Anales científicos UNALM, Vol. 70 N° 3, 2009
ISSN 0255-0407
Recibido: 09/07/2008
Aceptado: 21/07/2008

Modelos lineales mixtos con estructura de correlación en el análisis de datos longitudinales. Un caso aplicado

Ana Vargas P.[1]

Resumen

Objetivo. Identificar variables relevantes del entorno de la UNALM para caracterizar diversas tendencias que Es usual que los resultados de un experimento con medidas repetidas a través del tiempo en una misma unidad experimental sea analizado utilizando el análisis de varianza de parcelas divididas (análisis clásico), pese a que en la mayoría de los casos no siempre es posible asumir independencia entre las mediciones porque los datos longitudinales normalmente presentan correlación entre las medidas hechas a una misma unidad experimental. Una alternativa de análisis, con datos provenientes de un diseño longitudinal, es utilizar un modelo lineal mixto, que incluya una estructura de correlación entre las mediciones de cada unidad experimental. En este trabajo se aplican el análisis de varianza de parcelas dividas y modelos lineales mixtos con estructura de correlación entre las mediciones, para el análisis de muestras de leche tomadas con el fin de determinar si existe diferencias en el nivel de acidez entre muestras de leche tomadas en cuatro localidades distintas a través del tiempo. Para la estimación del modelo lineal mixto se eligió como estructura de correlación entre las mediciones de cada muestra la unstructured (no estructurada) por presentar un mejor ajuste (BIC=-13.2) entre cuatro distintas estructuras evaluadas (simetría compuesta, AR (1), toeplitz). Al comparar los resultados de la interacción tratamiento y tiempo del ANVA obtenido con el análisis clásico (parcelas divididas) muestra una no significancia (p-valor=0.1082) a diferencia del obtenido incluyendo la estructura de correlación ajustado a un modelo lineal mixto (p-valor=0.0406).
Palabras clave:

Abstract

It is usual to apply split-plot analysis to analyze experiments with repeated measures, despite of not being possible to assume independence between measures, because of longitudinal data generally have correlation between measures done at the same experimental unit. An alternative of analysis for this data is to include variance-covariance structure of repeated measures. This work applies two methodologies: classic (split-plot) and mixed linear models with variance-covariance structure of correlations between measures to determine whether there are differences between acidity levels of samples of milk come from of four different localities a long of the time. For estimation of mixed model was considered as variance-covariance structure of correlation "unstructured", because the fit obtained was the most appropriate between four structure evaluated (BIC=-13.2). Comparing results between both methodologies, we concluded there isn't significance in the interaction between treatment and time if we apply split-plot analysis (p-value<0.1); however there is significance if we apply mixed linear models with covariance structure of correlations (p-value=0.0406.
Key words:

1. Introducción [1]

1.1.- Fundamentación del problema de investigación

Las medidas repetidas involucran el registro de la variable respuesta o dependiente bajo diversas condiciones. En un contexto experimental, estas condiciones pueden ser: diferentes tratamientos u ocasiones de medidas antes, durante o después de la intervención. En un contexto observacional, las medidas se registran en distintos intervalos temporales. En ambos contextos cuando el interés es estudiar la respuesta a través del tiempo, el diseño se concibe como longitudinal.

Los estudios con datos longitudinales son usados principalmente por tres razones: a) incrementa la sensibilidad considerando la variación dentro de cada unidad experimental, b) estudia cambios a través del tiempo, pues las mediciones repetidas proporcionan información sobre la tendencia en el tiempo de la variable respuesta bajo diferentes condiciones de tratamiento y c) uso eficiente de unidades experimentales, ya que por lo general se requiere menos unidades experimentales.

Un estudio con estas características fue realizado por Silvia Peralta[2] quién registró evaluaciones del nivel de acidez (ºD) en muestras de leche, en cuatro momentos: a las 3 horas (tiempo 1), 6 horas (tiempo 2), 9 horas (tiempo 3) y 12 horas (tiempo 4), en cuatro localidades diferentes: Trapiche, Cerro, Huatocay y Macas, con el objeto de determinar si existen diferencias en los nivel de acidez en función del tiempo entre estas localidades.

1.2.- Formulación del problema de investigación

El problema a resolver en la presente investigación corresponde a dar respuesta a las siguientes interrogantes: ¿existe diferencias en los niveles de acidez, en muestras de leche de cabra a lo largo del tiempo, entre las localidades de Macas, Huatocay, Trapiche y Cerro?, además de ¿las conclusiones obtenidas con el método tradicional de análisis univariado son las mismas que con el método de análisis de modelos mixtos?

[1] Facultad de Economia y Planificación, Universidad Nacional Agrária La Molina. E-mail: anavargasp@hotmail.com

1.3.- Objetivos de la investigación

Por tanto este trabajo se propone:

Determinar si existe diferencias en el nivel de acidez (en muestras de leche de cabra) en función del tiempo en las localidades de estudio, utilizando el análisis de varianza de parcela divididas (método clásico).

Determinar si existen diferencias en el nivel de acidez (en muestras de leche de cabra) en función del tiempo en las localidades de estudio, aplicando un modelo lineal mixto con una estructura de correlación entre las mediciones de una misma muestra ajustado utilizando SAS®.

Comparar los resultados del análisis de varianza obtenido con el análisis clásico y con el análisis de modelos lineales mixtos.

1.4. Justificación de la investigación

Es usual que los resultados de un estudio como el descrito anteriormente, es decir, con medidas repetidas sea analizado a través del Análisis Univariado, ANOVA clásico (parcelas dividas), pero la aplicación de este requiere satisfacer los supuestos de normalidad independencia y esfericidad, éste último se refiere a que las diferencias de varianzas correspondientes a las distintas ocasiones de medidas son iguales.

