ILIANA MARÍA PATERNINA ORTEGA
EDUARDO CASTAÑO TOSTADO
MARIO SANTANA CIBRIÁN

UNIVERSIDAD AUTÓNOMA DE QUERÉTARO

PATERNINAO@GMAIL.COM

03

UN ANÁLISIS COMPARATIVO DE MODELAJE ESTADÍSTICO EN ESTUDIOS DE VIDA DE ANAQUEL SENSORIAL

A COMPARATIVE ANALYSIS OF STATISTICAL MODELING IN SENSORY SHELF LIFE STUDIES

RESUMEN

El presente artículo describe un modelo estadístico para datos provenientes de análisis sensoriales de alimentos, teniendo en cuenta la estructura de agrupación de tipo clúster que éstos poseen, así como el patrón de asociación en el tiempo para el conjunto de datos. Se hace uso del enfoque de las ecuaciones estimantes generalizadas, que ajusta modelos de media marginal con la ventaja de que solamente es necesaria la especificación correcta de las medias marginales para que los estimadores de los parámetros sean consistentes y asintóticamente normales. Para una base de datos específica donde se evalúa sensorialmente un yogurt, se encontró que la probabilidad de rechazo esperada sobre el tiempo es no lineal y creciente; reportándose un 13 % de rechazo poblacional esperado a las cero horas de almacenamiento del producto.

Palabras Clave: IMA, vida de anaquel sensorial, clúster, autocorrelación, ecuaciones estimantes generalizadas, marginal.

ABSTRACT

This article describes a statistical model for data recovered from sensorial food analyses, taking into account the cluster-type grouping structure that they have, as well as the pattern of association over time for the data set, adjusting a marginal model using the generalized estimating equations approach, a method that adjusts marginal mean models with the advantage that only the correct specification of the marginal means is necessary for the parameter estimators to be consistent and asymptotically normal. For a specific database, where a yogurt is sensorially evaluated, it was found that the probability of rejection over time is nonlinear and increasing; reporting a 13% expected population rejection at zero hours of product storage.

Keywords: IMA, sensory shelf life, cluster, autocorrelation, generalized estimating equations, marginal.

INTRODUCCIÓN

El análisis de sobrevivencia es el conjunto de herramientas estadístico-matemáticas que se han utilizado en las Ciencias de los Alimentos para realizar las inferencias respectivas a la vida de anaquel y la vida de anaquel sensorial de alimentos [1], [2], [3] y [4]. Hough *et al.* [1] fueron los primeros en aplicar análisis de sobrevivencia para estimar la vida de anaquel sensorial considerando la naturaleza censurada de los datos; de allí en adelante, se ha aplicado tal metodología, lo que ha resultado en la creación de diversas aplicaciones [2] y [3].

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, de manera que, en este tipo de estudios y en una primera aproximación, será pertinente considerar que los datos de cada evaluador forman un clúster con una estructura de autocorrelación y pueden ser usados en el modelaje estadístico.

Por otro lado, cabe señalar que, aparte de la estructura de agrupación y de autocorrelación mencionadas, existe la posibilidad de considerar otras circunstancias en el modelaje que representan factores inherentes en los individuos, para reconocer así diferencias entre evaluadores; por ejemplo, características sociodemográficas que pueden influenciar la evaluación del producto. Aquellos factores y otros inherentes de los individuos serán considerados en futuras investigaciones.

Para incorporar los anteriores aspectos y analizar datos provenientes de análisis sensoriales de alimentos, se hará uso del análisis de datos longitudinales, lo que 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.

Así las cosas, en el presente trabajo se lleva a cabo un modelaje estadístico de datos provenientes de análisis sensorial de alimentos, incorporando la estructura de agrupación de tipo clúster que éstos poseen, así como el patrón de asociación en el tiempo para el conjunto de datos.

