Universidad Autónoma de Querétaro
Facultad de Ingeniería
Maestría en Ingeniería Matemática

ALTERNATIVAS DE MODELAJE ESTADÍSTICO EN ESTUDIOS DE VIDA DE ANAQUEL SENSORIAL DE ALIMENTOS

Tesis

Que como parte de los requisitos para obtener el Grado de

Maestro en Ciencias (Ingeniería Matemática)

Presenta

Iliana María Paternina Ortega

Dirigido por

Dr. Eduardo Castaño Tostado

Dr. Eduardo Castaño Tostado
Presidente

Dr. Mario Santana Cibrián
Secretario

Msc. Wilfrido Jacobo Paredes García
Vocal

Msc. Juan Antonio Villeda Resendiz
Suplente

Msc. Sara Silva Hernández
Suplente

Centro Universitario, Querétaro, Qro.
Fecha de aprobación por el Consejo Universitario (mes y año)
México

Dedicatoria

A Dios.

Agradecimientos

Agradezco ante todo a Dios, por darme la vida y brindarme todas las oportunidades para alcanzar mi postgrado.
Al Consejo Nacional de Ciencia y Tecnología, CONACYT, por la beca que me otorgó para para la realización de mis estudios de maestría.

Índice general

Índice de figuras

Lista de Tablas

Resumen

En esta tesis se describe un modelo estadístico para datos provenientes del análisis sensorial de alimentos, teniendo en cuenta la estructura de respuestas agrupadas en cada una de las personas que realizan la evaluación, así como el patrón de asociación en el tiempo para el conjunto de datos. Se hace uso del enfoque de las ecuaciones de estimación generalizadas que ajusta modelos de media marginal con la ventaja de que solamente la especificación correcta de las medias marginales es necesaria para que los estimadores de los parámetros sean consistentes y asintóticamente normales. Se encontró que, para una base de datos específica, donde se evalúa la vida de anaquel sensorial de un yogurt, la probabilidad de rechazo esperada sobre el tiempo es no lineal y creciente, y la estructura de correlación de trabajo más adecuada fue la no estructurada. Se realizó también un análisis comparativo entre el modelado realizado a partir del enfoque de las ecuaciones de estimación generalizadas y el modelado actual que usa técnicas de análisis de sobrevivencia. Se puede destacar que, al hacer uso del enfoque de las ecuaciones de estimación generalizadas se trabaja con las respuestas directas de los consumidores y no es necesaria una codificación o transformación de las mismas para su tratamiento estadístico, como es el caso de la aplicación de las técnicas de análisis de sobrevivencia, que es enfoque utilizado a la fecha.

Palabras claves: Datos longitudinales, autocorrelación, vida de anaquel sensorial.

Abstract

This thesis describes a statistical model for data from food sensory analysis, taking into account the structure of responses grouped in each of the people who carry out the evaluation, as well as the pattern of association over time for the set of data. The generalized estimation equations approach that fits marginal mean models is used with the advantage that only the correct specification of the marginal means is necessary for the parameter estimators to be consistent and asymptotically normal. It was found that for a specific database, where the sensory shelf life of a yogurt is evaluated, the expected rejection probability over time was found to be nonlinear and increasing, and the most appropriate working correlation structure was unstructured.

A comparative analysis was performed between the modeling carried out using the generalized estimation equations approach and the current modeling that uses survival analysis techniques, where it can be highlighted that using the generalized estimation equations approach, we work with the direct responses of consumers and it is not necessary to encode or transform them for their statistical treatment, as is the case of the application of survival analysis techniques, the approach used to date.
Keywords: Longitudinal data, autocorrelation, sensory shelf life.

Capítulo 1

Introducción

La característica principal de un estudio longitudinal es que los individuos son medidos repetidamente a través del tiempo, en contraste con los estudios transversales en donde una sola respuesta es medida por cada individuo. De hecho, el objetivo de un estudio longitudinal es la caracterización de los cambios en la respuesta de interés a lo largo del tiempo. Son variadas las situaciones en donde mediciones repetidas se hacen sobre la misma unidad, formándose así un clúster de observaciones correlacionadas, correspondiente a cada individuo participante. Por ejemplo, los estudios médicos en donde se suele hacer un seguimiento de pacientes en el tiempo, pueden generar este tipo de información, lo que constituye una de las aplicaciones más encontradas en torno al análisis de datos longitudinales.

En estudios de vida de anaquel sensorial la evaluación hecha por evaluadores humanos para rechazar o aceptar un producto alimenticio en diferentes tiempos de almacenamiento juega un papel especial. Las respuestas de cada evaluador al paso del tiempo de almacenamiento del producto alimenticio son registradas y, aunque espaciadas en el tiempo, pueden considerarse como un clúster de datos (datos longitudinales), además de que éstos pueden contener un patrón de asociación en el tiempo. Este tipo de información puede ser analizada mediante técnicas de datos longitudinales. No obstante, a la fecha se presentan modelos que utilizan técnicas de análisis de sobrevivencia para hacer las inferencias respectivas a la vida de anaquel sensorial de alimentos. Este es el caso del trabajo propuesto por Hough et al. (2003) donde se ignora la estructura de clúster en los datos de este tipo de estudios.

En el presente trabajo se lleva a cabo un modelaje estadístico alternativo de datos provenientes de análisis sensorial de alimentos incorporando la estructura de agrupación de tipo clúster que éstos poseen, así como su patrón de asociación en

el tiempo. Este enfoque permitirá modelar la probabilidad de rechazo por parte de los consumidores a través del tiempo, generando así, alternativas de modelaje estadístico para estudios de vida de anaquel sensorial de alimentos.

El trabajo se organiza de la siguiente manera. Primero se realiza un resumen de cómo es utilizado el análisis de sobrevivencia para realizar las inferencias respectivas a la vida de anaquel sensorial de alimentos, comprendido en el capítulo de antecedentes. Segundo, se presentan la hipótesis y los objetivos de la investigación; en tercer lugar, se presenta un capítulo de fundamentos teóricos, donde se muestran los principales temas que aborda el análisis de datos longitudinales y que son utilizados para el desarrollo de la presente investigación; en cuarto lugar, se presenta la metodología que se utilizó para el desarrollo de los objetivos, finalizando con los capítulos de resultados y discusión, así como el capítulo de conclusiones.

1.1. Descripción del Problema

En los estudios de análisis sensorial de alimentos, diferentes muestras de un producto alimenticio son evaluadas repetidamente a través del tiempo por consumidores, formándose así un clúster de observaciones correlacionadas, correspondiente a cada individuo participante. Estas características no se incorporan, a la fecha, en el modelado estadístico actual de vida de anaquel sensorial.
La presente investigación pretende dar respuesta al interrogante:
¿cómo incorporar la estructura tipo clúster y el patrón de asociación en el tiempo de los datos provenientes de análisis sensorial de alimentos, para realizar inferencias respecto a la vida de anquel sensorial de alimentos?

1.2. Justificación

En el desarrollo de un producto alimenticio intervienen muchos factores que influyen en su adquisición y/o consumo, como son características físicas, químicas y microbiológicas. Del mismo modo intervienen las características sensoriales que presenta el alimento y aquellas que puede percibir el consumidor.

Especialistas en las Ciencias de Alimentos como Hough y Garita (2012) señalan que, cuando se habla de vida de anaquel de alimentos, en la gran mayoría de los casos se hace referencia a la vida de anaquel sensorial de alimentos. De manera que, la estimación de la vida de anaquel sensorial se ha convertido en un tema de

investigación extenso, continuo y muy usado en la industria de alimentos, buscando ampliar los tiempos de comercialización al máximo, y al mismo tiempo, garantizar la frescura del producto.

Así pues, los productores de alimentos deben confiar en metodologías precisas para la estimación tanto de la vida de anaquel como de la vida de anaquel sensorial, ya que estimaciones no confiables repercuten en la perdida de alimento en el anaquel y/o el retiro de productos alimenticios antes de tiempo. Esto tiene consecuencias tanto económicas como sociales, considerando que en el mundo más de 820 millones de personas padecen hambre.

Por lo tanto, la presente investigación se convierte en una herramienta fundamental para mostrar otro enfoque estadístico que permite hacer inferencias en torno a la vida de anaquel sensorial de alimentos, establecer comparativos y tomar mejores decisiones.

Capítulo 2

Antecedentes

La vida de anaquel sensorial de un producto alimenticio se define como el período durante el cual las características sensoriales y de funcionamiento son los previstos por el fabricante. El producto es consumible o útil durante este período, proporcionando al usuario final las características sensoriales, de rendimiento y beneficios previstos.

Para estimar la vida de anaquel sensorial de un producto alimenticio, los científicos de alimentos recurren a la evaluación sensorial de alimentos, que se puede definir como una disciplina científica usada para evocar, medir, analizar e interpretar reacciones sensoriales ante características de los alimentos que son percibidas mediante los sentidos de la vista, el olfato, el gusto, el tacto y el oído. Para llevar a cabo evaluaciones sensoriales, se puede usar un grupo de evaluadores entrenados (panel entrenado) o un grupo de consumidores quienes realizan evaluaciones de las características sensoriales de un conjunto de muestras almacenadas a diferentes tiempos.

Existen dos estrategias principales para almacenar y evaluar productos durante un experimento de vida de anaquel sensorial, el diseño básico y el diseño en reversa. El diseño básico es el más simple, y consiste en almacenar un sólo lote grande del producto bajo condiciones normales y evaluar muestras del producto en varios tiempos de almacenamiento (Lawless y Heymann, 2010). Por otro lado, el diseño en reversa consiste en evaluar un conjunto de muestras con diferentes tiempos de almacenamiento, todas juntas en un sólo instante de evaluación.

Una vez recolectados los resultados del análisis sensorial del producto alimenticio, la metodología que se utiliza para estimar la vida de anaquel sensorial hace uso del

análisis de sobrevivencia de la siguiente manera:

- Los datos obtenidos se preparan para el análisis de acuerdo a lo ilustrado en la Tabla 2.1. Para cada consumidor la aceptación o rechazo es incluida en una fila que indica sí la persona aceptó (sí) o rechazó (no) la muestra de cada tiempo de almacenamiento. Debido a que los consumidores evalúan un número limitado de muestras con diferentes tiempos de almacenamiento, el tiempo exacto de almacenamiento en que la persona rechaza el producto no puede ser observado exactamente, lo que resulta en un dato censurado.

Tabla 2.1: Algunos tipos de censura para datos que provienen de análisis sensorial de alimentos.

Tiempo de Almacenamiento								
Consumidor	t_1	t_2	t_3	t_4	t_5	t_6	t_7	Censura
1	sí	sí	sí	sí	no	no	no	Intervalo: $t_4 - t_5$
2	sí	sí	sí	sí	sí	sí	sí	Derecha: $> t_7$
3	sí	sí	no	sí	no	no	no	Intervalo: $t_2 - t_5$
4	sí	no	sí	sí	no	no	no	Izquierda: $\leq t_5$
5	no	no	sí	sí	no	sí	sí	No se considera

Note de la Tabla 2.1 que a los consumidores 1 y 3 se les asigna un tiempo censurado como intervalo, pero es claro que los dos consumidores son bastante diferentes al tener en cuenta su consistencia en las respuestas a lo largo del tiempo de almacenamiento, como lo hace notar Hough (2010). En efecto, el consumidor 1 se considera un consumidor consistente en sus respuestas a lo largo del tiempo de almacenamiento, pero los consumidores 3 y 4 se consideran consumidores inconsistentes, uno en mayor o menor grado que el otro, de tal manera que se podría asociar un grado de inconsistencia al tiempo censurado.

- Dada la variable aleatoria, T, tiempo de almacenamiento en que el consumidor rechaza la muestra, se define la función de sobrevivencia, $S(t)$, como la probabilidad de que un consumidor acepte un producto almacenado un período de tiempo mayor a t, esto es: $S(t) = P(T > t)$. De manera alternativa, se define la función de distribución acumulada, $F(t)$ como la probabilidad de que un consumidor rechace un producto almacenado antes del tiempo t, esto es $F(t) = P(T \leq t)$.

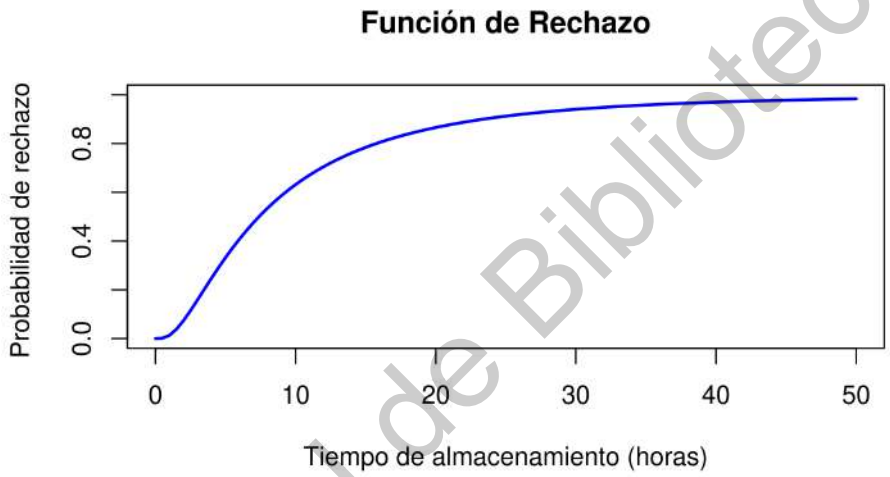
- Se estima la función de sobrevivencia, a través de la función de verosimilitud:

$$L = \prod_{i \in R} S(r_i) \cdot \prod_{i \in L} (1 - S(l_i)) \cdot \prod_{i \in I} (S(l_i) - S(r_i)),$$

donde R, L e I son los conjuntos de las observaciones censuradas a la derecha, izquierda y por intervalo, respectivamente.

- Las estimaciones de los parámetros de máxima verosimilitud son calculados y usados para graficar la proporción de consumidores que rechazan el producto contra el tiempo de almacenamiento, como se muestra en la Figura 2.1.

Figura 2.1: Ejemplo de un gráfico que representa la función de distribución acumulada $F(t)$



Hough et al. (2003) fueron los primeros en aplicar análisis de sobrevivencia para hacer inferencias respecto a la vida de anaquel sensorial de alimentos a partir del material desarrollado por Meeker y Escobar (1998). Se encuentran aplicaciones de esto en Jacobo et al. (2010) donde se estima la vida de anaquel sensorial de la pulpa de aguacate y mango procesada con alta presión hidrostática, de lo que se señala que un aspecto significativo de la metodología de análisis de sobrevivencia es la simplicidad del enfoque sensorial utilizado en el estudio; mientras que en Cruz et al. (2010) se predice la vida de anaquel sensorial de un yogurt probiótico.

En el análisis estadístico dominante, los datos originales se codifican en una variable continua de tiempo que hace necesaria la censura para su manejo estadístico y que implica un manejo opcional de grados de inconsistencia en el tiempo censurado asociado.

Capítulo 3

Hipótesis y Objetivos

3.1. Hipótesis

La estimación de la vida de anaquel sensorial de alimentos mediante la metodología de análisis de sobrevivencia ocasiona que se pueda perder información relevante en las estimaciones de la vida de anaquel sensorial de alimentos. Una metodología basada en datos longitudinales ofrece una alternativa para resolver este problema.