Sin embargo en el análisis de un experimento con medidas repetidas, por ejemplo, ocurre con cierta frecuencia que la matriz de covarianzas no se ajusta al supuesto restrictivo de esfericidad, o de igualdad de varianzas para cualquier par de diferencias entre tratamientos. Por lo que utilizar el análisis de varianza clásico en estos casos conlleva a que la probabilidad de rechazar la hipótesis nula cuando es cierta, esté por encima del criterio fijado por el investigador.

Entre las alternativas propuestas de análisis estadístico para los diseños de medidas repetidas cuando no se cumple el supuesto de esfericidad se encuentra el enfoque de modelo lineal mixto, el cual permite modelar la estructura de la matriz de covarianzas y sus diferencias entre los grupos en función de la descripción de los datos. Bajo este enfoque, la estructura de la matriz de covarianzas más adecuada se selecciona previamente mediante algún criterio de ajuste, como las medidas: AIC de Akaike o el BIC de Schwarz, etc.

Este trabajo por tanto analiza la información lograda por Silvia Peralta, bajo dos formas de análisis, y compara sus resultados, ignorando, no por menos importante, los análisis pos-anva.

Cochran y Cox (1957) discuten el diseño clásico de parcelas divididas como alternativa para el análisis de diseño de medidas repetidas.

Greenhouse y Geisser (1959) pusieron de manifiesto que si la matriz de varianza-covarianza tiene forma arbitraria los distintos estadísticos asociados con las medidas repetidas se siguen distribuyendo de acuerdo con la F ordinaria, pero con los grados de libertad corregidos en función de la desviación de la matriz de varianza-covarianza del patrón de uniformidad requerido, proponiendo la utilización de una prueba F con grados de libertad ajustados.

Huynh Feldt (1970) demostraron además que la razón F (estadístico de prueba) también se distribuye como F siempre y cuando las diferencias entre las varianzas de cualquier par de tratamientos son iguales. Por tanto proponen la extensión de la utilización del análisis clásico en casos cuando es posible probar que se cumple la condición señalada.

Alternativamente Cole y Grizzle (1966) partiendo del hecho de que las observaciones tratadas con diseño de parcelas divididas están correlacionadas, y por tanto son esencialmente de naturaleza multivariada, han sugerido que el procedimiento adecuado para analizar tales diseños es el análisis multivariado, muy recomendado cuando existe una grave desviación del supuesto de esfericidad y la muestra es grande (Jensen, 1982)

Otra alternativa desarrollada es el de enfoque del modelo lineal mixto, el cual permite modelar la estructura de la matriz de covarianzas y sus diferencias entre los grupos en función de la descripción de los datos (Cnaan, Laird y Slasor, 1997: Laird y Ware, 1982; Littell, Milliken, Stroup y Wolfinger, 1996).

De esta forma, la estructura de la matriz de covarianzas más adecuada se selecciona previamente mediante algún criterio, como el AIC de Akaike o el BIC de Schwarz. No obstante, pese a ser un método robusto cuando se selecciona la matriz de covarianza más adecuada, la principal dificultad de este procedimiento radica precisamente en la modelización de estas matrices de covarianzas. (Wolfinger, 1996).

SAS® (PROC MIXED) ofrece varias estructuras de covarianzas a ajustar entre ellas: simetría compuesta, no estructurada, autorregresivo de primer orden o de coeficientes aleatorios, etc.

2. Materiales y métodos Introducción

2.1.- Modelo lineal mixto LMM

En un modelo lineal (sólo con efectos fijos) el valor esperado de la variable respuesta normal y es de la forma:

$E(y) = X\beta,$ donde: $y \sim N(X\beta, R)$

En un modelo lineal mixto (LMM), es decir con una mixtura de efectos fijos y aleatorios, con variable respuesta normal y, es de la forma:

$$E(y/u) = X\beta + Zu, \ldots\ldots (1)$$

Donde:

β: vector de los efectos fijos.

u: vector de efectos aleatorios que ocurre en el vector de respuesta y, además se supone que $u \sim N(0, D)$.

X y Z son matrices conocidas del modelo (matrices del diseño).

Por tanto $y \sim N(X\beta, ZDZ^T + R)$, a partir del cual observamos que los efectos fijos se involucraran en la estimación de la media y los efectos aleatorios en la estimación de la varianza de y.

McCullogh (2001) muestra como los datos longitudinales pueden ser arregladas dentro de los modelos lineales mixtos (LMM), por ejemplo para el modelo estadístico para varios tratamientos (sin iteracciones), indica:

$E(y_{ijk}/u_{ij}) = \mu + \alpha_i + \gamma_k + u_{ij}$, para: i=1, ..., t j=1, ..., m k=1, ..., n.

Donde:

$E(y_{ijk}/u_{ij})$: respuesta esperada en el k-ésimo tiempo de la j-ésima unidad experimental asignada al i-ésimo tratamiento condicionado al valor de u_{ij}.

μ: media general.

α_i: efecto fijo del tratamiento i-ésimo.

u_{ij}: efecto aleatorio de la unidad experimental j, asignado al tratamiento i

λ_k: efecto fijo del tiempo k.

El cual puede ser expresado en términos matriciales como: $E(y/u) = X\beta + Zu,$ para $\beta = \begin{bmatrix} \mu \\ \{_c\,\alpha_i\}_{i=1}^{t} \\ \{_c\,\gamma_k\}_{k=1}^{n} \end{bmatrix}$, $u = \{_c\,\{_c\,b_{ij}\}_{i=1}^{t}\}_{j=1}^{m}$ vector aleatorio para el efecto de la unidad experimentalo, con X matriz diseño particionada como: $X = [1_{tmn} \quad I_t \otimes 1_{mn} \quad 1_{tm} \otimes I_n \quad I_t \otimes 1_m \otimes I_n]$ y $Z = I_{tm} \otimes 1_n$

2.2.- Modelo estadístico para medidas repetidas

Considere que se ha recogido información de unidades experimentales las cuales han sido asignadas aleatoriamente a un tratamiento del factor en estudio y las mediciones de la variable respuesta han sido hechas en tiempos fijados igualmente espaciados sobre cada unidad experimental.