FUNDAMENTOS TEÓRICOS

La característica que define un estudio longitudinal es que los individuos son medidos repetidamente a través del tiempo, en contraste con los estudios transversales, en que una sola respuesta es medida para cada individuo. Los estudios longitudinales pueden distinguir cambios en el tiempo dentro de los individuos (*ageing effects*)

de las diferencias entre las personas en sus líneas base (*cohort effects*).

Los datos longitudinales pueden ser recolectados prospectivamente, siguiendo los sujetos a través del tiempo; o retrospectivamente, extrayendo múltiples medidas de cada persona de un archivo histórico. En el caso de los estudios sensoriales de alimentos, éstos se pueden recolectar de manera prospectiva, que corresponde a un diseño tipo básico, y realizando las diferentes evaluaciones en un solo momento del tiempo, conocido como diseño en reversa, ver [5] para mayores referencias.

Hay tres enfoques para el modelaje de datos longitudinales discretos y continuos: usando extensiones de los modelos lineales generalizados, que son modelos marginales, modelos de efectos aleatorios y modelos de transición. En el presente trabajo se usa el enfoque de los modelos marginales.

En los modelos marginales, la regresión de la variable respuesta, Yij, sobre las variables explicativas es modelada separadamente de la correlación dentro de las personas, dado que los valores repetidos probablemente no son independientes; este análisis debe incluir también hipótesis acerca de la forma de la correlación. El enfoque de modelo marginal tiene la ventaja de modelar separadamente la media y la covarianza. En la regresión, modelamos la esperanza marginal, (Yij), en función de las variables explicativas sin tener en cuenta la dependencia entre observaciones. Este tipo de modelaje es bastante práctico y llamativo para un científico de alimentos porque le dará información concerniente del efecto del tiempo de almacenamiento sobre la aceptabilidad o rechazo del producto.

En esta primera aproximación del uso de datos longitudinales en estudios de vida de anaquel sensorial, tiene sentido físico suponer que las respuestas repetidas de cada consumidor pueden no tener una dependencia fuerte entre ellas, debido a que los científicos de alimentos diseñan sus pruebas sensoriales con el fin de que la respuesta de un consumidor en una i-ésima cata, no se vea alterada por las anteriores o posteriores, ejemplo de ello es la utilización de diseños en reversa, la utilización de cabinas sensoriales especiales para las degustaciones y el uso de limpiadores de paladar, entre otros [5].

Específicamente, un modelo marginal tiene los siguientes supuestos:

a. La esperanza marginal de la respuesta, (Yij)=μij, depende de las variables explicativas **xij** por h(μij) = **xij**'**β**, donde h es una función de enlace conocida, tal como logit para las respuestas binarias o log para conteos;

b. La varianza marginal depende de la media marginal de acuerdo a $Va(Yij)=\upsilon(\mu ij)\phi$, donde υ es una función de varianza conocida y ϕ es un parámetro de escala que podría ser estimado.

c. La correlación entre Yij y Yik es una función de la media marginal y quizás de parámetros adicionales $\boldsymbol{\alpha}$, es decir, $Cor(Yij,Yik) = \rho(\mu ij,\mu ik;\boldsymbol{\alpha})$, donde $\rho(\cdot)$ es una función conocida.

De manera que el modelo marginal está dado por:

$$logit(\mu_{ij}) = log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = log\left(\frac{p(Y_{ij}=1|T=t_{ij})}{1-p(Y_{ij}=1|T=t_{ij})}\right) = \alpha+\beta*t_{ij},\ i=1,2,\dots,n\ y\ j=1,2,\dots,m. \quad (1)$$

donde Yij = 1 denota que el i-ésimo consumidor rechazó la muestra del producto en el j-ésimo tiempo de almacenamiento, tij corresponde al j-ésimo tiempo de almacenamiento del i-ésimo consumidor, α y β son los parámetros a estimar del modelo que son calculados a través de las ecuaciones estimantes generalizadas, GEE por sus siglas en inglés, Generalized Estimating Equations.

Las GEE, son un método general para analizar datos recolectados en clústeres donde:

- Las observaciones dentro de un clúster pueden estar correlacionadas.
- Las observaciones en clústeres separados son independientes.
- Una transformación monótona de la esperanza está linealmente relacionada a las variables explicativas.
- La varianza es una función de la esperanza.

