factor medio de 1.0026 respecto al Tiempo de recuperación de una paciente mujer, así mismo notemos que ser un miembro de la comunidad NARP o el no pertenecer a un grupo étnico disminuye el Tiempo de recuperación en factores de 0.52 y 0.34 respecto al grupo indígena, en este modelo se dejó la etnia ROM contrario al modelo AFT clásico dándole una a priori en estimación de 0, y la cual al modelar por MCMC dio como resultado que el pertenecer a esta etnia no afecta de alguna manera el Tiempo de recuperación del paciente.

Igual que en el modelo AFT clásico, se tiene que los exponenciales de los coeficientes son cercanos a 1 por lo cual los efectos de las variables no muy grandes o importantes para

el Tiempo de recuperación de un paciente COVID-19 a pesar de ser significativas para el modelo.

3.3 Discusión

Este trabajo representó un ejercicio académico que permitió el estudio de una de las situaciones más difíciles que el mundo ha tenido que enfrentar en los últimos años, como miembros de una comunidad académica el aporte que podemos hacer es promover el estudio de temas relacionados con la pandemia para que a futuro nuestra Universidad pueda aportar de una mejor manera en la búsqueda de soluciones oportunas, así como aportar evidencias que guíen los procesos de toma de decisiones.

Particularmente en nuestro contexto nacional, se tienen pocas referencias de estudios relacionados con la pandemia desde la óptica de la bioestadística, es un poco mayor la presencia de análisis de datos desde la perspectiva epidemiológica debido principalmente a la escasa presencia de esta disciplina en el ámbito académico universitario e investigativo. Estudiar el Tiempo de recuperación de un paciente y variables relacionadas a las que se les puede dar seguimiento en el marco de una pandemia puede permitir a un gobierno y a la población en general establecer las medidas necesarias en infraestructura y prevención para optimizar los recursos y mejorar la atención a los pacientes, además, repetir estos estudios en diferentes periodos de tiempo permite actualizar de manera informada las medidas establecidas y priorizar aquellas poblaciones que presenten mayor vulnerabilidad.

El uso de análisis de supervivencia para estudiar el tiempo hasta la presentación de un evento no es un tema de reciente interés, de hecho, está muy consolidado en el ámbito clínico para el estudio de tratamientos médicos incluso por encima del ámbito industrial

donde se sitúa su origen. Es de notar que este tipo de herramientas de análisis hubiese sido muy relevantes implementarlas en lo que fueron las primeras etapas de la pandemia de COVID-19, dado que era ese el momento donde se poseía información limitada y era necesario actuar en consonancia con los pocos datos a disposición, lo inesperado del problema y la poca cultura de la información en nuestro país condujo a que si bien se cuenta con una base de datos que da cuenta desde el primer caso hasta el más reciente, ésta carece de información importante para soportar estudios de mayor envergadura, en particular no hay registro de variables como: tiempo exacto de hospitalización, síntomas, comorbilidades, tipo de servicio médico a disposición, información socioeconómica, identificación de reinfección y estatus frente al esquema de vacunación, por mencionar algunas.

Finalmente advertir que para establecer el evento de recuperación del virus para el estudio del Tiempo de recuperación se usó la variable Tipo de recuperación de la base de datos Casos positivos para COVID-19 en Colombia del INS la cual informaba sobre la no presencia del virus a través de la realización de una prueba PCR con resultado negativo, desafortunadamente debido a las dificultades de infraestructura y gran cantidad de pruebas que se acumularon en el país, el tiempo exacto de recuperación no se conoce para la gran mayoría de individuos registrados en la base de datos, razón por la cual existe una alta cantidad de censura de individuos, hecho que claramente limita la capacidad de análisis.

4. Conclusiones

El análisis de las variables a disposición que podrían constituirse en potenciales factores de riesgo para el contagio por COVID-19 como son: el Sexo, la Etnia y el Clima del municipio de residencia mostraron diferentes resultados; en cuanto al número de contagios, acorde a lo mostrado por las Figuras 7, 8, 12 y 14, se concluye que ni el Sexo ni el clima describen diferencias significativas en la incidencia de esta enfermedad. En cuanto al número de muertes asociadas al COVID encontramos que el Sexo masculino evidenció diferencias considerables en gran parte del período de tiempo y con mayores diferencias en los tres picos de la pandemia (Figura 9), en cuanto a la Etnia se encontró que los Rom fue la más afectada seguida por los Indígenas y contrario a lo acontecido en otros países la comunidad NARP que agrupa a los afrodescendientes no describió los mayores indicadores entre las minorías que hacen presencia en Colombia (Figura 13). Se debe resaltar que los Rom son la minoría con menos personas identificadas como tal, según el último censo nacional en Colombia hay 2649 personas pertenecientes a la etnia Rom o gitanos, siendo un 0.006% de la población colombiana.

En cuanto a la Edad se tiene que al explorar diferencias debidas a la Edad al discriminar por Sexo no se encontraron diferencias en la distribución del número de contagios ni en cuanto a la mortalidad por rangos de Edad (Figura 7,10 y 11). En particular, la población contagiada en hombres y mujeres se distribuyó de manera similar con un comportamiento asimétrico a derecha, donde las edades de mayor contagio están entre los 20 y los 60 años (Figura 7), mientras que la distribución de la población que falleció muestra un

comportamiento asimétrico a izquierda, concentrándose en pacientes entre los 40 y 60 años tanto en hombres como en mujeres (Figuras 10 y 11).

En cuanto al ajuste de modelos de supervivencia no paramétricos para la variable Tiempo de recuperación se utilizó el método Kaplan-Meier para ajustar las curvas de supervivencia a periodos de tiempo de aproximadamente 2 meses, las mayores diferencias en estas curvas se observaron en los 3 primeros periodos de tiempo (Figuras 15 y 19) que corresponden a los meses Febrero-marzo-abril y Mayo-junio de 2020, en todas ellas como es usual la gráfica muestra un comportamiento decreciente pero evidencia que las probabilidades de no recuperarse pasados los 30 días siguen siendo considerables (0.30 – 0.75).

También usando Kaplan-Meier, las curvas de supervivencia ajustadas para los demás periodos de tiempo muestran un comportamiento similar a la de Julio-agosto de 2020 (Figura 21) donde la probabilidad de supervivencia decrece lentamente mostrando que en estos periodos la probabilidad de recuperarse después de 30 días es baja y se mantiene así hasta el final del período en consideración, es decir a los 62 días (Apéndices del A al F).

Al ajustar las curvas de supervivencia por Kaplan-Meier discriminando por las variables Edad y Etnia, se apreció que se muestra una mayor afectación a nivel de recuperación para los que tienen más de 50 años (Figura 17), también para aquellos que pertenecen a una minoría étnica (Figura 18) mientras que los menores de 15 años y jóvenes entre 16 y 25 años describen probabilidades de recuperación más altas (Figura 18).

Los modelos de tiempo de falla acelerado (AFT) se ajustaron con el fin de identificar las variables que pudieran tener efecto sobre el Tiempo de recuperación, en primera instancia varios modelos fueron probados utilizando diferentes distribuciones y utilizando criterios

estadísticos (AIC y BIC) se concluyó que el mejor ajuste se consiguió al utilizar una distribución gamma generalizada. Así al ajustar los modelos AFT gamma generalizados el de menor AIC después del que involucra todas las variables fue el que involucró las covariables Grupo de edad y Etnia con un AIC de 19435.47. De lo anterior se concluye que las variables Grupo de edad y Etnia muestran ser significativas pero la interpretación del efecto en términos prácticos no es relevante, debido a los bajos factores de aceleración o desaceleración obtenidos.