Sea Y_{ijk} la medición en el tiempo k (k=1, ..., n) sobre la unidad experimental j (j=1, ..., m) asignado al tratamiento i (i=1, ..., t). El modelo estadístico para estas medidas es:

$E(y_{ijk}/b_{ij}) = \mu + \alpha_i + b_{ij} + \gamma_k + \alpha\gamma_{ik}$, ...(2)

donde:

$E(y_{ijk}/b_{ij})$: respuesta esperada en el k-ésimo tiempo de la j-ésima unidad experimental asignada al i-ésimo tratamiento condicionado al valor de b_{ij}.

μ: media general.

α_i: efecto fijo del tratamiento i-ésimo.

b_{ij}: efecto aleatorio de la unidad experiemntal j signada al tratamiento i

γ_k: efecto fijo del tiempo k.

$\alpha\lambda_{ik}$: efecto de la interacción del tratamiento i con tiempo k.

Para la descripción de (2) de la forma dada en (1), siguiendo la notación de McCulloch (2001) se expresa como:

$y = X\beta + Zu + e,$ para $\beta = \begin{bmatrix} \mu \\ \{_c\,\alpha_i\}_{i=1}^{t} \\ \{_c\,\gamma_k\}_{k=1}^{n} \\ \{_c\,\{_c\,\alpha\gamma_{ik}\}_{i=1}^{t}\}_{k=1}^{n} \end{bmatrix}$,

$u = \{_c\,\{_c\,b_{ij}\}_{i=1}^{t}\}_{j=1}^{m}$ con X, particionado como: $X = [1_{tmn} \quad I_t \otimes 1_{mn} \quad 1_{tm} \otimes I_n \quad I_t \otimes 1_m \otimes I_n]$, y $Z = I_{tm} \otimes 1_n$.

Donde: $\otimes$ representa el producto directo (Kronecker).

Por lo tanto:

$E(y) = X\beta$ y

$V(y) = ZDZ^T + R = (I_{tm} \otimes 1_n)(D \otimes 1)(I_{tm} \otimes 1_n^T) + R = D \otimes J_n + R$, ...(3)

Donde: J_n es una matriz cuadrada de unos de dimensión "n".

Si para el modelo (2) asumimos una estructura de correlación entre las unidades experimentales, y la misma varianza para cada unidad experimental, tomamos:

$\text{var}(u_l) = \sigma_u^2 \quad para\ todo\ l, \quad l = 1,...,tm$ y

$corr(u_{l,}u_{l'}) = \rho_u \quad para\ todo\ l \neq l'$

Entonces, D es una matriz con elementos σ_u^2 en la diagonal y elementos $\rho_u\sigma_u^2$ fuera de la diagonal, por tanto (3), es simplificada como:

$V(y) = \sigma_u^2[(1-\rho_u)I_{tm} + \rho_u J_{tm}] \otimes J_n + R$ .... (4)

2.3.- Análisis de varianza univariado para mediciones repetidas.

El uso del modelo de parcelas divididas para el análisis de datos longitudinales es utilizado cuando es posible suponer que las correlaciones entre unidades experimentales y dentro de unidades experimentales son constantes, esta estructura de covarianza es llamado simetría compuesta.

En el modelo (2), (4) tiene una matriz R, definida por:

$R = I_{tm} \otimes R_0$,

donde: R_0 es la matriz de covarianzas, de dimensión n × n, es decir la estructura de R_0 es de la forma:

$$R_0 = \begin{bmatrix} \sigma_e^2 & 0 & \dots & 0 \\ 0 & \sigma_e^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_e^2 \end{bmatrix}_{n\times n} = \sigma_e^2 I_n$$

Si son constantes las correlaciones dentro de las unidades experimentales y entre unidades experimentales, entonces (4) se simplifica a:

$V(y) = \sigma_u^2[(1-\rho_u)I_{tm} + \rho_u J_{tm}] \otimes J_n + I_{tm} \otimes R_0$ . . . . (5)

Si son constantes las correlaciones dentro de las unidades experimentales y no existe correlación entre unidades experimentales, entonces en (5) se debe hacer $\rho_u = 0$, por tanto queda simplificado en:

$V(y) = I_{tm} \otimes (\sigma_u^2 J_n + R_0) = I_{tm} \otimes (\sigma_u^2 J_n + \sigma_e^2 I_n)$ . . . . (6)

Luego: $\text{var}(y_{ijk}) = \sigma_u^2 + \sigma_e^2$ y $\text{cov}(y_{ijk}, y_{ijk'}) = \sigma_u^2$, consecuentemente la correlación entre dos mediciones de una misma unidad experimental en (5) es:

$\rho = \dfrac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$, llamado correlación intra-clase.