3.1.1. Objetivo general

Identificar estrategias alternativas de modelaje estadístico para el análisis de datos provenientes de evaluaciones sensoriales, explorando la inclusión de la inconsistencia y la autocorrelación en las respuestas de los evaluadores participantes en un estudio de sensorial de alimentos.

3.1.2. Objetivos específicos

- Proporcionar información comparativa entre dos enfoques estadísticos para estimar la vida de anaquel sensorial de alimentos.
- Analizar cómo la estimación de la vida de anaquel sensorial cambia con el nuevo enfoque de modelaje.

Capítulo 4

Fundamentación Teórica

4.1. Introducción

Los estudios transversales donde una sóla medida es tomada en un momento en el tiempo son ampliamente utilizados y existen diversas aplicaciones de los mismos. Sin embargo, éstos no poseen la capacidad de reflejar cambios en el tiempo. En contraste los estudios longitudinales miden una misma variable de interés repetidamente en el tiempo generando la posibilidad de estudiar y analizar patrones a lo largo del tiempo. Para el análisis de medidas repetidas que siguen una distribución normal aproximada, existen diferentes técnicas como se menciona en Zeger y Liang (1986).

Pero, al enfrentarse con respuestas que no siguen una distribución normal, hay un obstáculo mayor que se debe a la falta de una rica clase de distribuciones multivariadas, de manera que el uso de la función de verosimilitud para hacer estimaciones es limitado. Aún cuando la respuesta es binaria y el análisis de verosimilitud es posible, los cálculos pueden ser difíciles. Además, no incorporar la correlación de las respuestas puede conducir a una estimación incorrecta de los parámetros del modelo de regresión, particularmente cuando tales correlaciones son grandes (Zeger y Liang, 1986).

Debido a las complicaciones mencionadas antes, a finales de los años ochenta Zeger y Liang (1986) y Liang y Zeger (1986) propusieron una metodología para datos longitudinales discretos y continuos que usa el enfoque de cuasi-verosimilitud (Wedderburn, 1974). Concretamente se especifica que una función conocida de la esperanza marginal de la variable dependiente es una función lineal de las co-variables, y asume que la varianza es una función conocida de la media.

Además, se especifica una matriz de correlación para las observaciones de cada sujeto, de manera que esta configuración conduce a las Ecuaciones de Estimación Generalizadas (GEE, por sus siglas en inglés: Generalized Estimating Equations); mismas que dan estimadores consistentes de los coeficientes de regresión y de sus varianzas bajo suposiciones débiles sobre la correlación real entre las observaciones de un sujeto (Ware, 1985), sólo se requiere especificar correctamente los primeros dos momentos de la distribución subyacente de datos y trata la correlación como parámetros de ruido.

Esta última descripción de la metodología es la que aparece en Zeger y Liang (1986). Por otro lado, cabe mencionar que en Liang y Zeger (1986) se describe una metodología ligeramente diferente para un contexto más limitado. Ellos suponen una distribución marginal de la variable dependiente seguido de un modelo lineal generalizado y proponen un modelo de trabajo (conocido por su nombre en inglés "working") en que las observaciones para un individuo se suponen independientes, para luego generalizar el modelo de trabajo independiente (en inglés conocido como "Independence working model") y explicar la correlación, surgiendo así las ecuaciones de estimación generalizadas. Más detalles de este enfoque pueden verse en Wedderburn (1974).
El objetivo de este capítulo es dar los fundamentos teóricos necesarios para entender el enfoque de las ecuaciones de estimación generalizadas.

4.2. Notación

Sea Y_{ij} la variable respuesta y $\mathbf{x_{ij}}$ un vector de longitud p, de variables explicativas observadas en un tiempo t_{ij}, para observaciones $j = 1, \ldots, n_i$ sobre el sujeto $i = 1, \ldots, m$.

La media y la varianza de Y_{ij} se representan por $E(Y_{ij}) = \mu_{ij}$ y $Var(Y_{ij}) = \upsilon_{ij}$, respectivamente. El conjunto de respuestas repetidas para el sujeto i se guardan en un n_i-vector, $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})$ con media $E(\mathbf{Y}_i) = \boldsymbol{\mu_i}$ y matriz de varianzas-covarianzas $Var(\mathbf{Y_i}) = V_i$ de tamaño $n_i \times n_i$, donde el $jk-$elemento de V_i es la covarianza entre Y_{ij} y Y_{ik}, denotada por $Cov(Y_{ij}, Y_{ik}) = \upsilon_{ijk}$. R_i denota la matriz de correlación de tamaño $n_i \times n_i$ de $\mathbf{Y}_i$.

Las respuestas para todas las unidades se representan por $\mathbf{Y} = (\mathbf{Y}_i, \ldots, \mathbf{Y}_m)$, el cual es un $N-$vector con $N = \sum_{i=1}^{m} n_i$.

La mayoría de los análisis longitudinales se basan en un modelo de regresión como el modelo lineal,

$$\begin{aligned} Y_{ij} &= \beta_1 x_{ij1} + \beta_2 x_{ij2} + \ldots + \beta_p x_{ijp} + \epsilon_{ij} \\ &= \mathbf{x_{ij}}'\boldsymbol{\beta} + \epsilon_{ij}, \end{aligned}$$

donde $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ es un $p-$vector de coeficientes de regresión desconocidos y ϵ_{ij} es una variable aleatoria de media cero que representa la desviación de la respuesta a la predicción del modelo $\mathbf{x_{ij}}'\boldsymbol{\beta}$.

En notación matricial, la ecuación de regresión para el $i-$ésimo sujeto toma la forma:

$$\mathbf{Y_i} = X_i\boldsymbol{\beta} + \boldsymbol{\epsilon_i},$$

donde X_i es una matriz de orden $n_i \times p$, con $\mathbf{x_{ij}}$ en la $j-$ésima fila y $\boldsymbol{\epsilon_{ij}} = (\epsilon_{i1}, \ldots, \epsilon_{in_i})$.
En estudios longitudinales, la unidad experimental natural no es la medida individual Y_{ij}, sino $\mathbf{Y_i}$ el vector de medidas sobre un sujeto.
Los siguientes términos se usarán de manera indistinta, según el contexto: sujeto, unidad experimental, persona, animal e individuo.

4.3. Enfoques para análisis de datos longitudinales

Con una observación en cada unidad experimental, se está limitado a modelar el promedio poblacional de $\mathbf{Y}$, llamado la media marginal de la respuesta, pero si hay mediciones repetidas, hay varios enfoques diferentes que se pueden adoptar mismos que se describen a continuación.
El primero es modelar la media marginal como en un estudio transversal, pero como los valores repetidos probablemente no son independientes, el análisis marginal debe incluir hipótesis acerca de la correlación. Este enfoque se conoce como el enfoque de **modelo marginal**, y tiene la ventaja de que modela separadamente la media y la covarianza y puede hacer inferencias validas, aun cuando sea asumida una forma incorrecta de $V(\mathbf{Y_i})$.

Un segundo enfoque, el **modelo de efectos aleatorios**, el cual supone que la correlación surge entre las respuestas repetidas porque los coeficientes de regresión

varían de un individuo a otro. Aquí, se modela la esperanza condicional de Y_{ij} dados los coeficientes específicos del individuo, $\boldsymbol{\beta_i}$, por

$$E(Y_{ij}|\boldsymbol{\beta_i}) = (\mathbf{x'_{ij}}\boldsymbol{\beta_i}). \tag{4.1}$$

El último enfoque, conocido como **modelo de transición**, tiene por objetivo modelar la esperanza condicional, $E(Y_{ij}|Y_{ij-1}, \ldots, Y_{i1}, \mathbf{x_{ij}})$ como una función explícita de $\mathbf{x_{ij}}$ y de las respuestas pasadas.
Con los tres enfoques descritos, se puede modelar tanto la dependencia de la respuesta de las variables explicativas como la autocorrelación entre las respuestas. Es factible aplicar los demás enfoques pero por restricciones de tiempo en la presente investigación se utilizó sólo el enfoque marginal, en cambio se proponen éstos como trabajo a futuro.

4.4. Modelos lineales generales para datos longitudinales

En esta sección se desarrolla un marco del modelo lineal general para datos longitudinales, en el que las inferencias que se hacen sobre los parámetros de regresión de interés primario reconocen la estructura de correlación probable en los datos. Dos formas de alcanzar este objetivo es construir modelos paramétricos explícitos de la estructura de covarianza, o usar métodos de inferencia que son robustos la especificación incorrecta de la estructura de covarianza.

4.4.1. Modelo lineal general con errores correlacionados

Sea y_{ij}, $j = 1, \ldots, n$ la sucesión observada de medidas sobre el $i-$ésimo de m individuos, y t_j, $j = 1, \ldots, n$, los correspondientes tiempos en que son tomadas las medidas. Asociados con cada y_{ij} están los valores x_{ik}, $k = 1, 2, \ldots, p$, de p variables explicativas. Se asume que las y_{ij} son realizaciones de variables aleatorias Y_{ij} que siguen el modelo de regresión:

$$Y_{ij} = \beta_1 x_{ij1} + \ldots + \beta_p x_{ijp} + \epsilon_{ij}, \tag{4.2}$$

donde ϵ_{ij} son sucesiones aleatorias de longitud n, asociadas con cada uno de los m sujetos. Estas ϵ_{ij}, en el modelo lineal clásico, deben ser variables aleatorias mutuamente independientes distribuidas normal $N(0, \sigma^2)$, pero en el contexto longitudinal se espera que estas ϵ_{ij} estén correlacionadas dentro de los individuos.

Ahora considérese la formulación matricial de lo anterior, con tal de hacer un análisis formal del modelo lineal.

Sea $\mathbf{y_i} = (y_{i1}, \ldots, y_{in})$ la sucesión observada de medidas sobre el $i-$ésimo sujeto y $\mathbf{y} = (\mathbf{y_1}, \ldots, \mathbf{y_m})$ el conjunto completo de $N = nm$ medidas de m unidades. Sea X una matriz $N \times p$ de variables explicativas, con $\{n(i-1)+j\}-$ésima fila $(x_{ij1}, \ldots, x_{ijp})$.
Sea $\sigma^2 V$ una matriz diagonal por bloques con bloques no nulos de orden $n \times n$, cada una representando la matriz de varianza, para el vector de mediciones sobre un único sujeto. Entonces, el modelo lineal general para datos longitudinales trata a $\mathbf{y}$ como una realización de un vector aleatorio gaussiano multivariante $\mathbf{Y}$, con

$$\mathbf{Y} \sim MNV(X\boldsymbol{\beta}, \sigma^2 V). \tag{4.3}$$

Si se quiere usar el enfoque robusto para analizar los datos generados por el modelo (4.3), la estructura diagonal de bloque de $\sigma^2 V$ es crucial, porque se usa la replicación entre unidades para estimarla sin hacer suposiciones paramétricas sobre su forma. Dos posibles modelos para $\sigma^2 V$ pueden ser los siguientes.

- Supuesto de estructura de correlación uniforme
  En este modelo se supone que hay una correlación positiva ρ, entre cualesquiera dos medidas del mismo individuo. Matricialmente, esto se puede expresar así:

  $$V_0 = (1-\rho)I + \rho J, \tag{4.4}$$

  donde I denota la mariz identidad de $n \times n$, y J la matriz de orden $n \times n$ cuyas entradas son todas iguales a 1.

- Supuesto de estructura de correlación exponencial
  En este modelo, V_0 tiene como $jk-$ésimo elemento a $v_{jk} = Cov(Y_{ij}, Y_{ik})$, dado por:

  $$v_{jk} = \sigma^2 \exp(-\phi|t_j - t_k|). \tag{4.5}$$

  Se entiende que la correlación entre un par de mediciones en la misma unidad decae hacia cero a medida que aumenta la separación de tiempo entre las mediciones.

Un caso particular de este modelo es cuando las observaciones en los diferentes tiempos están igualmente espaciadas, entonces el $jk-$ésimo elemento $v_{jk} = Cov(Y_{ij}, Y_{ik})$

toma la forma:

$$v_{jk} = \sigma^2 \rho^{|j-k|}, \tag{4.6}$$

con $t_{j+1} - t_j = d$ y donde $\rho = \exp(-\phi d)$ es la correlación entre observaciones sucesivas sobre el mismo individuo.

En la investigación desarrollada no fue de interés particular modelar la estructura de correlación, por ello no se profundiza en el modelado paramétrico de las mismas, se refiere al lector a Diggle et al. (2002)

4.4.2. Estimación mínimos cuadrados ponderados

Volviendo a la expresión (4.3), considérese el problema de estimar los parámetros de regresión $\boldsymbol{\beta}$. El estimador de mínimos cuadrados ponderados de $\boldsymbol{\beta}$, con una matriz de pesos W, es el valor
$\widetilde{\boldsymbol{\beta}}_W$ que minimiza la forma cuadrática:

$$(\mathbf{y} - X\boldsymbol{\beta})' W (\mathbf{y} - X\boldsymbol{\beta}). \tag{4.7}$$

Al realizar algunos cálculos matriciales se obtiene que:

$$\widetilde{\boldsymbol{\beta}}_W = (X'WX)^{-1} X'W\mathbf{y}. \tag{4.8}$$

Debido a que $\mathbf{y}$ es una realización de un vector $\mathbf{Y}$ con $E(\mathbf{Y}) = X\boldsymbol{\beta}$, el estimador $\widetilde{\boldsymbol{\beta}}_W$ es insesgado, cualquiera que se la elección de W.
Además, dado que $Var(\mathbf{Y}) = \sigma^2 V$, entonces

$$Var(\widetilde{\boldsymbol{\beta}}_W) = \sigma^2 \{(X'WX)^{-1} X'W\} V \{WX(X'WX)^{-1}\}. \tag{4.9}$$

Si W es la matriz identidad, el estimador de mínimos cuadrados ponderados dado anteriormente se reduce a:

$$\widetilde{\boldsymbol{\beta}}_I = (X'X)^{-1} X'\mathbf{y}, \tag{4.10}$$

mismo que, en efecto coincide con el estimador de mínimos cuadrados ordinarios con

$$Var(\widetilde{\boldsymbol{\beta}}_I) = \sigma^2 (X'X)^{-1} X'VX(X'X)^{-1}. \tag{4.11}$$

4.4.3. Estimación máxima verosimilitud bajo hipótesis Gaussiana

Una estrategia para la estimación de parámetros en el modelo lineal general es considerar la estimación simultánea de los parámetros de interés, $\boldsymbol{\beta}$, y de los parámetros de covarianza, σ^2 y V_0, utilizando la función de verosimilitud. [1]
La log-verosimilitud para los valores observados $\mathbf{y}$ viene dada por:

$$L(\boldsymbol{\beta}, \sigma^2, V_0) = -0.5\{nm\log(\sigma^2) + m\log(|V_0|) + \sigma^2(\mathbf{y} - X\boldsymbol{\beta})'V^{-1}(\mathbf{y} - X\boldsymbol{\beta})\} \quad (4.12)$$

Para una V_0 dada, el estimador de máxima verosimilitud para $\boldsymbol{\beta}$ es el estimador de mínimos cuadrados ponderados:

$$\widehat{\boldsymbol{\beta}}(V_0) = (X'V^{-1}X)^{-1}X'V^{-1}\mathbf{y}. \quad (4.13)$$

El cual resulta de establecer $W = V^{-1}$ como la matriz de pesos y reemplazar en (4.8). Ahora, al sustituir esta última expresión en (4.12) queda:

$$L(\widehat{\boldsymbol{\beta}}(V_0), \sigma^2, V_0) = -0.5\{nm\log(\sigma^2) + m\log(|V_0|) + \sigma^2\mathrm{RSS}(V_0)\}, \quad (4.14)$$

donde

$$\mathrm{RSS}(V_0) = \{\mathbf{y} - X\widehat{\boldsymbol{\beta}}(V_0)\}'V^{-1}\{\mathbf{y} - X\widehat{\boldsymbol{\beta}}(V_0)\}.$$

Al derivar la expresión (4.14) respecto de σ^2 e igualando a cero, entonces el estimador de máxima verosimilitud para σ^2, con V_0 fija, es:

$$\widehat{\sigma}^2(V_0) = \mathrm{RSS}(V_0)/(nm), \quad (4.15)$$

Si se sustituye (4.13) y (4.15) en (4.12) se obtiene que la log-verosimilitud para V_0 está dada por:

$$L_r(V_0) = L\{\widehat{\boldsymbol{\beta}}(V_0), \widehat{\sigma}^2(V_0), V_0\} = -0.5m\{n\log\mathrm{RSS}(V_0) + \log(|V_0|)\}. \quad (4.16)$$

Finalmente, al maximizar $L_r(V_0)$ se obtiene $\widehat{V}_0$, sustituyendo en 4.13 y 4.15, los estimadores de máxima verosimilitud son:

$$\widehat{\boldsymbol{\beta}} \equiv \widehat{\boldsymbol{\beta}}(\widehat{V}_0) \text{ y } \widehat{\sigma}^2 \equiv \widehat{\sigma}^2(\widehat{V}_0).$$

[1] Recordar que V es una matrix diagonal por bloques, con bloques comunes no nulos iguales a V_0.

