Distribución Weibull geométrica aplicada a datos de supervivencia de pacientes diagnosticados con cáncer colorrectal a través de modelos aditivos generalizados de localización, forma y escala.

Trabajo final presentado por:
Fernanda Isabel Arce Núñez.

Proyecto de titulación para optar al título de:
Ingeniero Estadístico.

Profesora guía:
Claudia Navarro Villarroel, Ph.D.

Valparaíso, Chile, Diciembre de 2018.

Agradecimientos

En primer lugar, doy infinitamente gracias a Dios, por haberme dado fuerza y valor para culminar esta etapa de mi vida.

Agradezco también la confianza y el apoyo brindado por parte de mis padres, que sin duda alguna a lo largo del trayecto de mi vida me han demostrado su amor, corrigiendo mis faltas y celebrando mis triunfos.

Agradezco a cada una de las personas que fueron parte de mi vida durante estos años lejos de casa, por acogerme, apoyarme y ser un pilar fundamental en mi desarrollo personal y profesional en estos últimos 6 años.

A los docentes del Instituto de Estadística, por su apoyo, comprensión y confianza, en especial a mi profesora guía, doctora Claudia Navarro Villarroel, quien en estos últimos años ha sido una fuente de motivación y de apoyo, además de ser un ejemplo de superación.

Finalmente, quiero agradecer y dedicar este proyecto de título a mi abuela, por ser una madre y también amiga; un ejemplo de vigor y humildad durante los años que pudo acompañarme, y que hoy mantengo en mi corazón como un hermoso recuerdo.

Resumen

El estudio del tiempo de vida de organismos o maquinarias ha sido de gran interés para los investigadores de las diferentes áreas de la ciencia, una de las metodologías propuestas recientemente es el uso de nuevas distribuciones como es el caso de la distribución Weibull Geométrica (WG). En este proyecto se presenta una aplicación de la distribución WG a datos de supervivencia de pacientes diagnosticados con cáncer colorrectal, y una comparación de esta distribución con la distribución de Weibull. La estimación de los parámetros se llevó a cabo utilizando el método de máxima verosimilitud y el algoritmo EM. Asimismo, se aplicó un modelo GAMLSS para estudiar el efecto de las variables edad, género, tipo del tumor, estadio del tumor y localización del tumor sobre el tiempo de supervivencia de los pacientes. De este modelo se obtuvo que todas las variables influyen en el comportamiento de cada uno de los parámetros de la distribución WG, además se observa que la distribución WG es más flexible, por consiguiente, al comparar las distribuciones anteriormente mencionadas, se observa que al utilizar la distribución WG se obtiene un mejor ajuste.

Abstract

The study of organisms and machinery lifespan has been of great interest for researchers from different scientific fields. One of the methodologies recently proposed is the use of new distributions such as the Weibull-geometric distribution (WG). This project presents an application of WG distribution to survival data of patients diagnosed with colorectal cancer, as well as a comparison of this distribution with the Weibull distribution. The estimation of the parameters was carried out using the Maximum likelihood estimation and the EM algorithm. In addition the GAMLSS model was applied to study he effect of different variables such as age, gender, tumor type, tumor stage, and tumor location on the survival time of patients. From this model, it was obtained that all the variables influence the behavior of each WG distribution parameter. Also, it is observed that WG distribution is more flexible, hence, when comparing the previously mentioned distributions, it is observed that the use of WG distribution provides a better fitting.

Abreviaturas y Notación

En esta sección se presentan algunas abreviaturas utilizadas en el desarrollo de este trabajo, el conocimiento de estos conceptos ayudará a optimizar la comprensión de este proyecto de titulación.

WG	Distribución Weibull Geométrica.
GAMLSS	Modelos aditivos generalizados de localización forma y escala (Generalised Additive Models for Location Scale and Shape).
EMV	Estimador Máximo Verosímil.
EG	Distribución Exponencial Geométrica.
GEG	Distribución Geométrica Exponencial Generalizada.
MSKCC	Centro de Cáncer Memorial Sloan Kettering (Memorial Sloan Kettering Cancer Center).
ASCO	American Society of Clinical Oncology.
AIC	Criterio de información de Akaike (Akaike Information Criterion).

Índice general

Índice de figuras

Índice de cuadros

Objetivos

Los objetivos de este trabajo de titulación se presentan a continuación.

Objetivo general.

El objetivo principal de este proyecto es aplicar la distribución Weibull geométrica y los modelos GAMLSS a datos de pacientes diagnosticados de cáncer colorrectal.

Objetivos específicos.

(i) Presentar la metodología para la aplicación de la distribución Weibull geométrica a datos de supervivencia.

(i) Presentar la metodología para la utilización de los modelos GAMLSS.

(i) Determinar si existe relación entre el tiempo de supervivencia de un paciente y algunas características clínicas del diagnostico, tales como, género, tipo de tumor, localización del tumor, estadio del tumor y edad.

Capítulo 1

Introducción

El estudio de la duración de la vida de los organismos, dispositivos, las estructuras o materiales es de gran importancia para muchas de las disciplinas aplicadas, tales como: la ingeniería, las ciencias médicas, las ciencias biológicas, seguros y finanzas, entre otras. Una de sus aplicaciones es el uso de modelos estadísticos, utilizando la distribución de fallas.

En ocasiones, las condiciones físicas del mecanismo de fallas pueden conducir a una distribución específica, pero, en la mayoría de los casos, la elección se realiza en base al mejor ajuste que presenten las observaciones reales de los tiempos hasta un acontecimiento esperado en una distribución. En estudios de confiabilidad, uno de los métodos más utilizados para disminuir los posibles candidatos de estas distribuciones está dado por la forma y la monotonicidad de la función de la tasa de falla, pues permite identificar algunas características importantes de las observaciones que conducen a conclusiones sobre el tiempo de vida.

La generalización de las distribuciones es una práctica antigua y ha sido considerada tan importante como otros problemas prácticos de la estadística. Desde entonces, han surgido muchas investigaciones con respecto al estudio de la duración de vida aplicados en distintas áreas, debido a la necesidad de encontrar respuestas, definir modelos con más flexibilidad, y explicar cómo surge el fenómeno de la vida en áreas como la física, la informática, la medicina, la ingeniería, entre otros. Algunas de las distribuciones aplicadas a datos de tiempo de vida o supervivencia son las distribuciones Exponencial, Weibull y Gamma, no obstante, son muy limitadas en sus características y no muestran amplia flexibilidad.

En la literatura, algunos autores han realizado estudios sobre la función de la tasa de fallas, por ejemplo, Davis (1952) estudia esta función en datos de fallas obtenidas de operaciones realizadas por máquinas y personas (en donde la función de tasa de fallas es de tipo creciente) comparando estos datos con las distribuciones de frecuencia que surgen de una teoría de falla exponencial o normal. Luego, Lomax (1954) realiza un estudio de la función de fallas de negocios, en donde es razonable esperar probabilidades condicionales monótonamente decrecientes, demostrando que, las probabilidades condicionales de falla para los datos aplicados en su estudio presentaron un buen ajuste en funciones exponenciales como hiperbólicas.

A raíz de esta necesidad es que autores como Marshall y Olkin (1997) estudian un nuevo método para agregar un parámetro a familias de distribuciones, que fue aplicado a familias Exponenciales y Weibull. Otros autores que estudiaron algunas modificaciones para la distribución exponencial fueron Adammis y Loukas (1998), que proponen la distribución Exponencial Geométrica (EG), aplicando esta nueva distribución a dos conjuntos de datos. El primero correspondía a la cantidad de fallas sucesivas del sistema de aire acondicionado de cada miembro de una flota de 13 aviones a reacción Boeing 720 estos datos fueron estudiados previamente por Proschan (1963), y un segundo conjunto de datos que correspondían a 109 observaciones sobre el período entre los sucesivos desastres de la minería del carbón; en ambos casos se atribuyen las fallas a errores humanos, y se demuestra que la distribución EG ajusta casi a la totalidad de los datos estudiados. Del mismo modo Gupta y Kundu (1999), plantean que las distribuciones de Gamma y Weibull son una de las más utilizadas para modelar datos de tiempo de vida y ambas distribuciones permiten variaciones en la tasa de riesgo, dependiendo del parámetro de forma, que da una ventaja adicional sobre la distribución exponencial, la cual posee solo una tasa de riesgo constante. Sin embargo, estas distribuciones presentan desventajas: Particularmente en la distribución Gamma, el cálculo de la función de distribución o la función de supervivencia no es fácil si el parámetro de forma no es un número entero, mientras que, en la distribución de Weibull los estimadores de máxima verosimilitud de los parámetros pueden no comportarse correctamente para todos los valores de estos.

Es así que, Gupta y Kundo proponen una nueva versión de la distribución exponencial, llamada exponencial generalizada de tres parámetros (ubicación, escala y forma) y comparan sus propiedades con las dos distribuciones antes mencionadas, distribución de Weibull y distribución de Gamma. En su estudio, demuestran que esta distribución tiene una tasa de peligro constante o decreciente según el parámetro de forma, además, presenta propiedades similares a la distribución de Gamma y Weibull, concluyendo que, el modelo exponencial generalizado se puede utilizar como una alternativa posible para analizar cualquier conjunto de datos sesgados.

El estudio de las modificaciones de la distribución de Weibull se convirtió en un tema de interés para algunos investigadores. Particularmente, se buscaba modelar tasas de fallas con curva en forma de bañera, ya que, la distribución de Weibull no proporciona un ajuste paramétrico razonable para el fenómeno de modelado con tasas de falla no monótonas, como las tasas de falla en forma de bañera invertidas, que son comunes en la fiabilidad y los estudios biológicos. Por ejemplo, estas curvas de tasas de falla se pueden observar en el curso de una enfermedad cuya mortalidad alcanza un punto máximo después de un período y luego disminuye gradualmente. Los modelos de vida útil que presentan tasas de fallas en forma de bañera invertidas son muy útiles en el análisis de supervivencia. Es por este motivo que Silva et al. (2010) proponen la distribución Geométrica exponencial generalizada (GEG), que se adapta a una tasa de falla con forma bañera, decreciente y creciente, dependiendo del valor de sus parámetros.

El uso de las distribuciones continuas para el estudio de tiempo de vida es muy común y generalmente uno de los métodos más utilizados en diferentes áreas de la ciencia o ingeniería. Sin embargo, las distribuciones discretas también son utilizadas en el estudio de las tasas de fallas. En la literatura se pueden encontrar estudios sobre el uso de las distribuciones binomial, Geométrica y Poisson, entre otras. Sin embargo, estas distribuciones no son muy flexibles para adaptarse a muchos tipos de datos discretos. Es por esto que algunos autores han propuesto métodos para construir nuevas distribuciones discretas con una mayor flexibilidad, con el fin de modelar estos datos con un mejor ajuste. Uno de los métodos más comunes para construir nuevas distribuciones discretas es discretizar las distribuciones continúas conocidas. De hecho, hay dos técnicas generales para formar nuevas distribuciones discretas utilizando una distribución continua de referencia. Un ejemplo de esto es la distribución normal discreta (Lisman y Van Zuylen, 1972). Además, otras distribuciones propuestas en la literatura son, la distribución de Weibull discreta aditiva (Bebbington, Lai, Wellington, y Zitikis, 2012), la distribución discreta exponencial Weibull (Nekoukhou y Bidram, 2015) y las distribuciones discretas beta exponencial (Nekoukhou, Alamatsaz, Bidram, y Aghajani, 2015). Otro método para construir una nueva distribución discreta es definir la versión ponderada de una distribución, método utilizado por Patil, Rao, y Ratnaparkhi (1986), quienes estudiaron algunos modelos generales que conducen a distribuciones ponderadas con funciones de ponderación no necesariamente delimitadas. Aplicando su estudio en análisis de datos relacionados con las poblaciones humanas y el manejo de la vida silvestre.

Una de las propuestas más recientes encontradas en la literatura es la generalización de la distribución Geométrica ponderada (GWG) (Najarzadegan y Alamatsaz, 2017). En este estudio se explica que la distribución propuesta puede verse como la generalización de la distribución Geométrica ponderada, la distribución exponencial generalizada discreta y la distribución Geométrica clásica. Este estudio es aplicado a las frecuencias de rango de modelado de grafemas en cuatro idiomas eslavos (ruso, eslovaco, esloveno y ucraniano), datos

que fueron utilizados anteriormente por Makcutek (2008), y comparan los resultados con modelos como la distribución exponencial Geométrica.

Otras propuestas de distribuciones discretas son algunas modificaciones de la distribución de Weibull. La primera fue estudiada por Nakagama y Osaki(1975) que fue llamada distribución Weibull discreta tipo 1 (DW(1)). Luego, Dattero (1984) propone una nueva versión de la distribución Weibull discreta, llamada, distribución Weibull discreta tipo 2 (DW(2)). Un último ejemplo, Bebbington y cols (2012) introducen la distribución aditiva discreta de Weibull que posee virtudes como, la capacidad de tratamiento matemático y la capacidad de producir funciones de tasa de riesgo en forma de bañera. Además, concluyen que la elección de un modelo discreto o continuo puede afectar las características de la curva de confiabilidad como resultado de diferentes estimaciones de parámetros y del comportamiento de la función de tasa de riesgo.

Existen diferentes métodos para el estudio de supervivencia utilizando algunas distribuciones conocidas como la distribución exponencial o Weibull. Algunos de estos métodos es el uso de modelos lineales. Para el caso particular de la familia exponencial, es recomendable utilizar los modelos lineales generalizados, modelos aditivos o modelos mixtos. Estudios como los realizados por los autores Barajas y Naranjo en el año 2007, muestran que el uso de estos modelos tienen buenos resultados a la hora de ser aplicados a tiempos de supervivencia de pacientes que padezcan alguna patología.

El modelado de datos de supervivencia censurados casi siempre se realiza mediante la regresión de riesgos proporcionales de Cox. Sin embargo, el uso de modelos paramétricos para tales datos puede tener algunas desventajas. Por ejemplo, los riesgos no proporcionales, una dificultad potencial con los modelos de Cox, a veces se pueden manejar de una manera sencilla, y la visualización de la función de riesgo es mucho más fácil.

En este proyeto se proponen la distribución Weibull geométrica para el análisis de tiempos de supervivencia en pacientes diagnosticados con cáncer colorrectal, mediante uso y aplicación de los modelos GAMLSS, los que serán descritos en el capítulo siguiente.