Uno de los propósitos de este trabajo fue explorar la alternativa bayesiana, en este sentido se logró estimar en primera instancia la curva de supervivencia a partir de dos distribuciones a priori no informativas la gamma dependiente y la gamma independiente; al comparar el comportamiento de la curva de supervivencia dada por el estimador de Kaplan Meier con la a posteriori de la gamma dependiente lucen muy similares, mientras que la curva obtenida al usar como a priori la gamma independiente evidenció una gran diferencia a partir de los 40 días, con esto se observa como en este enfoque la escogencia de la distribución es fundamental para encontrar una distribución a posteriori con buen ajuste al fenómeno de interés. Sobre los Intervalos de credibilidad que se muestran en las Figuras 27 y 28 se destaca su precisión durante todo el período de estudio lo que garantiza un buen recurso para hacer pronósticos.

Adicionalmente a lo anterior, para evaluar el efecto de factores que caracterizan la población se ajustó un modelo de falla acelerado usando una mezcla gaussiana clásica para la distribución de los errores (CGM AFT) para los datos del primer período de la pandemia (febrero-abril 2020); el resultado mostró una estimación diferente para el efecto de las variables Sexo y Etnia en el Tiempo de recuperación en relación con el uso del AFT clásico

(Tabla 5 y 6) con lo cual tampoco se logró identificar variables cuyo efecto de cuenta de aceleración o retroceso del proceso de recuperación.

En nuestro caso no se contó con el criterio especializado para orientar el ajuste por métodos bayesianos sólo se hizo uso de fuentes bibliográficas, pero tal como se postula dentro de las ventajas del uso de Estadística bayesiana lo ideal es involucrar la mayor cantidad de información posible sobre el contexto adicional a los datos. No obstante, las curvas de supervivencia obtenidas lucen muy próximas con lo cual no es posible destacar una de las dos metodologías como más eficiente que la otra. Los métodos clásicos suponen la validación de supuestos, en el caso de un análisis de supervivencia esto conlleva a validar que tanto el evento de interés como la censura se dan de manera independiente, es decir, los individuos censurados tienen la misma probabilidad de presentar el evento que los no censurados, tal vez este sería el escenario donde el enfoque bayesiano tome ventaja pues para iniciar el modelo no se exige tal validación.

Referencias Bibliográficas

Abdul-Fatawu, M. (2020). Accelerated Failure Time Models: An Application in Insurance Attrition [Modelos de tiempo de falla acelerado: una aplicación en la deserción de seguros]. The Journal of Risk Management and Insurance, Bangkok, Thailand: The University.24(2). hal-02953269.

Alboukadel Kassambara, Marcin Kosinski, Przemyslaw Biecek y Scheipl Fabian. (2021). A Package for Drawing Survival Curves using 'ggplot2' in R [Un paquete para dibujar curvas de supervivencia usando 'ggplot2' en R]. R package versión 0.4.9. https://CRAN.R-project.org/package=survminer.

Bogaerts,K., Komárek, A. y Lesaffre E. (2017). Survival Analysis with Interval-Censored Data: A Practical Approach with Examples in R [Análisis de supervivencia con datos censurados por intervalos: un enfoque práctico con ejemplos en R], SAS, and BUGS. Chapman & Hall/CRC Press. ISBN: 978-1-42-007747-6.

Box, G. E. P. y Tiao, G. C. (1973). Bayesian Inference in Statistical Analysis [Inferencia bayesiana en análisis estadístico]. John Wiley & Sons, New York.

Brilleman, S., Elci, E., Buros Novik, J y Wolfe, R. (2020). Bayesian Survival Analysis Using the rstanarm R Package [Análisis de supervivencia bayesiana con el paquete de R rstanarm]. Cornell University. https://arxiv.org/abs/2002.09633.

Castro-Kuriss, C. (2018). Análisis de Sobrevida mediante el software R. ResearchGate. https://doi.org/10.13140/RG.2.2.28360.62720 .

Clark, T., Bradburn, M. y Love, S. (2003). Survival Analysis Part I: Basic concepts and first analyses. [Análisis de supervivencia parte I: Conceptos básicos y primeros análisis]. *Br J Cancer*. 89. 232–238. https://doi.org/10.1038/sj.bjc.6601118.

Coghlan, A. (2017). A Little Book of R For Bayesian Statistics [Un pequeño libro de R para estadísticas bayesianas]. Wellcome Trust Sanger Institute, Cambridge, U.K.

Correa, J.C y Barreara, C. J. (2018). Introducción a la Estadística Bayesiana. Fondo Editorial ITM.

Departamento Administrativo Nacional de Estadísticas (DANE). (2018). Censo Nacional de Población y Vivienda 2018-Colombia. DANE Información para Todos. https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/censo-nacional-de-poblacion-y-vivenda-2018.

Diaz, H., España, Guido., Castañeda, N., Rodriguez, L., y de la Hoz-Restrepo, F. (2021). Dynamical characteristics of the COVID-19 epidemic: Estimation from cases in Colombia [Características dinámicas de la epidemia de COVID-19: estimación a partir de casos en Colombia]. International Journal of Infectious Diseases. 105, 26-31. https://doi.org/10.1016/j.ijid.2021.01.053.

Gayathri, T., Ravanan, R. y Lakshmi, M. (2021). Modeling the recovery time of patients with coronavirus disease 2019 using an accelerated failure time model [Modelamiento del tiempo de recuperación de pacientes con enfermedad por coronavirus 2019 utilizando un modelo de tiempo de falla acelerado]. Journal of International Medical Research. 49(8), 1–7. https://doi.org/10.1177/03000605211040263.

Instituto Nacional de Salud. (2020-2021). Casos positivos de COVID-19 en Colombia. [Conjunto de Datos]. https://www.datos.gov.co/Salud-y-Protecci-n-Social/Casos-positivos-de-COVID-19-en-Colombia/gt2j-8ykr/data .

Ibrahim, J.G., Chen, Ming-Hui. y Sinha, D. (2005). Bayesian Survival Analysis. [Análisis de supervivencia bayesiano]. Journal of The American Statistical Association - J AMER STATIST ASSN. 99. https://doi.org/10.1002/0470011815.b2a11006.

Instituto Nacional de Salud. (2021). COVID-19 en Colombia. https://www.ins.gov.co/Noticias/Paginas/Coronavirus.aspx.

Jackson, C. H. (2016). flexsurv: a platform for parametric survival modeling in R [flexsurv: una plataforma para el modelado de supervivencia paramétrico en R]. Journal of statistical software, 70. https://doi.org/10.18637/jss.v070.i08.

Johnson, A., Ott, M y Dogucu, M. (2021). Bayes Rules! An Introduction to Bayesian Modeling with R [¡Reglas de Bayes! Introducción al modelado bayesiano con R]. CRC Press Taylor & Francis Group. https://www.bayesrulesbook.com/index.html.

Khan, F., Ali, S., Saeed, A., Kumar, R. y Khan, A. W. (2021). Forecasting daily new infections deaths and recovery cases due to COVID-19 in Pakistan by using Bayesian Dynamic Linear Models [Pronóstico diario de muertes por nuevas infecciones y casos de recuperación debido a COVID-19 en Pakistán mediante el uso de modelos lineales dinámicos bayesianos]. PLoSONE. 16(6): e0253367. https://doi.org/10.1371/journal.pone.0253367.

Komárek, A. (2020). A Package for Bayesian Survival Regression with Flexible Error and Random Effects Distributions in R [Regresión Bayesiana de Supervivencia con

Distribuciones Flexibles de Error y Efectos Aleatorios en R]. R package version 3.3. https://CRAN.R-project.org/package=bayesSurv.

Linasari, D. (2021). Survival Analysis of Covid 19 Patients from Two Hospitals in Cimahi, Indonesia [Análisis de supervivencia de pacientes con Covid 19 de dos hospitales en Cimahi, Indonesia]. Advances in Health Sciences Research, 37. https://doi.org/10.2991/ahsr.k.210723.041.