Huynh-Feldt sugirieron también el uso del análisis de parcelas divididas en una situación menos estricta que la de simetría compuesta, si es posible suponer que la diferencia para todos los pares posibles de observaciones tomadas en diferentes períodos en un mismo sujeto es la misma, es decir:

$\text{var}(y_{ijk} - y_{ijk'}) = 2\lambda \quad \forall k \neq k' \quad y \quad \lambda > 0$

Por tanto, será suficiente realizar un análisis de parcelas divididas si es posible suponer que por lo menos se cumple la condición de Huynh-Feldt sobre las mediciones repetidas. Una prueba para probar esta condición es la prueba de Mauchly, (Ho: cumple con la condición de Huynh-Feldt), si no existe evidencia de rechazar la hipótesis planteada, entonces las pruebas F en el análisis de varianzas es válida, de lo contrario se hace necesario ajustar los grados de libertad del estadístico F en el Anva, según Grenhouse y Geisser. Sin embargo la prueba de Mauchly es recomendada sólo si se tiene muestras grandes en cada grupo, Kuehl (2001) recomienda no confiar por completo en esta prueba, y que más bien la decisión para realizar un análisis de varianza univariado podría basarse en la experiencia del investigador así como también en las características específicas del material de investigación.

2.4.- Algunas estructuras de covarianzas alternativas.

a) La estructura más simple es la que asume independencia dentro de las mediciones de cada unidad experimental en el tiempo e independencia entre las mediciones de las unidades experimentales, es decir en (6) tiene $\sigma_u^2 = 0$. Bajo esta estructura se cumplen los supuestos del ANVA clásico, entonces: $V(y) = I_{tm} \otimes (R_0) = I_{tm} \otimes (\sigma_e^2 I_n)$. En este caso se requerirá estimar, en la estructura de covarianza, sólo el parámetro σ_e^2.

La estructura más compleja de covarianza es en la cual las observaciones para cada par de mediciones dentro de una misma unidad experimental tiene su propia y única correlación, asumiendo no correlación entre unidades experimentales y σ_u^2 constante conocida, entonces en $V(y) = I_{tm} \otimes (\sigma_u^2 J_n + R_0)$, donde:

$$R_0 = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_{kk}^2 \end{bmatrix}$$, se tiene $k + k(k-1)/2$ parámetros a estimar.

c) La estructura de simetría compuesta fue descrita en 3.1, y es la forma más sencilla de modelar la correlación entre mediciones de una misma unidad experimental pues supone que éstas son constantes. Para esta estructura se requerirá estimar: σ_u^2 y σ_e^2.

d) Es muy frecuente esperar que las correlaciones entre observaciones estén en función de su rezago en el tiempo, es decir las observaciones en tiempos más próximos tienen una correlación más alta que en tiempos más lejanos, un modelo que puede describir tal relación es el modelo autoregresivo de primer orden.

Asumiendo σ_u^2 constante conocida, entonces $V(y) = I_{tm} \otimes (\sigma_u^2 J_n + R_0)$ donde:

$$R_0 = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho^{k-1} \\ \rho & 1 & \cdots & \rho^{k-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{k-1} & \rho^{k-2} & \cdots & 1 \end{bmatrix}$$ y

$\sigma^2 = \dfrac{\sigma_e^2}{(1-\rho)^2}$. En este caso se requerirá estimar: ρ y σ_e^2.

e) La estructura Toeplitz considera también que observaciones en tiempo cercanos son más altas que en tiempos lejanos, pero cada correlación es única para cada par de mediciones dentro de cada sujeto.

Asumiendo σ_u^2 constante conocida, $V(y) = I_{tm} \otimes (\sigma_u^2 J_n + R_0)$, donde:

$$R_0 = \sigma_e^2 \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \cdots & \rho_{k-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \cdots & 1 \end{bmatrix}$$. En este caso se requerirá estimar: $\rho_1, \rho_2, ..., \rho_{k-1}$ y σ_e^2.

2.5.- Sobre el método de estimación.

Los parámetros de covarianzas son estimados usando métodos basados en verosimilitud,

como el máximo verosímil (ML) o el método máximo verosímil restringido (REML), los cuales en mucho de los casos derivan en procedimientos iterativos que maximizan la función de parámetros (verosimilitud). En esta aplicación en particular se hizo uso del método máximo verosímil (ML), la cual ha sido implementado como un método específico en las rutinas de SAS® para la estimación de modelos mixtos.

2.6.- Procedimiento de análisis de datos

Se realizó un análisis exploratorio datos, que consistió en el gráfico de perfiles con la finalidad de observar posibles diferencias entre las localidades y se estimó la matriz de correlación de Pearson para cada par de mediciones para observar la posible relación entre las mediciones.

Posteriormente se realizó el análisis de varianza utilizando el diseño de parcelas divididas, y se verificó el cumplimiento de la condición de Huynh-Feldt a través de la prueba de Mauchly. Al rechazarse la condición se procedió a realizar los ajustes necesarios en el estadístico de prueba en el análisis de varianza univariada (según Greenhouse y Geisser). Asimismo se realizó el análisis a través de contrastes polinomiales para analizar la tendencia en el tiempo.

Posteriormente se realizó el análisis utilizando modelos lineales mixtos ($\rho_u = 0$), eligiendo una de las cuatro siguientes estructuras de covarianzas: simétrica compuesta, autorregresiva de primer orden, Toeplitz y no estructurada, partir del estudio de medidas de bondad de ajuste como BIC y AIC.

3. Resultados

3.1.- Análisis exploratorio de los datos.

La figura 1 muestra el gráfico de perfiles para las medias de tratamientos a lo largo de los cuatro tiempos estudiados; parece que el nivel de acidez promedio en las muestras que proceden de Cerro fuesen mayores a las nueve horas (tiempo 3) y doce horas (tiempo 4) que en el resto de lugares, de forma similar la acidez promedio de las muestras de leche que proceden de Huatocay parecen ser menores a lo largo de los cuatro tiempos observados, asimismo parece no existir diferencias en el promedio de acidez a las 3 horas de observación (tiempo 1) entre los cuatro lugares de procedencia de las muestras de leche.

La matriz de correlación mostrada en el cuadro 1, genera la sospecha que las mediciones de las unidades experimentales (muestras de leche) están relacionadas a través del tiempo además se observa que esta relación es más fuerte en mediciones en tiempos cercanos y va disminuyendo conforme los tiempos son más distantes, como se aprecia en la tabla 1.