4.4.4. Estimación robusta de errores estándar

En este contexto, el término robusto significa que, si bien las estimaciones de los parámetros pueden derivarse haciendo una suposición de que la estructura de covarianza es de una forma particular, los errores estándar asociados se calculan de tal manera que sigan siendo válidos independientemente de la estructura de covarianza verdadera.

La idea esencial del enfoque robusto para inferencias en torno a $\boldsymbol{\beta}$ es usar el estimador de mínimos cuadrados generalizado $\widetilde{\boldsymbol{\beta}}_W$, definido en (4.8) como $\widetilde{\boldsymbol{\beta}}_W$ junto con una matriz de varianza estimada:

$$\widehat{R}_W = \{(X'WX)^{-1}X'W\}\widehat{V}\{WX(X'WX)^{-1}\}, \tag{4.17}$$

donde $\widehat{V}$ es consistente para V, cualquiera que sea la verdadera estructura de covarianza.

En (4.17) el parámetro de escala σ^2 fue re-absorbido en V. Luego, para la inferencia se considera que

$$\widetilde{\boldsymbol{\beta}}_W \sim MNV(\boldsymbol{\beta}, \widehat{R}_W) \tag{4.18}$$

En este enfoque, a W^{-1} se le llama **matriz de varianza de trabajo**, para distinguir esta de la matriz de varianza verdadera, V.

En general, para W^{-1} se usa una forma simple, que trate de capturar la estructura cualitativa de V. Sin embargo, la diferencia crucial entre este enfoque y el enfoque de modelar paramétricamente la estructura de covarianza, es que una estimación pobre de W afectará solamente la eficiencia de las inferencias para $\boldsymbol{\beta}$, no su validez (Diggle et al., 2002). Intervalos de confianza y pruebas de hipótesis derivados del supuesto (4.18) son asintóticamente correctas sin importar la forma verdadera de V (White, 1982; Liang y Zeger, 1986).

El enfoque robusto suele ser satisfactorio cuando los datos consisten en secuencias cortas, esencialmente completas, de mediciones observadas en un conjunto común de tiempos en muchas unidades experimentales, y se tiene cuidado en la elección de una matriz de varianza de trabajo.
Por otro lado, por el hecho de que la estructura de correlación puede ser difícil de identificar en la práctica, puede ser de interés preguntarse qué tanta eficiencia se

puede perder al utilizar una W incorrecta. Para esto, el lector interesado puede referirse a Diggle et al. (2002). Por otro lado, si el interés principal del estudio longitudinal recae principalmente en identificar la estructura de correlación más que modelar la relación entre la esperanza marginal de la variable respuesta y covariables, se puede consultar Diggle et al. (2002). En el caso de la presente investigación, se dió poca atención a los mecanismos por los cuales la asociación dentro de los sujetos pudo haber surgido; en cambio, se hace uso del criterio de información de cuasi-verosimilitud (QIC, por sus siglas en inglés Quasi-likelihood information criterion) para la elección del tipo de estructura de correlación más adecuada, como lo describe Pan (2001, 2002) y que se desarrolla en la sección 4.6.

4.5. Cuasi-verosimilitud

Esta sección describe brevemente los aspectos del enfoque de cuasi-verosimilitud que se utiliza en el desarrollo de las ecuaciones de estimación generalizadas para el análisis de datos longitudinales.

En el enfoque de cuasi-verosimilitud, se especifican sólo las relaciones entre la media de la respuesta y las co-variables, y entre la media y la varianza. Esta extensión es importante en el contexto de datos longitudinales porque, excepto para respuestas aproximadamente normales, existen pocas elecciones para la distribución conjunta de los valores repetidos para cada sujeto (Zeger y Liang, 1986). De manera que al utilizar el enfoque de cuasi-verosimilitud y especificando solo la estructura entre media y covarianza, pueden desarrollarse métodos que son aplicables a diferentes tipos de variables respuestas.

Considere $n_i = 1$ para todo i, entonces se asume cada dato de los individuos como un escalar, luego el índice j desaparecerá. Se define $\boldsymbol{\mu_i}$ como la esperanza de $\mathbf{Y_i}$ y se supone que:

$$\boldsymbol{\mu_i} = h(X_i\boldsymbol{\beta}), \tag{4.19}$$

donde $\boldsymbol{\beta}$ es el vector de parámetros y la inversa de h es la función de enlace. En el enfoque de cuasi-verosimilitud, la varianza de $\mathbf{Y_i}$ se expresa como una función conocida g, de la esperanza $\boldsymbol{\mu_i}$, esto es:

$$Var(\mathbf{Y_i}) = g(\boldsymbol{\mu_i})/\phi, \tag{4.20}$$

donde ϕ es un parámetro de escala. El enfoque de cuasi-verosimilitud recae sobre

las inferencias de $\boldsymbol{\beta}$. Por lo tanto, ϕ se trata como un parámetro de ruido.

El estimador de cuasi-verosimilitud es la solución del sistema de ecuaciones score-like:

$$S_k(\boldsymbol{\beta}) = \sum_{i=1}^{m} \frac{\partial \boldsymbol{\mu_i}}{\partial \beta_k} Var(\mathbf{Y_i})^{-1}(\mathbf{Y_i} - \boldsymbol{\mu_i}) = 0, \quad k = 1, \ldots, p. \tag{4.21}$$

Nota 1. Las ecuaciones en (4.21) son en efecto las ecuaciones score para $\boldsymbol{\beta}$ cuando $\mathbf{Y_i}$ sigue una distribución de la familia exponencial. La solución de estas se puede obtener por mínimos cuadrados iterativamente reponderados.

4.6. Modelos lineales generalizados para datos longitudinales

En esta sección se discute el modelo marginal, como una extensión de modelos lineales generalizados, donde también caben los modelos de efectos aleatorios y de transición, mencionados anteriormente.

4.6.1. Modelos marginales

En un modelo marginal, la regresión de la respuesta sobre las variables explicativas se modela por separado de la correlación dentro del individuo. En la regresión, se modela la esperanza marginal $E(Y_{ij})$, en función de variables explicativas. La esperanza marginal se refiere a la respuesta promedio sobre la subpoblación que comparte un valor común de x. Específicamente, un modelo marginal tiene los siguientes supuestos.

1. La esperanza marginal de la respuesta, $E(Y_{ij}) = \mu_{ij}$, depende de variables explicativas, x_{ij}, por $h(\mu_{ij}) = \mathbf{x_{ij}}'\beta$ donde h es una función de enlace conocida como logit para respuestas binarias o log para conteos;

2. La varianza marginal depende de la media marginal según $Var(Y_{ij}) = v(\mu_{ij})\phi$, donde v es una función de varianza conocida y ϕ es un parámetro de escala que puede ser necesario estimar.

3. La correlación entre Y_{ij} y Y_{ik} es una función de las medias marginales y quizás de los parámetros adicionales $\boldsymbol{\alpha}$, es decir, $Corr(Y_{ij}, Y_{ik}) = \rho(\mu_{ij}, \mu_{ik}; \boldsymbol{\alpha})$ donde $\rho(\cdot)$ es una función conocida.

En este enfoque marginal, se especifican los dos primeros momentos de la respuesta para cada individuo. En el caso de respuestas que se distribuyan normal, los primeros dos momentos están completamente determinados por la verosimilitud, pero no es el caso con otros miembros de la familia de los modelos lineales generalizados. Para especificar la verosimilitud completamente, es necesario formular hipótesis adicionales de momentos de mayor orden.

Aún con hipótesis adicionales la verosimilitud es con frecuencia difícil de manejar y se involucran muchos parámetros de ruido además de $\boldsymbol{\alpha}$ y $\boldsymbol{\beta}$. Es aquí donde se utiliza el enfonde de las ecuaciones de estimación generalizadas o GEE.

4.6.2. Ecuaciones de estimación generalizadas

Para aplicar el enfoque de cuasi-verosimilitud al análisis de datos longitudinales, se debe considerar la media y la covarianza del vector de respuestas $\mathbf{Y_i}$, para el $i-$ésimo individuo. Entonces, se procede como en la sección 4.5 pero adicionando $R_i(\boldsymbol{\alpha})$ la matriz de correlación de trabajo de tamaño $n_i \times n_i$, para cada $\mathbf{Y_i}$. Esta última matriz, $R_i(\boldsymbol{\alpha})$, se asume que está completamente especificada por un vector de parámetros desconocidos $\boldsymbol{\alpha}$, que es el mismo para todos los individuos.

Al seguir el enfoque de cuasi-verosimilitud, la matriz de covarianza de trabajo para $\mathbf{Y_i}$ está dada por:

$$V_i = A_i^{1/2} R_i(\boldsymbol{\alpha}) A_i^{1/2}/\phi, \tag{4.22}$$

donde A_i es una matriz diagonal de tamaño $n_i \times n_i$, con $g(\mu_{ij})$ como el $j-$ésimo elemento en la diagonal.

La extensión de la ecuación (4.21) al caso multivariado viene dado por:

$$\sum_{i=1}^{m} D_i' Var(\mathbf{Y_i})^{-1} S_i = 0, \text{ con } S_i = (\mathbf{Y_i} - \boldsymbol{\mu_i}) \text{ y } D_i = \left(\frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta}}\right) \tag{4.23}$$

Nota 2. La ecuación (4.23) se reduce a la ecuación de cuasi-verosimilitud (4.21), cuando $n_i = 1$, para todo i. Además, para respuestas Gausianas, la ecuación (4.23) es la ecuación score para $\boldsymbol{\beta}$. De manera más general, $U_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = D_i' Var(\mathbf{Y_i})^{-1} S_i$ es equivalente a la función de estimación que fue sugerida por Wedderburn (1974), con la complicación adicional de que éstas no sólo dependen de $\boldsymbol{\beta}$ sino también de $\boldsymbol{\alpha}$, dado que $Var(\mathbf{Y_i}) = Var(\mathbf{Y_i}; \boldsymbol{\beta}, \boldsymbol{\alpha})$.

Mientras las ecuaciones de estimación ahora dependen tanto de $\boldsymbol{\alpha}$ como de $\boldsymbol{\beta}$, éstas pueden ser re-expresadas como una función de $\boldsymbol{\beta}$ nada más, al reemplazar $\boldsymbol{\alpha}$ en las ecuaciones 4.22 y 4.23 por un estimador $m^{1/2}-$consistente, $\widehat{\boldsymbol{\alpha}}(Y, \boldsymbol{\beta}, \phi)$ y entonces se sustituye ϕ en $\widehat{\boldsymbol{\alpha}}$ por un estimador $m^{1/2}-$consistente $\widehat{\phi}(\mathbf{y}, \boldsymbol{\beta})$.

En consecuencia, para cualquier $R_i(\boldsymbol{\alpha})$, el estimador $\widehat{\boldsymbol{\beta}}_R$ de $\boldsymbol{\beta}$ se define como la solución de

$$\sum_{i=1}^{m} U_i\{\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}}[\boldsymbol{\beta}, \widehat{\phi}(\boldsymbol{\beta})]\} = 0. \tag{4.24}$$

Por lo tanto, para cualquier $R_i(\boldsymbol{\alpha})$ dada, la estimación de $\boldsymbol{\beta}$, denotada por $\hat{\boldsymbol{\beta}}_R$, está definida como la solución de:

$$\sum_{i=1}^{m} U_i\{\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}[\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta})]\} = 0.$$

Bajo condiciones de regularidad leves, Liang y Zeger (1986) mostraron que cuando $m \to \infty$, $\widehat{\boldsymbol{\beta}}_R$ es un estimador consistente de $\boldsymbol{\beta}$ y que $m^{1/2}(\widehat{\boldsymbol{\beta}}_R - \boldsymbol{\beta})$ es asintóticamente Gaussiana multivariante con matriz de covarianza V_R dada por

$$\begin{aligned} V_R &= \lim_{m\to\infty} m \left(\sum_{i=1}^{m} D_i' V_i^{-1} D_i\right)^{-1} \left[\sum_{i=1}^{m} D_i' V_i^{-1} \mathrm{cov}(\mathbf{y_i}) V_i^{-1} D_i\right] \left(\sum_{i=1}^{m} D_i' V_i^{-1} D_i\right)^{-1} \\ &= \lim_{m\to\infty} m(V_1^{-1} V_0 V_1^{-1}), \end{aligned} \tag{4.25}$$

donde la covarianza de $\mathbf{y_i}$ es la covarianza real en lugar de la supesta. V_R puede ser estimada consistentemente sinevaluar de forma directa $\mathrm{Cov}(\mathbf{y_i})$. Esto se logra al reemplazar $\mathrm{Cov}(\mathbf{y_i})$ por $S_i S_i'$, además de $\boldsymbol{\alpha}, \boldsymbol{\beta}$ y ϕ por sus estimadores en (4.25).

Para solucionar las GEE para $\widehat{\boldsymbol{\beta}}_R$, se resuelve iterativamente para el coeficiente de regresión, la correlación y el parámetro de escala $\boldsymbol{\alpha}$ y ϕ (Lian y Zeger, 1986).

Hay varias elecciones para la matriz de correlación de trabajo $R_i(\alpha)$. A continuación se enuncian algunas de las más usadas y que aparecen en los paquetes estadísticos para el análisis de datos longitudinales.