Martinez, J. (2017). Análisis de Supervivencia en R. http://rstudio-pubs-static.s3.amazonaws.com/316989_83cbe556125645b698c9ff6cf88c4c1a.html#1_introducción.

Mesa Páez, L. O., Rivera Lozano, M. y Romero Davila, J. A. (2011). Descripción general de la Inferencia Bayesiana y sus aplicaciones en los procesos de gestión. Universidad del Rosario.

Ministerio de Vivienda, Corporación Financiera Internacional Grupo Banco Mundial, Cámara Colombiana de la Construcción, y State secretariat for economic affairs. (2020). Mapa de Clasificación del Clima en Colombia según la Temperatura y la Humedad Relativa y listado de municipios (Anexo No. 2). https://docplayer.es/40184546-Anexo-no-2-mapa-de-clasificacion-del-clima-en-colombia-segun-la-temperatura-y-la-humedad-relativa-y-listado-de-municipios.html.

Mollazehi, M., Mollazehi, M. y Abdel-Salam, A. (2020). Modeling Survival Time to Recovery from COVID-19: A Case Study on Singapore [Modelado del tiempo de supervivencia hasta la recuperación de COVID-19: un estudio de caso sobre Singapur]. Research Square. https://doi.org/10.21203/rs.3.rs-18600/v2.

Pere, R (2005). Conceptos básicos del análisis de supervivencia. Cirugía Española, Corporación Sanitaria Parc Taulí. Sabadell. Barcelona, España. 78(4), 222–230. https://doi.org/10.1016/S0009-739X(05)70923-4.

San José, B., Pérez, E. y Madero, R. (2009). Métodos estadísticos en estudios de supervivencia. Anales de Pediatría Continuada. 7(1), 55-59. https://doi.org/10.1016/S1696-2818(09)70453-6.

Therneau, T. (2022). A Package for Survival Analysis in R [Un paquete para análisis de supervivencia en R]. R package version 3.3-1, https://CRAN.R-project.org/package=survival.

tok.wiki. (s. f.). Modelo de tiempo de falla acelerado Especificación del modelo y Problemas estadísticos. Wiki. https://hmong.es/wiki/Accelerated_failure_time_model.

Van der Pas, S. y Castillo, I. (2021). A package for Bayesian Survival Analysis for Right Censored in R [Un paquete para análisis de supervivencia bayesiano para censura por la derecha en R]. https://CRAN.R-project.org/package=BayesSurvival.

Villers, S., Vásquez, C.F. y Ramirez, L.A. (2021). Modelos de Supervivencia. https://carlosfernandovg.github.io/supervivencia_y_series_FC2021-1/index.html#licencia.

World Health Organization. (2021). WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int.

World Health Organization. (2021). COVID-19 Weekly Epidemiological Update [Actualización epidemiológica semanal de COVID-19]. 69. https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---7-december-2021.

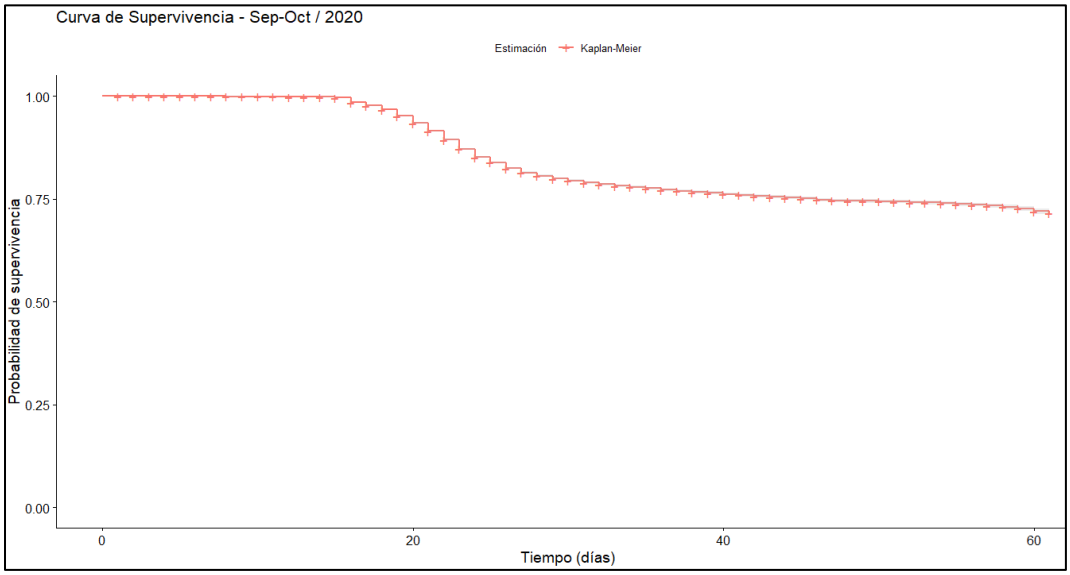
Xiang Gao. y Qunfeng Dong. (2020). A primer on Bayesian estimation of prevalence of COVID-19 patient outcomes [Introducción a la estimación bayesiana de la prevalencia de los resultados de los pacientes con COVID-19]. *JAMIA Open.* 3(4). 628–631. https://doi.org/10.1093/jamiaopen/ooaa062.

Yadav, S.K. y Akhter, Y. (2021). Statistical Modeling for the Prediction of Infectious Disease Dissemination With Special Reference to COVID-19 Spread [Modelamiento estadístico para la predicción de la diseminación de enfermedades infecciosas con especial referencia a la diseminación de COVID-19]. Front. Public Health. 9:645405. https://doi.org/10.3389/fpubh.2021.645405.