Capítulo 2

Metodología

2.1 Introducción.

Los análisis de supervivencia se han realizado bajo diferentes metodologías, desde descriptivos hasta modelos de regresión, pero no con las metodologías clásicas. En primer lugar, los tiempos de supervivencia son positivos, y para efectos de pronóstico, se ha demostrado que el uso de la regresión lineal simple no es la mejor opción (Clark et al, 2003). Es por esto que, se propone la distribución Weibull geométrica y los modelos aditivos generalizados para localización, forma y escala, también conocidos por la sigla de su nombre en ingles GAMLSS (Generalized Additive Models for Location Scale and Shape) para el estudio del tiempo de supervivencia en pacientes con cáncer colorrectal. En este capitulo se describen los modelos GAMLSS, se definen la distribución Weibull geométrica, sus propiedades y los métodos que serán utilizados para la estimación de los parámetros de esta distribución.

2.1.1 Modelo GAMLSS.

Los modelos aditivos generalizados para localización, forma y escala (GAMLSS) son modelos de regresión semiparamétricos. Son paramétricos, ya que requieren un supuesto de distribución paramétrica para la variable de respuesta, y "semi" en el sentido de que al modelar los parámetros de la distribución, como funciones de las variables explicativas, puede ser necesario utilizar funciones de suavizado no paramétricas (Rigby y Stasinopoulos, 2005).

Los modelos GAMLSS fueron introducidos por Rigby y Stasinopoulos (2005), como una nueva propuesta para solucionar algunas limitantes asociadas a los modelos lineales y modelos lineales generalizados.

Para aplicar los modelos lineales generalizados, es necesario que la variable respuesta siga una distribución perteneciente a la familia exponencial. En los modelos GAMLSS, este supuesto es eliminado y es reemplazado por el supuesto de una familia de distribución general, que incluye distribuciones continuas y discretas altamente sesgadas o kurtóticas. Los modelos GAMLSS permiten el modelado de la media (o ubicación), además de otros parámetros de la distribución de la variable respuesta como, funciones no paramétricas lineales o no lineales, paramétricas o suaves de variables explicativas o efectos aleatorios (Rigby & Stasinopoulos, 2005).

Por lo tanto, estos modelos son especialmente adecuados para modelar una variable de respuesta que no sigue una distribución de la familia exponencial (por ejemplo, leptocúrtica, platicúrtica, sesga o con presencia de sobredispersión en la variable respuesta), y aquellas variables que muestran heterogeneidad (Rigby & Stasinopoulos, 2005).

Los modelos GAMLSS se definen como modelos de regresión donde todos los parámetros de la distribución de la variable respuesta pueden ser modelados como funciones de factores evaluados (Rigby & Stasinopoulos, 2005). Además, son un modelo de regresión general que permite estudiar la flexibilidad en la elección de la variable respuesta, modelar los parámetros de la distribución, incluir interceptos aleatorios o ajustar modelos no lineales.

Estos modelos asumen que las variables respuesta y_i (con $i = 1, ..., n$) son independientes con función de densidad de probabilidad $f(y_i|\theta_i)$, donde $\theta_i = (\mu_i, \sigma_i, \nu_i, \tau_i)$, corresponde al vector de parámetros de la distribución. Rigby y Stasinopoulos (2005) definen la formulación original de un modelo GAMLSS de la siguiente manera.

Sea $y^T = (y_1, y_2, ..., y_n)$ el vector de longitud n de la variable de respuesta. También para $k = 1, 2, ..., p$. Sean $g_k(.)$ funciones conocidas de enlace monotónico que relacionan los parámetros de distribución con el predictor lineal η_k:

$$g_k(\theta_k) = \eta_k = X_k\beta_k + \sum_{j=1}^{J_k} Z_k\gamma_k, \tag{2.1}$$

donde, θ_k y η_k con vectores de longitud n, por ejemplo, $\theta_k^T = (\theta_{1k}, \theta_{2k}, ..., \theta_{nk})$, $\beta_k^T = (\beta_{1k}, \beta_{2k}, ..., \beta_{nk})$ es un vector de parámetros de tamaño J_k'; X_k es una matriz de diseño fija, conocida y de dimensión $n \times J_k'$, Z_{jk} es una matriz de diseño fija conocida de orden $n \times q_{jk}$ y γ_{jk} una variable aleatoria de dimensión q_{jk}. Entonces, se define la ecuación (2.1)

como el modelo GAMLSS.

Los vectores γ_{jk} para $j = 1, 2, ..., j_k$ pueden ser combinados en un solo vector γ_k con una única matriz de diseño Z_k; sin embargo, la formulación expuesta en (2.1) es adecuada para el ciclo de adaptación presente en el algoritmo RS (véase la sección 2.1.3) y permite que, las combinaciones de diferentes tipos de términos de efectos aleatorios aditivos se incorporen fácilmente en el modelo. Del mismo modo, si $k = 1, 2, ..., p$, $J_k = 0$, el modelo se reduce a un modelo totalmente paramétrico dado por,

$$g_k(\theta_k) = \eta_k = X_k\beta_k,$$

Si $Z_{jk} = I_n$, donde I_n es una matriz de identidad y $\gamma_{jk} = \mathbf{h}_{jk} = h_{jk}(x_{jk})$ para todas las combinaciones de j y k en el modelo (2.1), se obtiene que:

$$g_k(\theta_k) = \eta_k = X_k\beta_k + \sum_{j=1}^{j_k} h_{jk}(x_{jk}), \tag{2.2}$$

donde, x_{jk} para $j = 1, 2, ..., j_k$ y $k = 1, 2, ..., p$ son vectores de dimensión n. La función h_{jk} es una función desconocida de la variable exploratoria X_{jk} y $\mathbf{h}_{jk} = h_{jk}(x_{jk})$ es el vector que evalúa la función h_{jk} en x_{jk}. Se asume que que los vectores explicativos x_{jk} son conocidos. De este modo, se define el modelo (2.2) como un modelo GAMLSS semiparamétrico.

Los parámetros μ_i y σ_i por lo general se definen como los parámetros de escala y ubicación, mientras que los parámetros ν_i y τ_i corresponde a los parámetros de forma (asimetría y kurtosis) aunque el modelo puede aplicarse de manera más general a los parámetros de cualquier distribución de la población, y puede generalizarse a más de cuatro parámetros de distribución.

Para las familias de distribuciones de población con un máximo de dos parámetros de forma (ν y τ) se tiene que:

$$g_1(\mu) = \eta_1 = X_1\beta_1 + \sum_{j=1}^{J_1} Z_{j1}\gamma_{j1}$$

$$g_2(\sigma) = \eta_2 = X_2\beta_2 + \sum_{j=1}^{J_2} Z_{j2}\gamma_{j2}$$

$$g_3(\nu) = \eta_3 = X_3\beta_3 + \sum_{j=1}^{J_3} Z_{j3}\gamma_{j3}$$

$$g_4(\tau) = \eta_4 = X_4\beta_4 + \sum_{j=1}^{J_4} Z_{j4}\gamma_{j4}$$

Los modelos GAMLSS (2.1) son más generales que los modelos lineales, modelos lineales generalizados, modelos aditivos generalizados y modelos aditivos mixtos, ya que la distribución de la variable dependiente no se limita a la familia exponencial y todos los parámetros de la distribución son modelados (Rigby y Stasinopoulos, 2005).

2.1.2 Estimación del modelo.

En los modelos GAMLSS, la estimación se lleva a cabo maximizando la siguiente función de verosimilitud (2.3):

$$l_p = l - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k}\lambda_{jk}\gamma_{jk}^T\mathbf{G}_{jk}\gamma_{jk} \tag{2.3}$$

donde $l = \sum_{i=1}^{n}\log f(y_i|\theta_i)$ es un logaritmo de la función de verosimilitud de la variable respuesta y_i para cada uno de los parámetros de la distribución, λ_{jk} son los parámetros de penalización y $\mathbf{G}_{jk}$ es una matriz simétrica que depende los parámetros λ_{jk}. Esta estimación puede realizarse a través del algoritmo CG propuesto por Cole y Green, o utilizando el algoritmo RS propuesto por Rigby y Stasinopoulos en el año 2005. A continuación se expone una breve descripción de los algoritmos.

2.1.3 Algoritmos CG y RS.

Los algoritmos CG y RS fueron estudiados en el año 2005 por Rigby y Stasinopoulos en su articulo *"Generalized additive models for location, scale and shape"*. Estos autores describen ambos algoritmos de la siguiente manera.

Primero, el algoritmo CG (Cole y Green, 1992), utiliza las primeras y segundas derivadas cruzadas de la función de verosomilitud de los parámetros de la distribución ($\theta = \mu, \sigma, \nu, \tau$). Mientras que, el algoritmo RS que se utilizará, es una generalización del algoritmo usado por Rigby y Stasinopoulos (1996). Este algoritmo es más simple, dado que no utiliza las derivadas cruzadas, siendo más adecuado para ajustar los modelos aditivos de dispersión y media. La utilización del algoritmo CG, es ideal cuando se trabaja con las distribuciones binomial negativa, gamma, gaussiana inversa, logística y normal. Sin embargo, el algoritmo RS se ha utilizado con éxito para ajustar todas las distribuciones presente en el listado de distribuciones disponibles en el paquete **gamlss** del software estadístico R; aunque ocasionalmente puede ser lento para converger.

Es necesario tener en cuenta que, el algoritmo RS no es un caso especial del algoritmo CG y que el objetivo de los algoritmos es maximizar la función de verosimilitud penalizada l_p.

Las principales ventajas de los dos algoritmos son:

(a) Permiten diferentes diagnósticos de modelo para cada parámetro de la distribución.

(b) Presentan una fácil adición de distribuciones adicionales.

(c) Permiten una fácil adición de términos adicionales aditivos.

(d) Los valores iniciales se encuentran fácilmente, ya que solo requieren valores iniciales para θ.

Esencialmente, el algoritmo RS tiene un ciclo externo que maximiza la función de verosimilitud con respecto a β_k y γ_{jk} (para $j = 1, ..., J_k$) en el modelo sucesivamente para cada θ_k, y a su vez, para $k = 1, ..., p$. En cada cálculo del algoritmo se utilizan los valores actualizados de todas las cantidades. El algoritmo RS no es un caso especial del algoritmo CG porque en el algoritmo RS se actualiza la matriz de peso diagonal W_{kk} dentro del ajuste de cada parámetro θ_k, mientras que en el algoritmo CG todas las matrices de peso W_{ks} para $k = 1, 2, ..., p$ y $s = 1, 2, ..., p$ se evalúan después de ajustar todos θ_k para $k = 1, 2, ..., p$.

De este modo, Rigby y Stasinopoulos (2005) definen el algoritmo RS como sigue:

Paso 1: Inicio: Se inician los valores ajustados $\theta_k^{(1,1)}$ y los efectos aleatorios $\gamma_{jk}^{(1,1,1)}$, para $j = 1, ..., J_k$ y $k = 1, 2, ..., p$. Se evalúan los predictores lineales iniciales $\eta_k^{(1,1)} = g_k(\theta_k^{(1,1)})$, para $k = 1, 2..., p$.
Paso 2: Iniciar el ciclo externo $r = 1, 2, ...$ Hasta la convergencia. Para $k = 1, 2, ..., p$:

(a) Iniciar el ciclo interno $i = 1, 2, ...$ hasta la convergencia.

  (i) Evaluar la corriente $u_k^{(r,i)}, W_{kk}^{(r,i)}$ y $z_k^{(r,i)}$;

  (ii) Iniciar el ciclo de adaptación $m = 1, 2...$ hasta la convergencia;

  (iii) Regresar los residuos parciales actuales $\epsilon_{jk}^{(r,i,m)} = z_k^{(r,i)} - \sum_{j=1}^{JK} Z_{jk}\gamma_{jk}^{(r,i,m)}$ contra la matriz de diseño X_k, utilizando los pesos iterativos $W_{kk}^{(r,i)}$ para obtener las estimaciones actualizadas de los parámetros $\beta^{(r,i,m+1)}$;

  (iv) Para $j = 1, 2, ...J_k$ suaviza los residuos parciales $\epsilon_{jk}^{(r,i,m)} = z_k^{(r,i)} - X_k\beta_k^{(r,i,m+1)} - \sum_{t=1,t=j}^{JK} Z_{tk}\gamma_{tk}^{(r,i,c)}$, utilizando la matriz de reducción (suavizado) Sjk dada por la ecuación

$$S_{jk} = Z_{jk}(Z_{jk}^T W_{kk} Z_{jk} + G_{jk})^{-1} Z_{jk}^T W_{kk}$$

  para obtener el término del predictivo aditivo actualizado $Z_{jk}\gamma_{jk}^{(r,i,m+1)}$;

(v) Finalizar el ciclo de adaptación, en la convergencia de $\beta_{k^{(r,i,.)}}$ y $Z_{jk}\gamma_{jk}^{(r,i,.)}$ y establecer $\beta_k^{(r,i+1)} = \beta_k^{(r,i,.)}$ y $\gamma_{jk}^{(r,i+1)} = \gamma_{jk}^{(r,i,.)}$ para $j = 1, 2, ..., Jk$ y de otro modo actualizar m y continuar el ciclo de adaptación;

(vi) calcular el $\eta_k^{(r,i+1)}$ y $\theta_k^{(r,i+1)}$ actualizados;

(b) Terminar el ciclo interno en la convergencia de $\beta_k^{r,\cdot}$ y los términos predictivos aditivos $Z_{jk}\gamma_{jk}^{(r,.)}$ y establecer $\beta_k^{(r+1,1)} = \beta_k^{r,\cdot}$, $\gamma_{jk}^{(r+1,1)} = \gamma_{jk}^{(r+1,1)} = \gamma_{jk}^{(r,.)}$ para $j = 1, 2, ..., J_k$, $\eta_k^{(r+1,1)} = \eta_k^{(r,.)}$ y $\theta_k^{(r+1,1)} = \theta_k^{(r,.)}$; De lo contrario, actualizar i y continuar el ciclo interno.

Paso 3: Actualizar el valor de k.
Paso 4: Terminar el ciclo externo solo si el cambio en la probabilidad es suficientemente pequeño, de lo contrario actualice r y continúe el ciclo exterior.

En efecto, luego de estimar los valores para los parámetros del modelo y comprobar si estos son significativos, es necesario realizar un estudio de los residuos del modelo generado.