Estructura de correlación independiente: Se asume que la correlación entre las

medidas de un mismo clúster es cero. Específicamente, se asume que:

$$\text{corr}(Y_{it}, Y_{it'}) = 0, \quad t \neq t'.$$

Estructura de correlación auto-regresiva: Se asume que la correlación entre medidas consecutivas de un clúster (sujeto o individuo) decrece conforme se incrementa la separación en el tiempo. Específicamente, se asume que:

$$\text{corr}(Y_{it}, Y_{it'}) = \alpha^{|t-t'|}, \quad t \neq t'.$$

Estructura de correlación no estructurada: Se asume que todas las correlaciones deben ser estimadas de los datos de forma independiente.

$$\text{corr}(Y_{it}, Y_{it'}) = \alpha_{tt'}, \quad t \neq t'.$$

Estructura de correlación intercambiable: esta estructura asume que todos los pares de correlaciones en el mismo clúster tienen el mismo valor, esto es:

$$\text{corr}(Y_{it}, Y_{it'}) = \alpha, \quad t \neq t'.$$

No hay una prueba estadística que pueda evaluar cuál es la estructura de correlación de trabajo correcta, pero una vía para la elección de la más adecuada es entender el diseño del estudio y los datos. Por ejemplo, si el número de unidades por clúster es pequeño en un diseño balanceado y completo, entonces una matriz no estructurada es la recomendada (Pan, 2001, 2002). Por otro lado, si existe mucha incertidumbre en cómo seleccionar la estructura más adecuada a partir de información empírica, el estadístico QIC puede ser usado para comparar estructuras de correlación, prefiriendo al QIC más pequeño. El marco teórico respecto del QIC, se desarrolla a continuación.

4.6.3. Criterio de información de Cuasi-verosimilitud

Se conoce que en el contexto de los modelos lineales generalizados, el criterio de información de Akaike (AIC, por sus siglas en inglés que traducen Akaike Information Criterion) juega un papel muy importante en la selección del modelo; no obstante este hace uso de la función de verosimilitud y, por lo tanto no es posible usarlo en el enfoque de análisis de datos longitudinales. En respuesta a esto Pan (2001) propuso una modificación de aquel, donde la verosimilitud es reemplazada por la cuasi-verosimilitud y se hace un ajuste apropiado para el término de penalización.

Para una variable de respuesta Y y un vector de co-variables X, bajo un modelo lineal general $g(\mu) = \beta' X$ con $g(\cdot)$ la función de enlace $\mu = E(Y)$, el AIC viene dado por:

$$\text{AIC} = 2p - 2\ln(L), \tag{4.26}$$

donde L es la verosimilitud y p es la dimensión de $\boldsymbol{\beta}$.

Pan (2001), propuso reemplazar la verosimilitud L en (4.26) por la cuasi-verosimilitud Q bajo el modelo de trabajo independiente (the working independence model), esto es:

$$\text{QIC} = -2Q(\widehat{\mu}; I) + 2\text{traza}(\widehat{\Omega}_I^{-1}\widehat{V}_R), \tag{4.27}$$

donde I representa la estructura de covarianza independiente usada para calcular la cuasi-verosimilitud. Además, $\widehat{\mu} = g^{-1}(x\widehat{\boldsymbol{\beta}})$ con g^{-1} la inversa de la función de enlace. Las estimaciones de los coeficientes $\widehat{\boldsymbol{\beta}}$ y el estimador robusto de la varianza $\widehat{V}_R$, se obtienen a partir de una estructura de covarianza general de trabajo R. El estimador de varianza $\widehat{\Omega}_I^{-1}$ se obtiene bajo la hipótesis de una estructura de correlación independiente.
La cuasi-verosimilitud Q en (4.27) tiene la forma general

$$Q(\mu) = \int_y^{\mu} \frac{y - t}{\phi V(t)} dt, \tag{4.28}$$

donde ϕ es un parámetro de dispersión. La varianza de las observaciones respuesta es una función de la media μ y se denota por $V(\mu)$. El valor de $V(\mu)$ es dado en la Tabla 4.1 para alguna de las distribiciones comúnmente usadas de la familia exponencial. Al sustituir $V(\mu)$ en (4.28) con el valor correspondiente en la Tabla 4.1, se puede calcular el valor de la cuasi-verosimilitud $Q(\mu)$ el cual también es listado en la Tabla 4.1.

Tabla 4.1: Funciones de varianza y cuasi-verosimilitud para distribuciones comúnmente usadas en la familia exponencial.

Distribución	$V(\mu)$	$Q(\mu)$
Bernoulli	$\mu(1-\mu)$	$y\ln(\frac{\mu}{1-\mu}) + \ln(1-\mu)$
Normal	1	$-\frac{1}{2}\sum(y-\mu)^2$
Poisson	μ	$y\ln(\mu) - \mu$
Gamma	μ^2	$-(y/\mu + \ln(\mu))$
Binomial negativa	$\mu + \mu^2$	$y(\ln(\mu) - 2\ln(\mu+1))$
Gaussiana inversa	μ^3	$-\frac{y}{2\mu^2} + \frac{1}{\mu}$

4.7. Técnicas de diagnóstico para ecuaciones de estimación generalizadas

Una fuente para detectar problemas en el ajuste de los modelos de regresión son los gráficos de probabilidad semi-normal con envolventes simuladas, los cuales se estudian ampliamente en un contexto clásico en Atkinson (1981). Estos gráficos fueron extendidos para diagnosticar el ajuste de modelos de regresión con medidas repetidas en Venezuela et al. (2007), mismos que se describirán en la siguiente sección.

Así mismo, se han extendido una serie de medidas de bondad de ajuste al contexto de análisis de datos longitudinales (Zheng, 2000), y se presentarán en la sección 4.7.2.

4.7.1. Gráficos de probabilidad semi-normales

Los gráficos de probabilidad semi-normales con envolturas basadas en simulaciones,y fueron desarrollados por Venezuelta et al. (2007). Son útiles para identificar valores atípicos y examinar la adecuación de los modelos ajustados cuando las distribuciones marginales pertenecen a la familia exponencial.

El algoritmo para construir los gráficos de probabilidad semi-normales con envolturas basadas en simulaciones en el contexto GEE (Venezuela et al., 2007) difiere del algoritmo original de Atkinson (1981). En específico, éste último basó sus cálculos en un residuo de Jackknife, mientras que Venezuela et al. (2007) propusieron el uso del residuo común estandarizado.

El algoritmo propuesto por Venezuela et al. (2007), es el siguiente:

1. Estimar el modelo GEE con los datos originales.
2. Para $k = 1, \ldots, K$ (por defecto $K = 25$)
    a) Para cada cluster i, $i = 1, 2, \ldots, n$ simular un vector de respuestas de dimension $T \times 1$ (T es el tamaño del cluster, $t = 1, \ldots, T$) usando el vector media estimado y la matriz de correlación sobre la base del modelo ajustado con los datos originales.
    b) Para las respuestas simuladas ajustar el mismo modelo con las mismas covariables.

c) Estimar los residuales estandarizados S_{it} y ordenarlos. Sea S_{mk} el $m-$ésimo valor absoluto ordenado del residual estandarizado perteneciente a la $k-$ésima observacion, $m = n \cdot T$.

d) Calcular el mínimo de los valores absolutos de los residuales, denotado por $|S|_{it}$. Análogamente calcular la mediana $|S|_{0.5,k}$, y el máximo $|S|_{n\cdot T,k}$.

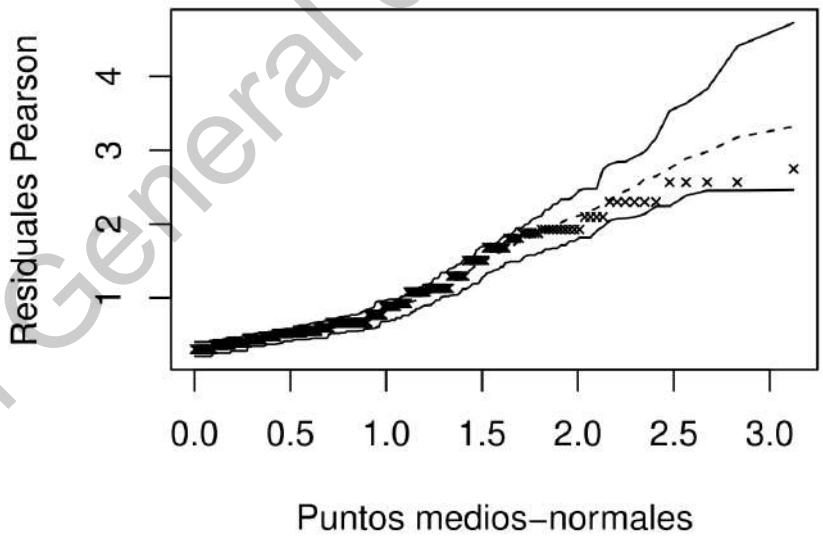
e) Graficar $|S|_{1,k}, |S|_{2,k}, |S|_{3,k}, \cdots |S|_{n\cdot T,k}$ y los valores absolutos ordenados de los residuales estandarizados de los datos originales versus los puntos half-normal

$$\Phi^{-1}\left(\frac{l + nT - \frac{1}{8}}{2nT + \frac{1}{2}}\right),$$

donde $\Phi(\cdot)$ es la función de distribución acumulada de la distribución normal estandar ($l = 1, 2, \ldots, nT$).

Un ejemplo de este tipo de gráficos se puede ver en la Figura 4.1

Figura 4.1: Ejemplo de un gráfico de probabilidad semi-normal.



Nota 3. La existencia de puntos que caen por fuera de la envoltura simulada, indica que el modelo ajustado es inapropiado. Sí hay puntos atípicos, éstos aparecerán en la parte superior del gráfico half-normal, separados de los demás puntos.

4.7.2. Medidas de bondad de ajuste aplicadas a datos longitudinales

Para complementar el diagnóstico de los modelos de regresión marginales, se requieren también medidas de bondad de ajuste. En este caso, se presentan dos extensiones de medidas de bondad de ajuste aplicadas en modelos lineales generalizados, como son la reducción proporcional de la entropía, el coeficiente de determinación R^2 y el coeficiente de correlación de concordancia, discutidos ampliamente en Zheng (2000).

Medidas de reducción proporcional en variación H_{marg} y R^2_{marg}

Sea $\pi_{tk} = P(Y_t = k|X)$ la probabilidad basada en el modelo de que una respuesta categórica en el tiempo t sea igual a k, y $\hat{\pi}_{tk}$ su estimación, además sea, $\alpha_k = P(Y = k)$ la probabilidad marginal de la respuesta k y $\hat{\alpha}_k$ su estimación. La extensión de la medida de reducción proporcional en entropía, H para un modelo marginal está dada por:

$$H_{marg} = 1 - \frac{\sum_{t=1}^{T}\sum_{i=i}^{n}\sum_{k=1}^{K}\hat{\pi}_{itk}\log(\hat{\pi}_{itk})}{nT\sum_{k=1}^{K}\hat{\alpha}_k\log(\hat{\alpha}_k)}. \tag{4.29}$$

La medida H_{marg} puede ser interpretada como la reducción proporcional en entropía debido al modelo de interés y se reduce a su análogo en el contexto de un modelo lineal general para cuando $T = 1$. La medida H_{marg} puede tomar valores en el intervalo $[0, 1]$. El valor cero indica que no existe asociación entre la respuesta y las variables predictoras y el valor de uno indica cuando la variable respuesta en cada punto del tiempo cae en una categoría con probabilidad estimada 1, lo que significa una predicción perfecta si el modelo ajustado es correcto (Zheng, 2000).

En Zheng (2000) se presenta una medida análoga al coeficiente de determinación R^2, el cual toma la forma

$$R^2_{marg} = 1 - \frac{\sum_{t=1}^{T}\sum_{i=1}^{n}(Y_{it} - \hat{Y}_{it})^2}{\sum_{t=1}^{T}\sum_{i=1}^{n}(Y_{it} - \bar{Y})^2}, \quad \text{con } \bar{Y} = \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}Y_{it}. \tag{4.30}$$

La medida R^2_{marg} puede tomar valores en el intervalo $[0, 1]$. El valor de 1 indica que existe una predicción perfecta; el valor de 0 inidica cuando no existe asociación entre la variable respuesta y las variables predictoras.

Coeficiente de correlación de concordancia r_c

Este coeficiente resume la concordancia entre la respuesta y el valor ajustado de manera simultánea en todos los puntos en el tiempo.

Dado $\bar{\hat{Y}} = \frac{1}{Tn}\sum_{i=1}^{n}\sum_{t=1}^{T}\hat{Y}_{it}$, el coeficiente de correlación de concordancia r_c, para un modelo marginal se define como:

$$r_c = \frac{2\sum_{i=1}^{n}\sum_{t=1}^{T}(Y_{it}-\bar{Y})(\hat{Y}_{it}-\bar{\hat{Y}})}{\sum_{i=1}^{n}\sum_{t=1}^{T}(Y_{it}-\bar{Y})^2+\sum_{i=1}^{n}\sum_{t=1}^{T}(\hat{Y}_{it}-\bar{\hat{Y}})^2}. \tag{4.31}$$

El valor de este coeficiente oscila entre -1 y 1, al igual que su contraparte en GLM, indicando un ajuste perfecto cuando toma el valor de 1 y un ajuste perfecto inverso cuando es igual a -1.

4.7.3. Bondad de ajuste para GEE con respuestas binarias repetidas

Horton et al. (1999) desarrollaron un estadístico que utiliza deciles de riesgo predichos que pueden ser vistos como una extensión de lo propuesto por Hosmer y Lemeshow (1980) para regresión logística ordinaria. El estadístico tiene una distribución Chi-Cuadrado aproximada cuando el modelo es correctamente especificado y es apropiado tanto para co-variables categóricas como continuas.
Los autores formaron G grupos de aproximadamente igual tamaño basados en las probabilidades estimadas

$$\hat{\pi}_{ij} = \left(\frac{e^{\hat{\beta}_0+x'_{ij}\hat{\boldsymbol{\beta}}_1}}{1+e^{\hat{\beta}_0+x'_{ij}\hat{\boldsymbol{\beta}}_1}}\right).$$

Se definen $(G-1)$ grupos indicadores

$$I_{ijg} = \begin{cases} 1 & \text{si } \hat{\pi}_{ij} \text{ está en el grupo } g \\ & \qquad\qquad g = 1, 2, \cdots, G-1 \\ 0, & \text{en otro caso.} \end{cases}$$

Los autores consideraron el modelo alternativo

$$\text{logit}(\pi_{ij}) = \beta_0 + X'_{ij}\hat{\boldsymbol{\beta}} + \gamma_1 I_{ij,1} + \cdots + \gamma_{G-1} I_{ij,G-1},$$

y para probar la bondad de ajuste se contrasta

$$H_0 : \gamma_1 = \cdots = \gamma_{G-1} = 0,$$

usando el estadístico score

$$\chi^2 = \mathbf{u}_2(\hat{\boldsymbol{\beta}}, 0)'\{\widehat{\text{Var}}[\mathbf{u}_2(\hat{\boldsymbol{\beta}}, 0)]\}^{-1}\mathbf{u}_2(\hat{\boldsymbol{\beta}}, 0)$$

Aquí χ^2 se distribuye como χ^2_{G-1} bajo el modelo nulo donde:

$$\mathbf{u}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \begin{bmatrix} \mathbf{u}_1(\boldsymbol{\beta}, \boldsymbol{\gamma}) \\ \mathbf{u}_2(\boldsymbol{\beta}, \boldsymbol{\gamma}) \end{bmatrix} = \sum_{i=1}^{m} \begin{bmatrix} \mathbf{D}'_{1i}\mathbf{V}_i^{-1}[\mathbf{Y}_i - \boldsymbol{\pi}_i(\boldsymbol{\beta}, \boldsymbol{\gamma})] \\ \mathbf{D}'_{2i}\mathbf{V}_i^{-1}[\mathbf{Y}_i - \boldsymbol{\pi}_i(\boldsymbol{\beta}, \boldsymbol{\gamma})] \end{bmatrix}, \tag{4.32}$$

y

$$\mathbf{D}_{1i} = \frac{\partial[\boldsymbol{\pi}_i(\boldsymbol{\beta}, \boldsymbol{\gamma})]}{\partial \boldsymbol{\beta}}, \quad \mathbf{D}_{2i} = \frac{\partial[\boldsymbol{\pi}_i(\boldsymbol{\beta}, \boldsymbol{\gamma})]}{\partial \boldsymbol{\gamma}}, \quad \boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_{G-1}]'.$$

El vector $\boldsymbol{\beta}$ se obtiene al solucionar $\mathbf{u}_1(\hat{\boldsymbol{\beta}}, 0) = 0$. Los autores notaron que a pesar que el estadístico es fácilmente interpretable, puede perder importantes desviaciones del ajuste y puede sólo evaluar de forma directa covariables que están en el modelo. En específico el estadístico podría tener bajo poder en muestras pequeñas y bajo poder para detectar alternativas específicas, pero podría tener amplio poder para detectar una variedad de alternativas generales.