Apéndices

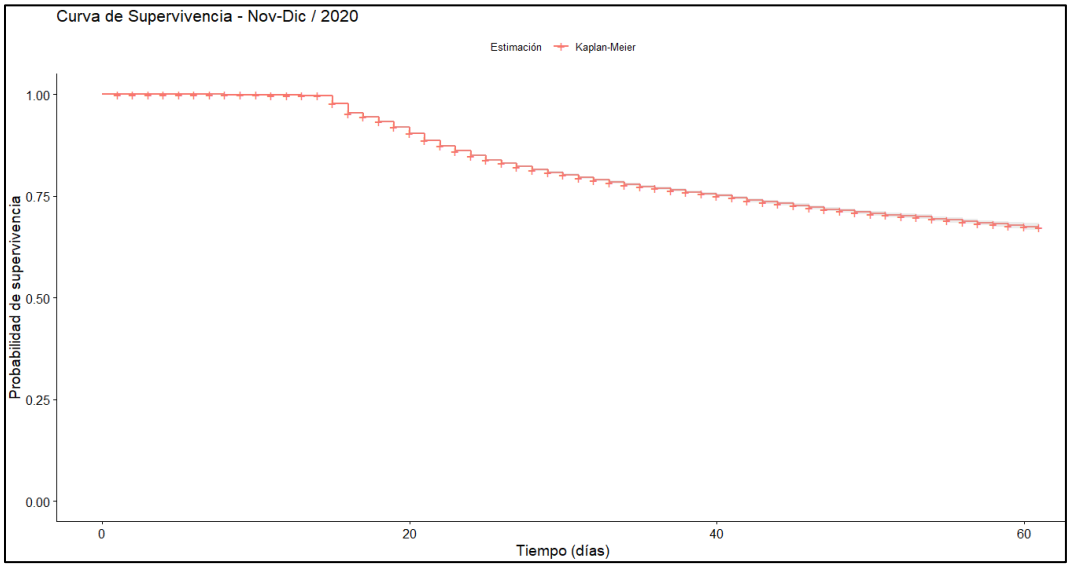
Apéndice A

Ajuste Kaplan Meier septiembre-octubre 2020



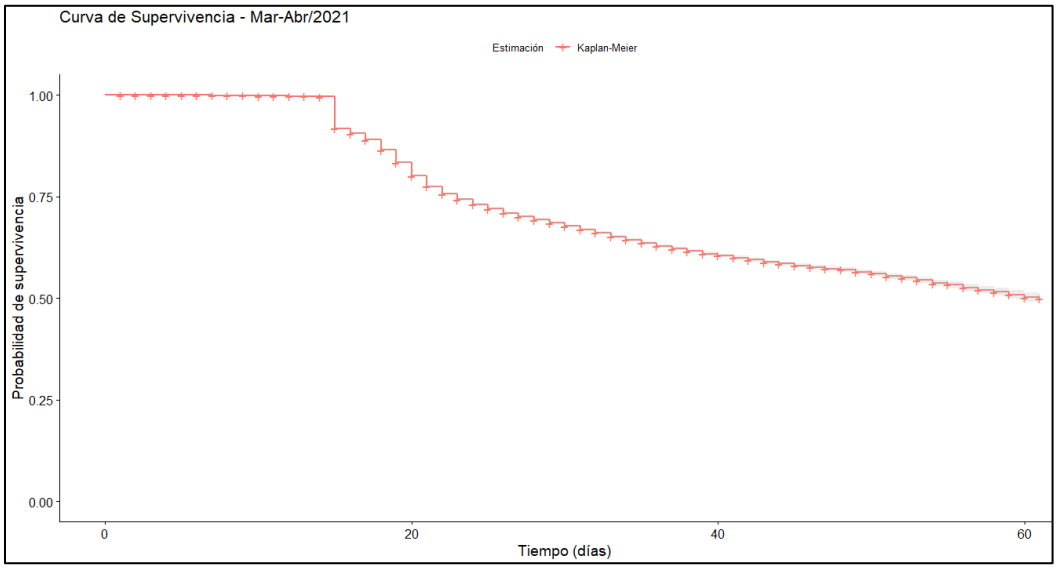
Apéndice B

Ajuste Kaplan Meier noviembre-diciembre 2020



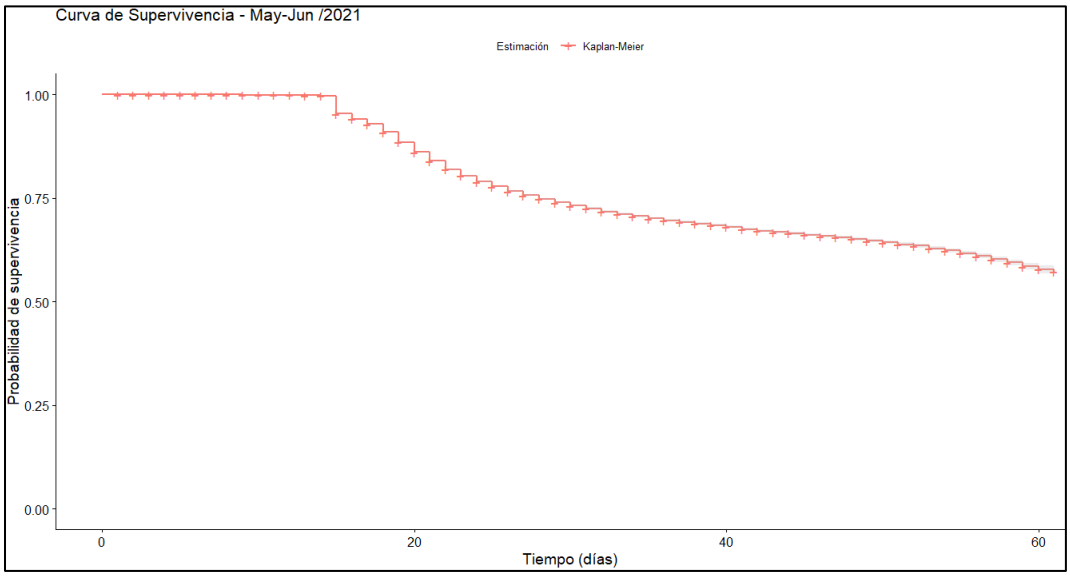
Apéndice C

Ajuste Kaplan Meier marzo-abril 2021



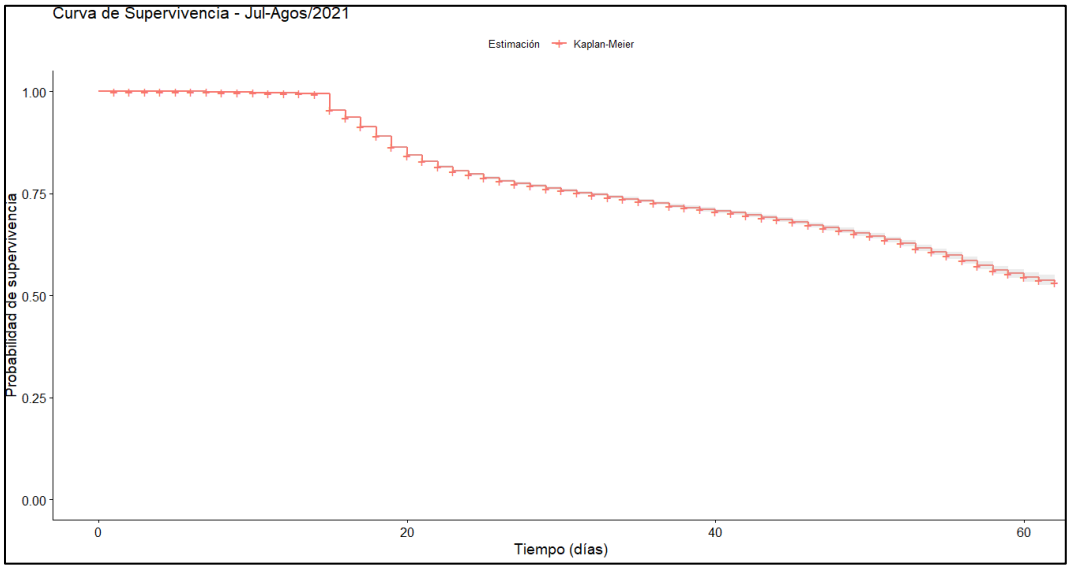
Apéndice D

Ajuste Kaplan Meier mayo-junio 2021



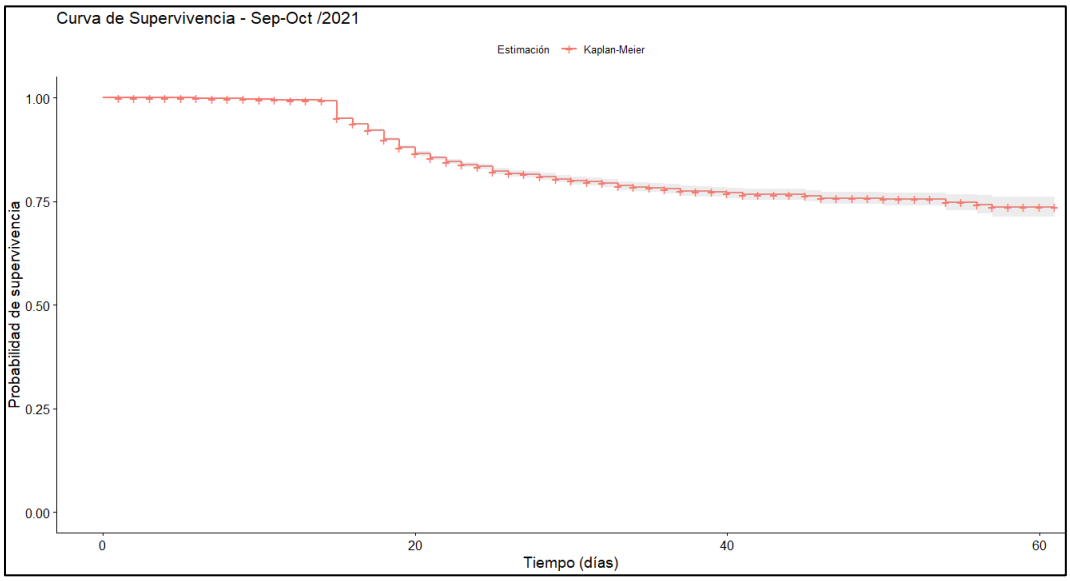
Apéndice E

Ajuste Kaplan Meier julio-agosto 2021



Apéndice F

Ajuste Kaplan Meier septiembre-octubre 2021



Apéndice G

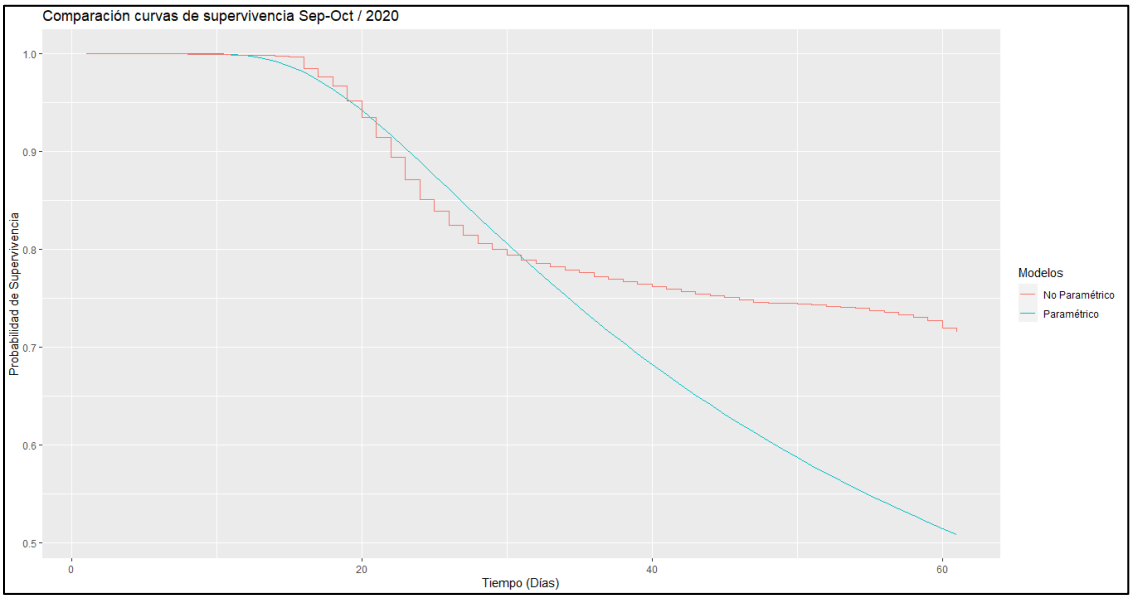
Criterios de bondad de ajuste AIC y BIC 2020-2021

Año	Periodo		Gamma Generalizada	Log Normal	Log Logística	Weibull	Exponencial
2020	02-04	AIC	19522.02	19764.92	19985.91	20287.52	22809,45
		BIC	19543.65	19779.34	20000.33	20301.94	22816,66
	05-06	AIC	230148.4	230376.7	230325.9	232190.1	285316.7
		BIC	230177.9	230345.6	230396.4	232209.7	285326.5
	07-08	AIC	562528.6	562655.3	566233.5	575654.1	673630.9
		BIC	562562.0	562677.5	566255.8	575676.4	673642.1