2.1.4 Residuos de un modelo GAMLSS.

Los residuos, y especialmente los gráficos de residuos, desempeñan un papel central en la verificación de los modelos estadísticos. En los modelos lineales, los residuos se distribuyen normalmente y se pueden estandarizar para tener varianzas iguales. En situaciones de regresión no normales, como la regresión logística o el análisis loglineal, los residuos, tal como se definen generalmente, pueden estar tan lejos de la normalidad y tener las mismas variaciones. En un modelo de regresión simple dado por, $y_i = \beta_0 + \beta_1 x_i + e_i$, se definen los residuos como la diferencia entre los valores observados y los ajustados $\hat{\varepsilon} = y_i - \hat{y}_i$ donde $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ para $i = 1, 2..., n$. algunas veces los $\hat{\varepsilon}$ son llamados residuos crudos para distinguirlos de los residuos estandarizados los cuales son definidos como $(y_i - \hat{y}_i)/\hat{\sigma}\sqrt{(1 - h_{ii})}$, donde h_{ii} son los valores de la diagonal de la matriz $\mathbf{H}$, (Álvarez, Palamarchuk y Riaño, 2017).

El problema con los residuos crudos es que son dificiles de generalizar a otras distribuciones diferentes a la distribución normal. Por otro lado, en el caso de los modelos lineales generalizados se utilizan los residuos de desvianza o los residuos de Pearson. Desafortunadamente, los residuos de desvianza no están bien definidos con múltiples parámetros para la distribución de la variable respuesta y, mientras que los residuos de Pearson pueden estar lejos de una distribución normal y tampoco son apropiados para datos altamente sesgados o kurtóticos. Es por esto que los modelos GAMLSS utilizan los *cuantiles aleatorios residuales* (Dunn y Smyth, 1996).

Los cuantiles aleatorios residuales fueron propuestos por Dunn y Smyth en el año 1996, y se definen de la siguiente manera.

Sea $F(y;\mu,\phi)$ la función de distribución acumulada de $P(\mu,\phi)$. Si F es continua, entonces la $F(y_i;\mu_i,\phi)$ se distribuye uniformemente en el intervalo de la unidad. En este caso, los cuantiles residuales se definen por:

$$r_{q,i} = \Phi^{-1}F(y_i;\hat{\mu}_i,\hat{\phi}),$$

donde $\Phi()$ es la función de distribución acumulada de la normal estándar. Además, la variabilidad de muestreo en $\hat{\mu}_i$ y $\hat{\phi}$, $r_{q,i}$ también sigue esta distribución. Esto implica que la distribución de $r_{q,i}$ converge hacía una distribución normal estándar si β y ϕ se estimaron consecuentemente. La definición anterior es un caso especial del residuo crudo de Cox y Snell del año 1968 (Dunn y Smyth, 1996).

Si F no es continua, se requiere una definición más general de los cuantiles residuales. Sea $\alpha_i = lim_{y\longrightarrow y_i}F(y;\hat{\mu}_i)$ y $b_i = F(y_i;\hat{\mu}_i,\hat{\phi})$. Se define el cuantil aleatorio residual para y_i por

$$r_{q,i} = \Phi^{-1}(u_i), \tag{2.4}$$

donde u_i es una variable aleatoria uniforme en el intervalo $(a_i, b_i]$. Nuevamente, $r_{q,i}$ son exactamente normal estándar, también se cumple la normalidad en la variabilidad de muestreo en $\hat{\mu}_i$ y $\hat{\phi}$.

Los modelos GAMLSS pueden ser aplicados en algunos software como Mathematical o R. En este proyecto se utilizará el software R para la aplicación de estos modelos. Con el objetivo de complementar el estudio de los modelos GAMLSS. En la siguiente sección podrá encontrar un breve resumen del paquete **gamlss** en el software R.

2.1.5 Modelos GAMLSS en software R.

La Tabla 2.1 muestra una variedad de familias de uno, dos, tres y cuatro parámetros de distribuciones implementadas en el paquete ***gamlss*** en el software estadítico R. Los libros de (1994, 1995); Johnson, Kotz y Kemp (2005) son referencias para la mayoría de las distribuciones en la Tabla 2.1 (Pérez, 2016).

Distribución	Nombre en el paquete de R	μ	σ	ν	τ
Beta	BE()	logit	logit	-	-
Binomial	BI()	logit	-	-	-
Exponencial	EXP()	logit	-	-	-
Gamma	GA()	logit	logit	-	-
Gamma generalizada	GG()	logit	logit	Identidad	-
Weibull	WEI()	logit	logit	-	-
Weibull Geomeétrica	WG()	log	log	logit	-

Tabla 2.1: Distribuciones disponibles en el paquete gamlss del software estadístico R.

Hay que tener en cuenta que el listado de la Tabla 2.1 presenta solo algunas de las distribuciones disponibles en el paquete, dado que son cerca de 50 distribuciones disponibles para trabajar en el software estadístico R. En este proyecto se realizará un estudio de supervivencia utilizando la distribución Weibull geométrica, disponible en el paquete ***gamlss***. En este proyecto se utilizará la distribución Weibull geométrica, la que es presentada en la sección 2.2.

2.2 Definición de la distribución Weibull Geométrica.

2.2.1 Distribución Weibull.

La distribución de Weibull fue establecida por el físico suizo llamado Waloddi Weibull. Esta distribución es conocida por su uso en el modelamiento de variables de tiempo de vida o tiempo de falla (Planco y Horna, 2015).

Se dice que una variable aleatoria y de tipo continuo tiene distribución Weibull de parámetros α y β con función de densidad:

$$g(y, \beta, \alpha) = \alpha\beta^{\alpha} y^{\alpha-1} e^{-(\beta y)^{\alpha}}, \tag{2.5}$$

donde $y \in \mathbb{R}_+$, $\beta \in \mathbb{R}_+$ y $\alpha \in \mathbb{R}_+$.

Se llama tiempo de vida al tiempo transcurrido hasta la ocurrencia de un suceso de interés.

2.2.2 Distribución Geométrica.

En un experimento binomial se tiene una serie de eventos idénticos e independientes, en donde cada uno origina un éxito o un fracaso. Si interesa el número z de pruebas hasta la observación del primer éxito, entonces z sigue una distribución Geométrica con función

de probabilidad:

$$P(z;p) = (1-p)p^{z-1} \tag{2.6}$$

donde $z \in \mathbb{N}$ y $p \in (0,1)$.

Notar que las pruebas pueden ocurrir indefinidamente, y que, z es un ejemplo de variable aleatoria discreta que puede tomar un número infinito, pero contable, de valores (Badii y Castillo, 2009).

2.2.3 Distribución Weibull Geométrica.

La distribución Weibull geométrica fue estudiada y propuesta por Barreto-souza, Lemos y Cordeiro en el año 2011. estos autores definen la distribución y sus propiedades del modo siguiente:

Sea $\{y_i\}_{(i=1)}^{z}$ variables independientes identicamente distribuidas Weibull con densidad dada en (2.5) Los parámetros β y α corresponden al parámetro de escala y al parámetro de forma, respectivamente. Además, z es una variable que sigue una distribución Geométrica con función de probabilidad (2.6).
Sea $x = min(\{y_i\}_{(i=1)}^{z})$, entonces la función de densidad será:

$$f(x;p,\beta,\alpha) = \alpha\beta^{\alpha}(1-p)x^{\alpha-1}e^{-(\beta x)^{\alpha}}\{1-pe^{-(\beta x)^{\alpha}}\}^{-2}, x > 0, \tag{2.7}$$

Esta última expresión define a la distribución Weibull Geométrica con notación $WG(p,\beta,\alpha)$.

Para entender la interpretación de la distribución WG, se plantea el siguiente ejemplo: Se considera la situación de un paciente diagnosticado con una enfermedad terminal (cáncer), que debe ser sometido a una serie de sesiones como tratamiento para combatir dicha patología. La recuperación o muerte del paciente ocurre luego de una cantidad finita de sesiones o terapias, z. Estas terapias serán consideras de un solo tipo, por ejemplo, quimioterapias. Las observaciones de y representan el tiempo transcurrido hasta la recuperación o muerte del paciente.

Es de gran importancia tener en cuenta algunas de las propiedades de la distribución Weibull geométrica, es por esto que, a continuación se describe la distribución, sus propiedades y su comportamiento según el valor de sus parámetros.

2.2.4 Comportamiento de la distribución Weibull geométrica.

Ante algunos cambios en el valor de los parámetros esta distribución presenta las siguientes propiedades (Barreto-souza, Lemos y Cordeiro, 2011), si p toma los valores:

- $p = 0$, la distribución WG, es simplemente la distribución Weibull.

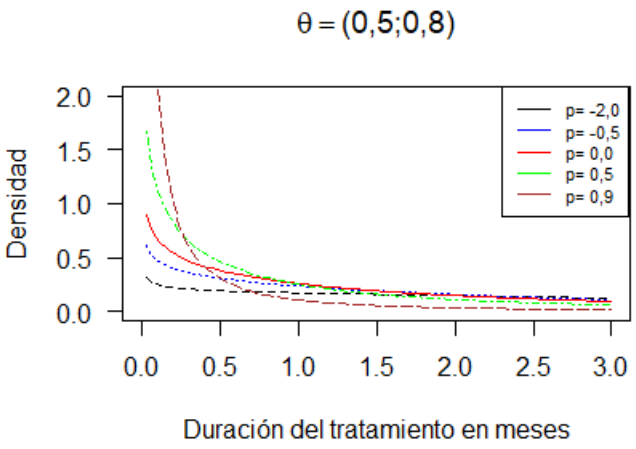
- $p \to -1$ la distribución tiende a una distribución degenerada en cero, por lo tanto, el parámetro p se define también como un parámetro de concentración

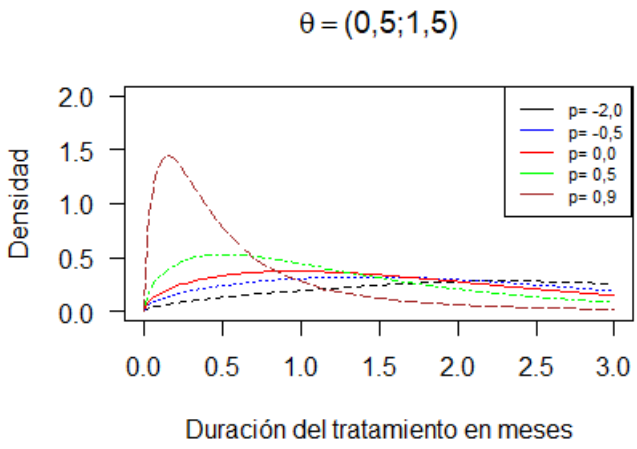
- $\alpha = 1$ y $0 < p < 1$, el modelo corresponde a la distribución Exponencial Geométrica (EG).

- $\alpha = 1$ y $p < 1$ se obtiene la distribución Exponencial Geométrica Extendida (EEG).

- Para $-1 \leq p < 1$, la densidad de WG es unimodal si $\alpha > 1$, y es estrictamente decreciente si $\alpha = 1$.

- Para $p < -1$, la densidad de la distribución WG puede ser unimodal. Por ejemplo, la distribución exponencial geométrica extendida. ($\alpha = 1$) es unimodal si $p < -1$.





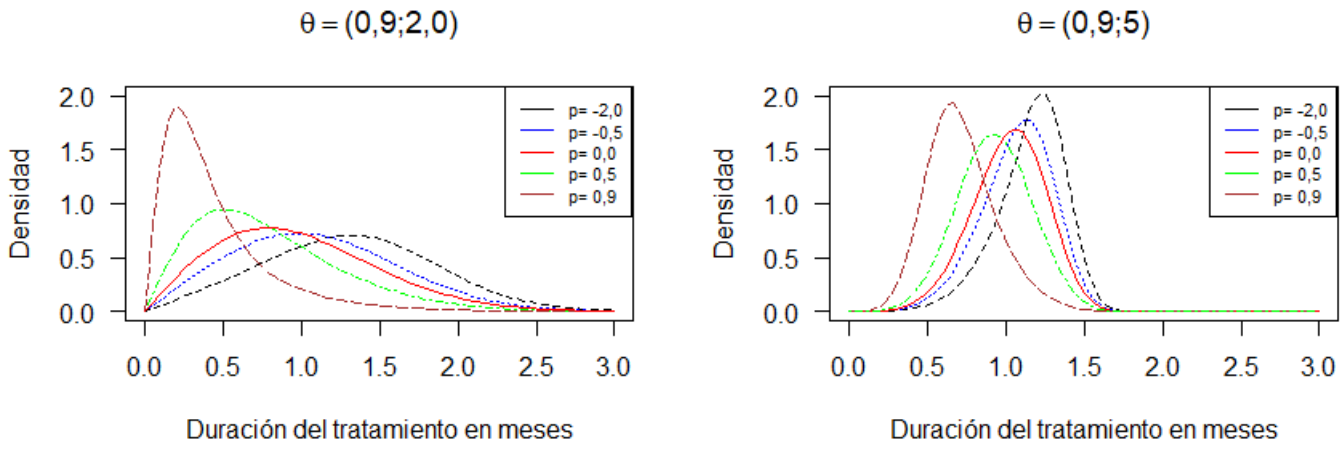


Figura 2.1: Comportamiento de la función de densidad para los diferentes valores del párametros p, con $\theta = (\beta, \alpha)$

Desde la Figura 2.1, se observa que, a medida que los valores de los parámetros cambian, la forma de la distribución WG se vuelve más flexible, este comportamiento puede ser de gran utilidad cuando se trabaja con datos altamente curtóticos.

- Para $\mid p \mid \leq 1$, es fácil probar que la densidad de la distribución WG se puede escribir como una mezcla infinita de distribuciones Weibull con el mismo parámetro de forma α. Si $\mid z \mid < 1$ y $k > 0$, tenemos la representación en serie:

$$(1-z)^{-k} = \sum_{j=0}^{\infty} \frac{\Gamma(k+j)}{\Gamma(k)j!} z^j \tag{2.8}$$

- Si $\mid p \mid \leq 1$, ampliamos $\{1 - pe^{-(\beta x)^\alpha}\}^{-2}$ y luego la función de densidad (2.3) se puede reducir a:

$$f(x; p, \beta, \alpha) = \alpha\beta^\alpha (1-p) x^{\alpha-1} e^{-(\beta x)^\alpha} \sum_{j=0}^{\infty} (j+1) p^j e^{-j(\beta x)^\alpha} \tag{2.9}$$

Usando la densidad de Weibull dada en (2.1), se obtiene:

$$f(x; p, \beta, \alpha) = (1-p) \sum_{j=0}^{\infty} p^j g(x; \beta(j+1)^{1/\alpha}, \alpha). \tag{2.10}$$

Algunas propiedades estadísticas (función de distribución acumulada, percentiles, función generadora de momentos, momentos factoriales, entre otros) de la distribución WG para $\mid p \mid \leq 1$ se puede obtener a partir de la ultima ecuación.