Como se mencionó antes, hay varios enfoques para modelar datos longitudinales, y así extender los modelos lineales generalizados (GLM, por sus siglas en inglés, Generalized Lineal Models) a datos longitudinales, los modelos de efectos mixtos y de transición que especifican completamente la distribución conjunta dentro de los clústeres vía variables latentes o dinámicas condicionales; pero con la presencia de efectos aleatorios, la estimación de verosimilitud necesita la integración sobre las distribuciones de efectos aleatorios, que quizás sean numéricamente intratables. Así que, comparados con estos enfoques, el método GEE ajusta modelos de media marginal, con la ventaja de que solamente es necesaria la especificación correcta de las medias marginales para que los estimadores de los parámetros $\hat{\alpha}$ y $\hat{\beta}$ sean consistentes y asintóticamente normales.

En la metodología de las ecuaciones estimantes generalizadas, el usuario puede fijar una es-

tructura de correlación, que con frecuencia es llamada, matriz de correlación de trabajo. Algunas estructuras de correlación de trabajo incluyen, intercambiable, la independiente, no estructurada y autoregresiva.

METODOLOGÍA

Para los análisis estadísticos, se hace uso del paquete geepack [5] disponible en el *Software* R; todas las tablas y figuras presentadas son de creación propia.

El paquete *geepack* trabaja bajo el enfoque de las Ecuaciones Estimantes Generalizadas, mencionadas antes (GEE por sus siglas en inglés, Generalized Estimating Equations), que son un método general para analizar datos recolectados en clústeres que surgen de medir repetidamente a los individuos a través del tiempo. A continuación se describe el uso de dicha función colocando los nombres en inglés, ya que así está predeterminado en el *software* R.

LA FUNCIÓN GEEGLM

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

family: la función de varianza es especificada por el argumento *family* y es identificada por el nombre de la distribución correspondiente en un modelo lineal generalizado. En la Tabla 1 se especifican las familias más representativas con las que se cuentan y sus respectivas funciones de varianza, (μ).

Tabla.1 Opciones del argumento *family* en la función *geeglm*.

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

corstr: las estructuras de correlación de trabajo ("*working*") predefinidas son especificadas con este argumento, en la Tabla 2 se muestran las diferentes opciones de estructuras de correlación y sus respectivas funciones de correlación, $R(\alpha)$.

Tabla 2. Opciones del argumento *corstr* en la función *geeglm*.

Nombre	Función de correlación
Independence	$COR(Yit,Yit') = 0$, $t \neq t'$
Exchangeable	$COR(Yit,Yit') = \alpha$, $t \neq t'$
ar1	$COR(Yit,Yit') = \alpha\|t-t'\|$, $t \neq t'$
Unstructured	$COR(Yit,Yit') = \alpha tt'$, $t \neq t'$

Los datos que provienen de estudios sensoriales de alimentos y a los cuales se les realizan los análisis estadísticos en el presente trabajo son de dominio público y se pueden encontrar en [5], a su vez, se pueden descargar en archivo .xlsx de la página web del editor.

Para la generación de la base de datos, los tecnólogos de alimentos básicamente seleccionan una muestra de consumidores, a los cuales se les pide prueben un conjunto de muestras de cierto alimento, con diferentes tiempos de almacenamiento, y respondan "sí" o "no" a la pregunta "¿normalmente consumiría este producto?"; de esto, generan un conjunto de datos como los mostrados en la Tabla 3.