	09-10	AIC	327218.5	334352	341958.9	346801.2	387624.7
		BIC	327251.5	334374	341980.8	346823.2	387635.7
	11-12	AIC	357218.2	365200.6	373833.3	379162.3	425497.8
		BIC	357251.9	365223.1	373855.8	379184.8	425509.1
2021	01-02	AIC	668885.8	699522	715376.8	733499.5	799552.1
		BIC	668918.9	699544	715398.9	733521.6	799563.1
	03-04	AIC	697976.8	721249.8	738127.1	759469.7	868552.7
		BIC	698011.1	721272.7	738150.0	759492.6	868564.1
	05-06	AIC	1494516	1084790	1111909	1136075	1292222
		BIC	1494552	1084814	1111933	1136099	1292234
	07-08	AIC	356261.2	364791.0	373202.5	380786.4	423297.0
		BIC	356293.7	364812.7	373224.1	380808.0	423307.8
	09-10	AIC	51451.51	52097.35	53228.39	54188.95	61018.08
		BIC	51479.53	52116.03	53247.06	54207.63	61027.42
	11-12	AIC	50970.50	51127.36	52025.62	52806.83	62886.52
		BIC	51002.58	51148.74	52047.01	52828.22	62897.21

Nota. El apéndice presenta una tabla con los diferentes valores de AIC y BIC de los modelos ajustados en cada periodo con las diferentes distribuciones, el periodo mayo-junio de 2021 es el único que difiere en la distribución mejor ajustada.