Capítulo 5

Metodología

Los datos usados en este trabajo provienen de estudios sensoriales de alimentos, son de dominio público y se pueden encontrar en Hough (2010); a su vez, pueden ser descargados en archivo .xlsx de la página web del editor.

Para la generación de la base de datos, los tecnólogos de alimentos, seleccionan una muestra de consumidores los cuales prueban ciegamente un conjunto de muestras de cierto alimento, con diferentes tiempos de almacenamiento, De manera concreta se les pide respondan sí o no a la pregunta ¿normalmente consumiría este producto?. Esto genera un conjunto de respuestas como las mostradas en la Tabla 5.1.

Tabla 5.1: Ilustración de un conjunto de datos proveniente del análisis sensoriales de alimentos.

	Tiempos de almacenamiento					
Consumidor	t_0	t_1	t_2	t_3	$\cdots$	t_n
1	no	no	sí	sí	$\cdots$	no
2	sí	sí	sí	sí	$\cdots$	no
3	sí	no	sí	no	$\cdots$	si
4	no	sí	sí	no	$\cdots$	no
5	sí	sí	sí	sí	$\cdots$	no
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	sí	no	no	no	$\cdots$	no

Para analizar los datos longitudinales provenientes del análisis sensorial de alimentos se llevó a cabo tres etapas:

Etapa 1: Ajuste del modelo estadístico

En esta etapa se ajustó un modelo estadístico para la variable aleatoria de interés, mediante el enfoque de modelos marginales, y se aplicó el cirterio de información de cuasi-verosimilitud para determinar la estructura de correlación de trabajo más adecuada.

En un modelo marginal, el interés es modelar la esperanza poblacional de cierta variable aleatoria como una función de variables explicativas. En el caso del presente estudio, el interés recayó sobre la variable aleatoria $\mathbf{Y}$: respuesta del consumidor a la pregunta ¿consumiría normalmente este producto?

Esta es una variable aleatoria binaria tal que

$$\mathbf{Y} = 0, \text{ si la respuesta es sí}$$
$$\mathbf{Y} = 1, \text{ si la respuesta es no}$$

En el estudio se contó con la variable explicativa:

$$t\text{: tiempo de almacenamiento (medido en horas).}$$

A su vez, se codificó una variable explicativa adicional, que no es medida directamente en el estudio sensorial, sino que fue identificada como una característica que presentan los consumidores en la tendencia de sus respuestas. En específico, se clasificó a los consumidores en dos categorías:

-consumidores consistentes: aquellos consumidores que una vez rechazan el producto en el i-ésimo tiempo, siguen rechazando en los $i+1$-ésimos tiempos subsecuentes.

-consumidores inconsistentes: aquellos consumidores que rechazan en el tiempo cero horas de almacenamiento (producto fresco) y/o rechazan en el i-ésimo tiempo, pero en los tiempos subsecuentes pueden volver a aceptar.

Entonces se utilizó la variable binaria consistencia del consumidor, que se denota por C, y

$$C = 0, \text{ si el consumidor es consistente}$$
$$C = 1, \text{ si el consumidor es inconsistente}$$

Se adoptó una función de enlace logit, quedando explícitamente el modelo marginal:

$$\log\left(\frac{E[Y_{ij}]}{(1-E[Y_{ij}])}\right) = \beta_0 + \beta_1 * t_j + \beta_2 * c + \beta_3 * (c * t_j). \tag{5.1}$$

Esto implica:

$$E[Y_{ij}] = \mu_{ij} = \frac{\exp(\beta_0 + \beta_1 * t_j + \beta_2 * c + \beta_3 * (c * t_j))}{1 + \exp(\beta_0 + \beta_1 * t_j + \beta_2 * c + \beta_3 * (c * t_j))}, \tag{5.2}$$

y dado que:

$$E[Y_{ij}] = \mu_{ij} = P(Y_{ij} = 1 | T = t_j, C = c), \tag{5.3}$$

entonces:

$$P(Y_{ij} = 1 | T = t_j, C = c) = \frac{\exp(\beta_0 + \beta_1 * t_j + \beta_2 * c + \beta_3 * (c * t_j))}{1 + \exp(\beta_0 + \beta_1 * t_j + \beta_2 * c + \beta_3 * (c * t_j))}. \tag{5.4}$$

donde Y_{ij} representa la respuesta del $i-$ésimo consumidor en el $j-$ésimo tiempo de almacenamiento, con $i = 1, 2, \ldots, 50$ y $j = 1, 2, \ldots, 7$.

El modelo marginal fue apropiado para examinar la probabilidad de rechazo de los consumidores como una función del tiempo de almacenamiento, el tipo de consumidor y la interacción de este con el tiempo de almacenamiento. Se usó la función *geeglm* del paquete *geepack* del software **R**, con las estructuras de correlación independiente, no estructurada, intercambiable y autoregresiva de orden 1, para estimar los coeficientes en el modelo (5.1). Para lo respectivo al ajuste del modelo, se usó el paquete *geepack* (Halekoh et al., 2006) disponible en el software **R**.

El paquete *geepack* trabaja bajo el enfoque de las ecuaciones de estimación generalizadas. En este paquete, se hace uso de la función *geeglm* para hacer los ajustes respectivos a los parámetros del modelo que se desee ajustar. la función *geeglm*, tiene características muy similares a la función *glm* o *lm*, también incorporadas en el software **R**.

La función *geeglm* es la función principal que utiliza la biblioteca *geepack* de **R** para hacer las estimaciones de los parámetros correspondientes al modelo estadístico y entre otros argumentos tiene los siguientes:

- *family*: la función de varianza es especificada por este argumento y es identificada por el nombre de la distribución correspondiente en un modelo lineal generalizado. En la Tabla 5.2 se especifican las familias más representativas

con las que se cuentan y sus respectivas funciones de varianza

- *corstr*: las estructuras de correlación de trabajo predefinidas son especificadas con este argumento, en la Tabla 5.3 se muestran las diferentes opciones de estructuras de correlación y sus respectivas funciones de correlación $R(\alpha)$.

Tabla 5.2: Opciones del argumento *family* en la función *geeglm*.

Nombre	Función de varianza
Gaussian	Identidad
Binomial	$\mu(1-\mu),\ \mu \in (0,1)$
Poisson	$\mu,\ \mu > 0$
Gamma	$\mu^2, \mu > 0$

Tabla 5.3: Opciones del argumento *corstr* en la función geeglm.

Nombre	Función de correlación
Independence	$COR(Y_{it}, Y_{it'}) = 0, \quad t \neq t'$
Exchangeable	$COR(Y_{it}, Y_{it'}) = \alpha, \quad t \neq t'$
ar1	$COR(Y_{it}, Y_{it'}) = \alpha^{\lvert t-t' \rvert}, \quad t \neq t'$
Unstructured	$COR(Y_{it}, Y_{it'}) = \alpha_{tt'}, \quad t \neq t'$

Para el cálculo del QIC, se utilizó la función *model.sel*, incorporada al paquete *MuMIn*.

Etapa 2: La evaluación del modelo

En esta etapa se aplicó lo mencionado en el la sección 4.7. En específico, se ajustó el gráfico de probabilidades semi-normal, con la función *hnp*, incorporada al paquete *hnp* del software **R**. Se realizó el test chi cuadrado propuesto por Horton et al. (1999) y se halló las medidas de resumen de bondad de ajuste, propuestas por Zheng (2000).

Etapa 3: Análisis comparativo entre los enfoques estadísticos: análisis de sobrevivencia y análisis de datos longitudinales, para realizar inferencias respecto de la vida de anaquel sensorial de alimentos

En esta última etapa se realizó un análisis comparativo entre los dos tipos de enfoques estadísticos, señalando similitudes, diferencias, ventajas y desventajas entre ellos.

Capítulo 6

Resultados y discusión

6.1. Análisis longitudinal

Se procedió a realizar las estimaciones de los parámetros y sus respectivos errores estándar. Un resumen de los resultados se presenta en la Tabla 6.1.

Tabla 6.1: Estimaciones de los parámetros de los modelos ajustados.

Parámetros	Estimación[a]	Error estándar	Estimación[b]	Error estándar	Estimación[c]	Error estándar	Estimación[c]	Error estándar
intercepto	-2.31	0.34	-2.21	0.34	-2.29	0.33	-2.38	0.33
tiempo	0.09	0.01	0.09	0.01	0.09	0.01	0.09	0.01
consistencia 1	0.85	0.40	0.76	0.40	0.78	0.40	0.90	0.39
tiempo: consistencia 1	-0.04	0.02	-0.04	0.02	-0.03	0.02	-0.03	0.02

Estructura de correlación: [a] Independiente, [b] Intercambiable, [c] Autorregresivo de orden uno, y [c] No estructurado.

Para seleccionar el mejor modelo, según la estructura de correlación de trabajo más adecuada, se utilizó el criterio de información de cuasi-verosimilitud o QIC. Los resultados resumen en la Tabla 6.2.

Tabla 6.2: Criterio de información de cuasi verosimilitud para los diferentes modelos ajustados.

Modelo	Estructura de correlación	QIC
Modelo 1	no estructurada	375
Modelo 2	auto regresivo de orden 1	377
Modelo 3	independiente	377
Modelo 4	intercambiable	378

La correlación no estructurada tuvo el valor más pequeño de QIC, aunque no difiere en gran medida de los demás, se puede considerar que este es el mejor. Además, en general si el número de unidades por clúster es pequeño en un diseño balanceado y completo, entonces una matriz no estructurada es la recomendada (Pan, 2001, 2002).

Una vez seleccionada la estructura de correlación en la Tabla 6.3 se puede ver el resumen del ajuste de los parámetros con el enfoque de modelos marginales. Con base a esto, se sugiere que el patrón de cambio en el rechazo del producto difiere según el tipo de consumidor (consistente o inconsistente).

Tabla 6.3: Resumen del ajuste GEE, con el modelo de correlación no estructurado.

Parámetros	Estimación	Error estándar	Wald	$Pr(>\|W\|)$
intercepto	-2.38	0.33	51.71	<0.001
tiempo	0.09	0.01	46.45	<0.001
consistencia 1	0.90	0.39	5.18	0.02
tiempo:consistencia 1	-0.03	0.02	4.34	0.04

Como se mencionó antes, los errores estándar reportados son robustos (o empíricos) también conocidos en la literatura como errores tipo "sándwichz son estimadores que funcionan de manera óptima en diseños balanceados cuando el número de participantes es grande y hay pocas mediciones repetidas (Rogers y Stoner, 2015). Consideramos que en esta aplicación en particular, este es el mejor tipo de error que se puede utilizar.

Sin embargo, en la Tabla 6.4, se reportan los errores tipo Naive y se puede ver que son bastante cercanos. En general, si la estructura de correlación de trabajo se ha especificado correctamente, se espera que los errores estándar empíricos y basados en el modelo sean relativamente similares (Fitzmaurice et al., 2009).

Tabla 6.4: Errores estándar tipo Naive y Robusto para el ajuste GEE con estructura de correlación no estructurada.

Parámetros	Estimación	Error estándar tipo Naive	Error estándar tipo Robusto
intercepto	-2.38	0.32	0.33
tiempo	0.09	0.01	0.01
consistencia	0.90	0.42	0.40
tiempo:consistencia	-0.03	0.02	0.02

Al tener en cuenta lo anterior, se escogió el modelo con correlación no estructurada. Así, el modelo descrito en la Ec. (5.1), viene dado por:

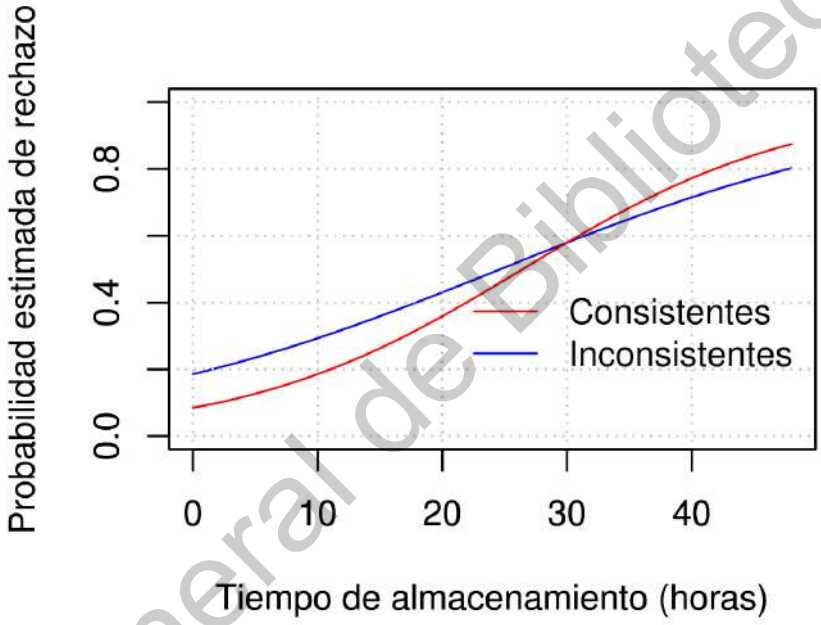
$$\log\left(\frac{E[Y_{ij}]}{(1-E[Y_{ij}])}\right) = -2.38 + 0.09 * t_j + 0.9 * c - 0.03 * (c * t_j) \qquad (6.1)$$

Por lo tanto, un posible modelo para describir la probabilidad de rechazo, dependiendo de la consistencia del consumidor, del tiempo de almacenamiento del producto y de la interacción entre estos, se puede escribir como:

$$\widehat{P}(\mathbf{Y}=1|\mathbf{T}=t,\mathbf{C}=c)=\frac{e^{(-2.38+0.09*t+0.9*c-0.03*(c*t))}}{1+e^{(-2.38+0.09*t+0.9*c-0.03*(c*t))}}. \tag{6.2}$$

Para observar el comportamiento de la probabilidad estimada de rechazo a través del tiempo de almacenamiento, para las dos subpoblaciones de consumidores, se puede ver la Figura 6.1.