Tabla 3. Ilustración de un conjunto de datos que se generan en análisis sensoriales de alimentos

	Tiempos de almacenamiento					
Consumidor	t_0	t_1	t_2	t_3	…	t_n
1	no	no	si	si	…	no
2	si	si	si	si	…	no
3	si	no	si	no	…	si
4	no	si	si	no	…	no
5	si	si	si	si	…	no
⋮	⋮	⋮	⋮	⋮	⋮	⋮
N	si	No	no	no	…	no

RESULTADOS Y DISCUSIÓN

Un modelo marginal logístico con la misma estructura media y la función de varianza binomial $(\mu it) = \mu it(1 - \mu it)$, utilizando el enfoque GEE, es ajustado con diferentes estructuras de correlación; los resultados obtenidos se muestran a continuación:

Utilizando una estructura de correlación independiente:

	intercepto	tiempo
Estimación	-1.86816	0.07457
Error estándar	0.20197	0.00856
Wald	85.6	76.0
Pr(>|W|)	<2e-16 ***	<2e-16 ***

Utilizando una correlación intercambiable:

	intercepto	tiempo
Estimación	-1.8431	0.0745
Error estándar	0.2020	0.0086
Wald	83.2	75.0
Pr(>|W|)	<2E-16 ***	<2E-16 ***

Utilizando una correlación autorregresiva de primer orden:

	intercepto	tiempo
Estimación	-1.88795	0.07497
Error estándar	0.20124	0.00868
Wald	88.0	74.7
Pr(>|W|)	<2E-16 ***	<2E-16 ***

Utilizando una correlación no estructurada:

	intercepto	tiempo
Estimación	-1.92043	0.07236
Error estándar	0.19851	0.00804
Wald	93.6	80.9
Pr(>|W|)	<2E-16 ***	<2E-16 ***

A su vez, se ajusta un modelo logístico mediante el enfoque de modelos lineales generalizados, lo que sería apropiado si no hubiera en los datos la estructura de tipo clúster, y no hubiera dispersión en las probabilidades de respuesta para los consumidores con los mismos valores de covariables; los resultados se muestran abajo:

	intercepto	Time
Estimación	-1.86816	0.07457
Error estándar	0.21080	0.00845
Wald	--8.86	8.83
Pr(>|W|)	<2E-16 ***	<2E-16 ***

Todos los resultados anteriores se resumen en las Tablas 4 y 5, y se puede apreciar que, en todos los casos, las estimaciones de los parámetros de cada modelo ajustado son cercanas y los errores estándar son aproximadamente iguales.

Para seleccionar el mejor modelo, Pan propuso un método de selección de modelo para GEE, que llamó Criterio de Información de Cuasiverosimilitud (QIC por sus siglas en inglés Quasilikelihood information criterion) [7] y [8], que también puede ser utilizado para seleccionar la mejor estructura de correlación de trabajo en el análisis GEE. Los resultados del coeficiente QIC se resumen en la Tabla 6.

Tabla 4. Estimaciones de los parámetros de los modelos ajustados.

	intercepto	time
Estimación	-1.87*	0.07*
Error estándar	0.21	0.01
Estimación	-1.87[1]	0.07[1]
Error estándar	0.20	0.01
Estimación	-1.84[2]	0.07[2]
Error estándar	0.20	0.01

*Usando el enfoque de modelos lineales generalizados y el enfoque de las ecuaciones estimantes generalizadas con estructura de correlación: 1Independiente, 2Intercambiable.

Tabla 5. Estimaciones de los parámetros de los modelos ajustados.

	intercepto	Time
Estimación	1.89[3]	0.07[3]
Error estándar	0.20	0.01
Estimación	-1.92[4]	0.07[4]
Error estándar	0.20	0.01

Usando el enfoque de las ecuaciones estimantes generalizadas con estructura de correlación: 3Autorregresivo de orden uno y 4No estructurado.

Tabla 6. Criterio de información de cuasiverosimilitud para los diferentes modelos ajustados.

Modelo	Estructura de correlación	Criterio de información de cuasiverosimilitud
modelo.uns	no estructurada	375
modelo.ar	Autorregresivo de orden 1	376
modelo.ind	independiente	376
modelo.exc	intercambiable	376

La estructura de correlación no estructurada tuvo el valor más pequeño de QIC, aunque no difiere en gran medida de los demás, se puede considerar que el modelo bajo la estructura de correlación no estructurada es el mejor. Además, en general, sí el número de unidades por clúster es pequeño en un diseño balanceado y completo, entonces una matriz no estructurada es la recomendada [7] y [8].