2.2.5 Propiedades de la distribución Weibull Geométrica.

Para una variable aleatoria X que sigue una distribución Weibull Geométrica $WG(p, \alpha, \beta)$, se definen las siguientes funciones y propiedades (Barreto-Souza, Lemos y Cordeiro, 2011):

Moda: La moda $x_0 = \beta^{-1} u^{1/\alpha}$ se obtiene al resolver la ecuación no lineal:

$$u + p^{-1} e^u (u + \frac{1-\alpha}{\alpha}) = \frac{1-\alpha}{\alpha} \tag{2.11}$$

Función de distribución acumulada

$$F(x) = \frac{1 - e^{-(\beta x)^\alpha}}{1 - p e^{-(\beta x)^\alpha}}, \quad x > 0 \tag{2.12}$$

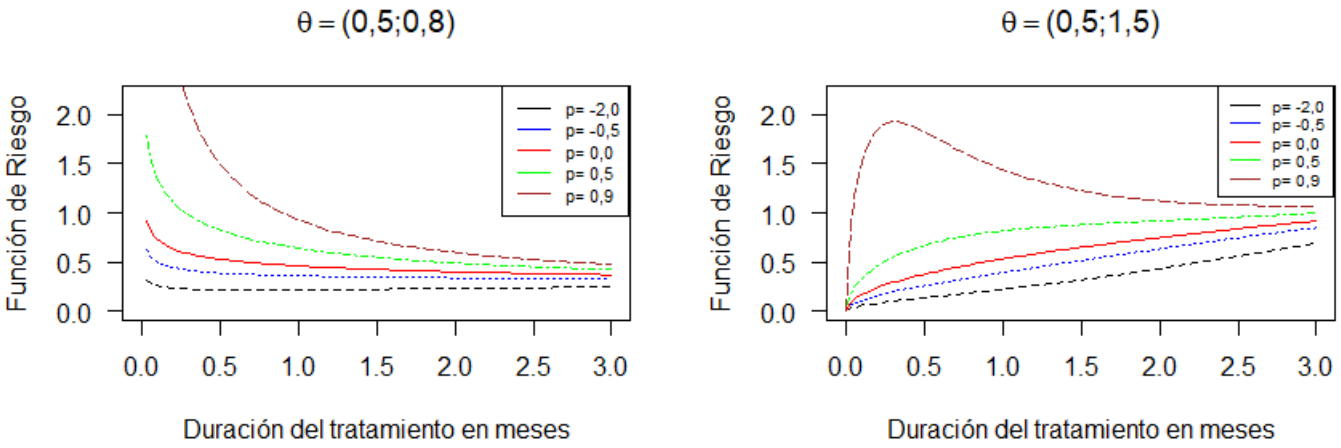
Función de supervivencia

$$S(x) = \frac{(1-p) e^{-(\beta x)^\alpha}}{1 - p e^{-(\beta x)^\alpha}}, \quad x > 0 \tag{2.13}$$

Función de riesgo

$$h(x) = \alpha \beta^\alpha x^{\alpha-1} \{1 - p e^{-(\beta x)^\alpha}\}^{-1}, \quad x > 0, \tag{2.14}$$

La función de riesgo está disminuyendo para $0 < \alpha \leq 1$ y $-1 \leq p < 1$. Sin embargo, para otros valores de parámetros, puede tomar diferentes formas.



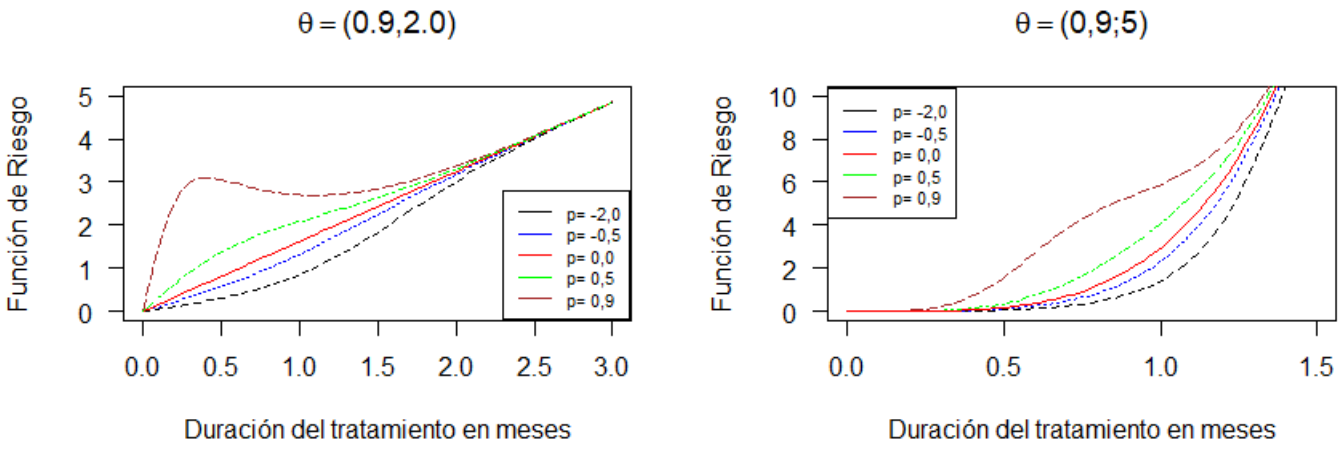


Figura 2.2: Comportamiento de la función de riesgo para los diferentes valores del parámetro p, con $\theta = (\beta, \alpha)$

La Figura 2.2 muestra el comportamiento de la función de riesgo cuando varían los valores de los parámetros de la distribución WG. Observar que, para los diferentes valores de p la forma de la función de riesgo flexible al igual que la función de densidad.

2.2.6 Cuantiles y momentos.

Análogamente, se definen los cuantiles y momentos de la distribución WG como sigue (Barreto-Souza, Lemos y Cordeiro,2011):

Cuantil
El cuantil $\gamma(x_\gamma)$ se define por:

$$x_\gamma = \beta^{-1}\{log(\frac{1-p\gamma}{1-\gamma})\}^{1/\alpha}. \tag{2.15}$$

En particular, la mediana es simplemente $x_{0,5} = \beta^{-1}\log(1-p)^{1/\alpha}$

Momentos
El momento *r-ésimo* está dado por:

$$E(x^r) = (1-p)\beta^{-r}\Gamma(r/\alpha+1)\Phi(p, r/r, 1), \tag{2.16}$$

donde Φ es una función trascendente de la función de Lerch's y definida por:

$$\Phi(z,s,a) = \{\Gamma(s)\}^{-1}\int_0^\infty t^{s-1}e^{-at}(1-ze^{-t})^{-1}dt,$$

para $z < 1$ y $s > 0$.

Para valores de $\mid p \mid \leq 1$ la expresión para el momento *r-ésimo* se reduce a:

$$E(X_r) = \frac{(1-p)\Gamma(r/\alpha+1)}{p\beta^r}L(p; r/\alpha), \tag{2.17}$$

donde $L(p;a) = \sum_{j=1}^{\infty} p^j j^{-a}$ es la función de polilogaritmo de Euler. Cuando $p \to -1$ los coeficientes de asimetría y curtosis tienden a cero, ya que la distribución WG converge a una distribución degenerada en cero.

Estadísticos de orden.

Dada una muestra x_n, las observaciones de la muestra ordenadas de menor a mayor serán:

$$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$$

Cuando la distribución de la variable aleatoria es continua, la probabilidad de coincidencia en valores es 0 y con probabilidad 1 se tiene que:

$$x_{(1)} < ... < x_{(n)}$$

Los estadísticos de orden aparecen en muchas áreas de la teoría y la práctica estadística, pueden utilizarse para definir, por ejemplo, la mediana o los cuartiles. (Moreno, 2016).

Barreto-souza, Lemos y Cordeiro (2011) definen la densidad del i-ésimo estadístico de orden del siguiente modo:

Sean $x_1, x_2, ..., x_n$ las variables aleatorias de distribución $WG(p, \beta, \alpha)$. La densidad del i-ésimo estadístico de orden llamado $x_{i:n}$ está dado por:

$$f_{i:n}(x) = \frac{\alpha\beta^{\alpha}(1-p)^{n-i+1}}{B(1, n-i+1)} x^{\alpha-1} e^{-(n-i+1)(\beta x)^{\alpha}} \frac{\{1-e^{-(\beta x)^{\alpha}}\}^{i-1}}{\{1-pe^{-(\beta x)^{\alpha}}\}^{n+1}}, \tag{2.18}$$

para $x > 0$, donde,

$$B(a, b) = \int_0^{\infty} w^{a-1}(1-w)^{b-1} dw, \tag{2.19}$$

es la función Beta. Luego la función de probabilidad del i-ésimo estadístico de orden Weibull con parámetros β y α

$$g_{i:n}(x) = \frac{\alpha\beta^{\alpha}}{B(q, n-i+1)} x^{\alpha-1} e^{-(n-i+1)(\beta x)^{\alpha}} \{q - e^{-(\beta x)^{\alpha}}\}^{i-1} \tag{2.20}$$

La ecuación (2.18) puede ser planteada en términos de $g_{i:n}(x)$

$$f_{i:n} = (1-p)^{n-i+1}\{1 - pe^{-(\beta x)^{\alpha}}\} g_{i:n}(x) \tag{2.21}$$

Se puede reescribir la función de densidad como una mezcla de densidades de los estadísticos de orden. Usando la ecuación (2.8) en (2.14) se obtiene:

$$f_{i:n}(x) = (1-p)^{n-i+1} \frac{n!(n+j-i)!}{(n+j)!(n-i)!} \sum_{j=0}^{\infty} \binom{n+j}{n} p^j g_{i:n+j}(x) \tag{2.22}$$

Algunas propiedades de los estadísticos de orden de la distribución Weibull Geométrica se pueden obtener utilizando la ecuación (2.18) para valores de $\mid p \mid \leq 1$.

2.2.7 Entropía de Rényi y Shannon

La entropía cuantifican la dispersión, incertidumbre, aleatoriedad o cantidad de información que contiene una variable aleatoria (Lessa, 2014). De este modo, se describe la entropía de Rényi y Shannon, propuestas por Barreto-souza, Lemos y Cordeiro en el año 2011.

Entropía de Rényi.

Este método define la entropía de una variable aleatoria como:

$$I_R(\gamma) = 1/(1-\gamma)\log\textstyle\int_{\mathbb{R}} f^\gamma(x)dx, \text{ para } \gamma > 0 \text{ y } \gamma \neq 1.$$

En este caso se puede obtener la siguiente expresión:

$$\int_0^\infty f^\gamma(x;p,\beta,\alpha)dx = \frac{[\alpha\beta(1-p)]^\gamma}{\Gamma(2\gamma)}\sum_{j=0}^{\infty} p^j \frac{\Gamma(2\gamma+j)}{j!}\int_0^\infty x^{(\alpha-1)^\gamma} e^{-(\gamma+j)(\beta x)^\alpha}dx \tag{2.23}$$

si $(\alpha-1)(\gamma-1) \geq 0$, la expresión se reduce a:

$$\int_0^\infty f^\gamma(x;p,\beta,\alpha)dx = \frac{\Gamma(\alpha)[\alpha(1-p)]^\gamma}{\beta^{\alpha(1-\gamma)}\Gamma(2\gamma)}\sum_{j=0}^{\infty} p^j \frac{\Gamma(2\gamma+j)}{j!(\alpha+j)} E(Y_j^{(\alpha-1)(\gamma-1)}) \tag{2.24}$$

donde Y_j sigue una distribución gamma con el parámetro de escala $(\gamma+j)^{1/\alpha}$ y el parámetro de forma α. Por lo tanto:

$$I_R(\gamma) = \frac{1}{1-\gamma}\log\frac{[\alpha(1-p)]^\gamma\Gamma(\gamma(\alpha-1)+1))}{\beta^{1-\gamma}\Gamma(2\gamma)}\sum_{j=0}^{\infty}\frac{p^j\Gamma(2\gamma+j)}{j!(\alpha+j)^{(\alpha-1)(\gamma-1)/\alpha+1}} \tag{2.25}$$

Entropía de Shannon.

La entropía de Shannon está definida por $E[-\log[f(X)]]$. Este es un caso especial derivado de $\lim_{\gamma\to 1} I_R(\gamma)$. Por lo tanto,

$$E[-\log f(X)] = -\log[\alpha\beta^\alpha(1-p)] - (\alpha-1)E[log(X)] + \beta^\alpha E(X^\alpha) + 2E\log[1-pe^{-(\beta X)^\alpha}] \tag{2.26}$$

Además, se puede probar que

$$E[\log(X)] = \psi(1)/\alpha - \log\beta - \frac{1-p}{\alpha}\sum_{j=1}^{\infty} p^j\log(j+1), \tag{2.27}$$

$$E\log[1-pe^{-(\beta X)^\alpha}] = 1 - \frac{log(1-p)}{p} \tag{2.28}$$

$$E(X^\alpha) = \frac{(1-p)}{p\beta\alpha}\log(1-p), \tag{2.29}$$

Por lo tanto, la entropía de Shannon se reduce a

$$E[-\log f(X)] = 2-\log[\alpha\beta(1-p)]-\frac{\alpha-1}{\alpha}[\psi(1)-(1-p)\sum_{j=1}^{\infty}p^j\log(j+1)]+(1-\frac{3}{p})\log(1-p) \tag{2.30}$$

Luego definir algunas de las propiedades de la distribución Weibull geométrica, se procede a estudiar los métodos para la estimación de los parámetros de la distribución WG. En este caso, se utilizarán el método del estimador máximo verosímil y el algoritmo EM.