Figura 6.1: Estimación de la probabilidad de rechazo por parte de los dos grupos de consumidores a través del tiempo de almacenamiento.



Se puede notar que en la subpoblación de consumidores consistentes, la probabilidad de rechazo al tiempo cero horas de almacenamiento es inferior al de la subpoblación de consumidores inconsistentes.

También se puede apreciar que la principal diferencia entre los dos grupos de consumidores es que los consumidores inconsistentes tienen una alta probabilidad de rechazar el producto fresco, casi el 20 %. De hecho, en la subpoblación de consumidores inconsistentes, la probabilidad de rechazar el producto, durante las primeras 24 horas de almacenamiento, se estima aproximadamente un 10 % más que la probabilidad de rechazar en la subpoblación de consumidores consistentes. Para contrastar esto último se reportaron los valores estimados de probabilidad de rechazo en 6 tiempos de almacenamiento menores o iguales a 25 horas, junto con sus intervalos del 95 % de confianza, ver Tabla 6.5 y 6.6.

Tabla 6.5: Estimaciones de la probabilidad de rechazo y sus intervalos de confianza del 95 % para la subpoblación de consumidores inconsistentes.

Tiempo(horas)	0	5	10	15	20	25
Lim. Inf.	0.13	0.17	0.23	0.28	0.34	0.40
Estimación Puntual	**0.19**	**0.24**	**0.29**	**0.36**	**0.43**	**0.51**
Lim. Sup.	0.26	0.30	0.35	0.41	0.48	0.56

Tabla 6.6: Estimaciones de la probabilidad de rechazo y sus intervalos de confianza del 95 %, para la subpoblación de consumidores consistentes.

Tiempo(horas)	0	5	10	15	20	25
Lim. Inf.	0.05	0.07	0.12	0.17	0.24	0.32
Estimación Puntual	**0.08**	**0.13**	**0.19**	**0.26**	**0.36**	**0.47**
Lim. Sup.	0.15	0.21	0.28	0.38	0.49	0.61

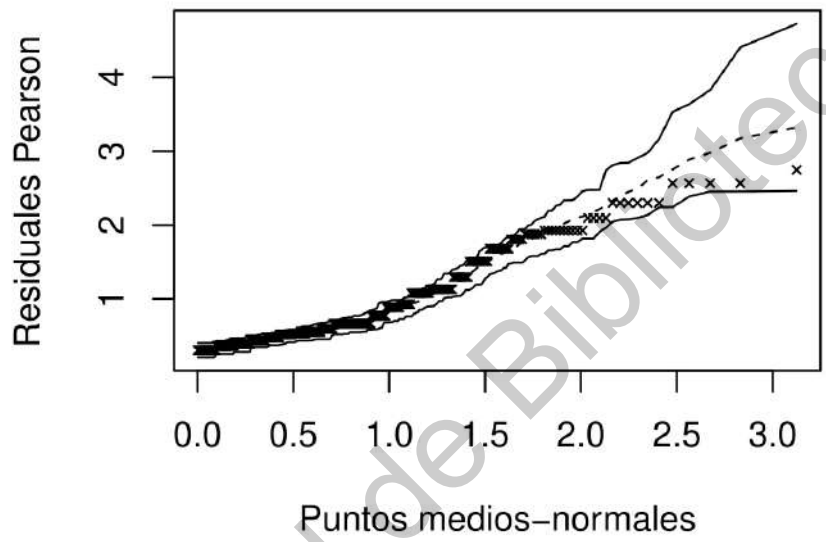
Al momento de hacer el reclutamiento de los consumidores que van a participar del análisis sensorial del producto alimenticio, se establecen unos criterios mínimos para que puedan ser parte de la muestra. Por ejemplo, en el caso del análisis sensorial que se realizó para generar la base de datos utilizada en este trabajo, se le preguntó al individuo si consumía yogurt por lo menos una vez a la semana y, si la respuesta era positiva, ingresaba a la muestra. Consideramos que una vez el individuo entra en la muestra, todas sus respuestas posteriores deben ser tenidas en cuenta, aún si rechaza el producto al tiempo cero, porque éste es un consumidor potencial del producto en cuestión, en el análisis realizado se consideraron todos los individuos incluidos cuatro que rechazaron al tirmpo cero horas de almacenamiento.
Por otro lado, cuando se realizan los análisis sensoriales de productos alimenticios, lo que se busca es conocer la aceptación y/o rechazo de un producto que va a ser lanzado al mercado, o que es un producto regular y se le han hecho modificaciones en su contenido, en su empaquetado u otra característica que podría modificar sus características sensoriales. Entonces, consideramos que una vez el individuo cumpla con los requisitos mínimos de selección establecidos por el analista sensorial, ingresa a la muestra y no debería ser retirado de la misma, pues es parte del grupo potencial de consumidores del producto alimenticio. Por lo tanto, resulta más realista tener en cuenta las respuestas de todos los consumidores reclutados.

6.2. Evaluación del modelo marginal

Al seguir lo propuesto por Venezuela et al. (2007), se usó el paquete *hnp* del software **R** para generar el grafico de probabilidad semi-normal, ver Figura 6.2. Lo que

evidenció que no existen puntos cayendo por fuera de la envoltura simulada, de tal manera que no hay datos con una influencia desproporcionada en el ajuste del modelo.

Figura 6.2: Gráfico de probabilidad semi-normal.



Por otro lado, se calculó el estadístico χ^2 propuesto por Horton et al. (1999), el cual usa deciles de riesgo predichos y que puede ser visto como una extensión del estadístico de bondad de ajuste de Hosmer y Lemeshow (1980).
Para su cálculo se utilizaron $G = 10$ grupos definidos por los deciles de los valores ajustados por el modelo. La prueba de bondad de ajuste dejó un p-valor de 0.05565, lo que muestra que no hay evidencia fuerte en contra de la hipótesis nula de no carencia de ajuste.

6.2.1. Comparación de modelos marginales

Para complementar la evaluación del modelo propuesto se ajustaron otros modelos bajo la misma estructura de correlación de trabajo, la no estructurada, permitiendo diferentes relaciones entre las variables explicativas. Se hallaron las medidas de bondad de ajuste resumidas en Zheng (2000) permitiendo una comparación más exhaustiva de los diferentes modelos. Se compararon los siguientes cuatro modelos:

Modelo 1: incluye como único factor principal el tiempo.

Modelo 2: incluye el tiempo como factor principal y un factor cuadrático del mismo.

Modelo 3: incluye dos factores principales, el tiempo y la consistencia.

Modelo 4: incluye los factores principales, tiempo y consistencia, así como la interacción entre los mismos.

En la Tabla 6.7 se resumen los valores ajustados de los parámetros involucrados en cada modelo, así como sus significancias. Se puede ver que el modelo 1 y 4, resultan adecuados para explicar a la variable respuesta Y. Ambos modelos indican que el rechazo del producto tiende a aumentar conforme el tiempo de almacenamiento aumenta y, para el último modelo, que la inconsistencia del consumidor contribuye a aumentar la probabilidad de rechazo del producto.

Tabla 6.7: Resumen de las estimaciones de los parámetros para los Modelos 1,2,3 y 4.

Modelos	Parámetros	Estimación	Error estándar	Wald	$Pr(>\|W\|)$
1	intercepto	-1.92	0.20	93.59	$< 2e^{-16}$
	tiempo	0.07	0.01	80.93	$< 2e^{-16}$
2	intercepto	-2.14	0.29	54.32	$< 2e^{-10}$
	tiempo	0.10	0.03	11.83	$5.8e^{-4}$
	tiempo^2	-6.60e^{-04}	5.70e^{-04}	1.31	0.25
3	intercepto	-2.10	0.28	55.25	$< 2e^{-10}$
	tiempo	0.07	0.01	80.78	$< 2e^{-16}$
	consistencia	0.36	0.297	1.51	0.22
4	intercepto	-2.38	0.33	51.71	$< 2e^{-10}$
	tiempo	0.09	0.01	46.45	$< 2e^{-10}$
	consistencia	0.90	0.40	5.18	0.02
	tiempo: consistencia	-0.03	0.02	4.34	0.04

La Tabla 6.8 resume las medidas de entropía marginal,H_{marg}, R^2_{marg} y el coeficiente de correlación de concordancia, r_c.

Tabla 6.8: Medidas de bondad de ajuste para los modelos 1,2,3 y 4.

	Modelo 1	Modelo 2	Modelo 3	Modelo 4
H_{marg}	0.216	0.220	0.224	**0.227**
R^2_{marg}	0.270	**0.275**	0.271	0.271
r_c	0.422	0.428	0.427	**0.432**

En negrilla se resaltan los valores más altos por medida.

De la Tabla 6.8 se puede observar que las magnitudes de cada una de las medidas allí resumidas, son bastante cercanas entre sí, y se resalta que entre los modelos 1 y 4, que son los más adecuados según los resultados de la Tabla 6.8, el modelo 4 se ajusta un poco mejor que el primero. Además, desde el punto de vista práctico, el modelo 4 incorpora el tipo de consumidor según su consistencia o no al responder a lo largo de la evaluación sensorial, aspecto que ayuda a entender un poco más el comportamiento de la variable respuesta Y; esto a diferencia de sólo incorporar el tiempo de almacenamiento al modelo, que dice más a la hora de interpretar el rechazo o la aceptación del producto alimenticio (yogurt).

Cabe mencionar que la base de datos utilizada en el presente estudio fue generada para ilustrar el uso del análisis de sobrevivencia para estimar la vida de anaquel sensorial de un yogurt comercial (Hough, 2010). Aquí sólo se midió la variable respuesta Y (descrita anteriormente) y no se consideró la inconsistencia o no, de manera explícita en el análisis sensorial. Esto pudiera ser relevante a la hora de modelar el rechazo o aceptación del producto alimenticio y, a su vez, pudiera ser significativo a la hora de estimar la vida de anaquel sensorial del producto. Variadas investigaciones (Costell et al., 2010; Hough, 2010) que concuerdan que existen factores que pueden favorecer o no a la aceptación de un producto alimenticio. Como lo mencionan Costell et al. (2010) "El proceso por el cual el hombre acepta o rechaza la comida/alimentos es de naturaleza multidimensional". Por ejemplo, características del consumidor, como su genética, su género, su grupo de edad, su estado fisiológico y psicológico, influencian en la decisión del consumidor a aceptar o rechazar los alimentos (Shepherd, 1989).

Por lo tanto, creemos que se puede formular un modelo mucho más robusto que el presentado aquí, si se incorporan variables explicativas como las mencionadas anteriormente. Lo propio se puede hacer con el enfoque de análisis de sobrevivencia.

6.3. Comparación del enfoque longitudinal y el enfoque de análisis de sobrevivencia

En Hough (2010) se ejemplifica el uso del análisis de sobrevivencia para estimar la vida de anaquel sensorial de un yogurt comercial y en la presente investigación se emplearon los mismos datos que en dicha investigación. Como se describió en los antecedentes, Hough (2010) codificó el vector de respuestas binarias de cada consumidor bajo ciertos criterios, y le asoció un intervalo de tiempo (Tabla 2.1).

Luego de tener para cada consumidor los intervalos de tiempos asociados, se estimó un modelo para la función de rechazo teniendo en cuenta lo descrito por Meeker y Escobar (1998).

Como consecuencia de la codificación realizada, el vector de respuestas del consumidor desaparece como tal, en el análisis hecho por Hough (2010) y lo que se analiza es el tiempo de rechazo como intervalo censurado, a diferencia del enfoque descrito aquí, donde la variable aleatoria de interés es la respuesta del consumidor. Entonces, el enfoque de análisis de datos longitudinales permite modelar directamente la respuesta del consumidor y utilizar el tiempo de almacenamiento como una variable explicativa dentro del modelo.

Los resultados y las interpretaciones de ambos enfoques difieren debido a que las variables aleatorias modeladas son distintas. Más aún, en el enfoque de análisis de datos longitudinales, el tiempo de almacenamiento entra como una variable explicativa junto con la consistencia del consumidor.

En efecto, con un modelo marginal cuando el tiempo de vida de anaquel sensorial fuera especulado por el analista de alimentos se puede responder a la pregunta **¿cuál es la probabilidad de rechazo poblacional?**.

Con el enfoque de análisis de sobrevivencia se encontró que, para una probabilidad acumulada del 50 % de rechazo del consumidor, la vida de anaquel sensorial estimada (para un modelo log-normal) es hasta de 19.8 h, con un intervalo de confianza del 95 % entre 14.8 h y 26.6 h. En contraste con el enfoque de análisis de datos longitudinales, que se puede inferir lo siguiente:

- para el grupo de consumidores consistentes, cuando la vida de anaquel sensorial del producto fuera especulada en 20 h, existe una probabilidad de rechazo aproximada del 36 % con un intervalo de confianza del 95 % entre 24 % y 49 %.
- para el grupo de consumidores inconsistentes, cuando la vida de anaquel sensorial del producto fuera especulada en 20 h, existe una probabilidad de rechazo aproximada del 43 % con un intervalo de confianza del 95 % entre 34 % y 48 %.
- para el grupo agregado sin distinguir entre el tipo de consumidor, cuando la vida de anaquel sensorial del producto fuera especulada en 20 h, existe una probabilidad de rechazo aproximada del 38 % con un intervalo de confianza del 95 % entre 31 % y 46 %.

Por otro lado, las pruebas de bondad de ajuste para datos de sobrevivencia, en el contexto de análisis sensorial de alimentos, se realizan con comparaciones visuales entre las diferentes distribuciones ajustadas y los datos experimentales, o bien comparando las log verosimilitudes; en cambio, para el contexto de datos longitudinales, hay pruebas de bondad de ajuste, como es el caso del test chi-cuadrado propuesto por Horton (1999), y otros más (Evans y Li, 2005); lo que permite una mejor evaluación de los modelos ajustados, mediante el enfoque de análisis de datos longitudinales.

A su vez, el análisis de las evaluaciones sensoriales, mediante el enfoque aquí discutido, comparte varias ventajas del mismo enfoque de análisis de sobrevivencia, como son las inferencias respectivas a la vida de anaquel sensorial siguen estando basadas en las respuestas de los consumidores; y en este caso específico, no se hace necesaria el proceso de codificación de esas respuestas, y mucho menos trabajar con datos censurados.

6.3.1. Simulación

Para profundizar un poco más en los resultados de aplicar estos dos tipos de modelado a datos provenientes de análisis sensorial de alimentos y establecer una comparación más amplia, se generaron diferentes escenarios hipotéticos de lo que serían las respuestas de consumidores en un análisis sensorial de alimentos, donde se responde a la pregunta: ¿normalmente consumiría este producto? Se realizó un estudio de simulación que consistió en simular cadenas de Markov no homogéneas con dos estados posibles 1 = no y 0 = si. La hipótesis inicial fue que las respuestas de los consumidores se pueden simular mediante una cadena no homogénea, pensando que sólo la respuesta actual depende de la inmediatamente anterior, cumpliéndose así la propiedad de Markov. Se consideró que la cadena de respuestas no es homogénea porque la probabilidad de responder si/no al tiempo t, cambia conforme el tiempo avanza. Esto se debe a que la frescura del producto depende del tiempo. Cabe resaltar que en la simulación sólo se generaron las respuestas de los consumidores, no se incluyó la simulación de ninguna variable explicativa. Se analizaron seis escenarios diferentes, descritos a continuación:

- El primer escenario fue motivado por la base de datos con la que se realizó la presente investigación. Se simularon respuestas de individuos tanto consistentes como inconsistentes, presentándose 8 individuos que rechazaron al tiempo 0 horas de almacenamiento.