Teniendo en cuenta lo anterior, escogemos el modelo con estructura de correlación no estruc-

turada, y así un modelo marginal análogo a la Ec. (1), vendría dado por:

$$logit(\hat{\mu}_{ij}) = log\left(\frac{\hat{\mu}_{ij}}{1-\hat{\mu}_{ij}}\right) = -1.92 + 0{,}07t_{ij}, i = 1, 2, \ldots, 50 \; y \; j = 1, 2, 3, \ldots, 7. \quad (2)$$

donde Yij=1 denota que el i-èsimo consumidor rechazó la muestra del producto en el j-èsimo tiempo de almacenamiento. De manera que:

$$\hat{\mu}_{ij} = \hat{P}(\boldsymbol{Y} = 1|\boldsymbol{T} = t) = \frac{e^{-1.92+0{,}07*t}}{1+e^{-1.92+0{,}07*t}} \quad (3)$$

Por tanto:

$$\hat{P}(\boldsymbol{Y} = 1|\boldsymbol{T} = t) = \frac{e^{-1.92+0{,}07*t}}{1+e^{-1.92+0{,}07*t}} \quad (4)$$

El modelo estimado en la expresión (4) se puede representar mediante la Fig. 1, que representa la probabilidad de rechazo estimada en el tiempo.

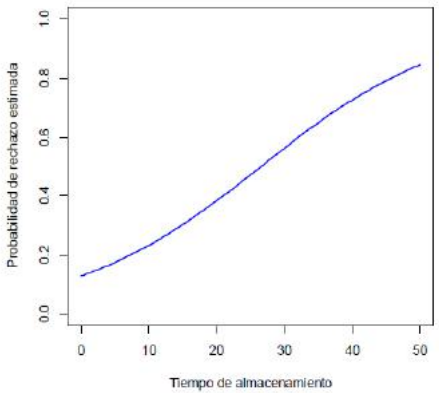

Figura 1. Estimación de la probabilidad de rechazo por parte de los consumidores en el tiempo.

Con el enfoque GEE se puede verificar que, en el tiempo cero, ya hay una probabilidad de rechazo estimada aproximado del 13%, que nos lleva a pensar que el modelo marginal reconoce la historia completa del consumidor respecto a su percepción del producto a través del tiempo, a diferencia del modelo para datos censurados, que otorga un 0% de rechazo al tiempo cero de almacenamiento, esto se puede ver en la Fig 2.

En particular, al tiempo cero existen cuatro consumidores que manifestaron rechazo, consumidores que son retirados del análisis estadístico realizado con el enfoque de análisis de sobrevivencia; en cambio, en el modelaje aquí presentado no fue necesario retirarlos, de manera que el modelo marginal tiene en cuenta la posibilidad de que haya consumidores que no acepten el producto fresco pero que puedan aceptarlo eventualmente.

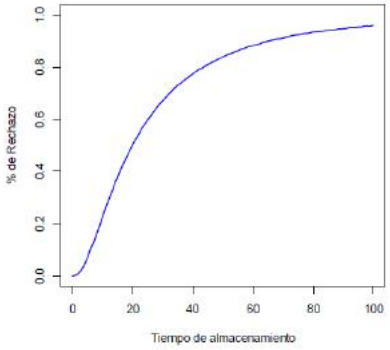

Figura 2. Función de distribución acumulada para la vida de anaquel sensorial.

La Tabla 7 muestran las estimaciones de las vidas de anaquel estimadas mediante el modelaje de datos censurados para 10, 20 y 25 % de rechazo.

Tabla 7. Estimaciones de la vida de anaquel sensorial y sus intervalos de confianza del 95%.

vida de anaquel sensorial (horas)	Error estándar	Límite Inferior	Límite Superior	Porcentaje de rechazo
6.03	1.33	3.91	9.30	10
9.08	1.67	6.33	13.02	20
10.60	1.83	7.56	14.87	25

Mediante el análisis de datos longitudinales, se estiman las probabilidades de rechazo para las vidas de anaquel sensorial de 6, 9 y 11 horas, que se pueden ver en la Tabla 8.