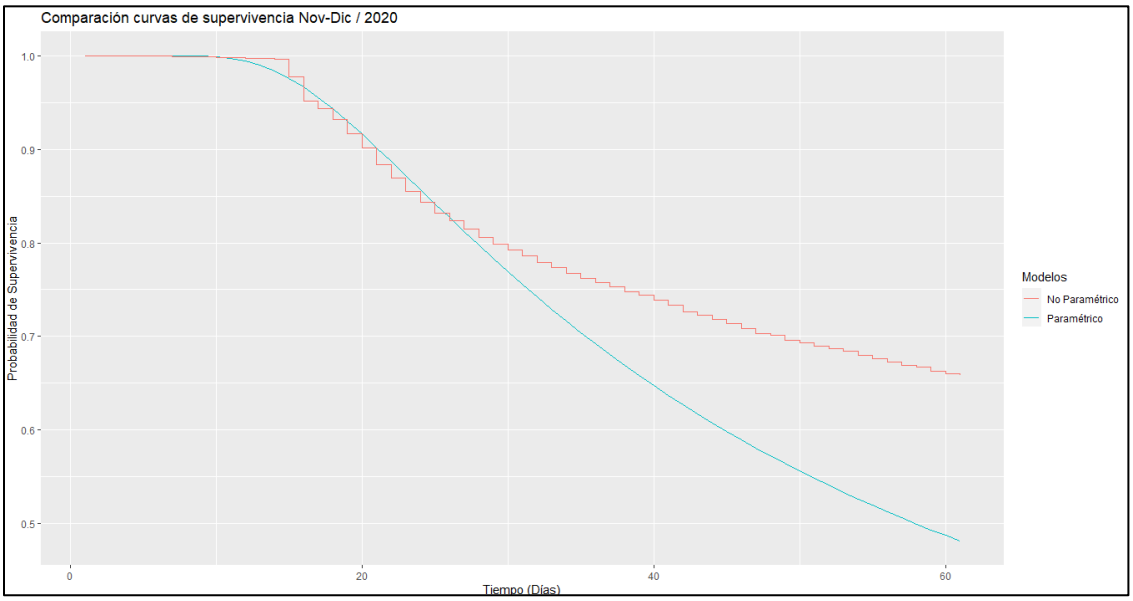
Apéndice H

Comparación Kaplan Meier y modelo paramétrico septiembre-octubre 2020



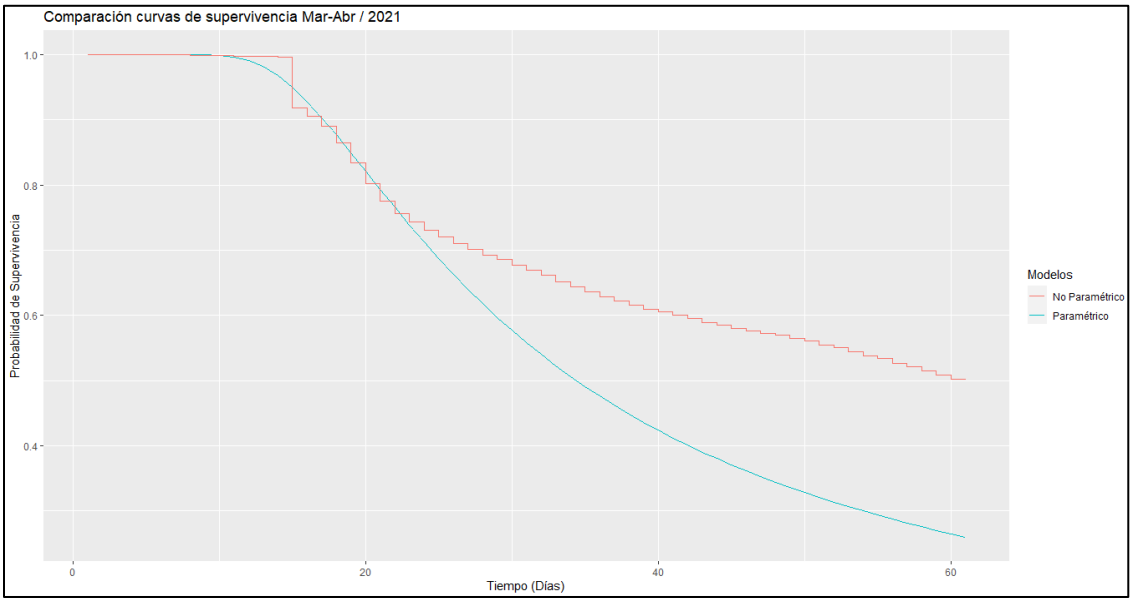
Apéndice I

Comparación Kaplan Meier y modelo paramétrico noviembre-diciembre 2020



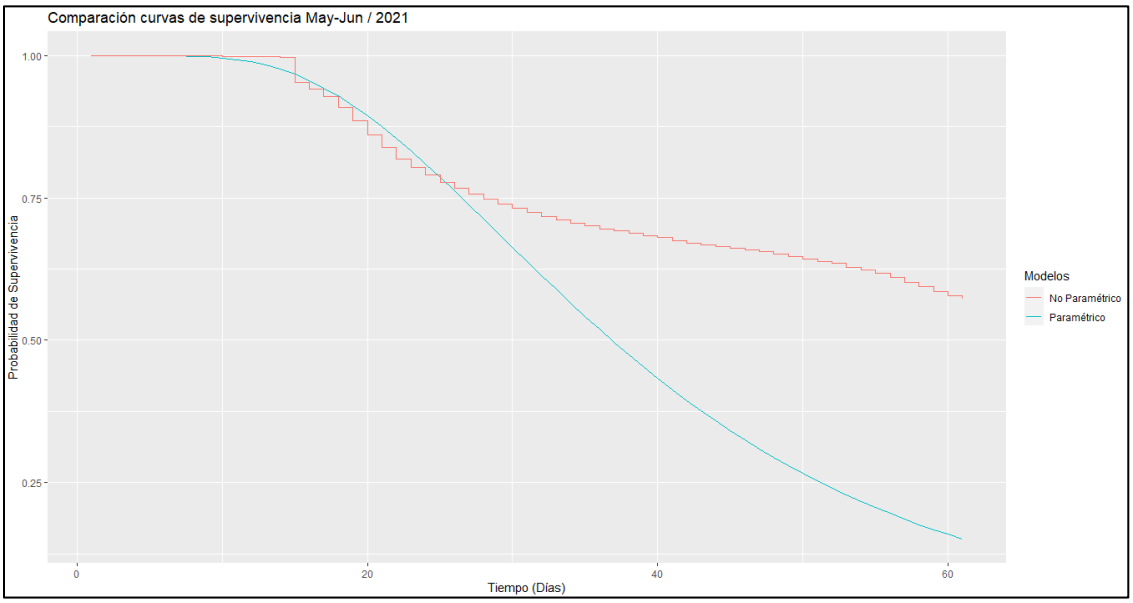
Apéndice J

Comparación Kaplan Meier y modelo paramétrico marzo-abril 2021



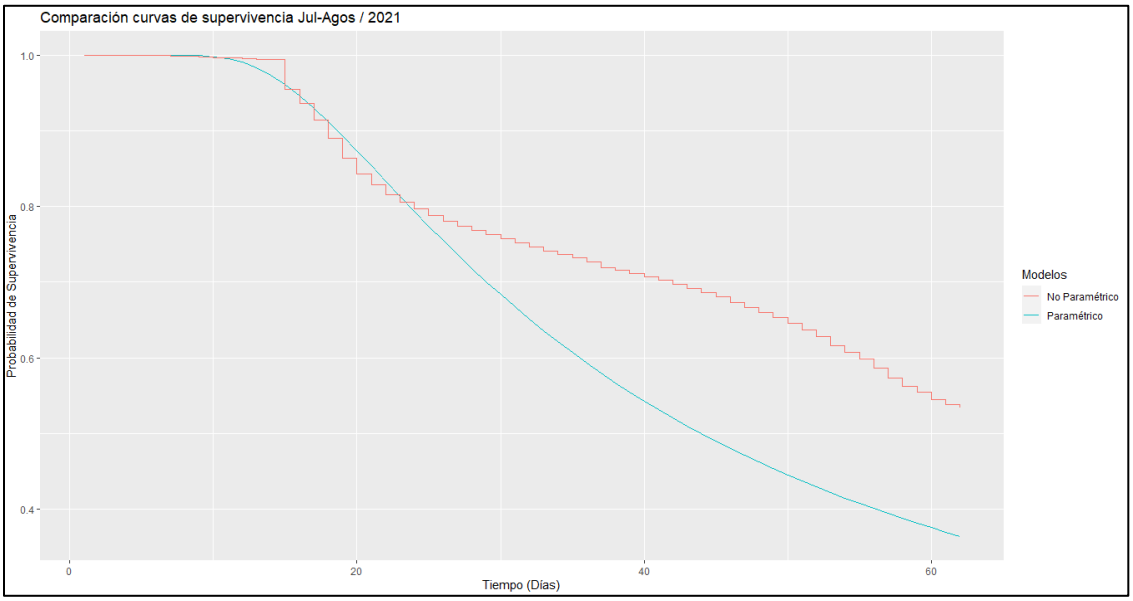
Apéndice K

Comparación Kaplan Meier y modelo paramétrico mayo-junio 2021



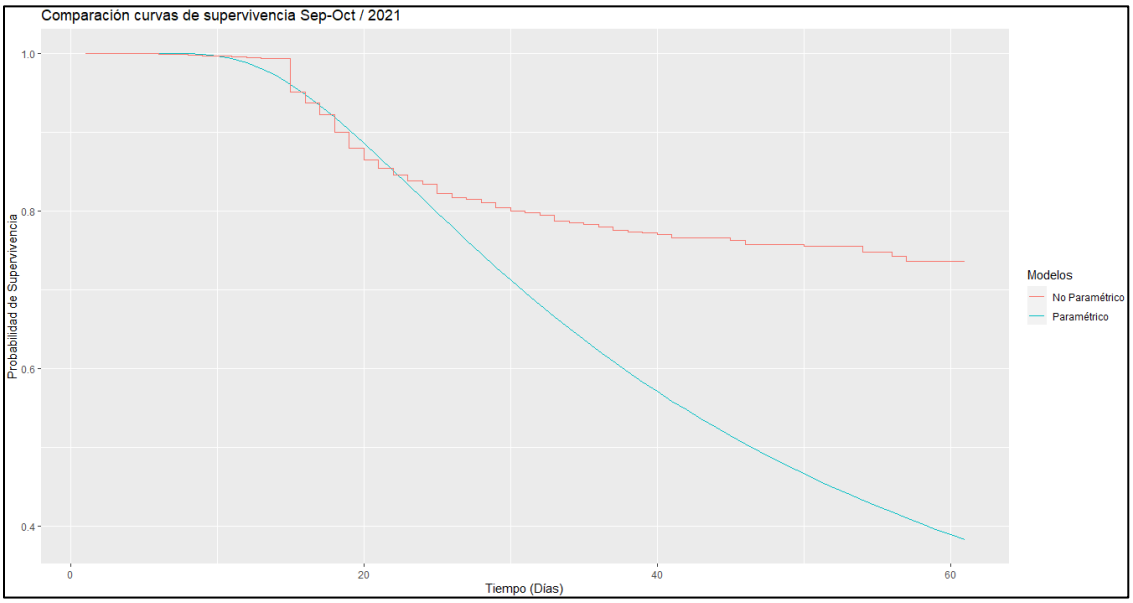
Apéndice L

Comparación Kaplan Meier y modelo paramétrico julio-agosto 2021



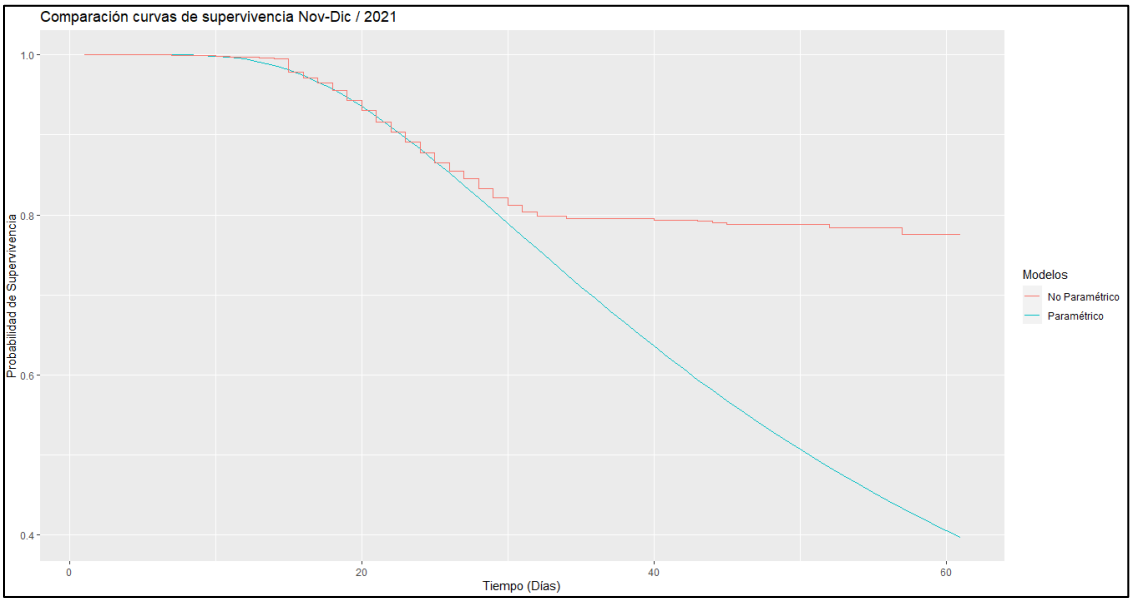
Apéndice M

Comparación Kaplan Meier y modelo paramétrico septiembre-octubre 2021



Apéndice N

Comparación Kaplan Meier y modelo paramétrico noviembre-diciembre 2021



Apéndice O

Código en R de gráficas, salidas y ajustes de modelos

El código que se usó para la creación de gráficas, la realización de ajustes y demás procesos de este trabajo se encuentran a continuación, siendo “dcovid” la base de datos:

Librerias

```
library(e1071)
library(ggplot2)
library(survival)
library(KMsurv)
```

```
library(survMisc)
library(survmine)
library(flexsurv)
library(coda)
```

```
library(bayesSur)
library(BayesSurvival)
```

Estadística Descriptiva

Histogramas Contagio Edad por Sexo

```
f<-filter(dcovid,dcovid$Sexo=="F")
m<-filter(dcovid,dcovid$Sexo=="M")
hist(f$Edad, main="Histograma Edad Pacientes Mujeres COVID-19", xlab="Años de Edad",ylab="Frecuencia" )
hist(m$Edad, main="Histograma Edad Pacientes Hombres COVID-19", xlab="Años de Edad",ylab="Frecuencia" )
```

Diagrama de Líneas Contagio por Sexo

```
dcovid$mes_anno <- format(as.Date(dcovid$fechanotificacion), "%Y-%m")
dmes<-split(dcovid, f = dcovid$mes_anno) # división por meses
```

```
table(dmes$`2020-03`$Sexo)
table(dmes$`2020-04`$Sexo)