2.2.8 Estimación de parámetros.

Sea $x = (x_1, ..., x_n)$ una muestra aleatoria de la distribución Weibull geométrica con un vector de parámetros desconocido $\theta = (p, \beta, \alpha)$. La función de log verosimilitud $\ell = \ell(\theta; x)$ para θ es:

$$\ell = n[\log\alpha + \alpha\log\beta + \log(1-p) + (\alpha-1)\sum_{i=}^{n}\log(x_i) - \sum_{i=1}^{n}(\beta x_i)^{\alpha} - 2\sum_{i=1}^{n} log[1-pe^{-(\beta x_i)^{\alpha}}]. \tag{2.31}$$

Los componentes de la función Score $U(\theta) = (\partial\ell/\partial p, \partial\ell/\partial\beta, \partial\ell/\partial\alpha)^T$ son:
$\frac{\partial\ell}{\partial p} = -n(1-p)^{-1} + 2\sum_{i=1}^{n} e^{-(\beta x_i)^{\alpha}}[1-pe^{-(\beta x_i)^{\alpha}}]^{-1}$,
$\frac{\partial\ell}{\partial\beta} = -n\alpha\beta^{-1} - \alpha\beta^{\alpha-1}\sum_{i=1}^{n} x_i^{\alpha}\{1-2pe^{-(\beta x_i)^{\alpha}}[1-pe^{-(\beta x_i)^{\alpha}}]^{-1}\}$;
$\frac{\partial\ell}{\partial\alpha} = -n\alpha^{-1} + \sum_{i=1}^{n}\log(\beta x_i) - \sum_{i=1}^{n}(\beta x_i)^{\alpha}\log(\beta x_i)\{1+2pe^{-((\beta x_i)^{\alpha})}[1-pe^{-(\beta x_i)^{\alpha}}]^{-1}\}$,

La estimación de máxima verosimilitud $\hat{\theta}$ de θ se puede determinar numéricamente a partir de las ecuaciones no lineales $U(\theta) = 0$.

2.2.9 Algoritmo Esperanza-Maximización.

El algoritmo Esperanza-Maximización, también conocido como algoritmo EM, es una técnica de optimización que se utiliza para encontrar estimadores máximo verosimil en modelos probabilísticos que dependen de variables no observables. (Alcazar, 2006).
Este algoritmo alterna pasos de esperanza (paso E) y un paso de maximización (paso M),

- Paso E: se calcula la esperanza de la verosimilitud mediante la incorporación de variables latentes como si fueran observables,
- Paso M: se calculan estimadores de máxima verosimilitud de los parámetros mediante la maximización de la verosimilitud esperada del paso E.

Los valores que se obtienen en el paso M se utilizan para comenzar el siguiente paso E, y así sucesivamente.

Se propone este método para el cálculo de los estimadores de la distribución WG, para esto se considera que $p \in (0,1)$, y se define una distribución hipotética de datos completos con función de densidad:

$$f(x,z;\theta) = \alpha\beta^{\alpha}(1-p)zx^{\alpha-1}p^{z-1}e^{-z(\beta x)^{\alpha}}, \tag{2.32}$$

para $x \in \mathbb{R}_+$,$\beta \in \mathbb{R}_+$, $\alpha \in \mathbb{R}_+$, $p \in (0,1)$ y $z \in \mathbb{N}$.

Según este algoritmo, el "paso E" del algoritmo EM requiere la esperanza condicional de $(Z|X;\theta^{(r)})$, donde $\theta^{(r)} = (p^{(r)}, \beta^{(r)}, \alpha^{(r)})$ es la estimación actual de θ. Desde,

$$P(z|x;\theta) = zp^{z-1}e^{-(z-1)(\beta x)^{\alpha}} \times \{1 - pe{-}(\beta x)^{\alpha}\}^2$$

para $z \in \mathbb{N}$, se tiene:

$$E(Z|X;\theta) = \{1 + pe^{-(\beta x)^{\alpha}}\}\{1 - pe^{-(\beta x)^{\alpha}}\}^{-1}.$$

El ciclo EM se completa con el "paso M" utilizando la estimación de máxima verosimilitud sobre θ, reemplazando las Z faltantes por sus esperanzas condicionales dadas anteriormente.

Por lo tanto, una iteración EM será:

$$p^{(r+1)} = 1 - \frac{n}{\sum_{i=0}^{n} w_i^{(r)}}, \tag{2.33}$$

$$\beta^{r+1} = n\{\sum_{i=1}^{n} x_i^{\alpha^{(r+1)}} w_i^{(r)}\}^{-1/\alpha^{(r+1)}} \tag{2.34}$$

donde $\alpha^{(r+1)}$ es la solución de la ecuación no lineal

$$\frac{n}{\alpha^{(r+1)}} + \sum_{i=1}^{n} \log x_i - n\frac{\sum_{i=1}^{n} w_i^{(r)} x_i^{\alpha^{(r+1)}} log x_i}{\sum_{i=1}^{n} w_i^{(r)} x_i^{\alpha^{(r+1)}}} = 0 \tag{2.35}$$

donde

$$w_i^{(r)} = \frac{1 + p^{(r)}e^{-(\beta^{(r)}x_i)^{\alpha^{(r)}}}}{1 - p^{(r)}e^{-\beta^{(r)}x_i^{\alpha^{(r)}}}}. \tag{2.36}$$

2.2.10 Inferencia.

Para la estimación de intervalos y pruebas de hipótesis para los parámetros de la distribución Weibull Geométrica, se obtiene la matriz de información $J_n = J_n(\theta)$

$$J_n = \begin{pmatrix} J_{pp} & J_{p\beta} & J_{p\alpha} \\ J_{p\beta} & J_{\beta\beta} & J_{\beta\alpha} \\ J_{p\alpha} & J_{\beta\alpha} & J_{\alpha\alpha} \end{pmatrix}$$

donde

$$-J_{pp} = \frac{\partial^2 \ell}{\partial p^2} = 2\sum_{i=1}^{n} T^{(i)}_{0,0,2,2} - n(1-p)^{-2},$$

$$-J_{p\beta} = \frac{\partial^2 \ell}{\partial p \partial \beta} = 2\alpha\beta^{\alpha-1}\sum_{i=1}^{n}(pT^{(i)}_{0,0,2,2} + T^{(i)}_{1,0,1,1}),$$

$$-J_{p\alpha} = \frac{\partial^2 \ell}{\partial p \partial \alpha} = 2\beta^{\alpha}\sum_{i=1}^{n}(pT^{(i)}_{1,1,2,2} + T^{(i)}_{1,1,1,1}),$$

$$-J_{\beta\beta} = \frac{\partial^2 \ell}{\partial \beta^2} = -n\alpha\beta^{-2} - \alpha(\alpha-1)\beta^{\alpha-2}\sum_{i=1}^{n}(T^{(i)}_{1,0,0,0} + 2pT^{(i)}_{1,0,1,1}) + 2\alpha^2\beta^{2\alpha-2}p\sum_{i=1}^{n}(pT^{(i)}_{2,0,2,2} + T^{(i)}_{2,0,1,1}),$$

$$-J_{\beta\alpha} = \frac{\partial^2 \ell}{\partial \alpha \partial \beta} = n\beta^{-1} - \beta^{\alpha-1}\sum_{i=1}^{n}(\alpha T^{(i)}_{1,1,0,0} + T^{(i)}_{1,0,0,0})(1 + 2pT^{(i)}_{0,0,1,1}) + 2p\alpha\beta^{2\alpha-1}\sum_{i=1}^{n}(pT^{(i)}_{2,1,2,2}T^{(i)}_{2,1,1,1}),$$

$$-J_{\alpha\alpha} = \frac{\partial^2 \ell}{\partial \alpha^2} = -n\alpha^{-2} + \sum_{i=1}^{n}(2p^2\beta^{2\alpha}T^{(i)}_{2,2,2,2} + 2p\beta^{2\alpha}T^{(i)}_{2,2,1,1} - \beta^{\alpha}T^{(i)}_{1,2,0,0} - 2p\beta^{\alpha}T^{(i)}_{1,2,1,1}),$$

donde

$$T^{(i)}_{j,k,l,m} = T^{(i)}_{j,k,l,m}(x_i, \theta) = x_i^{\alpha j}\{\log(\beta x_i)\}^k\{-pe^{-(\beta x_i)^{\alpha}}\}^{-1}e^{-m(\beta x)^{\alpha}},$$

para $j, k, l, m \in \{0, 1, 2\}$ e $i = 1, ..., n$. En condiciones que se cumplen para $\theta = (p, \beta, \alpha)$, la distribución asintótica de $\sqrt{n}(\hat{\theta} - \theta)$ es multivariante normal $N_3(0, K(\theta)^{-1})$, donde $K(\theta) = \lim_{n\to\infty} n^{-1}J_n(\theta)$ es la matriz de información de la unidad. Este comportamiento

asintótico sigue siendo válido si $K(\theta)$ es reemplazado por la matriz de información evaluada en $\hat{\theta}$.
Se puede usar la distribución aproximada multivariante normal $N_3(0, J_n(\hat{\theta})^{-1})$ de $\hat{\theta}$ para construir regiones de confianza aproximadas para algunos parámetros y para las funciones de peligro y supervivencia.
De este modo, un Intervalo Asintótico de Confianza (IAC) $100(1-\gamma)\,\%$ para cada parámetro θr viene dado por:

$$IAC_r = (\hat{\theta}_r - z_{\gamma/2}\sqrt{\hat{J}^{\theta_r\theta_r}}, \hat{\theta}_r + z_{\gamma/2}\sqrt{\hat{J}^{\theta_r\theta_r}}), \tag{2.37}$$

donde $\hat{J}^{\theta_r\theta_r}$ representa el elemento diagonal (r, r) de $J_n(\hat{\theta})^{-1}$ para $r = 1, 2, 3$ y $z_{\gamma/2}$ es el cuantil $1 - \gamma/2$ de la distribución normal estándar.
El estadístico Razón de Verosimilitud (RV) es útil para comparar la distribución de WG con algunos de sus submodelos especiales. Considerar la partición $\theta = (\theta_1^T, \theta_2^T)^T$ de la distribución WG, donde θ_1 es un subconjunto de parámetros de interés y θ_2 es un subconjunto de los parámetros restantes. Dadas las hipotesis nula y alternativa,

$$H_0 : \theta_1 = \theta_1^{(0)} \qquad \text{v/s} \qquad H_1 : \theta_1 \neq \theta_1^{(0)}$$

El estadístico de razón de verosimilitud para probar la hipótesis nula está dada por

$$w = 2\{\ell(\hat{\theta}) - \ell(\tilde{\theta})\},$$

donde $\tilde{\theta}$ y $\hat{\theta}$ son los estimadores de máxima verosimilitud bajo las hipótesis nula y alternativa, respectivamente. El estadístico w es asintóticamente (como $n \rightarrowtail \infty$) distribuido como χ_k^2, donde k es la dimensión del subconjunto θ_1 de interés.

Luego de describir la metodología que se utilizará en este proyecto. Se procede a, explicar los métodos utilizados para la selección de los datos clínicos de los pacientes con cáncer, describir el conjunto de datos seleccionados para la aplicación y definir las variables utilizadas presentes en el conjunto de datos de pacientes con cáncer colorrectal.

Capítulo 3

Aplicación

3.1 Introducción.

Según los datos entregados por el instituto nacional del cáncer de Estados Unidos, para el año 2018 se estima que se diagnosticarán 1.735.350 nuevos casos de cáncer en el país y que 609.640 personas morirán a causa de esta enfermedad. Además, se asume que el número de nuevos pacientes diagnosticados de cáncer (incidencia del cáncer) es de 439 por cada 100.000 pacientes por año, mientras que, el número de muertes por cáncer es de 163 por cada 100.000 pacientes por año (según los datos de muertes del año 2011 a 2015).

Para el año 2016, se estimaron 15,5 millones casos de supervivientes de cáncer en los Estados Unidos y se prevé que aumente a 20,3 millones para 2026. Del mismo modo, se estima que cerca de 38,4 % de pacientes recibirán un diagnóstico de cáncer en algún momento de sus vidas (según datos del año 2013 a 2015).

Bajo esta premisa, se trabajó con un conjunto de datos de pacientes diagnosticados de cáncer colorrectal. Este conjunto de datos contenía 1134 observaciones clínicas, que correspondían a pacientes de los laboratorios clínicos del servicio de diagnóstico molecular en el Centro de Cáncer Memorial Sloan Kettering (MSKCC), desde abril de 2014 hasta septiembre de 2016. Estos datos fueron utilizados por Yaeger et al.(2018) en su articulo *Clinical Sequencing Defines the Genomic Landscape of Metastatic Colorectal Cancer.* Para seleccionar los datos de supervivencia de los pacientes, se utilizarán los criterios de inclusión, exclusión y eliminación, los que se describen la sección 3.2.

3.2 Criterios de selección del conjunto de datos.

Según lo planteado en la sección 2, para que un conjunto de datos siga una distribución Weibull geométrica, estos deben cumplir con algunas normas o supuestos. Por esta razón, para la selección de datos se utilizaron criterios de inclusión, exclusión y eliminación.

3.2.1 Criterio de inclusión.

- Se consideran a todos los pacientes que fueron diagnosticados de cáncer colorrectal.

3.2.2 Criterio de exclusión.

Se debe tener en cuenta que, según lo propuesto por Barreto-souza, Lemos y Cordeiro los eventos ocurridos en la variable de estudio deben ser del mismo tipo.

- Dichos eventos serán el tratamiento al que se someta un grupo de pacientes; es por esto que se excluyen todos aquellos pacientes que no fueron sometidos a tratamiento de quimioterapia.

3.2.3 Criterio de eliminación.

- Serán eliminados aquellos pacientes que tengan tiempo de supervivencia igual a 0.

Luego de definir los criterios que se utilizarán para la selección de pacientes, se procede a estudiar el conjunto de datos reales de los pacientes con cáncer, y, posteriormente, aplicar los criterios anteriormente mencionados.

3.3 Conjunto de datos reales.

El cáncer se origina cuando las células en el cuerpo comienzan a crecer en forma descontrolada. La mayoría de las célula del cuerpo puede convertirse en cáncer y ramificarse, esto se conoce como metástasis. En particular, el cáncer colorrectal se origina en el colon o el recto, es por esto que, también es conocido como cáncer de colon o cáncer de recto (rectal), dependiendo del lugar donde se origine (American Cancer Society, 2018).

Antes de trabajar con los datos clínicos de los pacientes, es necesario definir las variables que serán parte de este estudio.

3.3.1 Conceptos claves y definición de variables.

Primeramente se presentan algunos conceptos técnicos sobre el cáncer colorrectal, estas definiciones fueron extraídas desde la página oficial del Instituto Nacional del Cáncer de Estados Unidos (https://www.cancer.gov/).