- En el segundo escenario, se tienen las mismas características que el escenario anterior, sólo que las matrices de transición que generaron el proceso cambiaron un poco (ver Anexo A.1).
- En el tercer escenario se simularon respuestas de individuos consistentes y que no rechazan al tiempo cero horas.
- En el cuarto escenario se simularon respuestas de individuos tanto consistentes como inconsistentes.
- En el quinto y sexto se simularon respuestas de individuos consistentes únicamente y que no rechazan al tiempo cero horas. Estos dos escenarios se diferenciaron en sus primeras cinco matrices de transición, buscando que los individuos simulados en el sexto escenario rechazaran en horas más tempranas.

Más detalles de los escenarios de simulación se pueden ver en Anexo A.1.

A través de estos escenarios de simulación se buscó verificar cómo los cambios de cada uno de ellos afectan los ajustes de los modelos con los dos enfoques. escenarios.

Tabla 6.9: Resumen de algunos resultados obtenidos con los diferentes escenarios de simulación.

Escenario	Tamaño de la cadena	Enfoque longitudinal		Enfoque análisis de sobrevivencia
		Estructura de Correlación	$P(Y=1\|t=0)$	Distribución
1	corta con tiempos	no estructurada	0.093	Lognormal
2	de 0, 4, 8, 12, 24, 36	no estructurada	0.145	Lognormal
3	y 48 horas	**ar1**	**0.108**	**Lognormal**
4	larga con tiempos de 0, 2, 4, 6, 8,	intercambiable	0.082	Lognormal
5	10, 12, 14, 16, 24, 36	**ar1**	**0.022**	**Logística**
6	y 48 horas	**ar1**	**0.52**	**Gausiana**

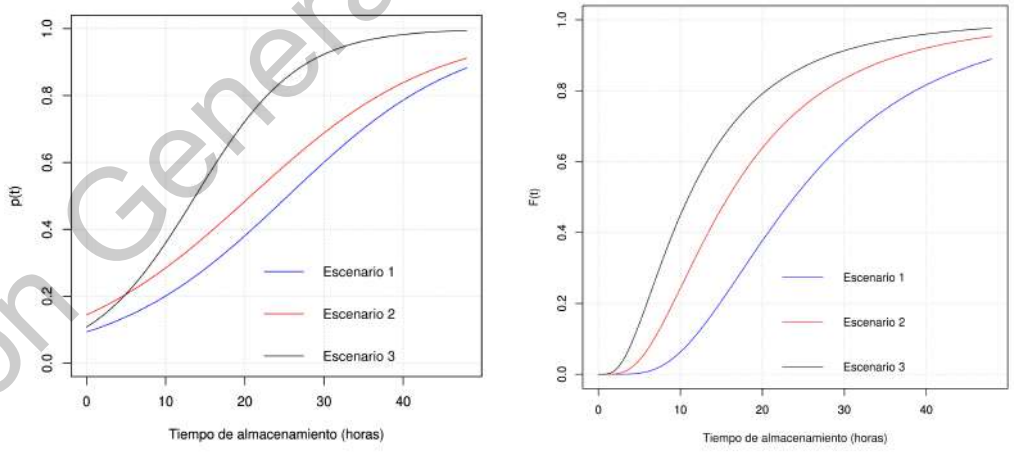
Se resalta en negrilla los escenarios donde se simularon individuos únicamente consistentes.

De la Tabla 6.9 nóte que en los escenarios donde se simulan individuos sólo consistentes, para el enfoque longitudinal, se favorece la estructura de correlación ar(1),

es decir, para estos escenarios se supone que las correlaciones (positivas) disminuyen con el tiempo a medida que aumenta la separación entre pares de medidas repetidas, una característica que efectivamente se puede suponer de las cadenas generadas en estos escenarios ya que esta estructura resulta apropiada cuando las mediciones se realizan en intervalos de tiempo aproximadamente iguales y, cuando se supone que dos mediciones que están una al lado de la otra en el tiempo estarán bastante correlacionadas, pero que a medida que las mediciones se alejen cada vez más, su correlación disminuirá. Esta situación puede ser inducida por el patrón definido en las cadenas de Markov, de manera que la probabilidad de aceptar disminuye conforme el tiempo. Por otro lado, en el enfoque de análisis de sobrevivencia, nótese como la distribución de probabilidad que se ajusta a los datos cambia de log-normal a logística y gaussiana; estas dos últimas distribuciones menos frecuentes en el ajuste de datos sobrevivencia, donde prevalece la distribución log-normal, así que también suponemos que el modelo con el enfoque de análisis de sobrevivencia es sensible de alguna manera a los diferentes escenarios. Para ver con más detalle cómo cambian los ajustes de los modelos de escenario en escenario, se muestran las Figuras 6.3a, 6.3b, 6.4a y 6.4b.

Figura 6.3: Comparación de $p(t)$ y $F(t)$ para los escenarios de simulación 1, 2 y 3.

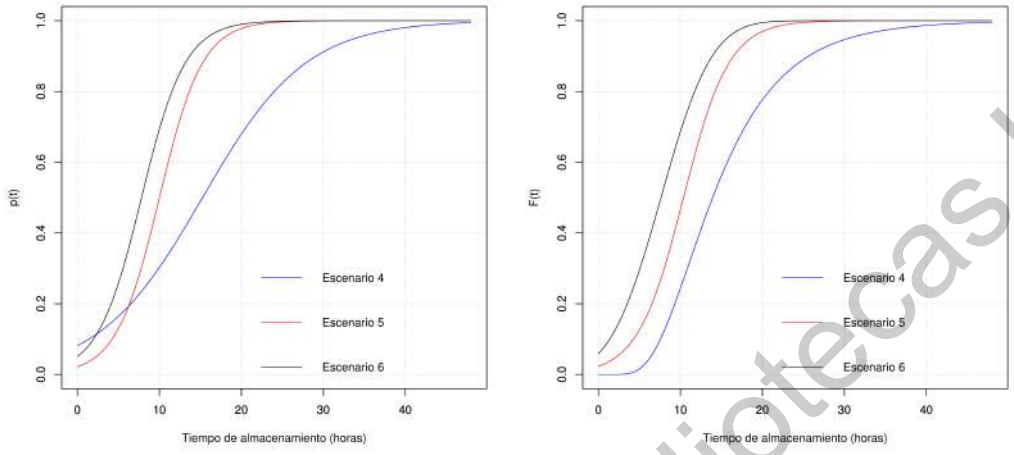
(a) Gráficas de $p(t)$ (b) Gráficas de $F(t)$



En las Figuras 6.3a y 6.3b se grafican las funciones $p(t)$ y $F(t)$ para los escenarios 1, 2 y 3 con la característica común de ser escenarios donde se simularon cadenas cortas, con 7 tiempos de almacenamiento. Note que para ambos enfoques las funciones $p(t)$ y $F(t)$ para el escenario 3 están por encima de las $p(t)$ y $F(t)$ de los otros dos escenarios. En las Figuras 6.4a y 6.4b se presentan las gráficas

Figura 6.4: Comparación de $p(t)$ y $F(t)$ para los escenarios de simulación 4, 5 y 6.

(a) Gráficas de $p(t)$ (b) Gráficas de $F(t)$



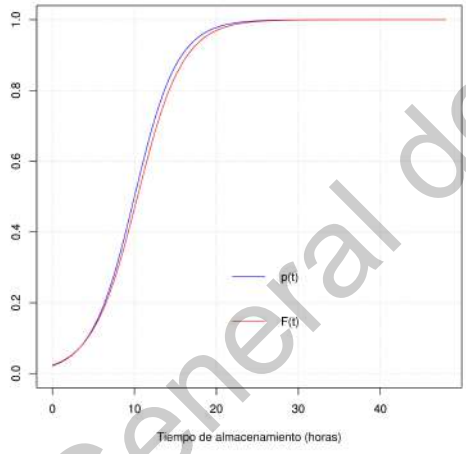
de $p(t)$ y $F(t)$ ahora para los escenarios 4, 5 y 6. En estas gráficas también se puede notar como prevalece en los dos enfoques que las funciones $p(t)$ y $F(t)$ del escenario 6 están por encima de los otros dos escenarios. Específicamente, el orden presente en las gráficas de $p(t)$ y $F(t)$ para los diferentes escenarios se está viendo influenciado por las distribuciones a tiempo t, denotadas por μ^t, que están directamente relacionadas con la distribución inicial y las matrices de transición definidas para cada escenario.

Es claro que los dos enfoques capturan, de alguna manera, la variabilidad presente en los escenarios de simulación propuestos. Note cómo en la comparación de las gráficas de $p(t)$ y $F(t)$ para los tres primeros escenarios éstas se presentan en el mismo orden, así como para los escenarios 4, 5 y 6. Entonces, podríamos rescatar que los dos enfoques son sensibles a las diferentes características de los escenarios de simulación.

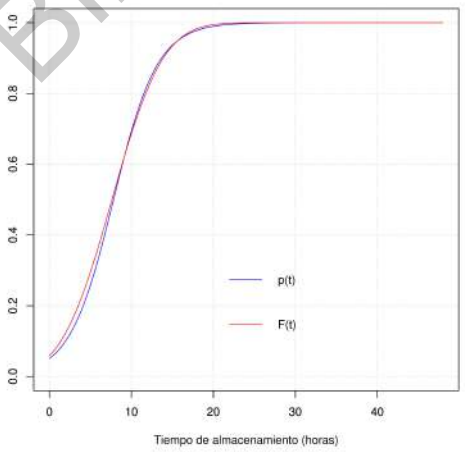
Una última observación, no menos importante, a partir del análisis de la simulación es que en los escenarios 5 y 6 las funciones $p(t)$ y $F(t)$ presentaron bastante similitud a lo largo del tiempo, esto se puede apreciar en las Figuras 6.5a y 6.5b. Esta simulitud se presentó en los escenarios donde se simularon individuos consistentes y la cadena simulada correspondió a doce tiempos de almacenamiento, de manera que podemos ver que hay escenarios en donde las gráficas de $p(t)$ y $F(t)$ se asemejan y pudiese haber una relación entre las mismas. Sin embargo, esto debe ser estudiado con mayor detalle ya que las interpretaciones que arroja p(t) y F(t) difieren sustancialmente.

Figura 6.5: Gráficas de $p(t)$ y $F(t)$ para los escenarios de simulación 5 y 6.

(a) Escenario 5



(b) Escenario 6



Capítulo 7

Conclusiones

El uso del análisis de datos longitudinales como alternativa de análisis estadístico para realizar inferencias en torno a la vida de anaquel sensorial de alimentos es viable y ofrece otra perspectiva de modelado. Diferentes modelos marginales permiten modelar las respuestas directas de los consumidores, al tiempo que evitan la censura de las mismas, Esto, potencialmente podría generar estimaciones más precisas de la percepción sensorial que tiene el consumidor en torno al producto alimenticio.

A través del análisis de simulación, se pudo constatar que el modelado a partir del análisis de datos longitudinales, aplicado a observaciones que provienen de estudios sensoriales de alimentos, es sensible a los diferentes cambios que pudiese haber en los patrones de las diferentes respuestas de los consumidores a lo largo del tiempo.

El enfoque utilizado en la presente investigación tiene fuertes bases teóricas y ofrece un modelado completo de los datos, desde el ajuste de los parámetros del modelo hasta la evaluación del mismo.

La investigación realizada demuestra que existe por lo menos un enfoque estadístico, diferente al de análisis de sobrevivencia, útil para realizar las inferencias respectivas a la vida de anaquel sensorial de alimentos.

7.1. Limitaciones

Una limitante fue la base de datos utilizada en los análisis, ya que es una base de datos carente de variables explicativas, que indudablemente las hay y pueden ofrecer

mejores ajustes al modelo propuesto.

7.2. Trabajo a futuro

- Sería importante que se pudiera implementar el enfoque de análisis longitudinales en estudios sensoriales de alimentos con una base de datos más completa, que incorpore variables explicativas relevantes.
- Un enfoque alternativo interesante por estudia sería aplicar el enfoque de modelos de transición; donde se modelaría la esperanza condicional de la variable respuesta dada respuestas pasadas.
- Otra labor pendiente es la de estudiar si existe una relación entre la función $p(t)$, obtenida con el enfoque de análisis de datos longitudinales, y la distribución de probabilidad $F(t)$ que se ajusta a partir del análisis de sobrevivencia.

Apéndice A

Anexos

A.1.

Las simulaciones de las cadenas de Markov no homogéneas, se realizaron usando la teoría contenida en Haggstrom (2002).

Descripción de los escenarios de simulación.

Escenario 1: En este escenario se presentaron 49 individuos consistentes y 51 inconsistentes de los cuales 8 rechazaron al tiempo cero horas.

La distribución inicial está dada por μ^0 =(0.95,0.05) y las matrices de transición se definieron así:

$$P^1 = \begin{bmatrix} 0.95 & 0.05 \\ 0.9 & 0.1 \end{bmatrix}, \quad P^2 = \begin{bmatrix} 0.9 & 0.1 \\ 0.7 & 0.3 \end{bmatrix}, \quad P^3 = \begin{bmatrix} 0.75 & 0.25 \\ 0.9 & 0.1 \end{bmatrix},$$

$$P^4 = \begin{bmatrix} 0.4 & 0.6 \\ 0.5 & 0.5 \end{bmatrix}, \quad P^5 = \begin{bmatrix} 0.2 & 0.8 \\ 0.4 & 0.6 \end{bmatrix}, \quad P^6 = \begin{bmatrix} 0.1 & 0.9 \\ 0.2 & 0.8 \end{bmatrix}$$

y para $n \geq 7$

$$P^n = \begin{bmatrix} \frac{1}{4(n-2)} & 1 - \frac{1}{4(n-2)} \\ \frac{1}{4(n-2)} & 1 - \frac{1}{4(n-2)} \end{bmatrix}$$

De tal manera que: $\mu^1 = (0.95, 0.05)$, $\mu^2 = (0.89, 0.11)$, $\mu^3 = (0.77, 0.23)$, $\mu^4 = (0.42, 0.58)$, $\mu^5 = (0.32, 0.68)$, $\mu^6 = (0.17, 0.83)$, $\mu^7 = (0.05, 0.95)$, $\mu^8 = (0.04, 0.96)$, $\mu^9 = (0.04, 0.96)$, $\mu^{10} = (0.03, 0.97)$.
Cuando $n \to \infty, \mu^n \to (0, 1)$.