table(dmes$`2020-05`$Sexo)
table(dmes$`2020-06`$Sexo)
table(dmes$`2020-07`$Sexo)
table(dmes$`2020-08`$Sexo)
table(dmes$`2020-09`$Sexo)
table(dmes$`2020-10`$Sexo)
table(dmes$`2020-11`$Sexo)
table(dmes$`2020-12`$Sexo)
table(dmes$`2021-01`$Sexo)
```

```
table(dmes$`2021-02`$Sexo)
table(dmes$`2021-03`$Sexo)
table(dmes$`2021-04`$Sexo)
table(dmes$`2021-05`$Sexo)
table(dmes$`2021-06`$Sexo)
table(dmes$`2021-07`$Sexo)
table(dmes$`2021-08`$Sexo)
table(dmes$`2021-09`$Sexo)
table(dmes$`2021-10`$Sexo)
table(dmes$`2021-11`$Sexo)
table(dmes$`2021-12`$Sexo)
```

Gráfico de líneas

```
df <- data.frame(
 "Meses" = as.character(c("20-03","20-04","20-05","20-06","20-07","20-08","20-09","20-
        10","20-11","20-12","20-03","20-04","20-05","20-06","20-07","20-08","20-
        09","20-10","20-11","20-12","21-01","21-02","21-03","21-04","21-05","21-
        06","21-07","21-08","21-09","21-10","21-11","21-12","21-01","21-02","21-
        03","21-04","21-05","21-06","21-07","21-08","21-09","21-10","21-11","21-12")),
 "Género"=
        as.character(c("M","M","M","M","M","M","M","M","M","M","M","F","F","F","F","F",
        "F","F","F","F","F",M","M","M","M","M","M","M","M","M","M","M","M","M","F","F
        ","F","F","F","F","F","F","F","F","F","F"),
 "Número_Pacientes"=
        c(965,3681,16559,51732,137235,136168,104300,127508,108137,178500,
        948,3313,14135,44104,128140,139066,110448,134513,118061,199769,
        190950,54501,99992,230122,286457,396247,181138,39286,19409,20500,30213,83
        951,220823,62028,113726,263975,326472,453371,209524,45673,22456,24017,357
        23,107707))
ggplot(df, aes(x=Meses, y= Número.de.Pacientes, group = Género, colour =Género),
        ylim=c(0,500000)) + ggtitle("Pacientes COVID-19 por meses y discriminando por
        género 2020-2021") + geom_line() + geom_point( size=2, shape=21, fill="white") +
        theme_minimal()
```