Figura 1. Gráfica de perfiles de cada grupo experimental en cada hora del estudio.

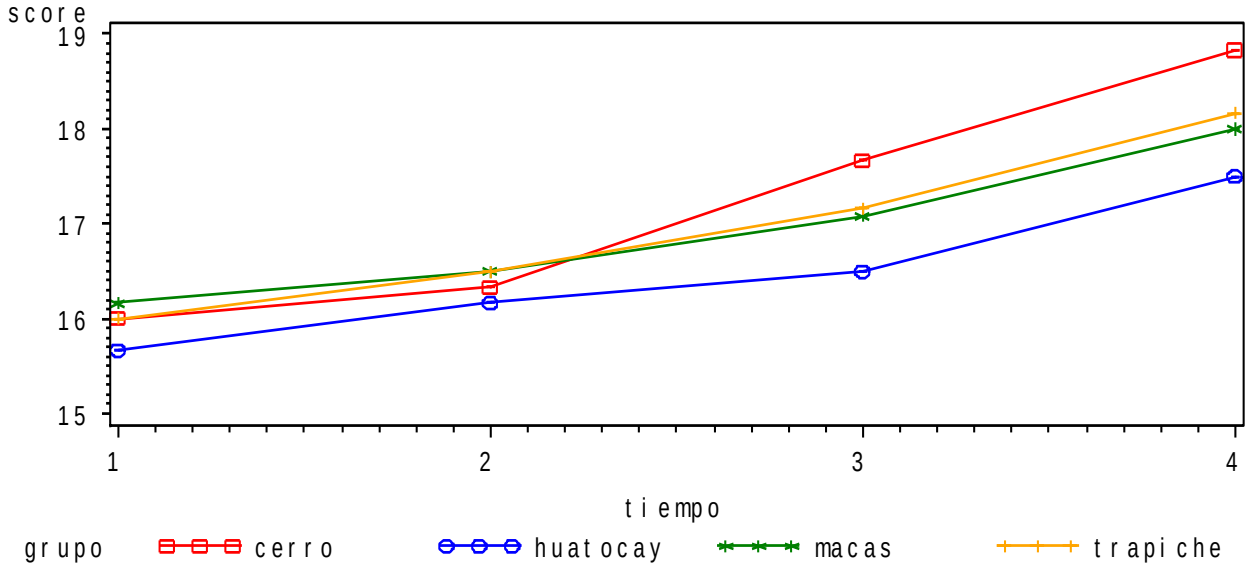

Tabla 1. Coeficientes de correlación de Pearson para cada par de mediciones.

Pearson Correlation Coefficients, N = 24				
t1		**t2**	**t3**	**t4**
t1	1.00000	0.96375	0.90335	0.86144
t2	0.96375	1	0.8658	0.86609
t3	0.90335	0.8658	1.00000	0.95826
t4	0.86144	0.86609	0.95826	1.00000

3.2.- Análisis clásico

En la tabla 2, se presenta la prueba de esfericidad de Mauchly que permite establecer si se cumple o no la condición de Huynh-Feldt para realizar el ajuste al estadístico F en el análisis de varianza de parcelas divididas para medidas repetidas.

Tabla 2. Test de Esferecidad de Mauchly's para probar la condición de Huynh-Feldt.

Sphericity Tests				
	Mauchly's			
Variables	DF	Criterion	Chi-Square	Pr > ChiSq
Transformed Variates	5	0.007515	32.877398	<.0001
Orthogonal Components	5	0.066711	18.199636	0.0027

No es posible suponer la condición de Huynh-Feldt, según los resultados de esta prueba (p-valor<0.0001), por lo que como fue descrito anteriormente, se hace necesario realizar ajustes al estadístico F_0 del análisis de varianzas (parcelas divididas), para aproximar mejor los niveles de significancia de las pruebas, el cuadro 3 muestra los ajustes bajo los criterios de Greenhouse y Geisser (G-G) y Huynh-Feldt (H-F) que proporciona SAS®.

Tabla 3. Análisis de varianza de medidas repetidas mediante el método clásico y con F ajustado.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
grupo	3	3.53515625	1.17838542	0.38	0.7728
Error	8	25.06250000	3.13281250		

Source	DF	Type III SS	Mean Square	F Value	Pr > F	Adj Pr > F G - G	Adj Pr > F H - F
tiempo	3	32.45182292	10.81727431	96.60	<.0001	<.0001	<.0001
tiempo*grupo	9	1.87630208	0.20847801	1.86	0.1082	0.1922	0.1491
Error(tiempo)	24	2.68750000	0.11197917				

Greenhouse-Geisser Epsilon	0.4316
Huynh-Feldt Epsilon	0.6729

Los resultados de la tabla 3 presenta dos columnas con los p-valores ajustados bajo estos dos criterios: G-G y H-F, ambas pruebas son llevadas a cabo estimando una cantidad conocida como ε (que de acuerdo al procedimiento de Greenhouse-Geisser este valor es estimado en 0.4316 y al procedimiento de Huynh-Feldt este valor es estimado en 0.6729) para luego ser multiplicado por el estadístico de prueba antes de obtener el p-valor.