Escenario 2: En este escenario se presentaron 44 individuos consistentes y 56

inconsistentes de los cuales 7 rechazaron al tiempo cero horas. La distribución inicial está dada por $\mu^0 = (0.95, 0.05)$ y las matrices de transición se definieron así:

$$P^1 = \begin{bmatrix} 0.95 & 0.05 \\ 0.9 & 0.1 \end{bmatrix}, \quad P^2 = \begin{bmatrix} 0.9 & 0.1 \\ 0.7 & 0.3 \end{bmatrix}, \quad P^3 = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix},$$

$$P^4 = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}, \quad P^5 = \begin{bmatrix} 0.2 & 0.8 \\ 0.4 & 0.6 \end{bmatrix}, \quad P^6 = \begin{bmatrix} 0.1 & 0.9 \\ 0.1 & 0.9 \end{bmatrix}$$

y para $n \geq 7$

$$P^n = \begin{bmatrix} \frac{1}{4(n-2)} & 1 - \frac{1}{4(n-2)} \\ \frac{1}{4(n-2)} & 1 - \frac{1}{4(n-2)} \end{bmatrix}$$

De tal manera que: $\mu^1 = (0.95, 0.05)$, $\mu^2 = (0.89, 0.11)$, $\mu^3 = (0.5, 0.5)$, $\mu^4 = (0.5, 0.5)$, $\mu^5 = (0.2, 0.8)$, $\mu^6 = (0.18, 0.82)$, $\mu^7 = (0.05, 0.95)$, $\mu^8 = (0.04, 0.96)$, $\mu^9 = (0.04, 0.96)$, $\mu^{10} = (0.03, 0.97)$.
Cuando $n \to \infty, \mu^n \to (0, 1)$.

Escenario 3: Para este escenario se generó una muestra de tamaño 100 de solo individuos consistentes , a partir de una cadena no homogénea con tiempos de almacenamiento $0, 4, 8, 12, 24, 36$, y 48 horas. La distribución inicial está dada por $\mu^0 = (1, 0)$, lo que implicó que no hubo individuos simulados que rechazaran al tiempo cero horas; y las matrices de transición se definieron así:

$$P^1 = \begin{bmatrix} 0.9 & 0.1 \\ 0 & 1 \end{bmatrix}, \quad P^2 = \begin{bmatrix} 0.8 & 0.2 \\ 0 & 1 \end{bmatrix}, \quad P^3 = \begin{bmatrix} 0.5 & 0.5 \\ 0 & 1 \end{bmatrix},$$

$$P^4 = \begin{bmatrix} 0.6 & 0.4 \\ 0 & 1 \end{bmatrix}, \quad P^5 = \begin{bmatrix} 0.25 & 0.75 \\ 0 & 1 \end{bmatrix}, \quad P^6 = \begin{bmatrix} 0.25 & 0.75 \\ 0 & 1 \end{bmatrix}$$

y para $n \geq 7$

$$P^n = \begin{bmatrix} \frac{1}{4(n-2)} & 1 - \frac{1}{4(n-2)} \\ \frac{1}{4(n-2)} & 1 - \frac{1}{4(n-2)} \end{bmatrix}$$

De tal manera que: $\mu^1 = (0.9, 0.1)$, $\mu^2 = (0.72, 0.18)$, $\mu^3 = (0.36, 0.64)$, $\mu^4 = (0.22, 0.78)$, $\mu^5 = (0.05, 0.95)$, $\mu^6 = (0.01, 0.99)$, $\mu^7 = (0.05, 0.95)$, $\mu^8 = (0.04, 0.96)$, $\mu^9 = (0.04, 0.96)$, $\mu^{10} = (0.03, 0.97)$.
Cuando $n \to \infty, \mu^n \to (0, 1)$.

Escenario 4: Se simularon cadenas de Markov no homogéneas más largas, específicamente de tamaño 12, simulando tiempos de almacenamiento de $0, 2, 4, 6, 8, 10, 12, 14, 16, 24, 36$ y 48 horas. Se generó un grupo de 100 indi-

viduos tanto consistentes como inconsistentes, con matrices de transición:

$$P^1 = \begin{bmatrix} 0.95 & 0.05 \\ 0 & 1 \end{bmatrix}, \quad P^2 = \begin{bmatrix} 0.85 & 0.15 \\ 0.9 & 0.1 \end{bmatrix}, \quad P^3 = \begin{bmatrix} 0.8 & 0.2 \\ 0.9 & 0.1 \end{bmatrix},$$

$$P^4 = \begin{bmatrix} 0.7 & 0.3 \\ 0.8 & 0.2 \end{bmatrix}, \quad P^5 = \begin{bmatrix} 0.6 & 0.4 \\ 0.7 & 0.3 \end{bmatrix}, \quad P^6 = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

$$P^7 = \begin{bmatrix} 0.4 & 0.6 \\ 0.5 & 0.5 \end{bmatrix}, \quad P^8 = \begin{bmatrix} 0.3 & 0.7 \\ 0.3 & 0.7 \end{bmatrix}, \quad P^9 = \begin{bmatrix} 0.2 & 0.8 \\ 0.2 & 0.8 \end{bmatrix}$$

$$P^{10} = \begin{bmatrix} 0.1 & 0.9 \\ 0.1 & 0.9 \end{bmatrix}, \quad P^{11} = \begin{bmatrix} 0.05 & 0.95 \\ 0.05 & 0.95 \end{bmatrix},$$

y para $n \geq 12$

$$P^n = \begin{bmatrix} \frac{1}{4(n-2)} & 1 - \frac{1}{4(n-2)} \\ 0 & 1 \end{bmatrix}$$

La distribución inicial está dada por $\mu^0 = (1, 0)$, con el fin de no eliminar individuos de la muestra a la hora de aplicar el enfoque de análisis de sobrevivencia y realizar la comparación con los 100 individuos. Recuerde que al establecer $\mu^0 = (1, 0)$, implica que todos los individuos simulados aceptarán en el tiempo cero horas.

Luego:

$\mu^1 = (0.95, 0.05)$, $\mu^2 = (0.85, 0.15)$, $\mu^3 = (0.81, 0.19)$, $\mu^4 = (0.72, 0.28)$, $\mu^5 = (0.63, 0.37)$, $\mu^6 = (0.5, 0.5)$, $\mu^7 = (0.45, 0.55)$, $\mu^8 = (0.3, 0.7)$, $\mu^9 = (0.2, 0.8)$, $\mu^{10} = (0.1, 0.9)$, $\mu^{11} = (0.05, 0.95)$.
Cuando $n \to \infty, \mu^n \to (0, 1)$.

En esta simulación resultaron tan sólo 18 consumidores consistentes y el resto inconsistentes.

Escenario 5: Se simularon cadenas de Markov no homogéneas de tamaño 12, con los mismos tiempos de almacenamiento anteriores, pero ahora simulando respuestas de individuos consistentes únicamente y que no rechazan al tiempo cero horas, por ello la distribución inicial es $\mu^0 = (1, 0)$. Fue generada una muestra de tamaño 100, con las siguientes matrices de transición:

$$P^1 = \begin{bmatrix} 0.95 & 0.05 \\ 0 & 1 \end{bmatrix}, \quad P^2 = \begin{bmatrix} 0.9 & 0.1 \\ 0 & 1 \end{bmatrix}, \quad P^3 = \begin{bmatrix} 0.85 & 0.15 \\ 0 & 1 \end{bmatrix},$$

$$P^4 = \begin{bmatrix} 0.8 & 0.2 \\ 0 & 1 \end{bmatrix}, \quad P^5 = \begin{bmatrix} 0.75 & 0.25 \\ 0 & 1 \end{bmatrix}, \quad P^6 = \begin{bmatrix} 0.5 & 0.5 \\ 0 & 1 \end{bmatrix}$$

$$P^7 = \begin{bmatrix} 0.45 & 0.55 \\ 0 & 1 \end{bmatrix}, \quad P^8 = \begin{bmatrix} 0.4 & 0.6 \\ 0 & 1 \end{bmatrix}, \quad P^9 = \begin{bmatrix} 0.3 & 0.7 \\ 0 & 1 \end{bmatrix}$$

$$P^{10} = \begin{bmatrix} 0.2 & 0.8 \\ 0 & 1 \end{bmatrix}, \quad P^{11} = \begin{bmatrix} 0.1 & 0.9 \\ 0 & 1 \end{bmatrix},$$

y para $n \geq 12$

$$P^n = \begin{bmatrix} \frac{1}{4(n-2)} & 1 - \frac{1}{4(n-2)} \\ 0 & 1 \end{bmatrix}$$

De tal maner que:

$\mu^1 = (0.95, 0.05)$, $\mu^2 = (0.86, 0.14)$, $\mu^3 = (0.73, 0.27)$, $\mu^4 = (0.58, 0.42)$, $\mu^5 = (0.44, 0.56)$, $\mu^6 = (0.22, 0.78)$, $\mu^7 = (0.01, 0.99)$, $\mu^8 = (0.04, 0.96)$, $\mu^9 = (0.01, 0.99)$, $\mu^{10} = (0.002, 0.998)$, $\mu^{11} = (0.0002, 0.9998)$.
Cuando $n \to \infty, \mu^n \to (0, 1)$.

Escenario 6: Siguiendo con la simulación de una cadena más larga, simulando 12 tiempos de almacenamiento, se generó un grupo de 100 individuos consistentes que no rechazan al tiempo cero horas ($\mu^0 = (1, 0)$), pero con matrices de transición diferentes, en el sentido que se otorga un poco más de probabilidad de pasar de aceptar a rechazar durante las primeras horas de almacenamiento

(estas probabilidades están en rojo en las matrices de transición).

$$P^1 = \begin{bmatrix} 0.9 & 0.1 \\ 0 & 1 \end{bmatrix}, \quad P^2 = \begin{bmatrix} 0.8 & 0.2 \\ 0 & 1 \end{bmatrix}, \quad P^3 = \begin{bmatrix} 0.75 & 0.25 \\ 0 & 1 \end{bmatrix},$$

$$P^4 = \begin{bmatrix} 0.7 & 0.3 \\ 0 & 1 \end{bmatrix}, \quad P^5 = \begin{bmatrix} 0.65 & 0.35 \\ 0 & 1 \end{bmatrix}, \quad P^6 = \begin{bmatrix} 0.5 & 0.5 \\ 0 & 1 \end{bmatrix}$$

$$P^7 = \begin{bmatrix} 0.45 & 0.55 \\ 0 & 1 \end{bmatrix}, \quad P^8 = \begin{bmatrix} 0.4 & 0.6 \\ 0 & 1 \end{bmatrix}, \quad P^9 = \begin{bmatrix} 0.3 & 0.7 \\ 0 & 1 \end{bmatrix}$$

$$P^{10} = \begin{bmatrix} 0.2 & 0.8 \\ 0 & 1 \end{bmatrix}, \quad \text{y} \quad P^{11} = \begin{bmatrix} 0.1 & 0.9 \\ 0 & 1 \end{bmatrix},$$

y para $n \geq 12$

$$P^n = \begin{bmatrix} \frac{1}{4(n-2)} & 1 - \frac{1}{4(n-2)} \\ 0 & 1 \end{bmatrix}$$

De tal manera que:

$\mu^1 = (0.9, 0.1)$, $\mu^2 = (0.72, 0.28)$, $\mu^3 = (0.54, 0.46)$, $\mu^4 = (0.38, 0.62)$, $\mu^5 = (0.25, 0.75)$, $\mu^6 = (0.12, 0.88)$, $\mu^7 = (0.06, 0.94)$, $\mu^8 = (0.02, 0.98)$, $\mu^9 = (0.007, 0.993)$, $\mu^{10} = (0.001, 0.999)$, $\mu^{11} = (0.0001, 0.9999)$.
Cuando $n \to \infty, \mu^n \to (0, 1)$.

Bibliografía

Atkinson, A. (1981). Two graphical displays for outlying and infuential observations in regression. *Biometrika*, 68(1):13-20.

Costell, E., Tárrega, A., y Bayarri, S. (2010). Food Acceptance: The Role of Consumer Perception and Attitudes. *Chemosensory Perception*, 3:42-50.

Cruz, A., Walter, E., Silva, R., Faria, J., Bolini, H., Pinheiro, H., y Sant'Ana, A. (2010). Survival analysis methodology to predict the shelf-life of probiotic flavored yogurt. *Food Research International*, 43(5):1444-1448.

Diggle, P., Heagerty, P., Liang, K., y Zeger, S. (2002). *Analysis of Longitudinal Data*. Oxford University Press.

Evans, S., y Li, L. (2005). A comparison of goodness of fit tests for the logistic GEE model. *Statistic in medicine*, 24:1245-1261.

Fitzmaurice, G., Davidian, M., Verbeke, G., y Molenberghs, G. (2009). *Longitudinal data analysis, Handbooks of Modern Statistical Methods*. Boca Raton:Taylor and Francis Group.

Halekoh, U., Højsgaard, S., y Yan, J. (2006). The R Package geepack for generalized estimating equations. *Journal of Statistical Software*, 15:1-11.

Haggstrom, O. (2002). *Finite Markov chains and algorithmic applications*. Cambridge University Press.

Horton, N., Bebchuk, J., Jones, C., Lipsitz, S., Catalano, P., Zahner, G., y Fitzmaurice, G. (1999). Goodness-of-fit for gee: an example with mental health service utilization. *Statistics in Medicine*, (18):213-222.

Hosmer, D., y Lemeshow, S. (1980). A goodness -of-fit test for the multiple logistic regression model. *Communications in Statistics*, (A10):1043-1069.

Hough, G. (2010). *Sensory Shelf Life Estimation of Food Products*. Boca Raton: CRC Press.

Hough, G., Langhor, K., Gómez, G., y Curia, A. (2003). Survival analysis applied to sensory shelf-life of foods. *Journal Food Science*, 68:359-362.

Hough, G., y Garitta, L. (2012). Methodology for sensory shelf-life estimation: A review. *Journal of Sensory Studies*, 23:137-147.

Jacobo-Velázquez, D., Ramos-Parra, P., y Hernández-Brenes, C. (2010). Survival analysis applied to the sensory shelf-life dating of high hydrostatic pressure processed avocado and mango pulps. *Journal Food Science*, 75(6):286-291.

Lawless, H., y Heymann, H. (2010). *Sensory evaluation of food:Principles and practices*. New York:Springer-Verlag.

Liang, Kung-Yee, y Zeger, Scott L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13-22.

Meeker, W., y Escobar, L. (1998). *Statistical methods for reliability data*. New York:John Wiley and Sons.

Pan, W. (2001). Akaikés Information Criterion in Generalized Estimating Equations. *Biometrics*, 57(1):120-125.

Pan, W. (2002). Goodness-of-Fit Tests for GEE with Correlated Binary Data. *Scandinavian Journal of Statistics*, 29(1):101-110.

Rogers, P., y Stoner, J. (2015). Modification of the sandwich estimator in generalized estimating equations with correlated binary outcomes in rare event and small sample settings. *American Journal of Applied Mathematics and Statistics*, 3(6):243-251.

Shepherd, R. (1989). *Factors in influencing food preferences and choice*. Chichester: Wiley.

Venezuela, K., Booter, A., y Carneiro, M. (2007). Diagnostic techniques in generalized estimating equations. *Journal of Statistical Computation and Simulation*, 77(10):879-888.

Ware, J. (1985). Linear models for the analysis of several measurements in longitudinal studies. *American Statistician*, 39:95-101.

Wedderburn, R. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61:439-447.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrics*, 50:1-25.

Zeger, Scott L., y Liang, Kung-Yee. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42:121-130.

Zheng, B. (2000). Summarizing the goodness of fit of generalized linear models for longitudinal data. *Statistics in medicine*, 19:1265-1275.