- **Cáncer Colorrecatal:** Cáncer de colon o de recto ubicado en el extremo inferior del tracto digestivo.
- **Quimioterapia:** La quimioterapia es una terapia empleada en el tratamiento del cáncer que consiste en emplear diversos fármacos para destruir células cancerígenas y reducir o eliminar completamente la enfermedad.
- **Estado o estadio del tumor:** El tamaño de un tumor maligno depende, en gran medida, de la etapa de desarrollo en que se encuentre. Estas etapas o estadios son:
  - Estadio 0 o cáncer in situ: Se considera que los tumores clasificados en esta etapa son aquellos que son considerados como benignos, pues, no ha crecido más allá de la capa interna (mucosa) del colon o del recto.
  - Estadio I: El cáncer ha crecido, atravesado la mucosa e invadido la capa muscular del colon o el recto. No se ha diseminado a los tejidos cercanos o ganglios linfáticos.
  - Estadio II: Este estadio se divide en dos tipos:
    - Estadio II A: El cáncer ha crecido y atravesado la pared del colon o del recto.
    - Estadio II B: Ocurre cuando, el cáncer ha crecido a través de las capas musculares hasta llegar al revestimiento del abdomen.

    En ambos casos, el cáncer no se ha diseminado a los tejidos o ganglios linfáticos cercanos.
  - Estadio III: Este estadio se divide en tres tipos:
    - Estadio III A: El cáncer ha crecido a través del revestimiento interno del intestino, este se ha diseminado hacia 1 a 3 ganglios linfáticos. No se ha diseminado a otras partes del cuerpo.
    - Estadio III B: El cáncer ha crecido a través de la pared intestinal o en los órganos circundantes y en 1 a 3 ganglios linfáticos. No se ha diseminado a otras partes del cuerpo.
    - Estadio III C: El cáncer se ha diseminado a 4 o más ganglios linfáticos, pero no a otras partes distantes del cuerpo.
  - Estadio IV: De igual modo este estadio se divide en tres tipos:
    - Estadio IV A: el cáncer se ha diseminado a una sola parte distante del cuerpo, como el hígado o los pulmones.

    - Estadio IV B: el cáncer se ha diseminado a más de 1 parte del cuerpo.
    - Estadio IV C: el cáncer se ha diseminado al peritoneo. También puede haberse diseminado a otras partes y órganos.

- **Tipo de tumor según ramificación:** Esta variable clasifica los tumores según su estado de metástasis o ramificación.

    - Metástasis: proceso de propagación de un foco canceroso a un órgano distinto de aquel en que se inició. Ocurre generalmente por vía sanguínea o linfática.
    - Primario : tumor que no ha presentado metástasis.

- **Localización del tumor:** Variable que indica si el tumor se encuentra al lado izquierdo o derecho del intestino grueso.

- **Supervivencia:** Tiempo de duración del tratamiento registrado para cada paciente. El tiempo medido corresponde a los meses transcurridos desde el inicio del tratamiento hasta la recuperación, abandono de tratamiento o muerte del paciente.

- **Edad:** Edad del paciente al momento del diagnóstico.

- **Género:** Variable que indica el sexo del paciente (femenino o masculino).

3.3.2 Datos de supervivencia.

Se seleccionó un total de 6 variables del conjunto de datos original, y se realizó un análisis descriptivo de estas. En primer lugar, se estudió la cantidad de hombres y mujeres presentes en el conjunto de datos. De este modo, la Tabla 3.1 muestra que 616 pacientes eran hombres (54,32 %), mientras que 518 pacientes eran mujeres (45,68 %).

Variable		Cantidad de pacientes	Porcentaje (%)
Género	Femenino	518	45,68
	Masculino	616	54,32
	Total	1.134	100,00
Tipo de tumor	Metatástico	533	47,00
	Primario	601	53,00
	Total	1134	100,00
Localización del tumor	I	41,00	3,62
	II	133	11,73
	III	274	24,16
	IV	686	60,49
	Total	1.134	100,00
Quimioterapia	Si	593	52,29
	No	541	47,71
	Total	1.134	100,00

Tabla 3.1: Análisis de frecuencia para las variables género, tipo de tumor, localización del tumor y tratamiento de quimioterapia.

De igual manera se realizó un análisis de frecuencia para la cantidad de tumores primarios y aquellos que hayan presentado ramificación en alguna parte del cuerpo del paciente. La tabla 3.1 muestra que 53,00 % son tumores primarios. Por otro lado, anteriormente se definieron los tipos de estadios del cáncer, de este modo, el análisis de frecuencia realizado mostró que el 60,49 % de los tumores estaban en estadio IV, es decir habrían presentado algún tipo de ramificación. Luego, se realizó un análisis de frecuencia para la localización del tumor en el intestino grueso, según este análisis, se observa que el 69,55 % de los tumores estaban localizados al lado izquierdo del colon.

Por último, se analizó la cantidad de pacientes que fueron sometidos a tratamiento de quimioterapia. En la Tabla 3.1 se muestra que 541 pacientes realizaron este tratamiento, pero es necesario tener en cuenta que también pueden tener un tratamiento alterativo además de las sesiones de quimioterapia.

Variable	Media	Desviación estándar	Mínimo	Máximo
Superviviencia (meses)	38,31	34,22	0,00	292,93
Edad	54,62	12,80	13,00	93,00

Tabla 3.2: Medidas descriptivas para el tiempo de supervivencia de los pacientes en meses.

Luego se estudió el tiempo de supervivencia de los pacientes, en donde se puede observar que el tiempo promedio de supervivencia para los pacientes es de 38,31 meses. Además, se observa que, la edad promedio de los pacientes fue de 54,62 años. Estos análisis serán de ayuda para la selección de pacientes en la sección 3.3.3 se presentan los datos seleccionados para la aplicación de la distribución Weibull geométrica.

3.3.3 Conjunto de datos seleccionados.

Luego de aplicar los criterios de selección expuestos en la sección 3.2, se seleccionaron 527 pacientes que fueron diagnosticados de cáncer colorrectal y fueron sometidos a tratamiento de quimioterapia. Luego, se realizó un análisis descriptivo estos pacientes, con el objetivo de conocer como fueron distribuidos según sexo, tipo de tumor (primario o metástico), lugar del tumor, edad y estadio del tumor.

Variable		Cantidad de pacientes	Porcentaje (%)
Género	Femenino	236	45,78
	Masculino	291	55,21
	Total	527	100
Tipo de tumor	Metatástico	399	75,71
	Primario	128	24,29
	Total	527	100
Localización del tumor	I	9	1,71
	II	51	9,68
	III	150	28,46
	IV	317	60,15
	Total	527	100

Tabla 3.3: Análisis de frecuencia de las variables seleccionadas para los 527 pacientes seleccionados.

Según el estado de propagación del cáncer existen dos tipos de tumores, aquellos que han presentado metástasis y los que aún no presentan propagaciones, que son conocidos como tumores primarios. Según la Tabla 3.3, el 75,71 % de los pacientes seleccionados tienen tumores con presencia de metástasis, mientras que el 24,29 % de los pacientes tenían tumores

primarios. Además del estado de propagación del cáncer, existe, adicionalmente, otra categoría para identificar el tipo de tumor presente en los pacientes diagnosticados con cáncer colorrectal. Este tipo de clasificación es llamada ëstadio del cáncer.̈ Según el análisis de frecuencias realizado, se puede observar que el 60,15 % de los pacientes presentan tumores de estadio IV, lo que significa que el cáncer se ha desiminado a 1 o más partes del cuerpo.

En la Tabla 3.3 también se puede observar que, según la localización del tumor en los pacientes, el 76,35 % de ellos tenían tumores ubicados en el lado izquierdo del colon, mientras que el 23,65 % de los pacientes presentaba el tumor alojado en el lado derecho.

Variable	Media	Desviación estándar	Mínimo	Máximo
Superviviencia (meses)	49,51	39,23	1,67	292,93
Edad	53,06	11,95	20,00	83,00

Tabla 3.4: Medidas descriptivas para el tiempo de supervivencia en meses de los pacientes seleccionados para la aplicación.

Según la información entregada por American Society of Clincal Oncology (ASCO) el riesgo de desarrollar cáncer colorrectal aumenta con la edad. El cáncer colorrectal puede aparecer en adultos jóvenes y adolescentes, pero la mayoría de los casos de cáncer colorrectal se presentan en personas mayores de 50 años.

Para el caso particular de los pacientes seleccionados para este estudio, la edad promedio al momento del diagnóstico es de 53,06 años, siendo 20 y 83 años el mínimo y el máximo, respectivamente (véase la Tabla 3.4). Mientras que, el tiempo de supervivencia de los 527 pacientes seleccionados, en promedio, fue de 49,51 meses.

Por otra parte, ASCO informa que los hombres presentan un riesgo ligeramente mayor a desarrollar este tipo de cáncer. Para el cáncer de colon, la edad promedio de las personas al momento de recibir el diagnóstico es de 68 años en los hombres y de 72 años en las mujeres. Para el cáncer de recto, es de 63 años tanto para los hombres como para las mujeres. La tabla 3.5 muestra los resultados del análisis descriptivo para la edad de diagnóstico de los pacientes con cáncer colorrectal según género. Observe que, la edad mínima de diagnóstico es de 4 años menos en hombres que en mujeres, mientras que el promedio es ligeramente cercano, siendo 52,33 años para mujeres y 53,65 años para hombres.

Género	Media	Desviación estándar	Mínimo	Máximo
Femenino	52,33	11,76	24,00	76,00
Masculino	53,65	12,10	20,00	83,00

Tabla 3.5: Estadística descriptiva para la edad de los pacientes según género.

La figuras 3.1 muestran el comportamiento de la edad de los pacientes con cáncer colorrectal, según género.

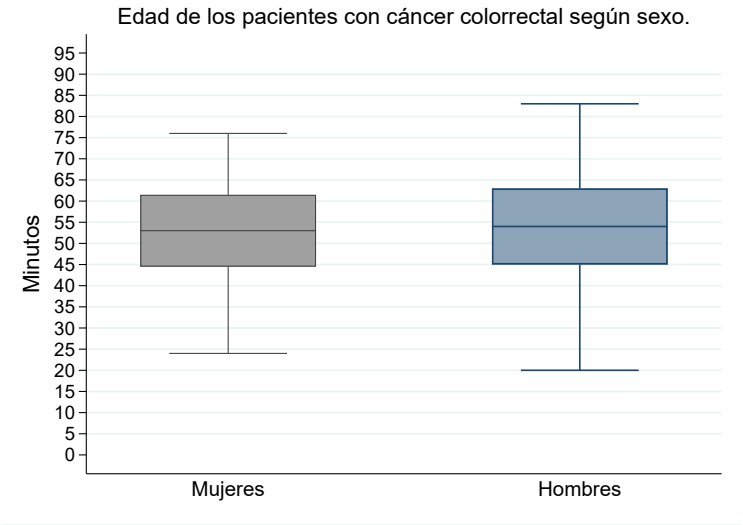


Figura 3.1: Boxplot para la edad de los pacientes diagnosticados con cáncer colorrectal.

Además, la figura 3.2 corresponde a un boxplot para el tiempo de supervivencia de los pacientes seleccionados, en donde se puede observar una cantidad considerable de datos atípicos. Además, se observa que el 50 % de los pacientes presentan tiempo de supervivencia entre 25 y 65 meses.

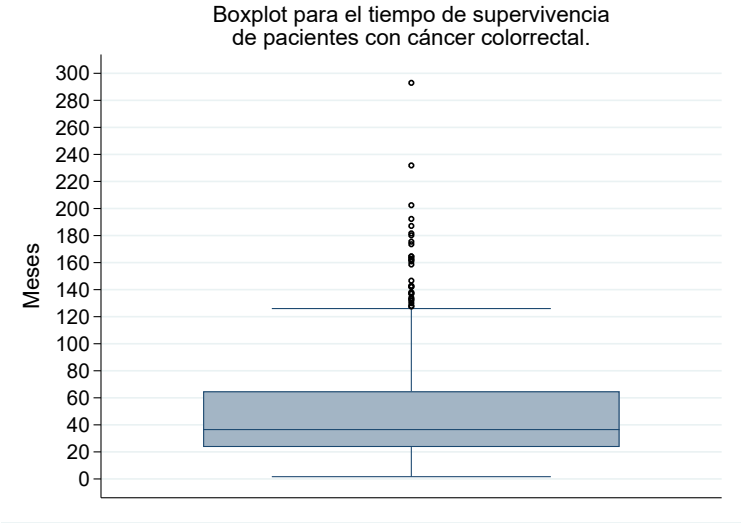


Figura 3.2: Boxplot para el tiempo de supervivencia de los pacientes diagnosticados con cáncer colorrectal.

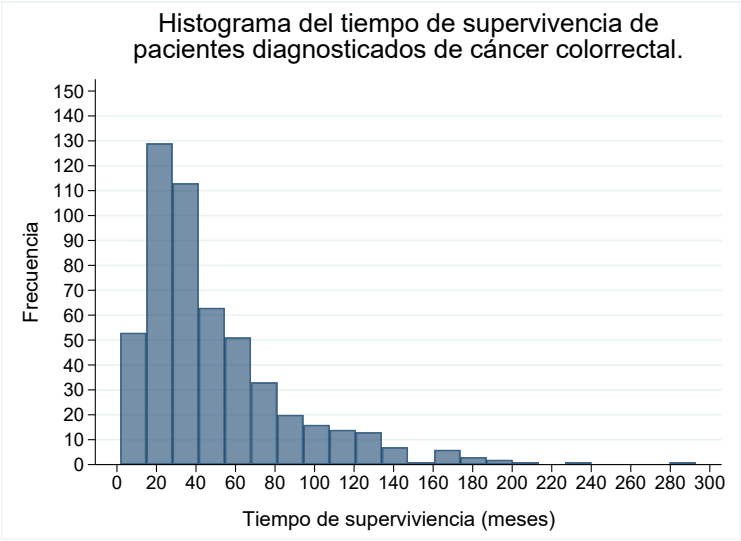


Figura 3.3: Boxplot para el tiempo de supervivencia de los pacientes diagnosticados con cáncer colorrectal.