A partir de los resultados de la tabla 3, la prueba F bajo los dos criterios de ajuste, muestra una interacción no significativa entre tiempo y lugar (p-valor>0.1), pero si una significancia para el factor tiempo (horas) indicando que el nivel de acidez (respuesta) cambia en al menos uno de los cuatro tiempos de medición (p-valor<0.0001), asimismo se observa no significancia entre los cuatro grupos experimentales observados (p-valor 0.7728)

3.3.- Análisis de contrastes polinomiales para el tiempo

Siendo que el tiempo es significativo y es un factor cuantitativo, una forma de observar la naturaleza de su tendencia es utilizando contrastes polinomiales y así describir el efecto de este factor. La tabla 4 muestra los anva para los contrastes polinomiales lineales, cuadrático y cúbico ajustados al factor tiempo, de este cuadro se observa que no existe interacción entre el contraste lineal de tiempo y el lugar (tratamiento); de la misma manera, las interacciones de tiempo cuadrático y tiempo cúbico con el lugar no son significativas, es posible por tanto concluir que no existe diferencias significativas entre los niveles de acidez promedio de las muestras de leche de los cuatro lugares.

Tabla 4. Análisis de varianza con contrastes polinomiales para las mediciones de acidez en las muestras de leche.

Repeated Measures Analysis of Variance
Analysis of Variance of Contrast Variables

tiempo_N represents the nth degree polynomial contrast for tiempo

Contrast Variable: tiempo_1 (lineal)

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Tiempo lineal	1	31.35651042	31.35651042	116.90	<.0001
Tiempo lineal*lugar	3	1.50703125	0.50234375	1.87	0.2125
Error	8	2.14583333	0.26822917		

Contrast Variable: tiempo_2 (cuadrático)

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Tiempo cuadrático	1	1.09505208	1.09505208	32.35	0.0005
Tiempo cuadrático*lugar	3	0.05598958	0.01866319	0.55	0.6614
Error	8	0.27083333	0.03385417		

Contrast Variable: tiempo_3 (cúbico)

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Tiempo cúbico	1	0.00026042	0.00026042	0.01	0.9323
Tiempo cúbico*lugar	3	0.31328125	0.10442708	3.08	0.0901
Error	8	0.27083333	0.03385417		

Sin embargo en los contrastes individuales vistos, las pruebas F se basan en varianzas del error con 8 grados de libertad, mientras que las pruebas para los mismos contrastes, a partir del análisis de parcelas divididas obtenidas en el cuadro 3, están basadas en varianzas del error con 24 grados de libertas. Además es posible percatarse que se tienen estimaciones diferentes de la varianza del error, por ejemplo en el análisis univariado el cuadrado medio del error es de 0.1119, con el contraste de ajuste lineal la estimación de la varianza del error es de 0.2682 es decir más del doble.

3.4.- Análisis utilizando modelos mixtos y estructuras de covarianzas entre las medidas repetidas

Con estructura de covarianza: simetría compuesta

Las salidas para este análisis en SAS® son presentadas en el cuadro 5, los valores estimados son:

$$\hat{R}_0 = \begin{bmatrix} 0.07465 & 0 & 0 & 0 \\ 0 & 0.07465 & 0 & 0 \\ 0 & 0 & 0.07465 & 0 \\ 0 & 0 & 0 & 0.07465 \end{bmatrix} \quad \hat{\sigma}_u^2 = 0.5035 \; .$$

Luego $\text{var}(y_{ijk}, y_{ijk'}) = \sigma_u^2 + \sigma_e^2 = 0.57815$ y

$$\text{cov}(y_{ijk}, y_{ijk'}) = \sigma_u^2 = 0.5035$$

El test de la razón de verosimilitud del modelo nulo, compara el ajuste de la estructura de covarianza de simetría compuesta con un modelo con independencia de errores en su estructura (modelo clásico). Siendo el p-valor <0.0001 se puede concluir que el modelo de independencia no se ajusta tan bien como el de simetría compuesta.

Tabla 5. Análisis de modelos mixtos estimando una estructura de simetría compuesta en la correlación entre las mediciones.

Covariance Parameter Estimates		
Cov Parm	Subject	Estimate
CS	unid(grupo)	0.5035
Residual		
Null Model	Likelihood	Ratio Test
DF	Chi-Square	Pr > ChiSq
1	58.28	<.0001

Con estructura de covarianza: autorregresivo de primer orden AR(1)

Las salidas para este análisis en SAS® son presentadas en la tabla 6, los parámetros estimados de la estructura de covarianza son:

$$\hat{R}_0 = \frac{0.6535}{(1-0.9395^2)} \begin{bmatrix} 1 & 0.9395 & 0.9394^2 & 0.9395^3 \\ 0.9395 & 1 & 0.9395 & 0.9395^2 \\ 0.9395^2 & 0.9395 & 1 & 0.9395 \\ 0.9395^3 & 0.9395^2 & 0.9395 & 1 \end{bmatrix}$$

En el test de la razón de verosimilitud del modelo nulo, p-valor <0.0001 concluyendo entonces que el modelo de independencia no se ajusta tan bien como el autorregresivo de primer orden.

Tabla 6.- Análisis de modelos mixtos estimando una estructura de covarianza autoregresiva de primer orden entre las mediciones.

Covariance Parameter Estimates		
Cov Parm	Subject	Estimate
AR(1)	unid(grupo)	0.9395
Residual		0.6535
Null Model	Likelihood	Ratio Test
DF	Chi-Square	Pr > ChiSq
1	71.23	<.0001

Con estructura de covarianza: Toeplitz

Las salidas para este análisis son presentadas en la tabla 7, los parámetros estimados de la estructura de covarianza es:

$$\hat{R}_0 = \sigma_e^2 \begin{bmatrix} 1 & 0.6756 & 0.6411 & 0.5835 \\ 0.6756 & 1 & 0.6756 & 0.6411 \\ 0.6411 & 0.6756 & 1 & 0.6756 \\ 0.5835 & 0.6411 & 0.6756 & 1 \end{bmatrix} \quad \text{y}$$

$\hat{\sigma}_e^2 = 0.7131$.