Diagramas de Barras Contagios Etnia

#POBLACIÓN	#ENFERMOS	sc1=P1-n1
P1=1905617	n1=74030	sc2=P2-n2
P2=2649	n2=111	sc3=P3-n3
P3=4671160	n3=122435	sc4=P4-n4
P4=41679068	n4=5075444	

```
datos = c (n1,sc1 ,n2,sc2, n3,sc3,n4,sc4)
tabla = cbind ( expand.grid ( list ( Estado = c ( " Infectado" ," No infectado " ) , Etnia = c ("Indigena","Rom" , "NARP", "Otro"))) , count = datos )
tetnia=ftable ( xtabs ( count~ Estado+Etnia  , tabla ))
petnia=prop.table(tetnia, margin = 2)
dpetnia<-data.frame(petnia)
```

Número de enfermos por cada mil personas pertenecientes a la etnia (Tasas).

```
dpetnia$Tasa<-(dpetnia$Freq)*1000
tetnia=ftable(xtabs( dpetnia$Tasa~ dpetnia$Estado+dpetnia$Etnia  , dpetnia$dpetnia ))
```

Gráfico Tasas

```
barplot(tetnia[1,], main = "Tasa de Pacientes covid-19 por pertenencia etnica", xlab = "Etnia", ylab = "Número de enfermos por cada mil", ylim=c(0,150), col = c("darkblue"), beside = TRUE, names.arg = c("Indigena","Rom" , "NARP", "Otro"))
```

Tasas Contagios por Clima

Marzo -abril- mayo

```
dcl1<-split(df1, f = df1$Clima)
V<-c("CALIDO HUMEDO","CALIDO SECO","FRIO","TEMPLADO")
y<-vector()
for(i in 1:4){
        mun1<-data.frame (ftable(dcl1[[V[i]]]$MUNICIPIO,dcl1[[V[i]]]$DEPARTAMENTO))
        mun1<-filter(mun1,mun1$Freq!=0)
        mun1 = rename(mun1, c(Var1="MUNICIPIO"))
        mun1 = rename(mun1, c(Var2="DEPARTAMENTO"))
        mun1<-merge(x = mun1, y = dfclima)
        mun1$Tasa<-mun1$Freq/mun1$`POBLACIÓN - 2020`*1000
        y[i]=mean(mun1$Tasa) }
table=cbind( expand.grid ( list ( CLIMA = V)), count = y )
```

Gráfico

```
plot(table, ylab="Tasa Media", main="Tasa Media por Clima Mar-Abril-May",ylim=c(0,15))
```

Supervivencia clásica

Para el primer periodo de tiempo Febrero, Marzo y Abril

```
df1<-filter(dcovid, fechainiciosintomas >= as.Date("2020-02-01") &
        fechainiciosintomas<=as.Date("2020-04-30"))
df1$delta =ifelse(!is.na(df1$fecharecuperacion) & df1$fecharecuperacion<=as.Date("2020-04-
        30")& df1$`Tipo de recuperación`=="PCR",1,0)
df1$tiempo1=
```

```
ifelse(is.na(df1$fecharecuperacion) & !is.na(df1$fechamuerte) & df1$fechamuerte < =
        as.Date("2020-04-30"),df1$fechamuerte-df1$fechainiciosintomas,
        ifelse(!is.na(df1$fecharecuperacion) & df1$fecharecuperacion < = as.Date("2020-04-30"),
        df1$fecharecuperacion - df1$fechainiciosintomas, as.Date("2020-04-30")-
        df1$fechainiciosintomas))
```

Kaplan- Meier

```
dcovid.km1 <- survfit(Surv(df1$tiempo1+1, df1$delta) ~ 1, data = df1, type = "kaplan-meier")
```

Gráfico de la curva

```
ggsurvplot(fit = dcovid.km1, data = df1, conf.int = T, title = "Curva de Supervivencia - Feb-Mar-
        Abr" , xlab = "Tiempo (días)",ylab="Probabilidad de supervivencia" ,legend.title =
        "Estimación" , surv.median.line = "hv",legend.labs = "Kaplan-Meier")
```

Curvas por grupos

Edad

```
survfit(Surv(df1$tiempo1+1, df1$delta) ~ df1$Gedad, df1) %>% ggsurvplot(title =
        "Supervivencia por Grupos de edad", conf.int = T, surv.median.line = "hv", legend.title =
        "Grupo Edad",legend.labs = c("[26,50]",">50","[16,25]","[0,15]"), xlab = "Tiempo
        (días)",ylab="Probabilidad de supervivencia",pval = T)
```

Modelos Paramétricos

Elección mejor ajuste de distribución

```
Dist <- c("exp", "weibull", "llogis","lnorm","gengamma")
data.Surv <- Surv(df1$tiempo1+1, df1$delta)
```

```
model <- sapply(Dist, function(x) flexsurvreg(data.Surv ~ 1, dist = x), USE.NAMES = T, simplify
        = F)
AIC.model <- sapply(model, function(x) c(AIC = AIC(x), BIC = BIC(x)), simplify = T)
```

AFT

Etnia y Sexo

```
aft1<-flexsurvreg(Surv(df1$tiempo1+1,df1$delta)~df1$Sexo+df1$Etnia, data = df1, dist =
        "gengamma")
```

Supervivencia Bayesiana

Modelamiento a priori gamma dependiente

```
CSBAYES <- BayesSurv(df = df1, time = "tiempo", event = "delta", prior = "Dependent")
```

Gráfico de curvas

```
gg <- PlotBayesSurv(bayes.surv.object = CSBAYES, object = "survival",xlab="Tiempo (Días)",
        ylab="Probabilidad de supervivencia",legend = F, color = "red")
        km <- survfit( Surv(df1$tiempo, df1$delta) ~ 1, data = df1 ) #Kaplan-Meier
        df.km <- data.frame(t = km$time, km = km$surv)
gg <- gg + geom_line(data = df.km, aes(x = t, y = km), colour = "black", size = 1, lty = 6)
gg <- gg + labs(title = "Curva de supervivencia Feb-Abr 2020 a priori Gamma Dependiente")
```

Modelo CGM AFT

A prioris

```
prior.eps <- list()
        ## A priori e hiperparametro de K
```

```
prior.eps$kmax <- 20
prior.eps$k.prior <- "uniform"
        ## A prioris para la media
prior.eps$mean.mu <- 3.38
prior.eps$var.mu <- 10^2
        ## Hiperparametros a priori para la distribución a priori de Dirichlet
        ## Valores de la mezcla
prior.eps$dirichlet.w <- 1
        ## Número de parámetros de regresión
nregres <- 4
        ## A priori betas
prior.beta <- list()
prior.beta$mean.prior <- rep(0, nregres)
prior.beta$var.prior <- rep(10^8, nregres)
        ##Valores iniciales
init<-list()
init$mixture <- c(1,  1, rep(0, prior.eps$kmax - 1),
                  3.38, rep(0, prior.eps$kmax - 1), ## Medias iniciales
                  (0.57)^2, rep(0, prior.eps$kmax - 1)) ## Varianzas iniciales
init$beta <- c(0.0001,-0.0198,0,-0.1325 )
```

Ajuste del modelo

```
simCGM <- bayessurvreg1( Surv(df1$tiempo1+1,df1$delta) ~ df1$Sexo + df1$Etnia, data = df1,
    dir = "/Desktop/TESIS FINAL", nsimul = list(niter = 1000, nthin = 50, nburn = 500,nwrite
    = 50), prior = prior.eps, prior.beta = prior.beta, init = init, store = list(y = FALSE, r =
    FALSE, u = FALSE))
```

Leer la información en los directorios de las simulaciones

```
parCGM <- files2coda(dir = "/Desktop/TESIS FINAL")
summary(parCGM)
```