Por último, se realizó un histograma del tiempo de supervivencia de los pacientes, con el objetivo de estimar el comportamiento de las tasas de supervivencia (o muerte). Según lo visualizado en la Figura 3.3 es posible sostener que la supervivencia (o muerte) de los pacientes está representada por tasas de fallas tempranas. Esta información es complementada con la Figura 3.4

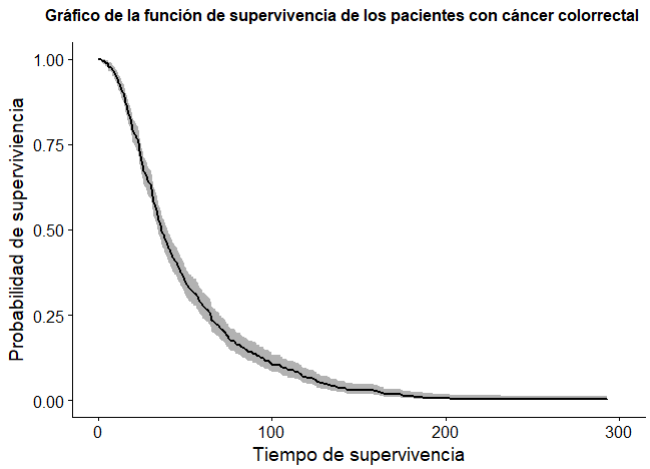


Figura 3.4: Gráfico de la función de supervivencia para el tiempo de vida de los pacientes seleccionados.

Por último, la Figura 3.4 muestra que se está trabajando con tasas de supervivencia de falla temprana. Luego de trabajar con los datos seleccionados, se procede a aplicar la distribución Weibull geométrica y contrastar esta distribución con la distribución de Weibull. El detalle de este análisis se presenta en la sección 3.4.

3.4 Aplicación de la distribución Weibull geométrica.

3.4.1 Estimación de parámetros.

En primer lugar, se calcularon los estimadores de la distribución Weibull Geométrica para los 527 datos de supervivencia de pacientes con cáncer colorrectal, utilizando el algoritmo EM planteado en la sección de metodología y el método de máxima verosimilitud. Además, se calcularon los estimadores máximo verosímil de los parámetros de la distribución Weibull, para posteriormente comparar ambas distribuciones.
La Tabla 3.6 muestra los resultados para los parámetros de ambas distribuciones.

Método						
Distribución	Algoritmo EM			EMV		
	β	α	p	β	α	p
Weibull	-	-	-	54,560	1,380	-
Weibull geométrica	0,008	2,092	0,922	0,007	2,108	0,929

Tabla 3.6: Valores para la estimación de los parámetros de las distribuciones Weibull y Weibull geométrica.

Al examinar los resultados de la Tabla 3.6, es posible apreciar que no existe una gran diferencia entre la estimación mediante el algoritmo EM y el método de de máxima verosimilitud (EMV), para la estimación de parámetros de la distribución Weibull geométrica. Además, recuerde que en la sección de 2.2.4 se estudió el comportamiento de la distribución Weibull geométrica según el valor de sus parámetros. En este análisis se observó que cuando los parámetros $\beta \approx 2$, $\alpha < 1$ y $p \approx 0{,}9$ la función de densidad de la distribución se vuelve más flexible para los datos extremos. En la figura 3.5 se puede observar que, al aplicar la distribución Weibull geométrica a los datos de supervivencia de pacientes con cáncer colorrectal, este supuesto se cumple. Además, se estima de manera gráfica que el ajuste de los datos utilizando la distribución Weibull geométrica es más apropiado que el ajuste de la distribución Weibull.

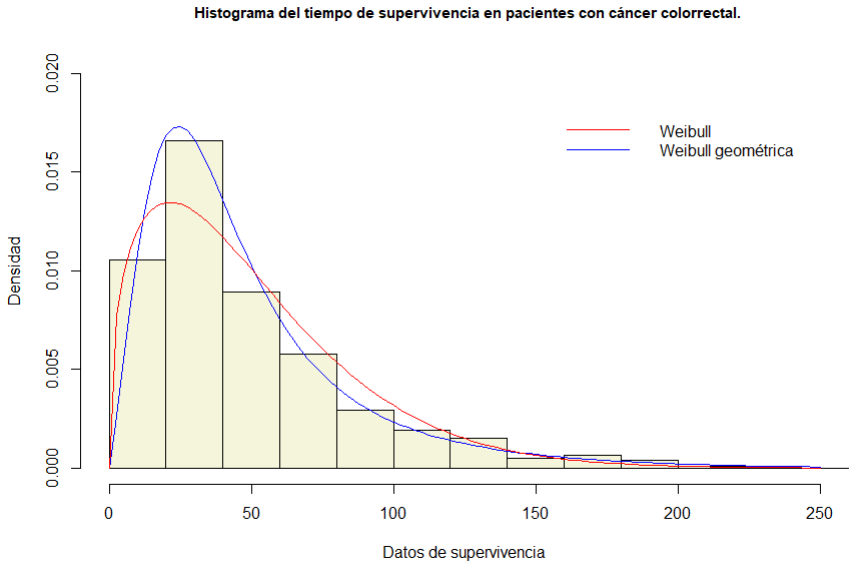


Figura 3.5: Histograma y ajuste de las distribuciones Weibull y Weibull geométrica.

Luego, se comparó el ajuste de los datos a ambas distribuciones utilizando gráficos de cuantil-cuantil. Estos gráficos se muestran en la figura 3.6. Note que, para la distribución Weibull los datos se alejan de la recta mas rápidamente en comparación con aquellos de la distribución Weibull geométrica.

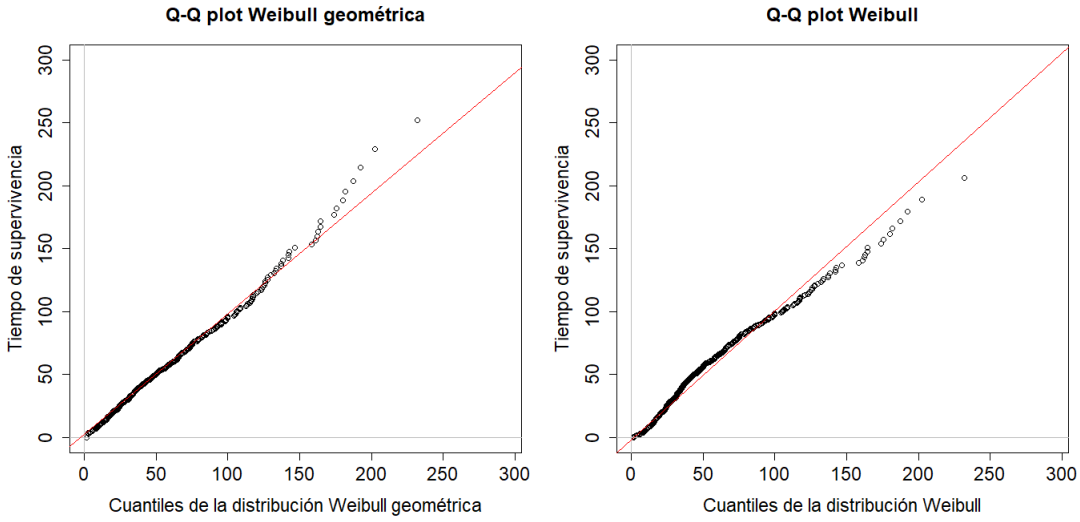


Figura 3.6: Q-Q plot para las distribuciones Weibull y Weibull geométrica.

Por último, se analizó el ajuste de los datos a ambas distribuciones utilizando el criterio de información de akaike. Los resultados de este análisis se muestran en la Tabla 3.7. Con esto se comprueba que la distribución Weibull geométrica es más flexible que la distribución Weibull, lo que es favorable a la hora de ajustar datos con valores extremos o atípicos. Por consiguiente, según los resultados obtenidos la distribución que mejor representa los datos de supervivencia de los pacientes es la distribución Weibull geométrica.

Distribución	AIC
Weibull geométrica	3.052,91
Weibull	3.244,67

Tabla 3.7: Valores obtenidos según el criterio de Akaike en las distribuciones Weibull y Weibull geométrica.

Análisis de residuos.

Luego de comprobar que la distribución Weibull geométrica es más apropiada que la distribución Weibull, se estudiaron los residuos de la distribución Weibull geométrica, con el objetivo de comprobar el cumplimiento de los supuestos de normalidad y homocedasticidad.

Para esto, se utilizó la prueba de Kolmogorov Smirnov y el test de Fligner Killeen para comprobar normalidad y homogeneidad de varianza, respectivamente. La tabla 3.9 muestra los resultados para las pruebas mencionadas anteriormente, ademas la Tabla 3.8 contiene un análisis descriptivo de los residuos.

Mínimo	1er cuantil	Mediana	Media	3er cuantil	Máximo
-216,851	-26,800	1,097	0,038	27,528	211,935

Tabla 3.8: Medidas descriptivas para los residuos de la distribución Weibull geométrica.

Observe que no se cumple el supuesto de normalidad según la prueba de Kolmogorov Smirnov, lo que indica que si se desea aplicar un modelo de regresión para esta distribución, se debe trabajar con modelos lineales generalizados. Por otro lado, la prueba de Fligner Killeen acepta el supuesto de homocedasticidad, esto puede comprobarse de manera gráfica en la Figura 3.7.

	Prueba K-S		Prueba de Fligner Killeen	
Variable	Estadistico D	Valor p	Estadistico χ^2	Valor p
Residuos	0,070	0,004	497,87	0,18

Tabla 3.9: Resultados para los test K-S, Shapiro Wilk y Fligner Killeen.

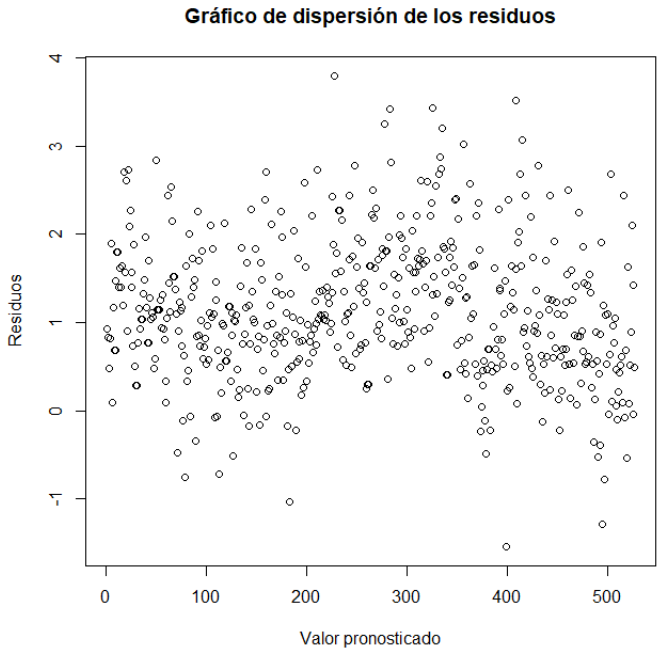


Figura 3.7: Gráfico de dispersión de los residuos de la distribución Weibull geométrica.

Al finalizar el estudio y aplicación de la distribución Weibull geométrica, y comprobar que es más flexible y apropiada que la distribución Weibull, el siguiente paso será aplicar un modelo GAMLS y estudiar la relación de las variables estudiabas en la sección 3.3.3 con el tiempo de supervivencia de los pacientes con cáncer colorrectal.

3.4.2 Modelo GAMLSS para la distribución Weibull geométrica.

Una de las ventajas de utilizar modelos GAMLSS, es que permite modelar cada uno de los parámetros de la distribución en estudio. Es por esto que, se aplicará el modelo GAMLSS utilizando la distribución Weibull geométrica para estudiar la relación que existe entre el tiempo de supervivencia de los pacientes diagnosticados con cáncer colorrectal y las variables: edad, estadio del tumor, sexo, tipo de tumor (primario o metastásico) y localización del tumor.

β				
	Estimación	Error estándar	Valor t	Valor p
(Intercepto)	1,061	0,018	57,868	1,64e-228
Género masculino (X_1)	-0,035	0,003	-9,212	7,87e-19
Edad (X_2)	0,004	0,001	30,687	1,04e-118
Lado derecho (X_3)	-0,118	0,004	-28,207	8,10e-107
Estadio II (X_4)	-0,289	0,017	-16,131	9,58e-48
Estadio III (X_5)	-0,340	0,016	-20,490	8,19e-69
Estadio IV (X_6)	-0,099	0,016	-6,027	3,16e-09
Tumor metastásico (X_7)	-0,993	0,009	-107,898	<2e-16
α				
	Estimación	Error estándar	Valor t	Valor p
(Intercepto)	-8,294	3,43e-05	-241675,369	<2e-16
Género masculino(X_1)	-0,545	1,03e-05	-52665,122	<2e-16
Edad (X_2)	0,033	3,33e-07	99355,027	<2e-16
Lado derecho (X_3)	-0,671	9,13e-06	-73561,481	<2e-16
Estadio II (X_4)	-1,002	4,15e-05	-24177,037	<2e-16
Estadio III (X_5)	-1,136	3,20e-05	-35485,326	<2e-16
Estadio IV (X_6)	-0,289	3,18e-05	-9095,501	<2e-16
Tumor metastásico (X_7)	3,088	7,38e-05	41861,397	<2e-16

Tabla 3.10: Estimaciones del modelo GAMLSS para el parámetro β, α.

p				
	Estimación	Error estándar	Valor t	Valor p
(Intercepto)	13,076	7,66e-06	1706746,740	<2e-16
Género masculino (X_1)	0,128	1,52e-06	84714,831	<2e-16
Edad (X_2)	-0,004	6,21e-08	-75263,033	<2e-16
Lado derecho (X_3)	0,127	1,69e-06	75088,718	<2e-16
Estadio II (X_4)	0,227	7,42e-06	30731,156	<2e-16
Estadio III (X_5)	0,126	6,96e-06	18174,912	<2e-16
Estadio IV (X_6)	0,155	6,89e-06	22569,853	<2e-16
tumor metastásico (X_7)	-12,108	0,029	-408,676	<2e-16

Tabla 3.11: Estimaciones del modelo GAMLSS para el parámetro p.