Tabla 8. Estimaciones de la probabilidad de rechazo y sus intervalos de confianza del 95%.

Vida de anaquel sensorial (horas)	Estimación probabilidad de rechazo	Error estándar	Límite Inferior	Límite Superior
6	0.18	0.027	0.13	0.24
9	0.22	0.030	0.16	0.28
11	0.25	0.031	0.18	0.31

Claramente, los resultados encontrados con el enfoque de análisis de datos longitudinales, di-

fieren de los encontrados con el análisis de sobrevivencia, debido a que se modelan variables aleatorias completamente diferentes. Con el modelo aquí presentado, podemos responder a la pregunta de cuál es la probabilidad de rechazo poblacional cuando el tiempo de vida de anaquel sensorial fuera especulado por el analista de alimentos.

Sin embargo, se puede notar que los resultados generados por el modelo marginal reflejan que, por debajo de las 8 horas de almacenamiento, las probabilidades de rechazo son superiores a 0.15, lo que representa una mayor probabilidad de rechazo en las primeras horas de almacenamiento que el modelo de análisis de sobrevivencia; por otro lado, para tiempos de almacenamiento por encima de las 12 horas, las probabilidades de rechazo son menores comparadas con los porcentajes de rechazo reportados por el modelaje de datos censurados.

CONCLUSIONES

El modelo marginal permite modelar las respuestas directas de los consumidores, evitando la censura de las mismas y presuponiendo que ello generará estimaciones más realistas de la percepción sensorial que tiene el consumidor en torno al producto alimenticio.

AGRADECIMIENTOS

Los autores agradecen al Consejo Nacional de Ciencia y Tecnología por su apoyo durante toda esta investigación.

REFERENCIAS

[1] G. Hough, K. Langhor, G. Gòmez, and A. Curia, "Survival analysis applied to sensory shelf-life of foods", *J. Food Sci*. Vol. 68, Nr. 1, pp. 359-362, 2003.

[2] D.A. Jacobo-Velasquez, P.A. Ramos-Parra, and C. HERNANDEZ-BRENES, "Survival analysis applied to the sensory shelf-life dating of high hydrostatic pressure processed avocado and mango pulps", *J. Food Sci*. Vol. 75, Nr. 6, pp. 286-291, 2010.

[3] A. Cruz, E. Waltert, R. Silva, J. Faria, H. Bolini, H. Pinheiro, and A. Santana, "Survival analysis methodology to predict the shelf-life of probiotic flavored yogurt", *Food Res*. Vol. 43, Issue 5, pp. 1444-1448, 2010.

[4] A. Giménez, F. Ares, G. Ares, "Sensory shelf-life estimation: A review of current methodological approaches", *Food Research International*. Vol. 49, Issue 5, pp. 311-325, 2012.

[5] G. Hough, *Sensory Shelf Life Estimation of Food Products*. UK: Taylor and Francis Group. 2010.

[6] U. Halekoh, S. Højsgaard, J. Yan, "The R Package geepack for Generalized estimating Equations", *Journal of Statistical Software*, Volume 15, Issue 2., pp. 1-11, 2006.

[7] W. Pan, "Akaike's Information Criterion in Generalized Estimating Equations", *Biometrics* Vol. 57, pp. 120-125, 2001.

[8] W. Pan, "Goodness of fit Tests for GEE with Correlated Binary Data", Scand J Statist Vol. 29, pp. 101-110, 2002.

[9] Y. Kwon, T. Park, A. Ziegler, and M. Paik, "Generalized estimating equations with stabilized working correlation structure", Computational Statistics and Data Analysis, Vol. 106, pp.1–11, 2017.

[10] Z. Liu, Methods and Applications of Longitudinal Data Analysis. Elsevier Inc. 2016.