En el test de la razón de verosimilitud del modelo nulo, p-valor <0.0001 concluyendo que el modelo de independencia no se ajusta tan bien como con el de una estructura Toeplitz.

Tabla 7.- Análisis de modelos mixtos estimando una estructura de covarianza Toeplitz de las mediciones.

Covariance Parameter Estimates		
Cov Parm	Subject	Estimate
TOEP(2)	unid(grupo)	0.6756
TOEP(3)	unid(grupo)	0.6411
TOEP(4)	unid(grupo)	0.5835
Residual		0.7131

Null Model Likelihood Ratio Test		
DF	Chi-Square	Pr > ChiSq
3	73.39	<.0001

Con covarianza no estructurada.

Las salidas para este análisis en SAS® son presentadas en la tabla 8, los parámetros estimados para esta estructura son:

$$\hat{R}_0 = \begin{bmatrix} 0.2778 & 0.3264 & 0.3819 & 0.5 \\ 0.3264 & 0.4028 & 0.4549 & 0.6181 \\ 0.3819 & 0.4549 & 0.5625 & 0.7396 \\ 0.5 & 0.6181 & 0.7396 & 1.0694 \end{bmatrix}$$

Al comparar con el modelo de independencia entre las mediciones, (test de la razón de verosimilitud nulo), concluimos que el modelo con matriz de covarianza no estructurada se ajusta mejor a los datos que con el modelo de independencia (p-valor < 0.0001).

Tabla 8. Análisis de modelos mixtos sin una estructura de covarianza de las mediciones.

Covariance Parameter Estimates		
Cov Parm	Subject	Estimate
UN(1,1)	unid(grupo)	0.2778
UN(2,1)	unid(grupo)	0.3264
UN(2,2)	unid(grupo)	0.4028
UN(3,1)	unid(grupo)	0.3819
UN(3,2)	unid(grupo)	0.4549
UN(3,3)	unid(grupo)	0.5625
UN(4,1)	unid(grupo)	0.5000
UN(4,2)	unid(grupo)	0.6181
UN(4,3)	unid(grupo)	0.7396
UN(4,4)	unid(grupo)	1.0694

Null Model Likelihood Ratio Test		
DF	Chi-Square	Pr > ChiSq
9	108.44	<.0001

3.5.- Selección de un modelo de covarianza apropiado.

Para la selección de un modelo con una estructura en particular se estudian las medidas de bondad de ajuste que proporciona el SAS®, la tabla 9, resume las medidas para cada uno de los cuatro modelos de covarianza estimados.

Tabla 9. Medidas de bondad de ajuste para el modelo mixto con cuatro estructuras de varianza.

Fit Statistics	(a) sim.comp.	(b) AR(1) toepl.	(c) no estruct.	(d)
Log Likelihood	-25.8	-19.3	-18.3	-0.7
Akaike's Information Criterion (AIC)	-27.8	-21.3	-22.3	-10.7
Schwarz's Bayesian Criterion (BIC)	-86.5	-22.3	-23.3	-13.2
-2 Log Likelihood (-2LL)	51.6	38.7	36.5	1.5

Se observa que el modelo con estructura de simetría compuesta (a) es el que presenta medidas de ajuste más altos que las otras tres estructuras probadas. Los modelos con estructuras (b) y (c) tienen medidas de ajustes bastante similares. El cuarto modelo sin estructura de covarianza (d), es el que proporciona un menor valor para AIC y BIC, (éstas medidas considera el número de parámetros estimados, a diferencia de las otras dos medidas presentadas), por tanto es seleccionando este modelo como de mejor ajuste entre los cuatro modelos estimados

3.6.- Análisis de la interacción.

La prueba F de los efectos fijos que proporciona SAS® por defecto, utiliza los grados de libertad del denominador como si se tratara de un modelo sin correlación entre las medidas, sin embargo los grados de libertad de la prueba F en el ANVA son a menudo afectados por estructuras de covarianzas más complejas de los errores. SAS® incluye en sus rutinas la corrección de Kenward's y Rogers para el estadístico F.

Tabla 10. Análisis de varianza para los efectos fijos con el ajuste de Kendward´s Rogers para el estadístico F.

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
grupo	3	12	0.56	0.6489
tiempo	3	10	57.95	<.0001
grupo*tiempo	9	12.8	2.9	0.0406

A partir de estos resultados y con un nivel de significancia del 5% no es posible descartar la presencia de interacción entre tratamiento y tiempo, indicando que el nivel medio de acidez de las muestras de leche es diferente en al menos un lugar a lo largo del tiempo.

4. Conclusiones

A partir de los resultados obtenidos podemos concluir:

Existe diferencias entre los niveles promedios de acidez para las muestras de leche del lugar Cerro, al observar el gráfico de perfiles figura1.

Es necesario ajustar el estadístico F del análisis de varianza de parcelas divididas de medidas repetidas ya que no se cumplió la condición de Huynh-Feldt según la prueba de Mauchly que resultó significativa (p-valor<0.0001).

No existe interacción entre tiempo y lugar (p-valor>0.1) según los resultados obtenidos del análisis de varianza de parcelas divididas (con estadístico F ajustado).

Las mediciones de muestras de leche muestran correlación entre los tiempos, justificando el estudio con el análisis de modelos lineales mixtos (cuadro 1).

Al evaluar el ajuste del modelo lineal mixto con cuatro estructuras de correlación a los datos, se concluye que el modelo sin una estructura específica de correlación (no estructurada) entre las mediciones y no correlación entre las unidades experimentales, se ajusta mejor a los datos, puesto que se obtiene mejores medidas de ajuste (AIC=-10.7 y BIC=-13.2), pese al mayor número de parámetros que requiere su estimación en comparación con las estimaciones de las otras tres estructuras.