En la Tabla 3.10 y 3.11 se puede observar que para los parámetros de la distribución Weibull geométrica, todas las variables estudiadas influyen en el tiempo de supervivencia de los pacientes con cáncer de colon. Recuerde que el modelo GAMLSS está definido por:

$$g_k(\theta_k) = \eta_k = X_k\beta_k.$$

Por lo tanto las ecuaciones que representan el modelo aplicado para cada parámetro de la distribución Weibull geométrica son:

$$log(\hat{\beta}) = 1,061 - 0,035 * X_1 + 0,004 * X_2 - 0,118 * X_3 - 0,289 * X_4 - 0,340 * X_5 - 0,099 * X_6 - 0,993 * X_7$$

$$log(\hat{\alpha}) = -8,294 - 0,545 * X_1 + 0,033 * X_2 - 0,671 * X_3 - 1,002 * X_4 - 1,136 * X_5 - 0,289 * X_6 + 3,088 * X_7$$

$$logit(\hat{p}) = 13,076 + 0,128 * X_1 - 0,004 * X_2 + 0,127 * X_3 + 0,227 * X_4 + 0,126 * X_5 + 0,155 * X_6 - 12,108 * X_7$$

Si se quiere analizar la influencia de los factores en el tiempo de supervivencia, basta con observar los valores obtenidos en la Tabla 3.10 para el parámetro β. Observe que, según los resultados obtenidos, los pacientes de género masculino tienen un valor diferencial de -0,035, esto indica que, el tiempo de supervivencia disminuye 0,035 veces. De igual modo si el tumor si encuentra al lado derecho del intestino grueso, el tiempo de supervivencia disminuirá en 0,118 veces.

3.4.3 Aspectos finales.

Un aspecto importante para definir si el modelo propuesto es apropiado, es el estudio de los residuos del modelo aplicado. Para esto, se estudiaron los supuestos de normalidad y homocedasticidad. Recuerde que en el caso particular de los modelos GAMLSS, se trabaja con los cuantiles aleatorios residuales. Por lo tanto, si se cumple el supuesto de normalidad, se puede concluir que el modelo planteado es un modelo con buen ajuste. La tabla 3.12 muestra que para el modelo GAMLSS aplicado, se cumplen los supuestos de normalidad y homocedasticidad.

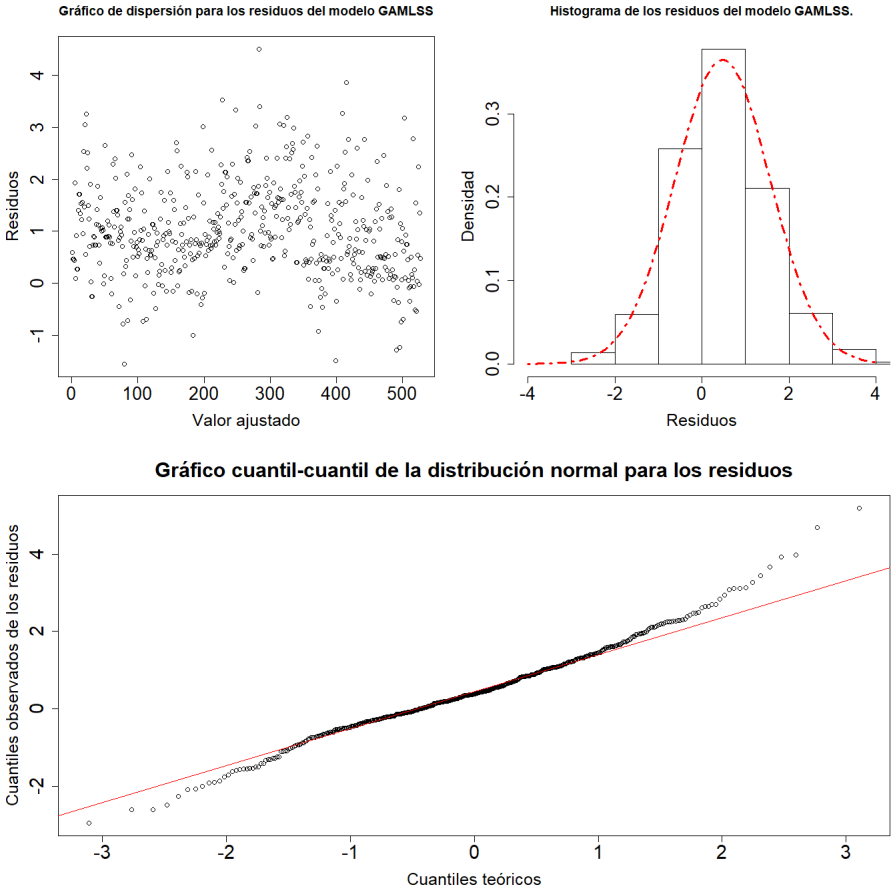


Figura 3.8: Diagnóstico de los residuos del modelo GAMLSS.

Asimismo, los gráficos de la Figura 3.8 avalan el cumplimiento de los supuestos ya mencionados. Observe que no se perciben patrones o agrupamientos en el gráfico de dispersión, de modo que, se concluye que los residuos del modelo son aleatorios. Además, el gráfico cuantil-cuantil y el histograma de la Figura 3.8 reafirman el supuesto de normalidad de los residuos.

	Prueba K-S		Prueba Fligner Killeen	
Variable	Estadítico D	Valor p	Estadístico χ^2	Valor p
Residuos	0,05	0,1323	482,2	0,3268

Tabla 3.12: Resultados para las pruebas KS y Fligner Killeen.

Capítulo 4

Conclusiones Finales

En este trabajo se comparó la distribución Weibull geométrica con la distribución de Weibull, utilizando datos clínicos de pacientes diagnosticados de cáncer colorrectal. En el estudio de la distribución Weibull geométrica, se pudo observar que esta distribución presenta mayor flexibilidad, por ende, es apropiada cuando se trabaja con datos altamente curtóticos.

Luego de comparar ambas distribuciones se puede concluir que, los datos clínicos de los pacientes seleccionados presentaron un mejor ajuste al trabajar con la distribución Weibull geométrica. Además, se estimaron los parámetros de la distribución WG mediante el método de estimadores máximo verosímil (EMV) y el algoritmo EM. Se pudo observar a través de esta estimación, que los valores obtenidos fueron relativamente cercanos, esto ocurre ya que ambos métodos buscan maximizar la función de verosimilitud.

Por otra parte, para el estudio de la relación del tiempo de supervivencia con las variables género, edad, tipo de tumor, estadio del tumor y localización del tumor, se utilizó un modelo GAMLSS. Estos modelos son muy útiles cuando se pretende modelas la media o parámetro de escala y además, los parámetros de forma y localización.

Al aplicar el modelo GAMLSS propuesto, se pudo observar en los resultados que, todas las variables estudiadas influyen significativamente en el tiempo de supervivencia de los pacientes, y además el comportamiento de los parámetros de forma y localización.

Aunque se ha demostrado que la distribución Weibull geométrica es una buena opción para trabajar con datos de tiempos de vida, no se descarta la posibilidad de estudiar otras distribuciones. Otro propuesta interesante para estudios futuros, es estudiar algún modelo o distribución para datos que presenten función de densidad con forma de bañera, dado que en este estudio los datos solo mostraron una tasa de supervivencia decreciente. Además, un tema interesante para estudios futuros es implementar la entropía de Rényi y Shannon expuestas en la sección de metodología en conjunto con nuevas aplicaciones en diferentes áreas de la ciencia.

Referencias

Adamidis, K. y Loukas, S. (1998). A lifetime distribution with decreasing failure rate. *Statistics and Probability Letters* , *39* (1), 35-42. Descargado de https://www.sciencedirect.com/science/article/pii/S0167715298000121

Adamidis, K., Dimitrakopoulou, T., & Loukas, S. (2005). On an extension of the exponential-geometric distribution. *Statistics and probability letters*, *73*(3), 259-269. Descargado de: https://www.sciencedirect.com/science/article/pii/S0167715205001100

Almalki, S. J., y Nadarajah, S. (2014). Modifications of the Weibull distribution: a review. *Reliability Engineering and System Safety* , *124*, 32-55. Descargado de https://www.sciencedirect.com/science/article/pii/S0951832013003074

Alvarez-Vaz, R., Palamarchuk, P., and Riano, E. (2017). Elaboración de patrones espirométricos en niños uruguayos mediante modelos GAM y GAMLSS: parte 2-Modelización de CVF y FEV por talla edad y sexo. Serie DT (17/2). Descargado de: https://www.colibri.udelar.edu.uy/jspui/handle/123456789/10986

Barajas, F. H., Naranjo-Dueñas, G., and Monsalve-Lugo, E. (2017). Estimación del rendimiento de orellana mediante modelos Gamlss. Revista de la Facultad de Ciencias, 6(1), 67-82. Descargado de https://revistas.unal.edu.co/index.php/rfc/article/view/61119.

Barreto-Souza, W., de Morais, A. L., y Cordeiro, G. M. (2011). The Weibull-geometric distribution. *Journal of Statistical Computation and Simulation*, *81* (5), 645-657. Descargado de https://www.tandfonline.com/doi/abs/10.1080/00949650903436554

Bebbington, M., Lai, C.-D., Wellington, M., y Zitikis, R. (2012). The discrete additive Weibull distribution: A bathtub-shaped hazard for discontinuous failure data. *Reliability En-*

gineering and System Safety , 106 , 37-44. Descargado de https://www.sciencedirect.com/scienc e/article/pii/S0951832012001147

Clark, T. G., Bradburn, M. J., Love, S. B., and Altman, D. G. (2003). Survival analysis part IV: further concepts and methods in survival analysis. *British journal of cancer*, 89(5), 781. Descargado de https://www.nature.com/articles/6601117

Davis, D. (1952). An analysis of some failure data. *Journal of the American Statistical Association*, *47* (258), 113-150. Descargado de https://amstat.tandfonline.com/doi /abs/10.1080/01621459.1952.10501160#.WuccTIgvzIU

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. Biometrics, 1041-1046.

Gleser, L. J. (1989). The gamma distribution as a mixture of exponential distributions. *The American Statistician* , *43* (2), 115-117. Descargado de https://www.tandfonline.com/doi/abs /10.1080/00031305.1989.10475632?journalCode=utas20

Gupta, R. D. y Kundu, D. (1999). Generalized exponential distributions. *Australian and New Zealand Journal of Statistics* , 41 (2), 173-188. Descargado de https://onlinelibrary.wiley.com/ doi/abs/10.1111/1467-842X.00072

Kemp, A. W. (1997). Characterizations of a discrete normal distribution. *Journal of Statistical Planning and Inference*, *63* (2), 223-229. Descargado de https://www.sciencedirect.com/ science/article/pii/S0378375897000207

Lisman, J. y Van Zuylen, M. (1972). Note on the generation of most probable frequency distributions. *Statistica Neerlandica*, *26* (1), 19-23. Descargado de https://onlinelibrary.wiley.com/ doi/abs/10.1111/j.1467-9574.1972.tb00152.x

Lomax, K. S. (1954). Business failures: Another example of the analysis of failure data. *Journal of the American Statistical Association*, *49* (268), 847-852. Descargado de https://www. tandfonline.com/doi/abs/10.1080/01621459.1954.10501239

Makcutek, J. (2008). A generalization of the geometric distribution and its application in quantitative linguistics. *Romanian Reports in Physics*, *60* (3), 501-509. Descargado de https://www.researchgate.net/publication/297574962_A_generalization_of_the_geometric_dis tribution_and_its_application_in_quantitative_linguistics

Marshall, A. W. y Olkin, I. (1997). A new method for adding a parameter to a family of dis-

tributions with application to the exponential and Weibull families. *Biometrika*, *84* (3), 641-652. Descargado de https://academic.oup.com/biomet/article-abstract/84/3/641/217183

McCullagh, y Nelder, J. A. (1983). Generalized linear models (Segunda ed.; Chapman y Hall, Eds.). Descargado de http://www.utstat.toronto.edu/ brunner/oldclass#/2201s11/readings/ glmbook.pdf

Mood, A. M. Graybill, F. A., y Boes, D. C. (1974). *Introduction to the theory of statistics* (Tercera ed.). McGraw-Hill Kogakusha.

Najarzadegan, H., y Alamatsaz, M. H. (2017). A new generalization of weighted geometric distribution and its properties. *Journal of the American Statistical Association*, *16* (4), 522-546. Descargado de https://www.atlantis-press.com/journals/jsta/25887940

Nascimento, A. D., Bourguignon, M., Zea, L. M., Santos-Neto, M., Silva, R. B., y Cordeiro, G. M. (2014). The gamma extended Weibull family of distributions. *Journal of Statistical Theory and Applications* , *13* (1), 1-16. Descargado de https://www.atlantis-press.com/proceedings/ jsta/11607

Nekoukhou, V. Alamatsaz, M., Bidram, H., y Aghajani, A. (2015). Discrete beta-exponential distribution. *Communications in Statistics-Theory and Methods* , 44 (10), 2079-2091.5

Nekoukhou, V. y Bidram, H. (2015). The exponentiated discrete Weibull distribution. *SORT*, *39* , 127-146. Descargado de https://upcommons.upc.edu/handle/2117/88523

Patil, G. Rao, C., y Ratnaparkhi, M. (1986). On discrete weighted distributions and their use in model choice for observed data. *Communications in Statistics-Theory and Methods*, *15* (3), 907-918. Descargado de https://www.tandfonline.com/doi/abs/10.1080/03610928608829159

Pérez, M. J. (2016). Modelos Aditivos Xeneralizados para Localización, Escala e Forma (GAMLSS).Descargado de: http://eio.usc.es/pub/mte/descargas/Proye
ctosFinMaster/Proyecto_1361.pdf

Proschan, F. (1963). Theoretical explanation of observed decreasing failure rate. *Technometrics*, *5* (3), 375-383. Descargado de https://amstat.tandfonline.com/doi/abs/10.1080/00401706.1963.10
490105#.WucrqogvzIU

Royston, P., and Parmar, M. K. (2002). Flexible parametric proportional hazards and proportional odds models for censored survival data, with application to prognostic modelling

and estimation of treatment effects. Statistics in medicine, 21(15), 2175-2197. Descargado de: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.1203

Silva, R. B. Barreto-Souza, W., y Cordeiro, G. M. (2010). A new distribution with decreasing, increasing and upside-down bathtub failure rate. *Computational Statistics and Data Analysis* , *54* (4), 935-944. Descargado de https://www.sciencedirect.com/science/article/pii/S0167947309003764

Xie, M., y Lai, C. D. (1996). Reliability analysis using an additive Weibull model with bathtub-shaped failure rate function. *Reliability Engineering and System Safety*, *52*(1), 87-93. Descargado de: https://www.sciencedirect.com/science/article/pii/0951832095001492

Yaeger, R., Chatila, W. K., Lipsyc, M. D., Hechtman, J. F., Cercek, A., Sanchez-Vega, F., and You, D. (2018). *Clinical sequencing defines the genomic landscape of metastatic colorectal cancer.* Cancer cell, 33(1), 125-136.