No presenta significancia la interacción entre lugar y tiempo con el análisis de parcelas divididas (p-valor=0.182) ni con el ajuste del estadístico F (G-G p-valor=0.1922 y H-F p-valor=0.1491). Sin embargo al utilizar modelos mixtos con una covarianza "no estructurada", no es posible descartar la presencia de interacción entre lugar y tiempo (p-valor=0.0406) si se considera un nivel de significancia del 5%.

5. Referencias bibliográficas

Littell R., Williken G., Stroup W. Wolfinger D., Schabenberger (2006). SAS® for mixed models. Segunda. Cary, NC: SAS Institute Inc

Kuehl R. (2001). Diseño de Experimentos. Thomson. 2da edición edición. México.

McCulloch C., Searle S. (2001) Generalizad, Linear and Mixed Models. John Wiley & Sons.

Cnaan, A., Laird, N.M. y Slasor, P. (1997). Using the general linear mixed model to analyze unbalanced repeated measures and longitudinal data. Statistics in Medicine, 16, 2.349-2.380.

Cole, V. y Grizzle, J (1996). Applications of multivariate análisis of variante to repeated measures experiments. Biometrics 41, 505-514.

Wolfinger, R. D. (1996). "Heterogeneous Variance Covariance Structures for Repeated Measures."Journal of Agricultural, Biological, and Environmental Statistics 1(2): 205–230.

Littell, R., Milliken, G., Stroup, W. y Wolfinger, D. (1996). SAS System for mixed models. Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1991) SAS® System for Linear Models. Tercera edición. Cary, NC: SAS Institute Inc.

Jensen, D. R. (1982). Efficiency and robustness in the use of repeated measurements. Biometrics, 38, 813-825.

Huynh, H. y Feldt, L. (1970). Conditions under which mean square ratios in repeated measurement designs have exact F-Distribution. Journal of the American Statistical Association, 65, 1.582-1.589.

Greenhouse, S. W. & Geisser, S. (1959). On methods in the analysis of profile data. Psychometrika, 24: 95-112.

Cochran, W. y Cox, M. (1957) Experimental Design, New Cork. John Wiley & Sons, Inc.

ANEXO

Sentencias en SAS trabajadas.

```
proc print data=sas1;
run;
***creando base de datos***;
data sas1_long;
set sas1;
array t(4) t1-t4;
do tiempo=1 to 4;
score=t(tiempo);
output;
end;
drop t1-t4;
run;
proc print data=sas1_long;
run;
***graficando las medias de cada tratamiento en el tiempo***;
proc sql;
create table ave as
select distinct mean (score) as score, tiempo as tiempo, grupo as grupo
from sas1_long
group by grupo, tiempo;
quit;
proc print data=ave;
run;
symbol1 i=std1mjt v=square c=red w=1.5;
symbol2 i=std1mjt v=circle c=blue w=1.5;
symbol3 i=std1mjt v=star c=green w=1.5;
symbol4 i=std1mjt v=plus c=orange w=1.5;
proc gplot data=ave;
plot score*tiempo=grupo;
run;
quit;
***matriz de correlaciones entre mediciones***;
proc corr data=sas1 nosimple noprob;
var t:;
run;
***analisis univariado de medidas repetidas***;
proc glm data=sas1;
class grupo;
```

```
model t1 t2 t3 t4=grupo/nouni;
repeated tiempo/nom printe;
run;
***contrastes polinomiales***;
proc glm data=sas1;
class grupo;
model t1 t2 t3 t4=grupo/nouni;
repeated tiempo polynomial /summary nom nou;
run;

***estimacion del modelo con estructura de cov.:
simetria compuesta***;
proc mixed data=sas1_long method=ML;
class grupo unid tiempo;
model score=grupo tiempo grupo*tiempo;
repeated tiempo /sub=unid(grupo) type=cs;
run;
***estimacion del modelo con estructura de cov.:
autorregresivo (1)***;
proc mixed data=sas1_long method=ML;
class grupo tiempo unid;
model score=grupo tiempo grupo*tiempo;
repeated tiempo /sub=unid(grupo) type=ar(1);
run;
proc mixed data=sas1_long method=ML;
class grupo tiempo unid;
model score=grupo tiempo grupo*tiempo;
repeated tiempo /sub=unid(grupo) type=ar(1);
random unid(grupo);
run;
***estimacion del modelo con estructura de cov.:
toeplitz***;
proc mixed data=sas1_long method=ML;
class grupo tiempo unid;
model score=grupo tiempo grupo*tiempo ;
repeated tiempo /sub=unid(grupo) type=toep;
run;
***estimacion del modelo con estructura de cov.: no
estructurada***;
proc mixed data=sas1_long method=ML;
class grupo tiempo unid;
model score=grupo tiempo grupo*tiempo;
repeated tiempo /sub=unid(grupo) type=un;
run;
**estimacion del modelo con estructura de cov.: no
estructurada***
****con ajuste de los grados de libertas para prueba
F de efectos fijos***;
proc mixed data=sas1_long method=ML;
class grupo tiempo unid;
model score=grupo tiempo grupo*tiempo /ddfm=kr;
repeated tiempo /sub=unid(grupo) type=un;
run;

Datos:
data sas1;
input unid grupo$ t1-t4;
datalines;
```

1	trapiche	16	16.5	17	18
2	trapiche	16.5	17	18	19.5
3	trapiche	15.5	16	16.5	17
4	cerro	17	17.5	19	20.5
5	cerro	15	15	16	16.5
6	cerro	16	16.5	18	19.5
7	huatocay	16	16.5	16.75	18
8	huatocay	16	16.5	17	17.5
9	huatocay	15	15.5	15.75	17
10	macas	16	16	17	17.5
11	macas	16	16.5	17	18
12	macas	16.5	17	17.25	18.5